Probability: Limit Theorems I

Charles Newman,
Transcribed by Ian Tobasco

Abstract. This is part one of a two semester course on measure theoretic probability. The course was offered in Fall 2011 at the Courant Institute for Mathematical Sciences, a division of New York University. The text is *Probability Theory* by S. R. S. Varadhan. Homework will be assigned on a regular basis.

Contents

CHAPTER 1

Introduction to Probability Theory

1. Probability Spaces

Lecture 1, 9/6/11

DEFINITION 1.1. A *measurable space* is a pair $(\Omega, \mathcal{B})$ where Ω is a set and $\mathcal{B}$ is a σ*-field* of subsets of Ω, i.e. $\mathcal{B}$ contains $\emptyset$ and is closed under complementation and countable unions.

It follows immediately from the definition that every σ-field also contains Ω and is closed under countable intersections.

PROPOSITION 1.2. [1] *Given any collection $\mathcal{F}$ of subsets of Ω, there exists a unique σ-field $\mathcal{B}$ such that*

- *$\mathcal{B} \supset \mathcal{F}$ and*
- *for any σ-field $\mathcal{G}$ with $\mathcal{G} \supset \mathcal{F}$, it follows that $\mathcal{G} \supset \mathcal{B}$.*

The σ-field $\mathcal{B}$ given by the proposition above is called the σ-field *generated* by $\mathcal{F}$ and is often denoted $\sigma(\mathcal{F})$. It is the smallest σ-field containing $\mathcal{F}$. In general it is difficult to say exactly whether a set belongs to a given σ-field. But the proposition above gives a convenient way to construct a σ-field from the interesting sets (whatever those may be).

EXAMPLES 1.3. Here are some first examples of measure spaces.

(1) A pair of dice:

$$\Omega_1 = \{(\omega_1, \omega_2) : \omega_i \in \{1, 2, \dots, 6\} \text{ for } i = 1, 2\},$$
$$\mathcal{B}_1 = \mathcal{P}(\Omega_1).$$

Or if dice are indistinguishable, replace $\mathcal{B}_1$ with

$$\mathcal{B}_1' = \{A : (\omega_1, \omega_2) \in A \implies (\omega_2, \omega_1) \in A\}.$$

(2) Number of cars passing an intersection (in a one-minute period):

$$\Omega_2 = \{0, 1, 2, \dots\},$$
$$\mathcal{B}_2 = \mathcal{P}(\Omega_2).$$

(3) Arbitrarily long (or infinite) sequences of tosses of a coin:

$$\Omega_3 = \{(\omega_1, \omega_2, \dots) : \omega_i \in \{0, 1\}\}.$$

We want to understand events of the form

$$\{\omega = (\omega_1, \omega_2, \dots) : \omega_1 = \varepsilon_1, \dots, \omega_k = \varepsilon_k,\ \varepsilon_1, \dots, \varepsilon_k \in \{0, 1\}\}.$$

So let $\mathcal{F}$ denote the collection of such events, and take $\mathcal{B}_3 = \sigma(\mathcal{F})$.

[1] This is exercise 1.7 in the text.

(4) Point processes in space (e.g. one could study the distribution of galaxies in the universe):

$$\Omega_4 = \left\{\omega : \omega \subset \mathbb{R}^3 \text{ s.t. } \forall \text{ bounded } \Lambda \subset \mathbb{R}^3,\ \omega \cap \Lambda \text{ is finite}\right\}.$$

Let Λ be an open, bounded subset of $\mathbb{R}^3$ and let

$$A = \{\omega : |\omega \cap \Lambda| = n\}$$

for fixed $n \in \{0, 1, 2, \dots\}$. Let $\mathcal{F}$ be the collection of all such A_n for all possible Λ, and take $\mathcal{B}_4 = \sigma(\mathcal{F})$.

DEFINITION 1.4. A *probability measure* P on $(\Omega, \mathcal{B})$ is a function from $\mathcal{B}$ to $[0, 1]$ such that

- $P(\Omega) = 1$
- P is *countably additive*, i.e. if $A_1, A_2, \dots$ are disjoint members of $\mathcal{B}$ then

$$P\left(\cup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

Note that countable additivity is equivalent to the following continuity property:[2] If $B_1 \supset B_2 \supset \cdots$ and $B_n \in \mathcal{B}$ for all n then

$$P\left(\cap_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} P(B_n).$$

Taking complements, we get another equivalent continuity property: If $C_1 \subset C_2 \subset \cdots$ and $C_n \in \mathcal{B}$ for all n then

$$P\left(\cup_{n=1}^{\infty} C_n\right) = \lim_{n\to\infty} P(C_n).$$

DEFINITION 1.5. A *probability space* is a triple $(\Omega, \mathcal{B}, P)$ as above.

If Ω is countable and $\mathcal{B} = \mathcal{P}(\Omega)$ then a probability measure is uniquely determined by the number $p(\omega) = P(\{\omega\})$ for each $\omega \in \Omega$, provided $p(\omega) \geq 0$ for each ω and $\sum_{\omega\in\Omega} p(\omega) = 1$. In this case, we have

$$P(A) = \sum_{\omega\in A} p(\omega)$$

for any $A \in \mathcal{B}$.

EXAMPLE 1.6. Recall the previous examples. For Ω_1, we could take $p(\omega) = 1/36$ for each ω, supposing we have fair, independent dice. For Ω_2, we could take $p(\omega) = e^{-\lambda}\frac{\lambda^\omega}{\omega!}$, the Poisson distribution with mean $\lambda > 0$. But (3) and (4) are not as straightforward.

What if $\Omega = \sigma(\mathcal{F})$ is uncountable? We may know what P should be on the smaller collection $\mathcal{F}$, e.g.

$$P(\{\omega : \omega_1 = \varepsilon_1, \dots, \omega_k = \varepsilon_k\}) = \left(\frac{1}{2}\right)^k$$

in the third example above. This gives rise to the following questions in the general context: Given Ω and $\mathcal{B} = \sigma(\mathcal{F})$, if P is first defined on $\mathcal{F}$, then

(1) Can the domain of P be extended to $\mathcal{B}$ such that P is a probability measure? (In particular, how can we guarantee countable additivity?)

[2]Proof of which is problem 1 from problem set 1 (also exercise 1.2 in the text).

(2) Is the extension to $\mathcal{B}$ unique?

Basic answers to these questions are in the following definitions and theorems.

DEFINITION 1.7. A *field* $\mathcal{F}$ existing of subsets of Ω is a collection closed under complementation and finite unions. A finitely addivitive function P from $\mathcal{F}$ to $[0,1]$ with $P(\Omega) = 1$ is called *countably additive on* $\mathcal{F}$ if for any decreasing sequence $A_1 \supset A_2 \supset \cdots$ with $A_j \in \mathcal{F}$ for all j and with $\cap_{n=1}^{\infty} A_n = \emptyset$, we have

$$\lim_{n\to\infty} P(A_n) = 0.$$

THEOREM 1.8 (Caratheodory Extension Theorem). *If $\mathcal{F}$ is a field and P is countably additive on $\mathcal{F}$ as above, then there exists a unique probability measure P' on $\mathcal{B} = \sigma(\mathcal{F})$ such that $P'|_{\mathcal{F}} = P$.*

See page four of the text for a discussion of the proof. To use the theorem, we need a field. Consider the next example.

EXAMPLE 1.9. The field of subsets of $\mathbb{R}$. Let $\mathcal{I}$ be the collection of intervals of the form $(a,b]$ with $a < b$ or $(-\infty, b]$ or (a,∞). Then the finite disjoint unions of intervals from $\mathcal{I}$ form a field.

Lecture 2, 9/13/11

EXERCISE 1.10. Check that the collection of finite disjoint unions of intervals in the example above indeed form a field.

DEFINITION 1.11. The σ-field generated by the field in the previous example is called the *Borel σ-field*, $\mathcal{B}$. Any element of $\mathcal{B}$ is called a *Borel subset* of $\mathbb{R}$.

To define a probability measure on $(\mathbb{R}, \mathcal{B})$, we'll define it first on intervals $(a,b]$, and then extend uniquely via the Caratheodory extension theorem.

DEFINITION 1.12. The *Lebesgue measure* is the unique measure λ such that $\lambda((a,b]) = b - a$.

REMARK. But λ is not a probability measure, as $\lambda(\mathbb{R}) = \infty$. We'll see next how to construct probability measures via the extension theorem.

Suppose $F(x)$ is a monotone increasing and right-continuous function on $\mathbb{R}$. Define the set function

$$\lambda_F((a,b]) = F(b) - F(a).$$

Note if $F(x) = x$, then $\lambda_F((a,b]) = \lambda((a,b])$ and so λ_F extends to the Lebesgue measure. Now suppose that $F(-\infty) = 0$ and $F(+\infty) = 1$. Then λ_F is a probability measure, for

$$\lambda_F(\mathbb{R}) = \lim_{\substack{b\to\infty \\ a\to-\infty}} F(b) - F(a) = 1 - 0 = 1$$

by our assumption.

EXAMPLE 1.13. Let $F : \mathbb{R} \to \mathbb{R}$ be given by

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}.$$

Note that

$$\lambda_F((-\infty, -\epsilon]) = F(-\epsilon) - F(-\infty) = 0 - 0 = 0,$$
$$\lambda_F((-\infty, 0]) = F(0) - F(-\infty) = 1,$$

and

$$\lambda_F((-\infty,\infty)) = 1.$$

λ_F generates the Dirac δ-measure (or point mass) which satisfies

$$\delta(U) = \begin{cases} 1 & 0 \in U \\ 0 & 0 \notin U \end{cases}.$$

In a similar way we can generate the point mass at x,

$$\delta_x(U) = \begin{cases} 1 & x \in U \\ 0 & x \notin U \end{cases}.$$

CLAIM. λ_F is countably additive on $\mathcal{F}$ iff F is right-continuous.

DEFINITION 1.14. A *distribution function* $F(x)$ is a right-continuous, non-decreasing function on $\mathbb{R}$ which satisfies $F(-\infty) = 0$ and $F(+\infty) = 1$.

PROPOSITION 1.15. *If P is a probability measure on $(\mathbb{R}, \mathcal{B})$, then $F(x) = P((-\infty, x])$ is a distribution function.*

THEOREM 1.16. *Conversely, for every distribution function F, there exists a unique probability measure on $(\mathbb{R}, \mathcal{B})$ with $P((-\infty, x]) = F(x)$.*

Now we'll see an important class of probability measures. Let $x_1, x_2, \dots$ be a sequence of distinct real numbers, and let $p_1, p_2, \dots \geq 0$ satisfy $\sum p_n = 1$. Define the measure

$$P(A) = \sum_{n: x_n \in A} p_n = \sum_{n=1}^{\infty} p_n \cdot 1_A(x_n).$$

These are called *discrete* probability measures on $\mathbb{R}$. What is the distribution function for P? We have

$$F(x) = P((-\infty, x]) = \sum_{n: x_n \leq x} p_n.$$

Here is an example of such a probaiblity measure.

EXAMPLE 1.17. The Poisson probability measure on $\mathbb{R}$ with rate $\lambda > 0$. Set $x_1 = 0, x_2 = 1, \dots, x_n = n-1, \dots$ and set $p_n = \frac{\lambda^{n-1}}{(n-1)!} e^{-\lambda}$. The discrete probability measure generated by these x_n, p_n is the *Poisson measure*.

EXAMPLE 1.18. Let $x_1, x_2, \dots$ be a denumeration of $\mathbb{Q}$. Let $p_n = 2^{-n}$. Let

$$P = \sum p_n \delta_{x_n} = \sum \frac{1}{2^n} \cdot \delta_{x_n}.$$

This is an important measure and has some strange properties. For example:

PROPOSITION 1.19. *The distribution function*

$$F(x) = \sum_{n: x_n \leq x} \frac{1}{2^n}$$

is strictly increasing.

PROOF. Fix x and let $\epsilon > 0$. The interval $(x, x+\epsilon]$ contains rational numbers, and hence has positive probability. Thus

$$F(x+\epsilon) - F(x) = P((x, x+\epsilon]) > 0.$$

□

This measure is a great source of counterexamples, and is worth remembering.

EXAMPLE 1.20. Suppose f is a non-negative, (Lebesgue)-integrable function with $\int_{-\infty}^{\infty} f(y)\, dy = 1$. (Such an f is known as a *density function.*) Define the distribution function

$$F(x) = \int_{-\infty}^{x} f(y)\, dy$$

and let P satisfy $P((-\infty, x]) = F(x)$. (Check that F is indeed a distribution function.) By the fundamental theorem of calculus,

$$\frac{d}{dx} F(x) = f(x)$$

almost everywhere. We'll see why later.

2. Measure Theoretic Integration

First we must identify which functions we expect to integrate.

DEFINITION 2.1. A (real-valued) *measurable function* on a measure space (Ω, Σ) is a map $f : \Omega \to \mathbb{R}$ such that for all Borel $B \subset \mathbb{R}$, $f^{-1}(B) \in \Omega$.

Generically, a measurable function has the property that the preimages of measurable sets are measurable.

DEFINITION 2.2. Let (Ω, Σ, P) be a probability space. A *random variable* (rv) on (Ω, Σ, P) is a measurable function.

So what is the difference here? In measure theory we care about the space, $(\mathbb{R}, \mathcal{B})$, and objects like measurable functions help you understand the space. In probability, we care about random variables (observables), and we choose the space Ω to help us understand the random variables. We usually supress the argument when writing a random variable, so we would write X to mean the random variable (measurable function) $X(a)$.

DEFINITION 2.3. A map $f : \Omega_1 \to \Omega_2$ between measure spaces is *measurable* if for all $B \in \Sigma_2$, $f^{-1}(B) \in \Sigma_1$. Such an f is called an Ω_2*-valued random variable.*

REMARK. In this course, the functons we'll discuss will be measurable. So it's nice to know that certain classes of functions (e.g. continuous functions) are measurable and that the usual operations (e.g. $+, \times, (\cdot)^{-1}, \circ, \int(\cdot), \frac{d}{dx}(\cdot), \lim(\cdot)$) preserve the property of measurability. Proof of these facts can be found in any standard text on measure theory.

EXAMPLE 2.4. Let X_1, X_2, X_3, X_4 be real-valued random variables (on some measure space (Ω, Σ, P)). Define the matrix-valued random variable

$$Y = \begin{pmatrix} X_1 & X_2 \\ X_3 & X_4 \end{pmatrix}.$$

Let $\Omega_{2\times 2}$ be the space of 2×2 matrices ($\Omega_{2\times 2} \cong \mathbb{R}^4$), then $Y : \Omega \to \Omega_{2\times 2}$ is measurable.

Here are the steps to building a theory of integration:

(1) Define the integral

$$\int_{\Omega} f(\omega)\, dP(\omega) = \int f\, dP$$

for *simple functions* f, i.e. finite linear combinations of indicator functions of measurable sets.
(2) Extend the definition to bounded functions.
(3) Extend to non-negative functions.
(4) Extend to integrable functions (suitably defined).

Here is step four. If f is measurable, define the functions

$$f_+(\omega) = f(\omega) \cdot 1_{[0,\infty)}(f(\omega)) = \max(f, 0)$$
$$f_-(\omega) = -f(\omega) \cdot 1_{(-\infty,0]}(f(\omega)) = -\min(f, 0).$$

Then f_+, f_- are measurable, and by the construction $f = f_+ - f_-$ and $|f| = f_+ + f_-$. We way that f is *integrable* if $\int_\Omega |f|\, dP < \infty$, and in that case we define

$$\int f\, dP = \int f_+\, dP - \int f_-\, dP.$$

EXAMPLE 2.5. Consider $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \sin x$. One might be tempted to say $\int f = 0$, but in fact f is not integrable under our definition. So $\int f$ is not well defined.

DEFINITION 2.6. If $X : \Omega \to \mathbb{R}$ is a real-valued random variable which is integrable, then we define its *expectation* to be the quantity

$$E[X] = \int_{\Omega} X(\omega)\, dP(\omega).$$

Here are some useful theorems of integration:

- Monotone Convergence Theorem: If $X_n \uparrow$, then $\lim_{n\to\infty} E[X_n] = E[\lim_{n\to\infty} X_n]$.
- Dominated Convergence Theorem: If $|X_n| \leq Y$ for integrable Y, then $\lim_{n\to\infty} E[X_n] = E[\lim_{n\to\infty} X_n]$.
- Bounded Convergence Theorem: If $|X_n| \leq M < \infty$, then $\lim_{n\to\infty} E[X_n] = E[\lim_{n\to\infty} X_n]$.
- Jensen's Inequality: If ϕ is convex, then $\phi(E[X_n]) \leq E[\phi(X_n)]$.

Lecture 3, 9/20/11

These theorems hold under pointwise convergence of $X_n \to X$, but they also hold under a weaker kind of convergence, in which the set where $X_n \not\to X$ has probability zero.

DEFINITION 2.7. If X_n is a sequence of (real-valued) random variables on $(\Omega, \mathcal{B}, P)$, then we say $X_n \to X$ *almost surely (a.s.)* if $P\{\omega : X_n(\omega) \to X(\omega)\} = 1$.

Here is a another important kind of convergence.

DEFINITION 2.8. If $P\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\} \to 0$ for all $\epsilon > 0$, then we say that $X_n \to X$ *in probability.*

NOTE. If we were donig analysis, we would say "almost everywhere" instead of "almost surely" and "in measure" instead of "in probability."

EXAMPLE 2.9. On $\{[0,1], \text{Borel sets}, \text{Lebesgue measure}\}$, consider the sequence of random variables

$$\begin{aligned} X_1 &= 1_{[0,1/2]} \\ X_2 &= 1_{[1/2,1]} \\ X_3 &= 1_{[0,1/4]} \\ X_4 &= 1_{[1/4,1/2]} \\ &\vdots \end{aligned}$$

and so on. The X_n do not converge a.e. to anything (consider any non-dyadic number in $[0,1]$). But they do converge in probability to $X \equiv 0$, for

$$P\{|X_n - X| > \epsilon\} \leq \frac{1}{2^{k(n)}}$$

with $k(n) \to \infty$ as $n \to \infty$.

3. Product Spaces and Product Measures

This is important for studying independence in the abstract setting. Let $(\Omega_1, \mathcal{B}_1, P_1)$, $(\Omega_2, \mathcal{B}_2, P_2)$ be probability spaces, and recall their *Cartesian product* $\Omega = \Omega_1 \times \Omega_2$ is the set of (ω_1, ω_2) with $\omega_i \in \Omega_i$.

There is a natural way to construct a σ-field $\mathcal{B}$ and probability measure P on $\Omega_1 \times \Omega_2$, which is analogous to going from one-dimension Lebesgue measure (length) on $\mathbb{R}^1$ to two-dimensional Lebesgue measure (area) on $\mathbb{R}^2$. The *product σ-field*, $\mathcal{B} = \mathcal{B}_1 \times \mathcal{B}_2$, is the σ-field generated by $\{A_1 \times A_2 : A_1 \in \mathcal{B}_1, A_2 \in \mathcal{B}_2\}$ (the measurable rectangles). Now we look to define a measure on $\mathcal{B}$. Let $\mathcal{F}$ be the field (not σ-field) of finite disjoint unions of the measurable rectangles. Then define P on $\mathcal{F}$ by first setting

$$P(A_1 \times A_2) = P_1(A_1) \cdot P_2(A_2)$$

for one measurable rectangle $A_1 \times A_2$ and then extending via finite additivity to all of $\mathcal{F}$.

LEMMA 3.1. *P defined as above is countably additive on $\mathcal{F}$.*

Thus P extends uniquely to a probability measure on $\mathcal{B}$, the product measure $P_1 \times P_2$. Furthermore, if $A \in \mathcal{B}$ (but not neccessarily in $\mathcal{F}$), then $P(A)$ can be evaluated by interated integration (in either order). This is corollary 1.11 in the text, and is a special case of (and leads to) the following theorem.

THEOREM 3.2 (Fubini). *If $f(\omega_1, \omega_2)$ is a (real-valued) integrable function on $(\Omega, \mathcal{B}, P) = (\Omega_1 \times \Omega_2, \mathcal{B}_1 \times \mathcal{B}_2, P_1 \times P_2)$, then the function $\omega_2 \mapsto f(\omega_1, \omega_2)$ is integrable for a.e. ω_1, the function $\omega_1 \mapsto \int_{\Omega_2} f(\omega_1, \omega_2)\, dP_2$ is integrable, and*

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2)\, dP = \int_{\Omega_1} \left[\int_{\Omega_2} f(\omega_1, \omega_2)\, dP_2 \right] dP_1.$$

Similarly,

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2)\, dP = \int_{\Omega_2} \left[\int_{\Omega_1} f(\omega_1, \omega_2)\, dP_1 \right] dP_2.$$

REMARK. The converse holds for non-negative functions, a result due to Tonelli.

4. Distributions and Expectations

DEFINITION 4.1. Suppose $(\Omega_1, \mathcal{B}_1, P)$ is a probability space and $(\Omega_2, \mathcal{B}_2)$ is a measure space. If $X : \Omega_1 \to \Omega_2$ is a Ω_2-valued random variable, then Q defined by

$$Q(A) = P\left(X^{-1}(A)\right)$$

for $A \in \mathcal{B}_2$ is a probability measure on $(\Omega_2, \mathcal{B}_2)$, called the *induced measure* or the *distribution* of X.

NOTE. In the analysis setting, one might see $Q = PX^{-1}$. In the probability setting, we can interpret $Q(A)$ as the probability that X takes values in A.

EXAMPLE 4.2. If X is real-valued ($(\Omega_2, \mathcal{B}_2) = (\mathbb{R}, \text{Borel sets})$), then Q is the probability measure on $\mathbb{R}$ whose distribution function is

$$F_Q(x) = Q((-\infty, x]) = P(X \leq x) = F_X(x).$$

F_X is the cumulative distribution function of X from classical probability. So the distribution of a random variable (in the new sense) is just a generalization of cumulative distribution to abstract probability space.

THEOREM 4.3 (Change of Variables). *If X is as in the above definition and h is a real-valued measureable function on $(\Omega_2, \mathcal{B}_2)$ then the function g given by $g(\omega_1) = h(X(\omega_1))$ is a real-valued random variable on $(\Omega_1, \mathcal{B}_1)$. g is integrable on $(\Omega_1, \mathcal{B}_1, P)$ iff h is integrable on $(\Omega_2, \mathcal{B}_2, Q)$ where Q is the distribution of X, and*

$$E[g] = E[h(X)] = \int_{\Omega_1} h(X(\omega_1))\, dP = \int_{\Omega_2} h(\omega_2)\, dQ.$$

EXAMPLE 4.4. If X is real-valued, then we have the usual formula

$$E[h(X)] = \int_{\Omega_1} h(X(\omega_1))\, dP = \int_{\mathbb{R}} h(\omega_2)\, dQ = \int_{-\infty}^{\infty} h(x)\, dF_X(x)$$

with the Lebesgue-Stieltjes integral on the right.

Here is the situation thus far. If X is a real-valued random variable on $(\Omega, \mathcal{B}, P)$, then it induces a measure $\alpha = PX^{-1}$ called the probability distribution of X. The distribution function of X is then $F(x) = \alpha((-\infty, x]) = P(X \leq x)$. And if X is integrable, its expectation (mean) is

$$E[X] = \int_{\Omega} X(\omega)\, dP = \int_{\mathbb{R}} x\, d\alpha = \int_{-\infty}^{\infty} x\, dF(x).$$

Finally, if g an α-integrable real-valued function on $\mathbb{R}$, then

$$E[g(X)] = \int_{\Omega} g(X(\omega))\, dP = \int_{\mathbb{R}} g(x)\, d\alpha = \int_{-\infty}^{\infty} g(x)\, dF(x).$$

DEFINITIONS 4.5. The *mth moment* of X for $m = 1, 2, 3, \ldots$ is

$$E[X^m] = \int_{\mathbb{R}} x^m\, d\alpha,$$

provided $E[|X|^m] = \int_{\mathbb{R}} |x|^m\, d\alpha < \infty$. The *variance* of X is

$$\text{Var}(X) = E\left[(X - E(X))^2\right] = E\left[X^2\right] - (E[X])^2$$

and the *standard deviation* of X is

$$\sigma(X) = \sqrt{\text{Var}(X)}.$$

Problem one on the second homework is to calculate the mean and standard deviation of a "shuffling random variable". Here is a hint.

PROPOSITION 4.6. *If $X_1, X_2, \ldots, X_n$ are integrable random variables then $\sum c_i X_i$ is integrable and $E\left[\sum c_i X_i\right] = \sum c_i E\left[X_i\right]$.*

CHAPTER 2

Weak Convergence

1. Characteristic Functions

NOTE. In probability, characteristic functions are not indicator functions.

Recall the following elementary facts about complex numbers:

- $i^2 = -1$
- $\mathbb{C} = \{a + bi \ a, b \in \mathbb{R}\}$
- $e^{ir} = \cos r + i \sin r$ for $r \in \mathbb{R}$.

EXERCISE 1.1. Define $e^{ir} = \sum \frac{(ir)^n}{n!}$. Prove the third fact using $i^2 = -1$ and Taylor series for $\cos r$ and $\sin r$.

DEFINITION 1.2. A complex-valued function $f(\omega) = a(\omega) + ib(\omega)$ is *integrable* if a and b are integrable, and then

$$\int f(\omega)\ dP = \int a(\omega)\ dP + i \int b(\omega)\ dP.$$

Thus if α is a probability measure on $\mathbb{R}$, then for $t \in \mathbb{R}$ we have

$$\int_{\mathbb{R}} e^{itx}\, d\alpha = \int_{\mathbb{R}} \cos tx\, d\alpha + i \int_{\mathbb{R}} \sin tx\, d\alpha.$$

If α is the probability distribution of some real-valued random variable X, then consider the complex-valued function

$$\phi(t) = \int_{\mathbb{R}} e^{itx}\, d\alpha = \int_{-\infty}^{\infty} e^{itx}\, dF_X = E\left[e^{itx}\right].$$

DEFINITION 1.3. The function ϕ above is called the *characteristic function* of α (or of F_X or of X).

Characteristic functions are important to many parts of this course. For example, a standard proof of the central limit theorem is via characteristic functions.

EXAMPLES 1.4.

(1) If X has uniform distribution on $[C, D]$, then

$$\int_{\mathbb{R}} e^{itx}\, d\alpha = \int_C^D e^{itx} \frac{1}{D-C}\, dx = \frac{1}{D-C}\left(\frac{e^{itD} - e^{itC}}{it}\right).$$

Note that

$$\lim_{t \to 0} \frac{1}{D-C}\left(\frac{e^{itD} - e^{itC}}{it}\right) = 1$$

as we expect (α is a probability measure). A special case is $C = -1/2$, $D = +1/2$, for which

$$\int_{\mathbb{R}} e^{itx}\, d\alpha = \frac{\sin \frac{t}{2}}{\frac{t}{2}}.$$

(2) If X is Binomial(n,p), then

$$E\left[e^{itx}\right] = \sum_{k=0}^{n} e^{itk}\binom{n}{k}p^k\left(1-p\right)^{n-k} = \left(pe^{it}+\left(1-p\right)\right)^n.$$

(3) If X is Poisson(λ), then

$$E\left[e^{itx}\right] = \sum_{k=0}^{\infty} e^{itk}e^{-\lambda}\frac{\lambda^k}{k!} = e^{\lambda\left(e^{it}-1\right)}.$$

(4) If X is exponential with mean μ, then

$$E\left[e^{itx}\right] = \frac{1}{1-i\mu t}.$$

(5) If X is Normal$\left(\mu,\sigma^2\right)$, then

$$E\left[e^{itx}\right] = e^{i\mu t-\frac{1}{2}\sigma^2t^2}.$$

EXERCISE 1.5. Verify these calculations.

What is the relation between characteristic functions and moments? Formally, by interchanging derivatives and integrals, we get

$$\frac{d}{dt}\phi\left(t\right) = \frac{d}{dt}\int e^{itx}\,d\alpha = \int\left(ix\right)e^{itx}\,d\alpha$$

and by setting $t=0$

$$\phi'\left(0\right) = \int ix\,d\alpha = iE\left[X\right].$$

More generally

$$\left(\frac{d}{dt}\right)^m e^{itx}\Big|_{t=0} = i^mx^m$$

and so

$$\phi^{(m)}\left(0\right) = i^mE\left[X^m\right],$$

at least formally. When is this justified?

PROPOSITION 1.6. *If $E\left[|X|^m\right]<\infty$, then ϕ is m-times (continuously) differentiable and*

$$\phi^{(m)}\left(0\right) = i^mE\left[X^m\right].$$

Proof of this will be problem two on the second problem set. Also see exercises 2.2, 2.4 of the text.

What about a converse? If ϕ is m-times (continuously) differentiable, does this imply that $E\left[|X|^m\right]<\infty$ and hence

$$E\left[X^m\right] = \frac{1}{i^m}\phi^{(m)}\left(0\right)?$$

Lecture 4, 9/27/11

The answer is yes for m even, but no for m odd. See exercise 2.3 of the text for the details.

Here is a classic theorem, which tells us how to determine if a complex-valued function is the characteristic function of a random variable.

THEOREM 1.7 (Bochner). *A complex-valued function $\phi\left(t\right)$ is the characteristic function of some α (or F or X) iff ϕ satisfies the following three properties:*

(1) $\phi\left(0\right)=1$

(2) *ϕ is continuous (for all t)*
(3) *ϕ is positive (semi-)definite; i.e. for any $n = 1, 2, 3, \ldots$ and any real $t_1, \ldots, t_n$, the $n \times n$ matrix*

$$(\phi(t_i - t_j))_{i,j=1}^n$$

should be positive (semi-)definite (self-adjoint with all positive eigenvalues). Equivalently, given $n \in \mathbb{N}$ and $t_1, \ldots, t_n \in \mathbb{R}$, ϕ should satisfy

$$\sum_{i,j=1}^n c_i \overline{c_j} \phi(t_i - t_j) \geq 0$$

for all complex $c_1, \ldots, c_n$.

See the text for the proof; the only if part is easier. This next proposition is used to finish the proof of the central limit theorem. Observe first that a distribution function F is determined uniquely by its values at its continuity points (F is right-continuous and increasing).

PROPOSITION 1.8. *A distribution α (or distribution function F) is uniquely determined by its characteristic function: if a, b are continuity points of F then*

$$\alpha((a,b]) = F(b) - F(a) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \phi(t)\, dt.$$

See pp. 20-21 of the text for the proof.

REMARK. Even if a, b are not continuity points, the limit above equals

$$\alpha((a,b)) + \frac{1}{2}\alpha(\{\alpha\}) + \frac{1}{2}\alpha(\{b\}).$$

This is consistent with the proposition.

2. Weak Convergence

We are working towards the central limit theorem, which is about the convergence of distributions to the normal distribution. We need to reinterpret the word "convergence" to acheive the theorem.

THEOREM 2.1. *Let α_n, α be probability measures on $\mathbb{R}$ and let F_n, F be the corresponding ditribution functions. The following are equivalent:*

(1) *$\lim_{n\to\infty} F_n(x) = F(x)$ for every x that is a continuity point of F.*
(2) *$\alpha_n([a,b]) \to \alpha([a,b])$ as $n \to \infty$ for all $a, b \in \mathbb{R}$ which are not atoms of α (i.e. $\alpha(\{a\}), \alpha(\{b\}) \neq 0$).*
(3) *$\lim_{n\to\infty} \int_{\mathbb{R}} f(x)\, d\alpha_n(x) = \int_{\mathbb{R}} f(x)\, d\alpha(x)$ for every bounded, continuous (real/complex-valued) function f on $\mathbb{R}$.*

REMARK. If X_n has distribution α_n and X has distribution α, then (3) says that $E[f(X_n)] \to E[f(X)]$.

DEFINITION 2.2. If α_n, α (F_n, F) satisfy any of the conditions above, we say α_n *converges weakly* to α, and we write $\alpha_n \Rightarrow \alpha$ ($F_n \Rightarrow F$).

THEOREM 2.3. *$\alpha_n \Rightarrow \alpha$ iff the characteristic functions ϕ_n, ϕ satisfy $\lim_{n\to\infty} \phi_n(t) = \phi(t)$ for every real t.*

This last interpretation of weak convergence is used to prove the central limit theorem.

REMARKS.

(1) If X_n, X have distributions α_n, α with $\alpha_n \Rightarrow \alpha$, we say X_n *converges in distribution (in law)* to X. As noted above, this means $E[f(X_n)] \to E[f(X)]$ for any bounded, continuous f.
(2) (1) or (2) in theorem 2.1 can be strengthened to: $\alpha_n(A) \to \alpha(A)$ for every Borel set A so that $\alpha\left(\overline{A} \backslash A^0\right) = 0$. (This is theorem 2.6 in the text.) Such a set A is called a *continuity set* of α.

EXAMPLE 2.4. Here is a simple example. Take $\alpha_n = \delta_{1-1/n}$ and $\alpha = \delta_1$. Then $\alpha_n \Rightarrow \alpha$, but $\alpha_n(\{1\}) = 0$ for all n so that $\alpha_n(\{1\}) \not\to \alpha(\{1\})$. Indeed, $\{1\}$ is not a continuity set of α ($\alpha\left(\overline{A} \backslash A^0\right) \neq 0$).

Consider the proof of theorems 2.1 and 2.3. (1) $\Longrightarrow$ (3) is fairly straightforward (see pp. 25-26 of the text). (3) $\Longrightarrow \phi_n(t) \to \phi(t)$ is trivial (take $f(x) = e^{itx}$). The interesting proof is that $\phi_n(t) \to \phi(t) \Longrightarrow$ (1), which follows from the next theorem.

THEOREM 2.5. *Suppose ϕ_n, $n = 1, 2, \ldots$, are characteristic functions (of probability measures α_n) and there exists a function ϕ so that $\phi_n(t) \to \phi(t)$ for each t. Suppose also that ϕ is continuous at $t = 0$. Then ϕ is a characteristic function of some probability measure α, and $\alpha_n \Rightarrow \alpha$.*

PROOF SKETCH. Part 1. (Steps 1-3 on pp. 26-27.) Let F_n be any sequence of distribution functions, then there exists a subsequence n_k and a "sub-distribution function" G (G is non-decreasing, right-continuous, with values in $[0, 1]$; but $G(-\infty) > 0$ and/or $G(+\infty) < 1$ are allowed) such that $F_{n_k}(x) \to G(x)$ at every continuity point of G. Proof of this fact uses boundedness of $\{F_n(x)\}$ for each x, countability of the rationals, diagonalization, and so forth. In some books, this type of convergence is called *vague convergence*.

EXAMPLE 2.6. For $\alpha_n = \frac{2}{3}\delta_1 + \frac{1}{4}\delta_n + \frac{1}{12}\delta_{-n}$, $F_n(x) \to G(x)$ with

$$G(x) = \begin{cases} \frac{1}{12} & x < 1 \\ \frac{3}{4} & x \geq 1 \end{cases}.$$

Part 2. (Steps 4-5 on pp. 27-28.) The key argument is to show that $\phi(t)$ is continuous at $t = 0$ implies that $G(-\infty) = 0$ and $G(+\infty) = 1$. This follows from the

LEMMA 2.7. *If F is a distribution function with characteristic function ϕ, then for any $T > 0$ we have*

$$1 - \left(F\left(\frac{2}{T}\right) - F\left(-\frac{2}{T}\right)\right) \leq 2\left(1 - \frac{1}{2T}\int_{-T}^{T} \phi(t)\, dt\right).$$

This lemma is proved by elementary manipulations and Fubini's theorem (see p. 27 of the text). Applying the lemma to the F_{n_k} and taking $k \to \infty$ yields that for small T such that $\pm 2/T$ are continuity points of G,

$$1 - \left(G\left(\frac{2}{T}\right) - G\left(\frac{-2}{T}\right)\right) \leq 2\left(1 - \frac{1}{2T}\int_{-T}^{T} \phi(t)\, dt\right).$$

This holds even though we don't yet know G is a distribution function (so we don't know ϕ is a characteristic function). But we do know ϕ is continuous at $t = 0$ and $\phi(0) = 1$, so $G(-\infty) = 0$ and $G(+\infty) = 1$. So G is a distribution function.

Now we have a subsequence F_{n_k} which converges to a distribution function G at continuity points. By what we've already shown, $\phi_{n_k}(t) \to \int e^{itx}\, dG$ for every t. But since $\phi_{n_k}(t) \to \phi(t)$, we have $\phi(t) = \int e^{itx}\, dG$. Now for any subsequence n'_k, by the same argument there exists a sub-subsequence n'_{k_j} so that $F_{n'_{k_j}}$ converges to a distribution function G'. But again $\phi_{G'} = \phi_G$ so that $G' = G$ and hence $F_{n'_{k_j}} \to G$. This implies $F_n \to G$; now rename G as F to complete the proof.

□

The arguments used in this proof are interesting in their own right.

DEFINITION 2.8. A collection $\mathcal{A}$ of probability distributions on $\mathbb{R}$ is called *totally bounded* if there exists a probability distribution α so that every sequence α_n from $\mathcal{A}$ has a subsequence $\alpha_{n_k} \Rightarrow \alpha$.

THEOREM 2.9. *A family of probability distributions $\mathcal{A}$ is totally bounded iff either of the following holds:*

(1) The family is tight, *i.e.*

$$\lim_{l \to \infty} \sup_{\alpha \in \mathcal{A}} \alpha(\{x : |x| \geq l\}) = 0.$$

(2) Let ϕ_α be the characteristic function of α. Then

$$\lim_{h \to 0} \sup_{|t| \leq h} |1 - \phi_\alpha(t)| = 0.$$

That (1) and (2) are equivalent is the content of lemma 2.7. That (1) implies $\mathcal{A}$ is totally bounded is part 1 of the proof of theorem 2.5. That (2) implies $\mathcal{A}$ is totally bounded is part 2 of the proof of theorem 2.5. Lecture 5, 10/4/11

CHAPTER 3

Independent Random Variables

1. Independence and Convolution

Recall the notion of independent random variables from classical probability. We need a more general definition to tackle the classic theorems of probability.

DEFINITIONS 1.1. A finite family of (real-valued) random variables $X_1, X_2, \cdots, X_n$ is *independent (jointly independent)* if

$$P(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n)$$

for $B_1, B_2, \ldots$ any Borel subsets of $\mathbb{R}$. An infinite family of random variables $\{X_\alpha\}$ is *independent* if every finite subfamily is. Events $A_1, A_2, \ldots$ are said to be *independent* if their indicators are independent, or equivalently if

$$P(C_1 \cap C_2 \cap \cdots \cap C_n) = P(C_1) \cdots P(C_n)$$

for each n and for every choice of $C_j = A_j$ or $(A_j)^c$.

NOTE.

(1) For a pair of events A_1, A_2 it suffices to check

$$P(A_1 \cap A_2) = P(A_1) P(A_2).$$

This is what is usually presented in a first course on probability.

EXERCISE 1.2. Prove this.

(2) It is possible for a family of events (or random variables) to be pairwise independent but not (jointly) independent. See the next example.

EXAMPLE 1.3. Fair coin tossing. Set

$$X_i = \begin{cases} +1 & i\text{th toss is heads} \\ -1 & i\text{th toss is tails} \end{cases},$$

then X_1, X_2 and $Y = X_1 X_2$ are pairwise independent but not jointly independent. It might seem non-intuitive that these are pairwise independent. But think of the conditional distributions. This really shows independence is a statistical property (and not a functional property).

Here is a second way to understand independence. Recall that if $X_1, \ldots, X_n$ are real-valued random variables on some $(\Omega, \mathbb{F}, P)$ we can regard $X = (X_1, \ldots, X_n)$ as an $\mathbb{R}^n$-valued random variable whose distribution (the *joint distribution* of $X_1, \ldots, X_n$ from elementary probability) is the measure μ on $\mathbb{R}^n$ so that $\mu(B) = P(\{\omega : X(\omega) \in B\})$. If B is a rectangle $B_1 \times \cdots \times B_n$ and if $X_1, \ldots, X_n$ are independent, then $\mu(B_1 \times \cdots \times B_n) = \mu_1(B_1) \cdots \mu_n(B_n)$ where μ_i is the distribution (on $\mathbb{R}$) of X_i. In this case, we write $\mu = \mu_1 \times \cdots \times \mu_n$. The converse is also true.

LEMMA 1.4. *$X_1, \ldots, X_n$ are independent iff the distribution of $(X_1, \ldots, X_n)$ is the product measure of the distributions of the individual X_is.*

We're working our way towards applying the theory to characteristic functions. This will lead to the classic theorems. Here is a tool we'll use quite often.

PROPOSITION 1.5. *Suppose $f_1, \ldots, f_n$ are (bounded) measurable functions on $\mathbb{R}$.*[1] *If $X_1, \ldots, X_n$ are independent (real-valued) random variables, then*

$$E\left[f_1\left(X_1\right) \cdots f_n\left(X_n\right)\right] = \Pi_{j=1}^n E\left[f_j\left(X_j\right)\right].$$

PROOF. Apply the above lemma and Fubini's theorem. □

An important case of this is when $f_j\left(X_j\right) = e^{it_j X_j}$ with real numbers t_j and $j = 1, \ldots, n$. Then, the proposition implies

$$E\left[e^{i(t_1 X_1 + \cdots + t_n X_n)}\right] = \Pi_{j=1}^n E\left[e^{it_j X_j}\right].$$

The object on the left is the (multivariate) characteristic function $\phi\left(t\right) = \phi\left(\left(t_1, \ldots, t_n\right)\right)$ of $X = (X_1, \ldots, X_n)$. Of course this definition does not require independence. The converse is also true (if the equality holds for all choices of t_j, then $X_1, \ldots, X_n$ are independent).

In this course, we'll mainly consider $t = t_1 = t_2 = \cdots = t_n$. In this case, we see that if $X_1, \ldots, X_n$ are independent, then the characteristic function of $X = X_1 + \cdots + X_n$ is

$$\begin{aligned}\phi_{X_1 + \cdots + X_n}\left(t\right) &= E\left[e^{it(X_1 + \cdots + X_n)}\right] \\ &= \Pi_{j=1}^n E\left[e^{itX_j}\right] \\ &= \Pi_{j=1}^n \phi_{X_j}\left(t\right).\end{aligned}$$

This is an important relation which we'll use time and time again. But note that even if the equality above holds for all t it may not be true that $X_1, \ldots, X_n$ are independent. (??)

The discussion above shows the utility of the characteristic functions. But in practice, distributions are closer to what we're actually interested in. So we ask: Is there a representation for the distribution of sums of independent random variables? The answer is yes, but unfortunately it's not quite as nice. By Fubini's theorem, one has that for $Z = X + Y$ (X, Y independent) and A any Borel subset of $\mathbb{R}$,

$$\mu_Z\left(A\right) = \int_{\mathbb{R}} \mu_X\left(A - y\right) d\mu_Y\left(y\right) = \int_{\mathbb{R}} \mu_Y\left(A - x\right) d\mu_X\left(x\right).$$

This is called the *convolution* of μ_X and μ_Y and is denoted by $\mu_X \star \mu_Y$. If X, Y have probability densities f_X, f_Y, then Z also has a density f_Z with

$$f_Z\left(z\right) = \int_{-\infty}^{\infty} f_X\left(z - y\right) f_Y\left(y\right) dy = \int_{-\infty}^{\infty} f_Y\left(z - x\right) f_X\left(x\right) dx.$$

Here is a standard and useful fact for independent X, Y, which follow from the relation $E\left[XY\right] = E\left[X\right] E\left[Y\right]$: If $E\left[X^2\right], E\left[Y^2\right] < \infty$, then

$$\operatorname{Var}\left(X + Y\right) = \operatorname{Var}\left(X\right) + \operatorname{Var}\left(Y\right).$$

[1] The assumption that $f_1, \ldots, f_n$ are bounded can be weakened to $E\left[\left|f_j\left(X_j\right)\right|\right] < \infty$ for each j.

For $X_1, \dots, X_n$ independent,

$$\operatorname{Var}(X_1 + \cdots + X_n) = \operatorname{Var}(X_1) + \cdots + \operatorname{Var}(X_n).$$

EXERCISE 1.6. Derive these results.

Many of the standard limit theorems of probability theory (e.g. the Law of Large Numbers, the Central Limit Theorem) concern sums of independent random variables $X_1 + \cdots + X_n$ as $n \to \infty$. In order to make sense of these theorems, we need to a probability space on which infinitely many independent random variables $X_1, X_2, \dots$ are defined (e.g. with some common prescribed distribution). This comes from the classical sampling techniques from statistics.

DEFINITION 1.7. If $X_1, X_2, \dots$ are independent random variables which share a common distributino, then they are said to be *independent, identically distributed (i.i.d.)* random variables.

EXAMPLE 1.8. Coin tossing where

$$X_j = \begin{cases} +1 & j\text{th toss is heads} \\ -1 & j\text{th toss is tails} \end{cases}.$$

There is a general argument from measure theory that gives the existence of such spaces.

THEOREM 1.9. *Let $\mu_1, \mu_2, \dots$ be (Borel) probability measures on $\mathbb{R}$, and let $\Omega = \{a = (x_1, x_2, \dots) : x_j \in \mathbb{R}\}$ be the space of infinite sequences of real numbers (sometimes denoted $\mathbb{R}^\infty$). Let $\mathbb{F}$ be the field of "finite dimensional cylinder sets", i.e. sets of the form $\{\omega = (x_1, x_2, \dots) : (x_1, \dots, x_n) \in A\}$ for $n \in \mathbb{N}$ and A a Borel subset of $\mathbb{R}^n$. Let Σ be the σ-field generated by $\mathbb{F}$. Then there exists a unique probability measure $\mu = \mu_1 \times \mu_2 \times \cdots$ on (Ω, Σ) such that*

$$\mu(\{\omega = (x_1, x_2, \dots) : (x_1, \dots, x_n) \in A\}) = (\mu_1 \times \cdots \times \mu_n)(A)$$

for any $n \in \mathbb{N}$ and any Borel subset A of $\mathbb{R}$.

REMARK.

(1) There is no convergence issue here. We're working on a probability space.
(2) We could have defined μ via any choice of n coordinates from ω, instead of just the first n. The result would be the same.

2. Weak Law of Large Numbers

Now we can consider i.i.d. random variables $X_1, X_2, \dots$ and study the *sample mean* $\frac{X_1 + \cdots + X_n}{n}$ as $n \to \infty$. We'll begin with the weaker result, where we'll assuming not only that $E(|X_1|) < \infty$ (so that $E(X_1) = m$ is defined) but also that $E(X_1^2) < \infty$ (finite variance). Write $S_n = X_1 + \cdots + X_n$. We first study the weak law of large numbers, which claims that $\frac{S_n}{n} \to m$ in probability, i.e. that $P\left(\left|\frac{S_n}{n} - m\right| > \delta\right) \to 0$ as $n \to \infty$.

NOTE. Usually convergence in probability and in distribution are not the same. But since the convergence here is to a constant (and not an arbitrary random variable), the weak law of large numbers equivalently claims that the distributions of $\frac{S_n}{n}$ converge in distribution to δ_m as $n \to \infty$.

EXERCISE 2.1. Check that $\alpha_n \Rightarrow \delta_0$ iff $P(|X_n| > \delta) \to 0$ as $n \to \infty$.

Later, we'll see the strong law of large numbers, where the convergence is almost surely. This takes much more work!

Recall the following result from measure theory.

PROPOSITION 2.2 (Chebyshev's Inequality). *For any random variable X with $E\left[X^2\right] < \infty$,*

$$P\left(|X| > a\right) \leq \frac{1}{a^2} E\left[X^2\right]$$

for any $a > 0$.

PROOF. We have

$$\begin{aligned} P\left(|X| > a\right) &= E\left[1_{\{|X|>a\}}\right] \\ &\leq E\left[\frac{X^2}{a^2} 1_{\{|X|>a\}}\right] \\ &\leq \frac{1}{a^2} E\left[X^2\right] \end{aligned}$$

and so we're done. □

Now apply this to $X = \frac{S_n}{n} - m$ with $a = \delta$ to get

$$\begin{aligned} P\left(\left|\frac{S_n}{n} - m\right| > \delta\right) &\leq \frac{1}{\delta^2} E\left[\left(\left(\frac{S_n}{n}\right) - m\right)^2\right] \\ &= \frac{1}{\delta^2} \mathrm{Var}\left(\frac{S_n}{n}\right) \\ &= \frac{1}{\delta^2} \mathrm{Var}\left(\frac{X_1}{n} + \cdots + \frac{X_n}{n}\right) \\ &= \frac{1}{\delta^2} n \cdot \mathrm{Var}\left(\frac{X_1}{n}\right) \\ &= \frac{1}{\delta^2} \frac{1}{n} \cdot \mathrm{Var}\left(X_1\right) \\ &= \frac{1}{n\delta^2} \sigma^2 \\ &\to 0 \end{aligned}$$

as $n \to \infty$. Note the second line follows as

$$m = E\left[X_1\right] = E\left[\frac{X_1 + \cdots + X_n}{n}\right] = E\left[\frac{S_n}{n}\right].$$

So we have obtained a first weak law of large numbers. (See theorem 3.2 in the text.)

THEOREM 2.3. *If $X_1, X_2, \ldots$ are i.i.d. with $E\left[X_1\right], E\left[X_1^2\right] < \infty$ and $m = E\left[X_i\right]$, then $\frac{Sn}{n} \to m$ in probability as $n \to \infty$.*

We can do better.

THEOREM 2.4. *If $X_1, X_2, \ldots$ are i.i.d. with $E\left[|X_1|\right] < \infty$ and $m = E\left[X_i\right]$, then $\frac{Sn}{n} \to m$ in probability as $n \to \infty$.*

There are two proofs.

First proof. This proof is by a "truncation" argument. Given $C > 0$, define $X_j^C = X_j \cdot 1_{\{|X_j| \leq C\}}$ and $Y_j^C = X_j - X_j^C = X_j \cdot 1_{\{|X_j| > C\}}$. Then

$$\frac{S_n}{n} = \frac{1}{n}\sum_{j=1}^{n} X_j^C + \frac{1}{n}\sum_{j=1}^{n} Y_j^C = \xi_n^C + \eta_n^C,$$

and

$$\begin{aligned} m &= E\left[X_1\right] \\ &= E\left[\frac{S_n}{n}\right] \\ &= E\left[X_1^C\right] + E\left[Y_1^C\right] \\ &= a_C + b_C \\ &= E\left[\xi_n^C\right] + E\left[\eta_n^C\right] \end{aligned}$$

we're we've introduced ξ_n^C, η_n^C and a_C, b_C for bookkeeping. To prove $\frac{S_n}{n} \to m$ it suffices to show that $E\left[\left|\frac{S_n}{n} \to m\right|\right] \to 0$. (Follows from Chebyshev's inequality.) Write

$$\begin{aligned} E\left[\left|\frac{Sn}{n} - m\right|\right] &= E\left[\left|\xi_n^C + \eta_n^C - a_C - b_C\right|\right] \\ &\leq E\left[\left|\xi_n^C - a_C\right|\right] + \frac{1}{n}\sum_{i=1}^{n} E\left[\left|Y_i^C\right| + \left|b_C\right|\right] \\ &= E\left[\left|\xi_n^C - a_C\right|\right] + E\left[\left|Y_1^C\right|\right] + \left|b_C\right| \\ &\leq \left(E\left[\left|\xi_n^C - a_C\right|^2\right]\right)^{1/2} + 2 \cdot E\left[\left|Y_1^C\right|\right] \end{aligned}$$

since by Cauchy-Schwarz $E\left[|W|\right] \leq \left(E\left[W^2\right]\right)^{1/2}$. Now

$$E\left[\left|\xi_n^C - a_C\right|^2\right] = \mathrm{Var}\left(\xi_n^C\right) = \left(\frac{1}{n}\right)^2 n \cdot \mathrm{Var}\left(X_1^C\right) \to 0$$

as $n \to \infty$, so that

$$\limsup_{n\to\infty} E\left[\left|\frac{Sn}{n} - m\right|\right] \leq 2 \cdot E\left[\left|Y_1^C\right|\right].$$

Now we claim $E\left[\left|Y_1^C\right|\right] \to 0$ as $C \to \infty$. This follows from the dominated convergence theorem since $|X_1| \cdot 1_{\{|X_1| > C\}} \to 0$ almost everywhere. This completes the proof. □

The idea of truncation will appear again when we prove the strong law of large numbers. The second proof is completely different. It uses an important line of reasoning that will show up time and time again (e.g. in the proof of the central limit theorem).

SECOND PROOF. Argue with characteristic functions. Let $\phi(t) = E\left[e^{itX_j}\right]$ (for any j), then set

$$\begin{aligned}\psi_n(t) &= E\left[e^{it\frac{S_n}{n}}\right] \\ &= E\left[e^{it\left(\frac{X_1+\cdots+X_n}{n}\right)}\right] \\ &= \cdots \\ &= \left(E\left[e^{i\frac{tX_1}{n}}\right]\right)^n \\ &= \left(\phi\left(\frac{t}{n}\right)\right)^n.\end{aligned}$$

Since $E[|X_1| < \infty]$, ϕ is differentiable with $\phi'(0) = iE[X_1] = im$. So by Taylor expansion,

$$\phi\left(\frac{t}{n}\right) = 1 + im\frac{t}{n} + o\left(\frac{1}{n}\right).$$

Then

$$\psi_n(t) = \left(1 + im\frac{t}{n} + o\left(\frac{1}{n}\right)\right)^n$$

and the limit as $n \to \infty$ is the same as of $\left(1 + im\frac{t}{n}\right)^n \to e^{imt}$. But the characteristic function of δ_m is e^{imt}, and hence $\frac{S_n}{n} \Rightarrow \delta_m$. □

We used the following result:

LEMMA 2.5. *If a_n is a sequence of complex numbers such that $na_n \to z \in \mathbb{C}$ (i.e. $a_n = z/n + o(1/n)$), then $(1 + a_n)^n \to e^z$.*

Lecture 6, 10/18/11

3. Strong Law of Large Numbers

The conclusion of the strong law of large numbers is that $\frac{S_n}{n} \to m$ almost surely. To prove almost sure limits, a basic (and very important) tool is the Borel-Cantelli lemma.

LEMMA 3.1 (Borel-Cantelli). *Let $\{A_n\}$ be any sequence of events in some probability space $(\Omega, \mathcal{F}, P)$. If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then*

$$P(\{A_n \text{ occurs for only finitely many } n\}) = 1$$

or equivalently

$$P(\{A_n \text{ occurs infinitely often}\}) = 0$$

where $\{A_n \text{ occurs infinitely often}\} = \{\omega : \omega \in A_n \text{ for infinitely many } n\text{'s}\}$. Conversely, if the A_n are (mutually) independent events with $\sum_{n=1}^{\infty} P(A_n) = \infty$, then

$$P(\{A_n \text{ occurs infinitely often}\}) = 1.$$

NOTE. Sometimes we write i.o. to mean "infinitely often".

PROOF. $\{A_n \text{ occurs i.o.}\} = \cap_{k=1}^{\infty}(\cup_{n=k}^{\infty} A_n)$, a decreasing limit as $k \to \infty$ of the sets $\cup_{n=k}^{\infty} A_n$. So by countable additivity we have

$$P(\{A_n \text{ occurs i.o.}\}) = \lim_{k\to\infty} P(\cup_{n=k}^{\infty} A_n) \le \lim_{k\to\infty} \sum_{n=k}^{\infty} P(A_n) = 0$$

since $\sum_{n=1}^{\infty} P(A_n) < \infty$ by assumption.

In the other direction, write

$$\{A_n \text{ occurs for only finitely many } n\} = \cup_{m=1}^{\infty} \left(\cap_{n=m}^{\infty} (A_n)^c\right),$$

which is an increasing limit as $k \to \infty$ of the sets $\cap_{n=m}^{\infty} (A_n)^c$. So

$$\begin{aligned} P(\{A_n \text{ occurs for only finitely many } n\}) &= \lim_{m\to\infty} P\left(\cap_{n=m}^{\infty} (A_n)^c\right) \\ &= \lim_{m\to\infty} \Pi_{n=m}^{\infty} P\left((A_n)^c\right) \\ &= \lim_{m\to\infty} \Pi_{n=m}^{\infty} (1 - P(A_n)) \end{aligned}$$

since by assumption the A_n are independent. Using the fact that $1 - x \leq e^{-x}$ for $x \geq 0$ (proof by calculus) we get

$$\begin{aligned} P(\{A_n \text{ occurs for only finitely many } n\}) &\leq \lim_{m\to\infty} e^{-\sum_{n=m}^{\infty} P(A_m)} \\ &= \lim_{m\to\infty} 0 \\ &= 0 \end{aligned}$$

and hence the claim. □

Exercise 3.2. Check the details of the proof by replacing $\Pi_{n=m}^{\infty} P\left((A_n)^c\right)$ with $\Pi_{n=m}^{l} P\left((A_n)^c\right)$ and letting $l \to \infty$ in the end.

Now we are equipped to prove a first strong law of large numbers.

Theorem 3.3. *If $X_1, X_2, \ldots$ are i.i.d. and $E\left[(X_1)^4\right] < \infty$, then $\frac{S_n}{n} \to m = E(X_1)$ almost surely.*

Proof. By replacing X_i by $X_i - m$, we may assume that $m = 0$. To show $\frac{S_n}{n} \to 0$ almost surely, it suffices to show that for every $\delta > 0$,

$$P\left(\left\{\left|\frac{S_n}{n}\right| \geq \delta \text{ for only finitely many } n\right\}\right) = 1$$

as

$$\left\{\omega : \frac{S_n(\omega)}{n} \to 0\right\} = \cap_{j=1}^{\infty} \left\{\omega : \left|\frac{S_n}{n}\right| \geq \frac{1}{j} \text{ for only finitely many } n\right\},$$

and as the events on the right form a decreasing sequence in j,

$$P\left(\left\{\frac{S_n}{n} \to 0\right\}\right) = \lim_{j\to\infty} P\left\{\left|\frac{S_n}{n}\right| \geq \frac{1}{j} \text{ for only finitely many } n\right\}.$$

Finally by Borel-Cantelli, it suffices to show $\sum_{n=1}^{\infty} P\left(\left|\frac{S_n}{n}\right| > \delta\right) < \infty$.

Since $E[X_1] = 0$ and $E\left[(X_1)^4\right] < \infty$, Chebyshev's inequality with $p = 4$ yields the estimate

$$P\left(\left|\frac{S_n}{n}\right| > \delta\right) \leq \frac{1}{\delta^4} E\left[\left|\frac{S_n}{n}\right|^4\right] = \frac{1}{n^4\delta^4} E\left[|S_n|^4\right].$$

So it will be enough to show $E\left[|S_n|^4\right] \leq B \cdot n^2$ for some constant B. Write

$$\begin{aligned} E\left[|S_n|^4\right] &= E\left[(X_1 + \cdots + X_n)^4\right] \\ &= E\Big[(X_1)^4 + \cdots + (X_n)^4 + 6(X_1)^2 (X_2)^2 + \cdots + 6(X_{n-1})^2 (X_n)^2 \\ &\quad + \text{other terms like } X_1 (X_2)^2 X_3 \text{ or } X_4 (X_7)^3 \text{ or } X_1 X_3 X_5 X_8\Big]. \end{aligned}$$

Now $E\left[(X_i)^4\right] = C < \infty$ for all i, and $E\left[(X_j)^2 (X_k)^2\right] = E\left[(X_j)^2\right] E\left[(X_k)^2\right] = \sigma^2 \cdot \sigma^2 = \sigma^4$ for $j \neq k$ where $\sigma^2 = \mathrm{Var}(X_1) < \infty$ (since $E[X_1] = 0$). On the other hand, $E\left[X_1 (X_2)^2 X_3\right] = E[X_1] E\left[(X_2)^2\right] E[X_3] = 0$ and similarly all other such terms have zero expectation. Thus

$$E\left[|S_n|^4\right] = nC + 6\binom{n}{2}\sigma^4 = nC + 3n(n-1)\sigma^4 \leq B \cdot n^2$$

for some $B < \infty$. This completes the proof. □

We want to claim the law of large numbers holds under weaker conditions. First we'll see some theorems about sums of independent (not neccessarily identically distributed) random variables X_j, which will say when $\sum_{j=1}^{\infty} X_j$ is convergent (almost surely). In order to do this, we need a technical lemma that improves the Chebyshev inequality:

$$P(\{|S_n| \geq l\}) \leq \frac{E\left((S_n)^2\right)}{l^2}.$$

LEMMA 3.4 (Kolmogorov's Inequality). *Suppose $X_1, \ldots, X_n$ are independent with $E\left[(X_j)^2\right] = (\sigma_j)^2 < \infty$ and $E[X_j] = 0$ for all j. Let $T_n(\omega) = \sup_{1 \leq k \leq n} |S_k(\omega)|$. Then,*

$$P(\{T_n \geq l\}) \leq \frac{E\left((S_n)^2\right)}{l^2}.$$

The proof has an interesting structure, which shows up again in the study of Markov chains and martingales.

PROOF. This is a "stopping-time" argument. Consider the first "time" that the sequence $|S_1|, |S_2|, \ldots$ is $\geq l$, and define E_j to be the event that that time is j. Explicitly we have $E_1 = \{|S_1| \geq l\}$, $E_2 = \{|S_1| < l,\ |S_2| \geq l\}$, and so on. Also, $\{T_n \geq l\} = \cup_{k=1}^n E_k$ and $1_{\{T_n \geq l\}} = \sum_{k=1}^n 1_{E_k}$. Now it suffices to show that $P(\{T_n \geq l\}) \leq \frac{1}{l^2} E\left[(S_n)^2 1_{\{T \geq l\}}\right]$, and by our choice of E_k we have $E\left[(S_n)^2 1_{\{T \geq l\}}\right] = \sum_{k=1}^n E\left[(S_n)^2 1_{E_k}\right]$. Now

$$\begin{aligned} P(\{T_n \geq l\}) &= E\left[1_{\{T_n \geq l\}}\right] \\ &= \sum_{k=1}^n E[1_{E_k}] \\ &\leq \sum_{k=1}^n \frac{1}{l^2} E\left[(S_k)^2 1_{E_k}\right] \end{aligned}$$

by our choice of E_k. Going forwards,

$$\begin{aligned} P(\{T_n \geq l\}) &\leq \sum_{k=1}^n \frac{1}{l^2} E\left[\left((S_k)^2 + (S_n - S_k)^2\right) 1_{E_k}\right] \\ &= \sum_{k=1}^n \frac{1}{l^2} E\left[\left((S_n)^2 - 2S_k(S_n - S_k)\right) 1_{E_k}\right]. \end{aligned}$$

But observe

$$E\left[S_k\left(S_n - S_k\right)1_{E_k}\right] = E\left[S_k 1_{E_k}\right] E\left[S_n - S_k\right] = 0$$

by independence and since $E[X_j] = 0$ for all j. So we've shown

$$P\left(\{T_n \geq l\}\right) \leq \sum_{k=1}^{n} \frac{1}{l^2} E\left[\left(S_n\right)^2 1_{E_k}\right],$$

and hence the result. □

Lecture 7, 10/31/11

Now we have Kolmogorov's two- and three-series theorems. These important results describe the convergence of independent series of random variables.

THEOREM 3.5 (Kolmogorov's two-series theorem). *Suppose $X_1, X_2, \ldots$ (on some $(\Omega, \mathcal{F}, P)$) are independent. If $\sum_{i=1}^{\infty} E[X_i]$ and $\sum_{i=1}^{\infty} Var(X_i)$ are convergent, then the series $\sum_{i=1}^{\infty} X_i(\omega)$ converges almost surely.*

REMARK. The two-series theorem gives a sufficient condition for almost sure convergence of series of independent r.v.s. The next theorem (the three-series theorem) will give a neccesary condition.

EXAMPLE 3.6. Recall from calculus that $\sum_{n=1}^{\infty} 1/n^s < \infty$ iff $s > 1$. But $\sum_{n=1}^{\infty} (-1)^n / n^s < \infty$ converges for $s > 0$. What if one takes random signs? Let $Y_1, Y_2, \ldots$ be independent with $P(Y_n = +1) = 1/2$, $P(Y_n = -1) = 1/2$ (e.g. fair coin-tossing). Does $\sum_{n=1}^{\infty} Y_n/n^s$ converge? The intuitive answer is that we still get enough cancellation to allow convergence. By the two-series theorem, the series converges a.s. for $s > 1/2$. Indeed, if we define $X_n = Y_n/n^s$, then $E[X_n] = 0$ and $\mathrm{Var}(X_n) = 1/n^{2s}$ so that $\sum_{n=1}^{\infty} \mathrm{Var}(X_n) < \infty$ if $2s > 1$. After we prove the three-series theorem, we'll be able to say that $\sum_{n=1}^{\infty} X_n$ is a.s. not convergent for $s \leq 1/2$.

PROOF. First, assume $E[X_i] = 0$ for all i. To show a.s. convergence, it suffices to prove for all $\delta > 0$,

$$\lim_{m,n\to\infty} P\left(\sup_{m<k\leq n} |S_k - S_m| \geq \delta\right) = 0$$

where $S_k = X_1 + \cdots + X_k$. By Kolmogorov's inequality applied to the random variables $X_{m+1}, X_{m+1} + X_{m+2}, \ldots, X_{m+1} + \cdots + X_n$,

$$\begin{aligned} P\left(\sup_{m<k\leq n} |S_k - S_m| \geq \delta\right) &\leq \frac{1}{\delta^2} \mathrm{Var}\left(X_{m+1} + \cdots + X_n\right) \\ &= \frac{1}{\delta^2} \cdot \sum_{i=m+1}^{n} \mathrm{Var}\left(X_i\right) \\ &\to 0 \end{aligned}$$

as $m, n \to \infty$.

If the means are non-zero, define $Y_i = X_i - E[X_i]$ so that $X_i = Y_i + E[X_i]$ with $E[Y_i] = 0$. Then apply the previous result to the Y_is and use that $\sum_i E[X_i] < \infty$. □

Suppose now we have independent X_i's to which the two-series theorem does not apply. Note that if $\sum_i \mathrm{Var}(X_i) < \infty$ but $\sum_i E[X_i]$ is divergent, we can still apply the two-series theorem to $Y_i = X_i - E[X_i]$ and argue that $\sum_i X_i = \sum_i (Y_i + E[X_i])$ is divergent. But suppose $\sum_i \mathrm{Var}(X_i) = +\infty$. We can then try a

truncation argument, letting $Y_i = X_i \cdot 1_{\{|X_i| \le C\}}$ with C a fixed constant. If the two-series theorem works for the Y_is, and if $\sum_{i=1}^{\infty} P(X_i \neq Y_i) = \sum_{i=1}^{\infty} P(|X_i| > C) < \infty$, then the Borel-Cantelli lemma implies $P(X_i = Y_i \text{ for all but finitely many } i) = 1$. And then $\sum_i X_i$ converges a.s. because $\sum_i Y_i$ does. This proves half of the next theorem.

THEOREM 3.7 (Kolmogorov's three-series thereom). *Let $X_1, X_2, \ldots$ be independent random variables, and let $C > 0$. Then $\sum_i X_i$ converges almost surely iff*

(1) $\sum_i P(|X_i| > C) < \infty$,
(2) $\sum_i E\left[X_i \cdot 1_{\{|X_i| \le C\}}\right] < \infty$, *and*
(3) $\sum_i Var\left(X_i \cdot 1_{\{|X_i| \le C\}}\right) < \infty$.

We already proved sufficiency; for necessity, see the text.

REMARK. It seems strange that we can take any $C > 0$ in the theorem above, for the series of truncated variables that result are different for different Cs. But as asserted above, they converge simultaneously. Sometimes we can make problems easier by choosing the right C.

Now we state and prove the strong law of large numbers.

THEOREM 3.8 (Strong law of large numbers). *If $X_1, X_2, \ldots$ are i.i.d. with $E[|X_1|] < \infty$, then $\frac{X_1 + \cdots + X_n}{n} \to E[X_1]$ almost surely.*

PROOF SKETCH. There are several steps.

(1) W.l.o.g., assume $E[X_1] = 0$.
(2) Let $Y_n = X_n \cdot 1_{\{|X_n| \le n\}}$ and $b_n = E[Y_n]$. Since $E[|X_1|] < \infty$, we get $\sum_n P(|X_1| > n) < \infty$. As $P(|X_1| > n) = P(|X_n| > n)$, we get $\sum_n P(|X_n| > n) < \infty$ and hence $\sum_n P(X_n \neq Y_n) < \infty$. So by Borel-Cantelli, $P(X_n \neq Y_n \text{ only finitely often}) = 1$. Thus if $\sum_n \frac{Y_n - b_n}{n}$ converges a.s., then so does $\sum_n \frac{X_n - b_n}{n}$.
(3) Since $E[X_n] = 0$, $b_n - E[X_n] = b_n = E\left[|X_1| \cdot 1_{\{|X_1| > n\}}\right] \to 0$ as $n \to \infty$ because $E[|X_1|] < \infty$. (This is by the dominated convergence theorem.)
(4) An elementary lemma about infinite series says that $\sum_n \frac{X_n - b_n}{n} < \infty$ implies $\frac{(X_1 - b_1) + \cdots + (X_n - b_n)}{n} \to 0$. Since $b_n \to 0$ implies $\frac{b_1 + \cdots + b_n}{n} \to 0$, we conclude $\frac{X_1 + \cdots + X_n}{n} \to 0$ as desired.
(5) It remains to show $\sum_n \frac{Y_n - b_n}{n} < \infty$. This is an application of the Kolmogorov three-series theorem. Conditions (1) and (2) in the theorem are immediate. Showing (3) requires an estimate based on the assumption that $E[|X_1|] < \infty$.

□

4. Kolmogorov's 0-1 Law

Consider random variables $X_1, X_2, \ldots$ on the infinite product space $(\Omega, \mathcal{B}, P) = \Pi_{i=1}^{\infty} (\mathbb{R}, \text{Borel sets}, \mu_i)$.

DEFINITIONS 4.1. Let $\mathcal{B}^n$ be the σ-field generated by all events of the form $\{X_j \in D\}$ for $j \ge n$ where D is any Borel set. Then $\mathcal{B}^n$ is called the *σ-field generated by $X_n, X_{n+1}, \ldots$*, or the *future σ-field from time n*. The *tail σ-field* $\mathcal{B}^\infty$ is the σ-field $\mathcal{B}^\infty = \cap_{n=1}^{\infty} \mathcal{B}^n$. Elements of $\mathcal{B}^\infty$ are called *tail events*.

One way to think about the future σ-field is to notice that if $F \in \mathcal{B}^n$, then the definition of F does not depend on $X_1, \ldots, X_{n-1}$. Similarly if $F \in \mathcal{B}^\infty$, then F does not depend on any finite collection of X_is.

EXAMPLE 4.2. Let $\omega = (\omega_1, \omega_2, \ldots)$ and $X_n(\omega) = \omega_n$. The events

$$E_1 = \left\{ \omega : \limsup_{n\to\infty} X_n(\omega) \geq 5 \right\},$$
$$E_2 = \left\{ \omega : \lim_{n\to\infty} \frac{X_1(\omega) + \cdots + X_n(\omega)}{n} = 2 \right\}$$

are tail events.

THEOREM 4.3 (Kolmogorov's 0-1 law). *If $X_1, X_2, \ldots$ are independent and A is a tail event, then either $P(A) = 0$ or $P(A) = 1$.*

EXAMPLE 4.4. Consider the tail event E_2 from the previous example, and suppose the X_i are i.i.d. events. Then if $E[X_1] = 2$, the strong law of large numbers says $P(E_2) = 1$. And if $E[X_1] \neq 2$, the strong law of large numbers says $P(E_2) = 0$. This agrees with the result of the Kolmogorov 0-1 law.

PROOF. The idea is to show that A is independent of A, since then $P(A) = P(A \cap A) = P(A) \cdot P(A)$ and so $(P(A))^2 - P(A) = 0$. Hence $P(A)(P(A) - 1) = 0$ and hence $P(A) = 0$ or $P(A) = 1$.

Let $\mathcal{B}^n$ be the σ-field generated by $X_1, \ldots, X_n$. $A \in B^\infty$ implies A is independent of every event in $\mathcal{B}^n$ for every n. Let $\mathcal{A} = \{\text{events that are independent of } A\}$, then $\mathcal{A}$ is a monotone class (closed under increasing/decreasing limits). Since $\mathcal{A}$ contains each of the fields $\cup_n \mathcal{B}^n$, it contains the whole σ-field $\mathcal{B}$ on the infinite product space. In particular we have $A \in \mathcal{A}$, so A is independent of itself. □

Lecture 8, 11/1/11

Here are some more examples of tail events:

EXAMPLES 4.5. Let $\{X_i\}$ be a sequence of independent events.

(1) $\left\{\omega : \lim_{n\to\infty} \frac{1}{n}\left(\sum X_i(\omega)\right) \leq 3\right\}$ is a tail event.
(2) $\left\{\omega : \limsup_{n\to\infty} \frac{1}{\sqrt{n \log\log n}} \left(\sum_{i=1}^n X_i(\omega) - na\right) = b\right\}$ is a tail event. Here a, b are constants.
(3) In fact, $\{\omega : \lim_{n\to\infty} X_i(\omega) \text{ exists}\}$ is a tail event. In most cases, this event has probability zero. The degenerate case is when the distributions approach a point mass.
(4) $\left\{\omega : \sup_n \frac{1}{n} \sum_{i=1}^n X_i(\omega) \leq 3\right\}$ is not a tail event.

5. Central Limit Theorem

Let $X_1, X_2, \ldots$ be i.i.d. r.v.'s and let $S_n = X_1 + \cdots + X_n$ be the nth partial sum. Our best law of large numbers says that if $\mu = E[|X_1|] < \infty$, then $S_n/n \to \mu$ with probability one. We can reinterpret this as saying $\frac{1}{n}(S_n - n\mu) \to 0$ with probability one. This suggests the question: how fast does $\frac{1}{n}(S_n - n\mu)$ tend to zero? The central limit theorem gives an answer, at least when $E\left[(X_1)^2\right] < \infty$. Then, as we'll see, $\frac{1}{n}(S_n - n\mu) \to 0$ roughly like order $1/\sqrt{n}$. More precisely, the distribution of $\frac{\sqrt{n}}{n}(S_n - n\mu) = \frac{1}{\sqrt{n}}(S_n - n\mu)$ does not tend to δ_0 as $n \to \infty$ (except in the trivial case where the common distribution of the X_is is δ_μ).

THEOREM 5.1 (Central limit theorem). *Let $X_1, X_2, \ldots$ be i.i.d. r.v.'s with $E\left[(X_1)^2\right] < \infty$ and call $\mu = E[X_1]$. Then the distribution of $\frac{1}{\sqrt{n}}(S_n - n\mu)$ converges weakly to the normal distribution with mean zero and variance $\sigma^2 = E\left[(X_1 - \mu)^2\right] = Var(X_1)$.*

REMARK. In the nontrivial case $\sigma \neq 0$, we can read the theorem as saying

$$P\left(\frac{S_n - n\mu}{\sqrt{n}} \leq u\right) \to \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}\, dx$$

for every $u \in \mathbb{R}$.

PROOF. The proof is an exercise in characteristic functions. Recall that

(1) The characteristic function of a normal r.v. W with mean μ and variance σ^2 is $\psi(t) = E\left[e^{itW}\right] = e^{i\mu t - \frac{1}{2}\sigma^2 t^2}$.
(2) Distributions of r.v.'s W_n converge weakly to the distribution of W iff $\psi_n(t) = E\left[e^{itW_n}\right] \to \psi(t) = E\left[e^{itW}\right]$ for all real t.
(3) If $Y_1, Y_2, \ldots, Y_n$ are independent, then $E\left[e^{it(\alpha_1 Y_1 + \cdots + \alpha_n Y_n)}\right] = \Pi_{i=1}^n E\left[e^{i(\alpha_i t)Y_i}\right]$.
(4) If $E\left[Y^2\right] < \infty$, then

$$\begin{aligned} E\left[e^{itY}\right] &= 1 + itE[Y] + \frac{(it)^2}{2} E\left[Y^2\right] + o\left(t^2\right) \\ &= 1 + iE[Y]\, t - \frac{E\left[Y^2\right]}{2} t^2 + o\left(t^2\right) \end{aligned}$$

as $t \to 0$. See exercise 2.4 in the text; see also the second proof of the weak law of large numbers (theorem 2.4).

Now the proof. By (1) and (2), we only need to prove that $E\left[e^{it\frac{1}{\sqrt{n}}(S_n - n\mu)}\right] \to e^{-\frac{1}{2}\sigma^2 t^2}$ for all $t \in \mathbb{R}$. To study $\frac{1}{\sqrt{n}}(S_n - n\mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu)$, let $Y_i = X_i - \mu$ and note that $E[Y_i] = E[X_i] - \mu = 0$ and $E\left[(Y_i)^2\right] = \mathrm{Var}(X_i) = \sigma^2$. Then by (3),

$$\begin{aligned} E\left[e^{it\frac{1}{\sqrt{n}}(S_n - n\mu)}\right] &= E\left[e^{it\sum_{i=1}^n \frac{Y_i}{\sqrt{n}}}\right] \\ &= \Pi_{i=1}^n E\left[e^{it\frac{Y_i}{\sqrt{n}}}\right] \\ &= \left(\phi\left(\frac{t}{\sqrt{n}}\right)\right)^n \end{aligned}$$

where $\phi(t) = E\left[e^{itY_1}\right]$. Finally by (1),

$$\phi(t) = 1 - \frac{\sigma^2}{2} t^2 + o\left(t^2\right)$$

as $t \to 0$, so for fixed $t \in \mathbb{R}$

$$\begin{aligned} \left[\phi\left(\frac{t}{\sqrt{n}}\right)\right]^n &= \left[1 - \frac{\sigma^2}{2}\frac{t^2}{n} + o\left(\frac{t^2}{n}\right)\right]^n \\ &= \left[1 - \frac{\sigma^2}{2}\frac{t^2}{n} + o\left(\frac{1}{n}\right)\right]^n \\ &\to e^{-\frac{1}{2}\sigma^2 t^2} \end{aligned}$$

by lemma 2.5. This completes the proof. □

There are a variety of extensions of the central limit theorem. In what follows we'll assume $X_1, X_2, \ldots$ are independent with $E\left[X_i^2\right] < \infty$, but not that they are identically distributed. To simplify the discussion, let's assume $E[X_i] = 0$ for all i so that $\mathrm{Var}(X_i) = E\left[X_i^2\right]$. (Of course, we can always replace X_i by $Y_i = X_i - E[X_i]$ to get to this case.) Let $S_n = X_1 + \cdots + X_n$ and let $s_n^2 = \mathrm{Var}(S_n) = \sum_{i=1}^n \mathrm{Var}(X_i) = \sum_{i=1}^n \sigma_i^2$. In the i.i.d. case, $\sigma_n^2 = n\sigma^2$ and the central limit theorem says (for $\sigma^2 \neq 0$) that $S_n/s_n \to N(0,1)$ in distribution.[2] In the extended setting, can we conclude $S_n/s_n \to N(0,1)$ in distribution? The next example shows we need some more assumptions.

EXAMPLE 5.2. Let $X_1', X_2', \ldots$ be i.i.d. and (to be concrete) with values ± 1 with probability 1/2. Let $X_j = \sigma_j X_j'$ with $\sum_{j=1}^\infty \sigma_j^2 < \infty$ (e.g. with $\sigma_j = j^{-2}$). Then by the Komogorov 2- or 3-series theorem, since $s_n^2 = \sum_{j=1}^n \mathrm{Var}(X_j) = \sum_{j=1}^n \sigma_2^2 \to s_\infty^2 = \sum_{j=1}^\infty \sigma_j^2 < \infty$ we have that $S_n \to \sum_{j=1}^\infty \sigma_j X_j'$ a.s. and so $S_n/s_n \to \frac{1}{s_\infty}\sum_{j=1}^\infty \sigma_j X_j'$ a.s. and hence also in distribution. But this limit is not $N(0,1)$-distributed.

There are (at least) two ways to see why the limit is not $N(0,1)$-distributed. The first is to consider the characteristic function of the limit. The second is to observe that in the $\sigma_j = j^{-2}$ case, the limit is a bounded random variable.

The example above went wrong because $\sum_{j=1}^\infty \sigma_j^2 < \infty$. So it seems we (at least) need to assume $\sum_{j=1}^\infty \sigma_j^2 = \infty$. Is this enough?

EXAMPLE 5.3. Let

$$X_j = \begin{cases} +j & \text{with probability } p_j/2 \\ -j & \text{with probability } p_j/2 \\ 0 & \text{with probability } 1 - p_j \end{cases}.$$

Then $\sigma_j^2 = E\left[(X_j)^2\right] = j^2 p_j$. Suppose $\sum_{j=1}^\infty j^2 p_j = \infty$ but also that $\sum_{j=1}^\infty P_j < \infty$ (e.g. $p_j = j^{-2}$). Then by Borel-Cantelli (or the 3-series theorem), $\sum_{j=1}^\infty X_j$ is a.s. convergent, since it is a.s. a finite sum. So $S_n/s_n \to 0$ a.s. and therefore not to $N(0,1)$ in distribution.

The problem in the example above is that although $s_\infty = \sum_{j=1}^\infty \sigma_j^2 = \infty$, the main contribution to $s_n = \sum_{j=1}^n \sigma_j^2$ comes from very large values of X_j. To avoid this phenomenon, one imposes the "Lindeberg condition": If α_i denotes the distribution of X_i, we require

$$\frac{1}{s_n^2}\sum_{i=1}^n \int_{|x| \geq \epsilon \cdot s_n} x^2 \, d\alpha_i \to 0$$

for every $\epsilon > 0$. This says that the contribution to the variance from very large values of X_j is negligible. (Recall for $E[X_i] = 0$, $\mathrm{Var}(X_i) = \int_{\mathbb{R}} x^2 \, d\alpha_i$.) As we'll see, Lindeberg's condition plus the requirement that $s_n \to \infty$ are enough to imply that $S_n/s_n \to N(0,1)$ in distribution. Lecture 9, 11/8/11

THEOREM 5.4 (Lindeberg's CLT). *Let $X_1, X_2, \ldots$ be independent with $E[X_j] = 0$ and $E\left[(X_j)^2\right] = Var(X_j) = \sigma_j^2 < \infty$ and let $S_n = X_1 + \cdots + X_n$ and $s_n^2 =$*

[2] $N(0,1)$ means the standard normal distribution, with mean 0 and variance 1.

$\sigma_1^2 + \cdots + \sigma_n^2 = Var(S_n)$. *Then* $S_n/s_n \Rightarrow N(0,1)$ *if* $s_n \to \infty$ *and if for all* $\epsilon > 0$,

$$\frac{1}{s_n^2}\sum_{j=1}^{n} E\left[X_j^2 \cdot 1_{\{|X_j/s_n|>\epsilon\}}\right] \to 0 \tag{5.1}$$

as $n \to \infty$.

REMARK. Condition (5.1) is called the *Lindeberg condition*.

PROOF. We have $\frac{S_n}{s_n} = \frac{X_1}{s_n} + \cdots + \frac{X_n}{s_n}$ and suppose X_j/s_n has the characteristic function $\phi_{n,j}$. ($\{\phi_{n,j}\}$ is an example of a triangular array.) We need to show $\Pi_{j=1}^n \phi_{n,j}(t) \to e^{-t^2/2}$ as $n \to \infty$ for any fixed $t \in \mathbb{R}$. The trick: as $\phi_{n,j}(t)$ is close to 1 for large n (by the Lindeberg condition), write

$$\phi_{n,j}(t) = (1 + (\phi_{n,j}(t) - 1)) = e^{\phi_{n,j}(t)-1} + O\left(\phi_{n,j}(t) - 1\right)^2$$

as $n \to \infty$. So replace $\phi_{n,j}$ with $\psi_{n,j}(t) = e^{\phi_{n,j}(t)-1}$; one can show it suffices to prove $\Pi_{j=1}^n \psi_{n,j}(t) = e^{\sum_{j=1}^n (\phi_{n,j}(t)-1)} \to e^{-t^2/2}$, or equivalently

$$K_n = \sum_{j=1}^{n}\left(\phi_{n,j}(t) - 1 + \frac{\sigma_j^2 t^2}{2s_n^2}\right) \to 0$$

for all t.

Observe that

$$|K_n| \le \sum_{j=1}^{n} \Big| E\left[e^{itX_j/s_n} - 1 + -it\frac{X_j}{s_n} + \frac{t^2}{2}\frac{(X_j)^2}{s_n^2}\right]\Big|,$$

since $E[X_j] = 0$ for all j. Now we'll write each term as an integral w.r.t. α_j, the distribution of X_j, break up the integral according to where $|x| < \epsilon s_n$ and $|x| \ge \epsilon s_n$ (with $\epsilon < 1$), and use the bounds

$$|e^{ir} - 1 - ir + \frac{r^2}{2}| \le \begin{cases} C|r|^3 & |r| < t \\ \tilde{C}|r|^2 & \text{otherwise} \end{cases}$$

with $r = tx/s_n$. So

$$\begin{aligned} |K_n| &\le Ct^3 \sum_{j=1}^{n} \int_{\{|x|<\epsilon s_n\}} \frac{|x|^3}{s_n^3}\, d\alpha_j + C't^2 \sum_{j=1}^{n} \int_{\{|x|\ge\epsilon s_n\}} \frac{x^2}{s_n^2}\, d\alpha_j \\ &= I + II. \end{aligned}$$

Notice the Lindeberg condition says exactly that $II \to 0$. And

$$\begin{aligned} \int_{\{|x|<\epsilon s_n\}} \frac{|x|^3}{s_n^3}\, d\alpha_j &= \int_{\{|x|<\epsilon s_n\}} \epsilon \frac{|x|^2}{s_n^2}\, d\alpha_j \\ &\le \epsilon \int_{\mathbb{R}} \frac{x^2}{s_n^2}\, d\alpha_j \\ &= \epsilon \frac{\sigma_j^2}{s_n^2} \end{aligned}$$

which gives $|K_n| \le \epsilon Ct^3$ after summing on j. Thus $\limsup_{n\to\infty} |K_n| \le D\epsilon$ for all $\epsilon > 0$, and so $\lim_{n\to\infty} |K_n| = 0$. This completes the proof. □

6. Triangular Arrays and Infinite Divisibility

The proof of the Lindeberg CLT used an example of a triangular array. Here is another example.

EXAMPLE 6.1. Binomial approximation of the Poisson distribution. Let S_n be the number of successes in n independent trials with probability p_n of success on each trial.

CLAIM. If $np_n \to \lambda \in (0, \infty)$, then $S_n \Rightarrow \text{Poisson}\,(\text{mean } \lambda)$.

PROOF. (By characteristic functions.) Write $S_n = X_{n,1} + \cdots + X_{n,n}$ with $X_{n,j}$ independent Bernoulli variables with parameter p_n. Then

$$\begin{aligned} E\left[e^{itS_n}\right] &= \left(E\left[e^{itX_{1,n}}\right]\right)^n \\ &= \left((1-p_n) + p_n e^{it}\right)^n \\ &= \left(1 + p_n\left(e^{it}-1\right)\right)^n \\ &= \left(1 + \frac{\lambda}{n}\left(e^{it}-1\right) + o\left(\frac{1}{n}\right)\right)^n \\ &\to e^{\lambda\left(e^{it}-1\right)} \end{aligned}$$

which is the characteristic function of Poisson (λ). Hence the claim. □

Now we have two examples of triangular arrays, one in which we get a normally distributed limit and one in which we get a Poisson limit. This is the beginning of the theory of triangular arrays.

DEFINITION 6.2. A *triangular array* consists of random variables

$$\begin{matrix} X_{1,1}, \ldots, X_{1,k_1} \\ X_{2,1}, \ldots, X_{2,k_2} \\ \vdots \end{matrix}$$

NOTE. Often $k_n = n$.

We'll consider the case where the random variables in each row are independent, and ask whether $S_n = X_{n,1} + \cdots + X_{n,k_n}$ has a limiting distribution as $n \to \infty$.

EXAMPLES 6.3.

(1) In Lindeberg CLT, $k_n = n$, $X_{n,j} = X_j/s_n$, and the limit is $N(0,1)$.
(2) In the binomial approximation of the Poisson distribution, $k_n = n$,

$$X_{n,j} = \begin{cases} 1 & \text{probability } p_n \\ 0 & \text{probability } (1-p_n) \end{cases},$$

and the limit is Poisson (λ).

Besides independence in each row, we want no individual summand to contribute too much (as $n \to \infty$). So we'll assume that for all $\delta > 0$,

$$\sup_{1 \le j \le k_n} P\left(|X_{n,j}| \ge \delta\right) \to 0$$

as $n \to \infty$. This condition is called *uniform infinitesimality* or *unifomly asymptotically negligible*, and is analogous to the Lindeberg condition (which says individual summands do not contribute too much to the variance).

Through all of this, we seek to answer the questions:

(1) What kinds of limits arise?
(2) How do we determine which limiting distribution occurs?

We start to answer (1) by defining a class of random variables (or distributions).

DEFINITION 6.4. A random variable Y (or its distribution) is called *infinitely divisible* if for every n there exsit i.i.d. random variables $X_{n,1}, \ldots, X_{n,n}$ so that $\sum_{j=1}^{n} X_{n,j}$ has the same distribution as Y.

NOTE. An equivalent definition says that for each n there is a characteristic function ϕ_n so that $(\phi_n(t))^n = E\left[e^{itY}\right]$ for all $t \in \mathbb{R}$. Note, however, that the nth root of a characteristic function is not always a characteristic function.

EXAMPLES 6.5.

(1) The characteristic function of the Poisson distribution is $\phi(t) = e^{\lambda\left(e^{it}-1\right)}$. Notice $\phi_n(t) = e^{\frac{\lambda}{n}\left(e^{it}-1\right)}$ is the characteristic function of the Poisson distribution with mean λ/n, and $(\phi_n)^n = \phi$ so the Poisson distribution is infinitely divisible.
(2) The characteristic function of the normal distribution is

$$\phi(t) = e^{i\mu t - \frac{\sigma^2}{2}t^2} = \left(e^{i\frac{\mu}{n}t - \frac{\sigma^2}{n}\frac{t^2}{2}}\right)^n.$$

So the normal distribution is infinitely divisible.
(3) The symmetric stable distribution has characteristic function

$$\phi(t) = e^{-c|t|^\alpha} = \left(e^{-\frac{c}{n}|t|^\alpha}\right)^n.$$

So it is infinitely divisible as well.

As it turns out, the set of infinitely divisible distributions is the answer to (1). Moreover, there is an explicit representation of the characteristic functions of infinitely divisible distributions due to Levy and Khinchine. But first, we'll discuss an important example.

DEFINITION 6.6. Let $X_1, X_2, \ldots$ be i.i.d. with common distribution α, and let N be Poisson (λ) and independent of the X_js. (All the r.v.s are defined on the same probability space.) Then the random variable $Y = X_1 + \cdots + X_N$ is called a *compound Poisson distribution.*

This is a particular example of a sum of a random number of i.i.d. random variables.

EXAMPLE 6.7. Let N be the number of mortgage defaults in a certain time period to a certain mortgage provider. Let X_i be the dollar amount of the ith default. Then Y is the total amount of all defaults in that time period.

Let's compute the characteristic function of a compound Poisson Y. We have

$$\begin{aligned}
E\left[e^{itY}\right] &= \sum_{k=0}^{\infty} E\left[e^{itY}\cdot 1_{\{N=k\}}\right] \\
&= \sum_{k=0}^{\infty} E\left[e^{it\sum_{j=1}^{k}X_k}\cdot 1_{\{N=k\}}\right] \\
&= \sum_{k=0}^{\infty} E\left[e^{it\sum_{j=1}^{k}X_k}\right] E\left[1_{\{N=k\}}\right] \\
&= \sum_{k=0}^{\infty} P\left(\{N=k\}\right)\cdot\left(E\left[e^{itX_1}\right]\right)^k \\
&= \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}\left(\phi_\alpha(t)\right)^k \\
&= e^{\lambda(\phi_\alpha(t)-1)} \\
&= e^{\int_{\mathbb{R}}\left(e^{itx}-1\right)\lambda\, d\alpha(x)} \\
&= e^{\int_{\mathbb{R}\setminus\{0\}}\left(e^{itx-1}\right) d[\lambda\alpha](x)}.
\end{aligned}$$

Notice we can replace $\lambda\alpha$ by any positive, finite measure M on $\mathbb{R}\setminus\{0\}$. So this distribution is infinitely divisible, for

$$e_\lambda(\alpha) = E\left[e^{itY}\right] = e^{\int_{\mathbb{R}\setminus\{0\}}\left(e^{itx-1}\right) dM(x)} = \left(e^{\int_{\mathbb{R}\setminus\{0\}}\left(e^{itx-1}\right) d\left[\frac{M}{n}\right](x)}\right)^n$$

for all n.

PROPOSITION 6.8. *If Y and Z are independent and infinitely divisible, then $Y+Z$ is infinitely divisible.*

So by adding a compound Poisson r.v. and a normal r.v., we get an infinitely divisible random variable with characteristic function

$$E\left[e^{itX}\right] = \exp\left[\int\left(e^{itx}-1\right) dM(x) + iat - \frac{\sigma^2t^2}{2}\right]$$

where M is any finite measure on $\mathbb{R}\setminus\{0\}$. This is a large class of infinitely divisible random variables. Can it be extended? First, define

$$\theta(x) = \begin{cases} x & |x|\le 1 \\ 1 & x>1 \\ -1 & x<1 \end{cases}.$$

(Alternatively we could use $\tilde{\theta}(x) = x/(1+x)^2$.) Then rewrite $E\left[e^{itX}\right]$ as

$$E\left[e^{itX}\right] = \exp\left[\int\left(e^{itx}-1-it\theta(x)\right) dM(x) + ia't - \frac{\sigma^2t^2}{2}\right]. \tag{6.1}$$

PROPOSITION 6.9. *The limit of infinitely divisible characteristic functions is an infinitely divisible characteristic function.*

From this we conclude that M need not be a finite measure near the origin, as long as it integrates against $e^{itx}-1-it\theta(x) \sim Cx^2$ for small x.

DEFINITION 6.10. An *admissible Levy measure* is a (possibly infinite) positive measure M on $\mathbb{R}\backslash\{0\}$ so that

- $\int_{|x|>\epsilon} dM < \infty$ for all $\epsilon > 0$,
- $\int_{|x|<K} x^2\, dM < \infty$ for all $K > 0$.

NOTE. These two conditions are equivalent to $\int \tilde{\theta}\, dM = \int \frac{x^2}{1+x^2}\, dM(x) < \infty$.

THEOREM 6.11 (Levy-Khintchine). *X is infinitely divisible iff* (6.1) *holds with $a \in \mathbb{R}$, $\sigma^2 \geq 0$, and M admissible.*

Lecture 10
11/15/11

REMARK. If M is a measure with finite total mass, then the integral term is exactly compound Poisson. The theorem allows mass to collect at 0 in the limit (but not so much that x^2 is not integrable). The term $\theta(x)$ ensures the integrand is quadratic near the origin and bounded near infinity.

EXAMPLE 6.12. The measure

$$dM(x) = \frac{c}{|x|^{1+\alpha}}\, dx$$

is a Levy measure for $0 < \alpha < 2$.

EXERCISE 6.13. Take $\sigma = 0$ in the example above. Then the term $it\theta(x)$ is unneccesary in the integral. Show that the resulting distributions for varying $0 < \alpha < 2$ are symmetric stable distributions.

We'll denote by $e\left(M, \sigma^2, a\right)$ the probability measure with characteristic function given by (6.1). The next theorem says that the set of distributions arising from triangular arrays with uniform infinitesimality is exactly the set of infinitely divisible distributions. It is taken from Frustedt-Gray, and combines the results of sections 3.7, 3.8 in that text.

THEOREM 6.14. *Let $(X_{n,j} : 1 \leq j \leq k_n)$ be a triangular array that is independent in each row and uniformly infinitesimal, i.e. suppose*

$$\sup_{1\leq j\leq k_n} P(|X_{n,j}| \geq \delta) \to 0$$

for all $\delta > 0$. Then if $S_n = \sum_{i=1}^n X_{n,k_n}$ converges in distribution, the limit must be infinitely divisible. In order that the limit is $e\left(M, \sigma^2, a\right)$, it is neccesary and sufficient that:

(1) For every bounded continuous function f which vanishes in a neighborhood around zero,

$$\sum_{j=1}^{k_n} E[f(X_{n,j})] \to \int f\, dM.$$

(2) Let

$$\sigma_n^\epsilon = \sqrt{\sum_{j=1}^{k_n} Var\left(X_{n,j} \cdot 1_{\{|X_{n,j}|\leq\epsilon\}}\right)},$$

then σ is given by

$$\lim_{\epsilon\to 0} \limsup_{n\to\infty} \sigma_n^\epsilon = \lim_{\epsilon\to 0} \liminf_{n\to\infty} \sigma_n^\epsilon = \sigma.$$

(3) a is given by

$$\lim_{n\to\infty}\sum_{j=1}^{k_n} E\left[\theta\left(X_{n,j}\right)\right] = a.$$

REMARK. The main idea of the proof is to replace the distribution $\alpha_{n,j}$ by the infinitely disivible distribution $e\left(\alpha'_{n,j}, 0, a''_{n,j}\right)$. Here, $\alpha'_{n,j}$ is $\alpha_{n,j}$ with centering and $a''_{n,j}$ is chosen to compensate for the centering. It turns out this produces the same limit, but computation is much easier with infinitely divisible distributions.

7. Law of the Iterated Logarithm

In this section we'll arrive at a fine (and technical) peice of information about sums of random variables, which will help to explain the relation between limits of distributions and almost sure limits. Let $X_1, X_2, \ldots$ be i.i.d. with $E[X_i] = 0$, $E\left[(X_i)^2\right] = 1$, and let $S_n = X_1 + \cdots + X_n$. Then $S_n/n \to 0$ almost surely while $S_n/\sqrt{n} \Rightarrow N(0,1)$. Does $S_n/\sqrt{n}$ have an almost sure limit? The answer is no.

LEMMA 7.1. *Define $Z(\omega) = \limsup_{n\to\infty} S_n(\omega)/\sqrt{n}$, then $Z(\omega) = +\infty$ almost surely.*

PROOF. Z is measurable w.r.t. the tail field. So Kolmogorov's 0-1 law implies that for any a, $P(Z \geq a)$ is either zero or one. But we can't have $P(Z \geq a) = 0$ for any $a < \infty$, for if so then $S_n/\sqrt{n} \leq 2a$ for all large n with probability one, and then $P(S_n/\sqrt{n} \leq 2a) \to 1$. This violates the central limit theorem. □

REMARKS.

(1) At first glance, the lemma looks like it contradicts the central limit theorem. But the central limit theorem implies that for some large n (not depending on ω) $S_n/\sqrt{n} \geq 100$ with very small probability, approximately $\int_{100}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$. The above lemma says that given ω, we can find a sequence N_j (depending on ω) going to infinity such that $S_{N_j}(\omega)/\sqrt{N_j(\omega)} \geq 100$. These are different questions. The central limit theorem says that the N_j's are sparse as $j \to \infty$, but the lemma says they still exist.
(2) Replacing X_j by $-X_j$ in the lemma shows that $\liminf_{n\to\infty} S_n/\sqrt{n} = -\infty$ almost surely.

The lemma says $\sup_{m\leq n}\{S_m/\sqrt{m}\} \to +\infty$ almost surely. How fast does this happen? The law of large numbers implies that this must diverge slower than $\sqrt{n}$. In fact, it diverges very very slowly, like $\sqrt{\log\log n}$.

THEOREM 7.2 (Law of the Iterated Logarithm). *Let X_i be i.i.d. with $E[X_i] = 0$ and $E\left[(X_i)^2\right] = 1$. Then,*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n}\sqrt{\log\log n}} = \sqrt{2}$$
$$\liminf_{n\to\infty} \frac{S_n}{\sqrt{n}\sqrt{\log\log n}} = -\sqrt{2}$$

almost surely.

PROOF SKETCH. We can get the second limit by replacing X_j with $-X_j$. So we'll only prove the first limit. As in the text, we'll assume that $E\left[|X|^{2+\alpha}\right] < \infty$ for some $\alpha > 0$. There are two parts.

Part I. Give the proof for the special case where the X_i are normal. Normality is only used to show that for $S_n/\sqrt{n}$ standard normal and $a > 0$,

(1) For $p < a^2/2$, there exists $C_p < \infty$ so that

$$P\left(\frac{S_k}{\sqrt{k}} \geq a\sqrt{\log\log k}\right) \leq C_p \left(\log k\right)^{-p}$$

for all large k.

(2) For $p > a^2/2$, there exists $C'_p < \infty$ so that

$$P\left(\frac{S_k}{\sqrt{k}} \geq a\sqrt{\log\log k}\right) \geq C'_p \left(\log k\right)^{-p}$$

for all large k.

These are precise estimates on the tail of the standard normal distribution.

Part II. Extend the proof to general distributions. Show that the estimates above are approximately correct with errors small enough to make no difference in the Borel-Cantelli arguments used in part I. The approximate validity of the estimates follows from an interesting extension of the central limit theorem, due to Berry and Esseen. See the following theorem.

We'll discuss part I now. Define $\phi(n) = \sqrt{n\log\log n}$, we need to show

$$I_{\text{upper}}: \ \limsup_n \frac{S_n}{\phi(n)} \leq \sqrt{2}$$
$$I_{\text{lower}}: \ \limsup_n \frac{S_n}{\phi(n)} \geq \sqrt{2}$$

almost surely. For both, consider blocks of S_j's for j between k_{n-1} and k_n with $k_n \approx \rho^n$ with $n > 1$ (so the blocks become large). Suppose we can show that for any $\lambda > \sqrt{2}$,

$$\sum_{n=1}^{\infty} P\left(\sup_{k_{n-1}\leq j\leq k_n} S_j \geq \lambda\phi\left(k_{n-1}\right)\right) < \infty. \tag{7.1}$$

Then by Borel-Cantelli,

$$\limsup_{n\to\infty} \frac{\sup_{k_{n-1}\leq j\leq k_n} S_j}{\phi\left(k_{n-1}\right)} \leq \lambda$$

almost surely. Since $\phi(n)$ is increasing in n, this implies that

$$\limsup_{n\to\infty} \frac{S_n}{\phi(n)} \leq \lambda$$

as desired. By the inequality of problem two, problem set eight (see p. 67), we can replace (7.1) by

$$\sum_{n=1}^{\infty} P\left(S_{k_n} \geq \tilde{\lambda}\phi\left(k_{n-1}\right)\right) < \infty$$

where $\sqrt{2} < \tilde{\lambda} \leq \lambda$. By our choice of $k_n \approx \rho^n$, with ρ sufficiently close to one, this can be replaced by

$$\sum_{n=1}^{\infty} P\left(S_{k_n} \geq \tilde{\tilde{\lambda}}\phi\left(k_n\right)\right) < \infty$$

with $\sqrt{2} < \tilde{\tilde{\lambda}}$ still. This follows from estimate (1) above.

The lower bound is somewhat similar, and uses the other half of Borel-Cantelli applied to $Y_n = S_{k_{n+1}} - S_{k_n}$.

□

THEOREM 7.3 (Berry-Esseen). *Let X_i be i.i.d with mean zero and variance one. If $E\left[|X|^{2+\alpha}\right] < \infty$ for some $\alpha > 0$, then there exists $C < \infty$, $\delta > 0$ so that*

$$\sup_{-\infty<b<\infty} \left|P\left(\frac{S_k}{\sqrt{k}} \geq b\right) - P\left(Z > b\right)\right| \leq \frac{C}{k^{\delta}}$$

where Z is the standard normal distribution.

The proof of the Berry-Esseen CLT is an extended exercise in characteristic functions. See pp. 69-71 of the text.

CHAPTER 4

Dependent Random Variables

1. Conditioning

Lecture 11
11/22/11

Let us begin with the discrete case. If X, Y are random variables with countable many values, we can define their *joint distribution function*

$$f(x,y) = P(X = x, Y = y).$$

Then the *conditional probability* of X given Y is

$$P\left(X = x \,\middle|\, Y = y\right) = \frac{f(x,y)}{P(Y=y)} = \frac{f(x,y)}{\sum_{x'} f(x',y)} = f\left(x \,\middle|\, y\right).$$

For each y-value, this gives a probability distribution on x-values that depends on y. (Later we will discuss a generalization, called "regular conditional distribution" that works for general real-valued random variables.) Given $f\left(x \,\middle|\, y\right)$, we can define the *conditional expectation*

$$E\left[X \,\middle|\, Y = y\right] = \sum_x x f\left(x \,\middle|\, y\right),$$

which depends on y. We'll want to look at this as a function of Y,

$$E\left[X \,\middle|\, Y\right] = \sum_x f\left(x \,\middle|\, Y\right).$$

This can be generalized to the notion of conditional expectation of X with respect to Σ, written $E\left[X \,\middle|\, \Sigma\right]$. Σ here is sometimes denoted $\sigma(Y)$, and is the σ-field generated by Y, i.e. the smallest σ-field for which Y is measurable. More generally, we have $E\left[X \,\middle|\, \mathcal{G}\right]$, where $\mathcal{G}$ is any sub-σ-field of the original σ-field in the probability space on which X was defined, and where X is any random variable with $E[|X|] < \infty$.

In the discrete case, we have the following property. Let A be a set of values taken on by Y, then

$$\begin{aligned} E\left[E\left[X \,\middle|\, Y\right] \cdot 1_{\{Y \in A\}}\right] &= \sum_{y \in A} E\left[X \,\middle|\, Y = y\right] P(Y = y) \\ &= \sum_{y \in A} \sum_x x \frac{f(x,y)}{P(Y=y)} P(Y = y) \\ &= E\left[X \cdot 1_{\{Y \in A\}}\right]. \end{aligned}$$

This identity will be a defining property of the conditional expectation $E\left[X \,\middle|\, Y\right]$ (likewise $E\left[X \,\middle|\, \Sigma\right]$), along with being a function of Y (or being Σ-measurable).

Definitions 1.1. Let $(\Omega, \mathcal{F}, P)$ be a probability space, let Σ be a sub-σ-field of $\mathcal{F}$, and let X be a ($\mathcal{F}$-measurable) random variable with $E[|X|] < \infty$. The

conditional expectation of X given Σ, denoted $E\left[X \,\middle|\, \Sigma\right]$, is any random variable W on $(\Omega, \mathcal{F}, P)$ with the following properties:

- W is Σ-measurable,
- For all $A \in \Sigma$, $E[X \cdot 1_A] = E[W \cdot 1_A]$.

Given $B \in \mathcal{F}$, we define the *conditional probability* of B given Σ by

$$P\left(B \,\middle|\, \Sigma\right) = E\left[1_B \,\middle|\, \Sigma\right].$$

NOTE. We can change a candidate W on a set of measure zero without affecting the result. Also note that $E\left[X \,\middle|\, Y\right]$ from before is $E\left[X \,\middle|\, \Sigma\right]$ where Σ is the pre-image of the Borel sets under Y. (The smallest σ-field on which Y is measurable.)

PROPOSITION 1.2. *If W and $\tilde{W}$ are both conditional expectations of X given Σ, then $W = \tilde{W}$ almost surely.*

So conditional expectation is defined exactly up to a set of measure zero (but is otherwise unique). We have yet to show the general existence of $E\left[X \,\middle|\, \Sigma\right]$. Here are some special cases:

(1) If X is already Σ-measurable, then we can take $E\left[X \,\middle|\, \Sigma\right] = X$.
(2) If X is independent of Σ, i.e. if for $A \in \Sigma$ we have $E[X \cdot 1_A] = E[X] \cdot E[1_A] = E[X] \cdot P(A)$, we can take $E\left[X \,\middle|\, \Sigma\right] = E[X]$. In this case, $E\left[X \,\middle|\, \Sigma\right]$ is almost surely a constant. But since X is independent of Σ, conditioning does not give any more information about X.
(3) If $E\left[X^2\right] < \infty$, then by regarding X as an element of $L^2(\Omega, \mathcal{F}, P)$ we can take $E\left[X \,\middle|\, \Sigma\right]$ as being obtained from X by orthogonal projection onto the subspace $L^2(\Omega, \Sigma, P) \subset L^2(\Omega, \mathcal{F}, P)$. See exercise 4.9 in the text.

This last condition $E\left[X^2\right] < \infty$ is unnatural from the view of probability theory. In what follows, we'll show that $E\left[X \,\middle|\, \Sigma\right]$ exists provided only that $E[|X|] < \infty$. Instead of starting with the L^2 case and approximating the L^1 case, we'll proceed via the Radon-Nikodym theorem.

DEFINITION 1.3. Let λ, μ be finite, non-negative measures on $(\Omega, \mathcal{F})$. We'll say λ is *absolutely continuous* w.r.t. μ ($\lambda \ll \mu$) if for any $A \in \mathcal{F}$ with $\mu(F) = 0$ we also have $\lambda(A) = 0$.

EXAMPLES 1.4. Consider $(\mathbb{R}, \text{Borel sets})$.

(1) If $\lambda_1(A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$ and $\mu_1(A) = \int_{A \cap [0,1]} dx$, then $\mu_1 \ll \lambda_1$ but not the other way.
(2) If $\lambda_2 = \sum_{i=1}^{\infty} 2^{-i} \delta_{1/i}$ and $\mu_2 = \delta_{1/3}$, then $\mu_2 \ll \lambda_2$ but not the other way.

Now consider a general measure space $(\Omega, \mathcal{F}, \mu)$. If f is a non-negative measurable function, we can define a measure by setting $\lambda(A) = \int_A f\, d\mu$. Then $\lambda \ll \mu$ automatically. The next theorem says that this is really the only example.

THEOREM 1.5 (Radon-Nikodym). *If $\lambda \ll \mu$, then there exists a non-negative $\mathcal{F}$-measurable function f (with $\int_\Omega f\, d\mu < \infty$) so that $\lambda(A) = \int_A f\, d\mu$ for any $A \in \mathcal{F}$. We call f the* Radon-Nikodym derivative *of λ w.r.t. μ and write $\frac{d\lambda}{d\mu} = f$.*

EXAMPLE 1.6. Let λ_i, μ_i be as in the previous example. Then

$$\frac{d\mu_1}{d\lambda_1}(x) = \begin{cases} \sqrt{2\pi} e^{x^2/2} & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\frac{d\mu_2}{d\mu_2}(x) = \begin{cases} 4 & x = \frac{1}{3} \\ 0 & x = \frac{1}{i},\ i \neq 3 \\ \text{anything} & \text{otherwise} \end{cases}.$$

PROPOSITION 1.7. *Let $(\Omega, \mathcal{F}, P)$ be a measure space and let Σ be a sub-σ-field. If X is $\mathcal{F}$-measurable with $E[|X|] < \infty$, then $E\left[X \,\middle|\, \Sigma\right]$ exists.*

PROOF. Write $X = X^+ - X^-$ where $X^+ = X \cdot 1_{\{X \geq 0\}}$ and $X^- = -X \cdot 1_{\{X \leq 0\}}$. Note X^+ and X^- are both integrable on $(\Omega, \mathcal{F}, P)$. Define measures λ^+ and λ^- on $(\Omega, \mathcal{F})$ by $\lambda^{\pm}(A) = \int_A X^{\pm}\, dP$ (so that $\frac{d\lambda^{\pm}}{dP} = X^{\pm}$). Now let $\tilde{P}, \widetilde{\lambda^{\pm}}$ be the restrictions of $P, \lambda^{\pm}$ to the smaller σ-field Σ. These are (finite) non-negative measures on (Ω, Σ). Since $\lambda^{\pm} \ll P$, $\widetilde{\lambda^{\pm}} \ll \tilde{P}$. By Radon-Nikodym, $\frac{d\widetilde{\lambda^{\pm}}}{d\tilde{P}}$ exists. So take

$$E\left[X \,\middle|\, \Sigma\right] = \frac{d\widetilde{\lambda^+}}{d\tilde{P}} - \frac{d\widetilde{\lambda^-}}{d\tilde{P}}.$$

The desired properties follow from Radon-Nikodym. □

THEOREM 1.8. *Conditional expectation satisfies the following almost surely:*

1. $E\left[E\left[X \,\middle|\, \Sigma\right]\right] = E[X]$, $E\left[1 \,\middle|\, \Sigma\right] = 1$
2. $X \geq 0 \implies E\left[X \,\middle|\, \Sigma\right] \geq 0$
3. $E\left[a_1X_1 + a_2X_2 \,\middle|\, \Sigma\right] = a_1E\left[X_1 \,\middle|\, \Sigma\right] + a_2E\left[X_2 \,\middle|\, \Sigma\right]$
4. $E\left[\left|E\left[X \,\middle|\, \Sigma\right]\right|\right] \leq E[|X|]$ *(conditional triangle inequality)*
5. *If Y is Σ-measurable (and bounded), then* $E\left[XY \,\middle|\, \Sigma\right] = YE\left[X \,\middle|\, \Sigma\right]$.
6. *If $\Sigma_2 \subset \Sigma_1 \subset \mathcal{F}$, then* $E\left[X \,\middle|\, \Sigma_2\right] = E\left[E\left[X \,\middle|\, \Sigma_1\right] \,\middle|\, \Sigma_2\right]$.
7. *If ϕ is a convex real-valued function on $\mathbb{R}$, then* $E\left[\phi(X) \,\middle|\, \Sigma\right] \geq \phi\left(E\left[X \,\middle|\, \Sigma\right]\right)$. *(conditional Jensen's inequality)*

REMARK. Taking the expectation of Jensen's inequality in (7) yields $E[\phi(X)] \geq E\left[\phi\left(E\left[X \,\middle|\, \Sigma\right]\right)\right]$.

PROOF OF (7). Observe (2) and (3) give that if $X \geq Y$, then $E\left[X \,\middle|\, \Sigma\right] \geq E\left[Y \,\middle|\, \Sigma\right]$ almost surely. So

$$E\left[\sup\{Y_i\} \,\middle|\, \Sigma\right] \geq E\left[Y_i \,\middle|\, \Sigma\right]$$

for all i, and so

$$E\left[\sup\{Y_i\} \,\middle|\, \Sigma\right] \geq \sup_i \left\{E\left[Y_i \,\middle|\, \Sigma\right]\right\}.$$

Recall any convex ϕ can be written as $\phi(x) = \sup_i \{a_ix + b_i\}$ for countably many a_i, b_i's. Hence

$$\begin{aligned} E\left[\phi(X) \,\middle|\, \Sigma\right] &= E\left[\sup_i \{a_iX + b_i\} \,\middle|\, \Sigma\right] \\ &\geq \sup_i \left\{E\left[a_iX + b_i \,\middle|\, \Sigma\right]\right\} \\ &= \sup_i \left\{a_iE\left[X \,\middle|\, \Sigma\right] + b_i\right\} \\ &= \phi\left(E\left[X \,\middle|\, \Sigma\right]\right) \end{aligned}$$

almost surely. □

Lecture 12
11/29/11

In our setup, $P\left(B \,\middle|\, \Sigma\right) = E\left[1_B \,\middle|\, \Sigma\right]$ is a Σ-measurable random variable and so depends on $\omega \in \Omega$ by default. We can indicate this by writing $P\left(B \,\middle|\, \omega\right)$. But we can also think of this as depending on B. The object $P\left(B \,\middle|\, \omega\right)$ has some nice properties, e.g. for countably many disjoint B_j's we have $P\left(\cup_j B_j \,\middle|\, \omega\right) = \sum_j P\left(B_j \,\middle|\, \omega\right)$ almost surely. (Follows from linearity of conditional expectation.) This hints that for fixed $\omega \in \Omega$, we might view this as a probability measure. But there is a technical problem. Since $P\left(B \,\middle|\, \omega\right)$ is only defined up to a null set, and since we may have uncountably many B's to consider (in the non-discrete setting), their corresponding null sets could "add up" to a non-null set. (The uncountable union of null sets can be non-null.) So it may be difficult to construct an object $P\left(B \,\middle|\, \omega\right)$ which, for a.e. ω, is a probability measure on the B's. Such a (nice) $P\left(B \,\middle|\, \omega\right)$ is called a *regular conditional probability*. These do not always exist, but have been found in many cases. The similar object $P\left(Y \in A \,\middle|\, \omega\right)$ is called a *regular conditional distribution* of Y given Σ.

2. Markov Chains and Random Walks

Let $(\mathcal{X}, \mathcal{F})$ be a measurable space. The *state space* $\mathcal{X}$ could be, for example, $\mathbb{R}$ or $\mathbb{R}^d$, or $\{0, 1, 2, \dots\}$, or $\mathbb{Z}^d$. These are all typical examples from the subject of Markov chains. To study (or construct) independent $\mathcal{X}$-valued random variables $X_0, X_1, X_2, \dots$, we can consider the infinite product space $(\mathcal{X}^\infty, \mathcal{F}^\infty, \Pi_{i=1}^\infty \nu_i) = (\mathcal{X}, \mathcal{F}, \nu_1) \times (\mathcal{X}, \mathcal{F}, \nu_2) \times \cdots$ where $\nu_0, \nu_1, \dots$ are any probability measures on $(\mathcal{X}, \mathcal{F})$. But this is really boring. Often we want to study $X_0, X_1, \dots$ which have a more interesting dependence structure. Here, the subscript k may represent (discrete) time. For example, one might want to construct a measure μ on $(\mathcal{X}^\infty, \mathcal{F}^\infty)$ so that

$$\mu\left(A_0 \times A_1 \times A_2\right) = \left[\int_{A_0} \pi_0\left(dx_0\right) \left[\int_{A_1} \pi_1\left(x_0; dx_1\right) \left[\int_{A_2} \pi_2\left(x_0, x_1; dx_2\right)\right]\right]\right]$$

where π_k is a regular conditional distribution of X_k given $\sigma\left(X_0, \dots, X_{k-1}\right)$. Note π_0 is usually denoted as μ_0.

Definitions 2.1. A sequence $X_0, X_1, \dots$ is called a *Markov process* if for each k, $\pi_k\left(x_0, \dots, x_{k-1}; \cdot\right)$ depends only on x_{k-1}, i.e. it equals some $\pi_{k-1,k}\left(x_{k-1}; \cdot\right)$. The $\pi_{k-1,k}\left(x_{k-1}; A\right)$'s are called *transition probabilities*. If $\pi_{k-1,k}\left(x; A\right)$ does not depend on k, i.e. it equals $\pi\left(x, A\right)$ for all k, then the process is called *time-homogeneous* and is said to have *stationary transition probabilities*.

Thus Markov processes have very short memory, in that the future depends on on the present and not on the past. In a time-homogeneous Markov process, the probability measure μ on $(\mathcal{X}^\infty, \mathcal{F}^\infty)$ is completely determined by $\pi\left(x, A\right)$ and the *initial distribution* μ_0 (for X_0). This can be proved using the Kolmogorov extension theorem.

In a general Markov process, the *l-step transition probability* from X_k to X_{k+l}, denoted $\pi_{k,k+l}\left(x, A\right)$, is defined by

$$\begin{aligned} \pi_{k,k+l}\left(x_k, A\right) &= P\left(X_{k+l} \in A \,\middle|\, \sigma\left(X_k\right)\right) \\ &= P\left(X_{k+l} \in A \,\middle|\, \sigma\left(X_0, \dots, X_k\right)\right). \end{aligned}$$

The second equality is essentially the defining property of a Markov process. The l-step transition probability is given in terms of the 1-step transition probabilities

by

$$\begin{aligned}&\pi_{k,k+l}(x_k, A)\\ &= \int_{x_{k+1}\in\mathcal{X}} \cdots \int_{x_{k+l-1}\in\mathcal{X}} \int_{x_{k+l}\in A} \pi_{k,k+1}(x_k; dx_{k+1})\, \pi_{k+1,k+2}(x_{k+1}; dx_{k+2})\\ &\quad \cdots \pi_{k+l-1,k+l}(x_{k+l-1}, dx_{k+l}).\end{aligned}$$

This easily yields

THEOREM 2.2 (Chapman-Kolmogorov Equations). *For $k < m < n$,*

$$\pi_{k,n}(x, dw) = \int_{y\in\mathcal{X}} \pi_{k,m}(x, dy)\, \pi_{m,n}(y, dw).$$

In a time-homogeneous process, we have

$$\begin{aligned}\pi_{k,k+l}(x, dw) &= \text{some } \pi^{(l)}(x, dw)\\ &= \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \pi(x, dy_1) \cdots \pi(y_{l-1}, dw)\end{aligned}$$

and the Chapman-Kolmogorov equations become

$$\pi^{(l_1+l_2)}(x, dw) = \int_{y\in\mathcal{X}} \pi^{(l_1)}(x, dy)\, \pi^{(l_2)}(y, dw).$$

Here are some important examples of time-homogeneous Markov processes.

EXAMPLES 2.3.

(1) Finite state space. Say $\mathcal{X} = \{1, \ldots, n\}$, then
$$\pi(i, \{j\}) = P\left(X_{k+1} = j \,\middle|\, X_k = i\right)$$
is a $n \times n$ *transition matrix* M with $M_{ij} = \pi(i, \{j\})$. The l-step transition probabilities are given by $\pi^{(l)}(i, \{j\}) = \left(M^l\right)_{ij}$, and the Chapman-Kolmogorov equations are
$$\left(M^{k+l}\right)_{ij} = \left[M^k \cdot M^l\right]_{ij} = \sum_{p=1}^{n} \left(M^k\right)_{ip} \left(M^l\right)_{pj}.$$

(2) Random walk on $\mathbb{Z}^d$. Now $\mathcal{X} = \mathbb{Z}^d$, and one takes $\xi_1, \xi_2, \ldots$ to be i.i.d. $\mathbb{Z}^d$-valued random variables. In particular, the *simple symmetric random walk* has $\xi_1 = (1, 0, \ldots, 0)$, or $(-1, 0, \ldots, 0)$, or $(0, 1, \ldots, 0)$, or $(0, -1, \ldots, 0)$, and so on, each with probability $1/(2d)$. Given X_0 with distribution μ_0, e.g. $\mu_0 = \delta_{(0,\ldots,0)}$, we set $X_n = X_0 + \sum_{i=1}^n \xi_i$. Here,
$$\pi(x, \{y\}) = \begin{cases} \frac{1}{2d} & \text{if } y \text{ is a nearest neighbor of } x \\ 0 & \text{otherwise} \end{cases}.$$

(3) Random walk on $\mathbb{R}$. $X_n = X_0 + \sum_{i=1}^n W_i$ where W_i's are i.i.d. real-valued r.v.s with common distribution μ on $\mathbb{R}$, independent of X_0. Here
$$\begin{aligned}\pi(x, A) &= \mu(A - x)\\ &= \mu(\{z \in \mathbb{R} : z = y - x \text{ for some } y \in A\}).\end{aligned}$$
A special case is with $\{W_i\}$ being i.i.d., standard normal r.v.s. Then
$$\pi(x, A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-(y-x)^2/2}\, dy.$$

This is closed related to Brownian motion.

3. Transience and Recurrence

Lecture 13
12/6/11

Consider a time-homogenous Markov chain $\{X_i\}$ on $\mathcal{X}$ with transition probability $\pi(\cdot,\cdot)$. If X_n has distribution $\mu_n(dx)$, then X_{n+1} will have distribution $\mu_{n+1}(dx) = \int_{x\in\mathcal{X}} \mu_n(dx)\,\pi(x,dy)$.

DEFINITION 3.1. A probability measure μ on $\mathcal{X}$ is called an *invariant measure* if

$$\int_{x\in\mathcal{X}} \mu(dx)\,\pi(x,dy) = \mu(dy).$$

So if $\mu_0 = \mu$ is invariant, then one has $\mu_1 = \mu_0$, $\mu_2 = \mu_0$, etc. If μ_0 is not invariant, then in general μ_n will depend on n. However, if μ_n has a limit $\overline{\mu}$ which is a probability measure, then that limit must be an invariant distribution. See sections 4.6, 4.7 in the text for more information. Some Markov chains have a unique invariant distribution, some have many, and some have none. This is related to the notions of transience and recurrence, which we study now.

DEFINITION 3.2. A r.v. $\tau : \mathcal{X}^\infty \to \{0,1,2,\ldots,\infty\}$ such that $\{\omega : \tau(\omega) = n\}$ is measurable w.r.t. $\sigma(X_0,\ldots,X_n)$ for all n is said to be a *stopping time*.

EXAMPLE 3.3. If $\mathcal{X}$ is countable (e.g. $\mathbb{R}^d$) and $y \in \mathcal{X}$, let τ_y be the time n of the first visit ($n > 0$) to y. If y is never visited, set $\tau_y = \infty$. The r.v. τ_y is an example of a stopping time.

From here on we assume $\mathcal{X}$ is countable.

DEFINITIONS 3.4. A (time homogeneous) Markov chain $\{X_i\}$ is called *irreducible* if $P_x(\tau_y < \infty) > 0$ for all $x, y \in \mathcal{X}$. (Here, P_x is the probability distribution for $\{X_i\}$ when $X_0 = x$.) A state x is called *transient* if $P_x(\tau_x < \infty) < 1$ (or $P_x(T_x = \infty) > 0$); a state is called *recurrent* if it is not transient. A state x is called *positive recurrent* if $E_x[\tau_x] < \infty$, and is called *null recurrent* if it is recurrent but not positive recurrent.

LEMMA 3.5. *In an irreducible Markov chain, all states are of the same type.*

The proof uses the following "renewal property" of Markov chains: let $\mathcal{F}_n = \sigma(X_0,\ldots,X_n)$, then

$$P_x\left\{X_{\tau_x+1} = x_1,\ldots,X_{\tau_x+n} = x_n \,\middle|\, \mathcal{F}_{\tau_x}\right\} = P_x\{X_1 = x_1,\ldots,X_n = x\}.$$

This property is used throughout the study of Markov chains.

PROPOSITION 3.6. *Let $\{X_i\}$ be a Markov chain and let N be the number of visits to y (possibly including a visit at $n = 0$). Then,*

$$G(x,y) = \sum_{n=0}^{\infty} \pi^{(n)}(x,y) = E_x[N].$$

PROOF. We have $N = \sum_{n=0}^{\infty} 1_{\{X_n=y\}}$ and $E_x\left[1_{\{X_n=y\}}\right] = P_x(X_n = y) = \pi^{(n)}(x,y)$. Now just sum up the expectations. □

With this formula in hand we can prove the following theorem.

THEOREM 3.7. *Let $f(x,y) = P_x(\tau_y < \infty)$. An irreducible Markov chain is transient iff $G(x,y) < \infty$ for all x, y (equivalently, for some x, y). Also, we have the relations*

$$G(x,y) = f(x,y)\, G(y,y),$$
$$G(x,x) = \frac{1}{1 - f(x,x)}.$$

PROOF. By the renewal property, after each return to x the probablity of never returning again is $1 - f(x,x)$. Thus

$$P_x\{\text{exactly } n \text{ returns to } x\} = (f(x,x))^n (1 - f(x,x)).$$

Summing up gives the formula for $G(x,x)$, from which the rest follows. □

EXAMPLE 3.8. Simple symmetric random walk on $\mathbb{R}^d$. We have recurrence/transience if $\sum_m \pi^{(m)}(0,0)$ is infinite/finite. Since $\pi^{(m)}(0,0) = 0$ if m is odd we only need to estimate $\pi^{(2n)}(0,0)$. For $d = 1, 2$ this can be done directly, by an application of Stirling's formula. For $d \geq 3$, one can use Fourier series to attack the problem indirectly. The conclusion for general d is that $\pi^{(2n)}(0,0) \sim Cn^{-d/2}$ as $n \to \infty$. So we have recurrence in $d = 1, 2$ and transience otherwise.

Now we discuss periodicity.

DEFINITION 3.9. An irreducible Markov chain is said to be *aperiodic* if, for all $x \in \mathcal{X}$, $\pi^{(n)}(x,x) > 0$ for large enough n. An irreducible Markov chain is said to be *periodic with period* $d > 1$ if, for all $x \in \mathcal{X}$, $\pi^{(n)}(x,x) = 0$ for n not divisible by d and $\pi^{(n)}(x,x) > 0$ for n divisible by d and large.

LEMMA 3.10. *Every irreducible Markov chain is either periodic with period $d > 1$ or aperiodic.*

EXAMPLE 3.11. The simple symmetric random walk on $\mathbb{R}^d$ is periodic with period $d = 2$.

Recall a recurrent chain is postiive recurrent if $E_x[\tau_x] < \infty$ or else null recurrent. Also recall if $\pi^{(n)}(x,y) \to q(y)$, a probability density, as $n \to \infty$ then the limit is the limit of the distributions of X_n with $X_0 = x$ and q is an invariant distribution. Is there a relationship between positive recurence and the existence of an invariant distribution?

THEOREM 3.12. *For recurrent aperiodic chains, null recurrence implies that $\pi^{(n)}(x,y) \to 0$ as $n \to \infty$ for all x, y. Posivite recurrence implies that $\pi^{(n)}(x,y) \to (E_y(\tau_y))^{-1} = q(y)$ as $n \to \infty$, and q is an invariant distribution. In particular, $\sum_{y \in \mathcal{X}} q(y) = 1$.*

PROOF SKETCH. The proof is elementary but long. Here are the salient points. For large n, $\pi^{(n)}(x,y)$ is (approximately) the probaiblity that the chain is at y after many steps. Since $E_x[\tau_x]$ is the mean return time to x, $\pi^{(n)}(x,y)$ is also (approximately) the asymptotic fraction of time spent at y, which should be $1/$(mean return time). □

There are analogous theorems in the periodic case.