

UNIVERSITY OF
CAMBRIDGE

MATHEMATICS TRIPOS

Part II

Probability and Measure

Michaelmas, 2018

Lectures by

E. BREUILLARD

Notes by

QIANGRU KUANG

Contents

1	Lebesgue measure	2
1.1	Boolean algebra	2
1.2	Jordan measure	2
1.3	Lebesgue measure	4
2	Abstract measure theory	9
3	Integration and measurable functions	17
4	Product measures	25
5	Foundations of probability theory	28
6	Independence	32
6.1	Useful inequalities	36
7	Convergence of random variables	38
8	L^p spaces	43
9	Hilbert space and L^2-methods	48
10	Fourier transform	55
11	Gaussians	63
12	Ergodic theory	66
12.1	The canonical model	67
	Index	73

1 Lebesgue measure

1.1 Boolean algebra

Definition (Boolean algebra). Let X be a set. A *Boolean algebra* on X is a family of subsets of X which

1. contains \emptyset ,
2. is stable under finite unions and complementation.

Example.

- The *trivial Boolean algebra* $\mathcal{B} = \{\emptyset, X\}$.
- The *discrete Boolean algebra* $\mathcal{B} = 2^X$, the family of all subsets of X .
- Less trivially, if X is a topological space, the family of *constructible sets* forms a Boolean algebra, where a constructible set is the finite union of locally closed set, i.e. a set $E = U \cap F$ where U is open and F is closed.

Definition (finitely additive measure). Let X be a set and \mathcal{B} a Boolean algebra on X . A *finitely additive measure* on (X, \mathcal{B}) is a function $m : \mathcal{B} \rightarrow [0, +\infty]$ such that

1. $m(\emptyset) = 0$,
2. $m(E \cup F) = m(E) + m(F)$ where $E \cap F = \emptyset$.

Example.

1. Counting measure: $m(E) = \#E$, the cardinality of E where \mathcal{B} is the discrete Boolean algebra of X .
2. More generally, given $f : X \rightarrow [0, +\infty]$, define for $E \subseteq X$,

$$m(E) = \sum_{e \in E} f(e).$$

3. Suppose $X = \coprod_{i=1}^N X_i$, then define $\mathcal{B}(X)$ to be the unions of X_i 's. Assign a weight $a_i \geq 0$ to each X_i and define $m(E) = \sum_{i: X_i \subseteq E} a_i$ for $E \in \mathcal{B}$.

1.2 Jordan measure

This section is a historic review and provides intuition for Lebesgue measure theory. We'll gloss over details of proofs in this section.

Definition. A subset of \mathbb{R}^d is called *elementary* if it is a finite union of *boxes*, where a box is a set $B = I_1 \times \cdots \times I_d$ where each I_i is a finite interval of \mathbb{R} .

Proposition 1.1. Let $B \subseteq \mathbb{R}^d$ be a box. Let $\mathcal{E}(B)$ be the family of elementary subsets of B . Then

1. $\mathcal{E}(B)$ is a Boolean algebra on B ,
2. every $E \in \mathcal{E}(B)$ is a disjoint finite union of boxes,
3. if $E \in \mathcal{E}(B)$ can be written as disjoint finite union in two ways, $E = \bigcup_{i=1}^n B_i = \bigcup_{j=1}^m B'_j$, then $\sum_{i=1}^n |B_i| = \sum_{j=1}^m |B'_j|$ where $|B| = \prod_{i=1}^d |b_i - a_i|$ if $B = I_1 \times \dots \times I_d$ and I_i has endpoints a_i, b_i .

Following this, we can define a finitely additive measure corresponding to our intuition of length, area, volume etc:

Proposition 1.2. Define $m(E) = \sum_{i=1}^n |B_i|$ if E is any elementary set and is the disjoint union of boxes $B_i \subseteq \mathbb{R}^d$. Then m is a finitely additive measure on $\mathcal{E}(B)$ for any box B .

Definition. A subset $E \subseteq \mathbb{R}^d$ is Jordan measurable if for any $\varepsilon > 0$ there are elementary sets A, B , $A \subseteq E \subseteq B$ and $m(B \setminus A) < \varepsilon$.

Remark. Jordan measurable sets are bounded.

Proposition 1.3. If a set $E \subseteq \mathbb{R}^d$ is Jordan measurable, then

$$\sup_{A \subseteq E \text{ elementary}} \{m(A)\} = \inf_{B \supseteq E \text{ elementary}} \{m(B)\}.$$

In which case we define the Jordan measure of E as

$$m(E) = \sup_{A \subseteq E} \{m(A)\}.$$

Proof. Take $A_n \subseteq E$ such that $m(A_n) \uparrow \sup$ and $B_n \supseteq E$ such that $m(B_n) \downarrow \inf$. Note that

$$\inf \leq m(B_n) = m(A_n) + m(B_n \setminus A_n) \leq \sup + m(B_n \setminus A_n) \leq \sup + \varepsilon$$

for arbitrary $\varepsilon > 0$ so they are equal. □

Exercise.

1. If B is a box, the family $\mathcal{J}(B)$ of Jordan measurable subsets of B is a Boolean algebra.
2. A subset $E \subseteq [0, 1]$ is Jordan measurable if and only if $\mathbf{1}_E$, the indicator function on E , is Riemann integrable.

1.3 Lebesgue measure

Although Jordan measure corresponds to the intuition of length, area and volume, it suffers from a few severe problems and issues:

1. unbounded sets in \mathbb{R}^d are not Jordan measurable.
2. $\mathbf{1}_{\mathbb{Q} \cap [0,1]}$ is not Riemann integrable, and therefore $\mathbb{Q} \cap [0,1]$ is not Jordan measurable.
3. pointwise limits of Riemann integrable function $f_n := \mathbf{1}_{\frac{1}{n}\mathbb{Z} \cap [0,1]} \rightarrow \mathbf{1}_{\mathbb{Q} \cap [0,1]}$ is not Riemann integrable.

The idea of Lebesgue is to use countable covers by boxes.

Definition. A subset $E \subseteq \mathbb{R}^d$ is *Lebesgue measurable* if for all $\varepsilon > 0$, there exists a countable union of boxes C with $E \subseteq C$ and $m^*(C \setminus E) < \varepsilon$, where m^* , the *Lebesgue outer measure*, is defined as

$$m^*(E) = \inf \left\{ \sum_{i \geq 1} |B_i| : E \subseteq \bigcup_{i \geq 1} B_i, B_i \text{ boxes} \right\}$$

for every subset $E \subseteq \mathbb{R}^d$.

Remark. wlog in these definitions we may assume that boxes are open.

Proposition 1.4. *The family \mathcal{L} of Lebesgue measurable subsets of \mathbb{R}^d is a Boolean algebra stable under countable unions.*

Lemma 1.5.

1. m^* is monotone: if $E \subseteq F$ then $m^*(E) \leq m^*(F)$.
2. m^* is countably subadditive: if $E = \bigcup_{n \geq 1} E_n$ where $E_n \subseteq \mathbb{R}^d$ then

$$m^*(E) \leq \sum_{n \geq 1} m^*(E_n).$$

Proof. Monotonicity is obvious. For countable subadditivity, pick $\varepsilon > 0$ and let $C_n = \bigcup_{i \geq 1} C_{n,i}$ where $C_{n,i}$ are boxes such that $E_n \subseteq C_n$ and

$$\sum_{i \geq 1} |C_{n,i}| \leq m^*(E_n) + \frac{\varepsilon}{2^n}.$$

Then

$$\sum_{n \geq 1} \sum_{i \geq 1} |C_{n,i}| \leq \sum_{n \geq 1} (m^*(E_n) + \frac{\varepsilon}{2^n}) = \varepsilon + \sum_{n \geq 1} m^*(E_n)$$

and $E \subseteq \bigcup_{n \geq 1} C_n = \bigcup_{n \geq 1} \bigcup_{i \geq 1} C_{n,i}$ so

$$m^*(E) \leq \varepsilon + \sum_{n \geq 1} m^*(E_n)$$

for all $\varepsilon > 0$. □

Remark. Note that m^* is *not* additive on the family of all subsets of \mathbb{R}^d . However, it will be on \mathcal{L} , as we will show later.

Lemma 1.6. *If A, B are disjoint compact subsets of \mathbb{R}^d then*

$$m^*(A \cup B) = m^*(A) + m^*(B).$$

Proof. \leq by the previous lemma so need to show \geq . Pick $\varepsilon > 0$. Let $A \cup B \subseteq \bigcup_{n \geq 1} B_n$ where B_n are open boxes such that

$$\sum_{n \geq 1} |B_n| \leq m^*(A \cup B) + \varepsilon.$$

wlog we may assume that the side lengths of each B_n are $< \frac{\alpha}{2}$, where

$$\alpha = \inf\{\|x - y\|_1 : x \in A, y \in B\} > 0.$$

where the inequality comes from the fact that A and B are compact and thus closed. wlog we may discard the B_n 's that do not intersect $A \cup B$. Then by construction

$$\sum_{n \geq 1} |B_n| = \sum_{n \geq 1, B_n \cap A = \emptyset} |B_n| + \sum_{n \geq 1, B_n \cap B = \emptyset} |B_n| \geq m^*(A) + m^*(B)$$

so

$$\varepsilon + m^*(A \cup B) \geq m^*(A) + m^*(B)$$

for all ε . □

Lemma 1.7. *If $E \subseteq \mathbb{R}^d$ has $m^*(E) = 0$ then $E \in \mathcal{L}$.*

Definition (null set). A set $E \subseteq \mathbb{R}^d$ such that $m^*(E) = 0$ is called a *null set*.

Proof. For all $\varepsilon > 0$, there exist $C = \bigcup_{n \geq 1} B_n$ where B_n are boxes such that $E \subseteq C$ and $\sum_{n \geq 1} |B_n| \leq \varepsilon$. But

$$m^*(C \setminus E) \leq m^*(C) \leq \varepsilon.$$

□

Lemma 1.8. *Every open subset of \mathbb{R}^d and every closed subset of \mathbb{R}^d is in \mathcal{L} .*

We will prove the lemma using the fact that the family of Lebesgue measurable subsets is stable under countable union, which itself *does not* use this lemma. This lemma, however, will be used to show the stability under complementation. Since the proof is quite technical (it has more to do with general topology than measure theory), for brevity and fluency of ideas we present the proof the main proposition first.

Proof of Proposition 1.4. It is obvious that $\emptyset \in \mathcal{L}$. To show it is stable under countable unions, start with $E_n \in \mathcal{L}$ for $n \geq 1$. Need to show $E := \bigcup_{n \geq 1} E_n \in \mathcal{L}$.

Pick $\varepsilon > 0$. By assumption there exist $C_n = \bigcup_{i \geq 1} B_{n,i}$ where $B_{n,i}$ are boxes such that $E_n \subseteq C_n$ and

$$m^*(C_n \setminus E_n) < \frac{\varepsilon}{2^n}.$$

Now

$$E = \bigcup_{n \geq 1} E_n \subseteq \bigcup_{n \geq 1} C_n =: C$$

so C is again a countable union of boxes and $C \setminus E \subseteq \bigcup_{n \geq 1} C_n \setminus E_n$. so

$$m^*(C \setminus E) \leq \sum_{n \geq 1} m^*(C_n \setminus E_n) \leq \sum_{n \geq 1} \frac{\varepsilon}{2^n} = \varepsilon$$

by countable subadditivity so $E \in \mathcal{L}$.

To show it is stable under complementation, suppose $E \in \mathcal{L}$. By assumption there exist C_n a countable union of boxes with $E \subseteq C_n$ and $m^*(C_n \setminus E) \leq \frac{1}{n}$. wlog we may assume the boxes are open so C_n is open, C_n^c is closed so $C_n^c \in \mathcal{L}$. Thus $\bigcup_{n \geq 1} C_n^c \in \mathcal{L}$ by first part of the proof.

But

$$m^*(E^c \setminus \bigcup_{n \geq 1} C_n^c) \leq m^*(E^c \setminus C_n^c) = m^*(C_n \setminus E) \leq \frac{1}{n}$$

so $m^*(E^c \setminus \bigcup_{n \geq 1} C_n^c) = 0$ so $E^c \setminus \bigcup_{n \geq 1} C_n^c \in \mathcal{L}$ since it is a null set. But

$$E^c = (E^c \setminus \bigcup_{n \geq 1} C_n^c) \cup \bigcup_{n \geq 1} C_n^c,$$

both of which are in \mathcal{L} so $E^c \in \mathcal{L}$. □

Proof of Lemma 1.8. Every open set in \mathbb{R}^d is a countable union of boxes so is in \mathcal{L} . It is more subtle for closed sets. The key observation is that every closed set is the countable union of compact subsets so we are left to show compact sets of \mathbb{R}^d are in \mathcal{L} .

Let $F \subseteq \mathbb{R}^d$ be compact. For all $k \geq 1$, there exist O_k a countable union of open sets such that $F \subseteq O_k := \bigcup_{i \geq 1} O_{k,i}$ where $O_{k,i}$ are open boxes such that

$$\sum_{i \geq 1} |O_{k,i}| \leq m^*(F) + \frac{1}{2^k}.$$

By compactness there exist a finite subcover so we can assume O_k is a finite union of open boxes. Moreover, wlog assume that

1. the side lengths of $O_{k,i}$ are $\leq \frac{1}{2^k}$.
2. for each i , $O_{k,i}$ intersects F .
3. $O_{k+1} \subseteq O_k$ (by replacing O_{k+1} with $O_{k+1} \cap O_k$ iteratively).

Then $F = \bigcap_{k \geq 1} O_k$ and we are left to show $m^*(O_k \setminus F) \rightarrow 0$. By additivity on disjoint compact sets,

$$m^*(F) + m^*(\overline{O}_i \setminus O_{i+1}) = m^*(F \cup (\overline{O}_i \setminus O_{i+1}))$$

so

$$m^*(F) + m^*(\overline{O}_i \setminus O_{i+1}) \leq m^*(\overline{O}_i) \leq \sum_{j \geq 1} |O_{i,j}| \leq m^*(F) + \frac{1}{2^i}$$

so $m^*(\overline{O}_i \setminus O_{i+1}) \leq \frac{1}{2^i}$. Finally,

$$m^*(O_k \setminus F) = m^*\left(\bigcup_{i \geq k} (O_i \setminus O_{i+1})\right) \leq \sum_{i \geq k} m^*(O_i \setminus O_{i+1}) \leq \sum_{i \geq k} \frac{1}{2^i} = \frac{1}{2^{k-1}}.$$

□

The result we're working towards is

Proposition 1.9. *m^* is countably additive on \mathcal{L} , i.e. if $(E_n)_{n \geq 1}$ where $E_n \in \mathcal{L}$ are pairwise disjoint then*

$$m^*\left(\bigcup_{n \geq 1} E_n\right) = \sum_{n \geq 1} m^*(E_n).$$

Lemma 1.10. *If $E \in \mathcal{L}$ then for all $\varepsilon > 0$ there exists U open, F closed, $F \subseteq E \subseteq U$ such that $m^*(U \setminus E) < \varepsilon$ and $m^*(E \setminus F) < \varepsilon$.*

Proof. By definition of \mathcal{L} , there exists a countable union of open boxes $E \subseteq \bigcup_{n \geq 1} B_n$ such that $m^*(\bigcup_{n \geq 1} B_n \setminus E) < \varepsilon$. Just take $U = \bigcup_{n \geq 1} B_n$ which is open.

For F do the same with $E^c = \mathbb{R}^d \setminus E$ in place of E . □

Proof of Proposition 1.9. First we assume each E_n is compact. By a previous lemma m^* is additive on compact sets so for all $N \in \mathbb{N}$,

$$m^*\left(\bigcup_{n=1}^N E_n\right) = \sum_{n=1}^N m^*(E_n).$$

In particular

$$\sum_{n=1}^N m^*(E_n) \leq m^*\left(\bigcup_{n \geq 1} E_n\right)$$

since m^* is monotone. Take $N \rightarrow \infty$ to get one inequality. The other direction holds by countable subadditivity of m^* .

Now assume that each E_n is a bounded subset in \mathcal{L} . By the lemma there exists $K_n \subseteq E_n$ closed, so compact, such that $m^*(E_n \setminus K_n) \leq \frac{\varepsilon}{2^n}$. Since K_n 's are disjoint, by the previous case

$$m^*\left(\bigcup_{n \geq 1} K_n\right) = \sum_{n \geq 1} m^*(K_n)$$

then

$$\begin{aligned}
 & \sum_{n \geq 1} m^*(E_n) \\
 & \leq \sum_{n \geq 1} m^*(K_n) + m^*(E_n \setminus K_n) \\
 & \leq m^*\left(\bigcup_{n \geq 1} K_n\right) + \sum_{n \geq 1} \frac{\varepsilon}{2^n} \\
 & \leq m^*\left(\bigcup_{n \geq 1} E_n\right) + \varepsilon
 \end{aligned}$$

so one direction of inequality. Similarly the other direction holds by countable subadditivity of m^* .

For the general case, note that $\mathbb{R}^d = \bigcup_{n \in \mathbb{Z}^d} A_n$ where A_n is bounded and in \mathcal{L} , for example by taking A_n to be product of half open intervals of unit length. Write E_n as $\bigcup_{m \in \mathbb{Z}^d} E_n \cap A_m$ so just apply the previous results to $(E_n \cap A_m)_{n \geq 1, m \in \mathbb{Z}^d}$. \square

Definition (Lebesgue measure). m^* when restricted to \mathcal{L} is called the *Lebesgue measure* and is simply denoted by m .

Example (Vitali counterexample). Although \mathcal{L} is pretty big (it includes all open and closed sets, countable unions and intersections of them, and has cardinality at least $2^{\mathfrak{c}}$ where \mathfrak{c} is the continuum, by considering a null set with cardinality \mathfrak{c} , and each subset thereof), it does not include every subset of \mathbb{R}^d .

Consider $(\mathbb{Q}, +)$, the additive subgroup of $(\mathbb{R}, +)$. Pick a set of representative E of the cosets of $(\mathbb{Q}, +)$. Choose it inside $[0, 1]$. For each $x \in \mathbb{R}$, there exists a unique $e \in E$ such that $x - e \in \mathbb{Q}$ (here we require axiom of choice). Claim that $E \notin \mathcal{L}$ and m^* is not additive on the family of all subsets of \mathbb{R}^d .

Proof. Pick distinct rationals p_1, \dots, p_N in $[0, 1]$. The sets $p_i + E$ are pairwise disjoint so if m^* were additive then we would have

$$m^*\left(\bigcup_{i=1}^N p_i + E\right) = \sum_{i=1}^N m^*(p_i + E) = Nm^*(E)$$

by translation invariance of m^* . But then

$$\bigcup_{i=1}^N p_i + E \subseteq [0, 2]$$

since $E \subseteq [0, 1]$ so by monotonicity of m^* have

$$m^*\left(\bigcup_{i=1}^N p_i + E\right) \leq 2$$

so for all $Nm^*(E) \leq 2$ so $m^*(E) = 0$. But

$$[0, 1] \subseteq \bigcup_{q \in \mathbb{Q}} E + q = \mathbb{R},$$

by countable subadditivity of m^* ,

$$1 = m^*([0, 1]) \leq \sum_{q \in \mathbb{Q}} m^*(E + q) = 0.$$

Absurd.

In particular $E \notin \mathcal{L}$ as m^* is additive on \mathcal{L} .

□

2 Abstract measure theory

In this chapter we extend measure theory to arbitrary set. Most part of the theory is developed by Fréchet and Carathéodory.

Definition (σ -algebra). A σ -algebra on a set X is a Boolean algebra stable under countable unions.

Definition (measurable space). A *measurable space* is a couple (X, \mathcal{A}) where X is a set and \mathcal{A} is a σ -algebra on X .

Definition (measure). A *measure* on (X, \mathcal{A}) is a map $\mu : \mathcal{A} \rightarrow [0, \infty]$ such that

1. $\mu(\emptyset) = 0$,
2. μ is countably additive (also known as σ -additive), i.e. for every family $(A_n)_{n \geq 1}$ of disjoint subsets in \mathcal{A} , have

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

The triple (X, \mathcal{A}, μ) is called a measure space.

Example.

1. $(\mathbb{R}^d, \mathcal{L}, m)$ is a measure space.
2. $(X, 2^X, \#)$ where $\#$ is the counting measure.

Proposition 2.1. Let (X, \mathcal{A}, μ) be a measure space. Then

1. μ is monotone: $A \subseteq B$ implies $\mu(A) \leq \mu(B)$,
2. μ is countably subadditive: $\mu(\bigcup_{n \geq 1} A_n) \leq \sum_{n \geq 1} \mu(A_n)$,
3. upward monotone convergence: if

$$E_1 \subseteq E_2 \subseteq \dots \subseteq E_n \subseteq \dots$$

then

$$\mu\left(\bigcup_{n \geq 1} E_n\right) = \lim_{n \rightarrow \infty} \mu(E_n) = \sup_{n \geq 1} \mu(E_n).$$

4. downward monotone convergence: if

$$E_1 \supseteq E_2 \supseteq \dots \supseteq E_n \supseteq \dots$$

and $\mu(E_1) < \infty$ then

$$\mu\left(\bigcap_{n \geq 1} E_n\right) = \lim_{n \rightarrow \infty} \mu(E_n) = \inf_{n \geq 1} \mu(E_n).$$

Proof.

1.

$$\mu(B) = \mu(A) + \underbrace{\mu(B \setminus A)}_{\geq 0}$$

by additivity of μ .

2. See example sheet. The idea is that every countable union $\bigcup_{n \geq 1} A_n$ is a disjoint countable union $\bigcup_{n \geq 1} B_n$ where for each n , $B_n \subseteq A_n$. It then follows by σ -additivity.

3. Let $E_0 = \emptyset$ so

$$\bigcup_{n \geq 1} E_n = \bigcup_{n \geq 1} (E_n \setminus E_{n-1}),$$

a disjoint union. By σ -additivity,

$$\mu\left(\bigcup_{n \geq 1} E_n\right) = \sum_{n \geq 1} \mu(E_n \setminus E_{n-1})$$

but for all N , by additivity of μ ,

$$\sum_{n=1}^N \mu(E_n \setminus E_{n-1}) = \mu(E_N)$$

so take limit. The supremum part is obvious.

4. Apply the previous result to $E_1 \setminus E_n$.

□

Remark. Note the $\mu(E_1) < \infty$ condition in the last part. Counterexample: $E_n = [n, \infty) \subseteq \mathbb{R}$.

Definition (σ -algebra generated by a family). Let X be a set and \mathcal{F} be some family of subsets of X . The intersection of all σ -algebras on X containing \mathcal{F} is a σ -algebra, called the σ -algebra *generated* by \mathcal{F} and is denoted by $\sigma(\mathcal{F})$.

Proof. Easy check. See example sheet.

□

Example.

1. Suppose $X = \coprod_{i=1}^N X_i$, i.e. X admits a finite partition. Let $\mathcal{F} = \{X_1, \dots, X_n\}$, then $\sigma(\mathcal{F})$ consists of all subsets that are unions of X_i 's.
2. Suppose X is countable and let \mathcal{F} be the collection of all singletons. Then $\sigma(\mathcal{F}) = 2^X$.

Definition (Borel σ -algebra). Let X be a topological space. The σ -algebra generated by open subsets of X is called the *Borel σ -algebra* of X , denoted by $\mathcal{B}(X)$.

Proposition 2.2. *If $X = \mathbb{R}^d$ then $\mathcal{B}(X) \subseteq \mathcal{L}$. Moreover every $A \in \mathcal{L}$ can be written as a disjoint union $A = B \cup N$ where $B \in \mathcal{B}(X)$ and N is a null set.*

Proof. We've shown that \mathcal{L} is a σ -algebra and contains all open sets so $\mathcal{B}(X) \subseteq \mathcal{L}$. Given $A \in \mathcal{L}$, $A^c \in \mathcal{L}$ so for all $n \geq 1$ there exists C_n countable unions of (open) boxes such that $A^c \subseteq C_n$ and $m^*(C_n \setminus A^c) \leq \frac{1}{n}$. Take $C = \bigcap_{n \geq 1} C_n \in \mathcal{B}(X)$. Thus $B := C^c \in \mathcal{B}(X)$ and $m(A \setminus B) = 0$ because $A \setminus B = C \setminus A^c$. \square

Remark.

1. It can be shown that $\mathcal{B}(\mathbb{R}^d) \subsetneq \mathcal{L}$. In fact $|\mathcal{L}| \geq 2^{\mathfrak{c}}$ and $|\mathcal{B}(\mathbb{R}^d)| = \mathfrak{c}$.
2. If \mathcal{F} is a family of subsets of a set X , the Boolean algebra generated by \mathcal{F} can be explicitly described as

$$\mathcal{B}(\mathcal{F}) = \{\text{finite unions of } F_1 \cap \dots \cap F_N : F_i \in \mathcal{F} \text{ or } F_i^c \in \mathcal{F}\}.$$

3. However, this is not so for $\sigma(\mathcal{F})$. There is no “simple” description of σ -algebra generated by \mathcal{F} . (c.f. Borel hierarchy in descriptive set theory and transfinite induction)

Definition (π -system). A family \mathcal{F} of subsets of a set X is called a *π -system* if it contains \emptyset and it is closed under finite intersection.

Proposition 2.3 (measure uniqueness). *Let (X, \mathcal{A}) be a measurable space. Assume μ_1 and μ_2 are two finite measures (i.e. $\mu_i(X) < \infty$) such that $\mu_1(F) = \mu_2(F)$ for every $F \in \mathcal{F}$ where \mathcal{F} is a π -system with $\sigma(\mathcal{F}) = \mathcal{A}$. Then $\mu_1 = \mu_2$.*

For \mathbb{R}^d , we only have to check open boxes.

Proof. We state first the following lemma:

Lemma 2.4 (Dynkin lemma). *If \mathcal{F} is a π -system on X and \mathcal{C} is a family of subsets of X such that $\mathcal{F} \subseteq \mathcal{C}$ and \mathcal{C} is stable under complementation and disjoint countable unions. Then $\sigma(\mathcal{F}) \subseteq \mathcal{C}$.*

Let $\mathcal{C} = \{A \in \mathcal{A} : \mu_1(A) = \mu_2(A)\}$. Then \mathcal{C} is clearly stable under complementation as

$$\mu_i(A^c) = \mu_i(X \setminus A) = \mu_i(X) - \mu_i(A).$$

\mathcal{C} is also clearly stable under countable disjoint unions by σ -additivity. Thus by Dynkin lemma, $\mathcal{C} \supseteq \sigma(\mathcal{F}) = \mathcal{A}$.

Proof of Dynkin lemma. Let \mathcal{M} be the smallest family of subsets of X containing \mathcal{F} and stable under complementation and countable disjoint union (2^X is such a family and taking intersection). Sufficient to show that \mathcal{M} is a σ -algebra, as then $\mathcal{M} \subseteq \mathcal{C}$ implies $\sigma(\mathcal{F}) \subseteq \mathcal{C}$.

It suffices to show \mathcal{M} is a Boolean algebra. Let

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

\mathcal{M}' again is stable under countable disjoint unions and complementation because

$$A^c \cap B = (B^c \cup (A \cap B))^c$$

as a disjoint union so is in \mathcal{M} .

As $\mathcal{M}' \supseteq \mathcal{F}$, by minimality of \mathcal{M} , have $\mathcal{M} = \mathcal{M}'$. Now let

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

The same argument shows that $\mathcal{M}'' = \mathcal{M}$. Thus \mathcal{M} is a Boolean algebra and a σ -algebra. □

□

Proposition 2.5 (uniqueness of Lebesgue measure). *Lebesgue measure is the unique translation invariant measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that*

$$\mu([0, 1]^d) = 1.$$

Proof. Exercise. Hint: use the π -system \mathcal{F} made of all boxes in \mathbb{R}^d and dissect a cube into dyadic pieces. Then approximate and use monotone. □

Remark.

1. There is *no* countably additive translation invariant measure on \mathbb{R} defined on all subsets of \mathbb{R} . (c.f. Vitali's counterexample).
2. However, the Lebesgue measure can be extended to a finitely additive measure on all subsets of \mathbb{R} (proof requires Hahn-Banach theorem. See IID Linear Analysis).

Recall the construction of Lebesgue measure: we take boxes in \mathbb{R}^d , and define elementary sets, which is the Boolean algebra generated by boxes. Then we can define Jordan measure which is finitely additive. However, this is not countably additive but analysis craves limits so we define Lebesgue measurable sets, by introducing the outer measure m^* , which is built from the Jordan measure. Finally we restrict this outer measure to \mathcal{L} . We also define the Borel σ -algebra, which is the same as the σ -algebra generated by the boxes. We show that the Borel σ -algebra is contained in \mathcal{L} , and every element in \mathcal{L} can be written as a disjoint union of an element in the Borel σ -algebra and a measure zero set.

Suppose \mathcal{B} is a Boolean algebra on a set X . Let μ be a finitely additive measure on \mathcal{B} . We are going to construct a measure on $\sigma(\mathcal{B})$.

Theorem 2.6 (Carathéodory extension theorem). *Assume that μ is countably additive on \mathcal{B} , i.e. if $B_n \in \mathcal{B}$ disjoint is such that $\bigcup_{n \geq 1} B_n \in \mathcal{B}$ then $\mu(\bigcup_{n \geq 1} B_n) = \sum_{n \geq 1} \mu(B_n)$ and assume that μ is σ -finite, i.e. there exists $X_m \in \mathcal{B}$ such that $X = \bigcup_{m \geq 1} X_m$ and $\mu(X_m) < \infty$, then μ extends uniquely to a measure on $\sigma(\mathcal{B})$.*

Proof. For any $E \subseteq X$, let

$$\mu^*(E) = \inf \left\{ \sum_{n \geq 1} \mu(B_n) : E \subseteq \bigcup_{n \geq 1} B_n, B_n \in \mathcal{B} \right\}$$

and call it the outer measure associated to μ . Define a subset $E \subseteq X$ to be μ^* -measurable if for all $\varepsilon > 0$ there exists $C = \bigcup_{n \geq 1} B_n$ with $B_n \in \mathcal{B}$ such that $E \subseteq C$ and

$$\mu^*(C \setminus E) \leq \varepsilon.$$

We denote by \mathcal{B}^* the set of μ^* -measurable subsets. Claim that

1. μ^* is countably subadditive and monotone.
2. $\mu^*(B) = \mu(B)$ for all $B \in \mathcal{B}$.
3. \mathcal{B}^* is a σ -algebra and contains all μ^* -null sets and \mathcal{B} .
4. μ^* is σ -additive on \mathcal{B}^* .

Then existence follows from the proposition as $\mathcal{B}^* \supseteq \sigma(\mathcal{B})$: μ^* will be a measure on \mathcal{B}^* and thus on $\sigma(\mathcal{B})$. Uniqueness follows from a similar proof for Lebesgue measure via Dynkin lemma.

Proof. This will be very easy as we only need to adapt our previous work to the general case. Note that in a few occasion we used properties of \mathbb{R}^d , such as openness of some sets, so be careful.

1. Same.
2. $\mu^*(B) \leq \mu(B)$ for all $B \in \mathcal{B}$ by definition of μ^* . For the other direction, for all $\varepsilon > 0$, there exist $B_n \in \mathcal{B}$ such that $B \subseteq \bigcup_{n \geq 1} B_n$ and $\sum_{n \geq 1} \mu(B_n) \leq \mu^*(B) + \varepsilon$. But

$$B = \bigcup_{n \geq 1} B_n \cap B = \bigcup_{n \geq 1} C_n$$

where $C_n := B_n \cap B \setminus \bigcup_{i < n} B_i \cap B$ and so $C_n \in \mathcal{B}$. Thus by countable additivity

$$\mu(B) = \sum_{n \geq 1} \mu(C_n) \leq \sum_{n \geq 1} \mu(B_n) \leq \mu^*(B) + \varepsilon$$

3. μ^* -null sets and \mathcal{B} are obviously in \mathcal{B}^* . Thus it is left to show that \mathcal{B}^* is a σ -algebra. Stability under countable union is exactly the same and then we claim that \mathcal{B}^* is stable under complementation. This is the bit where we used closed/open sets in \mathbb{R}^d in the original proof. Here we use a lemma as a substitute.

| **Lemma 2.7.** Suppose $B_n \in \mathcal{B}$ then $\bigcap_{n \geq 1} B_n \in \mathcal{B}^*$.

Proof. First claim that if $E = \bigcap_{n \geq 1} I_n$ where $I_{n+1} \subseteq I_n$ and $I_n \in \mathcal{B}$ such that $\mu(I_1) < \infty$ then $\mu^*(E) = \lim_{n \rightarrow \infty} \mu(I_n)$ and $E \in \mathcal{B}^*$: by additivity of μ on \mathcal{B} ,

$$\sum_{n=1}^N \mu(I_n \setminus I_{n+1}) = \mu(I_1) - \mu(I_N)$$

which converges as $N \rightarrow \infty$ (because $\mu(I_{n+1}) \leq \mu(I_n)$), so

$$\sum_{n \geq N} \mu(I_n \setminus I_{n+1}) \rightarrow 0$$

as $N \rightarrow \infty$. But LHS is greater than $\mu^*(I_N \setminus E)$ because $I_N \setminus E = \bigcup_{n \geq N} I_n \setminus I_{n+1}$. Therefore $E \in \mathcal{B}^*$ and

$$\mu(I_n) \leq \underbrace{\mu^*(I_n \setminus E)}_{\rightarrow 0} + \underbrace{\mu^*(E)}_{\leq \mu(I_n)}$$

so

$$\lim_{n \rightarrow \infty} \mu(I_n) = \mu^*(E).$$

Now for the actual lemma, let $E = \bigcap_{n \geq 1} I_n$ where $I_n \in \mathcal{B}$. wlog we may assume $I_{n+1} \subseteq I_n$. By σ -finiteness assumption, $X = \bigcup_{m \geq 1} X_m$ where $X_m \in \mathcal{B}$ with $\mu(X_m) < \infty$ so

$$E = \bigcup_{m \geq 1} E \cap X_m.$$

By the claim for all m , $E \cap X_m \in \mathcal{B}^*$ so $E \in \mathcal{B}^*$. □

From the lemma we can derive that \mathcal{B}^* is also stable under complementation: given $E \in \mathcal{B}^*$, for all n there exist $C_n = \bigcup_{i \geq 1} B_{n,i}$ where $B_{n,i} \in \mathcal{B}$ such that $E \subseteq C_n$ and $\mu^*(C_n \setminus E) \leq \frac{1}{n}$. Now

$$E^c = \left(\bigcup_{n \geq 1} C_n^c \right) \cup \left(E^c \setminus \bigcup_{n \geq 1} C_n^c \right)$$

but C_n^c is a countable intersection $\bigcap_{i \geq 1} B_{n,i}^c$ and $E^c \setminus \bigcup_{n \geq 1} C_n^c$ is μ^* -null so by the lemma, $C_n^c \in \mathcal{B}^*$. Therefore their union is also in \mathcal{B}^* . Since we've shown that null sets are in \mathcal{B}^* , $E^c \in \mathcal{B}^*$.

4. We want to show μ^* is countably additive on \mathcal{B}^* . Recall that μ is σ -finite: there exists $X_m \in \mathcal{B}$ such that $X = \bigcup_{m \geq 1} X_m$, $\mu(X_m) < \infty$. We say $E \subseteq X$ is *bounded* if there exists m such that $E \subseteq X_m$. It is then enough to show countable additivity for bounded sets by the same argument as before: write $X = \bigcup_{m \geq 1} \tilde{X}_m$ where $\tilde{X}_m = X_m \setminus \bigcup_{i < m} X_i \in \mathcal{B}$ so this is a disjoint union. Then if $E = \bigcup_{n \geq 1} E_n$ as a disjoint union then

$$E = \bigcup_{n \geq 1} \bigcup_{m \geq 1} (E_n \cap \tilde{X}_m)$$

which is also a countable disjoint union.

Given E , if we can show finite additivity then

$$\sum_{n=1}^N \mu^*(E_n) = \mu^*\left(\bigcup_{n=1}^N E_n\right) \leq \mu^*(E) \leq \sum_{n \geq 1} \mu^*(E_n)$$

take limit as $N \rightarrow \infty$ to have equality throughout.

It suffices to prove finite additivity when E and F are countable intersections of sets from \mathcal{B} : $E, F \in \mathcal{B}^*$ so for $\varepsilon > 0$ there exists C, D countable intersections of sets from \mathcal{B} such that $C \subseteq E, D \subseteq F$ and

$$\begin{aligned} \mu^*(E) &\leq \mu^*(C) + \varepsilon \\ \mu^*(F) &\leq \mu^*(D) + \varepsilon \end{aligned}$$

As $E \cap F = \emptyset$ and $C \subseteq E, D \subseteq F, C \cap D = \emptyset$ so by finite additivity,

$$\mu^*(E) + \mu^*(F) \leq 2\varepsilon + \mu^*(C \cup D) \leq 2\varepsilon + \mu^*(E \cup F).$$

As usual, reverse holds by subadditivity.

Finally for $E = \bigcap_{n \geq 1} I_n, F = \bigcap_{n \geq 1} J_n$ bounded, wlog assume $I_{n+1} \subseteq I_n, J_{n+1} \subseteq J_n$. $\mu(I_n), \mu(J_n) < \infty$. Now use claim 3,

$$\begin{aligned} \mu^*(E) &= \lim_{n \rightarrow \infty} \mu^*(I_n) \\ \mu^*(F) &= \lim_{n \rightarrow \infty} \mu^*(J_n) \end{aligned}$$

so

$$\mu^*(E) + \mu^*(F) = \lim_{n \rightarrow \infty} \mu(I_n) + \mu(J_n) = \lim_{n \rightarrow \infty} (\mu(I_n \cup J_n) + \mu(I_n \cap J_n))$$

But

$$\begin{aligned} \bigcap_{n \geq 1} (I_n \cap J_n) &= E \cap F = \emptyset \\ \bigcap_{n \geq 1} (I_n \cup J_n) &= E \cup F \end{aligned}$$

so by claim 3

$$\begin{aligned} \lim_{n \rightarrow \infty} \mu(I_n \cap J_n) &= 0 \\ \lim_{n \rightarrow \infty} \mu(I_n \cup J_n) &= \mu^*(E \cup F) \end{aligned}$$

which finishes the proof.

□

□

Remark. We prove that every set in \mathcal{B}^* is a disjoint union $E = F \cup N$ where $F \in \sigma(\mathcal{B})$ and N is μ^* -null.

Definition (completion). We say that \mathcal{B}^* is the *completion* of $\sigma(\mathcal{B})$ with respect to μ .

Example. \mathcal{L} is the completion of $\mathcal{B}(\mathbb{R}^d)$ in \mathbb{R}^d .

3 Integration and measurable functions

Definition (measurable function). Let (X, \mathcal{A}) be a measurable space. A function $X \rightarrow \mathbb{R}$ is called *measurable* or *\mathcal{A} -measurable* if for all $t \in \mathbb{R}$,

$$\{x \in X : f(x) < t\} \in \mathcal{A}.$$

Remark. The σ -algebra generated by intervals $(-\infty, t)$ where $t \in \mathbb{R}$ is the Borel σ -algebra of \mathbb{R} , denote $\mathcal{B}(\mathbb{R})$. Thus for every measurable function $f : X \rightarrow \mathbb{R}$, the preimage $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}(\mathbb{R})$. However, it is *not* true that $f^{-1}(L) \in \mathcal{A}$ for any $L \in \mathcal{L}$.

Remark. If f is allowed to take the values $+\infty$ and $-\infty$ we will say that f is measurable if additionally $f^{-1}(\{+\infty\}) \in \mathcal{A}$ and $f^{-1}(\{-\infty\}) \in \mathcal{A}$.

More generally,

Definition (measurable map). Suppose (X, \mathcal{A}) and (Y, \mathcal{B}) are measurable spaces. A map $f : X \rightarrow Y$ is *measurable* if for all $B \in \mathcal{B}$, $f^{-1}(B) \in \mathcal{A}$.

Proposition 3.1.

1. The composition of measurable maps is measurable.
2. If $f, g : (X, \mathcal{A}) \rightarrow \mathbb{R}$ are measurable functions then $f + g, fg$ and λf for $\lambda \in \mathbb{R}$ are also measurable.
3. If $(f_n)_{n \geq 1}$ is a sequence of measurable functions on (X, \mathcal{A}) then so are $\sup_n f_n, \inf_n f_n, \limsup_n f_n$ and $\liminf_n f_n$.

Proof.

1. Obvious.
2. Follow from 1 once it's shown that $+$: $\mathbb{R}^2 \rightarrow \mathbb{R}$ and \times : $\mathbb{R}^2 \rightarrow \mathbb{R}$ are measurable (with respect to Borel sets). The sets

$$\begin{aligned} \{(x, y) : x + y < t\} \\ \{(x, y) : xy < t\} \end{aligned}$$

are open in \mathbb{R}^2 and hence Borel.

3. $\inf_n f_n(x) < t$ if and only if

$$x \in \bigcup_n \{x : f_n(x) < t\}$$

and similar for sup. Similarly $\limsup_n f_n(x) < t$ if and only if

$$x \in \bigcup_{m \geq 1} \bigcap_{k \geq 1} \bigcup_{n \geq k} \{x : f_n(x) < t - \frac{1}{m}\}.$$

□

Proposition 3.2. $f = (f_1, \dots, f_d) : (X, \mathcal{A}) \rightarrow (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ where $d \geq 1$ is measurable if and only if each $f_i : X \rightarrow \mathbb{R}$ is measurable.

Proof. One direction is easy: suppose f is measurable then

$$\{x : f_i(x) < t\} = f^{-1}(\{y \in \mathbb{R}^d : y_i < t\}),$$

which is open so f_i is measurable.

Conversely, suppose f_i is measurable. Then

$$f^{-1}\left(\prod_{i=1}^d [a_i, b_i]\right) = \bigcap_{i=1}^d \{x : a_i \leq f_i(x) \leq b_i\}$$

As the boxes generate the Borel sets, done. □

Example.

1. Let (X, \mathcal{A}) be a measurable space and $E \subseteq X$. Then $E \in \mathcal{A}$ if and only if $\mathbf{1}_E$, the indicator function on E , is \mathcal{A} -measurable.
2. If $X = \coprod_{i=1}^N X_i$ and \mathcal{A} is the Boolean algebra generated by the X_i 's. A function $f : (X, \mathcal{A}) \rightarrow \mathbb{R}$ is measurable if and only if f is constant on each X_i . In this case the vector space of measurable functions has dimension N .
3. Every continuous function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is measurable.

Definition (Borel measurable). If X is a topological space, $f : X \rightarrow \mathbb{R}$ is *Borel* or *Borel measurable* if it is $\mathcal{B}(X)$ -measurable.

Definition (simple function). A function f on (X, \mathcal{A}) is called *simple* if

$$f = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$$

for some $a_i \geq 0$ and $A_i \in \mathcal{A}$.

Of course simple functions are measurable.

Lemma 3.3. *If a simple function can be written in two ways*

$$f = \sum_{i=1}^n a_i \mathbf{1}_{A_i} = \sum_{j=1}^s b_j \mathbf{1}_{B_j}$$

then

$$\sum_{i=1}^n a_i \mu(A_i) = \sum_{j=1}^s b_j \mu(B_j)$$

for any measure μ on (X, \mathcal{A}) .

Proof. Example sheet 1. □

Definition (integral of a simple function with respect to a measure). The μ -integral of f is defined by

$$\mu(f) := \sum_{i=1}^n a_i \mu(A_i).$$

Remark.

1. The lemma says that the integral is well-defined.
2. We also use the notation $\int_X f d\mu$ to denote $\mu(f)$.

Proposition 3.4. μ -integral satisfies, for all simple functions f and g ,

1. *linearity:* for all $\alpha, \beta \geq 0$, $\mu(\alpha f + \beta g) = \alpha \mu(f) + \beta \mu(g)$.
2. *positivity:* if $g \leq f$ then $\mu(g) \leq \mu(f)$.
3. if $\mu(f) = 0$ then $f = 0$ μ -almost everywhere, i.e. $\{x \in X : f(x) \neq 0\}$ is a μ -null set.

Proof. Obvious from definition and lemma. □

Definition. If $f \geq 0$ and measurable on (X, \mathcal{A}) , define

$$\mu(f) = \sup\{\mu(g) : g \text{ simple}, g \leq f\} \in [0, +\infty].$$

Remark. This is consistent with the definition for f simple, due to positivity.

Definition (integrable). If f is an arbitrary measurable function on (X, \mathcal{A}) we say f is μ -integrable if

$$\mu(|f|) < \infty.$$

Definition (integral with respect to a measure). If f is μ -integrable, then we define its μ -integral by

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

where $f^+ = \max\{0, f\}$ and $f^- = (-f)^+$.

Note.

$$\begin{aligned} |f| &= f^+ + f^- \\ f &= f^+ - f^- \end{aligned}$$

Theorem 3.5 (monotone convergence theorem). Let $(f_n)_{n \geq 1}$ be a sequence of measurable functions on a measure space (X, \mathcal{A}, μ) such that

$$0 \leq f_1 \leq f_2 \leq \dots \leq f_n \leq \dots$$

Let $f = \lim_{n \rightarrow \infty} f_n$. Then

$$\mu(f) = \lim_{n \rightarrow \infty} \mu(f_n).$$

Lemma 3.6. If g is a simple function on (X, \mathcal{A}, μ) , the map

$$\begin{aligned} m_g : \mathcal{A} &\rightarrow [0, \infty] \\ E &\mapsto \mu(\mathbf{1}_E g) \end{aligned}$$

is a measure on (X, \mathcal{A}) .

Proof. Write $g = \sum_{i=1}^r a_i \mathbf{1}_{A_i}$ so $g \mathbf{1}_E = \sum_{i=1}^r a_i \mathbf{1}_{A_i \cap E}$ so

$$\mu(\mathbf{1}_E g) = \sum_{i=1}^r a_i \mu(A_i \cap E).$$

By a question on example sheet this is well-defined. Then σ -additivity follows immediately from σ -additivity of μ . \square

Proof of monotone convergence theorem. $f_n \leq f_{n+1} \leq f$ by assumption so

$$\mu(f_n) \leq \mu(f_{n+1}) \leq \mu(f)$$

by definition of integral so

$$\lim_{n \rightarrow \infty} \mu(f_n) \leq \mu(f),$$

although RHS may be infinite.

Let g be any simple function with $g \leq f$. Need to show that $\mu(g) \leq \lim_{n \rightarrow \infty} \mu(f_n)$. Pick $\varepsilon > 0$. Let

$$E_n = \{x \in X : f_n(x) \geq (1 - \varepsilon)g(x)\}.$$

Then $X = \bigcup_{n \geq 1} E_n$ and $E_n \subseteq E_{n+1}$. So we may apply upward monotone convergence for sets to measure m_g and get

$$\lim_{n \rightarrow \infty} m_g(E_n) = m_g(X) = \mu(g \mathbf{1}_X) = \mu(g).$$

But

$$(1 - \varepsilon)m_g(E_n) = \mu((1 - \varepsilon)g \mathbf{1}_{E_n}) \leq \mu(f_n)$$

because $(1 - \varepsilon)g \mathbf{1}_{E_n}$ is a simple function smaller than f_n . Taking limit,

$$(1 - \varepsilon)\mu(g) \leq \lim_{n \rightarrow \infty} \mu(f_n)$$

which holds for all ε . So

$$\mu(g) \leq \lim_{n \rightarrow \infty} \mu(f_n).$$

\square

Lemma 3.7. *If $f \geq 0$ is a measurable function on (X, \mathcal{A}) then there is a sequence of simple functions $(g_n)_{n \geq 1}$*

$$0 \leq g_n \leq g_{n+1} \leq f$$

such that for all $x \in X$, $g_n(x) \uparrow f(x)$.

Notation. $g_n \uparrow f$ means that $\lim_{n \rightarrow \infty} g_n(x) = f(x)$ and $g_{n+1} \geq g_n$.

Proof. We can take

$$g_n = \frac{1}{2^n} \lfloor 2^n \min\{f, n\} \rfloor$$

pointwise. Check that $\lfloor 2y \rfloor \geq 2\lfloor y \rfloor$ for all $y \geq 0$. □

Proposition 3.8. *Basic properties of the integral (for positive functions): suppose $f, g \geq 0$ are measurable on (X, \mathcal{A}, μ) .*

1. *linearity: for all $\alpha, \beta \geq 0$, $\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g)$.*
2. *positivity: if $0 \leq f \leq g$ then $\mu(f) \leq \mu(g)$.*
3. *if $\mu(f) = 0$ then $f = 0$ μ -almost everywhere.*
4. *if $f = g$ μ -almost everywhere then $\mu(f) = \mu(g)$.*

Proof.

1. Follows from the same property for simple functions and from Lemma 3.7 combined with monotone convergence theorem.
2. Obvious from definition.
- 3.

$$\{x \in X : f(x) \neq 0\} = \bigcup_{n \geq 0} \{x \in X : f(x) > \frac{1}{n}\}$$

set $g_n = \frac{1}{n} \mathbf{1}_{\{x \in X : f(x) > 1/n\}}$ which is simple and $g_n \leq f$ so by definition of integral $\mu(g_n) \leq \mu(f)$ so $\mu(g_n) = 0$, i.e. $\mu(\{x : f(x) > \frac{1}{n}\}) = 0$.

4. Note that if $E \in \mathcal{A}$, $\mu(E^c) = 0$ then

$$\mu(h \mathbf{1}_E) = \mu(h)$$

for all h simple. Thus it holds for all $h \geq 0$ measurable. Now take $E = \{x : f(x) = g(x)\}$. □

Proposition 3.9 (linearity of integral). *Suppose f, g are μ -integrable functions and $\alpha, \beta \in \mathbb{R}$. Then $\alpha f + \beta g$ is μ -integrable and*

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

Proof. We have shown the case when $\alpha, \beta \geq 0$ and $f, g \geq 0$. In the general case, use the positive and negative parts. □

Lemma 3.10 (Fatou's lemma). *Suppose $(f_n)_{n \geq 1}$ is a sequence of measurable functions on (X, \mathcal{A}, μ) such that $f_n \geq 0$ for all n . Then*

$$\mu(\liminf_{n \rightarrow \infty} f_n) \leq \liminf_{n \rightarrow \infty} \mu(f_n).$$

Remark. We may not have equality: let $f_n = \mathbf{1}_{[n, n+1]}$ on $(\mathbb{R}, \mathcal{L}, m)$. Then $\mu(f_n) = 1$ but $\lim_{n \rightarrow \infty} f_n = 0$.

Proof. Let $g_n := \inf_{k \geq n} f_k$. Then $g_{n+1} \geq g_n \geq 0$ so by monotone convergence theorem, $\mu(g_n) \uparrow \mu(g)$ as $n \rightarrow \infty$ where $g = \lim_{n \rightarrow \infty} g_n = \liminf_{n \rightarrow \infty} f_n$ and $g_n \leq f_n$ so $\mu(g_n) \leq \mu(f_n)$ for all n . Take $n \rightarrow \infty$,

$$\mu(g) \leq \liminf_{n \rightarrow \infty} \mu(f_n).$$

□

In both monotone convergence theorem and Fatou's lemma we assumed that the sequence of functions is nonnegative. There is another version of convergence theorem where we replace nonnegativity by domination:

Theorem 3.11 (Lebesgue's dominated convergence theorem). *Let $(f_n)_{n \geq 1}$ be a sequence of measurable functions on (X, \mathcal{A}, μ) and g a μ -integrable function on X . Assume $|f_n| \leq g$ for all n (domination assumption) and assume for all $x \in X$, $\lim_{n \rightarrow \infty} f_n(x) = f(x)$. Then f is μ -integrable and*

$$\mu(f) = \lim_{n \rightarrow \infty} \mu(f_n).$$

This allows us to swap limit and integral.

Proof. $|f_n| \leq g$ so $|f| \leq g$ so $\mu(|f|) \leq \mu(g) < \infty$ and f is integrable. Note that $g + f_n \geq 0$ so by Fatou's lemma,

$$\mu(\liminf_{n \rightarrow \infty} (g + f_n)) \leq \liminf_{n \rightarrow \infty} \mu(g + f_n).$$

But $\liminf_{n \rightarrow \infty} (g + f_n) = g + f$ and by linearity $\mu(g + f_n) = \mu(g) + \mu(f_n)$, so

$$\mu(g) + \mu(f) \leq \mu(g) + \liminf_{n \rightarrow \infty} \mu(f_n),$$

i.e.

$$\mu(f) \leq \liminf_{n \rightarrow \infty} \mu(f_n).$$

Do the same with $g - f_n$ in place of $g + f_n$, get

$$\mu(-f) \leq \liminf_{n \rightarrow \infty} \mu(-f_n) = - \limsup_{n \rightarrow \infty} \mu(f_n)$$

so

$$\mu(f) = \lim_{n \rightarrow \infty} \mu(f_n).$$

□

Corollary 3.12 (exchanging integral and summation). *Let (X, \mathcal{A}, μ) be a measure space and let $(f_n)_{n \geq 1}$ be a sequence of measurable functions on X .*

1. *If $f_n \geq 0$ then*

$$\mu\left(\sum_{n \geq 1} f_n\right) = \sum_{n \geq 1} \mu(f_n).$$

2. *If $\sum_{n \geq 1} |f_n|$ is μ -integrable then $\sum_{n \geq 1} f_n$ is μ -integrable and*

$$\mu\left(\sum_{n \geq 1} f_n\right) = \sum_{n \geq 1} \mu(f_n).$$

Proof.

1. Let $g_N = \sum_{n=1}^N f_n$, then $g_N \uparrow \sum_{n \geq 1} f_n$ as $N \rightarrow \infty$ so the result follows from monotone convergence theorem.
2. Let $g = \sum_{n \geq 1} |f_n|$ and g_N as above. Then $|g_N| \leq g$ for all N so the domination assumption holds. The result thus follows from dominated convergence theorem.

□

Corollary 3.13 (differentiation under integral sign). *Let (X, \mathcal{A}, μ) be a measure space. Let $U \subseteq \mathbb{R}$ be an open set and let $f : U \times X \rightarrow \mathbb{R}$ be such that*

1. *$x \mapsto f(t, x)$ is μ -integrable for all $t \in U$,*
2. *$t \mapsto f(t, x)$ is differentiable for all $x \in X$,*
3. *domination: there exists $g : X \rightarrow \mathbb{R}$ μ -integrable such that for all $t \in U, x \in X$,*

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

Then $x \mapsto \frac{\partial f}{\partial t}(t, x)$ is μ -integrable for all $t \in U$ and if we set $F(t) = \int_X f(t, x) d\mu$ then F is differentiable and

$$F'(t) = \int_X \frac{\partial f}{\partial t}(t, x) d\mu.$$

Proof. Pick $h_n > 0, h_n \rightarrow 0$ and define

$$g_n(t, x) := \frac{1}{h_n}(f(t + h_n, x) - f(t, x)).$$

Then

$$\lim_{n \rightarrow \infty} g_n(t, x) = \frac{\partial f}{\partial t}(t, x).$$

By mean value theorem, there exists $\theta_{t, n, x} \in [t, t + h_n]$ such that

$$g_n(t, x) = \frac{\partial f}{\partial t}(\theta, x)$$

so

$$|g_n(t, x)| \leq g(x)$$

by domination assumption. Now apply dominated convergence theorem. \square

Remark.

1. If $f : [a, b] \rightarrow \mathbb{R}$ is continuous where $a < b$ in \mathbb{R} , then f is m -integrable (where m is the Lebesgue measure) and $m(f) = \int_a^b f(x)dx$ is the Riemann integral. In general if f is only assumed to be bounded, then f will be Riemann integrable if and only if the points of discontinuity of f is an m -null set. See example sheet 2.
2. If $g \in \text{GL}_d(\mathbb{R})$ and $f \geq 0$ is Borel measurable on \mathbb{R}^d , then

$$m(f \circ g) = \frac{1}{|\det g|} m(f).$$

See example sheet 2. In particular m is invariant under linear transformation whose determinant has absolute value 1, e.g. rotation.

Remark. In each of monotone convergence theorem, Fatou's lemma and dominated convergence theorem, we can replace pointwise assumption by the corresponding μ -almost everywhere. The same conclusion holds. Indeed, let

$$E = \{x \in X : \text{assumptions hold at } x\}$$

so E^c is a μ -null set. Replace each f_n (and similarly g etc) by $\mathbf{1}_E f_n$. Then assumptions then hold everywhere as $\mu(f \mathbf{1}_E) = \mu(f)$ for all f measurable.

4 Product measures

Definition (product σ -algebra). Let (X, \mathcal{A}) and (Y, \mathcal{B}) be measurable spaces. The σ -algebra of subsets of $X \times Y$ generated by the product sets $E \times F$ where $E \in \mathcal{A}, F \in \mathcal{B}$ is called the *product σ -algebra* of \mathcal{A} and \mathcal{B} and is denoted by $\mathcal{A} \otimes \mathcal{B}$.

Remark.

1. By analogy with the notion of product topology, $\mathcal{A} \otimes \mathcal{B}$ is the smallest σ -algebra of subsets of $X \times Y$ making the two projection maps measurable.
2. $\mathcal{B}(\mathbb{R}^{d_1}) \otimes \mathcal{B}(\mathbb{R}^{d_2}) = \mathcal{B}(\mathbb{R}^{d_1+d_2})$. See example sheet. However this is not so for $\mathcal{L}(\mathbb{R}^d)$.

Lemma 4.1. *If $E \subseteq X \times Y$ is $\mathcal{A} \otimes \mathcal{B}$ -measurable then for all $x \in X$, the slice*

$$E_x = \{y \in Y : (x, y) \in E\}$$

is in \mathcal{B} .

Proof. Let

$$\mathcal{E} = \{E \subseteq X \times Y : E_x \in \mathcal{B} \text{ for all } x \in X\}.$$

Note that \mathcal{E} contains all product sets $A \times B$ where $A \in \mathcal{A}, B \in \mathcal{B}$. \mathcal{E} is a σ -algebra: if $E \in \mathcal{E}$ then $E^c \in \mathcal{E}$ and if $E_n \in \mathcal{E}$ then $\bigcup E_n \in \mathcal{E}$ since $(E^c)_x = (E_x)^c$ and $(\bigcup E_n)_x = \bigcup (E_n)_x$. \square

Lemma 4.2. *Assume (X, \mathcal{A}, μ) and (Y, \mathcal{B}, ν) are σ -finite measure spaces. Let $f : X \times Y \rightarrow [0, +\infty]$ be $\mathcal{A} \otimes \mathcal{B}$ -measurable. Then*

1. *for all $x \in X$, the function $y \mapsto f(x, y)$ is \mathcal{B} -measurable.*
2. *for all $x \in X$, the map $x \mapsto \int_Y f(x, y) d\nu(y)$ is \mathcal{A} -measurable.*

Proof.

1. In case $f = \mathbf{1}_E$ for $E \in \mathcal{A} \otimes \mathcal{B}$ the function $y \mapsto f(x, y)$ is just $y \mapsto \mathbf{1}_{E_x}(y)$, which is measurable by the previous lemma.
More generally, the result is true for simple functions and thus for all measurable functions by taking pointwise limit.
2. By the same reduction we may assume $f = \mathbf{1}_E$ for some $E \in \mathcal{A} \otimes \mathcal{B}$. Now let $Y = \bigcup_{m \geq 1} Y_m$ with $\nu(Y_m) < \infty$. Let

$$\mathcal{E} = \{E \in \mathcal{A} \otimes \mathcal{B} : x \mapsto \nu(E_x \cap Y_m) \text{ is } \mathcal{A}\text{-measurable for all } m\}.$$

\mathcal{E} contains all product sets $E = A \times B$ where $A \in \mathcal{A}, B \in \mathcal{B}$ because $\nu(E_x \cap Y_m) = \mathbf{1}_{x \in A} \nu(B \cap Y_m)$. \mathcal{E} is stable under complementation:

$$\nu((E^c)_x \cap Y_m) = \nu(Y_m) - \nu(Y_m \cap E_x)$$

where LHS is ν -measurable. \mathcal{E} is stable under disjoint countable union: let $E = \bigcup_{n \geq 1} E_n$ where $E_n \in \mathcal{E}$ disjoint. Then by σ -additivity

$$\nu(E_x \cap Y_m) = \sum_{n \geq 1} \nu((E_n)_x \cap Y_m)$$

which is \mathcal{A} -measurable.

The product sets form a π -system and generates the product measure so by Dynkin lemma $\mathcal{E} = \mathcal{A} \otimes \mathcal{B}$. □

Definition (product measure). Let (X, \mathcal{A}, μ) and (Y, \mathcal{B}, ν) be measure spaces and μ, ν σ -finite. Then there exists a unique *product measure*, denoted by $\mu \otimes \nu$, on $\mathcal{A} \otimes \mathcal{B}$ such that for all $A \in \mathcal{A}, B \in \mathcal{B}$,

$$(\mu \otimes \nu)(A \times B) = \mu(A)\nu(B).$$

Proof. Uniqueness follows from Dynkin lemma. For existence, set

$$\sigma(E) = \int_X \nu(E_x) d\mu(x).$$

σ is well-defined because $x \mapsto \nu(E_x)$ is \mathcal{A} -measurable by lemma 2. σ is countably-additive: suppose $E = \bigcup_{n \geq 1} E_n$ where $E_n \in \mathcal{A} \otimes \mathcal{B}$ disjoint, then

$$\sigma(E) = \int_X \nu(E_x) d\mu(x) = \int_X \sum_{n \geq 1} \nu((E_n)_x) d\mu(x) = \sum_{n \geq 1} \int_X \nu((E_n)_x) d\mu(x) = \sum_{n \geq 1} \sigma(E_n)$$

by a corollary of MCT. □

Theorem 4.3 (Tonelli-Fubini). *Let (X, \mathcal{A}, μ) and (Y, \mathcal{B}, ν) be σ -finite measure spaces.*

1. *Let $f : X \times Y \rightarrow [0, +\infty]$ be $\mathcal{A} \otimes \mathcal{B}$ -measurable. Then*

$$\int_{X \times Y} f(x, y) d(\mu \otimes \nu) = \int_X \int_Y f(x, y) d\nu(y) d\mu(x) = \int_X \int_X f(x, y) d\mu(x) d\nu(y).$$

2. *If $f : X \times Y \rightarrow \mathbb{R}$ is $\mu \otimes \nu$ -integrable then for μ -almost everywhere x , $y \mapsto f(x, y)$ is ν -integrable, and for ν -almost everywhere y , $x \mapsto f(x, y)$ is μ -integrable and*

$$\int_{X \times Y} f(x, y) d(\mu \otimes \nu) = \int_X \int_Y f(x, y) d\nu(y) d\mu(x) = \int_X \int_X f(x, y) d\mu(x) d\nu(y).$$

Without the nonnegativity or integrability assumption, the result is false in general. For example for $X = Y = \mathbb{N}$, let $\mathcal{A} = \mathcal{B}$ be discrete σ -algebras and

$\mu = \nu$ counting measure. Let $f(n, m) = \mathbf{1}_{n=m} - \mathbf{1}_{n=m+1}$. Check that

$$\sum_{n \geq 1} f(n, m) = 0$$
$$\sum_{m \geq 1} f(n, m) = \begin{cases} 0 & n \geq 2 \\ 1 & n = 1 \end{cases}$$

so

$$\sum_{n \geq 1} \sum_{m \geq 1} f(n, m) \neq \sum_{m \geq 1} \sum_{n \geq 1} f(n, m).$$

Proof.

1. The result holds for $f = \mathbf{1}_E$ where $E \in \mathcal{A} \otimes \mathcal{B}$ by the definition of product measure and lemma 2, so it holds for all simple functions. Now take limits and apply MCT.
2. Write $f = f^+ - f^-$ and apply 1.

□

Note.

1. The Lebesgue measure m_d on \mathbb{R}^d is equal to $m_1 \otimes \cdots \otimes m_1$, because it is true on boxes and extend by uniqueness of measure.
2. $E \in \mathcal{A} \otimes \mathcal{B}$ if is $\mu \otimes \nu$ -null if and only if for μ -almost every x , $\nu(E_x) = 0$.

5 Foundations of probability theory

Modern probability theory was founded by Kolmogorov, who formulated the axioms of probability theory in 1933 in his thesis *Foundations on the Theory of Probability*. He defined a probability space to be a measure space $(\Omega, \mathcal{F}, \mathbb{P})$. The interpretation is as follow: Ω is the universe of possible *outcomes*. However, we wouldn't be able to assign probability to every single outcome unless the space is discrete. Instead we are interested in studying some subsets of Ω , which are called *events* and contained in \mathcal{F} . Finally \mathbb{P} is a probability measure with $\mathbb{P}(\Omega) = 1$. Thus for $A \in \mathcal{F}$, $\mathbb{P}(A \text{ occurs}) \in [0, 1]$. Thus finite additivity of \mathbb{P} says that if A and B never occurs simultaneously then $\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B)$. σ -additivity is slightly more difficult to justify and it is perhaps to see the equivalent notion *continuity*: if $A_{n+1} \supseteq A_n$ and $\bigcap_{n \geq 1} A_n = \emptyset$ then $\mathbb{P}(A_n) \rightarrow 0$ as $n \rightarrow \infty$.

Definition (probability measure, probability space). Let Ω be a set and \mathcal{F} a σ -algebra on Ω . A measure μ on (Ω, \mathcal{F}) is called a *probability measure* if $\mu(\Omega) = 1$ and the measure space $(\Omega, \mathcal{F}, \mu)$ is called a *probability space*.

Definition (random variable). A measurable function $X : \Omega \rightarrow \mathbb{R}$ is called a *random variable*.

We usually use a capital letter to denote a random variable.

Definition (expectation). If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space then the \mathbb{P} -integral is called *expectation*, denoted \mathbb{E} .

Definition (distribution/law). A random variable $X : \Omega \rightarrow \mathbb{R}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ determines a Borel measure μ_X on \mathbb{R} defined by

$$\mu_X((-\infty, t]) = \mathbb{P}(X \leq t) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq t\})$$

and μ_X is called the *distribution* of X , or the *law* of X .

Note. μ_X is the image of \mathbb{P} under

$$\begin{aligned} \Omega &\rightarrow \mathbb{R} \\ \omega &\mapsto X(\omega) \end{aligned}$$

Definition (distribution function). The function

$$\begin{aligned} F_X : \mathbb{R} &\rightarrow [0, 1] \\ t &\mapsto \mathbb{P}(X \leq t) \end{aligned}$$

is called the *distribution function* of X .

Proposition 5.1. *If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $X : \Omega \rightarrow \mathbb{R}$ is a random variable then F_X is non-decreasing, right-continuous and it determines μ_X uniquely.*

Proof. Given $t_n \downarrow t$,

$$F_X(t_n) = \mathbb{P}(X \leq t_n) \rightarrow \mathbb{P}\left(\bigcap_{n \geq 1} \{X \leq t_n\}\right) = \mathbb{P}(\{X \leq t\}) = F_X(t)$$

by downward monotone convergence for sets. Uniqueness follows from Dynkin lemma applied to the π -system $\{\emptyset\} \cup \{(-\infty, t]\}_{t \in \mathbb{R}}$. \square

Conversely,

Proposition 5.2. *If $F : \mathbb{R} \rightarrow [0, 1]$ is a non-decreasing right-continuous function with*

$$\begin{aligned} \lim_{t \rightarrow -\infty} F(t) &= 0 \\ \lim_{t \rightarrow +\infty} F(t) &= 1 \end{aligned}$$

then there exists a unique probability measure μ on \mathbb{R} such that

$$F(t) = \mu((-\infty, t])$$

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

Remark. The measure μ is called the Lebesgue-Stieltjes measure on \mathbb{R} associated to F . Furthermore for all $a, b \in \mathbb{R}$,

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

We can also construct Lebesgue measure this way.

Proof. Uniqueness is the same as above. For existence, we use the lemma

Lemma 5.3. *Let*

$$\begin{aligned} g : (0, 1) &\rightarrow \mathbb{R} \\ y &\mapsto \inf\{x \in \mathbb{R} : F(x) \geq y\} \end{aligned}$$

then g is non-decreasing, left-continuous and for all $x \in \mathbb{R}, y \in (0, 1), g(y) \leq x$ if and only if $F(x) \geq y$.

Proof. Let

$$I_y = \{x \in \mathbb{R} : F(x) \geq y\}.$$

Clearly if $y_1 \geq y_2$ then $I_{y_1} \subseteq I_{y_2}$ so $g(y_2) \leq g(y_1)$ so g is non-decreasing. I_y is an interval of \mathbb{R} because if $x > x_1$ and $x_1 \in I_y$ then $F(x) \geq F(x_1) \geq y$ so $x \in I_y$. So I_y is an interval with endpoints $g(y)$ and $+\infty$. But F is right-continuous so $g(y) = \min I_y$ and the minimum is obtained. Thus $I_y = [g(y), +\infty)$.

This means that $x \geq g(y)$ if and only if $x \in I_y$ if and only if $F(x) \geq y$.

Finally for left-continuity, suppose $y_n \uparrow y$ then $\bigcap_{n \geq 1} I_{y_n} = I_y$ by definition of I_y so $g(y_n) \rightarrow g(y)$. \square

Remark. If F is continuous and strictly increasing then $g = F^{-1}$.

Now back to the proposition. Set $\mu = g_* m$ where m is the Lebesgue measure on $(0, 1)$. μ is a probability measure as g is Borel-measurable. By the lemma

$$\mu((a, b]) = m(g^{-1}(a, b]) = m((F(a), F(b))) = F(b) - F(a).$$

□

Proposition 5.4. *If μ is a Borel probability measure on \mathbb{R} then there exists some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X on Ω such that $\mu_X = \mu$.*

In fact, one can even pick $\Omega = (0, 1)$, $\mathcal{F} = \mathcal{B}(0, 1)$ and $\mathbb{P} = m$, the Lebesgue measure.

Proof. For the first claim set $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$, $\mathbb{P} = \mu$ and $X(x) = x$.

For the second claim, set $F(t) = \mu((-\infty, t])$ and take $X = g$ where g is the auxillary function defined in the previous lemma, namely

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

Check that $\mu_X = \mu$:

$$\begin{aligned} \mu_X((a, b]) &= \mathbb{P}(X \in (a, b]) \\ &= m(\{\omega \in (0, 1) : a < X(\omega) \leq b\}) \\ &= m(\{\omega \in (0, 1) : F(a) < \omega < F(b)\}) \end{aligned}$$

□

Remark. If μ is a Borel probability measure on \mathbb{R} such that $\mu = f dt$ for some $f \geq 0$ measurable, we say that μ has a *density* (with respect to Lebesgue measure) and f is called the *density of μ* . Here $\mu = f dt$ means that $\mu((a, b]) = \int_a^b f(t) dt$.

Example.

1. uniform distribution on $[0, 1]$:

$$\begin{aligned} f(t) &= \mathbf{1}_{[0,1]}(t) \\ F(t) &= \mu((-\infty, t] \cap [0, 1]) \end{aligned}$$

2. exponential distribution of rate λ :

$$\begin{aligned} f_\lambda(t) &= \lambda e^{-\lambda t} \mathbf{1}_{t \geq 0} \\ F_\lambda(t) &= \int_{-\infty}^t f_\lambda(s) ds = \mathbf{1}_{t \geq 0} (1 - e^{-\lambda t}) \end{aligned}$$

3. Gaussian distribution with standard deviation σ and mean m :

$$\begin{aligned} f_{\sigma, m}(t) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-m)^2}{2\sigma^2}\right) \\ F_{\sigma, m}(t) &= \int_{-\infty}^t \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(s-m)^2}{2\sigma^2}\right) ds \end{aligned}$$

Definition (mean, moment, variance). If X is a random variable then

1. $\mathbb{E}(X)$ is called the *mean*,
2. $\mathbb{E}(X^k)$ is called the *kth-moment* of X ,
3. $\text{Var}(X) = \mathbb{E}((X - \mathbb{E}X)^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ is called the *variance*.

Remark. Suppose $f \geq 0$ is measurable and X is a random variable. Then

$$\mathbb{E}(f(X)) = \int_{\mathbb{R}} f(x) d\mu_X(x)$$

where by definition of $\mu_X = X_*\mathbb{P}$.

6 Independence

Independence is the key notion that makes probability theory distinct from (abstract) measure theory.

Definition (independence). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A sequence of events $(A_n)_{n \geq 1}$ is called *independent* or *mutually independent* if for all $F \subseteq \mathbb{N}$ finite,

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

Definition (independent σ -algebra). A sequence of σ -algebras $(\mathcal{A}_n)_{n \geq 1}$ where $\mathcal{A}_n \subseteq \mathcal{F}$ is called *independent* if for all $A_n \in \mathcal{A}_n$, the family $(A_n)_{n \geq 1}$ is independent.

Remark.

1. To prove that $(\mathcal{A}_n)_{n \geq 1}$ is an independent family, it is enough to check the independence condition for all A_n 's with $A_n \in \Pi_n$ where Π_n is a π -system generating \mathcal{A}_n . The proof is an application of Dynkin lemma. For example for σ -algebras $\mathcal{A}_1, \mathcal{A}_2$, suffices to check

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

for all $A_1 \in \Pi_1, A_2 \in \Pi_2$. Fix $A_2 \in \Pi_2$, look at the measures

$$\begin{aligned} A_1 &\mapsto \mathbb{P}(A_1 \cap A_2) \\ A_1 &\mapsto \mathbb{P}(A_1)\mathbb{P}(A_2) \end{aligned}$$

on \mathcal{A}_1 . They coincide on Π_1 by assumption and hence everywhere on \mathcal{A}_1 . Subsequently consider \mathcal{A}_2 .

Notation. Suppose X is a random variable. Denote by $\sigma(X)$ the smallest σ -subalgebra \mathcal{A} of \mathcal{F} such that X is \mathcal{A} -measurable, i.e.

$$\sigma(X) = \sigma(\{\omega \in \Omega : X(\omega) \leq t\}_{t \in \mathbb{R}}).$$

Definition (independence). A sequence of random variables $(X_i)_{i \geq 1}$ is called *independent* if the sequence of σ -subalgebras $(\sigma(X_i))_{i \geq 1}$ is independent.

Remark. This is equivalent to the condition that for all $(t_i)_{i \geq 1}$, for all n ,

$$\mathbb{P}((X_1 \leq t_1) \cap \dots \cap (X_n \leq t_n)) = \prod_{i=1}^n \mathbb{P}(X_i \leq t_i).$$

Yet another equivalent formulation is

$$\mu_{(X_1, \dots, X_n)} = \bigotimes_{i=1}^n \mu_{X_i}$$

as Borel probability measures on \mathbb{R}^n . "The joint law is the same as the product of individual laws".

Note. Note that independence is a property of a *family* so pairwise independence is necessary but not sufficient for independence. A famous counterexample is *Berstein's example*: take X and Y to be random variables for two independent fair coins flips. Set $Z = |X - Y|$. Then $Z = 0$ if and only if $X = Y$. Check that

$$\mathbb{P}(Z = 0) = \mathbb{P}(Z = 1) = \frac{1}{2}$$

and each pair (X, Y) , (X, Z) and (Y, Z) is independent. But (X, Y, Z) is not independent.

Proposition 6.1. *If X and Y are independent random variables, $X \geq 0, Y \geq 0$ then*

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Proof. Essentially Tonelli-Fubini:

$$\begin{aligned} \mathbb{E}(XY) &= \int_{\mathbb{R}^2} xy d\mu_{X,Y}(x, y) = \int_{\mathbb{R}^2} d\mu_X(x) d\mu_Y(y) \\ &= \left(\int_{\mathbb{R}} x d\mu_X(x) \right) \left(\int_{\mathbb{R}} y d\mu_Y(y) \right) \\ &= \mathbb{E}(X)\mathbb{E}(Y) \end{aligned}$$

□

Remark. As in Tonelli-Fubini, we may require XY to be integrable instead and the same conclusion holds.

Example. Let $\Omega = (0, 1)$, $\mathcal{F} = \mathcal{B}(0, 1)$, $\mathbb{P} = m$ the Lebesgue measure. Write the decimal expansion of $\omega \in (0, 1)$ as

$$\omega = 0.\varepsilon_1\varepsilon_2\dots$$

where $\varepsilon_i(\omega) \in \{0, \dots, 9\}$. Choose a convention so that each ω has a well-defined expansion (to avoid things like $0.099\dots = 0.100\dots$). Now let $X_n(\omega) = \varepsilon_n(\omega)$. Claim that the $(X_n)_{n \geq 1}$ are iid. random variables uniformly distributed on $\{0, \dots, 9\}$, where “iid.” stands for *independently and identically distributed*.

Proof. Easy check. For example $X_1(\omega) = \lfloor 10\omega \rfloor$ so

$$\mathbb{P}(X_1 = i_1) = \frac{1}{10}.$$

Similarly for all n

$$\mathbb{P}(X_1 = i_1, \dots, X_n = i_n) = \frac{1}{10^n}, \quad \mathbb{P}(X_n = i_n) = \frac{1}{10}$$

so

$$\mathbb{P}(X_1 = i_1, \dots, X_n = i_n) = \prod_{k=1}^n \mathbb{P}(X_k = i_k).$$

□

Remark.

$$\omega = \sum_{n \geq 1} \frac{X_n(\omega)}{10^n}$$

is distributed according to Lebesgue measure so if we want we can construct Lebesgue measure as the law of this random variable.

Proposition 6.2 (infinite product of product measure). *Let $(\Omega_i, \mathcal{F}_i, \mu_i)_{i \geq 1}$ be a sequence of probability spaces, $\Omega = \prod_{i \geq 1} \Omega_i$ and \mathcal{E} be the Boolean algebra of cylinder sets, i.e. sets of the form*

$$A \times \prod_{i \geq n} \Omega_i$$

for some $A \in \bigotimes_{i=1}^n \mathcal{F}_i$. Set $\mathcal{F} = \sigma(\mathcal{E})$, the infinite product σ -algebra. Then there is a unique probability measure μ on (Ω, \mathcal{F}) such that it agrees with product measures on all cylinder sets, i.e.

$$\mu\left(A \times \prod_{i > n} \Omega_i\right) = \left(\bigotimes_{i=1}^n \mu_i\right)(A)$$

for all $A \in \bigotimes_{i=1}^n \mathcal{F}_i$.

Proof. Omitted. See example sheet 3. □

Lemma 6.3 (Borel-Cantelli). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(A_n)_{n \geq 1}$ a sequence of events.*

1. *If $\sum_{n \geq 1} \mathbb{P}(A_n) < \infty$ then*

$$\mathbb{P}\left(\limsup_n A_n\right) = 0.$$

2. *Conversely, if $(A_n)_{n \geq 1}$ are independent and $\sum_{n \geq 1} \mathbb{P}(A_n) = \infty$ then*

$$\mathbb{P}\left(\limsup_n A_n\right) = 1.$$

Note that $\limsup_n A_n$ is also called A_n i.o. meaning “infinitely often”.

Proof.

1. Let $Y = \sum_{n \geq 1} \mathbf{1}_{A_n}$ be a random variable. Then

$$\mathbb{E}(Y) = \sum_{n \geq 1} \mathbb{E}(\mathbf{1}_{A_n}) = \sum_{n \geq 1} \mathbb{P}(A_n).$$

Since $Y \geq 0$, recall that we prove that $\mathbb{E}(Y) < \infty$ implies that $Y < \infty$ almost surely, i.e. \mathbb{P} -almost everywhere.

2. Note that

$$\left(\limsup_n A_n\right)^c = \bigcup_N \bigcap_{n \geq N} A_n^c$$

so

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{n \geq N} A_n^c\right) &\leq \mathbb{P}\left(\bigcap_{n=N}^M A_n^c\right) \\
&= \prod_{n=N}^M \mathbb{P}(A_n^c) = \prod_{n=N}^M (1 - \mathbb{P}(A_n)) \\
&\leq \prod_{n=N}^M \exp(-\mathbb{P}(A_n)) \\
&\leq \exp\left(-\sum_{n=N}^M \mathbb{P}(A_n)\right) \\
&\rightarrow 0
\end{aligned}$$

as $M \rightarrow \infty$. Thus

$$\mathbb{P}\left(\bigcap_{n \geq N} A_n^c\right) = 0$$

for all N so

$$\mathbb{P}\left(\bigcup_N \bigcap_{n \geq N} A_n^c\right) = 0.$$

□

Definition (random/stochastic process, filtration, tail σ -algebra, tail event). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X_n)_{n \geq 1}$ a sequence of random variables.

1. $(X_n)_{n \geq 1}$ is sometimes called a *random process* or *stochastic process*.

2.

$$\mathcal{F}_n = \sigma(X_1, \dots, X_n) \subseteq \mathcal{F}$$

is called the associated *filtration*. $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$.

3.

$$\mathcal{C} = \bigcap_{n \geq 1} \sigma(X_n, X_{n+1}, \dots)$$

is called the *tail σ -algebra* of the process. Its elements are called *tail events*.

Example. Tail events are those not affected by the first few terms in the sequence of random variables. For example,

$$\{\omega \in \Omega : \lim_n X_n(\omega) \text{ exists}\}$$

is a tail event, so is

$$\{\omega \in \Omega : \limsup_n X_n(\omega) \geq T\}.$$

Theorem 6.4 (Kolmogorov 0–1 law). *If $(X_n)_{n \geq 1}$ is a sequence of mutually*

independent random variables then for all $A \in \mathcal{C}$,

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

Proof. Pick $A \in \mathcal{C}$. Fix n . For all $B \in \sigma(X_1, \dots, X_n)$,

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

as \mathcal{C} is independent of $\sigma(X_1, \dots, X_n)$. The measures $B \mapsto \mathbb{P}(A)\mathbb{P}(B)$ and $B \mapsto \mathbb{P}(A \cap B)$ coincide on each \mathcal{F}_n so on $\bigcup_{n \geq 1} \mathcal{F}_n$. Hence they coincide on $\sigma(\bigcup_{n \geq 1} \mathcal{F}_n) \supseteq \mathcal{C}$ so

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

so

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

□

6.1 Useful inequalities

Proposition 6.5 (Cauchy-Schwarz). *Suppose X, Y are random variables then*

$$\mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2) \cdot \mathbb{E}(Y^2)}.$$

Proof. For all $t \in \mathbb{R}$,

$$0 \leq \mathbb{E}((|X| + t|Y|)^2) = \mathbb{E}(X^2) + 2t\mathbb{E}(|XY|) + t^2\mathbb{E}(Y^2)$$

so viewed as a quadratic in t , the discriminant is nonpositive, i.e.

$$(\mathbb{E}(|XY|))^2 - \mathbb{E}(X^2) - \mathbb{E}(Y^2) \leq 0.$$

□

Proposition 6.6 (Markov). *Let $X \geq 0$ be a random variable. Then for all $t \geq 0$,*

$$t\mathbb{P}(X \geq t) \leq \mathbb{E}(X).$$

Proof.

$$\mathbb{E}(X) \geq \mathbb{E}(X\mathbf{1}_{X \geq t}) \geq \mathbb{E}(t\mathbf{1}_{X \geq t}) = t\mathbb{P}(X \geq t)$$

□

Proposition 6.7 (Chebyshev). *Let Y be a random variable with $\mathbb{E}(Y^2) < \infty$, then for all $t \in \mathbb{R}$,*

$$t^2\mathbb{P}(|Y - \mathbb{E}(Y)| \geq t) \leq \text{Var } Y.$$

$\mathbb{E}(Y^2) < \infty$ implies that $\mathbb{E}(|Y|) < \infty$ by Cauchy-Schwarz, so $\text{Var } Y < \infty$. The converse is more subtle.

Proof. Apply Markov to $X = |Y - \mathbb{E}(Y)|^2$.

□

Theorem 6.8 (strong law of large numbers). *Let $(X_n)_{n \geq 1}$ be a sequence of iid. random variables. Assume $\mathbb{E}(|X_1|) < \infty$. Let*

$$S_n = \sum_{k=1}^n X_k,$$

then $\frac{1}{n}S_n$ converges almost surely to $\mathbb{E}(X_1)$.

Proof. We prove the theorem under a stronger condition: we assume $\mathbb{E}(X_1^4) < \infty$. This implies, by Cauchy-Schwarz, $\mathbb{E}(X_1^2), \mathbb{E}(|X_1|) < \infty$. Subsequently $\mathbb{E}(|X_1|^3) < \infty$. The full proof is much harder but will be given later when we have developed enough machinery.

wlog we may assume $\mathbb{E}(X_1) = 0$ by replacing X_n with $X_n - \mathbb{E}(X_1)$. Have

$$\mathbb{E}(S_n^4) = \sum_{i,j,k,\ell} \mathbb{E}(X_i X_j X_k X_\ell).$$

All terms vanish because $\mathbb{E}(X_i) = 0$ and $(X_i)_{i \geq 1}$ are independent, except for $\mathbb{E}(X_i^4)$ and $\mathbb{E}(X_i^2 X_j^2)$ for $i \neq j$. For example,

$$\mathbb{E}(X_i X_j^3) = \mathbb{E}(X_i) \cdot \mathbb{E}(X_j^3) = 0$$

for $i \neq j$. Thus

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

By Cauchy-Schwarz,

$$\mathbb{E}(X_i^2 X_j^2) \leq \sqrt{\mathbb{E}X_i^4 \cdot \mathbb{E}X_j^4} = \mathbb{E}X_1^4$$

so

$$\mathbb{E}(S_n^4) \leq (n + 6 \cdot \frac{n(n-1)}{2}) \mathbb{E}X_1^4$$

and asymptotically,

$$\mathbb{E}\left(\frac{S_n}{n}\right)^4 = O\left(\frac{1}{n^2}\right)$$

so

$$\mathbb{E}\left(\sum_{n \geq 1} \left(\frac{S_n}{n}\right)^4\right) = \sum_{n \geq 1} \mathbb{E}\left(\frac{S_n}{n}\right)^4 < \infty.$$

Hence $\sum_{n \geq 1} \left(\frac{S_n}{n}\right)^4 < \infty$ almost surely and it follows that

$$\lim_{n \rightarrow \infty} \frac{S_n}{n} = 0$$

almost surely. □

Strong law of large numbers has a very important statistical implication: we can sample the mean of larger number of iid. to detect an unknown law, at least the mean.

7 Convergence of random variables

Definition (weak convergence). A sequence of probability measures $(\mu_n)_{n \geq 1}$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is said to *converge weakly* to a measure μ if for all $f \in C_b(\mathbb{R}^d)$, the set of continuous bounded functions on \mathbb{R}^d ,

$$\lim_{n \rightarrow \infty} \mu_n(f) = \mu(f).$$

Example.

1. Let $\mu_n = \delta_{1/n}$ be the *Dirac mass* on \mathbb{R}^d , i.e. for $x \in \mathbb{R}^d$, δ_x is the Borel probability measure on \mathbb{R}^d such that

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

then $\mu_n \rightarrow \delta_0$.

2. Let $\mu_n = \mathcal{N}(0, \sigma_n^2)$, Gaussian distribution with standard deviation σ_n , where $\sigma_n \rightarrow 0$, then again $\mu_n \rightarrow \delta_0$. Indeed,

$$\begin{aligned} \mu_n(f) &= \int f(x) d\mu_n(x) \\ &= \int f(x) \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{x^2}{2\sigma_n^2}\right) dx \\ &= \int f(x\sigma_n) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \end{aligned}$$

$\sigma_n \rightarrow 0$ so $f(x\sigma_n) \rightarrow f(0)$ so by dominated convergence theorem, $\mu_n(f) \rightarrow f(0) = \delta_0(f)$.

Definition (convergence of random variable). A sequence $(X_n)_{n \geq 1}$ of \mathbb{R}^d -valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ is said to converge to a random variable X

1. *almost surely* if

$$\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)$$

for \mathbb{P} -almost every ω .

2. *in probability* or *in measure* if for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(\|X_n - X\| > \varepsilon) = 0.$$

Note that all norms on \mathbb{R}^d are equivalent so we don't have to specify one.

3. *in distribution* or *in law* if $\mu_{X_n} \rightarrow \mu_X$ weakly, where $\mu_X = X_*\mathbb{P}$ is the law of X , a Borel probability measure on \mathbb{R}^d . Equivalently, for all $f \in C_b(\mathbb{R}^d)$, $\mathbb{E}(f(X_n)) \rightarrow \mathbb{E}(f(X))$.

Proposition 7.1. $1 \implies 2 \implies 3$.

Proof.

1. $1 \implies 2$:

$$\mathbb{P}(\|X_n - X\| > \varepsilon) = \mathbb{E}(\mathbf{1}_{\|X_n - X\| > \varepsilon})$$

so if $X_n \rightarrow X$ almost surely then

$$\mathbf{1}_{\|X_n - X\| > \varepsilon} \rightarrow 0$$

\mathbb{P} -almost everywhere so by dominated convergence theorem $\mathbb{P}(\|X_n - X\| > \varepsilon) \rightarrow 0$.

2. $2 \implies 3$: given $f \in C_b(\mathbb{R}^d)$, need to show that $\mu_{X_n}(f) \rightarrow \mu_X(f)$. But

$$\mu_{X_n}(f) - \mu_X(f) = \mathbb{E}(f(X_n) - f(X)).$$

To bound this, note that f is continuous and \mathbb{R}^d is locally compact so it is locally uniformly continuous. In particular for all $\varepsilon > 0$ exists $\delta > 0$ such that if $\|x\| < \frac{1}{\varepsilon}$ and $\|y - x\| < \delta$ then $|f(y) - f(x)| < \varepsilon$. Thus

$$\begin{aligned} |\mathbb{E}(f(X_n) - f(X))| &\leq \mathbb{E}(\mathbf{1}_{\|X_n - X\| < \delta} \underbrace{\mathbf{1}_{\|X\| < 1/\varepsilon} |f(X_n) - f(X)|}_{< \varepsilon}) \\ &\quad + 2 \underbrace{\|f\|_\infty}_{< \infty} (\mathbb{P}(\|X_n - X\| \geq \delta) + \mathbb{P}(\|X\| \geq \frac{1}{\varepsilon})) \end{aligned}$$

so

$$\limsup_{n \rightarrow \infty} |\mathbb{E}(f(X_n) - f(X))| \leq \varepsilon + 2\|f\|_\infty \underbrace{\mathbb{P}(\|X\| > \frac{1}{\varepsilon})}_{\rightarrow 0 \text{ as } \varepsilon \rightarrow 0}.$$

which is 0 as ε is arbitrary. □

Remark. When $d = 1$, 3 is equivalent to $F_{X_n}(x) \rightarrow F_X(x)$ for all x as $n \rightarrow \infty$ where F_X is continuous. See example sheet 3.

The strict converses do not hold but we can say something weaker.

Proposition 7.2. *If $X_n \rightarrow X$ in probability then there is a subsequence $(n_k)_k$ such that $X_{n_k} \rightarrow X$ almost surely as $k \rightarrow \infty$.*

Proof. We know for all $\varepsilon > 0$, $\mathbb{P}(\|X_n - X\| > \varepsilon) \rightarrow 0$ as $n \rightarrow \infty$. So for all k exists n_k such that

$$\mathbb{P}(\|X_{n_k} - X\| > \frac{1}{k}) \leq \frac{1}{2^k}$$

so

$$\sum_{k \geq 1} \mathbb{P}(\|X_{n_k} - X\| > \frac{1}{k}) < \infty$$

so by the first Borel-Cantelli lemma

$$\mathbb{P}(\|X_{n_k} - X\| > \frac{1}{k} \text{ i.o.}) = 0.$$

This means that with probability 1, $\|X_{n_k} - X\| \rightarrow 0$ as $k \rightarrow \infty$. \square

Definition (convergence in mean). Let $(X_n)_{n \geq 1}$ and X be \mathbb{R}^d -valued integrable random variables. We say that $(X_n)_n$ *converges in mean* or *in L^1* to X if

$$\lim_{n \rightarrow \infty} \mathbb{E}(\|X_n - X\|) = 0.$$

Remark.

1. If $X_n \rightarrow X$ in mean then $X_n \rightarrow X$ in probability by Markov inequality:

$$\varepsilon \cdot \mathbb{P}(\|X_n - X\| > \varepsilon) \leq \mathbb{E}(\|X_n - X\|).$$

2. The converse is false. For example take $\Omega = (0, 1)$, $\mathcal{F} = \mathcal{B}(\Sigma)$ and \mathbb{P} Lebesgue measure. Let $X_n = n \mathbf{1}_{[0, \frac{1}{n}]}$. $X_n \rightarrow 0$ almost surely but $\mathbb{E}X_n = 1$.

When does convergence in probability imply convergence in mean? We need some kind of domination assumption.

Definition (uniformly integrable). A sequence of random variables $(X_n)_{n \geq 1}$ is *uniformly integrable* if

$$\lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{E}(\|X_n\| \mathbf{1}_{\|X_n\| \geq M}) = 0.$$

Remark. If $(X_n)_{n \geq 1}$ are *dominated*, namely exists an integrable random variable $Y \geq 0$ such that $\|X_n\| \leq Y$ for all n then $(X_n)_n$ is uniformly integrable by dominated convergence theorem:

$$\mathbb{E}(\|X_n\| \mathbf{1}_{\|X_n\| \geq M}) \leq \mathbb{E}(Y \mathbf{1}_{Y \geq M}) \rightarrow 0$$

as $M \rightarrow \infty$.

Theorem 7.3. Let $(X_n)_{n \geq 1}$ be a sequence of \mathbb{R}^d -valued integrable random variable. Let X be another random variable. Then TFAE:

1. X is integrable and $X_n \rightarrow X$ in mean,
2. $(X_n)_{n \geq 1}$ is uniformly integrable and $X_n \rightarrow X$ in probability.

Proof.

- 1 \implies 2: Left to show uniform integrability:

$$\begin{aligned} \mathbb{E}(\|X_n\| \mathbf{1}_{\|X_n\| \geq M}) &\leq \mathbb{E}(\|X_n - X\| \mathbf{1}_{\|X_n\| \geq M}) + \mathbb{E}(\|X\| \mathbf{1}_{\|X_n\| \geq M}) \\ &\leq \mathbb{E}(\|X_n - X\|) + \mathbb{E}(\|X\| \mathbf{1}_{\|X_n\| \geq M} (\mathbf{1}_{\|X\| \leq \frac{M}{2}} + \mathbf{1}_{\|X\| > \frac{M}{2}})) \\ &\leq \mathbb{E}(\|X_n - X\|) + \mathbb{E}(\|X\| \mathbf{1}_{\|X_n - X\| \geq \frac{M}{2}} \mathbf{1}_{\|X\| \leq \frac{M}{2}}) + \mathbb{E}(\|X\| \mathbf{1}_{\|X\| \geq \frac{M}{2}}) \\ &\leq \mathbb{E}(\|X_n - X\|) + \frac{M}{2} \mathbb{P}(\|X_n - X\| \geq \frac{M}{2}) + \mathbb{E}(\|X\| \mathbf{1}_{\|X\| \geq \frac{M}{2}}) \end{aligned}$$

Take lim sup,

$$\limsup_{n \rightarrow \infty} \mathbb{E}(\|X_n\| \mathbf{1}_{\|X_n\| \geq n}) \leq 0 + 0 + \mathbb{E}(\|X\| \mathbf{1}_{\|X\| \geq \frac{M}{2}}) \rightarrow 0$$

by dominated convergence theorem.

- 2 \implies 1: Prove first that X is integrable. By the previous proposition, we can find a subsequence $(n_k)_k$ such that $X_{n_k} \rightarrow X$ almost surely. By Fatou's lemma,

$$\mathbb{E}(\|X\| \mathbf{1}_{\|X\| \geq M}) \leq \liminf_{k \rightarrow \infty} \mathbb{E}(\|X_{n_k}\| \mathbf{1}_{\|X_{n_k}\| \geq M})$$

which goes to 0 as $M \rightarrow \infty$ by uniform integrability assumption. Thus

$$\mathbb{E}(\|X\|) \leq M + \mathbb{E}(\|X\| \mathbf{1}_{\|X\| \geq M}) < \infty$$

for M sufficiently large. Thus X is integrable.

To show convergence in mean, we use the same trick of splitting into small and big parts.

$$\begin{aligned} \mathbb{E}(\|X_n - X\|) &= \mathbb{E}((\mathbf{1}_{\|X_n - X\| \leq \varepsilon} + \mathbf{1}_{\|X_n - X\| > \varepsilon}) \|X_n - X\|) \\ &\leq \varepsilon + \mathbb{E}(\mathbf{1}_{\|X_n - X\| > \varepsilon} \|X_n - X\| (\mathbf{1}_{\|X_n\| \leq M} + \mathbf{1}_{\|X_n\| > M})) \\ &\leq \varepsilon + \dagger + \ddagger \end{aligned}$$

where

$$\begin{aligned} \dagger &= \mathbb{E}(\|X_n - X\| \mathbf{1}_{\|X_n - X\| > \varepsilon} \mathbf{1}_{\|X_n\| \leq M}) \\ &\leq \mathbb{E}((M + \|X\|) \mathbf{1}_{\|X_n - X\| > \varepsilon} (\mathbf{1}_{\|X\| \leq M} + \mathbf{1}_{\|X\| > M})) \\ &\leq 2M\mathbb{P}(\|X_n - X\| > \varepsilon) + 2\mathbb{E}(\|X\| \mathbf{1}_{\|X\| > M}) \end{aligned}$$

so

$$\limsup_{n \rightarrow \infty} \dagger \leq 2\mathbb{E}(\|X\| \mathbf{1}_{\|X\| > M}) \rightarrow 0$$

as $M \rightarrow \infty$. On the other hand

$$\begin{aligned} \ddagger &= \mathbb{E}(\|X_n - X\| \mathbf{1}_{\|X_n - X\| > \varepsilon} \mathbf{1}_{\|X_n\| > M}) \\ &\leq \mathbb{E}((\|X_n\| + \|X\|) \mathbf{1}_{\|X_n\| \geq M}) \\ &\leq \mathbb{E}(\|X_n\| \mathbf{1}_{\|X_n\| \geq M} + \|X\| \mathbf{1}_{\|X\| > M} + \|X\| \mathbf{1}_{\|X_n\| \geq M} \mathbf{1}_{\|X\| \leq M}) \\ &\leq 2\mathbb{E}(\|X_n\| \mathbf{1}_{\|X_n\| \geq M}) + \mathbb{E}(\|X\| \mathbf{1}_{\|X\| > M}) \end{aligned}$$

taking lim sup,

$$\limsup_{n \rightarrow \infty} \ddagger \leq 2 \limsup_{n \rightarrow \infty} \mathbb{E}(\|X_n\| \mathbf{1}_{\|X_n\| \geq M}) + \mathbb{E}(\|X\| \mathbf{1}_{\|X\| > M}) \rightarrow 0 + 0$$

as $M \rightarrow \infty$.

□

Definition. We say that a sequence of random variables $(X_n)_{n \geq 1}$ is *bounded* in L^p if there exists $C > 0$ such that $\mathbb{E}(\|X_n\|^p) \leq C$ for all n .

Proposition 7.4. *If $p > 1$ and $(X_n)_{n \geq 1}$ is bounded in L^p then $(X_n)_{n \geq 1}$ is uniformly integrable.*

Proof.

$$M^{p-1} \mathbb{E}(\|X_n\| \mathbf{1}_{\|X_n\| \geq M}) \leq \mathbb{E}(\|X_n\|^p \mathbf{1}_{\|X_n\| \geq M}) \leq \mathbb{E}(\|X_n\|^p) \leq C$$

so

$$\limsup_{n \rightarrow \infty} \mathbb{E}(\|X_n\| \mathbf{1}_{\|X_n\| \geq M}) \leq \frac{C}{M^{p-1}} \rightarrow 0$$

as $M \rightarrow \infty$. □

This provides a sufficient condition for uniform integrability and thus convergence in mean.

8 L^p spaces

Recall that $\varphi : I \rightarrow \mathbb{R}$ is convex means that for all $x, y \in I$, for all $t \in [0, 1]$,

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

Proposition 8.1 (Jensen inequality). *Let I be an open interval of \mathbb{R} and $\varphi : I \rightarrow \mathbb{R}$ a convex function. Let X be a random variable $(\Omega, \mathcal{F}, \mathbb{P})$. Assume X is integrable and takes values in I . Then*

$$\mathbb{E}(\varphi(X)) \geq \varphi(\mathbb{E}(X)).$$

Remark.

1. As $X \in I$ almost surely and I is an interval, have $\mathbb{E}(X) \in I$.
2. We'll show that $\varphi(X)^-$ is integrable so

$$\mathbb{E}(\varphi(X)) = \mathbb{E}(\varphi(X)^+) - \mathbb{E}(\varphi(X)^-)$$

with the possibility that both sides are infinity.

Lemma 8.2. *TFAE:*

1. φ is convex,
2. there exists a family \mathcal{F} of affine functions $(x \mapsto ax + b)$ such that $\varphi = \sup_{\ell \in \mathcal{F}} \ell$ on I .

Proof.

1. $2 \implies 1$: every ℓ is convex and the supremum of ℓ is convex:

$$\ell(tx + (1-t)y) = t\ell(x) + (1-t)\ell(y) \leq t \sup_{\ell \in \mathcal{F}} \ell(x) + (1-t) \sup_{\ell \in \mathcal{F}} \ell(y)$$

so

$$\begin{aligned} \varphi(tx + (1-t)y) &= \sup_{\ell \in \mathcal{F}} \ell(tx + (1-t)y) \\ &\leq t \sup_{\ell \in \mathcal{F}} \ell(x) + (1-t) \sup_{\ell \in \mathcal{F}} \ell(y) \\ &= t\varphi(x) + (1-t)\varphi(y) \end{aligned}$$

2. We need to show that for all $x_0 \in I$ we can find an affine function

$$\ell_{x_0}(x) = \theta_{x_0}(x - x_0) + \varphi(x_0),$$

where θ_{x_0} is morally the slope at x_0 , such that $\varphi(x) \geq \ell_{x_0}(x)$ for all $x \in I$. Then have $\varphi = \sup_{x_0 \in I} \ell_{x_0}$.

To find θ_{x_0} observe that for all $x < x_0 < y$ where $x, y \in I$, have

$$\frac{\varphi(x_0) - \varphi(x)}{x_0 - x} \leq \frac{\varphi(y) - \varphi(x_0)}{y - x_0}.$$

Indeed this is the convexity of φ on $[x, y]$ with $t = \frac{x_0 - x}{y - x}$. This holds for all $x < x_0, y > x_0$ so there exists $\theta \in \mathbb{R}$ such that

$$\frac{\varphi(x_0) - \varphi(x)}{x_0 - x} \leq \theta \leq \frac{\varphi(y) - \varphi(x_0)}{y - x_0}.$$

Then just set $\ell_{x_0}(x) = \theta(x - x_0) + \varphi(x_0)$. By construction $\varphi(x) \geq \ell_{x_0}(x)$ for all $x \in I$. □

Proof of Jensen inequality. Let $\varphi(x) = \sup_{\ell \in \mathcal{F}} \ell(x)$ where ℓ affine. then

$$\mathbb{E}(\varphi(X)) \geq \mathbb{E}(\ell(X)) = \ell(\mathbb{E}(X))$$

for all $\ell \in \mathcal{F}$ so take supremum,

$$\mathbb{E}(\varphi(X)) \geq \sup_{\ell \in \mathcal{F}} \ell(\mathbb{E}(X)) = \varphi(\mathbb{E}(X)).$$

Also for the remark,

$$-\varphi = -\sup_{\ell \in \mathcal{F}} \ell = \inf_{\ell \in \mathcal{F}} (-\ell)$$

so $\varphi^- = (-\varphi)^+ \leq |\ell|$ for all $\ell \in \mathcal{F}$. Then

$$\varphi(X)^- \leq |\ell(X)| \leq |a||X| + |b|.$$

As X is integrable, $\varphi(X)^-$ is integrable. □

Jensen inequality is for probability space only. The following applies to all measure spaces.

Proposition 8.3 (Minkowski inequality). *Let (X, \mathcal{A}, μ) be a measure space and f, g measurable functions on X . Let $p \in [1, \infty)$ and define the p -norm*

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

Then

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

Proof. wlog assume $\|f\|_p, \|g\|_p \neq 0$. Need to show

$$\left\| \frac{\|f\|_p}{\|f\|_p + \|g\|_p} \frac{f}{\|f\|_p} + \frac{\|g\|_p}{\|f\|_p + \|g\|_p} \frac{g}{\|g\|_p} \right\|_p \leq 1.$$

Suffice to show for all $t \in [0, 1]$, for all F, G measurable such that $\|F\|_p = \|G\|_p = 1$, have

$$\|tF + (1-t)G\|_p \leq 1$$

“the unit ball is convex”. For this note that

$$\begin{aligned} [0, +\infty) &\rightarrow [0, +\infty) \\ x &\mapsto x^p \end{aligned}$$

is convex if $p \geq 1$ so

$$|tF + (1-t)G|^p \leq t|F|^p + (1-t)|G|^p$$

and

$$\int_X |tF + (1-t)G|^p d\mu \leq t \underbrace{\int_X |F|^p d\mu}_{=1} + (1-t) \underbrace{\int_X |G|^p d\mu}_{=1} = 1.$$

□

Proposition 8.4 (Hölder inequality). *Suppose (X, \mathcal{A}, μ) is a measure space and let f, g be measurable functions on X . Given $p, q \in (1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$,*

$$\int_X |fg| d\mu \leq \left(\int_X |f|^p d\mu \right)^{1/p} \left(\int_X |g|^q d\mu \right)^{1/q}$$

with equality if and only if there exists $(\alpha, \beta) \neq (0, 0)$ such that $\alpha|f|^p = \beta|g|^q$ μ -almost everywhere.

Lemma 8.5 (Young inequality). *For all $p, q \in (1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$, for all $a, b \geq 0$, have*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Proof of Hölder inequality. wlog assume $\|f\|_p, \|g\|_q \neq 0$. By scaling by factors $(\alpha, \beta) \neq (0, 0)$ wlog $\|f\|_p = \|g\|_q = 1$. Then by Young inequality,

$$|fg| \leq \frac{1}{p}|f|^p + \frac{1}{q}|g|^q$$

so

$$\int_X |fg| d\mu \leq \frac{1}{p} \int_X |f|^p d\mu + \frac{1}{q} \int_X |g|^q d\mu = \frac{1}{p} + \frac{1}{q} = 1.$$

□

Remark. Apply Jensen inequality to $\varphi(x) = x^{p'/p}$ for $p' > p$, we have

$$\mathbb{E}(|X|^p)^{1/p} \leq \mathbb{E}(|X|^{p'})^{1/p'},$$

so the function $p \mapsto \mathbb{E}(|X|^p)^{1/p}$ is non-decreasing. This can be used, for example, to show that if X has finite p' th moment then it has finite p th moment for $p' \geq p$.

Definition. Let (X, \mathcal{A}, μ) be a measure space.

- For $p \geq 1$,

$$\mathcal{L}^p(X, \mathcal{A}, \mu) = \{f : X \rightarrow \mathbb{R} \text{ measurable such that } |f|^p \text{ is } \mu\text{-integrable}\}.$$

- For $p = \infty$,

$$\mathcal{L}^\infty(X, \mathcal{A}, \mu) = \{f : X \rightarrow \mathbb{R} \text{ measurable such that } \text{essup } |f| < \infty\}$$

where

$$\operatorname{ess\,sup} |f| = \inf\{t : |f(x)| \leq t \text{ for } \mu\text{-almost every } x\}.$$

Lemma 8.6. $\mathcal{L}^p(X, \mathcal{A}, \mu)$ is an \mathbb{R} -vector space.

Proof. For $p < \infty$ use Minkowski inequality. Similar we can check that $\|f + g\|_\infty \leq \|f\|_\infty + \|g\|_\infty$ for all f, g . \square

Definition. We say f and g are μ -equivalent and write $f \equiv_\mu g$ if for μ -almost every x , $f(x) = g(x)$.

Check that this is an equivalence relation stable under addition and multiplication.

Definition (L^p -space). Define

$$L^p(X, \mathcal{A}, \mu) = \mathcal{L}^p(X, \mathcal{A}, \mu) / \equiv_\mu$$

and if $[f]$ denotes the equivalence class of f under \equiv_μ we define

$$\|[f]\|_p = \|f\|_p.$$

Proposition 8.7. For $p \in [1, \infty]$, $L^p(X, \mathcal{A}, \mu)$ is a normed vector space under $\|\cdot\|_p$ and it is complete, i.e. it is a Banach space.

Proof. If $f \equiv_\mu g$ then $\|f\|_p = \|g\|_p$ so $\|\cdot\|_p$ on L^p is well-defined. Triangle inequality follows from Minkowski inequality and linearity is obvious so $\|\cdot\|_p$ is indeed a norm.

For completeness, pick $(f_n)_n$ a Cauchy sequence in $\mathcal{L}^p(X, \mathcal{A}, \mu)$. Need to show that there exists $f \in \mathcal{L}^p$ such that $\|f_n - f\|_p \rightarrow 0$ as $n \rightarrow \infty$. This then implies that $[f_n] \rightarrow [f]$ in L^p .

We can extract a subsequence $n_k \uparrow \infty$ such that $\|f_{n_{k+1}} - f_{n_k}\|_p \leq 2^{-k}$. Let

$$S_K = \sum_{k=1}^K |f_{n_{k+1}} - f_{n_k}|$$

then

$$\|S_K\|_p \leq \sum_{k=1}^K \|f_{n_{k+1}} - f_{n_k}\|_p \leq \sum_{k=1}^K 2^{-k} \leq 1$$

so by monotone convergence,

$$\lim_{K \rightarrow \infty} \int_X |S_K|^p d\mu = \int_X |S_\infty|^p d\mu,$$

i.e. $S_\infty \in \mathcal{L}^p$. In particular for μ -almost everywhere x , $|S_\infty(x)| < \infty$, i.e.

$$\sum_{k \geq 1} |f_{n_{k+1}}(x) - f_{n_k}(x)| < \infty.$$

Hence $(f_{n_k}(x))_k$ is a Cauchy sequence in \mathbb{R} . By completeness of \mathbb{R} , the limit exists and set $f(x)$ to be it. When this limit does not exist set $f(x) = 0$.

We then have, in case $p < \infty$, by Fatou's lemma

$$\|f_n - f\|_p \leq \liminf_{k \rightarrow \infty} \|f_n - f_{n_k}\|_p \leq \varepsilon$$

for any ε for n sufficiently large. Thus

$$\lim_{n \rightarrow \infty} \|f_n - f\|_p = 0.$$

When $p = \infty$, we use the fact that if $f_n \rightarrow f$ μ -almost everywhere then

$$\|f\|_\infty \leq \liminf_{n \rightarrow \infty} \|f_n\|_\infty.$$

Proof. Let $t > \limsup_{n \rightarrow \infty} \|f_n\|_\infty$. Then exists $n_k \uparrow \infty$ such that

$$\|f_{n_k}\|_\infty = \operatorname{ess\,sup} |f_{n_k}| = \inf\{s \geq 0 : |f_{n_k}(x)| \leq s \text{ for } \mu\text{-almost every } x\} < t$$

for all k . Thus for all k , for μ -almost every x , $|f_{n_k}(x)| < t$. But by σ -additivity of μ we can *swap the quantifiers*, i.e. for μ -almost every x , for all k , $|f_{n_k}(x)| < t$. Thus for μ -almost every x , $\|f(x)\|_\infty \leq t$. \square

\square

Proposition 8.8 (approximation by simple functions). *Let $p \in [1, \infty)$. Let V be the linear span of all simple functions on X . Then $V \cap \mathcal{L}^p$ is dense in \mathcal{L}^p .*

Proof. Note that $g \in \mathcal{L}^p$ implies $g^+, g^- \in \mathcal{L}^p$. Thus by writing $f = f^+ - f^-$ and using Minkowski inequality, suffice to show $f \geq 0$ is the limit of a sequence of simple functions.

Recall there exist simple functions $0 \leq g_n \leq f$ such that $g_n(x) \uparrow f(x)$ for μ -almost every x . Then

$$\lim_{n \rightarrow \infty} \|g_n - f\|_p^p = \lim_{n \rightarrow \infty} \int_X |g_n - f|^p d\mu = 0$$

by dominated convergence theorem ($|g_n - f| \leq g_n + f \leq 2f$ so $g_n - f$ is μ -integrable). \square

Remark. When $X = \mathbb{R}^d$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^d)$ and μ is the Lebesgue measure, $C_c(\mathbb{R}^d)$, the space of continuous functions with compact support is dense in $\mathcal{L}^p(X, \mathcal{A}, \mu)$ when $p \in [1, \infty)$ (this does not hold for $p = \infty$: a constant nonzero function has no noncompact support). See example sheet. In fact, $C_c^\infty(\mathbb{R}^d)$ suffices.

9 Hilbert space and L^2 -methods

Definition (inner product). Let V be a complex vector space. A *Hermitian inner product* on V is a map

$$\begin{aligned} V \times V &\rightarrow \mathbb{C} \\ (x, y) &\mapsto \langle x, y \rangle \end{aligned}$$

such that

1. $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$ for all $\alpha, \beta \in \mathbb{C}$, for all $x, y, z \in V$.
2. $\langle y, x \rangle = \overline{\langle x, y \rangle}$.
3. $\langle x, x \rangle \in \mathbb{R}$ and $\langle x, x \rangle \geq 0$, with equality if and only if $x = 0$.

Definition (Hermitian norm). The *Hermitian norm* is defined as $\|x\| = \sqrt{\langle x, x \rangle}$.

Lemma 9.1. *Properties of norm:*

1. *linearity:* $\|\lambda x\| = |\lambda| \|x\|$ for all $\lambda \in \mathbb{C}, x \in V$,
2. *Cauchy-Schwarz:* $|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$,
3. *triangle inequality:* $\|x + y\| \leq \|x\| + \|y\|$
4. *parallelogram identity:* $\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2)$

Proof. Exercise. For reference see author's notes on IID Linear Analysis. \square

Corollary 9.2. $(V, \|\cdot\|)$ is a normed vector space.

Definition (Hilbert space). We say $(V, \|\cdot\|)$ is a *Hilbert space* if it is complete.

Example. Let $V = L^2(X, \mathcal{A}, \mu)$ where (X, \mathcal{A}, μ) is a measure space. Then we can define

$$\langle f, g \rangle = \int_X f \bar{g} d\mu$$

which is well-defined (i.e. finite) by Cauchy-Schwarz. The axioms are easy to check, with positive-definiteness given by

$$0 = \langle f, f \rangle = \int_X |f|^2 d\mu$$

if and only if $f = 0$ μ -almost everywhere so $f = 0$.

Proposition 9.3. *Let H be a Hilbert space and let \mathcal{C} be a closed convex subset of H . Then for all $x \in H$, there exists unique $y \in \mathcal{C}$ such that*

$$\|x - y\| = d(x, \mathcal{C})$$

where by definition

$$d(x, \mathcal{C}) = \inf_{c \in \mathcal{C}} \|x - c\|.$$

This y is called the orthogonal projection of x on \mathcal{C} .

Proof. Let $c_n \in \mathcal{C}$ be a sequence such that $\|x - c_n\| \rightarrow d(x, \mathcal{C})$. Let's show that $(c_n)_n$ is a Cauchy sequence. By parallelogram identity,

$$\left\| \frac{x - c_n}{2} + \frac{x - c_m}{2} \right\|^2 + \left\| \frac{x - c_n}{2} - \frac{x - c_m}{2} \right\|^2 = \frac{1}{2}(\|x - c_n\|^2 + \|x - c_m\|^2)$$

so

$$\underbrace{\left\| \frac{x - \frac{c_n + c_m}{2}}{\in \mathcal{C}} \right\|^2}_{\geq d(x, \mathcal{C})^2} + \frac{1}{4} \|c_n - c_m\|^2 = \frac{1}{2} \underbrace{(\|x - c_n\|^2 + \|x - c_m\|^2)}_{\rightarrow d(x, \mathcal{C})^2}$$

so

$$\lim_{n, m \rightarrow \infty} \|c_n - c_m\| = 0$$

i.e. $(c_n)_n$ is Cauchy. By completeness exist $\lim_{n \rightarrow \infty} c_n = y \in H$. As \mathcal{C} is closed, $y \in \mathcal{C}$. As $\|x - c_n\| \rightarrow d(x, \mathcal{C})$, $\|x - y\| \rightarrow d(x, \mathcal{C})$. This shows existence of y .

For uniqueness, use parallelogram identity

$$\underbrace{\left\| x - \frac{y + y'}{2} \right\|^2}_{\geq d(x, \mathcal{C})^2} + \frac{1}{4} \|y - y'\|^2 = \frac{1}{2} (\|x - y\|^2 + \|x - y'\|^2) = d(x, \mathcal{C})^2$$

so $\|y - y'\| = 0$. □

Corollary 9.4. *Suppose $V \leq H$ is a closed subspace of a Hilbert space H .*

Then

$$H = V \oplus V^\perp$$

where

$$V^\perp = \{x \in H : \langle x, v \rangle = 0 \text{ for all } v \in V\}$$

is the orthogonal of V .

Proof. $V \cap V^\perp = 0$ by positivity of inner product. If $x \in H$ then there exists a unique $y \in V$ such that $\|x - y\| = d(x, V)$. Need to show that $x - y \in V^\perp$.

For all $z \in V$,

$$\|x - y - z\| \geq \|x - y\|$$

as $y + z \in V$. Thus

$$\|x - y\|^2 + \|z\|^2 - 2 \operatorname{Re} \langle x - y, z \rangle \geq \|x - y\|^2.$$

Rearrange,

$$2 \operatorname{Re}\langle x - y, z \rangle \leq \|z\|^2$$

for all $z \in V$. Now substitute tz for z where $t \in \mathbb{R}_+$, have

$$t \cdot 2 \operatorname{Re}\langle x - y, z \rangle \leq t^2 \|z\|^2.$$

For $t = 0$,

$$\operatorname{Re}\langle x - y, z \rangle \leq 0$$

Similarly replace z by $-z$ to conclude $\operatorname{Re}\langle x - y, z \rangle = 0$. Finally replace z by $e^{i\theta}z$ to have $\langle x - y, z \rangle = 0$ for all z . Thus $x - y \in V^\perp$. \square

Definition (bounded linear form). A linear form $\ell : H \rightarrow \mathbb{C}$ is *bounded* if there exists $C > 0$ such that $|\ell(x)| \leq C\|x\|$ for all $x \in H$.

Remark. ℓ bounded is equivalent to ℓ continuous.

Theorem 9.5 (Riesz representation theorem). *Let H be a Hilbert space. For every bounded linear form ℓ there exists $v_0 \in H$ such that*

$$\ell(x) = \langle x, v_0 \rangle$$

for all $x \in H$.

Proof. By boundedness of ℓ , $\ker \ell$ is closed so write

$$H = \ker \ell \oplus (\ker \ell)^\perp.$$

If $\ell = 0$ then just pick $v_0 = 0$. Otherwise pick $x_0 \in (\ker \ell)^\perp \setminus \{0\}$. But $(\ker \ell)^\perp$ is spanned by x_0 : indeed for any $x \in (\ker \ell)^\perp$,

$$\ell(x) = \frac{\ell(x)}{\ell(x_0)} \ell(x_0)$$

so

$$\ell\left(x - \frac{\ell(x)}{\ell(x_0)} x_0\right) = 0$$

so $x - \frac{\ell(x)}{\ell(x_0)} x_0 \in \ker \ell \cap (\ker \ell)^\perp = 0$. Now let

$$v_0 = \frac{\overline{\ell(x_0)}}{\|x_0\|^2} x_0$$

and observe that $\ell(x) - \langle x, v_0 \rangle$ vanishes on $\ker \ell$ and on $(\ker \ell)^\perp = \mathbb{C}x_0$. Thus it is identically zero. \square

Definition (absolutely continuous, singular measure). Let (X, \mathcal{A}) be a measurable space and let μ, ν be two measures on (X, \mathcal{A}) .

1. μ is *absolutely continuous* with respect to ν , denoted $\mu \ll \nu$, if for every $A \in \mathcal{A}$, $\nu(A) = 0$ implies $\mu(A) = 0$.
2. μ is *singular*, denoted $\mu \perp \nu$, if exists $\Omega \in \mathcal{A}$ such that $\mu(\Omega) = 0$ and

| $\nu(\Omega^c) = 0$.

Example.

1. Let ν be the Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $d\mu = f d\nu$ where $f \geq 0$ is a Borel function then $\mu \ll \nu$.
 2. If $\mu = \delta_{x_0}$, the Dirac mass at $x_0 \in \mathbb{R}$, then $\mu \perp \nu$.
- Non-examinable theorem and proof:

Theorem 9.6 (Radon-Nikodym). *Assume μ and ν are σ -finite measures on (X, \mathcal{A}) .*

1. *If $\mu \ll \nu$ then there exists $g \geq 0$ measurable such that $d\mu = g d\nu$, namely*

$$\mu(A) = \int_A g(x) d\nu(x)$$

for all $A \in \mathcal{A}$. g is called the density of μ with respect to ν or Radon-Nikodym derivative, sometimes denoted by $g = \frac{d\mu}{d\nu}$.

2. *For any μ, ν σ -finite, μ decomposes as*

$$\mu = \mu_a + \mu_s$$

where $\mu_a \ll \nu$ and $\mu_s \perp \nu$. Moreover this decomposition is unique.

Proof. Consider $H = L^2(X, \mathcal{A}, \mu + \nu)$, which is a Hilbert space. First assume μ and ν are finite. Consider the linear form

$$\ell : H \rightarrow \mathbb{R}$$

$$f \mapsto \mu(f) = \int_X f d\mu$$

ℓ is bounded by Cauchy-Schwarz and finiteness of the measures:

$$|\mu(f)| \leq \mu(|f|) \leq (\mu + \nu)(|f|) \leq \sqrt{(\mu + \nu)(X)} \cdot \sqrt{\int_X |f|^2 d(\mu + \nu)} = C \cdot \|f\|_H$$

so by Riesz representation theorem, there exists $g_0 \in L^2(X, \mathcal{A}, \mu + \nu)$ such that

$$\mu(f) = \int_X f g_0 d(\mu + \nu). \quad (*)$$

Claim that for $(\mu + \nu)$ -almost every x , $0 \leq g_0(x) \leq 1$: take $f = \mathbf{1}_{\{g_0 < 0\}}$ and plug it into $(*)$,

$$0 \leq \mu(\{g_0 < 0\}) = \int_X \underbrace{\mathbf{1}_{\{g_0 < 0\}} g_0}_{\leq 0} d(\mu + \nu) \leq 0$$

so equality throughout. Thus $g_0 \geq 0$ $(\mu + \nu)$ -almost everywhere. Similarly take $f = \mathbf{1}_{\{g_0 > 1 + \varepsilon\}}$ for $\varepsilon > 0$ and plug it into $(*)$,

$$\begin{aligned} (\mu + \nu)(\{g_0 > 1 + \varepsilon\}) &\geq \mu(\{g_0 > 1 + \varepsilon\}) \\ &= \int_X \mathbf{1}_{\{g_0 > 1 + \varepsilon\}} g_0 d(\mu + \nu) \\ &\geq (1 + \varepsilon)(\mu + \nu)(\{g_0 > 1 + \varepsilon\}) \end{aligned}$$

so must have

$$(\mu + \nu)(\{g_0 > 1 + \varepsilon\}) = 0$$

i.e. $g_0 \leq 1$ $(\mu + \nu)$ -almost everywhere.

Now set $\Omega = \{x \in X : g_0 \in [0, 1)\}$ so on Ω^c , $g_0 = 1$ $(\mu + \nu)$ -almost everywhere. Then (*) is equivalent to

$$\int_X f(1 - g_0)d\mu = \int_X fg_0d\nu$$

for all $f \in L^2(X, \mathcal{A}, \mu + \nu)$. Hence this holds for all $f \geq 0$. Now f be $\frac{f}{1-g_0}\mathbf{1}_\Omega$, get

$$\int_\Omega fd\mu = \int_\Omega f\frac{g_0}{1-g_0}d\nu. \quad (**)$$

Set

$$\begin{aligned} \mu_a(A) &= \mu(A \cap \Omega) \\ \mu_s(A) &= \mu(A \cap \Omega^c) \end{aligned}$$

Clearly $\mu = \mu_a + \mu_s$. Claim that this is the required result, i.e.

1. $\mu_a \ll \nu$,
2. $\mu_s \perp \nu$,
3. $d\mu_a = gd\nu$ where $g = \frac{g_0}{1-g_0}\mathbf{1}_\Omega$.

Proof.

1. If $\nu(A) = 0$ set $f = \mathbf{1}_A$ and plug into (**) to get $\mu(A \cap \Omega) = 0$, namely $\mu_a(A) = 0$.
2. Set $f = \mathbf{1}_{\Omega^c}$. On Ω^c , $g_0 = 1$ $(\mu + \nu)$ -almost everywhere. Plug this into (*) to get $\nu(\Omega^c) = 0$. But $\mu_s(\Omega) = 0$ so $\mu_s \perp \nu$.
3. (**) is equivalent to $d\mu_a = gd\nu$ where $g = \frac{g_0}{1-g_0}\mathbf{1}_\Omega$.

□

This settles part 2 of the theorem, and also part 1 as if $\mu \ll \nu$ then $\mu = \mu_a$.

If μ and ν are not finite but only σ -finite, use the old trick of partition X into countably many μ - and ν -finite sets and take their intersections. Suppose we get a disjoint countable union $X = \bigcup_n X_n$. Then $\mu = \sum_n \mu|_{X_n}$ where for each n we can write

$$\mu|_{X_n} = (\mu|_{X_n})_a + (\mu|_{X_n})_s.$$

Then set

$$\begin{aligned} \mu_a &= \sum_n (\mu|_{X_n})_a \\ \mu_s &= \sum_n (\mu|_{X_n})_s \end{aligned}$$

Remains to check uniqueness of decomposition. Suppose μ can be decomposed in two ways

$$\mu = \mu_a + \mu_s = \mu'_a + \mu'_s.$$

As $\mu_s, \mu'_s \perp \nu$ there exists $\Omega_0, \Omega'_0 \in \mathcal{A}$ such that

$$\begin{aligned} \mu_s(\Omega_0) &= 0, \nu(\Omega_0^c) = 0 \\ \mu'_s(\Omega'_0) &= 0, \nu((\Omega'_0)^c) = 0 \end{aligned}$$

Set $\Omega_1 = \Omega_0 \cap \Omega'_0$. Check that

$$\begin{aligned} \mu_s(\Omega_1) &= \mu'_s(\Omega_1) = 0 \\ \nu(\Omega_1^c) &= \nu(\Omega_0^c \cup \Omega_0 \cap \Omega_0'^c) = 0 \end{aligned}$$

Now $\mu_a, \mu'_a \ll \nu$ so

$$\mu_a(\Omega_1^c) = \mu'_a(\Omega_1^c) = 0.$$

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

$$\mu_a(A) = \mu_a(A \cap \Omega_1) = \mu(A \cap \Omega_1) = \mu'_a(A \cap \Omega_1) = \mu'_a(A)$$

so $\mu_a = \mu'_a$ and hence $\mu_s = \mu'_s$. □

Proposition 9.7. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let \mathcal{G} be a σ -subalgebra of \mathcal{F} and X a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Assume X is integrable, then there exists a random variable Y on $(\Omega, \mathcal{G}, \mathbb{P})$ such that*

$$\mathbb{E}(\mathbf{1}_A X) = \mathbb{E}(\mathbf{1}_A Y)$$

for all $A \in \mathcal{G}$. Moreover Y is unique almost surely.

If you are perplexed by why it is nontrivial to recover a random variable on \mathcal{G} from one on \mathcal{F} , as $\mathcal{G} \subseteq \mathcal{F}$ so it seems that we can easily restrict to a “sub-random variable”. But this reasoning makes absolutely no sense as random variables are functions from the space. In other words, $\text{id} : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\Omega, \mathcal{G}, \mathbb{P})$ is measurable but its inverse is not, and it is not obvious that X has a pushforward.

Definition (conditional expectation). Y as above is called the *conditional expectation* of X with respect to \mathcal{G} , denote by

$$Y = \mathbb{E}(X|\mathcal{G}).$$

Proof. wlog assume $X \geq 0$. Set $\mu(A) = \mathbb{E}(\mathbf{1}_A X)$ for all $A \in \mathcal{G}$. μ is finite by integrability of X and is a measure on (Ω, \mathcal{G}) . Moreover $\mu \ll \mathbb{P}$. Thus by Radon-Nikodym there exists $g \geq 0$ \mathcal{G} -measurable such that

$$\mu(A) = \int_A g d\mathbb{P} = \mathbb{E}(\mathbf{1}_A g).$$

Set $Y = g$.

Uniqueness is shown in example sheet 3. □

Remark. In case $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ then Y is the orthogonal projection of X onto $L^2(\Omega, \mathcal{G}, \mathbb{P})$. It is well-defined since $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a Hilbert space and $L^2(\Omega, \mathcal{G}, \mathbb{P})$ is a closed subspace. In this case TFAE:

1. $\mathbb{E}((X - Y)\mathbf{1}_A) = 0$ for all $A \in \mathcal{G}$,
2. $\mathbb{E}((X - Y)h) = 0$ for all h simple on $(\Omega, \mathcal{G}, \mathbb{P})$,
3. $\mathbb{E}((X - Y)h) = 0$ for all $h \geq 0$ \mathcal{G} -measurable,
4. $\mathbb{E}((X - Y)h) = 0$ for all $h \in L^2(\Omega, \mathcal{G}, \mathbb{P})$.

Remark. Special case when $\mathcal{G} = \{\emptyset, B, B^c, \Omega\}$ where $B \in \mathcal{F}$:

$$\mathbb{E}(\mathbf{1}_A | \mathcal{G})(\omega) = \begin{cases} \mathbb{P}(A|B) & \omega \in B \\ \mathbb{P}(A|B^c) & \omega \in B^c \end{cases}$$

where

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Proposition 9.8 (non-examinable). *Properties of conditional expectation:*

1. *linearity:* $\mathbb{E}(\alpha X + \beta Y | \mathcal{G}) = \alpha \mathbb{E}(X | \mathcal{G}) + \beta \mathbb{E}(Y | \mathcal{G})$.
2. *if X is \mathcal{G} -measurable then $\mathbb{E}(X | \mathcal{G}) = \mathbb{E}(X)$.*
3. *positivity: if $X \geq 0$ then $\mathbb{E}(X | \mathcal{G}) \geq 0$.*
4. $\mathbb{E}(\mathbb{E}(X | \mathcal{G}) | \mathcal{H}) = \mathbb{E}(X | \mathcal{H})$ if $\mathcal{H} \subseteq \mathcal{G}$.
5. *if Z is \mathcal{G} -measurable and bounded then $\mathbb{E}(XZ | \mathcal{G}) = Z \cdot \mathbb{E}(X | \mathcal{G})$.*

10 Fourier transform

Definition (Fourier transform). Let $f \in L^1(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), dx)$ where dx is the Lebesgue measure. The function

$$\begin{aligned} \hat{f} : \mathbb{R}^d &\rightarrow \mathbb{C} \\ u &\mapsto \int_{\mathbb{R}^d} f(x) e^{i\langle u, x \rangle} dx \end{aligned}$$

where $\langle u, x \rangle = u_1 x_1 + \dots + u_d x_d$, is called the *Fourier transform* of f .

Proposition 10.1.

1. $|\hat{f}(u)| \leq \|f\|_1$.
2. \hat{f} is continuous.

Proof. 1 is clear. 2 follows from dominated convergence theorem. \square

Definition. Given a finite Borel measure μ on \mathbb{R}^d , its *Fourier transform* is given by

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

Again $|\hat{\mu}(u)| \leq \mu(\mathbb{R}^d)$ and $\hat{\mu}$ is continuous.

Remark. If X is an \mathbb{R}^d -valued random variable with law μ_X then $\hat{\mu}_X$ is called the *characteristic function* of X .

Example.

1. Normalised Gaussian distribution on \mathbb{R} : $\mu = \mathcal{N}(0, 1)$. $d\mu = g dx$ where

$$g(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$

Claim that

$$\hat{\mu}(u) = \hat{g}(u) = e^{-u^2/2},$$

i.e. $\hat{g} = \sqrt{2\pi}g$. This is the defining characteristic of Gaussian distribution.

Proof. Since $(u \mapsto |iu \cdot e^{iu x} e^{-x^2/2}|) \in L^1(\mathbb{R})$, we can differentiate under

integral sign to get

$$\begin{aligned}
 \frac{d}{du} \hat{g}(u) &= \frac{d}{du} \int_{\mathbb{R}} e^{iux} e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} \\
 &= \int_{\mathbb{R}} ixe^{iux} e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} \\
 &= - \int_{\mathbb{R}} ie^{iux} g'(x) dx \\
 &= i \int_{\mathbb{R}} (e^{iux})' g(x) dx \\
 &= -u \int_{\mathbb{R}} e^{iux} g(x) dx \\
 &= -u \hat{g}(u)
 \end{aligned}$$

Thus

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

so

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

But

$$\hat{g}(0) = \int_{\mathbb{R}} g(x) dx = 1$$

so $\hat{g}(u) = e^{-u^2/2}$ as required. \square

2. d -dimensional version: $\mu = \mathcal{N}(0, I_d)$. $d\mu(x) = G(x)dx$ where $dx = dx_1 \cdots dx_d$ and

$$G(x) = \prod_{i=1}^d g(x_i) = \frac{e^{-\|x\|^2/2}}{(\sqrt{2\pi})^d}.$$

Then

$$\begin{aligned}
 \hat{G}(u) &= \int_{\mathbb{R}^d} e^{i(u,x)} G(x) dx \\
 &= \prod_{i=1}^d \int_{\mathbb{R}} g(x_i) e^{iu_i x_i} dx_i \\
 &= \prod_{i=1}^d \hat{g}(u_i) \\
 &= e^{-\|u\|^2/2}
 \end{aligned}$$

Theorem 10.2 (Fourier inversion formula).

1. If $f \in L^1(\mathbb{R}^d)$ is such that $\hat{f} \in L^1(\mathbb{R}^d)$ then f is continuous (i.e. f equals to a continuous function almost everywhere) and

$$f(x) = \frac{1}{(2\pi)^d} \hat{f}(-x).$$

2. If μ is a finite Borel measure on \mathbb{R}^d such that $\hat{\mu} \in L^1(\mathbb{R}^d)$ then μ has a continuous density with respect to Lebesgue measure, i.e. $d\mu = f dx$ with

$$f(x) = \frac{1}{(2\pi)^d} \hat{\mu}(-x).$$

Remark. In other words

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

where $\hat{f}(u)$ are *Fourier coefficients* and $e^{-i\langle u, x \rangle}$ are called *Fourier modes*, which are *characters* $e^{i\langle u, \cdot \rangle} : \mathbb{R}^d \rightarrow \{z \in \mathbb{C} : |z| = 1\}$. Informally this says that every f can be written as an “infinite linear combination” of Fourier modes.

Proof. (!UPDATE: 1 does not quite reduce to 2 as $\hat{f} = \hat{f}^+ - \hat{f}^-$ does not quite hold. Instead write $\hat{f} = a\hat{\mu}_X - b\hat{\mu}_Y$ where $a = \|f^+\|_1, d\mu_x = \frac{f}{\|f^+\|_1} dx$)

1 reduces to 2 by considering $f = f^+ - f^-$. In 2 we may assume wlog μ is a probability measure so is the law of some random variable X . Let $f(x) = \frac{1}{(2\pi)^d} \hat{\mu}(-x)$. We need to show $\mu = f dx$, which is equivalent to for all $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\mu(A) = \int_{\mathbb{R}^d} f \mathbf{1}_A dx.$$

Let $h = \mathbf{1}_A$ and wlog assume A is a bounded Borel set. The trick is to introduce an independent Gaussian random variable $N \sim \mathcal{N}(0, I_d)$ with law $G dx$. We have

$$\int_{\mathbb{R}^d} h(x) d\mu(x) = \mathbb{E}(h(X)) = \lim_{\sigma \rightarrow 0} \mathbb{E}(h(X + \sigma N))$$

by dominated convergence theorem. But

$$\begin{aligned} \mathbb{E}(h(X + \sigma N)) &= \mathbb{E} \left(\int_{\mathbb{R}^d} h(X + \sigma x) G(x) dx \right) \\ &= \mathbb{E} \left(\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} h(X + \sigma x) e^{i\langle u, x \rangle} G(u) \frac{du dx}{(\sqrt{2\pi})^d} \right) \end{aligned}$$

as

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

So by a change of variable $y = \sigma x$,

$$\begin{aligned} \mathbb{E}(h(X + \sigma N)) &= \mathbb{E} \left(\int \int h(X + y) e^{i\langle u, y/\sigma \rangle} G(u) \frac{du}{(\sqrt{2\pi})^d} \frac{dy}{\sigma^d} \right) \\ &= \mathbb{E} \left(\int \int h(z) e^{i\langle u/\sigma, z-x \rangle} G(u) \frac{du}{(\sqrt{2\pi}\sigma^2)^d} dz \right) \\ &= \int \int h(z) e^{i\langle u/\sigma, z \rangle} \hat{\mu}_X\left(-\frac{u}{\sigma}\right) G(u) \frac{du}{(\sqrt{2\pi}\sigma^2)^d} dz \quad \text{Tonelli-Fubini} \\ &= \frac{1}{(2\pi)^d} \int \int \hat{\mu}_X(u) e^{-i\langle u, z \rangle} h(z) e^{-\sigma^2 \|u\|^2/2} du dz \end{aligned}$$

□

We want a condition to ensure $\hat{f} \in L^1(\mathbb{R}^d)$. Clearly continuity is necessary. Here we use a generally principle in Fourier analysis: Fourier transform converts decay at infinity to smoothness. In Fourier inversion formula, if u is large then the Fourier character has fast oscillation. Thus if \hat{f} decays fast at infinity then the Fourier coefficients also decays fast, and the resulting transform is smoother.

Proposition 10.3. *If f, f' and f'' exists (for example if f is C^2) and are in L^1 then $\hat{f} \in L^1$.*

Proof. We prove the case $d = 1$. The general case follows from Tonelli-Fubini. We show first that $f, f' \in L^1$ implies that $\hat{f}(u) = \frac{i}{u} \hat{f}'(u)$. This easily follows from integration by parts:

$$\begin{aligned}\hat{f}(u) &= \int f(x)e^{iux} dx \\ &= \frac{1}{iu} \int f(x)(e^{iux})' dx \\ &= -\frac{1}{iu} \int f'(x)e^{iux} dx\end{aligned}$$

so in particular $|\hat{f}(u)| \leq \frac{1}{|u|} \|f'\|_1$.

Thus if $f, f', f'' \in L^1$ then $\hat{f}(u) = -\frac{1}{u^2} \hat{f}''(u)$ so

$$|\hat{f}(u)| \leq \frac{\|f''\|_1}{|u^2|}.$$

As $\int_1^\infty \frac{1}{|u|^2} du < \infty$, $\hat{f} \in L^1$. □

Definition (convolution). Given two Borel measures μ and ν on \mathbb{R}^d , we define their *convolution* $\mu * \nu$ as the image of $\mu \otimes \nu$ under the addition map

$$\begin{aligned}\Phi : \mathbb{R}^d \times \mathbb{R}^d &\rightarrow \mathbb{R}^d \\ (x, y) &\mapsto x + y\end{aligned}$$

i.e. $\mu * \nu = \Phi_*(\mu \otimes \nu)$

Thus given $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\Phi_*(\mu \otimes \nu)(A) = \mu \otimes \nu(\{(x, y) : x + y \in A\}).$$

Example. Given X, Y independent random variables and μ, ν be laws of X, Y respectively, then $\mu * \nu$ is the law of $X + Y$.

Definition (convolution). If $f, g \in L^1(\mathbb{R}^d)$ define their *convolution* $f * g$ by

$$(f * g)(x) = \int_{\mathbb{R}^d} f(x-t)g(t)dt.$$

This is well defined by Fubini: $f, g \in L^1$ so

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |f(x-t)g(t)| dt dx < \infty$$

and

$$\|f * g\|_1 = \int \left| \int f(x-t)g(t) dt \right| dx \leq \int \int |f(x-t)g(t)| dt dx \leq \|f\|_1 \cdot \|g\|_1$$

Therefore $(L^1(\mathbb{R}^d), *)$ forms a *Banach algebra*.

Remark. If μ, ν are two finite Borel measures on \mathbb{R}^d and if $\mu, \nu \ll dx$, i.e. absolutely continuous, then by Radon-Nikodym there exist $f, g \in L^1(\mathbb{R}^d)$ such that

$$\begin{aligned} d\mu &= f dx \\ d\nu &= g dx \end{aligned}$$

then $\mu * \nu \ll dx$ and

$$d(\mu * \nu) = (f * g) dx.$$

Proposition 10.4 (Gaussian approximation). *If $f \in L^p(\mathbb{R}^d)$ where $p \in [1, \infty)$ then*

$$\lim_{\sigma \rightarrow 0} \|f * G_\sigma - f\|_p = 0$$

where $G_\sigma = \mathcal{N}(0, \sigma^2 I_d)$, i.e.

$$G_\sigma(x) = \frac{1}{(\sqrt{2\pi\sigma^2})^d} e^{-\frac{|x|^2}{2\sigma^2}}.$$

Lemma 10.5 (continuity of translation in L^p). *Suppose $p \in [1, \infty)$ and $f \in L^p$. Then*

$$\lim_{t \rightarrow 0} \|\tau_t(f) - f\|_p = 0$$

where $\tau_t(f)(x) = f(x+t)$, $t \in \mathbb{R}^d$.

Proof. Example sheet. □

Proof of Gaussian approximation.

$$(f * G_\sigma - f)(x) = \int_{\mathbb{R}^d} G_\sigma(t)(f(x-t) - f(x)) dt = \mathbb{E}(f(x - \sigma N) - f(x))$$

where $N \sim \mathcal{N}(0, I_d)$ is Gaussian with density G_1 . Then

$$\|f * G_\sigma - f\|_p^p \leq \mathbb{E}(\|f(x + \sigma N) - f(x)\|_p^p) = \mathbb{E}(\|\tau_{\sigma N}(f) - f\|_p^p)$$

by Jensen's inequality and convexity of $x \mapsto x^p$. By the lemma,

$$\lim_{\sigma \rightarrow 0} \|\tau_{\sigma N}(f) - f\|_p = 0.$$

As $\|\tau_{\sigma N}(f) - f\|_p \leq 2\|f\|_p$, apply dominated convergence theorem to get the required result. □

Proposition 10.6.

- If $f, g \in L^1(\mathbb{R}^d)$ then

$$\widehat{f * g} = \hat{f} \cdot \hat{g}.$$

- If μ, ν are finite Borel measure then

$$\widehat{\mu * \nu} = \hat{\mu} \cdot \hat{\nu}.$$

Proof. 1 reduces to 2 by writing

$$f(x)dx = f^+(x)dx - f^-(x)dx = a d\mu - b d\nu$$

for some probability measure μ, ν .

wlog we may assume μ and ν are laws of independent random variables X and Y . Then by a previous result $\mu * \nu$ is just the law of $X + Y$ so

$$\begin{aligned} \widehat{\mu * \nu}(u) &= \int e^{i\langle u, x+y \rangle} d\mu(x) d\nu(y) \\ &= \mathbb{E}(e^{i\langle u, X+Y \rangle}) = \mathbb{E}(e^{i\langle u, X \rangle} e^{i\langle u, Y \rangle}) \quad \text{homomorphism} \\ &= \mathbb{E}(e^{i\langle u, X \rangle}) \mathbb{E}(e^{i\langle u, Y \rangle}) \quad \text{as } X, Y \text{ are independent} \\ &= \hat{\mu}(u) \cdot \hat{\nu}(u). \end{aligned}$$

In short, this is precisely because $e^{i\langle u, - \rangle}$ are characters. \square

Theorem 10.7 (Lévy criterion). *Let $(X_n)_{n \geq 1}$ and X be \mathbb{R}^d -valued random variables. Then TFAE:*

1. $X_n \rightarrow X$ in law,
2. For all $u \in \mathbb{R}^d$, $\hat{\mu}_{X_n}(u) \rightarrow \hat{\mu}_X(u)$.

In particular if $\hat{\mu}_X = \hat{\mu}_Y$ for two random variables X and Y then $X = Y$ in law, i.e. $\mu_X = \mu_Y$.

Thus Fourier transform is an injection from Borel measure to certain function space.

Proof.

- 1 \implies 2: Clear by definition as $f(x) = e^{i\langle u, x \rangle}$ is continuous and bounded for all $u \in \mathbb{R}^d$.
- 2 \implies 1: Need to show that for all $g \in C_b(\mathbb{R}^d)$,

$$\mathbb{E}(g(X_n)) \rightarrow \mathbb{E}(g(X)).$$

wlog it's enough to check this for all $g \in C_c^\infty(\mathbb{R}^d)$. For the sufficiency see example sheet.

Note that for all $g \in C_c^\infty(\mathbb{R}^d)$, $\hat{g} \in L^1$ so by Fourier inversion formula

$$g(x) = \int \hat{g}(u) e^{-i\langle u, x \rangle} \frac{du}{(2\pi)^d}.$$

Hence

$$\begin{aligned}
\mathbb{E}(g(X_n)) &= \int \hat{g}(u) \mathbb{E}(e^{-i\langle u, X_n \rangle}) du \\
&= \int \hat{g}(u) \hat{\mu}_{X_n}(-u) \frac{du}{(2\pi)^d} \\
&\rightarrow \int \hat{g}(u) \hat{m}_X(-u) \frac{du}{(2\pi)^d} \\
&= \mathbb{E}(g(X))
\end{aligned}$$

by dominated convergence theorem. □

Theorem 10.8 (Plancherel formula).

1. If $f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ then $\hat{f} \in L^2(\mathbb{R}^d)$ and

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

2. If $f, g \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ then

$$\langle \hat{f}, \hat{g} \rangle_{L^2} = (2\pi)^d \langle f, g \rangle_{L^2}.$$

3. The Fourier transform

$$\begin{aligned}
\mathcal{F} : L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) &\rightarrow L^2(\mathbb{R}^d) \\
f &\mapsto \frac{1}{(\sqrt{2\pi})^d} \hat{f}
\end{aligned}$$

extends uniquely to a linear operator on $L^2(\mathbb{R}^d)$ which is an isometry.

Moreover

$$\mathcal{F} \circ \mathcal{F}(f) = \check{f}$$

where $\check{f}(x) = f(-x)$, for all $f \in L^2(\mathbb{R}^d)$.

Proof. First we prove 1 and 2 assuming $\hat{f}, \hat{g} \in L^1(\mathbb{R}^d)$. By Fourier inversion formula,

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

and in particular $\hat{f} \in L^2(\mathbb{R}^d)$.

Similarly for 2,

$$\begin{aligned} \langle \hat{f}, \hat{g} \rangle_{L^2} &= \int \hat{f}(u) \overline{\hat{g}(u)} du \\ &= \int \left(\int f(x) e^{i\langle u, x \rangle} dx \right) \overline{\hat{g}(u)} du \\ &= \int \int f(x) \overline{\hat{g}(u) e^{-i\langle u, x \rangle}} du dx \\ &= \int f(x) \overline{g(x)} dx (2\pi)^d \\ &= (2\pi)^d \langle f, g \rangle_{L^2} \end{aligned}$$

Now for the general case we use Gaussian as a mollifier. Consider

$$\begin{aligned} f_\sigma &= f * G_\sigma \\ g_\sigma &= g * G_\sigma \end{aligned}$$

and based on results and computations before,

$$\hat{f}_\sigma = \hat{f} \cdot \hat{G}_\sigma = \hat{f} e^{-\sigma \|u\|^2/2}.$$

As $\|\hat{f}\|_\infty \leq \|f\|_1$, $\hat{f}_\sigma \in L^1(\mathbb{R}^d)$. Thus $\hat{f}_\sigma \in L^2(\mathbb{R}^d)$ and $\|\hat{f}_\sigma\|_2^2 = (2\pi)^d \|f_\sigma\|_2^2$. But by Gaussian approximation we know that $f_\sigma \rightarrow f$ in $L^2(\mathbb{R}^d)$ as $\sigma \rightarrow 0$. Hence $\|f_\sigma\|_2 \rightarrow \|f\|_2$. Then

$$\|\hat{f}_\sigma\|_2^2 = \|\hat{f} \cdot \hat{G}_\sigma\|_2^2 = \int |\hat{f}(u)|^2 e^{-\sigma \|u\|^2} du \rightarrow \|f\|_2^2$$

as $\sigma \rightarrow 0$ by monotone convergence theorem. Thus

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

For 2,

$$\langle \hat{f}_\sigma, \hat{g}_\sigma \rangle = \int \hat{f} \hat{g} e^{-\sigma \|u\|^2} du \rightarrow \int \hat{f} \hat{g} du$$

as $\sigma \rightarrow 0$ by dominated convergence theorem as $\hat{f} \hat{g} \in L^1$. The result follows from Gaussian approximation.

For 3, $L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ is dense in $L^2(\mathbb{R}^d)$ because it contains $C_c(\mathbb{R}^d)$. Then extend by completeness: given $f \in L^2(\mathbb{R}^d)$, pick a sequence $f_n \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ such that $f_n \rightarrow f$ in $L_2(\mathbb{R}^d)$. Then define

$$\mathcal{F}f = \lim_{n \rightarrow \infty} \mathcal{F}f_n.$$

The limit exists as $L^2(\mathbb{R}^d)$ is complete. \mathcal{F} is well-defined as

$$\|\mathcal{F}f_n - \mathcal{F}f_m\|_2 = \|f_n - f_m\|_2$$

by 1. Finally,

$$\|\mathcal{F}f\|_2 = \|f\|_2$$

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

$$\mathcal{F} \circ \mathcal{F}(f) = \check{f}$$

for all f such that $f, \hat{f} \in L^1(\mathbb{R}^d)$. Thus by continuity this holds for $L^2(\mathbb{R}^d)$. \square

11 Gaussians

Definition (Gaussian). An \mathbb{R}^d -valued random variable X is called *Gaussian* if for all $u \in \mathbb{R}^d$, $\langle X, u \rangle$ is Gaussian, namely its law has the form $\mathcal{N}(m, \sigma^2)$ for some $m \in \mathbb{R}, \sigma > 0$.

Proposition 11.1. *The law of a Gaussian vector $X = (X_1, \dots, X_d) \in \mathbb{R}^d$ is uniquely determined by*

1. its mean $\mathbb{E}X = (\mathbb{E}X_1, \dots, \mathbb{E}X_d)$,
2. its covariance matrix $(\text{Cov}(X_i, X_j))_{ij}$ where

$$\text{Cov}(X_i, X_j) = \mathbb{E}((X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)).$$

Proof. If $d = 1$ then this just says that it is determined by its mean m and covariance σ^2 , which is obviously true. For $d > 1$, compute the characteristic function

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

but by assumption $\langle X, u \rangle$ is Gaussian in $d = 1$ so its law is determined

1. the mean $\mathbb{E}\langle X, u \rangle = \langle \mathbb{E}X, u \rangle$,
2. the variance $\text{Var}\langle X, u \rangle$. But

$$\text{Var}\langle X, u \rangle = \mathbb{E}((\langle X, u \rangle - \mathbb{E}\langle X, u \rangle)^2) = \sum_{i,j} u_i u_j \text{Cov}(X_i, X_j).$$

□

In particular this shows that $(\text{Cov}(X_i, X_j))_{ij}$ is a non-negative semidefinite symmetric matrix.

Proposition 11.2. *If X is a Gaussian vector then exists $A \in \mathcal{M}_d(\mathbb{R}), b \in \mathbb{R}^d$ such that X has the same law as $AN + b$ where $N = (N_1, \dots, N_d)$, $(N_i)_{i=1}^d$ are iid. $\mathcal{N}(0, 1)$.*

Proof. Take A such that

$$AA^* = (\text{Cov}(X_i, X_j))_{ij}$$

where A^* is the adjoint/transpose of A , and

$$b = (\mathbb{E}X_1, \dots, \mathbb{E}X_d).$$

Check that for all $u \in \mathbb{R}^d$,

$$\begin{aligned} \mathbb{E}\langle X, u \rangle &= \langle b, u \rangle \\ \text{Var}\langle X, u \rangle &= \langle AA^*u, u \rangle = \|A^*u\|_2^2 = \text{Var}\langle AN + b, u \rangle \end{aligned}$$

□

Proposition 11.3. *If (X_1, \dots, X_d) is a Gaussian vector then TFAE:*

1. X_i 's are independent.
2. X_i 's are pairwise independent.
3. $(\text{Cov}(X_i, X_j))_{ij}$ is a diagonal matrix.

Proof. 1 \implies 2 \implies 3 is obvious. 3 \implies 1 as we can choose A to be diagonal. Thus X has the same law as $(a_1 N_1, \dots, a_d N_d) + b$. \square

Theorem 11.4 (central limit theorem). *Let $(X_i)_{i \geq 1}$ be \mathbb{R}^d -valued iid. random variables with law μ . Assume they have second moment, i.e. $\mathbb{E}(\|X_1\|^2) < \infty$. Let $\mathbf{m} = \mathbb{E}(X_1) \in \mathbb{R}^d$ and*

$$Y_n = \frac{X_1 + \dots + X_n - n \cdot \mathbf{m}}{\sqrt{n}}.$$

Then Y_n converges in law to a central Gaussian on \mathbb{R}^d with law $\mathcal{N}(0, K)$ where

$$K_{ij} = (\text{Cov}(X_1))_{ij} = \left[\int_{\mathbb{R}^d} (x_i - \mathbb{E}(X_1))(x_j - \mathbb{E}(X_1)) d\mu(x) \right]_{ij}.$$

Proof. The proof is an application of Lévy criterion. Need to show $\hat{\mu}_{Y_n}(u) \rightarrow \hat{\mu}_Y(u)$ as $n \rightarrow \infty$ for all u , where $Y \sim \mathcal{N}(0, K)$. As

$$\hat{\mu}_{Y_n}(u) = \mathbb{E}(e^{i\langle Y_n, u \rangle}),$$

this is equivalent to show that for all u , $\langle Y_n, u \rangle$ converges in law to $\langle Y, u \rangle$. But

$$\langle Y_n, u \rangle = \frac{\langle X_1, u \rangle + \dots + \langle X_n, u \rangle - n\langle \mathbf{m}, u \rangle}{\sqrt{n}}$$

so we reduce the problem to 1-dimension case. By rescaling wlog $\mathbb{E}(X_1) = 0$, $\mathbb{E}(X_1^2) = 1$.

Now

$$\begin{aligned} \hat{\mu}_{Y_n}(u) &= \mathbb{E}(e^{iuY_n}) \\ &= \mathbb{E}(\exp(iu \frac{X_1 + \dots + X_n}{\sqrt{n}})) \\ &= \prod_{i=1}^n \mathbb{E}(\exp(iu \frac{X_i}{\sqrt{n}})) \\ &= (\mathbb{E}(\exp(iu \frac{X_1}{\sqrt{n}})))^n \\ &= (\hat{\mu}(\frac{u}{\sqrt{n}}))^n \end{aligned}$$

But $\mathbb{E}(X_1) = 0, \mathbb{E}(X_1^2) = 1$ so we can differentiate $\hat{\mu}$ under the integral sign

$$\begin{aligned}\hat{\mu}(u) &= \int_{\mathbb{R}} e^{iux} d\mu(x) \\ \frac{d}{du} \hat{\mu}(u) &= \int_{\mathbb{R}} ix e^{iux} d\mu(x) = i\mathbb{E}(X_1) \\ \frac{d^2}{du^2} \hat{\mu}(u) &= \int_{\mathbb{R}} -x^2 e^{iux} d\mu(x) = -\mathbb{E}(X_1^2)\end{aligned}$$

Taylor expand $\hat{\mu}$ around 0 to 2nd order,

$$\begin{aligned}\hat{\mu}(u) &= \hat{\mu}(0) + u\hat{\mu}'(0) + \frac{u^2}{2}\hat{\mu}''(0) + o(u^2) \\ &= 1 + 0 \cdot u - \frac{u^2}{2} + o(u^2)\end{aligned}$$

so

$$\hat{\mu}_{Y_n}(u) = \left(1 - \frac{u^2}{2n} + o\left(\frac{u^2}{n}\right)\right)^n \rightarrow e^{-u^2/2} = \hat{g}(u)$$

as $n \rightarrow \infty$ where g is the law of Y . □

12 Ergodic theory

Let (X, \mathcal{A}, μ) be a measure space. Let $T : X \rightarrow X$ be an \mathcal{A} -measurable map. We are interested in the trajectories of $T^n x$ for $n \geq 0$ and their statistical behaviour. In particular we are interested in those T preserving measure μ .

Definition (measure-preserving). $T : X \rightarrow X$ is *measure-preserving* if $T_*\mu = \mu$. (X, \mathcal{A}, μ, T) is called a *measure-preserving dynamical system*.

Definition (invariant function, invariant set, invariant σ -algebra).

- A measurable function $f : X \rightarrow \mathbb{R}$ is called *T-invariant* if $f = f \circ T$.
- A set $A \in \mathcal{A}$ is *T-invariant* if $\mathbf{1}_A$ is *T-invariant*.
-

$$\mathcal{I} = \{A \in \mathcal{A} : A \text{ is } T\text{-invariant}\}$$

is called the *T-invariant σ -algebra*.

Lemma 12.1. f is *T-invariant* if and only if f is \mathcal{I} -measurable.

Proof. Indeed for all $t \in \mathbb{R}$,

$$\{x \in X : f(x) < t\} = \{x \in X : f \circ T(x) < t\} = T^{-1}(\{x \in X : f(x) < t\}).$$

□

Definition (ergodic). T is *ergodic* with respect to μ , or that μ is *ergodic* with respect to T , if for all $A \in \mathcal{I}$, $\mu(A) = 0$ or $\mu(A^c) = 0$.

This condition asserts that \mathcal{I} is *trivial*, i.e. its elements are either null or conull.

Lemma 12.2. T is *ergodic* with respect to μ if and only if every invariant function f is almost everywhere constant.

Proof. Exercise. □

Example.

1. Let X be a finite space, $T : X \rightarrow X$ a map and $\mu = \#$ the counting measure, then T is measure preserving is equivalent to T being a bijection, and T is ergodic is equivalent to there does not exist a partition $X = X_1 \cup X_2$ such that both X_1 and X_2 are T -invariant, which is equivalent to for all $x, y \in X$, there exists n such that $T^n x = y$.
2. Let $X = \mathbb{R}^d / \mathbb{Z}^d$, \mathcal{A} the Borel σ -algebra and μ the Lebesgue measure. Given $a \in \mathbb{R}^d$, translation $T_a : x \mapsto x + a$ is measure-preserving. T_a is ergodic with respect to μ if and only if $(1, a_1, \dots, a_d)$, where a_i 's are coordinates of a , are linearly independent. See example sheet. (hint: Fourier transform)

3. Let $X = \mathbb{R}/\mathbb{Z}$ and again \mathcal{A} Borel σ -algebra and μ the Lebesgue measure. The doubling map $T : x \mapsto 2x - [2x]$ is ergodic with respect to μ . (hint: again consider Fourier coefficients). Intuitively in the graph of this function the preimage ε is two segments each of length $\varepsilon/2$, so measure-preserving.
4. Furstenberg conjecture: every ergodic measure μ on \mathbb{R}/\mathbb{Z} invariant under T_2, T_3 must be either Lebesgue or finitely supported.

12.1 The canonical model

Let $(X_n)_{n \geq 1}$ be an \mathbb{R}^d -valued stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $X = (\mathbb{R}^d)^{\mathbb{N}}$ and define the *sample path map*

$$\begin{aligned} \Phi : \Omega &\rightarrow X \\ \omega &\mapsto (X_n(\omega))_{n \geq 1} \end{aligned}$$

Let

$$\begin{aligned} T : X &\rightarrow X \\ (x_n)_{n \geq 1} &\mapsto (x_{n+1})_{n \geq 1} \end{aligned}$$

be the *shift map*. Let $x_n : X \rightarrow \mathbb{R}^d$ be the n th coordinate function and let $\mathcal{A} = \sigma(x_n : n \geq 1)$.

Note. \mathcal{A} is the infinite product σ -algebra $\mathcal{B}(\mathbb{R}^d)^{\otimes \mathbb{N}}$ of $\mathcal{B}(\mathbb{R}^d)^{\mathbb{N}}$.

Let $\mu = \Phi_* \mathbb{P}$, a probability measure on (X, \mathcal{A}) . This μ is called the *law* of the process $(X_n)_{n \geq 1}$. Now (X, \mathcal{A}, μ, T) is called the *canonical model* associated to $(X_n)_{n \geq 1}$.

Proposition 12.3 (stationary process). *TFAE:*

1. (X, \mathcal{A}, μ, T) is measure-preserving.
2. For all $k \geq 1$, the law of $(X_n, X_{n+1}, \dots, X_{n+k})$ on $(\mathbb{R}^d)^k$ is independent of n .

In this case we say that $(X_n)_{n \geq 1}$ is a stationary process.

Proof.

- $1 \implies 2$: $\mu = T_* \mu$ implies $\mu = T_*^n \mu$ for all n and this says law of $(X_i)_{i \geq 1}$ is the same as that of $(X_{i+n})_{i \geq 1}$.
- $1 \longleftarrow 2$: μ and $T_* \mu$ agree on cylinders $A \times (\mathbb{R}^d)^{\mathbb{N} \setminus F}$ for $F \subseteq \mathbb{N}$ finite, $A \in \mathcal{B}((\mathbb{R}^d)^F)$.

□

In some sense ergodic system is the study of stationary process.

Proposition 12.4 (Bernoulli shift). *If $(X_n)_{n \geq 1}$ are iid. then (X, \mathcal{A}, μ, T) is ergodic. It is called the Bernoulli shift associated to the law ν of X_1 . We have*

$$\mu = \nu^{\otimes \mathbb{N}}.$$

Proof. Claim that $\Phi^{-1}(\mathcal{T}) \subseteq \mathcal{C}$, the tail σ -algebra of $(X_n)_{n \geq 1}$. But Kolmogorov 0-1 law says that if $A \in \mathcal{T}$ then $\mathbb{P}(\Phi^{-1}(A)) = 0$ or 1 , so $\mu(A) = 0$ or 1 , thus μ is T -ergodic.

Given $A \in \mathcal{T}$, $T^{-1}A = A$ so

$$\begin{aligned} \Phi^{-1}(A) &= \{\omega \in \Omega : (X_n(\omega))_{n \geq 1} \in A\} \\ &= \{\omega \in \Omega : (X_n(\omega))_{n \geq 1} \in T^{-1}A\} \\ &= \{\omega \in \Omega : (X_{n+1}(\omega))_{n \geq 1} \in A\} \\ &= \{\omega \in \Omega : (X_{n+k}(\omega))_{n \geq 1} \in A\} \quad \text{for all } k \\ &\in \sigma(X_k, X_{k+1}, \dots) \end{aligned}$$

for μ -almost every x . □

Theorem 12.5 (von Neumann mean ergodic theorem). *Let (X, \mathcal{A}, μ, T) be a measure-preserving system. Let $f \in L^2(X, \mathcal{A}, \mu)$. Then the ergodic average*

$$S_n f = \frac{1}{n} \sum_{i=0}^{n-1} f \circ T^i$$

converges in L^2 to \bar{f} , a T -invariant function. In fact \bar{f} is the orthogonal projection of f onto $L^2(X, \mathcal{T}, \mu)$.

The intuition is as follow: if f is the indicator function of a set, then $S_n f$ is exactly the average of the time each orbit spending in A .

Proof. Hilbert space argument. Let $H = L^2(X, \mathcal{A}, \mu)$ and define

$$\begin{aligned} U : H &\rightarrow H \\ f &\mapsto f \circ T \end{aligned}$$

which is an isometry: because μ is T -invariant, $\int |f \circ T|^2 d\mu = \int |f|^2 d\mu$. Then by Riesz representation theorem it has an adjoint

$$\begin{aligned} U^* : H &\rightarrow H \\ x &\mapsto U^* x \end{aligned}$$

which satisfies $\langle U^* x, y \rangle = \langle x, Uy \rangle$ for all $y \in H$. Let

$$W = \{\varphi - \varphi \circ T : \varphi \in H\}$$

be the *coboundaries*. Let $f \in W$. Then

$$S_n f = \frac{1}{n} \sum_{i=0}^{n-1} (\varphi \circ T^i - \varphi \circ T^{i+1}) = \frac{\varphi - \varphi \circ T^n}{n} \rightarrow 0$$

as $n \rightarrow \infty$.

Let $f \in \overline{W}$ then again $S_n f \rightarrow 0$ because for all ε exists $g \in W$ such that $\|f - g\| < \varepsilon$. Then

$$\|S_n f - S_n g\| = \|S_n(f - g)\| \leq \|f - g\| \leq \varepsilon$$

so $\limsup_n \|S_n f\| \leq \varepsilon$.

Have $H = \overline{W} \oplus \overline{W}^\perp$ and $\overline{W}^\perp = W^\perp$. Claim W^\perp is exactly the T -invariant functions. The theorem then follows because if $f \circ T = f$ then $S_n f = f$ for all f .

Proof of claim.

$$\begin{aligned} W^\perp &= \{g \in H : \langle g, \varphi - U\varphi \rangle = 0 \text{ for all } \varphi \in H\} \\ &= \{g : \langle g, \varphi \rangle = \langle g, U\varphi \rangle \text{ for all } \varphi\} \\ &= \{g : \langle g, \varphi \rangle = \langle U^*g, \varphi \rangle \text{ for all } \varphi\} \\ &= \{g : U^*g = g\} \\ &= \{g : Ug = g\} \end{aligned}$$

where the last equality is by

$$\|Ug - g\|^2 = 2\|g\|^2 - 2\operatorname{Re}\langle g, Ug \rangle = 2\|g\|^2 - 2\operatorname{Re}\langle U^*g, g \rangle,$$

and this shows that W^\perp are exactly T -invariant functions. □

□

In fact we can do better:

Theorem 12.6 (Birkhoff (pointwise) ergodic theorem). *Let (X, \mathcal{A}, μ, T) be a measure-preserving system. Assume μ is finite (actually σ -finite suffices) and let $f \in L^1(X, \mathcal{A}, \mu)$. Then*

$$S_n f = \frac{1}{n} \sum_{i=0}^{n-1} f \circ T^i$$

converges μ -almost everywhere to a T -invariant function $\bar{f} \in L^1$. Moreover $S_n f \rightarrow \bar{f}$ in L^1 .

Corollary 12.7 (strong law of large numbers). *Let $(X_n)_{n \geq 1}$ be a sequence of iid. random variables. Assume $\mathbb{E}(|X_1|) < \infty$. Let*

$$S_n = \sum_{k=1}^n X_k,$$

then $\frac{1}{n} S_n$ converges almost surely to $\mathbb{E}(X_1)$.

Proof. Let (X, \mathcal{A}, μ, T) be the canonical model associated to $(X_n)_{n \geq 1}$, where $X = \mathbb{R}^\mathbb{N}$, $\mathcal{A} = \mathcal{B}(\mathbb{R})^{\otimes \mathbb{N}}$, T the shift operator and $\mu = \nu^{\otimes \mathbb{N}}$ where ν is the law of X_1 . It is a Bernoulli shift. Let

$$\begin{aligned} f : X &\rightarrow \mathbb{R} \\ x &\mapsto x_1 \end{aligned}$$

the first coordinate. Then $f \circ T^i(x) = x_{i+1}$ so

$$\frac{1}{n}(X_1 + \dots + X_n)(\omega) = S_n f(x)$$

where $x = (X_n(\omega))_{n \geq 1}$. Hence by Birkhoff ergodic theorem

$$\frac{1}{n}(X_1 + \dots + X_n) \rightarrow \bar{f} = \int f d\mu = \int x_1 d\mu = \mathbb{E}(X_1)$$

almost surely. \square

Remark. If T is ergodic then \bar{f} is almost everywhere constant. Hence $\bar{f} = \int f d\mu$.

Lemma 12.8 (maximal ergodic lemma). *Let $f \in L^1(X, \mathcal{A}, \mu)$ and $\alpha \in \mathbb{R}$.*

Let

$$E_\alpha = \{x \in X : \sup_{n \geq 1} S_n f(x) > \alpha\}$$

then

$$\alpha \mu(E_\alpha) \leq \int_{E_\alpha} f d\mu.$$

Lemma 12.9 (maximal inequality). *Let*

$$f_0 = 0$$

$$f_n = n S_n f = \sum_{i=0}^{n-1} f \circ T^i \quad n \geq 1$$

Let

$$P_N = \{x \in X : \max_{0 \leq n \leq N} f_n(x) > 0\}.$$

Then

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

Proof of maximal inequality. Set $F_N = \max_{0 \leq n \leq N} f_n$. Observe that for all $n \leq N$, $F_N \geq f_n$ and hence

$$F_N \circ T + f \geq f_n \circ T + f = f_{n+1}.$$

Now if $x \in P_n$ then

$$F_N(x) \leq \max_{0 \leq n \leq N} f_{n+1} \leq F_N \circ T + f$$

Integrate to get

$$\int_{P_N} F_N d\mu \leq \int_{P_N} F_N \circ T d\mu + \int_{P_N} f d\mu.$$

Note that $F_N(x) = 0$ if $x \notin P_N$ because $f_0 = 0$ so

$$\int_{P_N} F_N = \int_X F_N \leq \int_X F_N \circ T + \int_{P_N} f.$$

As μ is T -invariant, $\int F_N \circ T = \int F_N$ so

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

□

Proof of maximal ergodic lemma. Apply the maximal inequality to $g = f - \alpha$. Observe that

$$E_\alpha(f) = \bigcup_{N \geq 1} P_N(g)$$

and $S_n g = S_n f - \alpha$. Thus

$$\int_{E_\alpha(f)} (f - \alpha) d\mu \geq 0,$$

which is equivalent to

$$\alpha \mu(E_\alpha) \leq \int_{E_\alpha} f d\mu.$$

□

Proof of Birkhoff (pointwise) ergodic theorem. Let

$$\begin{aligned} \bar{f} &= \limsup_n S_n f \\ \underline{f} &= \liminf_n S_n f \end{aligned}$$

Observe that $\bar{f} = \bar{f} \circ T$, $\underline{f} = \underline{f} \circ T$: indeed

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

Need to show that $\bar{f} = \underline{f}$ μ -almost everywhere. This is equivalent to for all $\alpha, \beta \in \mathbb{Q}$, $\alpha > \beta$, the set

$$E_{\alpha, \beta}(f) = \{x \in X : \underline{f}(x) < \beta, \bar{f}(x) > \alpha\}$$

is μ -null, as then

$$\{x : \bar{f}(x) \neq \underline{f}(x)\} = \bigcup_{\alpha > \beta} E_{\alpha, \beta}$$

is μ -null by subadditivity. Observe that $E_{\alpha, \beta}$ is T -invariant. Apply the maximal ergodic theorem to $E_{\alpha, \beta}$ to get

$$\alpha \mu(E_{\alpha, \beta}(f)) \leq \int_{E_{\alpha, \beta}} f d\mu.$$

Dually

$$-\beta \mu(E_{\alpha, \beta}(f)) \leq \int_{E_{\alpha, \beta}} -f d\mu$$

so

$$\alpha\mu(E_{\alpha,\beta}) \leq \beta\mu(E_{\alpha,\beta}).$$

But $\alpha > \beta$ so $\mu(E_{\alpha,\beta}) = 0$.

We have proved that the limit $\lim_n S_n f$ exists almost everywhere, which we now define to be \bar{f} , and left to show $\bar{f} \in L^1$ and $\lim_n \|S_n f - \bar{f}\|_1 = 0$. This is an application of Fatou's lemma:

$$\begin{aligned} \int |\bar{f}| d\mu &= \int \liminf_n |S_n f| d\mu \\ &\leq \liminf_n \int |S_n f| d\mu \\ &\leq \liminf_n \|S_n f\|_1 \\ &\leq \|f\|_1 \end{aligned}$$

so $\bar{f} \in L^1$ where the last inequality is because

$$\|S_n f\|_1 \leq \frac{1}{n} (\|f\|_1 + \dots + \|f \circ T^{n-1}\|_1) = \|f\|_1.$$

Now to show $\|S_n f - \bar{f}\|_1 \rightarrow 0$, we truncate f . Let $M > 0$ and set $\varphi_M = f \mathbf{1}_{|f| < M}$. Note that

- $|\varphi_M| \leq M$ so $|S_n \varphi_M| \leq M$. Hence by dominated convergence theorem $\|S_n \varphi_M - \bar{\varphi}_M\|_1 \rightarrow 0$.
- $\varphi_M \rightarrow f$ μ -almost everywhere and also in L^1 by dominated convergence theorem.

Thus by Fatou's lemma,

$$\|\bar{\varphi}_M - \bar{f}\|_1 \leq \liminf_n \|S_n \varphi_M - S_n f\|_1 \leq \|\varphi_M - f\|_1$$

Finally

$$\begin{aligned} \|S_n f - \bar{f}\|_1 &\leq \|S_n f - S_n \varphi_M\|_1 + \|S_n \varphi_M - \bar{\varphi}_M\|_1 + \|\bar{\varphi}_M - \bar{f}\|_1 \\ &\leq \|f - \varphi_M\|_1 + \|S_n \varphi_M - \bar{\varphi}_M\|_1 + \|f - \varphi_M\|_1 \end{aligned}$$

so

$$\limsup \|S_n f - \bar{f}\|_1 \leq 2\|f - \varphi_M\|_1$$

for all M , so goes to 0 as $M \rightarrow \infty$. □

Remark.

1. The theorem holds if μ is only assumed to be σ -finite.
2. The theorem holds if $f \in L^p$ for $p \in [1, \infty)$. The $S_n f \rightarrow f$ in L^p .

Index

- L^p -space, 47
- π -system, 12
- σ -algebra, 10
 - independence, 33
 - invariant, 67
 - tail, 36
- Bernoulli shift, 69
- Birkhoff ergodic theorem, 70
- Boolean algebra, 2
- Borel σ -algebra, 12
- Borel measurable, 19
- Borel-Cantelli lemma, 35
- canonical model, 68
- Carathéodory extension theorem, 14
- central limit theorem, 65
- characteristic function, 56
- coboundary, 69
- completion, 17
- conditional expectation, 54
- convergence
 - almost surely, 39
 - in L^1 , 41
 - in distribution, 39
 - in mean, 41
 - in measure, 39
 - in probability, 39
- convolution, 59
- covariance matrix, 64
- density, 31, 52
- Dirac mass, 39
- distribution, 29
- distribution function, 29
- Dynkin lemma, 12
- ergodic, 67
- expectation, 29
- Fatou's lemma, 23
- filtration, 36
- Fourier transform, 56
- Furstenberg conjecture, 68
- Gaussian, 64
- Gaussian approximation, 60
- Hermitian norm, 49
- Hilbert space, 49
- Hölder inequality, 46
- iid., 34
- independence, 33
- infinite product σ -algebra, 35, 68
- inner product, 49
- integrable, 20
- integral with respect to a measure, 20
- invariant σ -algebra, 67
- invariant function, 67
- invariant set, 67
- io., 35
- Jensen inequality, 44
- Kolmogorov 0 – 1 law, 36, 69
- law, 29
- law of large numbers, 38
- law of larger numbers, 70
- Lebesgue measure, 8
- Lebesgue's dominated convergence theorem, 23
- Lévy criterion, 61
- maximal ergodic lemma, 71
- maximal inequality, 71
- mean, 32, 64
- measurable, 18
- measurable space, 10
- measure, 10
 - absolutely continuous, 51
 - finitely additive, 2
 - singular, 51
- measure-preserving, 67
- measure-preserving dynamical system, 67
- Minkowski inequality, 45
- moment, 32
- monotone convergence theorem, 20
- null set, 5
- orthogonal projection, 50
- Plancherel formula, 62
- probability measure, 29

- probability space, 29
- product σ -algebra, 26
 - infinite, 35, 68
- product measure, 27
 - infinite, 35
- Radon-Nikodym derivative, 52
- Radon-Nikodym theorem, 52
- random process, 36
- random variable, 29
 - independence, 33
- Riesz representation theorem, 51, 69
- sample path map, 68
- simple function, 19
- stationary process, 68
- stochastic process, 68
- strong law of large numbers, 38, 70
- tail event, 36, 69
- Tonelli-Fubini theorem, 27
- uniformly integrable, 41
- variance, 32
- von Neumann mean ergodic theorem, 69
- weak convergence, 39
- Young inequality, 46