ma414l3.tex
Lecture 3. 26.1.2012

9. The Borel-Cantelli lemmas and the zero-one law.

First, recall from Real Analysis the definition of the upper and lower limit, lim sup and lim inf, of a real sequence x_n:

$$\limsup x_n := \inf_n \sup_{k\geq n} x_k, = \lim_n \sup_{k\geq n} x_k$$

(the inf here is a lim as the sequence $\sup_{k\geq n} x_k$ is decreasing), and dually

$$\liminf x_n := \sup_n \inf_{k\geq n} x_k, = \lim_n \inf_{k\geq n} x_k.$$

Then $\limsup x_n = +\infty$ iff x_n is not bounded above, while if x_n is bounded above, $\limsup x_n$ is the unique real c such that for all $\epsilon > 0$

$$x_n \leq c + \epsilon \quad \text{for all large enough } n, \quad x_n \geq c - \epsilon \quad \text{for infinitely many } n,$$

and dually for lim inf. For background, see any good book on Real Analysis.

There is an exact analogy of this for sets, with union replacing sup and intersection replacing inf, for sets A_n rather than reals x_n. Write 'i.o.' for 'infinitely often'.

$$\limsup A_n := \cap_n \cup_{k\geq n} A_k = \{x : x \in A_n i.o\} \text{ or } \{A_n i.o.\},$$

$$\liminf A_n = \cup_n \cap_{k\geq n} A_k = \{A_n \text{ for all sufficiently large } n\}.$$

Then for the indicator functions, one has (check)

$$\limsup I_{A_n} = I_{\limsup A_n}, \qquad \liminf I_{A_n} = I_{\liminf A_n}.$$

Note. Write $f = O(g)$ for f/g bounded, $f = o(g)$ for $f/g \to 0$ (the notation is due to Edmund LANDAU (1877-1938)). By systematic use of the Landau O, o notation, and of limsup and liminf, one can eliminate all but the essential $\epsilon > 0$ from Analysis. One should do so: Analysis well done never uses a superfluous ϵ; then the ϵs that one meets are the signature of the hard proofs.

The following results are due to Borel in 1909, F. P. CANTELLI (1906-1985) in 1917.

Theorem (Borel-Cantelli lemmas). If A_n are events, $A := \limsup A_n = \{A_n \ i.o.\}$:
(i) If $\sum P(A_n) < \infty$, then $P(A) = 0$.
(ii) If $\sum P(A_n) = \infty$ and the A_n are independent, then $P(A) = 1$.

Proof. (i) $A = \limsup A_n = \cap_n \cup_{m=n}^{\infty} A_m$, so $A \subset \cup_{m=n}^{\infty} A_m$ for each n. So

$$P(A) \leq P(\cup_{m=n}^{\infty} A_m) \leq \sum_{m=n}^{\infty} P(A_m) \to 0 \qquad (n \to \infty)$$

(tail of a convergent series): $P(A) = 0$.
(ii) By the De Morgan laws, $A^c = \cup_n \cap_{m=n}^{\infty} A_m^c$. But for each n

$$\begin{aligned} P(\cap_{m=n}^{\infty} A_m^c) &= \lim_N P(\cap_{m=n}^{N} A_m^c) \qquad (\sigma\text{-additivity}) \\ &= \prod_{m=n}^{N} (1 - P(A_m)) \qquad (\text{independence}) \\ &\leq \prod_{m=n}^{N} \exp\{-P(A_m)\} \qquad (1 - x \leq e^{-x} \text{ for } x \geq 0) \\ &= \exp\{-\sum_{m=n}^{N} P(A_m)\} \to 0 \qquad (N \to \infty), \end{aligned}$$

as $\sum P(A_n)$ diverges. So $\cap_{m=n}^{\infty} A_m^c)$ is null. So their union $A^c = \cup_n \cap_{m=n}^{\infty} A_m^c$ is null, giving the result. //

Corollary (Zero-One Law). If the A_n are independent and $A := \limsup A_n$, $P(A) = 0$ or 1 according as $\sum P(A_n)$ converges or diverges.

10. Infinite product measures; replication and copies.

Independence corresponds to product measures; the construction of the product measure of two measure extends to finite products of measures by induction. We now consider the extension to infinite products. This will model generation of a sequence of independent identically distributed (iid) random variables, called *replications* or *copies*. Think of repeatedly tossing a coin, or repeatedly sampling in Statistics.

In fact the construction is a special case of a much more general construction (in which independence is not assumed), called the *Daniell-Kolmogorov* *theorem*, which we shall meet later in connection with Stochastic Processes (Ch. II). But for now, consider a sequence of measure spaces $(\Omega_n, \mathcal{A}_n, \mu_n)$, $n = 1, 2, \ldots)$. We form the cartesian product $\Omega := \Omega_1 \times \ldots \times \Omega_n \times \ldots$; thus Ω is the set of points $\omega = (\omega_1, \ldots, \omega_n, \ldots)$ – sequences whose nth elements ω_n is in Ω_n. Call a set $A \subset \Omega$ a *cylinder set* if it is of the form $A = A_1 \times \ldots \times A_n \times \ldots$, with all but finitely many of the A_n, say $A_{n_1}, \ldots, A_{n_k}$, equal to Ω_n. Define a measure μ on the class $\mathcal{C}$ of such cylinder sets by

$$\mu(A) := \mu_{n_1}(A_{n_1}) \times \ldots \times \mu_{n_k}(A_{n_k})$$

(thus $\mu(A)$ expresses independence on the cylinder sets). The measure μ extends uniquely to a measure on the σ-field $\mathcal{A} := \sigma(\mathcal{C})$ generated by the cylinder sets. The resulting probability space is called the *infinite product* of the coordinate probability spaces, written

$$(\Omega, \mathcal{A}, \mu) = \times_{n=1}^{\infty} (\Omega_n, \mathcal{A}_n, \mu_n).$$

Example: Infinite coin tossing and the uniform distribution.
Take the Lebesgue probability space $([0,1], \mathcal{L}, \mu)$ modelling the uniform distribution $U[0,1]$ on the unit interval (probability = length). For a random variable $X \sim U[0,1]$, take its dyadic expansion

$$X = \sum_{1}^{\infty} \epsilon_n / 2^n.$$

Thus $\epsilon_1 = 0$ iff $X \in [0, 1/2)$, 1 iff $X \in [1/2, 1)$ (or $[1/2, 1]$: we can omit 1, as it carries 0 probability). If $\epsilon_1, \ldots, \epsilon_{n-1}$ are already defined, on the dyadic intervals $[k/2^{n-1}, (k+1)/2^{n-1})$, split each interval into two halves: $\epsilon_n = 0$ on the left half, 1 on the right half. This construction shows inductively that $\epsilon_1, \ldots, \epsilon_n$ are independent, coin-tossing random variables (Bernoulli with parameter 1/2), for each n. So the ϵ_n are independent coin-tosses.

Conversely, given ϵ_n independent coin tosses, form $X := \sum_1^\infty \epsilon_n/2^n$. Then $X_n := \sum_1^n \epsilon_k/2^k \to X$ a.s. The distribution function of X_n has jumps $1/2^n$ at $k/2^n$, $k = 0, 1, \ldots, 2^n - 1$. This 'saw-tooth jump function' converges to x on $[0,1]$, the distribution function of $U[0,1]$. So $X \sim U[0,1]$. So:

If $X = \sum_1^\infty \epsilon_n/2^n$, $X \sim U[0,1]$ iff ϵ_n are independent coin tosses.

So the Lebesgue probability space models *both* length on the unit interval *and* infinitely many independent coin tosses. Incidentally, this shows that the hard Measure Theory content of construction of Lebesgue measure (Carathéodory's Extension Theorem, which we have quoted) is the same as that of the construction of the infinite product space for repeated coin tossing (which we have sketched above, and referred forward to the Daniell-Kolmogorov theorem – which we shall also quote).

The mathematics above yields infinite replication of coin-tosses ϵ_n from the uniform distribution $U[0,1]$. Take the ϵ_n, and rearrange them into a two-suffix array ϵ_{jk} (as with Cantor's proof of 1873 that the rationals are countable). The ϵ_{jk} are all independent, so the $X_j := \sum \epsilon_{jk}/2^k$ are independent, and $U[0,1]$ by above. So from *one* $U(0,1)$, we get in this way *infinitely many copies*.

If F is a distribution function (right-continuous; increasing from 0 at $-\infty$ to 1 at ∞), define its (left-continuous) inverse function by

$$F^{-1}(t) := \inf\{F(x) \geq t\} \qquad (0 < t < 1).$$

Then if $U \sim U[0,1]$, $X := F^{-1}(U) \sim F$. For, $\{X \leq x\} = \{F^{-1}(U) \leq x\} = \{U \leq F(x)\}$, which has probability $F(x)$ as U is uniform. By this means (called the *probability integral transformation*) we can pass from generating copies from the uniform distribution (say by Monte Carlo simulation) to generating copies from the distribution F. Since by above we can use *one* uniform to generate a sequence of independent copies of uniforms, we may then generate a sequence of independent copies drawn from F. In particular, from *one* uniform we can generate an *infinite sequence* of copies of *standard normals*. We shall see in Ch. II that from this we can generate *Brownian motion*, the prototypical stochastic process. So the Lebesgue probability space, from which we can draw a uniform, is all we need – e.g. to generate Brownian motion. So everything rests on Lebesgue measure (as it should!)

11. The Strong Law of Large Numbers

Chebyshev's inequality.

The next result is due to P. L. CHEBYSHEV (1821-1984) in 1867. Kolmogorov's inequality below (A. N. KOLMOGOROV (1903-1987) in 1925) reduces to it when $n = 1$.

Theorem (Chebychev's inequality). If X has mean μ and variance σ^2, and $\epsilon > 0$,

$$P(|X - \mu| \geq \epsilon) \leq \sigma^2/\epsilon^2.$$

Proof.

$$\sigma^2 = \int_\Omega |X-\mu|^2 dP \geq \int_{|X-\mu|\geq\epsilon} |X-\mu|^2 dP \geq \epsilon^2 P(|X-\mu| \geq e). \qquad //$$

Theorem (Kolmogorov's Inequality). If $(X_n)_1^\infty$ are independent random variables with mean 0 and finite variance, then for all $\epsilon > 0$

$$P(\max_{r=1}^n |X_1 + \ldots + X_r| > \epsilon) \leq \frac{1}{\epsilon^2} \sum_{r=1}^n var(X_r).$$

Proof. Write

$$A := \{\max_{r=1}^n |X_1+\ldots+X_r| > \epsilon\}, \quad A_r := \{|X_1+\ldots+X_s| \leq \epsilon \ (s<r),\ |X_1+\ldots+X_r| > \epsilon\}.$$

Then $A_1, \ldots, A_n$ are disjoint with union A. Write $I_r := I_{A_r}$; by definition of A_r and independence, $I_r, X_{r+1}, \ldots, X_n$ are independent. We have

$$\begin{aligned} \sum_1^n var\ X_r &= var(X_1 + \ldots + X_n) = E[(X_1 + \ldots + X_n)^2] \\ &\geq E[I_A(X_1 + \ldots + X_n)^2] = \sum_1^n E[I_{A_r}(X_1 + \ldots + X_n)^2] \end{aligned}$$

$$= \sum_1^n E[I_r(X_1+\ldots+X_r)^2 + I_r(X_{r+1}+\ldots+X_n)^2 + 2I_r(X_1+\ldots+X_r)(X_{r+1}+\ldots+X_n)]$$

$$= \sum_1^n E[I_r(X_1+\ldots+X_r)^2] + E[I_r]E[(X_{r+1}+\ldots+X_n)^2] + 2E[I_r(X_1+\ldots+X_r)]E[(X_{r+1}+\ldots+X_n)],$$

by independence. The second term is ≥ 0; the third term is 0 as each $E[X_k] = 0$. So

$$\sum_1^n var\ X_r \geq \sum_1^n E[I_r(X_1 + \ldots + X_r)^2].$$

But on A_r, $(X_1 + \ldots + X_r)^2 \geq \epsilon^2$, so this is $\geq \epsilon^2 \sum_1^r P(A_r) = \epsilon^2 P(A)$, giving Kolmogorov's inequality. //

Corollary (Kolmogorov). If $\sum var\ X_n < \infty$, then $\sum(X_n - E[X_n])$ converges a.s.

Proof. By Kolmogorov's inequality, for all $\epsilon > 0$, $m, p = 1, 2, \ldots$,

$$P(\max_{r=1,\ldots,p}|(X_{m+1}-E[X_{m+1}])+\ldots+(X_{m+r}-E[X_{m+r}])| > \epsilon) \leq \frac{1}{\epsilon^2}\sum_{r=m+1}^{m+p} var\ X_r.$$

Let $p \to \infty$: by continuity (equivalently, σ-additivity) of the probability measure P,

$$P(\sup_{r\geq 1}|\sum_{k=m+1}^{m+r}(X_k - E[X_k])| > \epsilon) \leq \frac{1}{\epsilon^2}\sum_{r=m+1}^{\infty} var\ X_r < \infty.$$

Let $m \to \infty$: the right is the tail of a convergent series, so

$$P(\lim_{m\to\infty}\sup_{r\geq 1}|\sum_{k=m+1}^{m+r}(X_k - E[X_k])| > \epsilon) = 0.$$

This says that $\sum(X_k - E[X_k])$ is a.s. convergent. //

Lemma (Kronecker). If $\sum x_n$ converges to s (finite), and $b_n \uparrow \infty$, then

$$\frac{1}{b_n}\sum_1^n x_k b_k \to 0 \qquad (n \to \infty).$$

Proof. Write $b_0 := 0$, $a_k := b_k - b_{k+1}$, $s_n := \sum_1^n x_k$. By Abel's Lemma (= partial summation),

$$\frac{1}{b_n}\sum_1^n x_k b_k = \frac{1}{b_n}\sum b_k(s_k - s_{k-1}) = s_n - \frac{1}{b_n}\sum_1^n a_k s_{k-1} = (s_n - s) - \frac{1}{b_n}\sum_1^n a_k(s_{k-1} - s).$$

As $b_n \uparrow$, $a_n \geq 0$; for all $\epsilon > 0$, there exists N such that $|s_n - s| < \epsilon$ for $n \geq N+1$. Then for $n \geq N$,

$$|\frac{1}{b_n}\sum_1^n x_k b_k| \leq \epsilon + |\frac{1}{b_n}\sum_1^N b_k(s_k - s_{k-1})| + \epsilon(b_n - b_N)/b_n \leq 2\epsilon + |\frac{1}{b_n}\sum_1^N b_k(s_k - s_{k-1})|.$$

Let $n \to \infty$: limsup of LHS $\leq 2\epsilon$, for all $\epsilon > 0$, so is 0, so LHS $\to 0$. //

Theorem (Kolmogorov). If X_n are independent and $\sum var(X_n)/b_n^2 < \infty$ with $b_n \uparrow \infty$, then for $S_n := \sum_1^n X_k$,

$$(S_n - E[S_n])/b_n \to 0 \quad (n \to \infty) \quad a.s.$$

Proof. By the Corollary to Kolmogorov's Inequality, $\sum(X_n - E[X_n])/b_n$ converges a.s.. Then use Kronecker's Lemma with $x_n = (X_n - E[X_n])/b_n$. //

Lemma. If X has mean μ and distribution function F,

$$\sum_1^\infty P(|X| \geq n) \leq E|X| \leq 1 + \sum_1^\infty P(|X| \geq n).$$

Proof. For $i \geq 0$, let $A_i := \{i \leq |X| < i+1\}$. Then

$$\sum iP(A_i) \leq E|X| = \int |X| dP = \sum_i \int_{A_i} dP < \sum (i+1) P(A_i) = 1 + \sum_i iP(A_i).$$

But

$$\sum_i iP(A_i) = \sum_i \sum_{j=1}^i 1P(A_i) = \sum_j \sum_{i \geq j} P(A_i) = \sum_j P(X \geq j). \quad //$$

Theorem (Strong Law of Large Numbers, Kolmogorov, 1933). For X_n iid, $(X_1 + \ldots + X_n)/n$ converges to a constant μ a.s. as $n \to \infty$ iff $E|X| < \infty$, and then $\mu = EX$.

Proof. If $E|X| < \infty$, write μ for EX. Truncate $|X_n|$ at n to obtain Y_n:

$$Y_n := X_n \quad (|X_n| < n), \quad 0 \quad (|X_n| \geq n).$$

By the Lemma,

$$\sum P(X_n \neq Y_n) = \sum P(|X_n| \geq n) = \sum P(|X_1| \geq n) \leq E|X_1| < \infty.$$

So by the first Borel-Cantelli lemma, a.s. only finitely many of the events $X_n \neq Y_n$) occur. So

$$\frac{1}{n} \sum_1^n (X_k - Y_k) \to 0 \qquad a.s.,$$

so it suffices to show that, writing $S_n := \sum_1^n Y_k$,

$$S_n/n = \frac{1}{n} \sum_1^n Y_k \to \mu \qquad a.s. \tag{$*$}$$

Now

$$\sum var(Y_n)/n^2 \leq \sum E[Y_n^2]/n^2 = \sum \frac{1}{n^2} \int_{|x| \leq n} dF(x),$$

where F is the common distribution function of the X_n. The RHS is

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{j=1}^{n} \int_{j-1<|x|\leq j} x^2 dF(x) = \sum_{j=1}^{\infty} \int_{j-1<|x|\leq j} x^2 dF(x) \sum_{n=j}^{\infty} 1/n^2,$$

interchanging the order of summation. But in the integral $x^2 \leq j|x|$, and(arguing as in the proof of the Integral Test for convergent series)

$$\sum_{i=j}^{\infty} 1/i^2 - 1/j^2 \leq \int_j^{\infty} dx/x^2 = 1/j \leq \sum_{i=j}^{\infty} 1/i^2 : \quad \sum_{i=j}^{\infty} 1/i^2 \leq 1/j + 1/j^2 \leq 2/j.$$

So

$$\sum var(Y_n)/n^2 \leq \sum E[Y_n^2]/n^2 \leq 2 \sum_{j=1}^{\infty} \int_{j-1<|x|\leq j} |x| dF(x) = 2E|X| < \infty.$$

Now $(*)$ follows by Kolmogorov's theorem above.

Conversely, if $\Sigma_1^n X_k/n \to \mu$ a.s., then also $\Sigma_1^{n-1} X_k/n = [(n-1)/n].\Sigma_1^{n-1} X_k/(n-1) \to \mu$ also. Subtracting, $X_n/n \to 0$ a.s. Since the events $(|X_n| \geq n)$ are independent, the second Borel-Cantelli lemma gives

$$\sum P(|X_1| \geq n) = \sum P(|X_n| \geq n) < \infty.$$

This gives $E|X| < \infty$ by the Lemma. The conclusion of the first part now applies, and this completes the proof. //

Note. 1. Kolmogorov's SLLN of 1933 completes the story begun with Bernoulli's theorem in 1713. It gives precise form to the intuitive idea of the 'Law of Averages' – e.g., thinking about a probability as a long-run frequency. What this essentially says is that (thinking of a random variable as its mean plus a random error) independent errors tend to cancel. Any form of the LLN is really a result about *cancellation.*

2. Independence is not needed here. Strongly dependent errors need not cancel, but weakly dependent errors do (weak dependence can be made precise in many ways!). *Pairwise independence* suffices (N. Etemadi, 1981; cf. [GS], 7.5). For details, see e.g. the Stochastic Processes link on my Imperial College homepage, Lecture 14.