II Probability and Measure

Ishan Nath, Michaelmas 2023

Based on Lectures by Dr. Sourav Sarkar

January 18, 2024

Contents

1   Introduction

Probability and measure tries to fix, redeem and formally define some areas of analysis and probability that we previously saw, but were not yet equipped to handle.

Some examples in analysis are:

1. Generalizing the notion of length in $\mathbb{R}$, area in $\mathbb{R}^2$ and volume in $\mathbb{R}^d$ to arbitrary subsets.

2. Riemann integration as defined has some flaws. Suppose we have a sequence of continuous functions $\{f_n\}$, with $0 \le f_n(x) \le 1$ for all $x \in [0, 1]$, and with $(f_n(x))$ being a monotonically decreasing sequence in n.

   Then we can define the pointwise limit $f(x)$. Given these functions are continuous, their integral exists. However it may be the case that f is not Riemann integrable. We want a theory of integration such that the set of integrable functions is closed, and satisfies

   $$\lim_{n \to \infty} \int_0^1 f_n(x) \, \mathrm{d}x = \int_0^1 f(x) \, \mathrm{d}x.$$

3. In a similar vein to the above, the Riemann integrable functions are not complete.

Some examples in probability are:

1. Defining the probability of an event A that depends on an infinite sequence. We saw before that, if $X_i \sim \mathrm{Ber}(1/2)$ for $i \in \mathbb{N}$, then

   $$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{n \to \infty} \frac{1}{2} \right) = 1.$$

   This is the strong law of large numbers. But how do we even define this probability?

2. What does it mean to draw a point uniformly from $[0, 1]$? What about uniformly from $N(0, 1)$?

3. Similarly to the analysis example, we want to define the expectation $\mathbb{E}$ for a random variable formally.

   We would also want that if $0 \le X_n \le 1$ and $X_n \downarrow X$, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.

We will define a new framework that will allow us to solve these problems.

2   Defining Measures

We begin by building up a list of definitions that we will culminate in us being able to define the measure of a reasonable collection of sets.

2.1 σ-algebras

Definition 2.1 (σ-algebra)**.** Let E be a set. A σ-algebra $\mathcal{E}$ is a collection of subset of E such that:

- $\emptyset \in \mathcal{E}$.

- If $A \in \mathcal{E}$, then $A^c \in \mathcal{E}$.

- If (A_n) is a sequence of elements in $\mathcal{E}$, then $\bigcup A_n \in \mathcal{E}$.

The pair $(E, \mathcal{E})$ is called a *measurable space.*

The smallest σ-algebra is $\mathcal{E} = (\emptyset, E)$, and the biggest is $\mathcal{E} = \mathcal{P}(E)$. Typically we will work with intermediary σ-algebras.

Note that as $\bigcap A_n = (\bigcup A_n^c)^c$, we can take countable intersections in a σ-algebra. Moreover, if $A, B \in \mathcal{E}$, then $B \setminus A = B \cap A^c \in \mathcal{E}$.

Definition 2.2 (Measures)**.** A *measure* μ on $(E, \mathcal{E})$ is a non-negative function $\mu : \mathcal{E} \to [0, \infty]$ such that:

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

- (Countable additivity) If (A_n) is a sequence of disjoint elements of $\mathcal{E}$, then

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

We call $(E, \mathcal{E}, \mu)$ a *measure space*

Let E be a countable set, then for all $A \subset E$,

$$\mu(A) = \mu\left(\bigcup_{x \in A} \{x\}\right) = \sum_{x \in A} \mu(\{x\}).$$

Define $m : E \to [0, \infty]$ such that $m(x) = \mu(\{x\})$. Such an m is called a *mass function*, or pmf in discrete probability, and measures μ are in one-to-one correspondence with mass functions m.

Here $\mathcal{E} = \mathcal{P}(E)$, and this is the theory in elementary discrete probability.

When $\mu(\{x\}) = 1$ for all x, μ is called a counting measure. Here $\mu(A) = |A|$.

For uncountable E however, the theory is not so simple and $\mathcal{E} = \mathcal{P}(E)$ is generally no feasible. Instead, measures are defined on σ-algebras generated by a smaller class of simple subsets of E.

If $\mathcal{A}$ is any collection of subsets of E,

$$\sigma(\mathcal{A}) = \{A \subset E \mid A \in \mathcal{E} \text{ for all } \sigma\text{-algebras } \mathcal{E} \supset \mathcal{A}\} = \bigcap_{\substack{\mathcal{E} \supset \mathcal{A} \\ \mathcal{E} \text{ a } \sigma\text{-algebra}}} \mathcal{E}.$$

This is the σ-algebra generated by $\mathcal{A}$, which can be thought of as the smallest σ-algebra containing $\mathcal{A}$. We can verify that $\sigma(\mathcal{A})$ is indeed a σ-algebra.

2.2   Types of Classes

The class $\mathcal{A}$ that we choose will usually satisfy some properties. In the following let E be a set and $\mathcal{A}$ a collection of subsets of E.

Definition 2.3. $\mathcal{A}$ is called a *ring* if:

- $\emptyset \in \mathcal{A}$.

- If $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$ and $B \setminus A \in \mathcal{A}$.

Note that if $A, B \in \mathcal{A}$ and $\mathcal{A}$ is a ring, then $A \triangle B = (A \setminus B) \cup (B \setminus A) \in \mathcal{A}$, so $A \cap B = (A \cup B) \setminus (A \triangle B) \in \mathcal{A}$.

Definition 2.4. $\mathcal{A}$ is called a *algebra* if:

- $\emptyset \in \mathcal{A}$.

- If $A, B \in \mathcal{A}$, then $A^c \in A$ and $A \cup B \in \mathcal{A}$.

If $\mathcal{A}$ is an algebra and $A, B \in \mathcal{A}$, then $A \cap B \in \mathcal{A}$ and $B \setminus A = B \cap A^c \in \mathcal{A}$. Hence every algebra is a ring. However the converse is not true: $\{\emptyset\}$ is a ring but not an algebra.

The general idea of the upcoming parts are:

- Define a set function on a suitable collection $\mathcal{A}$, for example length $\ell((a, b)) = b - a$ on the intervals.

- Extend this to a measure on $\sigma(\mathcal{A})$ (Caratheodory's extension theorem).

- Show that such an extension is unique (Dynkin's lemma).

Definition 2.5 (Set-function)**.** Let $\mathcal{A}$ be any collection of subsets of E such that $\emptyset \in \mathcal{A}$. A *set-function* μ is a map $\mu : \mathcal{A} \to [0, \infty]$ such that $\mu(\emptyset) = 0$. We say:

(i) μ is *increasing* if $\mu(A) \leq \mu(B)$ for all $A, B \in \mathcal{A}$ with $A \subseteq B$,

(ii) μ is *additive* if $\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B \in \mathcal{A}$ with A, B disjoint and $A \cup B \in \mathcal{A}$,

(iii) μ is *countably additive* if

$$\mu\left(\bigcup_n A_n\right) = \sum_n \mu(A_n)$$

for all A_n disjoint, $A_n \in \mathcal{A}$ and $\bigcup A_n \in \mathcal{A}$,

(iv) μ is *countably subadditive* if

$$\mu\left(\bigcup_n A_n\right) \leq \sum_n \mu(A_n)$$

for all $A_n \in \mathcal{A}$ and $\bigcup A_n \in \mathcal{A}$.

Note that if μ is countably additive on $\mathcal{A}$, and $\mathcal{A}$ is a ring, then μ satisfies all other properties.

2.3   Caratheodory's Extension Theorem

Now we are ready for the big theorem of this section.

Theorem 2.1 (Caratheodory). *Let $\mathcal{A}$ be a ring of subsets of E, and $\mu : \mathcal{A} \to [0, \infty]$ be a countably additive set function on $\mathcal{A}$. Then μ extends to a measure on $\sigma(\mathcal{A})$.*

Proof: Technical, long, confusing maybe. Let's do it.

For any $B \subseteq E$, define

$$\mu^*(B) = \inf\left\{\sum_{i \in I} \mu(A_i) \mid B \subseteq \bigcup_{i \in I} A_i, A_i \in \mathcal{A}\right\}.$$

If no such collection exists, let $\mu^*(B) = \infty$. Clearly $\mu^*(\emptyset) = 0$ and μ^* is increasing. So μ^* is an increasing set function on $\mathcal{P}(E)$.

Call a set $A \subseteq E$ μ^*-*measurable* if for all $B \subseteq E$,

$$\mu^*(B) = \mu^*(B \cap A) + \mu^*(B \cap A^c).$$

Define $\mathcal{M} = \{A \subseteq E \mid A \text{ is } \mu^* \text{ measurable}\}$. We will show that $\mathcal{M}$ is a σ-algebra that contains $\mathcal{A}$, and that $\mu^*|_{\mathcal{M}}$ is a measure on $\mathcal{M}$ that extends μ.

We will begin by showing μ^* is countably subadditive on $\mathcal{P}(E)$. Say that $B \subseteq \bigcup B_n$, and we will show $\mu^*(B) \leq \sum \mu^*(B_n)$.

We have nothing to prove if $\mu^*(B_n) = \infty$ for some n, so assume $\mu^*(B_n) < \infty$ for all n. Now for all $\varepsilon > 0$, let $(A_{n,m}) \in \mathcal{A}$ be such that

$$B_n \subseteq \bigcup_m A_{n,m}, \qquad \mu^*(B_n) + \frac{\varepsilon}{2^n} \geq \sum_m \mu(A_{n,m}).$$

But then $B \subseteq \bigcup B_n \subseteq \bigcup A_{n,m}$, and $(A_{n,m}) \in \mathcal{A}$, so

$$\mu^*(B) \leq \sum_n \sum_m \mu(A_{n,m}) \leq \sum_n \sum_n \left(\mu^*(B_n) + \frac{\varepsilon}{2^n} \right) = \sum_n \mu^*(B_n) + \varepsilon.$$

As $\varepsilon > 0$ is arbitrary, we get $\mu^*(B) \leq \sum \mu^*(B_n)$.

We will now show that μ^* extends μ, so for all $A \in \mathcal{A}$, $\mu^*(A) = \mu(A)$. We immediately have that $\mu^*(A) \leq \mu(A)$ by the definition of μ^*, so it suffices to show $\mu^*(A) \geq \mu(A)$.

As μ is countably additive on $\mathcal{A}$ and $\mathcal{A}$ is a ring, μ is countably subadditive on $\mathcal{A}$ and is increasing. Now let $A \subseteq \bigcup A_n$ with $A, A_n \in \mathcal{A}$. Then $A = \bigcup (A \cap A_n)$, so by countable subadditivity,

$$\mu(A) \leq \sum_n \mu(A \cap A_n) \leq \sum_n \mu(A_n).$$

So by taking the infimum over all such $\{A_n\}$, $\mu(A) \leq \mu^*(A)$.

Now we will show that $\mathcal{M} \supseteq \mathcal{A}$, so for all $A \in \mathcal{A}$ and $B \subseteq E$, that

$$\mu^*(B) = \mu^*(B \cap A) + \mu^*(B \cap A^c).$$

As μ^* is countably subadditive, we immediately get that $\mu^*(B) \leq \mu^*(B \cap A) + \mu^*(B \cap A^c)$. Hence we just need to show $\mu^*(B) \geq \mu^*(B \cap A) + \mu^*(B \cap A^c)$.

If $\mu^*(B) = \infty$, we have nothing to prove. So assume $\mu^*(B) < \infty$. For all $\varepsilon > 0$, there exists $(A_n) \in \mathcal{A}$ with

$$B \subseteq \bigcup_n A_n, \qquad \mu^*(B) + \varepsilon \geq \sum_n \mu(A_n).$$

Then $(B \cap A) \subseteq \bigcup (A_n \cap A)$, $(B \cap A^c) \subseteq \bigcup (A_n \cap A^c)$, so

$$\mu^*(B \cap A) \leq \sum_n \mu(A_n \cap A), \qquad \mu^*(B \cap A^c) \leq \sum_n \mu(A_n \cap A^c).$$

Hence we get that

$$\mu^*(B\cap A)+\mu^*(B\cap A^c) \leq \sum_n (\mu(A_n\cap A)+\mu(A_n\cap A^c)) = \sum \mu(A_n) \leq \mu^*(B)+\varepsilon.$$

As $\varepsilon > 0$ is arbitrary, we get $\mu^*(B \cap A) + \mu^*(B \cap A^c) \leq \mu^*(B)$, showing the two are equal.

Now we begin showing $\mathcal{M}$ is a σ-algebra, by first showing it is an algebra. Notice that clearly $\emptyset \in \mathcal{A}$, and from the symmetric definition we have that if $A \in \mathcal{M}$, then $A^c \in \mathcal{M}$.

To show it is an algebra, it suffices to show it is closed under finite intersection. Let $A_1, A_2 \in \mathcal{M}$, and $B \subseteq E$. Then using our condition multiple times,

$$\begin{aligned}
\mu^*(B) &= \mu^*(B \cap A_1) + \mu^*(B \cap A_1^c) \\
&= \mu^*(B \cap A_1 \cap A_2) + \mu^*(B \cap A_1 \cap A_2^c) + \mu^*(B \cap A_1^c) \\
&= \mu^*(B \cap (A_1 \cap A_2)) + \mu^*(B \cap (A_1 \cap A_2)^c \cap A_1) \\
&\quad + \mu^*(B \cap (A_1 \cap A_2)^c \cap A_1^c) \\
&= \mu^*(B \cap (A_1 \cap A_2)) + \mu^*(B \cap (A_1 \cap A_2)^c).
\end{aligned}$$

Hence $A_1 \cap A_2 \in \mathcal{M}$, and $\mathcal{M}$ is an algebra.

Finally we show that $\mathcal{M}$ is a σ-algebra and $\mu^*|_{\mathcal{M}}$ is a measure. As we have shown $\mathcal{M}$ is an algebra, it suffices to prove for all disjoint $(A_n) \in \mathcal{M}$, that $\bigcup A_n \in \mathcal{M}$, and that $\mu(A) = \sum \mu(A_n)$.

As before, for any $B \subseteq E$,

$$\begin{aligned}
\mu^*(B) &= \mu^*(B \cap A_1) + \mu^*(B \cap A_1^c) \\
&= \mu^*(B \cap A_1) + \mu^*(B \cap A_1^c \cap A_2) + \mu^*(B \cap A_1^c \cap A_2^c) \\
&= \mu^*(B \cap A_1) + \mu^*(B \cap A_2) + \mu^*(B \cap A_1^c \cap A_2^c) \\
&\vdots \\
&= \sum_{i=1}^n \mu^*(B \cap A_i) + \mu^*(B \cap A_1^c \cap \cdots \cap A_n^c) \\
&\geq \sum_{i=1}^n \mu^*(B \cap A_i) + \mu^*(B \cap A^c),
\end{aligned}$$

as $A = \bigcup A_i$, and μ^* is increasing. As $n \to \infty$,

$$\mu^*(B) \geq \sum_{i=1}^{\infty} \mu^*(B \cap A_i) + \mu^*(B \cap A^c)$$
$$\geq \mu(B \cap A) + \mu^*(B \cap A^c),$$

by countable subadditivity. The other direction is obvious, so $\mu^*(B) = \mu^*(B \cap A) + \mu^*(B \cap A^c)$.

Now μ^* is a measure as if we take $B = A$,

$$\mu^*(A) \geq \sum_{i=1}^{\infty} \mu^*(A_i) + \mu^*(\emptyset) = \sum_{i=1}^{\infty}(A_i),$$

and as μ^* is subadditive we get the two sides are equal.

2.4   Uniqueness of Extension

To address the uniqueness of extension, we introduce further subclasses of $\mathcal{P}(E)$. Let $\mathcal{A}$ be a collection of subsets of E.

1. $\mathcal{A}$ is called a π-*system* is $\emptyset \in \mathcal{A}$ and for all $A, B \in \mathcal{A}$, $A \cap B \in \mathcal{A}$.

2. $\mathcal{A}$ is called a *d-system* (or λ-system) if:

   - $E \in \mathcal{A}$.

   - If $A, B \in \mathcal{A}$ and $A \subseteq B$, then $B \setminus A \in \mathcal{A}$.

   - If $\{A_n\} \in \mathcal{A}$ and $A_n \uparrow A = \bigcup A_n$, then $A \in \mathcal{A}$.

   Here $A_n \uparrow A$ means $A_1 \subseteq A_2 \subseteq A_3 \cdots$, and $A = \bigcup A_n$.

   Note equivalently $\mathcal{A}$ is a d-system if:

   - $\emptyset \in \mathcal{A}$.

   - If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$.

   - If $\{A_n\} \in \mathcal{A}$ and A_n are disjoint, then $\bigcup A_n \in \mathcal{A}$.

We can prove that a π-system $\mathcal{A}$ which is also a d-system is actually a σ-algebra.

Lemma 2.1 (Dynkin's Lemma). *Let $\mathcal{A}$ be a π-system. Then any d-system that contains $\mathcal{A}$, also contains $\sigma(\mathcal{A})$.*

Proof: Define
$$\mathcal{D} = \bigcap_{\substack{\overline{\mathcal{D}} \text{ a d-system} \\ \overline{\mathcal{D}} \supseteq \mathcal{A}}} \overline{\mathcal{D}}.$$

Then $\mathcal{D}$ is itself a d-system. We will show that $\mathcal{D}$ is also a π-system, so it is a σ-algebra. As it is a σ-algebra containing $\mathcal{A}$, it must contain $\sigma(\mathcal{A})$.

To achieve this, define

$$\mathcal{D}' = \{B \in \mathcal{D} \mid B \cap A \in \mathcal{D} \text{ for all } A \in \mathcal{A}\}.$$

Then $\mathcal{A} \subseteq \mathcal{D}'$ as $\mathcal{A}$ is a π-system. Moreover, $\mathcal{D}'$ is also a d-system. Fix $A \in \mathcal{A}$.

- Then $E \cap A = A \in \mathcal{A} \subseteq \mathcal{D}$, so $E \in \mathcal{D}'$.

- If $B_1 \subseteq B_2$, and $B_1, B_2 \in \mathcal{D}'$, then $(B_2 \setminus B_1) \cap A = (B_2 \cap A) \setminus (B_1 \cap A)$.

  But by definition, as $B_1, B_2 \in \mathcal{D}'$, then $B_1 \cap A$ and $B_2 \cap A \in \mathcal{D}$. Since $\mathcal{D}$ is a d-system, $(B_2 \cap A) \setminus (B_1 \cap A) \in \mathcal{D}$.

  So $(B_2 \setminus B_1) \cap A \in \mathcal{D}$, hence $B_2 \setminus B_1 \in \mathcal{D}'$.

- Finally, if $\{B_n\} \in \mathcal{D}'$ and $B_n \uparrow B$, then $B_n \cap A \in \mathcal{D}$, and $(B_n \cap A) \uparrow (B \cap A)$. Since $\mathcal{D}$ is a d-system, $B \cap A \in \mathcal{D}$. So $B \in \mathcal{D}'$.

So $\mathcal{D}'$ is a d-system, and moreover $\mathcal{D}' \subseteq \mathcal{D}$. But from definition, as $\mathcal{A} \subseteq \mathcal{D}'$, $\mathcal{D} \subseteq \mathcal{D}'$. Therefore $\mathcal{D} = \mathcal{D}'$, and so for all $B \in \mathcal{D}$ and $A \in \mathcal{A}$, $B \cap A \in \mathcal{D}$.

Now we repeat this argument one level higher. Define

$$\mathcal{D}'' = \{B \in \mathcal{D} \mid B \cap A \in \mathcal{D} \text{ for all } A \in \mathcal{D}\}.$$

Then $\mathcal{A} \subseteq \mathcal{D}''$, from the above. We can similarly check that $\mathcal{D}''$ is a d-system, from which it follows that $\mathcal{D} = \mathcal{D}''$ and then that $\mathcal{D}$ is a π-system.

Therefore $\mathcal{D}$ is a σ-algebra, and so $\mathcal{D} \supseteq \sigma(\mathcal{A})$.

Now we get the uniqueness theorem.

Theorem 2.2 (Uniqueness of extension). *Let μ_1, μ_2 be measures on $(E, \mathcal{E})$ with $\mu_1(E) = \mu_2(E) < \infty$. Suppose that $\mu_1 = \mu_2$ on $\mathcal{A}$ for some π-system $\mathcal{A}$ generating $\mathcal{E}$. Then $\mu_1 = \mu_2$ on $\mathcal{E}$.*

Proof: Consider

$$\mathcal{D} = \{A \in \mathcal{E} \mid \mu_1(A) = \mu_2(A)\}.$$

Then by the hypothesis, $\mathcal{A} \subseteq \mathcal{D}$. Also $\mathcal{A}$ is a π-system. So it is enough to show that $\mathcal{D}$ is a d-system:

- $\emptyset \ni \mathcal{D}$, since $\mu_1(\emptyset) = \mu_2(\emptyset) = 0$.

- Let $A \in \mathcal{D}$. Then $\mu_1(A) = \mu_2(A)$. Therefore

$$\mu_1(A^c) = \mu_1(E) - \mu_1(A) = \mu_2(E) - \mu_2(A) = \mu_2(A^c).$$

  Hence $A^c \in \mathcal{D}$.

- If $\{A_n\} \in \mathcal{D}$ is disjoint, then

$$\mu_1\left(\bigcup_n A_n\right) = \sum_n \mu_1(A_n) = \sum_n \mu_2(A_n) = \mu_2\left(\bigcup_n A_n\right).$$

  So $\bigcup A_n \in \mathcal{D}$.

Hence $\mathcal{D}$ is a d-system.

Note if $A_n \uparrow A$, then $\mu(A) = \lim_{n \to \infty} \mu(A_n)$. We can use this to show that $\mathcal{D}$ is a d-system.

One question that arises is, how can we show that all sets of a σ-algebra $\mathcal{E}$ generated by $\mathcal{A}$ satisfy a certain property P? The answer is as follows: let

$$\mathcal{G} = \{A \subseteq \mathcal{E} \mid A \text{ has the property } P\},$$

and suppose all elements of $\mathcal{A}$ have the property $\mathcal{P}$. We can do the following:

- Show that $\mathcal{G}$ is a σ-algebra.

- Show that $\mathcal{G}$ is a d-system, and $\mathcal{A}$ is a π-system, and then use Dynkin's lemma.

- Monotone class theorem.

3   Lebesgue Measure

3.1   Borel Sets

Let E be a topological space, which we take to be Hausdorff. The σ-algebra generated by the set of open sets, i.e. $\sigma(\mathcal{A})$ where $\mathcal{A} = \{A \subseteq E \mid A \text{ open}\}$ is called the *Borel σ-algebra* $\mathcal{B}(E)$ of E.

We usually denote $\mathcal{B}(\mathbb{R})$ by $\mathcal{B}$. A measure μ on $(E, \mathcal{B}(E))$ is called a *Borel measure* on E. If $\mu(K) < \infty$ for all K compact, then μ is called a *Radon measure* on E.

If $\mu(E) = 1$, then μ is called a *probability measure* on E, and $(E, \mathcal{E}, \mu)$ is called a *probability space*. We usually use $(\Omega, \mathcal{F}, \mathbb{P})$ instead.

If $\mu(E) < \infty$, then μ is a *finite measure* on E. If there exists (E_n) in $\mathcal{E}$, such that $\mu(E_n) < \infty$ for all n and $E = \bigcup E_n$, then μ is called a *σ-finite measure*. Note we can extend the uniqueness theorem to σ-finite measures.

One of the main goals of this course is to construct a Borel measure μ on $\mathcal{B}(\mathbb{R}^d)$, such that

$$\mu\left(\prod_{i=1}^{d}(a_i, b_i)\right) = \prod_{i=1}^{d}(b_i - a_i),$$

corresponding to the usual notion of volume of rectangles. This measure will be called the *Lebesgue measure*. We begin by proving this for $d = 1$

3.2   Uniqueness of Lebesgue Measure

Theorem 3.1. *There exists a unique Borel measure μ on $\mathbb{R}$ such that, for all $a, b \in \mathbb{R}$ with $a < b$,*

$$\mu((a, b]) = b - a. \tag{3.1}$$

μ is called the Lebesgue measure on $\mathbb{R}$.

Proof: We first show existence. Consider the ring $\mathcal{A}$ of finite unions of disjoint intervals of this form: $A \in \mathcal{A}$ if and only if

$$A = (a_1, b_1] \cup (a_2, b_2] \cup \cdots \cup (a_n, b_n]$$

where $a_1 \leq b_1 \leq a_2 \leq b_2 \leq \cdots \leq a_n \leq b_n$. Note that $\sigma(\mathcal{A}) = \mathcal{B}$, as all open intervals are in $\alpha(\mathcal{A})$ and open intervals generate open sets.

Define for such $A \in \mathcal{A}$, the measure

$$\mu(A) = \sum_{i=1}^{n} (b_i - a_i).$$

This agrees with (3.1), and is additive and well-defined. So the existence of μ on $\mathcal{B}$ follows from Caratheodory's theorem, if we can show that μ is countably additive on $\mathcal{A}$.

Recall that a finitely additive set function on a ring A is countably additive if and only if, if $A_n \uparrow A$ with $A_n, A \in \mathcal{A}$, then $\mu(A_n) \uparrow \mu(A)$. In addition, if μ is finite and $A_n \downarrow A$, for $A_n, A \in \mathcal{A}$, then $\mu(A_n) \downarrow \mu(A)$. (This is proved in the example sheets).

So from this result, showing μ is countably additive on A is equivalent to showing that if $A_n \in \mathcal{A}$, and $A_n \downarrow \emptyset$, then $\mu(A_n) \downarrow 0$. We shall prove this by contradiction.

Suppose there exists $B_n \in \mathcal{A}$, with $B_n \downarrow \emptyset$ but for all n, $\mu(B_n) \geq 2\varepsilon > 0$. Then for each n, B_n can be approximated from within by $C_n \in \mathcal{A}$, such that $\overline{C_n} \subseteq B_n$ and

$$\mu(B_n \setminus C_n) \leq \frac{\varepsilon}{2^n}.$$

Then

$$\mu(B_n \setminus C_1 \cap C_2 \cap \cdots \cap C_n) \leq \mu\left(\bigcup_{i=1}^{n}(B_i \setminus C_i)\right) \leq \sum_{i=1}^{n} \mu(B_i \setminus C_i) \leq \sum_{i=1}^{n} \frac{\varepsilon}{2^i} \leq \varepsilon.$$

Since $\mu(B_n) \geq 2\varepsilon$ and $\mu(B_n \setminus (C_1 \cap \cdots \cap C_n)) \leq \varepsilon$, which implies $\mu(C_1 \cap \cdots \cap C_n) \geq \varepsilon$, and hence $C_1 \cap \cdots \cap C_n \neq \emptyset$.

So $K_n = \overline{C_1} \cap \cdots \cap \overline{C_n} \neq \emptyset$, (K_n) is a sequence of decreasing closed sets, each non-empty. So,

$$\bigcap_n K_n \neq \emptyset.$$

But then $\emptyset \neq \bigcap K_n \subseteq \bigcap B_n = \emptyset$, which is a contradiction.

Now we have to show uniqueness. Suppose μ, λ are two measures on $\mathcal{B}$ such that (3.1) holds for sets of the form $(a, b]$. Define the truncated measures, for $A \in \mathcal{B}$, as

$$\mu_n(A) = \mu(A \cap (n, n+1]), \qquad \lambda_n(A) = \lambda(A \cap (n, n+1]).$$

Then μ_n, λ_n are probability measures on $\mathcal{B}$, and $\mu_n = \lambda_n$ on the π-system of intervals of the form $(a, b]$. This generates $\mathcal{B}$, so by the uniqueness theorem for finite measures, $\mu_n = \lambda_n$ on $\mathcal{B}$.

Hence, for all $A \in \mathcal{B}$,

$$\mu(A) = \mu\left(\bigcup_n (A \cap (n, n_1])\right) = \sum_n \mu(A \cap (n, n+1]) = \sum_n \mu_n(A)$$
$$= \sum_n \lambda_n(A) = \lambda(A).$$

Remark. A set $B \in \mathcal{B}$ is called a *Lebesgue null set* if $\lambda(B) = 0$.

1. Any singleton set $\{x\}$. Note $\{x\} = \bigcap(x - 1/n, x]$, so

$$\lambda(\{x\}) = \lim_{n \to \infty} \lambda(x - 1/n, x] = \lim_{n \to \infty} 1/n = 0.$$

   In particular,
$$\lambda((a, b)) = \lambda([a, b]) = \lambda([a, b)) = b - a.$$

   Any countable set Q satisfies $\lambda(Q) = 0$, as they are countable unions of singletons.

   Moreover there exist uncountable sets C with $\lambda(C) = 0$. For example take the Cantor set.

2. The Lebesgue measure is translation invariant: for $x \in \mathbb{R}$, and $B \in \mathcal{B}$, define

$$B + x = \{b + x \mid b \in B\}.$$

   Define the translation Lebesgue measure as

$$\lambda_x(B) = \lambda(B + x).$$

   Then, this agrees with λ:

$$\lambda_x((a, b]) = \lambda((a, b] + x) = \lambda((a + x, b + x]) = b - a = \lambda((a, b]).$$

   These agree on the π-system of intervals, and hence $\lambda_x = \lambda$ on $\mathcal{B}$. So for all Borel sets B, $\lambda(B + x) = \lambda(B)$. Up to scaling, λ is the only translation-invariant measure on $\mathcal{B}$.

3. Caratheodory extends λ from A to not just $\sigma(A) = \mathcal{B}$, but actually to M, the set of outer-measurable sets $M \supseteq \mathcal{B}$, but how large is M? We will prove

$$M = \{A \cup N \mid A \in \mathcal{B}, N \subseteq B, B \in \mathcal{B}, \lambda(B) = 0\}.$$

We now show that $\mathcal{B}$ isn't too large, and in fact show that $\mathcal{B} \subsetneq \mathcal{P}(\mathbb{R})$. This fact was first shown by Vitali.

Consider $E = [0, 1)$ with addition modulo 1. Then, the Lebesgue measure is invariant under this operation, i.e. $\lambda(B) = \lambda(B + x)$.

For $x, y \in [0, 1)$, define an equivalence relation by $x \sim y$ if $x - y \in \mathbb{Q} \cap [0, 1)$. Using the axiom of choice, we can select one representative from each equivalence class and form the set S. We will show $S \notin \mathcal{B}$.

Note that the sets $\{S + q \mid q \in \mathbb{Q} \cap [0, 1)\}$ are all disjoint, with union $[0, 1)$. If S was a Borel set, then so would $S + q$. Moreover, by the translation invariance of λ, and countable additivity,

$$1 = \lambda([0, 1)) = \sum_{q \in \mathbb{Q} \cap [0,1)} \lambda(S + q) = \sum_{q \in \mathbb{Q} \cap [0,1)} \lambda(S),$$

giving a contradiction. So $S \notin \mathcal{B}$.

We can extend the above proof to show that $S \notin M_{\mathrm{leb}}$.

Theorem 3.2 (Banach-Kuratowski)**.** *Assuming the continuum hypothesis, there exists no measure μ on $\mathbb{P}([0, 1))$ such that*

$$\mu([0, 1)) = 1, \qquad \mu(\{x\}) = 0$$

for all $x \in [0, 1)$.

The proof is by Dudley.

Henceforth, whenever we are on a metric space E, we will work with $\mathcal{B}(E)$, which will be perfectly satisfactory.

4   Probability Measures

We have seen in previous courses the axiomatic definition of probability: we take a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where

- Ω is the set of possible outcomes in the sample space.

- $\mathcal{F}$ is the set of events.

- $\mathbb{P}$ is the probability function that satisfies a few properties.

In the following, we take $(\Omega, \mathcal{F}, \mathbb{P})$ to be $(E, \mathcal{E}, \mu)$, where μ is a probability measure. The following axioms were set by Kolmogorov:

1. $\mathbb{P}(\Omega) = 1$, $\mathbb{P}(\emptyset) = 0$.

2. $\mathbb{P}(E) \geq 0$ for all $E \in \mathcal{F}$.

3. $\mathbb{P}(\bigcup A_n) = \sum \mathbb{P}(A_n)$ for all A_n disjoint.

This implies that $\mathbb{P}$ is a measure. From this, we get the following properties about (finite) measures:

- $\mathbb{P}(\bigcup A_n) \leq \sum \mathbb{P}(A_n)$.

- $A_n \uparrow A \implies \mathbb{P}(A_n) \uparrow \mathbb{P}(A)$.

- $A_n \downarrow A \implies \mathbb{P}(A_n) \downarrow \mathbb{P}(A)$.

Definition 4.1 (Independence)**.** We say that $\{A_i \mid i \in I\}$, $A_i \in \mathcal{F}$ are *independent* if for all finite sets $J \subseteq I$, we have

$$\mathbb{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbb{P}(A_j).$$

We say that the σ-algebras $\{\mathcal{A}_i \mid i \in I\}$, $\mathcal{A}_i \subseteq \mathcal{F}$ are *independent* if $\{A_i \mid i \in I\}$ are independent for all $A_i \in \mathcal{A}_i$.

Theorem 4.1. *Let $\mathcal{A}_1, \mathcal{A}_2$ be π-systems contained in $\mathcal{F}$, such that*

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$$

for all $A_1 \in \mathcal{A}_1$, $A_2 \in \mathcal{A}_2$. Then $\sigma(\mathcal{A}_1)$ and $\sigma(\mathcal{A}_2)$ are independent.

Proof: Fix $A_1 \in \mathcal{A}_1$, and define for $A \in \sigma(\mathcal{A}_2)$,

$$\mu(A) = \mathbb{P}(A_1 \cap A), \qquad \nu(A) = \mathbb{P}(A_1)\mathbb{P}(A).$$

Then μ and ν are finite measures, and they agree on the π-system $\mathcal{A}_2$. Hence

---

by the uniqueness theorem, $\mu(A) = \nu(A)$ for all $A \in \sigma(\mathcal{A}_2)$.

Repeat the same argument, now by fixing $A_2 \in \sigma(\mathcal{A}_2)$. Then define

$$\mu'(A) = \mathbb{P}(A \cap A_2), \qquad \nu'(A) = \mathbb{P}(A)\mathbb{P}(A_2).$$

Then $\mu' = \nu'$ on $\mathcal{A}_1$ by the above, and hence by uniqueness $\mu' = \nu'$ on $\sigma(\mathcal{A}_1)$.

4.1 Borel-Cantelli Lemmas

Given a sequence of events (A_n), we may ask for the probability that infinitely many of them occur. For $A_n \in \mathcal{F}$, define

$$\limsup A_n = \bigcap_{n=1}^{\infty} \bigcup_{m \geq n} A_m = \{A_n \text{ happens infinitely often}\},$$

$$\liminf A_n = \bigcup_{n=1}^{\infty} \bigcap_{m \geq n} A_m = \{A_n \text{ eventually}\}.$$

Lemma 4.1. *If* $\sum \mathbb{P}(A_n) < \infty$, *then* $\mathbb{P}(A_n \text{ i.o.}) = 0$.

Proof:　Fix any $n \in \mathbb{N}$. Then,

$$0 \leq \mathbb{P}(A_m \text{ i.o.}) \leq \mathbb{P}\left(\bigcup_{m \geq n} A_m \right) \leq \sum_{m \geq n} \mathbb{P}(A_m).$$

Taking the limit as $n \to \infty$, $\mathbb{P}(A_n \text{ i.o.}) = 0$.

Note that the above lemma holds for any measure μ. We will now prove a partial converse.

Lemma 4.2. *Assume the events* (A_n) *are independent. Then, if* $\sum_n \mathbb{P}(A_n) = \infty$, *then* $\mathbb{P}(A_n \text{ i.o.}) = 1$.

Proof:　We will use the inequality $1 - a \leq e^{-a}$, for all $a \geq 0$. Now as (A_n) are independent, (A_n^c) are independent, so for all n and $N \geq n$,

$$0 \leq \mathbb{P}\left(\bigcap_{m=n}^{N} A_m^c \right) = \prod_{m=n}^{N} \mathbb{P}(A_m^c) = \prod_{m=n}^{N} (1 - \mathbb{P}(A_m)) \leq \exp\left(-\sum_{m=n}^{N} \mathbb{P}(A_m) \right).$$

Taking $N \to \infty$,

$$0 \leq \mathbb{P}\left(\bigcap_{m=n}^{\infty} A_m^c\right) \leq \lim_{N \to \infty} \mathbb{P}\left(\bigcap_{m=n}^{N} A_m^c\right) \leq \lim_{N \to \infty} \exp\left(-\sum_{m=n}^{N} \mathbb{P}(A_m)\right)$$

$$= \exp\left(-\sum_{m=n}^{\infty} \mathbb{P}(A_m)\right) = 0.$$

So we get

$$\mathbb{P}\left(\bigcap_{m=n}^{\infty} A_m^c\right) = 0 \implies \mathbb{P}\left(\bigcup_{m=n}^{\infty} A_m\right) = 1$$

for all n. Now let us define

$$\bigcup_{m=n}^{\infty} A_m = B_n,$$

then B_n decreases to

$$B_n \downarrow \bigcap_n B_n = \bigcap_n \bigcup_{m \geq n} A_m = \{A_n \text{ i.o.}\}.$$

So, as $\mathbb{P}(B_n) = 1$, $\mathbb{P}(A_n \text{ i.o.}) = \lim_{n \to \infty} \mathbb{P}(B_n) = 1$.

Remark. If the (A_n) is independent, then $\{A_n \text{ i.o.}\}$ is a 0/1 event. For all such tail events, the probability is 0 or 1. This is a special case of Kolmogorov's 0-1 law, which will be proven later.

5   Measurable Functions

Let $(E, \mathcal{E})$ and $(G, \mathcal{G})$ be two measurable spaces. A map $f : E \to G$ is called *measurable* if $f^{-1}(A) \in \mathcal{E}$ for all $A \in \mathcal{G}$.

When $(G, \mathcal{G}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we simply say that f is measurable. If E is a topological space and $\mathcal{E} = \mathcal{B}(E)$, then f is called *Borel*.

Note that preimages preserve set operations:

$$f^{-1}\left(\bigcup_{i \in I} A_i\right) = \bigcup_{i \in I} f^{-1}(A_i), \qquad f^{-1}(G \setminus A) = E \setminus f^{-1}(A).$$

So for any function f, $\{f^{-1}(A) \mid A \in \mathcal{G}\}$ is a σ-algebra on E, and $\{A \subset G \mid f^{-1}(A) \in \mathcal{E}\}$ is a σ-algebra on G.

Example 5.1.

1. If $\mathcal{G} = \sigma(\mathcal{A})$ and $f^{-1}(A) \in \mathcal{E}$ for all $A \in \mathcal{A}$, then $\{A \subset G \mid f^{-1}(A) \in \mathcal{E}\}$ is a σ-algebra containing $\mathcal{A}$, hence it contains $\sigma(\mathcal{A}) = \mathcal{G}$. Therefore f is measurable.

2. In particular, when $G = \mathbb{R}$, $\mathcal{G} = \mathcal{B}$, then $\mathcal{B} = \sigma(\mathcal{A})$ where $\mathcal{A} = \{(-\infty, y] \mid y \in \mathbb{R}\}$. Hence a function is Borel measurable if and only if $\{x \in E \mid f(x) < y\} \in \mathcal{E}$ for all $\in \mathbb{R}$.

3. If E is a topological space, and $f : E \to \mathbb{R}$ is continuous, then for $\mathcal{A} = \{U \mid U \text{ open}\}$, $f^{-1}(U) \in \mathcal{E}$. So f is Borel measurable.

4. For $A \subset E$, the indicator function $\mathbb{1}_A(x) : E \to \{0, 1\}$ is measurable if and only if $A \in \mathcal{E}$.

5. We can easily prove the composition of measurable functions is measurable.

6. For a family of functions $f_i : E \to G$, $i \in I$ we can make all (f_i) measurable with respect to the σ-algebra

$$\mathcal{E} = \sigma(f_i^{-1}(A) \mid A \in \mathcal{G}, i \in I).$$

$\mathcal{E}$ is called the σ-algebra generated by $\{f_i\}$.

Proposition 5.1. *If $f_1, f_2, \ldots$ are measurable real-valued functions, then*

$$f_1 + f_2, \ f_1 f_2, \ \inf_n f_n, \ \sup_n f_n, \ \liminf_n f_n, \ \limsup_n f_n$$

are all measurable.

Theorem 5.1 (Monotone Class Theorem)**.** *Let $(E, \mathcal{E})$ be a measurable space and $\mathcal{A}$ a π-system generating $\mathcal{E}$. Let $\mathcal{V}$ be a vector space of bounded functions $f : E \to \mathbb{R}$ such that:*

- *$1 \in \mathcal{V}$ and $1_A \in \mathcal{V}$ for all $A \in \mathcal{A}$.*

- *If $f_n \in V$ for all n and f is bounded with $0 \le f_n \uparrow f$, then $f \in V$.*

Then $\mathcal{V}$ contains all bounded measurable functions.

Proof: Let $\mathcal{D} = \{A \in \mathcal{E} \mid 1_A \in \mathcal{V}\}$. Then $\mathcal{D}$ is a d-system:

- $1 = 1_E \in \mathcal{V}$, so $E \in \mathcal{D}$.

- If $A \subset B$, $A, B \in \mathcal{D}$ then $1_{B \setminus A} = 1_B - 1_A \in \mathcal{V}$ as $\mathcal{V}$ is a vector space. Hence $B \setminus A \in \mathcal{D}$.

- If $A_n \uparrow A$ for $A_n \in \mathcal{D}$ then $1_{A_n} \uparrow 1_A$, hence $1_A \in \mathcal{V}$ and $A \in \mathcal{D}$.

It contains the π-system $\mathcal{A}$, so by Dynkin's lemma it contains $\sigma(\mathcal{A}) = \mathcal{E}$. Therefore $\mathcal{D} = \mathcal{E}$, and so $1_A \in \mathcal{V}$ for all $A \in \mathcal{E}$.

Since $\mathcal{V}$ is a vector space, it contains all finite linear combinations of measurable sets. So,
$$f_n = 2^{-n} \lfloor 2^n f \rfloor \in \mathcal{V}.$$

For any bounded non-negative measurable function, define

$$f_n(x) = 2^{-n} \lfloor 2^n f(x) \rfloor = 2^{-n} \sum_{j=0}^{\infty} j \mathbb{1}_{2^n f(x) \in [j, j+1)}$$

$$= 2^{-n} \sum_{j=0}^{K_n} j \mathbb{1}_{f^{-1}[j/2^n, (j+1)/2^n)}$$

for some finite K_n as f is bounded. Then, $f_n \le f \le f_n + 2^{-n}$, and f_n is measurable.

Therefore as $0 \le f_n \uparrow f$, $f_n \in \mathcal{V}$ and f is bounded and non-negative, then $f \in \mathcal{V}$.

Finally for general f bounded and measurable, $f = f^+ - f^-$, so as f^+, f^- are bounded, non-negative and measurable, they are in $\mathcal{V}$. As $\mathcal{V}$ is a vector space, so is f.

5.1   Image Measure

et $(E, \mathcal{E})$ and $(G, \mathcal{G})$ be two measurable spaces, and $f : E \to G$ a measurable function. Let μ be a measure on $(E, \mathcal{E})$.

Then μ induces an *image measure* or *pushforward measure* ν on G given by

$$\nu = \mu \circ f^{-1}, \text{ i.e. } \nu(A) = \mu(f^{-1}(A)).$$

This is well-defined and ν is a measure.

We will show that starting from a Lebesgue measure and taking an appropriate measurable function, we can get all probability measures (and Radon measures).

Lemma 5.1. *Let $g : \mathbb{R} \to \mathbb{R}$ be non-constant, right-continuous and increasing. Let*

$$g(\pm\infty) = \lim_{x \to \pm\infty} g(x), \qquad I = (g(-\infty), g(\infty)).$$

Define $f : I \to \mathbb{R}$ by

$$f(x) = \inf\{g \in \mathbb{R} \mid g(y) \geq x\}.$$

Then f is left-continuous, increasing and for all $x \in I, y \in \mathbb{R}$,

$$f(x) \leq y \iff x \leq g(y).$$

f is called a generalized inverse of g.

> **Proof:**   Fix $x \in I$. Define $J_x = \{y \in \mathbb{R} \mid g(y) \geq x\}$. Then J_x is non-empty and bounded below, hence $f(x)$ is defined.
>
> As g is increasing, if $y \in J_x$ and $y' \geq y$, then $g(y') \geq g(y) \geq x$, i.e $y' \in J_x$. Hence as g is right-continuous, if $y_n \in J_x$ with $y_n \downarrow y$, then $g(y) = \lim g(y_n) \geq x$, so $y \in J_x$.
>
> So $J_x = [f(x), \infty)$, and $x \leq g(y) \iff u \in J_x \iff f(x) \leq y$.
>
> Now if $x \leq x'$, we have $J_x \supseteq J_{x'}$, as $y \in J_{x'} \implies y \in J_x$. Thus $[f(x), \infty) \supseteq [f(x'), \infty)$, so $f(x) \leq f(x')$ and f is increasing.
>
> To show f is left continuous, for $x_n \uparrow x$, then $J_x = \bigcap J_{x_n}$. So $[f(x), \infty) = \bigcap_n [f(x_n), \infty)$ and $f(x_n) \to f(x)$.

Theorem 5.2. *Let $g : \mathbb{R} \to \mathbb{R}$ be non-constant, right-continuous and increasing. Then there exists a unique Radon measure μ_g on $\mathbb{R}$ such that, for all $a, b \in \mathbb{R}$ with $a < b$,*

$$\mu_g((a, b]) = g(b) - g(a),$$

and every Radon measure can be obtained in this way.

This measure μ_g is called the *Lebesgue-Stieltjes measure* associated with g.

Proof: Define I, f as in the previous lemma, and let λ be the Lebesgue measure on I. Then note f is Borel measurable since

$$f^{-1}((-\infty, z]) = \{x \in I \mid f(x) \leq z\} = \{x \in I \mid x \leq g(z)\} = (g(-\infty), g(z)] \in \mathcal{B}.$$

Since $\{(-\infty, z] \mid z \in \mathbb{R}\}$ generates $\mathcal{B}$, f is measurable.

Then the induced measure $\mu_g = \lambda \circ f^{-1}$ exists on $\mathcal{B}$, where $\mu_g(A) = \lambda(f^{-1}(A))$. Then

$$\mu_g((a, b]) = \lambda(f^{-1}(a, b]) = \lambda(\{x \mid a < f(x) \leq b\}) = \lambda(\{x \mid g(a) < x \leq g(b)\})$$
$$= \lambda((g(a), g(b)]) = g(b) - g(a).$$

By the uniqueness of extensions for σ-finite measures, μ_g is uniquely defined.

Conversely, if ν is any Radon measure on $\mathbb{R}$, define $g : \mathbb{R} \to \mathbb{R}$ as

$$g(y) = \begin{cases} \nu((0, y]) & y \geq 0, \\ -\nu((y, 0]) & y > 0. \end{cases}$$

As ν is Radon, g is well-defined. We can easily check that g is right continuous and g is increasing.

Lastly, $\nu((a, b]) = g(b) - g(a)$.

Example 5.2.

Fix $x \in \mathbb{R}$ and take $g = \mathbb{1}_{[x, \infty)}$. Then $\mu_g = \delta_x$, the Dirac measure at x, defined such that

$$\delta_x(A) = \begin{cases} 1 & x \in A, \\ 0 & x \notin A. \end{cases}$$

5.2 Random Variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(E, \mathcal{E})$ be a measurable space, with $X : \Omega \to E$ a measurable function. Then X is called a *random variable* in E.

The image measure $\mu_x = \mathbb{P} \circ X^{-1}$ is called the *law* or *distribution* of X. It is a measure on $(E, \mathcal{E})$.

If $E = \mathbb{R}$, then μ_x is uniquely determined by its values on the π-system $\{(-\infty, x] :$

$x \in \mathbb{R}\}$, given by

$$F_X(x) = \mu_x((-\infty, x]) = \mathbb{P} \circ X^{-1}((\infty, x]) = \mathbb{P}(\omega \in \Omega : X(\omega) \leq x) = \mathbb{P}(X \leq x).$$

The function F_X is called the *distribution function* of X, because it characterizes the distribution of X. By properties of probability measures,

1. F_X is increasing.

2. F_X is right continuous.

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

Any function $F : \mathbb{R} \to [0, 1]$ satisfying these conditions is called a distribution function.

Given any distribution function F, there exists a random variable X, such that $F = F_X$, i.e.

$$F(x) = F_X(x) = \mathbb{P}(X \leq x),$$

for all $x \in \mathbb{R}$.

Proof: Let $\Omega = (0, 1)$ and $\mathbb{P}$ be the Lebesgue measure $\lambda|_{(0,1)}$. Then lets consider a random variable on $(\Omega, \mathbb{B}(\Omega), \mathbb{P})$.

Let F be any distribution function. Then F is increasing and right continuous, so we can define

$$X(\omega) = \inf\{x \mid \omega \leq F(x)\} : (0, 1) \to \mathbb{R}.$$

Since X is a measurable function, X is a random variable. Now

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\omega \in \Omega, X(\omega) \leq x) = \mathbb{P}(\omega \in \Omega, w \leq F(x))$$
$$= \mathbb{P}((0, F(x)]) = F(x).$$

Definition 5.1. A countable family of random variables $(X_i, i \in I)$ is said to be *independent* if the family of σ-algebras $(\sigma(X_i), i \in I)$ is independent.

For a sequence of random variables $(X_n, n \in \mathbb{N})$, this sequence of random variables is independent if and only if

$$\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \cdots \mathbb{P}(X_n \leq x_n).$$

5.3   Rademacher Functions

Given a distribution function f_n we saw how to construct a random variable X corresponding to it.

Let us try to extend this construction: given an infinite sequence of distribution functions $F_1, F_2, \ldots$, we want to find independent random variables $(X_1, X_2, \ldots)$ corresponding to them.

Let $(\Omega, \mathcal{F}, \mathbb{P}) = ((0,1), \mathbb{B}(0,1), \lambda|_{(0,1)})$. Then any $\omega \in \Omega$ has a binary expansion. If we exclude representations ending in an infinite sequence of zeroes, then this representation is unique.

Define $R_n : \Omega \to \{0,1\}$ by $R_n(\omega) = \omega_n$, then n'th decimal digit. For example,

$$R_1 = \mathbb{1}_{(1/2, 1]},$$
$$R_2 = \mathbb{1}_{(1/4, 1/2)} + \mathbb{1}_{(3/4, 1]},$$

and in general the R_n's are finite sums of indicators of intervals. Hence they are random variables, which are called *Rademacher functions*indexRademacher function.

We claim that the R_i are all iid Ber$(1/2)$. Indeed,

$$\mathbb{P}(R_1 = x_1, R_2 = x_2, \ldots, R_n = x_n) = 2^{-n} = \mathbb{P}(R_1 = x_1) \cdots \mathbb{P}(R_n = x_n),$$

hence these are independent.

Now choose a bijection $m : \mathbb{N}^2 \to \mathbb{N}$, and define $Y_{k,n} = R_{m(k,n)}$, and set

$$Y_n = \sum_{k=1}^{\infty} 2^{-k} Y_{k,n}.$$

Then we claim (Y_n) are iid $U(0,1)$. Showing independence is easy, as Y_i are measurable functions of independent variables.

The law of Y_n is identifies on the π-system of intervals $(i/2^m, (i+1)/2^m]$ for $i = 0, 1, \ldots, 2^m - 1$. Now

$$\mathbb{P}\left(\frac{i}{2^m} < Y_n \leq \frac{i+1}{2^m}\right) = \mathbb{P}\left(\frac{i}{2^m} < \sum_{k=1}^{\infty} 2^{-k} Y_{k,n} \leq \frac{i+1}{2^m}\right)$$

$$= \mathbb{P}(Y_{1,n} = y_1, \ldots, Y_{m,n} = y_m) = \prod_{i=1}^{m} \mathbb{P}(Y_{i,n} = y_i)$$

$$= 2^{-m} = \lambda\left(\frac{i}{2^m}, \frac{i+1}{2^m}\right].$$

Hence μ_{Y_n} is $\lambda|_{(0,1)}$, so Y_i are iid $U(0,1)$.

Then as before, set $G_n(x) = F_n^-(x) = \inf\{y \mid x \leq F_n(y)\}$ are Borel functions, so setting $X_n = G_n(Y_n)$ we have $F_{X_n} = F_n$ and the (X_n) are independent.

5.4   Convergence

Let $(E, \mathcal{E}, \mu)$ be a measure space. Let $A \in \mathcal{E}$ be defined by some property. We say that the property holds *almost everywhere* if $\mu(A^c) = 0$.

If μ is a $\mathbb{P}$-measure, then A holds *almost surely* if $\mathbb{P}(A^c) = 0$, i.e. $\mathbb{P}(A) = 1$.

Then if $(f_n), f : (E, \mathcal{E}, \mu) \to (\mathbb{R}, \mathbb{B})$ are measurable, we say:

- $f_n \to f$ almost everywhere if $\mu(\{x \in E \mid f_n(x) \not\to f(x)\}) = 0$.

- $f_n \to f$ in μ-measure if, for all $\varepsilon > 0$, $\mu(\{x \in E \mid |f_n(x) - f(x)| > \varepsilon\}) \to 0$ as $n \to \infty$.

Theorem 5.3. *Let (f_n) be a sequence of measurable functions. Then, if $\mu(E) < \infty$, then $f_n \to 0$ almost everywhere if and only if $f_n \to 0$ in μ-measure.*

Proof: Fix $\varepsilon > 0$. Suppose $f_n \to 0$ almost everywhere. Then,

$$\mu(|f_n| \leq \varepsilon) \geq \mu\left(\bigcap_{m=n}^{\infty} \{|f_m| < \varepsilon\}\right) \uparrow \mu(|f_n| \leq \varepsilon \text{ eventually})$$

$$\geq \mu(f_n \to 0) = \mu(E) < \infty.$$

Hence we get

$$\lim_{n \to \infty} \mu(|f_n| \leq \varepsilon) = \mu(E),$$

so $f_n \to 0$ in μ-measure. The other direction is easy.

Theorem 5.4. *If $f_n \to 0$ in μ-measure, then there exists a subsequence (n_k) such that $f_{n_k} \to 0$ μ almost everywhere.*

Proof: Suppose $f_n \overset{\mu}{\to} 0$. Choosing $\varepsilon = 1/k$, then $\mu(|f_n| > 1/k) \to 0$ as $n \to \infty$. Hence we can get n_k such that

$$\mu\left(|f_{n_k}| > \frac{1}{k}\right) < \frac{1}{k^2}.$$

Choosing this for all k, we get a subsequence (n_k). Moreover,

$$\sum_k \mu\left(|f_{n_k}| > \frac{1}{k}\right) \leq \sum_k \frac{1}{k^2} < \infty.$$

Hence by Borel-Cantelli,

$$\mu\left(|f_{n_k}| > \frac{1}{k} \text{ i.o.}\right) = 0 \implies \mu(f_{n_k} \not\to 0) = 0,$$

hence $f_{n_k} \to 0$ μ almost everywhere.

Note that going to a subsequence is necessary, so convergence in μ-measure does not imply convergence almost everywhere.

For example, let (A_n) be independent with $\mathbb{P}(A_n) = 1/n$, and let $X_n = \mathbb{1}_{A_n}$. Then $X_n \xrightarrow{\mathbb{P}} 0$, but $\sum \mathbb{P}(A_n) = \infty$, so as they are independent, by Borel-Cantelli $\mathbb{P}(X_n > \varepsilon \text{ i.o.}) = 1$, hence $X_n \not\to 0$ almost everywhere.

For X, (X_n) a sequence of random variables, we say $X_n \xrightarrow{d} X$, or X_n converges to X in distribution if

$$F_{X_n}(t) \overset{n\to\infty}{\Longrightarrow} F_X(t)$$

for all t such that $t \mapsto F_X(t)$ is continuous.

Remark. If $X_n \xrightarrow{\mathbb{P}} X$, then $X_n \xrightarrow{d} X$.

Example 5.3.

Let (X_n) be iid $\mathrm{Exp}(1)$, so $\mathbb{P}(X_n > x) = e^{-x}$, for all $n \in \mathbb{N}$ and $x \geq 0$.

We will find a function $g : \mathbb{N} \to \mathbb{R}$ such that

$$\limsup \frac{X_n}{g(n)} = 1.$$

Note that for $\alpha > 0$,

$$\mathbb{P}(X_n > \alpha \log n) = e^{-\alpha \log n} = n^{-\alpha},$$

hence we get

$$\sum_n \mathbb{P}(X_n > \alpha \log n) < \infty \iff \alpha > 1.$$

So by Borel-Cantelli,

$$\mathbb{P}\left(\frac{X_n}{\log n} > 1 + \varepsilon \text{ i.o.}\right) = 0.$$

Also by Borel-Cantelli, as $\{X_n > \log n\}$ are independent events,

$$\mathbb{P}\left(\frac{X_n}{\log n} > 1 \text{ i.o.}\right) = 1,$$

hence
$$\limsup \frac{X_n}{\log n} = 1 \text{ a.s.}$$

5.5   Tail Events

Let (X_n) be a sequence of random variables. Define

$$\tau_n = \sigma\{X_{n+1}, X_{n+2}, \ldots\}, \qquad \tau = \bigcap_{n \in \mathbb{N}} \tau_n.$$

Then τ is a σ-algebra called the *tail σ-algebra*. It contains events that depend only on the limiting behaviour of the sequence.

Theorem 5.5 (Kolmogorov 0-1 law). *Let (X_n) be a sequence of independent random variables. Then for the tail σ-algebra τ, if $A \in \tau$, then $\mathbb{P}(A) = 0$ or 1.*

Hence if $Y : (\Omega, \tau) \to \mathbb{R}$ is measurable, then Y is almost surely constant.

Proof: Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Then $\mathcal{F}_n$ is generated by the π-system of sets

$$A = \{X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\},$$

and $\tau_n = \sigma(X_{n+1}, \ldots)$ is generated by the π-system of events

$$B = \{X_{n+1} \leq x_{n+1}, \ldots, X_{n+k} \leq x_{n+k}\},$$

for $k \in \mathbb{N}$. But by independence, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for all such A, B. Hence, by an earlier theorem $\mathcal{F}_n$ and τ_n are independent. But $\tau \subset \tau_n$. So $\mathcal{F}_n$ and τ_n are independent for all n.

Now consider $\bigcup \mathcal{F}_n$. This is a π-system that generates

$$\mathcal{F}_\infty = \sigma(X_n, n \in \mathbb{N}).$$

But $\bigcup \mathcal{F}_n$ and τ are independent, so using this theorem again, $\mathcal{F}_\infty$ and τ are independent.

However, $\tau \subseteq \mathcal{F}_\infty$. So for any $A \in \tau$ and $A \in \mathcal{F}_\infty$,

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

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

Finally, if Y is τ measurable, then $\{Y \leq y\} \in \tau$, so $\mathbb{P}(Y \leq y) = 0$ or 1. Then let

$$c = \inf\{y \mid \mathbb{P}(Y \leq y) = 1\}.$$

Then $\mathbb{P}(Y = c) = 1$.

Let X_i be iid, with $\mathbb{E}[X] < \infty$, then

$$\limsup \frac{\sum_{i=1}^n X_i}{n}, \qquad \liminf \frac{\sum_{i=1}^n X_i}{n}$$

are constants almost surely.

6   Integration

For $(E, \mathcal{E}, \mu)$ a measure space and $f : E \to \mathbb{R}$ measurable, and $f \geq 0$, we shall define the *integral* of f and write it as

$$\mu(f) = \int_E f \, \mathrm{d}\mu = \int_E f(x) \, \mathrm{d}\mu(x).$$

When $(E, \mathcal{E}, \mu) = (\mathbb{R}, \mathcal{B}, \lambda)$, we write it as $\int f(x) \, \mathrm{d}x$. For $(E, \mathcal{E}, \mu) = (\Omega, \mathcal{F}, \mathbb{P})$ and X a random variable, we define its *expectation*

$$\mathbb{E}[X] = \int_\Omega X \, \mathrm{d}\mathbb{P} = \int_\Omega X(\omega) \, \mathrm{d}\mathbb{P}(\omega).$$

To start, we say $f : E \to \mathbb{R}$ is *simple* if

$$f = \sum_{k=1}^m a_k \mathbb{1}_{A_k},$$

for $0 \leq a_k < \infty$ and $A_k \in \mathcal{E}$. Define the integral for such simple f as

$$\mu(f) = \sum_{k=1}^m a_k \mu(A_k).$$

This is well-defined: we can check for f, g simple, and $\alpha, \beta > 0$ that:

(a) $\mu(\alpha f + \beta g) = \alpha \mu(f) + \beta \mu(g)$,

(b) $f \leq g \implies \mu(f) \leq \mu(g)$,

(c) $f = 0$ a.e. $\iff \mu(f) = 0$.

> **Example 6.1.**
>
> We saw that if $X_n \xrightarrow{\mathbb{P}} X$, then $X_n \xrightarrow{d} X$. However, the converse is not true. Let $(\Omega, \mathcal{F}, \mathbb{P}) = ((0, 1], \mathcal{B}, \lambda)$, and
>
> $$f_1 = \mathbb{1}_{(0,1/2]}, \qquad f_2 = \mathbb{1}_{(1/2,1]}, \qquad f_3 = \mathbb{1}_{(0,1/3]},$$
> $$f_4 = \mathbb{1}_{(1/3,2/3]}, \qquad f_5 = \mathbb{1}_{(2/3,1]}, \qquad f_6 = \mathbb{1}_{(0,1/4]},$$
>
> and so on. Then $f_n \to 0$ in λ measure, but $f_n \not\to 0$ λ-almost everywhere.

Definition 6.1. For $f : E \to \mathbb{R}$ measurable with $f \geq 0$, define

$$\mu(f) = \sup\{\mu(g) \mid g \text{ simple}, g \leq f\}.$$

Clearly if $0 \leq f_1 \leq f_2$, then $\mu(f_1) \leq \mu(f_2)$.

For general $f : E \to \mathbb{R}$ measurable, let $f = f^+ - f^-$, and $|f| = f^+ + f^-$. We say f is *integrable* if $\mu(|f|) < \infty$, and define

$$\mu(f) = \mu(f^+) - \mu(f^-).$$

If one of $\mu(f^+)$ or $\mu(f^-)$ is ∞ and the other is finite, then we define $\mu(f)$ to be ∞ or $-\infty$ respectively. However f is not integrable.

Theorem 6.1 (Monotone Convergence Theorem). *Let $(f_n), f : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be measurable and non-negative.*

Suppose that $f_n \uparrow f$. Then $\mu(f_n) \uparrow \mu(f)$.

Proof: Recall that $\mu(f) = \sup\{\mu(g) \mid g \leq f, g \text{ simple}\}$. Letting $M = \sup_n(\mu(f_n))$, we have that $\mu(f_n) \uparrow M$, so we need to show that $M = \mu(f)$.

As $f_n \leq f$ for all n, $\mu(f_n) \leq \mu(f)$, hence by taking the supremum, $M \leq \mu(f)$.

Next we will show that $M \geq \mu(f)$. It is enough to show that $M \geq \mu(g)$, for all g simple with $g \leq f$. Let

$$g = \sum_{k=1}^{m} a_k \mathbb{1}_{A_k} \leq f,$$

and assume, without loss of generality, that (A_k) are disjoint. Define the approximation g_n as

$$g_n(x) = \underbrace{(2^{-n} \lfloor 2^n f_n(x) \rfloor)}_{\bar{f}_n} \wedge g(x).$$

Then g_n is simple with $g_n \leq \bar{f}_n \leq f_n \uparrow f$, and

$$g_n = \bar{f}_n \wedge g \uparrow f \wedge g = g.$$

This gives $g_n \uparrow g$ with g_n simple, $g_n \leq f_n$. Fix $\varepsilon \in (0, 1)$, and consider the sets

$$A_k(n) = \{x \in A_k \mid g_n(x) \geq (1 - \varepsilon) a_k\}.$$

Now $g = a_k$ on the set A_k, and $g_n \uparrow g$. Hence $A_k(n) \uparrow A_k$, and $\mu(A_k(n)) \uparrow \mu(A_k)$. Also,

$$\mathbb{1}_{A_k(n)}(1 - \varepsilon) a_k \leq \mathbb{1}_{A_k(n)} g_n \leq \mathbb{1}_{A_k} g_n,$$

so as $\mu(f)$ is increasing, we have

$$\mu(\mathbb{1}_{A_k(n)}(1 - \varepsilon) a_k) \leq \mu(\mathbb{1}_{A_k} g_n)$$
$$\implies (1 - \varepsilon) a_k \mu(A_k(n)) \leq \mu(\mathbb{1}_{A_k} g_n).$$

Finally, as

$$g_n = \sum_{k=1}^{m} \mathbb{1}_{A_k} g_n,$$

we get that

$$\mu(g_n) = \mu\left(\sum_{k=1}^{m} \mathbb{1}_{A_k} g_n\right) = \sum_{k=1}^{n} \mu(\mathbb{1}_{A_k} g_n)$$

$$\geq \sum_{k=1}^{m}(1-\varepsilon)a_k\mu(A_k(n)) \uparrow \sum_{k=1}^{m}(1-\varepsilon)a_k\mu(A_k) = (1-\varepsilon)\mu(g).$$

Since $(1-\varepsilon)\mu(g) \leq \lim \mu(g_n) \leq \lim \mu(f_n) \leq M$, we get that

$$\mu(g) \leq \frac{M}{1-\varepsilon}$$

for all ε, hence $\mu(g) \leq M$.

Theorem 6.2. *Let* $(f,g):(E,\mathcal{E},\mu) \to \mathbb{R}$ *be measurable and non-negative. Then for all* $\alpha, \beta \geq 0$,

(a) $\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g)$,

(b) $f \leq g \implies \mu(f) \leq \mu(g)$,

(c) $f = 0$ *a.e.* $\iff \mu(f) = 0$.

Proof: (a) Let $f_n = (2^{-n}\lfloor 2^n f\rfloor) \wedge n$, $g_n = (2^{-n}\lfloor 2^n g\rfloor) \wedge n$. Then f_n, g_n are simple with $f_n \uparrow f$, $g_n \uparrow g$.

Then $\alpha f_n + \beta g_n \uparrow \alpha f + \beta g$, so by monotone convergence theorem,

$$\mu(\alpha f_n + \beta g_n) = \alpha\mu(f_n) + \beta\mu(g_n) \uparrow \mu(\alpha f + \beta g).$$

Hence taking limits, $\alpha\mu(f) + \beta\mu(g) = \mu(\alpha f + \beta g)$.

(b) follows from definition.

For (c), if $f = 0$ almost everywhere, then if $0 \leq f_n \leq f$ for f_n simple, then $f_n = 0$ almost everywhere. So $\mu(f_n) = 0$, and hence $\mu(f) = 0$.

Conversely, if $\mu(f) = 0$, then if $f_n \uparrow f$, as $0 \leq \mu(f_n) \uparrow \mu(f) = 0$, we must have $\mu(f_n) = 0$ and hence $f_n = 0$ almost everywhere. But then as $f_n \uparrow f$, we also have $f = 0$ almost everywhere.

Theorem 6.3. *If $(f, g) : (E, \mathcal{E}, \mu) \to \mathbb{R}$ are integrable, then for all $\alpha, \beta \in \mathbb{R}$,*

(a) $\mu(\alpha f + \beta g) = \alpha \mu(f) + \beta \mu(g)$,

(b) $f \leq g \implies \mu(f) \leq \mu(g)$,

(c) $f = 0$ *a.e.* $\implies \mu(f) = 0$.

> **Proof:**   This is an exercise, which follows from letting $f = f^+ + f^-$ and $g = g^+ + g^-$.

Remark.

1. If $0 \leq f_n \uparrow f$ almost everywhere, then $\mu(f_n) \uparrow \mu(f)$.

2. From monotone convergence theorem,

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

for $0 \leq f_n \uparrow \lim f_n = f$. If $g_n \geq 0$, then applying this we get

$$\lim_{n \to \infty} \int_{k=1}^{n} g_k \, \mathrm{d}\mu = \int \left(\sum_{k=1}^{\infty} g_k \right) \mathrm{d}\mu,$$

or more elegantly,

$$\sum_{k=1}^{\infty} \mu(g_k) = \mu \left(\sum_{k=1}^{\infty} g_k \right).$$

This generalizes the countable additivity of μ to integrals of non-negative functions.

Lemma 6.1 (Fatou's Lemma). *Let $f_n : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be measurable and non-negative. Then,*

$$\mu(\liminf f_n) \leq \liminf \mu(f_n).$$

For example, if $f_n = \mathbb{1}_{(n,n+1]}$, then f_n are non-negative with $f_n \to 0$ as $n \to \infty$. Then,

$$\lim_{n \to \infty} \lambda(f_n) = 1 \geq \lambda(0) = 0.$$

> **Proof:**   For $k \geq n$,
> $$\inf_{m \geq n} f_m \leq f_k.$$

Hence,

$$\mu\left(\inf_{m \geq n} f_m\right) \leq \mu(f_k)$$

$$\implies \mu\left(\inf_{m \geq n} f_m\right) \leq \inf_{k \geq n} \mu(f_k) \leq \liminf \mu(f_k).$$

Let $g_n = \inf_{m \geq n} f_m$. Then $g_n \geq 0$ and $g_n \uparrow \sup_n g_n = \sup_n \inf_{m \geq n} f_m = \liminf f_n$, hence by monotone convergence theorem,

$$\mu(g_n) \uparrow \mu(\liminf f_n).$$

Then putting this back,

$$\mu(\liminf f_n) \leq \liminf \mu(f_n).$$

Theorem 6.4 (Dominated Convergence Theorem). *Let $(f_n) : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be measurable, and suppose*

$$|f_n| \leq g$$

for all n, for some integrable function g, and suppose that $f_n \to f$ as $n \to \infty$ on E. Then f, f_n are integrable for all n, and

$$\mu(f_n) \to \mu(f).$$

Proof: The limit function f is measurable, and $|f| \leq g$. Hence,

$$|f_n| \leq g \implies \mu(|f_n|) \leq \mu(g) < \infty,$$
$$\mu(|f|) \leq \mu(g) < \infty.$$

So f_n and f are integrable. Now, since $0 \leq g \pm f_n \to g \pm f$, by Fatou's,

$$\mu(g \pm f) \leq \liminf \mu(g \pm f_n).$$

Therefore,

$$\mu(g) + \mu(f) \leq \liminf(\mu(g) + \mu(f_n)) \leq \mu(g) + \liminf \mu(f_n),$$
$$\mu(g) - \mu(f) \leq \liminf(\mu(g) - \mu(f_n)) \leq \mu(g) - \limsup \mu(f_n).$$

As $\mu(g) < \infty$, we get

$$\mu(f) \leq \liminf \mu(f_n) \leq \limsup \mu(f_n) \leq \mu(f),$$

which proves $\mu(f_n) \to \mu(f)$.

Remark.

1. We can replace "everywhere" with "almost everywhere".

2. In fact, $\mu(|f_n - f|) \to 0$, by using dominated convergence theorem again.

3. If $(E, \mathcal{E}, \mu) = (\Omega, \mathcal{F}, \mathbb{P})$ and $X_n \to X$ $\mathbb{P}$-almost surely, and $|X_n| \le Y$ for some Y with $\mathbb{E}[Y] < \infty$, then

$$\mathbb{E}[X_n] \to \mathbb{E}[X], \qquad \mathbb{E}[|X_n - X|] \to 0.$$

   In particular, if $|X_n| \le M$ for some $M > 0$ and $M \in \mathbb{R}$, then $\mathbb{E}[|X_n - X|] \to 0$. This is the *bounded convergence theorem.*

4. If f_n are measurable on $[0, 1]$ with $|f_n| \le 1$, and $f_n \to f$ pointwise, then

$$\int f_n \, \mathrm{d}x \to \int f \, \mathrm{d}x.$$

   This is stronger than the Riemann integral, which required uniform convergence.

In fact, we can generalise this to $\mathbb{P}$ convergence:

Theorem 6.5. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $(X_n), X$ be random variables. Suppose $X_n \to X$ in $\mathbb{P}$-measure, and suppose $|X_n| \le Y$ for some integrable random variable Y.*

Then $\mathbb{E}[|X_n - X|] \to 0$ as $n \to \infty$.

Proof: Suppose $\mathbb{E}[|X_n - X|] \not\to 0$. Then there exists a subsequence (n_k) such that $\mathbb{E}[|X_{n_k} - X|] > \varepsilon$ for all k, for some $\varepsilon > 0$. Now $X_n \xrightarrow{\mathbb{P}} X$ implies $X_{n_k} \xrightarrow{\mathbb{P}} X$, hence there exists (n_{k_l}) such that $X_{n_{k_l}} \xrightarrow{\text{a.s.}} X$.

But then by dominated convergence theorem, $\mathbb{E}[|X_{n_{k_l}} - X|] \to 0$, which contradicts our assumptions.

6.1 Comparisons with the Riemann Integral

Let's look at the fundamental theorem of calculus first. With the Riemann integral, we have the following:

(a) Generalizing the fundamental theorem of calculus:

(1) Let $f : [a, b] \to \mathbb{R}$ be continuous, and set

$$F(t) = \int_a^t f(x) \, \mathrm{d}x.$$

Then F is differentiable on $[a, b]$ with $F' = f$.

(2) Let $F : [a, b] \to \mathbb{R}$ be differentiable and F' be continuous. Then

$$\int_a^b F'(x) \, \mathrm{d}x = F(b) - F(a).$$

(b) For Lebesgue integral functions, we can also say the following:

If $f : [a, b] \to \mathbb{R}$ is integrable, and

$$F(t) = \int_a^t f(x) \, \mathrm{d}x,$$

then

$$\lim_{h \to 0} \frac{F(t + h) - F(t)}{h} = \lim_{h \to 0} \frac{\int_t^{t+h} f(t) \, \mathrm{d}x}{h} = f(t),$$

almost everywhere.

(c) We also have the substitution formula: let $\phi : [a, b] \to \mathbb{R}$ be such that ϕ is increasing and continuously differentiable. Then for all Borel functions g, with $g \geq 0$ on $[\phi(a), \phi(b)]$ we have

$$\int_{\phi(a)}^{\phi(b)} g(y) \, \mathrm{d}y = \int_a^b g(\phi(x))\phi'(x) \, \mathrm{d}x.$$

> **Proof:** Let $\mathcal{V}$ be the set of all measurable functions g for which this condition holds. Then by linearity of integral, $\mathcal{V}$ is a vector space. Now,
>
> - $1 \in \mathcal{V}$ by the fundamental theorem of calculus, and similarly so is $\mathbb{1}_{(c,d]}$.
>
> - If $f_n \in \mathcal{V}$, and $f_n \uparrow f$ with $f_n \geq 0$, then by monotone convergence theorem, $f \in \mathcal{V}$.
>
> Therefore by monotone class theorem, this holds for all $g \geq 0$ measurable.

(d) A (bounded) Riemann integrable function $f : [a, b] \to \mathbb{R}$ is Lebesgue integrable in the following sense: if f is bounded on $[a, b]$, then f is Riemann integrable if and only if

$$\mathcal{D} = \{x \in [a, b] \mid f \text{ is not continuous at } x\}$$

has $\lambda(\mathcal{D}) = 0$. Hence, f is continuous almost everywhere. Such an f need not be Borel, and can be modified on a Lebesgue measure 0 set to make it Borel. That is, there exists $\tilde{f}$ Borel such that $\tilde{f} = f$ on A, with $\lambda(A^c) = 0$. Then,

$$\int \tilde{f}\,\mathrm{d}x = \int f\,\mathrm{d}x,$$

where the first integral is in the Lebesgue sense, and the latter is a Riemann integral.

(e) $\mathbb{1}_{\mathbb{Q}}$ has discontinuity set $\mathcal{D} = [0,1]$, which is clearly not null. Hence $\mathbb{1}_{\mathbb{Q}}$ is not Riemann integrable. However it is Lebesgue integrable, and as $\mathbb{1}_{\mathbb{Q}} = 0$ almost everywhere, $\lambda(1_{\mathbb{Q}}) = 0$.

Theorem 6.6. *Let $U \subseteq \mathbb{R}$ be open, and $f : U \times E \to \mathbb{R}$ such that:*

(i) *$x \mapsto f(t, x)$ is integrable for all $t \in U$.*

(ii) *$t \mapsto f(t, x)$ is differentiable for all $x \in E$.*

(iii) *There exists $g : E \to \mathbb{R}$ integrable such that for all $x \in E$ and $t \in U$,*

$$\left| \frac{\partial f}{\partial t}(t, x) \right| \le g(x).$$

Then $x \mapsto \frac{\partial f}{\partial t}(t, x)$ is integrable for all t, and $F : U \to \mathbb{R}$ defined by

$$F(t) = \int_E f(t, x)\mu(\mathrm{d}x)$$

is differentiable with

$$\frac{\mathrm{d}}{\mathrm{d}t} F(t) = \int \frac{\partial f}{\partial t}(t, x)\mu(\mathrm{d}x).$$

Proof: For $h_n \to 0$, set

$$g_n(x) = \frac{f(t + h_n, x) - f(t, x)}{h_n} - \frac{\partial f}{\partial t}(t, x).$$

For any fixed t, $g_n(x) \to 0$ for all $x \in E$, and

$$|g_n(x)| = \left| \frac{\partial f}{\partial t}(\tilde{t}, x) - \frac{\partial f}{\partial t}(t, x) \right| \le 2g(x),$$

which is integrable. Hence $\mu(g_n) \to 0$ by dominated convergence theorem,

that is,

$$\int \frac{f(t + h_n, x) - f(t, x)}{h_n} \mu(\mathrm{d}x) - \int \frac{\partial f}{\partial t}(t, x)\mu(\mathrm{d}x) \to 0$$

$$\implies F'(t) = \lim_{n \to \infty} \frac{F(t + h_n) - F(t)}{h_n} = \int \frac{\partial f}{\partial t}(t, x)\mu(\mathrm{d}x).$$

6.2   Integrals and Image Measures

Let $f : (E, \mathcal{E}, \mu) \to (G, \mathcal{G})$ be measurable with the image measure

$$\nu = \mu \circ f^{-1}$$

on $(G, \mathcal{G})$. If $g : (G, \mathcal{G}) \to \mathbb{R}$ is measurable and non-negative, then

$$\nu(g) = \int_G g(x)\,\mathrm{d}\nu(x) = \int_G g\,\mathrm{d}\mu \circ f^{-1} = \int_E g(f(x))\,\mathrm{d}\mu(x),$$

or analogously,

$$\mu \circ f^{-1}(g) = \mu(g \circ f).$$

In particular, for $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ non-negative, and g Borel and non-negative,

$$\mathbb{E}[g(X)] = \int_\Omega g(X(\omega))\,\mathrm{d}\mathbb{P}(\omega) = \int g(x)\,\mathrm{d}\mu_X(x),$$

where $\mu_X = \mathbb{P} \circ X^{-1}$ is the law of X.

6.3   Densities of Measures

For $f : (E, \mathcal{E}, \mu) \to \mathbb{R}$ measurable and non-negative, let

$$\nu(A) = \mu(f 1_A),$$

for all $A \in \mathcal{E}$. Then ν is a measure on $(E, \mathcal{E})$, as can be shown.

Then for any g measurable, with $g \geq 0$ on E, $\nu(g) = \mu(fg)$, i.e.

$$\int g\,\mathrm{d}\nu = \int gf\,\mathrm{d}\mu.$$

> **Proof:** This holds for indicator functions by definition, and it also holds for simple functions by additivity.
>
> Hence by monotone convergence theorem, it holds for non-negative measurable functions.

We say that ν has the density f with respect to μ.

In particular, for $\mu = \lambda$ and for all f Borel, there exists a Borel measure ν on $\mathbb{R}$ given by

$$\nu(A) = \int_A f(x)\,\mathrm{d}x,$$

and then for all g Borel, $g \geq 0$,

$$\nu(g) = \int f(x)g(x)\,\mathrm{d}x.$$

We say that ν has density f. Thus ν is a probability measure on $(\mathbb{R}, \mathbb{B})$ if and only if $\int f(x)\,\mathrm{d}x = 1$.

For $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$, if the law μ_x has density f_X with respect to λ, we call f_X the *probability density function* of X. Then,

$$\mathbb{P}(X \in A) = \mathbb{P} \circ X^{-1}(A) = \mu_X(A) = \int_A f_X(x)\,\mathrm{d}x,$$

for all $A \in \mathbb{B}$. Moreover for all g Borel, non-negative,

$$\mathbb{E}[g(X)] = \int g(x)\,\mathrm{d}\mu_X(x) = \int g(x)f_X(x)\,\mathrm{d}x.$$

7   Product Measures

Let $(E_1, \mathcal{E}_1, \mu_1)$ and $(E_2, \mathcal{E}_2, \mu_2)$ be two finite measure spaces. On the Cartesian product $E = E_1 \times E_2$, we can consider the set of 'rectangles':

$$\mathcal{A} = \{A_1 \times A_2 \mid A_1 \in \mathcal{E}_1, A_2 \in \mathcal{E}_2\}.$$

Then $\mathcal{A}$ is a π-system. Define the σ-algebra $\mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2 = \sigma(\mathcal{A})$.

One can show that if E_i are topological spaces with a countable basis, then

$$\mathcal{B}(E_1 \times E_2) = \mathcal{B}(E_1) \otimes \mathcal{B}(E_2).$$

Our goal is to construct a product measure on $(E_1 \times E_2, \mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2)$.

Lemma 7.1. *Let $f : E \to \mathbb{R}$ be $\mathcal{E}$-measurable. Then for all $x_1 \in E_1$, the function*

$$x_2 \mapsto f(x_1, x_2) : E_2 \to \mathbb{R}$$

is $\mathcal{E}_2$-measurable.

Proof: The above condition holds when $f = \mathbb{1}_A$, where $A = A_1 \times A_2$. Indeed,

$$\mathbb{1}_A(x_1, x_2) = \mathbb{1}_{A_1 \times A_2}(x_1, x_2) = \mathbb{1}_{A_1}(x_1) \cdot \mathbb{1}_{A_2}(x_2) = \begin{cases} \mathbb{1}_{A_2}(x_2) & x_1 \in A_1, \\ 0 & \text{else.} \end{cases}$$

Let $\mathcal{V}$ be the set of all functions for which $x_2 \mapsto f(x_1, x_2)$ is measurable for all x_1. Then f contains the constants and the indicator functions, by the above. Moreover if $f_n \in \mathcal{V}$ for all n, and $f_n \uparrow f$ then $f \in \mathcal{V}$.

Hence by monotone class theorem, $\mathcal{V}$ contains all bounded measurable functions. Hence if we let $f_n = (-n) \vee f \wedge n$, then $f_n \in \mathcal{V}$. As $f_n \to f$, we have $f \in \mathcal{V}$.

Lemma 7.2. *Let $f : (E, \mathcal{E}) \to \mathbb{R}$ be measurable. Suppose either:*

(i) *f is bounded, or*

(ii) *f is non-negative.*

Then the function

$$f_1(x_1) = \int_{E_2} f(x_1, x_2)\mu_2(\mathrm{d}x_2)$$

is $\mathcal{E}_1$-measurable, and either

(i) *bounded, or*

(ii) *non-negative, taking values in $[0, \infty]$.*

Remark. A function f taking values in $[0, \infty]$ is measurable if and only if $f^{-1}(\{\infty\}) \in \mathcal{E}$, and $f^{-1}(A) \in \mathcal{E}$ for any $A \in \mathcal{B}$.

Proof: We again use monotone class theorem. Let $f = \mathbb{1}_{A_1 \times A_2}$, then

$$f_1(x_1) = \int \mathbb{1}_{A_1 \times A_2}(x_1, x_2)\mu_2(\mathrm{d}x_2) = \mathbb{1}_{A_1}(x_1)\mu_2(A_2)$$

is $\mathcal{E}_1$-measurable. Moreover if $f_n \uparrow f$, we can use monotone convergence theorem, and the fact that the limit of measurable functions is measurable.

Then we get the result for all bounded measurable f from monotone class theorem. To extend to non-negative f we take a limit of non-negative bounded f_i.

Theorem 7.1. *There exists a unique measure $\mu = \mu_1 \otimes \mu_2$ on $\mathcal{E}$ such that*

$$\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2),$$

for all $A_1 \in \mathcal{E}_1, A_2 \in \mathcal{E}_2$.

Proof: Uniqueness follows as $\mathcal{A}$ is a π-system generating $\mathcal{E}$, and μ is a finite measure.

For existence, define the iterated integral

$$\mu(A) = \int_{E_1} \left(\int_{E_2} \mathbb{1}_A(x_1, x_2)\mu_2(\mathrm{d}x_2) \right) \mu_1(\mathrm{d}x_1).$$

This is well-defined by the previous lemmas. Clearly, $\mu(\emptyset) = 0$ and $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$. Hence it suffices to show that μ is a measure.

It is easy to see that μ is countably additive: if (A_i) are disjoint, and $A = \bigcup A_i$, then $\mathbb{1}_A = \sum \mathbb{1}_{A_i}$, so we can apply monotone convergence theorem, specifically the statement for non-negative series.

Hence μ is a measure, as desired.

Theorem 7.2 (Fubini-Tonelli). *Consider $(E, \mathcal{E}, \mu) = (E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2, \mu_1 \otimes \mu_2)$, where $\mu_i(E_i) < \infty$.*

1. *Let $f : E \to \mathbb{R}$ be measurable, and $f \geq 0$. Then,*

$$\mu(f) = \int_{E_1} \left(\int_{E_2} f(x_1, x_2) \mu_2(\mathrm{d}x_2) \right) \mu_1(\mathrm{d}x_1) \qquad (\dagger)$$

$$= \int_{E_2} \left(\int_{E_1} f(x_1, x_2) \mu_1(\mathrm{d}x_1) \right) \mu_2(\mathrm{d}x_2). \qquad (*)$$

2. *Let $f : E \to \mathbb{R}$ be μ-measurable. If we set*

$$A_1 = \left\{ x_1 \in E_1 \ \Big| \ \int_{E_2} |f(x_1, x_2)| \, \mathrm{d}\mu_2(x_2) < \infty \right\},$$

and define $f_1 : E_1 \to \mathbb{R}$ by

$$f_1(x_1) = \int_{E_2} f(x_1, x_2) \, \mathrm{d}\mu_2(x_2) \text{ for all } x_1 \in A_1,$$

and 0 otherwise, then $\mu_1(A_1^c) = 0$, and f_1 is μ-integrable with $\mu_1(f_1) = \mu(f)$.

Proof:

1. The identities $(\dagger)$ and $(*)$ hold for $f = \mathbb{1}_A$, by definition of the product measures. Hence they extend to simple functions by linearity, and for general functions $f \geq 0$ by monotone class theorem, and approximated by simple functions $\bar{f}_n$.

2. Define $L : E_1 \to [0, \infty]$ by

$$L(x_1) = \int_{E_2} |f(x_1, x_2)| \mu_2(\mathrm{d}x_2).$$

Then L is measurable and non-negative, so $A_1 = L^{-1}(\mathbb{R}) \in \mathcal{E}_1$. Hence, as

$$\int_{E_1} \left(\int_{E_2} |f(x_1, x_2)| \mu_2(\mathrm{d}x_2) \right) \mu_1(\mathrm{d}x_1) = \mu(|f|) < \infty,$$

we have $\mu_1(A_1^c) = 0$, as otherwise $\mu(|f|) = \mu_1(L) \geq \mu_1(L\mathbb{1}_{A_1^c}) = \infty$ if $\mu(A_1^c) > 0$.

Then setting

$$f_1^+(x_1) = \int_{E_2} f^+(x_1, x_2) \mu_2(\mathrm{d}x_2), \qquad f_1^-(x_1) = \int_{E_2} f^-(x_1, x_2) \mu_2(\mathrm{d}x_2),$$

we see that $f_1 = (f_1^+ - f_1^-)\mathbb{1}_{A_1} = f_1^+ \mathbb{1}_{A_1} - f_1^- \mathbb{1}_{A_1}$.

Also, as $\mu_1(f_1^+) = \mu(f^+) < \infty$ and similarly $\mu_1(f_1^-) < \infty$, we get

$$\mu(f) = \mu(f^+) - \mu(f^-) = \mu_1(f_1^+) - \mu_1(f_1^-) = \mu_1(f_1).$$

Remark.

1. The proof of 2. is symmetric in μ_i, f_i, so $\mu_1(f_1) = \mu(f) = \mu_2(f_2)$. So we can interchange the order of integrals whenever $f \geq 0$ or f is integrable.

2. This theorem extends to σ-finite measures μ.

3. Associativity is easy to check, i.e. $(\mathcal{E}_1 \otimes \mathcal{E}_2) \otimes \mathcal{E}_3 = \mathcal{E}_1 \otimes (\mathcal{E}_2 \otimes \mathcal{E}_3)$, and also $\mu_1 \otimes (\mu_2 \otimes \mu_3) = (\mu_1 \otimes \mu_2) \otimes \mu_3$.

   So we can define the n-fold product $\bigotimes \mu_i$ on $(E_1 \times \cdots \times E_n, \mathcal{E}_1 \otimes \cdots \otimes \mathcal{E}_n)$, and n-fold integrals. In particular, when $\mu = \lambda$, we have constructed the Lebesgue measure on $\mathbb{R}^n$.

Proposition 7.1. *Let $X_1, \ldots, X_n$ be random variables $X_i : (\Omega, \mathcal{F}, \mathbb{P}) \to (E_i, \mathcal{E}_i)$. Set $E = E_1 \times \cdots \times E_n$, $\mathcal{E} = \mathcal{E}_1 \otimes \cdots \otimes \mathcal{E}_n$. Consider*

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E}),$$
$$X(\omega) = (X_1(\omega), \ldots, X_n(\omega)).$$

Then X is $\mathcal{E}$-measurable and the following are equivalent:

(a) *$X_1, \ldots, X_n$ are independent.*

(b) *$\mu_X = \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}$.*

(c) *For all bounded measurable $f_i : E_i \to \mathbb{R}$,*

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

Proof: To show X is measurable, we need to show $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{E}$.

It is enough to show that $X^{-1}(A_1 \times \cdots \times A_n) \in \mathcal{F}$, where $A_i \in \mathcal{E}_i$. Then

this is

$$X^{-1}(A_1 \times \cdots \times A_n) = \{\omega \mid X_1(\omega) \in A_i, \ldots, X_n(\omega) \in A_n\} = \bigcap_{i=1}^{n} X_i^{-1}(A_i) \in \mathcal{F}.$$

(a) $\implies$ (b). First let $\nu = \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}$. It is enough to show that μ_X and ν agree on $A_1 \times \cdots \times A_n$, where $A_i \in \mathcal{E}_i$. Note that

$$\mu_X(A_1 \times \cdots \times A_n) = \mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n)$$
$$= \mu_{X_1}(A_1) \cdots \mu_{X_n}(A_n) = \nu(A).$$

Since $\mathcal{A}$ is a π-system generating $\mathcal{E}$, $\mu_X = \nu$.

(b) $\implies$ (a). Note that, by Fubini,

$$\mathbb{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] = \int \prod_{i=1}^{n} f_i(x_i) \, \mathrm{d}\mu_X(x) = \prod_{i=1}^{n} \int f_i(x_i) \, \mathrm{d}\mu_{X_i}(x_i) = \prod_{i=1}^{n} \mathbb{E}[f_i(X_i)].$$

(c) $\implies$ (a). Use $f_i = \mathbb{1}_{A_i}$, for $A_i \in \mathcal{E}_i$. Then,

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{E}[\mathbb{1}_{A_1 \times \cdots \times A_n}(X)] = \mathbb{E}\left[\prod_{i=1}^{n} \mathbb{1}_{A_i}(X_i)\right]$$
$$= \prod_{i=1}^{n} \mathbb{E}[\mathbb{1}_{A_i}(X_i)] = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).$$

8 L^p Spaces, Norms and Inequalities

Recall that a norm on a real vector space V is a function $\|\cdot\| : V \to [0, \infty)$ such that:

(i) $\|\lambda v\| = |\lambda| \|v\|$ for all $\lambda \in \mathbb{R}$, $v \in V$.

(ii) $\|v + w\| \le \|v\| + \|w\|$, for all $v, w \in V$.

(iii) $\|v\| = 0 \iff v = 0$.

8.1 L^p Norms

For $(E, \mathcal{E}, \mu)$ a measure space and $1 \le p \le \infty$, define

$$L^p = L^p(E, \mathcal{E}, \mu) = \{f : E \to \mathbb{R} \text{ measurable} \mid \|f\|_p < \infty\},$$

where

$$\|f\|_p = \left(\int |f(x)|^p \, \mathrm{d}\mu(x) \right)^{1/p}, \qquad \|f\|_\infty = \sup |f|.$$

Let us examine the conditions to be a norm.

(i) This holds true for $1 \le p < \infty$ by linearity of integral, and for $p = \infty$ it is clear.

(ii) This holds for $p = 1$ and ∞ easily. For $1 < p < \infty$, we can show this holds by Minkowski's inequality.

(iii) If $f = 0$, then clearly $\|f\|_p = 0$. But $\|f\|_p = 0$ only implies $f = 0$ almost everywhere.

We can fix this by defining the equivalence classes

$$[f] = \{g \mid g = f \text{ almost everywhere}\},$$

and then defining

$$\mathcal{L}^p = \{[f] \mid f \in L^p\}.$$

Then $\mathcal{L}^p$, for $1 \le p \le \infty$ are normed vector spaces.

8.2   Inequalities

We start off with a classic.

Theorem 8.1 (Markov/Chebyshev's inequality)**.** *Let $f \geq 0$ be measurable. Then for all $\lambda > 0$,*

$$\mu(\{x \in E \mid f(x) \geq \lambda\}) \leq \frac{\mu(f)}{\lambda}.$$

Proof: Note that $\lambda \mathbb{1}_{\{f \geq \lambda\}} \leq f$. Hence integrating with respect to μ, we get

$$\lambda \mu(f \geq \lambda) \leq \mu(f).$$

In particular, if $g \in L^p$ for $p < \infty$, then considering $f = |g|^p$ gives, for all $\lambda > 0$,

$$\mu(|g| \geq \lambda) \leq \frac{\mu(|g|^p)}{\lambda^p} < \infty.$$

This gives us tail estimates as $\lambda \to \infty$.

We now move on to another classic. Before that, we need to introduce convexity again.

Definition 8.1. For $I \subseteq \mathbb{R}$ an interval, say a function $f : I \to \mathbb{R}$ is *convex* if for all $x, y \in I$, and $t \in [0, 1]$,

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y).$$

Equivalently, for $x < z < y$,

$$\frac{f(z) - f(x)}{z - x} \leq \frac{f(y) - f(z)}{y - z}.$$

In particular, we can show that such an f is continuous on I, hence Borel measurable.

Lemma 8.1. *Let $f : I \to \mathbb{R}$ be convex, and m in the interior of I. Then there exists $a, b \in \mathbb{R}$ such that $f(x) \geq ax + b$, and $f(m) = am + b$.*

Proof: Let

$$a = \sup\left\{\frac{f(m) - f(x)}{m - x} \;\middle|\; x \in I, x < m\right\} < \infty.$$

Then for all $y > m, y \in I$ we have

$$\frac{f(m) - f(x)}{m - x} \leq a \leq \frac{f(y) - f(m)}{y - m},$$

hence setting $b = -am + f(m)$ gives $f(y) \geq ay - am + f(m)$ for all $y \geq m$, and $f(x) \geq ax - am + f(m)$ for all $x \leq m$.

Theorem 8.2 (Jensen's Inequality)**.** *Let X be an integrable random variable taking values in an interval $I \subseteq \mathbb{R}$, and let $f : I \to \mathbb{R}$ be convex. Then $\mathbb{E}[f(X)]$ is well-defined, and*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

Proof: If X is a constant almost everywhere, then there is nothing to prove, so assume otherwise.

Then $\mathbb{E}[X] = m$ is in the interior of I. Hence there exists $a, b \in \mathbb{R}$ such that $f(X) \geq aX + b$, with equality at m. In particular,

$$(f(X))^- \leq |a||X| + |b|,$$

hence $\mathbb{E}[f(X)^-] < \infty$, and $\mathbb{E}[f(X)]$ is well-defined, in $(-\infty, \infty]$.

We claim that $\mathbb{E}[f(X)] \geq a\mathbb{E}[X] + b$. If $\mathbb{E}[f(X)] = \infty$, then there is nothing to prove.

Otherwise, $f(X)$ and $aX + b$ are integrable random variables satisfying $f(X) \geq aX + b$, so taking the expectation,

$$\mathbb{E}[f(X)] \geq a\mathbb{E}[X] + b = am + b = f(m) = f(\mathbb{E}[X]).$$

Let's see an application.

Proposition 8.1. *If $X \in L^q$ for $q > p \geq 1$, then $X \in L^p$.*

Proof: The function $f(x) = |x|^{q/p}$ is convex, hence

$$\|X\|_p = \left(\mathbb{E}[|X|^p]\right)^{1/p} = \left(f(\mathbb{E}[|X|^p])\right)^{1/q} \leq \left(\mathbb{E}[f(|X|^p)]\right)^{1/q}$$
$$= \left(\mathbb{E}[|X|^q]\right)^{1/q} = \|X\|_q.$$

Hence $X \in L^q \implies X \in L^p$.

We can also show if $X \in L^\infty$, then $X \in L^p$ for $1 \leq p < \infty$. Hence,

$$L^\infty(\mathbb{P}) \subset L^q(\mathbb{P}) \subset L^p(\mathbb{P}) \subset L^1(\mathbb{P}),$$

for $1 < p < q < \infty$.

Theorem 8.3 (Hölder's Inequality)**.** *Let f, g be measurable and $1 \leq p < q \leq \infty$ be conjugate, that is, $\frac{1}{p} + \frac{1}{q} = 1$. Then,*

$$\mu(|fg|) \leq \|f\|_p \|g\|_q.$$

Proof: If p or q is 1 or ∞, then the statement is clear. So is when $\|f\|_p = 0$ or $\|g\|_q = 0$, so we exclude these cases.

Dividing both sides by $\|f\|_p$, we may assume that $\|f\|_p = 1$, i.e. $\int |f|^p \, \mathrm{d}\mu = 1$. So, define a probability measure $\mathbb{P}$ on $\mathcal{E}$ by

$$\mathbb{P}(A) = \int_A |f|^p \, \mathrm{d}\mu.$$

Also, for $h \geq 0$ measurable, we can define

$$\int h \, \mathrm{d}\mathbb{P} = \int h |f|^p \, \mathrm{d}\mu.$$

Then,

$$\mu(|fg|) = \mu(|fg| \mathbb{1}_{|f|>0}) = \int \frac{|f|^p |g|}{|f|^{p-1}} \mathbb{1}_{|f|>0} \, \mathrm{d}\mu$$

$$= \int \frac{|g|}{|f|^{p-1}} \mathbb{1}_{|f|>0} |f|^p \, \mathrm{d}\mu = \int \frac{|g|}{|f|^{p-1}} \mathbb{1}_{|f|>0} \, \mathrm{d}\mathbb{P}$$

$$= \mathbb{E}\left[\frac{|g|}{|f|^{p-1}} \mathbb{1}_{|f|>0} \right].$$

But then from Jensen's inequality, $\|X\|_p \leq \|X\|_q$, so this is

$$\mathbb{E}\left[\frac{|g|}{|f|^{p-1}} \mathbb{1}_{|f|>0} \right] \leq \left(\mathbb{E}\left[\frac{|g|^q}{|f|^{q(p-1)}} \mathbb{1}_{|f|>0} \right] \right)^{1/q} = \left(\int \frac{|g|^q}{|f|^p} \mathbb{1}_{|f|>0} \, \mathrm{d}\mathbb{P} \right)^{1/q}$$

$$= \left(\int |g|^q \mathbb{1}_{|f|>0} \, \mathrm{d}\mu \right)^{1/q} \leq \left(\int |g|^q \, \mathrm{d}\mu \right)^{1/q}$$

$$= \|g\|_q.$$

Remark. $p = q = 2$ gives the Cauchy-Schwarz inequality.

Theorem 8.4 (Minkowski's Inequality)**.** *For $p \in [1, \infty]$ and f, g measurable, we have*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

Proof: For $p = 1, \infty$ this is clear, as well as when $\|f\|_p = 0$ or $\|g\|_p = 0$. So assume else.

Then, since $|f + g|^p \leq 2^p(|f|^p + |g|^p)$, we have

$$\mu(|f + g|^p) \leq 2^p(\mu(|f|^p) + \mu(|g|^p)) < \infty,$$

so if $f, g \in L^p$, then $f + g \in L^p$. Letting q be the conjugate of p,

$$
\begin{aligned}
\int |f + g|^p \, \mathrm{d}\mu &= \int |f + g||f + g|^{p-1} \, \mathrm{d}\mu \\
&\leq \int |f + g|^{p-1}|f| \, \mathrm{d}\mu + \int |f + g|^{p-1}|g| \, \mathrm{d}\mu \\
&\leq \|f\|_p \|(f + g)^{p-1}\|_q + \|g\|_p \|(f + g)^{p-1}\|_q \\
&= \left(\int |f + g|^{(p-1)q} \, \mathrm{d}\mu \right)^{1/q} (\|f\|_p + \|g\|_p) \\
&= \|f + g\|_p^{p/q}(\|f\|_p + \|g\|_p),
\end{aligned}
$$

where we use Hölder's and the fact p, q are conjugate. Hence dividing we get
$\|f + g\|_p \leq \|f\|_p + \|g\|_p$.

Now we are ready to prove that L^p is complete.

Theorem 8.5. *Let $p \in [1, \infty]$, and let (f_n) be a sequence of functions in L^p such that, for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $m, n \geq N$,*

$$\|f_m - f_n\|_p < \varepsilon,$$

so (f_n) is Cauchy in L^p. Then there exists $f \in L^p$ such that $\|f_n - f\|_p \to 0$ as $n \to \infty$.

Proof: Assume $1 \leq p < \infty$, as $p = \infty$ is easy.

Then choose a subsequence (n_k) such that $\|f_{n_{k+1}} - f_{n_k}\|_p \leq 2^{-k}$. Then

$$S = \sum_{k=1}^{\infty} \|f_{n_{k+1}} - f_{n_k}\| < \infty.$$

By Minkowski's, for any $K \in \mathbb{N}$,

$$\left\| \sum_{k=1}^{K} |f_{n_{k+1}} - f_{n_k}| \right\| \leq \sum_{k=1}^{\infty} \|f_{n_{k+1}} - f_{n_k}\| = S < \infty.$$

Hence we get

$$\int \left(\sum_{k=1}^{K} |f_{n_{k+1}} - f_{n_k}| \right)^p \, \mathrm{d}\mu \leq S^p < \infty.$$

But notice that

$$\left(\sum_{k=1}^{K} |f_{n_{k+1}} - f_{n_k}| \right)^p \uparrow \left(\sum_{k=1}^{\infty} |f_{n_{k+1}} - f_{n_k}| \right)^p,$$

so by monotone convergence theorem,

$$\left\| \sum_{k=1}^{\infty} |f_{n_{k+1}} - f_{n_k}| \right\|_p < \infty.$$

In particular, we get

$$\sum_{k=1}^{\infty} |f_{n_{k+1}} - f_{n_k}| < \infty,$$

μ-almost everywhere. Let A be the set where this sum converges. Then $\mu(A^c) = 0$. For any $x \in A$, $(f_{n_k}(x))$ is Cauchy, and since $\mathbb{R}$ is complete, it converges to $f(x)$, say. Define $f(x) = 0$ for all $x \in A$.

Then f is measurable, with $f_{n_k} \to f$ μ-almost everywhere. Then,

$$\|f_n - f\|_p^p = \mu(|f_n - f|^p) = \mu(\liminf_k |f_n - f_{n_k}|^p)$$

$$\leq \liminf_k \mu(|f_n - f_{n_k}|^p) \leq \varepsilon^p,$$

by Fatou's lemma. Moreover,

$$\|f\|_p \leq \|f_N - f\|_p + \|f_N\|_p \leq \varepsilon^p + \|f_N\|_p < \infty,$$

so $f \in L^p$ and $f_n \to f$ in L^p.

Remark. One can check that any choice of vector spaces $V = C[0, 1]$, or simple functions, or finite limit of indicators of intervals, are dense in $L^p((0, 1), \mathcal{B}, \lambda)$, and so $\overline{(C[0, 1], \|\cdot\|_1)}$ is an L^1 space of Lebesgue integrable functions.

8.3   Hilbert Spaces

On a vector space V, a symmetric bilinear form $(u, v) \mapsto \langle u, v \rangle$ is called an *inner product* if $\langle v, v \rangle \geq 0$ for all $v \in V$, and $\langle v, v \rangle = 0$ if and only if $v = 0$.

Then $\sqrt{\langle v, v \rangle} = \|v\|$ is a norm.

If $(v, \|\cdot\|)$ is complete, it is called a *Hilbert space*.

Corollary 8.1. $\mathcal{L}^2$ *with the inner product* $\langle f, g \rangle = \int fg \, \mathrm{d}\mu$ *is a Hilbert space.*

There are a few basic geometric facts that follow:

1. Pythagoras: $\|f + g\|_2^2 = \|f\|_2^2 + \|g\|_2^2 + 2\langle f, g \rangle$.

2. Parallelogram law: $\|f + g\|_2^2 + \|f - g\|_2^2 = 2(\|f\|_2^2 + \|g\|_2^2)$.

We say f is *orthogonal* to g if $\langle f, g \rangle = 0$. Then $\|f + g\|_2^2 = \|f\|_2^2 + \|g\|_2^2$. For a subset $V \subseteq L^2$, we define the *orthogonal complement*

$$V^\perp = \{f \in L^2 \mid \langle f, v \rangle = 0 \text{ for all } v \in V\}.$$

A subset V is *closed* if $(f_n) \in V$ and $f_n \to f$ implies $f \in V$.

Theorem 8.6 (Orthogonal Projection)**.** *If V is a closed subspace of L^2, then for all $f \in L^2$, $f = v + w$ where $v \in V$ and $w \in V^\perp$.*

Moreover, $\|f - v\|_2 \leq \|f - g\|_2$ for all $g \in V$, with equality if and only if $g = v$ almost everywhere.

Proof: Define
$$d(f, V) = \inf_{g \in V} \|f - g\|_2.$$

Let $(g_n) \in V$ be a sequence such that $\|f - g_n\|_2 \to d(f, V)$. Now, by parallelogram law,

$$2(\|f - g_n\|^2 + \|f - g_m\|^2) = \|g_n - g_m\|^2 + 4\left\|f - \frac{g_n + g_m}{2}\right\|^2,$$

and as $(g_n + g_m)/2 \in V$, the latter term is at least $d(f, V)^2$. Hence taking limits as $n, m \to \infty$,
$$\|g_n - g_m\|^2 \to 0,$$

hence (g_n) is Cauchy.

As L^2 is complete and V is closed, $g_n \to v \in V$, and $\|f - g_n\|^2 \to \|f - v\|^2$, i.e. $d(f, V) = \|f - v\|$.

Then for any $h \in V$, $t \in \mathbb{R}$,

$$d(f, V)^2 \leq \|f - (v + th)^2\| = d(f, V)^2 - 2t\langle f - v, h \rangle + t^2 \|h\|^2.$$

Letting $t \downarrow 0$ and $t \uparrow 0$, we get $\langle f - v, h \rangle = 0$, hence $f - v \in V^\perp$.

> Now note $f = v + (f - v)$, so for any $g \in V$, $f - g = (f - v) + (v - g)$, hence
>
> $$\|f - g\|_2^2 = \|f - v\|_2^2 + \|v - g\|_2^2.$$
>
> Hence $\|f - g\|_2 \geq \|f - v\|_2$, with equality if and only if $\|v - g\|_2 = 0$, that is, $v = g$ almost everywhere.

8.4   Covariance

Take a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, with $\mathbb{E}[X] = \mathbb{E}[Y] = 0$.

Then the *covariance* is

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] = \langle X, Y \rangle.$$

We let the *variance* be $\text{Var}(X) = \text{Cov}(X, X)$. If X and Y are independent, then $\langle X, Y \rangle = 0$. However the converse is not true.

If $\mathcal{G}$ is a sub σ-algebra of $\mathcal{F}$, then $L^2(\Omega, \mathcal{G}, \mathbb{P})$ is a closed subspace of $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

For $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, a variant of the *conditional expectation* of X given $\mathcal{G}$, namely $\mathbb{E}[X|\mathcal{G}]$ is defined as the orthogonal projection of X on $L^2(\Omega, \mathcal{F}, \mathbb{P})$. This should be measurable with respect to $\mathcal{G}$, and satisfy $\|X - Y\|_2 \geq \|X - \mathbb{E}[X|\mathcal{G}]\|_2$ for all Y $\mathcal{G}$-measurable.

Now we want to define $\mathbb{E}[X|\mathcal{G}]$ if $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. Advanced probably.

Let $(G_i)_{i \in I}$ be a countable family of disjoint events whose union is Ω, and set $\mathcal{G} = \sigma(G_i \mid i \in I)$. Let X be integrable. Then the conditional expectation of X given $\mathcal{G}$ is given by

$$\mathbb{E}[X|G_i] = \frac{\mathbb{E}[X \mathbb{1}_{G_i}]}{\mathbb{P}(G_i)},$$

and then

$$Y = \sum_i \mathbb{E}[X|G_i] \mathbb{1}_{G_i}.$$

Then we can check that Y is $\mathcal{G}$-measurable, that $Y \in L^2(\Omega, \mathcal{G}, \mathbb{P})$ and Y is the orthogonal projection of X onto $L^2(\Omega, \mathcal{G}, \mathbb{P})$, if $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$.

8.5 L^p Convergence and Uniform Integrability

For a probability space, it is easy to see that $\mathbb{P}$-almost sure convergence does not imply L^p convergence.

Indeed, take $f_n = n\mathbb{1}_{(0,1/n)}$ on $((0,1), \mathcal{B}, \lambda_{(0,1)})$. Then $f_n \to 0$ almost surely, but

$$\mathbb{E}[|f_n|] = \mathbb{E}[f_n] = 1 \neq \mathbb{E}[0].$$

By dominated convergence theorem, we know that the following holds:

Theorem 8.7. *Let (X_n) be such that $X_n \xrightarrow{\mathbb{P}} X$, and $|X_n| \leq Y$ for some integrable random variable Y. Then $X_n \xrightarrow{L^1} X$, that is, $\mathbb{E}[|X_n - X|] \to 0$.*

We want to find other, minimal conditions, on (X_n) such that $X_n \xrightarrow{\mathbb{P}} X$ implies $X_n \xrightarrow{L^1} X$. This condition turns out to be uniform integrability: if $\mathbb{E}[|X_n|] \to \mathbb{E}[|X|]$. For $X \in L^1(\mathbb{P})$, define

$$I_X(\delta) = \sup\{\mathbb{E}[|X|\mathbb{1}_A] \mid A \in \mathcal{F}, \mathbb{P}(A) \leq \delta\}.$$

Then $I_X(\delta) \to 0$ as $\delta \to 0$.

If not, there is $\varepsilon > 0$ and $(A_n) \in \mathcal{F}$ such that $\mathbb{P}(A_n) \leq 2^{-n}$, and $\mathbb{E}[|X|\mathbb{1}_{A_n}] \geq \varepsilon$. Then $\sum \mathbb{P}(A_n) < \infty$, so by Borel-Cantelli, $\mathbb{P}(A_n \text{ i.o.}) = 0$. Then

$$\varepsilon \leq \mathbb{E}[|X|\mathbb{1}_{A_n}] \leq \mathbb{E}[|X|\mathbb{1}_{\bigcup_{m \geq n} A_m}] \to 0,$$

a contradiction.

Definition 8.2. Let $\mathcal{X}$ be a collection of random variables in $L^1(\mathbb{P})$. Define

$$I_{\mathcal{X}}(\delta) = \sup\{\mathbb{E}[|X|\mathbb{1}_A \mid X \in \mathcal{X}, A \in \mathcal{F}, \mathbb{P}(A) \leq \delta].$$

We say $\mathcal{X}$ is *uniformly integrable* if:

1. $\mathcal{X}$ is bounded in L^1, i.e. $I_{\mathcal{X}}(1) < \infty$.

2. $I_{\mathcal{X}}(\delta) \to 0$ as $\delta \to 0$.

Remark.

1. Any single integrable random variable is uniformly integrable, as is any finite collection of random variables.

   Moreover if

   $$\mathcal{X} = \{X \mid X \text{ a random variable such that } |X| \leq Y \text{ for some } Y \in L^1\},$$

   then $\mathcal{X}$ is uniformly integrable.

2. If $\mathcal{X}$ is bounded in L^p for some $p > 1$, then

$$\sup_{X \in \mathcal{X}} \mathbb{E}[X \mathbb{1}_A] \leq \|X\|_p (\mathbb{P}(A))^{1/q} \leq C\delta^{1/q},$$

for all A such that $\mathbb{P}(A) \leq \delta$. Hence $I_{\mathcal{X}}(\delta) \leq C\delta^{1/q} \to 0$ as $\delta \to 0$, so $\mathcal{X}$ is uniformly integrable.

Lemma 8.2 (Alternative Definition for Uniformly Integrable)**.**

$\mathcal{X}$ *is uniformly integrable* $\iff \sup_{X \in \mathcal{X}} \mathbb{E}[|X|\mathbb{1}_{|X|\geq K}] \to 0$ *as* $K \to \infty$.

Proof: $\implies$ Fix any $\varepsilon > 0$. Then there exists $\delta > 0$ such that $I_{\mathcal{X}}(\delta) < \varepsilon$. Choose $K < \infty$ such that

$$\frac{I_{\mathcal{X}}(1)}{\delta} = \frac{\sup_{X \in \mathcal{X}} \mathbb{E}[|X|]}{\delta} < K.$$

Then for any $X \in \mathcal{X}$,

$$\mathbb{P}(|X| \geq K) \leq \frac{\mathbb{E}[|X|]}{K} \leq \frac{I_{\mathcal{X}}(1)}{K} \leq \delta.$$

Hence with $A = \{|X| \geq K\}$,

$$\mathbb{E}[|X|\mathbb{1}_{|X|\geq K}] \leq I_X(\delta) < \varepsilon.$$

$\impliedby$ Note that

$$\mathbb{E}[|X|] = \mathbb{E}[|X|\mathbb{1}_{|X|\geq K}] + \mathbb{E}[|X|\mathbb{1}_{|X|\leq K}],$$

but the last part is at most K. Hence

$$\sup_{X \in \mathcal{X}} \mathbb{E}[|X|] \leq K + \sup_{X \in \mathcal{X}} \mathbb{E}[|X|\mathbb{1}_{|X|\geq K}].$$

Choose K large so that $\sup \mathbb{E}[|X|\mathbb{1}_{|X|\geq K}] < 1$. Then $\mathcal{X}$ is L^1 bounded.

Fix any $\varepsilon > 0$. Choose K such that $\mathbb{E}[|X|\mathbb{1}_{|X|\geq K}] < \varepsilon/2$ for all $X \in \mathcal{X}$. Then choose $\delta > 0$ such that $\delta \leq \varepsilon/2K$. Then for all $X \in \mathcal{X}$ and $A \in \mathcal{F}$ with $\mathbb{P}(A) \leq \delta$,

$$\mathbb{E}[|X|\mathbb{1}_A] = \mathbb{E}[|X|\mathbb{1}_A\mathbb{1}_{|X|\geq K}] + \mathbb{E}[|X|\mathbb{1}_A\mathbb{1}_{|X|\leq K}]$$
$$\leq \mathbb{E}[|X|\mathbb{1}_{|X|\geq K}] + K\mathbb{P}(A) \leq \varepsilon.$$

Theorem 8.8. *Let (X_n), X be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. The following are equivalent:*

(a) $X, X_n \in L^1$ *for all n and $X_n \overset{L^1}{\to} X$.*

(b) $\mathcal{X} = \{X_n \mid n \in \mathbb{N}\}$ *is uniformly integrable and $X_n \overset{\mathbb{P}}{\to} X$.*

Proof: (a) $\implies$ (b). By Markov,

$$\mathbb{P}(|X_n - X| > \varepsilon) \leq \frac{\mathbb{E}[|X_n - X|]}{\varepsilon} \to 0$$

as $n \to \infty$, so $X_n \overset{\mathbb{P}}{\to} X$.

Choose N such that $\mathbb{E}[|X_n - X|] < \varepsilon/2$, for all $n \geq N$, and δ such that $\mathbb{E}[|X|\mathbb{1}_A] \leq \varepsilon/2$ and $\mathbb{E}[|X_n|\mathbb{1}_A] \leq \varepsilon$ for all $n = 1, 2, \ldots, N-1$ whenever $\mathbb{P}(A) < \delta$. This is possible as any finite set of random variables is uniformly integrable. Then

$$\mathbb{E}[|X_n|\mathbb{1}_A] \leq \mathbb{E}[|X_n - X|] + \mathbb{E}[|X|\mathbb{1}_A] \leq \varepsilon$$

for all $n \geq N$, and it holds for all $n < N$ by assumption. Hence $\mathcal{X}$ is uniformly integrable.

(b) $\implies$ (a). As $X_n \overset{\mathbb{P}}{\to} X$, we can take a subsequence (n_k) such that $X_{n_k} \overset{a.s.}{\to} X$. Then,

$$\mathbb{E}[|X|] = \mathbb{E}[\liminf |X_{n_k}|] \leq \liminf \mathbb{E}[|X_{n_k}|] \leq \sup_{X_n \in \mathcal{X}} \mathbb{E}[|X_n|] < \infty.$$

So $X \in L^1$. Define the truncated random variables

$$X_n^K = (-K) \vee X_n \wedge K, \qquad X^K = (-K) \vee X \wedge K.$$

Then $X_n^K \overset{\mathbb{P}}{\to} X^K$, as $\mathbb{P}(|X_n^K - X^K| > \varepsilon) \leq \mathbb{P}(|X_n - X| > \varepsilon)$.

Since $|X_n^K| \leq K$ for all n, by bounded convergence theorem, $X_n^K \overset{L^1}{\to} X^K$. Now,

$$\mathbb{E}[|X_n - X|] \leq \mathbb{E}[|X_n - X_n^K|] + \mathbb{E}[|X - X^K|] + \mathbb{E}[|X_n^K - X^K|]$$
$$\leq \mathbb{E}[|X_n|\mathbb{1}_{|X_n|\geq K}] + \mathbb{E}[|X|\mathbb{1}_{|X|\geq K}] + \mathbb{E}[|X_n^K - X^K|] \leq \varepsilon$$

for all $n \geq N$, by choosing K large enough so that the first terms are at most $\varepsilon/3$, and N large enough so that the last term is at most $\varepsilon/3$.

9   Fourier Transforms

In this section, $L^p = L^p(\mathbb{R}^d)$ is the space of complex-valued functions on $\mathbb{R}^d$, i.e. $f : \mathbb{R}^d \to \mathbb{C}$ for which

$$\left(\int_{\mathbb{R}^d} |f(x)|^p \, \mathrm{d}x \right)^{1/p} < \infty.$$

For g with $|g|$ measurable, we define

$$\int g(x) \, \mathrm{d}\mu(x) = \int \Re(g(x)) \, \mathrm{d}\mu(x) + i \int \Im(g(x)) \, \mathrm{d}\mu(x).$$

Then notice that

$$\left| \int g(x) \, \mathrm{d}\mu(x) \right| \leq \int |g(x)| \, \mathrm{d}\mu(x).$$

Define an inner product on $L^2(\mu)$ by

$$\langle f, g \rangle = \int f(x) \overline{g(x)} \, \mathrm{d}\mu(x).$$

For any $y \in \mathbb{R}^d$,

$$\int f(x - y) \, \mathrm{d}x = \int f(y - x) \, \mathrm{d}x = \int f(x) \, \mathrm{d}x = \int f(-x) \, \mathrm{d}x,$$

and for any $a \in \mathbb{R}$, $a \neq 0$,

$$\int f(ax) \, \mathrm{d}x = \frac{1}{a^d} \int f(x) \, \mathrm{d}x.$$

The *Fourier transform* $\hat{f}$ of $f \in L^1(\mathbb{R}^d)$ is defined as

$$\hat{f}(u) = \int_{\mathbb{R}^d} f(x) e^{i\langle u, x \rangle} \, \mathrm{d}x,$$

for $u \in \mathbb{R}^d$, where

$$\langle u, x \rangle = \sum_{i=1}^{d} u_i x_i.$$

Then notice that

$$\sup_u |\hat{f}(u)| \leq \int |f(x)| \, \mathrm{d}x = \|f\|_1 < \infty,$$

so $\hat{f} \in L^\infty$. And for $u_n \to u$, notice that

$$f(x)e^{i\langle u_n,x\rangle} \to f(x)e^{i\langle u,x\rangle}, \qquad |f(x)e^{i\langle u_n,x\rangle}| \le |f(x)|.$$

Since $f \in L^1$, by dominated convergence theorem, $\hat{f}(u_n) \to \hat{f}(u)$. Moreover,

$$\lim_{\|u\|\to\infty} \hat{f}(u) = 0.$$

This is the *Riemann-Lebesgue lemma*

Hence $\hat{f} \in c_0(\mathbb{R}^d)$, the space of bounded continuous functions vanishing at $\pm\infty$. The map is one-to-one, but not onto.

For a finite or probability measure μ on $\mathbb{R}^d$, define similarly

$$\hat{\mu}(u) = \int e^{i\langle u,x\rangle}\,\mathrm{d}\mu(x),$$

for $u \in \mathbb{R}^d$. Then $\hat{\mu}$ is a bounded continuous function on $\mathbb{R}^d$, and $|\hat{\mu}(u)| \le \mu(\mathbb{R}^d) < \infty$. If μ has density f with respect to λ, then

$$\hat{\mu}(u) = \int e^{i\langle u,x\rangle} f(x)\,\mathrm{d}x = \hat{f}(u).$$

The *characteristic function* ϕ_X of a random variable X on $\mathbb{R}^d$ is the Fourier transform of the law $\mu_X = \mathbb{P} \circ X^{-1}$. So,

$$\phi_X(u) = \hat{\mu}_X(u) = \int e^{i\langle u,x\rangle}\,\mathrm{d}\mu_X(x) = \int e^{i\langle u,X\rangle}\,\mathrm{d}\mathbb{P} = \mathbb{E}[e^{i\langle u,X\rangle}].$$

In particular, if X has probability distribution function f, then

$$\phi_X(u) = \hat{f}(u).$$

Definition 9.1. For $f \in L^1(\mathbb{R}^d)$ with $\hat{f} \in L^1(\mathbb{R}^d)$, we say that the *Fourier inversion* holds for f if

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{f}(u)e^{-i\langle u,x\rangle}\,\mathrm{d}u,$$

almost everywhere in $\mathbb{R}^d$.

Remark.

1. The right hand side is continuous by dominated convergence theorem. Hence for f continuous, the equality is everywhere in $\mathbb{R}^d$.

2. $f \mapsto \hat{f}$ is one-to-one: if $f, g \in L^1$ with $\hat{f} = \hat{g}$, then $f - g \in L^1$ and $\widehat{f - g} = \hat{f} - \hat{g} = 0$. So by Fourier inversion, $f - g = 0$ almost everywhere.

A key concept in Fourier analysis is *convolution*.

For $f \in L^p(\mathbb{R}^d)$ with $1 \le p < \infty$ and ν a probability measure, then

$$f * \nu(x) = \int_{\mathbb{R}^d} f(x - y)\nu(\mathrm{d}y),$$

for $x \in \mathbb{R}^d$ where the integral exists, and 0 otherwise. Note that

$$\int |f * \nu(x)|^p \, \mathrm{d}x \le \int \left(\int |f(x - y)|\nu(\mathrm{d}y) \right)^p \mathrm{d}x \le \int \int |f(x - y)|^p \nu(\mathrm{d}y)$$

$$= \int \left(\int |f(x - y)|^p \, \mathrm{d}x \right) \nu(\mathrm{d}y) = \int \left(\int |f(x)|^p \, \mathrm{d}x \right) \nu(\mathrm{d}y)$$

$$= \|f\|_p^p < \infty.$$

Hence $f * \nu$ is defined almost everywhere, and $\|f * \nu\|_p \le \|f\|_p < \infty$. When ν has probability density function $g \in L^1$, then

$$f * \nu(x) = \int f(x - y)g(y) \, \mathrm{d}y = f * g(x).$$

For two probability measure μ, ν on $\mathbb{R}^d$, the convolution $\mu * \nu$ is a new probability measure defined as

$$\mu * \nu(A) = \int \int \mathbb{1}_A(x + y)\mu(\mathrm{d}x)\nu(\mathrm{d}y) = \mu \otimes \nu(x + y \in A) = \mathbb{P}(X + Y \in A),$$

where X, Y are independent and $X \sim \mu$, $Y \sim \nu$. If μ has probability distribution function $f \in L^1$, then

$$\mu * \nu(A) = \int \left(\int \mathbb{1}_A(x + y)f(x) \, \mathrm{d}x \right) \nu(\mathrm{d}y) = \int \left(\int \mathbb{1}_A(x)f(x - y) \, \mathrm{d}x \right) \nu(\mathrm{d}y)$$

$$= \int \mathbb{1}_A(x) \left(\int f(x - y)\nu(\mathrm{d}y) \right) \mathrm{d}x = \int \mathbb{1}_A(x)f * \nu(x) \, \mathrm{d}x.$$

So $\mu * \nu$ has the probability distribution function $f * \nu$.

We can check that the following are true:

- $\widehat{f * \nu}(u) = \hat{f}(u)\hat{\nu}(u)$, for all $f \in L^1$ and ν a probability measure.

- $\widehat{\mu * \nu}(u) = \hat{\mu}(u)\hat{\nu}(u)$ for all μ, ν probability measures.

The proof of this is by defining independent variables X, Y with $X \sim \mu$ and $Y \sim \nu$, and then $X + Y \sim \mu * \nu$. Then looking at the definition of the Fourier transform in terms of expectations.

9.1   Fourier Transform of Gaussians

If ϕ_z is the characteristic functions of $Z \sim N(0,1)$, i.e.

$$\phi_Z(u) = \int \frac{1}{\sqrt{2\pi}} e^{-z^2/2} e^{iuz} \, \mathrm{d}z,$$

then by a proven theorem, ϕ_Z is differentiable and can be differentiated under the integral sign, i.e.

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}u} \phi_Z(u) &= \frac{1}{\sqrt{2\pi}} \int \frac{\mathrm{d}}{\mathrm{d}u} \left(e^{-z^2/2} e^{iuz} \right) \mathrm{d}z = \frac{1}{\sqrt{2\pi}} \int ize^{iuz} e^{-z^2/2} \, \mathrm{d}z \\
&= \frac{i}{\sqrt{2\pi}} \int e^{iuz} (ze^{-z^2/2}) \, \mathrm{d}z = \frac{i}{\sqrt{2\pi}} \int iue^{iuz} e^{-z^2/2} \, \mathrm{d}z \\
&= -u \int \frac{1}{\sqrt{2\pi}} e^{iuz} e^{-z^2/2} \, \mathrm{d}z = -u\phi_Z(u).
\end{aligned}$$

Hence solving, as $\phi_Z(0) = 1$, we get

$$\phi_Z(u) = e^{-u^2/2}.$$

Consider for $t \in (0, \infty)$ the centred Gaussian random variables on $\mathbb{R}^d$, which has pdf

$$g_t(x) = \frac{1}{(2\pi t)^{d/2}} e^{-\|x\|^2/2t}.$$

If $(Z_1, \ldots, Z_d) = Z$ are iid $N(0,1)$, then $\sqrt{t}Z$ has density g_t. So,

$$\begin{aligned}
\hat{g}_t(u) &= \mathbb{E}[e^{i\langle u, \sqrt{t}Z \rangle}] = \mathbb{E}[e^{i \sum_1^d u_i \sqrt{t} Z_i}] \\
&= \mathbb{E}\left[\prod_{i=1}^d e^{iu_i \sqrt{t} Z_i} \right] = \prod_{i=1}^d \mathbb{E}[e^{iu_i \sqrt{t} Z_i}] \\
&= \prod_{i=1}^d \phi_{Z_i}(\sqrt{t} u_i) = \prod_{i=1}^d e^{-tu_i^2/2} = e^{-t\|u\|^2/2}.
\end{aligned}$$

Hence we get

$$\hat{g}_t(u) = e^{-t\|u\|^2/2} = \frac{(2\pi)^{-d/2}}{t^{d/2}}\left(\frac{t}{2\pi}\right)^{d/2}e^{-t\|u\|^2/2} = \frac{(2\pi)^{d/2}}{t^{d/2}}g_{1/t}(u).$$

Hence

$$\hat{\hat{g}}_t(u) = \frac{(2\pi)^{d/2}}{t^{d/2}}\hat{g}_{1/t}(u) = (2\pi)^d g_t(u).$$

Therefore,

$$g_t(x) = g_t(-x) = (2\pi)^{-d}\hat{\hat{g}}_t(-x) = (2\pi)^{-d}\int \hat{g}_t(u)e^{-i\langle u,x\rangle}\,\mathrm{d}u.$$

Therefore Fourier inversion holds for g_t.

9.2   Gaussian Convolution

For $f \in L^1(\mathbb{R}^d)$, we look at $f * g_t$, where

$$g_t(u) = \frac{1}{(2\pi t)^{d/2}}e^{-|u|^2/2t}.$$

Then this has the following properties:

1. $\|f * g_t\|_1 \leq \|f\|_1$.

2. $f * g_t$ is continuous.

3. $f * g_t$ is bounded.

4. $\widehat{f * g_t}(u) = \hat{f}(u)\hat{g}_t(u) = \hat{f}(u)e^{-t|u|^2/2}$.

5. $\widehat{f * g_t}$ is bounded and continuous.

6. $\|\widehat{f * g_t}\|_1 \leq c_t\|\hat{f}\|_\infty \leq c_t\|f\|_1$.

7. For μ a probability measure and any $t > 0$, $\mu * g_t$ is a Gaussian convolution.

Lemma 9.1. *Fourier inversion holds for Gaussian convolutions.*

Proof: Let $f \in L^1$ and $t > 0$. Then, using Fubini,

$$(2\pi)^d f * g_t(x) = (2\pi)^d \int f(x-y)g_t(y)\,\mathrm{d}y$$

$$= \int f(x-y)(2\pi)^d g_t(y)\,\mathrm{d}y$$

$$= \int f(x-y) \int \hat{g}_t(u)e^{-i\langle u,y\rangle}\,\mathrm{d}u\,\mathrm{d}y$$

$$= \int \int f(x-y)\hat{g}_t(u)e^{-i\langle u,y\rangle}\,\mathrm{d}u\,\mathrm{d}y$$

$$= \int \hat{g}_t(u)\left(\int f(x-y)e^{-i\langle u,y\rangle}\,\mathrm{d}y\right)\mathrm{d}u$$

$$= \int \hat{g}_t(u)\left(\int f(y)e^{-i\langle u,x-y\rangle}\,\mathrm{d}y\right)\mathrm{d}y$$

$$= \int \hat{g}_t(u)e^{-i\langle u,x\rangle}\left(\int f(y)e^{i\langle u,y\rangle}\,\mathrm{d}y\right)\mathrm{d}u$$

$$= \int \hat{g}_t(u)e^{-i\langle u,x\rangle}\hat{f}(u)\,\mathrm{d}u$$

$$= \int \hat{g}_t(u)\hat{f}(u)e^{-i\langle u,x\rangle}\,\mathrm{d}u$$

$$= \int \widehat{f * g_t}(u)e^{-i\langle u,x\rangle}\,\mathrm{d}u.$$

Lemma 9.2. *Let $f \in L^p(\mathbb{R}^d)$, and $1 \le p < \infty$. Then,*

$$\|f * g_t - f\|_p \overset{t\to 0}{\to} 0.$$

Proof: Given $\varepsilon > 0$, there exists $h \in c_c(\mathbb{R}^d)$ (continuous functions with compact support), such that $\|f-h\|_p \le \varepsilon/3$. Then by linearity of convolution,

$$\|f * g_t - h * g_t\|_p = \|(f-h) * g_t\|_{\le} \|f-h\|_p \le \varepsilon/3.$$

So,

$$\|f*g_t - f\|_p \le \|f*g_t - h*g_t\|_p + \|f-h\|_p + \|h*g_t - h\|_p \le \frac{2\varepsilon}{3} + \|h*g_t - h\|_p.$$

Hence it is enough to prove $\|h * g_t - h\|_p \to 0$.

So h is bounded, and h is supported on $[-M, M]$ for some $M > 0$. Define

$$e(y) = \int |h(x-y) - h(x)|^p \, \mathrm{d}x.$$

Then as $y \to 0$, $|h(x-y) - h(x)|^p \to 0$ as h is continuous. But note that

$$|h(x-y) - h(x)|^p \le 2^p \|h\|_\infty^p \mathbb{1}_{|x| \le M+1},$$

which is integrable. Hence by dominated convergence theorem, $e(y) \to 0$ as $y \to 0$. Also, by Jensen and Fubini,

$$\begin{aligned}
\|h * g_t - h\|_p^p &= \int \left| \int h(x-y) g_t(y) \, \mathrm{d}y - \int h(x) g_t(y) \, \mathrm{d}y \right|^p \mathrm{d}x \\
&= \int \left| \int (h(x-y) - h(x)) g_t(y) \, \mathrm{d}y \right|^p \mathrm{d}x \\
&\le \int \int |h(x-y) - h(x)|^p g_t(y) \, \mathrm{d}y \, \mathrm{d}x \\
&= \int e(y) g_t(y) \, \mathrm{d}y = \int e(y) \frac{1}{t^{d/2}} g_1\left(\frac{y}{\sqrt{t}}\right) \mathrm{d}y \\
&= \int e(\sqrt{t}y) g_1(y) \, \mathrm{d}y \to 0,
\end{aligned}$$

by dominated convergence theorem.

Theorem 9.1. *Let $f \in L^1(\mathbb{R}^d)$ and $\hat{f} \in L^1(\mathbb{R}^d)$. Then,*

$$f(x) = \frac{1}{(2\pi)^d} \int e^{-i\langle u, x\rangle} \hat{f}(u) \, \mathrm{d}u \quad \text{almost everywhere.}$$

Proof: Consider $f * g_t$ and

$$f_t(x) = \frac{1}{(2\pi)^d} \int e^{-i\langle x, u\rangle} \hat{f}(u) e^{-t|u|^2/2} \, \mathrm{d}u = \frac{1}{(2\pi)^d} \int e^{-i\langle x, u\rangle} \widehat{f * g_t}(u) \, \mathrm{d}u.$$

As Fourier inversion holds for $f * g_t$, $f * g_t = f_t$. Now notice $\|f_t - f\|_1 \overset{t\to 0}{\to} 0$, so there exists a sequence such that $f_{t_n} \to f_t$ almost everywhere.

But since $e^{-i\langle x, u\rangle} \hat{f}(u)$ is bounded by $|\hat{f}(u)|$ which is integrable, by dominated

convergence theorem,

$$f_t(x) \to \frac{1}{(2\pi)^d} \int e^{-i\langle x,u\rangle} \hat{f}(u) \, \mathrm{d}u,$$

almost everywhere. Hence

$$f(x) = \frac{1}{(2\pi)^d} \int e^{-i\langle x,u\rangle} \hat{f}(u) \, \mathrm{d}u \quad \text{almost everywhere.}$$

Theorem 9.2 (Plancherel). *For $f, g \in L^1 \cap L^2(\mathbb{R}^d)$,*

$$\|f\|_2 = \frac{1}{(2\pi)^{d/2}} \|\hat{f}\|_2 \qquad \text{and} \qquad \langle f, g\rangle_2 = \frac{1}{(2\pi)^{d/2}} \langle \hat{f}, \hat{g}\rangle.$$

Remark. Since $L^1 \cap L^2$ is dense in L^2, the linear operator $F_0 : L^1 \cap L^2 \to L^2$ given by $F_0(f) = (2\pi)^{-d/2} \hat{f}$ extends to an isometry $F : L^2 \to L^2$, which is onto by Fourier inversion.

Proof: First, assume $f \in L^1$ with $\hat{f} \in L^1$. Then $(x, u) \mapsto f(x)\hat{f}(u)$ is $\mathrm{d}x \, \mathrm{d}u$ integrable, so

$$(2\pi)^d \|f\|_2^2 = (2\pi)^d \int f(x)\overline{f(x)} \, \mathrm{d}x = \int \int \hat{f}(u) e^{-i\langle x,u\rangle} \overline{f(x)} \, \mathrm{d}u \, \mathrm{d}x < \infty$$

$$= \int \hat{f}(u) \overline{\left(\int f(x) e^{i\langle x,u\rangle} \, \mathrm{d}x \right)} \, \mathrm{d}u$$

$$= \int \hat{f}(u) \overline{\hat{f}(u)} \, \mathrm{d}u = \|\hat{f}\|_2^2.$$

Now let $f \in L^1 \cap L^2$. For $t > 0$, take $f_t = f * g_t \to f$ in L^2, so $\|f_t\|_2 \to \|f\|_2$. Also,

$$|\hat{f}_t(u)| = |\hat{f}(u)\hat{g}_t(u)| = |\hat{f}(u)|e^{-t|u|^2/2} \uparrow |\hat{f}(u)|,$$

as $t \to 0$, and hence

$$\|\hat{f}_t\|_2^2 = \int |\hat{f}_t(u)|^2 \, \mathrm{d}u \to \int |\hat{f}(u)|^2 \, \mathrm{d}u = \|\hat{f}\|_2^2,$$

by monotone convergence theorem. But $f_t = f * g_t \in L^1$, and $\hat{f}_t \in L^1$, so by our result,

$$(2\pi)^d \|f_t\|_2^2 = \|\hat{f}_t\|_2^2.$$

Let $t \to 0$. Then we get
$$(2\pi)^d \|f\|_2^2 = \|\hat{f}\|_2^2.$$
The proof for $\langle f, g \rangle$ is similar.

9.3   Characteristic Functions, Weak Convergence and CLT

For a random variable X, recall the characteristic function is

$$\phi_X(t) = \mathbb{E}[e^{itX}] = \hat{\mu}_X = \int e^{i\langle t,x \rangle}\, \mathrm{d}\mu_X(x).$$

For the Dirac measure δ_0, note that

$$\hat{\delta}_0 = \int e^{itx}\, \mathrm{d}\delta_0(x) = 1$$

is not integrable on $\mathbb{R}$, so Fourier inversion does not make sense. To circumvent this, we test μ on test functions f.

Remark. Two probability measures μ and ν on $\mathbb{R}^d$ coincide if and only if $\int f\, \mathrm{d}\mu = \int f\, \mathrm{d}\nu$ for all $f : \mathbb{R}^d \to \mathbb{R}$ bounded and continuous.

In fact, it is enough to have that this holds for all $f \in c_c^\infty$ (infinitely differentiable functions with compact support).

Definition 9.2. Let $(\mu_n), \mu$ be Borel probability measures on $\mathbb{R}^d$. Then μ_n *converges to μ weakly* if

$$\int f\, \mathrm{d}\mu_n \to \int f\, \mathrm{d}\mu,$$

for all $f : \mathbb{R}^d \to \mathbb{R}$ bounded and continuous.

Remark.

1. For a sequence of random variables (X_n) and random variable X, $X_n \to X$ weakly if $\mu_{X_n} \to \mu_X$ weakly.

2. A sequence (μ_n) can have at most one weak limit (by our previous remark).

3. If $X_n \to X$ weakly, and $h : \mathbb{R}^d \to \mathbb{R}^k$ is continuous, then $h(X_n) \to h(X)$ weakly.

4. To prove weak convergence it is sufficient to check the condition for all $f \in c_c^\infty$ (by a form of tightness argument: show there exists K compact such that $\mu_n(K^c) < \varepsilon$ for all n if $\mu_n \to \mu$ weakly).

5. When $d = 1$, this is equivalent to $X_n \xrightarrow{\text{d}} X$.

Theorem 9.3. *Let X be a random variable on $\mathbb{R}^d$. Then μ_X is uniquely determined by $\hat{\mu}_X = \phi_X$. Further, if $\phi_X \in L^1$, then μ_X has a bounded continuous pdf given by*

$$f_X(x) = \frac{1}{(2\pi)^d} \int \phi_X(u) e^{-i\langle x, u\rangle} \, \mathrm{d}u.$$

Proof: Take $Z \sim N(0, I_d)$, independent of X. Thus, $\sqrt{t}Z$ has pdf g_t, and $X + \sqrt{t}Z$ has pdf $\mu_X * g_t = f_t$. Then,

$$\hat{f}_t(u) = \hat{\mu}_X(u)\hat{g}_t(u) = \phi_X(u) e^{-t|u|^2/2}.$$

So by Fourier inversion of Gaussian convolutions,

$$f_t(x) = \frac{1}{(2\pi)^d} \int \phi_X(u) e^{-t|u|^2/2} e^{-i\langle u, x\rangle} \, \mathrm{d}u,$$

for all x. So f_t is uniquely determined by ϕ_X. Now for any g bounded and continuous, as $t \to 0$,

$$\int g(x) f_t(x) \, \mathrm{d}x = \mathbb{E}[g(X + \sqrt{t}Z)] \to \mathbb{E}[g(X)] = \int g(x) \mu_X(\mathrm{d}x).$$

Hence $\int g(x) \, \mathrm{d}\mu_X$ is uniquely determined by ϕ_X for all such g, so μ_X is uniquely determined by ϕ_X.

If $\phi_X \in L^1$, then

$$\phi_X(u) e^{-t|u|^2/2} e^{-i\langle u, x\rangle} \to \phi_X(u) e^{-i\langle u, x\rangle}.$$

By DCT, $f_t(x) \to f_X(x)$ for all x. In particular, $f_X(x) \geq 0$ for all x. Also, for any g bounded and continuous with compact support,

$$\int g(x) f_t(x) \, \mathrm{d}x \to \int g(x) f_X(x) \, \mathrm{d}x,$$

by DCT. But we already know this converges to $\int g(x) \, \mathrm{d}\mu_X$, so

$$\int g(x) \mu_X(\mathrm{d}x) = \int g(x) f_X(x) \, \mathrm{d}x,$$

for all g bounded continuous with compact support, hence μ_X has density f_X.

Theorem 9.4 (Levy). *Let $(X_n), X$ be random variables on $\mathbb{R}^d$ with $\phi_{X_n}(u) \to \phi_X(u)$ for all u. Then $X_n \to X$ weakly.*

Remark.

1. Levy's continuity theorem states that if $\phi_{X_n}(u) \to \phi(u)$ for some function ϕ that is continuous in a neighbourhood of 0, then ϕ is the characteristic function of some random variable X, and $X_n \to X$ weakly.

2. Cramér-Wold device: Let $(X_n), X$ be random variables on $\mathbb{R}^d$. Then $X_n \to X$ weakly if and only if
$$\langle u, X_n \rangle \xrightarrow{\mathrm{d}} \langle u, X \rangle$$
   for all $u \in \mathbb{R}^d$.

Proof: Let $g : \mathbb{R}^d \to \mathbb{R}$ be compactly supported and Lipschitz continuous, i.e. $|g(x) - g(y)| \leq C_g |x - y|$ for some C_g. It is enough to show that
$$\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)].$$

"It is always better to add a small Gaussian", as we have seen. Let $Z \sim N(0, I_d)$ independent of $(X_n), X$. Then for fixed $\varepsilon > 0$, choose t small enough so that
$$C_g \sqrt{t} \mathbb{E}[|Z|] \leq \frac{\varepsilon}{3}.$$

Then, we have
$$
\begin{aligned}
|\mathbb{E}[g(X_n)] - \mathbb{E}[g(X)]| \leq{}& |\mathbb{E}[g(X_n)] - \mathbb{E}[g(X_n + \sqrt{t}Z)] \\
&+ |\mathbb{E}[g(X)] - \mathbb{E}[g(X + \sqrt{t}Z)]| \\
&+ |\mathbb{E}[g(X_n + \sqrt{t}Z)] - \mathbb{E}[g(X + \sqrt{t}Z)]| \\
\leq{}& 2C_g \sqrt{t} \mathbb{E}[|Z|] + |\mathbb{E}[g(X_n + \sqrt{t}Z)] - \mathbb{E}[g(X + \sqrt{t}Z)]|.
\end{aligned}
$$

Notice that $X_n + \sqrt{t}Z$ has density $\mu_{X_n} * g_t = f_{t,n}$, and by Fourier inversion,
$$f_{t,n}(x) = \frac{1}{(2\pi)^d} \int \phi_{X_n}(u) e^{-t|u|^2/2} e^{-i\langle u, x\rangle} \, \mathrm{d}u,$$

so we get that
$$\mathbb{E}[g(X_n + \sqrt{t}Z)] = \frac{1}{(2\pi)^d} \iint g(x) \phi_{X_n}(u) e^{-t|u|^2/2} e^{-i\langle u, x\rangle} \, \mathrm{d}u \, \mathrm{d}x.$$

As $n \to \infty$, the integral converges to
$$\frac{1}{(2\pi)^d} \iint g(x) \phi_X(u) e^{-t|u|^2/2} e^{-i\langle u, x\rangle} \, \mathrm{d}u \, \mathrm{d}x,$$

> by DCT as the integrand converges and is bounded by $g(x)e^{-t|u|^2/2}$, which is $\mathrm{d}u\,\mathrm{d}x$ integrable. But this is just $\mathbb{E}[g(X + \sqrt{t}Z)]$, proving our result.

Theorem 9.5 (Central Limit Theorem). *Let (X_n) be iid random variables on $\mathbb{R}$ with $\mathbb{E}[X_i] = 0$, and $\mathrm{Var}(X_i) = 1$. Then for $S_n = X_1 + X_2 + \cdots + X_n$, we have*

$$\frac{S_n}{\sqrt{n}} \to Z \sim N(0, 1),$$

weakly, or in distribution.

> **Proof:** Set $\phi(u) = \phi_{X_1}(u) = \mathbb{E}[e^{iuX_1}]$. Then $\phi(0) = 1$, and since $\mathbb{E}[X_1^2] < \infty$ we can differentiate under the integral sign to get
>
> $$\phi'(u) = i\mathbb{E}[X_1 e^{iuX_1}], \qquad \phi''(u) = i^2\mathbb{E}[X_i^2 e^{iuX_1}],$$
>
> and then $\phi'(0) = 0$, $\phi''(0) = -1$. By Taylor's theorem, as $u \to 0$,
>
> $$\phi(u) = 1 - \frac{u^2}{2} + o(u^2).$$
>
> Let ϕ_n be characteristic function of $S_n/\sqrt{n}$. Then,
>
> $$\phi_n(u) = \mathbb{E}[iuS_n/\sqrt{n}] = \mathbb{E}[e^{iu(X_1 + \cdots + X_n)/\sqrt{n}}] = \left(\mathbb{E}[e^{iuX_1/\sqrt{n}}]\right)^n$$
> $$= \left(\phi\left(\frac{u}{\sqrt{n}}\right)\right)^n = \left(1 - \frac{u^2}{2n} + o\left(\frac{u^2}{n}\right)\right)^n$$
> $$= \left(1 - \frac{u^2}{2n} + o\left(\frac{1}{n}\right)\right)^n,$$
>
> as u fixed and as $n \to \infty$. The complex logarithm satisfies, as $z \to 0$, $\log(1 + z) = z + o(z)$. So,
>
> $$\log \phi_n(u) = n\log\left(1 - \frac{u^2}{2n} + o\left(\frac{1}{n}\right)\right) = n\left(-\frac{u^2}{2n} + o\left(\frac{1}{n}\right)\right) = -\frac{u^2}{2} + o(1).$$
>
> So, we have
> $$\phi_n(u) \to e^{-u^2/2} = \phi_Z(u).$$
> By Levy, this means that $S_n/\sqrt{n} \to Z$ weakly.

Remark. The central limit theorem in $\mathbb{R}^d$ can be proven similarly using the Cramér-Wold device, and by properties of multivariate Gaussians.

Recall that a random variable on $\mathbb{R}$ is Gaussian $N(\mu, \sigma)$ if it has density

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

for $\mu \in \mathbb{R}$ and $\sigma > 0$.

Definition 9.3. X in $\mathbb{R}^d$ is *Gaussian* if

$$\langle u, X \rangle = \sum_{i=1}^{d} u_i X_i$$

is Gaussian for all $u \in \mathbb{R}^d$.

For example, if $X_1, \ldots, X_n \sim N(0, 1)$ iid, then $X = (X_1, \ldots, X_n)$ is Gaussian in $\mathbb{R}^n$.

Proposition 9.1. *Let X be Gaussian in $\mathbb{R}^n$ and A be an $m \times n$ matrix, and $b \in \mathbb{R}^m$. Then:*

(a) $AX + b$ is Gaussian in $\mathbb{R}^m$.

(b) $X \in L^2$ and μ_X is determined by $\mu = \mathbb{E}[X]$ and $V = (\text{Cov}(X_i, X_j))_{i,j}$.

(c) $\phi_X(u) = e^{i\langle u, \mu \rangle - \langle u, Vu \rangle / 2}$, for all $u \in \mathbb{R}^n$.

(d) If V is invertible, then X has a pdf on $\mathbb{R}^n$ given by

$$f_X(x) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|V|^{1/2}} e^{-\langle x - \mu, V^{-1}(X - \mu) \rangle / 2}.$$

(e) If $X = (X_1, X_2)$ with $X_1, \in \mathbb{R}^{n_1}$, $X_2 \in \mathbb{R}^{n_2}$, then X_1, X_2 are independent if and only if $\text{Cov}(X_1, X_2) = 0$.

9.4   Weak Law of Large Numbers

For (X_i) iid with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) < \infty$, then for all $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \varepsilon\right) \leq \frac{1}{n^2 \varepsilon^2} \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{n \text{Var}(X_i)}{n^2 \varepsilon^2} \to 0,$$

by Chebyshev's inequality. So, we have $S_n/n \to \mu$ in probability.

In the next section we will work on proving the strong law of large numbers, which does not assume finite variance and strengthens the convergence to almost-sure convergence.

Proposition 9.2. *Let (X_n) be independent with $\mathbb{E}[X_n] = \mu$ and $\mathbb{E}[X_n^4] \leq M$ for all n Then $S_n/n \to \mu$ almost surely.*

Proof: Let $X'_n = X_n - \mu$. Then,

$$\mathbb{E}[X'^4_n] \leq 2^4(\mathbb{E}[X^4_n] + \mu^4) \leq 2^4(M + \mu^4) = M',$$

for all n. So we can assume that $\mu = 0$. Then using independence, for distinct indices i, j, k, l,

$$0 = \mathbb{E}[X_i X^3_j] = \mathbb{E}[X_i X_j X^2_k] = \mathbb{E}[X_i X_j X_k X_l].$$

Hence, we have

$$\mathbb{E}[S^4_n] = \mathbb{E}[(X_1 + \cdots + X_n)^4] = \mathbb{E}\left[\sum_{1 \leq i \leq n} X^4_i + 6 \sum_{1 \leq i < j \leq n} X^2_i X^2_j\right].$$

But note that $\mathbb{E}[X^2_i X^2_j] \leq \sqrt{\mathbb{E}[X^4_i]\mathbb{E}[X^4_j]} \leq M$, by Cauchy-Schwarz. Hence, we have

$$\mathbb{E}\left[\frac{S^4_n}{n^4}\right] \leq \frac{3M}{n^2}.$$

Now we have

$$\mathbb{E}\left[\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4\right] = \sum_{n=1}^{\infty}\mathbb{E}\left[\left(\frac{S_n}{n}\right)^4\right] \leq \sum_n \frac{3M}{n^2} < \infty.$$

So we have that

$$\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4 < \infty,$$

almost surely. Hence we have $S^4_n/n^4 \to 0$ almost surely, or $S_n/n \to 0$ almost surely.

10 Ergodic Theory

Definition 10.1. Let $(E, \mathcal{E}, \mu)$ be a σ-finite measure space. A measurable map $\theta : E \to E$ is called *measure-preserving* if

$$\mu \circ \theta^{-1} = \mu \text{ i.e. } \mu(\theta^{-1}(A)) = \mu(A) \text{ for all } A \in \mathcal{E}.$$

In this case, for all $f \in L^1$,

$$\int f \, \mathrm{d}\mu = \int f \circ \theta \, \mathrm{d}\mu.$$

A set $A \in \mathcal{E}$ is *θ-invariant* if $\theta^{-1}(A) = A$.

A measurable function f is *θ-invariant* if $f = f \circ \theta$.

The set of all θ-invariant sets $\mathcal{E}_0$ is a σ-algebra, and f is θ-invariant if and only if f is $\mathcal{E}_0$-measurable.

Definition 10.2. The map θ is called *ergodic* if $\mathcal{E}_0$ is μ-trivial, i.e. if $A \in \mathcal{E}_0$, then $\mu(A) = 0$ or $\mu(A^c) = 0$.

For a Markov chain, ergodicity implies irreducibility.

If f is θ-invariant, then θ is ergodic if and only if $f = c$, a constant, almost-surely. This is by showing $\mu(f^{-1}(-\infty, x)) = 0$ or $\mu(f^{-1}[x, \infty)) = 0$.

> **Example 10.1.**
>
> On $((0,1], \mathbb{B}, \lambda_{(0,1]})$, the maps:
>
> - $\theta_a(x) = x + a \pmod 1$,
> - $\theta(x) = 2x \pmod 1$,
>
> are measure preserving and ergodic, unless $a \in \mathbb{Q}$.

Theorem 10.1 (Birkhoff's Ergodic Theorem)**.** *Let $(E, \mathcal{E}, \mu)$ be σ-finite, $f \in L^1$ and $\theta : E \to E$ measure-preserving. Define $S_0 = 0$ and*

$$S_n = S_n(f) = f + f \circ \theta + f \circ \theta^2 + \cdots + f \circ \theta^{n-1}.$$

Then there exists a θ-invariant function $\bar{f}$ with $\mu(|\bar{f}|) \leq \mu(|f|)$ such that

$$\frac{S_n(f)}{n} \to \bar{f}$$

almost everywhere as $n \to \infty$.

If θ is ergodic, then $\bar{f} = c$ almost everywhere.

Lemma 10.1 (Maximal Ergodic Theorem)**.** *For $f \in L^1(\mu)$, set $S^* = S^*(f) = \sup_{n\geq 0} S_n(f)$ Then,*

$$\int_{\{S^*>0\}} f \, \mathrm{d}\mu \geq 0.$$

Proof: Set $S_n^* = \max_{0\leq m\leq n} S_m$. Then for $m = 1, 2, \ldots, n+1$,

$$S_m = f + S_{m-1} \circ \theta \leq f + S_n^* \circ \theta.$$

On $A_n = \{S_n^* > 0\}$, we have

$$S_n^* = \max_{1\leq m\leq n} S_m \leq \max_{1\leq m\leq n+1} S_m \leq f + S_n^* \circ \theta.$$

Hence integrating,

$$\int_{A_n} S_n^* \, \mathrm{d}\mu \leq \int_{A_n} f \, \mathrm{d}\mu + \int_{A_n} S_n^* \circ \theta \, \mathrm{d}\mu.$$

On A_n^c, we have $S_n^* = 0 \leq S_n^* \circ \theta$, as $S_n^* \geq 0$ since $S_0 = 0$. Thus,

$$\int_{A_n^c} S_n^* \, \mathrm{d}\mu \leq \int_{A_n^c} S_n^* \circ \theta \, \mathrm{d}\mu.$$

Adding these up,

$$\int_E S_n^* \, \mathrm{d}\mu \leq \int_{A_n} f \, \mathrm{d}\mu + \int_E S_n^* \circ \theta \, \mathrm{d}\mu = \int_{A_n} f \, \mathrm{d}\mu + \int_E S_n^* \, \mathrm{d}\mu.$$

Hence, as $S_n^* \in L^1$,

$$\int_{A_n} f \, \mathrm{d}\mu \geq 0,$$

for all n. But notice that

$$A_n = \{S_n^* > 0\} = \{\max_{1\leq m\leq n} S_m > 0\} = \bigcup_{m=0}^{n} \{S_m > 0\}$$

$$\uparrow \bigcup_{m=0}^{\infty} \{S_m > 0\} = \{\sup S_m > 0\}.$$

Hence $f\mathbb{1}_{A_n} \to f\mathbb{1}_{\{S^*>0\}}$. Hence as $|f\mathbb{1}_{A_n}| \leq |f|$ and $f \in L^1$, by DCT,

$$0 \leq \int_{A_n} f \, \mathrm{d}\mu \to \int_{\{S^*>0\}} f \, \mathrm{d}\mu,$$

proving the theorem.

Remark. Let μ be a finite measure. Then for $f \in L^1$ and any $\alpha > 0$, define

$$\bar{S}_k = \frac{S_k(f)}{f}, \qquad \bar{S}^* = \sup_{k \geq 0} \bar{S}_k.$$

Then,

$$\mu(\bar{S}^* > \alpha) \leq \frac{1}{\alpha} \int_{\{\bar{S}^* > \alpha\}} f \, d\mu \leq \frac{1}{\alpha} \int |f| \, d\mu.$$

This is stronger than Markov's inequality.

Remark. For μ a probability measure and $f \in L^1(\mu)$, we can show that

$$\left\{ \frac{S_n(f)}{n} \mid n \in \mathbb{N} \right\}$$

is uniformly integrable. Hence

$$\frac{S_n(f)}{n} \xrightarrow{L^1} \bar{f}.$$

If θ is ergodic, then this means $\bar{f} = \int f \, d\mu$ almost surely.

We now return to proving Birkhoff's ergodic theorem.

Proof: Note that

$$\mu(|S_n|) \leq \sum_{i=0}^{n-1} \mu(|f \circ \theta^i|) = \sum_{i=0}^{n-1} \mu(|f|) = n|f|.$$

So we have that

$$\mu\left(\left| \frac{S_n}{n} \right| \right) \leq \mu(|f|).$$

Therefore, we have

$$\mu(|\bar{f}|) = \mu\left(\liminf \left| \frac{S_n}{n} \right| \right) \leq \liminf \mu\left(\left| \frac{S_n}{n} \right| \right) \leq \mu(|f|).$$

This proves one part of the theorem. Note that

$$\frac{S_n \circ \theta}{n} = \frac{S_{n+1} - f}{n+1} \times \frac{n+1}{n}.$$

Hence we get that

$$\limsup \frac{S_n}{n} \circ \theta = \limsup \frac{S_{n+1}}{n+1} = \limsup \frac{S_n}{n}.$$

Similar results hold for $\liminf$. For $a < b$, define

$$D = D(a, b) = \left\{ \liminf_n \frac{S_n}{n} < a < b < \limsup_n \frac{S_n}{n} \right\}.$$

Then this is θ-invariant. We shall show that $\mu(D) = 0$. Assuming this, define

$$\Delta = \left\{ \liminf \frac{S_n}{n} < \limsup \frac{S_n}{n} \right\} = \bigcup_{\substack{a,b \in \mathbb{Q} \\ a < b}} D(a, b).$$

Now $\mu(D) = 0$ for all $a, b \implies \mu(\Delta) = 0$. Define

$$\bar{f} = \begin{cases} \liminf(S_n/n) = \limsup(S_n/n) & \text{on } \Delta^c, \\ 0 & \text{on } \Delta. \end{cases}$$

Then we have $S_n/n \to \bar{f}$ μ-almost everywhere and $\bar{f}$ is θ-invariant, since $\lim S_n/n$ is θ-invariant on Δ^c, and Δ is θ-invariant.

Now we will show that $D = D(a, b)$ has measure 0. Note that $\theta : D \to D$ by invariance, and θ is $\mu|_D$-measure preserving.

Also, we will assume without loss of generality that $b > 0$ by changing f to $-f$ if $b < 0$.

We shall apply maximal ergodic lemma on D with $\mu|_D$. For any $B \subseteq D$ measurable and $\mu(B) < \infty$, let $g = f - b\mathbb{1}_B$. Then $g \in L^1$, and on D,

$$S_n(g) = S_n(f) - bS_n(\mathbb{1}_B) \geq S_n(f) - nb > 0$$

for some n. Hence $S^*(g) = \sup S_n(g) > 0$ on D, so $\{S^*(g) > 0\} \cap D = D$. Then by maximal ergodic lemma on D,

$$\int_{\{S^*(g)>0\} \cap D} g \, \mathrm{d}\mu \geq 0,$$

or

$$0 \leq \int_D (f - b\mathbb{1}_B) \, \mathrm{d}\mu = \int_D f \, \mathrm{d}\mu - b\mu(B),$$

hence $b\mu(B) \leq \int_D f \, \mathrm{d}\mu$. Since μ is σ-finite, there exist measurable sets (B_n) such that $B_n \uparrow D$ and $\mu(B_n) < \infty$ for all n. Hence,

$$b\mu(D) = \lim_{n \to \infty} b\mu(B_n) \leq \int_D f \, \mathrm{d}\mu < \infty,$$

hence $\mu(D) < \infty$. Now since we now $\mu(D)$ is finite, a similar argument with $g' = (-f) - (-a)\mathbb{1}_D$, since D is θ-invariant we have $S_n(\mathbb{1}_D) = n\mathbb{1}_D$. We get that

$$(-a)\mu(D) \leq \int (-f)\,\mathrm{d}\mu.$$

Hence we have

$$b\mu(D) \leq \int_D f\,\mathrm{d}\mu \leq a\mu(D).$$

But $a < b$, and $\mu(D) < \infty$, hence $\mu(D) = 0$.

Theorem 10.2 (von Neumann's L^p Ergodic Theorem). *Let $\mu(E) < \infty$ and $1 \leq p < \infty$. Then for $f \in L^p(\mu)$,*

$$\frac{S_n(f)}{n} \to \bar{f}$$

in L^p, as $n \to \infty$.

Proof: Note that
$$\|f \circ \theta^n\|_p = \|f\|_p,$$

as θ is μ-measure preserving. So by Minkowski's inequality,

$$\left\|\frac{S_n(f)}{n}\right\|_p \leq \frac{1}{n}\sum_{i=0}^{n-1} \|f \circ \theta^i\|_p = \|f\|_p.$$

Now, let $\varepsilon > 0$ by given. Then choose $K < 0$ such that $\|f - f_K\|_p < \varepsilon/3$, where $f_K - (-K) \vee f \wedge (K)$.

But f_K is bounded and μ is a finite measure, so $f_K \in L^1(\mu)$. Hence by Birkhoff, there exists $\bar{f} \in L^1$ such that

$$\frac{S_n(f_K)}{n} \to \bar{f}_K$$

almost everywhere. Again, as $f \in L^p(\mu) \subseteq L^1(\mu)$ as μ is a finite measure, by Birkhoff there exists $\bar{f} \in L^1$ such that

$$\frac{S_n(f)}{n} \to \bar{f}$$

almost everywhere. Now by Fatou,

$$\|\bar{f} - \bar{f}_K\|_p^p = \int |\bar{f} - \bar{f}_K|^p \, \mathrm{d}\mu = \int \liminf_n \left| \frac{S_n(f)}{n} - \frac{S_n(f_K)}{n} \right|^p \mathrm{d}\mu$$

$$= \int \liminf_n \left| \frac{S_n(f - f_K)}{n} \right|^p \mathrm{d}\mu \leq \liminf \int \left| \frac{S_n(f - f_K)}{n} \right|^p \mathrm{d}\mu$$

$$= \liminf \left\| \frac{S_n(f - f_K)}{n} \right\|_p^p = \|f - f_K\|_p^p \leq \frac{\varepsilon}{3}.$$

By Minkowski,

$$\left\| \frac{S_n(f)}{n} - \bar{f} \right\|_p \leq \left\| \frac{S_n(f)}{n} - \frac{S_n(f_K)}{n} \right\|_p + \|\bar{f} - \bar{f}_K\|_p + \left\| \frac{S_n(f_K)}{n} - \bar{f}_K \right\|_p$$

$$\leq \frac{2\varepsilon}{3} + \left\| \frac{S_n(f_K)}{n} - \bar{f}_K \right\|_p.$$

So it is enough to show that the latter term converges to 0. But note that

$$\left\| \frac{S_n(f_K)}{n} - \bar{f}_K \right\|_p^p = \int \left| \frac{S_n(f_K)}{n} - \bar{f}_K \right|^p \mathrm{d}\mu \to 0,$$

by bounded convergence theorem, as $S_n(f_K)/n \to \bar{f}_K$ almost everywhere.

Remark.

1. As we have seen before, if μ is a probability measure and θ is ergodic, then $\bar{f}$ is constant almost surely, and

$$\frac{S_n(f)}{n} \to \mathbb{E}[f]$$

   μ-almost surely, and in L^1.

2. For θ measure-preserving and $f \in L^1$, we get a similar statement:

$$\frac{S_n(f)}{n} \to \mathbb{E}[f|\mathcal{E}_0],$$

   μ-almost surely and in L^1. For $f \in L^2$, this is just the projection of f on $L^2(\mathcal{E}_0)$.

10.1   Bernoulli Shifts

On the infinite product space $E = \mathbb{R}^{\mathbb{N}} = \{x = (x_1, x_2, \ldots) \mid x_i \in \mathbb{R}\}$, consider the cylinder sets

$$\mathcal{A} = \left\{ \prod_{n=1}^{\infty} A_n \mid A_n \in \mathbb{B} \text{ and } A_n = \mathbb{R} \text{ for all } n \geq N \right\}.$$

Then $\mathcal{A}$ is a π-system, and let $\mathcal{E} = \sigma(\mathcal{A})$.

Then we can check that $\mathcal{E} = \sigma(\mathcal{A}) = \sigma(f_n \mid n \in \mathbb{N})$ where $f_n : E \to \mathbb{R}$, $f_n(x) = x_n$ are the coordinate maps. $\mathcal{E}$ is the Borel σ-algebra generated by the topology of pointwise convergence.

Now consider a sequence of iid random variables (X_n) on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with common distribution or law $\mu_{X_n} = \mathbb{P} \circ X_n^{-1} = m$, for all n. The map $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ given by

$$X(\omega) = (X_1(\omega), X_2(\omega), \ldots)$$

is measurable. The image measure $\mathbb{P} \circ X^{-1} = \mu$ is a probability measure on $(E, \mathcal{E})$ that satisfies, for any $A = A_1 \times \cdots \times A_N \times \mathbb{R} \times \cdots \subset \mathcal{A}$, that

$$\begin{aligned}
\mu(A) = \mathbb{P} \circ X^{-1}(A) &= \mathbb{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_N \in A_N) \\
&= \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2) \cdots \mathbb{P}(X_N \in A_N) \\
&= m(A_1)m(A_2) \cdots m(A_N) = \prod_{n=1}^{\infty} m(A_i),
\end{aligned}$$

and is the unique probability measure on $\mathcal{E}$ such that

$$\mu(A) = \prod_{n=1}^{\infty} m(A_n).$$

Under μ, the coordinate maps f_n are iid with the law m. The probability space $(E, \mathcal{E}, \mu)$ is called the *canonical model* for an iid sequence of random variables with law m.

Define the *shift map* $\theta : E \to E$ by

$$\theta(x_1, x_2, \ldots) = (x_2, x_3, \ldots).$$

Theorem 10.3. *On $(E, \mathcal{E}, \mu)$, the shift map θ is measurable, measure-preserving and ergodic.*

Proof: Being measurable is obvious. To show it is measure-preserving, it is enough to show for $A = A_1 \times \cdots \times A_n \times \mathbb{R} \times \cdots$. Then note that

$$\mu \circ \theta^{-1}(A) = \mu(\mathbb{R} \times A_1 \times \cdots) = \prod_{i=1}^{\infty} m(A_i) = \mu(A).$$

Now we show ergodicity. Recall the tail σ-algebra $\tau = \bigcap \tau_n$, where

$$\tau_n = \sigma(x_{n+1}, x_{n+2}, \ldots) = \sigma(f_{n+1}, f_{n+2}, \ldots).$$

For $A = \prod A_n \in \mathcal{A}$, note that

$$\theta^{-n}(A) = \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{n \text{ times}} \times A_1 \times \cdots = \{x_{n+1} \in A_1, x_{n+2} \in A_2, \ldots\} \in \tau_n.$$

If $A \in \mathcal{E}_0$, then $\theta^{-1}(A) = A$, so $\theta^{-n}(A) = A$ for all n. So $A \in \tau_n$ for all n, hence $A \in \bigcap \tau_n = \tau$. Hence $\mathcal{E}_0 \subseteq \tau$.

But the (x_i) are iid, and hence τ is μ-trivial from Kolmogorov 0-1 law, so $\mathcal{E}_0$ is μ-trivial.

10.2   Strong Law of Large Numbers

Theorem 10.4. *Let m be a probability measure on $\mathbb{R}$ such that*

$$\int_{\mathbb{R}} |x| \, \mathrm{d}m(x) < \infty \text{ and } \int_{\mathbb{R}} x \, \mathrm{d}m(x) = \nu.$$

Let $(E, \mathcal{E}, \mu)$ be the canonical model, where the coordinate maps $f_n(x) = x_n$ are iid with law m. Then

$$\mu(\{x \mid (x_1 + \cdots + x_n)/n \to \nu \text{ as } n \to \infty\}) = 1.$$

Proof: Let $\theta : E \to E$ be the shift map. It is measure-preserving and ergodic.

Consider $f : E \to \mathbb{R}$ by $f(x) = x_1$. Then $f \in L^1(\mu)$, as

$$\int |f| \, \mathrm{d}\mu = \int |x_1| \, \mathrm{d}m(x_1) < \infty,$$

and also

$$S_n(f) = f + f \circ \theta + \cdots + f \circ \theta^{n-1} = x_1 + x_2 + \cdots + x_n.$$

Hence by Birkhoff and von-Neumann, as θ is ergodic,

$$\frac{S_n(f)}{n} = \frac{x_1 + \cdots + x_n}{n} \to \bar{f} = \int f \, \mathrm{d}\mu = \int x_1 \, \mathrm{d}m(x_1) = \nu,$$

μ-almost surely.

Theorem 10.5 (Kolmogorov SLLN). *Let (X_n) be a sequence of iid integrable random variables, with $\mathbb{E}[X_1] = \nu$. Set $S_n = \sum_{i=1}^{n} X_i$. Then*

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

almost surely as $n \to \infty$.

Proof: Let m be the law of X_n, and $\mu = \mathbb{P} \circ X^{-1}$, where $X : \Omega \to E$ is

$$X(\omega) = (X_1(\omega), X_2(\omega), \ldots).$$

Then we can apply the previous theorem.

Remark.

1. (μ_n) is a sequence of probability measures that converges weakly to μ, then (μ_n) is "tight": for all ε, there exists a compact set K such that $\mu_n(K') < \varepsilon$.

2. If (μ_n) is a sequence of probability measures that are tight, then there exists a subsequence (n_k) and a probability measure μ such that

$$\mu_{n_k} \overset{\text{weakly}}{\to} \mu.$$

   This is the Banach-Alaoglu theorem, or the Prokhorov theorem when applied to probability spaces directly.

3. The SLLN says if X_i are iid with $\mathbb{E}[X_1] = 0$, then

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

   almost surely. But the CLT says that if $\mathbb{E}[X_1^2] = 1$, then

$$\frac{S_n}{\sqrt{n}} \to N(0,1)$$

   weakly. In fact, we get

$$\frac{S_n}{n^{1/2}(\log n)^{1/2+\delta}} \to 0$$

almost surely, for any $\delta > 0$, and Khinchin showed that

$$\limsup \frac{S_n}{\sqrt{2n \log \log n}} \to 1$$

almost surely. This is the *law of the iterated logarithm*.

4. If $F_n \xrightarrow{\text{d}} F$, then there exists a probability space such that μ_{X_n} is F_n, and $X_n \to X$ almost surely.

5. If $X_n \xrightarrow{\text{d}} X$ and $Y_n \xrightarrow{\mathbb{P}} c$, then $(X_n, Y_n) \xrightarrow{\text{d}} (X, c)$, so $X_n + Y_n \xrightarrow{\text{d}} X + c$.

Index