Probability and Measure *

Zhiyuan Bai

Compiled on June 1, 2022

This document serves as a set of revision materials for the Cambridge Mathematical Tripos Part II course *Probability and Measure* in Michaelmas 2020. However, despite its primary focus, readers should note that it is NOT a verbatim recall of the lectures, since the author might have made further amendments in the content. Therefore, there should always be provisions for errors and typos while this material is being used.

Contents

*Based on the lectures under the same name taught by Prof. R. Nickl in Michaelmas 2020.

0 Introduction

There are three aims of this course:
Firstly, we want to replace the Riemann integeral by the more powerful and general Lebesgue integral. Why do we want to do it? The space of "integrable functions $[0,1] \to \mathbb{R}$" should naturally be the (topological) completion of the normed space $(C([0,1]), \|\cdot\|_1)$ where

$$\|f\|_1 = \int_0^1 |f(x)|\,\mathrm{d}x$$

It turns out, this space is not quite the space of Riemann integrable functions on $[0,1]$ – it, in fact, equals the space L^1 of Lebesgue integrable functions. We can use instead the square norm

$$\|f\|_2 = \sqrt{\int_0^1 |f(x)|^2\,\mathrm{d}x}$$

which will give us a corresponding space L^2 that turns out to be a Hilbert space.
The second motivation is to give a proper axiomisation of probability theory, as derived from the work of Kolmogorov. Measure theory will allow us to derive key results such that the Law(s) of Large Numbers, Central Limit Theorem, Ergodic Theorem, etc., from a perfectly rigorous foundation.
The third motivation is the problem of measure: Does every subset of $\mathbb{R}^d$ have a volume? The answer to it is, sadly or not, no. Characterising those subsets who do have volumes is then an interesting topic to discuss.

1 Measures

1.1 Measures and Measurable Sets

Definition 1.1. Let E be any set and let 2^E be the family of all subsets of E. A σ-algebra $\mathcal{E}$ on E is a subset of 2^E such that:
1. $\varnothing \in \mathcal{E}$.
2. $A \in \mathcal{E} \implies A^c = E - A \in \mathcal{E}$.
3. $(A_n)_n \in \mathcal{E} \implies \bigcup_n A_n \in \mathcal{E}$.
We call $(E, \mathcal{E})$ a measurable space. Elements $A \in \mathcal{E}$ are called measurable set in E.

We immediately have $(A_n)_n \in \mathcal{E} \implies \bigcap_n A_n = (\bigcup_n A_n^c)^c \in \mathcal{E}$ and also $A, B \in \mathcal{E} \implies B - A = B \cap A^c \in \mathcal{E}$.

Definition 1.2. A measure μ on $(E, \mathcal{E})$ is a function $\mu : \mathcal{E} \to [0, \infty]$ such that:
1. $\mu(\varnothing) = 0$.
2. If $(A_n)_n \in \mathcal{E}$ are disjoint, then

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

When the measure μ is identified, the triple $(E, \mathcal{E}, \mu)$ is called a measure space.

Remark. Here we use the convention that $\infty + x = \infty$ for any $x \in [0, \infty]$ and $\sum_n x_n = \infty$ (for $x_n \in [0, \infty]$) if either $x_m = \infty$ for some m or $\sum_n x_n$ diverges as a sequence in $[0, \infty)$.

Remark. When E is countable, any measure μ on $(E, 2^E)$ necessarily have

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

So μ corresponds to a "mass function" $m : E \to [0, \infty]$ by $m(x) = \mu(\{x\})$.
When E is not countable, we generally need to work with σ-algebras $\mathcal{E} \subsetneq 2^E$.

Definition 1.3. For any collection $\mathscr{A} \subset 2^E$, we define

$$\sigma(\mathscr{A}) = \{A \subset E : A \in \mathcal{E} \text{ for all } \sigma\text{-algebra } \mathcal{E} \supset \mathscr{A}\} = \bigcap_{\mathcal{E}\ \sigma\text{-algebra containing } \mathscr{A}} \mathcal{E}$$

to be the σ-algebra generated by $\mathscr{A}$.

It is easy to see that $\sigma(\mathscr{A})$ is indeed a σ-algebra.
We often choose the collection $\mathscr{A} \subset 2^E$ to be one that already has some nice properties, for example:

Definition 1.4. A collection $\mathscr{A} \subset 2^E$ is:
1. A ring if $\varnothing \in \mathscr{A}$ and $A, B \in \mathscr{A} \implies B - A, A \cup B \in \mathscr{A}$.
2. An algebra if $\varnothing \in \mathscr{A}$ and $A, B \in \mathscr{A} \implies A^c, A \cup B \in \mathscr{A}$.

Remark. Since $A \cap B = (A \cup B) - (A \Delta B)$ where $A \Delta B = (A - B) \cup (B - A)$, we see that $A \cap B \in \mathscr{A}$ whenever A, B belong to the ring $\mathscr{A}$.

1.2 Carathéodory's Extension Theorem

Definition 1.5. Suppose $\mathscr{A} \subset 2^E$ contains $\varnothing$. A function $\mu : \mathscr{A} \to [0, \infty]$ is called a set function if $\mu(\varnothing) = 0$.
We say μ is:
1. Increasing if $\forall A, B \in \mathscr{A}$ with $A \subset B$, we have $\mu(A) \leq \mu(B)$.
2. Additive if $\forall A, B \in \mathscr{A}$ with $A \cap B = \varnothing$, we have $\mu(A \cup B) = \mu(A) + \mu(B))$ whenever $A \cup B \in \mathscr{A}$.
3. Countably additive if $\forall (A_n)_n \in \mathscr{A}$ disjoint, we have

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

whenever $\bigcup_n A_n \in \mathscr{A}$.
4. Countably subadditive if $\forall (A_n)_n \in \mathscr{A}$ we have

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

whenever $\bigcup_n A_n \in \mathscr{A}$.

Remark. A countably additive set function on a σ-algebra is a measure. Conversely, using the trick of disjointification, one can show that a countably additive set function on a ring $\mathscr{A}$ is also additive, increasing and countably subadditive.
What is disjointification? Suppose $(A_n)_n \in \mathscr{A}$ with $\mathscr{A}$ a ring, we can write $\tilde{A}_n = \bigcup_{j=1}^n A_j$ So $\tilde{A}_n$ is an increasing sequence of sets. Define $B_1 = \tilde{A}_1, B_2 = \tilde{A}_2 - \tilde{A}_1, \ldots, B_n = \tilde{A}_n - \tilde{A}_{n-1}$, then $\bigcup_n A_n = \bigcup_n B_n$ and $B_n \in \mathscr{A}$. This trick would be quite useful in many other set operation contexts as well.

Theorem 1.1 (Carathéodory)**.** *Let $\mathscr{A} \subset 2^E$ be a ring and let $\mu : \mathscr{A} \to [0, \infty]$ be a countably additive set function, then μ extends to a measure on $\sigma(\mathscr{A})$.*

Proof. For $B \subset E$, we define the outer measure of B to be

$$\mu^*(B) = \inf\left\{\sum_n \mu(A_n) : A_n \in \mathscr{A}, B \subset \bigcup_n A_n\right\}$$

with the convention that $\inf \varnothing = \infty$. Clearly $\mu^*(\varnothing) = 0$ and μ^* is increasing on 2^E. Say a subset $A \subset E$ is μ^*-measurable if for any $B \subset E$ we have $\mu^*(B) = \mu^*(B \cap A) + \mu^*(B \cap A^c)$. Let $\mathcal{M}$ be the collection of μ^*-measurable subsets of E.
The plan is to show that $\mathcal{M}$ is a σ-algebra on which μ^* is a measure that extends μ, and restrict μ^* to $\sigma(\mathscr{A})$.
Step 1: We shall show that μ^* is countably subadditive. For any $B_n, B \subset E$ with $B \subset \bigcup_n B_n$, we need to show that $\mu^*(B) \leq \sum_n \mu^*(B_n)$. If $\mu^*(B_n) = \infty$ for some n then we are done. Assume this is not the case, then for any n and all $\epsilon > 0$, there is some $(A_{n,m})_{n,m} \in \mathscr{A}$ such that $B_n \subset \bigcup_m A_{n,m}$ and

$$\mu^*(B_n) + \frac{\epsilon}{2^n} \geq \sum_m \mu(A_{n,m})$$

Then $B \subset \bigcup_n B_n \subset \bigcup_{n,m} A_{n,m}$. We know that μ^* is increasing, so

$$\begin{aligned}\mu^*(B) &\leq \mu^*\left(\bigcup_{n,m} A_{n,m}\right) \leq \sum_{n,m} \mu(A_{n,m}) \\ &\leq \sum_n \mu^*(B_n) + \epsilon \sum_n 2^{-n} = \sum_n \mu^*(B_n) + \epsilon\end{aligned}$$

But ϵ is arbitrary, so $\mu^*(B) \leq \sum_n \mu^*(B_n)$.
Step 2: We shall show that μ^* extends μ. Take $A \in \mathscr{A}$. Say $A \subset \bigcup_n A_n$ for some $(A_n)_n \in \mathscr{A}$, then we can write $A = \bigcup_n (A \cup A_n)$. μ is countable subadditive and increasing since it is a countable additive set function, so

$$\mu(A) = \mu\left(\bigcup_n (A \cap A_n)\right) \leq \sum_n \mu(A_n)$$

Taking infimum shows that $\mu(A) \leq \mu^*(A)$. Conversely, $\mu^*(A) \leq \mu(A)$ by taking $A_1 = A, A_n = \varnothing$ for $n > 1$, so $\mu^* = \mu$ on $\mathscr{A}$.
Step 3: We shall show that $\mathcal{M} \supset \mathscr{A}$. Pick $A \in \mathscr{A}, B \subset E$, we shall show that $\mu^*(B) = \mu^*(B \cap A) + \mu^*(B \cap A^c)$. Countable subadditivity gives $\mu^*(B) \leq \mu^*(B \cap A) + \mu^*(B \cap A^c)$. Conversely, if $\mu^*(B) < \infty$ then we are done. Otherwise, for any $\epsilon > 0$, we can find $A_n \in \mathscr{A}$ such that $B \subset \bigcup_n A_n$ and

$$\mu^*(B) + \epsilon \geq \sum_n \mu(A_n)$$

Now $B \cap A \subset \bigcup_n (A_n \cap A), B \cap A^c \subset \bigcup_n (A_n \cap A^c)$, so (noting that $A_n \cap A^c = A_n - A \in \mathscr{A}$)

$$\begin{aligned} \mu^*(B \cap A) + \mu^*(B \cap A^c) &\leq \sum_n \mu(A_n \cap A) + \sum_n \mu(A_n \cap A^c) \\ &= \sum_n \mu(A_n) \leq \mu^*(B) + \epsilon \end{aligned}$$

Again ϵ is arbitrary so $\mu^*(B \cap A) + \mu^*(B \cap A^c) \leq \mu^*(B)$, hence $\mu^*(B \cap A) + \mu^*(B \cap A^c) = \mu^*(B)$.
Step 4: We shall show that $\mathcal{M}$ is an algebra. Clearly $\varnothing \in \mathcal{M}$ and $A \in M \implies A^c \in \mathcal{M}$. Suppose $A_1, A_2 \in \mathcal{M}$ and $B \subset E$, then

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

So $A_1 \cap A_2 \in \mathcal{M}$ which is sufficient to imply that $\mathcal{M}$ is an algebra.
Step 5: We shall show that $\mathcal{M}$ is a σ-algebra. For any $(A_n)_n \in \mathcal{M}$ disjoint, we want to show that $A = \bigcup_{n=1}^{\infty} A_n$ is an element of $\mathcal{M}$. For any $B \subset E$, we have $B \subset (B \cap A) \cup (B \cap A^c)$, so $\mu^*(B) \leq \mu^*(B \cap A) + \mu^*(B \cap A^c)$. Conversely, we write

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

For any $N \geq 1$. Taking $N \to \infty$ gives

$$\mu^*(B) \geq \sum_{n=1}^{\infty} \mu^*(B \cap A_n) + \mu^*(B \cap A^c) \geq \mu^*(B \cap A) + \mu^*(B \cap A^c)$$

by countable subadditivity. Consequently $\mu^*(B) = \mu^*(B \cap A) + \mu^*(B \cap A^c)$, so $A \in \mathcal{M}$, so $\mathcal{M}$ is a σ-algebra.

Step 6: Finally, we shall show that μ^* is a measure on $\mathcal{M}$. Let $A = \bigcup_n A_n$ with $(A_n)_n \in \mathcal{M}$, we have

$$\mu^*(A) \geq \sum_{n=1}^{\infty} \mu^*(A \cap A_n) + \mu^*(A \cap A^c) = \sum_{n=1}^{\infty} \mu^*(A_n)$$

using the inequality obtained in the last step. Combining this with countable subadditivity of μ^* shows that μ^* is countably additive on $\mathcal{M}$, which completes the proof. □

How about uniqueness?

Definition 1.6. A collection $\mathscr{A} \subset 2^E$ is:
1. A π-system if $\varnothing \in \mathscr{A}$ and $A, B \in \mathscr{A} \implies A \cap B \in \mathscr{A}$.
2. A d-system if $E \in \mathscr{A}$ and for any $A, B \in \mathscr{A}$ with $A \subset B$ and any increasing $(A_n)_n \in \mathscr{A}$, one has $B - A \in \mathscr{A}, \bigcup_n A_n \in \mathscr{A}$.

Remark. 1. One can show (on example sheet) that $\mathscr{A}$ is a σ-algebra iff $\mathscr{A}$ is both a π-system and a d-system.
2. By disjointification, $\mathscr{A} \subset 2^E$ is a d-system iff $E \in \mathscr{A}$ and for any $A, B \in \mathscr{A}$ with $A \subset B$ and any disjoint (as opposed to increasing) $(A_n)_n \in \mathscr{A}$, one has $B - A \in \mathscr{A}, \bigcup_n A_n \in \mathscr{A}$.

Lemma 1.2 (Dynkin)**.** *Let $\mathscr{A}$ be a π-system, then any d-system containing $\mathscr{A}$ also contains $\sigma(\mathscr{A})$.*

Before proving this lemma, let's see what it gets us.

Theorem 1.3 (Uniqueness of Extension)**.** *Let μ_1, μ_2 be measures on $(E, \mathcal{E})$ such that $\mu_1(E) = \mu_2(E) < \infty$. Suppose $\mu_1 = \mu_2$ on $\mathscr{A}$ where $\mathscr{A}$ is some π-system generating $\mathcal{E}$, then $\mu_1 = \mu_2$ on $\mathcal{E}$.*

Proof. Consider the family $\mathcal{D} = \{A \in \mathcal{E} : \mu_1(A) = \mu_2(A)\}$ of measurable subsets of E where $\mu_1 = \mu_2$ coincide. We know that $\mathscr{A} \subset \mathcal{D}$. We will show that $\mathcal{D} \supset \sigma(\mathscr{A})$ which implies the theorem. Clearly $E \in \mathcal{D}$. Also, if $A, B \in \mathcal{E}, A \subset B$, then $\mu_i(A) + \mu_i(B - A) = \mu_i(B) \leq \mu_i(E) < \infty$ for $i = 1, 2$, so $A, B \in \mathcal{D} \implies B - A \in \mathcal{D}$. If $A = \bigcup_n A_n$ with $(A_n)_n \in \mathcal{D}$ disjoint, then

$$\mu_1(A) = \mu_1\left(\bigcup_n A_n\right) = \lim_{N\to\infty} \sum_{n=1}^{N} \mu_1(A_n) = \lim_{N\to\infty} \sum_{n=1}^{N} \mu_2(A_n) = \mu_2(A)$$

Thus $\mathcal{D}$ is a d-system containing the π-system $\mathscr{A}$. By Dynkin's lemma, $D \supset \sigma(\mathscr{A})$, so $\mu_1 = \mu_2$ on $\mathcal{E} = \sigma(\mathscr{A})$. □

Proof of Lemma 1.2. Define

$$\mathcal{D} = \bigcap_{\mathcal{D}_\alpha \ d\text{-system containing } \mathscr{A}} \mathcal{D}_\alpha$$

Then $\mathcal{D}$ is a d-system. We shall show that it is also a π-system, whch implies the result.
Define a new collection of sets $\mathcal{D}' = \{B \subset \mathcal{D} : \forall A \in \mathscr{A}, B \cap A \in \mathcal{D}\}$. Then $\mathscr{A} \subset \mathcal{D}'$ since $\mathscr{A}$ is a π-system. Also, $\mathcal{D}'$ is a d-system: Indeed, $E \cap A = A \in \mathscr{A}$

whenever $A \in \mathscr{A}$, and if $B_1, B_2 \in \mathcal{D}'$ are such that $B_1 \subset B_2$, then for any $A \in \mathscr{A}$ we have $(B_2 - B_1) \cap A = (B_2 \cap A) - (B_1 \cap A) \in \mathcal{D}$ since $\mathcal{D}$ is a d-system. So $B_2 - B_1 \in \mathcal{D}'$. Finally, if $B_n \uparrow B$ (i.e. $(B_n)_n$ is increasing and $\bigcup_n B_n = B$) with $(B_n)_n \in \mathcal{D}'$, then for any $A \in \mathscr{A}$ we have $B_n \cap A \uparrow B \cap A \in \mathcal{D} \implies B \in \mathcal{D}'$ since $\mathcal{D}$ is a d-system. We conclude that $\mathcal{D}'$ is a d-system containing the π-system $\mathscr{A}$, so necessarily $\mathcal{D} \subset \mathcal{D}'$. But $\mathcal{D}' \subset \mathcal{D}$ by definition, so $\mathcal{D}' = \mathcal{D}$, i.e. $B \cap A \in \mathcal{D}$ for all $A \in \mathscr{A}$.
Let $\mathcal{D}'' = \{B \in \mathcal{D} : \forall A \in \mathcal{D}, B \cap A \in \mathcal{D}\}$ and repeating essentially the same argument gives $D'' = D$, so $\mathcal{D}$ is a π-system, which completes the proof. □

Definition 1.7. A measure μ on $(E, \mathcal{E})$ is finite if $\mu(E) < \infty$, and σ-finite if E can be covered by a countable collection $(E_n)_n \in \mathcal{E}$ with $\mu(E_n) < \infty$ for all n.

Remark. The uniqueness theorem we proved applies to finite measures, but it quite easily generalises to σ-finite measures using techniques that we will use later when defining Lebesgue measure.

1.3 The Lebesgue Measure

Definition 1.8. Let E be a topological space, then the σ-algebra generated by the open sets of E is called the Borel σ-algebra $\mathcal{B}(E)$ on E.
A measure μ on $(E, \mathcal{B}(E))$ is called a Borel measure.

When $E = \mathbb{R}$ with the usual topology, we sometimes abbreviate $\mathcal{B} = \mathcal{B}(\mathbb{R})$.

Proposition 1.4. *There exists a unique Borel measure μ on $(\mathbb{R}, \mathcal{B})$ such that $\mu((a, b]) = b - a$.*

Proof. Consider the ring $\mathscr{A}$ of finite union of disjoint half-open intervals $(a_1, b_1] \sqcup \cdots \sqcup (a_n, b_n]$ which generates $\sigma(\{(a, b) : b < a\})$ (example sheet) which equals $\mathcal{B}$ as any open set in $\mathbb{R}$ is a countable union of open intervals. Consider the set function

$$\mu((a_1, b_1] \sqcup \cdots \sqcup (a_n, b_n]) = \sum_{i=1}^{n} (b_i - a_i)$$

This is obviously well-defined and finitely additive. Indeed μ is countably additive: In example sheet, we have shown that a finitely additive finite-valued set function ν on a ring $\mathscr{A}$ is countably additive iff $\nu(A_n) \to 0$ for any sequence $(A_n)_n$ in $\mathscr{A}$ decreasing to $\varnothing$. Suppose there is some $B_n \in \mathscr{A}$ decreasing to $\varnothing$ is such that $\mu(B_n)$ does not tend to 0, then there is some $\epsilon > 0$ such that $\mu(B_n) \geq 2\epsilon$ for all n (after possibly passing to a subsequence). Choose $C_n \in \mathscr{A}$ such that $\bar{C}_n \subset B_n$ and $\mu(B_n - C_n) \leq \epsilon 2^{-u}$ (e.g. if $B_n = (a_1, b_1] \sqcup \cdots \sqcup (a_r, b_r]$ then take $C_n = (a_1 + \epsilon 2^{-u}/r, b_1] \sqcup \cdots \sqcup (a_r + \epsilon 2^{-u}/r, b_r]$). Now $B_n - (C_1 \cap \cdots \cap C_n) \subset (B_1 - C_1) \cup \cdots \cup (B_n - C_n)$, so by finite additivity (hence finite subadditivity).

$$\begin{aligned} \mu(B_n - (C_1 \cap \cdots \cap C_n)) &\leq \mu((B_1 - C_1) \cup \cdots \cup (B_n - C_n)) \\ &\leq \sum_{i \leq n} \mu(B_i - C_i) \leq \epsilon \sum_i 2^{-i} \leq \epsilon \end{aligned}$$

Consequently, $\mu(C_1 \cap \cdots \cap C_n) \geq \mu(B_n) - \mu(B_n - (C_1 \cap \cdots \cap C_n)) \geq \epsilon$, in particular $C_1 \cap \cdots \cap C_n \neq \varnothing$. Hence $\bar{C}_1 \cap \cdots \cap \bar{C}_n = K_n$ is nonempty. K_n is a

decreasing sequence of compact sets, therefore

$$\varnothing \neq \bigcap_n K_n \subset \bigcap_n B_n = \varnothing$$

Contradiction. Therefore μ has to be countably additive, therefore extends to a Borel measure on $\sigma(\mathscr{A}) = \mathcal{B}$.
As for uniqueness, let λ be any other measure on $\mathcal{B}$ such that $\lambda((a,b]) = b - a$. For any $n \in \mathbb{N}$, consider the measures $\mu_n, \lambda_n : \mathcal{B} \to [0,\infty]$ defined by $\mu_n(A) = \mu(A \cap (n, n+1]), \lambda_n = \lambda(A \cap (n, n+1])$ which are finite measures that coincide on $\mathscr{A}$ which is in particular a π-system. Applying Theorem 1.3 on λ_n versus μ_n shows that $\lambda_n = \mu_n$ on $\mathcal{B}$. But any $A \in \mathcal{B}$ can be decomposed into the disjoint measurable sets $A = \bigcup_{n\in\mathbb{Z}}(A \cap (n, n+1])$, so $\mu = \lambda$ on $\mathcal{B}$. □

Definition 1.9. This unique measure on $\mathcal{B}$ is called the Lebesgue measure on $(\mathbb{R}, \mathcal{B})$.

Remark. 1. A set $A \in \mathcal{B}$ is called (Lebesgue) null if $\mu(A) = 0$. Singletons (hence in general any countable subsets of $\mathbb{R}$) are examples of this.
2. We defined the measure μ on $\mathcal{B}$, but of course the proof of Theorem 1.1 tells us that we can define μ instead on a slightly bigger σ-algebra $\mathcal{M}$ consists of "outer Lebesgue measurable sets". In example sheet, we have shown that $\mathcal{M}$ consists of subsets of the form $A \cup N$ for some $A \in \mathcal{B}, N \subset B \in \mathcal{B}$ with $\mu(B) = 0$. So $\mathcal{M}$ can be obtained from adding every subset of null sets to $\mathcal{B}$.
3. Lebesgue measure is translation-invariant: Given any $B \in \mathcal{B}$ (or indeed $\mathcal{M}$), we have $\mu(B + x) = \mu(B)$ for any $x \in \mathbb{R}$ (this can be proved, e.g. by the uniqueness theorem).

One might ask whether any or all of the inclusions $\mathcal{B} \subset \mathcal{M} \subset 2^{\mathbb{R}}$ are strict. Like it or not, none of $\mathcal{B}$ and $\mathcal{M}$ equals $2^{\mathbb{R}}$. For $x, y \in (0,1]$, write $x \sim y$ if $(x - y) \bmod 1 \in \mathbb{Q} \cap (0,1]$. Using the axiom of choice, we can collect a set of representatives S for the equivalence classes under this equivalence relation. Write $S + q = \{s + q : s \in S\}$, then $(0,1] = \bigcup_{q\in\mathbb{Q}\cap(0,1]}(S + q)$ is a disjoint union of cosets. Suppose $\mathcal{M} = 2^{\mathbb{R}}$, then everything is Lebesgue measurable. By translation invariance of μ, we have $\mu(S + q) = \mu(S)$, but then

$$\sum_{q\in\mathbb{Q}\cap(0,1]} \mu(S) = \sum_{q\in\mathbb{Q}\cap(0,1]} \mu(S+q) = \mu\left(\bigcup_{q\in\mathbb{Q}\cap(0,1]} (S+q)\right) = \mu((0,1]) = 1$$

Contradiction. Therefore $\mathcal{M} \neq 2^{\mathbb{R}}$.
What if the axiom of choice is not assumed? Turns out, the existence of non-Lebesgue measurable sets cannot be proved nor disproved without the axiom of choice.
But is there any way at all to find a measure on $(\mathbb{R}, 2^{\mathbb{R}})$ that is remotely sensible?

Theorem 1.5 (Banach-Kuratowski)**.** *Assume the continuum hypothesis, then there exists no measure μ on $2^{(0,1]}$ such that $\mu((0,1]) = 1$ and $\mu(\{x\}) = 0$ for any $x \in (0,1]$.*

1.4 Probability Measures

Definition 1.10. Let $(E, \mathcal{E}, \mu)$ be a measure space. It is called a probability space if $\mu(E) = 1$.

We normally use the symbols $(\Omega, \mathcal{F}, \mathbb{P})$ to denote a probability space, where Ω is taken as the sample space, $\mathcal{F}$ the space of events and $\mathbb{P}$ the probability. The definitive properties of measures translates to the axioms of probability theory (Kolmogorov, 1933), namely:
1. $\mathbb{P}(\Omega) = 1$.
2. $\mathbb{P}(E) \geq 0$.
3. For any disjoint $(A_n)_n \in \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_n A_n\right) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$$

But if we want to do probability, it's hard to get pass the concept of independence.

Definition 1.11. The events $(A_n)_n \in \mathcal{F}$ are independent if for any finite set J of indices we have

$$\mathbb{P}\left(\bigcap_{i\in J} A_i\right) = \prod_{i\in J} \mathbb{P}(A_i)$$

Definition 1.12. Sub-σ-algebras $(\mathcal{F}_i)_i \subset \mathcal{F}$ are said to be independent if for any finite set J of indices, any choice of events $A_i \in \mathcal{F}_i$ for $i \in J$ are independent.

Theorem 1.6. *Let $\mathscr{A}_1, \mathscr{A}_2$ be π-systems in $\mathcal{F}$ such that any $A_1 \in \mathscr{A}_1, A_2 \in \mathscr{A}_2$ has $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$, then $\sigma(\mathscr{A}_1)$ and $\sigma(\mathscr{A}_2)$ are independent.*

Proof. Fix $A_1 \in \mathscr{A}_1$, define $\mu(A) = \mathbb{P}(A_1 \cap A), \vartheta(A) = \mathbb{P}(A_1)\mathbb{P}(A)$ which are finite measures with $\mu(\Omega) = \mathbb{P}(A_1) = \vartheta(\Omega)$. They also coincide on the π-system $\mathscr{A}_2$, hence on $\sigma(A_2)$ by Theorem 1.3. Repeat the argument the other way around completes the proof. □

To prove asympototic results in probability, it is useful to consider

Definition 1.13.

$$\limsup_n A_n = \bigcap_n \bigcup_{m\geq n} A_m = \{A_n \text{ infinitely often}\} = \{A_n \text{ i.o.}\}$$

$$\liminf_n A_n = \bigcup_n \bigcap_{m\geq n} A_m = \{A_n \text{ eventually}\}$$

Lemma 1.7 (First Borel-Cantelli Lemma)**.** *If $\sum_n \mathbb{P}(A_n) < \infty$, then we have $\mathbb{P}(A_n \text{ i.o.}) = 0$.*

Proof.

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

as $n \to \infty$. □

Lemma 1.8 (Second Borel-Cantelli Lemma)**.** *Suppose $(A_n)_n \in \mathcal{F}$ are independent. If $\sum_n \mathbb{P}(A_n) = \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 1$.*

Proof. We shall use the fact that $1 - a \leq e^{-a}$ for all a. Take $a = a_n = \mathbb{P}(A_n)$. For any $n, N \geq n$, we have (by independence of A_n's hence A_n^c's),

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

as $N \to \infty$. We know that $\bigcap_{m=n}^{N} A_m^c$ decreases to $\bigcap_{m=n}^{\infty} A_m^c$ as $N \to \infty$, therefore $\mathbb{P}\left(\bigcap_{m\geq n} A_m^c\right) = 0$ for any n, consequently,

$$\mathbb{P}(A_n \text{ i.o.}) = 1 - \mathbb{P}\left(\bigcup_n \bigcap_{m\geq n} A_m^c\right) = 1 - 0 = 1$$

as desired. □

2 Measurable Functions and Random Variables

2.1 Measurable Functions; Monotone Class Theorem

Definition 2.1. Let $(E, \mathcal{E}), (G, \mathcal{G})$ be measurable spaces. A function $f : E \to G$ is $(\mathcal{E}, \mathcal{G})$-measurable if $f^{-1}(A) \in \mathcal{E}$ whenever $A \in \mathcal{G}$. When $(G, \mathcal{G}) = (\mathbb{R}, \mathcal{B})$, we simply say f is measurable. If in addition E is a topological space and $\mathcal{E} = \mathcal{B}(E)$, we say f is Borel measurable.

Most of the time the context is understood and we will just use "measurable" and not bother too much with the slightly confusing terminology.
Note that the inverse image preserve set operations

$$f^{-1}\left(\bigcup_i A_i\right) = \bigcup_i f^{-1}(A_i), f^{-1}(G - A) = E - f^{-1}(A)$$

So indeed for any $f : E \to G$, $\{f^{-1}(A) : A \in \mathcal{G}\}$ would itself be a σ-algebra. Likewise, $\{A \subset G : f^{-1}(A) \in \mathcal{E}\}$ is also a σ-algebra. In particular, we can use the following idea to check measurability: If $\mathcal{G} = \sigma(\mathscr{A})$ and $f^{-1}(A) \in \mathcal{E}$ for all $A \in \mathscr{A}$, then $\{A : f^{-1}(A) \in \mathcal{E}\}$ is a σ-algebra containing $\mathscr{A}$ and hence $\mathcal{G}$, therefore f is measurable.
We can use this on $(G, \mathcal{G}) = (\mathbb{R}, \mathcal{B})$ with $\mathscr{A} = \{(-\infty, y] : y \in \mathbb{R}\}$, which translates to the criterion that if $\{x \in E : f(x) \leq y\} \in \mathcal{E}$ for all $y \in \mathbb{R}$, then f is Borel measurable.
If E is a topological space and $f : E \to \mathbb{R}$ is continuous, then $f^{-1}(U)$ is open in E whenever $U \subset \mathbb{R}$ is open. Since open sets in $\mathbb{R}$ generate $\mathcal{B}$, the above criterion shows that f is Borel measurable.
If $A \in \mathcal{E}$ is measurable, the indicator 1_A function of A is obviously measurable. In example sheet, one will show that whenever $(f_n)_n$ are measurable functions $E \to \mathbb{R}$, so are $f_1 + f_2, f_1 f_2$ (hence also linear combinations of f_1 and f_2), $\inf_n f_n, \sup_n f_n, \liminf_n f_n$ and $\limsup_n f_n$. In addition, when $(f_n)_n$ are measurable, $\{x \in E : f_n(x) \text{ converges as } n \to \infty\}$ is a measurable subset of E.
Besides, given any collection of maps $f_i : E \to G$ with $(E, \mathcal{E}), (G, \mathcal{G})$ measurable spaces, it is easy to see that they are all measurable if (and only if) $\mathcal{E}$ contains

the σ-algebra $\sigma(\{f_i\}_i) = \sigma(\{f_i^{-1}(A) : A \in \mathcal{G}, i \in I\})$. This is known as the σ-algebra generated by $\{f_i\}_i$, which can be alternatively characterised as the "smallest" σ-algebra on E such that $\{f_i\}_i$ are all measurable.

Definition 2.2. For a sequence $(x_n)_n \in \mathbb{R}$, we write $x_n \uparrow x$ if (x_n) is monotone increasing and $x_n \to x$. For functions $f_n, f : E \to \mathbb{R}$, we write $f_n \uparrow f$ if $f_n(x) \uparrow f(x)$ for all $x \in E$.

For real-valued functions f, g, we write $f \wedge g = \min\{f, g\}, f \vee g = \max\{f, g\}$. One thing to note is that if f, g are measurable, so are $f \wedge g = 1_{(f-g)^{-1}((-\infty,0])} f + 1_{(f-g)^{-1}((0,\infty))} g$ and $f \vee g = 1_{(f-g)^{-1}((-\infty,0])} g + 1_{(f-g)^{-1}((0,\infty))} f$.

Theorem 2.1 (Monotone Class Theorem)**.** *Let $(E, \mathcal{E})$ be a measurable space and $\mathscr{A}$ a π-system generating $\mathcal{E}$. Let V be a vector space of bounded maps $f : E \to \mathbb{R}$ (i.e. $\sup_{x \in E} |f(x)| < \infty$) such that:*
1. $1 = 1_E \in V$ and $\forall A \in \mathscr{A}, 1_A \in V$.
2. Whenever $f_n \in V$ are such that $0 \le f_n \uparrow f$, then $f \in V$.
Then V contains all bounded measurable functions.

Proof. Define $\mathcal{D} = \{A \in \mathcal{E}, 1_A \in V\}$ which contains $A = E$. We also have $1_{B-A} = 1_B - 1_A \in V$ whenever $A, B \in \mathcal{D}$ since V is a vector space. For any $A_n \uparrow A$ with $A_n \in \mathcal{D}$, $0 \le 1_{A_n} \uparrow 1_A \in V$, so $A \in \mathcal{D}$. These make $\mathcal{D}$ a d-system that contains the π-system $\mathscr{A}$, so by Lemma 1.2 we have $\mathcal{D} \supset \mathcal{E}$. But we have $\mathcal{D} \subset \mathcal{E}$ by definition, so $\mathcal{E} = \mathcal{D}$, i.e. $1_A \in V$ for any $A \in \mathcal{E}$.
Let $f : E \to [0, \infty)$ be any bounded nonnegative measurable map. Define

$$f_n(x) = n \wedge 2^{-n} \lfloor 2^n f(x) \rfloor = 2^{-n} \sum_{j=0}^{n2^n} j 1_{A_{j,n}}$$

where

$$\forall j < n2^n, A_{j,n} = f^{-1}\left(\left(\frac{j}{2^n}, \frac{j+1}{2^n}\right]\right) \in \mathcal{E}, A_{n2^n,n} = f^{-1}((n, \infty)) \in \mathcal{E}$$

Then clearly $f_n \in V$ and $0 \le f_n \uparrow f$, hence $f \in V$.
For bounded measurable map f that is not necessarily nonnegative, we take $f = f^+ - f^-$ where $f^+ = f \vee 0, f^- = -(f \wedge 0)$ which completes the proof. $\square$

2.2 Image Measures

Definition 2.3. Let $(E, \mathcal{E}), (G, \mathcal{G})$ be measurable spaces and let $f : E \to G$ be measurable. For any measure μ on E, the measure $\nu(A) = \mu(f^{-1}(A))$ on $\mathcal{G}$ is called the image measure of μ under f.

One can check (example sheet) that this is indeed a measure on $\mathcal{G}$.

Lemma 2.2. *Let $g : \mathbb{R} \to \mathbb{R}$ be nonconstant, right-continuous and nondecreasing. Let $I = (g(-\infty), g(\infty))$ and $f : I \to \mathbb{R}$ given by $f(x) = \inf\{y \in \mathbb{R} : x \le g(y)\}$, then f is left-continuous, nondecreasing, and $f(x) \le y \iff x \le g(y)$.*

Proof. For $x \in I$, the set $J_x = \{y \in \mathbb{R} : x \le g(y)\}$ is nonempty, so the infimum exists in $\mathbb{R}$. If $y \in J_x$ and $y' \ge y$, then $x \le g(y) \le g(y')$, so $y' \in J_x$. Also, for any sequence $y_k \in J_x$ with y_n decreasing to y, by right continuity we have

$x \le g(y) \implies y \in J_x$, hence $J_x = [f(x), \infty)$, so $f(x) \le y \iff y \in J_x \iff x \le g(y)$, as desired. Next note that $x' \le x \implies J_{x'} \subset J_x \implies f(x') \le f(x)$, so f is nondecreasing. Finally, if x_n increases to x, then $J_x = \bigcap_n J_{x_n}$, hence $f(x_n) \to f(x)$ as $n \to \infty$. □

Definition 2.4. Let E be a locally compact Hausdorff topological space (e.g. $\mathbb{R}^d$), a Radon measure μ on E is a measure on $(E, \mathcal{B}(E))$ that is finite on any compact subsets of E.

Theorem 2.3. *Let $g \in \mathbb{R} \to \mathbb{R}$ be as in the preceding lemma, then there exists a unique Radon measure μ_g on $(\mathbb{R}, \mathcal{B})$ such that $\mu_g((a, b]) = g(b) - g(a)$. Moreover, any Radon measure on $\mathbb{R}$ can be represented in this way.*

Remark. A measure μ_g as such is called the Lebesgue-Stieltjes measure with distribution function g.

Proof. If f, I are as in the lemma, and μ the Lebesgue measure on $\mathbb{R}$, then $\{x \in I : f(x) \le z\} = \{x \in I : x \le g(z)\} = (g(-\infty), g(z)] \in \mathcal{B}(I)$ by the preceding lemma. The intervals $\{(-\infty, z] z \in \mathbb{R}\}$ generate $\mathcal{B}(\mathbb{R})$, so f is measurable and the image measure $\mu_g(A) = \mu \circ f^{-1}(A)$ exists on $\mathcal{B}(\mathbb{R})$. We also hav $\mu_g((a, b]) = \mu(\{x : f(x) > a, f(x) \le b\}) = \mu((g(a), g(b)]) = g(b) - g(a)$ by the lemma. By the very same uniqueness argument we used for μ, we see that μ_g is the unique such measure. It is also Radon since any compact subset of $\mathbb{R}$ is bounded.
Conversely, let ν be any Radon measure, then we can define

$$g(y) = \begin{cases} \nu((0, y]) & \text{for } y \ge 0 \\ -\nu((y, 0]) & \text{for } y \le 0 \end{cases}$$

which is clearly nondecreasing, right-continuous, and satisfies $\nu((a, b]) = g(b) - g(a)$. Then by uniqueness we have $\nu = \mu_g$. □

Example 2.1. Consider the Dirac point mass measure δ_x at $x \in \mathbb{R}$ is defined by $\delta_x(A) = 1_{x \in A}$. It is a Radon measure with distribution function $g = 1_{[x, \infty)}$.

2.3 Random Variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(E, \mathcal{E})$ a measure space.

Definition 2.5. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(E, \mathcal{E})$ a measure space. A measurable map $X : \Omega \to E$ is called a random variable (r.v.) in E. The image measure on $\mathcal{E}$ obtained from X is called the law or distribution of X.

We usually take $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B})$, in which case we simply call it a (real) random variable. In this case, the distribution of X is determined by the π-system $\{(-\infty, x] : x \in \mathbb{R}\}$ and $F_X(x) = \mu_X((-\infty, x]) = \mathbb{P}(\omega \in \Omega : X(\omega) \le x) = \mathbb{P}(X \le x)$ (by convention). $F_X(x)$ is called the cumulative distribution function (cdf) of X.
By properties of measures, we immediately have the increasingness of F and

$$\lim_{z \downarrow x} F(z) = \lim_{z \to x^+} F(z) = F(x), \lim_{z \to \infty} F(z) = \mathbb{P}(\Omega) = 1, \lim_{z \to \infty} F(z) = \mathbb{P}(\varnothing) = 0$$

By a mild confusion of notation, we say that any F with these properties a cdf. To justify it, note that if $\Omega = (0,1), \mathcal{F} = \mathcal{B}((0,1))$, and $\mu = \mu|_{(0,1)}$ be the Lebesgue measure restricted to $(0,1)$, then we know from Lemma 2.2 that the random variable $X(\omega) = \inf\{x : \omega \leq F(x)\}, \omega \in (0,1)$ has distribution function

$$\mathbb{P}(\omega \in \Omega : X(\omega) \leq x) = \mathbb{P}(\omega \leq F(x)) = F(x) - 0 = F(x)$$

So every cdf is the cdf of some random variable.

Definition 2.6. We say a countable family of (real) random variables $\{X_i\}_i$ is independent if the family of σ-algebras $\{\sigma(X_i) : i \in I\}$ is independent.

One can check (in example sheet) that this is equivalent to requiring that $\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \cdots \mathbb{P}(X_n \leq x_n)$ for all $x_i \in \mathbb{R}$ and finite subsets $X_1, \ldots, X_n \in \{X_i\}_i$.

Example 2.2. Consider the probability space $((0,1), \mathcal{B}((0,1)), \mu|_{(0,1)})$. Recall that any $\omega \in \Omega = (0,1)$ has a binary expansion $\omega = 0.\omega_1\omega_2\omega_3\cdots$ which is unique if we require the expansion to not include trailing zeros (so $1/2 = 0.01111\ldots$ instead of $1/2 = 0.10000\cdots$). The Rademacher random variables are defined as $R_n : (0,1) \to \{0,1\}, R_n(\omega) = \omega_n$, so for example $R_1 = 1_{(1/2,1)}, R_2 = 1_{(1/4,1/2]} + 1_{(3/4,1)}, R_3 = 1_{(1/8,1/4]} + 1_{(3/8,1/2]} + 1_{(5/8,3/4]} + 1_{(7/8,1)}$. So all the R_n are measurable as a sum of indicators of Borel sets, and we have $\mathbb{P}(R_n = 1) = \mathbb{P}(R_n = 0) = 1/2$.
They are actually independent: For any $x_i \in \{0,1\}^n$, we have

$$\mathbb{P}(R_1 = x_1, \ldots, R_n = x_n) = 2^{-n} = \left(\frac{1}{2}\right)^n = \mathbb{P}(R_1 = x_1) \cdots \mathbb{P}(R_n = x_n)$$

Hence the R_n are all independent. Note that they are also identically distributed. We call family of independent random variables that share the same distribution i.i.d. (independently and identically distributed) random variables. Now take a bijection $m : \mathbb{N}^2 \to \mathbb{N}$ (e.g. $m(k,n) = 2^{k-1}(2n-1)$) and define new variables $Y_{n,k} = R_{m(k,n)}$ (which are i.i.d.) and $Y_n = \sum_k 2^{-k} Y_{n,k}$ which are also i.i.d.. But what exactly are the distributions of Y_n? Consider the intervals $(i/2^m, (i+1)/2^m]$ with rational endpoints. Then

$$\begin{aligned}\mathbb{P}\left(\frac{i}{2^m} < Y_n \leq \frac{i+1}{2^m}\right) &= \mathbb{P}\left(\frac{i}{2^m} < \sum_k 2^{-k} Y_{n,k} \leq \frac{i+1}{2^m}\right) \\ &= \mathbb{P}(Y_{n,1} = y_1, \ldots, Y_{n,m} = y_m) = 2^{-m}\end{aligned}$$

where $(y_1, \ldots, y_m) \in \{0,1\}^m$ corresponds to the binary expansion of $y = i/2^m$. Easy to see that these intervals form a π-system generating $\mathcal{B}((0,1))$, so we conclude that the distribution of Y_n equals $\mu|_{(0,1)}$, hence $\{Y_n : n \in \mathbb{N}\}$ are i.i.d. $\mathrm{Unif}(0,1)$.

Proposition 2.4. *Let $\{F_n\}_n$ be any sequence of cdf's on $\mathbb{R}$, then there exists a sequence of independent random variables $\{X_n\}_n$ on $((0,1), \mathcal{B}((0,1)), \mu|_{(0,1)})$ with $F_{X_n} = F_n$ for all n.*

Proof. We already know from the preceding example that there exists i.i.d. $\mathrm{Unif}(0,1)$ random variables $\{Y_n\}_n$ on the said space. Take $g_n(y) = \inf\{x : y \leq F_n(x)\}$ and $X_n = g_n(Y_n)$ which are independent random variables with distributions $\mathbb{P}(X_n \leq x) = \mathbb{P}(g_n(Y_n) \leq x) = \mathbb{P}(Y_n \leq F_n(x)) = F_n(x)$ by Lemma 2.2. □

2.4 Convergence

We want to talk about convergence or asymptotic results in probability theory under this framework of measure theory. But first, we need to establish some new concept in order to accomodate the intuitive idea of convergence of random variables.

Definition 2.7. Let $(E, \mathcal{E}, \mu)$ be a measure space. A property (in the form of a measurable subset) $A \in \mathcal{E}$ is said to hold almost everywhere (a.e. or μ-a.e.) if $\mu(A^c) = 0$.

When $(E, \mathcal{E}, \mu) = (\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, we (just as a convention) use the synonym "almost surely" (or a.s., $\mathbb{P}$-a.s.).

Definition 2.8. A sequence of measurable $f_n : (E, \mathcal{E}, \mu) \to \mathbb{R}$ is said to converge μ-a.e. to $f : E \to \mathbb{R}$ (or $f_n \to^{\text{a.e.}} f$) as $n \to \infty$ if $\{x \in E : f_n(x) \to f(x)\}$ holds μ-a.e.. If $f_n, f : (E, \mathcal{E}, \mu) \to \mathbb{R}$ are measurable, we say $f_n \to f$ in μ-measure (or $f_n \to^{\mu} f$) as $n \to \infty$ if for any $\epsilon > 0$, $\mu(x \in E : |f_n(x) - f(x)| > \epsilon) \to 0$.

One will show in example sheet that $\{x \in E : f_n(x) \to f(x)\}$ and $\{x \in E : |f_n(x) - f(x)| > \epsilon\}$ are always measurable. Correspondingly, we can establish the same notions for a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and its random variables, in which case we replace a.e. by a.s., μ by $\mathbb{P}$ and "in measure" by "in probability". For real random variables, however, we have another notion of convergence.

Definition 2.9. For (real) random variables X_n, X, we say $X_n \to X$ in distribution (or $X_n \to^{\text{d}} X$) if $F_{X_n}(t) \to F_X(t)$ whenever F_X is continuous at t.

Theorem 2.5. *Let $(f_n)_n$ be measurable on $(E, \mathcal{E}, \mu)$.*
1. If $\mu(E) < \infty$ and $f_n \to^{\text{a.e.}} 0$, then $f_n \to^{\mu} 0$.
2. If $f_n \to^{\mu} 0$, then $f_{n_k} \to^{\text{a.e.}} 0$ for a subsequence (f_{n_k}) of (f_n).

Proof. 1. For any $\epsilon > 0$,

$$\mu(|f_n| \le \epsilon) \ge \mu\left(\bigcap_{m \ge n} \{|f_m| \le \epsilon\}\right) = \mu\left(\bigcap_{m \ge n} A_m\right), A_m = \{|f_m| \le \epsilon\}$$
$$\uparrow \mu\left(\bigcup_k \bigcap_{m \ge k} A_m\right) \ge \mu(f_n \to 0, n \to \infty) = \mu(E) < \infty$$

But $\mu(|f_n| \le \epsilon) \le \mu(E)$ for all n, so we must have (by taking complements) $\limsup_n \mu(|f_n| < \epsilon) = 0$, i.e. $f_n \to^{\mu} 0$.
2. For every m, select a subsequence n_k of indices such that $\sum_k \mu(|f_{n_k}| > 1/m) < \infty$ (e.g. such that $\mu(|f_{n_k}| > 1/m) < 1/k^2$). Then by Lemma 1.7, $\mu(|f_{n_k}| > 1/m \text{ i.o.}) = 0$, so $f_{n_k} \to^{\text{a.e.}} 0$. □

Remark. 1. In the first part of the theorem, the requirement of $\mu(E)$ being finite is indeed necessary, since $1_{(n,\infty)} \to^{\text{a.e.}} 0$ as $n \to \infty$, but $\mu(x : |f_n(x)| > \epsilon) = \mu((n, \infty)) = \infty$.
2. In general, we do need to pass to a strict subsequence in the second part of the theorem. Take independent $\{A_n : n \in \mathbb{N}\}$ such that $\mathbb{P}(A_n) = 1/n \to 0$ (this is possible since we know there exists a sequence of independent uniform

variables), then $\mathbb{P}(1_{A_n} > \epsilon) = \mathbb{P}(A_n) = 1/n \to 0$, so $1_{A_n} \to^{\mathbb{P}} 0$. Nonetheless, $\sum_n \mathbb{P}(A_n) = \infty$, so by Lemma 1.8 we have $\mathbb{P}(1_{A_n} = 1 \text{ i.o.}) = 1$, so 1_{A_n} does not $\mathbb{P}$-almost surely converge to 0.
3. One can show further that if $X_n \to^{\mathbb{P}} X$ then $X_n \to^{\mathrm{d}} X$. Moreover, if $X_n \to^{\mathrm{d}} X$, then there exists $\tilde{X}_n, \tilde{X} : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ such that $\tilde{X}_n$ the same distributionn as X_n, $\tilde{X}$ the same distribution as X and $\tilde{X}_n \to \tilde{X}$ a.s. as $n \to \infty$.

Example 2.3. Let $(X_n)_n$ be i.i.d. with cdf $\mathbb{P}(X_n \leq x) = 1 - e^{-x}$ and consider $A_n = \{X_n \geq \alpha \log n\}, \alpha > 0$, so that $\mathbb{P}(A_n) = e^{-\alpha \log n} = n^{-\alpha}$, hence $\sum_n \mathbb{P}(A_n) < \infty$ iff $\alpha > 1$, hence by Lemma 1.7 and Lemma 1.8 we have

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

Hence $\limsup_n X_n / \log n = 1$ a.s..

2.5 Kolmogorov's Zero-One Law

Definition 2.10. For any sequence $(X_n)_n$ of random variables, we set

$$\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, \ldots), \mathcal{T} = \bigcap_n \mathcal{T}_n$$

$\mathcal{T}$ is called the tail σ-algebra of the sequence $(X_n)_n$.

Theorem 2.6 (Kolmogorov)**.** *Let $(X_n)_n$ be independent random variables with tail σ-algebra $\mathcal{T}$. If $A \in \mathcal{T}$, then $\mathbb{P}(A) \in \{0, 1\}$. Furthermore, if $Y : (\Omega, \mathcal{T}, \mathbb{P}) \to \mathbb{R}$ is measurable, then Y is a.s. constant.*

Proof. The σ-algebra $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ is generated by the π-system $\mathscr{A} = \{X_1 \leq x_1, \ldots, X_n \leq x_n : x_i \in \mathbb{R}\}$. Correspondingly, $\mathcal{T}_n$ is generated by the π-system of sets $\mathscr{B} = \{X_{n+1} \leq x_{n+1}, \ldots, X_{n+k} \leq x_{n+k} : k \in \mathbb{N}, x_i \in \mathbb{R}\}$. Now $\mathscr{A}$ and $\mathscr{B}$ are independent, therefore $\mathcal{F}_n$ and $\mathcal{T}_n$ are also independent by Theorem 1.6. Then $\bigcup_n \mathcal{F}_n$ is a π-system generating $\mathcal{F}_\infty = \sigma((X_n)_n)$. Consequently $\mathcal{F}_\infty, \mathcal{T}$ have to be independent. But $\mathcal{T} \subset \mathcal{T}_n \subset \mathcal{F}_\infty$, so any $A \in \mathcal{T} \subset \mathcal{F}_\infty$ would have to be independent to itself. That is, $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2$, so $\mathbb{P}(A)$ can only be 0 or 1.
Finally, if Y is $\mathcal{T}$-measurable, then $Y^{-1}((-\infty, y]) = \{Y \leq y\} \in \mathcal{T}$ has probability either 0 or 1, so $Y = \inf\{y : F_Y(y) = 1\}$ a.e.. □

3 Integration

3.1 The Lebesgue Integral

For a measure space $(E, \mathcal{E}, \mu)$ and a measurable function $f : E \to \mathbb{R}$ (and hence $\mathbb{R}^d$ by putting everything pointwise), our aim is to define its (Lebesgue) integral

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

when f is nonnegative and then when f is "integrable".
After this is done, we will define the expectation of a random variable X (with range in $\mathbb{R}^d$) as

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

Definition 3.1. A function $f : E \to \mathbb{R}$ is simple if it has the form $f = \sum_{k=1}^{m} a_k 1_{A_k}$ for $A_k \in \mathcal{E}, m \in \mathbb{N}, a_k \in \mathbb{R}_{\geq 0}$.
For such f, we set its integral to be

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

One can check that this is well-defined (i.e. independent of the way we express f in the said form) and is linear (in the sense that for any $\alpha, \beta > 0$ and f, g simple one has $\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g)$), monotonic (if $f \leq g$ are simple, then $\mu(f) \leq \mu(g)$), and that $\mu(f) = 0$ iff f is zero almost everywhere.

Definition 3.2. For a general measurable nonnegative function $f : E \to [0, \infty]$, we set $\mu(f) = \sup\{\mu(g) : g \text{ simple}, g \leq f\}$.

Remark. We will see that we only need to take such supremum over a particular class of simple function, namely $f_n(x) = n \wedge 2^{-n}\lfloor 2^n f(x) \rfloor$ which we used in the proof of Theorem 2.1.
So, intuitively, while Riemann integral try to sum up the area under the graph of f by vertical stripes, Lebesgue integral does it by collecting horizontal "scraps" cut out (in this case) at $j/2^n$ for $j = 0, \ldots n2^n$.

Note that μ is still monotonic under this extended definition of integral, so we are comfortable writing

Definition 3.3. For a general measurable function, $f : E \to [-\infty, \infty]$ is measurable, we set $f^+ = f \vee 0, f^- = -(f \wedge 0)$, then $f = f^+ - f^-$ and $|f| = f^+ + f^-$. We say f is (Lebesgue) integrable if $\mu(|f|) < \infty$. If this were the case, we define its integral as $\mu(f) = \mu(f^+) - \mu(f^-)$.

Theorem 3.1 (Monotone Convergence Theorem)**.** *Let $f_n, f : (E, \mathcal{E}, \mu) \to [0, \infty]$ be measurable and suppose $f_n \uparrow f$, then $\mu(f_n) \uparrow \mu(f)$.*

Proof. $\mu(f_n)$ is increasing since f_n is increasing, so $\mu(f_n) \uparrow M = \sup_n \mu(f_n)$ for some $M \in [0, \infty]$. But $\mu(f_n) \leq \mu(f)$ for all n since $f_n \uparrow f$, consequently $M \leq \mu(f)$.
To complete the proof, if suffices to show that any simple $g \leq f$ has $\mu(g) \leq M$. WLOG we take $g = \sum_{n=1}^{m} a_k 1_{A_k}$ with A_k disjoint. Define a new sequence of simple functions $g_n = f_n \wedge g$ where $f_n = 2^{-n}\lfloor 2^n f(x) \rfloor$ (check it is indeed simple!), then $g_n \uparrow f \wedge g = g$. Fix $\epsilon \in (0, 1)$ and consider

$$A_k(n) = \{x \in A_k : g_n(x) \geq (1 - \epsilon)g(x)\} = \{x \in A_k : g_n(x) \geq (1 - \epsilon)a_k\}$$

Since $g_n \uparrow g$, we must have $A_k(n) \uparrow A_k$ as $n \to \infty$, so $\mu(A_k(n)) \uparrow \mu(A_n)$. $1_{A_k} g_n \geq (1 - \epsilon)a_k 1_{A_k(n)}$ by definition, so $\mu(1_{A_k} g(n)) \geq (1 - \epsilon)a_k \mu(A_k(n))$. $g_n = f_n \wedge g$ is zero outside $\bigcup_k A_k$, so $g_n = \sum_k 1_{A_k} g_n$, therefore (since μ is linear on nonnegative simple functions)

$$\mu(g_n) = \sum_{k=1}^{m} \mu(1_{A_k} g(n)) \geq (1 - \epsilon) \sum_{k=1}^{m} a_k \mu(A_k(n))$$

$$\uparrow (1 - \epsilon) \sum_{k=1}^{m} a_k \mu(A_k) = (1 - \epsilon)\mu(g)$$

So $(1 - \epsilon)\mu(g) \leq \mu(g_n) \leq \mu(f_n) \leq M$. As $\epsilon \in (0, 1)$ is arbitrary, we must have $\mu(g) \leq M$ which is what we wanted. □

Corollary 3.2. *Any sequence $(g_n)_n$ of nonnegative measurable functions has*

$$\sum_n \mu(g_n) = \mu\left(\sum_n g_n\right)$$

Remark. 1. Alternatively, the result of the theorem can be phrased as

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

as long as $0 \le f_n \uparrow f$.
2. The requirement $f_n \uparrow f$ can be weakened to $f_n \uparrow f$ a.e..
3. If $(f_n)_n$ are measurable functions and $f_n \uparrow f$ but $\mu(f_1) > -\infty$, then the theorem also holds. The condition $\mu(f_1) > -\infty$ is, however, necessary: Take the example $f_n = -1_{(n,\infty)}$.
4. Symmetrically, suppose $f_n \downarrow f$ (exactly what you think it means) and $\mu(f_1) < \infty$, then $\mu(f_n) \downarrow \mu(f)$.

The power of this theorem allows us to generalise some properties of the integral from simple functions to nonnegative measurable functions in general.

Proposition 3.3. *Let $f, g \ge 0$ be measurable functions, then:*
(a) $\mu(\alpha f + \beta g) = \alpha\mu(g) + \beta\mu(g)$ for all $\alpha, \beta \ge 0$.
(b) $\mu(f) = 0 \iff f = 0$ a.e..

Proof. Consider $f_n = (2^{-n}\lfloor 2^n f\rfloor) \wedge n, g_n = (2^{-n}\lfloor 2^n g\rfloor) \wedge n$ which are simple functions with $f_n \uparrow f, g_n \uparrow g$.
(a) We have $\alpha f_n + \beta g_n \uparrow \alpha f + \beta g$. Also, $\mu(\alpha f_n + \beta g_n) = \alpha\mu(f_n) + \beta\mu(g_n)$ since f_n, g_n are simple. Theorem 3.1 then finishes the proof.
(b) If $f = 0$ a.e., then $f_n = 0$ a.e. for all n, so $0 = \mu(f_n) \uparrow \mu(f) \implies \mu(f) = 0$ by Theorem 3.1. Conversely, if $\mu(f) = 0$, then $0 \le \mu(f_n) \le \mu(f) = 0 \implies \mu(f_n) = 0$ for all n. This means that $f_n = 0$ a.e. for all n. But $f_n \uparrow f$, so $f = 0$ almost everywhere. □

Theorem 3.4. *Let $f, g : E \to \mathbb{R}$ be integrable functions, then:*
(a) For any $\alpha, \beta \in \mathbb{R}$ we have $\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g)$.
(b) If $f = 0$ almost everywhere, then $\mu(f) = 0$.
(c) $\mu(g) \le \mu(f)$ whenever $g \le f$.

Proof. (a) For $\alpha \ge 0$, αf is also integrable and $\mu(\alpha f) = \mu(\alpha f^+ - \alpha f^-) = \alpha\mu(f^+) - \alpha\mu(f^-) = \alpha\mu(f)$. Combining this with $\mu(-f) = -\mu(f)$ gives $\mu(\alpha f) = \alpha\mu(f)$ for $\alpha < 0$. The identity $(f+g)^+ + f^- + g^- = (f+g)^- + f^+ + g^+$ shows that $\mu(f+g) = \mu(f) + \mu(g)$. Combining them gives the result.
(b) If $f = 0$ a.e., then $f^+ = 0$ a.e. and $f^- = 0$ a.e., therefore $\mu(f) = \mu(f^+) - \mu(f^-) = 0 - 0 = 0$.
(c) If $f \le g$, then $\mu(g) - \mu(f) = \mu(g - f) \ge 0$, so $\mu(f) \le \mu(g)$. □

Remark. There are obviously many integrable functions f with $\mu(f) = 0$ but f is not zero a.e.. However, to check if $f = 0$ a.e. we can always check if $\mu(|f|) = 0$ or if $\mu(f1_A) = 0$ for all $A \in \mathscr{A}$ where $\mathscr{A}$ is a π-system containing E that generates $\mathcal{E}$.

3.2 Dominated Convergence Theorem

Recall that for a real sequence (x_n) we have the notions

$$\liminf_n x_n = \sup_n \inf_{m \geq n} x_m, \limsup_n x_n = \inf_n \sup_{m \geq n} x_m$$

In particular, $x_n \to x$ iff $\liminf_n x_n = \limsup_n x_n = x$.
For sequences f_n of functions, we define their lim, sup, lim inf, lim sup in a pointwise way. It is easy to see that (if those notions turns out to be well-defined) $\inf f_n, \sup f_n$ and hence $\liminf f_n, \limsup f_n$ stays measurable given that (f_n) are all measurable.

Lemma 3.5 (Fatou)**.** *Suppose $f_n : (E, \mathcal{E}, \mu) \to [0, \infty)$ are measurable, then $\mu(\liminf_n f_n) \leq \liminf_n \mu(f_n)$.*

If $f_n \to f$ pointwise, then Fatou's lemma implies that $\mu(f) \leq \liminf_n \mu(f_n)$.

Proof. For any $k \geq n$, we have $\inf_{m \geq n} f_m \leq f_k$, so $\mu(\inf_{m \geq n} f_m) \leq \mu(f_k)$ for all $k \geq n$, hence

$$\mu\left(\inf_{m \geq n} f_m\right) \leq \inf_{k \geq n} \mu(f_k) \leq \sup_n \inf_{k \geq n} \mu(f_k) = \liminf_n \mu(f_n)$$

Also, we have $\inf_{m \geq n} f_m \uparrow \sup_n \inf_{m \geq n} f_m$, so by Theorem 3.1 we have

$$\mu\left(\inf_{m \geq n} f_m\right) \uparrow \mu\left(\sup_n \inf_{m \geq n} f_m\right) = \mu\left(\liminf_n f_n\right)$$

Combining the two gives $\mu(\liminf_n f_n) \leq \liminf_n \mu(f_n)$. □

Theorem 3.6 (Dominated Convergence Theorem)**.** *Let $f_n, f : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be measurable and suppose $f_n \to f$ pointwise on E and that there is some g integrable such that $|f_n| \leq g$ for all n, then f_n, f are integrable and $\mu(f_n) \to \mu(f)$ as $n \to \infty$.*

Proof. We have $|f| \leq g$ by taking limits, so $\mu(|f|) \leq \mu(g) < \infty$. It follows that f, f_n are all integrable. Now write $0 \leq g \pm f_n \to g \pm f = \liminf_n(g \pm f_n)$. By Fatou's lemma,

$$\begin{aligned}\mu(g) \pm \mu(f) = \mu(g \pm f) &= \mu\left(\liminf_n (g \pm f_n)\right) \\ &\leq \liminf_n \mu(g \pm f_n) = \mu(g) + \liminf_n \mu(\pm f_n)\end{aligned}$$

So

$$\limsup_n \mu(f_n) = -\liminf_n(-\mu(f_n)) \leq \mu(f) \leq \liminf_n \mu(f_n)$$

But we also have $\limsup_n \mu(f_n) \geq \liminf_n \mu(f_n)$, so equality must hold and $\mu(f_n) \to \mu(f)$. □

Example 3.1. Take $E = [0, 1]$ and suppose $f_n \to f$ pointwise and $g = \sup_n \|f_n\|_\infty$ (i.e. f_n is uniformly bounded), then $\mu(g) = g < \infty$, so $\mu(f_n) \to \mu(f)$ by the preceding theorem.

So the "higher" viewpoint of Lebesgue integral has made such strong convergence theorems much easier to prove.

3.3 The Wonders of Calculus

For Riemann integrals, we can evaluate integrals by seeking antiderivatives. Naturally, we want this – or even better, one that applies to a more general setting – to be true for Lebesgue integral as well.

Theorem 3.7. *Let $U \subset \mathbb{R}$ be open and $f : U \times E \to \mathbb{R}$ a function such that:*
1. $\forall t \in U, x \mapsto f(t,x)$ is integrable.
2. $\forall x \in E, t \mapsto f(t,x)$ is differentiable (with derivative say $\dot{f}(t,x)$).
3. $|\dot{f}(t,x)| \leq g(x)$ globally for some $g : E \to \mathbb{R}$ integrable.
Then $x \mapsto \dot{f}(t,x)$ is integrable for all t. Moreover, if we set

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

is differentiable and

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

Proof. For $h_n \to 0$, set

$$g_n(x) = \frac{f(t+h_n) - f(t,x)}{h_n} - \dot{f}(t,x) \to 0$$

as $n \to \infty$ pointwise on E. By mean value theorem there is some $\tilde{t}_n \in (t, t+h_n)$ such that $|g_n(x)| = |\dot{f}(\tilde{t}_n, x) - \dot{f}(t,x)| \leq 2|g(x)|$ which is μ-integrable. So by the Theorem 3.6 we have $\mu(g_n) \to \mu(0) = 0$, i.e.

$$\frac{F(t+h_n) - F(t)}{h_n} = \int_E \frac{f(t+h_n) - f(t,x)}{h_n}\,\mathrm{d}\mu(x) \to \int_E \dot{f}(t,x)\,\mathrm{d}\mu(x)$$

as desired. □

Remark. 1. One can show that any Riemann integrable function on $[0,1]$ is (Lebesgue) integrable on $([0,1], \mathcal{M}, \mu)$ with

$$\mu(x) = \int_0^1 f(x)\,\mathrm{d}x$$

But note that there are Riemann-integrable functions that are not Borel measurable, which in particular proves that the inclusion $\mathcal{B} \subset \mathcal{M}$ is strict. There are, of course, Lebesgue integrable functions that are not Riemann integrable: Take $1_{\mathbb{Q}}$.
2. For continuous functions on $[0,1]$ (which are in particular measurable), the fundamental theorem of calculus for Lebesgue integrals is true and can be proved in almost the same way as that for the Riemann integral. In fact, Lebesgue proved a more general assertion (known as the Lebesgue differentiation theorem): If $f : [0,1] \to \mathbb{R}$ is Lebesgue integrable and

$$F(x) = \int_0^x f(y)\,\mathrm{d}\mu(y)$$

then

$$\lim_{h\to\infty} \frac{F(x+h) - F(x)}{h} = \lim_{h\to 0} \frac{1}{h}\int_x^{x+h} f(y)\,\mathrm{d}\mu(y) = f(x) \text{ a.e.}$$

Proposition 3.8. *Let $\phi : [a, b] \to \mathbb{R}$ be continuously differentiable and strictly increasing, then for any Borel measurable $g : [\phi(a), \phi(b)] \to [0, \infty]$, we have*

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

Proof. The proposition is true for $g = 1_{(\phi(a_i),\phi(b_i))}$ with $a \le a_i < b_i \le b$ by fundamental theorem of calculus. The general case can be proved by an approximation argument like one we used in the proof of Theorem 2.1. □

More generally (as you'll prove in example sheet), if $f : (E, \mathcal{E}, \mu) \to (G, \mathcal{G})$ is measurable and $\nu(A) = \mu(f^{-1}(A))$ is the image measure on $\mathcal{G}$, then for any measurable $g : \mathcal{G} \to [0, \infty]$ we have

$$\int_G g\,\mathrm{d}\nu = \int_E g \circ f\,\mathrm{d}\mu$$

In particular, if $X = (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ is a random variable with distribution μ_X and if $g : \mathbb{R} \to [0, \infty]$ is Borel measurable, then

$$\mathbb{E}g(X) = \int_\Omega g \circ X\,\mathrm{d}\mathbb{P} = \int_{\mathbb{R}} g\,\mathrm{d}\mu_X$$

Proposition 3.9. *Suppose $f : (E, \mathcal{E}, \mu) \to \mathbb{R}_{\ge 0}$ is measurable and $A \in \mathcal{E}$. Define $\nu_f(A) = \mu(f1_A)$, then ν_f is a measure on $(E, \mathcal{E})$ and*

$$\int_E g\,\mathrm{d}\nu_f = \int_E gf\,\mathrm{d}\mu$$

for any measurable $g : (E, \mathcal{E}) \to [0, \infty]$.

Proof. Exercise (example sheet). □

Definition 3.4. We call ν_f the density of f with respect to μ. In particular, if ν_f is the distribution of a random variable X, we say f is the probability density function (pdf) of X.

So the preceding proposition translates to

$$\mathbb{E}g(X) = \int_\Omega g \circ X\,\mathrm{d}\mathbb{P} = \int_{\mathbb{R}} g\,\mathrm{d}\mu_X = \int_{\mathbb{R}} g\,\mathrm{d}\nu_f = \int_{\mathbb{R}} gf\,\mathrm{d}\mu$$

for measurable $g : \mathbb{R} \to [0, \infty]$.

3.4 Product Measures; Fubini's Theorem

For finite measure spaces $(E_1, \mathcal{E}_1, \mu_1), (E_2, \mathcal{E}_2, \mu_2)$, the family of subsets $\mathscr{A} = \{A_1 \times A_2 : A_1 \in \mathcal{E}_1, A_2 \in \mathcal{E}_2\}$ is a π-system on $E_1 \times E_2$.

Definition 3.5. $\mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2 = \sigma(\mathscr{A})$ is called the product σ-algebra on $E = E_1 \times E_2$.

Naturally, we want to define a measure μ on $(E, \mathcal{E})$ by extending $\mu(A_1 \times A_2) = \mu(A_1)\mu(A_2)$ for $A_1 \times A_2 \in \mathscr{A}$. To do this, we need some ground work.

Lemma 3.10. *For any $\mathcal{E}$-measurable $f : E \to \mathbb{R}$ and any $x_1 \in E_1$, then map $x_2 \mapsto f(x_1, x_2)$ is $\mathcal{E}_2$-measurable.*

The mirror version of this lemma is of course also true by symmetry.

Proof. We shall use Theorem 2.1. Let V be the vector space of bounded $\mathcal{E}$-measurable $f : E \to \mathbb{R}$ such that the lemma is true. Then $1_E \in V$ and for any $A_1 \times A_2 \in \mathscr{A}$ we have $1_{A_1 \times A_2}(x_1, x_2) = 1_{A_1}(x_1)1_{A_2}(x_2)$, so $1_{A_1 \times A_2} \in V$. Suppose we have $0 \leq f_n \uparrow f$ for $(f_n)_n \in V$, then (fixing $x \in E_1$) $f_n(x, \cdot) \uparrow f(x, \cdot)$, so $f(x, \cdot)$ is also $\mathcal{E}_2$-measurable as $f_n(x, \cdot)$ are.
Theorem 2.1 then implies that V contains all bounded measurable $f : E \to \mathbb{R}$, i.e. any bounded measurable $f : E \to \mathbb{R}$ satisfies the lemma. The limit $f_n = (-n) \vee (f \wedge n) \uparrow f$ implies the case for unbounded f. □

Lemma 3.11. *Suppose we have a measurable $f : (E, \mathcal{E}) \to [-\infty, \infty]$ that is either bounded or nonnegative. Define*

$$f^1(x_1) = \int_{E_2} f(x_1, x_2) \, \mathrm{d}\mu_2(x_2)$$

Then f^1 is $\mathcal{E}_1$-measurable and is bounded if f is.

Proof. Note that f^1 is well-defined as an integral by the preceding lemma. We again invoke Theorem 2.1 on the vector space of bounded measurable $f : E \to \mathbb{R}$ such that the lemma is true. For $f = 1_{A_1 \times A_2}$, we have $f^1(x_1) = 1_{A_1}(x_1)\mu_2(A_2)$ which is measurable and bounded (as E_2 is finite).
By Theorem 3.1, we know that if $0 \leq f_n \uparrow f, f_n \in V$, then

$$\begin{aligned} f^1(x_1) &= \int_{E_2} f(x_1, x_2) \, \mathrm{d}\mu_2(x_2) \\ &= \lim_{n\to\infty} \int_{E_2} f_n(x_1, x_2) \, \mathrm{d}\mu_2(x_2) = \lim_{n\to\infty} f_n^1(x_1) \end{aligned}$$

So f^1 has to be $\mathcal{E}_1$-measurable, nonnegative and bounded since we have $\|f^1\|_\infty \leq \|f\|_\infty \mu_2(E_2) < \infty$. Consequently V contains all bounded measurable functions. As for nonnegative ones, we use the limit $f \wedge n \uparrow f$. □

Theorem 3.12 (Product Measure)**.** *There exists a unique measure μ on $(E, \mathcal{E})$ such that $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ for any $A_1 \in \mathcal{E}_1, A_2 \in \mathcal{E}_2$.*

Proof. The uniqueness is just a consequence of Theorem 1.3 since $\mathscr{A} = \{A_1 \times A_2 : A_1 \in \mathcal{E}_1, A_2 \in \mathcal{E}_2\}$ is a π-system. For existence, we define

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

for $A \in \mathcal{E}$. This is well-defined by the preceding lemma. It is obviously finitely additive. As for countable additivity, just note that we have $1_{\bigcup_n A_n} = \sum_n 1_{A_n}$ is the increasing limit of $\sum_{n \leq N} 1_{A_n}$ as $N \to \infty$ and use Theorem 3.1 twice. □

Definition 3.6. This unique measure $\mu = \mu_1 \otimes \mu_2$ is called the product measure of μ_1 and μ_2.

Remark. The construction is, of course, perfectly symmetric in μ_1, μ_2.

Consider the product σ-algebra $\bar{\mathcal{E}} = \mathcal{E}_2 \otimes \mathcal{E}_1$ and the product measure $\bar{\mu} = \mu_2 \otimes \mu_1$. We have $\bar{\mu}(\bar{f}) = \mu(f)$ if $\bar{f} : E_2 \times E_1 \to [0, \infty]$ is defined as $\bar{f}(x_2, x_1) = f(x_1, x_2)$ because of the simple reason that $1_{A_1}(x_1)1_{A_2}(x_2) = 1_{A_2}(x_2)1_{A_1}(x_1)$. In general,

Theorem 3.13 (Fubini)**.** *Let $(E, \mathcal{E}, \mu) = (E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2, \mu_1 \otimes \mu_2)$ be a product space.*
(a) If $f : (E, \mathcal{E}, \mu) \to [0, \infty]$ is nonnegative and measurable, then

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

(b) If $f : (E, \mathcal{E}, \mu) \to \mathbb{R}$ is μ-integrable, set

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

and

$$f^1(x_1) = \int_{E_2} f(x_1, x_2) \, \mathrm{d}\mu_2(x_2)$$

for $x_1 \in A_1$ and 0 otherwise. Then $\mu_1(E_1 - A_1) = 0$ and f^1 is μ_1-integrable with $\mu_1(f^1) = \mu(f)$.

In particular, we can "switch the orders of integrals".

Proof. (a) We know it is true for $f = 1_{A_1 \times A_2}$, hence also 1_A for all $A \in \mathcal{E}$ since $\mathscr{A}$ is a π-system. We can further extend this to all simple functions by the linearity of integrals. For general f, we take $f_n = 2^{-n} \lfloor 2^n f \rfloor \wedge n$ and the limit $f_n \uparrow f$ allows us to conclude the result by Theorem 3.1.
(b) The function

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

is $\mathcal{E}_1$-measurable by Lemma 3.11. By (a), we know that

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

In particular $\mu_1(E_1 - A_1) = 0$ and f_1 is μ_1-integrable. Now we can decompose

$$(f^1)^{\pm}(x_1) = \int_{E_2} f^{\pm}(x_1, x_2) \, \mathrm{d}\mu_2(x_2)$$

Then $f^1 = ((f^1)^+ - (f^1)^-)1_A$, so $\mu(f) = \mu(f^+) - \mu(f^-) = \mu_1((f^1)^+) - \mu_1((f^1)^-) = \mu_1(f_1)$. □

Remark. 1. The theorem extends to σ-finite measures (but not to general measure spaces), in particular to Lebesgue measure on $\mathbb{R}$.
2. By a π-system argument argument, we find $(\mathcal{E}_1 \otimes \mathcal{E}_2) \otimes \mathcal{E}_3 = \mathcal{E}_1 \otimes (\mathcal{E}_2 \otimes \mathcal{E}_3)$, so by induction we can extend the definition of product measure from 2 to n components. In particular, we can construct the measure space $(\mathbb{R}^n, \mathcal{B}^{\otimes n}, \mu^{\otimes n})$ where we would have

$$\int_{\mathbb{R}^n} f \, \mathrm{d}\mu^{\otimes n} = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(x_1, \ldots, x_n) \, \mathrm{d}\mu_1(x_1) \cdots \mathrm{d}\mu_n(x_n)$$

for suitable $f : \mathbb{R}^n \to \mathbb{R}$.

What does this mean in the probability world?

3.5 Product Probability Spaces and Independence

Proposition 3.14. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(E_i, \mathcal{E}_i)$ for $i = 1, \ldots, n$ be measurable spaces with product space $(E, \mathcal{E}) = (\prod_i E_i, \bigotimes_i \mathcal{E}_i)$. Let $X_i : (\Omega, \mathcal{F}) \to (E_i, \mathcal{E}_i)$ be random variables, then $X(\omega) = (X_1(\omega), \ldots, X_n(\omega))$ defines a random variable $(\Omega, \mathcal{F}) \to (E, \mathcal{E})$. Under this setting, the followings are equivalent:*
(a) $X_1, \ldots, X_n$ are independent.
(b) $\mu_X = \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}$.
(c) If $f_i : E_i \to \mathbb{R}$ is bounded and $\mathcal{E}_i$-measurable, then

$$\mathbb{E}\left(\prod_{k=1}^{n} f_k(X_k)\right) = \prod_{k=1}^{n} \mathbb{E} f_k(X_k)$$

Proof. (a) $\implies$ (b): Set $\nu = \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}$ which, on the π-system $\mathscr{A} = \{\prod_k A_k : A_k \in \mathcal{E}_k\}$ generating $\mathcal{E}$, satisfies

$$\begin{aligned}\mu_X(A) = \mathbb{P}(X_k \in A_k, \forall k) &= \mathbb{P}\left(\bigcap_{k=1}^{n} \{X_k \in A_k\}\right) \\ &= \prod_{k=1}^{n} \mathbb{P}(X_k \in A_k) = \prod_{k=1}^{n} \mu_{X_k}(A_k) = \nu(A)\end{aligned}$$

Hence $\mu_X = \nu$ on $\mathcal{E}$ by Theorem 1.3.
(b) $\implies$ (c): By Theorem 3.13 we have

$$\begin{aligned}\mathbb{E}\left(\prod_{k=1}^{n} f_n(X_n)\right) &= \int_E \prod_{k=1}^{n} f_k(x_k)\, \mathrm{d}\mu_{X_n}(x_n) \\ &= \prod_{k=1}^{n} \int_{E_k} f_k(x_k)\, \mathrm{d}\mu_{X_n}(x_n) = \prod_{k=1}^{n} \mathbb{E} f_n(X_n)\end{aligned}$$

(c) $\implies$ (a): Take $f_k = 1_{A_k}$ for $A_k \in \mathcal{E}_k$. □

One can, in fact, construct (countably) infinite product probability spaces: Suppose $(\Omega_i, \mathcal{F}_i, \nu_i)_{i=1}^{\infty}$ is a sequence of probability spaces. Set $\Omega = \prod_{i=1}^{\infty} \Omega_i$, we can consider the π-system of "cylinders"

$$\mathscr{C} = \left\{ C \subset \Omega : C = A \times \prod_{i>n} \Omega_i, A \in \bigotimes_{i=1}^{n} \mathcal{F}_i, n \in \mathbb{N} \right\}$$

$\mathcal{F} = \bigotimes_i \mathcal{F}_i = \sigma(\mathscr{C})$ is then the σ-algebra we want to take on Ω. By Theorem 1.3 there is at most one (probability) measure $\nu = \bigotimes_i \nu_i$ on $\mathcal{F}$ such that

$$\nu\left(A \times \prod_{i>n} \Omega_i\right) = \nu_1 \otimes \cdots \otimes \nu_n(A)$$

The existence of such a measure is known as Kolmogorov's extension theorem. In particular, this allows us to realise an infinite sequence of independent random variables with distributions ν_i by just taking coordinate projections $X_n((\omega_1, \omega_2, \ldots)) = \omega_n$.

4 L^p-norms and L^p-spaces

4.1 The L^p-norm; Various Inequalities

Definition 4.1. Let $(E, \mathcal{E}, \mu)$ be a measure space and let $p \in [1, \infty]$. The L^p-space on E is defined by

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

where $\|\cdot\|_p$ is the L^p norm given by

$$\|f\|_p = \left(\int_E |f(x)|^p \, \mathrm{d}\mu(x)\right)^{1/p}$$

for $p < \infty$ and

$$\|f\|_\infty = \operatorname*{ess\,sup}_E |f| = \inf\{\lambda \in \mathbb{R} : |f| \leq \lambda \text{ a.e.}\}$$

Clearly L^p is a vector space for any p. Also, if $\mu(E) < \infty$, then $\|f\|_p \leq \|f\|_\infty (\mu(E))^{1/p}$, so (in this case) $L^\infty \subset L^p$ for any p.

Definition 4.2. We say $f_n \to f$ in L^p as $n \to \infty$ if $\|f_n - f\|_p \to 0$ as $n \to \infty$.

It is not quite clear that these $\|\cdot\|_p$ are indeed norms – the triangle inequality for it is not trivial at all. So we've got some inequality to derive.

Theorem 4.1 (Markov's Inequality (aka Chebyshev's Inequality))**.** *Suppose $f : E \to \mathbb{R}_{\geq 0}$ is measurable, then for any $\lambda \geq 0$,*

$$\mu(x \in E : f(x) \geq \lambda) = \mu(f \geq \lambda) \leq \frac{\mu(f)}{\lambda}$$

Proof. Integrate the inequality $\lambda 1_{f \geq \lambda} \leq f$. □

So if $g \in L^p(\mu)$, then applying this inequality to $f = |g|^p$ gives

$$\mu(|g| > \lambda) = \mu(|g|^p > \lambda^p) \leq \frac{\|g\|_p^p}{\lambda^p} = O(\lambda^{-p})$$

as $\lambda \to \infty$.

Definition 4.3. Let $I \subset \mathbb{R}$ be an interval. A map $c : I \to \mathbb{R}$ is called convex if $\forall x, y \in I, t \in [0, 1], c(tx + (1-t)y) \leq tc(x) + (1-t)c(y)$, or equivalently, $\forall x < t < y, x, t, y \in I$ we have

$$\frac{c(t) - c(x)}{t - x} \leq \frac{c(y) - c(t)}{y - t}$$

Easily any convex function is continuous, hence Borel measurable.

Lemma 4.2. *For $c : I \to \mathbb{R}$ convex and continuous and m in the interior of I, there exists $a, b \in \mathbb{R}$ such that $c(x) \geq ax + b$ for all $x \in I$ with equality when $x = m$.*

Proof. Set

$$a = \sup\left\{\frac{c(m)-c(x)}{m-x} : x < m, x \in I\right\}$$

which exists and is finite by convexity of c. Let $b = c(m) - am$ so that $c(m) = am + b$. For $y > m$, we have

$$\forall x < m, \frac{c(m)-c(x)}{m-x} \leq \frac{c(y)-c(m)}{y-m} \implies a \leq \frac{c(y)-c(m)}{y-m}$$

Rearranging gives $c(y) \geq ay + b$.
For $x < m$, we know by definition that

$$\frac{c(m)-c(x)}{m-x} \leq a \implies c(x) \geq ax + b$$

which completes the proof. □

Theorem 4.3 (Jensen's Inequality)**.** *Let X be an integrable random variable (i.e. $\mathbb{E}|X| < \infty$) taking values in $I \subset \mathbb{R}$. Let $c : I \to \mathbb{R}$ be convex, then $\mathbb{E}c(X)$ is well-defined and $\mathbb{E}c(X) \geq c(\mathbb{E}X)$*

Proof. Assume first that $m = \mathbb{E}X$ is in the interior of I, then by the preceding lemma we have $c(X) \geq aX + b$ and hence $\mathbb{E}c^-(X) \leq |a|\mathbb{E}|X| + b < \infty$ where as usual $c^- = -(c \wedge 0)$. So $\mathbb{E}c(X) = \mathbb{E}c^+(X) - \mathbb{E}c^-(X)$ is well-defined in $(-\infty, \infty]$. Integrating the preceding $c(X) \geq aX + b$ then gives $\mathbb{E}c(X) \geq a\mathbb{E}X + b = c(\mathbb{E}X)$ as desired.
If $\mathbb{E}X$ is on the boundary of I, then $X = \mathbb{E}X$ a.s., so $\mathbb{E}c(X) = c(\mathbb{E}X)$. □

For $1 \leq p < q < \infty$, if we set $c(x) = |x|^{q/p}$ then by Jensen's inequality

$$\|X\|_p = c(\mathbb{E}(|X|^p))^{1/q} \leq (\mathbb{E}c(|X|^p))^{1/q} = (\mathbb{E}(|X|^q))^{1/q} = \|X\|_q$$

So $L^q(\mathbb{P}) \subset L^p(\mathbb{P})$.

Theorem 4.4 (Hölder's Inequality)**.** *Let $p, q \in [1, \infty]$ be such that $p^{-1} + q^{-1} = 1$, then $\mu(|fg|) \leq \|f\|_p \|g\|_q$ for any $f, g : E \to \mathbb{R}$ measurable.*

Proof. If $\|f\|_p = \infty$ or $\|g\|_q = \infty$ then we are done. If $f = 0$ or $g = 0$ a.e. then we are done as well. Exclusin these cases, we can assume $\|f\|_p \neq 0$. WLOG $\|f\|_p = 1$, then

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

is a probability measure on E with density $|f|^p$. Then we have

$$\begin{aligned}
\mu(|fg|) &= \int_E |fg| \, \mathrm{d}\mu = \int_E |g| \frac{|f|^p}{|f|^{p-1}} 1_{|f|>0} \, \mathrm{d}\mu = \mathbb{E}\left[\frac{|g|}{|f|^{p-1}} 1_{|f|>0}\right] \\
&\leq \left(\mathbb{E}\left[\frac{|g|^q}{|f|^{q(p-1)}} 1_{|f|>0}\right]\right)^{1/q} = \left(\mathbb{E}\left[\frac{|g|^q}{|f|^p} 1_{|f|>0}\right]\right)^{1/q} \\
&= \left(\int_E \frac{|g|^q}{|f|^p} 1_{|f|>0} |f|^p \, \mathrm{d}\mu\right)^{1/q} \leq \mu(|g|^q)^{1/q} = \|g\|_q = \|f\|_p \|g\|_q
\end{aligned}$$

by (the discussed consequence of) Jensen's inequality. □

Finally,

Theorem 4.5 (Minkowski's Inequality)**.** *Let $p \in [1,\infty]$ and $f, g : E \to \mathbb{R}$ measurable, then $\|f+g\|_p \le \|f\|_p + \|g\|_p$.*

This is our desired triangle inequality for $\|\cdot\|_p$.

Proof. The cases $p = 1, \infty$ are clear. It is also clear if either $\|f+g\|_p = 0$, $\|f\|_p = \infty$, or $\|g\|_p = \infty$.
Assume otherwise, then $\|f+g\|_p \in (0,\infty)$, in particular $\mu(|f+g|^p) \in (0,\infty)$. For $p^{-1} + q^{-1} = 1$, we have $\||f+g|^{p-1}\|_q = \mu(|f+g|^p)^{1-p^{-1}} \in (0,\infty)$. By Hölder's inequality,

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

Since we assumed $\|f+g\|_p \in (0,\infty)$, we have $\|f+g\|_p = \|f+g\|_p^{p-p/q} \le \|f\|_p + \|g\|_p$ which is the desired inequality. □

We want to make L^p a normed vector space with respect to $\|\cdot\|_p$. Recall that a normed vector space is defined as

Definition 4.4. Let V be a (real) vector space. A map $V \to [0,\infty), v \mapsto \|v\|$ is called a norm if:
1. $\|u+v\| \le \|u\| + \|v\|$ for any $u, v \in V$.
2. $\|\alpha v\| = |\alpha|\|v\|$ for any $v \in V, \alpha \in \mathbb{R}$.
3. $\|v\| = 0$ if and only if $v = 0$.

What we've done cleared up the first condition; The second condition follows quite immmediately from the definition of $\|\cdot\|_p$. However, the third condition is not quite true: For $\|f\|_p = 0$ we only need $f = 0$ a.e. instead of $f = 0$. This means that we actually want to consider the new vector space

Definition 4.5. Define the equivalence relation $\sim$ by $f \sim g \iff f = g$ a.e.. We can define the quotient $\mathcal{L}^p(\mu) = L^p(\mu)/\sim$ which, as one can check easily, is a normed vector space under $[f]+[g] = [f+g], \alpha[f] = [\alpha f], \|[f]\|_p = \|f\|_p$ (for $f, g \in L^p, \alpha \in \mathbb{R}$).

The distinction of f and $[f]$, and that of L^p and $\mathcal{L}^p$, are often notationally surpressed. Unless otherwise specified, when we write f we automatically mean $[f]$ and when we write L^p we automatically mean $\mathcal{L}^p$.

4.2 Completeness of $\mathcal{L}^p$

Definition 4.6. A normed vector space $(V, \|\cdot\|)$ is complete if every Cauchy sequence in it converges. A complete normed vector space is called a Banach space.

Theorem 4.6. *$\mathcal{L}^p$ is a Banach space.*

Proof. The case $p = \infty$ is easy and left as exercise. Let $(f_n)_n$ be a Cauchy sequence, then we can find a subsequence n_k such that $\|f_{n_{k+1}} - f_{n_k}\|_p \leq 2^{-k}$, so $S = \sum_{k=1}^{\infty} \|f_{n_{k+1}} - f_{n_k}\|_p < \infty$. Consequently,

$$\forall K \in \mathbb{N}, \left\| \sum_{k=1}^{K} |f_{n_{k+1}} - f_{n_k}| \right\|_p \leq \sum_{k=1}^{K} \|f_{n_{k+1}} - f_{n_k}\|_p \leq S$$

Therefore $\left\|\sum_{k=1}^{\infty} |f_{n_{k+1}} - f_{n_k}|\right\|_p \leq S$ by Theorem 3.1. In particular, we have $\sum_{n=1}^{\infty} |f_{n_{k+1}} - f_{n_k}| < \infty$ a.e., hence the series

$$\sum_{k=1}^{K} (f_{n_{k+1}}(x) - f_{n_k}(x)) = f_{n_{K+1}}(x) - f_{n_0}(x)$$

converges absolutely as $K \to \infty$ for $x \in A$ where $\mu(E - A) = 0$. But $f_{n_0}(x)$ is fixed, so for $x \in A$, $f_{n_{K+1}}(x)$ tends to some value as $K \to \infty$. Set that to be $f(x)$, and set $f(x) = 0$ for $x \notin A$, then $f_{n_k} \to f$ a.e.. Now let $\epsilon > 0$ be fixed. Choose N large enough such that $\|f_n - f_m\|_p^p < \epsilon$ for all $m, n \geq \mathbb{N}$. Then by Lemma 3.5,

$$\begin{aligned} \|f_n - f\|_p^p = \mu(|f_n - f|^p) &= \mu\left(\lim_{k\to\infty} |f_n - f_{n_k}|^p\right) \\ &= \mu\left(\liminf_{k\to\infty} |f_n - f_{n_k}|^p\right) \leq \liminf_{k\to\infty} \mu(|f_n - f_{n_k}|^p) < \epsilon \end{aligned}$$

whenever $n \geq N$. f is then in L^p since $\|f\|_p \leq \|f - f_N\|_p + \|f_N\|_p < \infty$ for large enough N. We then have $f_n \to f$ in L^p, as desired. □

Remark. In particular, if the $\|\cdot\|_1$-Cauchy sequence f_n is in $C[0,1]$ or $V = \{f : [0,1] \to \mathbb{R} \text{ simple}\}$, then the theorem means that there exists $f : L^1(\mu)$ such that $f_n \to f$ in $L^1(\mu)$. Conversely, one can also show that $C[0,1]$ and V are both dense in $L^1(\mu)$. In this sense, $L^1(\mu)$ is the completion of $C[0,1]$ (and also V) under the $\|\cdot\|_1$ norm. This gives a satisfactory answer of "what should the space of integrable functions $[0,1] \to \mathbb{R}$ be".

4.3 $\mathcal{L}^2$ as a Hilbert Space

Definition 4.7. A symmetric bilinear map $(v, w) \mapsto \langle v, w\rangle, V \times V \to \mathbb{R}$ on a vector space V is called an inner product if $\langle v, v\rangle \geq 0$ with equality holds iff $v = 0$.
A vector space equipped with an inner product is called a Hilbert space if it is complete as a normed vector space under the norm $\|v\| = \sqrt{\langle v, v\rangle}$.

Note that the triangle inequality for $\|\cdot\|$ is due to the Cauchy-Schwartz inequality.
We can then make $\mathcal{L}^2(\mu)$ a Hilbert space by giving it the inner product

$$\langle f, g\rangle = \int_E fg \, \mathrm{d}\mu$$

We have $\|f+g\|_2^2 = \|f\|_2^2 + 2\langle f, g\rangle + \|g\|_2^2$. This motivates the notion of orthogonality: We say f is orthogonal to g in $\mathcal{L}^2$ if $\langle f, g\rangle = 0$.
We also jahve the parallelogram law: $\|f+g\|_2^2 + \|f-g\|_2^2 = 2(\|f\|_2^2 + \|g\|_2^2)$.
Note that these properties follow directly from the fact that $\mathcal{L}^2$ is equipped with an inner product. The fact of $\mathcal{L}^2$ being a Hilbert space allows us to do even more stuff:

Definition 4.8. Let $\mathcal{H}$ be a Hilbert space and $V \subset \mathcal{H}$ (which we almost always set to be a subspace), we define its orthogonal complement to be $V^\perp = \{f \in \mathcal{H} : \forall v \in V, \langle f, v\rangle = 0\}$. A subset $V \subset \mathcal{H}$ is closed if any $f_n \in V$ with $f_n \to f$ in $\mathcal{H}$ has $f \in V$.

Theorem 4.7 (Orthogonal Projection)**.** *Let $V \subset \mathcal{H}$ be a closed subspace, then for each $f \in \mathcal{H}$, there exists a unique orthogonal decomposition $f = v + u, v \in V, u \in V^\perp$. Furthermore, such that $\|f - v\| \le \|f - g\|$ for all $g \in V$ with equality if and only if $g = v$.*

We call v the orthogonal projection of f onto V.

Proof. Define the "minimal distance" between f and V as $d(f, V) = \inf_{g\in V} \|f - g\|$. Take a sequence $g_n \in V$ such that $\|f - g_n\| \to d(f, V)$ as $n \to \infty$. By the parallelogram law,

$$\begin{aligned} 2\|f-g_n\|^2 + 2\|f-g_m\|^2 &= \left\| 2\left(f - \frac{g_m+g_n}{2}\right)\right\|^2 + \|g_n - g_m\|^2 \\ &\le 4d(f,V)^2 + \|g_n - g_m\|^2 \end{aligned}$$

So

$$0 \le \|g_n - g_m\|^2 \le 2\|f-g_n\|^2 + 2\|f-g_m\|^2 - 4d(f,V)^2$$

which can be made arbitrarily small by increasing $\min\{n, m\}$. Hence $(g_n)_n$ is Cauchy and therefore converges to some $g \in \mathcal{H}$ by completeness. $g \in V$ as V is closed. As $\|\cdot\|$ is continuous in itself, we have $\|g - f\| = d(f, V)$, so we can take $v = g$. Also, for any h, the map $F : t \mapsto \|f - (v + th)\|^2, t \in \mathbb{R}, h \in V$ is minimised at $t = 0$. So necessarily $0 = F'(0) = -2\langle f - v, h\rangle$. But h is arbitrary, so $u = f - v \in V^\perp$, i.e. $f = v + u$ is an orthogonal composition.
As for uniqueness, suppose $f = w + z$ for some other $w \in V, z \in V^\perp$, then $v - w + u - z = 0$, so $0 = \|v - w + u - z\|^2 = \|v - w\|^2 + \|u - z\|^2 \implies v = w, u = z$. □

We can interpret the structure of $\mathcal{L}^2(\mu)$ more concretely in the case where $(E, \mathcal{E}, \mu) = (\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

Definition 4.9. The covariance of random variables X, Y in $L^2(\mathbb{P})$ is defined as $\operatorname{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)) = \langle X - \mathbb{E}X, Y - \mathbb{E}Y\rangle$. The variance of a random variable $X \in L^2(\mathbb{P})$ is $\operatorname{var}(X) = \operatorname{cov}(X, X)$.

We declare $X \perp Y$ in $\mathcal{L}^2(\mathbb{P})$ if $\mathbb{E}X = \mathbb{E}Y = \langle X, Y\rangle = 0$.

Definition 4.10. If $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$ generated by a countable family $(G_i)_{i\in I} \in \mathcal{F}$ that partitions Ω, then $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ would be a closed subspace of $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. The conditional expectation of a random variable $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ given $\mathcal{G}$ is defined as

$$Y = \sum_{i\in I} \mathbb{E}(X|G_i)1_{G_i}, \mathbb{E}(X|G_i) = \frac{\mathbb{E}(X1_{G_i})}{\mathbb{P}(G_i)}$$

Note that if $\mathbb{P}(G_i) = 0$ then $1_{G_i} = 0$ a.e., so we simply remove that term from the sum.

Proposition 4.8. *The distribution of Y coincides with that of the orthogonal projection of X onto $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$.*

Proof. Omitted but good exercise. □

4.4 Convergence in $\mathcal{L}^1(\mathbb{P})$

In this section, we will focus on $\mathcal{L}^1 = \mathcal{L}^1(\mathbb{P}) = \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, but everything generalises easily to any finite measure space.

Theorem 4.9. *Let $(X_n)_n$ be random variables such that $X_n \to^{\mathbb{P}} X$ as $n \to \infty$ and $|X_n| \leq C$ uniformly for some $C \in \mathbb{R}_{\geq 0}$ (consequently $(X_n)_n \in \mathcal{L}^1$), then $X_n \to X$ in $\mathcal{L}^1$.*

Proof. We know that $X_{n_k} \to X$ a.s. by Theorem 2.5 for some subsequence X_{n_k} of X_n. So $|X| = \lim_{n\to\infty} |X_{n_k}| \leq C$ a.s., therefore $X \in \mathcal{L}^1$. Now for any $\epsilon > 0$, $\mathbb{E}|X_n - X| = \mathbb{E}(|X_n - X|1_{|X_n - X| > \epsilon/2}) + \mathbb{E}(|X_n - X|1_{|X_n - X| \leq \epsilon/2}) \leq 2C\mathbb{P}(|X_n - X| > \epsilon/2) + \epsilon/2 < \epsilon$ whenever n is large enough. □

Lemma 4.10. *Let $X \in \mathcal{L}^1$ and*

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

Then $I_X(\delta) \downarrow 0$ (i.e. $I_X(\delta)$ decreases to 0) as $\delta \downarrow 0$.

Proof. Suppose not, then there is some $\epsilon > 0$ and some $(A_n)_n \in \mathcal{F}$ such that $\mathbb{E}(|X|1_{A_n}) \geq \epsilon > 0$ and $\sum_n \mathbb{P}(A_n) < \infty$. By Lemma 1.7 we have $\mathbb{P}(A_n \text{ i.o.}) = 0$, so Theorem 3.6 gives

$$\epsilon \leq \mathbb{E}(|X|1_{\bigcup_{m\geq n} A_m}) \to \mathbb{E}(|X|1_{A_n \text{ i.o.}}) = 0$$

A contradiction. □

What if we require this limit to hold uniformly in a certain collection $\mathscr{X} \subset \mathcal{L}^1$ of random variables? We can define

$$I_{\mathscr{X}}(\delta) = \sup_{X\in\mathscr{X}} I_X(\delta) = \sup\{\mathbb{E}(|X|1_A) : X \in \mathscr{X}, A \in \mathcal{F}, \mathbb{P}(A) \leq \delta\}$$

where it is clear that $I_{\mathscr{X}}(1) < \infty$ iff $\mathscr{X}$ is bounded in $\mathcal{L}^1$.

Definition 4.11. $\mathscr{X} \subset \mathcal{L}^1$ is uniformly integrable if it is bounded in $\mathcal{L}^1$ and $I_{\mathscr{X}}(\delta) \downarrow 0$ as $\delta \downarrow 0$.

Recall that $\|X\|_p \leq \|X\|_q$ whenever $p \leq q$ in a probability space.

Lemma 4.11. *If $\mathscr{X}$ is bounded in $\mathcal{L}^p(\mathbb{P})$ for some $1 < p \leq \infty$, then $\mathscr{X}$ is uniformly integrable.*

Proof. Let q be such that $p^{-1} + q^{-1} = 1$. Then for any $A \in \mathcal{F}$ with $\mathbb{P}(A) \leq \delta$, we have

$$\mathbb{E}(|X|1_A) \leq \|X\|_p(\mathbb{P}(A))^{1/q} \leq \|X\|_p\delta^{1/q}$$

by Hölder's inequality. □

Remark. This is however not true for $p = 1$. A counterexample can be found by taking $\mathbb{P}$ as the Lebesgue measure on $(0,1)$ and $\mathscr{X} = \{X_n = n1_{(0,1/n)}\}$.

Lemma 4.12. *$\mathscr{X}$ is uniformly integrable if and only if* $\sup\{\mathbb{E}(|X|1_{|X|>K}) : X \in \mathscr{X}\} \to 0$ *as* $K \to \infty$.

Proof. For the "only if" direction, for any $\epsilon > 0$, choose δ such that $I_{\mathscr{X}}(\delta) < \epsilon$ and K large enough such that $I_{\mathscr{X}}(1) \leq K\delta$. By Markov's inequality, $\mathbb{P}(|X| > K) \leq (\mathbb{E}|X|)/K \leq I_{\mathscr{X}}(1)/K \leq \delta$. Therefore $\mathbb{E}(|X|1_{|X|>K}) \leq I_{\mathscr{X}}(\delta) < \epsilon$.
Conversely, again fix $\epsilon > 0$. Take K such that $\sup_{X\in\mathscr{X}} \mathbb{E}(|X|1_{|X|>K}) < \epsilon/2$, then $\mathscr{X}$ is bounded in $\mathcal{L}^1$ since

$$\mathbb{E}|X| \leq \mathbb{E}(|X|1_{|X|\leq K}) + \mathbb{E}(|X|1_{|X|>K}) \leq K + \epsilon/2 < \infty$$

for any $X \in \mathscr{X}$. Also, if $A \in \mathcal{F}$ has $\mathbb{P}(A) \leq \delta$, then

$$\mathbb{E}(|X|1_A) = \mathbb{E}(|X|1_A 1_{|X|>K}) + \mathbb{E}(|X|1_A 1_{|X|\leq K}) < \frac{\epsilon}{2} + K\mathbb{P}(A) \leq \frac{\epsilon}{2} + K\delta < \epsilon$$

for $\delta < \epsilon/(2K)$. □

Theorem 4.13. *Let $X, (X_n)_n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, then the followings are equivalent:*
(a) $X, (X_n)_n \in \mathcal{L}^1$, and $X_n \to X$ in $\mathcal{L}^1$.
(b) $\mathscr{X} = (X_n)_n$ is uniformly integrable and $X_n \to^{\mathbb{P}} X$.

Proof. (a) $\implies$ (b): By Markov's inequality, for any $\epsilon > 0$, we have $\mathbb{P}(|X_n - X| > \epsilon) \leq \epsilon^{-1}\mathbb{E}|X_n - X| \to 0$ as $n \to \infty$, so $X_n \to^{\mathbb{P}} X$. As for uniform integrability, fix $\epsilon > 0$ and choose $N = N_\epsilon$ such that $\mathbb{E}|X_n - X| < \epsilon/2$ for all $n > N$. Observe that any finite collection of random variables have to be uniformly integrable, so there is some $\delta > 0$ such that whenever $A \in \mathcal{F}$ has $\mathbb{P}(A) \leq \delta$, we have $\mathbb{E}(|X|1_A) < \epsilon/2, \mathbb{E}(|X_n|1_A) < \epsilon/2$ whenever $n \leq N$. Then $\mathbb{E}(|X_n|1_A) \leq \mathbb{E}(|X_n - X|1_A) + \mathbb{E}(|X|1_A) < \epsilon/2 + \epsilon/2 = \epsilon$ whenever $n > N$ which implies the uniform integrability of $\mathscr{X}$.
(b) $\implies$ (a): By Theorem 2.5, we know that there is a subsequence $(X_{n_k})_k$ of $(X_n)_n$ such that $X_{n_k} \to X$ a.s. as $k \to \infty$. By Fatou's lemma,

$$\mathbb{E}|X| = \mathbb{E}\liminf_{k\to\infty} |X_{n_k}| \leq \liminf_{k\to\infty} \mathbb{E}|X_{n_k}| \leq I_{\mathscr{X}}(1) < \infty$$

So $X \in \mathcal{L}^1$. For $K > 0$, we take $X_n^K = (-K)\vee(X_n \wedge K), X^K = (-K)\vee(X \wedge K)$, then $X_n^K \to^{\mathbb{P}} X^K$. By Theorem 4.9, as $|X_n^K| \leq K$, we know that $X_n^K \to X^K$ in $\mathcal{L}^1$. Now,

$$\mathbb{E}|X_n - X| \leq \mathbb{E}|X_n - X_n^K| + \mathbb{E}|X_n^K - X^K| + \mathbb{E}|X^K - X|$$

which can be made arbitrarily small for large n, K due to Lemma 4.12. Consequently, $X_n \to X$ in $\mathcal{L}^1$. □

5 Fourier Transforms

In this section, we will work exclusively with $L^p = L^p(\mathbb{R}^d, \mathcal{B}^{\otimes d}, \mu^{\otimes d}), 1 \leq p < \infty$ on which we apply a slight modification and say

$$L^p = \left\{ f : \mathbb{R}^d \to \mathbb{C} \text{ measurable} : \|f\|_p^p = \int_{\mathbb{R}^d} |f(x)|^p \, \mathrm{d}x < \infty \right\}$$

in order to include all complex-valued measurable functions. Here (and thereafter), we use the shorthand $\mathrm{d}x = \mathrm{d}\mu^{\otimes d}(x)$. One can easily check that all results we have obtained on L^p-spaces still apply. In addition, L^p defined in this way also has the additional structure of a complex vector space.
For a complex-valued measurable function $\mathbb{R}^d \to \mathbb{C}$, one can define its integral as

$$\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \operatorname{Re} f(x)\,\mathrm{d}x + i\int_{\mathbb{R}^d} \operatorname{Im} f(x)\,\mathrm{d}x$$

Given that both integrals in the right hand side are well-defined (which will happen if $f \in L^1$). We also have the inequality

$$\left|\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x\right| \le \int_{\mathbb{R}^d} |f(x)|\,\mathrm{d}x$$

on all these L^p and the inner product

$$\langle f, g\rangle_{L^2} = \int_{\mathbb{R}^d} f(x)\overline{g(x)}\,\mathrm{d}x$$

on L^2. Recall also that $\mu^{\otimes d}$ is translation-invariant for any d.

5.1 Definitions; Convolution

Fourier had the idea of writing "nice" functions in the form $f = \sum_k c_k(f)e^{ikx}$, which looks as if one is doing an orthogonal decomposition of f in $L^2((0, 2\pi))$. A continuous, and higher dimensional, analogue of this will be our object of enquiry in this section.

Definition 5.1. Given $f \in L^1(\mathbb{R}^d)$, we define its Fouier transform

$$\hat{f} : \mathbb{R}^d \to \mathbb{C}$$

by

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

Since $|e^{i\langle u,x\rangle}| = 1$, we see $|\hat{f}(u)| \le \|f\|_1$ for all u. Also, if $u_n \to u$ in $\mathbb{R}^d$, then $\hat{f}(u_n) \to \hat{f}(u)$ by Theorem 3.6, so $\hat{f} : \mathbb{R}^d \to \mathbb{C}$ is a bounded continuous function. We will prove the following results:

Theorem 5.1 (Fourier Inversion)**.** *Suppose $f, \hat{f} \in L^1$, then*

$$f(x) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \hat{f}(u)e^{-i\langle x,u\rangle}\,\mathrm{d}u$$

almost everywhere.

Remark. 1. The identity holds everywhere f is continuous, and as one can check there is at most one continuous $f \in [f]$.
2. Consequently, $\hat{f} = 0$ a.e., then $f = 0$ a.e..

Theorem 5.2 (Plancherel's Theorem)**.** *For $f \in L^1 \cap L^2$,*

$$\|\hat{f}\|_2 = (2\pi)^{d/2}\|f\|_2$$

We will also prove the Central Limit Theorem using Fourier transforms. But first, let's get some more definitions.

Definition 5.2. For a probability measure μ on $\mathbb{R}^d$, we define its Fourier transform as

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

We have $|\hat{\mu}(u)| \leq \mu(\mathbb{R}^d) < \infty$. Also, by Proposition 3.9, if $\mu = \nu_f$ arises from the pdf f and the Lebesgue measure on $\mathbb{R}^d$, then $\hat{\nu}_f = \hat{f}$. Closely related to this is the characteristic function of a random variable on $\mathbb{R}^d$ given by $\phi_X(u) = \mathbb{E}e^{i\langle u,X\rangle} = \hat{\mu}_X(u)$ where μ_X is the distribution of X.
A key concept in Fourier analysis is the notion of convolution of functions and measures.

Definition 5.3. For $f \in L^p, 1 \leq p < \infty$ and ν a probability measure on $\mathbb{R}^d$, we set

$$\nu * f(x) = f * \nu(x) = \int_{\mathbb{R}^d} f(x-y) \, \mathrm{d}\nu(y) = \nu(f(x-\cdot))$$

By Jensen's inequality and Theorem 3.13,

$$\begin{aligned}\int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} |f(x-y)| \, \mathrm{d}\nu(y) \right)^p \mathrm{d}x &\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |f(x-y)|^p \, \mathrm{d}\nu(y) \, \mathrm{d}x \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |f(x-y)|^p \, \mathrm{d}x \, \mathrm{d}\nu(y) = \|f\|_p^p\end{aligned}$$

So $\|f * \nu\|_p \leq \|f\|_p$, in particular $f * \nu \in L^p$ and hence is finite a.e..

Definition 5.4. When $\nu = \nu_g$ has a density $g \in L^1$, then $\mathrm{d}\nu(x) = g(x) \, \mathrm{d}x$ and we write $f * g$ for $f * \nu_g$.
For two probability measures μ, ν on $\mathbb{R}^d$, we define their convolution to be the new measure

$$\mu * \nu(A) = \int_{\mathbb{R}^d \times \mathbb{R}^d} 1_A(x+y) \, \mathrm{d}\mu(x) \, \mathrm{d}\nu(y) = \mu \otimes \nu(X + Y \in A)$$

with $X \sim \mu, Y \sim \nu$ independent.

Again, by Fubini's theorem, if μ has density $f \in L^1$ then

$$\mu * \nu(A) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} 1_A(x-y) f(x) \, \mathrm{d}x \, \mathrm{d}\nu(y) = \int_{\mathbb{R}^d} 1_A(x)(f * \nu)(x) \, \mathrm{d}x$$

So $\mu * \nu$ has density $f * \nu \in L^1$. A key observation here is that for $f \in L^1$,

$$\begin{aligned}\widehat{f * \nu}(u) &= \int_{\mathbb{R}^d} (f * \nu)(x) e^{i\langle u,x\rangle} \, \mathrm{d}x = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(x-y) e^{i\langle u, x-y+y\rangle} \, \mathrm{d}\nu(y) \, \mathrm{d}x \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(x-y) e^{i\langle u,x-y+y\rangle} \, \mathrm{d}\nu(y) \, \mathrm{d}x \\ &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} f(x-y) e^{i\langle u,x-y\rangle} \, \mathrm{d}x \right) e^{i\langle u,y\rangle} \, \mathrm{d}\nu(y) \\ &= \hat{f}(u) \int_{\mathbb{R}^d} e^{i\langle u,y\rangle} \, \mathrm{d}\nu(y) = \hat{f}(u)\hat{\nu}(u)\end{aligned}$$

by Fubini's theorem. Likewise, suppose $X \sim \mu, Y \sim \nu$ are independent, then

$$\widehat{\mu * \nu} = \mathbb{E}(e^{i\langle u, X+Y\rangle}) = \mathbb{E}(e^{i\langle u,X\rangle})\mathbb{E}(e^{i\langle u,Y\rangle}) = \hat{\mu}(u)\hat{\nu}(u)$$

5.2 The Gaussians and Fourier Inversion Formula

Recall from probability that the density of a Gaussian $N(0,t)$ random variable on $\mathbb{R}$ is $g_t(x) = (2\pi t)^{-1/2} e^{-|x|^2/(2t)}$. In particular, for a standard Gaussian $X \sim N(0,1)$, the characteristic function $\phi_X(u) = \mathbb{E}e^{iuX}$ satisfies

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}u}\phi_X(u) &= \frac{1}{\sqrt{2\pi}}\frac{\mathrm{d}}{\mathrm{d}u}\int_{\mathbb{R}} e^{iux}e^{-x^2/2}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{iux}ixe^{-x^2/2}\,\mathrm{d}x \\ &= -u\int_{\mathbb{R}} e^{iux}\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x = -u\phi_X(u)\end{aligned}$$

Consequently $\phi_X(u) = \phi_X(0)e^{-u^2/2} = e^{-u^2/2}$. So in dimension $d = 1$ we have $\hat{g}_1 = \sqrt{2\pi}g_1$. This turns out to generalise to higher dimensions, and they made the Gaussians occupy a special place in probability theory.
Let $Z_1, \ldots, Z_d$ be i.i.d. $N(0,1)$ random variables, then we can consider the random variable $Z = (Z_1, \ldots, Z_d)$ on $\mathbb{R}^d$. Then, for $t > 0$, the random variable $\sqrt{t}Z$ (whose distribution we shall denote as $N(0, tI_d)$) has pdf

$$g_t(x) = \frac{1}{(2\pi t)^{d/2}}\exp\left(-\frac{|x|^2}{2t}\right)$$

The Fourier transform of this is

$$\hat{g}_t(u) = \mathbb{E}e^{i\langle u, \sqrt{t}Z\rangle} = \mathbb{E}\prod_{j=1}^{d} e^{iu_j\sqrt{t}Z_j} = \prod_{j=1}^{d}\mathbb{E}e^{iu_j\sqrt{t}Z_j} = \prod_{j=1}^{d} e^{-u_j^2 t/2} = e^{-|u|^2 t/2}$$

In other words, $\hat{g}_t = (2\pi)^{d/2}t^{-d/2}g_{1/t}$. Applying Fourier transform again gives $\hat{\hat{g}}_t = (2\pi)^d g_t$, in other words

$$g_t = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\hat{g}(u)e^{-i\langle u,x\rangle}\,\mathrm{d}u$$

Hence Theorem 5.1 holds for Gaussians. This fact shall assist us in proving the theorem in general.

Definition 5.5. A Gaussian convolution of $f \in L^1(\mathbb{R}^d)$ is

$$f * g_t = \int_{\mathbb{R}^d} f(\cdot - y)g_t(y)\,\mathrm{d}y$$

$f * g_t$ is continuous on $\mathbb{R}^d$ with $\|f * g_t\|_1 \le \|f\|_1$ as g_t is a density. Also,

$$|f * g_t(x)| \le \frac{1}{(2\pi t)^{d/2}}\int_{\mathbb{R}^d}|f(x-y)|e^{-|x|^2/2t}\,\mathrm{d}x \le \frac{\|f\|_1}{(2\pi t)^{d/2}}$$

So $f * g_t \in L^1 \cap L^\infty$. On the other hand, $\widehat{f * g_t}(u) = \hat{f}(u)e^{-|u|^2 t/2}$, so

$$\|\widehat{f * g_t}\|_1 \le \|f\|_1\int_{\mathbb{R}^d} e^{-|u|^2 t/2}\,\mathrm{d}u \le \left(\frac{2\pi}{t}\right)^{d/2}\|f\|_1$$

and $\|\widehat{f * g_t}\|_\infty \le \|f\|_1$. This means that $\widehat{f * g_t} \in L^1 \cap L^\infty$ as well.
For any probability measure μ, by writing $g_t = g_{t/2} * g_{t/2}$ we see $\mu * g_t = \mu * g_{t/2} * g_{t/2}$ is also a Gaussian convolution by associativity of convolution product (since $\mu * g_{t/2}$ is now an L^1 function).

Lemma 5.3. *Theorem 5.1 holds for Gaussian convolutions.*

Proof. For $f \in L^1$ and $t > 0$, by Fourier inversion for g_t and Fubini's theorem,

$$\begin{aligned}(2\pi)^d f * g_t(x) &= (2\pi)^d \int_{\mathbb{R}^d} f(x-y) g_t(y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(x-y)\hat{g}_t(u)e^{-i\langle u,y\rangle}\,\mathrm{d}u\,\mathrm{d}y \\ &= \int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} f(x-y)e^{i\langle u,x-y\rangle}\,\mathrm{d}y\right)\hat{g}_t(u)e^{-i\langle u,x\rangle}\,\mathrm{d}u \\ &= \int_{\mathbb{R}^d} \hat{f}(u)\hat{g}_t(u)e^{-i\langle u,x\rangle}\,\mathrm{d}u = \int_{\mathbb{R}^d} \widehat{f * g_t}(u)e^{-i\langle u,x\rangle}\,\mathrm{d}u\end{aligned}$$

as desired. □

How would we push this further? The idea is to use the observation that $f * g_t \approx f$ as $t \to 0$, so the set of Gaussian convolutions should be dense in L^1. Indeed,

Lemma 5.4. *For $f \in L^p(\mathbb{R}^d)$, we have $\|f * g_t - f\|_p \to 0$ as $t \to \infty$.*

Proof. From example sheet, we know that for all $\epsilon > 0$, there is some continuous and compactly supported $h : \mathbb{R}^d \to \mathbb{C}$ such that $\|f - h\|_p < \epsilon$. Then by linearity of convolution, we see that $\|f * g_t - h * g_t\|_p = \|(f-h) * g_t\|_p \le \|f - h\|_p$. Next, we define

$$e(y) = \int_{\mathbb{R}^d} |h(x-y) - h(x)|^p\,\mathrm{d}x$$

which satisfies $0 \le e(y) \le 2^p\|h\|_p^p$ for all y. $e(y) \to 0$ as $y \to 0$ by Theorem 3.6. Since g_t is a density, Jensen's inequality and Theorem 3.13 gives

$$\begin{aligned}\|h * g_t - h\|_p^p &= \int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d}(h(x-y)-h(x))g_t(y)\,\mathrm{d}y\right|^p\,\mathrm{d}x \le \int_{\mathbb{R}^d} e(y)g_t(y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}^d} e(\sqrt{t}z)g_1(z)\,\mathrm{d}z \to 0\end{aligned}$$

as $t \to 0$ again by Theorem 3.6. Finally, by Minkowski's inequality, we have

$$\|f * g_t - f\|_p \le \|f * g_t - h * g_t\|_p + \|h * g_t - h\|_p + \|h - f\|_p$$

which can be arbitrarily small for suitable choices of ϵ and t. □

Proof of Theorem 5.1. Set

$$\begin{aligned}f_t(x) &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \hat{f}(u)e^{-|u|^2t/2}e^{-i\langle u,x\rangle}\,\mathrm{d}u \\ &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \widehat{f * g_t}(u)e^{-i\langle u,x\rangle}\,\mathrm{d}u = f * g_t\end{aligned}$$

By the preceding lemma, $\|f_t - f\|_1 \to 0$ as $t \to 0$. Using the idea in the proof of Theorem 4.6, we know that there is a sequence $t_n \to 0$ such that $f_{t_n} \to f$ a.e.. Meanwhile, $\hat{f}(u)e^{-|u|^2t/2} \to \hat{f}(u)$ as $t \to \infty$. By Theorem 3.6,

$$f_t \to \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \hat{f}(u)e^{-i\langle x,u\rangle}\,\mathrm{d}u$$

So f_{t_n} converges to both sides of the desired equality a.e., hence the theorem. □

Proof of Theorem 5.2. First assume that $\hat{f} \in L^1$, then by Theorem 5.1 we have $f, \hat{f} \in L^1 \cap L^\infty$, so $f, \hat{f} \in L^2$. The function $(x,u) \mapsto f(x)\hat{f}(u)$ is measurable on $\mathbb{R}^d \times \mathbb{R}^d$, so we can write (by Theorem 5.1 and Theorem 3.13)

$$\begin{aligned}(2\pi)^d \|f\|_2^2 = (2\pi)^d \int_{\mathbb{R}^d} f(x)\overline{f(x)}\,\mathrm{d}x &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \hat{f}(u) e^{-i\langle u,x\rangle}\,\mathrm{d}u \right) \overline{f(x)}\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} \hat{f}(u) \overline{\int_{\mathbb{R}^d} f(x) e^{i\langle u,x\rangle}\,\mathrm{d}x}\,\mathrm{d}u = \int_{\mathbb{R}^d} \hat{f}(u)\overline{\hat{f}(u)}\,\mathrm{d}u = \|\hat{f}\|_2^2\end{aligned}$$

To extend this to the general case, we take the Gaussian convolution $f_t = f * g_t$. As $f_t \to f$ in L^2 by Lemma 5.4, $\|f_t\|_2 \to \|f\|_2$. This means that $\|f_t\|_2 \to \|f\|_2$ as $t \to 0$. Also, $|\hat{f}_t| = |\hat{f}|e^{-|u|^2 t/2} \uparrow |\hat{f}|$, so by Theorem 3.1 $\|\hat{f}_t\|_2^2 \uparrow \|f\|_2^2$ as $t \to 0$. Now both f_t and $\hat{f}_t$ are both in L^1 for any t, so

$$(2\pi)^d \|f\|_2^2 = \lim_{t\to\infty} (2\pi)^d \|f_t\|_2^2 = \lim_{t\to\infty} \|\hat{f}_t\|_2^2 = \|\hat{f}\|_2^2$$

as desired. □

6 Limit Theorems

6.1 Weak Convergence and Characteristic Functions

Definition 6.1. We say a set of Borel probability measures $(\mu_n)_n$ on $\mathbb{R}^d$ converge weakly to a Borel probability measure μ on $\mathbb{R}^d$ if $\mu_n(f) \to \mu(f)$ for any bounded continuous $f : \mathbb{R}^d \to \mathbb{R}$.
If $X_n \sim \mu_n, X \sim \mu$ are random variables (with range in $\mathbb{R}^n$), we say $X_n \to X$ weakly.

In example sheet, you will show that one can replace "bounded and continuous" by "bounded and Lipschitz". Also, if μ, ν are Borel probability measures on $\mathbb{R}^d$ with $\mu(f) = \nu(f)$ for all continuous bounded $f : \mathbb{R}^d \to \mathbb{R}$, then $\mu = \nu$. You will also show that if $d = 1$, then $\mu_{X_n} \to \mu_X$ weakly iff $X_n \to^{\mathrm{d}} X$ as $n \to \infty$. In fact, since every probability measure (on $\mathbb{R}^d$) is in particular a Radon measure, it must concentrate its mass on compact subsets of $\mathbb{R}^d$. From here, one can show that $\mu_n \to \mu$ weakly iff $\mu_n(f) \to \mu(f)$ for any infinitely differentiable $f : \mathbb{R}^d \to \mathbb{R}$ with compact support.

Theorem 6.1 (Lévy's Criterion)**.** *(a) Let X be a random variable with range in $\mathbb{R}^d$, then its distribution μ_X is uniquely determined by $\phi_X = \hat{\mu}_X$. Also, if $\phi_X \in L^1$, then μ_X has a pdf in the form*

$$f_X(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \phi_X(u) e^{-i\langle u,x\rangle}\,\mathrm{d}u$$

(b) If $(X_n)_n$ are random variables in $\mathbb{R}^d$ such that $\phi_{X_n} \to \phi_X$ pointwise, then $X_n \to X$ weakly.

Proof. (a) Take $Z \sim N(0, I_d)$ (with density g_1) independent of everything (this is possible since we can replace $(\Omega, \mathcal{F}, \mathbb{P})$ by $(\Omega, \mathcal{F}, \mathbb{P}) \times ((0,1), \mathcal{B}, \mu)$ WLOG as the statement is only about distributions of random variables instead of

the random variables themselves as functions). Then $\sqrt{t}Z$ has density g_t and $X + \sqrt{t}Z$ has density $f_t = \mu_X * g_t \in L^1$. By Theorem 5.1,

$$f_t(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \phi_X(u) e^{-|u|^2 t/2} e^{-i\langle u, x\rangle} \, \mathrm{d}u$$

which depends only on the characteristic function ϕ_X of X. Theorem 4.9 (or alternatively Theorem 3.6) shows that for any continuous bounded $g : \mathbb{R}^d \to \mathbb{R}$ we have

$$\int_{\mathbb{R}^d} g(x) f_t(x) \, \mathrm{d}x = \mathbb{E}g(X + \sqrt{t}Z) \to \mathbb{E}g(X) = \int_{\mathbb{R}^d} g(x) \, \mathrm{d}\mu_X(x)$$

as $t \to 0$. By uniqueness of limits, we know that ϕ_X determines $\mu_X(g)$ for all bounded continuous g, hence μ_X.
Next, if $\phi_X \in L^1$ then f_X is necessarily well-defined with $f_t \to f_X$ pointwise by Theorem 3.6. Also $|f_t| \leq (2\pi)^{-d} \|\phi_X\|_1$ and $f_t = \mu_X * g_t \geq 0$, so f_X has range in $\mathbb{R}_{\geq 0}$. For any bounded continuous g with compact support,

$$\int_{\mathbb{R}^d} g \, \mathrm{d}\mu_X = \lim_{t \to \infty} \int_{\mathbb{R}^d} g(x) f_t(x) \, \mathrm{d}x = \int_{\mathbb{R}^d} g(x) f_X(x) \, \mathrm{d}x$$

by Theorem 3.6. This implies that f_X is the pdf of μ_X.
(b) Let $(X_n)_n$ be such that $\phi_{X_n}(u) \to \phi_X(u)$. Suppose $g : \mathbb{R}^d \to \mathbb{R}$ is compactly supported and Lipschitz (say with Lipschitz constant $\|g\|_{\mathrm{Lip}}$), then in particular it's bounded and continuous. For $\epsilon > 0$, choose $t > 0$ small enough such that $\sqrt{t}\|g\|_{\mathrm{Lip}} \mathbb{E}|Z| \leq \epsilon/3$. Then $\mathbb{E}|g(X_n + \sqrt{t}Z) - g(X_n)| \leq \sqrt{t}\|g\|_{\mathrm{Lip}} \mathbb{E}|Z| \leq \epsilon/3$ and similarly $\mathbb{E}|g(X + \sqrt{t}Z) - g(X)| \leq \epsilon/3$. Now

$$\begin{aligned}
|\mu_{X_n}(g) - \mu_X(g)| &= |\mathbb{E}g(X_n) - \mathbb{E}g(X)| \\
&\leq |\mathbb{E}g(X_n + \sqrt{t}Z) - \mathbb{E}g(X_n)| \\
&\quad + |\mathbb{E}g(X + \sqrt{t}Z) - \mathbb{E}g(X)| \\
&\quad + |\mathbb{E}g(X_n + \sqrt{t}Z) - \mathbb{E}g(X + \sqrt{t}Z)| \\
&\leq \frac{2\epsilon}{3} + |\mathbb{E}g(X_n + \sqrt{t}Z) - \mathbb{E}g(X + \sqrt{t}Z)|
\end{aligned}$$

So it suffices to show that $|\mathbb{E}g(X_n + \sqrt{t}Z) - \mathbb{E}g(X + \sqrt{t}Z)| < \epsilon/3$ for large enough n, or equivalently $\mathbb{E}g(X_n + \sqrt{t}Z) \to \mathbb{E}g(X + \sqrt{t}Z)$ as $n \to \infty$. Indeed, by (a) we have

$$\begin{aligned}
\mathbb{E}g(X_n + \sqrt{t}Z) &= \int_{\mathbb{R}^d} g(x)(\mu_{X_n} * g_t)(x) \, \mathrm{d}x \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} g(x) \left(\int_{\mathbb{R}^d} \phi_{X_n}(u) e^{-|u|^2 t/2} e^{-i\langle x, u\rangle} \, \mathrm{d}u \right) \mathrm{d}x \\
&\to \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} g(x) \phi_X(u) e^{-|u|^2 t/2} e^{-i\langle x, u\rangle} \, \mathrm{d}u \, \mathrm{d}x \\
&= \int_{\mathbb{R}^d} g(x)(\mu_X * g_t)(x) \, \mathrm{d}x = \mathbb{E}g(X + \sqrt{t}Z)
\end{aligned}$$

as $n \to \infty$ by Theorem 3.6. This completes the proof. □

6.2 More on Multivariate Gaussians

We have seen some properties of the Gaussian $N(0, tI_d)$ random variables. As you know, there is a much bigger class of multivariate Gaussians we like to consider on $\mathbb{R}^d$.
Recall that a one-dimensional Gaussian $N(\mu, \sigma^2)$ random variable (in $\mathbb{R}$) is one that has pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

Definition 6.2. A random variable X with range in $\mathbb{R}^d$ is Gaussian if for all $u \in \mathbb{R}^d$, $\langle u, X\rangle$ is a one-dimensional Gaussian random variable.

Theorem 6.2. *Let $X = (X_1, \ldots, X_n)$ be a Gaussian r.v. in $\mathbb{R}^n$, then:*
(a) If A is a $m \times n$ matrix and $b \in \mathbb{R}^m$, then $AX + b$ is Gaussian in $\mathbb{R}^m$.
(b) $X \in L^2$ and $\mu = \mathbb{E}X, V = \mathrm{var}(X) = (\mathrm{cov}(X_i, X_j))_{i,j}$ both exists.
(c) $\phi_X(u) = e^{i\langle u,\mu\rangle} - \langle u, Vu\rangle/2$ for all $u \in \mathbb{R}^n$.
(d) If V is invertible, then the pdf of X is given by

$$f_X(x) = (2\pi)^{-n/2} (\det V)^{-1/2} \exp\left(-\frac{\langle x-\mu, V^{-1}(x-\mu)\rangle}{2}\right)$$

(e) If $X = (X_{(1)}, X_{(2)})$ and $\mathrm{cov}(X_{(1)}, X_{(2)}) = 0$, then the multivariate normals $X_{(1)}, X_{(2)}$ are independent.

Proof. Let's not waste time on this. Exercise. □

6.3 Sums of Independent Random Variables

Let $X_1, \ldots, X_n$ be i.i.d. random variables with $\mathrm{var}(X_i) = 1, \mathbb{E}X_i = 0$, then for any $\epsilon > 0$ we have (by Markov's inequality)

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| > \epsilon\right) \le \frac{n^{-2}\,\mathrm{Var}(\sum_i X_i)}{\epsilon^2} = \frac{1}{n\epsilon^2} \to 0$$

as $n \to \infty$. This is known as the weak law of large numbers. By some casual manipulation one can then conclude

Theorem 6.3 (Weak Law of Large Numbers)**.** *For i.i.d. random variables $X_1, \ldots, X_n$ with finite mean and variance, we have $n^{-1}\sum_i X_i \to^{\mathbb{P}} \mathbb{E}X$ as $n \to \infty$.*

There are obvious directions to which we want to improve this theorem: Do we really need the finite variance assumption? Can we replace convergence in probability by convergence a.s.?

Theorem 6.4 (Strong (but not quite) Law of Large Numbers)**.** *Let $(X_n)_n$ be independent random variables with $\mathbb{E}X_n = \mu$ and $\mathbb{E}(X_n^4) \le M$ for all n. Let $S_n = \sum_{i=1}^n X_i$, then $S_n/n \to \mu$ a.s. as $n \to \infty$.*

We will prove an even stronger version of the strong law of large numbers (where we drop the finite 4^{th} moment assumption) after we've proved the ergodic theorem.

Proof. Assume WLOG that $\mu = 0$ (since shifting by a constant would not prevent the 4^{th} moment from being finite). Note that X_n, X_n^2, X_n^3 are all integrable as $L^4(\mathbb{P}) \subset L^p(\mathbb{P})$ for $1 \leq p \leq 4$. Also, by independence we have $\mathbb{E}(X_i X_j^3) = \mathbb{E}(X_i X_j X_k^2) = \mathbb{E}(X_i X_j X_k X_l) = 0$ for any distinct i, j, k, l. By Cauchy-Schwartz inequality (which follows from Hölder's inequality), we have (for $i \neq j$) $\mathbb{E}(X_i^2 X_j^2) = \mathbb{E}(X_i^2)\mathbb{E}(X_j^2) \leq \sqrt{\mathbb{E}(X_i^4)}\sqrt{\mathbb{E}(X_j^4)} \leq M$. Combining these results gives

$$\mathbb{E}(S_n^4) = \mathbb{E}\left(\sum_{i\leq n} X_i^4\right) + 6\mathbb{E}\left(\sum_{1\leq i<j\leq n} X_i^2 X_j^2\right) \leq nM + 3n(n-1)M \leq 3n^2 M$$

So

$$\mathbb{E}\left(\sum_n \left(\frac{S_n}{n}\right)^4\right) \leq 3M \sum_n \frac{1}{n^2} < \infty$$

In particular, $\sum_n (S_n/n)^4$ is a.s. finite, so we must have $S_n/n \to 0$ a.s.. □

It is natural to follow up with an enquiry on the rate of such convergence. This is answered by the central limit theorem.

Theorem 6.5 (Central Limit Theorem)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables on $\mathbb{R}$ such that $\mathbb{E}X_i = 0$ and $\operatorname{var}(X_i) = 1$, then for $S_n = \sum_{i=1}^n X_i$ we have $n^{-1/2} S_n \to^{\mathrm{d}} Z$ where $Z \sim N(0,1)$.*

Equivalently (proved in example sheet) $n^{-1/2} S_n \to Z$ weakly in $\mathbb{R}$.

Proof. By Theorem 6.1, it suffices to show that $\phi_n \to \phi_Z$ pointwise as $n \to \infty$ where ϕ_n is the characteristic function of $n^{-1/2} S_n$.
Let $\phi(u) = \mathbb{E}e^{iuX_1}$, then $\phi(0) = 1$. Note that $\mathbb{E}|X_1| \leq \sqrt{\mathbb{E}|X_1|^2} = 1 < \infty$, so $\phi'(u) = i\mathbb{E}X_1 e^{iuX_1} \implies \phi'(0) = i\mathbb{E}X_1 = 0$ by Theorem 3.6. Likewise $\phi''(u) = i^2 \mathbb{E}X_1^2 e^{iuX_1} \implies \phi''(0) = -1$. By Taylor's theorem, we have $\phi(u) = 1 - u^2/2 + o(u^2)$ as $u \to 0$. So for fixed u,

$$\phi_n(u) = \mathbb{E}e^{iun^{-1/2}(X_1 + \cdots + X_n)} = \phi(n^{-1/2}u)^n = \left(1 - \frac{u^2}{2n} + o\left(\frac{u^2}{n}\right)\right)^n$$

as $n \to \infty$. Recall that $\log(1+z) = z + o(z)$ as $z \to \infty$ where log is the principal complex logarithm, so

$$\log \phi_n(u) = n \log\left(1 - \frac{u^2}{2n} + o\left(\frac{u^2}{n}\right)\right) = -\frac{u^2}{2} + o(1)$$

Therefore $\phi_n(u) \to e^{-u^2/2} = \mathbb{E}e^{iuZ} = \phi_Z(u)$ as $n \to \infty$. □

Remark. 1. One can prove an analogous d-dimensional version of this theorem with basically the same way, or by using the Cramér-Wold device: $X_n \to^{\mathrm{d}} X$ in $\mathbb{R}^d$ (with componentwise cdf's) if and only if $\langle u, X_n\rangle \to^{\mathrm{d}} \langle u, X\rangle$ in $\mathbb{R}$ for all $u \in \mathbb{R}^d$.
2. The theorem in particular implies that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X\right) \to^{\mathrm{d}} Z$$

which is nonzero. One might ask whether I can push this to the strength of convergence in probability or convergence a.s.. However, neither is in general true: One can prove that

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1 \text{ a.s.}, \liminf_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = -1 \text{ a.s.}$$

This known as the law of the iterated logarithm.

In practice, the following results are often quite useful in the application of the central limit theorem:

Proposition 6.6. *1. (Continuous mapping theorem) If $X_n \to X$ weakly in $\mathbb{R}^d$ with X_n taking values in $U \subset \mathbb{R}^d$ and $g : U \to \mathbb{R}$ is continuous, then $g(X_n) \to g(X)$ weakly.*
2. (Slutsky's lemma) If $X_n \to X$ weakly in $\mathbb{R}^d$ and $Y_n \to^{\mathbb{P}} c$ in $\mathbb{R}^k$ for some (a.s.) constant c, then $(X_n, Y_n) \to (X, c)$ weakly in $\mathbb{R}^d \times \mathbb{R}^k$. In particular (combined with continuous mapping theorem), if $k = 1$ then $X_n + Y_n \to X + c, X_n Y_n \to Xc$ weakly.

Proof. 1. We will prove the case where $U = \mathbb{R}^d$. The general case is analogous. For any bounded continuous $f : \mathbb{R}^d \to \mathbb{R}$, $f \circ g : \mathbb{R}^d \to \mathbb{R}$ is bounded and continuous, so by the weak convergence $X_n \to X$,

$$\int_{\mathbb{R}^d} f \,\mathrm{d}\mu_{g(X_n)} = \int_{\mathbb{R}^d} f \circ g \,\mathrm{d}\mu_{X_n} \to \int_{\mathbb{R}^d} f \circ g \,\mathrm{d}\mu_X = \int_{\mathbb{R}^d} f \,\mathrm{d}\mu_{g(X)}$$

2. First, we shall prove that $X_n \to X$ weakly and $|X_n - Y_n| \to^{\mathbb{P}} 0$, then $Y_n \to X$ weakly as $n \to \infty$. Indeed, for any bounded and Lipschitz f (say with Lipschitz constant $\|f\|_{\text{Lip}}$) we have $|\mathbb{E}f(Y_n) - \mathbb{E}f(X)| \le |\mathbb{E}f(X_n) - \mathbb{E}f(X)| + |\mathbb{E}f(Y_n) - \mathbb{E}f(X_n)|$. We certainly have $|\mathbb{E}f(X_n) - \mathbb{E}f(X)| \to 0$ as $n \to \infty$ since $X_n \to X$ weakly. As for the second term, we have for any $\epsilon > 0$ and $\delta > 0$,

$$\begin{aligned} |\mathbb{E}f(Y_n) - \mathbb{E}f(X_n)| &= |\mathbb{E}((f(Y_n) - f(X_n))(1_{|X_n - Y_n| \le \delta} + 1_{|X_n - Y_n| > \delta}))| \\ &\le \|f\|_{\text{Lip}}\delta + \|f\|_\infty \mathbb{P}(|X_n - Y_n| > \delta) < \epsilon \end{aligned}$$

For large enough n and small enough δ.
Slutsky's lemma follows directly from this, with the observation that $|(X_n, Y_n) - (X_n, c)| = |Y_n - c| \to^{\mathbb{P}} 0$ and that $(X_n, c) \to (X, c)$ weakly since for bounded continuous $f : \mathbb{R}^d \times \mathbb{R}^k$, the map $x \mapsto f(x, c)$ is also bounded and continuous (on $\mathbb{R}^d$), so $\mathbb{E}f(X_n, c) \to \mathbb{E}f(X, c)$ by the weak convergence $X_n \to X$. □

7 Ergodic Theory

Definition 7.1. A dynamical system is a duple (E, θ) consisting of a state space E and a transformation $\theta : E \to E$.

We write θ^n to denote the n-fold composition of θ with itself. In ergodic theory, one studies the long term "statistical" behaviour of the orbits, i.e. subsets of E that has the form $\{\theta^n(x) : x \in E\}$. Usually, we consider the case where E has the structure of a measure space, say with measure μ.

One can ask many questions about a dynamical system. For example, one can ask whether the "averages" in it behave nicely, i.e. whether

$$\frac{1}{N}\sum_{k=1}^{N} 1_A \circ \theta^k(x) = \frac{1}{N}\sum_{k=1}^{N} 1_{\theta^k(x)\in A}$$

stablises a.e. as $N \to \infty$ given that θ is somehow "invariant" under μ.

Example 7.1 (Boltzmann's Ergodic Hypothesis)**.** Suppose one has a closed system E with the motion of a particle in it move according to the transformation θ. Boltzmann asked whether, fixing a region $A \subset E$, one can draw certain conclusions about the average frequency a particle visits A. Naturally, this should be proportional to the "volume" of A, which is what we study in measure theory!

7.1 Ergodicity

Definition 7.2. Let $(E, \mathcal{E}, \mu)$ be a measure space. A measurable map $\theta : E \to E$ is called measure-preserving if $\mu(\theta^{-1}(A)) = \mu(A)$ for all $A \in \mathcal{E}$.

In this case, if $f : E \to \mathbb{R}$ is measurable, then necessarily

$$\int_E f \circ \theta \, \mathrm{d}\mu = \int_E f \, \mathrm{d}\mu$$

Definition 7.3. A measurable map $f : E \to \mathbb{R}$ is called invariant (under θ) if $f = f \circ \theta$. $A \in \mathcal{E}$ is called invariant if $\theta^{-1}(A) = A$.

One will prove in example sheet that the collection $\mathcal{E}_\theta$ of invariant subsets $A \in \mathcal{E}$ is a σ-algebra on E, and $f : E \to \mathbb{R}$ is invariant if and only if f is $\mathcal{E}_\theta$-measurable.

Definition 7.4. We say θ is (μ-)ergodic if any $A \in \mathcal{E}_\theta$ has either $\mu(A) = 0$ or $\mu(E - A) = 0$.

It is a fact you'll prove on the example sheet that if θ is ergodic and f is invariant for θ, then f is constant a.e..

Example 7.2. 1. On the n-torus $E = (0,1]^n, \mathcal{E} = \mathcal{B}((0,1])^{\otimes n}, \mu = \mu|_{(0,1]}^{\otimes n}$, the translation map

$$\theta_a(x_1, \ldots, x_n) = (x_1 + a_1, \ldots, x_n + a_n) \pmod 1$$

is measure preserving by translation-invariance of μ. For $n = 1$, θ_a is ergodic if and only if a is irrational (see example sheet).
2. On $E = (0,1], \mathcal{E} = \mathcal{B}((0,1]), \mu = \mu|_{(0,1]}$, the map $x \mapsto 2x \pmod 1$ is ergodic for μ (example sheet).

A third example, which will be important to what we'll do in a moment, is the following: Recall from earlier that on the infinite product space $E = \mathbb{R}^{\mathbb{N}}$ we can consider the σ-algebra $\mathcal{E}$ generated by the π-system of "cylinder sets" $\mathscr{C} = \{\prod_{n\in\mathbb{N}} A_n : A_n \in \mathcal{B}, \exists N \in \mathbb{N}, \forall n > N, A_n = \mathbb{R}\}$ (or, equivalently, $\mathcal{E} = \sigma(\mathscr{C}) = \sigma(X_n : n \in \mathbb{N})$ where $X_n(x) = x_n$). Let m be any probability

measure on $\mathbb{R}$. We have shown earlier that there exists an infinite sequence $(Y_n)_n$ of i.i.d. real random variables with distribution m defined on $(\Omega, \mathcal{F}, \mathbb{P}) = ((0,1), \mathcal{B}, \mu|_{(0,1)})$. This sequence can then be considered as a random variable with range in $(E, \mathcal{E}) = (\mathbb{R}^{\mathbb{N}}, \sigma(\mathscr{C}))$ by $Y : \Omega \to E, Y(\omega) = (Y_n(\omega))_n$ which is $(\mathcal{F}, \mathcal{E})$-measurable since any cylinder sets have measurable preimage. The image measure $\mu(A) = \mu_m(A) = \mathbb{P}(Y^{-1}(A))$ on $(E, \mathcal{E})$ then satisfies

$$\forall A = \prod_{n\in\mathbb{N}} A_n \in \mathscr{C}, \mu\left(\prod_{n\in\mathbb{N}} A_n\right) = \prod_{n\in\mathbb{N}} m(A_n)$$

We call $(E, \mathcal{E}, \mu)$ the "canonical model" for this infinite sequence of i.i.d. random variables.

Theorem 7.1. *On $(E, \mathcal{E}, \mu)$, the Bernoulli shift map $\theta(x_1, x_2, \ldots) = (x_2, x_3, \ldots)$ is measure-preserving and ergodic.*

Proof. For $A \in \mathscr{C}$ of the form $A = \prod_{i\le N}(-\infty, x_i] \times \mathbb{R} \times \mathbb{R} \times \cdots$, we can check

$$\begin{aligned}\mu(A) &= \mathbb{P}(Y_1 \le x_1, \ldots, Y_N \le x_N) = m((-\infty, x_1]) \times \cdots \times m((-\infty, x_N]) \\ &= \mathbb{P}(Y_2 \le x_1, \ldots, Y_{N+1} \le x_N) = \mu(\theta^{-1}(A))\end{aligned}$$

So θ is measure-preserving (by extending this equality via Lemma 1.2). To prove ergodicity, let $X_m : x \mapsto x_m$ be the projection map on E as usual. Recall that the tail σ-algebra for $(X_m)_m$ is defined as $\mathcal{T} = \bigcap_n \mathcal{T}_n, \mathcal{T}_n = \sigma(X_m : m \ge n+1)$. For $A \in \mathscr{C}$, we have

$$\theta^{-n}(A) = \{x : \forall k \ge 1, X_{n+k}(x) \in A_k\} = \bigcap_{k=1}^{\infty} X_{n+k}^{-1}(A_k) \in \mathcal{T}_n$$

As $\mathscr{C}$ generates $\mathcal{E}$, θ^n has to be $(\mathcal{T}_n, \mathcal{E})$-measurable, which in turn means that $\theta^{-n}(A) \in \mathcal{T}_n$ for all $A \in \mathcal{E}$.
Suppose A is invariant, then $A = \theta^{-n}(A) \in \mathcal{T}_n$ for all n, so $A \in \mathcal{T}$. By Theorem 2.6, $\mu(A)$ is either 0 or 1 for any invariant set $A \in \mathcal{E}_\theta$, hence θ is ergodic. □

7.2 Ergodic Theorems

Let $(E, \mathcal{E}, \mu)$ be a σ-finite measure space with $\theta : E \to E$ a measure-preserving transformation. For a measurable $f : E \to \mathbb{R}$, we set $S_0 = 0$ and

$$S_n = S_n(f) = \sum_{k=0}^{n-1} f \circ \theta^k$$

Theorem 7.2 (Birkhoff's Ergodic Theorem)**.** *Let $f \in L^1(\mu)$, then there is some θ-invariant $\bar{f} : (E, \mathcal{E}) \to \mathbb{R}$ such that $\mu(|\bar{f}|) \le \mu(|f|)$ and $S_n(f)/n \to \bar{f}$ a.e. as $n \to \infty$.*

Lemma 7.3 (Maximal Ergodic Lemma)**.** *For $f \in L^1(E, \mathcal{E}, \mu)$, we set $S^* = S^*(f) = \sup_{n\ge 0} S_n(f)$, then*

$$\int_{\{S^*>0\}} f \, \mathrm{d}\mu \ge 0$$

Proof. Define $S_n^* = \max_{0\le m\le n} S_m$, then $S_m \le S_n^*$ and thus $S_{m+1} = S_m\circ\theta + f \le S_n^*\circ\theta + f$ for all $0 \le m \le n$. Let $A_n = \{S_n^* > 0\} \uparrow \{S^* > 0\}$. On A_n, $S_n^* = \max_{1\le m\le n} S_m$ (note that we have removed S_0 from the set on which we are taking the maximum), so $S_n^* = \max_{1\le m\le n} S_m \le \max_{0\le m\le n} S_{m+1} \le S_n^*\circ\theta + f$ on A. Integrating this inequality with respect to $\mathrm{d}\mu$ gives (noting that $S_n^* = 0$ in A_n^c)

$$\begin{aligned}\int_E S_n^*\,\mathrm{d}\mu = \int_{A_n} S_n^*\,\mathrm{d}\mu &\le \int_{A_n} S_n^*\circ\theta\,\mathrm{d}\mu + \int_{A_n} f\,\mathrm{d}\mu \\ &\le \int_E S_n^*\circ\theta\,\mathrm{d}\mu + \int_{A_n} f\,\mathrm{d}\mu = \int_E S_n^*\,\mathrm{d}\mu + \int_{A_n} f\,\mathrm{d}\mu\end{aligned}$$

as θ is measure-preserving. By dominated convergence theorem (using the fact that $f \in L^1$), we have

$$0 \le \int_{A_n} f\,\mathrm{d}\mu = \int_E 1_{A_n} f\,\mathrm{d}\mu \to \int_E 1_{S^*>0} f\,\mathrm{d}\mu = \int_{\{S^*>0\}} f\,\mathrm{d}\mu$$

which is what we wanted. □

Proof of Theorem 7.2. Observe that

$$\limsup_{n\to\infty} \frac{S_n(f)}{n} = \limsup_{n\to\infty} \frac{S_n(f)\circ\theta}{n}, \liminf_{n\to\infty} \frac{S_n(f)}{n} = \liminf_{n\to\infty} \frac{S_n(f)\circ\theta}{n}$$

are invariant functions. Consequently,

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

are measurable and invariant since the preimages of measurable sets under invariant functions are measurable and invariant.
Suppose $D_{a,b} \ne \varnothing$ and assume WLOG that $b > 0$. Take $B \in \mathcal{E}, B \subset D$ such that $\mu(B) < \infty$ and set $g = f - b1_B \in L^1(\mu)$. We have $S_n(g) = S_n(f) - bS_n(1_B) \ge S_n(t) - bn > 0$ on D for some n, so $\{S^*(g) > 0\} \subset D$. Note that $\mu|_D$ is still θ-invariant as $\mu|_D(A) = \mu(A\cap D) = \mu(\theta^{-1}(A\cap D)) = \mu(\theta^{-1}(A)\cap\theta^{-1}(D)) = \mu(\theta^{-1}(A)\cap D) = \mu|_D(\theta^{-1}(A))$. So we can safely apply Lemma 7.3 to D (and g), which gives (since we already know that $\{S^*(g) > 0\} \subset D$)

$$0 \le \int_D g\,\mathrm{d}\mu = \int_D f\,\mathrm{d}\mu - b\mu(B) \implies b\mu(B) \le \int_D f\,\mathrm{d}\mu$$

Repeating this argument with $-f, -a$ gives $(-a)\mu(B) \le \mu|_D(-f)$. Combining them and we obtain

$$b\mu(B) \le \int_D f\,\mathrm{d}\mu \le a\mu(B)$$

which means that $\mu(B) = 0$, i.e. any finite measure subset of D must be null. But E is σ-finite, so necessarily $\mu(D) = 0$.
Now

$$\Delta = \left\{\liminf_{n\to\infty} \frac{S_n}{n} < \limsup_{n\to\infty} \frac{S_n}{n}\right\} = \bigcup_{a,b\in\mathbb{Q}, a<b} D_{a,b}$$

is null. We define

$$\bar{f} = \begin{cases} \liminf_n S_n/n = \limsup_n S_n/n & \text{on } \Delta^c \\ 0 & \text{on } \Delta \end{cases}$$

Then $S_n/n \to \bar{f}$ a.e. as $n \to \infty$.
To get the bound on $\mu(|\bar{f}|)$, note that $\mu(|f \circ \theta^n|) = \mu(|f|)$, so $\mu(|S_n|) \leq n\mu(|f|) \leq n\mu(|f|)$ and therefore

$$\mu(|\bar{f}|) = \mu\left(\liminf_{n\to\infty} \frac{S_n}{n}\right) \leq \liminf_{n\to\infty} \mu\left(\left|\frac{S_n}{n}\right|\right) \leq \mu(|f|)$$

as desired. □

Theorem 7.4 (von Neumann's L^p Ergodic Theorem)**.** *Assume $\mu(E) < \infty$ and $1 \leq p < \infty$. Then for all $f \in L^p(\mu)$, there is some invariant $\bar{f} \in L^p(\mu)$ such that $S_n(f)/n \to \bar{f}$ as $n \to \infty$ in $L^p(\mu)$.*

Proof. Since θ is measure-preserving,

$$\|f \circ \theta^i\|_p^p = \int_E |f|^p \circ \theta^i \, \mathrm{d}\mu = \int_E |f|^p \, \mathrm{d}\mu = \|f\|_p^p$$

By Minkowski's inequality, we have $\|S_n(f)/n\|_p \leq \|f\|_p$ for any $f \in L^p(\mu)$. For any $\epsilon > 0$, choose K large enough such that $f_K = (-K) \vee (f \wedge K)$ has $\|f - f_K\|_p < \epsilon/3$. This is possible due to Theorem 3.6 and that $|f(x)|^p 1_{|f|>K} \downarrow 0$ a.e. as $K \to \infty$. Now f_K is bounded and hence in $L^1(\mu)$ as $\mu(E) < \infty$. By Theorem 7.2, we know that $S_n(f_K)/n \to \bar{f}_K$ a.e. for some invariant $\bar{f}_K$. Since $|S_n(f_K)/n| \leq K$ for all n, we know that $S_n(f_K)/n \to \bar{f}_K$ in L^p by Theorem 4.9 (and some manipulation). Choose $n \geq N$ large enough such that $\|\bar{f}_K - S_n(f_K)/n\| < \epsilon/3$. Finally, let $\bar{f}$ be as in Theorem 7.2 (noting that $L^p(\mu) \subset L^1(\mu)$ as μ is finite), then

$$\begin{aligned} \|\bar{f} - \bar{f}_K\|_p^p &= \int_E \lim_{n\to\infty} \left|\frac{S_n(f)}{n} - \frac{S_n(f_K)}{n}\right|^p \mathrm{d}\mu \\ &\leq \liminf_{n\to\infty} \int_E \left|\frac{S_n(f - f_K)}{n}\right|^p \mathrm{d}\mu \leq \|f - f_K\|_p^p < \left(\frac{\epsilon}{3}\right)^p \end{aligned}$$

Collecting them all and using Minkowski's inequality, we conclude that

$$\left\|\frac{S_n(f)}{n} - \bar{f}\right\|_p \leq \left\|\frac{S_n(f - f_K)}{n}\right\|_p + \left\|\frac{S_n(f_K)}{n} - \bar{f}_K\right\|_p + \left\|\bar{f}_K - \bar{f}\right\|_p < \epsilon$$

for large n. □

To deduce the unconditional version of the law of large number, we apply these ergodic theorems to the Bernoulli shift map $\theta(x_1, x_2, \ldots) = (x_2, x_3, \ldots)$ on the canonical space $(E, \mathcal{E}, \mu) = (\mathbb{R}^{\mathbb{N}}, \sigma(\mathscr{C}), \mu_m)$.

Theorem 7.5 (Strong(est!) Law of Large Numbers)**.** *Let m be a probability measure on $\mathbb{R}$ such that*

$$\int_{\mathbb{R}} |y| \, \mathrm{d}m(y) < \infty$$

Then

$$\mu\left(x \in \mathbb{R}^{\mathbb{N}} : \frac{x_1 + \cdots + x_n}{n} \to \nu = \int_{\mathbb{R}} y \, \mathrm{d}m(y) \text{ as } n \to \infty\right) = 1$$

Proof. Applying the ergodic theorems to $f(x) = x_1$ gives

$$\frac{S_n(f)}{n} = \frac{x_1 + \cdots + x_n}{n} \to \bar{f}$$

both a.e. and in $L^1(\mu)$ for some $\bar{f}$ invariant hence constant a.e. since θ is ergodic. The constant equals to $\mu(\bar{f}) = \lim_n \mu(S_n(f)/n) = \nu$ which completes the proof. □

Theorem 7.6 (Kolmogorov-Khinchine)**.** *Let $(X_n)_n$ be a sequence of i.i.d. r.v.'s defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}|X_1| < \infty, \mathbb{E}X_1 = \nu$. If $S_n = \sum_{i=1}^n X_i$ then $S_n/n \to \nu$ as $n \to \infty$.*

Proof. Just replicate our construction of $(\mathbb{R}^{\mathbb{N}}, \sigma(\mathscr{C}), \mu_m)$ and cast the preceding theorem. □

These are indeed the best we can do with the law of large numbers: $\mathbb{E}|X_1| < \infty$ is necessary to the desired limit result.

Proposition 7.7. *Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}|X_1| = \infty$, then S_n/n does not converge a.s. to any finite limit.*

Proof. Suppose to the contrary that S_n/n converges a.s. to some limit, then

$$\frac{S_{n-1}}{n} = \frac{n-1}{n}\frac{S_{n-1}}{n-1}$$

converges to the same limit, hence

$$\frac{X_n}{n} = \frac{S_n - S_{n-1}}{n} \to 0$$

a.s. as $n \to \infty$. But for any non-negative r.v. Y we have $\mathbb{E}Y \leq \sum_n \mathbb{P}(Y > n)$ (see example sheet), so for $Y = |X_1|$ we have

$$\infty = \mathbb{E}|X_1| \leq \sum_n \mathbb{P}(|X_1| > n) = \sum_n \mathbb{P}(|X_n| > n)$$

by Lemma 1.8 we have $\mathbb{P}(|X_n|/n > 1 \text{ i.o.}) = 1$, contradiction. □