PROBABILITY AND MEASURE

J. R. NORRIS

Contents

1. Measures 3
2. Measurable functions and random variables 11
3. Integration 18
4. Norms and inequalities 28
5. Completeness of L^p and orthogonal projection 31
6. Convergence in $L^1(\mathbb{P})$ 34
7. Fourier transforms 37
8. Gaussian random variables 43
9. Ergodic theory 44
10. Sums of independent random variables 48

These notes are intended for use by students of the Mathematical Tripos at the University of Cambridge. Please send corrections to `j.r.norris@statslab.cam.ac.uk`.

Schedule

Measure spaces, σ-algebras, π-systems and uniqueness of extension, statement *and proof* of Carathéodory's extension theorem. Construction of Lebesgue measure on $\mathbb{R}$, Borel σ-algebra of $\mathbb{R}$, existence of a non-measurable subset of $\mathbb{R}$. Lebesgue–Stieltjes measures and probability distribution functions. Independence of events, independence of σ-algebras. Borel–Cantelli lemmas. Kolmogorov's zero–one law.

Measurable functions, random variables, independence of random variables. Construction of the integral, expectation. Convergence in measure and convergence almost everywhere. Fatou's lemma, monotone and dominated convergence, uniform integrability, differentiation under the integral sign. Discussion of product measure and statement of Fubini's theorem.

Chebyshev's inequality, tail estimates. Jensen's inequality. Completeness of L^p for $1 \leq p \leq \infty$. Hölder's and Minkowski's inequalities, uniform integrability.

L^2 as a Hilbert space. Orthogonal projection, relation with elementary conditional probability. Variance and covariance. Gaussian random variables, the multivariate normal distribution.

The strong law of large numbers, proof for independent random variables with bounded fourth moments. Measure preserving transformations, Bernoulli shifts. Statements *and proofs* of the maximal ergodic theorem and Birkhoff's almost everywhere ergodic theorem, proof of the strong law.

The Fourier transform of a finite measure, characteristic functions, uniqueness and inversion. Weak convergence, statement of Lévy's continuity theorem for characteristic functions. The central limit theorem.

Appropriate books

P. Billingsley *Probability and Measure.* Wiley 2012 (£90.00 hardback).

R.M. Dudley *Real Analysis and Probability.* Cambridge University Press 2002 (£40.00 paperback).

R.T. Durrett *Probability: Theory and Examples.* (£45.00 hardback).

D. Williams *Probability with Martingales.* Cambridge University Press 1991 (£31.00 paperback).

1. Measures

1.1. **Definitions.** Let E be a set. A *σ-algebra* $\mathcal{E}$ on E is a set of subsets of E, containing the empty set $\emptyset$ and such that, for all $A \in \mathcal{E}$ and all sequences $(A_n : n \in \mathbb{N})$ in $\mathcal{E}$,

$$A^c \in \mathcal{E}, \quad \bigcup_n A_n \in \mathcal{E}.$$

The pair $(E, \mathcal{E})$ is called a *measurable space.* Given $(E, \mathcal{E})$, each $A \in \mathcal{E}$ is called a *measurable set.*

A *measure* μ on $(E, \mathcal{E})$ is a function $\mu : \mathcal{E} \to [0, \infty]$, with $\mu(\emptyset) = 0$, such that, for any sequence $(A_n : n \in \mathbb{N})$ of disjoint elements of $\mathcal{E}$,

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

This property is called *countable additivity.* The triple $(E, \mathcal{E}, \mu)$ is called a *measure space.*

1.2. **Discrete measure theory.** Let E be a countable set and let $\mathcal{E}$ be the set of all subsets of E. A *mass function* is any function $m : E \to [0, \infty]$. If μ is a measure on $(E, \mathcal{E})$, then, by countable additivity,

$$\mu(A) = \sum_{x \in A} \mu(\{x\}), \quad A \subseteq E.$$

So there is a one-to-one correspondence between measures and mass functions, given by

$$m(x) = \mu(\{x\}), \quad \mu(A) = \sum_{x \in A} m(x).$$

This sort of measure space provides a toy version of the general theory, where each of the results we prove for general measure spaces reduces to some straightforward fact about the convergence of series. This is all one needs to do elementary discrete probability and discrete-time Markov chains, so these topics are usually introduced without discussing measure theory.

Discrete measure theory is essentially the only context where one can define a measure explicitly, because, in general, σ-algebras are not amenable to an explicit presentation which would allow us to make such a definition. Instead one specifies the values to be taken on some smaller set of subsets, which generates the σ-algebra. This gives rise to two problems: first to know that there is a measure extending the given set function, second to know that there is not more than one. The first problem, which is one of construction, is often dealt with by Carathéodory's extension theorem. The second problem, that of uniqueness, is often dealt with by Dynkin's π-system lemma.

1.3. **Generated σ-algebras.** Let $\mathcal{A}$ be a set of subsets of E. Define

$$\sigma(\mathcal{A}) = \{A \subseteq E : A \in \mathcal{E} \quad \text{for all } \sigma\text{-algebras } \mathcal{E} \text{ containing } \mathcal{A}\}.$$

Then $\sigma(\mathcal{A})$ is a σ-algebra, which is called the *σ-algebra generated by $\mathcal{A}$* . It is the smallest σ-algebra containing $\mathcal{A}$.

1.4. **π-systems and d-systems.** Let $\mathcal{A}$ be a set of subsets of E. Say that $\mathcal{A}$ is a *π-system* if $\emptyset \in \mathcal{A}$ and, for all $A, B \in \mathcal{A}$,

$$A \cap B \in \mathcal{A}.$$

Say that $\mathcal{A}$ is a *d-system* if $E \in \mathcal{A}$ and, for all $A, B \in \mathcal{A}$ with $A \subseteq B$ and all increasing sequences $(A_n : n \in \mathbb{N})$ in $\mathcal{A}$,

$$B \setminus A \in \mathcal{A}, \quad \bigcup_n A_n \in \mathcal{A}.$$

Note that, if $\mathcal{A}$ is both a π-system and a d-system, then $\mathcal{A}$ is a σ-algebra.

Lemma 1.4.1 (Dynkin's π-system lemma)**.** *Let $\mathcal{A}$ be a π-system. Then any d-system containing $\mathcal{A}$ contains also the σ-algebra generated by $\mathcal{A}$.*

Proof. Denote by $\mathcal{D}$ the intersection of all d-systems containing $\mathcal{A}$. Then $\mathcal{D}$ is itself a d-system. We shall show that $\mathcal{D}$ is also a π-system and hence a σ-algebra, thus proving the lemma. Consider

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

Then $\mathcal{A} \subseteq \mathcal{D}'$ because $\mathcal{A}$ is a π-system. Let us check that $\mathcal{D}'$ is a d-system: clearly $E \in \mathcal{D}'$; next, suppose $B_1, B_2 \in \mathcal{D}'$ with $B_1 \subseteq B_2$, then for $A \in \mathcal{A}$ we have

$$(B_2 \setminus B_1) \cap A = (B_2 \cap A) \setminus (B_1 \cap A) \in \mathcal{D}$$

because $\mathcal{D}$ is a d-system, so $B_2 \setminus B_1 \in \mathcal{D}'$; finally, if $B_n \in \mathcal{D}', n \in \mathbb{N}$, and $B_n \uparrow B$, then for $A \in \mathcal{A}$ we have

$$B_n \cap A \uparrow B \cap A$$

so $B \cap A \in \mathcal{D}$ and $B \in \mathcal{D}'$. Hence $\mathcal{D} = \mathcal{D}'$.

Now consider

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

Then $\mathcal{A} \subseteq \mathcal{D}''$ because $\mathcal{D} = \mathcal{D}'$. We can check that $\mathcal{D}''$ is a d-system, just as we did for $\mathcal{D}'$. Hence $\mathcal{D}'' = \mathcal{D}$ which shows that $\mathcal{D}$ is a π-system as promised. □

1.5. **Set functions and properties.** Let $\mathcal{A}$ be any set of subsets of E containing the empty set $\emptyset$. A *set function* is a function $\mu : \mathcal{A} \to [0, \infty]$ with $\mu(\emptyset) = 0$. Let μ be a set function. Say that μ is *increasing* if, for all $A, B \in \mathcal{A}$ with $A \subseteq B$,

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

Say that μ is *additive* if, for all disjoint sets $A, B \in \mathcal{A}$ with $A \cup B \in \mathcal{A}$,

$$\mu(A \cup B) = \mu(A) + \mu(B).$$

Say that μ is *countably additive* if, for all sequences of disjoint sets $(A_n : n \in \mathbb{N})$ in $\mathcal{A}$ with $\bigcup_n A_n \in \mathcal{A}$,

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

Say that μ is *countably subadditive* if, for all sequences $(A_n : n \in \mathbb{N})$ in $\mathcal{A}$ with $\bigcup_n A_n \in \mathcal{A}$,

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

1.6. **Construction of measures.** Let $\mathcal{A}$ be a set of subsets of E. Say that $\mathcal{A}$ is a *ring* on E if $\emptyset \in \mathcal{A}$ and, for all $A, B \in \mathcal{A}$,

$$B \setminus A \in \mathcal{A}, \quad A \cup B \in \mathcal{A}.$$

Say that $\mathcal{A}$ is an *algebra* on E if $\emptyset \in \mathcal{A}$ and, for all $A, B \in \mathcal{A}$,

$$A^c \in \mathcal{A}, \quad A \cup B \in \mathcal{A}.$$

Theorem 1.6.1 (Carathéodory's extension theorem)**.** *Let $\mathcal{A}$ be a ring of subsets of E and let $\mu : \mathcal{A} \to [0, \infty]$ be a countably additive set function. Then μ extends to a measure on the σ-algebra generated by $\mathcal{A}$.*

Proof. For any $B \subseteq E$, define the *outer measure*

$$\mu^*(B) = \inf \sum_n \mu(A_n)$$

where the infimum is taken over all sequences $(A_n : n \in \mathbb{N})$ in $\mathcal{A}$ such that $B \subseteq \bigcup_n A_n$ and is taken to be ∞ if there is no such sequence. Note that μ^* is increasing and $\mu^*(\emptyset) = 0$. Let us say that $A \subseteq E$ is *μ^*-measurable* if, for all $B \subseteq E$,

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

Write $\mathcal{M}$ for the set of all μ^*-measurable sets. *We shall show that $\mathcal{M}$ is a σ-algebra containing $\mathcal{A}$ and that μ^* restricts to a measure on $\mathcal{M}$, extending μ.* This will prove the theorem.

Step I. We show that μ^* is countably subadditive. Suppose that $B \subseteq \bigcup_n B_n$. We have to show that

$$\mu^*(B) \leq \sum_n \mu^*(B_n).$$

It will suffice to consider the case where $\mu^*(B_n) < \infty$ for all n. Then, given $\varepsilon > 0$, there exist sequences $(A_{nm} : m \in \mathbb{N})$ in $\mathcal{A}$, with

$$B_n \subseteq \bigcup_m A_{nm}, \quad \mu^*(B_n) + \varepsilon/2^n \geq \sum_m \mu(A_{nm}).$$

Now

$$B \subseteq \bigcup_n \bigcup_m A_{nm}$$

so

$$\mu^*(B) \leq \sum_n \sum_m \mu(A_{nm}) \leq \sum_n \mu^*(B_n) + \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, we are done.

Step II. We show that μ^* extends μ. Since $\mathcal{A}$ is a ring and μ is countably additive, μ is countably subadditive and increasing. Hence, for $A \in \mathcal{A}$ and any sequence $(A_n : n \in \mathbb{N})$ in $\mathcal{A}$ with $A \subseteq \bigcup_n A_n$,

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

On taking the infimum over all such sequences, we see that $\mu(A) \leq \mu^*(A)$. On the other hand, it is obvious that $\mu^*(A) \leq \mu(A)$ for $A \in \mathcal{A}$.

Step III. We show that $\mathcal{M}$ contains $\mathcal{A}$. Let $A \in \mathcal{A}$ and $B \subseteq E$. We have to show that

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

By subadditivity of μ^*, it is enough to show that

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

If $\mu^*(B) = \infty$, this is clearly true, so let us assume that $\mu^*(B) < \infty$. Then, given $\varepsilon > 0$, we can find a sequence $(A_n : n \in \mathbb{N})$ in $\mathcal{A}$ such that

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

Then

$$B \cap A \subseteq \bigcup_n (A_n \cap A), \quad B \cap A^c \subseteq \bigcup_n (A_n \cap A^c)$$

so

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

Since $\varepsilon > 0$ was arbitrary, we are done.

Step IV. We show that $\mathcal{M}$ is an algebra. Clearly $E \in \mathcal{M}$ and $A^c \in \mathcal{M}$ whenever $A \in \mathcal{M}$. Suppose that $A_1, A_2 \in \mathcal{M}$ and $B \subseteq E$. Then

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

Hence $A_1 \cap A_2 \in \mathcal{M}$.

Step V. We show that $\mathcal{M}$ is a σ-algebra and that μ^* restricts to a measure on $\mathcal{M}$. We already know that $\mathcal{M}$ is an algebra, so it suffices to show that, for any sequence of disjoint sets $(A_n : n \in \mathbb{N})$ in $\mathcal{M}$, for $A = \bigcup_n A_n$ we have

$$A \in \mathcal{M}, \quad \mu^*(A) = \sum_n \mu^*(A_n).$$

So, take any $B \subseteq E$, then

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

Note that $\mu^*(B \cap A_1^c \cap \cdots \cap A_n^c) \geq \mu^*(B \cap A^c)$ for all n. Hence, on letting $n \to \infty$ and using countable subadditivity, we get

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

The reverse inequality holds by subadditivity, so we have equality. Hence $A \in \mathcal{M}$ and, setting $B = A$, we get

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

□

1.7. **Uniqueness of measures.**

Theorem 1.7.1 (Uniqueness of extension). *Let μ_1, μ_2 be measures on $(E, \mathcal{E})$ with $\mu_1(E) = \mu_2(E) < \infty$. Suppose that $\mu_1 = \mu_2$ on $\mathcal{A}$, for some π-system $\mathcal{A}$ generating $\mathcal{E}$. Then $\mu_1 = \mu_2$ on $\mathcal{E}$.*

Proof. Consider $\mathcal{D} = \{A \in \mathcal{E} : \mu_1(A) = \mu_2(A)\}$. By hypothesis, $E \in \mathcal{D}$; for $A, B \in \mathcal{E}$ with $A \subseteq B$, we have

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

so, if $A, B \in \mathcal{D}$, then also $B \setminus A \in \mathcal{D}$; if $A_n \in \mathcal{D}, n \in \mathbb{N}$, with $A_n \uparrow A$, then

$$\mu_1(A) = \lim_n \mu_1(A_n) = \lim_n \mu_2(A_n) = \mu_2(A)$$

so $A \in \mathcal{D}$. Thus $\mathcal{D}$ is a d-system containing the π-system $\mathcal{A}$, so $\mathcal{D} = \mathcal{E}$ by Dynkin's lemma. □

1.8. **Borel sets and measures.** Let E be a Hausdorff topological space. The σ-algebra generated by the set of open sets is E is called the *Borel* σ-algebra of E and is denoted $\mathcal{B}(E)$. The Borel σ-algebra of $\mathbb{R}$ is denoted simply by $\mathcal{B}$. A measure μ on $(E, \mathcal{B}(E))$ is called a *Borel* measure on E. If moreover $\mu(K) < \infty$ for all compact sets K, then μ is called a *Radon* measure on E.

1.9. **Probability measures, finite and σ-finite measures.** If $\mu(E) = 1$ then μ is a *probability measure* and $(E, \mathcal{E}, \mu)$ is a *probability space*. The notation $(\Omega, \mathcal{F}, \mathbb{P})$ is often used to denote a probability space. If $\mu(E) < \infty$, then μ is a *finite* measure. If there exists a sequence of sets $(E_n : n \in \mathbb{N})$ in $\mathcal{E}$ with $\mu(E_n) < \infty$ for all n and $\bigcup_n E_n = E$, then μ is a *σ-finite* measure.

1.10. **Lebesgue measure.**

Theorem 1.10.1. *There exists a unique Borel measure μ on $\mathbb{R}$ such that, for all $a, b \in \mathbb{R}$ with $a < b$,*

$$\mu((a, b]) = b - a.$$

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

Proof. (*Existence.*) Consider the ring $\mathcal{A}$ of finite unions of disjoint intervals of the form

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

We note that $\mathcal{A}$ generates $\mathcal{B}$. Define for such $A \in \mathcal{A}$

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

Note that the presentation of A is not unique, as $(a, b] \cup (b, c] = (a, c]$ whenever $a < b < c$. Nevertheless, it is easy to check that μ is well-defined and additive. We aim to show that μ is countably additive on $\mathcal{A}$, from which the existence of a Borel measure extending μ follows by Carathéodory's extension theorem.

By additivity, it suffices to show that, if $A \in \mathcal{A}$ and if $(A_n : n \in \mathbb{N})$ is an increasing sequence in $\mathcal{A}$ with $A_n \uparrow A$, then $\mu(A_n) \to \mu(A)$. Set $B_n = A \setminus A_n$ then $B_n \in \mathcal{A}$ and $B_n \downarrow \emptyset$. By additivity again, it suffices to show that $\mu(B_n) \to 0$. Suppose, in fact, that for some $\varepsilon > 0$, we have $\mu(B_n) \geq 2\varepsilon$ for all n. For each n we can find $C_n \in \mathcal{A}$ with $\bar{C}_n \subseteq B_n$ and $\mu(B_n \setminus C_n) \leq \varepsilon 2^{-n}$. Then

$$\mu(B_n \setminus (C_1 \cap \cdots \cap C_n)) \leq \mu((B_1 \setminus C_1) \cup \cdots \cup (B_n \setminus C_n)) \leq \sum_{n \in \mathbb{N}} \varepsilon 2^{-n} = \varepsilon.$$

Since $\mu(B_n) \geq 2\varepsilon$, we must have $\mu(C_1 \cap \cdots \cap C_n) \geq \varepsilon$, so $C_1 \cap \cdots \cap C_n \neq \emptyset$, and so $K_n = \bar{C}_1 \cap \cdots \cap \bar{C}_n \neq \emptyset$. Now $(K_n : n \in \mathbb{N})$ is a decreasing sequence of bounded non-empty closed sets in $\mathbb{R}$, so $\emptyset \neq \bigcap_n K_n \subseteq \bigcap_n B_n$, which is a contradiction.

(*Uniqueness.*) Let λ be any measure on $\mathcal{B}$ with $\lambda((a,b]) = b - a$ for all $a < b$. Fix n and consider

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

Then μ_n and λ_n are probability measures on $\mathcal{B}$ and $\mu_n = \lambda_n$ on the π-system of intervals of the form $(a,b]$, which generates $\mathcal{B}$. So, by Theorem 1.7.1, $\mu_n = \lambda_n$ on $\mathcal{B}$. Hence, for all $A \in \mathcal{B}$, we have

$$\mu(A) = \sum_n \mu_n(A) = \sum_n \lambda_n(A) = \lambda(A).$$

□

The condition which characterizes Lebesgue measure μ on $\mathcal{B}$ allows us to check that μ is *translation invariant*: define for $x \in \mathbb{R}$ and $B \in \mathcal{B}$

$$\mu_x(B) = \mu(B + x), \quad B + x = \{b + x : b \in \mathcal{B}\}$$

then $\mu_x((a,b]) = (b+x) - (a+x) = b - a$, so $\mu_x = \mu$, that is to say $\mu(B+x) = \mu(B)$.

The restriction of Lebesgue measure to $\mathcal{B}((0,1])$ has another sort of translation invariance, where now we understand $B + x$ as the subset of $(0,1]$ obtained after translation by x and reduction modulo 1. This can be checked by a similar argument.

If we inspect the proof of Carathéodory's Extension Theorem, and consider its application in Theorem 1.10.1, we see we have constructed not only a Borel measure μ but also an extension of μ to the set of outer measurable sets $\mathcal{M}$. In this context, the extension is also called Lebesgue measure and $\mathcal{M}$ is called the Lebesgue σ-algebra. In fact, the Lebesgue σ-algebra can be identified also as the set of all sets of the form $A \cup N$, where $A \in \mathcal{B}$ and $N \subseteq B$ for some $B \in \mathcal{B}$ with $\mu(B) = 0$. Moreover $\mu(A \cup N) = \mu(A)$ in this case.

1.11. **Existence of a non-Lebesgue-measurable subset of $\mathbb{R}$.** For $x, y \in [0,1)$, let us write $x \sim y$ if $x - y \in \mathbb{Q}$. Then $\sim$ is an equivalence relation. Using the Axiom of Choice, we can find a subset S of $[0,1)$ containing exactly one representative of each equivalence class. We will show that S cannot be Lebesgue measurable.

Set $Q = \mathbb{Q} \cap [0,1)$ and, for each $q \in Q$, define $S + q = \{s + q \pmod 1 \colon s \in S\}$. It is an easy exercise to check that the sets $S + q$ are all disjoint and their union is $[0,1)$. On the other hand, the Lebesgue σ-algebra and Lebesgue measure on $(0,1]$ are translation invariant for addition modulo 1. Hence, if we suppose that S is Lebesgue measurable, then so is $S + q$, with $\mu(S+q) = \mu(S)$. But then

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

which is impossible. Hence S is not Lebesgue measurable.

1.12. **Independence.** A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ provides a model for an experiment whose outcome is subject to chance, according to the following interpretation:

Ω is the set of possible outcomes
$\mathcal{F}$ is the set of observable sets of outcomes, or *events*
$\mathbb{P}(A)$ is the probability of the event A.

Relative to measure theory, probability theory is enriched by the significance attached to the notion of independence. Let I be a countable set. Say that a family $(A_i : i \in I)$ of events is *independent* if, for all finite subsets $J \subseteq I$,

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

Say that a family $(\mathcal{A}_i : i \in I)$ of sub-σ-algebras of $\mathcal{F}$ is *independent* if the family $(A_i : i \in I)$ is independent whenever $A_i \in \mathcal{A}_i$ for all i. Here is a useful way to establish the independence of two σ-algebras.

Theorem 1.12.1. *Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be π-systems contained in $\mathcal{F}$ and suppose that*

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

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

Proof. Fix $A_1 \in \mathcal{A}_1$ and define for $A \in \mathcal{F}$

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

Then μ and ν are measures which agree on the π-system $\mathcal{A}_2$, with $\mu(\Omega) = \nu(\Omega) = \mathbb{P}(A_1) < \infty$. So, by uniqueness of extension, for all $A_2 \in \sigma(\mathcal{A}_2)$,

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

Now fix $A_2 \in \sigma(\mathcal{A}_2)$ and repeat the argument with

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

to show that, for all $A_1 \in \sigma(\mathcal{A}_1)$,

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

□

1.13. **Borel-Cantelli lemmas.** Given a sequence of events $(A_n : n \in \mathbb{N})$, we may ask for the probability that infinitely many occur. Set

$$\limsup A_n = \bigcap_n \bigcup_{m \geq n} A_m, \quad \liminf A_n = \bigcup_n \bigcap_{m \geq n} A_m.$$

We sometimes write $\{A_n \text{ infinitely often}\}$ as an alternative for $\limsup A_n$, because $\omega \in \limsup A_n$ if and only if $\omega \in A_n$ for infinitely many n. Similarly, we write $\{A_n \text{ eventually}\}$ for $\liminf A_n$. The abbreviations i.o. and ev. are often used.

Lemma 1.13.1 (First Borel–Cantelli lemma). *If $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$.*

Proof. As $n \to \infty$ we have

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

□

We note that this argument is valid whether or not $\mathbb{P}$ is a probability measure.

Lemma 1.13.2 (Second Borel–Cantelli lemma). *Assume that the events $(A_n : n \in \mathbb{N})$ are independent. If $\sum_n \mathbb{P}(A_n) = \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 1$.*

Proof. We use the inequality $1 - a \leq e^{-a}$. The events $(A_n^c : n \in \mathbb{N})$ are also independent. Set $a_n = \mathbb{P}(A_n)$. For all n and for $N \geq n$ with $N \to \infty$ we have

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

Hence $\mathbb{P}(\bigcap_{m \geq n} A_m^c) = 0$ for all n, and so $\mathbb{P}(A_n \text{ i.o.}) = 1 - \mathbb{P}(\bigcup_n \bigcap_{m \geq n} A_m^c) = 1$. □

2. Measurable functions and random variables

2.1. **Measurable functions.** Let $(E, \mathcal{E})$ and $(G, \mathcal{G})$ be measurable spaces. A function $f : E \to G$ is *measurable* if $f^{-1}(A) \in \mathcal{E}$ whenever $A \in \mathcal{G}$. Here $f^{-1}(A)$ denotes the *inverse image* of A by f

$$f^{-1}(A) = \{x \in E : f(x) \in A\}.$$

In the case $(G, \mathcal{G}) = (\mathbb{R}, \mathcal{B})$ we simply call f a measurable function on E. In the case $(G, \mathcal{G}) = ([0, \infty], \mathcal{B}([0, \infty]))$ we call f a non-negative measurable function on E. This terminology is convenient but it has the consequence that some non-negative measurable functions are not (real-valued) measurable functions. If E is a topological space and $\mathcal{E} = \mathcal{B}(E)$, then a measurable function on E is called a *Borel* function. For any function $f : E \to G$, the inverse image preserves set operations

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

Therefore, the set $\{f^{-1}(A) : A \in \mathcal{G}\}$ is a σ-algebra on E and $\{A \subseteq G : f^{-1}(A) \in \mathcal{E}\}$ is a σ-algebra on G. In particular, if $\mathcal{G} = \sigma(\mathcal{A})$ and $f^{-1}(A) \in \mathcal{E}$ whenever $A \in \mathcal{A}$, then $\{A : f^{-1}(A) \in \mathcal{E}\}$ is a σ-algebra containing $\mathcal{A}$ and hence $\mathcal{G}$, so f is measurable. In the case $G = \mathbb{R}$, the Borel σ-algebra is generated by intervals of the form $(-\infty, y], y \in \mathbb{R}$, so, to show that $f : E \to \mathbb{R}$ is Borel measurable, it suffices to show that $\{x \in E : f(x) \leq y\} \in \mathcal{E}$ for all y.

If E is any topological space and $f : E \to \mathbb{R}$ is continuous, then $f^{-1}(U)$ is open in E and hence measurable, whenever U is open in $\mathbb{R}$; the open sets U generate $\mathcal{B}$, so *any continuous function is measurable.*

For $A \subseteq E$, the *indicator function* 1_A *of* A is the function $1_A : E \to \{0, 1\}$ which takes the value 1 on A and 0 otherwise. Note that *the indicator function of any measurable set is a measurable function.* Also, *the composition of measurable functions is measurable.*

Given any family of functions $f_i : E \to G, i \in I$, we can make them all measurable by taking

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

Then $\mathcal{E}$ is the *σ-algebra generated by* $(f_i : i \in I)$.

Proposition 2.1.1. *Let $(f_n : n \in \mathbb{N})$ be a sequence of non-negative measurable functions on E. Then the functions $f_1 + f_2$ and $f_1 f_2$ are measurable, and so are the following functions:*

$$\inf_n f_n, \quad \sup_n f_n, \quad \liminf_n f_n, \quad \limsup_n f_n.$$

The same conclusion holds for real-valued measurable functions provided the limit functions are also real-valued.

Theorem 2.1.2 (Monotone class theorem)**.** *Let $(E, \mathcal{E})$ be a measurable space and let $\mathcal{A}$ be a π-system generating $\mathcal{E}$. Let $\mathcal{V}$ be a vector space of bounded functions $f : E \to \mathbb{R}$ such that:*

(i) *$1 \in \mathcal{V}$ and $1_A \in \mathcal{V}$ for all $A \in \mathcal{A}$;*
(ii) *if $f_n \in \mathcal{V}$ for all n and f is bounded with $0 \le f_n \uparrow f$, then $f \in \mathcal{V}$.*

Then $\mathcal{V}$ contains every bounded measurable function.

Proof. Consider $\mathcal{D} = \{A \in \mathcal{E} : 1_A \in \mathcal{V}\}$. Then $\mathcal{D}$ is a d-system containing $\mathcal{A}$, so $\mathcal{D} = \mathcal{E}$. Since $\mathcal{V}$ is a vector space, it contains all finite linear combinations of indicator functions of measurable sets. If f is a bounded and non-negative measurable function, then the functions $f_n = 2^{-n}\lfloor 2^n f \rfloor, n \in \mathbb{N}$, belong to $\mathcal{V}$ and $0 \le f_n \uparrow f$, so $f \in \mathcal{V}$. Finally, any bounded measurable function is the difference of two non-negative such functions, hence in $\mathcal{V}$. □

2.2. **Image measures.** Let $(E, \mathcal{E})$ and $(G, \mathcal{G})$ be measurable spaces and let μ be a measure on $\mathcal{E}$. Then any measurable function $f : E \to G$ induces an *image measure* $\nu = \mu \circ f^{-1}$ on $\mathcal{G}$, given by

$$\nu(A) = \mu(f^{-1}(A)).$$

We shall construct some new measures from Lebesgue measure in this way.

Lemma 2.2.1. *Let $g : \mathbb{R} \to \mathbb{R}$ be non-constant, right-continuous and non-decreasing. Set $g(\pm\infty) = \lim_{x \to \pm\infty} g(x)$ and write $I = (g(-\infty), g(\infty))$. Define $f : I \to \mathbb{R}$*

by $f(x) = \inf\{y \in \mathbb{R} : x \leq g(y)\}$. Then f is left-continuous and non-decreasing. Moreover, for $x \in I$ and $y \in \mathbb{R}$,

$$f(x) \leq y \quad \textit{if and only if} \quad x \leq g(y).$$

Proof. Fix $x \in I$ and consider the set $J_x = \{y \in \mathbb{R} : x \leq g(y)\}$. Note that J_x is non-empty and is not the whole of $\mathbb{R}$. Since g is non-decreasing, if $y \in J_x$ and $y' \geq y$, then $y' \in J_x$. Since g is right-continuous, if $y_n \in J_x$ and $y_n \downarrow y$, then $y \in J_x$. Hence $J_x = [f(x), \infty)$ and $x \leq g(y)$ if and only if $f(x) \leq y$. For $x \leq x'$, we have $J_x \supseteq J_{x'}$ and so $f(x) \leq f(x')$. For $x_n \uparrow x$, we have $J_x = \cap_n J_{x_n}$, so $f(x_n) \to f(x)$. So f is left-continuous and non-decreasing, as claimed. □

Theorem 2.2.2. *Let $g : \mathbb{R} \to \mathbb{R}$ be non-constant, right-continuous and non-decreasing. Then there exists a unique Radon measure dg on $\mathbb{R}$ such that, for all $a, b \in \mathbb{R}$ with $a < b$,*

$$dg((a, b]) = g(b) - g(a).$$

Moreover, we obtain in this way all non-zero Radon measures on $\mathbb{R}$.

The measure dg is called the *Lebesgue-Stieltjes measure* associated with g.

Proof. Define I and f as in the lemma and let μ denote Lebesgue measure on I. Then f is Borel measurable and the induced measure $dg = \mu \circ f^{-1}$ on $\mathbb{R}$ satisfies

$$dg((a, b]) = \mu(\{x : f(x) > a \text{ and } f(x) \leq b\}) = \mu((g(a), g(b)]) = g(b) - g(a).$$

The argument used for uniqueness of Lebesgue measure shows that there is at most one Borel measure with this property. Finally, if ν is any Radon measure on $\mathbb{R}$, we can define $g : \mathbb{R} \to \mathbb{R}$, right-continuous and non-decreasing, by

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

Then $\nu((a, b]) = g(b) - g(a)$ whenever $a < b$, so $\nu = dg$ by uniqueness. □

2.3. **Random variables.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(E, \mathcal{E})$ be a measurable space. A measurable function $X : \Omega \to E$ is called a *random variable in E*. It has the interpretation of a quantity, or state, determined by chance. Where no space E is mentioned, it is assumed that X takes values in $\mathbb{R}$. The image measure $\mu_X = \mathbb{P} \circ X^{-1}$ is called the *law* or *distribution* of X. For real-valued random variables, μ_X is uniquely determined by its values on the π-system of intervals $((-\infty, x] : x \in \mathbb{R})$, given by

$$F_X(x) = \mu_X((-\infty, x]) = \mathbb{P}(X \leq x).$$

The function F_X is called the *distribution function of X*.

Note that $F = F_X$ is increasing and right-continuous, with

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

Let us call any function $F : \mathbb{R} \to [0,1]$ satisfying these conditions a *distribution function*.

Let $\Omega = (0,1)$. Write $\mathcal{F}$ for the Borel σ-algebra on Ω and $\mathbb{P}$ for the restriction of Lebesgue measure to $\mathcal{F}$. Then $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Let F be any distribution function. Define $X : \Omega \to \mathbb{R}$ by

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

Then, by Lemma 2.2.1, X is a random variable and $X(\omega) \leq x$ if and only if $\omega \leq F(x)$. So

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

Thus every distribution function is the distribution function of a random variable.

A countable family of random variables $(X_i : i \in I)$ is said to be *independent* if the family of σ-algebras $(\sigma(X_i) : i \in I)$ is independent. For a sequence $(X_n : n \in \mathbb{N})$ of real valued random variables, this is equivalent to the condition

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \dots \mathbb{P}(X_n \leq x_n)$$

for all $x_1, \dots, x_n \in \mathbb{R}$ and all n. A sequence of random variables $(X_n : n \geq 0)$ is often regarded as a *process* evolving in time. The σ-algebra generated by $X_0, \dots, X_n$

$$\mathcal{F}_n = \sigma(X_0, \dots, X_n)$$

contains those events depending (measurably) on $X_0, \dots, X_n$ and represents what is known about the process by time n.

2.4. **Rademacher functions.** We continue with the particular choice of probability space $(\Omega, \mathcal{F}, \mathbb{P})$ made in the preceding section. Provided that we forbid infinite sequences of 0's, each $\omega \in \Omega$ has a unique binary expansion

$$\omega = 0.\omega_1\omega_2\omega_3 \dots .$$

Define random variables $R_n : \Omega \to \{0,1\}$ by $R_n(\omega) = \omega_n$. Then

$$R_1 = 1_{(\frac{1}{2},1]}, \quad R_2 = 1_{(\frac{1}{4},\frac{1}{2}]} + 1_{(\frac{3}{4},1]}, \quad R_3 = 1_{(\frac{1}{8},\frac{1}{4}]} + 1_{(\frac{3}{8},\frac{1}{2}]} + 1_{(\frac{5}{8},\frac{3}{4}]} + 1_{(\frac{7}{8},1]}.$$

These are called the *Rademacher functions.* The random variables $R_1, R_2, \dots$ are independent and *Bernoulli*, that is to say

$$\mathbb{P}(R_n = 0) = \mathbb{P}(R_n = 1) = 1/2.$$

The strong law of large numbers (proved in §10) applies here to show that

$$\mathbb{P}\left(\left\{\omega \in (0,1) : \frac{|\{k \leq n : \omega_k = 1\}|}{n} \to \frac{1}{2}\right\}\right) = \mathbb{P}\left(\frac{R_1 + \dots + R_n}{n} \to \frac{1}{2}\right) = 1.$$

This is called Borel's normal number theorem: *almost every point in* $(0,1)$ *is* normal, *that is, has 'equal' proportions of* 0*'s and* 1*'s in its binary expansion.*

We now use a trick involving the Rademacher functions to construct on $\Omega = (0,1)$, not just one random variable, but an infinite sequence of independent random variables with given distribution functions.

Proposition 2.4.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space of Lebesgue measure on the Borel subsets of $(0, 1)$. Let $(F_n : n \in \mathbb{N})$ be a sequence of distribution functions. Then there exists a sequence $(X_n : n \in \mathbb{N})$ of independent random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that X_n has distribution function $F_{X_n} = F_n$ for all n.*

Proof. Choose a bijection $m : \mathbb{N}^2 \to \mathbb{N}$ and set $Y_{k,n} = R_{m(k,n)}$, where R_m is the mth Rademacher function. Set

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

Then $Y_1, Y_2, \ldots$ are independent and, for all n, for $i2^{-k} = 0.y_1 \ldots y_k$, we have

$$\mathbb{P}(i2^{-k} < Y_n \leq (i+1)2^{-k}) = \mathbb{P}(Y_{1,n} = y_1, \ldots, Y_{k,n} = y_k) = 2^{-k}$$

so $\mathbb{P}(Y_n \leq x) = x$ for all $x \in [0, 1]$. Set

$$G_n(y) = \inf\{x : y \leq F_n(x)\}$$

then, by Lemma 2.2.1, G_n is Borel and $G_n(y) \leq x$ if and only if $y \leq F_n(x)$. So, if we set $X_n = G_n(Y_n)$, then $X_1, X_2, \ldots$ are independent random variables on Ω and

$$\mathbb{P}(X_n \leq x) = \mathbb{P}(G_n(Y_n) \leq x) = \mathbb{P}(Y_n \leq F_n(x)) = F_n(x).$$

□

2.5. **Convergence of measurable functions and random variables.**

Let $(E, \mathcal{E}, \mu)$ be a measure space. A set $A \in \mathcal{E}$ is sometimes defined by a property shared by its elements. If $\mu(A^c) = 0$, then we say that this property holds *almost everywhere* (or *a.e.*). When $(E, \mathcal{E}, \mu)$ is a probability space, we say instead that the property holds *almost surely* (or *a.s.*). Thus, for a sequence of measurable functions $(f_n : n \in \mathbb{N})$, we say f_n *converges to* f *almost everywhere* to mean that

$$\mu(\{x \in E : f_n(x) \not\to f(x)\}) = 0.$$

If, on the other hand, we have that

$$\mu(\{x \in E : |f_n(x) - f(x)| > \varepsilon\}) \to 0, \quad \text{for all } \varepsilon > 0,$$

then we say f_n *converges to* f *in measure* or *in probability* when $\mu(E) = 1$. For a sequence $(X_n : n \in \mathbb{N})$ of (real-valued) random variables there is a third notion of convergence. We say that X_n *converges to* X *in distribution* if $F_{X_n}(x) \to F_X(x)$ as $n \to \infty$ at all points $x \in \mathbb{R}$ where F_X is continuous. Note that the last definition does not require the random variables to be defined on the same probability space.

Theorem 2.5.1. *Let $(f_n : n \in \mathbb{N})$ be a sequence of measurable functions.*

(a) *Assume that $\mu(E) < \infty$. If $f_n \to 0$ a.e., then $f_n \to 0$ in measure.*

(b) *If $f_n \to 0$ in measure, then $f_{n_k} \to 0$ a.e. for some subsequence (n_k).*

Proof. (a) Suppose $f_n \to 0$ a.e.. For each $\varepsilon > 0$,

$$\mu(|f_n| \leq \varepsilon) \geq \mu\left(\bigcap_{m \geq n} \{|f_m| \leq \varepsilon\}\right) \uparrow \mu(|f_n| \leq \varepsilon \text{ ev.}) \geq \mu(f_n \to 0) = \mu(E).$$

Hence $\mu(|f_n| > \varepsilon) \to 0$ and $f_n \to 0$ in measure.

(b) Suppose $f_n \to 0$ in measure, then we can find a subsequence (n_k) such that

$$\sum_k \mu(|f_{n_k}| > 1/k) < \infty.$$

So, by the first Borel–Cantelli lemma,

$$\mu(|f_{n_k}| > 1/k \text{ i.o.}) = 0$$

so $f_{n_k} \to 0$ a.e.. □

Theorem 2.5.2. *Let X and $(X_n : n \in \mathbb{N})$ be real-valued random variables.*

(a) *If X and $(X_n : n \in \mathbb{N})$ are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $X_n \to X$ in probability, then $X_n \to X$ in distribution.*

(b) *If $X_n \to X$ in distribution, then there are random variables $\tilde{X}$ and $(\tilde{X}_n : n \in \mathbb{N})$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\tilde{X}$ has the same distribution as X, $\tilde{X}_n$ has the same distribution as X_n for all n, and $\tilde{X}_n \to \tilde{X}$ almost surely.*

Proof. Write S for the subset of $\mathbb{R}$ where F_X is continuous.

(a) Suppose $X_n \to X$ in probability. Given $x \in S$ and $\varepsilon > 0$, there exists $\delta > 0$ such that $F_X(x - \delta) \geq F_X(x) - \varepsilon/2$ and $F_X(x + \delta) \leq F_X(x) + \varepsilon/2$. then there exists N such that, for all $n \geq N$, we have $\mathbb{P}(|X_n - X| > \delta) \leq \varepsilon/2$, which implies

$$F_{X_n}(x) \leq \mathbb{P}(X \leq x + \delta) + \mathbb{P}(|X_n - X| > \delta) \leq F_X(x) + \varepsilon$$

and

$$F_{X_n}(x) \geq \mathbb{P}(X \leq x - \delta) - \mathbb{P}(|X_n - X| > \delta) \geq F_X(x) - \varepsilon.$$

(b) Suppose now that $X_n \to X$ in distribution. Take $(\Omega, \mathcal{F}, \mathbb{P})$ to be the interval $(0, 1)$ equipped with its Borel σ-algebra and Lebesgue measure. Define for $\omega \in (0, 1)$

$$\tilde{X}_n(\omega) = \inf\{x \in \mathbb{R} : \omega \leq F_{X_n}(x)\}, \quad \tilde{X}(\omega) = \inf\{x \in \mathbb{R} : \omega \leq F_X(x)\}.$$

Then $\tilde{X}$ has the same distribution as X, and $\tilde{X}_n$ has the same distribution as X_n for all n. Write Ω_0 for the subset of $(0, 1)$ where $\tilde{X}$ is continuous. Since $\tilde{X}$ is non-decreasing, $(0, 1) \setminus \Omega_0$ is countable, so $\mathbb{P}(\Omega_0) = 1$. Since F_X is non-decreasing, $\mathbb{R} \setminus S$ is countable, so S is dense. Given $\omega \in \Omega_0$ and $\varepsilon > 0$, there exist $x^-, x^+ \in S$ with $x^- < \tilde{X}(\omega) < x^+$ and $x^+ - x^- < \varepsilon$, and there exists $\omega^+ \in (\omega, 1)$ such that $\tilde{X}(\omega^+) \leq x^+$. Then $F_X(x^-) < \omega$ and $F_X(x^+) \geq \omega^+ > \omega$. So there exists N such that, for all $n \geq N$, we have $F_{X_n}(x^-) < \omega$ and $F_{X_n}(x^+) \geq \omega$, which implies $\tilde{X}_n(\omega) > x^-$ and $\tilde{X}_n(\omega) \leq x^+$, and hence $|\tilde{X}_n(\omega) - \tilde{X}(\omega)| < \varepsilon$. □

2.6. **Tail events.** Let $(X_n : n \in \mathbb{N})$ be a sequence of random variables. Define

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

Then $\mathcal{T}$ is a σ-algebra, called the *tail σ-algebra* of $(X_n : n \in \mathbb{N})$. It contains the events which depend only on the limiting behaviour of the sequence.

Theorem 2.6.1 (Kolmogorov's zero-one law)**.** *Suppose that $(X_n : n \in \mathbb{N})$ is a sequence of independent random variables. Then the tail σ-algebra $\mathcal{T}$ of $(X_n : n \in \mathbb{N})$ contains only events of probability* 0 *or* 1*. Moreover, any $\mathcal{T}$-measurable random variable is almost surely constant.*

Proof. Set $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$. Then $\mathcal{F}_n$ is generated by the π-system of events

$$A = \{X_1 \le x_1, \dots, X_n \le x_n\}$$

whereas $\mathcal{T}_n$ is generated by the π-system of events

$$B = \{X_{n+1} \le x_{n+1}, \dots, X_{n+k} \le x_{n+k}\}, \quad k \in \mathbb{N}.$$

We have $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for all such A and B, by independence. Hence $\mathcal{F}_n$ and $\mathcal{T}_n$ are independent, by Theorem 1.12.1. It follows that $\mathcal{F}_n$ and $\mathcal{T}$ are independent. Now $\bigcup_n \mathcal{F}_n$ is a π-system which generates the σ-algebra $\mathcal{F}_\infty = \sigma(X_n : n \in \mathbb{N})$. So by Theorem 1.12.1 again, $\mathcal{F}_\infty$ and $\mathcal{T}$ are independent. But $\mathcal{T} \subseteq \mathcal{F}_\infty$. So, if $A \in \mathcal{T}$,

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A)$$

so $\mathbb{P}(A) \in \{0, 1\}$.

Finally, if Y is any $\mathcal{T}$-measurable random variable, then $F_Y(y) = \mathbb{P}(Y \le y)$ takes values in $\{0, 1\}$, so $\mathbb{P}(Y = c) = 1$, where $c = \inf\{y : F_Y(y) = 1\}$. □

2.7. **Large values in sequences of independent identically distributed random variables.** Consider a sequence $(X_n : n \in \mathbb{N})$ of independent random variables, all having the same distribution function F. Assume that $F(x) < 1$ for all $x \in \mathbb{R}$. Then, almost surely, the sequence $(X_n : n \in \mathbb{N})$ is unbounded above, so $\limsup_n X_n = \infty$. A way to describe the occurrence of large values in the sequence is to find a function $g : \mathbb{N} \to (0, \infty)$ such that, almost surely,

$$\limsup_n X_n / g(n) = 1.$$

We now show that $g(n) = \log n$ is the right choice when $F(x) = 1 - e^{-x}$. The same method adapts to other distributions.

Fix $\alpha > 0$ and consider the event $A_n = \{X_n \ge \alpha \log n\}$. Then $\mathbb{P}(A_n) = e^{-\alpha \log n} = n^{-a}$, so the series $\sum_n \mathbb{P}(A_n)$ converges if and only if $\alpha > 1$. By the Borel–Cantelli Lemmas, we deduce that, for all $\varepsilon > 0$,

$$\mathbb{P}(X_n / \log n \ge 1 \text{ i.o.}) = 1, \quad \mathbb{P}(X_n / \log n \ge 1 + \varepsilon \text{ i.o.}) = 0.$$

Hence, almost surely,

$$\limsup_n X_n / \log n = 1.$$

3. Integration

3.1. **Definition of the integral and basic properties.** Let $(E, \mathcal{E}, \mu)$ be a measure space. We shall define for non-negative measurable functions f on E, and (under a natural condition) for (real-valued) measurable functions f on E, the *integral* of f, to be denoted

$$\mu(f) = \int_E f d\mu = \int_E f(x)\mu(dx).$$

When $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B})$ and μ is Lebesgue measure, the usual notation is

$$\mu(f) = \int_{\mathbb{R}} f(x)dx.$$

For a random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the integral is usually called instead the *expectation* of X and written $\mathbb{E}(X)$.

A *simple* function is one of the form

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

where $0 \leq a_k < \infty$ and $A_k \in \mathcal{E}$ for all k, and where $m \in \mathbb{N}$. For simple functions f, we define

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

where we adopt the convention $0.\infty = 0$. Although the representation of f is not unique, it is straightforward to check that $\mu(f)$ is well defined and, for simple functions f, g and constants $\alpha, \beta \geq 0$, we have

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

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

(c) $f = 0$ a.e. if and only if $\mu(f) = 0$.

We define the integral $\mu(f)$ of a non-negative measurable function f by

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

This is consistent with the definition for simple functions by property (b) above. Note that, for all non-negative measurable functions f, g with $f \leq g$, we have $\mu(f) \leq \mu(g)$. For any measurable function f, set $f^+ = f \vee 0$ and $f^- = (-f) \vee 0$. Then $f = f^+ - f^-$ and $|f| = f^+ + f^-$. If $\mu(|f|) < \infty$, then we say that f is *integrable* and define

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

Note that $|\mu(f)| \leq \mu(|f|)$ for all integrable functions f. We sometimes define the integral $\mu(f)$ by the same formula, even when f is not integrable, but when one of $\mu(f^-)$ and $\mu(f^+)$ is finite. In such cases the integral takes the value ∞ or $-\infty$.

Here is the key result for the theory of integration. For $x \in [0, \infty]$ and a sequence $(x_n : n \in \mathbb{N})$ in $[0, \infty]$, we write $x_n \uparrow x$ to mean that $x_n \leq x_{n+1}$ for all n and $x_n \to x$ as $n \to \infty$. For a non-negative function f on E and a sequence of such functions $(f_n : n \in \mathbb{N})$, we write $f_n \uparrow f$ to mean that $f_n(x) \uparrow f(x)$ for all $x \in E$.

Theorem 3.1.1 (Monotone convergence). *Let f be a non-negative measurable function and let $(f_n : n \in \mathbb{N})$ be a sequence of such functions. Suppose that $f_n \uparrow f$. Then $\mu(f_n) \uparrow \mu(f)$.*

Proof. Case 1: $f_n = 1_{A_n}, f = 1_A$.
The result is a simple consequence of countable additivity.
Case 2: *f_n simple, $f = 1_A$.*
Fix $\varepsilon > 0$ and set $A_n = \{f_n > 1 - \varepsilon\}$. Then $A_n \uparrow A$ and

$$(1 - \varepsilon)1_{A_n} \leq f_n \leq 1_A$$

so

$$(1 - \varepsilon)\mu(A_n) \leq \mu(f_n) \leq \mu(A).$$

But $\mu(A_n) \uparrow \mu(A)$ by Case 1 and $\varepsilon > 0$ was arbitrary, so the result follows.

Case 3: *f_n simple, f simple.*
We can write f in the form

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

with $a_k > 0$ for all k and the sets A_k disjoint. Then $f_n \uparrow f$ implies

$$a_k^{-1} 1_{A_k} f_n \uparrow 1_{A_k}$$

so, by Case 2,

$$\mu(f_n) = \sum_k \mu(1_{A_k} f_n) \uparrow \sum_k a_k \mu(A_k) = \mu(f).$$

Case 4: *f_n simple, $f \geq 0$ measurable.*
Let g be simple with $g \leq f$. Then $f_n \uparrow f$ implies $f_n \wedge g \uparrow g$ so, by Case 3,

$$\mu(f_n) \geq \mu(f_n \wedge g) \uparrow \mu(g).$$

Since g was arbitrary, the result follows.
Case 5: *$f_n \geq 0$ measurable, $f \geq 0$ measurable.*
Set $g_n = (2^{-n}\lfloor 2^n f_n \rfloor) \wedge n$ then g_n is simple and $g_n \leq f_n \leq f$, so

$$\mu(g_n) \leq \mu(f_n) \leq \mu(f).$$

But $f_n \uparrow f$ forces $g_n \uparrow f$, so $\mu(g_n) \uparrow \mu(f)$, by Case 4, and so $\mu(f_n) \uparrow \mu(f)$. □

Second proof. Set $M = \sup_n \mu(f_n)$. We know that

$$\mu(f_n) \uparrow M \leq \mu(f) = \sup\{\mu(g) : g \text{ simple }, g \leq f\}$$

so it will suffice to show that $\mu(g) \leq M$ for all simple functions

$$g = \sum_{i=1}^{m} a_k 1_{A_k} \leq f.$$

Without loss of generality, we may assume that the sets A_k are disjoint. Define functions g_n by

$$g_n(x) = \left(2^{-n}\lfloor 2^n f_n(x)\rfloor\right) \wedge g(x), \quad x \in E.$$

Then g_n is simple and $g_n \leq f_n$ for all n. Fix $\varepsilon \in (0,1)$ and consider the sets

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

Now $g_n \uparrow g$ and $g = a_k$ on A_k, so $A_k(n) \uparrow A_k$, and so $\mu(A_k(n) \uparrow \mu(A_k)$ by countable additivity. Also, we have

$$1_{A_k} g_n \geq (1-\varepsilon)a_k 1_{A_k(n)}$$

so

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

Finally, we have

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

and the integral is additive on simple functions, so

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

But $\mu(g_n) \leq \mu(f_n) \leq M$ for all n and $\varepsilon \in (0,1)$ is arbitrary, so we see that $\mu(g) \leq M$, as required. □

Theorem 3.1.2. *For all non-negative measurable functions f, g and all constants $\alpha, \beta \geq 0$,*

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

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

(c) $f = 0$ *a.e.* *if and only if* $\mu(f) = 0$.

Proof. Define simple functions f_n, g_n by

$$f_n = (2^{-n}\lfloor 2^n f\rfloor) \wedge n, \quad g_n = (2^{-n}\lfloor 2^n g\rfloor) \wedge n.$$

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

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

We know that $\mu(\alpha f_n + \beta g_n) = \alpha\mu(f_n) + \beta\mu(g_n)$, so we obtain (a) on letting $n \to \infty$. As we noted above, (b) is obvious from the definition of the integral. If $f = 0$ a.e., then $f_n = 0$ a.e., for all n, so $\mu(f_n) = 0$ and $\mu(f) = 0$. On the other hand, if $\mu(f) = 0$, then $\mu(f_n) = 0$ for all n, so $f_n = 0$ a.e. and $f = 0$ a.e.. □

Theorem 3.1.3. *For all integrable functions f, g and all constants $\alpha, \beta \in \mathbb{R}$,*

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

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

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

Proof. We note that $\mu(-f) = -\mu(f)$. For $\alpha \geq 0$, we have

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

If $h = f + g$ then $h^+ + f^- + g^- = h^- + f^+ + g^+$, so

$$\mu(h^+) + \mu(f^-) + \mu(g^-) = \mu(h^-) + \mu(f^+) + \mu(g^+)$$

and so $\mu(h) = \mu(f)+\mu(g)$. That proves (a). If $f \leq g$ then $\mu(g)-\mu(f) = \mu(g-f) \geq 0$, by (a). Finally, if $f = 0$ a.e., then $f^\pm = 0$ a.e., so $\mu(f^\pm) = 0$ and so $\mu(f) = 0$. □

Note that in Theorem 3.1.3(c) we lose the reverse implication. The following result is sometimes useful:

Proposition 3.1.4. *Let $\mathcal{A}$ be a π-system containing E and generating $\mathcal{E}$. Then, for any integrable function f,*

$$\mu(f 1_A) = 0 \textit{ for all } A \in \mathcal{A} \quad \textit{implies} \quad f = 0 \textit{ a.e.}.$$

Here are some minor variants on the monotone convergence theorem.

Proposition 3.1.5. *Let $(f_n : n \in \mathbb{N})$ be a sequence of non-negative measurable functions. Then*

$$f_n \uparrow f \textit{ a.e.} \quad \Longrightarrow \quad \mu(f_n) \uparrow \mu(f).$$

Proposition 3.1.6. *Let $(g_n : n \in \mathbb{N})$ be a sequence of non-negative measurable functions. Then*

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

This reformulation of monotone convergence makes it clear that it is the counterpart for the integration of functions of the countable additivity property of the measure on sets.

3.2. **Integrals and limits.** In the monotone convergence theorem, the hypothesis that the given sequence of functions is non-decreasing is essential. In this section we obtain some results on the integrals of limits of functions without such a hypothesis.

Lemma 3.2.1 (Fatou's lemma). *Let $(f_n : n \in \mathbb{N})$ be a sequence of non-negative measurable functions. Then*

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

Proof. For $k \geq n$, we have

$$\inf_{m \geq n} f_m \leq f_k$$

so

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

But, as $n \to \infty$,

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

so, by monotone convergence,

$$\mu(\inf_{m \geq n} f_m) \uparrow \mu(\liminf f_n).$$

□

Theorem 3.2.2 (Dominated convergence). *Let f be a measurable function and let $(f_n : n \in \mathbb{N})$ be a sequence of such functions. Suppose that $f_n(x) \to f(x)$ for all $x \in E$ and that $|f_n| \leq g$ for all n, for some integrable function g. Then f and f_n are integrable, for all n, and $\mu(f_n) \to \mu(f)$.*

Proof. The limit f is measurable and $|f| \leq g$, so $\mu(|f|) \leq \mu(g) < \infty$, so f is integrable. We have $0 \leq g \pm f_n \to g \pm f$ so certainly $\liminf(g \pm f_n) = g \pm f$. By Fatou's lemma,

$$\mu(g) + \mu(f) = \mu(\liminf(g + f_n)) \leq \liminf \mu(g + f_n) = \mu(g) + \liminf \mu(f_n),$$

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

Since $\mu(g) < \infty$, we can deduce that

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

This proves that $\mu(f_n) \to \mu(f)$ as $n \to \infty$. □

3.3. **Transformations of integrals.**

Proposition 3.3.1. *Let $(E, \mathcal{E}, \mu)$ be a measure space and let $A \in \mathcal{E}$. Then the set $\mathcal{E}_A$ of measurable subsets of A is a σ-algebra and the restriction μ_A of μ to $\mathcal{E}_A$ is a measure. Moreover, for any non-negative measurable function f on E, we have*

$$\mu(f 1_A) = \mu_A(f|_A).$$

In the case of Lebesgue measure on $\mathbb{R}$, we write, for any interval I with $\inf I = a$ and $\sup I = b$,

$$\int_{\mathbb{R}} f 1_I(x)dx = \int_I f(x)dx = \int_a^b f(x)dx.$$

Note that the sets $\{a\}$ and $\{b\}$ have measure zero, so we do not need to specify whether they are included in I or not.

Proposition 3.3.2. *Let $(E, \mathcal{E})$ and $(G, \mathcal{G})$ be measure spaces and let $f : E \to G$ be a measurable function. Given a measure μ on $(E, \mathcal{E})$, define $\nu = \mu \circ f^{-1}$, the image measure on $(G, \mathcal{G})$. Then, for all non-negative measurable functions g on G,*

$$\nu(g) = \mu(g \circ f).$$

In particular, for a G-valued random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, for any non-negative measurable function g on G, we have

$$\mathbb{E}(g(X)) = \mu_X(g).$$

Proposition 3.3.3. *Let $(E, \mathcal{E}, \mu)$ be a measure space and let f be a non-negative measurable function on E. Define $\nu(A) = \mu(f 1_A), A \in \mathcal{E}$. Then ν is a measure on E and, for all non-negative measurable functions g on E,*

$$\nu(g) = \mu(fg).$$

In particular, to each non-negative Borel function f on $\mathbb{R}$, there corresponds a Borel measure μ on $\mathbb{R}$ given by $\mu(A) = \int_A f(x)dx$. Then, for all non-negative Borel functions g,

$$\mu(g) = \int_{\mathbb{R}^n} g(x)f(x)dx.$$

We say that μ *has density f (with respect to Lebesgue measure).*

If the law μ_X of a real-valued random variable X has a density f_X, then we call f_X a *density function* for X. Then $\mathbb{P}(X \in A) = \int_A f_X(x)dx$, for all Borel sets A, and, for for all non-negative Borel functions g on $\mathbb{R}$,

$$\mathbb{E}(g(X)) = \mu_X(g) = \int_{\mathbb{R}} g(x)f_X(x)dx.$$

3.4. **Fundamental theorem of calculus.** We show that integration with respect to Lebesgue measure on $\mathbb{R}$ acts as an inverse to differentiation. Since we restrict here to the integration of continuous functions, the proof is the same as for the Riemann integral.

Theorem 3.4.1 (Fundamental theorem of calculus)**.**

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

$$F_a(t) = \int_a^t f(x)dx.$$

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

(b) *Let $F : [a, b] \to \mathbb{R}$ be differentiable with continuous derivative f. Then*

$$\int_a^b f(x)dx = F(b) - F(a).$$

Proof. Fix $t \in [a, b)$. Given $\varepsilon > 0$, there exists $\delta > 0$ such that $|f(x) - f(t)| \le \varepsilon$ whenever $|x - t| \le \delta$. So, for $0 < h \le \delta$,

$$\left|\frac{F_a(t+h) - F_a(t)}{h} - f(t)\right| = \frac{1}{h}\left|\int_t^{t+h} (f(x) - f(t))dx\right|$$
$$\le \frac{1}{h}\int_t^{t+h} |f(x) - f(t)|dx \le \frac{\varepsilon}{h}\int_t^{t+h} dx = \varepsilon.$$

Hence F_a is differentiable on the right at t with derivative $f(t)$. Similarly, for all $t \in (a, b]$, F_a is differentiable on the left at t with derivative $f(t)$. Finally, $(F - F_a)'(t) = 0$ for all $t \in (a, b)$ so $F - F_a$ is constant (by the mean value theorem), and so

$$F(b) - F(a) = F_a(b) - F_a(a) = \int_a^b f(x)dx.$$

□

Proposition 3.4.2. *Let $\phi : [a, b] \to \mathbb{R}$ be continuously differentiable and strictly increasing. Then, for all non-negative Borel functions g on $[\phi(a), \phi(b)]$,*

$$\int_{\phi(a)}^{\phi(b)} g(y)dy = \int_a^b g(\phi(x))\phi'(x)dx.$$

The proposition can be proved as follows. First, the case where g is the indicator function of an interval follows from the Fundamental Theorem of Calculus. Next, show that the set of Borel sets B such that the conclusion holds for $g = 1_B$ is a d-system, which must then be the whole Borel σ-algebra by Dynkin's lemma. The identity extends to simple functions by linearity and then to all non-negative measurable functions g by monotone convergence, using approximating simple functions $(2^{-n}\lfloor 2^n g\rfloor) \wedge n$.

A general formulation of this procedure, which is often used, is given in the monotone class theorem Theorem 2.1.2.

3.5. **Differentiation under the integral sign.** Integration in one variable and differentiation in another can be interchanged subject to some regularity conditions.

Theorem 3.5.1 (Differentiation under the integral sign)**.** *Let $U \subseteq \mathbb{R}$ be open and suppose that $f : U \times E \to \mathbb{R}$ satisfies:*

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

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

(iii) *for some integrable function g, for all $x \in E$ and all $t \in U$,*

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

Then the function $x \mapsto (\partial f/\partial t)(t,x)$ is integrable for all t. Moreover, the function $F : U \to \mathbb{R}$, defined by

$$F(t) = \int_E f(t,x)\mu(dx),$$

is differentiable and

$$\frac{d}{dt}F(t) = \int_E \frac{\partial f}{\partial t}(t,x)\mu(dx).$$

Proof. Take any sequence $h_n \to 0$ and set

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

Then $g_n(x) \to 0$ for all $x \in E$ and, by the mean value theorem, $|g_n| \leq 2g$ for all n. In particular, for all t, the function $x \mapsto (\partial f/\partial t)(t,x)$ is the limit of measurable functions, hence measurable, and hence integrable, by (iii).Then, by dominated convergence,

$$\frac{F(t+h_n) - F(t)}{h_n} - \int_E \frac{\partial f}{\partial t}(t,x)\mu(dx) = \int_E g_n(x)\mu(dx) \to 0.$$

□

3.6. **Product measure and Fubini's theorem.** Let $(E_1, \mathcal{E}_1, \mu_1)$ and $(E_2, \mathcal{E}_2, \mu_2)$ be *finite* measure spaces. The set

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

is a π-system of subsets of $E = E_1 \times E_2$. Define the *product σ-algebra*

$$\mathcal{E}_1 \otimes \mathcal{E}_2 = \sigma(\mathcal{A}).$$

Set $\mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2$.

Lemma 3.6.1. *Let $f : E \to \mathbb{R}$ be $\mathcal{E}$-measurable. Then, for all $x_1 \in E_1$, the function $x_2 \mapsto f(x_1, x_2) : E_2 \to \mathbb{R}$ is $\mathcal{E}_2$-measurable.*

Proof. Denote by $\mathcal{V}$ the set of bounded $\mathcal{E}$-measurable functions for which the conclusion holds. Then $\mathcal{V}$ is a vector space, containing the indicator function 1_A of every set $A \in \mathcal{A}$. Moreover, if $f_n \in \mathcal{V}$ for all n and if f is bounded with $0 \leq f_n \uparrow f$, then also $f \in \mathcal{V}$. So, by the monotone class theorem, $\mathcal{V}$ contains all bounded $\mathcal{E}$-measurable functions. The rest is easy. □

Lemma 3.6.2. *Let f be a bounded or non-negative measurable function on E. Define for $x_1 \in E_1$*

$$f_1(x_1) = \int_{E_2} f(x_1, x_2)\mu_2(dx_2).$$

If f is bounded then $f_1 : E_1 \to \mathbb{R}$ is a bounded $\mathcal{E}_1$-measurable function. On the other hand, if f is non-negative, then $f_1 : E_1 \to [0, \infty]$ is also an $\mathcal{E}_1$-measurable function.

Proof. Apply the monotone class theorem, as in the preceding lemma. Note that finiteness of μ_2 is needed for the boundedness of f_1 when f is bounded. □

Theorem 3.6.3 (Product measure)**.** *There exists a unique measure $\mu = \mu_1 \otimes \mu_2$ on $\mathcal{E}$ such that*

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

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

Proof. Uniqueness holds because $\mathcal{A}$ is a π-system generating $\mathcal{E}$. For existence, by the lemmas, we can define

$$\mu(A) = \int_{E_1} \left(\int_{E_2} 1_A(x_1, x_2)\mu_2(dx_2) \right) \mu_1(dx_1)$$

and use monotone convergence to see that μ is countably additive. □

Proposition 3.6.4. *Let $\hat{\mathcal{E}} = \mathcal{E}_2 \otimes \mathcal{E}_1$ and $\hat{\mu} = \mu_2 \otimes \mu_1$. For a function f on $E_1 \times E_2$, write $\hat{f}$ for the function on $E_2 \times E_1$ given by $\hat{f}(x_2, x_1) = f(x_1, x_2)$. Let f be a non-negative $\mathcal{E}$-measurable function. Then $\hat{f}$ is a non-negative $\hat{\mathcal{E}}$-measurable function and $\hat{\mu}(\hat{f}) = \mu(f)$.*

Theorem 3.6.5 (Fubini's theorem)**.**

(a) *Let f be a non-negative $\mathcal{E}$-measurable function. Then*

$$\mu(f) = \int_{E_1} \left(\int_{E_2} f(x_1, x_2)\mu_2(dx_2) \right) \mu_1(dx_1).$$

(b) *Let f be a μ-integrable function. Define*

$$A_1 = \{x_1 \in E_1 : \int_{E_2} |f(x_1, x_2)|\mu_2(dx_2) < \infty\}$$

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

$$f_1(x_1) = \int_{E_2} f(x_1, x_2)\mu_2(dx_2)$$

for $x_1 \in A_1$ and $f_1(x_1) = 0$ otherwise. Then $\mu_1(E_1 \setminus A_1) = 0$ and f_1 is μ_1-integrable with $\mu_1(f_1) = \mu(f)$.

Note that the *iterated integral* in (a) is well defined, for all bounded or non-negative measurable functions f, by Lemmas 3.6.1 and 3.6.2. Note also that, in combination with Proposition 3.6.4, Fubini's theorem allows us to interchange the order of integration in multiple integrals,whenever the integrand is non-negative or μ-integrable.

Proof. The conclusion of (a) holds for $f = 1_A$ with $A \in \mathcal{E}$ by definition of the product measure μ. It extends to simple functions on E by linearity of the integrals. For f non-negative measurable, consider the sequence of simple functions $f_n = (2^{-n}\lfloor 2^n f \rfloor) \wedge n$. Then (a) holds for f_n and $f_n \uparrow f$. By monotone convergence $\mu(f_n) \uparrow \mu(f)$ and, for all $x_1 \in E_1$,

$$\int_{E_2} f_n(x_1, x_2)\mu_2(dx_2) \uparrow \int_{E_2} f(x_1, x_2)\mu_2(dx_2)$$

and hence

$$\int_{E_1} \left(\int_{E_2} f_n(x_1, x_2)\mu_2(dx_2) \right) \mu_1(dx_1) \uparrow \int_{E_1} \left(\int_{E_2} f(x_1, x_2)\mu_2(dx_2) \right) \mu_1(dx_1).$$

Hence (a) extends to f.

Suppose now that f is μ-integrable. By Lemma 3.6.2, the function

$$x_1 \mapsto \int_{E_2} |f(x_1, x_2)|\mu_2(dx_2) : E_1 \to [0, \infty]$$

is $\mathcal{E}_1$-measurable, and it is then integrable because, using (a),

$$\int_{E_1} \left(\int_{E_2} |f(x_1, x_2)|\mu_2(dx_2) \right) \mu_1(dx_1) = \mu(|f|) < \infty.$$

Hence $A_1 \in \mathcal{E}_1$ and $\mu_1(E_1 \setminus A_1) = 0$. We see also that f_1 is well defined and, if we set

$$f_1^{(\pm)}(x_1) = \int_{E_2} f^{\pm}(x_1, x_2)\mu_2(dx_2)$$

then $f_1 = (f_1^{(+)} - f_1^{(-)})1_{A_1}$. Finally, by part (a),

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

as required. □

The existence of product measure and Fubini's theorem extend easily to σ-finite measure spaces. The operation of taking the product of two measure spaces is associative, by a π-system uniqueness argument. So we can, by induction, take the product of a finite number, without specifying the order. The measure obtained by taking the n-fold product of Lebesgue measure on $\mathbb{R}$ is called *Lebesgue measure on* $\mathbb{R}^n$. The corresponding integral is written

$$\int_{\mathbb{R}^n} f(x)dx.$$

3.7. **Laws of independent random variables.** Recall that a family $X_1, \dots, X_n$ of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be independent if the family of σ-algebras $\sigma(X_1), \dots, \sigma(X_n)$ is independent.

Proposition 3.7.1. *Let $X_1, \dots, X_n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, with values in $(E_1, \mathcal{E}_1), \dots, (E_n, \mathcal{E}_n)$ say. Set $E = E_1 \times \cdots \times E_n$ and $\mathcal{E} = \mathcal{E}_1 \otimes \cdots \otimes \mathcal{E}_n$. Consider the function $X : \Omega \to E$ given by $X(\omega) = (X_1(\omega), \dots, X_n(\omega))$. Then X is $\mathcal{E}$-measurable. Moreover, the following are equivalent:*

(a) *$X_1, \dots, X_n$ are independent;*

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

(c) *for all bounded measurable functions $f_1, \dots, f_n$ we have*

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

Proof. Set $\nu = \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}$. Consider the π-system $\mathcal{A} = \{\prod_{k=1}^n A_k : A_k \in \mathcal{E}_k\}$. If (a) holds, then for all $A \in \mathcal{A}$ we have

$$\mu_X(A) = \mathbb{P}(X \in A) = \mathbb{P}(\cap_{k=1}^n \{X_k \in A_k\}) = \prod_{k=1}^n \mathbb{P}(X_k \in A_k) = \prod_{k=1}^n \mu_{X_k}(A_k) = \nu(A)$$

and since $\mathcal{A}$ generates $\mathcal{E}$ this implies that $\mu = \nu$ on $\mathcal{E}$, so (b) holds. If (b) holds, then by Fubini's theorem

$$\mathbb{E}\left(\prod_{k=1}^n f_k(X_k)\right) = \int_E \prod_{k=1}^n f_k(x_k)\mu_{X_k}(dx_k) = \prod_{k=1}^n \int_{E_k} f_k(x_k)\mu_{X_k}(dx_k) = \prod_{k=1}^n \mathbb{E}(f_k(X_k))$$

so (c) holds. Finally (a) follows from (c) by taking $f_k = 1_{A_k}$ with $A_k \in \mathcal{E}_k$. □

4. Norms and inequalities

4.1. **L^p-norms.** Let $(E, \mathcal{E}, \mu)$ be a measure space. For $1 \le p < \infty$, we denote by $L^p = L^p(E, \mathcal{E}, \mu)$ the set of measurable functions f with finite *L^p-norm*:

$$\|f\|_p = \left(\int_E |f|^p d\mu\right)^{1/p} < \infty.$$

We denote by $L^\infty = L^\infty(E, \mathcal{E}, \mu)$ the set of measurable functions f with finite *L^∞-norm*:

$$\|f\|_\infty = \inf\{\lambda : |f| \le \lambda \text{ a.e.}\}.$$

Note that $\|f\|_p \le \mu(E)^{1/p}\|f\|_\infty$ for all $1 \le p < \infty$. For $1 \le p \le \infty$ and $f_n, f \in L^p$, we say that f_n *converges to f in L^p* if $\|f_n - f\|_p \to 0$.

4.2. **Chebyshev's inequality.** Let f be a non-negative measurable function and let $\lambda \geq 0$. We use the notation $\{f \geq \lambda\}$ for the set $\{x \in E : f(x) \geq \lambda\}$. Note that

$$\lambda 1_{\{f \geq \lambda\}} \leq f$$

so on integrating we obtain *Chebyshev's inequality*

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

Now let g be any measurable function. We can deduce inequalities for g by choosing some non-negative measurable function ϕ and applying Chebyshev's inequality to $f = \phi \circ g$. For example, if $g \in L^p, p < \infty$ and $\lambda > 0$, then

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

So we obtain the *tail estimate*

$$\mu(|g| \geq \lambda) = O(\lambda^{-p}), \quad \text{as } \lambda \to \infty.$$

4.3. **Jensen's inequality.** Let $I \subseteq \mathbb{R}$ be an interval. A function $c : I \to \mathbb{R}$ is *convex* if, for all $x, y \in I$ and $t \in [0, 1]$,

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

Lemma 4.3.1. *Let $c : I \to \mathbb{R}$ be convex and let m be a point in the interior of I. Then there exist $a, b \in \mathbb{R}$ such $c(x) \geq ax + b$ for all x, with equality at $x = m$.*

Proof. By convexity, for $m, x, y \in I$ with $x < m < y$, we have

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

So, fixing an interior point m, there exists $a \in \mathbb{R}$ such that, for all $x < m$ and all $y > m$

$$\frac{c(m) - c(x)}{m - x} \leq a \leq \frac{c(y) - c(m)}{y - m}.$$

Then $c(x) \geq a(x - m) + c(m)$, for all $x \in I$. □

Theorem 4.3.2 (Jensen's inequality)**.** *Let X be an integrable random variable with values in I and 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. The case where X is almost surely constant is easy. We exclude it. Then $m = \mathbb{E}(X)$ must lie in the interior of I. Choose $a, b \in \mathbb{R}$ as in the lemma. Then $c(X) \geq aX + b$. In particular $\mathbb{E}(c(X)^-) \leq |a|\mathbb{E}(|X|) + |b| < \infty$, so $\mathbb{E}(c(X))$ is well defined. Moreover

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

□

We deduce from Jensen's inequality *the monotonicity of L^p-norms with respect to a probability measure.* Let $1 \leq p < q < \infty$. Set $c(x) = |x|^{q/p}$, then c is convex on $\mathbb{R}$. So, for any $X \in L^p(\mathbb{P})$,

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

In particular, $L^p(\mathbb{P}) \supseteq L^q(\mathbb{P})$.

4.4. **Hölder's inequality and Minkowski's inequality.** For $p, q \in [1, \infty]$, we say that p and q are *conjugate indices* if

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Theorem 4.4.1 (Hölder's inequality). *Let $p, q \in (1, \infty)$ be conjugate indices. Then, for all measurable functions f and g, we have*

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

Proof. The cases where $\|f\|_p = 0$ or $\|f\|_p = \infty$ are obvious. We exclude them. Then, by multiplying f by an appropriate constant, we are reduced to the case where $\|f\|_p = 1$. So we can define a probability measure $\mathbb{P}$ on $\mathcal{E}$ by

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

For measurable functions $X \geq 0$,

$$\mathbb{E}(X) = \mu(X|f|^p), \quad \mathbb{E}(X) \leq \mathbb{E}(X^q)^{1/q}.$$

Note that $q(p-1) = p$. Then

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

□

Theorem 4.4.2 (Minkowski's inequality). *For $p \in [1, \infty)$ and measurable functions f and g, we have*

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

Proof. The cases where $p = 1$ or where $\|f\|_p = \infty$ or $\|g\|_p = \infty$ are easy. We exclude them. Then, since $|f + g|^p \leq 2^p(|f|^p + |g|^p)$, we have

$$\mu(|f + g|^p) \leq 2^p\{\mu(|f|^p) + \mu(|g|^p)\} < \infty.$$

The case where $\|f + g\|_p = 0$ is clear, so let us assume $\|f + g\|_p > 0$. Observe that

$$\||f + g|^{p-1}\|_q = \mu(|f + g|^{(p-1)q})^{1/q} = \mu(|f + g|^p)^{1-1/p}.$$

So, by Hölder's inequality,

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

The result follows on dividing both sides by $\||f+g|^{p-1}\|_q$. □

4.5. Approximation in L^p.

Theorem 4.5.1. *Let $\mathcal{A}$ be a π-system on E generating $\mathcal{E}$, with $\mu(A) < \infty$ for all $A \in \mathcal{A}$, and such that $E_n \uparrow E$ for some sequence $(E_n : n \in \mathbb{N})$ in $\mathcal{A}$. Define*

$$V_0 = \left\{ \sum_{k=1}^{n} a_k 1_{A_k} : a_k \in \mathbb{R}, A_k \in \mathcal{A}, n \in \mathbb{N} \right\}.$$

Let $p \in [1, \infty)$. Then $V_0 \subseteq L^p$. Moreover, for all $f \in L^p$ and all $\varepsilon > 0$ there exists $v \in V_0$ such that $\|v - f\|_p \leq \varepsilon$.

Proof. For all $A \in \mathcal{A}$, we have $\|1_A\|_p = \mu(A)^{1/p} < \infty$, so $1_A \in L^p$. Hence $V_0 \subseteq L^p$ because L^p is a vector space.

Write V for the set of all $f \in L^p$ for which the conclusion holds. By Minkowski's inequality, V is a vector space. Consider for now the case $E \in \mathcal{A}$ and define $\mathcal{D} = \{A \in \mathcal{E} : 1_A \in V\}$. Then $\mathcal{A} \subseteq \mathcal{D}$ so $E \in \mathcal{D}$. For $A, B \in \mathcal{D}$ with $A \subseteq B$, we have $1_{B \setminus A} = 1_B - 1_A \in V$, so $B \setminus A \in \mathcal{D}$. For $A_n \in \mathcal{D}$ with $A_n \uparrow A$, we have $\|1_A - 1_{A_n}\|_p = \mu(A \setminus A_n)^{1/p} \to 0$, so $A \in \mathcal{D}$. Hence $\mathcal{D}$ is a d-system and so $\mathcal{D} = \mathcal{E}$ by Dynkin's Lemma. Since V is a vector space it then contains all simple functions. For $f \in L^p$ with $f \geq 0$, consider the sequence of simple functions $f_n = (2^{-n}\lfloor 2^n f \rfloor) \wedge n \uparrow f$. Then, $|f|^p \geq |f - f_n|^p \to 0$ pointwise so, by dominated convergence, $\|f - f_n\|_p \to 0$. Hence $f \in V$. Hence, as a vector space, $V = L^p$.

Returning to the general case, we now know that, for all $f \in L^p$ and all $n \in \mathbb{N}$, we have $f 1_{E_n} \in V$. But $|f|^p \geq |f - f 1_{E_n}|^p \to 0$ pointwise so, by dominated convergence, $\|f - f 1_{E_n}\|_p \to 0$, and so $f \in V$. □

5. Completeness of L^p and orthogonal projection

5.1. **$\mathcal{L}^p$ as a Banach space.** Let V be a vector space. A map $v \mapsto \|v\| : V \to [0, \infty)$ is a *norm* if

(i) $\|u + v\| \leq \|u\| + \|v\|$ for all $u, v \in V$,
(ii) $\|\alpha v\| = |\alpha|\|v\|$ for all $v \in V$ and $\alpha \in \mathbb{R}$,
(iii) $\|v\| = 0$ implies $v = 0$.

We note that, for any norm, if $\|v_n - v\| \to 0$ then $\|v_n\| \to \|v\|$.

A symmetric bilinear map $(u, v) \mapsto \langle u, v \rangle : V \times V \to \mathbb{R}$ is an *inner product* if $\langle v, v \rangle \geq 0$, with equality only if $v = 0$. For any inner product, $\langle ., . \rangle$, the map $v \mapsto \sqrt{\langle v, v \rangle}$ is a norm, by the Cauchy–Schwarz inequality.

Minkowski's inequality shows that each L^p space is a vector space and that the L^p-norms satisfy condition (i) above. Condition (ii) also holds. Condition (iii) fails, because $\|f\|_p = 0$ does not imply that $f = 0$, only that $f = 0$ a.e.. For $f, g \in L^p$, write $f \sim g$ if $f = g$ almost everywhere. Then $\sim$ is an equivalence relation. Write $[f]$ for the equivalence class of f and define

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

Note that, for $f \in L^2$, we have $\|f\|_2^2 = \langle f, f\rangle$, where $\langle .,.\rangle$ is the symmetric bilinear form on L^2 given by

$$\langle f, g\rangle = \int_E fg d\mu.$$

Thus $\mathcal{L}^2$ is an inner product space. The notion of convergence in L^p defined in §4.1 is the usual notion of convergence in a normed space.

A normed vector space V is *complete* if every Cauchy sequence in V converges, that is to say, given any sequence $(v_n : n \in \mathbb{N})$ in V such that $\|v_n - v_m\| \to 0$ as $n, m \to \infty$, there exists $v \in V$ such that $\|v_n - v\| \to 0$ as $n \to \infty$. A complete normed vector space is called a *Banach space*. A complete inner product space is called a *Hilbert space*. Such spaces have many useful properties, which makes the following result important.

Theorem 5.1.1 (Completeness of L^p). *Let $p \in [1, \infty]$. Let $(f_n : n \in \mathbb{N})$ be a sequence in L^p such that*

$$\|f_n - f_m\|_p \to 0 \quad \text{as } n, m \to \infty.$$

Then there exists $f \in L^p$ such that

$$\|f_n - f\|_p \to 0 \quad \text{as } n \to \infty.$$

Proof. Some modifications of the following argument are necessary in the case $p = \infty$, which are left as an exercise. We assume from now on that $p < \infty$. Choose a subsequence (n_k) such that

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

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

$$\|\sum_{k=1}^{K} |f_{n_{k+1}} - f_{n_k}|\|_p \leq S.$$

By monotone convergence this bound holds also for $K = \infty$, so

$$\sum_{k=1}^{\infty} |f_{n_{k+1}} - f_{n_k}| < \infty \quad \text{a.e..}$$

Hence, by completeness of $\mathbb{R}$, f_{n_k} converges a.e.. We define a measurable function f by

$$f(x) = \begin{cases} \lim f_{n_k}(x) & \text{if the limit exists,} \\ 0 & \text{otherwise.} \end{cases}$$

Given $\varepsilon > 0$, we can find N so that $n \geq N$ implies

$$\mu(|f_n - f_m|^p) \leq \varepsilon, \quad \text{for all } m \geq n,$$

in particular $\mu(|f_n - f_{n_k}|^p) \leq \varepsilon$ for all sufficiently large k. Hence, by Fatou's lemma, for $n \geq N$,

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

Hence $f \in L^p$ and, since $\varepsilon > 0$ was arbitrary, $\|f_n - f\|_p \to 0$. □

Corollary 5.1.2. *We have*

(a) $\mathcal{L}^p$ *is a Banach space, for all* $1 \leq p \leq \infty$*,*

(b) $\mathcal{L}^2$ *is a Hilbert space.*

5.2. **$\mathcal{L}^2$ as a Hilbert space.** We shall apply some general Hilbert space arguments to L^2. First, we note *Pythagoras' rule*

$$\|f + g\|_2^2 = \|f\|_2^2 + 2\langle f, g\rangle + \|g\|_2^2$$

and the *parallelogram law*

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

If $\langle f, g\rangle = 0$, then we say that f and g are *orthogonal*. For any subset $V \subseteq L^2$, we define

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

A subset $V \subseteq L^2$ is *closed* if, for every sequence $(f_n : n \in \mathbb{N})$ in V, with $f_n \to f$ in L^2, we have $f = v$ a.e., for some $v \in V$.

Theorem 5.2.1 (Orthogonal projection)**.** *Let V be a closed subspace of L^2. Then each $f \in L^2$ has a decomposition $f = v + u$, with $v \in V$ and $u \in V^\perp$. Moreover, $\|f - v\|_2 \leq \|f - g\|_2$ for all $g \in V$, with equality only if $g = v$ a.e..*

The function v is called (*a version of*) the *orthogonal projection of f on V*.

Proof. Choose a sequence $g_n \in V$ such that

$$\|f - g_n\|_2 \to d(f, V) = \inf\{\|f - g\|_2 : g \in V\}.$$

By the parallelogram law,

$$\|2(f - (g_n + g_m)/2)\|_2^2 + \|g_n - g_m\|_2^2 = 2(\|f - g_n\|_2^2 + \|f - g_m\|_2^2).$$

But $\|2(f-(g_n+g_m)/2)\|_2^2 \geq 4d(f,V)^2$, so we must have $\|g_n-g_m\|_2 \to 0$ as $n,m \to \infty$. By completeness, $\|g_n - g\|_2 \to 0$, for some $g \in L^2$. By closure, $g = v$ a.e., for some $v \in V$. Hence

$$\|f - v\|_2 = \lim_n \|f - g_n\|_2 = d(f,V).$$

Now, for any $h \in V$ and $t \in \mathbb{R}$, we have

$$d(f,V)^2 \leq \|f - (v + th)\|_2^2 = d(f,V)^2 - 2t\langle f - v, h\rangle + t^2\|h\|_2^2.$$

So we must have $\langle f - v, h\rangle = 0$. Hence $u = f - v \in V^\perp$, as required. □

5.3. **Variance, covariance and conditional expectation.** In this section we look at some L^2 notions relevant to probability. For $X, Y \in L^2(\mathbb{P})$, with means $m_X = \mathbb{E}(X), m_Y = \mathbb{E}(Y)$, we define *variance*, *covariance* and *correlation* by

$$\begin{aligned}\operatorname{var}(X) &= \mathbb{E}[(X - m_X)^2],\\ \operatorname{cov}(X,Y) &= \mathbb{E}[(X - m_X)(Y - m_Y)],\\ \operatorname{corr}(X,Y) &= \operatorname{cov}(X,Y)/\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}.\end{aligned}$$

Note that $\operatorname{var}(X) = 0$ if and only if $X = m_X$ a.s.. Note also that, if X and Y are independent, then $\operatorname{cov}(X,Y) = 0$. The converse is generally false. For a random variable $X = (X_1, \dots, X_n)$ in $\mathbb{R}^n$, we define its *covariance matrix*

$$\operatorname{var}(X) = (\operatorname{cov}(X_i, X_j))_{i,j=1}^n.$$

Proposition 5.3.1. *Every covariance matrix is non-negative definite.*

Suppose now we are given a countable family of disjoint events $(G_i : i \in I)$, whose union is Ω. Set $\mathcal{G} = \sigma(G_i : i \in I)$. Let X be an integrable random variable. The *conditional expectation* of X given $\mathcal{G}$ is given by

$$Y = \sum_i \mathbb{E}(X|G_i)1_{G_i},$$

where we set $\mathbb{E}(X|G_i) = \mathbb{E}(X1_{G_i})/\mathbb{P}(G_i)$ when $\mathbb{P}(G_i) > 0$, and define $\mathbb{E}(X|G_i)$ in some arbitrary way when $\mathbb{P}(G_i) = 0$. Set $V = L^2(\mathcal{G}, \mathbb{P})$ and note that $Y \in V$. Then V is a subspace of $L^2(\mathcal{F}, \mathbb{P})$, and V is complete and therefore closed.

Proposition 5.3.2. *If $X \in L^2$, then Y is a version of the orthogonal projection of X on V.*

6. Convergence in $L^1(\mathbb{P})$

6.1. **Bounded convergence.** We begin with a basic, but easy to use, condition for convergence in $L^1(\mathbb{P})$.

Theorem 6.1.1 (Bounded convergence). *Let $(X_n : n \in \mathbb{N})$ be a sequence of random variables, with $X_n \to X$ in probability and $|X_n| \leq C$ for all n, for some constant $C < \infty$. Then $X_n \to X$ in L^1.*

Proof. By Theorem 2.5.1, X is the almost sure limit of a subsequence, so $|X| \leq C$ a.s.. For $\varepsilon > 0$, there exists N such that $n \geq N$ implies

$$\mathbb{P}(|X_n - X| > \varepsilon/2) \leq \varepsilon/(4C).$$

Then

$$\mathbb{E}|X_n - X| = \mathbb{E}(|X_n - X| 1_{|X_n - X| > \varepsilon/2}) + \mathbb{E}(|X_n - X| 1_{|X_n - X| \leq \varepsilon/2}) \leq 2C(\varepsilon/4C) + \varepsilon/2 = \varepsilon.$$

□

6.2. **Uniform integrability.**

Lemma 6.2.1. *Let X be an integrable random variable and set*

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

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

Proof. Suppose not. Then, for some $\varepsilon > 0$, there exist $A_n \in \mathcal{F}$, with $\mathbb{P}(A_n) \leq 2^{-n}$ and $\mathbb{E}(|X|1_{A_n}) \geq \varepsilon$ for all n. By the first Borel–Cantelli lemma, $\mathbb{P}(A_n \text{ i.o.}) = 0$. But then, by dominated convergence,

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

which is a contradiction. □

Let $\mathcal{X}$ be a family of random variables. For $1 \leq p \leq \infty$, we say that $\mathcal{X}$ is *bounded in L^p* if $\sup_{X \in \mathcal{X}} \|X\|_p < \infty$. Let us define

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

Obviously, $\mathcal{X}$ is bounded in L^1 if and only if $I_{\mathcal{X}}(1) < \infty$. We say that $\mathcal{X}$ is *uniformly integrable* or *UI* if $\mathcal{X}$ is bounded in L^1 and

$$I_{\mathcal{X}}(\delta) \downarrow 0, \quad \text{as } \delta \downarrow 0.$$

Note that, by Hölder's inequality, for conjugate indices $p, q \in (1, \infty)$,

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

Hence, if $\mathcal{X}$ is bounded in L^p, for some $p \in (1, \infty)$, then $\mathcal{X}$ is UI. The sequence $X_n = n1_{(0,1/n)}$ is bounded in L^1 for Lebesgue measure on $(0, 1)$, but not uniformly integrable.

Lemma 6.2.1 shows that any single integrable random variable is uniformly integrable. This extends easily to any finite collection of integrable random variables. Moreover, for any integrable random variable Y, the set

$$\mathcal{X} = \{X : X \text{ a random variable}, |X| \leq Y\}$$

is uniformly integrable, because $\mathbb{E}(|X|1_A) \leq \mathbb{E}(Y1_A)$ for all A.

The following result gives an alternative characterization of uniform integrability.

Lemma 6.2.2. *Let $\mathcal{X}$ be a family of random variables. Then $\mathcal{X}$ is UI if and only if*

$$\sup\{\mathbb{E}(|X|1_{|X|\geq K}) : X \in \mathcal{X}\} \to 0, \quad \textit{as } K \to \infty.$$

Proof. Suppose $\mathcal{X}$ is UI. Given $\varepsilon > 0$, choose $\delta > 0$ so that $I_{\mathcal{X}}(\delta) < \varepsilon$, then choose $K < \infty$ so that $I_{\mathcal{X}}(1) \leq K\delta$. Then, for $X \in \mathcal{X}$ and $A = \{|X| \geq K\}$, we have $\mathbb{P}(A) \leq \delta$ so $\mathbb{E}(|X|1_A) < \varepsilon$. Hence, as $K \to \infty$,

$$\sup\{\mathbb{E}(|X|1_{|X|\geq K}) : X \in \mathcal{X}\} \to 0.$$

On the other hand, if this condition holds, then, since

$$\mathbb{E}(|X|) \leq K + \mathbb{E}(|X|1_{|X|\geq K}),$$

we have $I_{\mathcal{X}}(1) < \infty$. Given $\varepsilon > 0$, choose $K < \infty$ so that $\mathbb{E}(|X|1_{|X|\geq K}) < \varepsilon/2$ for all $X \in \mathcal{X}$. Then choose $\delta > 0$ so that $K\delta < \varepsilon/2$. For all $X \in \mathcal{X}$ and $A \in \mathcal{F}$ with $\mathbb{P}(A) < \delta$, we have

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

Hence $\mathcal{X}$ is UI. □

Here is the definitive result on L^1-convergence of random variables.

Theorem 6.2.3. *Let X be a random variable and let $(X_n : n \in \mathbb{N})$ be a sequence of random variables. The following are equivalent:*

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

(b) *$\{X_n : n \in \mathbb{N}\}$ is UI and $X_n \to X$ in probability.*

Proof. Suppose (a) holds. By Chebyshev's inequality, for $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| > \varepsilon) \leq \varepsilon^{-1}\mathbb{E}(|X_n - X|) \to 0$$

so $X_n \to X$ in probability. Moreover, given $\varepsilon > 0$, there exists N such that $\mathbb{E}(|X_n - X|) < \varepsilon/2$ whenever $n \geq N$. Then we can find $\delta > 0$ so that $\mathbb{P}(A) \leq \delta$ implies

$$\mathbb{E}(|X|1_A) \leq \varepsilon/2, \quad \mathbb{E}(|X_n|1_A) \leq \varepsilon, \quad n = 1, \dots, N.$$

Then, for $n \geq N$ and $\mathbb{P}(A) \leq \delta$,

$$\mathbb{E}(|X_n|1_A) \leq \mathbb{E}(|X_n - X|) + \mathbb{E}(|X|1_A) \leq \varepsilon.$$

Hence $\{X_n : n \in \mathbb{N}\}$ is UI. We have shown that (a) implies (b).

Suppose, on the other hand, that (b) holds. Then there is a subsequence (n_k) such that $X_{n_k} \to X$ a.s.. So, by Fatou's lemma, $\mathbb{E}(|X|) \leq \liminf_k \mathbb{E}(|X_{n_k}|) < \infty$. Now, given $\varepsilon > 0$, there exists $K < \infty$ such that, for all n,

$$\mathbb{E}(|X_n|1_{|X_n|\geq K}) < \varepsilon/3, \quad \mathbb{E}(|X|1_{|X|\geq K}) < \varepsilon/3.$$

Consider the uniformly bounded sequence $X_n^K = (-K) \vee X_n \wedge K$ and set $X^K = (-K) \vee X \wedge K$. Then $X_n^K \to X^K$ in probability, so, by bounded convergence, there exists N such that, for all $n \geq N$,

$$\mathbb{E}|X_n^K - X^K| < \varepsilon/3.$$

But then, for all $n \geq N$,

$$\mathbb{E}|X_n - X| \leq \mathbb{E}(|X_n| 1_{|X_n| \geq K}) + \mathbb{E}|X_n^K - X^K| + \mathbb{E}(|X| 1_{|X| \geq K}) < \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, we have shown that (b) implies (a). □

7. Fourier transforms

7.1. **Definitions.** In this section (only), for $p \in [1, \infty)$, we will write $L^p = L^p(\mathbb{R}^d)$ for the set of *complex-valued* Borel measurable functions on $\mathbb{R}^d$ such that

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

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

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

Here, $\langle .,. \rangle$ denotes the usual inner product on $\mathbb{R}^d$. Note that $|\hat{f}(u)| \leq \|f\|_1$ and, by the dominated convergence theorem, $\hat{f}(u_n) \to \hat{f}(u)$ whenever $u_n \to u$. Thus $\hat{f}$ is a continuous bounded (complex-valued) function on $\mathbb{R}^d$.

For $f \in L^1(\mathbb{R}^d)$ with $\hat{f} \in L^1(\mathbb{R}^d)$, we say that the *Fourier inversion formula* holds for f if

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

for almost all $x \in \mathbb{R}^d$. For $f \in L^1 \cap L^2(\mathbb{R}^d)$, we say that the *Plancherel identity* holds for f if

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

The main results of this section establish that, for all $f \in L^1(\mathbb{R}^d)$, the inversion formula holds whenever $\hat{f} \in L^1(\mathbb{R}^d)$ and the Plancherel identity holds whenever $f \in L^2(\mathbb{R}^d)$.

The *Fourier transform* $\hat{\mu}$ of a finite Borel measure μ on $\mathbb{R}^d$ is defined by

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

Then $\hat{\mu}$ is a continuous function on $\mathbb{R}^d$ with $|\hat{\mu}(u)| \leq \mu(\mathbb{R}^d)$ for all u. The definitions are consistent in that, if μ has density f with respect to Lebesgue measure, then

$\hat{\mu} = \hat{f}$. The *characteristic function* ϕ_X of a random variable X in $\mathbb{R}^d$ is the Fourier tranform of its law μ_X. Thus

$$\phi_X(u) = \hat{\mu}_X(u) = \mathbb{E}(e^{i\langle u,X\rangle}), \quad u \in \mathbb{R}^d.$$

7.2. **Convolutions.** For $p \in [1,\infty)$ and $f \in L^p(\mathbb{R}^d)$ and for a probability measure ν on $\mathbb{R}^d$, we define the *convolution* $f * \nu \in L^p(\mathbb{R}^d)$ by

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

whenever the integral exists, setting $f * \nu(x) = 0$ otherwise. By Jensen's inequality and Fubini's theorem,

$$\begin{aligned}\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d}|f(x-y)|\nu(dy)\right)^p dx &\leq \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}|f(x-y)|^p\nu(dy)dx \\ = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}|f(x-y)|^p dx\nu(dy) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}|f(x)|^p dx\nu(dy) = \|f\|_p^p < \infty.\end{aligned}$$

Hence, the integral defining the convolution exists for almost all x, and then

$$\|f * \nu\|_p = \left(\int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d} f(x-y)\nu(dy)\right|^p dx\right)^{1/p} \leq \|f\|_p.$$

In the case where ν has a density function g, then we write $f * g$ for $f * \nu$.

For probability measures μ, ν on $\mathbb{R}^d$, we define the *convolution* $\mu * \nu$ to be the distribution of $X + Y$ for independent random variables X, Y having distributions μ, ν. Thus

$$\mu * \nu(A) = \int_{\mathbb{R}^d\times\mathbb{R}^d} 1_A(x+y)\mu(dx)\nu(dy), \quad A \in \mathcal{B}.$$

Note that, if μ has density function f, then by Fubini's theorem

$$\begin{aligned}\mu * \nu(A) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} 1_A(x+y)f(x)dx\nu(dy) \\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} 1_A(x)f(x-y)dx\nu(dy) = \int_{\mathbb{R}^d} 1_A(x)f * \nu(x)dx\end{aligned}$$

so $\mu * \nu$ has density function $f * \nu$.

It is easy to check using Fubini's theorem that $\widehat{f * \nu}(u) = \hat{f}(u)\hat{\nu}(u)$ for all $f \in L^1(\mathbb{R}^d)$ and all probability measures ν on $\mathbb{R}^d$. Similarly, we have

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

7.3. **Gaussians.** Consider for $t \in (0,\infty)$ the centred Gaussian probability density function g_t on $\mathbb{R}^d$ of variance t, given by

$$g_t(x) = \frac{1}{(2\pi t)^{d/2}} e^{-|x|^2/(2t)}.$$

The Fourier transform $\hat{g}_t$ may be identified as follows. Let Z be a standard one-dimensional normal random variable. Since Z is integrable, by Theorem 3.5.1, the characteristic function ϕ_Z is differentiable and we can differentiate under the integral sign to obtain

$$\phi_Z'(u) = \mathbb{E}(iZe^{iuZ}) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{iux} ixe^{-x^2/2} dx = -u\phi_Z(u)$$

where we integrated by parts for the last equality. Hence

$$\frac{d}{du}(e^{u^2/2}\phi_Z(u)) = 0$$

so

$$\phi_Z(u) = \phi_Z(0)e^{-u^2/2} = e^{-u^2/2}.$$

Consider now d independent standard normal random variables $Z_1, \ldots, Z_d$ and set $Z = (Z_1, \ldots, Z_d)$. Then $\sqrt{t}Z$ has density function g_t. So

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

Hence $\hat{g}_t = (2\pi)^{d/2}t^{-d/2}g_{1/t}$ and $\hat{\hat{g}}_t = (2\pi)^d g_t$. Then

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

so the Fourier inversion formula holds for g_t.

7.4. **Gaussian convolutions.** By a *Gaussian convolution* we mean any convolution $f * g_t$ of a function $f \in L^1(\mathbb{R}^d)$ with a Gaussian g_t and with $t \in (0,\infty)$. We note that $f * g_t$ is a continuous function and that

$$\|f * g_t\|_1 \le \|f\|_1, \quad \|f * g_t\|_\infty \le (2\pi)^{-d/2}t^{-d/2}\|f\|_1.$$

Also $\widehat{f * g_t}(u) = \hat{f}(u)\hat{g}_t(u)$ and we know $\hat{g}_t$ explicitly, so

$$\|\widehat{f * g_t}\|_1 \le (2\pi)^{d/2}t^{-d/2}\|f\|_1, \quad \|\widehat{f * g_t}\|_\infty \le \|f\|_1.$$

A straightforward calculation (using the parallelogram identity in $\mathbb{R}^d$) shows that $g_s * g_s = g_{2s}$ for all $s \in (0,\infty)$. Then, for any probability measure μ on $\mathbb{R}^d$ and any $t = 2s \in (0,\infty)$, we have $\mu * g_s \in L^1(\mathbb{R}^d)$ and hence $\mu * g_t = \mu * (g_s * g_s) = (\mu * g_s) * g_s$ is a Gaussian convolution.

Lemma 7.4.1. *The Fourier inversion formula holds for all Gaussian convolutions.*

Proof. Let $f \in L^1(\mathbb{R}^d)$ and let $t > 0$. We use the Fourier inversion formula for g_t and Fubini's theorem to see that

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

□

Lemma 7.4.2. *Let $f \in L^p(\mathbb{R}^d)$ with $p \in [1,\infty)$. Then $\|f * g_t - f\|_p \to 0$ as $t \to 0$.*

Proof. Given $\varepsilon > 0$, there exists a continuous function h of compact support such that $\|f-h\|_p \leq \varepsilon/3$. Then $\|f * g_t - h * g_t\|_p = \|(f-h) * g_t\|_p \leq \|f-h\|_p \leq \varepsilon/3$. Set

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

Then $e(y) \leq 2^p\|h\|_p^p$ for all y and $e(y) \to 0$ as $y \to 0$ by dominated convergence. By Jensen's inequality and then bounded convergence,

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

as $t \to 0$. Now $\|f * g_t - f\|_p \leq \|f * g_t - h * g_t\|_p + \|h * g_t - h\|_p + \|h - f\|_p$. So $\|f * g_t - f\|_p < \varepsilon$ for all sufficiently small $t > 0$, as required. □

7.5. **Uniqueness and inversion.**

Theorem 7.5.1. *Let $f \in L^1(\mathbb{R}^d)$. Define for $t > 0$ and $x \in \mathbb{R}^d$*

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

Then $\|f_t - f\|_1 \to 0$ as $t \to 0$. Moreover, the Fourier inversion formula holds whenever $f \in L^1(\mathbb{R}^d)$ and $\hat{f} \in L^1(\mathbb{R}^d)$.

Proof. Consider the Gaussian convolution $f * g_t$. Then $\widehat{f * g_t}(u) = \hat{f}(u)e^{-|u|^2t/2}$. So $f_t = f * g_t$ by Lemma 7.4.1 and so $\|f_t - f\|_1 \to 0$ as $t \to 0$ by Lemma 7.4.2.

Now, if $\hat{f} \in L^1(\mathbb{R}^d)$, then by dominated convergence with dominating function $|\hat{f}|$,

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

as $t \to 0$ for all x. On the other hand, $f_{t_n} \to f$ almost everywhere for some sequence $t_n \to 0$. Hence the inversion formula holds for f. □

7.6. **Fourier transform in $L^2(\mathbb{R}^d)$.**

Theorem 7.6.1. *The Plancherel identity holds for all $f \in L^1 \cap L^2(\mathbb{R}^d)$. Moreover there is a unique Hilbert space automorphism F on $\mathcal{L}^2$ such that*

$$F[f] = [(2\pi)^{-d/2}\hat{f}]$$

for all $f \in L^1 \cap L^2(\mathbb{R}^d)$.

Proof. Suppose to begin that $f \in L^1$ and $\hat{f} \in L^1$. Then the inversion formula holds and $f, \hat{f} \in L^\infty$. Also $(x,u) \mapsto f(x)\hat{f}(u)$ is integrable on $\mathbb{R}^d \times \mathbb{R}^d$. So, by Fubini's theorem, we obtain the Plancherel identity for f:

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

Now let $f \in L^1 \cap L^2$ and consider for $t > 0$ the Gaussian convolution $f_t = f * g_t$. We consider the limit $t \to 0$. By Lemma 7.4.2, $f_t \to f$ in L^2, so $\|f_t\|_2 \to \|f\|_2$. We have $\hat{f}_t = \hat{f}\hat{g}_t$ and $\hat{g}_t(u) = e^{-|u|^2 t/2}$, so $\|\hat{f}_t\|_2^2 \uparrow \|\hat{f}\|_2^2$ by monotone convergence. The Plancherel identity holds for f_t because $f_t, \hat{f}_t \in L^1$. On letting $t \to 0$ we obtain the identity for f.

Define $F_0 : \mathcal{L}^1 \cap \mathcal{L}^2 \to \mathcal{L}^2$ by $F_0[f] = [(2\pi)^{-d/2}\hat{f}]$. Then F_0 preserves the $\mathcal{L}^2$ norm. Since $\mathcal{L}^1 \cap \mathcal{L}^2$ is dense in $\mathcal{L}^2$, F_0 then extends uniquely to an isometry F of $\mathcal{L}^2$ into itself. Finally, by the inversion formula, F maps the set $V = \{[f] : f \in L^1 \text{ and } \hat{f} \in L^1\}$ into itself and $F^4[f] = [f]$ for all $[f] \in V$. But V contains all Gaussian convolutions and hence is dense in $\mathcal{L}^2$, so F must be onto $\mathcal{L}^2$. □

7.7. **Weak convergence and characteristic functions.** Let μ be a Borel probability measure on $\mathbb{R}^d$ and let $(\mu_n : n \in \mathbb{N})$ be a sequence of such measures. We say that μ_n *converges weakly* to μ if $\mu_n(f) \to \mu(f)$ as $n \to \infty$ for all continuous bounded functions f on $\mathbb{R}^d$. Given a random variable X in $\mathbb{R}^d$ and a sequence of such random variables $(X_n : n \in \mathbb{N})$, we say that X_n *converges weakly* to X if μ_{X_n} converges weakly to μ_X. There is no requirement that the random variables are defined on a common probability space. Note that a sequence of measures can have at most one weak limit, but if X is a weak limit of the sequence of random variables $(X_n : n \in \mathbb{N})$, then so is any other random variable with the same distribution as X. In the case $d = 1$, weak convergence is equivalent to convergence in distribution, as defined in Section 2.5.

Theorem 7.7.1. *Let X be random variable in $\mathbb{R}^d$. Then the distribution μ_X of X is uniquely determined by its characteristic function ϕ_X. Moreover, in the case where ϕ_X is integrable, μ_X has a continuous bounded density function given by*

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

Moreover, if $(X_n : n \in \mathbb{N})$ is a sequence of random variables in $\mathbb{R}^d$ such that $\phi_{X_n}(u) \to \phi_X(u)$ as $n \to \infty$ for all $u \in \mathbb{R}^d$, then X_n converges weakly to X.

Proof. Let Z be a random variable in $\mathbb{R}^d$, independent of X, and having the standard Gaussian density g_1. Then $\sqrt{t}Z$ has density g_t and $X + \sqrt{t}Z$ has density given by the Gaussian convolution $f_t = \mu_X * g_t$. We have $\hat{f}_t(u) = \phi_X(u)e^{-|u|^2 t/2}$ so, by the Fourier inversion formula,

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

By bounded convergence, for all continuous bounded functions g on $\mathbb{R}^d$, as $t \to 0$

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

Hence ϕ_X determines μ_X uniquely.

If ϕ_X is integrable, then $|f_t(x)| \le (2\pi)^{-d}\|\phi_X\|_1$ for all x and by dominated convergence with dominating function $|\phi_X|$, we have $f_t(x) \to f_X(x)$ for all x. Hence $f_X(x) \ge 0$ for all x and, for g continuous of compact support, by bounded convergence,

$$\int_{\mathbb{R}^d} g(x)\mu_X(dx) = \lim_{t\to 0} \int_{\mathbb{R}^d} g(x) f_t(x) dx = \int_{\mathbb{R}^d} g(x) f_X(x) dx$$

which implies that μ_X has density f_X, as claimed.

Suppose now that $(X_n : n \in \mathbb{N})$ is a sequence of random variables such that $\phi_{X_n}(u) \to \phi_X(u)$ for all u. We shall show that $\mathbb{E}(g(X_n)) \to \mathbb{E}(g(X))$ for all integrable functions g on $\mathbb{R}^d$ whose derivative is bounded, which implies that X_n converges weakly to X. Given $\varepsilon > 0$, we can choose $t > 0$ so that $\sqrt{t}\|\nabla g\|_\infty \mathbb{E}|Z| \le \varepsilon/3$. Then $\mathbb{E}|g(X + \sqrt{t}Z) - g(X)| \le \varepsilon/3$ and $\mathbb{E}|g(X_n + \sqrt{t}Z) - g(X)| \le \varepsilon/3$. On the other hand, by the Fourier inversion formula, and dominated convergence with dominating function $|g(x)|e^{-|u|^2 t/2}$, we have

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

as $n \to \infty$. Hence $|\mathbb{E}(g(X_n)) - \mathbb{E}(g(X))| < \varepsilon$ for all sufficiently large n, as required. □

There is a stronger version of the last assertion of Theorem 7.7.1 called Lévy's continuity theorem for characteristic functions: *if $\phi_{X_n}(u)$ converges as $n \to \infty$, with limit $\phi(u)$ say, for all $u \in \mathbb{R}$, and if ϕ is continuous in a neighbourhood of 0, then ϕ is the characteristic function of some random variable X, and $X_n \to X$ in distribution.* We will not prove this.

8. Gaussian random variables

8.1. **Gaussian random variables in $\mathbb{R}$.** A random variable X in $\mathbb{R}$ is *Gaussian* if, for some $\mu \in \mathbb{R}$ and some $\sigma^2 \in (0, \infty)$, X has density function

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

We also admit as Gaussian any random variable X with $X = \mu$ a.s., this degenerate case corresponding to taking $\sigma^2 = 0$. We write $X \sim N(\mu, \sigma^2)$.

Proposition 8.1.1. *Suppose $X \sim N(\mu, \sigma^2)$ and $a, b \in \mathbb{R}$. Then* (a) $\mathbb{E}(X) = \mu$, (b) $\operatorname{var}(X) = \sigma^2$, (c) $aX + b \sim N(a\mu + b, a^2\sigma^2)$, (d) $\phi_X(u) = e^{iu\mu - u^2\sigma^2/2}$.

8.2. **Gaussian random variables in $\mathbb{R}^n$.** A random variable X in $\mathbb{R}^n$ is *Gaussian* if $\langle u, X \rangle$ is Gaussian, for all $u \in \mathbb{R}^n$. An example of such a random variable is provided by $X = (X_1, \dots, X_n)$, where $X_1, \dots, X_n$ are independent $N(0,1)$ random variables. To see this, we note that

$$\mathbb{E}\, e^{i\langle u, X \rangle} = \mathbb{E} \prod_k e^{iu_k X_k} = e^{-|u|^2/2}$$

so $\langle u, X \rangle$ is $N(0, |u|^2)$ for all $u \in \mathbb{R}^n$.

Theorem 8.2.1. *Let X be a Gaussian random variable in $\mathbb{R}^n$. Let A be an $m \times n$ matrix and let $b \in \mathbb{R}^m$. Then*

(a) *$AX + b$ is a Gaussian random variable in $\mathbb{R}^m$,*

(b) *$X \in L^2$ and μ_X is determined by $\mu = \mathbb{E}(X)$ and $V = \operatorname{var}(X)$,*

(c) *$\phi_X(u) = e^{i\langle u, \mu \rangle - \langle u, Vu \rangle/2}$,*

(d) *if V is invertible, then X has a density function on $\mathbb{R}^n$, given by*

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

(e) *suppose $X = (X_1, X_2)$, with X_1 in $\mathbb{R}^{n_1}$ and X_2 in $\mathbb{R}^{n_2}$, then*

$$\operatorname{cov}(X_1, X_2) = 0 \quad \textit{implies} \quad X_1, X_2 \textit{ independent.}$$

Proof. For $u \in \mathbb{R}^n$, we have $\langle u, AX + b \rangle = \langle A^T u, X \rangle + \langle u, b \rangle$ so $\langle u, AX + b \rangle$ is Gaussian, by Proposition 8.1.1. This proves (a).

Each component X_k is Gaussian, so $X \in L^2$. Set $\mu = \mathbb{E}(X)$ and $V = \operatorname{var}(X)$. For $u \in \mathbb{R}^n$ we have $\mathbb{E}(\langle u, X \rangle) = \langle u, \mu \rangle$ and $\operatorname{var}(\langle u, X \rangle) = \operatorname{cov}(\langle u, X \rangle, \langle u, X \rangle) =$

$\langle u, Vu\rangle$. Since $\langle u, X\rangle$ is Gaussian, by Proposition 8.1.1, we must have $\langle u, X\rangle \sim N(\langle u,\mu\rangle, \langle u, Vu\rangle)$ and $\phi_X(u) = \mathbb{E}\, e^{i\langle u,X\rangle} = e^{i\langle u,\mu\rangle - \langle u,Vu\rangle/2}$. This is (c) and (b) follows by uniqueness of characteristic functions.

Let $Y_1,\dots,Y_n$ be independent $N(0,1)$ random variables. Then $Y = (Y_1,\dots,Y_n)$ has density

$$f_Y(y) = (2\pi)^{-n/2}\exp\{-|y|^2/2\}.$$

Set $\tilde{X} = V^{1/2}Y + \mu$, then $\tilde{X}$ is Gaussian, with $\mathbb{E}(\tilde{X}) = \mu$ and $\mathrm{var}(\tilde{X}) = V$, so $\tilde{X} \sim X$. If V is invertible, then $\tilde{X}$ and hence X has the density claimed in (d), by a linear change of variables in $\mathbb{R}^n$.

Finally, if $X = (X_1, X_2)$ with $\mathrm{cov}(X_1, X_2) = 0$, then, for $u = (u_1, u_2)$, we have

$$\langle u, Vu\rangle = \langle u_1, V_{11}u_1\rangle + \langle u_2, V_{22}u_2\rangle,$$

where $V_{11} = \mathrm{var}(X_1)$ and $V_{22} = \mathrm{var}(X_2)$. Then $\phi_X(u) = \phi_{X_1}(u_1)\phi_{X_2}(u_2)$ so X_1 and X_2 are independent. □

9. Ergodic theory

9.1. **Measure-preserving transformations.**

Let $(E, \mathcal{E}, \mu)$ be a measure space. A measurable function $\theta : E \to E$ is called a *measure-preserving transformation* if

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

A set $A \in \mathcal{E}$ is *invariant* if $\theta^{-1}(A) = A$. A measurable function f is *invariant* if $f = f \circ \theta$. The set of all invariant sets forms a σ-algebra, which we denote by $\mathcal{E}_\theta$. Then f is invariant if and only if f is $\mathcal{E}_\theta$-measurable. We say that θ is *ergodic* if $\mathcal{E}_\theta$ contains only sets of measure zero and their complements.

Here are two simple examples of measure preserving transformations.

(i) *Translation map on the torus.* Take $E = [0,1)^n$ with Lebesgue measure on its Borel σ-algebra, and consider addition modulo 1 in each coordinate. For $a \in E$ set

$$\theta_a(x_1,\dots,x_n) = (x_1 + a_1,\dots,x_n + a_n).$$

(ii) *Bakers' map.* Take $E = [0,1)$ with Lebesgue measure. Set

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

Proposition 9.1.1. *If f is integrable and θ is measure-preserving, then $f \circ \theta$ is integrable and*

$$\int_E f\,d\mu = \int_E f \circ \theta\, d\mu.$$

Proposition 9.1.2. *If θ is ergodic and f is invariant, then $f = c$ a.e., for some constant c.*

9.2. **Bernoulli shifts.** Let m be a probability measure on $\mathbb{R}$. In §2.4, we constructed a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which there exists a sequence of independent random variables $(Y_n : n \in \mathbb{N})$, all having distribution m. Consider now the infinite product space

$$E = \mathbb{R}^{\mathbb{N}} = \{x = (x_n : n \in \mathbb{N}) : x_n \in \mathbb{R} \text{ for all } n\}$$

and the σ-algebra $\mathcal{E}$ on E generated by the coordinate maps $X_n(x) = x_n$

$$\mathcal{E} = \sigma(X_n : n \in \mathbb{N}).$$

Note that $\mathcal{E}$ is also generated by the π-system

$$\mathcal{A} = \{\prod_{n\in\mathbb{N}} A_n : A_n \in \mathcal{B} \text{ for all } n,\ A_n = \mathbb{R} \text{ for sufficiently large } n\}.$$

Define $Y : \Omega \to E$ by $Y(\omega) = (Y_n(\omega) : n \in \mathbb{N})$. Then Y is measurable and the image measure $\mu = \mathbb{P} \circ Y^{-1}$ satisfies, for $A = \prod_{n\in\mathbb{N}} A_n \in \mathcal{A}$,

$$\mu(A) = \prod_{n\in\mathbb{N}} m(A_n).$$

By uniqueness of extension, μ is the unique measure on $\mathcal{E}$ having this property. Note that, under the probability measure μ, the coordinate maps $(X_n : n \in \mathbb{N})$ are themselves a sequence of independent random variables with law m. The probability space $(E, \mathcal{E}, \mu)$ is called the *canonical model* for such sequences. Define the *shift map* $\theta : E \to E$ by

$$\theta(x_1, x_2, \dots) = (x_2, x_3, \dots).$$

Theorem 9.2.1. *The shift map is an ergodic measure-preserving transformation.*

Proof. The details of showing that θ is measurable and measure-preserving are left as an exercise. To see that θ is ergodic, we recall the definition of the tail σ-algebras

$$\mathcal{T}_n = \sigma(X_m : m \geq n+1), \quad \mathcal{T} = \bigcap_n \mathcal{T}_n.$$

For $A = \prod_{n\in\mathbb{N}} A_n \in \mathcal{A}$ we have

$$\theta^{-n}(A) = \{X_{n+k} \in A_k \text{ for all } k\} \in \mathcal{T}_n.$$

Since $\mathcal{T}_n$ is a σ-algebra, it follows that $\theta^{-n}(A) \in \mathcal{T}_n$ for all $A \in \mathcal{E}$, so $\mathcal{E}_\theta \subseteq \mathcal{T}$. Hence θ is ergodic by Kolmogorov's zero-one law. □

9.3. **Birkhoff's and von Neumann's ergodic theorems.** Throughout this section, $(E, \mathcal{E}, \mu)$ will denote a measure space, on which is given a measure-preserving transformation θ. Given an measurable function f, set $S_0 = 0$ and define, for $n \geq 1$,

$$S_n = S_n(f) = f + f \circ \theta + \dots + f \circ \theta^{n-1}.$$

Lemma 9.3.1 (Maximal ergodic lemma)**.** *Let f be an integrable function on E. Set $S^* = \sup_{n\geq 0} S_n(f)$. Then*

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

Proof. Set $S^*_n = \max_{0\leq m\leq n} S_m$ and $A_n = \{S^*_n > 0\}$. Then, for $m = 1, \dots, n$,

$$S_m = f + S_{m-1} \circ \theta \leq f + S^*_n \circ \theta.$$

On A_n, we have $S^*_n = \max_{1\leq m\leq n} S_m$, so

$$S^*_n \leq f + S^*_n \circ \theta.$$

On A^c_n, we have

$$S^*_n = 0 \leq S^*_n \circ \theta.$$

So, integrating and adding, we obtain

$$\int_E S^*_n d\mu \leq \int_{A_n} f d\mu + \int_E S^*_n \circ \theta d\mu.$$

But S^*_n is integrable, so

$$\int_E S^*_n \circ \theta d\mu = \int_E S^*_n d\mu < \infty$$

which forces

$$\int_{A_n} f d\mu \geq 0.$$

As $n \to \infty$, $A_n \uparrow \{S^* > 0\}$ so, by dominated convergence, with dominating function $|f|$,

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

□

Theorem 9.3.2 (Birkhoff's almost everywhere ergodic theorem)**.** *Assume that $(E, \mathcal{E}, \mu)$ is σ-finite and that f is an integrable function on E. Then there exists an invariant function $\bar{f}$, with $\mu(|\bar{f}|) \leq \mu(|f|)$, such that $S_n(f)/n \to \bar{f}$ a.e. as $n \to \infty$.*

Proof. The functions $\liminf_n (S_n/n)$ and $\limsup_n (S_n/n)$ are invariant. Therefore, for $a < b$, so is the following set

$$D = D(a, b) = \{\liminf_n (S_n/n) < a < b < \limsup_n (S_n/n)\}.$$

We shall show that $\mu(D) = 0$. First, by invariance, we can restrict everything to D and thereby reduce to the case $D = E$. Note that either $b > 0$ or $a < 0$. We can interchange the two cases by replacing f by $-f$. Let us assume then that $b > 0$.

Let $B \in \mathcal{E}$ with $\mu(B) < \infty$, then $g = f - b1_B$ is integrable and, for each $x \in D$, for some n,

$$S_n(g)(x) \geq S_n(f)(x) - nb > 0.$$

Hence $S^*(g) > 0$ everywhere and, by the maximal ergodic lemma,

$$0 \leq \int_D (f - b1_B) d\mu = \int_D f d\mu - b\mu(B).$$

Since μ is σ-finite, there is a sequence of sets $B_n \in \mathcal{E}$, with $\mu(B_n) < \infty$ for all n and $B_n \uparrow D$. Hence,

$$b\mu(D) = \lim_{n\to\infty} b\mu(B_n) \leq \int_D f d\mu.$$

In particular, we see that $\mu(D) < \infty$. A similar argument applied to $-f$ and $-a$, this time with $B = D$, shows that

$$(-a)\mu(D) \leq \int_D (-f) d\mu.$$

Hence

$$b\mu(D) \leq \int_D f d\mu \leq a\mu(D).$$

Since $a < b$ and the integral is finite, this forces $\mu(D) = 0$. Set

$$\Delta = \{\liminf_n (S_n/n) < \limsup_n (S_n/n)\}$$

then Δ is invariant. Also, $\Delta = \bigcup_{a,b\in\mathbb{Q},a<b} D(a,b)$, so $\mu(\Delta) = 0$. On the complement of Δ, S_n/n converges in $[-\infty, \infty]$, so we can define an invariant function $\bar{f}$ by

$$\bar{f} = \begin{cases} \lim_n (S_n/n) & \text{on } \Delta^c \\ 0 & \text{on } \Delta. \end{cases}$$

Finally, $\mu(|f \circ \theta^n|) = \mu(|f|)$, so $\mu(|S_n|) \leq n\mu(|f|)$ for all n. Hence, by Fatou's lemma,

$$\mu(|\bar{f}|) = \mu(\liminf_n |S_n/n|) \leq \liminf_n \mu(|S_n/n|) \leq \mu(|f|).$$

□

Theorem 9.3.3 (von Neumann's L^p ergodic theorem)**.** *Assume that $\mu(E) < \infty$. Let $p \in [1, \infty)$. Then, for all $f \in L^p(\mu)$, $S_n(f)/n \to \bar{f}$ in L^p.*

Proof. We have

$$\|f \circ \theta^n\|_p = \left(\int_E |f|^p \circ \theta^n d\mu \right)^{1/p} = \|f\|_p.$$

So, by Minkowski's inequality,

$$\|S_n(f)/n\|_p \leq \|f\|_p.$$

Given $\varepsilon > 0$, choose $K < \infty$ so that $\|f - g\|_p < \varepsilon/3$, where $g = (-K) \vee f \wedge K$. By Birkhoff's theorem, $S_n(g)/n \to \bar{g}$ a.e.. We have $|S_n(g)/n| \leq K$ for all n so, by bounded convergence, there exists N such that, for $n \geq N$,

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

By Fatou's lemma,

$$\begin{aligned}\|\bar{f} - \bar{g}\|_p^p &= \int_E \liminf_n |S_n(f-g)/n|^p d\mu \\ &\leq \liminf_n \int_E |S_n(f-g)/n|^p d\mu \leq \|f - g\|_p^p.\end{aligned}$$

Hence, for $n \geq N$,

$$\begin{aligned}\|S_n(f)/n - \bar{f}\|_p &\leq \|S_n(f-g)/n\|_p + \|S_n(g)/n - \bar{g}\|_p + \|\bar{g} - \bar{f}\|_p \\ &< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon.\end{aligned}$$

□

10. Sums of independent random variables

10.1. **Strong law of large numbers for finite fourth moment.** The result we obtain in this section will be largely superseded in the next. We include it because its proof is much more elementary than that needed for the definitive version of the strong law which follows.

Theorem 10.1.1. *Let $(X_n : n \in \mathbb{N})$ be a sequence of independent random variables such that, for some constants $\mu \in \mathbb{R}$ and $M < \infty$,*

$$\mathbb{E}(X_n) = \mu, \quad \mathbb{E}(X_n^4) \leq M \quad \text{for all } n.$$

Set $S_n = X_1 + \cdots + X_n$. Then

$$S_n/n \to \mu \quad \text{a.s., as } n \to \infty.$$

Proof. Consider $Y_n = X_n - \mu$. Then $Y_n^4 \leq 2^4(X_n^4 + \mu^4)$, so

$$\mathbb{E}(Y_n^4) \leq 16(M + \mu^4)$$

and it suffices to show that $(Y_1 + \cdots + Y_n)/n \to 0$ a.s.. So we are reduced to the case where $\mu = 0$.

Note that X_n, X_n^2, X_n^3 are all integrable since X_n^4 is. Since $\mu = 0$, by independence,

$$\mathbb{E}(X_i X_j^3) = \mathbb{E}(X_i X_j X_k^2) = \mathbb{E}(X_i X_j X_k X_l) = 0$$

for distinct indices i, j, k, l. Hence

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

Now for $i < j$, by independence and the Cauchy–Schwarz inequality

$$\mathbb{E}(X_i^2 X_j^2) = \mathbb{E}(X_i^2)\mathbb{E}(X_j^2) \le \mathbb{E}(X_i^4)^{1/2}\mathbb{E}(X_j^4)^{1/2} \le M.$$

So we get the bound

$$\mathbb{E}(S_n^4) \le nM + 3n(n-1)M \le 3n^2 M.$$

Thus

$$\mathbb{E}\sum_n (S_n/n)^4 \le 3M \sum_n 1/n^2 < \infty$$

which implies

$$\sum_n (S_n/n)^4 < \infty \quad \text{a.s.}$$

and hence $S_n/n \to 0$ a.s.. □

10.2. **Strong law of large numbers.**

Theorem 10.2.1. *Let m be a probability measure on $\mathbb{R}$, with*

$$\int_{\mathbb{R}} |x| m(dx) < \infty, \qquad \int_{\mathbb{R}} x m(dx) = \nu.$$

Let $(E, \mathcal{E}, \mu)$ be the canonical model for a sequence of independent random variables with law m. Then

$$\mu(\{x : (x_1 + \cdots + x_n)/n \to \nu \text{ as } n \to \infty\}) = 1.$$

Proof. By Theorem 9.2.1, the shift map θ on E is measure-preserving and ergodic. The coordinate function $f = X_1$ is integrable and $S_n(f) = f + f \circ \theta + \cdots + f \circ \theta^{n-1} = X_1 + \cdots + X_n$. So $(X_1 + \cdots + X_n)/n \to \bar{f}$ a.e., for some invariant function $\bar{f}$, by Birkhoff's ergodic theorem. Moreover, this convergence holds also in L^1 by von Neumann's ergodic theorem. Since θ is ergodic, $\bar{f} = c$ a.e., for some constant c and then $c = \mu(\bar{f}) = \lim_n \mu(S_n/n) = \nu$. □

Theorem 10.2.2 (Strong law of large numbers)**.** *Let $(Y_n : n \in \mathbb{N})$ be a sequence of independent, identically distributed, integrable random variables with mean ν. Set $S_n = Y_1 + \cdots + Y_n$. Then*

$$S_n/n \to \nu \quad \textit{a.s., as } n \to \infty.$$

Proof. In the notation of Theorem 10.2.1, take m to be the law of the random variables Y_n. Then $\mu = \mathbb{P} \circ Y^{-1}$, where $Y : \Omega \to E$ is given by $Y(\omega) = (Y_n(\omega) : n \in \mathbb{N})$. Hence

$$\mathbb{P}(S_n/n \to \nu \text{ as } n \to \infty) = \mu(\{x : (x_1 + \cdots + x_n)/n \to \nu \text{ as } n \to \infty\}) = 1.$$

□

10.3. **Central limit theorem.**

Theorem 10.3.1 (Central limit theorem)**.** *Let* $(X_n : n \in \mathbb{N})$ *be a sequence of independent, identically distributed, random variables with mean* 0 *and variance* 1*. Set* $S_n = X_1 + \cdots + X_n$*. Then, for all* $x \in \mathbb{R}$*, as* $n \to \infty$*,*

$$\mathbb{P}\left(\frac{S_n}{\sqrt{n}} \leq x\right) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$

Proof. Set $\phi(u) = \mathbb{E}(e^{iuX_1})$. Since $\mathbb{E}(X_1^2) < \infty$, we can differentiate $\mathbb{E}(e^{iuX_1})$ twice under the expectation, to show that

$$\phi(0) = 1, \quad \phi'(0) = 0, \quad \phi''(0) = -1.$$

Hence, by Taylor's theorem, as $u \to 0$,

$$\phi(u) = 1 - u^2/2 + o(u^2).$$

So, for the characteristic function ϕ_n of $S_n/\sqrt{n}$,

$$\phi_n(u) = \mathbb{E}(e^{iu(X_1+\cdots+X_n)/\sqrt{n}}) = \{\mathbb{E}(e^{i(u/\sqrt{n})X_1})\}^n = (1 - u^2/2n + o(u^2/n))^n.$$

The complex logarithm satisfies, as $z \to 0$,

$$\log(1+z) = z + o(|z|)$$

so, for each $u \in \mathbb{R}$, as $n \to \infty$,

$$\log \phi_n(u) = n \log(1 - u^2/2n + o(u^2/n)) = -u^2/2 + o(1).$$

Hence $\phi_n(u) \to e^{-u^2/2}$ for all u. But $e^{-u^2/2}$ is the characteristic function of the $N(0,1)$ distribution, so $S_n/\sqrt{n} \to N(0,1)$ in distribution by Theorem 7.7.1, as required. □