Notes for the course of Functional Analysis, PhD in Statistics

Annalisa Cesaroni

Contents

These notes are intended for the first year students of the PhD course in Statistics, at University of Padova. They are not exhaustive, nor complete, but they could serve as a basis of the study of the arguments presented during the course of Functional Analysis. The topics are presented in a quite informal way, trying to reach also students without a specific preparation in mathematics. Only few proofs are provided and for the others bibliographical references are provided. At the end of each section some

exercises are proposed, more or less simple to solve. In the appendix there are the (sketchy) solutions to the problems.

Chapter 1

Measure theory and integration

1.1 Measure space

We fix a set X and we define $\mathcal{P}(X)$ the set of all subsets of X.

Definition 1.1.1. *$\Sigma \subset \mathcal{P}(X)$ is a $\sigma-$algebra on X if*

- *it is closed by complement, that is if $A \in \Sigma$ then $X \backslash A \in \Sigma$,*
- *it is closed by countable union, that is if $(A_i)_i$ is a sequence of elements in Σ then $\cup_{i=1}^{\infty} A_i \in \Sigma$.*

Let $C \subseteq \mathcal{P}(X)$, then $\Sigma(\mathcal{C})$, the σ-algebra generated by $\mathcal{C}$ is the smallest $\sigma-$algebra which contains all the elements in $\mathcal{C}$ (and then all countable intersections and countable unions of elements in $\mathcal{C}$).

The smallest possible σ-algebra on X is given by $\Sigma = \{\varnothing, X\}$, and the largest possible σ-algebra on X is $\Sigma = \mathcal{P}(X)$.

Definition 1.1.2. *$\mathcal{B}(\mathbb{R})$ is the σ-algebra on $\mathbb{R}$ generated by all the intervals $\mathcal{C} = \{(a, b) \mid a, b \in \mathbb{R}\}$. $\mathcal{B}(\mathbb{R}^N)$ is the σ-algebra on $\mathbb{R}^N$ generated by all the pluri-rectangulars $\mathcal{C} = \{\Pi_{i=1}^{N}(a_i, b_i) \mid a_i, b_i \in \mathbb{R}\}$.*

Remark 1.1.3. Note that $\sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R})$ also when $\mathcal{C} = \{(a, b] \mid a, b \in \mathbb{R}\}$, since $(a, b) = \cup_n \left(a, b - \frac{1}{n}\right]$, or when $\mathcal{C} = \{[a, b) \mid a, b \in \mathbb{R}\}$, since $(a, b) = \cup_n \left[a + \frac{1}{n}, b\right)$, or when $\mathcal{C} = \{[a, b] \mid a, b \in \mathbb{R}\}$ again because $(a, b) = \cup_n \left[a + \frac{1}{n}, b - \frac{1}{n}\right]$. Analogously $\sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R})$ when $\mathcal{C} = \{(a, +\infty) \mid a \in \mathbb{R}\}$, since $(a, b] = (a, +\infty) \cap (-\infty, b]$, and $(-\infty, b] = \mathbb{R} \backslash (b, +\infty)$ and so on.

Definition 1.1.4. *Let Σ be a σ-algebra on X. A function $\mu : \Sigma \to [0, +\infty]$ is a measure if*

- *$\mu(\varnothing) = 0$,*
- *it is σ-additive, that is if $(A_i)_i$ is a sequence of elements in Σ with $A_i \cap A_j = \varnothing$ for $i \neq j$ then $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{+\infty} \mu(A_i)$.*

(X, Σ, μ) is called a measure space.

If $\mu(X) < +\infty$, then μ is a finite measure (a probability measure if $\mu(X) = 1$). Usually measure spaces with probability measures are denoted with Ω (in place of X), the σ-algebra is $\mathcal{F}$ (in place of Σ) and the measure is $\mathbb{P}$ (in place of μ).

If $X = \cup_i A_i$, with $\mu(A_i) < +\infty$ for all i, μ is σ-finite.

If $X = \mathbb{R}^n$, $n \geqslant 1$ and $\Sigma = \mathcal{B}(\mathbb{R}^n)$, then μ is called a Borel measure.

Example 1.1.5. Let $x_0 \in \mathbb{R}$, and define the measure on $\mathcal{P}(\mathbb{R})$ as $\delta_{x_0}(A) = \begin{cases} 1 & x_0 \in A \\ 0 & x_0 \notin A \end{cases}$.

Then δ_{x_0} is called Dirac measure centered at x_0.

Proposition 1.1.6 (Monotonicity, subadditivity, continuity)**.** *Let μ be a measure on Σ. Then*

(i) *if $A \subset B$, $A, B \in \Sigma$, then $\mu(A) \leqslant \mu(B)$ (monotonicity with respect to inclusion);*

(ii) *if $(A_i)_i$ is a sequence of elements in Σ then $\mu(\cup_{i=1}^{\infty} A_i) \leqslant \sum_{i=1}^{+\infty} \mu(A_i)$;*

(iii) *if $(A_i)_i$ is a sequence of elements in Σ with $A_i \subseteq A_{i+1}$ then $\mu(\cup_{i=1}^{\infty} A_i) = \lim_{i \to +\infty} \mu(A_i)$;*

(iv) *if* $(A_i)_i$ *is a sequence of elements in* Σ *with* $A_i \supseteq A_{i+1}$ *and* $\mu(A_{i_0}) < +\infty$ *for some* i_0*, then* $\mu(\cap_{i=1}^{\infty} A_i) = \lim_{i\to+\infty} \mu(A_i)$.

Proof. (i) Observe that $B = A \cup (B\backslash A)$, so by σ-additivity $\mu(B) = \mu(A) + \mu(B\backslash A) \geqslant \mu(A)$.

(ii) Let $B_1 = A_1$ and $B_i = A_i \backslash \cup_{k=1}^{i-1} A_k$ then B_i are disjoint and

$$\mu(\cup_i A_i) = \mu(\cup_i B_i) = \sum_{i=1}^{+\infty} \mu(B_i) \leqslant \sum_{i=1}^{+\infty} \mu(A_i).$$

(iii) Let $B_1 = a_1$ and $B_i = A_i\backslash A_{i-1}$ then

$$\mu(\cup_i A_i) = \mu(\cup_i B_i) = \sum_{i=1}^{+\infty} \mu(B_i) = \lim_{n\to+\infty} \sum_{i=1}^{n} \mu(B_i) = \mu(A_n).$$

(iv) Let $F_i = A_{i_0}\backslash A_i$ for $i > i_0$. Then $\mu(A_{i_0}) = \mu(F_i) + \mu(A_i)$, $F_i \subseteq F_{i+1}$ and $\cup_i F_i = A_{i_0} \setminus \cap_i A_i$. Therefore by 1), we get

$$\mu(A_{i_0}) = \mu(\cap_i A_i) + \lim_i \mu(F_i) = \mu(\cap_i A_i) + \lim_i (\mu(A_{i_0}) - \mu(A_i))$$

and we cancel $\mu(A_{i_0})$ from both sides.

□

Definition 1.1.7. *Let* (X, Σ, μ) *a measure space. The completion of* Σ *with respect to* μ *is the* σ*-algebra*

$$\mathcal{M} = \{A \subseteq X \mid \exists B, C \in \Sigma, \mu(C) = 0, B \subseteq A, A\backslash B \subseteq C\}.$$

Definition 1.1.8. *Let* (X, Σ, μ) *a measure space. A property holds* **almost everywhere** *if there exists* $N \in \Sigma$ *with* $\mu(N) = 0$ *such that the property holds for all* $x \in X\backslash N$.

Proposition 1.1.9. *Let* Σ *be a* σ*-algebra on* X *and* $\mu : \Sigma \to [0, +\infty]$ *with* $\mu(\varnothing) = 0$*. Then they are equivalent:*

(i) μ *is* σ*-additive: if* $(A_i)_i$ *is a sequence of elements in* Σ *with* $A_i \cap A_j = \varnothing$ *for* $i \neq j$ *then* $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{+\infty} \mu(A_i)$,

(ii) μ *is additive: if* $A, B \in \Sigma$ *and* $A \cap B = \varnothing$ *then* $\mu(A \cap B) = \mu(A) + \mu(B)$
and
μ *is countable subadditive: if* $(A_i)_i$ *is a sequence of elements in* Σ *then* $\mu(\cup_{i=1}^{\infty} A_i) \leqslant \sum_{i=1}^{+\infty} \mu(A_i)$*;*

(iii) μ *is additive: if* $A, B \in \Sigma$ *and* $A \cap B = \varnothing$ *then* $\mu(A \cap B) = \mu(A) + \mu(B)$
and
μ *is continuous on increasing sequence of sets: if* $(A_i)_i$ *is a sequence of elements in* Σ *with* $A_i \subseteq A_{i+1}$ *then* $\mu(\cup_{i=1}^{\infty} A_i) = \lim_{i\to+\infty} \mu(A_i)$.

Proof. The fact that (i) implies (ii) and that (i) implies (iii) has been proved in Proposition 1.1.6. We prove that (ii) implies (i). We consider a sequence $(A_i)_i$ of elements in Σ with $A_i \cap A_j = \varnothing$ for $i \neq j$. Then by (ii) we get that $\mu(\cup_{i=1}^{\infty} A_i) \leqslant \sum_{i=1}^{+\infty} \mu(A_i)$. On the other hand by additivity and monotonicity (which is a consequence of additivity) we get that for every n, $\mu(\cup_{i=1}^{\infty} A_i) \geqslant \mu(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu(A_i)$. Sending $n \to +\infty$ we conclude $\mu(\cup_{i=1}^{\infty} A_i) \geqslant \sum_{i=1}^{+\infty} \mu(A_i)$.

We prove that (iii) implies (i). We consider a sequence $(A_i)_i$ of elements in Σ with $A_i \cap A_j = \varnothing$ for $i \neq j$. We define $B_i = \cup_{j=1}^{i} A_j$. Then $\cup_i B_i = \cup_i A_i$. Note that by additivity $\mu(B_i) = \sum_{j=1}^{i} \mu(A_j)$ and that $B_1 \subseteq B_2 \subseteq B_3 \ldots$. Therefore by (iii) and additivity we get

$$\mu(\cup_{i=1}^{\infty} A_i) = \mu(\cup_{i=1}^{\infty} B_i) = \lim_{i\to+\infty} \mu(B_i) = \lim_{i\to+\infty} \sum_{j=1}^{i} \mu(A_j) = \sum_{j=1}^{+\infty} \mu(A_j).$$

□

1.2 Borel measures on $\mathbb{R}$ and cumulative distribution functions

Let $F : \mathbb{R} \to \mathbb{R}$ be an increasing function which is right continuous, that is $\lim_{x\to a^+} F(x) = F(a)$. We define for all $a, b \in \mathbb{R}$,

$$\mu_F(a, b] = F(b) - F(a) \qquad \mu_F(\varnothing) = 0.$$

Then for every set $C \subset \mathbb{R}$ we define

$$\mu_F^*(C) = \inf\{\sum_i F(b_i) - F(a_i) \mid C \subseteq \cup_i (a_i, b_i]\}.$$

Note that since F is increasing, we get that for sequences $a_1 < b_1 < a_2 < b_2 < \cdots < a_i < b_i < a_{i+1} < b_{i+1} \ldots$, we obtain

$$\mu_F^*(\cup_i (a_i, b_i]) = \sum_i F(b_i) - F(a_i).$$

Observe that if we define $\mathcal{C} = \{(a, b], a, b \in \mathbb{R}\}$, then $\Sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R})$. Note that if $F_1 = F_2 + c$ for some constant then $\mu_{F_1}^* = \mu_{F_2}^*$. Also the viceversa is true: if $\mu_{F_1}^* = \mu_{F_2}^*$, then $F_1 = F_2 + c$ for some constant c.

Remark 1.2.1. Note that F monotone increasing implies that $\mu_F(a, b] \geqslant 0$, and moreover, since F is right continuous, then

$$\mu_F\left(\cup_n (a + 1/n, b]\right) = \mu_F(a, b] = F(b) - F(a) = F(b) - \lim_n F(a + 1/n) = \lim_n \mu_F(a + 1/n, b].$$

Reasoning as before, it is possible to see that, at least when restricted to $\mathcal{C}$, there holds that μ_F has positive values, is additive and is continuous with respect to increasing sequences of sets (which is enough to get σ-additivity if μ_F is defined on a σ-algebra, see Proposition 1.1.9).

We recall that F is monotone increasing and then $\lim_{x\to+\infty} F(x) = \sup F$ and $\lim_{x\to-\infty} F(x) = \inf F$ (we say that if $F(\mathbb{R})$ is unbounded from above, $\sup F = +\infty$ and if $F(\mathbb{R})$ is unbounded from below, $\inf F = -\infty$).

We may extend μ_F^* to intervals obtained by unions and intersections of elements in $\mathcal{C}$, and using additivity and continuity. In particular we get

$$\begin{aligned}
\mu_F^*(a, +\infty) &= \mu_F^*\left(\cup_n (a, a + n]\right) = \lim_n F(a + n) - F(a) = \sup F - F(a) \\
\mu_F^*(-\infty, b] &= \mu_F^*\left(\cup_n (b - n, b]\right) = \lim_n F(b) - F(b - n) = F(b) - \inf F \\
\mu_F^*(a, b) &= \mu_F^*\left(\cup_{n\geqslant n_0} (a, b - 1/n]\right) = \lim_n F(b - 1/n) - F(a) = \lim_{x\to b^-} F(x) - F(a) \\
\mu_F^*(-\infty, b) &= \mu_F^*\left((-\infty, b - 1] \cup (b - 1, b)\right) = \mu_F^*((-\infty, b - 1]) + \mu_F^*((b - 1, b)) \\
&= \lim_{x\to b^-} F(x) - F(b - 1) + F(b - 1) - \inf F = \lim_{x\to b^-} F(x) - \inf F \\
\mu_F^*[a, b) = &= \mu_F^*[(a - 1, b)\backslash(a - 1, a)] = \mu_F^*(a - 1, b) - \mu_F^*(a - 1, a) \\
&= \lim_{x\to b^-} F(x) - F(a - 1) - \lim_{x\to a^-} F(x) + F(a - 1) = \lim_{x\to b^-} F(x) - \lim_{x\to a^-} F(x) \\
\mu_F^*[a, b] &= \mu_F^*[[a, b + 1)\backslash(b, b + 1)] = \mu_F^*[a, b + 1) - \mu_F^*(b, b + 1) \\
&= F(b) - \lim_{x\to a^-} F(x) \\
\mu_F^*[a, +\infty) &= \sup_F - \lim_{x\to a^-} F(x).
\end{aligned}$$

Note that

$$\begin{aligned}
\mu_F^*(\mathbb{R}) &= \mu_F^*\left(\cup_n (a - n, b + n]\right) = \lim_n F(b + n) - F(a - n) = \sup F - \inf F \\
\mu_F^*(\{a\}) &= \mu_F^*\left((c, a]\backslash(c, a)\right) \\
&= \mu_F^*((c, a]) - \mu_F^*((c, a)) = F(a) - F(c) - (\lim_{x\to a^-} F(x) - F(c) = F(a) - \lim_{x\to a^-} F(x).
\end{aligned}$$

Theorem 1.2.2. (i) *There exists a unique Borel measure μ_F which coincides with μ_F^* on intervals $(a, b]$. This measure is σ-finite and it is finite if and only if $\sup F - \inf F < +\infty$.*

(ii) *Given a Borel measure on $\mathbb{R}$ which is σ-finite, there exists F monotone increasing and right continuous such that $\mu = \mu_F$. F is unique up to addition of constants: that is if $\mu = \mu_F = \mu_G$ then there exists $c \in \mathbb{R}$ such that $F(x) = G(x) + c$ for all x. t*

Proof. (i) The proof is based on the Caratheodory criterion, and we refer to [3, Theorem 1.14, Theorem 1.16]. As for the σ- finiteness it is sufficient to observe that $\mu_F(-n,n] = F(n) - F(-n) < +\infty$ and $\mathbb{R} = \cup_n(-n,n]$. Moreover, since $\mu_F(\mathbb{R}) = \sup F - \inf F$, we conclude that F is finite iff $\sup_F - \inf F < +\infty$.

(ii) We want to construct F. Put $F(0) = 0$ and

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

Observe that if $b > a \geqslant 0$, $F(b) - F(a) = \mu(0,b] - \mu(0,a] = \mu(0,b]\backslash(0,a] = \mu(a,b] \geqslant 0$, if $0 \geqslant b > a$, then $F(b) - F(a) = -\mu(b,0] + \mu(a,0] = \mu(a,0]\backslash(b,0] = \mu(a,b] \geqslant 0$ and finally if $a < 0 < b$, then $F(b) - F(a) = \mu(0,b] + \mu(a,0] = \mu(a,b] \geqslant 0$. So F is increasing.

We check that it is right continuous. First of all observe that for $a > 0$, $\lim_{x\to a^+} F(x) = \lim_n F(a + 1/n) = \lim_n \mu(0, a + 1/n] = \mu(\cap_n(0, a + 1/n]) = \mu(0,a] = F(a)$. If $a = 0$ $\lim_{x\to 0^+} F(x) = \lim_n F(1/n) = \lim_n \mu(0, 1/n] = \mu(\cap_n(0, 1/n]) = \mu(\varnothing) = 0 = F(0)$. Finally if $a < 0$, then $\lim_{x\to a^+} F(x) = \lim_n F(a + 1/n) = -\lim_n \mu(a + 1/n, 0] = -\mu(\cup_n(a + 1/n, 0]) = -\mu(a,0] = F(a)$.

Finally we already checked that $\mu(a,b] = F(b) - F(a)$ and then we conclude that $\mu = \mu_F$.

Assume now that there exists a right continuous increasing function G such that $\mu = \mu_G$. Then for $x > 0$, $F(x) = \mu(0,x] = \mu_G(0,x] = G(x) - G(0)$ and for $x < 0$ then $F(x) = -\mu(x,0] = \mu_G(x,0] = -(G(0) - G(x)) = G(x) - G(0)$. So, this implies that $F(x) = G(x) - G(0)$ (for $x = 0$ this is trivially verified).

□

Definition 1.2.3. *Let μ be a finite Borel measure. The function $F(x)$ associated to the measure μ and normalized in order to have* $\inf F = 0$ *is called the cumulative distribution function of the measure μ. It is easy to check that $F(x) := \mu(-\infty, x]$.*

1.3 The Lebesgue measure on $\mathbb{R}$ and $\mathbb{R}^n$.

Definition 1.3.1. *Let $F(x) = x$ for all x, then $\overline{\mu}_F$ is called* **Lebesgue measure**. *We indicate with $\mathcal{L}$. We denote with $\mathcal{M}(\mathbb{R})$ the completion of $\mathcal{B}(\mathbb{R})$ with respect to $\mathcal{L}$, and we call it the Σ-algebra of Lebesgue measurable sets.*

Proposition 1.3.2. *The Lebesgue measure*

(i) *associates to each interval its length,*

(ii) *is translation invariant, that is $\mathcal{L}(A + x) = \mathcal{L}(A)$ for all $x \in \mathbb{R}$, $A \in \mathcal{M}$,*

(iii) *is homogenous, that is $\mathcal{L}(\lambda A) = \lambda\mathcal{L}(A)$ for all $\lambda > 0$, $A \in \mathcal{M}$,*

(iv) *assigns measure 0 to points, and so also to countable sets (e.g. $\mathbb{Q}$),*

(v) *it is σ-finite, since $\mathbb{R} = \cup_{n\in\mathbb{N}}(-n,n)$ and $\mathcal{L}(-n,n) = 2n$.*

Proof. The proof is immediate by definitions and σ-additivity. Exercise. □

Measurable sets in $\mathbb{R}$ which contain at least one interval (they are called sets with non empty interior) have positive measure. On the other hand sets which are given by countable union of isolated points have measure zero. Nevertheless there are sets with empty interior in $\mathbb{R}$ (so that do not contain any interval) and with positive measure (almost full measure).

Example 1.3.3 (A set of positive measure which does not contain any interval)**.** Let (r_n) be an enumeration of $\mathbb{Q} \cap [0,1]$ and fix $\varepsilon > 0$ small.

Set $A = \cup_n(r_n - \varepsilon 2^{-n}, r_n + \varepsilon 2^{-n})$. Then by subadditivity, $\mathcal{L}(A) \leqslant \sum_n 2\varepsilon 2^{-n} = 4\varepsilon$. Moreover $B = [0,1]\backslash A$ is a set which does not contain any interval (otherwise it should contain some rational number but $\mathbb{Q} \cap [0,1] \subseteq A$), and moreover $\mathcal{L}(B) \geqslant 1 - 4\varepsilon > 0$.

Not all the subsets of $\mathbb{R}$ are contained in $\mathcal{M}(\mathbb{R})$, so there are sets which are not measurable. This is due to the fact that if we want to define a measure μ on the intervals of $\mathbb{R}$ such that $\mu([0,1]) = 1$, $\mu(A \cup B) = \mu(A) + \mu(B)$ if $A \cap B = \varnothing$ and $\mu(A) = \mu(B)$ if B can be obtained translating and rotating A, then the σ- algebra of measurable sets cannot be $\mathcal{P}(\mathbb{R})$.

Example 1.3.4 (A set which is not (Lebesgue) measurable)**.** We say that $x, y \in [0,1]$ are equivalent if $x - y \in \mathbb{Q}$. Let $P \in [0,1]$ a set such that P consists of exactly one representative point from each equivalence class (this set exists by the axiom of choice). In particular this means that if $p, p' \in P$, $p \neq p'$, then $p - p' \notin \mathbb{Q}$. We claim that P provides the required example of a non measurable set. We prove it by contradiction, showing that it is not possible for P to be measurable.

For each $q \in \mathbb{Q} \cap [0,1]$, define

$$P_q = [(P+q) \cap [0,1)] \cup [(P+q) \backslash [0,1)) - 1] = \{p + q,\ p \in P \cap [0, 1-q)\} \cup \{p + q - 1,\ p \in P \cap [1-q, 1)\}.$$

So P_q is obtained by considering $P + q$ and then shifting back of 1 unit the part of $P + q$ which is outside the interval $[0, 1)$.

First of all we observe that $\mathcal{L}(P) = \mathcal{L}(P_q)$. Indeed $[(P+q) \cap [0,1)] \cap [(P+q) \backslash [0,1)) - 1] = \varnothing$, since if $p + q \in [0,1)$ for some $p \in P$ and $p' + q - 1 \in [0,1)$ for some $p' \in P$, then necessarily $p + q \neq p' + q - 1$, since $p, p' \in [0,1)$.

Moreover we observe that if $r \neq q \in \mathbb{Q} \cap [0,1)$, then $P_r \cap P_q = \varnothing$. Indeed assume it is not true and $x \in P_r \cap P_q$, this means that $x = p + r = p' + q$, for some $p, p' \in P$ or $x = p + r = p' + q - 1$, or $x = p + r - 1 = p' + q$. In any case we get that $p - p' \in \mathbb{Q}$, which implies that $p = p'$ by definition of the set P and so $r = q$.

Finally we observe that $\cup_{q \in \mathbb{Q} \cap [0,1)} P_q = [0,1)$. Indeed take $x \in [0,1)$, then there exists $p \in P$ such that x is equivalent to P, which means that there exists $q \in \mathbb{Q}$ such that $x = p + q$. In particular this implies that $q \in (0,1]$ and $x \in P_q$.

We conclude by σ-additivity that

$$1 = \mathcal{L}([0,1)) = \mathcal{L}(\cup_{q \in \mathbb{Q} \cap [0,1)} P_q) = \sum_{q \in \mathbb{Q} \cap [0,1)} \mathcal{L}(P_q) = \sum_{q \in \mathbb{Q} \cap [0,1)} \mathcal{L}(P) = \begin{cases} 0 & \text{if } \mathcal{L}(P) = 0 \\ +\infty & \text{if } \mathcal{L}(P) > 0 \end{cases}$$

which is not possible.

It is possible to define the Lebesgue measures on $\mathbb{R}^n$ as the product measure of the Lebesgue measure on $\mathbb{R}$. It is a Borel maesure and we denote with $\mathcal{M}$ the Σ-algebra of Lebesgue measurable sets. We refer to [3, Section2.6].

Proposition 1.3.5. *The Lebesgue measure on $\mathbb{R}^n$*

(i) *associates to each n-parallelepiped its volume,*

(ii) *is translation invariant, that is $\mathcal{L}(A + x) = \mathcal{L}(A)$ for all $x \in \mathbb{R}^n$, $A \in \mathcal{M}$,*

(iii) *is n-homogenous, that is $\mathcal{L}(\lambda A) = \lambda^n \mathcal{L}(A)$ for all $\lambda > 0$, $A \in \mathcal{M}$, in particular $\mathcal{L}(B(0,r)) = r^n \mathcal{L}(B(0,1))$, where $B(0,r)$ is the ball if radius r centered at 0,*

(iv) *it is σ-finite, since $\mathbb{R}^n = \cup_{k \in \mathbb{N}} B(0,k)$ and $\mathcal{L}B(0,k) = k^n \mathcal{L}(B(0,1))$.*

From now on, for simplicity we will indicate $|A| = \mathcal{L}(A)$.

1.4 Measurable functions and random variables

Definition 1.4.1. *Let (X, Σ, μ) be a measure space, and let $f : X \to \mathbb{R}$ be a function. Then f is measurable if for all $t \in \mathbb{R}$,*

$$A(t) := \{x \in X \mid f(x) > t\} = f^{-1}(t, +\infty) \in \Sigma.$$

In particular we will be interested in the case in which $(X, \Sigma, \mu) = (\mathbb{R}^n, \mathcal{M}, \mathcal{L})$. In this case saying that $f : \mathbb{R}^n \to \mathbb{R}$ is measurable is equivalent to require that for all $A \in \mathcal{B}(\mathbb{R})$, $f^{-1}(A) \in \mathcal{M}$.

Example 1.4.2. Let $A \in \mathcal{M}$ and define the characteristic function of A as

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

Then χ_A is measurable. Indeed $A(t) = \varnothing$ for $t \geqslant 1$, $A(t) = \mathbb{R}^n$ for $t \leqslant 0$ and $A(t) = A$ for $t \in (0,1)$.

Example 1.4.3 (Random variables)**.** If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space (that is a measure space endowed with a probability measure), the measurable functions, that is functions $f : \Omega \to \mathbb{R}$ such that for all $t \in \mathbb{R}$, $A(t) := \{\omega \in \Omega \mid f(\omega) > t\} \in \mathcal{F}$, are called **random variables**. Usually random variables are indicated with X instead of f.

There is a notion of convergence of measurable functions which is quite used in probability.

Definition 1.4.4 (Convergence in measure)**.** *Let f_n, f be measurable functions defined on the measure space (X, Σ, μ). Then f_n converge to f in measure if for every $\varepsilon > 0$*

$$\lim_n \mu\{x \in X \mid |f_n(x) - f(x)| \geqslant \varepsilon\} = 0.$$

If we are in a probability space, this convergence is called **convergence in probability**, *since it reads*

$$\lim_n \mathbb{P}\{\omega \in \Omega \mid |X_n(\omega) - X(\omega)| \geqslant \varepsilon\} = 0.$$

1.5 Lebesgue integral

Definition 1.5.1. *Let $k \geqslant 1$, $A_1, \dots A_K$ a finite family of disjoint sets in $\mathcal{M}$ and $c_1, \dots c_k > 0$. The function $\phi(x) = \sum_{i=1}^k c_i \chi_{A_i}(x)$ is called* **simple function**. *It is a measurable (positive) function and we define its integral as*

$$\int_{\mathbb{R}^N} \phi(x)dx = \sum_{i=1}^k c_i \mathcal{L}(A_i).$$

Definition 1.5.2 (Lebesgue integral)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a measurable function such that $f(x) \geqslant 0$ for all x. Then*

$$\int_{\mathbb{R}^n} f(x)dx = \sup\left\{\int_{\mathbb{R}^n} \phi(x)dx \mid \phi \text{ simple function with } \phi \leqslant f\right\}.$$

If f is not positive we define its positve part $f^+(x) = \max(f(x), 0)$ and its negative part $f^-(x) = \max(-f(x), 0)$ and we define

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

Note that $\int_{\mathbb{R}^n} |f(x)|dx = \int_{\mathbb{R}^n} f^+(x)dx + \int_{\mathbb{R}^n} f^-(x)dx$.

Since $f^+ \leqslant |f|, f^- \leqslant |f|$, we have that

$$\left|\int_{\mathbb{R}^n} f(x)dx\right| < +\infty \quad \text{iff} \quad \int_{\mathbb{R}^n} |f(x)|dx < +\infty.$$

We denote

$$L^1(\mathbb{R}^n) := \{f : \mathbb{R}^n \to \mathbb{R} \mid \ f \text{ is measurable and } \int_{\mathbb{R}^n} |f(x)|dx < +\infty\}.$$

If $A \in \mathcal{M}$, then we define

$$L^1(A) = \left\{f : \mathbb{R}^n \to \mathbb{R} \mid f \text{ is measurable and } \int_{\mathbb{R}^n} |f(x)|\chi_A(x) = \int_A |f(x)|dx < +\infty\right\}.$$

Proposition 1.5.3. *The following properties hold.*

- *If $f = 0$ almost everywhere then $\int_{\mathbb{R}^n} f(x) = 0$. If $\int_{\mathbb{R}^n} |f(x)|dx = 0$ then $f = 0$ almost everywhere.*
- *If f, g are measurable functions such that $f = g$ almost everywhere, then $\int_{\mathbb{R}^n} f(x)dx = \int_{\mathbb{R}^n} g(x)dx$.*
- *If $f, g \in L^1(\mathbb{R}^n)$, $\alpha, \beta \in \mathbb{R}$, then $\int_{\mathbb{R}^n} \alpha f(x) + \beta g(x)dx = \alpha \int_{\mathbb{R}^n} f(x)dx + \beta \int_{\mathbb{R}^n} g(x)dx$.*
- *If $f, g \in L^1(\mathbb{R}^n)$, and $f \leqslant g$ then $\int_{\mathbb{R}^n} f(x)dx \leqslant \int_{\mathbb{R}^n} g(x)dx$.*

Proof. The proof is obtained by exploiting definitions, see [3, Section 2..2] □

Remark 1.5.4. [On the definition of L^1] Note that due to the previous proposition, in particular the fact that if f, g are measurable functions such that $f = g$ almost everywhere, then $\int_{\mathbb{R}^n} f(x)dx = \int_{\mathbb{R}^n} g(x)dx$, **we identify functions in $L^1(\mathbb{R}^n)$ which coincide almost everywhere.** So a function f in $L^1(\mathbb{R}^n)$ is actually a class of equivalence of functions, we do not distinguish functions which are different on sets of measure zero.

Theorem 1.5.5 (Monotone convergence)**.** *Let $f_k : \mathbb{R}^n \to \mathbb{R}$ measurables , positive, i.e. $f_k \geqslant 0$ for all k, and such that $f_k(x) \leqslant f_{k+1}(x)$ for all x and for all k. Then*

$$\lim_k \int_{\mathbb{R}^n} f_k(x)dx = \int_{\mathbb{R}^n} \lim_k f_k(x)dx.$$

Proof. See [3, Theorem 2.14]. □

Proposition 1.5.6. *An equivalent definition of the Lebesgue integral (which can be very useful) is the following. Let $f : \mathbb{R}^n \to \mathbb{R}$ measurable and positive. Let for every $t > 0$ $F(t) = \mathcal{L}(A(t)) = \mathcal{L}\{x \mid f(x) > t\}$. F is called the* **repartition function of** *f. Then*

$$\int_{\mathbb{R}^n} f(x)dx = \int_0^{+\infty} F(t)dt.$$

Proof. See [3, Proposition 6.24] □

1.6 Decomposition of measures.

Definition 1.6.1. *Let ν, ρ be measures defined on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.*

ν is **absolutely continuous** *with respect to $\mathcal{L}$, and we write $\nu << \mathcal{L}$ if $\nu(A) = 0$ for all $A \in \mathcal{B}$ such that $\mathcal{L}(A) = 0$.*

ρ is **singular** *with respect to $\mathcal{L}$, and we write $\rho \perp \mathcal{L}$, if there exist $A, B \in \mathcal{B}$, $A \cap B = \varnothing$, $A \cup B = \mathbb{R}^n$, such that $\mathcal{L}(A) = 0$ and $\rho(B) = 0$.*

Example 1.6.2. Let $x_0 \in \mathbb{R}$ and consider the Dirac measure δ_{x_0} centered at x_0. Then it is singular with respect to $\mathcal{L}$. Indeed fix $A = \mathbb{R}\backslash\{x_0\}$, $B = \{x_0\}$, and observe that $\mathcal{L}(B) = 0$ and $\delta_{x_0}(A) = 0$.

Proposition 1.6.3. *Let $f \geqslant 0$, measurable and such that $\int_{-M}^{M} f(x)dx < +\infty$ for all $M > 0$. Define the function*

$$\nu_f : \mathcal{M} \to [0, +\infty] \quad as \quad \nu_f(A) = \int_A f(x)dx.$$

Then ν_f is a measure on $(\mathbb{R}^n, \mathcal{M})$, which is σ-finite and which is absolutely continuous with respect to $\mathcal{L}$. If $f \in L^1(\mathbb{R}^n)$ the measure is finite.

Proof. First of all we show that it is a measure. Observe that $f(x)\chi_{\varnothing}(x) = 0$ almost everywhere, then $\nu_f(\varnothing) = 0$. Let $A_i \in \mathcal{M}$ which are pairwise disjoint. Define the simple function $\phi_k(x) = \sum_{i=1}^k \chi_{A_i}(x)$. Note that $\lim_k \phi_k(x) = \chi_{\cup_i A_i}(x)$. Moreover $0 \leqslant f(x)\phi_k(x) \leqslant f(x)\phi_{k+1}(x)$ and so by the monotone convergence theorem we get

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

Observe that

$$\lim_k \int_{\mathbb{R}^n} \phi_k(x)f(x)dx = \lim_k \int_{\mathbb{R}^n} \sum_{i=1}^k \phi_i(x)f(x)dx = \lim_k \sum_{i=1}^k \int_{\mathbb{R}^n} \phi_i(x)f(x)dx$$

$$= \lim_k \sum_{i=1}^k \int_{A_i} f(x)dx = \lim_k \sum_{i=1}^k \nu_f(A_i) = \sum_{i=1}^{+\infty} \nu_f(A_i)$$

and

$$\int_{\mathbb{R}^n} \lim_k \phi_k(x)f(x)dx = \int_{\mathbb{R}^n} \chi_{\cup_i A_i}(x)f(x)dx = \nu_f(\cup_i A_i).$$

Therefore we get that ν_f is a measure.

Since $\nu_f(B(0,k)) = \int_{B(0,k)} f(x)dx < +\infty$ by assumption, then ν_f is σ-finite.

Finally, note that if $A \in \mathcal{M}$ and $\mathcal{L}(A) = 0$, this implies that $\chi_A(x) = 0$ almost everywhere. Therefore also $f(x)\chi_A(x) = 0$ almost everywhere, which implies $\nu_f(A) = 0$.

□

Example 1.6.4. Let $f(x) = e^{-|x|^2}$. Then $f \in L^1(\mathbb{R}^n)$ and the measure ν_f is called the Gaussian measure. Note that it is a finite measure, and $\int_{\mathbb{R}^n} e^{-|x|^2}dx = \pi^{n/2}$, see [3, Prop. 2.53].

Theorem 1.6.5 (Lebesgue-Radon-Nikodym theorem)**.** *Let μ a Borelian measure on $\mathbb{R}^n$ which is $\sigma-$finite. Then there exist a unique $\nu << \mathcal{L}$ (absolutely continuous part) and a unique $\rho \perp \mathcal{L}$ (singular part) such that $\mu = \nu + \rho$.*

Moreover there exists $f \geqslant 0$, measurable and such that $\int_{B_R} f(x)dx < +\infty$ for all $R > 0$, for which $\nu = \nu_f$.
f is called the **density** *of ν, or the Radon-Nikodym derivative of ν and can be obtained (if the measure ν is regular) as $f(x) = \lim_{r\to 0} \frac{\nu(B(x,r))}{\mathcal{L}(B(x,r))}$.*

Proof. For the proof we refer to [3, Section 3.2]. □

1.7 Push forward of measures and laws of random variables

Definition 1.7.1 (Push forward of a measure)**.** *Let (X, Σ, μ) be a measure space, and let $f : X \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathcal{L})$ be a measurable function. Then the push forward of the measure μ by the function f is the Borel measure $f_\sharp\mu$ defined as follows: for all $A \in \mathcal{B}(\mathbb{R})$,*

$$f_\sharp\mu(A) = \mu\{x \in X, f(x) \in A\}.$$

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable (see Section 2.4). Then **the law** $\mathbb{L}_X$ of X is the push forward of the probability measure $\mathbb{P}$ by X: that is for every $A \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{L}_X(A) = \mathbb{P}(\{\omega \ | X(\omega) \in A\}).$$

The cumulative distribution function associated to such Borel measure is defined as

$$F_X(x) = \mathbb{P}(\{\omega \ | X(\omega) \leqslant x\}).$$

The law identifies the (main properties of) random variable, and often the random variables can be described just in terms of their laws.

Remark 1.7.2 (The cumulative distribution function)**.** If X is an (absolutely) continuous random variable, $\mathbb{L}_X$ is an absolutely continuous measure and F_X is an absolutely continuous function. The density of f_X with respect to the Lebesgue measure is

$$f_X(x) = F'_X(x) = \lim_{h\to 0} \frac{F(x+h) - F(x)}{h} \qquad \text{for a.e.} x \in \mathbb{R}.$$

If X is a discrete random variable, $\mathbb{L}_X$ is a singular measure with respect to the Lebesgue measure and F_X is a monotone piecewise constant function.

More generally if F_X is the cumulative distribution function associated to a random variable, then F a right continuous, monotone increasing function, which we normalize to have $\inf F_X = 0$ (and obviously $\sup F = 1$). F_X has at most countably many discontinuity points, that are those for which $F(a) > \lim_{x\to a^-} F(x)$, or equivalently for which

$$\mathbb{P}(\{\omega \ | X(\omega) = a\}) > 0.$$

We define

$$F^d_X(x) := \sum_{y\leqslant x} \mathbb{P}(\{\omega \ | X(\omega) = a\}).$$

Note that F_d is a monotone increasing function, which is a.e. constant and has jumps only at discontinuity points of F_X.

So the function $F_X - F^d_X$ is a continuous function, and it is easy to check it is still monotone increasing. A deep result in mathematical analysis (see [3, Thm 3.23]) states that monotone increasing functions F are differentiable a.e.- that is for a.e. $a \in \mathbb{R}$ there exists $F'(a) = \lim_{h\to 0} \frac{F(a+h)-F(a)}{h}$ and moreover $F'(a) \geqslant 0$ a.e. So we define the absolutely continuous part of F_X as

$$F^{ac}_X(x) = \int_{-\infty}^x F'_X(y)dy = \int_{-\infty}^x (F_X - F^d_X)'(y)dy.$$

So, $F'_X(x)$ is the density of the absolutely continuous measure $\mu_{F^{ac}_X}$.

It is possible to prove that in general

$$F_X(x) = F_X^d(x) + F_X^{ac}(x) + F_X^s(x)$$

where F_X^s is a continuous and increasing function, whose derivative is zero in almost all x, but it can be not identically zero (a typical example is the devil's staircase function, or the Cantor function).

The three functions F_X^d, F_X^{ac}, F_X^s are all increasing, but are of very different nature:

- F_X^d can only increase by jumps and it is constants between two consecutive jumps,
- F_X^{ac} is a "nice" function with the property of being the integral of its derivative, which coincide with the distribution density,
- F_X^s is a function quite hard to imagine (continuous, increasing with zero derivative a.e.).

We typically deal with real random variables such that the singular part F_X^s of their distribution function is identically zero.

Moreover, we see that a real random variable is discrete if and only if $F_X = F_X^d$ and it is absolutely continuous if and only if $F_X = F_X^{ac}$ and in this case $f_X(x) = F_X'(x)$.

Remark 1.7.3 (Joint law)**.** If X, Y are random variables on the same probability space, that is $X, Y : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$, we may define the joint law $\mathbb{L}_{X,Y}$ as the push forward of the probability measure $\mathbb{P}$ with respect to the map $(X, Y) : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R} \times \mathbb{R}$ which associates to ω the pair $(X(\omega), Y(\omega))$. Therefore

$$\mathbb{L}_{X,Y} = (X, Y)_\sharp \mathbb{P} \qquad \mathbb{L}_{X,Y}(A \times B) = \mathbb{P}\{\omega \in \Omega, X(\omega) \in A, Y(\omega) \in B\}.$$

The joint cumulative distribution function as

$$F_{X,Y}(x, y) = \mathbb{P}(\{\omega \mid X(\omega) \leqslant x\} \cap \{\omega \mid Y(\omega) \leqslant y\}).$$

If X, Y are independent then $F_{X,Y}(x, y) = F_X(x)F_Y(y)$. Two random variables X and Y are jointly continuous if there exists a nonnegative function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ such that for any measurable set $A \subseteq \mathbb{R}^2$ there holds

$$\mathbb{P}(\{\omega \mid (X(\omega), Y(\omega)) \in A\}) = \int_A f_{X,Y}(x, y)dxdy.$$

The function $f_{X,Y}(x, y)$ is called the joint probability density function and is obtained as

$$f_{X,Y}(x, y) = \frac{d^2}{dxdy} F_{X,Y}(x, y) \qquad \text{a.e..}$$

Given the joint probability density function it is possible to recover the density functions of X and Y as the marginals:

$$f_X(x) = \int_\infty^{+\infty} f_{X,Y}(x, y)dy \qquad f_Y(y) = \int_\infty^{+\infty} f_{X,Y}(x, y)dx.$$

On the other hand, given the marginals f_X, f_Y, there is not a unique associated joint probability density function (apart from the case in which X, Y are independent, in which case $f_{X,Y}(x, y) = f_X(x)f_Y(y)$).

Remark 1.7.4. Some examples of widely used random variables/laws:

- the Dirac measure δ_c centered at c is the law associated to the constant random variable c (so the random variable X such that $X(\omega) = c$ almost surely).
- the **gamma law** with parameters a, b is an absolutely continuous measure with density $f(x) = \Gamma(a)^{-1} b^a x^{a-1} e^{-bx} \chi_{(0,+\infty)}(x)$
- the **chi-square law** is a gamma distribution with parameters $n/2, 1/2$,
- the **normal or Gaussian law** with parameters μ, σ is an absolutely continuous random variable, with density $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma}}$,
- the **standard normal law** is a normal law with parameters $0, 1$, that is an absolutely continuous measure with density $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$,
- the **binomial law** of parameters n, p is a singular measure, and it is given by $\sum_{k=0}^n \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \delta_k$ where δ_k is the Dirac measure centered at k,
- the **Poisson law** of parameter λ is a singular measure, and it is given by $e^{-\lambda} \sum_{k=0}^{+\infty} \frac{\lambda^k}{k!} \delta_k$ where δ_k is the Dirac measure centered at k.

1.8 The space of laws of random variables. p-moments of a random variable

We restrict to consider the family of all Borel measures in $\mathbb{R}$ which are laws of some random variable, that is

$$\mathcal{P}(\mathbb{R}) = \{\mu \mid \text{there exists a probability space } (\Omega, \mathcal{F}, \mathbb{P}) \text{ and } X : \Omega \to \mathbb{R} \text{ random variable s.t. } \mu = \mathbb{L}_X\}.$$

Note that all $\mu \in \mathcal{P}(\mathbb{R})$ are Borel measures in $\mathbb{R}$ such that $\mu(\mathbb{R}) = 1$ (so they are finite). Observe also that given $\mu \in \mathcal{P}(\mathbb{R})$ there are several probability spaces $(\Omega, \mathbb{P}, \mathcal{F})$ and several $X : \Omega \to \mathbb{R}$ random variables s.t. $\mu = \mathbb{L}_X$. In any case, the law determines the most important feature (from a measure theoretic/analytic point of view) of the random variable, on the other hand we loose completely the information about what is the sample space (that is sample space, the set of possible outcomes of an experiment), and which are the events which have been measured. To every $\mu \in \mathcal{P}(\mathbb{R})$ is uniquely associated its cumulative distribution function

$$F(x) = \mu(-\infty, x].$$

Moreover, if μ is absolutely continuous, its density $f(x)$ coincides almost everywhere with $F'(x)$.

Following the same approach used to define the Lebesgue integral, it is possible to define in $\mathbb{R}$ the integration with respect to a general $\mu \in \mathcal{P}(\mathbb{R})$ of μ-measurable functions $g : \mathbb{R} \to \mathbb{R}$ (that is if for all $t \in \mathbb{R}$, $g^{-1}(t, +\infty)$ is a set contained in the completion of $\mathcal{B}(\mathbb{R})$ with respect to μ).

A more intuitive way to define integration of continuous or monotone functions is via the Lebesgue-Stiltjes integral. Let $g : \mathbb{R} \to \mathbb{R}$ continuous (so it is surely μ-measurable). We define

$$\int_{\mathbb{R}} g(x)d\mu = \sup_{M>0} \sup \left\{ \sum_{i=0}^{k} (\min_{[x_i, x_{i+1}]} g)(F(x_{i+1}) - F(x_i)), -M \leqslant x_0 < x_1 < \cdots < x_k < x_{k+1} \leqslant M \right\}.$$

It is possible to show that if μ is an absolutely continuous measure with density f, then

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

On the other hand, if μ is associated to a discrete random variable, so $F = F^d$, with jumps given by the countable or finite set of points $(a_i)_i$, then

$$\int_{\mathbb{R}} g(x)d\mu = \sum_i g(a_i)(F(a_i) - F(a_i^-)).$$

Definition 1.8.1. *The nth-moment of a random variable X is given by $\mathbb{E}(X^n) := \int_{\mathbb{R}} x^n d\mathbb{L}_X$, more precisely*

- *if X is a (asbsolutely) continuous random variable (whose associated law has density f) then*

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

- *if X is a discrete random variable (taking values on $\mathbb{Z}$),*

$$\mathbb{E}(X^n) = \sum_{k \in \mathbb{Z}} k^n P(\omega \mid X(\omega) = k).$$

Note that $\mathbb{E}(X^n) < +\infty$ if and only if $\mathbb{E}(|X|^n) < +\infty$.

We recall that the moment for $n = 1$, that is $\mathbb{E}(X)$, is called the **mean**, *whereas $\mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$ is called the* **variance**.

1.9 Modes of convergence for random variables

We have several notion of convergence in the space of random variables and in the space $\mathcal{P}(\mathbb{R})$.

Definition 1.9.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X_n, X : \Omega \to \mathbb{R}$ be real random variables.*

- *X_n converges to X in probability if for every $\varepsilon > 0$, $\lim_n \mathbb{P}(|X_n - X| > \varepsilon) = 0$,*

– X_n *converges in mean to* X *if* $\mathbb{E}(|X_n - X|) \to 0$

– X_n *converges in mean square to* X *if* $\mathbb{E}((X_n - X)^2) \to 0$

– X_n *converges in distribution to* X *if* $\mathbb{E}(g(X_n)) \to \mathbb{E}(g(X))$ *for every bounded continuous function* g.

We recall also a notion of convergence on the space $\mathcal{P}(\mathbb{R})$.

Definition 1.9.2 (Weak convergence)**.** *Let* $\mu_n, \mu \in \mathcal{P}(\mathbb{R})$ *with* $\mu_n = \mathbb{L}_{X_n}, \mu = \mathbb{L}_X$ *with* $X_n, X : \Omega \to \mathbb{R}$ *random variables on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$.

Then μ_n *converges weakly to* μ *if* X_n *converges in distribution to* X *that is* $\int_{\mathbb{R}} g(x)d\mu_n \to \int_{\mathbb{R}} g(x)d\mu$ *for every bounded continuous function* g.

Theorem 1.9.3 (Prokhorov's theorem)**.** *Let* X_n *be a sequence of random variables which are* **tight** *in the following sense: for every* $\varepsilon > 0$ *there exist* $n_\varepsilon > 0$ *and a compact set* K_ε *(so a bounded closed set) such that* $\mathbb{P}\{\omega, X_n(\omega) \in K_\varepsilon\} \geqslant 1 - \varepsilon$ *for all* $n \geqslant n_\varepsilon$. *Then, there exists a random variable* X *such that, up to a subsequence,* $X_n \to X$ *in distribution.*

The same statement can be stated in the space $\mathcal{P}(\mathbb{R})$.

Theorem 1.9.4 (Prokhorov's theorem for laws)**.** *Let* $\mu_n \in \mathcal{P}(\mathbb{R})$ *be a sequence of probability measures which is* **tight** *in the following sense: for every* $\varepsilon > 0$ *there exist* $n_\varepsilon > 0$ *and a compact set* K_ε *(so a bounded closed set) such that* $\mu_n(K_\varepsilon) \geqslant 1 - \varepsilon$ *for all* $n \geqslant n_\varepsilon$ *(recall that* $\mu_n(\mathbb{R}) = 1$ *for all* n*). Then, there exists* $\mu \in \mathcal{P}(\mathbb{R})$ *such that, up to a subsequence,* $\mu_n \to \mu$ *weakly.*

1.10 Problems

(i) Let $f : \mathbb{R} \to \mathbb{R}$ be a monotone function. Show that f is Lebesgue measurable.

(ii) Consider the right continuous increasing function on $\mathbb{R}$

$$F(x) = \begin{cases} x & x < 0 \\ x + 1 & x \geqslant 0. \end{cases}$$

Which is the Borel measure associated to this function?

Chapter 2

Spaces of random variables with finite p-moment.

2.1 The Banach spaces M^p of random variables with finite moments

2.1.1 Banach spaces

Let X be a vectorial space on $\mathbb{R}$ (this means that it is closed by summation and by multiplication by scalars, that is if $x, y \in X$, $\lambda, \mu \in \mathbb{R}$, then $\lambda x + \mu y \in X$).

Definition 2.1.1. *A norm $\|\cdot\| : X \to [0, +\infty)$ is a function such that*

- *$\|x\| \geqslant 0$ for all $x \in X$ and $\|x\| = 0$ iff $x = 0$ (positivity);*
- *$\|\lambda x\| = |\lambda| \|x\|$ for all $x \in X, \lambda \in \mathbb{R}$ (homogeneity);*
- *$\|x + y\| \leqslant \|x\| + \|y\|$ (triangular inequality).*

$(X, \|\cdot\|)$ is a normed space.

Example 2.1.2. On $\mathbb{R}^n$ we may define the euclidean norm $|x| = \sqrt{x_1^2 + \cdots + |x_n|^2}$.

A norm induces on X a metric structure on X in the following way.

Definition 2.1.3 (Metric structure and notion of convergence)**.** *Let $(X, \|\cdot\|)$ be a normed space. We define a distance between elements in X as*

$$d(x, y) = \|x - y\|.$$

Note that this is a good definition, since it is positive, zero only if $x = y$, and satisfies the triangular inequality, that is $d(x, z) \leqslant d(x, y) + d(y, z)$ for all x, y, z.

We define the balls associated to this distance as follows: we fix a center $x_0 \in X$ and a radius $r > 0$ and we set

$$B(x_0, r) = \{x \in X \mid \|x - x_0\| < r\}.$$

A set $A \subseteq X$ is **open** *if for all $x \in A$ there exists $r > 0$ such that $B(x, r) \subseteq A$. A set C is* **closed** *is $X \backslash C$ is open.*

Let $(x_n)_n$ a sequence of element in X and $x \in X$. Then

$$\lim_n x_n = x \qquad \text{iff } \lim_{n \to +\infty} \|x_n - x\| = 0.$$

Proposition 2.1.4. *The following are equivalent:*

i) C is closed

ii) for every sequence (x_n) of elements in C such that there exists $x \in X$ with $\lim_n x_n = x$, there holds that $x \in C$.

Proof. Assume that C is closed and ii) is false. Then there exists (x_n) of elements in C such that $\lim_n x_n = x \notin C$. This implies that there exists $r > 0$ such that $B(x,r) \subseteq X \backslash C$. Therefore $x_n \notin B(x,r)$ for all n, which means that $\|x_n - x\| \geqslant r$ for all n, in contradiction with the fact that $\lim_n x_n = x$.

Assume that ii) holds and assume that C is not closed. So there exists $x \notin C$ such that for all $r > 0$ there holds that $B(x,r) \cap C \neq \varnothing$. Let $x_n \in C$ such that $x_n \in B(x, \frac{1}{n}) \cap C$. So $\|x_n - x\| < \frac{1}{n}$ and then $\lim_n x_n = x$. But this would imply $x \in C$. □

Definition 2.1.5 (Banach space)**.**
A sequence $(x_n)_n$ in X is a Cauchy sequence if $\lim_{n,m} \|x_n - x_m\| = 0$.
A normed space is called a Banach space if all the Cauchy sequences have limit in X.

Remark 2.1.6. Note that if $(x_n)_n$ is a sequence which converge to $x \in X$, then it is also a Cauchy sequence, since by triangular inequality $\|x_n - x_m\| \leqslant \|x_n - x\| + \|x - x_m\|$ and then $0 \leqslant \lim_{n,m\to+\infty} \|x_n - x_m\| \leqslant \lim_{m,n\to+\infty} \|x_n - x\| + \|x - x_m\| = 0$.

The viceversa is not always true. Let's think e.g. of the case $X = \mathbb{Q}$ and the euclidean norm. Define (x_n) as follows: $x_0 = 1$, $x_1 = 1,01$, $x_2 = 1,01001$, $x_3 = 1,010010001$, $x_4 = 1,01001000100001$ and so on, that is $x_n = 1,1010010001\ldots1\,0\overset{n}{\ldots}0\,1$. It is easy to check that $x_n \in \mathbb{Q}$ for all n, that $x_n \to x$ (so $(x_n)_n$ is a Cauchy sequence, but this can also be checked directly) and that $x \notin \mathbb{Q}$. This implies that $(\mathbb{Q}, |\cdot|)$ is not a Banach space.

An important theorem in Banach spaces (more generally in complete metric spaces) is the **contraction lemma,** or **Banach-Caccioppoli theorem**:

Theorem 2.1.7. *Let $(X, \|\cdot\|)$ a Banach space and $F : X \to X$ such that there exists $0 < a < 1$ for which*

$$\|F(x) - F(y)\| \leqslant a\|x - y\| \qquad \forall x, y \in X.$$

(F is a contraction) Then the map F admits a unique fixed point, that is a point such that $\bar{x} = F(\bar{x})$.

Proof. See problem 1 at the end of the chapter. □

Definition 2.1.8. *Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be two Banach space.*

A **linear operator** *is a map $T : X \to Y$ such that $T(\alpha x + \beta y) = \alpha T(x) + \beta T(y)$ for all $\alpha, \beta \in \mathbb{R}$, $x, y \in X$.*

A **bounded operator** *is a map $T : X \to Y$ such that*

$$\|T\| = \sup_{\{x \in X \| x\| \leqslant 1\}} \|Tx\| < +\infty.$$

If this quantity if finite, it is called the norm of T.

A **continuous operator** *is a map $T : X \to Y$ such that*

$$\lim Tx_n = Tx \qquad \text{for all sequences } x_n \text{ such that } \lim_n x_n = x.$$

Proposition 2.1.9. *A linear operator $T : X \to Y$ is continuous if and only if it is bounded.*

Proof. Assume that T is bounded, then

$$\|Tx_n - Tx\| = \|T(x_n - x)\| = \|x_n - x\| T\left(\frac{x_n - x}{\|x_n - x\|}\right) \leqslant \|x_n - x\|\|T\|.$$

Therefore if $\|x_n - x\| \to 0$, then also $\|Tx_n - Tx\| \to 0$.

Assume that T is continuous, and we want to prove that T is bounded. Assume by contradiction that it is not true. So for every $n \in \mathbb{N}$ there exists $x_n \in X$ such that $\|x_n\| = 1$ and $\|Tx_n\| \geqslant n$. Define $y_n = \frac{x_n}{n}$. Then $\|y_n\| = \frac{\|x_n\|}{n} = \frac{1}{n} \to 0$. This implies that $y_n \to 0$. Observe that by linearity $Ty_n = \frac{1}{n}Tx_n$ and then $\|Ty_n\| = \frac{1}{n}\|Tx_n\| \geqslant \frac{n}{n} = 1$. Therefore $y_n \to 0$ but $Ty_n \not\to 0$, in contradiction with continuity. □

Theorem 2.1.10. *The set of all bounded linear operators between two Banach spaces X, Y, is a Banach space $\mathcal{B}(X,Y)$, with norm $\|T\|$ as defined above.*

Proof. See [2, Theorem 2.12]. □

Example 2.1.11. Let $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$ both with the euclidean norm. Let $\mathbf{A} \in M_{m\times n}(\mathbb{R})$ be a $n \times m$ matrix. Then

$$Tx = \mathbf{A}x = (\sum_{j=1}^{n} a_{ij}x_j)_{i=1,\dots,m}$$

is a bounded linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$.

Theorem 2.1.12 (Uniform boundedness principle, or Banach-Steinhaus theorem)**.** *Let T_n be a sequence of bounded linear operators between the Banach spaces X and Y, that is $T_n \in \mathcal{B}(X, Y)$ for all n. Assume that for all $x \in X$ there exists $C_x \in \mathbb{R}$ such that $\sup_n \|T_n x\| \leqslant C_x$.*

Then there exists $C \in \mathbb{R}$ such that $\|T_n\| \leqslant C$ for all n.

In particular this implies that if the sequence $T_n x$ is convergent for every $x \in X$, then $Tx := \lim_n T_n x$ defines a bounded linear operator.

Proof. See [2, Theorem 4.1]. □

2.1.2 Spaces of random variables with finite moments

We fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and we consider the random variables $X : \Omega \to \mathbb{R}$. We introduce the spaces of random variables with finite p-moment (see definition in Section 2.7)

$$M^p = \{X \text{ random variable } \mathbb{E}(|X|^p) < +\infty\}$$

with associated norm $\|X\|_p = (\mathbb{E}(|X|^p))^{1/p}$.

Definition 2.1.13. *Let $p > 1$. Then the conjugate exponent of p is the number $q > 1$ such that $1/p+1/q = 1$. In particular the conjugate exponent of 2 is 2.*

We say that the conjugate exponent of 1 is $+\infty$.

Lemma 2.1.14 (Young inequality)**.** *Let p, q be conjugate exponents. Then $xy \leqslant x^p/p + y^q/q$ for all $x, y \geqslant 0$.*

Proof. Fix $x > 0$ and consider $\sup_{y\geqslant 0}(xy - y^q/q)$. First of all observe that the supremum is actually a maximum, since $\lim_{y\to+\infty} xy - y^q/q = -\infty$. Differentiating in y, we get that the unique point where the derivative is 0 is given by $y = x^{1/(q-1)}$. This is the maximum. Therefore for all $y \geqslant 0$, $xy - y^q/q \leqslant x^{1+1/(q-1)} - x^{q/(q-1)}/q = x^p/p$, since $p = q/(q-1)$. □

Proposition 2.1.15. *Let $X \in M^p$ and $Y \in M^q$, with q conjugate exponent of p, then*

$$\mathbb{E}(|XY|) \leqslant \mathbb{E}(|X|^p))^{1/p}(\mathbb{E}(|Y|^q))^{1/q}.$$

Moreover if $X, Z \in M^p$, then

$$\mathbb{E}(|X + Z|^p))^{1/p} \leqslant \mathbb{E}(|X|^p))^{1/p} + \mathbb{E}(|Z|^p))^{1/p}$$

Proof. It is sufficient to apply the Young inequality to $x = |X|\mathbb{E}(|X|^p))^{-1/p}$ and to $y = |Y|\mathbb{E}(|Y|^q))^{-1/q}$ and one obtain

$$\frac{|X|}{\mathbb{E}(|X|^p))^{1/p}}\frac{|Y|}{\mathbb{E}(|Y|^q))^{1/q}} \leqslant \frac{|X|^p}{p\mathbb{E}(|X|^p)} + \frac{|Y|^q}{q\mathbb{E}(|Y|^q)}.$$

By applying $\mathbb{E}$ to both term we conclude

$$\frac{\mathbb{E}(|XY|)}{\mathbb{E}(|X|^p))^{1/p}\mathbb{E}(|Y|^q))^{1/q}} \leqslant \frac{1}{p} + \frac{1}{q} = 1.$$

Observe that if $X, Y \in M^p$ the $X + Y \in M^p$. This is due to the fact that

$$\frac{|X + Y|^p}{2^p} = \left|\frac{X}{2} + \frac{Y}{2}\right|^p \leqslant \frac{|X|^p}{2} + \frac{|Y|^p}{2}$$

by the convexity of the function $r \mapsto r^p$ on $[0, +\infty)$ when $p \geqslant 1$. Now we observe that

$$|X + Y|^p = |X + Y||X + Y|^{p-1} \leqslant |X||X + Y|^{p-1} + |X||X + Y|^{p-1}$$

and that $|X+Y|^{p-1} \in M^q$ where $q = \frac{p}{p-1}$ is the conjugate exponent of p. Moreover

$$\mathbb{E}\left(|X+Y|^{p-1}\right)^q = \mathbb{E}|X+Y|^p. \tag{2.1.1}$$

So by Holder inequality applied to f and $|f+g|^{p-1}$ we get

$$\mathbb{E}|X||X+Y|^{p-1} \leqslant (\mathbb{E}|X|^p)^{\frac{1}{p}} (\mathbb{E}|X+Y|^p)^{\frac{p-1}{p}}$$

and analogously by Holder inequality applied to f and $|f+g|^{p-1}$ we get

$$\mathbb{E}|Y||X+Y|^{p-1}dx \leqslant (\mathbb{E}|Y|^p)^{\frac{1}{p}} (\mathbb{E}|X+Y|^p)^{\frac{p-1}{p}}.$$

Integrating (2.1.1) and using the previous inequalities we get

$$\begin{aligned}\mathbb{E}|X+Y|^p &\leqslant (\mathbb{E}|X|^p)^{\frac{1}{p}} (\mathbb{E}|X+Y|^p)^{\frac{p-1}{p}} \\ &\quad + (\mathbb{E}|Y|^p)^{\frac{1}{p}} (\mathbb{E}|X+Y|^p)^{\frac{p-1}{p}} \\ &= (\mathbb{E}|X+Y|^p)^{\frac{p-1}{p}} \left[(\mathbb{E}|X|^p)^{\frac{1}{p}} + (\mathbb{E}|Y|^p)^{\frac{1}{p}}\right]\end{aligned}$$

from which we deduce the statement by dividing both sides by $(\mathbb{E}|X+Y|^p)^{\frac{p-1}{p}}$. □

Theorem 2.1.16. *The space M^p with the norm $\|X\|_p$ for $p \in [1, +\infty)$ is a Banach space.*

We recall the Jensen inequality:

Lemma 2.1.17 (Jensen's inequality)**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a convex function, then for every random variable X*

$$\mathbb{E}(g(X)) \geqslant g(\mathbb{E}(X)).$$

Theorem 2.1.18. *There holds that $M^k \subseteq M^n$ for every $1 \leqslant n \leqslant k$. Moreover if $X \in M^k$ then $(\mathbb{E}(|X|^n))^{\frac{1}{n}} \leqslant (\mathbb{E}(|X|^k))^{\frac{1}{k}}$ for all $n \leqslant k$.*

Proof. Let $1 \leqslant n \leqslant k$, $g(x) = |x|^{\frac{k}{n}}$. Since $\frac{k}{n} \geqslant 1$, the function g is convex. Let $X \in M^k$ and we apply Jensen's inequality to the random variable $|X|^n$, observing that $g(|X|^n) = |X|^k$,

$$\mathbb{E}(|X|^k) = \mathbb{E}(g(|X|^n)) \geqslant g(\mathbb{E}(|X|^n)) = (\mathbb{E}(|X|^n))^{\frac{k}{n}}.$$

□

Example 2.1.19. $T : M^k \to \mathbb{R}$ such that $T(X) = \mathbb{E}(X)$ is a bounded linear operator.

If we consider $X \in M^2$, then $T_X : M^2 \to \mathbb{R}$ defined as $T_X(y) = \mathbb{E}(XY)$ is again a bounded linear operator.

2.2 Hilbert space M^2 and conditional expectation

2.2.1 Hilbert spaces

Hilbert spaces are spaces where it is possible to define the notions of length and orthogonality, which allow to work with the elements geometrically, as if they were vectors in Euclidean space. First of all we recalls some basic definitions.

Definition 2.2.1. *A set X is a vector space on $\mathbb{R}$ (a real vector space) if it is a set equipped with two operations, vector addition (which allows to add two vectors $x, y \in X$ to obtain another vector $x + y \in X$) and scalar multiplication (which allows us to "scale" a vector $x \in X$ by a real number c to obtain a vector $cx \in X$). Moreover we require that X contains a neutral element for the vector addiction, that is an element $0 \in X$ such that $0 + x = x$ for every $x \in X$ and $x - x = 0$.*

A **scalar product** *on X is a function $(\cdot, \cdot) : X \times X \to \mathbb{R}$ such that*

- *$(x, x) \geqslant 0$ for all x and $(x, x) = 0$ iff $x = 0$;*
- *it is symmetric $(x, y) = (y, x)$ for all $x, y \in X$;*
- *it is linear, that is $(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$ for all $x, y, z \in X, \alpha, \beta \in \mathbb{R}$.*

We associate to a scalar product a norm in this way $\|x\| = \sqrt{(x,x)}$.

Proposition 2.2.2. *The function $\|\cdot\| : X \to [0,+\infty)$ defined as $\|x\| = \sqrt{(x,x)}$ is a norm. Moreover the scalar product is continuous, that is if $x_n \to x$ in X and $y \in X$, then $(x_n, y) \to (x, y)$ in $\mathbb{R}$.*

Proof. Positivity and homogeneity are obvious. To prove the triangle inequality one first need to to prove the Cauchy Schwartz inequality $|(x,y)| \leqslant \|x\|\|y\|$. See [2, Theorem 5.1].

The continuity is an easy consequence of the Cauchy Schwartz inequality:

$$|(x_n - x, y)| \leqslant \|x_n - x\|\|y\|.$$

□

Definition 2.2.3 (Hilbert space)**.** *A space X with a scalar product which induces on X a norm such that X is a Banach space is called Hilbert space.*

Proposition 2.2.4 (Parallelogram identity)**.** *For every $x, y \in H$, there holds*

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2.$$

Proof. By definition and by linearity and symmetry of the scalar product $\|x+y\|^2 = (x+y, x+y) = (x,x) + 2(x,y) + (y,y) = \|x\|^2 + 2(x,y) + \|y\|^2$, and $\|x-y\|^2 = (x+y, x+y) = \|x\|^2 - 2(x,y) + \|y\|^2$. It is sufficient to sum. □

Example 2.2.5. In $\mathbb{R}^n$ we define the scalar product $(x,y) = x_1y_1 + x_2y_2 + \cdots + x_ny_n$. The euclidean norm is the norm associated to this scalar product. So $\mathbb{R}^n$ with this scalar product is a Hilbert space. This is the basic example of Hilbert space of finite dimension.

2.2.2 Orthogonality and projections in Hilbert spaces

Definition 2.2.6 (Orthogonal space)**.** *We say that $x, y \in X$ are orthogonal if $(x,y) = 0$.*

If $S \subseteq X$ is a subset of X, we define the orthogonal subspace

$$S^\perp = \{x \in X \mid (x,s) = 0 \ \forall s \in S\}.$$

This a vectorial subspace of X.

Example 2.2.7. *If we consider $S \subset M^2$ the subspace of constant random variables, then $S^\perp = \{X \in M^2 \ |\mathbb{E}(X) = 0\}$.*

Theorem 2.2.8 (Orthogonal projection)**.** *Let $V \subseteq H$ be a closed subspace of a Hilbert space, $V \neq \{0\}$ and let $h \in H$.*

Then there exists a unique element $v \in V$ at minimal distance from h, that is such that $\|h - v\| = \min_{w \in V} \|h - w\|$. Moreover there exists a unique element $s \in V^\perp$ such that $h = v + s$.

The map $Pr_V : H \to V$ which associate $h \to v$ is called the orthogonal projection of H in V and it is a bounded linear operator of norm 1.

Proof. We consider the minimization problem $\min_{w \in V} \|h - w\|$ and we show that it admits a solution which is unique. Since $\|h - w\| \geqslant 0$ we get that $\inf_{w \in V} \|h - w\| = \delta \geqslant 0$. Let $v_n \in V$ such that $\delta \leqslant \|v_n - h\| \leqslant \delta + 1/n$. Then $(v_n)_n$ is a Cauchy sequence, since by parallelogram identity and linearity

$$\|v_n - v_m\|^2 = 2\|v_n - h\|^2 + 2\|v_m - h\|^2 - \|(v_n + v_m) - 2h\|^2 \leqslant 2(\delta + 1/n)^2 + 2(\delta + 1/m)^2 - 4\|h - (v_n + v_m)/2\|^2.$$

We conclude by recalling that since $(v_n + v_m)/2 \in V$ then $\|h - (v_n + v_m)/2\| \geqslant \delta$,

$$\|v_n - v_m\|^2 \leqslant 2(\delta + 1/n)^2 + 2(\delta + 1/m)^2 - 4\delta^2 = 4\delta/n + 4\delta/m + 1/n^2 + 1/m^2 \to 0 \qquad \text{as } n, m \to +\infty.$$

Since H is a Banach space there exists $v \in H$ such that $\lim_n v_n = v$ and since V is closed then $v \in V$. By continuity, we conclude that $\|v - h\| = \delta = \inf_{w \in V} \|h - w\|$. v is the unique minimizer. Indeed if it were not the case, there would exists $v' \in V$ with $\|v - h\| = \|v' - h\| = \delta$. By parallelogram identity

$$\|v - v'\|^2 = 2\|v - h\|^2 + 2\|v' - h\|^2 - 4\|(v + v')/2 - h\|^2 \leqslant 2\delta^2 + 2\delta^2 - 4\delta^2 = 0$$

which implies $\|v - v'\| = 0$.

Let $w \in V$. We claim that $(h - v, w) = 0$. Since v is the point at minimum distance, then the function $\lambda \to \|h - v + \lambda w\|^2$ has minimum in $\lambda = 0$. Differentiating the function in λ it should be that the derivative in 0 is 0. $\frac{\|h-v+\lambda w\|^2}{d\lambda} = \frac{(h-v+\lambda w, h-v+\lambda w)}{d\lambda} = 2(h - v, w)$. Therefore $(h - v, w) = 0$. This means that $h - v \in V^\perp$.

Let $v = Pr_V(h), v' = Pr_V(h')$ and let $\alpha, \beta \in \mathbb{R}$. Then $\alpha v + \beta v' \in V$ and $\alpha v + \beta v' - \alpha h - \beta h' \in V^\perp$. Therefore by uniqueness $Pr_V(\alpha h + \beta h') = \alpha v + \beta v'$. Then Pr_V is linear. Moreover since $(Pr_V h - h, Pr_V h) = 0$,

$$\|h\|^2 = \|h - Pr_V h + Pr_V h\|^2 = (h - Pr_V h + Pr_V h, h - Pr_V h + Pr_V h) = \|h - Pr_V h\|^2 + \|Pr_V h\|^2.$$

This implies that for all h with $\|h\| \leqslant 1$, $\|Pr_V h\|^2 = \|h\|^2 - \|h - Pr_V h\|^2 \leqslant 1$. So Pr_V is bounded. Moreover if $h \in V$, then $Pr_V h = h$. Therefore $\|Pr_V\| = 1$.

□

Definition 2.2.9 (Orthonormal set)**.** *A set $\{u_i, i \in I\}$ of elements in H is an orthonormal set if $\|u_i\| = 1$ for all i and $(u_i, u_j) = 0$ for all $i \neq j$.*

Proposition 2.2.10. *Let $\{u_i, i \in I\}$ be a orthonormal set. Then the following are equivalent*

- *if $(x, u_i) = 0$ for all i, then $x = 0$*
- *$\|x\|^2 = \sum_i |(x, u_i)|^2$ for all $x \in H$,*
- *for all $x \in H$, $x = \sum_i (x, u_i)u_i$, (where the convergence is with respect to the norm of H).*

An orthonormal set for which one of the previous conditions hold is called an orthonormal basis. Every Hilbert space admits a orthonormal basis.

Proof. See [3, Proposition 5.28]. □

Definition 2.2.11 (Separable space)**.** *H is separable if it admits a countable orthonormal basis.*

Theorem 2.2.12 (Computation of the orthogonal projection)**.** *Let V be a closed subspace of H and let $\{v_i, i \in I\}$ be an orthonormal basis of V. Then for all $h \in H$,*

$$Pr_V(h) = \sum_{i \in I} (h, v_i)v_i.$$

Proof. See [2, Theorem 5.10]. □

Theorem 2.2.13 (Parseval theorem)**.** *Let $\{u_i, i \in I\}$ be a countable orthonormal set in H. The following are equivalent*

- *if $(h, u_i) = 0$ for all i then $h = 0$,*
- *for each $h \in H$ there holds $h = \sum_i (h, u_i)u_i$, which means that $\lim_n \|h - \sum_{i=1}^n (h, u_i)u_i\| = 0$,*
- *for each $h \in H$, $\|h\|^2 = \sum_i |(h, u_i)|^2$.*

In particular $\{u_i, i \in I\}$ is an orthonormal basis of H.

2.2.3 Conditional expectation

We fix a probability space $(\Omega, \mathbb{P}, \mathcal{F})$ and we define the space

$$M^2 = \{X : (\Omega, \mathbb{P}, \mathcal{F}) \to \mathbb{R} \mid \ X \text{ random variable with } \mathbb{E}(X^2) < +\infty\}.$$

Recall that X is a random variable if $X^{-1}(A) \in \mathcal{F}$ for every $A \in \mathcal{B}$ (so for every A in the $\sigma-$algebra of Borel sets. Given X random variable, we define $\sigma(X) \subseteq \mathcal{F}$, that is **the σ-algebra generated by** X, as the minimal σ- algebra contained in $\mathcal{F}$ which contains all the elements $X^{-1}(A) = \{\omega \in \Omega \mid X(\omega) \in A\}$ for every $A \in \mathcal{B}$. So it is the minimal σ-algebra which assures that X is measurable.

Note that if X is a constant random variable, so $X(\omega) = c$ for all $\omega \in \Omega$, then $X^{-1}(A) = \Omega$ if $c \in A$, and $X^{-1}(A) = \varnothing$ if $c \notin A$. So in this case $\sigma(X) = \{\varnothing, \Omega\}$, which is the minimal possible σ-algebra.

We define on M^2 the scalar product

$$(X, Y) = \mathbb{E}(XY) = \int_{\mathbb{R}} xy d\mathbb{L}_{(X,Y)}(x, y)$$

and the induced norm is

$$\|X\| = \sqrt{\mathbb{E}(X^2)}.$$

It is possible to prove that M^2 with this norm and this scalar product is a Hilbert space. Observe that we are actually considering class of equivalence of random variables, since we are identifying two random variables X, Y such that $\mathbb{P}(\omega \mid X(\omega) = Y(\omega)) = 1$.

We consider a σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, and consider the probability space $(\Omega, \mathbb{P}, \mathcal{G})$. On this space we may define the space

$$M^2_{\mathcal{G}} = \{X : (\Omega, \mathbb{P}, \mathcal{G}) \to \mathbb{R} \mid \ X \text{ random variable with } \mathbb{E}(X^2) < +\infty\}.$$

Note that $M^2_{\mathcal{G}}$ is a closed subspace of M^2.

Definition 2.2.14 (Conditional expectation)**.** *We define the* **conditional expectation of X given $\mathcal{G}$** *as the orthogonal projection of $X \in M^2$ in the space $M^2_{\mathcal{G}}$ as defined and characterized in Theorem 2.2.8 that is*

$$\mathbb{E}(X|\mathcal{G}) = Pr_{M^2_{\mathcal{G}}}(X),$$

or equivalently $\mathbb{E}(X|\mathcal{G})$ is the unique random variable in $M^2_{\mathcal{G}}$ such that

$$\mathbb{E}(X - \mathbb{E}(X|\mathcal{G}))^2 = \min_{Z \in M^2_{\mathcal{G}}} \mathbb{E}(X - Z)^2.$$

In particular $\mathbb{E}(X|\mathcal{G})$ is the minimum mean squared predictor of X based on the information contained in $\mathcal{G}$.

Note that $X - \mathbb{E}(X|\mathcal{G})$ is orthogonal to every element of $M^2_{\mathcal{G}}$ that is

$$\mathbb{E}(XY) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})Y) \qquad \forall\ Y \in M^2_{\mathcal{G}}.$$

In particular, since constant random variables are in $M^2_{\mathcal{G}}$ for every $\mathcal{G}$, we get $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|\mathcal{G}))$.

Remark 2.2.15 (Conditioning with respect to a random variable X)**.** A particular case of the previous definition is the following. Let us consider a random variable $X \in M^2$, and let $\mathcal{G} = \sigma(X)$ as before. It is possible to show that in this case every $\mathcal{G}$ measurable random variable is a Borel function of X, which means that

$$M^2_{\mathcal{G}} := \{h(X), \text{ for } h : \mathbb{R} \to \mathbb{R}, \text{borelian function}\}.$$

$h : \mathbb{R} \to \mathbb{R}$ is a Borel function if for all borelian set $B \subseteq \mathcal{B}(\mathbb{R})$, the set $h^{-1}(B) := \{x \in \mathbb{R}\ h(x) \in B\}$ is in the Borel σ-algebra (Note that this condition is slightly stronger than asking that h is measurable, since measurable functions satisfies $h^{-1}(B) := \{x \in \mathbb{R}\ h(x) \in B\} \in \mathcal{M}$, that is are elements of the σ-algebra of measurable sets (given by sets which differs from Borel sets by subsets of sets of zero Lebesgue measure).

In this case $\mathbb{E}(Y|\sigma(X)) = \mathbb{E}(Y|X)$ is the best predictor of Y given X. In particular $\mathbb{E}(Y|X)$ the unique Borel function $h(X)$ which minimizes $\mathbb{E}(Y - h(X))^2$:

$$\mathbb{E}[(Y - \mathbb{E}(Y|X))^2] = \mathbb{E}[(Y - h(X))^2] = \min_{f:\mathbb{R}\to\mathbb{R}, \text{borelian}} \mathbb{E}[(Y - f(X))^2]$$

and moreover

$$\mathbb{E}(Y f(X)) = \mathbb{E}(h(X)f(X)) \qquad \forall f : \mathbb{R} \to \mathbb{R}.\ \text{borelian}.$$

Note that solving this minimization problem can be very difficult, so in general we consider a reduced problem, adding some conditions on the functions f on which we are minimizing.

The simplest case is the case in which we consider the minimization problem among *linear functions*: that is

$$\min_{f:\mathbb{R}\to\mathbb{R}, \text{linear}} \mathbb{E}[(Y - f(X))^2].$$

$h : \mathbb{R} \to \mathbb{R}$ is linear if and only if there exists $a, b \in \mathbb{R}$ such that $h(r) = ar + b$. So the problem reduced to a finite dimensional problem: given $X \in M^2$ we want to find for all Y, $a, b \in \mathbb{R}$ for which it is minimal $\mathbb{E}((Y - a - bX)^2)$. So, the **linear least square estimator** is given by

$$L(Y|X) = a + bX,$$

where a, b are the optimal values which minimize $\mathbb{E}((Y - a - bX)^2)$. This problem can be restated exactly as a projection problem: we define S as the space generated by $X, 1$ in M^2, that is $S = \{Z = aX + b \in M^2, a \in \mathbb{R}, b \in \mathbb{R}\}$ and we want to find $Pr_S(Y)$.

In order to solve the problem, first of all we choose an orthonormal basis of S. A basis of S is given by $\{1, X\}$. Observe that if $\mathbb{E}(X) = (X, 1) \neq 0$, we have that X and 1 are not orthogonal, so we substitute X with the element $X - \mathbb{E}(X)$ which is orthogonal to 1. Moreover we have to normalize this element by choosing $c \in \mathbb{R}$ such that $c^2\mathbb{E}(X - \mathbb{E}(X))^2 = 1$. Since $\mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = Var(X)$, it is sufficient to choose $c = \sqrt{VarX}$. Therefore an orthonormal basis of S is given by $1, \frac{X-\mathbb{E}(X)}{\sqrt{Var(X)}}$. Recalling Theorem 2.2.12, we get

$$Pr_S(Y) = (Y, 1)1 + \left(Y, \frac{X - \mathbb{E}(X)}{\sqrt{Var(X)}}\right) \frac{X - \mathbb{E}(X)}{\sqrt{Var(X)}}.$$

So the linear least square estimator coincides with

$$L(Y|X) = \mathbb{E}(Y) + \frac{Cov(X, Y)}{Var(X)}(X - \mathbb{E}(X))$$

where $Cov(X,Y) = \mathbb{E}(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$. Finally we compute the average error

$$\begin{aligned}\mathbb{E}(Y - L(Y|X))^2 = Var(Y) + \frac{Cov^2(X, Y)}{Var^2(X)} VarX - 2\frac{Cov(X, Y)}{Var(X)} Cov(X, Y) \\ = VarY - \frac{Cov^2(X, Y)}{Var(X)} = \frac{Var(Y)Var(X) - Cov^2(XY)}{Var(X)}.\end{aligned}$$

In general the best linear predictor is different from the general minimum mean squared predictor. Let $Y = X^2 + Z$ with X, Z independent and both normals with mean 0 and variance 1. Then $\mathbb{E}(Y|X) = X^2$, whereas $L(Y|X) = 1$ (check it!).

Remark 2.2.16 (Conditioning with respect to a constant random variable)**.** A very simple case to compute $\mathbb{E}(Y|\sigma(X)) = \mathbb{E}(Y|X)$ is the case in which $X \equiv k$ (that is X is constant). In this case $\sigma(X) = \{\varnothing, \Omega)$ and the space

$$M^2_{\mathcal{G}} := \{\text{constant random variables}\}.$$

So, $\mathbb{E}(Y|X)$ is the unique constant c such that

$$\mathbb{E}[(Y - c)^2] = \min_{\lambda \in \mathbb{R}} \mathbb{E}[(Y - \lambda)^2]$$

and moreover

$$\lambda\mathbb{E}(Y) = \mathbb{E}(Y\lambda) = \mathbb{E}(c\lambda) = c\lambda \qquad \forall \lambda \in \mathbb{R}.$$

It is immediate to verify that $c = \mathbb{E}(Y|\mathcal{G}) = \mathbb{E}(Y)$. Another simple case is the case in which $X = \chi_A$, for some $A \in \mathcal{F}$ which means that $\chi_A(\omega) = 1$ if $\omega \in A$ and $\chi_A(\omega) = 0$ if $\omega \notin A$. It is simple to see that in this case $\sigma(\chi_A) = \{\varnothing, \Omega, A, \Omega\backslash A\}$. In this case

$$M^2_{\mathcal{G}} := \{a\chi_A + b\chi_{\Omega\backslash A} = (a - b)\chi_A + b \qquad a, b \in \mathbb{R}\}.$$

So, $\mathbb{E}(Y|A)$ is obtained by solving the finite dimensional minimization problem

$$\min_{a,b \in \mathbb{R}} \mathbb{E}[(Y - a\chi_A - b)^2].$$

Since $M^2_{\mathcal{G}}$ is a finite dimensional space (of dimension 2), we compute a orthonormal basis of it. We start from the basis given by $\{1, \chi_A\}$ and we orthonormalize it by Gram-Schmidt procedure. Let $X_1 = 1$ and $X_2 = \frac{\chi_A - \mathbb{P}(A)}{\sqrt{\mathbb{P}(A)(1-\mathbb{P}(A))}}$. Note that $\mathbb{E}|X_1|^2 = 1 = \mathbb{E}|X_2|^2$ and moreover $\mathbb{E}(X_1X_2) = 0$. Therefore by Theorem 2.2.12 we deduce that

$$\mathbb{E}(Y|A) = \mathbb{E}(YX_1)X_1 + \mathbb{E}(YX_2)X_2 = \mathbb{E}(Y) + \frac{\mathbb{E}(Y\chi_A)}{\mathbb{P}(A)(1 - \mathbb{P}(A))}\chi_A - \mathbb{E}(Y)\frac{\mathbb{P}(A)}{\mathbb{P}(A)(1 - \mathbb{P}(A))} =$$

$$= \frac{\mathbb{E}(Y\chi_A)}{\mathbb{P}(A)(1 - \mathbb{P}(A))}\chi_A - \frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)}\mathbb{E}(Y).$$

2.3 Metric spaces of laws of random variables and basics of optimal transport

Up to now we fixed a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and we considered the random variables $X : \Omega \to \mathbb{R}$. with finite p-moment $M^p = \{X \text{ random variable } \mathbb{E}(|X|^p) < +\infty\}$ with associated norm $\|X\|_p = (\mathbb{E}(|X|^p))^{1/p}$. We showed that these spaces are Banach spaces (as normed linear spaces of random variables defined on a fixed probability space Ω) and in case $p = 2$ are also Hilbert.

Another point of view is possible. Actually we may directly work on spaces of Borel measures on $\mathbb{R}$ which are laws of some random variable. In this way, we have not to fix a given probability space (so a sample set), and we have much more freedom.

2.3.1 Space of probability measures (laws of random variables)

Let us recall the definition of the space of laws of random variables:

$$\begin{aligned}\mathcal{P}(\mathbb{R}) &= \{\mu \mid \text{there exists a probability space } (\Omega, \mathcal{F}, \mathbb{P}) \text{ and } X : \Omega \to \mathbb{R} \text{ random variable s.t. } \mu = \mathbb{L}_X\} \\ &= \{\mu \mid \text{Borel measure on } \mathbb{R} \text{ s.t. } \mu(\mathbb{R}) = 1\}.\end{aligned}$$

The second equality is completely not obvious: it is the consequence of the following result (see e.g. [1, Proposition 9.1.11]).

First we recall some definition. On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we say that $A \in \mathcal{F}$ is an atom if $\mathbb{P}(A) > 0$ and for all $B \in \mathcal{F}$ with $B \subseteq A$ and $\mathbb{P}(B) < \mathbb{P}(A)$, it holds $\mathbb{P}(B) = 0$. So in an atomless probability space for any $A \in \mathcal{F}$ with $\mathbb{P}(A) > 0$ there exists $B \subseteq A$, $B \in \mathcal{F}$, with $0 < \mathbb{P}(B) < \mathbb{P}(A)$.

Proposition 2.3.1. *Let μ be a Borel measure on $\mathbb{R}^n$, with $\mu(\mathbb{R}^n) = 1$. Then there exists an atomless probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}^n$ such that $\mathbb{L}_X = \mu$.*

In particular we get that the space of probability measures on $\mathbb{R}^n$ coincide with the space of all laws associated to some random variable (with values in $\mathbb{R}^n$). One of the most used is the **Total variation** distance:

$$d_{TV}(\mu, \nu) = 2 \sup_{A \in \mathcal{B}(\mathbb{R})} |\mu(A) - \nu(A)| = 2 \inf_{\mathbb{L}_X = \mu, \mathbb{L}_Y = \nu} \mathbb{P}\{\omega \in \Omega | X(\omega) \neq Y(\omega)\}.$$

Another important distance is the Lévy-Prokhorov distance, which is the distance associated to the weak convergence of probability measures.

$$d_{LP}(\mu, \nu) = \inf \left\{ \varepsilon > 0 : \inf_{\mathbb{L}_X = \mu, \mathbb{L}_Y = \nu} \mathbb{P}\{\omega \in \Omega | |X(\omega) - Y(\omega)| > \varepsilon\} < \varepsilon \right\}.$$

2.3.2 Couplings between measures and deterministic couplings

We introduce the following definition

Definition 2.3.2 (Coupling between measures)**.** *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$. A coupling π between μ and ν is a probability measure $\pi \in \mathcal{P}(\mathbb{R}^2)$ such that the first marginal of π is μ and the second marginal is ν, that is for all $A \in \mathcal{B}(R)$ it holds $\pi(A \times \mathbb{R}) = \mu(A)$, $\pi(\mathbb{R} \times A) = \nu(A)$. We denote $\Pi(\mu, \nu)$ the family of all couplings between μ, ν.*
For any $\pi \in \Pi(\mu, \nu)$, it is possible to find $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space, $X, Y : \Omega \to \mathbb{R}$ random variables, such that $\pi = \mathbb{L}_{(X,Y)}$.
In optimal transport theory a coupling $\pi \in \Pi(\mu, \nu)$ is also called a transport plan between μ and ν.

A particular class of couplings are the one associated to transport maps:

Definition 2.3.3 (Deterministic coupling)**.** *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ and $\psi : (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu) \to \mathbb{R}$ be a measurable map (e.g a monotone map or a continuous map). Then ψ is a transport map if $\psi_\sharp \mu = \nu$, that is for all $A \in \mathcal{B}(\mathbb{R})$, it holds*

$$\nu(A) = \mu\{x, \psi(x) \in A\}.$$

We associate to ψ a coupling called deterministic coupling and defined as $(Id, \psi)_\sharp \mu$ where $(Id, \psi) : \mathbb{R} \to \mathbb{R} \times \mathbb{R}$ is defined as $(Id, \psi)(x) = (x, \psi(x))$.

If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, with $X : \Omega \to \mathbb{R}$ random variable with law $\mathbb{L}_X = \mu$, then $Y = \psi(X)$ is a random variable with law ν and $(Id, \psi)_\sharp \mu = \mathbb{L}_{(X, \psi(X))}$.

Obviously the previous definition can be extended to the case $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$. In case of dimension 1, nonetheless, we may also use the cumulative distribution function.

Let $\mu \in \mathcal{P}(\mathbb{R})$ and let $F_\mu(x) = \mu(-\infty, x]$ the associated cumulative distribution function. Then $F_\mu : \mathbb{R} \to [0,1]$ is monotone increasing function and right continuous. We may define a pseudo-inverse of F_μ as follows:

$$F_\mu^-(x) := \inf\{t, F_\mu(t) \geqslant x\} \qquad F_\mu^- : [0,1] \to \mathbb{R}.$$

It is easy to show the following properties, by using the definition:

Lemma 2.3.4. *Let $\mathcal{L}_{|[0,1]}$ be the Lebesgue measure restricted to the interval $[0,1]$.*

(i) *$F_\mu^-(x) \leqslant a$ if and only if $F_\mu(a) \geqslant x$ and $F_\mu^-(x) > a$ if and only if $F_\mu(a) < x$.*

(ii) *$(F_\mu^-)_\sharp \mathcal{L}_{|[0,1]} = \mu$.*

(iii) *Let $\eta = (F_\mu^-, F_\nu^-)_\sharp \mathcal{L}_{|[0,1]}$. Then $\eta \in \Pi(\mu, \nu)$ and*

$$\eta((-\infty, a] \times (-\infty, b]) = \min(F_\mu(a), F_\nu(b)).$$

(iv) *If F_μ is continuous, then $(F_\mu)_\sharp \mu = \mathcal{L}_{|[0,1]}$. In particular $(F_\nu^- \circ F_\mu)_\sharp \mu = (F_\nu^-)_\sharp (F_\mu)_\sharp \mu = \nu$.*

Proof. (i) $F_\mu^-(x) \leqslant a$ is equivalent to say that $a \geqslant \inf\{t, F_\mu(t) \geqslant x\}$ which is equivalent to $F_\mu(a) \geqslant x$. Moreover, $F_\mu^-(x) > a$ is equivalent to say that $a < \inf\{t, F_\mu(t) \geqslant x\}$ which is equivalent to $F_\mu(a) < x$.

(ii) By definition

$$(F_\mu^-)_\sharp \mathcal{L}_{|[0,1]}(-\infty, a] = \left|\left\{x \in [0,1], F_\mu^-(x) \leqslant a\right\}\right| = |\{x \in [0,1], F_\mu(a) \geqslant x\}| = F_\mu(a).$$

Therefore F_μ is the cumulative distribution function associated to $(F_\mu^-)_\sharp \mathcal{L}_{|[0,1]}$, which therefore coincides with μ.

(iii) By definition and the previous properties

$$\eta((-\infty, a] \times (-\infty, b]) = \left|\left\{x \in [0,1], F_\mu^-(x) \leqslant a, F_\nu^-(x) \leqslant b\right\}\right| = |\{x \in [0,1], F_\mu(a) \geqslant x, F_\nu(b) \geqslant x\}|.$$

(iv) Note that since F_μ is continuous, for $0 < a < 1$, $\{x \in F_\mu(x) \leqslant a\} = (-\infty, x_a]$ where $F_\mu(x_a) = a$. This implies that

$$(F_\mu)_\sharp \mu[0, a] = \mu\{x \in F_\mu(x) \leqslant a\} = \mu(-\infty, x_a] = F_\mu(x_a) = a = \mathcal{L}_{|[0,1]}[0, a].$$

□

2.3.3 Monge and Kantorovich optimal transport problem

We will define for $p \in [1, +\infty)$ the subspace:

$$\begin{aligned} \mathcal{P}_p(\mathbb{R}) &= \{\mu \mid \exists(\Omega, \mathbb{P}, \mathcal{F}) \ , X : \Omega \to \mathbb{R} \text{ random variable s.t. } X \in M^p, \mu = \mathbb{L}_X\} \\ &= \{\mu \in \mathcal{P}(\mathbb{R}), \int_{\mathbb{R}} |x|^p d\mu < +\infty\}. \end{aligned}$$

It is quite easy, arguing as for M^p spaces (and using Jensen inequality), to show that for $1 \leqslant p \leqslant q$ it holds

$$\mathcal{P}_q(\mathbb{R}) \subseteq \mathcal{P}_p(\mathbb{R}) \subseteq \mathcal{P}_1(\mathbb{R}) \subseteq \mathcal{P}(\mathbb{R}).$$

The optimal transport problem as stated by Monge in 1781 (as a problem of optimal transportation and optimal allocation of resources) can be rephrased in modern language as follows. We are given two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R})$ and a convex cost $c(x, y)$, that from now on we fix to be $|x-y|^p$ for some $p \geqslant 1$, measuring the cost of transporting one unit of mass from x to y . The optimal transport problem is how to transport μ to ν (so finding a transport map ψ such that $\psi_\sharp \mu = \nu$) whilst minimizing the cost:

$$\inf\left\{\int_{\mathbb{R}} |x - \psi(x)|^p d\mu \qquad \psi : \mathbb{R} \to \mathbb{R}, \text{ measurable, and such that } \psi_\sharp \mu = \nu\right\}. \tag{2.3.1}$$

The problem with this formulation is that in general we cannot assume that the set of transport maps is nonempty: so it is not sure that there exists at least one map ψ such that $\psi_\sharp \mu = \nu$. For example if $\mu = \delta_{x_0}$ and $\nu = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$, it is easy to see that no transport map may exist.

Such maps exists always in two basic cases.

If $\mu = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ and $\nu = \frac{1}{N}\sum_{i=1}^N \delta_{y_i}$, then we consider any map ψ such that $\psi(x_i) = y_j$ for some i and some j and we are done. In this case the Monge problem can be rewritten as follows: let $\sigma : \{1, 2, \dots, N\} \to \{1, 2, \dots, N\}$ a injective map (it is one of the possible permutations of the indexes), then

$$\min_\sigma \sum_i |x_i - y_{\sigma(i)}|^p.$$

Another case in which there exists a transport map is when F_μ is continuous, as we have seen in Lemma 2.3.4, in this case $\psi = F_\nu^- \circ F_\mu$.

Note that even if the transport maps exist, the optimal transport map may not be unique.

Example 2.3.5. [Book shift] Let us consider $\mu = \mathcal{L}_{[0,1]}$ and $\nu = \mathcal{L}_{[1,2]}$. A transport map is $\psi_1(t) = t + 1$, but also $\psi_2(t) = 2 - t$. Let us compute the cost associated to them for $p = 1$:

$$\int_{\mathbb{R}} |x - \psi_1(x)| d\mu = \int_0^1 |x - (1+x)| dx = 1 \qquad \int_{\mathbb{R}} |x - \psi_2(x)| d\mu = \int_0^1 |x - (2-x)| dx = \int_0^1 (2 - 2x) dx = 2 - 1 = 1.$$

Actually it is possible to prove that in this case both ψ_1, ψ_2 are optimal for the Monge problem with $p = 1$. Let us compute the cost associated to them for $p = 2$:

$$\int_{\mathbb{R}} |x - \psi_1(x)|^2 d\mu = \int_0^1 |x - (1+x)|^2 dx = 1$$

$$\int_{\mathbb{R}} |x - \psi_2(x)|^2 d\mu = \int_0^1 |x - (2-x)|^2 dx = \int_0^1 (2 - 2x)^2 dx = 4 - 1 + \frac{4}{3} > 1.$$

In this case ψ_1 is surely better than ψ_2. We will see in the following that in case $p > 1$, if $\mu << \mathcal{L}$, then the optimal transport map is unique and coincide with ψ_1 which is monotone.

Since the Monge problem has not always a solution, Kantorovich proposed a relaxation of it around 1940: instead of minimizing on deterministic couplings, we may minimize on all possible couplings between measures μ, ν:

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} |x - y|^p d\pi(x, y). \tag{2.3.2}$$

It is equivalent to restrict, when we consider the coupling $c(x, y) = |x - y|^p$ to $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$. Since a transport map induces a deterministic coupling, which in particular is a coupling it holds

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} |x-y|^p d\pi(x, y) \leqslant \inf \left\{ \int_{\mathbb{R}} |x - \psi(x)|^p d\mu \qquad \psi : \mathbb{R} \to \mathbb{R}, \text{ measurable, and such that } \psi_\sharp \mu = \nu \right\}.$$

Looking at the problem of minimizing the cost with coupling is an alternative way to describe the displacement of the particles of μ: instead of prescribing for each x the destination $\psi(x)$ of the particle located at x, for every x, y we specify how many particles go from x to y: that is $\pi(A \times B)$ is the amount of mass moving from A to B. Obviously this formulation allows for more general movements, since it may happen that a particle move to different destinations.

We end the section looking at the counterpart of the previous problem in the case of random variables. Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$, and consider X, Y random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{L}_X = \mu$, $\mathbb{L}_Y = \nu$. In particular $X, Y \in M_p$. Note that we are not prescribing the joint law of (X, Y). Therefore

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} |x - y|^p d\pi(x, y) = \inf_{\{\mathbb{L}_X = \mu, \mathbb{L}_Y = \nu\}} \mathbb{E}|X - Y|^p = \inf_{\{\mathbb{L}_X = \mu, \mathbb{L}_Y = \nu\}} \int |x - y|^p d\mathbb{L}_{(X,Y)}(x, y).$$

We consider the case $p = 2$.

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} |x - y|^2 d\pi(x, y) = \inf_{\{\mathbb{L}_X = \mu, \mathbb{L}_Y = \nu\}} \mathbb{E}|X - Y|^2 = \inf_{\{\mathbb{L}_X = \mu, \mathbb{L}_Y = \nu\}} \int |x - y|^2 d\mathbb{L}_{(X,Y)}(x, y).$$

Define $x_0 = \mathbb{E}(X) = \int_{\mathbb{R}} x d\mu$, and $y_0 = \mathbb{E}(Y) = \int_{\mathbb{R}} x d\nu$ we have that

$$\begin{aligned}
\mathbb{E}|X - Y|^2 &= \mathbb{E}|(X - x_0) - (Y - y_0) + (x_0 - y_0)|^2 \\
&= \mathbb{E}|X - x_0|^2 + \mathbb{E}|Y - y_0|^2 + |x_0 - y_0|^2 + \\
&\quad + 2\mathbb{E}((X - x_0)(x_0 - y_0)) + 2\mathbb{E}((Y - y_0)(x_0 - y_0)) - 2\mathbb{E}((X - x_0)(Y - y_0)) \\
&= \mathrm{Var}X + \mathrm{Var}Y + |x_0 - y_0|^2 - 2\mathrm{Cov}(X, Y).
\end{aligned}$$

Therefore the optimization problem

$$\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathbb{R}\times\mathbb{R}}|x-y|^2d\pi(x,y)=\inf_{\{\mathbb{L}_X=\mu,\mathbb{L}_Y=\nu\}}\mathbb{E}|X-Y|^2$$

can be restated as

$$\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathbb{R}\times\mathbb{R}}|x-y|^2d\pi(x,y)=C(\mu,\nu)-2\sup_{\{\mathbb{L}_X=\mu,\mathbb{L}_Y=\nu\}}\mathrm{Cov}(X,Y),$$

where $C(\mu,\nu)$ is a constant depending on μ,ν. More precisely: let $x_0=\int_{\mathbb{R}}xd\mu$, $y_0=\int_{\mathbb{R}}xd\nu$, it holds

$$C(\mu,\nu)=\int_{\mathbb{R}}(x-x_0)^2d\mu+\int_{\mathbb{R}}(x-y_0)^2d\nu+|x_0-y_0|^2=\int_{\mathbb{R}}x^2d\mu+\int_{\mathbb{R}}x^2d\nu-2\int_{\mathbb{R}}xd\mu\int_{\mathbb{R}}xd\nu.$$

The optimal coupling is obtained by finding the joint law between X,Y which maximizes the covariance, that is which guarantees maximal dependance between the two random variables with given laws (we will see that it will be obtained when Y is an increasing function of X).

Remark 2.3.6. Optimal transport problem has several economic interpretation where π is a matching between different actors of an economy and c is a sort of compatibility condition between agents x and y or the opposite of a utility function.

An optimal matching problem which is very famous is that of the stable marriage. Let us consider a population of women, with distribution μ and a population of men with distribution ν. A coupling γ is a coupling between women and men, so a set of marriages. We define $c_w(x,y)$ as the interest of woman x towards man y, and analogously $c_m(x,y)$, so that the utility function is c_w+c_m.

Finally the problem (2.3.2) can be restated as a linear optimization problem under convex constraints. We express the constraint $\pi\in\Pi(\mu,\nu)$ as follows:

$$\sup_{f,g\in C_b(\mathbb{R})}\left\{\int_{\mathbb{R}}f(x)d\mu+\int_{\mathbb{R}}g(x)d\nu-\int_{\mathbb{R}\times\mathbb{R}}(f(x)+g(y))d\pi\right\}=\begin{cases}0 & \pi\in\Pi(\mu,\nu)\\ +\infty & \text{elsewhere}\end{cases}.$$

Therefore we may rewrite (2.3.2) as:

$$\inf_{\pi\in\mathcal{P}(\mathbb{R}\times\mathbb{R})}\int_{\mathbb{R}\times\mathbb{R}}|x-y|^pd\pi(x,y)+\sup_{f,g\in C_b(\mathbb{R})}\left\{\int_{\mathbb{R}}f(x)d\mu+\int_{\mathbb{R}}g(x)d\nu-\int_{\mathbb{R}\times\mathbb{R}}(f(x)+g(y))d\pi\right\}$$

interchanging sup and inf

$$=\sup_{f,g\in C_b(\mathbb{R})}\left\{\int_{\mathbb{R}}f(x)d\mu+\int_{\mathbb{R}}g(x)d\nu+\inf_{\pi\in\mathcal{P}(\mathbb{R}\times\mathbb{R})}\int_{\mathbb{R}\times\mathbb{R}}(|x-y|^p-f(x)-g(y))d\pi(x,y)\right\}.$$

We rewrite

$$\inf_{\pi\in\mathcal{P}(\mathbb{R}\times\mathbb{R})}\int_{\mathbb{R}\times\mathbb{R}}(|x-y|^p-f(x)-g(y))d\pi(x,y)=\begin{cases}0 & f(x)+g(y)\leqslant|x-y|^p\ \forall x,y\in\mathbb{R}\\ -\infty & \text{elsewhere}\end{cases}.$$

Therefore we have that

$$\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathbb{R}\times\mathbb{R}}|x-y|^pd\pi(x,y)=\sup_{f,g\in C_b(\mathbb{R}),f(x)+g(y)\leqslant|x-y|^p}\int_{\mathbb{R}}f(x)d\mu+\int_{\mathbb{R}}g(x)d\nu.$$

If $\bar\pi$ is an optimal transport plan, then there exist $\bar f,\bar g$ optimal function (which are called Kantorovich potentials) such that $\bar f(x)+\bar g(y)=|x-y|^p$ for $(x,y)\in\mathrm{supp}\ \bar\pi$.

2.3.4 Wasserstein spaces

Let $p\geqslant 1$, $\mu,\nu\in\mathcal{P}_p(\mathbb{R})$: we define the p-Wasserstein distance between μ,ν as

$$W_p(\mu,\nu)=\left(\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathbb{R}\times\mathbb{R}}|x-y|^pd\pi(x,y)\right)^{\frac{1}{p}}.$$

The limit case $p = +\infty$ is given by

$$W_\infty(\mu,\nu) = \sup_{\pi\in\Pi(\mu,\nu)} \{|x-y|, (x,y) \in \text{supp}\pi\}$$

where the support of a measure defined in (X,Σ) is the largest (closed) subset of X for which every open ball centered at any point of the set has positive measure.

We collect in the following proposition some results (see [4, Chapter 1]), which are completely non trivial:

Proposition 2.3.7. *Let $\mu,\nu \in \mathcal{P}_p(\mathbb{R})$.*

(i) *There exists always at least one coupling $\pi \in \Pi(\mu,\nu)$ such that*

$$W_p(\mu,\nu)^p = \int_{\mathbb{R}\times\mathbb{R}} |x-y|^p d\pi(x,y).$$

We denote such couplings as $\Pi^o(\mu,\nu)$.

(ii) *If $p > 1$, there exists $K \subseteq \mathbb{R}$, closed bounded set, such that $\mu(K) = 1 = \nu(K)$ and $\mu << \mathcal{L}$, then the optimal coupling π between μ,ν is unique, and coincides with a deterministic coupling: that is there exists $\psi : \mathbb{R} \to \mathbb{R}$ measurable such that $\pi = (1,\psi)_\sharp\mu$, or equivalently*

$$W_p(\mu,\nu)^p = \int_{\mathbb{R}} |x-\psi(x)|^p d\mu.$$

(iii) *If $p = 2$ and $\mu << \mathcal{L}$ then the optimal coupling π between μ,ν is unique, and coincides with a deterministic monotone coupling, that is there exists $\psi : \mathbb{R} \to \mathbb{R}$ measurable such that $\psi = u'$ where $u : \mathbb{R} \to \mathbb{R}$ is a convex function and $\pi = (1,u')_\sharp\mu$, i.e.*

$$W_2(\mu,\nu)^2 = \int_{\mathbb{R}} |x-u'(x)|^2 d\mu.$$

In particular $\psi = F_\nu^- \circ F_\mu$.

(iv) *If $p = 1$ then*

$$W_1(\mu,\nu) = \sup_{\phi:\mathbb{R}\to\mathbb{R}, |\phi(x)-\phi(y)|\leqslant|x-y|} \left(\int_{\mathbb{R}} \phi(x) d\mu - \int_{\mathbb{R}} \phi(x) d\nu \right).$$

We have a description more accurate in dimension 1. The basic idea behind the proof of this theorem is the idea of monotonicity. If π transports mass from x_1 to y_1 and from $x_2 > x_1$ to y_2 we expect $y_2 > y_1$, else it would have been cheaper to transport from x_1 to y_2 and from x_2 to y_1.

Proposition 2.3.8. *Let $\mu,\nu \in \mathcal{P}_p(\mathbb{R})$, and F_μ, F_ν the associated cumulative functions. Let us define on $\mathbb{R}\times\mathbb{R}$ the measure π whose cumulative distribution function is $H(x,y) = \pi((-\infty,x]\times(-\infty,y]) = \min(F_\mu(x), F_\nu(y))$. Then $\pi \in \Pi^o(\mu,\nu)$ and*

$$W_p(\mu,\nu)^p = \int_{\mathbb{R}\times\mathbb{R}} |x-y|^p d\pi(x,y) = \int_0^1 |F_\mu^-(t) - F_\nu^-(t)|^p dt.$$

If F_μ is continuous, then π is a deterministic coupling, associated to $\psi = F_\nu^- \circ F_\mu$. In particular for $p = 1$ it holds

$$W_1(\mu,\nu)^p = \int_{\mathbb{R}\times\mathbb{R}} |x-y| d\pi(x,y) = \int_0^1 |F_\mu^-(t) - F_\nu^-(t)| dt = \int_{\mathbb{R}} |F_\mu(x) - F_\nu(x)| dx.$$

It is not easy, but possible to show that actually W_p is distance. We have a notion of convergence with respect to this metric. Moreover the space $\mathcal{P}_p(\mathbb{R})$ with this metric is a complete metric space.

Proposition 2.3.9. (i) *W_p is a distance: that is $W_p(\mu,\nu) = 0$ if and only if $\mu = \nu$, $W_p(\mu,\nu) = W_p(\nu,\mu)$, and finally the triangular inequality holds $W_p(\mu,\rho) \leqslant W_p(\mu,\nu) + W_p(\nu,\rho)$.*

(ii) *Let $\mu_n, \mu \in \mathcal{P}_p(\mathbb{R})$. Then $W_p(\mu_n,\mu) \to 0$ iff $\mu \to \mu$ weakly and $\lim_n \int_{\mathbb{R}} |x|^p d\mu_n = \int_{\mathbb{R}} |x|^p d\mu$.*

(iii) *$(\mathcal{P}_p(\mathbb{R}), W_p)$ is a complete metric space.*

(iv) *If $q > p$, then $W_q(\mu,\nu) \geqslant W_p(\mu,\nu)$.*

One of the most popular applications of optimal transport is the barycenter problem, providing dimension-free rates of statistical estimation. Wasserstein barycenters are a type of average of probability measures defined using the optimal transport geometry, and allow to average data that can be represented as probability distributions on $\mathbb{R}$ (or $\mathbb{R}^d$), a setting that commonly arises in machine learning and statistics.

Let us consider just empirical barycenters. So, given $\mu_1, \dots, \mu_k \in \mathcal{P}_2(\mathbb{R})$ we define the barycenter of this family as $\bar{\mu}$ realizing (if it exists) the minimum of

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R})} \frac{1}{k} \sum_{i=1}^{k} W_2^2(\mu_i, \mu).$$

2.4 Problems

(i) Let $(X, \|\cdot\|)$ a Banach space and $F : X \to X$ such that there exists $0 < a < 1$ for which

$$\|F(x) - F(y)\| \leqslant a\|x - y\| \qquad \forall x, y \in X.$$

(F is a contraction)

(a) Show that the map F is continuous.

(b) Let $x_0 \in X$. Define $x_1 = F(x_0)$, $x_2 = F(x_1)$ and so on $x_n = F(x_{n-1})$. Prove that

$$\|x_n - x_{n+1}\| \leqslant a^n \|x_0 - x_1\|.$$

Deduce that $(x_n)_n$ is a Cauchy sequence.

(c) Let $\bar{x} = \lim_n x_n$, where (x_n) has been defined in the previous step. Show that $F(\bar{x}) = \bar{x}$. So, $\bar{x}$ is a fixed point of F.

(d) Show that the map F admits a unique fixed point, that is a point such that $\bar{x} = F(\bar{x})$.

This is called **Banach-Caccioppoli theorem**.

(ii) Let $X_n, Y_n \in \mathcal{H}$ such that $X_n \to X$ and $Y_n \to Y$. Show that

- $\mathbb{E}(X_n) \to \mathbb{E}(X)$,
- $(X_n, Y_n) = \mathbb{E}(X_n Y_n) \to \mathbb{E}(XY) = (X, Y)$,
- $Cov(X_n, Y_n) = \mathbb{E}(X_n Y_n) - \mathbb{E}(X_n)\mathbb{E}(Y_n) \to Cov(XY) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
- $Var(X_n) = Cov(X_n, X_n) \to Var(X) = Cov(X, X)$.

(iii) Consider $X, Y, Z \in \mathcal{H}$ and assume X, Z are not constant. Compute the least linear quadratic estimator $L(Y|X, Z)$. Show that $L(Y|X, Z) = L(Y|X) + L(Y|Z - L(Z|X)) - \mathbb{E}(Y)$. (Hint: look at Remark 2.2.15).

Chapter 3

Element of Fourier analysis and the Central Limit Theorem

Fourier Analysis has several important applications in mathematics and statistics, in particular in data analysis and estimation. Loosely speaking, Fourier analysis refers to the tool used to compress complex data into exponential functions (or trigonometric functions). So, it permits to analyze data in terms of their frequency components. Two of the central ingredients of Fourier Analysis are the convolution operator and the Fourier transform.

In this last section we will consider also functions taking complex values, that is $f : \mathbb{R} \to \mathcal{C}$. In this case f can be written in terms of 2 real functions f_1, f_2 which correspond to the real and imaginary part of f, that is $f(x) = f_1(x) + if_2(x)$.

We recall also the formula for the complex exponential

$$e^{ix} = \cos x + i \sin x.$$

3.1 Convolution operator

Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be measurable functions and we define the convolution between f and g as the function

$$f * g(x) := \int_{\mathbb{R}} f(x-y)g(y)dy \qquad (\text{ or equivalently } = \int_{\mathbb{R}} f(y)g(x-y)dy)$$

for all x such that the integral exists finite. Note that $f * g$ is a function of x!
Intuitively: let $x \in \mathbb{R}^n$ and consider the function $y \to f(x-y)$. This is the same as the function f, but we have to shift the graph of f by x and then flip it around the axis $y = x$. Assume that f is a smooth function which is positive only in a neighborhood of 0 and null elsewhere, with integral 1. Computing $f * g(x)$ we are taking a sort of weighted average of the values of g near the point x (weighted by the values of g)..

Basic properties of the convolution are the following. For the proof we refer to the Section 8.2 in [3].

(i) $f * g(x) = g * f(x)$ and $(f * g) * h(x) = f * (g * h)(x)$,

(ii) The support of a function h is the closure of the set of points where $h \neq 0$. The support of $f * g$ is contained in the closure of the sum of the support of f and the support of g.

One of the main important features of the convolution operator is that it has *regularizing properties.*

Proposition 3.1.1. *If $f \in L^1(\mathbb{R}^n)$ and $g \in C^k(\mathbb{R}^n)$ bounded and with bounded derivatives up to order k, then $f * g \in C^k(\mathbb{R}^n)$ and for every $i \in \{1, \dots, n\}$ and $h \in \{1, \dots, k\}$, $\partial^h_{x_i}(f * g)(x) = f * (\partial^h_{x_i} g)(x)$.*

Let

$$g(x) = \begin{cases} ce^{\frac{1}{|x|^2-1}} & |x| \leqslant 1 \\ 0 & \text{elsewhere} \end{cases}$$

where $c > 0$ is chosen such that $\int_{\mathbb{R}} g(x)dx = 1$. Note that $g \in C^\infty(\mathbb{R})$ and $g(x) = 0$ for $|x| \geqslant 1$.

Let $t > 0$ and consider $g_t(x) = tg\left(\frac{x}{t}\right)$. Then $\int_{\mathbb{R}} g_t(x)dx = 1$ (by change of variable formula!) and $g_t(x) = 0$ if $|x| \geqslant t$.

As $t \to 0$ g_t becomes more and more concentrated at $x = 0$. Observe that by its properties, g_t **is the density function of the law of an absolutely continuous random variable** X_t**.**

Proposition 3.1.2 (Approximation of the Dirac measure and regularization by convolution)**.** *Let X_t be the continuous random variable with density given by g_t as defined before. Then X_t converges in distribution as $t \to 0^+$ to the* **discrete** *random variable X_0 with associated distribution the Dirac measure δ_0 (that is $X \equiv 0$ almost surely).*

Proof. To prove the convergence in distribution we need to show that for every f which is continuous and bounded there holds

$$\lim_{t\to 0^+} \int_{\mathbb{R}} f(x)g_t(x)dx = \delta_0(f) = f(0).$$

By definition and changing the variable posing $y = \frac{x}{t}$

$$\int_{\mathbb{R}} f(x)g_t(x)dx = \int_{-t}^{t} f(x)g_t(x)dx = c\int_{-1}^{1} f(ty)e^{-\frac{1}{|y|^2-1}}dy.$$

Sending $t \to 0$ and applying the dominated convergence theorem we conclude.

□

The convolution is also useful to compute density functions of the sum of independent random variables.

Theorem 3.1.3. *Let X and Y be independent absolutely continuous random variables and let f, g the density functions of the laws of X, Y. So $Z = X + Y$ is a continuous random variable with density function given by $f * g$.*

Remark 3.1.4. The same statement holds also with discrete random variables, substituting the integral with sum and convolution with a discrete convolution. That is if X, Y are discrete independent random variables, then $X + Y = Z$ is discrete random variable and the following holds: for every $n \in \mathbb{Z}$,

$$\mathbb{P}(Z = n) = \sum_{-\infty}^{+\infty} \mathbb{P}(X = k)\mathbb{P}(Y = n - k).$$

The proof of this formula can be checked easily in the case of random variables taking a finite number of values.

Proof. Observe that for every a, b, by independence

$$\mathbb{P}(X \leqslant a, Y \leqslant b) = \mathbb{P}(X \leqslant a)\mathbb{P}(Y \leqslant b) = \int_{-\infty}^{a} f(x)dy \int_{-\infty}^{b} g(y)dy.$$

So in particular we get

$$\mathbb{P}(X + Y \leqslant t) = \mathbb{P}(X \leqslant x, Y \leqslant y, x + y \leqslant t) = \int_{(x,y)\in\mathbb{R}^2, x+y\leqslant t} f(x)g(y)dxdy$$

where the integral is an integral computed in $\mathbb{R}^2$. We change variables to (z, w) where $x = z$ and $w = x + y$ (so $y = w - z$). So we get that $z \in \mathbb{R}$ and $w \leqslant t$:

$$\mathbb{P}(X + Y \leqslant t) = \int_{(x,y)\in\mathbb{R}^2, x+y\leqslant t} f(x)g(y)dxdy = \int_{-\infty}^{t}\int_{\mathbb{R}} f(z)g(w - z)dwdz = \int_{-\infty}^{t} f * g(z)dz$$

where in the last equality we use the definition of convolution. □

3.2 Fourier transform

The Fourier transform is an isometry among Hilbert spaces as we will see (so a bijection which maintains the distance) and in some sense it can be interpreted as a generalization of the Fourier serie in non periodic context.

Let $f \in L^1(\mathbb{R})$. We define the Fourier transform of f as the complex valued function

$$\hat{f}(x) = \int_{\mathbb{R}} f(y) e^{ixy} dy.$$

It can be generalized to several dimension: if $f \in L^1(\mathbb{R}^n)$ then

$$\hat{f}(x) = \int_{\mathbb{R}^n} f(y) e^{ix\cdot y} dy.$$

Observe that since $|e^{ixy}| = 1$ for all $x, y \in \mathbb{R}$, $|\hat{f}(x)| \leqslant \int_{\mathbb{R}} |f(y)| e^{ixy} dy \leqslant \int_{\mathbb{R}} |f(y)| dy = \|f\|_{L^1}$. More precisely we get the following result (see for the proof [3], Section 8.3), stating that the Fourier transform sends integrable functions in bounded continuous functions.

Proposition 3.2.1 (Riemann Lebesgue lemma)**.** *Let $f \in L^1(\mathbb{R})$. Then $\hat{f} \in C(\mathbb{R})$ and moreover $\lim_{|x|\to+\infty} \hat{f}(x) = 0$, $\|\hat{f}\|_\infty \leqslant \|f\|_{L^1}$.*

Other important properties of the Fourier transform are stated in the following proposition.

Proposition 3.2.2. *Let $f, g \in L^1(\mathbb{R})$. Then*

(i) *$\widehat{(f * g)} = \hat{f}\hat{g}$. So the Fourier transform of a convolution is the product of the Fourier transform.*

(ii) *If $|x|^k f \in L^1(\mathbb{R})$, then $\hat{f} \in C^k(\mathbb{R})$ and $d_x^k \hat{f}(x) = [\widehat{(iy)^k f}]$.*

(iii) *If $f \in C^k(\mathbb{R})$, $d_x^k f(x) \in L^1$, $\lim_{|x|\to+\infty} d_x^n f(x) = 0$ for $n \leqslant k$, then $\widehat{(d_x^n f)}(x) = (-ix)^n \hat{f}(x)$ for all $n \leqslant k$.*

Proof. (i) By definition, properties of the exponential and changing at the end variables (from (y,t) to (s,t) where $s = y - t$)

$$\begin{aligned}
\widehat{(f*g)}(x) &= \int_{\mathbb{R}} f * g(y) e^{ixy} dy = \int_{\mathbb{R}}\int_{\mathbb{R}} f(t) g(y-t) e^{ixy} dt dy \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} f(t) g(y-t) e^{ix(y-t)} e^{ixt} dt dy \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} f(t) g(s) e^{ixs} e^{ixt} dt ds = \hat{f}(x)\hat{g}(x).
\end{aligned}$$

(ii) We get that

$$d_x \hat{f}(x) = d_x \int_{\mathbb{R}} f(y) e^{ixy} dy = \int_{\mathbb{R}} d_x f(y) e^{ixy} dy = \int_{\mathbb{R}} f(y)(iy) e^{ixy} dy = \widehat{(iyf)}(x).$$

Repeat the argument we conclude with the result for every $k \in \mathbb{N}$.

(iii) We integrate by parts and we have that

$$\widehat{d_y f}(x) = \int_{\mathbb{R}} d_y f(y) e^{ixy} dy = \Big[f(y) e^{ixy} \Big]_{-\infty}^{+\infty} - \int_{\mathbb{R}} f(y)(ix) e^{ixy} dy = -ix\hat{f}(x).$$

Iterating the procedure we conclude.

□

The previous proposition has a very important consequence:

$$\text{let } a > 0 \text{ and } f_a(x) = e^{-a|x|^2}, \text{ then } \widehat{f}_a(x) = \sqrt{\frac{\pi}{a}} e^{-\frac{|x|^2}{4a}}. \tag{3.2.1}$$

More generally in $\mathbb{R}^n$, if $f_a(x) = e^{-a|x|^2}$, for $x \in \mathbb{R}^n$, then $\widehat{f}_a(x) = \sqrt{\frac{\pi^n}{a^n}} e^{-\frac{|x|^2}{4a}}$.

We prove (3.2.1). Observe that by the previous proposition, items (ii) and (iii) we get that

$$d_x\widehat{f_a}(x) = \int_{\mathbb{R}} e^{-a|y|^2}(iy)e^{ixy}dy = \int_{\mathbb{R}} \frac{-i}{2a}d_y(e^{-a|y|^2})e^{ixy}dy = -\frac{i}{2a}\widehat{d_y f_a}(x) = -\frac{x}{2a}\hat{f}_a(x).$$

So the function $\hat{f}_a = \phi$ satisfies $\phi'(x) = -\frac{x}{2a}\phi(x)$, integrating we get that $(\log\phi(x))' = -\frac{x^2}{4a} + c$ and then $\phi(x) = ke^{-\frac{1}{4a}x^2}$. Finally to compute k we need to compute $\phi(0) = \hat{f}_a(0)$.

$$\hat{f}_a(0) = \int_{\mathbb{R}} e^{-a|y|^2}e^0 dy = \sqrt{\frac{\pi}{a}}.$$

Proposition 3.2.3. *Let $f, g \in L^1(\mathbb{R})$, then*

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

Proof. By definition and by changing the order of integration (thanks to Fubini Tonelli theorem)

$$\int_{\mathbb{R}} \hat{f}(x)g(x)dx = \int_{\mathbb{R}}\int_{\mathbb{R}} f(y)g(x)e^{ixy}dydx = \int_{\mathbb{R}} f(x)\hat{g}(x)dx.$$

□

For $f \in L^1(\mathbb{R})$ we may define also the anti transform of f as follows:

$$\check{f}(x) = \frac{1}{2\pi}\int_{\mathbb{R}} f(y)e^{-ixy}dy = \frac{1}{2\pi}\hat{f}(-x).$$

Obviously, this operator satisfies the same properties as the Fourier transform.

Theorem 3.2.4 (Fourier inversion theorem)**.** *Let $f \in L^1(\mathbb{R})$ such that also $\hat{f} \in L^1(\mathbb{R})$. Then f is continuous and bounded (that is, it coincides almost everywhere with a continuous function) and $\check{\hat{f}} = f = \hat{\check{f}}$.*

In particular if $f, g \in L^1(\mathbb{R})$ with $\hat{f} = \hat{g}$, then $f = g$ almost everywhere.

Proof. We give a sketch of the proof, for the rigorous demonstration we refer to [3], Theorem 8.26.

We have that

$$\int_{\mathbb{R}} \hat{f}(y)e^{-ixy}dy = \lim_{\varepsilon\to 0}\int_{\mathbb{R}} \hat{f}(y)e^{-ixy}e^{-\varepsilon y^2}dy = \lim_{\varepsilon\to 0}\int_{\mathbb{R}}\int_{\mathbb{R}} f(z)e^{iyz}dze^{-ixy}e^{-\varepsilon y^2}dy =$$

by changing the order of integration

$$= \lim_{\varepsilon\to 0}\int_{\mathbb{R}} f(z)\int_{\mathbb{R}} e^{iy(z-x)}e^{-\varepsilon y^2}dydz.$$

Now we observe that

$$\int_{\mathbb{R}} e^{iy(z-x)}e^{-\varepsilon y^2}dy = \widehat{e^{-\varepsilon y^2}}(z-x)$$

and then by (3.2.1) we conclude

$$\int_{\mathbb{R}} e^{iy(z-x)}e^{-\varepsilon y^2}dy = \widehat{e^{-\varepsilon y^2}}(z-x) = \frac{\sqrt{\pi}}{\sqrt{\varepsilon}}e^{-(x-z)^2/4\varepsilon}.$$

We substitute in the previous integral and we get

$$2\pi\check{\hat{f}}(x) = \int_{\mathbb{R}} \hat{f}(y)e^{-ixy}dy = \lim_{\varepsilon\to 0}\int_{\mathbb{R}} f(z)\frac{\sqrt{\pi}}{\sqrt{\varepsilon}}e^{-(x-z)^2/4\varepsilon}dz = 2\sqrt{\pi}\lim_{\varepsilon\to 0}\int_{\mathbb{R}} f\left(x - \frac{\xi}{2\sqrt{\varepsilon}}\right)e^{-\xi^2}d\xi$$

$$= 2\sqrt{\pi}f(x)\int_{\mathbb{R}} e^{-\xi^2}d\xi = 2\pi f(x).$$

The last conclusion comes from the fact that $\widehat{(f-g)} = \hat{f} - \hat{g} = 0$. Therefore $f - g \in L^1(\mathbb{R})$ is such that $\widehat{(f-g)} = 0 \in L^1(\mathbb{R})$, which implies by the inversion theorem that $f - g = \check{\widehat{(f-g)}} = 0$. □

Using the inversion theorem, we get also the following result:

Corollary 3.2.5. *Let*

$$\mathcal{S} = \{g : \mathbb{R} \to \mathbb{R},\ g \in C^\infty,\ \forall k\ \sup_x ||x|^k g(x)| \leqslant C_k,\ |x|^k g(x) \in L^1(\mathbb{R})\}.$$

Then the Fourier transform is a bijection of $\mathcal{S}$ into itself.

Note that for all $a > 0$, $e^{-ax^2} \in \mathcal{S}$.

Proof. By Proposition 3.2.2, we get that if $g \in \mathcal{S}$ then $\hat{g} \in C^\infty$ and moreover $x^k \hat{g}$ is bounded continuous and integrable, so in particular $\hat{g} \in \mathcal{S}$. The conclusion comes from the inversion theorem. □

Lemma 3.2.6. *The set $\mathcal{S}$ is dense in the space $C_0(\mathbb{R}) = \{g \in \mathcal{C}(\mathbb{R}) \mid \lim_{|x|\to+\infty} g(x) = 0\}$ (with respect to $\|\cdot\|_\infty$ norm).*

For this lemma we refer to [3, Proposition 8.17].

Theorem 3.2.7. *Let $f_n, f \in L^1(\mathbb{R})$. Assume that $\hat{f}_n \to \hat{f}$ pointwise and that there exists $C > 0$ such that $\|f_n\|_{L^1} \leqslant C$ for all n. Then $f_n \rightharpoonup f$ vaguely in $L^1(\mathbb{R})$, that is for all $g \in C_0(\mathbb{R})$, there holds $\lim_n \int_{\mathbb{R}} f_n(x)g(x)dx = \int_{\mathbb{R}} f(x)g(x)dx$.*

Proof. Let $g \in C_0(\mathbb{R})$. Then by Lemma 3.2.6 there exists $g_k \in \mathcal{S}$ such that $\sup_{x\in\mathbb{R}} |g_k(x) - g(x)| \leqslant \frac{1}{k}$.

Since $g_k \in \mathcal{S}$ then $g_k = \check{\hat{g}}_k$ by Corollary 3.2.5. Therefore we get

$$\int_{\mathbb{R}} (f_n - f)(x)g_k(x)dx = \int_{\mathbb{R}} (f_n - f)(x)\hat{\check{g}}_k(x)dx = \int_{\mathbb{R}}\int_{\mathbb{R}} (f_n - f)(x)\check{g}_k(y)e^{ixy}dydx$$

exchanging the order of integration

$$= \int_{\mathbb{R}}\int_{\mathbb{R}} (f_n - f)(x)\check{g}_k(y)e^{ixy}dxdy = \int_{\mathbb{R}} (\hat{f}_n - \hat{f})(y)\check{g}_k(y)dy.$$

Since $\sup_{y\in\mathbb{R}} |\hat{f}_n(y) - \hat{f}(y)| \leqslant \|f_n - f\|_{L^1} \leqslant \|f_n\|_{L^1} + \|f\|_{L^1} \leqslant C + \|f\|_{L^1}$ and $g_k \in L^1$, we get that $|(\hat{f}_n - \hat{f})(y)\check{g}_k(y)| \leqslant C + \|f\|_{L^1}|g_k| \in L^1$. Moreover $\hat{f}_n(y) - \hat{f}(y) \to 0$ as $n \to +\infty$ by assumption, then by the Lebesgue dominated convergence we conclude that

$$\lim_n \int_{\mathbb{R}} (f_n - f)(x)g_k(x)dx = 0$$

for all $k > 0$. Using the fact that $\sup_{x\in\mathbb{R}} |g_k(x) - g(x)| \leqslant \frac{1}{k}$ we get

$$\left|\int_{\mathbb{R}} (f_n - f)(x)g(x)dx\right| \leqslant \left|\int_{\mathbb{R}} (f_n - f)(x)(g_k - g)(x)dx\right| + \left|\int_{\mathbb{R}} (f_n - f)(x)g_k(x)dx\right|$$

$$\leqslant \int_{\mathbb{R}} |f_n(x) - f(x)||g_k(x) - g(x)|dx + \left|\int_{\mathbb{R}} (f_n - f)(x)g_k(x)dx\right|$$

$$\leqslant \frac{1}{k}\|f_n - f\|_{L^1} + \left|\int_{\mathbb{R}} (f_n - f)(x)g_k(x)dx\right| \leqslant \frac{1}{k}(C + \|f\|_{L^1}) + \left|\int_{\mathbb{R}} (f_n - f)(x)g_k(x)dx\right|.$$

Therefore we conclude that for all $k \in \mathbb{N}$,

$$\lim_n \left|\int_{\mathbb{R}} (f_n - f)(x)g(x)dx\right| \leqslant \frac{1}{k}(C + \|f\|_{L^1})$$

which gives the conclusion sending $k \to +\infty$. □

3.3 Characteristic functions of random variables

Let X be a random variable, with associated $\mathbb{P}_X$ probability distribution. The characteristic function of X is defined as the (complex valued) function

$$\phi_X(t) = \mathbb{E}(e^{itX}).$$

More precisely

– if X is a (asbsolutely) continuous random variable (with density f) then

$$\phi_X(t) = \int_{\mathbb{R}} e^{itx} f(x)dx = \widehat{f}(t).$$

So in this case the characteristic function of X is the Fourier transform of the density function f associated to X.

– if X is a discrete random variable (taking values on $\mathbb{Z}$),

$$\phi_X(t) = \sum_{k\in\mathbb{Z}} e^{ikt} P(\omega \mid X(\omega) = k).$$

Note that ϕ_X is a continuous function such that $\phi(0) = 1$.

Proposition 3.3.1. *If X_1, X_2 are independent random variables, then the characteristic function of $X_1 + X_2$ satisfies*

$$\phi_{X_1+X_2}(t) = \phi_{X_1}(t)\phi_{X_2}(t).$$

Proof. We consider only the case in which X_1, X_2 are absolutely continuous random variables (for the other case the argument is similar). The probability density of the sum of X_1 and X_2 is given by the convolution between the density of X_1 and the density of X_2 by Theorem 3.1.3. Then the Fourier transform of a convolution is the product of the Fourier transforms, see Proposition 3.2.2. □

The characteristic function associated to a random variable characterizes completely the random variable, and moreover the functional from the spaces of random variables with the convergence in distribution to the space of characteristic functions with the pointwise convergence is continuous, in the sense that if a sequence of random variables is converging in distribution to a random variable, then the same holds for the characteristic functions (and viceversa).

Theorem 3.3.2. *Let X_n be a family of random variables.*

(i) *If X_n are converging in distribution to X, then $\phi_{X_n}(t) \to \phi_X(t)$ for every t.*

(ii) *If $\phi_{X_n}(t) \to \phi(t)$ for every t, where ϕ is a function continuous at $t = 0$, then ϕ is the characteristic function of a random variable X and X_n converge in distribution to X.*

Proof. (i) $X_n \to X$ in distribution for every bounded continuous function g it holds

$$\mathbb{E}(g(X_n)) \to \mathbb{E}(g(X)).$$

So, taking for every t, $g_t(y) = e^{ity}$ (which is bounded and continuous), we get $\phi_{X_n}(t) \to \phi_X(t)$.

(ii) We prove this part theorem only in the case of absolutely continuous random variables X_n, with associated densities f_n. The general case can be obtained similarly.

We claim that X_n are tight. If the claim is true, then by Theorem 1.9.3, up to a subsequence we get that X_{n_k} converge in distribution to a random variable X. By (i), we get that $\phi_{X_{n_k}}(t) \to \phi_X(t)$ for every t and so $\phi(t) = \phi_X(t)$. Since the limit is unique (does not depend on subsequences), we conclude the convergence of the whole sequence of X_n.

So to conclude it is sufficient to show that X_n are tight. Since we are assuming X_n to have a density f_n, we get that $\phi_{X_n}(t) = \hat{f}_n(t)$. Fix $\delta > 0$ and consider

$$\frac{1}{2\delta}\int_{-\delta}^{\delta}(1-\phi_{X_n}(t))dt = \frac{1}{2\delta}\int_{-\delta}^{\delta}(1-\hat{f}_n(t))dt = \frac{1}{2\delta}\int_{-\delta}^{\delta}\int_{\mathbb{R}}(1-e^{iyt})f_n(y)dydt$$

$$= \frac{1}{2\delta}\int_{\mathbb{R}}\int_{-\delta}^{\delta}(1-e^{iyt})dtf_n(y)dy = \frac{1}{2\delta}\int_{\mathbb{R}}\left[2\delta - \frac{2\sin\delta y}{y}\right]f_n(y)dy = \int_{\mathbb{R}}\left[1-\frac{\sin\delta y}{\delta y}\right]f_n(y)dy$$

$$\geqslant \frac{1}{2}\int_{|\delta y|\geqslant 2} f_n(y)dy = \frac{1}{2}\mathbb{P}\left(|X_n| \geqslant \frac{2}{\delta}\right).$$

Hence

$$\mathbb{P}\left(|X_n| \geqslant \frac{2}{\delta}\right) \leqslant \frac{1}{\delta}\int_{-\delta}^{\delta}(1-\phi_{X_n}(t))dt \to \frac{1}{\delta}\int_{-\delta}^{\delta}(1-\phi(t))dt.$$

Since ϕ is continuous and $\phi(0)=1$, we get that for every $\varepsilon>0$ there exists δ such that $(1-\phi(t)) \leqslant \varepsilon/4$ for $t\in[-\delta,\delta]$. So

$$\frac{1}{\delta}\int_{-\delta}^{\delta}(1-\phi(t))dt \leqslant \frac{\varepsilon}{2}.$$

We fix $\varepsilon>0$, we choose δ as above, and $K_\varepsilon = \{|x| \leqslant \frac{2}{\delta}\}$ and then we choose $\bar{n}$ such that $\frac{1}{\delta}\int_{-\delta}^{\delta}(1-\phi_{X_n}(t))dt \leqslant \varepsilon$ for all $n \geqslant \bar{n}$. This gives the desired tightness: $\mathbb{P}(|X_n| \in K_\varepsilon) \geqslant 1-\varepsilon$ for all $n \geqslant \bar{n}$. □

3.4 The Central Limit Theorem

We conclude showing that actually the Central Limit theorem can be interpreted as a result in Fourier analysis. The theorem says that if we have a sufficiently large sample of observations- randomly produced in a way that does not depend on the values of the other observations- the probability distribution of the observed averages will closely approximate a normal distribution.

Theorem 3.4.1 (Central Limit theorem)**.** *Let X_n be independent identically distributed random variables with (common) mean μ and a variance σ. Then $\frac{X_1+\cdots+X_n-\mu}{\sqrt{n}\sigma}$ converges in distribution to the normal random variable with mean 0 and variance 1.*

We are not going to prove in full generality this theorem, but we are just giving an idea of what is going on in the case in which every X_i is an absolutely continuous random variable with density f. Up to a renormalization we may assume that the mean of X_i is 0 and the variance is 1.

Proposition 3.4.2. *Let $f:\mathbb{R}\to[0,+\infty)$ such that*

$$\int_{\mathbb{R}} f(x)dx = 1, \qquad \int_{\mathbb{R}} xf(x)dx = 0 \qquad \int_{\mathbb{R}} x^2 f(x)dx = 1.$$

Let $f^{*n} := f * \cdots * f$ (the convolution of f by itself n times).

Then $f_n(x) := \sqrt{n}f^{*n}(\sqrt{n}x)$ converges vaguely as $n\to+\infty$ to $\frac{e^{-x^2/2}}{\sqrt{2\pi}}$.

Proof. The first assumption on f implies that $\widehat{f}(0)=1$. Moreover, recalling Proposition 3.2.2, item ii, we get that the second and third assumption on f imply that $f\in C^2$. Moreover

$$\widehat{d_x f}(0) = \int_{\mathbb{R}}(iy)f(y)dy = 0 \qquad \widehat{d_x^2 f}(0) = \int_{\mathbb{R}}(-iy)^2 f(y)dy = -1.$$

By Taylor theorem we conclude that for $x\to 0$,

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

We compute now $\widehat{f_n}(x)$. We have that

$$\widehat{f_n}(x) = \int_{\mathbb{R}} f_n(y)e^{ixy}dy = \int_{\mathbb{R}} \sqrt{n}f^{*n}(\sqrt{n}y)e^{ixy}dy =$$

changing variable $z=\sqrt{n}y$

$$= \int_{\mathbb{R}} f^{*n}(z)e^{ix\frac{z}{\sqrt{n}}}dz = \int_{\mathbb{R}} f^{*n}(z)e^{i\frac{x}{\sqrt{n}}z}dz = \widehat{f^{*n}}\left(\frac{x}{\sqrt{n}}\right)$$

and recalling by Proposition 3.2.2, item i, that $\widehat{f^{*n}}(x) = (\widehat{f}(x))^n$ we conclude that

$$\widehat{f_n}(x) = \widehat{f^{*n}}\left(\frac{x}{\sqrt{n}}\right) = \left(\widehat{f}\left(\frac{x}{\sqrt{n}}\right)\right)^n.$$

So, we get for x fixed and $n \to +\infty$

$$\widehat{f_n}(x) = \left(\widehat{f}\left(\frac{x}{\sqrt{n}}\right)\right)^n = \left(1 - \frac{x^2}{2n} + o\left(\frac{1}{n}\right)\right)^n = e^{n\log\left(1-\frac{x^2}{2n}+o\left(\frac{1}{n}\right)\right)}.$$

Recalling that for x fixed and $n \to +\infty$, we get $\log\left(1 - \frac{x^2}{2n} + o\left(\frac{1}{n}\right)\right) = -\frac{x^2}{2n} + o\left(\frac{1}{n}\right)$ we get

$$\widehat{f_n}(x) = e^{-\frac{x^2}{2}+o(1)}$$

and therefore $\lim_n \widehat{f}_n(x) = e^{-\frac{x^2}{2}}$. By (3.2.1) with $a = \frac{1}{2}$ we have that $\widehat{\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}} = e^{-\frac{x^2}{2}}$. Therefore

$$\lim_n \widehat{f_n}(x) = \widehat{\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}}.$$

Moreover $\|f_n\|_1 = 1$ for all n. So, we may apply Theorem 3.2.7 to obtain that f_n is converging vaguely to $\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$. □

3.5 Problems

(i) Let $c > 0$, and

$$h_c(x) = \begin{cases} 1 & |x| \leqslant c \\ 0 & \text{elsewhere} \end{cases}.$$

Compute $h_c * h_c$. Then compute $h_c * h_c * h_c$. What we can say about the regularity of these functions?

(ii) Let $X_1, X_2, \ldots X_n$ are n independent continuous random variables with the same distribution (and so with the same density function f). Assume that $\mathbb{E}(X_i) = \mu$ and $\mathbb{E}(X_i - \mu)^2 = \sigma^2$. Show that the density function of $\frac{X_1+\cdots+X_n-\mu}{\sqrt{n}\sigma}$ is given by $\sqrt{n}\sigma f^{*n}(x\sqrt{n}\sigma + \mu n)$, where $f^{*n}(x)$ is the convolution of f with itself repeated n times.

(iii) (a) Compute the Fourier transform of $g(x) = e^{-x}\chi_{(0,+\infty)}(x)$.
Recall the following formulas (obtained by integration by parts):

$$\int e^{-y}\sin(xy)dy = -\frac{1}{x^2+1}e^{-y}(x\cos xy + \sin xy) + c$$

$$\int e^{-y}\cos(xy)dy = \frac{1}{x^2+1}e^{-y}(x\sin xy - \cos xy) + c.$$

(b) Compute the Fourier transform of $f(x) = xe^{-x}\chi_{(0,+\infty)}(x)$ (that is the characteristic function of the Gamma distribution).
Use item (a) and Proposition 3.2.2.

Bibliography

[1] V. I. Bogachev *Measure theory, vol I, II* Springer-Verlag, Berlin, 2007.

[2] A. Bressan *Lecture notes on Functional Analysis, with applications to Linear Partial Differential Equations* Graduate Studies in Mathematics, vol 143, AMS, 2013.

[3] G. Folland *Real Analysis: modern tecniques and their applications.* Wiley 1999 (2nd ed).

[4] F. Santambrogio *Optimal transport for applied mathematicians. Calculus of variations, PDEs, and modeling.* Birkhäuser/Springer, Cham, 2015.

Solutions to problems Chapter 1

(i) Let $f : \mathbb{R} \to \mathbb{R}$ be a monotone function. Show that f is Lebesgue measurable.

It is sufficient to show that for all $c \in \mathbb{R}$ the set $\{x \in \mathbb{R} \mid f(x) > c\}$ is measurable.

Assume that f is monotone increasing (if it is monotone decreasing the argument is analogous). Let $c \in \mathbb{R}$. If $f(x) \leqslant c$ for all $x \in \mathbb{R}$ then $\{x \in \mathbb{R} \mid f(x) > c\}$ is the empty set and we are done.

Assume now that there exists $\bar{x} \in \mathbb{R}$ such that $f(\bar{x}) > c$. By monotonicity we get that $f(y) > c$ for all $y > \bar{x}$. We consider now the set $A_c = \{x \in \mathbb{R} \,|f(x) > c\}$. Our aim is to show that this is a measurable set.

We observed that by monotonicity, if $x \in A_c$, then $[x, +\infty) \subseteq A_c$. So, if A_c is not bounded from below, this implies that $A_c = \mathbb{R}$ and so we are done. Assume now that A_c is bounded from below and define $x_c = \inf A_c$. For all $x > x_c$ we get that $f(x) > c$ and $f(x) \leqslant c$ for all $x < x_c$. This implies that $A_c = (x_c, +\infty)$ if $f(x_c) \leqslant c$, and $A_c = [x_c, +\infty)$ if $f(x_c) > c$. In both cases, $A_c \in \mathcal{M}$.

Note that actually we get something more: for all c, we get that A_c is a Borel set, so the function f is Borel measurable.

(ii) Consider the right continuous increasing function on $\mathbb{R}$

$$F(x) = \begin{cases} x & x < 0 \\ x + 1 & x \geqslant 0. \end{cases}$$

Which is the Borel measure associated to this function?

We define $\mu_F(a, b] = F(b) - F(a)$, and then we extend it to a measure on the Borel σ-algebra. Given F as in the statement, we get that $\mu_F(a, b] = b - a$ if $a < b < 0$, $\mu_F(a, b] = b + 1 - (a + 1) = b - a$ if $0 \leqslant a < b$, whereas if $a < 0 \geqslant b$, then $\mu_F(a, b] = b + 1 - a = b - a + 1$. Therefore $\mu_F = \mathcal{L} + \delta_0$.

Solutions to problems Chapter 2

(i) **Banach-Caccioppoli theorem**

(a) Let (x_n) be a sequence in X which is converging to x. Then $0 \leqslant \|F(x_n) - F(x)\| \leqslant a\|x_n - x\|$, and so $\lim_{n\to+\infty} F(x_n) = F(x)$ since $\lim_{n\to+\infty} x_n = x$.

(b) By the property of the function F and the definition of the we get that

$$\|x_{n+1} - x_n\| = \|F(x_n) - F(x_{n-1})\| \leqslant a\|x_n - x_{n-1}\| =$$

$$= a\|F(x_{n-1}) - F(x_{n-2})\| \leqslant a^2\|x_{n-1} - x_{n-2}\| \leqslant ... \leqslant a^n\|x_1 - x_0\|.$$

Let $n > m$. Then,by using the triangular inequality, we get

$$\|x_n - x_m\| \leqslant \|x_n - x_{n-1}\| + \|x_{n-1} - x_{n-2}\| + \cdots + \|x_{m+1} - x_m\|.$$

By using the previous inequality and recalling that $\sum_{i=0}^{n} a^i = \frac{1-a^{n+1}}{1-a}$, we get

$$\|x_n - x_m\| \leqslant (a^n + a^{n-1} + \cdots + a^m)\|x_0 - x_1\| \leqslant= \frac{a^{m+1} - a^{n+1}}{1-a}|x_0 - x_1\|.$$

Since $0 < a < 1$, we get that $a^{n+1}, a^{m+1} \to 0$ as $n, m \to +\infty$. So in particular the previous inequality implies that (x_n) is a Cauchy sequence.

(c) Since (x_n) is a Cauchy sequence, and the space is complete, it is converging to some point x. Using the continuity of F we have that $\lim_n F(x_n) = F(x)$. But we recall that $\lim_n F(x_n) = \lim_n x_{n-1} = x$. So $F(x) = x$.

(d) Let x, z such that $F(x) = x$ and $F(z) = z$. The by the property of F, and recalling that $a < 1$,

$$\|x - z\| = \|F(x) - F(z)\| \leqslant a\|x - z\| < \|x - z\|.$$

This is not possible unless $\|x - z\| = 0$, which implies $z = x$.

(ii)
- $\mathbb{E}(X_n) = (X_n, 1) \to (X, 1) = \mathbb{E}(X)$, by continuity of the scalar product (as a consequence of Cauchy Schwartz inequality).
- $(X_n, Y_n) = (X_n - X, Y_n - Y) + (X, Y_n) + (X_n, Y) - (X, Y)$. We conclude observing that $(X_n - X, Y_n - Y) \to 0$, $(X_n, Y) \to (X, Y)$ and $(X, Y_n) \to (X, Y)$.
- the convergence of covariance and variance are immediate consequences of the first two items.

(iii) Recalling Remark 2.2.15 we have that

$$L(Y|X, Z) = Pr_S(Y) = a + bX + cZ$$

where S is the space with basis $1, X, Z$.
Observe that by the same argument $L(Z|X) = Pr_T(Z)$ where T is the space with a basis given by $1, X$. In particular by Theorem 2.2.8 we have that $Z - L(Z|X) \in T^{\perp}$ and arguing as in Remark 2.2.15 $L(Z|X) = \mathbb{E}(Z) + \frac{Cov(X,Z)}{Var(X)}(X - \mathbb{E}(X))$.
An orthonormal basis of S can be therefore obtained by considering an orthonormal basis of T, which is given by $1, \frac{X-\mathbb{E}(X)}{\sqrt{Var(X)}}$ as proved in Remark 2.2.15 and then adding the element $k(Z - L(Z|X))$ where k is such that $\mathbb{E}(k(Z - L(Z|X))^2 = 1$. Since $\mathbb{E}((Z - L(Z|X))^2 = \frac{Var(Z)Var(X)-Cov^2(X,Z)}{Var(X)}$ as proved in Remark 2.2.15, we get that $k = \frac{\sqrt{VarX}}{\sqrt{Var(Z)Var(X)-Cov^2(X,Z)}}$.
So, as in Remark 2.2.15,

$$\begin{aligned} L(Y|X, Z) &= \mathbb{E}(Y) + \frac{Cov(X, Y)}{Var(X)}(X - \mathbb{E}(X))) \\ &\quad + \frac{Var(X)Cov(Z, Y) - Cov(X, Z)Cov(X, Y)}{Var(Z)Var(X) - Cov^2(X, Z)}(Z - L(Z|X)) \\ &= \mathbb{E}(Y) \\ &\quad + \frac{Var(Z)Cov(X, Y) - Cov(Z, Y)Cov(X, Z)}{Var(Z)Var(X) - Cov^2(X, Z)}(X - \mathbb{E}(X))) \\ &\quad + \frac{Var(X)Cov(Z, Y) - Cov(X, Z)Cov(X, Y)}{Var(Z)Var(X) - Cov^2(X, Z)}(Z - \mathbb{E}(Z)). \end{aligned}$$

Observe that

$$\mathbb{E}(Y) + \frac{Cov(X, Y)}{Var(X)}(X - \mathbb{E}(X))) = L(Y|X)$$

and moreover

$$\mathbb{E}(Y) + \frac{Var(X)Cov(Z, Y) - Cov(X, Z)Cov(X, Y)}{Var(Z)Var(X) - Cov(X, Z)}(Z - L(Z|X)) = L(Y|Z - L(Z|X)).$$

This conclude the proof.

Solutions to problems Chapter 3

(i) Let $c>0$, and $h_c(x)=\begin{cases}1 & |x|\leqslant c\\ 0 & \text{elsewhere}\end{cases}$. Compute $h_c*h_c(x)$. Then compute $h_c*h_c*h_c$. What we can say about the regularity of these functions?
By definition of h_c

$$h_c*h_c(x)=\int_{\mathbb{R}}h_c(x-y)h_c(y)dy=\int_{-c}^{c}h_c(x-y)dy=|[-c,c]\cap[x-c,x+c]|$$

where we indicated with $|[-c,c]\cap[x-c,x+c]|$ the length of the intersection between the two intervals. Since

$$[-c,c]\cap[x-c,x+c]=\begin{cases}\varnothing & x\geqslant 2c \text{ or } x\leqslant -2c\\ [-c,x+c] & -2c<x<0\\ [x-c,c] & 0<x<2c\end{cases}$$

we conclude that

$$h_c*h_c(x)=\begin{cases}0 & x\geqslant 2c \text{ or } x\leqslant -2c\\ x+2c & -2c<x<0\\ 2c-x & 0<x<2c.\end{cases}$$

The graph is a triangular.. Then again by definition

$$h_c*h_c*h_c(x)=\int_{\mathbb{R}}(h_c*h_c)(x-y)h_c(y)dy=\int_{-c}^{c}(h_c*h_c)(x-y)dy$$

$$=\int_{[-c,c]\cap[x-2c,x+2c]}(h_c*h_c)(x-y)dy.$$

We observe that $h_c*h_c*h_c(x)=h_c*h_c*h_c(-x)$ so it is sufficient to compute the function for x positive and then symmetrize it (as an even function). If $x>3c$ then $h_c*h_c*h_c(x)=0$. If $x\in(2c,3c)$ then $[-c,c]\cap[x-2c,x+2c]=[x-2c,c]$ with $x-2c>0$ and so

$$h_c*h_c*h_c(x)=\int_{x-2c}^{c}h_c*h_c(y)dy=\frac{(4c-x)^2}{2}-\frac{c^2}{2}.$$

If $x\in(c,2c)$ then $[-c,c]\cap[x-2c,x+2c]=[x-2c,c]$ with $x-2c<0$ and so

$$h_c*h_c*h_c(x)=\int_{x-2c}^{0}h_c*h_c(y)dy+\int_{0}^{c}h_c*h_c(y)dy=\frac{4c^2-x^2}{2}+\frac{3}{2}c^2.$$

If $x\in(0,c)$ then $[-c,c]\cap[x-2c,x+2c]=[-c,c]$ and so

$$h_c*h_c*h_c(x)=\int_{-c}^{c}h_c*h_c(y)dy=3c^2.$$

(ii) Let $X_1,X_2,\dots X_n$ are n independent continuous random variables with the same distribution (and so with the same density function f). Assume that $\mathbb{E}(X_i)=\mu$ and $\mathbb{E}(X_i-\mu)^2=\sigma^2$. Show that the density function of $\frac{X_1+\cdots+X_n-\mu}{\sqrt{n}\sigma}$ is given by $\sqrt{n}\sigma f^{*n}\left(x\sqrt{n}\sigma+\mu n\right)$, where $f^{*n}(x)$ is the convolution of f with itself repeated n times.
By Theorem 3.1.3 we get that the density function associated to the sum of X_1,X_2 is $f*f$. Then again by the theorem, the density function associated to the sum of X_1+X_2 with X_3 is $(f*f)*f=f^{*3}$ and so on.
By linearity $\mathbb{E}(X_1+\cdots+X_n)=n\mu$ and by independence we get $\mathbb{E}((X_1+\cdots+X_n-\mu n)^2)=n\sigma^2$. So the sum as $Z=\frac{X_1+\cdots+X_n-\mu n}{\sqrt{n}\sigma}$, we get that Z has $\mathbb{E}(Z)=0$ and $\mathbb{E}(Z^2)=1$ (so it has mean 0 and variance 1).
f^{n*} is the density associated to $X_1+\dots X_n$, we get that $\sqrt{n}\sigma f^{*n}\left(x\sqrt{n}\sigma+\mu n\right)$ is the density associated to Z. Indeed we compute, changing variable,

$$\int_{\mathbb{R}}x\sqrt{n}\sigma f^{*n}\left(x\sqrt{n}\sigma+\mu n\right)dx=\int_{\mathbb{R}}\frac{y-\mu n}{\sqrt{n}\sigma}f^{*n}(y)dy=\frac{1}{\sqrt{n}\sigma}\mathbb{E}(X_1+\cdots+X_n-n\mu)=0$$

$$\int_{\mathbb{R}}x^2\sqrt{n}\sigma f^{*n}\left(x\sqrt{n}\sigma+\mu n\right)dx=\int_{\mathbb{R}}\frac{(y-\mu n)^2}{n\sigma^2}f^{*n}(y)dy=1.$$

(iii) (a) Compute the Fourier transform of $g(x) = e^{-x}\chi_{(0,+\infty)}(x)$.
Recall the following formulas (obtained by integration by parts):

$$\int e^{-y}\sin(xy)dy = -\frac{1}{x^2+1}e^{-y}(x\cos xy + \sin xy) + c$$

$$\int e^{-y}\cos(xy)dy = \frac{1}{x^2+1}e^{-y}(x\sin xy - \cos xy) + c.$$

(b) Compute the Fourier transform of $f(x) = xe^{-x}\chi_{(0,+\infty)}(x)$ (that is the characteristic function of the Gamma distribution).
Use item a. and Proposition 3.2.2.

(a) By definition and using the primitive of the functions $e^{-y}\cos xy$ and $e^{-y}\sin xy$, we get

$$\hat{g}(x) = \int_{\mathbb{R}} g(y)e^{ixy}dy = \int_0^{+\infty} e^{-y}e^{ixy}dy = \int_0^{+\infty} e^{-y}\cos xydy + i\int_0^{+\infty} e^{-y}\sin xydy$$

$$= \frac{1}{x^2+1} + i\frac{x}{x^2+1}.$$

(b) By Proposition 3.2.2,

$$d_x\hat{g}(x) = \int_{\mathbb{R}} (iy)g(y)e^{ixy}dy = i\int_{\mathbb{R}} f(y)e^{ixy}dy = i\hat{f}(x).$$

Therefore

$$\hat{f}(x) = -i\left(\frac{1}{x^2+1} + i\frac{x}{x^2+1}\right)' = -i\left(\frac{-2x}{(x^2+1)^2} - i\frac{x^2-1}{(x^2+1)^2}\right)$$

$$= \frac{1-x^2}{(x^2+1)^2} + i\frac{2x}{(x^2+1)^2} = \left(\frac{1+ix}{1+x^2}\right)^2 = (1-ix)^{-2}$$

where the last identity is obtained by using the fact that $\frac{1}{1-ix} = \frac{1+ix}{1+x^2}$.