Graduate Probability Theory

Yiqiao YIN
Statistics Department
Columbia University
Notes in LaTeX

December 12, 2017

Abstract

This is the lecture note from Probability Theory class offered in Mathematics Department at Columbia University. The materials come from conventional graduate level probability text book, *Probability: Theory and Examples*, by Rick Durrett, and *Notes in Probability Theory*, by Varadhan. The document lands on Professor Ivan Corwin's work in Q-TASEP. [1]

[1] I want to specially thank Professor Corwin and his TA Xuan Wu for the assistance of this notes. Their knowledge provide crucial guidance for the formulation of this document.

This note is dedicated to Professor Ivan Corwin.

Contents

1 MEASURE THEORY

Go back to Table of Contents. Please click TOC

The first chapter starts with definitions and results from measure theory. The purpose is to provide an introduction for readers who are new to this field.

1.1 Probability Spaces

A **probability space** is a triple $(\Omega, \mathcal{F}, \mathcal{P})$ where Ω is a set of "outcomes," $\mathcal{F}$ is a set of "events," and $\mathcal{P} : \mathcal{F} \to [0, 1]$ is a function that assigns probabilities to events. We assume that $\mathcal{F}$ is a σ-**field** (or σ-**algebra**), i.e., a (nonempty) collection of subsets of Ω that satisfy

(i) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, and

(ii) if $A_i \in \mathcal{F}$ is a countable sequence of sets, then $\cup_i A_i \in \mathcal{F}$.

Here and in what follows, countable means finite or countably infinite. Since $\cap_i A_i = (\cup_i A_i^c)^c$, it follows that a σ-**field** is closed under countable intersections. **Algebra** is a collection of $\mathcal{A}$ of sets closed under finite unions and complementation. **Semi-algebra** is a collection $\mathcal{S}$ of sets closed under intersection and such that $S \in \mathcal{S}$ implies that S^c is a finite disjoint union of sets in $\mathcal{S}$. (Example: empty set plus sets of form $(a_1, b_1] \times \cdots \times (a_d, b_d] \in \mathbb{R}^d$.) The **Borel** σ-**algebra** $\mathcal{B}$ on a topological space is the smallest σ-algebra containing all open sets.

Let us notate the following. Say collection of non-empty sets $\mathcal{P}$ is a π-system if closed under intersection. Say a collection of sets $\mathcal{L}$ is a λ-system if (i) $\Omega \in \mathcal{L}$; (ii) if $A, B \in \mathcal{L}$ and $A \subset B$, then $B - A \in \mathcal{L}$; (iii) if $A_n \in \mathcal{L}$ and $A_n \uparrow A$ then $A \in \mathcal{L}$. Such terminologies are introduced again before Theorem 2.1.3. Notice that λ-system is closed under complements and disjoint countable unions described above in (i) - (iii), while a σ-**algebra** is closed under complements and arbitrary countable unions. That is, every σ-**algebra** is a λ-system, but not vice versa.

Without $\mathcal{P}$, $(\Omega, \mathcal{F})$ is called a measurable space, i.e., it is a space on which we can put a measure. A measure is a nonnegative countably additive set function; that is, a function $\mu : \mathcal{F} \to \mathbb{R}$ with

(i) $\mu(A) \geq \mu(\emptyset) = 0$ for all $A \in \mathcal{F}$, and

(ii) if $A_i \in \mathcal{F}$ is a countable sequence of disjoint sets, then

$$\mu(\cup_i A_i = \sum_i \mu(A_i)$$

If $\mu(\Omega) = 1$, we call μ a **probability measure**, which is always denoted by $\mathcal{P}$ in this document.

The next result gives some consequences of the definition of a measure that we will need later.

Theorem 1.1.1. *Let μ be a measure on $(\Omega, \mathcal{F})$*

(i) ***monotonicity****. If $A \subset B$, then $\mu(A) \leq \mu(B)$.*

(ii) ***subadditivity****. If $A \subset \cup_{m=1}^{\infty} A_m$, then $\mu(A) \leq \sum_{m=1}^{\infty} \mu(A_m)$.*

(iii) ***continuity from below****. If $A_i \uparrow A$ (i.e., $A_1 \subset A_2 \subset \ldots$ and $\cup_i A_i = A$), then $\mu(A_i) \uparrow \mu(A)$.*

(iv) ***continuity from above****. If $A_i \downarrow A$ (i.e., $A_1 \subset A_2 \subset \ldots$ and $\cap_i A_i = A$), with $\mu(A_1) < \infty$, then $\mu(A_i) \downarrow \mu(A)$.*

Proof. (i) Let $B - A = B \cap A^c$ be the difference of the two sets. Using $+$ to denote the disjoint union, $B = A + (B - A)$ so

$$\mu(B) = \mu(A) + \mu(B - A) \geq \mu(A).$$

(ii) Let $A'_n = A_n \cap A$, $B_1 = A'_1$ and for $n > 1$, $B_n = A'_n - \cup_{m=1}^{n-1}(A'_m)^c$. Since the B_n are disjoint and have union A have using (i) of the definition of measure, $B_m \subset A_m$, and (i) of this theorem

$$\mu(A) = \sum_{m=1}^{\infty} \mu(B_m) \leq \sum_{m=1}^{\infty} \mu(A_m)$$

(iii) Let $B_n = A_n - A_{n-1}$. Then the B_n are disjoint and have $\cup_{m=1}^{\infty} = A$, $\cup_{m=1}^{n} B_m = A_n$ so

$$\mu(A) = \sum_{m=1}^{\infty} \mu(B_m) = \lim_{n\to\infty} \sum_{m=1}^{n} \mu(B_m) = \lim_{n\to\infty} \mu(A_n)$$

(iv) $A_1 - A_n \uparrow A_1 - A$ so (iii) implies $\mu(A_1 - A_n) \uparrow \mu(A_1 - A)$. Since $A_1 \supset B$, we have $\mu(A_1 - B) = \mu(A_1) - \mu(B)$ and it follows that $\mu(A_n) \downarrow \mu(A)$.

□

Example 1.1.2. Discrete probability spaces. Let $\Omega =$ a countable set, i.e., finite or countably infinite. Let $\mathcal{F} =$ the set of all subsets of Ω. Let

$$\mathcal{P}(A) = \sum_{\omega \in A} p(\omega) \text{ where } p(\omega) \geq 0 \text{ and } \sum_{\omega \in \Omega} p(\omega) = 1$$

This is the most general probability measure on this space. Many cases when Ω is a finite set, we have $p(\omega) = 1/|\Omega|$ where $|\Omega| =$ the number of points in Ω.

Example 1.1.3. Measures on the real line. Measures on $(\mathbb{R}, \mathcal{R})$ are defined by giving probability a **Stieltjes measure function** with the following properties:

(i) F is nondecreasing.

(ii) F is right continuous, i.e., $\lim_{y \downarrow x} F(y) = F(x)$.

Theorem 1.1.4. *Associated with each Stieltjes measure function F there is a unique measure μ on $(\mathbb{R}, \mathcal{R})$ with $\mu((a, b]) = F(b) - F(a)$*

$$\mu((a, b]) = F(b) - F(a) \tag{1.1}$$

*When $F(x) = x$ the resulting measure is called **Lebesgue measure**. Moreover, if you can measure, then you can integrate. That is, if $(\Omega, \mathcal{F})$ is a measure space with a measure μ with $\mu(\Omega) < \infty$) and $f : \Omega \to \mathbb{R}$ is $\mathcal{F}$-measurable, then we can define $\int f d\mu$ (for non-negative f, also if both $f \vee 0$ and $-f \wedge 0$ and have finite integrals.)*

The details of the proof (using Caratheéodory Extention Theorem) is referenced from Appendix A1. in the text [4].

A collection $\mathcal{S}$ of sets is said to be a semialgebra if (i) it is closed under intersection, i.e., $S, T \in \mathcal{S}$ implies $S \cap T \in \mathcal{S}$, and (ii) if $S \in \mathcal{S}$ then S^c is a finite disjoint union of sets in $\mathcal{S}$.

Example 1.1.5. $\mathcal{S}^d =$ the empty set plus all sets of the form

$$(a_1, b_1] \times \cdots \times (a_d, b_d] \subset \mathbb{R}^d \text{ where } -\infty \leq a_i < b_i \leq \infty$$

The definition in 1.1 gives the value of μ on the semialgebra $\mathcal{S}_1$. To go from semialgebra to σ-algebra we use an intermediate step. A collection $\mathcal{A}$ of subsets of Ω is called an algebra (or field) if $A, B \in \mathcal{A}$ implies A^c and $A \cup B$ are in $\mathcal{A}$. Since $A \cap B = (A^c \cup B^c)^c$, it follows that $A \cap B \in \mathcal{A}$. Obviously a σ-algebra is an algebra. An example in which the converse is false is:

Example 1.1.6. Let $\Omega = \mathbb{Z} =$ the integers . $\mathcal{A} =$ the collection $A \subset \mathbb{Z}$ so that A or A^c is finite is an algebra.

Lemma 1.1.7. *If $\mathcal{S}$ is a semialgebra then $\bar{\mathcal{S}} = \{$finite disjoint unions of sets in $\mathcal{S}\}$ is an algebra, called algebra generated by $\mathcal{S}$.*

Proof. Suppose $A = +_i S_i$ and $B = +_j T_j$, where $+$ denotes disjoint union and we assume the index sets are finite. Then $A \cap B = +_{i,j} S_i \cap T_j \in \bar{\mathcal{S}}$. As for complements, if $A = +_i S_i$ then $A^c = \cap_i S_i^c$. The definition of $\mathcal{S}$ implies $S_i^c \in \bar{\mathcal{S}}$. We have shown that $\bar{\mathcal{S}}$ is closed under intersection, so it follows by induction that $A^c \in \bar{\mathcal{S}}$.

□

Example 1.1.8. Let $\Omega = \mathbb{R}$ and $\mathcal{S} = \mathcal{S}_1$ then $\bar{S}_1 =$ the empty set plus all sets of the form

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

Given a set function μ on $\mathcal{S}$ we can extend it to $\bar{\mathcal{S}}$ by

$$\mu(+_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i)$$

By a measure on an algebra $\mathcal{A}$, we mean a set function μ with

(i) $\mu(A) \ge \mu(\emptyset) = 0$ for all $A \in \mathcal{A}$, and

(ii) if $A_i \in \mathcal{A}$ are disjoint and their union is in $\mathcal{A}$, then

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

μ is said to be σ-finite if there is a sequence of sets $A_n \in \mathcal{A}$ so that $\mu(A_n) < \infty$ and $\cup_n A_n = \Omega$. Letting $A_1' = A_1$ and for $n \ge 2$,

$$A_n' = \cup_{m=1}^n A_m \text{ or } A_n' = A_n \cap (\cap_{m=1}^{n-1} A_m^c) \in \mathcal{A}$$

we can without loss of generality assume that $A_n \uparrow \Omega$ or the A_n are disjoint.

The next result helps us to extend a measure defined on a semi-algebra $\mathcal{S}$ to the σ-algebra it generates, $\sigma(\mathcal{S})$.

Theorem 1.1.9. *Let $\mathcal{S}$ be a semialgebra and let μ defined on $\mathcal{S}$ have $\mu(\emptyset) = 0$. Suppose (i) if $S \in \mathcal{S}$ is a finite disjoint union of sets $S_i \in \mathcal{S}$ then $\mu(S) = \sum_i \mu(S_i)$, and (ii) if S_i, $S \in \mathcal{S}$ with $S = +_{i \ge 1} S_i$ then $\mu(S) \ge \sum_{i \ge 1} \mu(S_i)$. Then μ has a unique extension $\bar{\mu}$ that is a measure on $\bar{\mathcal{S}}$ the algebra generated by $\mathcal{S}$. If $\bar{\mu}$ is sigma-finite then there is a unique extension ν that is a measure on $\sigma(\mathcal{S})$.*

In (ii) above, and in what follows, $i \ge 1$ indicates a countable union, while a plain subscript i or j indicates a finite union. Proofs see A.1.1.4 in text [4].

Lemma 1.1.10. *Suppose only that (i) holds.*

(a) If $A, B_i \in \bar{\mathcal{S}}$ with $A = +_{n=1}^n B_i$ then $\bar{\mu}(A) = \sum_i \bar{\mu}(B_i)$.

(b) If $A, B_i \in \bar{\mathcal{S}}$ with $A \subset \cup_{i=1}^n B_i$ then $\bar{\mu}(A) \le \sum_i \bar{\mu}(B_i)$.

Proof. Observe that it follows from the definition that if $A = +_i B_i$ is a finite disjoint union of sets in $\bar{\mathcal{S}}$ and $B_i = +_j S_{i,j}$, then

$$\bar{\mu}(A) = \sum_{i,j} \mu(S_{i,j} = \sum_i \bar{\mu}(B_i)$$

To prove (b), we begin with the case $n = 1$, $B_1 = B$. $B = A + (B \cap A^c)$ and $B \cap A^c \in \bar{\mathcal{S}}$, so

$$\bar{\mu}(A) \leq \bar{\mu}(A) + \bar{\mu}(B \cap A^c) = \bar{\mu}(B)$$

To handle $n > 1$ now, let $F_k = B_1^c \cap \cdots \cap B_{k-1}^c \cap B_k$ and note

$$\begin{aligned} \cup_i B_i &= F_1 + \cdots + F_n \\ A = A \cap (\cup_i B_i) &= (A \cap F_1) + \cdots + (A \cap F_n) \end{aligned}$$

so using (a), (b) with $n = 1$, and (a) again

$$\bar{\mu}(A) = \sum_{k=1}^n \bar{\mu}(A \cap F_k) \leq \sum_{k=1}^n \bar{\mu}(F_k) = \bar{\mu}(\cup_i B_i)$$

□

Proof of Theorem 1.1.4. Let $\mathcal{S}$ be the semi-algebra of half-open intervals $(a, b]$ with $-\infty \leq a < b \leq \infty$. To define μ and $\mathcal{S}$, we begin by observing that

$$F(\infty) = \lim_{x \uparrow \infty} F(x) \text{ and } F(-\infty) = \lim_{x \downarrow -\infty} F(x) \text{ exist}$$

and $\mu((a, b]) = F(b) - F(a)$ makes sense for all $-\infty \leq a < b \leq \infty$ since $F(\infty) > -\infty$ and $F(-\infty) < \infty$.

If $(a, b] = +_{i=1}^n (a_i, b_i]$ then after relabeling the intervals we must have $a_1 = a$, $b_n = b$, and $a_i = b_{i-1}$ for $2 \leq i \leq n$, so condition (i) in Theorem 1.1.9 holds. To check (ii), suppose first that $-\infty < a < b < \infty$, and $(a, b] \subset \cup_{i \geq 1} (a_i, b_i]$ where (without loss of generality) $-\infty < a_i < b_i < \infty$. Pick $\delta > 0$ so that $F(a + \delta) < F(a) + \epsilon$ and pick η_i so that

$$F(b_i + \eta_i) < F(b_i) + \epsilon 2^{-i}$$

The open intervals $(a_i, b_i + \eta_i)$ cover $[a + \delta, b]$, so there is a finite subcover (α_j, β_j), $1 \leq j \leq J$. Since $(a + \delta, b] \subset \cup_{j=1}^J (\alpha_j, \beta_j]$, (b) in Lemma 1.1.10 implies

$$F(b) - F(a + \delta) \leq \sum_{j=1}^J F(\beta_j) - F(\alpha_j) \leq \sum_{i=1}^\infty (F(b_i + \eta_i) - F(\alpha_i))$$

and since ϵ is arbitrary, we have proved the result in the case $-\infty < a < b < \infty$. To remove the last restriction, observe that if $(a, b] \subset \cup_i (a_i, b_i]$ and $(A, B] \subset (a, b]$ has $-\infty < A < B < \infty$, then we have

$$F(B) - F(A) \leq \sum_{i=1}^\infty (F(b_i) - F(a_i))$$

Since the last result holds for any finite $(A, B] \subset (a, b]$, the desired result follows.

□

The next goal is to prove a version of Theorem 1.1.4 for $\mathbb{R}^d$. The first step is to introduce the assumptions on the defining function F. By analogy with the case $d = 1$ it is natural to assume:

(i) It is nondecreasing, i.e., if $x \leq y$ (meaning $x_i \leq y_i$ for all i) then $F(x) \leq F(y)$.

(ii) F is right continuous, i.e., $\lim_{y \downarrow x} F(y) = F(x)$ (here $y \downarrow x$ means each $y_i \downarrow x_i$).

However this time it is not enough. Consider the following F

$$F(x_i, x_y) = \begin{cases} 1 & \text{if } x_1, x_2 \geq 1 \\ 2/3 & \text{if } x_1 \geq 1 \text{ and } 0 \leq x_2 < 1 \\ 2/3 & \text{if } x_2 \geq 1 \text{ and } 0 \leq x_2 < 1 \\ 0 & \text{otherwise} \end{cases}$$

A little thought shows that

$$\begin{aligned} \mu((a_1, b_1] \times (a_2, b_2]) &= \mu((-\infty, b_1] \times (-\infty, b_2]) - \mu((-\infty, a_1] \times (-\infty, b_2]) \\ &\quad -\mu((-\infty, b_1] \times (-\infty, a_2]) + \mu((-\infty, a_1] \times (-\infty, a_2]) \\ &= F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2) \end{aligned}$$

Using this with $a_1 = a_2 = 1 - \epsilon$ and $b_1 = b_2 = 1$ and letting $\epsilon \to 0$ we see that

$$\mu(\{1, 1\}) = 1 - 2/3 - 2/3 + 0 = -1/3$$

Similar reasoning shows that $\mu(\{1, 0\}) = \mu(\{0, 1\}) = 2/3$.
To formulate the third and final condition for F to define a measure, let

$$A = (a_1, b_1] \times \cdots \times (a_d, b_d]$$

$$V = \{a_1, b_1\} \times \cdots \times \{a_d, b_d\}$$

where $-\infty < a_i < b_i < \infty$. To emphasize that ∞'s are allowed, we will call A a finite rectangle. Then $V =$ the vertices of the rectangle A. If $v \in V$, let

$$\text{sgn}(v) = (-1)^{\#\text{ of } a's \text{ in } v}$$

$$\triangle_A F = \sum_{v \in V} \text{sgn}(v) F(v)$$

We will let $\mu(A) = \triangle_A F$, so we must assume

(iii) $\triangle_A F \geq 0$ for all rectangles A.

Theorem 1.1.11. *Suppose $F : \mathbb{R}^d \to [0, 1]$, satisfies (i) - (iii) given above. Then there is a unique probability measure μ on $(\mathbb{R}^d, \mathcal{R}^d)$ so that $\mu(A) = \triangle_A F$ for all finite rectangles.*

Example 1.1.12. Suppose $F(x) = \prod_{i=1}^d F_i(x)$, where the F_i satisfy (i) and (ii) of Theorem 1.1.4. In this case,

$$\triangle_A F = \prod_{i=1}^d (F_i(b_i) - F_i(a_i))$$

When $F_i(x) = x$ for all i, the resulting measure is Lebesgue measure on $\mathbb{R}^d$.

Proof. We let $\mu(A) = \triangle_A F$ for all finite rectangles and then use monotonicity to extend the definition to $\mathcal{S}_d$. To check (i) of Theorem 1.1.9, call $A = +_k B_k$ a regular subdivision of A if there are sequences $a_i = \alpha_{i,0} < \alpha_{i,1} \cdots < \alpha_{i,n_i} = b_i$ so that each rectangle B_k has the form

$$(\alpha_{1,j_1 - 1}, \alpha_{1,j_1}] \times \cdots \times (\alpha_{d,j_d - 1}, \alpha_{d,j_d}] \text{ where } 1 \leq j_i \leq n_i$$

When $F_i(x) = x$ for all i, the resulting measure is Lebesgue measure on $\mathbb{R}^d$.

Proof. We let $\mu(A) = \triangle_A F$ for all finite rectangles and then use monotonicity to extend the definition to $\mathcal{S}_d$. Check (i) of Theorem 1.1.9, write $A = +_k B_k$ a regular subdivision of A if there are sequences $a_i = a_{i,0} < \alpha_{i,1} \cdots < \alpha_{i,n_i} = b_i$.

It is easy to see that for regular subdivisions $\lambda(A) = \sum_k \lambda(B_k)$. To extend this result to a general finite subdivision $A = +_j A_j$, subdivide further to get a regular one.

The proof of (ii) is almost identical that in Theorem 1.1.4. Let

$$(x, y) = (x_1, y_1) \times \cdots \times (x_d, y_d)$$
$$(x, y] = (x_1, y_1] \times \cdots \times (x_d, y_d]$$
$$[x, y] = [x_1, y_1] \times \cdots \times [x_d, y_d]$$

for $x, y \in \mathbb{R}^d$. Suppose that $-\infty < a < b < \infty$, where the inequalities mean that each component is finite, and suppose $(a, b] \subset \cup_{i \geq 1}(a^i, b^i]$, where (without loss of generality) $-\infty < a^i < b^i < \infty$. Let $\mathbf{1} = (1, ..., 1)$, pick $\delta > 0$ so that

$$\mu((a, b^i + \eta_i \mathbf{1}]) < \mu((a^i, b^i]) + \epsilon$$

and pick η_i so that

$$\mu((a, b^i + \eta_i \mathbf{1}]) < \mu((a^i, b^i]) + \epsilon 2^{-i}$$

The open rectangles $(a^i, b^i + \eta_i \mathbf{1})$ cover $[a + \delta \mathbf{1}, b]$, so there is a finite subcover (α^j, β^j), $1 \leq j \leq J$. Since $(a + \delta \mathbf{1}, b] \subset \cup_{j=1}^{j}(\alpha^j, \beta^j]$, (b) in Lemma 1.1.10 implies

$$\mu([a + \delta \mathbf{1}, b]) \leq \sum_{j=1}^{J} \mu((\alpha^j, \beta^j]) \leq \sum_{i=1}^{\infty} \mu((a^i, b^i + \eta_i \mathbf{1}])$$

So, by the choice of δ and η_i,

$$\mu((a, b]) \leq 2\epsilon + \sum_{i=1}^{\infty} \mu((a^i, b^i])$$

and since ϵ is arbitrary, we have proved the result in case $-\infty < a < b < \infty$. The proof can now be completed exactly as before.

□

1.2 Distributions

Probability spaces become more interesting when we define random variables on them. A real valued function X on Ω is said to be a random variable if for every Borel set $B \subset \mathbb{R}$ we have $X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathcal{F}$. When we need to emphasize the σ-field, we will say that X is $\mathcal{F}$-measurable or write $X \in \mathcal{F}$. If Ω is a discrete probability space, then any function $X : \Omega \to \mathbb{R}$ is a random variable. A second trivial, but useful, type of example or a random variable is the indicator function of a set $A \in \mathcal{F}$:

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

If X is a random variable, then X inducse a probability measure on $\mathbb{R}$ called its distribution by setting $\mu(A) = P(X \in A)$ for Borel sets A. Using the notation introduced above, the right-hand side can be written as $P(X^{-1}(A))$. In words, we pull $A \in \mathcal{R}$ back to $X^{-1}(A) \in \mathcal{F}$ and then take P of that set.

To check that μ is a probability measure we observe that if the A_i are disjoint then using the definition of μ; the fact that X lands in the union if and only if it lands in

one of the A_i; the fact if the sets $A_i \in \mathcal{R}$ are disjoint then the events $\{X \in A_i\}$ are disjoint; and the definition of μ again; we have:

$$\mu(\cup_i A_i) = P(X \in \cup_i A_i) = P(\cup_i \{X \in A_i\}) = \sum_i P(X \in A_i) = \sum_i \mu(A_i)$$

The distribution of a random variable X is usually described by giving its distribution function, $F(x) = P(X \leq x)$.

Theorem 1.2.1. *Any distribution function F has the following properties:*

(i) *F is nondecreasing.*

(ii) $\lim_{x\to\infty} F(x) = 1$, $\lim_{x\to\infty} F(x) = 0$.

(iii) *F is right continuous, i.e.* $\lim_{y\downarrow x} F(y) = F(x)$.

(iv) *If* $F(x-) = \lim_{y\uparrow x} F(y)$, *then* $F(x-) = P(X < x)$.

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

Proof. (i) Note that if $x \leq y$, then $\{X \leq x\} \subset \{X \leq y\}$, and then use (i) in Theorem 1.1.1 monotonicity to imply that $P(X \leq x) \leq P(X \leq y)$.

(ii) We observe that if $x \uparrow \infty$, then $\{X \leq x\} \uparrow \Omega$, while if $x \downarrow -\infty$, then $\{X \leq x\} \downarrow \emptyset$ and then use (iii) and (iv) in Theorem 1.1.1 continuity from below and above for conclusion.

(iii) We observe that if $y \downarrow x$, then $\{X \leq y\} \downarrow \{X \leq x\}$.

(iv) We observe that if $y \uparrow x$, then $\{X \leq y\} \uparrow \{X < x\}$.

(v) Note that $P(X = x) = P(X \leq x) - P(X < x)$ and use (iii) and (iv).

□

Theorem 1.2.2. *If F satisfies (i), (ii), and (iii) in Theorem 1.2.1, then it is the distribution function of some random variable.*

Proof. Let $\Omega = (0, 1)$, $\mathcal{F} =$ Borel sets, and $P =$ Lebesgue measure. If $\omega \in (0, 1)$, let

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

Once we show that

$$(\star)\{\omega : X(\omega) \leq x\} = \{\omega : \omega \leq F(x)\}$$

the desired result follows immediately since $P(\omega : \omega \leq F(x)) = F(x)$. Recall P is Lebesgue measure. To check $(\star)$, we observe that if $\omega \leq F(x)$ then $X(\omega \leq x)$, since $x \notin \{y : F(y) < \omega\}$. On the other hand, if $\omega > F(x)$, then since F is right continuous, there is an $\epsilon > 0$ so that $F(x + \epsilon) < \omega$ and $X(\omega) \geq x + \epsilon > x$.

□

If X and Y induce the same distribution μ on $(\mathbb{R}, \mathcal{R})$, we say X and Y are equal in distribution. in view of Theorem 1.2.2, this holds if and only if X and Y have the same distribution function, i.e., $P(X \leq x) = P(Y \leq x)$ for all x. When X and Y have the same distribution, we like to write

$$X \stackrel{d}{=} Y$$

and we also write $X =_d Y$ for same meaning.

When the distribution function $F(x) = P(X \leq x)$ has the form

$$F(x) = \int_{-\infty}^{x} f(y)dy \tag{1.2}$$

and we say that X has density function f. It is often to think of $f(x)$ as being $P(X = x)$ although

$$P(X = x) = \lim_{\epsilon \to 0} \int_{x-\epsilon}^{x+\epsilon} f(y)dy = 0$$

Example 1.2.3. Uniform distribution on (0,1). $f(x) = 1$ for $x \in (0, 1)$ and 0 otherwise. Distribution function:

$$F(x) = \begin{cases} 0 & x \leq 0 \\ 0 & \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

Example 1.2.4. Exponential distribution with rate λ. $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and 0 otherwise. Distribution function:

$$F(x) = \begin{cases} 0 & x \leq 0 \\ 1 - e^{-x} & x \geq 0 \end{cases}$$

Example 1.2.5. Standard normal distribution

$$f(x) = (2\pi)^{-1/2} \exp(-x^2/2)$$

In this case, there is no closed form expression for $F(x)$, but we have the following bounds that are useful for large x:

Theorem 1.2.6. *For $x > 0$,*

$$(x^{-1} - x^{-3}) \exp(-x^2/2) \leq \int_x^{\infty} \exp(-y^2/2)dy \leq x^{-1} \exp(-x^2/2)$$

Proof. Changing variables $y = x + z$ and using $\exp(-z^2/2) \leq 1$ gives

$$\int_x^{\infty} \exp(-y^2/2)dy \leq \exp(-x^2/2) \int_0^{\infty} \exp(-xz)dz = x^{-1} \exp(-x^2/2)$$

For the other direction, we observe

$$\int_0^{\infty} (1 - 3y^{-4}) \exp(-y^2/2)dy = (x^{-1} - x^{-3}) \exp(-x^2/2)$$

□

1.3 Random Variables

In this section, we develop some results that will help us later to prove that quantities we define are random variables, i.e., they are measurable. Since most of what we have to say is true for random elements of an arbitrary measurable space $(S, \mathcal{S})$ and the proofs are the same (sometimes easier), we will develop our results in that generality. First, we need a definition. A function $X : \Omega \to S$ is said to be a measurable map from $(\Omega, \mathcal{F})$ to $(S, \mathcal{S})$ if

$$X^{-1}(B) \equiv \{\omega : W(\omega) \in B\} \in \mathcal{F} \text{ for all } B \in \mathcal{S}$$

Is $(S, \mathcal{S}) = (\mathbb{R}^d, \mathcal{R}^d)$ and $d > 1$ then X is called a random vector. If $d = 1$, X is called a random variable, or r.v. for short. For example, let $(M, \mathcal{A})$ be a measurable

space. Let $S \subseteq M$ be a subset. Consider the function $\mathbb{1}_S : M \to \mathbb{R}$ taking elements in S to 1 and elements outside S to 0. Equip $\mathbb{R}$ with, say, the Borel σ-algebra. Then $\mathbb{1}_S$ is measurable if and only if $S \in \mathcal{S}$.

The next result is useful for proving that maps are measurable.

Theorem 1.3.1. *If $\{\omega : X(\omega) \in A\} \in \mathcal{F}$ for all $A \in \mathcal{A}$ and $\mathcal{A}$ generates $\mathcal{S}$ (i.e., $\mathcal{S}$ is the smallest σ-field that contains $\mathcal{A}$), then X is measurable.*

Proof. Write $\{X \in B\}$ as shorthand for $\{\omega : X(\omega) \in B\}$, we have

$$\{X \in \cup_i B_i\} = \cup_i \{X \in B_i\}$$

$$\{X \in B^c\} = \{X \in B\}^c$$

So the class of sets $\mathcal{B} = \{B : \{X \in B\} \in \mathcal{F}\}$ is a σ-field. Since $\mathcal{B} \subset \mathcal{A}$ and $\mathcal{A}$ generates $\mathcal{S}$, then $\mathcal{B} \supset \mathcal{S}$.

□

It follows from the two equations displayed in the previous proof that if $\mathcal{S}$ is a σ-field, then $\{\{X \in B\} : B \in \mathcal{S}\}$ is a σ-field. It is the smallest σ-field on Ω that makes X a measurable map. It is called the σ-field generated by X and denoted $\sigma(X)$. For future reference we note that

$$\sigma(X) = \{\{X \in B\} : B \in \mathcal{S}\} \tag{1.3}$$

Example 1.3.2. If $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{R})$ then possible choices of $\mathcal{A}$ in Theorem 1.3.1 are $\{(-\infty, x] : x \in \mathbb{R}\}$ or $\{(-\infty, x) : x \in \mathbb{Q}\}$ where $\mathbb{Q} =$ the rationals.

Example 1.3.3. If $(S, \mathcal{S}) = \mathbb{R}^d, \mathcal{R}^d)$, a useful choice of $\mathcal{A}$ is

$$\{(a_1, b_1) \times \cdots \times (a_d, b_d) : -\infty < a_i < b_i < \infty\}$$

or occasionally the larger collection of open sets.

Theorem 1.3.4. *If $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ and $f : (S, \mathcal{S}) \to (T, \mathcal{T})$ are measurable maps, then $f(X)$ is a measurable map from $(\Omega, \mathcal{F})$ to $(T, \mathcal{T})$*

Proof. Let $B \in \mathcal{T}$. $\{\omega : f(X(\omega)) \in B\} = \{\omega : X(\omega) \in f^{-1}(B)\} \in \mathcal{F}$, since by assumption $f^{-1}(B) \in \mathcal{S}$.

□

From Theorem 1.3.4, it follows immediately that if X is a random variable then so is cX for all $c \in \mathbb{R}$, X^2, $\sin(X)$, etc. The next result shows why we wanted to prove Theorem 1.3.4 for measurable maps.

Theorem 1.3.5. *If $X_1, ..., X_n$ are random variables and $f : (\mathbb{R}^n, \mathcal{R}^n) \to \mathbb{R}, \mathcal{R})$ is measurable, then $f(X_1, ..., X_n)$ is a random variable.*

Proof. In view of Theorem 1.3.4, it suffices to show that $(X_1, ..., X_n)$ is a random vector. To do this, we observe that if $A_1, ..., A_n$ are Borel sets then

$$\{(X_1, ..., X_n) \in A_1 \times \cdots \times A_n\} = \cap_i \{X_i \in A_i\} \in \mathcal{R}$$

Since sets of the form $A_1 \times \cdots \times A_n$ generate $\mathcal{R}^n$, the desired result follows from Theorem 1.3.1.

□

Theorem 1.3.6. *If $X_1, ..., X_n$ are random variables then $X_1 + \cdots + X_n$ is a random variable.*

Proof. In view of Theorem 1.3.5 it suffices to show that $f(x_1, ..., x_n) = x_1 + \cdots + x_n$ is measurable. To do this, we use Example 1.3.2 and note that $\{x : x_1 + \cdots + x_n < a\}$ is an open set and hence is in $\mathcal{R}^n$.

□

Theorem 1.3.7. *If $X_1, X_2, ...$ are random variables then so are*

$$\inf_n X_n \quad \sup_n X_n \quad \limsup_n X_n \quad \liminf_n X_n$$

Proof. Since the infimum of a sequence is $< a$ if and only if some term is $< a$ (if all terms are $\geq a$ then the infimum is), we have

$$\{\inf_n X_n < a\} = \cup_n \{X_n < a\} \in \mathcal{F}$$

A similar argument shows $\{\sup_n X_n > a\} = \cup_n \{X_n < a\} \in \mathcal{F}$. For the last two, we observe

$$\liminf_{n\to\infty} X_n = \sup_n \big(\inf_{m \geq n} X_m \big)$$

$$\limsup_{n\to\infty} X_n = \inf_n \big(\sup_{m \geq n} X_n \big)$$

To complete the proof in the first case, note that $Y_n = \inf_{m \geq n} X_m$ is a random variable for each n so $\sup_n Y_n$ is as well.

□

From Theorem 1.3.7, we see that

$$\Omega_0 \equiv \{\omega : \lim_{n\to\infty} X_n \text{ exists } \} = \{\omega : \limsup_{n\to\infty} X_n - \liminf_{n\to\infty} X_n = 0\}$$

is a measurable set. (Here $\equiv$ indicates that the first equality is a definition.) If $P(\Omega_0) = 1$, we say that X_n converges almost surely, or a.s. for short. This type of convergence is called almost everywhere in measure theory. To have a limit defined on the whole space, it is convenient to let

$$X_\infty = \limsup_{n\to\infty} X_n$$

but this random variable may take the value $+\infty$ or $-\infty$. To accommodate this and some other headaches, we will generalize the definition of random variable.

A function whose domain is a set $D \in \mathcal{F}$ and whose range is $\mathbb{R}^* \equiv [-\infty, \infty]$ is said to be a random variable if for all $B \in \mathcal{R}^*$ we have $X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathcal{F}$. Here $\mathcal{R}^* =$ the Borel subsets of $\mathbb{R}^*$ with $\mathbb{R}^*$ given the usual topology, i.e., the one generated by intervals of the form $[-\infty, a), (a, b)$ and $(b, \infty]$ where $a, b \in \mathbb{R}$. Please note that the extended real line $(\mathbb{R}^*, \mathcal{R}^*)$ is a measurable space, so all the results above generalize immediately.

1.4 Integration

Let μ be a σ-finite measure on $(\Omega, \mathcal{F})$. we will be primarily interested in the special case μ is a probability measure, but we will sometimes need to integrate with respect to infinite measure and it is no harder to develop the results in general.

This section we will define $\int f d\mu$ for a class of measurable functions. This is a four-step procedure:

1. Simple functions
2. Bounded functions
3. Nonnegative functions
4. General functions

This sequence of four steps is also useful in proving integration formulas.

Step 1. φ is said to be a simple function if $\varphi(\omega) = \sum_{i=1}^{n} a_i 1_{A_i}$ and A_i are disjoint sets with $\mu(A_i) < \infty$. If φ is a simple function, we let

$$\int \varphi d\mu = \sum_{i=1}^{n} a_i \mu(A_i)$$

The representation of φ is not unique since we have not supposed that the a_i are distinct. However, it is easy to see that the last definition does not contradict itself.

We will prove the next three conclusions four times, but before we can state them for the first time, we need a definition. $\varphi \geq \psi$ μ-almost everywhere (or $\varphi \geq \psi$ μ-a.e.) means $\mu(\{\omega : \varphi(\omega) < \psi(\omega)\}) = 0$. When there is no doubt about what measure we are referring to, we drop the μ.

Lemma 1.4.1. *Let φ and ψ be simple functions.*
(i) If $\varphi \geq 0$ a.e., then $\int \varphi d\mu \geq 0$.
(ii) For any $a \in \mathbb{R}$, $\int a\varphi d\mu = a \int \varphi d\mu$.
(iii) $\int \varphi + \psi d\mu = \int \varphi d\mu + \int \psi d\mu$.

Proof. (i) and (ii) are immediate consequences of the definition. To prove (iii), suppose

$$\varphi = \sum_{i=1}^{m} a_i 1_{A_i} \text{ and } \psi = \sum_{j=1}^{n} b_j 1_{B_j}$$

To make the supports of the two functions the same we let $A_0 = \cup_i B_i - \cup_i A_i$, let $B_0 = \cup_i A_i - \cup_i B_i$, and let $a_0 = b_0 = 0$. Now

$$\varphi + \psi = \sum_{i=0}^{m} \sum_{j=0}^{n} (a_i + b_j) 1_{(A_i \cap B_j)}$$

and the $A_i \cap B_j$ are pairwise disjoint, so

$$\begin{aligned} \int (\varphi + \psi) d\mu &= \sum_{i=0}^{m} \sum_{j=0}^{n} (a_i + b_j) \mu(A_i \cap B_j) \\ &= \sum_{i=0}^{m} \sum_{j=0}^{n} a_i \mu(A_i \cap B_j) + \sum_{i=0}^{m} \sum_{j=0}^{n} b_j \mu(A_i \cap B_j) \\ &= \sum_{i=0}^{m} a_i \mu(A_i) + \sum_{j=0}^{n} b_j \mu(B_j) \\ &= \int \varphi d\mu + \int \psi d\mu \end{aligned}$$

We prove (i)-(iii) three more times as we generalize our integral. As a consequence of (i)-(iii), we get three more useful properties. To keep from repeating their proofs, which do not change, we will prove

Lemma 1.4.2. *If (i), and (iii) hold, then we have:*
(iv) If $\varphi \leq \psi$ a.e., then $\int \varphi d\mu \leq \int \psi d\mu$.
(v) If $\varphi = \psi$ a.e., then $\int \varphi d\mu = \int \psi d\mu$.
If, in addition, (ii) holds when $a = -1$ we have
(vi) $|\int \phi d\mu| \leq \int |\phi| d\mu$.

Proof. By (iii), $\int \varphi d\mu = \int \phi d\mu + \int(\psi - \phi)d\mu$ and the second integral is ≥ 0 by (i), so (iv) holds. $\varphi = \psi$ a.e. implies $\varphi \leq \psi$ a.e. and $\psi \leq \varphi$ a.e. so (v) follows from two applications of (iv). To prove (vi) now, notice that $\phi \leq |\phi|$ so (iv) implies $\int \phi d\mu \leq \int |\phi| d\mu$. $-\phi \leq |\phi|$, so (iv) and (ii) imply - $-\int \phi d\mu \leq \int |\phi| d\mu$. Since $|y| = \max(y, -y)$, the result follows.

□

Step 2. Let E be a set with $\mu(E) < \infty$ and let f be a bounded function that vanishes on E^c. To define the integral of f, we observe that if φ, ψ are simple functions that have $\varphi \leq f \leq \psi$, then we want to have

$$\int \varphi d\mu \leq \int f d\mu \leq \int \psi d\mu$$

so we let

$$\int f d\mu = \sup_{\phi \leq f} \int \varphi d\mu = \inf_{\psi \geq f} \int \psi d\mu \tag{1.4}$$

Starting from here, we assume that φ and ψ vanish on E^c. To justify the definition, we have to prove that the sup and inf are equal. It follows from (iv) in Lemma 1.4.2 that

$$\sup_{\phi \leq f} \int \varphi d\mu \leq \inf_{\psi \geq f} \int \psi d\mu$$

To prove the other inequality, suppose $|f| \leq M$ and let

$$E_k = \left\{ x \in E : \frac{kM}{n} \geq f(x) > \frac{(k-1)M}{n} \right\} \text{ for } -n \leq k \leq n$$

$$\psi_n(x) = \sum_{k=-n}^{n} \frac{kM}{n} 1_{E_k} \ \varphi_n(x) = \sum_{k=-n}^{n} \frac{(k-1)M}{n} 1_{E_k}$$

By definition, $\psi_n(x) - \varphi_n(x) = (M/n)1_E$, so

$$\int \psi_n(x) - \varphi_n(x) d\mu = \frac{M}{n} \mu(E)$$

Since $\varphi_n(x) \leq f(x) \leq \psi_n(x)$, it follows from (iii) in Lemma 1.4.1 that

$$\begin{aligned} \sup_{\phi \leq f} \int \varphi d\mu \geq \int \varphi_n d\mu &= -\frac{M}{n} \mu(E) + \int \psi_n d\mu \\ &\geq -\frac{M}{n} \mu(E) + \inf_{\psi \geq f} \int \psi d\mu \end{aligned}$$

The last inequality holds for all n, which completes the proof.

□

Lemma 1.4.3. *Let E be a set with $\mu(E) < \infty$. If f and g are bounded functions that vanish on E^c then:*
(i) If $f \geq 0$ a.e., then $\int f d\mu > 0$.
(ii) For any $a \in \mathbb{R}$, $\int af d\mu = a \int f d\mu$.
(iii) $\int f + g d\mu = \int f d\mu + \int g d\mu$.
(iv) If $g \leq f$ a.e., then $\int g d\mu \leq \int f d\mu$.
(v) If $g = f$ a.e., then $\int g d\mu = \int f d\mu$.
(vi) $|\int f d\mu| \leq \int |f| d\mu$.

Proof. Since we can take $\phi \equiv 0$, (i) is clear from definition. To prove (ii), we observe that if $a > 0$, then $a\varphi \leq af$ if and only if $\varphi \leq f$, so

$$\int afd\mu = \sup_{\phi \leq f} \int a\varphi d\mu = \sup_{\phi \leq f} a \int \varphi d\mu = a \sup_{\phi \leq f} = \int \varphi d\mu = a \int f d\mu$$

For $a < 0$, we observe that $a\varphi \leq af$ if and only if $\varphi \geq f$, so

$$\int afd\mu = \sup_{\phi \geq f} \int a\varphi d\mu = \sup_{\phi \geq f} a \int \varphi d\mu = a \inf_{\phi \geq f} \int \varphi d\mu = a \int f d\mu$$

To prove (iii), we observe that if $\psi_1 \geq f$ and $\psi_2 \geq g$, then $\psi_1 + \psi_2 \geq f + g$ so

$$\inf_{\psi \geq f+g} \int \psi d\mu \leq \inf_{\psi_1 \geq f, \psi_2 \geq g} \int \psi_1 + \psi_2 d\mu$$

Using linearity for simple functions, it follows that

$$\begin{aligned} \int f + g d\mu &= \inf_{\psi \geq f+g} \int \psi d\mu \\ &\leq \inf_{\psi_1 \geq f, \psi_2 \geq g} \int \psi_1 d\mu + \int \psi_2 d\mu \\ &= \int f d\mu + \int g d\mu \end{aligned}$$

To prove the other inequality, observe that the last conclusion applied to $-f$ and $-g$ and (ii) imply

$$-\int f + g d\mu \leq -\int f d\mu - \int g d\mu$$

(iv)-(vi) follow from (i)-(iii) by Lemma 1.4.2.

□

Notation. We define the integral of f over the set E:

$$\int_E f d\mu \equiv \int f \cdot 1_g d\mu$$

Step 3.
If $f \geq 0$, then we let

$$\int f d\mu = \sup\{\int h d\mu : 0 \leq h \leq f, h \text{ if bounded and } \mu(\{x : h(x) > 0\}) < \infty\}$$

The last definition is nice since it is clear that this is well defined. The next result will help us compute the value of the integral.

Lemma 1.4.4. *Let $E_n \uparrow \Omega$ have $\mu(E_n) < \infty$ and let $a \wedge b = \min(a, b)$. Then*

$$\int_{E_n} f \wedge n d\mu \uparrow \int f d\mu \text{ as } n \uparrow \infty$$

Proof. It is clear that from (iv) in Lemma 1.4.3 that the left-hand side increases as n does. Since $h = (f \wedge n)1_{E_n}$ is a possibility in the sup, each term is smaller than the integral on the right. To prove that the limit is $\int f d\mu$, observe that if $0 \leq h \leq f$, $h \leq M$, and $\mu(\{x : h(x) > 0\}) < \infty$, then for $n \geq M$ using $h \leq M$, (iv) and (iii),

$$\int_{E_n} f \wedge n d\mu \geq \int_{E_n} h d\mu = \int h d\mu - \int_{E_n^c} h d\mu$$

Now $0 \leq \int_{E_n^c} h d\mu \leq M\mu(E_n^c \cap \{x : h(x) > 0\}) \to 0$ as $n \to \infty$, so

$$\liminf_{n\to\infty} \int_{E_n} f \wedge n d\mu \geq \int h d\mu$$

which proves the desired result since h is an arbitrary member of the class that defines the integral of f.

□

Lemma 1.4.5. *Suppose $f, g \geq 0$.*
(i) $\int f d\mu \geq 0$
(ii) If $a > 0$, then $\int af d\mu = a \int f d\mu$.
(iii) $\int f + g d\mu = \int f d\mu + \int g d\mu$
(iv) If $0 \leq g \leq f$ a.e., then $\int g d\mu \leq \int f d\mu$.
(v) If $0 \leq g = f$ a.e., then $\int g d\mu = \int f d\mu$.

Proof. (i) is trivial from the definition. (ii) is clear, since when $a > 0$, $ah \leq af$ if and only if $h \leq f$ and we have $\int ah d\mu = a \int h d\mu$ for h in the defining class. For (iii), we observe that if $f \geq h$ and $g \geq k$, then $f + g \geq h + k$ so taking the sup over h and k in the defining classes for f and g gives

$$\int f + g d\mu \geq \int f d\mu + \int g d\mu$$

To prove the other direction, we observe $(a + b) \wedge n \leq (a \wedge n) + (b \wedge n)$ so (iv) from Lemma 1.4.3 and (iii) from Lemma 1.4.4 imply

$$\int_{E_n} (f + g) \wedge n d\mu \leq \int_{E_n} f \wedge n d\mu + \int_{E_n} g \wedge n d\mu$$

Letting $n \to \infty$ and using Lemma 1.4.4 gives (iii). As before, (iv) and (v) follow from (i), (iii), and Lemma 1.4.2.

□

Step 4. We say f is integrable if $\int |f| d\mu < \infty$. Let

$$f^+(x) = f(x) \vee 0 \text{ and } f^-(x) = (-f(x)) \vee 0$$

where $a \vee b = \max(a, b)$. Clearly,

$$f(x) = f^+(x) - f^-(x) \text{ and } |f(x)| = f^+(x) + f^-(x)$$

We define the integral of f by

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu$$

The right-hand side is well defined since f^+, $f^- \leq |f|$ and we have (iv) in Lemma 1.4.5. Finally, we prove our six properties. It is useful to know:

Lemma 1.4.6. *If $f = f_1 - f_2$ where $f_1, f_2 \geq 0$ and $\int f_1 d\mu < \infty$, then*

$$\int f d\mu = \int f_1 d\mu - \int f_2 d\mu$$

Proof. $f_1 + f^- = f_2 + f^+$ and all four functions are ≥ 0, so by (iii) of Lemma 1.4.5,

$$\int f_1 d\mu + \int f^- d\mu = \int f_1 + f^- d\mu = \int f_2 + f^+ d\mu = \int f_2 d\mu + \int f^+ d\mu$$

Rearranging gives the desired equation.

□

Theorem 1.4.7. *Suppose f and g are integrable*
(i) If $f \geq 0$ a.e., then $\int f d\mu \geq 0$.
(ii) For all $a \in \mathbb{R}$, $\int af d\mu = a \int f d\mu$.
(iii) $\int f + g d\mu = \int f d\mu + \int g d\mu$.
(iv) If $g \leq f$ a.e., then $\int g d\mu \leq \int f d\mu$.
(v) If $g = f$ a.e., then $\int g d\mu = \int f d\mu$.
(vi) $|\int f d\mu| \leq \int |f| d\mu$.

Proof. (i) is trivial. (ii) is clear since if $a > 0$, then $(af)^+ = a(f^+)$, and so on. To prove (iii), observe that $f + g = (f^+ + g^+) - (f^- + g^-)$, so using Lemma 1.4.5 and Lemma 1.4.6

$$\begin{aligned} \int f + g d\mu &= \int f^+ + g^+ d\mu - \int f^- + g^- d\mu \\ &= \int f^+ d\mu + \int g^+ d\mu - \int f^- d\mu - \int g^- d\mu \end{aligned}$$

As usual, (iv) - (vi) follow from (i)-(iii) and Lemma 1.4.2.

□

Notation for special cases:

(a) $(\Omega, \mathcal{F}, \mu) = (\mathbb{R}^d, \mathcal{R}^d, \lambda)$, we write $\int f(x)dx$ for $\int f d\lambda$.

(b) When $(\Omega, \mathcal{F}, \mu) = (\mathbb{R}, \mathcal{R}, \lambda)$ and $E = [a, b]$, we write $\int_a^b f(x)dx$ for $\int_E f d\lambda$.

(c) When $(\Omega, \mathcal{F}, \mu) = (\mathbb{R}, \mathcal{R}, \mu)$ with $\mu((a, b]) = G(b) - G(a)$ for $a < b$, we write $\int f(x)dG(x)$ for $\int f d\mu$.

(d) When Ω is a countable set, $\mathcal{F} =$ all subsets of $Omega$, and μ is counting measure, we write $\sum_{i \in \Omega} f(i)$ for $\int f d\mu$.

1.5 Properties of the Integral

In this section, we develop properties of the integral defined in the last section.

Theorem 1.5.1. ***Jensen's inequality***. *Suppose φ is convex, that is,*

$$\lambda\varphi(x) + (1 - \lambda)\varphi(y) \geq \varphi(\lambda x + (1 - \lambda)y)$$

for all $\lambda \in (0, 1)$ and $x, y \in \mathbb{R}$. If μ is a probability measure, and f and $\varphi(f)$ are integrable then

$$\varphi\Big(\int f d\mu\Big) \leq \int \varphi(f) d\mu$$

Proof. Let $c = \int f d\mu$ and $l(x) = ax + b$ be a linear function that has $l(c) = \varphi(c)$ and $\varphi(x) \geq l(x)$. To see that such a function exists, recall that convexity implies

$$\lim_{h \downarrow 0} \frac{\varphi(c) - \varphi(c - h)}{h} \leq \lim_{h \downarrow 0} \frac{\varphi(c + h) - \varphi(c)}{h}$$

The limits exist since the sequences are monotone. If we let a be any number between the two limits and let $l(x) = a(x - c) + \varphi(c)$, then l has the desired properties. With the existence of l established, the rest is easy. (iv) in Theorem 1.4.7 implies

$$\int \varphi(f) d\mu \geq \int (af + b) d\mu = a \int f d\mu + b = l\Big(\int f d\mu\Big) = \varphi\Big(\int f d\mu\Big)$$

since $c = \int f d\mu$ and $l(c) = \phi(c)$.

□

Let $||f||_p = (\int |f|^p d\mu)^{1/p}$ for $1 \leq p < \infty$, and notice $||cf||_p = |c| \cdot ||f||_p$ for any real number c.

Theorem 1.5.2. ***Hölder's inequality***. *If $p, q \in (1, \infty)$ with $1/p + 1/q = 1$ then*

$$\int |fg| d\mu \leq ||f||_p ||g||_q$$

Proof. If $||f||_p$ or $||g||_q = 0$, then $|fg| = 0$ a.e., so it suffices to prove the result when $||f||_p$ and $||g||_q > 0$ or by dividing both sides by $||f||_p ||g||_q$, when $||f||_p = ||g||_q = 1$. Fix $y \geq 0$ and let

$$\varphi(x) = x^p/p + y^q/q - xy \text{ for } x \geq 0$$

$$\varphi'(x) = x^{p-1} - y \text{ and } \varphi''(x) = (p-1)x^{p-2}$$

so φ has a minimum at $x_0 = y^{1/(p-1)}$. $q = p/(p-1)$ and $x_0^p = y^{p/(p-1)} = y^q$ so

$$\varphi(x_0) = y^q(1/p + 1/q) - y^{(p-1)}y = 0$$

Since x_0 is the minimum, it follows that $xy \leq x^p/p + y^q/q$. Letting $x = |f|$, $y = |g|$, and integrating

$$\int |fg| d\mu \leq \frac{1}{p} + \frac{1}{q} = 1 = ||f||_p ||g||_q$$

□

Remark 1.5.3. The special case $p = q = 2$ is called Cauchy-Schwarz inequality. One can give a direct proof of the result in the case by observing that for any θ,

$$0 \leq \int (f + \theta g)^2 d\mu = \int f^2 d\mu + \theta\big(2 \int fg d\mu\big) + \theta^2\big(\int g^2 d\mu\big)$$

so the quadratic $a\theta^2 + b\theta + c$ on the right-hand side has at most one real root. Recalling the formula for the roots of a quadratic

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

we see $b^2 - 4ac \leq 0$, which is the desired result.

Our next goal is to give conditions that guarantee

$$\lim_{n\to\infty} \int f_n d\mu = \int \big(\lim_{n\to\infty} f_n\big) d\mu$$

First, we need a definition. We say that $f_n \to f$ in measure, i.e., for any $\epsilon > 0$, $\mu(\{x : |f_n(x) - f(x)| > \epsilon\}) \to 0$ as $n \to \infty$. On a space of finite measure, this is a weaker assumption than $f_n \to f$ a.e., but the next result is easier to prove in the greater generality.

Theorem 1.5.4. ***Bounded convergence theorem***. *Let E be a set with $\mu(E) < \infty$. Suppose f_n vanishes on E^c, $|f_n(x)| \leq M$, and $f_n \to f$ in measure. Then*

$$\int f d\mu = \lim_{n\to\infty} \int f_n d\mu$$

Example 1.5.5. Consider the real line $\mathbb{R}$ equipped with the Borel sets $\mathcal{R}$ and Lebesgue measure λ. The functions $f_n(x) = 1/n$ on $[0, n]$ and 0 otherwise on show that the conclusion of Theorem 1.5.4 does not hold when $\mu(E) = \infty$.

Proof. Let $\epsilon > 0$, $G_n = \{x : |f_n(x) - f(x)| < \epsilon\}$ and $B_n = E - G_n$. Using (iii) and (vi) from Theorem 1.4.7,

$$\begin{aligned} |\int f d\mu - \int f_n d\mu| &= |\int (f - f_n) d\mu| \leq \int |f - f_n d\mu \\ &= \int_{G_n} |f - f_n| d\mu + \int_{B_n} |f - f_n| d\mu \\ &= \epsilon\mu(E) + 2M\mu(B_n) \end{aligned}$$

$f_n \to f$ in measure implies $\mu(B_n) \to 0$. $\epsilon > 0$ is arbitrary and $\mu(E) < \infty$, so the proof is complete.

□

Theorem 1.5.6. ***Fatou's lemma***. *If $f_n \geq 0$, then*

$$\liminf_{n\to\infty} \int f_n d\mu \geq \int \big(\liminf_{n\to\infty} f_n\big) d\mu$$

Example 1.5.7. Example 1.5.5 shows that we may have strict inequality in Theorem 1.5.6. The functions $f_n(x) = n1_{(0,1/n]}(x)$ on $(0, 1)$ equipped with the Borel sets and Lebesgue measure show that this can happen on a space of finite measure.

Proof. Let $g_n(x) = \inf_{m\geq n} f_m(x)$. $f_n(x) \geq g_n(x)$ and as $n \uparrow \infty$,

$$g_n(x) \uparrow g(x) = \liminf_{n\to\infty} f_n(x)$$

Since $\int f_n d\mu \geq \int g_n d\mu$, it suffices then to show that

$$\liminf_{n\to\infty} \int g_n d\mu \geq \int g d\mu$$

Let $E_m \uparrow \Omega$ be sets of finite measure. Since $g_n \geq 0$ and for fixed m

$$(g_n \wedge m) \cdot 1_{E_m} \to (g \wedge m) \cdot 1_{E_m} \ a.e.$$

the bounded convergence theorem, Theorem 1.5.4, implies

$$\liminf_{n\to\infty} \int g_n d\mu \geq \int_{E_m} g_n \wedge m d\mu \to \int_{E_m} g \wedge m d\mu$$

Taking the sup over m and using Lemma 1.4.4 gives the desired result.

□

Theorem 1.5.8. ***Monotone convergence theorem***. *If $f_n \geq 0$ and $f_n \uparrow f$, then*

$$\int f_n d\mu \uparrow \int f d\mu$$

Proof. Fatou's lemma, Theorem 1.5.6, implies $\liminf \int f_n d\mu \geq \int f d\mu$. On the other hand, $f_n \leq f$ implies $\limsup \int f_n d\mu \leq \int f d\mu$.

□

Theorem 1.5.9. ***Dominated convergence theorem***. *If $f_n \to f$ a.e., $|f_n| \leq g$ for all n, and g is integrable, then $\int f_n d\mu \to \int f d\mu$.*

Proof. $f_n + g \geq 0$ so Fatou's lemma implies

$$\liminf_{n\to\infty} \int f_n + g d\mu \geq \int f + g d\mu$$

Subtracting $\int g d\mu$ from both sides gives

$$\liminf_{n\to\infty} \int f_n d\mu \geq \int f d\mu$$

Applying the last result to $-f_n$, we get

$$\limsup_{n\to\infty} \int f_n d\mu \leq \int f d\mu$$

and the proof is complete.

□

1.6 Expected Value

We now specialize to integration with respect to a probability measure P. If $X \geq 0$ is a random variable on $(\Omega, \mathcal{F}, \mathcal{P})$ then we define its expected value to be $EX = \int X dP$, which always makes sense, but may be ∞. To reduce the general case to the nonnegative case, let $x^+ = \max\{x, 0\}$ be the positive part and let $x^- = \max\{-x, 0\}$ be the negative part of x. We declare that EX exists and set $EX = EX^+ - EX^-$ whenever the subtraction makes sense, i.e., $EX^+ < \infty$ or $EX^- < \infty$.

EX is often called the mean of X and denoted by μ. EX is defined by integrating X, so it has all the properties that integrals do. From Lemma 1.4.5 and Theorem 1.4.7 and the trivial observation that $E(b) = b$ for any real number b, we get the following

Theorem 1.6.1. *Suppose $X, Y \geq 0$ or $E[X], E[Y] < \infty$.*
(a) $E(X + Y) = EX + EY$.
(b) $E(aX + b) = aE(X) + b$ for any real numbers a, b.
(c) If $X \geq Y$, then $EX \geq EY$.

Theorem 1.6.2. ***Jensen's inequality***. *Suppose φ is convex, that is,*

$$\lambda\varphi(x) + (1 - \lambda)\varphi(y) \geq \varphi(\lambda x + (1 - \lambda)y)$$

for all $\lambda \in (0, 1)$ and $x, y \in \mathbb{R}$. Then

$$E(\varphi(X)) \geq \varphi(EX)$$

provided both expectations exist, i.e., $E[X]$ and $E[\varphi(X)] < \infty$.

Theorem 1.6.3. ***Hölder's inequality***. *If $p, q \in [1, \infty]$ with $1/p + 1/q = 1$ then*

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

Here $||X||_p = (E[X]^r)^{1/r}$ for $r \in [1, \infty)$; $||X||_\infty = \inf\{M : P(|X| > M) = 0\}$.

To state our next result, we need some notation. If we only integrate over $A \subset \Omega$, we write

$$E(X; A) = \int_A X dP$$

Theorem 1.6.4. ***Chebyshev's inequality***. *Suppose $\varphi : \mathbb{R} \to \mathbb{R}$ has $\varphi \geq 0$, let $A \in \mathcal{R}$ and let $i_A = \inf\{\varphi(y) : y \in A\}$.*

$$i_A P(X \in A) \leq E(\varphi(X); X \in A) \leq E\varphi(X)$$

Proof. The definition of i_A and the fact that $\phi \geq 0$ imply that

$$i_A 1_{(X \in A)} \leq \varphi(X) 1_{(X \in A)} \leq \varphi(X)$$

So taking expected values and using part (c) of Theorem 1.6.1 gives the desired result.

□

Remark 1.6.5. Some authors call this result Markov's inequality and use the name Chebyshev's inequality for the special case in which $\varphi(x) = x^2$ and $A = \{x : |x| \geq a\}$:

$$a^2 P(|X| \leq a) \leq EX^2 \tag{1.5}$$

Theorem 1.6.6. ***Fatou's lemma***. *If $X_n \geq 0$, then*

$$\liminf_{n\to\infty} EX_n \geq E(\liminf_{n\to\infty} X_n)$$

Theorem 1.6.7. ***Monotone convergence theorem***. *If $0 \leq X_n \uparrow X$, then $EX_n \uparrow EX$.*

Theorem 1.6.8. ***Dominated convergence theorem***. *If $X_n \to X$ a.s., $|X_n| \leq Y$ for all n, and $EY < \infty$, then $EX_n \to EX$.*

The special case of Theorem 1.6.7 in which Y is constant is called bounded convergence theorem.

Theorem 1.6.9. *Suppose $X_n \to X$ a.s., Let g, h be continuous functions with*
(i) $g \geq 0$ and $g(x) \to \infty$ as $|x| \to \infty$,
(ii) $|h(x)|/g(x) \to 0$ as $|x| \to \infty$, and
(iii) $Eg(X_n) \leq K < \infty$ for all n. Then $Eh(X_n) \to Eh(X)$.

Proof. By subtracting a constant from h, we can suppose without loss of generality that $h(0) = 0$. Pick M large so that $P(|X| = M) = 0$ and $g(x) > 0$ when $|x| \geq M$. Given a random variable Y, let $\bar{Y} = Y 1_{(|Y| \leq M)}$. Since $P(|X| = M) = 0$, $\bar{X}_n \to \bar{X}$ a.s. Since $h(\bar{X}_n)$ is bounded and h is continuous, it follows from the bounded convergence theorem that

$$Eh(\bar{X}_n) \to Eh(\bar{X}) \tag{a}$$

To control the effect of the truncation, we use the following:

$$|Eh(\bar{Y}) - Eh(Y)| \leq E|h(\bar{Y}) - h(Y)| \leq E(h(Y)|; |Y| > M) \leq \epsilon_M Eg(Y) \tag{b}$$

where $\epsilon_M = \sup\{|h(x)|/g(x) : |x| \leq M\}$. To check the second inequality, note that when $|Y| \leq M$, $\bar{Y} = Y$, and we have supposed $h(0) = 0$. The third inequality follows from the definition of ϵ_M.

Taking $Y = X_n$ in (b) and using (iii), it follows that

$$|Eh(\bar{X}_n) - Eh(X_n)| \leq K_{\epsilon M} \tag{c}$$

To estimate $|Eh(\bar{X}) - Eh(X)|$, we observe that $g \geq 0$ and g is continuous, so Fatou's lemma implies

$$Eg(X) \leq \liminf_{n\to\infty} Eg(X_n) \leq K$$

Taking $Y = X$ in (b) gives

$$|Eh(\bar{X}) - Eh(X)| \leq K_{\epsilon M} \tag{d}$$

The triangle inequality implies

$$\begin{aligned} |Eh(X_n) - Eh(X)| &\leq |Eh(X_n) - Eh(\bar{X}_n)| \\ &\quad + |Eh(\bar{X}_n) - Eh(\bar{X})| + |Eh(\bar{X} - Eh(X)| \end{aligned}$$

Taking limits and using (a), (c), (d), we have

$$\limsup_{n\to\infty} |Eh(X_n) - Eh(X)| \leq 2K_{\epsilon M}$$

which proves the desired result since $K < \infty$ and $\epsilon_M \to 0$ as $M \to \infty$.

□

Integrating over $(\Omega, \mathcal{F}, \mathcal{P})$ is nice in theory, but to do computations we have to shift to a space on which we can do calculus in most cases, we will apply the next result with $S = \mathbb{R}^d$.

Theorem 1.6.10. ***Change of variables formula.*** *Let X be a random element of $(S, \mathcal{S})$ with distribution μ, i.e., $\mu(A) = P(X \in A)$. If f is a measurable function from $(S, \mathcal{S})$ to $(\mathbb{R}, \mathcal{R})$ so that $f \geq 0$ or $E[f(X)] < \infty$, then*

$$Ef(X) = \int_S f(y)\mu(dy)$$

Remark 1.6.11. To explain the name, write h for X and $P \cdot h^{-1}$ for μ to get

$$\int_\Omega f(h(\omega))dP = \int_S f(y)d(P \cdot h^{-1})$$

Proof. We prove this result by verifying it in four increasingly more general special cases that parallel the way that integral was defined Section 4.

CASE 1: INDICATOR FUNCTIONS. If $B \in \mathcal{S}$ and $f = 1_B$ then recalling the relevant definitions shows

$$E1_B(X) = P(X \in B) = \mu(B) = \int_S 1_B(y)\mu(dy)$$

CASE 2: SIMPLE FUNCTIONS. Let $f(x) = \sum_{m=1}^n c_m 1_{B_m}$ where $c_m \in \mathbb{R}$, $B_m \in \mathcal{S}$. The linearity of expected value, the result of Case 1, and the linearity of integration imply

$$\begin{aligned} Ef(x) &= \sum_{m=1}^n c_m E1_{B_m}(X) \\ &= \sum_{m=1}^n c_m \int_S 1_{B_m}(y)\mu(dy) = \int_S f(y)\mu(dy) \end{aligned}$$

CASE 3: NONNEGATIVE FUNCTIONS. Now if $f \geq 0$ and we let

$$f_n(x) = (|2^n f(x)|/2^n) \wedge n$$

where $|x| =$ the largest integer $\leq x$ and $a \wedge b = \min\{a, b\}$, then the f_n are simple and $f_n \uparrow f$, so using the result for simple functions and the monotone convergence theorem:

$$Ef(x) = \lim_n Ef_n(X) = \lim_n \int_S f_n(y)\mu(dy) = \int_S f(y)\mu(dy)$$

CASE 4: INTEGRABLE FUNCTIONS. The general case now follows by writing $f(x) = f(x)^+ - f(x)^-$. The condition $E|f(X)| < \infty$ guarantees that $Ef(X)^+$ and

$Ef(X)^-$ are finite. So using the result for nonnegative functions and linearity of expected value and integration:

$$\begin{aligned} Ef(X) = Ef(X)^+ - Ef(X)^- &= \int_S f(y)^+ \mu(dy) - \int_S f(y)^- \mu(dy) \\ &= \int_S f(y)\mu(dy) \end{aligned}$$

which completes the proof.

□

A consequence of Theorem 1.6.9 is that we can compute expected values of functions of random variables by performing integrals on the real line. If k is a positive integer then EX^k is called the kth moment of X. The first moment EX is usually called the mean and denoted by μ. If $EX^2 < \infty$ then the variance of X is defined to be $\text{var}(X) = E(X-\mu)^2$. To compute the variance use the following:

$$\begin{aligned} \text{var}(X) &= E(X-\mu)^2 \\ &= EX^2 - 2\mu EX + \mu^2 = EX^2 - \mu^2 \end{aligned} \tag{1.6}$$

From this it is immediate that

$$\text{var}(X) \le EX^2 \tag{1.7}$$

Here EX^2 is the expected value of X^2. When we want the square of EX, we will write $(EX)^2$. Since $E(aX+b) = aEX + b$ by (b) of Theorem 1.6.1, it follows easily from the definition that

$$\begin{aligned} \text{var}(aX+b) &= E(aX+b-E(aX+b))^2 \\ &= a^2 E(X-EX)^2 = a^2 \text{var}(X) \end{aligned} \tag{1.8}$$

Now we turn to concrete examples.

Example 1.6.12. If X has an exponential distribution with rate 1 then

$$EX^k = \int_0^\infty x^k e^{-x} dx = k!$$

So the mean of X is 1 and variance is $EX^2 - (EX)^2 = 2 - 1^2 = 1$. If we let $Y = X/\lambda$, Y has density $\lambda e^{-\lambda y}$ for $y \ge 0$, the exponential density with parameter λ. From (b) of Theorem 1.6.1 and equation 1.8, it follows that Y has mean $1/\lambda$ and variance $1/\lambda^2$.

Example 1.6.13. If X has a standard normal distribution

$$EX = \int x(2\pi)^{-1/2} \exp(-x^2/2) dx = 0 \text{ (by symmetry)}$$

$$\text{var}(X) = EX^2 = \int x^2 (2\pi)^{-1/2} \exp(-x^2/2) dx = 1$$

If we let $\sigma > 0$, $\mu \in \mathbb{R}$, and $Y = \sigma X + \mu$, then (b) of Theorem 1.6.1 and 1.8 imply $EY = \mu$ and $\text{var}(Y) = \sigma^2$. Y has density

$$(2\pi\sigma^2)^{-1/2} \exp(-(y-\mu)^2/2\sigma^2)$$

the normal distribution with mean μ and variance σ^2.

Example 1.6.14. We say that X has a Bernoulli distribution with parameter p if $P(X = 1) = p$ and $P(X = 0) = 1 - p$. Clearly,

$$EX = p \cdot 1 + (1 - p) \cdot 0 = p$$

Since $X^2 = X$, we have $EX^2 = EX = p$ and

$$\text{var}(X) = EX^2 - (EX)^2 = p - p^2 = p(1 - p)$$

Example 1.6.15. We say that X has a Poisson distribution with parameter λ if

$$P(X = k) = e^{-\lambda}\lambda^k/k! \text{ for } k = 0, 1, 2, ...$$

To evaluate the moments of the Poisson random variable, we use a little inspiration to observe that for $k \geq 1$

$$\begin{aligned} E(X(X-1)\dots(X-k+1)) &= \sum_{j=k}^{\infty} j(j-1)\dots(j-k+1)e^{-\lambda}\frac{\lambda^j}{j!} \\ &= \lambda^k \sum_{j=k}^{\infty} e^{-\lambda}\frac{\lambda^{j-k}}{(j-k)!} = \lambda^k \end{aligned}$$

where the equalities follow from the facts that (i) $j(j-1)\dots(j-k+1) = 0$ when $j < k$, (ii) canceling part of the factorial, (iii) the fact that Poisson distribution has total mass 1. Using the last formula, it follows that $EX = \lambda$ while

$$\text{var}(X) = EX^2 - (EX)^2 = E(X(X-1)) + EX - \lambda^2 = \lambda$$

Example 1.6.16. N is said to have a geometric distribution with success probability $p \in (0, 1)$ if

$$P(N = k) = p(1-p)^{k-1} \text{ for } k = 1, 2, ...$$

N is the number of independent trials needed to observe an event with probability p. Differentiating the identity

$$\sum_{k=0}^{\infty}(1-p)^k = 1/p$$

and

$$-\sum_{k=1}^{\infty} k(1-)^{k-1} = -1/p^2$$

$$\sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2} = 2/p^3$$

From this it follows that

$$EN = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = 1/p$$

$$EN(N-1) = \sum_{k=1}^{\infty} k(k-1)p(1-p)^{k-1} = 2(1-p)/p^2$$

$$\begin{aligned} \text{var}(N) &= EN^2 - (EN)^2 = EN(N-1) + EN - (EN)^2 \\ &= \frac{2(1-p)}{p^2} + \frac{p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2} \end{aligned}$$

1.7 Product Measures, Fubini's Theorem

Let $(X, \mathcal{A}, \mu_1)$ and $(Y, \mathcal{B}, \mu_2)$ be two σ-finite measure spaces. Let

$$\Omega = X \times Y = \{(x, y) : x \in X, y \in Y\}$$

$$\mathcal{S} = \{A \times B : A \in \mathcal{A}, B \in \mathcal{B}\}$$

Sets in $\mathcal{S}$ are called rectangles. It is easy to see that $\mathcal{S}$ is a semi-algebra:

$$(S \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$$

$$(A \times B)^c = (A^c \times B) \cup (A \times B^c) \cup (A^c \times B^c)$$

Let $\mathcal{F} = \mathcal{A} \times \mathcal{B}$ be the σ-algebra generated by $\mathcal{S}$.

Theorem 1.7.1. *There is a unique measure μ on $\mathcal{F}$ with*

$$\mu(A \times B) = \mu_1(A)\mu_2(B)$$

Notation. μ is often denoted by $\mu_1 \times \mu_2$.

Proof. By Theorem 1.1.9 it is enough to show that if $A \times B = +_i(A_i \times B_i)$ is a finite or countable disjoint union then

$$\mu(A \times B) = \sum_i \mu(A_i \times B_i)$$

For each $x \in A$, let $I(x) = \{i : x \in A_i\}$. $B = +_{i \in I(x)} B_i$, so

$$1_A(x)\mu_2(B) = \sum_i 1_{A_i}(x)\mu_2(B_i)$$

Integrating with respect to μ_1 and

$$\mu_1(A)\mu_2 = \sum_i \mu_1(A_i)\mu_2(B_i)$$

which proves the result.

□

Using Theorem 1.7.1 and induction, it follows that if $(\Omega - i, \mathcal{F}_i, \mu_i)$, $i = 1, ..., n$, are σ-finite measures spaces and $\Omega = \Omega_1 \times \cdots \times \Omega_n$, there is a unique measure μ on the σ-algebra $\mathcal{F}$ generated by sets of the form $A_1 \times \cdots \times A_n$, $A_i \in \mathcal{F}_i$, that has

$$\mu(A_1 \times \cdots \times A_n) = \prod_{m-1}^{n} \mu_m(A_m)$$

When $(\Omega_i, \mathcal{F}_i, \mu_i) = (\mathbb{R}, \mathcal{R}, \lambda)$ for all i, the result is Lebesgue measure on the Borel subsets of n dimensional Euclidean space $\mathbb{R}^n$.

Returning to the case in which $(\Omega, \mathcal{F}, \mu)$ is the product of two measure spaces, $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$, our next goal is to prove:

Theorem 1.7.2. ***Fubini's Theorem.*** *If $f \geq 0$ or $\int |f| d\mu < \infty$ then*

$$(\star) \int_X \int_Y f(x, y)\mu_2(dy)\mu_1(dx) = \int_{X \times Y} f d\mu = \int_Y \int_X f(x, y)\mu_1(dx)\mu_2(dy)$$

Proof. We will prove only the first equality, since the second follows by symmetry. Two technical things that need to be proved before we can assert that the first integral makes sense are:

When x is fixed, $y \to f(x,y)$ is $\mathcal{B}$ measurable.

$x \to \int_Y f(x,y)\mu_2(dy)$ is $\mathcal{A}$ measurable.

We begin with the case $f = 1_E$. Let $E_x = \{y : (x,y) \in E\}$ be the cross-section at x.

Lemma 1.7.3. *If $E \in \mathcal{F}$ then $E_x \in \mathcal{B}$.*

Proof. $(E^c)_x = (E_x)^c$ and $(\cup_i E_i)_x = \cup_i (E_i)_x$, so if $\mathcal{E}$ is the collection of sets E for which $E_x \in \mathcal{B}$, then $\mathcal{E}$ is a σ-algebra. Since $\mathcal{E}$ contains the rectangles, the result follows.

□

Lemma 1.7.4. *If $E \in \mathcal{F}$ then $g(x) \equiv \mu_2(E_x)$ is $\mathcal{A}$ measurable and*

$$\int_X g d\mu_1 = \mu(E)$$

Notice that it is not obvious that the collection of sets for which the conclusion is true is a σ-algebra since $\mu(E_1 \cup E_2) = \mu(E_1) + \mu(E_2) - \mu(E_1 \cap E_2)$. Dynkin's $\pi - \lambda$ Theorem was tailor-made for situations like this.

Proof. If conclusions hold for E_n and $E_n \uparrow E$, then Theorem 1.3.5 and the monotone convergence theorem imply that they hold for E. Since μ_1 and μ_2 are σ-finite, it is enough then to prove the result for $E \subset F \times G$ with $\mu_1(F) < \infty$ and $\mu_2(G) < \infty$, or taking $\Omega = F \times G$ we can suppose without loss of generality that $\mu(\Omega) < \infty$. Let $\mathcal{L}$ be the collection of sets E for which the conclusions hold. We will now check that $\mathcal{L}$ is a λ-system. Property (i) of a λ-system is trivial. (iii) follows from the first sentence in the proof. To check (ii) we observe that

$$\mu_2((A-B)_x) = \mu_2(A_x - B_x) = \mu(A_x) - \mu_2(B_x)$$

and integrating over x gives the second conclusion. Since $\mathcal{L}$ contains the rectangles, a π-system that generates $\mathcal{F}$, the desired result follows from the $\pi - \lambda$ theorem.

□

We are now ready to prove Theorem 1.7.2 by verifying it in four increasingly more general special cases.

CASE 1. If $E \in \mathcal{F}$ and $f = 1_E$ then $(\star)$ follows from Lemma 1.7.4

CASE 2. Since each integral is linear in f, it follows that $(\star)$ holds for simple function functions.

CASE 3. Now if $f \geq 0$ and we let $f_n(x) = ([2^n f(x)]/2^n) \wedge n$, where $[x] =$ the largest integer $\leq x$, then the f_n are simple and $f_n \uparrow f$, so it follows from the monotone convergence theorem that $(\star)$ holds for all $f \geq 0$.

CASE 4. The general case now follows by writing $f(x) = f(x)^+ - f(x)^-$ and applying Case 3 to f^+, f^-, and $|f|$.

The following examples are to illustrate why the various hypotheses of Theorem 1.7.2 are needed.

Example 1.7.5. Let $X = Y = \{1, 2, ...\}$ with $\mathcal{A} = \mathcal{B} =$ all subsets and $\mu_1 = \mu_2 =$ counting measure . For $m \geq 1$, let $f(m,m) = 1$ and $f(m+1,m) = -1$, and let $f(m,n) = 0$ otherwise. We claim that

$$\sum_m \sum_n f(m,n) = 1 \text{ but } \sum_n \sum_m f(m,n) = 0$$

The following illustrates the equation:

$$\begin{array}{cccccc} & \vdots & \vdots & \vdots & \vdots & \\ & 0 & 0 & 0 & 1 & \dots \\ \uparrow & 0 & 0 & 1 & -1 & \dots \\ n & 0 & 1 & -1 & 0 & \dots \\ & 1 & -1 & 0 & 0 & \dots \\ & & m & \rightarrow & & \end{array}$$

That is, if we sum the columns first, the first one gives us a 1 and the others 0, while if we sum the rows each one gives us a 0.

It is worth noting that it is possible that a set A may not be measurable with respect to the product σ-field, but nevertheless the section A_{ω_1} and A_{ω_2} are all measurable, $P_2(A_{\omega_1})$ and $P_1(A_\omega 2)$ are measurable functions, but

$$\int P_1(A_{\omega_2})dP_2 \neq \int P_2(A_{\omega_1})dP_1$$

Example 1.7.6. Let $X = (0,1)$, $Y = (1,\infty)$, both equipped with the Borel sets and Lebesgue measure. Let $f(x,y) = e^{-xy} - 2e^{-2xy}$.

$$\int_0^1 \int_1^\infty f(x,y)dydx = \int_0^1 x^{-1}(e^{-x}e^{-2x})dx > 0$$

$$\int_1^\infty \int_0^1 f(x,y)dxdy = \int_1^\infty y^{-1}(e^{-2y} - 2^{-y})dy < 0$$

Example 1.7.7. Let $X = (0,1)$ with $\mathcal{A} =$ the Borel sets and $\mu_1 =$ Lebesgue measure. Let $Y = (0,1)$ with $\mathcal{B} =$ all subsets and $\mu_2 =$ counting measure. Let $f(x,y) = 1$ if $x = y$ and 0 otherwise

$$\int_Y f(x,y)\mu_2(dy) = 1 \text{ for all } x \text{ so } \int_X \int_Y f(x,y)\mu_2(dy)\mu_1(dx) = 1$$

$$\int_X f(x,y)\mu_1(dx) = 0 \text{ for all } y \text{ so } \int_Y \int_X f(x,y)\mu_1(dy)\mu_2(dx) = 0$$

The last example shows that measurability is important or may be that some of the axioms of set theory are not as innocent as they seem.

Example 1.7.8. By the axiom of choice and the continuum hypothesis one can define an order relation $<'$ on $(0,1)$ so that $\{x : x <' y\}$ is countable for each y. Let $X = Y = (0,1)$, let $\mathcal{A} = \mathcal{B} =$ the Borel sets and $\mu_1 = \mu_2 = Lebesguemeasure$. Let $f(x,y) = 1$ if $x <' y$, $= 0$ otherwise. Since $\{x : x <' y\}$ and $\{y : x <' y\}^c$ are countable

$$\int_X f(x,y)\mu_1(dx) = 0 \text{ for all } y$$

$$\int_Y f(x,y)\mu_2(dy) = 1 \text{ for all } x$$

1.8 Laplace Method

1.8.1 Laplace's Method for Analytic Approximation of Integrals

Instead of approximating just the posterior with normal distribution, we can use *Laplace's method* to approximate integrals of a smooth function times the posterior

$h(\theta)p(\theta|y)$. The approximation is proportional to a (multivariate) normal density in θ, and its integral is just

$$\text{approximation of } \mathbb{E}(h(\theta)|y): \ h(\theta_0)p(\theta_0|y)(2\pi)^{d/2}|-\mu''(\theta_0)|^{1/2},$$

where d is the dimension of θ, $u(\theta) = \log(h(\theta)p(\theta|y))$, and θ_0 is the point at which $u(\theta)$ is maximized.

If $h(\theta)$ is a fairly smooth function, this approximation can be reasonable, due to the approximate normality of the posterior distribution, $p(\theta|y)$, for large sample sizes [8].

1.8.2 Banach Space

Let $(X, |\cdot|)$ be a Banach Space and

$$X_b := \{(x_n)_{n\in\mathbb{N}} \subset X : ||(x_n)_{n\in\mathbb{N}}||_{X_b} < \infty\}$$

with

$$||(X_n)_{n\in\mathbb{N}}||_{X_b} := \sup_{n\in\mathbb{N}}|x_n|.$$

Proof. We can prove the above claim of Banach Space by doing the following. Suppose $(z_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in X_b, with $z_n = (x_{n,m})_{m\in\mathbb{N}}$ for each n. Fix $\epsilon > 0$. Then there is some $N \in \mathbb{N}$ such that

$$||z_{n_1} - z_{n_2}||_{X_b} < \epsilon$$

for all $n_1, n_2 \geq N$. But then for each $m \in \mathbb{N}$ we have

$$|x_{n_1,m} - x_{n_2,m}| \leq ||z_{n_1} - z_{n_2}||_{X_b} < \epsilon,$$

This means that the sequence $(x_{n,m})_{n\in\mathbb{N}}$ is Cauchy in X, and thus convergent to some $x_m \in X$. This gives us a sequence $z = (x_m)_{m\in\mathbb{N}}$ in X. Now we have to show two things:

(1) $z \in X_b$, and

(2) $||z_n - z||_{X_b} \to 0$ as $n \to \infty$.

The first is okay. Given $m \in \mathbb{N}$, there is some $n \in \mathbb{N}$ such that $|x_m - x_{n,m}| < \epsilon$, and thus

$$|x_m| < |x_{n,m}| + \epsilon \leq ||z_n||_{X_b} + \epsilon \leq \sup_{n\in\mathbb{N}}||z_n||_{X_b} + \epsilon,$$

and we know $\sup_{n\in\mathbb{N}}||z_n||_{X_b} < \infty$ since $(z_n)_{n\in\mathbb{N}}$ is Cauchy. Since $m \in \mathbb{N}$ was arbitrary, we have

$$\sup_{m\in\mathbb{M}}|x_m| \leq \sup_{n\in\mathbb{N}}||z_n||_{X_b} + \epsilon < \infty,$$

thus $z \in X_b$, so (1) is shown. Now to show (2), since $(z_n)_{n\in\mathbb{N}}$ is Cauchy, there is some $N' \in \mathbb{N}$ such that $||z_{n_1} - z_{n_2}||_{X_b} < \epsilon/2$ whenever $n_1, n_2 \geq N'$. Now let $n \geq N'$ be given. Then for any $m \in \mathbb{N}$, there is some $n' \geq N'$ such that $|x_m - x_{n',m}| < \epsilon/2$. Thus we have

$$|x_m - x_{n,m}| \leq |x_m - x_{n',m}| + |x_{n',m} - x_{n,m}| < \epsilon/2 + \epsilon/2 = \epsilon.$$

Now taking the supremum over m, we have

$$||z - z_n||_{X_b} \leq \epsilon,$$

and thus (2) is shown.

□

Let X be a Banach space with norm $||\cdot||_X$. Define

$$||f||_p := \left(\sum_{n=1}^{\infty} ||f(n)||_X^p \right)^{1/p}$$

for any function $f : \mathbb{N} \to X$. Let $l^p := \{f : \mathbb{N} \to X : ||f||_p < \infty\}$. Show that l^p is also a Banach space.

Proof. Choose a k such that $||f_i - f_j||_p < 1$ for $i, j \geq k$. Then, for every fixed $N \in \mathbb{N}$, you have

$$\left(\sum_{n=1}^{N} ||f_i(n) - f_k(n)||^p \right)^{1/p} \leq ||f_i - f_k||_p < 1$$

for every $i > k$. Since there are only finitely many terms in the sum, you can take the limit $i \to \infty$, and obtain

$$\left(\sum_{n=1}^{N} ||f(n) - f_k(n)||^p \right)^{1/p} \leq 1$$

Since that inequality holds for all N, you have $(f - f_k) \in l^p$ and $||f - f_k||_p \leq 1$. Since l^p is a vector space and $f_k \in l^p$, it follows that $f = f_k + (f - f_k) \in l^p$. Showing that $f_k \to f$ in l^p can be done same way.

□

1.8.3 Minkowski's Theorem

In mathematics, Minkowski's theorem is the statement that any convex set in $\mathbb{R}^n$ which is symmetric with respect to the origin and with volume greater than $2^n d(L)$ contains a non-zero lattice point. The theorem was proved by Hermann Minkowski in 1889 and became the foundation of the branch of number theory called the geometry of numbers.

Suppose that L is a lattice of determinant $d(L)$ in the n-dimensional real vector space $\mathbb{R}^n$ and S is a convex subset of $\mathbb{R}^n$ that is symmetric with respect to the origin, meaning that if x is in S then $-x$ is also in S. Minskowski's theorem states that if the volume of S is strictly greater than $2^n d(L)$, then S must contain at least one lattice point other than the origin. (Since the set S is symmetric, it would then contain at least three lattice points: the origin 0 and a pair of points $\pm x$, where $x \in L \setminus 0$.)

For example, the simplest case of a lattice is the set $\mathbb{Z}^n$ of all points with integer coefficients; its determinant is 1. For $n = 2$, the theorem claims that a convex figure in the plane symmetric about the origin and with area greater than 4 encloses at least one lattice point in addition to the origin. The area bound is sharp: if S is the interior of the square with vertices $(\pm 1, \pm 1)$ then S is symmetric and convex, has area 4, but the only lattice point it contains is the origin. This observation generalizes to every dimension n.

Theorem 1.8.1. ***Minkowski theorem***. *The theorem states that*

$$||f + g||_p \leq ||f||_p + ||g||_p \leq 2^{p-1}(||f||_p + ||g||_p)$$

Proof. Consider $||f||_p$ and $||g||_p$ finite. We have the following

$$\begin{aligned} ||f+g||_p^p &= \int |f+g|^p d\mu \\ &= \int |f+g||f+g|^{p-1} d\mu \\ &\leq \int |f||f+g|^{p-1} + \int |g||f+g|^{p-1} \\ &\leq (||f||_p + ||g||_q) \frac{||f+g||_p^p}{||f+g||_p} \\ &\leq ||f||_p + ||g||_p \end{aligned}$$

so we conclude that l^p has non-zero elements with zero semi-norm. The identity $f = g$ a.e. and tends to norm. Thus, we have a Banach space. That is,

$$L^\infty := \{f : ||f||_\infty < \infty\} \text{ while } ||f||_\infty = \inf\{\mu \geq 0 : \mu\{q : |f(x)| > \mu\} = 0\}$$

□

1.8.4 Riesz-Fischer Theorem

In his Note, Riesz (1907, p. 616) states [13] states the following result (translated here to modern language at one point: the notation $L^2([a,b])$ was not used in 1907).

"Let $\{\varphi\}$ be an orthonormal system in $L^2([a,b])$ and $\{a_n\}$ a sequence of reals. The convergence of the series $\sum a_n^2$ is a necessary and sufficient condition for the existence of a function f such that

$$\int_a^b f(x)\varphi_n(x)dx = a_n \forall n$$

"

Today, this result of Riesz is a special case of basic facts about series of orthogonal vectors in Hilbert spaces. The most common form of the theorem states that a measurable function on $[-\pi, \pi]$ is square integrable if and only if the corresponding Fourier series converges in the space L^2. This means that if the Nth partial sum of the Fourier series corresponding to a square-integrable function f is given by

$$S_N f(x) = \sum_{n=-N}^{N} F_n e^{inx},$$

where F_n, the nth Fourier coefficient, is given by

$$F_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} dx,$$

then

$$\lim_{N\to\infty} ||S_N f - f||_2 = 0,$$

where $||\cdot||_2$ is the L^2-norm.

The Riesz-Fischer theorem also applies in a more general setting. Let R be an inner product space consisting of functions (for example, measurable functions on the line, analytic functions in the unit disc; in old literature, sometimes called Euclidean Space),

and let $\{\varphi_n\}$ be an orthonormal system in R (e.g. Fourier basis, Hermite or Laguerre polynomials, etc. - see orthogonal polynomials), not necessarily complete (in an inner product space, an orthonormal set is complete if no nonzero vector is orthogonal to every vector in the set).

2 LAW OF LARGE NUMBERS

Go back to Table of Contents. Please click TOC

2.1 Independence

Measure theory ends and probability begins with the definition of independence. We begin with what is hopefully a familiar definition and work our way up to definition for current settings. Two events A and B are independent if $P(A \cap B) = P(A)P(B)$. Two random variables X and Y are independent if for all $C, D \in \mathcal{R}$.

$$P(X \in C, Y \in D) = P(X \in C)P(Y \in D)$$

i.e., the events $A = \{X \in C\}$ and $B = \{Y \in D\}$ are independent. Two σ-fields $\mathcal{F}$ and $\mathcal{G}$ are independent if for all $A \in \mathcal{F}$ and $B \in \mathcal{G}$ the events A and B are independent.

We take things in the opposite order when we say what it means for several things to be independent. We begin by reducing to the case of finitely many objects. An infinite collection of objects (σ-fields, random variables, or sets) is said to be independent if every finite subcollection is.

σ-fields $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_n$ are independent if whenever $A_i \in \mathcal{F}_i$ for $i = 1, ..., n$, we have

$$\mathcal{P}(\cap_{i=1}^n A_i) = \prod_{i=1}^n \mathcal{P}(A_i)$$

Random variables $X_1, ..., X_n$ are independent if whenever $B_i \in \mathcal{R}$ for $i = 1, ..., n$ we have

$$P(\cap_{i=1}^n \{X_i \in B_i\}) = \prod_{i=1}^n P(X_i \in B_i)$$

Sets $A_1, ..., A_n$ are independent if whenever $I \subset \{1, ..., n\}$ we have

$$P(\cap_{i \in I} A_i) \prod_{i \in I} P(A_i)$$

At first glance, it might seem that the last definition does not match the other two. However, if you think about it for a minute, you will see that if the indicator variables 1_{A_i}, $1 \leq i \leq n$ are independent and we take $B_i = \{1\}$ for $i \in I$, and $B_i \in \mathbb{R}$ for $i \notin I$ then the condition in the definition results.

Example 2.1.1. Let X_1, X_2, X_3 be independent random variables with

$$P(X_i = 0) = P(X_i = 1) = 1/2$$

Let $X_1 = \{X_2 = X_3\}$, $A_2 = \{X_3 = X_1\}$ and $A_3 = \{X_1 = X_2\}$. These events are pairwise independent since if $i \neq j$ then

$$P(A_i \cap A_j) = P(X_1 = X_2 = X_3) = 1/4 = P(A_i)P(A_j)$$

but they are not independent since

$$P(A_1 \cap A_2 \cap A_3) = 1/4 \neq 1/8 = P(A_1)P(A_2)P(A_3)$$

In order to show that random variables X and Y are independent, we have to check that $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for all Borel sets A and B.

Collections of sets $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_n \subset \mathcal{F}$ are said to be independent if whenever $A_i \in \mathcal{A}_i$ and $I \subset \{1, ..., n\}$ we have $P(\cap_{i \in I} A_i) = \prod_{i \in I} P(A_i)$. If each collection is a single set i.e., $\mathcal{A}_i = \{A_i\}$, this definition reduces to the one for sets.

Lemma 2.1.2. *Without loss of generality we can suppose each $\mathcal{A}_i$ contains Ω. In this case the condition is equivalent to*

$$P(\cap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i) \text{ whenever } A_i \in \mathcal{A}_i$$

since we can set $A_i = \Omega$ for $i \notin I$.

Proof. If $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_n$ are independent and $\bar{\mathcal{A}}_i = \mathcal{A}_i \cup \{\Omega\}$ then $\bar{\mathcal{A}}_1, \bar{\mathcal{A}}_2, ..., \bar{\mathcal{A}}_n$ are independent, since if $A_i \in \bar{\mathcal{A}}_i$ and $I = \{j : A_j = \Omega\} \cap_i A_i = \cap_{i \in I} A_i$.

□

Let us notate the following. Say collection of sets $\mathcal{P}$ is a π-system if closed under intersection. Say a collection of sets $\mathcal{L}$ is a λ-system if (i) $\Omega \in \mathcal{L}$; (ii) if $A, B \in \mathcal{L}$ and $A \subset B$, then $B - A \in \mathcal{L}$; (iii) if $A_n \in \mathcal{L}$ and $A_n \uparrow A$ then $A \in \mathcal{L}$.

Theorem 2.1.3. $\pi - \lambda$ ***Theorem***. *If $\mathcal{P}$ is a π-system and $\mathcal{L}$ is a λ-system that contains $\mathcal{P}$ then $\sigma(\mathcal{P}) \subset \mathcal{L}$.*

Theorem 2.1.4. *Suppose $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_n$ are independent and each $\mathcal{A}_i$ is a π-system. Then $\sigma(\mathcal{A}_1), \sigma(\mathcal{A}_2), ..., \sigma(A_n)$ are independent.*

The proof of Theorem 2.1.4 is based on Dynkin's $\pi - \lambda$ theorem. To state this result, we need definitions. We say that $\mathcal{A}$ is a π-system if it is closed under intersection, i.e., if $A, B \in \mathcal{A}$ then $A \cap B \in \mathcal{A}$. We say that $\mathcal{L}$ is a λ-system if: (i) $\Omega \in \mathcal{L}$. (ii) If $A, B \in \mathcal{L}$ and $A \subset B$, then $B - A \in \mathcal{L}$. (iii) If $A_n \in \mathcal{L}$ and $A_n \uparrow A$ then $A \in \mathcal{A}$.

Proof. Let $A_2, ..., A_n$ be sets with $A_i \in \mathcal{A}_i$, let $F = A_2 \cap \cdots \cap A_n$ and let $\mathcal{L} = \{A : P(A \cap F) = P(A)P(F)\}$. Since $P(\Omega \cap F) = P(\Omega)P(F)$, $\Omega \in \mathcal{L}$. To check (ii) of the definition of a λ-system, we note that if $A, B \in \mathcal{L}$ with $A \subset B$ then $(B - A) \cap F = (B \cap F) - (A \cap F)$. So using (i) in Theorem 1.1.1, the fact $A, B \in \mathcal{L}$ and then (i) in Theorem 1.1.1 again

$$\begin{aligned} P((B - A) \cap F) &= P(B \cap F) - P(A \cap F) = P(B)P(F) - P(A)P(F) \\ &= \{P(B) - P(A)\}P(F) = P(B - A)P(F) \end{aligned}$$

and we have $B - A \in \mathcal{L}$. To check (iii) let $B_k \in \mathcal{L}$ with $B_k \uparrow B$ and note that $(B_k \cap F) \uparrow (B \cap F)$ so using (iii) in Theorem 1.1.1, the fact that $B_k \in \mathcal{L}$, and then (iii) in Theorem 1.1.1 again

$$P(B \cap F) \lim_k P(B_k \cap F) = \lim_k P(B_k)P(F) = P(B)P(F)$$

Applying the $\pi - \lambda$ theorem now gives $\mathcal{L} \supset \sigma(\mathcal{A}_1)$. It follows that if $A_1 \in \sigma(\mathcal{A}_1)$ and $A_i \in \mathcal{A}_i$ for $2 \leq i \leq n$ then

$$P(\cap_{i=1}^n A_i) = P(A_1)P(\cap_{i=2}^n A_i) = \prod_{i=1}^n P(A_i)$$

Using Lemma 2.1.2 now, we have ($\star$) If $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_n$ are independent then $\sigma(\mathcal{A}_1)$, $\mathcal{A}_2, ..., \mathcal{A}_n$ are independent. Applying ($\star$) to $\mathcal{A}_2, ..., \mathcal{A}_n$, $\sigma(\mathcal{A}_1)$ (which are independent since the definition is unchanged by permuting the order) shows that $\sigma(\mathcal{A}_2), \mathcal{A}_3, ..., \mathcal{A}_n$, $\sigma(\mathcal{A}_1)$ are independent, and after n iterations we have the desired result.

□

Remark 2.1.5. The reader should note that it is not easy to show that if $A, B \in \mathcal{L}$ then $A \cap B \in \mathcal{L}$, or $A \cup B \in \mathcal{L}$, but it is easy to check that if $A, B \in \mathcal{L}$ with $A \subset B$ then $B - A \in \mathcal{L}$.

Theorem 2.1.6. *In order for $X_1, ..., X_n$ to be independent, it is sufficient that for all $x_1, ..., x_n \in (-\infty, \infty]$*

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

Proof. Let $\mathcal{A}_i =$ the sets of the form $\{X_i \leq x_i\}$. Since

$$\{X_i \leq x\} \cup \{X_i \leq y\} = \{X_i \leq x \wedge y\},$$

where $(x \wedge y)_i = x_i \wedge y_i = \min\{x_i, y_i\}$. $\mathcal{A}_i$ is a π-system. Since we have allowed $x_i = \infty$, $\Omega \in \mathcal{A}_i$.

□

Theorem 2.1.7. *Suppose $\mathcal{F}_{i,j}$, $1 \leq i \leq n$, $1 \leq j \leq m(i)$ are independent and let $\mathcal{G}_i = \sigma(\cup_i \mathcal{F}_{i,j})$. Then $\mathcal{G}_1, ..., \mathcal{G}_n$ are independent.*

Proof. Let $\mathcal{A}_i$ be the collection of sets of the form $\cap_j A_{i,j}$ where $A_{i,j} \in \mathcal{F}_{i,j}$. $\mathcal{A}_i$ is a π-system that contains Ω and contains $\cup_j \mathcal{F}_{i,j}$ so Theorem 2.1.4 implies $\sigma(\mathcal{A}_i) = \mathcal{G}_i$ are independent.

□

Theorem 2.1.8. *If for $1 \leq i \leq n$, $1 \leq j \leq m(i)$, $X_{i,j}$ are independent and $f_i : \mathbb{R}^{m(i)} \to \mathbb{R}$ are measurable then $f_i(X_{i,1}, ..., X_{i,m(i)})$ are independent.*

Proof. Let $\mathcal{F}_{i,j} = \sigma(X_{i,j})$ and $\mathcal{G}_i = \sigma(\cup_j \mathcal{F}_{i,j})$. Since $f_i(X_{i,1}, ..., X_{i,m(i)}) \in \mathcal{G}_i$, the results come from Theorem 2.1.7.

□

A concrete special case of Theorem 2.1.8 that we will use is: if $X_1, ..., X_n$ are independent then $X = X_1$ and $Y = X_2 \dots X_n$ are independent. Later, when we study sums $S_m = X_1 + \cdots + X_m$ of independent random variables $X_1, ..., X_n$, we will use Theorem 2.1.8 to conclude that if $m < n$ then $S_n - S_m$ is independent of the indicator function of the event $\{\max_{1 \leq k \leq m} S_k > x\}$.

The next goal is to obtain formulas for the distribution and expectation of independent random variables.

Theorem 2.1.9. *Suppose $X_1, ..., X_n$ are independent random variables and X_i has distribution μ_i, then $(X_1, ..., X_n)$ has distribution $\mu_1 \times \cdots \times \mu_n$.*

Proof. Using the definitions of (i) $A_1 \times \cdots \times A_n$, (ii) independence, (iii) μ_i, and (iv) $\mu_1 \times \cdots \times \mu_n$

$$\begin{aligned} P((X_1, ..., X_n) \in A_1 \times \cdots \times A_n) &= P(X_1 \in A_1, ..., X_n \in A_n) \\ &= \prod_{i=1}^{n} P(X_i \in A_i) \\ &= \prod_{i=1}^{n} \mu_i(A_i) \\ &= \mu_1 \times \cdots \times \mu_n(A_1 \times \cdots \times A_n) \end{aligned}$$

The last formula shows that the distribution of $(X_1, ..., X_n)$ and the measure $\mu_1 \times \dots \mu_n$ agree on sets of the form $A_1 \times \cdots \times A_n$, a π-system that generates $\mathcal{R}^n$. So theorem 2.1.3 implies they must agree.

□

Theorem 2.1.10. *Suppose X and Y are independent and have distributions μ and ν. If $h : \mathbb{R}^2 \to \mathbb{R}$ is a measurable function with $h \geq 0$ or $E|h(X,Y)| < \infty$ then*

$$Eh(X,Y) = \iint h(x,y)\mu(dx)\nu(dy)$$

In particular, if $h(x,y) = f(x)g(y)$ where $f, g : \mathbb{R} \to \mathbb{R}$ are measurable functions with $f, g \geq 0$ or $E|f(X)|$ and $E|g(Y)| < \infty$ then

$$Ef(X)g(Y) = Ef(X) \cdot Eg(Y)$$

Proof. Using Theorem 1.6.9 and Fubini's theorem (Theorem 1.7.2 we have

$$Eh(X,Y) = \int_{\mathbb{R}^2} hd(\mu \times \nu) = \iint h(x,y)\mu(dx)\nu(dy)$$

To prove the second result, we start with the result when $f, g \geq 0$. In this case, using the first result, the fact that $g(y)$ does not depend on x and then Theorem 1.6.9 twice we get

$$\begin{aligned} Ef(X)g(Y) &= \iint f(x)g(y)\mu(dx)\nu(dy) \\ &= \int g(y) \int f(x)\mu(dx)\nu(dy) \\ &= \int Ef(X)g(y)\nu(dy) \\ &= Ef(X)Eg(Y) \end{aligned}$$

Applying the result for nonnegative f and g to $|f|$ and $|g|$, shows $E|f(X)g(Y)| = E|f(X)|E|g(Y)| < \infty$, and we can repeat the last argument to prove the desired result.

□

Theorem 2.1.11. *If $X_1, ..., X_n$ are independent and have (a) $X_i \geq 0$ for all i, or (b) $E|X_i| < \infty$ for all i then*

$$E\big(\prod_{i=1}^{n} X_i\big) = \prod_{i=1}^{n} EX_i$$

i.e., the expectation on the left exists and has the value given on the right.

Proof. $X = X_1$ and $Y = X_2 \dots X_n$ are independent by Theorem 2.1.8 so taking $f(x) = |x|$ and $g(y) = |y|$ we have $E|X_1 \dots X_n = E|X_1|E|X_2 \dots X_n|$, and it follows by induction that if $1 \leq m \leq n$

$$E|X_m \dots X_n| = \prod_{i=m}^{n} E|X_k|$$

If the $X_i \geq 0$, then $|X_i| = X_i$ and the desired result follows from the special case $m = 1$. To prove the result in general note that the special case $m = 2$ implies $E|Y| = E|X_2 \dots X_n| < \infty$, so using Theorem 2.1.10 with $f(x) = x$ and $g(y) = y$ shows $E(X_1 \dots X_n) = EX_1 \dots E(X_2 \dots X_n)$, and the desired result follows by induction.

□

Example 2.1.12. It can happen that $E(XY) = EX \cdot EY$ without the variables being independent. Suppose the joint distribution of X and Y is given by the following table

		Y		
		1	0	−1
	1	0	a	0
X	0	b	c	b
	−1	0	a	0

where $a, b > 0$, $c \geq 0$, and $2a + 2b + c = 1$. Things are arranged so that $XY \equiv 0$. Symmetry implies $EX = 0$ and $EY = 0$, so $E(XY) = 0 = EXEY$. The random variables are not independent since

$$P(X = 1, Y = 1) = 0 \leq ab = P(X = 1)P(Y = 1)$$

Two random variables X and Y with EX^2, $EY^2 < \infty$ that have $EXY = EXEY$ are said to be uncorrelated. The finite second moments are needed so that we know $E|XY| < \infty$ by the Cauchy-Schwarz inequality.

Theorem 2.1.13. *If X and Y are independent, $F(X) = P(X \leq x)$, and $G(y) = P(Y \leq y)$, then*

$$P(X + Y \leq z) = \int F(z - y)dG(y)$$

The integral on the right-hand side is called the convolution of F and G and is denoted $F \star G(z)$. The meaning of $dG(y)$ will be explained in the proof.

Proof. Let $h(x, y) = 1_{(x+y\leq z)}$. Let μ and ν be the probability measures with distribution functions F and G. Since for fixed y

$$\int h(x, y)\mu(dx) = \int 1_{(\infty, z-y]}(x)\mu(dx) = F(z - y)$$

using Theorem 2.1.10 gives

$$\begin{aligned} P(X + Y \leq z) &= \iint \mathbf{1}_{(x+y\leq z)}\mu(dx)\nu(dy) \\ &= \int F(z - y)\nu(dy) \\ &= \int F(z - y)dG(y) \end{aligned}$$

The last equality is just a change of notation. We regard $dG(y)$ as a shorthand for "integrate with respect to the measure ν with distribution function G."

□

Theorem 2.1.14. *Suppose that X with density f and Y with distribution function G are independent. Then $X + Y$ has density*

$$h(x) = \int f(x - y)dG(y)$$

When Y has density g, the last formula can be written as

$$h(x) = \int f(x - y)g(y)dy$$

Proof. From Theorem 2.1.13, the definition of density function, and Fubini's theorem (Theorem 1.7.2), which is justified since everything is nonnegative, we get

$$\begin{aligned} P(X+Y\leq z) &= \int F(z-y)dG(y) \\ &= \iint_{-\infty}^{z} f(x-y)dxdG(y) \\ &= \int_{-\infty}^{z}\int f(x-y)dG(y)dx \end{aligned}$$

The last equation says that $X+Y$ has density $h(x)=\int f(x-y)dG(y)$. The second formula follows from the first when we recall the meaning of $dG(y)$ given in the proof of Theorem 2.1.13.

□

Example 2.1.15. The gamma density with parameters α and λ is given by

$$f(x)=\begin{cases} \lambda^{\alpha}x^{\alpha-1}e^{-\lambda x}/\Gamma(\alpha) & \text{for } x\geq 0 \\ 0 & \text{for } x<0 \end{cases}$$

where $\Gamma(\alpha)=\int_0^{\infty}x^{\alpha-1}e^{-x}dx$.

Theorem 2.1.16. *If $X=Gamma(\alpha,\lambda)$, and $Y=Gamma(\beta,\lambda)$ are independent, then $X+Y$ is $Gamma(\alpha+\beta,\lambda)$. Consequently if $X_1,...,X_n$ are independent exponential(λ) r.v. s, then $X_1+\cdots+X_n$, has a $Gamma(n,\lambda)$ distribution.*

Proof. Write $f_{X+Y}(z)$ for the density function of $X+Y$ and use Theorem 2.1.14

$$\begin{aligned} f_{X+Y}(x) &= \int_0^x \frac{\lambda^{\alpha}(x-y)^{\alpha-1}}{\Gamma(\alpha)}e^{-\lambda(x-y)}\frac{\lambda^{\beta}y^{\beta-1}}{e}^{-\lambda y}dy \\ &= \frac{\lambda^{\alpha+\beta}e^{-\lambda x}}{\Gamma(\alpha)\Gamma(\beta)}\int_0^x (x-y)^{\alpha-1}y^{\beta-1}dy \end{aligned}$$

so it suffices to show the integral is $x^{\alpha+\beta-1}\Gamma(\alpha)/\Gamma(\alpha+\beta)$. To do this, we begin by changing variables $y=xu$, $dy=xdu$ to get

$$x^{\alpha+\beta-1}\int_0^1(1-u)^{\alpha-1}u^{\beta-1}du=\int_0^x(x-y)^{\alpha-1}y^{\beta-1}dy \tag{2.1}$$

There are two ways to complete the proof at this point. The soft convolution is to note that we have shown that the density $f_{X+Y}(x)=c_{\alpha,\beta}e^{-\lambda}\lambda^{\alpha+\beta}x^{\alpha+\beta-1}$ where

$$c_{\alpha,\beta}=\frac{1}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1(1-u)^{\alpha-1}u^{\beta-1}du$$

There is only one norming constant $c_{\alpha,\beta}$ that makes this a probability distribution, so recalling the definition of the beta distribution, we must have $c_{\alpha,\beta}=1/\Gamma(\alpha+\beta)$.

The less elegant approach for those who cannot remember the definition of the beta is to prove the last equality by calculus. Rewriting 2.1 with the right-hand side on the left, multiplying each side by $\exp(-x)$, integrating from 0 to ∞, and then using

Fubini's Theorem on the right we have

$$\begin{aligned}\Gamma(\alpha+\beta)\int_0^1 (1-u)^{\alpha-1}u^{\beta-1}du &= \int_0^\infty\int_0^x y^{\beta-1}e^{-y}(x-y)^{\alpha-1}e^{-(x-y)}dydx \\ &= \int_0^\infty y^{\beta-1}e^{-y}\int_x^\infty (x-y)^{\alpha-1}e^{-(x-y)}dxdy \\ &= \Gamma(\alpha)\Gamma(\beta)\end{aligned}$$

which gives the first result. The second follows from the fact that a Gamma$(1,\lambda)$ is an exponential with parameter λ and induction.

□

Example 2.1.17. Normal distribution. We introduced the normal density with mean μ and variance a,

$$(2\pi a)^{-1/2}\exp(-(x-u)^2/2a).$$

Theorem 2.1.18. *If $X = Normal(\mu, a)$, and $Y = Normal(\nu, b)$ are independent, then $X + Y = Normal(0, b)$.*

Proof. It is enough to prove the result for $\mu = \nu = 0$. Suppose $Y_1 = \text{Normal}(0, a)$ and $Y_2 = \text{Normal}(0, b)$. Then Theorem 2.1.14 implies

$$f_{Y_1+Y_2}(z) = \frac{1}{2\pi\sqrt{ab}}\int e^{-x^2/2a}e^{-(z-x)^2/2b}dx$$

Dropping the constant in front, the integral can be rewritten as

$$\begin{aligned}\int \exp\big(-\frac{bx^2+ax^2-2axz+az^2}{2ab}\big)dx &= \int \exp\big(-\frac{a+b}{ab}\big\{x^2-\frac{2a}{a+b}xz+\frac{a}{a+b}z^2\big\}\big)dx \\ &= \int \exp\big(-\frac{a+b}{2ab}\big\{(x-\frac{a}{a+b}z)^2+\frac{ab}{(a+b)^2}z^2\big\}\big)dx\end{aligned}$$

since $\{a/(a+b)\}^2 + \{a/(a+b)\} = ab/(a+b)^2$. Factoring out the term that does not depend on x, the last integral

$$\begin{aligned}&= \exp\big(-\frac{z^2}{2(a+b)}\big)\int \exp\big(-\frac{a+b}{2ab}\big(x-\frac{a}{a+b}z\big)^2\big)dx \\ &= \exp\big(-\frac{z^2}{2(a+b)}\big)\sqrt{2\pi ab/(a+b)}\end{aligned}$$

since the last integral is the normal density with parameters $\mu = az/(a+b)$ and $\sigma^2 = ab/(a+b)$ without its proper normalizing constant. Reintroducing the constant we dropped at the beginning,

$$f_{Y_1+Y_2}(z) = \frac{1}{2\pi\sqrt{ab}}\sqrt{2\pi ab/(a+b)}\exp\big(-\frac{z^2}{2(a+b)}\big)$$

□

The last question that we have to address before we can study independent random variables is: Do they exist? (If they do not exist, then there is no point in studying them!) If we are given a finite number of distribution functions F_i, $1 \le i \le n$, it is easy to construct independent random variables $X_1, ..., X_n$ with $P(X_i \le x) = F_i(x)$. Let $\Omega = \mathbb{R}^n$, $\mathcal{F} = \mathcal{R}^n$, $X_i(\omega_1, ..., \omega_n) = \omega_i$ (the ith coordinate of $\omega \in \mathbb{R}^n$), and let P be the measure on $\mathcal{R}^n$ that has

$$P((a_1, b_1] \times \cdots \times (a_n, b_n]) = (F_1(b_1) - F_1(a_1)) \ldots (F_n(b_n) - F_n(a_n))$$

If μ_i is the measure with distribution function F_i, then $P = \mu_1 \times \cdots \times \mu_n$.

To construct an infinite sequence $X_1, X_2, ...$ of independent random variables with given distribution functions, we want to perform the last construction on the infinite product space

$$\mathbb{R}^{\mathbf{N}} = \{(\omega_1, \omega_2, ...) : \omega_i \in \mathbb{R}\} = \{\text{functions } \omega : \mathbb{N} \to \mathbb{R}\}$$

where $\mathbb{N} = \{1, 2, ...\}$ and $\mathbb{N}$ stands for natural numbers. We define $X_i(\omega) = \omega_i$ and we equip $\mathbb{R}^{\mathbf{N}}$ with the product σ-field $\mathcal{R}^{\mathbf{N}}$, which is generated by the finite dimensional sets = sets of the form $(\omega : \omega_i \in B_i, 1 \leq i \leq n\}$ where $B_i \in \mathcal{R}$. It is clear how we want to define P for finite dimensional sets.

Theorem 2.1.19. ***Kolmogorov's extension theorem****. Suppose we are given probability measures μ_n on $(\mathbb{R}^n, \mathcal{R}_n)$ that are consistent, that is,*

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

Then there is a unique probability measure P on $(\mathbb{R}^{\boldsymbol{N}}, \mathcal{R}^{\boldsymbol{N}})$, with

$$P(\omega : \omega_i \in (a_i, b_i], 1 \leq i \leq n) = \mu_n((a_1, b_1] \times \cdots \times (a_n, b_n])$$

In what follows we need to construct sequences of random variables that take values in other measurable spaces $(S, \mathcal{S})$. Unfortunately, Theorem 2.1.19 is not valid for arbitrary measurable spaces.

Theorem 2.1.20. *If S is a Borel subset of a complete separable metric space M, and S is the collection of Borel subsets of S, then $(S, \mathcal{S})$ is nice.*

Proof. We begin with the special case $S = [0, 1)^{\mathbf{N}}$ with metric

$$\rho(x, y) = \sum_{n=1}^{\infty} |x_n - y_n|/2^n$$

If $x = (x^1, x^2, x^3, ...)$, expand each component in binary $x^j = x_1^j x_2^j x_3^j \dots$ (taking the expansion with an infinite number of 0's). Let

$$\varphi_0(x) = x_1^1 x_2^1 x_1^2 x_3^1 x_2^2 x_1^3 x_4^1 x_3^2 x_2^3 x_1^4 \dots$$

To treat the general case, we observe that by letting

$$d(x, y) = \rho(x, y)/(1 + \rho(x, y))$$

We can suppose that the metric has $d(x, y) < 1$ for all x, y. Let $q_1, q_2, ...$ be a countable dense set in S. Let

$$\psi(x) = (d(x, q_1), d(x, q_2), ...).$$

$\psi : S \to [0, 1)^N$ is continuous and 1-1. $\varphi_0 \circ \psi$ gives the desired mapping.

□

2.2 Weak Laws of Large Numbers

In this section, we will prove several "weak laws of large numbers." The first order of business is to define the mode of convergence that appears in the conclusions of the theorem. We say that Y_n converges to Y in probability if for all $\epsilon > 0$, $P(|Y_n - Y| > \epsilon) \to 0$ as $n \to \infty$.

Our first set of weak laws come from computing variances and using Chebyshev's inequality. A family of random variables X_i, $i \in I$ with $EX_i^2 < \infty$ is said to be uncorrelated if we have

$$E(X_iX_j) = EX_iEX_j \text{ whenever } i \neq j$$

They key to our weak law for uncorrelated random variables is

Theorem 2.2.1. *Let $X_1, ..., X_n$ have $E(X_i^2) < \infty$ and be uncorrelated. Then*

$$var(X_1 + \cdots + X_n) = var(X_1) + \cdots + var(X_n)$$

where var(Y) = the variance of Y.

Proof. Let $\mu_i = EX_i$ and $S_n = \sum_{i=1}^n X_i$. Since $ES_n = \sum_{i=1}^n \mu_i$, using the definition of the variance, writing the square of the sum as the product of two copies of the sum, and then expanding, we have

$$\begin{aligned}
\text{var}(S_n) &= E(S_n - ES_n)^2 \\
&= E\Big(\sum_{i=1}^n (X_i - \mu_i)\Big)^2 \\
&= E\Big(\sum_{i=1}^n \sum_{j=1}^n (X_i - \mu_i)(X_j - \mu_j)\Big) \\
&= \sum_{i=1}^n E(X_i - \mu_i)^2 + 2\sum_{i=1}^n \sum_{j=1}^{i-1} E((X_i - \mu_i)(X_j - \mu_j))
\end{aligned}$$

where in the last equality we have separated out the diagonal terms $i = j$ and used the fact that the sum over $1 \leq i < j \leq n$ is the same as the sum over $1 \leq j < i \leq n$.

The first sum is $\text{var}(X_1) + \cdots + \text{var}(X_n)$ so we want to show that the second sum is zero. To do this, we observe

$$\begin{aligned}
E((X_i - \mu_i)(X_j - \mu_j)) &= EX_iX_j - \mu_iEX_j - \mu_jEX_i + \mu_i\mu_j \\
&= EX_iX_j - \mu_i\mu_j = 0
\end{aligned}$$

since X_i and X_j are uncorrelated.

□

In words, Theorem 2.2.1 says that for uncorrelated random variables the variance of the sum is the sum of the variances. The second ingredient is the following consequence:

$$\text{var}(cY) = c^2\text{var}(Y)$$

The third and final ingredient is

Lemma 2.2.2. *If $p > 0$ and $E|Z_n|^P \to 0$, then $Z_n \to 0$ in probability.*

Proof. Chebyshev's inequality, Theorem 1.6.6 with $\varphi(x) = x^p$ and $X = |Z_n|$ implies that if $\epsilon > 0$ then $P(|Z_n| \geq \epsilon) \leq \epsilon^{-p}E|Z_n|^p \to 0$.

□

Theorem 2.2.3. *L^2 **weak law**. Let $X_1, X_2, ...$ be uncorrelated random variables with $EX_i = \mu$ and $var(X_i) \leq C < \infty$. If $S_n = X_1 + \ldots X_n$ then as $n \to \infty$, $S_n/n \to \mu$ in L^2 and in probability.*

Proof. To prove L^2 convergence, observe that $E(S_n/n) = \mu$, so

$$E(S_n/n - \mu)^2 = \text{var}(S_n/n) = \frac{1}{n^2}(\text{var}(X_1) + \cdots + \text{var}(X_n)) \leq \frac{C_n}{n^2} \to 0$$

To conclude there is also convergence in probability, we apply the Lemma 2.2.2 to $Z_n = S_n/n - \mu$.

□

The most important special case of Theorem 2.2.3 occurs when $X_1, X_2, ...$ are independent random variables that all have the same distribution. In the jargon, they are independent and identically distributed or i.i.d. for short. Theorem 2.2.3 tells us in this case if $EX_i^2 < \infty$ then S_n/n converges to $\mu = EX_i$ in probability as $n \to \infty$. Below, we will see that $E|X_i| < \infty$ is sufficient for the last conclusion, but for the moment we will concern ourselves with consequences of the weaker result.

Example 2.2.4. Polynomial approximation. Let f be a continuous function on $[0, 1]$, and let

$$f_n(x) = \sum_{m=0}^{n} \binom{n}{m} x^m (1-x)^{n-m} f(m/n) \text{ where } \binom{n}{m} = \frac{n!}{m!(n-m)!}$$

be the **Bernstein polynomial of degree** n associated with f. Then as $n \to \infty$

$$\sup_{x \in [0,1]} |f_n(x) - f(x)| \to 0$$

Proof. First observe that if S_n is the sum of n independent random variables with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$ then $EX_i = p$, $\text{var}(X_i) = p(1-p)$ and

$$P(S_n = m) = \binom{n}{m} p^m (1-p)^{n-m}$$

so $Ef(S_n/n) = f_n(p)$. Theorem 2.2.3 tells us that as $n \to \infty$, $S_n/n \to p$ in probability. The last two observations motivate the definition of $f_n(p)$, but to prove the desired conclusion we have to use the proof of Theorem 2.2.3 rather than the result itself.

Combining the proof of Theorem 2.2.3 with our formula for the variance of X_i and the fact that $p(1-p) \leq 1/4$ when $p \in [0, 1]$, we have

$$P(|S_n/n - p| > \delta) \leq \frac{\text{var}(S_n/n)}{\delta^2} = \frac{p(1-p)}{n\delta^2} \leq \frac{1}{4n\delta^2}$$

To conclude now that $Ef(S_n/n) \to f(p)$, let $M = \sup_{x \in [0,1]} |f(x)|$, let $\epsilon > 0$, and pick $\delta > 0$ so that if $|x - y| < \delta$ then $|f(x) - f(y)| < \epsilon$. (This is possible since a continuous function is uniformly continuous on each bounded interval.) Now using Jensen's inequality, Theorem 1.6.2, gives

$$|Ef(S_n/n) - f(p)| \leq E|f(S_n/n) - f(p)| \leq \epsilon + 2MP(|S_n/n - p| > \delta)$$

Letting $n \to \infty$, we have $\limsup_{n\to\infty} |Ef(S_n/n) - f(p)| \leq \epsilon$, but ϵ is arbitrary so this gives the desired result.

□

Example 2.2.5. A high-dimensional cube is almost the boundary of a ball. Let $X_1, X_2, ...$ be independent and uniformly distributed on $(-1, 1)$. Let $Y_i = X_i^2$, which are independent since they are functions of independent random variables. $EY_i = 1/3$ and $\text{var}(Y_i) \le EY_i^2 \le 1$, so Theorem 2.2.3 implies

$$(X_1^2 + \cdots + X_n^2)/n \to 1/3 \text{ in probability as } n \to \infty$$

Let $A_{n,\epsilon} = \{x \in \mathbb{R}^n : (1-\epsilon)\sqrt{n/3} < |x| < (1+\epsilon)\sqrt{n/3}\}$ where $|x| = (x_1^2 + \cdots + x_n^2)^{1/2}$. If we let $|S|$ denote the Lebesgue measure of S then the last conclusion implies that for any $\epsilon > 0$, $|A_{n,\epsilon} \cap (-1,1)^n|/2^n \to 1$, or, in other words, most of the volume of the cube $(-1,1)^n$ comes from $A_{n,\epsilon}$, which is almost the boundary of the ball of radius $\sqrt{n/3}$.

Many classical limit theorems in probability concern arrays $X_{n,k}$, $1 \le k \le n$ of random variables and investigate the limiting behavior of their row sums $S_n = X_{n,1} + \cdots + X_{n,n}$. In most cases, we assume that the random variables on each row are independent, but for the next trivial (but useful) result we do not need that assumption. Indeed, here S_n can be any sequence of random variables.

Theorem 2.2.6. *Let $\mu_n = ES_n$, $\sigma_n^2 = var(S_n)$. If $\sigma_n^2/b_n^2 \to 0$, then*

$$\frac{S_n - \mu_n}{b_n} \to 0 \text{ in probability}$$

Proof. Our assumptions imply $E((S_n - \mu_n)/b_n)^2 = b_n^{-2}\text{var}(S_n) \to 0$, so the desired conclusion follows from Lemma 2.2.2.

□

We now give three applications of Theorem 2.2.6. For these three examples, the following calculation is useful:

$$\sum_{m=1}^{n} \frac{1}{m} \ge \int_1^n \frac{dx}{x} \ge \sum_{m=2}^{n} \frac{1}{m}$$

$$\log n \le \sum_{m=1}^{n} \frac{1}{m} \le 1 + \log n \tag{2.2}$$

Example 2.2.7. Coupon collector's problem. Let $X_1, X_2, ...$ be i.i.d. uniform on $\{1, 2, ..., n\}$. To motivate the name, think of collecting baseball cards (or coupons). Suppose that the ith item we collect is chosen at random from the set of possibilities and is independent of the previous choices. Let $\tau_k^n = \inf\{m : |\{X_1, ..., X_m\}| = k\}$ be the first time we have k different items. In this problem, we are interested in the asymptotic behavior of $T_n = \tau_n^n$, the time to collect a complete set. It is easy to see that $\tau_1^n = 1$. To make later formulas work out nicely, we will set $\tau_0^n = 0$. For $1 \le k \le n$, $X_{n,k} \equiv \tau_k^n - \tau_{k-1}^n$ represents the time to get a choice different from our first $k-1$, so $X_{n,k}$ has a geometric distribution with parameter $1 - (k-1)/n$ and is independent of the earlier waiting times $X_{n,j}$, $1 \le j < k$. If X has a geometric distribution with parameter p, then $EX = 1/p$ and $\text{var}(X) \le 1/p^2$. Using the linearity of expected value, bounds on $\sum_{m=1}^n 1/m$ in 2.2, and Theorem 2.2.1 we see that

$$ET_n = \sum_{k=1}^{n} \big(1 - \frac{k-1}{n}\big)^{-1} = n \sum_{m=1}^{n} m^{-1} \sim n \log n$$

$$\text{var}(T_n) \le \sum_{k=1}^{n} \big(1 - \frac{k-1}{n}\big)^{-2} = n^2 \sum_{m=1}^{n} m^{-2} \le n^2 \sum_{m=1}^{\infty} m^{-2}$$

Taking $b_n = n \log n$ and using Theorem 2.2.6, it follows that

$$\frac{T_n - n\sum_{m=1}^{n} m^{-1}}{n \log n} \to 0 \text{ in probability}$$

and hence $T_n/(n \log n) \to 1$ in probability.

For a concrete example, take $n = 365$, i.e., we are interested in the number of people we need to meet until we have seen someone with every birthday. In this case the limit theorem says it will take about $365 \log 365 = 2153.46$ tries to get a complete set. Note that the number of trials is 5.89 times the number of birthdays.

Example 2.2.8. Random permutations. Let Ω_n consist of the $n!$ permutations (i.e., one-to-one mappings from $\{1, ..., n\}$ onto $\{1, ..., n\}$) and make this into a probability space by assuming all the permutations are equally likely. This application of the weak law concerns the cycle structure of a random permutation π, so we begin by describing the decomposition of a permutation into cycles. Consider the sequence $1, \pi(1), \pi(\pi(1)), ...$ Eventually, $\pi^k(1) = 1$. When it does, we say the first cycle is completed and has length k. To start the second cycle, we pick the smallest integer i not in the first cycle and look at i, $\pi(i)$, $\pi(\pi(i))$, ... until we come back to i. We repeat the construction until all the elements are accounted for. For example, if the permutation is

i	1	2	3	4	5	6	7	8	9
$\pi(i)$	3	9	6	8	2	1	5	4	7

then the cycle decomposition is (136)(2975)(48).

Let $X_{n,k} = 1$ if a right parenthesis occurs after the kth number in the decomposition, $X_{n,k} = 0$ otherwise and let $S_n = X_{n,1} + \cdots + X_{n,n} =$ the number of cycles. (In the example, $X_{9,3} = X_{9,7} = X_{9,9} = 1$, and the other $X_{9,m} = 0$.) We claim that

Lemma 2.2.9. *$X_{n,1}, ..., X_{n,n}$ are independent and $P(X_{n,j} = 1) = \frac{1}{n-j+1}$.*

Intuitively, this is true since, independent of what has happened so far, there are $n-j+1$ values that have not appeared in the range, and only one of them will complete the cycle.

Proof. To prove this, it is useful to generate the permutation in a special way. Let $i_1 = 1$. Pick j_1 at random from $\{1, ..., n\}$ and let $\pi(i_1) = j_1$. If $j_1 \neq 1$, let $i_2 = j_1$. If $j_1 = 1$, let $i_2 = 2$. In either case, pick j_2 at random from $\{1, ..., n\} - \{j_1\}$. In general, if $i_1, j_1, ..., k_{k-1}, j_{k-1}$ have been selected and we have set $\pi(i_l) = j_l$ for $1 \leq l < k$, then (a) if $j_{k-1} \in \{i_1, ..., i_{k-1}\}$ so a cycle has just been completed, we let $i_k = \inf(\{1, ..., n\} - \{i_1, ..., i_{k-1}\})$ and (b) if $j_{k-1} \notin \{i_1, ..., j_{k-1}\}$ we let $i_k = j_{k-1}$. In either case we pick j_k at random from $\{1, ..., n\} - \{j_1, ..., j_{k-1}\}$ and let $\pi(i_k) = j_k$.

The construction above is tedious to write, but we can claim with a clear conscience that $X_{n,1}, ..., X_{n,n}$ are independent and $P(X_{n,k} = 1) = 1/(n - j + 1)$ since when we pick j_k there are $n - j + 1$ values in $\{1, ..., n\} - \{j_1, ..., j_{k-1}\}$ and only one of them will complete the cycle.

To check the conditions of Theorem 2.2.6, now note

$$ES_n = 1/n + 1/(n-1) + \cdots + 1/2 + 1$$

$$\text{var}(S_n) = \sum_{k=1}^{n} \text{var}(X_{n,k}) \leq \sum_{k=1}^{n} E(X_{n,k}^2) = \sum_{k=1}^{n} E(X_{n,k}) = ES_n$$

where the results on the second line follow from Theorem 2.2.1, the fact that $\text{var}(Y) \leq EY^2$, and $X_{n,k}^2 = X_{n,k}$. Now $ES_n \sim \log n$, so if $b_n = (\log n)^{.5+\epsilon}$ with $\epsilon > 0$, the conditions of Theorem 2.2.6 are satisfied and it follows that

$$\frac{S_n - \sum_{m=1}^{n} m^{-1}}{(\log n)^{\cdot 5+\epsilon}} \to 0 \text{ in probability}$$

Taking $\epsilon = 0.5$ we have that $S_n/\log n \to 1$ in probability, but $(\star)$ says more.

Example 2.2.10. An occupancy problem. Suppose we put r balls at random in n boxes, i.e., all n^r assignments of balls to boxes have equal probability. Let A_i be the event that the ith box is empty and $N_n =$ the number of empty boxes. It is easy to see that

$$P(A_i) = (1-1/n)^r \text{ and } EN_n = n(1-1/n)^r$$

A little calculus (take logarithms) shows that if $r/n \to c$, $EN_n/n \to e^{-c}$. To compute the variance of N_n, we observe that

$$EN_n^2 = E\Big(\sum_{m=1}^n 1_{A_m}\Big)^2 = \sum_{1\le k,m\le n} P(A_k \cap A_m)$$

$$\begin{aligned}
\text{var}(N_n) &= EN_n^2 - (EN_n)^2 \\
&= \sum_{1\le k,m\le n} P(A_k\cap A_m) - P(A_k)P(A_m) \\
&= n(n-1)\{(1-2/n)^r - (1-1/n)^{2r}\} + n\{(1-1/n)^r - (1-1/n)^{2r}\}
\end{aligned}$$

The first term comes from $k \neq m$ and the second from $k = m$. Since $(1-2/n)^r \to e^{-2c}$ and $(1-1/n)^r \to e^{-c}$, it follows easily from the last formula that $\text{var}(N_n/n) = \text{var}(N_n)/n^2 \to 0$. Taking $b_n = n$ in Theorem 2.2.6 now we have

$$N_n/n \to e^{-c} \text{ in probability}$$

To truncate a random variable X at level M means to consider

$$\bar{X} = X1_{(|X|\le M)} = \begin{cases} X & \text{if } |X| \le M \\ 0 & \text{if } |X| > M \end{cases}$$

To extend the weak law to random variables without a finite second moment, we will truncate and then use Chebyshev's inequality. We begin with a very general but also very useful result. Its proof is easy because we have assumed what we need for the proof. Later we will have to work a little to verify the assumptions in special cases, but the general result serves to identify the essential ingredients in the proof.

Theorem 2.2.11. ***Weak law for triangular arrays***. *For each n let $X_{n,k}$, $1 \le k \le n$, be independent. Let $b_n > 0$ with $b_n \to \infty$, and let $\bar{X}_{n,k} = X_{n,k}1_{(|X_{n,k}|\le b_n)}$. Suppose that as $n \to \infty$*

(i) $\sum_{k=1}^n P(|X_{n,k}| > b_n) \to 0$, *and*

(ii) $b_n^{-2}\sum_{k=1}^n E\bar{X}_{n,k}^2 \to 0$. *If we let* $eS_n = X_{n,1} + \cdots + X_{n.n}$ *and put* $a_n = \sum_{k=1}^n E\bar{X}_{n,k}$, *then*

$$(S_n - a_n)/b_n \to 0 \text{ in probability}$$

Proof. Let $\bar{S}_n = \bar{X}_{n,1} + \cdots + \bar{X}_{n,n}$. Clearly,

$$P\left(\left|\frac{S_n - a_n}{b_n}\right| > \epsilon\right) \le P(S_n \neq \bar{S}_n) + P\left(\left|\frac{S_n - a_n}{b_n}\right| > \epsilon\right)$$

To estimate the first term, we note that

$$P(S_n \neq \bar{S}_n) \le P(\cup_{k=1}^n \{\bar{X}_{n,k} \neq X_{n,k}\}) \le \sum_{k=1}^n P(|X_{n,k}| > b_n) \to 0$$

by (i). For the second term, we note that Chebyshev's inequality, $a_n = E\bar{S}_n$, Theorem 2.2.1, and $\text{var}(X) \leq EX^2$ imply

$$\begin{aligned} P\left(\left|\frac{\bar{S}_n - a_n}{b_n}\right| > \epsilon\right) &\leq \epsilon^{-2} E\left|\frac{\bar{X}_n - a_n}{b_n}\right|^2 \\ &= \epsilon^{-2} b_n^{-2} \text{var}(\bar{S}_n) \\ &= (b_n\epsilon)^{-2} \sum_{k=1}^{n} \text{var}(\bar{X}_{n,k}) \\ &\leq (b_n\epsilon)^{-2} \sum_{k=1}^{n} E(\bar{X}_{n,k})^2 \to 0 \end{aligned}$$

Theorem 2.2.12. ***Weak law of large numbers.*** *Let $X_1, X_2, ...,$ be i.i.d. with*

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

Let $S_n = X_1 + \cdots + X_n$ and let $\mu_n = E(X_1 1_{(|X_1| \leq n)})$. Then $S_n/n - \mu_n \to 0$ in probability.

Remark 2.2.13. The assumption in the theorem is necessary for the existence of constants a_n so that $S_n/n - a_n \to 0$.

Proof. We will apply Theorem 2.2.11 with $X_{n,k} = X_k$ and $b_n = n$. To check (i), we note

$$\sum_{k=1}^{n} P(|X_{n,k}| > n) = nP(|X_i| > n) \to 0$$

by assumption. To check (ii), we need to show $n^{-2} \cdot nE\bar{X}_{n,1}^2 \to 0$. To do this ,we need the following result, which will be useful several times below.

Lemma 2.2.14. *If $Y \geq 0$ and $p > 0$, then $E(Y^P) = \int_0^\infty py^{p-1}P(Y > y)dy$.*

Proof. Using the definition of expected value, Fubini's theorem (for nonnegative random variables, and then calculating the result integrals gives

$$\begin{aligned} \int_0^\infty py^{p-1}P(Y > y)dy &= \int_0^\infty \int_\Omega py^{p-1} 1_{(Y>y)} dP dy \\ &= \int_\Omega \int_0^\infty py^{p-1} 1_{(Y>y)} dy dP \\ &= \int_\Omega \int_0^\infty py^{p-1} dy dP \\ &= \int_\Omega Y^P dP = EY^P \end{aligned}$$

which is the desired result.

□

Returning to the proof of Theorem 2.2.12, we observe that Lemma 2.2.14 and the fact that $\bar{X}_{n,1} = X_1 1_{(|X_1| \leq n)}$ imply

$$E(\bar{X}_{n,1}^2) = \int_0^\infty 2yP(|\bar{X}_{n,1}| > y)dy \leq \int_0^n 2yP(|X_1| > y)dy$$

since $P(|\bar{X}_{n,1}| > y) = 0$ for $y \geq n$ and $= P(|X_1| > y) - P(|X_1| > n)$ for $y \leq n$. We claim that $yP(|X_1| > y) \to 0$ implies

$$E(\bar{X}_{n,1}^2)/n = \frac{1}{n}\int_0^n 2yP(|X_1| > y)dy \to 0$$

as $n \to \infty$. Intuitively, this holds since the right-hand side is the average of $g(y) = 2yP(|X_1| > y)$ over $[0, n]$ and $g(y) \to 0$ as $y \to \infty$. To spell out the details, note that $0 \le g(y) \le 2y$ and $g(y) \to 0$ as $y \to \infty$, so we must have $M = \sup g(y) < \infty$. If we let $\epsilon_K = \sup\{g(y) : y > K\}$ then by considering the integrals over $[0, K]$ and $[K, n]$ separately

$$\int_0^n 2yP(|X_1| > y)dy \le KM + (n - K)\epsilon_K$$

Dividing by n and letting $n \to \infty$, we have

$$\limsup_{n\to\infty} \frac{1}{n} \int_0^n 2yP(|X_1| > y)dy \le \epsilon_K$$

Since K is arbitrary and $\epsilon_K \to 0$ as $K \to \infty$, the desired result follows.

□

Finally, we have the weak law in its most familiar form.

Theorem 2.2.15. *Let $X_1, X_2, ...$ be i.i.d. with $E|X_i| < \infty$. Let $S_n = X_1 + \cdots + X_n$ and let $\mu = EX_1$. Then $S_n/n \to \mu$ in probability.*

Remark 2.2.16. Applying Lemma 2.2.14 with $p = 1-\epsilon$ and $\epsilon > 0$, we see that $xP(|X_1| > x) \to 0$ implies $E|X_1|^{1-\epsilon} < \infty$, so the assumption in is not much weaker than finite mean.

Proof. Two applications of the dominated convergence theorem imply

$$xP(|X_1| > x) \le E(|X_1|1_{(|X_1|>x)}) \to 0 \text{ as } x \to \infty$$

$$\mu_n = E(X_1 1_{(|X_1|\le n)}) \to E(X_1) = \mu \text{ as } n \to \infty$$

Using Theorem 2.2.12, we see that if $\epsilon > 0$ then $P(|S_n/n - \mu_n| > \epsilon/2) \to 0$. Since $\mu_n \to \mu$, it follows that $P(|S_n/n - \mu| > \epsilon) \to 0$.

Example 2.2.17. For an example where the weak law does not hold, suppose $X_1, X_2, ...$ are independent and have a Cauchy distribution:

$$P(X_i \le x) \int_{-\infty}^{x} \frac{dt}{\pi(1 + t^2)}$$

As $x \to \infty$,

$$P(|X_1| > x) = 2 \int_x^\infty \frac{dt}{\pi(1 + t^2)} \sim \frac{2}{\pi} \int_x^\infty t^{-2}dt = \frac{2}{\pi}x^{-1}$$

From the necessity of the condition above, we can conclude that there is no sequence of constants μ_n so that $S_n/n - \mu_n \to 0$. We will see later that S_n/n always has the same distribution as X_1.

As the next example shows, we can have a weak law in some situations in which $E|X| = \infty$.

Example 2.2.18. The "St. Petersburg paradox". Let $X_1, X_2, ...$ be independent random variables with

$$P(X_i = 2^j) = 2^{-j} \text{ for } j \ge 1$$

In words, you win e^j dollars if it takes j tosses to get a heads. The paradox here is that $EX_1 = \infty$, but you clearly wouldn't pay an infinite amount to play this game. An application of Theorem 2.2.11 will tell us how much we should pay to play the game n times.

In this example, $X_{n,k} = X_k$. To apply 2.2.11, we have to pick b_n. To do this, we are guided by the principle that in checking (ii) we want to take b_n as small as we can and have (i) hold. With this in mind, we observe that if m is an integer

$$P(X_1 \geq 2^m) = \sum_{j=m}^{\infty} 2^{-j} = 2^{-m+1}$$

Let $m(n) = \log_2 n + K(n)$ where $K(n) \to \infty$ and is chosen so that $m(n)$ is an integer (and hence the displayed formula is valid). Letting $b_n = 2^{m(n)}$, we have

$$nP(X_1 \geq b_n) = n3^{-m(n)+1} = 2^{-K(n)+1} \to 0$$

proving (i). To check (ii), we observe that if $\bar{X}_{n,k} = X_k 1_{(|X_k| \leq b_n)}$ then

$$E\bar{X}_{n,k}^2 = \sum_{j=1}^{m(n)} 2^{2j} \cdot 2^{-j} \leq 2^{m(n)} \sum_{k=0}^{\infty} 2^{-k} = 2b_n$$

So the expression in (ii) is smaller than $2n/b_n$, which $\to 0$ since

$$b_n = 2^{m(n)} = n2^{K(n)} \text{ and } K(n) \to \infty$$

The last two steps are to evaluate a_n and to apply Theorem 2.2.11

$$E\bar{X}_{n,k} = \sum_{j=1}^{m(n)} 2^j 2^{-j} = m(n)$$

so $a_n = nm(n)$. We have $m(n) = \log n + K(n)$ (here and until the end of the example all logs are base 2), so if we pick $K(n)/\log n \to 0$ then $a_n/n \log n \to 1$ as $n \to \infty$. Using Theorem 2.2.11 now, we have

$$\frac{S_n - a_n}{n2^{K(n)}} \to 0 \text{ in probability}$$

If we suppose that $K(n) \leq \log\log n$ for large n then the last conclusion holds with the denominator replaced by $n \log n$, and it follows that $S_n/(n \log n) \to 1$ in probability.

Returning to our original question, we see that a fair price for playing n times is $\$ \log_2 n$ per play. When $n = 1024$, this is \$10 per play. Nicolas Bernoulli wrote in 1713, "There ought not to exist any even halfway sensible person who would not sell the right of playing the game for 40 ducates (per play)." If the wager were 1 ducat, one would need $2^{40} \approx 10^{12}$ plays to start to break even.

2.3 Borel-Cantelli Lemmas

In A_n is a sequence of subsets of Ω, we let

$$\limsup A_n = \lim_{m \to \infty} \cup_{n=m}^{\infty} A_n = \{\omega \text{ that are in infinitely many } A_n\}$$

(the limit exists since the sequence is decreasing in m) and let

$$\liminf A_n = \lim_{m \to \infty} \cap_{n=m}^{\infty} A_n = \{\omega \text{ that are in all but finitely many } A_n\}$$

(the limit exists since the sequence is increasing in m). The names lim sup and lim inf can be explained by noting that

$$\limsup_{n \to \infty} 1_{A_n} = A_{(\limsup A_n)} \quad \liminf_{n \to \infty} 1_{A_n} = 1_{(\liminf A_n)}$$

Theorem 2.3.1. ***Borel-Cantelli lemma.*** *If* $\sum_{n=1}^{\infty} P(A_n) < \infty$ *then*

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

Proof. Let $N = \sum_k 1_{A_k}$ be the number of events that occur. Fubini's theorem implies $EN = \sum_k P(A_k) < \infty$, so we must have $N < \infty$ a.s.

□

Theorem 2.3.2. *$X_n \to X$ in probability if and only if for every subsequence $X_{n(m)}$ there is a further subsequence $X_{n(m_k)}$ that converges almost surely to X.*

Proof. Let ϵ_k be a sequence of positive numbers that $\downarrow 0$. For each k, there is an $n(m_k) > n(m_{k-1})$ so that $P(|X_{n(m_k)} - X| > \epsilon_k) \le 2^{-k}$. Since

$$\sum_{k=1}^{\infty} P(|X_{n(m_k)} - X| > \epsilon) < \infty$$

the Borel-Cantelli lemma implies $P(|X_{n(m_k)} - X| > \epsilon_k \text{ i.o.}) = 0$ i.e., $X_{n(m_k)} \to X$ a.s. To prove the second conclusion, we note that if for every subsequence $X_{n(m)}$ there is a further subsequence $X_{n(m_k)}$ that converges almost surely to X then we can apply the next lemma to the sequence of numbers $y_n = P(|X_n - X| > \delta)$ for any $\delta > 0$ to get the desired result.

□

Theorem 2.3.3. *Let y_n be a sequence of elements of a topological space. If every subsequence $y_{n(m)}$ has a further subsequence $y_{n(m_k)}$ that converges to y then $y_n \to y$.*

Proof. If $y_n \not\to y$ then there is an open set G containing y and a subsequence $y_{n(m)}$ with $y_{n(m)} \notin G$ for all m, but clearly no subsequence of $y_{n(m)}$ converges to y.

□

Remark 2.3.4. Since there is a sequence of random variables that converges in probability but not a.s., it follows from Theorem 2.3.3 that a.s. convergence does not come from a metric, or even from a topology.

Theorem 2.3.2 allows us to upgrade convergence in probability to convergence almost surely. An example of the usefulness of this is

Theorem 2.3.5. *If f is continuous and $X_n \to X$ in probability then $f(X_n) \to f(X)$ in probability. If, in addition, f is bounded then $Ef(X_n) \to Ef(X)$.*

Proof. If $X_{n(m)}$ is a subsequence then Theorem 2.3.2 implies there is a further subsequence $X_{n(m_k)} \to X$ almost surely. Since f is continuous, we know that $f(X_{n(m_k)}) \to f(X)$ almost surely and 2.3.2 implies $f(X_n) \to f(X)$ in probability. If f is bounded then the bounded convergence theorem implies $Ef(X_{n(m_k)}) \to Ef(X)$, and applying Theorem 2.3.3 to $y_n = Ef(X_n)$ gives the desired result.

□

The converse of the Borel-Cantelli lemma is trivially false.

Example 2.3.6. Let $\Omega = (0, 1)$, $\mathcal{F} =$ Borel sets, $P =$ Lebesgue measure. If $A_n = (0, a_n)$ where $a_n \to 0$ as $n \to \infty$ then $\limsup A_n = \emptyset$, but if $a_n \ge 1/n$, we have $\sum a_n = \infty$.

The example just given suggests that for general sets we cannot say much more than the next result.

Theorem 2.3.7. ***The second Borel-Cantelli lemma.*** *If the events A_n are independent then $\sum P(A_n) = \infty$ implies $P(A_n \text{ i.o.}) = 1$.*

Proof. Let $M < N < \infty$. Independence and $1 - x \leq e^{-x}$ imply

$$\begin{aligned} P(\cap_{n=M}^{N} A_n^c) &= \prod_{n=M}^{N} (1 - P(A_n)) \\ &\leq \prod_{n=M}^{N} \exp(-P(A_n)) \\ &= \exp\Big(- \sum_{n=M}^{N} P(A_n) \Big) \to 0 \text{ as } N \to \infty \end{aligned}$$

So $P(\cup_{n=M}^{\infty} A_n) = 1$ for all M, and since $\cup_{n=M}^{\infty} A_n \downarrow \limsup A_n$ it follows that $P(\limsup A_n) = 1$.

□

A typical application of the second Borel-Cantelli lemma is:

Theorem 2.3.8. *If $X_1, X_2, ...$ are i.i.d. with $E|X_i| = \infty$, then $P(|X_n| \geq n \text{ i.o.}) = 1$. So if $S_n = X_1 + \cdots + X_n$ then $P(\lim S_n/n \text{ exists } \in (-\infty, \infty)) = 0$.*

Proof. From Lemma 2.2.14, we get

$$E|X_1| = \int_0^{\infty} P(|X_1| > x)dx \leq \sum_{n=0}^{\infty} P(|X_1| > n)$$

Since $E|X_1| = \infty$ and $X_1, X_2, ...$ are i.i.d., it follows from the second Borel-Cantelli lemma that $P(|X_n| \geq n \text{ i.o.}) = 1$. To prove the second claim, observe that

$$\frac{S_n}{n} - \frac{S_{n+1}}{n+1} = \frac{S_n}{n(n+1)} - \frac{X_{n+1}}{n+1}$$

and on $C \equiv \{\omega : \lim_{n\to\infty} S_n/n \text{ exists } \in (-\infty, \infty)\}$, $S_n/(n(n+1)) \to 0$. So, on $C \cap \{\omega : |X_n| \geq n \text{ i.o.}\}$, we have

$$\left| \frac{S_n}{n} - \frac{S_{n+1}}{n+1} \right| > 2/3 \text{ i.o.}$$

contradicting the fact that $\omega \in C$. From the last observation, we conclude that

$$\{\omega : |X_n| \geq n \text{ i.o. } \} \cap C = \emptyset$$

and since $P(|X_n| \geq n \text{ i.o.}) = 1$; it follows that $P(C) = 0$.

□

Theorem 2.3.8 shows that $E|X_i| < \infty$ is necessary for the strong law of large numbers. The next result extends the second Borel-Cantelli lemma and sharpens its conclusion.

Theorem 2.3.9. *If $A_1, A_2, ...$ are pairwise independent and $\sum_{n=1}^{\infty} P(A_n) = \infty$ then as $n \to \infty$*

$$\sum_{m=1}^{n} 1_{A_m} \Big/ \sum_{m=1}^{n} P(A_m) \to 1 \text{ a.s.}$$

Proof. Let $X_m = 1_{A_m}$ and let $S_n = X_1 + \cdots + X_n$. Since the A_m pairwise independent, the X_m are uncorrelated and hence Theorem 2.2.1 implies

$$\text{var}(S_n) = \text{var}(X_1) + \cdots + \text{var}(X_n)$$

while $\text{var}(X_m) \leq E(X_m^2) = E(X_m)$, since $X_m \in \{0, 1\}$, so $\text{var}(S_n) \leq E(S_n)$. Chebyshev's inequality implies

$$(\star) P(|S_n - ES_n| > \delta ES_n) \leq \text{var}/(\delta ES_n)^2 \leq 1/(\delta^2 ES_n) \to 0$$

as $n \to \infty$. (Since we have assumed $ES_n \to \infty$.)

The last computation shows that $S_n/ES_n \to 1$ in probability. To get almost sure convergence, we have to take subsequences. Let $n_k = \inf\{n : ES_n \geq k^2\}$. Let $T_k = S_{n_k}$ and note that the definition and $EX_m \leq 1$ imply $k^2 \leq ET_k \leq k^2 + 1$. Replacing n by n_k in $(\star)$ and using $ET_k \geq k^2$ shows

$$P(|T_k ET_k| > \delta ET_k) \leq 1/(\delta^2 k^2)$$

So $\sum_{k=1}^{\infty} P(|T_k - ET_k| > \delta ET_k) < \infty$, and the Borel-Cantelli lemma implies $P(|T_k - ET_k| > \delta ET_k \text{i.o.}) = 0$. Since δ is arbitrary, it follows that $T_k/ET_k \to 1$ a.s. To show $S_n/ES_n \to 1$ a.s., pick an ω so that ${}_k(\omega)/ET_k \to 1$ and observe that if $n_k \leq n < n_{k+1}$ then

$$\frac{T_k(\omega)}{ET_{k+1}} \leq \frac{S_n(\omega)}{ES_n} \leq \frac{T_{k+1}(\omega)}{ET_k}$$

To show that the terms at the left and right ends $\to 1$, we rewrite the last inequalities as

$$\frac{ET_k}{ET_{k+1}} \cdot \frac{T_k(\omega)}{ET_k} \leq \frac{S_n(\omega)}{ES_n)} \leq \frac{T_{k+1}(\omega)}{ET_{k+1}} \cdot \frac{ET_{k+1}}{ET_k}$$

From this, we see it is enough to show that $ET_{k+1}/ET_k \to 1$, but this follows from

$$k^2 \leq ET_k \leq ET_{k+1} \leq (k+1)^2 + 1$$

and the fact that $\{(k+1)^2 + 1\}/k^2 = 1 + 2/k + 2/k^2 \to 1$.

□

Example 2.3.10. Record values. Let $X_1, X_2, ...$ be a sequence of random variables and think of X_k as the distance for an individual's kth high jump or shot-put toss so that $A_k = \{X_k > \sup_{j<k} X_j\}$ is the event that better with more experience or that injuries may occur, we will suppose that $X_1, X_2, ...$ are i.i.d. with a distribution $F(x)$ that is continuous. Even though it may seem that the occurrence of a record at time k will make it less likely that one will occur at time $k + 1$, we claim

$$\text{The } A_k \text{ are independent with } P(A_k) = 1/k.$$

To prove this, we start by observing that since F is continuous $P(X_j = X_k) = 0$ for any $j \neq k$, so we can let $Y_1^n > Y_2^n > \cdots > Y_n^n$ be the random variables $X_1, ..., X_n$ put into decreasing order and define a random permutation of $\{1, ..., n\}$ by $\pi_n(i) = j$ if $X_i = Y_j^n$, i.e., if the ith random variable has rank j. Since the distribution of $(X_1, ..., X_n)$ is not affected by changing the order of the random variables, it is easy to see:

(a) The permutation π_n is uniformly distributed over the set of $n!$ possibilities.

Proof of (a). This is "obvious" by symmetry, but if one wants to hear more, we can argue as follows. Let π_n be the permutation induced by $(X_1, ..., X_n)$, and let σ_n be

a randomly chosen permutation of $\{1, ..., n\}$ independent of the X sequence. Then we can say two things about the permutation induced by $X_{\sigma(1)}, ..., X_{\sigma(n)})$: (i) it is $\pi_n \circ \sigma_n$, and (ii) it has the same distribution as π_n. The desired result follows now by noting that if π is any permutation, $\pi \circ \sigma_n$, is uniform over the $n!$ possibilities.

□

Once you believe (a), the rest is easy:
(b) $P(A_n) = P(\pi_n(n) = 1) = 1/n$.
(c) If $m < n$ and $i_{m+1}, ..., i_n$ are distinct elements of $\{1, ..., n\}$ then

$$P(A_m | \pi_n(j) = i_j \text{ for } m + 1 \le j \le n) = 1/m$$

Intuitively, this is true since if we condition on the ranks of $X_{m+1}, ..., X_n$ then this determines the set of ranks available for $X_1, ..., X_m$, but all possible orderings of the ranks are equally likely and hence there is probability $1/m$ that the smallest rank will end up at m.

Proof of (c). If we let σ_m be a randomly chosen permutation of $\{1, ..., m\}$ then (i) $\pi_n \circ \sigma_m$ ahs the same distribution as π_n, and (ii) since the application of σ_m randomly rearranges $\pi_n(1), ..., \pi_n(m)$ the desired result follows.

□

If we let $m_1 < m_2 < \cdots < m_k$ then it follows from (c) that

$$P(A_{m_1} | A_{m_2} \cap \cdots \cap A_{m_k}) = P(A_{m_1})$$

And the claim follows by induction.

Using Theorem 2.3.9 and by now the familiar fact that $\sum_{m=1}^{n} 1/m \sim \log n$, we have

Theorem 2.3.11. *If $R_n = \sum_{m=1}^{n} 1_{A_m}$ is the number of records at time n then as $n \to \infty$,*

$$R_n / \log n \to 1 \text{ a.s.}$$

Note that the last result is independent of the distribution F (as long as it is continuous).

Remark 2.3.12. Let $X_1, X_2, ???$ be i.i.d. with a distribution that is continuous. Let Y_i be the number of $j \le i$ with $X_j > X_i$. It follows from (a) that Y_i are independent random variables with $P(Y_i = j) = 1/i$ for $0 \le j < i - 1$.

Comic relief. Let $X_0, X_1, ...$ be i.i.d. and imagine they are the offers you get for a car you are going to sell. Let $N = \inf\{n \ge 1 : X_n > X_0\}$. Symmetry implies exactly $1/(n + 1)$. (When the distribution is continuous this probability is exactly $1/(n + 1)$, but our distribution now is general and ties go to the first person who calls.)

$$EN = \sum_{n=0}^{\infty} P(N > n) \ge \sum_{n=0}^{\infty} \frac{1}{n+1} = \infty$$

So the expected time you have to wait until you get an offer better than the first one is ∞.

Example 2.3.13. Head runs. Let X_n, $n \in \mathbb{Z}$, be i.i.d. with $P(X_n = 1) = P(X_n = -1)1/2$. Let $l_n = \max\{m : X_{n-m+1} = \cdots = X_n = 1\}$ be the length of the run of $+1$'s at time n, and let $L_n = \max_{1 \le m \le n} l_m$ be the longest run at time n. We use a two-sided sequence so that for all n, $P(_n = k) = (1/2)^{k+1}$ for $k \ge 0$. Since $l_1 < \infty$, the result we are going to prove

$$L_n / \log_2 n \to 1 \text{ a.s.} \tag{2.3}$$

is also true for a one-sided sequence. To prove 2.3, we begin by observing

$$P(l_n \ge (1 + \epsilon) \log_2 n) \le n^{-(1+\epsilon)}$$

For any $\epsilon > 0$, so it follows from the Borel-Cantelli lemma that $l_n \leq (1+\epsilon)\log_2 n$ for $n \geq N_\epsilon$. Since ϵ is arbitrary, it follows that

$$\limsup_{n\to\infty} L_n / \log_2 n \leq 1 \text{ a.s.}$$

To get a result in the other direction, we break the first n trials into disjoint blocks of length $[(1-\epsilon)\log_2 n]+1$, on which the variables are all 1 with probability

$$2^{-[(1-\epsilon)\log_2 n]???1} \geq n^{-(1-\epsilon)/2},$$

to conclude that if n is large enough so that $[n/\{[(1-\epsilon)\log_2 n]+1\}] \geq n/\log_2 n$

$$P(L_n \leq (1-\epsilon)\log_2 n) \leq (1???n^{-(1-\epsilon)}/2)^{n/(\log_2 n)} \leq \exp(-n^\epsilon/2\log_2 n)$$

Which is summable, so the Borel-Cantelli lemma implies

$$\liminf_{n\to\infty} L_n / \log_2 n \geq 1 \text{ a.s.}$$

2.4 Strong Law of Large Numbers

We are now ready for the next part. This is proved in Etemadi (1981) [5].

Theorem 2.4.1. ***Strong law of large numbers***. *Let $X_1, X_2, ...$ be pairwise independent identically distributed random variables with $E|X_i| < \infty$. Let $EX_i = \mu$ and $S_n = X_1 + \cdots + X_n$. Then $S_n/n \to \mu$ a.s. as $n \to \infty$.*

Proof. As in the proof of the weak law of large numbers, we begin by truncating.

Lemma 2.4.2. *Let $Y_k = X_k 1_{(|X_k| \leq k)}$ and $T_n = Y_1 + \cdots + Y_n$. It is sufficient to prove that $T_n/n \to \mu$ a.s.*

Proof. $\sum_{k=1}^\infty P(|X_k| > k) \leq \int_0^\infty P(|X_1| > t)dt = E|X_1| \leq \infty$ so $P(X_k \neq Y_k)$ i.o.) $= 0$. This shows that $|S_n(\omega) - T_n(\omega)| \leq R(\omega)$ a.s. for all n, from which the desired result follows.

□

The second step is not so intuitive.

Lemma 2.4.3. $\sum_{k=1}^\infty var(Y_k)/k^2 \leq 4E|X_1| < \infty$.

Proof. To bound the sum, we observe

$$\text{var}(Y_k) \leq E(Y_k^2) = \int_0^\infty 2yP(|Y_k| > y)dy \leq \int_0^k 2yP(|X_1| > y)dy$$

So using Fubini's theorem (since everything is ≥ 0 and the sum is just an integral with respect to counting measure on $\{1, 2, ...\}$)

$$\begin{aligned}\sum_{k=1}^\infty E(Y_k^2)/k^2 &\leq \sum_{k=1}^\infty k^{-2} \int_0^\infty 1_{(y<k)} 2yP(|X_1| > y)dy \\ &= \int_0^\infty \left\{ \sum_{k=1}^\infty k^{-2} 1_{(y\leq k)} \right\} 2yP(|X_1| > y)dy\end{aligned}$$

Since $E|X_1| = \int_0^\infty P(|X_1| > y)dy$, we can complete the proof by showing

Lemma 2.4.4. *If $y \geq 0$ then $2y\sum_{k>y} k^{-2} \leq 4$.*

Proof. We begin with the observation that if $m \geq 2$ then

$$\sum_{k \geq m} k^{-2} \leq \int_{m-1}^{\infty} x^{-2} dx = (m-1)^{-1}$$

When $y \geq 1$, the sum starts with $k = [y] + 1 \geq 2$, so

$$2y \sum_{k > y} k^{-2} \leq 2y/[y] \leq 4$$

since $y/[y] \leq 2$ for $y \geq 1$ (the worst case being y close to 2). To cover $0 \leq y < 1$, we note that in this case

$$2y \sum_{k > y} k^{-2} \leq 2\left(1 + \sum_{k=1}^{\infty} k^{-2}\right) \leq 4$$

This establishes Lemma 2.4.4 which completes the proof of Lemma 2.4.3 and of the theorem.

□

The first two steps, Lemmas 2.4.2 and 2.4.3 above, are standard. Etemadi's inspiration was that since X_n^+, $n \geq 1$, and X_n^-, $n \geq 1$, satisfy the assumptions of the theorem and $X_n = X_n^+ - X_n^-$, we can without loss of generality suppose $X_n \geq 0$. As in the proof of Theorem 2.3.9, we will prove the result first for a subsequence and then use monotonicity to control the values in between. This time, however, we let $\alpha > 1$ and $k(n) = [\alpha]^n]$. Chebyshev's inequality implies that if $\epsilon > 0$

$$\begin{aligned}
\sum_{n=1}^{\infty} P(|T_{k(n)}???ET_{k(n)}| > \epsilon k(n)) &\leq \epsilon^{-2} \sum_{n=1}^{\infty} \mathrm{var}(T_{k(n)})/k(n)^2 \\
&= \epsilon^{-2} \sum_{n=1}^{\infty} k(n)^{-2} \sum_{m=1}^{k(n)} \mathrm{var}(Y_m) \\
&= \epsilon^{-2} \sum_{m=1}^{k(n)} \mathrm{var}(Y_m) \\
&= \epsilon^{-2} \sum_{m=1}^{\infty} \mathrm{var}(Y_m) \sum_{n:k(n) \geq m} k(n)^{-2}
\end{aligned}$$

where we have used Fubini's theorem to interchange the two summations of nonnegative terms. Now $k(n) = [\alpha^n]$ and $[\alpha^n] \geq \alpha^n/2$ for $n \geq 1$, so summing the geometric series and noting that the first term is $\leq m^{-2}$:

$$\sum_{n:\alpha^n \geq m} [\alpha^n]^{-2} \leq 4 \sum_{n:\alpha^n \geq m} \alpha^{-2n} \leq 4(1-\alpha^{-2})^{-1} m^{-2}$$

Combining the computations shows

$$\sum_{n=1}^{\infty} P(|T_{k(n)}???ET_{k(n)}| > \epsilon k(n)) \leq 4(1-\alpha^{-2})^{-2} \sum_{m=1}^{\infty} E(Y_m^2) m^{-2} < \infty$$

by Lemma 2.4.4. Since ϵ is arbitrary $(T_{k(n)}???ET_{k(n)})/k(n) \to 0$ a.s. The dominated convergence theorem implies $EY_k \to EX_1$ as $k \to \infty$, so $ET_{k(n)}/k(n) \to EX_1$ and we have shown $T_{k(n)}/k(n) \to EX_1$ a.s. To handle the intermediate values, we observe that if $k(n) \leq m < k(n+1)$

$$\frac{T_{k(n)}}{k(n+1)} \leq \frac{T_m}{m} \leq \frac{T_{k(n+1)}}{k(n)}$$

(here we use $Y_i \geq 0$), so recalling $k(n) = [\alpha^n]$, we have $k(n+1)/k(n) \to \alpha$ and

$$\frac{1}{\alpha} EX_1 \leq \liminf_{n\to\infty} T_m/m \leq \limsup_{m\to\infty} T_m/m \leq \alpha EX_1$$

Since $\alpha > 1$ is arbitrary, the proof is complete.

□

The next result shows that the strong law holds whenever EX_i exists.

Theorem 2.4.5. *Let $X_1, X_2, \ldots$ be i.i.d. with $EX_i^+ = \infty$ and $EX_i^- < \infty$. If $S_n = X_1 + \cdots + X_n$ then $S_n/n \to \infty$ a.s.*

Proof. Let $M > 0$ and $X_i^M = X_i \wedge M$. The X_i^M are i.i.d. with $E|X_i^M| < \infty$, so if $S_n^M = X_1^M + \cdots + X_n^M$ then Theorem 2.3.3 implies $S_n^M/n \to EX_i^M$. Since $X_i \geq X_i^M$, it follows that

$$\liminf_{n\to\infty} S_n/n \geq \lim_{n\to\infty} S_n^M/n = EX_i^M$$

The monotone convergence theorem implies $E(X_i^M)^+ \uparrow EX_i^+ = \infty$ as $M \uparrow \infty$, so $EX_i^M = E(X_i^M)^+ - E(X_i^M) \uparrow \infty$, and we have $\liminf_{n\to\infty} S_n/n \geq \infty$, which implies the desired result.

□

The rest of this section is devoted to applications of the strong law of large numbers.

Example 2.4.6. Renewal theory. Let $X_1, X_2, ???$ be i.i.d. with $0 < X_i < \infty$. Let $T_n = X_1 + \cdots + X_n$ and think of T_n as the time of nth occurrence of some event. For a concrete situation, consider a diligent janitor who replaces a light bulb the instant it burns out. Suppose the first bulb is put in at time 0 and let X_i be the lifetime of ith bulb. In this interpretation, T_n is the time the nth light bulb burns out and $N_t = \sup\{n : T_n \leq t\}$ is the number of light bulbs that have burnt out by time t.

Theorem 2.4.7. *If $EX_1 = \mu \leq \infty$ then as $t \to \infty$*

$$N_t/t \to 1/\mu \text{ a.s. } (1/\infty = 0)$$

Proof. By Theorems 2.4.1 and 2.4.5, $T_n/n \to \mu$ a.s. From the definition of N_l, it follows that $T(N_l) \leq t < T(N_l + 1)$, so dividing through by N_l gives

$$\frac{T(N_l)}{N_t} \leq \frac{t}{N_t} \leq \frac{T(N_l+1)}{N_t+1} \cdot \frac{N_l+1}{N_t}$$

To take the limit, we note that since $T_n < \infty$ for all n, we have $N_t \uparrow \infty$ as $t \to \infty$. The strong law of large numbers implies that for $\omega \in \Omega_0$ with $P(\Omega_0) = 1$, we have $T_n(\omega)/n \to \mu$, $N_t(\omega) \uparrow \infty$, and hence

$$\frac{T_{N_t(\omega)}(\omega)}{N_t(\omega)} \to \mu \text{ and } \frac{N_t(\omega)+1}{N_t(\omega)} \to 1$$

From this it follows that for $\omega \in \Omega_0$ that $t/N_t(\omega) \to \mu$ a.s.

□

The last argument shows that if $X_n \to X_\infty$ a.s. and $N(n) \to \infty$ a.s. then $X_{N(n)} \to X_\infty$ a.s. We have written this out with care because the analogous result for convergence in probability is false.

Example 2.4.8. Empirical distribution functions. Let $X_1, X_2, ...$ be i.i.d. with distribution F and let

$$F_n(x) = n^{-1} \sum_{m=1}^{n} 1_{(X_m \le x)}$$

$F_n(x)$ = the observed frequency of values that are $\le x$, hence the name given above. The next result shows that F_n converges uniformly to F as $n \to \infty$.

Theorem 2.4.9. ***The Glivenko-Cantelli theorem.*** *As $n \to \infty$,*

$$\sup_x |F_n(x) - F(x)| \to 0 \text{ a.s.}$$

Proof. Fix x and let $Y_n = 1_{(X_n \le x)}$. Since the Y_n are i.i.d. with $EY_n = P(X_n \le x) = F(x)$, the strong law of large numbers implies that $F_n(x) = n^{-1} \sum_{m=1}^{n} Y_m \to F(x)$ a.s. In general, if F_n is a sequence of nondecreasing functions that converges pointwise to a bounded and continuous limit F then $\sup_x |F_n(x) - F(x)| \to 0$. However, the distribution function $F(x)$ may have jumps, so we have to work a little harder.

Again, fix x and let $Z_n = 1_{(X_n < x)}$. Since the Z_n are i.i.d. with $EZ_n = P(X_n < x) = F(x-) = \lim_{y \uparrow x} F(y)$, the strong law of large numbers implies that $F_n(x-) = n^{-1} \sum_{m=1}^{n} Z_m \to F(x-)$ a.s. For $1 \le j \le k-1$ let $x_{j,k} = \inf\{y : F(y) \ge j/k\}$. The pointwise convergence of $F_n(x)$ and $F_n(x-)$ imply that we can pick $N_k(\omega)$ so that if $n \ge N_k(\omega)$ then

$$|F_n(x_{j.k}) - F(x_{j.k})| < k^{-1} \text{ and } |F_n(x_{j,k}-) - F(x_{j,k}-)| < k^{-1}$$

for $1 \le j \le k-1$. If we let $x_{0,k} = -\infty$ and $x_{k,k} = \infty$, then the last two inequalities hold for $j = 0$ or k. If $x \in (x_{j-1,k}, x_{j,k})$ with $1 \le j \le k$ and $n \ge N_k(\omega)$, then using the monotonicity of F_n and F, and $F(x_{j,k}-) - F(x_{j-1,k}) \le k^{-1}$, we have

$$F_n(x) \le F_n(x_{j,k}-) \le F(x_{j,k}-) + k^{-1} \le F(x_{j-1,k}) + 2k^{-1} \le F(X) + 2k^{-1}$$

$$F_n(x) \ge F_n(x_{j-1,k}) \ge F(x_{j-1,k}-) - k^{-1} \ge F(x_{j,k}-) - 2k^{-1} \ge F(X) - 2k^{-1}$$

so $\sup_x |F_n(x) - F(x)| \le 2k^{-1}$, and we have proved the result.

□

Example 2.4.10. Shannon's theorem. Let $X_1, X_2, ..., \in \{1, ..., r\}$ be independent with $P(X_i = k) = p(k) > 0$ for $1 \le k \le r$. Here we are thinking of $1, ..., r$ as the letters of an alphabet, and $X_1, X_2, ...$ are the successive letters produced by an information source. In this i.i.d. case, it is the proverbial monkey at a typewriter. Let $\pi_n(\omega) = p(X_1(\omega)) \dots p(X_n(\omega))$ be the probability of the realization we observed in the first n trials. Since $\log \pi_n(\omega)$ is a sum of independent random variables, it follows from the strong law of large numbers that

$$-n^{-1} \log \pi_n(\omega) \to H \equiv e \sum_{k=1}^{r} p(k) \log p(k) \text{ a.s.}$$

The constant H is called the **entropy** of the source and is a measure of how random it is. The last result is the **asymptotic equipartition property**: If $\epsilon > 0$ then as $n \to \infty$

$$P\{\exp(-n(H+\epsilon)) \le \pi_n(\omega) \le \exp(-n(H-\epsilon))\} \to 1$$

2.5 Convergence of Random Series

In this section, we will pursue a second approach to the strong law of large numbers based on the convergence of random series. This approach has the advantage that it leads and estimates on the rate of convergence under moment assumptions.

Let $\mathcal{F}'_n = \sigma(X_n, X_{n+1}, ...)$ = the future after time n = the smallest σ-field with respect to which all the X_m, $m \geq n$ are measurable. Let $\mathcal{R} = \cap_n \mathcal{F}'_n$ = the remote future, or **tail σ-field**. Intuitively, $A \in \mathcal{T}$ if and only if changing a finite number of values does not affect the occurrence of the event.

Example 2.5.1. If $B_n \in \mathcal{R}$ then $\{X_n \in B_n \text{ i.o.}\} \in \mathcal{T}$. If we let $X_n = 1_{A_n}$ and $B_n = \{1\}$, this example becomes $\{A_n \text{ i,o,}\}$.

Example 2.5.2. Let $S_n = X_1 + \cdots + X_n$. It is easy to check that
$\{\lim_{n\to\infty} S_n \text{ exists}\} \in \mathcal{T}$,
$\{\limsup_{n\to\infty} S_n > 0\} \notin \mathcal{T}$,
$\{\limsup_{n\to\infty} S_n/c_n > x\} \in \mathcal{T}$ if $c_n \to \infty$.
The next result shows that all examples are trivial.

Theorem 2.5.3. ***Kolmogorov's 0-1 law***. *If $X_1, X_2, ...$ are independent and $A \in \mathcal{T}$ then $P(A) = 0$ or 1.*

Proof. We will show that A is independent of itself, that is, $P(A \cap A) = P(A)P(A)$, so $P(A) = P(A)^2$, and hence $P(A) = 0$ or 1. We will sneak up on this conclusion in two steps:

(a) $A \in \sigma(X_1, ..., X_k)$ and $B \in \sigma(X_{k+1}, X_{k+2}, ...)$ are independent.

Proof of (a). If $B \in \sigma(X_{k+1}, ..., X_{k+j})$ for some j, this follows from Theorem 2.1.7. Since $\sigma(X_1, ..., X_k)$ and $\cup_j \sigma(X_{k+1}, ..., X_{k+j})$ are π-systems that contain Ω (a) follows from Theorem 2.1.7.

(b) $A \in \sigma(X_1, X_2, ...)$ and $B \in \mathcal{T}$ are independent.

Proof of (b). Since $\mathcal{T} \subset \sigma(X_{k+1}, X_{k+2}, ..)$, if $A \in \sigma(X_1, ..., X_k)$ for some k, this follows from (a). $\cup_k \sigma(X_1, ..., X_k)$ and $\mathcal{T}$ are π-systems that contain Ω, so 9b) follows from Theorem 2.1.7.

Since $\mathcal{T} \subset \sigma(X_1, X_2, ...)$, (b) implies an $A \in \mathcal{T}$ is independent of itself and Theorem 2.5.3 follows.

□

Theorem 2.5.4. ***Kolmogorov's maximal inequality***. *Suppose $X_1, ..., X_n$ are independent with $EX_i = 0$ and $var(X_i) < \infty$. If $S_n = X_1 + \cdots + X_n$ then*

$$P\left(\max_{1\leq k\leq n} |S_k| \geq x\right) \leq x^{-2} var(S_n)$$

Remark 2.5.5. Under the same hypotheses, Chebyshev's inequality (Theorem 1.6.4) gives only

$$P(|S_n| \geq x) \leq x^{-2}\text{var}(S_n)$$

Proof. Let $A_k = \{|S_k| \geq x \text{ but } |S_j| < x \text{ for } j < k\}$, i.e., we break things down according to the time that $|S_k|$ first exceeds x. Since the A_k are disjoint and $(S_n - S_k)^2 \geq 0$,

$$\begin{aligned} ES_n^2 &\geq \sum_{k=1}^{n} \int_{A_k} S_n^2 dP \\ &= \sum_{k=1}^{n} \int_{A_k} S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2 dP \\ &\geq \sum_{k=1}^{n} \int_{A_k} S_k^2 dP + \sum_{k=1}^{n} \int 2S_k 1_{A_k} \cdot (S_n - S_k) dP \end{aligned}$$

$S_k 1_{A_k} \in \sigma(X_1, ..., X_k)$ and $S_n - S_k \in \sigma(X_{k+1}, ..., X_n)$ are independent by Theorem 2.1.8, so using Theorem 2.1.11 and $E(S_n - S_k) = 0$ shows

$$\int 2S_k 1_{A_k} \cdot (S_n S_k) dP = E(2S_k 1_{A_k}) \cdot E(S_n - S_k) = 0$$

Using now the fact that $|S_k| \geq x$ on A_k and the A_k are disjoint,

$$ES_n^2 \geq \sum_{k=1}^{n} \int_{A_k} S_k^2 dP \geq \sum_{k=1}^{n} x^2 P(A_k) = x^2 P\left(\max_{1 \leq k \leq n} |S_k| \geq x \right)$$

Theorem 2.5.6. *Suppose $X_1, X_2, ...$ are independent and have $EX_n = 0$. If*

$$\sum_{n=1}^{\infty} var(X_n) < \infty$$

then with probability one $\sum_{n=1}^{\infty} X_n(\omega)$ converges.

Proof. Let $S_N = \sum_{n=1}^{N} X_n$. From Theorem 2.5.4, we get

$$P\left(\max_{M \leq m \leq N} |S_m - S_M| > \epsilon \right) \leq \epsilon^{-2} \text{var}(S_N - S_M) = \epsilon^{-2} \sum_{n=M+1}^{N} \text{var}(X_n)$$

Letting $N \to \infty$ in the last result, we get

$$P\left(\sup_{m \geq M} |S_m - S_M| > \epsilon \right) \epsilon^{-2} \sum_{n=M+1}^{\infty} \text{var}(X_n) \to 0 \text{ as } M \to \infty$$

If we let $w_M = \sup_{m,n \geq M} |S_m - S|n|$ then $w_M \downarrow$ as $M \uparrow$ and

$$P(w_M > 2\epsilon) \leq P\left(\sup_{m \geq M} |S_m - S_M| > \epsilon \right) \to 0$$

as $M \to \infty$ so $\omega_M \downarrow 0$ almost surely. But $w_M(\omega)$ is a Cauchy sequence and hence $\lim_{n \to \infty} S_n(\omega)$ exists, so the proof is complete.

□

Example 2.5.7. Let $X_1, X_2, ...$ be independent with

$$P(X_n = n^{-\alpha}) = P(X_n = -n^{-\alpha}) = 1/2$$

$EX_n = 0$ and $\text{var}(X_n) = n^{-2\alpha}$ so if $\alpha > 1/2$ it follows from Theorem 2.5.6 that $\sum X_n$ converges. The theorem below shows that $\alpha > 1/2$ is also necessary for this conclusion. Note that there is absolute convergence, i.e., $\sum |X_n| < \infty$, if and only if $\alpha > 1$.

Theorem 2.5.6 is sufficient for all of our applications, but our treatment would not be complete if we did not mention the last word on convergence of random series.

Theorem 2.5.8. ***Kolmogorov's three-series theorem***. *Let $X_1, X_2, ...$ be independent. Let $A > 0$ and let $Y_i = X_i 1_{(|X_i| \leq A)}$. In order that $\sum_{n=1}^{\infty} X_n$ converges a.s., it is necessary and sufficient that*

$$(i) \sum_{n=1}^{\infty} P(|X_n| > A) < \infty, \quad (ii) \sum_{n=1}^{\infty} EY_n \text{ converges, and } (iii) \sum_{n=1}^{\infty} var(Y_n) < \infty$$

Proof. We will prove the necessity as an application of the central limit theorem. To prove the sufficiency, let $\mu_n = EY_n$. (iii) and Theorem 2.5.6 imply that $\sum_{n=1}^{\infty}(Y_n - \mu_n)$ converges a.s. Using (ii) now gives that $\sum_{n=1}^{\infty} Y_n$ converges a.s. (i) and the Borel-Cantelli lemma imply $P(X_n \neq Y_n \text{ i.o.}) = 0$, so $\sum_{n=1}^{\infty} X_n$ converges a.s.

□

The link between convergence of series and the strong law of large numbers is provided by

Theorem 2.5.9. ***Kronecker's lemma***. *If $a_n \uparrow \infty$ and $\sum_{n=1}^{\infty} x_n/a_n$ converges then*

$$a_n^{-1}\sum_{m=1}^{n} x_m \to 0$$

Proof. Let $a_0 = 0$, $b_0 = 0$, and for $m \geq 1$, let $b_m = \sum_{k=1}^{m} x_k/a_k$. Then $x_m = a_m(b_m - b_{m-1})$ and so

$$\begin{aligned} a_n^{-1}\sum_{m=1}^{n} x_m &= a_n^{-1}\left\{\sum_{m=1}^{n} a_m b_m - \sum_{m=1}^{n} a_m b_{m-1}\right\} \\ &= a_n^{-1}\left\{a_n b_n + \sum_{m=2}^{n} a_{m-1}b_{m-1} - \sum_{m=1}^{n} a_m b_{m-1}\right\} \\ &= b_n - \sum_{m=1}^{n}\frac{(a_m - a_{m-1})}{a_n} b_{m-1} \end{aligned}$$

(Recall $a_0 = 0$.) By hypothesis, $b_n \to b_\infty$ as $n \to \infty$. Since $a_m - a_{m-1} \geq 0$, the last sum is an average of $b_0, ..., b_n$. Intuitively, if $\epsilon > 0$ and $M < \infty$ are fixed and n is large, the average assigns mass $\geq 1 - \epsilon$ to the b_m with $m \geq M$, so

$$\sum_{m=1}^{n}\frac{(a_m - a_{m-1})}{a_n} b_{m-1} \to b_\infty$$

To argue formally, let $B = \sup|b_n|$, pick M so that $|b_m - b_\infty| < \epsilon/2$ for $m \geq M$, then pick N so that $a_M/a_n < \epsilon/4B$ for $n \geq N$. Now if $n \geq N$, we have

$$\begin{aligned} \left|\sum_{m=1}^{n}\frac{(a_m - a_{m-1})}{a_n} b_{m-1} - b_\infty\right| &\leq \sum_{m=1}^{n}\frac{(a_m - a_{m-1})}{a_n}|b_{m-1} - b_\infty| \\ &\leq \frac{a_M}{a_n}\cdot 2B + \frac{a_n - a_M}{a_n}\cdot\frac{\epsilon}{2} < \epsilon \end{aligned}$$

proving the desired result since ϵ is arbitrary.

□

Theorem 2.5.10. ***The strong law of large numbers***. *Let $x_1, X_2, ...$ be i.i.d. random variables with $E|X_i| < \infty$. Let $EX_i = \mu$ and $S_n = X_1 + \cdots + X_n$. Then $S_n/n \to \mu$ a.s. as $n \to \infty$.*

Proof. Let $Y_k = X_k 1_{(|X_k| \leq k)}$ and $T_n = Y_1 + \cdots + Y_n$. By (a) in the proof of Theorem 2.4.1 imply

$$\sum_{k=1}^{\infty}\text{var}(Z_k)/k^2 \leq \sum_{k=1}^{\infty} EY_k^2/k^2 < \infty$$

Applying Theorem 2.5.6, we conclude that $\sum_{k=1}^{\infty} Z_k/k$ converges a.s. so Theorem 2.5.9 implies

$$n^{-1}\sum_{k=1}^{n}(Y_k - EY_k) \to 0 \text{ and hence } \frac{T_n}{n} - n^{-1}\sum_{k=1}^{n} EY_k \to 0 \text{ a.s.}$$

The dominated convergence theorem implies $EY_k \to \mu$ as $k \to \infty$. From this, it follows easily that $n^{-1}\sum_{k=1}^{n} EY_k \to \mu$ and hence $T_n/n \to \mu$.

□

Theorem 2.5.11. *Let $X_1, X_2, \ldots$ be i.i.d. random variables with $EX_i = 0$ and $EX_i^2 = \sigma^2 < \infty$. Let $S_n = X_1 + \cdots + X_n$. If $\epsilon > 0$ then*

$$S_n/n^{1/2}(\log n)^{1/2+\epsilon} \to 0 \text{ a.s.}$$

Remark 2.5.12. Kolmogorov's test, later theorem will show that

$$\limsup_{n\to\infty} S_n/n^{1/2}(\log\log n)^{1/2} = \sigma\sqrt{2} \text{ a.s.}$$

so the last result is not far from the best possible.

Proof. Let $a_n = n^{1/2}(\log n)^{1/2+\epsilon}$ for $n \geq 2$ and $a_1 > 0$.

$$\sum_{n=1}^{\infty} \text{var}(X_n/a_n) = \sigma^2\left(\frac{1}{a_1^2} + \sum_{n=1}^{\infty} \frac{1}{n(\log n)^{1+2\epsilon}}\right) < \infty$$

so applying Theorem 2.5.6 we get $\sum_{n=1}^{\infty} X_n/a_n$ converges a.s. and the indicated result follows from Theorem 2.5.9.

□

The next result due to Marcinkiewicz and Zygmund treats the situation in which $EX_i^2 = \infty$ but $E|X_i|^P < \infty$ for some $1 < p < 2$.

Theorem 2.5.13. *Let $X_1, X_2, \ldots$ be i.i.d. with $EX_1 = 0$ and $E|X_1|^p < \infty$ where $1 < p < 2$. If $S_n = X_1 + \cdots + X_n$ then $S_n/n^{1/p} \to 0$ a.s.*

Proof. Let $Y_k = X_k 1_{(|X_k| \leq k^{1/p})}$ and $T_n = Y_1 + \cdots + Y_n$.

$$\sum_{k=1}^{\infty} P(Y_k \neq X_k) = \sum_{k=1}^{\infty} P(|X_k|^p > k) \leq E|X_k|^p < \infty$$

so the Borel-Cantelli lemma implies $P(Y_k \neq X_k \text{ i.o.}) = 0$, and it suffices to show $T_n/n^{1/p} \to 0$. Using $\text{var}(Y_m) \leq E(Y_m^2)$, Lemma 2.2.14 with $p = 2$, $P(|Y_m| > y) \leq P(|X_1| > y)$, and Fubini's theorem (everything is ≥ 0) we have

$$\begin{aligned}
\sum_{m=1}^{\infty} \text{var}(Y_m/m^{1/p}) &\leq \sum_{m=1}^{\infty} EY_m^2/m^{2/p} \\
&\leq \sum_{m=1}^{\infty}\sum_{n=1}^{m} \int_{(n-1)^{1/p}}^{n^{1/p}} \frac{2y}{m^{2/p}} P(|X_1| > y)dy \\
&= \sum_{n=1}^{\infty} \int_{(n-1)^{1/p}}^{n^{1/p}} \sum_{m=n}^{\infty} \frac{2y}{m^{2/p}} P(|X_1| > y)dy
\end{aligned}$$

To bound the integral, we note that for $n \geq 2$ comparing the sum with the integral of $x^{-2/p}$

$$\sum_{m=n}^{\infty} m^{-2/p} \leq \frac{p}{2-p}(n-1)^{(p-2)/p} \leq Cy^{p-2}$$

when $y \in [(n-1)^{1/p}, n^{1/p}]$. Since $E[X_1]^p = \int_0^{\infty} px^{p-1}P(|X_i| > x)dx < \infty$, it follows that

$$\sum_{m=1}^{\infty} \text{var}(Y_m/m^{1/p}) < \infty$$

If we let $\mu_n = EY_m$ and apply Theorem 2.5.6 and Theorem 2.5.9 it follows that

$$n^{-1/p}\sum_{m=1}^{n}(Y_m - \mu_m) \to 0 \text{ a.s.}$$

To estimate μ_n, we note that since $EX_m = 0$, $\mu_m = -E(X_i; |X_i| > m^{1/p})$, so

$$\begin{aligned} |\mu_m| &\le E(|X|; |X_i| > m^{1/p}) = m^{1/p}E(|X|/m^{1/p}; |X_i| > m^{1/p}) \\ &\le m^{1/p}E((|X|/m^{1/p})^p; |X_i| > m^{1/p}) \\ &\le m^{-1+1/p}p^{-1}E(|X_i|^p; |X_i| > m^{1/p}) \end{aligned}$$

Now $\sum_{m=1}^{n} m^{-1+1/p} \le Cn^{1/p}$ and $E(|X_i|^p; |X_i| > m^{1/p}) \to 0$ as $m \to \infty$, so $n^{-1/p}\sum_{m=1}^{n}\mu_m \to 0$ and the desired result follows.

□

The St. Petersburg game, discussed earlier, is a situation in which $EX_i = \infty$, $S_n/n\log_2 n \to 1$ in probability but

$$\limsup_{n\to\infty} S_n/(n\log_2 n) = \infty \text{ a.s.}$$

The next result, due to Feller (1946) [6], shows that when $E|X_1| = \infty$, S_n/a_n cannot converge almost surely to a nonzero limit. In Theorem 2.3.8 we considered the special $a_n = n$.

Theorem 2.5.14. *Let $X_1, X_2, \ldots$ be i.i.d. with $E|X_1| = \infty$ and let $S_n = X_1 + \cdots + X_n$. Let a_n be a sequence of positive numbers with a_n/n increasing. Then $\limsup_{n\to\infty} |S_n|/a_n = 0$ or ∞ according as $\sum_n P(|X_1| \ge a_n) < \infty$ of ∞.*

Proof. Since $a_n/n \uparrow$, $a_{kn} \ge ka_n$ for any integer k. Using this and $a_n \uparrow$,

$$\sum_{n=1}^{\infty} P(|X_1| \ge ka_n) \ge \sum_{n=1}^{\infty} P(|X_1| \ge a_{kn}) \ge \frac{1}{k}\sum_{m=k}^{\infty} P(|X_1| \ge a_m)$$

The last observation shows that if the sum is infinite, $\limsup_{n\to\infty} |X_n|/a_n = \infty$. Since $\max\{|S_{n-1}|, |S_n|\} \ge |X_n|/2$, it follows that $\limsup_{n\to\infty} |S_n|/a_n = \infty$.

To prove the other half, we begin with the identity

$$(\star)\quad \sum_{m=1}^{\infty} mP(a_{m-1} \le |X_i| < a_m) = \sum_{n=1}^{\infty} P(|X_i| \ge a_{n-1})$$

To see this, write $m = \sum_{n=1}^{m} 1$ and then use Fubini's theorem. We now let $Y_n = X_n 1_{(|X_n|<a_n)}$, and $T_n = Y_1 + \cdots + Y + n$. When the sum is finite, $P(Y_n \ne X_n \text{ i.o.}) = 0$, and it suffices to investigate the behavior of the T_n. To do this, we let $a_0 = 0$ and compute

$$\begin{aligned} \sum_{n=1}^{\infty} \text{var}(Y_n/a_n) &\le \sum_{n=1}^{\infty} EY_n^2/a_n^2 \\ &= \sum_{n=1}^{\infty} a_n^{-2}\sum_{m=1}^{n}\int_{[a_{m-1},a_m)} y^2 dF(y) \\ &= \sum_{m=1}^{\infty}\int_{[a_{m-1},a_m)} y^2 dF(y)\sum_{n=m}^{\infty} a_n^{-2} \end{aligned}$$

Since $a_n \ge na_m/m$, we have $\sum_{n=m}^{\infty} a_n^{-2} \le (m^2/a_m^2)\sum_{n=m}^{\infty} n^{-2} \le Cma_m^{-2}$, so

$$\le C\sum_{m=1}^{\infty} m\int_{[a_{m-1},a_m)} dF(y)$$

Using $(\star)$ now, we conclude $\sum_{n=1}^{\infty} \text{var}(Y_n/a_n) < \infty$.

The last step is to show $ET_n/a_n \to 0$. To begin we note that if $E|X_i| = \infty$, $\sum_{n=1}^{\infty} P(|X_i| > a_n) < \infty$, and $a_n/n \uparrow$ we must have $a_n/n \uparrow \infty$. To estimate ET_n/a_n now, we observe that

$$\begin{aligned} \left| a_n^{-1} \sum_{m=1}^{n} EY_m \right| &\leq a_n^{-1} n \sum_{m=1}^{n} E(|X_m|; |X_m| < a_m) \\ &\leq \frac{na_N}{a_n} + \frac{n}{a_n} E(|X_i|; a_N \leq |X_i| < a_n) \end{aligned}$$

where the last inequality holds for any fixed N. Since $a_n/n \to \infty$, the first term converges to 0. Since $m/a_m \downarrow$, the second is

$$\begin{aligned} &\leq \sum_{m=N+1}^{n} \frac{m}{a_m} E(|X_i|; a_{m-1} \leq |X_i| < a_m) \\ &\leq \sum_{m=N+1}^{\infty} mP(a_{m-1} \leq |X_i| < a_m) \end{aligned}$$

$(\star)$ shows that the sum is finite, so it is small if N is large and the desired result follows.

□

2.6 Large Deviations

Let $X_1, X_2, ...$ be i.i.d. and let $S_n = X_1 + \cdots + X_n$. In this section, we will investigate the rate at which $PP(S_n > na) \to 0$ for $a > \mu = EX_i$. We will ultimately conclude that if the **moment-generating function** $\varphi(\theta) = E \exp(\theta X_i) < \infty$ for some $\theta > 0$, $P(S_n \geq na) \to 0$ exponentially rapidly and we will identify

$$\gamma(a) = \lim_{n \to \infty} \frac{1}{n} \log P(S_n \geq na)$$

Our first step is to prove that the limit exists. This is based on an observation that will be useful several times below. Let $\pi_n = P(S_n \geq na)$.

$$\pi_{m+n} \geq P(S_m \geq ma, S_{n+m} - S_m \geq na) = \pi_m \pi_n$$

since S_m and $S_{n+m} - S_m$ are independent. Letting $\gamma_n = \log \pi_n$ transforms multiplication into addition.

Lemma 2.6.1. *If $\gamma_{m+n} \geq \gamma_m + \gamma_n$ then as $n \to \infty$, $\gamma_n/n \to \sup_m \gamma_m/m$.*

Proof. Clearly, $\limsup \gamma_n/n \leq \sup \gamma_m/m$. To complete the proof, it suffices to prove that for any $m \liminf \gamma_n/n \geq \gamma_m/m$.Writing $n = km + l$ with $0 \leq l < m$ and making repeated use of the hypothesis gives $\gamma_n \geq k\gamma_m + \gamma_l$. Dividing by $n = km + l$ gives

$$\frac{\gamma(n)}{n} \geq \left(\frac{km}{km + l} \right) \frac{\gamma(m)}{m} + \frac{\gamma(l)}{n}$$

Letting $n \to \infty$ and recalling $n = km + l$ with $0 \leq l < m$ gives the desired result.

□

Lemma 2.6.1 implies that $\lim_{n \to \infty} \frac{1}{n} \log P(S_n \geq na) = \gamma(a)$ exists ≤ 0. It follows from the formula for the limit that

$$P(S_n \geq na) \leq e^{n\gamma(a)} \tag{2.4}$$

The last two observations gives us some useful information about $\gamma(a)$.

The conclusions above are valid for any distribution. For the rest of this section, we will suppose:

$$\text{(H1) } \varphi(\theta) = E\exp(\theta X_i) < \infty \text{ for some } \theta > 0$$

Let $\theta_+ = \sup\{\theta : \phi(\theta) < \infty\}$, $\theta_- = \inf\{\theta : \phi(\theta) < \infty\}$ and note that $\phi(\theta) < \infty$ for $\theta \in (\theta_-, \theta_+)$. (H1) implies that $EX_i^+ < \infty$ so $\mu = EX^+ - EX^- \in [-\infty, \infty)$. If $\theta > 0$ Chebyshev's inequality implies

$$e^{\theta na} P(S_n \geq na) \leq E\exp(\theta S_n) = \varphi(\theta)^n$$

or letting $\kappa(\theta) = \log \varphi(\theta)$

$$P(S_n \geq na) \leq \exp(-n\{a\theta - \kappa(\theta)\}) \tag{2.5}$$

Out first goal is to show:

Lemma 2.6.2. *If $a > \mu$ and $\theta > 0$ is small then $a\theta - \kappa(\theta) > 0$.*

Proof. $\kappa(0) = \log \varphi(0) = 0$, so it suffices to show that (i) κ is continuous at 0, (ii) differentiable on $), \theta_+)$, and (iii) $\kappa'(\theta) \to \mu$ as $\theta \to 0$. For then

$$a\theta - \kappa(\theta) = \int_0^\theta a - \kappa'(x)dx > 0$$

for small θ.

Let $F(x) = P(X_i \leq x)$. To prove (i) we note that if $0 < \theta < \theta_0 < \theta_-$

$$e^{\theta x} \leq 1 + e^{\theta_0 x} (\star)$$

so by the dominated convergence theorem as $\theta \to 0$

$$\int e^{\theta x} dF \to \int 1 dF = 1$$

To prove (ii) we note that if $|h| < h_0$ then

$$|e^{hx} - 1| = \left| \int_0^{hx} e^y dy \right| \leq |hx| e^{h_0 x}$$

so an application of the dominated convergence theorem shows that

$$\begin{aligned} \varphi'(\theta) &= \lim_{h \to 0} \frac{\varphi(\theta + h) - \varphi(\theta)}{h} \\ &= \lim_{h \to 0} \int \frac{e^{hx} - 1}{h} e^{\theta x} dF(x) \\ &= \int x e^{\theta x} dF(x) \text{ for } \theta \in (0, \theta_+) \end{aligned}$$

From the last equation, it follows that $\kappa(\theta) = \log \phi(\theta)$ has $\kappa'(\theta) = \phi'(\theta)/\phi(\theta)$. Using $(\star)$ and the dominated convergence theorem gives (iii) and the proof is complete.

□

Having found an upper bound on $P(S_n \geq na)$, it is natural to optimize it by finding the maximum of $\theta a - \kappa(\theta)$:

$$\frac{d}{d\theta}\{\theta a - \log \varphi(\theta)\} = a - \varphi'(\theta)/\varphi(\theta)$$

so (assuming things are nice) the maximum occurs when $a = \varphi'(\theta)/\varphi(\theta)$. To turn the parenthetical clause into a mathematical hypothesis we begin by defining

$$F_\theta(x) = \frac{1}{\varphi(\theta)} \int_{-\infty}^{x} e^{\theta y} dF(y)$$

whenever $\phi(\theta) < \infty$. It follows from the proof of Lemma 2.6.2 that if $\theta \in (\theta_-, \theta_+)$, F_θ is a distribution function with mean

$$\int x dF_\theta(x) = \frac{1}{\varphi(\theta)} \int_{-\infty}^{\infty} x e^{\theta x} dF(x) = \frac{\varphi'(\theta)}{\varphi(\theta)}$$

Repeating the proof in Lemma 2.6.1, it is easy to see that if $\theta \in (\theta_-, \theta_+)$ then

$$\phi''(\theta) = \int_{-\infty}^{\infty} x^2 e^{\theta x} dF(x)$$

So we have

$$\frac{d}{d\theta} \frac{\varphi'(\theta)}{\varphi(\theta)} = \frac{\varphi''(\theta)}{\varphi(\theta)} - \left(\frac{\varphi'(\theta)}{\varphi(\theta)} \right)^2 = \int x^2 dF_\theta(x) - \left(\int x dF_\theta(x) \right)^2 \geq 0$$

since the last expression is the variance of F_θ. If we assume

(H2) the distribution F is not a point mass at μ

then $\varphi'(\theta)/\varphi(\theta)$ is strictly increasing and $a\theta - \log \phi(\theta)$ is concave. Since we have $\varphi'(0)/\varphi(0) = \mu$, this shows that for each $a > \mu$ there is at most one $\theta_a \geq 0$ that solves $a = \varphi'(\theta_a)/\varphi(\theta_a)$, and this value of θ maximizes $a\theta - \log \varphi(\theta)$. Before discussing the existence of θ_a we will consider some examples.

Example 2.6.3. Normal distribution.

$$\int e^{\theta x} (2\pi)^{-1/2} \exp(-x^2/2) dx = \exp(\theta^2/2) \int (2\pi)^{-1/2} \exp(-(x-\theta)^2/2) dx$$

The integrand in the last integral is the density of a normal distribution with mean θ and variance 1, so $\varphi(\theta) = \exp(\theta^2/2)$, $\theta \in (-\infty, \infty)$. In this case, $\varphi'(\theta)/\varphi(\theta) = \theta$ and

$$F_\theta(x) = e^{-\theta^2/2} \int_{-\infty}^{x} e^{\theta y} (2\pi)^{-1/2} e^{-y^2/2} dy$$

is a normal distribution with mean θ and variance 1.

Example 2.6.4. Exponential distribution with parameter λ. If $\theta < \lambda$

$$\int_0^{\infty} e^{\theta x} \lambda e^{-\lambda x} dx = \lambda/(\lambda - \theta)$$

$\varphi'(\theta)\varphi(\theta) = 1/(\lambda - \theta)$ and

$$F_\theta(x) = \frac{\lambda}{\lambda - \theta} \int_0^{x} e^{\theta y} \lambda e^{-\lambda y} dy$$

is an exponential distribution with parameter $\lambda - \theta$ and hence mean $1/(\lambda - \theta)$.

Example 2.6.5. Coin flips. $P(X_i = 1) = P(X_i = -1) = 1/2$

$$\varphi(\theta) = (e^\theta + e^{-\theta})/2$$
$$\varphi'(\theta)/\varphi(\theta) = (e^\theta - e^{-\theta})/(e^\theta + e^{-\theta})$$

$F_\theta(\{x\})/F(\{x\}) = e^{\theta x}/\phi(\theta)$ so

$$F_\theta(\{1\}) = e^\theta/(e^\theta + e^{-\theta}) \text{ and } F_\theta(\{-1\}) = e^{-\theta}/(e^\theta + e^{-\theta})$$

Example 2.6.6. Perverted exponential. Let $g(x) = Cx^{-3}e^{-x}$ for $x \geq 1$, $g(x) = 0$ otherwise, and choose C so that g is a probability density. In this case,

$$\varphi(\theta) = \int e^{\theta x} g(x) dx < \infty$$

if and only if $\theta \leq 1$, and when $\theta \leq 1$, we have

$$\frac{\varphi'(\theta)}{\varphi(\theta)} \leq \frac{\varphi'(1)}{\varphi(1)} = \int_1^\infty Cx^{-2}dx \Big/ \int_1^\infty Cx^{-3}dx = 2$$

Recall $\theta_+ = \sup\{\theta : \varphi(\theta) < \infty\}$. In examples with normal distribution and exponential distribution, we have $\phi'(\theta)/\phi(\theta) \uparrow \infty$ as $\theta \uparrow \theta_+$ so we can solve $a = \phi'(\theta)/\phi(\theta)$ for any $a > \mu$. In example coin flips, $\phi'(\theta)\phi(\theta) \uparrow 1$ as $\theta \to \infty$, but we cannot hope for much more since F and hence F_θ is supported on $\{-1, 1\}$.

Theorem 2.6.7. *Suppose in addition to (H1) and (H2) that there is a $\theta_a \in (0, \theta_+)$ so that $a = \varphi'(\theta_a)/\varphi(\theta_a)$. Then, as $n \to \infty$,*

$$n^{-1} \log P(S_n \geq na) \to -a\theta_a + \log \varphi(\theta_a)$$

Proof. The fact that the lim sup of the left-hand side $\leq$ the right-hand side follows from 2.5. To prove the other inequality, pick $\lambda \in (\theta_a, \theta_+)$, let $X_1^\lambda, X_2^\lambda, ...$ be i.i.d. with distribution F_λ and let $S_n^\lambda = X_1^\lambda + \cdots + X_n^\lambda$. writing dF/dF_λ for the Radon-Nikodym derivative of the associated measures, it is immediate from the definition that $dF/dF_\lambda = e^{-\lambda x}\varphi(\lambda)$. If we let F_λ^n and F^n denote the distributions of S_n^λ and S_n, then

Lemma 2.6.8. $\frac{dF^n}{dF_\lambda^n} = e^{-\lambda x}\varphi(\lambda)^n$.

Proof. we will prove this by induction. The result holds when $n = 1$. For $n > 1$, we note that

$$\begin{aligned}
F^n &= F^{n-1} \star F(z) \\
&= \int_{-\infty}^{\infty} dF^{n-1}(x) \int_{-\infty}^{z-x} dF(y) \\
&= \int dF_\lambda^{n-1}(x) \int dF_\lambda(y) 1_{(x+y\leq z)} e^{-\lambda(z+y)} \varphi(\lambda)^n \\
&= E\big(1_{(S_{n-1}^\lambda + X_n^\lambda \leq z)} e^{-\lambda(S_{n-1}^\lambda + X_n^\lambda)} \varphi(\lambda)^n\big) \\
&= \int_{-\infty}^{x} dF_\lambda^n(u) e^{-\lambda u} \varphi(\lambda)^n
\end{aligned}$$

where in the last two equalities we have used Theorem 1.6.10 for $(S_{n-1}^\lambda, X_n^\lambda)$ and S_n^λ.

□

If $v > a$, then the lemma and monotonicity imply

$$(\star) P(S_n \geq na) \geq \int_{na}^{nv} e^{-\lambda x}\varphi(\lambda)^n dF_\lambda^n(x) \geq \varphi(\lambda)^n e^{-\lambda nv}(F_\lambda^n(nv) - F_\lambda^n(na))$$

F_λ has mean $\varphi'(\lambda)/\varphi(\lambda)$, so if we have $a < \varphi'(\lambda)/\varphi(\lambda) < \nu$, then the weak law of large numbers implies

$$F_\lambda^n(n\nu) - F_\lambda^n(na) \to 1 \text{ as } n \to \infty$$

From the last conclusion and $(\star)$ it follows that

$$\liminf_{n\to\infty} n^{-1} \log P(S_n > na) \geq -\lambda\nu + \log \phi(\lambda)$$

Since $\lambda > \theta_a$ and $\nu > a$ are arbitrary, the proof is complete.

□

To get a feel for what the answers look like, we consider our examples. To prepare for the computations, we recall some important information:

$$\kappa(\theta) = \log \phi(\theta)\ \kappa'(\theta) = \phi'(\theta)/\phi(\theta)\ \theta_a \text{ solves } \kappa'(\theta_a) = a$$

$$\gamma(a) = \lim_{n\to\infty} (1/n) \log P(S_n \geq na) = -a\theta_a + \kappa(\theta_a)$$

Normal distribution

$$\kappa(\theta) = \theta^2/2\ \kappa'(\theta) = \theta\ \theta_a = a$$

$$\gamma(a) = -a\theta_a + \kappa(\theta_a) = -a^2/2$$

Exponential distribution with $\lambda = 1$

$$\kappa(\theta) = -\log(1-\theta)\ \kappa'(\theta) = 1/(1-\theta)\ \theta_a = 1 - 1/a$$

$$\gamma(a) = -a\theta_a + \kappa(\theta_a) = -a + 1 + \log a$$

Theorem 2.6.9. *Suppose $x_0 = \infty$, $\theta_+ < \infty$, and $\varphi'(\theta)/\varphi(\theta)$ increases to a finite limit a_0 as $\theta \uparrow \theta_+$. If $a_0 \leq a < \infty$*

$$n^{-1} \log P(S_n \geq na) \to -a\theta_+ + \log \varphi(\theta_+)$$

i.e., $\gamma(a)$ is linear for $a \geq a_0$.

Proof. Since $(\log \varphi(\theta))' = \varphi'(\theta)/\varphi(\theta)$, integrating from 0 to θ_+ shows that $\log(\varphi(\theta_+)) < \infty$. Letting $\theta = \theta - +$ in 2.5 shows that the lim sup of the left-hand side $\leq$ the right-hand side. To get the other direction we will use the transformed distribution F_λ, for $\lambda = \theta_+$. Letting $\theta \uparrow \theta_+$ and using the dominated convergence theorem for $x \leq 0$ and the monotone convergence theorem for $x \geq 0$, we see that F_λ has mean a_0. From $(\star)$ in the proof of the Theorem 2.6.7, we see that if $a_0 \leq a < \nu = a + 3\epsilon$

$$P(S_n \geq na) \geq \varphi(\lambda)^n e^{-n\lambda\nu}(F_\lambda^n(n\nu) - F_\lambda^n(na))$$

and hence

$$\frac{1}{n} \log P(S_n \geq na) \geq \log \varphi(\lambda) - \lambda\nu + \frac{1}{n} \log P(S_n^\lambda \in (na, n\nu])$$

Letting $X_1^\lambda, X_2^\lambda, \ldots$ be i.i.d. with distribution F_λ and $S_n^\lambda = X_1^\lambda + \cdots + X_n^\lambda$, we have

$$\begin{aligned} P(S_n^\lambda \in (na, n\nu]) &\geq P\{S_{n-1}^\lambda \in ((a_0-\epsilon)n, (a_0+\epsilon)n]\} \\ &\quad \cdot P\{X_n^\lambda \in ((a-a_0+\epsilon)n, (a-a_0+2\epsilon)n]\} \\ &\geq \frac{1}{2} P\{X_n^\lambda \in ((a-a_0+\epsilon)n, (a-a_0+\epsilon)(n+1)]\} \end{aligned}$$

for large n by the weak law of large numbers. To get a lower bound on the right-hand side of the last equation, we observe that

$$\limsup_{n\to\infty} \frac{1}{n} \log P(X_1^\lambda \in ((a-a_0+\epsilon)n, (a-a_0+\epsilon)(n+1)]) = 0$$

for if the lim sup was < 0, we would have $E \exp(\eta X_1^\lambda) < \infty$ for some $\eta > 0$ and hence $E \exp((\lambda+\eta)X_1) < \infty$, contradicting the definition of $\lambda = \theta_+$. To finish the argument now, we recall that Theorem 2.6.1 implies that

$$\limsup_{n\to\infty} \frac{1}{n} \log P(S_n \geq na) = \gamma(a)$$

exists, so our lower bound on the lim sup is good enough.

□

3 CENTRAL LIMIT THEOREMS

Go back to Table of Contents. Please click TOC

3.1 The De Moivre-Laplace Theorem

Let $X_1, X_2, ...$ be i.i.d. with $P(X_1 = 1) = P(X_1 = -1) = 1/2$ and let $S_n = X_1 + \cdots + X_n$. In words, we are betting \$1 on the flipping of a fair coin and S_n is our winnings at time n. If n and k are integers

$$P(S_{2n} = 2k) = \binom{2n}{n+k} 2^{-2n}$$

since $S_{2n} = 2k$ if and only if there are $n + k$ flips that are $+1$ and $n - k$ flips are -1 in the first $2n$. The first factor gives the number of such outcomes and the second the probability of each one. **Stirling's formula** (see Feller, Vol. I. (1968), p.52) [7] tells us

$$n! \sim n^n e^{-n} \sqrt{2\pi n} \text{ as } n \to \infty \tag{3.1}$$

where $a_n \sim b_n$ means $a_n/b_n \to 1$ as $n \to \infty$, so

$$\begin{aligned} \binom{2n}{n+k} &= \frac{(2n)!}{(n+k)!(n-k)!} \\ &\sim \frac{(2n)^{2n}}{(n+k)^{n+k}(n-k)^{n-k}} \cdot \frac{(2\pi(2n))^{1/2}}{(2\pi(n+k))^{1/2}(2\pi(n-k))^{1/2}} \end{aligned}$$

and we have

$$\binom{2n}{n+k} 2^{-2n} \sim \big(1+\frac{k}{n}\big)^{-n-k} \cdot \big(1-\frac{k}{n}\big)^{-n+k} \cdot (\pi n)^{-1/2} \cdot \big(1+\frac{k}{n}\big)^{-1/2} \cdot \big(1-\frac{k}{n}\big)^{k} \tag{3.2}$$

The first two terms on the right are

$$= \big(1-\frac{k^2}{n^2}\big)^{-n} \cdot \big(1+\frac{k}{n}\big)^{-k} \cdot \big(1-\frac{k}{n}\big)^{k}$$

A little calculus shows that:

Lemma 3.1.1. *If $c_j \to 0$, $a_j \to \infty$ and $a_j e_j \to \lambda$ then $(1 + c_j)^{a_j} \to e^{\lambda}$.*

Proof. As $x \to 0$, $\log(1 + x)/x \to 1$, so $a_j \log(1 + c_j) \to \lambda$ and the desired result follows.

□

Theorem 3.1.2. *If $2k/\sqrt{2n} \to x$ then $P(S_{2n} = 2k) \sim (\pi n)^{-1/2} e^{-x^2/2}$.*

Our next step is to compute

$$P(a\sqrt{2n} \le S_{2n} \le b\sqrt{2n}) = \sum_{m \in [a\sqrt{2n}, b\sqrt{2n}] \cap 2\mathbb{Z}} P(S_{2n} = m)$$

Changing variables $m = x\sqrt{2n}$, we have that the above is

$$\approx \sum_{x \in [a,b] \cap (2\mathbb{Z}/\sqrt{2n}} (2\pi)^{-1/2} e^{-x^2/2} \cdot (2/n)^{1/2}$$

where $2\mathbb{Z}/\sqrt{2n} = \{2z/\sqrt{2n} : z \in \mathbb{Z}\}$. We have multiplied and divided by $\sqrt{2}$ since the space between points in the sum is $(2/n)^{1/2}$, so if n is large the sum above is

$$\approx \int_a^b (2\pi)^{-1/2} e^{-x^2/2} dx$$

The integrand is the density of the (standard) normal distribution, so changing notation we can write the last quantity as $P(a \leq \mathcal{X} \leq b)$ where $\mathcal{X}$ is a random variable with the distribution.

It is not hard to fill in the details to get:

Theorem 3.1.3. ***The De Moirvre-Laplace Theorem.*** *If $a < b$ then as $m \to \infty$*

$$P(a \leq S_m/\sqrt{m} \leq b) \to \int_a^b (2\pi)^{-1/2} e^{-x^2/2} dx$$

(To remove the restriction to even integers observe $S_{2n+1} = S_{2n} \pm 1$). The last result is a special case of the central limit theorem.

3.2 Weak Convergence

In this section, we will define the type of convergence that appears in the central limit theorem and explore some of its properties. A sequence of distribution functions is said to **converge weakly** to a limit F (written $F_n \Rightarrow F0$ if $F_n(y) \to F(y)$ for all y that are continuity points of F. A sequence of random variables X_n is said to **converge weakly** or **converge in distribution** to a limit X_∞ (written $X_n \Rightarrow X_\infty$ if their distribution functions $F_n(x) = P(X_n \leq x)$ converge weakly. To see that convergence at continuity points to identify the limit, observe that F is right continuous, the discontinuities of F are at most a countable set.

Two examples of weak convergence are:

Example 3.2.1. Let $x_1, X_2, ...$ be i.i.d. with $P(X_i = 1) = P(X_i = -1) = 1/2$ and let $S_n = X_1 + \cdots + X_n$. Then Theorem 3.1.3 implies

$$F_n(y) = P(S_n/\sqrt{n} \leq y) \to \int_{-\infty}^y (2\pi)^{-1/2} e^{-x^2/2} dx$$

Example 3.2.2. Let $X_1, X_2, ...$ be i.i.d. with distribution F. The Glivenko-Cantelli theorem (Theorem 2.4.9) implies that for almost every ω,

$$F_n(y) = n^{-1} \sum_{m=1}^n 1_{(X_m(\omega) \leq y)} \to F(y) \text{ for all } y$$

In the last two examples convergence occurred for all y, even though in the second case the distribution function could have discontinuities. The next example shows why we restrict our attention to continuity points.

Example 3.2.3. Let X have distribution F. Then $X + 1/n$ has distribution

$$F_n(x) = P(X + 1/n \leq x) = F(x - 1/n)$$

As $n \to \infty$, $F_n(x) \to F(x-) = \lim_{y \uparrow x} F(y)$ so convergence only occurs at continuity points.

Example 3.2.4. Waiting for rare events. Let X_p be the number of trials needed to get a success ina sequence of independent trials with success probability p. Then $P(X_p \geq n) = (1-p)^{n-1}$ for $n = 1, 2, 3, ...$ and it follows from Lemma 3.1.1 that as $p \to 0$,

$$P(pX_p > x) \to e^{-x} \text{ for all } x \geq 0$$

In words, pX_p converges weakly to an exponential distribution.

Example 3.2.5. Birthday problem. Let $X_1, X_2, ...$ be independent and uniformly distributed on $\{1, ..., N\}$, and let $T_N = \min\{n : X_n = X_m \text{ for some } m < n\}$.

$$P(T_N > n) = \prod_{m=2}^{n} \left(1 - \frac{m-1}{N}\right)$$

When $N = 365$ this is the probability that two people in a group of size n do not have the same birthday (assuming all birthdays are equally likely). It is easy to see that

$$P(T_N/N^{1/2} > x) \to \exp(-x^2/2) \text{ for all } x \geq 0$$

Taking $N = 365$ and noting $22/\sqrt{365} = 1.1515$ and $(1.1515)^2/2 = 0.6630$, this says that

$$P(T_{365} > 22) \approx e^{-0.6630} \approx 0.5151$$

This answer is 2% smaller than the true probability 0.524.

Before giving out sixth example, we need a simple result called **Scheffè's Theorem**. Suppose we have probability densities f_n, $1 \leq n \leq \infty$, and $f_n \to f_\infty$ pointwise as $n \to \infty$. Then for all Borel sets B

$$\begin{aligned} \left| \int_B f_n(x)dx - \int_B f_\infty(x)dx \right| &\leq \int |f_n(x) - f_\infty(x)|dx \\ &= 2\int (f_\infty(x) - f_n(x))^+ dx \to 0 \end{aligned}$$

by the dominated convergence theorem, the equality following from the fact that the $f_n \geq 0$ and have integral = 1. Writing μ_n for the corresponding measures, we have shown that the **total variation norm**

$$||\mu_n - \mu_\infty|| \equiv \sup_B |\mu_n(B) - \mu_\infty(B)| \to 0$$

a conclusion stronger than weak convergence. (Take $B = (-\infty, x]$.) The example $\mu_n = a$ point mass at $1/n$ (with $1/\infty = 0$) shows that we may have $\mu_n \Rightarrow \mu_\infty$ with $||\mu_n - \mu_\infty|| = 1$ for all n.

Lemma 3.2.6. *V_{n+1} has density function*

$$fV_{n+1}(x) = (2n+1)\binom{2n}{n}x^n(1-x)^n$$

Proof. There are $2n+1$ ways to pick the observation that falls at x, then we have to pick n indices for observations $< x$, which can be done in $\binom{2n}{n}$ ways. Once we have decided on the indices that will land $< x$ and $> x$¡ the probability the corresponding random variables will do what we want is $x^n(1-x)^n$, and the probability density that the remaining one will land at x is 1. If you don't like the previous sentence compute the probability $X_1 < x - \epsilon, ..., X_n < x - \epsilon, x - \epsilon < X_{n+1} < x + \epsilon$, $X_{n+2} > x + \epsilon, ..., X_{2n+1} > x + \epsilon$ then let $\epsilon \to 0$.

□

To compute the density function of $Y_n = 2(V_{n+1} - 1/2)\sqrt{2n}$, simply change variables $x = 1/2 + y/2\sqrt{2n}$, $dx = dy/2\sqrt{2n}$ to get

$$\begin{aligned} f_{Y_n}(y) &= (2n+1)\binom{2n}{n}\left(\frac{1}{2}+\frac{y}{2\sqrt{2n}}\right)^n\left(\frac{1}{2}-\frac{y}{2\sqrt{2n}}\right)^n\frac{1}{2\sqrt{2n}} \\ &= \binom{2n}{n}2^{-2n}\cdot(1-y^2/2n)^n\cdot\frac{2n+1}{2n}\cdot\sqrt{\frac{n}{2}} \end{aligned}$$

The first factor is $P(S_{2n} = 0)$ for a simple random walk so Theorem 3.1.2 imply that

$$f_{Y_n}(y) \to (2\pi)^{-1/2}\exp(-y^2/2) \text{ as } n \to \infty$$

Here and in what follows we write $P(Y_n = y)$ for the density function of Y_n. Using Scheffe's theorem now, we conclude that Y_n converges weakly to a standard normal distribution.

The next result is useful for proving things about weak convergence.

Theorem 3.2.7. *If $F_n \Rightarrow F_\infty$ then there are random variables Y_n, $1 \le n \le \infty$, with distribution F_n so that $Y_n \to Y_\infty$ a.s.*

Proof. Let $\Omega = (0,1)$, $\mathcal{F} =$ Borel sets, $P =$ Lebesgue measure, and let $Y_n(x) = \sup\{y : F_n(y) < x\}$. By Theorem 1.2.2, Y_n has distribution F_n. we will now show that $Y_n(x) \to Y_\infty(x)$ for all but a countable number of x. To do this, it is convenient to write $Y_n(x)$ as $F_n^{-1}(x)$ and drop the subscript when $n = \infty$. We begin by identifying the exceptional set. Let $a_x = \sup\{y : F(y) < x\}$, $b_x = \inf\{y : F(y) > x\}$, and $\Omega_0 = \{x : (a_x, b_x) = \emptyset\}$ where (a_x, b_x) is the open interval with the indicated endpoints. $\Omega - \Omega_)$ is countable since the (a_x, b_x) are disjoint and each nonempty interval contains a different rational number. If $x \in \Omega_0$ then $F(y) < x$ for $y < F^{-1}(x)$ and $F(z) > x$ for $z > F^{-1}(x)$. To prove that $F_n^{-1}(x) \to F^{-1}(x)$ for $x \in \Omega_0$, there are two things to show:

(a) $\liminf_{n\to\infty} F_n^{-1}(x) \ge F^{-1}(x)$

Proof of (a). Let $y < F^{-1}(x)$ be such that F is continuous at y. Since $x \in \Omega_0$, $F(y) < x$ and if n is sufficiently large $F_n(y) < x$, i.e., $F_n^{-1}(x) \ge y$. Since this holds for all y satisfying the indicated restrictions, the result follows.

(b) $\limsup_{n\to\infty} F_n^{-1}(x) \le F^{-1}(x)$

Proof of (b). Let $y > F^{-1}(x)$ be such that F is continuous at y. Since $x \in \Omega_0$, $F(y) > x$ and if n is sufficiently large $F_n(y) > x$, i.e., $F_n^{-1}(x) \le y$. Since this holds for all y satisfying the indicated restrictions, the result follows and we have completed the proof.

□

Theorem 3.2.8. *$X_n \Rightarrow X_\infty$ if and only if for every bounded continuous function g we have $Eg(X_n) \to Eg(X_\infty)$.*

Proof. Let Y_n have the same distribution as X_n and converge a.s. Since g is continuous $g(Y_n) \to g(Y_\infty)$ a.s. and the bounded convergence theorem implies

$$Eg(X_n) = Eg(Y_n) \to Eg(Y_\infty) = Eg(X_\infty)$$

To prove the converse let

$$g_{x,\epsilon}(y) = \begin{cases} 1 & y \le x \\ 0 & y \ge x+\epsilon \\ \text{linear} & x \le y \le x+\epsilon \end{cases}$$

Since $g_{x,\epsilon}(y) = 1$ for $y \leq x$, $g_{x,\epsilon}$ is continuous, and $g_{x,\epsilon}(y) = 0$ for $y > x + \epsilon$,

$$\limsup_{n\to\infty} P(X_n \leq x) \leq \limsup_{n\to\infty} Eg_{x,\epsilon}(X_n) = Eg_{x,\epsilon}(X_\infty) \leq P(X_\infty \leq x + \epsilon)$$

Letting $\epsilon \to 0$ gives $\limsup_{n\to\infty} P(X_n \leq x) \leq P(X_\infty \leq x)$. The last conclusion is valid for any x. To get the other direction, we observe

$$\limsup_{n\to\infty} P(X_n \leq x) \geq \limsup_{n\to\infty} Eg_{x,\epsilon}(X_n) = Eg_{x,\epsilon}(X_\infty) \leq P(X_\infty \geq x + \epsilon)$$

Letting $\epsilon \to 0$ gives $\liminf_{n\to\infty} P(X_n \leq x) \geq P(X_\infty < x) = P(X_\infty \leq x)$ if x is a continuity point. The results for the lim sup and the lim inf combine to give the desired result.

□

Theorem 3.2.9. ***Continuous mapping theorem***. *Let g be a measurable function and $D_g = \{x : g \text{ is discontinuous at } x\}$. If $X_n \Rightarrow X_\infty$ and $P(X_\infty \in D_g) = 0$ then $g(X_n) \Rightarrow g(X)$. If in addition g is bounded then $Eg(X_n) \to Eg(X_\infty)$.*

Remark 3.2.10. D_g is always a Borel set.

Proof. Let $Y_n =_d X_n$ with $Y_n \to Y_\infty$ a.s. If f is continuous then $D_{f\circ g} \subset D_g$ so $P(Y_\infty \in D_{f\circ g}) = 0$ and it follows that $f(g(Y_n)) \to f(g(Y_\infty)$ a.s. If, in addition, f is bounded then the bounded convergence theorem implies $Ef(g(Y_n)) \to Ef(g(Y_\infty))$. Since this holds for all bounded continuous functions, it follows from Theorem 3.2.8 that $g(X_n) \Rightarrow g(X_\infty)$.

The second conclusion is easier. Since $P(Y_\infty \in D_g) = 0$, $g(Y_n) \to g(Y_\infty)$ a.s., and the desired result follows from the bounded convergence theorem.

□

The next result provides a number of useful alternative definitions of weak convergence.

Theorem 3.2.11. *The following statements are equivalent:*

(i) $X_n \Rightarrow X_\infty$

(ii) For all open sets G, $\liminf_{n\to\infty} P(X_n \in G) \geq P(X_\infty \in G)$.

(iii) For all closed sets K, $\limsup_{n\to\infty} P(X_n \in K) \leq P(X_\infty \in K)$.

(iv) For all sets A with $P(X_\infty \in \partial A) = 0$, $\lim_{n\to\infty} P(X_n \in A) = P(X_\infty \in A)$.

Remark 3.2.12. To help remember the directions of the inequalities in (ii) and (iii), consider the special case in which $P(X_n = x_n) = 1$. In this case, if $x_n \in G$ and $x_n \to x_\infty \in \partial G$, then $P(X_n \in G) = 1$ for all n but $P(X_\infty \in G) = 0$. Letting $K = G^c$ gives an example for(iii).

Proof. We will prove four things and leave it to the reader to check that we have proved the result given above.

(i) implies (ii): Let Y_n have the same distribution as X_n and $Y_n \to Y_\infty$ a.s. Since G is open

$$\liminf_{n\to\infty} 1_G(Y_n) \geq 1_G(Y_\infty)$$

so Fatou's Lemma implies

$$\liminf_{n\to\infty} P(Y_n \in G) \geq P(Y_\infty \in G)$$

(ii) is equivalent to (iii): This follows easily from: A is open if and only if A^c is closed and $P(A) + P(A^c) = 1$.

(ii) and (iii) imply (iv): Let $K = \bar{A}$ and $G = A^0$ be the closure and interior of A respectively. The boundary of A, $\partial A = \bar{A} - A^0$ and $P(X_\infty \in \partial A) = 0$ so

$$P(X_\infty \in K) = P(X_\infty \in A) = P(X_\infty \in G)$$

Using (ii) and (iii) now

$$\limsup_{n\to\infty} P(X_n \in A) \leq \limsup_{n\to\infty} P(X_n \in K) \leq P(X_\infty \in K) = P(X_\infty \in A)$$

$$\liminf_{n\to\infty} P(X_n \in A) \geq \liminf_{n\to\infty} P(X_n \in G) \geq P(X_\infty \in G) = P(X_\infty \in A)$$

(iv) implies (i): Let x be such that $P(X_\infty = x) = 0$, i.e., x is a continuity point of F, and let $A = (-\infty, x]$.

□

The next result is useful in studying limits of sequences of distributions.

Theorem 3.2.13. ***Helly's selection theorem.*** *For every sequence F_n of distribution functions, there is a subsequence $F_{n(k)}$ and a right continuous nondecreasing function F so that $\lim_{k\to\infty} F_{n(k)}(y) = F(y)$ at all continuity points y of F.*

Remark 3.2.14. The limit may not be a distribution function. For example if $a+b+c = 1$ and $F_n(x) = a1_{(x\geq n)} + b1_{(x\geq -n)} + cG(x)$ where G is a distribution function, then $F_n(x) \to F(x) = b + cG(x)$,

$$\lim_{x\downarrow -\infty} F(x) = b \text{ and } \lim_{x\uparrow\infty} F(x) = b + c = 1 - a$$

In words, an amount of mass a escapes to $+\infty$, and mass b escapes to $-\infty$. The type of convergence that occurs in Theorem 3.2.13 is sometimes called **vague convergence**, and will be denoted here by $\Rightarrow_v$.

Proof. The first step is a diagonal argument. Let $q_1, q_2, ...$ be an enumeration of the rationals. Since for each k, $F_m(q_k) \in [0,1]$ for all m,there is a sequence $m_k(i) \to \infty$ that is a subsequence of $m_{k-1}(j)$ (let $m_0(j) =\equiv j$) so that

$$F_{m_k(i)}(q_k) \text{ converges to } G(q_k) \text{ as } i \to \infty$$

Let $F_{n(k)} = F_{m_k(k)}$. By construction $F_{n(k)}(q) \to G(q)$ for all rational q. The function G may not be right continuous but $F(x) = \inf\{G(q) : q \in \mathbb{Q}, q > x\}$ is since

$$\begin{aligned} \lim_{x_n\downarrow x} F(x_n) &= \inf\{G(q) : q \in \mathbb{Q}, q > x_n \text{ for some } n\} \\ &= \inf\{G(q) : q \in \mathbb{Q}, q > x\} = F(x) \end{aligned}$$

To complete the proof, let x be a continuity point of F. Pick rationals r_1, r_2, s with $f_1 < r_2 < x < s$ so that

$$F(x) - \epsilon < F(r_1) \leq F(r_2) \leq F(x) \leq F(s) < F(x) + \epsilon$$

Since $F_{n(k)}(r_2) \to G(r_2) \geq F(r_1)$, and $F_{n(k)}(s) \to G(s) \leq F(s)$ it follows that if k is large

$$F(x) - \epsilon < F_{n(k)}(r_2) \leq F_{n(k)}(x) \leq F_{n(k)}(s) < F(x) + \epsilon$$

which is the desired conclusion.

□

The last result raises a question: When can we conclude that no mass is lost in the limit in Theorem 3.2.13?

Theorem 3.2.15. *Every subsequential limit is the distribution function of a probability measure if and only if the sequence F_n is **tight**, i.e., for all $\epsilon > 0$ there is an M_ϵ so that*

$$\limsup_{n\to\infty} 1 - F_n(M_\epsilon) + F_n(-M_\epsilon) \leq \epsilon$$

Proof. Suppose the sequence is tight and $F_{n(k)} \Rightarrow_v F$. Let $r < -M_\epsilon$ and $s > M_\epsilon$ be continuity points of F. Since $F_n(r) \to F(r)$ and $F_n(s) \to F(s)$, we have

$$\begin{aligned} 1 - F(s) + F(r) &= \lim_{k\to\infty} 1 - F_{n(k)}(s) + F_{n(k)}(r) \\ &\leq \limsup_{n\to\infty} 1 - F_n(M_\epsilon) + F_n(-M_\epsilon) \leq \epsilon \end{aligned}$$

The last result implies $\limsup_{x\to\infty} 1 - F(x) + F(-x) \leq \epsilon$. Since ϵ is arbitrary it follows that F is the distribution function of a probability measure.

To prove the converse now suppose F_n is not tight. In this case, there is an $\epsilon > 0$ and a subsequence $n(k) \to \infty$ so that

$$1 - F_{n(k)}(k) + F_{n(k)}(-k) \geq \epsilon$$

for all k. By passing to a further subsequence $F_{n(k_j)}$ we can suppose that $F_{n(k_j)} \Rightarrow_v F$. Let $r < 0 < s$ be continuity points of F.

$$\begin{aligned} 1 - F(s) + F(r) &= \lim_{j\to\infty} 1 - F_{n(k_j)}(s) + F_{n(k_j)}(r) \\ &\geq \liminf_{j\to\infty} 1 - F_{n(k_j)}(k_j) + F_{n(k_j)}(-k_j) \geq \epsilon \end{aligned}$$

Letting $s \to \infty$ and $r \to -\infty$, we see that F is not the distribution function of a probability measure.

□

The following sufficient condition for rightness is often useful.

Theorem 3.2.16. *If there is a $\varphi \geq 0$ so that $\varphi(x) \to \infty$ as $|x| \to \infty$ and*

$$C = \sup_n \int \varphi(x) dF_n(x) < \infty$$

then F_n is tight.

Proof. $1 - F_n(M) + F_n(-M) \leq C/\inf_{|x|\geq M} \varphi(x)$.

□

Theorem 3.2.17. *If each subsequence of X_n has a further subsequence that converges to X then $X_n \Rightarrow X$.*

3.3 Characteristic Functions

This long section has five small parts. The first three are required reading, the last two are optional. In the first part, we show that the characteristic function $\varphi(t) = E\exp(itX)$ determines $F(x) = P(X \leq x)$, and we give recipes for computing F from φ. In the second part, we relate weak convergence of distributions to the behavior of the corresponding characteristic functions. In the third part, we relate the behavior of $\varphi(t)$ at 0 to the moments of X. In the fourth part, we prove Polya's criterion and use it to construct some famous and some strange examples of characteristic functions. Finally, in the fifth part, we consider the moment problem, i.e., when is a distribution characterized by its moments.

3.3.1 Definition, Inversion Formula

If X is a random variable we define its **characteristic function (ch.f.)** by

$$\varphi(t) = Ee^{itX} = E\cos tX + iE\sin tX$$

The last formula requires taking the expected value of a complex valued random variable but as the second equality may suggest no new theory is required. If Z is complex valued we define $EZ = E(\mathrm{Re}Z) + iE(\mathrm{Im}Z)$ where $\mathrm{Re}(a+bi) = a$ is the **real part** and $\mathrm{Im}(a+bi) = b$ is the **imaginary part**. Some other definitions we will need are: the **modulus** of the complex number $z = a+bi$ is $|a+bi| = (a^2+b^2)^{1/2}$, and the **complex conjugate** of $z = a+bi$, $\bar{z} = a-bi$.

Theorem 3.3.1. *All characteristic functions have the following properties:*
(a) $\varphi(0) = 1$,
(b) $\varphi(-t) = \overline{\varphi(t)}$,
(c) $|\varphi(t)| = |Ee^{itX}| \le E|e^{itX}| = 1$
(d) $|\varphi(t+h) - \varphi(t)| \le E|e^{thX} - 1|$, *so* $\varphi(t)$ *is uniformly continuous on* $(-\infty, \infty)$.
(e) $Ee^{it(aX+b)} = e^{itb}\varphi(at)$

Proof. (a) is obvious. For (b) we note that

$$\varphi(-t) = E(\cos(-tX) + i\sin(-tX)) = E(\cos(tX) - i\sin(tX))$$

(c) follows since $\varphi(x,y) = (x^2+y^2)^{1/2}$ is convex.

$$\begin{aligned} |\varphi(t+h) - \varphi(h)| &= |E(e^{i(t+h)X} - e^{itX})| \\ &\le E|e^{i(t+h)X} - e^{itX}| \\ &= E|e^{ihX} - 1| \end{aligned}$$

so uniform convergence follows from the bounded convergence theorem. For (e) we note $Ee^{it(aX+b)} = e^{itb}Ee^{i(ta)X} = e^{itb}\varphi(at)$.

The main reason for introducing characteristic functions is the following:

Theorem 3.3.2. *If* X_1 *and* X_2 *are independent and have ch.f.'s* φ_1 *and* φ_2 *then* $X_1 + X_2$ *has ch.f.* $\varphi_1(t)\varphi_2(t)$.

Proof.

$$Ee^{it(X_1+X_2)} = E(e^{itX_1}e^{itX_2}) = Ee^{itX_1}Ee^{itX_2}$$

since e^{itX_1} and e^{itX_2} are independent.

□

Example 3.3.3. Coin flips. If $P(X=1) = P(X=-1) = 1/2$ then

$$Ee^{itX} = (e^{it} + e^{-it})/2 = \cos t$$

Example 3.3.4. Poisson distribution. If $P(X=k) = e^{-\lambda}\lambda^k/k!$ for $k = 0, 1, 2, ...$

$$Ee^{itX} = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k e^{itk}}{k!} = \exp(\lambda(e^{it}-1))$$

Example 3.3.5. Normal distribution

$$\text{Density } (2\pi)^{-1/2}\exp(-x^2/2)$$

$$\text{Ch.F. } \exp(-t^2/2)$$

Combining this result with (e) of Theorem 3.2.9, we see that a normal distribution with mean μ and variance σ^2 has ch.f. $\exp(i\mu t - \sigma^2 t^2/2)$. Similar scalings can be

applied to other examples so we will often just give the ch.f. for one member of the family.

Physics Proof.

$$\int e^{itx}(2\pi)^{-1/2}e^{-x^2/2}dx = e^{-t^2/2}\int (2\pi)^{-1/2}e^{-(x-it)^2/2}dx$$

Math Proof. Now that we have cheated and figured out the answer we can verify it by a formal calculation that gives very little insight into why it is true. Let

$$\varphi(t) = \int e^{itx}(2\pi)^{-1/2}e^{-x^2/2}dx = \int \cos tx(2\pi)^{-1/2}e^{-x^2/2}dx$$

since $i\sin tx$ is an odd function. Differentiating with respect to t and then integrating by parts gives

$$\begin{aligned}\varphi'(t) &= \int -x\sin tx(2\pi)^{-1/2}e^{-x^2/2}dx \\ &= -\int t\cos tx(2\pi)^{-1/2}e^{-x^2/2}dx \\ &= -t\varphi(t)\end{aligned}$$

This implies $\frac{d}{dt}\{\varphi(t)\exp(t^2/2)\} = 0$ so $\varphi(t)\exp(t^2/2) = \varphi(0) = 1$.

In the next three examples, the density is 0 outside the indicated range.

Example 3.3.6. Uniform distribution on (a,b).

$$\text{Density } 1/(b-a)\ x \in (a,b)$$

$$\text{Ch.f. } (e^{itb} - e^{ita})/it(b-a)$$

In the special case $a = -c$, $b = c$ the ch.f. is $(e^{ite} - e^{-ite})/2cit = (\sin ct)/ct$.

Proof. Once you recall that $\int_a^b e^{\lambda x}dx = (e^{\lambda b} - e^{\lambda a})/\lambda$ holds for complex λ, this is immediate.

□

Example 3.3.7. Triangular distribution

$$\text{Density } 1 - |x|\ x \in (-1,1)$$

$$\text{Ch.f. } 2(1-\cos t)/t^2$$

Proof. To see this, notice that if X and Y are independent and uniform $(-1/2, 1/2)$ then $X + Y$ has a triangular distribution, Using previous example and Theorem 3.3.2, it follows that the desired ch.f. is

$$\{(e^{it/2} - e^{-it/2})/it\}^2 = \{2\sin(t/2)/t\}^2$$

Using the trig identity $\cos 2\theta = 1 - 2\sin^2\theta$ with $\theta = t/2$ converts the answer into the form given above.

□

Example 3.3.8. Exponential distribution

$$\text{Density } e^{-x}\ x \in (0,\infty)$$

$$\text{Ch.f. } 1/(1-it)$$

Proof. Integrating gives

$$\int_0^\infty e^{itx}e^{-x}dx = \frac{e^{(it-1)x}}{it-1}\Big|_0^\infty = \frac{1}{1-it}$$

since $\exp((it-1)x) \to 0$ as $x \to \infty$.

□

For the next result we need the following fact which follows from the fact that $\int f d(\mu+\nu) = \int f d\mu + \int f d\nu$.

Lemma 3.3.9. *If $F_1, ..., F_n$ have ch.f. $\varphi_1, ..., \varphi_n$ and $\lambda_i \geq 0$ have $\lambda_1 + \cdots + \lambda_n = 1$ then $\sum_{i=1}^n \lambda_i F_i$ has ch.f. $\sum_{i=1}^n \lambda_i \varphi_i$.*

Example 3.3.10. Bilateral exponential

$$\text{Density } \frac{1}{2}e^{-|x|} \; x \in (-\infty, \infty)$$

$$\text{Ch.f. } 1/(1+t^2)$$

Proof. This follows from Lemma 3.3.9 with F_1 the distribution of an exponential random variable X, F_2 the distribution of $-X$, and $\lambda_1 = \lambda_2 = 1/2$ then using (b) of Theorem 3.3.1 we see the desired ch.f. is

$$\frac{1}{2(1-it)} + \frac{1}{2(1+it)} = \frac{(1+it)+(1-it)}{2(1+t^2)} = \frac{1}{(1+t^2)}$$

Theorem 3.3.11. ***The inversion formula***. *Let $\varphi(t) = \int e^{itx}\mu(dx)$ where μ is a probability measure. If $a < b$ then*

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

Remark 3.3.12. The existence of the limit is part of the conclusion. If $\mu = \delta_0$, a point mass at 0, $\varphi(t) \equiv 1$. In this case, if $a = -1$ and $b = 1$, the integrand is $(2\sin t)/t$ and the integral does not converge absolutely.

Proof. Let

$$I_T = \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\varphi(t)dt = \int_{-T}^{T}\int \frac{e^{-ita} - e^{-itb}}{it} e^{-ita}\mu(dx)dt$$

The integrand may look bad near $t = 0$ but if we observe that

$$\frac{e^{-ita} - e^{-tb}}{it} = \int_a^b e^{-ity}dy$$

we see that the modulus of the integrand is bounded by $b - a$. Since μ is a probability measure and $[-T, T]$ is a finite interval it follows from Fubini's theorem, $\cos(-x) = \cos x$, and $\sin(-x) = -\sin x$ that

$$\begin{aligned} I_t &= \int\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} e^{-itx} dt\mu(dx) \\ &= \int \left\{ \int_{-T}^{T} \frac{\sin(t(x-a))}{t}dt - \int_{-T}^{T} \frac{\sin(t(x-b))}{t}dt \right\} \mu(dx) \end{aligned}$$

Introducing $R(\theta, T) = \int_{-T}^{T} (\sin\theta t)/t dt$, we can write the last result as

$$(\star)\; I_T = \int \{R(x-a, T) - R(x-b, T)\}\mu(dx)$$

If we let $S(T) = \int_0^T (\sin x)/x dx$ then for $\theta > 0$ changing variables $t = x/\theta$ shows that

$$R(\theta, T) = 2\int_0^{T\theta} \frac{\sin x}{x}dx = 2S(T\theta)$$

while for $\theta < 0$, $R(\theta, T) = -R(|\theta|, T)$. Introducing the function sgn x, which is 1 if $x > 0$, -1 is $x < 0$, we can write the last two formulas together as

$$R(\theta, T) = 2(\text{sgn } \theta)S(T|\theta|)$$

As $T \to \infty$, $S(T) \to \pi/2$, so we have $R(\theta, T) \to \pi \text{sgn } \theta$ and

$$R(x-a, T) - R(x-b, T) \to \begin{cases} 2\pi & a < x < b \\ \pi & x = a \text{ or } x = b \\ 0 & x < a \text{ or } x > b \end{cases}$$

$|R(\theta, T)| \leq 2 \sup_y S(y) < \infty$, so using the bounded convergence theorem with $(\star)$ implies

$$(2\pi)^{-1} I_T \to \mu(a, b) + \frac{1}{2}\mu(\{a, b\})$$

proving the desired result.

□

Theorem 3.3.13. *If $\int \varphi(t)|dt < \infty$ then μ has bounded continuous density*

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

Proof. As we observed in the proof of Theorem 3.3.11

$$\left| \frac{e^{-ita} - e^{-itb}}{it} \right| = \left| \int_a^b e^{-ity} dy \right| \leq |b - a|$$

so the integral in Theorem 3.3.11 converges absolutely in this case and

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

The last result implies μ has no point masses and

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

by Fubini's theorem, so the distribution μ has density function

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

The dominated convergence theorem implies f is continuous and the proof is complete.

□

Example 3.3.14. Polya's distribution

$$\text{Density } (1 - \cos x)/\pi x^2$$

$$\text{Ch.f. } (1 - |t|)^+$$

Proof. Theorem 3.3.13 implies

$$\frac{1}{2\pi} \int \frac{2(1 - \cos s)}{s^2} e^{-isy} ds = (1 - |y|)^+$$

Now let $s = x$, $y = -t$.

□

Example 3.3.15. The Cauchy distribution

$$\text{Density } 1/\pi(1+x^2)$$

$$\text{Ch.f. } \exp(-|t|)$$

Proof. Theorem 3.3.13 implies

$$\frac{1}{2\pi}\int \frac{1}{1+s^2}e^{-isy}ds = \frac{1}{2}e^{|-y|}$$

Now let $s = x$, $y = -t$ and multiply each side by 2.

□

3.3.2 Weak Convergence

Our next step toward the central limit theorem is to relate convergence of characteristic functions to weak convergence.

Theorem 3.3.16. ***Continuity theorem***. *Let μ_n, $1 \le n \le \infty$ be probability measures with ch.f. φ_n. (i) If $\mu_n \Rightarrow \mu_\infty$ then $\varphi_n(t) \to \varphi_\infty(t)$ for all t. (ii) If $\varphi_n(t)$ converges point-wise to a limit $\varphi(t)$ that is continuous at 0, then the associated sequence of distributions μ_n is tight and converges weakly to the measure μ with characteristic function φ.*

Remark 3.3.17. To see why continuity of the limit at 0 is needed in (ii), let μ_n have a normal distribution with mean 0 and variance n. In this case $\varphi_n(t) = \exp(-nt^2/2) \to 0$ for $t \neq 0$, and $\varphi_n(0) = 1$ for all n, but the measures do not converge weakly since $\mu_n((-\infty, x]) \to 1/2$ for all x.

Proof. (i) is easy. e^{itx} is bounded and continuous so if $\mu_n \Rightarrow \mu_\infty$ then Theorem 3.2.8 implies $\varphi_n(t) \to \varphi_\infty(t)$. To prove (ii), our first goal is to prove tightness. We begin with some calculations that may look mysterious but will prove to be very useful.

$$\int_{-u}^{u} 1 - e^{itx}dt = 2u - \int_{-u}^{u}(\cos tx + i\sin tx)dt = 2u - \frac{2\sin ux}{x}$$

Dividing both sides by u, integrating $\mu_n(dx)$, and using Fubini's theorem on the left-hand side gives

$$u^{-1}\int_{-u}^{u}(1-\varphi_n(t))dt = 2\int\left(1 - \frac{\sin ux}{ux}\right)\mu_n(dx)$$

To bound the right-hand side, we note that

$$|\sin x| = \left|\int_0^x \cos(y)dy\right| \le |x| \text{ for all x}$$

so we have $1 - (\sin ux/ux) \ge 0$. Discarding the integral over $(-2/u, 2/u)$ and using $|\sin ux| \le 1$ on the rest, the right-hand side is

$$\ge 2\int_{|x|\ge 2/u}\left(1 - \frac{1}{|ux|}\right)\mu_n(dx) \ge \mu_n(\{x : |x| > 2/u\})$$

Since $\varphi(t) \to 1$ as $t \to 0$,

$$u^{-1}\int_{-u}^{u}(1-\varphi(t))dt \to 0 \text{ as } u \to 0$$

Pick u so that the integral is $< \epsilon$. Since $\varphi_n(t) \to \varphi(t)$ for each g, it follows from the bounded convergence theorem that for $n \geq N$

$$2\epsilon \geq u^{-1} \int_{-u}^{u} (1 - \varphi_n(t))dt \geq \mu_n\{x : |x| > 2/u\}$$

Since ϵ is arbitrary, the sequence μ_n is tight.

To complete the proof now we observe that if $\mu_{n(k)} \Rightarrow \mu$, then it follows from the first sentence of the proof that μ has ch.f. φ. The last observation and tightness imply that every subsequence has a further subsequence that converges to μ. I claim that this implies the whole sequence converges to μ. To see this, observe that we have shown that if f is bounded and continuous then every subsequence of $\int f d\mu_n$ has a further subsequence that converges to $\int f d\mu$, so Theorem 2.3.2 implies that the whole sequence converges to that limit. This shows $\int f d\mu_n \to \int f d\mu$ for all bounded continuous functions f so the desired result follows from Theorem 3.2.8.

□

3.3.3 Moments and Derivatives

In the proof of Theorem 3.3.16, we derived the inequality

$$\mu\{x : |x| > 2\mu\} \leq \mu^{-1} \int_{-\mu}^{\mu} (1 - \varphi(t))dt \tag{3.3}$$

which shows that the smoothness of the characteristic function at 0 is related to the decay of the measure at ∞.

Lemma 3.3.18.

$$\left| e^{ix} - \sum_{m=0}^{n} \frac{(ix)^m}{m!} \right| \leq \min\left(\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right) \tag{3.4}$$

The first term on the right is the usual order of magnitude we expect in the correction term. The second is better for large $|x|$ and will help us prove the central limit theorem without assuming finite third moments.

Proof. Integrating by parts gives

$$\int_0^x (x-s)^n e^{is} ds = \frac{x^{n+1}}{n+1} + \frac{i}{n+1} \int_0^x (x-s)^{n+1} e^{is} ds$$

When $n = 0$, this says

$$\int_0^x e^{is} ds = x + i \int_0^x (x-s) e^{is} ds$$

The left-hand side is $(e^{ix} - 1)/i$, so rearranging gives

$$e^{ix} = 1 + ix + i^2 \int_0^x (x-s) e^{is} ds$$

Using the result for $n = 1$ now gives

$$e^{ix} = 1 + ix + \frac{i^2 x^2}{2} + \frac{i^3}{2} \int_0^x (x-s)^2 e^{is} ds$$

and iterating we arrive at

$$(a)\ e^{ix} - \sum_{m=0}^{n} \frac{(ix)^m}{m!} = \frac{i^{n+1}}{n!} \int_0^x (x-s)^n e^{is} ds$$

To prove the result now it only remains to estimate the "error term" on the right-hand side. Since $|e^{is}| \leq 1$ for all s,

$$(b) \quad \left| \frac{i^{n+1}}{n!} \int_0^x (x-s)^n e^{is} ds \right| \leq |x|^{n+1}/(n+1)!$$

The last estimate is good when x is small. The next is designed for large x. Integrating by parts

$$\frac{i}{n} \int_0^x (x-s)^n e^{is} ds = -\frac{x^n}{n} + \int_0^x (x-s)^{n-1} e^{is} ds$$

Noticing $x^n/n = \int_0^x (x-s)^{n-1} ds$ now gives

$$\frac{i^{n+1}}{n!} \int_0^x (x-s)^n e^{is} ds = \frac{i^n}{(n-1)!} \int_0^x (x-s)^{n-1} (e^{is} - 1) ds$$

and since $|e^{ix} - 1| \leq 2$, it follows that

$$(c) \quad \left| \frac{i^{n+1}}{n!} \int_0^x (x-s)^n e^{is} ds \right| \leq \left| \frac{2}{(n-1)!} \int_0^x (x-s)^{n-1} ds \right| \leq 2|x|^n/n!$$

Combining (a), (b), and (c) we have the desired result.

□

Taking expected values, using Jensen's inequality, applying Theorem 3.3.2 to $x = tX$, gives

$$\begin{aligned} \left| Ee^{itX} - \sum_{m=0}^{n} E\frac{(itX)^m}{m!} \right| &\leq E\left| e^{itX} - \sum_{m=0}^{n} \frac{(itX)^m}{m!} \right| \\ &\leq E \min(|tX|^{n+1}, 2|tX|^n) \end{aligned} \tag{3.5}$$

where in the second step we have dropped the denominators to make the bound simpler.

Theorem 3.3.19. *If $E|X|^2 < \infty$ then*

$$\varphi(t) = 1 + itEX - t^2 E(X^2)/2 + o(t^2)$$

Proof. The error term is $\leq t^2 E(|t| \cdot |X|^3 \wedge 2|X|^2)$. The variable in parentheses is smaller than $2|X|^2$ and converges to 0 as $t \to \infty$, so the desired conclusion follows from the dominated convergence theorem.

□

Remark 3.3.20. The point of the estimate in 3.5 which involves the minimum of two terms rather than just the first one which would result from a naive application of Taylor series, is that we get the conclusion in Theorem 3.3.19 under the assumption $E|X|^2 < \infty$, i.e., we do not have to assume $E|X|^3 < \infty$.

Theorem 3.3.21. *If $\limsup_{h \downarrow 0} \{\varphi(h) - 2\varphi(0) + \varphi(-h)\}/h^2 > -\infty$, then $E|X|^2 < \infty$.*

Proof. $(e^{ihx} - 2 + e^{-ihx})/h^2 = -2(1 - \cos hx)/h^2 \leq 0$ and $2(1 - \cos hx)/h^2 \to x^2$ as $h \to 0$ so Fatou's lemma and Fubini's theorem imply

$$\begin{aligned} \int x^2 dF(x) &\leq 2 \liminf_{h \to 0} \int \frac{1 - \cos hx}{h^2} dF(x) \\ &= -\limsup_{h \to 0} \frac{\varphi(h) - 2\varphi(0) + \varphi(-h)}{h^2} < \infty \end{aligned}$$

which proves the desired result.

3.3.4 Polya's Criterion

The next result is useful for constructing examples of ch.f.'s.

Theorem 3.3.22. ***Polya's criterion.*** *Let $\varphi(t)$ be real nonnegative and have $\varphi(0) = 1$, $\varphi(t) = \varphi(-t)$, and φ is decreasing and convex on $(0, x)$ with*

$$\lim_{t\downarrow 0}\varphi(t) = 1,\ \lim_{t\uparrow\infty}\varphi(t) = 0$$

Then there is a probability measure ν on $(0, \infty)$, so that

$$(\star)\ \varphi(t) = \int_0^\infty \left(1 - \left|\frac{t}{s}\right|\right)^+ \nu(ds)$$

and hence φ is a characteristic function.

Remark 3.3.23. Before we get lost in the details of the proof, the reader should note that $(\star)$ displays φ as a convex combination of ch.f.'s of the form, so an extension of Lemma 3.3.9 (to be proved below) implies that this is a ch.f.

The assumption that $\lim_{t\to 0}\varphi(t) = 1$ is necessary because the function $\varphi(t) = 1_{\{0\}}(t)$ which is 1 at 0 and 0 otherwise satisfies all the other hypotheses. We could allow $\lim_{t\to\infty}\varphi(t) = c > 0$ by having a point mass of size c at 0, but we leave this extension to the reader.

Proof. Let φ' be the right derivative of ϕ, i.e.,

$$\varphi'(t) = \lim_{h\downarrow 0}\frac{\varphi(t+h) - \varphi(t)}{h}$$

Since φ is convex this exists and is right continuous and increasing. So we can let μ be the measure on $(0, \infty)$ with $\mu(a, b] = \varphi'(b) - \varphi'(a)$ for all $0 \le a < b < \infty$, and let ν be the measure on $(0, \infty)$ with $d\nu/d\mu = s$.

Now $\varphi'(t) \to 0$ as $t \to \infty$ (for if $\varphi'(t) \downarrow -\epsilon$ we would have $\varphi(t) \le 1 - \epsilon t$ for all t, so we have

$$-\varphi'(s) = \int_s^\infty r^{-1}\nu(dr)$$

Integrating again and using Fubini's theorem we have for $t \ge 0$

$$\begin{aligned}\varphi(t) &= \int_t^\infty\int_s^\infty r^{-1}\nu(dr)ds = \int_t^\infty r^{-1}\int_t^r ds\nu(dr)\\ &= \int_t^\infty\left(1 - \frac{t}{r}\right)\nu(dr) = \int_0^\infty\left(1 - \frac{t}{r}\right)^+\nu(dr)\end{aligned}$$

Using $\varphi(-t) = \varphi(t)$ to extend the formula to $t \le 0$ we have $(\star)$. Setting $t = 0$ in $(\star)$ shows ν has total mass 1.

If φ is piecewise linear, ν has a finite number of atoms and the result follows from example and Lemma 3.3.9. To prove the general result, let ν_n be a sequence of measures on $(0, \infty)$ with a finite number of atoms that converges weakly to ν and let

$$\varphi_n(t) = \int_0^\infty \left(1 - \left|\frac{t}{s}\right|\right)^+ \nu_n(ds)$$

Since $s \to (1 - |t/s|)^+$ is bounded and continuous, $\varphi_n(t) \to \varphi(t)$ and the desired result follows from (ii) of Theorem 3.3.16.

□

Example 3.3.24. $\exp(-|t|^\alpha)$ is a characteristic function for $0 < \alpha < 2$.

Proof. A little calculus shows that for any β and $|x| < 1$

$$(1-x)^\beta = \sum_{n=0}^{\infty} \binom{\beta}{n} (-x)^n$$

where

$$\binom{\beta}{n} = \frac{\beta(\beta-1)\dots(\beta-n+1)}{1\cdot 2\dots n}$$

Let $\psi(t) = 1-(1-\cos t)^{\alpha/2} = \sum_{n=1}^{\infty} c_n(\cos t)^n$ where

$$c_n = \binom{\alpha/2}{n}(-1)^{n+1}$$

$c_n \geq 0$ (here we use $\alpha < 2$), and $\sum_{n=1}^{\infty} c_n = 1$ (take $t = 0$ in the definition of ψ). $\cos t$ is a characteristic function so an easy extension of Lemma 3.3.9 shows that ψ is a ch.f. We have $1-\cos t \sim t^2/2$ as $t \to 0$, so

$$1-\cos(t\cdot 2^{1/2}\cdot n^{-1/\alpha}) \sim n^{-2/\alpha}t^2$$

Using Lemma 3.1.1 and (ii) of Theorem 3.3.16 now, it follows that

$$\exp(-|t|^\alpha) = \lim_{n\to\infty}\{\psi(t\cdot 2^{1/2}\cdot n^{-1/\alpha})\}^n$$

is a ch.f.

□

Example 3.3.25. For some purposes, it is nice to have an explicit example of two ch.f.'s that agree on $[-1,1]$. We know that $(1-|t|)^+$ is the ch.f. of the density $(1-\cos x)/\pi x^2$. Define $\psi(t)$ to be equal to φ on $[-1,1]$ and periodic with period 2, i.e., $\psi(t) = \psi(t+2)$. The Fourier series for ψ is

$$\psi(u) = \frac{1}{2} + \sum_{n=-\infty}^{\infty} \frac{2}{\pi^2(2n-1)^2}\exp(i(2n-1)\pi u)$$

The right-hand side is the ch.f. of a discrete distribution with

$$P(X=0) = 1/2 \text{ and } P(X=(2n-1)\pi) = 2\pi^{-2}(2n-1)^{-2}\ n \in \mathbb{Z}.$$

3.3.5 The Moment Problem

Suppose $\int x^k dF_n(x)$ has a limit μ_k for each k. Then the sequence of distributions is tight by Theorem 3.2.16 and every subsequential limit has the moments μ_k, so we can conclude the sequence converges weakly if there is only one distribution with these moments. It is easy to see that this is true if F is concentrated on a finite interval $[-M,M]$ since every continuous function can be approximated uniformly on $[-M,M]$ by polynomials. The result is false in general.

Counterexample 1. Heyde (1963) [11] Consider the **lognormal density**

$$f_0(x) = (2\pi)^{-1/2}x^{-1}\exp(-(\log x)^2/2)\ x \geq 0$$

and for $-1 \leq a \leq 1$ let

$$f_a(x) = f_0(x)\{1+a\sin(2\pi\log x)\}$$

To see that f_a is a density and has the same moments as f_0, it suffices to show that

$$\int_0^\infty x^r f_0(x) \sin(2\pi \log x) dx = 0 \text{ for } r = 0, 1, 2, ...$$

Changing variables $x = \exp(s + r)$, $s = \log x - r$, $ds = dx/x$ the integral becomes

$$\begin{aligned} & (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp(rs + r^2) \exp(-(s+r)^2/2) \sin(2\pi(s+r)) ds \\ = \; & (2\pi)^{-1/2} \exp(r^2/2) \int_{-\infty}^{\infty} \exp(-s^2/2) \sin(2\pi s) ds \\ = \; & 0 \end{aligned}$$

The two equalities holding because r is an integer and the integrand is odd. From the proof, it should be clear that we could let

$$g(x) = f_0(x)\Big\{1 + \sum_{k=1}^{\infty} a_k \sin(k\pi \log x)\} \text{ if } \sum_{k=1}^{\infty} |a_k| \le 1$$

to get a large family of densities having the same moments as the lognormal.

The moments of the lognormal are easy to compute. Recall that if $\mathcal{X}$ has the standard normal distribution, then we have $\exp(\mathcal{X})$ has the lognormal distribution.

$$\begin{aligned} EX^n &= E\exp(n\mathcal{X}) \\ &= \int e^{nx} (2\pi)^{-1/2} e^{-x^2/2} dx \\ &= e^{n^2/2} \int (2\pi)^{-1/2} e^{-(x-n)^2/2} dx \\ &= \exp(n^2/2) \end{aligned}$$

since the last integrand is the density of the normal with mean n and variance 1. Somewhat remarkably there is a family of discrete random variables with these moments. Let $a > 0$ and

$$P(Y_a = ae^k) = a^{-k} \exp(-k^2/2)/c_a \text{ for } k \in \mathbb{Z}$$

where c_a is chosen to make the total mass 1.

$$\begin{aligned} \exp(-n^2/2) EY_a^n &= \exp(-n^2/2) \sum_k (ae^k)^n a^{-k} \exp(-k^2/2)/c_a \\ &= \sum_k a^{-(k-n)} \exp(-(k-n)^2/2)/c_a \\ &= 1 \end{aligned}$$

by the definition of c_a.

The lognormal density decays like $\exp(-(\log x)^2/2)$ as $|x| \to \infty$. The next counterexample has more rapid decay. Since the exponential distribution, e^{-x} for $x \ge 0$, is determined by its moments we cannot hope to do much better than this.

Counterexample 2. Let $\lambda \in (0, 1)$ and for $-1 \le a \le 1$ let

$$f_{a,\lambda}(x) = c_\lambda \exp(-|x|^\lambda)(1 + a \sin(\beta |x|^\lambda \text{sgn}(x))\}$$

where $\beta = \tan(\lambda\pi/2)$ and $1/c_\lambda = \int \exp(-|x|^\lambda) dx$. To prove that these are density functions and that for a fixed value of λ they have the same moments, it suffices to show

$$\int x^n \exp(-|x|^\lambda) \sin(\beta |x|^\lambda \text{sgn}(x)) dx = 0 \text{ for } n = 0, 1, 2, ...$$

This is clear for even n since the integrand is odd. To prove the result for odd n, it suffices to integrate over $[0, \infty)$. Using the identity

$$\int_0^\infty t^{p-1} e^{-qt} dt = \Gamma(p)/q^p \text{ when Re } q > 0$$

with $p = (n+1)/\lambda$, $q = 1 + \beta i$, and changing variables $t = x^\lambda$, we get

$$\begin{aligned} & \Gamma((n+1)/\lambda)/(1+\beta i)^{(n+1)/\lambda} \\ = & \int_0^\infty x^{\lambda\{(n+1)/\lambda - 1\}} \exp(-(1+\beta i)x^\lambda) \lambda x^{\lambda-1} dx \\ = & \lambda \int_0^\infty x^n \exp(-x^\lambda) \cos(\beta x^\lambda) dx - i\lambda \int_0^\infty x^n \exp(-x^\lambda) \sin(\beta x^\lambda) dx \end{aligned}$$

Since $\beta = \tan(\lambda\pi/2)$

$$(1+\beta i)^{(n+1)/\lambda} = (\cos \lambda\pi/2)^{-(n+1)/\lambda} (\exp(i\lambda\pi/2))^{(n+1)/\lambda}$$

The right-hand side is real since $\lambda < 1$ and $(n+1)$ is even, so

$$\int_0^\infty x^n \exp(-x^\lambda) \sin(\beta x^\lambda) dx = 0$$

A useful sufficient condition for a distribution to be determined by its moments is.

Theorem 3.3.26. *If* $\limsup_{k\to\infty} \mu_{2k}^{1/2k}/2k = r < \infty$ *then there is at most one d.f.* F *with* $\mu_k = \int x^k dF(x)$ *for all positive integers* k.

Remark 3.3.27. This is slightly stronger than **Carleman's condition**

$$\sum_{k=1}^\infty 1/\mu_{2k}^{1/2k} = \infty$$

which is also sufficient for conclusion of Theorem 3.3.26.

Proof. Let F be any d.f. with the moments μ_k and let $\nu_k = \int |x|^k dF(x)$. The Cauchy-Schwarz inequality implies $\nu_{2k+1}^2 \le \mu_{2k}\mu_{2k+2}$ so

$$\limsup_{k\to\infty} (\nu_k^{1/k})/k = r < \infty$$

Taking $x = tX$ in Lemma 3.3.2 and multiplying by $e^{i\theta X}$, we have

$$\left| e^{i\theta X} \left(e^{itX} - \sum_{m=0}^{n-1} \frac{(itX)^m}{m!} \right) \right| \le \frac{|tX|^n}{n!}$$

Taking expected values

$$\left| \varphi(\theta + t) - \varphi(\theta) - t\varphi'(\theta) \cdots - \frac{t^{n-1}}{(n-1)!} \varphi^{(n-1)}(\theta) \right| \le \frac{|t|^n}{n!} \nu_n$$

Using the last result, the fact that $\nu_k \le (r+\epsilon)^k k^k$ for large k, and the trivial bound $e^k \ge k^k/k!$ (expand the left-hand side in its power series), we see that for any θ

$$(\star) \varphi(\theta + t) = \varphi(\theta) + \sum_{m=1}^\infty \frac{t^m}{m!} \varphi^{(m)}(\theta) \text{ for } |t| < 1/er$$

Let G be another distribution with the given moments and ψ its ch.f. Since $\varphi(0) = \psi(0) = 1$, it follows from $(\star)$ and the induction that $\varphi(t) = \psi(t)$ for $|t| \le k/3r$ for all k, so the two ch.f.'s coincide and the distributions are equal.

□

Theorem 3.3.28. *Suppose $\int x^k dF_n(x)$ has a limit μ_k for each k and*

$$\limsup_{k\to\infty} \mu_{2k}^{1/2k} 2k < \infty$$

then F_n converges weakly to the unique distribution with these moments.

Our results so far have been for the so-called **Hamburger moment problem**. If we assume *a priori* that the distribution is concentrated on $[0,\infty)$, we have the **Stieltjes moment problem**. There is a 1-1 correspondence between $X \geq 0$ and symmetric distributions on $\mathbb{R}$ given by $X \to \xi X^2$ where $\xi \in \{-1,1\}$ is independent of X and takes its two values with equal probability. From this we see that

$$\limsup_{k\to\infty} \nu_k^{1/2k}/2k < \infty$$

is sufficient for there to be a unique distribution on $[0,\infty)$ with the given moments.

Counterexample 3. Let $\lambda \in (0,1/2)$, $\beta = \tan(\lambda\pi)$, $-1 \leq a \leq 1$ and

$$f_a(x) = c_\lambda \exp(-x^\lambda)(1 + a\sin(\beta x^\lambda)) \text{ for } x \geq 0$$

where $1/c_\lambda = \int_0^\infty \exp(-x^\lambda)dx$.

By imitating the calculation in Counterexample 2, it is easy to see that the f_a are probability densities that have the same moments. This example seems to be due to Stoyanov (1987) p. 92-93 [15]. The special case $\lambda = 1/4$ is widely known.

3.4 Central Limit Theorems

We are now ready for the main business of the chapter. We will first prove the central limit theorem for

3.4.1 i.i.d. Sequences

Theorem 3.4.1. *Let $X_1, X_2, \ldots$ be i.i.d. with $EX_i = \mu$, $var(X_i) = \sigma^2 \in (0,\infty)$. If $S_n = X_1 + \cdots + X_n$ then*

$$(S_n - n\nu)/\sigma n^{1/2} \Rightarrow \chi$$

where χ has the standard normal distribution.

This notation is non-standard but convenient. To see the logic note that the square of a normal has a chi-squared distribution.

Proof. By considering $X_i' = X_i - \mu$, it suffices to prove the result when $\mu = 0$. From Theorem 3.3.19

$$\varphi(t) = E\exp(itX_1) = 1 - \frac{\sigma^2 t^2}{2} + o(t^2)$$

so

$$E\exp(itS_n/\sigma n^{1/2}) = \left(1 - \frac{t^2}{2n} + o(n^{-1})\right)^n$$

From 3.1.1 it should be clear that the last quantity $\to \exp(-t^2/2)$ as $n \to \infty$, which with Theorem 3.3.16 completes the proof. However, Lemma 3.1.1 is a fact about real numbers.

Theorem 3.4.2. *If $c_n \to c \in \mathbb{C}$ then $(1 + c_n/n)^n \to e^c$.*

Proof. The proof is based on two simple facts:

Lemma 3.4.3. *Let $z_1, ..., z_n$ and $\omega_1, ..., \omega_n$ be complex numbers of modulus $\leq \theta$. Then*

$$\left| \prod_{m=1}^{n} z_m - \prod_{m=1}^{n} \omega_m \right| \leq \theta^{n-1} \sum_{m=1}^{n} |z_m - \omega_m|$$

Proof. The result is true for $n = 1$. To prove it for $n > 1$ observe that

$$\begin{aligned} \left| \prod_{m=1}^{n} z_m - \prod_{m=1}^{n} \omega_m \right| &\leq \left| z_1 \prod_{m=2}^{n} z_m - z_1 \prod_{m=2}^{n} \omega_m \right| + \left| z_1 \prod_{m=2}^{n} \omega_m - \omega_1 \prod_{m=2}^{n} \omega_m \right| \\ &\leq \theta \left| \prod_{m=2}^{n} z_m - \prod_{m=2}^{n} \omega_m \right| + \theta^{n-1} |z_1 - \omega_1| \end{aligned}$$

and use induction.

□

Lemma 3.4.4. *If b is a complex number with $|b| \leq 1$ then $|e^b - (1 + b)| \leq |b|^2$.*

Proof. $e^b - (1 + b) = b^2/2! + b^3/3! + b^4/4! + \dots$ so if $|b| \leq 1$ then

$$|e^b - (1 + b)| \leq \frac{|b|^2}{2}(1 + 1/2 + 1/2^2 + \dots) = |b|^2$$

□

Proof. Theorem 3.4.2. Let $z_m = (1 + c_n/n)$, $\omega_m = \exp(c_n/n)$, and $\gamma > |c|$. For large n, $|c_n| < \gamma$. Since $1 + \gamma/n \leq \exp(\gamma/n)$, it follows 3.4.3 and 3.4.4 that

$$|(1 + c_n/n)^n - e^{c_n}| \leq \left(e^{\gamma/n} \right)^{n-1} n \left| \frac{c_n}{n} \right|^2 \leq e^{\gamma} \frac{\gamma^2}{n} \to 0$$

as $n \to \infty$.

□

Example 3.4.5. Roulette. A roulette wheel has slots numbered 1-36 (18 red and 18 black) and two slots numbered 0 and 00 that are painted green. Players can bet \$1 that the ball will land in a red (or black) slot and win \$1 if it does. If we let X_i be the winnings on the ith play then $X_1, X_2, ...$ are i.i.d. with $P(X_i = 1) = 18/38$ and $P(X_i = 01) = 20/38$.

$$EX_i = -1/19 \text{ and } \text{var}(X) = EX^2 - (EX)^2 = 1 - (1/19)^2 = 0.9972$$

We are interested in

$$P(S_n \geq 0) = P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \geq \frac{-n\mu}{\sigma\sqrt{n}} \right)$$

Taking $n = 361 = 19^2$ and replacing σ by 1 to keep computations simple,

$$\frac{-n\mu}{\sigma\sqrt{n}} = \frac{361 \cdot (1/19)}{\sqrt{361}} = 1$$

So the central limit theorem and our table of the normal distribution in the back of the book tells us that

$$P(S_n \geq 0) \approx P(\chi \geq 1) = 1 - 0.8413 = 0.1587$$

In words, after 361 spins of the roulette wheel the casino will have won \$19 of your money on the average, but there is a probability about 0.16 that you will be ahead.

Example 3.4.6. Coin flips. Let $X_1, X_2, ...$ be i.i.d. with $P(X_i = 0) = P(X_i = 1) = 1/2$. If $X_i = 1$ indicates that a heads occurred on the ith toss then $S_n = X_1 + \cdots + X_n$ is the total number of heads at time n.

$$EX_i = 1/2 \textbf{ and } \text{var}(X) = EX^2 - (EX)^2 = 1/2 - 1/4 = 1/4$$

So the central limit theorem tells us $(S_n - n/2)/\sqrt{n/4} \Rightarrow \chi$. Our table of the normal distribution tells us that

$$P(\chi > 2) = 1 - 0.9773 = 0.0227$$

so $P(|\chi| \le 2) = 1 - 2(0.0227) = 0.9546$, or plugging into the central limit theorem

$$0.95 \approx P((S_n - n/2))/\sqrt{n/4} \in [-2, 2]) = P(S_n - n/2 \in [-\sqrt{n}, \sqrt{n}])$$

Taking $n = 10,000$ this says that 95% of the time the number of heads will be between 4900 and 5100.

Example 3.4.7. Normal approximation to the binomial. Let $X_1, X_2, ...$ and S_n be as in the previous example. To estimate $P(S_{16} = 8)$ using the central limit theorem, we regard 8 as the interval $[7.5, 8.5]$. Since $\mu = 1/2$, and $\sigma\sqrt{n} = 2$ for $n = 16$

$$\begin{aligned} P(|S_{16} - 8| \le 0.5) &= P\left(\frac{|S_n - n\mu|}{\sigma\sqrt{n}} \le 0.25\right) \\ &\approx P(|\chi| \le 0.25) \\ &= 2(0.5987 - 0.5) = 0.1974 \end{aligned}$$

Even though n is small, this agrees well with the exact probability

$$\binom{16}{8} 2^{-16} = \frac{13 \cdot 11 \cdot 10 \cdot 9}{65,536} = 0.1964$$

The computations above motivate the **histogram correction**, which is important in using the normal approximation for small n. For example, if we are going to approximate $P(S_{16} \le 11)$, then we regard this probability as $P(S_{16} \le 11.5)$. one obvious reason for doing this is to get the same answer if we regard $P(S_{16} \le 11) = 1 - P(S_{16} \ge 12)$.

Example 3.4.8. Normal approximation to the Poisson. Let Z_λ have a Poisson distribution with mean λ. If $X_1, X_2, ...$ are independent and have Poisson distributions with mean 1, then $S_n = X_1 + \cdots + X_n$ has a Poisson distribution with mean n. Since $\text{var}(X_i) = 1$, the central limit theorem implies:

$$S_n - n)/n^{1/2} \Rightarrow \chi \text{ as } n \to \infty$$

To deal with values of λ that are not integers, let N_1, N_2, N_3 be independent Poisson with means $[\lambda]$, $\lambda - [\lambda]$, and $[\lambda] + 1 - \lambda$. If we let $S_{[\lambda]} = N_1$, Z_λ, $Z_\lambda = N_1 + N_2$ and $S_{[\lambda]+1} = N_1 + N_2 + N_3$ then $S_{[\lambda]} \le Z_\lambda \le S_{[\lambda]+1}$ and using the limit theorem for the S_n it follows that

$$(Z_\lambda - \lambda)/\lambda^{1/2} \Rightarrow \chi \text{ as } \lambda \to \infty$$

Example 3.4.9. Pairwise independence is good enough for the strong law of large numbers (see Theorem 2.4.1). It is not good enough for the central limit theorem. Let $\xi_1, \xi_2, ...$ be i.i.d. with $P(\xi_i = 1) = P(\xi_i = -1) = 1/2$. We will arrange things so that for $n \ge 1$

$$S_{2^n} = \xi_1(1 + \xi_2) \ldots (1 + \xi_{n+1}) = \begin{cases} \pm 2^n & \text{with prob } 2^{-n-1} \\ 0 & \text{with prob } 1 - 2^{-n} \end{cases}$$

To do this we let $X_1 = \xi_1$, $X_2 = \xi_1\xi_2$, and for $m = 2^{n-1} + j$, $0 < j \le 2^{n-1}$, $n \ge 2$ let $X_m = X_j\xi_{n+1}$. Each X_m is a product of a different set of ξ_j's so they are pairwise independent.

3.4.2 Triangular Arrays

Theorem 3.4.10. ***The Lindeberg-Feller theorem.*** *For each n, let $X_{n,m}$, $1 \le m \le n$, be independent random variables with $EX_{n,m} = 0$. Suppose*
(i) $\sum_{m=1}^{n} EX_{n,m}^2 \to \sigma^2 > 0$
(ii) For all $\epsilon > 0$, $\lim_{n\to\infty} \sum_{m=1}^{n} E(|X_{n,m}|^2; |X_{n,m}| > \epsilon) = 0$.
Then $S_n = X_{n,1} + \cdots + X_{n,n} \Rightarrow \sigma_\chi$ as $n \to \infty$.

Remark 3.4.11. In words, the theorem says that a sum of a large number of small independent effects has approximately a normal distribution. To see that Theorem 3.4.10 contains our first central limit theorem, let $Y_1, Y_2, ...$ be i.i.d. with $EY_i = 0$ and $EY_i^2 = \sigma^2 \in (0, \infty)$, and let $X_{n,m} = Y_m/n^{1/2}$. Then $\sum_{m=1}^{n} EX_{n,m}^2 = \sigma^2$ and if $\epsilon > 0$

$$\begin{aligned}\sum_{m=1}^{n} E(|X_{n,m}|^2; |X_{n,m}| > \epsilon) &= nE(|Y_1/n^{1/2}|^2; |Y_1/n^{1/2}| > \epsilon) \\ &= E(|Y_1|^2; |Y_1| > \epsilon n^{1/2}) \to 0\end{aligned}$$

by the dominated convergence theorem since $EY_1^2 < \infty$.

Proof. Let $\varphi_{n,m}(t) = E\exp(itX_{n,m})$, $\sigma_{n,m}^2 = EX_{n,m}^2$. By Theorem 3.3.16, it suffices to show that

$$\prod_{m=1}^{n} \varphi_{n,m}(t) \to \exp(-t^2\sigma^2/2)$$

Let $z_{n,m} = \varphi_{n,m}(t)$ and $\omega_{n,m} = (1 - t^2\sigma_{n,m}^2/2)$. By 3.5

$$\begin{aligned}|z_{n,m} - \omega_{n,m}| &\le E(|tX_{n,m}|^3) \wedge 2|tX_{n,m}|^2) \\ &\le E(|tX_{n,m}|^3; |X_{n,m}| \le \epsilon) + E(2|tX_{n,m}|^2; |X_{n,m}| > \epsilon) \\ &\le \epsilon t^3 E(|X_{n,m}|^2; |X_{n,m}| \le \epsilon) + 2t^2 E(|X_{n,m}|^2|; |X_{n,m}| > \epsilon)\end{aligned}$$

Summing $m = 1$ to n, letting $n \to \infty$, and using (i) and (ii) gives

$$\limsup_{n\to\infty} \sum_{m=1}^{n} |z_{n,m} - \omega_{n,m}| \le \epsilon t^3 \sigma^2$$

Since $\epsilon > 0$ is arbitrary, it follows that the sequence converges to 0. Our next step is to use Lemma 3.4.3 with $\theta = 1$ to get

$$\left| \prod_{m=1}^{n} \varphi_{n,m}(t) - \prod_{m=1}^{n} (1 - t^2\sigma_{n,m}/2) \right| \to 0$$

To check the hypotheses of Lemma 3.4.3, note that since $\varphi_{n,m}$ is a ch.f. $|\varphi_{n,m}(t)| \le 1$ for all n, m. For the terms in the second product we note that

$$\sigma_{n,m}^2 \le \epsilon^2 + E(|X_{n,m}|^2; |X_{n,m}| > \epsilon)$$

and ϵ is arbitrary so (ii) implies $\sup_m \sigma_{n,m}^2 \to 0$ and thus if n is large $1 \le 1 - t^2\sigma_{n,m}^2/2 > -1$ for all m.

To complete the proof now, we apply exercise with $c_{m,n} = -t^2\sigma_{n,m}^2/2$. We have just shown $\sup_m \sigma_{n,m}^2 \to 0$. (i) implies

$$\sum_{m=1}^{n} c_{m,n} \to -\sigma^2 t^2/2$$

so $\prod_{m=1}^{n}(1 - t^2\sigma_{n,m}^2/2) \to \exp(-t^2\sigma^2/2)$ and the proof is complete.

□

Example 3.4.12. Cycles in a random permutation and record values. Continuing the analysis of previous examples, let $Y_1, Y_2, ...$ be independent with $P(Y_m = 1) = 1/m$, and $P(Y_m = 0) = 1 - 1/m$. $EY_m = 1/m$ and $\text{var}(Y_m) = 1/m - 1/m^2$. So if $S_n = Y_1 + \cdots + Y_n$ then $ES_n \sim \log n$ and $\text{var}(S_n) \sim \log n$. Let

$$X_{n,m} = (Y_m - 1/m)/(\log n)^{1/2}$$

$EX_{n,m} = 0$, $\sum_{m=1}^n EX_{n,m}^2 \to 1$, and for any $\epsilon > 0$

$$\sum_{m=1}^n E(|X_{n,m}|^2; |X_{n,m}| > \epsilon) \to 0$$

since the sum is 0 as soon as $(\log n)^{-1/2} < \epsilon$. Applying Theorem 3.4.10 now gives

$$(\log n)^{-1/2}\left(S_n - \sum_{m=1}^n \frac{1}{m}\right) \Rightarrow \chi$$

Observing

$$\sum_{m=1}^{n-1} \frac{1}{m} \geq \int_1^n x^{-1} dx = \log n \geq \sum_{m=2}^n \frac{1}{m}$$

shows $|\log n - \sum_{m=1}^n 1/m| \leq 1$ and the conclusion can be written as

$$(S_n - \log n)/(\log n)^{1/2} \Rightarrow \chi$$

Example 3.4.13. The converse of the three series theorem. Recall the set up of Theorem 2.5.8. Let $X_1, X_2, ...$ be independent, let $A > 0$ and let $Y_m = X_m 1_{(|X_m| \leq A)}|$. In order that $\sum_{n=1}^\infty X_n$ converges (i.e. $\lim_{N\to\infty} \sum_{n=1}^N X_n$ exists) it is necessary that:

$$\text{(i) } \sum_{n=1}^\infty P(|X_n| > A) < \infty, \text{ (ii) } \sum_{n=1}^\infty EY_n \text{ converges, and (iii) } \sum_{n=1}^\infty \text{var}(Y_n) < \infty$$

Proof. The necessity of the first condition is clear. For if that sum is infinite, $P(|X_n| > A \text{ i.o.}) < 0$ and $\lim_{n\to\infty} \sum_{m=1}^n X_m$ cannot exist. Suppose next that the sum in (i) is finite but the sum in (iii) is infinite. Let

$$c_n = \sum_{m=1}^n \text{var}(Y_m) \text{ and } X_{n,m} = (Y_m - EY_m)/c_n^{1/2}$$

$EX_{n,m} = 0$, $\sum_{m=1}^n EX_{n,m}^2 = 1$, and for any $\epsilon > 0$

$$\sum_{m=1}^n E(|X_{n,m}|^2; |X_{n,m}| > \epsilon) \to 0$$

since the sum is 0 as soon as $2A/c_n^{1/2} < \epsilon$. Applying Theorem 3.4.10 now gives that if $S_n = X_{n,1} + \cdots + X_{n,n}$ then $S_n \Rightarrow \chi$. Now

(i) if $\lim_{n\to\infty} \sum_{m=1}^n X_m$ exists, $\lim_{n\to\infty} \sum -m = 1^n Y_m$ exists.

(ii) if we let $T_n = (\sum_{m\leq n} Y_m)/c_n^{1/2}$ then $T_n \Rightarrow 0$. The last two results and exercise imply $(S_n - T_n) \Rightarrow \chi$. Since

$$S_n - T_n = -\left(\sum_{m\leq n} EY_m\right)/c_n^{1/2}$$

is not random, this is absurd.

Finally, assume the series in (i) and (iii) are finite. Theorem 2.5.6 implies that $\lim_{n\to\infty} \sum_{m=1}^n (Y_m - EY_m)$ exists, so if $\lim_{n\to\infty} \sum_{m=1}^n X_m$ and hence $\lim_{n\to\infty} \sum_{m=1}^n Y_m$ does, taking differences shows that (ii) holds.

Example 3.4.14. Infinite variance. Suppose $X_1, X_2, ...$ are i.i.d. and have $P(X_1 > x) = P(X_1 < -x)$ and $P(|X_1| > x) = x^{-2}$ for $x \geq 1$.

$$E|X_1|^2 = \int_0^\infty 2xP(|X_1| > x)dx = \infty$$

but it turns out that when $S_n = X_1 + \cdots + X_n$ is suitably normalized it converges to a normal distribution. Let

$$Y_{n,m} = X_m 1_{(|X_m| \leq n^{1/2} \log\log n)}$$

The truncation level $c_n = n^{1/2} \log\log n$ is chosen large enough to make

$$\sum_{m=1}^{n} P(Y_{n,m} \neq X_m) \leq nP(|X_1| > c_n) \to 0$$

However, we want the variance of $Y_{n,m}$ to be as small as possible, so we keep the truncation close to the lowest possible level.

Our next step is to show $EY_{n,m}^2 \sim \log n$. For this we need upper and lower bounds. Since $P(|Y_{n,m}| > x) \leq P(|X_1| > x)$ and is 0 for $x > c_n$, we have

$$\begin{aligned} EY_{n,m}^2 &\leq \int_0^{c_n} 2yP(|X_1| > y)dy \\ &= 1 + \int_1^{c_n} 2/ydy \\ &= 1 + 2\log c_n \\ &= 1 + \log n + 2\log\log\log n \\ &= \log n \end{aligned}$$

In the other direction, we observe $P(|Y_{n,m}| > x) = P(|X_1| > x) - P(|X_1| > c_n)$ and the right-hand side is $\geq (1 - (\log\log n)^{-2})P(|X_1| > x)$ when $x \leq \sqrt{n}$ so

$$EY_{n,m}^2 \geq (1 - (\log\log n)^{-2}) \int_1^{\sqrt{n}} 2/ydy \sim \log n$$

If $S_n' = Y_{n,1} + \cdots + Y_{n,n}$ then $\text{var}(S_n') \sim n\log n$, so we apply Theorem 3.4.10 to $X_{n,m} = Y_{n,m}/(n\log n)^{1/2}$. Things have been arranged so that (i) is satisfied. Since $|Y_{n,m}| \leq n^{1/2}\log\log n$, the sum in (ii) is 0 for large n, and it follows that $S_n'/(n\log n)^{1/2} \Rightarrow \chi$. Since the choice of c_n guarantees $P(S_n \neq S_n') \to 0$, the same result holds for S_n.

Remark 3.4.15. In Section 3.6, we will see that if we replace $P(|X_1| > x) = x^{-2}$ in example by $P(|X_1| > x) = x^{-\alpha}$ where $0 < \alpha < 2$, then $S_n/n^{1/\alpha} \Rightarrow$ to a limit which is not χ. The last word on convergence to the normal distribution is the next result due to Lévy.

Theorem 3.4.16. *Let $X_1, X_2, ...$ be i.i.d. and $S_n = X_1 + \cdots + X_n$. In order that there exist constants a_n and $b_n > 0$ so that $(S_n - a_n)/b_n \Rightarrow \chi$, it is necessary and sufficient that*

$$y^2 P(|X_1| > y)/E(|X_1|^2; |X_1| \leq y) \to 0$$

A proof can be found in Gnedenko and Kolmogorov (1954) [9], a reference that contains the last word on many results about sums of

3.4.3 Primate Divisors (Erdos-Kac)

Our aim here is to prove that an integer picked at random from $\{1, 2, ..., n\}$ has about

$$\log\log n + \chi(\log\log n)^{1/2}$$

prime divisors. Since $\exp(e^4) = 5.15 \times 10^{23}$, this result does not apply to most numbers we encounter in "everyday life." The first step in deriving this result is to give a

Second proof of Theorem 3.4.10. The first step is to let

$$h_n(\epsilon) = \sum_{m=1}^{n} E(X_{n,m}^2; |X_{n,m}| > \epsilon)$$

and observe

Lemma 3.4.17. *$h_n(\epsilon) \to 0$ for each fixed $\epsilon > 0$ so we can pick $\epsilon_n \to 0$ so that $h_n(\epsilon_n) \to 0$.*

Proof. Let N_m be chosen so that $h_n(1/m) \le 1/m$ for $n \ge N_m$ and $m \to N_m$ is increasing. Let $\epsilon_n = 1/m$ for $N_m \le n < N_{m+1}$, and $= 1$ for $n < N_1$. When $N_m \le n < N_{m+1}$, $\epsilon_n = 1/m$, so $|h_n(\epsilon_n)| = |h_n(1/m)| \le 1/m$ and the desired result follows.

□

Let $X'_{n,m} = X_{n,m} 1_{(|X_{n,m}| > \epsilon_n)}$, $Y_{n,m} = X_{n,m} 1_{(|X_{n,m}| \le \epsilon_n)}$, and $Z_{n,m} = Y_{n,m} - EY_{n,m}$. Clearly $|Z_{n,m} \le 2\epsilon_n$. Using $X_{n,m} = X'_{n,m} + Y_{n,m}$, $Z_{n,m} = Y_{n,m} - EY_{n,m}$, $EY_{n,m} = -EX'_{n,m}$, the variance of the sum is the sum of the variances, and $\text{var}(W) \le EW^2$, we have

$$\begin{aligned} E\Big(\sum_{m=1}^{n} X_{n,m} - \sum_{m=1}^{n} Z_{n,m}\Big)^2 &= E\Big(\sum_{m=1}^{n} X'_{n,m} - EX'_{n,m}\Big)^2 \\ &= \sum_{m=1}^{n} E(X'_{n,m} - EX'_{n,m})^2 \\ &\le \sum_{m=1}^{n} E(X'_{n,m})^2 \to 0 \end{aligned}$$

as $n \to \infty$, by the choice of ϵ_n.

Let $S_n = \sum_{m=1}^{n} X_{n,m}$ and $T_n = \sum_{m=1}^{n} Z_{n,m}$. The last computation shows $S_n - T_n \to 0$ in L^2 and hence in probability by Lemma 2.2.2. Thus, by exercise, it suffices to show $T_n \Rightarrow \sigma_\chi$. (i) implies $ES_n^2 \to \sigma^2$. We have just shown that $E(S_n - T_n)^2 \to 0$, so the triangle inequality for the L^2 norm implies $ET_n^2 \to \sigma^2$. To compute higher moments, we observe

$$T_n^r = \sum_{k=1}^{r} \sum_{r_i} \frac{r!}{r_1! \dots r_k!} \frac{1}{k!} \sum_{i_j} Z_{n,i_1}^{r_1} \dots Z_{n,i_k}^{r_k}$$

where $\sum_{r_i}$ extends over all k-tuples of positive integers with $r_1 + \dots + r_k = r$ and $\sum_{i_j}$ extends over all k-tuples of distinct integers with $1 \le i \le n$. If we let

$$A_n(r_1, ..., r_k) = \sum_{i_j} = EZ_{n,i_1}^{r_1} \dots EZ_{n,i_k}^{r_k}$$

then

$$ET_n^r = \sum_{k=1}^{r} \sum_{r_i} \frac{r!}{r_1! \dots r_k!} \frac{1}{k!} A_n(r_1, ..., r - K)$$

To evaluate the limit of ET_n^r we observe:
(a) If some $r_j = 1$, then $A_n(r_1, ..., r_k) = 0$ since $EZ_{n,i_j} = 0$.
(b) If all $r_j = 2$ then

$$\sum_{1_j} EZ_{n,i_1}^2 \dots EZ_{n,i_k}^2 \leq \Big(\sum_{m=1}^n EZ_{n,m}^2 \Big)^k \to \sigma^{2k}$$

To argue the other inequality, we note that for any $1 \leq a < b \leq k$ we can estimate the sum over all the $i_1, ..., i_k$ with $i_a = i_b$ by replacing EZ_{n,i_a}^2 by $(2\epsilon_n)^2$ to get (the factor $\binom{k}{2}$ giving the number of ways to pick $1 \leq a < b \leq k$)

$$\Big(\sum_{m=1}^n EZ_{n,m}^2 \Big)^k - \sum_{i_j} EZ_{n,i_1}^2 \dots EZ_{n,i_k}^2 \leq \binom{k}{2} (2\epsilon_n)^2 \Big(\sum_{m=1}^n EZ_{n,m}^2 \Big)^{k-1} \to 0$$

(c) If all the $r_i \geq 2$ but some $r_j > 2$ then using

$$E|Z_{n,i_j}|^{r_j} \leq (2\epsilon_n)^{r_j - 2} EZ_{n,i_k}|^{r_k}$$

we have

$$\begin{aligned} |A_n(r_1, ..., r_k)| &\leq \sum_{i_j} E|Z_{n,i_1}|^{r_1} \dots E|Z_{n,i_k}|^{r_k} \\ &\leq (2\epsilon_n)^{r-2k} A_n(2, ..., 2) \to 0 \end{aligned}$$

When r is odd, some r_j must be $= 1$ or ≥ 3 so $ET_n^r \to 0$ by (a) and (c). If $r = 2k$ is even, (a) - (c) imply

$$ET_n^r \to \frac{\sigma^{2k}(2k)!}{2^k k!} = E(\sigma_\chi)^r$$

and the result follows from Theorem 3.3.28.

□

Turning to the result for prime divisors, let P_n denote the uniform distribution on $\{1, ..., n\}$. If $P_\infty(A) \equiv \lim P_n(A)$ exists, the limit is called the density of $A \subset \mathbb{Z}$. Let A_p be the set of integers divisible by p. Clearly, if p is a prime $P_\infty(A_p) = 1/p$ and $q \neq p$ is another prime

$$P_\infty(A_p \cap A_q = 1/pq] P_\infty(A_p) P_\infty(A_q)$$

Even though P_∞ is not a probability measure (since $P(|i|) = 0$ for all i), we can interpret this as saying that the events of being divisible by p and q are independent. Let $\delta_p(n) = 1$ if n is divisible by p, and $= 0$ otherwise, and

$$g(n) = \sum_{p \leq n} \delta_p(n) \text{ be the number of prime divisors of } n$$

this and future sums on p being over the primes. Intuitively, the $\delta_p(n)$ behave like X_p that are i.i.d. with

$$P(X_p = 1) = 1/p \text{ and } P(X_p = 0) = 1 - 1/p$$

The mean and variance of $\sum_{p \leq n} X_p$ are

$$\sum_{p \leq n} 1/p \text{ and } \sum_{p \leq n} 1/p(1 - 1/p)$$

respectively. It is known that

$$(\star) \sum_{p \le n} 1/p = \log\log n + O(1)$$

(see Hardy and Wright, 1959, Chapter XXII) [10], while anyone can see $\sum_{[} 1/p^2 < \infty$, so applying Theorem 3.4.10 to X_p and making a small leap of faith gives us:

Theorem 3.4.18. ***Erdös-Kac central limit theorem.*** *As* $n \to \infty$

$$P_n\big(m \le n : g(m) - \log\log n \le x(\log\log n)^{1/2}\big) \to P(\chi \le x)$$

Proof We begin by showing that we can ignore the primes "near" n. Let

$$\alpha_n = n^{1/\log\log n}$$

$$\log \alpha_n = \log n / \log\log n$$

$$\log\log \alpha_n = \log\log n - \log\log\log n$$

The sequence α_n has two nice properties:
(a) $\big(\sum_{\alpha_n < p \le n} 1/p\big)/(\log\log n)^{1/2} \to 0$ by $(\star)$
Proof of (a). By $(\star)$

$$\begin{aligned} \sum_{\alpha_n < p \le n} 1/p &= \sum_{p \le n} 1/p - \sum_{p \le \alpha_n} 1/p \\ &= \log\log n - \log\log \alpha_n + O(1) \\ &= \log\log\log n + O(1) \end{aligned}$$

(b) If $\epsilon > 0$, then $\alpha_n \le n^\epsilon$ for large n and hence $\alpha_n'/n \to 0$ for all $r < \infty$.
Proof of (b). $1/\log\log n \to 0$ as $n \to \infty$. Let $g_n(m) = \sum_{p \le \alpha_n} \delta_p(m)$ and let E_n denote expected value w.r.t. P_n.

$$E_n\Big(\sum_{\alpha_n < p \le n} \delta_p\Big) = \sum_{\alpha_n < p \le n} P_n(m : \delta_p(m) = 1) \le \sum_{\alpha_n < p \le n} 1/p$$

so by (a) it is enough to prove the result for g_n. Let

$$S_n = \sum_{p \le \alpha_n} X_p$$

where the X_p are the independent random variables introduced above. Let $b_n = ES_n$ and $a_n^2 = \mathrm{var}(S_n)$. (a) tells us that b_n and a_n^2 are both

$$\log\log n + o((\log\log n)^{1/2})$$

so it suffices to show

$$P_n(m : g_n(m) - b_n \le x a_n) \to P(\chi \le x)$$

An application of Theorem 3.4.10 shows $(S_n - b_n)/a_n \Rightarrow \chi$, and since $|X_p| \le 1$ it follows from the second proof of Theorem 3.4.10 that

$$E((S_n - b_n)/a_n)^r \to E\chi^r \text{ for all } r$$

Using notation from that proof (and replacing i_j by p_j)

$$ES_n^r = \sum_{k=1}^{r} \sum_{r_i} \frac{r!}{r_1! \ldots r_k!} \frac{1}{k!} \sum_{p_j} E(X_{p_1}^{r_1} \ldots X_{p_k}^{r_k})$$

Since $X_p \in \{0,1\}$, the summand is

$$E(X_{p_1}, \ldots, X_{p_k}) = 1/(p_1 \ldots p_k)$$

A little thought reveals that

$$E_n(\delta_{p_1} \ldots \delta_{p_k}) \leq \frac{1}{n}[n/(p_1 \ldots p_k)]$$

The two moments differ by $\leq 1/n$, so

$$\begin{aligned} |E(S_n^r) - E_n(g_n^r)| &= \sum_{k=1}^{r} \sum_{r_i} \frac{r!}{r_1! \ldots r_k!} \frac{1}{k!} \sum_{p_j} \frac{1}{n} \\ &\leq 13n\Big(\sum_{p \leq \alpha_n} 1\Big)^r \\ &\leq \frac{\alpha_n^r}{n} \to 0 \end{aligned}$$

by (b). Now

$$E(S_n - b_n)^r = \sum_{m=0}^{r} \binom{r}{m} E S_n^m (-b_n)^{r-m}$$

$$E(g_n - b_n)^r = \sum_{m=0}^{r} \binom{r}{m} E g_n^m (-b_n)^{r-m}$$

so subtracting and using our bound on $|E(S_n^r) - E_n(g_n^r)|$ with $r = m$

$$|E(S_n - b_n)^r - E(g_n - b_n)^r| \leq \sum_{m=0}^{r} \binom{r}{m} \frac{1}{n} \alpha_n^m b_n^{r-m} = (\alpha_n + b_n)^r / n \to 0$$

since $b_n \leq \alpha_n$. This is more than enough to conclude that

$$E((g_n - b_n)/a_n)^r \to E\chi^r$$

and the desired result follows from Theorem 3.3.28.

□

3.4.4 Rates of Convergence (Berry-Essen)

Theorem 3.4.19. *Let $X_1, X_2, \ldots$ be i.i.d. with $EX_i = 0$, $EX_i^2 = \sigma^2$, and $E|X_i|^3 = \rho < \infty$. If $F_n 9x)$ is the distribution of $(X_1 + \cdots + X_n)/\sigma\sqrt{n}$ and $\mathcal{N}(x)$ is the standard normal distribution, then*

$$|F_n(x) - \mathcal{N}| \leq 3\rho/\sigma^3\sqrt{n}$$

Remark 3.4.20. The reader should note that the inequality holds for all n and x, but since $\rho \geq \sigma^3$, it only has nontrivial content for $n \geq 10$. It is easy to see that the rate cannot be faster than $n^{-1/2}$. When $P(X_i = 1) = p(X_i = -1) = 1/2$, symmetry imply that

$$F_{2n}(0) = \frac{1}{2}[1 + P(S_{2n} = 0)] = \frac{1}{2}(1 + (\pi n)^{-1/2}) + o(n^{-1/2})$$

Proof Since neither side of the inequality is affected by scaling, we can suppose without loss of generality that $\sigma^2 = 1$. The first phase of the argument is to derive an inequality that relates the difference between the two distributions to the distance between their ch.f.'s. Polya's density

$$h_L(x) = \frac{1 - \cos Lx}{\pi L x^2}$$

has ch.f. $\omega_L(\theta) = (1 - |\theta/:|)^+$ for $|\theta| \leq L$. we will use H_L for its distribution function. we will convolve the distributions under consideration with H_L to get ch.f. that have compact support. The first step is to show that convolution with H_L does not reduce the difference between the distributions too much.

Lemma 3.4.21. *Let F and G be distribution functions with $G'(x) \leq \lambda < \infty$. Let $\triangle(x) = F(x) - G(x)$, $\eta = \sup|\triangle(x)|$, $\triangle \star H_L$, and $\eta_P = \sup|\triangle_L(x)|$. Then*

$$\eta_L \geq \frac{\eta}{2} - \frac{12\lambda}{\pi L} \text{ or } \eta \leq 2\eta_L + \frac{24\lambda}{\pi L}$$

Proof. $\triangle$ goes to 0 at $\pm\infty$, G is continuous, and F is a d.f., so there is an x_0 with $\triangle(x_0) = \eta$ or $\triangle(x_0-) = -\eta$. By looking at the d.f.'s of (-2) times the r.v.'s in the second case, we can suppose without loss of generality that $\triangle(x_0) = \eta$. Since $G'(x) \leq \lambda$ and F is nondecreasing, $\triangle(x_0 + s) \geq \eta - \lambda s$. Letting $\delta = \eta/2\lambda$, and $t = x_0 + \delta$, we have

$$\triangle(t - x) \geq \begin{cases} (\eta/2) + \lambda x & \text{for } |x| \leq \delta \\ -\eta & \text{otherwise} \end{cases}$$

To estimate the convolution $\triangle_L$, we observe

$$2\int_\delta^\infty h_L(x)dx \leq 2\int_\delta^\infty 2/(\pi L x^2)dx = 4/(\pi L\delta)$$

Looking at $(-\delta, \delta)$ and its complement separately and noticing that symmetry implies $\int_{-\delta}^{\delta} x h_L(x)dx = 0$, we have

$$\eta_L \geq \triangle_L(t) \geq \frac{\eta}{2}\big(1 - \frac{4}{\pi L\delta}\big) - \eta\frac{4}{\pi L\delta} = \frac{6\eta}{2} - \frac{6\eta}{\pi L\delta} = \frac{\eta}{2} - \frac{12\lambda}{\pi L}$$

which proves the lemma.

□

Lemma 3.4.22. *Let K_1 and K_2 be d.f. with mean 0 whose ch.f. $\mathcal{K}_i$ are integrable*

$$K_1(x) - K_2(x) = (2\pi)^{-1}\int -e^{-itx}\frac{\mathcal{K}_1(t) - \mathcal{K}_2(t)}{it}dt$$

Proof. Since the $\mathcal{K}_i$ are integrable, the inversion formula, Theorem 3.3.11, implies that the density $k_i(x)$ has

$$k_i(y) = (2\pi)^{-1}\int e^{-ity}\mathcal{K}_i(t)dt$$

Subtracting the last expression with $i = 2$ from the one with $i = 1$, then integrating from a to x and letting $\triangle K = K_1 - K_2$ gives

$$\begin{aligned} \triangle K(x) - \triangle K(a) &= (2\pi)^{-1}\int_a^x\int e^{-ity}\{\mathcal{K}_1(t) - \mathcal{K}_2(t)\}dtdy \\ &= (2\pi)^{-1}\int\{e^{-ita} - e^{-itx}\}\frac{K_1(t) - K_2(t)}{it}dt \end{aligned}$$

the application of Fubini's theorem being justified since the $\mathcal{K}_i$ are integrable in t and we are considering a bounded interval in y.

The factor $1/it$ could cause problems near zero, but we have supposed that the K_i have mean 0, so $\{1 - \mathcal{K}_i(t)\}/t \to 0$, and hence $(\mathcal{K}_1(t) - \mathcal{K}_2(t))/it$ is bounded and continuous. The factor $1/it$ improves the integrability for large t so $(\mathcal{K}_1(t) - \mathcal{K}_2(t))/it$ is integrable. Letting $a \to -\infty$ and using the Riemann-Lebesgue lemma proves the result.

□

Let φ_F and φ_G be the ch.f.'s of F and G. Applying Lemma 3.4.22 to $F_L = F \star H_L$ and $G_L = G \star \star H_L$, gives

$$\begin{aligned} \triangle K(x) - \triangle K(a) &= (2\pi)^{-1} \int_a^x \int e^{-ity} \{\mathcal{K}_1(t) - \mathcal{K}_2(t)\} dt dy \\ &= (2\pi)^{-1} \int \{e^{-ita} - e^{-itx}\} \frac{\mathcal{K}_1(t) - \mathcal{K}_2(t)}{it} dt \end{aligned}$$

because $|\omega_L(t)| \leq 1$. Using Lemma 3.4.21 now, we have

$$|F(x) - G(x)| \leq \frac{1}{\pi} \int_{-L}^{L} |\psi_F(\theta) - \psi_G(\theta)| \frac{d\theta}{|\theta|} + \frac{24\lambda}{\pi L}$$

where $\lambda = \sup_x G'(x)$. Plugging in $F = F_n$ and $G = \mathcal{N}$ gives

$$|F_n(x) - \mathcal{N}(x)| \leq \frac{1}{\pi} \int_{-L}^{L} |\psi^n(\theta/\sqrt{n}) - \psi(\theta)| \frac{d\theta}{|\theta|} + \frac{24\lambda}{\pi L} \tag{3.6}$$

and it remains to estimate the right-hand side. This phase of the argument is fairly routine, but there is a fair amount of algebra. To save the reader from trying to improve the inequalities along the way in hopes of getting a better bound, we would like to observe that we have used the fact that $C = 3$ to get rid of the cases $n \leq 9$, and we use $n \geq 10$ in (e).

To estimate the second term in 3.6, we observe that

$$\text{(a) } \sup_x G'(x) = G'(0) = (2\pi)^{-1/2} = 0.39894 < 2/5$$

For the first, we observe that if $|\alpha|$, $|\beta| \leq \gamma$

$$\text{(b) } |\alpha^n - \beta^n| \leq \sum_{m=0}^{n-1} |\alpha^{n-m}\beta^m - \alpha^{n-m-1}\beta m + 1| \leq n|\alpha_\beta|\gamma^{n-1}$$

Using 3.5 now gives (recall we are supposing $\sigma^2 = 1$)

$$\text{(c) } |\varphi(t) - 1 + t^2/2| \leq \rho|t|^3/6$$

so if $t^2 \leq 2$

$$\text{(d) } |\varphi(t)| \leq 1 - t^2/2 + \rho|t|^3/6$$

Let $L = 4\sqrt{n}/3\rho$. If $|\theta| \leq L$, then by (d) and the fact $\rho|\theta|/\sqrt{n} \leq 4/3$

$$\begin{aligned} |\varphi(\theta/\sqrt{n})| &\leq 1 - \theta^2/2n + \rho|\theta|^3/6n^{3/2} \\ &\leq 1 - 5\theta^2/18n \\ &\leq \exp(-5\theta^2/18n) \end{aligned}$$

since $1 - x \leq e^{-x}$. We will now apply (b) with

$$\alpha = \varphi(\theta/\sqrt{n})\ \beta = \exp(-\theta^2/2n)\ \gamma = \exp(-5\theta^2/18n)$$

Since we are supposing $n \geq 10$

$$\text{(e) } \gamma^{n-1} \leq \exp(-\theta^2/4)$$

For the other part of (b), we write

$$n|\alpha - \beta| \leq n|\varphi(\theta/\sqrt{n}) - 1 + \theta^2/2n| + n|1 - \theta^2/2n - \exp(-\theta^2/2n)|$$

To bound the first term on the right-hand side, observe that (c) implies

$$n|\varphi(\theta)/\sqrt{n}) - 1 + \theta^2/2n| \leq \rho|\theta|^3/6n^{1/2}$$

For the second term, note that if $0 < x < 1$, then we have an alternating series with decreasing terms, so

$$|e^{-x} - (1-x)| = \left| -\frac{x^2}{2!} + \frac{x^3}{3!} - \dots \right| \leq \frac{x^2}{2}$$

Taking $x = \theta^2/2n$, it follows that for $|\theta| \leq L \leq \sqrt{2n}|$

$$n|1 - \theta^2/2n - \exp(-\theta^2/2n)| \leq \theta^4/8n$$

Combining this with our estimate on the first term gives

$$\text{(f) } n|\alpha - \beta| \leq \rho|\theta|^3/6n^{1/2} + \theta^4/8n$$

Using (f) and (e) in (b), gives

$$\begin{aligned} \frac{1}{|\theta|}|\varphi^n(\theta/\sqrt{n}) - \exp(-\theta^2/2)| &\leq \exp(-\theta^2/4)\left\{\frac{\rho\theta^2}{6n^{1/2}} + \frac{|\theta|^3}{8n}\right\} \\ &\leq \frac{1}{L}\exp(-\theta^2/4)\left\{\frac{2\theta^2}{9} + \frac{|\theta|^3}{18}\right\} \end{aligned}$$

since $\rho/\sqrt{n} = 4/3L$, and $1/n = 1/\sqrt{n} \cdot 1/\sqrt{n} \leq 4/3L \cdot 1/3$ since $\rho \geq 1$ and $n \geq 10$. Using the last result and (a) in Lemma 3.4.22 gives

$$\pi L|F_n(x) - \mathcal{N}(x)| \leq \int \exp(-\theta^2/4)\left\{\frac{2\theta^2}{9} + \frac{|\theta|^3}{18}\right\}d\theta + 9.6$$

Recalling $L = 4\sqrt{n}/3\rho$, we see that the last result is of the form $|F_n(x) - \mathcal{N}(x)| \leq C\rho/\sqrt{n}$. To evaluate the constant, we observe

$$\int (2\pi a)^{-1/2} x^2 \exp(-x^2/2a)dx = a$$

and writing $x^3 = 2x^2 \cdot x/2$ and integrating by parts

$$\begin{aligned} 2\int_0^\infty x^3 \exp(-x^2/4)dx &= 2\int_0^\infty 4x \exp(-x^2/4)dx \\ &= -16e^{-x^2/4}\Big|_0^\infty = 16 \end{aligned}$$

This gives us

$$|F_n(x) - \mathcal{N}(x)| \leq \frac{1}{\pi} \cdot \frac{3}{4}\left(\frac{2}{9} \cdot 2 \cdot \sqrt{4\pi} + \frac{16}{18} + 9.6\right)\frac{\rho}{\sqrt{n}} < 3\frac{\rho}{\sqrt{n}}$$

For the last step, you have to get your calculator or trust Feller.

□

3.5 Local Limit Theorems

We saw that if $X_1, X_2, \dots$ are i.i.d. with $P(X_1 = 1) = P(X_1 = -1) = 1/2$ and k_n is a sequence of integers with $2k_n/(2n)^{1/2} \to x$, then

$$P(S_{2n} = 2k_n) \sim (\pi n)^{-1/2} \exp(-x^2/2)$$

In this section, we will prove two theorems that generalize the last result. We begin with two definitions. A random variable X has a **lattice distribution** if there are constants b and $h > 0$ so that $P(X \in b + h\mathbb{Z}) = 1$, where $b + h\mathbb{Z} = \{b + hz : z \in \mathbb{Z}\}$.

The largest h for which the last statement holds is called the **span** of the distribution.

Example 3.5.1. If $P(X = 1) = P(X = -1) = 1/2$, then X has a lattice distribution with span 2. When h is 2, one possible choice is $b = -1$.

The next result relates the last definition to the characteristic function. To check (ii) in its statement, note that in last example $E(e^{itX}) = \cos t$ has $|\cos(t)| = 1$ when $t = n\pi$.

Theorem 3.5.2. *Let $\varphi(t) = Ee^{itX}$. There are only three possibilities.*

(i) $|\varphi(t)| < 1$ for all $t \neq 0$.

(ii) There is a $\lambda > 0$ so that $|\varphi(\lambda)| = 1$ and $|\varphi(t)| < 1$ for $0 < t < \lambda$. In this case, X has a lattice distribution with span $2\pi/\lambda$.

(ii) $|\varphi(t)| = 1$ for all t. In this case, $X = b$ a.s. for some b.

Proof. We begin (ii). It suffices to show that $|\varphi(t)| = 1$ if and only if $P(X \in b + (2\pi/t)\mathbb{Z}) = 1$ for some b. First if $P(X \in b + (2\pi/t)\mathbb{Z}) = 1$, then

$$\varphi(t) = Ee^{itX} = e^{itb} \sum_{n \in \mathbb{Z}} e^{i2\pi n} P(X = b + (2\pi/t)n) = e^{itb}$$

Conversely, if $|\varphi(t)| = 1$, then there is equality in the inequality $|Ee^{itX}| \leq E|e^{itX}|$, so the distribution of e^{itX} must be concentrated at some point e^{itb}, and $P(X \in b + (2\pi/t)\mathbb{Z}) = 1$.

To prove trichotomy now, we suppose that (i) and (ii) do not hold, that is, there is a sequence $t_n \downarrow 0$ so that $|\varphi(t_n)| = 1$. The first paragraph shows that there is a b_n so that $P(X \in b_n + (2\pi)/t_n)\mathbb{Z}) = 1$. Without loss of generality, we can pick $b_n \in (-\pi/t_n, \pi/t_n]$. As $n \to \infty$, $P(X \notin (-\pi/t_n, \pi/t_n]) \to 0$, so it follows that $P(X = b_n) \to 1$. This is only possible if $b_n = b$ for $n \geq N$, and $P(X = b) = 1$.

□

We call the three cases in Theorem 3.5.2 (i) **nonlattice**, (ii) **lattice**, and (iii) **degenerate**. The reader should notice that this means that lattice random variables are by definition nondegenerate. Before we turn to the main business of this section, we would like to introduce one more special case. If X is a lattice distribution and we can take $b = 0$, i.e. $P(X \in h\mathbb{Z}) = 1$, then X is said to be **arithmetic**. In this case, if $\lambda = 2\pi/h$ then $\varphi(\lambda) = 1$ and φ is periodic: $\varphi(t + \lambda) = \varphi(t)$.

Our first local limit theorem is for the lattice case. Let $X_1, X_2, ...$ be i.i.d. with $EX_i = 0$, $EX_i^2 = \sigma^2 \in (0, \infty)$, and having a common lattice distribution with span h. If $S_n = X_1 + \cdots + X_n$ and $P(X_i \in b + h\mathbb{Z}) = 1$ then $P(S_n \in nb + h\mathbb{Z}) = 1$. We put

$$\rho_n(x) = P(S_n/\sqrt{n} = x) \text{ for } x \in \mathcal{L}_n = \{(nb + hz)/\sqrt{n} : z \in \mathbb{Z}\}$$

and

$$n(x) = (2\pi\sigma^2)^{-1/2} \exp(-x^2/2\sigma^2) \text{ for } x \in (-\infty, \infty)$$

Theorem 3.5.3. *Under the hypotheses above, as $n \to \infty$*

$$\sup_{x \in \mathcal{L}_n} \left| \frac{n^{1/2}}{h} p_n(x) - n(x) \right| \to 0$$

Remark 3.5.4. To explain the statement, note that if we followed the approach in example, then we would conclude that for $x \in \mathcal{L}_n$,

$$p_n(x) \approx \int_{-x-h/2\sqrt{n}}^{x+h/2\sqrt{h}} n(y)dy \approx \frac{h}{\sqrt{n}} n(x)$$

Proof. Let Y be a random variable with $P(Y \in a+\theta\mathbb{Z}) = 1$ and $\psi(t) = E\exp(itY)$. It follows from part (iii) of exercise that

$$P(Y = x) = \frac{1}{2\pi/\theta} \int_{-\pi/\theta}^{\pi/\theta} e^{-itx}\psi(t)dt$$

Using this formula with $\theta = h/\sqrt{n}$, $\psi(t) = E\exp(itS_n/\sqrt{n}) = \psi^n(t/\sqrt{n})$, and then multiplying each side by $1/\theta$ gives

$$\frac{n^{1/2}}{h} p_n(x) = \frac{1}{2\pi} \int_{-\pi\sqrt{n}/h}^{\pi\sqrt{n}/h} e^{-itx}\varphi^n(t/\sqrt{n})dt$$

Using the inversion formula, Theorem 3.3.13, for $n(x)$, which has ch.f. $\exp(-\sigma^2t^2/2)$, gives

$$n(x) = \frac{1}{2\pi} \int e^{-itx} \exp(-\sigma^2t^2/2)dt$$

Subtracting the last two equations gives (recall $\pi > 1$, $|e^{-itx}| \le 1$)

$$\begin{aligned} \left| \frac{n^{1/2}}{h} p_n(x) - n(x) \right| \quad \le \quad & \int_{-\pi\sqrt{n}/h}^{\pi\sqrt{n}/h} |\varphi^n(t/\sqrt{n}) - \exp(-\sigma^2t^2/2)|dt \\ & + \int_{\pi\sqrt{n}/h}^{\infty} \exp(-\sigma^2t^2/2)dt \end{aligned}$$

The right-hand side is independent of x, so to prove Theorem 3.5.2 it suffices to show that it approaches 0. The second integral clearly $\to 0$. To estimate the first integral, we observe $\varphi^n(t/\sqrt{n}) \to \exp(-\sigma^2t^2/2)$, so the integrand goes to 0 and it is now just a question of "applying the dominated convergence theorem".

To do this, we will divide the integral into three pieces. The bounded convergence theorem implies that for any $A < \infty$ the integral over $(-A, A)$ approaches 0. To estimate the integral over $(-A, A)^c$, we observe that since $EX_i = 0$ and $EX_i^2 = \sigma^2$, formula 3.5 and the triangle inequality imply that

$$|\varphi(u)| \le |1 - \sigma^2u^2/2| + \frac{u^2/2}{E}(\min(|u| \cdot |X|^3, 6|X|^2))$$

The last expected value $\to 0$ as $u \to 0$. This means we can pick $\delta > 0$ so that if $|u| < \delta$, it is $\le \sigma^2/2$ and hence

$$|\varphi(u)| \le 1 - \sigma^2u^2/2 + \sigma^2u^2/4 = 1 - \sigma^2u^2/4 \le \exp(-\sigma^2u^2/4)$$

since $1 - x \le e^{-x}$. Applying the last result to $u = t/\sqrt{n}$, we see that for $t \le \delta\sqrt{n}$

$$(\star)\ |\varphi(t/\sqrt{n})^n| \le \exp(-\sigma^2t^2/4)$$

So the integral over $(-\delta\sqrt{n}, \delta\sqrt{n}) - (-A, A)$ is smaller than

$$2\int_A^{\delta\sqrt{n}} \exp(-\sigma^2t^2/4)dt$$

which is small if A is large.

To estimate the rest of the integral we observe that since x has span h, Theorem 3.5.2 implies $|\varphi(u)| \neq 1$ for $u \in [\delta, \pi/h]$. φ is continuous, so there is an $\eta < 1$ so that $|\varphi(u)| \leq \eta < 1$ for $|u| \in [\delta, \pi/h]$. Letting $u = t/\sqrt{n}$ again, we see that the integral over $[-\pi\sqrt{n}/h, \pi\sqrt{n}/h]$. Letting $u = t/\sqrt{n}$ again, we see that the integral over $[-\pi\sqrt{n}/h, \pi\sqrt{n}/h] - (-\delta\sqrt{n}, \delta\sqrt{n})$ is smaller than

$$2\int_{\delta\sqrt{n}}^{\pi\sqrt{n}/h} \eta^n + \exp(-\sigma^2 t^2/2)dt$$

which $\to 0$ as $n \to \infty$. This completes the proof.

□

We turn now to the nonlattice case. Let $X_1, X_2, ...$ be i.i.d. with $EX_i = 0$, $EX_i^2 = \sigma^2 \in (0, \infty)$, and having a common characteristic function $\varphi(t)$ that has $|\varphi(t)| < 1$ for all $t \neq 0$. Let $S_n = X_1 + \cdots + X_n$ and $n(x) = (2\pi\sigma^2)^{-1/2} \exp(-x^2/2\sigma^2)$.

Theorem 3.5.5. *Under the hypotheses above, if $x_n/\sqrt{n} \to x$ and $a < b$,*

$$\sqrt{n}P(S_n \in (x_n + a, x_n + b)) \to (b - a)n(x)$$

Remark 3.5.6. The proof of this result has to be a little devious because the assumption above does not give us much control over the behavior of φ. For a bad example, let $q_1, q_2, ...$ be an enumeration of the positive rationals that has $q_n \leq n$. Suppose

$$P(X = q_n) = P(X = -q_n) = 1/2^{n+1}$$

In this case $EX = 0$, $EX^2 < \infty$, and the distribution is nonlattice. However, the characteristic function has $\limsup_{t\to\infty} |\varphi(t)| = 1$.

Proof. To tame bad ch.f.'s, we use a trick. Let $\delta > 0$

$$h_0(y) = \frac{1}{\pi} \cdot \frac{1 - \cos \delta y}{\delta y^2}$$

be the density of the Polya's distribution and let $h_\theta(x) = e^{i\theta x} h_0(x)$. If we introduce the Fourier transform

$$\hat{g}(u) = \int e^{iuy} g(y)dy$$

then it follows that

$$\hat{h}_0(u) = \begin{cases} 1 - |u/\delta| & \text{if } |u| \leq \delta \\ 0 & \text{otherwise} \end{cases}$$

and it is easy to see that $\hat{h}_\theta(u) = \hat{h}_0(u + \theta)$. We will show that for any θ

$$\text{(a) } \sqrt{n}Eh_\theta(S_n - x_n) \to n(x) \int h_\theta(y)dy$$

Before proving (a), we will show it implies Theorem 3.5.5. Let

$$\mu_n(A) = \sqrt{n}P(S_n - x_n \in A), \text{ and } \mu(A) = n(x)|A|$$

where $|A|$ = the Lebesgue measure of A. Let

$$\alpha_n = \sqrt{n}Eh_0(S_n - x_n) \text{ and } \alpha = n(x) \int h_0(y)dy = n(x)$$

Finally, define probability measures by

$$\nu_n(B) = \frac{1}{\alpha_n}\int_B h_0(y)\mu_n(dy), \text{ and } \nu(B) = \frac{1}{\alpha}\int_B h_0(y)\mu(dy)$$

Taking $\theta = 0$ in (a) we see $\alpha_n \to \alpha$ and so (a) implies

$$\text{(b)} \int e^{i\theta y}\nu_n(dy) \to \int e^{i\theta y}\nu(dy)$$

Since this holds for all θ, it follows from Theorem 3.3.16 that $\nu_n \Rightarrow \nu$. Now if $|a|$, $|b| < 2\pi/\delta$, then the function

$$k(y) = \frac{1}{h_0(y)} \cdot 1_{(a,b)}(y)$$

is bounded and continuous a.s. with respect to ν so it follows from Theorem 3.2.9 that

$$\int k(y)\nu_n(dy) \to \int k(y)\nu(dy)$$

Since $\alpha_n \to \alpha$, this implies

$$\sqrt{n}P(S_n \in (x_n + a, x_n + b)) \to (b - a)n(x)$$

which is the conclusion of Theorem 3.5.5

Turning now to the proof of (a), the inversion formula, Theorem 3.3.13, implies

$$h_0(x) = \frac{1}{2\pi}\int e^{-iux}h_0(u)du$$

Recalling the definition of h_θ, using the last result, and changing variables $u = \nu+\theta$, we have

$$\begin{aligned} h_\theta(x) = e^{i\theta x}h_0(x) &= \frac{1}{2\pi}\int e^{-i(\alpha-\theta)x}\hat{h}_0(u)du \\ &= \frac{1}{2\pi}\int e^{-i\nu x}\hat{h}_\theta(\nu)d\nu \end{aligned}$$

since $\hat{h}_\theta(\nu) = \hat{h}_0(\nu + \theta)$. Letting F_n be the distribution of $S_n - x_n$ and integrating gives

$$\begin{aligned} Eh_\theta(S_n - x_n) &= \frac{1}{2\pi}\int\int e^{-iux}\hat{h}_\theta(u)dudF_n(x) \\ &= \frac{1}{2\pi}\int\int e^{-iux}dF_n(x)\hat{h}_\theta(u)du \end{aligned}$$

by Fubini's theorem. (Recall that $\hat{h}_\theta(u)$ has compoact support and F_n is a distribution function.) Using (e) of Theorem 3.3.1, we see that the last expression

$$= \frac{1}{2\pi}\int \varphi(-u)^n e^{iux_n}\hat{h}_\theta(u)du$$

To take the limist as $n \to \infty$ of this integral, let $[-M, M]$ be an interval with $\hat{h}_\theta(u) = 0$ for $u \notin [-M, M]$. By $(\star)$ above, we can pick δ so that for $|u| < \delta$

$$\text{(c)} \ |\varphi(u)| \le \exp(-\sigma^2u^2/4)$$

Let $I = [-\delta, \delta]$ and $J = [-M, M] - I$. Since $|\varphi(u)|, 1$ for $u \neq 0$ and φ is continuous, there is a constant $\eta < 1$ so that $|\varphi(u)| \le \eta < 1$ for $u \in J$. Since $|\hat{h}_\theta(u)| \le 1$, this implies that

$$\left|\frac{\sqrt{n}}{2\pi}\int_J \varphi(-u)^n e^{i\alpha x_n}\hat{h}_\theta(u)du\right| \le \frac{\sqrt{n}}{2\pi}\cdot 2M\eta^n \to 0$$

as $n\to\infty$. For the integral over I, change variables $u = t/\sqrt{n}$ to get

$$\frac{1}{2\pi}\int_{-\delta\sqrt{n}}^{\delta\sqrt{n}} \varphi(-t/\sqrt{n})^n e^{itx_n/\sqrt{n}}\hat{h}_\theta(t/\sqrt{n})dt$$

The central limit theorem implies $\varphi(-t/\sqrt{n})^n \to \exp(-\sigma^2 3t^2/2)$. Using (c) now and the dominated convergence theorem gives (recall $x_n/\sqrt{n}\to x$)

$$\begin{aligned}\frac{\sqrt{n}}{2\pi}\int_I \varphi(-u)^n e^{iux_n}\hat{h}_\theta(u)du &\to \frac{1}{2\pi}\int \exp(-\sigma^2t^2/2)e^{itx}\hat{h}_\theta(0)dt\\ &= n(x)\hat{h}_\theta(0)\\ &= n(x)\int h_\theta(y)dy\end{aligned}$$

by the inversion formula, Theorem 3.3.13, and the definition of $\hat{h}_\theta(0)$. This proves (a) and completes the proof of Theorem 3.5.5.

3.6 Poisson Convergence

3.6.1 The Basic Limit Theorem

Our first result is sometimes facetiously called the "weak law of small numbers" or the "law of rare events." These names derive from the fact that the Poisson appears as the limit of a sum of indicators of events that have small probabilities.

Theorem 3.6.1. *For each n, let $X_{n,m}$, $1\le m\le n$ be independent random variables with $P(X_{n,m}=1)=p_{n,m}$, $P(X_{n,m}=0)=1-p_{n,m}$. Suppose*
(i) $\sum_{m=1}^n p_{n,m}\to\lambda\in(0,\infty)$, and
(ii) $\max_{1\le m\le n} p_{n,m}\to 0$. If $S_n = X_{n,1}+\cdots+X_{n,n}$ then $S_n \Rightarrow Z$ where Z is Poisson(λ).

Here Poisson(λ) is shorthand for Poisson distribution with mean λ, that is,

$$P(Z=k) = e^{-\lambda}\lambda^k/k!$$

Note that in the spirit of the Lindeberg-Feller theorem, no single term contributes very much to the sum. In contrast to that theorem, the contributions, when positive, are not small.

First proof. Let $\varphi_{n,m}(t) = E(\exp(itX_{n,m})) = (1-p_{n,m}) + p_{n,m}e^{it}$. Then

$$E\exp(itS_n) = \prod_{m=1}^n (1+p_{n,m}(e^{it}-1))$$

Let $0\le p\le 1$. $|\exp(p(e^{it}-1))| = \exp(p\mathrm{Re}(e^{it}-1)) \le 1$ and $|1+p(e^{it}-1)|\le 1$ since it is on the line segment connecting 1 to e^{it}. Using Lemma 3.4.3 with $\theta=1$ and then Lemma 3.4.4, which is valid when $\max_m p_{n,m}\le 1/2$ since $|e^{it}-1|\le 2$,

$$\begin{aligned}&\left|\exp\left(\sum_{m=1}^n p_{n,m}(e^{it}-1)\right) - \prod_{m=1}^n\{1+p_{n,m}(e^{it}-1)\}\right|\\ \le\ &\sum_{m=1}^n |\exp(p_{n,m}(e^{it}-1)) - \{1+p_{n,m}(e^{it}-1)\}|\\ \le\ &\sum_{m=1}^n p_{n,m}^2|e^{it}-1|^2\end{aligned}$$

Using $|e^{it} - 1| \leq 2$ again, it follows that the last expression

$$\leq 4\left(\max_{1\leq m\leq n} p_{n,m}\right)\sum_{m=1}^{n} p_{n,m} \to 0$$

by assumptions (i) and (ii). The last conclusion and $\sum_{m=1}^{n} p_{n,m} \to \lambda$ imply

$$E\exp(itS_n) \to \exp(\lambda(e^{it} - 1))$$

To complete the proof now, we consult example for the ch.f. of the Poisson distribution and apply Theorem 3.2.8

□

We will now consider some concrete situations in which Theorem 3.6.1 can be applied. In each case we are considering a situation in which $p_{n,m}c/n$, so we approximate the distribution of the sum by a Poisson with mean c.

Example 3.6.2. In a calculus class with 400 students, the number of students who have their birthday on the day of the final exam has approximately a Poisson distribution with mean $400/365 = 1.096$. This means that the probability no one was born on that date is about $e^{-1.096} = 0.334$. Similar reasoning shows that the number of babies born on a given day or the number of people who arrive at a bank between 1:15 and 1:30 should have a Poisson distribution.

Example 3.6.3. Suppose we roll two dice 36 times. The probability of "double ones" (one on each die) is 1/36, so the number of times this occurs should have approximately a Poisson distribution with mean 1. Comparing the Poisson approximation with exact probabilities shows that the agreement is good even though the number of trials is small.

k	0	1	2	3
Poisson	0.3678	0.3678	0.1839	0.0613
exact	0.3627	0.3730	0.1865	0.0604

After we give the second proof of Theorem 3.6.1, we will discuss rates of convergence. Those results will show that for large n the largest discrepancy occurs for $k = 1$ and is about $1/2en$ (=0.0051 in this case).

Example 3.6.4. Let $\xi_{n,1}, ..., \xi_{n,n}$ be independent and uniformly distributed over $[-n, n]$. Let $X_{n,m} = 1$ if $\xi_{n,m} \in (a, b)$, $= 0$ otherwise. S_n is the number of points that land in (a, b). $p_{n,m} = (b-a)/2n$ so $\sum_m p_{n,m} = (b-a)/2$. This shows that (i) and (ii) in Theorem 3.6.1 hold, and we conclude that $S_n \Rightarrow Z$, a Poisson r.v. with mean $(b-a)/2$. A two-dimensional version of the last theorem might explain why a Poisson distribution, As Feller, Vol. I (1968), pp. 160-161 reports [7], the area was divided into 576 areas of 1/4 square kilometers each. The total number of hits was 537 for an average of 0.9323 per cell. The following table compares N_k the number of cells with k hits with the predictions of the Poisson approximation.

k	0	1	2	3	4	≥ 5
N_k	229	211	93	35	7	1
Poisson	226.74	211.39	98.54	30.62	7.14	1.57

For other observations fitting a Poisson distribution, see Feller, Vol. I (1968), Section VI.7 [7].

Our second proof of Theorem 3.6.1 requires a little more work but provides information about the rate of convergence. We begin by defining the **total variation distance** between two measures on a countable set S.

$$||\mu - \nu|| \equiv \frac{1}{2}\sum_z |\mu(z) - \nu(z)| = \sup_{A \subset S} |\mu(A) - \nu(A)|$$

The first equality is a definition. To prove the second, note that for any A

$$\sum_z |\mu(z) - \nu(z)| \geq |\mu(A) - \nu(A)| + |\mu(A^c) - \nu(A^c)| = 2|\mu(A) - \nu(A)|$$

and there is equality when $A = \{z : \mu(z) \geq \nu(z)\}$.

Lemma 3.6.5. *If $\mu_1 \times \mu_2$ denotes the product measure on $\mathbb{Z} \times \mathbb{Z}$ that has $(\mu_1 \times \mu_2)(x, y) = \mu_1(x)\mu_2(y)$, then*

$$||\mu_1 \times \mu_2 - \nu_1 \times \nu_2|| \leq ||\mu_1 - \nu_1|| + ||\mu_2 - \nu_2||$$

Proof.

$$\begin{aligned} 2||\mu_1 \times \mu_2 - \nu_1 \times \nu_2|| &= \sum_{x,y} |\mu_1(x)\mu_2(y) - \nu_1(x)\nu_2(y)| \\ &= \sum_y \mu_y(y) \sum_x |\mu_1(x) - \nu_1(x)| + \sum_x \nu_1(x) \sum_y |\mu_2(y) - \nu_2(y)| \\ &= 2||\mu_1 - \nu_1|| + 2||\mu_2 - \nu_2|| \end{aligned}$$

which gives the desired result.

□

Lemma 3.6.6. *If $\mu_1 * \mu_2$ denotes the convolution of μ_1 and μ_2, that is,*

$$\mu_1 * \mu_2(x) = \sum_y \mu_1(x - y)\mu_2(y)$$

then $||\mu_1 * \mu_2 - \nu_1 * \nu_2|| \leq ||\mu_1 \times \mu_2 - \nu_1 \times \nu_2||$.

Proof.

$$\begin{aligned} 2||\mu_1 * \mu_2 - \nu_1 * \nu_2|| &= \textstyle\sum_x |\sum_y \mu_1(x - y)\mu_y(y) - \sum_y \nu_1(x - y)\nu_2(y)| \\ &\leq \textstyle\sum_x \sum_y |\mu_1(x - y)\mu_2(y) - \nu_1(x - y)\nu_2(y)| \\ &= 2||\mu_1 \times \mu_2 - \nu_1 \times \nu_2|| \end{aligned}$$

which gives the desired result.

□

Lemma 3.6.7. *Let μ be the measure with $\mu(1) = p$ and $\mu(0) = 1 - p$. Let ν be a Poisson distribution with mean p. Then $||\mu - \nu|| \leq p^2$.*

Proof.

$$\begin{aligned} 2||\mu - \nu|| &= |\mu(0) - \nu(0)| + |\mu(1) - \nu(1)| + \textstyle\sum_{n \geq 2} \nu(n) \\ &= |1 - p - e^{-p}| + |p - pe^{-p}| + 1 - e^{-p}(1 + p) \end{aligned}$$

since $1 - x \leq e^{-x} \leq 1$ for $x \geq 0$, the above

$$\begin{aligned} &= e^{-p} - 1 + p + p(1 - e^{-p}) + 1 - e^{-p} - pe^{-p} \\ &= 2p(1 - e^{-p}) \leq 2p^2 \end{aligned}$$

which gives the desired result.

□

Second proof of Theorem 3.6.1. Let $\mu_{n,m}$ be the distribution of $X_{n,m}$. Let μ_n be the distribution of S_n. Let $\nu_{n,m}$, ν_n, and ν be Poisson distributions with means $p_{n,m}$, $\lambda_n = \sum_{m\leq n} p_{n,m}$, and λ, respectively. Since $\mu_n = \mu_{n,1} * \cdots * \mu_{n,n}$ and $\nu_n = \nu_{n,1} * \cdots * \nu_{n,n}$, Lemmas 3.6.5 3.6.6, and 3.6.7 imply

$$||\mu_n - \nu_n|| \leq \sum_{m=1}^{n} ||\mu_{n,m} - \nu_{n,m}|| \leq 2\sum_{m=1}^{n} p_{n,m}^2 \tag{3.7}$$

Using the definition of total variation distance now gives

$$\sup_A |\mu_n(A) - \nu_n(A)| \leq \sum_{m=1}^{n} p_{n,m}^2$$

Assumptions (i) and (ii) imply that the right-hand side $\to 0$. Since $\nu_n \Rightarrow \nu$ as $n \to \infty$, the result follows.

□

Remark 3.6.8. The proof above is due to Hodges and Le Cam (1960) [12]. By different methods, C. Stein (1987) (see (43) on p.89) [14] has proved

$$\sup_A |\mu_n(A) - \nu_n(A)| \leq (\lambda \vee 1)^{-1} \sum_{m=1}^{n} p_{n,m}^2$$

Rates of convergence. When $p_{n,m} 1/n$, 3.7 becomes

$$\sup_A |\mu_n(A) - \nu_n(A)| \leq 1/n$$

To assess the equality of this bound, we will compare the Poisson and binomial probabilities for k successes.

k	Poisson	Binomial
0	e^{-1}	$(1-\frac{1}{n})^n$
1	e^{-1}	$n \cdot n^{-1}(1-\frac{1}{n})^{n-1} = (1-\frac{1}{n})^{n-1}$
2	$e^{-1}/e!$	$\binom{n}{2} n^{-2}(1-\frac{1}{n})^{n-2} = (1-\frac{1}{n})^{n-1}/2!$
3	$e^{-1}/3!$	$\binom{n}{3} n^{-3}(1-\frac{1}{n})^{n-3} = (1-\frac{2}{n})(1-\frac{1}{n})^{n-2}/3!$

Since $(1-x) \leq e^{-x}$, we have $\mu_n(0) - \nu_n(0) \leq 0$. Expanding

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots$$

gives

$$(n-1)\log\left(1-\frac{1}{n}\right) = -\frac{n-1}{n} - \frac{n-1}{2n^2} - \cdots = -1 + \frac{1}{2n} + O(n^{-2})$$

So

$$n\left((1-\frac{1}{n})^{n-1} - e^{-1}\right) = ne^{-1}(\exp\{1/2n + O(n^{-2}\} - 1) \to e^{-1}/2$$

and it follows that

$$n(\mu_n(1) - \nu_n(1)) \to e^{-1}/2$$

$$n(\mu_n(2) - \nu_n(2)) \to e^{-1}/4$$

For $k \geq 3$, using $(1-2/n) \leq (1-1/n)^2$ and $(1-x) \leq e^{-x}$ shows $\mu_n(k) - \nu_n(k) \leq 0$, so

$$\sup_{A\subset\mathbb{Z}}|\mu_n(A)-\nu_n(A)|\approx 3/4en$$

There is a large literature on Poisson approximations for dependent events. Here we consider the following.

3.6.2 Two Examples with Dependence

Example 3.6.9. Let π be a random permutation of $\{1,2,...,n\}$, let $X_{n,m}=1$ if m is a fixed point (0 otherwise), and let $S_n=X_{n,1}+\cdots+X_{n,n}$ be the number of fixed points. We want to compute $P(S_n=0)$. Let $A_{n,m}=\{X_{n,m}=1\}$. The inclusion-exclusion formula implies

$$\begin{aligned} P(\cup_{m=1}^n A_m) &= \sum_m P(A_m)-\sum_{l<m}P(S_l\cap A_m)+\sum_{k<l<m}P(A_k\cap A_l\cap A_m)-\dots \\ &= n\cdot\frac{1}{n}-\binom{n}{2}\frac{(n-2)!}{n!}+\binom{n}{3}\frac{(n-3)!}{n!}-\dots \end{aligned}$$

since the number of permutations with k specified fixed points is $(n-k)!$. Canceling some factorials gives

$$P(S_n>0)=\sum_{m=1}^n\frac{(-1)^{m-1}}{m!}\text{ so }P(S_n=0)=\sum_{m=0}^n\frac{(-1)^m}{m!}$$

Recognizing the second sum as the first $n+1$ terms in the expansion of e^{-1} gives

$$\begin{aligned} |P(S_n=0)-e^{-1}| &= \Big|\sum_{m=n+1}^\infty\frac{(-1)^m}{m!}\Big| \\ &\le \frac{1}{(n+1)!}\Big|\sum_{k=0}^\infty(n+2)^{-k}\Big| \\ &= \frac{1}{(n+1)!}\cdot\left(1-\frac{1}{n+2}\right)^{-1} \end{aligned}$$

a much better rate of convergence than $1/n$. To compute the other probabilities, we observe that by considering the locations of the fixed points

$$\begin{aligned} P(S_n=k) &= \binom{n}{k}\frac{1}{n(n-1)\dots(n-k+1)}P(S_{n-k}=0) \\ &= \frac{1}{k!}P(S_{n-k}=0)\to e^{-1}/k! \end{aligned}$$

Example 3.6.10. Occupancy problem. Suppose that r balls are placed at random into n boxes. It follows from the Poisson approximation to the binomial that if $n\to\infty$ and $r/n\to c$, then the number of balls in a given box will approach a Poisson distribution with mean c. The last observation should explain why the fraction of empty boxes approached e^{-c} in example. Here we will show:

Theorem 3.6.11. *If $ne^{-r/n}\to\lambda\in[0,\infty)$ the number of empty boxes approaches a Poisson distribution with mean λ.*

Proof. To see where the answer comes from, notice that in the Poisson approximation the probability that a given box is empty is $e^{-r/n}\approx\lambda/n$, so if the occupancy of the various boxes were independent, the result would follow from Theorem 3.6.1. To prove the result, we begin by observing

$$P(\text{boxes } i_1, i_2, ..., i_k \text{ are empty}) = \left(1 - \frac{k}{n}\right)^r$$

If we let $p_m(r,n)$ = the probability exactly m boxes are empty when r balls are put in n boxes, then $P(\text{no empty box}) = 1 - P(\text{at least one empty box})$. So by inclusion-exclusion

$$\text{(a) } p_0(r,n) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \left(1 - \frac{k}{n}\right)^r$$

By considering the locations of the empty boxes

$$\text{(b) } p_m(r,n) = \binom{n}{m} \left(1 - \frac{m}{n}\right)^r p_0(r, n-m)$$

To evaluate the limit of $p_m(r,n)$ we begin by showing that if $ne^{-r/n} \to \lambda$ then

$$\text{(c) } \binom{n}{m} (1 - \frac{m}{n})^r \to \lambda^m / m!$$

One half of this is easy. Since $(1-x) \le e^{-x}$ and $ne^{-r/n} \to \lambda$

$$\text{(d) } \binom{n}{m} (1 - \frac{m}{n})^r \le \frac{n^m}{m!} e^{-mr/n} \to \lambda^m / m!$$

For the other direction, observe $\binom{n}{m} \ge (n-m)^m / m!$ so

$$\binom{n}{m} (1 - \frac{m}{n})^r \ge (1 - \frac{m}{n})^{m+r} n^m / m!$$

Now $(1 - m/n)^m \to 1$ as $n \to \infty$ and $1/m!$ is a constant. To deal with the rest, we note that if $0 \le t \le 1/2$ then

$$\begin{aligned} \log(1-t) &= -t - t^2/2 - t^3/3 \dots \\ &\ge -t^2 - \tfrac{t^2}{2}(1 + 2^{-1} + 2^{-2} + \dots) \\ &= -t - t^2 \end{aligned}$$

so we have

$$\log(n^m (1 - \frac{m}{n})^r) \ge m \log n - rm/n - r(m/n)^2$$

Our assumption $ne^{-r/n} \to \lambda$ means

$$r = n \log n - n \log \lambda + o(n)$$

so $r(m/n)^2 \to 0$. Multiplying the last display by m/n and rearranging gives $m \log n - rm/n \to m \log \lambda$. Combining the last two results shows

$$\inf_{n \to \infty} n^m (1 - \frac{m}{n})^r \ge \lambda^m$$

and (c) follows. From (a), (c), and the dominated convergence theorem (using (d) to get the domination), we get

$$\text{(e) if } ne^{-r/n} \to \lambda \text{ then } p_0(r,n) \to \sum_{k=0}^{\infty} (-1)^k \frac{\lambda^k}{k!} = e^{-\lambda}$$

For fixed m, $(n-m)e^{-r/(n-m)} \to \lambda$, so it follows from (c) that $p_0(r, n-m) \to e^{-\lambda}$. Combining this with (b) and (c) completes the proof.

□

Example 3.6.12. Coupon collector's problem. Let $X_1, X_2, ...$ be i.i.d. uniform on $\{1, 2, ..., n\}$ and $T_n = \inf\{m : \{X_1, ..., X_m\} = \{1, 2, ..., n\}\}$. Since $T_n \leq m$ if and only if m balls fill up all n boxes, it follows from Theorem 3.6.11 that

$$P(T_n - n \log n \leq nx) \to \exp(e^{-x})$$

Proof. If $r = n \log n + nx$ then $ne^{-r/n} \to e^{-x}$.

□

Note that T_n is the sum of n independent random variables, but T_n does not converge to the normal distribution. The problem is that the last few terms in the sum are of order n, so the hypotheses of the Lindeberg-Feller theorem are not satisfied.

For a concrete instance of the previous result consider: What is the probability that in a village of $2190 (= 6 \cdot 365)$ people all birthdays are represented? Do you think the answer is much different for $1825 (= 5 \cdot 365)$ people?

Solution. Here $n = 365$, so $365 \log 365 = 2153$ and

$$\begin{aligned} P(T_{365} \leq 2190) &= P((T_{365} - 2153)/365 \leq 37/365) \\ &\approx \exp(-e^{-0.1014}) \\ &= \exp(-0.9036) \\ &= 0.4051 \end{aligned}$$

$$\begin{aligned} p(T_{365} \leq 1825) &= P((T_{365} - 2153)/365 \leq -328/365) \\ &\approx \exp(-e^{0.8986}) \\ &= \exp(-2.4562) \\ &= 0.085 \end{aligned}$$

As we observed, if we let

$$\tau_k^n = \inf\{m : |\{X_1, ..., X_m\}| = k\}$$

then $\tau_1^n = 1$ and for $2 \leq k \leq n$, $\tau_k^n - \tau_{k-1}^n$ are independent and have a geometric distribution with parameter $1 - (k-1)/n$.

3.6.3 Poisson Processes

Theorem 3.6.1 generalizes trivially to give the following result.

Theorem 3.6.13. *Let $X_{n,m}$, $1 \leq m \leq n$ be independent non-negative integer valued random variables with $P(X_{n,m} = 1) = p_{n,m}$, $P(X_{n,m} \geq 2) = \epsilon_{n,m}$.*
(i) $\sum_{m=1}^n p_{n,m} \to \lambda \in (0, \infty)$,
(ii) $\max_{1 \leq m \leq n} p_{n,m} \to 0$, and
(iii) $\sum_{m=1}^n \epsilon_{n,m} \to 0$.
If $S_n = X_{n,1} + \cdots + X_{n,n}$ then $S_n \Rightarrow Z$ where Z is Poisson(λ).

Proof. Let $X'_{n,m} = 1$ if $X_{n,m} = 1$, and 0 otherwise. Let $S'_n = X'_{n,1} + \cdots + X'_{n,n}$. (i) - (ii) and Theorem 3.6.1 imply $S'_n \Rightarrow Z$, (iii) tells us $P(S_n \neq S'_n) \to 0$, and the result follows from the converging together lemma.

□

The next result, which uses Theorem 3.6.13, explains why the Poisson distribution comes up so frequently in applications. Let $N(s,t)$ be the number of arrivals at a bank or an ice cream parlor in the time intervals $(s,t]$. Suppose

(i) the numbers of arrivals in disjoint intervals are independent,

(ii) the distribution of $N(s,t)$ only depends on $t-s$,

(iii) $P(N(0,h)=1)=\lambda h+o(h)$, and

(iv) $P(N(0,h)\geq 2)=o(h)$.

Here, the two $o(h)$ stand for functions $g_1(h)$ and $g_2(h)$ with $g_i(h)/h\to 0$ as $h\to 0$.

Theorem 3.6.14. *If (i) - (iv) hold, then $N(0,t)$ has a Poisson distribution with mean λt.*

Proof. Let $X_{n,m}=N((m-1)t/n,mt/n)$ for $1\leq m\leq n$ and apply Theorem 3.6.13.

□

A family of random variables N_t, $t\geq 0$ satisfying

(i) if $0=t_0<t_1<\cdots<t_n$, $N(t_k)-N(t_{k-1})$, $1\leq k\leq n$ are independent,

(ii) $N(t)-N(s)$ is Poisson$(\lambda(t-s))$,

is called a **Poisson process with rate** λ. To understand how N_t behaves, it is useful to have another method to construct it. Let $\xi_1,\xi_2,...$ be independent random variables with $P(\xi_i>t)=e^{-\lambda t}$ for $t\geq 0$. Let $T_n=\xi_1+\cdots+\xi_n$ and $N_t=\sup\{n: T_n\leq t\}$ where $T_0=0$. In the language of renewal theory (see Theorem 2.4.7, T_n is the time of the nth arrival and N_t is the number of arrivals by time t. To check that N_t is a Poisson process, we begin by recalling (Theorem 2.1.16):

$$f_{T_n}(s)=\frac{\lambda^n s^{n-1}}{(n-1)!}e^{-\lambda s} \text{ for } s\geq 0$$

that is, the distribution of T_n has a density given by the right-hand side. Now and for $n\geq 1$

$$\begin{aligned} P(N_t=n)=P(T_n\leq t<T_{n+1} &= \int_0^t P(T_n=s)P(\xi_{n+1}>t-s)ds \\ &= \int_0^t \frac{\lambda^n s^{n-1}}{(n-1)!}e^{-\lambda s}e^{-\lambda(t-s)}dx \\ &= e^{-\lambda t}\tfrac{(\lambda t)^n}{n!} \end{aligned}$$

The last two formulas show that N_t has a Poisson distribution with mean λt. To check that the number of arrivals in disjoint intervals is independent, we observe

$$P(T_{n+1}\geq u|N_t=n)=P(T_{n+1}\geq u,T_n\leq t)/P(N_t=n)$$

To compute the numerator, we observe

$$\begin{aligned} P(T_{n+1}\geq u,T_n\leq t) &= \int_0^t f_{T_n}(s)P(\xi_{n+1}\geq u-s)ds \\ &= \int_0^t \frac{\lambda^n s^{n-1}}{(n-1)!}e^{-\lambda s}e^{-\lambda(u-s)}ds \\ &= e^{-\lambda u}\tfrac{(\lambda t)^n}{n!} \end{aligned}$$

The denominator is $P(N_t=n)=e^{-\lambda t}(\lambda t)^n/n!$, so

$$P(T_{n+1} \geq u | N_t = n) = e^{-\lambda u}/e^{-\lambda t} = e^{-\lambda(u-t}$$

or, rewriting things, $P(T_{n+1} - t \geq s | N_t = n) = e^{-\lambda s}$. Let $T_1' = T_{N(t)+1} - t$, and $T_k' = T_{N(t)+k} - T_{N(t)+k-1}$ for $k \geq 2$. The last computation shows that T_1' is independent of N_t. If we observe that

$$P(T_n \leq t, T_{n+1} \geq u, T_{n+k} - T_{n+k-1} \geq \nu_k, k = 2, ..., K) = P(T_n \leq t, T_{n+1} \geq u) \prod_{k=1}^{K} P(\xi_{n+k} \geq \nu_k)$$

then it follows that

(a) $T_1', T_2', ...$ are i.i.d. and independent of N_t. The last observation shows that the arrivals after time t are independent of N_t and have the same distribution as the original sequence. From this it follows easily that

(b) If $0 = t_0 < t_1 \cdots < t_n$ then $N(t_i) - N(t_{i-1})$, $i = 1, ..., n$ are independent. To see this, observe that the vector $(N(t_2) - N(t_1), ..., N(t_n) - N(t_{n-1}))$ is $\sigma(T_k', k \geq 1)$ measurable and hence is independent of $N(t_1)$. Then use induction to conclude

$$P(N(t_i) - N(t_{i-1}) = k_i, i = 1, 2, ..., n) = \prod_{i=1}^{n} \exp(-\lambda(t_i - t_{i-1})) \frac{\lambda(t_i - t_{i-1})^{k_i}}{k_i!}$$

Remark 3.6.15. The key to the proof of (a) is the lack of memory property of the exponential distribution:

$$(*)\ P(T > t + s | T > t) = P(T > s)$$

which implies that the location of the first arrival after t is independent of what occurred before time t and has an exponential distribution.

4 RANDOM WALKS

Go back to Table of Contents. Please click TOC

Let $X_1, X_2, ...$ be i.i.d. taking values in $\mathbb{R}^d$ and let $S_n = X_1 + \cdots + X_n$. S_n is a **random walk**. In the last chapter, we were primarily concerned with the distribution of S_n. In this one, we will look at properties of the sequence $S_1(\omega), S_2(\omega), ...$ For example, does the last sequence return to (or near) infinitely often? The first section introduces stopping times, a concept that will be very important in this and the next two chapters. After the first section, the remaining three can be read in any order or skipped without much loss.

4.1 Stopping Times

Most of the results in this section are valid for i.i.d. X's taking values in some nice measurable space $(S, \mathcal{S})$ and will be proved in that generality. For several reasons, it is convenient to use the special probability space from the proof of Kolmogorov's extension theorem:

$$\Omega = \{(\omega_1, \omega_2, ...) : \omega_i \in S\}$$

$$\mathcal{F} = \mathcal{S} \times \mathcal{S} \times \dots$$

$$P = \mu \times \mu \times \dots \mu \text{ is the distribution of } X_i$$

$$X_n(\omega) = \omega_n$$

Throughout this section, we will suppose (w.l.o.g.) that our random variables are constructed on this special space.

Before taking up our main topic, we will prove a 0-1 law that, in the i.i.d. case, generalizes Kolmogorov's. To state the new 0-1 law we need two definitions. A **finite permutation** of $\mathbf{N} = \{1, 2, ...\}$ is a map π from $\mathbf{N}$ onto $\mathbf{N}$ so that $\pi(i) \neq i$ for only finitely many i. If π is a finite permutation of $\mathbf{N}$ and $\omega \in S^{\mathbf{N}}$ we define $(\pi\omega)_i = \omega_{\pi(i)}$. In words, the coordinates of ω are rearranged according to π. Since $X_i(\omega) = \omega_i$ this is the same as rearranging the random variables. An event A is **permutable** if $\pi^{-1}A \equiv \{\omega : \pi\omega \in A\}$ is equal to A for any finite permutation π, or in other words, if its occurrence is not affected by rearranging finitely many of the random variables. The collection of permutable events is a σ-field. It is called the **exchangeable** σ-field and denoted by $\mathcal{E}$.

To see the reason for interest in permutable events, suppose $S = \mathbf{R}$ and let $S_n(\omega) = X_1(\omega) + \cdots + X_n(\omega)$. Two examples of permutable events are

(i) $\{\omega : S_n(\omega) \in B \text{ i.o.}\}$ (ii) $\{\omega : \limsup_{n\to\infty} S_n(\omega)/c_n \geq 1\}$

In each case, the event is permutable because $S_n(\omega) = S_n(\pi\omega)$ for large n. The list of examples can be enlarged considerably by observing:

(iii) All events in the tail σ-field $\mathcal{T}$ are permutable.

To see this, observe that if $A \in \sigma(X_{n+1}, X_{n+2}, ...)$ then the occurrence of A is unaffected by a permutation of $X_1, ..., X_n$. (i) shows that the converse of (iii) is false. The next result shows that for an i.i.d. sequence there is no difference between $\mathcal{E}$ and $\mathcal{T}$. They are both trivial.

Theorem 4.1.1. ***Hewitt-Savage 0-1 Law.*** *If $X_1, X_2, ...$ are i.i.d. and $A \in \mathcal{E}$ then $P(A) \in \{0, 1\}$.*

Proof. Let $A \in \mathcal{E}$. As in the proof of Kolmogorov's 0-1 law, we will show A is independent of itself, i.e., $P(A) = P(A \cap A) = P(A)P(A)$ so $P(A) \in \{0, 1\}$. Let $A_n \in \sigma(X_1, ..., X_n)$ so that

$$(a) P(A_n \triangle A) \to 0$$

Here $A \triangle B = (A - B) \cup (B - A)$ is the symmetric difference. The existence of the A_n's is proved in part ii of Lemma A.2.1 in text [4]. A_n can be written as $\{\omega : (\omega_1, ..., \omega_n) \in B_n\}$ with $B_n \in \mathcal{S}^n$. Let

$$\pi(j) = \begin{cases} j + n & \text{if } 1 \le j \le n \\ j - n & \text{if } n + 1 \le j \le 2n \\ j & \text{if } j \ge 2n + 1 \end{cases}$$

Observing that π^2 is the identity (so we do not have to worry about whether to write π or π^{-1}) and the coordinates are i.i.d. (so the permuted coordinates are) gives

$$(b) P(\omega : \omega \in A_n \triangle A) = P(\omega : \pi\omega \in A_n \triangle A)$$

Now $\{\omega : \pi\omega \in A\} = \{\omega : \omega \in A\}$, since A is permutable, and

$$\{\omega : \pi\omega \in A_n\} = \{\omega : (\omega_{n+1}, ..., \omega_{2n}) \in B_n\}$$

If we use A'_n to denote the last event then we have

$$(c) \{\omega : \pi\omega \in A_n \triangle A\} = \{\omega : \omega \in A'_n \triangle A\}$$

Combining (b) and (c) gives

$$(d) P(A_n \triangle A) = P(A'_n \triangle A)$$

It is easy to see that

$$|P(B) - P(C)| \le |P(B \triangle C)|$$

so (d) implies $P(A_n)$, $P(A'_n) \to P(A)$. Now $A - C \subset (A - B) \cup (B - C)$ and with a similar inequality for $C - A$ implies $A \triangle C \subset (A \triangle B) \cup (B \triangle C)$. The last inequality, (d), and (a) imply

$$P(A_n \triangle A'_n) \le P(A_n \triangle A) + P(A \triangle A'_n) \to 0$$

The last result implies

$$\begin{aligned} 0 &\le P(A_n) - P(A_n \cap A'_n) \\ &\le P(A_n \cup A'_n) - P(A_n \cap A'_n) \\ &= P(A_n \triangle A'_n) \to 0 \end{aligned}$$

so $P(A_n \cup A'_n) \to P(A)$. But A_n and A'_n are independent, so

$$P(A_n \cap A'_n) = P(A_n)P(A'_n) \to P(A)^2$$

This shows $P(A) = P(A)^2$, and proves Theorem 4.1.1.

□

A typical application of Theorem 4.1.1 is

Theorem 4.1.2. *For a random walk on* $\mathbf{R}$*, there are only four possibilities, one of which has probability one.*

(i) $S_n = 0$ *for all* n*. (ii)* $S_n \to \infty$*. (iii)* $S_n \to -\infty$*. (iv)* $-\infty = \liminf S_n < \limsup S_n = \infty$*.*

Proof. Theorem 4.1.1 implies $\limsup S_n$ is a constant $c \in [-\infty, \infty]$. Let $S'_n = S_{n+1} - X_1$. Since S'_n has the same distribution as S_n, it follows that $c = c - X_1$. If c is finite, subtracting c from both sides we conclude $X_1 \equiv 0$ and (i) occurs. Turning the last statement around, we see that if $X_1 \not\equiv 0$ then $c = -\infty$ or ∞. The same analysis applies to the $\liminf$. Discarding the impossible combination $\limsup S_n = -\infty$ and $\liminf S_n = +\infty$, we have proved the result.

□

The special case in which $P(X_i = 1) = P(X_i = -1) = 1/2$ is called **simple random walk**. Since a simple random walk cannot skip over any integers, it follows from either exercise above that with probability one it visits every integer infinitely many times.

Let $\mathcal{F}_n = \sigma(X_1, ..., X_n) =$ the information known at time n. A random variable N taking values in $\{1, 2, ...\} \cup \{\infty\}$ is said to be a **stopping time** or an **optional random variable** if for every $n < \infty$, $\{N = n\} \in \mathcal{F}_n$. If we think of S_n as giving the (logarithm of the) price of a stock at time n, and N as the time we sell it, then the last definition says that the decision to sell at time n must be based on the information known at that time. The last interpretation gives one explanation for the second name. N is a time at which we can exercise an option to buy a stock.

The canonical example of a stopping time is $N = \inf\{n : S_n \in A\}$, the **hitting time of A**. To check that this is a stopping time, we observe that

$$\{N = n\} = \{S_1 \in A^c, ..., S_{n-1} \in A^c, S_n \in A\} \in \mathcal{F}_n$$

Two concrete examples of hitting times that have appeared above are

Example 4.1.3. $N = \inf\{k : |S_k| \geq x\}$ from the proof of Theorem 2.5.4.

Example 4.1.4. If the $X_i \geq 0$ and $N_t = \sup\{n : S_n \leq t\}$ is the random variable that first appeared in example in Section 2.4, then $N_t + 1 = \inf\{n : S_n > t\}$ is a stopping time.

Theorem 4.1.5. *Let $X_1, X_2, ...$ be i.i.d., $\mathcal{F}_n = \sigma(X_1, ..., X_n)$ and N be a stopping time with $P(N < \infty) > 0$. Conditional on $\{N < \infty\}$, $\{X_{N+n}, n \geq 1\}$ is independent of $\mathcal{F}_n$ and has the same distribution as the original sequence.*

Proof. We can refer to Theorem A.1.5 from text [4]. It is enough to show that if $A \in \mathcal{F}_N$ and $B_j \in \mathcal{S}$ for $1 \leq j \leq k$, then

$$(A, N < \infty, X_{N+j} \in B_j, 1 \leq j \leq k) = P(A \cap \{N < \infty\}) \prod_{j=1}^{k} \mu(B_j)$$

where $\mu(B) = P(X_i \in B)$. The method ("divide and conquer") is one that we will see many times below. We break things down according to the value of N in order to replace N by n and reduce to the case of a fixed time.

$$\begin{aligned} P(A, N = n, X_{N+j} \in B_j, 1 \leq j \leq k) &= P(A, N = n, X_{n+j} \in B_j, 1 \leq j \leq k) \\ &= P(A \cap \{N = n\}) \prod_{j=1}^{k} \mu(B_j) \end{aligned}$$

and since $A \cap \{N = n\} \in \mathcal{F}_n$ and that σ-field is independent of $X_{n+1}, ..., X_{n+k}$. Summing over n now gives the desired result.

□

To delve further into properties of stopping times, we recall that we have supposed $\Omega = S^{\mathbf{N}}$ and define the **shift** $\theta : \Omega \to \Omega$ by

$$(\theta\omega)(n) = \omega(n+1) n = 1, 2, ...$$

In words, we drop the first coordinate and shift the others one place to the left. The iterates of θ are defined by composition. Let $\theta^1 = \theta$, and for $k \geq 2$, let $\theta^k = \theta \cdot \theta^{k-1}$. Clearly, $(\theta^k\omega)(n) = \omega(n+k), n = 1, 2, ...$ To extend the last definition to stopping times, we let

$$\theta^N\omega = \begin{cases} \theta^n\omega & \text{on } \{N = n\} \\ \triangle & \text{on } \{N = \infty\} \end{cases}$$

Here $\triangle$ is an extra point that we add to Ω.

Example 4.1.6. Returns to 0. For a concrete example of the use of θ, suppose $S = \mathbb{R}^d$ and let

$$\tau(\omega) = \inf\{n : \omega_1 + \cdots + \omega_n = 0\}$$

where $\inf \emptyset = \infty$, and we set $\tau(\triangle) = \infty$. If we let $\tau_2(\omega) = \tau(\omega) + \tau(\theta^\tau\omega)$, then on $\{\tau < \infty\}$,

$$\begin{aligned} \tau(\theta^\tau\omega) &= \inf\{n : (\theta^\tau(\omega) + \cdots + (\theta^\tau\omega)_n = 0\} \\ &= \inf\{n : \omega_{\tau+1} + \cdots + \omega_{\tau+n} = 0\} \end{aligned}$$

So τ_2 is the time of the second visit to 0 (and thanks to the conventions $\theta^\infty\omega = \triangle$ and $\tau(\triangle) = \infty$, this is true for all ω). The last computation generalizes easily to show that if we let

$$\tau_n(\omega) = \tau_{n-1}(\omega) + \tau(\theta^{\tau_{n-1}}\omega)$$

then τ_n is the time of the nth visit to 0.

If we have any stopping time T, we can define its iterates by $T_0 = 0$ and

$$T_n(\omega) = T_{n-1}(\omega) + T(\theta^{T_{n-1}}\omega) \text{ for } n \geq 1$$

If we assume $P = \mu \times \mu \times \mu \ldots$ then

$$P(T_n < \infty) = P(T < \infty)^n \tag{4.1}$$

Proof. we will prove this by induction. The result is trivial when $n = 1$. Suppose now that it is valid for $n - 1$. Applying Theorem 4.1.5 to $N = T_{n-1}$, we see that $T(\theta^{T_{n-1}}) < \infty$ is independent of $T_{n-1} < \infty$ and has the same probability as $T < \infty$, so

$$\begin{aligned} P(T_n < \infty) &= P(T_{n-1} < \infty, T(\theta^{T_{n-1}}\omega) < \infty) \\ &= P(T_{n-1} < \infty)P(T < \infty) \\ &= P(T < \infty)^n \end{aligned}$$

by induction hypothesis.

□

Letting $t_n = T(\theta^{T_{n-1}})$, we can extend Theorem 4.1.5 to

Theorem 4.1.7. *Suppose $P(T < \infty) = 1$. Then the "random vectors"*

$$V_n = (t_n, X_{T_{n-1}+1}, ..., X_{T_n})$$

are independent and identically distributed.

Proof. It is clear from Theorem 4.1.5 and V_n and V_1 have the same distribution. The independence follows from Theorem 4.1.5 and induction since $V_1, ..., V_{n-1} \in \mathcal{F}(T_{n-1})$.

□

Example 4.1.8. Ladder variables. Let $\alpha(\omega) = \inf\{n : \omega_1 + \cdots + \omega_n > 0\}$ where $\inf \emptyset = \infty$, and set $\alpha(\triangle) = \infty$. Let $\alpha_0 = 0$ and let

$$\alpha_k(\omega) = \alpha_{k-1}(\omega) + \alpha(\theta^{\alpha_{k-1}}\omega)$$

for $k \geq 1$. At time α_k, the random walk is at a record high value.

A famous result about stopping times for random walks is:

Theorem 4.1.9. ***Wald's equation***. *Let $X_1, X_2, ...$ be i.i.d. with $\mathbb{E}|X_i| < \infty$. Iff N is a stopping time with $\mathbb{E}N < \infty$, then $\mathbb{E}S_n = \mathbb{E}X_1\mathbb{E}N$.*

Proof. First suppose the $X_i \geq 0$.

$$\mathbb{E}S_n = \int S_n dP = \sum_{n=1}^{\infty} \int S_n 1_{\{N=n\}} dP = \sum_{n=1}^{\infty}\sum_{m=1}^{n} \int X_m 1_{\{N=n\}} dP$$

Since the $X_i \geq 0$, we can interchange the order of summation (i.e., use Fubini's theorem) to conclude that the last expression

$$= \sum_{m=1}^{\infty}\sum_{n=m}^{\infty} \int X_m 1_{\{N=n\}} dP = \sum_{m=1}^{\infty} \int X_m 1_{\{N\geq m\}} dP$$

Now $\{N \geq m\} = \{N \leq m-1\}^c \in \mathcal{F}_{m-1}$ and is independent of X_m, so the last expression

$$= \sum_{m=1}^{\infty} \mathbb{E}X_m P(N \geq m) = \mathbb{E}X_1\mathbb{E}N$$

To prove the result in general, we run the last argument backwards. If we have $\mathbb{E}N < \infty$ then

$$\infty > \sum_{m=1}^{\infty} \mathbb{E}|X_m| P(N \geq m) = \sum_{m=1}^{\infty}\sum_{n=m}^{\infty} \int |X_m| \mathbf{1}_{\{N=n\}} dP$$

The last formula shows that the double sum converges absolutely in one order, so Fubini's theorem gives

$$\sum_{m=1}^{\infty}\sum_{n=m}^{\infty} \int X_m \mathbf{1}_{\{N=n\}} dP = \sum_{n=1}^{\infty}\sum_{n=m}^{n} \int X_m \mathbf{1}_{\{N=n\}} dP$$

Using the independence of $\{N \geq m\} \in \mathcal{F}_{m-1}$ and X_m and rewriting the last identity, it follows that

$$\sum_{m=1}^{\infty} \mathbb{E}X_m P(N \geq m) = \mathbb{E}S_N$$

Since the left-hand side is $\mathbb{E}N\mathbb{E}X_1$, the proof is complete.

□

Example 4.1.10. Simple random walk. Let $X_1, X_2, ...$ be i.i.d. with $P(X_i = 1) = 1/2$ and $P(X_i = -1) = 1/2$. Let $a < 0 < b$ be integers and let $N = \inf\{n : S_n \notin (a, b)\}$. To apply Theorem 4.1.9, we have to check that $\mathbb{E}N < \infty$. To do this, we observe that if $x \in (a, b)$, then

$$P(x + S_{b-a} \notin (a, b)) \geq 2^{-(b-a)}$$

since $b - a$ steps of size $+1$ in a row will take us out of the interval. Iterating the last inequality, it follows that

$$P(N > n(b-a)) \leq (1 - 2^{-(b-a)})^n$$

so $\mathbb{E}N < \infty$. Applying Theorem 4.1.9 now gives $\mathbb{E}S_n = 0$ or

$$bP(S_n = b) + aP(S_n = a) = 0$$

Since $P(S_n = b) + P(S_n = a) = 1$, it follows that $(b-a)P(S_n = b) = -a$, so

$$P(S_n = b) = \frac{-a}{b-a} \; P(S_N = a) = \frac{b}{b-a}$$

Letting $T_a = \inf\{n : S_n = a\}$, we can write the last conclusion as

$$P(T_a < T_b) = \frac{b}{b-a} \text{ for } a < 0 < b \tag{4.2}$$

Setting $b = M$ and letting $M \to \infty$ gives

$$P(T_a < \infty) \geq P(T_a < T_m) \to 1$$

for all $a < 0$. From symmetry (and the fact that $T_0 \equiv 0$), it follows that

$$P(T_x < \infty) = 1 \text{ for all } x \in \mathbb{Z} \tag{4.3}$$

Our final fact about T_x is that $\mathbb{E}T_x = \infty$ for $x \neq 0$. To prove this, note that if $\mathbb{E}T_x < \infty$ then Theorem 4.1.5 would imply

$$x = \mathbb{E}S_{T_x} = \mathbb{E}X_1 \mathbb{E}T_x = 0$$

In Section 4.3, we will compute the distribution of T_1 and show that

$$P(T_1 > t) \sim Ct^{-1/2}$$

Theorem 4.1.11. ***Wald's second equation***. *Let $X_1, X_2, ...$ be i.i.d. with $\mathbb{E}X_n = 0$ and $\mathbb{E}X_n^2 = \sigma^2 < \infty$. If T is a stopping time with $\mathbb{E}T < \infty$, then $\mathbb{E}S_T^2 = \sigma^2 \mathbb{E}T$*

Proof. Using the definitions and then taking expected value

$$S_{T\wedge n}^2 = S_{T\wedge(n-1)}^2 + (2X_n S_{n-1} + X_n^2)1_{(T \geq n)}$$

$$\mathbb{E}S_{T\wedge n}^2 = \mathbb{E}S_{T\wedge(n-1)}^2 + \sigma^2 P(T \geq n)$$

since $\mathbb{E}X_n = 0$ and X_n is independent of S_{n-1} and $1_{(T\geq n)} \in \mathcal{F}_{n-1}$. The expectation of $S_{n-1}X_n$ exists since both random variables are in L^2. From the last equality and induction we get

$$\mathbb{E}T_{T\wedge n}^2 = \sigma^2 \sum_{m=1}^{n} P(T \geq m)$$

$$\mathbb{E}(S_{T\wedge n} - S_{T\wedge m})^2 = \sigma^2 \sum_{k=m+1}^{n} P(T \geq n)$$

The second equality follows from the first applied to X_{m+1}, X_{m+2}, ... The second equality implies that $S_{T\wedge n}$ is a Cauchy sequence in L^2, so letting $n \to \infty$ in the first, it follows that $\mathbb{E}S_T^2 = \sigma^2 \mathbb{E}T$.

□

Example 4.1.12. Simple random walk, II. Continuing previous example, we investigate $N = \inf\{S_n \notin (a, b)\}$. We have shown that $\mathbb{E}N < \infty$. Since $\sigma^2 = 1$, it follows from Theorem 4.1.11 and 4.2 that

$$\mathbb{E}N = \mathbb{E}S_N^2 = a^2 \frac{b}{b-a} + b^2 \frac{-a}{b-a} = -ab$$

If $b = L$ and $a = -L$, $\mathbb{E}N = L^2$.

An amusing consequence of Theorem 4.1.11 is

Theorem 4.1.13. *Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{E}X_n = 0$ and $\mathbb{E}X_n^2 = 1$, and let $T_c = \inf\{n \geq 1 : |S_n| > cn^{1/2}\}$.*

$$\mathbb{E}T_c \begin{cases} < \infty & \text{for } c < 1 \\ = \infty & \text{for } c \geq 1 \end{cases}$$

Proof. One half of this is easy. If $\mathbb{E}T_c < \infty$ then, the previous exercise implies $\mathbb{E}T_c = \mathbb{E}(S_{T_c}^2) > c^2\mathbb{E}T_c$, a contradiction if $c \geq 1$. To prove the other direction, we let $\tau = T_c \wedge n$ and observe $S_{\tau-1}^2 \leq c^2(\tau - 1)$, so using the Cauchy-Schwarz inequality

$$\mathbb{E}\tau = \mathbb{E}S_\tau^2 = \mathbb{E}S_{\tau-1}^2 + 2\mathbb{E}(S_{\tau-1}X_\tau) + \mathbb{E}X_\tau^2 \leq c^2\mathbb{E}\tau + 2c(\mathbb{E}\tau\mathbb{E}X_\tau^2)^{1/2} + \mathbb{E}X_\tau^2$$

To complete the proof now, we will show

Lemma 4.1.14. *If T is a stopping time with $\mathbb{E}T = \infty$, then*

$$\mathbb{E}X_{T\wedge n}^2/\mathbb{E}(T \wedge n) \to 0$$

Theorem 4.1.13 follows, for if $\epsilon < 1 - c^2$ and n is large, we will have $\mathbb{E}\tau \leq (c^2 + \epsilon)\mathbb{E}\tau$, a contradiction.

Proof. We begin by writing

$$\mathbb{E}(X_{T\wedge n}^2) = \mathbb{E}(X_{T\wedge n}^2; X_{T\wedge n}^2 \leq \epsilon(T \wedge n)) + \sum_{j=1}^{n} \mathbb{E}(X_j^2; T \wedge n = j, X_j^2 > \epsilon j)$$

The first term is $\leq \epsilon\mathbb{E}(T \wedge n)$. To bound the second, choose $N \geq 1$ so that for $n \geq N$

$$\sum_{j=1}^{n} \mathbb{E}(X_j^2; X_j^2 > \epsilon j) < n\epsilon$$

This is possible since the dominated convergence theorem implies $\mathbb{E}(X_j^2; X_j^2 > \epsilon j) \to 0$ as $j \to \infty$. For the first part of the sum, we use a trivial bound

$$\sum_{j=1}^{N} \mathbb{E}(X_j^2; T \wedge n = j, X_j^2 > \epsilon j) \leq N\mathbb{E}X_1^2$$

To bound the remainder of the sum, we note (i) $X_j^2 \geq 0$; (ii) $\{T \wedge n \geq j\}$ is $\in \mathcal{F}_{j-1}$ and hence is independent of $X_j^2 1_{(X_j^2 > \epsilon j)}$, (iii) use some trivial arithmetic, (iv) use Fubini's theorem and enlarge the range of j, (v) use the choice of N and a trivial inequality

$$\begin{aligned}
\sum_{j=N}^{n} \mathbb{E}(X_j^2; T \wedge n = j, X_j^2 > \epsilon j) &\leq \sum_{j=N}^{n} \mathbb{E}(X_j^2; T \wedge n \geq j, X_j^2 > \epsilon j) \\
&= \sum_{j=N}^{n} P(T \wedge n \geq j)\mathbb{E}(X_j^2; X_j^2 > \epsilon j) \\
&= \sum_{j=N}^{n} \sum_{k=1}^{\infty} P(T \wedge n = k)\mathbb{E}(X_j^2; X_j^2 > \epsilon j) \\
&\leq \sum_{k=N}^{\infty} \sum_{j=1}^{k} P(T \wedge n = k)\mathbb{E}(X_j^2; X_j^2 > \epsilon j) \\
&\leq \sum_{k=N}^{\infty} \epsilon k P(T \wedge n = k) \\
&\leq \epsilon\mathbb{E}(T \wedge n)
\end{aligned}$$

Combining our estimates shows

$$\mathbb{E}X_{T\wedge n}^2 \leq 2\epsilon\mathbb{E}(T\wedge n) + N\mathbb{E}X_1^2$$

Letting $n \to \infty$ and noting $\mathbb{E}(T\wedge n) \to \infty$, we have

$$\limsup_{n\to\infty} \mathbb{E}X_{T\wedge n}^2/\mathbb{E}(T\wedge n) \leq 2\epsilon$$

where ϵ is arbitrary.

□

4.2 Recurrence

Throughout this section, S_n will be will be a random walk, that is, $S_n = X_1 + \cdots + X_n$ where $X_1, X_2, ...$ are i.i.d., and we will investigate the question mentioned at the beginning of the chapter. Does the sequence $S_1(\omega), S_2(\omega),$ ... return (or near) 0 infinitely often? The answer to the last question is either Yes or No, and the random walk is called recurrent or transient accordingly. We begin with some definition.

The number $x \in \mathbb{R}^d$ is said to be a **recurrent value** for the random walk S_n if for every $\epsilon > 0$, $P(||S_n - x|| < \epsilon i.o.) = 1$. Here $||x|| = \sup|X_i|$. The reader will see the reason for this choice of norm in later lemma.s The Hewitt-Savage 0-1 law implies that if the last probability is < 1, it is 0. Our first result shows that to know the set of recurrent values, it is enough to check $x = 0$. A number x is said to be a **possible value** of the random walk if for any $\epsilon > 0$, there is an n so that $P(|||S_n - x|| < \epsilon) > 0$.

Theorem 4.2.1. *The set $\mathcal{V}$ of recurrent values is either $\emptyset$ or a closed subgroup of $\mathbf{R}^d$. In the second case, $\mathcal{V} = \mathcal{U}$, the set of possible values.*

Proof. Suppose $\mathcal{V} \neq \emptyset$. It is clear that $\mathcal{V}^c$ is open, so $\mathcal{V}$ is closed. To prove this is a group, we will first show that

($\star$) if $x \in \mathcal{U}$ and $y \in \mathcal{V}$ then $y - x \in \mathcal{V}$.

This statement has been formulated so that once it is established, the result follows easily. Let

$$p_{\delta,m}(z) = P(||S_n - z \geq \delta \text{ for all } n \geq m)$$

If $y - x \notin \mathcal{V}$, there is an $\epsilon > 0$ and $m \geq 1$ so that $p_{2\epsilon,m}(y - x) > 0$. Since $x \in \mathcal{U}$, there is a k so that $P(||S_k - x|| < \epsilon) > 0$. Since

$$P(||S_n - S_k - (y - x)|| \geq 2\epsilon \text{ for all } n \geq k + m) = p_{2\epsilon,m}(y - x)$$

and is independent of $\{||S_k - x|| < \epsilon\}$, it follows that

$$p_{\epsilon,m+k}(y) \geq P(||S_k - x|| < \epsilon)p_{2\epsilon,m}(y - x) > 0$$

contradicting $y \in \mathcal{V}$, so $y - x \in \mathcal{V}$.

To conclude that $\mathcal{V}$ is a group when $\mathcal{V} \notin \emptyset$, let $q, r \in \mathcal{V}$, and observe: (i) taking $x = y = r$ in ($\star$)shows $0 \in \mathcal{V}$, (ii) taking $x = r$, $y = 0$ shows $-r \in \mathcal{V}$, and (iii) taking $x = -r$, $y = q$ shows $q + r \in \mathcal{V}$. To prove that $\mathcal{V} = \mathcal{U}$ now, observe that if $u \in \mathcal{U}$ taking $x = u$, $y = 0$ shows $-u \in \mathcal{V}$, and since $\mathcal{V}$ is a group, it follows that $u \in \mathcal{V}$.

□

If $\mathcal{V} = \emptyset$, the random walk is said to be **transient**; otherwise it is called **recurrent**. Before plunging into the technicalities needed to treat a general random walk, we begin by analyzing the special case Polya considered in 1921. Legend has it that Polya thought of this problem wandering around in a park near Zürich when he noticed that he kept encountering the same young couple. History does not record what the young couple thought.

Example 4.2.2. Simple Random walk on $\mathbf{Z}^d$.

$$P(X_i = e_j) = P(X_i = -e_j) = 1/2d$$

for each of the d unit vectors e_j. To analyze this case, we begin with a result that is valid for any random walk. Let $\tau_0 = 0$ and $\tau_n = \inf\{m > \tau_{n-1} : S_m = 0\}$ be the time of the nth return to 0. From 4.1, it follows that

$$P(\tau_n < \infty) = P(\tau_1 < \infty)^n$$

a fact that leads easily to:

Theorem 4.2.3. *For any random walk, the following are equivalent:*

$$(i)P(\tau_1 < \infty = 1, (ii)P(S_m = 0, i.o) = 1, \text{ and } (iii) \sum_{m=0}^{\infty} P(S_m = 0) = \infty$$

Proof. If $P(\tau_1 < \infty) = 1$, then $P(\tau_n < \infty) = 1$ for all n and $P(S_m = 0 \text{ i.o.}) = 1$. Let

$$V = \sum_{m=0} \infty 1_{(S_m=0)} = \sum_{n=0}^{\infty} 1_{(\tau_n<\infty)}$$

be the number of the visits to 0, counting the visit at time 0. Taking expected value and using Fubini's theorem to put the expected value inside the sum:

$$\begin{aligned} \mathbb{E}V &= \sum_{m=0}^{\infty} p(S_m = 0) \\ &= \sum_{n=0}^{\infty} P(\tau_n < \infty) \\ &= \sum_{n=0}^{\infty} P(\tau_1 < \infty)^n \\ &= \frac{1}{1 - P(\tau_1 < \infty)} \end{aligned}$$

The second equality shows that (ii) implies (iii) and, in combination with the last two, shows that if (i) is false, then (iii) is false (i.e., (iii) implies (i)).

□

Theorem 4.2.4. *Simple random walk is recurrent in $d \leq 2$ and transient in $d \geq 3$.*

Proof. Let $\rho_d(m) = P(S_m = 0)$. $\rho_d(m)$ is 0 if m is odd. From Theorem 3.1.3, we get $\rho_1(2n) \sim (\pi n)^{(-1/2}$ as $n \to \infty$. This and Theorem 4.2.3 gives the result in one dimension. Our next step is

Simple random walk is recurrent in two dimensions. Note that in order for $S_{2n} = 0$, we must for some $0 \leq m \leq n$ have m up steps, m down steps, $n - m$ to the left, and $n - m$ to the right, so

$$\begin{aligned} \rho_2(2n) &= 4^{-2n} \sum_{m=0}^{n} \frac{2n!}{m!m!(n-m)!(n-m)!} \\ &= 4^{-2n} \binom{2n}{n} \sum_{m=0}^{n} \binom{n}{m} \binom{n}{m-n} \\ &= 4^{-2n} \binom{2n}{n}^2 \\ &= \rho_1(2n)^2 \end{aligned}$$

To see the next-to-last equality, consider choosing n students from a class with n boys and n girls and observe that for some $0 \le m \le n$, you must choose m boys and $n-m$ girls. Using the asymptotic formula $\rho_1(2n) \sim (\pi n)^{-1/2}$, we get $\rho_2(2n) \sim (\pi n)^{-1}$. Since $\sum n^{-1} = \infty$, the result follows from Theorem 4.2.3.

Remark 4.2.5. For a direct proof of $\rho_2(2n) = \rho_1(2n)^2$, note that if T_n^1 and T_n^2 are independent, one-dimensional random walks, then T_n jumps from x to $x+(1,1)$, $x+(1,-1)$, $x+(-1,1)$, and $x+(-1,-1)$ with equal probability, so rotating T_n by 45 degrees and dividing by $\sqrt{2}$ gives S_n.

Simple random walk is transient in three dimensions. Intuitively, this holds since the probability of being back at 0 after 2n steps is $\sim cn^{-3/2}$, and this is summable. We will not compute the probability exactly but will get an upper bound of the right order of magnitude. Again, since the number of steps in the directions $\pm e_i$ must be equal for $i = 1, 2, 3$,

$$\begin{aligned}
\rho_3(2n) &= 6^{-2n} \sum_{j,k} \frac{(2n)!}{(j!k!(n-j-k)!)^2} \\
&= 2^{-2n} \binom{2n}{n} \sum_{j.k} \left(3^{-n} \frac{n!}{j!k!(n-j-k)!} \right)^2 \\
&\le 2^{-2n} \binom{2n}{n} \max_{j,k} 3^{-n} \frac{n!}{j!k!(n-j-k)!}
\end{aligned}$$

where in the last inequality we have used the fact that if $a_{j.k}$ are ≥ 0 and sum to 1, then $\sum_{j.k} a_{j.k}^2 \le \max_{j.k} a_{j.k}$. Our last step is to show

$$\max_{j.k} 3^{-n} 3^{-n} \frac{n!}{j!k!(n-j-k)!} \le Cn^{-1}$$

To do this, we note that (a) if any of the numbers j, k or $n-j-k$ is $< [n/3]$, increasing the smallest number and decreasing the largest number decreases the denominator (since $x(1-x)$ is maximized at $1/2$), so the maximum occurs when all three numbers are as close as possible to $n/3$; (b) Stirling's formula implies

$$\frac{n!}{j!k!(n-j-k)!} \sim \frac{n^n}{j^j k^k (n-j-k)^{n-j-k}} \cdot \sqrt{\frac{n}{jk(n-j-k)}} \cdot \frac{1}{2\pi}$$

Taking j and k within 1 of $n/3$ the first term on the right is $\le C3^n$, and the desired result follows.

Simple random walk is transient in $d > 3$. Let $T_n = (S_n^1, S_n^2, S_n^3)$, $N(0) = 0$ and $N(n) = \inf\{m > N(n-1) : T_m \ne T_{N(n-1)}\}$. It is easy to see that $T_{N(n)}$ is a three dimensional simple random walk. Since $T_{N(n)}$ returns infinitely often to 0 with probability 0 and the first three coordinates are constant in between the $N(n)$, S_n is transient.

□

Remark 4.2.6. Let $\pi_d = P(S_n = 0 \text{ for some } n \ge 1)$ be the probability that simple random walk on $\mathbf{Z}^d$ returns to 0. The last display in the proof of Theorem 4.2.3 implies

$$\sum_{n=0}^{\infty} p(S_{2n} = 0) = \frac{1}{1-\pi_d} \tag{4.4}$$

In $d = 3$, $P(S_{2n} = 0) \sim Cn^{-3/2}$ so $\sum_{n=N}^{\infty} P(S_{2n} = 0) \sim C' N^{-1/2}$, and the series converges rather slowly.

The rest of the section is devoted to proving the following facts about random walks:
- S_n is recurrent in $d = 1$ if $S_n/n \to 0$ in probability.
- S_n is recurrent in $d = 2$ if $S_n/n^{1/2} \Rightarrow$ a nondegenerate normal distribution.
- S_n is transient in $d \geq 3$ if it is "truly three-dimensional."

Lemma 4.2.7. *If* $\sum_{n=1}^{\infty} P(||S_n|| < \epsilon) < \infty$, *then* $P(||S_n|| < \epsilon \text{ i.o.}) = 0$. *If* $\sum_{n=1}^{\infty} P(||S_n|| < \epsilon) = \infty$ *then* $P(||S_n|| < 2\epsilon \text{ i.,o.}) = 1$.

Proof. The first conclusion follows from the Borel-Cantelli lemma. To prove the second, let $F = \{||S_n|| < \epsilon \text{ i.o.}\}^c$. Breaking things down according to the last time $||S_n|| < \epsilon$,

$$\begin{aligned} P(F) &= \sum_{m=0}^{\infty} P(||S_m|| < \epsilon, ||S_n|| \geq \epsilon \text{ for all } n \geq m+1) \\ &\geq \sum_{m=0}^{\infty} p(||S_m|| < \epsilon, ||S_n - S_m|| \geq 2\epsilon \text{ for all } n \geq m+1) \\ &= \sum_{m=0}^{\infty} P(||_m|| < \epsilon)\rho_{2\epsilon,1} \end{aligned}$$

where $\rho_{s,k} = P(||S_n|| \geq \delta \text{ for all } n \geq k)$. Since $P(F) \leq 1$, and

$$\sum_{m=0}^{\infty} P(||S_m|| < \epsilon) = \infty$$

it follows that $\rho_{2\epsilon,1} = 0$. To extend his conclusion to $\rho_{2\epsilon,k}$ with $k \geq 2$, let

$$A_m = \{||S_m|| < \epsilon, ||S_n|| \geq \epsilon \text{ for all } n \geq m+k\}$$

Since any ω can be in at most k of the A_m, repeating the argument above gives

$$k \geq \sum_{m=0}^{\infty} P(A_m) \geq \sum_{m=0}^{\infty} P(||S_m|| < \epsilon)\rho_{2\epsilon,k}$$

So $\rho_{2\epsilon,k} = P(||S_n|| \geq 2\epsilon \text{ for all } j \geq k) = 0$, and since k is arbitrary, the desired conclusion follows.

□

Lemma 4.2.8. *Let* m *be an integer* ≥ 2,

$$\sum_{n=0}^{\infty} P(||S_n|| < m\epsilon) \leq (2m)^d \sum_{n=0}^{\infty} P(||S_n|| < \epsilon)$$

Proof. We begin by observing

$$\sum_{n=0}^{\infty} P(||S_n|| < m\epsilon) \leq \sum_{n=0}^{\infty} \sum_{k} P(S_n \in k\epsilon + [0,\epsilon)^d)$$

where to the inner sum is over $k \in \{-m, ..., m-1\}^d$. If we let

$$T_k = \inf\{l \geq 0 : S_; \in k\epsilon + [0,\epsilon)^d\}$$

then breaking things down according to the value of T_k and using Fubini's theorem gives

$$\begin{aligned} \sum_{n=0}^{\infty} P(S_n \in k\epsilon + [0,\epsilon)^d) &= \sum_{n=0}^{\infty} \sum_{l=0}^{n} P(S_n \in k\epsilon + [0,\epsilon)^d, T_k = l) \\ &\leq \sum_{l=0}^{\infty} \sum_{n=l}^{\infty} P(||S_n - S_l|| < \epsilon, T_k = l) \end{aligned}$$

Since $\{T_k = l\}$ and $\{||S_n - S_l|| < \epsilon\}$ are independent, the last sum

$$= \sum_{m=0}^{\infty} P(T_k = m) \sum_{j=0}^{\infty} P(||S_j|| < \epsilon) \leq \sum_{j=0}^{\infty} P(||S_j|| < \epsilon)$$

Since there are $(2m)^d$ values of k in $\{-m, ..., m-1\}^d$, the proof is complete.

□

Combining Lemmas 4.2.7 and 4.2.8 gives:

Theorem 4.2.9. *The convergence (resp. divergence) of $\sum_n P(||S_n|| < \epsilon)$ for a single value of $\epsilon > 0$ is sufficient for transience (resp. recurrence).*

Theorem 4.2.10. ***Chung-Fuchs theorem**. Suppose $d = 1$. If the weak law of large numbers holds in the form $S_n/n \to 0$ in probability, then S_n is recurrent.*

Proof. Let $u_n(x) = P(|S_n| < x)$ for $x > 0$. Lemma 4.2.8 implies

$$\sum_{n=0}^{\infty} u_n(1) \geq \frac{1}{2m} \sum_{n=0}^{\infty} u_n(m) \geq \frac{1}{2m} \frac{n=0}{Am} u_n(n/A)$$

for any $A < \infty$ since $u_n(x) \geq 0$ and is increasing in x. By hypothesis $u_n(n/A) \to 1$, so letting $m \to \infty$ and noticing the right-hand side is $A/2$ times the average of the first Am terms

$$\sum_{n=0}^{\infty} u_n(1) \geq A/2$$

Since A is arbitrary, the sum must be ∞, and the desired conclusion follows from Theorem 4.2.9.

□

Theorem 4.2.11. *If S_n is a random walk in $\mathbb{R}^2$ and $S_n/n^{1/2} \Rightarrow$ a nondegenerate normal distribution, then S_n is recurrent.*

5 MARTINGALES

Go back to Table of Contents. Please click TOC

A martingale X_n can be thought of as the fortune at time n of a player who is betting on a fair game; submartingales (supermartingales) as the outcome of betting on a favorable (unfavorable) game. There are two basic facts about martingales. The first is that you cannot make money betting on them (read §5.2), and in particular if you choose to stop playing at some bounded time N, then your expected winnings $\mathbb{E}_N$ are equal to your initial fortune X_0. (We are supposing for the moment that X_0 is not random.) Our second fact, discussed in §5.2 as well, concerns submartingales. To use a heuristic we learned from Mike Brennan, "They are the stochastic analogues of nondecreasing sequences and so if they are bounded above (to be precise, $\sup_n \mathbb{E}X_n^+ < \infty$) they converge almost surely."

5.1 Conditional Expectation

We begin with a definition that is important for this chapter and the next one. After giving the definition, we will consider several examples to explain it. Given are a probability space $(\Omega, \mathcal{F}_0, P)$, a σ-field $\mathcal{F} \subset \mathcal{F}_0$, and a random variable $X \in \mathcal{F}_0$ with $\mathbb{E}|X| < \infty$. We define the **conditional expectation of X given $\mathcal{F}$**, $\mathbb{E}(X|\mathcal{F})$, to be any random variable Y that has

(i) $Y \in \mathcal{F}$, that is, is $\mathcal{F}$ measurable, and

(ii) for all $A \in \mathcal{F}$, $\int_A X dP = \int_A Y dP$.

Any Y satisfying (i) and (ii) is said to be a **version of** $\mathbb{E}(X|\mathcal{F})$. The first thing to be settled is that the conditional expectation exists and is unique. We tackle the second claim first, but start with a technical point.

Lemma 5.1.1. *If Y satisfies (i) and (ii), then it is integrable.*

Proof. Letting $A = \{Y > 0\} \in \mathcal{F}$, using (ii) twice, and then adding

$$\int_A Y dP = \int_A X dP \leq \int_A |X| dP$$

$$\int_{A^c} -Y dP = \int_{A^c} -X dP \leq \int_{A^c} |X| dP$$

So we have $\mathbb{E}|Y| \leq \mathbb{E}|X|$.

□

Uniqueness. If Y' also satisfies (i) and (ii), then

$$\int_A Y dP = \int_A Y' dP \text{ for all } A \in \mathcal{F}$$

Taking $A = \{Y - Y' \geq \epsilon < 0\}$, we see

$$0 = \int_A X - X dP = \int_A Y - Y' dP \geq \epsilon P(A)$$

so $P(A) = 0$. Since this holds for all ϵ, we have $Y \leq Y'$ a.s., and interchanging the roles of Y and Y', we have $Y = Y'$ a.s. Technically, all equalities such as $Y = \mathbb{E}(X|\mathcal{F})$ should be written as $Y = \mathbb{E}(X|\mathcal{F})$ a.s., but we have ignored this point in previous chapters and will continue to do so.

Existence. To start, we recall ν is said to be **absolutely continuous with respect to** μ (abbreviated $\nu << \mu$) if $\mu(A) = 0$ implies $\nu(A) = 0$, and we use [4] Theorem A.4.6:

Radon-Nikodym theorem. Let μ and ν be σ-finite measures on $(\Omega, \mathcal{F})$. If $|nu << \mu$, there is a function $f \in \mathcal{F}$ so that for all $A \in \mathcal{F}$,

$$\int_A f d\mu = \nu(A)$$

f is usually denoted $d\nu/d\mu$ and called the **Radon-Nikodym derivative**.

The last theorem easily gives the existence of conditional expectation. Suppose first that $X \geq 0$. Let $\mu = P$ and

$$\nu(A) = \int_A X dP \text{ for } A \in \mathcal{F}$$

The dominated convergence theorem implies $\nu << \mu$. The Radon-Nikodym derivative $d\nu/d\mu \in \mathcal{F}$ and for any $A \in \mathcal{F}$ has

$$\int_A X dP = \nu(A) \int_A \frac{d\nu}{d\mu} dP$$

Taking $A = \Omega$, we see that $d\nu/d\mu \geq 0$ is integrable, and we have shown that $d\nu/d\mu$ is a version of $\mathbb{E}(X|\mathcal{F})$.

To treat the general case now, write $X = X^+ - X^-$, let $Y_1 = \mathbb{E}(X^+|\mathcal{F})$ and $Y_2 = \mathbb{E}(X^-|\mathcal{F})$. Now $Y_1 - Y_2 \in \mathcal{F}$ is integrable, and for all $A \in \mathcal{F}$ we have

$$\begin{aligned} \int_A X dP &= \int_A X^+ dP - \int_A X^- dP \\ &= \int_A Y_1 dP - \int_A Y_2 dP \\ &= \int_A (Y_1 - Y_2) dP \end{aligned}$$

This shows $Y_1 - Y_2$ is a version of $\mathbb{E}(X|\mathcal{F}$ and completes the proof.

□

5.1.1 Examples

Intuitively, we think of $\mathcal{F}$ as describing the information we have at our disposal - for each $A \in \mathcal{F}$, we know whether or not A has occurred. $\mathbb{E}(X|\mathcal{F})$ is then our "best guess" of the value of X given the information we have. Some examples should help to clarify this and connect $\mathbb{E}(X|\mathcal{F})$ with other definitions of conditional expectation.

Example 5.1.2. If $X \in \mathcal{F}$, then $\mathbb{E}(X|\mathcal{F}) = X$; that is, if we know X, then our "best guess" is X itself. Since X always satisfies (ii), the only thing that can keep X from being $\mathbb{E}(X|\mathcal{F})$ is condition (i). A special case of this example is $X = c$, where c is a constant.

Example 5.1.3. At the other extreme from perfect information is no information. Suppose X is independent of $\mathcal{F}$, that is, for all $B \in \mathcal{R}$ and $A \in \mathcal{F}$,

$$P(\{X \in B\} \cap A) = P(X \in B)P(A)$$

We claim that, in this case, $\mathbb{E}(X|\mathcal{F}) = \mathbb{E}X$; that is, if you do not know anything about X, then the best guess is the mean $\mathbb{E}X$. To check the definition, note that $\mathbb{E}X \in \mathcal{F}$ so (i). To verify (ii), we observe that if $A \in \mathcal{F}$, then since X and $1_A \in \mathcal{F}$ are independent, Theorem 2.1.11 implies

$$\int_A X dP = \mathbb{E}(1_A) = \mathbb{E}X\mathbb{E}1_A = \int_A \mathbb{E} dP$$

The reader should note that here and in what follows the game is "guess and verify."

Example 5.1.4. In this example, we relate the new definition of conditional expectation to the first one taught in undergraduate school. Suppose $\Omega_1, \Omega_2, ...$ is a finite or infinite partition of Ω into disjoint sets, each of which has positive probability, and let $\mathcal{F} = \sigma(\Omega_1, \Omega_2, ...)$ be the σ-field generated by these sets. Then

$$\mathbb{E}(X|\mathcal{F}) = \frac{\mathbb{E}(X;\Omega_i)}{P(\Omega_i)} \text{ on } \Omega_i$$

In words, the information in Ω_i tells us which element of the partition our outcome lies in, and given this information, the best guess for X is the average value of X over Ω_i. To prove our guess is correct, observe that the proposed formula is constant on each Ω_i, so it is measurable with respect to $\mathcal{F}$. To verify (ii), it is enough to check the equality for $A = \Omega_i$, but this is trivial:

$$\int_{\Omega_i} \frac{\mathbb{E}(X;\Omega_i)}{P(\Omega_i)} dP = \mathbb{E}(X;\Omega_i) = \int_{\Omega_i} X dP$$

A degenerate but important special case is $\mathcal{F} = \{\emptyset, \Omega\}$, the trivial σ-field. In this case, $\mathbb{E}(X|\mathcal{F}) = \mathbb{E}X$.

To continue the connection, let

$$P(A|\mathcal{G}) = \mathbb{E}(1_A|\mathcal{G})$$

$$P(A|B) = P(A \cap B)/P(B)$$

and observe that in the last example $P(A|\mathcal{F}) = P(A|\Omega_i)$ on Ω_i.

Example 5.1.5. To continue making connection with definitions of conditional expectation from undergraduate probability, suppose X and Y have joint density $f(x, y)$, that is,

$$P((X, Y) \in B) = \int_B f(x, y)dxdy \text{ for } B \in \mathcal{R}^2$$

and suppose for simplicity that $\int f(x, y)dx > 0$ for all y. We claim that in this case, if $\mathbb{E}|g(X)| < \infty$ then $\mathbb{E}(g(X)|Y) = h(Y)$, where

$$h(y) = \int g(x)f(x, y)dx \Big/ \int f(x, y)dx$$

To "guess" this formula, note that treating the probability densities $P(Y = y)$ as if they were real probabilities

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f(x, y)}{\int f(x, y)dx}$$

so, integrating against the conditional probability density, we have

$$\mathbb{E}(g(X)|Y = y) = \int g(x)P(X = x|Y = y)dx$$

To "verify" the proposed formula now, observe $h(Y) \in \sigma(Y)$ so (i) holds. To check (ii), observe that if $A \in \sigma(Y)$ then $A = \{\omega : Y(\omega) \in B\}$ for some $B \in \mathcal{R}$, sop

$$\begin{aligned} \mathbb{E}(h(Y); A) &= \int_B \int h(y)f(x, y)dxdy \\ &= \int_B \int g(x)f(x, y)dxdy \\ &= \mathbb{E}(g(X)1_b(y)) \\ &= \mathbb{E}(g(X); A) \end{aligned}$$

Remark 5.1.6. To drop the assumption that $\int f(x,y)dx > 0$, define h by

$$h(y)\int f(x,y)dx = \int g(x)f(x,y)dx$$

(i.e., h can be anything where $\int f(x,y)dx = 0$), and observe that this is enough for the proof.

Example 5.1.7. Suppose X and Y are independent. Let φ be a function with $\mathbb{E}|\varphi(X,Y)| < \infty$ and let $g(x) = \mathbb{E}(\varphi(x,Y))$. We will now show that

$$\mathbb{E}(\varphi(X,Y)|X) = g(X)$$

Proof. It is clear that $g(X) \in (X)$. To check (ii), note that if $A \in \sigma(X)$, then $A = \{X \in C\}$, so using the change of variables formula (Theorem 1.6.10) and the fact that the distribution of X, Y) is product measure (Theorem 2.1.9, then the definition of g, and change of variables again,

$$\begin{aligned}\int_A \phi(X,Y)dP &= \mathbb{E}\{\phi(X,Y)1_C(X)\} \\ &= \iint \phi(x,y)1_C(x)\nu(dy)\mu(dx) \\ &= \int 1_C(x)g(x)\mu(dx) \\ &= \int_A g(X)dP\end{aligned}$$

which proves the desired result.

5.1.2 Properties

Conditional expectation has many of the same properties that ordinary expectation does.

Theorem 5.1.8. *There are the following properties:*

(a) Conditional expectation is linear:

$$\mathbb{E}(aX + Y|\mathcal{F}) = a\mathbb{E}(X|\mathcal{F}) + \mathbb{E}(Y|\mathcal{F}) \tag{5.1}$$

(b) If $X \le Y$, then

$$\mathbb{E}(X|\mathcal{F}) \le \mathbb{E}(Y|\mathcal{F}) \tag{5.2}$$

(c) If $X_n \ge 0$ and $X_n \uparrow X$ with $\mathbb{E}X < \infty$, then

$$\mathbb{E}(X_n|\mathcal{F}) \uparrow \mathbb{E}(X|\mathcal{F}) \tag{5.3}$$

Proof. To prove (a), we need to check that the right-hand side is a version of the left. It clearly is $\mathcal{F}$-measurable. To check (ii), we observe that if $A \in \mathcal{F}$, then by linearity of the integral and the defining properties of $\mathbb{E}(X|\mathcal{F})$ and $\mathbb{E}(Y|\mathcal{F})$,

$$\begin{aligned}\int_A \{a\mathbb{E}(X|\mathcal{F}) + \mathbb{E}(Y|\mathcal{F})\}dP &= a\int_A \mathbb{E}(X|\mathcal{F})dP + \int_A \mathbb{E}(Y|\mathcal{F})dP \\ &= a\int_A XdP + \int_A YdP \\ &= \int_A aX + YdP\end{aligned}$$

which proves 5.1.

Use the definition

$$\int_A \mathbb{E}(X|\mathcal{F})dP = \int_A XdP \leq \int_A YdP = \int_A \mathbb{E}(Y|\mathcal{F})dP$$

Letting $A = \{\mathbb{E}(X|\mathcal{F}) - \mathbb{E}(Y|\mathcal{F}) \geq \epsilon > 0\}$, we see that the indicated set has probability 0 for all $\epsilon > 0$, and we have proved 5.2.

Let $Y_n = X - X_n$. It suffices to show that $\mathbb{E}(Y_n|\mathcal{F}) \downarrow 0$. Since $Y_n \downarrow$, 5.2 implies $Z_n \equiv \mathbb{E}(Y_n|\mathcal{F}) \downarrow$ a limit Z_∞. If $A \in \mathcal{F}$, then

$$\int_A Z_n dP = \int_A Y_n dP$$

Letting $n \to \infty$, noting $Y_n \downarrow 0$, and using the dominated convergence theorem gives that $\int_A Z_\infty dP = 0$ for all $A \in \mathcal{F}$, so $Z_\infty \equiv 0$.

□

Theorem 5.1.9. *If φ is convex and $\mathbb{E}|X|$, $\mathbb{E}|\varphi(X)| < \infty$, then*

$$\varphi(\mathbb{E}(X|\mathcal{F})) \leq \mathbb{E}(\varphi(X)|\mathcal{F}) \tag{5.4}$$

Proof. If φ is linear, the result is trivial, so we will suppose φ is not linear. We do this so that if we let $S = \{(a,b) : a,b \in \mathbb{Q}, ax + b \leq \varphi(x) \text{ for all } x\}$, then $\varphi(x) = \sup\{ax + b : (a,b) \in S\}$. If $\varphi(x) \geq ax + b$, then 5.1 and 5.2 imply

$$\mathbb{E}(\varphi(X)|\mathcal{F}) \geq a\mathbb{E}(X|\mathcal{F}) + b \text{ a.s.}$$

Taking the sup over $(a,b) \in S$ gives

$$\mathbb{E}(\varphi(X)|\mathcal{F}) \geq \varphi(\mathbb{E}(X|\mathcal{F})) \text{ a.s.}$$

which proves the desired result.

□

Theorem 5.1.10. *Conditional expectation is a contraction in L^p, $p \geq 1$.*

Proof. 5.4 implies $|\mathbb{E}(X|\mathcal{F})|^p \leq \mathbb{E}(|X|^p|\mathcal{F})$. Taking expected values gives

$$\mathbb{E}(|\mathbb{E}(X|\mathcal{F})|^p) \leq \mathbb{E}(\mathbb{E}(|X|^p|\mathcal{F})) = \mathbb{E}|X|^p$$

□

In the last equality, we have used an identity that is an immediate consequence of the definition (use property (ii) in the definition with $A = \Omega$).

$$\mathbb{E}(\mathbb{E}(Y|\mathcal{F})) = \mathbb{E}(Y) \tag{5.5}$$

Conditional expectation also has properties, like 5.5, that have no analogue for "ordinary" expectation.

Theorem 5.1.11. *If $\mathcal{F} \subset \mathcal{F}$ and $\mathbb{E}(X|\mathcal{F}) \in \mathcal{F}$, then $\mathbb{E}(X|\mathcal{F}) = \mathbb{E}(X|\mathcal{G})$.*

Proof. By assumption, $\mathbb{E}(X|\mathcal{G}) \in \mathcal{F}$. To check the other part of the definition, we note that if $A \in \mathcal{F} \subset \mathcal{G}$, then

$$\int_A XdP = \int_A E(X|\mathcal{G})dP$$

□

Theorem 5.1.12. *If $\mathcal{F}_1 \subset \mathcal{F}_2$, then (i) $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_1)|\mathcal{F}_2) = \mathbb{E}(X|\mathcal{F}_1)$, (ii) $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_2|\mathcal{F}_1) = \mathbb{E}(X|\mathcal{F}_1)$.*

Proof. Once we notice that $\mathbb{E}(X|\mathcal{F}_1) \in \mathcal{F}_2$, (i) follows from Example 5.1.2. To prove (ii), notice that $\mathbb{E}(X|\mathcal{F}_1) \in \mathcal{F}_1$, and if $A \in \mathcal{F}_1 \subset \mathcal{F}_2$, then

$$\int_A \mathbb{E}(X|\mathcal{F}_1 dP = \int_A X dP = \int_A \mathbb{E}(X|\mathcal{F}_2) dP$$

□

Theorem 5.1.13. *If $X \in \mathcal{F}$ and $\mathbb{E}|Y|$, $\mathbb{E}|XY| < \infty$, then*

$$\mathbb{E}(XY|\mathcal{F}) = X\mathbb{E}(Y|\mathcal{F}).$$

Proof. The right-hand side $\in \mathcal{F}$, so we have to check (ii). To do this, we use the usual four-step procedure. First, suppose $X = 1_B$ with $B \in \mathcal{F}$. In this case, if $A \in \mathcal{F}$,

$$\int_A 1_B \mathbb{E}(Y|\mathcal{F}) dP = \int_{A\cap B} \mathbb{E}(Y|\mathcal{F}) dP = \int_{A\cap B} Y dP = \int_A 1_B Y dP$$

so (ii) holds. The last result extends to simple X by linearity. If $X, Y \geq 0$, let X_n be simple random variables that $\uparrow X$, and use the monotone convergence theorem to conclude that

$$\int_A X\mathbb{E}(Y|\mathcal{F}) dP = \int_A XY dP$$

To prove the result in general, split X and Y into their positive and negative parts.

□

Theorem 5.1.14. *Suppose $\mathbb{E}X^2 < \infty$. $\mathbb{E}(X|\mathcal{F})$ is the variable $Y \in \mathcal{F}$ that minimizes the "mean square error" $\mathbb{E}(X - Y)^2$.*

Remark 5.1.15. This result gives a "geometric interpretation" of $\mathbb{E}(X|\mathcal{F})$, $L^2(\mathcal{F}_0 = \{Y \in \mathcal{F}_0 : \mathbb{E}Y^2 < \infty\}$ is a Hilbert space, and $L^2(\mathcal{F})$ is a closed subspace. In this case, $\mathbb{E}(X|\mathcal{F})$ is the projection of X onto $L^2(\mathcal{F})$. That is, the point in the subspace closest to X.

Proof

We begin by observing that if $Z \in L^2(\mathcal{F})$, then Theorem 5.1.13 implies

$$Z\mathbb{E}(X|\mathcal{F}) = \mathbb{E}(ZX|\mathcal{F})$$

($\mathbb{E}|XZ| < \infty$ by the Cauchy-Schwarz inequality.) Taking expected values gives

$$\mathbb{E}(Z\mathbb{E}(X|\mathcal{F})) = \mathbb{E}(\mathbb{E}(ZX|\mathcal{F})) = \mathbb{E}(ZX)$$

or, rearranging,

$$\mathbb{E}[Z(X - \mathbb{E}(X|\mathcal{F}))] = 0 \text{ for } Z \in L^2(\mathcal{F})$$

If $Y \in L^2(\mathcal{F})$ and $Z = \mathbb{E}(X|\mathcal{F}) - Y$, then

$$\mathbb{E}(X - Y) = \mathbb{E}\{X - \mathbb{E}(X|\mathcal{F}) + Z\}^2 = \mathbb{E}\{X - \mathbb{E}(X|\mathcal{F})\}^2 + \mathbb{E}Z^2$$

since the cross-product term vanishes. From the last formula, it is easy to see $\mathbb{E}(X-Y)^2$ is minimized when $Z = 0$.

□

5.2 Martingales, A.S. Convergence

In this section we will define martingales and their cousins supermartingales and submartingales, and take the first steps in developing their theory. Let $\mathcal{F}_n$ be a **filtration**, that is, an increasing sequence of σ-fields. A sequence X_n is said to be **adapted** to $\mathcal{F}_n$ if $X_n \in \mathcal{F}_n$ for all n. If X_n is sequence with

(i) $\mathbb{E}|X_n| < \infty$,

(ii) X_n is adapted to $\mathcal{F}_n$,

(iii) $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$ for all n,

then X is said to be a **martingale** (with respect to $\mathcal{F}_n$). If in the last definiition, $=$ is replaced by $\leq$ or $\geq$, then X is said to be a **supermartingale** or . respectively.

Example 5.2.1. Simple random walk. Consider the successive tosses of a fair coin and let $\xi_n = 1$ if the nth toss is heads and $\xi_n = -1$ if the nth toss is tails. Let $X_n = \xi_1 + \cdots + \xi_n$ and $\mathcal{F}_n = \sigma(\xi_1, ..., \xi_n)$ for $n \geq 1$, $X_0 = 0$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$. I claim that $X_n \in \mathcal{F}_n$, $\mathbb{E}|X_n| < \infty$, and ξ_{n+1} is independent of $\mathcal{F}_n$, so using the linearity of conditional expectation, 5.1, and Example 5.1.3,

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) = \mathbb{E}(X_n|\mathcal{F}_n) + \mathbb{E}(\xi_{n+1}|\mathcal{F}_n) + \mathbb{E}(\xi_{n+1}|\mathcal{F}_n) = X_n + \mathbb{E}\xi_{n+1} = X_n$$

Example 5.2.2. Superharmonic functions. If the coin tosses considered above have $P(\xi_n = 1) \leq 1/2$ then the computation just completed shows $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \leq X_n$, i.e. X_n is a supermartingale. In this case, X_n corresponds to betting on an unfavorable game so there is a supermartingale. The name comes from the fact that if f is superharmonic (i.e., f has continuous derivatives of order ≤ 2 and $\partial^2 f/\partial x_1^2 + \cdots + \partial^2 f/\partial x_d^2 \leq 0$), then

$$f(x) \leq \frac{1}{|B(0,r)|} \int_{B(x,r)} f(y)dy$$

where $B(x,r) = \{y : |x - y| \leq r\}$ is the ball of radius r, and $|B(0,r)|$ is the volume of the ball of radius r.

Theorem 5.2.3. *If X_n is a supermartingale then for $n > m$, $\mathbb{E}(X_n|\mathcal{F}_m) \leq X_m$.*

Proof. The definition gives the result for $n = m + 1$. Suppose $n = m + k$ with $k \geq 2$. By Theorem 5.1.8,

$$\mathbb{E}(X_{m+k}|\mathcal{F}_m) = \mathbb{E}(\mathbb{E}(X_{m+k}|\mathcal{F}_{m+k-1})|\mathcal{F}_m) \leq \mathbb{E}(X_{m+k-1}|\mathcal{F}_m)$$

by the definition and 5.1. The desired result now follows by induction.

□

Theorem 5.2.4. *(i) If X_n is a submartingale, then for $n > m$, $\mathbb{E}(X_n|\mathcal{F}_m) \geq X_m$. (ii) If X_n is a martingale then for $n > m$, $\mathbb{E}(X_n|\mathcal{F}_m) = X_m$.*

Proof. To prove (i), note that $-X_n$ is a supermartingale and use 5.1. For (ii), observe that X_n is a supermartingale and a submartingale.

□

Remark 5.2.5. The idea in the proof of Theorem 5.2.4 can be used many times below. To keep from repeating ourselves, we will just state the result for either supermartingales or submartinagles and leave it to the reader to translate the result for the other two.

Theorem 5.2.6. *If X_n is a martingale w.r.t. $\mathcal{F}_n$ and φ is a convex function with $\mathbb{E}|\varphi(X_n)| < \infty$ for all n, then $\varphi(X_n)$ is a submartingale w.r.t. $\mathcal{F}_n$. Consequently, if $p \geq 1$ and $\mathbb{E}|X_n|^p < \infty$ for all n, then $|X_n|^p$ is a submartingale w.r.t. $\mathcal{F}_n$.*

Proof. By Jensen's inequality and the definition,

$$\mathbb{E}(\varphi(X_{n+1})|\mathcal{F}_n) \geq \varphi(\mathbb{E}(X_{n+1}|\mathcal{F}_n)) = \varphi(X_n)$$

□

Theorem 5.2.7. *If X_n is a submartingale w.r.t. $\mathcal{F}_n$ and φ is an increasing convex function with $\mathbb{E}|\varphi(X_n)| < \infty$ for all n, then $\varphi(X_n)$ is a submartingale w.r.t. $\mathcal{F}_n$. Consequently (i) If X_n is a submartingale, then $(X_n - a)^+$ is a submartingale, (ii) If X_n is a supermartingale, then $X_n \wedge a$ is a supermartingale.*

Proof. By Jensen's inequality and the assumptions,

$$\mathbb{E}(\varphi(X_{n+1})|\mathcal{F}_n) \geq \varphi(\mathbb{E}(X_{n+1}|\mathcal{F}_n)) \geq \varphi(X_n)$$

□

Let $\mathcal{F}_n$, $n \geq 0$ be a filtration. H_n, $n \geq 1$ is said to be a **predictable sequence** if $H_n \in \mathcal{F}_{n-1}$ for all $n \geq 1$. In words, the value of H_n may be predicted (with certainty) from the information available at time $n-1$. In this section, we will be thinking of H_n as the amount of money a gambler will bet at time n. This can be based on the outcomes at times $12, ..., n-1$, but not on the outcome at time n!

Once we start thinking of H_n as a gambling system, it is natural to ask how much money we would make if we used it. For concreteness, let us suppose that game consists of flipping a coin and that for each dollar you bet, you win one dollar when the coin comes up heads and lose your dollar when the coin comes up tails. Let X_n be the net amount of money you would have won at time n if you had bet one dollar each time. If you bet according to a gambling system H, then your winnings at time n would be

$$(H \cdot X)_n = \sum_{m=1}^{n} H_m(X_m - X_{m-1})$$

since $X_m - X_{m-1} = +1$ or -1 when the mth toss results in a win or loss, respectively.

Let $\xi_m = X_m - X_{m-1}$. A famous gambling system called the "martingale" is defined by $H_1 = 1$ and for $n \geq 2$, $H_n = 2H_{n-1}$ if $\xi_{n-1} = -1$ and $H_n = 1$ if $\xi_{n-1} = 1$. In words, we double our bet when we lose, so that if we lose k times and then win, our net winnings will be $-1 - 2 ... - 2^{k-1} + 2^k = 1$. This system seems to provide us within a "sure thing" as long as $P(\xi_m = 1) > 0$. However, the next result says there is no system for beating an unfavorable game.

Theorem 5.2.8. *Let X_n, $n \geq 0$, be a supermartingale. If $H_n \geq 0$ is predictable and each H_n is bounded then $(H \cdot X)_n$ is a supermartingale.*

Proof. Using the fact that conditional expectation is linear, $(H \cdot X)_n \in \mathcal{F}_n$, $H_n \in \mathcal{F}_{n-1}$, we have

$$\begin{aligned} \mathbb{E}((H \cdot X)_{n+1}|\mathcal{F}_n) &= (H \cdot X)_n + \mathbb{E}(H_{n+1}(X_{n+1} - X_n)|\mathcal{F}_n) \\ &= (H \cdot X)_n + H_{n+1}\mathbb{E}((X_{n+1} - X_n)|\mathcal{F}_n) \leq (H \cdot X)_n \end{aligned}$$

since $\mathbb{E}((X_{n+1} - X_n)|\mathcal{F}_n) \leq 0$ and $H_{n+1} \geq 0$.

Remark 5.2.9. The same result is obviously true for submartingales and for martingales (in the last case, without the restriction $H_n \geq 0$).

The notion of a stopping time, introduced in §4, is closely related to the concept of a gambling system. Recall that a random variable N is said to be a **stopping time** if $\{N = n\} \in \mathcal{F}_n$ for all $n < \infty$. If you think of N as the time a gambler stops gambling, then the condition above says that the decision to stop at time n must be measurable with respect to the information he has at that time. If we let $H_n = 1_{\{N \geq n\}}$, then $\{N \geq n\} = \{N \leq n-1\}^c \in \mathcal{F}_{n-1}$, so H_n is predictable, and it follows from Theorem 5.2.8 that $(H \cdot X)_n = X_{N \wedge n} - X_0$ is a supermartingale. Since the constant sequence $Y_n = X_0$ is a supermartaingle and the sum of two supermartingales is also, we have:

Theorem 5.2.10. *If N is a stopping time and X_n is a supermartingale, then $X_{N\wedge n}$ is a supermartingale.*

Although you cannot make money with gambling systems, you can prove theorems with them. Suppose X_n, $n \geq 0$, is a submartingale. Let $a < b$, let $N_0 = -1$, and for $k \geq 1$ let

$$N_{2k-1} = \inf\{m > N_{2k-2} : X_m \leq a\}$$
$$N_{2k} = \inf\{m > N_{2k-1} : X_m \geq b\}$$

The N_j are stopping times, and $\{N_{2k-1} < m \leq N_{2k}\} = \{N_{2k-1} \leq m-1\} \cap \{N_{2k} \leq m-1\}^c \in \mathcal{F}_{m-1}$, so

$$H_m = \begin{cases} 1 & \text{if } N_{2k-1} < m \leq N_{2k} \text{ for some } k \\ 0 & \text{otherwise} \end{cases}$$

The N_j are stopping times, and $\{X(N_{2k-1}) \leq a$ and $X(N_{2k}) \geq b$, so between times N_{2k-1} and N_{2k}, X_m crosses from below a to above b. H_m is a gambling system that tries to take advantage of these "upcrossings." In stock market terms, we buy when $X_m \leq a$ and sell when $X_m \geq b$, so every time an upcrossing is completed, we make a profit of $\geq (b-a)$. Finally, $U_n = \sup\{k : N_{2k} \leq n\}$ is the number of upcrossings completed by time n.

Theorem 5.2.11. ***Upcrossing inequality***. *If X_m, $m \geq 0$, is a submartingale, then*

$$(b-a)\mathbb{E}U_n \leq \mathbb{E}(X_n - a)^+ - \mathbb{E}(X_0 - a)^+$$

Proof. Let $Y_m = a + (X_m - a)^+$. By Theorem 5.2.7, Y_m is a submartingale. Clearly, it upcrosses $[a, b]$ the same number of times that X_m does, and we have $(b-a)U_n \leq (H \cdot Y)_n$, since each upcrossing results in a profit $\geq (b-a)$, and a final incomplete upcrossing (if there is one) makes a nonnegative contribution to the right-hand side. Let $K_m = 1 - H_m$. Clearly, $Y_n - Y_0 = (H \cdot Y)_n + (K \cdot Y)_n$, and it follows from Theorem 5.2.8 that $\mathbb{E}(K \cdot Y)_n \geq \mathbb{E}(K \cdot Y)_0 = 0$, so $\mathbb{E}(H \cdot Y)_n \leq \mathbb{E}(Y_n - Y_0)$, proving the desired inequality.

□

From the upcrossing inequality, we easily get

Theorem 5.2.12. ***Martingale convergence theorem***. *If X_n is a submartingale with $\sup \mathbb{E}X_n^+ < \infty$, then as $n \to \infty$, X_n convergence a.s. to a limit X with $\mathbb{E}|X| < \infty$.*

Proof. Since $(X-a)^+ \leq X^+ + |a|$, Theorem 5.2.11 implies that

$$\mathbb{E}U_n \leq (|a| + \mathbb{E}X_n^+)/(b-a)$$

As $n \uparrow \infty$, $U_n \uparrow U$ the number of upcrossings of $[a, b]$ by the whole sequence, so if $\sup \mathbb{E}X_n^+ < \infty$, then $\mathbb{E}U < \infty$ and hence $U < \infty$ a.s. Since the last conclusion holds for all rational a and b,

$$\cup_{a,b\in\mathbb{Q}}\{\liminf X_n < a < b < \limsup X_n\} \text{ has probability } 0$$

and hence $\limsup X_n = \liminf X_n$ a.s., that is, $\lim X_n$ exists a.s. Fatou's lemma guarantees $\mathbb{E}X^+ \leq \liminf \mathbb{E}X_n^+ < \infty$, so $X < \infty$ a.s. To see $X > -\infty$, we observe that

$$\mathbb{E}X_n^- = \mathbb{E}X_n^+ - \mathbb{E}X_n \leq \mathbb{E}X_n^+ - \mathbb{E}X_0$$

(since X_n is a submartingale), so another application of Fatou's lemma shows

$$\mathbb{E}X^- \leq \liminf_{n\to\infty} \mathbb{E}X_n^- \leq \sup_n \mathbb{E}X_n^+ - \mathbb{E}X_0 < \infty$$

and completes the proof.

□

A special case of Theorem 5.2.12 is

Theorem 5.2.13. *If $X_n \geq 0$ is a supermartingale, then as $n \to \infty$, $X_n \to X$ a.s. and $\mathbb{E}X \leq \mathbb{E}X_0$.*

Proof. $Y_n = -X_n \leq 0$ is a submartingale with $\mathbb{E}Y_n^+ = 0$. Since $\mathbb{E}X_0 \geq \mathbb{E}X_n$, the inequality follows from Fatou's lemma.

□

In the next section, we will give several applications of the last two results. We close this one by giving two "counterexamples."

Example 5.2.14. The first shows that the assumptions of Theorem 5.2.13 do not guarantee convergence in L^1. Let S_n be a symmetric simple random walk with $S_0 = 1$, that is, $S_n = S_{n-1} + \xi_n$ where $\xi_1, \xi_2, ...$ are i.i.d. with $P(\xi_i = 1) = P(\xi_i = -1) = 1/2$. Let $N = \inf\{n : S_n = 0\}$ and let $X_n = S_{N\wedge n}$. Theorem 5.2.10 implies X_n converges to a limit $X_\infty < \infty$ that must be $\equiv 0$, since convergence to $k > 0$ is impossible. (If $X_n = k > 0$, then $X_{n+1} = k \pm 1$.) Since $\mathbb{E}X_n = \mathbb{E}X_0 = 1$ for all n and $X_\infty = 0$, convergence cannot occur in L^1.

Theorem 5.2.15. ***Doob's decomposition***. *Any submartingale X_n, $n \geq 0$, can be written in a unique way as $X_n = M_n + A_n$, where M_n is a martingale and A_n is a predictable increasing sequence with $A_0 = 0$.*

Proof. We want $X_n = M_n + A_n$, $\mathbb{E}(M_n|\mathcal{F}_{n-1}) = M_{n-1}$, and $A_n \in \mathcal{F}_{n-1}$. So we must have

$$\begin{aligned} \mathbb{E}(X_n|\mathcal{F}) &= \mathbb{E}(M_n|\mathcal{F}_{n-1}) + \mathbb{E}(A_n|\mathcal{F}_{n-1}) \\ &= M_{n-1} + A_n \\ &= X_{n-1} - A_{n-1} + A_n \end{aligned}$$

and it follows that

(a) $A_n - A_{n-1} = \mathbb{E}(X_n|\mathcal{F}_{n-1}) - X_{n-1}$

(b) $M_n = X_n - A_n$

Now $A_0 = 0$ and $M_0 = X_0$ by assumption, so we have A_n and M_n defined for all time, and we have proved uniqueness. To check that our recipe works, we observe that $A_n - A_{n-1} \geq 0$ since X_n is a submartingale and induction shows $A_n \in \mathcal{F}_{n-1}$. To see that M_n is a martingale, we use (b), $A_n \in \mathcal{F}_{n-1}$ and (a):

$$\begin{aligned} \mathbb{E}(M_n|\mathcal{F}_{n-1}) &= \mathbb{E}(X_n - A_n|\mathcal{F}_{n-1}) \\ &= \mathbb{E}(X_n|\mathcal{F}_{n-1}) - A_n \\ &= X_{n-1} - A_{n-1} \\ &= M_{n-1} \end{aligned}$$

which completes the proof.

□

5.3 Examples

In this section, we will apply the martingale convergence theorem to generalize the second Borel-Cantelli lemma and to study Polya's urn scheme, Radon Nikodym derivatives, and branching processes. The four topics are independent of each other and are taken up in the order indicated.

5.3.1 Bounded Increments

Our first result shows that martingales with bounded increments either converge or oscillate between $+\infty$ and $-\infty$.

Theorem 5.3.1. *Let $x_1, x_2, \ldots$ be a martingale with $|X_{n+1} - X_n| \leq M < \infty$. Let*

$$C = \{\lim X_n \text{ exists and is finite}\}$$

$$D = \{\limsup X_n = +\infty \text{ and } \liminf X_n = -\infty\}$$

Then $P(C \cup D) = 1$.

Proof. Since $X_n - X_0$ is a martingale, we can without loss of generality suppose that $X_0 = 0$. Let $0 < K < \infty$ and let $N = \inf\{n : X_n \leq -K\}$. $X_{n \wedge N}$ is a martingale with $X_{n \wedge N} \geq -K - M$ a.s. so applying Theorem 5.2.15 to $X_{n \wedge N} + K + M$ shows $\lim X_n$ exists on $\{N = \infty\}$. Letting $K \to \infty$, we see that the limit exists on $\{\liminf X_n > -\infty\}$. Applying the last conclusion to $-X_n$, we see that $\lim X_n$ exists on $\{\limsup X_n < \infty\}$ and the proof is complete.

□

Theorem 5.3.2. ***Second Borel-Cantelli lemma II.*** *Let $\mathcal{F}_n$, $n \geq 0$ be a filtration with $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and A_n, $n \geq 1$ a sequence of events with $A_n \in \mathcal{F}_n$. Then*

$$\{A_n \text{ i.o.}\} = \left\{ \sum_{n=1}^{\infty} P(A_n | \mathcal{F}_{n-1}) = \infty \right\}$$

Proof. If we let $X_0 = 0$ and $X_n = \sum_{m=1}^{n} 1_{A_m} - P(A_m | \mathcal{F}_{m=1})$ for $n \geq 1$, then X_n is a martingale with $|X_n - X_{n-1}| \leq 1$. Using the notation of Theorem 5.3.1, we have

$$\text{on } C, \sum_{n=1}^{\infty} 1_{A_n} = \infty \text{ if and only if } \sum_{n=1}^{\infty} P(A_n | \mathcal{F}_{n-1}) = \infty$$

$$\text{on } D, \sum_{n=1}^{\infty} 1_{A_n} = \infty \text{ and } \sum_{n=1}^{\infty} P(A_n | \mathcal{F}_{n-1}) = \infty$$

Since $P(C \cup D) = 1$, the result follows.

□

5.3.2 Polya's Urn Scheme

An urn contains r red and g green balls. At eeach time we draw a ball out, then replace it, and add c more balls of the color drawn. Let X_n be the fraction of green balls after the nth draw. To check that X_n is a martingale, note that if these are i red balls and j green balls at time n, then

$$X_{n+1} = \begin{cases} (j+c)/(i+j+c) & \text{with probability } j/(i+j) \\ j/(i+j+c) & \text{with probability } i/(i+j) \end{cases}$$

and we have

$$\frac{i+c}{i+j+c} \cdot \frac{j}{i+j} + \frac{j}{i+j+c} \cdot \frac{i}{i+j} = \frac{(j+c+i)j}{(i+j+c)(i+j)} = \frac{j}{i+j}$$

Since $X_n \geq 0$, Theorem 5.2.13 implies that $X_n \to X_\infty$ a.s. To compute the distribution of the limit, we observe (a) the probability of getting green on the first m draws then red on the next $l = n - m$ draws is

$$\frac{g}{g+1} \cdot \frac{g+c}{g+r+c} \cdots \frac{g+(m-1)c}{g+r+(m-1)c} \cdot \frac{r}{g+r+mc} \cdots \frac{r+(l-1)c}{g+r+(n-1)c}$$

and 9b) any other outcome of the first n draws with m green balls drawn and l red balls drawn has the same probability since the denominator remains the same and be the number of green balls after the nth draw has been completed and the new ball has been added. It follows from (a) and (b) that

$$P(G_n = m+1) = \binom{n}{m} \frac{m!(n-m)!}{(n+1)!} = \frac{1}{n+1}$$

so X_∞ has a uniform distribution on (0,1).

If we suppose that $c = 1$, $g = 2$, and $r = 1$, then

$$P(G_n = m+2) = \frac{n!}{m!(n-m)!} \frac{(m+1)!(n-m)!}{(n+2)!/2} \to 2x$$

if $n \to \infty$ and $m/n \to x$. In general, the distribution of X_∞ has density

$$\frac{\Gamma((g+r)/c)}{\Gamma(g/c)\Gamma(r/c)} x^{(g/c)-1}(1-x)^{(r/c)-1}$$

This is the **beta distribution** with parameters g/c and r/c.

5.3.3 Radon-Nikodym Derivatives

Let μ be a finite measure and ν a probability measure on $(\Omega, \mathcal{F})$. Let $\mathcal{F}_n \uparrow \mathcal{F}$ be σ-fields (i.e., $\sigma(\cup \mathcal{F}_n) = \mathcal{F}$). Let μ_n and ν_n be the restrictions of μ and ν to $\mathcal{F}_n$.

Theorem 5.3.3. *Suppose $\mu_n << \nu_n$ for all n. Let $X_n = d\mu_n/d\nu_n$ and let $X = \limsup X_n$. Then*

$$\mu(A) = \int_A X d\nu + \mu(A \cap \{X = \infty\})$$

Remark 5.3.4. $\mu_r(A) \equiv \int_A X d\nu$ is a measure $<< \nu$. Since Theorem 5.2.13 implies $\nu(X) = \infty) = 0$, $\mu_s(A) \equiv \mu(A \cap \{X = \infty\})$ is a singular w.r.t. ν. Thus $\mu = \mu_r + \mu_s$ gives the Lebesgue decomposition of μ and $X_\infty = d\mu_r/d\nu$, ν-a.s. Here and in the proof we need to keep track of the measure to which the a.s. refers. Please refer to Durrett [4] Page 242 for detailed proof.

5.3.4 Branching Processes

Let ξ_i^n, i, $n \geq 1$, be i.i.d. nonnegative integer-valued random variables. Define a sequence Z_n, $n \geq 0$ by $Z_0 = 1$ and

$$Z_{n+1} = \begin{cases} \xi_1^{n+1} + \cdots + \xi_{Z_n}^{n+1} & \text{if } Z_n > 0 \\ 0 & \text{if } Z_n = 0 \end{cases} \tag{5.6}$$

Z_n is called a **Galton-Wawtson process**. The idea behind the definitions is that Z_n is the number of individuals in the nth generation, and each member of the nth generation gives birth independently to an identically distributed number of children. $p_k = P(\xi_i^n = k)$ is called the **offspring distribution**.

Lemma 5.3.5. *Let $\mathcal{F}_n = \sigma(\xi_i^m : i \geq 1, 1 \leq m \leq n)$ and $\mu = \mathbb{E}\xi_i^m \in (0, \infty)$. Then Z_n/μ^n is a martingale w.r.t. $\mathcal{F}_n$.*

Proof. Clearly, $Z_n \in \mathcal{F}_n$,

$$\mathbb{E}(Z_{n+1}|\mathcal{F}_n) = \sum_{k=1}^{\infty} \mathbb{E}(Z_{n+1} 1\{Z_n = k\}|\mathcal{F}_n)$$

by the linearity of conditional expectation, 5.1, and the monotone convergence theorem, 5.3. On $\{Z_n = k\} = \xi_1^{n+1} + \cdots + \xi_k^{n+1}$, so the sum is

$$\sum_{k=1}^{\infty} \mathbb{E}((\xi_1^{n+1} + \cdots + \xi_k^{n+1}) 1_{\{Z_n=k\}} | \mathcal{F}_n) = \sum_{k=1}^{\infty} 1_{\{Z_n=k\}} \mathbb{E}(\xi_1^{n+1} + \cdots + \xi_k^{n+1} | \mathcal{F}_n)$$

by Theorem 5.1.13. Since each ξ_j^{n+1} is independent of $\mathcal{F}_n$, the last expression

$$= \sum_{k=1}^{\infty} 1_{\{Z_n=k\}} k\mu = \mu Z_n$$

Dividing both sides by μ^{n+1} now gives the desired result.

□

Theorem 5.3.6. *If $\mu < 1$ then $Z_n = 0$ for all n sufficiently large, so $Z_n/\mu^n \to 0$.*

Proof. $\mathbb{E}(Z_n/\mu^n) = \mathbb{E}(Z_0) = 1$, so $\mathbb{E}(Z_n) = \mu^n$. Now $Z_n \geq 1$ on $\{Z_n > 0\}$ so

$$P(Z_n > 0) \leq \mathbb{E}(Z_n; Z_n > 0) = \mathbb{E}(Z_n) = \mu^n \to 0$$

exponentially fast if $\mu < 1$.

□

5.4 Martingales and Markov Chain

One of the ways of specifying the joint distribution of a sequence $X_0, X_1, ..., X_n$ of random variables is to specify the distribution of X_0 and, for each $j \geq 1$, specify the conditional distribution of X_j given the σ-field $\mathcal{F}_{j-1}$ generated by $X_0, X_1, ..., X_{j-1}$. Equivalently, instead of the conditional distributions one can specify the conditional expectations $\mathbb{E}[f(X_j)|\mathcal{F}_{j-=1}]$ for $1 \leq j \leq n$. Let us write

$$h_{j-1}(X_0, X_1, ..., X_{j-1}) = \mathbb{E}[f(X_j)|\mathcal{F}_{j-1}] - f(X_{j-1})$$

so that, for $1 \leq j \leq n$,

$$\mathbb{E}[\{f(X_j - f(X_{j-1}) - h_{j-1}(X_0, X_1, ..., X_{j-1})\}|\mathcal{F}_{j-1}] = 0$$

or

$$Z_j^f = f(X_j) - f(X_0) - \sum_{i=1}^{j} h_{i-1}(X_0, X_1, ..., X_{i-1})$$

is a martingale for every f. It is not difficult to see that the specification of $\{h_i\}$ for each f is enough to determine all the successive conditional expectations and therefore the conditional distributions. If, in addition, the initial distribution of X_0 is specified, then the distribution of $(X_0, X_1, ..., X_n)$ is completely determined.

If, for each j and f, the corresponding $h_{j-1}(X_0, X_1, ..., X_{j-1})$ is a function $_{j-1}(_{j-1})$ of X_{j-1} only, then the distribution of $(X_0, X_1, ..., X_n)$ is Markov and the transition probabilities are seen to be given by the relation

$$\begin{aligned} h_{j-1}(X_{j-1}) &= \mathbb{E}[[f(X_j) - f(X_{j-1})]|\mathcal{F}_{j-1}] \\ &= \int [f(y) - f(X_{j-1})] \pi_{j-1,j}(X_{j-1}, dy) \end{aligned}$$

In the case of a stationary Markov chain, the relationship is

$$\begin{aligned} h_{j-1}(X_{j-1}) &= h(X_{j-1}) \\ &= \mathbb{E}[[f(X_j) - f(X_{j-1})]|\mathcal{F}_{j-1}] \\ &= \int [f(y) - f(X_{j-1})]\pi(X_{j-1}, dy) \end{aligned}$$

If we introduce the linear transformation (transition operator)

$$(\Pi f)(x) = \int f(y)\pi(x, dy), \tag{5.7}$$

then

$$h(x) = ([\Pi - I]f)(x).$$

Remark 5.4.1. In the case of a Markov chain on a countable state space,

$$(\Pi f)(x) = \sum_y \pi(x, y)f(y)$$

and

$$h(x) = [\Pi - I]f(x) = \sum_y [f(y) - f(x)]\pi(x, y).$$

Remark 5.4.2. The measure P_x on the space $(\Omega, \mathcal{F})$ of sequences $\{x_j : j \geq 0\}$ from the state space X that corresponds to the Markov process with transition probability $\pi(x, dy)$ and initial state x can be characterized as the unique measure on $(\Omega, \mathcal{F})$ such that

$$P_x\{\omega : x_0 = x\} = 1,$$

and for every bounded measurable function f defined on the state space X,

$$f(x_n) - f(x_0) - \sum_{j=1}^{n} h(x_{j-1})$$

is a martingale with respect to $(\Omega, \mathcal{F}_n, P_x)$, where

$$h(x) = \int_x [f(y) - f(x)]\pi(x, dy).$$

Let $A \subset X$ be a measurable subset and let $\tau_A = \inf\{n \geq 0 : x_n \in A\}$ be the first entrance time into A. It is easy to see that τ_A is a stopping time. It need not always be true that $P_x\{\tau_A < \infty\} = 1$. But $U_A(x) = P_x\{\tau_A < \infty\}$ is a well-defined measurable function of x that satisfies $0 \leq U(x) \leq 1$ for all x and is the exit probability from the set A^c. By its very definition $U_A(x) \equiv 1$ on A, and if $x \notin A$, by the Markov property,

$$U_A(x) = \pi(x, A) + \int_{A^c} U_A(y)\pi(x, dy) = \int_X U_A(y)\pi(x, dy).$$

In other words, U_A satisfies $0 \leq U_A \leq 1$ and is a solution of

$$\begin{aligned} (\Pi - I)V &= 0 \qquad \text{on } A^c \\ V &= 1 \qquad \text{on } A. \end{aligned} \tag{5.8}$$

Theorem 5.4.3. *Among all nonnegative solutions V of equation 5.8 $U_A(x) = P_x\{\tau_A < \infty\}$ is the smallest. If $U_A(x) = 1$, then any bounded solution of the equation*

$$\begin{aligned} (\Pi - I)V &= 0 \qquad \textit{on } A^c \\ V &= f \qquad \textit{on } A. \end{aligned} \tag{5.9}$$

is equal to

$$V(x) = \mathbb{E}^{P_x}\{f(x_{\tau_A})\}. \tag{5.10}$$

In particular, if $U_A(x) = 1$ for all $x \notin A$, then any bounded solution V of equation 5.9 is unique and is given by formula 5.9.

Proof. First we establish that any nonnegative solution V of 5.8 dominates U_A. Let us replace V by $W = \min(V, 1)$. Then $0 \leq W \leq 1$ everywhere, $W(x) = 1$ for $x \in A$, and for $x \notin A$,

$$(\Pi W)(x) = \int_X W(y)\pi(x, dy) \leq \int_X V(y)\pi(x, dy) = V(x).$$

Since $\Pi W \leq 1$ as well, we conclude that $\Pi W \leq W$ on A^c. On the other hand, it is obvious that $\Pi W \leq 1 = W$ on A. Since we have shown that $\Pi W \leq W$ everywhere, it follows that $\{W(x_n)\}$ is a supermartingale with respect to $(\Omega, \mathcal{F}_n, P_x)$. In particular, for any bounded stopping time τ,

$$\mathbb{E}^{P_x}\{W(x_\tau)\} \leq \mathbb{E}^{P_x}\{W(x_0)\} = W(x).$$

While we cannot take $\tau = \tau_A$ (since τ_A may not be bounded), we can always take $\tau = \tau_N = \min(\tau_A, N)$ to conclude

$$\mathbb{E}^{P_x}\{W(x_{\tau_N})\} \leq \mathbb{E}^{P_x}\{W(x_0)\} = W(x).$$

While we cannot take $\tau = \tau_A$ (since τ_A may not be bounded), we can always take $\tau = \tau_N = \min(\tau_A, N)$ to conclude

$$\mathbb{E}^{P_x}\{W(x_{\tau_N}\} \leq \mathbb{E}^{P_x}\{W(x_0)\} = W(x)$$

Let $N \to \infty$. On the set $\{\omega : \tau_A(\omega)\}$, $\tau_N \uparrow \tau_A$ and $W(x_{\tau_N}) \to W(x_{\tau_A}) = 1$. Since W is nonnegative and bounded,

$$\begin{aligned} W(x) &\geq \limsup_{N\to\infty} \mathbb{E}^{P_x}\{W(x_{\tau_N}\} \\ &\geq \limsup_{N\to\infty} \int -\tau_A < \infty W(x_{\tau_N}) dP_x \\ &= P_x\{\tau_A < \infty\} = U_A(x) \end{aligned}$$

Since $V(x) \geq W(x)$, it follows that $V(x) \geq U_A(x)$.

For a bounded solution V of 5.9, let us defined $h = (\Pi - I)V$, which will be a function vanishing on A^c. We know that

$$V(x_n) - V(x_0) - \sum_{j=1}^{n} h(x_{j-1})$$

is a martingale with respect to $(\Omega, \mathcal{F}_n, P_x)$, and let us use the stopping theorem with $\tau_N = \min(\tau_A, N)$. Since $h(x_{j-1}) = 0$ for $j \leq \tau_A$, we obtain

$$V(x) = \mathbb{E}^{P_x}\{V(x_{\tau_N})\}.$$

If we now make the assumption that $U_A(x) = P_x\{\tau_A < \infty\} = 1$, let $N \to \infty$, and use the bounded convergence theorem, it is easy to see that

$$V(x) = \mathbb{E}^{P_x}\{f(x_{\tau_A})\},$$

which proves 5.10 and the rest of the theorem.

□

6 MARKOV CHAINS

Go back to Table of Contents. Please click TOC

This section we refer to text [16]. One of the ways of generating a sequence of dependent random variables is to think of a system evolving in time. We have time points that are discrete say $T = 0, 1, ..., N, ...$ The state of the system is described by a point x in the state space $\mathcal{X}$ of the system. The state space $\mathcal{X}$ comes with a natural σ-field of subsets $\mathcal{F}$. At time 0 the system is in a random state and its distribution is specified by a probability distribution μ_0 on $(\mathcal{X}, \mathcal{F})$. At successive times $T = 1, 2, ...,$ the system changes its state and given the past history $(x_0, ..., x_k)$ of the states of the system at times $T = 0, ..., k-1$ the probability that system finds itself at time k in a subset. $A \in \mathcal{F}$ is given by $\pi_k(x_0, ..., x_{k-1}; A)$. For each $(x_0, ..., x_{k-1})$, π_k defines a probability measure on $(\mathcal{X}, \mathcal{F})$ and for each $A \in \mathcal{F}$, $\pi_k(x_0, ..., x_{k-1}; A)$ is assumed to be a measurable function of $(x_0, ..., x_{k-1})$, on the space $(\mathcal{X}^k, \mathcal{F}^k)$ which is the product of k copies of the space $(\mathcal{X}, \mathcal{F})$ with itself. We can inductively define measures μ_k on $(\mathcal{X}^{k+1}, \mathcal{F}^{k+1})$ that describe the probability distribution of the entire history $(x_0, ..., x_k)$ of the system through time k. To go from μ_{k-1} to μ_k we think of $\mathcal{X}^{k+1}, \mathcal{F}^{k+1})$ as the product of $(\mathcal{X}^k, \mathcal{F}^k)$ with $(\mathcal{X}, \mathcal{F})$ and construct on $(\mathcal{X}^{k+1}, \mathcal{F}^{k+1})$ a probability measure with marginal μ_{k-1} on $(\mathcal{X}^k, \mathcal{F}^k)$ and conditionals $\pi_k(x_0, ..., x_{k-1}; \cdot)$ on the fibers $(x_1, ..., x_{k-1}) \times \mathcal{X}$. This will define μ_k and the induction can proceed. We may stop at some finite terminal time N or go on indefinitely. If we do go on indefinitely, we will have a consistent family of finite dimensional distributions $\{\mu_k\}$ on $(\mathcal{X}^{k+1}, \mathcal{F}^{k+1})$ and we may try to use Kolmogorov's theorem to construct a probability measure P on the space $(\mathcal{X}^\infty, \mathcal{F}^\infty)$ of sequences $\{x_j : j \geq 0\}$ representing the total evolution of the system for all times.

Remark 6.0.1. However Kolmogorov's theorem requires some assumptions on $(\mathcal{X}, \mathcal{F})$ that are satisfied if $\mathcal{X}$ is a complete separable metric space and $\mathcal{F}$ are the Borel sets. However, in the present context, there is a result known as Tulcea's theorem that proves the existence of a P on $(\mathcal{X}^\infty, \mathcal{F}^\infty)$ for any choice of $(\mathcal{X}, \mathcal{F})$, exploiting the fact that the consistent family of finite dimensional distributions μ_k arise from well defined successive regular conditional probability distributions.

An important subclass is generated when the transition probability depends on the past history only through the current state. In other words

$$\pi_k(x_0, ..., x_{k-1}; \cdot) = \pi_{k-1,k}(x_{k-1}; \cdot)$$

In such a case the process is called a Markov Process with transition probabilities $\pi_{k-1,k}(\cdot, \cdot)$. An even smaller subclass arises when we demand that $\pi_{k-1,k}(\cdot, \cdot)$ be the same for different values of k. A single transition probability $\pi(x, A)$ and the initial distribution μ_0 determine the entire process i.e. the measure P on $(\mathcal{X}^\infty, \mathcal{F}^\infty)$. Such processes are called time-homogeneous Markov Processes or Markov Processes with stationary transition probabilities.

Chapman-Kolmogorov Equations: If we have the transition probabilities $\pi_{k,k+1}$ of transition from time k to $k+1$ of a Markov Chain it is possible to obtain directly the transition probabilities from time k to $k+l$ for any $l \geq 2$. We do it by induction on l. Define

$$\pi_{k,k+l+1}(x, A) = \int_{\mathcal{X}} \pi_{k,k+l}(x, dy)\pi_{k+l,k+l+1}(y, A) \tag{6.1}$$

or equivalently, in a more direct fashion

$$\pi_{k,k+l+1}(x, A) = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \pi_{k,k+1}(x, dy_{k+1}) \dots \pi_{k+l,k+l+1}(y_{k+l}, A)$$

Theorem 6.0.2. *The transition probabilities $\pi_{k,m}(\cdot,\cdot)$ satisfy the relations*

$$\pi_{k,n}(x,A) = \int_{\mathcal{X}} \pi_{k,m}(x,dy)\pi_{m,n}(y,A) \tag{6.2}$$

for any $k < m < n$ and for the Markov Process defined by the one step transition probabilities $\pi_{k,k+1}(\cdot,\cdot)$, for any $n > m$

$$P[x_n \in A|\Sigma_m] = \pi_{m,n}(x_m, A) \text{ a.e.}$$

where Σ_m is the σ-field of past history upto time m generated by the coordinates $x_0, x_1, ..., x_m$.

In Durrett [4], there is a more strict description for *transition probability*. Let $(S,\mathcal{S})$ be a measurable space. A function $p : S \times \mathcal{S} \to \mathbb{R}$ is said to be a **transition probability** if:

(i) For each $x \in S$, $A \to p(x,A)$ is a probability measure on $(S,\mathcal{S})$.

(ii) For each $A \in \mathcal{F}$, $x \to p(x,A)$ is a measurable function.

We say X_n is a Markov chain (w.r.t. $\mathcal{F}_n$) with transition probability p if

$$P(X_{n+1} \in B|\mathcal{F}_n) = p(X_n, B)$$

Given a transition probability p and an **initial distribution** μ on $(S,\mathcal{S})$, we can define a consistent set of finite dimensional distributions by

$$P(X_j \in B_j, 0 \le j \le n) = \int_{B_0} \mu(dx_0) \int_{b_1} p(x_0, dx_1) \cdots \int_{B_n} p(x_{n-1}, dx_n) \tag{6.3}$$

Now we can proceed to prove Theorem 6.0.2:

Proof. The identity is basically algebra. The multiple integral can be carried out by iteration in any order and after enough variables are integrated we get our identity. To prove that the conditional probabilities are given by the right formula we need to establish

$$P[\{x_n \in A\} \cap B] = \int_B \pi_{m,n}(x_m, A)dP$$

for all $B \in \Sigma_m$ and $A \in \mathcal{F}$. We write

$$\begin{aligned}
P[\{x_n \in A\} \cap B] &= \int_{\{x_n \in A\} \cap B} dP \\
&= \int \cdots \int_{\{x_n \in A\} \cap B} d\mu(x_0)\pi_{0,1}(x_0, dx_1) \ldots \pi_{m-1,m}(x_{m-1}, dx_m) \\
&\quad \pi_{m,m+1}(x_m, dx_{m-1}) \ldots \pi_{n-1,n}(x_{n-1}, dx_n) \\
&= \int \cdots \int_B d\mu(x_0)\pi_{0,1}(x_0, dx_1) \ldots \pi_{m-1,m}(x_{m-1}, dx_m) \\
&\quad \pi_{m,m+1}(x_m, dx_{m-1}) \ldots \pi_{n-1,n}(x_{n-1}, A) \\
&= \int \cdots \int_B d\mu(x_0)\pi_{0,1}(x_0, dx_1) \ldots \pi_{m-1,m}(x_{m-1}, dx_m)\pi_{m,n}(x_m, A) \\
&= \int_B \pi_{m,n}(x_m, A)dP
\end{aligned}$$

and we are done.

□

Remark 6.0.3. If the chain has stationary transition probabilities then the transition probabilities $\pi_{m,n}(x,dy)$ from time m to time n depend only on the difference $k = n-m$ and are given by what are usually called the k step transition probabilities. They are defined inductively by

$$\pi^{(k+1)}(x,A) = \int_{\mathcal{X}} \pi^{(k)}(x,dy)\pi(y,A)$$

and satisfy the Chapman-Kolmogorov equations

$$\pi^{(k+l)}(x,A)=\int_{\mathcal{X}}\pi^{(k)}(x,dy)\pi^{(l)}(x,A)=\int_{\S}^{'}pi^{(;)}(x,dy)\pi^{(k)}(y,A)$$

Suppose we have a probability measure P on the product space $X\times Y\times Z$ with the product σ-field. The Markov property in this context refers to equality

$$\mathbb{E}^P[g(z)|\Sigma_{x,y}]=E^P[g(z)|\Sigma_y]\text{ a.e. }P \tag{6.4}$$

for bounded measurable functions g on Z, where we have used $\Sigma_{x,y}$ to denote the σ-field generated by projection on to $X\times Y$ and Σ_y the corresponding σ-field by projection on to Y. The Markov property in the reverse direction is the similar condition for bounded measurable functions f on X.

$$\mathbb{E}^P[f(x)|\Sigma_{y,z}]=\mathbb{E}^P[f(x)|\Sigma_y]\text{ a.e. }P \tag{6.5}$$

They look different. But they are both equivalent to the symmetric condition

$$\mathbb{E}^P[f(x)g(z)|\Sigma_y]=\mathbb{E}^P[f(x)|\Sigma_y]\mathbb{E}^P[g(z)|\Sigma_y]\text{ a.e. }P \tag{6.6}$$

which says that given the present, the past and future are conditionally independent. In view of the symmetry it sufficient to prove the following:

Theorem 6.0.4. *For any P on $(X\times Y\times Z)$ the relations 6.4 and 6.6 are equivalent.*

Proof. Let us fix f and g. Let us denote the common value in 6.4 by $\hat{g}(y)$. Then

$$\begin{aligned}
\mathbb{E}^P[f(x)g(z)|\Sigma_y] &= \mathbb{E}^P[\mathbb{E}^P[f(x)g(z)|\Sigma_{x,y}]|\Sigma_y] && \text{a.e. } P\\
&= \mathbb{E}^P[f(x)\mathbb{E}^P[g(z)|\Sigma_{x,y}]|\Sigma_y] && \text{a.e. } P\\
&= \mathbb{E}^P[f(x)\hat{g}(y)|\Sigma_y] && \text{a.e. } P\\
&= \mathbb{E}^P[f(x)|\Sigma_y]\hat{g}(y) && \text{a.e. } P\\
&= \mathbb{E}^P[f(x)|\Sigma_y]\mathbb{E}^P[g(z)|\Sigma_y] && \text{a.e. } P
\end{aligned}$$

which is 6.6. Conversely, we assume 6.6 and denote by $\bar{g}(x,y)$ and $\hat{g}(y)$ the expressions on the left and right side of 6.4. Let $b(y)$ be a bounded measurable function on Y.

$$\begin{aligned}
\mathbb{E}^P[f(x)b(y)\bar{g}(x,y)] &= \mathbb{E}^P[f(x)b(y)g(z)]\\
&= \mathbb{E}^P[b(y)\mathbb{E}^P[f(x)g(z)|\Sigma_y]]\\
&= \mathbb{E}^P[b(y)\{\mathbb{E}^P[f(x)|\Sigma_y]\}\{\mathbb{E}^P[g(z)|\Sigma_y]\}]\\
&= \mathbb{E}^P[b(y)\{\mathbb{E}^P[f(x)|\Sigma_y]\}\hat{g}(y)]\\
&= \mathbb{E}^P[f(x)b(y)\hat{g}(y)]
\end{aligned}$$

Since f and b are arbitrary this implies that $\bar{g}(x,y)=\hat{g}(y)$ a.e. P.

□

Before we move on, it is worth our attention to review a few related theorems from Durrett [4].

Theorem 6.0.5. *X_n is a Markov chain (with respect to $\mathcal{F}_n=\sigma(X_0,X_1,...,X_n)$) with transition probability p.*

Proof. Let $A=\{X_0\in B_0,X_1\in B_1,...,X_n\in B_n\}$, $B_{n+1}=B$, and observe that using the definition of the integral, the definition of A, and the definition of P_μ

$$\begin{aligned}
\int_A 1_{(X_{n+1}\in B)}dP_\mu &= P_\mu(A,X_{n+1}\in B)\\
&= P_\mu(X_0\in B_0,X_1\in B_1,...,X_n\in B_n,X_{n+1}\in B)\\
&= \int_{B_0}\mu(dx_0)\int_{B_1}p(x_0,dx_1)\cdots\int_{B_n}p(x_{n-1},dx_n)p(x_n,B_{n+1})
\end{aligned}$$

We would like to assert that the last expression is

$$= \int_A p(X_n, B) dP_\mu$$

To do this, replace $p(X_n, B_n)$ by a general function $f(x_n)$. If f is an indicator function, the desired equality is true. Linearity implies that it is valid for simple functions, and the bounded convergence theorem implies that it is valid for bounded measurable f, for example, $f(x) = p(x, B_{n+1})$.

The collection of sets for which

$$\int_A 1_{(X_{n+1} \in B)} dP_\mu = \int_A p_n(X_n, B) dP_\mu$$

holds is a λ-system, and the collection for it has been proved is a π-system, and the collection for which it has been proved is a π-system, so it follows from the $\pi - \lambda$ theorem, Theorem 2.1.3, that the equality is true for all $A \in \mathcal{F}_n$. This shows that

$$P(X_{n+1} \in B | \mathcal{F}_n) = p(X_n, B)$$

and proves the desired result.

From this foundation, we have shown that given a sequence of transition probabilities and an initial distribution, we can construct a Markov chain. Conversely, we have

Theorem 6.0.6. *If X_n is a Markov chain with transition probabilities p and initial distribution μ, then the finite dimensional distributions are given by 6.3.*

Proof. Our first step is to show that if X_n has transition probability p, then for any bounded measurable f

$$\mathbb{E}(f(X_{n+1}) | \mathcal{F}_n) = \int p(X_n, dy) f(y) \tag{6.7}$$

Theorem 6.0.7. ***Monotone class theorem.*** *Let $\mathcal{A}$ be a π-system that contains Ω and let $\mathcal{H}$ be a collection of real-valued functions that satisfies:*

(i) If $A \in \mathcal{A}$, then $1_A \in \mathcal{H}$.

(ii) If $f, g \in \mathcal{H}$, then $f + g$, and $cf \in \mathcal{H}$ for any real number c.

(iii) If $f_n \in \mathcal{H}$ are nonnegative and increase to a bounded function f, then $f \in \mathcal{H}$.

Then $\mathcal{H}$ contains all bounded functions measurable with respect to $\sigma(\mathcal{A})$.

Proof. The assumption $\Omega \in \mathcal{A}$, (ii), and (iii) imply that $\mathcal{G} = \{A : 1_{A \in \mathcal{H}}$ is a λ-system, so by (i) and the $\pi - \lambda$ theorem, Theorem 2.1.3, $\mathcal{G} \supset \sigma(\mathcal{A})$. (ii) implies that $\mathcal{H}$ contains all simple functions, and (iii) implies that $\mathcal{H}$ contains all bounded measurable functions.

□

The desired conclusion is a consequence of the next result. Let $\mathcal{H}$ = the collection of bounded functions for which the identity holds.

□

6.1 Stopping Times and Renewal Times

One of the important notions in the analysis of Markov Chains is the idea of stopping times and renewal times. A function

$$\tau(\omega) : \Omega \longrightarrow \{n : n \geq 0\}$$

is a random variable defined on the set $\Omega = \chi^{\infty}$ such that for every $n \geq 0$ the set $\{\omega : \tau(\omega) = n\}$ (or equivalently for each $n \geq 0$ the set $\{\omega : \tau(\omega) \leq n\}$) is measurable with respect to the σ-field $\mathcal{F}_n$ generated by $X_j : 0 \leq j \leq n$. It is not necessary that $\tau(\omega < \infty$ for every ω. Such random variable τ are called stopping times. Examples of stopping times are, constant times $n \geq 0$, the first visit to a state x, or the second visit to a state x. The important thing is that in order to decide if $\tau \leq n$ i.e. to know if whatever is supposed to happen did happen before time n the chain need be observed only up to time n. Examples of τ that are not stopping times are easy to find. The last time a site is visited is not a stopping time nor is the first time such that at the next time one is in a state x. An important fact is that the Markov property extends to stopping times. Just as we have σ-fields $\mathcal{F}_n$ associated with constant times, we do have a σ-field $\mathcal{F}_\tau$ associated to any stopping time. This is the information we have when we observe the chain up to time τ. Formally

$$\mathcal{F}_\tau = \{A : A \in \mathcal{F}^{\infty} \text{ and } \cap\{\tau \leq n\} \in \mathcal{F}_n \text{ for each } n\}$$

One can check from the definition that τ is $\mathcal{F}_\tau$ measurable and so is X_τ on the set $\tau < \infty$. If τ is the time of first visit to y then τ is a stopping time and the event that the chain visits a state z before visiting y is $\mathcal{F}_\tau$ measurable.

Lemma 6.1.1. ***Strong Markov Property***. *At any stopping time τ the Markov property holds in the sense that the conditional distribution of $X_{\tau+1}, ..., X_{\tau+n}, ...$ conditioned on $\mathcal{F}_\tau$ is the same as the original chain starting from the state $x = X_\tau$ on the set $\tau < \infty$. In other words,*

$$P_x\{X_{\tau+1} \in A_1, ..., X_{\tau+n} \in A_n | \mathcal{F}_\tau\} = \int_{A_1} \cdots \int_{A_n} \pi(X_\tau, dx_1) \ldots \pi(x_{n-1}, dx_n)$$

a.e. on $\{\tau < \infty\}$.

Proof. Let $A \in \mathcal{F}_\tau$ be given with $A \subset \{\tau < \infty\}$. Then

$$\begin{aligned}
& P_x\{A \cap \{X_{\tau+1} \in A_1, ..., X_{\tau+n} \in A_n\}\} \\
= & \sum_k P_x\{A \cap \{\tau = k\} \cap \{X_{k+1} \in A_1, ..., X_{k+n} \in A_n\}\} \\
= & \sum_k \int_{A \cap \{\tau = k\}} \int_{A_1} \cdots \int_{A_n} \pi(X_k, dx_{k+1}) \ldots \pi(x_{k+n-1}, dx_{k+n}) dP_x \\
= & \int_A \int_{A_1} \cdots \int_{A_n} \pi(X_\tau, dx_1) \ldots \pi(x_{n-1}, dx_n) dPX
\end{aligned}$$

We have used the fact that if $A \in \mathcal{F}_\tau$ then $A \cap \{\tau = k\} \in \mathcal{F}_k$ for every $k \geq 0$.

□

Remark 6.1.2. If $X_\tau = y$ a.e. with respect to P_x on the set $\tau < \infty$, then at time τ, when it is finite, the process starts afresh with no memory of the past and will have conditionally the same probabilities in the future as P_y. At such times the process renews itself and these times are called renewal times.

6.2 Countable State Space

From the point of view of analysis a particularly simple situation is when the state space χ is a countable set. It can be taken as the integers $\{x : x \geq 1\}$. Many applications fall in this category and an understanding of what happens in this situation will tell us what to expect in general. The one step transition probability is a matrix $\pi(x, y)$ with nonnegative entries such that $\sum_y \pi(x, y) = 1$ for each x. Such matrices are called stochastic matrices. The n step transition matrix is just the n-th power of the matrix defined inductively by

$$\pi^{(n+1)}(x, y) = \sum_z \pi^{(n)}(x, z)\pi(z, y)$$

To be consistent one defines $\pi^{(0)}(x, y) = \delta_{x,y}$ which is 1 if $x = y$ and 0 otherwise. The problem is to analyze the behavior for large n of $\pi^{(n)}(x, y)$. A state x is said to communicate with a state y if $\pi^{(n)}(x, y) > 0$ for some $n \geq 1$. We will assume for simplicity that every state communicates with every other state. Such Markov Chains are called **irreducible**. Let us first limit ourselves to the study of irreducible chains. Given an irreducible Markov chain with transition probabilities $\pi(x, y)$ we define $f_n(x)$ as the probability of returning to x for the first time at the n-th step assuming that the chain starts from the state x. Using the convention that P_x refers to the measure on sequences for the chain starting from x and $\{X_j\}$ are the successive positions of the chain

$$\begin{aligned} f_n(x) &= P_x\{X_j \neq x \text{ for } 1 \leq j \leq n-1 \text{ and } X_n = x\} \\ &= \sum_{\substack{y_1 \neq x \dots \\ y_{n-1} \neq x}} \pi(x, y_1)\pi(y_1, y_2) \dots \pi \end{aligned}$$

Since $f_n(x)$ are probabilities of disjoint events $\sum_n f_n(x) \leq 1$. The state x is called **transient** if $\sum_n f_n(x) < 1$ and **recurrent** if $\sum_n f_n(x) = 1$. The recurrent case is divided into two situations. If we denote by $\tau_x = \inf\{n \geq 1 : X_n = x\}$, the time of first visit to x, then recurrence is $P_x\{\tau_x < \infty\} = 1$. A recurrent state x is called **Positive recurrent** if

$$\mathbb{E}^{P_x}\{\tau_x\} = \sum_{n \geq 1} n f_n(x) < \infty$$

and **null recurrent** if

$$\mathbb{E}^{P_x}\{\tau_x\} = \sum_{n \geq 1} n f_n(x) = \infty$$

Lemma 6.2.1. *If for a (not necessarily irreducible) chain starting from x, the probability of ever visiting y is positive, then so is the probability of visiting y before returning to x.*

Proof. Assume that for the chain starting from x the probability of visiting y before returning to x is zero. But when it returns to x it starts afresh and so will not visit y until it returns again. This reasoning can be repeated and so the chain will have to visit x infinitely often before visiting y. But this will use up all the time and so it cannot visit y at all.

□

Lemma 6.2.2. *For an irreducible chain all states x are of the same type.*

Proof. Let x be recurrent and y be given. Since the chain is irreducible, for some k, $\pi^{(k)}(x, y) > 0$. By the previous lemma, for the chain starting from x, there is a positive probability of visiting y before returning to x. But this will use up all the time and so it cannot visit y at all.

□

Lemma 6.2.3. *For an irreducible chain all states x are of the same type.*

Proof. Let x be recurrent and y be given. Since the chain is irreducible, for some k, $\pi^{(k)}(x,y) > 0$. By the previous lemma, for the chain starting from x, there is a positive probability of visiting y before returning to x. After each successive return to x, the chain starts afresh and there is a fixed positive probability of visiting y before the next return to x. Since there are infinitely many returns to x, y will be visited infinitely many times as well. Or y is also a recurrent state.

We now prove that if x is positive recurrent, then so is y. We saw already that the probability $p = P_x\{\tau_y < \tau_x\}$ of visiting y before returning to x is positive. Clearly,

$$\mathbb{E}^{P_x}\{\tau_x\} \geq P_x\{\tau_y < \tau_x\}\mathbb{E}^{P_y}\{\tau_x\} \text{ and therefore } \mathbb{E}^{P_y}\{\tau_x\} \leq \frac{1}{p}\mathbb{E}^{P_x}\{\tau_x\} < \infty.$$

On the other hand, we can write

$$\begin{aligned}
\mathbb{E}^{P_x}\{\tau_y\} &\leq \int_{\tau_y<\tau_x} \tau_x dP_x + \int_{\tau_x<\tau_y} \tau_y dP_x \\
&= \int_{\tau_y<\tau_x} \tau_x dP_x + \int_{\tau_x<\tau_y} \{\tau_x + \mathbb{E}^{P_x}\{\tau_y\}\} dP_x \\
&= \int_{\tau_y<\tau_x} \tau_x dP_x + \int_{\tau_x<\tau_y} \tau_x dP_x + (1-p)\mathbb{E}^{P_x}\{\tau_y\} \\
&= \int \tau_x dP_x + (1-p)\mathbb{E}^{P_x}\{\tau_y\}
\end{aligned}$$

by the renewal property at the stopping time τ_x. Therefore

$$\mathbb{E}^{P_x}\{\tau_y\} \leq \frac{1}{p}\mathbb{E}^{P_x}\{\tau_x\}.$$

We also have

$$\mathbb{E}^{P_y}\{\tau_y\} \leq \mathbb{E}^{P_y}\{\tau_x\} + \mathbb{E}^{P_x}\{\tau_y\} \leq \frac{2}{p}\mathbb{E}^{P_x}\{\tau_x\}$$

proving that y is positive recurrent.

□

We have the following theorem regarding transience.

Theorem 6.2.4. *An irreducible chain is transient if and only if*

$$G(x,y) = \sum_{n=0}^{\infty} \pi^{(n)}(x,y) < \infty \text{ for all } x, y$$

Moreover, for any two states x and y,

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

where $f(x,y) = P_x\{\tau_y < \infty\}$.

Proof. Each time the chain returns to x there is a probability $1 - f(x,x)$ of never returning. The number of returns then has the geometric distribution

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

and the expected number of returns is given by

$$\sum_{k=1}^{\infty} \pi^{(k)}(x,x) = \frac{f(x,x)}{1-f(x,x)}.$$

The left-hand side comes from the calculation

$$\mathbb{E}^{P_x} \sum_{k=1}^{\infty} \chi_{\{X\}}(X_k) = \sum_{k=1}^{\infty} \pi^{(k)}(x,x),$$

and the right-hand side from the calculation of the mean of a geometric distribution. Since we count the visit at time 0 as a visit to x, we add 1 to both sides to get our formula. If we want to calculate the expected number of visits to y when we start from x, first we have to get to y and the probability of that is $f(x,y)$. Then by the renewal property it is exactly the same as the expected number of visits to y starting from y, including the visit at time 0 and that equals $G(y,y)$.

□

7 INTEGRABLE PROBABILITY

Go back to Table of Contents. Please click TOC

7.1 q-TASEP

We introduce q-TASEP in this section. One can study moments of q-TASEP using markov dualities. It is an interacting particle system throughout continuous time which was introduced by Borodin and Corwin a few years ago [1] [2] [3]. It is a simple model for a traffic. Consider particles, or cars, sitting on a one-dimensional lattice. The rule is that each particle jump to the right by one with exponential waiting time with rate to be $1-q^{\text{gap}}$ where $q \in (0,1)$ fixed. Later we can take various limits and we can discuss when $q<0$. The gap is defined as the empty spaces between you and the next particles. There can only be one particle per space and this is forced the the definition of $1-q^{\text{gap}}$ since the rate will go to 0 if the gap goes to 0. If gap goes to infinity, the rate would go to 1. Henceforth, q monitors the safety window between you and the car in front of you. Let us assume a particle $X_1(t)$, $X_2(t)$ (left of $X_1(t)$), $X_3(t)$, ..., etc. and there are total X^N particles to be

$$\mathbf{X}^N = \{\vec{x} = (\infty = x_0 > x_1 > \cdots > x_N), x_i \in \mathbb{Z}\} \text{ while } X_+ = \infty$$

This gives us an intuitive way of describing the system. Now we can pursue the same notion from the moment generator. Given a function $f : \mathbf{X}^N(\text{ state space }) \to \mathbb{C}$

$$(L^{q-\text{TASEP}} f)(\vec{x}) \sum_{i=1}^{N} \underbrace{(1-q^{X_{i-1}-X_i-1})}_{\text{rate of change}} \underbrace{(\underbrace{f(\vec{x}_i^+)}_{\text{old value}} - f(\vec{x}))}_{\text{effect of change}}$$

$$\text{while } \vec{x}_i^+ = (x_0, x_1, ..., x_i+1, ..., x_n)$$

Given function of the Markov process and one evolves the Markov process instant in time, the generator of a Markov process describes how this function would change in expectation. What would happen in instant of time? Each of the time these jumps can happen and they are independent with exponential waiting time. That is, we want to take into the possibility that particle 1 through N moves. Then we need to consider the effect of the change. If each of the particle responses to the move from X_i to X_i+1. The second parenthesis describes the state after the jump has occurred.

If you have a continuous time Markov process on a countable space $\mathbf{X}$ (i.e., a state space), you can define a semi-group $(S_t f)(x) = \mathbb{E}^x[f(x(t))]$ which notates $(S_t f)(x) = \mathbb{E}[f(x(t))|X(0) = x]$. It tells us how expectations of functions evolve according to Markov process. The generator of Markov process is then defined as

$$L := \lim_{t\to\infty} \frac{S_t - \text{Id}}{t}, e^{tL} = S_t$$

In other words, we can look at the amount of small amount of time S_t on the expectation of a function. The change of that function we then can assume is proportionate to t. Under some weak hypothesis, one can recover the semi-group by taking exponentials on the generator. Some related consequences that worth our attention is the following.

$$\frac{d}{dt} S_t = S_t L = L S_t (\text{commute})$$

or written more precisely

$$\frac{d}{dt}\mathbb{E}^x[f(x(t))] = \mathbb{E}^x[(Lf)(x(t))] = L\mathbb{E}^x[f(x(t))]$$

Example 7.1.1. What if we consider $f_n(\vec{x}) = q^{X_n+n}$ while $1 \leq n < N$. Then question is $\frac{d}{dt}\mathbb{E}^{\vec{x}}[f_n(\vec{x}(t))] =?$

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}^{\vec{x}}[f_n(\vec{x}(t))] &= \mathbb{E}^{\vec{x}}[(L^{q-\text{TASEP}}f_n)(\vec{x}(t))], \text{ differentiation gives the} \\
&\quad \text{generator of the Markov process} \\
&= \mathbb{E}^{\vec{x}}[(1-q^{X_{n-1}-X_n-1})(q^{X_n+n+1}-q^{X_n+1})], \text{ rate multipled by effect} \\
&\quad \text{we can factor out the term } q^{X_n} \\
&= (1-q) \bigtriangledown \mathbb{E}^{\vec{x}}[f_n(\vec{x}(t))], \text{ results of the facorization} \\
&\quad \text{which cancels out the same term in the first parenthesis}
\end{aligned}$$

while

$$(\bigtriangledown f)(n) = f(n-1) - f(n), \text{ a discrete derivative}$$

$$\text{that acts on a function of variable } n$$

This is a very interesting description. Say that one is interested in the computation of the expectation of f_n, one can find that this time-related derivative can be described by $f(n-1)$ and also $f(n)$. This gives us a triangular system of ODEs for the target expectation that we were interested in. If one has a triangular system of ODEs with initial data, then one can just diagonalize the triangular matrix and one can solve this equation in a simple manner.

However, if one just choose an arbitrary function for state space, does it still solve the expectation in a close form? The answer is no. That is, if we have $f_n(\vec{x}) = \tilde{q}^{X_n+n}$, then one we fail to continue at the factorization step since

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}^{\vec{x}}[f_n(\vec{x}(t))] &= \mathbb{E}^{\vec{x}}[(L^{q-\text{TASEP}}f_n)(\vec{x}(t))], \text{ differentiation gives the} \\
&\quad \text{generator of the Markov process} \\
&= \mathbb{E}^{\vec{x}}[(1-q^{X_{n-1}-X_n-1})(q^{X_n+n+1}-q^{X_n+1})], \text{ rate multiplied by effect} \\
&\quad \text{we can not factor out the term since } q^{X_n} \text{ and } \tilde{q}^{X_n} \text{ cannot cancel out}
\end{aligned}$$

That is, we can define a step to be $X_i(0) = -i$, while $i = 1$. In graph, imagine a one-dimensional lattice with an arbitrary point to be 0. Then we would have an infinite traffic jam. In that case, $f_n(x(0)) = \mathbf{1}_{n \geq 1}$ which is simply the indicator function.

Definition 7.1.2. Suppose there is $X(t) \in \mathbf{X}$ and independent $y(t) \in \mathbf{Y}$ and a function $H : \mathbf{X} \times \mathbf{Y} \to \mathbb{C}$. Then x, y are dual w.r.t. H if

$$L^{\mathbf{X}}H(x,y) = L^{\mathbf{Y}}H(x,y), \forall x, y$$

This duality definition, Definition 7.1.2 implies that $\mathbb{E}^{\mathbf{X}}[H(x(t),y)] = \mathbb{E}^{\mathbf{Y}}[H(x,y(t))]$. In other words, the function H does not care which coordinates one goes first. The results hold the same in expectation.

We have the following implication.

$$\begin{aligned}
\tfrac{d}{dt}\mathbb{E}^{\mathbf{X}}[H(x(t),y)] &= L^{\mathbf{Y}}\mathbb{E}^{\mathbf{Y}}[H(x,y(t))] \\
&= L^{\mathbf{Y}}\mathbb{E}^{\mathbf{Y}}[H(x(t),y)], \text{ change to } x(t)
\end{aligned}$$

$\mathbf{Y}$ can be a space parameter, such a kin to the subscript N in the Example 7.1.1. Then if we have a duality, a dual system, then the time derivative of such an expectation is actually equal to the generator $L^{\mathbf{Y}}$ of the dual system applied to the same expectation. In the Example 7.1.1, $\bigtriangledown$ can be thought of as the generator $L^{\mathbf{Y}}$.

7.2 q-Boson

The dual system to q-TASEP is called q-Boson particle system for $q \in (0,1)$. Consider the graph of bins 0 to N. Of each bin, there can be nonnegative objects. In continuous time, one can move one particle from one bin to another (say the one left to it). Then the rate is given $1 - q^{Y_i}$ while Y_i is the height of the bin, i.e. the number of objects in that bin. Suppose for now that there is nothing coming in from right and nothing going out from left (i.e. a closed system).

The state space, $\mathbf{Y}^N = \{\vec{y} = (y_0, ..., y_n), y_i \in \mathbb{Z} \geq 0\}$. Then

$$(L^{q-\text{Boson}} h)(\vec{y}) := \sum_{i=1}^{N} (1 - q^{Y_i})(h(\vec{y}^{\,i,i-1}) - h(\vec{y}))$$

$$\text{while } y^{i,i-1} = (..., y_{i-1} + 1, y_i - 1, ...)$$

These definitions lead to the following theorem:

Theorem 7.2.1. ***The Duality***. *q-TASEP, $\vec{x}(t)$, q-Boson process, $\vec{y}(t)$, are dual with respect to*

$$H(\vec{x}, \vec{y}) = \prod_{i=0}^{N} q^{(x_i+1)y_i}.$$

Recall $x_0 = +\infty$, so such product $\prod_{i=0}^{N} q^{(x_i+1)y_i} = 0$ if $y_0 > 0$ since then we are dealing with $q^{+\infty}$.

If one consider state space with one particle, then the duality would be between q-TASEP and q-Boson with one particle. That means every term in the product $\prod q^{(x_i=1)y_i}$ would be equal to q^0 except for the location that one particle is. In that case,

$$H(\vec{x}, \vec{y}) = \prod_{i=0}^{N} q^{(x_i+1)y_i} \to \prod q^{x_n+n}$$

and the generator of the q-Boson particle would just be $(1-q) \bigtriangledown \mathbb{E}^{\vec{x}}[f_n(\vec{x}(t))]$. This is the generalized result of Example 7.1.1. We can prove this result.

Proof. $L^{q-\text{TASEP}} H(x,y) = L^{q-\text{Boson}} H(x,y)$, which is the form of Theorem 7.2.1. Then

$$\begin{aligned}
L^{q-\text{TASEP}} H(\underbrace{x}_{\uparrow}, y) &= L^{q-\text{Boson}} H(x, \underbrace{y}_{\uparrow}), \text{ same form of the generator apply on} \\
&\quad \text{different coordinates, notated with } \uparrow \\
&\Rightarrow \mathbb{E}^{\vec{x}}[\textstyle\prod_{i=0}^{N} q^{(x_i(t)+i)y_i}], \text{ by choosing } x_i \text{ and } y_i \\
&\quad \text{will allow us to identify } q-\text{TASEP} \\
&= L^{q-\text{Boson}} \\
&\quad \text{so that there exists a triangular inequality}
\end{aligned}$$

then we will get the implication as discussed above. In words, if we want to compute the derivative, all we need is to solve the q-Boson generator with specific initial data.

□

Proposition 7.2.2. *Fix $\vec{x} \in \mathbf{X}^N$. If $h : \mathbb{R}_h \times \mathbf{Y}^N \to \mathbb{R}$ solves that*
(1) $\forall y \in \mathbf{Y}^N$ with $t \geq 0$, then $\frac{d}{dt}h(t,y) = L^{q-Boson}h(t,\vec{y})$
(2) $\forall y$, $h(0,\vec{y}) = \mathbb{E}^{\vec{x}}[H(x(0),y)]$
Then $h(t,\vec{y}) = \mathbb{E}^{\vec{x}}[H(x(t),y)]$

Restrict q-Boson with k particles and index $\vec{y}$ by ordered locations of particles. In other words, we can imagine a state space with $N = 3$, and there is one particle for at $n = 1$, three particles at $n = 2$, and one particle at $n = 3$; and let this be associated with $\vec{n} = (n_1 \geq n_2 \geq \cdots \geq n_k)$, and it can be (3,2,2,2,1). In terms of n variables, we have

$$(L^{q-\text{Boson}}h)(\vec{n}) = \sum_{\text{clusters of } \vec{n}} (1-q^{c_i})[\bigtriangledown]_{c_i,1+c_i}h$$

while $n = (3,2,2,2,1)$ (move last particle first, say the 3rd "2") and $c = (1,3,1)$ and consider $([\bigtriangledown]_i f)(n_1,...,n_k) =$ apply $\bigtriangledown$ to i^{th} coordinate. Reserve a state $W_{20}^k = \{\vec{n} = (n_1 \geq \ldots n_k \geq 0) : n_i \in \mathbb{Z}\}$, which is a chamber of ordered elements that are non-negative. The proposition here gives us a way to solve problems with this type, such as Theorem 7.2.1. The generator, which is the R.H.S. of $(L^{q-\text{Boson}}h)(\vec{n}) = \sum_{\text{clusters of } \vec{n}}(1-q^{c_i})[\bigtriangledown]_{c_i,1+c_i}h$, is not separable and not homogeneous. In other words, the term depends on the state of $\vec{n}$. This will be handy when one is trying to solve this. Then the idea is that we treat this thing as almost constant coefficient-inseparable. That is, if one imagines the state space of the q-Boson system, say $(L^{q-\text{Boson}}h)(\vec{n})$ and imagine having each of the cluster with size no bigger than one (a diffuse system of particles with each $n_i \geq \ldots n_k \geq 0$), then the generator of the q-Boson system is very simple. Away from the boundary, the generator is just $\sum_{i=1}^N (1-q)[\bigtriangledown]h$ (the generator away from the boundary) and then we can look for the solution to the system that is correct and away from the boundary but also try to impose certain boundary conditions on solutions which will affect the clusters near the boundaries.

Proposition 7.2.3. *Suppose there is function $\mu : \mathbb{R}_a \mathbb{R}_{\geq 0}^k \to \mathbb{R}$ that solves*
(1) $\forall \vec{n} \in W_{\geq 0}^k$ (which can be replaced by $\mathbb{R}^k$) such that

$$\frac{d}{dt}\mu(t,\vec{n}) = \sum_{i=1}^{k}(1-q) \bigtriangledown \mu(t,\vec{n})$$

(2) $\forall t \forall \vec{n}, n_i = n_{i+1} 1 \leq ik-1$, then there is the following boundary condition

$$(\bigtriangledown_i - q\bigtriangledown_{i+1})\mu\Big|_{\vec{n}} = 0$$

If one is looking for solution to translate the convolution equation and say the solution has a certain relation near the boundary, then
(3) $\forall n \in W_{\geq 0}, \mu(0,\vec{n}) = h_0(\vec{n}) = h_0(\vec{y}(\vec{n}))$, then

$$\forall \vec{n} = W_{\geq 0}^*, t > 0, \mu(t,\vec{n}) = h(t,\vec{n}) = h(t,\vec{y}(\vec{n}))$$

This is considered a very subtle thing since this does not appear in other equations. Imagine one had a solution that satisfies (1) and (3).

Proof. Consider a cluster $n_1 \geq n_2 \geq \ldots$. That means we can take $\frac{d}{dt}\mu(t,\vec{n}) = \sum(1-q) \bigtriangledown \mu(t,\vec{n})$ and we can progressively replace that by using the boundary condition:$(1-q)(\bigtriangledown_1 + \cdots + \bigtriangledown_c)\mu$ and $\bigtriangledown_1 \mu = q\bigtriangledown_2 = \cdots = q^{c+1} \bigtriangledown_k \mu$, so we can write $(1-q)(q^{c+1} \bigtriangledown_c + \cdots + \bigtriangledown_c)\mu$. Thus, we replace the whole summation by

$$\underbrace{(1-q)(q^{c+1}+q^{c+2}+\cdots+1)}_{(1-q^c)\bigtriangledown_c}\bigtriangledown_c\mu$$

which is the first time in the q-Boson generator given in $(L^{q-\text{Boson}}h)(\vec{y}) := \sum_{i=1}^{N}(1-q^{Y_i})(h(\vec{y}^{\,i,i-1}) - h(\vec{y}))$.

We can write down the solutions by the help of the above equation and we can check (1)-(3). We are interested in

$$h_0(\vec{n}) = \prod_{i=1}^{k}\mathbf{1}_{n_i\geq 1}$$

if we solve Proposition 7.2.3 with h_0

$$\mu(t,\vec{n}) = \mathbb{E}^{\text{Step } q-\text{TASEP}}[\prod_{i=1}^{k}q^{X_n+n_i}]$$

Theorem 7.2.4. *For $n_i \geq n_k + \dots n_k + 0$,*

$$\mu(t,\vec{n}) = \mathbb{E}^{Step\ q-TASEP}[\prod_{i=1}^{k}q^{X_n+n_i}] = \mu(t,\vec{n}$$

$$= \frac{(-1)^k q^{\frac{k(k-1)}{2}}}{(2\pi_i)^k}\underbrace{\int\cdots\int}_{k}\prod_{1\leq A\leq B\leq N}\frac{Z_A-Z_B}{Z_A-qZ_B}\prod\frac{e^{(q-1)tZ_j}}{(1-z_j)^{n_j}}\frac{dz_j}{z_j}$$

Sketch. We have

$$\mu_z(t,n) = \frac{e^{(n+1)tz}}{(1-z)^n}$$

and we write $\frac{d}{dt}\mu_t(+n) = (1-q)\bigtriangledown\mu_z(t,n)$, then $\mu_{z_1}(t,n)\dots\mu_{z_n}(t,n_k)$. If take derivative, we would have arrived the derivation side of the product.

Consider $n_1 = n_2$, then $\bigtriangledown_1 q\bigtriangledown_2$ apply to integral, it will brings out $(z_1 - qZ_2)$ into integral. This is like taking $\int dz_1\int dz_2(z_1-z_2)\underbrace{G(z_1,z_2)}_{\text{sym. function in } z_1,z_2}$. Then this can be checked to be 0, that is,

$$\int dz_1\int dz_2(z_1-z_2)\underbrace{G(z_1,z_2)}_{\text{sym. function in } z_1,z_2} = 0$$

In principle, these formulas completely characterized the joint distribution the vector $\{q^{x_n(t)+n}\}_{n\geq 1}$. The next challenge is to use the knowledge of the moments to compute the asymptotic. Then we can understand a larger system, a long-term behavior.

As a summary, we can derive the following results. Denote

$$(a_n q)_\infty = \prod_{i=0}^{\infty}(1-q^k n)$$

and

$$e_q(z) = \frac{1}{(1-q)(z,q)_\infty}$$

also given e_q-Laplace which then we know the distribution, then we have the following:

Theorem 7.2.5. *For $S \in \mathbb{C} \setminus \{\mathbb{R}\}, q \in (0,1), t \geq 0, n \geq 1$, then we have*

$$\mathbb{E}^{q-TASEP}[e_q(\zeta q^{X_n(t)+n})] = det(1+K_\zeta)_{L^2(small\ contour\ of\ 1)}$$

and then

$$K_\zeta(w,w') = \int_{i\mathbb{R}+1/2} \frac{\prod}{\sin(-\pi\zeta)}(1-(1-q)\zeta)^\zeta \frac{e^{tw(q^\zeta-1)}(\frac{(q-q^\zeta w)_n}{(q_i-w)_\infty})}{q^\zeta w - w'} \frac{ds}{d\pi_i'}$$

The kernel only involves n and t. As they grow, the integral formula would grow depends on the largest value of the integrand. Let us say we have some contour

$$\det(1+K)_{L^2(Y)} = 1 + \sum_{l=1}^{\infty} \frac{1}{l!} \int \frac{dw}{k\pi} \cdots \int \frac{dw_l}{2\pi} \det(K(w_i,w_j))_{i,j=1}^l$$

which is the generalization of Von Koch formula.

Corollary 7.2.6. *For $\kappa > 0$, $\exists c_\kappa t_\kappa$ such that*

$$\mathbb{P}(\frac{X_{t_\kappa(t)} - tc_\kappa}{t^{1/3} d\kappa} \leq r) = F_{GUE}(r)$$

One can look deep into the system with infinitely sequence on the left of the origin and let us say one wants to keep track how many objects he has passed the toll booth at the origin. This is equivalent as keeping track of the location of various particles far into the process. This is saying that one can prove L.L.N. involving the where the toll booth is moving with fluctuations with order $t^{1/3}$, which is a surprising fact from conventional Gaussian statistics.

The study of this field can be used to study other processes. For instance, one can take $u \to 1$, then this will evolve to the O'conor York semi-distribution process.

As summary, we used duality to identify certain expectations of q-TASEP which solved to close evolution equations. Then we reduced the evolution equations into nice forms in which we can write down moment formulas which allows us to recover distribution information.

Index

References

[1] Borodin, Alexei, Ivan, Corwin, and Sasamoto, Tomohiro (2013), 'Spectral theory for the q-Boson particle system", *The Annals of Probability*, `https://arxiv.org/pdf/1308.3475.pdf`.

[2] Borodin, Alexei, Ivan, Corwin, and Sasamoto, Tomohiro (2014), "From Duality to Determinants for q-TASEP and ASEP", *The Annals of Probability*, 42(6), 2314-2382. `https://arxiv.org/pdf/1207.5035.pdf`.

[3] Alexei Borodin, Ivan Corwin, Leonid Petrov, Tomohiro Sasamoto (2016), "Spectral theory for interacting particle systems solvable by coordinate Bethe ansatz", `https://arxiv.org/pdf/1407.8534.pdf`.

[4] Durrett, Rick, "Probability: Theory and Examples".

[5] Etemadi, N. (1981), "An elementary proof of the strong law of large numbers. *Z. Warsch. verw. Gebiete.*, 55, 119-122.

[6] Feller W. (1946), "A Limit Theorem for Random Variables with Infinite Moments", *Amer. J. Math*, 68, 257-262.

[7] Feller W. (1968), "An Introduction to Probability Theory and its Applications", *Vol. I*, third edition, John Wiley and Sons, New York.

[8] Gelman, "Bayesian Data Analysis", 3rd Edition. CRC Press, 2014.

[9] Gnedenko, B.V., and Kolmogorov, A.V. (1954), "Limit distributions for sums of independent random variables",Addison-Weslesy, Reading, MA.

[10] Hardy, G.H., and E.M. Wright (1959), "An Introduction to the Theory of numbers," 4th edition. *Oxford University Press*, London.

[11] Heyde, C.C. (1967), "On the influence of moments on the rate of convergence to the normal distribution", *Z. Warsch. verw. Gebiete*. 8, 12-18.

[12] Hodges, J.L. Jr., and L. Le Cam (1960), "The Poisson approximation to the binomial distribution", *Ann. Math. Statist.* 31, 737-740.

[13] Riesz, Frigyes (1907), "Sur les systèmes orthogonaux de fonctions", *Comptes rendus de l'Académie des sciences*, 144, 615 - 619.

[14] Stein C. (1987), "Appriximate computation of expectations", *IMS Lecture Notes*, Vol. 7.

[15] Stoyanov, J. (1987), "Counterexamples in probability", *John Wiley and Sons*, New York.

[16] Varadhan, "Probability Theory".