Stat 205B Lecture Notes of May 14, 2002

Prepared by Gábor Pete

Disclaimer: These lecture notes have been only lightly proofread. Please inform Prof. Peres `<peres@stat.berkeley.edu>` *of errors.*

1 The main ergodic theorems

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $T : \Omega \longrightarrow \Omega$ a measure preserving transformation, meaning two things: it is measurable, and $\mathbb{P}(T^{-1}(A)) = \mathbb{P}(A)$ for all $A \in \mathcal{F}$. Here we took the inverse because $T(A)$ is not necessarily measurable. The set of T-invariant sets, $\mathcal{I} = \{A \in \mathcal{F} : T^{-1}A \subseteq A\}$ is a sub-σ-algebra of $\mathcal{F}$. Note that $A = T^{-1}A$ almost everywhere for $A \in \mathcal{I}$. For any function $f \in L^1(\Omega)$, let us define the measurable functions

$$S_n f(\omega) := \sum_{j=0}^{n-1} f(T^j \omega), \quad M_n f(\omega) := \max_{1 \le j \le n} S_j f(\omega) \quad \text{and} \quad Mf(\omega) := \sup_{j \ge 1} S_j f(\omega).$$

The most important result in ergodic theory is the following Pointwise Ergodic Theorem due to G. Birkhoff (1931). He formulated and proved it stimulated by J. von Neumann's much simpler L^2 version.

Theorem 1.1. (Birkhoff's Pointwise Ergodic Theorem) *For $f \in L^1(\Omega)$,*

$$\frac{1}{n} S_n f(\omega) \xrightarrow{\text{a.s.}} \overline{f}(\omega) = \mathbb{E}(f \,|\, \mathcal{I})(\omega).$$

The original proof was 50 pages long. Then Kakutani and Yosida formulated their Maximal Ergodic Theorem in 1939, having a much cleaner 10 page proof, from which Birkhoff's theorem can be deduced nicely.

Theorem 1.2. (Maximal Ergodic Theorem) *For $f \in L^1(\Omega)$,*

$$\int_{\{Mf > 0\}} f \, d\mathbb{P} \ge 0.$$

Similar maximal inequalities are very important in all parts of analysis, and usually they can be considered as improvements to Markov's inequality. Just to mention two examples: we saw the different L^p maximal inequalities for (sub/super)martingales. The Hardy-Littlewood maximal function associated to a real valued Lebesgue-measurable function $f \in L^1_{\text{loc}}(\mathbb{R}^d, \lambda)$ is

$$Mf(x) = \sup_r S_r |f(x)| = \sup_r \frac{1}{\lambda B_r(x)} \int_{B_r(x)} |f(y)| \, d\lambda(y),$$

and the corresponding maximal inequality is

$$\lambda\{x : Mf(x) > \alpha\} \le \frac{3^d}{\alpha} \int |f(x)| \, d\lambda(x)$$

for any $\alpha > 0$.

Now we deduce Birkhoff's theorem from the Maximal Ergodic Theorem. It could be done more shortly, in one step, but instead we will first prove the existence of the a.s. limit, then will identify the limit.

Proof of Birkhoff's theorem. For $a < b$ rationals let us define the measurable set $\Omega_{a,b} = \{\liminf_n \frac{S_n f}{n} < a,\ \text{and}\ \limsup_n \frac{S_n f}{n} > b\}$. It is easy to check that $T\Omega_{a,b} \subseteq \Omega_{a,b}$. Now suppose that $\mathbb{P}(\Omega_{a,b}) > 0$ for some $a, b \in \mathbb{Q}$, and define $\tilde{\mathbb{P}}(\cdot) = \mathbb{P}(\cdot \,|\, \Omega_{a,b})$. It is clear that $\Omega_{a,b} \subseteq \{M(f-b) > 0\} \cap \{M(a-f) > 0\}$, so the Maximal Ergodic Theorem applied to $(\Omega_{a,b}, \mathcal{F}|_{\Omega_{a,b}}, \tilde{\mathbb{P}})$ gives

$$\int_{\Omega_{a,b}} (f-b)\, d\tilde{\mathbb{P}} \geq 0 \qquad \text{and} \qquad \int_{\Omega_{a,b}} (a-f)\, d\tilde{\mathbb{P}} \geq 0.$$

Summing up these inequalities we get $\int_{\Omega_{a,b}} (a-b)\, d\tilde{\mathbb{P}} \geq 0$, which contradicts to $a - b < 0$ and $\tilde{\mathbb{P}}(\Omega_{a,b}) = 1$. Thus we have

$$\mathbb{P}\left(\bigcup_{a<b\in\mathbb{Q}} \Omega_{a,b} \right) = 1,$$

which means a.s. convergence.

Now write $\overline{f} = \lim_n \frac{S_n f}{n}$ for this a.s. limit. What can this limit be? First of all, note that $\overline{f}$ is invariant: $\overline{f} = \overline{f} \circ T$. Or, to say the same thing differently: it is $\mathcal{I}$-measurable.

Lemma 1.3. *For any $f \in L^1(\Omega)$ and measure preserving T, $\int_\Omega f \circ T\, d\mathbb{P} = \int_\Omega f\, d\mathbb{P}$. More generally, for any invariant set $B \in \mathcal{I}$, $\int_B f \circ T\, d\mathbb{P} = \int_B f\, d\mathbb{P}$.*

Proof. The first statement is true for any $f = \mathbf{1}_A$, $A \in \mathcal{F}$, by the definition of a measure-preserving transformation. Then we can pass to general f's by the "standard machine": approximate $f \geq 0$ by step functions, and use the Monote Convergence Theorem, and then write $f \in L^1$ as $f = f_+ - f_-$. The second statement follows by noticing that $(f\mathbf{1}_B) \circ T = (f \circ T)\mathbf{1}_B$ a.s. for $B \in \mathcal{I}$. □

This lemma implies that $\int_B S_n f\, d\mathbb{P} = n \int_B f\, d\mathbb{P}$ for $B \in \mathcal{I}$. For $f \geq 0$ we can now apply Fatou's lemma to get $\int_B \overline{f}\, d\mathbb{P} \leq \int_B f\, d\mathbb{P}$, and for bounded f we can apply the Dominated Convergence Theorem to get $\int_B \overline{f}\, d\mathbb{P} = \int_B f\, d\mathbb{P}$. So it is reasonable to expect that $\overline{f} = \mathbb{E}(f \,|\, \mathcal{I})$, the unique $\mathcal{I}$-measurable function that gives the same integral on each invariant set as f.

To actually prove this claim, set $g = f - \mathbb{E}(f \,|\, \mathcal{I})$. Since $\mathbb{E}(f \,|\, I)$ is T-invariant, we have to prove that $\overline{g} = \lim \frac{S_n g}{n}$ equals 0 almost surely; note that we know the existence of the limit from the existence of $\overline{f}$. Let us proceed similarly as before: take $\Omega_\epsilon = \{\overline{g} > \epsilon\} \in \mathcal{I}$ for some $\epsilon > 0$, and consider the restriction of our dynamical system to Ω_ϵ. If $\mathbb{P}(\Omega_\epsilon) > 0$, then we have a decent measurable dynamical system, and the Maximal Ergodic Theorem gives $\int_{\Omega_\epsilon} (g - \epsilon)\, d\mathbb{P} \geq 0$. If $\mathbb{P}(\Omega_\epsilon) = 0$, then the same inequality is trivial. Hence

$$\epsilon\mathbb{P}(\Omega_\epsilon) \leq \int_{\Omega_\epsilon} g\, d\mathbb{P} = \int_{\Omega_\epsilon} \mathbb{E}(g \,|\, \mathcal{I})\, d\mathbb{P} = 0,$$

where in the first equality we used $\Omega_\epsilon \in \mathcal{I}$ and the definition of conditional expectation, while the second one follows simply from the definition of g. Thus we have $\mathbb{P}(\Omega_\epsilon) = 0$. Similarly, $\mathbb{P}(\overline{g} < -\epsilon) = 0$. These show that $\overline{g} = 0$ a.s., and the proof is complete. □

The general belief after 1939 was that the maximal theorem and Birkhoff's theorem were basically equivalent. As a shock came Garsia's three-line proof of the maximal theorem in 1965.

Garsia's proof of the Maximal Ergodic Theorem. The sets $\{M_n f > 0\}$ increase up monotonicly to $\{Mf > 0\}$, so by the Dominated Convergence Theorem it is enough to prove $\int_{\{M_n f > 0\}} f\, d\mathbb{P} \geq 0$. Note that $f + [M_{n-1}(f \circ T)]^+ = M_n f$, and $f + [M_n(f \circ T)]^+ \geq M_n f$, hence

$$\begin{aligned}\int_{\{M_n f>0\}} f\,d\mathbb{P} &\geq \int_{\{M_n f>0\}} \left\{M_n f - [M_n(f\circ T)]^+\right\} d\mathbb{P}\\ &= \int_{\{M_n f>0\}} \left\{[M_n f]^+ - [M_n(f\circ T)]^+\right\} d\mathbb{P}\\ &\geq \int_\Omega [M_n f]^+\,d\mathbb{P} - \int_\Omega [M_n(f\circ T)]^+\,d\mathbb{P} = 0,\end{aligned}$$

where in the last step we used Lemma 1.3 again. □

2 Information Theory

We closed the semester with a very brief introduction to information theory by Peter Ralph.

Given a random variable X on $(\Omega,\mathcal{B},\mathbb{P})$, and a sub-$\sigma$-algebra $\mathcal{F}\subseteq\mathcal{B}$, let us define the conditional probability, information and entropy as

$$\begin{aligned}\mathbb{P}(X\,|\,\mathcal{F}) &= \mathbb{E}\left(\mathbf{1}_{X^{-1}(X(\omega))}\,|\,\mathcal{F}\right),\\ I(X\,|\,\mathcal{F}) &= -\log\mathbb{P}(X\,|\,\mathcal{F}),\\ H(X\,|\,\mathcal{F}) &= \mathbb{E}I(X\,|\,\mathcal{F}).\end{aligned}$$

In particular, if X is a discrete variable, $\mathbb{P}(X=x)=p_x$, and $\mathcal{F}=\mathcal{B}$, then $\mathbb{P}(X\,|\,\mathcal{F})$ is the random variable $\omega\mapsto p_{X(\omega)}$, and

$$H(X)=H(X\,|\,\mathcal{F})=-\sum_x p_x\log p_x.$$

If $\mathcal{F}=\sigma(Y)$ for another r.v. Y, then

$$H(X\,|\,Y)=-\sum_{x,y}\mathbb{P}(X=x,Y=y)\log\mathbb{P}(X=x\,|\,Y=y)=H(X,Y)-H(Y).$$

So

$$H(X,Y)\leq H(X)+H(Y),$$

with equality if and only if X and Y are independent.

The most important theorem about entropy is probably the following:

Theorem 2.1. (Shannon – McMillan – Breiman) *For a stationary ergodic sequence $(X_n)_{-\infty}^{\infty}$ on a countable state space, with $H(X_0)<\infty$,*

$$-\frac{1}{n}\log I(X_1,\ldots,X_n)\xrightarrow[L^1]{\text{a.s.}}H=H(X_0\,|\,X_{-1},X_{-2},\ldots).$$

An equivalent reformulation in the case of $X_i\in S$, where S is a finite set: to capture $1-\epsilon$ probability mass of the possible outcomes $(X_1,\ldots,X_n)$, we need at least around $\exp((H\pm\epsilon)n)$ sequences. E.g. in the case of an i.i.d. uniform sequence, we have $H=\log|S|$, so we need almost all possible outcomes. Thus larger entropy can be interpreted as a larger degree of random independence in the sequence. Among all probability distributions X on a finite set, the uniform one has the largest entropy $H(X)$, and among all probability densities on $\mathbb{R}$ with $\mathbb{E}X^2=1$, the standard normal has this maximizing property.

It is also possible to define the entropy of a measure preserving transformation, which notion is central in the theory of dynamical systems.