Probability Theory

Albert Zhang

August 23, 2021

Contents

1 Measure Theory **3**
1.1 The Problem of Measure . . . 3
1.2 Probability Spaces . . . 4
1.2.1 Constructing the Uniform Measure . . . 5
1.2.2 Caratheodory's Extension Theorem* . . . 8
1.3 Random Variables & Distributions . . . 10
1.4 Integration: Construction . . . 12
1.5 Integration: Tools . . . 15
1.5.1 Inequalities . . . 15
1.5.2 Convergence Theorems . . . 16
1.6 Expectation . . . 19
1.6.1 Inequalities . . . 19
1.6.2 Convergence Theorems . . . 19
1.6.3 Change of Measure . . . 20
1.7 Product Measures . . . 22

2 Laws of Large Numbers **25**
2.1 Independence . . . 25
2.1.1 Infinite Sequences of Random Variables . . . 27
2.2 Weak Laws of Large Numbers . . . 29
2.2.1 Truncation . . . 30
2.3 Strong Law of Large Numbers . . . 34
2.3.1 Borel-Cantelli Lemmas . . . 34
2.3.2 Subsequence Method . . . 38
2.3.3 Convergence of Random Series . . . 40
2.3.4 An Application of the Strong Law . . . 43

3 Central Limit Theorems **45**
3.1 Weak Convergence . . . 45
3.2 Characteristic Functions . . . 51
3.2.1 Moments & Derivatives . . . 51
3.2.2 Invertibility . . . 54
3.3 Central Limit Theorems . . . 57
3.3.1 Lindeberg's CLT . . . 59
3.3.2 Higher Dimensions . . . 61
3.4 Poisson Limit Theorem . . . 63

4 Martingales **64**
4.1 Radon-Nikodym Derivative . . . 64
4.2 Conditional Expectation . . . 65
4.2.1 Geometry of Conditional Expectations . . . 67
4.2.2 Regular Conditional Probabilities* . . . 68
4.3 Martingales . . . 69
4.3.1 Martingale Convergence . . . 72
4.3.2 Lp Inequalities . . . 74
4.3.3 L1 Theory & Uniformly Integrable Martingales . . . 76
4.3.4 Optional Stopping Theorems . . . 78
4.3.5 Reverse Martingales . . . 80
4.4 Applications to Random Walks . . . 84

5 Markov Processes **88**
5.1 Random Walks . . . 88
5.1.1 Combinatorics of Random Walks . . . 88
5.1.2 Recurrence & Transience of Random Walks . . . 91
5.2 Construction & Definitions . . . 95
5.2.1 Markov Kernels . . . 95

6 Brownian Motion **96**
6.1 Definition & Construction . . . 96

7 Concentration of Measure **97**
7.1 The Moment Method . . . 97
7.2 The Truncation Method . . . 99

1 Measure Theory

To develop probability more rigorously, measure theory is the usual choice for foundations. We start with basic measure theory constructions, and then move on to develop concepts such as random variables and their expectation in this context.

1.1 The Problem of Measure

We need to have some notion of measure that is well-defined and satisfies some desired "axioms" of how volume should behave. For example, perhaps we would like to have translation invariance:

$$\mu(E) = \mu(E + x).$$

Another desirable axiom could be disjoint countable additivity. That is, if we have disjoint sets $(E_n)_{n=1}^{\infty}$, with $E = \cup_n E_n$, then

$$\mu(E) = \sum_{n=1+}^{\infty} \mu(E_n).$$

But it turns out that there exists sets for which a measure would be "hard" to define given these axioms. We go over the construction of such a set known as the Vitali set. In particular, consider the base set $I = [-1, 2]$. Define an equivalence relation on I given by

$$x \sim y \iff x - y \in \mathbb{Q}.$$

Consider the restriction of the classes in $[0, 1]$. Note that the size of each class is countable. For each equivalence class B, pick $x_B \in B \cap [0, 1]$ (check that there always exists an x to be picked, note that this has to do with axiom of choice). Now consider

$$E = \{x_B : B \text{ is an equivalence class}\}.$$

Can we determine the size of E? To answer this, we first represent $[-1, 2]$ in terms of E. Note that

$$[0, 1] \subseteq \bigcup_{q \in [-1,1] \cap \mathbb{Q}} (E + q) \subseteq [-1, 2]$$

Since the $(E + q)$ disjoint for each $q \in [-1, 1] \cap \mathbb{Q}$, then we'd have

$$\infty \cdot \mu(E) = \sum_{q \in [-1,1] \cap \mathbb{Q}} \mu(E + q) = \mu\left(\bigcup_{q \in [-1,1] \cap \mathbb{Q}} (E + q) \right) \leq \mu([-1, 2]) < \infty.$$

This implies $\mu(E) = 0$, but then

$$0 = \sum_{q \in [-1,1] \cap \mathbb{Q}} \mu(E + q) = \mu\left(\bigcup_{q \in [-1,1] \cap \mathbb{Q}} (E + q) \right) \geq \mu([0, 1]) > 0,$$

clearly a contradiction. It follows that we cannot define a measure on every set that satisfies both translational invariance and disjoint countable additivity as axioms!

1.2 Probability Spaces

Recall from measure theory that a *measure space* is a triple

$$(\Omega, \mathcal{F}, \mu),$$

where Ω is the base set, $\mathcal{F}$ is the σ-algebra, and $\mu : \mathcal{F} \to [0, \infty]$ is our measure. A *probability measure* $\mu = \mathbb{P}$ is just the special case where $\mathbb{P}(\Omega) = 1$. The σ-algebra $\mathcal{F}$ by definition is a collection of subsets of Ω that satisfies

(i) $\Omega \in \mathcal{F}$.
(ii) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
(iii) If $(A_n) \subset \mathcal{F}$ is a countable sequence of sets, then $\cup_n A_n \in \mathcal{F}$.

Note that (ii) and (iii) imply (i).

Example 1.1. Verify that each of the following are valid σ-algebras.

- $\mathcal{F} = \{\varnothing, \Omega\}$
- $\mathcal{F} = 2^\Omega$
- $\mathcal{F} = \{A \subset \Omega : A \text{ is countable or co-countable}\}$

For the definition of a measure, we will only assume the axioms of nonnegativity and countable disjoint additivity. In particular, $\mu : \mathcal{F} \to [0, \infty]$ is a measure if it satisfies

(i) $\mu(A) \geq \mu(\varnothing) = 0$ for every $A \in \mathcal{F}$.
(ii) If A_n is a countable sequence of disjoint sets and $A = \sqcup_n A_n$, then

$$\mu(A) = \sum_n \mu(A_n).$$

Note: here and henceforth we may use $\sqcup_n A_n$ to denote disjoint union.

Example 1.2. Let $\Omega = \{1, 2, \ldots, n\}$ and $\mathcal{F} = 2^\Omega$. Then all possible ways to define a measure on $(\Omega, \mathcal{F})$ can be obtained by assigning a measure to each singleton. That is,

$$\mu(\{k\}) = p_k$$

for $k \in \Omega$.

Surprisingly, many familiar properties follow from this set of axioms alone.

Theorem 1.3 (Properties of measure)
Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.

(i) *monotonicity.* If $A \subset B$, then $\mu(A) \leq \mu(B)$.
(ii) *subadditivity.* If $A = \cup_n A_n$, then $\mu(A) \leq \sum_n \mu(A_n)$.
(iii) *continuity from below.* If $A_n \uparrow A$, then $\mu(A_n) \uparrow \mu(A)$.
(iv) *continuity from above.* If $A_n \downarrow A$ and $\mu(A_k) < \infty$ for some finite k, then $\mu(A_n) \downarrow \mu(A)$.

Proof. Monotonicity. Since A and $B \setminus A$ are disjoint, we have

$$\mu(A) \leq \mu(A) + \mu(B \setminus A) = \mu(B).$$

Subadditivity. Disjointify the sets as follows:

$$\begin{aligned} B_1 &= A_1 \\ B_2 &= A_2 \setminus A_1 \\ &\vdots \\ B_n &= A_n \setminus (\cup_{k=1}^{n-1} A_k) \end{aligned}$$

Then $A = \cup_n B_n$, so we have

$$\mu(A) = \mu(\cup_n B_n) = \sum_{n=1}^{\infty} \mu(B_n) \leq \sum_{n=1}^{\infty} \mu(A_n).$$

Continuity from below. Disjointify the set using (B_n) as from before. We have $A = \cup_n B_n$, and so countable additivity gives us

$$\mu(A) = \sum_{n=1}^{\infty} \mu(B_n) = \lim_{N\to\infty} \sum_{n=1}^{N} \mu(B_n) = \lim_{N\to\infty} \mu(A_N).$$

Continuity from above. The proof follows from (iii) by taking complements (*exercise:* check that the assumption $\mu(A_k) < \infty$ is necessary by coming up with an example where the property fails without it). □

We now talk about some important σ-algebras. For any $\mathcal{A} \subset 2^\Omega$, we define the σ-algebra *generated* by $\mathcal{A}$ as the smallest σ-algebra which contains $\mathcal{A}$. Explicitly,

$$\sigma(\mathcal{A}) := \bigcap_{\mathcal{F}:\mathcal{F} \text{ is a } \sigma\text{-algebra containing } \mathcal{A}} \mathcal{F}.$$

Exercise 1.4. Check that this definition is well-defined, and that $\sigma(\mathcal{A})$ is a σ-algebra.

A common tactic in proving that two σ-algebras are the same is through their generating sets. In particular, suppose $X, Y \subset 2^\Omega$, and we want to show that

$$\sigma(X) = \sigma(Y).$$

Then it suffices to show that

$$X \subseteq \sigma(Y), \quad Y \subseteq \sigma(X).$$

The *Borel* σ-algebra is one generated by the collection of open sets in any topological space. The most common case is in $\Omega = \mathbb{R}$, and we denote $\mathcal{B} = \mathcal{B}(\mathbb{R})$ the Borel σ-algebra of $\mathbb{R}$ under the usual topology.

Exercise 1.5. Show that the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ is generated by the collection of open intervals. Likewise show that it can be alternatively generated by the collection of closed intervals (hint: $\mathbb{Q}$ is dense in $\mathbb{R}$).

1.2.1 Constructing the Uniform Measure

We will now attempt to define the uniform measure on $\mathcal{B}(\mathbb{R})$. Before we begin, a rough outline:

1. First define it on "intervals" for a "semi-algebra".
2. Extend the definition to an "algebra".

3. Extend to a σ-algebra, namely $\mathcal{B}(\mathbb{R})$.

A *semi-algebra* $\mathcal{S}$ has the following properties:

- If $S_1, S_2 \in \mathcal{S}$, then $S_1 \cap S_2 \in \mathcal{S}$.
- If $S \in \mathcal{S}$, then S^c is a finite disjoint union of sets in $\mathcal{S}$.

It's not too hard to check that the collection of all clopen intervals forms a semi-algebra, i.e.

$$\mathcal{S} = \{(a, b] : a, b \in \mathbb{R}\}.$$

Now let $\mu((a, b]) := b - a$. We claim that μ is countably additive. Let $A = \cup_n B_n$ be a disjoint union, where all sets are in $\mathcal{S}$. Clearly

$$\mu(A) \geq \sum_n \mu(B_n).$$

For the other direction, we use compactness. Since

$$A' = \left[a + \frac{1}{m}, b\right] \subset A$$

is compact, and

$$B_n = (c_n, d_n] \subset (c_n, d_n + \epsilon/2^n) = B'_n,$$

we have that

$$A' \subset \cup_{j=1}^k B'_{n_j}.$$

Then we have a finite subcovering, so we may write

$$b - \left(a + \frac{1}{m}\right) \leq \sum_{j=1}^k [d'_{n_j} - c'_{n_j}] \leq \epsilon + \sum_{j=1}^k [d_{n_j} - c_{n_j}]$$

Now, since ϵ and m are arbitrary, we have

$$\mu(A) \leq \sum_n \mu(B_n).$$

We continue by extending our measure μ from a semi-algebra to an algebra. An *algebra* $\mathcal{A}$ satisfies

- $A \in \mathcal{A}$ implies $A^c \in \mathcal{A}$.
- $A_1, A_2 \in \mathcal{A}$ implies $A_1 \cap A_2 \in \mathcal{A}$.

We claim that the collection of all finite disjoint unions of clopen intervals,

$$\mathcal{A} = \{\sqcup_{i=1}^n A_i : A_i \in \mathcal{S}\}$$

is an algebra. For finite intersections, note that

$$(\cup_{i=1}^n A_i) \cap \left(\cup_{j=1}^m B_j\right) = \sqcup_{i,j} (A_i \cap B_j)$$

is a finite disjoint union of clopen intervals, and thus belongs to $\mathcal{A}$.

For complements, we wish to show $\cap_{i=1}^n A_i^c$ belongs to $\mathcal{A}$. Since each $A_i \in \mathcal{S}$, we know that A_i^c is a disjoint union of clopen intervals, and thus belongs to $\mathcal{A}$. Since we've shown that finite intersections belong to $\mathcal{A}$, we are done.

We now must extend our measure μ to $\mathcal{A}$. By a *measure on an algebra*, we mean a set function μ which satisfies

(i) $\mu(A) \geq \mu(\varnothing) = 0$ for all $A \in \mathcal{A}$.
(ii) If $A_n \in \mathcal{A}$ and $\sqcup_{n=1}^{\infty} A_n = A \in \mathcal{A}$, then

$$\mu(\sqcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n).$$

The first thing we need to check that the measures assigned to sets are well-defined. Suppose

$$A = \cup_{i=1}^{k} B_i = \cup_{j=1}^{l} B_j,$$

where $B_i, C_j \in \mathcal{S}$. We wish to prove

$$\mu(A) = \sum_{i=1}^{k} \mu(B_i) \stackrel{?}{=} \sum_{j=1}^{l} \mu(C_j).$$

Note that

$$B_i = \sqcup_{j=1}^{l} [B_i \cap C_j],$$

and similarly,

$$C_j = \sqcup_{i=1}^{k} [C_j \cap B_i].$$

Then by countable additivity of the measure defined on $\mathcal{S}$, we have

$$\sum_{i=1}^{k} \mu(B_i) = \sum_{i=1}^{k} \sum_{j=1}^{l} \mu(B_i \cap C_j) = \sum_{j=1}^{k} \mu(C_j),$$

verifying that μ is well-defined.

We will next prove that μ is countably additive on $\mathcal{A}$. Note that finite additivity follows immediately by decomposing each set into a finite disjoint union as in the definition of $\mathcal{A}$. Then, this immediately implies monotonicity, which we'll need to show countable additivity.

Let (A_i) be a sequence in $\mathcal{A}$ such that $\sqcup_i A_i = A \in \mathcal{A}$. By monotonicity, we have

$$\mu(A) \geq \sum_{i=1}^{n} \mu(A_i).$$

Taking the limit $n \to \infty$, we deduce that

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

To prove the other direction, write

$$A = \cup_{j=1}^{k} C_j, \quad C_j \in \mathcal{S}$$
$$A_i = \cup_{l=1}^{m_i} C_l^{(i)}, \quad C_l^{(i)} \in \mathcal{S}.$$

Then it suffices to show, for each j,

$$\mu(C_j) \leq \sum_{i=1}^{\infty} \mu(C_j \cap A_i).$$

since we can just use finite disjoint additivity while summing over each j. We may write

$$C_j \cap A_i = \cup_{l=1}^{m_i} [C_j \cap C_l^{(i)}] = \cup_{i=1}^{\infty} C_j \cap A_i.$$

The rest of the proof is just an exercise in set manipulation, and left as an exercise.

Now, we've extended μ from $\mathcal{S}$ to $\mathcal{A}$. Our next step is to extend μ to $\sigma(\mathcal{A})$ using Caratheodory's extension theorem, which is left as optional reading in the next section.

1.2.2 Caratheodory's Extension Theorem*

We devote this section to conveying the main ideas of the extension theorem. As such, some of the more tedious details may be left out. For a complete treatment, see [Dur19] or any other introductory text on measure theory.

We say a measure μ on space $(\Omega, \mathcal{F})$ is *σ-finite* if there exists a countable covering of Ω by finite measure sets in $\mathcal{F}$. Note that probability measures are trivially σ-finite. It turns out that this condition is necessary to prove uniqueness of extension, as the following example shows.

Example 1.6. Consider the semi-algebra $\mathcal{S} = \{(a, b] \cap \mathbb{Q} : a, b \in \mathbb{R}\}$ on the space $\Omega = \mathbb{Q}$. Then $\sigma(\mathcal{S}) = 2^{\mathbb{Q}}$. Note that the cardinality of each element of $\mathcal{S}$ is either ∞ or 0, so we can define a measure μ on $\mathcal{S}$ by

$$\mu(A) = \begin{cases} \infty & \text{if } |A| = \infty \\ 0 & \text{o.w.} \end{cases}$$

Now, we want to construct two distinct extensions $\mu_1 \neq \mu_2$ that agree with μ on $\mathcal{S}$. One possible example is

- μ_1 is the cardinality of a set (counting measure).
- $\mu_2 = 2\mu_1$.

Note now that μ is not σ-finite. Thus the σ-finite condition is necessary in the extension theorem.

Theorem 1.7

Given a countably additive measure on an algebra $\mathcal{A}$, it can be extended to a measure on $\sigma(\mathcal{A})$. If μ is σ-finite on $\mathcal{A}$, then the extension is unique.

Proof of Uniqueness. We use the good set principle. This is a general strategy for showing some desired property is true for a sigma algebra. We show that the class of sets satisfying the property is closed under σ-algebra operations. But then any σ-algebra containing the σ-algebra generated by some class $\mathcal{A}$ must contain the whole of $\sigma(\mathcal{A})$. A result of this type that we will use is Dynkin's $\pi - \lambda$ theorem.

A *π-system* is a collection of sets closed under finite intersection. A *λ-system* is a collection G of sets satisfying

(i) $\Omega \in G$.
(ii) $A \subset B$ and $A, B \in G$ implies $B \setminus A \in G$.
(iii) $A_i \in G$ and $A_i \uparrow A$ implies $A \in G$.

Lemma 1.8 (Dynkin's $\pi - \lambda$ theorem)

Let $\mathcal{P}$ be a π-system that is contained in a λ-system $\mathcal{L}$. Then

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

Proof. See [Dur19, A.1.4]. Note that this lemma is just a stronger version of the good set principle, in which instead of using λ-systems, we use the finer class of σ-algebras. □

We will first prove the case where $\mu(X) < \infty$. With this tool, note that a semi-algebra is a π-system, and let

$$\mathcal{L} = \{A : \mu_1(A) = \mu_2(A)\},$$

where μ_1 and μ_2 are the two extensions that we wish to show are equal. We know that $\mathcal{S} \subset \mathcal{L}$, so it suffices to show that $\mathcal{L}$ is a λ-system. A σ-algebra is also a λ-system, so given any π-system $\mathcal{P}$, we know that $\sigma(\mathcal{P})$ is the smallest λ-system containing $\mathcal{P}$. To verify $\mathcal{L}$ is a λ-system:

(i) $\Omega \in \mathcal{L}$ because $\Omega \in \mathcal{A}$.

(ii) Let $A \subset B$ and $A, B \in \mathcal{L}$. Then $\mu_1(A) = \mu_2(A)$ and $\mu_1(B) = \mu_2(B)$. Then, since $\mu_1(\Omega) = \mu(\Omega) = \mu_2(\Omega) < \infty$, we have

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

(iii) If $A_i \in \mathcal{L}$ and $A_i \uparrow A$, then we use continuity from below to see that

$$\mu_1(A_i) \to \mu_1(A)$$
$$\mu_2(A_i) \to \mu_2(A),$$

for $A \in \mathcal{L}$.

Exercise: Modify the proof slightly to include the σ-finite case. □

Sketch of Existence. Let $B \subset \mathbb{R}$. How do we define $\mu(B)$? We could try to approximate B by a union of intervals. This is known as *outer measure*. In particular, let

$$\mu^*(B) := \inf_{\substack{(A_i) \subset \mathcal{A} \\ B \subset \cup_i A_i}} \sum_{i=1}^{\infty} \mu(A_i).$$

Then, we say that a set E is measurable if it satisfies the Caratheodory criterion. That is,

$$\mu^*(F) = \mu^*(F \cap E) + \mu^*(F \cap E^c)$$

for all $F \subset \Omega$. The geometric intuition behind this definition is that E is deemed measurable if and only if it "cleanly" cuts every test set F.

First, we begin by showing that this new definition extends the old one. That is, if $A \in \mathcal{A}$, then $\mu^*(A) = \mu(A)$ and A is measurable. Then, we check that our outer measure has the following properties:

(i) The empty set is a null-set,

$$\mu_*(\varnothing) = 0.$$

(ii) Monotonicity, for $A \subset B$

$$\mu^*(A) \leq \mu * (B).$$

(iii) Countable subadditivity. That is,

$$\mu^*(\cup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \mu^*(A_n).$$

Then, we proceed by showing that the class $\mathcal{M}$ of μ^*-measurable sets is a σ-algebra, and we finish off the proof by upgrading countable subadditivity to countable additivity. This verifies that the restriction of μ^* to $\mathcal{M}$ is a measure. For the details, see [Dur19, A.1]. □

1.3 Random Variables & Distributions

We now consider functions between measure spaces. Suppose we have two measure spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$. We say that a function $f : \Omega_1 \to \Omega_2$ is *measurable* if

$$f^{-1}(A_2) \in \mathcal{F}_1 \quad \forall A_2 \in \mathcal{F}_2.$$

If $(\Omega_2, \mathcal{F}_2) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then f will be called a *random variable*. Denote the space of $\mathcal{F}$-measurable (here $\mathcal{F}$ emphasizes the domain) functions by $m\mathcal{F}$.

In particular, if we have a random variable X on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then we may talk about the probabilities

$$\mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B)), \quad B \in \mathcal{B}(\mathbb{R}),$$

since the preimages of Borel sets are measurable by definition.

Lemma 1.9
Suppose $\mathcal{F}_2 = \sigma(\mathcal{A})$. Then to check that f is measurable, it suffices to check that $f^{-1}(B) \in \mathcal{F}_1$ for each $B \in \mathcal{A}$.

Proof. We use the good set principle. Define

$$\mathcal{F}' = \{B : f^{-1}(B) \in \mathcal{F}_1\}.$$

Because inverses preserve complements and countable unions, we see that $\mathcal{F}'$ is a σ-algebra. Now, since $\mathcal{F}' \supset \mathcal{A}$, it follows that $\mathcal{F}' \supset \sigma(\mathcal{A})$. □

Exercise 1.10. Verify the following closure properties of measurability.

(i) If $f_1, f_2 \in m\mathcal{F}$, then $f_1 \circ f_2 \in m\mathcal{F}$.
(ii) If $f_1, f_2 \in m\mathcal{F}$, then (f_1, f_2) is also measurable.
(iii) Any continuous function is measurable from $(\Omega_1, \mathcal{B}(\Omega_1))$ to $(\Omega_2, \mathcal{B}(\Omega_2))$.
(iv) If $f_1, \ldots, f_n \in m\mathcal{F}$, then $f_1 + \cdots + f_n \in m\mathcal{F}$.
(v) Suppose $f_n \in m\mathcal{F}$ and $f_n \to f$ pointwise. Then $f \in m\mathcal{F}$.

We say a mapping $F : \mathbb{R} \to [0, 1]$ is a *distribution* function if it satisfies:

(i) F is nondecreasing.
(ii) F is right continuous.
(iii) $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.

From Caratheodory's extension theorem, it turns out that for any given distribution function F, there is a unique measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, satisfying

$$\mu([a, b]) = F(b) - F(a). \tag{1}$$

In the other direction, given a measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we can show that

$$F(b) := \mu(-\infty, b]$$

is a distribution function.

Now, suppose we have a measurable function f between spaces $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2)$. Then notice that f induces a measure μ_2, called the *pushforward*, on $(\Omega_2, \mathcal{F}_2)$ given by

$$\mu_2(A_2) = \mu_1(f^{-1}(A_2)).$$

Naturally, this induces a distribution function.

Conversely, it turns out that we can construct a random variable given a distribution function induced by μ. We know that we can construct a measure on $\mathbb{R}$ via Caratheodory's extension theorem. So then we take

$$I : (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$$

the identity map to be our random variable. But perhaps we wish to come up with a method that does not resort to the heavy machinery of Caratheodory. Indeed there are more straightforward ways to do this.

So say we are given a distribution function F. We want to construct a random variable X that satisfies

$$\mathbb{P}(X^{-1}(-\infty, y]) = F(y)$$

for every y. Consider the definition $X(F(y)) = y$. This works only if F is continuous, and is never flat. For general F, we need to work a little harder. In particular, one can check that the construction

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

gives us a random variable.

If two random variables X and Y induce the same distribution μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then we say that they are equal in distribution, and write

$$X \overset{d}{=} Y.$$

Then (1) tells us that to check two variables agree in distribution, it suffices to check that $\mathbb{P}(X \leq x) = \mathbb{P}(Y \leq x)$ for all x.

1.4 Integration: Construction

Suppose we are given a measure space $(\Omega, \mathcal{F}, \mu)$, where μ is a σ-finite measure. We wish to define the integral of measurable functions $f : \Omega \to \mathbb{R}$, i.e.

$$\int_\Omega f d\mu.$$

The strategy we will use, which can be generalized to problems in many contexts, is as follows:

1. Establish for simple functions.
2. Extend to bounded functions.
3. Extend to nonnegative functions.
4. Extend to general functions.

Step 1: We define a *simple* function to be a finite linear combination of indicators,

$$f = \sum_{i=1}^n c_i \mathbf{1}_{A_i},$$

for disjoint measurable sets A_i with $\mu(A_i) < \infty$ for all i. Note that the disjointness assumption is made without loss of generality. The intuitive definition of the integral is then

$$\int f d\mu = \sum_{i=1}^n c_i \mu(A_i).$$

To check that this is well-defined, suppose

$$f = \sum_{i=1}^n c_i \mathbf{1}_{A_i} = \sum_{j=1}^m d_j \mathbf{1}_{B_j}.$$

Then consider the sets $A_i \cap B_j$, and observe that if $\mu(A_i \cap B_j) > 0$, then $c_i = d_j$. Then we may write

$$f = \sum_{i,j} c_i \mathbf{1}_{A_i \cap B_j}.$$

It follows that

$$\sum_{i=1}^n c_i \mu(A_i) = \sum_{i,j} c_i \mu(A_i \cap B_j) = \sum_{j=1}^m d_j \mu(B_j).$$

The following properties are easy to check.

Proposition 1.11

Let f be a simple function.

(i) $f \overset{a.e.}{\geq} 0 \implies \int f \geq 0$.
(ii) $\int af = a \int f$.
(iii) $\int (f + g) = \int f + \int g$.
(iv) $f \overset{a.e.}{\leq} g \implies \int f \leq \int g$.
(v) $f \overset{a.e.}{=} g \implies \int f = \int g$.
(vi) $|\int f| \leq \int |f|$

Step 2: Now we consider *bounded* functions f, which we define as satisfying

- $|f| \leq M < \infty$, and
- f vanishes outside a set E where $\mu(E) < \infty$.

We have two options, to approximate from above or from below. But it turns out that both methods are equivalent. First note that

$$\sup_{\substack{g \text{ simple} \\ g \leq f}} \int g \leq \inf_{\substack{h \text{ simple} \\ h \geq f}} \int h. \tag{2}$$

To prove equality, it suffices to find g, h for every $\epsilon > 0$ with

$$\int h - \int g < \epsilon.$$

To do this, we will construct an h such that $h - f < \epsilon$ pointwise, and similarly a g with $f - g < \epsilon$ pointwise. In particular, partition the range of f into intervals $(I_n)_{n=1}^m$ each of size at most ϵ. Then define

$$A_n = f^{-1}(I_n),$$

and let

$$h = \sum_{n=1}^{m} \sup I_n \mathbf{1}_{A_n}$$
$$g = \sum_{n=1}^{m} \inf I_n \mathbf{1}_{A_n}.$$

Now, we have

$$\int h - \int g \leq 2\epsilon\mu(E),$$

and since ϵ is arbitrary, we have equality in (2).

Exercise 1.12. Check Proposition 1.11 for bounded functions.

Step 3: We now extend the integral for all nonnegative functions. Note that it only really makes sense to approximate a potentially unbounded function by bounded functions from below. So consider the definition

$$\int f = \sup \left\{ \int h : h \text{ is bounded, } 0 \leq h \leq f \right\}.$$

Once again, we'd like to verify the properties in Proposition 1.11. We prove linearity, and leave the rest as easy exercises.

First, note that the direction

$$\int f + \int g \leq \int (f + g)$$

is the same as before, following directly from the definition. The other direction is trickier since we do not have access to approximations from above. So consider the growing sets E_n with $E_n \uparrow \Omega$ and $\mu(E_n) < \infty$ (note that these exist only because we assumed our measure to be σ-finite). Let $f \wedge n$ be the pointwise minimum between f and n. Define $h_n = (f \wedge n)\mathbf{1}_{E_n}$. We will show that

$$\int h_n \uparrow \int f.$$

First note that

$$\lim_{n\to\infty} \int h_n \leq \int f.$$

To prove the other direction, produce from the definition a bounded g with

$$\int f \leq \int g + \epsilon.$$

If $|g| \leq M$, then for $n \geq M$, we have

$$\int h_n \geq \int_{E_n} g = \int g - \int_{E_n^c} g.$$

But since

$$\cap_n E_n^c = (\cup_n E_n)^c = \varnothing,$$

it follows that

$$0 \leq \int_{E_n^c} g \leq M\mu(E_n^c) \to 0.$$

From this we deduce that

$$\int f \leq \int g + \epsilon \leq \lim_{n\to\infty} \int h_n + \epsilon,$$

and since ϵ is arbitrary, we conclude that

$$\int (f \wedge n)\mathbf{1}_{E_n} \uparrow \int f. \tag{3}$$

Now, returning to the proof of linearity, note that since $(a+b)\wedge n \leq (a\wedge n)+(b\wedge n)$, we have

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

Letting $n \to \infty$, this becomes

$$\int (f+g) \leq \int f + \int g.$$

With both sides now proven, linearity follows.

Exercise 1.13. Check the rest of Proposition 1.11 for nonnegative functions.

Step 4: For arbitrary measurable functions f, we will define $\int f$ only when $\int |f| < \infty$ or when $f \geq 0$ (in which case the integral may be infinite). We consider the positive and negative parts of f,

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

respectively, where $a \vee b := \max(a, b)$. Note that

$$f = f^+ - f^-,$$
$$|f| = f^+ + f^-.$$

So, we define the general integral by

$$\int f = \int f^+ - \int f^-.$$

Exercise 1.14. Once more, verify the properties in Proposition 1.11.

1.5 Integration: Tools

1.5.1 Inequalities

Theorem 1.15 (Markov's inequality)
Let $f \geq 0$ be measurable with respect to μ. Then

$$\mu(f \geq \epsilon) \leq \epsilon^{-1} \int f d\mu.$$

Proof. Simply note that

$$\int f d\mu \geq \int_{f \geq \epsilon} \epsilon d\mu = \epsilon \mu(f \geq \epsilon),$$

to get the desired inequality. □

Theorem 1.16 (Jensen's inequality)
Let φ be a convex function, 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 if f and $\varphi(f)$ are integrable with respect to a probability meaure μ,

$$\varphi\left(\int f d\mu\right) \leq \int \varphi(f) d\mu.$$

Remark. Note that we require μ to be a probability measure. In general the inequality does not hold for infinite measures, for consider the Lebesgue measure on $[1, \infty)$ with $f(x) = x^{-1}$ and $\varphi(x) = x^2$.

Proof. Let $M = \int f d\mu$. Convexity allows us to produce a linear function $l(x) = ax + b$ such that $l(M) = \varphi(M)$ and $\varphi(x) \geq l(x)$ for all x. It follows that

$$\int \varphi(f) d\mu \geq \int (af + b) d\mu = aM + b = l(M) = \varphi(M) = \varphi\left(\int f d\mu\right),$$

as desired. □

Theorem 1.17 (Hölder's inequality)
Let $p, q \in (1, \infty)$ be complements in the sense that $\frac{1}{p} + \frac{1}{q} = 1$. Then

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

Proof. Note that the case where $\|f\|_p = 0$ or $\|g\|_q = 0$ is trivial, as is the case where either one is infinite, so assume $\|f\|_p, \|g\|_q \in (0, \infty)$. By homogeneity, we may further assume that $\|f\|_p = \|g\|_q = 1$. Consider the function

$$\varphi(x) = \frac{x^p}{p} + \frac{y^q}{q} - xy, \quad x \geq 0.$$

Through some basic calculus, one can show that φ is convex and entirely nonnegative. Thus letting $x = |f|$, $y = |g|$, and integrating, we get

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

which is what we wanted. □

Exercise 1.18. Let $\|f\|_\infty = \inf\{M : \mu(\{x : |f(x)| > M\}) = 0\}$. Show that

$$\int |fg| d\mu \leq \|f\|_1 \|g\|_\infty.$$

1.5.2 Convergence Theorems

We now discuss the behavior of integration of sequences of functions f_n under the limit. In particular, let

$$f = \lim_{n \to \infty} f_n.$$

be the almost sure limit of a sequence of measurable functions f_n. Note that wlog we may assume f is measurable by defining f to be zero on the complement of the set on which f_n converges. We will examine the validity of the statement:

$$\int f_n \to \int f.$$

First, we examine a notion of convergence that will lay the intuitive groundwork for our first convergence result. We say $f_n \to f$ *in measure* if for every $\epsilon > 0$,

$$\lim_{n \to \infty} \mu(|f_n - f| > \epsilon) \to 0,$$

and we may write $f_n \xrightarrow{\mu} f$ to denote this. It turns out that, if $\mu(\Omega) < \infty$, then almost sure convergence implies convergence in measure. However, the converse is false.

Exercise 1.19. Verify that if $\mu(\Omega) < \infty$, then almost sure convergence implies convergence in measure. Come up with an example in which we have convergence in measure, but not almost sure convergence.

Example 1.20. Consider the functions $f_n = \mathbf{1}_{[-n,n]}$ which converge almost surely to $f = \mathbf{1}_{\mathbb{R}}$, but not in measure.

Back to the main question. If $f_n \to f$ almost surely and we are on a σ-finite space, we also have convergence in measure. In particular, we may ask if $f_n \xrightarrow{\mu} f$ implies

$$\int f_n \to \int f.$$

It turns out that the answer is no, for we have the counterexample

$$f_n(x) = \begin{cases} n & x \in [0, 1/n) \\ 0 & \text{o.w.} \end{cases}$$

which converges in measure to 0, but the integrals do not converge.

However, if we enforced the condition that f_n were bounded, we would eliminate the potential for mass to "escape to infinity". This is exactly the content of the next theorem.

Theorem 1.21 (Bounded convergence theorem)
Suppose $\mu(\Omega) < \infty$, or that f_n vanishes outside a set of finite measure, say E. If $|f_n| \leq M$ and $f_n \xrightarrow{\mu} f$, then

$$\int f_n \to \int f.$$

Proof. Let $\epsilon > 0$. Then the set

$$B_{n,\epsilon} = \{|f_n - f| > \epsilon\}$$

has measure $\mu(B_{n,\epsilon})$ converging to 0 as $n \to \infty$. It's not too hard to see that $|f| \leq M$, so one has

$$\begin{aligned}\left|\int (f - f_n)\right| &\leq \int |f - f_n| \\ &= \int_{B_{n,\epsilon}} |f - f_n| + \int_{B_{n,\epsilon}^c} |f - f_n| \\ &\leq 2M\mu(B_{n,\epsilon}) + \epsilon\mu(E).\end{aligned}$$

Taking $n \to \infty$, the first term vanishes. Since ϵ is arbitrary, the second term vanishes as well, and the claim follows. □

From this we get the following result, which roughly says that we can only lose mass in the limit.

Theorem 1.22 (Fatou's lemma)
If $f_n \geq 0$, then

$$\liminf_{n\to\infty} \int f_n \geq \int \liminf_{n\to\infty} f_n.$$

Proof. Let $g_n = \inf_{m \geq n} f_n$. Then clearly $f_n \geq g_n$, so $\int f_n \geq \int g_n$. Furthermore, since the g_n are monotone, they converge to some limit $g_n \uparrow g$. Thus

$$\liminf_n \int f_n \geq \liminf_n \int g_n = \lim_n \int g_n.$$

So it suffices to show that

$$\lim_n \int g_n \geq \int g.$$

We know that $g_n \wedge m \uparrow g \wedge m$, so by the bounded convergence theorem we have

$$\int g_n \wedge m \uparrow \int g \wedge m.$$

Then one has

$$\lim_n \int g_n \geq \lim_n \int g_n \wedge m = \int g \wedge m.$$

By (3), we can take the limit $m \to \infty$ to get

$$\lim_n \int g_n \geq \int g,$$

as desired. □

Theorem 1.23 (Monotone convergence theorem)
Suppose $0 \leq f_n \uparrow f$. Then

$$\lim_{n\to\infty} \int f_n \uparrow \int f.$$

Proof. By monotonicity of the integral, we have

$$\int f \geq \lim_n \int f_n.$$

By Fatou's lemma, we get the other direction

$$\int f = \int \liminf_n f_n \leq \liminf_n \int f_n = \lim_n \int f_n.$$

Thus our conclusion follows. □

Theorem 1.24 (Dominated convergence theorem)
Let $f_n \to f$ almost surely. If $|f_n| \leq g$ and $\int g < \infty$, then

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

Proof. Since $g + f_n \geq 0$ and $g + f_n \xrightarrow{a.s.} g + f$, we have by Fatou's lemma that

$$\liminf_n \int (g + f_n) \geq \int (g + f)$$

which implies

$$\liminf_n \int f_n \geq \int f.$$

Similarly, we can apply this to $-f_n$ to get the other direction

$$\limsup_n \int f_n \leq \int f.$$

Together, this implies that $\lim_n \int f_n$ exists, and is in fact $\int f$. □

1.6 Expectation

Now, we can translate all of the above results into the language of probability theory. In particular if we have a random variable X on a measure space $(\Omega, \mathcal{F}, \mathbb{P})$, we define its *expectation* as

$$\mathbb{E}X = \int_\Omega X d\mathbb{P}.$$

Then, we may restate our inequalities and convergence theorems as follows.

1.6.1 Inequalities

Theorem 1.25 (Markov's inequality)
If $X \geq 0$ then

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

Theorem 1.26 (Jensen's inequality)
If φ is convex then

$$\mathbb{E}\varphi(X) \geq \varphi(\mathbb{E}X),$$

provided $\mathbb{E}|X|, \mathbb{E}|\varphi(X)| < \infty$.

Theorem 1.27 (Hölder's inequality)
Let $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then

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

where $\|X\|_r = (\mathbb{E}|X|^r)^{1/r}$ for $r \in [1, \infty)$ and $\|X\|_\infty = \inf\{M : \mathbb{P}(|X| > M) = 0\}$.

1.6.2 Convergence Theorems

Theorem 1.28 (Bounded convergence theorem)
If $|X_n| \leq M$ for all n and $X_n \xrightarrow{a.s.} X$, then

$$\mathbb{E}X_n \to \mathbb{E}X.$$

Theorem 1.29 (Fatou's Lemma)
If $X_n \geq 0$, then

$$\liminf_{n\to\infty} \mathbb{E}X_n \geq \mathbb{E} \liminf_{n\to\infty} X_n.$$

Theorem 1.30 (Monotone convergence theorem)
If $0 \leq X_n \uparrow X$, then

$$\mathbb{E}X_n \uparrow \mathbb{E}X.$$

Theorem 1.31 (Dominated convergence theorem)
If $X_n \xrightarrow{a.s.} X$, and $|X_n| \leq Y$ for all n, with $\mathbb{E}Y < \infty$, then

$$\mathbb{E}X_n \to \mathbb{E}X.$$

1.6.3 Change of Measure

We now turn to the concept of *change of variables* (or change of measure) for random variables. This allows us to compute expectations on a space that's easier to work with. In particular, let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (S, \mathcal{S})$ be a random variable with distribution μ, i.e. $\mu(A) = \mathbb{P}(X \in A)$ for $A \in \mathcal{B}(\mathbb{R})$. Then, suppose we have a function $f : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Here is the picture:

$$(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} (S, \mathcal{S}, \mu) \xrightarrow{f} (\mathbb{R}, \mathcal{B}(\mathbb{R})).$$

Then the idea is we can perform a change of variables to do calculus on the intermediate space rather than the original.

Theorem 1.32 (Change of variables)
Let X and f be as described above.

$$\mathbb{E}f(X) = \int_\Omega f(X(\omega))d\mathbb{P} = \int_S f(x)\mu(dx).$$

Note: we use $\mu(dx)$ instead of $d\mu$ to emphasize the variable of integration.

Proof. We prove the statement for four increasingly general classes of functions.

Step 1: For an indicator function $\mathbf{1}_B$, where $B \in \mathcal{S}$, we have by definition that

$$\mathbb{E}\mathbf{1}_B(X) = \int_\Omega \mathbf{1}\{X \in B\}d\mathbb{P} = \mathbb{P}(X \in B) = \mu(B) = \int_\mathcal{S} \mathbf{1}_B(x)\mu(dx).$$

Step 2: For a simple function $f = \sum_{k=1}^n c_k \mathbf{1}_{B_k}$, where $c_k \in \mathbb{R}$ and $B_k \in \mathcal{S}$, we can apply linearity of the integral to get

$$\begin{aligned}
\mathbb{E}f(X) &= \sum_{k=1}^n c_k \mathbb{E}\mathbf{1}_{B_k}(X) \\
&= \sum_{k=1}^n c_k \int_\mathcal{S} \mathbf{1}_{B_k}(x)\mu(dx) \\
&= \int_\mathcal{S} f(x)\mu(dx).
\end{aligned}$$

Step 3: Let $f \geq 0$ be a nonnegative measurable function. We can approximate it by simple functions from below using

$$f_n(x) = \frac{\lfloor 2^n f(x) \rfloor}{2^n} \wedge n.$$

In particular, each f_n is simple and we have $0 \leq f_n \uparrow f$. We may then apply MCT twice to get

$$\mathbb{E}f(X) = \lim_n \mathbb{E}f_n(X) = \lim_n \int_{\mathcal{S}} f_n(x)\mu(dx) = \int_{\mathcal{S}} f(x)\mu(dx).$$

Step 4: Suppose f is integrable, i.e. $\mathbb{E}|f(X)| < \infty$. Then we do the usual by writing

$$f = f^+ - f^-$$

and since integrability guarantees that both $\mathbb{E}f(X)^+$ and $\mathbb{E}f(X)^-$ are finite, we can apply linearity to get

$$\begin{aligned}
\mathbb{E}f(X) &= \mathbb{E}f(X)^+ - \mathbb{E}f(X)^- \\
&= \int_{\mathcal{S}} f(x)^+\mu(dx) - \int_{\mathcal{S}} f(x)^-\mu(dx) \\
&= \int_{\mathcal{S}} f(x)\mu(dx).
\end{aligned}$$

□

1.7 Product Measures

Suppose we have two σ-finite measure spaces $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$. Consider the measurable space

$$(\Omega_1 \times \Omega_2, \sigma(\mathcal{F}_1 \times \mathcal{F}_2)).$$

Our goal is to construct the product measure μ for this product space.

Theorem 1.33 (Product measures)
There is a unique measure μ on the product space with

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

for every $A \in \mathcal{F}_1$ and $B \in \mathcal{F}_2$.

Proof sketch. From the Caratheodory extension theorem, it suffices to check countable additivity on $\mathcal{F}_1 \times \mathcal{F}_2$. That is, if $A\times = \sqcup_i (A_i \times B_i)$, then we require that

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

The strategy is to project onto lower dimensions. For any $x \in A$, the set of all y with $(x, y) \in A\times B$ is just B. Let's consider all A_i's that contain x. Then the corresponding B_i's form a disjoint partition of B. Then we know that

$$\mu_2(B) = \sum_{x \in A_i} \mu_2(B_i)$$

for every $x \in A$. Then

$$\mathbf{1}_{x\in A}\mu_2(B) = \sum_i \mathbf{1}_{A_i}\mu_2(B_i).$$

Integrating both sides with respect to μ_1 and applying MCT, we get

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

which is what we wanted to show. □

Having constructed the product measure, we wish to compute integrals on the product space. To build up to this, we start with the following claim.

Lemma 1.34
Let $E \in \mathcal{F}_1 \times \mathcal{F}_2$. For any $X \in \Omega_1$, the slice

$$E_x = \{y \in \Omega_2 : (x, y) \in E\}$$

is measurable.

Proof. Let

$$\mathcal{G} = \{E \in \mathcal{F}_1 \times \mathcal{F}_2 : E_x \text{ is measurable}\}$$

If $E \in \mathcal{G}$, then $(E_x)^c = (E^c)_x \in \mathcal{G}$. Also, if (E_n) is a countable sequence in $\mathcal{G}$, then likewise

$$\cup_{n=1}^{\infty}(E_i)_x = (\cup_{n=1}^{\infty}E_n)_x \in \mathcal{G}.$$

Thus $\mathcal{G}$ is a σ-algebra. But clearly $\mathcal{G}$ contains all the rectangles $E_1 \times E_2 \in \mathcal{F}_1 \times \mathcal{F}_2$, which is a generating set. Then $\mathcal{G} \supset \mathcal{F}_1 \times \mathcal{F}_2$, and our claim follows. □

Lemma 1.35
For any $E \in \mathcal{F}_1 \times \mathcal{F}_2$, we have

$$\mu(E) = \int_{\Omega_1} \mu_2(E_x) d\mu_1, \tag{4}$$

where $\mu_2(E_x)$ is a measurable function from $\Omega_1 \to \mathbb{R}$.

Proof. Let $\mathcal{A}$ be the collection of sets E for which $\mu_2(E_x)$ is measurable and (4) holds. It turns out that the vanilla good set principle is not good enough, so we will make use of the $\pi - \lambda$ theorem. In particular, we show that $\mathcal{A}$ is a λ-system:

(i) Clearly $E = \Omega_1 \times \Omega_2$ belongs to $\mathcal{A}$, since $E_x = \Omega_2$, and we have

$$\mu(E) = \mu(\Omega_1 \times \Omega_2) = \mu_1(\Omega_1)\mu_2(\Omega_2) = \int_{\Omega_1} \mu_2(\Omega_2) d\mu_1.$$

(ii) Suppose we have $A \supset B$ with both sets in $\mathcal{A}$. Since μ_1 and μ_2 are σ-finite, we will first assume that A is of finite measure, and extend afterwards. We write (note that finiteness justifies the subtraction here)

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

which proves measurability. Integrating gives (4). Now, to extend to the infinite case, note that we can rewrite $A = \cup_n A_n$ and $B = \cup_n B_n$ as countable disjoint unions of finite measure sets. Then we have

$$\mu_2((\cup_n A_n \setminus \cup_n B_n)_x) = \mu_2((\cup_n (A_n \setminus B_n))_x).$$

Since each $\cup_{n=1}^{N}(A_n \setminus B) \in \mathcal{A}$ by additivity and linearity, we can just use the next part to conclude that $A \setminus B \in \mathcal{A}$.

(iii) Suppose the conclusions hold for E_n and $E_n \uparrow E$. Then $(E_n)_x \uparrow E_x$, so in particular $\mu_2((E_n)x) \uparrow \mu_2(E_x)$ pointwise. Since a limit of measurable functions is measurable, it follows that $\mu_2(E_x)$ is measurable. Then, by MCT we get

$$\mu(E) = \lim_n \mu(E_n) = \lim_n \int_{\Omega_1} \mu_2((E_n)x) d\mu_1 = \int_{\Omega_1} \mu_2(E_x) d\mu_1,$$

and so $E \in \mathcal{A}$.

Now, since $\mathcal{A}$ contains all the rectangles, a π-system which generates $\mathcal{F}_1 \times \mathcal{F}_2$, we are done by Dynkin's theorem. □

With this in hand, we can prove an important result which allows us to swap the order of integration.

Theorem 1.36 (Fubini's theorem)
Let $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ be a measurable function with $f \geq 0$ or $\int |f| < \infty$. Then

(i) $\int f(\cdot, y) d\mu_1$ is measurable with respect to μ_2.
(ii) $\int f(x, \cdot) d\mu_2$ is measurable with respect to μ_1.

In particular, we have

$$\int_{\Omega_1} \int_{\Omega_2} f(x,y)\mu_2(dy)\mu_1(dx) = \int_{\Omega_1 \times \Omega_2} f d\mu = \int_{\Omega_2} \int_{\Omega_1} f(x,y)\mu_x(dx)\mu_2(dy) \tag{5}$$

Proof. Note that it suffices to prove just the first equality, for the second follows by symmetry. We verify the claim for four increasingly general cases.

Step 1: If $E \in \mathcal{F}_1 \times \mathcal{F}_2$, then the claim follows by Lemma 1.35.

Step 2: Since measurability is preserved under linear combinations, we get (i) and (ii). Linearity of the integral gives us (5).

Step 3: Suppose $f \geq 0$. Then using the same construction from before,

$$f_n(x) = \frac{\lfloor 2^n f(x) \rfloor}{2^n} \wedge n,$$

measurability follows since it is preserved under limits, and (5) follows by an application of MCT.

Step 4: If f is integrable, write $f = f^+ - f^-$. Then the claims follow by applying the previous step to f^+, f^-, and $|f|$. □

Exercise 1.37. Extend Fubini's theorem to higher dimensions.

Example 1.38. Let $\Omega_1 \times \Omega_2 = \mathbb{N} \times \mathbb{N}$ and μ_1, μ_2 be the counting measure. Consider the function $f(x, y)$ which looks like

0	0	0	0	-1
0	0	0	-1	1
0	0	-1	1	0
0	-1	1	0	0
-1	1	0	0	0

where the x and y axes are oriented as usual. It's easy to check that Fubini's theorem fails, as f is not integrable.

2 Laws of Large Numbers

Having laid the measure theoretic foundations, we now move into the first major topic of interest: laws of large numbers. That is, suppose we have sequences of random variables (X_n) with $S_n = X_1 + \cdots X_n$. We may often wish to analyze the behavior of S_n in the limit, with respect to various notions of convergence. To guarantee desirable behavior we will usually have to assume some level of "independence" between our random variables, so we start by developing a rigorous framework for independence. Then, we will discuss the various weak and strong laws of large numbers, which capture in probability and almost sure convergence, respectively.

2.1 Independence

Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose $\mathcal{F}_1, \ldots, \mathcal{F}_n \subset \mathcal{F}$ are sub-σ-algebras. We say that they are *independent* if whenever $A_i \in \mathcal{F}_i$, we have

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

Now, suppose $A_1, \ldots, A_n \in \mathcal{F}$ are events. We say that they are *independent* if whenever $I \subset \{1, \ldots, n\}$, we have

$$\mathbb{P}(\cap_{i\in I} A_i) = \prod_{i=1}^n \mathbb{P}(A_i).$$

Finally, suppose $X_1, \ldots, X_n$ are random variables. We say that they are *independent* if whenever $B_i \in \mathcal{B}(\mathbb{R})$, we have

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

Exercise 2.1. Check that the above definitions are equivalent in following senses:

(i) Random varibles $X_1, \ldots, X_n$ are independent iff the *induced* σ-algebras

$$\sigma(X_1), \ldots, \sigma(X_n)$$

are independent, where we define

$$\sigma(X_1) := \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}.$$

(ii) Events $A_1, \ldots, A_n$ are independent iff their indicators $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_n}$ are independent.

Example 2.2 (Pairwise but not mutually independent)**.** Consider two independent coin tosses. Let A_1 be the event where the first coin comes up heads, A_2 be the event where the second coin comes up heads, and A_3 be the event that the two coins come up the same. It's easy to check that these events are pairwise independent, but not mutually independent. Likewise, the same is true for their indicator random variables $\mathbf{1}_{A_i}$, $i = 1, 2, 3$.

We say that collections of set $\mathcal{A}_1, \ldots, \mathcal{A}_n \subset \mathcal{F}$ are *independent* if whenever $A_i \in \mathcal{A}_i$ and $I \subset \{1, \ldots, n\}$, we have $\mathbb{P}(\cap_{i\in I} A_i) = \prod_{i\in I} \mathbb{P}(A_i)$.

Theorem 2.3

Suppose $\mathcal{A}_1, \ldots, \mathcal{A}_2 \subset \mathcal{F}$ are independent π-systems. Then $\sigma(\mathcal{A}_1), \ldots, \sigma(\mathcal{A}_n)$ are independent.

Proof. Wlog we can assume that each $\mathcal{A}_i$ contains Ω, since if the $\mathcal{A}_i$'s are independent, then so are the $\overline{\mathcal{A}}_i = \mathcal{A}_i \cup \{\Omega\}$, and clearly $\sigma(\mathcal{A}_i) = \sigma(\overline{\mathcal{A}}_i)$. Then the independence condition becomes

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

whenever $A_i \in \mathcal{A}_i$, since we can set $A_i = \Omega$ for any $i \notin I$.

Now, we will make use of the $\pi - \lambda$ theorem. Use F to denote $A_2 \cap \cdots \cap A_n$. Let $\mathcal{L}$ be the class of all $A \subset \Omega$ such that

$$\mathbb{P}(A \cap F) = \mathbb{P}(A \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A) \prod_{i=2}^n \mathbb{P}(A_i) = \mathbb{P}(A)\mathbb{P}(F),$$

for any combination of $A_i \in \mathcal{A}_i$ for $2 \leq i \leq n$. Clearly $\mathcal{L} \supset \mathcal{A}_1$. We check that $\mathcal{L}$ is a λ-system:

(i) $\Omega \in \mathcal{L}$ since $\Omega \in \mathcal{A}_1$.

(ii) If $A \subset B$ are both in $\mathcal{L}$, then since

$$(B \setminus A) \cap F = (B \cap F) \setminus (A \cap F),$$

we have

$$\mathbb{P}((B \setminus A) \cap F) = \mathbb{P}(B \cap F) - \mathbb{P}(A \cap F) = (\mathbb{P}(B) - \mathbb{P}(A))\mathbb{P}(F) = \mathbb{P}(B \setminus A)\mathbb{P}(F),$$

so $B \setminus A \in \mathcal{L}$.

(iii) Suppose $B_k \in \mathcal{L}$ with $B_k \uparrow B$. Then $(B_k \cap F) \uparrow (B \cap F)$, so we have

$$\mathbb{P}(B \cap F) = \lim_k \mathbb{P}(B_k \cap F) = \lim_k \mathbb{P}(B_k)\mathbb{P}(F) = \mathbb{P}(B)\mathbb{P}(F),$$

so $B \in \mathcal{L}$.

Thus by Dynkin's theorem, we deduce that $\sigma(\mathcal{A}_1) \subset \mathcal{L}$. Now, we've shown that

$$\sigma(\mathcal{A}_1), \mathcal{A}_2, \ldots, \mathcal{A}_n$$

are independent. We can rearrange and iterate to conclude

$$\sigma(\mathcal{A}_1), \sigma(\mathcal{A}_2), \ldots, \sigma(\mathcal{A}_n)$$

are independent. □

Corollary 2.4

To check that $X_1, \ldots, X_n$ are independent, it suffices to check for all $x_1, \ldots, x_n \in (-\infty, \infty]$ that

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

Proof. The sets $\{X_i \leq x\}$ form a π-system which generates $\sigma(X_i)$, so the result follows immediately from Theorem 2.3. □

Exercise 2.5. Suppose $X_{i,j}$ are independent, where $1 \leq i \leq n$, $1 \leq j \leq m(i)$, and let $f_i : \mathbb{R}^{m(i)} \to \mathbb{R}$ be measurable. Show that over $1 \leq i \leq n$,

$$f_i(X_{i,1}, \ldots, X_{i,m(i)})$$

are independent.

2.1.1 Infinite Sequences of Random Variables

So far we've only discussed independence with regards to a finite collection $(X_n)_{n=1}^m$ of random variables. But what if we wish to consider an infinite sequence $(X_n)_{n\geq 1}$, say, some stochastic process indexed by time?

For starters, we'd like to define a sequence $(X_n)_{n\geq 1}$ of random variables to be *independent* if for every finite collection $I \subset \mathbb{N}$, the random variables $(X_i)_{i\in I}$ are independent.

One question we may ask if whether we can even construct such sequences of random variables. For finite sequences, this can be done as follows. Suppose we are given the distributions μ_i for $1 \leq i \leq n$. Then let $(\Omega, \mathcal{F}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, and set

$$X_i(\omega_1, \ldots, \omega_n) = \omega_i.$$

Then, we may define $\mathbb{P}$ to be the product measure $\mu_1 \times \cdots \times \mu_n$ on $\mathcal{B}(\mathbb{R}^n)$ given by

$$\mathbb{P}((a_1, b_1] \times \cdots \times (a_n, b_n]) = \mu_1((a_1, b_1]) \cdots \mu_n((a_n, b_n]).$$

For infinite sequences, the existence is trickier. First, we define the infinite product space $\mathbb{R}^{\mathbb{N}}$ equipped with the product σ-algebra

$$\mathcal{B}(\mathbb{R})^{\mathbb{N}} := \sigma(\{\cap_{i=1}^n \{X_i \in B_i\} : n \in \mathbb{N}, B_i \in \mathcal{B}(\mathbb{R})\}).$$

Then the existence of an extension of probability measures to $\mathcal{B}(\mathbb{R}^{\mathbb{N}})$ is a consequence of the following theorem.

Theorem 2.6 (Kolmogorov's extension theorem)
Let $\mathbb{P}_n$ be probability measures on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ that are *consistent*, that is,

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

for all intervals $(a_i, b_i] \subset \mathbb{R}$. Then there exists a unique probability measure $\mathbb{P}$ on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R})^{\mathbb{N}})$ with

$$\mathbb{P}(\cap_{i=1}^n \{X_i \in (a_i, b_i]\}) = \mathbb{P}_n((a_1, b_1] \times \cdots \times (a_n, b_n]).$$

Proof. The idea is to use Caratheodory's extension theorem. See [Dur19, A.3.1]. □

With the existence out of the way, we're ready to exhibit a remarkable result which roughly says that any event which only depends on the behavior of a sequence of independent random variables (X_n) in the limit, must have probability 0 or 1. Results of this type are known as *0-1 laws*. Later we will see an example of this in the Borel-Cantelli lemmas.

We start with some notation. Define

$$\mathcal{T}_n = \sigma(X_n, X_{n+1}, \ldots)$$

to be the smallest σ-algebra with respect to which all the X_m are measurable for $m \geq n$. In words, this is the observable future after time n. Then let

$$\mathcal{T} = \cap_{n=1}^{\infty} \mathcal{T}_n$$

be the *tail σ-algebra*. In words, this is the observable "remote" future, and $A \in \mathcal{T}$ iff changing a finite number of values does not affect the occurence of the event.

Exercise 2.7. Check the following examples:

- For $B_n \in \mathcal{B}(\mathbb{R})$, the event $\{X_n \in B_n \text{ i.o.}\} \in \mathcal{T}$, where i.o. stands for "infinitely often".
- For $S_n = X_1 + \cdots + X_n$, the event $\{\lim_{n\to\infty} S_n \text{ exists}\} \in \mathcal{T}$.

Then, verify the following counterexample:

- For $S_n = X_1 + \cdots + X_n$, the event $\{\limsup_{n\to\infty} S_n > 0\} \notin \mathcal{T}$.

Theorem 2.8 (Kolmogorov's 0-1 law)
Suppose $X_1, X_2, \ldots$ are independent. Then $\mathcal{T}$ is $\mathbb{P}$-trivial, i.e. for all $A \in \mathcal{T}$,

$$\mathbb{P}(A) = 0 \text{ or } 1.$$

Proof. The idea of is to show that any $A \in \mathcal{T}$ is independent of itself. For then we'd have $\mathbb{P}(A) = \mathbb{P}(A)^2$, which forces $\mathbb{P}(A)$ to be 0 or 1. We do this in three steps.

Step 1: $\sigma(X_1, \ldots, X_k)$ and $\sigma(X_{k+1}, \ldots X_{k+j})$ are independent. To see this, let

$$\mathcal{A}_1 = \{\cap_{i=1}^{k} A_i : A_i \in \sigma(X_i)\},$$
$$\mathcal{A}_2 = \{\cap_{i=k+1}^{k+j} A_i : A_i \in \sigma(X_i)\}.$$

These are independent π-systems that contain $\cup_{i=1}^{k}\sigma(X_i)$ and $\cup_{i=k+1}^{k+j}\sigma(X_i)$, respectively. The claim then follows from Theorem 2.3.

Step 2: $\sigma(X_1, \ldots, X_k)$ and $\sigma(X_{k+1}, X_{k+2}, \ldots)$ are independent. Note that

$$\sigma(X_1, \ldots, X_k),$$
$$\cup_{j\geq 1}\sigma(X_{k+1}, \ldots, X_{k+j})$$

are independent π-systems, with the latter containing $\{X_i^{-1}(B) : B \in \mathcal{B}(\mathbb{R}), i \geq k+1\}$, a generating set for $\sigma(X_{k+1}, X_{k+2}, \ldots)$. The claim follows again from Theorem 2.3.

Step 3: $\sigma(X_1, X_2, \ldots)$ and $\mathcal{T}$ are independent. Since $\mathcal{T} \subset \sigma(X_{k+1}, X_{k+2}, \ldots)$ for every k, it is independent of each $\sigma(X_1, \ldots, X_k)$. Using a similar argument from before, we note that $\cup_{k\geq 1}\sigma(X_1, \ldots, X_k)$ and $\mathcal{T}$ are independent π-systems, with the former containing $\{X_i^{-1}(B) : B \in \mathcal{B}(\mathbb{R}), i \geq 1\}$, a generating set for $\sigma(X_1, X_2, \ldots)$. One last time, the claim follows from Theorem 2.3.

Now, since $T \subset \sigma(X_1, X_2, \ldots)$, it follows that any $A \in \mathcal{T}$ is independent of itself, and the theorem follows. □

Note that this theorem implies the event

$$\left\{\lim_{n\to\infty} S_n/n \text{ exists}\right\}$$

has probability 0 or 1 for iid X_i and $S_n = X_1 + \cdots + X_n$.

2.2 Weak Laws of Large Numbers

We say that random variables $(X_n)_{n\geq 1}$ are *uncorrelated* if for each $i, j \geq 1$,

$$\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\mathbb{E}[X_j].$$

In this case, it should be a familiar property from undergraduate probability that

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

It should also be a familiar fact that independence implies uncorrelated but not the other way around. To see this, let μ_1 and μ_2 be the distributions of X and Y respectively. If they are independent, then the product measure $\mu_1 \times \mu_2$ is just the joint distribution of (X, Y), since

$$\mu(A \times B) = \mathbb{P}((X, Y) \in A \times B) = \mathbb{P}(X \in A, Y \in B) = \mu_1(A)\mu_2(B),$$

and by Theorem 1.33 there is only one extension to the product space. Then by Fubini's theorem, we can write

$$\begin{aligned}\mathbb{E}[X]\mathbb{E}[Y] &= \left(\int x\mu_1(dx)\right)\left(\int y\mu_2(dy)\right) \\ &= \int\int xy\mu_1(dx)\mu_2(dy) \\ &= \int xy\mu = \mathbb{E}[XY].\end{aligned}$$

Example 2.9. Consider $(X, Y) \in \{(0, 1), (1, 0), (0, -1), (-1, 0)\}$ each with probability $1/4$. Then X and Y are uncorrelated but clearly not independent.

Theorem 2.10 (Weak law I)
Suppose (X_n) are uncorrelated, with $\mathbb{E}X_i = \mu$ and $\mathbb{E}X_i^2 \leq C < \infty$. Then

$$\frac{S_n}{n} \to \mu$$

in probability.

Proof. Let $\epsilon > 0$. By Chebyshev's inequality, we have

$$\mathbb{P}(|S_n/n - \mu| > \epsilon) < \frac{\mathrm{Var}(S_n/n)}{\epsilon^2} \leq \frac{Cn}{n^2\epsilon^2} \to 0$$

as $n \to \infty$. □

Example 2.11 (Polynomial approximation)**.** Given a continuous function $f : [0, 1] \to \mathbb{R}$, we wish to find a sequence of polynomials (f_n) which uniformly approximate f. That is, for $\epsilon > 0$ and sufficiently large n,

$$|f_n(x) - f(x)| < \epsilon$$

for $x \in [0, 1]$. Consider

$$f_n(x) = \sum_{m=0}^{n} \binom{n}{m} x^m (1 - x)^{n-m} f\left(\frac{m}{n}\right).$$

Note that we can alternately write

$$f_n(x) = \mathbb{E}\left[f\left(\frac{S_n}{n}\right)\right],$$

where $S_n \sim \text{Bin}(n, x)$. Then by the weak law of large numbers 2.10,

$$\mathbb{P}(A) := \mathbb{P}\left(\left|\frac{S_n}{n} - x\right| > \epsilon\right) \to 0$$

as $n \to \infty$. Now, since f is uniformly continuous on $[0,1]$ and also bounded, $|f| < M$, and for $\epsilon > 0$ we may pick $\delta > 0$ so that $|x - y| < \delta$ implies $|f(x) - f(y)| < \epsilon$. Then we have

$$\begin{aligned}\left|\mathbb{E}\left[f\left(\frac{S_n}{n}\right)\right] - f(p)\right| &\le \mathbb{E}\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right|\right] \\ &= \mathbb{E}\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right| \mathbf{1}_A\right] + \mathbb{E}\left[\left|f\left(\frac{S_n}{n}\right) - f(p)\right| \mathbf{1}_{A^c}\right] \\ &\le 2M\mathbb{P}(A) + \epsilon\end{aligned}$$

which, as $n \to \infty$, vanishes uniformly for $x \in [0,1]$.

2.2.1 Truncation

Now, we may wish to prove a law of large numbers without the second moment assumption. A common strategy is to truncate our random variables to have finite second moments, and then take limits. In particular, we define the *truncation* of a random variable X at level M to be

$$\overline{X} = X\mathbf{1}_{|X| \le M},$$

where the level M will sometimes be implicit in the notation or explicitly indexed, depending on the context.

First, we'll introduce a series of lemmas involving triangular arrays of random variables. In particular, suppose we have random variables $X_{n,i}$, with $1 \le i \le n$ for each n. Then we define

$$\begin{aligned}S_1 &= X_{1,1} \\ S_2 &= X_{2,1} + X_{2,2} \\ &\vdots \\ S_n &= X_{n,1} + \cdots + X_{n,n},\end{aligned}$$

where each row $\{X_{n,1}, \ldots, X_{n,n}\}$ consists of independent random variables.

Lemma 2.12 (Weak law for triangular arrays)
For a triangular array $X_{n,i}$, consider the truncations $\overline{X}_{n,i} = X_{n,i}\mathbf{1}_{|X_{n,i}| \le b_n}$, where the b_n are constants with $b_n \to \infty$ and

(i) $\sum_{i=1}^n \mathbb{P}(|X_{n,i}| > b_n) \to 0$, and
(ii) $b_n^{-2} \sum_{i=1}^n \mathbb{E}\overline{X}_{n,i}^2 \to 0$.

Let $a_n = \sum_{i=1}^n \mathbb{E}\overline{X}_{n,i}$. Then

$$\frac{S_n - a_n}{b_n} \to 0$$

in probability.

Proof. Write $\overline{S}_n = \overline{X}_{n,1} + \cdots + \overline{X}_{n,n}$. Then by Chebyshev,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\overline{S}_n - a_n}{b_n}\right| > \epsilon\right) &< \epsilon^{-2}\mathbb{E}\left|\frac{\overline{S}_n - a_n}{b_n}\right|^2 \\ &= \epsilon^{-2} b_n^{-2}\mathrm{Var}(\overline{S}_n) \\ &= \epsilon^{-2} b_n^{-2}\sum_{i=1}^n \mathrm{Var}(\overline{X}_{n,i}) \\ &\leq \epsilon^{-2} b_n^{-2}\sum_{i=1}^n \mathbb{E}\overline{X}_{n,i}^2,\end{aligned}$$

which goes to 0 by (ii). Now, to replace the $\overline{S}_n$ with S_n, note that

$$\mathbb{P}(S_n \neq \overline{S}_n) \leq \mathbb{P}(\cup_{i=1}^n \{X_{n,i} \neq \overline{X}_{n,i}\}) \leq \sum_{i=1}^n \mathbb{P}(|X_{n,i}| > b_n) \to 0.$$

Then, we have

$$\mathbb{P}\left(\left|\frac{S_n - a_n}{b_n}\right| > \epsilon\right) \leq \mathbb{P}(S_n \neq \overline{S}_n) + \mathbb{P}\left(\left|\frac{\overline{S}_n - a_n}{b_n}\right| > \epsilon\right) \to 0,$$

as desired. □

Lemma 2.13
Let (X_i) be iid with

$$x\mathbb{P}(|X_1| > x) \to 0$$

as $x \to \infty$. Let $\mu_n = \mathbb{E}[X_1 \mathbf{1}_{|X_1| \leq n}]$. Then

$$\frac{S_n}{n} - \mu_n \to 0$$

in probability.

Proof. We will use Lemma 2.12 with $X_{n,i} = X_i$ and $b_n = n$. Then (i) is immediate, since

$$\sum_{i=1}^n \mathbb{P}(|X_{n,i}| > n) = n\mathbb{P}(|X_1| > n) \to 0.$$

To check (ii), we will first show the following useful result: If $Y \geq 0$ and $p > 0$ then

$$\mathbb{E}Y^p = \int_0^\infty p y^{p-1}\mathbb{P}(Y > y)dy. \tag{6}$$

By Fubini's theorem, we have

$$\begin{aligned}\int_0^\infty p y^{p-1}\mathbb{P}(Y > y)dy &= \int_0^\infty \int_\Omega p y^{p-1}\mathbf{1}_{Y>y} d\mathbb{P}dy \\ &= \int_\Omega \int_0^Y p y^{p-1} dy d\mathbb{P} \\ &= \int_\Omega Y^p d\mathbb{P} = \mathbb{E}Y^p,\end{aligned}$$

which proves (6).

Now, consider the truncations $\overline{X}_{n,i} = X_i \mathbf{1}_{|X_i| \le n}$. By (6) we have

$$\mathbb{E}\overline{X}_{n,i}^2 = \int_0^\infty 2x\mathbb{P}(|\overline{X}_{n,i}| > x)dx \le \int_0^n 2x\mathbb{P}(|X_i| > x)dx$$

For $\epsilon > 0$, there is an N such that $2x\mathbb{P}(|X_i| > x) < \epsilon$ for $x > N$. So if we let n get large, we have

$$n^{-1}\mathbb{E}\overline{X}_{n,i}^2 \le \frac{1}{n}\int_0^n 2x\mathbb{P}(|X_i| > x)dy \to O(\epsilon).$$

But since ϵ is arbitrary, we see that

$$n^{-2}\sum_{i=1}^n \mathbb{E}\overline{X}_{n,i}^2 = n^{-1}\mathbb{E}\overline{X}_{n,1}^2 \to 0,$$

which completes the proof. □

Theorem 2.14 (Weak law II)
Let (X_i) be iid with $\mathbb{E}|X_i| < \infty$ and $\mathbb{E}X_1 = \mu$. Then

$$\frac{S_n}{n} \to \mu$$

in probability.

Proof. We may apply the dominated convergence theorem to see that

$$\mu_n = \mathbb{E}[X_1 \mathbf{1}_{|X_1| \le n}] \to \mathbb{E}X_1 = \mu.$$

A second application of dominated convergence theorem tells us that

$$x\mathbb{P}(|X_1| > x) \le \mathbb{E}[|X_1|\mathbf{1}_{|X_1|>x}] \to 0.$$

Thus, if we pick N sufficiently large so that $|\mu - \mu_n| < \epsilon/2$ for all $n > N$, we have

$$\mathbb{P}(|S_n/n - \mu| > \epsilon) \le \mathbb{P}(|S_n/n - \mu_n| > \epsilon/2).$$

Then by Lemma 2.13 this goes to zero, and the claim follows. □

Exercise 2.15. Prove Theorem 2.14 from scratch using truncation, but avoiding the full generality of triangular arrays.

Let (X_n) be iid, nonnegative random variables with $\mathbb{E}X_1 = \infty$. If we define $Y_i = X_i \mathbf{1}_{|X_i|<M}$ and $S_n' = Y_1 + \cdots + Y_n$, then by the weak law 2.14,

$$\frac{S_n'}{n} \xrightarrow{\mathbb{P}} \mathbb{E}Y_i.$$

But since $\mathbb{E}Y_i$ can be made arbitrarily large by the monotone convergence theorem, we have for any c,

$$\mathbb{P}(S_n/n > c) \to 1 \text{ as } n \to \infty.$$

So in some sense, we could write

$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} \infty.$$

Then we may wish to ask: at what rate does S_n grow? Let's examine an explicit example.

Example 2.16 (St. Petersburg paradox)**.** Suppose $X_i \sim X$ are iid random variables with

$$\mathbb{P}(X = 2^j) = \frac{1}{2^j}, \quad j \geq 1,$$

Then $\mathbb{E}X = \infty$, and $S_n/n \xrightarrow{\mathbb{P}} \infty$ in the sense mentioned above. One can think of this as a betting game where you win 2^j dollars if it takes j tosses to get a heads. Although you expect to win ∞, it's clear that one shouldn't actually pay ∞ to play this game. To find out how much we should really pay, we will use Lemma 2.12 to find a more precise growth rate of S_n.

Let $X_{n,k} = X_k$. We have to pick the b_n small while still ensuring that (i) and (ii) hold. With this in mind, first note that for integer m,

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

Then, we let $m(n) = \log_2 n + K(n)$, where $K(n) \to \infty$ is chosen so that $m(n)$ is an integer. If we let $b_n = 2^{m(n)}$, we have

$$n\mathbb{P}(X \geq b_n) = n2^{-m(n)+1} = 2^{-K(n)+1} \to 0,$$

which establishes (i). In checking (ii), recall that $\overline{X}_{n,k} = X_k \mathbf{1}_{|X_k| \leq b_n}$, so

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

and so the expression in (ii) vanishes, since

$$b_n^{-2} \sum_{k=1}^{n} \mathbb{E}\overline{X}_{n,k}^2 = 2n/b_n = 2n \cdot 2^{-m(n)} = 2^{-K(n)+1} \to 0.$$

Now, we evaluate a_n,

$$a_n = \sum_{k=1}^{n} \mathbb{E}\overline{X}_{n,k} = n \sum_{j=1}^{m(n)} 2^j 2^{-j} = nm(n) = n \log n + nK(n).$$

So it remains to pick $K(n)$. Note that

$$\frac{a_n}{b_n} = \frac{n \log n + nK(n)}{n2^{K(n)}} = \frac{\log n + K(n)}{2^{K(n)}},$$

from which we see that by taking $K(n) = \Theta(\log \log n)$ this ratio converges to 1. Then, by Lemma 2.12, we can say that

$$\frac{S_n}{n \log n} \xrightarrow{\mathbb{P}} 1.$$

So to play n time, we should only be willing to pay $n \log_2 n$ dollars, or $\log_2 n$ dollars per play. Therefore, it's worthy to note that the expectation may not always be the best, or even a good, utility function. The idea is that we, as human beings, don't have infinite time or resources, so we need to incorporate an aspect of risk-aversion. This is encapsulated by the weak law analysis, which ensures us with high probability that the growth rate is $O(n \log n)$.

2.3 Strong Law of Large Numbers

Suppose we have random variables X_i. Then, under various regularity assumptions, we've shown in the previous section that

$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} 0.$$

We now wish upgrade this to the statement

$$\frac{S_n}{n} \xrightarrow{a.s.} 0.$$

We will begin by building up to the full strong law of large numbers, in which we only have finite first moments, by developing the Borel-Cantelli lemmas. With this we will be able to obtain fourth and second moment strong laws. Then we will discuss two approaches to the strong law, namely the subsequence method, and convergence of random series.

2.3.1 Borel-Cantelli Lemmas

In what is to follow, we will need the Borel-Cantelli lemmas. Let $A_n \in \mathcal{F}$ be a sequence of event. We define the event

$$\{A_n \text{ i.o.}\} = \{\omega : \omega \text{ in } A_n \textit{ infinitely often}\} = \lim_{m\to\infty} \cup_{n=m}^{\infty} A_n = \limsup_{n\to\infty} A_n.$$

Correspondingly, we define

$$\{A_n \text{ ev.}\} = \{\omega : \omega \text{ in every } A_n \textit{ eventually}\} = \lim_{m\to\infty} \cap_{n=m}^{\infty} A_n = \liminf_{n\to\infty} A_n.$$

Here, the lim sup and lim inf is shorthand for the relationship

$$\limsup_{n\to\infty} \mathbf{1}_{A_n} = \mathbf{1}_{\limsup A_n}$$
$$\liminf_{n\to\infty} \mathbf{1}_{A_n} = \mathbf{1}_{\liminf A_n}.$$

Recall Kolmogorov's 0-1 law. Since $\{A_n \text{ i.o.}\} \in \mathcal{T}$, it must have probability either 0 or 1. Indeed, the Borel-Cantelli lemmas provide conditions to tell us whether $\mathbb{P}(A_n \text{ i.o.})$ is 0 or 1.

Theorem 2.17 (Borel-Cantelli Lemma I)
If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$.

Proof. Since $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, we know that tail sums vanish,

$$\sum_{n=N}^{\infty} \mathbb{P}(A_n) \to 0.$$

Now, because the events $B_N := \cup_{n\geq N} A_n$ are decreasing, i.e. $B_1 \supset B_2 \supset \cdots$, we have that $\mathbb{P}(B_N) \downarrow \mathbb{P}(\cap_{N\geq 1} B_N)$. By union bound, we have

$$\mathbb{P}(B_N) \leq \sum_{n=N}^{\infty} \mathbb{P}(A_n) \to 0$$

as $N \to \infty$. Thus,

$$\mathbb{P}(A_n \text{ i.o.}) = \mathbb{P}\left(\cap_{N\geq 1} B_N\right) = 0,$$

as desired. □

Theorem 2.18 (Borel-Cantelli Lemma II)
If the events A_n are independent and $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 1$.

Proof. Using the identity $1 - x \leq e^{-x}$, we have

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

Therefore $\mathbb{P}(\cup_{n=M}^{\infty} A_n) = 1$ for all M. Since $\cup_{n=M}^{\infty} A_n \downarrow \{A_n \text{ i.o.}\}$, the claim follows. □

Theorem 2.19 (Easy strong law)
Let (X_n) be iid with $\mathbb{E}X_i = \mu$ and $\mathbb{E}X_i^4 < \infty$. Then

$$\frac{S_n}{n} \xrightarrow{a.s.} \mu.$$

Proof. We may wlog assume $\mu = 0$. Then, we can expand

$$\mathbb{E}S_n^4 = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right)^4 = \mathbb{E}\sum_{1 \leq i,j,k,l \leq n} X_i X_j X_k X_l.$$

Terms of the form $\mathbb{E}[X_i^3 X_j]$, $\mathbb{E}[X_i^2 X_j X_k]$, and $\mathbb{E}[X_i X_j X_k X_l]$ are just 0 by independence. We are left with

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i^4\right] + \mathbb{E}\left[\sum_{i \neq j} X_i^2 X_j^2\right] = n\mathbb{E}[X_1^4] + n^2 \mathbb{E}[X_1^2]\mathbb{E}[X_2^2].$$

By Markov's inequality, we have

$$\mathbb{P}(|S_n/n| > \epsilon) < \epsilon^{-4}\mathbb{E}(S_n/n)^4 = O(\epsilon^{-4} n^{-2}).$$

Then by the Borel-Cantelli lemma, we see that $\mathbb{P}(|S_n/n| > \epsilon \text{ i.o.}) = 0$. Since ϵ is arbitrary, this implies almost sure convergence. □

Exercise 2.20. Make an attempt to reproduce the above proof but only under second or third moment assumptions.

Now, finite fourth moments are a rather strong assumption, and in practice may often be hard to compute. It turns out that under only second moment assumptions, we can use the naive Markov bound method to obtain almost sure convergence for subsequences. We can then attempt to control the oscillations between elements of the subsequence, and upgrade to almost sure convergence for the whole sequence.

Theorem 2.21 (2nd moment strong law)
Suppose (X_n) are uncorrelated and identically distributed random variables with $\mathbb{E}X_1 = \mu$ and $\mathbb{E}X_1^2 \leq C < \infty$. Then

$$\frac{S_n}{n} \xrightarrow{a.s.} \mu.$$

Proof. We may assume $\mu = 0$. We will prove the claim first for a subsequence. In particular, suppose $k(n) = n^2$. Then by Markov's inequality,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\frac{S_{k(n)}}{k(n)}\right| > \epsilon\right) = O\left(\sum_{n=1}^{\infty} \frac{1}{n^2}\right) < \infty.$$

Hence by the Borel-Cantelli lemma, $\mathbb{P}(|S_{k(n)}/k(n)| > \epsilon \text{ i.o.}) = 0$, which implies

$$S_{k(n)}/k(n) \xrightarrow{a.s.} 0.$$

Now, to show almost sure convergence for the whole sequence, it suffices to show that it is "well-behaved" between

$$\frac{S_{k(n)}}{k(n)} \leftrightarrow \frac{S_{k(n+1)}}{k(n+1)}.$$

Towards this end, define

$$D(n) = \sup_{k(n) \leq i < k(n+1)} |S_i - S_{k(n)}|.$$

Then it suffices to show that

$$\frac{D(n)}{k(n)} \xrightarrow{a.s.} 0. \tag{7}$$

By a union bound and Markov's inequality,

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{D(n)}{k(n)}\right| > \epsilon\right) &\leq \sum_{i=k(n)}^{k(n+1)} \mathbb{P}\left(\left|\frac{S_i - S_{k(n)}}{k(n)}\right| > \epsilon\right) \\
&\leq C \sum_{i=k(n)}^{k(n+1)} \frac{i - k(n)}{\epsilon^2 k(n)^2} \\
&\leq O\left(\frac{[k(n+1) - k(n)]^2}{k(n)^2}\right) = O(n^{-2}).
\end{aligned}$$

Then (7) follows by the Borel-Cantelli lemma. From this we have that, for $k(n) \leq i < k(n+1)$,

$$\left|\frac{S_i}{i}\right| \leq \left|\frac{S_i}{k(n)}\right| \leq \left|\frac{S_{k(n)} + D(n)}{k(n)}\right| \leq \left|\frac{S_{k(n)}}{k(n)}\right| + \left|\frac{D(n)}{k(n)}\right| \to 0$$

almost surely. □

As we will see, the strong law actually holds under assuming finite first moments. An application of the second Borel-Cantelli lemma says that this is the best we can do.

Theorem 2.22
If (X_n) are iid with $\mathbb{E}|X_1| = \infty$, then $\mathbb{P}(\lim_{n\to\infty} S_n/n \in (-\infty, \infty)) = 0$.

Proof. First, note that by (6),

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

Then the second Borel-Cantelli lemma tells us that $\mathbb{P}(|X_n| \geq n \text{ i.o.}) = 1$. Let $B = \{\lim_{n\to\infty} S_n/n \in (-\infty, \infty)\}$. Note 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 B, the first term goes to 0. Therefore on $B \cap \{|X_n| \geq n \text{ i.o.}\}$, we have

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

which contradicts convergence. Therefore $B \cap \{|X_n| \geq n \text{ i.o.}\} = \varnothing$, and since $\mathbb{P}(|X_n| \geq n \text{ i.o.}) = 1$, it follows that $P(B) = 0$. □

Before we move on to proving the strong law under first moment assumptions, we consider an extension to the second Borel-Cantelli lemma. The proof strategy is similar to that of Theorem 2.21, and will be seen again in the proof of the strong law under first moment assumptions.

Theorem 2.23

If (E_n) are pairwise uncorrelated events and $\sum_{n=1}^{\infty} \mathbb{P}(E_n) = \infty$, then as $n \to \infty$,

$$\frac{\sum_{i=1}^n \mathbf{1}_{E_i}}{\sum_{i=1}^n \mathbb{P}(E_i)} \xrightarrow{a.s.} 1.$$

Proof. Let $S_n = \sum_{i=1}^n \mathbf{1}_{E_i}$ and $p_i = \mathbb{P}(E_i)$. Then we wish to show that $S_n/\mathbb{E}S_n \to 1$ almost surely. Fixing an ϵ, let $a_n = \mathbb{P}(|S_n - \mathbb{E}S_n| > \epsilon \mathbb{E}S_n)$. Then it suffices to show that $\sum a_n < \infty$ by Borel-Cantelli. Chebyshev's inequality gives

$$\begin{aligned}
\mathbb{P}(|S_n - \mathbb{E}S_n| > \epsilon \mathbb{E}S_n) &\leq \frac{\mathrm{Var}(S_n)}{\epsilon^2(\mathbb{E}S_n)^2} \\
&= \frac{\sum_{i=1}^n p_i(1-p_i)}{\epsilon^2(\mathbb{E}S_n)^2} \\
&\leq \frac{1}{\epsilon^2 \mathbb{E}S_n}.
\end{aligned}$$

However, there's no reason to expect that this would converge. So, we turn to a subsequence. In particular, write $\mathbb{E}(S_n) = g_n$, and let $k(n)$ be the least element such that $g_{k(n)} \geq n^2$. Then $n^2 \leq g_{k(n)} \leq n^2 + 1$. This time Chebyshev's inequality gives us a summable series $\sum a_{k(n)}$, so we have that

$$\frac{S_{k(n)}}{k(n)} \xrightarrow{a.s.} 1.$$

It remains to control the intermediate terms $k(n) \leq m < k(n+1)$, between

$$\frac{S_{k(n)}}{g_{k(n)}} \leftrightarrow \frac{S_{k(n+1)}}{g_{k(n+1)}}.$$

Notice that $S_{k(n)} \leq S_m \leq S_{k(n+1)}$ since the $\mathbf{1}_{E_i}$ are nonnegative. Then we have

$$\frac{g_{k(n)}}{g_{k(n+1)}} \cdot \frac{S_{k(n)}}{g_{k(n)}} = \frac{S_{k(n)}}{g_{k(n+1)}} \leq \frac{S_m}{g_m} \leq \frac{S_{k(n+1)}}{g_{k(n)}} \leq \frac{S_{k(n+1)}}{g_{k(n+1)}} \cdot \frac{g_{k(n+1)}}{g_{k(n)}}.$$

Since $n^2 \leq g_{k(n)} \leq g_{k(n+1)} \leq (n+1)^2 + 1$, we see that the left and right ends both converge almost surely to 1. Hence by sandwiching we see that $S_n/g_n \to 1$ almost surely. □

2.3.2 Subsequence Method

Finally, we will prove our first strong law under first moment assumptions. The main idea is to first prove it for a subsequence, whose terms are not too sparse, and then control the oscillations among the intermediate terms. This technique was already seen in our proof of the strong law 2.21 under second moment assumptions as well as in Theorem 2.23. Indeed, outside of some technical details, the proof will be much the same idea once we truncate.

Theorem 2.24 (Strong law of large numbers)
Let (X_n) be pairwise independent, identically distributed random variables with $\mathbb{E}|X_1| < \infty$ and $\mathbb{E}X_1 = \mu$. Then

$$\frac{S_n}{n} \xrightarrow{a.s.} \mu.$$

Proof. Without loss of generality, we will assume $X_i \geq 0$. Our first step is to truncate,

$$\overline{X}_i := X_i \mathbf{1}_{|X_i| \leq i}.$$

Claim: It suffices to prove $\overline{S}_n/n \to \mu$ almost surely.

Proof of claim. Note that $\overline{X}_i = X_i$ for all large enough i because

$$\sum_{i=1}^{\infty} \mathbb{P}(\overline{X}_i \neq X_i) = \sum_{i=1}^{\infty} \mathbb{P}(|X_i| > i) \leq \int_0^{\infty} \mathbb{P}(|X_1| > t)dt = \mathbb{E}|X_1| < \infty$$

Hence by Borel-Cantelli $\mathbb{P}(\overline{X}_i \neq X_i \text{ i.o.}) = 0$. Then almost surely, we have a bound

$$|\overline{S}_n(\omega) - S_n(\omega)| \leq R(\omega) < \infty$$

for all n, which vanishes in the limit. Thus, if $\overline{S}_n/n \to \mu$ almost surely, then $S_n/n \to mu$ almost surely as well. □

As usual, we will consider the Markov bound,

$$\mathbb{P}\left(\left|\frac{\overline{S}_n - \mathbb{E}\overline{S}_n}{n}\right| > \epsilon\right) \leq \frac{\mathrm{Var}(\overline{S}_n)}{\epsilon^2 n^2}.$$

But this will not be summable over n, so we will work with the subsequence

$$k_n = \alpha^n,$$

where $\alpha > 1$ is some fixed constant to be chosen later. Then we have

$$\frac{\mathrm{Var}(\overline{S}_{k_n})}{k_n^2} \leq \frac{\sum_{i=1}^{k_n} \mathbb{E}\overline{X}_i^2}{k_n^2},$$

and by (6), we have the approximation

$$\mathbb{E}\overline{X}_i^2 \approx \sum_{j=1}^{i} j\mathbb{P}(X_i > j).$$

So our goal is to show convergence of

$$\sum_{n=1}^{\infty} \frac{\sum_{i=1}^{k_n}\sum_{j=1}^{i} j\mathbb{P}(X > j)}{k_n^2}.$$

For any particular j, the coefficient of $j\mathbb{P}(X > j)$ is at most $\mathbf{1}_{k_n \geq j}/k_n$, so the sum is comparable to

$$\sum_{j=1}^{\infty}\sum_{n=1}^{\infty} \frac{j\mathbb{P}(X > j)}{k_n}\mathbf{1}_{k_n \geq j}$$

We can bound the inside terms,

$$\begin{aligned}\sum_{n=1}^{\infty} \frac{1}{k_n}\mathbf{1}_{k_n \geq j} &= \sum_{n=1}^{\infty} \frac{1}{\alpha^n}\mathbf{1}_{\alpha_n > j} \\ &\approx \frac{1}{j}.\end{aligned}$$

Therefore, the whole thing is comparable to $\sum_{j=1}^{\infty} \mathbb{P}(X > j) \approx \mathbb{E}X < \infty$. Then by the usual argument, we have

$$\frac{\mathbb{E}\overline{S}_{k_n}}{k_n} \to \mathbb{E}X$$

almost surely. Now, since $\overline{X}_i \geq 0$, we have monotonicity of $\overline{S}_m$, so we may do the sandwiching,

$$\frac{k_n}{k_{n+1}} \cdot \frac{\overline{S}_{k_n}}{k_n} \leq \frac{\overline{S}_m}{m} \leq \frac{\overline{S}_{k_{n+1}}}{k_{n+1}} \cdot \frac{k_{n+1}}{k_n}.$$

This time, we have a fudge factor of α, which we will now leverage. By selecting α arbitrarily close to 1, we see that the lim sup and lim inf are both equal to $\mathbb{E}X$, which completes the sandwiching. And so we are done. □

By repeating the truncation technique, we get the following, which says that the strong law holds whenever $\mathbb{E}X$ exists.

Corollary 2.25
Let (X_n) be iid with $\mathbb{E}X_i^+ = \infty$ and $\mathbb{E}X_i^- < \infty$. Then $S_n/n \to \infty$ almost surely.

Proof. Let $M > 0$ and consider the truncations $X_i^{(M)} = X_i \wedge M$. Then $\mathbb{E}|X_i^{(M)}| < \infty$, so if $S_n^{(M)}) = \sum_{i=1}^{n} X_i^{(M)}$ then Theorem 2.24 implies $S_n^{(M)}/n \to \mathbb{E}X_i^{(M)}$ almost surely. Since $X_i \geq X_i^{(M)}$, we have

$$\liminf_{n\to\infty} \frac{S_n}{n} \geq \lim_{n\to\infty} \frac{S_n^{(M)}}{n} = \mathbb{E}X_i^{(M)}.$$

Then, an application of the monotone convergence theorem tells us that $\mathbb{E}(X_i^{(M)})^+ \uparrow \mathbb{E}X_i^+$, so in particular $\mathbb{E}X_i^{(M)} \uparrow \infty$, which implies $S_n/n \to \infty$ almost surely. □

2.3.3 Convergence of Random Series

We go over a second approach to the strong law which involves convergence of random series. This has the added benefit of allowing us to determine rates of convergence.

Theorem 2.26 (Kolmogorov's maximal inequality)
Let (X_n) be independent with $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 < \infty$. Then we have the following bound on the total oscillation,

$$\mathbb{P}\left(\max_{1\le k\le n} |S_k| \ge \epsilon\right) \le \epsilon^{-2}\mathrm{Var}(S_n).$$

Proof. Define T_k to be the event where k is the smallest index such that $|S_k| \ge \epsilon$, and let $A_n = \cup_{k=1}^n T_k$. Then

$$\mathrm{Var}(S_n) = \mathbb{E}S_n^2 \ge \mathbb{E}[S_n^2 \mathbf{1}_{A_n}] = \mathbb{E}[S_n^2 \sum_{k=1}^n \mathbf{1}_{T_k}].$$

Each term becomes

$$\begin{aligned}
\mathbb{E}[S_n^2\mathbf{1}_{T_k}] &= \mathbb{E}\left[[S_k + (S_n - S_k)]^2\mathbf{1}_{T_k}\right] \\
&= \mathbb{E}\left[S_k^2\mathbf{1}_{T_k}\right] + \mathbb{E}\left[(S_n - S_k)^2\mathbf{1}_{T_k}\right] + \mathbb{E}\left[S_k(S_n - S_k)\mathbf{1}_{T_k}\right] \\
&\ge \mathbb{E}[S_k^2\mathbf{1}_{T_k}] \\
&\ge \epsilon^2\mathbb{P}(T_k),
\end{aligned}$$

where the third term on the second line vanishes since $S_n - S_k \in \sigma(X_{k+1}, \dots, X_n)$ is independent from $S_k\mathbf{1}_{T_k} \in \sigma(X_1, \dots, X_k)$. Since $\{\max_{1\le k\le n}|S_k| \ge \epsilon\}$ is just a disjoint union of the T_k, we sum over all k to obtain

$$\mathrm{Var}(S_n) \ge \epsilon^2 \sum_{k=1}^n \mathbb{P}(T_k) = \epsilon^2\mathbb{P}\left(\max_{1\le k\le n}|S_k| \ge \epsilon\right),$$

which proves the claim. □

Theorem 2.27
If (X_n) are independent with $\mathbb{E}X_i = 0$ and $\sum_{i=1}^\infty \mathrm{Var}(X_i) < \infty$, then

$$\sum_{k=1}^\infty X_k$$

converges almost surely.

Proof. We will show that almost surely, the (S_n) forms a Cauchy sequence. From the maximal inequality 2.26, we get

$$\mathbb{P}\left(\sup_{n\le k\le m}|S_k - S_n| > \epsilon\right) \le \epsilon^{-2}\sum_{k=n}^m \mathrm{Var}(X_k).$$

By continuity from below when taking $m \to \infty$, we have

$$\mathbb{P}(A_n) := \mathbb{P}\left(\sup_{n\le k}|S_k - S_n| > \epsilon\right) \le \epsilon^{-2}\sum_{k=n}^\infty \mathrm{Var}(X_k) \to 0.$$

Let $B_n = \{\sup_{k_1,k_2 \geq n} |S_{k_1} - S_{k_2}| \leq 2\epsilon\}$, and note that $B_n \geq 1 - \mathbb{P}(A_n) \to 1$. Since ϵ is arbitrary we see that S_n is Cauchy with probability one. Thus (S_n) converges almost surely. □

Corollary 2.28
Suppose (X_n) are iid with $\mathbb{E}X_i = 0$ and $\mathbb{E}|X_i| < \infty$. Then

$$\sum_{k=1}^{\infty} \frac{X_k}{k}$$

converges almost surely.

Proof. To apply the previous result, we consider the truncations $\overline{X}_k = X_k \mathbf{1}_{|X_k| \leq k}$. From the proof of Theorem 2.24 we have that

$$\sum_{k=1}^{\infty} \frac{\mathbb{E}\overline{X}_k^2}{k^2} < \infty.$$

Then the variance series converges, and we have by Theorem 2.27 that

$$\sum_{k=1}^{\infty} \frac{\overline{X}_k}{k}$$

converges almost surely. Recall the claim from the proof of the strong law 2.24, in which we showed that $X_k \neq \overline{X}_k$ only finitely often. The same argument shows that

$$\sum_{k=1}^{\infty} \frac{\overline{X}_k}{k} \text{ converges } \iff \sum_{k=1}^{\infty} \frac{X_k}{k} \text{ converges,}$$

which finished the proof. □

Lemma 2.29 (Kronecker's lemma)
Suppose $a_n \uparrow \infty$ and y_n are real sequences such that

$$\sum_{k=1}^{n} \frac{y_k}{a_k}$$

converges. Then

$$a_n^{-1} \sum_{k=1}^{n} y_k \to 0.$$

Proof. Let $b_n = \sum_{k=1}^n y_k/a_k$, and suppose $b_n \to b$. Then $y_k = (b_k - b_{k-1})a_k$, so we may write

$$\begin{aligned} a_n^{-1} \sum_{k=1}^n y_k &= a_n^{-1} \sum_{k=1}^n (b_k - b_{k-1}) a_k \\ &= a_n^{-1} \left[b_n a_n - \sum_{k=1}^{n-1} b_k (a_k - a_{k-1}) \right] \\ &= b_n - a_n^{-1} \sum_{k=1}^{n-1} b_k (a_k - a_{k-1}). \end{aligned}$$

Note that the second term at the end is a weighted average of the b_n. Intuitively, since $b_n \to b$, the difference should then go to 0. □

Exercise 2.30. Complete the proof above using standard real analysis techniques.

Theorem 2.31 (Strong law of large numbers)
Let (X_n) be iid with $\mathbb{E}|X_i| < \infty$ and $\mathbb{E}X_i = \mu$. Then

$$\frac{S_n}{n} \xrightarrow{a.s.} \mu.$$

Proof. Without loss of generality, we may assume $\mu = 0$. Then the result follows from Corollary 2.28 and Kronecker's lemma. □

But we can say more using our results on random series. The following result says that S_n is more than sublinear, it is $o(n^{1/2}(\log n)^{1/2+\epsilon})$.

Theorem 2.32 (Sharper strong law)
Let (X_n) be iid with $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = \sigma^2 < \infty$. Then for any $\epsilon > 0$,

$$\frac{\sum_{i=1}^n X_i}{n^{1/2}(\log n)^{1/2+\epsilon}} \xrightarrow{a.s.} 0.$$

Proof. For $a_n = \sqrt{n}(\log n)^{1/2+\epsilon}$, note that by an integral comparison test,

$$\sum_{k=1}^{\infty} \mathrm{Var}(X_k/a_k) < \infty.$$

Then we apply Theorem 2.27 along with Kronecker's lemma to get the desired result. □

Theorem 2.33
Let (X_n) be iid with $\mathbb{E}X_i = 0$ and $\mathbb{E}|X_i|^p < \infty$ for some $p \in (1, 2)$. Then

$$\frac{S_n}{n^{1/p}} \xrightarrow{a.s.} 0.$$

Proof. Consider the truncations $\overline{X}_k = X_k \mathbf{1}_{|X_k| \le k^{1/p}}$. We have

$$\begin{aligned}\sum_{k=1}^{\infty} \mathbb{P}(|X_k| > k^{1/p}) &= \sum_{k=1}^{\infty} \mathbb{P}(|X_k|^p > k) \\ &= \sum_{k=1}^{\infty} \mathbb{P}(|X_1|^p > k) \\ &\le \mathbb{E}|X_1|^p < \infty.\end{aligned}$$

Then Borel-Cantelli implies $\mathbb{P}(X_k \neq \overline{X}_k \text{ i.o.}) = 0$. Now, it suffices to check that

$$\sum_{k=1}^{\infty} \frac{\operatorname{Var}(\overline{X}_k)}{k^{2/p}} < \infty.$$

Exercise. Complete the proof. □

2.3.4 An Application of the Strong Law

Let (X_n) be iid with distribution F. We define the *empirical distribution functions* F_n given by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \le x}.$$

In words, $F_n(x)$ is just the observed frequency of values that are $\le x$. Note that F_n is a random variable on $\sigma(X_1, \ldots, X_n)$.

Theorem 2.34 (Glivenko-Cantelli lemma)
Almost surely, we have

$$\sup_y |F_n(y) - F(y)| \to 0.$$

Proof. Fix $\epsilon > 0$. We divide the real line into intervals depending on F in the following way. Let

$$\begin{aligned} z_0 &= -\infty \\ z_1 &= \inf\{y : F(y) \ge \epsilon\} \\ &\vdots \\ z_n &= \inf\{y : F(y) \ge n\epsilon\} \\ &\vdots \\ z_{1/\epsilon} &= \infty. \end{aligned}$$

Then it suffices to show that for any i,

$$|F_n(x) - F(x)| \to 0$$

uniformly for $x \in [z_i, z_{i+1})$. Let $W_j = \mathbf{1}_{X_j \le z_i}$ and $W_j' = \mathbf{1}_{X_j < z_{i+1}}$. Then $\mathbb{E}W_j = F(z_i)$ and $\mathbb{E}W_j' = F(z_{i+1})$. Applying the strong law, we have

$$F_n(z_i) = \frac{1}{n} \sum_{j=1}^{n} W_j \xrightarrow{a.s.} F(z_i)$$

and similarly $F_n(z_{i+1}) \xrightarrow{a.s.} F(z_{i+1})$. So, F_n converges to F at the endpoints. By monotonicity and sandwiching, we get uniform convergence on the interval $[z_i, z_{i+1})$, and hence uniform convergence on the entire real line. $\square$

3 Central Limit Theorems

3.1 Weak Convergence

Let F_n, F be distribution functions. We say that F_n converges *weakly* to F, if $F_n(x) \to F(x)$ for all x where F is continuous at x, and write $F_n \Rightarrow F$ if this is the case. For random variables X_n, X we write

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

or that $X_n \to X$ *in distribution*, if $F_n \to F$ weakly where F_n, F are the distributions of X_n, X respectively.

Example 3.1. Let X has distribution F. Then $X + 1/n$ has distribution

$$F_n(x) = \mathbb{P}(X + 1/n \le x) = F(x - 1/n).$$

Since we just need to check continuity points of F, it follows that $X + 1/n \xrightarrow{d} X$.

Theorem 3.2
If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$.

Proof. Write

$$\mathbb{P}(X_n \le z) \le \mathbb{P}(X \le z + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon).$$

Taking $n \to \infty$ and noting that ϵ is arbitrary,

$$\limsup_{n\to\infty} \mathbb{P}(X_n \le z) \le \mathbb{P}(X \le z).$$

To show the reverse direction, we write

$$\mathbb{P}(X_n \le z) \ge \mathbb{P}(X \le z - \epsilon) - \mathbb{P}(|X_n - X| > \epsilon).$$

Playing the same game, we have

$$\liminf_{n\to\infty} \mathbb{P}(X_n \le z) \ge \mathbb{P}(X \le z).$$

The claim follows. □

Since weak convergence is defined without regards to the underlying probability spaces, one should expect the converse to be generally false. Indeed, we can just take iid Bernoullis B_n, B and $B_n \xrightarrow{d} B$ but clearly B_n does not converge in probability to B.

Theorem 3.3 (Skorokhod representation)
If $F_n \to F$ weakly, then there exists a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with random variables X_n, X on this space, such that $X_n \sim F_n$, $X \sim F$, and $X_n \xrightarrow{a.s.} X$.

Proof. Let $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}, \mathbb{P})$. Define X_n and X to be the generalized inverses of F_n and F. Namely,

$$X_n(\omega) = F_n^{-1}(\omega) = \sup\{y : F_n(y) < \omega\}$$
$$X(\omega) = F^{-1}(\omega) = \sup\{y : F(y) < \omega\}.$$

Also define the complementary inverses,

$$G_n^{-1}(\omega) = \inf\{z : F_n(z) > \omega\}$$
$$G^{-1}(\omega) = \inf\{z : F(z) > \omega\}.$$

Then, consider the set $A = \{\omega : F^{-1}(a) < G^{-1}(a)\}$. This set is countable, and hence has measure zero. Now, through some standard real analysis, we can show that $X_n(\omega) \to X(\omega)$ on A^c, and so $X_n \xrightarrow{a.s.} X$. □

Exercise 3.4. Fill in the details of the above proof. That is, check that A is countable, and that $X_n \to X$ on A^c.

Theorem 3.5
$X_n \xrightarrow{d} X$ if and only if for any bounded continuous function g,

$$\mathbb{E}g(X_n) \to \mathbb{E}g(X).$$

Proof. Suppose $X_n \xrightarrow{d} X$ and we have some continuous bounded function g. Let Y_n, Y be random variables with the same distributions as X_n, X, such that $Y_n \to Y$, virtue of the previous theorem. Then by continuity, $g(Y_n) \xrightarrow{a.s.} g(Y)$. By the bounded convergence theorem,

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

To show the converse, consider the smoothed out indicators

$$g_{z,\epsilon}(y) = \begin{cases} 1 & y \le z \\ 0 & y \ge x + \epsilon \\ \text{linear} & z \le y \le z + \epsilon \end{cases}$$

where the linear segment is defined so that $g_{z,\epsilon}$ is continuous. Then we have

$$\limsup_{n\to\infty} \mathbb{P}(X_n \le z) \le \limsup_{n\to\infty} \mathbb{E}g_{z,\epsilon}(X_n) = \mathbb{E}g_{z,\epsilon}(X) \le \mathbb{P}(X \le z + \epsilon).$$

Since ϵ is arbitrary, this gives $\limsup_{n\to\infty} \mathbb{P}(X_n \le z) \le \mathbb{P}(X \le z)$. To get the other direction, just replace z with $z - \epsilon$. Thus we conclude that $X_n \xrightarrow{d} X$. □

Virtue of Theorem 3.5, we can define the notion of convergence of measures, or sometimes called converence in law. For any topological space $(\Omega, \mathcal{F})$ and a sequence of measures μ_n, μ on $(\Omega, \mathcal{F})$, we write

$$\mu_n \xrightarrow{\text{law}} \mu$$

if for all bounded continuous $g : \Omega \to \mathbb{R}$, we have $\int g d\mu_n \to \int g d\mu$.

Theorem 3.6 (Continuous mapping theorem)
Suppose $X_n \xrightarrow{d} X$ and g is a measurable function with $C_g = \{\text{continuity points of } g\}$. If $\mathbb{P}(X \in C_g) = 1$, then $g(X_n) \xrightarrow{d} g(X)$.

Proof. By Theorem 3.3, we can pass to random variables Y_n, Y with the same distributions as X_n, X, such that $Y_n \xrightarrow{a.s.} Y$. With probability one, $Y \in C_g$, and so almost surely we have

$$g(Y_n) \to g(Y).$$

But then since almost sure implies in probability implies in distribution convergence, we have

$$g(Y_n) \xrightarrow{d} g(Y),$$

and so $g(X_n) \xrightarrow{d} g(X)$. □

Theorem 3.7 (Portmanteau's lemma)
The following are equivalent:

(i) $F_n \to F$ weakly, where F_n, F induce distributions μ_n, μ respectively.
(ii) For any open U, $\liminf \mu_n(U) \geq \mu(U)$.
(iii) For any closed V, $\limsup \mu_n(V) \leq \mu(V)$.
(iv) For any A such that $\mu(\partial A) = 0$, $\mu_n(A) \to \mu(A)$.

Proof. (i) $\implies$ (ii): First use Skorokhod's representation to get Y_n, Y with distributions F_n, F respectively. Then define

$$f_n = \mathbf{1}_{Y_n \in U}$$
$$f = \mathbf{1}_{Y \in U}.$$

Since $Y_n \xrightarrow{a.s.} Y$ and U is open, we know that $f_n \xrightarrow{a.s.} f$. Then, by Fatou's lemma, we have

$$\liminf \mu_n(U) = \liminf \int f_n \geq \int \liminf f_n = \int f = \mu(U).$$

(ii) $\iff$ (iii): Take complements. Note that we also have the converse.

(ii),(iii) $\implies$ (iv): Write $\overline{A} = A^o \cup \partial A$ and combine the inequalities from (ii) and (iii). In particular, write

$$\liminf \mu_n(A^o) \geq \mu(A^o) = \mu(A)$$
$$\limsup \mu_n(\overline{A}) \leq \mu(\overline{A}) = \mu(A),$$

and by sandwiching, we see that $\lim \mu_n(A) = \mu(A)$.

(iv) $\implies$ (i): Let $A = (-\infty, x]$ for a continuity point x of F. Then $F(x) = F(x^-)$ and so $\mu(\partial A) = \mu(\{x\}) = 0$. Therefore,

$$F_n(x) = \mu_n(x) \to \mu(x) = F(x),$$

and hence $F_n \to F$ weakly. □

Theorem 3.8 (Helly's selection theorem)
Given a sequence of distributions (F_n), there exists a subsequence n_k and a right continuous non-decreasing function F such that

$$F_{n_k} \to F$$

at all continuity points of F.

Remark. Note that F may not be a distribution function. In fact, its mass $F(\infty)$ can be seen to be anywhere between 0 and 1.

Proof. Fix an x. Since $0 \leq F_n(x) \leq 1$, we can pick a subsequence so that $F_{n_k}(x)$ converges. To extend to the real line, enumerate the rationals $\{q_n\}_{n\geq 1}$. For each rational q_n, construct a further subsequence so that the distributions converge on that subsequence for $\{q_1, \ldots, q_n\}$. Then consider the diagonal subsequence, which we denote by n_k^k. By construction $F_{n_k^k}(q)$ converges to some $G(q)$ for every $q \in \mathbb{Q}$. The function G is only defined for rationals, so consider

$$F(x) = \inf_{q>x} G(q).$$

We claim that this is the desired limit. Clearly it is non-decreasing. Note that it is right continuous since

$$\begin{aligned}\lim_{x_n \to x^+} F(x_n) &= \inf\{G(q) : q \in \mathbb{Q}, q > x_n \text{ for some } n\} \\ &= \inf\{G(q) : q \in \mathbb{Q}, q > x\} = F(x).\end{aligned}$$

Now, let x be a continuity point of F. We will use a sandwiching argument to complete the proof. Pick rationals r_1, r_2, s with $r_1 < r_2 < x < s$ so that

$$F(x) - \epsilon < F(r_1) \leq F(r_2) \leq F(x) \leq F(s) < F(x) + \epsilon.$$

Then, since $F_{n_k^k}(r_2) \to G(r_2) \geq F(r_1)$ and $F_{n_k^k}(s) \to G(s) \leq F(s)$, we have for sufficiently large k that

$$F(x) - \epsilon < F_{n_k^k}(r_2) \leq F_{n_k^k}(x) \leq F_{n_k^k}(s) < F(x) + \epsilon.$$

Therefore $F_{n_k^k}(x) \to F(x)$ for all continuity points x, as desired. □

Now we may ask: when does F has mass 1, i.e. when is no mass lost in the limit? We say that the distributions (F_n) are *tight* if for every $\epsilon > 0$ there is an M_ϵ such that for all n,

$$\limsup_{n\to\infty} 1 - F_n(M_\epsilon) + F_n(-M_\epsilon) \leq \epsilon.$$

That is, we can locate most of the mass for all the F_n to be in the interval $[-M_\epsilon, M_\epsilon]$ in the limit.

Theorem 3.9
Every subsequential limit (if it exists) of the distributions (F_n) is a probability distribution if and only if the sequence (F_n) is tight.

Proof. Let n_k be the subsequence in question. We may assume that $\pm M_\epsilon$ are continuity points of F, since the discontinuity set is countable. Then we have

$$F(\infty) - F(-\infty) \geq F(M_\epsilon) - F(-M_\epsilon) \leftarrow F_{n_k}(M_\epsilon) - F_{n_k}(-M_\epsilon) \geq 1 - \epsilon.$$

Since ϵ is arbitrary, it follows that F has mass 1.

For the converse, suppose F_n is not tight. Then there is an $\epsilon > 0$ and a subsequence n_k such that

$$F_{n_k}(k) - F_{n_k}(-k) \leq 1 - \epsilon$$

for all k. Pass to a further subsequence n_{k_j} so that $F_{n_{k_j}} \to F$ at all continuity point of F, virtue of Theorem 3.8. Let $r < 0 < s$ be continuity points of F, and write

$$\begin{aligned} F(s) - F(r) &= \lim_{j\to\infty} F_{n_{k_j}}(s) - F_{n_{k_j}}(r) \\ &\leq \limsup_{j\to\infty} F_{n_{k_j}}(k_j) - F_{n_{k_j}}(-k_j) \leq 1 - \epsilon. \end{aligned}$$

Letting $r \to -\infty$ and $s \to \infty$ we conclude that F cannot be the distribution function of a probability measure. □

Now, it turns out that weak convergence can be *metrized*, i.e. there exists a metric ρ such that $\rho(F_n, F) \to 0$ if and only if $F_n \Rightarrow F$.

Theorem 3.10 (Levy metric)
The function ρ given by

$$\rho(F, G) = \inf\{\epsilon > 0 : F(x-\epsilon) - \epsilon \leq G(x) \leq F(x+\epsilon) + \epsilon \text{ for all } x\}$$

is a metric on the space of distributions. Furthermore, it metrizes weak convergence.

Proof. First, note that $\rho(F, F) = 0$ since we can take ϵ arbitrarily close to 0, and

$$F(x-\epsilon) - \epsilon \leq F(x-\epsilon) \leq F(x) \leq F(x+\epsilon) \leq F(x+\epsilon) + \epsilon$$

holds by monotonicity for all x. Now, suppose $\rho(F, G) = 0$. Then for all x we can take $\epsilon \to 0$ and get by right continuity that

$$G(x) \leq \lim_{\epsilon\to 0} F(x+\epsilon) + \epsilon = F(x),$$

and similarly for the other side,

$$G(x) = \lim_{\epsilon\to 0} G(x+\epsilon) \geq \lim_{\epsilon\to 0} F(x) - \epsilon = F(x).$$

Thus $F \overset{d}{=} G$ if and only if $\rho(F, G) = 0$.

Next, note that $F(x-\epsilon) - \epsilon \leq G(x)$ holds for all x if and only if $F(x) \leq G(x+\epsilon) + \epsilon$ for all x. Similarly, $G(x) \leq F(x+\epsilon) + \epsilon$ holds for all x if and only if $G(x-\epsilon) - \epsilon \leq F(x)$ for all x. Therefore we see that $\rho(F, G) = \rho(G, F)$.

Finally, to show the triangle inequality, let $\epsilon_1 = \rho(F, G)$ and $\epsilon_2 = \rho(G, H)$. Then we have

$$\begin{aligned} F(x - (\epsilon_1 + \epsilon_2)) - (\epsilon_1 + \epsilon_2) &\leq G(x - \epsilon_2) - \epsilon_2 \leq H(x) \\ F(x + (\epsilon_1 + \epsilon_2)) + (\epsilon_1 + \epsilon_2) &\geq G(x + \epsilon_2) + \epsilon_2 \geq H(x), \end{aligned}$$

so we clearly have $\rho(F, H) \leq \epsilon_1 + \epsilon_2 = \rho(F, G) + \rho(G, H)$. This completes the proof that ρ is a metric.

Now, to show that ρ metrizes weak convergence, first suppose $\rho(F_n, F) < \epsilon_n \to 0$. Then

$$F(x) \leftarrow F(x - \epsilon_n) - \epsilon_n \leq F_n(x) \leq F(x + \epsilon_n) + \epsilon_n \to F(x)$$

for all continuity points of F. Therefore $F \xrightarrow{d} F$.

Conversely, suppose $F_n \xrightarrow{d} F$. Pick continuity points $\{x_1, x_2, \ldots, x_k\}$ of F such that $F(x_1) < \epsilon$, $F(x_k) > 1 - \epsilon$, and $|x_i - x_{i+1}| < \epsilon$ for $1 \leq i \leq k$. Then we can pick N large enough so that for $n \geq N$ we have $|F_n(x_i) - F(x_i)| < \epsilon$ for $1 \leq i \leq k$. This implies

$$F(x_i - \epsilon) - \epsilon \leq F(x_i) - \epsilon \leq F_n(x_i) \leq F(x_i) + \epsilon \leq F(x_i + \epsilon) + \epsilon,$$

and so for $x_i < x < x_{i+1}$ we have

$$F_n(x) \leq F_n(x_{i+1}) \leq F(x_{i+1}) + \epsilon \leq F(x + \epsilon) + \epsilon.$$

Similarly,

$$F(x - \epsilon) - \epsilon \leq F(x_i) - \epsilon \leq F_n(x_i) \leq F_n(x).$$

Now, if $x < x_1$ then

$$F_n(x) \leq F_n(x_1) \leq F(x_1) + \epsilon \leq 2\epsilon \leq F(x + 2\epsilon) + 2\epsilon,$$

and $F(x - \epsilon) - \epsilon \leq 0 \leq F_n(x)$. Similarly, for $x > x_k$, we have

$$F_n(x) \leq 1 \leq F(x) + 2\epsilon \leq F(x + 2\epsilon) + 2\epsilon,$$

and

$$F(x - 2\epsilon) - 2\epsilon \leq 1 - 2\epsilon \leq F(x_k) - \epsilon \leq F_n(x_k) \leq F_n(x).$$

Thus for all x we've shown that

$$F(x - 2\epsilon) - 2\epsilon \leq F_n(x) \leq F(x + 2\epsilon) + 2\epsilon,$$

and so $\rho(F_n, F) \leq 2\epsilon$ for $n \geq N$. In particular $\rho(F_n, F) \to 0$. □

Exercise 3.11 (Ky Fan metric)**.** Show that the function α given by

$$\alpha(X, Y) = \inf\{\epsilon \geq 0 : \mathbb{P}(|X - Y| > \epsilon) \leq \epsilon\}$$

is a metric on the space of random variables which metrizes convergence in probability. Furthermore, if $\alpha(X, Y) = \alpha$ then $\rho(F, G) \leq \alpha$ where X and Y follow distributions F and G, respectively.

Exercise 3.12. Using Theorem 3.10 and the previous exercise, show the following topological consequence: If every subsequence of X_n has a further subsequence which converges weakly to X, then $X_n \xrightarrow{d} X$.

Remark. We can use Theorems 3.9 and 3.8 to see that if (F_n) is a tight family of distributions, then every subsequence has a further subsequence which converges weakly to some probability distribution. Then, if all these subsequential limits are the same, we can use the previous exercise to see that $X_n \xrightarrow{d} X$.

3.2 Characteristic Functions

Suppose X is a random variable with distribution F. Then for any t, we define the *characteristic function*

$$\varphi(t) = \mathbb{E}e^{itX}.$$

Remark. This is just the Fourier transform of X.

Proposition 3.13

We have the following properties of φ:

(a) $\varphi(0) = 1$.
(b) $\varphi(t) = \overline{\varphi(-t)}$.
(c) $|\varphi(t)| \leq 1$.
(d) φ is uniformly continuous.

Proof. (a)-(c): These are easy exercises. (d): Write

$$\begin{aligned} |\varphi(t+h) - \varphi(t)| &= |\mathbb{E}[e^{i(t+h)X} - e^{itX}]| \\ &= |\mathbb{E}e^{itX}(e^{ihX} - 1)| \\ &\leq \mathbb{E}|e^{itX}(e^{ihX} - 1)| \\ &= \mathbb{E}|e^{ihX} - 1|, \end{aligned}$$

where the third line is due to Jensen's inequality with the convex function $f(x) = |x|$ on $\mathbb{C}$. Then note that by the bounded convergence theorem, this quantity goes to 0 uniformly in t. □

Now, suppose φ_X and φ_Y are characteristic functions of independent random variables X and Y, respectively. Then we can write

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

and so $\varphi_{X+Y} = \varphi_X \varphi_Y$.

3.2.1 Moments & Derivatives

Another property we may care about is differentiability. Suppose φ were differentiable, and that we could apply a variant of the Leibniz rule to get

$$\varphi'(t) = \mathbb{E}\left[\frac{\partial}{\partial t} e^{itX}\right] = \mathbb{E}\left[iXe^{itX}\right].$$

It turns out that this is only true under certain regularity conditions, which we now exhibit.

Theorem 3.14

If $\mathbb{E}|X| < \infty$, then φ is continuously differentiable with

$$\varphi'(t) = \mathbb{E}[iXe^{itX}].$$

Proof. Write

$$\begin{aligned}\left|\frac{\varphi(t+h)-\varphi(t)}{h}-\mathbb{E}[ixe^{itx}]\right| &= \left|\mathbb{E}\left[\frac{e^{i(t+h)X}-e^{itX}}{h}\right]-\mathbb{E}[ixe^{itX}]\right| \\ &= \left|\mathbb{E}\left[iXe^{itX}\left(\frac{e^{ihX}-1}{ihX}-1\right)\right]\right| \\ &= \left|\mathbb{E}\left[iXe^{itX}\left(\frac{\int_0^h e^{iuX}-1du}{h}\right)\right]\right| \\ &\leq \mathbb{E}\left[|iXe^{itX}|\cdot\sup_{0\leq u\leq h}|e^{iuX}-1|\right].\end{aligned}$$

Now observe that the inside terms are dominated by $2|X|$, so by the dominated convergence theorem this goes to zero as $h\to 0$. We can apply the same argument to see that

$$|\varphi'(t+h)-\varphi'(t)| = \mathbb{E}\left[iXe^{itX}\left(e^{ihX}-1\right)\right]\to 0$$

as $h\to 0$. Thus φ' is continuous. □

Corollary 3.15
If $\mathbb{E}|X|^k<\infty$, then $\varphi\in C^k$ with

$$\varphi^{(k)}(t)=\mathbb{E}\left[(iX)^k e^{itX}\right].$$

So far we've gone from moments to derivatives. It turns out that the other direction is possible too.

Lemma 3.16
If φ is twice differentiable then $\mathbb{E}X^2<\infty$.

Proof. A basic exercise in calculus implies

$$\frac{2-2\mathbb{E}[\cos(hx)]}{h^2}=\frac{2\varphi(0)-\varphi(h)-\varphi(-h)}{h^2}\to\varphi''(0),$$

as well as

$$\frac{1-\cos(\theta)}{\theta^2}\to\frac{1}{2}\text{ as }\theta\to 0.$$

So, we know that by Fatou's lemma,

$$-\varphi''(0)=\lim_{h\to 0}\mathbb{E}\left[\frac{2-2\cos(hx)}{h^2}\right]\geq\mathbb{E}X^2.$$

But once we have finite second moments, we have by Corollary 3.15 that $\varphi''(0)=\mathbb{E}[(iX)^2]=-\mathbb{E}X^2$. □

An intuitive picture to have in mind is that with more moments, we will have lighter tails, and this is equivalent to higher degrees of smoothness in the characteristic function.

Now, if $\mathbb{E}|X|^k < \infty$, we have the formula

$$\varphi^{(k)}(0) = \mathbb{E}(iX)^k,$$

which we can use to compute k-th moments. In particular, expanding φ in a Taylor series about 0 leads to

$$\varphi(t) = \sum_{k=0}^{n} \frac{\mathbb{E}(itX)^k}{k!} + o(t^n).$$

In what is to come, it will be useful to have a more precise handle on the error term, which is the content of the next result.

Lemma 3.17 (Taylor remainder bound)
Integration by parts shows

$$|\mathcal{E}_n(x)| := \left| e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!} \right| \leq \min\left(\frac{2|x|^n}{n!}, \frac{|x|^{n+1}}{(n+1)!} \right). \tag{8}$$

In particular, by taking expectations and applying Jensen's inequality, we get

$$\left| \varphi(t) - \sum_{k=0}^{n} \frac{\mathbb{E}(itX)^k}{k!} \right| \leq \mathbb{E}\left[\min\left(\frac{2|tX|^n}{n!}, \frac{|tX|^{n+1}}{(n+1)!} \right) \right]. \tag{9}$$

Proof. Integrating by parts, we have

$$\int_0^x e^{is}(x-s)^n ds = -\frac{e^{is}(x-s)^{n+1}}{n+1}\Big|_0^x + \frac{i}{n+1}\int_0^x e^{is}(x-s)^{n+1} ds,$$

so for $n = 0$, we get

$$\int_0^x e^{is} = \frac{e^{ix}-1}{i}.$$

By induction, this extends to

$$e^{ix} = 1 + ix + \frac{(ix)^2}{2!} + \cdots + \frac{(ix)^n}{n!} + \underbrace{\frac{i^{n+1}}{n!}\int_0^x e^{is}(x-s)^n ds}_{\mathcal{E}_n(x)}$$

Note that we already have the bound

$$|\mathcal{E}_n(x)| \leq \frac{|x|^{n+1}}{(n+1)!}.$$

This will be more useful for $|x|$ small. The second term we will need for $|x|$ large. Integrating by parts again gives

$$\frac{i}{n}\int_0^x e^{is}(x-s)^n ds = -\frac{x^n}{n} + \int_0^x e^{is}(x-s)^{n-1} ds.$$

Note that $x^n/n = \int_0^x (x-s)^{n-1} ds$, and so we have

$$\begin{aligned} |\mathcal{E}_n(x)| &\leq \left| \frac{i^n}{(n-1)!}\int_0^x (x-s)^{n-1}(e^{is}-1) ds \right| \\ &\leq \left| \frac{2}{(n-1)!}\int_0^x (x-s)^{n-1} ds \right| \leq \frac{2|x|^n}{n!}, \end{aligned}$$

as needed. □

3.2.2 Invertibility

Now we turn to the matter of Fourier inversion. That is, given φ, we'd like to reconstruct X.

Theorem 3.18 (Inversion formula)
Let $\varphi(t) = \int e^{itx}\mu(dx)$ where μ is a probability measure. Then for $a < b$,

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

Proof. Through Fubini, we have

$$\begin{aligned}
\underbrace{\int_{-T}^{T}\int_{a}^{b} \varphi(t)e^{-iut}dudt}_{I_T} &= \int_{-T}^{T} \varphi(t)\frac{e^{-iat}-e^{-ibt}}{it}dt \\
&= \int_{-T}^{T} \frac{e^{-iat}-e^{-ibt}}{it}\int e^{itx}d\mu(x) \\
&= \int\int_{-T}^{T} \frac{e^{-it(x-a)}-e^{-it(x-b)}}{it}dtd\mu(x) \\
&= \int 2\int_{0}^{T} \frac{\sin(t(x-a))-\sin(t(x-b))}{t}dtd\mu(x)
\end{aligned}$$

Using contour integration from complex analysis, we can show that

$$\lim_{T\to\infty}\int_0^T \frac{\sin x}{x}dx \to \frac{\pi}{2}.$$

Then by a change of variables, we get

$$\lim_{T\to\infty}\int_0^T \frac{\sin(\theta x)}{x}dx \to \operatorname{sgn}(\theta)\frac{\pi}{2}.$$

So we have

$$I_T \to 2\pi\mu(a,b) + \pi\mu(\{a,b\}),$$

which gives us what we want. □

Now we deal with the question of uniqueness. Suppose F_1 and F_2 are distributions with characterisitic functions $\varphi_1 = \varphi_2$. Then we'd like for the previous theorem to imply that $\mu_1 = \mu_2$. To do this, it suffices to show that μ_1 and μ_2 agree on all singletons. Consider the countable set

$$A = \{x : \mu_1(x) > 0 \text{ or } \mu_2(x) > 0\}.$$

Then for all $a, b \in A^c$, we have $\mu_1([a,b]) = \mu_2([a,b])$ by Theorem 3.18. Sending $a \to -\infty$, we have $F_1(b) = F_2(b)$ for all $b \in A^c$. Since A^c is dense, $F_1 = F_2$ everywhere by right continuity. Therefore our inversion formula provides an injective mapping from the characteristic functions to distribution functions.

Corollary 3.19
If $\varphi(t)$ is integrable, then μ has a bounded continuous density function given by

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

Proof. We'd like to show for all $a \leq b$ that

$$\mu([a,b]) = \int_a^b f(x)dx$$

for some continuous bounded function f. To start, we will show that there are no atomic masses. For any $a < x < b$, from our inversion formula we have the bound

$$\mu(\{x\}) \leq \mu(a,b) + \frac{1}{2}\mu(\{a,b\}) \leq \frac{b-a}{2\pi}\int |\varphi(t)|dt.$$

But by taking $|b-a| \to 0$, we see that $\mu(\{x\}) = 0$. Thus

$$\int_a^b \underbrace{\frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi(t)e^{-iut}dt}_{f(u)}\, du = \mu(a,b) = \mu([a,b]).$$

Various properties of $f(u)$ are left as an exercise. □

Exercise 3.20. For the density f in the above proof, show that:

- f is real valued.
- $f \geq 0$ almost surely.
- f is continuous.
- f is bounded.

Theorem 3.21

Let F_n be distributions with characteristic functions φ_n. Then the F_n converge weakly if and only if $\varphi_n \to \varphi$, where φ is continuous at zero.

Proof. Recall that if $F_n \Rightarrow F$, then Theorem 3.5 says that for every continuous bounded g,

$$\mathbb{E}g(X_n) \to \mathbb{E}g(X),$$

where $X_n \sim F_n$ and $X \sim F$. Since $x \mapsto e^{itx}$ is a bounded continuous function, it follows that

$$\varphi_n(t) \to \varphi(t).$$

Furthermore, since $\varphi(t)$ is the characteristic function of F, it is continuous at zero.

Conversely, suppose $\varphi_n \to \varphi$ and φ is continuous at 0. We will show that $\{F_n\}$ is a tight sequence. Consider the identity

$$\frac{1}{2u}\int_{-u}^{u}(1 - e^{itx})dt = 1 - \frac{\sin(ux)}{ux} \geq 0.$$

Integrating both sides with respect to μ_n, we can use Fubini to get

$$\begin{aligned}\frac{1}{2u}\int_{-u}^{u}(1-\varphi_n(t))dt &= \int_{-u}^{u}\left(1 - \frac{\sin(ux)}{ux}\right)d\mu_n(x) \\ &\geq \frac{1}{2}\mu_n(|x| \geq 2/u)\end{aligned}$$

By the bounded convergence theorem,

$$\frac{1}{2u}\int_{-u}^{u}(1-\varphi_n(t))dt \to \frac{1}{2u}\int_{-u}^{u}(1-\varphi(t))dt,$$

and by continuity of φ at 0, we know that $\varphi(t) \to 1$ as $t \to 0$, and so for all u sufficiently small,

$$1-\epsilon \leq \varphi(t) \leq 1+\epsilon.$$

Therefore,

$$\frac{1}{2u}\int_{-u}^{u}(1-\varphi(t))dt \leq \epsilon.$$

Thus for any $\epsilon > 0$, we can set $M_\epsilon = 2/u$ to satisfy the tightness condition.

Now, by Theorems 3.8 and 3.9 we have for every subsequence a further subsequence $F_{n_k} \Rightarrow F_\infty$. By the forward direction we proved earlier, $\varphi_{n_k} \to \varphi_\infty$. But we already know that $\varphi_n \to \varphi$, so $\varphi_\infty = \varphi$, and in particular because characteristic functions are unique, $F_\infty = F$ for every subsequential limit. So now we've shown that every subsequence of F_n admits a further subsequence which converges weakly, all to F. Passing this statement to pointwise every continuity point x of F, we conclude that $F_n \Rightarrow F$. □

Example 3.22. Let F_n be the distribution of a Gaussian $\mathcal{N}(0,n)$. One can compute

$$\varphi_n(t) = e^{-t^2 n/2},$$

and so $\varphi_n \to \mathbf{1}_{\{0\}}$, which is clearly not continuous at zero. Therefore, note that the F_n do not converge in distribution, as mass is lost in the limit.

3.3 Central Limit Theorems

In the most vanilla setting, one can prove a barebones version of the central limit theorem for iid Bernoulli random variables by analyzing binomial coefficients:

Exercise 3.23. Let $X_i \sim 2\text{Ber}(1/2) - 1$ be iid with $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = 1$. Show that

$$\frac{S_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1).$$

Theorem 3.24 (Central Limit Theorem)
Let (X_n) be iid with $\mathbb{E}X_n = \mu$ and $\mathbb{E}X_n^2 = \sigma^2 \in (0,\infty)$. Then

$$\frac{S_n - n\mu}{\sigma n^{1/2}} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

Proof. We may assume that $\mathbb{E}X_1 = 0$. Since $\varphi_Z(t) = e^{-t^2/2}$ is continuous at 0, by Theorem 3.21 it suffices to show that

$$\varphi_{S_n/\sqrt{n}}(t) \to e^{-t^2/2}. \tag{10}$$

First note by scaling and independence, we have

$$\varphi_{S_n/\sqrt{n}}(t) = \varphi_{S_n}(t/\sqrt{n}) = \left[\varphi_{X_1}(t/\sqrt{n})\right]^n.$$

By Lemma 3.17, we have the estimate

$$\left|\varphi_{X_1}(t/\sqrt{n}) - \mathbb{E}\sum_{m=0}^{2}\left(\frac{itX}{\sqrt{n}}\right)^m\right| \lesssim \mathbb{E}\min\left(\left(\frac{itX}{\sqrt{n}}\right)^2, \left(\frac{itX}{\sqrt{n}}\right)^3\right),$$

where we have tossed out some constant factors. Note that the bound can be further simplified to $o(t^2/n)$. Putting things together, we have

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

Suppose $c_n \to c \in \mathbb{C}$, then we claim that

$$\left(1 + \frac{c_n}{n}\right)^n \to e^c.$$

Note that this is just an extension of the fact in $\mathbb{R}$, and we will prove this in the next two lemmas. With this in hand, we can conclude (10), as well as the proof of the theorem. □

Lemma 3.25
Suppose we have complex numbers $z_1, \dots, z_n$ and $w_1, \dots, w_n$ with $|z_k|, |w_k| \le \theta$ for all k. Then

$$\left|\prod_{k=1}^n z_k - \prod_{k=1}^n w_k\right| \le \theta^{n-1}\sum_{k=1}^n |z_k - w_k|$$

Proof. We write

$$\left|\prod_{k=1}^{n} z_k - \prod_{k=1}^{n} w_k\right| \leq \left|z_1 \prod_{k=2}^{n} z_k - z_1 \prod_{k=2}^{n} w_k\right| + \left|z_1 \prod_{k=2}^{n} w_k - w_1 \prod_{k=2}^{n} w_k\right|$$
$$\leq \theta \left|\prod_{k=2}^{n} z_m - \prod_{k=2}^{n} w_k\right| + \theta^{n-1}|z_1 - w_1|.$$

Proceeding by induction, the desired inequality follows. □

Lemma 3.26
If $c_n \to c \in \mathbb{C}$, then

$$\left(1 + \frac{c_n}{n}\right)^n \to e^c.$$

Proof. Note that it suffices to show

$$\left|\left(1 + \frac{c_n}{n}\right)^n - e^{c_n}\right| \to 0.$$

Using the facts that $e^x \geq 1 + x$ for $x \in \mathbb{R}$ and $|e^z - (1 + z)| \leq |z|^2$ for $z \in \mathbb{C}$ (*exercise*), we have by Lemma 3.25 that

$$\left|\left(1 + \frac{c_n}{n}\right)^n - (e^{c_n/n})^n\right| \leq (e^{|c_n|/n})^{n-1} \cdot n \cdot \left|\frac{c_n}{n}\right|^2 \to 0$$

as $n \to \infty$. □

Finally, we show that the central limit theorem cannot be upgraded to convergence in probability.

Proposition 3.27
Let (X_n) be iid as in Theorem 3.24. Then it cannot be the case that

$$Z_n := \frac{S_n - n\mu}{\sigma n^{1/2}} \xrightarrow{p} Z = \mathcal{N}(0, 1).$$

Proof. Without loss of generality, assume $\mu = 0$ and $\sigma = 1$. Suppose for contradiction that $Z_n \xrightarrow{p} Z$. The idea is to consider $\sqrt{2}Z_{2n} - Z_n$. On one hand, this is equal to

$$\frac{X_{n+1} + X_{n+2} + \cdots + X_{2n}}{\sqrt{n}},$$

which converges in distribution to Z. On the other hand, we know that it must converge in probability to

$$\sqrt{2}Z - Z = (\sqrt{2} - 1)Z,$$

which is clearly a contradiction. □

3.3.1 Lindeberg's CLT

Theorem 3.28 (Lindeberg-Feller CLT)
Suppose we have a triangular array X_{nk} of random variables, where for each n, $1 \le k \le n$ and $X_{n1}, \dots, X_{nn}$ are independent, and $\mathbb{E}X_{nk} = 0$. Further suppose
(i) $\sum_{k=1}^n \mathbb{E}X_{nk}^2 \to \sigma^2 > 0$.
(ii) For all $\epsilon > 0$, $\lim_{n\to\infty} \sum_{k=1}^n \mathbb{E}[|X_{nk}|^2 \mathbf{1}_{|X_{nk}|>\epsilon}] = 0$.
Then

$$S_n = X_{n1} + \cdots + X_{nn} \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Remark. To see that this is a generalization of Theorem 3.24, let $X_{nk} = X_k/\sqrt{n}$. It's easy to check that the two conditions hold, and thus we have

$$\frac{S_n}{n^{1/2}} \to \mathcal{N}(0, \sigma^2) = \sigma\mathcal{N}(0, 1)$$

Proof. It suffices to show that

$$\prod_{k=1}^n \varphi_{nk}(t) \to \exp\left(-\frac{t^2\sigma^2}{2}\right).$$

So we write

$$\begin{aligned}\left|\underbrace{\varphi_{nk}(t)}_{z_{nk}} - \underbrace{\left(1 - \frac{t^2\sigma_{nk}^2}{2}\right)}_{w_{nk}}\right| &\le \mathbb{E}[\min(|tX_{nk}|^3, 2|tX_{nk}|^2)] \\ &\le \mathbb{E}[|tX_{nk}|^3 \mathbf{1}_{|X_{nk}|<\epsilon}] + \mathbb{E}[2|tX_{nk}|^2 \mathbf{1}_{|X_{nk}|\ge\epsilon}] \\ &\le \epsilon t^3 \mathbb{E}[|X_{nk}|^2 \mathbf{1}_{|X_{nk}|<\epsilon}] + 2t^2 \mathbb{E}[|X_{nk}|^2 \mathbf{1}_{|X_{nk}|\ge\epsilon}].\end{aligned}$$

Using the assumptions (i) and (ii), we have

$$\limsup_{n\to\infty} \sum_{k=1}^n |z_{nk} - w_{nk}| \le \epsilon t^3 \sigma^2,$$

and because ϵ is arbitrary, the limit is just zero. Then, we aim to use Lemma 3.25 with $\theta = 1$. Clearly $|\varphi_{nk}(t)| \le 1$. On the other hand, write

$$\sigma_{nk}^2 \le \epsilon^2 + \mathbb{E}[X_{nk}^2 \mathbf{1}_{|X_{nk}|>\epsilon}],$$

and so we see that $\sup_k \sigma_{nk}^2 \to 0$ as $n \to 0$. Thus for sufficiently large n, we have $|1 - t^2\sigma_{nk}^2/2| \le 1$. Now, applying Lemma 3.25 gives

$$\left|\prod_{k=1}^n \varphi_{nk}(t) - \prod_{k=1}^n (1 - t^2\sigma_{nk}^2/2)\right| \to 0.$$

It remains to show that the second term converges to $\exp(-t^2\sigma^2/2)$. Using (i) and the approximation $\log(1 - x) \approx -x$ as $x \to 0$, we have

$$\log \prod_{k=1}^n (1 - t^2\sigma_{nk}^2/2) = \sum_{k=1}^n \log(1 - t^2\sigma_{nk}^2/2) \approx -\sum_{k=1}^n t^2\sigma_{nk}^2/2 \to -t^2\sigma^2/2.$$

Exponentiating, we deduce that

$$\prod_{k=1}^{n}(1-t^2\sigma_{nk}^2/2) \to \exp\left(-\frac{t^2\sigma^2}{2}\right),$$

as desired. □

Theorem 3.29 (Kolmogorov's three series theorem)
Let (X_n) be independent and $A > 0$. Define $\overline{X}_i = X_i\mathbf{1}_{|X_i|<A}$. Then $\sum_{i=1}^n X_i$ converges almost surely if and only if the following conditions are satisfied:

(i) $\sum_{i=1}^\infty \mathbb{P}(|X_i| > A) < \infty$.
(ii) $\sum_{i=1}^\infty \mathbb{E}\overline{X}_i$ converges.
(iii) $\sum_{i=1}^\infty \mathrm{Var}(\overline{X}_i) < \infty$.

Proof. We first prove the "if" direction. It suffices to show that $\sum \overline{X}_i$ converges almost surely, since we can use (i) to apply Borel-Cantelli to see that $\mathbb{P}(X_n \neq \overline{X}_n \text{ i.o.}) = 0$. By Theorem 2.27, condition (iii) says $\sum \mathrm{Var}(\overline{X}_i) < \infty$ which implies that

$$\sum_{i=1}^{\infty}(\overline{X}_i - \mathbb{E}\overline{X}_i)$$

converges almost surely. Condition (ii) then gives almost sure convergence of $\sum \overline{X}_i$.

Now, we prove the converse. By Borel-Cantelli, we immediately deduce (i). We next prove (iii). Suppose for contradiction that (iii) does not hold, and let $c_n = \sum_{i=1}^n \mathrm{Var}(\overline{X}_i) \to \infty$. Then define

$$\overline{X}_{ni} = \frac{\overline{X}_i - \mathbb{E}\overline{X}_i}{\sqrt{c_n}}.$$

It's straightforward to verify the conditions of Lindeberg-Feller, so we have

$$\frac{\sum_{i=1}^n \overline{X}_i - \sum_{i=1}^n \mathbb{E}\overline{X}_i}{\sqrt{c_n}} \xrightarrow{d} \mathcal{N}(0,1).$$

But by hypothesis, $\sum \overline{X}_i$ converges almost surely, so $\sum_{i=1}^n \overline{X}_i/\sqrt{c_n} \to 0$, which implies $\sum_{i=1}^n \mathbb{E}\overline{X}_i/\sqrt{c_n} \xrightarrow{d} \mathcal{N}(0,1)$, a contradiction since a sequence of constants cannot converge in distribution to a Gaussian.

Finally, supposing (i) and (iii) hold, we know by Theorem 2.27 that

$$\sum_{i=1}^{\infty}(\overline{X}_i - \mathbb{E}\overline{X}_i)$$

exists. Then as usual, by Borel-Cantelli $\sum_{i=1}^\infty X_i$ converging implies $\sum_{i=1}^\infty \overline{X}_i$ converges. Hence (ii) holds. □

Example 3.30. Let (X_n) be iid and symmetric about zero. Suppose $\mathbb{P}(|X_n| > x) = x^{-2}$ for $x > 1$. Then we have

$$\mathbb{E}X^2 = \int 2x\mathbb{P}(|X| > x) = 2\int \frac{1}{x} = \infty.$$

Thus, in order to apply CLT, we will need to truncate and build triangular arrays. We keep in mind that the truncation level should be chosen small while still letting us to apply the Borel-Cantelli strategy. Consider

$$X_i^n = X_i \mathbf{1}_{|X_i| \leq \sqrt{n} \log\log n} \text{ for } i \leq n.$$

We have

$$\begin{aligned}
\mathbb{E}[(X_1^n)^2] &= (1 + o(1)) \int_1^{\sqrt{n}\log\log n} 2x\mathbb{P}(|X| > n) \\
&\approx \int_1^{\sqrt{n}\log\log n} \frac{2}{x} \\
&= \log n + 2\log\log\log n.
\end{aligned}$$

Exercise: prove that we have the lower bound

$$\mathbb{E}[(X_1^n)^2] = \log n + o(\log n),$$

which implies that our upper bound is essentially tight. Now, we apply Lindeberg-Feller to the triangular array:

$$\frac{X_1^n}{\sqrt{n\log n}}, \frac{X_2^n}{\sqrt{n\log n}}, \ldots, \frac{X_n^n}{\sqrt{n\log n}}.$$

Using our sharp bound, we have

$$\sum_{i=1}^n \mathbb{E}[(X_1^n)^2] = n \cdot \frac{\log n + o(\log(n))}{n\log n} \to 1.$$

For the second condition, note that for large n,

$$\mathbb{P}\left(\frac{X_i \mathbf{1}_{|X_i| \leq \sqrt{n}\log\log n}}{\sqrt{n\log n}} > \epsilon \right) = 0.$$

Thus by Lindeberg-Feller,

$$\frac{\overline{S}_n}{\sqrt{n\log n}} \xrightarrow{d} \mathcal{N}(0,1),$$

and by Borel-Cantelli, this is true of the non-truncated variables as well.

3.3.2 Higher Dimensions

Let $(\mathcal{S}, d)$ be a nice, separable metric space. We first extend weak convergence to this context. We say that $X_n \xrightarrow{d} X$ if

$$\mathbb{E}f(X_n) \to \mathbb{E}f(X)$$

for all bounded continuous functions f.

Theorem 3.31 (Portmanteau)
Suppose X_n, X be random variables taking values in an appropriate metric space $(\mathcal{S}, d)$. Then the following are equivalent:

1. $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all bounded continuous functions, i.e. $X_n \xrightarrow{d} X$.
2. For closed K, $\limsup \mathbb{P}(X_n \in K) \le \mathbb{P}(X \in K)$.
3. For open U, $\liminf \mathbb{P}(X_n \in U) \ge \mathbb{P}(X \in U)$.
4. If A is a set with $\mathbb{P}(\partial A) = 0$, then $\mathbb{P}(X_n \in A) \to \mathbb{P}(A)$.
5. Let f is a bounded measurable function, and D_f are the discontinuity points of f. If $\mathbb{P}(X \in D_f) = 0$, then $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$.

Remark. There is an extension of Skorokhod's representation to n dimensions, but involves more care in the construction.

Proof. (i) $\implies$ (ii) $\implies$ (iii) $\implies$ (iv) are the same proofs as before. (v) $\implies$ (i) is trivial. It remains to prove (iv) $\implies$ (v). Suppose $|f| \le K$. Partition the interval $[-K, K]$ into ϵ length subintervals, denoted I_i. □

Now, consider the special case of $\mathbb{R}^d$. If $X = (X_1, \dots, X_d)$ is a random vector, then we can define its *distribution function* by $F(x) = \mathbb{P}(X \le x)$, where $\{X \le x\}$ means $\cap_{i=1}^d \{X_i \le x_i\}$. As usual, we require F to satisfy:

(i) F is monotonic nondecreasing, i.e. if $x \le y$, then $F(x) \le F(y)$, where $x \le y$ means $x_i \le y_i$ in each coordinate.
(ii) $\lim_{x \to \infty} F(x) = 1$, where $x \to \infty$ means each coordinate goes to ∞ simultaneously.
(iii) $\lim_{x_i \to \infty} F(x) = 0$ for any $1 \le i \le d$.
(iv) F is right continuous, i.e. $\lim_{y \downarrow x} F(y) = F(x)$.

However, for $d > 1$, we need the additional condition: every rectangle has positive mass.

Naturally, we may extend our definition of weak convergence and say that $F_n \Rightarrow F$ if

$$F_n(x) \to F(x)$$

for all continuity points x of F. Using similar but more delicate methods, one can show that $X_n \xrightarrow{d} X$ if and only if $F_n \Rightarrow F$.

Remark. In $d = 1$, there were at most countably many discontinuity points. For $d > 1$, this is generally not true. Consider the joint random variable (X, Y) where $X = 0$ and $Y \sim \text{Unif}[0, 1]$. Instead, this property can be modified as follows: the set of all discontinuity points lie in the union of countably many hyperplanes.

Characteristic functions can also be defined for higher dimensions. Let

$$X = (X_1, \dots, X_d).$$

Then the characteristic function $\varphi : \mathbb{R}^d \to \mathbb{C}$ is given by

$$\varphi(t_1, \dots, t_d) = \mathbb{E}e^{i \sum_{k=1}^d t_k X_k} = \mathbb{E}e^{i\langle t, X \rangle}.$$

We state the extensions to three core theorems of this section, but leave their proofs to [Dur19].

Theorem 3.32 (Inversion formula)
If $A = [a_1, b_1] \times \cdots \times [a_d, b_d]$ with $\mu(\partial A) = 0$, then

$$\mu(A) = \lim_{T\to\infty} (2\pi)^{-d} \int_{[-T,T]^d} \varphi(t) \prod_{j=1}^{d} \frac{e^{-it_j a_j} - e^{-it_j b_j}}{is} dt$$

Theorem 3.33 (Convergence theorem)

Suppose $(X_n)_{n=1}^{\infty}$ are random vectors with characteristic functions φ_n. Then $X_n \xrightarrow{d} X$ if and only if $\varphi_n(t) \to \varphi_\infty(t)$ for all $t \in \mathbb{R}^d$.

Theorem 3.34 (Central limit theorem)
Let (X_n) be iid random vectors with $\mathbb{E}X_n = \mu \in \mathbb{R}^d$ and finite covariances,

$$\Sigma_{ij} = \mathbb{E}[(X_{n,i} - \mu_i)(X_{n,j} - \mu_j)] < \infty.$$

Then

$$\frac{S_n - n\mu}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

3.4 Poisson Limit Theorem

Theorem 3.35
Consider a triangular array of independent random variables X_{nk} for $1 \le k \le n$.

Proof. Through a straightforward computation, □

4 Martingales

4.1 Radon-Nikodym Derivative

Given an unsigned measure m, we can construct more unsigned measures by defining

$$m_f(E) := \int \mathbf{1}_E f dm,$$

where f is some nonnegative m-measurable function. Furthermore, it is not hard to show using the monotone convergence theorem that for any nonnegative g,

$$\int g dm_f = \int g f dm.$$

We may express this relationship shorthand as $dm_f = f dm$. We say that m_f is "differentiable" with respect to m and call f the *Radon-Nikodym derivative* of m_f with respect to m, writing

$$f = \frac{dm_f}{dm}.$$

Exercise 4.1 (Uniqueness)**.** Let m be σ-finite. For two nonnegative functions f, g show that $m_f = m_g$ if and only if $f = g$ for m-almost every x.

Exercise 4.2 (Relationship to classical derivative)**.** Let m be the Lebesgue measure on $\mathbb{R}$, and μ an unsigned measure that is differentiable with respect to m. If the Radon-Nikodym derivative $\frac{d\mu}{dm}$ is continuous, show that the map $x \mapsto \mu((-\infty, x])$ is differentiable and $\frac{d}{dx}\mu((-\infty, x]) = \frac{d\mu}{dm}(x)$ for all x. In probabilistic terms, $\frac{d\mu}{dm}$ is the probability density of the distribution μ with respect to m.

Now, consider a *signed measure* $\mu : \mathcal{F} \to [-\infty, \infty]$, satisfying

1. $\mu(\varnothing) = 0$.
2. μ cannot take both values $-\infty$ or ∞.
3. If $A = \sqcup_i E_i$, then $\sum_i \mu(E_i) = \mu(A)$, with the sum being absolutely convergence (and hence rearrangeable) if the series is finite.

We say f is μ-measurable if

$$\int f^- < \infty, \quad \int f^+ = \infty.$$

Then $\nu(A) = \int_A f d\mu$ is a signed measure.

Theorem 4.3 (Radon-Nikodym theorem)

Suppose m is an unsigned σ-finite measure, and μ is a signed σ-finite measure. Then the following are equivalent:

1. μ is *absolutely continuous* with respect to m, i.e. $\mu(E) = 0$ whenever $m(E) = 0$, and we write $\mu \ll m$.
2. There exists $f \geq 0$ such that $\mu = m_f$.

Proof. A long winded measure-theoretic detour, see [Dur19, A.4]. □

4.2 Conditional Expectation

Consider a random variable X on a probability space $(\Omega, \mathcal{F}_0, \mathbb{P})$, with $\mathbb{E}|X| < \infty$. For a σ-algebra $\mathcal{F} \subset \mathcal{F}_0$, we define the *conditional expectation* by $\mathbb{E}[X|\mathcal{F}]$ to be a $\mathcal{F}$-measurable function satisfying

$$\int_S \mathbb{E}[X|\mathcal{F}]d\mathbb{P} = \int_S X d\mathbb{P}$$

for all $S \in \mathcal{F}$. Intuitively, $\mathbb{E}[X|\mathcal{F}]$ is the maximum information about X that is observable given the "field of information" $\mathcal{F}$.

Lemma 4.4
If $\mathbb{E}|X| < \infty$, then

$$\int |\mathbb{E}[X|\mathcal{F}]| < \infty.$$

Proof. Define $A = \{\mathbb{E}[X|\mathcal{F}] > 0\}$. Then we have

$$\begin{aligned}
\int |\mathbb{E}[X|\mathcal{F}]| &= \int_A \mathbb{E}[X|\mathcal{F}] - \int_{A^c} \mathbb{E}[X|\mathcal{F}] \\
&= \int_A X - \int_{A^c} X \\
&\leq \int_A |X| + \int_{A^c} |X| \\
&= \int |X| < \infty.
\end{aligned}$$

□

Theorem 4.5 (Existence)
Conditional expectation, as defined, exists.

Proof. First assume $X \geq 0$. Define a measure μ' on $(\Omega, \mathcal{F})$ with

$$\mu'(E) = \int_E X d\mu_1,$$

since $E \in \mathcal{F}_0 \subset \mathcal{F}$. It's easy to check $\mu' \ll \mu_1$, and so by the Radon-Nikodym theorem, there exists a $g \in m\mathcal{F}$ such that

$$\mu'(E) = \int_E g d\mu_1.$$

Then we define $g = \mathbb{E}[X|\mathcal{F}]$, which we know is unique almost surely. □

Proposition 4.6

We have the following properties of conditional expectation:

(i) If $B \in \mathcal{F}$, and $X_1 = X_2$ on B, then $\mathbb{E}[X_1|\mathcal{F}] \overset{a.s.}{=} \mathbb{E}[X_2|\mathcal{F}]$ on B.

(ii) If $X, Y \in \mathcal{F}_0$, then
$$\mathbb{E}[aX + bY|\mathcal{F}] \overset{a.s.}{=} a\mathbb{E}[X|\mathcal{F}] + b\mathbb{E}[Y|\mathcal{F}].$$

(iii) If $X \leq Y$, then $\mathbb{E}[X|\mathcal{F}] \leq \mathbb{E}[Y|\mathcal{F}]$.

Exercise 4.7. Check that the properties above hold.

Exercise 4.8. State and prove versions of Fatou's lemma, monotone convergence theorem, and dominated convergence theorem for conditional expectation.

Theorem 4.9 (Jensen's inequality for conditional expectation)

Let X be a random variable and φ a convex function with $\mathbb{E}|X|, \mathbb{E}|\varphi(X)| < \infty$. Then
$$\mathbb{E}[\varphi(X)|\mathcal{G}] \geq \varphi(\mathbb{E}[X|\mathcal{G}]).$$

Proof. Let $l(x) = ax + b$ be an arbitrary affine map bounded by φ. For all $A \in \mathcal{G}$, we have by Jensen's inequality that
$$\int_A \mathbb{E}[\varphi(X)|\mathcal{G}] = \int_A \varphi(X) \geq \int_A (aX + b) = \int_A (a\mathbb{E}[X|\mathcal{G}] + b)$$
Therefore $\mathbb{E}[\varphi(X)|\mathcal{G}] \geq a\mathbb{E}[X|\mathcal{G}] + b$, and taking supremum over all such affine l, it follows that $\mathbb{E}[\varphi(X)|\mathcal{G}] \geq \varphi(\mathbb{E}[X|\mathcal{G}])$. □

Theorem 4.10 (Tower property)

Suppose we have σ-algebras $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{F}$. Then
$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1] = \mathbb{E}[X|\mathcal{G}_1].$$

Remark. The intuition is as follows: if one takes a sequence of coarser σ-algebras, then the one that determines the conditional expectation is the finest one.

Proof. For any $A \in \mathcal{G}_1 \subset \mathcal{G}_2$, we have
$$\int_A \mathbb{E}[\mathbb{E}[X|\mathcal{G}_2]|\mathcal{G}_1] = \int_A \mathbb{E}[X|\mathcal{G}_2] = \int_A X = \int_A \mathbb{E}[X|\mathcal{G}_1]$$
Then, since $\mathbb{E}[X|\mathcal{G}_1]$ is $\mathcal{G}_2$ measurable, the desired conclusion follows by uniqueness of conditional expectation. □

Lemma 4.11

Suppose $\mathcal{G} \subset \mathcal{F}$, and X is a random variable independent of $\mathcal{G}$, i.e. $\mathcal{G}$ and $\sigma(X)$ are independent σ-algebras. Then
$$\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X].$$

Proof. For all $A \in \mathcal{G}$, we have

$$\int_A \mathbb{E}[X|\mathcal{G}] = \int_A X = \int X \mathbf{1}_A \overset{\text{fubini}}{=} \mathbb{E}[X]\mathbb{P}(A) = \int_A \mathbb{E}[X],$$

and thus by uniqueness we have $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$ almost surely. □

Lemma 4.12
If $Y \in \mathcal{G}$, and $\mathbb{E}|X|, \mathbb{E}|XY| < \infty$, then

$$\mathbb{E}[XY|\mathcal{G}] = Y\mathbb{E}[X|\mathcal{G}].$$

Proof. First suppose $Y = \mathbf{1}_B$, with $B \in \mathcal{G}$. Then for $A \in \mathcal{G}$,

$$\int_A \mathbf{1}_B \mathbb{E}[X|\mathcal{G}] = \int_{A\cap B} \mathbb{E}[X|\mathcal{G}] = \int_{A\cap B} X = \int_A \mathbf{1}_B X.$$

To extend to general cases, do the usual by starting first with simple functions, then nonnegative functions by the monotone convergence theorem, and finally general integrable functions by splitting into positive and negative parts. □

4.2.1 Geometry of Conditional Expectations

For now, assume $X \in L^2$. Let $\mathcal{G} \subset \mathcal{F}$, and let Y be a $\mathcal{G}$ measurable random variable in L^2. First, note that $\mathbb{E}[X|\mathcal{G}] \in L^2$, since by Jensen's inequality we have

$$\mathbb{E}[X^2|\mathcal{G}] \geq (\mathbb{E}[X|\mathcal{G}])^2,$$

and so the claim follows by taking expectations. So all our variables live in L^2. In particular, the conditional expectation can be seen as an L^2 projection of X onto the space spanned by $\mathcal{G}$:

Lemma 4.13
Let Y be any $\mathcal{G}$-measureable random variable. Then the conditional expectation is the minimizer of the mean square error,

$$\mathbb{E}(X - Y)^2 \geq \mathbb{E}(X - \mathbb{E}[X|\mathcal{G}])^2,$$

and furthermore it is the orthogonal projection of X onto $\mathcal{G}$,

$$\mathbb{E}[(X - \mathbb{E}[X|\mathcal{G}])Y] = 0.$$

Proof. Write

$$\begin{aligned}\mathbb{E}(X - Y)^2 &= \mathbb{E}(X - \mathbb{E}[X|\mathcal{G}] + \mathbb{E}[X|\mathcal{G}] - Y)^2 \\ &= \mathbb{E}(X - \mathbb{E}[X|\mathcal{G}])^2 + \mathbb{E}(\mathbb{E}[X|\mathcal{G}] - Y)^2 + \mathbb{E}[2(X - \mathbb{E}[X|\mathcal{G}])(\mathbb{E}[X|\mathcal{G}] - Y)].\end{aligned}$$

Then, using a previous lemma, note that we can compute the cross term,

$$\begin{aligned}\mathbb{E}[2(X - \mathbb{E}[X|\mathcal{G}])(\mathbb{E}[X|\mathcal{G}] - Y)] &= \mathbb{E}\left[\mathbb{E}[2(X - \mathbb{E}[X|\mathcal{G}])\underbrace{(\mathbb{E}[X|\mathcal{G}] - Y)}_{m\mathcal{G}}|\mathcal{G}]\right] \\ &= (\mathbb{E}[X|\mathcal{G}] - Y) \cdot \underbrace{\mathbb{E}[2(X - \mathbb{E}[X|\mathcal{G}])|\mathcal{G}]}_{=0} = 0.\end{aligned}$$

Orthorgonality follows in a line:

$$\mathbb{E}[(X - \mathbb{E}[X|\mathcal{G}])Y] = \mathbb{E}[XY] - \mathbb{E}[\mathbb{E}[XY|\mathcal{G}]] = 0,$$

where we used the fact that $Y\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[XY|\mathcal{G}]$ since Y is $\mathcal{G}$-measurable. □

4.2.2 Regular Conditional Probabilities*

We define regular conditional probabilities. Denote $\mathbb{P}(A|\mathcal{G}) := \mathbb{E}[\mathbf{1}_A|\mathcal{G}]$. Also, for any $A_1, A_2, \ldots$ disjoint, we can write

$$\mathbb{P}(\sqcup_{i=1}^{\infty} A_i|\mathcal{G}) = \sum_{i=1}^{\infty} \mathbb{P}(A_i|\mathcal{G}).$$

This follows from applying the monotone convergence theorem to the function $\sum_{i=1}^{n} \mathbf{1}_{A_i}$.

Remark. Caveat: It is natural to desire for a family of sets, that $\mathbb{P}(A|\mathcal{G})$ is "regular" for all $A \in \mathcal{F}$. That is, there are potential issues to ensure that the equation above holds for all families of sets.

So the definition of regular conditional probability that we want is as follows. Consider the setting

$$(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R})).$$

Let $f : \Omega \times \mathcal{B}(\mathbb{R}) \to [0, 1]$. Then f is a *regular conditional probability* if it satisfies the following properties:

(a) For all $A \in \mathcal{B}(\mathbb{R})$, the function $f(\cdot, A) : \Omega \to [0, 1]$ is a version of $\mathbb{P}(A|\mathcal{G})$.

(b) The function $f(\omega, \cdot)$ is almost surely a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Remark. The key thing that makes the construction possible is the existence of a countably dense set in the codomain space.

4.3 Martingales

Suppose we have a stochastic process $(X_n, n \in \mathbb{N})$, each belonging to $(\Omega, \mathcal{F}, \mathbb{P})$, satisfying $\mathbb{E}|X_n| < \infty$. A *filtration* is a sequence of σ-algebras,

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}.$$

Since $X_n \in L^1$, we can talk about $\mathbb{E}[X_n|\mathcal{F}_m]$ for $m \leq n$. We will assume $X_n \in m\mathcal{F}_n$. For $m > n$, $\mathbb{E}[X_n|\mathcal{F}_m] = X_n$. We say that (X_n) is *predictable* if $X_n \in m\mathcal{F}_{n-1}$.

Suppose $(X_n, n \in \mathbb{N})$ is adapted to $(\mathcal{F}_n)$. Then we say that (X_n) is a *martingale* if

$$\mathbb{E}[X_n|\mathcal{F}_{n-1}] = X_{n-1}$$

for all n.

Example 4.14. Let $Z_1, Z_2, \ldots$ be iid Ber(± 1), and define $X_n = \sum_{i=1}^n Z_i$, with $\mathcal{F}_n = \sigma(Z_1, \ldots, Z_n)$. Then we can check

$$\begin{aligned}\mathbb{E}[X_n|\mathcal{F}_{n-1}] &= \mathbb{E}[Z_n + X_{n-1}|\mathcal{F}_{n-1}] \\ &= X_{n-1} + \underbrace{\mathbb{E}[Z_n|\mathcal{F}_{n-1}]}_{=\mathbb{E}[Z_n]=0},\end{aligned}$$

and so (X_n) is a martingale.

Example 4.15. Consider (Z_n) and (X_n) from the previous example. We show that (X_n) is a submartingale:

$$\begin{aligned}\mathbb{E}[X_n^2|\mathcal{F}_{n-1}] &= \mathbb{E}[(X_{n-1} + Z_n)^2|\mathcal{F}_{n-1}] \\ &= \mathbb{E}[X_{n-1}^2|\mathcal{F}_{n-1}] + \mathbb{E}[Z_n^2|\mathcal{F}_{n-1}] + 2\mathbb{E}[X_n Z_n|\mathcal{F}_{n-1}] \\ &= X_{n-1}^2 + 1 + 0 \geq X_{n-1}^2.\end{aligned}$$

Exercise 4.16. Keeping the notation from the above examples, check that $X_n^2 - n$ is a martingale.

We now discuss a way to generate new martingales from old. The method uses predictable processes. Suppose we have a predictable process H_i and X_n is a martingale. Then consider

$$M_n := \sum_{i=1}^n H_i(X_i - X_{i-1}).$$

Note that for $H_i = 1$, we get the original martingale back. We claim that for general (H_i), the process (M_n) is a martingale, that is adapted to the same filtration as (X_n). We say that (M_n) is the *martingale transform* of (X_n) with respect to (H_n), denoted $M_n = H_n \cdot X_n$.

Corollary 4.17
If (X_n) is a submartingale, and $H_n \geq 0$, then the process (M_n) is also a submartingale. A similar statement is true for supermartingales.

Proof. We write

$$\mathbb{E}[M_n|\mathcal{F}_{n-1}] = H_n(\mathbb{E}[X_n|\mathcal{F}_{n-1}] - X_{n-1}) + M_{n-1} \geq M_{n-1},$$

where the first term is nonnegative by the assumptions. □

We define a *stopping time* $\tau : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{N} \cup \{\infty\}$ to be a random variable such that the event $\{\tau \leq n\}$ is $\mathcal{F}_n$-measurable.

Exercise 4.18. Consider the symmetric random walk, $S_n = \sum_{i=1}^n X_i$, where X_i are iid Ber(± 1). Check whether each of the following are examples of stopping times or not:

- (First hitting time) $\tau_1 = \min\{n > 0 : S_n = 0\}$.
- (Last hitting time) $\tau_\infty = \max\{n > 0 : S_n = 0\}$.

Lemma 4.19

Let τ be stopping time and X_n be a martingale. Then the process

$$Z_n := X_{n\wedge\tau}$$

is a martingale.

Proof. The strategy is to pick a predictable sequence (H_n) such that (Z_n) is the martingale transform of X_n. In particular, pick $H_n = \mathbf{1}_{\tau \geq n}$. Since $\{\tau \leq n-1\}$ is $\mathcal{F}_{n-1}$ measurable, the H_n's are predictable. Then we know that the transform,

$$\sum_{i=1}^n H_i(X_i - X_{i-1}) = X_{n\wedge\tau} - X_0,$$

is a martingale. So since $X_0 \in m\mathcal{F}_0$, it follows that $X_{n\wedge\tau}$ is a martingale. □

Exercise 4.20. Adapt the result above for submartingales & supermartingales.

Lemma 4.21

If X_n is a submartingale and $\tau_1 \leq \tau_2$ almost surely for stopping times $\tau_1 \& \tau_2$, then

$$\mathbb{E}[X_{n\wedge\tau_1}] \leq \mathbb{E}[X_{n\wedge\tau_2}].$$

Proof. First, we show that $\mathbb{E}[X_{n\wedge\tau_1}] \leq \mathbb{E}[X_n]$. Let $M_n = X_{n\wedge\tau} - X_0 = H_n \cdot X_n$. Consider $Z_n = (1 - H_n) \cdot X_n$. This is also a submartingale. We have

$$Z_n + M_n = X_n - X_0,$$

and taking expectations...

Now, define $Y_n = X_{n\wedge\tau_2}$. We know that Y_n is a submartingale, and so by what we've shown above, we have

$$\mathbb{E}[Y_{n\wedge\tau_1}] \leq \mathbb{E}[Y_n] = \mathbb{E}[X_{n\wedge\tau_2}].$$

But the left hand side is just $\mathbb{E}[X_{n\wedge\tau_1\wedge\tau_2}] = \mathbb{E}[X_{n\wedge\tau_1}]$, which is what we wanted to show. □

Remark. The same argument applied with supermartingales shows that if X_n is a martingale, then $\mathbb{E}[X_{n\wedge\tau}] = \mathbb{E}[X_n] = \mathbb{E}[X_0]$ for all n and stopping times τ. However, it is not generally true that $\mathbb{E}[X_\tau] = \mathbb{E}[X_0]$, since we could have $\tau = \infty$. But if we're given that $\tau < \infty$ almost surely, then $X_{n\wedge\tau} \to X_\tau$ as $n \to \infty$. Hence we could apply a convergence theorem to see that

$$\mathbb{E}[X_\tau] = \lim_{n\to\infty} \mathbb{E}[X_{n\wedge\tau}] = \lim_{n\to\infty} \mathbb{E}[X_0] = \mathbb{E}[X_0].$$

Now, given a convex function φ and a martingale (X_n), the process $\varphi(X_n)$ is a submartingale. Indeed, by conditional Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}[\varphi(X_n)|\mathcal{F}_{n-1}] &\geq \varphi(\mathbb{E}[X_n|\mathcal{F}_{n-1}]) \\ &= \varphi(X_{n-1}).\end{aligned}$$

Exercise 4.22. Adapt the above result for the case where (X_n) is a submartingale.

Theorem 4.23 (Doob's maximal inequality)
Consider the convex function $\varphi(x) = x^+ = \max(x, 0)$. Suppose (X_n) is a submartingale. Then for any $a > 0$,

$$\mathbb{P}\left(\max_{0\leq k\leq n} X_k \geq a\right) \leq \frac{\mathbb{E}[X_n^+]}{a}.$$

Proof. Define the stopping time

$$\tau = \{\inf_{0\leq k\leq n} X_k \geq a\},$$

and let $E = \{\max_{0\leq k\leq n} X_k \geq a\}$. Then $X_\tau \geq a$ on E, and so

$$a\mathbb{P}(E) \leq \mathbb{E}[X_\tau \mathbf{1}_E] \leq \mathbb{E}[X_n \mathbf{1}_E].$$

Since $X_n = X_\tau$ on E^c, it follows that

$$a\mathbb{P}(E) \leq \mathbb{E}[X_n] \leq \mathbb{E}[X_n^+],$$

from which the claim follows. □

Remark. Recall Kolmogorov's maximal inequality 2.26. This is a corollary of Doob's inequality by considering the submartingale $M_n = S_n^2$, which gives

$$\begin{aligned}\mathbb{P}\left(\max_{k\leq n} |S_k| \geq x\right) &= \mathbb{P}\left(\max_{k\leq n} |S_k|^2 \geq x^2\right) \\ &\leq \frac{\mathbb{E}|S_n|^2}{x^2} = \frac{\mathrm{Var}(S_n)}{x^2}.\end{aligned}$$

4.3.1 Martingale Convergence

Theorem 4.24 (Upcrossing lemma)
Let (X_n) be a submartingale, and pick an interval $[a, b]$, with $a < b$. Define (where n is even):

$$\begin{aligned}
\tau_0 &= \inf\{t : X_t \le a\} \\
\tau_1 &= \inf\{t \ge \tau_0 : X_t \ge b\} \\
\tau_2 &= \inf\{t \ge \tau_1 : X_t \le a\} \\
&\vdots \\
\tau_n &= \inf\{t \ge \tau_{n-1} : X_t \le a\} \\
\tau_{n+1} &= \inf\{t \ge \tau_n : X_t \ge b\} \\
&\vdots
\end{aligned}$$

If $N(a, b, n)$ is the number of crossings up to time n. Then

$$\mathbb{E}N(a, b, n) \le \frac{\mathbb{E}(X_n - a)^+ - \mathbb{E}(X_0 - a)^+}{b - a}.$$

Proof. The strategy is to come up with a predictable process H_n which acts as an indicator for the segments $(\tau_n, \tau_{n+1}]$. In particular, define

$$H(i) = \mathbf{1}_{\tau_0 < i \le \tau_1} + \mathbf{1}_{\tau_2 < i \le \tau_3} + \cdots.$$

Then consider the process

$$\overline{X}_n = (X_n - a)^+ + a,$$

which we know is a submartingale since $(x - a)^+ + a$ is convex and increasing. Therefore the transforms $Y = H \cdot \overline{X}$ and $Z = (1 - H) \cdot \overline{X}$ are submartingales. Then we have

$$\begin{aligned}
\mathbb{E}[Y_n - Y_0] &\ge 0 \\
\mathbb{E}[Z_n - Z_0] &\ge 0,
\end{aligned}$$

and

$$Y_n - Y_0 \ge (b - a)N(a, b, n) + \underbrace{(\overline{X}_n - \overline{X}_{\tau^*})}_{\ge 0}.$$

Together, this implies

$$\begin{aligned}
(b - a)\mathbb{E}[N(a, b, n)] &\le \mathbb{E}[Y_n - Y_0] \\
&\le \mathbb{E}[\overline{X}_n - \overline{X}_0] \\
&= \mathbb{E}[(X_n - a)^+ - (X_0 - a)^+],
\end{aligned}$$

which is what we wanted. □

Theorem 4.25 (Martingale convergence theorem)
Let X_n be a submartingale such that $\sup_n \mathbb{E}X_n^+ < \infty$. Then X_n converges almost surely to a random variable X with $\mathbb{E}|X| < \infty$.

Remark. The intuition is that if X_n does not converge, then it will oscillate infinitely often between some thresholds a and b. This is where we can apply the upcrossing lemma.

Proof. Fix $a < b$. Then by the upcrossing lemma we have

$$\mathbb{E}[N(a,b,n)] \leq \frac{\mathbb{E}[(X_n - a)^+]}{b-a} \leq \frac{\mathbb{E}[X_n^+ + |a|]}{b-a}.$$

Sending $n \to \infty$, we have $\mathbb{E}[N(a,b,\infty)] < \infty$. Thus $N(a,b,\infty) < \infty$ almost surely. Now, choosing $a < b \in \mathbb{Q}$ one can write

$$\mathbb{P}\left(\cap_{a<b\in\mathbb{Q}}\{N(a,b,\infty) < \infty\}\right) = 1.$$

Suppose for contradiction that $\limsup X_n > \liminf X_n$. Then we can find rationals $q_1 < q_2$ sandwiched:

$$\liminf X_n < q_1 < q_2 < \limsup X_n$$

and so X_n will oscillate infinitely often between q_1 and q_2, contradiction. Hence $X_n \to X$ almost surely.

It remains to show $\mathbb{E}|X| < \infty$. Note that since $X_n \to X$, we have $X_n^+ \to X^+$ almost surely. Then by Fatou's lemma, we have

$$\mathbb{E}X^+ \leq \liminf \mathbb{E}X_n^+ < \infty.$$

Similarly $\mathbb{E}X^- < \infty$, and we conclude $\mathbb{E}|X| < \infty$. □

Corollary 4.26
If $X_n \geq 0$ is a supermartingale, then $X_n \xrightarrow{a.s.} X$ with $\mathbb{E}|X| < \infty$.

Proof. Consider the submartingale $-X_n$ and apply the martingale convergence theorem. □

Theorem 4.27 (Bounded increments)
Suppose X_n is a martingale with *bounded increments*, i.e.

$$|X_n - X_{n-1}| < K$$

for all n. Define

$$A = \{X_n \text{ converges}\}$$
$$B = \{\limsup X_n = \infty, \liminf X_n = -\infty\}.$$

Then $\mathbb{P}(A \cup B) = 1$.

Proof. Consider the stopping time $\tau = \inf\{t : X_t \geq M\}$. Then $Y_n = X_{\tau\wedge n}$ is a martingale. By bounded increments we know that $Y_n^+ \leq M + K$, and so $\mathbb{E}Y_n^+ \leq M + K$ for all n. Hence we may apply the martingale convergence theorem to see that Y_n converges almost surely. Then X_n converges almost surely on the event $\tau = \infty$. Then we have

$$\mathbb{P}\left(\cap_{M=1}^{\infty}\{Y_n^{(M)} \text{ converges}\}\right) = 1,$$

which implies X_n converges on the event

$$\cup_{M=1}^{\infty}\{\tau_M = \infty\} = \{\limsup X_n < \infty\}.$$

Now, since X_n is a martingale, we can consider $-X_n$ to see that X_n converges on the event $\{\liminf X_n > -\infty\}$. Thus X_n converges or else we must have B. □

Example 4.28 (Absence of bounded increments)**.** Consider the random variables

$$Z_n = \begin{cases} 2^n & \text{w.p. } 2^{-n} \\ -(1-2^{-n})^{-1} & \text{w.p. } 1-2^{-n} \end{cases}$$

Then $X_n = \sum_{i=1}^{n} Z_i$ is a martingale. By Borel-Cantelli, $Z_n = 2^n$ for only finitely many n, and so $Z_n = -(1-2^{-n})^{-1} \approx -1$. Therefore $\lim X_n = -\infty$ almost surely.

4.3.2 Lp Inequalities

Theorem 4.29
Suppose X_n is a submartingale with $\mathbb{E}|X_n|^p < \infty$ for all n. Then for $p > 1$, we have

$$\mathbb{E}\left[\left(\max_{0\le k\le n} X_k^+\right)^p\right] \le \left(\frac{p}{p-1}\right)^p \mathbb{E}(X_n^+)^p.$$

Proof. Let $Y_n = \max_{0\le k\le n} X_k^+$. We will need to also truncate, so assume without loss of generality that Y_n are bounded random variables. By (6), Doob's inequality, Fubini, and Holder's inequality, we have

$$\begin{aligned}
\mathbb{E}Y_n^p &= \int p\lambda^{p-1}\mathbb{P}(Y_n > \lambda) \\
&\le \int_0^\infty p\lambda^{p-1}\left(\lambda^{-1}\int X_n^+ \mathbf{1}_{Y_n\ge\lambda} dP\right) d\lambda \\
&= \int X_n^+ \int_0^{Y_n} p\lambda^{p-2} d\lambda dP \\
&= \frac{p}{p-1}\int X_n^+ Y_n^{p-1} dP \\
&\le \frac{p}{1-p}(\mathbb{E}|X_n^+|^p)^{1/p}(\mathbb{E}|Y_n|^p)^{1/q}.
\end{aligned}$$

Dividing both sides by $\mathbb{E}|Y_n|^{p/q}$ and raising both sides to the p-th power, we obtain the claim. For non-truncated random variables applying the monotone convergence theorem suffices. □

Theorem 4.30 (L^p convergence)
Let X_n be a martingale with $\sup_n \mathbb{E}|X_n|^p < \infty$. Then there exists a random variable X such that $X_n \to X$ a.s. and in L^p.

Proof. By Holder's inequality,

$$\mathbb{E}|X_n| \le (\mathbb{E}|X_n|^p)^{1/p},$$

and so we may apply the martingale convergence theorem to see that there exists an X such that $X_n \xrightarrow{a.s.} X$. Using Doob's L^p-inequality on $|X_k|$, we have

$$\mathbb{E}\left[\max_{k\le n}|X_k|^p\right] \le \left(\frac{p}{p-1}\right)^p \mathbb{E}|X_n|^p.$$

By sending $n \to \infty$, it follows that

$$\mathbb{E}\left[\sup_{n\ge 1}|X_n|^p\right] < \infty.$$

Since $X_n \xrightarrow{a.s.} X$, we have $|X_n - X| \le 2\sup_n |X_n|$ almost surely. Thus by the bound above and dominated convergence theorem we can conclude that

$$\mathbb{E}|X_n - X|^p \to 0.$$

□

Example 4.31 (Counterexample for $p = 1$)**.** Consider the symmetric random walk $S_n = X_1 + \cdots + X_n$ started at $S_0 = 1$, where $X_i \sim \text{Ber}(\pm 1)$. Let

$$\tau = \inf\{t \ge 0 : S_t = 0\},$$

and $\overline{S}_n = S_{n\wedge\tau}$ be the associated martingale. Note that $\mathbb{E}\overline{S}_n = \mathbb{E}\overline{S}_0 = 1$. On the other hand, we know that $\overline{S}_n \xrightarrow{a.s.} S$ for some random variable S. But $|\overline{S}_{n+1} - \overline{S}_n| = 1$ unless $\overline{S}_n = 0$. Therefore $\overline{S}_n \xrightarrow{a.s.} 0$. But then $\overline{S}_n$ cannot converge to 0 in L^1, for then we'd get the contradiction $\mathbb{E}\overline{S}_n \to 0$.

It can be shown that

$$\mathbb{P}\left(\max_{i\le n}\overline{S}_i > m\right) \approx \frac{1}{m},$$

and so

$$\mathbb{E}[\max \overline{S}_i] \approx \sum_{m=1}^{\infty}\frac{1}{m} = \infty.$$

So no such maximal inequality can exist.

Let X_n be a submartingale adapted to $\mathcal{F}_n$. Then there exists a unique predictable, nonnegative process A_n with $A_0 = 0$ and

$$X_n = M_n + A_n,$$

where M_n is a martingale also adapted to $\mathcal{F}_n$. This is known as *Doob's decomposition*.

Example 4.32. Consider the submartingale $S_n = \sum_{i=1}^n Z_i$ where $Z_i \sim \text{Ber}(p)$. Then consider the modified process

$$M_n = \sum_{i=1}^n \overline{Z}_i, \quad \overline{Z}_i = Z_i - p.$$

This is a martingale, with the Doob decomposition

$$S_n = M_n + np.$$

Theorem 4.33 (Doob's decomposition)
The decomposition

$$X_n = M_n + A_n$$

exists and is unique.

Proof. A priori suppose such a decomposition exists. Then we must have

$$\mathbb{E}[X_n|\mathcal{F}_{n-1}] = \underbrace{M_{n-1}}_{X_{n-1}-A_{n-1}} + A_n,$$

which upon rearranging gives

$$\mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}] = A_n - A_{n-1} \geq 0.$$

So consider the following definition:

$$A_n := \sum_{i=1}^{n} \mathbb{E}[X_i - X_{i-1}|\mathcal{F}_{i-1}].$$

Since each term is $\mathcal{F}_{n-1}$-measurable, we know A_n is predictable. By reversing the construction, it's easy to see that $M_n = X_n - A_n$ is a martingale. □

4.3.3 L1 Theory & Uniformly Integrable Martingales

We say that a family of random variables $\{X_i\}_{i\in I}$ is *uniformly integrable* (U.I.) if given $\epsilon > 0$, there exists an M such that

$$\mathbb{E}\left[|X_i|\mathbf{1}_{|X_i|>M}\right] < \epsilon$$

for all $i \in I$. In words, this is saying that the mass of the random variables are uniformly concentrated around zero.

Exercise 4.34. Show that the collection $(X_i)_{i\in I}$ is U.I. if and only if the following conditions hold:

(i) $\sup_{i\in I} \mathbb{E}|X_i| < \infty$.
(ii) For every $\epsilon > 0$, there is a $\delta > 0$ such that for every $A \in \mathcal{F}$ with $\mathbb{P}(A) < \delta$, we have

$$\mathbb{E}[|X_i|\mathbf{1}_A] < \epsilon$$

for all $i \in I$.

Example 4.35. A trivial example of a U.I. family is a collection (X_n) that is dominated by an integrable random variable, i.e. $|X_i| \leq Y \in L^1$ for all i.

Lemma 4.36
Given a probability space $(\Omega, \mathcal{F}_0, \mathbb{P})$ and an $X \in L^1$, the family

$$\{\mathbb{E}[X|\mathcal{F}] : \mathcal{F} \subset \mathcal{F}_0 \text{ is a } \sigma\text{-algebra}\}$$

is U.I.

Proof. Let $\epsilon > 0$. Then since $X \in L^1$, there exists $\delta > 0$ such that $\mathbb{E}[|X|\mathbf{1}_A] < \epsilon$ whenever $\mathbb{P}(A) < \delta$ (*exercise*). Then by Markov's and Jensen's inequalities we have

$$\begin{aligned}\mathbb{P}(|\mathbb{E}[X|\mathcal{F}]| > M) &\leq \frac{\mathbb{E}|\mathbb{E}[X|\mathcal{F}]|}{M} \\ &\leq \frac{\mathbb{E}[|X||\mathcal{F}]}{M} = \frac{\mathbb{E}|X|}{M}\end{aligned}$$

for any σ-algebra $\mathcal{F} \subset \mathcal{F}_0$. So, by picking $M > \mathbb{E}|X|/\delta$, we have

$$\begin{aligned}\mathbb{E}[|\mathbb{E}[X|\mathcal{F}]|\mathbf{1}_{|\mathbb{E}[X|\mathcal{F}]|>M}] &\leq \mathbb{E}[\mathbb{E}[|X||\mathcal{F}]\mathbf{1}_{|\mathbb{E}[X|\mathcal{F}]|>M}] \\ &= \mathbb{E}[\mathbb{E}[|X|\mathbf{1}_{|\mathbb{E}[X|\mathcal{F}]|>M}|\mathcal{F}] \\ &= \mathbb{E}[|X|\mathbf{1}_{|\mathbb{E}[X|\mathcal{H}]|>M}] < \epsilon,\end{aligned}$$

which proves uniform integrability. □

Theorem 4.37

If X_n is a martingale then the following are equivalent:

1. X_n is U.I.
2. $X_n \xrightarrow{a.s.} X \in L^1$.
3. $X_n = \mathbb{E}[X|\mathcal{F}_n]$.

Proof. (i) $\implies$ (ii): The U.I. condition implies

$$\mathbb{E}|X_n| = \mathbb{E}[|X_n|\mathbf{1}_{|X_n|\leq M}] + \mathbb{E}[|X_n|\mathbf{1}_{|X_n|>M}] \leq M + \epsilon,$$

hence by the martingale convergence theorem there is an X for which $X_n \xrightarrow{a.s.} X$.
(ii) $\implies$ (iii): It suffices to prove

$$\int_A X_n = \int_A X$$

for all $A \in \mathcal{F}_n$. By the martingale property, we know

$$\int_A X_n = \int_A X_m \quad \text{for } m \geq n.$$

(iii) $\implies$ (i): To see that X_n is a martingale, write

$$\begin{aligned}\mathbb{E}[X_n|\mathcal{F}_{n-1}] &= \mathbb{E}[\mathbb{E}[X|\mathcal{F}_n]|\mathcal{F}_{n-1}] \\ &= \mathbb{E}[X|\mathcal{F}_{n-1}] = X_{n-1}.\end{aligned}$$

The fact that X_n is U.I. follows from Lemma 4.36. □

Corollary 4.38

If X is a random variable and we define the *Doob martingale* $X_n := \mathbb{E}[X|\mathcal{F}_n]$, then

$$X_n \xrightarrow{a.s.} Y \quad \text{and} \quad X_n \xrightarrow{L^1} Y.$$

Furthermore if $\mathcal{F}_\infty = \sigma(\cup_{n=0}^\infty \mathcal{F}_n)$, then $Y = \mathbb{E}[X|\mathcal{F}_\infty]$.

Proof. It suffices to show that for all $A \in \mathcal{F}_\infty$,

$$\int_A X = \int_A Y.$$

However, this is a bit hard to show, so we will instead show this for all $B \in \cup_{n=0}^{\infty} \mathcal{F}_n$, and then invoke Dynkin's $\pi - \lambda$ theorem. Indeed, for $m \geq n$ we can write

$$\begin{aligned} \int_B X &= \int_B X_n \\ &= \int_B X_m \to \int_B Y, \end{aligned}$$

by the previous theorem. □

4.3.4 Optional Stopping Theorems

We wish to determine conditions under which

$$\mathbb{E}X_0 = \mathbb{E}X_\tau,$$

where X_n is a martingale and τ is a stopping time.

Lemma 4.39
If X_n is U.I. and τ is a stopping time, then $\overline{X}_n = X_{n\wedge\tau}$ is also U.I.

Proof. Note that

$$\mathbb{E}\left[|\overline{X}_n|\mathbf{1}_{|\overline{X}_n|>M}\right] \leq \mathbb{E}\left[|X_n|\mathbf{1}_{|\overline{X}_n|>M}\mathbf{1}_{\tau\geq n}\right] + \mathbb{E}\left[|X_\tau|\mathbf{1}_{|\overline{X}_n|>M}\right].$$

Since $|X_{n\wedge\tau}|$ is a submartingale, $\mathbb{E}|X_{n\wedge\tau}| \leq \mathbb{E}|X_n|$. Therefore if X_n is U.I. we have

$$\sup_n \mathbb{E}|X_{n\wedge\tau}| \leq \sup_n \mathbb{E}|X_n| < \infty.$$

□

Theorem 4.40 (Optional stopping theorems)
For a martingale X_n and a stopping time τ, we have

$$\mathbb{E}X_\tau = \mathbb{E}X_0,$$

provided that any one of below are satisfied:

(i) X_n is U.I.
(ii) $\tau \leq k$ a.s.
(iii) $|X_n| \leq k$ a.s. for each n, and $\tau < \infty$ a.s.
(iv) X_n has bounded increments, i.e. for all n,

$$|X_n - X_{n-1}| \leq B \quad a.s.$$

and $\mathbb{E}\tau < \infty$.
(v) $\mathbb{E}[|X_n - X_{n-1}||\mathcal{F}_{n-1}] \leq B$ a.s. and $\mathbb{E}\tau < \infty$.

Remark. In the sub/supermartingale cases, the equality can be replaced by an inequality.

Proof of (i). By the previous lemma, we know that $X_{n\wedge\tau}$ is U.I. So together with Theorem 4.37 we can write

$$\mathbb{E}X_0 = \mathbb{E}X_{0\wedge\tau} = \mathbb{E}X_{n\wedge\tau} \to \mathbb{E}X_\tau.$$

□

Proof of (ii). Recall that $\mathbb{E}X_0 = \mathbb{E}X_{n\wedge\tau} = \mathbb{E}X_n$. Taking $n = k$ gives us what we want. □

Proof of (iii). Consider the stopping time $\tau \wedge n$. Since $\tau < \infty$ we know that $X_{\tau\wedge n} \to X_\tau$ almost surely, and so by the dominated convergence theorem,

$$\mathbb{E}X_0 = \mathbb{E}X_{\tau\wedge n} \to \mathbb{E}X_\tau.$$

□

Proof of (iv). Note that $X_{n\wedge\tau} \le k\tau \in L^1$, so by the dominated convergence theorem we have

$$\mathbb{E}X_0 = \mathbb{E}X_{n\wedge\tau} \to \mathbb{E}X_\tau.$$

□

Proof of (v). It suffices to show uniform integrability of $X_{\tau\wedge n}$. First, note that

$$|X_{\tau\wedge n}| \le |X_0| + \sum_{m=0}^{\infty} |X_{m+1} - X_m|\mathbf{1}_{\tau>m}.$$

Then by dominated convergence theorem it's enough to show that the right side has finite expectation. To see this, write

$$\begin{aligned}\mathbb{E}[|X_{m+1} - X_m|\mathbf{1}_{\tau>m}] &= \mathbb{E}[\mathbb{E}[|X_{m+1} - X_m||\mathcal{F}_m]\mathbf{1}_{\tau>m}] \\ &\le B\mathbb{P}(\tau > m),\end{aligned}$$

and so we have

$$\begin{aligned}\mathbb{E}\left[\sum_{m=0}^{\infty} |X_{m+1} - X_m|\mathbf{1}_{\tau>m}\right] &\le B\sum_{m=0}^{\infty} \mathbb{P}(\tau > m) \\ &= B\mathbb{E}N < \infty.\end{aligned}$$

□

Example 4.41 (Gambler's ruin)**.** Consider the simple random walk S_n and stopping time

$$\tau = \inf\{t : S_t = a \text{ or } b\}.$$

We clearly have bounded increments,

$$|S_n - S_{n-1}| = |X_n| \le 1,$$

and furthermore by stochastic dominance we have $\mathbb{E}\tau < \infty$. Hence by the optional stopping theorem, we know $\mathbb{E}S_0 = \mathbb{E}S_\tau$. In particular, we have

$$0 = \mathbb{E}S_0 = \mathbb{E}S_\tau = p_a \cdot a + (1 - p_a) \cdot b \implies p_a =$$

Alternatively, suppose we wish to find $\tau_{\{a,b\}}$.

Example 4.42. Let $\tau = \tau_0$ be the hitting time of 0. We start at $S_0 = 1$, and consider $S^* = \max_{i<\tau} S_i$. It turns out that $\mathbb{E}S^* = \infty$. We have

$$\mathbb{P}(S^* \geq x) = \mathbb{P}(\tau_x < \tau_0) = 1/x.$$

But then

$$\mathbb{E}S^* = \sum_{i=1}^{\infty} \mathbb{P}(S^* \geq i) = \infty.$$

4.3.5 Reverse Martingales

Consider a reversed filtration $\mathcal{F}_0 \supset \mathcal{F}_1 \supset \mathcal{F}_2 \supset \cdots$, and the process X_n adapted to this reversed filtration. Then we say X_n is a *reversed martingale* if

$$\mathbb{E}[X_i|\mathcal{F}_{i+1}] = X_{i+1}.$$

Note that a reverse martingale is nothing but the Doob martingale, since

$$\mathbb{E}[X_0|\mathcal{F}_2] = \mathbb{E}[\mathbb{E}[X_0|\mathcal{F}_1]|\mathcal{F}_2] = \mathbb{E}[X_1|\mathcal{F}_2] = X_2,$$

and so by induction we have $X_i = \mathbb{E}[X_0|\mathcal{F}_i]$ for all i.

Theorem 4.43 (Reverse martingale convergence theorem)
Suppose X_n is a reverse martingale. Then there exists an $X \in L^1$ satisfying
(i) $X_n \xrightarrow{a.s.} X$.
(ii) $X_n \xrightarrow{L^1} X$.
(iii) $X = \mathbb{E}[X_0|\mathcal{F}_\infty]$, where $\mathcal{F}_\infty = \cap_{i=0}^{\infty} \mathcal{F}_i$.

Proof. (i): Fixing n, note that

$$X_n, X_{n-1}, \ldots, X_0$$

is a forward martingale. Let $N(q_1, q_2, n)$ be the number of crossings up to time n. Then the upcrossing lemma gives us

$$\mathbb{E}[N(q_1, q_2, n)] \leq \frac{\mathbb{E}(X_0 - q_1)^+}{q_2 - q_1}.$$

Since this is a uniform bound over all n, we have shown that the number of upcrossings between any two rationals is finite almost surely. Hence, as in the proof of the martingale convergence theorem, we've shown the existence of an almost sure limit.
(ii): Almost sure together with U.I. implies L^1 convergence, so it suffices to show X_n is U.I. Since $X_i = \mathbb{E}[X_0|\mathcal{F}_i]$, the result follows from Lemma 4.36.
(iii): Note that X is $\mathcal{F}_\infty$ measurable. Then for any $B \in \mathcal{F}_\infty \subset \mathcal{F}_i$, we have

$$\int_B X_0 = \int_B X_i$$

for all i. But since $X_n \to X$ in L^1, we have

$$\int_B X_i \to \int_B X.$$

Therefore

$$\int_B \mathbb{E}[X_0|\mathcal{F}_\infty] = \int_B X$$

for all B, which implies $X = \mathbb{E}[X_0|\mathcal{F}_\infty]$. □

Corollary 4.44
If $\mathcal{F}_n \downarrow \mathcal{F}_\infty$, then for any integrable Y,

$$\mathbb{E}[Y|\mathcal{F}_n] \xrightarrow{a.s.} \mathbb{E}[Y|\mathcal{F}_\infty] \quad \text{and} \quad \mathbb{E}[Y|\mathcal{F}_n] \xrightarrow{L^1} \mathbb{E}[Y|\mathcal{F}_\infty].$$

Recall Kolmogorov's 0-1 law, which states that the tail σ-algebra is trivial. What follows is a generalization of that. In particular, we say a σ-algebra (on the sequence space generated by $X_1, X_2, \ldots$) is *permutation-invariant* if

$$\mathcal{E}_n := \mathcal{F}_n = \sigma(A : A \text{ is invariant under permutation of the first } n \text{ coordinates}).$$

In other words, a collection of sequences $(x_1, x_2, \ldots)$ is in $\mathcal{F}_n$ if and only if after permutating any of the first n coordinates the collection is still in $\mathcal{F}_n$. Importantly, note that

$$\mathcal{F}_n \supset \mathcal{F}_{n+1}$$

for all n, so we have a filtration and we can define $\mathcal{E} := \mathcal{F}_\infty = \cap_{n\geq 1}\mathcal{F}_n$ to be the *exchangeable σ-algebra*.

Exercise 4.45. Show that the event

$$\left\{\sum_{i=1}^{n} X_i > t\right\}$$

belongs to $\mathcal{F}_n$.

Before proving our 0-1 law, we will need the following general fact about conditional expectations:

Lemma 4.46
Suppose $X \in L^2$ and X is independent of $\mathcal{F}$. If $\mathbb{E}[X|\mathcal{G}]$ is $\mathcal{F}$ measurable, then $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$.

Proof. Let $Y = \mathbb{E}[X|\mathcal{G}]$. Then conditional Jensen implies $Y \in L^2$. By independence we have

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[Y]^2.$$

By the orthogonality principle of conditional expectations, we have

$$\mathbb{E}[(X - Y)Y] = 0,$$

and so $\mathbb{E}[Y^2] = 0$, and in particular $\text{Var}(Y) = 0$. Thus $\mathbb{E}[X|\mathcal{G}]$ must be degenerate, and precisely equal to it's expectation $\mathbb{E}[X]$. □

Theorem 4.47 (Hewitt-Savage 0-1 law)
Let X_n be iid on $(\Omega, \mathcal{F}, \mathbb{P})$. Then any event $A \in \mathcal{E}$ must have $\mathbb{P}(A) = 0$ or 1.

Proof. As in the proof of Kolmogorov's 0-1 law, the strategy is to show $\mathcal{E}$ is independent of itself. Fix k and some $\phi : \mathbb{R}^k \to \mathbb{R}$ bounded. We will show that

$$\phi(X_1, X_2, \ldots, X_k)$$

is independent of $\mathcal{E}$. To see that this implies $\mathcal{E}$ is independent of itself, take ϕ as indicators. This implies that the π-system

$$\cup_{k=1}^{\infty} \underbrace{\sigma(X_1, \dots, X_k)}_{\mathcal{G}_k}$$

is independent of $\mathcal{E}$. Then by Theorem 2.3, it follows that $\mathcal{E}$ is independent of $\sigma(\cup_k \mathcal{G}_k) \supset \mathcal{E}$. The first step is to symmetrize $\phi(X_1, \dots, X_k)$. Let's work with the symmetrization

$$A_n(\phi) = \frac{1}{(n)_k} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \phi(X_{i_1}, X_{i_2}, \dots, X_{i_k}),$$

where $(n)_k = n(n-1)\cdots(n-k+1)$. Note that $A_n(\phi)$ is $\mathcal{S}_n$ invariant, and so $A_n(\phi) \in m\mathcal{E}_n$. Then

$$\begin{aligned} A_n(\phi) &= \mathbb{E}[A_n(\phi)|\mathcal{E}_n] \\ &= \frac{1}{(n)_k} \sum \mathbb{E}[\phi(X_{i_1}, \dots, X_{i_k})|\mathcal{E}_n] \\ &= \mathbb{E}[\phi(X_1, \dots, X_k)|\mathcal{E}_n]. \end{aligned}$$

Since $\mathcal{E}_n$'s are decreasing, we see that $A_n(\phi)$ is a reverse martingale. Therefore by the reverse martingale convergence theorem, we know

$$A_n(\phi) \to \mathbb{E}[\phi(X_1, \dots, X_k)|\mathcal{E}].$$

Then we want to show that the limit is also $\mathbb{E}[\phi(X_1, \dots, X_k)]$. By the previous lemma, it suffices to show that $\mathbb{E}[\phi(X_1, \dots, X_k)|\mathcal{E}]$ is independent of $(X_1, \dots, X_k)$. To do this, firstnote that there are $k(n-1)_{k-1}$ terms in $A_n(\phi)$ involving X_1, so

$$\frac{1}{(n)_k} \sum_{\text{contains } X_1} \phi(X_{i_1}, \dots, X_{i_k}) \le \frac{k(n-1)_{k-1}}{(n)_k} \sup \phi \to 0.$$

This shows that

$$\mathbb{E}[\phi(X_1, \dots, X_k)|\mathcal{E}] \in m\sigma(X_2, X_3, \dots).$$

Repeating the argument for the other indices gives

$$\mathbb{E}[\phi(X_1, \dots, X_k)|\mathcal{E}] \in m\sigma(X_{k+1}, X_{k+2}, \dots).$$

And since our random variables are iid, the claim follows. □

A sequence of random variables X_n is said to be *exchangeable* if for any n and $\pi \in \mathcal{S}_n$, we have

$$(X_1, X_2, \dots, X_n) \stackrel{d}{=} (X_{\pi(1)}, X_{\pi(2)}, \dots, X_{\pi(n)}).$$

Example 4.48. The following examples are exchangeable:

- X_n iid.
- Fix a random $\theta \in \text{Unif}[0,1]$. Given θ, define our sequence to be

  $$X^{(\theta)} = (\text{Ber}(\theta), \text{iid}).$$

  Note that each X_n is marginally $\text{Ber}(1/2)$, but they are only independent when conditioned on θ. Indeed, check that

  $$\mathbb{E}[X_1 X_2] = \mathbb{E}_\theta[\mathbb{E}[X_1 X_2|\theta]] = \mathbb{E}[\theta^2] = 1/3 \neq 1/4 = \mathbb{E}[X_1]\mathbb{E}[X_2],$$

  so our X_n are positively correlated, and therefore cannot be independent.

The next result essentially says that the only types of exchangeable sequences are the ones exhibited in the example above.

Theorem 4.49 (de Finetti)
Given an exchangeable sequence X_n and the exchangeable σ-algebra Σ, the random variables $X_1, X_2, \ldots$ are iid when conditioned on Σ.

Proof. Let $f : \mathbb{R}^{k-1} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be bounded functions. Denoting $I_{n,k}$ to be the set of all sequences of distinct indices $1 \leq i_1, \ldots, i_k \leq n$, then we have

$$\begin{aligned}(n)_{k-1}A_n(f)nA_n(g) &= \sum_{i \in I_{n,k-1}} f(X_{i_1}, \ldots, X_{i_{k-1}}) \sum_m g(X_m) \\ &= \sum_{i \in I_{n,k}} f(X_{i_1}, \ldots, X_{i_{k-1}})g(X_{i_k}) \\ &\quad + \sum_{i \in I_{n,k-1}} \sum_{j=1}^{k-1} f(X_{i_1}, \ldots, X_{i_{k-1}})g(X_{i_j}).\end{aligned}$$

Using the notation from the previous theorem, let $\phi(x_1, \ldots, x_k) = f(x_1, \ldots, x_{k-1})g(x_k)$. Also let $\phi_j(x_1, \ldots, x_{k-1}) = f(x_1, \ldots, x_{k-1})g(x_j)$. Then rearranging gives

$$A_n(\phi) = \frac{n}{n-k+1} A_n(f)A_n(g) - \frac{1}{n-k+1} \sum_{j=1}^{k-1} A_n(\phi_j).$$

Applying the reverse martingale convergence theorem we have

$$\mathbb{E}[f(X_1, \ldots, X_{k-1})g(X_k)|\mathcal{E}] = \mathbb{E}[f(X_1, \ldots, X_{k-1})|\mathcal{E}]\mathbb{E}[g(X_k)|\mathcal{E}].$$

Thus, by induction we have

$$\mathbb{E}\left[\prod_{j=1}^{k} f_j(X_j)|\mathcal{E}\right] = \prod_{j=1}^{k} \mathbb{E}[f_j(X_j)|\mathcal{E}],$$

which is the desired conclusion. □

4.4 Applications to Random Walks

Suppose $X_1, X_2, \ldots$ are iid. Then we consider the random walk $S_n = S_0 + X_1 + \cdots + X_n$, where S_0 is a constant, and let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. We first derive some general results, before moving on to simple random walks.

Example 4.50 (Linear martingale)**.** If $\mathbb{E}X_i = \mu$, then $M_n = S_n - n\mu$ is a martingale.

Example 4.51 (Quadratic martingale)**.** If $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = \sigma^2 \in (0, \infty)$, then $M_n = S_n^2 - n\sigma^2$ is a martingale.

Example 4.52 (Exponential martingale)**.** Consider the moment-generating function (mgf) $\phi(\theta) = \mathbb{E}\exp(\theta X_i)$. Suppose it is finite, then $M_n = \exp(\theta S_n)/\phi(\theta)^n$ is a martingale.

Exercise 4.53. Check that each of the three examples above are in fact martingales.

Theorem 4.54 (Wald's identity)
In the setting above, suppose $S_0 = 0$, and N a stopping time such that $\mathbb{E}N < \infty$. Then we have

$$\mathbb{E}S_N = \mu\mathbb{E}N.$$

Proof. Using the linear martingale $M_n = S_n - n\mu$, we note that

$$\mathbb{E}[|M_{n+1} - M_n||\mathcal{F}_n] = \mathbb{E}|X_i - \mu|,$$

so we may apply the optional stopping theorem to get

$$\mathbb{E}[S_N - N\mu] = \mathbb{E}[S_0 - 0\mu] = 0,$$

from which the conclusion follows. □

We now focus our attention to the special case of simple random walks. That is, let

$$\mathbb{P}(X_i = 1) = p, \quad \mathbb{P}(X_i = -1) = q = 1 - p.$$

The case $p = 1/2$ was already seen in the gambler's ruin example, and so the results will be repeated here for completeness.

Theorem 4.55 (Symmetric simple random walk)
In the symmetric case of $p = 1/2$, suppose $S_0 = x$ and let $N = T_{a,b} = \min\{n : S_n \notin (a, b)\}$. Then

1. $\mathbb{P}(S_N = a) = \frac{b-x}{b-a}$ and $\mathbb{P}(S_N = b) = \frac{x-a}{b-a}$.
2. $\mathbb{E}N = (b - x)(x - a)$.

Next, under the symmetric setting with $S_0 = 0$, we consider the alternate stopping time $T_1 = \min\{n : S_n = 1\}$. First, we settle the matter of finiteness.

Lemma 4.56
We have $\mathbb{P}(T_1 < \infty) = 1$ and $\mathbb{E}T_1 = \infty$.

Proof. Since $\mathbb{E}[T_{1,-M}] = M < \infty$ by Theorem 4.55, the hitting time $T_{1,-M}$ is finite almost surely for every M. Therefore

$$\mathbb{P}(T_1 = \infty) \leq \mathbb{P}(T_{-M} < T_1) = \frac{1}{M+1} \to 0,$$

and so $\mathbb{P}(T_1 < \infty) = 1$. Next, note that for any M, we have

$$\mathbb{E}[T_1] \geq \mathbb{E}[T_{1,-M}] = M.$$

Hence $\mathbb{E}[\tau_1] = \infty$. □

Now, we wish to determine the distribution of T_1. To do this we will first compute its moment-generating function.

Theorem 4.57

Let S_n be the symmetric simple random walk with $S_0 = 0$. Then we have the moment-generating function

$$\mathbb{E}s^{T_1} = \frac{1 - \sqrt{1-s^2}}{s}.$$

Proof. Let $\theta > 0$. Note that

$$\phi(\theta) = \mathbb{E}e^{\theta X_n} = \frac{e^\theta + e^{-\theta}}{2} \geq 1.$$

Then consider $M_n = e^{\theta S_n}/\phi(\theta)^n$, the exponential martingale. Then the martingale $M_{n\wedge T_1}$ is bounded, since

$$|M_{n\wedge T_1}| = \frac{e^{\theta S_{n\wedge T_1}}}{\phi(\theta)^{n\wedge T_1}} \leq e^\theta.$$

Hence by the bounded convergence theorem we have

$$\mathbb{E}M_{n\wedge T_1} \to \mathbb{E}M_{T_1}$$

Combined with optional stopping theorem on $M_{n\wedge T_1}$, this implies $\mathbb{E}M_{T_1} = \mathbb{E}M_{0\wedge T_1} = 1$, and hence

$$e^{-\theta} = \mathbb{E}\phi(\theta)^{-T_1}.$$

To get the generating function, set

$$\phi(\theta) = \frac{e^\theta + e^{-\theta}}{2} = \frac{1}{s}.$$

Letting $x = e^\theta$, we can do some algebra and solve a quadratic to get

$$x = \frac{1 \pm \sqrt{1-s^2}}{s}.$$

We want the solution that is 0 when $s = 0$, which is the form of the mgf we are looking for. □

Exercise 4.58. Invert the generating function and find that

$$\mathbb{P}(T_1 = 2n-1) = \frac{1}{2n-1} \cdot \binom{2n}{n} 2^{-2n}.$$

Now, we move on to the asymmetric case, i.e. $p \neq q$. In contrast to before, note that $S_n - (p-q)n$ is a martingale, so optional stopping theorem does not immediately give us what we want. Instead, we will have to work with exponential martingales to compute hitting probabilities and expected hitting times.

Theorem 4.59 (Asymmetric simple random walk)
For the setting in which $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = -1) = q = 1 - p$ with $p \neq q$, the following are true:

(a) If $\varphi(y) = [(1-p)/p]^y$, then $\varphi(S_n)$ is a martingale.

(b) For $S_0 = x$ and $a < x < b$,

$$\mathbb{P}(T_a < T_b) = \frac{\varphi(b) - \varphi(x)}{\varphi(b) - \varphi(a)} \quad \text{and} \quad \mathbb{P}(T_b < T_a) = \frac{\varphi(x) - \varphi(a)}{\varphi(b) - \varphi(a)}.$$

(c) Suppose $p \in (1/2, 1)$ and $S_0 = 0$. For $a < 0$ we have

$$\mathbb{P}(\min_{n \geq 0} S_n \leq a) = \mathbb{P}(T_a < \infty) = \left(\frac{1-p}{p}\right)^{-a}.$$

On the other hand, for $b > 0$ we have $\mathbb{P}(T_b < \infty) = 1$ and $\mathbb{E}T_b = b/(p-q)$.

Proof of (a). Write

$$\begin{aligned}\mathbb{E}[\varphi(S_{n+1})|\mathcal{F}_n] &= p\left(\frac{1-p}{p}\right)^{S_n+1} + (1-p)\left(\frac{1-p}{p}\right)^{S_n-1} \\ &= (1-p+p)\left(\frac{1-p}{p}\right)^{S_n} = \varphi(S_n).\end{aligned}$$

□

Proof of (b). Since $\varphi(S_{T_{a,b} \wedge n})$ is bounded, we may apply the bounded convergence theorem along with optional stopping theorem to get

$$\varphi(x) = \mathbb{E}\varphi(S_{T_{a,b} \wedge n}) \to \mathbb{E}\varphi(S_{T_{a,b}}) = \mathbb{P}(T_a < T_b)\varphi(a) + \mathbb{P}(T_a > T_b)\varphi(b).$$

Rearranging gives the desired hitting probabilities. □

Proof of (c). Note that $T_a < \infty$ if and only if $T_a < T_b$ for some b, so by taking the limit as $b \to \infty$, we get the formula

$$\mathbb{P}(T_a < \infty) = \lim_{b \to \infty} \mathbb{P}(T_a < T_b) = \frac{1}{\varphi(a)} = \left(\frac{1-p}{p}\right)^{-a}.$$

Similarly, note that $T_b < \infty$ if and only if $T_b < T_a$ for some a, so we can take the limit as $a \to -\infty$ to get

$$\mathbb{P}(T_b < \infty) = \lim_{a \to -\infty} \mathbb{P}(T_b < T_a) = 1.$$

Finally, to compute the expected hitting time, we use the linear martingale $M_n = S_n - (p-q)n$. Since $T_b \wedge n$ is bounded, we can apply optional stopping theorem to get

$$0 = \mathbb{E}[S_{T_b \wedge n} - (p-q)(T_b \wedge n)].$$

Rearranging gives

$$\mathbb{E}[T_b \wedge n] = \frac{\mathbb{E}[S_{T_b \wedge n}]}{p-q}$$

By the monotone convergence theorem we know the left side limits to $\mathbb{E}T_b$. To deal with the right side, note that

$$\min_m S_m \leq S_{T_b \wedge n} \leq b,$$

and our estimate on $\mathbb{P}(\min_m S_m \leq a)$ implies $\mathbb{E}[\min_m S_m] > -\infty$. Hence we may apply the dominated convergence theorem and conclude that

$$\mathbb{E}[T_b] = \frac{\mathbb{E}S_{T_b}}{p-q} = \frac{b}{p-q}.$$

□

5 Markov Processes

We begin with a primer using random walks, before moving on to discuss Markov processes in generality. After developing the general theory, we will briefly introduce Poisson processes, renewal processes, and continuous time Markov chains.

5.1 Random Walks

At the end of the last chapter, we saw applications of martingales to computing the hitting times and probabilities of random walks (i.e. gambler's ruin). We now continue that discussion as a means to build a bridge to the rest of the chapter.

5.1.1 Combinatorics of Random Walks

Consider the sequence $S_0, S_1, S_2, \ldots, S_n$ as being represented by a *path* with segments

$$(k-1, S_{k-1}) \to (k, S_k).$$

For a path from $(0,0)$ to (n,x), we define $a = (n+x)/2$ to be the number of positive steps and $b = (n-x)/2$ to be the number of negative steps. So notice we have $n = a+b$ and $x = a-b$. Provided $-n \leq x \leq n$ and $n-x$ is even, then the number of paths from $(0,0)$ to (n,x) is

$$N_{n,x} = \binom{n}{a}.$$

Otherwise if $n-x$ is odd, there cannot exist any paths.

Lemma 5.1 (Reflection principle)
Let $x, y > 0$. Then the number of paths from $(0,x)$ to (n,y) that touch 0 at some time is equal to the number of paths from $(0,-x)$ to (n,y).

Proof. A quick sketch shows that the two collection of paths are in bijection via reflecting the path about 0 until the first time it hits 0. □

Theorem 5.2 (Ballot theorem)
Suppose that two candidates A and B get α and β votes, respectively, where $b < a$. Then the probability that A always strictly leads B is

$$\frac{a-b}{a+b}.$$

Proof. The reflection principle tells us that the number of paths from $(1,1)$ to (n,x) that never touch 0 is equal to the number of paths from $(1,-1)$ to (n,x). Hence the number of paths from $(1,1)$ to (n,x) which never touch 0 is given by

$$\begin{aligned} N_{n-1,x-1} - N_{n-1,x+1} &= \binom{n-1}{a-1} - \binom{n-1}{a} \\ &= \frac{a-b}{a+b} N_{n,x}. \end{aligned}$$

And so the probability can be found by dividing by $N_{n,x}$. □

Using the ballot theorem, we now aim to determine the distribution of the first hitting time of zero, for the case of simple random walks.

Lemma 5.3
Suppose S_n is a simple random walk started at $S_0 = 0$. Then

$$\mathbb{P}(S_1 \neq 0, S_2 \neq 0, \dots, S_{2n\neq 0}) = \mathbb{P}(S_{2n} = 0).$$

Proof. First, note that by symmetry,

$$\begin{aligned}\mathbb{P}(S_1 \neq 0, S_2 \neq 0, \dots, S_{2n\neq 0}) &= 2\mathbb{P}(S_1 > 0, S_2 > 0, \dots, S_{2n} > 0). \\ &= \sum_{r=1}^{\infty} \mathbb{P}(S_1 > 0, S_2 > 0, \dots, S_{2n} = 2r).\end{aligned}$$

By the ballot theorem, the terms at the end are given by

$$\frac{1}{2}(p_{2n-1,2r-1} - p_{2n-1,2r+1}),$$

where $p_{n,x} := \mathbb{P}(S_n = x)$. Summing over r we get a telescoping sum, and so

$$\mathbb{P}(S_1 \neq 0, S_2 \neq 0, \dots, S_{2n\neq 0}) = 2 \cdot \frac{1}{2} p_{2n-1,1} = \mathbb{P}(S_{2n} = 0),$$

as desired. □

Now, using Stirling's approximation, we have

$$\mathbb{P}(S_{2n} = 0) = \binom{2n}{n} \frac{1}{2^{2n}} \sim \frac{1}{\sqrt{n}}.$$

Let T be the first return time to 0. It follows that

$$\begin{aligned}\mathbb{P}(T > 2n) &= \mathbb{P}(S_1 \neq 0, S_2 \neq 0, \dots S_{2n\neq 0}) \\ &= \mathbb{P}(S_{2n} = 0) \sim \frac{1}{\sqrt{n}}.\end{aligned}$$

Similarly, we can compute

$$\begin{aligned}\mathbb{P}(T = 2n) &= \mathbb{P}(T > 2n - 2) - \mathbb{P}(T > 2n) \\ &= \mathbb{P}(S_{2n-2} = 0) - \mathbb{P}(S_{2n} = 0) \sim \frac{1}{n^{3/2}},\end{aligned}$$

which, as expected, is the derivative of the previous expression.

Next, we compute the last hitting time of 0, i.e. let

$$L_{2n} = \sup\{m \leq 2n : S_m = 0\}.$$

Lemma 5.4
Let $u_{2m} = \mathbb{P}(S_{2m} = 0)$. Then

$$\mathbb{P}(L_{2n} = 2k) = u_{2k} u_{2n-2k}.$$

Proof. Note that

$$\mathbb{P}(L_{2n} = 2k) = \mathbb{P}(S_{2k=0}, S_{2k+1} \neq 0, \ldots, S_{2n} \neq 0),$$

and we can view the random walk as being restarted at time $2k$, and so the result follows from the previous lemma. □

With the discrete density in hand, we now derive the limiting density. Such a result is known as an arcsine law.

Theorem 5.5 (Arcsine law for last hitting time)
For $0 < a < b < 1$,

$$\mathbb{P}(a \leq L_{2n}/2n \leq b) \to \int_a^b \frac{1}{\pi\sqrt{x(1-x)}} dx.$$

Remark. Note that the density is symmetric about $1/2$, so in particular

$$\mathbb{P}(L_{2n}/2n \leq 1/2) \to 1/2.$$

In other words, if two people were to bet on a coin flip every day of the year, then with probability $1/2$ one of the players would be ahead for the last six months of the year!

Proof. From the previous lemma, we can take $k/n \to x$ to get

$$n\mathbb{P}(L_{2n} = 2k) \to \frac{1}{\pi\sqrt{x(1-x)}}.$$

The rest of the proof involves taking limits of densities, and can be found in [Dur19, 4.9] □

Exercise 5.6. The following problem contains the idea which finishes the proof above. Suppose Z_n are integer valued random variables and $\delta_n \downarrow 0$ are positive constants. Further, for some probability density f suppose that

$$\mathbb{P}(Z_n = z_n)/\delta_n \to f(x)$$

almost surely for every x, where z_n is any sequence of integers with $z_n\delta_n \to x$. Prove that $Z_n\delta_n \xrightarrow{d} X$ for X with density f.

Theorem 5.7 (Arcsine law for time above zero)
Let r_{2n} be the number of segments $(k-1, S_{k-1}) \to (k, S_k)$ that lie above zero. Then

$$\mathbb{P}(r_{2n} = 2k) = u_{2k}u_{2n-2k},$$

and hence for $0 < a < b < 1$,

$$\mathbb{P}(a \leq r_{2n}/2n \leq b) \to \int_a^b \frac{1}{\pi\sqrt{x(1-x)}} dx.$$

Proof. See [Dur19, 4.9]. □

5.1.2 Recurrence & Transience of Random Walks

Let $X_1, X_2, \dots$ be iid random variables in $\mathbb{Z}^d$, and $S_n = X_1 + \cdots + X_n$ be the corresponding random walk. A number $x \in \mathbb{R}^d$ is said to be *recurrent* if

$$\mathbb{P}(S_n = x \text{ i.o.}) = 1,$$

and we denote V to be the set of all recurrent sites. Note that the event $\{S_n = x \text{ i.o.}\} \in \mathcal{E}$, so the Hewitt-Savage law implies that the above probability must either be 0 or 1. Let a site $y \in \mathbb{R}^d$ be called *reachable* if

$$\mathbb{P}(S_n = y) > 0 \text{ for some } n,$$

and similarly denote U to be the set of all reachable sites. If $V = \varnothing$, then we say that the random walk is *transient*, otherwise we say it is *recurrent*. The following result says that any *irreducible* class of sites, i.e. every site is reachable from sites within the class, must either be entirely recurrent or entirely transient.

Theorem 5.8
Either $V = \varnothing$ or $V = U$.

Proof. First, we will show that if $x \in V$ and $y \in U$, then $x - y \in V$. If $y \in U$, then there exists k such that $\mathbb{P}(S_k = y) > 0$. So conditioned on the event $\{S_k = y\}$, we know that we hit x infinitely often. But note that

$$S_{k+1} - S_k, S_{k+2} - S_k, \dots$$

has the same distribution as the original random walk. Therefore S_n will hit $x - y$ infinitely often with probability 1. Since $V \subset U$, it follows that if $x \in V \neq \varnothing$, then $0 = x - x \in V$. But then $-x = 0 - x \in V$. Finally, if $y \in U$, then $-y = 0 - y \in V$, and so $y \in V$. Thus $U = V$. □

Exercise 5.9 (Transience of biased simple random walk)**.** Consider the biased simple random walk with $\mathbb{P}(X_i = 1) = p > 1/2$ and $\mathbb{P}(X_i = -1) = 1 - p$. Show that it is transient.

Theorem 5.10
If $\mathbb{E}X_i \neq 0$, then S_n is transient.

Proof. By strong law, we have

$$\frac{S_n}{n} \xrightarrow{a.s.} \mathbb{E}X_i,$$

so $S_n \to \pm\infty$ depending on the sign of $\mathbb{E}X_i$. Thus S_n cannot hit any site infinitely often. □

Let $\tau_0 = 0$, and τ_n be the n-th return time to 0. Then intuitively by memorylessness we should have

$$\mathbb{P}(\tau_n < \infty) = \mathbb{P}(\tau_1 < \infty)^n. \tag{11}$$

To make this rigorous, we will define the *strong Markov* property. Suppose we have an iid sequence $X_1, X_2, \dots$ and τ is a stopping time. Under the natural filtration

$$\mathcal{F}_n = \sigma(X_1, \dots, X_n),$$

consider $\mathcal{F}_\tau = \{A \in \mathcal{F}_\infty : A \cap \{\tau \leq t\} \in \mathcal{F}_t \text{ for all } t \geq 0\}$.

Theorem 5.11 (Strong Markov property)
If $\mathbb{P}(\tau < \infty) > 0$, then conditioning on $\{\tau < \infty\}$, the future $(X_{\tau+1}, X_{\tau+2}, \ldots)$ is independent of the past $\mathcal{F}_\tau$. Furthermore, the future is equal in distribution to $(X_1, X_2, \ldots)$.

Proof. Let $A \in \mathcal{F}_\tau$ and $B_1, B_2, \ldots B_k$ be Borel sets. Then note that to prove both claims, it suffices to show that

$$\begin{aligned}&\mathbb{P}(A, \{\tau < \infty\}, X_{\tau+1} \in B_1, X_{\tau+2} \in B_2, \ldots, X_{\tau+k} \in B_k)\\ &= \mathbb{P}(A, \{\tau < \infty\}) \prod_{i=1}^{k} \mathbb{P}(X_i = B_i).\end{aligned}$$

Splitting up the cases for τ, we can rewrite the left-hand side as

$$\begin{aligned}\sum_i \mathbb{P}(\underbrace{A \cap \{\tau = i\}}_{\in \mathcal{F}_i}, \cap_{j=1}^{k} X_{i+j} \in B_j) &= \sum_i \mathbb{P}(A \cap \{\tau = i\}) \prod_{j=1}^{k} \mathbb{P}(X_j \in B_j)\\ &= \mathbb{P}(A \cap \{\tau < \infty\}) \prod_{j=1}^{k} \mathbb{P}(X_j \in B_j).\end{aligned}$$

□

Now, to verify (11), note that by the strong Markov property,

$$\begin{aligned}\mathbb{P}(\tau_n < \infty) &= \mathbb{P}(\tau_n < \infty, \tau_{n-1} < \infty)\\ &= \mathbb{P}(\tau_{n-1} < \infty)\mathbb{P}(\tau_n < \infty | \tau_{n-1} < \infty)\\ &= \mathbb{P}(\tau_{n-1} < \infty)\mathbb{P}(\tau_1 < \infty | \tau_0 < \infty).\end{aligned}$$

Hence the result follows by induction. With this heuristic verified, let's see some consequences.

Theorem 5.12
The following are equivalent:
(i) S_n is recurrent.
(ii) $\mathbb{P}(\tau_1 < \infty) = 1$.
(iii) $\sum_{n=0}^{\infty} \mathbb{P}(S_n = 0) = \infty$.

Proof. (i) $\implies$ (ii): Since $\mathbb{P}(S_n \text{ hits } 0 \text{ i.o.}) = 1$, we must have $\mathbb{P}(\tau_1 < \infty) = 1$.
(ii) $\implies$ (iii): By (11) we know that $\mathbb{P}(\tau_n < \infty) = 1$ for all n. Then we can write

$$\begin{aligned}\sum_{n=0}^{\infty} \mathbb{P}(S_n = 0) &= \mathbb{E}\left[\sum_{n=0}^{\infty} \mathbf{1}_{S_n = 0}\right]\\ &= \mathbb{E}N\\ &= \sum_k \mathbb{P}(N \geq k)\\ &= \sum_k \underbrace{\mathbb{P}(\tau_k < \infty)}_{1} = \infty.\end{aligned}$$

where N is the total number of returns to 0.
(iii) $\implies$ (i): Using what we have above, we know that

$$\sum_k \mathbb{P}(\tau_1 < \infty)^k = \infty.$$

But then $\mathbb{P}(\tau_1 < \infty) = 1$, and so $\mathbb{P}(\tau_n < \infty) = 1$ for all n. Thus 0 must be hit infinitely often with probability 1, so S_n is recurrent. □

The next result acts as a converse to Theorem 5.10.

Theorem 5.13
If $\mathbb{E}X_i = 0$ and $S_n/n \xrightarrow{p} 0$, then S_n is recurrent.

Proof. First, we prove a useful lower bound:

$$\sum_{n=1}^{\infty} \mathbb{P}(S_n = 0) \geq \frac{1}{2k+1} \sum_{m=1}^{\infty} \mathbb{P}(|S_m| \leq k). \tag{12}$$

Note that this is essentially an average

$$\frac{1}{2k+1} \sum_{m=1}^{\infty} \sum_{j=-k}^{k} \mathbb{P}(S_m = j) = \frac{1}{2k+1} \sum_{j=-k}^{k} \sum_{m=1}^{\infty} \mathbb{P}(S_m = j),$$

so it suffices to write

$$\begin{aligned}
\sum_{m=1}^{\infty} \mathbb{P}(S_m = j) &= \sum_m \sum_l \mathbb{P}(T_j = l)\mathbb{P}(S_{m-l} = 0) \\
&= \sum_l \mathbb{P}(T_j = l) \sum_m \mathbb{P}(S_{m-l} = 0) \\
&= \mathbb{P}(T_j < \infty) \sum_m \mathbb{P}(S_m = 0) \\
&\leq \sum_m \mathbb{P}(S_m = 0).
\end{aligned}$$

Since this holds for every j, clearly it holds for the average of j from $-k$ to k. And so (12) follows.

Now, to show recurrence we just need to show that the right side of (12) diverges. We write

$$\begin{aligned}
\frac{1}{2k+1} \sum_{m=1}^{\infty} \mathbb{P}(|S_m| \leq k) &\geq \frac{1}{2k+1} \sum_{m=1}^{cK} \mathbb{P}(|S_m| \leq k) \\
&\geq \frac{1}{2k+1} \sum_{m=1}^{cK} \mathbb{P}\left(\frac{|S_m|}{m} \leq \frac{1}{c}\right),
\end{aligned}$$

for some large constant c. By the assumption, we know

$$\mathbb{P}\left(\frac{|S_m|}{m} \leq \frac{1}{c}\right) \to 1.$$

Combining this with the above, we see that

$$\sum_m \mathbb{P}(S_m = 0) \geq \frac{c}{2}.$$

But since c is any arbitrarily large constant, it follows that S_n is recurrent. □

Now, we turn to the case of simple random walks in $\mathbb{Z}^d$, where at each step S_n jumps to each of its neighbors uniformly at random. For example, in $d = 2$:

$$X_i \sim \text{Unif}\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\}.$$

Theorem 5.14
The simple random walk S_n is recurrent for $d \leq 2$ and transient for $d = 3$.

Proof. Recall that for $d = 1$, we have shown in the previous section that

$$\sum_n \mathbb{P}(S_{2n} = 0) \sim \sum_n \frac{1}{\sqrt{n}} = \infty,$$

giving us the result for one dimension.

For $d = 2$, we use a nice trick. Consider two independent one-dimensional simple random walks,

$$(S_n^{(1)}, S_n^{(2)}).$$

Note that by rotating by $\pi/4$, this is equivalent to the standard random walk in $d = 2$. Then we have

$$\mathbb{P}\left(S_{2n} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \mathbb{P}(S_{2n}^{(1)} = 0)^2 \sim \frac{1}{n},$$

which is still not summable. Hence the S_n is recurrent in this case as well.

For $d = 3$, intuitively we should have $\mathbb{P}(S_{2n} = 0) \sim n^{-3/2}$, which is summable. The details are left as an exercise or can be found in [Dur19, 5.4]. □

Exercise 5.15. Try computing each $\mathbb{P}(S_{2n} = 0)$ in two dimensions, and then using this show that the random walk is recurrent for $d = 2$. Stirling's approximation may be useful. Do the same for $d = 3$.

Corollary 5.16
The simple random walk S_n is transient for $d > 3$.

Proof. We simply extend the result for $d = 3$ using a projection idea. Let $T_n = (S_n^{(1)}, S_n^{(2)}, S_n^{(3)})$ be the random vector that keeps track of the first three coordinates of S_n, and let $N(n) = \inf\{m > N(n-1) : T_m \neq T_{N(n-1)}\}$ be the select indices on which one of the first three coordinates is chosen. Then $T_{N(n)}$ is a simple random walk in $d = 3$ which returns infinitely often to 0 with probability 0. Hence S_n cannot possibly return to 0 infinitely often itself, so it is transient as well. □

5.2 Construction & Definitions

First, we have a brief discussion on existence. Suppose we wish to construct a stochastic process $(X_i, i \in I)$, where I is perhaps a large index set, which may be countable or uncountable. Suppose each X_i exists on a space $(S_i, \mathcal{S}_i)$. Then we're looking to construct a probability measure on the sequence space

$$(\Omega_0, \mathcal{F}_\infty) = (S^I, \mathcal{S}^I).$$

For any finite subset $F \subset I$, the distribution of $X_F := (X_i, i \in F)$, known as a *finite-dimensional distribution* (FDD) is determined by projection of the law of $(X_i, i \in I)$ on the big product space. In particular, it satisfies the consistency condition in Kolmogorov's extension theorem, which in turn guarantees the existence of a unique probability measure $\mathbb{P}$ on the big product space which agrees with all the FDDs.

Remark. The construction via Kolmogorov's extension theorem is rarely used for two reasons:

- Often we can explicitly construct the underlying probability space and its process X.
- The space $\mathbb{R}^I$ may not be the best space on which to define the measure of interest.

We provide two examples illustrating these two points.

Example 5.17. To construct an iid sequence of uniform distributions on $[0,1]$, define $\Omega = [0,1]$, $X : \Omega \to \{0,1\}^{\mathbb{N}}$, where X maps ω to the binary expansion of ω, denoted by $X(\omega) = (X_i, i \in \mathbb{N})$. Then the X_i's are iid Ber(1/2) variables, so for disjoint subsets $I = \{i_1, i_2, \ldots\}$ and $J = \{j_1, j_2, \ldots\}$, define $U_1 = \sum 2^{-n} X_{i_n}$ and $U_2 = \sum 2^{-n} X_{j_n}$. Note that U_1, U_2 are iid Unif$[0,1]$. Repeat this construction for an infinite partition of $\mathbb{N}$ into disjoint subsets.

Example 5.18. For $I = [0,1]$, we typically cannot define a useful measure on the space $\mathbb{R}^I$ which measures events such as $\{$the sample path is continuous$\}$, since this event does not belong to the product measure, since every product measureable subset of $\mathbb{R}^I$ is determined by some countable collection of variables (X_{i_n}), and continuity cannot be characterized by any countable set.

5.2.1 Markov Kernels

Let $(S_1, \mathcal{S}_1)$ and $(S_2, \mathcal{S}_2)$ be measurable spaces. A *Markov kernel*, or *transition probability function*, is a function $P : S_1 \times \mathcal{S}_2 \to [0,1]$ which satisfies

(i) For each $x \in S$, the map $A \mapsto P(x, A)$ is a probability measure on $(S_2, \mathcal{S}_2)$.
(ii) For each $A \in \mathcal{S}$, the map $x \mapsto P(x, A)$ is $\mathcal{S}_1$-measurable.

For a Markov kernel P and a non-negative or bounded jointly measurable function $g : S_1 \times S_2 \to \mathbb{R}$, we define the function $(Pg) : S_1 \to \mathbb{R}$ by

$$(Pg)(x_1) := \int_{S_2} g(x_1, x_2) P(x_1, dx_2),$$

meaning that the second variable x_2 of $g(x_1, x_2)$ is integrated out with respect to the probability measure $P(x_1, \cdot)$.

6 Brownian Motion

Brownian motion is deeply related to the heat equation and diffusions, and can be used as a model for many physical phenomena ranging from the motion of pollen grains to the stock market. Also, it is the beginning of stochastic calculus.

6.1 Definition & Construction

A one-dimensional, real-valued stochastic process $\{B_t, t \geq 0\}$ is called a *Brownian motion* if it has the following properties:

(i) *Independent increments.* If $t_0 < t_1 < \cdots < t_n$, then $B_{t_0}, B_{t_1} - B_{t_0}, \ldots, B_{t_n} - B_{t_{n-1}}$ are independent.

(ii) *Gaussian increments.* For all $s, t \geq 0$ the increments are Gaussian,

$$\mathbb{P}(B_{s+t} - B_s \in A) = \int_A (2\pi t)^{-1/2} \exp(-x^2/2t) dx.$$

(iii) *Continuity.* The map $t \mapsto B_t$ is continuous with probability one.

Proposition 6.1

Suppose $\{B_t, t \geq 0\}$ is a Brownian motion. Then we have the properties:

- *Translation invariance.* The process $\{B_t - B_0, t \geq 0\}$ is a Brownian motion independent of B_0.
- *Brownian scaling.* If $B_0 = 0$, then for $t > 0$,

$$\{B_{st}, s \geq 0\} \stackrel{d}{=} \{t^{1/2} B_s, s \geq 0\},$$

where equal in distribution here means that for all $s_1 < \cdots < s_n$,

$$(B_{s_1 t}, \ldots, B_{s_n t}) \stackrel{d}{=} (t^{1/2} B_{s_1}, \ldots, t^{1/2} B_{s_n}).$$

Proof. Clearly $\{B_t - B_0, t \geq 0\}$ is a Brownian motion if $\{B_t, t \geq 0\}$ is. To show independence of B_0, let $\mathcal{A}_1 = \sigma(B_0)$ and $\mathcal{A}_2$ be the events of the form

$$\{B_{t_1} - B_{t_0} \in A_1, \ldots, B_{t_n} - B_{t_{n-1}} \in A_n\}.$$

Since $\mathcal{A}_1$ and $\mathcal{A}_2$ are independent π-systems, translation invariance follows from Theorem 2.3.

The Brownian scaling property for $n = 1$ follows from Gaussian increments, and we can extend it to $n > 1$ with independent increments. □

For Brownian motion starting at $B_0 = 0$, we have the following equivalent definition, i.e. $\{B_t, t \geq 0\}$ is a real-valued process satisfying the following properties:

(i)' *Gaussian process.* For every $t_0 < t_1 < \ldots < t_n$ the distribution of

$$\begin{bmatrix} B_{t_0} & B_{t_1} & \cdots & B_{t_n} \end{bmatrix}^\top$$

is multivariate Gaussian.

(ii)' *Second order statistics.* For all $s, t \geq 0$, we have $\mathbb{E}B_s = 0$ and $\mathbb{E}B_s B_t = s \wedge t$.

(iii)' *Continuity.* The map $t \mapsto B_t$ is continuous with probability one.

Exercise 6.2. Check that the two characterizations are equivalent.

7 Concentration of Measure

Suppose we have a collection of scalar random variables $X_1, \dots, X_n$. We may often wish to analyze the distribution of the sum

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

It turns out that, assuming that our X_i have a sufficient amount of regularity and independence, which will be quantified throughout this section, the probability will sharply concentrate in a relatively narrow range.

The reason we only consider scalars for now is for sake of intuition, as the techniques seen here can be extended for the analogous problems in random matrix theory.

7.1 The Moment Method

The first moment method should be a familiar application of Markov's inequality,

$$\mathbb{P}(|S_n| \geq \lambda) \leq \frac{1}{\lambda} \sum_{i=1}^{n} \mathbb{E}|X_i|, \tag{13}$$

as should the second moment method, an application of Chebyshev's inequality,

$$\mathbb{P}(|S_n| \geq \lambda) \leq \frac{1}{\lambda^2} \sum_{i=1}^{n} \mathrm{Var}(X_i), \tag{14}$$

where we have assumed that the X_i are pairwise independent.

Exercise 7.1. Come up with examples of random variables $X_1, \dots, X_n$ in which (13) and (14) are tight.

We can play a similar game with k-th moments, by assuming k-wise independence. We'd have to some combinatorial bookkeeping with the terms in

$$\mathbb{E}|S_n|^k = \sum_{1 \leq i_1, \dots, i_k \leq n} \mathbb{E} X_{i_1} \dots X_{i_k},$$

and after some algebra involving Stirling's formula, we can arrive at the large deviation bound

$$\mathbb{P}(|S_n| \geq \lambda\sqrt{n}) \leq 2 \left(\frac{\sqrt{ek/2}}{\lambda} \right)^k. \tag{15}$$

But instead of dwelling on this, we can often obtain much better bounds using exponential moments, namely by considering the moment generating function $\mathbb{E}e^{tS_n}$. The following is a bound for a single random variable which will be useful.

Lemma 7.2 (Crude Hoeffding's lemma)
If X is a scalar random variable taking values in $[a, b]$, then for any $t > 0$,

$$\mathbb{E}e^{tX} \leq e^{t\mathbb{E}X} \left(1 + O\left(t^2 \mathrm{Var}(X) e^{O(t(b-a))} \right) \right), \tag{16}$$

and in particular,

$$\mathbb{E}e^{tX} \leq e^{t\mathbb{E}X} e^{O(t^2 \mathrm{Var}(X))} \leq e^{t\mathbb{E}X} e^{O(t^2 (b-a)^2)}. \tag{17}$$

Proof. Note that we can subtract the mean from X, a, b and assume that $\mathbb{E}X = 0$. Furthermore, by normalizing X we can assume that $b - a = 1$. Then $X = O(1)$, and we have the Taylor expansion

$$\begin{aligned} e^{tX} &= 1 + tX + (tX)^2 \left(\frac{1}{2} + \frac{tX}{3!} + \cdots \right) \\ &= 1 + tX + O\left(t^2 X^2 e^{O(t)} \right). \end{aligned}$$

Taking expectations gives us

$$\mathbb{E}e^{tX} = 1 + O\left(t^2 \mathrm{Var}(X) e^{O(t)} \right),$$

proving (16). To get the other bound, note that $\mathrm{Var}(X) \leq (b-a)^2$, and consider the function

$$f(x) = \frac{1 + x^2 e^x}{e^{x^2}},$$

which can be shown to be bounded. With $x = t(b-a)$, this gives us (17). □

Using some calculus, we can sharpen Hoeffding's lemma to the following explicit bound:

Lemma 7.3 (Sharp Hoeffding's lemma)
As in the setting above, we have the sharp bound

$$\mathbb{E}e^{tX} \leq e^{t\mathbb{E}X} e^{t^2 (b-a)^2 / 8}. \tag{18}$$

Proof. Without loss of generality, we again assume that $\mathbb{E}X = 0$. By Jensen's inequality, we have

$$e^{tx} \leq \frac{b-x}{b-a} e^{ta} + \frac{x-a}{b-a} e^{tb}.$$

Letting $\theta = -\frac{a}{b-a} > 0$ and taking expectations we get

$$\begin{aligned} \mathbb{E}e^{tX} &\leq (1-\theta) e^{ta} + \theta e^{tb} \\ &= \end{aligned}$$

□

Theorem 7.4 (Chernoff bound)
Let $X_1, \ldots, X_n$ be independent scalar random variables with $|X_i| \leq K$ almost surely, with means μ_i and variances σ_i^2. Then for any $\lambda > 0$, we have

$$\mathbb{P}(|S_n - \mu| \geq \lambda \sigma) \leq C \max\left(e^{-c\lambda^2}, e^{-c\lambda\sigma/K} \right), \tag{19}$$

where $C, c > 0$ are constants, $\mu := \sum_{i=1}^n \mu_i$, and $\sigma^2 := \sum_{i=1}^n \sigma_i^2$.

Proof. We may assume that $\mu_i = 0$ and $K = 1$. It then suffices to prove the upper tail bound

$$\mathbb{P}(S_n \geq \lambda \sigma) \leq C \max\left(e^{-c\lambda^2}, e^{-c\lambda\sigma} \right).$$

Note that by independence,

$$\mathbb{E}e^{tS_n} = \prod_{i=1}^{n} \mathbb{E}e^{tX_i}.$$

By (17) and the fact that $|X| \leq 1$, we have

$$\mathbb{E}e^{tX_i} \leq e^{(O(t^2\sigma_i^2)},$$

and together this gives us

$$\mathbb{E}e^{tS_n} \leq e^{O(t^2\sigma^2)}.$$

By Markov's inequality, one has

$$\mathbb{P}(S_n \geq \lambda\sigma) \leq e^{O(t^2\sigma^2)-t\lambda\sigma}.$$

Optimizing over t subject to the constraint $t \in [0,1]$ gives us (19). □

Exercise 7.5. By letting t take values in some larger interval than $[0,1]$, show that the term $e^{-c\lambda\sigma/K}$ in the Chernoff bound can be replaced with $(\lambda K/\sigma)^{-c\lambda\sigma/K}$, which is better for when $\lambda K \gg \sigma$.

Corollary 7.6 (Hoeffding bound)
Let $X_1, \ldots, X_n$ be independent random variables taking values in intervals $[a_i, b_i]$, respectively. Then

$$\mathbb{P}(|S_n - \mathbb{E}S_n| \geq \lambda\sigma) \leq Ce^{-c\lambda^2},$$

where $C, c > 0$ are constants and $\sigma^2 := \sum_{i=1}^{n} |b_i - a_i|^2$.

Proof. This follows from Chernoff's bound and the fact that $\mathrm{Var}(S_n) \leq \sum_{i=1}^{n} |b_i - a_i|^2$. □

7.2 The Truncation Method

Proposition 7.7
Let $X_1, \ldots, X_n$ be iid copies of a sub-Gaussian random variable X, i.e.

$$\mathbb{P}(|X| \geq t) \leq Ce^{-ct^2}$$

for all $t > 0$ and some $C, c > 0$. Then for any sufficiently large A, we have

$$\mathbb{P}(|S_n - n\mu| \geq An) \leq C_A e^{-c_A n}$$

for some constants C_A, c_A depending on A, C, c. Furthermore, $c_A = O(A)$.

Proof. Without loss of generality, we may assume $\mathbb{E}X = 0$. Let

$$X_i = \sum_{m=0}^{\infty} X_{i,m},$$

where we define $X_{i,0} = X_i \mathbf{1}_{|X_i| \leq 1}$, and $X_{i,m} = X_i \mathbf{1}_{2^{m-1} < |X_i| \leq 2^m}$ for $m \geq 1$. For visual purposes, consider the array

$$\begin{array}{cccc} X_{1,0} & X_{1,1} & X_{1,2} & \dots \\ X_{2,0} & X_{2,1} & X_{2,2} & \dots \\ X_{3,0} & X_{3,1} & X_{3,2} & \dots \\ \vdots & \vdots & \vdots & \end{array}$$

Then we write the decomposition

$$S_n = \sum_{m=0}^{\infty} S_{n,m},$$

where we define $S_{n,m} = \sum_{i=1}^{n} X_{i,m}$. By a union bound, we have (the factor of 100 is much larger than needed; we just need a factor large enough to normalize $\sum m^{-2}$)

$$\mathbb{P}(|S_n| \geq An) \leq \sum_{m=0}^{\infty} \mathbb{P}\left(|S_{n,m}| \geq \frac{An}{100(m+1)^2}\right).$$

Now, aiming to use a Chernoff bound for each of the terms in the sum, first note that the sub-Gaussian hypothesis gives us

$$\begin{aligned} |\mathbb{E}[X_{i,m}]|, \mathbb{E}[X_{i,m}^2] &\leq \mathbb{P}(|X| \geq 2^{m-1}) 2^{2m} \\ &\leq C' e^{-c' 2^{2m}}. \end{aligned}$$

From this we get bounds on the mean and variance $S_{n,m}$,

$$|\mu|, \sigma \leq n C' e^{-c' 2^{2m}}.$$

Hence, applying the variant of Chernoff in Exercise 7.5 with $\lambda = \frac{An}{100(m+1)^2 \sigma}$, we get

$$\mathbb{P}(|S_{n,m}| \geq \lambda \sigma) \leq C \max\left(e^{-c\lambda^2}, (\lambda K/\sigma)^{-c\lambda\sigma/K}\right).$$

Plugging in our estimates, we get that

$$\begin{aligned} e^{-c\lambda^2} = e^{-c\frac{A^2 n}{m^4} e^{c' 2^{2m}}} &\leq C'' e^{-c'' An} \\ (\lambda K/\sigma)^{-c\lambda\sigma/K} = \left(\frac{2^m}{m^2} A e^{c' 2^{2m}}\right)^{-c\frac{An}{m^2} 2^{-m}} &\leq O(e^{-cAn}), \end{aligned}$$

as required. □

Exercise 7.8. Generalize the sub-Gaussian hypothesis for $p > 1$:

$$\mathbb{P}(|X| \geq t) \leq C e^{-ct^p}.$$

Furthermore, check that the conclusion fails for the case $0 < p \leq 1$.

References

[Dur19] *Probability: Theory and Examples*, fifth edition, Cambridge University Press, 2019.