Probability and Measure

Course notes for STAT 571

Adam B Kashlak
Mathematical & Statistical Sciences
University of Alberta
Edmonton, Canada, T6G 2G1

March 25, 2022

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/.

Contents

Preface

> It is hard to explain just how a single sight of a tangible object with measurable dimensions could so shake and change a man.
>
> *The Case of Charles Dexter Ward*
> H. P. Lovecraft (1941)

The following are course notes for measure theory and probability theory. These are intended for students who have background in mathematics and probability, but have not seen measure theory yet. Hence, the first half of these notes focus on measure theory with the second half discuss how probability theory fits into the setting of measure theory. It was only about 100 years ago that mathematicians like Kolmogorov and Von Mises were trying to formalize probability theory while in France, Borel, Baire, and Lebesgue where working on analysis, and across the channel, Fisher, Pearson, Jeffreys, and their contemporaries were bringing rigour to the field of statistics. Hence, the material in this course is still relatively young.

The sources I used to put these notes together are quite numerous. The main textbooks I relied on are *Real Analysis and Probability* by RM Dudley and *Probability and Measure* by Patrick Billingsley. I first learned measure theory apart from probability theory in a course at McGill University from Dr Paul Koosis. He used W Rudin's text *Real and Complex Analysis*. Later, as a PhD student, I learned more formal probability theory from reading the course notes of Dr James R Noris from the University of Cambridge. I also attended lectures in advanced probability theory from Dr Perla Sousi and Dr Alan Sola whose course notes I often review. Lastly, I also have referenced the previous iteration of this course taught by Dr Michael Kouritzin here at the University of Alberta. I hope you will find these notes as useful to your own understanding of the subject as I have found the notes written by others for my own understanding.

Adam B Kashlak
Edmonton, Canada
January 2022

Chapter 1

Measure Theory

Introduction

We innately understand the concept of a measure in the context of lengths, areas, and volumes. In a mathematical context, a measure assigns a non-negative value to a set. For example, on the plane $\mathbb{R}^2$, we can consider the area of a subset $A \subset \mathbb{R}^2$. In this case, A needs to be *measureable*, which we will make more precise below. If A is a rectangle, then we can say its area is the length times the width. If A is a union of disjoint rectangles, we can sum the area of each individual rectangle to get the total area of A.

On the real line $\mathbb{R}$, we can similarly say that the measure of an interval is the length of that interval. Hence, the measure of $[a, b]$ for $-\infty < a < b < \infty$ is just $b-a$. However, given a probability distribution function $\Phi(x) = \mathrm{P}(X < x)$ on the real line, we can also define a measure of $[a, b]$ to be $\Phi(b) - \Phi(a)$. In this case, the measure will always take a value between 0 and 1. This is an example of a probability measure on $\mathbb{R}$.

Notation

The set $\mathbb{R}$ denotes the real numbers and $\mathbb{R}^p$ is the space of p-dimensional real valued vectors. Also, $\mathbb{Z}$ is the set of integers, $\mathbb{Q}$ is the set of rational numbers, and $\mathbb{C}$ is the set of complex numbers. $\emptyset$ is the empty set or null set. Typically, Ω will be the space we are working in—e.g. $\mathbb{R}$ or $\mathbb{R}^p$. For a set $A = \{x \in \Omega : x \in A\}$, the complement $A^c = \{x \in \Omega : x \notin A\}$. A collection of sets $\{A_i\}_{i=1}^{\infty}$ is said to be pairwise disjoint if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

1.1 Measures and σ-fields

Formally, we need to define what a measure is and what sets can be measured.

Definition 1.1.1 (σ-field)**.** *For some set Ω, a σ-field $\mathcal{F}$ is a collection of sets $A \subseteq \Omega$ such that*

1. $\emptyset, \Omega \in \mathcal{F}$

finite intersections

2. *if* $A \in \mathcal{F}$ *then* $A^c \in \mathcal{F}$
3. *for a countable collection of sets* $\{A_i\}_{i=1}^{\infty}$ *such that* $A_i \in \mathcal{F}$ *for* $i = 1, \ldots, \infty$, $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Note that this definition implies that a σ-field also contains countable intersections of sets. Indeed, if $\{A_i\}_{i=1}^{\infty} \in \mathcal{F}$ then

$$\left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c \;\Rightarrow\; \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}.$$

This follows from *De Morgan's laws*.

Definition 1.1.2 (Measure)**.** *For a measure space* $(\Omega, \mathcal{F})$*, a measure* $\mu : \mathcal{F} \to \mathbb{R}^+$ *such that*

1. $\mu(\emptyset) = 0$
2. μ *is countably additive—i.e. for any pairwise disjoint countable collection of sets* $\{A_i\}_{i=1}^{\infty}$*, then*

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

Definition 1.1.3. *There are a few special cases of measures* μ *that will be of interest. For a measure space* $(\Omega, \mathcal{F}, \mu)$*, we say that*

- μ *is a probability measure if* $\mu(\Omega) = 1$.
- μ *is a finite measure if* $\mu(\Omega) < \infty$.
- μ *is a* σ*-finte measure if* $\Omega = \bigcup_{i=1}^{\infty} A_i$ *such that* $\mu(A_i) < \infty$ *for all* i.

Note that if μ *is a probability measure, then we say that* $(\Omega, \mathcal{F}, \mu)$ *is a probability space. In this case, the measure is often written as* P *instead of* μ.

The next question is how to construct a measure space $(\Omega, \mathcal{F}, \mu)$ to work with. Given a set Ω, we can define the *power set* $\mathcal{P}(\Omega)$ to be the set of all subsets of Ω. Hence, $\mathcal{F} \subset \mathcal{P}$ for any σ-field on Ω. Typically, $\mathcal{P}(\Omega)$ is much to large to work with. A notable example is the finite space with counting measure.

Example 1.1.4 (Counting Measure)**.** *Let* $\Omega = \{1, \ldots, n\}$*, then the power set* $\mathcal{P}(\Omega)$*, sometimes denoted as* 2^{Ω}*, contains all* 2^n *subsets of* $\{1, \ldots, n\}$*. The counting measure* μ *counts the number of elements in a set* $A \in \mathcal{P}(\Omega)$*. If we normalize this measure, then we can think of it as a uniform distribution on the integers from 1 to* n*. For example,*

- $\mu(\{1, 3, 7\}) = 3$
- $\frac{1}{n}\mu(\{1, 3, 7\}) = \frac{3}{n}$

Instead of just assigning a weight of $1/n$ *to each integer, we could, for example, assign a binomial probability* $\binom{n}{i} p^i (1-p)^i$ *for some* $p \in (0, 1)$*. The same can be done with the Poission probabilities,* $\mathrm{e}^{-\lambda}\lambda^i / i!$ *for some* $\lambda > 0$*, and taking* $n \to \infty$.

1.2 Constructing σ-fields and measures

Consider starting with $\Omega = \mathbb{R}$ and $\mathcal{I}$ the set of all open intervals of the form (a, b) with $-\infty < a < b < \infty$. Then, the *length* of the interval can be its measure. More formally, we define $\lambda((b-a)) = b - a$. This λ will be the famous Lebesgue measure. However, the set $\mathcal{I}$ is not a σ-field as $(a, b) \cup (c, d) \notin \mathcal{I}$ for any $-\infty < a < b < c < d < \infty$. Hence, starting from $\mathcal{I}$ (or really any set of subsets of Ω), how do we construct a sensible σ-field $\mathcal{F}$ to work with? Furthermore, how do we extend a measure on $\mathcal{I}$ to a measure on $\mathcal{F}$? Lastly, is such an extension unique?

Let Ω be some set and $\mathcal{A}$ a set of subsets of Ω not a σ-field. Then, we can consider the smallest σ-field that contains $\mathcal{A}$ defined as

$$\sigma(\mathcal{A}) := \{B \subseteq \Omega \,:\, B \in \mathcal{F},\ \forall\, \mathcal{F} \text{ such that } \mathcal{A} \subset \mathcal{F}\}\,.$$

Furthermore, let μ be a measure on $\mathcal{A}$. Then, we want to show (1) that μ can be extended to a measure on $\sigma(\mathcal{A})$ and (2) that this extension is unique.

To make sure that $\sigma(\mathcal{A})$ is actually interesting, we will consider sets of sets $\mathcal{A}$ that are *semirings*, *rings*, or *fields*.

Definition 1.2.1 (Semiring)**.** *A collection of sets $\mathcal{A}$ of Ω is a semiring if $\emptyset \in \mathcal{A}$ and for all $A, B \in \mathcal{A}$ then $A \cap B \in \mathcal{A}$ and $B \setminus A = \bigcup_{i=1}^{n} C_i$ where $C_i \in \mathcal{A}$ for $i = 1, \ldots, n$.*

Definition 1.2.2 (Ring)**.** *A collection of sets $\mathcal{A}$ of Ω is a ring if $\emptyset \in \mathcal{A}$ and for all $A, B \in \mathcal{A}$ both $B \setminus A \in \mathcal{A}$ and $A \cup B \in \mathcal{A}$.*

Definition 1.2.3 (Field)**.** *A ring $\mathcal{A}$ is a field if $\Omega \in \mathcal{A}$.*

Note that fields and σ-fields are sometimes referred to as algebras and σ-algebras, respectively. For more on why semirings are a thing, see [Dudley(2002)], section 3.2.

Definition 1.2.4 (Set Functions)**.** *For a general set function $\mu : \mathcal{A} \to \mathbb{R}^+$ (i.e. not necessarilly a measure) and $A, B \in \mathcal{A}$, we say that*

- *μ is increasing if for $A \subset B$, $\mu(A) \leq \mu(B)$.*
- *μ is additive if for A, B disjoint, $\mu(A \cup B) = \mu(A) + \mu(B)$.*
- *μ is countably additive if for $\{A_i\}_{i=1}^{\infty}$ pairwise disjoint with $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$, $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.*
- *μ is countably subadditive if for $\{A_i\}_{i=1}^{\infty}$ with $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$, $\mu(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$.*

Note that such a set function μ is sometimes referred to as a *pre-measure* when it is countably additive and $\mu(\emptyset) = 0$.

1.2.1 Existence

In this subsection, we prove what is typically referred to as the Carathéodory Extension Theorem. A similar proof can be found in [Billingsley(2008)] Chapter 1 Section 3 but is restricted to probability measures.

First, we define the *outer measure* for μ and $\mathcal{A}$ to be

$$\mu^*(E) = \inf \sum_i \mu(A_i), \text{ for any } E \subset \Omega$$

where the infimum is over all finite and countably infinite collections $\{A_i\}$ such that $E \subseteq \bigcup_i A_i$. Furthermore, let $\mathcal{M}$ be the set of all μ^*-measureable sets where a set B is said to be μ^*-measureable if

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

for all $E \subseteq \Omega$. Our aim is to show that μ^* is the *correct* way to extend μ from $\mathcal{A}$ to $\sigma(\mathcal{A})$. We also want to show that $\mathcal{M}$ is a σ-field and that it contains $\sigma(\mathcal{A})$.

Theorem 1.2.1 (Carathéodory Extension Theorem)**.** *Let $\mathcal{A}$ be a ring on Ω and μ be a pre-measure. Then, μ extends to a measure on $\sigma(\mathcal{A})$.*

Proof. This proof will proceed in multiple steps. We assume that the $B \subseteq \Omega$ below have finite measure $\mu^*(B) < \infty$. Otherwise, the results can still be shown to trivially hold.

(1) We first prove a few properties of μ^*.

1. $\mu^*(\emptyset) = 0$, which follows from μ being a pre-measure.
2. μ^* is non-negative for all $B \subset \Omega$, which follows from the non-negativity of μ.
3. μ^* is monotone. Let $B_1, B_2 \in \mathcal{A}$ and $B_1 \subset B_2$, then for any $\{A_i\}$ such that $B_2 \subseteq \bigcup_i A_i$, $B_1 \subseteq \bigcup_i A_i$. Therefore $\mu^*(B_1) \leq \mu^*(B_2)$.
4. μ^* is countably subadditive. For $\{B_i\}_{i=1}^\infty$ and a given $\varepsilon > 0$, let $B_i \subseteq \bigcup_j A_{ij}$ for $A_{ij} \in \mathcal{A}$ such that $\sum_j \mu(A_{ij}) \leq \mu^*(B_i) + \varepsilon 2^{-i}$. As $\bigcup_{i=1}^\infty B_i \subseteq \bigcup_{i,j} A_{ij}$ and μ^* is monotone and μ is subadditive,

$$\mu^*\left(\bigcup_{i=1}^\infty B_i\right) \leq \mu\left(\bigcup_{i,j} A_{ij}\right) \leq \sum_{i,j} \mu(A_{ij}) \leq \sum_{i,j} \mu^*(B_{ij}) + \varepsilon.$$

   As $\varepsilon > 0$ is arbitrary, this implies μ^* is countably subadditive.

(2) Check that μ and μ^* coincide on $\mathcal{A}$. For any $A \in \mathcal{A}$, we immediately have that $\mu^*(A) \leq \mu(A)$ since $A \subseteq A$. For the reverse, if $A \subset \bigcup_i A_i$, then by countable subadditivity and monotonicity $\mu(A) \leq \sum_i \mu(A \cap A_i) \leq \sum_i \mu(A_i)$. Thus, $\mu(A) \leq \mu^*(A)$ and finally $\mu(A) = \mu^*(A)$.

(3) Check that $\mathcal{A} \subset \mathcal{M}$—i.e. for any $A \in \mathcal{A}$, we need to show that A is μ^*-measureable. Hence, for any $A \in \mathcal{A}$ and all $E \subseteq \Omega$, we want

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

As $\mu^*(E \cap A) + \mu^*(E \cap A^c) \geq \mu^*(E)$ by subadditivity, it suffices to show that $\mu^*(E \cap A) + \mu^*(E \cap A^c) \leq \mu^*(E)$. For some $\varepsilon > 0$, choose $\{A_i\}$ such that $E \subseteq \bigcup_i A_i$ and $\sum_i \mu(A_i) \leq \mu^*(E) + \varepsilon$. Furthermore,

$$\begin{aligned} E \cap A &\subseteq \textstyle\bigcup_i (A \cap A_i) \\ E \cap A^c &\subseteq \textstyle\bigcup_i (A^c \cap A_i). \end{aligned}$$

Thus,

$$\mu^*(E \cap A) + \mu^*(E \cap A^c) \leq \sum_i \mu(A \cap A_i) + \sum_i \mu(A^c \cap A_i) = \sum_i \mu(A_i) \leq \mu^*(E) + \varepsilon,$$

which shows that $\mathcal{A} \subset \mathcal{M}$ since ε was arbitrary.

(4) Show that $\mathcal{M}$ is a σ-field. We first check that $\mathcal{M}$ is a field. $\emptyset \in \mathcal{M}$ as $\emptyset \in \mathcal{A}$. For Ω and any $E \subseteq \Omega$,

$$\mu^*(E \cap \Omega) + \mu^*(E \cap \emptyset) = \mu^*(E)$$

and hence $\Omega \in \mathcal{M}$. Next, since $A \cap B = (A^c \cup B^c)^c$, we will show that $\mathcal{M}$ is closed under intersections. For $B_1, B_2 \in \mathcal{M}$ and any $E \subset \Omega$,

$$\begin{aligned} \mu^*(E) &= \mu^*(B_1 \cap E) + \mu^*(B_1^c \cap E) \\ &= \mu^*(B_2 \cap B_1 \cap E) + \mu^*(B_2 \cap B_1^c \cap E) + \\ &\quad \mu^*(B_2 \cap^c B_1 \cap E) + \mu^*(B_2^c \cap B_1^c \cap E) \\ &\geq \mu^*(B_2 \cap B_1 \cap E) + \\ &\quad \mu^*\left(\{B_2 \cap B_1^c \cap E\} \cup \{B_2 \cap^c B_1 \cap E\} \cup \{B_2^c \cap B_1^c \cap E\}\right) \\ &= \mu^*(\{B_2 \cap B_1\} \cap E) + \mu^*(\{B_2 \cap B_1\}^c \cap E) \\ &\geq \mu^*(E). \end{aligned}$$

Hence, the $B_2 \cap B_1 \in \mathcal{M}$. Lastly, since $B \setminus A = B \cap A^c$, we need to show $\mathcal{M}$ is closed under complementation, which trivially follows from the definition as for any $B \in \mathcal{M}$,

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

To extend from a field to a σ-field, we need to show that for a countable pairwise disjoint collection $\{B_i\}$ in $\mathcal{M}$, that $\bigcup_i B_i \in \mathcal{M}$.[1] Let $B = \bigcup_{i=1}^{\infty} B_i$. Proceeding, once

[1] It suffices to consider pairwise disjoint sets as countable unions of arbitrary sets $\bigcup_i A_i = \bigcup_i B_i$ where $B_i = A_i \setminus (\bigcup_{j=1}^{i} B_j)$ are pairwise disjoint.

again, from the definition,

$$\begin{aligned}\mu^*(E) &= \mu^*(E \cap B_1) + \mu^*(E \cap B_1^c) \\ &= \mu^*(E \cap B_1) + \mu^*(E \cap B_2) + \mu^*(E \cap B_1^c \cap B_2^c) \\ &= \sum_{i=1}^{n} \mu^*(E \cap B_i) + \mu^*\left(E \cap \{\textstyle\bigcap_{i=1}^n B_i^c\}\right).\end{aligned}$$

By monotonicity, subadditivity, and taking $n \to \infty$, we get

$$\mu^*(E) \geq \sum_{i=1}^{\infty} \mu^*(E \cap B_i) + \mu^*(E \cap B^c) \geq \mu^*(E \cap B) + \mu^*(E \cap B^c) \geq \mu^*(E).$$

Thus, $\mathcal{M}$ is closed under countable unions. Finally, choosing $E = B$ above, we have

$$\mu^*(E) = \sum_{i=1}^{\infty} \mu^*(E \cap B_i)$$

and thus μ^* is countably additive.

(5) The Conclusion. What we have from all of the above is that μ^* is a set function on the power set $\mathcal{P}(\Omega)$, but it is also a measure in $\mathcal{M} \subset \mathcal{P}(\Omega)$. Furthermore, since $\mathcal{A} \subset \mathcal{M}$ and $\mathcal{M}$ is a σ-field, we have that $\sigma(\mathcal{A}) \subseteq \mathcal{M}$. Lastly, since μ^* is a measure on $\mathcal{M}$ it is also a measure on any sub-σ-field. Hence, it is a measure on $\sigma(\mathcal{A})$. □

Example 1.2.5 (Lebesgue Measure)**.** *We can construct Lebesgue measure on the half-open unit interval* $(0, 1]$ *by considering* $\mathcal{I}$ *to be the set of all finite disjoint unions of half open intervals of the form* $(a, b]$ *for* $0 \leq a < b \leq 1$ *along with the empty set* $\emptyset$ *for length 0 intervals. That is,* $A \in \mathcal{I}$ *is of the form* $A = \bigcup_{j=1}^n I_j$ *where* $\{I_j\}_{j=1}^n$ *are pairwise disjoint half-open intervals. Then,* $\lambda(A) = \sum_{j=1}^n \lambda(I_j)$ *is a set function where* $\lambda(I_j)$ *is just the length of the interval* I_j*.*

We can check that $\mathcal{I}$ *is, in fact, a ring and that* λ *is a pre-measure. Thus, the Carathéodory extension theorem tells us that* λ *can be extended to a* σ*-field. In this case, we have*

$$\mathcal{I} \subset \sigma(\mathcal{I}) \subset \mathcal{M} \subset \mathcal{P}((0, 1]).$$

The σ*-field* $\sigma(\mathcal{I})$ *is the Borel* σ*-field and is often written as* $\mathcal{B}$*. The set* $\mathcal{M}$ *contains all Lebesgue measurable subsets of the unit interval, and it is strictly larger than* $\mathcal{B}$*. Also, the power set* $\mathcal{P}((0, 1])$ *contains subsets of* $(0, 1]$ *that are not Lebesgue measurable. It's very non-trivial to construct sets that fall into these categories, but this leads to some very interesting excursions.*

1.2.2 Uniqueness

Given such an extension as above, we wish to know whether or not it is unique. That is, if μ_1 and μ_2 are measures on $\sigma(\mathcal{A})$ and if $\mu_1(A) = \mu_2(A)$ for any $A \in \mathcal{A}$, then is it also true that $\mu_1(B) = \mu_2(B)$ for any $B \in \sigma(\mathcal{A})$? Answer this question, we require two more definitions.

Definition 1.2.6 (π-system)**.** *A collection of subsets $\mathcal{A}$ is called a π-system if $\emptyset \in \mathcal{A}$ and for $A, B \in \mathcal{A}$, then $A \cap B \in \mathcal{A}$.*

Definition 1.2.7 (λ-system)**.** *A collection of subsets $\mathcal{L}$ is called a λ-system if $\Omega \in \mathcal{L}$ and*

- *for $A, B \in \mathcal{L}$ with $A \subset B$, then $B \setminus A \in \mathcal{L}$.*
- *for $\{A_i\}_{i=1}^{\infty}$ pairwise disjoint, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{L}$.*

Hence, a λ-system is very similar to a σ-field except it is only closed under countable *disjoint* unions. Note that if $\mathcal{A}$ is a field, then it is a π-system. The theorem we want to prove is as follows.

Theorem 1.2.2 (Uniqueness of Extension)**.** *Let μ_1 and μ_2 be σ-finite measures on $\sigma(\mathcal{A})$ where $\mathcal{A}$ is a π-system. Then if $\mu_1(A)$ and $\mu_2(A)$ agree for all $A \in \mathcal{A}$, then μ_1 and μ_2 agree on $\sigma(\mathcal{A})$.*[2]

To prove this theorem, we will make use of the Dynkin π-λ Theorem and the fact that if $\mathcal{F}$ is both a π-system and a λ-system, then it is a σ-field.[3]

Theorem 1.2.3 (Dynkin π-λ Theorem)**.** *Let $\mathcal{A}$ be a π-system, $\mathcal{L}$ be a λ-system, and $\mathcal{A} \subset \mathcal{L}$. Then, $\sigma(\mathcal{A}) \subset \mathcal{L}$.*

Proof. Let $\mathcal{L}_0$ be the smallest λ-system such that $\mathcal{A} \subset \mathcal{L}_0$. Then, we have that $\mathcal{L}_0 \subseteq \mathcal{L}$. We aim to show that $\mathcal{L}_0$ is also a π-system and hence also a σ-field so that necessarily $\sigma(\mathcal{A}) \subset \mathcal{L}_0$. Thus, we need $\mathcal{L}_0$ to be closed under intersections.

Let $\mathcal{L}' = \{B \in \mathcal{L}_0 : B \cap A \in \mathcal{L}_0 \ \forall A \in \mathcal{A}\}$. Then, $\mathcal{A} \in \mathcal{L}'$ as $\mathcal{A}$ is a π-system. We will show that $\mathcal{L}'$ is also a λ-system. Indeed, $\Omega \in \mathcal{L}'$ as $\mathcal{A} \subset \mathcal{L}_0$ and

- if $B_1, B_2 \in \mathcal{L}'$ such that $B_1 \subset B_2$, then we have for any $A \in \mathcal{A}$ that $B_1 \cap A, B_2 \cap A \in \mathcal{L}_0$. Thus, $(B_2 \cap A) \setminus (B_1 \cap A) = (B_2 \setminus B_1) \cap A \in \mathcal{L}_0$. Thus, $B_2 \setminus B_1 \in \mathcal{L}'$.
- if $\{B_i\}_{i=1}^{\infty} \in \mathcal{L}'$ are pairwise disjoint, then for all i and $A \in \mathcal{A}$, $A \cap B_i \in \mathcal{L}_0$ thus $\bigcup_{i=1}^{\infty}(A \cap B_i) = A \cap (\bigcup_{i=1}^{\infty} B_i) \in \mathcal{L}_0$. Hence, $\bigcup_{i=1}^{\infty} B_i \in \mathcal{L}'$.

By definition $\mathcal{L}' \subset \mathcal{L}_0$, but as $\mathcal{L}_0$ is minimal the reverse is true, and thus $\mathcal{L}' = \mathcal{L}_0$. Hence, $\mathcal{L}_0$ contains all intersections with sets in $\mathcal{A}$.

Next, let $\mathcal{L}'' = \{B \in \mathcal{L}_0 : B \cap C \in \mathcal{L}_0 \ \forall C \in \mathcal{L}_0\}$. Thus, $\mathcal{L}_0 = \mathcal{L}'$ implies that $\mathcal{A} \subset \mathcal{L}''$. Using the same arguements as for L', it can be shown that L'' is a λ-system and thus $\mathcal{L}'' = \mathcal{L}_0$. This implies that $\mathcal{L}_0$ is closed under intersections and hence a π-system and hence a σ-field and hence contains $\sigma(\mathcal{A})$. □

Proof of Theorem 1.2.2 for finite measures. (This is the easier proof for finite measures)

If we additionally assume that $\mu_1(\Omega) = \mu_2(\Omega) < \infty$ (i.e. μ_1, μ_2 are finite measures), then the proof is much simpler. This condition is immediately true for probability measures, which is the main focus of this course.

[2] By *agree*, we mean $\mu_1(A) = \mu_2(A)$ if finite and $\mu_1(A)\infty \leftrightarrow \mu_2(A) = \infty$.

[3] **Exercise:** Show that this fact is true, or see Lemma 6 in Section 3 of [Billingsley(2008)].

Let $\mathcal{L} = \{B \subset \Omega : \mu_1(B) = \mu_2(B)\}$. We will show that $\mathcal{L}$ is a λ-system and then apply Theorem 1.2.3 to prove this theorem. By assumption, $\Omega \in \mathcal{L}$. Secondly, if $A, B \in \mathcal{L}$ with $A \subset B$, then

$$\mu_1(B \setminus A) + \mu_1(A) = \mu_1(B) = \mu_2(B) = \mu_2(B \setminus A) + \mu_2(A) < \infty.$$

Hence, $B \setminus A \in \mathcal{L}$. Lastly, for $\{A_i\}_{i=1}^{\infty}$ pairwise disjoint with $A_i \in \mathcal{L}$,

$$\mu_1\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu_1(A_i) = \sum_{i=1}^{\infty} \mu_2(A_i) = \mu_2\left(\bigcup_{i=1}^{\infty} A_i\right) < \infty.$$

Hence, $\bigcup_{i=1}^{\infty} A_i \in \mathcal{L}$. Thus, $\mathcal{L}$ is a λ-system and contains $\mathcal{A}$. Hence, $\mathcal{L}$ contains $\sigma(\mathcal{A})$. Thus, μ_1 and μ_2 agree on $\sigma(\mathcal{A})$. □

Proof of Theorem 1.2.2 for σ-finite measures. (This is the more general proof)

For any A in $\mathcal{A}$ such that $\mu_1(A) = \mu_2(A) < \infty$, we define $\mathcal{L}_A$ to be the collection of sets $B \subseteq \Omega$ such that $\mu_1(A \cap B) = \mu_2(A \cap B)$. By a similar argument as above,[4] we have that $\mathcal{L}_A$ is a λ-system and hence $\sigma(\mathcal{A}) \subset \mathcal{L}_A$ by Theorem 1.2.3.

By σ-finiteness, we can decompose $\Omega = \bigcup_{i=1}^{\infty} A_i$ where $A_i \in \mathcal{A}$ and $\mu_1(A_i) = \mu_2(A_i) < \infty$ for all i. Thus, for any $B \in \sigma(\mathcal{A})$ and any n, we can similarly write

$$\mu_1\left(\bigcup_{i=1}^{n}(B \cap A_i)\right) = \sum_{i=1}^{n} \mu_1\left(B \cap A_i\right) - \sum_{i<j} \mu_1\left(B \cap A_i \cap A_j\right) + \dots$$

by the inclusion-exclusion formula. This same formula holds for μ_2. Furthermore, as $\mathcal{A}$ is a π-system, we have that $A_i \cap A_j \in \mathcal{A}$ and further intersections. Thus, the $\mu_1\left(\bigcup_{i=1}^{n}(B \cap A_i)\right) = \mu_2\left(\bigcup_{i=1}^{n}(B \cap A_i)\right)$ for any finite n. Letting $n \to \infty$ shows that $\mu_1(B) = \mu_2(B)$ concluding the proof. □

Remark 1.2.8 (Probability Spaces and π-systems)**.** *Condition (1) for Theorem 1.2.2 tells us that we can extend π-systems to σ-algebras. In the context of probability, a π-system tells us that if we have two events, then we can also consider the joint event. For example, when rolling two fair dice, D_1 and D_2, we can note that* $\mathrm{P}\left(D_1 + D_2 = 8\right) = 5/36$ *and that* $\mathrm{P}\left(D_1 = 0 \mod 2\right) = 1/2$*. Thus, we can consider the probability of the intersection of both events* $\mathrm{P}\left(\{D_1 + D_2 = 8\} \cap \{D_1 = 0 \mod 2\}\right) = 1/12$*.*

Remark 1.2.9 (σ-finiteness)**.** *Without σ-finiteness in Condition (2) for Theorem 1.2.2, uniqueness can fail. One example of this is to take $\Omega = (0, 1]$, $\mathcal{A}$ to be all finite unions of half open intervals $(a, b]$, and μ the set function that assigns 0 to the emptyset and ∞ to any non-empty element of $\mathcal{A}$. In this case, μ^* simply assigns ∞ to any subset of Ω that is non-empty. However, the counting measure also assigns 0 to $\emptyset$ and ∞ to any element of $\mathcal{A}$. However, it assigns finite values to finte sets like $\{0.25, 0.5, 0.75\}$ and hence does not coincide with the outer measure μ^*.*

[4] **Exercise:** Check the conditions to verify this.

1.2.3 Completeness

In this section, we want to *complete* a measure. That is, if some arbitrary set E only differs from a measureable set A on a set of measure zero, then we wish to assign the same measure to both sets. However, our σ-field may not contain all sets that *should* be measure zero. This is made more precise below.[5]

Definition 1.2.10 (Symmetric Difference)**.** *For two sets A, B, the symmetric difference is $A\Delta B = (A \setminus B) \cap (B \setminus A)$.*

For a measure space $(X, \mathcal{F}, \mu)$ with $\mathcal{F} \subset \mathcal{P}(X)$, we can define the outer measure μ^* on $\mathcal{P}(X)$ as above:

$$\mu^*(B) = \inf\{\mu(A) : B \subset A\} \text{ for any } B \in \mathcal{P}(X).$$

Then, we can define the set of μ-null sets to be $\mathcal{N}_\mu \subset \mathcal{P}(X)$ where $\mathcal{N} = \{N \subset X : \mu^*(N) = 0\}$. A measure space $(X, \mathcal{F}, \mu)$ is *complete* if $\mathcal{N} \subset \mathcal{F}$. Also, $\mathcal{N}_\mu$ is a ring.[6]

The completion of σ-field $\mathcal{F}$ with respect to μ is $\mathcal{F} \vee \mathcal{N}_\mu = \{A \cup N : A \in \mathcal{F}, N \in \mathcal{N}\}$. In [Dudley(2002)], Proposition 3.3.2, it is proven that this completion is equal to $\{B \subseteq X : \exists A \in \mathcal{F} \text{ s.t. } A\Delta B \in \mathcal{N}_\mu\}$ and that this set is the smallest σ-field that contains both $\mathcal{F}$ and $\mathcal{N}_\mu$. Thus, we can define the completed measure space to be $(X, \mathcal{F} \vee \mathcal{N}_\mu, \bar{\mu})$ where $\bar{\mu}(A \cup N) = \mu(A)$ for $A \in \mathcal{F}$ and $N \in \mathcal{N}_\mu$.

1.3 Lebesgue Measure

Theorems 1.2.1 and 1.2.2 allow us to construct Lebesgue measure, which is a central tool of measure theory. As noted before, a standard way to construct Lebesgue measure on $(0, 1]$ or on $\mathbb{R}$ is to begin with the set of finite unions of half-open intervals $(a, b]$.[7] Both cases are of interest as the Lebesgue measure on the unit interval coresponds to the uniform distribution and Lebesgue on $\mathbb{R}$ is necessary for defining probability density functions.

The standard notation for Lebesgue measure is λ. Often, this is used for both the premeasure assigning lengths to unions of intervals from $\mathcal{A}$ and the outer measure on the Borel σ-field $\mathcal{B}$. We will also denote $\mathcal{M}_\lambda$ to be the set of all Lebesgue measureable sets. It's worth noting that $\mathcal{B} \subset \mathcal{M}_\lambda$. In fact, $\mathcal{M}_\lambda$ is the completion of $\mathcal{B}$ with $\mathcal{N}_\lambda$.

It is also of interest that $\lambda((a, b]) = b - a$ and that λ is the only such measure with this property. Indeed, by Theorem 1.2.2 and the fact that the set of half open intervals $\mathcal{A} = \{(a, b] : a < b\}$ form a π-system, any σ-finite measure μ such that $\mu((a, b]) = b - a$ must coincide with Lebesgue measure on $\sigma(\mathcal{A})$ being the Borel σ-field $\mathcal{B}$.

[5] See [Dudley(2002)], section 3.3 for more details on completion of measures.

[6] **Exercise:** Try to show this.

[7] In [Dudley(2002)] Section 3.2, he considers just half-open intervals, which form a semiring.

1.3.1 Non-Measureable Sets

A classic example of a subset of the unit interval that is not Lebesgue measureable is the Vitali set[8] For $x, y \in (0,1]$, we define addition modulo 1 so that

$$x + y = \begin{cases} x + y & \text{if } x + y \leq 1 \\ x + y - 1 & \text{if } x + y > 1 \end{cases},$$

which can be thought of as wrapping the unit interval into a circle. The set $\mathcal{L}$ of Lebesgue measureable subsets of $(0,1]$ such that $\lambda(A + x) = \lambda(A)$ is a λ-system[9] where $A + x = \{y \in (0,1] : y - x \in A\}$. As $\mathcal{A} \subset \mathcal{L}$, then $\sigma(\mathcal{A}) = \mathcal{B} \subset \mathcal{L}$ due to Dykin's π-λ Theorem. Thus, Lebesgue measure is translation invariant for any Borel set.

Next, we say that $x \sim y$ if $x - y \in \mathbb{Q}$. Hence, we can decompose $(0,1]$ into disjoint equivalence classes. Let the set $H \subset (0,1]$ contain one point from each of these equivalence classes.[10] Let $r_1, r_2 \in \mathbb{Q}$. Since no two points in H are equivalent, $H + r_1 = H + r_2$ is only true if $r_1 = r_2$. Thus, we can write $(0,1] = \bigcup_{r \in \mathbb{Q}} (H + r)$, which is a countable disjoint union.

Finally, by countable additivity, $1 = \lambda((0,1]) = \sum_{r \in \mathbb{Q}} \lambda(H + r)$. However, this leads to a contradiction as if $\lambda(H) = 0$, then the above equation becomes $0 = 1$. Otherwise, if $\lambda(H) > 0$, then the above becomes $1 = \infty$. Hence, the set H lies in $\mathcal{P}((0,1])$ but is not Lebesgue measureable.

Remark 1.3.1 (Fun Fact!)**.** *Lebesgue measure on $\mathbb{R}$ is characterized by being translation invariant. This can be extended into $\mathbb{R}^n$. However, there is no analogue of Lebesgue measure in infinite dimensions. Indeed, it can be proven that the only locally finite and translation-invariant Borel measure μ on Ω is the trivial measure, with $\mu(A) = 0$ for every measurable set A. See https://en.wikipedia.org/wiki/Infinite-dimensional_Lebesgue_measure.*

1.4 Product Measure, Briefly

Now that we have Lebesgue measure λ on $\mathbb{R}$, it is natural to extend it to $\mathbb{R}^p$. The easiest way to attempt this is to consider *rectangles*. That is, For half-open intervals $(a, b]$ and $(c, d]$ on $\mathbb{R}$, we can consider the rectangle $A = (a, b] \times (c, d]$ whose measure (i.e. area) is simply $\lambda^{(2)}(A) = \lambda((a, b])\lambda((c, d]) = (b - a)(d - c)$. This can be extended to higher dimensional Euclidean space by defining

$$\lambda^{(k)}\left((a_1, b_1] \times \ldots \times (a_p, b_p]\right) = \prod_{i=1}^{p} \lambda((a_i, b_i]) = \prod_{i=1}^{p} (b_i - a_i).$$

[8] https://en.wikipedia.org/wiki/Vitali_set

[9] **Exercise:** Check this claim.

[10] Constructing H relies on the Axiom of Choice. That is, if we have a decomposition of $\{A_\theta : \theta \in \Theta\}$. of some set Ω, then there exists a set C that contains one point from each A_θ. The AoC is typically assumed true in the standard approach to measure theory.

As the set of rectangles in $\mathbb{R}^p$ form a π-system, we can argue as before to construct Lebsegue measure on $\mathbb{R}^p$. Similar to λ on $\mathbb{R}$, p-dimensional Lebesgue measure is the only translation invariant measure on $\mathbb{R}^p$—i.e. $\lambda^{(k)}(A+x) = \lambda^{(k)}(A)$ for some $x \in \mathbb{R}^p$ where $A + x = \{y \in \mathbb{R}^p : y = a + x \text{ for some } a \in A\}$.

Ultimately, we will want to show more generally that for two measure spaces $(\mathbb{X}, \mathcal{X}, \mu)$ and $(\mathbb{Y}, \mathcal{Y}, \nu)$ that we can rigorously define the product space $(\mathbb{X} \times \mathbb{Y}, \mathcal{X} \times \mathcal{Y}, \pi)$ where the measure π is uniquely defined by $\pi(A \times B) = \mu(A)\nu(B)$ for $A \in \mathcal{X}$ and $B \in \mathcal{Y}$. This will be explored in more detail in the next chapter.

It is proven in [Dudley(2002)] Proposition 4.1.7. that the product of two Borel σ-fields $\mathcal{B}(\mathbb{X}) \times \mathcal{B}(\mathbb{Y})$ is contained within the Borel σ-field on the product space $\mathcal{B}(\mathbb{X} \times \mathbb{Y})$. Furthermore, in most *nice* settings[11] like when $\mathbb{X} = \mathbb{Y} = \mathbb{R}$, there two σ-fields coincide.

1.5 Independence

For random variables and statistics problems, we have an intuitive understanding of the concept of independence. In some sense, probabilistic measure theory is classical measure theory with the concept of independence inserted into the σ-fields. This will be revisited once we consider random variables, but for now, we work with the probability space $(\Omega, \mathcal{F}, \mu)$.

Definition 1.5.1 (Independence for Sets)**.** *For a countable collection of sets A_i, $i \in I$, we say that the collection is independent if for all finite subsets $J \subset I$, we have*

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

This coincides with the idea that sets A_i are *events* that may occur with some probability. For example, consider drawing a single card from a standard deck of 52 cards. Then, let $A_1 = \{\text{card is red}\}$, $A_2 = \{\text{card is a heart or club}\}$, $A_3 = \{\text{card is a Queen}\}$. This gives the following probablities:

$$\begin{aligned} \mu(A_1) &= 1/2 & \mu(A_1 \cap A_2) &= 1/4 \\ \mu(A_2) &= 1/2 & \mu(A_1 \cap A_3) &= 1/26 \\ \mu(A_3) &= 1/13 & \mu(A_2 \cap A_3) &= 1/26 \\ & & \mu(A_1 \cap A_2 \cap A_3) &= 1/52 \end{aligned}$$

Definition 1.5.2 (Independence for σ-fields)**.** *For a countable collection of σ-fields $\mathcal{F}_i \subset \mathcal{F}$, $i \in I$, we say that this collection of σ-fields is independent if any set of sets $\{A_i \in \mathcal{F}_i : i \in I\}$ is independent in the sense of the previous definition.*

We can use the notion of a π-system to construct such independent σ-fields

[11] that is, when $(\mathbb{X}, \mathcal{X})$ and $(\mathbb{Y}, \mathcal{Y})$ are *second countable*. This includes all seperable metric spaces. See https://en.wikipedia.org/wiki/Second-countable_space

Theorem 1.5.1. *Let $\mathcal{A}_1, \mathcal{A}_2 \subset \mathcal{F}$ be π-systems. If $\mu(A_1 \cap A_2) = \mu(A_1)\mu(A_2)$ for any $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$, then $\sigma(\mathcal{A}_1)$ and $\sigma(\mathcal{A}_2)$ are independent.*

Proof. For a fixed $A_1 \in \mathcal{A}_1$, we can define two measures for $B \in \mathcal{F}$ as

$$\nu_1(B) = \mu(A_1 \cap B) \text{ and } \nu_2(B) = \mu(A_1)\mu(B).$$

By assumption, $\nu_1(A_2) = \nu_2(A_2)$ for any $A_2 \in \mathcal{A}_2$. Hence, by Theorem 1.2.2, they must coincide on $\sigma(\mathcal{A}_2)$. Therefore, $\mu(A_1 \cap B_2) = \mu(A_1)\mu(B_2)$ for a fixed A_1 and any $B_2 \in \sigma(\mathcal{A}_2)$.

This argument can be repeated by fixing an element $B_2 \in \sigma(\mathcal{A}_2)$ to get that $\mu(B_1 \cap B_2) = \mu(B_1)\mu(B_2)$ for $B_i \in \sigma(\mathcal{A}_i)$. □

Chapter 2

Functions, Random Variables, and Integration

Introduction

2.1 Simple Functions and Random Variables

We briefly introduce the concept of a simple random variable to set the stage for more general measurable functions. In these notes, we will use P to denote our probability measure rather than μ for more general measures (albeit usually σ-finite).

For a probability space $(\Omega, \mathcal{F}, P)$, we can define a simple random variable $X : \Omega \to \mathbb{R}$ as a real valued function that only takes on a finite number of values $x_1, \ldots, x_p$ and such that the set

$$\{\omega \in \Omega \,:\, X(\omega) = x_i\} \in \mathcal{F}.$$

One way to write such a function is to finitely partition Ω into disjoint sets $\{A_i\}_{i=1}^p$—i.e. $\bigcup_{i=1}^p A_i = \Omega$ and $A_i \cap A_j = \emptyset$—and write

$$X(\omega) = \textstyle\sum_{i=1}^p x_i \mathbf{1}[\omega \in A_i]\,.$$

Then the *probability that* $X = x_i$ can be equivalently written as

$$\mathrm{P}\left(X = x_i\right) = \mathrm{P}\left(\{\omega \,:\, X(\omega) = x_i\}\right) = \mathrm{P}\left(A_i\right).$$

Furthermore, this allows us to define the *expectation* of the simple random variable X to be

$$\mathrm{E}X = \textstyle\sum_{i=1}^p x_i \mathrm{P}\left(X = x_i\right).$$

Example 2.1.1 (Binary Steps)**.** *Let* $\Omega = (0, 1]$ *and* $A_1 = (0, 0.25], A_2 = (0.25, 0.5], A_3 = (0.5, 0.75], A_4 = (0.75, 1]$ *and* $x_i = (i-1)/4$*. Then, for the probability space* $((0, 1], \mathcal{B}, \lambda)$*,* $\lambda(A_i) = 0.25$ *for any* $i = 1, 2, 3, 4$*. Thus, the simple random variable* $X^{(4)}$ *as above takes on the values of* $0, 0.25, 0.5, 0.75$ *each with probability 25%. The expectation is* $\mathrm{E}X^{(4)} = (0.25 + 0.5 + 0.75)/4 = 0.375$*.*

If we take the number of partitions from 4 to infinity, this simple random variable $X^{(2^m)}$ converges to the Uniform distribution. The method of convergence will be discussed in a later section.

These same ideas can be modified to get a simple function from $(\Omega, \mathcal{F}, \mu)$ to $\mathbb{R}$ defined as $f(\omega) = \sum_{i=1}^{p} x_i \mathbf{1}[\omega \in B_i]$ for $B_i \in \mathcal{F}$. Then, we write the integral of f to be

$$\int f d\mu := \sum_{i=1}^{p} x_i \mu(B_i).$$

The sets B_i need not be disjoint, but given a simple function, we can define it in terms of disjoint B_i.

Exercise: Let $f, g : \Omega \to \mathbb{R}$ be simple functions. Check that $f + g$, fg, $\max\{f, g\}$, and $\min\{f, g\}$ are all simple functions.

Exercise: Check that the integral defined above is linear for non-negative functions—i.e. for simple non-negative functions $f, g : \Omega \to \mathbb{R}^+$ and scalar $c > 0$, show that

$$\int (f + g) d\mu = \int f d\mu + \int g d\mu \quad \text{and} \quad \int cf d\mu = c \int f d\mu.$$

2.2 Measurable Functions and Random Variables

To extend the above idea of a simple random variable, we want to replace the finite x_i with any Borel set $B \subset \mathbb{R}$. However, we can also consider general functions mapping from one measure space to another.

We begin with two *measurable* spaces[1] $(\mathbb{X}, \mathcal{X})$ and $(\mathbb{Y}, \mathcal{Y})$. Let f be a function that maps from $\mathbb{X}$ to $\mathbb{Y}$, then we can consider f applied to sets. For $A \subset \mathbb{X}$ and $B \subset \mathbb{Y}$,

$$f(A) = \{y \in \mathbb{Y} : y = f(x) \text{ for some } x \in A\}$$
$$f^{-1}(B) = \{x \in \mathbb{X} : y = f(x) \text{ for some } y \in B\}.$$

This allows us to define what it means to be a measurable function.

Definition 2.2.1 (Measurable Function)**.** *A function $f : \mathbb{X} \to \mathbb{Y}$ is said to be measurable (with repect to $\mathcal{X}/\mathcal{Y}$, that is) if $f^{-1}(B) \in \mathcal{X}$ for any $B \in \mathcal{Y}$.*

Typically, the σ-fields of interest are the Borel σ-fields and it is sometimes writen $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ when we have a topological space.[2] Moreover, the space $(\mathbb{Y}, \mathcal{Y})$ is typically taken to be $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ or $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$. In this case, we say that f is *Borel measurable*. If we replace $\mathcal{B}(\mathbb{R})$ with $\mathcal{M}_\lambda(\mathbb{R})$, the set of Lebesgue measurable subsets of $\mathbb{R}$, then we say f is Lebesgue measurable. The set of Lebesgue measurable functions gives us a much larger collection of function to define integrals.

Measurablity is a property of functions that is preserved under a variety of operations and transformations. Here are some useful facts that can be verified:

[1] Note that measurable spaces do not have a measure specified otherwise they would be measure spaces. See https://en.wikipedia.org/wiki/Measurable_space

[2] That is, when we can define the set of open sets to *sigma-fy* into the Borel sets.

1. Inverse images of set functions preserve set operations. That is, for $f : \mathbb{X} \to \mathbb{Y}$ and $A, A_i \subset \mathbb{Y}$,

$$f^{-1}(\bigcup_i A_i) = \bigcup_i f^{-1}(A_i) \text{ and } f^{-1}(\mathbb{Y} \setminus A) = \mathbb{X} \setminus f^{-1}(A).$$

   For a measurable set function f, this implies that $\{f^{-1}(B) : B \in \mathcal{Y}\}$ is a σ-field and is contained in $\mathcal{X}$. Hence, we want $\mathcal{Y}$ to be no larger than $\mathcal{X}$ to have measurable functions. Furthermore, this can be used to show that the measurability of f can be established by looking only at a collection of sets $\mathcal{A} \subset \mathcal{Y}$ that generate $\mathcal{Y}$. For example, letting $\mathcal{A}$ be the set of all half-lines $A_t = (-\infty, t]$ for $t \in \mathbb{R}$ will generate $\mathcal{B}(\mathbb{R})$. Thus, f is measurable as long as the sets $\{x : f(x) \leq t\}$ are measurable.

2. For any $A \in \mathcal{X}$, the indicator functions $f(x) = \mathbf{1}[x \in A]$ are measurable. The σ-field generated by f^{-1} is simply $\{\emptyset, A, A^c, \mathbb{X}\} \subset \mathcal{X}$.

3. For measurable functions $f, g : \mathbb{X} \to \mathbb{R}$, the functions $f + g$ and fg are measurable. This follows from [Dudley(2002)] Proposition 4.1.7. as mentioned in the earlier discussion on product measures.

4. For measurable functions, $\{f_i\}_{i=1}^{\infty}$ from $\mathbb{X}$ to $\mathbb{R}$, the following are also measurable: $\sup_i f_i$, $\inf_i f_i$, $\limsup_i f_i$, $\liminf_i f_i$, and also $\lim_i f_i$ if it exists for all x.

   Proof Sketch. In set notation, $\{x : \sup_i f_i(x) \leq t\} = \bigcap_i \{x : f_i(x) \leq t\}$ where the righthand side is a countable intersection of measurable sets and hence measurable. Similarly, $\{x : \inf_i f_i(x) \leq t\} = \bigcup_i \{x : f_i(x) \leq t\}$ and $\limsup_i f_i = \inf_i \sup_{j \geq i} f_i$ and $\liminf_i f_i = \sup_i \inf_{j \geq i} f_i$. If the limit exists then it coincides with the limsup and liminf. □

5. Let $f : \mathbb{X} \to \mathbb{R}$ be a continuous function, then it is measurable.

   Proof. If U is an open set in $\mathbb{R}$, then $f^{-1}(U)$ is open in $\mathbb{X}$.[3] Thus, the set $f^{-1}(U)$ is measurable. Since the open sets of $\mathbb{R}$ generate $\mathcal{B}$, the function f is measurable. □

6. Given a collection of functions $f_i : \mathbb{X} \to \mathbb{Y}$, we can *make* them measurable by constructing the measurable space $(\mathbb{X}, \mathcal{X})$ where $\sigma(\{f_i\}_{i \in I}) \subseteq \mathcal{X}$ where $\sigma(\{f_i\}_{i \in I})$ is the σ-field generated by the sets $f_i^{-1}(B)$ for all i and $B \in \mathcal{Y}$.

All of the above is valid for measurable random variables, which are merely measurable functions from Ω to $\mathbb{R}$ or otherwise.

Definition 2.2.2 (Almost Everywhere / Almost Surely)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. For two functions $f, g : \Omega \to \mathbb{R}$, we say that $f = g$ a.e. (almost everywhere) when the set $N = \{\omega : f(\omega) \neq g(\omega)\}$ has measure $\mu(N) = 0$. In probability theory, "almost everywhere" is replaced with "almost surely" abbreviated a.s. and it is equivalently is written "with probability 1" or wp1.*

[3] This is the definition of a continous function between two topological spaces.

Example 2.2.3 (Equal Almost Everywhere)**.** *Let* $([0,1], \mathcal{B}, \lambda)$ *be the standard measure space of Borel sets on the unit interval with Lebesgue measure. A classical example of being equal almost everywhere is* $f(t) = 0$ *for all* $t \in [0,1]$ *and* $g(t) = 0$ *on* $[0,1] \setminus \mathbb{Q}$ *and* $g(t) = 1$ *on* $[0,1] \cap \mathbb{Q}$ *where* $\mathbb{Q}$ *is the set of rational numbers.*

This is because $\lambda(\mathbb{Q}) = 0$*. This fact can be proven*[4] *by beginning with* $\{q_m\}_{m=1}^{\infty}$ *the enumerated set of rational numbers. Try surrounding each* q_m *with an interval* $(q_m - \varepsilon 2^{-m-1}, q_m + \varepsilon 2^{-m-1})$*.*

2.3 Integration

We have already seen that simple functions can be integrated. Now, we want to extend this idea to any measurable function. The following theorem allows us to work theoretically with simple functions on a π-sysmtem and then pass to all measurable functions in the limit. In what follows we will consider measurable functions mapping from $(\Omega, \mathcal{F})$ to $[-\infty, \infty]$, which is called the extended real line. This allows us to handle sets such as $f^{-1}(\infty)$, for example.

We also require some new notation. For a sequence of functions f_i that are increasing and converge to f for every ω, we write $f_i \uparrow f$. This implies that $f_i(\omega) \to f(\omega)$ and that $f_i(\omega) \leq f_{i+1}(\omega)$ for all ω.

Theorem 2.3.1. *Let* $(\Omega, \mathcal{F})$ *be a measurable space and* $\mathcal{A}$ *a* π*-system that generates* $\mathcal{F}$*. Let* $\mathcal{V}$ *a the linear space of functions such that* $\mathcal{V}$ *contains*

1. *all indicators* $\mathbf{1}_\Omega$ *and* $\mathbf{1}_A$ *for each* $A \in \mathcal{A}$
2. *all functions* f *such that there exists a sequence* $f_i \in \mathcal{V}$ *such that* $f_i \uparrow f$*.*

Then, $\mathcal{V}$ *contains all measurable functions.*[5]

Proof. First, $\mathcal{V}$ contains $\mathbf{1}_A$ for each $A \in \mathcal{A}$. Letting $\mathcal{L} = \{B \in \mathcal{F} : \mathbf{1}_B \in \mathcal{V}\}$, we can show that $\mathcal{L}$ is a λ-system and hence $\mathcal{L} = \mathcal{F}$, so every indicator function $\mathbf{1}_B$ is in $\mathcal{V}$.

For any measurable non-negative f, we can write $f_i = 2^{-i}\lfloor 2^i f \rfloor$ for $i \in \mathbb{N}$. Each f_i in a finite linear combination of indicator functions and hence $f_i \in \mathcal{V}$. Furthermore, $f_i \uparrow f$ and hence $f \in \mathcal{V}$.

For a general measurable function f, we can write it as $f = f^+ - f^-$ where f^+, f^- are non-negative measurable functions. □

Definition 2.3.1 (Integral of a Measurable Function)**.** *For a measurable non-negative function* f *on the measure space* $(\Omega, \mathcal{F}, \mu)$ *and mapping into* $[-\infty, \infty]$*, we define the integral to be*

$$\int f d\mu = \sup \left[\sum_i \left\{ \inf_{\omega \in A_i} f(\omega) \right\} \mu(A_i) \right]$$

[4] **Exercise:** Try it yourself!

[5] In [Billingsley(2008)] Theorem 13.5, decreasing sequences $f_i \downarrow f$ are used to handle the non-positive functions. Very often, measure theory is developed with everything being non-negative to avoid such annoyances.

where the supremum is taken over all finite partitions of Ω into sets A_i.

Inside the square brackets is the integral of the simple function that assigns a value of $\inf_{\omega \in A_i} f(\omega)$ for the set A_i. Hence, for a non-negative f, we consider all simple functions g such that $0 \leq g \leq f$. To extend this to all measurable functions, we write

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

where $f = f^+ - f^-$ and f^+, f^- are non-negative measurable functions.

Remark 2.3.2 (Dealing with infinities)**.** *In the above definition, we may need the following conventions:*

$$0 \times \infty = 0, \quad c \times \infty = \infty$$

*for $c > 0$. Thus, if $f = 0$ on a set A where $\mu(A) = \infty$, then that term in the above definition is just 0 as desired. Also, $\infty - \infty$ is undefined. Hence, $\int f d\mu$ is undefined if $\int f^+ d\mu$ and $\int f^- d\mu$ are both ∞. In the case that both $\int f^+ d\mu$ and $\int f^- d\mu$ are finite, we say that f is **integrable**.*

Theorem 2.3.2. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and let $f \geq 0$ be measureable, and let $f_n \geq 0$ for $n \in \mathbb{N}$ be a sequence of measureable simple functions such that $f_n \uparrow f$. Then, $\int f_n d\mu \uparrow \int f d\mu$.*

Proof. Let g be any simple measureable function such that $0 \leq g \leq f$. Since $f_n \uparrow f$ and f is integrable, $\int f_n d\mu \uparrow c \in [0, \infty]$. We aim to show that $\int g \, d\mu \leq c$.

Indeed, we can write $g = \sum_{i \in \mathcal{I}} a_i \mathbf{1}_{A_i}$ where the A_i are a disjoint partition of Ω. And similarly, we can write $f_n = \sum_{j \in \mathcal{J}} b_j \mathbf{1}_{B_j}$. Consequently,

$$f_n = \sum_{i \in \mathcal{I}} f_n \mathbf{1}_{A_i} = \sum_{i,j} b_j \mathbf{1}_{A_i \cap B_j},$$

and $\int f_n d\mu = \sum_{i \in \mathcal{I}} \int_{A_i} f_n d\mu$. Hence, we want to show that for each $i \in \mathcal{I}$ that

$$\lim_{n \to \infty} \int_{A_i} f_n d\mu \geq a_i \mu(A_i) \tag{2.3.1}$$

to conclude the proof. If $a_i = 0$, then inequality 2.3.1 must hold. If $a_i > 0$, we can divide by a_i. Hence, without loss of generality, we take $a_i = 1$ and take $g = \mathbf{1}_A$ for some set A. For any $\varepsilon > 0$, let $C_n = \{x \in A : f_n(x) > 1 - \varepsilon\}$. Then, $C_n \uparrow A$ in the sense that $C_1 \subseteq \ldots \subseteq C_n \subseteq C_{n+1} \subseteq \ldots \subseteq A$. By countable additivity, $\mu(C_{n+1}) = \mu(C_1) + \sum_{m=1}^{n} \mu(C_{m+1} \setminus C_m)$ and $\mu(C_n) \uparrow \mu(A)$. Since $\int f_n d\mu \geq (1 - \varepsilon)\mu(C_n)$, we have that $c \geq (1 - \varepsilon)\mu(A) = (1 - \varepsilon) \int g \, d\mu$. Taking ε to zero gives $c \geq \int g \, d\mu$. As this holds for any simple function g, $c \geq \int f \, d\mu$. But since $\int f_n d\mu \leq \int f d\mu$, we have $\int f d\mu = c$, which completes the proof. □

Following from the previous discussion on almost eveywhere equality, we can prove a similar a.e. equality for integrals. This is important, because it means that most theorems regarding integrals only require conditions to hold almost everywhere. This will be seen in the three theorems in the next subsection.

Theorem 2.3.3. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and let $f, g : \Omega \to [-\infty, \infty]$ be measureable and $f = g$ a.e. Then, f is integrable if and only if g is integrable. If f and g are integrable, then $\int f d\mu = \int g d\mu$.*

Proof. Let $f = g$ on $\Omega \setminus N$ where $\mu(N) = 0$. For any measureable function $h : \Omega \to [-\infty, \infty]$, we can consider when $\int_\Omega h d\mu$ and $\int_{\Omega \setminus N} h d\mu$ coincide. This is true when h is an indicator function, say, $h = \mathbf{1}_A$ as $\mu(A) = \mu(A \setminus N)$. Thus, the integrals are also equal when h is a simple function. As a consequence of the previous Theorem, we can take non-negative simple functions h to any non-negative measureable function. Lastly, using the above definition that $\int h d\mu = \int h^+ d\mu - \int h^- d\mu$ shows that the integrals will coincide for any integrable measureable function.

To complete the proof, we note that

$$\int f d\mu = \int_{\Omega \setminus N} f d\mu = \int_{\Omega \setminus N} g d\mu = \int g d\mu.$$

□

This result basically tells us that we can modify functions on a set of measure zero without breaking anything.

2.3.1 Three Important Convergence Theorems

When I first learned these results in a graduate measure theory class taught be Prof Paul Koosis at McGill University, I recall him saying over and over again that these are the most important results to learn. Basically every subsequent proof used these to some extent.

In what follows, we are interested in how to handle a sequence of integrals $\int f_i d\mu$ of measureable functions as $i \to \infty$. Under what conditions does it converge to some $\int f d\mu$? From a probability perspective, $\int X_i d\mu$ is the expectation of some random variable X_i, so you can think of the following as theorems about convergence of the mean of a sequence of random variables.

Theorem 2.3.4 (Monotone Convergence)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $\{f_i\}_{i=1}^\infty$ be measurable functions from Ω to $\mathbb{R}$ such that $f_i \uparrow f$ a.e. and $\int f_1 d\mu > -\infty$. Then, $\int f_i d\mu \uparrow \int f d\mu$.*

Proof. First, we need to check that f is measureable.[6] For $c \in \mathbb{R}$, we consider the sets $(c, \infty]$, which generate the Borel σ-field. Since $f_i \uparrow f$, $f^{-1}((c, \infty]) = \bigcup_{i=1}^\infty f_i^{-1}((c, \infty])$ and $f_i^{-1}((c, \infty]) \in \mathcal{F}$, we have that f is measureable.

Now, we assume that $f_1 \geq 0$, and for each f_i we take simple functions g_{ij} such that $g_{ij} \uparrow f_i$. Thus, by Theorem 2.3.2, $\int g_{ij} d\mu \uparrow \int f_i d\mu$. Furthermore, let $g_i^\star = \max\{g_{1i}, \ldots, g_{ii}\}$.[7] These $g_i^\star$ are simple functions and $g_i^\star \uparrow f$. Once again, Theorem 2.3.2

[6] Note that measureabiliy of f is not assumed in the theorem but implied via the convergence condition.

[7] Note that we set $j = i$ in this expression and take the max over the first i functions.

implies that $\int g_i^\star d\mu \uparrow \int f d\mu$. But since $g_i^\star \leq f_i$ by construction, $\int g_i^\star d\mu \leq \int f_i d\mu \leq \int f d\mu$. Thus, $\int f_i d\mu \uparrow \int f d\mu$.

Now, we assume that $f \leq 0$. In this case $f_i \uparrow f$ implies that $-f_i \downarrow -f$. Writing $h = -f$ and $h_i = -f_i$, we have that $0 \leq \int h d\mu \leq \int h_i d\mu$. Next, note that $0 \leq h_1 - h_i \uparrow h_1 - h$. Applying the above result gives that $\int (h_1 - h_i) d\mu \uparrow \int (h_1 - h) d\mu$. Since all of the h have finite integrals, we are allowed to subtract to get that $\int h_i d\mu \downarrow \int h d\mu$ and thus $\int f_i d\mu \uparrow \int f d\mu$.

For a general function $f = f^+ - f^-$, we have that $f_i^+ \uparrow f^+$ and $f_i^- \downarrow f^-$ and $\int f^- d\mu < \infty$. So by the above special cases, $\int f_i^+ d\mu \uparrow \int f^+ d\mu$ and $\int f_i^- d\mu \downarrow \int f^- d\mu$ and finally $\int f_i d\mu \uparrow \int f d\mu$. □

Remark 2.3.3. *From Theorem 2.3.3, we only require $f_i \uparrow f$ to hold almost everywhere to establish the result. Hence convergence can fail on a set of measure (probability) zero and we still have convergence of the integrals.*

Secondly, we can redo the above proof for $f_i \downarrow f$ with $\int f_1 d\mu < \infty$ to get a similar result for decreasing sequences.

Theorem 2.3.5 (Fatou's Lemma)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $\{f_i\}_{i=1}^\infty$ be non-negative measurable functions from Ω to $\mathbb{R}$. Then, $\int \liminf f_i d\mu \leq \liminf \int f_i d\mu$.*

Proof. Recall that $\liminf_{i\to\infty} f_i = \sup_j \inf_{i \geq j} f_i$. Hence, let $g_j = \inf\{f_i : i \geq j\}$. Then, $g_j \uparrow \liminf_{i\to\infty} f_i$ and $f_1 \geq 0$ by assumption, so Theorem 2.3.4 says that $\int g_j d\mu \uparrow \int \liminf_{i\to\infty} f_i d\mu$. By construction, $g_j \leq f_i$ for any $i \geq j$, and thus, $\int g_j d\mu \leq \int f_i d\mu$ for any $i \geq j$, and subsequently, $\int g_j d\mu \leq \inf_{i \geq j} \int f_i d\mu$. Taking $j \to \infty$, gives $\lim_{j\to\infty} \int g_j d\mu = \int \liminf f_i d\mu \leq \liminf \int f_i d\mu$. □

Theorem 2.3.6 (Dominated Convergence)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $\{f_i\}_{i=1}^\infty$ and g be absolutely integrable. If $|f_i| \leq g$ for all i and $f_i(\omega) \to f(\omega)$ for each $\omega \in \Omega$ (i.e. pointwise convergence), then f is absolutely integrable and $\int f_i d\mu \to \int f d\mu$.*

Proof. Let $f_i^\wedge = \inf\{f_j : j \geq i\}$ and $f_i^\vee = \sup\{f_j : j \geq i\}$. Then, $f_i^\wedge \leq f_i \leq f_i^\vee$. We have that $f_i^\wedge \uparrow f_i$ and that $\int f_1^\wedge d\mu \geq -\int g \, d\mu > -\infty$, so Theorem 2.3.4 implies that $\int f_i^\wedge d\mu \uparrow \int f d\mu$.

Doing the same for $f_i^\vee$, we have that $f_i^\vee \downarrow f$ and hence that $\int f_i^\vee d\mu \downarrow \int f d\mu$. Since $\int f_i^\wedge d\mu \leq \int f_i d\mu \leq \int f_i^\vee d\mu$, we have the desired result that $\int f_i d\mu \to \int f d\mu$. □

2.3.2 Lebesgue-Stieltjes measure

Given two measureable spaces, $(\mathbb{X}, \mathcal{X})$ and $(\mathbb{Y}, \mathcal{Y})$, and a measureable function $\psi : \mathbb{X} \to \mathbb{Y}$, the function ψ can induce an image measure. Let μ be a measure on $\mathcal{X}$, then we can define $\nu = \mu \circ \psi^{-1}$ to be a measure on $\mathcal{Y}$. That is, for a set $B \in \mathcal{Y}$, we define $\nu(B) = \mu(\psi^{-1}(B))$. This allows us to turn Lebesgue measure into Lebesgue-Stieltjes measures. The most obvious application of such is the cumulative distribution function for a probability distribution.

Theorem 2.3.7. *Let $F : \mathbb{R} \to \mathbb{R}$ be non-constant, right-continuous, and non-decreasing. Then, there exists a unique measure dF on $\mathbb{R}$ such that for all $a, b \in \mathbb{R}$ with $a < b$,*

$$dF((a, b]) = F(b) - F(a).$$

Proof. Let $F(\infty) = \lim_{x\to\infty} F(x)$ and $F(-\infty) = \lim_{x\to-\infty} F(x)$. We define an open interval $I = (F(-\infty), F(\infty))$ and define $g(y) = \inf\{x \in \mathbb{R} : y \leq F(x)\}$. We want to define dF to be $\lambda \circ g^{-1}$ where λ is Lebesgue measure on $\mathbb{R}$, so we need to show that this makes sense.

We first show that g is left continuous and non-decreasing and for $y \in I$ and $x \in \mathbb{R}$, $g(y) \leq x$ if and only if $y \leq F(x)$. To show this, fix a $y \in I$ and consider $J_y = \{x \in \mathbb{R} : y \leq F(x)\}$. As F is non-decreasing, if $x \in J_y$ and $x' \geq x$ then $x' \in J_y$. As F is right continuous, if $x_n \in J_y$ and $x_n \downarrow x$, then $x \in J_y$. Therefore, $J_y = [g(y), \infty)$. And furthermore, $g(y) \leq x$ if and only if $y \leq F(x)$. Secondly, for $y \leq y'$, we have that $J_{y'} \subseteq J_y$ and thus $g(y) \leq g(y')$. So for $y_n \uparrow y$, we have that $J_y = \bigcap_n J_{y_n}$ and further that $g(y_n) \to g(y)$, which implies that g is left continuous and non-decreasing.

From the above, we have that g is Borel measureable (see *useful fact 1* in Section 2.2). And thus defining $dF = \lambda \circ g^{-1}$ gives us that

$$dF((a, b]) = \lambda\left(\{y : g(y) > a,\ g(y) \leq b\}\right) = \lambda((F(a), F(b)]) = F(b) - F(a).$$

Furthermore, this measure, dF, is unique by using the same arguments as before for Lebesgue measure. □

In the case that $F : \mathbb{R} \to [0, 1]$ such that the interval $I = [0, 1]$, we have a cumulative distribution function, which induces a measure on the real line. This allows us to do things like integrate with respect to such measures—i.e. take an expectation.

Definition 2.3.4 (Radon Measure)**.** *Let $(\Omega, \mathcal{B}, \mu)$ be a measure space where $\mathcal{B}$ is the Borel σ-field. The measure μ is said to be a Radon measure if $\mu(K) < \infty$ for all compact $K \in \mathcal{B}$.*

Note that 'most' measures you will encounter in practice are Radon measures.

Going one step beyond the above proof, we can note that dF is a Radon measure and, more interestingly, that every non-zero Radon measure on $\mathcal{B}(\mathbb{R})$ can be written as $dF = \lambda \circ g^{-1}$ for some F.

Indeed, if μ is a Radon measure on $\mathbb{R}$, then we can define F as

$$F(x) = \begin{cases} \mu((0, x]) & \text{if } x \geq 0 \\ -\mu((x, 0]) & \text{if } x < 0 \end{cases}.$$

Thus, $F(b) - F(a) = \mu((a, b])$ for $a < b$ and hence $\mu = dF$ by uniqueness.[8]

[8] Question to consider: Why is μ being Radon necessary?

2.4 Product Measure, in detail

Now that we have defined the integral, we can more formally construct product measures. First, given two σ-fields $\mathcal{X}$ and $\mathcal{Y}$, we will denote the *product σ-field* to be $\mathcal{X} \times \mathcal{Y}$, which is the σ-field generated by the rectangles $A \times B$ for $A \in \mathcal{X}$ and $B \in \mathcal{Y}$. Sets of the form $A \times B$ for $A \in \mathcal{X}$ and $B \in \mathcal{Y}$ are called *rectangles*. The collection of all rectangles will be denoted as $\mathcal{R}$. Our goal is to prove the following existence and uniqueness theorem:

Theorem 2.4.1 (Existence and Uniqueness of Product Measure)**.** *Let $(\mathbb{X}, \mathcal{X}, \mu)$ and $(\mathbb{Y}, \mathcal{Y}, \nu)$ be σ-finite measure spaces. Let π be a set function on $\mathcal{X} \times \mathcal{Y}$ such that for $A \in \mathcal{X}$ and $B \in \mathcal{Y}$, $\pi(A \times B) = \mu(A)\nu(B)$. Then, π extends uniquely to a measure on $(\mathbb{X} \times \mathbb{Y}, \mathcal{X} \times \mathcal{Y})$ such that for any $E \in \mathcal{X} \times \mathcal{Y}$,*

$$\pi(E) = \iint \mathbf{1}_E(x, y) d\mu(x) d\nu(y) = \iint \mathbf{1}_E(x, y) d\nu(y) d\mu(x).$$

We will approach this proof by first proving the result for finite measures μ and ν and then extending it to σ-finite measures. It can be shown that the set function π where $\pi(A \times B) = \mu(A)\nu(B)$ is countably additive on $\mathcal{R}$.[9] By including finite unions of rectangles, the collection $\mathcal{R}$ can be extended to a field $\mathcal{A}$.

First, we will prove a similar theorem to Dynkin's π-λ theorem involving montone classes.

Definition 2.4.1 (Monotone Class)**.** *A collection of subsets $\mathcal{M}$ of Ω is said to be monotone if*

1. *for $\{A_i\}_{i=1}^{\infty}$ such that $A_i \in \mathcal{M}$ and $A_i \uparrow A = \bigcup_{i=1}^{\infty} A_i$, then $A \in \mathcal{M}$,*
2. *for $\{A_i\}_{i=1}^{\infty}$ such that $A_i \in \mathcal{M}$ and $A_i \downarrow A = \bigcap_{i=1}^{\infty} A_i$, then $A \in \mathcal{M}$.*

Note that if a field $\mathcal{A}$ is also monotone, then it is a σ-field. Furthermore, recall that we defined a field to be such that $\emptyset, \Omega \in \mathcal{A}$, if $B, A \in \mathcal{A}$ then $B \setminus A \in \mathcal{A}$, and if $B, A \in \mathcal{A}$ then $A \cup B \in \mathcal{A}$. Instead, we can replace $B \setminus A \in \mathcal{A}$ with $A^c \in \mathcal{A}$. This is because $B \setminus A = B \cap A^c$.

Theorem 2.4.2 (Monotone Class Theorem)**.** *Let $\mathcal{A}$ be a field and $\mathcal{M}$ be monotone such that $\mathcal{A} \subset \mathcal{M}$. Then, $\sigma(\mathcal{A}) \subseteq \mathcal{M}$.*

Proof. In this proof, we will show that $\sigma(\mathcal{A}) \subset m(\mathcal{A})$ where $m(\mathcal{A})$ is the smallest monotone class that contains $\mathcal{A}$—i.e. the intersection of all monotone classes than contain $\mathcal{A}$. This is done by showing that $m(\mathcal{A})$ is a field and thus a σ-field and thus contains $\sigma(\mathcal{A})$ as $\sigma(\mathcal{A})$ is minimal.

Step 1 is to show that $m(\mathcal{A})$ is closed under taking complements. Let $\mathcal{F} = \{A \in m(\mathcal{A}) : A^c \in m(\mathcal{A})\}$. This means that $\mathcal{A} \in \mathcal{F}$ since $\mathcal{A}$ is closed under complements. Furthermore, for $\{A_i\}_{i=1}^{\infty}$ such that $A_i \in \mathcal{F}$ and $A_i \uparrow A = \bigcup_{i=1}^{\infty} A_i$, then $A_i^c \in \mathcal{F}$ and

[9] Exercise: Try to prove this fact.

$A_i^c \downarrow A^c = \bigcap_{i=1}^{\infty} A_i^c = (\bigcup_{i=1}^{\infty} A_i)^c$. Therefore, $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$. Thus, $\mathcal{F}$ is monotone and $\mathcal{F} \subseteq m(\mathcal{A})$. Doing the same for $A_i \downarrow A$ shows that $\mathcal{F} = m(\mathcal{A})$ by minimality of $m(\mathcal{A})$, and therefore, $m(\mathcal{A})$ is closed under complementation.

Step 2 is to show that $m(\mathcal{A})$ is closed under finite unions. Let

$$\mathcal{G}_1 = \{A \in m(\mathcal{A}) \,:\, A \cup B \in m(\mathcal{A}) \text{ for all } B \in \mathcal{A}\}.$$

As with $\mathcal{F}$ from the previous paragraph, we note that $\mathcal{A} \in \mathcal{G}_1$. Also, $\mathcal{G}_1$ is monotone, since $(\bigcup_{i=1}^{\infty} A_i) \cup B = \bigcup_{i=1}^{\infty}(A_i \cup B) \in m(\mathcal{A})$ and $(\bigcap_{i=1}^{\infty} A_i) \cup B = \bigcap_{i=1}^{\infty}(A_i \cup B) \in m(\mathcal{A})$. Thus by minimality, $m(\mathcal{A}) = \mathcal{G}_1$ and thus $m(\mathcal{A})$ is closed under finite unions with elements of $\mathcal{A}$. Now, let

$$\mathcal{G}_2 = \{B \in m(\mathcal{A}) \,:\, A \cup B \in m(\mathcal{A}) \text{ for all } A \in m(\mathcal{A})\}.$$

Once again, $\mathcal{A} \subset \mathcal{G}_2$, since if $B \in \mathcal{A}$, then $A \cup B \in m(\mathcal{A})$ for any $A \in m(\mathcal{A})$ from the argument with $\mathcal{G}_1$. $\mathcal{G}_2$ is also monotone. Hence $\mathcal{G}_2 = m(\mathcal{A})$, which means $m(\mathcal{A})$ is closed under finite unions, and hence it is a field, and hence it is a σ-field. Thus, $\sigma(A) \subseteq m(\mathcal{A})$ by minimality. □

We will also require a lemma that allows us to swap the above order of integration for finite measures.

Lemma 2.4.2. *Let $(\mathbb{X}, \mathcal{X}, \mu)$ and $(\mathbb{Y}, \mathcal{Y}, \nu)$ be finite measure spaces, and let*

$$\mathcal{F} = \left\{ E \subset X \times Y \,:\, \iint \mathbf{1}_E(x,y) d\mu(x) d\nu(y) = \iint \mathbf{1}_E(x,y) d\nu(y) d\mu(x) \right\}.$$

Then, $\mathcal{X} \times \mathcal{Y} \subset \mathcal{F}$.

Proof. First, let $E = A \times B$ for $A \in \mathcal{X}$ and $B \in \mathcal{Y}$, i.e. $E \in \mathcal{R}$. Then,

$$\begin{aligned}\iint \mathbf{1}_E(x,y) d\mu(x) d\nu(y) &= \mu(A) \int \mathbf{1}_B(y) d\nu(y) = \mu(A)\nu(B) \\ &= \nu(B)\mu(A) = \nu(B) \int \mathbf{1}_A(x) d\mu(x) = \iint \mathbf{1}_E(x,y) d\nu(y) d\mu(x).\end{aligned}$$

Therefore, $\mathcal{R} \subset \mathcal{F}$. Also, for disjoint $R_1, R_2 \in \mathcal{R}$, $\mathbf{1}_{R_1 \cup R_2} = \mathbf{1}_{R_1} + \mathbf{1}_{R_2}$. Hence, $\mathcal{F}$ contains finite disjoint unions of rectangles. This implies that the field generated by the set of rectangles $\mathcal{A} \subset \mathcal{F}$.[10]

We next consider $\{E_i\}_{i=1}^{\infty}$ with $E_i \in \mathcal{F}$. If $E_i \uparrow E$ then monotone convergence implies that

$$\iint \mathbf{1}_{E_i}(x,y) d\mu(x) d\nu(y) \uparrow \iint \mathbf{1}_E(x,y) d\mu(x) d\nu(y)$$

and

$$\iint \mathbf{1}_{E_i}(x,y) d\nu(y) d\mu(x) \uparrow \iint \mathbf{1}_E(x,y) d\nu(y) d\mu(x)$$

Thus, $E \in \mathcal{F}$, and the same holds if $E_i \downarrow E$. Therefore, $\mathcal{F}$ is a monotone class. Finally, applying the monotone class theorem shows that $\mathcal{X} \times \mathcal{Y} = \sigma(\mathcal{A}) \subset \mathcal{F}$. □

[10] See [Dudley(2002)] Proposition 3.2.3, which states that for a semi-ring such as $\mathcal{R}$, the collection of all finite disjoint unions of elements of $\mathcal{R}$ is a ring. And $\mathbb{X} \times \mathbb{Y} \in \mathcal{R}$.

Proof of Theorem 2.4.1. We first consider the case that μ and ν are finite measures. We begin with $\pi(A \times B) = \mu(A)\nu(B)$ for $A \times B \in \mathcal{R}$. Then, we extend π to the set function

$$\pi(E) := \iint \mathbf{1}_E(x,y) d\mu(x) d\nu(y)$$

for any $E \in \mathcal{X} \times \mathcal{Y}$. The above lemma says that we can define this set function and the order of integration can be reversed for any $E \in \mathcal{X} \times \mathcal{Y}$. Linearity of the integral implies that the set function π is finitely additive, and monotone convergence further implies that π is countably additive. Thus, π is a measure on $\mathcal{X} \times \mathcal{Y}$.

To show that π is unique, let ρ be some other set function such that $\rho(A \times B) = \mu(A)\nu(B)$ for $A \times B \in \mathcal{R}$. Let $\mathcal{M} = \{E \subset \mathbb{X} \times \mathbb{Y} : \pi(E) = \rho(E)\}$. Then, $\mathcal{M}$ is a monotone class, because for $E_i \uparrow E = \bigcup_{i=1}^{\infty} E_i$, we can rewrite $E = \bigcup_{i=1}^{\infty} D_i$ where $D_1 = E_1$ and $D_i = E_i \setminus E_{i-1}$ for $i \geq 2$ are disjoint. Thus, by countable additivity $\pi(E) = \rho(E)$, so $E \in \mathcal{M}$. Arguing similarly for $E_i \downarrow E$ shows that $\mathcal{M}$ is monotone. Hence, application of the monotone class theorem implies that $\mathcal{X} \times \mathcal{Y} \subset \mathcal{M}$. Thus, π is unique on $\mathcal{X} \times \mathcal{Y}$ for finite measures μ and ν.

Now let μ and ν be σ-finite measures. Let $\{A_i\}_{i=1}^{\infty}$ and $\{B_i\}_{i=1}^{\infty}$ be disjoint partitions of $\mathbb{X}$ and $\mathbb{Y}$, respectively, such that $\mu(A_i) < \infty$ and $\nu(B_i) < \infty$. Then, for any $E \in \mathcal{X} \times \mathcal{Y}$, we define $E_{ij} = E \cap (A_i \times B_j)$. Thus, from the above finite measure case,

$$\iint \mathbf{1}_{E_{i,j}}(x,y) d\mu(x) d\nu(y) = \iint \mathbf{1}_{E_{i,j}}(x,y) d\nu(y) d\mu(x).$$

Summing over all i and j, we can extend π using monotone convergence again to get

$$\pi(E) = \iint \mathbf{1}_E(x,y) d\mu(x) d\nu(y) = \iint \mathbf{1}_E(x,y) d\nu(y) d\mu(x)$$

for any $E \in \mathcal{X} \times \mathcal{Y}$. Monotone convergence implies that π is countably additive and hence a measure on $\mathcal{X} \times \mathcal{Y}$. For any other measure ρ such that $\rho(A \times B) = \mu(A)\nu(B)$, countably additivity and uniqueness for finite measures implies that

$$\pi(E) = \sum_{i,j} \pi(E_{i,j}) = \sum_{i,j} \rho(E_{i,j}) = \rho(E)$$

for any $E \in \mathcal{X} \times \mathcal{Y}$. Hence, the extension of π to $\mathcal{X} \times \mathcal{Y}$ is unique. □

2.4.1 The Fubini-Toneli Theorem

The above existence and uniqueness theorem for product measures allows us to swap the order of integration for indicator functions. The following important theorem allows us to similarly swap the order of integration for measureable functions in product spaces.

Theorem 2.4.3 (Fubini-Toneli Theorem)**.** *Let $(\mathbb{X}, \mathcal{X}, \mu)$ and $(\mathbb{Y}, \mathcal{Y}, \nu)$ be σ-finite measure spaces, and let $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ be measurable with respect to $\mathcal{X} \times \mathcal{Y}$ such that either $f \geq 0$ or $\iint |f| d(\mu \times \nu) < \infty$. Then,*

$$\int f \, d(\mu \times \nu) = \iint f(x,y) d\mu(x) d\nu(y) = \iint f(x,y) d\nu(y) d\mu(x).$$

Also, $\int f(x,y) d\mu(x)$ is $\mathcal{Y}$-measurable and $\int f(x,y) d\nu(y)$ is $\mathcal{X}$-measurable.

Proof. This result is immediate if f is a simple function due to the above existence and uniqueness theorem for product measure and the linearity of the integral. That is,

$$\iint \sum_{i=1}^{n} \mathbf{1}_{E_i}(x,y)d\mu(x)d\nu(y) = \iint \sum_{i=1}^{n} \mathbf{1}_{E_i}(x,y)d\nu(y)d\mu(x).$$

Secondly, applying the monotone convergence theorem to simple functions implies that the above holds for non-negative measureable functions.

If instead we assume that $\iint |f| d(\mu \times \nu) < \infty$, then $f = f^+ - f^-$ and the above holds for both f^+ and f^-. That is, both $\int f^+(x,y)d\mu(x) < \infty$ for ν-almost-every y and $\int f^+(x,y)d\nu(y) < \infty$ for μ-almost-every x and similarly for f^-. Therefore,

$$\int |f|(x,y)d\nu(y) = \int f^+(x,y)d\nu(y) + \int f^-(x,y)d\nu(y) < \infty \quad (\mu\, a.e.)$$

and thus

$$\int f(x,y)d\nu(y) = \int f^+(x,y)d\nu(y) - \int f^-(x,y)d\nu(y) < \infty \quad (\mu\, a.e.).$$

As Theorem 2.3.3 tells us that we only require finiteness to occur almost everywhere to have the integral exist, we can integrate both sides of the above with respect to μ to get

$$\iint f(x,y)d\nu(y)d\mu(x) = \iint f^+(x,y)d\nu(y)d\mu(x) - \iint f^-(x,y)d\nu(y)d\mu(x).$$

The same can be done swapping the role of μ and ν to conclude the theorem. □

Remark 2.4.3. *The above theorem lets us swap the order of integration for the product of two measure spaces. This can be extended by induction to the finite product of n measure spaces. For the sake of stochastic processes, we will have to consider the infinite product of probability spaces, which will be discussed later.*

2.4.2 Infinite Product Probabilities

As noted in the previous remark, induction allows us to extend from the product of two measure spaces to the product of n measure spaces. However, in fields like statistics, we often want to take $n \to \infty$. In general, a countably infinite product of finite or σ-finite measure spaces may not retain its finiteness or σ-finiteness. However, we will show in this section that a countable product of probability spaces will still be a probability space.

For notation, let $(\Omega_n, \mathcal{F}_n, P_n)$ for $n \in \mathbb{N}$ be a sequence of probability spaces. Then, $\Omega := \otimes_{n=1}^{\infty}\Omega_n$ consists of elements of the form $\omega = \{\omega_i\}_{n=1}^{\infty}$ where $\omega_n \in \Omega_n$. This is a space of sequences; for example, if $\Omega_n = \mathbb{R}$ for all n, then Ω would be the space of all real-valued sequences. We define the set $\mathcal{R}$ to be the space of all finite dimensional rectangles in Ω. That is, for $R \in \mathcal{R}$, $R \subseteq \Omega$ and we can write $R = \bigotimes_{n=1}^{\infty} A_n$ for $A_n \in \mathcal{F}_n$ but we also require $A_n = \Omega_n$ for all but a finite number of n. As for finite

product spaces, it can be shown that $\mathcal{R}$ is a semi-ring. Furthermore, we will denote the field generated by $\mathcal{R}$ to be $\mathcal{S}$. We define P to be a set function on $\mathcal{R}$ such that $P(R) = \prod_{n=1}^{\infty} P_n(A_n)$. Note that this infinite product must converge as all but a finite number of these $P_n(A_n) = 1$. Lastly, we require a σ-field $\mathcal{F}$ on Ω. This is constructed by considering the projections $\varpi_n : \Omega \to \Omega_n$ where $\varpi_n((\omega_1, \omega_2, \ldots)) = \omega_n$. Then, $\mathcal{F}$ is the smallest σ-field such that all ϖ_n are measurable mappings from $(\Omega, \mathcal{F})$ into $(\Omega_n, \mathcal{F}_n)$. That is, $\mathcal{F}$ contains all sets of the form $\varpi_n^{-1}(A_n)$ for $A_n \in \mathcal{F}_n$. We can write $\mathcal{F} = \sigma(\varpi_1, \varpi_2, \ldots)$.[11] Note that sometimes sets of the form $\varpi_n^{-1}(A_n)$ are called cylinder sets. For the product of two measure spaces, $(\mathbb{X} \times \mathbb{Y}, \mathcal{X} \times \mathcal{Y}, \pi)$ and a set $A \in \mathcal{X} \times \mathcal{Y}$, for any $x \in \mathbb{X}$, we denote $A_x := \{y \in \mathbb{Y} : (x, y) \in A\}$. This can be thought of as slicing set A along x. Furthermore, it can be shown that $A_x \in \mathcal{Y}$ for any $x \in \mathbb{X}$.

Theorem 2.4.4 (Existence and Uniqueness of Infinite Product Probabilities)**.** *The set function P on $\mathcal{R}$ extends uniquely to a probability measure on $\mathcal{F}$.*

We first need a lemma that gives us a condition for a finitely additive set function on a field to be a countably additive set function on a field. That is, we eventually want to show that P is a pre-measure on $\mathcal{S}$.

Lemma 2.4.4. *Let μ be a finitely additive set function on a field $\mathcal{S}$. Then, μ is countably additive if and only if for any sequence $A_i \downarrow \emptyset$ with $A_i \in \mathcal{S}$, then $\mu(A_i) \to 0$.*

Proof Exercise. Try to show this yourself by replacing the A_i with disjoint sets. Also, recall that if a series converges, then the summands have to go to zero. □

Proof of Theorem 2.4.4. First, we want to extend P from $\mathcal{R}$ to the field $\mathcal{S}$ and show that it makes sense. It can be shown[12] that any $S \in \mathcal{S}$ can be written as a disjoint union of elements of $\mathcal{R}$. Hence, we can write

$$S = \bigcup_{i=1}^{k} R_i = \bigcup_{i=1}^{k} \bigotimes_{n=1}^{\infty} A_{i,n}$$

where $A_{i,n} \in \mathcal{F}_n$ and all but of finite number of these $A_{i,n}$ are equal to Ω_n. Thus, there exists an $m \in \mathbb{N}$ such that $A_{i,n} = \Omega_n$ for all i and for all $n > m$. Thus, we can treat $P(S)$ as a finite product measure on $\Omega_1 \times \ldots \times \Omega_k$ and it is finitely additive on $\mathcal{S}$.

Carathéodory's Extension Theorem tells us that if P is countably additive on the field $\mathcal{S}$ then it extend to a measure on $\mathcal{F} = \sigma(\mathcal{S})$. To show that P is countably additive, we apply the above lemma via a contradiction argument. That is, for some decreasing sequence $\{A_i\}_{i=1}^{\infty}$ such that for some $\varepsilon > 0$, $P(A_i) > \varepsilon$ for all i, then $\bigcap_{i=1}^{\infty} A_i \neq \emptyset$.

Let $P^{(0)}$ be the set function P on $\mathcal{S}$. For $n \geq 1$, we define $\Omega^{(n)} := \bigotimes_{m>n} \Omega_m$, and similarly we let $\mathcal{S}^{(n)}$ be the set of disjoint unions of rectangles on $\Omega^{(n)}$ and $P^{(n)}$ to be

[11] In probability theory, we often think of a sequence of random variables $X_1, X_2, \ldots$, and we can consider the smallest σ-field such that these are all measurable. This is denoted as $\sigma(X_1, X_2, \ldots)$.

[12] See [Dudley(2002)] Proposition 8.2.1.

P defined on $\mathcal{S}^{(n)}$. These are all concerned with the *tail* of the sequence of spaces. For any subset $E \subseteq \Omega$ and $(x_1, \ldots, x_n) \in \Omega_1 \times \ldots \times \Omega_n$, we can define

$$E^{(n)}(x_1, \ldots, x_n) = \left\{ \{x_m\}_{m>n} \in \Omega^{(n)} \,:\, \{x_m\}_{m \geq 1} \in E \right\}$$

which is the set of all tail sequences such that the entire sequence lies in E.

From the above discussion, for any $E \in \mathcal{S}$, then there exists an N such that we can write $E = F \times \bigotimes_{n>N} \Omega_n$ for some $F \subseteq \bigotimes_{n=1}^{N} \Omega_n$—i.e. E is a finite product F and a trivial tail sequence. Thus, we can decompose $F = \bigcup_{i=1}^{k} F_i = \bigcup_{i=1}^{k} \bigotimes_{n=1}^{N} F_{i,n}$ where $F_{i,n} \in \mathcal{F}_n$, that is, F is a finite union of N-*dimensional* rectangles. For any choice of $m < N$ and $(x_1, \ldots, x_m)$, $E^{(m)}(x_1, \ldots, x_m) = G \times \Omega^{(N)}$ for $G = \bigcup_{i:x_i \in F_{i,n}} \bigotimes_{n=m}^{N} F_{i,n}$. This implies that $E^{(n)}(x_1, \ldots, x_n) \in \mathcal{S}^{(n)}$. Thus, $P^{(n)}$ is defined on $\mathcal{S}^{(n)}$.

Through application of Fubini-Toneli theorem, we have

$$\begin{aligned} P(E) &= \int \mathbf{1}_E \, dP_1 \times \ldots \times dP_n \times dP^{(n)} \\ &= \int P^{(n)} \left(E^{(n)}(x_1, \ldots, x_n) \right) \, dP_1 \times \ldots \times dP_n. \end{aligned}$$

Returning to the A_i from above, let $\{A_i\}_{i=1}^{\infty}$ be a decrease sequence such that $P(A_i) > \varepsilon > 0$. Then, for each A_i, we further define $F_i := \left\{ x_1 \in \Omega_1 \,:\, P^{(1)}(A_i^{(1)}(x_1)) > \varepsilon/2 \right\}$, which is the set of $x_1 \in \Omega_1$ such that the set of tail sequences has $P^{(1)}$-measure greater than $\varepsilon/2$. Using the above integral formula with $n = 1$, we can set $E = A_i$ to get

$$\begin{aligned} \varepsilon < P(A_i) &= \int P^{(1)}(A_i^{(1)}(x_1)) \, dP_1(x_1) \\ &= \int_{F_i} P^{(1)}(A_i^{(1)}(x_1)) \, dP_1(x_1) + \int_{\Omega_1 \setminus F_i} P^{(1)}(A_i^{(1)}(x_1)) \, dP_1(x_1) \\ &\leq P_1(F_i) + \varepsilon/2, \end{aligned}$$

because on $\Omega_1 \setminus F_i$, $P^{(1)}(A_i^{(1)}(x_1)) < \varepsilon/2$ by construction and, of course, $P^{(1)}(A_i^{(1)}(x_1)) \leq 1$ on F_i. Note that we are relying on P being a probability measure at this step.

The conclusion of the above derivation is that $P_1(F_i) > \varepsilon/2$ for all i, and since the A_i's are decreasing, so are the F_i's. Furthermore, P_1 is a countably additive probability measure on $(\Omega_1, \mathcal{F}_1)$. Hence,

$$P_1\left(\bigcap_{i=1}^{\infty} F_i \right) = \int \mathbf{1}_{\bigcap_{i=1}^{\infty} F_i} dP_1 = \int \inf_i \mathbf{1}_{F_i} dP_1 = \inf_i \int \mathbf{1}_{F_i} dP_1 = \inf_i P_1(F_i) \geq \varepsilon/2$$

where we can swap the infimum and integral using monotone convence since $\mathbf{1}_{\bigcap_{i=1}^{m} F_i} \downarrow \mathbf{1}_{\bigcap_{i=1}^{\infty} F_i}$. Thus, applying the above lemma results in $\bigcap_{i=1}^{\infty} F_i \neq \emptyset$.

Next, we can fix some $y_1 \in \bigcap_{i=1}^{\infty} F_i$, and we can define

$$G_i := \{x_2 \in \Omega_2 \,:\, P^{(2)}(A_i^{(2)}(y_1, x_2)) > \varepsilon/4\},$$

which is all of the $x_2 \in \Omega_2$ such that the set of tail sequences after fixing $x_1 = y_1$ have a measure greater than $\varepsilon/4$. Thus, redoing the above with G_i replacing F_i, we find that the G_i are also decreasing, but with $P_2(G_i) > \varepsilon/4$. Hence, the $\bigcap_i G_i \neq \emptyset$, so we can fix a $y_2 \in \bigcap_i G_i$.

Continuing via induction, we can construct a sequence $\{y_i\}_{i=1}^{\infty}$ with $y_i \in \Omega_i$ such that $P^{(n)}(A_i^{(n)}(y_1, \ldots, y_n)) \geq \varepsilon/2^n$ for all i. Lastly, we need to show that the sequence $\{y_i\}_{i=1}^{\infty} \in A_j$ for all $j = 1, \ldots, \infty$. Thus, for each j, we can select an $n_j \in \mathbb{N}$ large enough so that for all $(x_1, \ldots, x_{n_j}) \in \Omega_1 \times \ldots, \Omega_{n_j}$, we have $A_j^{(n_j)}(x_1, \ldots, x_{n_j})$ is either $\emptyset$ or $\Omega^{(n)}$, which is possible since each $A_j \in \mathcal{S}$. Thus, we must have that for n large enough $A_j^{(n)}(y_1, \ldots, y_n) = \Omega^{(n)}$. Since $(y_1, y_2, \ldots) \in A_j$ for all j, $\bigcap_{j=1^{\infty}} A_j \neq \emptyset$. Thus, P is countably additive on $\mathcal{S}$ and thus we can extend it uniquely to $\mathcal{F}$. $\square$

Chapter 3

Probability Theory

Introduction

In this chapter, are aim to to prove the Law of Large Numbers and the Ergodic theorem. To get there, we will require two preliminary sections discussing some theory around L^p spaces and what it means for a sequence of measures to converge to another measure.

3.1 L^p spaces

L^p spaces are a standard example of Banach spaces, which are complete normed linear space. We have already seen these when discussing the space of all absolutely integrable functions. More generally, we have the following definition.

Definition 3.1.1 (L^p space)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $f : \Omega \to [-\infty, \infty]$ a measurable function, then we say $f \in L^p(\Omega, \mathcal{F}, \mu)$, for $1 \leq p < \infty$ if*

$$\int |f|^p d\mu < \infty.$$

For $p = \infty$, we say $f \in L^\infty(\Omega, \mathcal{F}, \mu)$ if $\inf\{t \in [-\infty, \infty] : |f| \leq t \ \mu\, a.e.\} < \infty$.

This definition allows us to write down the L^p norm, which will be shown to be a norm below:

$$\begin{aligned} \|f\|_p &= \left[\int |f|^p d\mu\right]^{1/p} && 1 \leq p < \infty \\ \|f\|_\infty &= \inf\{t \in [-\infty, \infty] : |f| \leq t \ \mu\, a.e.\} && p = \infty \\ &= \inf\{t \in [-\infty, \infty] : \mu(\{|f| > t\}) = 0\} \end{aligned}$$

The L^∞ norm is sometimes referred to as the *essential supremum*. For $1 < p, q < \infty$, we say that p and q are conjugate indices if $p^{-1} + q^{-1} = 1$. In this context, we also say that 1 and ∞ are conjugates.

The following subsections contain many important inequalities in the theory of L^p spaces. In the following theorems, we assume f and g are measurable functions in $(\Omega, \mathcal{F}, \mu)$ unless stated otherwise.

3.1.1 Markov / Chebyshev and Jensen's inequalities

The following results are ambiguously referred to as Markov's inequality, Chebyshev's inequality, and Chernoff's inequality. They allow us to bound the measure of a set by an integral of a measurable function. In the context of probability theory, we are bounding a tail probability by the moments of a random variable.

Theorem 3.1.1 (Markov's Inequality)**.** *Let f be a non-negative measurable function and $t > 0$. Then, denoting $\{f > t\} := \{\omega \in \Omega : f(\omega) > t\}$,*

$$\mu(\{f > t\}) \leq t^{-1} \int f d\mu.$$

Proof. Noting that $t\mathbf{1}_{\{f>t\}} \leq f$, then by monotonicity of the integral,

$$t\mu(\{f > t\}) = \int t\mathbf{1}_{\{f>t\}} d\mu \leq \int f d\mu,$$

which proves the theorem. □

There are two useful inequalities resulting from Markov's inequality, which are

1. Chebyshev's Inequality: For f measurable and $m \in \mathbb{R}$,

$$\mu(\{|f - m| > t\}) \leq t^{-2} \int (f - m)^2 d\mu.$$

   For probability measures and random variables this is $P(|X - \mathrm{E}X| > t) \leq \mathrm{Var}(X)/t^2$.

2. Chernoff's Inequality: For f measurable and $\eta \in \mathbb{R}$,

$$\mu(\{f > t\}) \leq \mathrm{e}^{-\eta t} \int \mathrm{e}^{\eta f} d\mu.$$

   In probability theory, the right hand side becomes the moment generating function or the Laplace transform.

For Jensen's inequality, we need the definition of a convex function.

Definition 3.1.2 (Convex Functions on $\mathbb{R}$)**.** *Let $I \subseteq \mathbb{R}$ be an interval. A function $\phi : I \to \mathbb{R}$ is convex if for all $t \in [0, 1]$ and all $x, y \in I$,*

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

Theorem 3.1.2 (Jensen's Inequality)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space and X an integrable random variable[1] such that $X : \Omega \to I \subseteq \mathbb{R}$. For any convex $\phi : I \to \mathbb{R}$,*

$$\phi\left(\int X d\mu\right) \leq \int \phi(X) d\mu,$$

which is $\phi(\mathrm{E}[X]) \leq \mathrm{E}[\phi(X)]$.

[1] i.e. a measurable function in $L^1(\Omega, \mathcal{F}, P)$

Proof. For some $c \in I$, if $X = c$, P-a.e., then the result is immediate. Otherwise, let $m = \mathrm{E}X$ be the mean of X, which lies in the interior of interval I. Then, we can choose $a, b \in \mathbb{R}$ such that $\phi(x) \geq ax + b$ for all $x \in I$ with equality at $x = m$.[2] Then, $\phi(X) \geq aX + b$, and

$$\phi(\mathrm{E}[X]) = am + b = \mathrm{E}[aX + b] \leq \mathrm{E}[\phi(X)].$$

To check that $\mathrm{E}[\phi(X)]$ is well defined (i.e. not $\infty - \infty$), we note that $\phi = \phi^+ - \phi^-$ where ϕ^- is concave and $\phi^-(x) \leq |a||x| + |b|$. Hence, $\mathrm{E}[\phi(X)] \leq |a|\mathrm{E}|x| + |b| < \infty$. □

3.1.2 Hölder and Minkowski's Inequalities

Theorem 3.1.3 (Hölder's inequality)**.** *Let $p, q \in [1, \infty]$ be conjugate indices and f and g be measurable, then $\|fg\|_1 \leq \|f\|_p \|g\|_q$.*

Proof. If either $\|f\|_p = 0, \infty$ or $\|g\|_q = 0, \infty$, then the result is immediate. Hence, for f such that $0 < \|f\|_p < \infty$, we can normalize and without loss of generality assume that $\|f\|_p = 1$. Hence, we can define a probability measure P on $\mathcal{F}$ such that for any $A \in \mathcal{F}$,

$$P(A) := \int_A f \, d\mu.$$

Then, using Jensen's inequality and noting that $q(p-1) = p$,

$$\begin{aligned}
\|fg\|_1 = \int |fg| d\mu &= \int \frac{|g|}{|f|^{p-1}} \mathbf{1}_{|f|>0} |f|^p d\mu \\
&\leq \left[\int \frac{|g|^q}{|f|^{q(p-1)}} \mathbf{1}_{|f|>0} |f|^p d\mu \right]^{1/q} \\
&\leq \left[\int |g|^q d\mu \right]^{1/q} = \|f\|_p \|g\|_q.
\end{aligned}$$

□

The most famous version of Hölder's inequality is the Cauchy-Schwarz inequality, which is just the setting where $p = q = 2$.

Corollary 3.1.3 (Cauchy-Schwarz Inequality)**.** *For measurable f and g, $\|fg\|_1 \leq \|f\|_2 \|g\|_2$.*

Minkowski's inequality shows that $\|\cdot\|_p$ is subadditive, which is one of the conditions for $\|\cdot\|_p$ to be a norm.

Theorem 3.1.4 (Minkowski's inequality)**.** *Let $p \in [1, \infty]$ and f and g be measurable, then $\|f + g\|_p \leq \|f\|_p + \|g\|_p$.*

[2] This is a property of convex functions. Try to show it yourself!

Proof. If either $\|f\|_p = \infty$ or $\|g\|_p = \infty$, then we are done. If $\|f+g\|_p = 0$, then we are done. If $p = 1$, then $|(f+g)(\omega)| \leq |f(\omega)| + |g(\omega)|$, and the result follows quickly.

In all other cases, note that

$$|f+g|^p = 2^p \left|\frac{f+g}{2}\right|^p \leq 2^p \left(\frac{1}{2}|f|^p + \frac{1}{2}|g|^p\right) = 2^{p-1}\left(|f|^p + |g|^p\right).$$

This implies that

$$\int |f+g|^p d\mu \leq 2^{p-1}\int |f|^p d\mu + 2^{p-1}\int |g|^p d\mu < \infty$$

Thus, $f+g \in L^p(\Omega, \mathcal{F}, \mu)$ if both $f, g \in L^p(\Omega, \mathcal{F}, \mu)$.

Assuming $\|f+g\|_p > 0$ and $p > 1$ and p, q conjugates, we have that

$$\||f+g|^{p-1}\|_q = \left[\int |f+g|^{(p-1)q} d\mu\right]^{1/q} = \left[\int |f+g|^p d\mu\right]^{\frac{p-1}{p}} = \|f+g\|_p^{p-1}.$$

Finally, using the above equality, we have

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

Dividing both sides by $\|f+g\|_p^{p-1}$ finishes the proof. □

A nice application of the above results is showing that simple functions can approximate any L^p functions. In the previous chapter, we used simple functions to approximate measurable functions and their integrals (for example, see Theorem 2.3.2).

Theorem 3.1.5 (L^p approximation)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and let $\mathcal{A}$ be a π-system such that $\sigma(\mathcal{A}) = \mathcal{F}$ and $\mu(A) < \infty$ for all $A \in \mathcal{A}$ and there exists $A_i \uparrow \Omega$, $A_i \in \mathcal{A}$.*[3] *Let the collection of simple functions be*

$$V_0 := \left\{\sum_{i=1}^n a_i \mathbf{1}_{A_i} \,:\, a_i \in \mathbb{R}, A_i \in \mathcal{A}, n \in \mathbb{N}\right\}.$$

For $p \in [1, \infty)$, $V_0 \subset L^p$, and for all $f \in L^p$ and all $\varepsilon > 0$, there exists a $v \in V_0$ such that $\|f - v\|_p < \varepsilon$.

Proof. For any $A \in \mathcal{A}$, $\|\mathbf{1}_A\|_p = \left(\int \mathbf{1}_A d\mu\right)^{1/p} = \mu(A)^{1/p} < \infty$. Thus, $\mathbf{1}_A \in L^p$ for all $A \in \mathcal{A}$. Since L^p is a linear space, $V_0 \subset L^p$.

Next, let $V \subseteq L^p$ be all of the f that can be approximated by some $v \in V_0$. Let f be approximated by v_f and g by v_g, then by Minkowski's inequality,

$$\|(f+g) - (v_f + v_g)\|_p \leq \|f - v_f\|_p + \|g - v_g\|_p \leq 2\varepsilon.$$

[3] For example, Lebesgue measure λ with the A being half-open bounded intervals and $\mathcal{F} = \mathcal{B}$.

Hence, V is also a linear space.

Now, we assume $\Omega \in \mathcal{A}$ (i.e. $\mu(\Omega) < \infty$). Let $\mathcal{L} = \{B \in \Omega : \mathbf{1}_B \in V\}$, which we will show is, in fact, a λ-system. We know that $\mathcal{A} \subseteq \mathcal{L}$ and thus $\Omega \in \mathcal{L}$. Next, for $A, B \in \mathcal{L}$ such that $A \subseteq B$, then $\mathbf{1}_{B\setminus A} = 1_B - 1_A \in V$ since V is linear, so $B \setminus A \in \mathcal{L}$. Lastly, for $\{A_i\}_{i=1}^{\infty}$ pairwise disjoint with $A_i \in \mathcal{L}$, let $A = \bigcup_{i=1}^{\infty} A_i$ and $B_j = \bigcup_{i=1}^{j} A_i$. Then, $B_j \uparrow A$ and $\|\mathbf{1}_A - \mathbf{1}_{B_j}\|_p = \mu(A \setminus B_j)^{1/p} \to 0$. Therefore, $A \in \mathcal{L}$, and thus $\mathcal{L}$ is a λ-system. By Dynkin's π-λ theorem, $\mathcal{F} \subseteq \mathcal{L}$ and thus $\mathbf{1}_B \in V$ for any $B \in \mathcal{F}$. Therefore, for any nonnegative $f \in L^p$, we can construct simple functions $f_n = \min\{n, 2^{-n}\lfloor 2^n f \rfloor\}$ such that $f_n \uparrow f$. Then, $|f - f_n|^p \to 0$ pointwise and $|f - f_n|^p \leq |f|^p$. Hence, by dominated convergence, $\|f - f_n\|_p \to 0$. Therefore, $f \in V$ and, by the linearity of V, $V = L^p$.[4]

Lastly, for general Ω, we have by assumption a sequence $A_i \uparrow \Omega$. Hence, for any $f \in L^p$, we have that $f\mathbf{1}_{A_i} \in V$, and similarly to above, $|f - f\mathbf{1}_{A_i}|^p \to 0$ pointwise and $|f - f\mathbf{1}_{A_i}|^p \leq |f|^p$. Therefore, $\|f - f\mathbf{1}_{A_i}\|_p \to 0$ by dominiated convergence. Therefore, $f \in V$. □

3.2 Convergence in Probability & Measure

In this section, we will discuss many types of convergence. Argueably, there are too many different types of convergence all with their own special properties and hierachies from weakest to strongest notion.

3.2.1 Convergence of Measure

Note that in this section, all measures are probability measures unless otherwise stated.

Given a measurable space $(\Omega, \mathcal{F})$ and a sequence of probability measures $\{P_i\}_{i=1}^{\infty}$, what do we want $P_i \to P$ to mean? At a minimum, we want $P_i(A) \to P(A)$ for any $A \in \mathcal{F}$, which is sometimes called setwise convergence, but this often is not enough. One of the most used notions of covergence of measure is weak convergence, which requires us to switch to a metric space.

Definition 3.2.1 (Weak Convergence of Measure)**.** *Let S be a metric space and $\mathcal{S}$ be the Borel σ-field on S. Then, for a measure P and a sequence $\{P_i\}_{i=1}^{\infty}$, we say that P_i converges weakly to P, i.e. $P_i \Rightarrow P$, if*

$$\int f\, dP_i \to \int f\, dP$$

for all f bounded continuous real-valued functions on S. We will write $f \in \mathcal{C}_B(\mathbb{R})$ for such functions.

The notion of weak convergence is tied into the topology of S generated by the metric d. Furthermore, having measures converge implies a *closeness* between measures P_i and P_{i+1}. For a finite collection of $f_1, \ldots, f_n \in \mathcal{C}_B(\mathbb{R})$, we can define an ε-neighbourhood of P to be all of the measures Q such that $|\int f_i dP - \int f_i dQ| < \varepsilon$. There are many equivalent notions of weak convergence, which are listed in the *Portmanteau Theorem.*

[4] i.e. as before write a general $f = f^+ - f^-$.

Theorem 3.2.1 (Portmanteau Theorem)**.** *For P and P_i on $(S, \mathcal{S})$, the following are equivalent:*

1. $P_i \Rightarrow P$
2. $\int f\, dP_i \to \int f\, dP$ *for all bounded uniformly continuous functions* f
3. $\limsup_i P_i(C) \leq P(C)$ *for all closed* C
4. $\liminf_i P_i(U) \geq P(U)$ *for all open* U
5. $\lim_i P_i(A) = P(A)$ *for all* $A \in \mathcal{S}$ *such that* $P(\partial A) = 0$[5]

Proof. (1)→(2) If convergence holds for every $f \in \mathcal{C}_B$ then is certainly hold for all bounded uniformly continous f.

(2)→(3) For any C closed and $\varepsilon > 0$ there exists a $\delta > 0$ such that for $C_\delta = \{x \in S : d(x, C) < \delta\}$, we have $P(C_\delta) < P(C) + \varepsilon$ as $C_\delta \downarrow C$ as $\delta \to 0^+$. Then, we can define an f such that $f = 1$ on C, $f = 0$ on $S \setminus C_\delta$, and f is uniformly continuous and $0 \leq f \leq 1$.[6] Then, by assumption (2), we have that

$$P_i(C) \leq \int f\, dP_i \to \int f\, dP \leq P(C_\delta) < P(C) + \varepsilon.$$

Thus, taking the lim sup and ε to zero gives $\limsup_i P_i(C) \leq P(C)$.

(3)→(1) Let $f \in \mathcal{C}_B$. Our goal is to show that $\limsup_i \int f\, dP_i \leq \int f\, dP$ and similarly for lim inf to show (1) holds. As f is bounded, we can shift and scale it, and without loss of generality, we assume that $0 < f < 1$. Then, for any choice of $n \in \mathbb{N}$, we can define nested closed sets $C_j = \{x \in S : f(x) \geq j/n\}$ for $j = 0, 1, \ldots, n$ and cut f into pieces to get

$$\sum_{j=1}^{n} \frac{j-1}{n} P(C_{j-1} \setminus C_j) \leq \int f\, dP \leq \sum_{j=1}^{n} \frac{j}{n} P(C_{j-1} \setminus C_j).$$

Noting that $P(C_{j-1} \setminus C_j) = P(C_{j-1}) - P(C_j)$, the above becomes

$$\frac{1}{n} \sum_{j=1}^{n} P(C_j) \leq \int f\, dP \leq \frac{1}{n} + \frac{1}{n} \sum_{j=1}^{n} P(C_j).$$

Thus,

$$\limsup_i \int f\, dP_i \leq \frac{1}{n} + \frac{1}{n} \sum_{j=1}^{n} \limsup_i P_i(C_j) \leq \frac{1}{n} + \frac{1}{n} \sum_{j=1}^{n} P(C_j) \leq \frac{1}{n} + \int f\, dP.$$

Taking $n \to \infty$ gives $\limsup_i \int f\, dP_i \leq \int f\, dP$. Replacing f with $-f$ gives $\liminf_i \int f\, dP_i \geq \int f\, dP$. Thus the lim sup and lim inf coincide proving that (3)→(1).

The equivalence of (4) and (5) with the rest is omitted, but can be found in most probability textbooks. □

[5] where ∂A is the boundary of A, which is $\partial A = \bar{A} \cap \bar{A}^c$ where $\bar{A}$ is the closure of A.

[6] See Urysohn's Lemma or Theorem 1.2 in [Billingsley(2013)]

Remark 3.2.2 (Other ways measures converge)**.** *As noted about P_i converges weakly if $\int f dP_i \to \int f dP$ for each $f \in \mathcal{C}_B$. Changing the space where f lives changes the convergence. Convergence in the Radon metric is*

$$\sup_f \left\{ \int f dP - \int f dP_i \right\} \to 0$$

where the sup is taken over all continuous functions $f : S \to [-1, 1]$. If we take the sup over all measurable $f : S \to [-1, 1]$, then we have convergence in total variation. If the sup is over all Lipschitz $f : S \to [-1, 1]$ with a Lipschitz constant of 1, then this is convergence in the 1-Wasserstein metric.

3.2.2 Convergence of Random Variables

Convergence in Distribution

In contrast to convergence of measures, let $(\Omega, \mathcal{F}, \mu)$ be a probability space and $(S, \mathcal{S})$ be a metric space with Borel sets as above. Then, for a random variable (i.e. measurable function) $X : \Omega \to S$, we can define a probability measure

$$P(A) = \mu(X^{-1}(A)), \quad A \in \mathcal{S}.$$

This is the distribution of X. For a sequence of random variables $\{X_i\}_{i=1}^{\infty}$, we say that X_i converges to X in distribution—denoted $X_i \xrightarrow{\text{d}} X$—means that $P_i \Rightarrow P$.[7] The expectation of a random variable can be written in multiple ways due to change of variables:

$$\text{E}[X] = \int_\Omega X(\omega) d\mu(\omega) = \int_S x dP(x).$$

We also often write $P_i(A)$ as $\text{P}(X_i \in A)$, which is a bit ambiguous. Note that above Portmanteau theorem can be rephrased for random variables as follows.

Convergence in Probability

Given the same setup as above, we say that X_i converges in probabilty to X—denoted $X_i \xrightarrow{\text{P}} X$—if for all $\varepsilon > 0$

$$\mu\left(\{\omega \in \Omega : d(X_i(\omega), X(\omega)) > \varepsilon\}\right) \to 0.$$

This means that the measure of the set of ω where $X_i(\omega)$ and $X(\omega)$ differ by more than ε goes to zero as $i \to \infty$. Note that this is often written in shorthand as $\text{P}(d(X_i, X) > \varepsilon) \to 0$. Convergence in probability is closely connected to the metric d on $(S, \mathcal{S})$.

[7] Note that since we only really care about weak convergence of measure, the metric space $(S, \mathcal{S})$ must be fixed, but the initial probabilty space $(\Omega, \mathcal{F}, \mu)$ is allowed to change.

Convergence Almost Surely

Given the same setup as above, we say that X_i converges almost surely to X—denoted $X_i \xrightarrow{\text{a.s.}} X$—if

$$\mu\left(\{\omega \in \Omega \,:\, X_i(\omega) \to X(\omega)\}\right) = 0.$$

This was already mentioned in the section on important integral convergence theorems. It implies that X_i converges to X pointwise except on a set of measure zero.

Convergence in L^p

Given the same setup as above, we say that X_i converges to X in L^p if

$$\mathrm{E}[d(X_i, X)^p] = \int d(X_i(\omega), X(\omega))^p d\mu(\omega) \to 0.$$

Here, we can think of $d(X_i(\omega), X(\omega))$ as a function from Ω to $\mathbb{R}^+$. In the case that we have real valued random variables, i.e. $S = \mathbb{R}$, then this is

$$\int |X_i - X|^p d\mu \to 0.$$

Hierarchy of Convergence Types

Some of the above types of convergence are stronger or weaker than others in the sense that one type implies another. Here is a short list of such implications:

- Convergence almost surely implies convgerence in probability.
- Convergence in probability implies convgence in distribution.
- For $1 \leq q < p \leq \infty$, convergence in L^p implies convergence in L^q.[8]
- For any $p \in [1, \infty]$, convergence in L^p implies convergence in probability.[9]

Also, consider what extra conditions are necessary to make almost sure convergence imply L^p convergence and vice versa.

3.2.3 Borel-Cantelli Lemmas

Let $(\Omega, \mathcal{F}, \mu)$ be a probability space as before. For $\{A_i\}_{i=1}^{\infty}$, $A_i \in \mathcal{F}$, then

$$\limsup_i A_i = \bigcap_{i=1}^{\infty} \bigcup_{j>i} A_j \quad \text{and} \quad \liminf_i A_i = \bigcup_{i=1}^{\infty} \bigcap_{j>i} A_j.$$

The set $\limsup_i A_i$ is sometimes referred to as A_i *infinitely often* or A_i i.o. This is because $\omega \in \limsup_i A_i$ implies that for any $N \in \mathbb{N}$ there exists an $n > N$ such that

[8] Try to show this using one of the inequalities from the previous section.
[9] Once again, try to show this using one of the inequalities from the previous section.

$\omega \in A_n$. Similarly, some write A_i *eventually* (or A_i ev.) for $\liminf_i A_i$. This is because for $\omega \in \liminf_i A_i$ then there exists an N large enough such that $\omega \in A_n$ for all $n \geq N$.

The Borel-Cantelli lemmas are a very useful tool to use when proving convergence theorems. We will use these to prove the strong law of large numbers.

Theorem 3.2.2 (1st Borel-Cantelli Lemma)**.** *Let $\{A_i\}_{i=1}^{\infty}$ with $A_i \in \mathcal{F}$. If $\sum_{i=1}^{\infty} \mu(A_i) < \infty$, then $\mu(\limsup_i A_i) = 0$.*[10]

Proof. Noting that if a summation converges, then the tail sum has to tend to zero, we have simply that

$$\mu(\limsup_i A_i) = \mu\left(\bigcap_{i=1}^{\infty}\bigcup_{j>i} A_i\right) \leq \mu\left(\bigcup_{j>i} A_i\right) \leq \sum_{j>i} \mu(A_i) \to 0$$

as $i \to \infty$ where the first inequality comes from monotonicity and the second comes from subadditivity. □

Theorem 3.2.3 (2nd Borel-Cantelli Lemma)**.** *Let $\{A_i\}_{i=1}^{\infty}$ be an independent collection with $A_i \in \mathcal{F}$. If $\sum_{i=1}^{\infty} \mu(A_i) = \infty$, then $\mu(\limsup_i A_i) = 1$.*[11]

Proof. Note that $1 - t \leq \mathrm{e}^{-t}$ for all $t \in \mathbb{R}$. One can check that the independence of the $\{A_i\}_{i=1}^{\infty}$ implies the independence $\{A_i^c\}_{i=1}^{\infty}$. Therefore, for any $i \in \mathbb{N}$ and $k \geq i$,

$$\mu\left(\bigcap_{j=i}^{k} A_j^c\right) = \prod_{j=i}^{k}[1 - \mu(A_j)] \leq \exp\left[-\sum_{j=i}^{k} \mu(A_j)\right].$$

Taking $k \to \infty$ takes the right hand side to zero. Hence, $\mu(\bigcap_{j>i} A_j^c) = 0$ for all i. Thus,

$$\mu(\limsup_i A_i) = \mu\left(\bigcap_{i=1}^{\infty}\bigcup_{j>i} A_j\right) = 1 - \mu\left(\bigcup_{i=1}^{\infty}\bigcap_{j>i} A_j^c\right) = 1.$$

□

3.2.4 Prohorov's Theorem

In this section, we just quickly state Prohorov's Theorem[12] to be used later in proving the Central Limit Theorem. Prohorov's Theorem discusses sequential compactness for a sequence of measures much as how the Bolzano-Weierstrauss Theorem[13] discusses compactness for bounded sequences in $\mathbb{R}^d$. See [Billingsley(2013)] Chapter 1, Section 6 for more.

[10] i.e. the set of ω that occur infinitely often as zero probability.
[11] i.e. the set of ω that occur infinitely often has probability 1.
[12] https://en.wikipedia.org/wiki/Prokhorov%27s_theorem
[13] https://en.wikipedia.org/wiki/Bolzano%E2%80%93Weierstrass_theorem

Definition 3.2.3 (Uniform Tightness)**.** *A collection of probability measures $\{\mu_i\}_{i\in I}$ in a metric space is said to be uniformly tight if for every $\varepsilon > 0$, there exists a compact set K_ε such that $\mu_i(K_\varepsilon) > 1 - \varepsilon$ for all i.*

Theorem 3.2.4 (Prohorov's Theorem)**.** *For a sequence of probability measures $\{\mu_i\}_{i=1}^\infty$, if the sequence is uniformly tight then it is relatively compact (sequentially compact)—i.e. for every subsequence μ_{i_k} there exists a weakly convergence subsubsequence $\mu_{i_{k_r}} \Rightarrow \mu$ for some probability measure μ depending on the subsequence.*

A nice convergence result relying on subsubsequences is the following proposition.

Proposition 3.2.4. *If $\{\mu_i\}_{i=1}^\infty$ and μ are probability measures such that for every subsequence μ_{i_k}, there exists a subsubsequence $\mu_{i_{k_r}} \Rightarrow \mu$, then $\mu_i \Rightarrow \mu$.*

Proof. Assume this is not the case, then there exists a continuous bounded function f such that $\int f d\mu_i \not\to \int f d\mu$. Thus, for some subsequence i_k and $\varepsilon > 0$,

$$\left| \int f d\mu_{i_k} - \int f d\mu \right| > \varepsilon$$

for all k. However, $\mu_{i_{k_r}} \Rightarrow \mu$ contradicts this. □

3.3 Law of Large Numbers

The goal of this section is to prove the strong law of large numbers, which is a pivotal result in probability and statistics. First, we will prove the weak law of large numbers with stronger than necessary assumptions. This is mainly to contrast how much more work in involved to prove the strong law with weaker assumptions.

In this section, we will consider an infinite collection of random variables $\{X_i\}_{i=1}^\infty$ from $(\Omega, \mathcal{F}, P)$ to $(\mathbb{R}, \mathcal{B})$. Hence, for $A \in \mathcal{B}$, we write

$$P(X \in A) := P(\{\omega \in \Omega \,:\, X(\omega) \in A\})$$

and $\mathrm{E}X = \int X(\omega) dP$. Furthermore, we define the partial sum $S_n = \sum_{i=1}^n X_i$, which is also a measurable random variable.

Before discussing the laws of large numbers, we need to define independence for random variables.

Definition 3.3.1. *For random variables X and Y on the same probability space $(\Omega, \mathcal{F}, P)$ but possibly with different codomains, $(\mathbb{X}, \mathcal{X})$ and $(\mathbb{Y}, \mathcal{Y})$ respectively, we say that X and Y are independent if*

$$P(\{X \in A\} \cap \{Y \in B\}) = P(X \in A)P(Y \in B)$$

for all $A \in \mathcal{X}$ and $B \in \mathcal{Y}$.

This definition can be extended to a finite collection of random variables $\{X_i\}_{i=1}^n$ *implying*

$$P\left(\bigcap_{i=1}^{n}\{X_i \in A_i\}\right) = \prod_{i=1}^{n} P(X_i \in A_i)$$

for all A_i*. We say that an infinite collection of random variables is independent if all finite collections are independent.*

Note that since $\{X \in A\}$ is shorthand for $\{\omega \in \Omega : X(\omega) \in A\} = X^{-1}(A)$, random variables X and Y are independent if and only if the σ-fields $\sigma(X)$ and $\sigma(Y)$ are independent as discussed in Chapter 1 of these notes.

We also need to rigorously define what it means to have radom variables that are identically distributed. For $X : \Omega \to \mathbb{R}$, the *law* or *distribution* of X is the measure induced by X. That is, $P \circ X^{-1}(A)$ for $A \in \mathcal{B}$, say. We say that X and Y are identically distributed if $P \circ X^{-1}$ and $P \circ Y^{-1}$ coincide almost surely. Note that if X and Y are identically distributed, then we are implying that they have the same domain and codomain.

Theorem 3.3.1 (Weak Law of Large Numbers)**.** *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space and* $\{X_i\}_{i=1}^{\infty}$ *be random variables (measurable functions) from* Ω *to* $\mathbb{R}$ *such that* $\mathrm{E}X_i = c \in \mathbb{R}$ *and* $\mathrm{E}X_i^2 = 1$ *for all* i *and* $\mathrm{E}[(X_i - c)(X_j - c)] = 0$ *for all* $i \neq j$*. Then* $S_n/n \xrightarrow{P} c$*.*

Proof. Without loss of generality, we assume $c = 0$. Otherwise, we can replace X_i with $X_i - c$. Then, for any $t > 0$, Chebyshev's inequality implies that

$$P\left(\frac{|S_n|}{n} \geq \varepsilon\right) \leq \frac{\mathrm{E}S_n^2}{t^2 n^2} = \frac{\sum_{i,j=1}^{n} \mathrm{E}X_i X_j}{t^2 n^2} = \frac{1}{nt^2} \to 0$$

as $n \to \infty$. □

Note that in the above proof, we only require that the X_i be uncorrelated (i.e. $\mathrm{E}[(X_i - c)(X_j - c)] = 0$) and not independent. In the next theorem, we require independence, but remove all second moment conditions. However, we still need the notation

$$\mathrm{Var}\,(X) = \int (X - \mathrm{E}X)^2 dP(\omega).$$

Theorem 3.3.2 (Strong Law of Large Numbers)**.** *Let* $\{X_i\}_{i=1}^{\infty}$ *be iid random variables from* Ω *to* $\mathbb{R}$*. If* $\mathrm{E}|X_i| < \infty$*, then* $S_n/n \xrightarrow{a.s.} c$ *for* $c = \mathrm{E}X_i$*. If* $\mathrm{E}|X_i| = \infty$*, then* S_n/n *does not converge to any finite value.*

Proof. We first quickly prove the second part of the theorem. Assume that $n^{-1}S_n \to c \in \mathbb{R}$ but also that $\mathrm{E}|X_i| = \infty$, and note that $n^{-1}X_n = n^{-1}(S_n - S_{n-1}) \to 0$. Since $\mathrm{E}|X_i| = \infty$, then $\sum_{n=0}^{\infty} \mathrm{P}\,(X_i > n) = \infty$ and Borel-Cantelli says that $|X_n| > n$ for infinitely many n. Thus

$$\mathrm{P}\left(\{\omega : n^{-1}(S_n(\omega) - S_{n-1}(\omega)) \to 0\}\right) = 0.$$

Thus, $n^{-1}S_n \not\to c \in \mathbb{R}$.

Most our effort will be for proving the above assuming $\mathrm{E}X_i = c \in \mathbb{R}$. Without loss of generality, we assume $X_i \geq 0$ for all i. This is doable as we can write a general $X = X^+ - X^-$ and independence of X and Y implies independence for X^+ and Y^+.[14] Also, we use F to denote the law of X, i.e. $F(x) = \mathrm{P}(X \leq x)$.

We define $Y_i = X_i \mathbf{1}_{X_i \leq i}$ and $T_n = \sum_{i=1}^n Y_i$ as bounded analogues to X_i and S_n, so that their variance is finite. Our goal is to use both Chebyshev and Borel-Cantelli. For any $\delta > 1$, we can define a non-decreasing integer sequence $k_n = \lfloor \delta^n \rfloor$. Then, $1 \leq k_n \leq \delta^n < k_n + 1 \leq 2k_n$ and $k_n^{-2} \leq 4\delta^{-2n}$ and furthermore

$$\sum_{n=1}^{\infty} k_n^{-2} \mathbf{1}_{k_n \geq i} \leq 4 \sum_{n=1}^{\infty} \delta^{-2n} \mathbf{1}_{\delta^n \geq i} \leq \frac{4}{i^2(1-\delta^{-2})} \leq c_0 i^{-2} \tag{3.3.1}$$

for some constant $c_0 > 0$. We also note that $\sum_{i=k+1}^{\infty} i^{-2} < \int_k^{\infty} x^{-2} dx = 1/k$. By Chebyshev's inequality, for any $t > 0$, there exists a constant c_1 depending on t and δ such that

$$\begin{aligned}
\sum_{n=1}^{\infty} P\left(|T_{k_n} - \mathrm{E}T_{k_n}| > tk_n\right) &\leq c_1 \sum_{n=1}^{\infty} k_n^{-2} \mathrm{Var}\left(T_{k_n}\right) && \text{[Chebyshev]} \\
&= c_1 \sum_{n=1}^{\infty} \frac{1}{k_n^2} \sum_{i=1}^{k_n} \mathrm{Var}\left(Y_i\right) \\
&= c_1 \sum_{i=1}^{\infty} \mathrm{Var}\left(Y_i\right) \sum_{\{k_n \geq i\}} \frac{1}{k_n^2} \\
&\leq c_2 \sum_{i=1}^{\infty} i^{-2} \mathrm{E}Y_i && \text{[Eqn 3.3.1]} \\
&= c_2 \sum_{i=1}^{\infty} i^{-2} \int_0^i x^2 dF(x) \\
&= c_2 \sum_{i=1}^{\infty} i^{-2} \left\{ \sum_{k=0}^{i-1} \int_k^{k+1} x^2 dF(x) \right\} \\
&\leq c_3 \sum_{k=0}^{\infty} \frac{1}{k+1} \int_k^{k+1} x^2 dF(x) \\
&\leq c_3 \sum_{k=0}^{\infty} x dF(x) = c_3 \mathrm{E}X_i < \infty. && [x/(k+1) < 1]
\end{aligned}$$

And thus $\sum_{n=1}^{\infty} P\left(|T_{k_n} - \mathrm{E}T_{k_n}| > tk_n\right) < \infty$. Hence, by Borel-Cantelli, $k_n^{-1}|T_{k_n} - \mathrm{E}T_{k_n}| \xrightarrow{\text{a.s.}} 0$. Since $\mathrm{E}Y_n \uparrow \mathrm{E}X_i$, we have that $k_n^{-1}\mathrm{E}T_{k_n} \uparrow \mathrm{E}X_i$ and in turn that $k_n^{-1}T_{k_n} \xrightarrow{\text{a.s.}} \mathrm{E}X_i$.

[14] Check that for any measureable $f, g : \mathbb{R} \to \mathbb{R}$ that X and Y independent implies that $f(X)$ and $g(Y)$ are independent.

To get back to X_i and S_n, we note that $\mathrm{E}X < \infty$ if and only if $\sum_{i=0}^{\infty} \mathrm{P}(X > i) < \infty$.[15] Thus $\sum_{i=1}^{\infty} \mathrm{P}(X_i \neq Y_i) = \sum_{i=1}^{\infty} \mathrm{P}(X_i > i) < \infty$ and Borel-Cantelli says that $\mathrm{P}(\limsup\{X_i \neq Y_i\}) = 0$ so for i large enough, $X_i = Y_i$ a.s. We define "large enough" to be $i > m(\omega)$.[16] Furthermore, $k_n^{-1} S_{m(\omega)} \to 0$ and $k_n^{-1} T_{m(\omega)} \to 0$ as $n \to$ meaning that the contribution of the terms where X_i and Y_i may not coincide becomes negligible. Hence, $k_n^{-1} S_{k_n} \xrightarrow{\text{a.s.}} \mathrm{E}X_i$, so we have almost sure convergence of a subsequence.

Finally, since $k_{n+1}/k_n \to \delta$, there exists an n large enough such that $1 \leq k_{n+1}/k_n < \delta^2$. Thus, for $k_n < i < k_{n+1}$,

$$k_n^{-1} S_{k_n} \leq \delta^2 \frac{S_i}{i} \leq \delta^4 k_{n+1}^{-1} S_{k_{n+1}} \quad \text{and}$$
$$\delta^{-2} \mathrm{E}X_i \leq \liminf_{i\to\infty} \frac{S_i}{i} \leq \limsup_{i\to\infty} \frac{S_i}{i} \leq \delta^2 \mathrm{E}X_i.$$

Thus, taking $\delta \downarrow 0$ concludes the proof. □

3.4 Central Limit Theorem

To discuss the central limit theorem, we must first discuss what a Gaussian random variable is.

Definition 3.4.1 (Gaussian Measure on $\mathbb{R}$)**.** *A Borel measure γ on $(\mathbb{R}, \mathcal{B})$ is said to be Gaussian with mean m and variance σ if*

$$\gamma((a,b]) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b \exp\left(-\frac{1}{2\sigma^2}(x-m)^2\right) d\lambda(x)$$

If $\gamma = \delta_m$, a Dirac measure at m, we say γ is a degenerate Gaussian measure.

Definition 3.4.2 (Gaussian Measure on $\mathbb{R}^d$)**.** *A Borel measure γ on $(\mathbb{R}^d, \mathcal{B})$ is said to be Gaussian if for all linear functionals $f : \mathbb{R}^d \to \mathbb{R}$, the induced measure $\gamma \circ f^{-1}$ on $(\mathbb{R}, \mathcal{B})$ is Gaussian.*

Definition 3.4.3 (Gaussian Random Variable)**.** *A random variable Z from a probability space $(\Omega, \mathcal{F}, \mu)$ to $(\mathbb{R}^d, \mathcal{B})$ is said to be Gaussian if $\gamma := \mu \circ Z^{-1}$ is a Gaussian measure on $(\mathbb{R}^d, \mathcal{B})$.*

For vectors $u, v \in \mathbb{R}^d$, we define the inner product (dot product) to be $\langle u, v\rangle = \sum_{i=1}^d u_i v_i$. Note that the inner product is bilinear. We also define $|u|^2 := \langle u, v\rangle$. A collection of random variables $\{X_i\}_{i=1}^{\infty}$ is said to be independent and identically distributed (iid) if X_i and X_j are independent for all $i \neq j$ and the induced measures for each X_i coincide.

There are many ways to prove the Central Limit Theorem, Theorem 3.4.2 below. In these notes, we will use the standard approach based on convergence of the characteristic

[15] Consider sets of the form $I_k = \{k < X \leq X + 1\}$ to show this.

[16] Note that this m depends on $\omega \in \Omega$ as for each ω there exists an $m(\omega)$ such that $X_i(\omega) = Y_i(\omega)$ for all $i > m(\omega)$.

function. For a probability measure μ on $(\mathbb{R}^d, \mathcal{B})$, the characteristic function (Fourier transform) $\tilde{\mu} : \mathbb{R}^d \to \mathbb{C}$ is defined as

$$\tilde{\mu}(t) := \int \exp\{i \langle x, t \rangle\} \, d\mu(x).$$

We can also invert the above transformation. That is, if $\tilde{\mu}$ is integrable with respect to Lebesgue measure on $\mathbb{R}^d$, then

$$p(x) = (2\pi)^{-d} \int \tilde{\mu}(t) \exp\{-i \langle x, t \rangle\} \, d\lambda(t) \quad \lambda \, a.e.$$

where $p(x)$ is the probability density function for the measure μ.

Characteristic functions determine probability measures as discussed below in Theorem 3.4.1. First, we must define the *convolution*.

Definition 3.4.4 (Convolution)**.** *For two measures μ and ν on $(\mathbb{R}^d, \mathcal{B})$, the convolution measure is defined as*

$$(\mu * \nu)(B) := \int \nu(B - x) d\mu(x)$$

for any $B \in \mathcal{B}$ where $B - x = \{y \in \mathbb{R}^d : y + x \in B\}$.

Note that it can be shown that the convolution operation, $*$, is associative and commutative. Also, the characteristic function of $\mu * \nu$ is $\tilde{\mu}\tilde{\nu}$. Lastly, it can also be shown that for two independent random variables X and Y with corresponding measures μ and ν that the measure induced by $X + Y$ is $\mu * \nu$.

Theorem 3.4.1. *Let μ and ν be probability measures on $(\mathbb{R}^d, \mathcal{B})$. If $\tilde{\mu} = \tilde{\nu}$ then $\mu = \nu$.*

Proof. Let γ_σ be a mean zero Gaussian measure on $\mathbb{R}^d$ with variance $\sigma^2 I$. We denote $\mu^{(\sigma)} = \mu * \gamma_\sigma$ and similarly for $\nu^{(\sigma)}$. It can be shown that the corresponding density functions for $\mu^{(\sigma)}$ and $\nu^{(\sigma)}$ is

$$p^{(\sigma)}(x) = (2\pi)^{-d} \int \tilde{\mu}(t) \exp\left\{-i \langle t, x \rangle - \frac{1}{2}\sigma^2 |t|^2\right\} d\lambda(t)$$
$$q^{(\sigma)}(x) = (2\pi)^{-d} \int \tilde{\nu}(t) \exp\left\{-i \langle t, x \rangle - \frac{1}{2}\sigma^2 |t|^2\right\} d\lambda(t).$$

Thus, if $\tilde{\mu} = \tilde{\nu}$ then $\mu^{(\sigma)} = \nu^{(\sigma)}$ for all $\sigma > 0$.

Next, we consider the limit as $\sigma \downarrow 0$. Let X be a random variable corresponding to μ and Z to γ_1. Then, the measure μ^σ is paired with the random variable $X + \sigma Z$. Thus, $X + \sigma Z \xrightarrow{\text{a.s.}} X$, that is, pointwise for almost all ω. Thus, this convergence holds in probability and thus in distribution, i.e. $\mu^{(\sigma)} \Rightarrow \mu$ as $\sigma \downarrow 0$.

Lastly, we have that $\mu^{(\sigma)} \Rightarrow \mu$ and $\nu^{(\sigma)} \Rightarrow \nu$. Since the limit is unique $\mu = \nu$. □

Theorem 3.4.2 (Central Limit Theorem)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, $\{X_n\}_{n=1}^\infty$ be iid random variables on $(\mathbb{R}^d, \mathcal{B})$ such that $\mathrm{E}X_n = 0$ and $\mathrm{E}|X_n|^2 < \infty$. Let $S_n = \sum_{j=1}^n X_j$. Then, $n^{-1/2} S_n \xrightarrow{d} Z$ where Z is a Gaussian random variable with zero mean and covariance Σ with jkth entry $\Sigma_{jk} = \mathrm{E}[X_{nj} X_{nk}]$.*

Lemma 3.4.5. *For a uniformly tight sequence of probability measures μ_i on $\mathbb{R}^d$, if for all $v \in \mathbb{R}^d$ $\tilde{\mu}_i(v) \to \tilde{\mu}(v)$, then $\mu_i \Rightarrow \mu$ where μ is a measure with characteristic function $\tilde{\mu}$.*

Proof. As the μ_i are uniformly tight, Prohorov's Theorem says that every subsequence μ_{i_k} has a convergence subsubsequence $\mu_{i_{k_r}}$. But all of these subsubsequences have characteristic functions that all converge to $\tilde{\mu}$. By uniqueness of characteristic functions, all subsubsequences converge to the same measure μ. Thus, by Proposition 3.2.4, $\mu_i \Rightarrow \mu$. □

Proof of Theorem 3.4.2. As the random vectors X_j are mean zero and independent $\mathrm{E}\langle X_j, X_k\rangle = 0$, for $j \neq k$. In turn, for any n,

$$\mathrm{E}|n^{-1/2}S_n|^2 = n^{-1}\mathrm{E}\left(\sum_{j,k=1}^{n}\langle X_j, X_k\rangle\right) = \mathrm{E}|X_j|^2.$$

For any $\varepsilon > 0$, there exists an $M_\varepsilon > 0$ such that $\mathrm{E}|X_j|^2/M_\varepsilon^2 < \varepsilon$. Thus, from Chebyshev's inequality, we have that $\mathrm{P}\left(|n^{-1/2}S_n| > M_\varepsilon\right) < \varepsilon$. This implies that the sequence $n^{-1/2}S_n$ is uniformly tight.

For a vector $v \in \mathbb{R}^d$, the random variables $\langle v, X_j\rangle$ are iid real-valued with $\mathrm{E}\langle v, X_j\rangle = 0$ and $\mathrm{E}\langle v, X_j\rangle^2 < \infty$. Let $h(v) := \mathrm{E}\exp(i\langle v, X_j\rangle)$. Then, $h(0) = 1$ and $\nabla h(0) = 0$ and $\nabla^2 h(0) = -\Sigma$. Thus, by Taylor's Theorem, we have

$$h(v) = 1 - \frac{1}{2}v^{\mathrm{T}}\Sigma v + o(\|v\|_2^2).$$

Thus, for any fixed vector v,

$$\mathrm{E}\exp\left\{i\left\langle n^{-1/2}S_n, v\right\rangle\right\} = h(n^{-1/2}v)^n = \left(1 - \frac{v^{\mathrm{T}}\Sigma v}{2n} + o\left[\frac{\|v\|_2^2}{n}\right]\right)^n$$
$$\to \exp\left\{-\frac{1}{2}v^{\mathrm{T}}\Sigma v\right\}, \quad n \to \infty.$$

This limit is the characteristic function for Z. Thus, by applying the above lemma, we conclude that $n^{-1/2}S_n \xrightarrow{\mathrm{d}} Z$. □

3.5 Ergodic Theorem

In this section, we prove two Ergodic Theorems; Birkhoff's is for convergence almost everywhere much like the SLLN being for almost sure convergence; von Neumann's is for L^p convergence. These results are very powerful and imply the SLLN from the previous section. Roughly, the egodicity applies to dynamical systems and stochastic processes that uniformly visit an entire space.

We first require the notion of a measure-preserving map, invariance and ergodicity.

Definition 3.5.1 (Measure Preserving Map)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space A mapping $T : \Omega \to \Omega$ is called measure preserving if*

$$\mu(T^{-1}(A)) = \mu(A), \quad \text{for all } A \in \mathcal{F}.$$

Definition 3.5.2 (Invariant Set and Function)**.** *A set $A \in \mathcal{F}$ is T-invariant if $T^{-1}(A) = A$. The set of all T-invariant sets forms a σ-field $\mathcal{F}_T$.*[17]

A measurable function f is invariant if $f = f \circ T$. f is invariant if and only if f is $\mathcal{F}_T$-measurable.[18]

Definition 3.5.3 (Ergodic Map)**.** *A mapping T is said to be ergodic if for any $A \in \mathcal{F}_T$, we have*

$$\mu(A) = 0 \quad \text{or} \quad \mu(A^c) = 0.$$

For Lebesgue measure on $(0, 1]$, two examples of measure preserving maps are the shift map

$$T(x) = x + a \mod 1$$

and Baker's Map[19]

$$T(x) = 2x - \lfloor 2x \rfloor.$$

Furthermore, it can be shown that

- If f is integrable and T is measure preserving then $f \circ T$ is integrable and

$$\int f d\mu = \int f \circ T d\mu.$$

- If T is ergodic and f is invariant, then $f = c$ μ-a.e. for some constant c.

3.5.1 Birkhoff and von Neumann's Theorems

In what follows, we let $(\Omega, \mathcal{F}, \mu)$ be a measure space, T be a measure preserving transformation, $f : \Omega \to \mathbb{R}$ a measurable function, and

$$S_n = S_n(f) = f + f \circ T + \ldots + f \circ T^{n-1}$$

where $S_0 = 0$.

Lemma 3.5.4 (Maximal Ergodic Lemma)**.** *Let f be integrable and $S^* = \sup_{n \geq 0} S_n(f)$. Then,*

$$\int_{S^* > 0} f d\mu \geq 0.$$

[17] Check this!
[18] Check this!
[19] https://en.wikipedia.org/wiki/Baker%27s_map

Proof. Let $S_n^* = \max_{0\leq m\leq n} S_m(f)$ and $A_n = \{\omega \in \Omega : S_n^*(\omega) > 0\}$. Then, for $1 \leq m \leq n$,

$$S_m = f + S_{m-1} \circ T \leq f + S_n^* \circ T.$$

Furthermore, on the set A_n,

$$S_n^* = \max_{0\leq m\leq n} S_m(f) \leq f + S_n^* \circ T.$$

On the set A_n^c, $S_n^* = 0 \leq S_n^* \circ T$. Thus, integrating both sides of the above gives

$$\int_\Omega S_n^* d\mu \leq \int_{A_n} f d\mu + \int_\Omega S_n^* \circ T d\mu.$$

Since S_n^* is integrable and T is measure preserving, $\int S_n^* d\mu = \int S_n^* \circ T d\mu < \infty$. Therefore, $\int_{A_n} f d\mu \geq 0$. As $n \to \infty$, $A_n \uparrow \{S_n^* > 0\}$, we have that

$$\int_{S^*>0} f d\mu = \lim_{n\to 0} \int_{A_n} f d\mu \geq 0$$

due to dominated convergence with $|f|$ as the dominating function. □

Theorem 3.5.1 (Birkhoff's Ergodic Theorem)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a σ-finite measure space and $f \in L^1(\Omega, \mathcal{F}, \mu)$. Then, there exists an invariant $\bar{f} \in L^1(\Omega, \mathcal{F}, \mu)$ such that*

$$\int |\bar{f}| d\mu \leq \int |f| d\mu$$

and $n^{-1} S_n(f) \to \bar{f}$ as $n \to \infty$ μ-a.e.

Proof. Both $\liminf_{n\to\infty} n^{-1} S_n(f)$ and $\limsup_{n\to\infty} n^{-1} S_n(f)$ are T-invariant. Indeed, $n^{-1} S_n(f) \circ T = n^{-1}[S_{n+1}(f) - f] = [(n+1)/n](n+1)^{-1} S_{n+1}(f) - n^{-1} f$. Thus, we can define a set for $a < b$

$$D_{a,b} = \left\{ \omega \in \Omega : \liminf_{n\to\infty} \frac{S_n(f)}{n} < a < b < \limsup_{n\to\infty} \frac{S_n(f)}{n} \right\}$$

which means that the lim inf and lim sup are separated, and this set $D_{a,b}$ is T-invariant. The goal of the proof is to show that $\mu(D_{a,b}) = 0$. Without loss of generality, we take $b > 0$. Otherwise, $a < 0$ and we multiply everything by -1.

For some $B \in \mathcal{F}$ such that $\mu(B) < \infty$, then we set $g = f - b\mathbf{1}_B$. Function g is integrable and for each $x \in D_{a,b}$, there is an n such that $S_n(g)(x) \geq S_n(f)(x) - nb \geq 0$ since $b < \limsup_{n\to 0} n^{-1} S_n(f)$. Thus, $S^*(g) > 0$ and the maximal Ergodic lemma 3.5.4 says that

$$0 \leq \int_D (f - b\mathbf{1}_B) d\mu = \int_D f d\mu - b\mu(B).$$

As μ is σ-finite, there exist such a sequence of sets $B_n \in \mathcal{F}$ such that $B_n \uparrow D_{a,b}$ and $\mu(B_n) < \infty$ for all n. Thus,

$$b\mu(D_{a,b}) = \lim_{n\to\infty} b\mu(B_n) \leq \int_{D_{a,b}} f d\mu.$$

This implies that $\mu(D_{a,b}) < \infty$.

Redoing the above argument for $-a$ and $-f$ results in $-a\mu(D_{a,b}) \leq \int_{D_{a,b}}(-f)d\mu$. Therefore

$$b\mu(D_{a,b}) \leq \int_{D_{a,b}} f d\mu \leq a\mu(D_{a,b})$$

and since $a < b$, we have that $\mu(D_{a,b}) = 0$.

Next, let

$$E = \{\omega \in \Omega : \liminf_{n\to\infty} n^{-1}S_n(f) < \limsup_{n\to\infty} n^{-1}S_n(f)\}.$$

Then, E is T-invariant as the lim inf and lim sup are. Furthermore, $E = \bigcup_{a,b\in\mathbb{Q},a<b} D_{a,b}$. Thus, $\mu(E) = 0$.

This means that $n^{-1}S_n(f)$ converges in $[-\infty, \infty]$ on E^c. Therefore, we define

$$\bar{f} := \begin{cases} \lim_{n\to\infty} n^{-1}S_n(f) & \omega \in E^c \\ 0 & \omega \in E \end{cases}.$$

Lastly, $\int |f \circ T^n| d\mu = \int |f| d\mu$ and thus $\int |S_n(f)| d\mu \leq n \int |f| d\mu$ for all n. Applying Fatou's lemma 2.3.5 gives

$$\int |\bar{f}| d\mu = \int \liminf_{n\to\infty} |n^{-1}S_n(f)| d\mu \leq \liminf_{n\to\infty} \int |n^{-1}S_n(f)| d\mu \leq \int |f| d\mu$$

finishing the proof. □

Theorem 3.5.2 (von Neumann's Ergodic Theorem)**.** *Let $\mu(\Omega) < \infty$ and $p \in [1, \infty)$. Then, for all $f \in L^p(\Omega, \mathcal{F}, \mu)$, there exists an $\bar{f} \in L^p$ such that $n^{-1}S_n(f) \to \bar{f}$ in L^p.*

Proof. We begin by noting that

$$\|f \circ T^n\|_p^p = \int |f|^p \circ T^n d\mu = \|f\|_p^p.$$

By the above and Minkowski's inequality, $\|n^{-1}S_n(f)\|_p \leq \|f\|_p$. Since $f \in L^p$, given a $\varepsilon > 0$, we can choose a $C > 0$ such that $\|f - g\|_p < \varepsilon/3$ with $g = \min[\max\{-C, f\}, C]$, i.e. g is f bounded above and below by C and $-C$. By Birkhoff's Theorem 3.5.1, $n^{-1}S_n(g) \to \bar{g}$ μ-a.e.

Next, we note that $|n^{-1}S_n(g)| \leq C$ for all n, and thus by dominated convergence[20] 2.3.6 there exists an N such that for all $n > N$,

$$\|n^{-1}S_n(g) - \bar{g}\|_p < \varepsilon/3.$$

Applying Fatou's Lemma 2.3.5 gives that

$$\|\bar{f} - \bar{g}\|_p^p = \int \liminf_{n\to\infty} |n^{-1}S_n(f-g)|^p d\mu \leq \liminf_{n\to\infty} \int |n^{-1}S_n(f-g)|^p d\mu = \|f - g\|_p^p.$$

Thus, for $n > N$,

$$\|n^{-1}S_n(f) - \bar{f}\|_p \leq \|n^{-1}S_n(f-g)\|_p + \|n^{-1}S_n(g) - \bar{g}\|_p + \|\bar{g} - \bar{f}\|_p < \varepsilon.$$

□

[20] Recall, $\mu(\Omega) < \infty$!

3.5.2 Law of Large Numbers, again

Let $(\Omega, \mathcal{F}, P)$ be a probability space with iid real-valued random variables $\{X_i\}_{i=1}^{\infty}$ with distribution function dF. Let $(S, \mathcal{S})$ be a measurable space with $S = \mathbb{R}^{\mathbb{N}}$, a countably infinite product of $\mathbb{R}$, and $\mathcal{S}$ generated by the π-system

$$\mathcal{A} = \left\{ \prod_{n\in\mathbb{N}} A_n \,:\, A_n \in \mathcal{B} \;\; \forall\, n, A_m = \mathbb{R} \text{ eventually} \right\}.$$

Let the random variable $X : \Omega \to \mathbb{R}^{\mathbb{N}}$ to be $X(\omega) = (X_1(\omega), X_2(\omega), \ldots)$. Then, X induces the measure

$$\mu(A) = P \circ X^{-1}(A) = \prod_{n\in\mathbb{N}} dF(A_n)$$

for $A = \prod A_n$. For a sequence $(x_1, x_2, \ldots)$, we can define the shift map $T : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ to be

$$T(x_1, x_2, \ldots) = (x_2, x_3, \ldots).$$

Proposition 3.5.5. *The shift map is measure preserving and ergodic.*

If you want to show this yourself, consider Kolmogorov's zero-one law.[21]

Theorem 3.5.3 (Strong Law of Large Numbers, again)**.** *Let $\{X_i\}_{i=1}^{\infty}$ be iid random variables from Ω to $\mathbb{R}$. If $\mathrm{E}|X_i| < \infty$, then $S_n/n \xrightarrow{a.s.} c$ for $c = \mathrm{E}X_i$.*

Proof. Let $f : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ by *taking the first coordinate*, that is, $f(X_1, X_2, \ldots) = X_1$. Then, for T being the shift map,

$$S_n = f + f \circ T + \ldots + f \circ T^{n-1} = X_1 + \ldots + X_n.$$

Thus, Birkhoff's Ergodic Theorem 3.5.1 says that there exists an invariant $\bar{f} \in L^1$ such that

$$n^{-1} S_n \to \bar{f} \text{ a.s.}$$

Since T is ergodic, the result from the beginning of this section states that $\bar{f} = c$, a constant, almost surely. Lastly, using von Neumann's Ergodic Theorem 3.5.2 with $p = 1$,

$$c = \int \bar{f} d\mu = \lim_{n\to\infty} \int n^{-1} S_n(f) d\mu = \mathrm{E}X_i.$$

□

[21] https://en.wikipedia.org/wiki/Kolmogorov%27s_zero%E2%80%93one_law

Bibliography

[Billingsley(2008)] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[Billingsley(2013)] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[Dudley(2002)] Richard M Dudley. *Real analysis and probability*, volume 74. Cambridge University Press, 2002.