CUNY Graduate Center

Math 83100 – Probability

Lecture Notes

Fall 2013

Christian Beneš

cbenes@brooklyn.cuny.edu
http://userhome.brooklyn.cuny.edu/cbenes/index.html

Math 83100 (Fall 2013) September 9, 2013
Prof. Christian Beneš

Lecture #1: Introduction; Probability Spaces; Random Variables; Distributions

Reference. Sections 1.1, 1.2, 1.3

1.1 Why probability?

You probably don't need to be convinced that probability is present everywhere in your daily life (the weather, results of sports competitions, the lottery are just some common examples). Probability also plays a major role in many of the other fields of knowledge:

- Biology: Random mutations are one of the causes of evolution and much deep research (including by the author of our textbook) in probability is done with this application in mind.
- Chemistry: Polymers can be modeled by certain types of random walks (see Chapter 3).
- Finance and economics: Many models exist for the behavior of stocks, but none is perfect, so many probabilists are still working on this problem.
- Physics: The last decade has seen many breakthroughs by probabilists in the field of statistical mechanics.
- Computer science: Markov chains are widely used (for instance in Monte Carlo algorithms)
- etc.

1.2 Notation

Ω will always denote the set of outcomes of an "experiment" and ω will denote an element of Ω. An outcome A ($A \subset \Omega$) of the experiment will occur with a certain probability $P(A)$, where P is a function on Ω with $0 \leq P(\cdot) \leq 1$.

The letters X, Y, Z will usually denote random variables (functions $X : \Omega \to E$)

$E[X]$ will denote the expectation (mean) of a random variable X.

Note 1.1. The definitions above require some care. There will be some restrictions on P which will impose restrictions on the sets A for which $P(A)$ is defined. Not every function $X : \Omega \to E$ will be a random variable nor will $E[X]$ be defined for every random variable.

1.3 Some big questions in probability

One of the big basic questions in probability is the following:

If $\{X_i\}_{i\geq 1}$ are independent, identically distributed random variables (i.i.d. r.v.'s) with same mean $\mu = E[X_i]$, and for $n \geq 1, S_n := \sum_{i-1}^{n} X_i$, then what can be said about S_n?(One can ask this question about the random variable S_n, or about the sequence of random variables $\{S_n\}_{n\geq 1}$, an example of a *stochastic process*.)

- Law of large numbers:
$$(1/n)S_n \to \mu$$
(What does it mean for a sequence r.v.'s like $Y_n = (1/n)S_n$ to converge to a constant μ or to a random variable Y?)

- Central Limit Theorem: If $\{X_i\}_{i\geq 1}$ are i.i.d. with mean μ and finite variance $\sigma^2 = E[(X_i - \mu)^2]$, then
$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \to Z,$$
where Z is the standard normal distribution. (What does it mean for a sequence of r.v.'s like Z_n to converge to a random variable Y?)

- Law of the iterated logarithm: If $\{X_i\}_{i\geq 1}$ are i.i.d. with mean μ and finite variance $\sigma^2 = E[(X_i - \mu)^2]$, then
$$\limsup_n (S_n - n\mu)/\sqrt{2\sigma^2 n \log\log n} = 1$$
and
$$\liminf_n (S_n - n\mu)/\sqrt{2\sigma^2 n \log\log n} = -1$$
(What is the "lim sup" of a sequence of random variables?)

1.4 Motivation for measure-theoretic approach

We would like a probability P to be a measure on a set $\mathcal{F}$ of subsets of Ω: For appropriate sets A, $A_i, i \geq 1$, if we know $P(A)$ and $P(A_i), i \geq 1$, we should also be able to compute:

1. $P(A^c)$
2. $P\left(\cup_{n=1}^{\infty} A_i\right)$.

Therefore, if $A, A_i \in \mathcal{F}$, we will need to automatically have $A^c, \cup_{n=1}^{\infty} A_i \in \mathcal{F}$.

1.5 Basic Definitions

Definition 1.1. A collection of subsets $\mathcal{F}$ of a set Ω is an *algebra* if

1. It is nonempty
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
3. If $A_1, A_2, \ldots, A_n \in \mathcal{F}$, then $\cup_{i=1}^n A_i \in \mathcal{F}$.

In other words, n algebra is a nonempty collection of subsets, closed under complementation and finite unions.

A collection of subsets $\mathcal{F}$ of a set Ω is a $\sigma - algebra$ if

1. It is nonempty
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
3. If $A_1, A_2, \ldots, \in \mathcal{F}$, then $\cup_{i=1}^\infty A_i \in \mathcal{F}$.

In other words, a sigma-algebra is a nonempty collection of subsets, closed under complementation and countable unions.

Note 1.2. Since $(\cap_{i=1}^\infty A_i)^c = \cup_{i=1}^\infty A_i^c$, a sigma-algebra is also closed under countable intersections. Also, since $\mathcal{F}$ is nonempty, $\Omega = A \cup A^c \in \mathcal{F}$ and $\emptyset = A \cap A^c \in \mathcal{F}$.

Example 1.1. $\{\emptyset, \Omega\}$ is a σ-algebra and so is the power set (the set of all subsets) 2^Ω of Ω. These are the smallest and largest σ-algebras of Ω. If $A \subset \Omega$, then $\{\emptyset, A, A^c, \Omega\}$ is a σ-algebra.

Note 1.3. A σ-algebra is clearly an algebra. However, the converse is not true: If $\Omega = \mathbb{R}$ and $\mathcal{A}$ is the collection of sets of the form

$$\cup_{i=1}^k (a_i, b_i], -\infty \leq a_i < b_i \leq \infty, k < \infty,$$

then it is easy to see that $\mathcal{A}$ is an algebra. However, for $i \geq 1, A_i = (0, 1 - \frac{1}{i}] \in \mathcal{A}$, but $\cup_{i=1}^\infty A_i = (0, 1) \notin \mathcal{A}$, so $\mathcal{A}$ is not a sigma-algebra

Definition 1.2. Given a collection $\mathcal{A}$ (not necessarily a σ-algebra) of subsets of Ω, the smallest σ-algebra containing $\mathcal{A}$ is called the *σ-algebra generated by $\mathcal{A}$* and is denoted by $\sigma(\mathcal{A})$. Note that Exercise 1.1.1 in Durrett guarantees the existence of $\sigma(\mathcal{A})$.

Definition 1.3. If Ω is a topological space, then the open subsets of Ω generate a σ-algebra $\mathcal{B}(\Omega)$, called the *Borel σ-algebra* of Ω. We will follow the notation of the textbook and write $\mathcal{B}(\mathbb{R}) = \mathcal{R}$ and $\mathcal{B}(\mathbb{R}^n) = \mathcal{R}^n$.

Definition 1.4. $(\Omega, \mathcal{F})$ is called a *measurable space*. A *measure* is a nonnegative, countably additive set function $\mu : \mathcal{F} \to \mathbb{R}_+ \cup \{\infty\}$ satisfying:

1. $\mu(\emptyset) = 0$

2. If $\{A_i\}_{i\geq 1}$ is a countable collection of disjoint sets, $\mu\left(\cup_{n=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$.

If $\mu(\Omega) = 1$, we write P for μ and call it a *probability measure*.

Note 1.4. It follows from 2. that $P(A^c) = 1 - P(A)$.

Definition 1.5. $(\Omega, \mathcal{F}, P)$ is called a *probability space*.

Example 1.2. (Finite probability space) Suppose $\Omega = \{\omega_1, \ldots, \omega_n\}$. Then we can always define a probability measure on $(\Omega, 2^{\Omega})$ by defining $P(\{\omega_i\}) = p_i$ with $\sum_{i=1}^{n} p_i = 1$. Note that this is a particular case of Example 1.1 in Durrett.

Theorem 1.1. A probability measure P on $(\Omega, \mathcal{F})$ satisfies the following properties: monotonicity, subadditivity, continuity from below and above.

Proof. This is a particular case of Theorem 1.1.1, the proof of which you should make sure you read. □

1.6 Random Variables, Distributions, and densities

Definition 1.6. Given two measurable spaces $(\Omega, \mathcal{F})$ and $(S, \mathcal{S}), X : \Omega \to S$ is a *measurable map* from $(\Omega, \mathcal{F})$ to $(S, \mathcal{S})$ if

$$X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathcal{F} \; \forall B \in \mathcal{S}.$$

In particular, if $X : (\Omega, \mathcal{F}) \to (\mathbb{R}^d, \mathcal{R}^d)$ is measurable, X is called a *random variable* (if $d = 1$) or a *random vector* (if $d > 1$).

Note 1.5. Every random variable X on a probability space $(\Omega, \mathcal{A}, P)$ induces a probability measure on $\mathbb{R}$ which is denoted P^X and called the *law* (or *distribution*) of X. It is defined for every $B \in \mathcal{R}$ by

$$P^X(B) := P\{\omega \in \Omega : X(\omega) \in B\} = P\{X \in B\} = P\{X^{-1}(B)\}.$$

In other words, the random variable X transforms the probability space $(\Omega, \mathcal{A}, P)$ into the probability space $(\mathbb{R}, \mathcal{R}, P^X)$:

$$(\Omega, \mathcal{A}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{R}, P^X).$$

Definition 1.7. A function $F : \mathbb{R} \to \mathbb{R}$ is called a *distribution function* if the following hold:

1. If $x \leq y$, then $F(x) \leq F(y)$.

2. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$

3. $\lim_{y\to x+} F(y) = F(x)$. In other words, F is right-continuous.

Theorem 1.2. If X is a random variable, then $F(x) = F_X(x) := P(X \leq x)$ is a distribution function. It is called the distribution function of X.

Proof. The proof follows almost directly from Theorem 1.1 □

Thinking about the distribution function of a random variable X gives us another (more graphical) way of thinking about the distribution of X.

Theorem 1.3. A distribution function F is the distribution function of some random variable X.

Proof. The key idea is to define X on $((0,1), \mathcal{R}(0,1), \mathcal{L})$ by setting, for $\omega \in (0,1), X(\omega) = \sup\{y : F(y) < \omega\}$. See Theorem 1.2.2 of Durrett. □

Note 1.6. In addition to properties 1-3 above, the distribution function of a random variable X also satisfies the following:

4. $\lim_{y\to x-} F(y) = P(X < x)$.

5. $P(X = x) = F(x) - F(x-)$.

Theorem 1.4. If F is a distribution function, there exists a unique probability measure P on $(\mathbb{R}, \mathcal{R})$ such that

$$P((a,b]) = F(b) - F(a)$$

for all $-\infty \leq a < b < \infty$.

Before proving this, we will need a definition and a big theorem:

Definition 1.8. A set function μ defined on an algebra $\mathcal{A}$ of subsets of Ω is

- *finitely additive* if for any finite collection of pairwise disjoint sets $\{A_i\}_{1\leq i\leq k}$,

$$\mu(\cup_{i=1}^k A_i) = \sum_{i=1}^k \mu(A_i).$$

- *countably additive* (or σ-additive) if for any countable collection of pairwise disjoint sets $\{A_i\}_{1\leq i\leq k}$ such that $\cup_{i=1}^\infty A_i \in \mathcal{A}$,

$$\mu(\cup_{i=1}^\infty A_i) = \sum_{i=1}^\infty \mu(A_i).$$

The next theorem (Theorem A.1.3 in Durrett) is important, but technical, so in order to move forward, I'll let you work through it on your own.

Theorem 1.5. (Caratheodory's Extension Theorem) Given an algebra $\mathcal{A}$ of subsets of Ω, consider $\mathcal{B} = \sigma(\mathcal{A})$, the smallest σ-algebra containing $\mathcal{A}$ and μ_0, a countably additive measure on $(\Omega, \mathcal{A})$. Then there exists a unique measure μ on $(\Omega, \sigma(\mathcal{A}))$ such that $\mu(A) = \mu_0(A)$ for $A \in \mathcal{A}$.

Proof of Theorem 1.4: Note that this is a somewhat different proof from Durrett's proof.

Define $\mathcal{A}$ to be the algebra formed by subsets of $\mathbb{R}$ of the form

$$A = \cup_{k=1}^{n}(a_k, b_k], \qquad -\infty \le a_k < b_k < \infty, n < \infty,$$

and on $\mathcal{A}$, define the (clearly finitely additive) set function P_0 by

$$P_0(A) = \sum_{k=1}^{n} F(b_k) - F(a_k).$$

We will first show that for $A_n \in \mathcal{A}$, if $A_n \downarrow \emptyset$, (i.e, for all $n \ge 1, A_{n+1} \subset A_n$ and $\cap_{k\ge1} A_n = \emptyset$), implies that $P_0(A_n) \to 0$, then P_0 is countably additive.

Indeed, suppose $A_1, A_2, \ldots$ are pairwise disjoint and $\cup_{k\ge1} A_k \in \mathcal{A}$. Then $B_n = \cup_{k\ge n+1} A_k \downarrow \emptyset$ as $n \to \infty$ and so by our assumption, $P(B_n) \to 0$ as $n \to \infty$. Therefore,

$$\begin{aligned}\sum_{i\ge1} P_0(A_i) &= \lim_{n\to\infty} \sum_{i=1}^{n} P_0(A_i) = \lim_{n\to\infty} P_0(\cup_{i=1}^{n} A_i) \\ &= \lim_{n\to\infty} P_0(\cup_{i=1}^{\infty} A_i) - P_0(\cup_{i=n+1}^{\infty} A_i) = \lim_{n\to\infty} P_0(\cup_{i=1}^{\infty} A_i) - P_0(B_n) = P_0(\cup_{i=1}^{\infty} A_i).\end{aligned}$$

So P_0 is countably additive.

Now let $A_1, A_2, \ldots \in \mathcal{A}$ satisfy $A_n \downarrow \emptyset$. We just need to show that $P_0(A_n) \to 0$.

We will first assume that there is an $N < \infty$ such that $A_n \in [-N, N]$ for all $n \ge 1$. By right-continuity of $F, P_0(x, b] \to P_0(a, b]$ as $x \uparrow a$, so for every $A_n \in \mathcal{A}$, there is a $B_n \in \mathcal{A}$ such that the closure $\bar{B}_n \subset A_n$ and $P_0(A_n) \le P_0(B_n) + \epsilon 2^{-n}$.

Since the $\bar{B}_n$ are closed, there is $n_0 < \infty$ such that

$$\cap_{n\ge1} A_n = \emptyset \Rightarrow \cap_{n\ge1} \bar{B}_n = \emptyset \Rightarrow \cap_{n=1}^{n_0} \bar{B}_n = \emptyset, .$$

(Why?)

So we get that

$$\begin{aligned}P_0(A_{n_0}) &= P_0\left(A_{n_0} \setminus \cap_{k=1}^{n_0} B_k\right) \le P_0\left(\cup_{k=1}^{n_0}(A_k \setminus B_k)\right) \\ &\le \sum_{k=1}^{n_0} P_0(A_k \setminus B_k) \le \sum_{k=1}^{n_0} \epsilon 2^{-k} \le \epsilon\end{aligned}$$

Therefore, $P_0(A_n) \to 0$.

If there is no N such that $A_n \subset [-N, N]$ for all $n \geq 1$, fix $\epsilon > 0$ and choose N such that $P_0[-N, N] > 1 - \epsilon/2$. Then repeat the procedure above with $A_n \cap [-N, N]$.

□

Note 1.7. Given a probability measure P on $(\mathbb{R}, \mathcal{R})$, we can define a distribution function by

$$F(x) = P((-\infty, x]).$$

We therefore have a one-to-one correspondence between:

- probability measures on $(\mathbb{R}, \mathcal{R})$
- distribution functions
- random variables

Math 83100 (Fall 2013) September 16, 2013
Prof. Christian Beneš

Lecture #2: Distributions

Reference. Sections 1.3 - 1.6

Definition 2.1. If two random variables X and Y induce the same probability measure $P^X = P^Y$ on $(\mathbb{R}, \mathcal{R})$, we say that X and Y are *equal in distribution* and write $X \stackrel{d}{=} Y$.

There are essentially 3 types of probability measures:

1. Discrete measures: There exists a set of real numbers $\{x_k\}_{k\geq 1}$ such that

$$P(x_k) := P(\{x_k\}) > 0 \text{ and } \sum_{k\geq 1} P(x_k) = 1.$$

2. Absolutely continuous measures: These are measures for which the corresponding distribution functions can be written as

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

with some nonnegative function f, called the *density* of the distribution function F.

3. Singular measures: These are measures for which the corresponding distribution functions are continuous, but have all their points of increase on a set of Lebesgue measure 0.

Theorem 2.1. (Lebesgue decomposition) Every distribution F can be written

$$F = c_1 F_1 + c_2 F_2 + c_3 F_3,$$

where $c_1, c_2, c_3 \geq 0, c_1 + c_2 + c_3 = 1$ and F_1 is discrete, F_2 absolutely continuous, and F_3 singular.

Example 2.1.

1. The pointmass at x_0:

$$F(x) = 1 \text{ for } x \geq x_0, F(x) = 0 \text{ for } x < x_0$$

is a discrete distribution.

2. The uniform distribution on the Cantor set (see Durrett, p. 11) is singular.

3. The Uniform distribution:

$$f(x) = \frac{1}{b-a}, a \leq x \leq b, a, b \in \mathbb{R}.$$

4. The Gamma distribution:

$$f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, x \geq 0, \alpha, \beta > 0.$$

   If $\alpha = 1$, this is the Exponential distribution.

5. The Cauchy distribution:

$$f(x) = \frac{\theta}{\pi(x^2 + \theta^2)}, x \in \mathbb{R}, \theta > 0.$$

   This is an example of a *heavy-tailed* distribution (its associated random variable X has no mean (see definition of $E[X]$ below)).

6. The normal distribution:

$$f(x) = (2\pi\sigma^2)^{-1/2}e^{-(x-\mu)^2/2\sigma^2}.$$

   The change of variables $y = \frac{x-\mu}{\sigma}$ transforms any normal density into the *standard normal density*, i.e., the density of a normal with mean 0, variance 1, usually called $\phi(x)$.

One often comes across $\Phi(x) = \int_{-\infty}^{x} \phi(x)\,dx$, the normal distribution function. Unfortunately, there is no closed-form expression for it, so the following estimate is often useful (note that $1 - \Phi(x) = P(X > x)$):

Lemma 2.1. If $x > 0$, then

$$(x + x^{-1})^{-1}\phi(x) \leq 1 - \Phi(x) \leq x^{-1}\phi(x).$$

Proof. Let $x > 0$. Since $\phi'(y) = -y\phi(y)$, we have

$$\phi(x) = \int_x^{\infty} y\phi(y)\,dy \geq x\int_x^{\infty} \phi(y)\,dy = x(1 - \Phi(x)).$$

Also, since $(y^{-1}\phi(y))' = -(1 + y^{-2})\phi(y)$,

$$x^{-1}\phi(x) = \int_x^{\infty} (1 + y^{-2})\phi(y)\,dy \leq (1 + x^{-2})\int_x^{\infty} \phi(y)\,dy = (1 + x^{-2})(1 - \Phi(x)).$$

□

Note 2.1. It's worth comparing this estimate with Durrett's, which is slightly different.

2.1 Random Variables

Recall that a *random variable* is a measurable (or $\mathcal{F}$-measurable) function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{R})$. Measurability is key in allowing us to integrate a random variable and this requirement means that not just any random function is a random variable.

Example 2.2. If $A \in \Omega$ is measurable, then

$$\mathbb{1}_A(\omega) := \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}$$

is a r.v.

Given 2 measurable spaces $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$, it isn't always straightforward to check if a function

$$X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$$

is measurable, i.e., to check if

$$X^{-1}(B) \in \mathcal{F} \text{ for every } B \in \mathcal{S}.$$

The following lemma and its corollary somewhat simplify this task.

Lemma 2.2. Let $\mathcal{A}$ be a family of sets such that $\sigma(\mathcal{A}) = \mathcal{S}$. Then $X : \Omega \to S$ is measurable iff

$$X^{-1}(B) \in \mathcal{F}$$

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

Proof. Necessity is obvious: If $B \in \mathcal{A}$, then $B \in \mathcal{S}$, so $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{S}$ implies $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{A}$.

To show sufficiency, define $\mathcal{B} = \{B \in \mathcal{S} : X^{-1}(B) \in \mathcal{F}\}$. (We will show that $\mathcal{B} = \mathcal{S}$.)

Then if $\{B_i\}_{i \geq 1} \in \mathcal{B}$ (implying that for any $i, X^{-1}(B_i) \in \mathcal{F}$), we can use the fact that

$$X^{-1}\left(\cup_{i \geq 1} B_i\right) = \cup_{i \geq 1} X^{-1}(B_i) \qquad \text{and} \qquad X^{-1}(B_1^c) = (X^{-1}(B_1))^c, \tag{1}$$

to see that

- $X^{-1}\left(\cup_{i \geq 1} B_i\right) = \cup_{i \geq 1} X^{-1}(B_i) \in \mathcal{F} \Rightarrow \cup_{i \geq 1} B_i \in \mathcal{B}$
- $X^{-1}(B_1^c) = X^{-1}(B_1)^c \in \mathcal{F} \Rightarrow B_1^c \in \mathcal{B}$,

so $\mathcal{B}$ is a σ-algebra.

Therefore,

$$\mathcal{A} \subset \mathcal{B} \subset \mathcal{S},$$

implying

$$\mathcal{S} = \sigma(\mathcal{A}) \subset \sigma(\mathcal{B}) = \mathcal{B} \subset \sigma(\mathcal{S}) = \mathcal{S},$$

so that $\mathcal{S} = \mathcal{B}$. Since we assumed $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}$, we now have $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{S}$, so X is measurable. □

Corollary 2.1. $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{R})$ is a random variable iff $\{\omega : X(\omega) \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$.

Proof. The sets $\{(-\infty, x] : x \in \mathbb{R}\}$ generate the Borel sets. □

Note 2.2. In Corollary 2.1, on can replace $\leq x$ by $< x$, $> x$, $\geq x$.

Lemma 2.3. If $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$, then

$$\mathcal{A} = \{X^{-1}(B) : B \in \mathcal{S}\}$$

is a σ-algebra. It is called the *σ-algebra generated by X*.

Proof. Suppose $\{A_i\}_{i\geq 1} \in \mathcal{A}$. Then there exist $\{B_i\}_{i\geq 1} \in \mathcal{S}$ with $A_i = X^{-1}(B_i)$.
By (1) above, the fact that $\cup_{i\geq 1} B_i \in \mathcal{S}$ implies

$$\cup_{i=1}^{n} A_i = \cup_{i\geq 1} X^{-1}(B_i) = X^{-1}\left(\cup_{i\geq 1} B_i\right) \in \mathcal{A}$$

and the fact that $B_1^c \in \mathcal{S}$ implies that

$$A_1^c = (X^{-1}(B_1))^c = X^{-1}(B_1^c) \in \mathcal{A}.$$

□

The next lemma is the general case of the fact that a Borel function (i.e., a Borel-measurable function) of a random variable is a random variable:

Lemma 2.4. Let $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ and $f : (S, \mathcal{S}) \to (T, \mathcal{T})$ be measurable. Then $Y = f \circ X : (\Omega, \mathcal{F}) \to (T, \mathcal{T})$ is measurable.

Proof. If $Y(\omega) = f(X(\omega))$, then for $B \in \mathcal{T}$, we have

$$\{\omega : Y(\omega) \in B\} = \{\omega : X(\omega) \in f^{-1}(B)\} \in \mathcal{F},$$

since $f^{-1}(B) \in \mathcal{S}$. □

We will now use the lemmas above to construct new random variables from collections of random variables.

Theorem 2.2. If $X_1, \ldots, X_n$ are random variables and $f : (\mathbb{R}^n, \mathcal{R}^n) \to (\mathbb{R}, \mathcal{R})$ is measurable, then $f(X_1, \ldots, X_n)$ is a random variable.

Proof. By Lemma 2.4, all we need to do is show that $(X_1, \ldots, X_n) : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{R}^n)$ is measurable, that is, that $(X_1, \ldots, X_n)$ is a random vector. Since n-dimensional Borels are generated by n-fold products of one-dimensional Borels, Lemma 2.2 implies that we just need to check that $(X_1, \ldots, X_n)^{-1}(A_1, \ldots, A_n) \in \mathcal{F}$ if $A_1, \ldots, A_n \in \mathcal{R}$:

$$\{(X_1, \ldots, X_n) \in (A_1, \ldots, A_n)\} = \cap_{i=1}^{n} \{X_i \in A_i\} \in \mathcal{F}.$$

□

Definition 2.2. The random variable $\limsup_n X_n$ is defined for every $\omega \in \Omega$ by

$$\limsup_n X_n(\omega) = \inf_n \sup_{m \geq n} X_m(\omega).$$

Similarly,

$$\liminf_n X_n = \sup_n \inf_{m \geq n} X_m$$

Corollary 2.2. Suppose $X_1, X_2, \ldots$ are random variables. Then the following are random variables as well:

1. $f(X_1)$, where $f : \mathbb{R} \to \mathbb{R}$ is continuous.
2. $\sum_{i=1}^{n} X_i$
3. $\sup_n X_n, \inf_n X_n$
4. $\limsup_n X_n, \liminf_n X_n$

Proof. 1. This follows from Lemma 2.4, since continuous functions are measurable.

2. By Theorem 2.2, we just need to check that $f(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$ is measurable and by Lemma 2.2 it's enough to show that $f^{-1}(-\infty, a) \in \mathcal{R}^n$, since intervals of the form $(-\infty, a)$ generate the Borels. But since $f^{-1}(-\infty, a) = \{(x_1, \ldots, x_n) : \sum_{i=1}^{n} x_i < a$ is an open set, it of course is a Borel set too.

3. This follows directly from

$$\{\omega : \sup X_n(\omega) > x\} = \cup_n \{\omega : X_n(\omega) > x\} \in \mathcal{F}$$

and the fact that

$$\inf X_n = -\sup(-X_n).$$

4. This follows directly from the definitions

$$\limsup_n X_n = \inf_n \sup_{m \geq n} X_m,$$

$$\liminf_n X_n = \sup_n \inf_{m \geq n} X_m$$

and from part 3.

□

Definition 2.3. If for $n \in \mathbb{N}, X_n : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{R})$ is a r.v., then

$$\Omega_0 := \{\omega : \lim_{n\to\infty} X_n(\omega) \text{ exists } \} = \{\omega : \limsup_n X_n(\omega) - \liminf_n X_n(\omega) = 0\}$$

is measurable. If $P(\Omega_0) = 1$, X_n *converges almost surely*. Since $\lim_n X_n$ may not be defined for all ω, we let $X_\infty := \limsup_n X_n$ and write

$$X_n \overset{\text{a.s.}}{\to} X_\infty.$$

2.2 Expectation

Definition 2.4. Let X be a random variable and let $X^+ := \max\{X, 0\}, X^- := -\min\{X, 0\}$. If $E[X^+] = \int X^+ \, dP < \infty$ and $E[X^-] = \int X^- \, dP < \infty$, we say that the expectation of X exists. In that case, the *expectation* is

$$E[X] := E[X^+] - E[X^-].$$

Note 2.3. Being an integral, expectation is linear.

2.2.1 Inequalities

Proposition 2.1. (Markov/Chebyshev's Inequality) Let $X \geq 0$ be a random variable. Then

$$P(X \geq \epsilon) \leq \frac{E[X]}{\epsilon}.$$

Proof.

$$E[X] \geq E[X \cdot \mathbb{1}_{X^{-1}[\epsilon,\infty)}] \geq \epsilon E[\mathbb{1}_{X\geq\epsilon}] = \epsilon \int_{X\geq\epsilon} dP = \epsilon P(X \geq \epsilon).$$

□

Note 2.4. There is no absolute consensus about whether to call the inequality above Markov's or Chebyshev's. I will always refer to it as Markov's inequality.

Although you have certainly seen the following definition before, I'll state it here for self-containment of the notes (or at least of its probabilistic material).

Definition 2.5. If X is a random variable, its *variance* is defined to be (when it exists)

$$Var(X) = E[(X - E[X])^2] (= E[X^2] - E[X]^2).$$

Proposition 2.2. (Chebyshev's inequality) For any random variable X,

$$P(|X - E[X]| \geq \epsilon) \leq \frac{Var(X)}{\epsilon^2}.$$

Proof. By Markov's inequality,

$$P(|X - E[X]| \geq \epsilon) = P((X - E[X])^2 \geq \epsilon^2) \leq \frac{E[(X - E[X])^2]}{\epsilon^2} = \frac{Var(X)}{\epsilon^2}.$$

□

2.3 Jensen's Inequality

Definition 2.6. A function $\phi : \mathbb{R} \to \mathbb{R}$ is called *convex* if for $x, y \in \mathbb{R}, 0 \leq p \leq 1$,

$$\phi(px + (1-p)y) \leq p\phi(x) + (1-p)\phi(y).$$

Theorem 2.3 (Jensen's inequality)**.** Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is convex. Suppose X is a random variable satisfying $E[|X|] < \infty$ and $E[|\phi(x)|] < \infty$. Then

$$\phi(E[X]) \leq E[\phi(X)].$$

Proof. If ϕ is convex, then for every $x_0 \in \mathbb{R}$, there is a $c(x_0)$ such that $\frac{\phi(x)-\phi(x_0)}{x-x_0} \geq c(x_0)$. Choosing $x_0 = E[X]$ and letting $x = X$, we get

$$\phi(X) \geq c(E[X])(X - E[X]) + \phi(E[X]).$$

Taking expectations on both sides concludes the proof. □

Example 2.3.

1. For any random variable,

   $$|E[X]| \leq E[|X|],$$

   provided the latter exists.

2. For any random variable for which $E[X^2] < \infty$,

   $$E[X]^2 \leq E[X^2].$$

   In particular, $Var(X) = E[X^2] - E[X]^2 \geq 0$. This example can be a good reminder of which way the inequality goes in Jensen's inequality (which can be easy to forget).

2.4 Convergence Theorems

Definition 2.7. If $\{X_n\}_{n\geq 1}$ and X are random variables, we say that $\{X_n\}$ *converges to* X *in probability* (and write $X_n \xrightarrow{P} X$) if for every $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) \to 0, \text{ as } n \to \infty.$$

We will see other types of convergence and how they are related later. For now we focus on the following implication:

Theorem 2.4. If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{P} X$.

Proof. Note that since P is continuous from above (See Theorem 1.1.1 in Durrett),

$$\begin{aligned} X_n \xrightarrow{a.s.} X \quad &\iff \quad P\left(\bigcup_{l\geq 1}\bigcap_{N\geq 1}\bigcup_{n\geq N}\left\{|X_n - X| \geq \frac{1}{l}\right\}\right) = 0 \\ &\iff \quad \forall \epsilon > 0, \lim_{N\to\infty} P\left(\bigcup_{n=N}^{\infty}\{|X_n - X| \geq \epsilon\}\right) = 0. \end{aligned}$$

This implies that $\lim_{N\to\infty} P(|X_N - X| \geq \epsilon) = 0$. □

We now turn to a number of results that tell us under what conditions, $X_n \to X$ (in some sense) implies $E[X_n] \to E[X]$.

Theorem 2.5. (Bounded Convergence) Suppose $\{X_n\}_{n\geq 1}$ and X are random variables such that $X_n \xrightarrow{P} X$ and for some $K > 0$, $|X_n| \leq K$ for all n. Then

$$E[|X_n - X|] \overset{n\to\infty}{\to} 0,$$

implying

$$E[X_n] \overset{n\to\infty}{\to} E[X].$$

Proof. First observe that $P(|X| \leq K) = 1$. Indeed, for all $k \in \mathbb{N}$,

$$P(|X| > K + \frac{1}{k}) \leq P(|X - X_n| > \frac{1}{k}\ \forall n) = 0.$$

Therefore,

$$P(|X| > K) = P(\cup_k \{|X| > K + \frac{1}{k}\}) = 0,$$

by subadditivity. Now one of two things can happen: Either $|X_n - X|$ is small, or it is large (remember it can't exceed $2K$), but with probability decaying to 0. Formally, fix $\epsilon > 0$ and choose N so that for all $n \geq N$,

$$P(|X_n - X| > \epsilon/3) < \frac{\epsilon}{3K}.$$

Then, for $n \geq N$,

$$\begin{aligned} E[|X_n - X|] &= E[|X_n - X|\mathbb{1}_{\{|X_n - X| > \epsilon/3\}}] + E[|X_n - X|\mathbb{1}_{\{|X_n - X| \leq \epsilon/3\}}] \\ &\leq 2KP(|X_n - X| > \epsilon/3) + \epsilon/3 \leq \epsilon. \end{aligned}$$

□

We give the following result without a proof.

Theorem 2.6. (Fatou's Lemma) Suppose $\{X_n\}_{n\geq 1}$ are random variables satisfying $X_n \geq 0$ for every n. Then

$$E[\liminf_{n\to\infty} X_n] \leq \liminf_{n\to\infty} E[X_n].$$

Theorem 2.7. (Reverse Fatou's Lemma) If $\{X_n\}_{n\geq 1}, X$ satisfy $X_n \leq X, \forall n$, $E[X] < \infty$, then

$$E[\limsup X_n] \geq \limsup E[X_n].$$

Proof. It follows from Fatou's Lemma that $E[\liminf(X - X_n) \leq \liminf E[X_n - X]$. Therefore,

$$E[X] - E[\liminf(-X_n)] \leq E[X] - \liminf E[-X_n],$$

implying that

$$E[\liminf(-X_n)] \geq \liminf E[-X_n].$$

The theorem now follows from the fact that $\liminf(-A_n) = \limsup A_n$. □

Theorem 2.8. (Monotone Convergence) Suppose $\{X_n\}_{n\geq 1}$ and X are random variables such that $E[X_n] < \infty \; \forall n, E[X] < \infty$, $X_n \overset{a.s}{\to} X$, and for all $n \geq 1, 0 \leq X_n \leq X_{n+1}$. Then $E[X_n] \to E[X]$.

Proof. Since $X_n \leq X$ for every n, $E[X_n] \leq E[X]$ for every n. Moreover, $\{E[X_n]\}_{n\geq 1}$ forms a bounded monotonic sequence, which must converge. Therefore,

$$\lim_{n\to\infty} E[X_n] \leq E[X].$$

Since $X = \lim_{n\to\infty} X_n$, Fatou's Lemma gives

$$E[X] = E[\liminf_{n\to\infty} X_n] \leq \liminf_{n\to\infty} E[X_n] = \lim_{n\to\infty} E[X_n].$$

The 2 inequalities prove the theorem. □

Why is the requirement that we have an increasing or a bounded sequence necessary? The following example shows that in general, the fact that $X_n \overset{a.s}{\to} X$ doesn't necessarily imply that $E[X_n] \to E[X]$. It is also an example showing that we may have a strict inequality in Fatou's Lemma.

Example 2.4. Consider the probability space $([0,1], \mathcal{R}[0,1], \mathcal{L}[0,1])$, where $\mathcal{L}[0,1]$ denotes Lebesgue measure. Then if we define

$$X_n(\omega) = \begin{cases} n, & 0 \leq \omega \leq 1/n \\ 0, & \text{otherwise} \end{cases},$$

$X_n \overset{a.s}{\to} X$, but $1 = E[X_n] \not\to E[X] = 0$.

The Bounded Convergence Theorem above is a particular case of the Dominated Convergence Theorem:

Theorem 2.9. (Dominated Convergence) Suppose $\{X_n\}_{n\geq 1}, X$, and Y are random variables such that $X_n \overset{a.s.}{\to} X, |X_n| \leq Y$ for all n and $E[Y] < \infty$. Then $E[|X_n - X|] \to 0$, implying $E[X_n] \to E[X]$.

Proof. $|X_n - X| \leq 2Y$, so by the Reverse Fatou's Lemma,

$$0 = E[\limsup |X_n - X|] \geq \limsup E[|X_n - X|],$$

so $E[|X_n - X|] \to 0$. The second part follows from Jensen's inequality. □

Lecture #3: L^p spaces

Reference. Section 1.6; see also David Williams' "Probability With Martingales", an excellent book which should be on reserve at the library

3.1 L^p Spaces

Definition 3.1. If $X : \Omega \to \mathbb{R}$ is a random variable, we define for all $p \in [1, \infty)$,

$$\|X\|_p := (E[|X|^p])^{1/p},$$

provided that the expectation exists.

$$\|X\|_\infty = \inf\{M \in \mathbb{R} : P(|X| > M) = 0\}.$$

The space $L^p = L^p(\Omega, \mathcal{F}, P)$ is the space of all random variables $X : \Omega \to \mathbb{R}$ such that $\|X\|_p < \infty$.

For $1 \leq p < \infty$, L^p is a vector space:

Proposition 3.3. For all $X, Y \in L^p$ and $c \in \mathbb{R}$,

1. $\|cX\|_p = |c|\|X\|_p$,
2. $X + Y \in L^p$.

Proof. 1. is trivial by linearity of integrals. For 2., if $1 \leq p < \infty$, note that $|X + Y|^p \leq (2 \max\{|X|, |Y|\})^p \leq 2^p(|X|^p + |Y|^p)$, from which we get

$$\|X + Y\|_p^p = E[|X + Y|^p] \leq 2^p(E[|X|^p] + E[|Y|^p]) < \infty$$

if $X, Y \in L^p$, in which case we also have $\|X + Y\|_p < \infty$. If $p = \infty$, 2. is obvious. □

L^p norms are monotone in the following sense (note that we don't know yet that $\|\cdot\|_p$ is a norm, but we will see soon that this is the case, at least if one considers the right version of L^p:

Proposition 3.4. Suppose $1 \leq p \leq r < \infty$ and $X \in L^r$. Then $X \in L^p$ and

$$\|X\|_p \leq \|X\|_r.$$

Proof. Define the truncated (thus bounded) random variable $X_n := (|X| \wedge n)^p$. Then $|X_n| \leq n^p$, so X_n and $X_n^{r/p}$ are in L^1. Jensen's inequality applied to $\phi(x) = x^{r/p}$ implies that

$$E[X_n]^{r/p} \leq E[(|X| \wedge n)^r] \leq E[|X|^r].$$

Since $X_n \overset{a.s}{\to} |X|^p$, we can apply the Monotone Convergence Theorem to obtain that

$$E[|X|^p]^{r/p} \leq E[|X|^r] \Rightarrow E[|X|^p]^{1/p} \leq E[|X|^r]^{1/r}$$

□

3.2 Two Important Inequalities

Lemma 3.1. If $X \geq 0$ is a random variable satisfying $E[X] = 0$, then $X = 0$, almost surely.

Proof. Define $A = \{\omega : X(\omega) > 0\}$ and $A_n = \{\omega : X(\omega) \geq 1/n\}$. Clearly, $A_n \uparrow A$ and

$$0 \leq X\mathbb{1}_{A_n} \leq X\mathbb{1}_A \leq X,$$

so

$$0 \leq E[X\mathbb{1}_{A_n}] \leq E[X\mathbb{1}_A] \leq E[X] = 0.$$

Therefore (since $X \geq 1/n$ over A_n),

$$\frac{1}{n}P(A_n) \leq E[X\mathbb{1}_{A_n}] = 0,$$

so $P(A_n) = 0$. But since $A_n \uparrow A, P(A_n) \to P(A)$. Therefore, $P(A) = 0$.

□

Theorem 3.1. (Hölder's inequality) Suppose $p \in [1, \infty], 1/q = 1 - 1/p$ and $X \in L^p, Y \in L^q$. Then $XY \in L^1$ and

$$E[|XY|] \leq \|X\|_p \|Y\|_q.$$

Proof. The case $p = 1$ is left as an easy homework exercise. Assume $1 < p < \infty$. Since for $a \geq 1$, $|X|^a$ is a random variable when X is, Lemma 3.1 implies that $|XY| = 0$ a.s. if $\|X\|_p = 0$ or $\|Y\|_q = 0$, so the theorem is true in that case. Suppose now that $\|X\|_p > 0$ and $\|Y\|_q > 0$ and define the normalized r.v.'s

$$\tilde{X} = \frac{X}{\|X\|_p} \qquad \text{and} \qquad \tilde{Y} = \frac{Y}{\|Y\|_q}.$$

Since $\phi(x) = \ln x$ is concave, we have, for $x, y > 0$,

$$\ln(x^{1/p}y^{1/q}) = \frac{1}{p}\ln x + \frac{1}{q}\ln y \leq \ln(x/p + y/q),$$

which, since ϕ is increasing, implies

$$x^{1/p}y^{1/q} \leq x/p + y/q.$$

Therefore, it follows from monotonicity of expectations that

$$E\left[|\tilde{X}\tilde{Y}|\right] = E\left[(|\tilde{X}|^p)^{1/p}(|\tilde{Y}|^q)^{1/q}\right] \leq E\left[\frac{1}{p}|\tilde{X}|^p + \frac{1}{q}|\tilde{Y}|^q\right] = \frac{1}{p} + \frac{1}{q} = 1.$$

□

Note 3.1. If $p = q(= \frac{1}{2})$, Hölder's inequality is known as the Cauchy-Bunyakovski-Schwarz inequality: If $X, Y \in L^2$,

$$E[|XY|]^2 \leq E[X^2]E[Y^2],$$

or equivalently, as we'll see later,

$$\langle X, Y \rangle \leq \|X\|_2 \|Y\|_2,$$

where $\langle \cdot, \cdot \rangle$ denotes inner product.

Theorem 3.2. (Minkowski's inequality) If $p \in [1, \infty]$ and $X, Y \in L^p$, then

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p.$$

Proof. The proof is straightforward if $p = 1$ and $p = \infty$, so suppose $1 < p < \infty$. Note that $\||X + Y|^{p-1}\|_q$ is well-defined, since $(|X + Y|^{p-1})^q = (|X + Y|^{p-1})^{\frac{p}{p-1}} = |X + Y|^p$.

The vector space property of L^p and Hölder's inequality give

$$\begin{aligned} E[|X + Y|^p] &\leq (E[|X||X + Y|^{p-1}] + E[|Y||X + Y|^{p-1}]) \\ &\leq (\|X\|_p + \|Y\|_p)\||X + Y|^{p-1}\|_q = (\|X\|_p + \|Y\|_p)E[|X + Y|^p]^{1/q}, \end{aligned}$$

so $E[|X + Y|^p]^{1-1/q} \leq \|X\|_p + \|Y\|_p$, implying that

$$E[|X + Y|^p]^{1/p} \leq \|X\|_p + \|Y\|_p.$$

□

3.3 More on L^p spaces

Definition 3.2. For $0 < p < \infty$, we say a sequence $\{X_n\} \in L^p$ of random variables converges to $X \in L^p$ in L^p (and write $X_n \xrightarrow{L^p} X$) if

$$E[|X_n - X|^p] \to 0, \text{ as } n \to \infty.$$

Proposition 3.5. If $X_n \xrightarrow{L^p} X$, then $X_n \xrightarrow{P} X$.

Proof. This is just Markov's inequality: For every ϵ,

$$P(|X_n - X| > \epsilon) = P(|X_n - X|^p > \epsilon^p) \leq \frac{E[|X_n - X|^p]}{\epsilon^p} \to 0.$$

□

In all that follows, $p \in [1, \infty)$.

We now know that L^p is a vector space and Minkowski's inequality tells us it is equipped with a semi-norm $\|\cdot\|_p$. If it were a norm under which the space is complete, we'd have a Banach space. So we need to check that $\|X\|_p = 0 \iff X = 0$. Unfortunately, it is easy to

construct nonzero random variables X such that $\|X\|_p = 0$. However, it is true (see Lemma 3.1) that

$$\|X\|_p = 0 \iff X = 0 \text{ a.s.}$$

We can therefore define for every $X \in L^p$ an equivalence class $[X]$ by saying $X \sim Y$ if $X = Y$ a.s. If we define $[L^p]$ to be the collection of equivalence classes $[X]$ for $X \in L^p$, and immediately revert to the previous notation, L^p is a normed vector space.

Now to completeness:

Definition 3.3. A sequence of random variables $\{X_n\} \in L^p$ is a *Cauchy sequence in L^p* if

$$\sup_{s,t \geq n} \|X_s - X_t\|_p \to 0 \text{ as } n \to \infty.$$

Theorem 3.3. For every Cauchy sequence $\{X_n\} \in L^p$, there exists a random variable $X \in L^p$ such that

$$X_n \overset{L^p}{\to} X.$$

Proof. Consider an increasing sequence $\{n_k\}_{k \geq 1}$ in $\mathbb{N}$ such that if $s_k, t_k \geq n_k, \|X_{s_k} - X_{t_k}\|_p < 2^{-k}$. Then by monotonicity of L^p norms,

$$E[|X_{s_k} - X_{t_k}|] = \|X_{s_k} - X_{t_k}\|_1 \leq \|X_{s_k} - X_{t_k}\|_p < 2^{-k},$$

which implies that

$$E[\sum_{k \geq 1} |X_{s_k} - X_{t_k}|] < \infty.$$

Therefore,

$$\sum_{k \geq 1} |X_{s_k} - X_{t_k}|$$

converges a.s., implying that

$$\sum_{k \geq 1} |X_{n_{k+1}} - X_{n_k}|$$

converges a.s. (indeed, if there were a set of non-zero probability on which the sum diverges, the expectation would be infinite), so $\lim_{k \to \infty} X_{n_k}$ exists a.s. If we define $X(\omega) = \limsup_{k \to \infty} X_{n_k}(\omega)$, then X is a random variable and $X_{n_k} \to X$ a.s.

Now for every $k, m, n \in \mathbb{N}$ such that $m \geq k$ and $n \geq n_k$,

$$E[|X_n - X_{n_m}|^p] \leq 2^{-pk}.$$

Letting $m \to \infty$, we get from Fatou's lemma that

$$E[|X_n - X|^p] \leq 2^{-pk} \; \forall k.$$

This shows that $X_n \to X$ in L^p and, since $X_n - X \in L^p$, so is X. □

Corollary 3.1. If $p \geq 1$, L^p is a Banach space.

Proof. We knew that L^p is a normed vector space and just proved that it is complete. □

3.4 L^2 is a Hilbert space

The case $p = 2$ is particularly important, since Hölder's inequality tells us we can construct an inner product on the space as follows:

Proposition 3.6. Suppose $X, Y \in L^2$. Then

$$\langle X, Y \rangle := E[XY]$$

is an inner product on L^2.

Proof. If $X, Y, U \in L^2$, the proposition follows from the fact that for $a, b \in \mathbb{R}$,

- $\langle aX + bY, U \rangle = a\langle X, U \rangle + b\langle Y, U \rangle$
- $\langle X, X \rangle \geq 0$
- $\langle X, X \rangle = 0 \iff X = 0$ (this holds by Lemma 3.1)

□

Definition 3.4. If $X, Y \in L^2$, we define

$$Cov(X, Y) := \langle X - E[X], Y - E[Y] \rangle$$

and

$$Var(X) := Cov(X, X) = \|X - E[X]\|_2^2.$$

The *angle θ between X and Y* is defined by

$$\cos\theta = \frac{\langle X, Y \rangle}{\|X\|_2 \|Y\|_2}.$$

(Note that $|\cos\theta| \leq 1$ by C-S-B.) The *correlation* between X and Y is

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}.$$

If $\langle X, Y \rangle = 0$, we say X and Y are *orthogonal*.

Proposition 3.7. If $Cov(X, Y) = 0$, then $Var(X + Y) = Var(X) + Var(Y)$.

Proof.

$$\begin{aligned} Var(X+Y) &= \langle X + Y - (E[X] + E[Y]), X + Y - (E[X] + E[Y]) \rangle \\ &= \langle X - E[X], X - E[X] \rangle + \langle Y - E[Y], Y - E[Y] \rangle - 2\langle X - E[X], Y - E[Y] \rangle \\ &= \langle X - E[X], X - E[X] \rangle + \langle Y - E[Y], Y - E[Y] \rangle = Var(X) + Var(Y) \end{aligned}$$

□

3.5 Computing Expectations

Theorem 3.4. (Change of variables formula) Let $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ be a measurable function (with distribution P^X defined by the relation $P^X(A) = P(X \in A)$). If $f : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{R})$ is measurable and $f \in L^1(S, \mathcal{S}, P^X)$, then

$$E[f(X)] = \int_S f(y) P^X(dy).$$

Proof. By definition, the result is true for indicator functions. Linearity implies it's true for simple functions. The monotone convergence theorem implies it holds for nonnegative functions and linearity again shows it holds for integrable functions. □

Definition 3.5. $E[X^k]$ is called the k^{th} *moment of* X.

Definition 3.6. The *moment generating function* of a random variable X is

$$M_X(t) = E[e^{Xt}] = \int_{\mathbb{R}} e^{xt} P^X(dx) = \int_{\mathbb{R}} e^{xt} dF(x).$$

Note 3.2. This is defined for every t, since e^{Xt} is a positive random variable. However, it may be infinite.

Regardless of X, $M_X(t)$ is finite at $t = 0$, where its value is 1. However, as the following example shows, this may be the only point at which $M_X(t)$ is finite.

Example 3.1. Suppose $P^X\{n\} = P^X\{-n\} = \frac{C}{n^2}, n \in \mathbb{N}$. Then

$$\int_{\mathbb{R}} e^{xt} P^X(dx) = \sum_{n \geq 1} e^{tn} \frac{C}{n^2} + \sum_{n \geq 1} e^{-tn} \frac{C}{n^2}.$$

If $t > 0$, the first sum is infinite, while if $t < 0$, the second is. So $M_X(t) < \infty$ only if $t = 0$.

Theorem 3.5. Suppose $M_X(t)$, the moment generating function of X, is finite for $t \in (-t_0, t_0), t_0 > 0$. Then

$$E[X^n] = M_X^{(n)}(0).$$

Proof. Suppose $M_X(t)$ is finite on $(-t_0, t_0), t_0 > 0$ and choose $t \in (-t_0, t_0)$. Then $e^{|xt|} \leq e^{xt} + e^{-xt}$, so since by assumption $e^{xt} + e^{-xt}$ is P^X-integrable, $\int_{\mathbb{R}} e^{|xt|} P^X(dx) < \infty$, so $\int_{\mathbb{R}} \sum_{k \geq 0} \frac{|xt|^k}{k!} P^X(dx)$ is defined. The dominated convergence theorem allows us to interchange the integral and sum, so

$$M_X(t) = \sum_{k \geq 0} \frac{t^k}{k!} \int_{\mathbb{R}} x^k P^X(dx) = \sum_{k \geq 1} \frac{t^k}{k!} E[X^k].$$

By uniqueness of Taylor expansions,

$$E[X^k] = M_X^{(k)}(0).$$

□

Example 3.2. Let Z have the standard normal distribution with density $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Then the moment generating function of Z can be found as follows:

$$M_Z(t) = E[e^t Z] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{x^2-2tx}{2}}\, dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{(x-t)^2-t^2}{2}}\, dx = e^{t^2/2}.$$

This allows us to find all even moments of Z (the odd moments are all obviously 0):

$$E[Z^{2n}] = M_Z^{2n}(0) = \prod_{k=1}^{n} (2k-1) = \frac{(2n)!}{2^n n!}.$$

Lecture #4: Independence; Laws of Large Numbers

Reference. Section 1.3

4.1 Independence

As is often the case with concepts that have definitions both in English and Mathematics, it's easy to mis-interpret the meaning of independence in the context of probability. Here's what it means in probability:

Definition 4.1. If $A, B \in \mathcal{F}$ and $P(A) \neq 0$, we define the *conditional probability of B given A* by

$$P(B|A) = \frac{P(AB)}{P(A)}.$$

We say that A and B are *independent* if

$$P(AB) = P(A)P(B).$$

Note 4.1. Our definitions mean that if A and B are independent,

$$P(B|A) = P(B).$$

In other words, the occurrence of A has no influence on the probability of B. This is to be contrasted with the statement "the occurrence of A has no influence on the occurrence of B", which is how one generally thinks of independence and does not correspond to the probabilistic meaning of the word. For example, in the experiment of the toss of two dice, the events $A = \{\text{the outcome of the first die is 3}\}$ and $B = \{\text{the sum of the outcomes of the dice is 7}\}$ are independent, which can be counterintuitive if one thinks of independence in non-probabilistic terms.

Definition 4.2. A collection of events $\{E_\alpha\}_{\alpha \in I} \subset \mathcal{F}$ is *independent* if for all $\{i_j\}_{1 \leq j \leq n} \in I$,

$$P(\bigcap_{j=1}^{n} E_{i_j}) = \prod_{j=1}^{n} P(E_{i_j}).$$

A collection of random variables $\{X_\alpha\}_{\alpha \in I}$ is *independent* if for all $\{i_j\}_{1 \leq j \leq n} \in I, \{B_{i_j}\}_{1 \leq j \leq n} \in \mathcal{R}$,

$$P(\bigcap_{j=1}^{n} X_{i_j}^{-1}(B_{i_j})) = \prod_{j=1}^{n} P(X_{i_j}^{-1}(B_{i_j})).$$

A collection of classes of events $\{\mathcal{C}_\alpha\}_{\alpha \in I} \subset \mathcal{F}$ is *independent* if for all $\{i_j\}_{1 \leq j \leq n} \in I, C_{i_j} \in \mathcal{C}_{i_j}$,

$$P(\bigcap_{j=1}^{n} C_{i_j}) = \prod_{j=1}^{n} P(C_{i_j}).$$

Pairwise independence in a collection of events, r.v.'s, or sigma-algebras refers to independence between any two elements of the collection.

Note 4.2. Pairwise independence doesn't necessarily imply independence. See Example 2.1.1 in Durrett.

Definition 4.3. A collection $\mathcal{P}$ of sets is a *π-system* if whenever $A, B \in \mathcal{P}$, we also have $A \cap B \in \mathcal{P}$.

A collection $\mathcal{L}$ of sets is a *λ-system* if

1. $\Omega \in \mathcal{L}$,
2. If $A, B \in \mathcal{L}$ and $A \subset B$, then $B \setminus A \in \mathcal{L}$,
3. If $A_n \in \mathcal{L}$ and $A_n \uparrow A$, then $A \in \mathcal{L}$.

Theorem 4.1. (Dynkin's $\pi - \lambda$ theorem) If $\mathcal{P}$ is a π-system, $\mathcal{L}$ is a λ-system, and $\mathcal{P} \subset \mathcal{L}$, then

$$\sigma(\mathcal{P}) \subset \mathcal{L}.$$

Theorem 4.2. Suppose classes of π-systems $\{\mathcal{P}_\alpha\}_{\alpha \in I} \subset \mathcal{F}$ are independent. Then $\{\sigma(\mathcal{P}_\alpha)\}_{\alpha \in I}$ are independent too.

Proof. See Durrett, p. 39.

□

Corollary 4.1. If for all $x_1, \ldots, x_n \in (-\infty, \infty]$,

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} P(X_i \leq x_i),$$

then $X_1, \ldots, X_n$ are independent.

Proof. Let $\mathcal{A}_i$ be the family of sets of the form $\{X_i^{-1}(-\infty, x]\}$, which is a π-system, since $X_i^{-1}(-\infty, x] \cap X_i^{-1}(-\infty, y] = X_i^{-1}(-\infty, x \wedge y]$. Since the sets $(-\infty, x]$ generate $\mathcal{R}$, the $\mathcal{A}_i$ generate $\sigma(X_i)$ (see Exercise 1.3.1), so $\sigma(\mathcal{A}_i) = \sigma(X_i)$. So if we assume that the $\mathcal{A}_i$ are independent, by Theorem 4.2, so are the $\sigma(X_i)$ and thus, by definition of independence of random variables, so are the X_i. □

4.2 Product Spaces

Definition 4.4. Suppose $(\Omega_i, \mathcal{F}_i, P_i)$ are probability spaces. Then

$$\Omega := \Omega_1 \times \cdots \times \Omega_n = \{(\omega_1, \ldots, \omega_n) : \omega_1 \in \Omega_1, \ldots, \omega_n \in \Omega_n\},$$

$$\mathcal{F} := \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$$

is the sigma-algebra generated by the collection of rectangles $\{A_1 \times \cdots \times A_n : A_1 \in \mathcal{F}_1, \ldots, A_n \in \mathcal{F}_n\}$.

Theorem 4.3. Given probability spaces $(\Omega_i, \mathcal{F}_i, P_i)$, there is a unique probability measure

$$P = P_1 \times \cdots \times P_n$$

on $(\Omega, \mathcal{F})$ such that if $A_i \in \mathcal{F}_i, 1 \leq i \leq n$, then

$$P_1 \times \cdots \times P_n(A_1 \times \cdots \times A_n) = P_1(A_1) \cdots P_n(A_n).$$

Proof. See proof of Theorem 1.7.1 in Durrett □

This probability measure is a particularly important one (one of many, of course) on $(\Omega, \mathcal{F})$ since, as we will now see, it generates the joint measure of an n-tuple of independent random variables.

Of course, in order for the notion of independence to make sense, we need the random variables to live on the same space. On the other hand, two given random variables don't need a priori to live on the same probability space (think, for instance, of X_1 as recording the value of a 6-sided die and X_2 as counting the number of heads in a flip (independent of the die throw) of a coin. Therefore, we may need to embed our probability spaces into a larger one (and thus use projections when we only care about a subset of the random variables).

Now in order to be able to talk about the independence of the X_i, we need to put them on a same space. We do this by embedding them in Ω: We define $\tilde{X}_i : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{R}, P^{X_i})$ by

$$\tilde{X}_i(\omega_1, \ldots, \omega_n) = X_i(\omega_i).$$

Note, in particular that if $B_i \in \mathcal{R}$ and $X_i^{-1}(B_i) = A_i$, then $\tilde{X}_i^{-1}(B_i) = \Omega_1 \times \ldots \Omega_{i-1} \times A_i \times \Omega_{i+1} \times \ldots \times \Omega_n$.

Since X_i and $\tilde{X}_i$ coincide on Ω_i and are thus essentially the same, we drop the $\tilde{}$ and revert to writing X_i, being aware that it is now defined on Ω, not just Ω_i (though for ease of notation we will use X_i both for the original r.v. and its embedded version in everything that follows).

Let $(\Omega, \mathcal{F}, P)$ be as above and define $(X_1, \ldots, X_n) : (\Omega, \mathcal{F}, P) \to (\mathbb{R}^n, \mathcal{R}^n, P^{(X_1, \ldots, X_n)})$ by

$$(X_1, \ldots, X_n)(\omega_1, \ldots, \omega_n) = (X_1(\omega_1), \ldots, X_n(\omega_n)).$$

(Note that on the right of this last equality, X_i has its original meaning.) Then it is easy to check that $(X_1, \ldots, X_n)$ is a random vector, i.e. a measurable function.

Theorem 4.4. Suppose $\{X_i\}_{1 \leq i \leq n}$ are independent random variables and that X_i has distribution P^{X_i}. Then $(X_1, \ldots, X_n) : (\Omega, \mathcal{F}, P) \to (\mathbb{R}^n, \mathcal{R}^n, P^{(X_1, \ldots, X_n)})$ has distribution $P^{X_1} \times \cdots \times P^{X_n}$. That is,

$$P^{(X_1, \ldots, X_n)} = P^{X_1} \times \cdots \times P^{X_n}.$$

Before proving this theorem, we state a consequence of the $\pi - \lambda$ theorem without proof (for the proof see Theorem A.1.5 in Durrett):

Theorem 4.5. Suppose $\mathcal{P}$ is a π-system and P_1, P_2 are probability measures that agree on $\mathcal{P}$. Then P_1 and P_2 agree on $\sigma(\mathcal{P})$.

Proof of Theorem 7.2 We show that $P^{(X_1,\ldots,X_n)}$ and $P^{X_1} \times \cdots \times P^{X_n}$ agree on the π-system of sets of the form $A_1 \times \ldots \times A_n$, where the A_i are Borel sets. Using Theorem 4.5, we then are done.

$$P^{(X_1,\ldots,X_n)}(A_1 \times \ldots \times A_n) = P((X_1,\ldots,X_n) \in (A_1 \times \ldots \times A_n)) = P(X_1 \in A_1, \ldots, X_n \in A_n).$$

Using independence, the definition of P^{X_i} and $P^{X_1} \times \cdots \times P^{X_n}$, we see that this last term is equal to

$$\prod_{i=1}^{n} P(X_i \in A_i) = \prod_{i=1}^{n} P^{X_i}(A_i) = P^{X_1} \times \cdots \times P^{X_n}(A_1 \times \ldots \times A_n).$$

□

4.3 Convolution

Definition 4.5. The convolution of two distribution functions F and G is

$$F * G(z) = \int F(z-y) dG(y),$$

where $dG(y)$ is the measure P^Y associated with the distribution function G.

Theorem 4.6. If X and Y have distribution functions F and G, respectively, then

$$P(X+Y \le z) = F * G(z).$$

Moreover, if X has density f, $X+Y$ has density

$$f_{X+Y}(z) = \int f(z-y) dG(y).$$

If Y also has a density g,

$$f_{X+Y}(z) = \int f(z-y) g(y)\, dy.$$

Proof.

$$P(X+Y \le z) = \iint \mathbb{1}_{\{x+y \le z\}} P^X(dx) P^Y(dy) = \int P(X \le z-y) P^Y(dy) = \int F(z-y) dG(y).$$

If X has a density, $F(x) = \int_{-\infty}^{x} f(y)\, dy$, so

$$P(X+Y \le z) = \int \int_{-\infty}^{z} f(x-y)\, dx\, dG(y) = \int_{-\infty}^{z} \int f(x-y)\, dG(y)\, dx.$$

To deal with the case where Y has a density, see Exercise 1.6.8 in Durrett.

□

Example 4.1. One can use convolution to show for instance:

- If X, Y are independent normal random variables with means μ_X, μ_Y and standard deviations σ_X, σ_Y, then $X + Y$ is normal with mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$
- If X, Y are independent and uniform on $[0, 1]$,

$$f_{X+Y}(a) = \begin{cases} a, & 0 \leq a \leq 1 \\ 2 - a, & 1 \leq a \leq 2 \\ 0, & \text{otherwise} \end{cases} .$$

Math 83100 (Fall 2013) September 17, 2008
Prof. Christian Beneš

Lecture #5: Independence, Laws of Large Numbers, Borel-Cantelli Lemmas

Reference. Sections 2.1, 2.2

5.1 Independence Revisited

We now summarize the results of the previous section.

Definition 5.1. If $(X_1, \dots, X_n)$ is an n-dimensional random vector, recall that its *distribution* $P^{(X_1,\dots,X_n)}$ is defined by

$$P^{(X_1,\dots,X_n)}(A) = P((X_1, \dots, X_n) \in A), \quad A \in \mathcal{R}^n.$$

Its *distribution function* F is defined by

$$F(x_1, \dots, x_n) = P(X_1 \leq x_1, \dots, X_n \leq x_n) = P^{(X_1,\dots,X_n)}(\{y \in \mathbb{R}^n : y_1 \leq x_1, \dots, y_n \leq x_n\}).$$

$X_1, \dots, X_n$ are independent if

$$P(X_1 \in A_1, \dots, X_n \in A_n) = P(X_1 \in A_1) \cdots P(X_n \in A_n)$$

for all $A_1, \dots, A_n \in \mathcal{R}$.

The previous section shows that $X_1, \dots, X_n$ are independent if and only if one of the following is satisfied:

- $$F(x_1, \dots, x_n) = F(x_1) \cdots F(x_n).$$
- $$P^{(X_1,\dots,X_n)} = P_1 \times \cdots \times P_n.$$
- If for $1 \leq i \leq n$, P_i has density f_i and $P^{(X_1,\dots,X_n)}$ has density f,
$$f(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n).$$

5.2 Fubini and Consequences

Recall

Theorem 5.1. (Fubini's theorem) If $(\Omega, F, P) = (\Omega_1, \mathcal{F}_1, P_1) \times (\Omega_2, \mathcal{F}_2, P_2)$, X is a random variable on (Ω, F, P) such that $X \geq 0$ or $\int |X| \, dP < \infty$, then

$$\int_{\Omega_1} \int_{\Omega_2} X \, dP_2 dP_1 = \int_{\Omega} X dP = \int_{\Omega_2} \int_{\Omega_1} X \, dP_1 dP_2.$$

Theorem 5.2. Suppose X and Y are independent random variables with distributions P^X and P^Y. If $h : \mathbb{R}^2 \to \mathbb{R}$ is measurable with $h \geq 0$ or $E[|h(X,Y)|] < \infty$, then

$$E[h(X,Y)] = \iint h(x,y)P^X(dx)P^Y(dy).$$

In particular, if $f, g : \mathbb{R} \to \mathbb{R}$ are measurable with $f, g \geq 0$ or $E[|f(X)|] < \infty, E[|g(Y)|] < \infty$, then

$$E[f(X)g(Y)] = E[f(X)]E[g(Y)].$$

Proof. This follows from the change of variables formula and Fubini. □

As an immediate corollary, we have the following:

Theorem 5.3. If $X_1, \ldots, X_n \geq 0 \ \forall \ 1 \leq i \leq n$ or $X_i \in L^1 \ \forall \ 1 \leq i \leq n$ and $X_1, \ldots, X_n$ are independent, then

$$E\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} E[X_i].$$

Note 5.1. The converse is not necessarily true. See Example 2.1.2 in Durrett, an example of dependent random variables X, Y satisfying $E[XY] = E[X]E[Y]$.

However, recalling the definition of correlation, we see that when $X, Y \in L^2, \rho(X,Y) = 0 \iff E[XY] = E[X]E[Y]$.

So $X, Y \in L^2$ are uncorrelated if they are independent, but they may be dependent despite being uncorrelated.

5.3 Kolmogorov Extension

We now know that finite n-tuples of random variables X_i with probability measure P_i exist: Just let $(\Omega, \mathcal{F}, P) = (\mathbb{R}^n, \mathcal{R}^n, P_1 \times \cdots P_n)$ and define $X_i(\omega_1, \ldots, \omega_n) = \omega_i$.

It is not obvious that we can do the same thing for infinite sequences of random variables, but it turns out to be true as well:

In the infinite-dimensional case, we define $\Omega = \mathbb{R}^{\mathbb{N}} = \{\omega : \mathbb{N} \to \mathbb{R}\}, \mathcal{F} = \mathcal{R}^{\mathbb{N}}$, wich is the sigma-algebra generated by the finite-dimensional sets $\{\omega : \omega_i \in B_i, B_i \in \mathcal{R}, 1 \leq i \leq n\}$. Kolmogorov showed that you can consistently define a product measure on $(\mathbb{R}^{\mathbb{N}}, \mathcal{R}^{\mathbb{N}})$:

Theorem 5.4 (Kolmogorov Extension Theorem)**.** Suppose we have a sequence of probability measures on $(\mathbb{R}^n, \mathcal{R}^n)$ that are consistent, i.e., for all $n \in \mathbb{N}$,

$$\mu_{n+1}((a_1, b_1] \times \cdots \times (a_n, b_n] \times \mathbb{R}) = \mu_{n+1}((a_1, b_1] \times \cdots \times (a_n, b_n]).$$

Then there exists a unique probability measure P on $(\mathbb{R}^{\mathbb{N}}, \mathcal{R}^{\mathbb{N}})$ which coincides with μ_n on all rectangles of the form $(a_1, b_1] \times \cdots \times (a_n, b_n]$.

In particular, if the X_i are independent,

$$\mu_n((a_1, b_1] \times \cdots \times (a_n, b_n]) = P_1(a_1, b_1] \cdots P_n(a_n, b_n].$$

For more details, see Section 2.1.4 in Durrett.

5.4 Weak Laws of Large Numbers

One of the big objectives in probability is that of understanding $\sum_{i=1}^n (X_i - \mu_i)$ if X_i are random variables with $E[X_i] = \mu_i$. A quick first glance indicates that $\sum_{i=1}^n (X_i - \mu_i)$ has mean zero and should be more and more spread out as n increases. It is then natural to ask if we can normalize it to get a nontrivial random variable

$$\frac{1}{n^a}\sum_{i=1}^n (X_i - \mu_i),$$

where $a > 0$ is some adequately chosen constant. We will answer this question in several steps and will start by focusing on the distribution of the sample mean $(1/n)\sum_{i=1}^n X_i$. We begin with a lemma giving two basic properties of variance, which will be useful when proving a weak version of the weak law of large numbers (WLLN).

Lemma 5.1. Let $X_1, \ldots, X_n$ have finite second moments and be pairwise uncorrelated. Then

1. $$Var(\sum_{i=1}^n X_i) = \sum_{i=1}^n Var(X_i).$$

2. $$Var(cX_1) = c^2 Var(X_1).$$

Proof. We write μ_i for $E[X_i]$.

1. $$\begin{aligned} Var(\sum_{i=1}^n X_i) &= E\left[(\sum_{i=1}^n (X_i - \mu_i))^2\right] = E\left[\sum_{i=1}^n \sum_{j=1}^n (X_i - \mu_i)(X_j - \mu_j)\right] \\ &= E\left[\sum_{i=1}^n (X_i - \mu_i)^2\right] + 2E\left[\sum_{i=1}^n \sum_{j-1}^{i-1} (X_i - \mu_i)(X_j - \mu_j)\right] = \sum_{i=1}^n Var(X_i), \end{aligned}$$

   since if $i \neq j, E[(X_i - \mu_i)(X_j - \mu_j)] = E[X_i X_j] - \mu_i E[X_j] - \mu_j E[X_i] + \mu_i \mu_j = 0$.

2. $$Var(cX_1) = E[(cX_1 - c\mu_1)^2] = c^2 E[(X_1 - \mu_1)^2] = c^2 Var(X_1).$$

□

Standard Notation. In probability, S_n is almost always used to denote a sum of random variables: If $X_1, \ldots, X_n$ are random variables,

$$S_n = \sum_{i=1}^n X_i.$$

In particular, if the X_i are i.i.d., S_n is called a *random walk*.

Theorem 5.5. (L^2 weak law) Let $X_1, \ldots, X_n$ be pairwise uncorrelated random variables with $E[X_i] = \mu$ and $Var(X_i) \leq C < \infty$. Then

$$\frac{S_n}{n} \xrightarrow{L^2} \mu.$$

Proof.

$$E\left[\left(\frac{S_n}{n} - \mu\right)^2\right] = Var\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \leq \frac{Cn}{n^2} = \frac{C}{n} \overset{n\to\infty}{\to} 0.$$

□

Corollary 5.1. Let $X_1, \ldots, X_n$ be pairwise uncorrelated random variables with $E[X_i] = \mu$ and $Var(X_i) \leq C < \infty$. Then

$$\frac{S_n}{n} \xrightarrow{P} \mu.$$

Proof. This follows from the fact that convergence in L^p implies convergence in probability. □

We know the conclusion $S_n/n \xrightarrow{P} \mu$ holds if the X_i have 2 (uniformly bounded) moments. Can we do better? It turns out we can, but (almost) one moment is necessary, as the next example shows:

Example 5.1. Recall that X is a standard Cauchy random variable if it has density function

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

Trying to compute

$$\int_{\mathbb{R}} \frac{x^a}{\pi(1+x^2)},$$

for $a \geq 0$, one sees that X has a moments iff $0 \leq a < 1$.

Using characteristic functions (which we will see soon), one can show that if $X_1, \ldots, X_n$ are independent standard Cauchy random variables, then S_n/n has the same distribution as X_1. Therefore,

$$P\left(\left|\frac{S_n}{n}\right| > \epsilon\right) = P(X_1 > \epsilon) = 1 - \frac{2}{\pi}\arctan(\epsilon) \not\to 0.$$

Theorem 5.6. (Weak Law of Large Numbers) Suppose $X_1, \ldots, X_n$ are i.i.d and that

$$xP(|X_1| > x) \to 0 \text{ as } x \to \infty.$$

Then if $\mu_n = E[X_1 \mathbb{1}_{\{|X_1| \leq n\}}]$,

$$\frac{S_n}{n} - \mu_n \xrightarrow{P} 0.$$

Note 5.2. : It turns out that the Cauchy distribution example is a borderline case. It behaves like $1/x^2$ as $x \to \infty$. Suppose a random variable X has density function behaving like $1/x^{2+\epsilon}$ for some $\epsilon > 0$, as $x \to \infty$. Then $P(|X_1| > x) \leq Cx^{-\epsilon-1}$, so $xP(|X_1| > x) \to 0$ as $x \to \infty$. This suggests that having one moment is almost sufficent for a weak law of large numbers to hold.

Note 5.3. Theorem 5.6 doesn't apply to random variables whose density decays like $1/x^2$, but it does to those that decay like $\frac{1}{x^2 \ln x}$, which are not L^1. This is why truncation is needed (since μ is not defined for such random variables).

Lemma 5.2. If $Y \geq 0$ and $p > 0$, then

$$E[Y^p] = \int_0^\infty p y^{p-1} P(Y > y)\, dy.$$

Proof. This is just Fubini's theorem. See proof of Lemma 2.2.8 in Durrett (it's probably more natural to go from right to left in his sequence of equalities). □

Proof of Theorem 5.6

For $n \geq 1, 1 \leq k \leq n$, let $\bar{X}_{n,k} = X_k \mathbb{1}_{\{|X_k| \leq n\}}$ and $\bar{S}_n = \sum_{k=1}^n \bar{X}_{n,k}$. Then $E[\bar{S}_n] = n\mu_n$. Our goal is to show that

$$P\left(\left|\frac{S_n - E[\bar{S}_n]}{n}\right| > \epsilon\right) \to 0, \text{ as } n \to \infty.$$

We have

$$P\left(\left|\frac{S_n - E[\bar{S}_n]}{n}\right| > \epsilon\right) \leq P(S_n \neq \bar{S}_n) + P\left(\left|\frac{\bar{S}_n - E[\bar{S}_n]}{n}\right| > \epsilon\right).$$

We will estimate both terms and show that they go to 0.

$$P(S_n \neq \bar{S}_n) \leq P(\cup_{k=1}^n \{\bar{X}_{n,k} \neq X_k\}) \leq nP(\bar{X}_{n,1} \neq X_1) = nP(X_1 > n) \to 0,$$

by hypothesis.

Now Chebyshev's inequality implies that

$$P\left(\left|\frac{\bar{S}_n - E[\bar{S}_n]}{n}\right| > \epsilon\right) \leq \left(\frac{1}{\epsilon n}\right)^2 Var(\bar{S}_n) \leq \left(\frac{1}{\epsilon n}\right)^2 \sum_{k=1}^n E[\bar{X}_{n,k}^2] = \left(\frac{1}{\epsilon n}\right)^2 nE[\bar{X}_{n,1}^2].$$

So all that is left is show that $n^{-1}E[\bar{X}_{n,1}^2] \to 0$, as $n \to \infty$. Lemma (5.2) implies that

$$\begin{aligned} E[\bar{X}_{n,1}^2] &= 2\int_0^\infty y P(|\bar{X}_{n,1}| > y)\, dy = 2\int_0^n y P(|\bar{X}_{n,1}| > y)\, dy \\ &= 2\int_0^n y \left(P(|X_1| > y) - P(|X_1| > n)\right)\, dy \leq 2\int_0^n y P(|X_1| > y)\, dy. \end{aligned}$$

Now $0 \le 2yP(|X_1| > y) \le 2y$ and $yP(|X_1| > y) \to 0$ as $y \to \infty$, so $\sup_y yP(|X_1| > y) = M < \infty$. We now define $\epsilon_n := \sup\{yP(|X_1| > y) : y \ge n^{1/2}\}$. Then

$$\int_0^n yP(|X_1| > y)\, dy = \int_0^{\sqrt{n}} yP(|X_1| > y)\, dy + \int_{\sqrt{n}}^n yP(|X_1| > y)\, dy \le M\sqrt{n} + n\epsilon_n.$$

So

$$n^{-1} \int_0^n yP(|X_1| > y)\, dy \le \frac{M}{\sqrt{n}} + \epsilon_n \to 0 \text{ as } n \to \infty,$$

implying that for every $\epsilon > 0$, $\left(\frac{1}{\epsilon n}\right)^2 nE[\bar{X}_{n,1}^2] \to 0$. This proves the theorem. □

We obtain from this a slightly weaker corollary which has the advantage of having nicer looking assumptions:

Theorem 5.7. (WLLN - Standard Form) If $X_1, \ldots, X_n$ are i.i.d. with $E[X_1] = \mu$, then

$$\frac{S_n}{n} \xrightarrow{P} \mu.$$

Proof. We need to show that if $E[|X_1|] < \infty$, then $xP(|X_1| > x) \to 0$ as $x \to \infty$ and that if μ_n is as in Theorem 5.6, then $\mu_n \to \mu$ as $n \to \infty$.

$|X_1|\mathbb{1}_{\{|X_1|>x\}} \to 0$ almost surely since for every $\omega, X_1(\omega)$ is bounded and $|X_1|\mathbb{1}_{\{|X_1|>x\}} \le |X_1| \in L^1$, so by the dominated convergence theorem, $xP(|X_1| > x) = xE[\mathbb{1}_{\{|X_1|>x\}}] \le E[|X_1|\mathbb{1}_{\{|X_1|>x\}}] \to 0$, as $x \to \infty$.

Also, $X_1\mathbb{1}_{\{|X_1|\le n\}} \to X_1$ almost surely and $|X_1\mathbb{1}_{\{|X_1|\le n\}}| \le |X_1| \in L^1$. So by the dominated convergence theorem, $\mu_n = E[X_1\mathbb{1}_{\{|X_1|\le n\}}] \to \mu$ as $n \to \infty$. Therefore,

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \le P\left(\left|\frac{S_n}{n} - \mu_n\right| > \frac{\epsilon}{2}\right) + \mathbb{1}_{\{|\mu_n - \mu| > \epsilon/2\}} \to 0.$$

□

Note 5.4. Why is independence necessary? Suppose $\{X_i\}_{i\ge1}$ satisfy $X_i(\omega) = X_j(\omega)$ for all i, j and $P(X_1 = 0) = 0, E[X] = 0$. Then $P(|S_n/n| > \epsilon) = P(|X_1| > \epsilon) \to 1$ as $\epsilon \to 0$.

Note 5.5. The WLLN tells us that

$$S_n = n\mu + \phi(n), \text{ where } \frac{\phi(n)}{n} \xrightarrow{P} 0.$$

Though this is a useful result, it is rather imprecise, as it tells us almost nothing about $\phi(n)$. Our goal in what follows will be to find a such that

$$\frac{\phi(n)}{n^a} \to \text{ nontrivial distribution },$$

and of course, to find the distribution as well.

5.5 Borel-Cantelli Lemmas

Definition 5.2. Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $(A_n)_{n\geq 1}$ be a sequence of events in $\mathcal{A}$.

$$\limsup_{n\to\infty} A_n := \bigcap_{n=1}^{\infty} \bigcup_{m\geq n} A_m = \{\omega : \forall m \geq 1 \; \exists \, n(\omega) \geq m \text{ such that } \omega \in A_{n(\omega)}\},$$

which can be interpreted probabilistically as

$$\{A_n \text{ occur infinitely often}\} =: \{A_n \text{ i.o.}\}.$$

$$\begin{aligned} \liminf_{n\to\infty} A_n := \bigcup_{n=1}^{\infty} \bigcap_{m\geq n} A_m &= \{\omega : \exists \, m(\omega) \geq 1 \text{ such that } \; \forall \, n \geq m(\omega), \; \omega \in A_n\} \\ &= \{\omega : \omega \in A_n \text{ for large enough } n\}. \end{aligned}$$

The probabilistic meaning of the liminf is

$$\{A_n \text{ occur eventually }\} =: \{A_n, \text{ ev.}\}.$$

Note that

$$\{A_n, \text{ ev.}\}^c = \{A_n^c, \text{ i.o.}\}.$$

Theorem 5.8 (First Borel-Cantelli Lemma)**.** Let A_n be a sequence of events in $(\Omega, \mathcal{A}, P)$. If

$$\sum_{n=1}^{\infty} P(A_n) < \infty,$$

then $P\{A_n \text{ i.o.}\} = 0$.

Theorem 5.9 (Second Borel-Cantelli Lemma)**.** If A_n are independent and if

$$\sum_{n=1}^{\infty} P(A_n) = \infty,$$

then $P(A_n \text{ i.o.}) = 1$.

Math 83100 (Fall 2013) October 15, 2013
Prof. Christian Beneš

Lecture #6: Borel-Cantelli Lemmas and Applications; Zero-One Laws

Reference. Sections 1.3, 1.4

Theorem 6.1 (First Borel-Cantelli Lemma)**.** Let A_n be a sequence of events in $(\Omega, \mathcal{A}, P)$. If

$$\sum_{n=1}^{\infty} P(A_n) < \infty,$$

then $P\{A_n \text{ i.o.}\} = 0$.

Proof. By Fubini, if

$$\sum_{n=1}^{\infty} E[\mathbb{1}_{A_n}] = \sum_{n=1}^{\infty} P(A_n) < \infty,$$

then

$$E[\sum_{n\geq 1} \mathbb{1}_{A_n}] < \infty,$$

implying that

$$\sum_{n=1}^{\infty} \mathbb{1}_{A_n} < \infty \text{ a.s.}$$

Moreover,

$$\sum_{n=1}^{\infty} \mathbb{1}_{A_n}(\omega) = \infty \iff \omega \in \limsup_{n\to\infty} A_n.$$

Together these give

$$\sum_{n=1}^{\infty} P(A_n) < \infty \quad \Rightarrow \quad P(A_n \text{ i.o.}) = 0.$$

□

Theorem 6.2 (Second Borel-Cantelli Lemma)**.** If A_n are independent and if

$$\sum_{n=1}^{\infty} P(A_n) = \infty,$$

then $P(A_n \text{ i.o.}) = 1$.

Proof.

$$P((\limsup A_n)^c) = P(\liminf A_n^c) = P\left(\bigcup_{m\geq 1}\bigcap_{n\geq m} A_n^c\right) \leq \sum_{m\geq 1} P\left(\bigcap_{n\geq m} A_n^c\right).$$

If we write $p_n = P(A_n)$ and use the fact that the A_n are independent, we get

$$P\left(\bigcap_{n\geq m} A_n^c\right) = \prod_{n\geq m}(1-p_n).$$

The Taylor expansion of e^{-x} shows that if $x \geq 0, 1-x \leq e^{-x}$, so that

$$\prod_{n\geq m}(1-p_n) \leq \exp\left(-\sum_{n\geq m} p_n\right) = 0,$$

since $\sum_{n\geq 1} p_n = \infty$. Therefore, $P(\limsup A_n) = 1$.

□

Example 6.1. (B-C 2 doesn't hold without the assumption of independence) Suppose that $\Omega = [0,1]$ and P is the uniform probability measure on $[0,1]$. For $n = 1, 2, \ldots$, let

$$A_n = \left[0, \tfrac{1}{n}\right]$$

Then

$$\{A_n \text{ i.o.}\} = \bigcap_{m=1}^{\infty} \bigcup_{n\geq m} A_n = \bigcap_{m=1}^{\infty} \left[0, \tfrac{1}{n}\right] = \{0\}$$

so $P(A_n \text{ i.o.}) = 0$. The A_n are not independent since if $1 < k < j$, then $A_k \cap A_j = A_j$, implying that

$$\frac{1}{k}\frac{1}{j} = P(A_k)P(A_j) \neq P(A_k \cap A_j) = P(A_j) = \frac{1}{j}.$$

However,

$$\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

Therefore, $\sum_{n=1}^{\infty} P(A_n) = \infty$ does not generally imply $P\{A_n \text{ i.o.}\} = 1$.

Example 6.2. (lim sup of a sequence of independent random variables) Suppose $\{X_n\}_{n\geq 1}$ are independent and exponentially distributed: If $x \geq 0, P(X_1 > x) = e^{-x}$. Then if $\alpha > 0, P(X_n > \alpha \log n) = n^{-\alpha}$. It then follows from the two Borel-Cantelli lemmas that

$$P(X_n > \alpha \log n, \text{ i.o}) = \begin{cases} 0 \text{ if } \alpha > 1 \\ 1 \text{ if } \alpha \leq 1 \end{cases}.$$

Now define $L = \limsup \frac{X_n}{\log n}$ (here, the lim sup is for sequences of real numbers and should be thought of as a random variable, defined omega by omega). We will show that $L = 1$, a.s. by showing that $P(L \geq 1) = 1$ and $P(L > 1) = 0$.

If $X_n > \log n$, i.o., then $L \geq 1$, so

$$1 = P(X_n > \log n, \text{ i.o.}) \leq P(L \geq 1).$$

On the other hand, if $k \in \mathbb{N}$,

$$P(L > 1) = P\left(\bigcup_{k\geq 1}\{L > 1 + \frac{2}{k}\}\right) \leq \sum_{k\geq 1} P\left(L > 1 + \frac{2}{k}\right) \leq \sum_{k\geq 1} P\left(\frac{X_n}{\log n} > 1 + \frac{1}{k}, \text{i.o.}\right) = 0.$$

If $\{X_n\}_{n\geq 1}$ are independent and normally distributed with density $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, one might expect a smaller lim sup than in the exponential case, since the density decays faster and large values are therefore less likely for the X_n. This is indeed the case. On can show as above that

$$\limsup \frac{X_n}{\sqrt{2\log n}} = 1.$$

We will see this method again later when looking for $\limsup S_n$, where $S_n = \sum_{i=1}^{n} X_i$.

6.1 Pólya's Theorem

Example 6.3 (Recurrence/Transience of Simple random walk)**.** If $\{e_1, \ldots, e_d\}$ is the canonical basis of $\mathbb{R}^d$, $\{X_i\}_{i\geq 1}$ are independent random vectors with distribution

$$P(X_i = \pm e_j) = \frac{1}{2d},$$

for all $j \in \{1, \ldots, d\}, S(0) = 0$, and for $n \geq 1, S(n) = \sum_{i=1}^{n} X_i$, then $\{S(n)\}_{n\geq 0}$ is called a d-dimensional simple random walk (SRW).

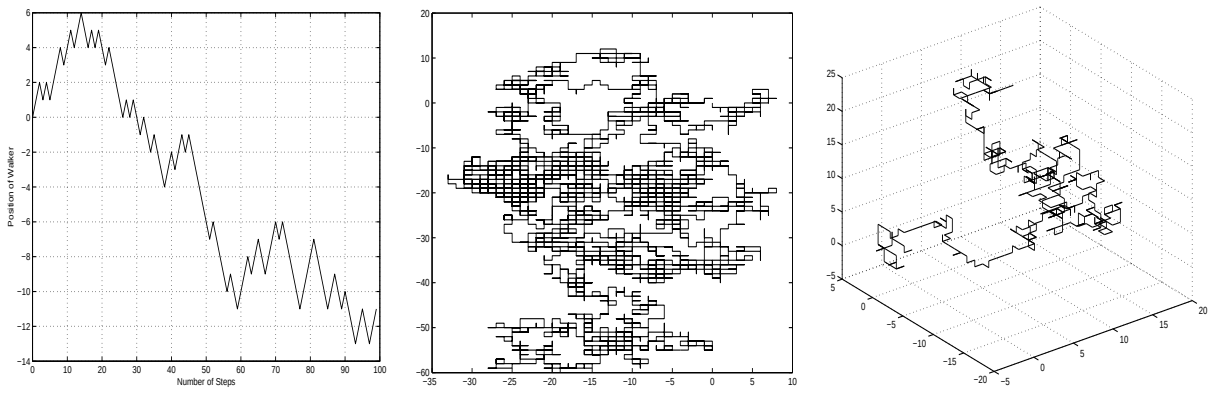

Figure 1: Simple random walks in 1, 2, and 3 dimensions

A random walk is called *recurrent* if has probability one of returning to the origin and *transient* otherwise. A natural question is: Is d-dimensional simple random walk transient or recurrent? You should be able to quickly convince yourself that

- If SRW is recurrent in dimension d_r, then it is recurrent in dimension d with $1 \leq d \leq d_r$.

- If SRW is transient in dimension d_t, then it is transient in dimension d with $d \geq d_t$.

Lemma 6.1. Simple random walk $\{S(n)\}$ is recurrent if and only if $\sum_{n=0}^{\infty} P(S(n) = 0) = \infty$.

Proof. Let

$$r^{(n)} = P(S(1) \neq 0, \ldots, S(n-1) \neq 0, S(n) = 0)$$

be the probability of a first return to 0 at time n and let

$$r = P(\bigcup_{n \geq 1} \{S(n) = 0\}) = \sum_{n \geq 1} r^{(n)}$$

be the probability of eventually returning to 0. Then if $1 \leq n_1 < \ldots < n_k$, using the fact that $P(S(n) = 0 | S(k) = 0) = P(S(n-k) = 0)$ and defining

$$A(n_1, \ldots, n_k) = \{S(n_i) = 0 \text{ for } 1 \leq i \leq k; S(l) \neq 0 \text{ for } l \leq n_k, l \notin \{n_i\}_{1 \leq i \leq k}\},$$

we get

$$P(A(n_1, \ldots, n_k)) = r^{(n_1)} \cdots r^{(n_k - n_{k-1})}$$

and for all $k \geq 1$,

$$\begin{aligned} P(S \text{ returns to } 0 \ k \text{ times }) &= P(S(n) = 0 \text{ for } k \text{ different values of } n \neq 0) \\ &= \sum_{\{n_1, \ldots, n_k \in \mathbb{N}^k : 1 \leq n_1 < \cdots < n_k\}} P(A(n_1, \ldots, n_k)) = r^k. \end{aligned}$$

This last equality follows from the fact that

$$\sum P(A(n_1, \ldots, n_k)) = \sum r^{(n_1)} \cdots r^{(n_k - n_{k-1})} = \sum_{1 \leq m_1, \ldots, m_k} r^{(m_1)} \cdots r^{(m_k)} = r^k.$$

Letting $k \to \infty$, we see that if $r < 1, P(S(n) = 0 \text{ i.o.}) = 0$. Moreover, if $r = 1$,

$$\begin{aligned} P(S(n) = 0, i.o.) &= 1 - P(\cup_{m \geq 1} \cap_{n \geq m} \{S(n) \neq 0\}) \\ &\geq 1 - \sum_{m \geq 1} P(\cap_{n \geq m} \{S(n) \neq 0\}) \geq 1 - \sum_{m \geq 1} (1 - r^{m-1}) = 1. \end{aligned}$$

Therefore,

$$P(S(n) = 0 \text{ i.o.}) = \begin{cases} 0, r < 1 \\ 1, r = 1 \end{cases}.$$

B-C 1 implies that if $\sum_{n=0}^{\infty} P(S(n) = 0) < \infty$, then $P(S(n) = 0, \text{i.o.}) = 0$.

So we've shown that $\sum_{n\geq 1} P(S(n)=0) < \infty \Rightarrow r < 1$. Let's show the converse, using a *first-passage decomposition* (where we decompose the event $\{S(n)=0\}$ into sub-events where the first visit to 0 is at time $1, 2, \ldots, n$). Assume $r < 1$.

$$\begin{aligned} P(S(n)=0) &= \sum_{k=0}^{n-1} P(S(1)\neq 0, \ldots, S(n-k-1)\neq 0, S(n-k)=0, S(n)=0) \\ &= \sum_{k=0}^{n-1} P(S(1)\neq 0, \ldots, S(n-k-1)\neq 0, S(n-k)=0) P(S(n)=0) \\ &= \sum_{k=0}^{n-1} r^{(n-k)} P(S(k)=0). \end{aligned}$$

Therefore, reversing the order of summation, we get for any $m \geq 1$,

$$\begin{aligned} \sum_{n=1}^{m} P(S(n)=0) &= \sum_{n=1}^{m}\sum_{k=0}^{n-1} r^{(n-k)} P(S(k)=0) = \sum_{k=0}^{m-1}\sum_{n=k+1}^{m} r^{(n-k)} P(S(k)=0) \\ &= \sum_{k=0}^{m-1} P(S(k)=0) \sum_{n=k+1}^{m} r^{(n-k)} \leq r \sum_{k=0}^{m-1} P(S(k)=0) \\ &\leq r \sum_{k=0}^{m} P(S(k)=0) = r\left(1 + \sum_{k=1}^{m} P(S(k)=0)\right). \end{aligned}$$

Therefore, since $r < 1$, we have for all $m \geq 1$,

$$\sum_{n=1}^{m} P(S(n)=0) \leq \frac{r}{1-r} < \infty.$$

Letting $m \to \infty$, we get

$$\sum_{n=1}^{\infty} P(S(n)=0) \leq \frac{r}{1-r} < \infty.$$

□

We now turn to the question of determining whether d-dimensional simple random walk is recurrent or transient. Let S_n^d denote a simple random walk in d dimensions and $A_n^d = \{S_n^d = 0\}$. Clearly, $P(S_{2n+1}^d = 0) = 0 \; \forall \; n \geq 0$, so we will focus on A_{2n}^d for $n \geq 0$.

- $\underline{d=1}$:

$$P(A_{2n}^1) = \binom{2n}{n}\left(\frac{1}{2}\right)^{2n} \sim \frac{(2n/e)^{2n}\sqrt{4\pi n}}{\left((n/e)^n\sqrt{2\pi n}\right)^2}\left(\frac{1}{2}\right)^{2n} = 2^{2n}\left(\frac{1}{2}\right)^{2n}\frac{1}{\sqrt{\pi n}} = \frac{1}{\sqrt{\pi n}}.$$

  In particular, there exists $C > 0$ such that for all $n \geq 1$,

$$P(A_{2n}^1) \leq \frac{C}{\sqrt{n}}.$$

Since $\sum_{n\geq 1} P(A^1_{2n}) = \infty$, Lemma 6.1 tells us that 1-dimensional SRW is recurrent.

- $\underline{d = 2:}$

$$\begin{aligned} P(A^2_{2n}) = &= \sum_{k=0}^{n} \frac{(2n)!}{k!k!(n-k)!(n-k)!}(\frac{1}{4})^{2n} \\ &= (\frac{1}{4})^{2n}\frac{(2n)!}{n!n!}\sum_{k=0}^{n}\frac{n!n!}{k!k!(n-k)!(n-k)!} \\ &= (\frac{1}{4})^{2n}\binom{2n}{n}\sum_{k=0}^{n}\binom{n}{k}^2 = \left((\frac{1}{2})^{2n}\binom{2n}{n}\right)^2. \end{aligned}$$

This is the square of the 1-d case, and so

$$P(A^2_{2n}) \sim \frac{1}{\pi n},$$

Again, there exists $C > 0$ such that for all $n \geq 1$,

$$P(A^2_{2n}) \leq \frac{C}{n}$$

and again, $\sum_{n\geq 1} P(A^2_{2n}) = \infty$, so Lemma 6.1 tells us that 2-dimensional SRW is recurrent.

- $\underline{d = 3:}$ We denote by $H_k = H_k(n)$ the event that $2k$ steps of $S^3[0, 2n]$ are taken in a direction parallel to e_1 or e_2. Then

$$\begin{aligned} P(A^3_{2n}) &= \sum_{k=0}^{n} P(A^3_{2n}; H_k) = \sum_{k=0}^{n} P(A^3_{2n}|H_k)P(H_k) \\ &= \sum_{k=0}^{n} P(A^2_{2k})P(A^1_{2n-2k})\binom{2n}{2k}\left(\frac{2}{3}\right)^{2k}\left(\frac{1}{3}\right)^{2(n-k)} \\ &\leq 2\left(\frac{2}{3}\right)^n + C3^{-2n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k}}\frac{1}{n-k}\binom{2n}{2k}2^{2k}. \end{aligned}$$

There are many ways to attack such a sum. Here's one not so elegant but intuitive version:

There are two essential components to this sum. On one hand, we can note that

$$C_1 n^{-3/2} \leq \frac{1}{\sqrt{k}}\frac{1}{n-k} \leq C_2 n^{-1/2}$$

and that $\frac{1}{\sqrt{k}}\frac{1}{n-k}$ is minimal when k is close to $n/2$ (in which case it's bounded above by $Cn^{-3/2}$. On the other hand, $\binom{2n}{2k}\left(\frac{2}{3}\right)^{2k}\left(\frac{1}{3}\right)^{2(n-k)}$ is minimal when k is close to 0 or to n (in which case it's very small).

Using this heuristic argument, you can show

Exercise 6.1. There exists a constant $C > 0$ such that for all $n \geq 1$,

$$P(A_{2n}^3) \leq Cn^{-3/2}.$$

This and Lemma 6.1 is all that is needed to show that 3-dimensional SRW is transient.

- $\underline{d \geq 4}$: The process obtained from S_n^d by considering only the times at which one of the first 3 components changes is a 3-dimensional random walk. Since this is transient, so is S_n^d. This can easily be made perfectly precise (see pp. 185-186 in Durrett).

We just proved

Theorem 6.3. Simple random walk is recurrent in dimensions $d \leq 2$ and transient in dimensions $d \geq 3$.

Math 83100 (Fall 2013) October 21, 2013
Prof. Christian Beneš

Lecture #7: Strong Law of Large Numbers

Reference. Sections 2.3-2.5

There are a number of different versions of laws of large numbers. We start with one that has relatively loose hypothesis and is easy to prove. Stronger hypotheses will lead to a much more involved proof based on generalities about random series later.

7.1 A Weak Strong Law of Large Numbers

Theorem 7.1 (Cantelli)**.** Suppose that $X_1, X_2, \ldots$ are independent and identically distributed L^4 random variables with common mean $E(X_1) = \mu$. Then

$$\frac{S_n}{n} \to \mu \quad \text{a.s.},$$

as $n \to \infty$.

Proof. [Note that this proof is essentially the same as that of Theorem 2.3.5 in Durrett, except for the last step which bypasses the use of Borel-Cantelli.] We write σ^2 for the common variance of the X_j and let

$$\tilde{S}_n = X_1 + \cdots + X_n - n\mu.$$

Our goal now is to estimate $E(\tilde{S}_n^4)$ and show that $\frac{E(\tilde{S}_n^4)}{n^4\epsilon^4}$ is summable in n.

If we write $Y_i = X_i - \mu$, then

$$\begin{aligned}\tilde{S}_n^4 &= (Y_1 + Y_2 + \cdots + Y_n)^4 \\ &= \sum_{j=1}^n Y_j^4 + C_1 \sum_{i\neq j} Y_i^3 Y_j + \binom{4}{2} \sum_{i<j} Y_i^2 Y_j^2 + C_2 \sum_{i\neq j\neq k} Y_i^2 Y_j Y_k + C_3 \sum_{i\neq j\neq k\neq \ell} Y_i Y_j Y_k Y_\ell.\end{aligned}$$

Since $Y_1, Y_2, \ldots$ are independent with $E(Y_1) = 0$, we see that

$$E\left(\sum_{i\neq j} Y_i^3 Y_j\right) = \sum_{i\neq j} E(Y_i^3 Y_j) = \sum_{i\neq j} E(Y_i^3)E(Y_j) = 0$$

as well as

$$E\left(\sum_{i\neq j\neq k} Y_i^2 Y_j Y_k\right) = \sum_{i\neq j\neq k} E(Y_i^2)E(Y_j)E(Y_k) = 0$$

and

$$E\left(\sum_{i\neq j\neq k\neq \ell} Y_i Y_j Y_k Y_\ell\right) = \sum_{i\neq j\neq k\neq \ell} E(Y_i)E(Y_j)E(Y_k)E(Y_\ell) = 0.$$

This gives

$$E(\tilde{S}_n^4) = \sum_{j=1}^{n} E(Y_j^4) + \binom{4}{2} \sum_{i<j} E(Y_i^2)E(Y_j^2).$$

Since $X_j \in L^4$, we see that $Y_j \in L^4$ and $\tilde{S}_n \in L^4$. Thus, if we write $E(Y_j^4) = M$, then since $E(Y_j^2) = \sigma^2$, we see that there exists some constant C such that

$$E(\tilde{S}_n^4) = nM + \binom{4}{2} \frac{n(n-1)}{2} \sigma^4 \leq Cn^2.$$

($C = M + 3\sigma^4$ works since

$$nM + \binom{4}{2} \frac{n(n-1)}{2} \sigma^4 = nM + 3n^2\sigma^4 - 3n\sigma^4 \leq nM + 3n^2\sigma^4 \leq n^2(M + 3\sigma^4).)$$

So, by Fubini,

$$E\left(\sum_{n\geq 1} \frac{\tilde{S}_n^4}{n^4} \right) = \sum_{n\geq 1} E\left(\frac{\tilde{S}_n^4}{n^4} \right) \leq \sum_{n\geq 1} \frac{C}{n^2} < \infty.$$

Therefore,

$$\sum_{n\geq 1} \left(\frac{\tilde{S}_n}{n} \right)^4 < \infty, \text{ a.s.},$$

which implies that $\frac{\tilde{S}_n}{n} \to 0$, a.s., and therefore that

$$\frac{S_n}{n} \to \mu, \text{ a.s.}$$

□

In order to understand the proof a bit better, it's worth thinking about what would fail in it if the X_i were L^2 or L^3.

7.2 Tail σ-algebras; Zero-One Law

As the Borel-Cantelli lemmas show, there are only two possible values for $P(\limsup A_n)$ if $\{A_n\}_{n\geq 1}$ is a sequence of independent events. It turns out that the lim sup is just a particular case of a larger family of events for which the probability can be only 0 or 1.

Recall that if $\{Y_n\}_{n\geq 1}$ is a collection of random variables, then $\sigma(\{Y_n\}_{n\geq 1})$ is the smallest sigma-algebra containing $Y_n^{-1}(B)$ for every $n \geq 1$, every $B \in \mathcal{R}$.

Definition 7.1. For a sequence $\{X_n\}_{n\geq 1}$ of random variables, consider $\mathcal{T}_n := \sigma(X_{n+1}, X_{n+2}, \ldots)$.

$$\mathcal{T} := \bigcap_{n\geq 1} \mathcal{T}_n.$$

is called the *tail σ-algebra of the sequence* $\{X_n\}_{n\geq 1}$. An event $E \in T$ is called a *tail event*

Example 7.1. The event

$$\left\{\sum_{n=1}^{\infty}\frac{X_n}{n}\text{ converges}\right\}$$

is a tail-event, since for every k,

$$\left\{\sum_{n=1}^{\infty}\frac{X_n}{n}\text{ converges}\right\}=\left\{\sum_{n=k}^{\infty}\frac{X_n}{n}\text{ converges}\right\}\in\mathcal{T}_k,$$

implying that

$$\left\{\sum_{n=1}^{\infty}\frac{X_n}{n}\text{ converges}\right\}\in\bigcap_{k\geq 1}\mathcal{T}_k=\mathcal{T}.$$

Example 7.2 (Homework problem)**.** Of the following events, the first 2 are tail-events while the last two are not:

1. $\{\limsup\frac{S_n}{n}<c\}$,
2. $\{\lim S_n \text{ exists}\}$,
3. $\{\forall n\geq 1, X_n=0\}$
4. $\{\lim S_n \text{ exists and is } <c\}$

We will now turn to showing that if T is a tail event for a sequence of independent random variables, the only possible values for $P(T)$ are 0 or 1. First a few lemmas:

Lemma 7.1. Suppose $\mathcal{I}$ is a π-system on Ω. If μ_1 and μ_2 are finite measures with $\mu_1(\Omega)=\mu_2(\Omega)<\infty$ on $(\Omega,\sigma(\mathcal{I}))$ that agree on $\mathcal{I}$, then they agree on $\sigma(\mathcal{I})$.

Proof. Let $\mathcal{L}=\{A\in\sigma(\mathcal{I}):\mu_1(A)=\mu_2(A)\}$. Then $\mathcal{L}$ is a λ-system. Indeed,

- $\Omega\in\mathcal{L}$.
- If $A,B\in\mathcal{L}, A\subseteq B$, then since the measures are finite,

  $$\mu_1(B\setminus A)=\mu_1(B)-\mu_1(A)=\mu_2(B)-\mu_2(A)=\mu_2(B\setminus A),$$

  so $B\setminus A\in\mathcal{L}$.
- If $A_n\in\mathcal{L}, A_n\uparrow A$, then $\mu_1(A)=\lim_{n\to\infty}\mu_1(A_n)=\lim_{n\to\infty}\mu_2(A_n)=\mu_2(A)$, since $A=\cup_{n\geq 1}A_n$. Therefore, $A\in\mathcal{L}$.

By hypothesis, $\mathcal{I}\subseteq\mathcal{L}$, so since $\mathcal{L}$ is a λ-system, Dynkin's $\pi-\lambda$ theorem implies that $\sigma(\mathcal{I})\subset\mathcal{L}$. So for every $A\in\sigma(\mathcal{I}),\mu_1(A)=\mu_2(A)$.

□

Lemma 7.2. Suppose $\mathcal{G} \subset \mathcal{F}$ and $\mathcal{H} \subset \mathcal{F}$ are sigma-algebras and $\mathcal{I}, \mathcal{J}$ are π-systems with $\mathcal{G} = \sigma(\mathcal{I}), \mathcal{H} = \sigma(\mathcal{J})$. Then $\mathcal{G}$ and $\mathcal{H}$ are independent iff $\mathcal{I}$ and $\mathcal{J}$ are independent.

Proof. Obviously, if $\mathcal{G}$ and $\mathcal{H}$ are independent, so are $\mathcal{I}$ and $\mathcal{J}$. Suppose $\mathcal{I}$ and $\mathcal{J}$ are independent. Let $I \in \mathcal{I}$. The measures μ_1 and μ_2 on $(\Omega, \mathcal{F})$ defined by $\mu_1(H) = P(I \cap H)$ and $\mu_2(H) = P(I)P(H)$ agree on $\mathcal{J}$ and have same total mass $P(I)$. Therefore, by Lemma 7.1, they agree on $\mathcal{H} = \sigma(\mathcal{J})$. Thus,

$$P(I \cap H) = P(I)P(H), \ \forall I \in \mathcal{I}, H \in \mathcal{H}.$$

We use the same argument one more time: For fixed $H \in \mathcal{H}$, the measures ν_1 and ν_2 on $(\Omega, \mathcal{F})$ defined by $\nu_1(G) = P(G \cap H)$ and $\nu_2(G) = P(G)P(H)$ agree on $\mathcal{I}$ and have same total mass $P(H)$. Therefore, they agree on $\mathcal{G} = \sigma(\mathcal{I})$. So

$$P(G \cap H) = P(G)P(H), \ \forall G \in \mathcal{G}, H \in \mathcal{H}.$$

□

Theorem 7.2 (Kolmogorov's Zero-One Law)**.** Suppose that $\{X_n\}_{n\geq 1}$ are independent random variables and $\mathcal{T}$ is the tail σ-algebra generated by $\{X_n\}_{n\geq 1}$. If $C \in \mathcal{T}$, then either $P(C) = 0$ or $P(C) = 1$.

Proof. Define $\mathcal{H}_n = \sigma(X_1, \ldots, X_n)$ and $\mathcal{H} = \sigma(\{X_n\}_{n\geq 1})$, so that $\mathcal{H}_n$ and $\mathcal{T}_n$ are independent σ-algebras. Indeed, they are generated by the independent π-systems

$$\{\omega : X_i(\omega) \leq x_i : 1 \leq i \leq n\}, x_i \in \mathbb{R} \cup \infty$$

and

$$\{\omega : X_j(\omega) \leq x_j : n+1 \leq j \leq n+r\}, r \in \mathbb{N}, x_j \in \mathbb{R} \cup \infty,$$

respectively. So by Lemma 7.2, $\mathcal{H}_n$ and $\mathcal{T}_n$ are independent.

Now since $\mathcal{T} \subset \mathcal{T}_n$, it is clear that $\mathcal{H}_n$ and $\mathcal{T}$ are independent.

Since $\mathcal{H}_n \subset \mathcal{H}_{n+1}$ for all $n \geq 1$, $\cup_n \mathcal{H}_n$ is a π-system which generates $\mathcal{H}$. Since $\cup_n \mathcal{H}_n$ and $\mathcal{T}$ are independent, Lemma 7.2 implies that $\mathcal{T}$ and $\mathcal{H}$ are.

Since $\mathcal{T} \subset \mathcal{H}, \mathcal{T}$ is independent of $\mathcal{T}$. Therefore, if $F \in \mathcal{T}$, $P(F) = P(F \cap F) = P(F)P(F)$. So $P(F) = 1$ or $P(F) = 0$.

□

7.3 Random Series

We now focus on random series, i.e.,

$$S_n = \sum_{k=1}^{n} X_k,$$

where $\{X_n\}_{n\geq 1}$ is a sequence of independent (not necessarily identically distributed) random variables. The $0-1$ law (which applies, as we saw in Lecture 11) implies that $P(\{S(n)/n\}$ converges$) = 0$ or 1. We will now look for criteria to determine when this probability is 0 and when it is 1.

An important tool in deriving criteria for convergence of random series is an extension of Chebyshev's inequality. Recall that according to Chebyshev, if S_n is a random variable with $E[S_n] = 0$ and $E[S_n^2] < \infty$, then for every $a > 0$,

$$P(|S_n| \geq a) \leq \frac{E[S_n^2]}{a^2}.$$

It turns out that if we think of S_n as a random series, then the same inequality holds for the maximum of the partial sums up to n:

Lemma 7.3. (Kolmogorov's Maximal Inequality) Let $X_1, \ldots, X_n$ be independent random variables with $E[X_i] = 0$ and $E[X_i^2] < \infty$ for all $i \leq n$. Then for every $a > 0$,

$$P(\max_{1\leq k\leq n} |S_k| \geq a) \leq \frac{E[S_n^2]}{a^2}.$$

Proof. Define $A = \{\max_{1\leq k\leq n} |S_k| \geq a\}$ and for $1 \leq k \leq n$,

$$A_k = \{|S_i| < a, i = 1, \ldots, k-1, |S_k| \geq a\}.$$

Then

$$A = \bigsqcup_{1\leq k\leq n} A_k,$$

where $\sqcup$ denotes the disjoint union. Let $R_{n,k} = \sum_{i=k+1}^{n} X_i$. Then, since $R_{n,k}$ and $S_k \mathbb{1}_{A_k}$ are independent,

$$E[S_n^2 \mathbb{1}_{A_k}] = E[(S_k + R_{n,k})^2 \mathbb{1}_{A_k}] = E[S_k^2 \mathbb{1}_{A_k}] + E[R_{n,k}^2 \mathbb{1}_{A_k}],$$

so $E[S_n^2 \mathbb{1}_{A_k}] \geq E[S_k^2 \mathbb{1}_{A_k}]$. This gives

$$E[S_n^2] \geq E[S_n^2 \mathbb{1}_A] = \sum_{k=1}^{n} E[S_n^2 \mathbb{1}_{A_k}] \geq \sum_{k=1}^{n} E[S_k^2 \mathbb{1}_{A_k}] \geq a^2 P(A).$$

□

Note 7.1. Of course, Chebyshev's inequality is just a particular case of Kolmogorov's Maximal Inequality.

Lemma 7.4. (Cauchy Criterion for almost sure convergence) A sequence $\{X_n\}_{n\geq 1}$of random variables converges almost surely if and only if

$$P(\sup_{k\geq 0} |X_{n+k} - X_n| \geq \epsilon) \overset{n\to\infty}{\to} 0$$

Proof. Homework 3. □

Theorem 7.3. (Kolmogorov-Khinchine) Let $\{X_n\}_{n\geq 0}$ be a sequence of random variables with $E[X_n] = 0$. Then if $\sum_{n\geq 1} E[X_n^2] < \infty, \sum_{n\geq 1} X_n$ converges almost surely.

Proof. Lemma 7.3 implies that

$$P(\sup_{k\geq 1} |S_{n+k} - S_k| \geq \epsilon) = \lim_{N\to\infty} P(\max_{1\leq k\leq N} |S_{n+k} - S_n| \geq \epsilon) \leq \lim_{N\to\infty} \frac{\sum_{k=n}^{n+N} E[X_k^2]}{\epsilon^2} = \frac{\sum_{k=n}^{\infty} E[X_k^2]}{\epsilon^2}.$$

By hypothesis, this last term goes to 0 as $n \to \infty$, so Lemma 7.4 implies that $\{S_n\}$ converges almost surely. □

7.4 Strong Law of Large Numbers

In order to prove the strong version of the strong law of large numbers, we will need, in addition to the results from the previous subsection, 3 lemmas. The first two are purely analytic and involve no probability. The third is a general fact about positive random variables.

Lemma 7.5. (Toeplitz) Consider a sequence of real numbers $\{a_n\}_{n\geq 1}$ with $a_1 > 0, a_n \geq 1 \ \forall \ n \geq 1$, and $\sum_{i=1}^{n} a_i \uparrow \infty$ as $n \to \infty$. Then if $\{x_n\}_{n\geq 1}$ is a sequence of real numbers with $\lim_{n\to\infty} x_n = x$,

$$\frac{1}{\sum_{i=1}^{n} a_i} \sum_{i=1}^{n} a_i x_i \to x, \text{ as } n \to \infty.$$

In particular,

$$\frac{1}{n} \sum_{i=1}^{n} x_i \to x, \text{ as } n \to \infty.$$

Proof. Let's write $b_n = \sum_{i=1}^{n} a_i$. Define $n_0 \in \mathbb{N}$ to be such that $|x_i - x| < \epsilon/2$ for all $i \geq n_0$ and $n_1 > n_0$ to satisfy

$$\frac{1}{b_{n_1}} \sum_{i=1}^{n_0} a_i |x_i - x| < \epsilon/2.$$

Then, if $n \geq n_1$,

$$\begin{aligned}\left|\frac{1}{b_n}\sum_{i=1}^{n} a_i x_i - x\right| &\leq \frac{1}{b_n}\sum_{i=1}^{n} a_i|x_i - x| = \frac{1}{b_n}\left(\sum_{i=1}^{n_0} a_i|x_i - x| + \sum_{i=n_0+1}^{n} a_i|x_i - x|\right) \\ &< \frac{1}{b_{n_1}}\sum_{i=1}^{n_0} a_i|x_i - x| + \frac{\epsilon}{2}\frac{1}{b_n}\sum_{i=n_0+1}^{n} a_i < \frac{\epsilon}{2} + \frac{\epsilon}{2}\frac{b_n - b_{n_0+1}}{b_n} \leq \epsilon.\end{aligned}$$

□

Lemma 7.6. (Kronecker) Let $\{b_n\}_{n\geq 1}$ be a sequence of real numbers satisfying $b_n > 0$ and $b_n \uparrow \infty$ as $n \to \infty$. If $\{x_n\}_{n\geq 1}$ be a sequence of real numbers such that $\sum_{n\geq 1} x_n$ converges, then

$$\frac{1}{b_n}\sum_{i=1}^{n} b_i x_i \to 0, \text{ as } n \to \infty.$$

In particular, if $\sum_{n\geq 1} x_n/n$ converges, then

$$\frac{1}{n}\sum_{i=1}^{n} x_i \to 0, \text{ as } n \to \infty.$$

Proof. Let $b_0 = s_0 = 0$, $s_n = \sum_{i=1}^{n} x_i$ and $x = \lim_{n\to\infty} s_n$. Then $\sum_{i=1}^{n} b_i x_i = \sum_{i=1}^{n} b_i(s_i - s_{i-1}) = b_n s_n + \sum_{i=1}^{n}(b_{i-1} - b_i)s_{i-1}$. Therefore,

$$\frac{1}{b_n}\sum_{i=1}^{n} b_i x_i = s_n - \frac{1}{b_n}\sum_{i=1}^{n}(b_i - b_{i-1})s_{i-1} \to 0,$$

since as $n \to \infty$ $s_n \to x$ and by Toeplitz's lemma, $\frac{1}{b_n}\sum_{i=1}^{n}(b_i - b_{i-1})s_{i-1} \to x$. □

Lemma 7.7. Suppose $X \geq 0$ is a random variable. Then

$$\sum_{n\geq 1} P(X \geq n) \leq E[X] \leq 1 + \sum_{n\geq 1} P(X \geq n).$$

Proof. Homework 3. □

Note 7.2. If X is integer-valued, the first inequality in this lemma becomes an equality.

Theorem 7.4. (Kolmogorov's strong law of large numbers) Let $\{X_i\}_{i\geq 1}$ be a sequence of i.i.d. random variables with $E[|X_1|] < \infty$. Then

$$\frac{S_n}{n} \xrightarrow{a.s.} E[X_1], \text{ as } n \to \infty.$$

Proof. We will assume that $E[X_1] = 0$. By Lemma 7.7 and Borel-Cantelli,

$$E[|X_1|] < \infty \iff \sum_{n\geq 1} P(|X_1| \geq n) < \infty \iff \sum_{n\geq 1} P(|X_n| \geq n) < \infty \iff P(|X_n| \geq n, \text{ i.o.}) = 0. \tag{2}$$

Define $\bar{X}_n = X_n \mathbb{1}_{|X_n|<n}$. Then $S_n/n \to 0$, a.s., if and only if

$$1/n \sum_{k=1}^{n} \bar{X}_k \to 0, \text{ a.s.}, \tag{3}$$

since by (2), for every ω, all but a finite number of terms in the two sums are the same. Since

$$E[\bar{X}_n] = E[X_1 \mathbb{1}_{\{|X_1|<n\}}] \to E[X_1] = 0,$$

Toeplitz's lemma implies that

$$\frac{1}{n} \sum_{k=1}^{n} E[\bar{X}_n] \to 0, \text{ as } n \to \infty. \tag{4}$$

Therefore, by (3) and (4), $S_n/n \to 0$, a.s. if and only if $1/n \sum_{k=1}^{n} (\bar{X}_k - E[\bar{X}_k]) \to 0$, a.s. By Kronecker's lemma, we'll be done if

$$\sum_{k=1}^{n} \left(\frac{\bar{X}_k - E[\bar{X}_k]}{k} \right) \to 0, \text{ a.s.},$$

and by Kolmogorov-Khinchine, if

$$\sum_{k\geq 1} \frac{Var(\bar{X}_k)}{k^2} \text{ converges.}$$

$$\begin{aligned}
\sum_{n\geq 1} \frac{Var(\bar{X}_n)}{n^2} &\leq \sum_{n\geq 1} \frac{E[\bar{X}_n^2]}{n^2} = \sum_{n\geq 1} E[\frac{1}{n^2}(X_n \mathbb{1}_{\{|X_n|<n\}})^2] = \sum_{n\geq 1} \frac{1}{n^2} E[X_1^2 \mathbb{1}_{\{|X_1|<n\}}] \\
&= \sum_{n\geq 1} \sum_{k=1}^{n} E[X_1^2 \mathbb{1}_{\{|X_1|\in[k-1,k)\}}] = \sum_{k=1}^{\infty} E[X_1^2 \mathbb{1}_{\{|X_1|\in[k-1,k)\}}] \sum_{n=k}^{\infty} \frac{1}{n^2} \\
&\leq C \sum_{k=1}^{\infty} \frac{1}{k} E[X_1^2 \mathbb{1}_{\{|X_1|\in[k-1,k)\}}] \leq C \sum_{k=1}^{\infty} E[|X_1| \mathbb{1}_{\{|X_1|\in[k-1,k)\}}] = CE[|X_1|] < \infty.
\end{aligned}$$

□

It turns out that this version of the strong law of large numbers is optimal, as the following theorem shows:

Theorem 7.5. Suppose $\{X_i\}_{i\geq 1}$ is a sequence of i.i.d. random variables such that

$$\frac{S_n}{n} \overset{a.s.}{\to} C < \infty, \text{ as } n \to \infty.$$

Then $X_1 \in L^1$ and $E[X_1] = C$.

Proof. We can write

$$\frac{X_n}{n} = \frac{S_n - S_{n-1}}{n} = \frac{S_n}{n} - \frac{n-1}{n}\frac{S_{n-1}}{n-1} \to 0 \text{ a.s.}$$

So $P(|X_n| > n, \text{ i.o.}) = 0$ and by Borel-Cantelli,

$$\sum_{n\geq 1} P(|X_1| > n) = \sum_{n\geq 1} P(|X_n| > n) < \infty.$$

Lemma 7.7 now implies that $E[|X_1|] < \infty$ and Theorem 7.4 implies that $E[X_1] = C$. □

Lecture #8: Law of Large Numbers; Law of the Iterated Logarithm

8.1 Another Law of Large Numbers

There is another version of the strong law of large numbers which doesn't require the random variables to be identically distributed, but needs two moments:

Theorem 8.1. Let $\{X_i\}_{i\geq 1}$ be a sequence of independent L^2 random variables. If there is a sequence of positive real numbers $\{b_n\}$ such that $b_n \uparrow \infty$ and

$$\sum_{n\geq 1} \frac{Var(X_n)}{b_n^2} < \infty,$$

then

$$\frac{S_n - E[S_n]}{b_n} \to 0, \text{ a.s.}$$

In particular, if

$$\sum_{n\geq 1} \frac{Var(X_n)}{n^2} < \infty,$$

then

$$\frac{S_n - E[S_n]}{n} \to 0, \text{ a.s.}$$

Proof. The idea is to write $\frac{S_n - E[S_n]}{b_n}$ as a sum of independent random variables and to use Kroncker's lemma. Clearly,

$$\frac{S_n - E[S_n]}{b_n} = \frac{1}{b_n}\sum_{k=1}^{n} b_k \frac{X_k - E[X_k]}{b_k}.$$

Now by Kolmogorov-Khinchine, since

$$\sum_{n\geq 1} E\left[\left(\frac{X_n - E[X_n]}{b_n}\right)^2\right] \text{ converges },$$

$$\sum_{n\geq 1} \frac{X_n - E[X_n]}{b_n} \text{ converges almost surely,}$$

so that we can conclude the proof using Kronecker's lemma. □

8.2 Law of the Iterated Logarithm

Suppose that $\{X_i\}$ is a sequence of independent mean 0, variance 1 random variables. Then one can show that

$$\limsup \frac{S_n}{n^{1/2-\epsilon}} = \infty \text{ and } \liminf \frac{S_n}{\sqrt{n}} = -\infty, \text{ a.s.} \tag{5}$$

Indeed, let $N > 0$ and $E_N = \{\limsup \frac{S_n}{n^{1/2-\epsilon}} \leq N\}$ Then for every $\delta > 0$,

$$\begin{aligned} P(E_N) &\leq P\left(\frac{S_n}{n^{1/2}-\epsilon} \leq N+\epsilon, \text{ eventually}\right) = P\left(\liminf\{\frac{S_n}{n^{1/2}-\epsilon} \leq N+\epsilon\}\right) \\ &\overset{\text{Fatou}}{\leq} \liminf P\left(\frac{S_n}{n^{1/2}-\epsilon} \leq N+\epsilon\right) < 1, \end{aligned}$$

where the last inequality is due to the central limit theorem (which we'll see soon, but which you've certainly seen in an undergraduate class).

Now since E_N is a tail event (see Example 2.5.2 in Durrett), this means that there are only two possibilities: $P(E_N) = 0$ or $P(E_N) = 1$. But we just showed that $P(E_N) < 1$. So $P(E_N) = 0$. Since this holds for arbitrary N, (5) follows. Note that the choice of $n^{1/2-\epsilon}$ above is fairly arbitrary. The argument works if we replace $n^{1/2} - \epsilon$ by any function $f(n)$ such that $\frac{f(n)}{n^{1/2}} \to 0$ as $n \to \infty$.

On the other hand, by Theorem 8.1, $\frac{S_n}{\sqrt{n}\log n} \to 0$, a.s., since $\sum_{n\geq 1} \frac{\mathrm{Var(X_n)}}{(\sqrt{n}\log n)^2} = \sum_{n\geq 1} \frac{1}{n\log^2 n} < \infty$.

This suggests that we might be able to find a non-trivial almost sure lim sup for the sequence S_n. Can we find a function $\psi(n)$ such that $\limsup \frac{S(n)}{\psi(n)} = 1$, a.s.? It turns out that we can.

First, we prove two ancillary results. The first, called the reflexion principle, is often needed, when dealing with stochastic processes:

Definition 8.1. A random variable X is said to be *symmetric* if for all $B \in \mathcal{R}, P(X \in B) = P(-X \in B)$.

Lemma 8.1. Let $\{X_i\}_{1\leq i\leq n}$ be independent symmetric random variables Then for every $a \in \mathbb{R}$,

$$P(\max_{1\leq k\leq n} S_k > a) \leq 2P(S_n > a).$$

Proof. The idea is that if $S_k > a$ for some $k \leq n$, then S_n has a probability of at least $1/2$ of being greater than a, since $S_n - S_k$ is symmetric. For $k \geq 1$, define

$$A_k = \{S_j \leq a \text{ for } 1 \leq j \leq k-1, S_k > a\},$$

the event that S is greater than a for the first time at time k. Then

$$P(\max_{1\leq k\leq n} S_k > a) = \sum_{k=1}^{n} P(A_k)$$

and

$$\begin{aligned} P(S_n > a) &= \sum_{k=1}^{n} P(S_n > a; A_k) \geq \sum_{k=1}^{n} P(S_n - S_k \geq 0; A_k) \\ &= \sum_{k=1}^{n} P(S_n - S_k \geq 0) P(A_k) \geq \frac{1}{2} \sum_{k=1}^{n} P(A_k) = \frac{1}{2} P(\max_{1 \leq k \leq n} S_k > a), \end{aligned}$$

where the second equality follows from the independence of $\{S_n - S_k \geq 0\}$ and A_k. □

Lemma 8.2. Suppose $S_n \sim N(0, \sigma_n^2)$ with $\sigma_n^2 \uparrow \infty$, as $n \to \infty$, and suppose $\{r_n\}_{n \geq 1}$ satisfies

$$\lim_{n \to \infty} \frac{r_n}{\sigma_n} \to \infty.$$

Then

$$P(S_n > r_n) \sim \frac{\sigma_n}{\sqrt{2\pi} r_n} \exp\left\{ -\frac{r_n^2}{2\sigma_n^2} \right\}.$$

Proof. First note that $S_n/\sigma_n \sim N(0,1)$. So

$$P(S_n > r_n) = P(\frac{S_n}{\sigma_n} > \frac{r_n}{\sigma_n}) \sim \frac{\sigma_n}{\sqrt{2\pi} r_n} \exp\left\{ -\frac{r_n^2}{2\sigma_n^2} \right\},$$

since we know from Lemma 2.1 that if Z is a standard normal random variable,

$$P(Z > x) \overset{x \to \infty}{\sim} \frac{1}{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

□

Theorem 8.2. (Law of the iterated logarithm) Let $\{X_i\}_{i \geq 1}$ be a sequence of i.i.d. random variables. with $E[X_1] = 0$ and $E[X_1^2] = \sigma^2 < \infty$. Then

$$P\left(\limsup_{n \to \infty} \frac{S_n}{\sqrt{2\sigma^2 n \log\log n}} = 1 \right) = 1$$

and

$$P\left(\liminf_{n \to \infty} \frac{S_n}{\sqrt{2\sigma^2 n \log\log n}} = -1 \right) = 1.$$

Note 8.1. Another formulation of the theorem is the following: For every $\epsilon > 0$,

$$P(S_n \geq (1-\epsilon)\sqrt{2\sigma^2 n \log\log n}, \text{ i.o.}) = P(S_n \leq -(1-\epsilon)\sqrt{2\sigma^2 n \log\log n}, \text{ i.o.}) = 1,$$

$$P(S_n \geq (1+\epsilon)\sqrt{2\sigma^2 n \log\log n}, \text{ i.o.}) = P(S_n \leq -(1+\epsilon)\sqrt{2\sigma^2 n \log\log n}, \text{ i.o.}) = 0.$$

Proof. We will prove the theorem in the case where $X_i \sim N(0,1)$. Let $\psi(n) = \sqrt{2n \log\log n}$. We will show that

$$P\left(\limsup_{n\to\infty} \frac{S_n}{\psi(n)} \leq 1\right) = 1 \tag{6}$$

and

$$P\left(\limsup_{n\to\infty} \frac{S_n}{\psi(n)} \geq 1\right) = 1. \tag{7}$$

Equation (6) will be somewhat more straightforward to prove than (7).

We begin by noting that

$$\begin{aligned}\left\{\limsup_{n\to\infty} \frac{S_n}{\psi(n)} \leq 1\right\} &= \left\{\lim_{n\to\infty} \sup_{m\geq n} \frac{S_m}{\psi(m)} \leq 1\right\} \\ &= \{\forall \epsilon > 0 \exists m_1(\epsilon) s.t. \forall m \geq m_1(\epsilon), S_m \leq (1+\epsilon)\psi(m)\}\end{aligned}$$

and

$$\left\{\limsup_{n\to\infty} \frac{S_n}{\psi(n)} \geq 1\right\} = \left\{\lim_{n\to\infty} \sup_{m\geq n} \frac{S_m}{\psi(m)} \geq 1\right\} = \{\forall \epsilon > 0, S_m \geq (1-\epsilon)\psi(m), \text{ i.o.},\}.$$

Therefore, we can show (6) by proving for every $\epsilon > 0$ that $P(S_m > (1+\epsilon)\psi(m), \text{ i.o.}) = 0$. We will now decompose $\mathbb{R}$ into intervals of exponentially increasing length and estimate the probability that over each interval $S_n > (1+\epsilon)\psi(n)$ for some n. These probabilities will be summable.

Let $\lambda = 1 + \epsilon$ and define $A_k = \{S_m > \lambda\psi(m) \text{ for some } m \in (\lambda^k, \lambda^{k+1}]\}$. Then we can use the fact that $S_n \sim N(0,n)$ and Lemmas 8.1 and 8.2 to see that

$$\begin{aligned}P(A_k) &= P(S_m > \lambda\psi(m), \text{ some } m \in (\lambda^k, \lambda^{k+1}]) \leq P(S_m > \lambda\psi(\lambda^k), \text{ some } m \leq \lambda^{k+1}) \\ &\leq 2P(S_{\lfloor\lambda^{k+1}\rfloor} > \lambda\psi(\lambda^k)) \leq C\frac{\sqrt{\lambda^k}}{\lambda\psi(\lambda^k)} \exp\left\{-\frac{\lambda^2}{2}\left(\frac{\psi(\lambda^k)}{\sqrt{\lfloor\lambda^k\rfloor}}\right)^2\right\} \\ &\leq C\exp\left\{-\lambda^2 \ln\ln\lambda^k\right\} \leq C\exp\left\{-\lambda\ln(k\ln\lambda)\right\} \leq C\exp\left\{-\lambda\ln k\right\} = Ck^{-\lambda}.\end{aligned}$$

Here, we assumed k is large enough that all quantities are defined. Since $\lambda > 1, \sum_{k\geq 1} P(A_k) < \infty$, implying via Borel-Cantelli that $P(A_k, \text{ i.o.}) = 0$.

We now turn to the proof of (7). The key difficulty is that the S_n are not independent, making the direct use of the second Borel-Cantelli Lemma impossible. The trick is to find independent random variables to which B-C 2 can be applied. These independent random variables will be increments of S_n over certain appropriately chosen intervals.

We let $\lambda = 1 - \epsilon$ and will show that

$$P(S_m \geq \lambda\psi(m), \text{ i.o.}) = 1.$$

From the work above and symmetry, we know that $P(S_m \leq -2\psi(m), \text{ i.o.}) = 0$ (the "2" is arbitrary; anything strictly greater than 1 works too). So since if $N \in \mathbb{N} \setminus \{1\}$,

$$\{S_{N^k} - S_{N^{k-1}} > \lambda\psi(N^k) + 2\psi(N^{k-1}), \text{ i.o.}\} \subseteq \{S_{N^k} > \lambda\psi(N^k), \text{ i.o.}\} \cup \{S_{N^{k-1}} \leq -2\psi(N^{k-1}), \text{ i.o.}\},$$

we have

$$P(S_{N^k} - S_{N^{k-1}} > \lambda\psi(N^k) + 2\psi(N^{k-1}),\ \text{i.o.}) \leq P(S_{N^k} > \lambda\psi(N^k),\ \text{i.o.}).$$

The expression $\lambda\psi(N^k)$ is not so convenient to work with. However,

$$\begin{aligned}
\lambda\psi(N^k) + 2\psi(N^{k-1}) &= \lambda\sqrt{2N^k \ln\ln N^k} + 2\sqrt{2N^{k-1}\ln\ln N^{k-1}} < \left(\lambda + \frac{2}{\sqrt{N}}\right)\sqrt{2N^k}\sqrt{\ln\ln N^k} \\
&= \left(\left(\lambda + \frac{2}{\sqrt{N}}\right)/\sqrt{1-\frac{1}{N}}\right)\sqrt{2(N^k - N^{k-1})}\sqrt{\ln\ln N^k} \\
&< \lambda'\sqrt{2(N^k - N^{k-1})\ln\ln N^k},
\end{aligned}$$

for some λ' with $\lambda' \in (\lambda, 1)$, for k large enough. So we'll be done if we can show that

$$P(S_{N^k} - S_{N^{k-1}} > \lambda'\sqrt{2(N^k - N^{k-1})\ln\ln N^k},\ \text{i.o.}) = 1. \tag{8}$$

By Lemma 8.2,

$$\begin{aligned}
P\left(S_{N^k} - S_{N^{k-1}} > \lambda'\sqrt{2(N^k - N^{k-1})\ln\ln N^k}\right) &\sim \frac{1}{\sqrt{2\pi\lambda'}\sqrt{2\ln\ln N^k}} e^{-\lambda'^2 \ln\ln N^k} \\
&\geq \frac{C}{\sqrt{\ln k}}(k\ln N)^{-\lambda'^2} \geq \frac{C}{k\ln k}.
\end{aligned}$$

Since $\displaystyle\sum_{k\geq 1}\frac{1}{k\ln k} = \infty$ and the increments $S_{N^k} - S_{N^{k-1}}$ are independent, the second Borel-Cantelli Lemma implies that (8) holds. □

Lecture #9: Convergence in Distribution, Weak Convergence; Characteristic Functions

Reference. Sections 3.2, 3.3

In order to derive one of the most important results of probability theory, the central limit theorem, we need to be comfortable with the notions of convergence in distribution and characteristic functions.

9.1 Convergence in Distribution & Weak Convergence

There is a mode of convergence of random variables which is radically different from those we've studied so far. It is *convergence in distribution*. Whereas the other modes we have seen require that the random variables are all defined on the same probability space, weak convergence concerns the laws of the random variables, so we don't need to assume that they are all defined on a common probability space.

Notation. We will denote by $\mathbb{C}(\mathbb{R})$ the space of bounded continuous functions $f : \mathbb{R} \to \mathbb{R}$.

Definition 9.1. A sequence $\{X_n\}_{n\geq 1}$ *converges in distribution* to the random variable X if

$$E[f(X_n)] \to E[f(X)], \text{ as } n \to \infty$$

for every $f \in \mathbb{C}(\mathbb{R})$. We write

$$X_n \xrightarrow{d} X.$$

Proposition 9.8. If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{d} X$.

Proof. Consider a continuous function f with $|f(x)| \leq C$ for some $C < \infty$. For every $N < \infty$, f is uniformly continuous on $[-N, N]$.

Fix $\epsilon > 0$ and let N be such that $P(|X| \geq N) < \frac{\epsilon}{6c}$. Choose $\delta > 0$ with $\delta < N$ such that if $|x| < N$ and $|x - y| \leq \delta$, then $|f(x) - f(y)| \leq \frac{\epsilon}{3}$. Also choose n_0 such that for all $n \geq n_0, P(|X_n - X| > \delta) < \frac{\epsilon}{6c}$.

$$\begin{aligned} E[|f(X_n) - f(X)|] \quad &\leq \quad E[|f(X_n) - f(X)|\mathbb{1}\{|X_n - X| \leq \delta; |X| \leq N\}] \\ &+ \quad E[|f(X_n) - f(X)|\mathbb{1}\{|X| > N\}] \\ &+ \quad E[|f(X_n) - f(X)|\mathbb{1}\{|X_n - X| > \delta\}] \\ &< \quad \frac{\epsilon}{3} + 2c\frac{\epsilon}{6c} + 2cP(|X_n - X| > \delta) \leq \epsilon. \end{aligned}$$

□

Recall that if $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$ is a random variable, then P^X, the law of X, is given by

$$P^X(B) = P\{X \in B\} \quad \text{for every } B \in \mathcal{B}$$

and defines a probability measure on $(\mathbb{R}, \mathcal{B})$. That is, if X is a random variable, then $(\mathbb{R}, \mathcal{B}, P^X)$ is a probability space.

Suppose that for each $n \geq 1$, $X_n : (\Omega_n, \mathcal{A}_n, P_n) \to (\mathbb{R}, \mathcal{B})$ and $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$ are random variables. The laws of each of these random variables induce probability measures on $(\mathbb{R}, \mathcal{B})$, say P^{X_n}, $n = 1, 2, 3, \ldots$, and P^X.

Definition 9.2. If $\{P_n\}_{n\geq 1}$ and P are probability measures, we say that P_n *converges weakly to* P if

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x) P_n(dx) = \int_{\mathbb{R}} f(x) P(dx)$$

for all $f \in \mathbb{C}(\mathbb{R})$. We write $P_n \Rightarrow P$.

Similarly, we say that the distribution functions F_n *converge weakly* to the distribution function F if

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x) dF_n(x) = \int_{\mathbb{R}} f(x) dF(x)$$

for every $f \in \mathbb{C}(\mathbb{R})$ and write $F_n \Rightarrow F$.

We say that X_n *converges weakly to* X or X_n *converges to* X *in law* if P^{X_n} converges weakly to P^X and write $X_n \Rightarrow X$.

Note 9.1. This definition is rather unintuitive. Fortunately, we will prove a number of characterizations of weak convergence which will be, in general, much easier to handle.

Weak convergence and convergence in distribution are the same thing:

Theorem 9.1. Let $\{X_n\}_{n\geq 1}$ be random variables with laws P^{X_n} and let X be a random variable with law P^X. Then $X_n \Rightarrow X$ if and only if $X_n \xrightarrow{d} X$

Proof. This follows immediately from the change of variables formula (see Lecture 3): If X_n has distribution P^{X_n} and X has distribution P^X, then

$$\int_{\mathbb{R}} f(x) P^{X_n}(dx) = \mathbb{E}(f(X_n)) \quad \text{and} \quad \int_{\mathbb{R}} f(x) P^X(dx) = \mathbb{E}(f(X)).$$

This equivalence establishes the theorem. □

Note 9.2. The reason for the name "convergence in law" should be clear. We are discussing convergence of the laws of random variables. The reason that it is also called convergence in distribution is the following. Since the law of a random variable is characterized by its distribution function, one could hope that convergence of the laws of the random variables is equivalent to convergence of the corresponding distribution functions. This is almost the case, as we will see.

Note 9.3. Since $f(x) = x$ is NOT bounded, we cannot infer whether or not $X_n \to X$ weakly from the convergence/non-convergence of $\mathbb{E}(X_n)$ to $\mathbb{E}(X)$.

Although the random variables X_n and X need not be defined on the same probability space in order for weak convergence to make sense, if they are defined on a common probability space, then we can talk about implication between types of convergence.

Note 9.4. Suppose that $\{X_n\}_{n\geq 1}$, and X are random variables defined on a common probability space $(\Omega, \mathcal{A}, P)$. The facts about convergence are:

$$\begin{array}{ccc} X_n \to X \text{ almost surely} & & \\ \Downarrow & & \\ X_n \to X \text{ in probability} & \Rightarrow & X_n \Rightarrow X \\ \Uparrow & & \\ X_n \to X \text{ in } L^p & & \end{array}$$

with no other implications holding in general.

The following example shows just how weak convergence in distribution really is.

Example 9.1. Suppose that $\Omega = \{a, b\}$, $\mathcal{F} = 2^\Omega$, and $P\{a\} = P\{b\} = 1/2$. Define the random variables Y and Z by setting

$$Y(a) = 1, \quad Y(b) = 0, \quad \text{and} \quad Z = 1 - Y.$$

Then

$$Y \neq Z \text{ almost surely} \quad \text{but} \quad P^Y = P^Z.$$

If n is odd, let $X_n = Y$, and if n is even, let $X_n = Z$. Since $P^Y = P^Z$, it is clear that $P^{X_n} \to P^Y$ meaning that $X_n \to Y$ in distribution. (In fact, $P^{X_n} = P^Y$ meaning that $X_n = Y$ in distribution for every n.) However, if $\epsilon > 0$ and n is even, then

$$P\{|X_n - Y| > \epsilon\} = P\{|Z - Y| > \epsilon\} = P\{Z = 1, Y = 0\} + P\{Z = 0, Y = 1\} = 1.$$

Thus, it is not possible for X_n to converge in probability to Y. Of course, this example should make sense. By construction, since $Z = 1 - Y$, the observed sequence of X_n will alternate between 1 and 0 or 0 and 1 (depending on whether a or b is first observed).

One would hope that if $P^{X_n} \to P^X$, then F_{X_n} would converge to F_X. This is not quite true as the following example shows:

Example 9.2. Suppose X has distribution F. Then if $Y_n = X + \frac{1}{n}, Y_n \xrightarrow{a.s.} X$, implying that $Y_n \Rightarrow X$. However,

$$F_{Y_n}(x) = P(X + \frac{1}{n} \leq x) = P(X \leq x - \frac{1}{n}) = F(x - \frac{1}{n}),$$

so $\lim_{n\to\infty} F_{Y_n}(x) = \lim_{y\uparrow x} F_Y(y)$. So F_{Y_n} converges to F_X only at points of continuity of F_X.

Definition 9.3. A sequence of distribution functions $\{F_n\}$ converges *in general* to the distribution F if $F_n(x) \to F(x)$ for all points x of continuity of F.

Theorem 9.2.

$$X_n \Rightarrow X \iff F_{X_n} \to F_X \text{ in general.}$$

Example 9.3. Let $X_1, X_2, \ldots$ be independent with $P(X_i = 1) = P(X_i = -1) = \frac{1}{2}$ and let X be such that $P(X = 0) = 1$. Then

$$\forall \epsilon > 0, P(\left|\frac{S_n}{n}\right| > \epsilon) \to 0 \iff \forall x > 0, F_{S_n/n}(x) \to 1 \text{ and } \forall x < 0, F_{S_n/n}(x) \to 0$$

$$\iff F_{S_n/n} \to F_X \text{ in general} \iff \frac{S_n}{n} \Rightarrow X.$$

So WLLN is equivalent to convergence of $\frac{S_n}{n}$ to X in law.

9.2 Characteristic Functions

Definition 9.4. If X is a random variable, then the *characteristic function* of X is given by

$$\phi_X(t) = E[e^{itX}] = E[\cos(tX)] + iE[\sin(tX)], \quad t \in \mathbb{R}.$$

Note 9.5. The following is a list of basic properties of characteristic functions. Their proofs all hold in 2 lines, so we omit them (see Durrett, pp. 90-91 if you are unsure how to derive them).

- $$\phi_X(0) = 1.$$
- For all $t \in \mathbb{R}$,
$$\phi_X(-t) = \overline{\phi_X(t)}.$$
- For all $t \in \mathbb{R}$,
$$|\phi_X(t)| \leq 1.$$
- For all $t \in \mathbb{R}$,
$$\phi_{aX+b}(t) = e^{itb}\phi_X(at).$$
- For all $t \in \mathbb{R}$,
$$\phi_{-X}(t) = \overline{\phi_X(t)}.$$
- ϕ_X is uniformly continuous.

Theorem 9.3. If $X_1, X_2, \ldots, X_n$ are independent random variables and

$$S_n = X_1 + X_2 + \cdots + X_n,$$

then

$$\phi_{S_n}(t) = \prod_{i=1}^{n} \phi_{X_i}(t).$$

In particular, if $X_1, X_2, \ldots, X_n$ are identically distributed, then

$$\phi_{S_n}(t) = [\phi_{X_1}(t)]^n .$$

Proof. By definition,

$$\phi_{S_n}(t) = E[e^{itS_n}] = E[e^{it(X_1+\cdots+X_n)}] = E[e^{itX_1} \cdots e^{itX_n}] = E[e^{itX_1}] \cdots E[e^{itX_n}] = \prod_{i=1}^{n} \phi_{X_i}(t)$$

where the second-to-last equality follows from the fact that X_i are independent. □

9.3 Moments and Derivatives

Theorem 9.4. If X has a moment of order n, then

$$\left| \phi_X(t) - \sum_{j=0}^{n} E\left[\frac{(itX)^j}{j!}\right] \right| \leq E\left[\min\left\{\frac{|tX|^{n+1}}{(n+1)!}, \frac{2|tX|^n}{n!}\right\}\right].$$

Note 9.6. In particular, for any t with $\lim_{n\to\infty} \frac{|t|^n E[|X|^n]}{n!} = 0, \phi(t) = \sum_{k\geq 0} \frac{(it)^k}{k!} E[X^k]$.

Proof. If we let $A_n = \int_0^x (x-s)^n e^{is}\, ds$, then $A_0 = \frac{e^{ix}-1}{i}$ and by integration by parts, we get for $n \geq 0, A_n = \frac{x^{n+1}}{n+1} + \frac{i}{n+1} A_{n+1}$, implying that for $n \geq 1, A_n = ix^n - inA_{n-1}$. By induction, we get

$$e^{ix} = \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^{n+1}}{n!} A_n, \tag{9}$$

so

$$\begin{aligned} e^{ix} &= \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^{n+1}}{n!}(ix^n - inA_{n-1}) = \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^{n+1}}{n!}(ix^n - in\int_0^x (x-s)^{n-1} e^{is}\, ds) \\ &= \sum_{k=0}^{n} \frac{(ix)^k}{k!} - \frac{(ix)^n}{n!} + \frac{ni^n}{n!} \int_0^x (x-s)^{n-1} e^{is}\, ds \\ &= \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^n}{(n-1)!} \left(\int_0^x (x-s)^{n-1} (e^{is} - 1)\, ds \right), \end{aligned} \tag{10}$$

where we used the fact that $x^n = n\int_0^x (x-s)^{n-1}\,ds$. Therefore, by (9) and (10),

$$e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!} = \frac{i^{n+1}}{n!}A_n = \frac{i^n}{(n-1)!}\int_0^x (x-s)^{n-1}(e^{is}-1)\,ds.$$

Now

$$\left|\frac{i^{n+1}}{n!}A_n\right| \le \frac{1}{n!}\int_0^x |x-s|^n\,ds \le \frac{|x|^{n+1}}{(n+1)!}$$

and

$$\left|\frac{i^n}{(n-1)!}\int_0^x (x-s)^{n-1}(e^{is}-1)\,ds\right| \le \frac{1}{(n-1)!}2\int_0^x |x-s|^{n-1}\,ds \le \frac{2|x|^n}{n!}.$$

So

$$\left|e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!}\right| \le \min\left\{\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!}\right\}.$$

The theorem follows. □

9.4 The Inversion Theorem

Definition 9.5. For $T \ge 0$, we define

$$S(T) := \int_0^T \frac{\sin x}{x}\,dx.$$

Lemma 9.1. 1. $\lim_{T\to\infty} S(T) = \frac{\pi}{2}$.

2. $\int_0^T \frac{\sin(\theta t)}{t}\,dt = sgn(\theta)S(T|\theta|)$

Proof. 1. This is just a creative application of Fubini's theorem (which is OK, since on $[0,T]\times[0,\infty), |\sin x e^{-ux}|$ is integrable) :

$$\begin{aligned}
\int_0^T \frac{\sin x}{x}\,dx &= \int_0^T \sin x \int_0^\infty e^{-ux}\,du\,dx = \int_0^\infty \int_0^T \sin x e^{-ux}\,dx\,du \\
&= \int_0^\infty \frac{1}{1+u^2}\left(1 - e^{-uT}(u\sin T + \cos T)\right)\,du \\
&= \int_0^\infty \frac{1}{1+u^2}\,du - \int_0^\infty \frac{e^{-uT}(u\sin T + \cos T)}{1+u^2}\,du. \\
&= \frac{\pi}{2} - T\int_0^\infty \frac{e^{-s}(s\sin T/T + \cos T)}{s^2+T^2}\,ds.
\end{aligned}$$

But

$$\left|T\int_0^\infty \frac{e^{-s}(s\sin T/T + \cos T)}{s^2+T^2}\,ds\right| \le \frac{T}{T^2}\int_0^\infty Cse^{-s}\,ds \to 0, \text{ as } T\to\infty.$$

2. Using the change of variables $|\theta|t = u$, we get $\int_0^T \frac{\sin(\theta t)}{t}\,dt = \frac{|\theta|}{\theta}\int_0^{|\theta|T}\frac{\sin(u)}{u}\,dt$

□

Theorem 9.5. If a probability measure μ has characteristic function $\phi(t) = \int_{\mathbb{R}} e^{itx}\mu(dx)$ and if $\mu(\{a\}) = \mu(\{b\}) = 0$, then

$$\mu(a,b] = \lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\phi(t)\,dt. \tag{11}$$

Proof. By Fubini's theorem,

$$I_T := \frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\phi(t)\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-T}^{T}\frac{e^{it(x-a)}-e^{it(x-b)}}{it}\,dt\,\mu(dx).$$

Using the fact that $\frac{\sin\theta t}{t}$ is even and $\frac{\cos\theta t}{t}$ is odd and that

$$\begin{aligned}\frac{e^{it(x-a)}-e^{it(x-b)}}{it} &= \frac{\cos(t(x-a))+i\sin(t(x-a))-(\cos(t(x-b))+i\sin(t(x-b)))}{it}\\ &= i\frac{\cos(t(x-b))-\cos(t(x-a))}{t}+\frac{\sin(t(x-a))-\sin(t(x-b))}{t},\end{aligned}$$

we get, via Lemma 9.1 2.,

$$\begin{aligned}I_T &= \frac{1}{\pi}\int_{-\infty}^{\infty}\int_0^T\frac{\sin(t(x-a))-\sin(t(x-b))}{t}\,dt\,\mu(dx)\\ &= \int_{-\infty}^{\infty}\frac{1}{\pi}(sgn(x-a)S(T|x-a|)-sgn(x-b)S(T|x-b|))\,\mu(dx).\end{aligned}$$

By Lemma 9.1 1., the integrand converges, as $T\to\infty$ to the function $f_{a,b}(x) = \mathbb{1}_{\{a<x<b\}}(x) + \frac{1}{2}\mathbb{1}_{x\in\{a,b\}}(x)$. The bounded convergence theorem now implies that $I_T \to \int_{-\infty}^{\infty} f_{a,b}(x)\,\mu(dx)$.

□

Corollary 9.1. The characteristic function characterizes the random variable: If $\phi_X(t) = \phi_Y(t)$ then X and Y have the same distribution, that is, the measures P^X and P^Y are the same.

Proof. Intervals $(a,b]$ with $a<b, \mu\{a\} = \mu\{b\} = 0$ form a π-system. This π-system generates the π-system of *all* intervals of the form $(a,b]$ with $a<b$. This is clear once we know that there are only countably many points a_i for which $\mu\{a_i\} \neq 0$. This can be proved by contradiction: Suppose there are uncountably many points a_i for which $\mu\{a_i\} > 0$. Then there must be uncountably many a_i for which $\mu\{a_i\} \in \left(\frac{1}{n+1}, \frac{1}{n}\right]$ for some $n\in\mathbb{N}$. But this implies that $\mu(\mathbb{R}) = \infty$. □

Example 9.4. If X has a Binomial(n,p) distribution ($X \sim Bin(n,p)$), that is, for $0\le k\le n$,

$$P(X=k) = \binom{n}{k}p^k(1-p)^{n-k},$$

then

$$\phi_X(u) = [pe^{iu} + 1 - p]^n.$$

It follows from Theorem 9.3 that if $X \sim Bin(n, p)$ and $Y \sim Bin(m, p)$ are independent, then $\phi_{X+Y} = [pe^{iu} + 1 - p]^n [pe^{iu} + 1 - p]^m = [pe^{iu} + 1 - p]^{n+m}$. Corollary 9.1 now implies that $X + Y \sim Bin(n + m, p)$. By induction, if $X_j \sim Bin(n_j, p)$ are independent, $\sum_{j=1}^{n} X_j \sim Bin(\sum_{j=1}^{n} n_j, p)$.

Example 9.5. If $X \sim N(\mu, \sigma^2)$,

$$\phi_X(t) = \exp\{i\mu t - \frac{\sigma^2 t^2}{2}\}.$$

Therefore, if $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent,

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) = \exp\{i\mu_1 t - \frac{\sigma_1^2 t^2}{2}\} \exp\{i\mu_2 t - \frac{\sigma_2^2 t^2}{2}\} = \exp\{i(\mu_1 + \mu_2)t - \frac{(\sigma_1^2 + \sigma_2^2)t^2}{2}\},$$

implying that $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. By induction, if $X_j \sim N(\mu_j, \sigma_j^2)$ are independent, $\sum_{j=1}^{n} X_j \sim N(\sum_{j=1}^{n} \mu_j, \sum_{j=1}^{n} \sigma_j^2)$.

Example 9.6. Suppose that $\lambda > 0$ and X_j are independent random variables with a Poisson distribution with parameter λ_j ($X_j \sim Po(\lambda_j)$), that is,

$$P\{X_j = k\} = \frac{\lambda_j^k e^{-\lambda_j}}{k!}, \quad k = 0, 1, 2, 3, \ldots.$$

Then

$$\phi_{X_j}(t) = \exp\{\lambda_j(e^{it} - 1)\}.$$

and if $S_n = \sum_{j=1}^{n} X_j$

$$\phi_{S_n}(t) = \exp\{\sum_{j=1}^{n} \lambda_j(e^{it} - 1)\},$$

so $S_n \sim Po(\sum_{j=1}^{n} \lambda_j)$.

Lecture #10: Central Limit Theorem

Reference. Sections 3.2, 3.3

10.1 Some Basic Facts from Complex Analysis

Definition 10.1. The residue $\mathrm{Res}(f, a)$ of a meromorphic function f at an isolated singularity a, is the coefficient a_{-1} of $(z-a)^{-1}$ in the Laurent series expansion of f around a. At a simple pole, the residue is given by:

$$\mathrm{Res}(f, a) = \lim_{z \to a} (z-a) f(z).$$

Theorem 10.1. Suppose D is a simply connected open subset of the complex plane, and $a_1, \ldots, a_n$ are finitely many points of D and f is a function which is defined and holomorphic on $D \setminus \{a_1, \ldots, a_n\}$. If γ is a rectifiable Jordan (i.e., closed, non-self-intersecting) curve in D which disconnects a_k from infinity for all $k \in \{1, \ldots, n\}$ but such that $\gamma \cap a_k = \emptyset$ for all $k \in \{1, \ldots, n\}$, then

$$\oint_\gamma f(z)\, dz = 2\pi i \sum_{k=1}^{n} \mathrm{Res}(f, a_k),$$

where $\mathrm{Res}(f, a_k)$ denotes the residue of f at a_k.

10.2 More on Characteristic Functions

Example 10.1. If $X \sim N(\mu, \sigma^2)$,

$$\phi_X(t) = \exp\left\{ i\mu t - \frac{\sigma^2 t^2}{2} \right\}.$$

Indeed, if $X \sim N(\mu, \sigma^2)$, $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$. Therefore, $X = \sigma Z + \mu$, implying that

$$\phi_X(t) = e^{i\mu t} \phi_Z(\sigma t). \tag{12}$$

Now

$$\begin{aligned}
\phi_Z(t) &= E[e^{itZ}] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{itx} e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \sum_{n \geq 0} \frac{(itx)^n}{n!} e^{-x^2/2}\, dx \\
&= \sum_{n \geq 0} \frac{(it)^n}{n!} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x^n e^{-x^2/2}\, dx = \sum_{n \geq 0} \frac{(it)^{2n}}{(2n)!} \frac{(2n)!}{2^n n!} = \sum_{n \geq 0} \frac{(-1)^n t^{2n}}{2^n n!} \\
&= \sum_{n \geq 0} \left(-\frac{t^2}{2} \right)^n \frac{1}{n!} = e^{-t^2/2},
\end{aligned}$$

where the second equality on the second line follows from a computation from the end of Lecture 3. The expression for $\phi_X(t)$ now follows from (12).

Example 10.2. If X has the standard Cauchy distribution with density

$$f(x) = \frac{1}{\pi(1+x^2)},$$

then

$$\phi_X(t) = E[e^{itX}] = \int_{\mathbb{R}} \frac{e^{itx}}{\pi(1+x^2)}\, dx.$$

We evaluate this integral by computing

$$\oint_{\gamma_R} \frac{e^{itz}}{\pi(1+z^2)}\, dz$$

along the curve γ_R composed of the line segment $[-R, R]$ and the semi-circle $C_R = \{z \in \mathbb{C} : |z| = R, \mathrm{Im}(z) \geq 0\}$, traversed counterclockwise, where $R > 1$. Then

$$\oint_{\gamma_R} \frac{e^{itz}}{\pi(1+z^2)}\, dz = \int_{-R}^{R} \frac{e^{itx}}{\pi(1+x^2)}\, dx + \int_{C_R} \frac{e^{itz}}{\pi(1+z^2)}\, dz. \tag{13}$$

Also, by the residue theorem,

$$\oint_{\gamma_R} \frac{e^{itz}}{\pi(1+z^2)}\, dz = 2\pi i \lim_{z\to i} \frac{e^{itz}(z-i)}{\pi(1+z^2)} = 2\pi i \lim_{z\to i} \frac{e^{itz}}{\pi(z+i)} = 2\pi i \frac{e^{-t}}{\pi(2i)} = e^{-t}. \tag{14}$$

Now if $|z| = |x+iy| = R$,

$$\left| \frac{e^{itz}}{\pi(1+z^2)} \right| \leq \frac{e^{-ty}}{\pi R^2},$$

so since $y \geq 0$, if $t \geq 0$,

$$\left| \int_{C_R} \frac{e^{itz}}{\pi(1+z^2)}\, dz \right| \leq \frac{1}{\pi R^2} \int_{C_R} dz = \frac{1}{R} \to 0, \text{ as } R \to \infty.$$

(13), (14), and continuity of Lebesgue integral now imply that

$$\int_{\mathbb{R}} \frac{e^{itx}}{\pi(1+x^2)}\, dx = e^{-t}.$$

If $t \leq 0$, we can repeat the same argument with γ_R' composed of the line segment $[-R, R]$ and the semi-circle $C_R = \{z \in \mathbb{C} : |z| = R, \mathrm{Im}(z) \leq 0\}$. Alternatively, we can just use the fact that $\phi_X(-t) = \bar{\phi}_X(t)$ for all t. This then gives

$$\phi_X(t) = e^{-|t|}.$$

Theorem 10.2. Let X be a random variable with distribution function F_X. Then ϕ_X is real-valued if and only if the measure dF_X is symmetric ($\int_B dF_X(x) = \int_{-B} dF_X(x) \,\forall\, B \in \mathcal{R}$).

Proof. Suppose dF_X is symmetric. Then $\int_{\mathbb{R}} \sin(tx)\, dF_X(x) = 0$, so $\phi_X(t) = E[\cos(tX)]$, which is real-valued.

Conversely, suppose ϕ_X is real-valued. Then

$$\phi_{-X}(t) = \overline{\phi_X(t)} = \phi_X(t).$$

By the inversion theorem, the measures P^X and P^{-X} are the same, so

$$P(X \in B) = P(-X \in B) = P(X \in -B).$$

□

Corollary 10.1. If X, Y are i.i.d., then $X - Y$ is symmetric.

Proof. Since X and Y are i.i.d.,

$$\phi_{X-Y}(t) = \phi_X(t)\phi_{-Y}(t) = \phi_X(t)\overline{\phi_X(t)} = |\phi_X(t)|^2 \in \mathbb{R}.$$

□

Note 10.1. It also follows immediately that if X and Y are independent, symmetric, then $X + Y$ is symmetric.

10.3 More on Weak Convergence

Theorem 10.3 (Helly's Theorem)**.** For every sequence $\{F_n\}$ of distribution functions, there exists a subsequence $\{F_{n_k}\}$ and a nondecreasing right-continuous function F such that $\lim_{k\to\infty} F_{n_k}(x) = F(x)$ at all continuity points of F.

Proof. The diagonal method gives a sequence $\{n_k\} \in \mathbb{N}$ such that $G(q) := \lim_{k\to\infty} F_{n_k}(q)$ exists for every $r \in \mathbb{Q}$.

Indeed, let $\mathbb{Q} = \{q_1, q_2, \ldots\}$. Then since $\{F_n(q_1)\}_{n\geq 1}$ is a bounded sequence, it contains a convergent subsequence $\{F_{n(1,j)}(q_1)\}_{j\geq 1}$ with $\{F_{n(1,j)}(q_1)\}_{j\geq 1} \overset{j\to\infty}{\to} G(q_1)$. There then is some subsequence $\{n(2,j)\}_{j\geq 1}$ of $\{n(1,j)\}_{j\geq 1}$ with $\{F_{n(2,j)}(q_2)\}_{j\geq 1} \overset{j\to\infty}{\to} G(q_2)$, and so on. If we let for $i \in \mathbb{N}, n_i = n(i,i)$, we see that for any $j \in \mathbb{N}$, the following is well defined:

$$G(q_j) = \lim_{i\to\infty} F_{n_i}(q_j).$$

For $x \in \mathbb{R}$, define $F(x) = \inf\{G(r) : r > x\}$. [Note that for $q \in \mathbb{Q}, F(q)$ may be different form $G(q)$.] Then

- F is obviously non-decreasing.
- For every $x \in \mathbb{R}, \epsilon > 0$, there exists $q \in \mathbb{Q}$ such that $x < q$ and $G(q) < F(x) + \epsilon$ (otherwise, $F(x)$ wouldn't be the inf it's defined to be). If $x \leq y < q, F(y) \leq G(q) < F(x) + \epsilon$. Therefore, F is right-continuous.

Suppose F is continuous at x. Choose $y < x$ such that $F(y) > F(x) - \epsilon$ and $r, s \in \mathbb{Q}$ such that $y < r < x < s$ and $G(s) < F(x) + \epsilon$. Since $F(x) - \epsilon < G(r) \leq G(s) < F(x) + \epsilon$ and since $F_n(r) \leq F_n(x) \leq F_n(s)$, we get that

$$\limsup_k F_{n_k}(x) \leq \limsup F_{n_k}(s) = G(s)$$

and

$$\liminf_k F_{n_k}(x) \geq \liminf F_{n_k}(s) = G(r).$$

Since this means that for every $\epsilon > 0$, $|\limsup_k F_{n_k}(x) - F(x)| < \epsilon$ and $|\liminf_k F_{n_k}(x) - F(x)| < \epsilon$, we get that $\lim_{k\to\infty} F_{n_k}(x) = F(x)$. □

Example 10.3. Let $X_n = n$, a.s. Then $F_n(x) = \mathbb{1}_{\{X \geq n\}}$ so for any sequence $\{n_k\}_{k\geq 1} \subset \mathbb{N}$, we have $\lim_{k\to\infty} F_{n_k}(x) = F(x) = 0$ for every $x \in \mathbb{R}$. So although the $F(x)$ from Helly's theorem satisfies $0 \leq |F(x)| \leq 1$ for every $x \in \mathbb{R}$, it may not be a distribution.

One main reason preventing the limit from being a distribution is that the mass in $\{F_n\}$ may run off to infinity. The following definition is trying to deal with that:

Definition 10.2. A sequence of probability measures μ_n on $\mathbb{R}$ is *tight* if for every $\epsilon > 0$, there exists an interval $(a, b]$ such that $\mu_n(a, b] > 1 - \epsilon \; \forall \, n$.

Theorem 10.4. A sequence of probability measures $\{\mu_n\}$ is tight if and only if for every sequence $\{n_k\}_{k\geq 1}$, there exists a subsequence $\{n_{k(j)}\}_{j\geq 1}$ and a probability measure μ such that $\mu_{n_{k(j)}} \Rightarrow \mu$ as $j \to \infty$.

Note 10.2. Among other things, this theorem says that if $\{\mu_n\}_{n\geq 1}$ is not tight, there exists a sequence μ_{n_k} of which no subsequence converges weakly to a probability measure.

Proof. $\underline{\Rightarrow:}$ Let $\{F_{n_k}\}_{k\geq 1}$ be the distribution functions associated with $\{\mu_{n_k}\}_{k\geq 1}$. By Helly's theorem, there exists a subsequence $\{F_{n_{k(j)}}\}$ with $F_{n_{k(j)}}(x) \to F(x)$, a non-decreasing, right-continuous function, at points of continuity of F. Now there exists a unique measure μ on $(\mathbb{R}, \mathcal{R})$ such that for all $a \leq b, \mu(a, b] = F(b) - F(a)$ (adapt the result of Chapter 2 for probability measures to the present case where μ is just a finite measure). By tightness, for every $\epsilon > 0$, we can find $a \leq b$ such that for all $n \geq 1, F_n(b) - F_n(a) > 1 - \epsilon$ and F is continuous at a and b. But since $F_n(b) \to F(b)$ and $F_n(a) \to F(a)$, we get $\mu(a, b] \geq 1 - \epsilon$. Therefore, μ is a probability measure. Theorem 9.2 now tells us that since $F_{n_{k(j)}}(x) \to F(x)$ at points of continuity of F, $\mu_{n_{k(j)}} \Rightarrow \mu$.

$\underline{\Leftarrow:}$ Suppose $\{\mu_n\}$ is not tight. Then we can find $\{n_k\}_{k\geq 1}$ such that for all $k \geq 1, \mu_{n_k}(-k, k] \leq 1 - \epsilon$. Now if $\{n_{k(j)}\}_{j\geq 1}$ is such that $\mu_{n_{k(j)}} \Rightarrow \mu$, we can find $(a, b]$ such that $\mu(a, b] > 1 - \epsilon$ and $\mu\{a\} = \mu\{b\} = 0$. Then there exists j_0 such that for all $j \geq j_0, (a, b] \subset (-k(j), k(j)]$, so

$$1 - \epsilon \geq \mu_{n_{k(j)}}(-k(j), k(j)] \geq \mu_{n_{k(j)}}(a, b] \to \mu(a, b].$$

Therefore, $\mu(a, b] \leq 1 - \epsilon$, which is a contradiction. □

Corollary 10.2. If $\{\mu_n\}$ is a tight sequence of probability measures and if each subsequence that converges weakly converges weakly to μ, then $\mu_n \Rightarrow \mu$.

Proof. Theorem 10.4 and the hypothesis imply that for every sequence $\{n_k\}_{k\geq 1}$, there exists a sequence $\{n_{k(j)}\}_{j\geq 1} \subset \{n_k\}_{k\geq 1}$ such that $\mu_{n_{k(j)}} \Rightarrow \mu$. Suppose $\mu_n \not\Rightarrow \mu$. Then there exists x with $\mu\{x\} = 0$ such that $F_n(x) \to F(x)$, so there exists $\epsilon > 0, \{n_k\}_{k\geq 1}$ such that $|F_{n_k}(x) - F_k(x)| \geq \epsilon$ for all $k \geq 1$. So there is no $\{n_{k(j)}\}_{j\geq 1}$ such that $F_{n_{k(j)}}(x) \to F(x)$. So $\mu_{n_{k(j)}} \not\Rightarrow \mu$. □

10.4 The Continuity Theorem

Theorem 10.5. Suppose $\{\mu_n\}, \mu$ are probability measures with characteristic functions $\{\phi_n\}, \phi$. Then

$$\mu_n \Rightarrow \mu \iff \phi_n(t) \to \phi(t) \,\forall\, t.$$

Proof. Suppose $\mu_n \Rightarrow \mu$. Then $\int_{\mathbb{R}} f(x)\,\mu_n(dx) \to \int_{\mathbb{R}} f(x)\,\mu(dx)$ for every bounded continuous function f. Therefore, for any $t \in \mathbb{R}$,

$$\begin{aligned}\phi_n(t) &= \int_{\mathbb{R}} e^{itx}\,\mu_n(dx) \\ &= \int_{\mathbb{R}} \cos(tx)\,\mu_n(dx) + i\int_{\mathbb{R}} \sin(tx)\,\mu_n(dx) \to \int_{\mathbb{R}} \cos(tx)\,\mu(dx) + i\int_{\mathbb{R}} \sin(tx)\,\mu(dx) = \phi(t).\end{aligned}$$

Now suppose that for all $t \in \mathbb{R}, \phi_n(t) \to \phi(t)$. We will show that $\{\mu_n\}$ is tight.

Since ϕ is continuous at 0 and $\phi(0) = 1$, for every $\epsilon > 0$, we can find u such that $u^{-1}\int_{-u}^{u}(1 - \phi(t))\,dt < \epsilon$ (note in particular that the integral is real-valued). Now since $\phi_n(t) \to \phi(t)$, the bounded convergence theorem implies that there exists $n_0 > 0$ such that for all $n \geq n_0$, $u^{-1}\int_{-u}^{u}(1 - \phi_n(t))\,dt < \epsilon$. Now we will try to relate this integral to μ_n. Fubini's theorem gives

$$\begin{aligned}u^{-1}\int_{-u}^{u}(1 - \phi_n(t))\,dt &= \int_{\mathbb{R}} u^{-1}\int_{-u}^{u}(1 - e^{itx})\,dt\,\mu_n(dx) = 2\int_{\mathbb{R}}\left(1 - \frac{\sin(ux)}{ux}\right)\mu_n(dx) \\ &\geq 2\int_{|x|\geq 2/u}\left(1 - \frac{1}{|ux|}\right)\mu_n(dx) \geq \mu_n\{x : |x| \geq 2/u\}.\end{aligned}$$

This implies that for all $n \geq n_0$,

$$\mu_n\{x : |x| \geq 2/u\} < 2\epsilon.$$

So we can find $K > 0$ such that for all $n \geq 1$,

$$\mu_n\{x : |x| \geq K\} < 2\epsilon.$$

So μ_n is tight.

Now suppose that $\{\mu_{n_k}\}$ is a subsequence of $\{\mu_n\}$ that converges to some measure ν. Then by the first part of our proof, ν has characteristic function $\lim \phi_n(t) = \phi(t)$. So since characteristic functions uniquely determine measures, μ and ν are the same measure. Corollary 10.2 now implies that $\mu_n \Rightarrow \mu$.

□

10.5 The central limit theorem

Definition 10.3. A function $h(t)$ is said to be *little-o of* $g(t)$, written $h(t) = o(g(t))$, if

$$\lim_{t\to 0+} \frac{h(t)}{g(t)} = 0.$$

In order to prove the central limit theorem, the following lemma will be of use:

Lemma 10.1. Let $z_1, \ldots, z_m$ and $w_1, \ldots, w_m$ be complex numbers of modulus less than or equal to 1. Then

$$\left| \prod_{i=1}^{n} z_i - \prod_{i=1}^{n} w_i \right| \leq \sum_{i=1}^{n} |z_i - w_i|.$$

Proof. By induction, using the fact that

$$\left| \prod_{i=1}^{n} z_i - \prod_{i=1}^{n} w_i \right| = (z_1 - w_1) \prod_{i=2}^{n} z_i + \left| \prod_{i=2}^{n} z_i - \prod_{i=2}^{n} w_i \right|.$$

□

Theorem 10.6 (Central Limit Theorem)**.** Suppose that $X_1, X_2, \ldots$ are independent and identically distributed L^2 random variables. If we write $E[X_1] = \mu$ and $Var(X_1) = \sigma^2$ for their common mean and variance, respectively, then as $n \to \infty$,

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \to Z$$

in distribution, where Z is normally distributed with mean 0 and variance 1.

Proof. For each $n = 1, 2, 3, \ldots$, let

$$Y_n = \frac{X_n - \mu}{\sigma}$$

so that $Y_1, Y_2, \ldots$, are independent and identically distributed with $E[Y_1] = 0$ and $Var(Y_1) = 1$. Furthermore, if we define S_n by

$$S_n = \frac{Y_1 + \cdots + Y_n}{\sqrt{n}},$$

then

$$S_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

and so the theorem will be proved if we can show $S_n \to Z$ in distribution. Using the properties of characteristic functions derived during Lecture 9, we find

$$\phi_{S_n}(t) = E[e^{itS_n}] = E[e^{i\frac{t}{\sqrt{n}}(Y_1+\cdots+Y_n)}) = \phi_{Y_1}(t/\sqrt{n}) \cdots \phi_{Y_n}(t/\sqrt{n}) = \left[\phi_{Y_1}(t/\sqrt{n})\right]^n.$$

By Theorem 9.4, if X has 2 moments,

$$\left|\phi_X(t) - (1 + itE[X] - \frac{1}{2}t^2E[X^2])\right| \leq E\left[\min\left\{\frac{|tX|^3}{3!}, \frac{2|tX|^2}{2!}\right\}\right].$$

$\min\left\{|t|\frac{|X|^3}{3!}, \frac{2|X|^2}{2!}\right\}$ is dominated by $X^2 \in L^1$ and as $t \to 0$, goes to 0 almost surely (since it is bounded by $|t|\frac{|X(\omega)|^3}{3!}$ for every ω), so $E[\min\left\{|t|\frac{|X|^3}{3!}, \frac{2|X|^2}{2!}\right\}] = g(t)$, with $g(t) \to 0$ as $t \to 0$. Therefore,

$$\left|\phi_X(t) - (1 + itE[X] - \frac{1}{2}t^2E[X^2])\right| \leq t^2g(t),$$

with $g(t) = o(1)$. In particular,

$$\left|\phi_{Y_1}(t) - (1 - \frac{1}{2}t^2)\right| \leq t^2g(t), \tag{15}$$

which, together with Lemma 10.1, gives for all $n > t^2/2$

$$\begin{aligned}\left|\phi_{S_n}(t) - (1 - \frac{t^2}{2n})^n\right| &= \left|\left[\phi_{Y_1}(t/\sqrt{n})\right]^n - (1 - \frac{t^2}{2n})^n\right| \\ &\leq n\left|\phi_{Y_1}(t/\sqrt{n}) - (1 - \frac{t^2}{2n})\right| \overset{(15)}{\leq} n\frac{t^2}{n}g(t/\sqrt{n}) \overset{n\to\infty}{\to} 0.\end{aligned}$$

Since $(1 - \frac{t^2}{2n})^n \to e^{-t^2/2}$ as $n \to \infty, \phi_{S_n}(t) \to e^{-t^2/2}$ as $n \to \infty$. Since $e^{-t^2/2}$ is the characteristic function of the standard normal distribution, the continuity theorem now implies that $S_n \Rightarrow Z$. □

Lecture #11: Conditional Expectation

Reference. Sections 2.5

11.1 Orthogonal Projection

Recall that we proved in Lecture 5 that $L^2(\Omega, \mathcal{F}, P)$ is a Hilbert space with inner product $\langle X, Y\rangle = E[XY]$. Unless noted otherwise, in this lecture, $||\cdot||$ means $||\cdot||_2$.

Theorem 11.1. Suppose $\mathcal{K}$ is a complete vector subspace of L^2. Then given $X \in L^2$, there exists a $Y \in \mathcal{K}$ such that $||X - Y|| = \inf\{||X - V|| : V \in \mathcal{K}\}$ and $\langle X - Y, Z\rangle = 0$ for every $Z \in \mathcal{K}$. These two properties are equivalent and if Y, Y' satisfy either one of them, $Y = Y'$ almost surely.

Definition 11.1. The random variable Y from the theorem is (a version of) the *orthogonal projection of X onto $\mathcal{K}$.*

Proof. Consider a sequence of random variables $\{Y_n\} \in \mathcal{K}$ such that $||X - Y_n|| \to \inf\{||X - V|| : V \in \mathcal{K}\}$. Then, by the parallelogram law ($||U + V||^2 + ||U - V||^2 = 2||U||^2 + 2||V||^2$),

$$||X - Y_r||^2 + ||X - Y_s||^2 = 2||X - (Y_r + Y_s)/2||^2 + 2||(Y_r - Y_s)/2||^2.$$

Since $||X - Y_r||^2 + ||X - Y_s||^2 \to 2\inf\{||X - V|| : V \in \mathcal{K}\}$ and $2||X - (Y_r + Y_s)/2||^2 \geq 2\inf\{||X - V|| : V \in \mathcal{K}\}$ (since $(Y_r + Y_s)/2 \in \mathcal{K}$),

$$\sup_{r,s\geq k} ||(Y_r - Y_s)/2|| \overset{k\to\infty}{\to} 0,$$

so $\{Y_n\}$ is a Cauchy sequence. Since $\mathcal{K}$ is assumed to be complete, there exists $Y \in \mathcal{K}$ such that $||Y - Y_k|| \to 0$. By the triangle inequality, $||X - Y|| \leq ||X - Y_n|| + ||Y_n - Y||$, so that (since $||X - Y_n|| \to \inf\{||X - V|| : V \in \mathcal{K}\}$ and $||Y_n - Y|| \to 0$)

$$||X - Y|| = \inf\{||X - V|| : V \in \mathcal{K}\}.$$

Now if $Z \in \mathcal{K}$, then for any $t \in \mathbb{R}, Y + tZ \in \mathcal{K}$, so

$$||X - Y - tZ|| \geq ||X - Y||,$$

implying that

$$\begin{aligned} \langle X - Y - tZ, X - Y - tZ\rangle \quad &\geq \quad \langle X - Y, X - Y\rangle \iff \langle -tZ, X - Y - tZ\rangle \geq 0 \\ &\iff \quad -t\langle Z, X - Y\rangle + t^2\langle Z, Z\rangle \geq 0. \end{aligned}$$

This inequality will hold for all $t \in \mathbb{R}$ if and only if $\langle Z, X - Y\rangle = 0$. □

11.2 Conditional Expectation

The section on orthogonal projection will be useful for the proof of the following theorem, which defines conditional expectation:

Theorem 11.2 (Definition of Conditional Expectation)**.** Suppose that $(\Omega, \mathcal{F}, P)$ is a probability space, and let X be a real-valued L^1 random variable. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. There exists a random variable Y such that

(i) Y is $\mathcal{G}$-measurable, i.e., $Y : (\Omega, \mathcal{G}, P) \to (\mathbb{R}, \mathcal{B})$,

(ii) $E[|Y|] < \infty$,

(iii) for every set $G \in \mathcal{G}$, we have
$$E[Y \mathbb{1}_G] = E[X \mathbb{1}_G].$$

Moreover, Y is almost surely unique. That is, if $\tilde{Y}$ is another random variable with these properties, then $P\{\tilde{Y} = Y\} = 1$. We call Y (a version of) the *conditional expectation of X given $\mathcal{G}$* and write
$$Y = E[X|\mathcal{G}].$$

Example 11.1. Here is a very simple explicit example which it may be helpful to keep in mind when losing track of the meaning of conditional expectation:
Suppose $\Omega = \{1, 2, 3, 4\}, \mathcal{F} = 2^\Omega$, and P is defined on Ω by

$$P(i) = \frac{16}{15} 2^{-i}.$$

Define on $(\Omega, \mathcal{F}, P)$ the random variable X by $X(\omega) = \omega$ and note that $E[X] = \frac{26}{15}$. Define also

$$\mathcal{F}_0 = \{\emptyset, \Omega\},$$
$$\mathcal{F}_1 = \{\emptyset, \{1\}, \{2, 3, 4\}, \Omega\},$$
$$\mathcal{F}_2 = \{\emptyset, \{1\}, \{2, 3, 4\}, \{2, 3\}, \{4\}, \{1, 4\}, \{1, 2, 3\}\}, \Omega\}.$$

Then
$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}.$$

Denote by Y_i the conditional expectation $E[X|\mathcal{F}_i]$. Since Y_1 is $\mathcal{F}_1$-measurable (by definition), Y_1 must be constant on the nontrivial elements of $\mathcal{F}_1$ (since for any $x \in \mathbb{R}, X^{-1}(\{x\}) \in \mathcal{F}_1$). So
$$Y_1(\omega) = y_1 \mathbb{1}\{\omega \in \{1\}\} + y_2 \mathbb{1}\{\omega \in \{2, 3, 4\}\},$$
which implies
$$y_1 P(\omega = 1) = E[Y_1 \mathbb{1}\{\omega = 1\}] = E[X \mathbb{1}\{\omega = 1\}] = \frac{8}{15},$$
$$y_2 P(\omega \in \{2, 3, 4\}) = E[Y_1 \mathbb{1}\{\omega \in \{2, 3, 4\}\}] = E[X \mathbb{1}\{\omega \in \{2, 3, 4\}\}] = \frac{6}{5},$$

from which we obtain that

$$P(Y_1 = 1) = \frac{8}{15}, \qquad P(Y_1 = \frac{18}{7}) = \frac{7}{15}.$$

Similarly, we find that Y_2 must be constant on the sets $\{1\}, \{2,3\}$, and $\{4\}$, so going through the same calculations give

$$P(Y_2 = 1) = \frac{8}{15}, \qquad P(Y_2 = \frac{14}{6}) = \frac{6}{15}, P(Y_2 = 4) = \frac{1}{15}.$$

We now turn to the trivial σ-algebra $\mathcal{F}_0$. Since Y_0 must be $\mathcal{F}_0$-measurable, Y_0 must be constant. Since $E[Y_0] = E[Y_0 \mathbb{1}\{\Omega\}] = E[X \mathbb{1}\{\Omega\}] = E[X]$, Y_0 is a random variable concentrated on $E[X]$. That is, $P(Y = \frac{26}{15}) = 1$.

Finally, let's examine $E[X|\mathcal{F}]$. Since $E[X|\mathcal{F}]$ and X have same expectation on every element of $\mathcal{F}$ and Ω is finite, $E[X|\mathcal{F}]$ and X are the same random variable.

This example suggests the following way of thinking about conditional expectation. As one refines the σ-algebras $\mathcal{G}$ going from $\{\emptyset, \Omega\}$ to $\mathcal{F}$, one goes from a coarse to a fine picture of the random variable X when looking at $E[Y|\mathcal{G}]$.

Proof of Theorem 11.2

Uniqueness: Let Y and $\tilde{Y}$ be versions of $E[X|\mathcal{G}]$. Then Y, $\tilde{Y} \in L^1(\Omega, \mathcal{G}, P)$ and

$$E[(Y - \tilde{Y})1_G] = 0 \quad \text{for every } G \in \mathcal{G}.$$

Suppose that Y and $\tilde{Y}$ are not almost surely equal and assume (without loss of generality) that we have

$$P\{Y > \tilde{Y}\} > 0.$$

Since

$$\{Y > \tilde{Y} + n^{-1}\} \uparrow \{Y > \tilde{Y}\}$$

we see that there exists an n such that

$$P\{Y > \tilde{Y} + n^{-1}\} > 0.$$

However, Y and $\tilde{Y}$ are both $\mathcal{G}$-measurable and so it follows that $\{Y > \tilde{Y} + n^{-1}\} \in \mathcal{G}$. Now if we choose $G = \{Y > \tilde{Y} + n^{-1}\}$ then

$$E[(Y - \tilde{Y})1_G] = E[(Y - \tilde{Y})1_{\{Y > \tilde{Y} + n^{-1}\}}] \geq n^{-1} P\{Y > \tilde{Y} + n^{-1}\} > 0.$$

This contradicts the fact that $E[(Y - \tilde{Y})1_G] = 0$ for *every* $G \in \mathcal{G}$, and so we are forced to conclude that $Y = \tilde{Y}$ almost surely.

Existence: We first deal with the case when $X \in L^2(\Omega, \mathcal{F}, P)$. We know that $L^2(\Omega, \mathcal{G}, P)$ is a Hilbert space and by Theorem 11.1, that there exists $Y \in L^2(\Omega, \mathcal{G}, P)$ such that

$$\langle X - Y, Z \rangle = 0 \, \forall \, Z \in L^2(\Omega, \mathcal{G}, P). \tag{16}$$

For $G \in \mathcal{G}$, $Z = \mathbb{1}\{G\} \in L^2(\Omega, \mathcal{G}, P)$, so that by (16), $E[(X - Y)\mathbb{1}\{G\}] = 0$, which implies that $E[X\mathbb{1}\{G\}] = E[Y\mathbb{1}\{G\}]$. Therefore, $Y = E[X|\mathcal{G}]$.

Before dealing with the L^1 case, we prove the following:

Lemma 11.1. If $X \geq 0$ almost surely, then $E[X|\mathcal{G}] \geq 0$ almost surely.

Proof. Let Y be a version of $E[X|\mathcal{G}]$. Assume that Y is not almost surely greater than or equal to 0 so that $P\{Y < 0\} > 0$. Thus, there exists some n such that

$$P\{Y < -n^{-1}\} > 0.$$

If we set $G = \{Y < -n^{-1}\}$, then $G \in \mathcal{G}$ (since Y is $\mathcal{G}$-measurable), so that

$$0 \leq E[X\mathbb{1}_G] = E[Y\mathbb{1}_G] < -n^{-1}P\{G\} < 0.$$

This is a contradiction and so we are forced to conclude that $Y \geq 0$ almost surely. □

Now suppose $X \in L^1$. Since we can write $X = X^+ - X^-$, we can suppose $X \geq 0$. Consider a sequence $\{X_n\} \uparrow X$ with X_n bounded. Then, of course, for every $n, X_n \in L^2$, so $Y_n = E[X_n|\mathcal{G}]$ exists. By the lemma, $0 \leq Y_n \uparrow$, so we can define $Y(\omega) = \limsup Y_n(\omega)$. Y is $\mathcal{G}$-measurable and $Y_n \uparrow Y$, a.s. The monotone convergence theorem (applied here only over $G \in \mathcal{G}$) therefore implies that $E[Y\mathbb{1}\{G\}] = E[X\mathbb{1}\{G\}]$.

□

Now that we know that conditional expectation exists for L^1 random variables we can revisit the proof of Lemma 11.1 and see that the same proof pulls through. This gives:

Theorem 11.3 (Monotonicity of Conditional Expectation)**.** If $X \geq 0$ almost surely, then $E[X|\mathcal{G}] \geq 0$ almost surely. In particular, if $Z \geq X$ almost surely, then $E[Z|\mathcal{G}] \geq E[X|\mathcal{G}]$ almost surely.

Another consequence of the results in this section is

Theorem 11.4 (Conditional expectation is a best predictor in the least-squares sense)**.** If $X \in L^2(\Omega, \mathcal{F}, P)$ and $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$, then $Y = E[X|\mathcal{G}]$ is the orthogonal projection of X onto $L^2(\Omega, \mathcal{G}, P)$. So Y is the best least-squares $\mathcal{G}$-measurable predictor of X (it minimizes $E[(X - Y)^2]$ among all $\mathcal{G}$-measurable random variables).

The next two theorems generalize what we already observed in the example at the beginning of this lecture:

Theorem 11.5. If X is $\mathcal{F}$-measurable, $\mathcal{G} \subseteq \mathcal{F}$, and $Y = E[X|\mathcal{G}]$, then $E[Y] = E[X]$.

Proof. This follows from property (iii) in Theorem 11.2 by noting that $\Omega \in \mathcal{G}$ since $\mathcal{G}$ is a σ-algebra. That is,

$$E[Y] = E[Y1_\Omega] = E[X1_\Omega] = E[X]$$

and the proof is complete. □

Theorem 11.6. If the random variable X is $\mathcal{F}$-measurable, then $E[X|\mathcal{F}] = X$ almost surely.

Proof. This follows immediately from the fact that X satisfies conditions (i)-(iii) in Theorem 11.2 and from almost sure uniqueness. □

Theorem 11.7 (Linearity of Conditional Expectation)**.** If Y_1 is a version of $E[X_1|\mathcal{G}]$ and Y_2 is a version of $E[X_2|\mathcal{G}]$, then $\alpha Y_1 + \beta Y_2$ is a version of $E[\alpha X_1 + \beta X_2|\mathcal{G}]$.

Proof. This result is also immediate from the definition of conditional expectation. □

Theorem 11.8 (Tower Property of Conditional Expectation)**.** If $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$, then

$$E[E[X|\mathcal{F}]\,|\,\mathcal{G}] = E[X|\mathcal{G}] = E[E[X|\mathcal{G}]\,|\,\mathcal{F}]$$

almost surely.

Proof. Suppose $Y = E[E[X|\mathcal{F}]\,|\,\mathcal{G}]$. Then for every $G \in \mathcal{G} \subset \mathcal{F}$,

$$E[Y\mathbb{1}\{G\}] = E[E[X|\mathcal{F}]\mathbb{1}\{G\}] = E[X\mathbb{1}\{G\}] = E[E[X|\mathcal{G}]\mathbb{1}\{G\}].$$

This proves the first equality. The second is immediate by Theorem 11.6 since $E[X|\mathcal{G}]$ is $\mathcal{G}$-measurable and therefore $\mathcal{F}$-measurable. □

In particular, we see that we can think of conditional expectations geometrically. Indeed, $E[\cdot|\mathcal{F}]$ is a projection:

Corollary 11.1.

$$E[E[X|\mathcal{F}]|\mathcal{F}] = E[X|\mathcal{F}].$$

Theorem 11.9. ["Taking out what is known"] If X is $\mathcal{F}$-measurable and $E[|XY|] < \infty$, then

$$E[XZ|\mathcal{F}] = XE[Z|\mathcal{F}]$$

almost surely.

Proof. If $A \in \mathcal{F}$, then for any $B \in \mathcal{F}$,

$$E[\mathbb{1}_A E[Z|\mathcal{F}]1_B] = E[E[Z|\mathcal{F}]\mathbb{1}_{A\cap B}] \overset{A\cap B\in\mathcal{F}}{=} E[Z\mathbb{1}_{A\cap B}] = E[(\mathbb{1}_A Z)\mathbb{1}_B].$$

Since $1_A E[X|\mathcal{F}]$ is $\mathcal{F}$-measurable, the equality holds when $X = \mathbb{1}_A$ and $A \in \mathcal{F}$. Using linearity and taking limits yields the equality whenever Z is $\mathcal{F}$-measurable and X and XZ are integrable. □

Theorem 11.10. If $\mathcal{G}$ is independent of $\sigma(\sigma(X), \mathcal{F})$, then

$$E[X|\sigma(\mathcal{F}, \mathcal{G})] = E[X|\mathcal{F}], a.s.$$

In particular, if X is independent of $\mathcal{G}$, then $E[X|\mathcal{G}] = E[X]$, a.s.

Proof. Suppose without loss of generality that $X \geq 0$. If $F \in \mathcal{F}$ and $G \in \mathcal{G}$, then $X\mathbb{1}_F$ and $\mathbb{1}_G$ are independent, so

$$E[X\mathbb{1}_{F\cap G}] = E[X\mathbb{1}_F\mathbb{1}_G] = E[X\mathbb{1}_F]P(G).$$

Now suppose $Y = E[X|\mathcal{F}]$. Since Y is $\mathcal{F}$-measurable, $Y\mathbb{1}_F$ is independent of $\mathcal{G}$, so $E[(Y\mathbb{1}_F)\mathbb{1}_G] = E[Y\mathbb{1}_F]P(G)$, so $E[X\mathbb{1}_{F\cap G}] = E[Y\mathbb{1}_{F\cap G}]$. Therefore, the measures on $\sigma(\mathcal{F}, \mathcal{G})$: $\nu_1 : I \mapsto E[X\mathbb{1}_I]$ and $\nu_2 : I \mapsto E[Y\mathbb{1}_I]$ have same finite total mass and agree on the sets of the form $F \cap G$, where $F \in \mathcal{F}$ and $G \in \mathcal{G}$, a π-system. Therefore, they agree on $\sigma(\mathcal{F}, \mathcal{G})$. □

Finally, we state (almost) without proofs the analogues of the main results about interchanging limits and expectations. Since conditional expectations are random variables, these are statements about almost sure limits.

Theorem 11.11 (Monotone Convergence)**.** If $X_n \geq 0$ for all n and $X_n \uparrow X$, then

$$\lim_{n\to\infty} E[X_n|\mathcal{G}] = E[X|\mathcal{G}],$$

almost surely.

Proof. Suppose $X_n \uparrow X$. Then if $Y_n = E[X_n|\mathcal{G}], Y_n \uparrow$ a.s. Let $Y = \limsup Y_n$. Then Y is $\mathcal{G}$-measurable and $Y_n \uparrow Y$. Since $E[Y_n \mathbb{1}_G] = E[X_n \mathbb{1}_G]$ for all $G \in \mathcal{G}$, the monotone convergence theorem implies that $E[Y \mathbb{1}_G] = E[X \mathbb{1}_G]$ for all $G \in \mathcal{G}$. So $Y = E[X|\mathcal{G}]$. Therefore, $\lim E[X_n|\mathcal{G}] = E[X|\mathcal{G}]$ a.s. □

Theorem 11.12 (Fatou's Lemma)**.** If $X_n \geq 0$ for all n, then

$$\mathbb{E}\left[\liminf_{n\to\infty} X_n \,|\, \mathcal{F}\right] \leq \liminf_{n\to\infty} E[X_n|\mathcal{F}],$$

almost surely.

Theorem 11.13 (Dominated Convergence)**.** Suppose $X \in L^1$ and $|Y_n(\omega)| \leq X(\omega)$ for all n. If $Y_n \to Y$ almost surely, then

$$\lim_{n\to\infty} E[Y_n|\mathcal{F}] = E[Y|\mathcal{F}],$$

almost surely.

Lecture #12: Martingales

Reference.

12.1 Martingales

Suppose that $(\Omega, \mathcal{F}, P)$ is a probability space. An increasing sequence of sub-σ-algebras $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}$ is called a *filtration*.

Example 12.1. If $X_0, X_1, X_2, \ldots$ is a sequence of random variables on $(\Omega, \mathcal{F}, P)$ and for $j \in \mathbb{N}$, we let $\mathcal{F}_j = \sigma(X_0, X_1, \ldots, X_j)$, then $\{\mathcal{F}_n,\, n \in \mathbb{N}\}$ is a filtration often called the *natural filtration*.

Definition 12.1. A stochastic process $\{Y_n\}_{n\in\mathbb{N}\cup\{0\}}$ is *adapted* to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}\cup\{0\}}$ if for all $n \in \mathbb{N} \cup \{0\}, Y_n$ is $\mathcal{F}_n$-measurable.

Definition 12.2. A sequence $X_0, X_1, X_2, \ldots$ of random variables is said to be a *supermartingale* with respect to the filtration $\{\mathcal{F}_n\}$ if for every $n \in \mathbb{N}$,

(i) $E[|X_n|] < \infty$,

(ii) X_n is $\mathcal{F}_n$-measurable, that is, X_n is adapted to $\mathcal{F}_n$, and

(iii) $E[X_{n+1}|\mathcal{F}_n] \leq X_n$.

A sequence $X_0, X_1, X_2, \ldots$ of random variables is said to be a *submartingale* with respect to the filtration $\{\mathcal{F}_n\}$ if for every $n \in \mathbb{N}$,

(i) $E[|X_n|] < \infty$,

(ii) X_n is $\mathcal{F}_n$-measurable, that is, X_n is adapted to $\mathcal{F}_n$, and

(iii) $E[X_{n+1}|\mathcal{F}_n] \geq X_n$.

$X_0, X_1, X_2, \ldots$ is a *martingale* with respect to the filtration $\{\mathcal{F}_n\}$ if it is a supermartingale and a submartingale with respect to the filtration $\{\mathcal{F}_n\}$.

Note 12.1. If the filtration used is the natural one, then we sometimes write the third condition as

$$E[X_{n+1}|X_0, X_1, \ldots, X_n] = X_n$$

instead. If the filtration isn't specified, we assume it's the natural one.

Theorem 12.1. If $\{X_n,\, n \in \mathbb{N}\}$ is a martingale, then $E[X_n] = E[X_0]$ for every $n \in \mathbb{N}$.

Proof. Since by the tower property of conditional expectations,

$$E[X_{n+1}] = E[E[X_{n+1}|\mathcal{F}_n]] = E[X_n],$$

we can use induction to conclude that

$$E[X_{n+1}] = E[X_n] = E[X_{n-1}] = \cdots = E[X_0]$$

as required. □

Example 12.2. Suppose that $S_n = \sum_{i=1}^{n} X_i$ is a one-dimensional simple, symmetric random walk. Then $E[S_n] = 0$ and $Var(S_n) = n$.

We now show that S_n is a martingale with respect to the natural filtration. To see that $\{S_n\}$ satisfies the first two conditions, note that S_n is bounded and so $S_n \in L^1$ for each n. Furthermore, S_n is by definition measurable with respect to $\mathcal{F}_n = \sigma(S_0, S_1, \ldots, S_n)$. As for the third condition, notice that

$$\begin{aligned} E[S_{n+1}|\mathcal{F}_n] &= E[X_{n+1} + S_n|\mathcal{F}_n] \\ &= E[X_{n+1}|\mathcal{F}_n] + E[S_n|\mathcal{F}_n]. \end{aligned}$$

Since X_{n+1} is independent of $X_1, X_2, \ldots, X_n$ we can use Theorem 11.10 to conclude that

$$E[X_{n+1}|\mathcal{F}_n] = E[X_{n+1}] = 0.$$

S_n is $\mathcal{F}_n$-measurable, so

$$E[S_n|\mathcal{F}_n] = S_n.$$

Therefore, $E[S_{n+1}|\mathcal{F}_n] = S_n$, which proves that $\{S_n, n = 0, 1, 2, \ldots\}$ is a martingale.

A quantity of great interest is the expected displacement of d-dimensional simple random walk. Unfortunately, it isn't too straightforward to compute $E[|S_n|]$. However, as the next proposition shows, martingale theory makes it easy to compute the expected *squared* displacement.

Proposition 12.9. If S_n is d-dimensional symmetric simple random walk, then

$$E[|S_n|^2] = n.$$

Proof. We just show that $M_n = |S_n|^2 - n$ is a martingale with respect to the filtration generated by S_n. Clearly $E[|M_n|] < \infty$ and M_n is $\mathcal{F}_n$-measurable.

$$\begin{aligned} E[M_{n+1}|\mathcal{F}_n] &= E[|S_{n+1}|^2 - (n+1)|\mathcal{F}_n] = E[|S_{n+1}|^2 - |S_n|^2 - ((n+1) - n)|\mathcal{F}_n] + |S_n|^2 - n] \\ &= E[|S_{n+1}|^2 - |S_n|^2 - 1|\mathcal{F}_n] + M_n \\ &= \frac{1}{2d}\sum_{i=1}^{d} \left((|S_n + e_i|^2 - |S_n|^2 - 1) + (|S_n - e_i|^2 - |S_n|^2 - 1)\right) + M_n \end{aligned}$$

Now it is easy to verify (just suppose $S_n = (x_1, \ldots, x_d)$) that regardless of the position of S_n,

$$(|S_n + e_i|^2 - |S_n|^2 - 1) + (|S_n - e_i|^2 - |S_n|^2 - 1) = 0.$$

This implies that $E[M_{n+1}|\mathcal{F}_n] = M_n$, so M_n is indeed a martingale. Theorem 12.1 now implies that $E[M_n] = E[M_0] = 0$, which proves the proposition. □

At this point, it may be worthwhile to look again at Kolmogorov extension in Lecture 5. We will be thinking of realizations of martingales as being determined by each ω. So now the random objects based on our probability space will be paths (equivalently, an infinite sequence of random variables). Moreover, we will sometimes need to throw in a few more random variables based on the same probability space, so thinking about all these objects being generated by one same ω will be particularly convenient.

12.2 Stopping Times

Definition 12.3. A function $T : \Omega \to \{0\} \cup \mathbb{N}$ is a *stopping time* (for the filtration $\mathcal{F}_n$) if

$$\{\omega : T(\omega) \leq n\} \in \mathcal{F}_n \, \forall \, n \in \{0\} \cup \mathbb{N} \cup \{\infty\}.$$

Proposition 12.10. T is a stopping time if and only if

$$\{\omega : T(\omega) = n\} \in \mathcal{F}_n \, \forall \, n \in \{0\} \cup \mathbb{N} \cup \{\infty\}.$$

Proof. Suppose T is a stopping time. Then

$$\{T = n\} = \{T \leq n\} \cap \{T \leq n-1\}^c \in \mathcal{F}_n,$$

since $\{T \leq n\} \in \mathcal{F}_n$ and $\{T \leq n-1\}^c \in \mathcal{F}_{n-1} \subset \mathcal{F}_n$.
Now suppose that $\{T = k\} \in \mathcal{F}_k$ for all $0 \leq k \leq n$. Then $\{T = k\} \in \mathcal{F}_n$ for all $0 \leq k \leq n$, so

$$\{T \leq n\} = \cup_{k=0}^n \{T = k\} \in \mathcal{F}_n.$$

□

Example 12.3. If X_n is an adapted process and $B \in \mathbb{R}$, then

$$T_1 = \inf\{n \geq 0 : X_n \in B\}$$

(with the convention $\inf \emptyset = \infty$) and

$$T_2 = n$$

are stopping times, while if $N \in \mathbb{N}$,

$$T_3 = \sup\{n \leq N : S_n \in B\}$$

is not. Indeed,

$$\{T_1 \leq n\} = \cup_{k=0}^n \{X_k \in B\} \in \mathcal{F}_n$$

and

$$\{T_2 \leq n\} = \Omega \in \mathcal{F}_n.$$

However, for $n < N$,

$$\{T_3 = n\} = \{X_n \in B\} \cap_{k=n+1}^{N} \{X_k \in B\}^c \notin \mathcal{F}_n.$$

Lemma 12.1. Suppose S and T are stopping times. Then $S \vee T$ and $S \wedge T$ are stopping times. In particular, $S \vee n$ and $S \wedge n$ are stopping times.

Proof. We prove that $S \wedge T$ is a stopping time and leave the other fact as an exercise. We first note that

$$\begin{aligned} \{S \wedge T = n\} &= \{S = n, T \geq n\} \cup \{S \geq n, T = n\} \\ &= (\{S = n\} \cap \{T \leq n-1\}^c) \cup (\{S \leq n-1\}^c \cap \{T = n\}) \end{aligned}$$

Since $\{T \leq n-1\}^c \in \mathcal{F}_{n-1} \subset \mathcal{F}_n, \{S \leq n-1\}^c \in \mathcal{F}_{n-1} \subset \mathcal{F}_n, \{T = n\} \in \mathcal{F}_n$, and $\{S = n\} \in \mathcal{F}_n$, we see that $\{S \wedge T = n\} \in \mathcal{F}_n$. □

12.3 Optional Stopping Theorem

Suppose that $\{S_n, n = 0, 1, 2, \ldots\}$ is a simple random walk starting from 0. As we just saw, $\{S_n^2 - n, n = 0, 1, 2, \ldots\}$ is a martingale. Hence,

$$E[S_n^2 - n] = E[S_0^2 - 0] = 0. \tag{17}$$

Suppose now that T is a stopping time. The question of whether or not we obtain an expression like (17) by replacing n with T is not an easy one. The one-word answer is "sometimes". The next example shows why it can't be "yes".

Example 12.4. Suppose that $\{S_n, n = 0, 1, 2, \ldots\}$ is a simple random walk starting from 0. Let $T = \inf\{k \geq 0 : S_k = 1\}$ denote the first time that the simple random walk reaches 1. Since S_n is a martingale, we know that

$$E[S_n] = E[S_0] = 0.$$

However, it turns out that $T < \infty$, a.s., so by definition, $S_T = 1$. Therefore,

$$E[S_T] = 1 \neq E[S_0].$$

In order to be able to claim $E[X_T] = E[X_0]$, we need some restrictions on T. The problem in the example above is that $E[T] = \infty$ (although $P(T < \infty) = 1$).

12.3.1 Martingale Transforms

Definition 12.4. A process $\{C_n\}_{n\in\mathbb{N}}$ is *previsible* if C_n is $\mathcal{F}_{n-1}$-measurable for all $n \geq 1$. If C is previsible and X is a supermartingale,

$$(C \bullet X)_n := \sum_{k=1}^{n} C_k(X_k - X_{k-1})$$

is the *martingale transform* of X.

Theorem 12.2. If $C \geq 0$ is previsible, there exists K such that $|C_n| \leq K\ \forall n \geq 1$, and X is a supermartingale, then $C \bullet X$ is a supermartingale . Moreover, if C is previsible, there exists K such that $|C_n| \leq K\ \forall n \geq 1$, and X is a martingale, then $C \bullet X$ is a martingale .

Proof. Let $Y_n = (C \bullet X)_n$. Then

$$E[Y_n - Y_{n-1}|\mathcal{F}_{n-1}] = E[C_n(X_n - X_{n-1})|\mathcal{F}_{n-1}] = C_n E[X_n - X_{n-1}|\mathcal{F}_{n-1}] \leq 0, \text{ a.s.},$$

where the last equality follows from the $\mathcal{F}_{n-1}$-measurability of C_n and Theorem 11.9 and the inequality follows from the fact that $C \geq 0$ and X is a super-martingale.

Now do the same thing again for a martingale to obtain the second statement. □

We saw earlier in Example 12.4 that one of the problems that can prevent a statement such as (17) from holding if n is replaced by a stopping time T is if $E[T] = \infty$. As has been the case all semester, whenever we want to deal with a random variable that is not L^1, we do the same thing: truncate.

Definition 12.5. Let T be a stopping time. Then the *process X stopped at T* is defined by

$$X_n^T(\omega) := X_{T(\omega)\wedge n}(\omega).$$

Theorem 12.3. If X is a supermartingale and T is a stopping time, then X^T is a supermartingale. In particular, $E[X_{T\wedge n}] \leq E[X_0]$.

If X is a martingale and T is a stopping time, then X^T is a martingale. In particular, $E[X_{T\wedge n}] = E[X_0]$.

Proof. Define $C_n^T(\omega) := \mathbb{1}\{T(\omega) \geq n\}$. Then

$$(C^T \bullet X)_n = \sum_{k=1}^{n} C_k^T(X_k - X_{k-1}) = X_{T\wedge n} - X_0.$$

Therefore, the processes $C^T \bullet X$ and $X^T - X_0$ are equal. In particular, since $C^T \geq 0$ is bounded for all n and

$$\{C_n^T = 0\} = \{T \leq n-1\} \in \mathcal{F}_{n-1},$$

so that C^T is previsible, Theorem 12.2 implies that if X is a supermartingale, so is $C^T \bullet X$. So X^T is a supermartingale.

Now do the same thing for a martingale. □

Theorem 12.4 (Doob's Optional Stopping Theorem)**.** Suppose that $\{X_n, n = 0, 1, 2, \ldots\}$ is a supermartingale and T is a stopping time. Then $E[|X_T|] < \infty$ and

$$E[X_T] \leq E[X_0]$$

if either of the following is satisfied:

1. T is bounded.
2. X is bounded and T is almost surely finite.
3. $E[T] < \infty$ and there exists $K > 0$ such that for all $\omega \in \Omega$, all $n \geq 1$,

$$|X_n(\omega) - X_{n-1}(\omega)| \leq K.$$

If either 1.-3. holds and X is a martingale, then $E[X_T] = E[X_0]$.

Note 12.2. As expected given the conclusion of the theorem, the random walk in Example (12.4) doesn't satisfy any of the three hypotheses.

Proof. Since for all n, X_n is L^1, so is $X_{T\wedge n}$, by Theorem 12.3, and for all $n \in \mathbb{Z}^+$,

$$E[X_{T\wedge n}] \leq E[X_0]. \tag{18}$$

If T is bounded, there exists $N < \infty$ such that $T(\omega) \leq N \forall \omega \in \Omega$, so

$$E[X_T] = E[X_{T\wedge N}] \leq E[X_0].$$

This proves 1.

If $T < \infty$, a.s., $X_{T\wedge n} \to X_T$, a.s., so if X is bounded, $E[X_{T\wedge n}] \to E[X_T]$ by BCT and we can take limits in (18) to prove 2.

Finally, if $E[T] < \infty$, then T is almost surely finite, so $X_{T\wedge n} \to X_T$, a.s. Also, if for all $\omega \in \Omega$, all $n \geq 1, |X_n(\omega) - X_{n-1}(\omega)| \leq K$,

$$|X_{T\wedge n} - X_0| = \sum_{k=1}^{T\wedge n} (X_k - X_{k-1})| \leq KT.$$

Since $E[KT] < \infty$, we can use the dominated convergence theorem to take limits in (18) and thus prove 3.

If X is a submartingale satisfying one of the conditions 1.-3., then $-X$ is a supermartingale satisfying one of the conditions 1.-3. So $E[-X_T] \leq E[-X_0]$, implying that $E[X_T] \geq E[X_0]$. Therefore, if X is a martingale (thus a submartingale and a supermartingale) satisfying one of the conditions 1.-3., then $E[X_T] = E[X_0]$. □

12.4 Maximal Inequalities

The Optional Sampling Theorem has immediate implications concerning the pathwise behavior of martingales, submartingales, and supermartingales. The most elementary of these concern the maxima of the sample paths, and so are called maximal inequalities.

Proposition 12.11. Let $\{X_n\}_{n\geq 0}$ be a sequence of random variables, and for each $n \geq 0$, define

$$M_n = \max_{0\leq m\leq n} X_m,$$

and

$$M_\infty = \sup_{0\leq m} X_m = \lim_{n\to\infty} M_n.$$

Then for any $\alpha > 0$ and $n \geq 1$,

- if $\{X_n\}_{n\geq 0}$ is a submartingale,

$$P(M_n \geq \alpha) \leq E[X_n \vee 0]/\alpha,$$

- if $\{X_n\}_{n\geq 0}$ is a nonnegative supermartingale.

$$P(M_\infty \geq \alpha) \leq E[X_0]/\alpha$$

Proof. Assume that $\{X_n\}_{n\geq 0}$ is a submartingale. Without loss of generality, we may assume that each $X_n \geq 0$, because if not we may replace the original submartingale X_n by the larger submartingale $X_n \vee 0$. Define $\tau = \inf\{n \geq 0 : X_n \geq \alpha\}$ to be the first time at which $X_n \geq \alpha$, with $\inf \emptyset = \infty$. Then for any nonrandom $n \geq 0$, the truncation $\tau \wedge n$ is a stopping time and so, by the Optional Sampling Theorem,

$$E[X_{\tau\wedge n}] \leq E[X_n].$$

But because the random variables X_m are nonnegative, and because $X_{\tau\wedge n} \geq \alpha$ on the event $\{\tau \leq n\}$,

$$E[X_n] \geq E[X_{\tau\wedge n}] \geq E[X_{\tau\wedge n}\mathbb{1}\{\tau \leq n\}] \geq \alpha E[\mathbb{1}\{\tau \leq n\}] = \alpha P(\tau \leq n) = \alpha P(M_n \geq \alpha).$$

This proves the first inequality. The proof of the second inequality is similar, but needs an additional limiting argument. First, for any finite $n \geq 0$, the same argument as in the first part of the proof shows that $P(M_n \geq \alpha) \leq E[X_0]/\alpha$. Now the random variables M_n are nondecreasing in n, and converge up to M_∞, so for any $\epsilon > 0$, the event $\{M_\infty \geq \alpha\}$ is contained in the event $\{M_n \geq \alpha-\epsilon\}$ for some n. But by the last inequality and the monotone convergence theorem, the probability of this is no larger than $E[X_0]/(\alpha - \epsilon)$. Since $\epsilon > 0$ may be taken arbitrarily small, the second inequality of the theorem follows. □

12.5 Martingale Convergence Theorem

The following is one of the important results from martingale theory. We'll see the proof next week.

Theorem 12.5. (Martingale Convergence) Let X be a supermartingale bounded in L^1. Then

$$X_\infty = \lim_{n\to\infty} X_n < \infty, a.s.$$

Lecture #13: Martingale Convergence

Reference. Sections 5.2, 5.3

13.1 Martingale Convergence Theorem

If we put some fairly gentle restrictions on martingales, they often end up converging almost surely. For this to happen (for any given ω), they can't wiggle much forever. In other words, given any two horizontal lines, the martingale path determined by ω can't cross these lines infinitely often if it is to converge. It is therefore natural to study the number of times a martingale path moves from below a line to above a higher line. This motivates the following definition:

Definition 13.1. Given $a, b \in \mathbb{R}$ with $a < b$ and a sequence of random variables $\{X_n\}_{n\geq 0}$, we define $U_N = U_N[a,b](\omega)$, the number of *upcrossings* of $[a,b]$ made by the path $X(\omega)$ by time N to be

$$\max\{k \geq 0 : \exists 0 \leq s_1 < t_1 < \cdots < s_k < t_k \leq N \text{ with } X_{s_i}(\omega) < a, X_{t_i}(\omega) > b \, \forall 1 \leq i \leq k\}.$$

Since for every ω, $\{U_N(\omega)\}_N$ is a nondecreasing sequence, we can define

$$U_\infty[a,b](\omega) := \lim_{N\to\infty} U_N[a,b](\omega).$$

The goal will be to show that under some conditions on a martingale X, for any a, b and almost all ω, $U_N[a,b](\omega)$ is finite.

For $x \in \mathbb{R}$, let $x^- := \max\{-x, 0\}$.

Lemma 13.1. (Doob's Upcrossing Inequality) Suppose X is a supermartingale. Then for all $a, b \in \mathbb{R}$ with $a < b$, $N \in \mathbb{N}$,

$$(b-a)E[U_N[a,b]] \leq E[(X_N - a)^-].$$

Proof. Define a sequence $\{C_i\}_{i\geq 1}$ of random variables based on $\{X_i\}_{i\geq 1}$ as follows: Let $C_1 := \mathbb{1}\{X_0 < a\}$ and for $n \geq 2$, define

$$C_n := \mathbb{1}\{C_{n-1} = 1\}\mathbb{1}\{X_{n-1} \leq b\} + \mathbb{1}\{C_{n-1} = 0\}\mathbb{1}\{X_{n-1} < a\}.$$

Then clearly, C is ≥ 0 and bounded (since $C = 0$ or 1) and previsible (by induction, since C_1 is $\mathcal{F}_0$-measurable and C_n is determined by C_{n-1} and X_{n-1}).

Now define $Y := C \bullet X$ (recall $(C \bullet X)_n := \sum_{k=1}^{n} C_k(X_k - X_{k-1})$). Then

$$Y_N(\omega) \geq (b-a)U_N[a,b](\omega) - (X_N - a)^-.$$

The properties of C and Theorem 12.2 imply that Y is a supermartingale. Therefore, $E[Y_N] \leq E[Y_1] = 0$. Thus,

$$0 \geq E[Y_N] \geq (b-a)E[U_N[a,b]] - E[(X_N - a)^-].$$

□

Corollary 13.1. Suppose X is a supermartingale bounded in L^1, that is, such that there exists $K < \infty$, such that $E[|X_n|] \leq K$ for all $n \geq 0$. Consider $a, b \in \mathbb{R}$ with $a < b$. Then

$$(b-a)E[U_\infty[a,b]] \leq |a| + \sup_n E[|X_n|] < \infty,$$

and so

$$P(U_\infty[a,b] = \infty) = 0.$$

Proof. The upcrossing inequality implies, for $N \in \mathbb{N}$,

$$(b-a)E[U_N[a,b]] \leq E[|X_N|] + |a|.$$

Using the monotone convergence theorem, we can let $N \to \infty$ to get the statement of the lemma.

And of course, random variables with finite expectation must be finite with probability 1. □

For a process X defined on $(\Omega, \mathcal{F}, P)$, we define X_∞ by $X_\infty(\omega) = \limsup X_n(\omega)$ for all $\omega \in \Omega$. Then X_∞ is $\mathcal{F}_\infty$-measurable.

Theorem 13.1. (Martingale Convergence) Let X be a supermartingale bounded in L^1. Then

$$X_\infty = \lim_{n\to\infty} X_n < \infty, a.s.$$

Proof. For $a, b \in \mathbb{Q}$ with $a < b$, define

$$\Lambda_{a,b} = \{\omega : \liminf X_n(\omega) < a < b < \limsup X_n(\omega)\}$$

and

$$\Lambda = \{\omega : \liminf X_n(\omega) < \limsup X_n(\omega)\}.$$

Then

$$\Lambda = \bigcup_{\{a,b\in\mathbb{Q}, a<b\}} \Lambda_{a,b}.$$

Since

$$\Lambda_{a,b} \subseteq \{\omega : U_\infty[a,b](\omega) = \infty\},$$

Corollary 13.1 implies that $P(\Lambda_{a,b}) = 0$, so that $P(\Lambda) = 0$. This proves that almost surely, $X_\infty = \lim_{n\to\infty} X_n$. Moreover,

$$E[|X_\infty|] = E[\liminf_{n\to\infty} |X_n|] \leq \liminf_{n\to\infty} E[|X_n|] < \infty,$$

by Fatou's lemma and since X is bounded in L^1. Therefore, $P(X < \infty) = 1$. □

Corollary 13.2. If X is a positive super-martingale, then

$$X_\infty = \lim_{n\to\infty} X_n < \infty, a.s.$$

Proof. If X is a positive super-martingale, then

$$E[|X_n|] = E[X_n] \leq E[X_0] = E[|X_0|] < \infty,$$

since for a super-martingale X, $E[|X_n|] < \infty$ for all n. So X is bounded in L^1 and we can apply Theorem 13.1 □

13.2 Gambler's Ruin for Random Walk

As we will see below, with the help of the optional stopping theorem it is fairly easy to compute the probability that one-dimensional (not necessarily symmetric) simple random walk started at 0 reaches integer x before integer y. Of course, this probability is nontrivial only if $xy < 0$.

First we find two martingales based on random walk, one of which will show that the stopping times we care about in the gambler's ruin formula are almost surely finite.

Lemma 13.2. Suppose $\{X_i\}_{i\geq 1}$ is a sequence of independent random variables with $X_i \geq 0$ and $E[X_i] = 1$ for all $i \geq 1$. Then if $M_0 = 1, \mathcal{F}_0 = \{\emptyset, \Omega\}, \mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, and

$$M_n = \prod_{i=1}^{n} X_i,$$

then M is a martingale with respect to $\{\mathcal{F}_n\}$.

Proof. It is clear that M_n is $\mathcal{F}_n$-measurable. Also, since the X_i are independent and positive,

$$E[|M_n|] = E[M_n] = \prod_{i=1}^{n} E[X_i] = 1 < \infty.$$

For $n \geq 1$,

$$E[M_n|\mathcal{F}_{n-1}] = E[M_{n-1}X_n|\mathcal{F}_{n-1}] = M_{n-1}E[X_n|\mathcal{F}_{n-1}] = M_{n-1}E[X_n] = M_{n-1}.$$

□

Lemma 13.3. 1. Suppose $\{X_i\}$ are independent random variables with $P(X_i = 1) = P(X_i = -1) = 1/2$. Then if $S_n = \sum_{i=1}^{n} X_i, \mathcal{F}_0 = \{\emptyset, \Omega\}, \mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, and for $t \in \mathbb{R}, M_0^t = 1$ and

$$M_n^t = \frac{e^{tS_n}}{\cosh^n t},$$

is a martingale with respect to $\{\mathcal{F}_n\}$.

2. Suppose $\{X_i\}$ are independent random variables with $P(X_i = 1) = 1 - P(X_i = -1) = p$. Then if $S_n = \sum_{i=1}^{n} X_i, \mathcal{F}_0 = \{\emptyset, \Omega\}, \mathcal{F}_n = \sigma(X_1, \ldots, X_n), M_0 = 1$, and for $n \geq 1$,

$$M_n = \left(\frac{1-p}{p}\right)^{S_n}$$

is a martingale with respect to $\{\mathcal{F}_n\}$.

Proof. 1.

$$M_n^t = \prod_{i=1}^{n} \frac{e^{tX_i}}{\cosh t}.$$

Since $\frac{e^{tX_i}}{\cosh t} \geq 0$ and $E[\frac{e^{tX_i}}{\cosh t}] = \frac{e^t + e^{-t}}{2\cosh t} = 1$, Lemma 13.2 implies that M_n^t is a martingale.

2. Since $\left(\frac{1-p}{p}\right)^{X_i} \geq 0$ and

$$M_n = \prod_{i=1}^{n} \left(\frac{1-p}{p}\right)^{X_i},$$

the fact that

$$E[X_i] = p\left(\frac{1-p}{p}\right)^{1} + (p-1)\left(\frac{1-p}{p}\right)^{-1} = 1,$$

together with Lemma 13.2, implies that M is a martingale.

□

Proposition 13.12. If S is one-dimensional symmetric simple random walk with $S_0 = 0$, $x, y \in \mathbb{N}$, and $T_1 = \inf\{n \geq 0 : S_n \notin (-y, x)\}, T_2 = \inf\{n \geq 0 : S_n \geq x\}$, then

$$P(T_1 < \infty) = P(T_2 < \infty) = 1.$$

Proof. We prove that $P(T_2 < \infty) = 1$. The result for T_1 follows since for all $\omega, T_1 \leq T_2$. We will write T for T_2.

Let $t \geq 0$ and consider $M_n^t = \frac{e^{tS_n}}{\cosh^n t}$. Since $T \wedge n$ is a bounded stopping time for S_n and thus for M_n^t, $0 \leq T$ implies

$$E[M_{T\wedge n}^t] = 1.$$

Since $\cosh t \geq 1$ and $e^{tS_{T\wedge n}} \leq e^{tx}, M_{T\wedge n}^t \leq e^{tx}$. Now as $n \to \infty, M_{T\wedge n}^t \to M_T^t$ (with $M_T^t(\omega) := 0$ if $T(\omega) = \infty$; this definition is the only sensible one since if $T = \infty$, $M_{T\wedge n}^t = M_n^t \leq \frac{e^{tx}}{\cosh^n t} \overset{n\to\infty}{\to} 0$).

By the bounded convergence theorem,

$$1 = E[M_T^t] = E\left[\frac{e^{tS_T}}{(\cosh t)^T}\right] = E\left[\frac{e^{xt}}{\cosh^T t}\right].$$

Therefore,

$$E\left[\frac{1}{\cosh^T t}\right] = e^{-xt} \,\forall\, t > 0.$$

If we let $t \downarrow 0$,

$$\frac{1}{\cosh^T t} \to \mathbb{1}\{T < \infty\}.$$

Therefore,

$$1 = \lim_{t\to 0} E\left[\frac{1}{\cosh^T t}\right] = \lim_{t\to 0} E[\mathbb{1}\{T < \infty\}] = P(T < \infty).$$

□

Theorem 13.2. (Gambler's ruin formula) Suppose that $\{S_n, n = 0, 1, 2, \ldots\}$ is a one-dimensional simple random walk started at $x \in \mathbb{N} \cup \{0\}$. That is, if $\{X_i\}_{i\geq 1}$ is a sequence of independent random variables with $P(X_i = 1) = 1 - P(X_i = -1) = p \in (0, 1)$, we define $S_0 = x$ and for $n \geq 1, S_n = S_0 + \sum_{i=1}^{n} X_i$. Let $N \geq x$ and $T = \inf\{n \geq 0 : S_n \in \{0, N\}\}$. (Note that we showed that if $p = 1/2$, this is almost surely finite; the same holds if $p \neq 1/2$.)

- If $p = 1/2$, then
$$P\{S_T = N\} = \frac{x}{N}.$$
- If $p \in (0, 1) \setminus \{1/2\}$, then
$$P\{S_T = N\} = \frac{1 - \left(\frac{1-p}{p}\right)^x}{1 - \left(\frac{1-p}{p}\right)^N}.$$

Proof. Suppose first that $p = 1/2$.

Note that $|S_{n\wedge T}| \leq N$ for all n and that this stopping time T satisfies $P(T < \infty) = 1$. Therefore, since S_n is a martingale, Theorem 12.3 implies that $S_{n\wedge T}$ is a bounded martingale, so by the optional stopping theorem (for a bounded martingale with almost surely finite stopping time),

$$E[S_T] = \lim_{n\to\infty} E[S_{T\wedge n}] = \mathbb{E}[S_0] = x.$$

Since

$$E[S_T] = 0 \cdot P\{S_T = 0\} + N \cdot P\{S_T = N\},$$

we get

$$N \cdot P\{S_T = N\} = x \quad \Longleftrightarrow \quad P\{S_T = N\} = \frac{x}{N}$$

Suppose now that $p \in (0, 1/2) \cup (1/2, 1)$.

We know that

$$Z_n = \left(\frac{1-p}{p}\right)^{S_n}$$

is a martingale. Therefore, by Theorem 12.3, $Z_{n\wedge T}$ is a bounded martingale
The optional stopping theorem implies $E[Z_T] = \lim_{n\to\infty} E[Z_{n\wedge T}] = E[Z_0]$ and so

$$\left(\frac{1-p}{p}\right)^x = E\left[\left(\frac{1-p}{p}\right)^{S_T}\right].$$

Now,

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1-p}{p}\right)^{S_T}\right] &= \left(\frac{1-p}{p}\right)^0 P\{S_T = 0\} + \left(\frac{1-p}{p}\right)^N P\{S_T = N\} \\ &= P\{S_T = 0\} + \left(\frac{1-p}{p}\right)^N (1 - P\{S_T = 0\})\end{aligned}$$

implying that

$$\left(\frac{1-p}{p}\right)^x = P\{S_T = 0\} + \left(\frac{1-p}{p}\right)^N (1 - P\{S_T = 0\}).$$

Solving for $P\{S_T = 0\}$ gives

$$P\{S_T = 0\} = \frac{\left(\frac{1-p}{p}\right)^x - \left(\frac{1-p}{p}\right)^N}{1 - \left(\frac{1-p}{p}\right)^N}.$$

We therefore find that

$$P\{S_T = N\} = 1 - \frac{\left(\frac{1-p}{p}\right)^x - \left(\frac{1-p}{p}\right)^N}{1 - \left(\frac{1-p}{p}\right)^N} = \frac{1 - \left(\frac{1-p}{p}\right)^x}{1 - \left(\frac{1-p}{p}\right)^N}$$

□

Example 13.1. Consider the standard North American roulette game that has 38 numbers of which 18 are black, 18 are red, and 2 are green. A bet on "black" pays even money. That is, a bet of \$1 on black pays \$1 plus your original \$1 if black does, in fact, appear. However, the probability that you win on a bet of "black" is $p = 18/38$. Suppose that $x = 100$ and $N = 200$. Then

$$P(S_T = 200) \approx 7 \cdot 10^{-10}.$$

13.3 Examples of convergent martingales

Example 13.2. (Branching Process) The Galton-Watson process is a common model for the evolution of a population where every individual generates k offspring ($k \in \mathbb{N} \cup \{0\}$) according to a distribution which is the same for all individuals. One of the key questions is whether a population will go extinct or not. As one might expect, the answer depends on the distribution of offspring.

Definition 13.2. If $\{X_i\}_{i,n\in\mathbb{N}}$ is a family of independent nonnegative integer-valued random variables, let $\{Z_n\}_{n\geq 0}$ be defined by $Z_0 = 1$ and for $n \geq 1$,

$$Z_{n+1} = \mathbb{1}\{Z_n > 0\}\sum_{i=1}^{Z_n} X_{i,n+1}.$$

Z is called a *Galton-Watson process.* $p_k = P(X_{i,n} = k)$ is the *offspring distribution.*

Lemma 13.4. If $\mathcal{F}_n = \sigma(X_{i,m} : i \geq 1, 1 \leq m \leq n)$ and $\mu = E[X_{i,m}] \in (0,\infty)$, then $\frac{Z_n}{\mu^n}$ is a martingale with respect to $\mathcal{F}_n$.

Proof. Clearly, $Z_n \in \mathcal{F}_n$. Linearity and the monotone convergence theorem imply the first of the four equalities in

$$\begin{aligned} E[Z_{n+1}|\mathcal{F}_n] &= \sum_{k=1}^{\infty} E[Z_{n+1}\mathbb{1}\{Z_n = k\}|\mathcal{F}_n] \\ &= \sum_{k=1}^{\infty} E[(X_{1,n+1} + \cdots + X_{k,n+1})\mathbb{1}\{Z_n = k\}|\mathcal{F}_n] = \sum_{k=1}^{\infty} \mathbb{1}\{Z_n = k\}k\mu = \mu E[Z_n], \end{aligned}$$

while the penultimate follows from the fact that $X_{i,n+1}$ is independent of $\mathcal{F}_n$ and that $\mathbb{1}\{Z_n = k\}$ is $\mathcal{F}_n$-measurable. Therefore,

$$E\left[\frac{Z_{n+1}}{\mu^{n+1}}\right] = \frac{Z_n}{\mu_n},$$

so Z_n/μ^n is a martingale. □

Corollary 13.3.

$$\frac{Z_n}{\mu^n} \text{ converges almost surely.}$$

Proof. Z_n/μ^n is a non-negative martingale, so by the martingale convergence theorem, it converges. □

Theorem 13.3. If $\mu < 1, Z_n \to 0$, almost surely.

Proof. $E[Z_n/\mu^n] = E[Z_0] = 1$, so

$$P(Z_n > 0) \leq E[Z_n\mathbb{1}\{Z_n > 0\}] = E[Z_n] = \mu^n \to 0.$$

Since for all $i \geq 0, Z_{n+i} = 0$ if $Z_n = 0$, we see that $Z_n \to 0$, almost surely. □

Theorem 13.4. If $\mu = 1$ and $P(X_{i,m} = 1) < 1$, then $Z_n \to 0$, almost surely.

Proof. If $\mu = 1, Z_n$ is a nonnegative martingale, so $Z_n \to Z_\infty$, a.s. For all $k > 0$,

$$P(\exists\, N > 0 \text{ s.t. } Z_n = k\, \forall\, n \geq N) = 0, \tag{19}$$

so $P(Z_\infty = k) = 0$, implying that $P(Z_\infty = 0) = 1$. Note that (19) is true for the following reason:

$$P(Z_n = k \,\forall\, n \geq N) = \prod_{n \geq N} P(Z_{n+1} = k | Z_n = k) = 0,$$

since $P(Z_{n+1} = k | Z_n = k) < 1$ (this follows from the fact that $\mu = 1$ and $P(X_{i,m} = 1) < 1$, which implies that there exists $r > 1$ such that $P(X_{i,m} = 1) > 0$, so that $P(Z_{n+1} = k | Z_n = k) < 1$). □

If $\mu > 1$, the probability of eternal survival of the species is greater than 0. That is,

$$P(Z_n > 0 \,\forall\, n > 0) > 0.$$

For details, see Section 5.3.3 in Durrett.

Example 13.3. (A bounded martingale with nontrivial limit) The previous example is an example of a very practical application of the martingale convergence theorem. However it leaves us with a pressing question: Can a martingale converge to something else than a trivial distribution? This example gives a somewhat artificial martingale, but shows that the limiting distribution can be interesting. Consider the sequence $\{X_i\}_{i \geq 1}$ of independent random variables whose distribution is as follows:

$$P(X_i = \frac{1}{2^{i+1}}) = P(X_i = -\frac{1}{2^{i+1}}) = \frac{1}{2},$$

$S_0 = \frac{1}{2}$, almost surely, and $S_n = S_0 + \sum_{i=1}^{n} X_i$. Then, if $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, clearly S_n is $\mathcal{F}_n$-measurable and $E[|S_n|] < \infty$, since $|S_n| < \infty$ almost surely. Also, since for every $n \geq 1, E[X_n] = 0$, $E[S_{n+1} | \mathcal{F}_n] = S_n + E[X_n] = S_n$.

Therefore, by the martingale convergence theorem, S_n must converge almost surely. We will now determine its limit by showing that if $a < b$ are dyadic rationals, then

$$P(S_\infty \in [a, b)) = b - a. \tag{20}$$

In particular, it will follow that if $A = \cup_{i=1}^{n} [a_i, b_i)$, where $0 \leq a_1 < b_1 < a_2 < b_2 < \cdots < a_n < b_n < 1, P(S_\infty \in A) = \mathcal{L}(A)$, where $\mathcal{L}$ is Lebesgue measure.

Since the family of sets of the same form as A is a π-system which generates the Borel sets, on which $P(S_\infty \in \cdot)$ agrees with Lebesgue measure, Theorem 4.5 implies that $P(S_\infty \in B) = \mathcal{L}(B)$ for all $B \in \mathcal{R}$.

Suppose a and b are two rationals with $-1 \leq a < b \leq 1$, of the form

$$a = \sum_{i \geq 1} \frac{a_i}{2^i} \text{ and } b = \sum_{i \geq 1} \frac{b_i}{2^i},$$

where $a_i, b_i \in \{-1, 1\}$.

We now prove equation (20):

It is easy to see that for $n \geq 1, c = 0, \ldots, 2^n - 1$,

$$S_\infty(\omega) \in \left[\frac{c}{2^n}, \frac{c+1}{2^n}\right] \iff S_n(\omega) = \frac{2c+1}{2^{n+1}}.$$

Therefore,

$$P\left(S_\infty(\omega) \in \left[\frac{c}{2^n}, \frac{c+1}{2^n}\right]\right) = P\left(S(n) = \frac{2c+1}{2^{n+1}}\right) = \frac{1}{2^n}. \tag{21}$$

Now suppose

$$a = \frac{c}{2^m} \text{ and } b = \frac{d}{2^n},$$

and suppose without loss of generality that $m < n$. Then the interval $[a, b)$ consists of $d - c2^{n-m}$ disjoint intervals of width $\frac{1}{2^n}$, that is,

$$\big[a, b\big) = \bigcup_{i=1}^{d-c2^{n-m}} \left[a + \frac{i-1}{2^n}, a + \frac{i}{2^n}\right).$$

In particular, it follows from (21) that

$$P(S_\infty \in [a, b)) = \sum_{i=1}^{d-c2^{n-m}} \frac{1}{2^n} = \big(d - c2^{n-m}\big)\frac{1}{2^n} = \frac{d}{2^n} - \frac{c}{2^m} = b - a.$$

The argument above now shows that S_∞ has the uniform distribution on $(0, 1)$.

Example 13.4. (Another bounded martingale for the road) Consider the sequence of random variables $X_0, X_1, \ldots$ defined by $X_0 = \frac{1}{2}$, a.s., and for $n \geq 1$,

$$P(X_{n+1} = \frac{X_n}{2}) = 1 - X_n,$$

$$P(X_{n+1} = \frac{1}{2} + \frac{X_n}{2}) = X_n.$$

Then X_n is a bounded martingale which therefore converges (one could also just invoke the fact that $X_n \geq 0$ for all $n \geq 0$). The symmetry of this martingale indicates that X_∞ should be symmetric. Now suppose that for some $x \in (0, 1), X_\infty(\omega) = x$ (say $x > \frac{1}{2}$). Then for every $\epsilon > 0$, there exists $N > 0$ such that for all $n \geq N, |X_n(\omega) - x| < \epsilon$. The same argument as in the previous example shows that this happens only with probability zero (for all n, X_{n+1} has probability bounded below to equal $X_n/2$, so eventually, this will have to happen, contradicting convergence to x). Therefore, the only values X_n can converge to are 1 and 0. Since X_∞ is symmetric, we get that

$$P(X_\infty = 1) = P(X_\infty = 0) = \frac{1}{2}.$$