Probability Theory

Richard F. Bass

Revised 2001.

1. Basic notions.

A *probability* or *probability measure* is a measure whose total mass is one. Because the origins of probability are in statistics rather than analysis, some of the terminology is different. For example, instead of denoting a measure space by $(X, \mathcal{A}, \mu)$, probabilists use $(\Omega, \mathcal{F}, \mathbb{P})$. So here Ω is a set, $\mathcal{F}$ is called a *σ-field* (which is the same thing as a σ-algebra), and $\mathbb{P}$ is a measure with $\mathbb{P}(\Omega) = 1$. Elements of $\mathcal{F}$ are called *events*. Elements of Ω are denoted ω.

Instead of saying a property occurs almost everywhere, we talk about properties occurring *almost surely*, written *a.s.*. Real-valued measurable functions from Ω to $\mathbb{R}$ are called *random variables* and are usually denoted by X or Y or other capital letters. We often abbreviate "random variable" by *r.v.*

We let $A^c = (\omega \in \Omega : \omega \notin A)$ (called the *complement* of A) and $B - A = B \cap A^c$.

Integration (in the sense of Lebesgue) is called *expectation* or *expected value*, and we write $\mathbb{E}\, X$ for $\int X d\mathbb{P}$. The notation $\mathbb{E}\,[X; A]$ is often used for $\int_A X d\mathbb{P}$.

The random variable 1_A is the function that is one if $\omega \in A$ and zero otherwise. It is called the *indicator* of A (the name characteristic function in probability refers to the Fourier transform). Events such as $(\omega : X(\omega) > a)$ are almost always abbreviated by $(X > a)$.

Given a random variable X, we can define a probability on $\mathbb{R}$ by

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A), \qquad A \subset \mathbb{R}. \tag{1.1}$$

The probability $\mathbb{P}_X$ is called the *law* of X or the *distribution* of X. We define $F_X : \mathbb{R} \to [0, 1]$ by

$$F_X(x) = \mathbb{P}_X((-\infty, x]) = \mathbb{P}(X \le x). \tag{1.2}$$

The function F_X is called the *distribution function* of X.

As an example, let $\Omega = \{H, T\}$, $\mathcal{F}$ all subsets of Ω (there are 4 of them), $\mathbb{P}(H) = \mathbb{P}(T) = \frac{1}{2}$. Let $X(H) = 1$ and $X(T) = 0$. Then $\mathbb{P}_X = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, where δ_x is point mass at x, that is, $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise. $F_X(a) = 0$ if $a < 0$, $\frac{1}{2}$ if $0 \le a < 1$, and 1 if $a \ge 1$.

Proposition 1.1. *The distribution function F_X of a random variable X satisfies:*

(a) F_X is nondecreasing;

(b) F_X is right continuous with left limits;

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

Proof. We prove the first part of (b) and leave the others to the reader. If $x_n \downarrow x$, then $(X \le x_n) \downarrow (X \le x)$, and so $\mathbb{P}(X_{\le} x_n) \downarrow \mathbb{P}(X \le x)$ since $\mathbb{P}$ is a measure. □

Note that if $x_n \uparrow x$, then $(X \le x_n) \uparrow (X < x)$, and so $F_X(x_n) \uparrow \mathbb{P}(X < x)$.

Any function $F : \mathbb{R} \to [0, 1]$ satisfying (a)-(c) of Proposition 1.1 is called a distribution function, whether or not is comes from a random variable.

Proposition 1.2. *Suppose F is a distribution function. There exists a random variable X such that $F = F_X$.*

Proof. Let $\Omega = [0,1]$, $\mathcal{F}$ the Borel σ-field, and $\mathbb{P}$ Lebesgue measure. Define $X(\omega) = \sup\{x : F(x) < \omega\}$. It is routine to check that $F_X = F$. □

In the above proof, essentially $X = F^{-1}$. However F may have jumps or be constant over some intervals, so some care is needed in defining X.

Certain distributions or laws are very common. We list some of them.

(a) *Bernoulli.* A random variable is Bernoulli if $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$ for some $p \in [0,1]$.

(b) *Binomial.* This is defined by $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, where n is a positive integer, $0 \le k \le n$, and $p \in [0,1]$.

(c) *Geometric.* For $p \in (0,1)$ we set $\mathbb{P}(X = k) = (1-p)p^k$. Here k is a nonnegative integer.

(d) *Poisson.* For $\lambda > 0$ we set $\mathbb{P}(X = k) = e^{-\lambda}\lambda^k/k!$ Again k is a nonnegative integer.

(e) *Uniform.* For some positive integer n, set $\mathbb{P}(X = k) = 1/n$ for $1 \le k \le n$.

If F is absolutely continuous, we call $f = F'$ the *density* of F. Some examples of distributions characterized by densities are the following.

(f) *Uniform on $[a,b]$.* Define $f(x) = (b-a)^{-1} 1_{[a,b]}(x)$. This means that if X has a uniform distribution, then

$$\mathbb{P}(X \in A) = \int_A \frac{1}{b-a} 1_{[a,b]}(x)\, dx.$$

(g) *Exponential.* For $x > 0$ let $f(x) = \lambda e^{-\lambda x}$.

(h) *Standard normal.* Define $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. So

$$\mathbb{P}(X \in A) = \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2} dx.$$

(i) $\mathcal{N}(\mu, \sigma^2)$. We shall see later that a standard normal has mean zero and variance one. If Z is a standard normal, then a $\mathcal{N}(\mu, \sigma^2)$ random variable has the same distribution as $\mu + \sigma Z$. It is an exercise in calculus to check that such a random variable has density

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}. \tag{1.3}$$

(j) *Cauchy.* Here

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

We can use the law of a random variable to calculate expectations.

Proposition 1.3. *If g is bounded or nonnegative, then*

$$\mathbb{E}\, g(X) = \int g(x)\, \mathbb{P}_X(dx).$$

Proof. If g is the indicator of an event A, this is just the definition of $\mathbb{P}_X$. By linearity, the result holds for simple functions. By the monotone convergence theorem, the result holds for nonnegative functions, and by linearity again, it holds for bounded g. □

If F_X has a density f, then $\mathbb{P}_X(dx) = f(x)\,dx$. So, for example, $\mathbb{E}\,X = \int xf(x)\,dx$ and $\mathbb{E}\,X^2 = \int x^2 f(x)\,dx$. (We need $\mathbb{E}\,|X|$ finite to justify this if X is not necessarily nonnegative.)

We define the *mean* of a random variable to be its expectation, and the *variance* of a random variable is defined by

$$\operatorname{Var} X = \mathbb{E}\,(X - \mathbb{E}\,X)^2.$$

For example, it is routine to see that the mean of a standard normal is zero and its variance is one.

Note

$$\operatorname{Var} X = \mathbb{E}\,(X^2 - 2X\mathbb{E}\,X + (\mathbb{E}\,X)^2) = \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2.$$

Another equality that is useful is the following.

Proposition 1.4. *If $X \geq 0$ a.s. and $p > 0$, then*

$$\mathbb{E}\,X^p = \int_0^\infty p\lambda^{p-1}\mathbb{P}(X > \lambda)\,d\lambda.$$

The proof will show that this equality is also valid if we replace $\mathbb{P}(X > \lambda)$ by $\mathbb{P}(X \geq \lambda)$.

Proof. Use Fubini's theorem and write

$$\int_0^\infty p\lambda^{p-1}\mathbb{P}(X > \lambda)\,d\lambda = \mathbb{E}\int_0^\infty p\lambda^{p-1}1_{(\lambda,\infty)}(X)\,d\lambda = \mathbb{E}\int_0^X p\lambda^{p-1}\,d\lambda = \mathbb{E}\,X^p.$$

□

We need two elementary inequalities.

Proposition 1.5. *Chebyshev's inequality If $X \geq 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}\,X}{a}.$$

Proof. We write

$$\mathbb{P}(X \geq a) = \mathbb{E}\left[1_{[a,\infty)}(X)\right] \leq \mathbb{E}\left[\frac{X}{a}1_{[a,\infty)}(X)\right] \leq \mathbb{E}\,X/a,$$

since X/a is bigger than 1 when $X \in [a, \infty)$. □

If we apply this to $X = (Y - \mathbb{E}\,Y)^2$, we obtain

$$\mathbb{P}(|Y - \mathbb{E}\,Y| \geq a) = \mathbb{P}((Y - \mathbb{E}\,Y)^2 \geq a^2) \leq \operatorname{Var} Y/a^2. \tag{1.4}$$

This special case of Chebyshev's inequality is sometimes itself referred to as Chebyshev's inequality, while Proposition 1.5 is sometimes called the Markov inequality.

The second inequality we need is Jensen's inequality, not to be confused with the Jensen's formula of complex analysis.

Proposition 1.6. *Suppose g is convex and and X and $g(X)$ are both integrable. Then*

$$g(\mathbb{E}\, X) \leq \mathbb{E}\, g(X).$$

Proof. One property of convex functions is that they lie above their tangent lines, and more generally their support lines. So if $x_0 \in \mathbb{R}$, we have

$$g(x) \geq g(x_0) + c(x - x_0)$$

for some constant c. Take $x = X(\omega)$ and take expectations to obtain

$$\mathbb{E}\, g(X) \geq g(x_0) + c(\mathbb{E}\, X - x_0).$$

Now set x_0 equal to $\mathbb{E}\, X$. □

If A_n is a sequence of sets, define (A_n i.o.), read "A_n infinitely often," by

$$(A_n \text{ i.o.}) = \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i.$$

This set consists of those ω that are in infinitely many of the A_n.

A simple but very important proposition is the Borel-Cantelli lemma. It has two parts, and we prove the first part here, leaving the second part to the next section.

Proposition 1.7. (Borel-Cantelli lemma) *If $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$.*

Proof. We have

$$\mathbb{P}(A_n \text{ i.o.}) = \lim_{n\to\infty} \mathbb{P}(\cup_{i=n}^{\infty} A_i).$$

However,

$$\mathbb{P}(\cup_{i=n}^{\infty} A_i) \leq \sum_{i=n}^{\infty} \mathbb{P}(A_i),$$

which tends to zero as $n \to \infty$. □

2. Independence.

Let us say two events A and B are *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. The events $A_1, \ldots, A_n$ are independent if $\mathbb{P}(A_1 \cap \cdots \cap A_n) = \prod_{i=1}^{n} \mathbb{P}(A_i)$.

Proposition 2.1. *If A and B are independent, then A^c and B are independent.*

Proof. We write

$$\mathbb{P}(A^c \cap B) = \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(B)(1 - \mathbb{P}(A)) = \mathbb{P}(B)\mathbb{P}(A).$$

□

We say two σ-fields $\mathcal{F}$ and $\mathcal{G}$ are independent if A and B are independent whenever $A \in \mathcal{F}$ and $B \in \mathcal{G}$. Two random variables X and Y are independent if the σ-field generated by X and the σ-field generated by Y are independent. (Recall that the σ-field generated by a random variable X is given by $\{(X \in A) : A \text{ a Borel subset of } \mathbb{R}\}$.) We define the independence of n σ-fields or n random variables in the obvious way.

Proposition 2.1 tells us that A and B are independent if the random variables 1_A and 1_B are independent, so the definitions above are consistent.

If f and g are Borel functions and X and Y are independent, then $f(X)$ and $g(Y)$ are independent. This follows because the σ-field generated by $f(X)$ is a sub-σ-field of the one generated by X, and similarly for $g(Y)$.

Let $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ denote the joint distribution function of X and Y. (The comma inside the set means "and.")

Proposition 2.2. *$F_{X,Y}(x,y) = F_X(x)F_Y(y)$ if and only if X and Y are independent.*

Proof. If X and Y are independent, the $1_{(-\infty,x]}(X)$ and $1_{(-\infty,y]}(Y)$ are independent by the above comments. Using the above comments and the definition of independence, this shows $F_{X,Y}(x,y) = F_X(x)F_Y(y)$.

Conversely, if the inequality holds, let $\mathcal{M}_y$ denote the collection of sets of the form $A \times (-\infty, y]$ for which $\mathbb{P}(X \in A, Y \leq y) = \mathbb{P}(X \in A)\mathbb{P}(Y \leq y)$. $\mathcal{M}_y$ contains all sets of the form $(-\infty, x]$. It is clear that $\mathcal{M}_y$ is a monotone class, so by the monotone class lemma, $\mathcal{M}_y$ contains the Borel σ-field.

For a fixed set A, let $\mathcal{M}_A$ denote the collection of sets of the form $A \times B$ for which $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$. Again, $\mathcal{M}_A$ is a monotone class and by the preceding paragraph contains the σ-field generated by the sets $(-\infty, y]$, hence contains the Borel sets. Therefore X and Y are independent. □

The following is known as the multiplication theorem.

Proposition 2.3. *If X, Y, and XY are integrable and X and Y are independent, then $\mathbb{E}\,XY = \mathbb{E}\,X\mathbb{E}\,Y$.*

Proof. Consider the random variables in $\sigma(X)$ (the σ-field generated by X) and $\sigma(Y)$ for which the multiplication theorem is true. It holds for indicators by the definition of X and Y being independent. It holds for simple random variables, that is, linear combinations of indicators, by linearity of both sides. It holds for nonnegative random variables by monotone convergence. And it holds for integrable random variables by linearity again. □

Let us give an example of independent random variables. Let $\Omega = \Omega_1 \times \Omega_2$ and let $\mathbb{P} = \mathbb{P}_1 \times \mathbb{P}_2$, where $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ are probability spaces for $i = 1, 2$. We use the product σ-field. Then it is clear that $\mathcal{F}_1$ and $\mathcal{F}_2$ are independent by the definition of $\mathbb{P}$. If X_1 is a random variable such that $X_1(\omega_1, \omega_2)$ depends only on ω_1 and X_2 depends only on ω_2, then X_1 and X_2 are independent.

This example can be extended to n independent random variables, and in fact, if one has independent random variables, one can always view them as coming from a product space. We will not use this fact. Later on, we will talk about countable sequences of independent r.v.s and the reader may wonder whether such things can exist. That it can is a consequence of the Kolmogorov extension theorem; see PTA, for example.

If $X_1, \ldots, X_n$ are independent, then so are $X_1 - \mathbb{E}\,X_1, \ldots, X_n - \mathbb{E}\,X_n$. Assuming everything is integrable,

$$\mathbb{E}\,[(X_1 - \mathbb{E}\,X_1) + \cdots (X_n - \mathbb{E}\,X_n)]^2 = \mathbb{E}\,(X_1 - \mathbb{E}\,X_1)^2 + \cdots + \mathbb{E}\,(X_n - \mathbb{E}\,X_n)^2,$$

using the multiplication theorem to show that the expectations of the cross product terms are zero. We have thus shown

$$\operatorname{Var}(X_1 + \cdots + X_n) = \operatorname{Var} X_1 + \cdots + \operatorname{Var} X_n. \tag{2.1}$$

We finish up this section by proving the second half of the Borel-Cantelli lemma.

Proposition 2.4. *Suppose A_n is a sequence of independent events. If $\sum_n \mathbb{P}(A_n) = \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 1$.*

Note that here the A_n are independent, while in the first half of the Borel-Cantelli lemma no such assumption was necessary.

Proof. Note

$$\mathbb{P}(\cup_{i=n}^N A_i) = 1 - \mathbb{P}(\cap_{i=n}^N A_i^c) = 1 - \prod_{i=n}^N \mathbb{P}(A_i^c) = 1 - \prod_{i=n}^N (1 - \mathbb{P}(A_i)).$$

By the mean value theorem, $1 - x \le e^{-x}$, so we have that the right hand side is greater than or equal to $1 - \exp(-\sum_{i=n}^{N} \mathbb{P}(A_i))$. As $N \to \infty$, this tends to 1, so $\mathbb{P}(\cup_{i=n}^{\infty} A_i) = 1$. This holds for all n, which proves the result. □

3. Convergence.

In this section we consider three ways a sequence of r.v.s X_n can converge.

We say X_n *converges to* X *almost surely* if $(X_n \not\to X)$ has probability zero. X_n *converges to* X *in probability* if for each ε, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$. X_n *converges to* X *in* L^p if $\mathbb{E}\,|X_n - X|^p \to 0$ as $n \to \infty$.

The following proposition shows some relationships among the types of convergence.

Proposition 3.1. *(a) If $X_n \to X$ a.s., then $X_n \to X$ in probability.*

(b) If $X_n \to X$ in L^p, then $X_n \to X$ in probability.

(c) If $X_n \to X$ in probability, there exists a subsequence n_j such that X_{n_j} converges to X almost surely.

Proof. To prove (a), note $X_n - X$ tends to 0 almost surely, so $1_{(-\varepsilon,\varepsilon)^c}(X_n - X)$ also converges to 0 almost surely. Now apply the dominated convergence theorem.

(b) comes from Chebyshev's inequality:

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - X|^p > \varepsilon^p) \le \mathbb{E}\,|X_n - X|^p/\varepsilon^p \to 0$$

as $n \to \infty$.

To prove (c), choose n_j larger than n_{j-1} such that $\mathbb{P}(|X_n - X| > 2^{-j}) < 2^{-j}$ whenever $n \ge n_j$. So if we let $A_i = (|X_{n_j} - X| > 2^{-i}$ for some $j \ge i)$, then $\mathbb{P}(A_i) \le 2^{-i+1}$. By the Borel-Cantelli lemma $\mathbb{P}(A_i \text{ i.o.}) = 0$. This implies $X_{n_j} \to X$ on the complement of $(A_i$ i.o.). □

Let us give some examples to show there need not be any other implications among the three types of convergence.

Let $\Omega = [0,1]$, $\mathcal{F}$ the Borel σ-field, and $\mathbb{P}$ Lebesgue measure. Let $X_n = e^n 1_{(0,1/n)}$. Then clearly X_n converges to 0 almost surely and in probability, but $\mathbb{E}\,X_n^p = e^{np}/n \to \infty$ for any p.

Let Ω be the unit circle, and let $\mathbb{P}$ be Lebesgue measure on the circle normalized to have total mass 1. Let $t_n = \sum_{i=1}^{n} i^{-1}$, and let $A_n = \{\theta : t_{n-1} \le \theta < t_n\}$. Let $X_n = 1_{A_n}$. Any point on the unit circle will be in infinitely many A_n, so X_n does not converge almost surely to 0. But $\mathbb{P}(A_n) = 1/2\pi n \to 0$, so $X_n \to 0$ in probability and in L^p.

4. Weak law of large numbers.

Suppose X_n is a sequence of independent random variables. Suppose also that they all have the same distribution, that is, $F_{X_n} = F_{X_1}$ for all n. This situation comes up so often it has a name, *independent, identically distributed*, which is abbreviated *i.i.d.*

Define $S_n = \sum_{i=1}^{n} X_i$. S_n is called a *partial sum process*. S_n/n is the average value of the first n of the X_i's.

Theorem 4.1. (Weak law of large numbers) *Suppose the X_i are i.i.d. and $\mathbb{E}\,X_1^2 < \infty$. Then $S_n/n \to \mathbb{E}\,X_1$ in probability.*

Proof. Since the X_i are i.i.d., they all have the same expectation, and so $\mathbb{E}\,S_n = n\mathbb{E}\,X_1$. Hence $\mathbb{E}\,(S_n/n - \mathbb{E}\,X_1)^2$ is the variance of S_n/n. If $\varepsilon > 0$, by Chebyshev's inequality,

$$\mathbb{P}(|S_n/n - \mathbb{E}\,X_1| > \varepsilon) \le \frac{\text{Var}\,(S_n/n)}{\varepsilon^2} = \frac{\sum_{i=1}^{n} \text{Var}\,X_i}{n^2\varepsilon^2} = \frac{n\text{Var}\,X_1}{n^2\varepsilon^2}. \tag{4.1}$$

Since $\mathbb{E}\, X_1^2 < \infty$, then $\operatorname{Var} X_1 < \infty$, and the result follows by letting $n \to \infty$. □

A nice application of the weak law of large numbers is a proof of the Weierstrass approximation theorem.

Theorem 4.2. *Suppose f is a continuous function on $[0,1]$ and $\varepsilon > 0$. There exists a polynomial P such that $\sup_{x\in[0,1]} |f(x) - P(x)| < \varepsilon$.*

Proof. Let

$$P(x) = \sum_{k=0}^{n} f(k/n) \binom{n}{k} x^k (1-x)^{n-k}.$$

Clearly P is a polynomial. Since f is continuous, there exists M such that $|f(x)| \le M$ for all x and there exists δ such that $|f(x) - f(y)| < \varepsilon/2$ whenever $|x - y| < \delta$.

Let X_i be i.i.d. Bernoulli r.v.s with parameter x. Then S_n, the partial sum, is a binomial, and hence $P(x) = \mathbb{E}\, f(S_n/n)$. The mean of S_n/n is x. We have

$$\begin{aligned} |P(x) - f(x)| &= |\mathbb{E}\, f(S_n/n) - f(\mathbb{E}\, X_1)| \le \mathbb{E}\, |f(S_n/n) - f(\mathbb{E}\, X_1)| \\ &\le M\, \mathbb{P}(|S_n/n - x| > \delta) + \varepsilon/2. \end{aligned}$$

By (4.1) the first term will be less than

$$M \operatorname{Var} X_1/n\delta^2 \le M x(1-x)/n\delta^2 \le M n\delta^2,$$

which will be less than $\varepsilon/2$ if n is large enough, uniformly in x. □

5. Techniques related to almost sure convergence.

Our aim is the strong law of large numbers (SLLN), which says that S_n/n converges to $\mathbb{E}\, X_1$ almost surely if $\mathbb{E}\, |X_1| < \infty$.

We first prove it under the assumption that $\mathbb{E}\, X_1^4 < \infty$.

Proposition 5.1. *Suppose X_i is an i.i.d. sequence with $\mathbb{E}\, X_i^4 < \infty$ and let $S_n = \sum_{i=1}^n X_i$. Then $S_n/n \to \mathbb{E}\, X_1$ a.s.*

Proof. By looking at $X_i - \mathbb{E}\, X_i$ we may assume that the X_i have mean 0. By Chebyshev,

$$\mathbb{P}(|S_n/n| > \varepsilon) \le \frac{\mathbb{E}\, (S_n/n)^4}{\varepsilon^4} = \frac{\mathbb{E}\, S_n^4}{n^4 \varepsilon^4}.$$

If we expand S_n^4, we will have terms involving X_i^4, terms involving $X_i^2 X_j^2$, terms involving $X_i^3 X_j$, terms involving $X_i^2 X_j X_k$, and terms involving $X_i X_j X_k X_\ell$, with i, j, k, ℓ all being different. By the multiplication theorem and the fact that the X_i have mean 0, the expectations of all the terms will be 0 except for those of the first two types. So

$$\mathbb{E}\, S_n^4 = \sum_{i=1}^{n} \mathbb{E}\, X_i^4 + \sum_{i \ne j} \mathbb{E}\, X_i^2 \mathbb{E}\, X_j^2.$$

By the finiteness assumption, the first term on the right is bounded by $c_1 n$. By Cauchy-Schwarz, $\mathbb{E}\, X_i^2 \le (\mathbb{E}\, X_i^4)^{1/2} < \infty$, and there are at most n^2 terms in the second term on the right, so this second term is bounded by $c_2 n^2$. Substituting, we have

$$\mathbb{P}(|S_n/n| > \varepsilon) \le c_3/n^2 \varepsilon^4.$$

Consequently $\mathbb{P}(|S_n/n| > \varepsilon \text{ i.o.}) = 0$ by Borel-Cantelli. Since ε is arbitrary, this implies $S_n/n \to 0$ a.s. □

Before we can prove the SLLN assuming only the finiteness of first moments, we need some preliminaries.

Proposition 5.2. *If $Y \geq 0$, then $\mathbb{E}\, Y < \infty$ if and only if $\sum_n \mathbb{P}(Y > n) < \infty$.*

Proof. By Proposition 1.4, $\mathbb{E}\, Y = \int_0^\infty \mathbb{P}(Y > x)dx$. $\mathbb{P}(Y > x)$ is nonincreasing in x, so the integral is bounded above by $\sum_{n=0}^\infty \mathbb{P}(Y > n)$ and bounded below by $\sum_{n=1}^\infty \mathbb{P}(Y > n)$. □

If X_i is a sequence of r.v.s, the tail σ-field is defined by $\cap_{n=1}^\infty \sigma(X_n, X_{n+1}, \ldots)$. An example of an event in the tail σ-field is $(\limsup_{n\to\infty} X_n > a)$. Another example is $(\limsup_{n\to\infty} S_n/n > a)$. The reason for this is that if $k < n$ is fixed,

$$\frac{S_n}{n} = \frac{S_k}{n} + \frac{\sum_{i=k+1}^n X_i}{n}.$$

The first term on the right tends to 0 as $n \to \infty$. So $\limsup S_n/n = \limsup(\sum_{i=k+1}^n X_i)/n$, which is in $\sigma(X_{k+1}, X_{k+2}, \ldots)$. This holds for each k. The set $(\limsup S_n > a)$ is easily seen not to be in the tail σ-field.

Theorem 5.3. (Kolmogorov 0-1 law) *If the X_i are independent, then the events in the tail σ-field have probability 0 or 1.*

This implies that in the case of i.i.d. random variables, if S_n/n has a limit with positive probability, then it has a limit with probability one, and the limit must be a constant.

Proof. Let $\mathcal{M}$ be the collection of sets in $\sigma(X_{n+1}, \ldots)$ that is independent of every set in $\sigma(X_1, \ldots, X_n)$. $\mathcal{M}$ is easily seen to be a monotone class and it contains $\sigma(X_{n+1}, \ldots, X_N)$ for every $N > n$. Therefore $\mathcal{M}$ must be equal to $\sigma(X_{n+1}, \ldots)$.

If A is in the tail σ-field, then for each n, A is independent of $\sigma(X_1, \ldots, X_n)$. The class $\mathcal{M}_A$ of sets independent of A is a monotone class, hence is a σ-field containing $\sigma(X_1, \ldots, X_n)$ for each n. Therefore $\mathcal{M}_A$ contains $\sigma(X_1, \ldots)$.

We thus have that the event A is independent of itself, or

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A) = \mathbb{P}(A)^2.$$

This implies $\mathbb{P}(A)$ is zero or one. □

The next proposition shows that in considering a law of large numbers we can consider truncated random variables.

Proposition 5.4. *Suppose X_i is an i.i.d. sequence of r.v.s with $\mathbb{E}\,|X_1| < \infty$. Let $X_n' = X_n 1_{(|X_n| \leq n)}$. Then*

(a) X_n converges almost surely if and only if X_n' does;

(b) If $S_n' = \sum_{i=1}^n X_i'$, then S_n/n converges a.s. if and only if S_n'/n does.

Proof. Let $A_n = (X_n \neq X_n') = (|X_n| > n)$. Then $\mathbb{P}(A_n) = \mathbb{P}(|X_n| > n) = \mathbb{P}(|X_1| > n)$. Since $\mathbb{E}\,|X_1| < \infty$, then by Proposition 5.2 we have $\sum \mathbb{P}(A_n) < \infty$. So by the Borel-Cantelli lemma, $\mathbb{P}(A_n \text{ i.o.}) = 0$. Thus for almost every ω, $X_n = X_n'$ for n sufficiently large. This proves (a).

For (b), let k (depending on ω) be the largest integer such that $X_k'(\omega) \neq X_k(\omega)$. Then $S_n/n - S_n'/n = (X_1 + \cdots + X_k)/n - (X_1' + \cdots + X_k')/n \to 0$ as $n \to \infty$. □

Next is Kolmogorov's inequality, a special case of Doob's inequality.

Proposition 5.5. *Suppose the X_i are independent and $\mathbb{E}\, X_i = 0$ for each i. Then*

$$\mathbb{P}(\max_{1 \leq i \leq n} |S_i| \geq \lambda) \leq \frac{\mathbb{E}\, S_n^2}{\lambda^2}.$$

Proof. Let $A_k = (|S_k| \geq \lambda, |S_1| < \lambda, \ldots, |S_{k-1}| < \lambda)$. Note the A_k are disjoint and that $A_k \in \sigma(X_1, \ldots, X_k)$. Therefore A_k is independent of $S_n - S_k$. Then

$$\begin{aligned}
\mathbb{E}\, S_n^2 &\geq \sum_{k=1}^n \mathbb{E}\,[S_n^2; A_k] \\
&= \sum_{k=1}^n \mathbb{E}\,[(S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2); A_k] \\
&\geq \sum_{k=1}^n \mathbb{E}\,[S_k^2; A_k] + 2\sum_{k=1}^n \mathbb{E}\,[S_k(S_n - S_k); A_k].
\end{aligned}$$

Using the independence, $\mathbb{E}\,[S_k(S_n - S_k)1_{A_k}] = \mathbb{E}\,[S_k 1_{A_k}]\mathbb{E}\,[S_n - S_k] = 0$. Therefore

$$\mathbb{E}\, S_n^2 \geq \sum_{k=1}^n \mathbb{E}\,[S_k^2; A_k] \geq \sum_{k=1}^n \lambda^2 \mathbb{P}(A_k) = \lambda^2 \mathbb{P}(\max_{1\leq k\leq n} |S_k| \geq \lambda).$$

Our result is immediate from this. □

The last result we need for now is a special case of what is known as Kronecker's lemma.

Proposition 5.6. *Suppose x_i are real numbers and $s_n = \sum_{i=1}^n x_i$. If $\sum_{j=1}^\infty (x_j/j)$ converges, then $s_n/n \to 0$.*

Proof. Let $b_n = \sum_{j=1}^n (x_j/j)$, $b_0 = 0$, and suppose $b_n \to b$. As is well known, this implies $(\sum_{i=1}^n b_i)/n \to b$. We have $n(b_n - b_{n-1}) = x_n$, so

$$\frac{s_n}{n} = \frac{\sum_{i=1}^n (ib_i - ib_{i-1})}{n} = \frac{\sum_{i=1}^n ib_i - \sum_{i=1}^{n-1}(i+1)b_i}{n} = b_n - \frac{\sum_{i=1}^n b_i}{n} \to b - b = 0.$$

□

6. Strong law of large numbers.

This section is devoted to a proof of Kolmogorov's strong law of large numbers. We showed earlier that if $\mathbb{E}\, X_i^2 < \infty$, where the X_i are i.i.d., then the weak law of large numbers (WLLN) holds: S_n/n converges to $\mathbb{E}\, X_1$ in probability. The WLLN can be improved greatly; it is enough that $x\mathbb{P}(|X_1| > x) \to 0$ as $x \to \infty$. Here we show the strong law (SLLN): if one has a finite first moment, then there is almost sure convergence.

First we need a lemma.

Lemma 6.1. *Suppose V_i is a sequence of independent r.v.s, each with mean 0. Let $W_n = \sum_{i=1}^n V_i$. If $\sum_{i=1}^\infty \operatorname{Var} V_i < \infty$, then W_n converges almost surely.*

Proof. Choose $n_j > n_{j-1}$ such that $\sum_{i=n_j}^\infty \operatorname{Var} V_i < 2^{-3j}$. If $n > n_j$, then applying Kolmogorov's inequality shows that

$$\mathbb{P}(\max_{n_j \leq i \leq n} |W_i - W_{n_j}| > 2^{-j}) \leq 2^{-3j}/2^{-2j} = 2^{-j}.$$

Letting $n \to \infty$, we have $\mathbb{P}(A_j) \leq 2^{-j}$, where

$$A_j = (\max_{n_j \leq i} |W_i - W_{n_j}| > 2^{-j}).$$

By the Borel-Cantelli lemma, $\mathbb{P}(A_j \text{ i.o.}) = 0$.

Suppose $\omega \notin (A_j \text{ i.o.})$. Let $\varepsilon > 0$. Choose j large enough so that $2^{-j+1} < \varepsilon$ and $\omega \notin A_j$. If $n, m > n_j$, then

$$|W_n - W_m| \le |W_n - W_{n_j}| + |W_m - W_{n_j}| \le 2^{-j+1} < \varepsilon.$$

Since ε is arbitrary, $W_n(\omega)$ is a Cauchy sequence, and hence converges. □

Theorem 6.2. (SLLN) *Let X_i be a sequence of i.i.d. random variables. Then S_n/n converges almost surely if and only if $\mathbb{E}\,|X_1| < \infty$.*

Proof. Let us first suppose S_n/n converges a.s. and show $\mathbb{E}\,|X_1| < \infty$. If $S_n(\omega)/n \to a$, then

$$\frac{S_{n-1}}{n} = \frac{S_{n-1}}{n-1}\frac{n-1}{n} \to a.$$

So

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

Hence $X_n/n \to 0$, a.s. Thus $\mathbb{P}(|X_n| > n \text{ i.o.}) = 0$. By the second part of Borel-Cantelli, $\sum \mathbb{P}(|X_n| > n) < \infty$. Since the X_i are i.i.d., this means $\sum_{n=1}^{\infty} \mathbb{P}(|X_1| > n) < \infty$, and by Proposition 4.1, $E|X_1| < \infty$.

Now suppose $\mathbb{E}\,|X_1| < \infty$. By looking at $X_i - \mathbb{E}\,X_i$, we may suppose without loss of generality that $\mathbb{E}\,X_i = 0$. We truncate, and let $Y_i = X_i 1_{(|X_i| \le i)}$. It suffices to show $\sum_{i=1}^{n} Y_i/n \to 0$ a.s., by Proposition 5.4.

Next we estimate. We have

$$\mathbb{E}\,Y_i = \mathbb{E}\,[X_i 1_{(|X_i| \le i)}] = \mathbb{E}\,[X_1 1_{(|X_1| \le i)}] \to \mathbb{E}\,X_1 = 0.$$

The convergence follows by the dominated convergence theorem, since the integrands are bounded by $|X_1|$. To estimate the second moment of the Y_i, we write

$$\begin{aligned}
\mathbb{E}\,Y_i^2 &= \int_0^\infty 2y\mathbb{P}(|Y_i| \ge y)\,dy \\
&= \int_0^i 2y\mathbb{P}(|Y_i| \ge y)\,dy \\
&\le \int_0^i 2y\mathbb{P}(|X_1| \ge y)\,dy,
\end{aligned}$$

and so

$$\begin{aligned}
\sum_{i=1}^{\infty} \mathbb{E}\,(Y_i^2/i^2) &\le \sum_{i=1}^{\infty} \frac{1}{i^2} \int_0^i 2y\mathbb{P}(|X_1| \ge y)\,dy \\
&= 2\sum_{i=1}^{\infty} \frac{1}{i^2} \int_0^\infty 1_{(y \le i)}\, y\mathbb{P}(|X_1| \ge y)\,dy \\
&= 2\int_0^\infty \sum_{i=1}^{\infty} \frac{1}{i^2} 1_{(y \le i)}\, y\mathbb{P}(|X_1| \ge y)\,dy \\
&\le 4\int_0^\infty \frac{1}{y} y\mathbb{P}(|X_1| \ge y)\,dy \\
&= 4\int_0^\infty \mathbb{P}(|X_1| \ge y)\,dy = 4\mathbb{E}\,|X_1| < \infty.
\end{aligned}$$

Let $U_i = Y_i - \mathbb{E}\,Y_i$. Then $\operatorname{Var} U_i = \operatorname{Var} Y_i \le \mathbb{E}\,Y_i^2$, and by the above,

$$\sum_{i=1}^{\infty} \operatorname{Var}(U_i/i) < \infty.$$

By Lemma 6.1 (with $V_i = U_i/i$), $\sum_{i=1}^n (U_i/i)$ converges almost surely. By Kronecker's lemma, $(\sum_{i=1}^n V_i)/n$ converges almost surely. Finally, since $\mathbb{E}\, Y_i \to 0$, then $\sum_{i=1}^n \mathbb{E}\, Y_i/n \to 0$, hence $\sum_{i=1}^n Y_i/n \to 0$. □

7. Uniform integrability. Before proceeding to some extensions of the SLLN, we discuss uniform integrability. A sequence of r.v.s is *uniformly integrable* if

$$\sup_i \int_{(|X_i|>M)} |X_i|\, d\mathbb{P} \to 0$$

as $M \to \infty$.

Proposition 7.1. *Suppose there exists $\varphi : [0,\infty) \to [0,\infty)$ such that φ is nondecreasing, $\varphi(x)/x \to \infty$ as $x \to \infty$, and $\sup_i \mathbb{E}\, \varphi(|X_i|) < \infty$. Then the X_i are uniformly integrable.*

Proof. Let $\varepsilon > 0$ and choose x_0 such that $x/\varphi(x) < \varepsilon$ if $x \geq x_0$. If $M \geq x_0$,

$$\int_{(|X_i|>M)} |X_i| = \int \frac{|X_i|}{\varphi(|X_i|)} \varphi(|X_i|) 1_{(|X_i|>M)} \leq \varepsilon \int \varphi(|X_i|) \leq \varepsilon \sup_i \mathbb{E}\, \varphi(|X_i|).$$

□

Proposition 7.2. *If X_n and Y_n are two uniformly integrable sequences, then $X_n + Y_n$ is also a uniformly integrable sequence.*

Proof. Since there exists M_0 such that $\sup_n \mathbb{E}\,[|X_n|; |X_n| > M_0] < 1$ and $\sup_n \mathbb{E}\,[|Y_n|; |Y_n| > M_0] < 1$, then $\sup_n \mathbb{E}\, |X_n| \leq M_0 + 1$, and similarly for the Y_n. Let $\varepsilon > 0$ and choose $M_1 > 4(M_0+1)/\varepsilon$ such that $\sup_n \mathbb{E}\,[|X_n|; |X_n| > M_1] < \varepsilon/4$ and $\sup_n \mathbb{E}\,[|Y_n|; |Y_n| > M_1] < \varepsilon/4$. Let $M_2 = 4M_1^2$.

Note $\mathbb{P}(|X_n| + |Y_n| > M_2) \leq (\mathbb{E}\, |X_n| + \mathbb{E}\, |Y_n|)/M_2 \leq \varepsilon/(4M_1)$ by Chebyshev's inequality. Then

$$\begin{aligned}\mathbb{E}\,[|X_n + Y_n|; |X_n + Y_n| > M_2] \leq\ & \mathbb{E}\,[|X_n|; |X_n| > M_1] \\ & + \mathbb{E}\,[|X_n|; |X_n| \leq M_1, |X_n + Y_n| > M_2] \\ & + \mathbb{E}\,[|Y_n|; |Y_n| > M_1] \\ & + \mathbb{E}\,[|Y_n|; |Y_n| \leq M_1, |X_n + Y_n| > M_2].\end{aligned}$$

The first and third terms on the right are each less than $\varepsilon/4$ by our choice of M_1. The second and fourth terms are each less than $M_1 \mathbb{P}(|X_n + Y_n| > M_2) \leq \varepsilon/2$. □

The main result we need in this section is Vitali's convergence theorem.

Theorem 7.3. *If $X_n \to X$ almost surely and the X_n are uniformly integrable, then $\mathbb{E}\, X_n \to \mathbb{E}\, X$.*

Proof. By the above proposition, $X_n - X$ is uniformly integrable and tends to 0 a.s., so without loss of generality, we may assume $X = 0$. Let $\varepsilon > 0$ and choose M such that $\sup_n \mathbb{E}\,[|X_n|; |X_n| > M] < \varepsilon$. Then

$$\mathbb{E}\, |X_n| \leq \mathbb{E}\,[|X_n|; |X_n| > M] + \mathbb{E}\,[|X_n|; |X_n| \leq M] \leq \varepsilon + \mathbb{E}\,[|X_n| 1_{(|X_n| \leq M)}].$$

The second term on the right goes to 0 by dominated convergence. □

8. Complements to the SLLN.

Proposition 8.1. *Suppose X_i is an i.i.d. sequence and $\mathbb{E}\,|X_1| < \infty$. Then $\mathbb{E}\,|(S_n/n) - \mathbb{E}\,X_1| \to 0$.*

Proof. Without loss of generality we may assume $\mathbb{E}\,X_1 = 0$. By the SLLN, $S_n/n \to 0$ a.s. So we need to show that the sequence S_n/n is uniformly integrable.

Pick M_1 such that $\mathbb{E}\,[|X_1|; |X_1| > M_1] < \varepsilon/2$. Pick $M_2 = M_1\mathbb{E}\,|X_1|/\varepsilon$. So

$$\mathbb{P}(|S_n/n| > M_2) \le \mathbb{E}\,|S_n|/nM_2 \le \mathbb{E}\,|X_1|/M_2 = \varepsilon/M_1.$$

We used here $\mathbb{E}\,|S_n| \le \sum_{i=1}^n \mathbb{E}\,|X_i| = n\mathbb{E}\,|X_1|$.

We then have

$$\begin{aligned}\mathbb{E}\,[|X_i|; |S_n/n| > M_2] &\le \mathbb{E}\,[|X_i| : |X_i| > M_1] + \mathbb{E}\,[|X_i|; |X_i| \le M_1, |S_n/n| > M_2] \\ &\le \varepsilon + M_1\mathbb{P}(|S_n/n| > M_2) \le 2\varepsilon.\end{aligned}$$

Finally,

$$\mathbb{E}\,[|S_n/n|; |S_n/n| > M_2] \le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\,[|X_i|; |S_n/n| > M_2] \le 2\varepsilon.$$

□

We now consider the "three series criterion." We prove the "if" portion here and defer the "only if" to Section 20.

Theorem 8.2. *Let X_i be a sequence of independent random variables., $A > 0$, and $Y_i = X_i 1_{(|X_i| \le A)}$. Then $\sum X_i$ converges if and only if all of the following three series converge: (a) $\sum \mathbb{P}(|X_n| > A)$; (b) $\sum \mathbb{E}\,Y_i$; (c) $\sum \operatorname{Var} Y_i$.*

Proof of "if" part. Since (c) holds, then $\sum (Y_i - \mathbb{E}\,Y_i)$ converges by Lemma 6.1. Since (b) holds, taking the difference shows $\sum Y_i$ converges. Since (a) holds, $\sum X_i$ converges by Proposition 5.4. □

9. Conditional expectation.

If $\mathcal{F} \subseteq \mathcal{G}$ are two σ-fields and X is an integrable $\mathcal{G}$ measurable random variable, the *conditional expectation* of X given $\mathcal{F}$, written $\mathbb{E}\,[X \mid \mathcal{F}]$ and read as "the expectation (or expected value) of X given $\mathcal{F}$," is any $\mathcal{F}$ measurable random variable Y such that $\mathbb{E}\,[Y; A] = \mathbb{E}\,[X; A]$ for every $A \in \mathcal{F}$. The *conditional probability* of $A \in \mathcal{G}$ given $\mathcal{F}$ is defined by $\mathbb{P}(A \mid \mathcal{F}) = \mathbb{E}\,[1_A \mid \mathcal{F}]$.

If Y_1, Y_2 are two $\mathcal{F}$ measurable random variables with $\mathbb{E}\,[Y_1; A] = \mathbb{E}\,[Y_2; A]$ for all $A \in \mathcal{F}$, then $Y_1 = Y_2$, a.s., or conditional expectation is unique up to a.s. equivalence.

In the case X is already $\mathcal{F}$ measurable, $\mathbb{E}\,[X \mid \mathcal{F}] = X$. If X is independent of $\mathcal{F}$, $\mathbb{E}\,[X \mid \mathcal{F}] = \mathbb{E}\,X$. Both of these facts follow immediately from the definition. For another example, which ties this definition with the one used in elementary probability courses, if $\{A_i\}$ is a finite collection of disjoint sets whose union is Ω, $\mathbb{P}(A_i) > 0$ for all i, and $\mathcal{F}$ is the σ-field generated by the A_is, then

$$\mathbb{P}(A \mid \mathcal{F}) = \sum_i \frac{\mathbb{P}(A \cap A_i)}{\mathbb{P}(A_i)} 1_{A_i}.$$

This follows since the right-hand side is $\mathcal{F}$ measurable and its expectation over any set A_i is $\mathbb{P}(A \cap A_i)$.

As an example, suppose we toss a fair coin independently 5 times and let X_i be 1 or 0 depending whether the ith toss was a heads or tails. Let A be the event that there were 5 heads and let $\mathcal{F}_i =$

$\sigma(X_1, \ldots, X_i)$. Then $\mathbb{P}(A) = 1/32$ while $\mathbb{P}(A \mid \mathcal{F}_1)$ is equal to $1/16$ on the event $(X_1 = 1)$ and 0 on the event $(X_1 = 0)$. $\mathbb{P}(A \mid \mathcal{F}_2)$ is equal to $1/8$ on the event $(X_1 = 1, X_2 = 1)$ and 0 otherwise.

We have

$$\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{F}\right]\right] = \mathbb{E}\, X \tag{9.1}$$

because $\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{F}\right]\right] = \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{F}\right]; \Omega\right] = \mathbb{E}\left[X; \Omega\right] = \mathbb{E}\, X$.

The following is easy to establish.

Proposition 9.1. *(a) If $X \geq Y$ are both integrable, then $\mathbb{E}\left[X \mid \mathcal{F}\right] \geq \mathbb{E}\left[Y \mid \mathcal{F}\right]$ a.s.*
(b) If X and Y are integrable and $a \in \mathbb{R}$, then $\mathbb{E}\left[aX + Y \mid \mathcal{F}\right] = a\mathbb{E}\left[X \mid \mathcal{F}\right] + \mathbb{E}\left[Y \mid \mathcal{F}\right]$.

It is easy to check that limit theorems such as monotone convergence and dominated convergence have conditional expectation versions, as do inequalities like Jensen's and Chebyshev's inequalities. Thus, for example, we have the following.

Proposition 9.2. (Jensen's inequality for conditional expectations) *If g is convex and X and $g(X)$ are integrable,*

$$\mathbb{E}\left[g(X) \mid \mathcal{F}\right] \geq g(\mathbb{E}\left[X \mid \mathcal{F}\right]), \quad \text{a.s.}$$

A key fact is the following.

Proposition 9.3. *If X and XY are integrable and Y is measurable with respect to $\mathcal{F}$, then*

$$\mathbb{E}\left[XY \mid \mathcal{F}\right] = Y\mathbb{E}\left[X \mid \mathcal{F}\right]. \tag{9.2}$$

Proof. If $A \in \mathcal{F}$, then for any $B \in \mathcal{F}$,

$$\mathbb{E}\left[1_A \mathbb{E}\left[X \mid \mathcal{F}\right]; B\right] = \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{F}\right]; A \cap B\right] = \mathbb{E}\left[X; A \cap B\right] = \mathbb{E}\left[1_A X; B\right].$$

Since $1_A \mathbb{E}\left[X \mid \mathcal{F}\right]$ is $\mathcal{F}$ measurable, this shows that (9.1) holds when $Y = 1_A$ and $A \in \mathcal{F}$. Using linearity and taking limits shows that (9.1) holds whenever Y is $\mathcal{F}$ measurable and $X$and XY are integrable. □

Two other equalities follow.

Proposition 9.4. *If $\mathcal{E} \subseteq \mathcal{F} \subseteq \mathcal{G}$, then*

$$\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{F}\right] \mid \mathcal{E}\right] = \mathbb{E}\left[X \mid \mathcal{E}\right] = \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{E}\right] \mid \mathcal{F}\right].$$

Proof. The right equality holds because $\mathbb{E}\left[X \mid \mathcal{E}\right]$ is $\mathcal{E}$ measurable, hence $\mathcal{F}$ measurable. To show the left equality, let $A \in \mathcal{E}$. Then since A is also in $\mathcal{F}$,

$$\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{F}\right] \mid \mathcal{E}\right]; A\right] = \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{F}\right]; A\right] = \mathbb{E}\left[X; A\right] = \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{E}\right]; A\right].$$

Since both sides are $\mathcal{E}$ measurable, the equality follows. □

To show the existence of $\mathbb{E}\left[X \mid \mathcal{F}\right]$, we proceed as follows.

Proposition 9.5. *If X is integrable, then $\mathbb{E}[X \mid \mathcal{F}]$ exists.*

Proof. Using linearity, we need only consider $X \geq 0$. Define a measure $\mathbb{Q}$ on $\mathcal{F}$ by $\mathbb{Q}(A) = \mathbb{E}[X; A]$ for $A \in \mathcal{F}$. This is trivially absolutely continuous with respect to $\mathbb{P}|_{\mathcal{F}}$, the restriction of $\mathbb{P}$ to $\mathcal{F}$. Let $\mathbb{E}[X \mid \mathcal{F}]$ be the Radon-Nikodym derivative of $\mathbb{Q}$ with respect to $\mathbb{P}|_{\mathcal{F}}$. The Radon-Nikodym derivative is $\mathcal{F}$ measurable by construction and so provides the desired random variable. □

When $\mathcal{F} = \sigma(Y)$, one usually writes $\mathbb{E}[X \mid Y]$ for $\mathbb{E}[X \mid \mathcal{F}]$. Notation that is commonly used (however, we will use it only very occasionally and only for heuristic purposes) is $\mathbb{E}[X \mid Y = y]$. The definition is as follows. If $A \in \sigma(Y)$, then $A = (Y \in B)$ for some Borel set B by the definition of $\sigma(Y)$, or $1_A = 1_B(Y)$. By linearity and taking limits, if Z is $\sigma(Y)$ measurable, $Z = f(Y)$ for some Borel measurable function f. Set $Z = \mathbb{E}[X \mid Y]$ and choose f Borel measurable so that $Z = f(Y)$. Then $\mathbb{E}[X \mid Y = y]$ is defined to be $f(y)$.

If $X \in L^2$ and $\mathcal{M} = \{Y \in L^2 : Y \text{ is } \mathcal{F}\text{-measurable}\}$, one can show that $\mathbb{E}[X \mid \mathcal{F}]$ is equal to the projection of X onto the subspace $\mathcal{M}$. We will not use this in these notes.

10. Stopping times.

We next want to talk about stopping times. Suppose we have a sequence of σ-fields $\mathcal{F}_i$ such that $\mathcal{F}_i \subset \mathcal{F}_{i+1}$ for each i. An example would be if $\mathcal{F}_i = \sigma(X_1, \ldots, X_i)$. A random mapping N from Ω to $\{0, 1, 2, \ldots\}$ is called a *stopping time* if for each n, $(N \leq n) \in \mathcal{F}_n$. A stopping time is also called an optional time in the Markov theory literature.

The intuition is that the sequence knows whether N has happened by time n by looking at $\mathcal{F}_n$. Suppose some motorists are told to drive north on Highway 99 in Seattle and stop at the first motorcycle shop past the second realtor after the city limits. So they drive north, pass the city limits, pass two realtors, and come to the next motorcycle shop, and stop. That is a stopping time. If they are instead told to stop at the third stop light before the city limits (and they had not been there before), they would need to drive to the city limits, then turn around and return past three stop lights. That is not a stopping time, because they have to go ahead of where they wanted to stop to know to stop there.

We use the notation $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. The proof of the following is immediate from the definitions.

Proposition 10.1.

(a) Fixed times n are stopping times.
(b) If N_1 and N_2 are stopping times, then so are $N_1 \wedge N_2$ and $N_1 \vee N_2$.
(c) If N_n is a nondecreasing sequence of stopping times, then so is $N = \sup_n N_n$.
(d) If N_n is a nonincreasing sequence of stopping times, then so is $N = \inf_n N_n$.
(e) If N is a stopping time, then so is $N + n$.

We define $\mathcal{F}_N = \{A : A \cap (N \leq n) \in \mathcal{F}_n \text{ for all } n\}$.

11. Martingales.

In this section we consider martingales. Let $\mathcal{F}_n$ be an increasing sequence of σ-fields. A sequence of random variables M_n is *adapted* to $\mathcal{F}_n$ if for each n, M_n is $\mathcal{F}_n$ measurable.

M_n is a *martingale* if M_n is adapted to $\mathcal{F}_n$, M_n is integrable for all n, and

$$\mathbb{E}[M_n \mid \mathcal{F}_{n-1}] = M_{n-1}, \qquad \text{a.s.}, \qquad n = 2, 3, \ldots. \tag{11.1}$$

If we have $\mathbb{E}[M_n \mid \mathcal{F}_{n-1}] \geq M_{n-1}$ a.s. for every n, then M_n is a submartingale. If we have $\mathbb{E}[M_n \mid \mathcal{F}_{n-1}] \leq M_{n-1}$, we have a supermartingale. Submartingales have a tendency to increase.

Let us take a moment to look at some examples. If X_i is a sequence of mean zero i.i.d. random variables and S_n is the partial sum process, then $M_n = S_n$ is a martingale, since $\mathbb{E}[M_n \mid \mathcal{F}_{n-1}] = M_{n-1} + \mathbb{E}[M_n - M_{n-1} \mid \mathcal{F}_{n-1}] = M_{n-1} + \mathbb{E}[M_n - M_{n-1}] = M_{n-1}$, using independence. If the X_i's have variance one and $M_n = S_n^2 - n$, then

$$\mathbb{E}[S_n^2 \mid \mathcal{F}_{n-1}] = \mathbb{E}[(S_n - S_{n-1})^2 \mid \mathcal{F}_{n-1}] + 2S_{n-1}\mathbb{E}[S_n \mid \mathcal{F}_{n-1}] - S_{n-1}^2 = 1 + S_{n-1}^2,$$

using independence. It follows that M_n is a martingale.

Another example is the following: if $X \in L^1$ and $M_n = \mathbb{E}[X \mid \mathcal{F}_n]$, then M_n is a martingale.

If M_n is a martingale and $H_n \in \mathcal{F}_{n-1}$ for each n, it is easy to check that $N_n = \sum_{i=1}^n H_i(M_i - M_{i-1})$ is also a martingale.

If M_n is a martingale and $g(M_n)$ is integrable for each n, then by Jensen's inequality

$$\mathbb{E}[g(M_{n+1}) \mid \mathcal{F}_n] \geq g(\mathbb{E}[M_{n+1} \mid \mathcal{F}_n]) = g(M_n),$$

or $g(M_n)$ is a submartingale. Similarly if g is convex and nondecreasing on $[0, \infty)$ and M_n is a positive submartingale, then $g(M_n)$ is a submartingale.

12. Optional stopping.

Note that if one takes expectations in (11.1), one has $\mathbb{E}\, M_n = \mathbb{E}\, M_{n-1}$, and by induction $\mathbb{E}\, M_n = \mathbb{E}\, M_0$. The theorem about martingales that lies at the basis of all other results is Doob's optional stopping theorem, which says that the same is true if we replace n by a stopping time N. There are various versions, depending on what conditions one puts on the stopping times.

Theorem 12.1. *If N is a bounded stopping time with respect to $\mathcal{F}_n$ and M_n a martingale, then $\mathbb{E}\, M_N = \mathbb{E}\, M_0$.*

Proof. Since N is bounded, let K be the largest value N takes. We write

$$\mathbb{E}\, M_N = \sum_{k=0}^K \mathbb{E}[M_N; N = k] = \sum_{k=0}^K \mathbb{E}[M_k; N = k].$$

Note $(N = k)$ is $\mathcal{F}_j$ measurable if $j \geq k$, so

$$\begin{aligned}\mathbb{E}[M_k; N = k] &= \mathbb{E}[M_{k+1}; N = k] \\ &= \mathbb{E}[M_{k+2}; N = k] = \ldots = \mathbb{E}[M_K; N = k].\end{aligned}$$

Hence

$$\mathbb{E}\, M_N = \sum_{k=0}^K \mathbb{E}[M_K; N = k] = \mathbb{E}\, M_K = \mathbb{E}\, M_0.$$

This completes the proof. □

The assumption that N be bounded cannot be entirely dispensed with. For example, let M_n be the partial sums of a sequence of i.i.d. rnadom variable that take the values ± 1, each with probability $\frac{1}{2}$.

The same proof as that in Theorem 12.1 gives the following corollary. If $N = \min\{i : M_i = 1\}$, we will see later on that $N < \infty$ a.s., but $\mathbb{E}\, M_N = 1 \neq 0 = \mathbb{E}\, M_0$.

Corollary 12.2. *If N is bounded by K and M_n is a submartingale, then $\mathbb{E}\, M_N \leq \mathbb{E}\, M_K$.*

Also the same proof gives

Corollary 12.3. *If N is bounded by K, $A \in \mathcal{F}_N$, and M_n is a submartingale, then $\mathbb{E}[M_N; A] \le \mathbb{E}[M_K; A]$.*

Proposition 12.4. *If $N_1 \le N_2$ are stopping times bounded by K and M is a right continuous martingale, then $\mathbb{E}[M_{N_2} \mid \mathcal{F}_{N_1}] = M_{N_1}$, a.s.*

Proof. Suppose $A \in \mathcal{F}_{N_1}$. We need to show $\mathbb{E}[M_{N_1}; A] = \mathbb{E}[M_{N_2}; A]$. Define a new stopping time N_3 by

$$N_3(\omega) = \begin{cases} N_1(\omega) & \text{if } \omega \in A \\ N_2(\omega) & \text{if } \omega \notin A. \end{cases}$$

It is easy to check that N_3 is a stopping time, so $\mathbb{E}\, M_{N_3} = \mathbb{E}\, M_K = \mathbb{E}\, M_{N_2}$ implies

$$\mathbb{E}[M_{N_1}; A] + \mathbb{E}[M_{N_2}; A^c] = \mathbb{E}[M_{N_2}].$$

Subtracting $\mathbb{E}[M_{N_2}; A^c]$ from each side completes the proof. □

13. Doob's inequalities.

The first interesting consequences of the optional stopping theorems are Doob's inequalities. If M_n is a martingale, denote $M_n^* = \max_{i \le n} |M_i|$.

Theorem 13.1. *If M_n is a martingale or a positive submartingale,*

$$\mathbb{P}(M_n^* \ge a) \le \mathbb{E}[|M_n|; M_n^* \ge a]/a \le \mathbb{E}\,|M_n|/a.$$

Proof. Let $N = \min\{j : |M_j| \ge a\} \wedge n$. Since $|\cdot|$ is convex, $|M_n|$ is a submartingale. If $A = (M_n^* \ge a)$, then $A \in \mathcal{F}_N$ and by Corollary 12.3

$$a\mathbb{P}(M_n^* \ge a) \le \mathbb{E}[|M_N|; A] \le \mathbb{E}[|M_n|; A] \le \mathbb{E}\,|M_n|.$$

□

For $p > 1$, we have the following inequality.

Theorem 13.2. *If $p > 1$ and $\mathbb{E}\,|M_i|^p < \infty$ for $i \le n$, then*

$$\mathbb{E}\,(M_n^*)^p \le \Big(\frac{p}{p-1}\Big)^p \mathbb{E}\,|M_n|^p.$$

Proof. Note $M_n^* \le \sum_{i=1}^n |M_n|$, hence $M_n^* \in L^p$. We write

$$\begin{aligned} \mathbb{E}\,(M_n^*)^p &= \int_0^\infty pa^{p-1}\mathbb{P}(M_n^* > a)\,da \le \int_0^\infty pa^{p-1}\mathbb{E}\,[|M_n|1_{(M_n^* \ge a)}/a]\,da \\ &= \mathbb{E}\int_0^{M_n^*} pa^{p-2}|M_n|\,da = \frac{p}{p-1}\mathbb{E}\,[(M_n^*)^{p-1}|M_n|] \\ &\le \frac{p}{p-1}(\mathbb{E}\,(M_n^*)^p)^{(p-1)/p}(\mathbb{E}\,|M_n|^p)^{1/p}. \end{aligned}$$

The last inequality follows by Hölder's inequality. Now divide both sides by the quantity $(\mathbb{E}\,(M_n^*)^p)^{(p-1)/p}$. □

14. Martingale convergence theorems.

The martingale convergence theorems are another set of important consequences of optional stopping. The main step is the upcrossing lemma. The number of upcrossings of an interval $[a,b]$ is the number of times a process crosses from below a to above b.

To be more exact, let

$$S_1 = \min\{k : X_k \le a\}, \qquad T_1 = \min\{k > S_1 : X_k \ge b\},$$

and

$$S_{i+1} = \min\{k > T_i : X_k \le a\}, \qquad T_{i+1} = \min\{k > S_{i+1} : X_k \ge b\}.$$

The number of upcrossings U_n before time n is $U_n = \max\{j : T_j \le n\}$.

Theorem 14.1. (Upcrossing lemma) *If X_k is a submartingale,*

$$\mathbb{E}\, U_n \le (b-a)^{-1}\mathbb{E}\,[(X_n - a)^+].$$

Proof. First assume that $a = 0$ and $X_k \ge 0$ for each k. Fix n and define $X_m = X_n$ for $m \ge n$. This will still be a submartingale. Define the S_i, T_i as above, and let $S_i' = S_i \wedge (n+1)$, $T_i' = T_i \wedge (n+1)$.

We write

$$\mathbb{E}\, X_{n+1} = \mathbb{E}\, X_{S_1'} + \sum_{i=0}^{\infty} \mathbb{E}\,[X_{T_i'} - X_{S_i'}] + \sum_{i=0}^{\infty} \mathbb{E}\,[X_{S_{i+1}'} - X_{T_i'}].$$

All the summands in the third term on the right are nonnegative since X_k is a submartingale, and

$$\sum_{i=0}^{\infty} (X_{T_i'} - X_{S_i'}) \ge (b-a)U_n.$$

So

$$\mathbb{E}\, U_n \le \mathbb{E}\, X_{n+1}/b. \tag{4.8}$$

Now let us remove the assumption that $a = 0$ and $X_k \ge 0$. The number of upcrossings of $[a,b]$ by X_k is the same as the number of upcrossings of $[0, b-a]$ by $Y_k = (X_k - a)^+$. So we merely apply (4.8) to the number of upcrossings of the interval $[0, b-a]$ by the process $(X_k - a)^+$. □

This leads to the martingale convergence theorem.

Theorem 14.2. *If X_n is a submartingale such that $\sup_n \mathbb{E}\, X_n^+ < \infty$, then X_n converges a.s. as $n \to \infty$.*

Proof. Let $U(a,b) = \lim_{n\to\infty} U_n$. For each a, b rational, by monotone convergence,

$$\mathbb{E}\, U(a,b) \le c(b-a)^{-1}\mathbb{E}\,(X_n - a)^+ < \infty.$$

So $U(a,b) < \infty$, a.s. Taking the union over all pairs of rationals a, b, we see that a.s. the sequence $X_n(\omega)$ cannot have $\limsup X_n > \liminf X_n$. Therefore X_n converges a.s., although we still have to rule out the possibility of the limit being infinite. Since X_n is a submartingale, $\mathbb{E}\, X_n \ge \mathbb{E}\, X_0$, and thus

$$\mathbb{E}\,|X_n| = \mathbb{E}\, X_n^+ + \mathbb{E}\, X_n^- = 2\mathbb{E}\, X_n^+ - \mathbb{E}\, X_n \le 2\mathbb{E}\, X_n^+ - \mathbb{E}\, X_0.$$

By Fatou's lemma, $\mathbb{E}\lim_n |X_n| \le \sup_n \mathbb{E}|X_n| < \infty$, or X_n converges a.s. to a finite limit. □

Corollary 14.3. *If X_n is a positive supermartingale or a martingale bounded above or below, X_n converges a.s.*

Proof. If X_n is a positive supermartingale, $-X_n$ is a submartingale bounded above by 0. Now apply Theorem 4.12.

If X_n is a martingale bounded above, by considering $-X_n$, we may assume X_n is bounded below. Looking at $X_n + M$ for fixed M will not affect the convergence, so we may assume X_n is bounded below by 0. Now apply the first assertion of the corollary. □

Proposition 14.4. *If X_n is a martingale with $\sup_n \mathbb{E}|X_n|^p < \infty$ for some $p > 1$, then the convergence is in L^p as well as a.s. This is also true when X_n is a submartingale. If X_n is a uniformly integrable martingale, then the convergence is in L^1. If $X_n \to X_\infty$ in L^1, then $X_n = \mathbb{E}[X_\infty \mid \mathcal{F}_n]$.*

X_n is a uniformly integrable martingale if the collection of random variables X_n is uniformly integrable.

Proof. The L^p convergence assertion follows by using Doob's inequality (Theorem 13.2) and dominated convergence. The L^1 convergence assertion follows since a.s. convergence together with uniform integrability implies L^1 convergence. Finally, if $j < n$, we have $X_j = \mathbb{E}[X_n \mid \mathcal{F}_j]$. If $A \in \mathcal{F}_j$,

$$\mathbb{E}[X_j; A] = \mathbb{E}[X_n; A] \to \mathbb{E}[X_\infty; A]$$

by the L^1 convergence of X_n to X_∞. Since this is true for all $A \in \mathcal{F}_j$, $X_j = \mathbb{E}[X_\infty \mid \mathcal{F}_j]$. □

15. Applications of martingales.

Let S_n be your fortune at time n. In a fair casino, $\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] = S_n$. If N is a stopping time, the optional stopping theorem says that $\mathbb{E}S_N = \mathbb{E}S_0$; in other words, no matter what stopping time you use and what method of betting, you will do not better on average than ending up with what you started with.

An elegant application of martingales is a proof of the SLLN. Let Y_i be i.i.d. Let $Z_n = \mathbb{E}[Y_1 \mid S_n, S_{n+1}, \ldots]$. We claim $Z_n = S_n/n$. Certainly S_n/n is $\sigma(S_n, \ldots)$ measurable. If $A \in \sigma(S_n, \ldots, S_N)$ for some n, then $A = ((S_n, \ldots, S_N) \in B)$ for some Borel subset B of $\mathbb{R}^{N-n+1}$. Since the Y_i are i.i.d., for each $k \le n$,

$$\mathbb{E}[Y_1; (S_n, \ldots, S_N) \in B] = \mathbb{E}[Y_k; (S_n, \ldots, S_N) \in B].$$

Summing over k and dividing by n,

$$\mathbb{E}[Y_1; (S_n, \ldots, S_N) \in B] = \mathbb{E}[S_n/n; (S_n, \ldots, S_N) \in B].$$

Therefore $\mathbb{E}[Y_1; A] = \mathbb{E}[S_n/n; A]$ for every $A \in \sigma(S_n, \ldots, S_N)$, and by letting $N \to \infty$, this holds for every $A \in \sigma(S_n, \ldots)$. Thus $Z_n = S_n/n$.

Fix N, let $X_k = Z_{N-k}$, and let $\mathcal{F}_k = \sigma(S_{N-k}, S_{N-k+1}, \ldots)$. It is easy to see that X_k is a martingale, and by Doob's upcrossing inequality, if U_n^X is the number of upcrossings of $[a, b]$ by X, then $\mathbb{E}U_N^X \le \mathbb{E}X_N^+/(b-a) \le \mathbb{E}|Z_0|/(b-a) = \mathbb{E}|Y_1|/(b-a)$. This differs by at most one from the number of upcrossings of $[a, b]$ by $Z_1, \ldots, Z_N$. By Fatou's lemma, the expected number of upcrossings of $[a, b]$ by $Z_1, \ldots$ is finite. Arguing as in the proof of the martingale convergence theorem, this says that $Z_n = S_n/n$ does not oscillate.

It is conceivable that $|S_n/n| \to \infty$. But by Fatou's lemma,

$$E[\lim |S_n/n|] \le \liminf \mathbb{E}|S_n/n| \le \liminf n\mathbb{E}|Y_1|/n = \mathbb{E}|Y_1| < \infty.$$

Another application of martingale techniques is Wald's identities.

Proposition 15.1. *Suppose the Y_i are i.i.d. with $\mathbb{E}\,|Y_1| < \infty$, N is a stopping time with $\mathbb{E}\,N < \infty$, and N is independent of the Y_i. Then $\mathbb{E}\,S_N = (\mathbb{E}\,N)(EY_1)$, where the S_n are the partial sums of the Y_i.*

Proof. $S_n - n(\mathbb{E}\,Y_1)$ is a martingale, so $\mathbb{E}\,S_{n\wedge N} = \mathbb{E}\,(n \wedge N)\mathbb{E}\,Y_1$ by optional stopping. The right hand side tends to $(\mathbb{E}\,N)(\mathbb{E}\,Y_1)$ by monotone convergence. $S_{n\wedge N}$ converges almost surely to S_N, and we need to show the expected values converge.

Note

$$\begin{aligned}|S_{n\wedge N}| &= \sum_{k=0}^{\infty} |S_{n\wedge k}| 1_{(N=k)} \le \sum_{k=0}^{\infty}\sum_{j=0}^{n\wedge k} |Y_j| 1_{(N=k)} \\ &= \sum_{j=0}^{n}\sum_{k>j}^{\infty} |Y_j| 1_{(N=k)} = \sum_{j=0}^{n} |Y_j| 1_{(N\ge j)} \le \sum_{j=0}^{\infty} |Y_j| 1_{(N\ge j)}.\end{aligned}$$

The last expression, using the independence, has expected value

$$\sum_{j=0}^{\infty} (\mathbb{E}\,|Y_j|)\mathbb{P}(N \ge j) \le (\mathbb{E}\,|Y_1|)(1 + \mathbb{E}\,N) < \infty.$$

So by dominated convergence, we have $\mathbb{E}\,S_{n\wedge N} \to \mathbb{E}\,S_N$. □

Wald's second identity is a similar expression for the variance of S_N.

We can use martingales to find certain hitting probabilities.

Proposition 15.2. *Suppose the Y_i are i.i.d. with $\mathbb{P}(Y_1 = 1) = 1/2$, $\mathbb{P}(Y_1 = -1) = 1/2$, and S_n the partial sum process. Suppose a and b are positive integers. Then*

$$\mathbb{P}(S_n \text{ hits } -a \text{ before } b) = \frac{b}{a+b}.$$

If $N = \min\{n : S_n \in \{-a, b\}\}$, then $\mathbb{E}\,N = ab$.

Proof. $S_n^2 - n$ is a martingale, so $\mathbb{E}\,S_{n\wedge N}^2 = \mathbb{E}\,n \wedge N$. Let $n \to \infty$. The right hand side converges to $\mathbb{E}\,N$ by monotone convergence. Since $S_{n\wedge N}$ is bounded in absolute value by $a + b$, the left hand side converges by dominated convergence to $\mathbb{E}\,S_N^2$, which is finite. So $\mathbb{E}\,N$ is finite, hence N is finite almost surely.

S_n is a martingale, so $\mathbb{E}\,S_{n\wedge N} = \mathbb{E}\,S_0 = 0$. By dominated convergence, and the fact that $N < \infty$ a.s., hence $S_{n\wedge N} \to S_N$, we have $\mathbb{E}\,S_N = 0$, or

$$-a\mathbb{P}(S_N = -a) + b\mathbb{P}(S_N = b) = 0.$$

We also have

$$\mathbb{P}(S_N = -a) + \mathbb{P}(S_N = b) = 1.$$

Solving these two equations for $\mathbb{P}(S_N = -a)$ and $\mathbb{P}(S_N = b)$ yields our first result. Since $\mathbb{E}\,N = \mathbb{E}\,S_N^2 = a^2\mathbb{P}(S_N = -a) + b^2\mathbb{P}(S_N = b)$, substituting gives the second result. □

Based on this proposition, if we let $a \to \infty$, we see that $\mathbb{P}(N_b < \infty) = 1$ and $\mathbb{E}\,N_b = \infty$, where $N_b = \min\{n : S_n = b\}$.

Next we give a version of the Borel-Cantelli lemma.

Proposition 15.3. *Suppose $A_n \in \mathcal{F}_n$. Then (A_n i.o.) and ($\sum_{n=1}^{\infty} \mathbb{P}(A_n \mid \mathcal{F}_{n-1}) = \infty$) differ by a null set.*

Proof. Let $X_n = \sum_{m=1}^{n} [1_{A_m} - \mathbb{P}(A_m \mid \mathcal{F}_{m-1})]$. Note $|X_n - X_{n-1}| \leq 1$. Also, it is easy to see that $\mathbb{E}\,[X_n - X_{n-1} \mid \mathcal{F}_{n-1}] = 0$, so X_n is a martingale.

We claim that for almost every ω either $\lim X_n$ exists and is finite, or else $\limsup X_n = \infty$ and $\liminf X_n = -\infty$. In fact, if $N = \min\{n : X_n \leq -k\}$, then $X_{n\wedge N} \geq -k - 1$, so $X_{n\wedge N}$ converges by the martingale convergence theorem. Therefore $\lim X_n$ exists and is finite on $(N = \infty)$. So if $\lim X_n$ does not exist or is not finite, then $N < \infty$. This is true for all k, hence $\liminf X_n = -\infty$. A similar argument shows $\limsup X_n = \infty$ in this case.

Now if $\lim X_n$ exists and is finite, then $\sum_{n=1}^{\infty} 1_{A_n} = \infty$ if and only if $\sum \mathbb{P}(A_n \mid \mathcal{F}_{n-1}) < \infty$. On the other hand, if the limit does not exist or is not finite, then $\sum 1_{A_n} = \infty$ and $\sum \mathbb{P}(A_n \mid \mathcal{F}_{n-1}) = \infty$. □

16. Weak convergence.

We will see later that if the X_i are i.i.d. with mean zero and variance one, then $S_n/\sqrt{n}$ converges in the sense

$$\mathbb{P}(S_n/\sqrt{n} \in [a, b]) \to \mathbb{P}(Z \in [a, b]),$$

where Z is a standard normal. If $S_n/\sqrt{n}$ converged in probability or almost surely, then by the zero-one law it would converge to a constant, contradicting the above. We want to generalize the above type of convergence.

We say F_n *converges weakly* to F if $F_n(x) \to F(x)$ for all x at which F is continuous. Here F_n and F are distribution functions. We say X_n *converges weakly* to X if F_{X_n} converges weakly to F_X. We sometimes say X_n *converges in distribution* or *converges in law* to X. Probabilities μ_n converge weakly if their corresponding distribution functions converges, that is, if $F_{\mu_n}(x) = \mu_n(-\infty, x]$ converges weakly.

An example that illustrates why we restrict the convergence to continuity points of F is the following. Let $X_n = 1/n$ with probability one, and $X = 0$ with probability one. $F_{X_n}(x)$ is 0 if $x < 1/n$ and 1 otherwise. $F_{X_n}(x)$ converges to $F_X(x)$ for all x except $x = 0$.

Proposition 16.1. *X_n converges weakly to X if and only if $\mathbb{E}\, g(X_n) \to \mathbb{E}\, g(X)$ for all g bounded and continuous.*

The idea that $\mathbb{E}\, g(X_n)$ converges to $\mathbb{E}\, g(X)$ for all g bounded and continuous makes sense for any metric space and is used as a definition of weak convergence for X_n taking values in general metric spaces.

Proof. First suppose $\mathbb{E}\, g(X_n)$ converges to $\mathbb{E}\, g(X)$. Let x be a continuity point of F, let $\varepsilon > 0$, and choose δ such that $|F(y) - F(x)| < \varepsilon$ if $|y - x| < \delta$. Choose g continuous such that g is one on $(-\infty, x]$, takes values between 0 and 1, and is 0 on $[x + \delta, \infty)$. Then $F_{X_n}(x) \leq \mathbb{E}\, g(X_n) \to \mathbb{E}\, g(X) \leq F_X(x + \delta) \leq F(x) + \varepsilon$.

Similarly, if h is a continuous function taking values between 0 and 1 that is 1 on $(-\infty, x - \delta]$ and 0 on $[x, \infty)$, $F_{X_n}(x) \geq \mathbb{E}\, h(X_n) \to \mathbb{E}\, h(X) \geq F_X(x - \delta) \geq F(x) - \varepsilon$. Since ε is arbitrary, $F_{X_n}(x) \to F_X(x)$.

Now suppose X_n converges weakly to X. If a and b are continuity points of F and of all the F_{X_n}, then $\mathbb{E}\, 1_{[a,b]}(X_n) = F_{X_n}(b) - F_{X_n}(a) \to F(b) - F(a) = \mathbb{E}\, 1_{[a,b]}(X)$. By taking linear combinations, we have $\mathbb{E}\, g(X_n) \to \mathbb{E}\, g(X)$ for every g which is a step function where the end points of the intervals are continuity points for all the F_{X_n} and for F_X. Since the set of points that are not a continuity point for some F_{X_n} or for F_X is countable, and we can approximate any continuous function on an interval by such step functions uniformly, we have $\mathbb{E}\, g(X_n) \to \mathbb{E}\, g(X)$ for all g such that the support of g is a closed interval whose endpoints are continuity points of F_X and g is continuous on its support.

Let $\varepsilon > 0$ and choose M such that $F_X(M) > 1 - \varepsilon$ and $F_X(-M) < \varepsilon$ and so that M and $-M$ are continuity points of F_X and of the F_{X_n}. By the above argument, $\mathbb{E}\,(1_{[-M,M]}g)(X_n) \to \mathbb{E}\,(1_{[-M,M]}g)(X)$,

where g is a bounded continuous function. The difference between $\mathbb{E}\,(1_{[-M,M]}g)(X)$ and $\mathbb{E}\,g(X)$ is bounded by $\|g\|_\infty \mathbb{P}(X \notin [-M,M]) \le 2\varepsilon\|g\|_\infty$. Similarly, when X is replaced by X_n, the difference is bounded by $\|g\|_\infty \mathbb{P}(X_n \notin [-M,M]) \to \|g\|_\infty \mathbb{P}(X \notin [-M,M])$. So for n large, it is less than $3\varepsilon\|g\|_\infty$. Since ε is arbitrary, $\mathbb{E}\,g(X_n) \to \mathbb{E}\,g(X)$ whenever g is bounded and continuous. □

Let us examine the relationship between weak convergence and convergence in probability. The example of $S_n/\sqrt{n}$ shows that one can have weak convergence without convergence in probability.

Proposition 16.2. *(a) If X_n converges to X in probability, then it converges weakly.*

(b) If X_n converges weakly to a constant, it converges in probability.

(c) (Slutsky's theorem) *If X_n converges weakly to X and Y_n converges weakly to a constant c, then $X_n + Y_n$ converges weakly to $X + c$ and X_nY_n converges weakly to cX.*

Proof. To prove (a), let g be a bounded and continuous function. If n_j is any subsequence, then there exists a further subsequence such that $X(n_{j_k})$ converges almost surely to X. Then by dominated convergence, $\mathbb{E}\,g(X(n_{j_k})) \to \mathbb{E}\,g(X)$. That suffices to show $\mathbb{E}\,g(X_n)$ converges to $\mathbb{E}\,g(X)$.

For (b), if X_n converges weakly to c,

$$\mathbb{P}(X_n - c > \varepsilon) = \mathbb{P}(X_n > c + \varepsilon) = 1 - \mathbb{P}(X_n \le c + \varepsilon) \to 1 - \mathbb{P}(c \le c + \varepsilon) = 0.$$

We use the fact that if $Y \equiv c$, then $c + \varepsilon$ is a point of continuity for F_Y. A similar equation shows $\mathbb{P}(X_n - c \le -\varepsilon) \to 0$, so $\mathbb{P}(|X_n - c| > \varepsilon) \to 0$.

We now prove the first part of (c), leaving the second part for the reader. Let x be a point such that $x - c$ is a continuity point of F_X. Choose ε so that $x - c + \varepsilon$ is again a continuity point. Then

$$\mathbb{P}(X_n + Y_n \le x) \le \mathbb{P}(X_n + c \le x + \varepsilon) + \mathbb{P}(|Y_n - c| > \varepsilon) \to \mathbb{P}(X \le x - c + \varepsilon).$$

So $\limsup \mathbb{P}(X_n + Y_n \le x) \le \mathbb{P}(X + c \le x + \varepsilon)$. Since ε can be as small as we like and $x - c$ is a continuity point of F_x, then $\limsup \mathbb{P}(X_n + Y_n \le x) \le \mathbb{P}(X + c \le x)$. The lim inf is done similarly. □

We say a sequence of distribution functions is *tight* if for each $\varepsilon > 0$ there exists M such that $F_n(M) \ge 1 - \varepsilon$ and $F_n(-M) \le \varepsilon$. A sequence of r.v.s is tight if the corresponding distribution functions are tight; this is equivalent to $\mathbb{P}(|X_n| \ge M) \le \varepsilon$.

Theorem 16.3. (Helly's theorem) *Let F_n be a sequence of distribution functions that is tight. There exists a subsequence n_j and a distribution function F such that F_{n_j} converges weakly to F.*

What could happen is that $X_n = n$, so that $F_{X_n} \to 0$; the tightness precludes this.

Proof. Let q_k be an enumeration of the rationals. Since $F_n(q_k) \in [0,1]$, any subsequence has a further subsequence that converges. Use diagonalization so that $F_{n_j}(q_k)$ converges for each q_k and call the limit $F(q_k)$. F is nondecreasing, and define $F(x) = \inf_{q_k \ge x} F(q_k)$. So F is right continuous and nondecreasing.

If x is a point of continuity of F and $\varepsilon > 0$, then there exist r and s rational such that $r < x < s$ and $F(s) - \varepsilon < F(x) < F(r) + \varepsilon$. Then

$$F_{n_j}(x) \ge F_{n_j}(r) \to F(r) > F(x) - \varepsilon$$

and

$$F_{n_j}(x) \le F_{n_j}(s) \to F(s) < F(x) + \varepsilon.$$

Since ε is arbitrary, $F_{n_j}(x) \to F(x)$.

Since the F_n are tight, there exists M such that $F_n(-M) < \varepsilon$. Then $F(-M) \le \varepsilon$, which implies $\lim_{x\to-\infty} F(x) = 0$. Showing $\lim_{x\to\infty} F(x) = 1$ is similar. Therefore F is in fact a distribution function. □

We conclude by giving an easily checked criterion for tightness.

Proposition 16.4. *Suppose there exists $\varphi : [0,\infty) \to [0,\infty)$ that is increasing and $\varphi(x) \to \infty$ as $x \to \infty$. If $c = \sup_n \mathbb{E}\,\varphi(|X_n|) < \infty$, then the X_n are tight.*

Proof. Let $\varepsilon > 0$. Choose M such that $\varphi(x) \geq c/\varepsilon$ if $x > M$. Then

$$\mathbb{P}(|X_n| > M) \leq \int \frac{\varphi(|X_n|)}{c/\varepsilon} 1_{(|X_n|>M)} d\mathbb{P} \leq \frac{\varepsilon}{c}\mathbb{E}\,\varphi(|X_n|) \leq \varepsilon.$$

□

17. Characteristic functions.

We define the *characteristic function* of a random variable X by $\varphi_X(t) = \mathbb{E}\,e^{itx}$ for $t \in \mathbb{R}$.

Note that $\varphi_X(t) = \int e^{itx}\mathbb{P}_X(dx)$. So if X and Y have the same law, they have the same characteristic function. Also, if the law of X has a density, that is, $\mathbb{P}_X(dx) = f_X(x)\,dx$, then $\varphi_X(t) = \int e^{itx} f_X(x)\,dx$, so in this case the characteristic function is the same as (one definition of) the Fourier transform of f_X.

Proposition 17.1. *$\varphi(0) = 1$, $|\varphi(t)| \leq 1$, $\varphi(-t) = \overline{\varphi(t)}$, and φ is uniformly continuous.*

Proof. Since $|e^{itx}| \leq 1$, everything follows immediately from the definitions except the uniform continuity. For that we write

$$|\varphi(t+h) - \varphi(t)| = |\mathbb{E}\,e^{i(t+h)X} - \mathbb{E}\,e^{itX}| \leq \mathbb{E}\,|e^{itX}(e^{ihX} - 1)| = \mathbb{E}\,|e^{ihX} - 1|.$$

$|e^{ihX} - 1|$ tends to 0 almost surely as $h \to 0$, so the right hand side tends to 0 by dominated convergence. Note that the right hand side is independent of t. □

Proposition 17.2. *$\varphi_{aX}(t) = \varphi_X(at)$ and $\varphi_{X+b}(t) = e^{itb}\varphi_X(t)$,*

Proof. The first follows from $\mathbb{E}\,e^{it(aX)} = \mathbb{E}\,e^{i(at)X}$, and the second is similar. □

Proposition 17.3. *If X and Y are independent, then $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$.*

Proof. From the multiplication theorem,

$$\mathbb{E}\,e^{it(X+Y)} = \mathbb{E}\,e^{itX}e^{itY} = \mathbb{E}\,e^{itX}\mathbb{E}\,e^{itY}.$$

□

Note that if X_1 and X_2 are independent and identically distributed, then

$$\varphi_{X_1-X_2}(t) = \varphi_{X_1}(t)\varphi_{-X_2}(t) = \varphi_{X_1}(t)\varphi_{X_2}(-t) = \varphi_{X_1}(t)\overline{\varphi_{X_2}(t)} = |\varphi_{X_1}(t)|^2.$$

Let us look at some examples of characteristic functions.

(a) *Bernoulli*: By direct computation, this is $pe^{it} + (1-p) = 1 - p(1 - e^{it})$.

(b) *Coin flip*: (i.e., $\mathbb{P}(X = +1) = \mathbb{P}(X = -1) = 1/2$) We have $\frac{1}{2}e^{it} + \frac{1}{2}e^{-it} = \cos t$.

(c) *Poisson*:

$$\mathbb{E}\,e^{itX} = \sum_{k=0}^{\infty} e^{itk} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum \frac{(\lambda e^{it})^k}{k!} = e^{-\lambda}e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}.$$

(d) *Point mass at a*: $\mathbb{E}\,e^{itX} = e^{ita}$. Note that when $a = 0$, then $\varphi \equiv 1$.

(e) *Binomial*: Write X as the sum of n independent Bernoulli r.v.s B_i. So

$$\varphi_X(t) = \prod_{i=1}^{n} \varphi_{B_i}(t) = [\varphi_{B_i}(t)]^n = [1 - p(1 - e^{it})]^n.$$

(f) *Geometric*:

$$\varphi(t) = \sum_{k=0}^{\infty} p(1-p)^k e^{itk} = p \sum ((1-p)e^{it})^k = \frac{p}{1-(1-p)e^{it}}.$$

(g) *Uniform on* $[a, b]$:

$$\varphi(t) = \frac{1}{b-a} \int_a^b e^{itx} dx = \frac{e^{itb} - e^{ita}}{(b-a)it}.$$

Note that when $a = -b$ this reduces to $\sin(bt)/bt$.

(h) *Exponential*:

$$\int_0^\infty \lambda e^{itx} e^{-\lambda x}\, dx = \lambda \int_0^\infty e^{(it-\lambda)x} dx = \frac{\lambda}{\lambda - it}.$$

(i) *Standard normal*:

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2} dx.$$

This can be done by completing the square and then doing a contour integration. Alternately, $\varphi'(t) = (1/\sqrt{2\pi}) \int_{-\infty}^{\infty} ixe^{itx} e^{-x^2/2} dx$. (do the real and imaginary parts separately, and use the dominated convergence theorem to justify taking the derivative inside.) Integrating by parts (do the real and imaginary parts separately), this is equal to $-t\varphi(t)$. The only solution to $\varphi'(t) = -t\varphi(t)$ with $\varphi(0) = 1$ is $\varphi(t) = e^{-t^2/2}$.

(j) *Normal with mean* μ *and variance* σ^2: Writing $X = \sigma Z + \mu$, where Z is a standard normal, then $\varphi_X(t) = e^{i\mu t} \varphi_Z(\sigma t) = e^{i\mu t - \sigma^2 t^2/2}$.

(k) *Cauchy*: We have

$$\varphi(t) = \frac{1}{\pi} \int \frac{e^{itx}}{1+x^2} dx.$$

This is a standard exercise in contour integration in complex analysis. The answer is $e^{-|t|}$.

18. Inversion formula.

We need a preliminary real variable lemma, and then we can proceed to the inversion formula, which gives a formula for the distribution function in terms of the characteristic function.

Lemma 18.1. *(a)* $\int_0^N (\sin(Ax)/x)\, dx \to \operatorname{sgn}(A)\pi/2$ *as* $N \to \infty$.
(b) $\sup_a |\int_0^a (\sin(Ax)/x)\, dx| < \infty$.

Proof. If $A = 0$, this is clear. The case $A < 0$ reduces to the case $A > 0$ by the fact that sin is an odd function. By a change of variables $y = Ax$, we reduce to the case $A = 1$. Part (a) is a standard result in contour integration, and part (b) comes from the fact that the integral can be written as an alternating series. □

Theorem 18.2. (Inversion formula) *Let* μ *be a probability measure and let* $\varphi(t) = \int e^{itx} \mu(dx)$. *If* $a < b$, *then*

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

The example where μ is point mass at 0, so $\varphi(t) = 1$, shows that one needs to take a limit, since the integrand in this case is $2 \sin t/t$, which is not integrable.

Proof. By Fubini,

$$\begin{aligned}\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt &= \int_{-T}^{T} \int \frac{e^{-ita} - e^{-itb}}{it} e^{itx} \mu(dx)\, dt \\ &= \int \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} e^{itx} dt\, \mu(dx).\end{aligned}$$

To justify this, we bound the integrand by the mean value theorem

Expanding e^{-itb} and e^{-ita} using Euler's formula, and using the fact that cos is an even function and sin is odd, we are left with

$$\int 2\Big[\int_0^T \frac{\sin(t(x-a))}{t} dt - \int_0^T \frac{\sin(t(x-b))}{t} dt \Big]\, \mu(dx).$$

Using Lemma 18.1 and dominated convergence, this tends to

$$\int [\pi \mathrm{sgn}\,(x-a) - \pi \mathrm{sgn}\,(x-b)] \mu(dx).$$

□

Theorem 18.3. *If $\int |\varphi(t)|\, dt < \infty$, then μ has a bounded density f and*

$$f(y) = \frac{1}{2\pi} \int e^{-ity} \varphi(t)\, dt.$$

Proof.

$$\begin{aligned}&\mu(a,b) + \frac{1}{2}\mu(\{a\}) + \frac{1}{2}\mu(\{b\}) \\ &\quad = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt \\ &\quad = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt \\ &\quad \le \frac{b-a}{2\pi} \int |\varphi(t)|\, dt.\end{aligned}$$

Letting $b \to a$ shows that μ has no point masses.

We now write

$$\begin{aligned}\mu(x, x+h) &= \frac{1}{2\pi} \int \frac{e^{-itx} - e^{-it(x+h)}}{it} \varphi(t) dt \\ &\quad \frac{1}{2\pi} \int \Big(\int_x^{x+h} e^{-ity} dy \Big) \varphi(t)\, dt \\ &\quad = \int_x^{x+h} \Big(\frac{1}{2\pi} \int e^{-ity} \varphi(y)\, dt \Big) dy.\end{aligned}$$

So μ has density $(1/2\pi) \int e^{-ity} \varphi(t)\, dt$. As in the proof of Proposition 17.1, we see f is continuous. □

A corollary to the inversion formula is the uniqueness theorem.

Theorem 18.3. *If $\varphi_X = \varphi_Y$, then $\mathbb{P}_X = \mathbb{P}_Y$.*

The following proposition can be proved directly, but the proof using characteristic functions is much easier.

Proposition 18.4. *(a) If X and Y are independent, X is a normal with mean a and variance b^2, and Y is a normal with mean c and variance d^2, then $X+Y$ is normal with mean $a+c$ and variance b^2+d^2.*

(b) If X and Y are independent, X is Poisson with parameter λ_1, and Y is Poisson with parameter λ_2, then $X+Y$ is Poisson with parameter $\lambda_1+\lambda_2$.

(c) If X_i are i.i.d. Cauchy, then S_n/n is Cauchy.

Proof. For (a),

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t) = e^{iat-b^2t^2/2}e^{ict-c^2t^2/2} = e^{i(a+c)t-(b^2+d^2)t^2/2}.$$

Now use the uniqueness theorem.

Parts (b) and (c) are proved similarly. □

19. Continuity theorem.

Lemma 19.1. *Suppose φ is the characteristic function of a probability μ. Then*

$$\mu([-2A,2A]) \geq A\Big|\int_{-1/A}^{1/A} \varphi(t)\,dt\Big| - 1.$$

Proof. Note

$$\begin{aligned}\frac{1}{2T}\int_{-T}^{T}\varphi(t)\,dt &= \frac{1}{2T}\int_{-T}^{T}\int e^{itx}\mu(dx)\,dt\\ &= \int\int \frac{1}{2T}1_{[-T,T]}(t)e^{itx}\,dt\mu(dx)\\ &= \int \frac{\sin Tx}{Tx}\mu(dx).\end{aligned}$$

Since $|(\sin(Tx))/Tx| \leq 1$ and $|\sin(Tx)| \leq 1$, then $|(\sin(Tx))/Tx| \leq 1/2TA$ if $|x| \geq 2A$. So

$$\begin{aligned}\Big|\int \frac{\sin Tx}{Tx}\mu(dx)\Big| &\leq \mu([-2A,2A]) + \int_{[-2A,2A]^c}\frac{1}{2TA}\mu(dx)\\ &= \mu([-2A,2A]) + \frac{1}{2TA}(1-\mu([-2A,2A])\\ &= \frac{1}{2TA} + \Big(1-\frac{1}{2TA}\Big)\mu([-2A,2A]).\end{aligned}$$

Setting $T = 1/A$,

$$\Big|\frac{A}{2}\int_{-1/A}^{1/A}\varphi(t)\,dt\Big| \leq \frac{1}{2} + \frac{1}{2}\mu([-2A,2A]).$$

Now multiply both sides by 2. □

Proposition 19.2. *If μ_n converges weakly to μ, then φ_n converges to φ uniformly on every finite interval.*

Proof. e^{itx} is continuous and bounded, so $\int e^{itx}\mu_n(dx) \to \int e^{itx}\mu(dx)$. (Look at the real and imaginary parts, and think of the μ_n and μ as the laws of some r.v.s X_n and X.) Since

$$|\varphi_n(t+h)-\varphi_n(t)| \leq \int |e^{ihx}-1|\mu_n(dx) \to \int |e^{ihx}-1|\mu(dx),$$

it follows easily that the φ_n are equicontinuous. Therefore the convergence is uniform on finite intervals. □

The interesting result of this section is the converse, Lévy's continuity theorem.

Theorem 19.3. *Suppose μ_n are probabilities, $\varphi_n(t)$ converges to a function $\varphi(t)$ for each t, and φ is continuous at 0. Then φ is the characteristic function of a probability μ and μ_n converges weakly to μ.*

Proof. Let $\varepsilon > 0$. Since φ is continuous at 0, choose δ small so that

$$\Big|\frac{1}{2\delta}\int_{-\delta}^{\delta}\varphi(t)\,dt - 1\Big| < \varepsilon.$$

Using the dominated convergence theorem, choose N such that

$$\frac{1}{2\delta}\int_{-\delta}^{\delta}|\varphi_n(t)-\varphi(t)|\,dt < \varepsilon$$

if $n \geq N$. So if $n \geq N$,

$$\begin{aligned}\Big|\frac{1}{2\delta}\int_{-\delta}^{\delta}\varphi_n(t)\,dt\Big| &\geq \Big|\frac{1}{2\delta}\int_{-\delta}^{\delta}\varphi(t)\,dt\Big| - \frac{1}{2\delta}\int_{-\delta}^{\delta}|\varphi_n(t)-\varphi(t)|\,dt\\ &\geq 1-2\varepsilon.\end{aligned}$$

By Lemma 19.1. with $A = 1/\delta$, for such n,

$$\mu_n[-2/\delta, 2/\delta] \geq 2(1-2\varepsilon) - 1 = 1 - 4\varepsilon.$$

This shows the μ_n are tight.

Let n_j be a subsequence such that μ_{n_j} converges weakly, say to μ. Then $\varphi_{n_j}(t) \to \varphi_\mu(t)$, hence $\varphi(t) = \varphi_\mu(t)$, or φ is the characteristic function of a probability μ. If μ' is any subsequential weak limit point of μ_n, then $\varphi_{\mu'}(t) = \varphi(t) = \varphi_\mu(t)$; so μ' must equal μ. Hence μ_n converges weakly to μ. □

We need the following estimate on moments.

Proposition 19.4. *If $\mathbb{E}\,|X|^k < \infty$ for an integer k, then φ_X has a continuous derivative of order k and*

$$\varphi^{(k)}(t) = \int (ix)^k e^{itx}\,\mathbb{P}_X(dx).$$

In particular, $\varphi^{(k)}(0) = i^k\,\mathbb{E}\,X^k$.

Proof. Write

$$\frac{\varphi(t+h)-\varphi(t)}{h} = \int \frac{e^{i(t+h)x} - e^{itx}}{h}\,\mathbb{P}(dx).$$

The integrand is bounded by $|x|$. So if $\int |x|\mathbb{P}_X(dx) < \infty$, we can use dominated convergence to obtain the desired formula for $\varphi'(t)$. As in the proof of Proposition 17.1, we see $\varphi'(t)$ is continuous. We do the case of general k by induction. Evaluating $\varphi^{(k)}$ at 0 gives the particular case. □

One nice application of the continuity theorem is a proof of the weak law of large numbers. Its proof is very similar to the proof of the central limit theorem, which we give in the next section.

Another nice use of characteristic functions and martingales is the following.

Proposition 19.5. *Suppose X_i is a sequence of independent r.v.s and S_n converges weakly. Then S_n converges almost surely.*

Proof. Suppose S_n converges weakly to W. Then $\varphi_{S_n}(t) \to \varphi_W(t)$ uniformly on compact sets by Proposition 19.2. Since $\varphi_W(0) = 1$ and φ_W is continuous, there exists δ such that $|\varphi_W(t) - 1| < 1/2$ if $|t| < \delta$. So for n large, $\varphi_{S_n}(t)| \geq 1/4$ if $|t| < \delta$.

Note

$$\mathbb{E}\left[e^{itS_n} \mid X_1, \ldots X_{n-1}\right] = e^{itS_{n-1}}\mathbb{E}\left[e^{itX_n} \mid X_1, \ldots, X_{n-1}\right] = e^{itS_{n-1}}\varphi_{X_n}(t).$$

Since $\varphi_{S_n}(t) = \prod \varphi_{X_i}(t)$, it follows that $e^{itS_n}/\varphi_{S_n}(t)$ is a martingale.

Therefore for $|t| < \delta$ and n large, $e^{itS_n}/\varphi_{S_n}(t)$ is a bounded martingale, and hence converges almost surely. Since $\varphi_{S_n}(t) \to \varphi_W(t) \neq 0$, then e^{itS_n} converges almost surely if $|t| < \delta$.

Let $A = \{(\omega, t) \in \Omega \times (-\delta, \delta) : e^{itS_n(\omega)} \text{ does not converge}\}$. For each t, we have almost sure convergence, so $\int 1_A(\omega, t)\,\mathbb{P}(d\omega) = 0$. Therefore $\int_{-\delta}^{\delta}\int 1_A\,d\mathbb{P}\,dt = 0$, and by Fubini, $\int\int_{-\delta}^{\delta} 1_A\,dt\,d\mathbb{P} = 0$. Hence almost surely, $\int 1_A(\omega, t)\,dt = 0$. This means, there exists a set N with $\mathbb{P}(N) = 0$, and if $\omega \notin N$, then $e^{itS_n(\omega)}$ converges for almost every $t \in (-\delta, \delta)$.

If $\omega \notin N$, by dominated convergence, $\int_0^a e^{itS_n(\omega)}dt$ converges, provided $a < \delta$. But

$$\int_0^a e^{itS_n(\omega)}\,dt = \frac{e^{iaS_n(\omega)} - 1}{iS_n(\omega)}$$

if $S_n(\omega) \neq 0$ and equals a otherwise. Since the left hand side converges and $e^{iaS_n(\omega)}$ converges, then $S_n(\omega)$ converges. □

20. Central limit theorem.

The simplest case of the central limit theorem (CLT) is the case when the X_i are i.i.d., with mean zero and variance one, and then the CLT says that $S_n/\sqrt{n}$ converges weakly to a standard normal. We first prove this case.

We need the fact that if c_n are complex numbers converging to c, then $(1 + (c_n/n))^n \to e^c$. We leave the proof of this to the reader, with the warning that any proof using logarithms needs to be done with some care, since $\log z$ is a multi-valued function when z is complex.

Theorem 20.1. *Suppose the X_i are i.i.d., mean zero, and variance one. Then $S_n/\sqrt{n}$ converges weakly to a standard normal.*

Proof. Since X_1 has finite second moment, then φ_{X_1} has a continuous second derivative. By Taylor's theorem,

$$\varphi_{X_1}(t) = \varphi_{X_1}(0) + \varphi'_{X_1}(0)t + \varphi''_{X_1}(0)t^2/2 + R(t),$$

where $|R(t)|/t^2 \to 0$ as $|t| \to 0$. So

$$\varphi_{X_1}(t) = 1 - t^2/2 + R(t).$$

Then

$$\varphi_{S_n/\sqrt{n}}(t) = \varphi_{S_n}(t/\sqrt{n}) = (\varphi_{X_1}(t/\sqrt{n}))^n = \left[1 - \frac{t^2}{2n} + R(t/\sqrt{n})\right]^n.$$

Since $t/\sqrt{n}$ converges to zero as $n \to \infty$, we have

$$\varphi_{S_n/\sqrt{n}}(t) \to e^{-t^2/2}.$$

Now apply the continuity theorem. □

Let us give another proof of this simple CLT that does not use characteristic functions. For simplicity let X_i be i.i.d. mean zero variance one random variables with $\mathbb{E}\,|X_i|^3 < \infty$.

Proposition 20.2. *With X_i as above, $S_n/\sqrt{n}$ converges weakly to a standard normal.*

Proof. Let $Y_1, \dots, Y_n$ be i.i.d. standard normal r.v.s that are independent of the X_i. Let $Z_1 = Y_2 + \cdots + Y_n$, $Z_2 = X_1 + Y_3 + \cdots + Y_n$, $Z_3 = X_1 + X_2 + Y_4 + \cdots + Y_n$, etc.

Let us suppose $g \in C^3$ with compact support and let W be a standard normal. Our first goal is to show

$$|\mathbb{E}\, g(S_n/\sqrt{n}) - \mathbb{E}\, g(W)| \to 0. \tag{20.1}$$

We have

$$\begin{aligned}\mathbb{E}\, g(S_n/\sqrt{n}) - \mathbb{E}\, g(W) &= \mathbb{E}\, g(S_n/\sqrt{n}) - \mathbb{E}\, g(\sum_{i=1}^n Y_i/\sqrt{n}) \\ &= \sum_{i=1}^n \Big[\mathbb{E}\, g\Big(\frac{X_i + Z_i}{\sqrt{n}}\Big) - \mathbb{E}\, g\Big(\frac{Y_i + Z_i}{\sqrt{n}}\Big)\Big].\end{aligned}$$

By Taylor's theorem,

$$\mathbb{E}\, g\Big(\frac{X_i + Z_i}{\sqrt{n}}\Big) = g(Z_i/\sqrt{n}) + g'(Z_i/\sqrt{n})\frac{X_i}{\sqrt{n}} + \frac{1}{2}g''(Z_i/\sqrt{n})X_i^2 + R_n,$$

where $|R_n| \le \|g'''\|_\infty |X_i|^3/n^{3/2}$. Taking expectations and using the independence,

$$\mathbb{E}\, g\Big(\frac{X_i + Z_i}{\sqrt{n}}\Big) = \mathbb{E}\, g(Z_i/\sqrt{n}) + 0 + \frac{1}{2}\mathbb{E}\, g''(Z_i/\sqrt{n}) + \mathbb{E}\, R_n.$$

We have a very similar expression for $\mathbb{E}\, g((Y_i + Z_i)/\sqrt{n})$. Taking the difference,

$$\Big|\mathbb{E}\, g\Big(\frac{X_i + Z_i}{\sqrt{n}}\Big) - \mathbb{E}\, g\Big(\frac{Y_i + Z_i}{\sqrt{n}}\Big)\Big| \le \|g'''\|_\infty \frac{\mathbb{E}\,|X_i|^3 + \mathbb{E}\,|Y_i|^3}{n^{3/2}}.$$

Summing over i from 1 to n, we have (20.1).

By approximating continuous functions with compact support by C^3 functions with compact support, we have (20.1) for such g. Since $\mathbb{E}\,(S_n/\sqrt{n})^2 = 1$, the sequence $S_n/\sqrt{n}$ is tight. So given ε there exists M such that $\mathbb{P}(|S_n/\sqrt{n}| > M) < \varepsilon$ for all n. By taking M larger if necessary, we also have $\mathbb{P}(|W| > M) < \varepsilon$. Suppose g is bounded and continuous. Let ψ be a continuous function with compact support that is bounded by one, is nonnegative, and that equals 1 on $[-M, M]$. By (20.1) applied to $g\psi$,

$$|\mathbb{E}\,(g\psi)(S_n/\sqrt{n}) - \mathbb{E}\,(g\psi)(W)| \to 0.$$

However,

$$|\mathbb{E}\, g(S_n/\sqrt{n}) - \mathbb{E}\,(g\psi)(S_n/\sqrt{n})| \le \|g\|_\infty \mathbb{P}(|S_n/\sqrt{n}| > M) < \varepsilon\|g\|_\infty,$$

and similarly

$$|\mathbb{E}\, g(W) - \mathbb{E}\,(g\psi)(W)| < \varepsilon\|g\|_\infty.$$

Since ε is arbitrary, this proves (20.1) for bounded continuous g. By Proposition 16.1, this proves our proposition. □

We give another example of the use of characteristic functions.

Proposition 20.3. *Suppose for each n the r.v.s X_{ni}, $i = 1, \dots, n$ are i.i.d. Bernoullis with parameter p_n. If $np_n \to \lambda$ and $S_n = \sum_{i=1}^n X_{ni}$, then S_n converges weakly to a Poisson r.v. with parameter λ.*

Proof. We write

$$\begin{aligned}\varphi_{S_n}(t) &= [\varphi_{X_{n1}}(t)]^n = \Big[1 + p_n(e^{it} - 1)\Big]^n \\ &= \Big[1 + \frac{np_n}{n}(e^{it} - 1)\Big]^n \to e^{\lambda(e^{it}-1)}.\end{aligned}$$

Now apply the continuity theorem. □

A much more general theorem than Theorem 20.1 is the Lindeberg-Feller theorem.

Theorem 20.4. *Suppose for each n, X_{ni}, $i = 1, \dots, n$ are mean zero independent random variables. Suppose*

(a) $\sum_{i=1}^n \mathbb{E}\, X_{ni}^2 \to \sigma^2 > 0$ *and*

(b) for each ε, $\sum_{i=1}^n \mathbb{E}\,[|X_{ni}^2; |X_{ni}| > \varepsilon] \to 0$.

Let $S_n = \sum_{i=1}^n X_{ni}$. Then S_n converges weakly to a normal r.v. with mean zero and variance σ^2.

Note nothing is said about independence of the X_{ni} for different n.

Let us look at Theorem 20.1 in light of this theorem. Suppose the Y_i are i.i.d. and let $X_{ni} = Y_i/\sqrt{n}$. Then

$$\sum_{i=1}^n \mathbb{E}\,(Y_i/\sqrt{n})^2 = \mathbb{E}\, Y_1^2$$

and

$$\sum_{i=1}^n \mathbb{E}\,[|X_{ni}|^2; |X_{ni}| > \varepsilon] = n\mathbb{E}\,[|Y_1|^2/n; |Y_1| > \sqrt{n}\varepsilon] = \mathbb{E}\,[|Y_1|^2; |Y_1| > \sqrt{n}\varepsilon],$$

which tends to 0 by the dominated convergence theorem.

If the Y_i are independent with mean 0, and

$$\frac{\sum_{i=1}^n \mathbb{E}\,|Y_i|^3}{(\operatorname{Var} S_n)^{3/2}} \to 0,$$

then $S_n/(\operatorname{Var} S_n)^{1/2}$ converges weakly to a standard normal. This is known as Lyapounov's theorem; we leave the derivation of this from the Lindeberg-Feller theorem as an exercise for the reader.

Proof. Let φ_{ni} be the characteristic function of X_{ni} and let σ_{ni}^2 be the variance of X_{ni}. We need to show

$$\prod_{i=1}^n \varphi_{ni}(t) \to e^{-t^2\sigma^2/2}. \tag{20.2}$$

Using Taylor series, $|e^{ib} - 1 - ib + b^2/2| \le c|b|^3$ for a constant c. Also,

$$|e^{ib} - 1 - ib + b^2/2| \le |e^{ib} - 1 - ib| + |b^2|/2 \le c|b|^2.$$

If we apply this to a random variable tY and take expectations,

$$|\varphi_Y(t) - (1 + t\mathbb{E}\, Y + t^2\mathbb{E}\, Y^2/2)| \le c(t^2\mathbb{E}\, Y^2 \wedge t^3\mathbb{E}\, Y^3).$$

Applying this to $Y = X_{ni}$,

$$|\varphi_{ni}(t) - (1 - t^2\sigma_{ni}^2/2)| \le c\mathbb{E}\,[t^3|X_{ni}|^3 \wedge t^2|X_{ni}|^2].$$

The right hand side is less than or equal to

$$c\mathbb{E}\left[t^3|X_{ni}|^3;|X_{ni}|\le\varepsilon\right]+c\mathbb{E}\left[t^2|X_{ni}|^2;|X_{ni}|>\varepsilon\right]$$
$$\le c\varepsilon t^3\mathbb{E}\left[|X_{ni}|^2\right]+ct^2\mathbb{E}\left[|X_{ni}|^2;|X_{ni}|\ge\varepsilon\right].$$

Summing over i we obtain

$$\sum_{i=1}^n|\varphi_{ni}(t)-(1-t^2\sigma_{ni}^2/2)|\le c\varepsilon t^3\sum\mathbb{E}\left[|X_{ni}|^2\right]+ct^2\sum\mathbb{E}\left[|X_{ni}|^2;|X_{ni}|\ge\varepsilon\right].$$

Note $|\varphi_{ni}(t)|\le 1$ and $|1-t^2\sigma_{ni}^2/2|\le 1$ because $\sigma_{ni}^2\le\varepsilon^2+\mathbb{E}\left[|X_{ni}^2|;|X_{ni}|>\varepsilon\right]<1/t^2$ if we take ε small enough and n large enough. So

$$\Big|\prod_{i=1}^n\varphi_{ni}(t)-\prod_{i=1}^n(1-t^2\sigma_{ni}^2/2)\Big|\le c\varepsilon t^3\sum\mathbb{E}\left[|X_{ni}|^2\right]+ct^2\sum\mathbb{E}\left[|X_{ni}|^2;|X_{ni}|\ge\varepsilon\right].$$

Since $\sup_i\sigma_{ni}^2\to 0$, then $\log(1-t^2\sigma_{ni}^2/2)$ is asymptotically equal to $-t^2\sigma_{ni}^2/2$, and so

$$\prod(1-t^2\sigma_{ni}^2/2)=\exp\Big(\sum\log(1-t^2\sigma_{ni}^2/2)\Big)$$

is asymptotically equal to

$$\exp\Big(-t^2\sum\sigma_{ni}^2/2\Big)=e^{-t^2\sigma^2/2}.$$

Since ε is arbitrary, the proof is complete. □

We now complete the proof of Theorem 8.2.

Proof of "only if" part of Theorem 8.2. Since $\sum X_n$ converges, then X_n must converge to zero a.s., and so $\mathbb{P}(|X_n|>A \text{ i.o.})=0$. By the Borel-Cantelli lemma, this says $\sum\mathbb{P}(|X_n|>A)<\infty$. We also conclude by Proposition 5.4 that $\sum Y_n$ converges.

Let $c_n=\sum_{i=1}^n\operatorname{Var}Y_i$ and suppose $c_n\to\infty$. Let $X_{nm}=(Y_m-\mathbb{E}\,Y_m)/\sqrt{c_n}$. Then $\sum_{m=1}^n\operatorname{Var}X_{nm}=(1/c_n)\sum_{m=1}^n\operatorname{Var}Y_m=1$. If $\varepsilon>0$, then for n large, we have $2A/\sqrt{c_n}<\varepsilon$. Since $|Y_m|\le A$ and hence $|\mathbb{E}\,Y_m|\le A$, then $|X_{nm}|\le 2A/\sqrt{c_n}<\varepsilon$. It follows that $\sum_{m=1}^n\mathbb{E}\left(|X_{nm}|^2;|X_{nm}|>\varepsilon\right)=0$ for large n. By Theorem 20.4, $\sum_{m=1}^n Y_m/\sqrt{c_n}$ converges weakly to a standard normal. However, $\sum_{m=1}^n Y_m$ converges, and $c_n\to\infty$, so the ratio must converge to 0, a contradiction. We conclude c_n does not converge to infinity.

Let $V_i=Y_i-\mathbb{E}\,Y_i$. Since $|V_i|<2A$, $\mathbb{E}\,V_i=0$, and $\operatorname{Var}V_i=\operatorname{Var}Y_i$, which is summable, by the "if" part of the three series criterion, $\sum V_i$ converges. Since $\sum Y_i$ converges, taking the difference shows $\sum\mathbb{E}\,Y_i$ converges. □

21. Framework for Markov chains.

Suppose $\mathcal{S}$ is a set with some topological structure that we will use as our state space. Think of $\mathcal{S}$ as being $\mathbb{R}^d$ or the positive integers, for example. A sequence of random variables $X_0,X_1,\ldots$, is a Markov chain if

$$\mathbb{P}(X_{n+1}\in A\mid X_0,\ldots,X_n)=\mathbb{P}(X_{n+1}\in A\mid X_n)\tag{21.1}$$

for all n and all measurable sets A. The definition of Markov chain has this information: to predict the probability that X_{n+1} is in any set, we only need to know where we currently are; how we got there gives no new additional intuition.

Let's make some additional comments. First of all, we previously considered random variables as mappings from Ω to $\mathbb{R}$. Now we want to extend our definition by allowing a random variable be a map X from Ω to $\mathcal{S}$, where $(X \in A)$ is $\mathcal{F}$ measurable for all open sets A. This agrees with the definition of r.v. in the case $\mathcal{S} = \mathbb{R}$.

Although there is quite a theory developed for Markov chains with arbitrary state spaces, we will confine our attention to the case where either $\mathcal{S}$ is finite, in which case we will usually suppose $\mathcal{S} = \{1, 2, \ldots, n\}$, or countable and discrete, in which case we will usually suppose $\mathcal{S}$ is the set of positive integers.

We are going to further restrict our attention to Markov chains where

$$\mathbb{P}(X_{n+1} \in A \mid X_n = x) = \mathbb{P}(X_1 \in A \mid X_0 = x),$$

that is, where the probabilities do not depend on n. Such Markov chains are said to have stationary transition probabilities.

Define the initial distribution of a Markov chain with stationary transition probabilities by $\mu(i) = \mathbb{P}(X_0 = i)$. Define the transition probabilities by $p(i, j) = \mathbb{P}(X_{n+1} = j \mid X_n = i)$. Since the transition probabilities are stationary, $p(i, j)$ does not depend on n.

In this case we can use the definition of conditional probability given in undergraduate classes. If $\mathbb{P}(X_n = i) = 0$ for all n, that means we never visit i and we could drop the point i from the state space.

Proposition 21.1. *Let X be a Markov chain with initial distribution μ and transition probabilities $p(i, j)$. then*

$$\mathbb{P}(X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0) = \mu(i_0)p(i_0, i_1) \cdots p(i_{n-1}, i_n). \tag{21.2}$$

Proof. We use induction on n. It is clearly true for $n = 0$ by the definition of $\mu(i)$. Suppose it holds for n; we need to show it holds for $n + 1$. For simplicity, we will do the case $n = 2$. Then

$$\begin{aligned}
&\mathbb{P}(X_3 = i_3, X_2 = i_2, X_1 = i_1, X_0 = i_0) \\
&\qquad = \mathbb{E}\left[\mathbb{P}(X_3 = i_3 \mid X_0 = i_0, X_1 = i_1, X_2 = i_2); X_2 = i_2, X_1 = i_i, X_0 = i_0\right] \\
&\qquad = \mathbb{E}\left[\mathbb{P}(X_3 = i_3 \mid X_2 = i_2); X_2 = i_2, X_1 = i_i, X_0 = i_0\right] \\
&\qquad = p(i_2, i_3)\mathbb{P}(X_2 = i_2, X_1 = i_1, X_0 = i_0).
\end{aligned}$$

Now by the induction hypothesis,

$$\mathbb{P}(X_2 = i_2, X_1 = i_1, X_0 = i_0) = p(i_1, i_2)p(i_0, i_1)\mu(i_0).$$

Substituting establishes the claim for $n = 3$. □

The above proposition says that the law of the Markov chain is determined by the $\mu(i)$ and $p(i, j)$. The formula (21.2) also gives a prescription for constructing a Markov chain given the $\mu(i)$ and $p(i, j)$.

Proposition 21.2. *Suppose $\mu(i)$ is a sequence of nonnegative numbers with $\sum_i \mu(i) = 1$ and for each i the sequence $p(i, j)$ is nonnegative and sums to 1. Then there exists a Markov chain with $\mu(i)$ as its initial distribution and $p(i, j)$ as the transition probabilities.*

Proof. Define $\Omega = \mathcal{S}^\infty$. Let $\mathcal{F}$ be the σ-fields generated by the collection of sets $\{(i_0, i_1, \ldots, i_n) : n > 0, i_j \in \mathcal{S}\}$. An element ω of Ω is a sequence $(i_0, i_1, \ldots)$. Define $X_j(\omega) = i_j$ if $\omega = (i_0, i_1, \ldots)$. Define $\mathbb{P}(X_0 = i_0, \ldots, X_n = i_n)$ by (21.2). Using the Kolmogorov extension theorem, one can show that $\mathbb{P}$ can be extended to a probability on Ω.

The above framework is rather abstract, but it is clear that under $\mathbb{P}$ the sequence X_n has initial distribution $\mu(i)$; what we need to show is that X_n is a Markov chain and that

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid X_0 = i_0, \ldots X_n = i_n) = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n) = p(i_n, i_{n+1}). \tag{21.3}$$

By the definition of conditional probability, the left hand side of (21.3) is

$$\begin{aligned}\mathbb{P}(X_{n+1} = i_{n+1} \mid X_0 = i_0, \ldots, X_n = i_n) &= \frac{\mathbb{P}(X_{n+1} = i_{n+1}, X_n = i_n, \ldots X_0 = i_0)}{\mathbb{P}(X_n = i_n, \ldots, X_0 = i_0)} \\ &= \frac{\mu(i_0) \cdots p(i_{n-1}, i_n) p(i_n, i_{n+1})}{\mu(i_0) \cdots p(i_{n-1}, i_n)} \\ &= p(i_n, i_{n+1})\end{aligned}$$

as desired.

To complete the proof we need to show

$$\frac{\mathbb{P}(X_{n+1} = i_{n+1}, X_n = i_n)}{\mathbb{P}(X_n = i_n)} = p(i_n, i_{n+1}),$$

or

$$\mathbb{P}(X_{n+1} = i_{n+1}, X_n = i_n) = p(i_n, i_{n+1}) \mathbb{P}(X_n = i_n). \tag{21.4}$$

Now

$$\begin{aligned}\mathbb{P}(X_n = i_n) &= \sum_{i_0, \cdots, i_{n-1}} \mathbb{P}(X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) \\ &= \sum_{i_0, \cdots, i_{n-1}} \mu(i_0) \cdots p(i_{n-1}, i_n)\end{aligned}$$

and similarly

$$\begin{aligned}\mathbb{P}(X_{n+1} &= i_{n+1}, X_n = i_n) \\ &= \sum_{i_0, \cdots, i_{n-1}} \mathbb{P}(X_{n+1} = i_{n+1}, X_n = i_n, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) \\ &= p(i_n, i_{n+1}) \sum_{i_0, \cdots, i_{n-1}} \mu(i_0) \cdots p(i_{n-1}, i_n).\end{aligned}$$

Equation (21.4) now follows. □

Note in this construction that the X_n sequence is fixed and does not depend on μ or p. Let $p(i, j)$ be fixed. The probability we constructed above is often denoted $\mathbb{P}^\mu$. If μ is point mass at a point i or x, it is denoted $\mathbb{P}^i$ or $\mathbb{P}^x$. So we have one probability space, one sequence X_n, but a whole family of probabilities $\mathbb{P}^\mu$.

Later on we will see that this framework allows one to express the Markov property and strong Markov property in a convenient way. As part of the preparation for doing this, we define the shift operators $\theta_k : \Omega \to \Omega$ by

$$\theta_k(i_0, i_1, \ldots) = (i_k, i_{k+1}, \ldots).$$

Then $X_j \circ \theta_k = X_{j+k}$. To see this, if $\omega = (i_0, i_1, \ldots)$, then

$$X_j \circ \theta_k(\omega) = X_j(i_k, i_{k+1}, \ldots) = i_{j+k} = X_{j+k}(\omega).$$

22. Examples.

Random walk on the integers

We let Y_i be an i.i.d. sequence of r.v.'s, with $p = \mathbb{P}(Y_i = 1)$ and $1 - p = \mathbb{P}(Y_i = -1)$. Let $X_n = X_0 + \sum_{i=1}^n Y_i$. Then the X_n can be viewed as a Markov chain with $p(i, i+1) = p$, $p(i, i-1) = 1 - p$, and $p(i, j) = 0$ if $|j - i| \neq 1$. More general random walks on the integers also fit into this framework. To check that the random walk is Markov,

$$\begin{aligned}\mathbb{P}(X_{n+1} = i_{n+1} \mid X_0 = i_0, \ldots, X_n = i_n) \\ &= \mathbb{P}(X_{n+1} - X_n = i_{n+1} - i_n \mid X_0 = i_0, \ldots, X_n = i_n) \\ &= \mathbb{P}(X_{n+1} - X_n = i_{n+1} - i_n),\end{aligned}$$

using the independence, while

$$\begin{aligned}\mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n) &= \mathbb{P}(X_{n+1} - X_n = i_{n+1} - i_n \mid X_n = i_n) \\ &= \mathbb{P}(X_{n+1} - X_n = i_{n+1} - i_n).\end{aligned}$$

Random walks on graphs

Suppose we have n points, and from each point there is some probability of going to another point. For example, suppose there are 5 points and we have $p(1,2) = \frac{1}{2}$, $p(1,3) = \frac{1}{2}$, $p(2,1) = \frac{1}{4}$, $p(2,3) = \frac{1}{2}$, $p(2,5) = \frac{1}{4}$, $p(3,1) = \frac{1}{4}$, $p(3,2) = \frac{1}{8}$, $p(3,3 = \frac{1}{2}$, $p(4,1) = 1$, $p(5,1) = \frac{1}{2}$, $p(5,5) = \frac{1}{2}$. The $p(i,j)$ are often arranged into a matrix:

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{2} & 0 & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \end{pmatrix}.$$

Note the rows must sum to 1 since

$$\sum_{j=1}^5 p(i,j) = \sum_{j=1}^5 \mathbb{P}(X_1 = j \mid X_0 = i) = \mathbb{P}(X_1 \in \mathcal{S} \mid X_0 = i) = 1.$$

Renewal processes

Let Y_i be i.i.d. with $\mathbb{P}(Y_i = k) = a_k$ and the a_k are nonnegative and sum to 1. Let $T_0 = i_0$ and $T_n = T_0 + \sum_{i=1}^n$. We think of the Y_i as the lifetime of the nth light bulb and T_n the time when the nth light bulb burns out. (We replace a light bulb as soon as it burns out.) Let

$$X_n = \min\{m - n : T_i = m \text{ for some } i\}.$$

So X_n is the amount of time after time n until the current light bulb burns out.

If $X_n = j$ and $j > 0$, then $T_i = n + j$ for some i but T_i does not equal $n, n+1, \ldots, n+j-1$ for any i. So $T_i = (n+1) + (j-1)$ for some i and T_i does not equal $(n+1), (n+1)+1, \ldots, (n+1)+(j-2)$ for any i. Therefore $X_{n+1} = j - 1$. So $(p(i, i-1) = 1$ if $i \geq 1$.

If $X_n = 0$, then a light bulb burned out at time n and X_{n+1} is 0 if the next light bulb burned out immediately and $j - 1$ if the light bulb has lifetime j. The probability of this is a_j. So $p(0, j) = a_{j+1}$. All the other $p(i,j)$'s are 0.

Branching processes

Consider k particles. At the next time interval, some of them die, and some of them split into several particles. The probability that a given particle will split into j particles is given by a_j, $j = 0, 1, \ldots$, where the a_j are nonnegative and sum to 1. The behavior of each particle is independent of the behavior of all the other particles. If X_n is the number of particles at time n, then X_n is a Markov chain. Let Y_i be i.i.d. random variables with $\mathbb{P}(Y_i = j) = a_j$. The $p(i, j)$ for X_n are somewhat complicated, and can be defined by $p(i, j) = \mathbb{P}(\sum_{m=1}^{i} Y_m = j)$.

Queues

We will discuss briefly the $M/G/1$ queue. The M refers to the fact that the customers arrive according to a Poisson process. So the probability that the number of customers arriving in a time interval of length t is k is given by $e^{-\lambda t}(\lambda t)^k/k!$ The G refers to the fact that the length of time it takes to serve a customer is given by a distribution that is not necessarily exponential. The 1 refers to the fact that there is 1 server.

Suppose the length of time to serve one customer has distribution function F with density f. The probability that k customers arrive during the time it takes to serve one customer is

$$a_k = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^k}{k!} f(t)\, dt.$$

Let the Y_i be i.i.d. with $\mathbb{P}(Y_i = k - 1) = a_k$. So Y_i is the number of customers arriving during the time it takes to serve one customer. Let $X_{n+1} = (X_n + Y_{n+1})^+$ be the number of customers waiting. Then X_n is a Markov chain with $p(0, 0) = a_0 + a_1$ and $p(i, j - 1 + k) = a_k$ if $j \geq 1, k > 1$.

Ehrenfest urns

Suppose we have two urns with a total of r balls, k in one and $r - k$ in the other. Pick one of the r balls at random and move it to the other urn. Let X_n be the number of balls in the first urn. X_n is a Markov chain with $p(k, k + 1) = (r - k)/r$, $p(k, k - 1) = k/r$, and $p(i, j) = 0$ otherwise.

One model for this is to consider two containers of air with a thin tube connecting them. Suppose a few molecules of a foreign substance are introduced. Then the number of molecules in the first container is like an Ehrenfest urn. We shall see that all states in this model are recurrent, so infinitely often all the molecules of the foreign substance will be in the first urn. Yet there is a tendency towards equilibrium, so on average there will be about the same number of molecules in each container for all large times.

Birth and death processes

Suppose there are i particles, and the probability of a birth is a_i, the probability of a death is b_i, where $a_i, b_i \geq 0$, $a_i + b_i \leq 1$. Setting X_n equal to the number of particles, then X_n is a Markov chain with $p(i, i + 1) = a_i$, $p(i, i - 1) = b_i$, and $p(i, i) = 1 - a_i - b_i$.

23. Markov properties.

Let us look at

$$\mathbb{P}^x(X_{n+1} = j \mid X_n = i) = \mathbb{P}^x(X_1 = j \mid X_0 = i).$$

Let $g(y) = \mathbb{P}^y(X_1 = j)$. We have

$$\mathbb{P}^x(X_1 = j, X_0 = k) = \begin{cases} \mathbb{P}^k(X_1 = j) & \text{if } x = k, \\ 0 & \text{if } x \neq k, \end{cases}$$

while

$$\begin{aligned} \mathbb{E}^x[g(X_0); X_0 = k] = \mathbb{E}^x[g(k); X_0 = k] &= \mathbb{P}^k(X_1 = j)\mathbb{P}^x(X_0 = k) \\ &= \begin{cases} \mathbb{P}^k(X_1 = j) & \text{if } x = k, \\ 0 & \text{if } x \neq k. \end{cases} \end{aligned}$$

It follows that $\mathbb{P}^x(X_1 = j \mid X_0) = g(X_0)$, and so

$$\mathbb{P}^x(X_1 = j \mid X_0 = i) = \mathbb{P}^i(X_1 = j).$$

The following are equivalent ways of writing this:

$$\begin{aligned}
\mathbb{E}^x[1_{\{j\}}(X_{n+1}) \mid X_n = i] &= \mathbb{E}^i[1_{\{j\}}(X_1)];\\
\mathbb{E}^x[1_{\{j\}}(X_1) \circ \theta_n \mid X_n = i] &= \mathbb{E}^i[1_{\{j\}}(X_1)];\\
\mathbb{E}^x[Y \circ \theta_n \mid X_n = i] &= \mathbb{E}^i[Y], \qquad Y = 1_{\{j\}}(X_1);\\
\mathbb{E}^x[Y \circ \theta_n \mid X_n] &= \mathbb{E}^{X_n}[Y];\\
\mathbb{E}^x[Y \circ \theta_n \mid \mathcal{F}_n] &= \mathbb{E}^{X_n}[Y].
\end{aligned}$$

The last one is a complicated way of writing the first. We generalize this.

Theorem 23.1. (Markov property) *If Y is bounded and measurable, then*

$$\mathbb{E}^x[Y \circ \theta_n \mid \mathcal{F}_n] = \mathbb{E}^{X_n}[Y], \qquad \text{a.s.}$$

for each n and x.

Proof. If we can prove this for $Y = f_1(X_1) \cdots f_m(X_m)$, then we will have it for all Y by linearity and taking limits. We use induction on m. The case $m = 1$ is just the string of equivalences above.

Suppose the result holds for m and we want to show it holds for $m + 1$. We have

$$\begin{aligned}
&\mathbb{E}^x[f_1(X_{n+1}) \cdots f_{m+1}(X_{n+m+1}) \mid \mathcal{F}_n]\\
&\qquad = \mathbb{E}^x[\mathbb{E}^x[f_{m+1}(X_{n+m+1}) \mid \mathcal{F}_{n+m}] f_1(X_{n+1}) \cdots f_m(X_{n+m}) \mid \mathcal{F}_n]\\
&\qquad \mathbb{E}^x[\mathbb{E}^{X_{n+m}}[f_{m+1}(X_1)] f_1(X_{n+1}) \cdots f_m(X_{n+m}) \mid \mathcal{F}_n]\\
&\qquad \mathbb{E}^x[f_1(X_{n+1}) \cdots f_{m-1}(X_{n+m-1}) h(X_{n+m}) \mid \mathcal{F}_n].
\end{aligned}$$

Here we used the result for $m = 1$ and we defined $h(y) = f_{n+m}(y)g(y)$, where $g(y) = \mathbb{E}^y[f_{m+1}(X_1)]$. Using the induction hypothesis, this is equal to

$$\begin{aligned}
\mathbb{E}^{X_n}[f_1(X_1) \cdots f_{m-1}(X_{m-1}) g(X_m)] &= \mathbb{E}^{X_n}[f_1(X_1) \cdots f_m(X_m) \mathbb{E}^{X_m} f_{m+1}(X_1)]\\
&= \mathbb{E}^{X_n}[f_1(X_1) \cdots f_m(X_m) \mathbb{E}[f_{m+1}(X_{m+1}) \mid \mathcal{F}_m]]\\
&= \mathbb{E}^{X_n}[f_1(X_1) \cdots f_{m+1}(X_{m+1})],
\end{aligned}$$

which is what we needed. □

Define $\theta_N(\omega) = (\theta_{N(\omega)})(\omega)$. The strong Markov property is the same as the Markov property, but where the fixed time n is replaced by a stopping time N.

Theorem 23.2. *If Y is bounded and measurable and N is a finite stopping time, then*

$$\mathbb{E}^x[Y \circ \theta_N \mid \mathcal{F}_N] = \mathbb{E}^{X_N}[Y].$$

Proof. We will show

$$\mathbb{P}^x(X_{N+1} = j \mid \mathcal{F}_N) = \mathbb{P}^{X_N}(X_1 = j).$$

Once we have this, we can proceed as in the proof of the Theorem 23.1 to obtain our result. To show the above equality, we need to show that if $B \in \mathcal{F}_N$, then

$$\mathbb{P}^x(X_{N+1} = j, B) = \mathbb{E}^x[\mathbb{P}^{X_N}(X_1 = j); B]. \tag{23.1}$$

Recall that since $B \in \mathcal{F}_N$, then $B \cap (N = k) \in \mathcal{F}_k$. We have

$$\begin{aligned}
\mathbb{P}^x(X_{N+1} = j, B, N = k) &= \mathbb{P}^x(X_{k+1} = j, B, N = k) \\
&= \mathbb{E}^x[\mathbb{P}^x(X_{k+1} = j \mid \mathcal{F}_k); B, N = k] \\
&= \mathbb{E}^x[\mathbb{P}^{X_k}(X_1 = j); B, N = k] \\
&= \mathbb{E}^x[\mathbb{P}^{X_N}(X_1 = j); B, N = k].
\end{aligned}$$

Now sum over k; since N is finite, we obtain our desired result. □

Another way of expressing the Markov property is through the Chapman-Kolmogorov equations. Let $p^n(i,j) = \mathbb{P}(X_n = j \mid X_0 = i)$.

Proposition 23.3. *For all i, j, m, n we have*

$$p^{n+m}(i,j) = \sum_{k \in \mathcal{S}} p^n(i,k) p^m(k,j).$$

Proof. We write

$$\begin{aligned}
\mathbb{P}(X_{n+m} = j, X_0 = i) &= \sum_k \mathbb{P}(X_{n+m} = j, X_n = k, X_0 = i) \\
&= \sum_k \mathbb{P}(X_{n+m} = j \mid X_n = k, X_0 = i) \mathbb{P}(X_n = k \mid X_0 = i) \mathbb{P}(X_0 = i) \\
&= \sum_k \mathbb{P}(X_{n+m} = j \mid X_n = k) p^n(i,k) \mathbb{P}(X_0 = i) \\
&= \sum_k p^m(k,j) p^n(i,k) \mathbb{P}(X_0 = i).
\end{aligned}$$

If we divide both sides by $\mathbb{P}(X_0 = i)$, we have our result. □

Note the resemblance to matrix multiplication. It is clear if P is the matrix made up of the $p(i,j)$, then P^n will be the matrix whose (i,j) entry is $p^n(i,j)$.

24. Recurrence and transience.

Let

$$T_y = \min\{i > 0 : X_i = y\}.$$

This is the first time that X_i hits the point y. Even if $X_0 = y$ we would have $T_y > 0$. We let T_y^k be the k-th time that the Markov chain hits y and we set

$$r(x,y) = \mathbb{P}^x(T_y < \infty),$$

the probability starting at x that the Markov chain ever hits y.

Proposition 24.1. $\mathbb{P}^x(T_y^k < \infty) = r(x,y)r(y,y)^{k-1}$.

Proof. The case $k = 1$ is just the definition, so suppose $k > 1$. Using the strong Markov property,

$$\begin{aligned}
\mathbb{P}^x(T_y < \infty) &= \mathbb{P}^x(T_y \circ \theta_{T_y^{k-1}} < \infty, T_y^{k-1} < \infty) \\
&= \mathbb{E}^x[\mathbb{P}^x(T_y \circ \theta_{T_y^{k-1}} < \infty \mid \mathcal{F}_{T_y^{k-1}}); T_y^{k-1} < \infty] \\
&= \mathbb{E}^x[\mathbb{P}^{X(T_y^{k-1})}(T_y < \infty); T_y^{k-1}] \\
&= \mathbb{E}^x[\mathbb{P}^y(T_y < \infty); T_y^{k-1} < \infty] \\
&= r(y,y)\mathbb{P}^x(T_y^{k-1} < \infty).
\end{aligned}$$

We used here the fact that at time T_y^{k-1} the Markov chain must be at the point y. Repeating this argument $k-2$ times yields the result. □

We say that y is recurrent if $r(y,y) = 1$; otherwise we say y is transient. Let

$$N(y) = \sum_{n=1}^{\infty} 1_{(X_n = y)}.$$

Proposition 24.2. *y is recurrent if and only if $\mathbb{E}^y N(y) = \infty$.*

Proof. Note

$$\begin{aligned}
\mathbb{E}^y N(y) &= \sum_{k=1}^{\infty} \mathbb{P}^y(N(y) \geq k) = \sum_{k=1}^{\infty} \mathbb{P}^y(T_y^k < \infty) \\
&= \sum_{k=1}^{\infty} r(y,y)^k.
\end{aligned}$$

We used the fact that $N(y)$ is the number of visits to y and the number of visits being larger than k is the same as the time of the k-th visit being finite. Since $r(y,y) \leq 1$, the left hand side will be finite if and only if $r(y,y) < 1$. □

Observe that

$$\mathbb{E}^y N(y) = \sum_n \mathbb{P}^y(X_n = y) = \sum_n p^n(y,y).$$

If we consider simple symmetric random walk on the integers, then $p^n(0,0)$ is 0 if n is odd and equal to $\binom{n}{n/2} 2^{-n}$ if n is even. This is because in order to be at 0 after n steps, the walk must have had $n/2$ positive steps and $n/2$ negative steps; the probability of this is given by the binomial distribution. Using Stirling's approximation, we see that $p^n(0,0) \sim c/\sqrt{n}$ for n even, which diverges, and so simple random walk in one dimension is recurrent.

Similar arguments show that simple symmetric random walk is also recurrent in 2 dimensions but transient in 3 or more dimensions.

Proposition 24.3. *If x is recurrent and $r(x,y) > 0$, then y is recurrent and $r(y,x) = 1$.*

Proof. First we show $r(y,x) = 1$. Suppose not. Since $r(x,y) > 0$, there is a smallest n and $y_1, \ldots, y_{n-1}$ such that $p(x,y_1)p(y_1,y_2)\cdots p(y_{n-1},y) > 0$. Since this is the smallest n, none of the y_i can equal x. Then

$$\mathbb{P}^x(T_x = \infty) \geq p(x,y_1)\cdots p(y_{n-1},y)(1 - r(y,x)) > 0,$$

a contradiction to x being recurrent.

Next we show that y is recurrent. Since $r(y,x) > 0$, there exists L such that $p^L(y,x) > 0$. Then

$$p^{L+n+K}(y,y) \geq p^L(y,x)p^n(x,x)p^K(x,y).$$

Summing over n,

$$\sum_n p^{L+n+K}(y,y) \geq p^L(y,x)p^K(x,y)\sum_n p^n(x,x) = \infty.$$

□

We say a subset C of $\mathcal{S}$ is closed if $x \in C$ and $r(x,y) > 0$ implies $y \in C$. A subset D is irreducible if $x, y \in D$ implies $r(x,y) > 0$.

Proposition 24.4. *Let C be finite and closed. Then C contains a recurrent state.*

From the preceding proposition, if C is irreducible, then all states will be recurrent.

Proof. If not, for all y we have $r(y,y) < 1$ and

$$\mathbb{E}^x N(y) = \sum_{k=1}^{\infty} r(x,y)r(y,y)^{k-1} = \frac{r(x,y)}{1-r(y,y)} < \infty.$$

Since C is finite, then $\sum_y \mathbb{E}^x N(y) < \infty$. But that is a contradiction since

$$\sum_y \mathbb{E}^x N(y) = \sum_y \sum_n p^n(x,y) = \sum_n \sum_y p^n(x,y) = \sum_n \mathbb{P}^x(X_n \in C) = \sum_n 1 = \infty.$$

□

Theorem 24.5. *Let $R = \{x : r(x,x) = 1\}$, the set of recurrent states. Then $R = \cup_{i=1}^{\infty} R_i$, where each R_i is closed and irreducible.*

Proof. Say $x \sim y$ if $r(x,y) > 0$. Since every state is recurrent, $x \sim x$ and if $x \sim y$, then $y \sim x$. If $x \sim y$ and $y \sim z$, then $p^n(x,y) > 0$ and $p^m(y,z) > 0$ for some n and m. Then $p^{n+m}(x,z) > 0$ or $x \sim z$. Therefore we have an equivalence relation and we let the R_i be the equivalence classes. □

Looking at our examples, it is easy to see that in the Ehrenfest urn model all states are recurrent. For the branching process model, suppose $p(x,0) > 0$ for all x. Then 0 is recurrent and all the other states are transient. In the renewal chain, there are two cases. If $\{k : a_k > 0\}$ is unbounded, all states are recurrent. If $K = \max\{k : a_k > 0\}$, then $\{0, 1, \ldots, K-1\}$ are recurrent states and the rest are transient.

For the queueing model, let $\mu = \sum k a_k$, the expected number of people arriving during one customer's service time. We may view this as a branching process by letting all the customers arriving during one person's service time be considered the progeny of that customer. It turns out that if $\mu \leq 1$, 0 is recurrent and all other states are also. If $\mu > 1$ all states are transient.

25. Stationary measures.

A probability μ is a stationary distribution if

$$\sum_x \mu(x)p(x,y) = \mu(y). \tag{25.1}$$

In matrix notation this is $\mu P = \mu$, or μ is the left eigenvector corresponding to the eigenvalue 1. In the case of a stationary distribution, $\mathbb{P}^\mu(X_1 = y) = \mu(y)$, which implies that $X_1, X_2, \ldots$ all have the same distribution. We can use (25.1) when μ is a measure rather than a probability, in which case it is called a stationary measure.

If we have a random walk on the integers, $\mu(x) = 1$ for all x serves as a stationary measure. In the case of an asymmetric random walk: $p(i, i+1) = p$, $p(i, i-1) = q = 1 - p$ and $p \neq q$, setting $\mu(x) = (p/q)^x$ also works.

In the Ehrenfest urn model, $\mu(x) = 2^{-r} \begin{pmatrix} r \\ x \end{pmatrix}$ works. One way to see this is that μ is the distribution one gets if one flips r coins and puts a coin in the first urn when the coin is heads. A transition corresponds to picking a coin at random and turning it over.

Proposition 25.1. *Let x be recurrent and let $T = T_x$. Set*

$$\mu(y) = \mathbb{E}^x \sum_{n=0}^{T-1} 1_{(X_n = y)}.$$

Then μ is a stationary measure.

The idea of the proof is that $\mu(y)$ is the expected number of visits to y by the sequence $X_0, \ldots, X_{T-1}$ while μP is the expected number of visits to y by $X_1, \ldots, X_T$. These should be the same because $X_T = X_0 = x$.

Proof. First, let $\overline{p}_n(x, y) = \mathbb{P}^x(X_n = y, T > n)$. So

$$\mu(y) = \sum_{n=0}^{\infty} \mathbb{P}^x(X_n = y, T > n) = \sum_{n=0}^{\infty} \overline{p}_n(x, y)$$

and

$$\sum_y \mu(y) p(y, z) = \sum_y \sum_{n=0}^{\infty} \overline{p}_n(x, y) p(y, z).$$

Second, we consider the case $z \neq x$. Then

$$\begin{aligned} &\sum_y \overline{p}_n(x, y) p(y, z) \\ &\quad = \sum_y \mathbb{P}^x(\text{hit } y \text{ in } n \text{ steps without first hitting } x \text{ and then go to } z \text{ in one step}) \\ &\quad = \overline{p}_{n+1}(x, z). \end{aligned}$$

So

$$\begin{aligned} \sum_n \mu(y) p(y, z) &= \sum_n \sum_y \overline{p}_n(x, y) p(y, z) \\ &= \sum_{n=0}^{\infty} \overline{p}_{n+1}(x, z) = \sum_{n=0}^{\infty} \overline{p}_n(x, z) \\ &= \mu(z) \end{aligned}$$

since $\overline{p}_0(x, z) = 0$.

Third, we consider the case $x = z$. Then

$$\begin{aligned} &\sum_y \overline{p}_n(x, y) p(y, z) \\ &\quad = \sum_y \mathbb{P}^x(\text{hit } y \text{ in } n \text{ steps without first hitting } x \text{ and then go to } z \text{ in one step}) \\ &\quad = \mathbb{P}^x(T = n + 1). \end{aligned}$$

Recall $\mathbb{P}^x(T=0)=0$, and since x is recurrent, $T<\infty$. So

$$\begin{aligned}\sum_y \mu(y)p(y,z) &= \sum_n \sum_y \overline{p}_n(x,y)p(y,z)\\ &= \sum_{n=0}^{\infty} \mathbb{P}^x(T=n+1) = \sum_{n=0}^{\infty} \mathbb{P}^x(T=n) = 1.\end{aligned}$$

On the other hand,

$$\sum_{n=0}^{T-1} 1_{(X_n=x)} = 1_{(X_0=x)} = 1,$$

hence $\mu(x)=1$. Therefore, whether $z\neq x$ or $z=x$, we have $\mu P(z)=\mu(z)$.

Finally, we show $\mu(y)<\infty$. If $r(x,y)=0$, then $\mu(y)=0$. If $r(x,y)>0$, choose n so that $p^n(x,y)>0$, and then

$$1=\mu(x)=\sum_y \mu(y)p^n(x,y),$$

which implies $\mu(y)<\infty$. □

We next turn to uniqueness of the stationary distribution.

Proposition 25.2. *If the Markov chain is irreducible and all states are recurrent, then the stationary measure is unique up to a constant multiple.*

Proof. Fix $a\in\mathcal{S}$. Let μ be the stationary measure constructed above and let ν be any other stationary measure.

Since $\nu=\nu P$, then

$$\begin{aligned}\nu(z) &= \nu(a)p(a,z)+\sum_{y\neq a}\nu(y)p(y,z)\\ &= \nu(a)p(a,z)+\sum_{y\neq a}\nu(a)p(a,y)p(y,z)+\sum_{x\neq a}\sum_{y\neq a}\nu(x)p(x,y)p(y,z)\\ &= \nu(a)\mathbb{P}^a(X_1=z)+\nu(a)\mathbb{P}^a(X_1\neq a, X_2=z)+\mathbb{P}^\nu(X_0\neq a, X_1\neq a, X_2=z).\end{aligned}$$

Continuing,

$$\begin{aligned}\nu(z) &= \nu(a)\sum_{m=1}^{n}\mathbb{P}^a(X_1\neq a, X_2\neq a,\ldots,X_{m-1}\neq a, X_m=z)\\ &\quad+\mathbb{P}^\nu(X_0\neq a, X_1\neq a,\ldots,X_{n-1}\neq a, X_n=z)\\ &\geq \nu(a)\sum_{m=1}^{n}\mathbb{P}^a(X_1\neq a, X_2\neq a,\ldots,X_{m-1}\neq a, X_m=z).\end{aligned}$$

Letting $n\to\infty$, we obtain

$$\nu(z)\geq\nu(a)\mu(z).$$

We have

$$\begin{aligned}\nu(a) &= \sum_x \nu(x)p^n(x,a)\geq\nu(a)\sum_x\mu(x)p^n(x,a)\\ &= \nu(a)\mu(a)=\nu(a),\end{aligned}$$

since $\mu(a)=1$ (see proof of Proposition 25.1). This means that we have equality and so

$$\nu(x)=\nu(a)\mu(x)$$

whenever $p^n(x,a) > 0$. Since $r(x,a) > 0$, this happens for some n. Consequently

$$\frac{\nu(x)}{\nu(a)} = \mu(x).$$

□

Proposition 25.3. *If a stationary distribution exists, then $\mu(y) > 0$ implies y is recurrent.*

Proof. If $\mu(y) > 0$, then

$$\begin{aligned}\infty = \sum_{n=1}^{\infty} \mu(y) &= \sum_{n=1}^{\infty} \sum_x \mu(x) p^n(x,y) = \sum_x \mu(x) \sum_{n=1}^{\infty} p^n(x,y) \\ &= \sum_x \mu(x) \sum_{n=1}^{\infty} \mathbb{P}^x(X_n = y) = \sum_x \mu(x) \mathbb{E}^x N(y) \\ &= \sum_x \mu(x) r(x,y)[1 + r(y,y) + r(y,y)^2 + \cdots].\end{aligned}$$

Since $r(x,y) \leq 1$ and μ is a probability measure, this is less than

$$\sum_x \mu(x)(1 + r(y,y) + \cdots) \leq 1 + r(y,y) + \cdots.$$

Hence $r(y,y)$ must equal 1. □

Recall that T_x is the first time to hit x.

Proposition 25.4. *If the Markov chain is irreducible and has stationary distribution μ, then*

$$\mu(x) = \frac{1}{\mathbb{E}^x T_x}.$$

Proof. $\mu(x) > 0$ for some x. If $y \in \mathcal{S}$, then $r(x,y) > 0$ and so $p^n(x,y) > 0$ for some n. Hence

$$\mu(y) = \sum_x \mu(x) p^n(x,y) > 0.$$

Hence by Proposition 25.3, all states are recurrent. By the uniqueness of the stationary distribution, $\mu(x)$ is a constant multiple of

$$\mu_x(y) = \sum_{n=0}^{\infty} \mathbb{P}^x(X_n = y, T_x > n).$$

Note

$$\begin{aligned}\sum_y \mu_x(y) &= \sum_y \sum_{n=0}^{\infty} \mathbb{P}^x(X_n = y, T_x > n) = \sum_n \sum_y \mathbb{P}^x(X_n = y, T_x > n) \\ &= \sum_n \mathbb{P}^x(T_x > n) = \mathbb{E}^x T_x.\end{aligned}$$

Therefore, since $\sum_y \mu(y) = 1$,

$$\mu(x) = \frac{\mu_x(x)}{\sum_y \mu_x(y)} = \frac{1}{\mathbb{E}^x T_x}.$$

□

We make the following distinction for recurrent states. If $\mathbb{E}^x T_x < \infty$, then x is said to be positive recurrent. If x is recurrent but $\mathbb{E}^x T_x = \infty$, x is null recurrent.

Proposition 25.5. *Suppose a chain is irreducible.*

(a) If there exists a positive recurrent state, then there is a stationary distribution,.

(b) If there is a stationary distribution, all states are recurrent.

(c) If there exists a transient state, all states are transient.

(d) If there exists a null recurrent state, all states are null recurrent.

Proof. To show (a), if x is positive recurrent, then there exists a stationary measure with $\mu(x) = 1$. Then $\overline{\mu}(y) = \mu(y)/\mathbb{E}^x T_x$ will be a stationary distribution.

For (b), suppose $\mu(x) > 0$ for some x. We showed this implies $\mu(y) > 0$ for all y. Then $0 < \mu(y) = 1/\mathbb{E}^y T_y$, which implies $\mathbb{E}^y T_y < \infty$.

We showed that if x is recurrent and $r(x, y) > 0$, then y is recurrent. So (c) follows.

Suppose there exists a null recurrent state. If there exists a positive recurrent or transient state as well, then by (a) and (b) or by (c) all states are positive recurrent or transient, a contradiction, and (d) follows. □

26. Convergence.

Our goal is to show that under certain conditions $p^n(x, y) \to \pi(y)$, where π is the stationary distribution. (In the null recurrent case $p^n(x, y) \to 0$.)

Consider a random walk on the set $\{0, 1\}$, where with probability one on each step the chain moves to the other state. Then $p^n(x, y) = 0$ if $x \neq y$ and n is even. A less trivial case is the simple random walk on the integers. We need to eliminate this periodicity.

Suppose x is recurrent, let $I_x = \{n \geq 1 : p^n(x, x) > 0\}$, and let d_x be the g.c.d. (greatest common divisor) of I_x. d_x is called the period of x.

Proposition 26.1. *If $r(x, y) > 0$, then $d_y = d_x$.*

Proof. Since x is recurrent, $r(y, x) > 0$. Choose K and L such that $p^K(x, y), p^L(y, x) > 0$.

$$p^{K+L+n}(y, y) \geq p^L(y, x)p^n(x, x)p^K(x, y),$$

so taking $n = 0$, we have $p^{K+L}(y, y) > 0$, or d_y divides $K + L$. So d_y divides n if $p^n(x, x) > 0$, or d_y is a divisor of I_x. Hence d_y divides d_x. By symmetry d_x divides d_y. □

Proposition 26.2. *If $d_x = 1$, there exists m_0 such that $p^m(x, x) > 0$ whenever $m \geq m_0$.*

Proof. First of all, I_x is closed under addition: if $m, n \in I_x$,

$$p^{m+n}(x, x) \geq p^m(x, x)p^n(x, x) > 0.$$

Secondly, if there exists N such that $N, N + 1 \in I_x$, let $m_0 = N^2$. If $m \geq m_0$, then $m - N^2 = kN + r$ for some $r < N$ and

$$m = r + N^2 + kN = r(N + 1) + (N - r + k)N \in I_x.$$

Third, pick $n_0 \in I_x$ and $k > 0$ such that $n_0 + k \in I_x$. If $k = 1$, we are done. Since $d_x = 1$, there exists $n_1 \in I_x$ such that k does not divide n_1. We have $n_1 = mk + r$ for some $0 < r < k$. Note $(m+1)(n_0+k) \in I_x$ and $(m + 1)n_0 + n_1 \in I_x$. The difference between these two numbers is $(m + 1)k - n_1 = k - r < k$. So now we have two numbers in I_k differing by less than or equal to $k - 1$. Repeating at most k times, we get two numbers in I_x differing by at most 1, and we are done. □

We write d for d_x. A chain is aperiodic if $d = 1$.

If $d > 1$, we say $x \sim y$ if $p^{kd}(x, y) > 0$ for some $k > 0$ We divide $\mathcal{S}$ into equivalence classes $\mathcal{S}_1, \ldots \mathcal{S}_d$. Every d steps the chain started in $\mathcal{S}_i$ is back in $\mathcal{S}_i$. So we look at $p' = p^d$ on $\mathcal{S}_i$.

Theorem 26.3. *Suppose the chain is irreducible, aperiodic, and has a stationary distribution π. Then $p^n(x,y) \to \pi(y)$ as $n \to \infty$.*

Proof. The idea is to take two copies of the chain with different starting distributions, let them run independently until they couple, i.e., hit each other, and then have them move together. So define

$$q((x_1,y_1),(x_2,y_2)) = \begin{cases} p(x_1,x_2)p(y_1,y_2) & \text{if } x_1 \neq y_1, \\ p(x_1,x_2) & \text{if } x_1 = y_1, x_2 = y_2, \\ 0 & \text{otherwise.} \end{cases}$$

Let $Z_n = (X_n, Y_n)$ and $T = \min\{i : X_i = Y_i\}$. We have

$$\begin{aligned} \mathbb{P}(X_n = y) &= \mathbb{P}(X_n = y, T \le n) + \mathbb{P}(X_n = y, T > n) \\ &= \mathbb{P}(Y_n = y, T \le n) + \mathbb{P}(X_n = y, T > n), \end{aligned}$$

while

$$\mathbb{P}(Y_n = y) = \mathbb{P}(Y_n = y, T \le n) + \mathbb{P}(Y_n = y, T > n).$$

Subtracting,

$$\begin{aligned} \mathbb{P}(X_n = y) - \mathbb{P}(Y_n = y) &\le \mathbb{P}(X_n = y, T > n) - \mathbb{P}(Y_n = y, T > n) \\ &\le \mathbb{P}(X_n = y, T > n) \le \mathbb{P}(T > n). \end{aligned}$$

Using symmetry,

$$|\mathbb{P}(X_n = y) - \mathbb{P}(Y_n = y)| \le \mathbb{P}(T > n).$$

Suppose we let Y_0 have distribution π and $X_0 = x$. Then

$$|p^n(x,y) - \pi(y)| \le \mathbb{P}(T > n).$$

It remains to show $\mathbb{P}(T > n) \to 0$. To do this, consider another chain $Z'_n = (X_n, Y_n)$, where now we take X_n, Y_n independent. Define

$$r((x_1,y_1),(x_2,y_2)) = p(x_1,x_2)p(y_1,y_2).$$

The chain under the transition probabilities r is irreducible. To see this, there exist K and L such that $p^K(x_1,x_2) > 0$ and $p^L(y_1,y_2) > 0$. If M is large, $p^{L+M}(x_2,x_2) > 0$ and $p^{K+M}(y_2,y_2) > 0$. So $p^{K+L+M}(x_1,x_2) > 0$ and $p^{K+L+M}(y_1,y_2) > 0$, and hence we have $r^{K+L+M}((x_1,x_2),(y_1,y_2)) > 0$.

It is easy to check that $\pi'(a,b) = \pi(a)\pi(b)$ is a stationary distribution for Z'. Hence Z'_n is recurrent, and hence it will hit (x,x), hence the time to hit the diagonal $\{(y,y) : y \in \mathcal{S}\}$ is finite. However the distribution of the time to hit the diagonal is the same as T. □

27. Gaussian sequences.

We first prove a converse to Proposition 17.3.

Proposition 27.1. *If $\mathbb{E}\, e^{i(uX+vY)} = \mathbb{E}\, e^{iuX}\mathbb{E}\, e^{ivY}$ for all u and v, then X and Y are independent random variables.*

Proof. Let X' be a random variable with the same law as X, Y' one with the same law as Y, and X', Y' independent. (We let $\Omega = [0,1]^2$, $\mathbb{P}$ Lebesgue measure, X' a function of the first variable, and Y' a function of the second variable defined as in Proposition 1.2.) Then $\mathbb{E}\, e^{i(uX'+vY')} = \mathbb{E}\, e^{iuX'}\mathbb{E}\, e^{ivY'}$. Since X, X' have the same law, they have the same characteristic function, and similarly for Y, Y'. Therefore (X', Y') has the

same joint characteristic function as (X, Y). By the uniqueness of the Fourier transform, (X', Y') has the same joint law as (X, Y), which is easily seen to imply that X and Y are independent. □

A sequence of random variables $X_1, \ldots, X_n$ is said to be jointly normal if there exists a sequence of independent standard normal random variables $Z_1, \ldots, Z_m$ and constants b_{ij} and a_i such that $X_i = \sum_{j=1}^m b_{ij} Z_j + a_i$, $i = 1, \ldots, n$. In matrix notation, $X = BZ + A$. For simplicity, in what follows let us take $A = 0$; the modifications for the general case are easy. The covariance of two random variables X and Y is defined to be $\mathbb{E}\,[(X - \mathbb{E}\,X)(Y - \mathbb{E}\,Y)]$. Since we are assuming our normal random variables are mean 0, we can omit the centering at expectations. Given a sequence of mean 0 random variables, we can talk about the covariance matrix, which is $\mathrm{Cov}\,(X) = \mathbb{E}\,XX^t$, where X^t denotes the transpose of the vector X. In the above case, we see $\mathrm{Cov}\,(X) = \mathbb{E}\,[(BZ)(BZ)^t] = \mathbb{E}\,[BZZ^tB^t] = BB^t$, since $\mathbb{E}\,ZZ^t = I$, the identity.

Let us compute the joint characteristic function $\mathbb{E}\,e^{iu^tX}$ of the vector X, where u is an n-dimensional vector. First, if v is an m-dimensional vector,

$$\mathbb{E}\,e^{iv^tZ} = \mathbb{E}\prod_{j=1}^m e^{iv_jZ_j} = \prod_{j=1}^m \mathbb{E}\,e^{iv_jZ_j} = \prod_{j=1}^m e^{-v_j^2/2} = e^{-v^tv/2}$$

using the independence of the Z's. So

$$\mathbb{E}\,e^{iu^tX} = \mathbb{E}\,e^{iu^tBZ} = e^{-u^tBB^tu/2}.$$

By taking $u = (0, \ldots, 0, a, 0, \ldots, 0)$ to be a constant times the unit vector in the jth coordinate direction, we deduce that each of the X's is indeed normal.

Proposition 27.2. *If the X_i are jointly normal and $\mathrm{Cov}\,(X_i, X_j) = 0$ for $i \neq j$, then the X_i are independent.*

Proof. If $\mathrm{Cov}\,(X) = BB^t$ is a diagonal matrix, then the joint characteristic function of the X's factors, and so by Proposition 27.1, the Xs would in this case be independent. □

28. Stationary processes.

In this section we give some preliminaries which will be used in the next on the ergodic theorem. We say a sequence X_i is *stationary* if $(X_k, X_{k+1}, \ldots)$ has the same distribution as $(X_0, X_1, \ldots)$.

One example is if the X_i are i.i.d. For readers who are familiar with Markov chains, another is if X_i is a Markov chain, π is the stationary distribution, and X_0 has distribution π.

A third example is rotations of a circle. Let Ω be the unit circle, $\mathbb{P}$ normalized Lebesgue measure on Ω, and $\theta \in [0, 2\pi)$. We let $X_0(\omega) = \omega$ and set $X_n(\omega) = \omega + n\theta \pmod{2\pi}$.

A fourth example is the Bernoulli shift: let $\Omega = [0, 1)$, $\mathbb{P}$ Lebesgue measure, $X_0(\omega) = \omega$, and $X_n(\omega)$ be binary expansion of ω from the nth place on.

Proposition 28.1. *If X_n is stationary, then $Y_k = g(X_k, X_{k+1}, \ldots)$ is stationary.*

Proof. If $B \subset \mathbb{R}^\infty$, let

$$A = \{x = (x_0, x_1, \ldots) : (g(x_0, \ldots), g(x_1, \ldots,), \ldots) \in B\}.$$

Then

$$\begin{aligned}\mathbb{P}((Y_0, Y_1, \ldots) \in B) &= \mathbb{P}((X_0, X_1, \ldots) \in A)\\ &= \mathbb{P}((X_k, X_{k+1}, \ldots) \in A)\\ &= \mathbb{P}((Y_k, Y_{k+1}, \ldots) \in B).\end{aligned}$$

□

We say that $T:\Omega\to\Omega$ is *measure preserving* if $\mathbb{P}(T^{-1}A)=\mathbb{P}(A)$ for all $A\in\mathcal{F}$.

There is a one-to-one correspondence between measure preserving transformations and stationary sequences. Given T, let $X_0=\omega$ and $X_n=T^n\omega$. Then

$$\mathbb{P}((X_k,X_{k+1},\ldots)\in A)=\mathbb{P}(T^k(X_0,X_1,\ldots)\in A)=\mathbb{P}((X_0,X_1,\ldots)\in A).$$

On the other hand, if X_k is stationary, define $\widehat{\Omega}=\mathbb{R}^\infty$, and define $\widehat{X}_k(\omega)=\omega_k$, where $\omega=(\omega_0,\omega_1,\ldots)$. Define $\widehat{P}$ on $\widehat{\Omega}$ so that the law of $\widehat{X}$ under $\widehat{P}$ is the same as the law of X under $\mathbb{P}$. Then define $T\omega=(\omega_1,\omega_2,\ldots)$. We see that

$$\begin{aligned}\mathbb{P}(A)&=\mathbb{P}((\omega_0,\omega_1,\ldots)\in A)=\widehat{\mathbb{P}}((\widehat{X}_0,\widehat{X}_1,\ldots)\in A)\\&=\mathbb{P}((X_0,X_1,\ldots)\in A)=\mathbb{P}((X_1,X_2,\ldots)\in A)\\&=\widehat{P}((\widehat{X}_1,\widehat{X}_2,\ldots)\in A)=\mathbb{P}((\omega_1,\omega_2,\ldots)\in A)\\&=\mathbb{P}(T\omega\in A)=\mathbb{P}(T^{-1}A).\end{aligned}$$

We say a set A is *invariant* if $T^{-1}A=A$ (up to a null set, that is, the symmetric difference has probability zero). The invariant σ-field $\mathcal{I}$ is the collection of invariant sets. A measure preserving transformation is *ergodic* if the invariant σ-field is trivial.

In the case of an i.i.d. sequence, A invariant means $A=T^{-n}A\in\sigma(X_n,X_{n+1},\ldots)$ for each n. Hence each invariant set is in the tail σ-field, and by the Kolmogorov 0-1 law, T is ergodic.

In the case of rotations, if θ is a rational multiple of π, T need not be ergodic. For example, let $\theta=\pi$ and $A=(0,\pi/2)\cup(\pi,3\pi/2)$. However, if θ is an irrational multiple of π, the T is ergodic. To see that, recall that if f is measurable and bounded, then f is the L^2 limit of $\sum_{k=-K}^K c_k e^{ikx}$, where c_k are the Fourier coefficients. So

$$\begin{aligned}f(T^nx)&=\sum c_k e^{ikx+ikn\theta}\\&=\sum d_k e^{ikx},\end{aligned}$$

where $d_k=c_ke^{ikn\theta}$. If $f(T^nx)=f(x)$ a.e., then $c_k=d_k$, or $c_ke^{ikn\theta}=c_k$. But θ is not a rational multiple of π, so $e^{ikn\theta}\neq 1$, so $c_k=0$. Therefore $f=0$ a.e. If we take $f=1_A$, this says that either A is empty or A is the whole space, up to sets of measure zero.

Our last example was the Bernoulli shift. Let X_i be i.i.d. with $\mathbb{P}(X=1)=\mathbb{P}(X_i=0)=1/2$. Let $Y_n=\sum_{m=0}^\infty 2^{-(m+1)}X_{n+m}$. So there exists g such that $Y_n=g(X_n,X_{n+1},\ldots)$. If A is invariant for the Bernoulli shift,

$$A=((Y_n,Y_{n+1},\ldots)\in B)=((X_n,X_{n+1},\ldots)\in C),$$

where $C=\{x:(g(x_0,x_1,\ldots),g(x_1,x_2,\ldots),\ldots)\in B\}$. this is true for all n, so A is in the invariant σ-field for the X_i's, which is trivial. Therefore T is ergodic.

29. The ergodic theorem.

The key to the ergodic theorem is the following maximal lemma.

Lemma 29.1. *Let X be integrable. Let T be a measure preserving transformation, let $X_j(\omega)=X(T^j\omega)$, let $S_k(\omega)=X_0(\omega)+\cdots+X_{k-1}(\omega)$, and $M_k(\omega)=\max(0,S_1(\omega),\ldots,S_k(\omega))$. Then $\mathbb{E}\,[X;M_k>0]\geq 0$.*

Proof. If $j\leq k$, $M_k(T\omega)\geq S_j(T\omega)$, so $X(\omega)+M_k(T\omega)\geq X(\omega)+S_j(T\omega)=S_{j+1}(\omega)$, or

$$X(\omega)\geq S_{j+1}(\omega)-M_k(T\omega),\qquad j=1,\ldots,k.$$

Since $S_1(\omega) = X(\omega)$ and $M_k(T\omega) \geq 0$, then

$$X(\omega) \geq S_1(\omega) - M_k(T\omega).$$

Therefore

$$\begin{aligned}\mathbb{E}\left[X(\omega); M_k > 0\right] &\geq \int_{(M_k>0)} [\max(S_1, \ldots, S_k)(\omega) - M_k(T\omega)] \\ &= \int_{(M_k>0)} [M_k(\omega) - M_k(T\omega)].\end{aligned}$$

On the set $(M_k = 0)$ we have $M_k(\omega) - M_k(T\omega) = -M_k(T\omega) \leq 0$. Hence

$$\mathbb{E}\left[X(\omega); M_k > 0\right] \geq \int [M_k(\omega) - M_k(T\omega)].$$

Since T is measure preserving, $\mathbb{E}\, M_k(\omega) - \mathbb{E}\, M_k(T\omega) = 0$, which completes the proof. □

Recall $\mathcal{I}$ is the invariant σ-field. The ergodic theorem says the following.

Theorem 29.2. *Let T be measure preserving and X integrable. Then*

$$\frac{1}{n}\sum_{m=0}^{n-1} X(T^m\omega) \to \mathbb{E}\left[X \mid \mathcal{I}\right],$$

where the convergence takes place almost surely and in L^1.

Proof. We start with the a.s. result. By looking at $X - \mathbb{E}\left[X \mid \mathcal{I}\right]$, we may suppose $\mathbb{E}\left[X \mid \mathcal{I}\right] = 0$. Let $\varepsilon > 0$ and $D = \{\limsup S_n/n > \varepsilon\}$. We will show $\mathbb{P}(D) = 0$.

Let $\delta > 0$. Since X is integrable, $\sum \mathbb{P}(|X_n(\omega)| > \delta n) = \sum \mathbb{P}(|X| > \delta n) < \infty$ (cf. proof of Proposition 5.1). By Borel-Cantelli, $|X_n|/n$ will eventually be less than δ. Since δ is arbitrary, $|X_n|/n \to 0$ a.s. Since

$$(S_n/n)(T\omega) - (S_n/n)(\omega) = X_n(\omega)/n - X_0(\omega)/n \to 0,$$

then $\limsup(S_n/n)(T\omega) = \limsup(S_n/n)(\omega)$, and so $D \in \mathcal{I}$. Let $X^*(\omega) = (X(\omega) - \varepsilon)1_D(\omega)$, and define S_n^* and M_n^* analogously to the definitions of S_n and M_n. On D, $\limsup(S_n/n) > \varepsilon$, hence $\limsup(S_n^*/n) > 0$.

Let $F = \cup_n(M_n^* > 0)$. Note $\cup_{i=0}^n(M_i^* > 0) = (M_n^* > 0)$. Also $|X^*| \leq |X| + \varepsilon$ is integrable. By Lemma 29.1, $\mathbb{E}\left[X^*; M_n^* > 0\right] \geq 0$. By dominated convergence, $\mathbb{E}\left[X^*; F\right] \geq 0$.

We claim $D = F$, up to null sets. To see this, if $\limsup(S_n^*/n) > 0$, then $\omega \in \cup_n(M_n^* > 0)$. Hence $D \subset F$. On the other hand, if $\omega \in F$, then $M_n^* > 0$ for some n, so $X_n^* \neq 0$ for some n. By the definition of X^*, for some n, $T^n\omega \in D$, and since D is invariant, $\omega \in D$ a.s.

Recall $D \in \mathcal{I}$. Then

$$\begin{aligned}0 \leq \mathbb{E}\left[X^*; D\right] &= \mathbb{E}\left[X - \varepsilon; D\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{I}\right]; D\right] - \varepsilon\mathbb{P}(D) = -\varepsilon\mathbb{P}(D),\end{aligned}$$

using the fact that $\mathbb{E}\left[X \mid \mathcal{I}\right] = 0$. We conclude $\mathbb{P}(D) = 0$ as desired.

Since we have this for every ε, then $\limsup S_n/n \leq 0$. By applying the same argument to $-X$, we obtain $\liminf S_n/n \geq 0$, and we have proved the almost sure result. Let us now turn to the L^1 convergence. Let $M > 0$, $X_M' = X1_{(|X|\leq M)}$, and $X_M'' = X - X_M'$. By the almost sure result,

$$\frac{1}{n}\sum X_M'(T^m\omega) \to \mathbb{E}\left[X_M' \mid \mathcal{I}\right]$$

almost surely. Both sides are bounded by M, so

$$\mathbb{E}\left|\frac{1}{n}\sum X'_M(T^m\omega) - \mathbb{E}\left[X'_M \mid \mathcal{I}\right]\right| \to 0. \tag{29.1}$$

Let $\varepsilon > 0$ and choose M large so that $\mathbb{E}\,|X''_M| < \varepsilon$; this is possible by dominated convergence. We have

$$\mathbb{E}\left|\frac{1}{n}\sum_{m=0}^{n-1} X''_M(T^m\omega)\right| \le \frac{1}{n}\sum \mathbb{E}\,|X''_M(T^m\omega)| = \mathbb{E}\,|X''_M| \le \varepsilon$$

and

$$\mathbb{E}\,|\mathbb{E}\left[X''_M \mid \mathcal{I}\right]| \le \mathbb{E}\left[\mathbb{E}\left[|X''_M| \mid \mathcal{I}\right]\right] = \mathbb{E}\,|X''_M| \le \varepsilon.$$

So combining with (29.1)

$$\limsup \mathbb{E}\left|\frac{1}{n}\sum X(T^m\omega) - \mathbb{E}\left[X \mid \mathcal{I}\right]\right| \le 2\varepsilon.$$

This shows the L^1 convergence. □

What does the ergodic theorem tell us about our examples? In the case of i.i.d. random variables, we see $S_n/n \to \mathbb{E}\,X$ almost surely and in L^1, since $\mathbb{E}\left[X \mid I\right] = \mathbb{E}\,X$. Thus this gives another proof of the SLLN.

For rotations of the circle with $X(\omega) = 1_A(\omega)$ and θ is an irrational multiple of π, $\mathbb{E}\left[X \mid \mathcal{I}\right] = \mathbb{E}\,X = \mathbb{P}(A)$, the normalized Lebesgue measure of A. So the ergodic theorem says that $(1/n)\sum 1_A(\omega + n\theta)$, the average number of times $\omega + n\theta$ is in A, converges for almost every ω to the normalized Lebesgue measure of A.

Finally, in the case of the Bernoulli shift, it is easy to see that the ergodic theorem says that the average number of ones in the binary expansion of almost every point in $[0, 1)$ is $1/2$.