INTRODUCTION TO ERGODIC THEORY. ERGODIC THEOREMS

1. Measure-preserving systems and ergodicity

Let X be a compact metric space, let $T : X \to X$ be a surjective measurable map (it will usually be continuous), and let μ be a regular Borel probability measure on X. The compactness of X is not essential, but every space we work with in this course is compact and so we make this assumption. The fact that $\mu(X) < \infty$ is much more important and most of our theorems fail completely in infinite measure spaces.

We say that the measure μ is T-invariant if $\int f\, d\mu = \int f \circ T\, d\mu$ for all measurable f, or equivalently if $\mu(T^{-1}(E)) = \mu(E)$ for all measurable sets E. That the first condition implies the second is obvious upon taking $f := 1_E$; to see that the second implies the first, approximate f by simple measurable functions and take limits. Another, slightly higher-brow, way of saying the same thing is that the map $U_T : L^2(X) \to L^2(X)$ defined by $f \mapsto f \circ T$ is an *isometry*. The fact that $\langle U_T f, U_T g\rangle = \langle f, g\rangle$ is immediate from the definition of T-invariance; to deduce T-invariance from the fact that U_T is an isometry, simply set $g = 1$.

A triple (X, μ, T) is called a *measure-preserving system*, and is the basic object of study in ergodic theory. Sometimes the underlying σ-algebra $\mathcal{F}$ for the measure μ is emphasised, but for us this will always be the Borel σ-algebra.

It is possible to prove the following nice recurrence result without additional assumptions on the system.

Theorem 1.1 (Poincáre Recurrence Theorem)**.** *Let $E \subseteq X$ be a measurable set and let $E' \subseteq E$ be the set of $x \in E$ for which there are infinitely many $n \geqslant 1$ with $T^n x \in E$. Then $\mu(E \setminus E') = 0$, or in other words almost all points in E recur back to E.*

Proof. Write $A_N := \bigcup_{n \geqslant N} T^{-n}E$ and set $S := \bigcap_N A_N$. Then $E' = E \cap S$. Now we have the nesting $A_0 \supseteq A_1 \supseteq \dots$, and by the T-invariance of μ we have $\mu(A_{N+1}) = \mu(T^{-1}A_N) = \mu(A_N)$. By the monotone convergence theorem it follows that $\mu(S) = \mu(A_0)$, and so $\mu(A_0 \setminus S) = 0$. But $E \subseteq A_0$, and therefore $\mu(E \setminus S)$ is zero as well. □

To prove more interesting results we focus on a very important class of transformations. The following definition is perhaps the most important one in the course.

Definition 1.2 (Ergodicity)**.** We say that T is ergodic, or that the measure μ is ergodic for T, if all T-invariant sets (that is to say measurable sets E with $T^{-1}E = E$), have measure 0 or 1.

It is not hard to see that T is ergodic if and only if the only T-invariant measurable functions (functions f satisfying $f = f \circ T$) are the functions which are constant almost everywhere. On the second example sheet I ask for a detailed proof of this fact.

Ergodicity seems like a rather weak property. Remarkably, it is exactly the property that allows us to make rigorous assertions of the form "times averages approximate space averages".

Let us first give some examples of ergodic measures.

Proposition 1.3 (Irrational circle rotations are ergodic)**.** *Suppose that α is irrational. Then the circle rotation $R_\alpha : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ is ergodic with respect to the Lebesgue measure μ.*

First proof. Suppose that $f : X \to \mathbb{R}$ is an R_α-invariant measurable function which is not constant a.e.. Each level set $\{x : s \leqslant f(x) \leqslant t\}$ must also be R_α-invariant: by passing to one of these level sets which is not constant a.e. we may assume that $f \in L^1(X)$. Let $\varepsilon > 0$. Since the space $C(X)$ of continuous functions is dense in $L^1(X)$, we may find a continuous function $\tilde{f}$ with $\|f - \tilde{f}\|_1 \leqslant \varepsilon$. Applying R_α^n and using the supposed invariance of f together with the rotation-invariance of $\|\cdot\|_1$, we have $\|f - \tilde{f} \circ R_\alpha^n\|_1 \leqslant \varepsilon$ for all n, and hence $\|\tilde{f} - \tilde{f} \circ R_\alpha^n\|_1 \leqslant 2\varepsilon$. Since the orbit $R_\alpha^n 0$ is dense in $\mathbb{R}/\mathbb{Z}$, it follows from this and the continuity of $\tilde{f}$ that in fact $\|\tilde{f} - \tilde{f} \circ R_t\| \leqslant 2\varepsilon$ for all $t \in \mathbb{R}/\mathbb{Z}$. Writing $c(\tilde{f})$ for the constant function $\int \tilde{f}\, d\mu$, this implies that

$$\begin{aligned}\|\tilde{f} - c(\tilde{f})\|_1 &= \int \big|\tilde{f}(x) - \int \tilde{f}(x+t)\, d\mu(t)\big|\, d\mu(x) \\ &\leqslant \int |\tilde{f}(x) - \tilde{f}(x+t)|\, d\mu(t) d\mu(x) \\ &\leqslant 2\varepsilon.\end{aligned}$$

Thus we have $\|f - c(\tilde{f})\|_1 \leqslant 3\varepsilon$. This implies that

$$|c(f) - c(\tilde{f})| = \big|\int (f - c(\tilde{f}))\, d\mu\big| \leqslant 3\varepsilon,$$

and hence by the triangle inequality $\|f - c(f)\|_1 \leqslant 6\varepsilon$. However $\varepsilon > 0$ was arbitrary, and so $\|f - c(f)\|_1 = 0$ which implies that $f = c(f)$ a.e., contrary to assumption.

Second proof. Replacing f with a level set if necessary, as before, we may assume that there is a T-invariant function $f \in L^2(X)$ which is not constant a.e.. By standard harmonic analysis (see the first example sheet) we have $\sigma_M f(\theta) := \sum_{|r| \leqslant M} a_r e^{2\pi i r\theta} \to f$ in L^2 as $M \to \infty$, where $a_r = \widehat{f}(r) := \int_0^1 f(\theta) e^{-2\pi i r\theta}\, d\theta$. Now it is clear that $(\sigma_M f) \circ T(\theta) = \sum_{|r| \leqslant M} a_r e^{2\pi i \alpha r} e^{2\pi i r\theta}$. But we have $(\sigma_M f) \circ T \to f \circ T = f$ in L^2, and hence

$$\|\sum_{|r| \leqslant M} a_r (1 - e^{2\pi i \alpha r}) e^{2\pi i r\theta}\|_{L^2(d\theta)} \to 0.$$

However by orthogonality we have

$$\|\sum_{|r| \leqslant M} b_r e^{2\pi i r\theta}\|_{L^2(d\theta)} = \sum_r |b_r|^2$$

for any $b_r \in \mathbb{C}$. Since $\alpha \notin \mathbb{Q}$ we have $e^{2\pi i r\alpha} \neq 1$ when $r \neq 0$, and so the only way to reconcile these two facts is to have $a_r = 0$ whenever $r \neq 0$. That is, f is constant a.e.□

Remark. The "spectral" proof that R_α is ergodic was rather cleaner than the more geometric first proof. That was because we had access to such a nice explicit basis of

$L^2(X)$ which interacted with translations, namely the collection of exponentials $e^{2\pi i r\theta}$. The more hands-on geometric proofs tend to be easier to generalise, even when the transformation T is given by a group action as here.

Proposition 1.4 (Doubling map is ergodic)**.** *The doubling map $T : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ is ergodic with respect to the Lebesgue measure μ.*

First proof. Let $D_{a,n}$ be a "standard dyadic interval at scale n", that is to say an interval of type $(\frac{a}{2^n}, \frac{a+1}{2^n})$, $a \in \mathbb{Z}$. If E is any measurable set then it is not hard to check that $\mu(T^{-n}E \cap D_{a,n}) = 2^{-n}\mu(E)$. Thus if E is T-invariant then $\mu(E \cap D_{a,n}) = 2^{-n}\mu(E)$ (or in other words the relative density of E on $D_{a,n}$ is $\mu(E)$). Unless $\mu(E) = 0$ or 1, this contradicts the Lebesgue density theorem. We may also argue directly, noting from the preceding discussion that if $\tilde{U}$ is any finite union of standard dyadic intervals then $\mu(E \cap \tilde{U}) = \mu(E)\mu(\tilde{U})$. Now for any $\varepsilon > 0$ there is an open set U with $E \subseteq U$ and $\mu(U \setminus E) \leqslant \varepsilon$. For each n write U_n for the union of all the standard dyadic intervals at scale n which are contained in U. We clearly have $U_1 \subseteq U_2 \subseteq \dots$, and since U is open we have $\bigcup_n U_n = U$. By the monotone convergence theorem we therefore have

$$\mu(E) = \mu(E \cap U) = \lim_{n\to\infty} \mu(E \cap U_n) = \mu(E) \lim_{n\to\infty} \mu(U_n) = \mu(E)\mu(U).$$

Thus if $\mu(E) \neq 0$ then $\mu(U) = 1$, and hence $\mu(E) \geqslant \mu(U) - \varepsilon = 1 - \varepsilon$. Since ε was arbitrary we must have $\mu(E) = 1$.

Second proof. Suppose that $f = f \circ T$, where $f \in L^2(X)$. Using Fourier analysis, with the same notation as before, we conclude that

$$(\sigma_M f) - (\sigma_M f) \circ T = \sum_{|r| \leqslant 2M} (1_{|r| \leqslant M} a_r - 1_{r \equiv 0 (\mathrm{mod}\, 2)} a_{r/2}) e^{2\pi i r\theta}.$$

Since this tends to 0 in L^2 as $M \to \infty$, we must have $a_r = a_{r/2}$ for all even $r \neq 0$ and $a_r = 0$ for all odd r. This forces $a_r = 0$ for all $r \neq 0$, whence f is constant a.e. □

2. Time averages vs space averages.

Let (X, μ, T) be a measure-preserving system with $\mu(X) = 1$, and suppose that $T : X \to X$ is ergodic. An "ergodic theorem" is, broadly speaking, any result stating that the time averages $S_N f := \mathbb{E}_{0 \leqslant n \leqslant N-1} f(T^n x)$ converge to the space average $\overline{f} := \int f \, d\mu$. There are various notions of convergence of functions that we might consider. Let us recall some of the most common.

Definition 2.1 (Modes of convergence)**.** Suppose that (X, μ) is a compact measure space. and suppose that $(f_N)_{N=1}^\infty$ is a sequence of real-valued functions. Suppose that f is another real-valued function. Then we say that

(i) $f_N \to f$ weakly in L^2 if, for any $g \in L^2(X)$, we have $\langle f_N - f, g \rangle \to 0$.
(ii) $f_N \to f$ in L^2 if $\|f_N - f\|_2 \to 0$.
(iii) More generally, if $p \geqslant 1$, we say that $f_N \to f$ in L^p if $\|f_N - f\|_p \to 0$. We will usually take $1 \leqslant p \leqslant 2$.
(iv) We say that $f_N \to f$ pointwise a.e. if $f_N(x) \to f(x)$ for all x outside a set of measure 0.

To orientate ourselves, let us establish some relations between these notions of convergence.

If $f_N \to f$ in L^2 then $f_N \to f$ weakly in L^2, since

$$\langle f_N - f, g\rangle \leqslant \|f_N - f\|_2 \|g\|_2$$

by the Cauchy-Schwarz inequality. The converse is not true, however. Indeed take $X = \mathbb{R}/\mathbb{Z}$ and $f_N(x) = e^{2\pi i N x}$. By Bessel's inequality we have $\sum_{N=1}^{\infty} |\langle f_N, g\rangle|^2 \leqslant \|g\|_2^2$, and so $\langle f_N, g\rangle \to 0$ as $N \to \infty$ for any fixed $g \in L^2(X)$. Thus $f_N \to 0$ weakly in L^2. It is clear that this convergence does not take place in L^2.

If $1 \leqslant p < 2$, convergence in L^p is weaker than convergence in L^2. This follows easily from the nesting of norms.

Pointwise convergence is "morally" the strongest notion of all. If there is some function $g \in L^p(X)$ such that $|f_n(x)|, |f(x)| \leqslant g(x)$ for almost all x then by the dominated convergence theorem we have that $f_n \to f$ in L^p. Indeed by convexity we have

$$|f_n(x) - f(x)|^p \leqslant 2^{p-1}(|f_n(x)|^p + |f(x)|^p) \leqslant 2^p g(x)^p,$$

an integrable function.

Pathologies can occur, however: the sequence of functions $f_n : [0,1] \to \mathbb{R}$ defined by $f_n(x) = n$ if $0 \leqslant x \leqslant 1/n$ and $f_n(x) = 0$ otherwise converges to zero a.e. but f_n does not converge to zero in any L^p norm, $p \geqslant 1$.

3. Time averages vs space averages in weak L^2

In this section we remark on the weakest theorem that might be called an ergodic theorem.

Theorem 3.1. *Suppose that (X, μ, T) is an ergodic measure-preserving system. Suppose that $f \in L^2(X)$. Then the time averages $S_N f(x) := \mathbb{E}_{0 \leqslant n < N} f(T^n x)$ converge to the constant function $\overline{f} := \int f \, d\mu$ in weak L^2.*

The proof uses an instance of the Banach-Alaoglu theorem, in this case the fact that the unit ball of $L^2(X)$ is compact in the weak topology (more accurately, the theorem tells us that the unit ball of $L^2(X)^*$ is compact in the weak*-topology, but L^2 is isomorphic to its own dual). This can be proved using Tychonov's theorem (see any functional analysis textbook), or alternatively by a diagonalisation argument almost identical to the one alluded to in the notes *A very brief review of measure theory.* The key point is that $L^2(X)$ is separable.

Proof. Suppose without loss of generality that $\|f\|_2 = 1$. Since the map $U_T : L^2(X) \to L^2(X)$ is an isometry we have

$$\|S_N f\|_2 = \|\frac{1}{N}(f + U_T f + \cdots + U_T^{N-1} f)\|_2 \leqslant 1, \tag{3.1}$$

that is to say $S_N f$ lies in the unit ball of $L^2(X)$. Note also that the averages $S_N f$ are almost T-invariant in the sense that

$$\|S_N f - S_N f \circ T\|_2 \leqslant \frac{1}{N}(\|f\|_2 + \|U_T^N f\|_2) \leqslant 2/N. \tag{3.2}$$

Suppose that some subsequence $(S_{N_k} f)_{k=1}^{\infty}$ converges weakly to $g \in L^2(X)$. From (3.2) we see that $(S_{N_k} f \circ T)_{k=1}^{\infty}$ also converges weakly to g, whence $g = g \circ T$ a.e.. Since T is ergodic, this implies that g is constant a.e.. Since

$$\int S_{N_k} f \, d\mu - \int g \, d\mu = \langle S_{N_k} f - g, 1 \rangle \to 0$$

and $\langle S_N f, 1 \rangle = \int S_N f \, d\mu = \overline{f}$ for all N, we must have $g = \overline{f}$.

We have shown that *any* subsequence of $(S_N f)_{N=1}^{\infty}$ which converges weakly in L^2 converges to $\overline{f}$. It is not hard to deduce from this and the compactness of the unit ball of $L^2(X)$ in the weak topology that $(S_N f)_{N=1}^{\infty}$ itself converges to $\overline{f}$ weakly in L^2. □

4. Time averages vs space averages in L^2: The mean ergodic theorem

The main theorem of this section completely supercedes that of the previous section, though the proof is more complicated.

Theorem 4.1 (Von Neumann's mean ergodic theorem for ergodic transformations)**.** *Suppose that (X, μ, T) is an ergodic m.p.s. Then for any $f \in L^2(X)$ we have $S_N f \to \overline{f}$ in L^2.*

There is, in fact, a version of the theorem for measure-preserving transformations which are not necessarily ergodic. Write $I \subseteq L^2(X)$ for the closed subspace consisting of T-invariant functions, and let $\pi(f)$ be the projection of f onto I.

Theorem 4.2 (Von Neumann's mean ergodic theorem again)**.** *Suppose that (X, μ, T) is a m.p.s. Then for any $f \in L^2(X)$ we have*

$$\|S_N f - \pi(f)\|_2 \to 0$$

as $N \to \infty$.

It is a simple matter to deduce Theorem 4.1 from 4.2, since in the ergodic case I consists of just the constant functions.

The proof of Von Neumann's ergodic theorem is most naturally discussed using the isometry $U_T : L^2(X) \to L^2(X)$ induced by T. In fact with little additional effort one can deal with contractions rather than simply isometries.

Theorem 4.3 (Von Neumann's mean ergodic theorem for Hilbert Spaces)**.** *Let H be a Hilbert space, and let $U : H \to H$ be a contraction (thus $\|Uf\| \leqslant \|f\|$ for all $f \in H$). Let $I \subseteq H$ be the closed subspace of U-invariant elements and let $\pi : H \to I$ be the projection map. Then the time averages $S_N f := \mathbb{E}_{0 \leqslant n < N} U^n f$ converge to $\pi(f)$ in H.*

Proof. Let us note that the adjoint U^* is also a contraction, since for any $f \in H$ we have

$$\|U^* f\|^2 = \langle U^* f, U^* f \rangle = \langle UU^* f, f \rangle \leqslant \|UU^* f\| \|f\| \leqslant \|U^* f\| \|f\|.$$

Furthermore if f is U-invariant then it is also U^*-invariant, since

$$\begin{aligned} \|f - U^* f\|^2 &= \|f\|^2 + \|U^* f\|^2 - \langle f, U^* f \rangle - \langle U^* f, f \rangle \\ &= \|f\|^2 + \|U^* f\|^2 - \langle Uf, f \rangle - \langle f, Uf \rangle \\ &= \|U^* f\|^2 - \|f\|^2 \leqslant 0. \end{aligned}$$

The key idea of the proof is to identify the orthogonal complement of I as the closed subspace spanned by *cocycles*. If $g \in H$ we write $\partial g = g - Ug$, and let M be the closed subspace of H spanned by all cocycles ∂g. It is clear that $I \subseteq M^\perp$, since if f is U-invariant we have

$$\langle \partial g, f\rangle = \langle g - Ug, f\rangle = \langle g, f\rangle - \langle g, U^* f\rangle = 0.$$

Suppose, conversely, that $f \in H$ is orthogonal to all cocycles. Then in particular we have $\langle f, \partial f\rangle = 0$. It follows that

$$\begin{aligned}\|f - Uf\|^2 &= \langle f, f - Uf\rangle + \langle f - Uf, f\rangle - \|f\|^2 + \|Uf\|^2 \\ &= -\|f\|^2 + \|Uf\|^2 \\ &\leqslant 0.\end{aligned}$$

Thus $f = Uf$, that is to say $f \in I$.

Now that we have this decomposition, let $f \in H$ be arbitrary. Then for any $\varepsilon > 0$ we may write

$$f = \pi(f) + \partial g + h$$

where $\|h\| \leqslant \varepsilon$. Taking time averages, we thus have

$$\|S_N f - \pi(f)\| \leqslant \|S_N(\partial g)\| + \|S_N h\|. \tag{4.1}$$

Since U is a contraction we have

$$\|S_N h\| \leqslant \varepsilon.$$

Now by telescoping the sum we see that

$$S_N(\partial g) = \frac{1}{N}(g - U^N g),$$

and so

$$\|S_N(\partial g)\| \leqslant \frac{2}{N}\|g\|.$$

Comparing with (4.1) we see that

$$\|S_N f - \pi(f)\| \leqslant \frac{2}{N}\|g\| + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, the result follows. □

We remark once again that Theorems 4.1 and 4.2 are simple consequences of this.

5. The almost-everywhere ergodic theorem

Pointwise ergodic theorems capture the essence of ergodic theory. We will give a more "nuts and bolts" proof than is customary; I find this gives more of a feeling for what is going on. Some pointers to slicker arguments are given on the second example sheet.

Students may wish to consult the notes *A primer on conditional expectation* at this point.

Theorem 5.1 (Birkhoff's almost-everywhere ergodic theorem)**.** *Suppose that (X, μ, T) is a m.p.s, and suppose that $f \in L^1(X)$. Let $\mathcal{F}_0$ be the σ-algebra of T-invariant measurable sets, and write $\pi : L^1(X) \to L^1(X)$ for the conditional expectation operator $f \mapsto \mathbb{E}(f|\mathcal{F}_0)$. Then $S_N f \to \pi(f)$ pointwise a.e. In particular if T is ergodic then the time-average $S_N f$ tends pointwise to the space-average $\overline{f} = \int f\, d\mu$ a.e.*

The key ingredient is a result called the *maximal ergodic theorem.* For any integer $L \geqslant 1$ we write $S_L^* f(x) := \max_{M \leqslant L} \mathbb{E}_{0 \leqslant m < M} f(T^m x)$. Write $S^* f(x) := \sup_L S_L^* f(x)$. In analysis jargon, the maximal ergodic theorem is the statement that the operator S^* satisfies a weak type $(1,1)$-inequality.

Proposition 5.2 (Maximal ergodic theorem)**.** *Let (X, μ, T) be a m.p.s. and let $\lambda > 0$. Then $\mu\{x : S^* f(x) \geqslant \lambda\} \leqslant 4\|f\|_1/\lambda$.*

Proof. We will apply the *Hardy-Littlewood maximal inequality* for functions $F : \mathbb{Z} \to \mathbb{R}$. Given such a function, define the (one-sided) maximal function F^* by

$$F^*(x) := \sup_{M \geqslant 1} \mathbb{E}_{0 \leqslant m < M} F(x+m).$$

The Hardy-Littlewood maximal inequality states, in analysis jargon, that the maximal operator $F \mapsto F^*$ is weak $(1,1)$-type. We remark that it is possible to generalise this theorem in many different ways, and there is a whole branch of analysis devoted to the study of such theorems. In this special case, however, the proof is quite simple.

Theorem 5.3 (Hardy-Littlewood maximal inequality for the line)**.** *Suppose that $\lambda > 0$. Then we have $\#\{x : F^*(x) \geqslant \lambda\} \leqslant 4\|F\|_{\ell^1(\mathbb{Z})}/\lambda$.*

Proof. Write $A \subseteq \mathbb{Z}$ for the set of all x such that $F^*(x) \geqslant \lambda$. Then associated to each $x \in A$ there is some $M = M(x) \geqslant 1$ such that $\mathbb{E}_{0 \leqslant m < M} F(x+m) \geqslant \lambda/2$. Write $I_x := \{x, \dots, x+M-1\}$; thus F has average at least $\lambda/2$ on each I_x.

At this point we use (a simple version of) the *Vitali covering lemma.* In this special case the result can be found in Hallard Croft's article in *Eureka 1968.* The claim is that there is a set $A' \subseteq A$ such that the intervals I_x, $x \in A'$ are disjoint, and such that $\sum_{x \in A'} |I_x| \geqslant \frac{1}{2}|\bigcup_{x \in A} I_x|$. Once this is proved, we have

$$\|F\|_{\ell^1(\mathbb{Z})} \geqslant \frac{\lambda}{2} \sum_{x \in A'} |I_x| \geqslant \frac{\lambda}{4} |\bigcup_{x \in A} I_x| \geqslant \frac{\lambda}{4}|A|,$$

which is the result we claimed.

It remains to prove the claim. First of all pass to a minimal subcollection $\{I_x : x \in A_0\}$ with the property that $\bigcup_{x \in A_0} I_x = \bigcup_{x \in A} I_x$. By a simple inspection this subcollection has the property that no point y lies in three of the I_x. Labelling the intervals as $I_j := \{a_j, \dots, b_j\}$ with $a_1 \leqslant a_2 \leqslant \dots \leqslant a_k$, it may now be seen that $b_1 \leqslant a_3, b_2 \leqslant a_4, \dots, b_{k-2} \leqslant a_k$. Thus the two collections $I_1 \cup I_3 \cup \dots$ and $I_2 \cup I_4 \cup \dots$ consist of disjoint intervals. Passing to whichever of these two collections has the greatest total measure, the result follows. □

We now let $\lambda > 0$ and return to the proof of the maximal ergodic theorem. Let $N \gg L \geqslant 1$ be integers. Let $x \in X$ be fixed. We shall apply the Hardy-Littlewood maximal inequality to a suitably truncated version of the function $n \mapsto f(T^n x)$, specifically

$F(n) := f(T^n x)$ if $0 \leqslant n < N+L$ and $F(n) = 0$ otherwise. We obtain, writing out the definition of F^*, the inequality

$$\#\{n \in [N] : \max_{M \leqslant L} \mathbb{E}_{0 \leqslant m < M} f(T^{n+m}x) \geqslant \lambda\} \leqslant \frac{4 \sum_{0 \leqslant j < N+L} |f(T^j x)|}{\lambda}.$$

Integrating over $x \in X$ and using the T-invariance of μ, we obtain

$$\int_X \#\{n \in [N] : \max_{M \leqslant L} \mathbb{E}_{0 \leqslant m < M} f(T^{n+m}x) \geqslant \lambda\}\, dx \leqslant \frac{C(N+L)\|f\|_1}{\lambda}.$$

Now the indicator function of the event $\max_{M \leqslant L} \mathbb{E}_{0 \leqslant m < M} f(T^{n+m}x) \geqslant \lambda$ is a measurable function of n and x on the space $\mathbb{Z} \times X$, so by Fubini's theorem we may swap the order and obtain

$$\sum_{n \in [N]} \mu\{x : \max_{M \leqslant L} \mathbb{E}_{0 \leqslant m < M} f(T^{n+m}x) \geqslant \lambda\} \leqslant \frac{C(N+L)\|f\|_1}{\lambda}.$$

By the T-invariance of μ we may drop the T^n, obtaining

$$N\mu\{x : \max_{M \leqslant L} \mathbb{E}_{0 \leqslant m < M} f(T^{m}x) \geqslant \lambda\} \leqslant \frac{C(N+L)\|f\|_1}{\lambda},$$

or in other words

$$\mu\{x : S_L^* f(x) \geqslant \lambda\} \leqslant \frac{C(N+L)\|f\|_1}{\lambda N}.$$

Letting $L, N \to \infty$, with N growing much faster than L, we obtain

$$\mu\{x : S^* f(x) \geqslant \lambda\} \leqslant \frac{C\|f\|_1}{\lambda},$$

which is the maximal ergodic theorem albeit with a slightly weaker constant. □

Proof of the pointwise ergodic theorem. To prove the pointwise ergodic theorem it suffices to show that, for any $\varepsilon > 0$, the set

$$E_\varepsilon := \{x : \limsup_{N \to \infty} |S_N f(x) - \pi(f)| \geqslant \varepsilon\}$$

has measure zero.

Let $\delta > 0$ be a parameter. In the proof of the von Neumann ergodic theorem we took a function $f \in L^2(X)$ and decomposed it as

$$f = \pi(f) + \partial g + h, \tag{5.1}$$

where $g \in L^2(X)$ and $\|h\|_2 \leqslant \delta$.

In the pointwise ergodic theorem we are operating under the weaker assumption that $f \in L^1(X)$, and to deal with this we must replace (5.1) by a slightly more refined decomposition. However $L^2(X)$ is dense in $L^1(X)$ (in fact $C(X)$ is dense in $L^1(X)$), and so for any $\delta > 0$ we may find $f_0 \in L^2(X)$ with $\|f - f_0\|_1 \leqslant \delta$. Applying (5.1) to this function we obtain a decomposition

$$f_0 = \pi(f_0) + \partial g_0 + h_0$$

with $\|h_0\|_2 \leqslant \delta$. Thus

$$f = \pi(f) + \partial g_0 + h_1,$$

where $h_1 := h_0 + (f - f_0) - (\pi(f) - \pi(f_0))$. Note that $\|h_1\|_1 \leqslant \|h_0\|_1 + 2\delta \leqslant 3\delta$.

It is important to have some control on g_0, which at present is only known to lie in $L^2(X)$. Since $L^\infty(X)$ is dense in $L^2(X)$, we may find $g \in L^\infty(X)$ with $\|g - g_0\|_2 \leqslant \delta$. Our decomposition of f then becomes

$$f = \pi(f) + \partial g + h$$

where $h = h_1 + \partial(g_0 - g)$; thus $\|h\|_1 \leqslant 3\delta + \|\partial(g - g_0)\|_2^{1/2} \leqslant 3\delta + (2\delta)^{1/2}$. Redefining δ, which after all was arbitrary, we may suppose that $\|h\|_1 \leqslant \delta$.

We clearly have, for any $x \in X$,

$$|S_N f(x) - \pi(f)| \leqslant |S_N(\partial g)(x)| + |S_N h(x)|. \tag{5.2}$$

The first term on the right may be analysed as before; the sum telescopes and we have

$$|S_N(\partial g)(x)| = |\frac{1}{N}(g(x) - g(T^{n-1}(x))|,$$

which tends to zero for all x since $g \in L^\infty(X)$. Therefore E_ε is contained, up to a set of measure zero, in the set

$$\{x : \limsup_{N\to\infty} |S_N h(x)| \geqslant \varepsilon\}.$$

This is of course contained in the set

$$\{x : |S^* h(x)| \geqslant \varepsilon\}$$

which, by the maximal ergodic theorem and the bound for $\|h\|_1$, has measure at most $4\delta/\varepsilon$. Since δ was arbitrary, we see that $\mu(E_\varepsilon) = 0$ as required. □

6. References

This material will be found in most books on ergodic theory. It is not so common to find the maximal ergodic theorem proved via the Hardy-Littlewood maximal inequality, as we did here. This argument may be found in volume 3 of Stein and Shakarchi.

It is about time I mentioned the very nice, if somewhat condensed, book of W. Parry entitled *Topics in ergodic theory*. I have used this book repeatedly in preparing the course.