Probability Theory and Stochastic Processes
2015-2016 UAB

Paul Jung

Notation used in the text:

Throughout the text := denotes a definition or "set equal to". We use the symbol $\equiv$ to indicate two different notations for the same object, or in the case of functions, identically equal to a constant.

$\biguplus$ denotes a disjoint union.

Lebesgue measure is denoted by m, $m(dx)$, dx, or dt.

$\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$

$\{X < x\} := \{\omega : X(\omega) < x\}$

For events A and B, we often use a comma for intersection: $\{A, B\} := \{A \cap B\}$.

For events A, B or random variables X, Y, we denote independence by $A \perp\!\!\!\perp B$ or $X \perp\!\!\!\perp Y$.

$X \in \mathcal{F}$ denotes that X is $\mathcal{F}$-measurable.

$\mathbb{N}_0 := \mathbb{N} \cup \{0\}$

u.i. stands for *uniformly integrable*.

i.i.d. stands for *independent and identically distributed.*

$\mu_n \Rightarrow \mu$ denotes weak convergence of distributions.

$a_n \simeq b_n$ denotes $\lim_{n\to\infty} a_n/b_n = 1$.

Contents

1 Measure Theory

We will begin with a brief review of measure theory. As this is meant only for review, many proofs will be omitted, but can be found in [RF10] or [Rud87].

Definition 1.1. Given a set Ω, a *σ-algebra* or *σ-field* is a subset $\mathcal{F} \subset \mathcal{P}(\Omega)$ (the power set of Ω) which satisfies

(i) $\emptyset \in \mathcal{F}$ (thus $\mathcal{F}$ is nonempty)

(ii) $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$

(iii) $A_n \in \mathcal{F}$ for all n implies that $\cup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

Definition 1.2. A *measure* is a non-negative function $\mu : \mathcal{F} \to [0, \infty]$ which satisfies

(i) $\mu(\emptyset) = 0$

(ii) countable additivity: $\sum_{n\in\mathbb{N}} \mu(A_n) = \mu\left(\biguplus_{n\in\mathbb{N}} A_n\right)$ where $\biguplus$ denotes the disjoint union.

Definition 1.3. A *measurable space* is a pair $(\Omega, \mathcal{F})$, and a *measure space* is a triplet $(\Omega, \mathcal{F}, \mu)$.

Example 1.4 (Lebesgue measure). Let $\mathcal{B}$ be the Borel σ-field on $\Omega = \mathbb{R}$, i.e., the smallest σ-field containing all open sets. Thus, any open set G is in $\mathcal{B}$, and any closed set F closed is in $\mathcal{B}$. Also, $G_\delta \in \mathcal{B}$ and $F_\sigma \in \mathcal{B}$ where G_δ is a countable intersection of open sets and F_σ is a countable union of closed sets. Furthermore, countable unions of G_δ sets form the collection $G_{\delta\sigma} := (G_\delta)_\sigma$ and countable intersections of F_σ form $F_{\sigma\delta}$. This procedure can be recursively applied to get all of $\mathcal{B}$.

Let $m(\cdot)$ be the measure which assigns to any open interval, its length

$$m((a, b)) = b - a.$$

This is Lebesgue measure, and we have that $(\mathbb{R}, \mathcal{B}, m)$ is a measure space. One can extend the σ-field to include sets which are not in $\mathcal{B}$, but such a discussion is better left for a course in real analysis. One can also restrict Lebesgue measure to obtain a measure space of the form $(\mathbb{R}, \sigma(\{A\}), m)$. Here $\sigma(\{\cdot\})$ is the smallest σ-field containing $\{\cdot\}$, e.g. $\sigma(\{A\}) = \{\emptyset, A, A^c, \Omega\}$.

Example 1.5 (Counting measure). Let $\Omega = \mathbb{Z}$. We can define the counting measure space $(\mathbb{Z}, \mathcal{P}(\mathbb{Z}), \mu)$ where for $A \in \mathcal{P}(\mathbb{Z})$, $\mu(A) := \#\{x \in \mathbb{Z} : x \in A\}$. Again, we can also define measures by restricting the σ-field, for example $(\mathbb{Z}, \{\emptyset, A, A^c, \mathbb{Z}\}, \mu)$.

Definition 1.6. The *norm* of a measure μ, denoted $\|\mu\|$, is $\mu(\Omega)$.

If $\|\mu\| < \infty$, this corresponds to a true norm for the vector space of finite signed measures on Ω.

1.1 Random Variables

Definition 1.7. A *probability space* is a measure space where the norm of the measure is 1. We generally denote probability spaces as $(\Omega, \mathcal{F}, \mathbf{P})$. Any element $\omega \in \Omega$ is called an *outcome* and any $A \in \mathcal{F}$ is called an *event*. The set Ω is often called the *sample space*.

Any measure μ with a nonzero, finite norm can be made into a probability measure by *normalizing*, i.e., scaling the measure by $\|\mu\|^{-1}$.

Example 1.8. Consider the set of all possible results from n fair coin tosses which result in H or T. Let $\Omega = \{H, T\}^n$, $\mathcal{F} = \mathcal{P}(\Omega)$, and $\mathbf{P}(A) = \frac{\#A}{2^n}$. Then for $A = \{\text{exactly one } T\}$ we have that $\mathbf{P}(A) = \frac{n}{2^n}$. Note that $\mathbf{P}$ is a normalized counting measure.

Example 1.9. If we count the number of H's appearing in each $\omega \in \Omega$ of the previous example and identify (or merge into a single element) all elements with the same number of H's, then we get the sample space $\tilde{\Omega} = \{0, 1, \ldots, n\}$ with $\tilde{\mathcal{F}} = \mathcal{P}(\tilde{\Omega})$. One can then check that

$$\tilde{\mathbf{P}}(A) = \sum_{k \in A} \binom{n}{k} \frac{1}{2^n}$$

is the measure induced by $\mathbf{P}$ under this identification.

Example 1.10. Let $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}([0, 1])$, and $\mathbf{P}(A) = m(A)$ (Lebesgue measure). This can be thought of as the measure corresponding to infinitely many fair coin tosses with $\omega \in \Omega$ given by $\omega = (\omega_1, \omega_2, \ldots)$ and $\omega_i = 0$ or 1 with the association that $H = 1$ and $T = 0$. In other words, we use the dyadic expansion of $\omega \in [0, 1]$. For example:

$$\omega = 0.010110\ldots = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} + 0 \cdot \frac{1}{8} + 1 \cdot \frac{1}{16} + 1 \cdot \frac{1}{32} + \ldots$$

Definition 1.11. Given two measurable spaces $(\Omega, \mathcal{F})$ and $(\mathbb{R}, \mathcal{B})$, we say that $X : \Omega \to \mathbb{R}$ is an *$\mathcal{F}$-measurable function* if $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}$. We often just say *measurable* when our choice of $\mathcal{F}$ is implicit. Probabilists call measurable functions on $(\Omega, \mathcal{F}, \mathbf{P})$, *random variables*.

Remarks 1.12.

1. It is important to look at *inverse* images of Borel sets from the range of X and not *forward* images of $A \in \mathcal{F}$ from the domain of X. In general, the image of a measurable function X may not be measurable with respect to the Borel σ-field on $\mathbb{R}$. For instance, set $\Omega \subset \mathbb{R}$ such that $\Omega \notin \mathcal{B}$ and consider the σ-field $\mathcal{F} = \{\Omega \cap B : B \in \mathcal{B}\}$. Then the identity map restricted to Ω (or inclusion map $X : \Omega \hookrightarrow \mathbb{R}$) is $\mathcal{F}$-measurable, but the forward image of Ω (under the map X) is Ω which is not a Borel set.

2. Since $\{(-\infty, c), c \in \mathbb{R}\}$, or alternatively $\{(-\infty, c], c \in \mathbb{R}\}$, generate the Borel sets, the measurability condition is equivalent to $X^{-1}((-\infty, c)) \in \mathcal{F}$ for all $c \in \mathbb{R}$, or alternatively $X^{-1}((-\infty, c]) \in \mathcal{F}$ for all $c \in \mathbb{R}$.

Exercise 1.1. If X and Y are random variables, show that $X + Y$ and $X \cdot Y$ are random variables.

Example 1.13 (Binomial and Bernoulli random variables). The probability space described in Example 1.9 is a canonical one for the following important random variable. A random variable which counts the number of heads out of n independent coin tosses is called a *binomial* random variable. In general, the coin tosses come up heads with probability $0 < p < 1$, in which case a binomial random variable $X \sim \text{Bin}(n, p)$ is described by the probabilities

$$\mathbf{P}(\{\omega : X(\omega) = k\}) = \binom{n}{k} p^k (1-p)^{n-k}.$$

The special instance $n = 1$ is called a *Bernoulli* random variable, in which case we write $X \sim \text{Ber}(p)$.

Example 1.14 (Geometric random variables). Consider a probability space similar to Example 1.10, except that we toss an infinite number of coins which may not be fair. Suppose each coin comes up heads independently with probability $0 < p < 1$. A random variable X which counts the number of tosses required in order to get our first H is called a *geometric* random variable and we write $X \sim \text{Geom}(p)$. It is described by the probabilities

$$\mathbf{P}(\{\omega : X(\omega) = k\}) = p(1-p)^{n-1}.$$

Definition 1.15. If $X : (\Omega, \mathcal{F}) \to (\overline{\mathbb{R}}, \mathcal{B})$ is measurable, then we say that X is an *extended* random variable (Here, $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$). If $\vec{X} : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}^{\otimes n})$ is measurable, then we say that $\vec{X}$ is a *random vector* ($\mathcal{B}^{\otimes n}$ is just the Borel σ-field on $\mathbb{R}^n$).

Proposition 1.16. *For a countable set of random variables $\{X_n, n \in \mathbb{N}\}$, we have that $\inf_{n\in\mathbb{N}} X_n$, $\sum_{n\in\mathbb{N}} X_n$, $\liminf_{n\in\mathbb{N}} X_n$ and $\limsup_{n\in\mathbb{N}} X_n$ are all extended random variables.*

Proof. We shall only prove the first claim. We begin by noting that

$$\{\omega : \inf_{n\in\mathbb{N}} X_n(\omega) < a\} = \cup_{n\in\mathbb{N}} \{\omega : X_n(\omega) < a\}. \tag{1.1}$$

Each set $\{\omega : X_n(\omega) < a\} \in \mathcal{F}$ since X_n is measurable by assumption. Thus, $\cup_{n\in\mathbb{N}} \{\omega : X_n(\omega) < a\} \in \mathcal{F}$ as it is a countable union of measurable sets. □

Example 1.17. It is important that the index set of $\inf_{n\in\mathbb{N}} X_n$ is countable. Consider the probability space $([0,1],\mathcal{B},m)$, a fixed subset $A\subset[0,1]$, and the random variables

$$X_t(\omega)=\begin{cases}0 & \text{if } t=\omega \text{ and } \omega\in A^c\\ 1 & \text{otherwise.}\end{cases}$$

Then $\inf_{t\in[0,1]} X_t(\omega)=\mathbf{1}_A(\omega)$ which is measurable only if $A\in\mathcal{B}$.

1.2 Expectation

In the rest of this section we consider general measure spaces $(E,\mathcal{F},\mu)$ which are finite (unless otherwise stated) so that in particular, $\mu(E)<\infty$.

Definition 1.18. A measurable function $\varphi(x)$ is *simple* if φ takes finitely many values. A simple function φ is called an *indicator function* if $\varphi(x)\in\{0,1\}$ for all $x\in E$. We write

$$\mathbf{1}_A(x)=\begin{cases}1, & x\in A\\ 0, & x\in A^c\end{cases}.$$

Definition 1.19 (Integral of simple functions)**.** If $\varphi: E\to\mathbb{R}$ is simple and takes values $\{a_1,\dots,a_n\}$, then we define its *Lebesgue integral* as

$$\begin{aligned}\int_E \varphi\,d\mu &\equiv \int_E \varphi\,\mu(dx)\\ &:=\sum_{j=1}^{n} a_j\mu(\varphi^{-1}(a_j)).\end{aligned}\tag{1.2}$$

Example 1.20. Let A be a measurable subset of E. Then

$$\begin{aligned}\int_E \mathbf{1}_A\,d\mu &= 1\mu(A)+0\mu(A^c)\\ &=\mu(A).\end{aligned}$$

The following lemma is what makes the Lebesgue integral work and what gives it power.

Lemma 1.21 (Simple Approximation Lemma)**.** *A function f is measurable if and only if there exists a sequence of simple functions $(\varphi_n, n\in\mathbb{N})$ such that $\varphi_n(x)\to f(x)$ for all $x\in E$. Moreover, this sequence satisfies $|\varphi_n|\le|f|$ for all n. If f is nonnegative, the sequence $(\varphi_n,n\in\mathbb{N})$ can be chosen to be non-decreasing.*

Definition 1.22 (Integral of measurable functions)**.** We proceed in three steps, by defining the integral first for bounded and nonnegative functions, then for nonnegative functions, and finally for general measurable functions.

(i) If $f \geq 0$ is a bounded, measurable function on E, we define

$$\int_E f\,d\mu := \sup\left\{\int_E \varphi\,d\mu : \ \varphi \text{ is simple and } \varphi \leq f\right\}. \tag{1.3}$$

(ii) If $g \geq 0$ is measurable on E, we define

$$\int_E g\,d\mu := \sup\left\{\int_E f\,d\mu : \ f \text{ is a bounded, measurable function and } 0 \leq f \leq g\right\}.$$

Note that $\int_E g\,d\mu = \infty$ is possible.

(iii) If g is measurable, then there exists measurable functions g_1 and g_2 such that $g_1 \geq 0$, $g_2 < 0$, and $g = g_1 + g_2$. For example, let $A = g^{-1}([0,\infty))$. Then A is measurable and $g = \mathbf{1}_A g + \mathbf{1}_{A^c} g$. Using this decomposition, we can then define

$$\int_E g\,d\mu := \int_E g_1\,d\mu - \int_E (-g_2)\,d\mu \tag{1.4}$$

whenever (at least) one of the integrals on the right is finite. If both integrals on the right are infinite, the integral is undefined.

Exercise 1.2. For bounded, measurable f,

$$\int_E f\,d\mu = \inf\left\{\int_E \varphi\,d\mu : \ \varphi \text{ is simple and } \varphi \geq f\right\}.$$

Exercise 1.3. If $(E,\mu) = ([0,1], m)$ and f is bounded and Riemann integrable, then

$$\int_E f\,m(dx) = \int_E f\,dx.$$

Henceforth, all integrals are with respect to measures and dx or dt will denote Lebesgue measure (we also continue to sometimes use $m(dx)$ to denote Lebesgue measure).

Theorem 1.23 (Linearity and monotonicity)**.** *If f and g are measurable functions on (E,μ), then*

$$\int_E af + bg\,d\mu = a\int_E f\,d\mu + b\int_E g\,d\mu.$$

If $f \leq g$, then

$$\int_E f\,d\mu \leq \int_E g\,d\mu.$$

In a typical course in measure theory, one would prove linearity and monotonicity in three steps for measurable f and g: first for bounded functions, next for nonnegative functions, and finally for general measurable functions.

Corollary 1.24 (Triangle Type Inequality)**.** *If f is measurable, then*

$$\left|\int_E f\,d\mu\right| \leq \int_E |f|\,d\mu.$$

Proof. Since $-|f| \leq f \leq |f|$, we have by monotonicity that

$$-\int_E |f|\,d\mu \leq \int_E f\,d\mu \leq \int_E |f|\,d\mu.$$

□

Definition 1.25 (Integral over subsets)**.** For any measurable subset $A \subset E$, we set

$$\int_A f\,d\mu := \int_E f\mathbf{1}_A\,d\mu.$$

Proposition 1.26. *If $E = \biguplus_{n=1}^{\infty} A_n$, then*

$$\int_E f\,d\mu = \sum_{n=1}^{\infty} \int_{A_n} f\,d\mu.$$

Proof. This follows from linearity of the integral and the countable additivity of μ. □

Definition 1.27. We say that $f(x) = g(x)$ *almost everywhere* on E with respect to the measure μ, denoted a.e., if $\mu(\{x : f(x) \neq g(x)\}) = 0$. If $\|\mu\| = 1$, then we say that $f = g$ *almost surely*, denoted a.s., on E. We say that two sets A and B are equal a.e. or a.s. if their indictor functions are equal a.e. or a.s.

Remark 1.28. It is often the case in both probability theory and measure theory that one refers to a function f when really one means the equivalence class of functions which are equal to f a.s. or a.e., since typically one only wants to know things "up to measure zero".

Corollary 1.29. *If $f \overset{a.e.}{=} g$, then*

$$\int_E f\,d\mu = \int_E g\,d\mu.$$

Proof. Let $A = \{x : f(x) = g(x)\}$. Then

$$\begin{aligned}
\int_E f\,d\mu &= \int_A f\,d\mu + \int_{A^c} f\,d\mu \\
&= \int_A g\,d\mu + \int_{A^c} f\,d\mu \\
&= \int_A g\,d\mu \\
&= \int_A g\,d\mu + \int_{A^c} g\,d\mu \\
&= \int_E g\,d\mu.
\end{aligned}$$

□

Definition 1.30 (Convergence almost everywhere, almost surely)**.** We say $(f_n, n \in \mathbb{N})$ converges *almost everywhere* to f and write $f_n \xrightarrow{\text{a.e.}} f$ whenever it converges pointwise except on a set of measure zero:

$$\mu(\{x : \lim_{n\to\infty} f_n(x) \neq f(x)\}) = 0.$$

If μ is a probability measure, we say *almost surely* and write $f_n \xrightarrow{\text{a.s.}} f$.

Theorem 1.31 (Bounded Convergence Theorem)**.** *If $(f_n, n \in \mathbb{N})$ is a sequence of measurable functions for which there exists an $M \in \mathbb{R}$ such that $|f_n| \leq M$ for all n and $f_n \xrightarrow{a.e.} f$, then*

$$\int_E f \, d\mu = \lim_{n\to\infty} \int_E f_n \, d\mu.$$

It is crucial in the above theorem that $(E, \mathcal{F}, \mu)$ is a finite measure space. This is not so for the next three results, but for consistency let us continue to assume the setting of finite measure spaces.

Lemma 1.32 (Fatou's Lemma)**.** *If $(f_n, n \in \mathbb{N})$ is a sequence of nonnegative, measurable functions on E, then*

$$\int_E \liminf_{n\to\infty} f_n \, d\mu \leq \liminf_{n\to\infty} \int_E f_n \, d\mu.$$

Theorem 1.33 (Monotone Convergence Theorem)**.** *If $(f_n, n \in \mathbb{N})$ is a sequence of measurable functions on E such that $f_n \nearrow f$ a.e., then*

$$\int_E f \, d\mu = \lim_{n\to\infty} \int_E f_n \, d\mu.$$

Theorem 1.34 (Dominated Convergence Theorem)**.** *If $(f_n, n \in \mathbb{N})$ is a sequence of measurable functions on E for which $|f_n| \leq g$ for some g such that $\int_E |g| \, d\mu < \infty$. If $f_n \xrightarrow{a.e.} f$, then*

$$\lim_{n\to\infty} \int_E f_n \, d\mu = \int_E f \, d\mu.$$

Definition 1.35. If $\|\mu\| = 1$, then we call the integral (if it exists) of a random variable X, the *expectation* of X, denoted

$$\mathbf{E}X := \int_E X(x) \, d\mu(x).$$

Our typical notation for a probability space is $(\Omega, \mathcal{F}, \mathbf{P})$ in which case the above becomes

$$\mathbf{E}X = \int_\Omega X(\omega) \, d\mathbf{P}.$$

Exercise 1.4. Show that if $g : \mathbb{R} \to \mathbb{R}$ is a Borel measurable function and X is a random variable, then $g(X)$ is a random variable. In particular, the following is well-defined:

$$\mathbf{E}(g(X)) = \int_\Omega g(X(\omega))\, d\mathbf{P}.$$

Having introduced probabilist's notation for integration, in the rest of the section, we shall couch some typical theorems from real analysis in this setting. First, however, let us introduce two more terms used in the language of probability.

Definition 1.36. Since $g(x) = x^p$ is a Borel measurable function, $g(X) = X^p$ is a random variable. The pth *moment* of X is given by

$$\mathbf{E}X^p = \int_\Omega X^p(\omega)\, d\mathbf{P}.$$

Definition 1.37. $\mathbf{E}(X^2 - (\mathbf{E}X)^2)$ is called the *variance* of X, denoted $\operatorname{Var} X$. It is easy to see that $\operatorname{Var} X$ is equivalent also to $\mathbf{E}(X - \mathbf{E}X)^2$

Because $\mathbf{E}X$ and $\operatorname{Var} X$ play a central role, they often are denoted by $\mu = \mathbf{E}X$ and $\sigma^2 = \operatorname{Var} X$. We acknowledge that μ is being used for several different objects, but the reader should be able to deduce the meaning in each case by context.

Theorem 1.38 (Hölder's Inequality)**.** *Suppose p and q are conjugate exponents, i.e.,*

$$\frac{1}{p} + \frac{1}{q} = 1, \quad 1 < p, q < \infty.$$

If X and Y are random variables, then

$$\mathbf{E}\,|XY| \leq \|X\|_p \, \|Y\|_q \,,$$

where $\|X\|_p := (\mathbf{E}\,|X|^p)^{\frac{1}{p}}$.

In the case where $p = q = 2$, Hölder's Inequality is just the Cauchy-Schwarz Inequality (one can check that $\mathbf{E}\,|XY|$ defines an inner product between X and Y).

Exercise 1.5 (Paley-Zygmund Inequality). Let $Y \geq 0$ with $\mathbf{E}Y^2 < \infty$. For $\theta \in [0, 1]$

$$\mathbf{P}(Y > \theta \mathbf{E}Y) \geq (1 - \theta)^2 \frac{(\mathbf{E}Y)^2}{\mathbf{E}Y^2}\,.$$

Hint: Use Hölder's Inequality on the product $Y \cdot \mathbf{1}_{\{Y > \theta \mathbf{E}Y\}}$.

Exercise 1.6. Let $\mathbf{E}|X|^k < \infty$. Then for $0 < j < k$ we have

$$\mathbf{E}|X|^j \leq \mathbf{E}(|X|^k)^{j/k} < \infty\,.$$

Exercise 1.7 (Minkowski's Inequality). For $p \geq 1$ and random variables X and Y, show that

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

Minkowski's Inequality shows that $\|\cdot\|_p$ is a norm for the space of all random variables with finite pth moment. In fact, it turns out that this normed linear space is complete, thus making $L^p(\Omega)$ a *Banach space*. This motivates the following definition.

Definition 1.39. We say that a random variable X is *integrable* with respect to $\mathbf{P}$ if $\mathbf{E}|X| < \infty$ and write $X \in L^1(\Omega) \equiv L^1(\Omega, \mathcal{F}, \mathbf{P})$. If $\mathbf{E}|X|^p < \infty$, then we say $X \in L^p(\Omega)$.

Example 1.40. If $X \in L^p(\Omega)$ and $Y \in L^q(\Omega)$ where p and q are conjugate exponents, then by Hölder's Inequality, $XY \in L^1(\Omega)$.

Example 1.41. Let Z be a random variable. Define random variables $Y \equiv 1$ and $X = |Z|^\alpha$ for $\alpha \geq 1$. Then we have that

$$\mathbf{E}|X \cdot 1| \leq (\mathbf{E}|1|^q)^{\frac{1}{q}} (\mathbf{E}|X|^p)^{\frac{1}{p}}.$$

Thus,

$$(\mathbf{E}|Z|^\alpha)^{\frac{1}{\alpha}} \leq (\mathbf{E}|Z|^{\alpha p})^{\frac{1}{\alpha p}}$$

or $\|Z\|_\alpha \leq \|Z\|_{\alpha p}$. Thus, for $1 \leq p < q < \infty$, we have that $\|Z\|_p \leq \|Z\|_q$. In particular, if $X \in L^q(\Omega)$, then $X \in L^p(\Omega)$. An immediate result of this fact is that for a random variable X, $\operatorname{Var} X < \infty$ implies $\mathbf{E}|X| < \infty$ and thus, $\mathbf{E}X < \infty$. *Note:* This sort of result does not hold for general measure spaces. It relies on the assumption that we are working on a probability (and thus finite) measure space.

Exercise 1.8. (a) Use summation-by-parts (the discrete analog of integration-by-parts) to show that when X takes values in $\mathbb{N}$,

$$\mathbf{E}X = \sum_{n \in \mathbb{N}} \mathbf{P}(\{\omega : X(\omega) \geq n\}).$$

(b) For a general random variable, use Example 1.41 to show also that $\operatorname{Var} X < \infty$ if and only if

$$\sum_{n \in \mathbb{N}} \mathbf{E}(|X| \mathbf{1}_{\{|X| > n\}}) < \infty.$$

Definition 1.42. A function $\varphi : \mathbb{R} \to \mathbb{R}$ is said to be *convex* on $\mathbb{R}$ if for all $x, y \in \mathbb{R}$ and $\lambda \in [0, 1]$

$$\varphi(\lambda x + (1 - \lambda)y) \leq \lambda \varphi(x) + (1 - \lambda)\varphi(y).$$

Theorem 1.43 (Jensen's Inequality)**.** *If $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function, then*

$$\mathbf{E}(\varphi(X)) \geq \varphi(\mathbf{E}X)$$

provided that both are finite.

Proof. Consider a line $\ell(x) = ax+b$ satisfying $\ell(x) \leq \varphi(x)$ and $\ell(\mathbf{E}X) = \varphi(\mathbf{E}X)$. For convex functions, such a line always exists. Then $\varphi(X) \geq aX + b$, thus by monotonicity and linearity of expectation,

$$\mathbf{E}\varphi(X) \geq \mathbf{E}(aX + b) = a\mathbf{E}X + b = \ell(\mathbf{E}X) = \varphi(\mathbf{E}X).$$

□

Example 1.44. Let $\varphi(t) = t^p$ for $t \geq 0$ and $p \geq 1$. Then φ is a convex function. Let Z be a random variable and define the random variable X as $X = |Z|^\alpha$, $\alpha \geq 1$. Then Jensen's Inequality gives us that $\mathbf{E}\varphi(X) \geq \varphi(\mathbf{E}X)$. Therefore,

$$\mathbf{E}\,|Z|^{\alpha p} \geq (\mathbf{E}\,|Z|^\alpha)^p.$$

By taking the αp^{th} root of both sides we get

$$(\mathbf{E}\,|Z|^{\alpha p})^{\frac{1}{\alpha p}} \geq (\mathbf{E}\,|Z|^\alpha)^{\frac{1}{\alpha}}.$$

Thus, we again have the result that $\|Z\|_{\alpha p} \geq \|Z\|_\alpha$.

Theorem 1.45 (Markov's Inequality)**.** *For a random variable X and $u > 0$, we have that*

$$\mathbf{P}(\{\omega : |X(\omega)| \geq u\}) \leq \frac{\mathbf{E}\,|X|}{u}.$$

Proof. We begin by noting that

$$u\mathbf{1}_{\{\omega:|X(\omega)|\geq u\}} \leq |X|.$$

We can then take the expectation of both sides to get

$$u\mathbf{E}(\mathbf{1}_{\{\omega:|X(\omega)|\geq u\}}) \leq \mathbf{E}\,|X|.$$

By rewriting the expectation on the left as a probability and diving both sides by u, we get

$$\mathbf{P}(\{\omega : |X(\omega)| \geq u\}) \leq \frac{\mathbf{E}\,|X|}{u}.$$

□

Remark 1.46. When $u > 0$, the event $\{|X| \geq u\}$ is the same as $\{X^2 \geq u^2\}$, thus an easy corollary is *Chebyshev's Inequality*:

$$\mathbf{P}(\{\omega : |X(\omega)| \geq u\}) \leq \frac{\mathbf{E}X^2}{u^2}.$$

Exercise 1.9. If φ is a strictly convex function, then $\varphi(\mathbf{E}X) = \mathbf{E}\varphi(X)$ implies that X is a.s. constant.

Theorem 1.47 (Transformation Theorem)**.** *If $f : \mathbb{R}^n \to \mathbb{R}$ is a Borel measurable function and $\{X_i, 1 \leq i \leq n\}$ are random variables, then $f(X_1, \ldots, X_n)$ is a random variable.*

We now want to discuss the idea of product measures. In order to simplify the mechanics while still providing insight, we will consider the case where $n = 2$.

Lemma 1.48 (Product measure)**.** *Let $(E_1, \mathcal{F}_1, \mu_1)$ and $(E_2, \mathcal{F}_2, \mu_2)$ be measure spaces. There is a unique measure μ on $E = E_1 \times E_2$ with σ-field*

$$\mathcal{F}_1 \otimes \mathcal{F}_2 := \sigma(\{A \times B : A \in \mathcal{F}_1, B \in \mathcal{F}_2\}) =$$

such that $\mu(A \times B) = \mu_1 \times \mu_2(A \times B) := \mu_1(A)\mu_2(B)$.

Using the lemma above, we may define the notion of a *product σ-field* and *product measure* as the ones described in the lemma. One should check that, in the case of Lebesgue measure, these coincide with our previous notion of $(\mathbb{R}^n, \mathcal{B}^{\otimes n}, m)$. This is done by confirming the countable additivity of Lebesgue measure on $\mathcal{B}^{\otimes n}$, and then utilizing Caratheodory's Extension Theorem (see [RF10]).

Theorem 1.49 (Fubini-Tonelli Theorem)**.** *Suppose $E = E_1 \times E_2$ is endowed with a product σ-field and product measure. For a measurable function f on E, if $f \geq 0$ or $f \in L^1(E, \mathcal{F}, \mu)$, then*

$$\int_E f(x,y)\,\mu(dx,dy) = \int_{E_2} \underbrace{\int_{E_1} f(x,y)\,\mu_1(dx)}_{(*)}\ \mu_2(dy).$$

Remarks 1.50.

1. It is possible that $f(x,\cdot)$ be measurable for each y and $f(\cdot,y)$ be measurable for each x, but f is still not measurable. In such cases, the theorem cannot hold since the integral on the left is not even well-defined.

2. For Tonelli ($f \geq 0$), both sides may be infinite.

3. Part of the theorem is that $(*)$, as a function of y, is measurable with respect to $(E_2, \mathcal{F}_2, \mu_2)$.

4. Note that the iterated integral on the right may be done in either order since the designation of E_1 was arbitrary.

5. Similar to Fatou, Monotone Convergence, and Dominated Convergence, it is not necessary here that the measure spaces be finite.

1.3 Distributions

"A property is probability-theoretical if and only if it is described in terms of a distribution." – M. Loève [Loè77]

Definition 1.51. The *distribution*[1] of a random variable X is a measure μ_X on $(\mathbb{R}, \mathcal{B})$ such that for $A \in \mathcal{B}$, $\mu_X(A) := \mathbf{P}(X^{-1}(A))$. In other words, it is the induced probability measure on $\mathbb{R}$ by the measure $\mathbf{P}$ on Ω. We will also sometimes say that a measure μ on $(\mathbb{R}, \mathcal{B})$ is a distribution if it is a probability measure, even if there is no a priori associated random variable.

Remark 1.52. A less widely used term for the induced measure on $\mathbb{R}$ is the *law* of a random variable X. However, the term *law* is also sometimes used in reference to the measure $\mathbf{P}$ on the measurable space $(\Omega, \mathcal{F})$. Due to its ambiguity, we do not use this terminology in the sequel.

The quote at the beginning of this section is in the context of M. Loève writing on the conceptual difference between probability theory and general measure theory. We add to this the claim that the notion of **a distribution is the heart of a random variable**. To explain this claim, consider that a measurable space $(\Omega, \mathcal{F})$ is required to have very little structure– just enough to define a measure on it (for example we cannot discuss addition or continuity in the set Ω since it may not be a group and may not have a topology). This is reasonable since events which occur in real life typically do not come with natural algebraic, topological, or geometric structures.

Introducing a random variable X into the picture (or random vector $\vec{X}$) allows us to push probabilities into a space in which we have a great deal of structure, namely $\mathbb{R}$ (respectively $\mathbb{R}^n$). We can then forget about $(\Omega, \mathcal{F}, \mathbf{P})$ and work with $(\mathbb{R}, \mathcal{B}, \mu_X)$ (respectively $(\mathbb{R}^n, \mathcal{B}^{\otimes n}, \mu_{\vec{X}})$). The values X takes are now thought of as varying according to the induced measure μ_X, which is the motivation behind calling it a random variable rather than a function. This is embodied by the shorthand notation

$$\mathbf{P}(X \in A) \equiv \mathbf{P}(\{\omega : |X(\omega)| \in A\}).$$

Remark 1.53. Since Ω is typically an abstract space to begin with, one often just sets $\Omega = \mathbb{R}$. Then we may also set $\mu_X = \mathbf{P}$ in which case we might as well set $X(\omega) = \omega$ (the identity function). This is taking the above discussion to the extreme, but this sort of thinking is sometimes helpful particularly when when we later encounter sequences of random variables taking values in an infinite product space $\mathbb{R}^{\mathbb{N}} = \prod_{i=1}^{\infty} \mathbb{R}$.

Exercise 1.10. Show that μ_X is a probability measure.

Definition 1.54. A collection of random variables $\{X_i, i \in I\}$ is said to be *identically distributed* (i.d.) if the distribution of X_i is the same for all $i \in I$.

[1] These should not be confused with distributions in the theory of partial differential equations.

Example 1.55. Let $\Omega = \{X, T\}^3$ and consider the Bernoulli random variables

$$X_i = \begin{cases} 1, & \text{if } H \text{ on the } i^{\text{th}} \text{ toss} \\ 0, & \text{otherwise} \end{cases}$$

for $i \in \{1, 2, 3\}$. Then

$$\Omega = \left\{ \begin{array}{ll} \omega_1 = (H, H, H), & \omega_5 = (H, T, T), \\ \omega_2 = (H, H, T), & \omega_6 = (T, H, T), \\ \omega_3 = (H, T, H), & \omega_7 = (T, T, H), \\ \omega_4 = (T, H, H), & \omega_8 = (T, T, T) \end{array} \right\}.$$

Then the distribution of X_i for $i \in \{1, 2, 3\}$ is given by

$$\mu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1.$$

Thus, these random variables are identically distributed with distribution $\mu = \mu_{X_i}$.

Example 1.56 (Exponential distribution with rate λ). We say that X has an exponential distribution with rate λ and write $X \sim \text{Exp}(\lambda)$ if

$$\mu_X(A) = \int_{A \cap [0,\infty)} \lambda e^{-\lambda x} dx \qquad \text{for all } A \in \mathcal{B}.$$

Example 1.57 (Normal or Gaussian distribution). We say that X has a normal or Gaussian distribution with mean μ_0 and variance σ^2 and write $X \sim N(\mu_0, \sigma^2)$ if

$$\mu_X(A) = \int_A \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma^2}\right) dx$$

Definition 1.58. The *distribution function* of X is defined to be

$$F_X(x) := \mathbf{P}(X \leq x) \equiv \mathbf{P}(\{\omega : X(\omega) \leq x\}) = \mu_X((-\infty, x]).$$

One should not confuse distribution functions, which are actual functions on $\mathbb{R}$, with distributions which are probability measures on $\mathbb{R}$. To make the distinction clear, one often calls F_X a *cumulative distribution function* or more simply a *cdf*.

The following is a standard result from measure theory.

Lemma 1.59 (Continuity of measure)**.** *Suppose $A_n, A \in \mathcal{F}$ for some (possibly infinite) measure space $(E, \mathcal{F}, \mu)$. If $A_n \nearrow A$, then*

$$\mu(A) = \lim_{n\to\infty} \mu(A_n).$$

If μ is a finite measure and $A_n \searrow A$, then

$$\mu(A) = \lim_{n\to\infty} \mu(A_n).$$

In the following, we use the notation

$$f(x+) := \lim_{y \to x^+} f(y) \text{ and } f(x-) := \lim_{y \to x^-} f(y).$$

Proposition 1.60 (Distribution function properties)**.**

(i) F_X *is nondecreasing, i.e.,* $x \leq y$ *implies* $F_X(x) \leq F_X(y)$*.*

(ii) F_X *is right-continuous, i.e.,* $F(x+) = F(x)$ *for all* x*.*

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

(iv) $\mathbf{P}(X = x) = \mu_X(\{x\}) = F(x) - F(x-)$*.*

Proof.

(i) Since $x \leq y$, we have $(-\infty, x] \subset (-\infty, y]$ which implies

$$F_X(x) = \mu_X((-\infty, x]) \leq \mu_X((-\infty, y]) = F_X(y).$$

(ii) Suppose $x_n \searrow x$. Then $\bigcap_n (-\infty, x_n] = (-\infty, x]$ which implies

$$\begin{aligned} F_X(x+) &= \lim_{n \to \infty} \mu_X((-\infty, x_n]) \\ &= \mu_X(\bigcap_n (-\infty, x_n]) = \mu_X((-\infty, x]) = F_X(x) \end{aligned}$$

where the second equality follows from continuity of measure.

(iii) From the fact that $\bigcap_{n \in \mathbb{N}} (-\infty, -n] = \emptyset$, we conclude that

$$\begin{aligned} \lim_{x \to -\infty} F_X(x) &= \lim_{n \to \infty} \mu_X((-\infty, -n]) \\ &= \mu_X(\bigcap_{n \in \mathbb{N}} (-\infty, -n]) = \mu_X(\emptyset) = 0. \end{aligned}$$

From $\bigcup_{n \in \mathbb{N}} (-\infty, n] = \mathbb{R}$, we conclude that

$$\begin{aligned} \lim_{x \to \infty} F_X(x) &= \lim_{n \to \infty} \mu_X((-\infty, n]) \\ &= \mu_X(\bigcup_{n \in \mathbb{N}} (-\infty, n])) = \mu_X(\mathbb{R}) = 1 \end{aligned}$$

where the second equalities in both of the above follow from continuity of measure.

(iv) This follows from the fact that $F_X(x-) = \mathbf{P}(X < x)$, which one can easily check.

□

Theorem 1.61 (Characterization by distribution functions)**.** *If a function F satisfies properties (i), (ii), and (iii) of Proposition 1.60, then it is the distribution function of some random variable X, i.e., there is an X such that $F = F_X$.*

Proof. We will construct an X using $(\Omega, \mathcal{F}, \mathbf{P}) = ([0,1], \mathcal{B}([0,1]), m)$, i.e., Lebesgue measure on the Borel subsets of $[0,1]$. We define

$$X(\omega) := \sup\{y : F(y) < \omega\}.$$

When F is continuous and strictly increasing, X is its inverse, and this is how one should think about it even when F is discontinuous or not strictly increasing. If we can show that

$$\{\omega : X(\omega) \leq x\} = \{\omega : \omega \leq F(x)\}, \tag{1.5}$$

then

$$\begin{aligned}\mathbf{P}(X \leq x) &= \mathbf{P}(\omega \leq F(x)) \\ &= m([0, F(x)]) = F(x),\end{aligned}$$

as desired. So now we show (1.5).

[$\supset$] Suppose that $\omega \leq F(x)$. Since X is nondecreasing, we have

$$X(\omega) \leq X(F(x)) \leq x.$$

The latter inequality is equivalent to saying

$$\sup\{y : F(y) < F(x)\} \leq x,$$

which is true since if y is such that $F(y) < F(x)$, then $y \leq x$ since F is nondecreasing.

[$\subset$] Suppose that $X(\omega) \leq x$. We must show $\omega \leq F(x)$. By way of contradiction, suppose $\omega > F(x)$. Let $x_n \searrow x$ with $x_n > x$ for all n. By right-continuity, $F(x_n) \searrow F(x)$. Choose an $N \in \mathbb{N}$ such that $\omega > F(x_N) \geq F(x)$. Then we have

$$\sup\{y : F(y) < \omega\} \geq x_N$$

because x_N is one such y. Then by definition, $X(\omega) \geq x_N > x$, a contradiction. □

Remark 1.62. Since F_X is nondecreasing, all discontinuities are jump-discontinuities. Therefore, any F_X with a discontinuity of height c at point x must be associated to a distribution μ_X which is partly made up by the point mass $c\delta_x$. The point x is sometimes referred to as an *atom*.

Example 1.63 (Uniform distribution). Perhaps the simplest distribution functions are those of continuous uniform random variables $X \sim \text{Unif}[a, b]$ and

discrete uniform random variables $X \sim \text{Unif}\{x_1, \ldots, x_n\}$. For the continuous case we have

$$F_X(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & x \in (a,b) \\ 1, & x \geq b. \end{cases}$$

In the discrete case we have

$$F_X(x) = \begin{cases} 0, & x < x_1 \\ \frac{k}{n}, & x_k \leq x < x_{k+1} \text{ and } 1 \leq k < n \\ 1, & x \geq x_n. \end{cases}$$

The use of the words "discrete" and "continuous" apply to more than just uniform random variables. In particular, we say that a random variable is *discrete* if its distribution can be written in the form

$$\sum_{n \in \mathbb{N}} p_n \delta_{x_n}$$

for a countable set of values $\{x_n, n \in \mathbb{N}\}$ which occur with probabilities $\{p_n, n \in \mathbb{N}\}$ summing to one (in the *finite* case, infinitely many of the probabilities are zero). A random variable is *continuous* if its distribution function F_X is continuous. However, it is often the case that when one says X is continuous, one really means the slightly stronger statement that its distribution is absolutely continuous, which we now discuss.

Definition 1.64. If μ and ν are measures on $(\mathbb{R}, \mathcal{B})$, we say that ν is *absolutely continuous* with respect to μ if for all $A \in \mathcal{B}$, $\mu(A) = 0 \;\Rightarrow\; \nu(A) = 0$, and we write $\nu \ll \mu$.

Before moving on, let us motivate the above definition. Given a random variable X, we have up until now, three different yet equivalent ways of describing a probability measure. Firstly, the abstract way, $(\Omega, \mathcal{F}, \mathbf{P})$. Secondly, using the measure μ_X on $(\mathbb{R}, \mathcal{B})$ induced by X. Finally, Theorem 1.61 tells us that the distribution function F_X uniquely determines the measure μ_X. In the rest of the section we will show that the notion of absolute continuity provides a fourth description of the probability measure that holds whenever a distribution is absolutely continuous with respect to Lebesgue measure.

Definition 1.65. A measure space $(E, \mathcal{F}, \mu)$ is *σ-finite* if there exists a countable collection $\{E_n, n \in \mathbb{N}\} \subset \mathcal{F}$ such that $E = \bigcup_{n \in \mathbb{N}} E_n$ and $\mu(E_n) < \infty$ for all n.

Theorem 1.66 (Radon-Nikodym Theorem)**.** *Suppose $(E, \mathcal{B}, \mu)$ is a σ-finite measure space. Then $\nu \ll \mu$ if and only if there is a measurable function $f \geq 0$ such that*

$$\nu(B) = \int_B f \, d\mu \quad \textit{for all } B \in \mathcal{B}. \tag{1.6}$$

The function f is called the Radon-Nikodym derivative and is unique μ-a.e.

Proof. For the proof we refer to [Rud87]. However, one can easily see the μ-a.e. uniqueness of the function f, for if f and g both satisfy (1.6), and differ on a set of positive μ-measure, then there is a set $B \in \mathcal{B}$ such that $\int_B (f-g)\,d\mu \neq 0$. Hence $\nu(B) = \int_B f\ d\mu \neq \int_B g\ d\mu = \nu(B)$, a contradiction. □

Definition 1.67. We say that a function F on $\mathbb{R}$ is *absolutely continuous* on an interval $[a,b]$ if for every $\epsilon > 0$, there exists a $\delta > 0$ such that for every $n \in \mathbb{N}$ and every collection $\{(a_k, b_k)\}_{k=1}^n$ of disjoint open subintervals of $[a,b]$ such that $\sum_{k=1}^n (b_k - a_k) < \delta$, we have $\sum_{k=1}^n |F(b_k) - F(a_k)| < \epsilon$.

Note that absolute continuity implies uniform continuity, which in turn implies continuity.

Theorem 1.68 (Fundamental Theorem of Calculus)**.** *A function F is absolutely continuous on $[a,b]$ if and only if $F'(x)$ exists a.e., $F' \in L^1([a,b])$, and*

$$F(x) - F(a) = \int_a^x F'(t)\,dt \qquad \text{for all} x \in [a,b].$$

As usual, dt denotes Lebesgue measure.

Definition 1.69. If X is a random variable with an absolutely continuous distribution function F_X, then its *density* (sometimes called *probability density function* or *pdf*) is $f_X(x) := F_X'(x)$.

The previous theorem shows that a probability distribution on $\mathbb{R}$ is absolutely continuous with respect to Lebesgue measure, precisely when its associated distribution function is absolutely continuous (thus affording us the dual use of this terminology). In particular, one gets that

$$\mu_X(A) = \int_A f_X(x)dx,$$

thus when F_X is absolutely continuous on $\mathbb{R}$, we have four equivalent ways of interpreting the probability measure.

Example 1.70 (Exponential and Normal densities). If X has an exponential distribution with rate λ then

$$f_X(x) = \lambda e^{-\lambda x}\mathbf{1}_{[0,\infty)}$$

and thus for $c > 0$,

$$F_X(c) = \mathbf{P}(X \leq c) = \int_0^c \lambda e^{-\lambda x}\,dx.$$

If X has a normal distribution with mean μ_0 and variance σ^2 then

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma^2}\right).$$

Example 1.71. Clearly, F_X must be continuous in order for it to be absolutely continuous , so there is no density for the random variable $X \sim \text{Ber}(0,1)$, which is a coin flip assigning 1 to heads and 0 to tails. It is not enough, however, that F_X is continuous, in order for X to have a density. Consider the continuous Cantor-Lebesgue function F on $[0,1]$, i.e., the Devil's Staircase, which can be extended to all of $\mathbb{R}$ by setting its value to 1 for $x > 1$ and 0 for $x < 0$. Then, it is a distribution function. Since $F'(x) = 0$ a.e.,

$$F(1) - F(0) = 1 \neq 0 = \int_0^1 F'(x)dx.$$

We see that F is not absolutely continuous thus has no associated density.

Exercise 1.11. If the distribution μ_X is absolutely continuous with density f_X, show that for any Borel measurable function h,

$$\mathbf{E}h(X) = \int_{\mathbb{R}} h(x) f_X(x)\, dx.$$

2 Bernoulli's Laws of Large Numbers

2.1 Independence and Convolution

"Measure theory ends and probability begins with the definition of independence."
–R. Durrett [Dur10]

Before defining independence, let us consider the motivation behind the way it is defined. Let us suppose some event $B \subset \Omega$ (with $\mathbf{P}(B) > 0$) is known occur (say to someone with extra or inside information). Conditioned on the information that B has occurred, the probability that B^c occurs must then be 0. On the other hand, it may be that $\mathbf{P}(B) < 1$, yet we know B occurs. It is natural then, under the assumption that B occurs, to normalize $\mathbf{P}$ by dividing by $\mathbf{P}(B)$. This gives us a new probability

$$\tilde{\mathbf{P}}(A) = \mathbf{P}(A|B) := \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$$

on Ω, called the **conditional probability**, where $\tilde{\mathbf{P}}(A)$ and $\mathbf{P}(A|B)$ are just two different notations for the same thing and are defined by the right-hand side. We read $\mathbf{P}(A|B)$ as the probability that A occurs given that B occurs. With this in mind, if we think of A and B as being independent of each other, then knowledge of B should not affect the probability of A occurring. So, we should expect $\mathbf{P}(A|B) = \mathbf{P}(A)$. This would then imply that $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. We now define independence in a variety of contexts.

These considerations motivate us to make the following definition of independence on the space $(\Omega, \mathcal{F}, \mathbf{P})$: we say that two events A and B are **independent** (with respect to $\mathbf{P}$), and we write $A \perp\!\!\!\perp B$, if

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

Given $(\Omega, \mathcal{F}, \mathbf{P})$, two sub-$\sigma$-fields of $\mathcal{F}$, say $\mathcal{G}$ and $\mathcal{H}$, are said to be **independent** if $A \perp\!\!\!\perp B$ for all $A \in \mathcal{G}$ and $B \in \mathcal{H}$. The notion of independent σ-fields is just an extension of independent events. To see this, note that if $A \perp\!\!\!\perp B$, then $A \perp\!\!\!\perp B^c$, for we have

$$\begin{aligned} \mathbf{P}(A \cap B^c) &= \mathbf{P}(A) - \mathbf{P}(A \cap B) = \mathbf{P}(A) - \mathbf{P}(A)\mathbf{P}(B) \\ &= \mathbf{P}(A)(1 - \mathbf{P}(B)) = \mathbf{P}(A)\mathbf{P}(B^c). \end{aligned}$$

Since $\sigma(\{A\}) = \{\emptyset, A, A^c, \Omega\}$ and $\sigma(\{B\}) = \{\emptyset, B, B^c, \Omega\}$, it follows that $A \perp\!\!\!\perp B$ implies that every element of $\sigma(\{A\})$ is independent of every element of $\sigma(\{B\})$.

We say that two random variables X and Y are **independent** if $\sigma(X) \perp\!\!\!\perp \sigma(Y)$, where for a random variable Z we define the **σ-field generated by** Z to be $\sigma(Z) := \{Z^{-1}(B) : B \in \mathcal{B}\}$ (one can check that this is a σ-field). Note that $\sigma(Z)$ is in fact the smallest σ-field that can be constructed from Ω that allows

Z to be measurable. Also, notice that if $X \perp\!\!\!\perp Y$, then

$$\begin{aligned}\mathbf{P}(X \in [a,b], Y \in (c,d)) &= \mathbf{P}(\underbrace{\{\omega : X(\omega) \in [a,b]\}}_{=X^{-1}([a,b])\in\sigma(X)} \cap \underbrace{\{\omega : Y(\omega) \in (c,d)\}}_{=Y^{-1}((c,d))\in\sigma(Y)}) \\ &= \mathbf{P}(X \in [a,b])\mathbf{P}(Y \in (c,d)).\end{aligned}$$

We could of course use any two Borel sets in place of $[a,b]$ and (c,d).

We say that a finite number of events $A_1, \ldots, A_n$ are **independent** if for every index set $I \subset \{1, \ldots, n\}$, we have

$$\mathbf{P}\Big(\bigcap_{i\in I} A_i\Big) = \prod_{i\in I} \mathbf{P}(A_i).$$

The events $A_1, \ldots, A_n$ are **pairwise independent** if for every $1 \le i < j \le n$,

$$\mathbf{P}(A_i \cap A_j) = \mathbf{P}(A_i)\mathbf{P}(A_j).$$

It is clear that independence implies pairwise independence.

Exercise 2.1. Show that the converse is not true. In other words, construct events which are pairwise independent, but not independent.

We say that finitely many σ-fields $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are **independent** if

$$\mathbf{P}\Big(\bigcap_{i=1}^{n} A_i\Big) = \prod_{i=1}^{n} \mathbf{P}(A_i)$$

for all $A_i \in \mathcal{F}_i$, $1 \le i \le n$. Defining this in terms of arbitrary index sets I, as above, is not necessary because we can always let some $A_i = \Omega$.

Finitely many random variables $X_1, \ldots, X_n$ are **independent** if $\sigma(X_1), \ldots, \sigma(X_n)$ are independent.

Lastly, whether we are talking about events, σ-fields, or random variables, an infinite collection is said to be **independent** if every finite subcollection is independent. For now, we assume that such infinite collections exist and address the existence issue later in Theorem 3.12.

Having defined independence eight times, let us develop some properties and see how it is useful. First, a result:

Theorem 2.1 (Independence Transformation Theorem)**.** *Suppose $\{X_i, i \in \mathbb{N}\}$ are independent random variables and*

$$f_i : \mathbb{R} \to \mathbb{R},\ i \in \mathbb{N},$$

are Borel measurable. Then $\{f_i(X_i), i \in \mathbb{N}\}$ are independent.

Proof. Since we must show that $\{\sigma(f_i(X_i)), i \in \mathbb{N}\}$ are independent, it suffices to show that $\sigma(f_i(X_i)) \subset \sigma(X_i)$ since $\{\sigma(X_i), i \in \mathbb{N}\}$ are independent. This means we must show

$$\{X^{-1}(f^{-1}(B)) : B \in \mathcal{B}\} \subset \{X^{-1}(B) : B \in \mathcal{B}\},$$

but this is clear for if B is a Borel set, so is $f^{-1}(B)$ since f is Borel measurable. Hence $X^{-1}(f^{-1}(B))$ is of the form $X^{-1}(B')$ where $B' \in \mathcal{B}$. □

By applying the same arguments, one can show

Corollary 2.2. *If $\{X_{ij}, (i,j) \in \mathbb{N}^2\}$ are independent, and*

$$f_i : \mathbb{R}^{m_i} \to \mathbb{R},\ i \in \mathbb{N},$$

are Borel measurable, then $\{f_i(X_{i1}, \ldots, X_{im_i}), i \in \mathbb{N}\}$ are independent.

Proposition 2.3 (Product measures and independence)**.** *If $\{X_i, i \in \mathbb{N}\}$ are independent with distributions $\{\mu_i, 1 \le i \le n\}$ respectively, then the random vector $\vec{X} = (X_1, \ldots, X_n)$ has distribution $\mu := \prod_{i=1}^n \mu_i$ on $(\mathbb{R}^n, \mathcal{B}^{\otimes n})$.*

Proof. First we check that μ coincides with the distribution when applied to product sets $A_1 \times \cdots \times A_n$, where $A_i \in \mathcal{B}$. We have

$$\begin{aligned} \mathbf{P}((X_1, \ldots, X_n) \in A_1 \times \cdots \times A_n) &= \mathbf{P}(X_1 \in A_1, \ldots, X_n \in A_n) \\ &= \mathbf{P}(X_1 \in A_1) \cdots \mathbf{P}(X_n \in A_n) \\ &= \mu_1(A_1) \cdots \mu_n(A_n) \\ &= \mu(A_1 \times \cdots \times A_n). \end{aligned}$$

Now we must show that $\mu(B) = \mathbf{P}(\vec{X} \in B)$ for arbitrary $B \in \mathcal{B}^{\otimes n}$. We have thus far only shown this for very special B, however, note that the Borel σ-field in $\mathbb{R}^n$ is generated by products $A_1 \times \cdots \times A_n$ of Borel sets in $\mathbb{R}$. Hence, since μ and the distribution coincide on a generating set, then they coincide on all elements of $\mathcal{B}^{\otimes n}$. □

Example 2.4. If $X \sim N(0,1)$ and $Y \sim N(0,1)$ are independent, then they are distributed as

$$\begin{aligned} \mu_X(A) &= \int_A \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, m(dx), \\ \mu_Y(B) &= \int_B \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, m(dy). \end{aligned}$$

By Theorem 2.3, the distribution of (X, Y) is given by

$$\mu_{(X,Y)}(A \times B) = (\mu_X \times \mu_Y)(A \times B) = \int\int_{A \times B} \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}\, dxdy.$$

As one might expect, we see that $\mu_{(X,Y)}$ turns out to be the two-dimensional standard Gaussian distribution.

More generally, if X and Y have absolutely continuous distribution functions F_X and F_Y, then they have densities f_X and f_Y. If $X \perp\!\!\!\perp Y$, by Theorem 2.3 one has

$$\mu_{(X,Y)}(B) = \int\int_B f_X f_Y\, dxdy. \tag{2.1}$$

Conversely, if $\mu_{(X,Y)}$ is given by the above equation, it is an easy exercise to see that $X \perp\!\!\!\perp Y$.

Definition 2.5. The *joint distribution* for a collection of random variables $\{X_1, \ldots, X_n\}$, is the distribution μ on $(\mathbb{R}^n, \mathcal{B}^{\otimes n})$ of the random vector

$$\vec{X} = (X_1, \ldots, X_n),$$

i.e., $\mu(A) = \mathbf{P}(\vec{X} \in A)$ for all $A \in \mathcal{B}^{\otimes n}$. Likewise, if

$$\mu(A) = \int_A f_{\vec{X}}(x_1, \ldots, x_n)\, m(d\vec{x})$$

for all $A \in \mathcal{B}^{\otimes n}$, then $f_{\vec{X}}$ is said to be a *joint density* for the collection $\{X_1, \ldots, X_n\}$

Remark 2.6. We often abuse notation by switching indiscriminately between $\vec{X}$ and $\{X_1, \ldots, X_n\}$. Similarly, we sometimes say "joint distribution" (when thought of as a collection) and other times simply say "distribution" (when thought of as a vector). Note that in the above definition, independence of the random variables is not required. See the examples below.

Corollary 2.7. *If $\{X_1, \ldots, X_n\}$ are independent and have densities $f_{x_1}, \ldots, f_{x_n}$, then the random vector $\vec{X} = (X_1, \ldots, X_n)$ has joint density $f_{\vec{X}}(x_1, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R}$ defined by*

$$f_{\vec{X}} = \prod_{i=1}^{n} f_{x_i}.$$

Example 2.8. If $X_1 \sim \text{Unif}(0, 1)$ and $X_1 = X_2$, i.e., the two random variables are not only identically distributed but in fact identical, then the random vector $\vec{X} = (X_1, X_2)$ has a joint distribution but no joint density since the distribution concentrates on the one-dimensional line $y = x$. This is despite the fact that X_1 and X_2 both have densities.

Example 2.9. For $X_1 \perp\!\!\!\perp X_2$, let $X_1 \sim N(0, 1)$ and let X_2 have a density defined by

$$f_{X_2}(x) = \begin{cases} \frac{2}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, & \text{if x} < 0 \\ 0, & \text{otherwise} \end{cases}$$

Also, let $X_3 = X_1$ if $X_1 > 0$ and $X_3 = X_2$ if $X_1 < 0$. Then, $X_3 \sim N(0, 1)$, but X_3 is not independent from X_1. Again, the random vector $\vec{X} = (X_1, X_3)$ has a joint distribution but no joint density since part of the distribution concentrates on the the line $y = x$ for $x > 0$. However, if one conditions on the event $X_1 < 0$ (or $X_3 < 0$), then the conditional distribution has a joint density. In fact, this conditional density is proportional to the density of a standard two-dimensional Gaussian random variable, restricted to the quadrant $x < 0, y < 0$.

Definition 2.10. Suppose the vector $\vec{X} = (X_1, X_2)$ has distribution μ. The probabilities $\mu(B)$ for all B of the form $(a, b) \times \mathbb{R}$, determine a *marginal distribution*, μ_1, defined by

$$\mu_1((a, b)) := \mu((a, b) \times \mathbb{R})$$

Remarks 2.11.

1. Given $\mu((a_1, b_1) \times \mathbb{R})$ and $\mu(\mathbb{R} \times (a_2, b_2))$ for all open (a_1, b_1) and (a_2, b_2), we then know the marginal distributions μ_1 and μ_2, but we still do not know the distribution μ for the random vector. However, in the special case where X_1 and X_2 are independent, one can easily extract μ, since then we obtain that

$$\begin{aligned} \mu(((a_1, b_1) \times \mathbb{R}) \cap (\mathbb{R} \times (a_2, b_2))) &= \mu((a_1, b_1) \times (a_2, b_2)) \\ &\underbrace{=}_{\text{by} \perp\!\!\!\perp} \mu_1((a_1, b_1))\mu_2(a_2, b_2)). \end{aligned}$$

2. If B_i is the support of μ_i, then μ is supported on $B_1 \times B_2$, but $B_1 \times B_2$ may not be its support. In other words, even if $\mu_i(A_i) > 0$ for both $A_i \subset B_i$, this does not imply that $\mu(A_1 \times A_2) > 0$.

3. If for $B \in \mathcal{B} \otimes \mathcal{B}$, the distribution μ is given by a density function

$$\mu(B) = \int \int_B f_{\vec{X}}(x_1, x_2) dx_1 dx_2,$$

   then for $B_1 \in \mathcal{B}$, the marginal distribution $\mu_1(B_1) = \int_{B_1} f_{X_1}\, dx_1$, where $f_{X_1} = \int_{\mathbb{R}} f_{\vec{X}}(x_1, x_2)\, dx_2$ is called the *marginal density*. If in addition $X_1 \perp\!\!\!\perp X_2$, then $f_{\vec{X}} = f_{X_1} \cdot f_{X_2}$. In fact when marginal densities exist, the previous statement is "if and only if".

4. If $\vec{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ with $n \geq 3$, then the marginal distributions described above are the one-point or one-dimensional marginals. More generally, the *k-point* or *k-dimensional marginals* are the distributions of vectors $(X_{j_1}, X_{j_2}, \ldots, X_{j_k})$ where $1 \leq k < n$.

Example 2.12. Suppose

$$\mu(A) = \int \int_A e^{-y} \mathbf{1}_{\{0<x<y\}}\, dxdy,$$

then

$$f_X(x) = \int_x^\infty e^{-y}\, dy = e^{-x} \text{ for } x > 0 \text{ and}$$

$$f_Y(y) = \int_0^y e^{-y}\, dx = e^{-y} \int_0^y 1\, dx = ye^{-y} \text{ for } y > 0.$$

Hence, since $e^{-y}\mathbf{1}_{\{0<x<y\}} \neq ye^{-y}e^{-x}1_{\{x>0,y>0\}}$, we can conclude that X and Y are not independent.

Proposition 2.13. *If X and Y are independent and $\mathbf{E}|X| < \infty$ and $\mathbf{E}|Y| < \infty$, then $\mathbf{E}|XY| < \infty$ and $\mathbf{E}XY \equiv \mathbf{E}(XY) = \mathbf{E}X\mathbf{E}Y$.*

Proof. First note that if we let $X = \mathbf{1}_A$ and $Y = \mathbf{1}_B$, then

$$\mathbf{E}XY = \mathbf{E}\mathbf{1}_{A\cap B} = \mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B) = \mathbf{E}X\mathbf{E}Y.$$

The idea is that the class of indicator random variables are in some sense the building blocks of all random variables. The next step is to use linearity to extend the result to all simple functions of the type $X = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}$ Then, using the Simple Approximation Lemma and the Monotone Convergence Theorem, one can show that the theorem holds for all $X \geq 0$ and $Y \geq 0$. Finally, considering the negative and nonnegative parts of arbitrary X and Y separately completes the proof. □

Independence is hugely important in probability theory and most of the fundamental theorems and basic models in the sequel are built on independent random variables at some level. These theorem and models are, however, only starting points for more complex models which require some level of dependence in order to be more realistic. Thus, we now quickly introduce the most basic tools for measuring dependence.

Definition 2.14.

(i) If $-\infty < \mathbf{E}XY = \mathbf{E}X\mathbf{E}Y < \infty$, then X and Y are said to be *uncorrelated*. Note that *uncorrelated* does not imply independence.

(ii) If $\infty > \mathbf{E}XY > \mathbf{E}X\mathbf{E}Y > -\infty$, then X and Y are *positively correlated*.

(iii) If $-\infty < \mathbf{E}XY < \mathbf{E}X\mathbf{E}Y < \infty$, then X and Y are *negatively correlated*.

Definition 2.15.

(i) The *covariance* of X and Y is

$$\mathrm{Cov}(X, Y) := \mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)] = \mathbf{E}XY - \mathbf{E}X\mathbf{E}Y.$$

(ii) The *correlation*[2] of X and Y is

$$\mathrm{Corr}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}\, X}\sqrt{\mathrm{Var}\, Y}}.$$

[2]The notion of correlation is extremely important in probability; it represents one of the first ways of measuring dependence, hence providing a foil to independence. The modern form of correlation is due to Pearson, but it is generally recognized (including by Pearson himself) that Galton invented this concept in 1888, after many writings on similar ideas. On a related note, Pearson is also the first person to use the term "standard deviation" [Sti89].

(iii) The *covariance matrix* Σ associated to a random vector $(X_1, \ldots, X_n) \in \mathbb{R}^n$ is

$$\begin{pmatrix} \operatorname{Cov}(X_1, X_1) & \operatorname{Cov}(X_1, X_2) & \ldots & \operatorname{Cov}(X_1, X_n) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Cov}(X_2, X_2) & \ldots & \operatorname{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_n, X_1) & \operatorname{Cov}(X_n, X_2) & \ldots & \operatorname{Cov}(X_n, X_n) \end{pmatrix},$$

where each entry $\Sigma_{ij} = \operatorname{Cov}(X_i, X_j)$.

Remarks 2.16.

1. It is easy to see that $\operatorname{Cov}(X, X) = \operatorname{Var} X$.
2. We have $\operatorname{Cov}(aX, bY) = ab \operatorname{Cov}(X, Y)$ which implies $\operatorname{Var}(cX) = c^2 \operatorname{Var} X$ and also
$$\underbrace{\sqrt{\operatorname{Var}(cX)}}_{\sigma_{cX}} = \underbrace{c\sqrt{\operatorname{Var}(X)}}_{c\sigma_X}$$
where σ_X is called the *standard deviation* of X.
3. We have $\operatorname{Cov}(X, Y+Z) = \operatorname{Cov}(Y+Z, X) = \operatorname{Cov}(X, Y) + \operatorname{Cov}(X, Z)$ which together with (b) shows that covariance is a symmetric, bilinear form.
4. A very common formula is
$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i,j} \operatorname{Cov}(X_i, X_j),$$
which is the same as summing all the entries of the covariance matrix for $(X_1, \ldots, X_n)$.

Exercise 2.2. Show that $|\operatorname{Corr}(X, Y)| \leq 1$. Also, find when $\operatorname{Corr}(X, Y) = 1$ and when $\operatorname{Corr}(X, Y) = -1$.

Exercise 2.3. Show that the covariance matrix Σ of any random vector $\vec{X}$ must be positive semi-definite, i.e.,

$$\mathbf{v}^T \Sigma \mathbf{v} \geq 0 \quad \text{for all } \mathbf{v} \in \mathbb{R}^n.$$

One way to do this is to consider the variance of the scalar random variable given by the dot product $\mathbf{v} \cdot \vec{X}$. Conversely, show that any positive semi-definite matrix Σ is the covariance matrix of some random vector.

Hint: for the converse direction, take a random vector $\vec{X}$ whose marginals are all independent and which each have variance one. Since Σ is positive definite, the square-root matrix is well-defined. Calculate the covariance matrix for the random vector $\sqrt{\Sigma}\vec{X}$.

Example 2.17 (Multivariate Gaussian distribution). Covariance matrices are especially useful when dealing with Gaussian distributions in high dimensions. In particular, if $\vec{\mu} = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$ and Σ is an $n \times n$ positive definite matrix then the n-dimensional Gaussian distribution with mean vector $\vec{\mu}$ and covariance matrix Σ has the density

$$(2\pi)^{-\frac{n}{2}}(\det\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right) \quad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

We are nearing our first big result in probability theory (both historically and in traditional pedagogy) which uses independence to analyze the limiting behavior of normalized sums of i.i.d. random variables. The limiting behavior is described by the so-called *Law of Large Numbers* attributed to Jakob Bernoulli [Ber13]. However, before moving to the limiting behavior, let us first present a method for obtaining a complete description of the distribution of sums of a fixed finite number of independent random variables.

Definition 2.18. If μ_X and μ_Y are distributions on $\mathbb{R}$ corresponding to independent random variables X and Y, then their *convolution*, defined in terms of the product measure by

$$\mu_X * \mu_Y(A) := \mu_X \times \mu_Y(\{(x,y) : x + y \in A\}),$$

is the distribution of the sum $X + Y$.

Remarks 2.19.

1. It is very important that $X+Y$ is a sum of *independent* random variables.
2. The convolution is a distribution on $\mathbb{R}$ even though it is defined in terms of a distribution on $\mathbb{R}^2$.
3. This notion extends to random vectors in $\vec{X}, \vec{Y} \in \mathbb{R}^n$ (as long as they are both in the same dimension n).

Exercise 2.4 (Poisson distribution). We say that $X \sim \text{Poiss}(\lambda)$ has a Poisson[3] distribution with mean λ if

$$\mathbf{P}(X = k) = e^{-k}\frac{\lambda^k}{k!} \text{for } k \in [0, \infty).$$

If $X \sim \text{Poiss}(\lambda)$ and $Y \sim \text{Poiss}(\kappa)$ are independent, use convolution to show $X + Y \sim \text{Poiss}(\lambda + \kappa)$. If $X \sim \text{Bin}(n,p)$ and $Y \sim \text{Bin}(m,p)$ are independent, use convolution to show $X + Y \sim \text{Bin}(n + m, p)$.

[3] As noted in [JKK05, p. 157], this distribution and its corresponding convergence theorem were most likely first discovered by A. De Moivre in 1711, well before Poisson's time. This an example of Stigler's Law, the notion that the name attached to a mathematical theorem is never the person that actually discovered the theorem.

Proposition 2.20 (Convolution is a semigroup)**.** *As an operation, convolution is commutative and associative. Moreover, δ_0 is an identity with respect to convolution of dsitributions.*

Proof. Commutativity and associativity follow from these properties for addition (of independent random variables). Similarly, since $X \equiv 0$ is an identity for addition of independent random variables, its distribution δ_0 is the identity for the operation of convolution. □

Remarks 2.21.

1. Convolution is not a group since the only possible inverse of X would be $-X$, but these are clearly not independent.

2. The *n-fold convolution* of μ_X with itself corresponds to the sum of n i.i.d. random variables which have the same distribution as X. It is denoted by $\nu = \mu_X^{*n}$. Equivalently we say that μ_X is the *nth convolution root* of ν, and we may write $\nu^{*1/n} = \mu_X$.

3. If the nth convolution root of ν exists for every $n \in \mathbb{N}$, we say that ν is an *infinitely divisible distribution.*

Exercise 2.5. Show that the nth convolution root of the Gaussian distribution $N(\mu, \sigma^2)$ is the Gaussian distribution $N(\mu/n, \sigma^2/n)$. In particular, every Gaussian distribution is infinitely divisible.

Proposition 2.22. *Let $A - y = \{z : z + x \in A\}$. The convolution of μ_X and μ_Y is given by*

$$\mu_X * \mu_Y(A) = \int_{\mathbb{R}} \mu_X(A - y)\mu_Y(dy).$$

If μ_X and μ_Y have associated densities f_X and f_Y, then $\mu_X * \mu_Y$ also has a density which is given by

$$f_{X+Y}(z) = \int_{\mathbb{R}} f_X(z - y) f_Y(y)\, dy.$$

Proof. Set $B := \{(x, y) : x + y \in A\}$. Then, using Tonelli's Theorem, we have

$$\begin{aligned}\mu_X * \mu_Y(A) &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathbf{1}_B(x, y)\mu_X(dx) \right) \mu_Y(dy) \\ &= \int_{\mathbb{R}} \mu_X(A - y)\, \mu_Y(dy).\end{aligned}$$

For the second part, integrate the function

$$g(z) := \int_{\mathbb{R}} f_X(z - y) f_Y(y)\, dy$$

over a set $A \in \mathcal{B}$ to get

$$\begin{aligned}\int_A g(x)\,dx &= \int_A \left(\int_{\mathbb{R}} f_X(x-y) f_Y(y)\,dy\right) dx \\ &= \int_{\mathbb{R}} \left(\int_A f_X(x-y)\,dx\right) f_Y(y)\,dy \\ &= \int_{\mathbb{R}} \mu_X(A-y) f_Y(y)\,dy.\end{aligned}$$

The right side is equal to $\mu_X * \mu_Y(A)$ by the first part of the proposition, thus g must be the density corresponding to the distribution $\mu_X * \mu_Y$. □

Example 2.23 (Gamma distribution). We say that X has a Gamma distribution with rate $\lambda > 0$ and shape parameter $\nu > 0$, and write $X \sim \text{Gamma}(\nu, \lambda)$, if

$$\mu_X(A) = \int_{A\cap[0,\infty)} \frac{\lambda^\nu x^{\nu-1}}{\Gamma(\nu)} e^{-\lambda x} dx \qquad \text{for all } A \in \mathcal{B}.$$

Note that when $\nu = 1$, this is just an exponential distribution with rate λ For $\nu \in \mathbb{N}$,

$$\Gamma(\nu) = (\nu-1)!$$

whereas for other values this is the well-known Gamma function.

If $X \sim \text{Gamma}(\nu_1, \lambda)$ and $Y \sim \text{Gamma}(\nu_2, \lambda)$ are independent, then by Proposition 2.22, $X + Y$ has a density given by

$$\begin{aligned}f_{X+Y}(z) &= \int_0^z \frac{\lambda^{\nu_1+\nu_2}}{\Gamma(\nu_1)\Gamma(\nu_2)} (z-y)^{\nu_1-1} e^{-\lambda(z-y)} y^{\nu_2-1} e^{-\lambda y}\,dy \\ &= \frac{\lambda^{\nu_1+\nu_2} e^{-\lambda z}}{\Gamma(\nu_1+\nu_2)} \int_0^z \frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} (z-y)^{\nu_1-1} y^{\nu_2-1}\,dy \\ &= \frac{\lambda^{\nu_1+\nu_2} e^{-\lambda z}}{\Gamma(\nu_1+\nu_2)}\end{aligned}$$

where the integral is seen to be equal to one by substituting $y = zu$ and $dy = z\,du$ to turn the integrand into a Beta function (or simply use the fact that the right side must be a density). Thus we see that the sum of two independent Gamma random variables with the same rate, gives us another Gamma random variable.

2.2 Weak Law of Large Numbers

Definition 2.24. We say the sequence $(X_n, n \in \mathbb{N})$ *converges in probability* to X, denoted by $X_n \xrightarrow{\text{pr}} X$, if for every $\epsilon > 0$, there exists an N such that $n \geq N$ implies that $\mathbf{P}(|X_n - X| > \epsilon) < \epsilon$.

Exercise 2.6. Show that Fatou's Lemma, the Dominated Convergence Theorem, and the Monotone Convergence Theorem all remain valid if we replace convergence a.s. with convergence in probability.

Exercise 2.7. Suppose a function $h : \mathbb{R} \to \mathbb{R}$ is continuous. If $X_n \xrightarrow{\text{pr}} X$, then $h(X_n) \xrightarrow{\text{pr}} h(X)$. If $X_n \xrightarrow{\text{a.s.}} X$, then $h(X_n) \xrightarrow{\text{a.s.}} h(X)$. These are versions of what is known as the *Continuous Mapping Theorem*[4].

Example 2.25. (The shrinking and revolving interval) Set

$$f_1(x) = \mathbf{1}_{[0,1]},\ f_2(x) = \mathbf{1}_{[1,1\frac{1}{2}]}, \ldots,\ f_n(x) = \mathbf{1}_{[\sum_{k=1}^{n-1}\frac{1}{k}, \sum_{k=1}^{n}\frac{1}{k}]}.$$

Considering the intervals above, modulo 1, we set

$$g_1(x) = \mathbf{1}_{[0,1]},\ g_2(x) = \mathbf{1}_{[0,\frac{1}{2}]}, \ldots,\ g_n(x) = \mathbf{1}_{[\sum_{k=1}^{n-1}\frac{1}{k} \pmod 1, \sum_{k=1}^{n}\frac{1}{k} \pmod 1]}$$

where modulo 1 simply means that we slide back to $[0, 1]$ (if the left endpoint becomes greater than the right endpoint, modulo 1, we split the interval in two in the natural way). Then for any fixed $\omega \in [0, 1], g_n(\omega) = 1$ for infinity many n. The sequence $(g_n, n \in \mathbb{N})$ converges pointwise nowhere, but if we let $g \equiv 0$, then

$$\mathbf{P}(|g_n - g| > \epsilon) = \frac{1}{n}.$$

Hence, the sequence converges in probability.

Example 2.26. Let the random variable $g_i = \mathbf{1}_{[i,i+1]}$. If $(\mathbb{R}, \mathcal{B}, \mu)$ is a probability space, then for all x, $\lim_{i\to\infty} g_i(x) = 0$, and hence $g_i \xrightarrow{\text{a.s.}} 0$. Also, $g_i \xrightarrow{\text{pr}} 0$. For each $\epsilon > 0$, simply choose N_ϵ large enough such that $\mu([-N_\epsilon, N_\epsilon]) > 1 - \epsilon$. Then for $n > N_\epsilon$, $\mathbf{P}(|g_n - 0| > \epsilon) < \epsilon$. Note that in this example, if one uses Lebesgue measure instead of the probability measure μ, then the sequence does not *converge in measure*, see for example [RF10].

Theorem 2.27 (Weak Law of Large Numbers, finite 2nd moments)**.** *If $\{X_n, n \in \mathbb{N}\}$ are independent and identically distributed (i.i.d.) and $\mathbf{E}X_1^2 < \infty$, then*

$$\frac{S_n}{n} \xrightarrow{pr} \mathbf{E}X_1$$

where $S_n = X_1 + \cdots + X_n$. In fact, the assumption of independence in the above can be weakened to $\text{Cov}(X_i, X_j) \leq 0$ for all $i \neq j$.

[4]This result and analogs were proved in [MW43].

Proof. By Chebyshev's Inequality, we obtain

$$\begin{aligned}
\mathbf{P}\left(\left|\frac{S_n}{n} - \mathbf{E}X_1\right| > \epsilon\right) &\leq \frac{\mathbf{E}\left|\frac{S_n}{n} - \mathbf{E}X_1\right|^2}{\epsilon^2} \\
&= \frac{\mathrm{Var}(\frac{S_n}{n})}{\epsilon^2} \ (\text{since } \mathbf{E}\,(S_n/n) = n\mathbf{E}X_1/n = \mathbf{E}X_1 < \infty) \\
&= \frac{1}{n^2\epsilon^2}\,\mathrm{Var}(X_1 + \cdots + X_n) \\
&= \frac{1}{n^2\epsilon^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i, X_j) \\
&\leq \frac{1}{n^2\epsilon^2} n\,\mathrm{Var}\,X_1 \ (\text{since } \mathrm{Cov}(X_i, X_j) \leq 0 \text{ for all } i \neq j) \\
&= \frac{1}{n\epsilon^2}\,\mathrm{Var}\,X_1 \xrightarrow{n\to\infty} 0.
\end{aligned}$$

□

Remarks 2.28.

1. The result is not generally valid for positively correlated random variables. If $\mathbf{E}X_1^2 < \infty$ and all X_i are identical, i.e., $X_i = X_1$ (think of n people observing the same coin toss), then $\frac{S_n}{n} = \frac{nX_1}{n} = X_1 \neq \mathbf{E}X_1$, unless $X_1 \equiv$ constant.

2. The assumption of being identically distributed can also be relaxed. For example, one can use $\{X_n, n \in \mathbb{N}\}$ which have bounded variances and are pairwise uncorrelated or negatively correlated. Then without changing the proof too much, one can obtain

$$\frac{S_n - \mathbf{E}(S_n)}{n} \xrightarrow{\text{pr}} 0.$$

Definition 2.29. We say that $(X_n, n \in \mathbb{N})$ *converges in* L^p and write $X_n \xrightarrow{L^p} X$, if

$$\lim_{n\to\infty} \mathbf{E}\,|X_n - X|^p = 0.$$

Remarks 2.30.

1. By Chebyshev's Inequality, for $p > 0$,

$$\mathbf{P}(|X_n - X| > \epsilon) \leq \frac{\mathbf{E}\,|X_n - X|^p}{\epsilon^p},$$

   and thus convergence in L^p implies convergence in probability.

2. In the proof of the Weak Law of Large Numbers (WLLN), we actually proved the stronger result $(\frac{S_n}{n}, n \in \mathbb{N})$ converges in L^2 to $\mathbf{E}X_1$.

3. The shrinking, revolving interval in Example 2.25 shows that it is possible for a sequence to converge in L^p, but not almost surely.

2.3 Strong Law of Large Numbers

Definition 2.31. If $(A_n, n \in \mathbb{N})$ is a sequence of events, we define the events

(i) $\{A_n \underbrace{\text{ infinitely often}}_{\text{i.o.}}\} := \limsup_{n\to\infty} A_n \equiv \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m$

(ii) $\{A_n \text{ eventually}\} := \liminf_{n\to\infty} A_n \equiv \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m$.

Example 2.32. Consider independent fair coin tossing events, and let

$$A_n = \{\text{Heads on the } n\text{-th toss}\}.$$

Then $\mathbf{P}(A_n \text{ i.o.}) = 1$ and $\mathbf{P}(A_n \text{ eventually}) = 0$.

Proposition 2.33 (Characterization of "a.s."). *We have that* $X_n \xrightarrow{a.s.} X$ *if and only if*

$$\mathbf{P}(|X_n - X| > \epsilon \ \ \textit{i.o.}) = 1 - \mathbf{P}(|X_n - X| \le \epsilon \ \ \textit{eventually}) = 0$$

for all $\epsilon > 0$.

Proof. Without loss of generality, we can let $X = 0$ by setting $Y_n = X_n - X$. Then

$$\{\omega : Y_n(\omega) \nrightarrow 0\} = \bigcup_{k=1}^{\infty} \{|Y_n| > \frac{1}{k} \text{ i.o.}\}.$$

Thus, $\mathbf{P}(Y_n \nrightarrow 0) = 0$ if and only if $\mathbf{P}(Y_n > \frac{1}{k} \text{ i.o.}) = 0$ for all $k \in \mathbb{N}$. □

Example 2.34. Let X_n be identical random variables, i.e., $X_1 = X_n$ for all n, and suppose X_1 is a Bernoulli random variable with $p = 1/2$. Also, let $\{Y_n, n \in \mathbb{N}\}$ be i.i.d. Bernoulli random variables with $p = 1/2$. If we set $A_n = \{X_n = 1\}$ and $B_n = \{Y_n = 1\}$, then

$$\begin{aligned} \mathbf{P}(A_n \text{ i.o.}) &= \frac{1}{2} \quad \text{and} \quad \mathbf{P}(A_n \text{ eventually}) = \frac{1}{2}. \\ \mathbf{P}(B_n \text{ i.o.}) &= 1 \quad \text{and} \quad \mathbf{P}(B_n \text{ eventually}) = 0. \end{aligned}$$

Lemma 2.35 (Borel-Cantelli Lemma). *If* $\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$, *then*

$$\mathbf{P}(A_n \textit{ i.o.}) = 0 \quad \textit{(Borel-Cantelli I).}$$

If in addition, the events $\{A_n, n \in \mathbb{N}\}$ *are independent, then* $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$ *implies that*

$$\mathbf{P}(A_n \textit{ i.o.}) = 1 \quad \textit{(Borel-Cantelli II).}$$

Proof. For the proof of the first part, we have

$$\begin{aligned}\mathbf{P}(A_n \text{ i.o.}) &= \mathbf{P}(\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m) \\ &= \lim_{n\to\infty} \mathbf{P}(\bigcup_{m=n}^{\infty} A_m), \text{ (by continuity of measure)} \\ &\leq \lim_{n\to\infty} \sum_{m=n}^{\infty} \mathbf{P}(A_m) \text{ (by subadditivity).}\end{aligned}$$

But $\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$ by assumption, which implies the tail of the series on the right-hand side goes to zero.

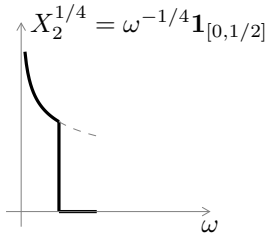
For the second part, we will use the bound $1 - x \leq e^{-x}$. Let $m \leq N < \infty$.

$$\begin{aligned}\lim_{N\to\infty} \mathbf{P}\left(\bigcap_{n=m}^{N} A_n^c\right) &= \prod_{n=m}^{N} (1 - \mathbf{P}(A_n)) \leq \lim_{N\to\infty} \prod_{n=m}^{N} \exp\left(-\mathbf{P}(A_n)\right) \\ &= \lim_{N\to\infty} \exp\left(-\sum_{n=m}^{N} \mathbf{P}(A_n)\right) = 0.\end{aligned}$$

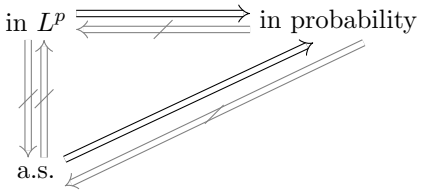
By continuity of measure this implies $\mathbf{P}\left(\bigcap_{n=m}^{\infty} A_n^c\right) = 0$, and by De Morgan's Law $\mathbf{P}\left(\left[\bigcup_{n=m}^{\infty} A_n\right]^c\right) = 0$. Thus $\mathbf{P}\left(\bigcup_{n=m}^{\infty} A_n\right) = 1$ for any m. Therefore $\mathbf{P}\left(\limsup A_n\right) = 1$. □

Exercise 2.8. Show using the Borel-Cantelli Lemma that $X_n \xrightarrow{\text{a.s.}} X$ implies that $X_n \xrightarrow{\text{pr}} X$. Conversely, if $X_n \xrightarrow{\text{pr}} X$, there exists $\{n_k\}$ such that we have the subsequential convergence $X_{n_k} \xrightarrow{\text{a.s.}} X$.

Example 2.36. Let $\Omega = [0, 1]$, fix $a > 0$, and set $X_n^a = \omega^{-a}\, \mathbf{1}_{[0,1/n]}$. This forms a sequence of random variables converging to 0 a.s. and in probability, but which does not belong to L^p when $ap > 1$. Therefore this sequence cannot converge in L^p when $ap > 1$.



Example 2.25, Remarks 2.30, Exercise 2.8, and Example 2.36 give the following convergence diagram:



Example 2.37 (Simple Random Walk). The usefuleness of the Borel-Cantelli Lemma is immediately displayed by the following. A *Rademacher* random variable has the distribution

$$\mathrm{Rad}(p) = \begin{cases} 1 & \text{with probability } p \\ -1 & \text{with probability } 1-p. \end{cases}$$

These are the same as Bernoulli random variables except the value 0 is replaced by -1. Let $\{X_n, n \in \mathbb{N}\}$ be a family of i.i.d. Rademacher random variables with parameter p, and let

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

We call the sequence $(S_n, n \in \mathbb{N})$, a one-dimensional **Simple Random Walk** with drift $p - \frac{1}{2}$.

Clearly S_n has the distribution $2\,\mathrm{Bin}(n,p) - n$ and by the WLLN,

$$\frac{S_n}{n} \xrightarrow{\text{pr}} 2p - 1.$$

We will show also that $p \neq 1/2$ implies

$$\mathbf{P}(S_n = 0 \text{ i.o.}) = 0.$$

Indeed, $S_n \neq 0$ for all odd n, and

$$\mathbf{P}(S_{2n} = 0) = \binom{2n}{n} p^n (1-p)^n = \frac{2n!}{n!n!} p^n (1-p)^n.$$

We next use *Stirling's formula*

$$\lim_{n\to\infty} \frac{(n/e)^n \sqrt{2\pi n}}{n!} = 1$$

to approximate the above expression by

$$\frac{1}{\sqrt{\pi n}} \left(4p(1-p)\right)^n.$$

Note that $\sum_{n\geq 1} \mathbf{P}(S_{2n} = 0) < \infty$ for $p \neq 1/2$, and thus by Borel-Cantelli I, $\mathbf{P}(S_n = 0 \text{ i.o.}) = 0$.

For $p = 1/2$, in fact, $\mathbf{P}(S_n = 0 \text{ i.o.}) = 1$, but we save this for later.

Theorem 2.38 (Strong Law of Large Numbers, finite 4th moments)**.** *If* $\{X_n, n \in \mathbb{N}\}$ *are i.i.d. and* $\mathbf{E}X_1^4 < \infty$*, then*

$$\frac{S_n}{n} \xrightarrow{a.s.} \mathbf{E}X_1$$

where $S_n = X_1 + \cdots + X_n$*.*

Proof. Without loss of generality assume that $\mathbf{E}X_1 = 0$, or use $Y_k = X_k - \mathbf{E}X_1$. By Borel-Cantelli I it is enough to show that for all $\epsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbf{P}\left(\left|\frac{S_n}{n}\right| > \epsilon\right) < \infty,$$

then it would follow that $\mathbf{P}\left(\left|\frac{S_n}{n}\right| > \epsilon \text{ i.o.}\right) = 0$ and therefore $\frac{S_n}{n} \xrightarrow{\text{a.s.}} 0$.

Now

$$\begin{aligned}
\mathbf{E}S_n^4 &= \sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{i=1}^{n} \mathbf{E}(X_i X_j X_k X_l) \\
&= \sum_{k=1}^{n} \mathbf{E}X_k^4 + 2 \sum_{1 \le j < k \le n} \mathbf{E}X_j^2 \, \mathbf{E}X_k^2 \\
&\le n\mathbf{E}X_1^4 + n^2 \left(\mathbf{E}X_1^2\right)^2 \le Cn^2
\end{aligned}$$

for some constant $C > 0$. Therefore

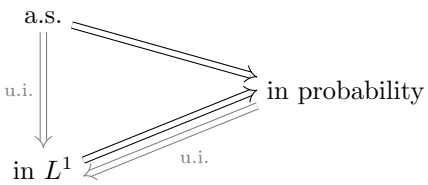
$$\mathbf{P}\left(\left|\frac{S_n}{n}\right|^4 > \epsilon^4\right) \le \frac{\mathbf{E}\left|\frac{S_n}{n}\right|^4}{\epsilon^4} = \frac{\mathbf{E}\left|S_n\right|^4}{\epsilon^4 n^4} \le \frac{C}{n^2\epsilon^4}.$$

To complete the proof, note that $\sum_{n=1}^{\infty} 1/n^2 < \infty$. □

2.4 Uniform Integrability and the L^1 Law of Large Numbers

We have up until now proven that $(\frac{S_n}{n}, n \in \mathbb{N})$ converges a.s. and in probability to its mean, under 2nd and 4th moment conditions, respectively. We will eventually see that the SLLN holds assuming only finite first moments. In this section, we first prove the WLLN under finite 1st moments. In fact, we will prove the stronger statement of convergence in L^1 under this minimal moment condition.

First, however, we will introduce a reasonable condition called *uniform integrability* (u.i.) by which either the SLLN or the WLLN, under only a finite 1st moment condition, implies the LLN in L^1. In particular, uniform integrability implies the following relations between convergence:



Exercise 2.9. Show that $X \in L^1(\Omega, \mathcal{F}, \mathbf{P})$ if and only if for all $\epsilon > 0$,

$$\text{there exists } \delta > 0 \text{ such that } \mathbf{E}\left(|X|\mathbf{1}_A\right) < \epsilon \text{ whenever } \mathbf{P}(A) \leq \delta. \tag{2.2}$$

Definition 2.39. We say $\{X_t, t \in T\}$ is *uniformly integrable* if (2.2) holds uniformly in t, i.e., for a given ϵ, the same δ works for all X_t.

Example 2.40. Consider the probability space $([0,1], \mathcal{B}, m)$ and let

$$X_k(\omega) = \begin{cases} k, & \omega \in [0, 1/k] \\ 0, & \text{otherwise.} \end{cases}$$

For each $\delta > 0$, and $A = [0, \delta]$, we have $\mathbf{E}\left(|X|\mathbf{1}_A\right) = 1$ whenever $k > 1/\delta$. Thus, the family $\{X_k, k \in \mathbb{N}\}$ is not uniformly integrable. Note that this is the case even though each random variable is individually bounded.

Our first observation regarding uniform integrability is the following. Instead of requiring small integrals over uniformly small regions in the domains of the random variables, we may instead look at increasingly unlikely regions in the ranges of the random variables.

Lemma 2.41 (Characterization of uniform integrability)**.** *The family* $\{X_t, t \in T\}$ *is uniformly integrable if and only if*

$$\underbrace{\lim_{n\to\infty}}_{integrable} \quad \underbrace{\sup_{t\in T}}_{uniformly} \quad \mathbf{E}\left(|X_t|\mathbf{1}_{\{|X_t|\geq n\}}\right) = 0.$$

Proof of lemma 2.41. [$\Longleftarrow$] Fix $\epsilon > 0$. We have that, uniformly in t,

$$\begin{aligned} \mathbf{E}\left(|X_t|\mathbf{1}_A\right) &= \mathbf{E}\left(|X_t|\mathbf{1}_A\left(\mathbf{1}_{\{|X_t|<n\}} + \mathbf{1}_{\{|X_t|\geq n\}}\right)\right) \\ &\leq n\mathbf{P}(A) + \underbrace{\mathbf{E}\left(|X_t|\mathbf{1}_A\mathbf{1}_{\{|X_t|\geq n\}}\right)}_{<\frac{\epsilon}{2} \text{ for } n \text{ large enough}} \end{aligned}$$

Choose $\delta < \frac{\epsilon}{2n}$. We have that if $\mathbf{P}(A) < \delta$ then for all $t \in T$,

$$\mathbf{E}\left(|X_t|\mathbf{1}_A\right) < n\frac{\epsilon}{2n} + \frac{\epsilon}{2} = \epsilon.$$

[$\Longrightarrow$] Begin by noting that uniform integrability implies

$$\sup_{t\in T} \mathbf{E}|X_t| < \infty. \tag{2.3}$$

To see this, choose $\epsilon = 1$ and find a δ for which (2.2) holds uniformly in t. Then $\sup_{t\in T} \mathbf{E}|X_t| \leq n$ for any n satisfying $1/n < \delta$. Next, by Markov's Inequality and (2.3), we have that there is a constant c for which

$$\mathbf{P}\left(|X_t| \geq n\right) \leq \frac{\mathbf{E}|X_t|}{n} \leq \frac{c}{n} \tag{2.4}$$

for all t. Denote $A_t^n := \{\omega : X_t > n\}$. By uniform integrability, for any $\epsilon > 0$ there exists $\delta > 0$ such that $\mathbf{E}\left(|X_t|\mathbf{1}_{A_t^n}\right) < \epsilon$ whenever $\mathbf{P}(A_t^n) < \delta$. Choosing N such that $c/N < \delta$ we have that $\mathbf{P}(A_t^n) < \delta$ for any $n \geq N$ by (2.4). Thus, for any $\epsilon > 0$ there exists N such that $\sup_{t\in T} \mathbf{E}\left(|X_t|\mathbf{1}_{A_t^n}\right) < \epsilon$ for $n \geq N$. □

Exercise 2.10. If $\{X_t, t \in T\}$ is uniform integrable and $X \in L^1(\mathbb{R})$, then $\{X + X_t, t \in T\}$ is also uniform integrable.

Exercise 2.11. If for some $p > 1$, $\mathbf{E}|X_t|^p < B < \infty$ for all $t \in T$ then $\{X_t, t \in T\}$ is uniform integrable. Example 2.40 shows this is not true for $p = 1$.

Theorem 2.42 (Vitali Convergence Theorem)**.** *Suppose $X_k \xrightarrow{pr} X$ and $\mathbf{E}|X_k| < \infty$ for each $k \in \mathbb{N}$. Then the following are equivalent:*

1. *$\{X_k, k \in \mathbb{N}\}$ is uniformly integrable.*
2. *$X_k \xrightarrow{L^1} X$.*
3. *$\mathbf{E}|X_k| \to \mathbf{E}|X|$.*

In addition, all these imply

4. *$\mathbf{E}X_k \to \mathbf{E}X$*

Proof. Note first that Fatou's Lemma implies $X \in L^1$ (see Exercise 2.6).

[$1 \Longrightarrow 2$] If $Y_k = X_k - X$, then by Exercise 2.10, $\{Y_k, k \in \mathbb{N}\}$ is uniformly integrable. Fix $n \in \mathbb{N}$, so that

$$\limsup_{k\to\infty} \mathbf{E}|Y_k| = \limsup_{k\to\infty} \left(\mathbf{E}\left(|Y_k|\mathbf{1}_{\{|Y_k|<n\}}\right) + \mathbf{E}\left(|Y_k|\mathbf{1}_{\{|Y_k|\geq n\}}\right)\right).$$

The first term on the right goes to zero by the convergence in probability version of the Bounded Convergence Theorem (again, see Exercise 2.6). The second

term on the right is arbitrarily small by Lemma 2.41 and the arbitrary choice of n.

[$2 \Longrightarrow 3$] We have

$$\begin{aligned} |\mathbf{E}\left(|X_k| - |X|\right)| &\leq \mathbf{E}\big||X_k| - |X|\big| \text{ (by Jensen's Inequality)} \\ &\leq \mathbf{E}\,|X_k - X| \to 0. \end{aligned}$$

[$2 \Longrightarrow 4$] This follows since

$$|\mathbf{E}X_k - \mathbf{E}X| \leq \mathbf{E}\,|X_k - X| \to 0.$$

[$3 \Longrightarrow 1$] Without loss of generality $X_k \geq 0$ (or consider X_k^+ and X_k^- separately). Define

$$X_k^{(n)} := X_k \mathbf{1}_{\{X_k \geq n\}} \quad \text{and} \quad X^{(n)} := X \mathbf{1}_{\{X \geq n\}}.$$

By Fatou's Lemma

$$\liminf_{k\to\infty} \mathbf{E}\left(X_k - X_k^{(n)}\right) \geq \mathbf{E}\left(\liminf_{k\to\infty}\left(X_k - X_k^{(n)}\right)\right) = \mathbf{E}X - \mathbf{E}X^{(n)}$$

By the assumption of statement (3),

$$\liminf_{k\to\infty} \mathbf{E}\left(X_k - X_k^{(n)}\right) = \mathbf{E}X - \limsup_{k\to\infty} \mathbf{E}(X_k^{(n)}).$$

Thus

$$\limsup_{k\to\infty} \mathbf{E}(X_k^{(n)}) \leq \mathbf{E}X^{(n)}.$$

Choose n_0 so that $n \geq n_0$ implies $\mathbf{E}X^{(n)} < \frac{\epsilon}{2}$, then choose k_0 such that $\mathbf{E}X_k^{(n)} < \epsilon$ for $k \geq k_0$. Finally choose n_1 so that $\mathbf{E}X_k^{(n)} < \epsilon$ for $1 \leq k \leq k_0$ and $n \geq n_1$, and let $N_\epsilon = \max(n_0, n_1)$. For all k, we have

$$\mathbf{E}\left(X_k \mathbf{1}_{\{X_k \geq N_\epsilon\}}\right) < \epsilon,$$

and by Lemma 2.41 we obtain uniform integrability. □

Exercise 2.12. Give a counterexample to show that even when $X_k \xrightarrow{\text{a.s.}} X$ and $\mathbf{E}|X_k| < \infty$ for each $k \in \mathbb{N}$, statement (4) in Theorem 2.42 does not imply statements (1)-(3).

Now, since an i.i.d. sequence of random variables in L^1 is easily seen to be uniformly integrable, either the SLLN or the WLLN implies the Law of Large Numbers in $L^1(\Omega, \mathcal{F}, \mathbf{P})$. As we mentioned before, in the next chapter, we will prove the SLLN under only a finite first moment condition (in fact, we will prove a much stronger result). But before doing so, let us end this chapter by proving the L^1 LLN under this moment assumption. The method of truncation used in the proof of the following result also has pedagogical purposes since similar methods are used throughout probability theory.

Theorem 2.43 (L^1 Law of Large Numbers, finite 1st moments)**.** *If* $\{X_n, n \in \mathbb{N}\}$ *are i.i.d. and* $\mathbf{E}|X_1| < \infty$*, then*

$$\frac{S_n}{n} \xrightarrow{L^1} \mathbf{E}X_1$$

where $S_n = X_1 + \cdots + X_n$*.*

Proof. Fix $B > 0$. We split up X_k into a bounded random variable $X_k^B := X_k \mathbf{1}_{\{|X_k| \le B\}}$ and a tail random variable $X_k^T := X_k \mathbf{1}_{\{|X_k| > B\}}$ so that $X_k = X_k^B + X_k^T$. We can also split up the mean as follows

$$\mu \equiv \mathbf{E}X_k = \mathbf{E}X_k^B + \mathbf{E}X_k^T \equiv \mu_B + \mu_T.$$

Then

$$\begin{aligned}\frac{1}{n}\mathbf{E}\left|\sum_{k=1}^n (X_k - \mu)\right| &\le \mathbf{E}\left|\frac{1}{n}\sum_{k=1}^n (X_k^B - \mu_B)\right| + \mathbf{E}\left|\frac{1}{n}\sum_{k=1}^n (X_k^T - \mu_T)\right| \\ &\le \left(\mathbf{E}\left|\frac{1}{n}\sum_{k=1}^n (X_k^B - \mu_B)\right|^2\right)^{1/2} + \mathbf{E}\left|\frac{1}{n}\sum_{k=1}^n (X_k^T - \mu_T)\right|.\end{aligned}$$

The first term on the right side goes to zero by following our proof of the WLLN under 2nd moment conditions. The second term on the right is bounded by

$$\frac{2n}{n}\mathbf{E}|X_1^T|.$$

Since our bound B was arbitrary and $X \in L^1$, we can make $\mathbf{E}|X_1^T|$ arbitrarily small by choosing large B. □

Exercise 2.13. Since convergence in L^1 implies convergence in probability, the above theorem also shows that the WLLN holds assuming only the 1st moment condition $\mathbf{E}|X_1| < \infty$. In fact, a version of the WLLN sometimes holds when $\mathbf{E}|X_1| = \infty$. Using truncation show that, when $\{X_i, i \in \mathbb{N}\}$ are i.i.d. with density

$$f_X(x) = \begin{cases} \frac{C}{(1+x^2)\log(1+x^2)} & x \in [-1,1]^c \\ 0 & x \in [-1,1] \end{cases}$$

one has

$$\frac{S_n}{n} \xrightarrow{\text{pr}} 0.$$

Hint: Use a varying truncation level.

3 The Ergodic Theorem

Our goal in this chapter is to prove the celebrated Ergodic Theorem in its a.s. version which is attributed to Birkhoff [Bir31] and sometimes to the pair Birkhoff-Khintchine[5] [Khi33]. As mentioned before, this result applies to certain dependent sequences of random variables and broadly generalizes the Law of Large Numbers.

3.1 Conditional Expectation

"Modern probability theory can be said to begin with the notions of conditioning and disintegration." – O. Kallenberg [Kal02]

Up until now we have dealt mostly with independent random variables. In order to analyze dependent random variables, one of the most useful tools (in addition to covariance and correlation) is the notion of conditional probability which takes a more general form through conditional expectations.

In the sequel, we often have several different σ-fields $\{F_k, k \in \mathbb{Z}\}$, and we will want to know which σ-fields a typical random variable X is measurable with respect to. We use the shorthand $X \in \mathcal{F}_k$ to denote that X is $\mathcal{F}_k$-measurable. Also, throughout this initial discussion of conditional expectation, it may be helpful to think of $\mathcal{F}$ and $\mathcal{F}_k$ as sub-σ-fields of some larger universal σ-field $\mathcal{G}$ for which all random variables are measurable with respect to.

Definition 3.1. If $\mathbf{E}\,|X| < \infty$, we define the *conditional expectation* of X, given the σ-field $\mathcal{F}$, as a random variable $Y \equiv \mathbf{E}(X|\mathcal{F})$ satisfying

(a) $Y \in \mathcal{F}$, i.e., $\sigma(Y) \subset \mathcal{F}$, and

(b) for all $A \in \mathcal{F}$

$$\int_A Y \, d\mathbf{P} = \int_A X \, d\mathbf{P}.$$

In general, there may be many Y satisfying the above, but they are all equal to each other a.s., as we will see. Any such Y is called a *version* of $\mathbf{E}(X|\mathcal{F})$.

We first want to know if such a random variable exists. As a first step, suppose that $X \geq 0$. Define a set function ν on $\mathcal{F}$ by

$$\nu(A) := \int_A X \, d\mathbf{P} \quad \text{for } A \in \mathcal{F}. \tag{3.1}$$

Clearly, ν is an absolutely continuous measure with respect to $\mathbf{P}$. By the Radon-Nikodym Theorem (for general σ-finite measures) there exists a Radon-Nikodym derivative, $\frac{d\nu}{d\mathbf{P}}$, which satisfies for each $A \in \mathcal{F}$

$$\int_A \frac{d\nu}{d\mathbf{P}} \, d\mathbf{P} = \int_A d\nu = \nu(A).$$

[5] Birkhoff proved his theorem for indicator functions, and Khintchine argued a generalized form of the result. It should also be noted that Von Neumann proved the so-called mean ergodic theorem a bit earlier (see [BK32]).

By the definition of ν, this equals $\int_A X\,d\mathbf{P}$. Since the set function in (3.1) is defined only on $\mathcal{F}$, it is not too difficult to see that $\frac{d\nu}{d\mathbf{P}} \in \mathcal{F}$. Thus, we have existence when $X \geq 0$. Now, leaving the details to the reader, one can extend this to general X by writing $X = X^+ - X^-$. Let us also show that $Y \equiv \mathbf{E}(X|\mathcal{F})$ is in $L^1(\Omega)$. Set $A := \{Y \geq 0\} \in \mathcal{F}$ so that

$$\begin{aligned}\mathbf{E}\,|Y| &= \int_A Y\,d\mathbf{P} + \int_{A^c} (-Y)\,d\mathbf{P} \\ &\overset{(b)}{=} \int_A X\,d\mathbf{P} + \int_{A^c} (-X)\,d\mathbf{P} \\ &\leq \mathbf{E}\,|X| \\ &< \infty.\end{aligned}$$

Next, we want to show that the conditional expectation is unique (up to a.s., i.e., up to a *version*). Suppose that Y and Z both satisfy (a) and (b) above. Choose $\epsilon > 0$ and let

$$A := \{Y - Z > \varepsilon\} \in \mathcal{F}.$$

Then, by Markov's Inequality,

$$\int_A Y - Z\,d\mathbf{P} \geq \varepsilon \mathbf{P}(A).$$

But,

$$\begin{aligned}\int_A Y - Z\,d\mathbf{P} &= \int_A Y\,d\mathbf{P} - \int_A Z\,d\mathbf{P} \\ &\overset{(b)}{=} \int_A X\,d\mathbf{P} - \int_A X\,d\mathbf{P} \\ &= 0.\end{aligned}$$

Thus $\mathbf{P}(A) = 0$ and, a.s., $Y \leq Z$. A similar argument shows that, a.s., $Z \leq Y$ from which we conclude $Y \overset{\text{a.s.}}{=} Z$.

Since $\mathbf{E}(X|\mathcal{F})$ is unique (up to a.s.), in consideration of Remark 1.28, we will henceforth use the notation $\mathbf{E}(X|\mathcal{F})$ with the understanding that we are implicitly taking a version of $\mathbf{E}(X|\mathcal{F})$.

Having established existence and uniqueness, let us move on to an example which aims to show that $\mathbf{E}(X|\mathcal{F})$ should be interpreted as our "best guess of X" given $\mathcal{F}$.

Example 3.2. Let $X = X_1 + X_2$ where X_i are i.i.d. random variables with Ber $\left(\frac{1}{2}\right)$ distribution. Then $X \sim \text{Bin}(2, \frac{1}{2})$ so that

$$X = \begin{cases} 0, & \text{with probability } 1/4 \\ 1, & \text{with probability } 1/2 \\ 2, & \text{with probability } 1/4 \end{cases}.$$

(a) Suppose $\mathcal{F}_1 = \{\emptyset, \Omega\}$. We claim that in this case $Y = \mathbf{E}(X|\mathcal{F}_1) \overset{\text{a.s.}}{=} \mathbf{E}X = 1$. We need to check that

$$\int_A Y \, d\mathbf{P} = \int_A X \, d\mathbf{P}$$

for $A = \emptyset$ and $A = \Omega$. Clearly, this holds true for $A = \emptyset$ since each integral is zero. For $A = \Omega$,

$$\int_\Omega 1 \, d\mathbf{P} = 1$$

and

$$\int_\Omega X \, d\mathbf{P} = \frac{1}{4}(0) + \frac{1}{2}(1) + \frac{1}{4}(2) = 1.$$

Now instead of just a.s., let $Y \equiv 1$ (for all ω), then Y is clearly $\mathcal{F}_1$-measurable.

One might ponder, why we cannot set, for some set B with $\mathbf{P}(B) = \frac{1}{2}$,

$$Y = \begin{cases} 2, & B \\ 0, & B^c \end{cases}.$$

This is because $\{Y > 1\}$ must be $\mathcal{F}_1$-measurable, but $\{Y > 1\} = B \notin \mathcal{F}_1$.

(b) Suppose that $\mathcal{F}_2 = \sigma(X)$. We now claim that $Y = \mathbf{E}(X|\mathcal{F}_2) \overset{\text{a.s.}}{=} X$. We can see that for $A \in \sigma(X)$,

$$\int_A Y \, d\mathbf{P} = \int_A X \, d\mathbf{P}$$

since $Y \overset{\text{a.s.}}{=} X$. Letting $Y \equiv X$, we see that $Y \in \mathcal{F}_2$.

(c) Suppose $\mathcal{F}_3 = \sigma(\{X_1 = 1\})$. We claim in this case that

$$Y = \mathbf{E}(X|\mathcal{F}_3) = \begin{cases} 3/2, & \text{on } \{X_1 = 1\} \\ 1/2, & \text{on } \{X_1 = 0\} \end{cases}.$$

We leave this as an exercise for the reader.

Exercise 3.1. When $\mathcal{F}$ is a finite σ-field it is generated by some partition $\pi = \{\pi_1, \ldots, \pi_n\}$. Show that in this case

$$\mathbf{E}(X|\mathcal{F}) = \sum_{k=1}^{n} \left(\int_{\pi_k} X \, d\mathbf{P} \right) \mathbf{1}_{\pi_k}(\omega).$$

Exercise 3.2. Generalizing the idea of (b) in Example 3.2, if X is measurable with respect to $\mathcal{F}$, then $\mathbf{E}(X|\mathcal{F}) \overset{\text{a.s.}}{=} X$.

Our next example illustrates that conditional expectation is a generalization of the "undergraduate" notion of conditional probability:

$$\mathbf{P}(A|B) := \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} \quad \text{whenever } \mathbf{P}(B) > 0.$$

Example 3.3 (Conditional probability). Let $X = \mathbf{1}_A$ and $\mathcal{F}_B := \sigma(B) = \{\emptyset, B, B^c, \Omega\}$. Then the conditional probability of A given B is the random variable

$$Y = \mathbf{P}(A|\mathcal{F}_B) := \mathbf{E}(\mathbf{1}_A|\mathcal{F}_B) = \begin{cases} \mathbf{P}(A|B), & \text{on } B \\ \mathbf{P}(A|B^c), & \text{on } B^c \end{cases}.$$

It is easily seen that integrating over all of Ω gives $\mathbf{E}Y = \mathbf{P}(A)$. Also,

$$\int_B Y \, d\mathbf{P} = \mathbf{P}(A|B)\mathbf{P}(B) = \mathbf{P}(A \cap B)$$

while

$$\begin{aligned} \int_B X \, d\mathbf{P} = \int_B \mathbf{1}_A \, d\mathbf{P} &= \int_\Omega \mathbf{1}_{A\cap B} \, d\mathbf{P} \\ &= \mathbf{P}(A \cap B). \end{aligned}$$

The case for B^c follows by symmetry and the case of the empty set is trivial.

Exercise 3.3. It is trivial that $\mathbf{P}(A|\mathcal{F}) \geq 0$ and $\mathbf{P}(\Omega|\mathcal{F}) = 1$, almost surely. Show also that conditional probability is countably additive[6]:

$$\mathbf{P}\left(\biguplus_{k=1}^{\infty} A_k|\mathcal{F}\right) = \sum_{k=1}^{\infty} \mathbf{P}(A_k|\mathcal{F}).$$

Conditional expectation is, by definition, an action which maps random variables to random variables. However, in the case of a finite σ-field, Exercise 3.1 shows that it can also be thought of as integrating over various subsets of Ω. This idea gives us a heuristic for general σ-fields, and the next exercise illustrates that many of the important properties of regular expectation carry over to conditional expectation.

Exercise 3.4. Prove the following properties of conditional expectation:

(a) (Consistency) $\mathbf{E}(\mathbf{E}(X|\mathcal{F})) = \mathbf{E}(X)$

(b) (Linearity) $\mathbf{E}(aX + bY|\mathcal{F}) \overset{\text{a.s.}}{=} a\mathbf{E}(X|\mathcal{F}) + b\mathbf{E}(Y|\mathcal{F})$.

(c) (Monotonicity) If $X \leq Y$, then a.s., $\mathbf{E}(X|\mathcal{F}) \leq \mathbf{E}(Y|\mathcal{F})$.

(d) (Monotone Convergence) If $X_n \geq 0$ such that $X_n \nearrow X$ with $\mathbf{E}X < \infty$, then

$$\mathbf{E}(X_n|\mathcal{F}) \xrightarrow{\text{a.s.}} \mathbf{E}(X|\mathcal{F}).$$

[6]One can take Exercise 3.3 further and try to find a version of $\mathbf{P}(\cdot|\mathcal{F})$ which, for each fixed $\omega \in \Omega$, is a bona fide probability measure with respect to some σ-field on Ω. These are called *regular conditional probabilities* and they do not always exist (see [Bre92, Sec. 4.3]). However, when they exist, the properties of probability measures and expectation immediately carry over to their conditional counterparts.

Our next result generalizes the idea behind Example 3.2 (b) and Exercise 3.2.

Lemma 3.4 (Factoring out a Random Variable)**.** *If $X \in \mathcal{F}$, $\mathbf{E}|Y| < \infty$, and $\mathbf{E}|XY| < \infty$, then $\mathbf{E}(XY|\mathcal{F}) = X\mathbf{E}(Y|\mathcal{F})$.*

Proof. Let $A, B \in \mathcal{F}$ and set $X = \mathbf{1}_B$. Then,

$$\begin{aligned}\int_A \mathbf{E}(XY|\mathcal{F})d\mathbf{P} &= \int_A \mathbf{1}_B Y d\mathbf{P} \\ &= \int_{A\cap B} Y d\mathbf{P} \\ &= \int_A \mathbf{1}_B \mathbf{E}(Y|\mathcal{F})d\mathbf{P} \\ &= \int_A X\mathbf{E}(Y|\mathcal{F})d\mathbf{P}.\end{aligned}$$

Now, simply extend the above to nonnegative simple functions and then use the Simple Approximation Lemma and Monotone Convergence Theorem to show that the result holds for all nonnegative random variables X and Y with the above assumptions. Finally, extend it to the general case by consider the nonnegative and negative parts of X and Y. □

The usefulness of the above proposition is seen in the proof of a conditional version of Jensen's Inequality.

Theorem 3.5 (Conditional Jensen's Inequality)**.** *Suppose φ is convex and $\mathbf{E}|X|, \mathbf{E}|\varphi(X)| < \infty$. Then a.s.,*

$$\varphi(\mathbf{E}(X|\mathcal{F})) \leq \mathbf{E}(\varphi(X)|\mathcal{F}).$$

Proof. Since we have monotonicity and linearity for conditional expectations, we can use the same idea as in the unconditional version of Jensen's Inequality. Namely, choose a random line $\ell_\omega(x) = A(\omega)x + B(\omega)$, which is $\mathcal{F}$-measurable, satisfying $\ell_\omega(x) \leq \varphi(x)$ a.s. and $\ell_\omega(\mathbf{E}(X|\mathcal{F})) \overset{\text{a.s.}}{=} \varphi(\mathbf{E}(X|\mathcal{F})$. Then by monotonicity and linearity, a.s.,

$$\mathbf{E}(\varphi(X)|\mathcal{F}) \geq A\mathbf{E}(X|\mathcal{F}) + B = \ell(\mathbf{E}(X|\mathcal{F})) = \varphi(\mathbf{E}(X\mathbf{1}_A|\mathcal{F})).$$

□

Corollary 3.6. *Conditional expectation is a (non-strict) contraction in $L^p(\Omega, \mathcal{F}, \mathbf{P})$ for $p \geq 1$, i.e.,*

$$\|\mathbf{E}(X|\mathcal{F})\|_{L^p} \leq \|X\|_{L^p}.$$

Proof. Let $\varphi = |x|^p$ which is convex. By the Conditional Jensen Inequality, we have a.s.

$$|\mathbf{E}(X|\mathcal{F})|^p \leq \mathbf{E}(|X|^p|\mathcal{F}).$$

After taking expectation and the pth root, we have

$$\|\mathbf{E}(X|\mathcal{F})\|_{L^p} \leq (\mathbf{E}\,|X|^p)^{\frac{1}{p}} = \|X\|_{L^p}.$$

□

Another instance of the usefulness of factoring out a random variable is the following result which gives us an alternate interpretation of conditional expectation when second moments are finite.

Proposition 3.7. *If $\mathbf{E}X^2 < \infty$, then $\mathbf{E}(X|\mathcal{F})$ is the random variable $Y \in \mathcal{F}$ which minimizes $\mathbf{E}(X-Y)^2$.*

Proof. As a warmup, let us consider the case where $\mathcal{F} = \{\Omega, \emptyset\}$. We want to minimize

$$\mathbf{E}(X-c)^2 = \mathbf{E}(X^2 - 2xX + c^2) = \mathbf{E}X^2 - 2c\mathbf{E}X + c^2.$$

The c which minimizes the expression is the value which solves

$$\frac{d}{dc}\mathbf{E}(X-c)^2 = -2\mathbf{E}X + 2c = 0,$$

which is $c = \mathbf{E}X$, since the second derivative is positive.

In general, let $Y = \mathbf{E}(X|\mathcal{F}) + Z$ such that $Z \in \mathcal{F}$. Then

$$\begin{aligned}\mathbf{E}(X-Y)^2 &= \mathbf{E}((X - \mathbf{E}(X|\mathcal{F})) - Z)^2 \\ &= \mathbf{E}(X - \mathbf{E}(X|\mathcal{F}))^2 + \mathbf{E}Z^2 - 2\mathbf{E}(Z(X - \mathbf{E}(X|\mathcal{F}))).\end{aligned}$$

By Lemma 3.4 and (a), we have that the last term on the right is zero. Thus, the expression is minimized when $Z = 0$. □

Finally, a third use of factoring gives us a generalization of (a).

Proposition 3.8. *If $\mathcal{F} \subset \mathcal{G}$, then*

$$\mathbf{E}(\mathbf{E}(X|\mathcal{F})|\mathcal{G}) \overset{a.s.}{=} \mathbf{E}(X|\mathcal{F}) \overset{a.s.}{=} \mathbf{E}(\mathbf{E}(X|\mathcal{G})|\mathcal{F}).$$

Proof. To see the first equality, use Exercise 3.2 which states that if $X \in \mathcal{F}$, then $\mathbf{E}(X|\mathcal{F}) = X$. Hence, since $\mathbf{E}(X|\mathcal{F}) \in \mathcal{G}$, we may "factor" it outside of the outer conditional expectation to get that,

$$\mathbf{E}(\mathbf{E}(X|\mathcal{F})|\mathcal{G}) \overset{\text{a.s.}}{=} \mathbf{E}(X|\mathcal{F}) \cdot 1.$$

To see the second equality, we use the definition of conditional expectation twice to see that for any $A \in \mathcal{F}$ (which implies $A \in \mathcal{G}$):

$$\begin{aligned}\int_A \mathbf{E}(\mathbf{E}(X|\mathcal{G})|\mathcal{F})\, d\mathbf{P} &= \int_A \mathbf{E}(X|\mathcal{G})\, d\mathbf{P} \\ &= \int_A X d\mathbf{P}\end{aligned}$$

□

Exercise 3.5. Find an example where

$$\mathbf{E}(\mathbf{E}(Y|\mathcal{F})|\mathcal{G}) \neq \mathbf{E}(\mathbf{E}(Y|\mathcal{G})|\mathcal{F}).$$

3.2 Stationary Sequences

In the sequel we let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

Definition 3.9. Let $X_0, X_1, \ldots$ be a sequence of random variables. We say that $\{X_n, n \geq 0\}$ is a (one-sided) *stationary sequence* if for all $k, m \in \mathbb{N}_0$, the random vectors $(X_0, \ldots, X_m)$ and $(X_k, \ldots, X_{k+m})$ have the same distribution. If $k \in \mathbb{Z}$ and the index set is $\mathbb{Z}$, then it is called a two-sided stationary sequence.

We see that for $m = 0$, this gives that the $X_0, X_1, \ldots$ are identically distributed random variables.

Example 3.10. Suppose $X_0, X_1, \ldots$ are i.i.d. random variables and X is another random variable which also has the same distribution and is independent from this sequence. Then $\{X_n, n \geq 0\}$ is stationary. This is seen by letting μ be the common distribution for each X_i so that

$$\mu_{(X_0,\ldots,X_m)} = \prod_{i=0}^{m} \mu = \mu_{(X_k,\ldots,X_{k+m})}.$$

Also if $Y_n = X$ for all n, then $\{Y_n, n \geq 0\}$ is also stationary. These two examples represent the two extremes of stationary sequences: complete independence and complete dependence. In both examples, the one-dimensional marginal distributions are all the same $X_0 \overset{\mathrm{d}}{=} X_n \overset{\mathrm{d}}{=} Y_0 \overset{\mathrm{d}}{=} Y_n$ for all n, the only difference is in the dependence structure.

One may wonder whether other stationary sequences even exist (or even if generic infinite sequences of independent random variables exist). This is answered by a measure-theoretic result of Kolmogorov which allows us to extend "consistent" finite sequences of random variables to infinite sequences. As a simple example, one may use Kolmogorov's result below to construct stationary Gaussian sequences.

Definition 3.11. If μ_n is a probability distribution on $(\mathbb{R}^n, \mathcal{B}^{\otimes n})$ for each $n \in \mathbb{N}$, then $\{\mu_n\}_{n\in\mathbb{N}}$ is said to be *consistent* if for every $n \in \mathbb{N}$ and Borel sets $B_1, B_2, \ldots \in \mathcal{B}^1$, one has that

$$\mu_{n+1}(B_1 \times \cdots \times B_n \times \mathbb{R}) = \mu_n(B_1 \times \cdots \times B_n).$$

The proof of the following can be found in [Dur10, Appendix A3].

Theorem 3.12 (Kolmogorov's Extension Theorem)**.** *Given a consistent family $\{\mu_n\}_{n\in\mathbb{N}}$ of probability distributions, there exists a unique probability measure $\mathbf{P}$ on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}})$ such that*

$$\mathbf{P}(\{\omega : \omega_i \in B_i \quad \textit{for all } i = 1, \ldots, n\}) = \mu_n(B_1 \times \cdots \times B_n).$$

The μ_n are called the finite-dimensional distributions (fdd's)*.*

Exercise 3.6. If $\{X_n, n \geq 0\}$ and $\{Y_n, n \geq 0\}$ are stationary sequences which are independent from each other, then the sequence formed by $Z_n = aX_n + bY_n$ for all $n \in \mathbb{N}_0$ and constants $a, b \in \mathbb{R}$ is also stationary.

Example 3.13 (Rotation of the circle). Let $\Omega = [0, 1)$, $\mathcal{F} = \mathcal{B}([0, 1))$, and $\mathbf{P} = m$. Also, let $\theta \in (0, 1)$. For $n \in \mathbb{N}_0$, define

$$X_n(\omega) := (\omega + n\theta) (\text{mod } 1)$$

where $x(\text{mod } 1) := x - \lfloor x \rfloor$. For example, $2.91(\text{mod } 1) = 0.91$. We think of $[0, 1)$ as the torus $\mathbb{T}$.

We claim that $\{X_n, n \geq 0\}$ is a stationary sequence. To see this, let $k, m \in \mathbb{N}_0$. It suffices to show that $\mu_{(X_0,\ldots,X_m)} = \mu_{(X_k,\ldots,X_{k+m})}$ holds on a generating set. We will show that

$$\mu_{(X_0,\ldots,X_m)}(B_0 \times \cdots \times B_m) = \mu_{(X_k,\ldots,X_{k+m})}(B_0 \times \cdots \times B_m)$$

where each B_i is a Borel set. We have

$$\begin{aligned}
\mu_{(X_0,\ldots,X_m)}(B_0 \times \cdots \times B_m) &= \mathbf{P}((X_0, \ldots, X_m) \in B_0 \times \cdots \times B_m) \\
&= \mathbf{P}(X_0 \in B_0, \ldots, X_m \in B_m) \\
&= \mathbf{P}(\{\omega : \omega \in B_0, \omega + \theta \in B_1, \ldots, \omega + m\theta \in B_m\}) \\
&= \mathbf{P}(\{B_0 \cap (B_1 - \theta) \cap \cdots \cap (B_m - m\theta)\} - k\theta) \\
&= \mathbf{P}((B_0 - k\theta) \cap (B_1 - (k+1)\theta) \cap \cdots \cap (B_m - (k+m)\theta)) \\
&= \mathbf{P}((X_k, \ldots, X_{k+m}) \in B_0 \times \cdots \times B_m) \\
&= \mu_{(X_k,\ldots,X_{k+m})}(B_0 \times \cdots \times B_m).
\end{aligned}$$

Up until now, we have typically viewed $\{X_n, n \in \mathbb{N}\}$ as a collection or family of random variables and the ordering of the index set was used only for taking limits. Indeed, this is the general setting under which the Law of Large Numbers

applies. However, if one views random variables being recorded successively in time, such as the (percentage) movements of a stock price recorded every five seconds, or the location of an electron randomly wandering in a wire over a period of time (see Example 2.37 for a typical model of these), then one may think of the indices of the random variables as representing time. We use the term **stochastic process** for a collection $\{X_t, t \in T\}$ of random variables indexed by a set of "times", T. In general, T can be of any cardinality, but for the rest of this section let us assume that T is countable, for simplicity. It is in the above context that we will view stationary sequences of random variables. Indeed, the use of the term "stationary" already indicates a time variable lurking in the background. We continue to use vector notation $(X_t, t \in T)$ in place of set notation $\{X_t, t \in T\}$ whenever the ordering of the index is important. We may also sometimes write (X_t) or even just $\vec{X}$ when the set of times T is understood.

Remark 3.14. When dealing with stationary sequences, or more generally stochastic processes, it is often helpful or intuitive to identify the measurable space $(\Omega, \mathcal{F})$ with $(\mathbb{R}^{\mathbb{N}_0}, \mathcal{B}^{\otimes \mathbb{N}_0})$ which is called the *sequence space* (this term also sometimes refers to just the set $\mathbb{R}^{\mathbb{N}_0}$, ignoring the σ-field). The reason for this is that although a sequence of events does not come with a natural algebraic or geometric structure (see Remark 1.53), it does come equipped with a time-ordering. Thus, being able to break down $\omega \in \Omega$ into a sequence $\omega = (\omega_0, \omega_1, \omega_2, \ldots)$ is useful and intuitive. Analogous to setting $X(\omega) = \omega$ in Remark 1.53, we now set $\vec{X} : \Omega \mapsto \Omega$ to be the identity map and X_k is the kth coordinate map, i.e., projection onto the kth coordinate:

$$\begin{aligned} \vec{X}(\omega) &= (X_0(\omega), X_1(\omega), X_2(\omega) \ldots) \\ &=: (\omega_0, \omega_1, \omega_2, \ldots) \\ &= \omega \in \Omega \end{aligned}$$

where Ω is the product space given by $\prod_{k=0}^{\infty} \mathbb{R}$ equipped with the σ-field $\mathcal{B}^{\otimes \mathbb{N}_0}$ generated by sets of the form $B_0 \times \cdots \times B_m \times \mathbb{R} \times \cdots$ (only finitely many coordinates are not equal to $\mathbb{R}$– see the Kolmogorov Extension Theorem above). These sets are called *cylinder sets*. This sort of thinking immediately gives the following proposition.

Proposition 3.15. *If $(X_n, n \geq 0)$ is stationary and $g : \mathbb{R}^{\mathbb{N}_0} \to \mathbb{R}$ is Borel measurable, then $(Y_n, n \geq 0)$ is stationary, where $Y_k := g(X_k, X_{k+1}, \ldots)$.*

Proof. Let $(\Omega, \mathcal{F}) = (\mathbb{R}^{\mathbb{N}_0}, \mathcal{B}^{\otimes \mathbb{N}_0})$ be the sequence space so that X_n just becomes a coordinate map, $X_n(\omega) = \omega_n$. If we have $\omega = (\omega_0, \omega_1, \ldots)$ then let us write $\tilde{\omega} = (\omega_1, \omega_2, \ldots)$ which has the same distribution as ω by the assumption that $(X_n, n \geq 0)$ is stationary. For any $B \in \mathcal{B}^{\otimes \mathbb{N}_0}$ we have

$$\begin{aligned} \mathbf{P}\left((Y_0(\omega), Y_1(\omega), \ldots) \in B\right) &= \mathbf{P}\left((Y_0(\tilde{\omega}), Y_1(\tilde{\omega}), \ldots) \in B\right) \\ &= \mathbf{P}\left((Y_1(\omega), Y_2(\omega), \ldots) \in B\right). \end{aligned}$$

□

Definition 3.16. A measurable map $\varphi : \Omega \to \Omega$, i.e., a map such that $\varphi^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{F}$, is said to be *measure-preserving* if

$$\mathbf{P}(\varphi^{-1}(A)) = \mathbf{P}(A)$$

for all $A \in \mathcal{F}$. Here $\varphi^{-1}(A) := \{\omega : \varphi(\omega) \in A\}$.

Remarks 3.17.

1. A word of warning: the assumption that φ is measure-preserving does not imply that, in the forward direction, $\mathbf{P}(\varphi(A)) = \mathbf{P}(A)$. First of all, by Remark 1.12, it may be that $\varphi(A) \notin \mathcal{F}$ even when $A \in \mathcal{F}$. However, even in the case that $\varphi(A) \in \mathcal{F}$, we may not have $\mathbf{P}(\varphi(A)) = \mathbf{P}(A)$. Consider $([0,1), \mathcal{B}, m)$ and the *doubling map*

   $$\varphi(x) := 2x (\mathrm{mod}\ 1).$$

   If $A = [0, 1/2]$, then $m(\varphi(A)) = 1$. One can also verify that φ is measure-preserving, since

   $$\varphi^{-1}([a,b]) = \left[\frac{a}{2}, \frac{b}{2}\right] \cup \left[\frac{a+1}{2}, \frac{b+1}{2}\right],$$

   and $\mathbf{P}([a,b]) = b - a$, just as $\mathbf{P}(\varphi^{-1}[a,b]) = \frac{b-a}{2} + \frac{(b+1)-(a+1)}{2} = b - a$. On the other hand, if φ is bijective and its inverse is measurable then one may use either the forward or backward direction; however such measure-preserving maps are not so interesting.

2. It is often the case in ergodic theory that one has a measure-preserving map between two different spaces $\varphi : \Omega \mapsto \Omega'$, in which case one requires that

   $$\mathbf{P}(\varphi^{-1}(A)) = \mathbf{P}'(A).$$

   However, in our setting we will always have $\Omega' = \Omega$ and $\mathbf{P}' = \mathbf{P}$.

Example 3.18. If $X \in \mathcal{F}$, then $X_n(\omega) := X(\varphi^n \omega)$ defines a stationary sequence, and the measure-preserving map φ is called a *shift map*. To see this, first note that from φ we get a sequence of measure-preserving maps as follows. Set $\varphi^0(\omega) = \omega$, $\varphi^1(\omega) = \varphi(\omega)$, $\varphi^2(\omega) = \varphi(\varphi(\omega))$, etc. Since φ is measure-preserving, φ^k is measure-preserving for each k:

$$\begin{aligned}
\mathbf{P}(\varphi^{-k}A) &= \mathbf{P}(\varphi^{-1}(\varphi^{-(k-1)}A)) \\
&= \mathbf{P}(\varphi^{-(k-1)}A) \\
&\vdots \\
&= \mathbf{P}(A).
\end{aligned}$$

Now, let $B \in \mathcal{B}^{\otimes n+1}$ and let $A := \{\omega : (X_0(\omega), \ldots, X_n(\omega)) \in B\}$. Then

$$\begin{aligned}
\mu_{(X_k,\ldots,X_{k+n})}(B) &= \mathbf{P}(\{\omega : (X_k(\omega), \ldots, X_{k+n}(\omega)) \in B\}) \\
&= \mathbf{P}(\{\omega : (X(\varphi^k(\omega)), \ldots, X(\varphi^{k+n}(\omega))) \in B\}) \\
&= \mathbf{P}(\{\omega : (X_0(\varphi^k(\omega)), \ldots, X_n(\varphi^k(\omega))) \in B\}) \\
&= \mathbf{P}(\{\omega : \varphi^k(\omega) \in A\}) \\
&= \mathbf{P}(\varphi^{-k}A) \\
&= \mathbf{P}(A) \\
&= \mu_{(X_0,\ldots,X_n)}(B).
\end{aligned}$$

In the special case where $\{X_n, n \geq 0\}$ forms an i.i.d. sequence, φ is called a *Bernoulli shift*[7].

Exercise 3.7. Let $\omega \in [0, 1)$ be given by its dyadic expansion as in Example 1.10. Suppose $\omega = \sum_{k=1}^{\infty} X_k/2^k$, where X_k are i.i.d. Ber $\left(\frac{1}{2}\right)$ random variables. Show that $\mathbf{P}$ is just Lebesgue measure m. Note that the doubling map on $\Omega = [0, 1)$ is precisely a shift map for the sequence $(X_0, X_1, X_2, \ldots)$. When the doubling map is applied to the above, it is also a Bernoulli shift.

Whenever $(Y_n, n \geq 0)$ is stationary, we can construct a probability measure $\mathbf{P}$ on some space such that the sequence $(X_n(\omega) = \omega_n, n \geq 0)$ has the same distribution as $(Y_n, n \geq 0)$. This is done by applying the idea of Remark 3.14 to Example 3.18. Simply let $\Omega := \mathbb{R}^{\mathbb{N}_0}$ and let $\mathbf{P}$ be exactly the distribution $\mu_{\vec{Y}}$ of $\vec{Y}$ on $\mathbb{R}^{\mathbb{N}_0}$ which exists by the Kolmogorov Extension Theorem. Let $\varphi : \Omega \to \Omega$ be the shift map

$$\varphi(\omega_0, \omega_1, \ldots) = (\omega_1, \omega_2, \ldots).$$

Finally set $X(\omega) := \omega_0$ so that $X_n(\omega) = \omega_n = X(\varphi^n\omega)$. In other words, every stationary sequence can be thought of as having marginal distributions of the form $X_n(\omega) = X(\varphi^n\omega)$. Thus

We shall now assume the following throughout the rest of this chapter: $(\Omega, \mathcal{F}, \mathbf{P})$ is our probability space, φ is $\mathbf{P}$-preserving, and for a given random variable X and a shift φ,

$$X_n(\omega) := X(\varphi^n\omega).$$

Definition 3.19. A set $A \in \mathcal{F}$ is said to be φ-*invariant* (or simply invariant when φ is understood) if $A \overset{\text{a.s.}}{=} \varphi^{-1}A$. Again, it is important to use the pullback φ^{-1} instead of the pushforward φ (as noted in Remark 3.17).

Definition 3.20. Given a measure-preserving map φ, we define the *invariant* σ-*field* as

$$\mathcal{I} \equiv \mathcal{I}_\varphi := \{A \in \mathcal{F} : A \overset{\text{a.s.}}{=} \varphi^{-1}A\}.$$

We leave it to the reader to check that $\mathcal{I}$ is indeed a σ-field.

[7]Some authors require also that X_n is a discrete random variable which can take only finitely many values.

Remark 3.21. One has that $X \in \mathcal{I}$ if and only if X is invariant with respect to φ, i.e.,

$$X \circ \varphi \overset{\text{a.s.}}{=} X.$$

To see this, choose $B \in \mathcal{B}$ and set $A := \{X \in B\}$. We break down the equation $A = \varphi^{-1}(A)$ as follows:

$$\begin{aligned}\varphi^{-1}(A) &= \{\omega : X(\varphi(\omega)) \in B\} \\ &\overset{\text{a.s.}}{=} \{\omega : X(\omega) \in B\}, \text{ whenever } X \overset{\text{a.s.}}{=} X \circ \varphi \\ &= A.\end{aligned}$$

Definition 3.22. A measure-preserving map φ is said to be *ergodic* if $\mathcal{I}$ is trivial, i.e., if for all $A \in \mathcal{I}$, one has $\mathbf{P}(A) \in \{0, 1\}$. In particular, if φ is not ergodic, then there exists an $A \in \mathcal{I}$ such that $0 < \mathbf{P}(A) < 1$.

Example 3.23. (Rotation of the circle, part II) Using the same probability space given in Example 3.13 with $\varphi(x) = x + \theta$, it is not hard to see that φ is not ergodic if $\theta \in (0, 1) \cap \mathbb{Q}$. We claim that conversely, if $\theta \in (0, 1) \setminus \mathbb{Q}$, φ is ergodic. To see this, we employ a basic fact from Fourier analysis which says that if $f \in L^2([0, 1))$, then

$$\lim_{n\to\infty} \sum_{k=-n}^{n} c_k e^{2\pi i k x} \xrightarrow{L^2} f(x)$$

where $c_k := \int_0^1 f(x) e^{2\pi i k x}\, dx$. By Exercise 2.8, we may choose a subsequence so that we have a.s. convergence, and we denote this limit by $\sum_{k\in\mathbb{Z}} c_k e^{2\pi i k x}$.

Pick $A \in \mathcal{I}$ and let $f := \mathbf{1}_A$. Then since $f \in \mathcal{I}$, it must be that $f \circ \varphi \overset{\text{a.s.}}{=} f$. We have the following a.s. equivalences:

$$\begin{aligned}\sum_{k\in\mathbb{Z}} c_k e^{2\pi i k x} &= f(x) \\ &= f(\varphi^n(x)) \\ &= \sum_{k\in\mathbb{Z}} c_k e^{2\pi i k (x+n\theta)} \\ &= \sum_{k\in\mathbb{Z}} c_k e^{2\pi i k n\theta} e^{2\pi i k x}.\end{aligned}$$

Hence for each k, $c_k = e^{2\pi i k n\theta} c_k$ for all $n \in \mathbb{N}$ implying that $c_k = 0$ for all $k \neq 0$ in the case that $\theta \in (0, 1) \setminus \mathbb{Q}$. Thus, $f \overset{\text{a.s.}}{=} c_0$, i.e., f is a.s. constant. It follows that $\mathbf{P}(A) \in \{0, 1\}$.

Proposition 3.24. *Let $g : \mathbb{R}^{\mathbb{N}_0} \to \mathbb{R}$ be Borel measurable and $X_n(\omega) := X(\varphi^n\omega)$. If $(X_n, n \geq 0)$ is an ergodic stationary process, then the sequence $(Y_n, n \geq 0)$ is also ergodic, where $Y_k := g(X_k, X_{k+1}, \ldots)$.*

Proof. The proof is similar to the proof of Proposition 3.15. □

3.3 Birkhoff's Ergodic Theorem

Theorem 3.25 (Ergodic Theorem)**.** *If φ is a measure-preserving map, $\varphi : \Omega \to \Omega$, and $X \in L^1(\Omega)$, then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X(\varphi^k(\omega)) \overset{a.s.}{=} \mathbf{E}(X|\mathcal{I}).$$

Remarks 3.26.

1. If $\{X_k, k \in \mathbb{N}\}$ are i.i.d. (where $X_k = X \circ \varphi^k$), then φ is ergodic and $\mathcal{I}$ is trivial. Thus $\mathbf{E}(X|\mathcal{I}) \overset{\text{a.s.}}{=} \mathbf{E}X$ and we get

   $$\frac{1}{n} \sum_{k=1}^{n} X_k(\omega) \xrightarrow{\text{a.s.}} \mathbf{E}X.$$

   This is just a restatement of the Strong Law of Large Numbers, but note that we have only assumed finite first moments.

2. The partial sums

   $$Y_n := \frac{1}{n} \sum_{k=1}^{n} X_k, \quad n \in \mathbb{N},$$

   are uniformly integrable, and hence, converge in $L^1(\Omega)$ and not just almost surely. To see the uniform integrability, note first that the $\{X_k, k \in \mathbb{N}\}$ are identically distributed and thus uniformly integrable. Therefore, given $\epsilon > 0$, we can choose δ such that $\mathbf{E}(|X_k|\mathbf{1}_A) < \epsilon$ whenever $\mathbf{P}(A) < \delta$, uniformly in k. Then by linearity and the triangle inequality,

   $$\begin{aligned} \int_\Omega |Y_n|\mathbf{1}_A d\mathbf{P} &\le \frac{1}{n} \sum_{k=1}^{n} \int_\Omega |X_k|\mathbf{1}_A d\mathbf{P} \\ &< \frac{1}{n} \sum_{k=1}^{n} \epsilon \\ &= \epsilon. \end{aligned}$$

Exercise 3.8. If $X \in L^p(\Omega)$ for $p > 1$, show that convergence in the Ergodic Theorem is in L^p as well. Hint: use the uniform integrability of $\{|X_k|^p, k \in \mathbb{N}\}$.

Before proving the Ergodic Theorem, we introduce the Maximal Ergodic Lemma, which will be used in the proof. This lemma was first proved by Hopf, but the simplified proof used these days is due to Garsia [Gar65].

Lemma 3.27 (Maximal Ergodic Lemma)**.** *Let*

$$S_n(\omega) := \sum_{k=0}^{n-1} \underbrace{X(\varphi^k(\omega))}_{X_k},$$

where φ is a measure-preserving map. If we define $M_n(\omega) := \max(S_1, \ldots, S_n, 0)$, then

$$\mathbf{E}(X\mathbf{1}_{\{M_n>0\}}) \geq 0.$$

Proof. We have for $1 \leq j \leq n$

$$M_n(\omega) \geq S_j(\omega).$$

Since the inequality holds for all ω, we also have for $1 \leq j \leq n$

$$M_n(\varphi(\omega)) \geq S_j(\varphi(\omega)).$$

We can add $X = X_0$ to both sides so that for $1 \leq j \leq n$

$$X(\omega) + M_n(\varphi(\omega)) \geq S_j(\varphi(\omega)) + X(\omega), \quad \text{or}$$

$$X(\omega) + M_n(\varphi(\omega)) \geq S_l(\omega), \quad \text{where } l = j + 1.$$

Hence, subtracting $M_n(\varphi(\omega))$ from both sides,

$$X(\omega) \geq S_l(\omega) - M_n(\varphi(\omega)), \text{ for all } 2 \leq l \leq n + 1.$$

Trivially, $X(\omega) \geq S_1(\omega) - M_n(\varphi(\omega))$ since $S_1 = X$, and $M_n \geq 0$. Hence, letting $A := \{M_n > 0\}$, the following holds for all $1 \leq l \leq n + 1$:

$$\int_A X_1(\omega)\, d\mathbf{P} \geq \int_A (S_l(\omega) - M_n \circ \varphi(\omega))\, d\mathbf{P}.$$

Since the above expression holds for each $S_l, 1 \leq l \leq n + 1$, it is also true for $M_n = \max(S_1, \ldots, S_n, 0)$, and so:

$$\begin{aligned}\int_A X_1(\omega) d\mathbf{P} &\geq \int_A (M_n(\omega) - M_n \circ \varphi(\omega)) d\mathbf{P} \\ &\geq \int_{A \cup \{M_n(\omega)=0\}} (M_n(\omega) - M_n \circ \varphi(\omega)) d\mathbf{P}\end{aligned}$$

But $A \cup \{M_n(\omega) = 0\} = \Omega$ so that, by the stationarity of φ, the right side above equals zero. Hence, we conclude that

$$\mathbf{E}(X\mathbf{1}_A) = \mathbf{E}(X\mathbf{1}_{\{M_n>0\}}) \geq 0.$$

□

Proof of the Ergodic Theorem. First set $X' = X - \mathbf{E}(X|\mathcal{I}) \in L^1$. Then, taking the conditional expectation with respect to $\mathcal{I}$ of both sides,

$$\mathbf{E}(X'|\mathcal{I}) \overset{\text{a.s.}}{=} \mathbf{E}(X|\mathcal{I}) - \mathbf{E}(X|\mathcal{I}) = 0,$$

since $\mathbf{E}(\mathbf{E}(X|\mathcal{I})|\mathcal{I}) \overset{\text{a.s.}}{=} \mathbf{E}(X|\mathcal{I})$. Also note that

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n-1} X' \circ \varphi^k(\omega) &= \frac{1}{n}\sum_{k=1}^{n-1}(X - \mathbf{E}(X|\mathcal{I})) \circ \varphi^k(\omega) \\ &\overset{\text{a.s.}}{=} \frac{1}{n}\sum_{k=1}^{n-1} X \circ \varphi^k(\omega) - \mathbf{E}(X|\mathcal{I}),\end{aligned}$$

since if $Y \in \mathcal{I}$, then $Y \circ \varphi \overset{\text{a.s.}}{=} Y$. Hence, without loss of generality, we can let $\mathbf{E}(X|\mathcal{I}) = 0$.

Now, set $X^* := \limsup_{n\to\infty} \frac{S_n}{n}$, so that $X^* \circ \varphi = X^*$, since

$$\begin{aligned}X^* \circ \varphi &= \limsup_{n\to\infty} \frac{S_n \circ \varphi}{n} \\ &= \limsup_{n\to\infty} \left(\frac{X + X \circ \varphi + \cdots + X \circ \varphi^n}{n} - \frac{X}{n} \right) \\ &= \limsup_{n\to\infty} \left(\frac{S_{n+1}}{n+1} \cdot \frac{n+1}{n} - \frac{X}{n} \right) \\ &= X^*.\end{aligned}$$

Hence, X^* is measurable with respect to $\mathcal{I}$. Now, for a given $\epsilon > 0$, set

$$D_\epsilon := \{X^* > \epsilon\}.$$

If we can show that $\mathbf{P}(D_\epsilon) = 0$ for all $\epsilon > 0$, then a.s. $\limsup_{n\to\infty} \frac{S_n}{n} \leq 0$. By symmetry, the same can be said of $\liminf_{n\to\infty} \frac{S_n}{n}$. Thus we will have shown that a.s. $\lim_{n\to\infty} \frac{S_n}{n} \geq 0$, proving the theorem. Let us now show that $P(D_\epsilon) = 0$. Set

1. $\tilde{X} := (X - \epsilon)\mathbf{1}_{D_\epsilon}$.
2. $\tilde{S}_n := \tilde{X} + \tilde{X} \circ \varphi + \cdots + \tilde{X} \circ \varphi^{n-1}$.
3. $\tilde{M}_n := \max(0, \tilde{S}_1, \ldots, \tilde{S}_n)$.

By the Maximal Ergodic Lemma,

$$\mathbf{E}(\tilde{X}\mathbf{1}_{\{\tilde{M}_n > 0\}}) \geq 0. \tag{3.2}$$

Moreover, since diving by n does not affect the property of being positive,

$$\begin{aligned}\{\tilde{M}_n > 0\} &= \{\max(\tilde{S}_1, \ldots, \tilde{S}_n)/n > 0\} \\ &\nearrow \{\sup_n \frac{\tilde{S}_n}{n} > 0\} \\ &= \{\sup_n \frac{S_n}{n} > \epsilon\} \cap D_\epsilon \qquad \text{(by the definition of } \tilde{X}) \\ &= \{\sup_n \frac{S_n}{n} > \epsilon\} \cap \{\limsup_{n\to\infty} \frac{S_n}{n} > \epsilon\}\end{aligned}$$

where the last line equals D_ϵ since

$$\{\limsup_{n\to\infty} \frac{S_n}{n} > \epsilon\} \subset \{\sup_n \frac{S_n}{n} > \epsilon\}.$$

We conclude from the above that $\{\tilde{M}_n > 0\} \nearrow D_\epsilon$. Now, by the Dominated Convergence Theorem,

$$\mathbf{E}(\tilde{X}\mathbf{1}_{\{\tilde{M}_n>0\}}) \to \mathbf{E}(\tilde{X}\mathbf{1}_{D_\epsilon}),$$

since $|\tilde{X}| \leq |X| + \epsilon$. Recalling that these expectations are nonnegative by (3.2), we have

$$\begin{aligned} 0 &\leq \mathbf{E}(\tilde{X}\mathbf{1}_{D_\epsilon}) \\ &= \mathbf{E}((X-\epsilon)\mathbf{1}_{D_\epsilon}) \\ &= \mathbf{E}(X\mathbf{1}_{D_\epsilon}) - \epsilon\mathbf{E}(\mathbf{1}_{D_\epsilon}) \\ &= \mathbf{E}(X\mathbf{1}_{D_\epsilon}) - \epsilon\mathbf{P}(D_\epsilon) \end{aligned}$$

Also, note that $\mathbf{E}(X\mathbf{1}_{D_\epsilon}) = 0$, since $X \in \mathcal{I}$ and $D_\epsilon \in \mathcal{I}$ implies

$$\begin{aligned} \mathbf{E}(X\mathbf{1}_{D_\epsilon}) &= \int_{D_\epsilon} X d\mathbf{P} \\ &= \int_{D_\epsilon} \mathbf{E}(X|\mathcal{I}) d\mathbf{P} \\ &= 0 \end{aligned}$$

where the last line follows by the assumption that $\mathbf{E}(X|\mathcal{I}) = 0$. Thus,

$$0 \leq -\epsilon\mathbf{P}(D_\epsilon)$$

which implies $\mathbf{P}(D_\epsilon) = 0$ as required. □

4 The Central Limit Theorem

For a Simple Random Walk S_n with drift $p - \frac{1}{2}$ (see Example 2.37), we know by the Strong Law of Large Numbers that

$$\frac{S_n - n(2p-1)}{n} \xrightarrow{\text{a.s.}} 0$$

which gives $n\mathbf{E}X_1 = n(2p-1)$ as a first order approximation to S_n. In particular, $S_n - n\mathbf{E}X_1$ is $o(n)$, i.e., grows slower than order n, so any second order approximation should identify the growth rate of $S_n - n\mathbf{E}X_1$. The Central Limit Theorem[8] (CLT), recognized as the most important result in probability, identifies this growth rate as order $\sqrt{n}$ and in fact, gives the following second order approximation for any $x \in \mathbb{R}$:

$$\lim_{n\to\infty} \mathbf{P}\left(Z_n \le x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-u^2}{2}}\, du, \quad Z_n := \frac{S_n - n(2p-1)}{2\sqrt{np(1-p)}}. \tag{4.1}$$

The above statement of the CLT (to be proved later) is the sort of version one would see in an undergraduate probability text, but it raises the question, "in what sense does $(Z_n, n \in \mathbb{N})$ converge to some random variable?" It turns out that $(Z_n, n \in \mathbb{N})$ does not converge in probability and therefore cannot converge a.s or in $L^p(\Omega), p \ge 1$. We need a different type of convergence to describe (4.1). Since

$$F_{Z_n}(x) = \mathbf{P}(Z_n \le x), \tag{4.2}$$

any such convergence should be akin to pointwise convergence of distribution functions.

4.1 Convergence in Distribution

Definition 4.1.

(i) We say a sequence of distributions $(\mu_n, n \in \mathbb{N})$ on $(\mathbb{R}, \mathcal{B})$ *converge weakly* to μ if for all bounded, continuous functions f:

$$\lim_{n\to\infty} \int_{\mathbb{R}} f d\mu_n = \int_{\mathbb{R}} f d\mu.$$

We use the notation $\mu_n \Rightarrow \mu$ for weak convergence of distributions.

(ii) We say that X_n *converges in distribution* to X, denoted $X_n \xrightarrow{\text{d}} X$, if the distributions $(\mu_{X_n}, n \in \mathbb{N})$ converges weakly to μ_X, the distribution of X, i.e., $\mathbf{E}f(X_n) \to \mathbf{E}f(X)$ for all bounded, continuous functions f on $\mathbb{R}$.

[8] The name "Central Limit Theorem" is due to Pólya in [Pól20]. Its history however dates back to De Moivre in 1733 who proved the Gaussian approximation to the binomial. Indeed, he was trying to find a refinement of Jakob Bernoulli's Law of Large Numbers proved earlier in 1710. The first "universality" result beyond just binomial random variables is attributed to Laplace in 1812 (the proof used generating functions). A nice historical account can be found in [Fis10].

Remark 4.2. The above definition lends itself to an immediate version of the Continuous Mapping Theorem (see Exercise 2.7). Assuming $h : \mathbb{R} \to \mathbb{R}$ is continuous, if $X_n \xrightarrow{\mathrm{d}} X$, then $h(X_n) \xrightarrow{\mathrm{d}} h(X)$. This follows since $f(h(x))$ is bounded and continuous whenever f is bounded and continuous, thus

$$\mathbf{E}(f \circ h(X_n)) \to \mathbf{E}(f \circ h(X)).$$

Example 4.3. Let $\mu_n = \delta_{x_n}$ and $\mu = \delta_x$ be point masses. Then, $\mu_n \Rightarrow \mu$ if and only if $x_n \to x$. This follows since we have for all bounded and continuous f that $\int f d\delta_{x_n} = f(x_n)$ converges to $\int f d\delta_x = f(x)$.

Example 4.4. Let

$$\mu_n = \sum_{k=1}^{n} \frac{1}{n} \delta_{\frac{k}{n}}$$

and let m be Lebesgue measure on $[0, 1]$. We have $\mu_n \Rightarrow m$ since if f is continuous, then

$$\int_0^1 f \, d\mu_n = \frac{1}{n} \sum_{k=1}^{n} f\left(\frac{k}{n}\right) \to \int_0^1 f(x) \, dx.$$

Next we note that convergence in distribution is weaker than convergence in probability, a.s., or L^p, as seen by the following exercise.

Exercise 4.1. Show that if $X_n \xrightarrow{\mathrm{pr}} X$, then $X_n \xrightarrow{\mathrm{d}} X$. Conversely, show that if C is a constant and $X_n \xrightarrow{\mathrm{d}} C$, then $X_n \xrightarrow{\mathrm{pr}} C$.

Our next result verifies that convergence in distribution is akin to pointwise convergence of distribution functions as remarked below (4.2).

Theorem 4.5 (Convergence in distribution characterization)**.** *$X_n \xrightarrow{d} X$ if and only if $F_{X_n}(x) \to F_X(x)$ at all $x \in \mathbb{R}$ which are continuity points of F_X.*

Remark 4.6. F_X is uniquely determined by values at continuity points since it is right-continuous by Proposition 1.60.

Lemma 4.7 (Skorokhod Representation Lemma)**.** *If $F_{X_n}(x) \to F_X(x)$ at all continuity points of F_X, then there exist $Y_n \overset{d}{=} X_n$, for all $n \in \mathbb{N}$, and $Y \overset{d}{=} X$ such that $Y_n \xrightarrow{a.s.} Y$.*

Proof. Similar to Theorem 1.61, let $(\Omega, \mathcal{F}, \mathbf{P}) = ([0, 1], \mathcal{B}([0, 1]), m)$ and for all $k \in \mathbb{N} \cup \{\infty\}$ let

$$Y_k(\omega) := \sup \{x : F_{X_k}(x) < \omega\}$$

and

$$Y(\omega) := \sup \{x : F_X(x) < \omega\}.$$

When F_{X_k} is continuous and strictly increasing, Y_k is the inverse of F_{X_k}, and this is how one should view it even when F_{X_k} is discontinuous or not strictly

increasing. By arguments in the proof of Theorem 1.61, the distribution function of Y_k is F_{X_k} and similarly for Y and F_X.

The proof proceeds by seeing that Y_n converges to Y at all of Y's continuity points, which in general is a different set than the continuity points of F_Y. But since Y is nevertheless nondecreasing, there are only countably many discontinuities which suffices to prove the result. We need only convert our assumption about convergence at continuity points of F_Y into convergence at continuity points of Y. Let Ω_c be the continuous points of Y (its complement, the discontinuities, consists of any height ω of $F_Y(x)$ corresponding to a flat interval where $F_Y' = 0$). To see that convergence occurs at any $\omega \in \Omega_c$, it is enough that

$$\limsup Y_n(\omega) \leq Y(\omega) \leq \liminf Y_n(\omega) \quad \text{for } \omega \in \Omega_c. \tag{4.3}$$

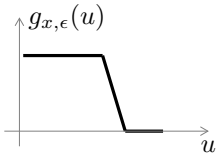
To see this, for $\omega \in \Omega_c$, pick $x_0 < Y(\omega)$ so that F_Y is continuous at x_0. Then by assumption

$$\lim_{n\to\infty} F_{Y_n}(x_0) = F_Y(x_0).$$

But we also know that $F_Y(x_0) < \omega$ since x_0 is not in a flat interval of F_Y, thus $x_0 \leq \liminf Y_n(\omega)$ by the definition of the $Y_n(\omega)$. Since this is true for all continuity points $x_0 < Y(\omega)$ (which is a dense set), it is also true for $Y(\omega)$. The other inequality in (4.3) is proved similarly. □

Proof of Theorem 4.5. $\Rightarrow$: Define a family of bounded and continuous functions

$$g_{x,\epsilon}(u) = \begin{cases} 1, & u \leq x \\ 0, & u \geq x + \epsilon \\ (x + \epsilon - u)/\epsilon, & x < u < x + \epsilon \end{cases}$$



so that

$$\limsup_{n\to\infty} \mathbf{P}(X_n \leq x) \leq \limsup_{n\to\infty} \mathbf{E} g_{x,\epsilon}(X_n) = \mathbf{E} g_{x,\epsilon}(X) \leq \mathbf{P}(X \leq x + \epsilon).$$

Letting $\epsilon \to 0$, by the continuity of measure we have

$$\limsup_{n\to\infty} \mathbf{P}(X_n \leq x) \leq \mathbf{P}(X \leq x).$$

At the same time,

$$\liminf_{n\to\infty} \mathbf{P}(X_n \leq x) \geq \liminf_{n\to\infty} \mathbf{E} g_{x-\epsilon,\epsilon}(X_n) \geq \mathbf{P}(X \leq x - \epsilon),$$

which in the limit as $\epsilon \to \infty$ gives, by the right-continuity of distribution functions,

$$\liminf_{n\to\infty} \mathbf{P}(X_n \le x) \ge \mathbf{P}(X < x).$$

$\Leftarrow$: By the preceding lemma we choose $Y_n \overset{\mathrm{d}}{=} X_n$ and $Y \overset{\mathrm{d}}{=} X$ such that $Y_n \xrightarrow{\text{a.s.}} Y$ and also $g(Y_n) \xrightarrow{\text{a.s.}} g(Y)$ for continuous g by Exercise 2.7. If g is a bounded continuous function, by Bounded Convergence Theorem

$$\mathbf{E}g(X_n) = \mathbf{E}g(Y_n) \to \mathbf{E}g(Y) = \mathbf{E}g(X).$$

□

Definition 4.8.

(i) A collection π of distributions on $\mathbb{R}$ is *relatively compact* (sequentially) if for every sequence $(\mu_n, n \in \mathbb{N}) \subset \pi$ there exists a subsequence $(\mu_{n_k}, k \in \mathbb{N})$ that converges weakly to some (probability) distribution which does not necessarily have to be in π.

(ii) π is *tight* if for any $\epsilon > 0$ there exists N such that $\mu([-N, N]) > 1 - \epsilon$ for all $\mu \in \pi$.

Theorem 4.9 (Prokhorov's Theorem)**.** *A family of distributions π is relatively compact if and only if π is tight.*

Proof. $\Leftarrow$: Suppose π is tight. Given $(\mu_n) \subset \pi$ we need to find (μ_{n_k}) converging weakly. Let F_n be the distribution function for μ_n and let (q_i) be an enumeration of $\mathbb{Q}$. For each set of numbers $Q_i = \{q_1, \ldots, q_i\}$, since the range $[0, 1]$ of any distribution function is compact, we can find a subsequence $(n_k^{(i)}, k \in \mathbb{N})$ such that $(F_{n_k^{(i)}})$ converges at each point of Q_i. It is also not hard to see

$$F_{n_k^{(i)}}(q_m) \le F_{n_k^{(i)}}(q_n) \quad \text{whenever} \quad q_m < q_n. \tag{4.4}$$

Using diagonalization set $(n_k) := (n_k{}^{(k)})$ so that (F_{n_k}) converges for all $q \in \mathbb{Q}$ and also preserves order in the sense of (4.4).

For each $q \in \mathbb{Q}$, set $G(q) := \lim_{k\to\infty} F_{n_k}(q)$ which is non-decreasing, but only defined on $\mathbb{Q}$. For general $x \in \mathbb{R}$, we extend using right-continuity $F(x) := \lim_{q \searrow x} G(q)$. By construction, for any $\epsilon > 0$ we can find N such that

$$F(N) - F(-N) = \lim_{k\to\infty} (F_{n_k}(N) - F_{n_k}(-N)) > 1 - \epsilon$$

which shows $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(s) = 1$. Thus F is a distribution function by Theorem 1.61. To complete the proof it is enough to show that $F_{n_k}(x) \to F(x)$ for all x which are continuity points of $F(x)$.

Take x, a continuity point of $F(x)$. Find two rationals q_1, q_2 for which $q_1 < x < q_2$. Since F is nondecreasing

$$F(q_1) \le \liminf F_{n_k}(x) \le \limsup F_{n_k}(x) \le F(q_2).$$

Since $F(x)$ is continuous at x, we can make $F(q_2) - F(q_1)$ arbitrarily small, thus $\lim F_{n_k}(x) = F(x)$.

$\Rightarrow$: Proof by contradiction: suppose that π is relatively compact. If π is not tight, then there exist $\epsilon > 0$ and $(\mu_n) \subset \pi$ with $\mu_n[-n, n] \leq 1 - \epsilon$ for all $n \in \mathbb{N}$. By relative compactness we can find (n_k) such that $\mu_{n_k} \implies \mu$. But notice that $\mu(a, b) \leq \liminf \mu_{n_k}(a, b) \leq 1 - \epsilon$ for any a, b, which cannot be if μ is a probability distribution. □

Prokhorov's Theorem is only useful if there is an easy way of checking tightness. Fortunately, there is.

Proposition 4.10 (Tightness criterion)**.** *If for some $p > 0$, $\sup_{t \in T} \mathbf{E}|X_t|^p < \infty$, then the family $\{X_t, t \in T\}$ is tight.*

Proof. We simply use Chebychev's Inequality to see that

$$\mathbf{P}(|X_t| \geq N) \leq \frac{\mathbf{E}|X_n|^p}{N^p} \leq \frac{c}{N^p}.$$

□

Example 4.11. Let $\{X_n, n \in \mathbb{N}\}$ be i.i.d. random variables with $\mathbf{E}X_1 = 0$ and $\mathbf{E}X_1^2 = 1$. As usual denote $S_n := \sum_{i=1}^n X_i$. Then

$$\mathbf{E}\left(\frac{S_n}{\sqrt{n}}\right)^2 = \frac{1}{n}\mathbf{E}S_n^2 = \frac{1}{n}n\mathbf{E}X_1^2 = 1,$$

implying that some subsequence of $(S_n, n \in \mathbb{N})$ must converge in distribution.

4.2 Characteristic Functions

We would next like a method for checking whether a sequence of random variables $(X_k, k \in \mathbb{N})$ converges in distribution. Historically, the first step in this direction is the so-called "moment method" which tells us that if all the moments converge,

$$\mathbf{E}X_n^k \xrightarrow{n \to \infty} M_k, \ k \in \mathbb{N},$$

then under some regularity conditions (see Section (4.4)), the sequence of moments (M_k) uniquely determines a distribution.

We will bypass the moment method for now and use another popular, and perhaps more powerful, way of checking convergence in distribution which is via characteristic functions[9] which are basically Fourier transforms of probability measures. In particular, the characteristic function of a random variable X, $\varphi_X(t)$, determines $F_X(x)$ and μ_X. Thus we will now have four different objects φ_X, F_X, μ_X, and (if it exists) f_X which determine the distribution of X. While characteristic functions are not as intuitive to use as moments, they are much easier to use, and over time one gets used to them.

[9]These should not be confused with the characteristic functions of real analysis, which we have been calling indicator functions above.

Definition 4.12. The *characteristic function* of a random variable X is

$$\varphi_X(t) := \mathbf{E}e^{itX} = \int_{\mathbb{R}} e^{itx}\mu_X(dx).$$

If a density exists, then $\varphi_X(t) = \int_{\mathbb{R}} e^{itx} f_X(x)\,dx$ is (up to a constant) the Fourier transform of $f_X(x)$.

Remark 4.13. It is easy to check that the following properties hold for any characteristic function φ_X:

1. $|\varphi_X(t)| \le \mathbf{E}|e^{itX}| = 1$.
2. $\varphi_X(0) = 1$.
3. $\varphi_X(-t) = \overline{\varphi(t)}$.
4. $\varphi_X(t)$ is uniformly continuous:

$$|\varphi_X(t+h) - \varphi_X(t)| \le \mathbf{E}|e^{itX}(e^{ihX} - 1)| = \mathbf{E}|e^{ihX} - 1| \xrightarrow{h\to 0} 0.$$

5. If X and Y are independent, then

$$\varphi_{X+Y}(t) = \mathbf{E}e^{it(X+Y)} = \mathbf{E}e^{itX}\mathbf{E}e^{itY} = \varphi_X(t)\varphi_Y(t).$$

6. $\varphi_{-X}(t) = \varphi_X(-t) = \overline{\varphi_X(t)}$, therefore if X is symmetric, i.e., $X \overset{\mathrm{d}}{=} -X$, then $\varphi_X(t)$ is real-valued.
7. $|\varphi_X|^2 = \varphi_X\overline{\varphi_X} = \varphi_X\varphi_{-X} = \varphi_{X+Y}$ where $Y \overset{\mathrm{d}}{=} -X$ and $Y \perp\!\!\!\perp X$.

Example 4.14. We compute the characteristic functions for some common distributions.

(a) $X \sim \mathrm{Rad}\left(\frac{1}{2}\right)$, that is $\mathbf{P}(X = -1) = \mathbf{P}(X = 1) = \frac{1}{2}$.

$$\varphi_X(t) = \mathbf{E}e^{itX} = \frac{1}{2}e^{it} + \frac{1}{2}e^{-it} = \cos t.$$

(b) $X \sim \mathrm{Poiss}(\lambda)$, that is $\mathbf{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k \ge 0$ (See).

$$\varphi_X(t) = \sum \frac{\lambda^k}{e^\lambda k!}e^{itk} = \frac{e^{\lambda e^{it}}}{e^\lambda} = e^{\lambda(e^{it}-1)}.$$

(c) $X \sim \mathrm{Exp}(\lambda)$, that is $f_X(x) = \lambda e^{-\lambda x}\mathbf{1}_{[0,\infty)}$.

$$\varphi_X(t) = \int_{\mathbb{R}} f_X(x)e^{itx}\,dx = \int_0^\infty \lambda e^{(-\lambda+it)x}\,dx = \frac{\lambda}{\lambda - it}.$$

(d) $X \sim \text{Unif}\,[-K, K]$, that is $f_X(x) = \frac{1}{2K}\mathbf{1}_{[-K,K]}(x)$.

$$\varphi_X(t) = \int_{-K}^{K} \frac{1}{2K} e^{itx}\, dx = \frac{e^{itK} - e^{-itK}}{2Kit} = \frac{\sin Kt}{Kt}.$$

(e) $X \sim N(0,1)$, $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

$$\varphi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{x^2}{2} + itx + \frac{t^2}{2}} e^{-\frac{t^2}{2}}\, dx = e^{-\frac{t^2}{2}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-it)^2}{2}}\, dx = e^{-\frac{t^2}{2}}.$$

where the last equality is obvious when t is pure imaginary, but needs work to prove it in general (we leave this to the reader in the next exercise).

Exercise 4.2.

(a) Show that $\varphi_X(t) = e^{-\frac{t^2}{2}}$ when $X \sim N(0,1)$.

(b) Suppose Z is the sum of two independent Unif$[-\frac{1}{2}, \frac{1}{2}]$ random variables so that it has density $f_Z(z) = \begin{cases} 1 - |x| & x \in [-1,1] \\ 0 & \text{otherwise} \end{cases}$.

Show that $\varphi_Z(z) = \frac{4\sin^2 \frac{t}{2}}{t^2} = \frac{2(1-\cos t)}{t^2}$.

(c) Find $\varphi_X(t)$ when $X \sim \text{Bin}(n,p)$.

Recall the Fourier Inversion Formula from real analysis: If both f and its Fourier transform $\hat{f}(t) := \int_{\mathbb{R}} e^{-2\pi itx} f(x)\, dx$ are in $L^1(\mathbb{R})$ then

$$\int_{\mathbb{R}} e^{2\pi itx} \hat{f}(t)\, dt \overset{\text{a.e.}}{=} f(x).$$

We have the following analog for characteristic functions, which importantly tells us that a distribution μ_X is determined by its characteristic function φ_X.

Theorem 4.15 (Lévy's Inversion Formula)**.** *If $a < b$ and φ is the characteristic function of μ, then*

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

Remarks 4.16.

1. We might be worried about a singularity at $t = 0$, but note that the integral kernel satisfies

$$\left|\frac{e^{-ita} - e^{-itb}}{it}\right| = \left|\int_a^b e^{-itx}\, dx\right| \le \int_a^b 1\, dx = b - a$$

2. We will need the Dirichlet Integral

$$\lim_{T\to\infty}\int_{-T}^{T}\frac{\sin\theta t}{t}\,dt=\begin{cases}\pi & \theta>0\\ -\pi & \theta<0\\ 0 & \theta=0.\end{cases}$$

Proof. By the definition of a characteristic function

$$\int_{-T}^{T}\varphi(t)\left[\frac{e^{-ita}-e^{-itb}}{it}\right]\,dt\;=\int_{-T}^{T}\int_{\mathbb{R}}e^{itx}\,\mu(dx)\left[\frac{e^{-ita}-e^{-itb}}{it}\right]\,dt.$$

The integrand above is bounded, thus we may apply Fubini's theorem to see that the above equals

$$\begin{aligned}&\int_{\mathbb{R}}\int_{-T}^{T}\frac{e^{it(x-a)}-e^{it(x-b)}}{it}\,dt\,\mu(dx)\\ =\;&\int_{\mathbb{R}}\int_{-T}^{T}\frac{[\cos((x-a)t)+i\sin((x-a)t)]}{it}\\ &\qquad-\frac{[\cos((x-b)t)+i\sin((x-b)t)]}{it}\,dt\,\mu(dx).\end{aligned}$$

The function

$$g(t)=\frac{\cos((x-a)t)-\cos((x-b)t)}{t}$$

is odd and goes to 0 as $t\to 0$ thus, the above is equal to

$$\begin{aligned}&\int_{\mathbb{R}}\left[\int_{-T}^{T}\frac{\sin(t(x-a))}{t}\,dt-\int_{-T}^{T}\frac{\sin(t(x-b))}{t}\,dt\right]\,\mu(dx)\\ =\;&\int_{\mathbb{R}}g(x)\,\mu(dx).\end{aligned}$$

Taking the limit as $T\to\infty$ we obtain by the Dirichlet Integral:

$$g(x)=\begin{cases}2\pi & a<x<b\\ \pi & x=a\;\text{ or }\;x=b\\ 0 & \text{otherwise}\end{cases}.$$

Therefore

$$\int_{\mathbb{R}}g(x)\mu(dx)=2\pi\mu((a,b))+\pi\mu(\{a,b\}).$$

□

Corollary 4.17.

(a) If $\varphi_X=\varphi_Y$, then $X\overset{d}{=}Y$.

(b) If φ_X is real-valued, then μ_X is symmetric, i.e.,

$$\mu((-b,-a)) = \mu((a,b)) \quad \textit{for all } a < b.$$

(c) If $\varphi_X(t) \in L^1(\mathbb{R})$, then X has a density $f_X(x)$ which is continuous.

Proof. Part (a) is obvious from the Inversion Formula. Part (b) follows from item (6) in Remark 4.13. For part (c), we have

$$\mu_X((a,b)) + \frac{1}{2}\mu_X(\{a,b\}) = \frac{1}{2\pi}\int_{\mathbb{R}} \varphi_X(t)\left(\int_a^b e^{-itx}dx\right)dt.$$

Since the function $(x,t) \mapsto \varphi(t)e^{-itx}$ is in $L^1([a,b]\times\mathbb{R})$, we may apply Fubini's theorem to obtain

$$= \int_a^b \underbrace{\frac{1}{2\pi}\int_{\mathbb{R}} \varphi_X(t)e^{-itx}dt}_{=:f_X(x)}\ dx,$$

where $f_X(x)$ is continuous by an argument similar to item (4) in Remark 4.13. □

We have already discussed the relationship between pointwise convergence $F_{X_n} \to F_X$ and convergence in distribution $X_n \xrightarrow{d} X$ via Theorem 4.5. We now want to know the relationship between pointwise convergence of characteristic functions $\varphi_{X_n} \to \varphi_X$ and $X_n \xrightarrow{d} X$.

Theorem 4.18 (Lévy's Continuity Theorem[10])**.** *If $X_n \xrightarrow{d} X$, then $\varphi_{X_n}(t) \to \varphi_X(t)$ for all $t \in \mathbb{R}$. Conversely, suppose (φ_n) are characteristic functions for the distributions (μ_n) and $\varphi_n(t) \to \varphi(t)$ for all $t \in \mathbb{R}$, where $\varphi(t)$ is some complex-valued function. If $\varphi(t)$ is continuous at 0, then*

(a) φ is the characteristic function for some distribution μ,

(b) $\mu_n \Rightarrow \mu$.

To prove Lévy's Continuity Theorem, we will need:

Lemma 4.19. *Let φ be the characteristic function for the distribution μ. Then for all $A > 0$, we have*

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

[10]This was proved in [Lév25].

Proof. By Fubini's theorem,

$$\begin{aligned}\frac{1}{2}\left|\int_{-T}^{T}\int_{\mathbb{R}} e^{itx}\mu(dx)dt\right| &= \left|\int_{\mathbb{R}}\frac{1}{2}\int_{-T}^{T} e^{itx}dt\ \mu(dx)\right| \\ &\leq \int_{\mathbb{R}}\frac{1}{2}\left|\int_{-T}^{T}\cos(xt)dt\right|\ \mu(dx) \\ &= \int_{\mathbb{R}}\frac{T}{T}\left|\frac{\sin(xT)}{x}\right|\mu(dx).\end{aligned}$$

Now for all $u \in \mathbb{R}$, $\left|\frac{\sin(u)}{u}\right| \leq 1$, and for $|u| \geq C$, $\left|\frac{\sin(u)}{u}\right| \leq \frac{1}{C}$. The right side above can be broken into

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

Setting $T = 1/A$ we get that the right side is bounded by $\frac{1}{A}(\frac{1}{2}\mu[-2A,2A] + \frac{1}{2})$ which is what was required. □

Proof of Theorem 4.18. For the first statement of the theorem, note that $X_n \xrightarrow{d} X$ means $\mu_{X_n} \Rightarrow \mu_X$ which in turn means $\int f\ d\mu_n \to \int f\ d\mu$ for all bounded and continuous f, thus we consider $f_t(x) = e^{itx}$ to get pointwise characteristic function convergence.

For the converse, first note that by Lemma 4.19, we have for every $n \in \mathbb{N}$,

$$\mu_n[-2A,2A] \geq A\left|\int_{-1/A}^{1/A}\varphi_n(t)dt\right| - 1$$

so that

$$\liminf_{n\to\infty}\mu_n[-2A,2A] \geq A\left|\int_{-1/A}^{1/A}\varphi(t)dt\right| - 1$$

using the Dominated Convergence Theorem. Then by continuity of φ at 0,

$$\lim_{A\to\infty}\liminf_{n\to\infty}\mu_n[-2A,2A] \geq 2\underbrace{\lim_{A\to\infty}\frac{1}{\frac{2}{A}}\left|\int_{-1/A}^{1/A}\varphi(t)dt\right|}_{=\varphi(0)=1} - 1 = 1.$$

which shows that the family $\{\mu_n, n \in \mathbb{N}\}$ is tight. By Prokhorov's Theorem, if we choose a sequence $(\mu_{n_k}, k \in \mathbb{N})$, then there is a further subsequence $(\mu_{n_{k_\ell}}, \ell \in \mathbb{N})$

that converges weakly to some distribution μ_0 possibly depending on the subsequences (n_k) and (n_{k_ℓ}). By the first part of the theorem, the characteristic function of the distribution μ_0 is

$$\varphi_0(t) := \lim_{\ell\to\infty} \varphi_{n_{k_\ell}}(t).$$

But by assumption

$$\varphi(t) = \lim_{n\to\infty} \varphi_n(t).$$

which shows that the limiting distribution μ_0, associated to any subsequence, always has the characteristic function $\varphi(t)$ and thus does not depend on our choice of (n_k) or (n_{k_ℓ}). □

Remark 4.20. The intuition gathered from the Continuity Theorem is that the tail behavior of μ, i.e., the mass of μ near infinity, is controlled by how nicely behaved its characteristic function φ is around the origin. The next example illustrates this from the side of bad behavior of φ at 0.

Example 4.21. Let

$$\mu_n = \begin{cases} \frac{1}{2n}dx & \text{on } [-n,n] \\ 0 & \text{on } [-n,n]^c \end{cases}$$

The characteristic function of μ_n is

$$\varphi_n(t) = \frac{\sin(nt)}{nt}, t \neq 0$$

which goes to 0 as $n \to \infty$ for all $t \neq 0$, while on the other hand $\varphi_n(0) = 1$ for all n. Mass is lost to infinity since the pointwise limiting function of the φ_n is not continuous at 0 and therefore does not qualify as a characteristic function.

To go even further with this idea we state as a general principle, that smoothness of the characteristic function at the origin corresponds to nice tail behavior of the distribution and vice versa. Another result in this direction is given by:

Proposition 4.22. *If*

$$\limsup_{h\to 0} \frac{\varphi(h) - 2\varphi(0) + \varphi(-h)}{h^2} > -\infty,$$

(which is true, e.g. if $\varphi''(0)$ exists), then $\mathbf{E}X^2 < \infty$.

Proof. Using Fatou's lemma,

$$\begin{aligned} \mathbf{E}X^2 &= \int x^2\mu(dx) \leq \liminf_{h\to 0} 2\int \frac{1-\cos(hx)}{h^2}\mu(dx) \\ &= -\limsup_{h\to 0} \int \frac{e^{ihx} - 2 + e^{-ihx}}{h^2}\mu(dx) < \infty \end{aligned}$$

□

Exercise 4.3. Show that if $\mathbf{E}|X|^k < \infty$, then $\varphi(t) \in C^k(\mathbb{R})$ and

$$\varphi^{(k)}(t) = \mathbf{E}((iX)^k e^{itX}).$$

Hint: prove it for $k = 1$ and use induction.

4.3 The Central Limit Theorem

We finally state what is widely regarded as the most important theorem in probability theory.

Theorem 4.23 (Central Limit Theorem)**.** *Let* $\{X_i, i \in \mathbb{N}\}$ *be i.i.d.,* $\mu = \mathbf{E}X_1$*, and* $\sigma^2 = \mathbf{E}X_1^2 < \infty$*. Then*

$$\frac{S_n - n\mu}{\sqrt{n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}} \xrightarrow{d} N(\mu, \sigma^2).$$

In final preparation for the proof, we have a couple technical lemmas.

Lemma 4.24.

$$\left| e^{ix} - \sum_{m=0}^{n} \frac{(ix)^m}{m!} \right| \leq \min\left\{ \frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right\}$$

Proof. We prove the lemma for $n = 2$, which is the case we use below, leaving the general case to the reader. We integrate by parts

$$\int_0^x (x-s)^n e^{is} ds = \frac{x^{n+1}}{n+1} + \frac{i}{n+1} \int_0^x (x-s)^{n+1} e^{is} ds.$$

Setting $n = 0$ yields

$$e^{ix} = 1 + ix + i^2 \int_0^x (x-s) e^{is} ds$$

and using $n = 1$, this becomes

$$e^{ix} = 1 + ix + \frac{i^2 x^2}{2} + \frac{i^3}{2} \int_0^x (x-s)^2 e^{is} ds.$$

Since $\left|\int_0^x (x-s)^2 e^{is} ds\right| \leq \int_0^x (x-s)^2 ds = \frac{x^3}{3}$, this proves the first estimate.

For the other estimate, we start again with

$$e^{ix} = 1 + ix + \frac{i^2 x^2}{2} + \frac{i^3}{2} \int_0^x (x-s)^2 e^{is} ds.$$

but now integrating the integral by parts yields

$$\int_0^x (x-s)^2 e^{is} ds = \frac{2}{i} \int_0^x (x-s)(e^{is} - 1) ds.$$

Since $\left|e^{is} - 1\right| \leq 2$, this latter integral is bounded by $\frac{x^2}{2}$, and we may conclude the second bound. □

Lemma 4.25. *If* $\mathbf{E}X^2 < \infty$, *then* $\varphi_X(t) = 1 + it\mathbf{E}X - \frac{t^2}{2}\mathbf{E}X^2 + o(t^2),\ t \to 0$.

Proof. Using $|\mathbf{E}Y| \leq \mathbf{E}|Y|$ and Lemma 4.24, we have

$$\begin{aligned}\left|\mathbf{E}e^{itX} - \mathbf{E}\left(1 + itX - \frac{t^2}{2}X^2\right)\right| &\leq \mathbf{E}\min\left(\frac{t^3|X|^3}{6}, t^2X^2\right) \\ &= t^2\mathbf{E}\min\left(\frac{t|X|^3}{6}, X^2\right).\end{aligned}$$

Since $\min\left(\frac{t|X|^3}{6}, X^2\right) \leq X^2 \in L^1$ and $\min\left(\frac{t|X|^3}{6}, X^2\right) \to 0$ as $t \to \infty$, Dominated Convergence proves the theorem. □

We are finally ready to prove the CLT. Since we are now equipped with the infrastructure of characteristic functions, the proof turns out to be quite short (and elegant).

Proof of the Central Limit Theorem. By translating our random variables by the mean, we may assume without loss of generality that $\mu = \mathbf{E}X_1 = 0$. Then with the help of Lemma 4.25, we have

$$\begin{aligned}\varphi_{S_n}(t) &= \mathbf{E}e^{itS_n/\sqrt{n}} = \mathbf{E}e^{i\frac{t}{\sqrt{n}}(X_1+\cdots+X_n)} \\ &= \left(\mathbf{E}e^{i\frac{t}{\sqrt{n}}X_1}\right)^n = \left(\varphi_{X_1}\left(\frac{t}{\sqrt{n}}\right)\right)^n \\ &= \left(1 - \frac{t^2}{2n}\sigma^2 + o\left(\frac{t^2}{n}\right)\right)^n \\ &\to e^{-\frac{t^2\sigma^2}{2}} = \varphi_{N(0,\sigma^2)}(t).\end{aligned}$$

By Lévy's Continuity Theorem, we have $\frac{S_n}{\sqrt{n}} \xrightarrow{d} N(0,\sigma^2)$. □

Exercise 4.4. State and prove a multivariate version of the CLT based off of Example 2.17.

Remark 4.26. Let us again emphasize that the CLT is a refinement of the WLLN. Let $\mathbf{E}X_1 = \mu$ and $\operatorname{Var} X_1 = \sigma^2$. Then we have that

$$\frac{S_n - n\mu}{\sqrt{n}} \xrightarrow{d} N(0,\sigma^2).$$

We can then multiply both sides by $C = \frac{1}{\sqrt{m}}$. Then we have

$$\begin{aligned}\frac{S_n - n\mu}{\sqrt{mn}} &\xrightarrow{\mathrm{d}} \frac{1}{\sqrt{m}}N(0,\sigma^2) \\ &= N(0,\sigma^2/m).\end{aligned}$$

Heuristically, if $m = n$ (which is only a heuristic since m is fixed), then the right hand side converges in distribution to zero (and by Exercise 4.1 also converges in probability to zero). If we have tightness in m, i.e., uniform bounds on m, we can make this heuristic rigorous. A simple proof is provided by characteristic functions, below.

Theorem 4.27 (Weak Law of Large Numbers, revisited)**.** *If $\mu = \mathbf{E}\,|X_1| < \infty$, $\{X_k, k \in \mathbb{N}\}$ are i.i.d., and $S_n = X_1 + \cdots + X_n$, then*

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

Proof. We start with the characteristic function for $\frac{S_n}{n}$. Then we have

$$\begin{aligned}
\varphi_{\frac{S_n}{n}}(t) &= \mathbf{E}e^{i\frac{t}{n}(X_1+\cdots X_n)} \\
&= \left(\varphi\left(\frac{t}{n}\right)\right)^n \\
&= \left(1 + \frac{it}{n}\mathbf{E}X_1 + o\left(\frac{1}{n}\right)\right)^n \\
&\to e^{it\mathbf{E}X_1}.
\end{aligned}$$

But, e^{itc} is the characteristic function of a constant c. Thus, by Lévy's Continuity Theorem,

$$\frac{S_n}{n} \xrightarrow{d} \mu$$

and, by Exercise 4.1,

$$\frac{S_n}{n} \xrightarrow{\text{pr}} \mu.$$

□

Exercise 4.5. Suppose $\sum_{k\geq 1} |a_k| < \infty$. If $\{X_k, k \geq 1\}$ are all independent and each has finite variance, then

$$\lim_{n\to\infty} \sum_{k=1}^{n} a_k X_k$$

converges in probability.

Let us end our discussion of the CLT with a statement (without proof) of one important generalization of the CLT, and a counterexample to the CLT in the case of infinite variance.

Theorem 4.28 (Lindeberg-Feller Theorem)**.** *Let $\{X_{n,m},\, 1 \leq m \leq n, n \in \mathbb{N}\}$ be a triangular array of random variables such that*

$$\mathbf{E}X_{n,m} = 0 \quad \textit{and} \quad \mathbf{E}X_{n,m}^2 = \sigma_{n,m}^2 < \infty.$$

Also, suppose

$$\sum_{m=1}^{n} \sigma_{n,m}^2 =: \sigma_n^2 \to \sigma^2.$$

If for each n, the random variables $\{X_{n,m}, 1 \le m \le n\}$ are independent and satisfy the Lindeberg condition:

$$\lim_{n\to\infty} \sum_{m=1}^{n} \mathbf{E}(X_{n,m}^2 \mathbf{1}_{\{|X_{n,m}|>\epsilon\}}) = 0 \quad \text{for all } \epsilon > 0,$$

then $S_n = \sum_{m=1}^{n} X_{n,m} \xrightarrow{d} N(0, \sigma^2)$.

Note that the Lindeberg-Feller Theorem does not require identical distribution of the random variables, only independence.

Next we look at a case where the CLT does not hold.

Example 4.29 (Symmetric stable distributions). Consider

$$f_X(x) = \begin{cases} \frac{\alpha}{2|x|^{\alpha+1}}, & |x| \ge 1 \\ 0, & |x| < 1 \end{cases}$$

for $0 < \alpha < 2$. Then we can see that f_X is a density since

$$2\int_1^\infty \frac{\alpha}{2} x^{-(\alpha+1)}\, dx = 1.$$

By the definition of the characteristic function (for symmetric random variables) we have that

$$\frac{1-\varphi_X(t)}{t^\alpha} = \int_1^\infty \frac{(1-\cos(tx))\alpha x^{-(\alpha+1)}}{t^\alpha}\, dx.$$

Let $u = xt$. Then $du = tdx$ and we have that

$$\frac{1-\varphi_X(t)}{t^\alpha} = \int_t^\infty \frac{1-\cos u}{u^{\alpha+1}} \alpha\, du < \infty. \tag{4.5}$$

Since $\cos(u)$ looks like $1 - \frac{u^2}{2}$ for small u and since $\alpha < 2$, one can see that as $t \to 0$, the integral above converges to some constant

$$c = \int_0^\infty \frac{1-\cos u}{u^{\alpha+1}} \alpha\, du < \infty.$$

Multiplying both sides of (4.5) by t^α, one gets

$$\varphi_X(t) = 1 - c\,|t|^\alpha + o(|t|^\alpha) \quad \text{as } t \to 0.$$

Let $\{X_k, k \in \mathbb{N}\}$ be i.i.d. random variables and

$$\frac{X_1 + \cdots + X_n}{n^{1/\alpha}} =: \frac{S_n}{n^{1/\alpha}}.$$

Then

$$\begin{aligned}\varphi_{\frac{S_n}{n^{1/\alpha}}}(t) &= \left(\varphi_X\left(\frac{t}{n^{1/\alpha}}\right)\right)^n \\ &= \left(1 - \frac{c\,|t|^\alpha}{n} + o\left(\frac{1}{n}\right)\right)^n \\ &\to e^{-c|t|^\alpha}.\end{aligned}$$

By Lévy's Continuity Theorem, $e^{-c|t|^\alpha}$ is the characteristic function of some random variable Y, so that

$$\frac{S_n}{n^{1/\alpha}} \xrightarrow{d} Y$$

(this can also be seen using Bochner's Theorem which is covered in the next section).

When the characteristic function φ_Y takes the form the $e^{-c|t|^\alpha}$, Y is said to have a *symmetric stable distribution*. For example, if $\alpha = 1$, then $e^{-c|t|}$ is the characteristic function of the Cauchy distribution which has density $(\pi(1 + x^2))^{-1}$. Note that, in line with Remark 4.20, $e^{-c|t|}$ is not differentiable at 0 since it has "heavy tails".

4.4 The Moment Method

Historically, the CLT was first verified for i.i.d. sequences of random variables having finite moments of all order. In this case, a more intuitive method for checking convergence in distribution is the so-called moment method. Throughout this section let $\mathbf{E}X^k = M_k$ for some random variable X. When the method is applicable, it requires that one check

$$\mathbf{E}X_n^k \xrightarrow{n\to\infty} M_k,\ k \in \mathbb{N}, \tag{4.6}$$

to deduce convergence in distribution to X.

When the limiting distribution has compact support $[-c, c]$, then the Stone-Weierstrass Theorem tells us the monomials $\{x^k, k \in \mathbb{N}\}$ form a basis for the vector space of continuous functions $C[-c, -c]$ equipped with the uniform topology. Thus, if two random variables satisfy $\mathbf{E}X^k = \mathbf{E}Y^k$ for all $k \in \mathbb{N}$, then by the Bounded Convergence Theorem,

$$\mathbf{E}f(X) = \mathbf{E}f(Y) \quad \text{for all } f \in C[-c, c]$$

which tells us that X and Y have the same distribution, i.e., the distribution of X is uniquely determined by its moments[11]. We will soon see that if uniquely determines the distribution of X, then (4.6) is enough to check convergence in distribution to X.

Of course, the Gaussian distribution does not have compact support, and in fact, the method is not in general true for distributions with unbounded

[11] This is often referred to as satisfying the "(determinant) moment problem."

support. A well-known sufficient criterion under which the method does hold is Carleman's Condition [Car22]:

$$\sum_{k\in\mathbb{N}} M_{2k}^{-1/2k} < \infty.$$

We will prove that under a slightly stronger assumption, which one can check that the Gaussian distribution satisfies, the moment method holds.

Theorem 4.30 (Criterion for the moment method)**.** *Suppose X has a distribution μ_X with finite moments of all orders:*

$$M_k := \int x^k \, d\mu_X < \infty \quad \textit{for all } k \in \mathbb{N}.$$

If the power series

$$\sum_{k\in\mathbb{N}} M_k x^k/k! \tag{4.7}$$

has a positive radius of convergence R, then μ_X is uniquely determined by its moments.

Proof. Denote the absolute moments by $A_k := \int |x|^k \, d\mu_X$. Since (4.7) has a positive radius of convergence $R > 0$, by looking at (4.7) strictly inside R, we see that $\lim_{k\to\infty} A_k = 0$.

By Lemma 4.24,

$$\left| \mathbf{E} e^{itX} e^{ihX} - \sum_{k=0}^{n} \mathbf{E}\left(e^{itX} \frac{(ihX)^k}{k!} \right) \right| \leq \frac{1}{(n+1)!} \mathbf{E}\left| hX^{n+1} \right|.$$

By Exercise 4.3 we also have

$$\left| \varphi_X(t+h) - \sum_{k=0}^{n} \frac{h^k}{k!} \varphi_X^{(k)}(t) \right| \leq \frac{1}{(n+1)!} \mathbf{E}\left| hX^{n+1} \right|. \tag{4.8}$$

When $|h| < R$, we have already seen that the right side goes to zero as $n \to \infty$ so that

$$\varphi_X(t+h) = \sum_{k=0}^{\infty} \frac{h^k}{k!} \varphi_X^{(k)}(t) \tag{4.9}$$

for $|h| < R$ and all $t \in \mathbb{R}$. Setting $t = 0$, we have that the characteristic function $\varphi_X(h)$ is determined by the sequence (M_k) at all $|h| < R$. Next set $|t| = R/2$. At those two values of t, (4.9) holds for all $|h| < R$ whereby we have extended the interval on which φ_X is determined. Continuing in this fashion proves the theorem. □

As we mentioned earlier, the determinacy of the moments extends to a method for checking convergence in distribution

Theorem 4.31 (The moment method)**.** *If the distribution of X is determined by its moment sequence (M_k), then* (4.6) *implies convergence in distribution of (X_n) to X.*

Proof. We start by observing that Proposition 4.10 and convergence of second moments of the sequence (X_n) implies tightness of the distributions $\{\mu_{X_n}, n \in \mathbb{N}\}$. Consider a subsequence along which we have weak convergence to some random variable Y:

$$X_{n_k} \xrightarrow{\mathrm{d}} Y.$$

By Exercise 2.11, $\{X_{n_k}^p, k \in \mathbb{N}\}$ are uniformly integrable for every $p \in \mathbb{N}$, thus

$$\lim_{k\to\infty} \mathbf{E}X_{n_k}^p = \mathbf{E}Y^p.$$

By the determinacy of moments, $X \stackrel{\mathrm{d}}{=} Y$. Since this argument holds for any subsequence we choose, it must be that all distributional limit points have the same distribution as X. □

4.5 Bochner's Theorem

We end this chapter with an important theorem concerning characteristic functions.

Definition 4.32. A function $f : \mathbb{R} \to \mathbb{C}$ is called positive semi-definite if for all $(t_1, \ldots, t_n) \in \mathbb{R}^n$ and $(z_1, \ldots, z_n) \in \mathbb{C}^n$, we have that

$$\sum_{j=1}^{n}\sum_{k=1}^{n} \overline{z_k} z_j f(t_k - t_j) \geq 0.$$

Note that this coincides with the definition for matrices since if we consider the matrix with entries $a_{kj} = f(t_k - t_j)$, we have for $z = (z_1, \ldots, z_n)$

$$z^* A z = \sum_{j,k=1}^{n} \overline{z_k} z_j f(t_k - t_j) \geq 0.$$

Proposition 4.33. *Every characteristic function φ is positive semi-definite.*

Proof.

$$\begin{aligned}
\sum_{j,k=1}^{n} \overline{z_k} z_j \varphi(t_k - t_j) &= \sum_{j,k=1}^{n} \int_{\mathbb{R}} \overline{z_k} z_j e^{i(t_k - t_j)x} \mu(dx) \\
&= \int_{\mathbb{R}} \sum_{j,k=1}^{n} \overline{z_k} z_j e^{it_k x} e^{-it_j x} \mu(dx) \\
&= \int_{\mathbb{R}} \left(\sum_{k=1}^{n} \overline{z_k e^{-it_k x}} \right) \left(\sum_{j=1}^{n} z_j e^{-it_j x} \right) \mu(dx) \\
&= \int_{\mathbb{R}} \left| \sum_{j=1}^{n} z_j e^{-it_j x} \right|^2 \mu(dx) \\
&\geq 0
\end{aligned}$$

□

Theorem 4.34 (Bochner's Theorem)**.** *If $\varphi(t)$ is a positive semi-definite function such that $\varphi(0) = 1$ and which is continuous at 0, then $\varphi(t)$ is the characteristic function for some probability distribution on $\mathbb{R}$.*

Exercise 4.6. If $f \in L^1$ is continuous, then

$$\int_{-T}^{T} \left(1 - \frac{|t|}{T}\right) f(t)\, dt = \frac{1}{T} \int_0^T \int_0^T f(t-s)\, ds\, dt.$$

Hint: Use $\mathbf{1}_{\{s<t\}}$ and the Fubini-Tonelli Theorem.

Lemma 4.35 (Properties of positive semi-definite functions)**.** *If $\varphi(t)$ and $\{\varphi_k(t), k \in \mathbb{N}\}$ are continuous, positive semi-definite functions and $\varphi_k(0) = 1$ for all $k \in \mathbb{R}$, then:*

(a) $\varphi(t)e^{ita}$ is positive semi-definite for all $a \in \mathbb{R}$.

(b) $|\varphi(t)| \leq 1$ and $\varphi(-t) = \overline{\varphi(t)}$ for all $t \in \mathbb{R}$.

(c) The mixture

$$\psi(t) := \sum_{k=1}^{m} p_k \varphi_k(t)$$

is a continuous, positive semi-definite function with $\psi(0) = 1$ for all probability vectors $p = (p_1, \ldots, p_m)$.

(d) For all $s, t \in \mathbb{R}$ $|\varphi(t) - \varphi(s)|^2 \leq 4\,|1 - \varphi(t-s)|$.

Proof. (a) For $z \in \mathbb{C}^n$ and $t \in \mathbb{R}^n$, we have

$$\begin{aligned}
&\sum_{j,k=1}^{n} \varphi(t_k - t_j) e^{i(t_k - t_j)a} z_j \overline{z_k} \\
&= \sum_{j,k=1}^{n} \varphi(t_k - t_j) z_j e^{it_j a} \overline{z_k e^{it_k a}} \\
&= \sum_{j,k=1}^{n} \varphi(t_k - t_j) v_j \overline{v_k} \quad \text{for } v = (z_1 e^{it_1 a}, \ldots, z_n e^{it_n a}) \\
&\geq 0.
\end{aligned}$$

(b) For $n = 2$, we have that

$$\sum_{j=1}^{2} \sum_{k=1}^{2} \varphi(t_k - t_j) z_j \overline{z_k} \geq 0,$$

and the 2×2 matrix with entries $a_{kj} = \varphi(t_k - t_j)$ is positive semi-definite. The two properties follow from properties of positive semi-definite matrices and the fact that diagonal entries are 1 since $\varphi(0) = 1$.

(c) The only property which is not obvious is whether ψ is positive semi-definite. This can be seen by

$$\begin{aligned}
\sum_{j,k=1}^{n} \psi(t_j - t_k) z_j \overline{z_k} &= \sum_{\ell=1}^{m} p_\ell \underbrace{\sum_{j,k=1}^{n} \varphi_\ell(t_k - t_j) z_j \overline{z_k}}_{\geq 0} \\
&\geq 0.
\end{aligned}$$

(d) For $n = 3$, with $t_1 = t$, $t_2 = s$, and $t_3 = 0$, we have that

$$\begin{aligned}
0 &\leq \det \begin{pmatrix} 1 & \varphi(t-s) & \varphi(t) \\ \overline{\varphi(t-s)} & 1 & \varphi(s) \\ \overline{\varphi(t)} & \overline{\varphi(s)} & 1 \end{pmatrix} \\
&= 1 - |\varphi(t) - \varphi(s)|^2 - |\varphi(t-s)|^2 - \varphi(s)\overline{\varphi(t)}(1 - \varphi(t-s)) - \overline{\varphi(s)}\varphi(t)(1 - \overline{\varphi(t-s)}) \\
&\leq 1 - |\varphi(t) - \varphi(s)|^2 - |\varphi(t-s)|^2 + 2\,|1 - \varphi(t-s)|\,.
\end{aligned}$$

Thus,

$$\begin{aligned}
|\varphi(t) - \varphi(s)|^2 &\leq 1 - |\varphi(t-s)|^2 + 2\,|1 - \varphi(t-s)| \\
&\leq 2\,|1 - \varphi(t-s)| + 2\,|1 - \varphi(t-s)| \\
&= 4\,|1 - \varphi(t-s)|\,.
\end{aligned}$$

□

Remark 4.36. Property (d) shows that if $\varphi(0) = c > 0$ and φ is continuous at zero and positive semi-definite then φ is uniformly continuous.

Proof of Theorem 4.34. By Remark 4.36, it is enough to show the result for continuous φ. We will first assume that $|\varphi(t)| \in L^1(\mathbb{R})$ (proving the general case later). In this case, we have already seen from the corollary to Lévy's Inversion Formula that φ is the characteristic function of a distribution that has a density.

We begin by defining our candidate density as

$$f(x) := \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t)\, dt. \tag{4.10}$$

We first see that $f \geq 0$. By Exercise 4.6,

$$\begin{aligned} f(x) &= \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} e^{-itx} \varphi(t) \left(1 - \frac{|t|}{T}\right) dt \\ &= \lim_{T\to\infty} \frac{1}{T} \frac{1}{2\pi} \int_0^T \int_0^T e^{-i(u-s)x} \varphi(u-s)\, du\, ds \\ &= \lim_{T\to\infty} \frac{1}{T} \frac{1}{2\pi} \int_0^T \int_0^T e^{isx} \varphi(u-s) e^{-iux}\, du\, ds \\ &\geq 0 \end{aligned}$$

where the inequality in the last line holds by the fact that φ is positive semi-definite and a Riemann approximation argument (since φ is continuous).

Now, define

$$f_\sigma(x) := e^{-\frac{\sigma^2 x^2}{2}} f(x).$$

Then using the Fubini-Tonelli Theorem and Example 4.14 part (e),

$$\begin{aligned} \int_{\mathbb{R}} e^{itx} f_\sigma(x)\, dx &= \int_{\mathbb{R}} e^{itx} \left(e^{-\frac{\sigma^2 x^2}{2}} f(x)\right) dx \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} e^{itx} \left(e^{-\frac{\sigma^2 x^2}{2}} \int_{\mathbb{R}} \varphi(s) e^{-isx}\, ds\right) dx \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \varphi(s) \int_{\mathbb{R}} e^{i(t-s)x} e^{-\frac{\sigma^2 x^2}{2}}\, dx\, ds \\ &= \int_{\mathbb{R}} \varphi(s) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-s)^2}{2\sigma^2}}\, ds \end{aligned}$$

By Lemma 4.35 part (b), evaluating the above at $t = 0$ gives

$$\int_{\mathbb{R}} f_\sigma(x)\, dx = \int_{\mathbb{R}} \varphi(s) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{s^2}{2\sigma^2}}\, ds \leq 1.$$

By Fatou, we have that

$$\int_{\mathbb{R}} \liminf_{\sigma\to 0} f_\sigma(x) \leq 1$$

which implies

$$\int_{\mathbb{R}} f(x)\,dx \leq 1.$$

We want to show that $\int_{\mathbb{R}} f(x)\,dx = 1$. The calculation above allows us to use the Dominated Convergence Theorem to calculate

$$\begin{aligned}\int_{\mathbb{R}} e^{itx} f(x)\,dx &= \lim_{\sigma\to 0}\int_{\mathbb{R}} \varphi(s)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(t-s)^2}{2\sigma^2}}\,ds \\ &= \lim_{\sigma\to 0}\int_{\mathbb{R}} \varphi(t-\sigma u)\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}\,du \\ &= \varphi(t).\end{aligned}$$

Evaluating at $t = 0$, and using our assumption $\varphi(0) = 1$, we arrive at

$$1 = \varphi(0) = \int_{\mathbb{R}} f(x)\,dx$$

which proves the theorem in the case $\varphi \in L^1$.

We can expand this result for $\varphi \notin L^1(\mathbb{R})$ by setting

$$\varphi_\sigma(t) := \int_{\mathbb{R}} \varphi(t)\frac{1}{\sqrt{2\pi\sigma^2}}e^{ity}e^{-\frac{y^2}{2\sigma^2}}\,dy.$$

By parts (a) and (c) of the previous lemma, together with another Riemann approximation argument, φ_σ is positive semi-definite. We have

$$\varphi_\sigma(t) = \varphi(t)e^{-\frac{\sigma^2 t^2}{2}} \in L^1(\mathbb{R})$$

which, by the first part of the proof, implies $\varphi_\sigma(t)$ is a characteristic function. Letting $\sigma \to 0$ and using Lévy's Continuity Theorem completes the proof. □

5 The Law of Small Numbers

The phrase "Law of Small Numbers"[12] was coined by L. Von Bortkiewicz in his treatise (in German) by the same name [vB98] which concerns applications of Poisson convergence to annual occurrences of suicides, insurable accidents, and military deaths due to horse kicks, among other things. Contrary to what one may think, this notion is an alternative to the Central Limit Theorem, as opposed to the Law of Large Numbers.

[12]Recently, this phrase has been used to mean something quite different. Popularized by Kahneman and Tversky [TK71], the phrase also describes the statistical fallacy of making conclusions based on sample sizes which are too small.

5.1 Poisson Convergence

In the Central Limit Theorem, weak convergence was only possible because we normalized the sums by $1/\sqrt{n}$. The idea behind Poisson convergence[13] is that in order for an unnormalized sum of n identically distributed random variables to converge in distribution, finitely many of them must comprise most of the sum, while the rest of the random variables take values very close to zero (in the easiest case of Poisson convergence, they actually are equal to 0). As L. Breiman [Bre92, Sec. 9.5] eloquently puts it, this is "the difference between the sum of uniformly small smears, versus the sum of occasionally large blips."

For a typical picture of Poisson convergence, consider the interval $[0, 1]$ divided into n equal parts with n large and let $\{X_k, 1 \le k \le n\}$ be i.i.d. Ber $\left(\frac{\lambda}{n}\right)$. Then note that for $S_n = X_1 + \cdots + X_n$, we have

$$\mathbf{E}S_n = \lambda \quad \text{and} \quad \operatorname{Var}(S_n) = n\left(\frac{\lambda}{n}\right) + n\left(\frac{\lambda}{n}\right)^2 \longrightarrow \lambda.$$

Recall from Exercise 2.4 and Example 4.14 part (b) that $X \sim \text{Poiss}(\lambda)$ if

$$\mathbf{P}(X = k) = e^{-k}\frac{\lambda^k}{k!} \text{ for } k \in [0, \infty)$$

and

$$\varphi_X(t) = e^{\lambda(e^{it}-1)}.$$

In particular, if $\{X_i,\, 1 \le i \le n\}$ are i.i.d. Poisson with mean λ, then:

$$\varphi_{X_1+\cdots+X_n}(t) = \varphi_{S_n}(t) = e^{n\lambda(e^{it}-1)}.$$

Hence, $S_n \sim \text{Poiss}(n\lambda)$.

Theorem 5.1 (Poisson Convergence, Law of Small Numbers)**.** *Suppose*

$$\{X_{n,m}, 1 \le m \le n, n \in \mathbb{N}\}$$

is a triangular array of independent Bernoulli random variables with parameters $\{p_{n,m}\}$*. If for some* $\lambda \in (0, \infty)$

1. $\lim_{n\to\infty} \sum_{m=1}^{n} p_{n,m} =: \lim_{n\to\infty} p_n = \lambda$,
2. $\lim_{n\to\infty} \max_{1\le m\le n} p_{n,m} =: \lim_{n\to\infty} \tilde{p}_n = 0$,

then $\sum_{m=1}^{n} X_{n,m}$ *converges in distribution, as* $n \to \infty$*, to a Poisson distribution with mean* λ*.*

Remarks 5.2.

1. Note that uniformly in n, the number of nonzero summands is bounded in distribution by some random variable (i.e., tightness). This can be deduced by the fact that the variance of the sum S_n converges as $n \to \infty$ (see Proposition 4.10).

[13] This result is due to Abraham De Moivre; see Exercise 2.4.

2. Like the Lindeberg-Feller Theorem, the random variables are not required to be identically distributed. Rather, the key is independence[14].

3. Unlike the Lindeberg-Feller Theorem, the sum:

$$\sum_{m=1}^{n} \underbrace{\mathbf{E}(X_{n,m}^2 \mathbf{1}_{\{|X_{n,m}|>\epsilon\}})}_{p_{n,m}} \nrightarrow 0.$$

   Rather, by condition 1 in the theorem, $\sum_{m=1}^{n} p_{n,m} \to \lambda$. In particular, in the Lindeberg-Feller Theorem, the variance of the sum is controlled by making the value of $X_{n,m}$ small (i.e., $X_{n,m} \sim \frac{1}{\sqrt{n}} \operatorname{Rad}\left(\frac{1}{2}\right)$). Here, in Poisson convergence, the variance of S_n is controlled by making the probabilities $p_{n,m}$ small while the values $X_{n,m}$ are actually relatively big.

Before we prove the Poisson Convergence Theorem, we require the following Lemma:

Lemma 5.3. *If z_i and $z_i' \in \mathbb{C}$ such that $|z_i| \leq 1$ and $|z_i'| \leq 1$, then:*

$$\left|\prod_{i=1}^{n} z_i - \prod_{i=1}^{n} z_i'\right| \leq \sum_{i=1}^{n} |z_i - z_i'|.$$

Proof. We induct on n. The case when $n = 1$ is trivial since the left-hand side of the inequality is the same as the right-hand side. For the inductive step, assume the inequality holds for $n-1$, and rewrite the left-hand side as

$$\begin{aligned}
\left|\prod_{i=1}^{n} z_i - \prod_{i=1}^{n} z_i'\right| &= \left|z_1 \prod_{i=2}^{n} z_i - z_1 \prod_{i=2}^{n} z_i' + z_1 \prod_{i=2}^{n} z_i' - z_1' \prod_{i=2}^{n} z_i'\right| \\
&\leq \left|z_1 \prod_{i=2}^{n} z_i - z_1 \prod_{i=2}^{n} z_i'\right| + \left|z_1 \prod_{i=2}^{n} z_i' - z_1' \prod_{i=2}^{n} z_i'\right| \\
&\leq |z_1| \left|\prod_{i=2}^{n} z_i - \prod_{i=2}^{n} z_i'\right| + \left|(z_1 - z_1') \prod_{i=2}^{n} z_i'\right| \\
&\leq \left|\prod_{i=2}^{n} z_i - \prod_{i=2}^{n} z_i'\right| + |(z_1 - z_1')| \\
&\leq \sum_{i=2}^{n} |z_i - z_i'| + |(z_1 - z_1')|, \text{ (by the inductive assumption)} \\
&= \sum_{i=1}^{n} |z_i - z_i'|.
\end{aligned}$$

□

[14] Actually, all that is required is asymptotic independence as can be seen in [Adl78]

Proof of Theorem 5.1. To prove the theorem, we need to show convergence in distribution, for which we use the characteristic function of S_n and Lévy's Continuity Theorem. We will prove a slightly easier version where $\sum_{m=1}^{n} p_{n,m} = \lambda$ (leaving the generalization to the reader). We show $\lim_{n\to\infty} \varphi_{S_n}(t) = \varphi_Z(t)$ where Z has our desired Poisson distribution:

$$\begin{aligned}
&|\varphi_{S_n}(t) - \varphi_Z(t)| \\
&= \left| \prod_{m=1}^{n} (e^0(1-p_{n,m}) + p_{n,m}e^{it}) - e^{\lambda(e^{it}-1)} \right| \\
&= \left| \prod_{m=1}^{n} [1 + p_{n,m}(e^{it}-1)] - e^{\lambda(e^{it}-1)} \right| \\
&= \left| \prod_{m=1}^{n} [1 + p_{n,m}(e^{it}-1)] - e^{\sum_{m=1}^{n} p_{n,m}(e^{it}-1)} \right| \\
&= \left| \prod_{m=1}^{n} [1 + p_{n,m}(e^{it}-1)] - \prod_{m=1}^{n} e^{p_{n,m}(e^{it}-1)} \right|
\end{aligned}$$

Since $|\varphi(t)| \le 1$, by Lemma 5.3, the above is bounded by

$$\begin{aligned}
&\sum_{m=1}^{n} \left| e^{p_{n,m}(e^{it}-1)} - (1 + p_{n,m}(e^{it}-1)) \right| \\
&\le \sum_{m=1}^{n} \frac{c(p_{n,m}(e^{it}-1))^2}{2} \quad \text{(for large } n\text{, } |p_{n,m}(e^{it}-1)| \le 1) \\
&\le 2c \sum_{m=1}^{n} p_{n,m}^2 \qquad\qquad \text{(since } |(e^{it}-1)| \le 2).
\end{aligned}$$

By the definition of $\tilde{p}_n$, this is bounded by

$$2c \sum_{m=1}^{n} (p_{n,m})\tilde{p}_n = 2c\tilde{p}_n \sum_{m=1}^{n} p_{n,m}$$

which goes to 0 since $\tilde{p}_n \to 0$. □

5.2 Poisson Processes

Theorem 5.1 can be used to construct a Poisson process. In fact, in this section we will give two constructions of Poisson point process– first on $\mathbb{R}$ using exponential random variables, and then on a more general space using Poisson random variables. Let us now describe the arrival times construction. First, we set $p_{n,m} = \frac{\lambda}{n}$. Note that trivially, the sum $\sum_{m=1}^{n} p_{n,m}$ goes to λ as $n \to \infty$, and the maximum taken over m also goes to zero as $n \to \infty$. We have thus met the two conditions of Theorem 5.1. To see the connection of that theorem to exponential random variables:

1. Consider the interval $[0,1]$ divided into n parts or "time intervals" and put a 1 or 0 on each of the n time intervals. At this discrete level, 1's are thought of as "arrivals" and 0's signify "no arrivals".

2. The k^{th} arrival to the right of the origin is $\tau_k^{(n)}$. At the n^{th} stage (of dividing the unit interval), each arrival $\tau_k^{(n)}$ has a Geom $\left(\frac{\lambda}{n}\right)$ distribution, i.e.,
$$\mathbf{P}(\tau_1^{(n)} = m) = \frac{\lambda}{n}\left(1 - \frac{\lambda}{n}\right)^{m-1}.$$

3. In order to take a continuum limit, we must consider arrivals for different n to exist on the same copy of the unit interval $[0,1]$. We divide the arrivals by n and, letting m depend on n, rewrite the geometric probabilities as
$$\mathbf{P}\left(\frac{\tau_1^{(n)}}{n} = \frac{m_n}{n}\right) = \frac{\lambda}{n}\left(1 - \frac{\lambda}{n}\right)^{m_n - 1}.$$
As $n \to \infty$, we require that $\frac{m_n}{n}$ takes some limiting value $t \in [0,1]$. In particular, if $\tau_k^{(n)}/n \xrightarrow{\text{d}} \tau_k$, as $n \to \infty$, then,
$$\mathbf{P}(\tau_1 = t) \simeq \left(1 - \frac{\lambda}{n}\right)^{nt} \frac{\lambda}{n} \to \lambda e^{-\lambda t} dt.$$

Exercise 5.1. Supposing $\tau_k^{(n)} \xrightarrow{d} \tau_k$, show that

(a) the difference $\tau_k - \tau_{k-1}$ is exponentially distributed with parameter λ, and

(b) the so-called *memoryless property*: $(\tau_k - \tau_{k-1}) \perp\!\!\!\perp \tau_{k-1}$.

A *Poisson process* on $\mathbb{R}^+$ is now defined as the set of random times $\{\tau_k, k \in \mathbb{N}\}$ (alternatively, the set of random points in $\mathbb{R}^+$) where
$$\tau_k := \sum_{j=1}^{k} T_j$$
and the $\{T_j, j \in \mathbb{N}\}$ are i.i.d. Exp(λ) random variables. The reason for the name of this process is due to the following theorem:

Theorem 5.4 (Poisson number of arrivals)**.** *Let $N(B)$ be the number of arrivals in a Borel set B. Then*
$$N(t) := N([0,t]) \sim \text{Poiss}(\lambda t).$$
In other words, $\mathbf{P}(N = k) = \dfrac{e^{-\lambda t}(\lambda t)^k}{k!}$.

Recall from Example 2.23, that the Exp(λ) distribution is a special case of a Gamma distribution such that $\nu = 1$, and that the sum of two independent Gamma(ν_1, λ) and Gamma(ν_2, λ) random variables is a Gamma($\nu_1 + \nu_2, \lambda$) random variable.

Proof of Theorem 5.4. We prove only the case where B is of the form $[0, t]$ by calculating

$$\begin{aligned} \mathbf{P}(N(t) = k) &= \mathbf{P}(N(t) \geq k) - \mathbf{P}(N(t) \geq k+1) \\ &= \mathbf{P}(\tau_k < t) - \mathbf{P}(\tau_{k+1} < t) \\ &= \int_0^t \lambda e^{-\lambda s} \frac{(\lambda s)^{k-1}}{(k-1)!} du - \int_0^t \lambda e^{-\lambda s} \frac{(\lambda s)^k}{k!} ds. \end{aligned}$$

Letting $u = e^{-\lambda s}$ and $dv = \lambda \frac{(\lambda s)^{k-1}}{(k-1)!} ds$, we get

$$\begin{aligned} \int_0^t u\, dv + \int_0^t v\, du &= [uv]_0^t \\ &= e^{-\lambda t} \frac{(\lambda t)^k}{k!}. \end{aligned}$$

□

Exercise 5.2. Conditioned on the event $N(t) = N([0, t]) = k$, the k points are uniformly distributed, i.e., the k unordered points $(X_1, \ldots, X_k) \sim \text{Unif}[0, t]^k$, or the k ordered points $(X^{(1)}, \ldots, X^{(k)})$, $X^{(1)} < \cdots < X^{(k)}$, are uniform on the simplex

$$\Delta^k := \big\{(t_1, \ldots, t_k) \in \mathbb{R}^k : 0 \leq t_1 < \cdots < t_k \leq t \quad \text{for all } j = 1, \ldots, k\big\}.$$

Hint: Try $k = 1$ first and then use induction.

We will now generalize the Poisson process. Recall that a stochastic process is just a family $\{X_t\}_{t \in T}$ of random variables indexed by a set T. Typically we think of $T \subset \mathbb{N}$ (discrete-time stochastic process) or $T \subset \mathbb{R}^+$ (continuous-time stochastic process), but one can easily extend to negative times and even more general spaces. When $T \subset \mathbb{R}^d$ for $d > 1$, we sometimes call the stochastic process a "random field," which is often seen in physical applications. We will take this even one step further by considering T to be the collection of Borel subsets of $\mathbb{R}^d$.

Definition 5.5. A *Poisson process* $N(d\lambda)$ on $\mathbb{R}^d$ with intensity λ is a collection of random variables $\{N_B, B \in T\}$ indexed by $T = \mathcal{B}^d$ such that

$$N_B \equiv N(B) \sim \text{Poiss}(\lambda m(B)),$$

and $N(B) \perp\!\!\!\perp N(A)$ whenever $A \cap B = \emptyset$.

For existence of such a process, one may use a weak limit construction, as before, from discrete grid approximations and $\mathrm{Ber}(\lambda/n^d)$ random variables where 1's converge to a set of points in $\mathbb{R}^d$. We prefer an alternative way using the idea behind Exercise 5.2. For this point of view, it is helpful to see Poisson processes as random measures.

Definition 5.6. Let $\mathcal{M}(\mathbb{R}^d)$ denote the set of all σ-finite measures on $\mathbb{R}^d$ (the measures on $\mathbb{R}^d$ for which $\mathbb{R}^d$ can be written as a countable union of finite-measure sets). A *random measure* ν on $(\mathbb{R}^d, \mathcal{B}^d)$ is a mapping from $(\Omega, \mathcal{F})$ to $\mathcal{M}(\mathbb{R}^d)$. In other words, for $\mathbf{P}$-almost every $\omega \in \Omega$, $\nu(\omega, \cdot)$ is a σ-finite measure on $(\mathbb{R}^d, \mathcal{B}^d)$, and for each $B \in \mathcal{B}^d$, $\nu(\cdot, B)$ is an $\overline{\mathbb{R}}$-valued random variable.

Remark 5.7. The space $\mathcal{M}(\mathbb{R}^d)$ is actually metrizable under the so-called vague topology and in fact, it is complete and separable under this metric. Once one has random elements of a complete separable metric space (Polish space) one can talk about convergence in distribution of such random elements (in this case random measures). For more details, see for example [Res87].

Example 5.8 (Poisson random measure). In this setting, we can think of $N(d\lambda)$ as a collection

$$\{N(\omega, B); \omega \in \Omega, B \in \mathcal{B}^d\},$$

where for fixed ω and for some random set $\{X_i(\omega) = x_i, i \in \mathbb{N}\}$,

$$N(\omega, \cdot) = \sum_{i=1}^{\infty} \delta_{x_i}$$

is a sum of delta measures. On the other hand, for fixed $B \in \mathcal{B}^d$,

$$N(\cdot, B) = \sum_{i=1}^{\infty} \delta_{X_i}(B) = \#\{i : X_i \in B\}$$

is a Poisson random variable with parameter $\lambda\, m(B)$. We will explicitly construct the set $\{X_i, i \in \mathbb{N}\}$ below.

Remark 5.9. Up until now, the intensity λ has been a constant, but one can generalize this by letting $\lambda(dx)$ be a (deterministic) measure on $\mathbb{R}^d$ which is absolutely continuous with respect to Lebesgue measure. The only thing that changes now is the parameter of the Poisson distributions associated to Borel sets:

$$N(B) \sim \mathrm{Poiss}(\lambda(B)).$$

In this general setting, $\lambda(dx)$ is called the *intensity measure*, *mean measure*, or sometimes *control measure*. Such a generalization is applicable when using Poisson processes on $\mathbb{R}^+$ to model the number of customers which arrive in a store on a given day. In this case, $\lambda(dx)$ can be varied according to the local "rate" at which customers arrive.

Technically, one can also consider intensity measures which have portions that are singular with respect to Lebesgue, but we will not be concerned with these here.

Theorem 5.10 (Existence of Poisson random measures)**.** *If $\lambda(dx)$ is a measure on $\mathbb{R}^d$ which absolutely continuous with respect to Lebesgue measure, then there exists a Poisson random measure on $\mathbb{R}^d$ with intensity measure $\lambda(dx)$.*

Proof. We prove the case when $\lambda(dx)$ is a finite measure on $S \subset \mathbb{R}^d$ with norm $\|\lambda\|$. In the general case, one may consider $\mathbb{R}^d$ to be a disjoint union of sets $\biguplus_i S_i = \mathbb{R}^d$ such that $\lambda(S_i) < \infty$, and simply glue together the Poisson random measures restricted to the various S_i.

By Kolmogov's Extension Theorem, we may construct an infinite sequence of independent S-valued random vectors $(Y_n, n \in \mathbb{N})$ with distribution $\lambda/\|\lambda\|$. Also let $Y \sim \text{Poiss}(\|\lambda\|)$ be independent of (Y_n). Our candidate random measure is given by the random sum of δ-measures:

$$N(d\lambda) := \sum_{i=1}^{Y} \delta_{Y_i}.$$

For each Borel subset $B \subset S$, we then have that $N(B) = \sum_{i=1}^{Y} \mathbf{1}_{\{Y_i \in B\}}$ is a well-defined random variable.

Given a mutually disjoint union of S, say $B_1, \ldots, B_k$, a calculation shows

$$\begin{aligned}\mathbf{P}\left(N(B_1) = n_1, \ldots, N(B_k) = n_k\right) &= \binom{n}{n_1, \ldots, n_k} \prod_{i=1}^{k} \left(\frac{\lambda(B_i)}{\|\lambda\|}\right)^{n_i} e^{-\|\lambda\|} \frac{\|\lambda\|^n}{n!} \\ &= \prod_{i=1}^{k} e^{-\lambda(B_i)} \frac{\lambda(B_i)^{n_i}}{n_i!}\end{aligned}$$

for non-negative integers $n_1, n_2, \ldots, n_k$ which sum to n. which shows that $N(d\lambda)$ has our desired properties. □

Exercise 5.3. Suppose $N(d\lambda)$ is a Poisson process on $[0, T]^2$ with constant intensity λ. You play a gambling game where you pick one of the points, say (X, Y), of this Poisson process uniformly at random. Conditioning on there being at least one point, and supposing you earn XY dollars, what is the fair price of this game? What if you get ear the minimum X coordinate of all points (still conditioned on at least one point)– what is the new fair price?

5.3 Bernstein's Block Method

Independence is typically not observed in real systems that are modeled by sequences of random variables. As such, one would like to know if the CLT and Poisson convergence hold for dependent sequences of random variables. General dependence is difficult to deal with, so we will also assume stationarity (recall that the Ergodic Theorem shows the SLLN holds for certain stationary sequences). The most common apparatus for extending probabilistic limit theorems from independent sequences to dependent sequences is Bernstein's block method introduced in [Ber27] (for a further discussion see [IL71, Ch. 18]). In this section, we apply this method in the context of Poisson convergence.

Theorem 5.11 (Poisson convergence for dependent arrays)**.** *Let*

$$\{X_{n,m}, 1 \le m \le n, n \in \mathbb{N}\}$$

be a triangular array of Bernoulli random variables such that each row is stationary. Assume for some $\lambda > 0$*,*

$$\lim_{n\to\infty} n\mathbf{P}(X_{n,1} = 1) =: \lim_{n\to\infty} np_n = \lambda \tag{5.1}$$

and for $1 \le m \le n$*,*

$$|\mathrm{Cov}(X_{n,1}, X_{n,m})| \le o\left(\frac{1}{n}\right)\theta^m, \tag{5.2}$$

for some $\theta \in (0,1)$*. Also assume that for every sequence of indices*

$$i_1 < \cdots < i_p < j_1 < \cdots < j_q$$

satisfying $j_1 \ge i_p + k$*,*

$$\left|\mathrm{Cov}\left(\prod_{\ell=1}^{p} X_{n,i_\ell}, \prod_{\ell=1}^{q} X_{n,j_\ell}\right)\right| \le C\theta^k \tag{5.3}$$

for some $C > 0$*. Then* $(\sum_{m=1}^{n} X_{n,m}, n \in \mathbb{N})$ *converges in distribution to a* Poiss(λ) *random variable.*

Proof. Below C denotes a positive constant that may change from line to line.

If the random variables in each row were independent then this is a special case of Theorem 5.1. We do not have independence, but assumptions (ii) and (iii) imply asymptotic independence which controls the dependence between $X_{n,1}, \cdots, X_{n,n}$. This will be done via Bernstein's classical small-large block method. More precisely, choose $0 < a < b < 1$, and for any $n \ge 1$, divide $\{1, \ldots, n\}$ into a sequence of pairs of alternating big intervals (blocks) of length $[n^b]$ and small blocks of length $[n^a]$ where $[\cdot]$ is the ceiling function. The number of pairs of big and small blocks is $B = [n/([n^a] + [n^b])] \simeq n^{1-b}$. There may be a leftover partial block L in the end which is is negligible since

$$\sum_{m\in L} \mathbf{P}(X_{n,m} = 1) \le \frac{Cn^b}{n}.$$

Thus we may henceforth assume $n = ([n^a] + [n^b])B$.

We denote by $\mathcal{B}_k$ and $\mathcal{S}_k$ for $k = 1, \cdots, B$, the elements of $\{1, \ldots, n\}$ in big blocks and small blocks, respectively. Let

$$Y_{n,k} = \sum_{m\in\mathcal{B}_k} X_{n,m} \quad \text{and} \quad Z_{n,k} = \sum_{m\in\mathcal{S}_k} X_{n,m},$$

where for each n, both $(Y_{n,k}, 1 \le k \le B)$ and $(Z_{n,k}, 1 \le k \le B)$ are sequences of identically distributed random variables. Let

$$S'_n = \sum_{k=1}^{B} Y_{n,k} \quad \text{and} \quad S''_n = \sum_{k=1}^{B} Z_{n,k}.$$

Similar to the argument for L, we have

$$\mathbf{E}S''_n \le C\frac{n^{1-b}n^a}{n} \le C\frac{1}{n^{b-a}}$$

so that (S''_n) converges in probability to zero. Thus we can ignore small blocks, and it is enough to show that (S'_n) converges to a Poiss(λ).

By (5.3), for $j < k$,

$$\mathrm{Cov}(Y_{n,j}, \prod_{i=k}^{\ell} Y_{n,i}) \le C\theta^{n^a} \tag{5.4}$$

which implies that for some $\bar{\theta} < 1$ (since ℓ grows at most linearly)

$$\left| \mathbf{E}\prod_{i=k}^{\ell} Y_{n,i} - \prod_{i=k}^{\ell} \mathbf{E}Y_{n,i} \right| \le C\bar{\theta}^{n^a}. \tag{5.5}$$

Next we note that (5.2) implies

$$\mathbf{P}(Y_{n,1} \ge 2) \le \sum_{m=1}^{[n^b]} \left([n^b] - m\right) \left(\frac{\lambda^2}{n^2} + o\left(\frac{1}{n}\right)\theta^m\right) \le Cn^{2b-2} + o(n^{b-1}).$$

Since $B \simeq n^{1-b}$, we have that $\sum_{k=1}^{B} \mathbf{P}(Y_{n,k} \ge 2)$ goes to zero as $n \to \infty$. By subtracting off the portion of each $Y_{n,k}$ which takes value greater than one, we may henceforth assume without loss of generality that each $Y_{n,k}$ takes value 1 or 0 only (one can check that 5.5 still holds for such modified random variables). We have

$$\mathbf{P}(Y_{n,1} = 1) = \mathbf{E}(Y_{n,1}) \simeq n^b p_n \simeq \lambda n^{b-1}.$$

We can combine this with (5.5) and apply the moment method to $\left(\sum_{i=1}^{B} Y_{n,i}, n \in \mathbb{N}\right)$ in order to conclude our result. □

6 Random Walk

Recall from Example 2.37 that if $S_n = X_1 + \cdots + X_n$, where $\{X_i\}$ are i.i.d. Rademacher random variables, then $(S_n, n \in \mathbb{N})$ is called a Simple Random Walk (SRW) on the integer lattice $\mathbb{Z}$. This in one of the most important models in probability, and there are countless research papers and even whole books written about this model. In Example 2.37, we proved one of the basic properties of SRW in the case that the X_i have distribution Rad(p) for $p > \frac{1}{2}$ (recall that we used Stirling's Formula combined with the first Borel-Cantelli Lemma to show that $\mathbf{P}(S_n = 0 \text{ i.o.}) = 0$). In the rest of this chapter we show some other basic properties of SRW on general integer lattices $\mathbb{Z}^d$ with standard basis $\{e_1, \ldots, e_d\}$. This is the process $(S_n, n \in \mathbb{N})$ where

$$S_n = X_1 + \cdots + X_n \tag{6.1}$$

and $\{X_i, i \in \mathbb{N}\}$ are i.i.d. random vectors taking the values $\{\pm e_i, 1 \leq i \leq d\}$ with some probability distribution.

In the case, that $\{X_i\}$ are i.i.d. with some other distribution, we call (S_n) a Random Walk In this chapter, we focus on SRW, but most of these properties have analogs for more general Random Walks, and we refer the interested reader to [LL10] for details.

6.1 Recurrence and Transience

Let us start by defining the phenomenon occurring in Example 2.37.

Definition 6.1. A Random Walk $(S_n, n \in \mathbb{N}_0)$ is said to be *recurrent* if

$$\mathbf{P}(S_n = 0 \text{ i.o.}) = 1$$

and it is said to be *transient* otherwise.

To tie up a loose end from Example 2.37, let us consider the case when $p = \frac{1}{2}$. Then $(S_n, n \in \mathbb{N}_0)$ is called a Simple *Symmetric* Random Walk (SSRW) on $\mathbb{Z}$.

Proposition 6.2. *A Simple Symmetric Random Walk on $\mathbb{Z}$ is recurrent.*

Proof. First, define $N := \sum_{n=1}^{\infty} \mathbf{1}_{\{S_n = 0\}}$. We then have that

$$\begin{aligned}
\mathbf{E}N &= \sum_{n=1}^{\infty} \mathbf{E}\mathbf{1}_{\{S_n = 0\}} \\
&= \sum_{n=1}^{\infty} \mathbf{P}(S_{2n} = 0) \\
&\geq c \sum_{n=1}^{\infty} \frac{1}{\sqrt{2n}} \\
&= \infty.
\end{aligned}$$

In this case, Stirling's Formula implies that

$$\mathbf{P}(S_{2n} = 0) \simeq \frac{c}{\sqrt{2n}}.$$

This suggests that maybe (S_n) visits the origin infinitely often a.s., but we need more information about how this expectation becomes infinite.

We therefore calculate

$$\begin{aligned}
\mathbf{P}(N \geq 2) &= \sum_{m=1}^{\infty} \mathbf{P}(S_1 \neq 0, \ldots, S_{m-1} \neq 0, S_m = 0, S_n = 0, \text{ for some } n > m) \\
&= \sum_{m=1}^{\infty} \mathbf{P}(S_1 \neq 0, \ldots, S_{m-1} \neq 0, S_m = 0, S_n - S_m = 0 \text{ for some } n > m) \\
&= \sum_{m=1}^{\infty} \mathbf{P}(S_1 \neq 0, \ldots, S_{m-1} \neq 0, S_m = 0)\mathbf{P}(\tilde{S}_{n-m} = 0 \text{ where } n > m).
\end{aligned}$$

where in the last line $(\tilde{S}_n)$ is an independent copy of (S_n). This last step follows by the independence of the increments of SRW, and the calculation here is an example of the so-called *Markov property* (see Section 6.3). Continuing we see that the above equals

$$\begin{aligned}
\sum_{m=1}^{\infty} \mathbf{P}(S_1 \neq 0, \ldots, S_{m-1} \neq 0, S_m = 0) \cdot \mathbf{P}(N \geq 1) &= \mathbf{P}(N \geq 1) \cdot \mathbf{P}(N \geq 1) \\
&= (\mathbf{P}(N \geq 1))^2.
\end{aligned}$$

The proof can easily be generalized to show that $\mathbf{P}(N \geq k) = (\mathbf{P}(N \geq 1))^k$.

There are now two possibilities for $\mathbf{P}(N \geq 1)$. Either (a) $\mathbf{P}(N \geq 1) < 1$ or (b) $\mathbf{P}(N \geq 1) = 1$. If (a) is true, then:

$$\mathbf{E}N = \sum_{k=1}^{\infty} \mathbf{P}(N \geq k) = \sum_{k=1}^{\infty} (\mathbf{P}(N \geq 1))^k < \infty,$$

which contradicts our observation that $\mathbf{E}N = \infty$. Hence, (b) must be true, in which case $\mathbf{P}(N \geq k) = 1^k = 1$ for all k. Therefore, $N = \infty$ a.s. so that (S_n) visits the origin infinitely often, almost surely. □

Remark 6.3. We can now see why the convergence of $(S_n/\sqrt{n}, n \in \mathbb{N})$ in the CLT cannot be in the almost sure sense. This is by observing that the above result, together with the CLT, implies that almost surely, every point of $\mathbb{R}$ is a limit point of $(S_n/\sqrt{n}, n \in \mathbb{N})$.

In general dimension d, if each X_i takes the values in $\{\pm e_i, 1 \leq i \leq d\}$ with equal probability $\frac{1}{2d}$ then we call (S_n) the SSRW on $\mathbb{Z}^d$. We have the following characterization of recurrence and transience:

Theorem 6.4. *Simple Symmetric Random Walk on $\mathbb{Z}^d$ is recurrent for $d = 1, 2$ and transient for $d \geq 3$.*

Proof. We have already proved recurrence for $d = 1$. For $d \geq 2$, we use the dichotomy of the two cases (a) and (b) presented in the proof in $d = 1$ after which we saw it was enough to figure out if $\mathbf{E}N$ was finite or infinite (where N is the number of visits to 0). For $d = 2$ we have

$$\begin{aligned}\mathbf{E}N &= \sum_{n=1}^{\infty} \mathbf{P}(S_{2n} = 0) \\ &= \sum_{n=1}^{\infty} \binom{2n}{n} \sum_{k=0}^{n} \frac{1}{4^{2n}} \binom{n}{k} \binom{n}{n-k} \qquad = \sum_{n=1}^{\infty} \binom{2n}{n} \frac{1}{4^{2n}} \binom{2n}{n}\end{aligned}$$

The second to last line follows by thinking of up and right as positive directions, and down and left as negative directions. To return to the origin, it must be that n of the $2n$ steps are in negative directions. For the inside summation, one conditions on there being $2k$ horizontal steps and notes that there must be exactly k positive steps and k negative steps. The last line follows since choosing m elements from $2n$ elements can be done be choosing k from the first n and then choosing $m - k$ from the next n (then set $m = n$). Applying Stirling's formula to the last line shows the above expectation to be similar to $\sum_{n\geq 1} 1/n$ which is infinite.

For $d = 3$,

$$\begin{aligned}\mathbf{E}N &= \sum_{n=1}^{\infty} \mathbf{P}(S_{2n} = 0) \\ &= \sum_{n=1}^{\infty} \binom{2n}{n} \sum_{k=0}^{n-j} \sum_{j=0}^{n} \frac{1}{6^{2n}} \binom{n}{j,k} \binom{n}{j,k}\end{aligned}$$

and Stirling's formula can be used to show that this is finite.

Finally, for $d > 3$, we represent the Random Walk by the d-dimensional vector $(X_1(n), \ldots, X_d(n))_{n\in\mathbb{N}_0}$ where the subscripts are the various coordinates and time is now in the argument. Consider only the subsequence of times $(n_k, k \in \mathbb{N}_0)$ at which a move in the first three coordinates is chosen, i.e., X_{n_k} takes a value $\{\pm e_1, \pm e_2, \pm e_3\}$, then $(X_1(n_k), X_2(n_k), X_3(n_k))_{k\in\mathbb{N}_0}$ is precisely a 3-dimensional SSRW. This returns to $(0, 0, 0)$ finitely many times a.s., thus the d-dimensional SSRW must also return to the origin finitely many times, almost surely. □

6.2 Stopping Times

The calculations involved in recurrence and transience of SRW were associated to the kth time that (S_n) hit the origin. In general, when using SRW as a model, we are often interested in the first time that the SRW hits other subsets of Z or $\mathbb{Z}^d$. For instance, if we use SRW as a model for movements in the stock market, we might be interested in the first time the Standard and Poor's index hits the level 2500. We introduce these sorts of random times in this section, and prove a fundamental theorem using them in the next section.

Definition 6.5. A *filtration* associated to a SRW $(S_n, n \in \mathbb{N}_0)$ (or more generally any discrete-time stochastic process $(S_n, n \in \mathbb{N}_0)$), is a sequence of increasing σ-fields:

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \cdots$$

such that $S_n \in \mathcal{F}_n$ for all $n \in \mathbb{N}_0$. In particular, the *minimal* filtration for (S_n) is given by letting $\mathcal{F}_n = \sigma(S_1, \ldots, S_n)$.

Given a filtration $(\mathcal{F}_n, n \in \mathbb{N}_0)$, a random variable τ taking values in $\mathbb{N}_0 \cup \{\infty\}$ and satisfying

$$\{\tau = n\} \in \mathcal{F}_n \tag{6.2}$$

is called a *stopping time*.

Remark 6.6. One typically interprets a σ-field as a collection of events for which one has knowledge about. If $Y \in \mathcal{F}_n$ then we interpret this as saying we are able to know what value Y takes by time n, and thus we may measure the probability of events such as

$$\{Y = c\}, \{Y < c\}, \text{ and } \{Y > c\},$$

i.e., these events should be measurable by time n[15]. On the other hand, if Y is random variable whose measurement depends on things that happen after time n, then $\{Y = c\}, \{Y < c\}$, and $\{Y > c\}$ should not be measurable by time n. This motivates why a filtration should be increasing, since as time passes we gain more and more knowledge. Due to the increasing nature of a filtration, the condition $\{\tau = n\} \in \mathcal{F}_n$ is equivalent to $\{\tau \leq n\} \in \mathcal{F}_n$ (one can easily check this), and we will often use and/or verify the latter condition instead.

Throughout the rest of this section we will assume that $(\mathcal{F}_n)$ is a filtration for (S_n) and τ is a stopping time with respect to $(\mathcal{F}_n)$.

Example 6.7. Perhaps the easiest example of a stopping time τ is the "first time that $S_n = 0, n \geq 1$" assuming that $S_0 = 0$:

$$\tau = \inf\{k > 0 : S_k = 0\}.$$

In other words, $\tau(\omega) = n$ if $S_i(\omega) \neq 0$ for all $0 < i < n$ and $S_n(\omega) = 0$. In the case $S_i(\omega) \neq 0$ for all $i > 0$, then $\tau(\omega) = \infty$. Clearly $\{\tau \leq n\} \in \mathcal{F}_n$ since $\mathcal{F}_n$ includes $\sigma(S_1, \ldots, S_n)$.

A generalization of the above stopping time for SRW on $\mathbb{Z}^d$ is

$$\tau = \inf\{k > 0 : S_k \in A\}$$

for some subset $A \subset \mathbb{Z}^d$. In this case, we also call τ the first *hitting time* of the set A.

[15] Philosophically speaking, this means we were able to record whether these events took place n steps into the future under very similar circumstances (which all occurred in the past), and thus we can assign a probability to these event occurring again n steps into the future under the current similar circumstances.

Example 6.8. An example of a stopping time which is not a hitting time is the following for SRW on $\mathbb{Z}$:

$$\tau = \inf\{k > 0 : S_k = 1, S_j = 0 \text{ for some } 0 < j < k\}$$

which is the first time that (S_n) hits 1 given that it has already returned to the origin at some positive time. Of course, we can also create increasingly elaborate stopping times by simply adding conditions that (S_n) must satisfy before it hits some set A.

Example 6.9. Let us illustrate that not all "random times" are stopping times since they may in general violate (6.2). Consider SRW on $\mathbb{Z}$ and let

$$T = \inf\{k > 0 : S_k = 0, S_k + 1 = 0\}$$

which is the first time that (S_n) hits 0 such that it is at 1 at its next step. If $(\mathcal{F}_n)$ is the minimal filtration for (S_n), then the event $\{T = n\} \notin \mathcal{F}_n$ although it is true that $\{T = n\} \in \mathcal{F}_{n+1}$. In particular, if we are using the minimal filtration, then the evaluation of a stopping time (the ability to check whether $T = n$) can only use the knowledge of the process $(S_k, k \in \mathbb{N}_0)$ up to time n.

By now, one should see that stopping times are special random times that use only information up to time n in order to be evaluated. As such, they are frequently used as a random index for the process (S_n). For instance, in Example 6.7 where τ is the hitting time of 0, one has $S_\tau = 0$. Since τ could take arbitrarily large values and can even be infinite with positive probability (as is the case of asymmetric SRW on $\mathbb{Z}$), it could be that $S_\tau \notin \mathcal{F}_n$ for all $n \in \mathbb{N}_0$. This motivates a new σ-field; the *stopping time σ-field* $\mathcal{F}_\tau$ associated to a stopping time τ is the collection of all events A such that

$$A \cap \{\tau = n\} \in \mathcal{F}_n.$$

6.3 The Markov and Martingale Properties

One of the most important properties that SRW possesses is the Markov property:

Definition 6.10. A discrete-time stochastic process $(S_n, n \in \mathbb{N}_0)$ with filtration $(\mathcal{F}_n, n \in \mathbb{N}_0)$ on a measurable space $(E, \mathcal{F})$ satisfies the *Markov property* if

$$\mathbf{P}(S_n \in A | \mathcal{F}_{n-1}) = \mathbf{P}(S_n \in A | \sigma(S_{n-1}))$$

for all $A \in \mathcal{F}$. In words, the probability that $S_n \in A$ depends only on its most recent location S_{n-1}, and not on its whole history $S_1, \ldots, S_n$.

Exercise 6.1. Show that the independent increments property: $S_n - S_{n_1} \perp\!\!\!\perp S_{n_1}$ implies the Markov property. In particular, SRW has the Markov property.

Exercise 6.2. Show that the Markov property is equivalent to:

$$\mathbf{E}\left(f(S_n)|\mathcal{F}_{n-1}\right) = \mathbf{E}(f(S_n)|\sigma(S_{n-1}))$$

for all bounded and measurable $f : \mathcal{S} \to \mathbb{R}$.

Since we will investigate more general processes with this property in a later chapter, let us quickly move on to another extremely important property. In the symmetric case, SSRW possesses the martingale property:

Definition 6.11. A discrete-time stochastic process $(S_n, n \in \mathbb{N}_0)$ with filtration $(\mathcal{F}_n, n \in \mathbb{N}_0)$ on a measurable space $(E, \mathcal{F})$ is a *martingale* with respect to $(\mathcal{F}_n, n \in \mathbb{N}_0)$ if

$$\mathbf{E}(S_{n+1}|\mathcal{F}_n) = S_n \tag{6.3}$$

In words, our best guess (expectation) of S_{n+1} using only information up to time n is S_n.

Remarks 6.12.

1. A consequence of (6.3) follows by taking the expectation of each side:

$$\mathbf{E}S_{n+1} = \mathbf{E}S_n = \cdots = \mathbf{E}S_0.$$

2. If the equals sign in (6.3) is replace by a $\leq$, then the process is called a supermartingale. If it is replaced by a $\geq$, then the process is a submartingale. The sub- and super- prefixes may seem to be intuitively reversed, but they are chosen in this way due to the relationship these processes have with subharmonic and superharmonic functions.

Exercise 6.3. Show that the SSRW on $\mathbb{Z}^d$ is a martingale.

Example 6.13. A first example of a nontrivial martingale is $(S_n^2 - n, n \in \mathbb{N})$ where (S_n) is SSRW on $\mathbb{Z}$ with ith increment X_i. We verify

$$\begin{aligned}\mathbf{E}(S_{n+1}^2|\mathcal{F}_n) &= \mathbf{E}(S_n^2 + 2S_nX_{n+1} + X_{n+1}^2|\mathcal{F}_n) \\ &= S_n^2 + 2S_n\mathbf{E}X_{n+1} + \mathbf{E}X_{n+1}^2\end{aligned}$$

where the conditioning is not present in the last line since $X_{n+1} \perp\!\!\!\perp \mathcal{F}_n$. Since $\mathbf{E}X_i = 0$ by symmetry, the martingale property is proved.

We now have a fundamental result

Theorem 6.14 (Stopping Time Theorem)**.** *If $(S_n, n \in \mathbb{N}_0)$ is a martingale and τ is a stopping time, both with respect to $(\mathcal{F}_n, n \in \mathbb{N}_0)$, then $(S_{n\wedge\tau}, n \in \mathbb{N}_0)$ is a martingale with respect to $(\mathcal{F}_n)$.*

Remark 6.15. In the case, that τ is bounded by some value B we have that $\tau \wedge n = \tau$ for all $n > B$, thus by Remark 6.12, $\mathbf{E}S_\tau = \mathbf{E}S_0$. This is not true when τ is unbounded since we may let $S_0 = 0$ and consider τ to be the first time that (S_n) hits some value $x \neq 0$.

We will use the following lemma

Lemma 6.16 (Martingale transformation lemma)**.** *Suppose* $(S_n, n \in \mathbb{N}_0)$ *is a martingale with respect to* $(\mathcal{F}_n, n \in \mathbb{N}_0)$ *and* $(Y_n, n \in \mathbb{N}_0)$ *is a sequence of bounded random variables satisfying* $Y_n \in \mathcal{F}_{n-1}, n \in \mathbb{N}$*. If* $M_0 = S_0$ *and*

$$M_n = M_0 + \sum_{k=1}^{n} Y_k(S_k - S_{k-1}),$$

then $(M_n, n \in \mathbb{N}_0)$ *is a martingale with respect to* $(\mathcal{F}_n, n \in \mathbb{N}_0)$*.*

Proof. Clearly $M_n \in \mathcal{F}_n$ so that $(\mathcal{F}_n)$ is a filtration for (M_n). We just need to check (6.3). Since Y_n is bounded, the conditional expectations make sense, and

$$\begin{aligned}\mathbf{E}(M_{n+1} - M_n|\mathcal{F}_n) &= \mathbf{E}(Y_n(S_n - S_{n-1})|\mathcal{F}_n)\\ &= Y_n\mathbf{E}(S_n - S_{n-1})|\mathcal{F}_n)\\ &= 0\end{aligned}$$

where the last line follows since (S_n) is a martingale with respect to $(\mathcal{F}_n)$. □

Proof of the Stopping Time Theorem. We can without loss of generality suppose that $S_0 = 0$, since otherwise we can simply subtract off the random variable S_0. Let

$$Y_k = \mathbf{1}_{\{\tau \geq k\}} = 1 - \mathbf{1}_{\{\tau < k\}}$$

so that

$$S_{\tau \wedge n} = S_\tau \mathbf{1}_{\{\tau < n\}} + S_n \mathbf{1}_{\{\tau \geq n\}} = \sum_{k=1}^{n} Y_k(S_k - S_{k-1}).$$

The theorem follows from applying the martingale transformation lemma. □

6.4 Large Deviations

We know that as $n \to \infty$, $\left(\frac{S_n - \mu n}{\sigma\sqrt{n}}, n \in \mathbb{N}\right)$ converges in distribution to a $N(0,1)$ where $S_n = X_1 + \cdots + X_n$. We now look at events $\{S_n \geq \mu + an, n \in \mathbb{N}\}$, where a is not necessarily the mean, and analyze

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}(S_n \geq an) = -I(a) < 0,$$

where $I(a) > 0$ is called the *rate function* of a Large Deviation Principle (LDP).

Theorem 6.17. *If* $\{X_i, i \in \mathbb{N}\}$ *are i.i.d. with a* $\mathrm{Ber}(\frac{1}{2})$ *distribution, then the rate function is given by*

$$I(a) = \begin{cases} \log 2 + a \log a + (1-a)\log(1-a) & \text{if } a \in [0,1] \\ \infty & \text{if } a > 1 \end{cases}$$

Proof. The cases $a = \frac{1}{2}$, $|a| > 1$, $|a| = 1$ are clear. For $a \in [0,1]$, we need only consider $a \in (\frac{1}{2}, 1)$ by symmetry. Let

$$M_n(a) = \max_{k \geq an} \binom{n}{k}.$$

Since $\mathbf{P}(S_n \geq an) = 2^{-n} \sum_{k \geq an} \binom{n}{k}$,

$$2^{-n} M_n(a) \leq \mathbf{P}(S_n \geq an) \leq (n+1) 2^{-n} M_n(a).$$

Take the log of both sides and divide by n. Letting $I(a) = -\frac{1}{2} \log M_n(a)$ gives the result. □

Remark 6.18. Fix $\epsilon > 0$. Then $\displaystyle\sum_{n=1}^{\infty} \mathbf{P}\left(\left| \frac{S_n}{n} - \frac{1}{2} \right| > \epsilon \right) < \infty$. By the Borel-Cantelli Lemma, this implies the SLLN. Also, at the mean $a = \frac{1}{2}$, we have $I(\frac{1}{2}) = 0$. One can also show that the curvature at $a = \frac{1}{2}$ is related to the CLT.

Definition 6.19. The moment-generating function of a random variable X is $\psi(t) = \mathbf{E}\, e^{tX}$ whenever this expectation exists.

Note that if $\psi(t) < \infty$ for all t, then $\psi(t) \in C^\infty(\mathbb{R})$, $\frac{d}{dt}\psi(t)\big|_{t=0} = \mathbf{E}X$, $\frac{d^2}{dt^2}\psi(t)\Big|_{t=0} = \mathbf{E}X^2$ and so on.

Theorem 6.20. *Suppose $\{X_i, i \in \mathbb{N}\}$ are i.i.d. with $\psi(t) := \psi_{X_1}(t) < \infty$ for all $t \in \mathbb{R}$, and $\mathbf{E}X_1 < 0$. Then*

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbf{P}(S_n \geq 0) = \inf_{t \in \mathbb{R}} \log \psi(t)$$

Corollary 6.21. *Let $Y_i = X_i + a$, so that $\mathbf{E}Y_1 < a$, and denote $\hat{S}_n = \sum_{i=1}^n Y_i$. Then*

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbf{P}(\hat{S}_n \geq an) = -\sup_{t \in \mathbb{R}} (at - \log \psi(t))$$

Proof of Theorem 6.20. If $\mathbf{P}(X_1 = c) = 1$ for some c then the proof follows from direct verification. If not, we have three cases to consider:

1. $\mathbf{P}(X_1 < 0) = 1$. Note that $\psi(t)$ is always decreasing and

   $$\inf_{t \in \mathbb{R}} \psi(t) = \lim_{t \to \infty} \psi(t) = 0$$

   which makes both sides of the equality in the theorem $-\infty$.

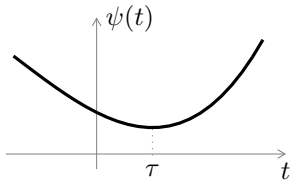
2. $\mathbf{P}(X_1 \leq 0) = 1$, $\mathbf{P}(X_1 = 0) > 0$. Here $\psi(t)$ is again always decreasing and

   $$\inf_{t \in \mathbb{R}} \psi(t) = \lim_{t \to \infty} \psi(t) = \mathbf{P}(X_1 = 0).$$

   On the other hand

   $$\mathbf{P}(S_n \geq 0) = \mathbf{P}(S_n = 0) = (\mathbf{P}(X_1 = 0))^n,$$

   from which statement of the theorem follows immediately.



3. $\mathbf{P}(X_1 > 0) > 0$, $\mathbf{P}(X_1 < 0) > 0$. Analysis of $\psi(t)$ yeilds: $\psi''(t) > 0$ for all t and $\psi'(0) < 0$. Additionally, $\lim_{t\to\pm\infty} \psi(t) = \infty$. Therefore $\psi(t)$ has a unique minimum at some point $\tau > 0$.

   We will prove the theorem statement in two steps. First by Chebyshev's inequality

$$\mathbf{P}(S_n \geq 0) = \mathbf{P}(e^{\tau S_n} \geq 1) \leq \frac{\mathbf{E}\, e^{\tau S_n}}{1} = (\psi(\tau))^n$$

   which renders

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}(S_n \geq 0) \leq \inf_{t\in\mathbb{R}} \log \psi(t).$$

   To prove the other inequality we define the *Cramer transform* (under current assumptions on X_1 and $\psi(t)$) of a random variable X to be a random variable $\hat{X}$ with distribution $\mu_{\hat{X}}$ with Radon-Nikodim derivative $\frac{d\mu_{\hat{X}}}{d\mu_X} = \frac{e^{\tau X}}{\psi(\tau)}$. Note how the definition depends on τ. Also $\mathbf{E}\hat{X} = 0$:

$$\psi_{\hat{X}}(t) = \int_{\mathbb{R}} e^{tx}\, d\mu_{\hat{X}} = \frac{1}{\psi(\tau)} \int_{\mathbb{R}} e^{(t+\tau)x}\, d\mu_X = \frac{\psi(\tau+t)}{\psi(\tau)}$$

   and $\mathbf{E}\hat{X} = \psi'_{\hat{X}}(0) = \frac{\psi'(\tau+t)}{\psi(\tau)} = 0$. Similarly

$$\mathbf{E}\hat{X}^2 = \psi''_{\hat{X}}(0) = \frac{\psi''(\tau+t)}{\psi(\tau)} \in (0, \infty).$$

   Now

$$\begin{aligned}
\mathbf{P}(S_n \geq 0) &= \int_{S_n\geq 0} \mu_X(dx_1) \dots \mu_X(dx_n) \\
&= \psi(\tau)^n \int_{S_n\geq 0} e^{-\tau(x_1+\dots+x_n)} \mu_{\hat{X}}(dx_1) \dots \mu_{\hat{X}}(dx_n) \\
&= \psi(\tau)^n \mathbf{E}\left(e^{-\tau \hat{S}_n} \mathbf{1}_{\{\hat{S}_n \geq 0\}}\right).
\end{aligned}$$

   It remains to show that the last expectation cannot decay too quickly as n gets large. Let $\hat{\sigma}$ be the standard deviation of $\hat{X}$. Then

$$\mathbf{E}\left(e^{-\tau \hat{S}_n} \mathbf{1}_{\{\hat{S}_n \geq 0\}}\right) \geq e^{-\tau 2\hat{\sigma}\sqrt{n}} \mathbf{P}(0 \leq \hat{S}_n \leq 2\hat{\sigma}\sqrt{n}) \geq \frac{1}{4} e^{-\tau 2\hat{\sigma}\sqrt{n}}$$

for large enough n (by the CLT). Therefore

$$\mathbf{P}(S_n \geq 0) \geq \psi(\tau)^n \frac{1}{4} e^{-\tau 2\hat{\sigma}\sqrt{n}}$$

which after taking the logarithm and dividing by n completes the proof.

Remarks 6.22.

1. $\psi(0) = 1$, and if $\mathbf{E}X_1 < a$ and $\psi'(0) < 0$ then $I(a) = -\inf_{t\in\mathbb{R}} \log \psi(t) > 0$.
2. One can also obtain exponential decay rates of $\mathbf{P}(S_n \geq an^\beta)$ for $1/2 < \beta < 1$ which are called *moderate deviations*.
3. If $X \sim N(0,1)$ then

$$\psi(t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{tx} e^{-\frac{x^2}{2}}\, dx = e^{\frac{t^2}{2}}$$

Exercise 6.4. Show that if $\mathbf{E}X_1 = a$ then $I(a) = 0$.

Example 6.23 (Exponential Tilting). Let $X \sim N(0,1)$ and $f(x) = \mathbf{1}_{\{x\geq 10\}}$. We will try to estimate $\mathbf{E}f(X) = \frac{1}{\sqrt{2\pi}} \int_{10}^{\infty} e^{-x^2/2}\, dx$. One way to do this is to use the SLLN: let $X_1, X_2, \ldots$ be i.i.d. $N(0,1)$, then

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i) \to \mathbf{E}f(X),$$

but this naive way is hard to carry out: $f(X_1) = 0$ with probability about $1 - 7.6 \cdot 10^{-24}$, and a *very large* n is required for an accurate estimate, as well as a high quality random number generator.

Instead we do the following:

$$\begin{aligned} \mathbf{E}f(X) &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x) e^{-\frac{x^2}{2}}\, dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x) e^{-\frac{(x-10)^2}{2}} e^{\frac{10^2}{2}} e^{-10x}\, dx \\ &= \mathbf{E}\left[f(Y)\, e^{50-10Y}\right] = \mathbf{E}\left[\mathbf{1}_{\{Y\geq 10\}}\, e^{50-10Y}\right] \end{aligned}$$

where $Y \sim N(10,1)$. Now we take $Y_1, Y_2, \ldots$ as i.i.d. $N(10,1)$ and again by the SLLN

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{Y_i\geq 10\}}\, e^{50-10Y_i} \to \mathbf{E}f(X),$$

and this is a realistically computable estimate.

□

7 Brownian Motion

7.1 Construction of the Process

Definition 7.1 (Gaussian Vector)**.** A vector $\mathbf{X} \in \mathbb{R}^d$ is *Gaussian* with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma_{d\times d}$ if for all $\theta \in \mathbb{R}^d$, the inner product $\theta \cdot \mathbf{X}$ is a one-dimensional Gaussian random variable.

Note that this is much stronger than simply saying each marginal of

$$\mathbf{X} = (X_1, \ldots, X_d)$$

is Gaussian. This can be seen from Example 2.9.

Proposition 7.2. *If $\mathbf{X}$ is a Gaussian vector in $\mathbb{R}^n$ then for any $m \times n$ matrix A, the random vector $\mathbf{Y} = A\mathbf{X}$ is a Gaussian vector in $\mathbb{R}^m$.*

Proof. For any $\theta \in \mathbb{R}^m$, $\theta\cdot\mathbf{Y} = (\theta A)\cdot\mathbf{X}$. Since the right side is a one-dimensional Gaussian, the result follows. □

Definition 7.3. The *characteristic function* of a random vector $\mathbf{X}$ is

$$\varphi_{\mathbf{X}}(\theta) := \mathbf{E}e^{i(\theta\cdot\mathbf{X})}.$$

Exercise 7.1. Check that a Gaussian vector has

$$\varphi(\theta) = \exp\left[i\theta^T\mu - \frac{1}{2}\theta^T\Sigma\theta\right].$$

Proposition 7.4. *If $\mathbf{X}$ is random vector with mean vector μ and covariance satisfying $det(\Sigma) \neq 0$ then $\mathbf{X}$ is Gaussian if and only if its joint density is*

$$f(\mathbf{x}) = \frac{1}{\sqrt{(\det\Sigma)(2\pi)^d}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right]. \tag{7.1}$$

Proof. ($\Rightarrow$) Suppose $Z := \theta\cdot\mathbf{X}$ is Gaussian with $\mathbf{E}Z = \theta^T\mu$. Also,

$$\operatorname{Var} Z = \operatorname{Var}(\theta_1X_1 + \cdots + \theta_nX_n) = \theta^T\Sigma\theta,$$

where Σ is the covariance matrix of $\mathbf{X}$. Then

$$\varphi_Z(1) = \mathbf{E}e^{iZ} = e^{i\theta^T\mu - \frac{1}{2}\theta^T\Sigma\theta}$$

Thus $\mathbf{X}$ has the desired density (apply Exercise 7.1 to a Gaussian vector with i.i.d. $N(0,1)$ marginals).

($\Leftarrow$) Just reverse the steps. □

Proposition 7.5. *If $\mathbf{X}$ is Gaussian, then its coordinate random variables are independent if and only if they are uncorrelated.*

Proof. The "only if" direction is obvious. For the "if" direction, note that when Σ is a diagonal matrix, the density in (7.1) is a product of the marginal densities. □

Exercise 7.2. Show that if $\mathbf{X}$ is a standard Gaussian vector in $\mathbb{R}^n$ (i.i.d. $N(0,1)$ marginals) and A is an $n \times n$ orthogonal matrix then $A\mathbf{X} \stackrel{\mathrm{d}}{=} \mathbf{X}$.

Definition 7.6 (Gaussian Process)**.** A process $(X_t, t \in [0,T])$ is a *Gaussian process* if all finite-dimensional distributions of (X_t) are jointly Gaussian, i.e., for $0 \le t_0 < \cdots < t_n \le T$, $(X_{t_0}, X_{t_1}, \ldots, X_{t_n})$ is a Gaussian vector. As we did with Poisson processes, we sometimes put time in the argument and write the process as $(X(t))$.

Remark 7.7. In this definition we think of $[0,T]$ as a time interval with the possibility that $T = \infty$. By Kolmogorov's Extension Theorem, the definition forms a well-defined probability measure on the space of functions, $\mathbb{R}^{[0,T]}$. The σ-field on $\mathbb{R}^{[0,T]}$ is the one that is generated by $A_{t_1} \times \cdots \times A_{t_n}$ which specifies that the function must pass through the Borel set A_{t_i} at time t_i. One can check that this generates sets in the σ-field which specify functions to pass through A_{t_i} for a countable collection of times $\{t_i\} \subset [0,T]$. On the other hand, the σ-field does not contain sets which specify functions at an uncountable collection of times.

Definition 7.8. Consider a Gaussian process $(B_t, t \in \mathbb{Q} \cap [0,T])$ with $\mathbf{E}B_t = 0$ and $\mathrm{Cov}(B_s, B_t) = s \wedge t$ for all $s, t \in [0,T]$ where $s \wedge t$ is the minimum of s and t. If we extend this to a process on $(B_t, t \in [0,T])$ which is a.s. continuous on $[0,T]$, then the process is called a *standard Brownian motion* (SBM) on $[0,T]$.

When $T < \infty$, SBM can alternatively be defined as a Gaussian random vector taking values in the Banach space $C[0,T]$, equipped with the supremum norm, satisfying $\mathbf{E}B_t = 0$ and $\mathrm{Cov}(B_s, B_t) = s \wedge t$ for all $s, t \in [0,T]$.

Remark 7.9. By the remark preceding this definition, we can see that the continuous functions on $[0,T]$ are not measurable with respect to the σ-field on $\mathbb{R}^{[0,T]}$ generated by finite-dimensional Borel sets. This is the reason for defining Brownian motion first on $\mathbb{Q} \cap [0,T]$ and then extending this to a distribution on the continuous functions on $[0,T]$.

Exercise 7.3. Show that in the definition of SBM, it is not enough to say "(B_t) is a.s. continuous for every $t \in [0,T]$". Hint: consider τ having an Exp(1) distribution independent of (B_t) and the process defined by

$$X(t) = \begin{cases} B(t); & t \neq \tau \\ 0; & t = \tau. \end{cases}$$

Definition 7.10 (Stationary and independent increments)**.** A process $(X_t, t \in [0,T])$ has *stationary and independent increments* if for any vector of times $\vec{t}$, the increments $(X_{t_{j+1}} - X_{t_j}, 1 \le j \le n)$ are independent and $X_t - X_s \stackrel{\mathrm{d}}{=} X_{t-s} - X_0$ for all $s, t \in [0,T]$.

The covariance property in the definition of SBM is often substituted by the condition that the process is Gaussian and has stationary and independent increments along with the normalization $\mathrm{Var}(B_1) = 1$. These are equivalent as we see next.

Proposition 7.11. *Suppose $(X(t), t \in [0,T])$ is a centered Gaussian Process (centered in the sense that $\mathbf{E}X(t) = 0$ for all $t \in [0,T]$). Then $(X(t))$ has stationary and independent increments with $\mathrm{Var}\, X(1) = 1$ if and only if $\mathrm{Cov}(X(s), X(t)) = s \wedge t$ for all $s, t \in [0,T]$.*

Proof. First the direction: ($\Longleftarrow$)

Assume $t > s$. To check the independence of increments consider the matrix A,

$$\begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \dots & -1 & 1 \end{pmatrix}$$

which gives

$$A(X(t_1), X(t_2), \dots, X(t_n)) = (X(t_2) - X(t_1), X(t_3) - X(t_2), \dots, X(t_n) - X(t_{n-1}).$$

Since $(X(t_1), X(t_2), \dots, X(t_n)$ is Gaussian it must be that

$$(X(t_2) - X(t_1), X(t_3) - X(t_2), \dots, X(t_n) - X(t_{n-1}))$$

is also Gaussian so to show independence of increments it is enough to show that increments are uncorrelated. Supposing $s < t \leq u < v$, we have

$$\begin{aligned} &\mathrm{Cov}(X(t) - X(s), X(v) - X(u)) \\ &= \mathrm{Cov}(X(t), X(v)) - \mathrm{Cov}(X(t), X(u)) - \mathrm{Cov}(X(s), X(v)) + \mathrm{Cov}(X(s), X(u)) \\ &= t - t - s + s. \end{aligned}$$

To show stationarity we see that

$$\begin{aligned} \mathrm{Var}(X(t) - X(s)) &= \mathrm{Var}(X(t)) + \mathrm{Var}(X(s)) - 2\,\mathrm{Cov}(X(s), X(t)) \\ &= t + s - 2s = t - s. \end{aligned}$$

Since the increments are independent, this shows $(X(t))$ has stationary increments.

Now the other direction: ($\Longrightarrow$)

First note that $\mathbf{E}X(1)^2 = 1$ implies by stationarity of increments that $\mathbf{E}X(s)^2 = s$. Now assume $t > s$,

$$\begin{aligned} &\mathrm{Cov}(X(t), X(s)) = \mathbf{E}[(X(t) - X(s) + X(s))(X(s))] \\ = \;& \mathbf{E}[(X(t) - X(s)(X(s) - X(0)] + \mathbf{E}X(s)^2 = 0 + s = s \wedge t. \end{aligned}$$

□

We can now construct a Standard Brownian Motion. We will do this in the case that $T < \infty$ so that we view SBM as a $C[0,T]$-valued random vector. Recall that if $f, g \in L^2[0,T]$ then Parseval's Identity tells us

$$\langle f, g\rangle = \int_0^T fg\, dm = \sum_{n\in\mathbb{N}} (\langle f, \psi_n\rangle\langle \psi_n, g\rangle)$$

where $(\psi_n, n \in \mathbb{N}_0)$ is an orthonormal basis of the space.

Consider now $f = \mathbf{1}_{[0,s]}$ and $g = \mathbf{1}_{[0,t]}$ with $0 \le s, t < T$. Then

$$\begin{aligned} s \wedge t &= \int_0^T \mathbf{1}_{[0,s\wedge t]}\, dm \\ &= \int_0^T fg\, dm \\ &= \sum_{n=0}^{\infty} \int_0^s \psi_n\, dm \int_0^t \psi_n\, dm. \end{aligned} \tag{7.2}$$

Let $\{Z_n, n \in \mathbb{N}_0\}$ be i.i.d. $N(0,1)$ and consider

$$B_t := \sum_{n=0}^{\infty} Z_n \int_0^t \psi_n\, dm. \tag{7.3}$$

A formal[16] calculation gives

$$\begin{aligned} \mathbf{E}(B_s B_t) &= \mathbf{E}\left(\sum_{n=0}^{\infty} Z_n \int_0^s \psi_n\, dm \sum_{k=0}^{\infty} Z_k \int_0^t \psi_k\, dm\right) \\ &= \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} \mathbf{E}\left[Z_n Z_k \int_0^s \psi_n\, dm \int_0^t \psi_k\, dm\right] \\ &= \sum_{n=0}^{\infty} \mathbf{E}\left(Z_n^2 \int_0^s \psi_n\, dm \int_0^t \psi_n\, dm\right) \\ &= s \wedge t, \end{aligned}$$

where the second to last equality comes from independence and the last equality comes from (7.2).

In order to show that the process defined by (7.3) is SBM, we need to rigorously check the above formal calculation, check that the process is Gaussian, and check that the process is a.s. continuous. For simplicity, let us fix $T = 1$ and the orthonormal basis $\psi_0 = 1$, $\psi_1 = \mathbf{1}_{[0,\frac{1}{2}]} - \mathbf{1}_{[\frac{1}{2},1]}$ on $[0,1]$ and

$$\psi_n = 2^{\frac{l}{2}} \psi_1(2^l t - k)$$

for $n = 2^l + k \ge 1$ with $0 \le k < 2^l$. It is not hard to check that this is indeed an orthonormal basis.

[16] "Formal" here means, formative in the sense that we do not check conditions needed for switching limits.

Lemma 7.12. *If $\{Z_n, n \in \mathbb{N}\}$ are i.i.d. $N(0,1)$, then there is a (finite) random variable Y such that*

$$|Z_n| \le Y\sqrt{\log n} \qquad \text{a.s. for all } n \ge 2.$$

Proof. For $x \ge 1$,

$$\begin{aligned}\mathbf{P}(|Z_n| \ge x) &= \frac{2}{\sqrt{2\pi}}\int_x^\infty e^{-u^2/2}du \le \sqrt{\frac{2}{\pi}}\int_x^\infty ue^{-u^2/2}du \\ &= \sqrt{\frac{2}{\pi}}e^{-x^2/2}\end{aligned}$$

Letting $\alpha > 1$,

$$\mathbf{P}(|Z_n| \ge \sqrt{2\alpha\log n}) \le \sqrt{\frac{2}{\pi}}e^{-\alpha\log n} = n^{-\alpha}\sqrt{\frac{2}{\pi}}$$

By Borel-Cantelli, $\mathbf{P}(|Z_n| \ge \sqrt{2\alpha\log n} \text{ i.o.}) = 0$ which also tells us that

$$M(\omega) := \sup_n \left\{\frac{|Z_n|}{\sqrt{\log(n)}} > 2\alpha\right\}$$

is an a.s. well-defined (or finite) random variable. Thus,

$$Y := \sup_{1\le n\le M}\frac{|Z_n|}{\sqrt{\log n}},$$

is a.s. finite. □

We can now establish the uniform absolute convergence in t of (7.3), which gives the a.s. continuity of (B_t). For large m and $L = 2^j \le m$, we have that a.s.

$$\begin{aligned}\sum_{n=m}^\infty |Z_n|\int_0^t \psi_n dm &\le \sum_{n=m}^\infty Y\sqrt{\log n}\int_0^t \psi_n dm \\ &\le \sum_{\ell=L}^\infty\sum_{k=0}^{2^\ell-1} Y\sqrt{\log(2^\ell+k)}\int_0^t \psi_{2^\ell+k}dm \\ &\le c\sum_{\ell=L}^\infty Y\sqrt{\ell+1}\ 2^{-(1+\ell/2)} \\ &\to 0 \qquad \text{as } L\to\infty.\end{aligned}$$

Next we check that (B_t) is a Gaussian process. Fix $t_1, \ldots, t_m$. We must

show that $(B_{t_1}, \ldots, B_{t_m})$ is a Gaussian vector:

$$\begin{aligned}
\mathbf{E}\exp\left(i\sum_{j=1}^{m}\theta_j B_{t_j}\right) &= \mathbf{E}\exp\left(i\sum_{j=1}^{m}\theta_j\left[\sum_{n=0}^{\infty}Z_n\int_0^{t_j}\psi_n dm\right]\right)\\
&= \prod_{n=0}^{\infty}\mathbf{E}\exp\left(iZ_n\underbrace{\sum_{j=1}^{m}\theta_j\int_0^{t_j}\psi_n dm}_{=:c_n}\right)\\
&= \prod_{n=0}^{\infty}\exp(-\frac{1}{2}c_n^2) = \exp\left(-\frac{1}{2}\sum_{n=0}^{\infty}c_n^2\right).
\end{aligned}$$

By Parseval's identity, we have that the above equals

$$\exp\left(-\frac{1}{2}\sum_{k=1}^{m}\sum_{j=1}^{m}\theta_j\theta_k t_j\wedge t_k\right),$$

which is the characteristic function of a Gaussian. This calculation also verifies our formal calculation that $\mathrm{Cov}(B_s, B_t) = s\wedge t$. This completes the construction.

7.2 Properties of Brownian Motion

Theorem 7.13. *$(B(t), t \geq 0)$ is a self-similar process with Hurst exponent $H = \frac{1}{2}$, i.e., $(B(t), t \geq 0)$ has the same distribution as $(c^{-1/2}B(ct), t \geq 0)$.*

Proof. We simply calculate

$$\begin{aligned}
\mathrm{Cov}\left(c^{-1/2}B(ct), c^{-1/2}B(cs)\right) &= 1/c[\mathrm{Cov}(B(ct), B(cs))]\\
&= [sc\wedge tc]/c\\
&= s\wedge t.
\end{aligned}$$

One can also easily see that $(c^{-1/2}B(ct), t \geq 0)$ is a.s. continuous. □

Exercise 7.4. (a) Show that

$$X(t) = \begin{cases} tB(1/t) & \text{if } t > 0\\ 0 & \text{if t=0}\end{cases}, \quad t \in [0, T]$$

defines a SBM on $[0, T]$.

(b) Show $(Y(s), s \geq 0)$, where $Y(s) = B(t_0 + s) - B(t_0)$, is a SBM.

A simple corollary of Exercise 7.4 is that a.s. $\lim_{t\to\infty} B(t)/t = 0$. To see this note that

$$\mathbf{P}\left(\lim_{t\to\infty}B(t)/t = 0\right) = \mathbf{P}\left(\lim_{t\to 0}X(t) = 0\right) = 1.$$

Proposition 7.14. $\mathbf{P}(\forall \varepsilon > 0, (B_t) \textit{ takes both signs in } (0, \varepsilon)) = 1$

Proof. Since

$$X(t) = \begin{cases} tB(1/t) & \text{if } t > 0 \\ 0 & \text{if t=0} \end{cases}$$

defines a SBM, we have that $(X(n), n \in \mathbb{N})$ is a symmetric Random Walk with $N(0,1)$ increments. A generalization of Proposition 6.2 shows that a one-dimensional Random Walk with finite, non-zero variance satisfies

$$\mathbf{P}(\limsup_{n\to\infty} X(n) = \infty) = \mathbf{P}(\liminf_{n\to\infty} X(n) = -\infty) = 1.$$

Thus

$$\mathbf{P}(\limsup_{n\to\infty}(nB(1/n)) = \infty) = \mathbf{P}(\liminf_{n\to\infty}(nB(1/n)) = -\infty) = 1.$$

□

Remark 7.15. This also shows that $\mathbf{P}[(B_t) \text{ is differentiable at } 0] = 0$ since $\lim_{n\to\infty} X(n) = \lim_{n\to\infty} B(1/n)/(1/n)$ does not exist almost surely. Moreover, by stationary increments:

$$\mathbf{P}\left(\lim_{t\to 0} \frac{B(t_0+t) - B(t_0)}{t} \text{ exists}\right) = 0 \quad \text{for all } t_0 \geq 0.$$

In other words for any fixed t_0, (B_t) is not differentiable at t_0 almost surely.

Now we prove a stronger result:

Theorem 7.16. $\mathbf{P}\left((B_t) \textit{ is nowhere differentiable}\right) = 1.$

Proof. It is enough to show this on $(0,1)$. For $m \geq 1$, $n \geq 4$ consider the event

$$A_{n,m} := \left\{\omega : \exists s \in \left(\frac{2}{n}, 1 - \frac{2}{n}\right) \text{ such that } |B(t) - B(s)| \leq m|t-s| \text{ for } t \text{ satisfying } |t-s| \leq \frac{1}{n}\right\}.$$

Note that

$$\left\{\omega : B(s,\omega) \text{ is differentiable at some } s \in (0,1)\right\}$$

is contained in $\cup_m \cup_n A_{n,m}$. We can enclose $A_{n,m}$ in

$$D_{n,m} := \left\{\omega : \exists k \text{ such that } 1 \leq k \leq n-2 \text{ and } \left|B\left(\frac{j}{n}\right) - B\left(\frac{j-1}{n}\right)\right| \leq \frac{2m}{n} \text{ for } j = k, k+1, k+2\right\},$$

using the following triangle inequality for $j = k, k+1, k+2$,

$$\left|B\left(\frac{j}{n}\right) - -B\left(s\right)\right| \leq \left|B\left(\frac{j}{n}\right) - B\left(\frac{j-1}{n}\right)\right| + \left|B\left(\frac{j-1}{n}\right) - B\left(s\right)\right|.$$

Then

$$\begin{aligned}\mathbf{P}(A_{n,m}) &\leq \mathbf{P}(D_{n,m}) \leq \sum_{k=1}^{n} \mathbf{P}\left(\left|B\left(\frac{k}{n}\right) - B\left(\frac{k-1}{n}\right)\right| \leq \frac{2m}{n}\right)^3 \\ &= n\mathbf{P}\left(\left|B\left(\frac{1}{n}\right)\right| \leq \frac{2m}{n}\right)^3 = n\mathbf{P}\left(|B(1)| \leq \frac{2m}{\sqrt{n}}\right)^3 \\ &\leq n\left(\frac{1}{\sqrt{2\pi}}\frac{4m}{\sqrt{n}}\right)^3 \xrightarrow{\text{as } n\to\infty} 0.\end{aligned}$$

Note that $A_{n,m} \subset A_{n+1,m}$, therefore $\mathbf{P}(\cup_m \cup_n A_{n,m}) = 0$. □

References

[Adl78] Robert J. Adler. Weak convergence results for extremal processes generated by dependent random variables. *The Annals of Probability*, 6(4):660–667, 1978.

[Ber13] Jakob Bernoulli. *Ars conjectandi*. 1713.

[Ber27] Serge Bernstein. Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Mathematische Annalen*, 97(1):1–59, 1927.

[Bir31] G. D. Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences U.S.A.*, 17(12):656–660, 1931.

[BK32] G. D. Birkhoff and B. O. Koopman. Recent contributions to the ergodic theory. *Proceedings of the National Academy of Sciences U.S.A.*, 18(3):279, 1932.

[Bre92] Leo Breiman. *Probability*. Society for Industrial and Applied Mathematics, 1992.

[Car22] Torsten Carleman. Sur le probleme des moments. *CR Acad. Sci. Paris*, 174:1680–1682, 1922.

[Dur10] R. Durrett. *Probability: theory and examples*. Cambridge Univ Pr, 2010.

[Fis10] Hans Fischer. *A history of the central limit theorem: From classical to modern probability theory*. Springer Science & Business Media, 2010.

[Gar65] Adriano M. Garsia. A simple proof of E. Hopf's maximal ergodic theorem. *Journal of Mathematics and Mechanics*, 14(3):381, 1965.

[IL71] I.A. Ibragimov and Y.V. Linnik. Independent and stationary sequences of random variables. 1971.

[JKK05] Norman L Johnson, Adrienne W Kemp, and Samuel Kotz. *Univariate discrete distributions*. John Wiley & Sons, 3rd edition, 2005.

[Kal02] Olav Kallenberg. *Foundations of modern probability*. Springer, 2002.

[Khi33] A Khintchine. Zu Birkhoffs lösung des ergodenproblems. *Mathematische Annalen*, 107(1):485–488, 1933.

[Lév25] Paul Lévy. *Calcul des probabilités*. Gauthier-Villars Paris, 1925.

[LL10] G.F. Lawler and V. Limic. *Random walk: a modern introduction*. Cambridge Univ Pr, 2010.

[Loè77] M. Loève. Probability theory I, 4th ed., 1977.

[MW43] Henry B. Mann and Abraham Wald. On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14(3):217–226, 1943.

[Pól20] Georg Pólya. Über den zentralen grenzwertsatz der wahrscheinlichkeitsrechnung und das momentenproblem. *Mathematische Zeitschrift*, 8(3):171–181, 1920.

[Res87] Sidney I. Resnick. *Extreme values, regular variation, and point processes*. Springer, 1987.

[RF10] Halsey L. Royden and Patrick Fitzpatrick. *Real analysis*. Pearson Education, 2010.

[Rud87] Walter Rudin. *Real and complex analysis*. McGraw-Hill Education, 1987.

[Sti89] Stephen M. Stigler. Francis galton's account of the invention of correlation. *Statistical Science*, pages 73–79, 1989.

[TK71] Amos Tversky and Daniel Kahneman. Belief in the law of small numbers. *Psychological bulletin*, 76(2):105, 1971.

[vB98] Ladislaus von Bortkiewicz. *Das Gesetz der kleinen Zahlen*. BG Teubner, 1898.