Probability and Measure

Cambridge University Mathematical Tripos: Part II

17th May 2024

Contents

1 Measures

1.1 Definitions

Definition. Let E be a (nonempty) set. A collection $\mathcal{E}$ of subsets of E is called a *σ-algebra* if the following properties hold:

- $\varnothing \in \mathcal{E}$;
- $A \in \mathcal{E} \implies A^c = E \setminus A \in \mathcal{E}$;
- if $(A_n)_{n\in\mathbb{N}}$ is a countable collection of sets in $\mathcal{E}$, $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{E}$.

Example. Let $\mathcal{E} = \{\varnothing, E\}$. This is a σ-algebra. Also, $\mathcal{P}(E) = \{A \subseteq E\}$ is a σ-algebra.

Remark. Since $\bigcap_n A_n = \left(\bigcup_n A_n^c\right)^c$, any σ-algebra $\mathcal{E}$ is closed under countable intersections as well as under countable unions. Note that $B \setminus A = B \cap A^c \in \mathcal{E}$, so σ-algebras are closed under set difference.

Definition. A set E with a σ-algebra $\mathcal{E}$ is called a *measurable space*. The elements of $\mathcal{E}$ are called *measurable sets*.

Definition. A *measure* μ is a set function $\mu : \mathcal{E} \to [0, \infty]$, such that $\mu(\varnothing) = 0$, and for a sequence $(A_n)_{n\in\mathbb{N}}$ such that the A_n are disjoint, we have

$$\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n)$$

This is the *countable additivity* property of the measure.

Remark. If E is countable, then for any $A \in \mathcal{P}(E)$ and measure μ, we have

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

Hence, measures are uniquely defined by the measure of each singleton. This corresponds to the notion of a probability mass function.

Definition. For a collection $\mathcal{A}$ of subsets of E, we define the σ-algebra $\sigma(A)$ *generated by* $\mathcal{A}$ by

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

So it is the smallest σ-algebra containing $\mathcal{A}$. Equivalently,

$$\sigma(\mathcal{A}) = \bigcap_{\mathcal{E}\supseteq\mathcal{A}, \mathcal{E} \text{ a } \sigma\text{-algebra}} \mathcal{E}$$

1.2 Rings and algebras

To construct good generators, we define the following.

Definition. $\mathcal{A} \subseteq \mathcal{P}(E)$ is called a *ring* over E if $\varnothing \in \mathcal{A}$ and $A, B \in \mathcal{A}$ implies $B \setminus A \in \mathcal{A}$ and $A \cup B \in \mathcal{A}$.

Rings are easier to manage than σ-algebras because there are only finitary operators.

Definition. $\mathcal{A}$ is called an *algebra* over E if $\varnothing \in \mathcal{A}$ and $A, B \in \mathcal{A}$ implies $A^c \in \mathcal{A}$ and $A \cup B \in \mathcal{A}$.

Remark. Rings are closed under symmetric difference $A \triangle B = (B \setminus A) \cup (A \setminus B)$, and are closed under intersections $A \cap B = A \cup B \setminus A \triangle B$. Algebras are rings, because $B \setminus A = B \cap A^c = (B^c \cup A)^c$. Not all rings are algebras, because rings do not need to include the entire space.

Proposition (Disjointification of countable unions)**.** Consider $\bigcup_n A_n$ for $A_n \in \mathcal{E}$, where $\mathcal{E}$ is a σ-algebra (or a ring, if the union is finite). Then there exist $B_n \in \mathcal{E}$ that are disjoint such that $\bigcup_n A_n = \bigcup_n B_n$.

Proof. Define $\widetilde{A}_n = \bigcup_{j \le n} A_j$, then $B_{n+1} = \widetilde{A}_n \setminus \widetilde{A}_{n-1}$. □

Definition. A *set function* on a collection $\mathcal{A}$ of subsets of E, where $\varnothing \in \mathcal{A}$, is a map $\mu : \mathcal{A} \to [0, \infty]$ such that $\mu(\varnothing) = 0$. We say μ is *increasing* if $\mu(A) \le \mu(B)$ for all $A \subseteq B$ in $\mathcal{A}$. We say μ is *additive* if $\mu(A \cup B) = \mu(A) + \mu(B)$ for disjoint $A, B \in \mathcal{A}$ and $A \cup B \in \mathcal{A}$. We say μ is *countably additive* if $\mu\big(\bigcup_n A_n\big) = \sum_n \mu(A_n)$ for disjoint sequences A_n where $\bigcup_n A_n$ and each A_n lie in $\mathcal{A}$. We say μ is *countably subadditive* if $\mu\big(\bigcup_n A_n\big) \le \sum_n \mu(A_n)$ for arbitrary sequences A_n under the above conditions.

Remark. A measure satisfies all four of the above conditions. Countable additivity implies the other conditions.

Theorem (Carathéodory's theorem)**.** Let μ be a countably additive set function on a ring $\mathcal{A}$ of subsets of E. Then there exists a measure $\mu^\star$ on $\sigma(\mathcal{A})$ such that $\mu^\star|_{\mathcal{A}} = \mu$.

We will later prove that this extended measure is unique.

Proof. For $B \subseteq E$, we define the *outer measure* $\mu^\star$ as

$$\mu^\star(B) = \inf\left\{ \sum_{n \in \mathbb{N}} \mu(A_n), A_n \in \mathcal{A}, B \subseteq \bigcup_{n \in \mathbb{N}} A_n \right\}$$

If there is no sequence A_n such that $B \subseteq \bigcup_{n \in \mathbb{N}} A_n$, we declare the outer measure $\mu^\star(B)$ to be ∞. We define the class

$$\mathcal{M} = \{A \subseteq E \mid \forall B \subseteq E, \mu^\star(B) = \mu^\star(B \cap A) + \mu^\star(B \cap A^c)\}$$

This is the class of *$\mu^\star$-measurable sets*.

Step 1. $\mu^\star$ is countably sub-additive on $\mathcal{P}(E)$. It suffices to prove that for $B \subseteq E$ and $B_n \subseteq E$ such that $B \subseteq \bigcup_n B_n$ we have

$$\mu^\star(B) \leq \sum_n \mu^\star(B_n) \tag{$\dagger$}$$

We can assume without loss of generality that $\mu^\star(B_n) < \infty$ for all n, otherwise there is nothing to prove. For all $\varepsilon > 0$ there exists a collection $A_{n,m}$ such that $B_n \subseteq \bigcup_m A_{n,m}$ and

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

Now, since $\mu^\star$ is increasing, and $B \subseteq \bigcup_n B_n \subseteq \bigcup_n \bigcup_m A_{n,m}$, we have

$$\mu^\star(B) \leq \mu^\star\left(\bigcup_{n,m} A_{n,m}\right) \leq \sum_{n,m} \mu(A_{n,m}) \leq \sum_n \mu^\star(B_n) + \sum_n \frac{\varepsilon}{2^n} = \sum_n \mu^\star(B_n) + \varepsilon$$

Since ε was arbitrary in the construction, ($\dagger$) follows by construction.

Step 2. $\mu^\star$ extends μ. Let $A \in \mathcal{A}$, and we want to show $\mu^\star(A) = \mu(A)$. We can write $A = A \cup \varnothing \cup \ldots$, hence $\mu^\star(A) \leq \mu(A) + 0 + \cdots = \mu(A)$ by definition of $\mu^\star$. We need to prove the converse, that $\mu(A) \leq \mu^\star(A)$. If $\mu^\star$ is infinite, there is nothing to prove. For the finite case, suppose there is a sequence A_n where $\mu(A_n) < \infty$ and $A \subseteq \bigcup_n A_n$. Then, $A = \bigcup_n (A \cap A_n)$, which is a union of elements of the ring $\mathcal{A}$. Since μ is a countably additive set function on $\mathcal{A}$, it is countably subadditive. Hence $\mu(A) \leq \sum_n \mu(A \cap A_n) \leq \sum_n \mu(A_n)$. Since the A_n were arbitrary, we have $\mu(A) \leq \mu^\star(A)$ as required.

Step 3. $\mathcal{M} \supseteq \mathcal{A}$. Let $A \in \mathcal{A}$. We must show that for all $B \subseteq E$, $\mu^\star(B) = \mu^\star(B \cap A) + \mu^\star(B \cap A^c)$. We have $B \subseteq (B \cap A) \cup (B \cap A^c) \cup \varnothing \cup \ldots$, hence by countable subadditivity ($\dagger$), $\mu^\star(B) \leq \mu^\star(B \cap A) + \mu^\star(B \cap A^c)$. It now suffices to prove the converse, that $\mu^\star(B) \geq \mu^\star(B \cap A) + \mu^\star(B \cap A^c)$. We can assume $\mu^\star(B)$ is finite, and assume there exists $A_n \in \mathcal{A}$ such that $B \subseteq \bigcup_n A_n$ and $\mu^\star(B) + \varepsilon \geq \sum_n \mu(A_n)$. Now, $B \cap A \subseteq \bigcup_n (A_n \cap A)$, and $B \cap A^c \subseteq \bigcup_n (A_n \cap A^c)$. All of the members of these two unions are elements of $\mathcal{A}$, since $A_n \cap A^c = A_n \setminus A$. Therefore,

$$\begin{aligned}
\mu^\star(B \cap A) + \mu^\star(B \cap A^c) &\leq \sum_n \mu(A_n \cap A) + \sum_n \mu(A_n \cap A^c) \\
&\leq \sum_n [\mu(A_n \cap A) + \mu(A_n \cap A^c)] \\
&\leq \sum_n \mu(A_n) \leq \mu^\star(B) + \varepsilon
\end{aligned}$$

Since ε was arbitrary, $\mu^\star(B) = \mu^\star(B \cap A) + \mu^\star(B \cap A^c)$ as required.

Step 4. $\mathcal{M}$ is an algebra. Clearly $\varnothing$ lies in $\mathcal{M}$, and by the symmetry in the definition of $\mathcal{M}$, complements lie in $\mathcal{M}$. We need to check $\mathcal{M}$ is stable under finite intersections. Let $A_1, A_2 \in \mathcal{M}$ and let $B \subseteq E$. We have

$$\mu^\star(B) = \mu^\star(B \cap A_1) + \mu^\star(B \cap A_1^c) = \mu^\star(B \cap A_1 \cap A_2) + \mu^\star(B \cap A_1 \cap A_2^c) + \mu^\star(B \cap A_1^c)$$

We can write $A_1 \cap A_2^c = (A_1 \cap A_2^c)^c \cap A_1$, and $A_1^c = (A_1 \cap A_2)^c \cap A_1^c$. Hence

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

which is the requirement for $A_1 \cap A_2$ to lie in $\mathcal{M}$.

Step 5. $\mathcal{M}$ is a σ-algebra and $\mu^\star$ is a measure on $\mathcal{M}$. It suffices now to show that $\mathcal{M}$ has countable unions and the measure respects these countable unions. Let $A = \bigcup_n A_n$ for $A_n \in \mathcal{M}$. Without loss of generality, let the A_n be disjoint. We want to show $A \in \mathcal{M}$, and that $\mu^\star(A) = \sum_n \mu^\star(A_n)$. By (†), we have $\mu^\star(B) \leq \mu^\star(B \cap A) + \mu^\star(B \cap A^c) + 0 + \ldots$ so we need to check only the converse of this inequality. Also, $\mu^\star(A) \leq \sum_n \mu^\star(A_n)$, so we need only check the converse of this inequality as well. Similarly to before,

$$\begin{aligned}
\mu^\star(B) &= \mu^\star(B \cap A_1) + \mu^\star(B \cap A_1^c) \\
&= \mu^\star(B \cap A_1) + \mu^\star(B \cap A_1^c \cap A_2) + \mu^\star(B \cap A_1^c \cap A_2^c) \\
&= \mu^\star(B \cap A_1) + \mu^\star(B \cap A_2) + \mu^\star(B \cap A_1^c \cap A_2^c) \\
&= \mu^\star(B \cap A_1) + \mu^\star(B \cap A_2) + \mu^\star(B \cap A_1^c \cap A_2^c \cap A_3) + \mu^\star(B \cap A_1^c \cap A_2^c \cap A_3^c) \\
&= \mu^\star(B \cap A_1) + \mu^\star(B \cap A_2) + \mu^\star(B \cap A_3) + \mu^\star(B \cap A_1^c \cap A_2^c \cap A_3^c) \\
&= \cdots \\
&= \sum_{n \leq N} \mu^\star(B \cap A_n) + \mu^\star(B \cap A_1^c \cap \cdots \cap A_N^c)
\end{aligned}$$

Since $\bigcup_{n \leq N} A_n \subseteq A$, we have $\bigcap_{n \leq N} A_n^c \supseteq A^c$. $\mu^\star$ is increasing, hence, taking limits,

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

By (†),

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

as required. Hence $\mathcal{M}$ is a σ-algebra. For the other inequality, we take the above result for $B = A$.

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

So $\mu^\star$ is countably additive on $\mathcal{M}$ and is hence a measure on $\mathcal{M}$. □

1.3 Uniqueness of extension

Definition. A collection $\mathcal{A}$ of subsets of E is called a *π-system* if $\varnothing \in \mathcal{A}$ and $A, B \in \mathcal{A} \implies A \cap B \in \mathcal{A}$.

Definition. A collection $\mathcal{A}$ of subsets of E is called a *d-system* if $E \in \mathcal{A}$, and if $B_1 \subset B_2$ are elements of $\mathcal{A}$, we have $B_2 \setminus B_1 \in \mathcal{A}$, and if $A_n \in \mathcal{A}$ and A_n is an increasing sequence of sets, we have $\bigcup_n A_n \in \mathcal{A}$.

Proposition. A d-system which is also a π-system is a σ-algebra.

Proof. Refer to the first example sheet. □

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

Proof. We define

$$\mathcal{D} = \bigcap_{\mathcal{D}' \text{ is a } d\text{-system; } \mathcal{D}' \supseteq \mathcal{A}} \mathcal{D}'$$

We can show this is a d-system. It suffices to prove that $\mathcal{D}$ is a π-system, because this is then a σ-algebra. We now define

$$\mathcal{D}' = \{B \in \mathcal{D} \mid \forall A \in \mathcal{A}, B \cap A \in \mathcal{D}\}$$

We can see that $\mathcal{D}' \supseteq \mathcal{A}$, as $\mathcal{A}$ is a π-system. We now show that $\mathcal{D}'$ is a d-system. Clearly $E \cap A = A \in \mathcal{A} \subseteq \mathcal{D}'$ hence $E \in \mathcal{D}'$. Let $B_1, B_2 \in \mathcal{D}'$ such that $B_1 \subseteq B_2$. Then $(B_2 \setminus B_1) \cap A = (B_2 \cap A) \setminus (B_1 \cap A)$, and since $B_i \cap A \in \mathcal{D}$ this difference also lies in $\mathcal{D}$, so $B_2 \setminus B_1 \in \mathcal{D}'$. Now, suppose B_n is an increasing sequence converging to B, and $B_n \in \mathcal{D}'$. Then $B_n \cap A \in \mathcal{D}$, and $\mathcal{D}$ is a d-system, we have $B \cap A \in \mathcal{D}$, so $B \in \mathcal{D}'$.

Hence $\mathcal{D}'$ is a d-system that contains $\mathcal{A}$, so $\mathcal{D} \subseteq \mathcal{D}'$, and $\mathcal{D}' \subseteq \mathcal{D}$ by construction of $\mathcal{D}'$, giving $\mathcal{D} = \mathcal{D}'$. We then define

$$\mathcal{D}'' = \{B \in \mathcal{D} \mid \forall A \in \mathcal{D}, B \cap A \in \mathcal{D}\}$$

Note that $\mathcal{A} \subseteq \mathcal{D}''$, because $\mathcal{D}' = \mathcal{D} \supseteq \mathcal{A}$. Running the same argument as before, we can show that $\mathcal{D}'' = \mathcal{D}$, and so $\mathcal{D}'' = \mathcal{D}$ is a π-system. □

Theorem (Uniqueness of extension)**.** Let μ_1, μ_2 be measures on a measurable space $(E, \mathcal{E})$, such that $\mu_1(E) = \mu_2(E) < \infty$. Suppose that μ_1 and μ_2 coincide on a π-system $\mathcal{A}$, such that $\mathcal{E} \subseteq \sigma(\mathcal{A})$. Then $\mu_1 = \mu_2$ on $\sigma(\mathcal{A})$, and hence on $\mathcal{E}$.

Proof. We define

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

This collection contains $\mathcal{A}$ by assumption. By Dynkin's lemma, it suffices to prove $\mathcal{D}$ is a d-system, because then $\mathcal{D} \supseteq \sigma(\mathcal{A}) \supseteq \mathcal{E}$ giving $\mathcal{D} = \mathcal{E}$. Note that $E \in \mathcal{D}$ by assumption. By additivity and finiteness of μ_i, for $B_1 \subseteq B_2$ elements of $\mathcal{D}$, we have $\mu_1(B_2 \setminus B_1) = \mu_1(B_2) - \mu_1(B_1) = \mu_2(B_2) - \mu_2(B_1) = \mu_2(B_2 \setminus B_1)$, where the subtractions are valid by finiteness of μ, so set differences lie in $\mathcal{D}$.

Now suppose B_n is an increasing sequence converging to B for $B_n \in \mathcal{D}$. This implies that $B \setminus B_n$ is a decreasing sequence converging to $\emptyset$, and by a result from the first example sheet we have $\mu_i(B \setminus B_n) \to \mu(\emptyset) = 0$. Since μ_i are finite, $\mu_i(B_n) \to \mu_i(B)$ as $n \to \infty$. Then, $\mu_1(B) = \lim_{n\in\mathbb{N}} \mu_1(B_n) = \lim_{n\in\mathbb{N}} \mu_2(B_n) = \mu_2(B)$, so $\mathcal{D}$ is closed under increasing sequences and hence is a d system. □

Remark. The above theorem applies to finite measures (μ such that $\mu(E) < \infty$) only. However, the theorem can be extended to measures that are σ-finite, for which $E = \bigcup_{n\in\mathbb{N}} E_n$ where $\mu(E_n) < \infty$.

1.4 Borel measures

Definition. Let (E, τ) be a Hausdorff topological space. The σ-algebra generated by the open sets of E is called the *Borel σ-algebra* on E, denoted $\mathcal{B}(E) = \sigma(\tau)$. We write $\mathcal{B} = \mathcal{B}(\mathbb{R})$. Members of $\mathcal{B}(E)$ are called *Borel sets*. A measure μ on $(E, \mathcal{B}(E))$ is called a *Borel measure on E*. A *Radon measure* is a Borel measure μ on E such that $\mu(K) < \infty$ for all $K \subseteq E$ compact. Note that in a Hausdorff space, compact sets are closed and hence measurable.

1.5 Lebesgue measure

We will construct a unique Borel measure μ on $\mathbb{R}^d$ such that

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

Initially, we will perform this construction for $d = 1$, and later we will consider product measures to extend this to higher dimensions.

Theorem (Construction of the Lebesgue measure)**.** There exists a unique Borel measure μ on $\mathbb{R}$ such that

$$a < b \implies \mu((a, b]) = b - a$$

Proof. Consider the subsets of $\mathbb{R}$ of the form

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

where the intervals in question are disjoint. The set $\mathcal{A}$ of such sets forms a ring and a π-system of Borel sets. This generates the same σ-algebra as that generated by finite unions of open intervals, by the first example sheet. Open intervals with rational endpoints generate $\mathcal{B}$, so $\sigma(A) \supseteq \mathcal{B}$. We define the set function μ on $\mathcal{A}$ by $\mu(A) = \sum_{i=1}^{n}(b_i - a_i)$. μ is additive, and well-defined since if $A = \bigcup_j C_j = \bigcup_k D_k$ for distinct disjoint unions, we can write $C_j = \bigcup_k (C_j \cap D_k)$ and $D_k = \bigcup_j (D_k \cap C_j)$, giving

$$\mu(A) = \mu\left(\bigcup_j C_j\right) = \sum_j \mu(C_j) = \sum_j \mu\left(\bigcup_j (C_j \cap D_k)\right) = \sum_j \sum_k \mu(C_j \cap D_k) = \mu\left(\bigcup_k D_k\right)$$

To prove the existence of μ on $\mathcal{B}$, we apply Carathéodory's extension theorem, and therefore must check that μ is countably additive on $\mathcal{A}$. Equivalently, by a question on an example sheet, it suffices to show that for all sequences $A_n \in \mathcal{A}$ such that A_n decreases to $\varnothing$, we have $\mu(A_n) \to 0$. Suppose this is not the case, so there exist $\varepsilon > 0$ and $B_n \in \mathcal{A}$ such that B_n decreases to $\varnothing$ but $\mu(B_n) \geq 2\varepsilon$ for infinitely many n (and so without loss of generality for all n). We can approximate B_n from within by a sequence C_n. Suppose $B_n = \bigcup_{i=1}^{N_n}(a_{ni}, b_{ni}]$, then define $C_n = \bigcup_{i=1}^{N_n}(a_{ni} + \frac{2^{-n}\varepsilon}{N_n}, b_{ni}]$. Note that the C_n lie in $\mathcal{A}$, and $\mu(B_n \setminus C_n) \leq 2^{-n}\varepsilon$ Since B_n is decreasing, we have $B_N = \bigcap_{n \leq N} B_n$, and

$$B_N \setminus (C_1 \cap \cdots \cap C_N) = B_n \cap \left(\bigcup_{n \leq N} C_n^c\right) = \bigcup_{n \leq N} B_N \setminus C_n \subseteq \bigcup_{n \leq N} B_n \setminus C_n$$

Since μ is increasing,

$$\mu(B_N \setminus (C_1 \cap \cdots \cap C_N)) \leq \mu\left(\bigcup_{n\leq N} B_n \setminus C_n\right) \leq \sum_{n\leq N} \mu(B_n \setminus C_n) \leq \sum_{n\leq N} 2^{-N}\varepsilon \leq \varepsilon$$

Since in addition $\mu(B_N) \geq 2\varepsilon$, additivity implies that $\mu(C_1 \cap \cdots \cap C_N) \geq \varepsilon$. This means that $C_1 \cap \cdots \cap C_N$ cannot be empty. We can add the left endpoints of the intervals, giving $K_N = \overline{C}_1 \cap \cdots \cap \overline{C}_N$. By Analysis I, K_N is a nested sequence of nonempty closed intervals and therefore there is a point $x \in \mathbb{R}$ such that $x \in K_N$ for all N. But $K_N \subseteq \overline{C}_N \subseteq B_N$, so $x \in \bigcap_N B_n$, which is a contradiction since $\bigcap_N B_N$ is empty. Therefore, a measure μ on $\mathcal{B}$ exists.

Now we prove uniqueness. Suppose μ, λ are measures such that the measure of an interval $(a, b]$ is $b - a$. We define new measures $\mu_n(A) = \mu(A \cap (n, n+1])$ and $\lambda_n(A) = \lambda(A \cap (n, n+1])$. These new measures are finite with total mass 1. Hence, we can use the uniqueness of extension theorem to show $\mu_n = \lambda_n$ on $\mathcal{B}$. We find

$$\mu(A) = \mu\left(\bigcup_n A \cap (n, n+1]\right) = \sum_{n\in\mathbb{Z}} \mu(A \cap (n, n+1]) = \sum_{n\in\mathbb{Z}} \mu_n(A) = \sum_{n\in\mathbb{Z}} \lambda_n(A) = \cdots = \lambda(A)$$

□

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

Remark. A singleton $\{x\}$ can be written as $\bigcap_n \left(x - \frac{1}{n}, x\right]$, hence $\mu(x) = \lim_n \frac{1}{n} = 0$. Hence singletons are null sets. In particular, $\mu((a, b)) = \mu((a, b]) = \mu([a, b)) = \mu([a, b])$. Any countable set $Q = \bigcup_q \{q\}$ is a null set. Not all null sets are countable; the Cantor set is an example.

The Lebesgue measure is *translation-invariant*. Let $x \in \mathbb{R}$, then the set $B + x = \{b + x \mid b \in B\}$ lies in $\mathcal{B}$ if and only if $B \in \mathcal{B}$, and in this case, it satisfies $\mu(B + x) = \mu(B)$. We can define the translated Lebesgue measure $\mu_x(B) = \mu(B + x)$ for all $B \in \mathcal{B}$, but since the Lebesgue measure is unique, $\mu_x = \mu$.

The class of outer measurable sets $\mathcal{M}$ used in Carathéodory's extension theorem is here called the class of Lebesgue measurable sets. This class can be shown to be

$$\mathcal{M} = \{M = A \cup N, A \in \mathcal{B}, N \subseteq B, B \in \mathcal{B}, \mu(B) = 0\} \supsetneq \mathcal{B}$$

1.6 Existence of non-measurable sets

Assuming the axiom of choice, there exists a non-measurable set of reals. Consider $E = (0, 1]$ with addition defined modulo one. By the same argument as before, the Lebesgue measure is translation-invariant modulo one. Consider the subgroup $Q = E \cap \mathbb{Q}$ of $(E, +)$. We define $x \sim y$ if $x - y \in Q$. Then, this gives equivalence classes $[x] = \{y \in E : x \sim y\}$ for all $x \in E$. Assuming the axiom of choice, we can select a representative of $[x]$ for each $x \in E$, and denote by S the set of such representatives. We can partition E into the union of its cosets, so $E = \bigcup_{q\in Q}(S + q)$ is a disjoint union.

Suppose S is a Borel set. Then $S + q$ is also a Borel set. We can therefore write

$$1 = \mu(E) = \mu\left(\bigcup_{q\in Q}(S + q)\right) = \sum_{q\in Q} \mu(S + q) = \sum_{q\in Q} \mu(S)$$

But no value for $\mu(S) \in [0, \infty]$ can be assigned to make this equation hold. Therefore S is not a Borel set.

One can further show that μ cannot be extended to all subsets $\mathcal{P}(E)$.

Theorem (Banach, Kuratowski)**.** Assuming the continuum hypothesis, there exists no measure μ on the set $\mathcal{P}((0,1])$ such that $\mu((0,1]) = 1$ and $\mu(\{x\}) = 0$ for $x \in (0,1]$.

1.7 Probability spaces

Definition. If a measure space $(E, \mathcal{E}, \mu)$ has $\mu(E) = 1$, we call it a *probability space*, and instead write $(\Omega, \mathcal{F}, \mathbb{P})$. We call Ω the outcome space or sample space, $\mathcal{F}$ the set of events, and $\mathbb{P}$ the probability measure.

The axioms of probability theory (Kolmogorov, 1933), are

(i) $\mathbb{P}(\Omega) = 1$;

(ii) $0 \leq \mathbb{P}(E) \leq 1$ for all $E \in \mathcal{F}$;

(iii) if A_n are a disjoint sequence of events in $\mathcal{F}$, then $\mathbb{P}\left(\bigcup_n A_n\right) = \sum_n \mathbb{P}(A_n)$.

This is exactly what is required by our definition: $\mathbb{P}$ is a measure on a σ-algebra.

Definition. Events $A_i, i \in I$ are *independent* if for all finite $J \subseteq I$, we have

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

σ-algebras $\mathcal{A}_i, i \in I$ are independent if for any $A_j \in \mathcal{A}_j$ where $J \subseteq I$ is finite, the A_j are independent.

Kolmogorov showed that these definitions are sufficient to derive the law of large numbers.

Proposition. Let $\mathcal{A}_1, \mathcal{A}_2$ be π-systems of sets in $\mathcal{F}$. Suppose $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\,\mathbb{P}(A_2)$ for all $A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2$. Then the σ-algebras $\sigma(\mathcal{A}_1), \sigma(\mathcal{A}_2)$ are independent.

This follows by uniqueness.

1.8 Borel–Cantelli lemmas

Definition. Let $A_n \in \mathcal{F}$ be a sequence of events. Then the *limit superior* of A_n is

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

The *limit inferior* of A_n is

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

Lemma (First Borel–Cantelli lemma)**.** Let $A_n \in \mathcal{F}$ be a sequence of events such that $\sum_n \mathbb{P}(A_n) < \infty$. Then $\mathbb{P}(A_n \text{ infinitely often}) = 0$.

Proof. For all n, we have

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

□

This proof did not require that $\mathbb{P}$ be a probability measure, just that it is a measure. Therefore, we can use this for arbitrary measures.

Lemma (Second Borel–Cantelli lemma)**.** Let $A_n \in \mathcal{F}$ be a sequence of independent events, and $\sum_n \mathbb{P}(A_n) = \infty$. Then $\mathbb{P}(A_n \text{ infinitely often}) = 1$.

Proof. By independence, for all $N \geq n \in \mathbb{N}$ and using $1 - a \leq e^{-a}$, we find

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

As $N \to \infty$, this approaches zero. Since $\bigcap_{m=n}^N A_m^c$ decreases to $\bigcap_{m=n}^\infty A_m^c$, by countable additivity we must have $\mathbb{P}\left(\bigcap_{m=n}^\infty A_m^c\right) = 0$. But then

$$\mathbb{P}(A_n \text{ infinitely often}) = \mathbb{P}\left(\bigcap_n \bigcup_{m \geq n} A_m\right) = 1 - \mathbb{P}\left(\bigcup_n \bigcap_{m \geq n} A_m^c\right) \geq 1 - \sum_n \mathbb{P}\left(\bigcap_{m \geq n} A_m^c\right) = 1$$

Hence this probability is equal to one. □

2 Measurable functions

2.1 Definition

Definition. Let $(E, \mathcal{E}), (G, \mathcal{G})$ be measurable spaces. A function $f : E \to G$ is called $\mathcal{E}$-$\mathcal{G}$-*measurable* if when $A \in \mathcal{G}$, we have $f^{-1}(A) \in \mathcal{E}$.

Informally, the preimage of a measurable set under a measurable function is measurable.

If $G = \mathbb{R}$ and $\mathcal{G} = \mathcal{B}$, we can just say that $f : (E, \mathcal{E}) \to G$ is measurable. Moreover, if E is a topological space and $\mathcal{E} = \mathcal{B}(E)$, we say f is Borel measurable.

Note that preimages f^{-1} commute with many set operations such as intersection, union, and complement. This implies that $\{f^{-1}(A) \mid A \in \mathcal{G}\}$ is a σ-algebra over E, and likewise, $\{A \mid f^{-1}(A) \in \mathcal{E}\}$ is a σ-algebra over G. Hence, if $\mathcal{A}$ is a collection of subsets of G generating $\mathcal{G}$ such that $f^{-1}(A) \in \mathcal{E}$ for all $A \in \mathcal{A}$, the class $\{A \mid f^{-1} \in \mathcal{E}\}$ is a σ-algebra that contains $\mathcal{A}$ and hence that contains $\mathcal{G}$. In particular, it suffices to check $f^{-1}(A) \in \mathcal{E}$ for all elements of a generator to conclude that f is measurable.

If $f : (E, \mathcal{E}) \to \mathbb{R}$, the collection $\mathcal{A} = \{(-\infty, y] : y \in \mathbb{R}\}$ generates $\mathcal{B}$ as is shown on the first example sheet. Hence f is measurable whenever $f^{-1}((-\infty, y]) = \{x \in E \mid f(x) \leq y\} \in \mathcal{E}$ for all $y \in \mathbb{R}$.

If E is a topological space and $\mathcal{E} = \mathcal{B}(E)$, then if $f : E \to \mathbb{R}$ is continuous, the preimages of open sets B are open, and hence Borel sets. The open sets in $\mathbb{R}$ generate the σ-algebra $\mathcal{B}$. Hence, continuous functions to the real line are measurable.

Example. Consider the indicator function $\mathbb{1}_A$ of a set A. This is measurable if and only if A is measurable, or equivalently $A \in \mathcal{E}$.

Example. The composition of measurable functions is measurable. Measurability is preserved under addition, multiplication, countable infimum, countable supremum, countable limit inferior, countable limit superior, and some other operations. Note that given a collection of maps $\{f_i : E \to (G, \mathcal{G}) \mid i \in I\}$, we can make them all measurable by taking $\mathcal{E}$ to be a large enough σ-algebra, for instance $\sigma\big(\{f_i^{-1}(A) \mid A \in \mathcal{G}, i \in I\}\big)$.

2.2 Monotone class theorem

Theorem. Let $\mathcal{A}$ be a π-system that generates the σ-algebra $\mathcal{E}$ over E. Let $\mathcal{V}$ be a vector space of bounded maps from E to $\mathbb{R}$ such that

(i) $\mathbb{1}_E \in \mathcal{V}$;
(ii) $\mathbb{1}_A \in \mathcal{V}$ for all $A \in \mathcal{A}$;
(iii) if f is bounded and $f_n \in \mathcal{V}$ are nonnegative functions that form an increasing sequence that converge pointwise to f on E, then $f \in \mathcal{V}$.

Then $\mathcal{V}$ contains all bounded measurable functions $f : E \to \mathbb{R}$.

Proof. Define $\mathcal{D} = \{A \in \mathcal{E} \mid \mathbb{1}_A \in \mathcal{V}\}$. This contains $\mathcal{A}$ by hypothesis, as well as E itself. We show $\mathcal{D}$ is a d-system, so that by Dynkin's lemma, $\mathcal{E} = \mathcal{D}$. Indeed, $E \in \mathcal{D}$ by assumption. For $A \subseteq B$ and $A, B \in \mathcal{D}$, we have $\mathbb{1}_{B \setminus A} = \mathbb{1}_B - \mathbb{1}_A$ which is well-defined and lies in $\mathcal{V}$ as $\mathcal{V}$ is a vector space. Finally, if $A_n \in \mathcal{D}$ increases to A, we have $\mathbb{1}_{A_n}$ increases pointwise to $\mathbb{1}_A$, which lies in $\mathcal{V}$ by the second hypothesis. Hence $\mathcal{E} = \mathcal{D}$.

Let $f : E \to \mathbb{R}$ be a bounded measurable function, which we will assume at first is nonnegative. We define

$$f_n = \sum_{j=0}^{n2^n} \frac{j}{2^n} \mathbb{1}_{A_{n_j}}; \quad A_{n_j} = \begin{cases} \left\{x \in E \mid \frac{j}{2^n} < f(x) \leq \frac{j+1}{2^n}\right\} = f^{-1}\left(\left(\frac{j}{2^n}, \frac{j+1}{2^n}\right]\right) \in \mathcal{E} & \text{if } j \neq n2^n \\ \{x \in E \mid n < f(x)\} = f^{-1}((n, \infty)) & \text{if } j = n2^n \end{cases}$$

Since f is bounded, for $n > \|f\|_\infty$, we have $f_n \leq f \leq f_n + 2^{-n}$. Hence $|f_n - f| \leq 2^{-n} \to 0$. By assumption, the limit of the f_n, which is exactly f, also lies in $\mathcal{V}$.

Now, by separating any bounded measurable function f into its positive and negative parts, we find that these two parts lie in $\mathcal{V}$, and so $f \in \mathcal{V}$ as required. □

2.3 Image measures

Definition. Let $f : (E, \mathcal{E}) \to (G, \mathcal{G})$ be a measurable function, and μ is a measure on $(E, \mathcal{E})$. Then the *image measure* $\nu = \mu \circ f^{-1}$ is obtained from assigning $\nu(A) = \mu(f^{-1}(A))$ for all $A \in \mathcal{G}$.

Lemma. Let $g : \mathbb{R} \to \mathbb{R}$ be an increasing, right-continuous function, and set $g(\pm\infty) = \lim_{z\to\pm\infty} g(z)$. On $I = (g(-\infty), g(+\infty))$ we define the *generalised inverse*

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

for $x \in I$. Then f is increasing, left-continuous, and $f(x) \le y$ if and only if $x \le g(y)$ for all $x \in I, y \in \mathbb{R}$.

Remark. f and g form a Galois connection.

Proof. Let $J_x = \{y \in \mathbb{R} \mid x \le g(y)\}$. Since $x > g(-\infty)$, J_x is nonempty and bounded below. Hence $f(x)$ is a well-defined real number. If $y \in J_x$, then $y' \ge y$ implies $y' \in J_x$ since g is increasing. Further, if y_n converges from the right to y, and all $y_n \in J_x$, we can take limits in $x \le g(y_n)$ to find $x \le \lim_n g(y_n) = g(y)$ since g is right-continuous. Hence $y \in J_x$. So $J_x = [f(x), \infty)$. Hence $f(x) \le y \iff x \le g(y)$ as required.

If $x \le x'$, we have $J_x \supseteq J_{x'}$ by definition, so $f(x) \le f(x')$. Similarly, if x_n converges from the left to x, we have $J_x = \bigcap_n J_{x_n}$, so $f(x_n) \to f(x)$ as $x_n \to x$. □

Theorem. Let $g : \mathbb{R} \to \mathbb{R}$ be an increasing, right-continuous function, and set $g(\pm\infty) = \lim_{z\to\pm\infty} g(z)$. Then there exists a unique Radon measure μ_g on $\mathbb{R}$ such that $\mu_g((a, b]) = g(b) - g(a)$ for all $a < b$. Further, all Radon measures can be obtained in this way.

Proof. We will show that the generalised inverse f as defined above is measurable. For all $z \in \mathbb{R}$, we find $f^{-1}((-\infty, z]) = \{x : f(x) \le z\} = \{x : x \le g(z)\} = [-g(\infty), g(z)]$ which is measurable. Since $\mathcal{B}$ is generated by these such sets, f is $\mathcal{B}(I)$-$\mathcal{B}$ measurable as required. Therefore, the image measure $\mu_g = \mu \circ f^{-1}$, where μ is the Lebesgue measure on I, exists. Then for any $-\infty < a < b < \infty$, we have

$$\begin{aligned}
\mu_g((a, b]) &= \mu(f^{-1}((a, b])) \\
&= \mu(\{x : a < f(x) \le f(b)\}) \\
&= \mu(\{x : g(a) < x \le g(b)\}) \\
&= g(b) - g(a)
\end{aligned}$$

This uniquely determines μ_g by the same argument as shown previously for the Lebesgue measure μ on $\mathbb{R}$. Since g maps into $\mathbb{R}$, $g(b) - g(a) \in \mathbb{R}$ so any compact set has finite measure as it is a subset of a closed bounded interval.

Conversely, let ν be a Radon measure on $\mathbb{R}$. Define

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

This is an increasing function in y, since ν is a measure. Since we are using right-closed intervals, g is right-continuous. Finally, $\nu((a,b]) = g(b) - g(a)$ which can be seen by case analysis and additivity of the measure ν. By uniqueness as before, this characterises ν in its entirety. □

Remark. Such image measures μ_g are called *Lebesgue–Stieltjes measures*, where g is the *Stieltjes distribution*.

Example. The *Dirac measure at* x, written δ_x, is defined by

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

This has Stieltjes distribution $g(x) = \mathbb{1}_{[x,\infty)}$.

2.4 Random variables

Definition. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $(E, \mathcal{E})$ be a measurable space. An *E-valued random variable* X is an $\mathcal{F}$-$\mathcal{E}$ measurable map $X : \Omega \to E$. When $E = \mathbb{R}$ or $\mathbb{R}^d$ with the Borel σ-algebra, we simply call X a random variable or random vector.
The *law* or *distribution* μ_X of a random variable X is given by the image measure $\mu_X = \mathbb{P} \circ X^{-1}$.
When E is the real line, this measure has a distribution function

$$F_X(z) = \mu_X((-\infty, z]) = \mathbb{P}(X^{-1}(-\infty, z]) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \leq z\}) = \mathbb{P}(X \leq z)$$

This uniquely determines μ_X by the π-system argument given above.

Using the properties of measures, we can show that any distribution function satisfies:

(i) F_X is increasing;

(ii) F_X is right-continuous;

(iii) $\lim_{z \to -\infty} F_X(z) = \mu_X(\varnothing) = 0$;

(iv) $\lim_{z \to \infty} F_X(z) = \mu_X(\mathbb{R}) = \mathbb{P}(\Omega) = 1$.

Given any function F_X satisfying each property, we can obtain a random variable X on $(\Omega, \mathcal{F}, \mathbb{P}) = ((0,1), \mathcal{B}((0,1)), \mu)$ by $X(\omega) = \inf\{x \mid \omega \leq f(x)\}$, and then F_X is the distribution function of X.

Definition. Consider a countable collection $(X_i : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E}))$ for $i \in I$. This collection of random variables is called *independent* if the σ-algebras $\sigma\left(\left\{X_i^{-1}(A) : A \in \mathcal{E}\right\}\right)$ are independent.

For $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B})$ we show on an example sheet that 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 finite subsets $\{X_1, \dots, X_n\}$ of the X_i.

2.5 Constructing independent random variables

We now construct an infinite sequence of independent random variables with prescribed distribution functions on $(\Omega, \mathcal{F}, \mathbb{P}) = ((0,1), \mathcal{B}, \mu)$ with μ the Lebesgue measure on $(0,1)$. We start with Bernoulli random variables.

Any $\omega \in (0,1)$ has a binary representation given by $(\omega_i) \in \{0,1\}^{\mathbb{N}}$, which is unique if we exclude infinitely long tails of zeroes from the binary representation. We can then define the *nth Rademacher function* $R_n(\omega) = \omega_n$ which extracts the nth bit from the binary expansion. Since each R_n can be given as the sum of 2^{n-1} indicator functions on measurable sets, they are measurable functions and are hence random variables. Their distribution is given by $\mathbb{P}(R_n = 1) = \frac{1}{2} = \mathbb{P}(R_n = 0)$, so we have constructed Bernoulli random variables with parameter $\frac{1}{2}$. We show they are independent. For a finite set $(x_i)_{i=1}^n$,

$$\mathbb{P}(R_1 = x_1, \dots, R_n = x_n) = 2^{-n} = \mathbb{P}(R_1 = x_1) \dots \mathbb{P}(R_n = x_n)$$

Therefore, the R_n are all independent, so countable sequences of independent random variables indeed exist. Now, take a bijection $m : \mathbb{N}^2 \to \mathbb{N}$ and define $Y_{nk} = R_{m(n,k)}$, which are independent random variables. We can now define $Y_n = \sum_k 2^{-k} Y_{nk}$. This converges for all $\omega \in \Omega$ since $|Y_{nk}| \leq 1$, and these are still independent. We show the Y_n are uniform random variables, by showing the distribution coincides with the uniform distribution on the π-system of intervals $\left(\frac{i}{2^m}, \frac{i+1}{2^{m+1}}\right]$ for $i = 0, \dots, 2^m - 1$, which generates $\mathcal{B}$.

$$\mathbb{P}\left(Y_n \in \left(\frac{i}{2^m}, \frac{i+1}{2^m}\right]\right) = \mathbb{P}\left(\frac{i}{2^m} < \sum_k 2^{-k} Y_{nk} \leq \frac{i+1}{2^n}\right) = 2^{-m} = \mu\left(\frac{i}{2^m}, \frac{i+1}{2^{m+1}}\right]$$

Hence $\mu_{Y_n} = \mu|_{(0,1)}$ by the uniqueness theorem, and so we have constructed an infinite sequence of independent uniform random variables Y_n. If F_n are probability distribution functions, taking the generalised inverse, we see that the $F_n^{-1}(Y_n)$ are independent and have distribution function F_n.

2.6 Convergence of measurable functions

Definition. We say that a property defining a set $A \in \mathcal{E}$ holds *μ-almost everywhere* if $\mu(A^c) = 0$ for a measure μ on $\mathcal{E}$. If $\mu = \mathbb{P}$, we say a property holds *$\mathbb{P}$-almost surely* or *with probability one*, if $\mathbb{P}(A) = 1$.

Definition. If f_n and f are measurable functions on $(E, \mathcal{E}, \mu)$, we say *f_n converges to f μ-almost everywhere* if $\mu(\{x \in E \mid f_n(x) \nrightarrow f(x)\}) = 0$. We say *$f_n$ converges to f in μ-measure* if for all $\varepsilon > 0$, $\mu(\{x \in E \mid |f_n(x) - f(x)| > \varepsilon\}) \to 0$ as $n \to \infty$. For random variables, we say $X_n \to X$ *$\mathbb{P}$-almost surely* or *in $\mathbb{P}$-probability*, written $X_n \to^p X$, respectively. If X_n, X take values in $\mathbb{R}$, we say $X_n \to X$ *in distribution*, written $X_n \to^d X$ if $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ at all points x for which the limit $x \mapsto \mathbb{P}(X \leq x)$ is continuous.

We can show that $X_n \to^p X \implies X^n \to^d X$.

Theorem. Let $f_n : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be measurable functions. Then,
(i) if $\mu(E) < \infty$, then $f_n \to 0$ almost everywhere implies that $f_n \to 0$ in measure;
(ii) if $f_n \to 0$ in measure, $f_{n_k} \to 0$ almost everywhere on some subsequence.

Proof. Let $\varepsilon > 0$.

$$\mu(|f_n| < \varepsilon) \geq \mu\left(\bigcap_{m \geq n} \{|f_m| \leq \varepsilon\}\right)$$

The sequence $\left(\bigcap_{m \geq n} \{|f_m| \leq \varepsilon\}\right)_n$ increases to $\bigcup_n \bigcap_{m \geq n} \{|f_m| \leq \varepsilon\}$. So by countable additivity,

$$\begin{aligned}\mu\left(\bigcap_{m \geq n} \{|f_m| \leq \varepsilon\}\right) &\to \mu\left(\bigcup_n \bigcap_{m \geq n} \{|f_m| \leq \varepsilon\}\right) \\ &= \mu(|f_n| \leq \varepsilon \text{ eventually}) \\ &\geq \mu(|f_n| \to 0) = \mu(E)\end{aligned}$$

Hence,

$$\liminf_n \mu(|f_n| \leq \varepsilon) \geq \mu(E) \implies \limsup_n \mu(|f_n| > \varepsilon) \leq 0 \implies \mu(|f_n| > \varepsilon) \to 0$$

For the second part, by hypothesis, we have

$$\mu\left(|f_n| > \frac{1}{k}\right) < \varepsilon$$

for sufficiently large n. So choosing $\varepsilon = \frac{1}{k^2}$, we see that along some subsequence n_k we have

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

Hence,

$$\sum_k \mu\left(|f_{n_k}| > \frac{1}{n}\right) < \infty$$

So by the first Borel–Cantelli lemma, we have

$$\mu\left(|f_{n_k}| > \frac{1}{k} \text{ infinitely often}\right) = 0$$

so $f_{n_k} \to 0$ almost everywhere. □

Remark. Condition (i) is false if $\mu(E)$ is infinite: consider $f_n = \mathbb{1}_{(n,\infty)}$ on $(\mathbb{R}, \mathcal{B}, \mu)$, since $f_n \to 0$ almost everywhere but $\mu(f_n) = \infty$. Condition (ii) is false if we do not restrict to subsequences: consider independent events A_n such that $\mathbb{P}(A_n) = \frac{1}{n}$, then $\mathbb{1}_{A_n} \to 0$ in probability since $\mathbb{P}\left(\mathbb{1}_{A_n} > \varepsilon\right) = \mathbb{P}(A_n) = \frac{1}{n} \to 0$, but $\sum_n \mathbb{P}(A_n) = \infty$, and by the second Borel–Cantelli lemma, $\mathbb{P}\left(\mathbb{1}_{A_n} > \varepsilon \text{ infinitely often}\right) = 1$, so $\mathbb{1}_{A_n} \nrightarrow 0$ almost surely.

Example. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent exponential random variables distributed by $\mathbb{P}(X_1 \le x) = 1 - e^{-x}$ for $x \ge 0$. Define $A_n = \{X_n \ge \alpha \log n\}$ where $\alpha > 0$, so $\mathbb{P}(A_n) = n^{-\alpha}$, and in particular, $\sum_n \mathbb{P}(A_n) < \infty$ if and only if $\alpha > 1$. By the Borel–Cantelli lemmas, we have for all $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{X_n}{\log n} \ge 1 \text{ infinitely often}\right) = 1; \quad \mathbb{P}\left(\frac{X_n}{\log n} \ge 1 + \varepsilon \text{ infinitely often}\right) = 0$$

In other words, $\limsup_n \frac{X_n}{\log n} = 1$ almost surely.

2.7 Kolmogorov's zero-one law

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. We can define $\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, \dots)$. Let $\mathcal{T} = \bigcap_{n\in\mathbb{N}} \mathcal{T}_n$ be the *tail σ-algebra*, which contains all events in $\mathcal{F}$ that depend only on the limiting behaviour of (X_n).

Theorem. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables. Let $A \in \mathcal{T}$ be an event in the tail σ-algebra. Then $\mathbb{P}(A) = 1$ or $\mathbb{P}(A) = 0$. If $Y : (\Omega, \mathcal{T}) \to (\mathbb{R}, \mathcal{B})$ is measurable, it is constant almost surely.

Proof. Define $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$ to be the σ-algebra generated by the first n elements of (X_n). This is also generated by the π-system of sets $A = (X_1 \le x_1, \dots, X_n \le x_n)$ for any $x_i \in \mathbb{R}$. Note that the π-system of sets $B = (X_{n+1} \le x_{n+1}, \dots, X_{n+k} \le x_{n+k})$, for arbitrary $k \in \mathbb{N}$ and $x_i \in \mathbb{R}$, generates $\mathcal{T}_n$. By independence of the sequence, we see that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for all such sets A, B, and so the σ-algebras $\mathcal{T}_n, \mathcal{F}_n$ generated by these π-systems are independent.

Let $\mathcal{F}_\infty = \sigma(X_1, X_2, \dots)$. Then, $\bigcup_n \mathcal{F}_n$ is a π-system that generates $\mathcal{F}_\infty$. If $A \in \bigcup_n \mathcal{F}_n$, we have $A \in \mathcal{F}_n$ for some n, so there exists $\overline{n}$ such that $B \in \mathcal{T}_{\overline{n}}$ is independent of A. In particular, $B \in \bigcap_n \mathcal{T}_n = \mathcal{T}$. By uniqueness, $\mathcal{F}_\infty$ is independent of $\mathcal{T}$.

Since $\mathcal{T} \subseteq \mathcal{F}_\infty$, if $A \in \mathcal{T}$, A is independent from A. So $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A)$, so $\mathbb{P}(A)^2 - \mathbb{P}(A) = 0$ as required.

Finally, if $Y : (\Omega, \mathcal{T}) \to (\mathbb{R}, \mathcal{B})$, the preimages of $\{Y \le y\}$ lie in $\mathcal{T}$, which give probability one or zero. Let $c = \inf\{y \mid F_Y(y) = 1\}$, so $Y = c$ almost surely. □

3 Integration

3.1 Notation

Let $f : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be an 'integrable' function, a notion we will define. We will then define the integral with respect to μ, either written $\mu(f)$ or $\int_E f \, \mathrm{d}\mu = \int_E f(x) \, \mathrm{d}\mu(x)$. If X is a random variable, we will define its expectation $\mathbb{E}[X] = \int_\Omega X \, \mathrm{d}\mathbb{P} = \int_\Omega X(\omega) \, \mathrm{d}\mathbb{P}(\omega)$.

3.2 Definition

We say that a function $f : (E, \mathcal{E}, \mu) \to \mathbb{R}$ is *simple* if it is of the form

$$f = \sum_{k=1}^{m} a_k \mathbb{1}_{A_k}; \quad a_k \ge 0; \quad A_k \in \mathcal{E}; \quad m \in \mathbb{N}$$

Definition. The *μ-integral* of a simple function f defined as above is

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

which is independent of the choice of representation of the simple function.

Remark. We have $\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g)$ for all nonnegative coefficients α, β and simple functions f, g. If $g \leq f$, $\mu(g) \leq \mu(f)$, so μ is increasing. If $f = 0$ almost everywhere, $\mu(f) = 0$.

For a general non-negative function $f : (E, \mathcal{E}, \mu) \to \mathbb{R}$, we define its μ-integral to be

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

which agrees with the above definition for simple functions. This operator takes values in the extended non-negative real line $[0, \infty]$. Now, for $f : (E, \mathcal{E}, \mu) \to \mathbb{R}$ measurable but not necessarily non-negative, we define $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$, so that $f = f^+ - f^-$ and $|f| = f^+ + f^-$.

Definition. A measurable function $f : (E, \mathcal{E}, \mu) \to \mathbb{R}$ is *μ-integrable* if $\mu(|f|) < \infty$. In this case, we define its integral to be

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

which is a well-defined real number.

3.3 Monotone convergence theorem

Theorem. Let $f_n, f : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be measurable and non-negative such that f_n increases pointwise to f, so $f_n(x) \leq f_{n+1}(x) \leq f(x)$ and $f_n(x) \to f(x)$ as $n \to \infty$. Then, $\mu(f_n) \to \mu(f)$ as $n \to \infty$.

Remark. This is a theorem that allows us to interchange a pair of limits, $\mu(f) = \mu(\lim_n f_n) = \lim_n \mu(f_n)$. Also, $g_n \geq 0$, $\mu\left(\sum_n g_n\right) = \sum_n \mu(g_n)$.

If we consider the approximating sequence $\widetilde{f_n} = 2^{-n}\lfloor 2^n f \rfloor$, as defined in the monotone class theorem, then this is a non-negative sequence converging to f. So in particular, $\mu(f)$ is equal to the limit of the integrals of these simple functions.

It suffices to require convergence of $f_n \to f$ almost everywhere, the general argument does not need to change. The non-negativity constraint is not required if the first term in the sequence f_0 is integrable, by subtracting f_0 from every term.

Proof. Recall that $\mu(f) = \sup\{\mu(g) \mid g \leq f, g \text{ simple}\}$. Since f_n is an increasing sequence of non-negative functions, $\mu(f_n)$ is an increasing sequence of nonnegative functions. So it converges to its (*extended* non-negative real) supremum $M = \sup_n \mu(f_n)$. Since $f_n \leq f$, $\mu(f_n) \leq \mu(f)$, so taking suprema, $M \leq \mu(f)$. If M is finite, $\sup_n \mu(f_n) = \lim_n \mu(f_n) \leq \mu(f)$. If M is infinite, we are already done.

Now, we need to show $\mu(f) \leq M$, or equivalently, $\mu(g) \leq M$ for all simple g such that $g \leq f$, so that taking suprema, $\mu(f) = \sup_g \mu(g) \leq M$. We define $g_n = \min(\overline{f}_n, g)$, where $\overline{f}_n$ is the nth approximation of f_n by simple functions from the monotone class theorem. Now, since f_n increases to f, $\overline{f}_n$ increases to f. In particular, $g_n = \min(\overline{f}_n, g)$ increases to $\min(f, g) = g$. Since $\overline{f}_n \leq f_n$ by definition, we have $g_n \leq f_n$ for all n.

Now let g be an arbitrary simple function of the form $g = \sum_{k=1}^m a_k \mathbb{1}_{A_k}$ where $a_k \geq 0$ and the $A_k \in \mathcal{E}$ are disjoint. For $\varepsilon > 0$, we define sets $A_k(n) = \{x \in A_k \mid g_n(x) \geq (1-\varepsilon)a_k\}$. Since $g = a_k$ on A_k, and since g_n increases to g, we must have $A_k(n)$ increases to A_k for all k. Since μ is a measure, $\mu(A_k(n))$ increases to $\mu(A_k)$ by countable additivity.

We have $g_n \mathbb{1}_{A_k} \geq g_n \mathbb{1}_{A_k(n)} \geq (1-\varepsilon) a_k \mathbb{1}_{A_k(n)}$ on E. Moreover, $g_n = \sum_{k=1}^m g_n \mathbb{1}_{A_k}$ since the A_k are disjoint and support g_n. Hence, $g_n \geq \sum_{k=1}^m (1-\varepsilon) a_k \mathbb{1}_{A_k(n)}$, and in particular, $\mu(g_n) \geq (1-\varepsilon) \sum_{k=1}^m a_k \mu(A_k(n))$. The right hand side increases to $(1-\varepsilon)\sum_{k=1}^m a_k \mu(A_k) = (1-\varepsilon)\mu(g)$. Hence

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

Since ε was arbitrary, this completes the proof. □

3.4 Linearity of integral

Theorem. Let $f, g : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be nonnegative measurable functions. Then $\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g)$ for all $\alpha, \beta \geq 0$. Further, if $g \leq f$, then $\mu(g) \leq \mu(f)$. Finally, $f = 0$ almost everywhere if and only if $\mu(f) = 0$.

Proof. If $\widetilde{f}_n, \widetilde{g}_n$ are the approximations of f and g by simple funtions from the monotone class theorem, $\alpha\widetilde{f}_n$ increases to αf and $\beta\widetilde{g}_n$ increases to βg, so $\alpha\widetilde{f}_n + \beta\widetilde{g}_n$ increases to $\alpha f + \beta g$. Integrating both sides and using the monotone convergence theorem, the result follows, since linearity of simple functions is simple to prove.

The second part $g \leq f \implies \mu(g) \leq \mu(f)$ has already been proven. Now, if $f = 0$ almost everywhere, its approximation $0 \leq \widetilde{f}_n$ increases to f almost everywhere, so must be exactly zero for all n. So $\mu(\widetilde{f}_n) = 0$ so $\mu(f) = 0$. Conversely, if $\mu(f) = 0$, then $0 \leq \mu(\widetilde{f}_n) \to 0$ gives $\mu(\widetilde{f}_n) = 0$ so $\widetilde{f}_n = 0$ almost everywhere. Since $0 = \widetilde{f}_n$ increases almost everywhere to f, f is zero almost everywhere. □

Remark. Functions such as $\mathbb{1}_{\mathbb{Q}}$ are integrable and have integral zero. They are 'identified' with the zero element in the theory of integration.

Theorem. Let $f, g : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be integrable functions. Then $\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g)$ for all $\alpha, \beta \in \mathbb{R}$; if $g \leq f$, then $\mu(g) \leq \mu(f)$; and if $f = 0$ almost everywhere, we have $\mu(f) = 0$.

Proof. Clearly, if f is integrable, so is αf, and $\mu(-f) = -\mu(f)$, by definition of the integral for a general function. We can explicitly check that for $\alpha \geq 0$, we have $\mu(\alpha f) = \mu((\alpha f)^+) - \mu((\alpha f)^-) = \alpha\mu(f^+) - \alpha\mu(f^-) = \alpha\mu(f)$. Define $h = f + g$. Then $h^+ + f^- + g^- = h^- + f^+ + g^+$, so by the previous theorem, $\mu(h^+) + \mu(f^-) + \mu(g^-) = \mu(h^-) + \mu(f^+) + \mu(g^+)$ and the result holds.

Finally, if $0 \leq f - g$, we have $0 \leq \mu(0) \leq \mu(f - g) = \mu(f) - \mu(g)$ so the result follows. If $f = 0$ almost everywhere, $f^+ = 0$ and $f^- = 0$ almost everywhere, so $\mu(f) = 0$. □

3.5 Fatou's lemma

Lemma. Let $f_n : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be nonnegative measurable functions. Then $\mu(\liminf_n f_n) \leq \liminf_n \mu(f_n)$.

Remark. Recall that $\liminf_n x_n = \sup_n \inf_{m \geq n} x_m$ and $\limsup_n x_n = \inf_n \sup_{m \geq n} x_m$. In particular, $\limsup_n x_n = \liminf_n x_n$ implies that $\lim_n x_n$ exists and is equal to $\limsup_n x_n$ and $\liminf_n x_n$. Hence, if the f_n converge to some measurable function f, we must have $\mu(f) \leq \liminf_n \mu(f_n)$.

Proof. We have $\inf_{m \geq n} f_m \leq f_k$ for all $k \geq n$, so by taking integrals, $\mu\big(\inf_{m \geq n} f_m\big) \leq \mu(f_k)$. Thus,

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

Note that $\inf_{m \geq n} f_m$ increases to $\sup_n \inf_{m \geq n} f_m = \liminf_n f_n$. By the monotone convergence theorem,

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

as required. □

3.6 Dominated convergence theorem

Theorem. Let $f_n, f : (E, \mathcal{E}, \mu)$ be measurable functions such that $|f_n| \leq g$ almost everywhere on E, and the dominating function g is μ-integrable, so $\mu(g) < \infty$. Suppose $f_n \to f$ pointwise (or almost everywhere) on E. Then f_n and f are also integrable, and $\mu(f_n) \to \mu(f)$ as $n \to \infty$.

Proof. Clearly $\mu(|f_n|) \leq \mu(g) < \infty$, so the f_n are integrable. Taking limits in $|f_n| \leq g$, we have $|f| \leq g$, so f is also integrable by the same argument. Now, $g \pm f_n$ is a nonnegative function, and converges pointwise to $g \pm f$. Since limits are equal to the limit inferior when they exist, by Fatou's lemma, we have

$$\mu(g) + \mu(f) = \mu(g + f) = \mu\left(\liminf_n (g + f_n)\right) \leq \liminf_n \mu(g + f_n) = \mu(g) + \liminf_n \mu(f_n)$$

Hence $\mu(f) \leq \liminf_n \mu(f_n)$. Likewise, $\mu(g) - \mu(f) \leq \mu(g) - \liminf_n \mu(f_n)$, so $\mu(f) \geq \limsup_n \mu(f_n)$, so

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

But since $\liminf_n \mu(f_n) \leq \limsup_n \mu(f_n)$, the result follows. □

Example. Let $E = [0, 1]$ with the Lebesgue measure. Let $f_n \to f$ pointwise and the f_n are uniformly bounded, so $\sup_n \|f_n\|_\infty \leq g$ for some $g \in \mathbb{R}$. Then since $\mu(g) = g < \infty$, the dominated convergence theorem implies that f_n, f are integrable and $\mu(f_n) \to \mu(f)$ as $n \to \infty$. In particular, no notion of uniform convergence of the f_n is required.

Remark. The proof of the fundamental theorem of calculus requires only the fact that

$$\int_x^{x+h} \mathrm{d}t = h$$

This is a fact which is obviously true of the Riemann integral and also of the Lebesgue integral. Therefore, for any continuous function $f : [0,1] \to \mathbb{R}$, we have

$$\underbrace{\int_0^x f(t)\,\mathrm{d}t}_{\text{Riemann integral}} = F(x) = \underbrace{\int_0^x f(t)\,\mathrm{d}\mu(t)}_{\text{Lebesgue integral}}$$

So these integrals coincide for continuous functions. We can show that all Riemann integrable functions are $\mu^\star$-measurable, where $\mu^\star$ is the outer measure of the Lebesgue measure, as defined in the proof of Carathéodory's theorem. However, there exist certain Riemann integrable functions that are not Borel measurable. We can find that a bounded $\mu^\star$-measurable function is Riemann integrable if and only if

$$\mu(\{x \in [0,1] \mid f \text{ is discontinuous at } x\}) = 0$$

The standard techniques of Riemann integration, such as substitution and integration by parts, extend to all bounded measurable functions by the monotone class theorem.

Theorem. Let $U \subseteq \mathbb{R}$ be an open set and $(E, \mathcal{E}, \mu)$ be a measure space. Let $f : U \times E \to \mathbb{R}$ be a map such that $x \mapsto f(t,x)$ is measurable, and $t \mapsto f(t,x)$ is differentiable where $\left|\frac{\partial f}{\partial t}\right| < g(x)$ for all $t \in U$, and g is μ-integrable. Then

$$F(t) = \int_E f(t,x)\,\mathrm{d}\mu(x) \implies F'(t) = \int_E \frac{\partial f}{\partial t}(t,x)\,\mathrm{d}\mu(x)$$

Proof. By the mean value theorem,

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

Note that g is μ-integrable. By differentiability of f, we have $g_h \to 0$ as $h \to 0$, so applying the dominated convergence theorem, $\mu(g_h) \to \mu(0) = 0$. By linearity of the integral,

$$\mu(g_h) = \frac{\int_E f(t+h,x) - f(t,x)\,\mathrm{d}\mu(x)}{h} - \int_E \frac{\partial f}{\partial t}(t,x)\,\mathrm{d}\mu(x)$$

Hence, $\frac{F(t+h)-F(t)}{h} - F'(t) \to 0$. □

Example. For a measurable function $f : (E, \mathcal{E}, \mu) \to (G, \mathcal{G})$, if $g : G \to \mathbb{R}$ is a nonnegative function, we show on an example sheet that

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

On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a G-valued random variable X, we then compute

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

Example (measures with densities)**.** If $f : (E, \mathcal{E}, \mu) \to \mathbb{R}$ is a nonnegative measurable function, we can define $\nu_f(A) = \mu(f\mathbb{1}_A)$ for any measurable set A, which is again a measure on $(E, \mathcal{E})$ by the monotone convergence theorem. In particular, if $g : (E, \mathcal{E}) \to \mathbb{R}$ is measurable, $\nu_f(g) = \int_E g(x)f(x)\,\mathrm{d}\mu(x) = \int_E g\,\mathrm{d}\nu(f)$. We call f the *density* of ν_f with respect to μ. If its integral is one, it is called a *probability density function*.

4 Product measures

4.1 Integration in product spaces

Let $(E_1, \mathcal{E}_1, \mu_1), (E_2, \mathcal{E}_2, \mu_2)$ be finite measure spaces. On $E = E_1 \times E_2$, we can consider the π-system of 'rectangles' $\mathcal{A} = \{A_1 \times A_2 \mid A_1 \in \mathcal{E}_1, A_2 \in \mathcal{E}_2\}$. Then we define the σ-algebra $\mathcal{E}_1 \otimes \mathcal{E}_2 = \sigma(\mathcal{A})$ on the product space. If the E_i are topological spaces with a countable base, then $\mathcal{B}(E_1 \times E_2) = \mathcal{B}(E_1) \otimes \mathcal{B}(E_2)$.

Lemma. Let $E = E_1 \times E_2, \mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2$. Let $f : (E, \mathcal{E}) \to \mathbb{R}$ be measurable. Then for all $x_1 \in E_1$, the map $(x_2 \mapsto f(x_1, x_2)) : (E_2, \mathcal{E}_2) \to \mathbb{R}$ is $\mathcal{E}_2$-measurable.

Proof. Let

$$\mathcal{V} = \{f : (E, \mathcal{E}) \to \mathbb{R} \mid f \text{ bounded, measurable, conclusion of the lemma holds}\}$$

This is a $\mathbb{R}$-vector space, and it contains $\mathbb{1}_E, \mathbb{1}_A$ for all $A \in \mathcal{A}$, since $\mathbb{1}_A = \mathbb{1}_{A_1(x_1)}\mathbb{1}_{A_2(x_2)}$. Now, let $0 \le f_n$ increase to f, $f_n \in \mathcal{V}$. Then $(x_2 \mapsto f(x_1, x_2)) = \lim_n (x_2 \mapsto f_n(x_1, x_2))$, so it is $\mathcal{E}_2$-measurable as a limit of a sequence of measurable functions. Then by the monotone class theorem, $\mathcal{V}$ contains all bounded measurable functions. This extends to all measurable functions by truncating the absolute value of f to $n \in \mathbb{N}$, then the sequence of such bounded truncations converges pointwise to f. □

Lemma. Let $E = E_1 \times E_2, \mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2$. Let $f : (E, \mathcal{E}) \to \mathbb{R}$ be measurable such that
(i) f is bounded; or
(ii) f is nonnegative.
Then the map $x_1 \mapsto \int_{E_2} f(x_1, x_2)\, \mathrm{d}\mu_2(x_2)$ is μ_1-measurable and is bounded or nonnegative respectively.

Remark. In case (ii), the map on x_1 may evaluate to infinity, but the set of values

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

lies in $\mathcal{E}_1$.

Proof. Let

$$\mathcal{V} = \{f : (E, \mathcal{E}) \to \mathbb{R} \mid f \text{ bounded, measurable, conclusion of the lemma holds}\}$$

This is a vector space by linearity of the integral. $\mathbb{1}_E \in \mathcal{V}$, since $\mathbb{1}_{E_1}\mu_2(E_2)$ is nonnegative and bounded. $\mathbb{1}_A \in \mathcal{V}$ for all $A \in \mathcal{A}$, because $\mathbb{1}_{A_1}(x_1)\mu_2(A_2)$ is $\mathcal{E}_1$-measurable, nonnegative, and bounded since it is at most $\mu_2(E_2) < \infty$. Now let f_n be a sequence of nonnegative functions that increase to f, where $f_n \in \mathcal{V}$. Then by the monotone convergence theorem,

$$\int_{E_2} \lim_{n\to\infty} f_n(x_1, x_2)\, \mathrm{d}\mu_2(x_2) = \lim_{n\to\infty} \int_{E_2} f_n(x_1, x_2)\, \mathrm{d}\mu_2(x_2)$$

is an increasing limit of $\mathcal{E}_1$-measurable functions, so is $\mathcal{E}_1$-measurable. It is bounded by $\mu_2(E_2)\|f\|_\infty$, or nonnegative as required. So $f \in \mathcal{V}$. By the monotone class theorem, the result for bounded functions holds. In case (ii), we can take a bounded approximation in $\mathcal{V}$ of an arbitrary measurable function f to conclude the proof. □

Theorem (product measure). Let $(E_1, \mathcal{E}_1, \mu_1), (E_2, \mathcal{E}_2, \mu_2)$ be finite measure spaces. There exists a unique measure $\mu = \mu_1 \otimes \mu_2$ on $(E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2)$ such that $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ for all $A_1 \in \mathcal{E}_1, A_2 \in \mathcal{E}_2$.

Proof. $\mathcal{A}$ generates $\mathcal{E} \otimes \mathcal{E}_2$, so by the uniqueness theorem, there can only be one such measure. We define

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

We have

$$\begin{aligned} \mu(A_1 \times A_2) &= \int_{E_1} \left(\int_{E_2} \mathbb{1}_{A_1}(x_1) \mathbb{1}_{A_2}(x_2) \, \mathrm{d}\mu_2(x_2) \right) \mathrm{d}\mu_1(x_1) \\ &= \int_{E_1} \mathbb{1}_{A_1}(x_1) \mu_2(A_2) \, \mathrm{d}\mu_1(x_1) \\ &= \mu_1(A_1)\mu_2(A_2) \end{aligned}$$

Clearly $\mu(\varnothing) = 0$, so it suffices to show countable additivity. Let A_n be disjoint sets in $\mathcal{E}_1 \otimes \mathcal{E}_2$. Then

$$\mathbb{1}_{(\bigcup_n A_n)} = \sum_n \mathbb{1}_{A_n} = \lim_{n \to \infty} \sum_{i=1}^{n} \mathbb{1}_{A_n}$$

Then by the monotone convergence theorem and the previous lemmas,

$$\begin{aligned} \mu\left(\bigcup_n A_n \right) &= \int_{E_1} \left(\int_{E_2} \lim_{n \to \infty} \sum_{i=1}^{n} \mathbb{1}_{A_i} \, \mathrm{d}\mu_2(x_2) \right) \mathrm{d}\mu_1(x_1) \\ &= \int_{E_1} \left(\lim_{n \to \infty} \int_{E_2} \sum_{i=1}^{n} \mathbb{1}_{A_i} \, \mathrm{d}\mu_2(x_2) \right) \mathrm{d}\mu_1(x_1) \\ &= \lim_{n \to \infty} \int_{E_1} \left(\int_{E_2} \sum_{i=1}^{n} \mathbb{1}_{A_i} \, \mathrm{d}\mu_2(x_2) \right) \mathrm{d}\mu_1(x_1) \\ &= \lim_{n \to \infty} \sum_{i=1}^{n} \int_{E_1} \left(\int_{E_2} \mathbb{1}_{A_i} \, \mathrm{d}\mu_2(x_2) \right) \mathrm{d}\mu_1(x_1) \\ &= \lim_{n \to \infty} \sum_{i=1}^{n} \mu(A_i) \\ &= \sum_{n=1}^{\infty} \mu(A_n) \end{aligned}$$

□

4.2 Fubini's theorem

Theorem. Let $(E, \mathcal{E}, \mu) = (E_1 \times E_2, \mathcal{E}_1 \otimes \mathcal{E}_2, \mu_1 \otimes \mu_2)$ be a finite measure space. Let $f : E \to \mathbb{R}$ be a nonnegative measurable function. Then

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

Now, let $f : E \to \mathbb{R}$ be a μ-integrable function (on the product measure). Let

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

Define f_1 by $f_1(x_1) = \int_{E_2} f(x_1, x_2) \, \mathrm{d}\mu_2(x_2)$ on A_1 and zero elsewhere. Then $\mu_1(A_1^c) = 0$ and $\mu(f) = \mu_1(f_1) = \mu_1(f_1 \mathbb{1}_{A_1})$, and defining A_2 symmetrically, $\mu(f) = \mu_2(f_2) = \mu_2(f_2 \mathbb{1}_{A_2})$.

Remark. If f is bounded, $A_1 = E_1$. Note, for $f(x_1, x_2) = \frac{x_1^2 - x_2^2}{(x_1^2 + x_2^2)^2}$ on $(0,1)^2$, we have $\mu_1(f_1) \neq \mu_2(f_2)$, but f is not Lebesgue integrable on $(0,1)^2$.

Proof. By the construction of the product measure $\mu(A)$ for rectangles $A = A_1 \times A_2$ in the π-system $\mathcal{A}$ generating $\mathcal{E}$, the identities in the first part of the theorem clearly hold for $f = \mathbb{1}_A$. By uniqueness, this extends to $\mathbb{1}_A$ for all $A \in \mathcal{E}$. Then, by linearity of the integral, this extends to simple functions. By the monotone convergence theorem, the first part of the theorem follows.

Now let f be μ-integrable. Let $h(x_1) = \int_{E_2} |f(x_1, x_2)| \, \mathrm{d}\mu_2(x_2)$. Then by the first part, $\mu_1(|h|) \leq \mu(|f|) < \infty$. So f_1 is μ_1-integrable. We have $\mu_1(A_1^c) = 0$, otherwise, we could compute a lower bound $\mu_1(|h|) \geq \mu_1(|h| \mathbb{1}_{A_1^c}) = \infty$, but it must be finite. Note that $f_1^{\pm} = \int_{E_2} f^{\pm}(x_1, x_2) \, \mathrm{d}\mu_2(x_2)$, and $\mu(f_1) = \mu_1(f_1^+) - \mu_1(f_1^-)$. Hence, by the first part, $\mu(f) = \mu(f^+) - \mu(f^-) = \mu_1(f_1^+) - \mu_1(f_1^-) = \mu_1(f_1)$ as required. □

Remark. The proofs above extend to σ-finite measures μ.

Let $(E_i, \mathcal{E}_i, \mu_i)$ be measure spaces with σ-finite measures. Note that $(\mathcal{E}_1 \otimes \mathcal{E}_2) \otimes \mathcal{E}_3 = \mathcal{E}_1 \otimes (\mathcal{E}_2 \otimes \mathcal{E}_3)$, by a π-system argument using Dynkin's lemma. So we can iterate the construction of the product measure to obtain a measure $\mu_1 \otimes \ldots \mu_n$, which is a unique measure on $\left(\prod_{i=1}^n E_i \bigotimes_{i=1}^n \mathcal{E}_i \right)$ with the property that the measure of a hypercube $\mu(A_1 \times A_n)$ is the product of the measures of its sides $\mu_i(A_i)$.

In particular, we have constructed the Lebesgue measure $\mu^n = \bigotimes_{i=1}^n \mu$ on $\mathbb{R}^n$. Applying Fubini's theorem, for functions f that are either nonnegative and measurable or μ^n-integrable, we have

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

4.3 Product probability spaces and independence

Proposition. Let $(\Omega, \mathcal{F}, \mathbb{P})$, and $(E, \mathcal{E}) = \left(\prod_{i=1}^n E_i, \bigotimes_{i=1}^n \mathcal{E}_i\right)$. Let $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ be a measurable function, and define $X(\omega) = (X_1(\omega), X_2(\omega), \dots, X_n(\omega))$. Then the following are equivalent.

(i) $X_1, \dots, X_n$ are independent random variables;
(ii) $\mu_X = \bigotimes_{i=1}^n \mu_{X_i}$;
(iii) for all bounded and measurable $f_i : E_i \to \mathbb{R}$, $\mathbb{E}\left[\prod_{i=1}^n f_i(X_i)\right] = \prod_{i=1}^n \mathbb{E}[f_i(X_i)]$.

Proof. (i) implies (ii). Consider the π-system $\mathcal{A}$ of rectangles $A = \prod_{i=1}^n A_i$ for $A_i \in \mathcal{E}_i$. Since μ_X is an image measure, Then

$$\mu_X(A_1 \times \cdots \times A_n) = \mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mathbb{P}(X_1) \dots \mathbb{P}(A_n) = \prod_{i=1}^n \mu_{X_i}(A_i)$$

So by uniqueness, the result follows.

(ii) implies (iii). By Fubini's theorem,

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^n f_i(X_i)\right] &= \mu_X\left(\prod_{i=1}^n f_i(x_i)\right) \\
&= \int_E f(x)\, \mathrm{d}\mu(x) \\
&= \int \cdots \int_{E_i} \left(\prod_{i=1}^n f_i(x_i)\right) \mathrm{d}\mu_{X_1}(x_1) \dots \mathrm{d}\mu_{X_2}(x_2) \\
&= \prod_{i=1}^n \int_{E_i} f_i(x_i)\, \mathrm{d}\mu_{X_i}(x_i) \\
&= \prod_{i=1}^n \mathbb{E}[f_i(X_i)]
\end{aligned}$$

(iii) implies (i). Let $f_i = \mathbb{1}_{A_i}$ for any $A_i \in \mathcal{E}_i$. These are bounded and measurable functions. Then

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

So the σ-algebras generated by the X_i are independent as required. □

5 Function spaces and norms

5.1 Norms

Definition. A *norm* on a real vector space is a map $\|\cdot\|_V : V \to \mathbb{R}$ such that
(i) $\|\lambda v\| = |\lambda| \cdot \|v\|$;
(ii) $\|u + v\| \leq \|u\| + \|v\|$;
(iii) $\|v\| = 0$ if and only if $v = 0$.

Definition. Let $(E, \mathcal{E}, \mu)$ be a measure space. We define $L^p(E, \mathcal{E}, \mu) = L^p(\mu) = L^p$ for the space of measurable functions $f : E \to \mathbb{R}$ such that $\|f\|_p$ is finite, where

$$\|f\|_p = \begin{cases} \left(\int_E |f(x)|^p \, \mathrm{d}\mu(x)\right)^{\frac{1}{p}} & 1 \leq p < \infty \\ \operatorname{ess\,sup} |f| = \inf\{\lambda > 0 \mid |f| \leq \lambda \text{ almost everywhere}\} & p = \infty \end{cases}$$

We must check that $\|\cdot\|_p$ as defined is a norm. Clearly (i) holds for all $1 \leq p \leq \infty$. Property (ii) holds for $p = 1$ and $p = \infty$, and we will prove later that this holds for other values of p. The last property does not hold: $f = 0$ implies $\|f\|_p = 0$, but $\|f\|_p = 0$ implies only that $|f|^p = 0$ almost everywhere, so f is zero almost everywhere on E. Therefore, to rigorously define the norm, we must construct the quotient space $\mathcal{L}^p$ of functions that coincide almost everywhere. We write $[f]$ for the equivalence class of functions that are equal almost everywhere. The functional $\|\cdot\|_p$ is then a norm on $\mathcal{L}^p$.

Proposition (Chebyshev's inequality, Markov's inequality)**.** Let $f : E \to \mathbb{R}$ be nonnegative and measurable. Then for all $\lambda > 0$,

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

Proof. Integrate the inequality $\lambda \mathbb{1}_{\{f \geq \lambda\}} \leq f$, which holds on E. □

Definition. Let $I \subseteq R$ be an interval. Then we say a map $c : I \to \mathbb{R}$ is *convex* if for all $x, y \in I$ and $t \in [0, 1]$, we have $c(tx + (1-t)y) \leq tc(x) + (1-t)c(y)$. Equivalently, for all $x < t < y$ and $x, y \in I$, we have $\frac{c(t)-c(x)}{t-x} \leq \frac{c(y)-c(t)}{y-t}$.

Since a convex function is continuous on the interior of the interval, it is Borel measurable.

Lemma. Let $I \subseteq R$ be an interval, and let $m \in I^\circ$. If c is convex on I, there exist a, b such that $c(x) \geq ax + b$, and $c(m) = am + b$.

Proof. Define $a = \sup\left\{\frac{c(m)-c(x)}{m-x} \mid x < m, x \in I\right\}$. This exists in $\mathbb{R}$ by the second definition of convexity. Let $y \in I$, and $y > m$. Then $a \leq \frac{c(y)-c(m)}{y-m}$, so $c(y) \geq ay - am + c(m) = ay + b$ where we

define $b = c(m) - am$. Similarly, for $y < m$, by definition of the supremum, $\frac{c(m)-c(y)}{m-y} \leq a$, we have $c(y) \geq ay + b$. □

Theorem (Jensen's inequality)**.** Let X be a random variable taking values in an interval $I \subseteq \mathbb{R}$, such that $\mathbb{E}[|X|] < \infty$. Let $c : I \to \mathbb{R}$ be a convex function. Then $c(\mathbb{E}[X]) \leq \mathbb{E}[c(X)]$.

Note that the integral $\mathbb{E}[c(X)]$ is defined as $\mathbb{E}[c^+(X)] - \mathbb{E}[c^-(X)]$, and this is well-defined and takes values in $(-\infty, \infty]$.

Proof. Define $m = \mathbb{E}[X] = \int_I z \, \mathrm{d}\mu_X(z)$. If $m \notin I^\circ$, X must equal m almost surely, and then the result follows. Now let $m \in I^\circ$. Applying the previous lemma, we find a, b such that $c^-(X) \leq |a| \cdot |X| + |b|$. Hence, $\mathbb{E}[c^-(X)] \leq |a|\mathbb{E}[|X|] + |b| < \infty$, and $\mathbb{E}[c(X)] = \mathbb{E}[c^+(X)] - \mathbb{E}[c^-(X)]$ is well-defined in $(-\infty, \infty]$. Integrating the inequality from the lemma, and using linearity of the integral,

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

□

Remark. If $1 \leq p < q < \infty$, $c(x) = |x|^{\frac{q}{p}}$ is a convex function. If X is a bounded random variable (so lies in $L^\infty(\mathbb{P})$), we then have

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

Using the monotone convergence theorem, this extends to all $X \in L^q(\mathbb{P})$ when $\|X\|_q$ is finite. In particular, $L^q(\mathbb{P}) \subseteq L^p(\mathbb{P})$ for all $1 \leq p \leq q \leq \infty$.

Theorem (Hölder's inequality)**.** Let f, g be measurable functions on $(E, \mathcal{E}, \mu)$. If p, q are *conjugate*, so $\frac{1}{p} + \frac{1}{q} = 1$ and $1 \leq p, q \leq \infty$, we have

$$\mu(|fg|) = \int_E |f(x)g(x)| \, \mathrm{d}\mu \leq \|f\|_p \cdot \|g\|_q$$

Remark. For $p = q = 2$, this is exactly the Cauchy–Schwarz inequality on L^2.

Proof. The cases $p = 1$ or $p = \infty$ are obvious. We can assume $f \in L^p$ and $g \in L^q$ without loss of generality since the right hand side would otherwise be infinite. We can also assume f is not equal to zero almost everywhere, otherwise this reduces to $0 \leq 0$. Hence, $\|f\|_p > 0$. Then, we can divide both sides by $\|f\|_p$ and then assume $\|f\|_p = 1$.

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

Note that we can set $|f|^p \, \mathrm{d}\mu = \mathrm{d}\mathbb{P}$, and since $L^q(\mathbb{P}) \subseteq L^1(\mathbb{P})$,

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

□

Theorem (Minkowski's inequality)**.** Let $f, g : (E, \mathcal{E}, \mu) \to \mathbb{R}$ be measurable functions. Then for all $1 \leq p \leq \infty$, we have $\|f + g\|_p \leq \|f\|_p + \|g\|_p$.

Proof. The results for $p = 1, \infty$ are clear. Suppose $1 < p < \infty$. We can assume without loss of generality that $f, g \in L^p$. We can integrate the pointwise inequality $|f + g|^p \leq 2^p(|f|^p + |g|^p)$ to deduce that $\|f + g\|_p^p \leq 2^p(\|f\|_p^p + \|g\|_p^p) < \infty$ so $f + g \in L^p$. We assume that $0 < \|f + g\|_p$, otherwise the result is trivial. Now, using Hölder's inequality with q conjugate to p,

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

Dividing both sides by $\|f + g\|_p^{\frac{p}{q}}$, we obtain $\|f + g\|_p \leq \|f\|_p + \|g\|_p$. □

So the L^p spaces are indeed normed spaces.

5.2 Banach spaces

Definition. A *Banach space* is a complete normed vector space.

Theorem ($\mathcal{L}^p$ is a Banach space)**.** Let $1 \leq p \leq \infty$, and let $f_n \in L^p$ be a Cauchy sequence, so for all $\varepsilon > 0$ there exists N such that for all $m, n \geq N$, we have $\|f_m - f_n\|_p < \varepsilon$. Then there exists a function $f \in L^p$ such that $f_n \to f$ in L^p, so $\|f_n - f\|_p \to 0$ as $n \to \infty$.

Proof. For this proof, we assume $p < \infty$; the other case is already proven in IB Analysis and Topology. Since f_n is Cauchy, using $\varepsilon = 2^{-k}$ we extract a subsequence f_{N_k} of L^p functions such that

$$S = \sum_{k=1}^{\infty} \left\|f_{N_{k+1}} - f_{N_k}\right\|_p \leq \sum_{k=1}^{\infty} 2^{-k} < \infty$$

By Minkowski's inequality, for any K, we have

$$\left\|\sum_{k=1}^{K} |f_{N_{k+1}} - f_{N_k}|\right\|_p \leq \sum_{k=1}^{K} \left\|f_{N_{k+1}} - f_{N_k}\right\|_p \leq S < \infty$$

By the monotone convergence theorem applied to $\left|\sum_{k=1}^{K}\left|f_{N_{k+1}}-f_{N_k}\right|\right|^p$ which increases to $\left|\sum_{k=1}^{\infty}\left|f_{N_{k+1}}-f_{N_k}\right|\right|^p$, we find

$$\left\|\sum_{k=1}^{\infty}\left|f_{N_{k+1}}-f_{N_k}\right|\right\|_p \leq S < \infty$$

Since the integral is finite, we see that $\sum_{k=1}^{\infty}\left|f_{N_{k+1}}-f_{N_k}\right|$ is finite almost everywhere. Then $\sum_{k=1}^{K}(f_{N_{k+1}}(x)-f_{N_k}(x)) = f_{N_{k+1}}(x)-f_{N_1}(x)$ converges in the real line for all x in a set A that has full measure, so $\mu(A^c)=0$. In particular, $f_{N_k}(x)$ is a Cauchy sequence of reals, so by completeness of the real line, we can define the limit

$$f(x)=\begin{cases}\lim_{k\to\infty} f_{N_k}(x) & x\in A\\ 0 & x\in A^c\end{cases}$$

so $f_{N_k}\to f$ as $k\to\infty$ almost everywhere. Now, by Fatou's lemma,

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

Since the f_n are Cauchy,

$$\|f\|_p\leq\underbrace{\|f-f_N\|_p}_{\leq\varepsilon}+\underbrace{\|f_N\|_p}_{<\infty}<\infty$$

so $f\in L^p$, and $\|f_n-f\|_p^p\leq\varepsilon^p$ for $n,N_k\geq N$, so $f_n\to f$ in L^p. □

Remark. If V is any of the spaces

$$C([a,b]);\quad \{f \text{ simple}\};\quad \{f \text{ a linear combination of indicators of intervals}\}$$

then V is dense in $L^1(\mu)$ where μ is the Lebesgue measure on $\mathcal{B}([a,b])$. So the completion $\overline{(V,\|\cdot\|)}$ is exactly $L^1(\mu)$.

5.3 Hilbert spaces

Definition. A symmetric bilinear form $\langle\cdot,\cdot\rangle : V\times V\to\mathbb{R}$ on a real vector space V is called an *inner product* if $\langle v,v\rangle\geq 0$ and $\langle v,v\rangle=0$ implies $v=0$. In this case, we can define a norm $\|v\|=\sqrt{\langle v,v\rangle}$. If $(V,\langle\cdot,\cdot\rangle)$ is complete, we say that it is a *Hilbert space*.

Corollary. The space $\mathcal{L}^2$ is a Hilbert space for the inner product $\langle f,g\rangle=\int_E fg\,\mathrm{d}\mu$.

Example. An analog of the Pythagorean theorem holds. Let $f,g\in L^2$, then $\|f+g\|_2^2=\|f\|_2^2+2\langle f,g\rangle+\|g\|_2^2$. We say f is *orthogonal* to g if $\langle f,g\rangle=0$. f and g are orthogonal if and only if $\|f+g\|_2^2=\|f\|_2^2+\|g\|_2^2$. For centred (mean zero) random variables X,Y, we have $\langle X,Y\rangle=\mathbb{E}[XY]=\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]=\mathrm{Cov}(X,Y)$ which vanishes when X and Y are orthogonal.

Example. The parallelogram identity holds: $\|f+g\|_2^2+\|f-g\|_2^2=2\left(\|f\|_2^2+\|g\|_2^2\right)$

Definition. Let $V \subseteq L^2(\mu)$. We define its *orthogonal complement* to be

$$V^\perp = \{f \in L^2(\mu) \mid \forall g \in V, \langle f, g\rangle = 0\}$$

We say that a subset V of $\mathcal{L}^2$ is *closed* if any sequence $f_n \in V$ that converges in $\mathcal{L}^2$, its limit f coincides almost everywhere with some $v \in V$.

Theorem. Let V be a closed linear subspace of $\mathcal{L}^2(\mu)$. Then for all $f \in \mathcal{L}^2$, there exists an orthogonal decomposition $f = v + u$ where $v \in V$ and $u \in V^\perp$ such that $\|f - v\|_2 \le \|f - g\|_2$ for all $g \in V$ with equality only if $v = g$ almost everywhere. We call v the *projection* of f onto V.

Proof. In this proof, we set $p = 2$ for all norms. We define $d(f, V) = \inf_{g\in V} \|g - f\|$, and let $g_n \in V$ be a sequence of functions such that $\|g_n - f\|$ converges to $d(f, V)$. By the parallelogram law,

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

Taking the limit superior as $n, m \to \infty$, $\limsup_{m,n} \|g_n - g_m\|^2 \le 4d(f, V) - 4d(f, V) = 0$. So the sequence g_n is Cauchy in L^2, so by completeness, it converges to some $v \in L^2$. Since V is closed, $v \in V$. In particular, $d(f, V) = \inf_{g\in V} \|g - f\| = \|v - f\|$.

Note that $d(f, V)^2 \le F(t) = \|f - (v + th)\|^2$ where $t \in \mathbb{R}$ and $h \in V$. So we obtain the first-order condition $F'(0) = 2\langle f - v, h\rangle = 0$ for all h. Defining $f - v = u$, we have $f = u + v$ and $u \in V^\perp$ since h was arbitrary.

For uniqueness, suppose $f = w + z$ with $w \in V$ and $z \in V^\perp$. Then $v - w + u - z = f - f = 0$, so taking norms, $0 = \|v - w + u - z\|^2 = \|v - w\|^2 + \|u - z\|^2$ so $v = w$ and $u = z$ (almost everywhere) by orthogonality. □

5.4 Convergence in probability and uniform integrability

Theorem (bounded convergence)**.** Let X_n be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $|X_n| \le C < \infty$ and they converge in probability to X. Then $X_n \to X$ in $L^1(\mathbb{P})$.

Proof. We know that $X_{n_k} \to X$ almost surely along a subsequence n_k. So $|X| = \lim_k |X_{n_k}| \le C < \infty$ almost surely. Then

$$\begin{aligned}
\mathbb{E}[|X_n - X|] &= \mathbb{E}\left[|X_n - X|\left(\mathbb{1}_{\left\{|X_n - X| > \frac{\varepsilon}{2}\right\}} + \mathbb{1}_{\left\{|X_n - x| \le \frac{\varepsilon}{2}\right\}}\right)\right] \\
&\le 2C\mathbb{P}\left(|X_n - X| \ge \frac{\varepsilon}{2}\right) + \frac{\varepsilon}{2} \\
&< \varepsilon
\end{aligned}$$

for sufficiently large n. □

If $X \in L^1(\mathbb{P})$, then as $\delta \to 0$,

$$I_X(\delta) = \sup\{\mathbb{E}[|X|\mathbb{1}_A] \mid \mathbb{P}(A) \le \delta\} \downarrow 0$$

Suppose this does not hold. Then there exists $\varepsilon > 0$ and a sequence of events $A_n \in \mathcal{F}$ such that $\mathbb{P}(A_n) \le 2^{-n}$ but $\mathbb{E}\left[|X|\mathbb{1}_{A_n}\right] \ge \varepsilon$. Since $\sum_n \mathbb{P}(A_n) < \infty$, by the first Borel–Cantelli lemma, we have $\mathbb{P}\left(\bigcap_n \bigcup_{m\ge n} A_m\right) = 0$. But $\mathbb{E}\left[|X|\mathbb{1}_{A_n}\right] \le \mathbb{E}\left[|X|\mathbb{1}_{\bigcup_{m\ge n} A_m}\right]$. Note that $\mathbb{1}_{\bigcup_{m\ge n} A_m} \to \mathbb{1}_{\bigcap_n \bigcup_{m\ge n} A_n}$, so $\mathbb{E}\left[|X|\mathbb{1}_{\bigcup_{m\ge n} A_m}\right] \to \mathbb{E}\left[|X|\mathbb{1}_{\bigcap_n \bigcup_{m\ge n}}\right]$ by the dominated convergence theorem, but this is equal to zero, giving a contradiction.

Definition. For a collection $\mathcal{X} \subseteq L^1(\mathbb{P})$ of random variables, we say $\mathcal{X}$ is *uniformly integrable* if it is bounded in $L^1(\mathbb{P})$, and

$$I_{\mathcal{X}}(\delta) = \sup\{\mathbb{E}[|X|\mathbb{1}_A] \mid \mathbb{P}(A) \le \delta, X \in \mathcal{X}\} \downarrow 0$$

Remark. Note that $X_n = n\mathbb{1}_{\left[0,\frac{1}{n}\right]}$ for the Lebesgue measure μ on $[0,1]$ is bounded in $L^1(\mathbb{P})$ but not uniformly integrable. If $\mathcal{X}$ is bounded in $L^p(\mathbb{P})$ for $p > 1$, then by Hölder's inequality,

$$\mathbb{E}[|X|\mathbb{1}_A] \le \underbrace{\|X\|_p}_{\text{bounded}} \cdot \underbrace{\mathbb{P}(A)^{\frac{1}{q}}}_{\le \delta^{\frac{1}{q}} \to 0}$$

Lemma. $\mathcal{X} \subseteq L^1(\mathbb{P})$ is uniformly integrable if and only if $\sup_{X\in\mathcal{X}} \mathbb{E}\left[|X|\mathbb{1}_{\{|X|>K\}}\right] \to 0$ as $K \to \infty$.

Proof. Let $\mathcal{X}$ be uniformly integrable. Applying Markov's inequality, as $K \to \infty$,

$$\mathbb{P}(|X| > K) \le \frac{\mathbb{E}[|X|]}{K} = \frac{\mathbb{E}[|X|\mathbb{1}_\Omega]}{K} \le \frac{I_{\mathcal{X}}(1)}{K} \to 0$$

Using the uniform integrability property using $A = \{|X| > K\}$, we obtain the required limit. Conversely, we have

$$\mathbb{E}[|X|] = \mathbb{E}\left[|X|\left(\mathbb{1}_{\{|X|\le K\}} + \mathbb{1}_{\{|X|>K\}}\right)\right] \le K + \frac{\varepsilon}{2}$$

for sufficiently large K. So $\mathcal{X}$ is bounded in $L^1(\mathbb{P})$ as required. Then for A such that $\mathbb{P}(A) \le \delta$,

$$\mathbb{E}\left[|X|\mathbb{1}_A\left(\mathbb{1}_{\{|X|\le K\}} + \mathbb{1}_{\{|X|>K\}}\right)\right] \le K\mathbb{P}(A) + \mathbb{E}\left[|X|\mathbb{1}_{\{|X|>K\}}\right] \le K\delta + \frac{\varepsilon}{2} < \varepsilon$$

for sufficiently small δ. □

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

(i) $X_n, X \in L^1(\mathbb{P})$ and $X_n \to X$ in $L^1(\mathbb{P})$.
(ii) $\{X_n \mid n \in \mathbb{N}\}$ is uniformly integrable, and $X_n \to X$ in probability.

Proof. (i) implies (ii). Using Markov's inequality,

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

so $X_n \to X$ in probability. Since any finite collection is uniformly integrable, so are X along with $X_1, \dots, X_N$ for each N. For the indices larger than N, we have

$$\mathbb{E}[|X_n| \mathbb{1}_A] \leq \mathbb{E}[|X_n - X| \mathbb{1}_A] + \mathbb{E}[|X| \mathbb{1}_A] \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$$

for sufficiently large N and sufficiently small δ, so all X_n are uniformly integrable.

(ii) implies (i). Along a subsequence, $X_n \to X$ almost surely. So

$$\mathbb{E}[|X|] = \mathbb{E}\left[\liminf_k |X_{n_k}|\right] \leq \liminf_k \mathbb{E}\left[|X_{n_k}|\right] \leq I_{\mathcal{X}}(1) < \infty$$

almost surely, so $X \in L^1(\mathbb{P})$. Next, we define random variables $g(X_n) = X_n^K = \max(-K, \min(K, X_n))$ and $g(X) = X^K = \max(-K, \min(K, X))$, where g is continuous. Then for some $\varepsilon' > 0$,

$$\mathbb{P}(|g(X_n) - g(X)| > \varepsilon) \leq \mathbb{P}(|X_n - X| > \varepsilon') \to 0$$

as $n \to \infty$, since $X_n \to X$ in probability and g is continuous. Then by bounded convergence, $X_n^K \to X^K$ in L^1, and so

$$\begin{aligned}
\mathbb{E}[|X_n - X|] &\leq \mathbb{E}\left[|X_n - X_n^K|\right] + \mathbb{E}\left[|X_n^K - X^K|\right] + \mathbb{E}\left[|X^K - X|\right] \\
&= \mathbb{E}\left[|X_n| \mathbb{1}_{\{|X_n| > k\}}\right] + \mathbb{E}\left[|X_n^K - X^K|\right] + \mathbb{E}\left[|X| \mathbb{1}_{\{|X| > K\}}\right] \\
&< \varepsilon
\end{aligned}$$

by choosing sufficiently large K and n. □

6 Fourier analysis

6.1 Fourier transforms

In this section, we will write $L^p(\mathbb{R}^d)$ for the set of measurable functions $f : \mathbb{R}^d \to \mathbb{C}$ such that $\|f\|_p = \left(\int_{\mathbb{R}^d} |f(x)|^p \, \mathrm{d}x\right)^{\frac{1}{p}} < \infty$. We can extend the integral as a complex linear map $L^1(\mathbb{R}) \to \mathbb{C}$ by defining

$$\int_{\mathbb{R}} (u + iv)(x) \, \mathrm{d}x = \int_{\mathbb{R}} u(x) \, \mathrm{d}x + i \int_{\mathbb{R}} v(x) \, \mathrm{d}x$$

Note that for some $u + iv = \alpha \in \mathbb{C}$ with $|\alpha| = 1$,

$$\left| \int_{\mathbb{R}^d} f(x) \, \mathrm{d}x \right| = \int_{\mathbb{R}^d} \alpha f(x) \, \mathrm{d}x = \int_{\mathbb{R}^d} u(x) \, \mathrm{d}x + i \int_{\mathbb{R}^d} v(x) \, \mathrm{d}x$$

But since the left hand side is real-valued, the $i \int_{\mathbb{R}^d} v(x) \, \mathrm{d}x$ term vanishes. So

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

Definition. Let $f \in L^1(\mathbb{R}^d)$. We define the *Fourier transform* $\hat{f}$ by

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

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

Remark. Note that $\left|\hat{f}(u)\right| \leq \|f\|_1$. Also, if $u_n \to u$, then $e^{i\langle u_n,x\rangle} \to e^{i\langle u,x\rangle}$. By the dominated convergence theorem with dominating function $|f|$, we have $\hat{f}(u_n) \to \hat{f}(u)$, so $\hat{f}$ is a continuous bounded function.

Definition. Let $f \in L^1(\mathbb{R}^d)$ such that $\hat{f} \in L^1(\mathbb{R}^d)$. Then we say that the *Fourier inversion formula* holds for f if

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

almost everywhere in $\mathbb{R}^d$.

Definition. Let $f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. Then the *Plancherel identity* holds for f if

$$\left\|\hat{f}\right\|_2 = (2\pi)^{\frac{d}{2}} \|f\|_2$$

We will show that the Fourier inversion formula holds whenever $\hat{f} \in L^1(\mathbb{R}^d)$, and the Plancherel identity holds for all $f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$.

Remark. Given the Plancherel identity, the Fourier transform is a linear isometry of $L^2(\mathbb{R}^d)$, by approximating any function in $L^2(\mathbb{R}^d)$ by integrable functions.

Definition. Let μ be a finite Borel measure on $\mathbb{R}^d$. We define the Fourier transform of the measure by

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

Note that $|\hat{\mu}(u)| \leq \mu(\mathbb{R}^d)$, and $\hat{\mu}$ is continuous by the dominated convergence theorem. If μ has a density f with respect to the Lebesgue measure, $\hat{\mu} = \hat{f}$.

Definition. Let X be an $\mathbb{R}^d$-valued random variable. The *characteristic function* φ_X is given by

$$\varphi_X(u) = \mathbb{E}\left[e^{i\langle u,X\rangle}\right] = \hat{\mu}_X(u)$$

where μ_X is the law of X.

6.2 Convolutions

Definition. Let $f \in L^1(\mathbb{R}^d)$ and ν be a probability measure on $\mathbb{R}^d$. We define their *convolution* $f * \nu$ by

$$(f * \nu)(x) = \begin{cases} \int_{\mathbb{R}^d} f(x-y)\,\mathrm{d}\nu(y) & \text{if } (y \mapsto f(x-y)) \in L^1(\nu) \\ 0 & \text{else} \end{cases}$$

Remark. If $1 \le p < \infty$, by Jensen's inequality,

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

So $f \in L^p(\mathbb{R}^d)$, we have $(y \mapsto f(x-y)) \in L^p(\nu)$ almost everywhere, and again by Jensen's inequality,

$$\|f * \nu\|_p^p = \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} f(x-y)\,\mathrm{d}\nu(y) \right|^p \mathrm{d}x \le \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} |f(x-y)|\,\mathrm{d}\nu(y) \right)^p \mathrm{d}x \le \|f\|_p^p$$

Hence $f \mapsto f * \nu$ is a contraction on $L^p(\mathbb{R}^d)$.

In the case where ν has a density g with respect to the Lebesgue measure, we write $f * g = f * \nu$.

Definition. For probability measures μ, ν on $\mathbb{R}^d$, their convolution $\mu * \nu$ is a probability measure on $\mathbb{R}^d$ given by the law of $X + Y$ where X, Y are independent random variables with laws μ and ν, so

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

If μ has density f with respect to the Lebesgue measure, $\mu * \nu$ has density $f * \nu$ with respect to the

Lebesgue measure. Indeed,

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

Proposition. $\widehat{f * \nu}(u) = \hat{f}(u)\hat{\nu}(u)$.

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

6.3 Fourier transforms of Gaussians

Definition. The *normal distribution* $N(0,t)$ is given by the probability density function

$$g_t(x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}$$

If φ_X is the characteristic function of a standard normal random variable, by integration by parts,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}u}\varphi_X(u) &= \frac{\mathrm{d}}{\mathrm{d}u}\int_{\mathbb{R}} e^{iux} g_1(x)\,\mathrm{d}x \\ &= \int_{\mathbb{R}} g_1(x)\frac{\mathrm{d}}{\mathrm{d}u} e^{iux}\,\mathrm{d}x \\ &= \frac{i}{\sqrt{2\pi}}\int_{\mathbb{R}} \underbrace{e^{iux}}_{v}\,\underbrace{x e^{-\frac{x^2}{2}}}_{w'}\,\mathrm{d}x \\ &= \frac{i^2}{\sqrt{2\pi}}\int_{\mathbb{R}} u e^{iux} e^{-\frac{x^2}{2}}\,\mathrm{d}x \\ &= -u\varphi_X(u)\end{aligned}$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}u}\left(e^{\frac{u^2}{2}}\varphi_X(u)\right) = u e^{\frac{u^2}{2}}\varphi_X(u) - e^{\frac{u^2}{2}} u\varphi_X(u) = 0$$

In particular, $\varphi_X(u) = \varphi_X(0)e^{-\frac{u^2}{2}} = e^{-\frac{u^2}{2}}$. In other words, $\hat{g}_1(u) = \sqrt{2\pi} g_1(u)$.

In $\mathbb{R}^d$, consider a Gaussian random vector $Z = (Z_1, \dots, Z_d)$ with independent and identically distributed entries $Z_i \sim N(0,1)$. Then, the joint probability density function of $\sqrt{t}Z$ is

$$g_t(x) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi t}} e^{-\frac{x_j^2}{2t}} = (2\pi t)^{-\frac{d}{2}} e^{-\frac{\|x\|^2}{2t}}$$

The Fourier transform of g_t is

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

which implies that in general, $\hat{g}_t(u) = (2\pi)^{\frac{d}{2}} t^{\frac{d}{2}} g_{\frac{1}{t}}(u)$. Taking the Fourier transform with respect to u, $\hat{\hat{g}}_t = (2\pi)^d g_t$, and since $g_t(-x) = g_t(x)$ and the Lebesgue measure is translation invariant, we have

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

so the Fourier inversion theorem holds for such Gaussian random vectors.

Definition. We say that a function on $\mathbb{R}^d$ is a *Gaussian convolution* if it is of the form

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

where $x \in \mathbb{R}^d, t > 0, f \in L^1(\mathbb{R}^d)$.

We can show that $f * g_t$ is continuous on $\mathbb{R}^d$, and $\|f * g_t\|_1 \le \|f\|_1$. Note that $\widehat{f * g_t}(u) = \hat{f}(u)e^{-\frac{\|u\|^2 t}{2}}$, so $\left\|\widehat{f * g_t}\right\|_\infty \le \|f\|_1$, giving $\left\|\widehat{f * g_t}\right\|_1 \le \|f\|_1 (2\pi)^{\frac{d}{2}} t^{-\frac{d}{2}} < \infty$.

Lemma. The Fourier inversion theorem holds for all Gaussian convolutions.

Proof. We can use the Fourier inversion theorem for $g_t(y)$ to see that

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

□

Remark. If μ is a finite measure, then $\mu * g_t = \mu * g_{\frac{t}{2}} * g_{\frac{t}{2}}$ with $\mu * g_{\frac{t}{2}} \in L^1$, so is also a Gaussian convolution.

Lemma (Gaussian convolutions are dense in L^p)**.** Let $f \in L^p$ where $1 \le p < \infty$. Then $\|f * g_t - f\|_p \to 0$ as $t \to 0$.

Proof. One can easily show that the space $C_c(\mathbb{R}^d)$ of continuous functions of compact support is dense in L^p. Hence, for all $\varepsilon > 0$, there exists $h \in C_c(\mathbb{R}^d)$ such that $\|f - h\|_p < \frac{\varepsilon}{3}$, and by properties of the convolution, we also obtain

$$\|f * g_t - h * g_t\|_p = \|(f - h) * g_t\|_p \le \|f - h\|_p < \frac{\varepsilon}{3}$$

So

$$\|f * g_t - f\|_p \le \|f * g_t - h * g_t\|_p + \|h * g_t + h\|_p + \|h - f\|_p < \frac{\varepsilon}{2} + \|h * g_t - h\|_p$$

so it suffices to prove the result for $f = h \in C_c(\mathbb{R}^d)$. We define a new map

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

Since h is bounded on its bounded support, the dominated convergence theorem implies that e is continuous at $y = 0$. Note that $e(y) \le 2^{p+1}\|h\|_p^p$. Hence, by Jensen's inequality,

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

□

Theorem (Fourier inversion)**.** Let $f \in L^1(\mathbb{R}^d)$ be such that $\hat{f} \in L^1(\mathbb{R}^d)$. Then for almost all $x \in \mathbb{R}^d$,

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

Remark. This proves that the Fourier transform is injective; $\hat{f} = \hat{g}$ implies $\widehat{f - g} = 0$ so by Fourier inversion, $f = g$ almost everywhere. The identity holds everywhere on $\mathbb{R}^d$ for the (unique) continuous representative f in its equivalence class.

Proof. The Fourier inversion theorem holds for the following Gaussian convolution for all t.

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

Now, since Gaussian convolutions are dense, $f * g_t \to f$ in L^1, so $f * g_t \to f$ in measure by Markov's inequality. Hence, along a subsequence, $f * g_{t_k} \to f$ almost everywhere. On the other hand, by the dominated convergence theorem with dominating function $|\hat{f}|$, the right hand side converges to $\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\langle u,x\rangle} \hat{f}(u)\,\mathrm{d}u$. So this is equal to $\lim_{t_k \to 0} f_{t_k}$ almost everywhere by uniqueness of limits. □

Theorem (Plancherel). Let $f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. Then $\|f\|_2 = (2\pi)^{-\frac{d}{2}} \left\|\hat{f}\right\|_2$.

Remark. By the Pythagorean identity, $\langle f, g\rangle = (2\pi)^{-d} \langle \hat{f}, \hat{g}\rangle$.

Proof. Initially, we assume $\hat{f} \in L^1$. In this case, $f, \hat{f} \in L^\infty$, and $(x,u) \mapsto f(x)\hat{f}(u)$ is integrable for the product Lebesgue measure $\mathrm{d}x \otimes \mathrm{d}u$ on $\mathbb{R}^d \times \mathbb{R}^d$, so Fubini's theorem for bounded functions applies.

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

To extend this result to general f, we take the Gaussian convolutions $f * g_t = f_t$ such that $f_t \to f$ in L^2. By the continuity of the norm, $\|f_t\|_2 \to \|f\|_2$. Since $\left|\hat{f}(u) e^{-\frac{|u|^2 t}{2}}\right|^2$ increases to $\left|\hat{f}(u)\right|^2$, we have by monotone convergence that $\left\|\hat{f}_t\right\|_2^2 \uparrow \left\|\hat{f}\right\|_2^2$. Therefore, since the Plancherel identity holds for the f_t,

$$\|f\|_2^2 = \lim_{t\to 0} \|f_t\|_2^2 = \lim_{t\to 0} (2\pi)^{-d} \left\|\hat{f}_t\right\|_2^2 = (2\pi)^{-d} \left\|\hat{f}\right\|_2^2$$

□

Remark. Since $L_1 \cap L_2$ is dense in L^2, we can extend the linear operator $F_0(f) = (2\pi)^{-\frac{d}{2}} \hat{f}$ to L^2 by continuity to a linear isometry $F : L^2 \to L^2$ known as the *Fourier–Plancherel transform*. One can show that F is surjective with inverse $F^{-1} : L^2 \to L^2$.

Example. Consider the Dirac measure δ_0 on $\mathbb{R}$, so $\hat{\delta}_0(u) = \int_{\mathbb{R}} e^{iux}\,\mathrm{d}\delta_0(x) = 1$. But the inverse Fourier transform would be $\frac{1}{2\pi} \int_{\mathbb{R}} e^{iux}\,\mathrm{d}u$ which is not a Lebesgue integrable function.

Theorem. Let X be a random vector in $\mathbb{R}^d$ with law μ_X. Then the characteristic function $\varphi_X = \hat{\mu}_X$ uniquely determines μ_X. In addition, if $\varphi_X \in L^1$, then μ_X has a probability density

function f_X which can be computed almost everywhere by $\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{-i\langle u,x\rangle}\varphi_X(u)\,\mathrm{d}u$.

Proof. Let $Z = (Z_1, \dots, Z_d)$ be a vector of independent and identically distributed random variables, independent of X, with $Z_j \sim N(0,1)$. Then $\sqrt{t}Z$ has probability density function g_t. Then $X+\sqrt{t}Z$ has probability density function $f_t = \mu_X * g_t$. This is a Gaussian convolution since $\mu_X * g_t = \mu_X * g_{\frac{t}{2}} * g_{\frac{t}{2}}$. Hence,

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

which is uniquely determined by φ_X. We show on an example sheet that two Borel probability measures μ, ν on $\mathbb{R}^d$ coincide if and only if $\mu(g) = \nu(g)$ for all $g : \mathbb{R}^d \to \mathbb{R}$ that are bounded, continuous, and have compact support. Now,

$$\int_{\mathbb{R}^d} g(x)f_t(x)\,\mathrm{d}x = \mathbb{E}\left[\underbrace{g(X+\sqrt{t}Z)}_{\to X \text{ a.s.}}\right]$$

Since $\left|g(X+\sqrt{t}Z)\right| \le \|g\|_\infty < \infty$, by the bounded convergence theorem, this converges to $\mathbb{E}[g(X)] = \int_{\mathbb{R}^d} g(x)\,\mathrm{d}\mu_X(x)$. So by uniqueness of limits, φ_X determines μ_X.

If $\varphi_X \in L^1$, by dominated convergence, $f_t(x)$ converges everywhere to some function f_X. In particular, since $\mu_X * g_t \ge 0$, the limit f_X is also nonnegative on $\mathbb{R}^d$. Then, for any bounded continuous function on compact support $g \in C_c^b(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d} g(x)f_X(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} g(x)\lim_{t\to 0}\underbrace{f_t(x)}_{\|\varphi_X\|_1}\mathrm{d}x = \lim_{t\to 0}\int_{\mathbb{R}^d} g(x)f_t(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} g(x)\,\mathrm{d}\mu_X(x)$$

by the dominated convergence theorem, since g has compact support. □

Definition. A sequence $(\mu_n)_{n\in\mathbb{N}}$ of Borel probability measures on $\mathbb{R}^d$ *converges weakly* to a Borel probability measure μ if $\mu_n(g) \to \mu(g)$ for all $g : \mathbb{R}^d \to \mathbb{R}$ bounded and continuous. If $(X_n)_{n\in\mathbb{N}}, X$ are random vectors with laws $(\mu_{X_n}), \mu_X$ such that μ_{X_n} converges weakly to μ_X, we say (X_n) converges weakly to X.

Remark. If $d = 1$, weak convergence is equivalent to convergence in distribution; this is proven on an example sheet. One can also show that convergence of $\mu_n(g)$ to $\mu(g)$ for all $g \in C_c^\infty(\mathbb{R}^d)$ suffices to show weak convergence, where $C_c^\infty(\mathbb{R}^d)$ is the space of smooth functions of compact support. This is equivalent to the notion of weak-* convergence on the function space $C_b(\mathbb{R}^d)$.

Theorem (Lévy's continuity theorem)**.** Let X_n, X be random vectors in $\mathbb{R}^d$, such that $\varphi_{X_n}(u) \to \varphi_X(u)$ for all u, as $n \to \infty$. Then $\mu_{X_n} \to \mu_X$ weakly.

Remark. The converse holds by definition of weak convergence, testing against the complex exponentials in the Fourier transform.

Proof. Let $Z = (Z_1, \dots, Z_d)$ be a vector of standard normal random variables, independent from each other, X_n, and X. Let $g \in C_c^\infty(\mathbb{R}^d)$. Then $g \in L^1(\mathbb{R}^d)$, and is Lipschitz by the mean value theorem, as its first derivative is bounded. Let $|g(x) - g(y)| \le \|g\|_{\text{Lip}}|x - y|$. Let $\varepsilon > 0$. Let $t > 0$ be sufficiently small such that $\sqrt{t}\|g\|_{\text{Lip}}\mathbb{E}[|Z|] < \frac{\varepsilon}{3}$. Then,

$$\begin{aligned}
|\mu_{X_n}(g) - \mu_X(g)| &= |\mathbb{E}[g(X_n)] - \mathbb{E}[g(X)]| \\
&\le \mathbb{E}\left[\left|g(X_n) - g(X_n + \sqrt{t}Z)\right|\right] + \mathbb{E}\left[\left|g(X) - g(X + \sqrt{t}Z)\right|\right] \\
&+ \left|\mathbb{E}\left[g(X_n + \sqrt{t}Z) - g(X + \sqrt{t}Z)\right]\right| \\
&\le 2\|g\|_{\text{Lip}}\sqrt{t}\mathbb{E}[|Z|] + \left|\mathbb{E}\left[g(X_n + \sqrt{t}Z) - g(X + \sqrt{t}Z)\right]\right| \\
&\le \frac{2\varepsilon}{3} + \left|\mathbb{E}\left[g(X_n + \sqrt{t}Z) - g(X + \sqrt{t}Z)\right]\right|
\end{aligned}$$

We show that the remaining term can be made less than $\frac{\varepsilon}{3}$ as $n \to \infty$. Let $f_{t,n}(x) = g_t * \mu_{X_n}$. Then, by Fourier inversion for Gaussian convolutions,

$$\begin{aligned}
\mathbb{E}\left[g(X_n + \sqrt{t}Z)\right] &= \int_{\mathbb{R}^d} g(x) f_{t,n}(x)\,\mathrm{d}x \\
&= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} g(x) \int_{\mathbb{R}^d} e^{-i\langle u,x\rangle}\varphi_{X_n}(u)e^{-\frac{|u|^2 t}{2}}\,\mathrm{d}u\,\mathrm{d}x
\end{aligned}$$

Since characteristic functions are bounded by 1, we can apply the dominated convergence theorem with dominating function $|g(x)|e^{-\frac{|u|^2 t}{2}}$ to find

$$\begin{aligned}
\mathbb{E}\left[g(X_n + \sqrt{t}Z)\right] &\to \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} g(x) \int_{\mathbb{R}^d} e^{-i\langle u,x\rangle}\varphi_X(u)e^{-\frac{|u|^2 t}{2}}\,\mathrm{d}u\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d} g(x) f_t(x)\,\mathrm{d}x \\
&= \mathbb{E}\left[g(X + \sqrt{t}Z)\right]
\end{aligned}$$

where $f_t = g_t * \mu_X$. So as $n \to \infty$, the difference between these two terms can be made less than $\frac{\varepsilon}{3}$ as required. □

Theorem (central limit theorem)**.** Let $X_1, \dots, X_n$ be independent and identically distributed random variables with $\mathbb{E}[X_i] = 0$ and $\operatorname{Var}(X_i) = 1$. Let $S_n = \sum_{i=1}^n X_n$. Then

$$\frac{1}{\sqrt{n}}S_n \xrightarrow{\text{weakly}} Z \sim N(0,1)$$

In particular,

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}S_n \le x\right) \to \mathbb{P}(Z \le x)$$

Proof. Let $X = X_1$. The characteristic function $\varphi(u) = \varphi_X(u) = \mathbb{E}\left[e^{iuX}\right]$ satisfies $\varphi(0) = 1$, $\varphi'(u) = i\mathbb{E}\left[Xe^{iuX}\right]$, $\varphi''(u) = i^2\mathbb{E}\left[X^2 e^{iuX}\right]$. We can find $\varphi'(0) = i\mathbb{E}[X] = 0$ and $\varphi''(0) = -\mathbb{E}\left[X^2\right] = -\operatorname{Var}(X) =$

-1. By Taylor's theorem, $\varphi(v) = 1 - \frac{v^2}{2} + o(v^2)$ as $v \to 0$. Now, denoting $\varphi_n(u) = \varphi_{\frac{1}{\sqrt{n}} S_n}(u)$, we can write

$$\begin{aligned}
\varphi_n(u) &= \mathbb{E}\left[e^{i\frac{u}{\sqrt{n}}(X_1+\cdots+X_n)}\right] \\
&= \prod_{j=1}^{n} \mathbb{E}\left[e^{i\frac{u}{\sqrt{n}}X_j}\right] \\
&= \left[\varphi\left(\frac{u}{\sqrt{n}}\right)\right]^n \\
&= \left[1 - \frac{u^2}{2n} + o\left(\frac{1}{n}\right)\right]^n
\end{aligned}$$

The complex logarithm satisfies $\log(1+z) = z + o(z)$, so by taking logarithms, we find

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

Hence, $\varphi_n(u) \to e^{-\frac{|u|^2}{2}} = \varphi_Z(u)$. So by Lévy's continuity theorem, the result follows. □

Remark. This theorem extends to $\mathbb{R}^d$ by using the next proposition, using the fact that $X_n \to X$ weakly in $\mathbb{R}^d$ if and only if $\langle X_n, v\rangle \to \langle X, v\rangle$ weakly in $\mathbb{R}$ for all $v \in \mathbb{R}^d$.

Definition. A random variable X in $\mathbb{R}^d$ is called a *Gaussian vector* if $\langle X_n, v\rangle$ are Gaussian for each $v \in \mathbb{R}^d$.

Proposition. Let X be a Gaussian vector in $\mathbb{R}^d$. Then $Z = AX + b$ is a Gaussian vector in $\mathbb{R}^m$ where A is an $m \times d$ matrix and $b \in \mathbb{R}^m$. Also, $X \in L^2(\mathbb{R}^d)$, and $\mu = \mathbb{E}[X]$ and $V = \operatorname{Cov}(X_i, X_j)$ exist and determine μ_X. The characteristic function is

$$\varphi_X(u) = e^{i\langle \mu, u\rangle - \frac{\langle u, Vu\rangle}{2}}$$

If V is invertible, then μ_X has a probability density function

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

Subvectors $X_{(1)}, X_{(2)}$ of X are independent if and only if $\operatorname{Cov}(X_{(1)}, X_{(2)}) = 0$.

Proposition. Let $X_n \to X$ weakly in $\mathbb{R}^d$ as $n \to \infty$. Then,

(i) if $h : \mathbb{R}^d \to \mathbb{R}^k$ is continuous, then $h(X_n) \to h(X)$ weakly;
(ii) if $|X_n - Y_n| \to 0$ in probability, then $Y_n \to X$ weakly;
(iii) if $Y_n \to c$ in probability where c is constant on Ω, then $(X_n, Y_n) \to (X, c)$ weakly in $\mathbb{R}^d \times \mathbb{R}^d$.

Remark. Combining parts (iii) and (i), $X_n + Y_n \to X + c$ weakly if $Y_n \to c$ in probability. If $d = 1$, then in addition $X_n Y_n \to cX$ weakly.

Proof. *Part (i).* This follows from the fact that gh is continuous for any test function g.

Part (ii). Let $g\colon \mathbb{R}^d \to \mathbb{R}$ be bounded and Lipschitz continuous. Then

$$|\mathbb{E}[g(Y_n)] - \mathbb{E}[g(X)]| \leq \underbrace{|\mathbb{E}[g(X_n)] - \mathbb{E}[g(X)]|}_{<\frac{\varepsilon}{3}} + \mathbb{E}[|g(X_n) - g(Y_n)|]$$

where the bound on $\mathbb{E}[g(X_n)] - \mathbb{E}[g(X)]$ holds for sufficiently large n. Then the remaining term is upper bounded by

$$\mathbb{E}[|g(X_n) - g(Y_n)|]\left(\mathbb{1}_{\left\{|X_n - Y_n| \leq \frac{\varepsilon}{3\|g\|_{\mathrm{Lip}}}\right\}} + \mathbb{1}_{\left\{|X_n - Y_n| > \frac{\varepsilon}{3\|g\|_{\mathrm{Lip}}}\right\}}\right)$$

$$\leq \|g\|_{\mathrm{Lip}} \frac{\varepsilon}{3\|g\|_{\mathrm{Lip}}} + 2\|g\|_\infty \mathbb{P}\left(|X_n - Y_n| > \frac{\varepsilon}{3\|g\|_{\mathrm{Lip}}}\right) < \frac{2\varepsilon}{3}$$

for sufficiently large n.

Part (iii). $|(X_n, c) - (X_n, Y_n)| = |Y_n - c| \to 0$ in probability. Also, $\mathbb{E}[g(X_n, c)] \to \mathbb{E}[g(X, c)]$ for all bounded continuous maps $g\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, so $(X_n, c) \to (X, c)$ weakly. Hence, by (ii), $(X_n, Y_n) \to (X, c)$ weakly. □

7 Ergodic theory

7.1 Laws of large numbers

Proposition. Let $(X_n)_{n\in\mathbb{N}}$ be independent and identically distributed random variables such that $\mathbb{E}[X_n] = 0$ and $\mathrm{Var}(X_n) = \sigma^2 < \infty$. Then $\frac{1}{n}\sum_{i=1}^n X_i \to 0$ in probability as $n \to \infty$.

Proof. By Chebyshev's inequality,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| > \varepsilon\right) \leq \frac{1}{n^2\varepsilon^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0$$

So $\frac{1}{n}\sum_{i=1}^n X_i \to \mathbb{E}[X_1]$ in probability. □

This is known as the weak law of large numbers. However, this result has several weaknesses, and we can provide stronger results.

Proposition. Let $(X_n)_{n\in\mathbb{N}}$ be independent random variables such that $\mathbb{E}[X_n] = \mu$ and $\mathbb{E}\left[X_n^4\right] \leq M$ for all n. Then $\frac{1}{n}\sum_{i=1}^n X_i \to \mu$ almost surely as $n \to \infty$.

Proof. Let $Y_n = X_n - \mu$. Then $\mathbb{E}[Y_n] = 0$, and $\mathbb{E}\left[Y_n^4\right] \leq 2^4\left(\mathbb{E}\left[X_n^4\right] + \mu^4\right) < \infty$. So we can assume $\mu = 0$. For distinct indices i, j, k, ℓ, by independence and the Cauchy–Schwarz inequality, we have

$$0 = \mathbb{E}\left[X_i X_j X_k X_\ell\right] = \mathbb{E}\left[X_i^2 X_j X_j\right] = \mathbb{E}\left[X_i^3 X_j\right]; \quad \mathbb{E}\left[X_i^2 X_j^2\right] \leq \sqrt{\mathbb{E}\left[X_i^4\right]}\sqrt{\mathbb{E}\left[X_j^4\right]} \leq M$$

So we can compute

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

Let $S_n = \sum_{i=1}^n X_i$. Then,

$$\mathbb{E}\left[\sum_{n=1}^\infty \left(\frac{S_n}{n}\right)^4\right] \le \sum_{n=1}^\infty \frac{1}{n^4} 3n^2M < \infty$$

Hence $\sum_{n=1}^\infty \left(\frac{S_n}{n}\right)^4 < \infty$ almost surely. But then $\left(\frac{S_n}{n}\right)^4 \to 0$ almost surely, so $\frac{S_n}{n} \to 0$ almost surely. □

7.2 Invariants

Let $(E, \mathcal{E}, \mu)$ be a σ-finite measure space.

Definition. A measurable transformation $\Theta : E \to E$ is *measure-preserving* if $\mu(\Theta^{-1}(A)) = \mu(A)$ for all $A \in \mathcal{E}$.

In this case, for any integrable function $f \in L^1(\mu)$, we have $\int_E f \, \mathrm{d}\mu = \int_E f \circ \Theta \, \mathrm{d}\mu$.

Definition. A measurable map $f : E \to \mathbb{R}$ is called Θ-*invariant* if $f \circ \Theta = f$. A set $A \in \mathcal{E}$ is Θ-invariant if $\Theta^{-1}(A) = A$, or equivalently, $\mathbb{1}_A$ is Θ-invariant.

The collection $\mathcal{E}_\Theta$ of Θ-invariant sets forms a σ-algebra over E. A function $f : E \to \mathbb{R}$ is invariant if and only if f is $\mathcal{E}_\Theta$-measurable; this is a question on an example sheet.

Definition. Θ is called *ergodic* if the Θ-invariant sets A satisfy either $\mu(A) = 0$ or $\mu(E \setminus A) = 0$.

If f is Θ-invariant and Θ is ergodic, then one can show that f is constant almost everywhere on E.

Example. Consider $(E, \mathcal{E}) = ((0,1], \mathcal{B})$ with the Lebesgue measure μ. The maps $\Theta_a(x) = x + a$ modulo 1 and $\Theta(x) = 2x$ modulo 1 are both measure-preserving, and ergodic unless $a \in \mathbb{Q}$. This is a question on an example sheet.

Lemma (maximal ergodic lemma)**.** Let $(E, \mathcal{E}, \mu)$ be a σ-finite measure space. Let $\Theta : E \to E$ be measure-preserving. For $f \in L^1(\mu)$, we define $S_0(f) = 0$ and $S_n(f) = \sum_{k=0}^{n-1} f \circ \Theta^k$. Let $S^\star = S^\star(f) = \sup_{n \ge 0} S_n(f)$. Then $\int_{\{S^\star > 0\}} f \, \mathrm{d}\mu \ge 0$.

Proof. Define $S_n^\star = \max_{k \le n} S_k$. Then clearly $S_n^\star \uparrow S^\star$, and $S_k \le S_n^\star$ for all $k \le n$. Note that $S_{k+1} = S_k \circ \Theta + f \le S_n^\star \circ \Theta + f$.

Define $A_n = \{S_n^\star > 0\}$, so $A_n \uparrow \{S^\star > 0\}$. On A_n, we have

$$S_n^\star = \max_{1 \le k \le n} S_k \le \max_{0 \le k \le n} S_{k+1} \le S_n^\star \circ \Theta + f$$

since $S_0 = 0$. We can integrate this inequality to find

$$\int_{A_n} S_n^\star \, \mathrm{d}\mu \leq \int_{A_n} S_n^\star \circ \Theta \, \mathrm{d}\mu + \int_{A_n} f \, \mathrm{d}\mu$$

On the complement A_n^c, we must have $S_n^\star = 0 \leq S_n^\star \circ \Theta$. Hence,

$$\int_E S_n^\star \, \mathrm{d}\mu \leq \int_E S_n^\star \circ \Theta \, \mathrm{d}\mu + \int_{A_n} f \, \mathrm{d}\mu$$

Since Θ is measure-preserving,

$$\int_E S_n^\star \, \mathrm{d}\mu \leq \int_E S_n^\star \, \mathrm{d}\mu + \int_{A_n} f \, \mathrm{d}\mu$$

so we obtain

$$\int_{A_n} f \, \mathrm{d}\mu \geq 0$$

Since $f\mathbb{1}_{A_n} \to f\mathbb{1}_{\{S^\star > 0\}}$ pointwise, and $|f\mathbb{1}_{A_n}| \leq |f| \in L^1(\mu)$, we can apply the dominated convergence theorem to show that

$$\int_{\{S^\star > 0\}} f \, \mathrm{d}\mu = \lim_{n\to\infty} \int_{A_n} f \, \mathrm{d}\mu \geq 0$$

as required. □

7.3 Ergodic theorems

Theorem (Birkhoff). Let $(E, \mathcal{E}, \mu)$ be a σ-finite measure space. Let $\Theta : E \to E$ be measure-preserving. For $f \in L^1(\mu)$, we define $S_0(f) = 0$ and $S_n(f) = \sum_{k=0}^{n-1} f \circ \Theta^k$. Then there exists a Θ-invariant integrable function $\overline{f} \in L^1(\mu)$ with $\mu\left(\left|\overline{f}\right|\right) \leq \mu(|f|)$ such that $\frac{S_n(f)}{n} \to \overline{f}$ almost everywhere.

The proof of Birkhoff's ergodic theorem is non-examinable.

Proof (non-examinable). Note that

$$\limsup_n \frac{S_n(f)}{n} = \limsup_n \frac{S_n(f) \circ \Theta}{n}$$

and the same holds for $\liminf_n$. Hence $\limsup_n \frac{S_n(f)}{n}$ and $\liminf_n \frac{S_n(f)}{n}$ are invariant functions. So they are $\mathcal{E}_\Theta$-measurable. Hence

$$D = D_{a,b} = \left\{ \liminf_n \frac{S_n(f)}{n} < a < b < \limsup_n \frac{S_n(f)}{n} \right\}$$

are measurable and invariant sets. Without loss of generality, let $b > 0$. Let $B \in \mathcal{E}$, where $B \subseteq D$ such that $\mu(B) < \infty$. Let $g = f - b\mathbb{1}_B \in L^1(\mu)$. Then,

$$S_n(g) = S_n(f) - bS_n(\mathbb{1}_B) \geq S_n(f) - bn$$

which is positive on D for some n by the definition of $\limsup_n$. We will apply the maximal ergodic lemma with $E = D$ and $\mu = \mu|_D$; Θ is still measure-preserving on this new measure since

$$\mu\Big|_D(A) = \mu(A \cap D) = \mu(\Theta^{-1}(A \cap D)) = \mu(\Theta^{-1}(A) \cap \Theta^{-1}(D)) = \mu(\Theta^{-1}(A) \cap D) = \mu\Big|_D(\Theta^{-1}(A))$$

Note that $\{S^\star > 0\} \subseteq D$ as we restrict our measure space to D, but by the previous inequality, $S^\star > 0$ on D. So $D = \{S^\star > 0\}$. Then the maximal ergodic lemma gives

$$0 \leq \int_{S^\star > 0} g \, \mathrm{d}\mu = \int_D g \, \mathrm{d}\mu = \int_D f \, \mathrm{d}\mu - b\mu(B)$$

Hence, $b\mu(B) \leq \int_D f \, \mathrm{d}\mu$. By σ-finiteness, this inequality extends to D; one can choose an approximating sequence $B_n \uparrow D$ where $\mu(B_n) < \infty$, then take limits to show $b\mu(D) = b \lim_n \mu(B_n) \leq \int_D f \, \mathrm{d}\mu$. Repeating the above argument for $-f$ and $-a$, we obtain $-a\mu(D) \leq \int_D -f \, \mathrm{d}\mu$. Combining these two inequalities gives

$$b\mu(D) \leq \int_D f \, \mathrm{d}\mu \leq a\mu(D)$$

But $a < b$, so $\mu(D) = 0$ or ∞, but f is integrable, so $\mu(D) = 0$. Now, define

$$\Delta = \left\{ \liminf_n \frac{S_n(f)}{n} < \limsup_n \frac{S_n(f)}{n} \right\} = \bigcup_{a<b\in\mathbb{Q}} D_{a,b}$$

By countable additivity,

$$\mu(\Delta) = \mu\left(\bigcup_{a<b\in\mathbb{Q}} D_{a,b} \right) = \sum_{a<b\in\mathbb{Q}} \mu(D_{a,b}) = 0$$

On Δ^c, $\frac{S_n}{n}$ converges in $[-\infty, \infty]$. We define the invariant function $\overline{f}$ by

$$\overline{f} = \begin{cases} \lim_n \frac{S_n}{n} & x \in \Delta^c \\ 0 & x \in \Delta \end{cases}$$

so $\frac{S_n}{f} \to \overline{f}$ almost everywhere as $n \to \infty$. Since $\mu(|f \circ \Theta^{n-1}|) = \mu(|f|)$, we have $\mu(|S_n|) \leq n\mu(|f|)$ and thus

$$\mu\left(\left|\overline{f}\right|\right) = \mu\left(\liminf_n \left|\frac{S_n}{n}\right|\right) \leq \liminf_n \mu\left(\left|\frac{S_n}{n}\right|\right) \leq \mu(|f|)$$

which is one of the results required by the theorem. In particular, $\mu\left(\left|\overline{f}\right|\right) < \infty$ so $\left|\overline{f}\right| < \infty$ almost everywhere. □

Theorem (von Neumann)**.** Let $(E, \mathcal{E}, \mu)$ be a finite measure space (not σ-finite). Let $\Theta : E \to E$ be measure-preserving. Let $f \in L^p(E)$ with $1 \leq p < \infty$. Then $\frac{S_n(f)}{n} \to \overline{f}$ in L^p.

Proof. Since Θ is measure-preserving, we have

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

Thus, by Minkowski's inequality, for all $f \in L^p$ we have

$$\left\| \frac{S_n(f)}{n} \right\|_p \leq \frac{1}{n} \sum_{i=0}^{n-1} \left\| f \circ \Theta^i \right\|_p = \|f\|_p$$

So $\frac{S_n(f)}{n}$ is a contraction in L^p. For each $K > 0$, we define $f_K = \max(\min(f, K), -K)$. Then

$$\|f - f_K\|_p^p = \int_E |f - f_K|^p \mathbb{1}_{|f|>K} \, \mathrm{d}\mu$$

Since $\mathbb{1}_{|f|>K}$ converges to zero pointwise, and $|f - f_K| \leq 2|f|^p \in L^1$, we find $\|f - f_K\|_p < \frac{\varepsilon}{3}$ by dominated convergence, for sufficiently large $K = K_\varepsilon$. As $|f_K| \leq K$, we have $\left|\frac{S_n(f_K)}{n}\right| \leq K$. Since μ is finite, $f_K \in L^1(\mu)$, so by Birkhoff's ergodic theorem, $\frac{S_n(f_K)}{n} \to \overline{f}_K$ almost everywhere for some invariant function $\overline{f}_K$. Note that $\overline{f}_k$ is bounded by K as $\frac{S_n(f_K)}{n}$ is bounded by K. By the bounded convergence theorem, we deduce that $\left\| \frac{S_n(f_K)}{n} - \overline{f}_K \right\| \to 0$ as $n \to \infty$. Further, this holds in L^p since

$$\left\| \frac{S_n(f_K)}{n} - \overline{f}_K \right\|_p \leq (2K)^{\frac{p-1}{p}} \left\| \frac{S_n(f_K)}{n} - \overline{f}_K \right\|_1 < \frac{\varepsilon}{3}$$

where the last inequality holds for sufficiently large n. Since μ is a finite measure, $L^p(\mu) \subseteq L^1(\mu)$, hence by Birkhoff's ergodic theorem, $\frac{S_n(f)}{n} \to \overline{f}$ almost everywhere as $f \to \infty$. Then, by the contraction property applied to $f - f_K$,

$$\begin{aligned}
\left\| \overline{f} - \overline{f}_K \right\|_p^p &= \int_E \left| \overline{f} - \overline{f}_K \right|^p \mathrm{d}\mu \\
&= \int_E \liminf_n \left| \frac{S_n(f) - S_n(f_K)}{n} \right|^p \mathrm{d}\mu \\
&\leq \liminf_n \int_E \left| \frac{S_n(f) - S_n(f_K)}{n} \right|^p \mathrm{d}\mu \\
&= \liminf_n \int_E \left| \frac{S_n(f - f_K)}{n} \right|^p \mathrm{d}\mu \\
&\leq \liminf_n \|f - f_K\|_p^p \\
&= \|f - f_K\|_p^p < \left(\frac{\varepsilon}{3}\right)^p
\end{aligned}$$

So in particular, $\overline{f} \in L^p$. Then by the triangle inequality,

$$\begin{aligned}
\left\| \frac{S_n(f)}{n} - \overline{f} \right\|_p &\leq \left\| \frac{S_n(f) - S_n(f_K)}{n} \right\|_p + \left\| \frac{S_n(f_K)}{n} - \overline{f}_K \right\|_p + \left\| \overline{f} - \overline{f}_K \right\|_p \\
&< \left\| \frac{S_n(f) - S_n(f_K)}{n} \right\|_p + \frac{2\varepsilon}{3} \\
&\leq \|f - f_K\|_p + \frac{2\varepsilon}{3} = \varepsilon
\end{aligned}$$

for sufficiently large n. □

7.4 Infinite product spaces

Let $E = \mathbb{R}^{\mathbb{N}} = \{x = (x_n)_{n\in\mathbb{N}}\}$ be the space of real sequences. Consider

$$\mathcal{C} = \left\{ A = \prod_{n=1}^{\infty} A_n \,\middle|\, A_n \in \mathcal{B}, \exists N \in \mathbb{N}, \forall n > N, A_n = \mathbb{R} \right\}$$

This forms a π-system, which generates the *cylindrical σ-algebra* $\sigma(\mathcal{C})$. One shows that $\sigma(\mathcal{C}) = \sigma(\{f_n \mid n \in \mathbb{N}\})$ where $f_n(x) = x_n$ are the coordinate projection functions on E. We can also show $\sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R}^{\mathbb{N}})$ for the product topology. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with marginal distributions $\mu_{X_n} = m$ for all n; this exists by an earlier theorem. We define a map $X : \Omega \to E$ by $X(\omega)_n = X_n(\omega)$. This is $\mathcal{F}$–$\sigma(\mathcal{C})$ measurable, since for all $A \in \mathcal{C}$, we have

$$X^{-1}(A) = \{\omega \mid X_1(\omega) \in A_1, \dots, X_N(\omega) \in A_n\} = \bigcap_{n=1}^{N} X_n^{-1}(A_n) \in \mathcal{F}$$

We denote $\mu = \mathbb{P} \circ X^{-1}$, which is the unique product probability measure in $\mathbb{R}^{\mathbb{N}}$ satisfying

$$\begin{aligned}
\mu\left(\prod_{n=1}^{\infty} A_n\right) &= \lim_{N\to\infty} \mu\left(\prod_{n=1}^{N} A_n\right) \\
&= \lim_{N\to\infty} \mathbb{P}(X_1 \in A_1, \dots, X_N \in A_N) \\
&= \lim_{N\to\infty} \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_N \in A_N) \\
&= \prod_{n=1}^{\infty} \mathbb{P}(X_n \in A_n) \\
&= \prod_{n=1}^{\infty} m(A_n)
\end{aligned}$$

Note that we need to use finiteness of N to exploit independence of the X_i. We call $(E, \mathcal{E}, \mu) = (\mathbb{R}^{\mathbb{N}}, \sigma(\mathcal{C}), m^{\mathbb{N}})$ the *canonical model* for an infinite sequence of random variables of law m.

Theorem. The shift map $\Theta : E \to E$ defined by $\Theta(x)_n = x_{n+1}$ is measure preserving and ergodic.

Proof. For $A \in \mathcal{C}$,

$$\begin{aligned}
\mu(A) &= \mathbb{P}(X_1 \in A_1, \dots, X_N \in A_N) \\
&= \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_N \in A_N) \\
&= \prod_{n=1}^{N} m(A_n) \\
&= \mathbb{P}(X_2 \in A_1) \cdots \mathbb{P}(X_{N+1} \in A_N) \\
&= \mu(\Theta^{-1}(A))
\end{aligned}$$

so Θ is measure-preserving. Recall that the tail σ-algebra is defined by $\mathcal{T} = \bigcap_n \mathcal{T}_n$ where $\mathcal{T}_n = \sigma(\{X_k \mid k \geq n+1\})$. Note that for all $A \in \mathcal{C}$, we have

$$\Theta^{-n}(A) = \left\{x \in \mathbb{R}^{\mathbb{N}} \mid (x_{n+1}, x_{n+2}, \dots) \in A\right\} \in \mathcal{T}_n$$

Now, if A is invariant, $A = \Theta^{-n}(A) \in \mathcal{T}_n$ for all n, so $A \in \mathcal{T}$. By Kolmogorov's zero-one law, $\mu(A) = 0$ or $\mu(A) = 1$ as required for ergodicity. □

We can apply Birkhoff's ergodic theorem to Θ. If $f \in L^1(\mu)$, then $\frac{S_n(f)}{n} \to \overline{f} \in L^1(\mu)$ almost surely. Since $\overline{f}$ is invariant and Θ is ergodic, $\overline{f}$ is almost surely constant. By von Neumann's L^p-ergodic theorem, convergence holds in fact in L^1.

7.5 Strong law of large numbers

Theorem. Let $\int_{\mathbb{R}} |x| \, \mathrm{d}m(x) < \infty$, and let $\int_{\mathbb{R}} x \, \mathrm{d}m(x) = \nu$. Then

$$\mu\left(\left\{x \in \mathbb{R}^{\mathbb{N}} \,\middle|\, \frac{x_1 + x_2 + \cdots + x_n}{n} \to \nu\right\}\right) = 1$$

Proof. Let $f(x) = x_1$. Then $f \in L^1(\mu)$, since $\int_E |f| \, \mathrm{d}\mu = \int_{\mathbb{R}} |x| \, \mathrm{d}m(x) < \infty$. So by Birkhoff's ergodic theorem,

$$\mu\left(\left\{\frac{x_1 + \cdots + x_n}{n} \to \nu\right\}\right) = \mu\left(\left\{\frac{S_n(f)}{n} \to \overline{f}\right\}\right) = 1$$

where we also use von Neumann's ergodic theorem to deduce that

$$\overline{f} = \mu(\overline{f}) = \lim_n \mu\left(\frac{S_n(f)}{n}\right) = \frac{n}{n}\nu = \nu$$

□

Theorem (strong law of large numbers). Let $(X_n)_{n \in \mathbb{N}}$ be independent and identically distributed random variables such that $\mathbb{E}[|X_1|] < \infty$. Then $\frac{1}{n}\sum_{i=1}^n X_i \to \mathbb{E}[X]$ almost surely.

Proof. Inject X from Ω to $E = \mathbb{R}^{\mathbb{N}}$ as before, and notice that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \to \mathbb{E}[X]\right) = \mu\left(\left\{x \,\middle|\, \frac{x_1 + \cdots + x_n}{n} \to \nu\right\}\right) = 1$$

□

Remark. The hypothesis $\mathbb{E}[|X|] < \infty$ cannot be weakened; we see on an example sheet that $\frac{1}{n}\sum_{i=1}^n X_i$ can exhibit various behaviours. Note that this notion of convergence is stronger than the weak convergence seen in the central limit theorem. The law of the iterated logarithm is that

$$\limsup_n \frac{X_1 + \cdots + X_n}{\sqrt{2n \log\log n}} = 1$$

almost surely, and -1 for the limit inferior. In particular, the central limit theorem does not hold almost surely.

Corollary. By von Neumann's ergodic theorem, in the strong law of large numbers, we have $\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right|\right] \to 0$ as $n \to \infty$.