PROBABILITY AND MEASURE, LECTURES NOTES MICHAELMAS 2019-2020, E. BREUILLARD

Lecture 1

0. Plan for the course

Useful material for the course include:

• Lecture notes by James Norris available on the course's website.
• the course syllabus approved by the Faculty Board (notice the asterisques signaling non-examinable material). Note that although we will follow the syllabus very closely, the material will not be presented in the order suggested in the "schedule" booklet: most importantly we first construct the Lebesgue measure (by hand directly on $\mathbb{R}^d$) and only then abstract Lebesgue's theory to handle arbitrary measure spaces (rather than present the abstract Carathéodory theorem first, as suggested in the schedule (last revised in 1991...), then specialize to the case of Lebesgue measure on $\mathbb{R}$ and wait until the existence of product measures is finally proved in order to define Lebesgue measure on $\mathbb{R}^d$).
• apart from the books listed in the syllabus, for the measure theory part of the course I would recommend: T. Tao's "An introduction to measure theory" as well as W. Rudin's classic "Real and Complex analysis". Have a look at Halmos's "Measure theory" as well, another classic. For the Ergodic Theory bit at the end, take a look at Einsiedler and Ward.
• a related Part II course is "Linear analysis"; following it is not mandatory, but it can be helpful to understand some of the concepts from a different perspective. A recommended follow-up to this course is the D-course "Analysis of functions".

A rough plan for the course is as follows:

Week 1 Lebesgue measure
Week 2 Abstract measure theory
Week 3 Integration
Week 4 Measure theoretic foundations of probability theory
Week 5 Modes of convergence of random variables
Week 6 L^p spaces, Hilbert space techniques.
Week 7 Fourier transform, gaussian laws, Central Limit Theorem
Week 8 Ergodic theory

The notes below are essentially a write-up of the actual lectures, meant to help the student revise her or his own notes. Occasionally there are some further discussion and explanation as well as few additional remarks that could not be given during the lectures.

1. Boolean algebras and finitely additive measures

Let X be a set.

Definition 1.1. *A* Boolean algebra *on X is a family $\mathcal{B}$ of subsets of X which*
(i) contains $\varnothing$
(ii) is stable under finite union and under complementation.

Note that obviously these assumptions imply that $X \in \mathcal{B}$ and that $\mathcal{B}$ is also stable under finite intersection $A \cap B$, set difference $A \backslash B$, symmetric difference $A \triangle B$.

Examples:

(1) the *trivial Boolean algebra* $\mathcal{B} = \{\varnothing, X\}$
(2) the *discrete Boolean algebra* $\mathcal{B} = 2^X$ = the family of all subsets of X.
(3) If X is a topological space $\mathcal{B}$ = the family of *constructible sets*, that is finite unions of *locally closed subsets* (recall that a locally closed subset is one of the form $O \cap F$, where O is open and F is closed).

Definition 1.2. *A* finitely additive measure *(or* mean*) on* $(X, \mathcal{B})$ *is a function* $m : \mathcal{B} \to [0, +\infty]$ *such that*

(i) $m(\varnothing) = 0$
(ii) $m(E \cup F) = m(E) + m(F)$ *if* $E \cap F = \varnothing$, *and* $E, F \in \mathcal{B}$.

Note that a finitely additive measure is also *sub-additive*, namely $m(E \cup F) \leqslant m(E) + m(F)$ for all $E, F \in \mathcal{B}$, and *monotone*: if $E \subset F$, then $m(E) \leqslant m(F)$.

Examples:

(1) if X is any set and $\mathcal{B} = 2^X$ is the discrete Boolean algebra, set $m(E) = \#E$, the number of elements in $E \subset X$, is called the *counting measure* on X.
(2) more generally if $f : X \to [0, +\infty]$ is a function, then $m(E) := \sum_{e \in E} f(e)$ is a finitely additive measure.
(3) If $X = \bigsqcup_i X_i$ is a finite partition of X and $\mathcal{B}$ is the Boolean algebra it generates (i.e. subsets in $\mathcal{B}$ are unions of X_i's), if we assign weights $a_i \geqslant 0$ to each X_i, we can set $m(E) := \sum_{i; X_i \subset E} a_i$ and get this way a finitely additive measure on $(X, \mathcal{B})$.

2. Jordan measure on $\mathbb{R}^d$

This is a notion defined by Camille Jordan in the 19th century. Of course the idea that one can compute the volume of a body by counting the number of small cubes needed to approximate it within a reasonable error goes back (at least) to Archimedes. The Jordan measure is an early attempt to formalise this idea. It will be one of the aims of the course to explain how this first attempt has been surpassed in the 20th century by the advent of Lebesgue measure and its subsequent generalisation to abstract measure spaces that forms *Measure Theory*. So let us first have a look at this notion.

Definition 2.1. *A subset of* $\mathbb{R}^d$ *is called* elementary *if it is a finite union of* boxes $B = I_1 \times \cdots \times I_d$, *where* I_i *is a finite interval of* $\mathbb{R}$.

Recall that a finite interval of $\mathbb{R}$ has the form $[a, b]$ or (a, b) or $(a, b]$ or $[a, b)$ for some reals $a \leqslant b$. Given a box B we can define its volume $|B|$ by setting

$$|B| = \prod_{i=1}^{d} |a_i - b_i|$$

if $B = I_1 \times \cdots \times I_d$, and $(a_i, b_i) \subset I_i \subset [a_i, b_i]$ for each i.

Proposition 2.2. *Let* $B = I_1 \times \cdots \times I_d \subset \mathbb{R}^d$ *be a box as above and* $\mathcal{E}(B)$ *the family of all elementary subsets of* B. *Then*

(a) $\mathcal{E}(B)$ *is a Boolean algebra,*
(b) every $E \in \mathcal{E}(B)$ *is a finite union of disjoint boxes,*

(c) if $E \in \mathcal{E}(B)$ is written in two different ways as a finite union of disjoint boxes, i.e.

$$E = \bigsqcup_{i=1}^{N} B_i = \bigsqcup_{j=1}^{N'} B'_j,$$

then

$$\sum_{i=1}^{N} |B_i| = \sum_{j=1}^{N'} |B'_j|$$

Proof sketch. It helps to first convince oneself that this is true in dimension $d = 1$. The general case isn't that much harder. Clearly the intersection of two boxes is a box and if $B_1 \subset B_2$ are boxes, then $B_2 \backslash B_1$ is a finite union of disjoint boxes. (a) follows. And (b) also, by induction on the number of boxes whose union is E. (c) too is pretty obvious: writing $E = \bigsqcup_{i,j} B_i \cap B'_j$ it is enough to prove it for one box, say $E = B_1 = I_1 \times \cdots \times I_d$, and this case is easily checked by refining the partition $(B'_j)_j$ into a grid partition of smaller boxes whose sides are arbitrary pieces of the partition of each I_i formed by the projections of the B'_j's onto the i-th coordinate. □

Proposition-Definition 2.3. *We may set*

$$m(E) := \sum_{i=1}^{N} |B_i|$$

for each $E \in \mathcal{E}(B)$ of the form $E = \bigsqcup_{i=1}^{N} B_i$ a disjoint union of boxes. Then m defines a finitely additive measure on the Boolean algebra $(B, \mathcal{E}(B))$.

Proof. It is well-defined and finitely additive by (c) of Proposition 2.2. □

Definition 2.4. *A subset $A \subset \mathbb{R}^d$ is called* Jordan measurable *if for all $\varepsilon > 0$ there exist elementary subsets E, F with $E \subset A \subset F$ such that*

$$m(F \backslash E) < \varepsilon.$$

Remark 2.5. Equivalently A is Jordan measurable if for each $\varepsilon > 0$ there is a finite union of boxes $F = \bigcup_{i=1}^{N} B_i$ containing A, such that $F \backslash A$ is contained in an elementary set of measure $< \varepsilon$.

Remark 2.6. Jordan measurable subsets of $\mathbb{R}^d$ are bounded (because so are elementary subsets).

Exercise/Example: If $f : [0,1] \to \mathbb{R}$ is a continuous function, then the subgraph $\{(x,y) \in \mathbb{R}^2, x \in [0,1], 0 \leqslant y \leqslant f(x)\}$ is Jordan measurable (thanks are due to Erik Ma for catching a mistake in the formulation of that example in an earlier version of these notes!)

Definition 2.7. *If $A \subset \mathbb{R}^d$ is Jordan measurable, we may define its measure $m(A)$ by*

$$m(A) = \inf\{m(F), A \subset F, F \text{ elementary}\}.$$

Note that $m(A)$ is also equal to $\sup\{m(E), A \supset E, E \text{ elementary}\}$, because by the defining property of Jordan measurability, for every $\varepsilon > 0$ there are elementary sets E, F with $E \subset A \subset F$ and $m(F \backslash E) < \varepsilon$. And by finite additivity of m, we have $m(E) = m(F) - m(F \backslash E) \geqslant m(A) - \varepsilon$.

Proposition 2.8. *If $B \subset \mathbb{R}^d$ is a box, then the class $\mathcal{J}(B)$ of Jordan measurable subsets of B forms a Boolean algebra and m is a finitely additive measure on $(B, \mathcal{J}(B))$.*

Proof. It is immediate from the definition that $\mathcal{J}(B)$ is stable under complementation. It is also clear that it is stable under finite unions. So we have a Boolean algebra. To see that it is finitely additive, i.e. $m(A \cup A') = m(A) + m(A')$ if A, A' are disjoint in $\mathcal{J}(B)$, use the previous remark to find for each $\varepsilon > 0$ elementary sets $E \subset A$ and $E' \subset A'$ with $m(E) \geqslant m(A) - \varepsilon$ and $m(E') \geqslant m(A') - \varepsilon$. Clearly E and E' are disjoint, so $m(E \cup E') = m(E) + m(E') \geqslant m(A) + m(A') - \varepsilon$. This yields $m(A \cup A') \geqslant m(A) + m(A')$ and the reverse inequality is clear by definition of m. □

The notions of Jordan measurability for sets and Riemann integrability for functions are tightly connected. In the example sheet, you will find the following

Exercise: Given a finite interval $[a, b]$ of $\mathbb{R}$, a subset $E \subset [a, b]$ is Jordan measurable if and only if the indicator funtion $1_E(x)$ is Riemann integrable.

Lecture 2

3. Lebesgue measurable sets

There are some issues with the Jordan measure:

(i) unbounded sets in $\mathbb{R}^d$ are not Jordan measurable.

(ii) many simple minded bounded sets are not Jordan measurable, e.g. $A := \mathbb{Q}\cap[0,1]$ is not Jordan measurable (indeed if $E \subset A \subset F$ with E, F elementary, then E must be finite and F must contain $[0,1]$, so $m(F\backslash E) = 1$).

(iii) as hinted in the exercise above, the integration theory associated to the notion of Jordan measurability is the good old notion of Riemann integrability. This has well-known shortcomings, for example pointwise limits of Riemann integrable functions are not necessarily Riemann integrable, e.g. $1_{[0,1]\cap\frac{1}{n!}\mathbb{Z}} =: f_n$ has $f_n \to f := 1_{\mathbb{Q}\cap[0,1]}$ pointwise, and the f_n's are Riemann integrable, while f isn't. By the same token, an infinite series of Riemann integrable functions may not be Riemann integrable, which was a major problem in the theory of Fourier series.

For all these reasons mathematicians at the end of the 19th century looked for another definition of measurability for subsets of $\mathbb{R}^d$ (and integrability for functions on $\mathbb{R}^d$) that would be more robust and give a sound basis to Analysis, which was until then mostly confined to continuous functions. This is what Henri Lebesgue achieved in 1901. His main idea: allow countable unions of boxes in Jordan's definition.

Definition 3.1. *For any subset $E \subset \mathbb{R}^d$ we can define its Lebesgue outer-measure as*

$$m^*(E) := \inf\{\sum_{n\geqslant 1} |B_n|, E \subset \bigcup_{n\geqslant 1} B_n, B_n \text{ a box in } \mathbb{R}^d\}.$$

Here $|B|$ is the volume of a box (i.e. the product of side lengths) as in the previous lecture. Note that m^* is translation invariant, namely $m^*(E+x) = m^*(E)$ for any subset $E \subset \mathbb{R}^d$ and any $x \in \mathbb{R}^d$.

Definition 3.2. *A subset $E \subset \mathbb{R}^d$ is called Lebesgue measurable if for all $\varepsilon > 0$ there is $C := \bigcup_n B_n$ a countable union of boxes B_n's, such that $E \subset C$ and*

$$m^*(C\backslash E) < \varepsilon.$$

Note that the family $\mathcal{L}$ is clearly invariant under translation: i.e. of $E \in \mathcal{L}$ then $E + x \in \mathcal{L}$ for every $x \in \mathbb{R}^d$. Clearly it also scales naturally: $m^*(\lambda E) = \lambda^d m^*(E)$ for all $\lambda \in (0, +\infty)$.

Remark 3.3. In the above definitions, we may always assume that the boxes are open (i.e. Cartesian products of d open intervals (a_i, b_i)). Indeed we can always change $B_n = \prod_1^d [a_i, b_i]$ into the slightly bigger $B'_n := \prod_1^d (a_i - \varepsilon_n, b_i + \varepsilon_n)$. This will only affect the volume of each ball by a small amount: $|B'_n| \leqslant |B_n| + \varepsilon 2^{-n}$ if ε_n is chosen small enough and $\varepsilon > 0$ is fixed but arbitrary, hence also the total sum will be $\sum_n |B'_n| \leqslant \varepsilon + \sum_n |B_n|$ and this will not change the definition of m^*.

Remark 3.4. Note that every Jordan measurable set is clearly Lebesgue measurable.

The main proposition today is:

Proposition 3.5. *(a) m^* extends m, namely $m^*(E) = m(E)$ if E is Jordan measurable.*

(b) The family $\mathcal{L}$ of Lebesgue measurable subsets of $\mathbb{R}^d$ forms a Boolean algebra stable under countable unions.

(c) m^* *is countably additive on* $(\mathbb{R}^d, \mathcal{L})$.

Countably additive means that $m^*(\bigcup_{n\geqslant 1} E_n) = \sum_{n\geqslant 1} m^*(E_n)$ for every countable (i.e. finite or countably infinite) family $\{E_n\}_n$ of pairwise disjoints subsets from $\mathcal{L}$.

From (b) we see that for example $\mathbb{Q}$ is a Lebesgue measurable subset of $\mathbb{R}$ and so is $\mathbb{Q} \cap [0,1]$. When restricted to the class $\mathcal{L}$ of Lebesgue measurable sets the outer-measure m^* is called the Lebesgue measure.

Not every subset of $\mathbb{R}^d$ is Lebesgue measurable (we will see an example shortly, assuming the axiom of choice). By the same token, we'll also see that m^* is not even finitely additive on all subsets of $\mathbb{R}^d$.

Remark 3.6. This remark can be skipped. There is an apparent asymmetry in our definition of a Lebesgue measurable set in that we only approximate the set E *from above* by a union of boxes and not also from below as in the original definition of Jordan measurability. Our definition of Lebesgue measurability is the exact analogue of the equivalent definition of Jordan measurability given in Remark 2.5 from the last lecture. If we were to also approximate from below by a countable union of boxes we would run into trouble: for example there are closed subsets of $[0,1]$ (hence Lebesgue measurable as we shall see) that have empty interior and positive Lebesgue measure (such as some Cantor sets, cf. the example sheet), clearly those cannot be approximated from below by unions of boxes. This asymmetry is responsible for the fact that, with our definition, it is not obvious that the complement of a Lebesgue measurable set is again Lebesgue measurable and we will have to work a bit to establish it.

It will take some time to prove Proposition 3.5 in full. Today we'll prove (a). Next time we'll prove that open and closed sets are in $\mathcal{L}$, and then establish that $\mathcal{L}$ is stable under complementation, and finally prove (b) and (c). Actually the hardest part of the proof of Proposition 3.5 will be to show that m^* is finitely additive. First we give some basic properties of m^*.

Lemma 3.7. *The set function* m^* *is*

(i) monotone, i.e. $A \subset B$ *implies* $m^*(A) \leqslant m^*(B)$,

(ii) countably sub-additive, i.e. for any countable family $\{A_n\}_n$ *of subsets of* $\mathbb{R}^d$,

$$m^*\Big(\bigcup_{n\geqslant 1} A_n\Big) \leqslant \sum_{n\geqslant 1} m^*(A_n).$$

Proof. (i) is obvious and (ii) is pretty clear as well. Indeed pick $\varepsilon > 0$ and let $C_n := \bigcup_{i\geqslant 1} B_{n,i}$ be a countable union of boxes such that $A_n \subset C_n$ and $\sum_{i\geqslant 1} |B_{n,i}| \leqslant m^*(A_n) + \varepsilon/2^n$. Then $\bigcup_n A_n \subset \bigcup_n C_n = \bigcup_{n,i} B_{n,i}$ and

$$m^*\Big(\bigcup_{n\geqslant 1} A_n\Big) \leqslant \sum_{n,i} |B_{n,i}| \leqslant \sum_{n\geqslant 1} m^*(A_n) + \varepsilon/2^n = \varepsilon + \sum_{n\geqslant 1} m^*(A_n),$$

and since ε is arbitrary we get:

$$m^*\Big(\bigcup_{n\geqslant 1} A_n\Big) \leqslant \sum_{n\geqslant 1} m^*(A_n)$$

as desired. □

Now we make the following remark: assertions (a) and (c) of Proposition 3.5 clearly imply that the Jordan measure m is countably additive on Jordan measurable sets. In particular, if $(E_n)_{n\geqslant 1}$ is a decreasing sequence of elementary sets with empty intersection, then

$$m(E_1) = \sum_{n\geqslant 1} m(E_n \backslash E_{n+1})$$

so

$$0 = m(E_1) - \lim_{N\to+\infty} \sum_{n=1}^{N-1} (m(E_n) - m(E_{n+1})) = \lim_{N\to+\infty} m(E_N)$$

This last property is known as the *continuity property* of the measure m. In fact countable additivity is equivalent to the conjunction of finite additivity and the continuity property (this is an Exercise in the Example sheet). So to sum up: Proposition 3.5 implies that the Jordan measure has the continuity property on elementary sets. Let us prove this directly, because this fact will be useful on the way to the proof of Proposition 3.5.

Lemma 3.8. *The Jordan measure has the continuity property on elementary sets, namely if $(E_n)_{n\geqslant 1}$ is a non-increasing (for set inclusion, i.e. $E_{n+1} \subset E_n$) sequence of elementary sets with empty intersection, then*

$$\lim_{n\to+\infty} m(E_n) = 0.$$

Proof. The proof uses a basic topological property of $\mathbb{R}^d$, namely the Heine-Borel property that bounded and closed sets are compact. We argue by contradiction. The limit exists, because the sequence $m(E_n)$ is non-increasing, so suppose it is positive, say $> 2\varepsilon$ for some $\varepsilon > 0$. Then $m(E_n) \geqslant 2\varepsilon$ for all n. Recall that elementary sets are finite union of boxes. Since the E_n may not be closed, we can shrink a bit the sides of each box making E_n and find a closed elementary set $F_n \subset E_n$ such that $m(E_n \backslash F_n) < \varepsilon/2^n$ for each $n \geqslant 1$. Then

$$m(E_n\backslash(F_1\cap\ldots\cap F_n) = m(\bigcup_{i=1}^{n}(E_n\backslash F_i)) \leqslant \sum_{1}^{n} m(E_n\backslash F_i) \leqslant \sum_{1}^{n} m(E_i\backslash F_i) \leqslant \varepsilon \sum_{1}^{n} 2^{-i} \leqslant \varepsilon$$

where we used the (finite) sub-additivity of m on elementary sets. In particular $m(F_1 \cap \ldots \cap F_n) \geqslant \varepsilon$, hence non-zero, for all n. But $\bigcap_n F_n$ is empty and the F_n's closed and bounded. Hence by Heine-Borel there is a finite N such that $\bigcap_1^N F_n$ is empty. This is a contradiction. □

We now begin the proof of Proposition 3.5 proper.

Proof of (a) in Prop. 3.5. We need to show that m^* extends m, the Jordan measure, on Jordan measurable sets. Note that we've already checked that $m(B) = |B|$ in the first lecture.

First of all it is clear from the definitions that $m^*(A) \leqslant m(A)$ for every Jordan measurable set A.

We have to prove the reverse inequality. To begin with let us assume that A is an elementary set. By definition of m^*, given $\varepsilon > 0$ there is a countable family of boxes B_n such that $A \subset \bigcup_n B_n$ and $\sum_n |B_n| \leqslant m^*(A) + \varepsilon$. Let $E_n = A\backslash(B_1 \cup \ldots \cup B_n)$. It is an elementary set. Moreover $E_{n+1} \subset E_n$ and $\bigcap_n E_n = \varnothing$. So Lemma 3.8 applies and we get $m(E_n) \to 0$. But

$$m(A) \leqslant m(A\backslash(B_1 \cup \ldots \cup B_n)) + m(B_1 \cup \ldots \cup B_n) \leqslant m(E_n) + \sum_i |B_i|$$

which implies that $m(A) \leqslant m^*(A) + \varepsilon$, and hence that $m(A) \leqslant m^*(A)$ as desired.

Finally if A is an arbitrary Jordan measurable set, then by definition for each $\varepsilon > 0$ there is an elementary set $E \subset A$ such that $m(A) \leqslant m(E) + \varepsilon$. But $m(E) = m^*(E)$ by the above, and $m^*(E) \leqslant m^*(A)$ by monotonicity. Since ε is arbitrary we conclude that $m(A) = m^*(A)$. □

Lecture 3

Recall our definition of the Lebesgue outer-measure m^* and of Lebesgue measurable sets from the last lecture. Recall that we proved that m^* coincides with the Jordan measure m on Jordan measurable sets. Today we will complete the proof of Proposition 3.5 by showing that the complement of a Lebesgue measurable set is Lebesgue measurable and that m^* is countably additive. We will also discuss an important example of non-measurable set.

We begin with the following

Proof that $\mathcal{L}$ is stable under countable unions. This is easy. If $A = \bigcup_{n\geqslant 1} A_n$ and each $A_n \in \mathcal{L}$, then for each $\varepsilon > 0$ there are countable unions of boxes $C_n := \bigcup_i B_{n,i}$ with $A_n \subset C_n$ such that $m^*(C_n \backslash A_n) \leqslant \varepsilon/2^n$. Then $\bigcup_n C_n$ too is a countable union of boxes and by sub-additivity of m^* (Lemma 3.7)

$$m^*(\bigcup_n C_n \backslash A) \leqslant \sum_n m^*(C_n \backslash A) \leqslant \sum_n \varepsilon/2^n = \varepsilon.$$

This shows that $A \in \mathcal{L}$. □

Lemma 3.9. *If $A = \bigcap_n E_n$ is a countable intersection of elementary sets E_n with $E_{n+1} \subset E_n$, then A is Lebesgue measurable and $m(E_n) \to m^*(A)$. In particular countable intersections of elementary sets are Lebesgue measurable.*

Proof. Recall that elementary sets are finite union of boxes. If A is a countable intersection of elementary sets, say $A = \bigcap E_n$, with E_n elementary, without loss of generality we can assume that $E_{n+1} \subset E_n$ (simply replace E_n by $E_1 \cap \ldots \cap E_n$). Then $A \subset E_n$ and $E_n \backslash A = \bigcup_{i\geqslant n} E_i \backslash E_{i+1}$. In particular

$$m^*(E_n \backslash A) \leqslant \sum_{i\geqslant n} m^*(E_i \backslash E_{i+1}) \tag{3.1}$$

by sub-additivity of m^*. But $E_i \backslash E_{i+1}$ is elementary and we have seen that $m = m^*$ on elementary sets, so

$$\sum_{i\geqslant n} m^*(E_i \backslash E_{i+1}) = \sum_{i\geqslant n} m(E_i \backslash E_{i+1}) = m(E_n) - \lim_{i\to+\infty} m(E_i) \leqslant m(E_n) < +\infty$$

In particular the right hand side in (3.1) is the remainder of a convergent series and, hence must tend to zero as n tends to infinity. This implies that A is Lebesgue measurable (by definition). Also note that by sub-additivity:

$$m^*(A) \leqslant m(E_n) = m^*(E_n) \leqslant m^*(E_n \backslash A) + m^*(A),$$

which means that $m(E_n) \to m^*(A)$ as $n \to +\infty$. □

Corollary 3.10. *Every open and every closed subset of $\mathbb{R}^d$ is Lebesgue measurable.*

Proof. Every open set is a countable union of open boxes, hence must be in $\mathcal{L}$ by stability of $\mathcal{L}$ under countable unions. Now if $C \subset \mathbb{R}^d$ is closed, then $C = \bigcup_n C \cap B_n$, where B_n is the closed box $[-n, n]^d$. So to prove that C is in $\mathcal{L}$ it is enough to assume that C is bounded, i.e. belongs to some open box B. But then $B \backslash C$ is open, hence a countable union of boxes B_n contained in B. But $B \backslash B_n$ is elementary. Hence $C = \bigcap_n B \backslash B_n$ is a countable intersection of elementary sets, hence in $\mathcal{L}$ by the previous lemma. □

Definition 3.11. *A subset $E \subset \mathbb{R}^d$ is called a null set if $m^*(E) = 0$.*

Lemma 3.12. *Null sets are Lebesgue measurable.*

Proof. This is clear: given any $\varepsilon > 0$ there is a countable union of boxes $C = \bigcup_i B_i$ such that $E \subset C$ and $\sum_i |B_i| < \varepsilon$. In particular

$$m^*(C\backslash E) \leqslant m^*(C) \leqslant \sum_i |B_i| < \varepsilon.$$

□

Proof of (b) in Prop. 3.5. We have already proved that $\mathcal{L}$ is stable under countable unions. We need to show that $\mathcal{L}$ is stable under complementation. Namely given $E \in \mathcal{L}$ we want to prove that $E^c \in \mathcal{L}$. Note first that it is enough to prove that $B\backslash E$ is in $\mathcal{L}$ for any box B, because $E^c = \bigcup_n [-n, n]^d \backslash E$ is a countable union of such sets and $\mathcal{L}$ is stable under countable unions. By definition of Lebesgue measurability for each $n \geqslant 1$ there is a countable union of boxes C_n such that $E \subset C_n$ and $m^*(C_n\backslash E) < 1/n$. But each $B\backslash C_n$ is a countable intersection of elementary sets, hence by Lemma 3.9, belongs to $\mathcal{L}$. So $F := \bigcup_n B\backslash C_n \in \mathcal{L}$. Note that $F \subset B\backslash E$ and that $(B\backslash E)\backslash F$ is a null set, because for each n we have

$$m^*((B\backslash E)\backslash F) \leqslant m^*((B\backslash E)\backslash B\backslash C_n)) \leqslant m^*(C_n\backslash E) < 1/n.$$

So $(B\backslash E)\backslash F$ is in $\mathcal{L}$ by the previous lemma, and hence $B\backslash E \in \mathcal{L}$. □

We can now show that we can approximate any Lebesgue measurable set by closed sets from below and open sets from above:

Proposition 3.13. *Assume that $E \subset \mathbb{R}^d$ is a Lebesgue measurable subset and let $\varepsilon > 0$. Then there exists a closed subset F and an open set U such that $F \subset E \subset U$ and*

$$m^*(U\backslash F) < \varepsilon.$$

Moreover E can be written as a disjoint union $E = B\backslash N$, where B is a countable intersection of open sets and N is a null set (and analogously $E = C \sqcup M$, where M is null and C a countable union of closed sets).

Proof. For U we can take a countable union of open boxes $U = \bigcup_i B_i$ such that $m^*(U\backslash E) < \varepsilon/2$ as given by the definition of Lebesgue measurability. Now that we know that the complement E^c is also in $\mathcal{L}$, we can do the same for E^c and find an open set $\Omega \supset E^c$ such that $m^*(\Omega\backslash E^c) < \varepsilon/2$. Then set $C = \Omega^c$. Clearly $\Omega\backslash E^c = E\backslash C$ and C is closed, so we are done.

Now letting $\varepsilon = 1/n$ for each n we get an open set $U_n \supset E$ with $m^*(U_n\backslash E) < 1/n$ and set $B = \bigcap_n U_n$. Then $B\backslash E$ is null (apply this to E^c to get the analogous statement in brackets). □

Proof of (c) in Prop. 3.5. We need to show that m^* is countably additive on $\mathcal{L}$. Namely given a countable family of pairwise disjoint subsets E_n from $\mathcal{L}$ we need to show that

$$m^*(\bigcup_1^\infty E_n) = \sum_{n \geqslant 1} m^*(E_n).$$

We will prove this after a series of initial reduction steps.

(i) *first reduction step:* wlog we may assume that each E_n is a bounded subset of $\mathbb{R}^d$. Indeed, we may decompose $\mathbb{R}^d$ into a countable disjoint union of bounded sets (e.g. let $B_n = [-n, n]^d$ exhaust $\mathbb{R}^d$ and write $\mathbb{R}^d = \bigsqcup_n X_n$, where $X_n := B_n\backslash(B_1 \cup \ldots \cup B_{n-1})$). Then $E := \bigsqcup_n E_n = \bigsqcup_{n,m} E_n \cap X_m$, and each $E_n \cap X_m$ is bounded, so if m^* is countably additive on bounded sets, then for each n, $\sum_m m^*(E_n \cap X_m) = m^*(E_n)$ and

$$m^*(E) = \sum_{m,n} m^*(E_n \cap X_m) = \sum_n m^*(E_n)$$

as desired.

(ii) *second reduction step:* it is enough to prove that m^* is finitely additive on bounded sets. Indeed, if the E_n's are pairwise disjoint and bounded, for every N

$$\sum_1^N m^*(E_n) = m^*(\bigcup_1^N E_n) \leqslant m^*(\bigcup_1^\infty E_n) \leqslant \sum_{n\geqslant 1} m^*(E_n)$$

where we used finite additivity on the left hand side and sub-additivity of m^* on the right hand side. Letting N tend to infinity we conclude that $m^*(\bigcup_1^\infty E_n) = \sum_{n\geqslant 1} m^*(E_n)$.

So we are left to show that m^* is finitely additive on bounded sets, or in other words that

$$m^*(E \cup F) = m^*(E) + m^*(F) \tag{3.2}$$

whenever E, F are disjoint bounded subsets from $\mathcal{L}$. This will follow from the finite additivity of m on elementary sets.

(iii) *third reduction step:* it is enough to prove (3.2) when both E and F are countable intersections of elementary sets. Indeed, since E is bounded, I claim that for each $\varepsilon > 0$, there is a countable intersection of elementary sets C such that $m^*(E\backslash C) < \varepsilon$ (this is quite clear from the definition: there is a box B containing E, and $B\backslash E$ belongs to $\mathcal{L}$ (we have already shown that $\mathcal{L}$ is a Boolean algebra!), so there is a countable union of boxes $U = \bigcup_i B_i$ containing $B\backslash E$ such that $m^*(U\backslash(B\backslash E)) < \varepsilon$. Just set $C := \bigcap_i (B\backslash(B\cap B_i))$). Now do the same for F to get $D \subset F$ a countable intersection of elementary sets, such that $m^*(F\backslash D) < \varepsilon$. Finally if we knew that $m^*(C \cup D) = m^*(C) + m^*(D)$, then we would get:

$$m^*(E)+m^*(F)-2\varepsilon \leqslant m^*(C)+m^*(D) = m^*(C\cup D) \leqslant m^*(E\cup F) \leqslant m^*(E)+m^*(F)$$

so letting $\varepsilon \to 0$ we would be done.

So it all remains to prove (3.2) under the assumption that both E and F are countable intersections of elementary sets. For this we will use Lemma 3.9. Let $E = \bigcap_n I_n$ and $F = \bigcap_n J_n$, with I_n and J_n elementary. Without loss of generality (replace I_n by $I_1 \cap \ldots \cap I_n$) we may assume that $I_{n+1} \subset I_n$ and similarly $J_{n+1} \subset J_n$. Now observe that

$$E \cup F = \bigcap_n (I_n \cup J_n)$$

is also a countable intersection of elementary sets. Besides $\bigcap_n (I_n \cap J_n) = E \cap F = \varnothing$. We know by Lemma 3.9 that $m(I_n), m(J_n), m(I_n \cup J_n)$ and $m(I_n \cap J_n)$ converge respectively to $m^*(E), m^*(F), m^*(E \cup F)$ and 0. However by finite additivity of m on elementary sets:

$$m(I_n) + m(J_n) = m(I_n \cup J_n) + m(I_n \cap J_n)$$

which implies (3.2) in the limit as $n \to +\infty$. This ends the proof of Proposition 3.5(c).

□

Lecture 4

Alright, we have now finished the proof of Proposition 3.5. Let's sum up: we have defined the class $\mathcal{L}$ of Lebesgue measurable subsets of $\mathbb{R}^d$. We have shown that it is a Boolean algebra stable under countable unions (and hence also countable intersections!). Furthermore we have shown that the outer-measure m^* (which makes sense for arbitrary subsets of $\mathbb{R}^d$) is additive and even countably additive on $\mathcal{L}$. We call it the Lebesgue measure. We've also shown along the way that null sets are Lebesgue measurable and that every open and every closed subset of $\mathbb{R}^d$ is Lebesgue measurable and actually every Lebesgue measurable set can be "approximated" (up to sets of arbitrarily small measure) from below by a closed subset and from above by an open subset. So one may wonder: is every subset of $\mathbb{R}^d$ Lebesgue measurable? Well, assuming the axiom of choice (which is something the overwhelming majority of mathematicians are willing to do), we will construct a counter-example, i.e. a non-measurable subset. This example was found by Guiseppe Vitali in the wake of Lebesgue's discovery.

Vitali's counter-example. We are going to construct a subset of $\mathbb{R}$, which is not Lebesgue measurable. The idea is to consider a set of representatives of the cosets of the additive group of $\mathbb{Q}$ inside $\mathbb{R}$. We could use any countable dense subgroup of $\mathbb{R}$ in place of $\mathbb{Q}$, but let's use $\mathbb{Q}$ as Vitali did to make it more concrete. We also restrict to $[0,1]$ for definiteness.

So let $E \subset [0,1]$ be a set of representatives of $(\mathbb{Q},+)$ in $(\mathbb{R},+)$, namely in each coset $x+\mathbb{Q}$ we pick an element lying in $[0,1]$. So E is such that for every $x \in \mathbb{R}$ there is a unique $e \in E$ such that $x-e \in \mathbb{Q}$. Of course it is the axiom of choice that allows us to assert that E is indeed a subset of $\mathbb{R}$.

Claim 1. m^* is not (finitely) additive on the family of all subsets of $\mathbb{R}^d$.

Claim 2. E is not Lebesgue measurable.

Proof. Note that, by construction, if $r_1, ..., r_N$ are N pairwise distinct rational numbers, then the subsets $r_i + E$ for $i = 1, ..., N$ are pairwise disjoint. So if m^* were finitely additive on the Boolean algebra of all subsets of $\mathbb{R}$, then we could write:

$$m^*(\bigcup_1^N (r_i + E)) = \sum_1^N m^*(r_i + E) = N m^*(E) \tag{3.3}$$

where we use the fact (this is clear, as we've already observed) that m^* is translation invariant. But if we assume that the r_i's belong to $[0,1]$ say, then $r_i + E \subset [0,2]$ and so by monotonicity of m^* we would get:

$$N m^*(E) \leqslant m^*([0,2]) = m([0,2]) = 2$$

because we've already proved that $m^* = m$ on elementary sets. Letting N tend to $+\infty$ this would mean that

$$m^*(E) = 0.$$

However by construction $[0,1] \subset \bigcup_{r\in\mathbb{Q}} (E+r)$ and hence, by countable sub-additivity of m^* we would get:

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

clearly a contradiction. This concludes the proof of Claim 1.

To see that $E \notin \mathcal{L}$ simply argue that otherwise $E + r \in \mathcal{L}$ for each $r \in \mathbb{Q}$ and thus (3.3) would be legitimate because we've shown (Prop. 3.5) that m^* is additive on $\mathcal{L}$. This would yield $m^*(E) = 0$ as above and lead to the same contradiction. This proves Claim 2.

□

Remark 3.14. Note that the Vitali set E must have positive outer measure, i.e. $m^*(E) > 0$ because null sets are Lebesgue measurable. Someone asked in class whether there are non-Lebesgue measurable sets of arbitrary (non-zero) measure. The answer is clearly yes, because given some scaling factor $\lambda > 0$, E is not in $\mathcal{L}$ if and only λE is not in $\mathcal{L}$, while $m^*(\lambda E) = \lambda m^*(E)$ can clearly achieve any value as λ varies. In fact, λE will be a Vitali set for the dense additive subgroup $\lambda^{-1}\mathbb{Q}$.

A logical aside: the axiom of choice is independent of the Zermelo-Frenkel axioms that form the foundation of most mathematics today. This means that neither it nor its negation can be proven assuming only the ZF axioms. This was a major result of 20th century Logic obtained by Kurt Gödel and Paul Cohen. In the 1970's the logician Robert Solovay went further to show that one can construct models of the real numbers in ZF in which all subsets of $\mathbb{R}$ are Lebesgue measurable.

4. Abstract measure theory

Now that we've understood the construction of the Lebesgue measure, we are ripe to lay the foundation of abstract measure theory. In the early 20-th century, after Lebesgue's ideas became widespread and accepted (despite some initial criticism) by most mathematicians, people started to understand that they were much more general, that Lebesgue's construction could be made to work in the abstract, not just on $\mathbb{R}^d$ but on any set, even without a topology. Even though Lebesgue himself seemed to have been somewhat reluctant to generalisations, it was soon recognized by people such as Felix Hausdorff, Constantin Carathéodory or Maurice Fréchet, that one could gain a lot from such a point of view. Later developments, such as Kolmogorov's axiomatic approach to probability theory, proved them right.

Let X be a set.

Definition 4.1. *A σ-algebra on a set X is a Boolean algebra of subsets of X, which is stable under countable unions.*

Note that (taking complements) it is clearly also stable under countable intersections. (the letter σ is for "countable", it's widespread notation, I don't know the rationale behind it).

Definition 4.2. *A measurable space is a couple $(X, \mathcal{A})$, where X is a set and $\mathcal{A}$ a σ-algebra on X.*

Definition 4.3. *A measure on $(X, \mathcal{A})$ is a map $\mu : \mathcal{A} \to [0, +\infty]$ such that*

(i) $\mu(\varnothing) = 0$

(ii) μ *is countably additive, i.e.*

$$\mu(\bigsqcup_{n\geqslant 1} E_n) = \sum_{n\geqslant 1} \mu(E_n)$$

if the E_n's are in $\mathcal{A}$ and pairwise disjoint.

A triple $(X, \mathcal{A}, \mu)$ is then called a measure space.

Examples

(1) $(\mathbb{R}^d, \mathcal{L}, m)$, where m is the Lebesgue measure, is a measure space (this is content of Proposition 3.5).
(2) If $A_0 \in \mathcal{L}$ then $m_0(E) := m(A_0 \cap E)$ defines another measure on $(\mathbb{R}^d, \mathcal{L})$.
(3) $(X, 2^X, \#)$ is a measure space (where 2^X is the discrete Boolean algebra and $\#$ counting measure).

(4) pick a sequence $(a_n)_{n\geqslant 1}$ of non-negative real numbers, then $(\mathbb{N}, 2^{\mathbb{N}}, \mu)$ is a measure space, where $\mu(I) = \sum_{i\in I} a_i$ defines the measure for every subset $I \subset \mathbb{N}$.

Proposition 4.4. *Let $(X, \mathcal{A}, \mu)$ be a measure space.*

(a) μ is monotone, i.e. for $A \subset B$ in $\mathcal{B}$ we have $\mu(A) \leqslant \mu(B)$.

(b) μ is countably sub-additive, i.e. $\mu(\bigcup_{n\geqslant 1} E_n) \leqslant \sum_{n\geqslant 1} \mu(E_n)$ for every sequence of sets $E_n \in \mathcal{A}$.

(c) We have upward monotone convergence, i.e. if $E_1 \subset E_2 \subset \ldots \subset E_n \subset \ldots$ are all in $\mathcal{A}$, then

$$\mu(\bigcup_n E_n) = \lim_{n\to+\infty} \mu(E_n) = \sup_{n\geqslant 1} \mu(E_n).$$

(d) and downward monotone convergence, i.e. if $E_1 \supset E_2 \supset \ldots \supset E_n \supset \ldots$ are all in $\mathcal{A}$, and if $\mu(E_1) < \infty$, then

$$\mu(\bigcap_n E_n) = \lim_{n\to+\infty} \mu(E_n) = \inf_{n\geqslant 1} \mu(E_n).$$

Caveat: note the extra condition in (d): it is necessary to assume that $\mu(E_1) < \infty$. On the real line with Lebesgue measure, you could take for instance $E_n = [n, +\infty)$ and see that $m(E_n) = +\infty$ for all n, while $\bigcap_n E_n$ is empty.

Proof. (a) write $\mu(B) = \mu(B\backslash A) + \mu(A)$.

(b) write $\bigcup E_n = \bigsqcup F_n$, where $F_n = E_n\backslash(E_1 \cup \ldots \cup E_{n-1})$, so $\mu(\bigcup_n E_n) = \sum \mu(F_n) \leqslant \sum \mu(E_n)$.

(c) set $E_0 = \varnothing$. Write $\bigcup E_n = \bigsqcup F_n$ as in (b). We have:

$$\sum_1^N \mu(F_n) = \sum_1^N \mu(E_n) - \mu(E_{n-1}) = \mu(E_N)$$

and by countable additivity of μ, we have $\mu(\bigcup E_n) = \sum_n \mu(F_n)$, so letting $N \to +\infty$ we get what we wanted.

(d) Apply (c) to $E_1\backslash E_n$. □

Definition 4.5. *Let $(X, \mathcal{A}, \mu)$ be a measure space. It is called finite if $\mu(X) < \infty$ and σ-finite if there is a countable sequence E_n of subsets from $\mathcal{A}$ such that $X = \bigcup_n E_n$ and $\mu(E_n) < \infty$ for all n.*

Example $(\mathbb{R}^d, \mathcal{L}, m)$ is σ-finite but not finite.

Definition 4.6. *A measure space $(X, \mathcal{A}, \mu)$ is called a probability space and μ a probability measure if $\mu(X) = 1$.*

Proposition-Definition 4.7. *If $\mathcal{F}$ is a family of subsets of X, then the intersection of all σ-algebras containing $\mathcal{F}$ is a σ-algebra, called the σ-algebra generated by $\mathcal{F}$ and denoted by $\sigma(\mathcal{F})$.*

Proof. This is an easy check, see the Example sheet. □

Example

(1) If $X = \bigsqcup_1^N X_i$ is a finite partition of X and $\mathcal{F} = \{X_1, \ldots, X_N\}$, then $\sigma(\mathcal{F})$ is the family of all unions of X_i's. It is the Boolean algebra generated by $\mathcal{F}$.

(2) if X is a countable set, $\mathcal{F}$ the family of singletons (i.e. one-element subsets), then $\sigma(\mathcal{F}) = 2^X$ the discrete Boolean algebra on X.

Now comes a very important definition.

Definition 4.8. *Suppose that X is a topological space (i.e. a set endowed with a collection of "open sets" that forms a topology). The σ-algebra generated by all open subsets is called the Borel σ-algebra of X and is denoted by $\mathcal{B}(X)$. Its elements are called Borel sets.*

Remark 4.9. We've shown that every open set is Lebesgue measurable and that $\mathcal{L}$ is a σ-algebra, so this means that $\mathcal{B}(\mathbb{R}^d)$ is contained in $\mathcal{L}$, i.e. every Borel subset of $\mathbb{R}^d$ is Lebesgue measurable. It is not very difficult to prove (see a course in Logic) that there are Lebesgue measurable sets that are not Borel. In fact the cardinal of $\mathcal{B}(\mathbb{R}^d)$ is strictly smaller than the cardinal of $\mathcal{L}$. The reason behind it is that every subset of a null set is null and hence Lebesgue measurable, while there it is not necessarily Borel. This gives at least $2^{2^{\mathrm{Card}(\mathbb{R})}}$ Lebesgue measurable sets, but there are at most $2^{\mathrm{Card}\,\mathbb{R}}$ Borel sets.

Proposition 4.10. *Let $X = \mathbb{R}^d$, then $\mathcal{B}(X) \subset \mathcal{L}$ and moreover, every $A \in \mathcal{L}$ can be written as a disjoint union $A = B \sqcup N$, where $B \in \mathcal{B}(X)$ and N is a null set.*

Proof. We have already proved these facts in Proposition 3.13 above (note that every closed set is Borel and so is every countable union of closed sets). □

Remark 4.11. The σ-algebra of Borel sets of a topological space X is usually much larger than the family of constructible sets (i.e. the Boolean algebra generated by open sets). More generally, if $\mathcal{F}$ is any family of subsets of a set X, then the Boolean algebra $\beta(\mathcal{F})$ generated by $\mathcal{F}$ can be explicitly described: the elements of $\beta(\mathcal{F})$ are all finite unions of the subsets of the form

$$F_1 \cap \ldots \cap F_n$$

where for each $i = 1, \ldots, n$ either F_i or its complement F_i^c lies in $\mathcal{F}$ (see the Example Sheet). This is notoriously not so for $\sigma(\mathcal{F})$. The process of taking alternatively countable unions and countable intersections ad libitum does not stabilize in finitely many steps: this leads to the notion of *Borel hierarchy* and a full description of $\mathcal{B}(X)$ requires transfinite induction. See a Logic and Set Theory course.

Definition 4.12 (Borel measure)**.** *A Borel measure on a topological space X is a measure on Borel σ-algebra of X.*

Lecture 5

We now come to the construction of measures on σ-algebras. As we have seen in the construction of the Lebesgue measure, it is often easy to build a finitely additive measure on a natural Boolean algebra (e.g. the Jordan measure on elementary subsets of a box) and it is then a goal to extend this measure to the induced σ-algebra and hope to get this way a countably additive measure on a larger class of sets. This is what we did to build the Lebesgue measure on Lebesgue measurable sets. It turns out that this idea can be performed in exactly the same way in complete generality in the setting of abstract measure spaces.

Let X be a set, $\mathcal{B}$ a Boolean algebra of subsets of X and μ a finitely additive measure on $\mathcal{B}$ (i.e. a finitely additive non-negative set function defined on $\mathcal{B}$).

Definition 4.13. *Say that μ has the continuity property if for any non-increasing sequence of sets $E_n \in \mathcal{B}$ with empty intersection such that $\mu(E_1) < \infty$, we have*

$$\lim_{n \to +\infty} \mu(E_n) = 0.$$

We've already see (in Proposition 4.4) that if $\mathcal{B}$ is a σ-algebra and μ a genuine measure (i.e. countably additive) on $\mathcal{B}$, then μ has the continuity property. So the continuity property is a necessary condition on μ for it to ever admit a possible extension to a genuine measure on the σ-algebra generated by $\mathcal{B}$. The content of the following theorem is that it is actually also sufficient.

Theorem 4.14. *(Carathéodory extension theorem) Let X be a set, $\mathcal{B}$ a Boolean algebra of subsets of X and μ a σ-finite finitely additive measure on $\mathcal{B}$ with the continuity property. Then μ extends uniquely to a measure μ^* on the σ-algebra $\sigma(\mathcal{B})$ generated by $\mathcal{B}$.*

This is also sometimes called the Hahn-Kolmogorov extension theorem (but this attribution is probably not quite right, because Kolmogorov himself attributes it to Caratheodory in his book, and there is another Kolmogorov extension theorem having to do with defining probability measures on infinite stochastic processes, which is related but quite different).

Here the σ-finite condition on μ means that there is a countable family $(X_n)_n$ of subsets of X such that $\bigcup_n X_n = X$ and for all n, $X_n \in \mathcal{B}$ and $\mu(X_n) < \infty$. Even though we will use it in the proof (and in all interesting examples I know it holds), the σ-finiteness assumption can be dropped for the existence part but is essential to the uniqueness part.

The construction mimics word-by-word the construction we have made for the Lebesgue measure. In particular we define

Definition 4.15. *the outer-measure μ^* of an arbitrary subset E of X by*

$$\mu^*(E) = \inf\{\sum_i \mu(B_i); E \subset \bigcup_i B_i, B_i \in \mathcal{B}\}$$

where the family $\{B_i\}$ is countable.

And we

Definition 4.16. *say that a subset $E \subset X$ is μ^*-measurable if for every $\varepsilon > 0$ it is contained in a countable union $C := \bigcup_n B_n$ of sets from $\mathcal{B}$ such that $\mu^*(C \backslash E) < \varepsilon$.*

The existence part of Caratheodory's theorem then follows from

Proposition 4.17. *Under the assumptions of Theorem 4.14, the family $\mathcal{B}^*$ of μ^* measurable subsets of X is a σ-algebra containing $\mathcal{B}$ (called the completion of $\mathcal{B}$ with respect to μ). The outer-measure μ^* is countably additive on $\mathcal{B}^*$ and coincides with μ on $\mathcal{B}$.*

We'll prove the uniqueness part of Theorem 4.14 next time as a consequence of Dynkin's lemma.

Proof. As it turns out we have already proven Proposition 4.17, because the proof we gave of Proposition 3.5 for the existence of the Lebesgue measure, works verbatim in our generalized setting. In fact the definitions we have given today were geared to make all previous arguments work in this abstract setting. This is the power of the axiomatic method! That said, it is still a good exercise to check this by yourselves. You will have to replace the words "boxes" and "elementary set" by the word "element of $\mathcal{B}$". At some point we considered bounded subsets of $\mathbb{R}^d$: replace this notion by the word "contained in X_n for some n", where X_n is any family of sets in $\mathcal{B}$ such that $\mu(X_n) < \infty$ and $X = \bigcup_n X_n$. At some point we also used the Heine-Borel property, but this was only to establish the continuity property for the Jordan measure, which we assume here. Everything else works verbatim. □

Note that $\mathcal{B}^*$ contains all null sets (i.e. sets with zero μ^*-measure). In the case of $\mathbb{R}^d$ and the Boolean algebra generated by elementary sets, $\mathcal{B}^*$ coincides with $\mathcal{L}$, while $\sigma(\mathcal{B})$ is the Borel σ-algebra.

Of course the main and defining example of use of Caratheodory's extension theorem is the construction of the Lebesgue measure from the Jordan measure, but we will see several more examples in this course, in probability theory in particular.

A side remark: Paul Halmos in his well-known book and James Norris in his lecture notes take a slightly different route to define μ^* and the notion of μ^*-measurability, which is closer to Lebesgue's original definition. A set is said to be μ^*-measurable if the sum of its outer measure and that of its complement equals $\mu(X)$ (assuming this is finite). These two approaches are equivalent (see the Example sheet). Ours sticks more closely to the intuitive idea that the measure of a set is given by the smallest number of cubes needed to cover it.

Another side remark: although Borel measures on abstract topological spaces may at first sight look much more complicated and rich than the good old interval endowed with Lebesgue measure, this is not so. It can be shown that if X is a compact metric space and μ a probability measure on its Borel σ-algebra $\mathcal{B}$ giving mass zero to each point, then there is a measure preserving (measurable) isomorphism between $(X, \mathcal{B}^*, \mu)$ and $([0,1], \mathcal{L}, m)$.

5. Uniqueness of measures

We now discuss π-systems and the problem of uniqueness of measures.

Definition 5.1. *Let X be any set. A family $\mathcal{F}$ of subsets of X is called a π-system if it*

(1) contains the empty set, and
(2) is stable under finite intersections.

So this is a weaker notion than being a Boolean algebra. The reason for introducing it is the following proposition and lemma that help in proving that two measures are the same: it is enough to check that they are the same on a π-system generating the σ-algebra.

Proposition 5.2. *(measure uniqueness) Let $(X, \mathcal{A})$ be a measurable space and μ_1, μ_2 be two finite measures on X such that $\mu_1(F) = \mu_2(F)$ for all $F \in \mathcal{F} \cup \{X\}$, where $\mathcal{F}$ is a π-system such that $\sigma(\mathcal{F}) = \mathcal{A}$. Then $\mu_1 = \mu_2$.*

For the proof we will require:

Lemma 5.3. *(Dynkin's lemma) If $\mathcal{F}$ is a π-system and $\mathcal{F} \subset \mathcal{C}$, where $\mathcal{C}$ is a family of subsets of X, which is stable under complementation and disjoint countable union, then $\sigma(\mathcal{F}) \subset \mathcal{C}$.*

Proof of Proposition 5.2. Let $\mathcal{C} = \{A \in \mathcal{A}, \mu_1(A) = \mu_2(A)\}$. Note that $\mathcal{C}$ is stable under complementation, because $\mu_1(X \backslash A) = \mu_1(X) = \mu_1(A)$, and also stable under disjoint countable unions by σ-additivity of both measures. By Dynkin's lemma, we conclude that $\sigma(\mathcal{F}) \subset \mathcal{C}$, so $\mathcal{C} = \mathcal{A}$. □

Remark 5.4. While the conclusion may fail in general if μ_1 and μ_2 are infinite measures, it very easy to see that it continues to hold under the following mild additional assumption: that there is a countable family of subsets $F_n \in \mathcal{F}$ each of finite measure and such that $X = \bigcup_n F_n$.

Proof of Dynkin's lemma. Let $\mathcal{M}$ be the smallest family of subsets of X containing $\mathcal{F}$ and stable under complementation and disjoint countable union (note that such an $\mathcal{M}$ exists, it is the intersection of all such families). We need to show that $\mathcal{M}$ is a Boolean algebra (note that this will clearly imply that $\mathcal{M}$ is a σ-algebra, because as we have already seen any countable union of sets from a Boolean algebra can be written as a disjoint countable union of such sets).

So let

$$\mathcal{M}' := \{A \in \mathcal{M}, A \cap B \in \mathcal{M} \ \forall B \in \mathcal{F}\}.$$

Then $\mathcal{M}'$ again is stable under complementation and disjoint countable union (note that $A^c \cap B \in \mathcal{M}$ because it is $(B^c \sqcup (A \cap B))^c$). And clearly, by definition of $\mathcal{M}'$, since $\mathcal{F}$ is a π-system and thus stable under intersection, we have $\mathcal{F} \subset \mathcal{M}'$. Now by minimality of $\mathcal{M}$ we conclude that $\mathcal{M}' = \mathcal{M}$.

Similarly set

$$\mathcal{M}'' = \{A \in \mathcal{M}, A \cap B \in \mathcal{M} \ \forall B \in \mathcal{M}\}.$$

Then again $\mathcal{M}''$ is stable under complementation and disjoint countable union, so by minimality $\mathcal{M}'' = \mathcal{M}$. So $\mathcal{M}$ is a Boolean algebra, and hence a σ-algebra as desired. □

Lecture 6

A simple consequence is:

Proposition 5.5. *Lebesgue measure is the unique translation invariant measure m on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that $m([0,1]^d) = 1$.*

Recall that $\mathcal{B}(\mathbb{R}^d)$ denotes the σ-algebra of Borel subsets of $\mathbb{R}^d$. And *translation invariant* means that $m(A + x) = m(A)$ for all $A \in \mathcal{B}(\mathbb{R}^d)$ and all $x \in \mathbb{R}^d$.

Proof. The fact that m is translation invariant is clear, because the outer measure m^* is obviously translation invariant. So we have to prove the uniqueness. Let μ be a measure as in the statement of the proposition. Let $\mathcal{F}$ be the family of all boxes in $\mathbb{R}^d$. Note that $\mathcal{F}$ is a π-system made of Borel sets such that $\sigma(\mathcal{F}) = \mathcal{B}(\mathbb{R}^d)$ (because every open set is a countable union of boxes). So by the previous proposition (and Remark 5.4) it is enough to show that $\mu = m$ on $\mathcal{F}$.

In fact it is enough to check that $\mu = m$ on $\mathcal{F}_d$, the family of *dyadic boxes* (i.e. boxes with side lengths of the form $\frac{k}{2^n}$, $k, n \in \mathbb{Z}$), because of upwards monotone convergence (every box is an increasing union of dyadic boxes).

Also, given a coordinate hyperplane H (i.e. $H = \{x \in \mathbb{R}^d, x_i = 0\}$ for some i) in $\mathbb{R}^d$ we have $\mu(H) = 0$ by translation invariance (otherwise some cube, a translate of $[0,1]^d$, would intersect H in a set of positive measure, and by translation invariance we could pack infinitely many translates of this intersection inside the same cube contradicting the finiteness of $\mu([0,1]^d)$).

Now we can write $[0,1]^d$ as a union of 2^{nd} translates of the dyadic box $[0, \frac{1}{2^n}]^d$, each having the same measure (by translation invariance, because the boundaries are contained in hyperplanes of measure zero). So by additivity $\mu([0, \frac{1}{2^n}]^d) = 2^{-nd}$. And since any dyadic box is an almost disjoint (i.e. with overlaps confined to hyperplanes, hence of measure zero) finite union of translates of such small cubes, by translation invariance again we get that $\mu(B) = m(B)$ for all dyadic boxes B as desired. □

We end with two remarks:

Remark 5.6. There are no countably additive translation invariant measure μ defined on the family of all subsets of $\mathbb{R}$ and such that $0 < \mu([0,1]) < \infty$ (because Vitali's counter-example would lead to a contradiction in exactly the same way as we have discussed already).

Remark 5.7. However (assuming the Axiom of Choice) there are finitely additive ones (one says that $\mathbb{R}/\mathbb{Z}$ is a discrete *amenable group*), but this requires some functional analysis (e.g. the Markov-Kakutani fixed point theorem).

6. Measurable functions

Definition 6.1. *Let $(X, \mathcal{A})$ be a measurable space. A function $f : X \to \mathbb{R}$ is said to be measurable with respect to $\mathcal{A}$ if for all $t \in \mathbb{R}$*

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

Following this definition, we make two initial remarks. The first is that if f is measurable, then the pre-image $f^{-1}(B)$ of any Borel subset $B \in \mathcal{B}(\mathbb{R})$ belongs to $\mathcal{A}$. This is clear, because on the one hand the family of all such subsets B is a σ-algebra and on the other hand the family of all intervals $(-\infty, t)$ for $t \in \mathbb{R}$, generates the σ-algebra of Borel subsets of $\mathbb{R}$.

The second remark is that it is sometimes convenient to extend the notion of measurability to functions that can take the value $+\infty$ or $-\infty$. In that case we say

that f is measurable (w.r.t $\mathcal{A}$) if additionally the sets $\{x \in X, f(x) = +\infty\}$ and $\{x \in X, f(x) = -\infty\}$ are in $\mathcal{A}$.

More generally we can define the notion of measurable map between any two measurable spaces.

Definition 6.2. *A map $f : X \to Y$ between two measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ is called measurable if $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$.*

To have a better feel for the notion of measurability of a map, let us give some examples:

Examples:

(1) every continuous function $f : \mathbb{R}^d \to \mathbb{R}$ is measurable: indeed $\{x \in R^d, f(x) < t\}$ is an open set.
(2) if $(X, \mathcal{A})$ is a measurable space and $E \subset X$ a subset, then $E \in \mathbb{A}$ if and only if the indicator function 1_E is $\mathcal{A}$-measurable.
(3) if $X = \sqcup_1^N X_i$ is a finite partition of a set X with non-empty pieces. Let $\mathcal{A}$ be the Boolean algebra generated by the pieces X_i's (so its elements are the unions of pieces). Then a function $f : X \to \mathbb{R}$ is measurable (with respect to $\mathcal{A}$) if and only if it is constant on each X_i. In this case we see that the set of all measurable functions on $(X, \mathcal{A})$ forms a real vector space of dimension N.

As these examples demonstrate, the notion of measurability of a function on X is very sensitive to the choice of σ-algebra $\mathcal{A}$. Being measurable with respect to $\mathcal{A}$ means, roughly speaking, that the value of the function at a point x depends only on the family of sets from $\mathcal{A}$ that contain x.

In Analysis, mathematicians often work with a single σ-algebra: the Borel σ-algebra (or its completion, the Lebesgue measurable sets) but consider various measures on this space. In Probability the opposite is true: the measure is given, while the σ-algebra may vary a lot. In Information Theory the sigma algebra can be interpreted as the precision at which one can understand a given function or signal, the coarser the subalgebra is (i.e. the fewer subsets of X it contains), the less information one can retrieve from a function measurable w.r.t this subalgebra.

The class of measurable functions is very handy and stable under several basic operations (a much wider range of operations than say for the class of piecewise continuous functions on $\mathbb{R}$):

Proposition 6.3. *(a) given three measurable spaces, the composition $f \circ g : X \to Z$ of two measurable maps $f : X \to Y$ and $g : Y \to Z$ is measurable.*
(b) the family of measurable functions on a measurable space $(X, \mathcal{A})$ forms an $\mathbb{R}$-algebra: namely if f, g are two such functions then so is $f + g$, fg and λf for any scalar $\lambda \in \mathbb{R}$.
(c) If $(f_n)_{n \geqslant 1}$ is a sequence of measurable functions on $(X, \mathcal{A})$, then $\limsup f_n$, $\liminf f_n$, $\inf f_n$, $\sup f_n$ are also measurable.

Proof. (a) is clear, (b) follows from (a) once it is shown that the maps

$$\mathbb{R}^2 \to \mathbb{R}$$
$$(x, y) \mapsto x + y$$

and

$$\mathbb{R}^2 \to \mathbb{R}$$
$$(x, y) \mapsto xy$$

are measurable (for the Borel σ-algebra on $\mathbb{R}^2$). This is the case, because the sets $\{(x, y) \in \mathbb{R}^2, x + y < t\}$ and $\{(x, y) \in \mathbb{R}^2, xy < t\}$ are both open, hence Borel.

(c) This follows from the following translations:

(i) $\inf f_n(x) < t \iff x \in \bigcup_n \{f_n(x) < t\}$ and $\inf f_n(x) = -\infty \iff x \in \bigcap_{k\geqslant 1}\bigcup_n \{x, f_n(x) < -k\}$.
(ii) $\sup f_n(x) < t \iff x \in \bigcup_{m\geqslant 1}\bigcap_n \{f_n(x) < t - \frac{1}{m}\}$
(iii) $\liminf f_n(x) < t \iff x \in \bigcup_{m\geqslant 1}\bigcap_k \bigcup_{n\geqslant k} \{x, f_n(x) < t - \frac{1}{m}\}$
(iv) $\limsup f_n(x) < t \iff x \in \bigcup_{m\geqslant 1}\bigcup_k \bigcap_{n\geqslant k} \{x, f_n(x) < t - \frac{1}{m}\}$

□

Proposition 6.4. *Let $(X, \mathcal{A})$ be a measurable space and $f : (X, \mathcal{A}) \to \mathbb{R}^d$ a map. Then f is $\mathcal{A}$-measurable if and only if each f_i is $\mathcal{A}$-measurable, where $f = (f_1, \dots, f_d)$.*

Proof. ($\Rightarrow$) note that $\{x \in X, f_i(x) < t\} = f^{-1}(\{y \in \mathbb{R}^d, y_i < t\})$ for each i. So if f is $\mathcal{A}$-measurable, so is each f_i.

($\Leftarrow$) conversely if f_i is measurable for all i, then $f^{-1}(\prod[a_i, b_i]) = \bigcap_1^d \{x \in X, a_i \leqslant f_i(x) \leqslant b_i\}$ is in $\mathcal{A}$. But boxes $\prod_1^d [a_i, b_i]$ generate the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$. So $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}(\mathbb{R}^d)$.

□

Definition 6.5. *If X is a topological space, a function $f : X \to \mathbb{R}$ is called Borel measurable (or simply Borel) if it is measurable with respect to the Borel σ-algebra $\mathcal{B}(X)$.*

Remark 6.6.
(i) the preimage of a Lebesgue measurable subset of $\mathbb{R}$ by a measurable function need not be measurable (this is due to the wealth of null sets, cf. 2nd example sheet)
(ii) the image of a measurable set under a measurable map need not be measurable (e.g. if the target space is endowed with the trivial σ-algebra (the one made of only $\varnothing$ and the whole set) then every map is measurable).
(iii) worse: the image of a Borel set by a continuous function need not be measurable. In fact there exists a Borel subset $B \subset \mathbb{R}^2$ such that $\pi_1(B) \subset \mathbb{R}$ is not Lebesgue measurable, where $\pi_1(x, y) = x$ is the projection onto the first coordinate (this was Lebesgue's famous mistake!).

7. Integration

We now begin the construction of the integral. As we will see, we are going to be able to integrate any non-negative measurable function defined on any measure space $(X, \mathcal{A}, \mu)$.

Definition 7.1. *A simple function on a measure space $(X, \mathcal{A}, \mu)$ is a function of the form $\sum_1^N a_i 1_{A_i}$, with $a_i \geqslant 0$, $A_i \in \mathcal{A}$ for each $i = 1, \dots, N$.*

Note that a simple function is non-negative and measurable. Equivalently, it is straightforward to verify that a simple function on $(X, \mathcal{A})$ is a measurable function admitting only finitely many values, all non-negative.

Example 7.2. *Let I be a countable set endowed with the discrete σ-algebra (i.e. all subsets of I) and a measure μ. Then a simple function is just a non-negative function on I admitting finitely many values. Moreover $\mu(f) = \sum_{i\in I} f(i)\mu(\{i\})$.*

Lemma 7.3. *If a simple function f has two representations $f = \sum_1^N a_i 1_{A_i} = \sum_1^M b_j 1_{B_j}$, then*

$$\sum_1^N a_i \mu(A_i) = \sum_1^M b_j \mu(B_j).$$

(We have taken the fonction $0 \cdot \infty = 0$ in case some A_i or B_j has infinite measure and the corresponding coefficient is zero.)

Proof. We omit the proof, this is an exercise in the first Example sheet. □

We can now *define the integral* of the simple function f with respect to μ by

$$\mu(f) := \sum_1^N a_i \mu(A_i).$$

We will also use another standard notation for the same quantity:

$$\int_X f d\mu$$

which is closer to the original leibnizian notation for the integral. Note that $\mu(f) \in [0, +\infty]$.

Having defined $\mu(f)$ for (non-negative) simple functions, we are ready to extend the definition to all non-negative measurable functions. We do this as follows: given a non-negative measurable function (i.e. $\forall x \in X,\ f(x) \geqslant 0$, we abbreviate this as $f \geqslant 0$), we define

$$\mu(f) := \sup\{\mu(g), g \leqslant f, g \text{ simple}\}. \tag{7.1}$$

Lecture 7

This is consistent with the case when f is simple, by (i) of the following

Proposition 7.4 (Positivity of the integral)**.** *Suppose f, g are non-negative measurable functions on $(X, \mathcal{A}, \mu)$.*

(i) (positivity) $f \geqslant g$ implies $\mu(f) \geqslant \mu(g)$,
(ii) (equality case) if $f \geqslant g$ and $\mu(f) = \mu(g)$ is finite, then $f = g$ almost everywhere.

To say that $f = g$ almost everywhere means that $\{x \in X, f(x) \neq g(x)\}$ has μ-measure 0, i.e. is a null set (note that this set belongs to $\mathcal{A}$ because f is measurable). It is often abbreviated as $f = g$ a.e., or equivalently for μ-a.e. $x \in X$, $f(x) = g(x)$.

Proof. First we verify that both items hold when f and g are assume to be simple functions. This is pretty obvious given the definition and Lemma 7.3: just note that $f - g$ is again a simple function and that $\mu(f - g) = \mu(f) - \mu(g)$. Then (i) is immediate for any non-negative measurable functions by (7.1). Let's prove (ii). If $A_n := \{x \in X, f(x) - g(x) > \frac{1}{n}\}$, then $f - g \geqslant \frac{1}{n} 1_{A_n}$ pointwise. In particular $\mu(f - g) \geqslant \frac{1}{n}\mu(A_n)$ by (i). However by (7.1) we have immediately $\mu(f) \geqslant \mu(g) + \mu(f - g)$ (we will see shortly below that there is in fact always equality, but this half is obvious at this stage). Since we assumed that $\mu(f) = \mu(g)$ we get $\mu(f - g) = 0$ and so we get $\mu(A_n) = 0$, and by subadditivity $\mu(\{x, f(x) > g(x)\}) \leqslant \sum_n \mu(A_n) = 0$. □

Note that the converse is clear: if f, g are non-negative measurable functions on $(X, \mathcal{A})$ such that $f = g$ μ-almost everywhere, then $\mu(f) = \mu(g)$. Indeed if $E = \{x \in X, f(x) = g(x)\}$, then $\mu(E^c) = 0$ and thus for every simple function h we will have $\mu(h) = \mu(h1_E)$ (by Lemma 7.3). In particular if $h \leqslant f$, then $h1_E \leqslant g$ and thus $\mu(h) = \mu(h1_E) \leqslant \mu(g)$, which yields $\mu(f) \leqslant \mu(g)$ by (7.1). By symmetry $\mu(f) = \mu(g)$.

Example 7.5. *(i) In the previous example, $\mu(f) = \sum_{i \in I} f(i)\mu(\{i\})$, and this holds also for every non-negative function (note that all functions are measurable, because the σ-algebra is the discrete one).*

(ii) when $(X, \mathcal{A}, \mu)$ is $(\mathbb{R}, \mathcal{L}, m)$, and f is a Lebesgue measurable function, then $m(f)$ is called the Lebesgue integral *of f. It coincides with the Riemann integral of f in case f is assumed Riemann integrable (cf. Example sheet).*

We defined the integral by means of simple functions. It is often useful to be able to approximate any non-negative measurable function by simple functions as follows:

Lemma 7.6. *Let $f \geqslant 0$ be a measurable function on the measure space $(X, \mathcal{A}, \mu)$. Then there is a sequence of simple functions g_n with $g_n \leqslant g_{n+1}$ such that $g_n \to f$ pointwise (i.e. $\forall x \in X, g_n(x) \to f(x)$).*

Proof. One can take, for example, $g_n = \frac{1}{2^n}\lfloor 2^n \min\{f(x), n\}\rfloor$, where $\lfloor x \rfloor$ denotes the largest integer less or equal to x (to see that $g_{n+1} \geqslant g_n$ note that $\lfloor 2y \rfloor \geqslant 2\lfloor y \rfloor$ for all $y \geqslant 0$). □

We now move on to the main result regarding Lebesgue's integration, namely Lebesgue's Monotone Convergence Theorem. This will be the key result, which will imply the next two important statements: Fatou's lemma and Lebesgue's Dominated Convergence Theorem. Together these three facts make Lebesgue's integration theory much more powerful and versatile than Riemann's. The scope and

generality in which these results hold (i.e. on abritrary measure spaces) make them ubiquitous in mathematics.

Theorem 7.7 (Monotone Convergence Theorem, MCT)**.** *Let $f_n \geqslant 0$ be a sequence of non-negative measurable functions on a measure space $(X, \mathcal{A}, \mu)$ such that*

$$0 \leqslant f_1 \leqslant f_2 \leqslant \ldots \leqslant f_n \leqslant \ldots$$

Let $f(x) = \lim_n f_n(x) \in [0, +\infty]$ for each $x \in X$. Then

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

Lemma 7.8. *If g is simple, the map*

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

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

Proof. We need to check σ-additivity of m_g. So let $E = \bigsqcup_n E_n$ a disjoint countable union of sets from $\mathcal{A}$. Then we have $\mu(1_E g) = \sum_1^N a_i \mu(E \cap A_i)$ if $g = \sum a_i 1_{A_i}$. But $E \mapsto \mu(E \cap A_i)$ is σ-additive for each i, since μ is, hence so is m_g. □

Proof of the MCT. We have $f_n \leqslant f_{n+1} \leqslant f$, so $\mu(f_n) \leqslant \mu(f_{n+1}) \leqslant \mu(f)$ by (i) in Proposition 7.4. Hence $\lim_{n \to +\infty} \mu(f_n)$ exists and is $\leqslant \mu(f)$.

We will now show the reverse inequality. To this end let g be a simple function with $g \leqslant f$. Pick $\varepsilon \in (0, 1)$ and let $E_n = \{x \in X, f_n(x) \geqslant (1 - \varepsilon) g(x)\}$. Then $X = \bigcup_n E_n$ and $E_n \subset E_{n+1})$. So we may apply upwards monotone convergence for sets to the measure m_g on $(X, \mathcal{A})$ and thus get:

$$\lim_{n \to +\infty} m_g(E_n) = m_g(X) = \mu(g)$$

but

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

where the last inequality follows from (i) of Proposition 7.4. We conclude that

$$(1 - \varepsilon)\mu(g) \leqslant \lim_{n \to +\infty} \mu(f_n)$$

holds for every simple function $g \leqslant f$ and for all $\varepsilon \in (0, 1)$. Letting ε tend to 0 we deduce that $\mu(f) \leqslant \lim_{n \to +\infty} \mu(f_n)$ as desired. □

We are now ready for

Lemma 7.9 (Fatou's lemma)**.** *Let $f_n \geqslant 0$ be a sequence of non-negative measurable functions (on a measure space $(X, \mathcal{A}, \mu)$). Then*

$$\mu(\liminf_{n \to +\infty} f_n) \leqslant \liminf_{n \to +\infty} \mu(f_n).$$

Remark 7.10. Strict inequality can occur. For example when $(X, \mathcal{A}, \mu) = (\mathbb{R}, \mathcal{L}, m)$ and we consider either of the following "moving bump" example: (i) $f_n = 1_{[n, n+1]}$, (ii) $f_n = \frac{1}{n} 1_{[0, n]}$, (iii) $f_n = n 1_{[\frac{1}{n}, \frac{2}{n}]}$, where each time $f_n \to 0$ pointwise, while $\mu(f_n) = 1$.

Proof of Fatou's lemma. Let $g_n := \inf_{k \geqslant n} f_k$ and $g = \liminf_{n \to +\infty} f_n$. Then $g_{n+1} \geqslant g_n \geqslant 0$ and $g_n \to g$ pointwise, so by the Monotone Convergence Theorem, we have $\mu(g_n) \to \mu(g)$. But $g_n \leqslant f_n$, so $\mu(g_n) \leqslant \mu(f_n)$ by positivity (Proposition 7.4). And thus: $\mu(g) \leqslant \liminf_{n \to +\infty} \mu(f_n)$. □

So far we have defined the integral for non-negative measurable functions only. In order to extend this definition to functions that can change sign, for any measurable function $f : X \to \mathbb{R}$, we set $f^+ := \max\{0, f\}$ and $f^- := (-f)^+$. Note then that $|f| = f^+ + f^-$ and $f = f^+ - f^-$. Moreover f^+, f^- and $|f|$ are clearly measurable.

Definition 7.11. *A measurable function $f : (X, \mathcal{A}) \to \mathbb{R}$ is said to be μ-integrable if $\mu(|f|) < \infty$. In this case its integral is defined by $\mu(f) := \mu(f^+) - \mu(f^-)$.*

So integrable functions as those measurable functions whose absolute value has finite integral. Beware that $f \geqslant 0$ can be measurable without being integrable, even though $\mu(f)$ is well-defined (and equals $+\infty$).

Proposition 7.12 (Linearity of the integral)**.** *Let f, g be measurable functions on $(X, \mathcal{A}, \mu)$ and let $\alpha, \beta \in \mathbb{R}$. If f, g are assumed integrable, then $\alpha f + \beta g$ is integrable and*

$$\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g). \tag{7.2}$$

Moreover (7.2) also holds if we replace the integrability assumption on f, g by the assumption that f, g and α, β are non-negative.

Proof. Write $f = f^+ - f^-$, $\alpha = \alpha^+ - \alpha^-$ and $\beta = \beta^+ - \beta^-$ and expand $\alpha f + \beta g$. It then becomes clear that the proof reduces to proving (7.2) in the case when f, g, α, β are all non-negative. That $\mu(\alpha f) = \alpha\mu(f)$ is obvious from the definition of the integral (by means of simple functions). So only $\mu(f + g) = \mu(f) + \mu(g)$ needs to be shown. This is clearly true for simple functions (by the previous lemma that any two different writings of a simple function give rise to the same value of their integral). The general case follows by the Monotone Convergence Theorem by approximating f and g by simple functions (Lemma 7.6). □

We are now ready for the following important theorem.

Theorem 7.13 (Lebesgue's Dominated Convergence Theorem)**.** *Let f and $(f_n)_{n\geqslant 1}$ be measurable functions on a measure space $(X, \mathcal{A}, \mu)$. Assume that there exists an integrable function g such that for all $x \in X$*

(i) $|f_n(x)| \leqslant g(x)$ *for all* $n \geqslant 1$,
(ii) $\lim_{n\to+\infty} f_n(x) = f(x)$.

Then $\lim_{n\to+\infty} \mu(f_n) = \mu(f)$ *and f is integrable.*

In this result lies the main advantage of Lebesgue's integration versus Riemann's: it allows to move the integral sign passed the limits and exchange the two. So it is very powerful a tool.

Proof. Since $|f_n| \leqslant g$ for all n, passing to the limit we get $|f| \leqslant g$. It follows that f is integrable and that $g + f_n \geqslant 0$ for all n. Then Fatou's lemma applies and yields

$$\mu(g) + \mu(f) = \mu(g + f) = \mu(\liminf_{n\to+\infty} g + f_n) \leqslant \liminf_{n\to+\infty} \mu(g + f_n) = \mu(g) + \liminf_{n\to+\infty} \mu(f_n)$$

where we have also used the linearity of the integral. But $\mu(g) < \infty$, so

$$\mu(f) \leqslant \liminf_{n\to+\infty} \mu(f_n).$$

What we have just done for f_n, we can do for $-f_n$. And this will give us:

$$-\mu(f) \leqslant -\limsup_{n\to+\infty} \mu(f_n).$$

This ends the proof. □

Lecture 8

We can upgrade slightly all three main statements (MCT, Fatou's lemma, Lebesgue's DCT) as follows. We have assumed that the limits (or liminf) hold pointwise, namely for each $x \in X$. It turns out that the conclusion of all three results continue to hold under the weaker assumption that the limits hold μ-almost everywhere. For example in the MCT, we can assume that the assumptions $f_n \geqslant 0$ and $f_{n+1} \geqslant f_n$ only hold μ-almost everywhere, or in Lebesgue's DCT, we can assume that each of the two assumptions $f(x) = \lim_n f_n(x)$ and $|f(x)| \leqslant g(x)$ only hold for μ-almost every x:

Proof. For the MCT for example we can set $E_n = \{x \in X, f_n(x) \geqslant 0\}$, $F_n = \{x \in X, f_{n+1}(x) \geqslant f_n(x)\}$. Then the complement of $E = \bigcap E_n \cap F_n$ is a μ-null set and we can set $f_n'(x) = 1_E f_n(x)$ and now apply the original MCT to f_n' to conclude that $\mu(1_E f) = \lim_n \mu(1_E f_n)$. But it follows from the definition of the integral (7.1) that $\mu(f 1_E) = \mu(f)$ and $\mu(f_n 1_E) = \mu(f_n)$ for all n. Proceed similarly for Fatou's lemma and the DCT. □

We now pass to two important corollaries of Lebesgue's Dominated Convergence Theorem.

Corollary 7.14 (Exchange of $\int$ and $\sum$ signs)**.** *Let $(X, \mathcal{A}, \mu)$ be a measure space and $(f_n)_{n\geqslant 1}$ a sequence of measurable functions.*

(i) if $f_n \geqslant 0$ for all n, then

$$\mu(\sum_{n\geqslant 1} f_n) = \sum_{n\geqslant 1} \mu(f_n)$$

(ii) if $\sum_{n\geqslant 1} |f_n|$ is μ-integrable (i.e. $\mu(\sum_{n\geqslant 1} |f_n|) < \infty$), then so is $\sum_{n\geqslant 1} f_n$, and

$$\mu(\sum_{n\geqslant 1} f_n) = \sum_{n\geqslant 1} \mu(f_n).$$

Proof. (i) Let $g_N = \sum_1^N f_n$, then g_N is a non-decreasing sequence of non-negative functions, so we can simply appy the Monotone Convergence Theorem to reach the deired conclusion.

(ii) Let $g = \sum_{n\geqslant 1} |f_n|$. Then $|g_N| \leqslant g$ for all N. So by Dominated Convergence, we get $\mu(g_N) \to_{N\to+\infty} \mu(\lim g_N)$. □

Corollary 7.15 (Differentiation under the $\int$ sign)**.** *Let $(X, \mathcal{A}, \mu)$ be a measure space and let $U \subset \mathbb{R}$ be an open set. Let $f : U \times X \to \mathbb{R}$ be such that*

(i) $x \mapsto f(t, x)$ is μ-integrable for every $t \in U$,
(ii) $t \mapsto f(t, x)$ is differentiable for every $x \in X$,
(iii) (domination) $\exists g : X \to \mathbb{R}$ a μ-integrable function such that for all $t \in U$, and all $x \in X$

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

Then $x \mapsto \frac{\partial f}{\partial t}(t, x)$ is μ-integrable for all $t \in U$. Moreover, setting

$$F(t) := \int_X f(x, t) d\mu(x)$$

F is differentiable on U, and

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

Proof. Let h_n a sequence of positive real numbers such that $\lim_n h_n = 0$. Set

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

which given $t \in U$ is defined as soon as $t + h_n \in U$, and in particular for all large enough n. By the Mean Value Theorem there exists $\theta_{t,x,n} \in [t, t+h_n]$ such that

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

Hence for each t, for all large enough n and for all $x \in X$,

$$|g_n(t,x)| \leqslant g(x).$$

Moreover for all $t \in U$ and $x \in X$, $\lim_n g_n(t,x) = \frac{\partial f}{\partial t}(t,x)$. But

$$\mu(g_n(t,x)) = \int_X g_n(t,x)d\mu(x) = \frac{1}{h_n}[F(t+h_n) - F(t)],$$

so by Dominated Convergence, we get

$$\lim_n \frac{1}{h_n}[F(t+h_n) - F(t)] = \int_X \frac{\partial f}{\partial t}(t,x)d\mu(x).$$

So F is differentiable at t and its derivative $F'(t)$ is equal to the right hand side. □

You will see several examples of use of these theorems in the Example sheets.

To end this section we make (without proof) several remarks that can be skipped in first reading:

Remark 7.16. If $f : [a,b] \to \mathbb{R}$ is continuous ($a < b$ real numbers), then it is integrable with respect to the Lebesgue measure m, and moreover

$$m(f) = \int_a^b f(x)dx$$

where the latter is ordinary Riemann integral of f. More generally one can show that if f is only assumed to be a bounded function, then f is Riemann integrable if and only if the set of points where f is not continuous has Lebesgue measure zero (see the Example sheet).

Remark 7.17 (Fundamental Theorem of Calculus)**.** We've learned in a basic course on differential calculus the Fundamental Theorem of Calculus, according to which a function $f : [a,b] \to \mathbb{R}$ assumed to be differentiable with continuous derivative (that is a C^1 function) satisfies

$$f(b) - f(a) = \int_a^b f'(x)dx. \tag{7.3}$$

What happens if we relax the hypotheses a bit? Well it depends and it is an interesting chapter of Analysis to determine under what conditions the Fundamental Theorem of Calculus still holds. It turns out that it is enough to assume that f is differentiable and f' is integrable w.r.t. Lebesgue measure (see Rudin's book). It is also enough to assume that f is a Lipschitz function: in that case it can be shown that f differentiable almost everywhere, its derivative is integrable and (7.3) holds. On the other hand it is not too difficult to construct examples of non-descreasing and non-constant continuous functions such that $f'(x) = 0$ almost everywhere for Lebesgue measure, so that clearly (7.3) fails (see the "Devil's staircase cosntructions" as in Ex 12 in the 1st Example sheet). More about this, and in particular a proof of the main result in this area, the Lebesgue Density Theorem (whose proof bares a lot of resemblance to the that of the pointwise ergodic theorem) is given in the Lent term D-course "Analysis of Functions".

Remark 7.18 (invariance of Lebesgue measure under affine maps)**.** We have shown that Lebesgue measure is invariant under translation. It also behaves very well under linear transformations. Namely if $g \in \mathrm{GL}_d(\mathbb{R})$ and $f \geqslant 0$ is a non-negative measurable function on $(\mathbb{R}^d, \mathcal{L}, m)$, then

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

In particular m is invariant by rotation on $\mathbb{R}^d$ (see the Example sheet).

More generally, one can use the previous remark to establish the following useful:

Change of variables formula: *Assume that U, V are open subsets of $\mathbb{R}^d$ and that $\phi : U \to V$ is a C^1-diffeomorphism, then*

$$\int_U f(\phi(x)) J_\phi(x) dx = \int_V f(x) dx$$

where dx is short for $dm(x)$ the Lebesgue measure, $f \geqslant 0$ is an arbitrary Borel measurable map on V and $J_\phi(x) := |\det d\phi(x)|$ is the Jacobian of ϕ, i.e. the absolute value of the determinant of its differential.

The details of the (slightly boring) proof can be found in Rudin's book as well as in many other textbooks. The same formula clearly holds (looking at positive and negative parts of f) if we change the assumption $f \geqslant 0$ and assume instead that f integrable on V with respect to Lebesgue.

Next we comment on a the relation between measure theory and linear functionals, which is at the basis of Bourbaki's approach to integration:

Remark 7.19 (Riesz Representation theorem)**.** Let X be a topological space, which is locally compact (i.e. every point has a compact neighborhood) and second countable (i.e. there is a countable basis of open sets for the topology). A *Radon measure* on X is a Borel measure μ, which is finite on every compact subset of X. If μ is a Radon measure, then the map:

$$C_c(X) \to \mathbb{R}$$
$$f \mapsto \mu(f) = \int_X f d\mu$$

defines a linear functional on the normed vector space of continuous and compactly supported functions $C_c(X)$ on X endowed with the supremum norm $\|f\|_\infty := \sup_{x \in X} |f(x)|$. Moreover $|\mu(f)| \leqslant \|f\|_\infty \mu(\mathrm{Supp}(f))$, where

$$\mathrm{Supp}(f) := \overline{\{x \in X, f(x) \neq 0\}}$$

is the (compact) *support* of f. And the functionial is non-negative, i.e. $f \geqslant 0 \Rightarrow \mu(f) \geqslant 0$.

It turns out that this is a characterization of Radon measures: every non-negative linear functional on $C_c(X)$ is of the form $\mu(f)$ for a certain Radon measure μ on X. This is called the Riesz representation theorem for locally compact spaces (see Rudin's book on Real and Complex analysis for a proof).

This functional analytic point of view leads to an integration theory (in which one *defines* a measure as a non-negative linear functional) that serves most purposes of analysis on locally compact spaces and has been the preferred route to present integration theory among mathematicians for a while (cf. Bourbaki's volumes on integration). Its drawback is that it is confined to locally compact spaces and hence less general than the route via abstract measure theory we have presented in these lectures, and unsuitable for much of probability theory.

8. Product measures

Definition 8.1 (Product σ-algebra)**.** *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces. Then the σ-algebra generated by all product sets $A \times B$ with $A \in \mathcal{A}$ and $B \in \mathcal{B}$ is called the product σ-algebra of $\mathcal{A}$ by $\mathcal{B}$ and is denoted by $\mathcal{A} \otimes \mathcal{B}$.*

Remark 8.2. (i) note that the family of product sets (i.e. $A \times B$, $A \in \mathcal{A}$ and $B \in \mathcal{B}$) forms a π-system.
(ii) by analogy with the notion of *product topology* in topology, the *product σ-algebra* is the smallest σ-algebra on $X \times Y$ making both projection maps (to X and to Y) measurable (this is a trivial check).
(iii) for the Borel σ-algebras on $\mathbb{R}^d$, we have the following nice compatibility:

$$\mathcal{B}(\mathbb{R}^{d_1}) \otimes \mathcal{B}(\mathbb{R}^{d_2}) = \mathcal{B}(\mathbb{R}^{d_1+d_2}).$$

However this is not so for the σ-algebra of Lebesgue measurable sets (the product of two copies of $\mathcal{L}(\mathbb{R})$ is not complete, so strictly smaller than $\mathcal{L}(\mathbb{R}^2)$, see the 2nd Example sheet).

Example: if $X = \bigsqcup_1^N X_i$ and $Y = \bigsqcup_1^M Y_j$ are finite partitions of two sets and $\mathcal{A}$ (resp. $\mathcal{B}$) is the Boolean algebra generated by this partition of X (resp. Y), then $\mathcal{A} \otimes \mathcal{B}$ is the Boolean algebra generated by the partition $(X_i \times Y_j)_{1 \leqslant i \leqslant N, 1 \leqslant j \leqslant M}$ of $X \times Y$.

The following lemma says that every vertical slice of a $\mathcal{A} \otimes \mathcal{B}$-measurable set is itself $\mathcal{B}$-measurable (and of course vice versa: every horizontal slice is $\mathcal{A}$-measurable).

Lemma 8.3. *If $E \subset X \times Y$ is $\mathcal{A} \otimes \mathcal{B}$-measurable, then for all $x \in X$ the slice $E_x := \{y \in Y, (x, y) \in E\}$ is $\mathcal{B}$-measurable.*

Proof. Note that $\mathcal{C} := \{E \subset X \times Y, E_x \in \mathcal{B} \forall x \in X\}$ contains all subsets of the form $E \times F$ with $E \in \mathcal{A}$ and $F \in \mathcal{B}$. And it is a σ-algebra. Indeed if $E \in \mathcal{C}$, then so does its complement E^c, because $(E^c)_x = \{y \in Y, (x, y) \in E^c\} = (E_x)^c$. Similarly if $E_n \in \mathcal{C}$ for all n, then $\bigcup_n E_n \in \mathcal{C}$ because $(\bigcup_n E_n)_x = \bigcup (E_n)_x$. This means that $\mathcal{A} \otimes \mathcal{B} \subset \mathcal{C}$. □

Lemma 8.4. *Assume that $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ are σ-finite measure spaces and let $f : X \times Y \to [0, +\infty]$ be $\mathcal{A} \otimes \mathcal{B}$-measurable. Then*
(a) for all $x \in X$, $y \mapsto f(x, y)$ is $\mathcal{B}$-measurable,
(b) $x \mapsto \int_Y f(x, y) d\nu(y)$ is $\mathcal{A}$-measurable.

Proof. (a) the special case $f = 1_E$ for any $E \in \mathcal{A} \otimes \mathcal{B}$ was exactly the content of the previous lemma. This implies the case when f achieves only finitely many values, i.e. when f is a simple function. The general case follows, because f is a limit of simple functions (see Lemma 7.6).

(b) By the same token we may assume that $f = 1_E$ for some $E \in \mathcal{A} \otimes \mathcal{B}$. Now we may write $Y = \bigcup_m Y_m$, where $Y_{m+1} \subset Y_m$ and $\nu(Y_m) < \infty$ for all m. Note that it is enough to show that $x \mapsto \nu(E_x \cap Y_m)$ is $\mathcal{A}$-measurable for each m. Indeed:

$$\int_Y 1_E(x, y) d\nu(y) = \nu(E_x) = \lim_m \nu(E_x \cap Y_m)$$

Then we can consider the family

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

and observe that
(i) $\mathcal{C}$ contains all $E = A \times B$, because $\nu(E_x \cap Y_m) = 1_A(x)\nu(B \cap Y_m)$
(ii) $\mathcal{C}$ is stable under complementation: $\nu((E^c)_x \cap Y_m) = \nu(Y_m) - \nu(Y_m \cap E_x)$

(iii) $\mathcal{C}$ is stable under disjoint countable unions: if $E = \bigsqcup_n E_n$, then

$$\nu(E_x \cap Y_m) = \sum_n \nu((E_n)_x \cap Y_m)$$

so by Dynkin's lemma we conclude that $\mathcal{C}$ is the σ-algebra generated by the π-system of product sets. So $\mathcal{C} = \mathcal{A} \otimes \mathcal{B}$. □

Lecture 9

Proposition-Definition 8.5 (Product measure)**.** *Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be σ-finite measure spaces. Then there exist a unique measure σ on the product σ-algebra $\mathcal{A} \otimes \mathcal{B}$ such that for all $A \in \mathcal{A}$ and all $B \in \mathcal{B}$ we have:*

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

The measure σ is called the product measure *and is usually denoted by $\mu \otimes \nu$.*

Proof. Let us first prove the existence of such a measure. For $E \in \mathcal{A} \otimes \mathcal{B}$ we set:

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

where E_x is the vertical slice defined above. It is well-defined because the function $x \mapsto \nu(E_x)$ is $\mathcal{A}$-measurable by Lemma 8.4 (b) applied to $f = 1_E$. It is also countably additive: this is a consequence of the Monotone Convergence Theorem:

$$\sigma(\bigsqcup_n E_n) = \int_X \sum_n \nu((E_n)_x) d\mu(x) = \sum_n \int_X \nu((E_n)_x) d\mu(x) = \sum_n \sigma(E_n).$$

As for uniqueness, it follows immediately from Dynkin's π-system uniqueness lemma (cf. Prop. 5.2), because the family of all product sets $A \times B$, $A \in \mathcal{A}$, $B \in \mathcal{B}$ clearly forms a π-system, which by definition generates $\mathcal{A} \otimes \mathcal{B}$. □

Remark 8.6. The σ-finiteness assumption is essentially a technical assumption that holds in most cases of interest, but it is important for the uniqueness clause in the previous definition.

Remark 8.7. Another route to establish the existence of the product measure is to apply Caratheodory's extension theorem to the Boolean algebra generated by product sets. This is possible and close to what we did in the first lecture regarding the Jordan measure (as one first needs to show that this Boolean algebra is made of disjoint finite unions of product sets and extend to measure to these sets), but also requires some work (and the Monotone Convergence Theorem) to establish that it has the continuity property.

Remark 8.8. One can also define the product measure on the product of more than two measurable spaces, i.e. $\mathcal{A} \otimes \mathcal{B} \otimes \mathcal{C}$ on $X \times Y \times Z$, simply by iterating the above construction. One checks easily (exercise!) that this operation is associative (i.e. $(\mathcal{A} \otimes \mathcal{B}) \otimes \mathcal{C} = \mathcal{A} \otimes (\mathcal{B} \otimes \mathcal{C})$) and the resulting σ-algebra, and product measure, is independent of the order in which the products are taken.

Example: The Lebesgue measure on Borel subsets of $\mathbb{R}^d$ is the (d-fold) product measure of the Lebesgue measure on $\mathbb{R}$.

Theorem 8.9 (Fubini-Tonelli theorem)**.** *Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be σ-finite measure spaces.*

(a) If $f : X \times Y \to [0, +\infty]$ is $\mathcal{A} \otimes \mathcal{B}$-measurable, then

$$\int_{X \times Y} f d\mu \otimes \nu = \int_X \Big(\int_Y f(x,y) d\nu(y) \Big) d\mu(x) = \int_Y \Big(\int_X f(x,y) d\mu(x) \Big) d\nu(y). \quad (8.1)$$

(b) If $f : X \times Y \to \mathbb{R}$ is $\mu \otimes \nu$-integrable, then for μ-almost every x, $y \mapsto f(x,y)$ is ν-integrable and $x \mapsto \int_Y f(x,y) d\nu(y)$ is μ-integrable. Moreover (8.1) holds.

Proof. (a) holds for $f = 1_E$ and any $E \in \mathcal{A} \otimes \mathcal{B}$ by Lemma 8.4. So it holds for simple functions and hence, thanks to the Monotone Convergence Theorem, for all non-negative measurable f.

(b) write $f = f^+ - f^-$ and apply (a) to f^+ and f^-. □

Remark 8.10. Note that $\mu\otimes\nu(|f|) = \mu\otimes\nu(f^+)+\mu\otimes\nu(f^-)$, so if this is finite, then both terms are finite and by (a) $\int_X \big(\int_Y f^{\pm}(x,y)d\nu(y)\big)d\mu(x) < \infty$, which implies that $\{x \in X, \int_Y f^{\pm}(x,y)d\nu(y) = \infty\}$ is a μ-null set.

Example: The assumption that f is $\mu\otimes\nu$-integrable is necessary in general to be able to swap the order of integration. A silly example is given by $X = Y = \mathbb{N}$ and $\mathcal{A} = \mathcal{B}$ =the discrete σ-algebra of all subsets with $\mu = \nu$ =counting measure, and for all $n, m \geqslant 1$,

$$f(n,m) = 1_{n=m} - 1_{n=m+1}$$

Clearly, $\forall m \geqslant 1$, $\sum_{n\geqslant 1} f(n,m) = 0$, while for all $n \geqslant 2$ $\sum_{m\geqslant 1} f(n,m) = 0$ and $\sum_{m\geqslant 1} f(1,m) = 1$. So we see that

$$\sum_n \sum_m f(n,m) \neq \sum_m \sum_n f(n,m).$$

Remark 8.11. This theorem applies in particular to Lebesgue measure on $\mathbb{R}^d$ provided the function f is Borel-measurable on $\mathbb{R}^d$. There is also a version of this theorem (and of Lemma 8.4) that holds for all Lebesgue measurable functions on $\mathbb{R}^d$ as well (see next remark).

Remark 8.12 (Completed version)**.** As mentioned above the product measure $\mu\otimes\nu$ on the product σ-algebra may not be complete (i.e. sub-null sets, i.e. subsets of $\mu\otimes\nu$-null sets from $\mathcal{A}\otimes\mathcal{B}$, may not be in $\mathcal{A}\otimes\mathcal{B}$) even though both μ and ν are complete. So it is often implicit to automatically complete $\mu\otimes\nu$ to the completed σ-algebra $\overline{\mathcal{A}\otimes\mathcal{B}}$ (i.e. the family of sets that are unions of subsets from $\mathcal{A}\otimes\mathcal{B}$ with sub-null sets, see the Example sheet 1). Then Theorem 8.9 continues to hold verbatim with $\overline{\mathcal{A}\otimes\mathcal{B}}$ in place of $\mathcal{A}\otimes\mathcal{B}$ (exercise!). On the other hand, the corresponding analogues of Lemma 8.3 and Lemma 8.4 (a) only hold for μ-almost every $x \in X$, because for example a set of the form $E = A \times B$, where $A \in \mathcal{A}$ is μ-null, while $B \notin \mathcal{B}$ will be $\mu\otimes\nu$-null (hence $\overline{\mathcal{A}\otimes\mathcal{B}}$-measurable) and yet its slices E_x for $x \in A$ will not be in $\mathcal{B}$. At any case Theorem 8.9 holds for Lebesgue measurable functions on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$.

9. Foundations of Probability Theory

In 1933 Kolmogorov published his famous treatise "Foundations of the Theory of Probability" (or rather "Grundbegriffe der Wahrscheinlichkeitsrechnung" as he wrote it first in German) in which he laid down the mathematical foundations of probability theory, making use, for the first time, of the then recent formalism of abstract measure theory and arguing that it is the right framework and language in which to develop rigourously the calculus of probabilities. We follow his footsteps in our lecture today.

For a lovely recent book with a historical point of view and a modern perspective on what probability and statistics are really about, I recommend "Ten great ideas about chance" by Diaconis and Skyrms (2018).

We will fix the ambient set Ω. Probabilists call it a universe. It is the set of all possible outcomes. An element of Ω is an outcome ω, namely one of many possible scenarios that might happen. In probability theory, we usually fix the universe once and for all.

The family $\mathcal{F}$ of subsets of Ω will be the family of all possible events, or subsets of possible outcomes that could take place. By assumption it is a σ-algebra of subsets of Ω. A subset in $\mathcal{F}$ is called an event.

Now the odds that an event, say $A \in \mathcal{F}$ occurs, is a number

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

called the probability of A. The natural axioms are:

$$(i)\ \mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B)$$

provided A and B are disjoint, i.e. they cannot occur simultaneously.

$$(ii)\ \mathbb{P}(\Omega) = 1, \mathbb{P}(\varnothing) = 0$$

and the continuity axiom: if A_n is a non-increasing sequence of events that cannot occur simultaneously (i.e. $A_{n+1} \subset A_n$ and $\bigcap_n A_n = \varnothing$), then

$$\lim_{n \to +\infty} \mathbb{P}(A_n) = 0.$$

The first two axioms turn $\mathbb{P}$ into a finitely additive measure on $(\Omega, \mathcal{F})$ and the last one (which is harder to justify empirically given that it depends on infinitely many events, but it still very reasonable to take for granted) make it a (countable additive) measure (recall that a finitely additive measure with the continuity property is countably additive, see Exple sheet).

Recall:

Definition 9.1. *A measure μ on a measurable space $(\Omega, \mathcal{F})$ is called a* probability measure *if $\mu(\Omega) = 1$.*

In probability theory it is customary to give different names to notions that we have already defined in measure theory. It's only a cultural difference, but the objects are the same. For example a measurable function is called a random variable and the integral is called expectation.

Definition 9.2. *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (namely a measure space with a measure $\mathbb{P}$ of total mass 1), a measurable function $f : \Omega \to \mathbb{R}$ is called a random variable and is usually denoted by a capital letter X or Y. Similarly the $\mathbb{P}$-integral is called the expectation and is denoted by $\mathbb{E}$. An event A is said to hold almost surely if $\mathbb{P}(A) = 1$.*

So for example we will write $\mathbb{E}(X)$ in place of $\int_X f d\mathbb{P}$, etc. More generally we can defined $\mathbb{R}^d$-valued random variables $X = (X_1, ..., X_d)$. These are the same thing as vectors of d real valued random variables.

Now comes an important definition.

Definition 9.3. *A random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ determines a Borel measure μ_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by*

$$\mu_X(A) = \mathbb{P}(X \in A)$$

for every Borel set $A \in \mathcal{B}(\mathbb{R})$. The measure μ_X is a probability measure called the law of X or the distribution of X.

In other words the probability distribution μ_X is the *image measure* of $\mathbb{P}$ under the map $X : \Omega \to \mathbb{R}$. In general, given a measurable function $f : (Y, \mathcal{A}) \to (Z, \mathcal{C})$ between two measurable spaces, and given a measure μ on $(Y, \mathcal{A})$ the image measure, denoted by $f_*\mu$ is the measure on $(Z, \mathcal{C})$ defined by

$$f_*\mu(C) := \mu(f^{-1}(C)). \tag{9.1}$$

Note that, clearly, if μ is a probability measure, so is $f_*\mu$.

The function $t \mapsto F_X(t) := \mathbb{P}(X \leqslant t)$ is called the distribution function of X. (the notation $\mathbb{P}(X \leqslant t)$ is a shorthand for $\mathbb{P}(\{\omega \in \Omega, X(\omega) \leqslant t\})$.

Remark 9.4. By the same token, one can define the *distribution* of an $\mathbb{R}^d$-valued random variable $X = (X_1, \ldots, X_d)$. This is the Borel probability distribution $\mu_{(X_1,...,X_d)}$ on $\mathbb{R}^d$ defined by $\mu_{(X_1,...,X_d)}(A) = \mathbb{P}((X_1, \ldots, X_d) \in A)$ for any Borel subset $A \in \mathcal{B}(\mathbb{R}^d)$.

Example 9.5 (Archimedes' theorem)**.** *Suppose we pick a point on the sphere uniformly at random. Call this random point ω. Consider the orthogonal projection of this point onto the north-south axis. What is the probability that it is closer to the center of the sphere than it is from either pole?*

Answer: $\frac{1}{2}$. We can formalize this problem within the above framework. Here Ω will be the Euclidean sphere, $\mathcal{F}$ the family of Borel subsets of the sphere (note that they are precisely the intersections of Borel sets of $\mathbb{R}^3$ with the sphere), and $\mathbb{P}$ will be the Lebesgue measure on the sphere, normalized so it has total mass 1 (to define it, one can for example take Lebesgue measure on $\mathbb{R}^3$ restricted to the unit ball minus $\{0\}$, renormalize it so it has total measure 1 and take its image (as we have just defined) under the projection map $\mathbb{R}^3\backslash\{0\} \to \Omega$, $x \mapsto x/\|x\|$.).

Then $(\Omega, \mathcal{F}, \mathbb{P})$ will be a probability space and the orthogonal projection $Y(\omega)$ will be measurable (Borel measurable, because it is in fact continuous). The distribution of $Y(\omega)$ can be easily computed (exercise!): it turns out that it is the uniform distribution *on the north-south axis, i.e. Lebesgue measure on that interval, renormalized so as to have total mass one. This is easily checked via calculus, and is a fairly remarkable fact, going back to Archimedes.*

So if $X(\omega)$ is the distance between $Y(\omega)$ and the north pole, then μ_Y is the uniform probability measure on the interval $[0, 2]$ and $F_X(t) = \mathbb{P}(X \leqslant t) = \frac{t}{2} 1_{t\in[0,2]}$.

Lecture 10

Proposition 9.6. *Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. Let $X : \Omega \to \mathbb{R}$ be a random variable. Then the distribution function F_X is non-decreasing and right-continuous. Moreover F_X determines μ_X uniquely.*

Proof. It is clear from the definition that F_X is non-decreasing. So see the continuity claim, let $t_n \geqslant t$ be a sequence such that $t_n \to t$. Then the events $\{X \leqslant t_n\} = \{\omega \in \Omega, X(\omega) \leqslant t_n\}$ form a non-increasing sequence of events whose intersection is exactly $\{X \leqslant t\}$. By downward monotone convergence for sets, we conclude that $\mathbb{P}(X \leqslant t_n) \to \mathbb{P}(X \leqslant t)$ as desired.

As for uniqueness, it follows from Dynkin's lemma, given that the family of all intervals $(-\infty, t]$ for $t \in \mathbb{R}$ forms, together with the empty set, a π-system generating the Borel σ-algebra $\mathcal{B}(\mathbb{R})$. □

Proposition 9.7. *Conversely, if $F : \mathbb{R} \to [0,1]$ is a non-decreasing, right continuous function with $\lim_{t\to-\infty} F(t) = 0$ and $\lim_{t\to+\infty} F(t) = 1$, then there exists a unique Borel probability measure μ_F on $\mathbb{R}$ such that for all $t \in \mathbb{R}$,*

$$F(t) = \mu_F((-\infty, t]).$$

Remark 9.8. The measure μ_F is called the Lebesgue-Stieltjes measure associated to F. For all $a < b$, we have

$$\mu_F((a, b]) = F(b) - F(a). \tag{9.2}$$

Proof of Proposition 9.7. The proof of uniqueness is the same as before (via Dynkin's π-system lemma). To show that such a measure exists, note that one can define μ_F by (9.2) on half open intervals and this defines a finitely additive measure on the Boolean algebra consisting of finite unions of disjoint half open intervals. To show that this extends to a well-defined Borel probability measure on $\mathbb{R}$, we need to apply the Carathéodory extension theorem. And for this we only need to verify that it has the continuity property. The proof that μ_F has the continuity property is exactly the same as the one we gave for the continuity of the Jordan measure on elementary sets (see Lemma 3.8, which was based on the Heine-Borel property). One only needs to check that given a finite union of half-open intervals of the form $(a, b]$ it is possible to shrink them a tiny bit and find a', a'' with $a < a' < a'' < b$ so that $F(a'')$ is arbitrarily close to $F(a)$ (this is possible thanks to the right continuity property of F) and $(a, b] \supset [a', b] \supset (a'', b]$. This way $\mu_F((a, b])$ and $\mu_F((a'', b])$ will be very close to each other, while $[a', b]$ is compact. The argument via Heine-Borel then applies without change. □

There is a more direct way to construct the Lebesgue-Stieltjes measure associated to F, which takes advantage of the fact that we have already defined the Lebesgue measure (instead of proving the existence via the extension theorem, i.e. via the same route used to construct the Lebesgue measure). And this is to view μ_F as the image measure of the Lebesgue measure on the interval $[0, 1]$ under the "inverse" of F, namely the function

$$\begin{aligned} g : (0,1) &\to \mathbb{R} \\ y &\mapsto g(y) := \inf\{t \in \mathbb{R}, F(t) \geqslant y\} \end{aligned}$$

Lemma 9.9. *Given a function $F : \mathbb{R} \to [0,1]$ as in Proposition 9.7, the "inverse function" g defined by (??) is non-decreasing, left continuous, and satisfies:*

$$\forall t \in \mathbb{R}, \forall y \in (0,1), g(y) \leqslant t \iff F(t) \geqslant y. \tag{9.3}$$

Proof. This is an easy check. First observe that

$$I_y := \{t \in \mathbb{R}, F(t) \geqslant y\}$$

is an interval in $\mathbb{R}$, because if $t < t' \in I_y$, then $y \leqslant F(t) \leqslant F(t')$, because F is non-decreasing, so $y \leqslant F(t'')$ and hence $t'' \in I_y$ for every $t'' \in (t, t')$. This means that I_y has the form

$$I_y = \text{“(”}g(y), +\infty),$$

where the bracket “(” could a priori be either open (or closed [. But F is right continuous, so the infinimum defining $g(y)$ is realized and $I_y = [g(y), +\infty)$. This shows (9.3). Also if $y_1 \leqslant y_2$, then $I_{y_2} \subset I_{y_1}$, so $g(y_1) \leqslant g(y_2)$. It remains to check left-continuity of g. This is clear because if $y_n < y$ and $y_n \to y$, then $\bigcap_n I_{y_n} = I_y$ by definition of I_y. So $g(y_n) \to g(y)$. □

Remark 9.10. If F is continuous and increasing, then $g = F^{-1}$ is the inverse of F (i.e. $F \circ g$ is the identity on $(0,1)$ and $g \circ F$ the identity on $\mathbb{R}$.)

Now let m be the Lebesgue measure on $(0,1)$ and set

$$\mu := g_* m,$$

be image of m under g (see (9.1)), that is:

$$\mu(A) := m(g^{-1}(A))$$

for every Borel subset $A \subset \mathbb{R}$. Then g is Borel measurable (because $\{y \in \mathbb{R}, g(y) \leqslant t\} = (0, F(t)] \in \mathcal{B}((0,1))$), so μ is a Borel measure, and

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

so by uniqueness, this is precisely the Lebesgue-Stieltjes measure constructed previously: $\mu = \mu_F$.

We have seen that a random variable (i.e. a measurable function defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$) gives rise to a Borel probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, called the *law* or *probability distribution* of X. It is interesting to note that conversely every Borel probability measure arises this way:

Proposition 9.11. *If μ is a Borel probability measure on $\mathbb{R}$, then there exist a probability space $(\Omega, \mathcal{F}, \mathbb{R})$ and a random variable X such that $\mu = \mu_X$. In fact one can take $\Omega = (0,1)$, $\mathcal{F}$ the Borel σ-algebra of the interval $(0,1)$ and $\mathbb{P}$ Lebesgue measure on $(0,1)$.*

Proof. The first assertion is obvious if we set $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$, $\mathbb{P} = \mu$ and define the random variable by $X(x) = x$. To see that one can also do this on $(0,1)$ we can use the distribution function $F(t) := \mu((-\infty, t])$ and define the random variable X as $X = g$, the inverse of F, as in the previous discussion, that is:

$$X(\omega) := \inf\{t \in \mathbb{R}, F(t) \geqslant \omega\} : (0,1) \to \mathbb{R}.$$

Then X is a random variable, because X is Borel measurable (we have seen that it is non-decreasing and left continuous) and $\mu_X = \mu$, because as we have seen:

$$\begin{aligned} \mathbb{P}(X \in (a,b]) &= m(\{\omega \in (0,1), a < X(\omega) \leqslant b\} \\ &= m(\{\omega \in (0,1), F(a) < \omega \leqslant F(b)\} \quad = F(b) - F(a) = \mu((a,b]). \end{aligned}$$

□

Remark 9.12. If there exists $f \geqslant 0$ measurable such that $\mu_X((a,b]) = \int_a^b f(t)dt$, we say that X (or μ_X) has a density (with respect to Lebesgue measure). And f is called the density function.

Example 9.13 (Some examples of Borel probability measures on $\mathbb{R}$)**.** .

(1) *the uniform distribution on* $[0,1]$ *has density* $f(x) = 1_{[0,1]}$ *and distribution function* $F(t) = \int_{-\infty}^{t} f(x)dx = t1_{[0,1]}$.
(2) *the exponential distribution with rate* $\lambda > 0$ *is defined as*
$$f(x) = \lambda e^{-\lambda x} 1_{x \geqslant 0},$$
while $F(t) = 1_{t\geqslant 0}(1 - \exp(-t/\lambda)$.
(3) *the gaussian distribution with standard deviation* $\sigma > 0$ *and mean* m *is defined by its density*
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-m)^2}{2\sigma^2}).$$
A wide-spread short-hand for the gaussian distribution is $\mathcal{N}(m, \sigma^2)$.
(4) *the Dirac mass* δ_m *at* $m \in \mathbb{R}$ *is the probability distribution defined as:*
$$\delta_m(A) = 1_{m\in A}$$
for any Borel subset $A \subset \mathbb{R}$. *Its distribution function is Heavyside's function* $H_m(t) := 1_{t\geqslant m}$.

Definition 9.14. *If* X *is a random variable (on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*) we define*

(1) $\mathbb{E}(X)$ *its mean (well-defined if* X *is* $\mathbb{P}$-integrable*, i.e. if* $\mathbb{E}(|X|) < \infty$*),*
(2) $\mathbb{E}(X^k)$ *its moment of order* $k \in \mathbb{N}$ *(well-defined if* X^k *is integrable),*
(3) $\mathbf{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}((X - \mathbb{E}(X))^2)$ *its variance (well-defined if* X^2 *is integrable).*

Remark 9.15. If $f \geqslant 0$ is Borel measurable, then $\mathbb{E}(f(X)) = \int f \circ X(\omega)d\mathbb{P}(\omega) = \int f(x)d\mu_X(x)$, because $\mu_X = X_*\mathbb{P}$.

Lecture 11

10. Independence

Today we discuss a central notion on probability theory, that of independence. This notion gives a distinctive flavour to probability theory, which, until now could have been mistaken for a sub-branch of measure theory.

Definition 10.1 (independence of events)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A sequence of events $(A_i)_{i\geqslant 1}$ is called mutually independent if for every finite subset $F \subset \mathbb{N}$ we have*

$$\mathbb{P}(\bigcap_{i\in F} A_i) = \prod_{i\in F} \mathbb{P}(A_i) \tag{10.1}$$

Remark 10.2. If $(A_i)_{i\geqslant 1}$ is an independent sequence of events, then so is $(B_i)_{i\geqslant 1}$, where for each i, B_i is either A_i or its complement A_i^c. This is easily checked, for example by writing $\mathbb{P}(\bigcap_{i\in F} B_i) = \mathbb{E}(\prod_{i\in F} 1_{B_i})$, where 1_{B_i} is either f_i or $1-f_i$ with $f_i := 1_{A_i}$, and then expanding the product.

Definition 10.3 (independence of subalgebras)**.** *A sequence of σ-subalgebras $(\mathcal{A}_i)_{i\geqslant 1}$ ($\mathcal{A}_i \subset \mathcal{F}$) is called mutually independent if for any event $A_i \in \mathcal{A}_i$ the sequence of events $(A_i)_i$ is mutually independent.*

Remark 10.4. Remark 10.2 above shows that if $(A_i)_{i\geqslant 1}$ is an independent family of events, then the sequence of subalgebras $(\mathcal{A}_i)_{i\geqslant 1}$ where $\mathcal{A}_i := \{\varnothing, A_i, A_i^c, \Omega\}$ forms an independent sequence.

Remark 10.5. A sooped-up version of the last two remarks is as follows. If $\Pi_i \subset \mathcal{A}_i$ is a π-system with $\sigma(\Pi_i) = \mathcal{A}_i$ for all i, then it is enough to check (10.1) for A_i's in Π_i to be able to claim that the $(\mathcal{A}_i)_i$ form an independent sequence. To see this (say in the simple case when there are only two subalgebras) consider, for some event $A_1 \in \Pi_1$ the maps $A \mapsto \mathbb{P}(A \cap A_1)$ and $A \mapsto \mathbb{P}(A)\mathbb{P}(A_1)$ and note that they both are measures on $(\Omega, \mathcal{A}_2)$ with the same total mass and that they coincide on the π-system Π_2. Hence by Dynkin's lemma, they must coincide on $\mathcal{A}_2$, so that (10.1) holds now for all $A_1 \in \Pi_1$ and all $A_2 \in \mathcal{A}_2$. Then apply this argument one more time using the π-system Π_1 instead and conclude that (10.1) holds for all $A_1 \in \mathcal{A}_1$ and all $A_2 \in \mathcal{A}_2$. A similar argument handles the general case of an arbitrary family of subalgebras.

Notation 1. *If X is a random variable, we will denote by $\sigma(X)$ the smallest σ-algebra of $\mathcal{F}$ making X $\mathcal{A}$-measurable, that is:*

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

Definition 10.6. *A sequence of random variables $(X_i)_{i\geqslant 1}$ is called (mutually-) independent if $(\sigma(X_i))_{i\geqslant 1}$ is mutually independent.*

Remark 10.7 (independence and product measure)**.** By the previous remark, this is equivalent to asking that for every finite subset F of indices and all $t_i \in \mathbb{R}$,

$$\mathbb{P}(X_i \leqslant t_i \forall i \in F) = \prod_{i\in F} \mathbb{P}(X_i \leqslant t_i)$$

or equivalently

$$\mu_{(X_{i_1},\dots,X_{i_d})} = \mu_{X_{i_1}} \otimes \cdots \otimes \mu_{X_{i_d}}$$

if $F = \{i_1, \dots, i_d\}$, where the above is the product of the laws μ_{X_i} of the X_i's, and $\mu_{(X_{i_1},\dots,X_{i_d})}$ is the law of the $\mathbb{R}^d$-valued random vector $(X_{i_1}, \dots, X_{i_d})$.

So independence of two random variables can be read off the probability distribution of the pair: two random variables are independent if and only if the distribution of the pair is the product of the two individual distributions.

Remark 10.8. If f_i is a measurable function for each index i, and $(X_i)_{i\geqslant 1}$ is a sequence of (mutually-) independent random variables, then $(f_i(X_i))_i$ is also a sequence of independent random variables. This is clear, because $\sigma(f(X)) \subset \sigma(X)$ for any random variable X and any measurable function f.

Proposition 10.9. *If X, Y are independent non-negative random variables, then*

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

The same holds if X and Y are independent and integrable (and this implies that XY is integrable).

Proof. This is an instance of the Fubini-Tonelli theorem applied to the function $f(x, y) = xy$ on $[0, +\infty)^2$ with the measure $\mu_{(X,Y)} = \mu_X \otimes \mu_Y$. □

Example: (Bernstein's example) This example shows that pairwise independence does not imply mutual independence. Let X, Y be two independent coin tosses. That is $X, Y \in \{-0, 1\}$ and

$$\mathbb{P}(X = 0) = \mathbb{P}(Y = 0) = \mathbb{P}(X = 1) = \mathbb{P}(Y = 1) = \frac{1}{2}.$$

Let $Z := |X - Y|$. Then it is straightforward to check that each of the pairs (X, Y), (Y, Z) and (X, Z) is an independent pair. But the triple (X, Y, Z) is *not* an independent triple, because:

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

while

$$\mathbb{P}((X, Y, Z) = (1, 1, 0)) = \frac{1}{4}$$

and

$$\mathbb{P}(X = 1 \text{ and } Y = 1)\mathbb{P}(Z = 0) = \frac{1}{8} \neq \frac{1}{4}.$$

Example: decimal expansion. Let $\Omega = (0, 1)$, $\mathbb{P} = m$ = Lebesgue measure, and $\mathcal{F} = \mathcal{B}(0, 1)$ the Borel σ-algebra. Clearly $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

Given $\omega \in (0, 1)$ we may look at its decimal expansion $\omega = 0.\epsilon_1\epsilon_2...$, where $\epsilon_n \in \{0, 1, ..., 9\}$. There is the usual indeterminacy of course that takes place when ω is a decimal number, i.e. of the form $a/10^b$ for integers a, b, in which case there can be two decimal expansions one of which ends with an infinite string of 9's: in that case, we can choose one of the two expansions, e.g. avoid infinite strings of 9.

Now set $X_n(\omega) = \epsilon_n$. This becomes a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ (it is plain to check that X_n is measurable).

Claim: The $(X_n)_n$'s form a sequence of independent random variables that are uniformly distributed in $\{0, 1, ..., 9\}$.

In particular this means that $\mathbb{P}(X_n = i) = \frac{1}{10}$ for each $n \geqslant 1$ and $i \in \{0, ..., 9\}$. The claim is easily checked: the set of ω's such that $X_n(\omega) = i$ is a union 10^{n-1} of intervals of length 10^{-n}. So we see that

$$\mathbb{P}(X_1 = i_1 \text{ and } X_2 = i_2 \text{ and } ... \text{ and } X_n = i_n) = \frac{1}{10^n} = \prod_{j=1}^{n} \mathbb{P}(X_j = i_j),$$

which means that the $(X_n)_n$'s are independent.

Note that

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

for all $\omega \in (0,1)$.

Remark 10.10. It turns out that this is another way to build the Lebesgue measure on $\mathbb{R}$. Pick independent and uniformly distributed random variables X_n on $\{0, ..., 9\}$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and define a random variable

$$Y := \sum_{n \geqslant 1} \frac{X_n}{10^n}.$$

Then, as the above example demonstrates, the law of Y is precisely the Lebesgue measure on $(0,1)$.

As an aside, one may further ask how to construct such a sequence (of independent random variables) from scratch? (in the example, we have defined such a sequence of independent variables, but our construction made use of Lebesgue measure). The following proposition enables us to find a single probability space on which to find a sequence of independent random variables with prescribed laws:

Proposition-Definition 10.11 (infinite product measure)**.** *Let $\{(\Omega_i, \mathcal{F}_i, \nu_i)\}_{i \geqslant 1}$ be a sequence of probability spaces. Let $\Omega = \prod_{i \geqslant 1} \Omega_i$. Let $\mathcal{C}$ be the Boolean algebra of* cylinder sets*, namely subsets of the form $B := A \times \prod_{i>n} \Omega_i$, where $A \subset \prod_{i=1}^n \Omega_i$ belongs to the product σ-algebra $\mathcal{F}_1 \otimes \ldots \otimes \mathcal{F}_n$. Finally let $\mathcal{F} = \sigma(\mathcal{C})$. Then there exists a unique probability measure ν on $(\Omega, \mathcal{F})$ such that*

$$\nu(B) = \nu_1 \otimes \ldots \otimes \nu_n(A),$$

for every cylinder set $B = A \times \prod_{i>n} \Omega_i$ as above.

Proof. Apply Carathéodory's extension theorem. See the 2nd example sheet. □

Now if μ_i is a Borel probability measure on $\mathbb{R}$, then we know by Proposition 9.11 that there exists a random variable Y_i on $((0,1), \mathcal{B}(0,1), m)$ whose law is precisely μ_i. If we let $\Omega_i = (0,1)$, $\mathcal{F}_i = \mathcal{B}(0,1)$ and $\nu_i = m$ the Lebesgue measure, then we may form the infinite product $\Omega = \prod \Omega_i$ as in the previous statement. Now we may set $X_i(\omega) = Y_i \circ \pi_i$, where $\pi_i : \Omega \to \Omega_i$ is the projection to the i-th coordinate. This will yield an infinite sequence $(X_i)_i$ of *independent* random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that X_i has law μ_i for each i.

Remark 10.12. The above is a special case of a more general theorem of Kolmogorov, the Kolmogorov extension theroem, that asserts that any family of measures μ_n defined $\prod_1^n \Omega_i$ and satisfying a necessary compatibility condition (the projection of μ_n to the first m coordinates is assumed to coincide with μ_m) gives rise to a unique measure on the infinite product, whose restriction to the cylinders coincide with the μ_n's.

We now pass to the Borel-Cantelli lemmas: let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(A_n)_{n \geqslant 1}$ a sequence of events.

Lemma 10.13 (1st Borel Cantelli lemma)**.** *If $\sum_{n \geqslant 1} \mathbb{P}(A_n) < \infty$, then*

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

The lim sup of a sequence of events is the event A such that $1_A = \limsup_n 1_{A_n}$. In other words

$$\limsup_n A_n := \{\omega \in \Omega, \omega \in A_n \text{ for infinitely many } n\}.$$

The next result is a sort of converse:

Lemma 10.14 (2nd Borel Cantelli lemma)**.** *If* $\sum_{n\geqslant 1}\mathbb{P}(A_n)=\infty$ *and the* $(A_n)_{n\geqslant 1}$*'s are mutually independent, then*

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

Proof of both lemmas. (1) $\mathbb{E}(\sum_n 1_{A_n})=\sum_n \mathbb{P}(A_n)<\infty$, so $\mathbb{P}(\sum_n 1_{A_n}=\infty)=0$. This proves the first lemma.

(2) $(\limsup_n A_n)^c=\bigcup_N\bigcap_{n\geqslant N}A_n^c$, so exploiting the independence of the events (and using Remark 10.2) we may write for any $N\leqslant M$:

$$\mathbb{P}(\bigcap_{n\geqslant N}A_n^c)\leqslant\mathbb{P}(\bigcap_{N\leqslant n\leqslant M}A_n^c)=\prod_N^M(1-\mathbb{P}(A_n))\leqslant\exp(-\sum_N^M\mathbb{P}(A_n))$$

as $1-x\leqslant\exp(-x)$ for all $x\in[0,1]$. But the right hand side tends to 0 as M tends to $+\infty$. So we get $\mathbb{P}(\bigcap_{n\geqslant N}A_n^c)=0$ for all N, hence $\mathbb{P}(\limsup_n A_n)=1$. This proves the second lemma. □

The independence assumption in the second lemma can be relaxed. For example it holds assuming only pairwise independence, while an even weaker assumption (small correlation between the events) implies that the $\limsup_n A_n$ has positive probability. See the 3rd Example sheet.

Example: the infinite monkey theorem. Imagine a monkey frantically typing at random on a typewriter. What is the probability that he will eventually type the Song of Songs? Answer: 1. Indeed the Song of Songs is a finite, say of length N, string of characters (in its English translation that is, and the typewriter is assumed to offer all the letters of the Latin alphabet). So if A_n is the event: "the string from the $nN+1$-st character to the $(n+1)N$-th character is exactly the Song of Songs", then the A_n's are (or quite close to be) independent events. And each happens with probability K^{-N}, where K is the number of keys on the typewriter. So the series $\sum_n\mathbb{P}(A_n)$ diverges and hence $\mathbb{P}(\limsup_n A_n)=1$.

Lecture 12

Definition 10.15. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.*

(a) a sequence of random variables $(X_n)_{n\geqslant 1}$ is called a random (or stochastic) process.

(b) the σ-algebra $\mathcal{F}_n := \sigma(X_1, \dots, X_n)$ (which, by definition is the σ-algebra generated by the events $\{\omega \in \Omega, X_i(\omega) \leqslant t_i\}$ for any i and $t_i \in \mathbb{R}$) is called the n-th term of the associated filtration. *We have $\mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \mathcal{F}$ for all n,*

(c) the σ-algebra $\mathcal{T} := \bigcap_{n\geqslant 1} \sigma(X_n, X_{n+1}, \dots)$ is called the tail *σ-algebra of the process. Its elements are called* tail events.

Example: the events "$(X_n)_n$ converges" or "$\limsup_n X_n \geqslant T$" are tail events.

Theorem 10.16 (Kolmogorov 0-1 law)**.** *Let (X_n) be a family of (mutually independent random variables. Then for all $A \in \mathcal{T}$ we have*

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

Proof. Let $A \in \mathcal{T}$. Given n consider $B \in \sigma(X_1, \dots, X_n)$. Then

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \tag{10.2}$$

because $\mathcal{T}$ is independent of B. Let now $\mathcal{F}_n := \sigma(X_1, \dots, X_n)$ and $\mathcal{F}_\infty := \sigma(X_1, \dots, X_n, \dots)$. So the measures $B \mapsto \mathbb{P}(A)\mathbb{P}(B)$ and $B \mapsto \mathbb{P}(A \cap B)$ coincide on $\bigcup_n \mathcal{F}_n$. As this is a π-system generating $\mathcal{F}_\infty$, the measures coincide on $\mathcal{F}_\infty$. This means that (10.2) holds for all $B \in \mathcal{F}_\infty$. But $\mathcal{T} \subset \mathcal{F}_\infty$. So it actually holds for $B = A$. This means:

$$\mathbb{P}(A) = \mathbb{P}(A)^2,$$

or in other words $\mathbb{P}(A) \in \{0, 1\}$. □

Example: Let $(X_n)_n$ be a sequence of i.i.d. (that is independent and identically distributed) random variables with common law μ (a Borel measure on $\mathbb{R}$). Assume that for all $T > 0$ we have $\mathbb{P}(X_1 \leqslant T) < 1$. Then $\limsup_n X_n = +\infty$ almost surely (that is $\mathbb{P}(\{\omega \in \Omega, \limsup_n X_n(\omega) = +\infty\}) = 1$).

Indeed, to see this note that by Kolmogorov's $0-1$ law, we have $\mathbb{P}(\limsup_n X_n = +\infty) \in \{0, 1\}$. But $\sum_n \mathbb{P}(X_n \geqslant T) = +\infty$ for all T. So by the 2nd Borel-Cantelli lemma we must have $\mathbb{P}(\limsup_n X_n \geqslant T) = 1$ for all T. Hence $\mathbb{P}(\limsup_n X_n = +\infty) = 1$.

Example (Very well approximable numbers):

Definition 10.17. *A real number $\alpha \in [0, 1]$ is called very-well approximable (VWA) if there exists $\epsilon > 0$ and infinitely many $q \in \mathbb{Z}\backslash\{0\}$ such that $\|q\alpha\| < 1/q^{1+\epsilon}$, where $\|x\| := \inf_{n\in\mathbb{Z}} |x - n|$.*

Proposition 10.18. *Lebesgue almost every $\alpha \in [0, 1]$ is not VWA.*

Proof. Fix $\epsilon > 0$. Let $\Omega = [0, 1]$, $\mathbb{P}$ Lebesgue measure, $\mathcal{F}$ the Borel σ-algebra. Let $A_q := \{\alpha \in [0, 1], \|q\alpha\| < 1/q^{1+\epsilon}\}$. Then $\mathbb{P}(A_q) \leqslant q1/q^{2+\epsilon}$, so $\sum_{q\geqslant 1} \mathbb{P}(A_q) < \infty$ and hence, by the first Borel-Cantelli lemma, $\mathbb{P}(\limsup_q A_q) = \mathbb{P}(\text{“}\alpha \text{ is VWA”}) = 0$. □

Here are three useful probabilistic inequalities:

(1) Cauchy-Schwarz: Let X, Y be two (real valued) random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then

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

(2) Markov's inequality: If $X \geqslant 0$ is a non-negative random variable, and $\lambda \geqslant 0$, then
$$\lambda \mathbb{P}(X \geqslant \lambda) \leqslant \mathbb{E}(X),$$
(3) Chebychev's inequality: Let Y be a random variable, and $\lambda \geqslant 0$, then
$$\lambda^2 \mathbb{P}(|Y - \mathbb{E}(Y)| \geqslant \lambda) \leqslant \mathbf{Var}(Y).$$

The proofs are straigtforward. Recall that the proof of Cauchy-Schwarz is a consequence of the fact that $\mathbb{E}((t|X| + |Y|)^2)$ is a quadratic polynomial in t, which is always non-negative: this implies that the discriminant $\Delta = 4((\mathbb{E}(|XY|)^2 - \mathbb{E}(X^2)\mathbb{E}(Y^2))$ is $\leqslant 0$, which yields the Cauchy-Schwarz inequality.

The proof of Markov's inequality is clear: $\mathbb{E}(X) \geqslant \mathbb{E}(X 1_{X \geqslant \lambda}) \geqslant \lambda \mathbb{E}(1_{X \geqslant \lambda}) = \lambda \mathbb{P}(X \geqslant \lambda)$. And the proof of Chebychev's follows from Markov's applied to $X = (Y - \mathbb{E}(Y))^2$ and recalling that the variance $\mathbf{Var}(Y)$ is equal to $\mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = \mathbb{E}(X)$.

The Markov and Chebychev inequalities are extermely useful to get upper bounds on the probability that a random variable is large, or deviates from its mean by a certain amount. Indeed it is often easier to estimate the expectation $\mathbb{E}(X)$ or the variance $\mathbf{Var}(Y)$, rather than to directly compute the probabilities on the left hand side of these inequalities.

We now prove a landmark result from probability theory, the law of large numbers:

Theorem 10.19 (Strong law of large numbers)**.** *Let $(X_n)_{n \geqslant 1}$ be a sequence of i.i.d. (i.e. independent and identically distributed) random variables with common law μ (= a Borel measure on $\mathbb{R}$). Assume that $\int_{\mathbb{R}} |x| d\mu(x) = \mathbb{E}(|X_1|)$ is finite. Then, almost surely,*
$$\frac{1}{n} \sum_{i=1}^{n} X_i \to_{n \to +\infty} \mathbb{E}(X_1) = \int_{\mathbb{R}} x d\mu(x).$$

Proof. We give a short proof under the additional assumption that X_1 has a finite moment of order 4. The proof assuming only a finite moment of order 1 is much more involved. We will eventually give a proof of it at the end of the course, as a corollary of the pointwise ergodic theorem. So let us assume that $\mathbb{E}(X_1^4) < \infty$.

Without loss of generality, we may assume that $\mathbb{E}(X_1) = 0$. Indeed we may set $Y_i = X_i - \mathbb{E}(X_1)$ and apply the result to Y_i instead (note that this is legitimate, because the Y_i's will also be i.i.d. and will have a finite moment of order 4, because $\mathbb{E}(Y_1^4) \leqslant \mathbb{E}(Y_1^2)^2 \leqslant \mathbb{E}(X_1^2)^2 < \infty$ by Cauchy-Schwarz.

Note further that X_1 has a finite moment of order 1, 2 and 3, i.e. X_1, X_1^2 and X_1^3 are integrable. This is because $\mathbb{E}(|X_1|) \leqslant \sqrt{\mathbb{E}(X_1^2)}$ by Cauchy-Schwarz, while $\mathbb{E}(X_1^2) \leqslant \sqrt{\mathbb{E}(X_1^4)}$ and $\mathbb{E}(X_1^3) \leqslant \sqrt{\mathbb{E}(X_1^2)\mathbb{E}(X_1^4)}$. (in fact it is not hard to prove that finiteness of a moment of order k implies finiteness of all moments of order $\leqslant k$).

Then we set $S_n = \sum_1^n X_i$ and compute
$$\mathbb{E}(S_n^4) = \sum_{i,j,k,l} \mathbb{E}(X_i X_j X_k X_l).$$
Exploiting independence of the X_i's and the fact that $\mathbb{E}(X_i) = 0$, we see that all terms vanish, except of the terms X_i^4 and the cross terms $X_i^2 X_j^2$. This leads to
$$\mathbb{E}(S_n^4) = \sum_1^n \mathbb{E}(X_i^4) + 6 \sum_{i<j} \mathbb{E}(X_i^2 X_j^2)$$

By Cauchy-Schwarz again, we have $\mathbb{E}(X_i^2 X_j^2) \leqslant \sqrt{\mathbb{E}(X_i^4)\mathbb{E}(X_j^4)} = \mathbb{E}(X_1^4)$. So

$$\mathbb{E}(S_n^4) \leqslant (n + 3n(n-1))\mathbb{E}(X_1^4)$$

and hence

$$\mathbb{E}((S_n/n)^4) = O(1/n^2),$$

which yields the convergence of the series

$$\sum_n \mathbb{E}((S_n/n)^4) = \mathbb{E}(\sum_n (S_n/n)^4).$$

Clearly this means that almost surely the series $\sum_n (S_n/n)^4$ converges, and hence that almost surely S_n/n tends to 0 as desired. □

The word "strong" refers to the fact that the convergence holds almost surely (i.e. $\mathbb{P}$-almost everywhere). There is also a weak law of large numbers, in which the convergence holds in a weaker sense (in probability) under a weaker assumption on the sequence $(X_n)_n$, see the Example sheet.

Lecture 13

11. Convergence of random variables

Given a sequence of random variables $(X_n)_n$ there are several different and non equivalent ways in which one may express the fact that they converge. The weakest type of convergence is called the "weak convergence" or "convergence in law".

Definition 11.1. *A sequence of probability measures μ_n on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is said to converge weakly to a measure μ if for every continuous and bounded function f on $\mathbb{R}^d$ we have:*

$$\mu_n(f) \to_{n\to+\infty} \mu(f). \tag{11.1}$$

Examples:

(1) $\mu_n = \delta_{1/n}$ is a sequence of Dirac masses at $1/n$. Clearly μ_n converges weakly to δ_0, because $f(1/n) \to f(0)$ by continuity of f.
(2) $\mu_n = \mathcal{N}(0, \sigma_n^2)$ a centered gaussian distribution with standard deviation σ_n such that $\sigma_n \to 0$. Then $\mu_n \to \delta_0$ as well, as follows from dominated convergence:

$$\mu_n(f) = \int f(x) \exp(-x^2/2\sigma_n)\sqrt{2\pi\sigma_n^2}dx = \int f(\sigma_n x) \exp(-x^2/2)/\sqrt{2\pi}dx \to f(0).$$

(3) $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{\frac{i}{n}}$ converges weakly to the uniform distribution on $[0,1]$.

Definition 11.2. *A sequence of $\mathbb{R}^d$-valued random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be converging to X*

(a) almost surely (or a.s.) if for $\mathbb{P}$-almost every $\omega \in \Omega$ we have $X_n(\omega) \to X(\omega)$,
(b) in probability (or in measure) if for all $\epsilon > 0$, as $n \to +\infty$,

$$\mathbb{P}(\|X_n - X\| > \epsilon) \to 0,$$

(c) in law (or in distribution) if μ_{X_n} converges weakly to μ_X.

Recall that μ_X is the law of the random variable X, namely the Borel measure on $\mathbb{R}$ defined by $\mu_X(B) = \mathbb{P}(X \in B)$. Here the norm $\|x\|$ is the Euclidean norm on $\mathbb{R}^d$ (or any other norm for that matter). We now show that $(a) \Rightarrow (b) \Rightarrow (c)$.

Proposition 11.3. *If a sequence of random variables $(X_n)_n$ converges almost surely to X, then it converges in probability to X. If a sequence converges in probability, then it converges in distribution.*

Proof. $(a) \Rightarrow (b)$. Write:

$$\mathbb{P}(\|X_n - X\| > \epsilon) = \mathbb{E}(1_{\|X - X_n\| > \epsilon}) \to_{n\to+\infty} 0$$

by Dominated Convergence.

For $(b) \Rightarrow (c)$, first note (recall Part IB) that every continuous and bounded function f on $\mathbb{R}^d$ is uniformly continuous on compact subsets. Recall that this means that for every $\epsilon > 0$ there is $\delta > 0$ such that if $\|x\| \leqslant 1/\epsilon$ and $\|y - x\| \leqslant \delta$, then $|f(x) - f(y)| \leqslant \epsilon$. So

$$\begin{aligned}
|\mu_{X_n}(f) - \mu_X(f)| &= |\mathbb{E}(f(X_n)) - \mathbb{E}(f(X))| \\
&\leqslant \mathbb{E}(1_{\|X_n - X\| \leqslant \delta} 1_{|X|<1/\epsilon} |f(X_n) - f(X)|) + 2\|f\|_\infty \mathbb{E}(1_{|X_n - X| \geqslant \delta} + 1_{\|X\| \geqslant 1/\epsilon}) \\
&\leqslant \epsilon + 2\|f\|_\infty (\mathbb{P}(\|X - X_n\| \geqslant \delta) + \mathbb{P}(\|X\| \geqslant 1/\epsilon))
\end{aligned}$$

where we have denoted as usual $\|f\|_\infty := \sup_{x\in\mathbb{R}} |f(x)|$. Letting n tend to infinity, this gives:

$$\limsup_n |\mu_{X_n}(f) - \mu_X(f)| \leqslant \epsilon + 2\|f\|_\infty \mathbb{P}(|X| \geqslant 1/\epsilon).$$

Finally letting $\epsilon \to 0$, we obtain the desired conclusion (weak convergence of μ_{X_n} towards μ_X.

□

Remark 11.4. When $d = 1$, i.e. for real valued random variables, there is a further characterization of convergence in law in terms of the distribution functions $F_X(x) = \mathbb{P}(X \leqslant x)$. Namely: $X_n \to X$ in law if and only if $F_{X_n}(x) \to F_X(x)$ for every x where $F_X(x)$ is continuous. See the 3rd Example Sheet.

Remark 11.5. To prove that a sequence of probability measures μ_n converges weakly to a probability measure μ, it is enough to check (11.1) for smooth and compactly supported functions f on $\mathbb{R}^d$ (exercise! or wait for the 4th ex. sheet).

Remark 11.6. The converse statements to those of the last proposition do not hold. For an example showing that (c) does not imply (b) consider a sequence of i.i.d. random variables X_n with common law μ and assume that μ is not a Dirac mass. Clearly μ_{X_n} converges to μ_X, because $\mu_{X_n} = \mu_X = \mu$ for all n. However $\mathbb{P}(\|X_n - X_0\| > \epsilon)$ is independent of $n \geqslant 1$ and non-zero if ϵ is small enough, because μ is not concentrated on a single point.

To see an example showing that (b) does not imply (a) consider the moving bump examples already discussed: $\Omega = (0,1)$, $\mathbb{P}$ Lebesgue, $\mathcal{F} = \mathcal{B}(0,1)$ and $X_{k,n} = 1_{[k/n,(k+1)/n]}$ for $k = 0, ..., n-1$. This is an array of random variables, that you may of course organize in a single sequence Y_m so that $X_{k+1,n}$ comes right after $X_{k,n}$ and $X_{0,n+1}$ right after $X_{n-1,n}$. Then it is clear that for every $\omega \in \Omega$ the sequence $Y_m(\omega)$ does not converge (it will take the value 0 and the value 1 infinitely often). But $\mathbb{P}(|Y_m| > \epsilon)$ tends to 0 as m tends to infinity, because $\mathbb{P}(|X_{k,n}| > \epsilon) = 1/n$ if $\epsilon \in (0,1)$.

Proposition 11.7. *If $X_n \to X$ in probability, then there is a subsequence $\{n_k\}_k$ such that*

$$X_{n_k} \longrightarrow_{k \to +\infty} X$$

almost surely.

Proof. The assumption says that $\mathbb{P}(\|X_n - X\| > \epsilon) \to 0$ as $n \to +\infty$. So for every $k \in \mathbb{N}$, there is n_k such that $\mathbb{P}(\|X_{n_k} - X\| > 1/k) \leqslant 1/2^k$. Hence

$$\sum_k \mathbb{P}(\|X_{n_k} - X\| > 1/k) < \infty$$

and by the first Borel-Cantelli lemma we know that with probability 1

$$\#\{k \in \mathbb{N}, \|X_{n_k} - X\| > 1/k\} < \infty$$

Hence $\lim_{k \to +\infty} \|X_{n_k} - X\| = 0$. □

We have seen so far three types of convergence: in law, in probability and almost sure. Here is a fourth:

Definition 11.8. *A sequence of integrable random variables $(X_n)_n$ is said to converge to X in $\mathbf{L}^1$ if*

$$\mathbb{E}(\|X_n - X\|) \to_{n \to +\infty} 0.$$

[Recall that a random variable Y is said to be integrable if $\mathbb{E}(|Y|) < \infty$.]

This notion is stronger than convergence in probability, but it does not imply almost sure convergence, nor is it implied by almost sure convergence. In fact we will soon examine in detail the difference between convergence in probability and convergence in $\mathbf{L}^1$, this will rely on the notion of uniform integrability.

Proposition 11.9. *If $X_n \to X$ in $\mathbf{L}^1$, then $X_n \to X$ in probability.*

Proof. This is clear from the Markov inequality:

$$\mathbb{P}(\|X_n - X\| > \epsilon) \leqslant \frac{1}{\epsilon}\mathbb{E}(\|X_n - X\|).$$

□

Remark 11.10. The converse is not true: again take $\Omega = (0,1)$, $\mathcal{F} = \mathcal{B}(0,1)$ and $\mathbb{P}$ Lebesgue measure, and $X_n := n1_{[0,1/n]}$. Clearly $X_n \to 0$ almost surely and hence in probability, but $\mathbb{E}(X_n) = 1$ for all n.

Note however that in this example X_n is not bounded. If X_n is bounded (i.e. $|X_n| \leqslant C$ for some $C > 0$ independent of n), then convergence in probability implies convergence in $\mathbf{L}^1$ (and hence is equivalent to it): indeed if not there would be $\epsilon > 0$ and a subsequence $\{n_k\}_k$ such that $\mathbb{E}(\|X_{n_k} - X\|) > \epsilon$ for all k. But one may then pass to an even finer subsequence that convergences almost surely by Proposition 11.7. This would contradict the Dominated Convergence Theorem. We are now going to define the right necessary and sufficient condition for us to be able to upgrade convergence in probability (or almost sure convergence) to convergence in $\mathbf{L}^1$.

Definition 11.11. *A sequence of integrable ($\mathbb{R}^d$-valued) random variables $(X_n)_n$ is said to be uniformly integrable (or U.I. for short) if*

$$\lim_{M\to+\infty} \limsup_{n\to+\infty} \mathbb{E}(\|X_n\|1_{\|X_n\|>M}) = 0.$$

Examples

(1) If $(X_n)_n$ is *dominated* in the sense that there is a random variable $Y \geqslant 0$, which is integrable, and such that

$$\|X_n\| \leqslant Y$$

for all n, then $(X_n)_n$ is uniformly integrable. Indeed

$$\mathbb{E}(\|X_n\|1_{\|X_n\|>M}) \leqslant \mathbb{E}(Y1_{Y>M})$$

and the right hand side tends to 0 as $M \to +\infty$ by Dominated Convergence.

(2) Given $p \in [1, +\infty]$, we say that a sequence $(X_n)_n$ is bounded in $\mathbf{L}^p$ if

$$\sup_n \mathbb{E}(\|X_n\|^p) < \infty.$$

If $(X_n)_n$ is bounded in $\mathbf{L}^p$ for some $p > 1$, then $(X_n)_n$ is uniformly integrable. Indeed, we may write:

$$\mathbb{E}(\|X_n\|1_{\|X_n\|>M}) \leqslant \frac{1}{M^{p-1}}\mathbb{E}(\|X_n\|^p) \leqslant \frac{1}{M^{p-1}} \sup_n \mathbb{E}(\|X_n\|^p)$$

But $p > 1$, so the right hand side tends to 0 as $n \to +\infty$.

Lecture 14

The main reason for introducing the notion of uniform integrability lies in the following theorem.

Theorem 11.12. *Let $(X_n)_n$ be a sequence of ($\mathbb{R}^d$-valued) integrable random variables. Let X be another random variable. Then the following are equivalent:*

(i) X is integrable and $X_n \to X$ in $\mathbf{L}^1$,
(ii) $(X_n)_n$ is uniformly integrable and $X_n \to X$ in probability.

We will need:

Lemma 11.13. *If Y is an integrable random variable and $(X_n)_n$ is U.I., then so is $(X_n + Y)_n$.*

Proof. This follows from the following calculation:

$$\begin{aligned}\mathbb{E}(\|X_n + Y\|1_{\|X_n+Y\|\geqslant M}) &\leqslant \mathbb{E}((\|X_n\| + \|Y\|)1_{\|X_n+Y\|\geqslant M}(1_{\|X_n\|\geqslant M/2} + 1_{\|X_n\|<M/2}))\\ &\leqslant \mathbb{E}((x_n + y)1_{x_n\geqslant M/2}) + \mathbb{E}((x_n + y)1_{y\geqslant M/2}1_{x_n<M/2})\\ &\leqslant \mathbb{E}((x_n + y)1_{x_n\geqslant M/2}) + \mathbb{E}(2y1_{y\geqslant M/2})\\ &\leqslant \mathbb{E}(2x_n1_{x_n\geqslant M/2}) + \mathbb{E}(3y1_{y\geqslant M/2})\end{aligned}$$

where we wrote x_n for $\|X_n\|$ and y for $\|Y\|$. In the second line, we used the triangle inequality so that $1_{\|X_n+Y\|\geqslant M}1_{\|X_n\|<M/2} \leqslant 1_{\|Y\|\geqslant M/2}1_{\|X_n\|<M/2}$. □

Proof of Theorem 11.12. $(i) \Rightarrow (ii)$. We know that $X_n \to X$ in probability, because it converges in $\mathbf{L}^1$ (see Remark 11.9). But $(X_n - X)_n$ is clearly U.I., because it tends to 0 in $\mathbf{L}^1$. By Lemma 11.13, we conclude that $(X_n)_n$ is U.I.

$(ii) \Rightarrow (i)$ Let us first check that X must be integrable. By Proposition 11.7, there is a subsequence $\{n_k\}_k$ such that $X_{n_k} \to X$ almost surely. By Fatou's lemma we get:

$$\mathbb{E}(\|X\|1_{\|X\|\geqslant M}) \leqslant \liminf_{k\to+\infty} \mathbb{E}(\|X_{n_k}\|1_{\|X_{n_k}\|\geqslant M}),$$

which by assumption tends to 0 as M tends to $+\infty$ and is in particular finite. Hence $\mathbb{E}(\|X\|) \leqslant \mathbb{E}(\|X\|1_{\|X\|\geqslant M}) + M < \infty$.

By Lemma 11.13 we have that $(X_n - X)_n$ is U.I. Now assume by way of contradiction that X_n does not converge to X in $\mathbf{L}^1$. Then there is $\epsilon > 0$ and a subsequence $\{n_k\}_k$ such that $\mathbb{E}(\|X_{n_k} - X\|) > \epsilon$ for all k. By Proposition 11.7 (given that X_n tends to X in probability), we may pass to an even finer subsequence and assume wlog that $X_{n_k} \to X$ as $k \to +\infty$. Since $(X_n - X)_n$ is U.I. there is $M > 0$ such that

$$\limsup_k \mathbb{E}(\|X_{n_k} - X\|1_{\|X_{n_k}-X\|>M}) < \epsilon$$

But by Dominated Convergence:

$$\limsup_k \mathbb{E}(\|X_{n_k} - X\|1_{\|X_{n_k}-X\|\leqslant M}) = 0$$

We conclude that

$$\limsup_k \mathbb{E}(\|X_{n_k} - X\|) < \epsilon,$$

which is a contradiction.

□

12. $\mathbf{L}^p$ SPACES

We begin with three inequalities of fundamental importance in Analysis.

(1) Jensen's inequality

Proposition 12.1 (Jensen's inequality)**.** *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and I an open interval of $\mathbb{R}$ and $X : \Omega \to I$ be a random variable. Assume that X is integrable and $\phi : I \to \mathbb{R}$ is convex. Then*

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

Recall:

Definition 12.2. *A function $\phi : I \to \mathbb{R}$ is said to be convex if $\forall x, y \in I$ and for all $t \in [0,1]$ one has:*

$$\phi(tx + (1-t)y) \leqslant t\phi(x) + (1-t)\phi(y).$$

Remark 12.3. $\mathbb{E}(X) \in I$ clearly and we will show that $\mathbb{E}(\phi(X)^-) < \infty$, where ϕ^- is the negative part of ϕ. So $\mathbb{E}(\phi(X))$ is always well defined as $\mathbb{E}(\phi(X)^+) - \mathbb{E}(\phi(X)^-)$ even though $\phi(X)$ may not be integrable.

In order to prove Jensen's inequality we require:

Lemma 12.4. *A function $\phi : I \to \mathbb{R}$ is convex if and only if*

$$\phi = \sup_{\ell \in \mathcal{F}} \ell$$

where $\mathcal{F}$ is some family of affine linear forms $x \mapsto ax + b$ (for various $a, b \in \mathbb{R}$).

Proof. The "if" part is clear, because each ℓ is convex, and hence so is their supremum. For the "only if" part note that for each $x_0 \in I$ we need to find $\ell_{x_0}(x) = \theta_{x_0}(x - x_0) + \phi(x_0)$ with $\theta_{x_0} \in \mathbb{R}$ such that $\phi(x) \geqslant \ell_{x_0}(x)$ for all x. Observe that since ϕ is convex, for all x, y with $x < x_0 < y$ we have

$$\frac{\phi(x_0) - \phi(x)}{x_0 - x} \leqslant \frac{\phi(y) - \phi(x_0)}{y - x_0},$$

indeed this is equivalent to $\phi(x_0) \leqslant t\phi(x) + (1-t)\phi(y)$ for $t = (x_0 - x)/(y - x)$. So there exists $\theta \in \mathbb{R}$ such that for all x, y with $x < x_0 < y$ we have:

$$\frac{\phi(x_0) - \phi(x)}{x_0 - x} \leqslant \theta \leqslant \frac{\phi(y) - \phi(x_0)}{y - x_0}.$$

So just set $\ell_{x_0}(x) = \theta(x - x_0) + \phi(x_0)$ and get $\phi(x) \geqslant \ell_{x_0}(x)$. □

We can now prove Jensen's inequality:

Proof of Proposition 12.1. Write

$$\phi(X) = \sup_{\ell \in \mathcal{F}} \ell(X)$$

and for all $\ell \in \mathcal{F}$,

$$\mathbb{E}(\phi(X)) \geqslant \mathbb{E}(\ell(X)) = \ell(\mathbb{E}(X))$$

So

$$\mathbb{E}(\phi(X)) \geqslant \phi(\mathbb{E}(X))$$

as desired.

Besides, $-\phi(x) = \inf_{\ell \in \mathcal{F}} -\ell(x)$, so

$$(-\phi(X))^+ \leqslant |-\ell(X)| \leqslant |\ell(X)| \leqslant a|X| + b,$$

so $\phi(X)^-$ is integrable.

□

Lecture 15

Definition 12.5 ($\mathbf{L}^p$-norm)**.** *Let $(X, \mathcal{A}, \mu)$ be a measure space and $f : X \to \mathbb{R}$ a measurable map. For $p \in [1, +\infty)$, we set*

$$\|f\|_p := \Big[\int_X |f|^p d\mu\Big]^{1/p},$$

while for $p = \infty$

$$\|f\|_\infty = essup|f| := \inf\{t \geqslant 0, |f(x)| \leqslant t \ \ \mu - a.e.\}.$$

Example: if $(X, \mathcal{A}, \mu) = (\mathbb{R}, \mathcal{L}, m)$ and $f = 1 + 1_{x=0}$, then $\sup_{x\in X} |f(x)| = 2$, but $\|f\|_\infty = 1$.

Remark 12.6. Note that for any measurable f, there is another measurable function g such that $f = g$ μ-a.e. and $\|f\|_\infty = \sup_{x\in X} |g(x)|$. Indeed take $g(x) = f(x)1_{|f(x)|\leqslant\|f\|_\infty}$.

(2) Minkowski's inequality

Proposition 12.7 (Minkowski's inequality)**.** *In this setting, for any $p \in [1, +\infty]$ we have:*

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

Proof. The case when $p = +\infty$ is obvious. So assume p is finite. Note that the inequality is equivalent to

$$\|t\frac{f}{\|f\|_p} + (1 - t)\frac{g}{\|g\|_p}\|_p \leqslant 1$$

where

$$t := \frac{\|f\|_p}{\|f\|_p + \|g\|_p}.$$

This means that we need to show that for every $t \in [0, 1]$ and every measurable functions F and G on X with $\|F\|_p = \|G\|_p = 1$ we have:

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

or in other words

$$\int_X |tF + (1 - t)G|^p d\mu \leqslant 1.$$

But $x \mapsto x^p$ is convex in $[0, +\infty)$ if $p \geqslant 1$. So we have:

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

Integrating over X, we obtained as desired:

$$\int_X |tF + (1 - t)G|^p d\mu \leqslant 1.$$

□

(3) Hölder's inequality

Let $p, q \in [1, +\infty]$. Assume that

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

Proposition 12.8 (Hölder's inequality)**.** *Let $(X, \mathcal{A}, \mu)$ be a measure space and let f, g be measurable functions. Then*

$$\int_X |fg| d\mu \leqslant \|f\|_p \cdot \|g\|_q.$$

Moreover, when p and q are finite, if equality holds (and the terms are finite), then there is $(\alpha, \beta) \neq (0, 0)$ such that $\alpha|f|^p = \beta|g|^q$ μ-amost everywhere.

The case when either p or q is infinite is obvious, by the positivity property of the μ-integral: $|fg| \leqslant \|f\|_\infty |g|$ holds μ-a.e., hence $\mu(|fg|) \leqslant \|f\|_\infty \mu(|g|)$. So wlog we can now assume that p and q are finite. For the proof, we need:

Lemma 12.9 (Young's inequality for products)**.** *Let $a, b \geqslant 0$, then*

$$ab \leqslant \frac{a^p}{p} + \frac{b^q}{q},$$

with equality if and only if $a^p = b^q$.

Proof. the function $-\log$ is strictly convex, so

$$-\log(\frac{a^p}{p} + \frac{b^q}{q}) \leqslant \frac{1}{p}(-\log)(a^p) + \frac{1}{q}(-\log)(b^q) = -\log ab,$$

with equality if and only if $a^p = b^q$. □

Proof of Proposition 12.8. Without loss of generality we may assume that $\int_X |f|^p d\mu \neq 0$ and that $\int_X |g|^q d\mu \neq 0$. Otherwise either f or g is zero μ-almost everywhere and the inequality is trivial. Similarly, we may further assume that both $\|f\|_p$ and $\|g\|_q$ are finite. By further multiplying f and g by a scalar multiple, we may assume that $\|f\|_p = 1$ and $\|g\|_q = 1$. Now we are in a position to apply Young's inequality for products (previous lemma):

$$|fg| \leqslant \frac{|f|^p}{p} + \frac{|g|^q}{q}. \tag{12.1}$$

And integrating, we get:

$$\int_X |fg| d\mu \leqslant \frac{1}{p} + \frac{1}{q} = 1$$

as desired. To see the equality case, note that it implies that (12.1) holds μ-almost everywhere, and thus that $|f|^p = |g|^q$ almost everywhere.

□

Remark 12.10.

(1) When $p = q = 2$ Hölder's inequality is also known as Cauchy-Schwarz (which we have already seen for random variables, but it also holds when μ is not a probability measure).

(2) Jensen's inequality implies that if X is a random variable, then the function

$$p \mapsto \mathbb{E}(|X|^p)^{1/p}$$

is non-decreasing. Indeed set $\phi(x) = x^{q/p}$, which is convex if $q > p$ and apply Jensen's inequality to it.

Definition 12.11. *We set $\mathcal{L}^p(X, \mathcal{A}, \mu) := \{f : X \to \mathbb{R}$ measurable, $\|f\|_p < \infty\}$.*

Note that $\mathcal{L}^p(X, \mathcal{A}, \mu)$ is a vector space (this follows from Minkowski's inequality when p is finite).

Let us introduce the following relation among measurable functions on a measure space $(X, \mathcal{A}, \mu)$. We will write $f \equiv g$ is $f(x) = g(x)$ holds μ-almost everywhere.

Lemma 12.12. *The relation $\equiv$ is an equivalence relation, which is stable under addition and multiplication.*

Proof. Clearly if $f \equiv g$ and $g \equiv h$, then $f \equiv h$, so the relation in transitive and hence an equivalence relation. It is also clear that if $f' \equiv g'$, then $f + f' \equiv g + g'$ and $ff' \equiv gg'$. □

Definition 12.13. *The $\mathbf{L}^p$-space associated to the measure space $(X, \mathcal{A}, \mu)$ is the quotient space $\mathcal{L}^p(X, \mathcal{A}, \mu) \mod \equiv$.*

In other words it is the set of equivalence classes $[f]$ (up to μ-measure zero) of functions f with finite $\mathcal{L}^p$ norm. The main reason why we pass to this quotient is that on the quotient the $\mathcal{L}^p$ norm becomes a genuinue norm. The value $\|f\|_p$ depends only on the class $[f]$ of the function and not on the function itself, i.e. if $f \equiv g$, then $\|f\|_p = \|g\|_p$. So it makes sense to talk about $\|[f]\|_p$ and we have that $\|[f]\|_p = 0$ implies $[f] = 0$. Thus we have:

Proposition 12.14 (Completeness of $\mathbf{L}^p$-spaces)**.** *For $p \in [1, +\infty]$ the norm $\|\cdot\|_p$ turns $\mathbf{L}^p(X, \mathcal{A}, \mu)$ into a normed vector space. Moreover it is a complete normed vector space (a.k.a. a Banach space).*

Recall that complete means that Cauchy sequences converge.

Proof. If $f \equiv g$, then $\|f\|_p = \|g\|_p$, to $\|\cdot\|_p$ descends to $\mathbf{L}^p$.

If $\|f\|_p = 0$, then $f \equiv 0$ by the properties of the μ-integral.

The triangle inequality holds $\|f + g\|_p \leqslant \|f\|_p + \|g\|_p$, by Minkowski's inequality when p is finite and clearly when $p = +\infty$. Also $\|\lambda f\|_p = |\lambda| \|f\|_p$ for all $\lambda \in \mathbb{R}$.

So all this makes $\mathbf{L}^p$ a normed vector space. It remains to show completeness. We first handle the case when p is finite. So suppose $([f_n])_n$ is a Cauchy sequence in $\mathbf{L}^p$. This means that for all $\epsilon > 0$ there is $n_0 > 0$ such that for all $n, m > n_0$ we have $\|f_n - f_m\|_p < \epsilon$. We set $\epsilon = 2^{-k}$ for an arbitrary integer $k \geqslant 0$. Then there is n_k such that for all $k \geqslant 0$

$$\|f_{n_{k+1}} - f_{n_k}\| \leqslant 2^{-k}.$$

Let

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

Then

$$\|S_K\|_p \leqslant \sum_{k=1}^{K} 2^{-k} \leqslant 1,$$

by Minkowski's inequality. So by Monotone Convergence we get

$$\int_X |S_K(x)|^p d\mu(x) \to_{K \to +\infty} \int_X |S_\infty|^p d\mu$$

and so $S_\infty \in \mathcal{L}^p$ and for μ-a.e. x we have $S_\infty(x) < \infty$. This means that

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

and hence that $(f_{n_k}(x))_k$ is itself a Cauchy sequence in $\mathbb{R}$. But $\mathbb{R}$ is complete, so $\lim_{k \to +\infty} f_{n_k}(x)$ exists in $\mathbb{R}$. Call it $f(x)$. This was defined only on those x such that $S_\infty(x) < \infty$. On the complement of this set (which is of μ-measure 0) we can set $f(x) = 0$. Then

$$\|f_n - f\|_p \leqslant \liminf_{k \to +\infty} \|f_n - f_{n_k}\|_p \leqslant \epsilon$$

by Fatou's lemma. This holds for all $n \geqslant n_0$. So we have shown that

$$\lim_n \|f_n - f\|_p = 0.$$

Finally the case when $p = +\infty$ can be handled in a similar way, except that in place of Fatou's lemma we use the following:

Fact: if $f_n \to f$ μ-a.e., then $\|f\|_\infty \leqslant \liminf_n \|f_n\|_\infty$.
Proof: Let $t > \liminf_n \|f_n\|$. Then there exists an increasing subsequence n_k with $\|f_{n_k}\|_\infty \leqslant t$. In other words for all k, μ-a.e. $|f_{n_k}(x)| \leqslant t$. Swapping the order ($\forall k \mu$-a.e. versus μ-a.e. $\forall k$, which is allowed by σ-subadditivity of the measure μ), this implies that μ-a.e. for all k we have $|f_{n_k}(x)| \leqslant t$. And hence that $|f(x)| \leqslant t$.
□

From now on, we will often abuse notation and stop distinguishing between an element of $\mathcal{L}^p$ and a representative of that element, namely a function in $\mathcal{L}^p$ whose equivalence class is that element. This should not cause confusion, but we should always keep in mind that elements in $\mathbf{L}^p$ are equivalence classes.

The next proposition is a useful technical device when dealing with $\mathbf{L}^p$-spaces, when p is finite (it fails when $p = +\infty$).

Proposition 12.15 (Approximation by simple functions)**.** *Let $p \in [1, \infty)$ and $(X, \mathcal{A}, \mu)$ a measure space. Let V be the linear span of simple functions on $(X, \mathcal{A})$. Then $V \cap \mathbf{L}^p$ is dense in $\mathbf{L}^p$.*

in other words: for every $\epsilon > 0$ and every $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ there is $g = g^+ - g^-$ with g^+ and g^- simple functions on $(X, \mathcal{A})$ such that $g \in \mathcal{L}^p$ and $\|f - g\|_p < \epsilon$.

Proof. Note that $g^+, g^- \leqslant |g|$. Hence if $g \in \mathcal{L}^p$, then g^+ and g^- belong to $\mathcal{L}^p$. Writing $f = f^+ - f^-$ and using Minkowski's inequality it is enough to assume that $f \geqslant 0$. We've shown (cf. Lemma 7.6) that there are simple functions g_n with $0 \leqslant g_n \leqslant f$ and $g_n(x) \to f(x)$ for every $x \in X$. But we have

$$|g_n - f|^p \leqslant (2f)^p$$

and the right hand side is integrable, so we may apply Lebesgue's Dominated Convergence Theorem to conclude that

$$\int_X |g_n - f|^p d\mu \to 0.$$

□

Remark 12.16. When $(X, \mathcal{A}, \mu) = (\mathbb{R}^d, \mathcal{L}, m)$, then smooth compactly supported functions $C_c^\infty(\mathbb{R}^d)$ form a dense subspace in $\mathbf{L}^p$ (see the 3rd example sheet).

Remark 12.17.
(1) If $\mu(X) < \infty$, then $\mathbf{L}^{p'} \subset \mathbf{L}^p$ if $p' \geqslant p$ (we have already seen how this follows from Jensen's inequality),
(2) if X is discrete and countable (i.e. $\mathcal{A} = 2^X$), then $\mathbf{L}^{p'} \subset \mathbf{L}^p$ for $p' \leqslant p$,
(3) in general (e.g. when $(X, \mathcal{A}) = (\mathbb{R}^d, \mathcal{L})$) these inclusions do not hold (in neither direction).

Lecture 16

13. Hilbert spaces and $\mathbf{L}^2$-methods

Let V be a complex vector space (possibly infinite dimensional).

Definition 13.1 (inner product)**.** *A* Hermitian inner product *on V is a map:*

$$V \times V \to \mathbb{C}$$
$$(x, y) \mapsto \langle x, y\rangle$$

with the following properties

(i) $\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$ *for all* $\alpha, \beta \in \mathbb{C}$ *and all* $x, y, z \in V$*,*
(ii) $\langle y, x\rangle = \overline{\langle x, y\rangle}$ *for all* $x, y \in V$*,*
(iii) for all $x \in V$*,* $\langle x, x\rangle \geqslant 0$ *with equality if and only if* $x = 0$*.*

Axioms (i) and (ii) make the inner product a *sesquilinear form* ("sesqui" means one-and-a-half, it is linear in the first variable and skew-linear in the second, that is $\langle x, \alpha y\rangle = \overline{\alpha}\langle x, y\rangle$, where $\overline{\alpha}$ is the complex conjugate). Note that (ii) implies that $\langle x, x\rangle$ is always real.

For real vector spaces, one has the same definition, but of course in that case (i) and (ii) simply mean that the inner product is a bilinear symmetric form, and the inner product is then called a *Euclidean inner product.*

In what follows V is a complex (resp. real) vector space endowed with a Hermitian (resp. Euclidean) inner product. For each $x \in V$ we set $\|x\| := \langle x, x\rangle$.

Lemma 13.2. *For any* $\alpha \in \mathbb{C}$ *and* $x, y \in V$ *we have:*

(a) $\|\alpha x\| = |\alpha|\|x\|$*,*
(b) (Cauchy-Schwarz inequality) $|\langle x, y\rangle| \leqslant \|x\| \cdot \|y\|$*,*
(c) (triangle inequality) $\|x + y\| \leqslant \|x\| + \|y\|$*,*
(d) (Parallelogram identity) $\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2)$*.*

Proof. For (a) note that $\|\alpha x\|^2 = \langle \alpha x, \alpha x\rangle = |\alpha|^2\|x\|^2$. The proof of (b) is as follows: for every $t \in \mathbb{R}$ we have:

$$\langle x + ty, x + ty\rangle = t^2\|y\|^2 + \|x\|^2 + 2\,\mathrm{Re}\langle x, y\rangle \geqslant 0,$$

this is a quadratic polynomial in t that does not change sign: its discriminant must henceforth be non-positive, i.e. $\Delta \leqslant 0$. But

$$\Delta = 4(\mathrm{Re}(\langle x, y\rangle))^2 - 4\|x\|^2 \cdot \|y\|^2$$

so we get

$$|\,\mathrm{Re}(\langle x, y\rangle)| \leqslant \|x\| \cdot \|y\|.$$

But for every $\theta \in \mathbb{R}$ we have $\langle e^{i\theta}x, y\rangle = e^{i\theta}\langle x, y\rangle$, so in particular, given x, y there always exists some $\theta \in \mathbb{R}$ such that $\langle e^{i\theta}x, y\rangle = |\langle x, y\rangle|$. We conclude that

$$|\langle x, y\rangle| = |\,\mathrm{Re}(\langle e^{i\theta}x, y\rangle)| \leqslant \|e^{i\theta}x\| \cdot \|y\| = \|x\| \cdot \|y\|$$

as desired. To prove (c) apply (b) as follows:

$$\|x + y\|^2 \leqslant \|x\|^2 + \|y\|^2 + 2\,\mathrm{Re}(\langle x, y\rangle) \leqslant (\|x\| + \|y\|)^2.$$

The proof of (d) is straighforward: simply expand the inner product of the sum and of the difference. □

Corollary 13.3. $(V, \|\cdot\|)$ *is a* normed vector space.

Recall that a normed vector space is simply a (real or complex) vector space endowed with a *norm*, i.e. a map $x \mapsto \|x\|$ from V to $[0, +\infty)$ satisfying axioms (a) and (c) above, and such that $\|x\| = 0$ if and only if $x = 0$.

A normed vector space is in particular a metric space, with the distance function defined as $\|x - y\|$. Recall further that a metric space is said to be *complete* if all Cauchy sequences converge.

Definition 13.4. *A hermitian (resp. Euclidean) vector space V is said to be a* Hilbert space *if $(V, \|\cdot\|)$ is complete.*

Example: Let $V = \mathbf{L}^2(X, \mathcal{A}, \mu)$ with inner product

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

This is well-defined, because as we have already seen (Cauchy-Schwarz) if both f, g are in $\mathbf{L}^2$, then $f\overline{g}$ is μ-integrable. It is plain to check that V then becomes a Hermitian vector space with this inner product. And we have shown previously that V is complete. So V is a *Hilbert space*. In fact it is the archetypal Hilbert space, as it can be shown that every Hilbert space is isomorphic to an $\mathbf{L}^2$ space.

Proposition 13.5 (orthogonal projection on closed convex sets)**.** *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{C} \subset \mathcal{H}$ a* closed convex *subset. Then there is a well-defined orthogonal projection to $\mathcal{C}$. This means that for every $x \in \mathcal{H}$ there is a unique $y \in \mathcal{C}$ such that*

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

We call y the *orthogonal projection* of x onto $\mathcal{C}$.

Proof. Pick $c_n \in \mathcal{C}$ such that $\|x - c_n\| \to d(x, \mathcal{C})$. By the parallelogram identity we have:

$$\|\frac{x - c_n}{2} + \frac{x - c_m}{2}\|^2 + \|\frac{x - c_n}{2} - \frac{x - c_m}{2}\|^2 = 2(\frac{\|x - c_n\|^2}{4} + \frac{\|x - c_m\|^2}{4}) \quad (13.1)$$

in other words:

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

Since $\mathcal{C}$ is convex, $\frac{c_n + c_m}{2} \in \mathcal{C}$, so we get:

$$\|x - \frac{c_n + c_m}{2}\| \geqslant d(x, \mathcal{C})$$

and it follows that $\|c_n - c_m\| \to 0$ as n and m tend to infinity. Hence the sequence $(c_n)_n$ is a Cauchy sequence in $\mathcal{H}$.

But we have assumed that $\mathcal{H}$ is complete. We conclude that the sequence $(c_n)_n$ converges to a point $y \in \mathcal{H}$. Since $\mathcal{C}$ is closed by assumption, we must have $y \in \mathcal{C}$ and $d(x, \mathcal{C}) = \|x - y\|$.

This shows the existence of y. The uniqueness follows directly from (13.1) replacing c_n and c_m by two minimizing elements. □

Corollary 13.6. *If $V \leqslant \mathcal{H}$ is a closed vector subspace, then $\mathcal{H} = V \oplus V^\perp$, where $V^\perp := \{x \in \mathcal{H}, \langle x, v\rangle = 0 \forall v \in V\}$.*

Note: even if V is not closed, $V^\perp$ is always a closed subspace. Indeed if $x_n \to x$, then $\langle x_n, v\rangle \to \langle x, v\rangle$, so

$$|\langle x_n, v\rangle - \langle x, v\rangle| \leqslant \|x_n - x\| \cdot \|v\|$$

and $x \in V^\perp$ if $x_n \in V^\perp$ for all n.

Proof. Note that $V \cap V^\perp = 0$, because $\langle x, x\rangle = 0$ implies $x = 0$.

Let now $x \in \mathcal{H}$ and y its projection to V as given by Proposition 13.5 Claim: $x - y \in V^\perp$. Indeed for all $z \in V$ we have $\|x - y - z\| \geqslant \|x - y\|$ as $y + z \in V$. So

$$\|x - y\|^2 + \|z\|^2 - 2\,\mathrm{Re}\langle x - y, z\rangle \geqslant \|x - y\|^2$$

and hence

$$2\,\mathrm{Re}\langle x - y, z\rangle \leqslant \|z\|^2$$

for all $z \in V$. In particular for all $t > 0$,

$$2\,\mathrm{Re}\langle x - y, tz\rangle \leqslant t^2\|z\|^2,$$

which letting $t \to 0$ yields:

$$\mathrm{Re}\langle x - y, z\rangle \leqslant 0.$$

But this holds for all $z \in V$ so in particular for $-z$, and hence $\mathrm{Re}\langle x - y, z\rangle = 0$. Changing z into $e^{i\theta}$ for a suitable angle $\theta \in \mathbb{R}$, we finally get $\langle x - y, z\rangle = 0$. Hence $x - y \in V^\perp$. □

Let $\mathcal{H}$ be a complex Hilbert space.

Definition 13.7. *A linear form $\ell : \mathcal{H} \to \mathbb{C}$ is called bounded if $\exists c > 0$ such that*

$$|\ell(x)| \leqslant c\|x\|$$

for all $x \in \mathcal{H}$.

Remark: a linear form is bounded if and only if it is continuous (an easy exercise!). Of course if $\mathcal{H}$ is a real Hilbert space, then linear forms are assumed to be $\mathbb{R}$-linear only and take values in $\mathbb{R}$.

Theorem 13.8 (Riesz representation theorem for Hilbert spaces)**.** *Let $\mathcal{H}$ be a Hilbert space and ℓ a bounded linear form on $\mathcal{H}$. Then there is a unique vector $v_0 \in \mathcal{H}$ such that*

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

for all $x \in \mathcal{H}$.

Proof. Uniqueness is clear, because if v_0 and v_0' are such, then by linearity of the inner product $\langle x, v_0 - v_0'\rangle = 0$ for all $x \in \mathcal{H}$ and in particular for $x = v_0 - v_0'$, which yields $v_0 - v_0' = 0$. We now prove the existence part. By the last corollary we have $\mathcal{H} = \ker\ell \oplus (\ker\ell)^\perp$, because $\ker\ell$ is a closed subspace of $\mathcal{H}$ (since ℓ is continuous). We may assume that ℓ is not identically zero (otherwise set $v_0 = 0$). Pick $x_0 \in (\ker\ell)^\perp \backslash \{0\}$. Then $\ell(x_0) \neq 0$ and

Claim: $(\ker\ell)^\perp = \mathbb{C}x_0$.

Indeed if $x \in (\ker\ell)^\perp$, then $\ell(x) = \alpha\ell(x_0)$ with $\alpha := \frac{\ell(x)}{\ell(x_0)}$. So $\ell(x - \alpha x_0) = 0$, that is $x - \alpha x_0 \in \ker\ell \cap (\ker\ell)^\perp = \{0\}$. Hence $x = \alpha x_0$.

Now write:

$$\ell(x) - \langle x, x_0\rangle\frac{\ell(x_0)}{\|x_0\|^2} = \ell(x) - \langle x, v_0\rangle,$$

where $v_0 := x_0\frac{\overline{\ell(x_0)}}{\|x_0\|^2}$. This linear form vanishes on $\ker\ell$ and on x_0, so also on $(\ker\ell)^\perp$ by the Claim above. Hence on all of $\mathcal{H}$.

□

Lecture 17

14. Conditional expectation

We now define a fundamental concept in probability theory, that of *conditional expectation*. Again the key to the proofs will be the existence of an orthogonal projection in Hilbert space as established in the previous lectures.

Proposition-Definition 14.1 (Conditional expectation)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{G} \subset \mathcal{F}$ be a σ-subalgebra. Let X be a real valued integrable random variable. Then there exists Y a $\mathcal{G}$-measurable and integrable random variable such that*

$$\mathbb{E}(1_A X) = \mathbb{E}(1_A Y) \tag{14.1}$$

for all $A \in \mathcal{G}$. Moreover Y is unique in the sense that if Y' is as above, then $Y = Y'$ almost surely. The random variable Y is called the conditional expectation of X with respect to $\mathcal{G}$ and is denoted by

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

Note that $\mathbb{E}(X|\mathcal{G})$ is a ($\mathcal{G}$-measurable) random variable (while $\mathbb{E}(X)$ was just a number). Intuitively $\mathbb{E}(X|\mathcal{G})$ is the average value of X "*knowing*" $\mathcal{G}$, that is given the information provided by $\mathcal{G}$. A good way to understand the idea of conditional expectation is to consider the special case when $\mathcal{G}$ is the Boolean algebra generated by a partition of the universe Ω into finitely many subsets from $\mathcal{F}$, namely $\Omega = \bigsqcup_1^N X_i$. Then $\mathbb{E}(X|\mathcal{G})$ is $\mathcal{G}$-measurable, so it is constant on each X_i. On X_i it equals the average value of $X(\omega)$ knowing that ω belongs to X_i, that is

$$Y(\omega) = \mathbb{E}(X|\mathcal{G})(\omega) = \frac{1}{\mathbb{P}(X_i)} \int X(\omega) 1_{X_i}(\omega) d\mathbb{P}(\omega).$$

Indeed it is very easy to check that (14.1) does hold for this Y and for any $A \in \mathcal{G}$, because A is then a finite union of X_i's.

Proof. (existence) We first prove the existence of conditional expectation assuming that X has a finite moment of order 2. In that case Y will be the orthogonal projection of X onto the closed subspace $\mathbf{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ of the Hilbert space $\mathbf{L}^2(\Omega, \mathcal{G}, \mathbb{P})$. The fact that $\mathbf{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ is closed follows from its completeness (this is an exercise in Exple Sheet no 4). Recall that on this Hilbert space the inner product between say W and Z is given by $\mathbb{E}(WZ)$. It is then clear that $\mathbb{E}(1_A Y) = \mathbb{E}(1_A X)$.

Now assume that X is integrable and non-negative. Then we can truncate and consider $X_n = X 1_{X \leqslant n}$. Then $X_n \in \mathbf{L}^2$ and we can let Y_n the orthogonal projection of X_n onto $\mathbf{L}^2(\Omega, \mathcal{G}, \mathbb{P})$. It is clear that $Y_{n+1} \geqslant Y_n \geqslant 0$ almost surely (indeed, for example if we let A be the event where $Y_{n+1} < Y_n$, then A is $\mathcal{G}$-measurable, and we get $\mathbb{E}((Y_{n+1} - Y_n)1_A) = \mathbb{E}((X_{n+1} - X_n)1_A) \geqslant 0$, which forces $(Y_{n+1} - Y_n)1_A = 0$ almost surely, or in other words $\mathbb{P}(A) = 0$). So we can denote by $Y = \lim_n Y_n$ and observe that my Monotone Convergence $\mathbb{E}(Y_n 1_A) \to \mathbb{E}(Y 1_A)$ and hence $\mathbb{E}(Y 1_A) = \mathbb{E}(X 1_A)$. This shows the existence in this case.

In the general case, we may write $X = X^+ - X^-$ and set $Y = Y^+ - Y^-$, where Y^+ is a conditional expectation for X^+ and Y^- for X^-.

(uniqueness) If Y_1 and Y_2 are two candidates, then $\mathbb{E}(1_A(Y_1 - Y_2)) = 0$ for all $A \in \mathcal{G}$. But this forces $Y_1 = Y_2$ almost surely (this was an exercise in Example sheet no 2.). □

We now list the key properties of conditional expectation:

(1) (linearity) if $\alpha, \beta \in \mathbb{R}$ and X, Y are random variables then almost surely

$$\mathbb{E}(\alpha X + \beta Y|\mathcal{G}) = \alpha \mathbb{E}(X|\mathcal{G}) + \beta \mathbb{E}(Y|\mathcal{G}),$$

(2) if X is already $\mathcal{G}$-measurable, then $\mathbb{E}(X|\mathcal{G}) = X$ a.s.
(3) (positivity) if $X \geqslant 0$ a.s., then $\mathbb{E}(X|\mathcal{G}) \geqslant 0$ a.s.,
(4) if $\mathcal{H} \subset \mathcal{G}$ is a sub-σ-algebra, then a.s.

$$\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(X|\mathcal{H}),$$

(5) if Z is a $\mathcal{G}$-measurable bounded random variable, then a.s.

$$\mathbb{E}(XZ|\mathcal{G}) = Z \cdot \mathbb{E}(X|\mathcal{G}),$$

(6) (independence) if X is independent from $\mathcal{G}$, then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$ a.s.,
(7) the Monotone Convergence Theorem, Fatou's lemma and the Dominated Convergence Theorem continue to hold for $\mathbb{E}(\cdot|\mathcal{G})$. (for example the analogue of the MCT states that if $X_{n+1} \geqslant X_n \geqslant 0$ and X_n converges almost surely to X, then $\mathbb{E}(X_n|\mathcal{G})$ converges almost surely to $\mathbb{E}(X|\mathcal{G})$, etc.)

Proof. The proofs are a simple application of Proposition 14.1. One uses uniqueness and the defining property (14.1) to check that the properties hold almost surely. For example to prove (3), set $A = \{\omega \in \Omega, Y(\omega) < 0\}$, where $Y = \mathbb{E}(X|\mathcal{G})$; then $\mathbb{E}(1_A Y) = \mathbb{E}(1_A X) \geqslant 0$, which forces $1_A Y = 0$ a.s., and hence $\mathbb{P}(A) = 0$. Item (6) follows from the fact that $\mathbb{E}(X1_A) = \mathbb{E}(X)\mathbb{E}(1_A)$ as X and 1_A will be independent if $A \in \mathcal{G}$. And the same for (7) : for $A \in \mathcal{G}$ write $\mathbb{E}(1_A X_n) = \mathbb{E}(1_A \mathbb{E}(X_n|\mathcal{G}))$, then use the ordinary MCT on both sides to conclude that $\mathbb{E}(1_A X_n)$ converges to $\mathbb{E}(1_A X)$ and that $\mathbb{E}(1_A \mathbb{E}(X_n|\mathcal{G}))$ converges to $\mathbb{E}(1_A \lim_n \mathbb{E}(X_n|\mathcal{G}))$. The uniqueness in Propostion 14.1 then implies that $\lim_n \mathbb{E}(X_n|\mathcal{G})$ must be (almost surely) the conditional expectation $\mathbb{E}(X|\mathcal{G})$. Fatou's lemma and the DCT then follow from the MCT by the same argument as in their original proof. □

15. The Fourier transform on $\mathbb{R}^d$

We now take a break from probability theory to go back to analysis on $\mathbb{R}^d$ and present a fundamental tool: the Fourier transform on $\mathbb{R}^d$. This tool will be crucial to establish later one of the corner stones of probability theory, namely the Central Limit theorem. It will also be crucial when discussing the gaussian distribution and gaussian vectors.

Recall that $\mathcal{B}(\mathbb{R}^d)$ is the σ-algebra of Borel subsets of $\mathbb{R}^d$. We denote by dx the Lebesgue measure on $\mathbb{R}^d$.

Definition 15.1. *Let $f \in \mathbf{L}^1(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), dx)$ and let $u \in \mathbb{R}^d$. We set*

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

where $\langle u, x\rangle = \sum_i u_i x_i$ is the standard Euclidean inner product. The function $\widehat{f}$ is called the Fourier transform *of f.*

Proposition 15.2. *(a) $|\widehat{f}(u)| \leqslant \|f\|_1$,*
(b) $u \mapsto \widehat{f}(u)$ is continuous.

Proof. a) is clear, b) follows from directly from the Dominated Convergence Theorem: if $u_n \to u$, then the functions $x \mapsto f(x)e^{iu_n x}$ converge pointwise to $f(x)e^{iux}$ and are dominated by the integrable function f. □

Definition 15.3. *Similarly if μ is a finite Borel measure on $\mathbb{R}^d$ and $u \in \mathbb{R}^d$ we set*

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

We call it the characteristic function of μ.

Again $|\widehat{\mu}(u)| \leqslant \mu(\mathbb{R}^d)$ and $u \mapsto \widehat{\mu}(u)$ is continuous for the same reasons.

Example: Let $\mu = \mathcal{N}(0,1)$ be a normalized (i.e. mean 0, standard deviation 1) gaussian measure, so that $d\mu(x) = g(x)dx$, where

$$g(x) := \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}).$$

Claim: $\widehat{g}(u) = \widehat{\mu}(u) = \exp(-\frac{u^2}{2}) = \sqrt{2\pi} g(u)$.

Proof. We can write

$$\widehat{g}(u) = \int_{\mathbb{R}} e^{iux} e^{-\frac{x^2}{2}} \frac{dx}{\sqrt{2\pi}}.$$

Differentiating (under the integral sign, which is legitimate by Corollary 7.15) with respect to u and integrating by parts, we obtain:

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

so

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

which implies that

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

But

$$\widehat{g}(0) = \int g(x)dx = 1,$$

so

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

□

This shows that the gaussian is *self-dual*, namely it is equal to its Fourier transform up to a scaling factor: $\widehat{g} = \sqrt{2\pi} g$. It can be shown that this property *characterizes* the gaussian distribution among all Borel probability measures on $\mathbb{R}$.

Example: If $\mu = \mathcal{N}(0, I_d)$ is an isotropic multivariate gaussian (here I_d is the $d \times d$ identity matrix, we will explain the rationale for this notation in a few lectures), in other words:

$$d\mu = g(x_1) \cdot \ldots \cdot g(x_d) dx_1 \ldots dx_d = G(x)dx,$$

where

$$G(x) = \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2}(x_1^2 + \ldots + x_d^2)). \tag{15.1}$$

Then

$$\widehat{G}(u) = \int_{\mathbb{R}^d} G(x) e^{i\langle u,x\rangle} dx_1 \ldots dx_d = \prod_1^d \int_{\mathbb{R}} g(x_i) e^{iux_i} dx_i = \exp(-\frac{1}{2}\|u\|^2)$$

where $\|u\|^2 = u_1^2 + \ldots + u_d^2$.

Lecture 18

Theorem 15.4 (Fourier Inversion Formula). *(1) If μ if a finite Borel measure on $\mathbb{R}^d$ such that $\widehat{\mu} \in \mathbf{L}^1(\mathbb{R}^d)$, then μ has a continuous density $\phi(x)$ with respect to Lebesgue measure, i.e. $d\mu = \phi(x)dx$, and*

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

(2) If $f \in \mathbf{L}^1(\mathbb{R}^d)$ is such that $\widehat{f} \in \mathbf{L}^1(\mathbb{R}^d)$, then

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

for Lebesgue almost every x.

Note that $x \mapsto \frac{1}{(2\pi)^d}\widehat{\widehat{f}}(-x)$ is a continuous function, being the Fourier transform of an $\mathbf{L}^1$ function.

Interpretation: this theorem says that the function f can be decomposed as a weighted sum, or integral, of "Fourier modes", i.e. the oscillatory functions $x \mapsto e^{i\langle u,x\rangle}$,

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

and the "weights" $\widehat{f}(u)$ are called the Fourier coefficients of f. The functions $\chi_u(x) := e^{-i\langle u,x\rangle}$ are called modes in physics and characters in mathematics. Their defining property is that they are group homomorphisms from $(\mathbb{R}^d, +)$ to the circle group $\{z \in \mathbb{C}, |z| = 1\}$, namely $\chi_u(x + y) = \chi_u(x)\chi_u(y)$ for all $x, y \in \mathbb{R}^d$.

Proof. (1) Without loss of generality, we may assume that μ is a probability measure (replace μ by $\mu/\mu(\mathbb{R}^d)$). And that μ is the law of some random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The main idea of the proof is to use gaussians to "mollify" X (that is replace X by the "smoother" distribution $X + \sigma N$ for an independent gaussian N and let σ tend to 0) and exploit the self-duality property of the gaussian distribution. To achieve this, let N be an independent random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, which is assumed to be a normalized standard gaussian $\mathcal{N}(0, I_d)$ with density $G(x)$ (defined in the Example above)

We need to show that for every A is a bounded Borel subsets of $\mathbb{R}^d$, we have:

$$\mathbb{P}(X \in A) = \int_{\mathbb{R}^d} 1_A(z)\phi(z)dz.$$

Let $h := 1_A$ and $\sigma > 0$. By the Dominated Convergence Theorem we have:

$$\lim_{\sigma\to 0} \mathbb{E}(h(X + \sigma N)) = \mathbb{E}(h(X)). \tag{15.2}$$

On the other hand:

$$\mathbb{E}(h(X+\sigma N)) = \mathbb{E}(\int h(X+\sigma x)G(x)dx) = \mathbb{E}(\int\int h(X+\sigma x)G(u)e^{-i\langle u,x\rangle}\frac{du}{(2\pi)^{d/2}}dx)$$

because $G(x) = \int_{\mathbb{R}^d} G(u)e^{-i\langle u,x\rangle}\frac{du}{(2\pi)^{d/2}}$ as follows from the computation in the above Example. Then setting $z = X + \sigma x$ we get

$$\begin{aligned}\mathbb{E}(h(X+\sigma N)) &= \mathbb{E}\Big[\int\int h(z)G(u)e^{-i\langle \frac{u}{\sigma}, z-X\rangle}\frac{du}{(2\pi\sigma^2)^{d/2}}dz\Big]\\ &= \int\int h(z)G(u)e^{i\langle \frac{u}{\sigma},z\rangle}\widehat{\mu_X}(-\frac{u}{\sigma})\frac{du}{(2\pi\sigma^2)^{d/2}}dz\\ &= \int\int h(z)G(\sigma u)e^{-i\langle u,z\rangle}\widehat{\mu_X}(u)\frac{du}{(2\pi)^{d/2}}dz.\end{aligned}$$

We have used Fubini to interchange $\mathbb{E}$ and $\int\int$ at the second line. This was legitimate because $h(z)G(u)e^{-i\langle \frac{u}{\sigma},z-X\rangle}$ is an integrable (w.r.t. $dz\otimes du\otimes d\mathbb{P}$) function. Now the integrand is dominated, because

$$|h(z)G(\sigma u)e^{-i\langle u,z\rangle}\widehat{\mu_X}(u)| \leqslant \frac{1}{(2\pi)^{d/2}}|\widehat{\mu_X}(u)|$$

which is integrable by assumption. So we may apply the Dominated Convergence Theorem: letting $\sigma\to 0$ and noting that $G(0) = 1/(2\pi)^{d/2}$, we conclude that:

$$\lim_{\sigma\to 0}\mathbb{E}(h(X+\sigma N)) = \int\int h(z)e^{-i\langle u,z\rangle}\widehat{\mu_X}(u)\frac{du}{(2\pi)^d}dz = \int h(z)\widehat{\widehat{\mu_X}}(-z)\frac{dz}{(2\pi)^d}.$$

Comparing this to (15.2) ends the proof.

(2) The proof of part (2) is entirely similar: write $f = f^+ - f^-$ and $f(x)dx = a\,d\mu_X(x) - b\,d\mu_Y(x)$ for some $a, b \geqslant 0$ and some independent $\mathbb{R}^d$-valued random variables X, Y. So that $a\,d\mu_X = f^+(x)dx$ and $b\,d\mu_Y(x) = f^-(x)dx$. One needs to show that

$$\int h(z)f(z)dz = \int h(z)\frac{1}{(2\pi)^d}\widehat{\widehat{f}}(-z)dz$$

for every bounded measurable $h \geqslant 0$. Perform the same proof as in (1) writing

$$\int h(z)f(z)dz = a\mathbb{E}(h(X)) - b\mathbb{E}(h(Y)),$$

then replacing X by $X+\sigma N$ and Y by $Y+\sigma N$, and compute

$$\lim_{\sigma\to 0} a\mathbb{E}(h(X+\sigma N)) - b\mathbb{E}(h(Y+\sigma N))$$

in two ways. At the end, when applying the Dominated Convergence theorem, use the integrability of $\widehat{f} = a\widehat{\mu_X} - b\widehat{\mu_Y}$ in place of that of $\widehat{\mu_X}$ (note that there is no reason for the latter to be integrable). □

Remark 15.5. We see from the above theorem the importance of the assumption $\widehat{f} \in \mathbf{L}^1$. How can one ensure that this is the case in practice? Well, it is enough that f has enough integrable derivatives. In the case of univariate functions a simple sufficient condition for this is to require that f be C^2 and f,f' and f'' be integrable (i.e. in $\mathbf{L}^1(\mathbb{R})$, we suppose here that we are over $\mathbb{R}$, but a similar condition can be given over $\mathbb{R}^d$). To see this, note that if f is C^1 and $f' \in \mathbf{L}^1$, then $\widehat{f}(u) = \frac{i}{u}\widehat{f'}(u)$. Indeed, integrating by parts, we have:

$$\widehat{f}(u) = \int f(x)e^{iux}dx = \frac{1}{iu}\int f(x)\frac{d}{dx}(e^{iux})dx = -\frac{1}{iu}\int f'(x)e^{iux}dx.$$

It follows that

$$|\widehat{f}(u)| \leqslant \frac{1}{|u|}\|f'\|_1.$$

Iterating this fact, we conlude that is f,f' and f'' are in $\mathbf{L}^1$, then $\widehat{f}(u) = -\frac{1}{u^2}\widehat{f''}(u)$, to

$$|\widehat{f}(u)| \leqslant \frac{1}{u^2}\|f''\|_1$$

and hence $\widehat{f} \in \mathbf{L}^1(\mathbb{R})$.

We now pass to an important operation one can make on functions or measures:

Definition 15.6 (Convolution product)**.** *Given two Borel measures on $\mathbb{R}^d$, say μ and ν, we may define their convolution $\mu * \nu$ as the image of the product measure $\mu \otimes \nu$ under the addition law on $(\mathbb{R}^d, +)$, namely:*

$$\mu * \nu = \Phi_*(\mu \otimes \nu)$$

where

$$\Phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$$
$$(x, y) \mapsto x + y$$

Example: If $\mu = \mu_X$ and $\nu = \mu_Y$ are the laws of two independent $\mathbb{R}^d$-valued random variables X and Y, then $\mu * \nu$ is the law of the random variable $X + Y$, indeed:

$$\mathbb{P}(X + Y \in A) = \mu_X \otimes \mu_Y(\Phi^{-1}(A)).$$

Definition 15.7. *Similarly, if f and g are in $\mathbf{L}^1(\mathbb{R}^d)$, we may define*

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

This is well-defined for (Lebesgue) almost every x, because the map

$$(x, t) \mapsto f(x - t)g(t)$$

belongs to $\mathbf{L}^1(\mathbb{R}^d \times \mathbb{R}^d)$, since

$$\int\int |f(x - t)g(t)|dtdx = \|f\|_1\|g\|_1 < \infty.$$

Therefore, by Fubini, $f * g(x)$ is well-defined and is finite for m-almost every x. Also

$$\|f * g\|_1 = \int |f * g(x)|dx \leqslant \|f\|_1\|g\|_1.$$

One says that $\mathbf{L}^1(\mathbb{R}^d)$ endowed with the convolution product $*$ is a *Banach algebra*.

Remark 15.8. If μ, ν have densities with respect to Lebesgue, that is $\mu = f dx$ and $\nu = g dx$ for some integrable densities f and g, then $\mu * \nu$ also has a density with respect to Lebesgue equal to $f * g$.

Proposition 15.9 (Gaussian approximation)**.** *If $f \in \mathbf{L}^p(\mathbb{R}^d)$ and $p \in [1, +\infty)$, then*

$$\lim_{\sigma \to 0} \|f * G_\sigma - f\|_p = 0,$$

where $G_\sigma(x) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp(-\frac{\|x\|^2}{2\sigma^2})$ is the density of a gaussian distribution $\mathcal{N}(0, \sigma^2 I_d)$.

To prove this, we need a lemma:

Lemma 15.10 (Continuity of translation in $\mathbf{L}^p$)**.** *Let $f \in \mathbf{L}^p(\mathbb{R}^d)$ and $p \in [1, +\infty)$, then*

$$\lim_{t \to 0} \|\tau_t(f) - f\|_p = 0,$$

where $\tau_t(f)(x) = f(x + t)$ is the "translation" by $t \in \mathbb{R}^d$.

Proof. This is an exercise in the 4th Example sheet. Use the density of $C_c(\mathbb{R}^d)$ in $\mathbf{L}^p$ (which itself follows from Ex. 13 in the 3rd Example sheet). □

Proof of Proposition 15.9. We can write

$$f * G_\sigma(x) - f(x) = \int G_\sigma(t)(f(x-t) - f(x))dt = \mathbb{E}(f(x - \sigma N) - f(x))$$

where N is a normalized gaussian $\mathcal{N}(0,1)$. Hence by Jensen's inequality (given that $x \mapsto x^p$ is convex)

$$|f * G_\sigma(x) - f(x)|^p \leqslant \mathbb{E}(|f(x - \sigma N) - f(x)|^p),$$

and hence

$$\|f * G_\sigma - f\|_p^p \leqslant \mathbb{E}(\|\tau_{-\sigma N}(f) - f\|_p).$$

By the lemma above we know that almost surely $\|\tau_{-\sigma N}(f) - f\|_p$ tends to 0 as $\sigma \to 0$. So by Dominated Convergence (licit because $\|\tau_{-\sigma N}(f) - f\|_p \leqslant 2\|f\|_p$) we get the desired conclusion. □

Lecture 19

Proposition 15.11. *(a) if μ, ν are Borel probability measures on $\mathbb{R}^d$, then $\widehat{\mu * \nu} = \widehat{\mu} * \widehat{\nu}$.*
(b) if $f, g \in \mathbf{L}^1(\mathbb{R}^d)$, then $\widehat{f * g} = \widehat{f}\widehat{g}$.

Proof. For (a) wlog we may assume that $\mu = \mu_X$ and $\nu = \mu_Y$ are the laws of two independent random variables. By definition, the law of $X+Y$ is precisely $\mu_X * \mu_Y$. But by independence of X and Y,

$$\mathbb{E}(e^{iu(X+Y)}) = \mathbb{E}(e^{iuX})\mathbb{E}(e^{iuY})$$

hence $\widehat{\mu_{X+Y}}(u) = \widehat{\mu_X}(u)\widehat{\mu_Y}(u)$ as desired. (b) reduces to (a) writing $f = f^+ - f^-$, $a\, d\mu = f^+(x)dx$ and $b\, d\nu = f^-(x)dx$ (where $a, b \geqslant 0$ are so that μ and ν are probability measures), doing the same for g and expanding the product. □

The Fourier transform yields a very handy criterion to check convergence in law of a sequence of random variables: it is equivalent to pointwise convergence of the Fourier transforms, namely:

Theorem 15.12 (Lévy's criterion)**.** *Let $(X_n)_{n\geqslant 1}$ and X be an $\mathbb{R}^d$-valued random variable. The following are equivalent:*

(i) $X_n \to X$ in law,
(ii) for all $u \in \mathbb{R}^d$, $\lim_{n\to+\infty} \widehat{\mu_{X_n}}(u) = \widehat{\mu_X}(u)$

In particular, if $\widehat{\mu_X} = \widehat{\mu_Y}$ for two random variables X and Y, then they coincide in law, i.e. $\mu_X = \mu_Y$.

Proof. $(i) \Rightarrow (ii)$ is by definition, because for each u, $x \mapsto e^{iux}$ is continuous and bounded. For the other direction, we need to show that for every continuous and bounded function g on $\mathbb{R}^d$ we have:

$$\mathbb{E}(g(X_n)) \to_{n\to+\infty} \mathbb{E}(g(X)).$$

By Ex. 3 in the 4th Example sheet, it is enough to prove this for every smooth and compactly supported function g on $\mathbb{R}^d$. But then both g and $\widehat{g}$ are in $\mathbf{L}^1(\mathbb{R}^d)$ (see Ex 6 in the 4th Example Sheet). So by Fourier inversion we get:

$$g(x) = \int \widehat{g}(u) e^{-i\langle u,x\rangle} \frac{du}{(2\pi)^d}$$

and thus

$$\mathbb{E}(g(X_n)) = \int \widehat{g}(u)\widehat{\mu_{X_n}}(-u)\frac{du}{(2\pi)^d}.$$

Now we can conclude by Dominated Convergence, since $|\widehat{\mu_{X_n}}(-u)| \leqslant 1$,

$$\lim_{n\to\infty} \mathbb{E}(g(X_n)) = \int \widehat{g}(u)\widehat{\mu_X}(-u)\frac{du}{(2\pi)^d}.$$

□

Example: Show that the $\mathcal{N}(m, \sigma^2)$ converges weakly to the Dirac mass δ_m as $\sigma \to 0$. [Wlog we may assume that $m = 0$. Then by Lévy's criterion, this boils down to showing that $\widehat{\mu_{X_\sigma}}(u) \to \widehat{\mu_X}(u)$ for every $u \in \mathbb{R}$, where X_σ is distributed according to $\mathcal{N}(0, \sigma^2)$ and $X = 0$ a.s. This is immediate, because $\widehat{\mu_{X_\sigma}}(u) = \exp(-\sigma^2 u^2/2)$.]

Remark 15.13 (Bochner's theorem)**.** The characteristic function $\widehat{\mu_X}$ of a random variable X is a continuous function equal to 1 at 0. It is easy to verify (exercise) that it is a positive definite function, namely that given any $u_1, \dots, u_N \in \mathbb{R}^d$ and scalars $t_1, \dots, t_N$ in $\mathbb{C}$ we have:

$$\sum_{i=1}^{N} t_i \overline{t_j} \widehat{\mu_X}(u_i - u_j)$$

is real and $\geqslant 0$. *Solomon Bochner* proved in the 1930's that this property characterizes the Fourier transform of Borel probability measures among all (complex valued) continuous functions of $\mathbb{R}^d$ equal to 1 at the origin.

The inversion formula given previously works under the assumptions that both f and its Fourier transform $\widehat{f}$ are integrable. It turns out that when f is square integrable (i.e. in $\mathbf{L}^2$), then one can make sense both of the Fourier transform and of the inversion formula even though neither f nor $\widehat{f}$ may be in $\mathbf{L}^1$. To do this we exploit the Hilbert space structure of $\mathbf{L}^2$ and the following is the main result:

Theorem 15.14 (Plancherel formula)**.** *(a) Let $f \in \mathbf{L}^1(\mathbb{R}^d) \cap \mathbf{L}^2(\mathbb{R}^d)$. Then $\widehat{f} \in \mathbf{L}^2(\mathbb{R}^d)$ and*

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

(b) if $f, g \in \mathbf{L}^1(\mathbb{R}^d) \cap \mathbf{L}^2(\mathbb{R}^d)$, then we have the so-called Plancherel formula*:*

$$\langle \widehat{f}, \widehat{g} \rangle_{\mathbf{L}^2(\mathbb{R}^d)} = (2\pi)^d \langle \widehat{f}, \widehat{g} \rangle_{\mathbf{L}^2(\mathbb{R}^d)}.$$

(c) The map

$$\begin{aligned} \mathcal{F} : \mathbf{L}^1(\mathbb{R}^d) \cap \mathbf{L}^2(\mathbb{R}^d) &\to \mathbf{L}^2(\mathbb{R}^d) \\ f &\mapsto \frac{1}{(2\pi)^{d/2}} \widehat{f} \end{aligned}$$

extends uniquely to a linear isometry of $\mathbf{L}^2(\mathbb{R}^d)$. Moreover $\mathcal{F} \circ \mathcal{F}(f)(x) = f(-x)$.

A "linear isometry" means that $\|\mathcal{F} f\|_2 = \|f\|_2$ for every $f \in \mathbf{L}^2(\mathbb{R}^d)$. It is that extension to all of $\mathbf{L}^2(\mathbb{R}^d)$ that we continue to call the Fourier transform, and the relation $\mathcal{F} \circ \mathcal{F}(f)(x) = f(-x)$ can be seen as the extension of the Fourier inversion formula to all of $\mathbf{L}^2(\mathbb{R}^d)$.

Proof. First we prove (a) and (b). Assume to begin with that $\widehat{f}$ and $\widehat{g}$ belong to $\mathbf{L}^1 \cap \mathbf{L}^2$. Compute:

$$\begin{aligned} \langle \widehat{f}, \widehat{g} \rangle &= \int \widehat{f}(u) \overline{\widehat{g}(u)} du = \int \int f(x) e^{i\langle u, x \rangle} \overline{\widehat{g}(u)} dx du \\ &= \int f(x) \overline{\widehat{\widehat{g}}(-x)} dx = (2\pi)^d \int f(x) \overline{g(x)} dx = (2\pi)^d \langle f, g \rangle \end{aligned}$$

where we have used the Fourier inversion formula in the second line. It was legitimate to swap the two integrals at the first line, because f and $\widehat{g}$ are integrable, so Fubini applies.

To handle the general case, we let $\sigma > 0$ and consider the convolution products $f_\sigma := f * G_\sigma$ and $g_\sigma := g * G_\sigma$, where G_σ is the density of a gaussian $\mathcal{N}(0, \sigma^2 I_d)$. In other words $G_\sigma(x) = \frac{1}{\sigma^d} G(\frac{x}{\sigma})$, where G is the density of the standard gaussian defined in (15.1). By Proposition 15.11, we may compute:

$$\widehat{f * G_\sigma} = \widehat{f} \widehat{G_\sigma} = \widehat{f} \exp(-\sigma \|u\|^2 / 2)$$

Clearly f_σ belongs to $\mathbf{L}^1 \cap \mathbf{L}^2$ (note that $\|\widehat{f}\|_\infty \leqslant \|f\|_1$) and so does $\widehat{g_\sigma}$. So by the above we conclude that for every $\sigma > 0$

$$\|\widehat{f_\sigma}\|_2^2 = (2\pi)^d \|f_\sigma\|_2^2.$$

But now by gaussian approximation (i.e. Lemma 15.9) we get:

$$\lim_{\sigma \to 0} \|f_\sigma\|_2 = \|f\|_2$$

and

$$\|\widehat{f_\sigma}\|_2^2 = \|\widehat{f}\widehat{G_\sigma}\|_2^2 = \int |\widehat{f}(u)|^2 \exp(-\sigma\|u\|^2/2) du$$

converges to $\|\widehat{f}\|_2^2$ as $\sigma \to 0$ by Monotone Convergence. We conclude that $\|\widehat{f}\|_2^2 = (2\pi)^d \|f\|_2^2$.

Now similarly

$$\langle \widehat{f_\sigma}, \widehat{g_\sigma} \rangle = \int \widehat{f}(u)\overline{\widehat{g}(u)} e^{-\sigma\|u\|^2}) du$$

converges $\int \widehat{f}(u)\overline{\widehat{g}(u)} du$ by Dominated Convergence (licit because $\widehat{f}\widehat{\widehat{g}}$ is integrable, given that $\widehat{f}$ and $\widehat{g}$ are in $\mathbf{L}^2$, cf. Cauchy-Schwarz). This ends the proof of (a) and (b).

We now turn to (c). The subspace $\mathbf{L}^1(\mathbb{R}^d) \cap \mathbf{L}^2(\mathbb{R}^d)$ is dense in $\mathbf{L}^2(\mathbb{R}^d)$. Indeed it contains the continuous and compactly supported functions $C_c(\mathbb{R}^d)$, which is already dense. So we can define $\mathcal{F}f$ in general as

$$\mathcal{F}f = \lim_n \mathcal{F}f_n$$

where $f_n \in \mathbf{L}^1 \cap \mathbf{L}^2$ and $f_n \to f$ in $\mathbf{L}^2$. This is well-defined, because on the one hand

$$\|\mathcal{F}f_n - \mathcal{F}f_m\|_2 = \|f_n - f_m\|_2$$

as follows from the Plancherel formula (part (a) of the proposition), which implies that $\mathcal{F}f_n$ is a Cauchy sequence, hence converges in $\mathbf{L}^2$. And on the other hand the limit does not depend on the choice of sequence $(f_n)_n$, because if $(f'_n)_n$ is another such, then

$$\|\mathcal{F}f_n - \mathcal{F}f'_n\|_2 = \|f_n - f'_n\|_2$$

so $\mathcal{F}f'_n$ and $\mathcal{F}f_n$ have the same limit. In the limit we get: $\|\mathcal{F}f\|_2 = \|f\|_2$.

Finally, from the Fourier Inversion Formula, we have

$$\mathcal{F} \circ \mathcal{F}f = f^\vee, \tag{15.3}$$

where $f^\vee(x) = f(-x)$ for every $f \in \mathbf{L}^1$ with $\widehat{f} \in \mathbf{L}^1$. But such functions are dense in $\mathbf{L}^2$ (they contain all of $C_c^\infty(\mathbb{R}^d)$ the smooth compactly supported functions). Hence (15.3) holds for all fucntions $f \in \mathbf{L}^2(\mathbb{R}^d)$. □

Remark 15.15 (smoothness/decay barter)**.** An important metamathematical fact to remember about the Fourier transform is that it exchanges smoothness for decay at infinity and vice versa. For example a very smooth (i.e. with many continuous derivatives) integrable function will have a Fourier transform that decays fast (polynomially with a degree that depends on the number of continuous derivatives) at infinity. Conversely if a function decays fast at infinity (e.g. is compactly supported), then its Fourier transform will be very smooth. The intuition behind this is as follows. The characters $x \mapsto e^{iux}$ are oscillatory functions that oscillate with frequency proportional to $1/u$. So if the decomposition of f as

$$f(x) = \int \widehat{f}(u) e^{-iux} du$$

has not too small Fourier coefficients $\widehat{f}(u)$ even for large values of u, then it means that the high frequencies occur a lot in the decomposition of f. And consequently f is not very smooth.

Remark 15.16 (uncertainty principle)**.** Further to this, it can be shown that it is impossible, unless f is identically zero, for f and its Fourier transform $\widehat{f}$ to be both compactly supported (f would be analytic and vanish on an open set, hence would be identically zero). There even is a (mathematical) *uncertainty principle* according to which if X is a random variable whose law has density $|f(x)|^2 \in \mathbf{L}^1(\mathbb{R})$, and Y is another random variable whose law has density $|\widehat{f}(x)|^2/(2\pi) \in \mathbf{L}^1(\mathbb{R})$, then

$$\mathbf{Var}X \cdot \mathbf{Var}Y \geqslant 1/(16\pi^2),$$

which can be interpreted as saying that X and Y cannot be both too localized.

Remark 15.17 (Schwartz space)**.** An interesting subspace of $\mathbf{L}^2(\mathbb{R}^d)$ is the space of smooth (i.e. C^∞) functions all of whose derivatives decay fast at infinity (i.e. faster that any polynomial). This is called the *Schwartz space*. And it can be shown that the Fourier transform $\mathcal{F}$ preserves the Schwartz space. See the Example sheet.

Lecture 20

16. Gaussian random variables

We now come back to probability theory, introduce gaussian random variables and their co-variance matrix, and then state and prove the Central Limit Theorem.

Definition 16.1. *An $\mathbb{R}^d$-valued random variable X is called* gaussian *if for every $u \in \mathbb{R}^d$ the inner product*

$$\langle X, u\rangle := u_1 X_1 + \ldots + u_d X_d$$

is a real valued gaussian random variable, i.e. has law $\mathcal{N}(m, \sigma^2)$ for some $m \in \mathbb{R}$ and some $\sigma \geqslant 0$. Recall the gaussian law $\mathcal{N}(m, \sigma^2)$ is the Borel probability measure on $\mathbb{R}$ whose density with respect to Lebesgue measure is given by

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-m)^2}{2\sigma^2})$$

if $\sigma > 0$ and when $\sigma = 0$ we take it to mean δ_m, the Dirac mass at m.

Example: if $N_1, \ldots, N_d$ are d independent normalized gaussians (distributed according to $\mathcal{N}(0,1)$), then $(N_1, \ldots, N_d)$ is a gaussian vector, because all linear combination $\alpha_1 N_1 + \ldots + \alpha_d N_d$ are gaussian (with mean zero and variance $\sum_j \alpha_j^2$, indeed it has the right characteristic function:

$$\mathbb{E}(\exp(iu(\sum_j \alpha_j N_j))) = \prod_1^d \mathbb{E}(\exp(iu\alpha_j N_j)) = \prod_1^d \exp(-u^2\alpha_j^2/2) = \exp(-(\sum_j \alpha_j^2)u^2/2).$$

$\mathbb{R}^d$-valued gaussian random variables are also called gaussian vectors. Their law is entirely characterized by their mean and their co-variance matrix:

Proposition 16.2. *The law of a gaussian vector is determined by*

(1) its mean: $\overline{X} := (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_d))$, and
(2) its covariance matrix: $(\mathbf{Cov}(X_i, X_j)_{1 \leqslant i,j \leqslant d}$

We often denote it by $\mathcal{N}(m, K)$, where $m \in \mathbb{R}^d$ is the mean, and $K \in M_d(\mathbb{R})$ is the covariance matrix.

The entries of the covariance matrix are the *correlation coefficients* between the coordinates of X, namely: $\mathbf{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))]$.

Proof. Let $\widehat{\mu_X}(u) := \mathbb{E}(e^{i\langle X,u\rangle})$ be the characteristic function of X. Clearly this is entirely determined by the family of laws $\mu_{\langle X,u\rangle}$ for u ranging in $\mathbb{R}^d$ (this fact holds for any random vector X). But as X is gaussian, we know that $\langle X, u\rangle$ is a real valued gaussian random variable. And the law of a real valued gaussian random variable is determined by its mean m and its variance σ^2. So the law of $\langle X, u\rangle$ is determined by $\mathbb{E}(\langle X, u\rangle) = \langle \overline{X}, u\rangle$ and by

$$\mathbf{Var}(\langle X, u\rangle) = \mathbb{E}(\langle X - \overline{X}, u\rangle^2) = \sum_{1 \leqslant i,j \leqslant d} u_i u_j \mathbf{Cov}(X_i, X_j). \tag{16.1}$$

□

Remark 16.3. Note that (16.1) shows that the covariance matrix of any random vector is positive semi-definite symmetric matrix.

The next proposition gives a way to construct an arbitrary gaussian vector out of d independent normalized real valued gaussians. It also shows that the image of a gaussian vector under an affine transformation of $\mathbb{R}^d$ is again a gaussian vector.

Proposition 16.4. *Let $N_1, \dots, N_d$ be d independent normalized gaussians each distributed as $\mathcal{N}(0,1)$. We write $\overrightarrow{N} = (N_1, \dots, N_d)$. Let $A \in M_d(\mathbb{R})$ a square $d \times d$ matrix and $\overrightarrow{b} \in \mathbb{R}^d$. Then*

$$A\overrightarrow{N} + \overrightarrow{b}$$

is a gaussian vector with mean $\overrightarrow{b}$ and covariance matrix AA^*. Moreover, given a gaussian vector X in $\mathbb{R}^d$, there is a vector $\overrightarrow{b} \in \mathbb{R}^d$ and a matrix A such that X and $AN + \overrightarrow{b}$ have the same law.

Proof. Note first that $A\overrightarrow{N} + \overrightarrow{b}$ is a gaussian vector, because for every $u \in \mathbb{R}^d$, $\langle u, A\overrightarrow{N} + \overrightarrow{b} \rangle$ is a linear combination of N_i's and a constant vector, so it is gaussian by the Example above. Clearly $\mathbb{E}(A\overrightarrow{N} + \overrightarrow{b}) = \overrightarrow{b}$. Moreover

$$\mathbf{Cov}(A\overrightarrow{N} + \overrightarrow{b}) = \mathbf{Cov}(A\overrightarrow{N}) = AA^*,$$

because by (16.1)

$$\begin{aligned}\langle u, \mathbf{Cov}(A\overrightarrow{N})u\rangle &= \mathbf{Var}(\langle A\overrightarrow{N}, u\rangle) = \mathbf{Var}(\langle \overrightarrow{N}, A^*u\rangle) \\ &= \sum_i (A^*u)_i^2 = \|A^*u\|^2 = \langle A^*u, A^*u\rangle = \langle u, AA^*u\rangle\end{aligned}$$

Finally if X is a gaussian vector, then set $\overrightarrow{b} = \mathbb{E}(X)$ and pick A so that $\mathbf{Cov}(X) = AA^*$. Then $A\overrightarrow{N} + \overrightarrow{b}$ will have the same law as X, because it has same mean and same covariance matrix. □

Remark 16.5. If $X = (N_1, \dots, N_d)$, then the covariance matrix of X is the identity matrix I_d. And the law of X is invariant under rotations. Indeed if $O \in O_d(\mathbb{R})$ is a rotation centered at the origin, then OX is again a gaussian vector, with identity covariance matrix. It can be shown that the only Borel probability laws on $\mathbb{R}^d$ invariant under rotation and with independent coordinates are gaussian laws $\mathcal{N}(0, \lambda I_d)$ for $\lambda \geqslant 0$. This provides a further interesting way in which the gaussian distribution arises naturally (see the 2019 exam...); for yet another see Exercise 17 in the 4th Example Sheet.

Remark 16.6. If $\det(A) \neq 0$, then $X = A\overrightarrow{N} + \overrightarrow{b}$ is a *non-degenerate* gaussian: its law has a density with respect to Lebesgue measure on $\mathbb{R}^d$. Its density can be easily computed and equals:

$$\frac{1}{(2\pi)^{d/2}|\det(A)|} \exp(-\frac{1}{2}\|A^{-1}(x - \overrightarrow{b})\|^2),$$

and note that $\|A^{-1}(x - \overrightarrow{b})\|^2 = \langle K^{-1}(x - \overrightarrow{b}), (x - \overrightarrow{b})\rangle$, where $K := \mathbf{Cov}(X)$.

The following characterizes gaussian vectors with independent coordinates:

Proposition 16.7. *Let $X = (X_1, \dots, X_d)$ be a gaussian vector. The following are equivalent:*

(a) the X_i's are independent random variables,
(b) the X_i's are pairwise independent,
(c) the covariance matrix $\mathbf{Cov}(X_i, X_j)$ is a diagonal matrix.

Proof. (a) implies (b) implies (c) are all clear. By Proposition 16.4 X has the same law as some $A\overrightarrow{N} + \overrightarrow{b}$, where $\overrightarrow{b} = \mathbb{E}(X)$ and A is any matrix such that $\mathbf{Cov}(X) = AA^*$. So if (c) holds, then we can choose A diagonal. This now clearly implies (a). □

We are now ready to state and prove one of the corner stones of probability theory, namely the Central Limit Theorem.

Theorem 16.8 (Central Limit Theorem)**.** *Let $(X_n)_{n\geqslant 1}$ be independent and identically distributed $\mathbb{R}^d$-valued random variables with common law μ. Assume that $\|X_1\|^2$ is integrable (we say that μ has a finite moment of order 2). Then*

$$Y_n := \frac{1}{\sqrt{n}}(X_1 + \ldots + X_n - n\mathbb{E}(X_1))$$

converges in law towards a gaussian distribution on $\mathbb{R}^d$ with mean 0 and same covariance matrix as X_1, namely $K_\mu := \mathbf{Cov}(X_1)$.

This means that a sum of n independent $\mathbb{R}^d$-valued random variables with common law μ tends to concentrate around the mean $n\int_{\mathbb{R}^d} x d\mu(x)$ with fluctuations of order $\sqrt{n}$, and the fluctuations are random and distributed according to a gaussian law determined by the covariance matrix of μ. The fact that the gaussian law arises this way and depends on μ in such a mild way (only via the covariance) is remarkable.

A historical aside: The phenomenon was discovered by de Moivre in the early 18th century and discussed in his book "The Doctrine of Chances" in which he applied the recently discovered Stirling formula on the asymptotics of $n!$ to derive the theorem in the special case of binomial random variables (i.e. X_n is 1 or 0 with probability p and $1-p$ respectively), see the 3rd Example sheet. The theorem was extended in the present form (perhaps assuming the X_i's were bounded) by Laplace later on, and then by Lyapunov to non uniformly distributed random variables (but this requires a further assumption on the growth of the variances). Throughout the 19th century the result was known as the "Law of errors". The term Central Limit Theorem was coined (in German: "Zentraler Grezwertsatz der Wahrscheinlichkeitsrechnung") by Pólya in the 20th century.

Proof. We give the usual proof via Fourier transform and characteristic functions. By Lévy's criterion, to show that $Y_n \to Y$ in law we need to prove pointwise convergence of characteristic functions $\widehat{\mu_{Y_n}}(u) \to \widehat{\mu_Y}(u)$, for each $u \in \mathbb{R}^d$ Since $\widehat{\mu_Y}(tu) = \widehat{\mu_{\langle Y,u\rangle}}(t)$, this is equivalent to showing that $\langle Y_n, u\rangle$ converges in law towards $\langle Y, u\rangle$ for each u. So without loss of generality, we may assume that $d = 1$.

Then again wlog (changing X_i into $(X_i - \mathbb{E}(X_i))/\sqrt{\mathbf{Var}(X_i)}$, we may assume that $\mathbb{E}(X_i) = 0$ and $\mathbb{E}(X_i^2) = 1$. Then we write:

$$\widehat{\mu_{Y_n}}(u) = \mathbb{E}(e^{iuY_n}) = \prod_{i=1}^{n} \mathbb{E}(e^{(iu\frac{X_i}{\sqrt{n}})}) = [\widehat{\mu}(\frac{u}{\sqrt{n}})]^n$$

where $\widehat{\mu}(u) := \mathbb{E}(e^{iuX_1})$. But X_1 is square integrable, so we may differentiate twice under the integral sign:

$$\frac{d^2}{du^2}\widehat{\mu}(u) = \int -x^2 e^{iux} d\mu(x),$$

which is a continuous function of u. Hence $\widehat{\mu}(u)$ is of class C^2, and we may write its Taylor expansion near $u = 0$ as follows:

$$\widehat{\mu}(u) = \widehat{\mu}(0) + u\widehat{\mu}'(0) + \frac{u^2}{2}\widehat{\mu}''(0) + o(u^2).$$

But note that

$$\widehat{\mu}'(0) = i\int x d\mu(x) = i\mathbb{E}(X_1) = 0,$$

and

$$\widehat{\mu}''(0) = -\int x^2 d\mu(x) = -1$$

Hence

$$\widehat{\mu}(u) = 1 - \frac{u^2}{2} + o(u^2).$$

Hence for each $u \in \mathbb{R}$, as n tends to $+\infty$,

$$(\widehat{\mu}(\frac{u}{\sqrt{n}}))^n = (1 - \frac{u^2}{2n} + o(\frac{u^2}{n}))^n = [\exp(-\frac{u^2}{2n} + o(\frac{u^2}{n}))]^n \to \exp(-\frac{u^2}{2}).$$

Finally we get:

$$\mathbb{E}(e^{iuY_n}) \to e^{-u^2/2} = \widehat{g}(u)$$

where $g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the density of the standard gaussian $\mathcal{N}(0,1)$. This concludes the proof.

□

Remark 16.9. A comment on proofs of the CLT. This is the slickest proof of the Central Limit Theorem. There are other proofs. We've already mentioned the original proof by de Moivre via Stirling's formula (see the 3rd Example sheet). But this seems to work only for finitely supported laws μ. For bounded random variables, one can use the method of moments (i.e. prove that all moments converge to the respective moment of a gaussian law); this works but it is messier than the proof via charateristic functions we have just given. Another approach is due to Lindenberg and consists in replacing each increment X_i by a gaussian one at a time and controlling the error terms (see Feller's book for Lindeberg's method). This leads to a generalized result, where the random variables are allowed to have different laws. Yet another approach it to use entropy: the gaussian maximizes the entropy among all laws with given variance. The great Soviet mathematician Linnik showed how to exploit this to give another proof of the Central Limit Theorem.

Lecture 21

17. Introduction to ergodic theory

In the last three lectures we give a very brief introduction to ergodic theory. This is a vast subject and we will only go as far as proving the Von Neumann Mean ergodic theorem as an application of the Hilbert space techniques developed earlier in the course and derive from it the (stronger) Birkhoff pointwise ergodic theorem. Given time constraints we will have to skip over several basic facts and important motivational examples. For a thorough and modern introduction accessible at this level we recommend the first few chapters of Einsiedler and Ward "Ergodic Theory" (Springer GTM).

Ergodic theory is the study of statistical properties of dynamical systems. In dynamics one studies iterations $T^n = T \circ \ldots \circ T$ of a self map $T : X \to X$ of a space X and one is interested in the behavior of orbits $\{T^n x\}_{n \geqslant 0}$. Is the orbit dense? does it accumulate onto some attractor? does it come back close to where it started? if so at what time and how often? etc. Ergodic theory is the study of these questions from a statistical point of view, where one assumes that the space X comes equipped with a T-invariant measure μ. The questions then become: does the orbit $\{T^n x\}_{n \geqslant 0}$ become equidistributed w.r.t to some measure, i.e. does the sequence of measures

$$\frac{1}{n} \sum_{i=0}^{n-1} \delta_{T^i x}$$

on X converge weakly to some measure on X? What is the behaviour of a typical orbit, i.e. of $\{T^n x\}_{n \geqslant 0}$ for μ-almost every x? Are there more than one T-invariant measure on X, can one classify them? etc.

We begin by introducing some standard terminology. Let $(X, \mathcal{A}, \mu)$ be a measure space. We will assume throughout that $\mu(X)$ is finite.

Definition 17.1 (measure preserving map)**.** *A measurable map $T : X \to X$ is called measure preserving if*

$$T_* \mu = \mu,$$

where $T_* \mu$ denotes the image measure. In other words: $\mu(T^{-1} A) = \mu(A)$ for all $A \in \mathcal{A}$. A measure space $(X, \mathcal{A}, \mu)$ together with a measure preserving map T is often called a measure preserving system.

Definition 17.2 (Invariant function and invariant σ-algebra)**.** *(1) A measurable function $f : X \to \mathbb{R}$ is called T-invariant if $f = f \circ T$.*

(2) A measurable subset $A \in \mathcal{A}$ is called T-invariant if $T^{-1} A = A$,

(3) $\mathcal{I} := \{A \in \mathcal{A}, T^{-1} A = A\}$ is a σ-subalgebra of $\mathcal{A}$ called the invariant σ-algebra.

Lemma 17.3. *For a measurable function $f : X \to \mathbb{R}$, TFAE:*

(i) f is T-invariant,

(ii) f is measurable with respect to $\mathcal{I}$

Proof. For $t \in \mathbb{R}$ we have:

$$T^{-1}(\{x \in X, f(x) < t\}) = \{x \in X, f \circ T(x) < t\}.$$

So if f is T-invariant this is also equal to $\{x \in X, f(x) < t\}$, so that (i) implies (ii). For the converse, note that if f is $\mathcal{I}$-measurable, then $\{x \in X, f(x) < t\}$ is in $\mathcal{I}$ for all t, and thus equals $\{x \in X, f \circ T(x) < t\}$. Hence f and $f \circ T$ have the same sublevel sets. But this clearly implies that $f = f \circ T$. □

Definition 17.4 (ergodic transformation)**.** *Given a measure space $(X, \mathcal{A}, \mu)$ and a measure preserving map $T : X \to X$, we say that T is ergodic with respect to μ (or equivalently that μ is ergodic with respect to T) if for all $A \in \mathcal{I}$ either $\mu(A) = 0$ or $\mu(A^c) = 0$.*

In other words a measure preserving system is ergodic if it cannot be written as the disjoint union of two non-trivial subsystems (i.e. invariant measurable subsets of positive measure). So it is a kind of irreducibility condition, and indeed one can often reduce the understanding of a measure preserving system to ergodic subsystems.

Exercise/Example: Let X be a finite set and $T : X \to X$ a self map. Take $\mathcal{A}$ the discrete Boolean algebra (i.e. all subsets of X) and μ the counting measure. Then

(1) T is measure preserving if and only if T is a bijection,
(2) T is ergodic if and only if for every $x, y \in X$, there is an integer $n \geqslant 0$ such that $T^n x = y$.

Lemma 17.5. *Let $(X, \mathcal{A}, \mu, T)$ be a measure preserving system. Then T is ergodic with respect to μ if and only if for every $\mathcal{I}$-measurable map f there is $a \in \mathbb{R}$ such that $f(x) = a$ for μ-almost every $x \in X$.*

Proof. This is Exercise 11 in the 4th Example Sheet. □

We now study two important examples:

Example 1: (circle rotation) Let $X = \mathbb{R}/\mathbb{Z}$ be the circle. Let $\mathcal{A}$ be the Borel σ-algebra and m Lebesgue measure (rather Lebegue measure on $[0, 1)$ identified naturally with $\mathbb{R}/\mathbb{Z}$). Fix $a \in \mathbb{R}$ and consider the self-map

$$\begin{aligned} T : X &\to X, \\ x &\mapsto x + a. \end{aligned}$$

Then T is measure preserving. We have:

Proposition 17.6. *T is ergodic w.r.t m if and only if a is irrational.*

Note that this is also equivalent to asking that there is a dense orbit (or that all orbits are dense).

Proof. The proof uses the Fourier transform. Let $f = 1_A$, where $A \in \mathcal{I}$. We compute the Fourier coefficients:

$$\begin{aligned} \widehat{f}(n) &= \int_{\mathbb{R}/\mathbb{Z}} e^{2i\pi nx} f(x) dx = \int_{\mathbb{R}/\mathbb{Z}} e^{2i\pi nT(x)} f \circ T(x) dx \\ &= \int_{\mathbb{R}/\mathbb{Z}} e^{2i\pi na} e^{2i\pi nx} f(x + a) dx = \int_{\mathbb{R}/\mathbb{Z}} e^{2i\pi na} e^{2i\pi nx} f(x) dx = e^{2i\pi na} \widehat{f}(n) \end{aligned}$$

where we have first used the fact that $m = dx$ is T-invariant, and then that $f \circ T = f$. If a is irrational, then $e^{2i\pi na} \neq 1$ when $n \neq 0$ so we must conclude that $\widehat{f}(n) = 0$ if $n \neq 0$. But a function on $\mathbb{R}/\mathbb{Z}$ all of whose Fourier coefficients are zero except when $n = 0$ is almost everywhere constant (to see this apply Parseval's formula). On the other hand if a is rational, say $a = \frac{p}{q}$, then T is not ergodic, because, for instance the union of intervals of the form $[\frac{k}{q}, \frac{k}{q} + \frac{1}{2q})$ for $k = 0, \ldots, q - 1$ is invariant under T and has measure $\frac{1}{2}$.

□

Example 2: (times 2 map on the circle) Again let $X = \mathbb{R}/\mathbb{Z}$ with Lebesgue measure m, but this time, we consider the map

$$\begin{aligned} T_2 : X &\to X, \\ x &\mapsto 2x \mod \mathbb{Z}. \end{aligned}$$

Note that T_2 is a measure preserving map (even though it dilates by a factor 2). Indeed, the preimage of a small interval of size s is made of two intervals of size $s/2$.

Proposition 17.7. *T_2 is ergodic.*

Proof. Let $f = 1_A$, where $A \in \mathcal{I}$. Once again we can compute its Fourier coefficients:

$$\begin{aligned} \widehat{f}(n) &= \int_{\mathbb{R}/\mathbb{Z}} e^{2i\pi nx} f(x)dx = \int_{\mathbb{R}/\mathbb{Z}} e^{2i\pi nT_2(x)} f \circ T_2(x)dx \\ &= \int_{\mathbb{R}/\mathbb{Z}} e^{4i\pi nx} f(x)dx = \widehat{f}(2n). \end{aligned}$$

where we have first used the fact that $m = dx$ is T_2-invariant, and then that $f \circ T_2 = f$. Iterating this relation we see that

$$\widehat{f}(2^k n) = \widehat{f}(n) \tag{17.1}$$

for all $n \in \mathbb{Z}$ and all $k \geqslant 1$. However $f \in \mathbf{L}^2(X, dx)$, so Parseval's identity reads:

$$\sum_{n\in\mathbb{Z}} |\widehat{f}(n)|^2 = \int_{\mathbb{R}/\mathbb{Z}} |f(x)|^2 dx = m(A).$$

This is finite, so by (17.1) we must have $\widehat{f}(n) = 0$ for all $n \neq 0$. Hence f is almost everywhere constant. In other words either $m(A) = 0$, or $m(A^c) = 0$, which means that T_2 is ergodic.

□

Remark 17.8. The Lebesgue measure is not the only T_2-invariant and ergodic Borel probability measure on $\mathbb{R}/\mathbb{Z}$. For example the Dirac mass δ_0 is invariant and so is $\frac{1}{2}(\delta_{1/3} + \delta_{2/3})$. But there are non-atomic invariant measures too. For example we may consider the random variable $X := \sum_{n\geqslant 1} \frac{\epsilon_n}{2^n}$, modulo $\mathbb{Z}$, where $(\epsilon_n)_{n\geqslant 1}$ is a sequence of i.i.d. random variables such that $\mathbb{P}(\epsilon_n = 0) = p$ and $\mathbb{P}(\epsilon_n = 1) = 1 - p$ for some $p \in (0, 1)$. Then $2X \mod \mathbb{Z}$ has the same law as $X \mod \mathbb{Z}$. Thus this law is therefore invariant under T_2. And it is not Lebesgue if $p \neq \frac{1}{2}$ and has no atoms. A famous conjecture of Furstenberg (still open!) asserts that the only Borel probability measure on $\mathbb{R}/\mathbb{Z}$ with no atoms that is invariant under both T_2 and T_3 is Lebesgue.

18. Canonical model for stochastic processes

In this section, starting with an arbitrary sequence of random variables, we are going to associate dynamical system $T : X \to X$, where X will be the space of all sequences. Then we investigate invariant measures on this system.

Let $(X_n)_{n\geqslant 1}$ be a sequence of $\mathbb{R}^d$-valued random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We define

$$\begin{aligned} \Phi : \Omega &\to (\mathbb{R}^d)^{\mathbb{N}} \\ \omega &\mapsto (X_n(\omega))_{n\geqslant 1} \end{aligned}$$

be the sample path map, it assigns to the outcome $\omega \in \Omega$ the full sequence, or sample path, $(X_n(\omega))_n$.

Now on the space of sequences $X := (\mathbb{R}^d)^{\mathbb{N}}$, we may define the shift map

$$T : (\mathbb{R}^d)^{\mathbb{N}} \to (\mathbb{R}^d)^{\mathbb{N}}$$
$$(x_n)_{n\geqslant 1} \mapsto (x_{n+1})_{n\geqslant 1}.$$

The space of sequences $X := (\mathbb{R}^d)^{\mathbb{N}}$ is also called the shift space. On it we can define the coordinate functions:

$$x_n : X \to \mathbb{R}^d$$
$$(x_n)_{n\geqslant 1} \mapsto x_n.$$

We endow X with a σ-algebra $\mathcal{A}$, which is called the product σ-algebra, and is defined as the smallest σ-algebra that makes all coordinate functions measurable. In other words $\mathcal{A} = \sigma(x_n, n \geqslant 1)$.

It is also the σ-algebra generated by the Boolean algebra of cylinder sets. A cylinder set is a subset of X of the form $\pi_F^{-1}(A)$, $F \subset \mathbb{N}$ is a finite set of indices, A is a Borel set in $(\mathbb{R}^d)^{|F|}$, and $\pi_F : X \to (\mathbb{R}^d)^{|F|}$ is the projection to the coordinates $i_1, ..., i_k$, if $F = \{i_1, \dots, i_k\}$.

The image measure $\mu := \Phi_*\mathbb{P}$ is a probability measure on $(X, \mathcal{A})$ and is called the law of the stochastic process $(X_n)_{n\geqslant 1}$.

The dynamical system $(X, \mathcal{A}, \mu, T)$ is called the canonical model associated to $(X_n)_n$.

Proposition-Definition 18.1. *Let $(X_n)_{n\geqslant 1}$ be a stochastic process and $(X, \mathcal{A}, \mu, T)$ its canonical model. Then the following are equivalent:*

(1) $(X, \mathcal{A}, \mu, T)$ is measure preserving,
(2) for all $k \geqslant 1$, the joint law of $(X_n, X_{n+1}, \dots, X_{n+k})$ is independent of n.

In this case, the process is called stationary.

Proof. Note that μ is the law of $(X_i)_{i\geqslant 1}$, while $T_*^n\mu$ is the law of $(X_{i+n})_{i\geqslant 1}$. So if $\mu = T_*\mu$, then for all n, $\mu = T_*^n\mu$ and the two laws coincide. Conversely if the laws coincide on cylinders, they must be equal on all of $\mathcal{A}$ by Dynkin's lemma (the cylinders form a π-system that generates). □

A special class of stationary processes are the so-called Bernoulli shifts:

Proposition-Definition 18.2. *If $(X_n)_{n\geqslant 1}$ is an i.i.d. process, then it is stationary and the canonical model $(X, \mathcal{A}, \mu, T)$ is ergodic. In this case we say that the measure preserving system $(X, \mathcal{A}, \mu, T)$ is a Bernoulli shift, and $\mu = \nu^{\otimes\mathbb{N}}$, where ν is the law of X_1.*

Proof. This is an application of Kolmogorov's 0-1 law. Clearly the process is stationary, so μ is T-invariant. We have to show that it is ergodic. So let $\mathcal{I}$ be the invariant σ-algebra. Note that $\Phi^{-1}\mathcal{I}$ is contained in the tail σ-algebra $\mathcal{T}$ of the process, namely $\mathcal{T} = \bigcap_k \sigma(X_k, X_{k+1}, \dots)$. Indeed, if $A \in \mathcal{I}$, then $A = T^{-1}A$ and

$$\begin{aligned}\Phi^{-1}(A) &= \{\omega \in \Omega, (X_n(\omega))_{n\geqslant 1} \in A\}\\ &= \{\omega \in \Omega, (X_n(\omega))_{n\geqslant 1} \in T^{-1}A\} = \{\omega \in \Omega, (X_{n+1}(\omega))_{n\geqslant 1} \in A\}\\ &= \{\omega \in \Omega, (X_{n+k}(\omega))_{n\geqslant 1} \in A\} \subset \sigma(X_{k+1}, X_{k+2}, \dots)\end{aligned}$$

for all $k \geqslant 1$. This means that $\Phi^{-1}(A) \in \mathcal{T}$.

But the $(X_n)_n$'s are i.i.d., so Kolmogorov's 0-1 law implies that $\mathcal{T}$ is trivial, and hence $\mu(A) \in \{0, 1\}$. So T is ergodic. □

Remark 18.3. Sometimes authors reserve the term Bernoulli shift (or Bernoulli scheme) to the case when ν is finitely supported on a finite abstract set (i.e. not necessarily part of $\mathbb{R}^d$).

Lecture 22

19. The mean ergodic theorem

Let $(X, \mathcal{A}, \mu, T)$ be a probability measure preserving system.

Theorem 19.1 (Mean ergodic theorem in $\mathbf{L}^2$)**.** *Let $f \in \mathbf{L}^2(X, \mathcal{A}, \mu)$. We set*

$$S_n(f) := \frac{1}{n} \sum_0^{n-1} f \circ T^i.$$

Then there is a T-invariant function $\bar{f}$ (in fact $\bar{f} = \mathbb{E}(f|\mathcal{I})$ the conditional expectation, where $\mathcal{I}$ is the σ-algebra of T-invariant subsets) such that

$$\lim_{n \to +\infty} S_n(f) = \bar{f}$$

where the convergence takes place in $\mathbf{L}^2(X, \mathcal{A}, \mu)$.

Remark: If $\mathcal{H}$ is a Hilbert space and $A : \mathcal{H} \to \mathcal{H}$ is a bounded linear map, then we may define its *adjoint* A^*, which is the linear map $y \mapsto A^* y$ defined to be the unique vector (as given by the Riesz representation theorem) such that

$$\langle Ax, y \rangle = \langle x, A^* y \rangle.$$

Observe that the operation $A \mapsto A^*$ is involutive, i.e. $A^{**} = A$. Moreover the operator norm of A and A^* coincide, because

$$\|A^* y\| = \sup_{\|x\|=1} |\langle x, A^* y \rangle| = \sup_{\|x\|=1} |\langle Ax, y \rangle| \leqslant \|A\| \cdot \|y\|$$

which gives $\|A^*\| \leqslant \|A\|$ and hence $\|A^*\| = \|A\|$ by symmetry (one can further prove that $\|AA^*\| = \|A\|^2$).

Proof of the mean ergodic theorem. We give the original proof, due to von Neumann (1932). It is based on a simple Hilbert space argument. We consider the Hilbert space $\mathcal{H} := \mathbf{L}^2(X, \mathcal{A}, \mu)$. Let

$$\begin{aligned} U : \mathcal{H} &\to \mathcal{H} \\ f &\mapsto f \circ T. \end{aligned}$$

It is clear that U is a linear operator on $\mathcal{H}$, which is bounded and in fact an isometry, that is:

$$\|Uf\| = \|f\|$$

for every $f \in \mathcal{H}$. This is clear, because by assumption μ is T-invariant, so

$$\|Uf\|^2 = \int_X |f \circ T|^2 d\mu = \int_X |f|^2 d\mu = \|f\|^2.$$

Let $W := \{\phi - U\phi, \phi \in \mathcal{H}\}$. This is a subspace of $\mathcal{H}$ (called the subspace of *co-boundaries*).

(a) If $f \in W$, then

$$S_n f = \frac{1}{n} \sum_0^{n-1} (\phi \circ T^i - \phi \circ T^{i+1}) = \frac{1}{n}(\phi - \phi \circ T^n)$$

obviously tends to 0 in $\mathcal{H}$ as $n \to +\infty$.

(b) if $f \in \overline{W}$ (the closure of W in $\mathcal{H}$), then we again have $S_n f \to 0$, because for every $\epsilon > 0$ we can find $g \in W$ with $\|f - g\| < \varepsilon$, and so:

$$\|S_n f - S_n g\| \leqslant \frac{1}{n} \sum_1^n \|f \circ T^i - g \circ T^i\| \leqslant \|f - g\| \leqslant \epsilon,$$

which implies that $\limsup \|S_n f\| \leqslant \epsilon$, and hence that $\limsup \|S_n f\| = 0$ since ϵ was arbitrary.

(c) By the orthogonal decomposition for closed subspaces of Hilbert space we have:

$$\mathcal{H} = \overline{W} \oplus W^{\perp}$$

and $\overline{W} = (W^{\perp})^{\perp}$. Since $S_n f \to 0$ if $f \in \overline{W}$, without loss of generality we may assume that $f \in W^{\perp}$. But we now observe the following:

$$\{g \in \mathcal{H}, g = Ug\} \subset W^{\perp} = \{g \in \mathcal{H}, \langle g, \phi - U\phi\rangle = 0 \,\forall \phi \in \mathcal{H}\} = \{g \in \mathcal{H}, \langle g, \phi\rangle = \langle U^* g, \phi\rangle \,\forall \phi \in \mathcal{H}\}$$
$$= \{g \in \mathcal{H}, g = U^* g\} \subset \{g \in \mathcal{H}, g = Ug\},$$

which is exactly the subspace of T-invariant functions in $\mathcal{H}$. To justify the last containment in the above formula, note that

$$\|g - Ug\|^2 = \|g\|^2 + \|Ug\|^2 - 2\,\mathrm{Re}\langle g, Ug\rangle = 2\|g\|^2 - 2\,\mathrm{Re}\langle U^* g, g\rangle,$$

which is clearly 0 if $g = U^* g$.

So if $g \in W^{\perp}$, then $g = g \circ T$ and $S_n g = g$ for all n. Hence the theorem holds with $\bar{f}$ = the orthogonal projection of f onto the closed subspace $W^{\perp}$ of T-invariant functions. □

The mean ergodic theorem gives convergence of ergodic averages in $\mathbf{L}^2$. It is easy to derive from it convergence in $\mathbf{L}^p$ for any $p \in [1, +\infty)$ if we assume that f is in $\mathbf{L}^p$ to begin with:

Corollary 19.2 (Mean ergodic theorem in $\mathbf{L}^p$)**.** *Let $p \in [1, +\infty)$. Let $(X, \mathcal{A}, \mu, T)$ be a probability measure preserving system. Let $f \in L^p(X, \mathcal{A}, \mu)$ and $S_n(f) := \frac{1}{n}\sum_0^{n-1} f \circ T^i$ as before. Then there is a T-invariant function $\bar{f}$ (in fact $\bar{f} = E(f|\mathcal{I})$, where $\mathcal{I}$ is the σ-algebra of T-invariant subsets) such that*

$$\lim_{n \to +\infty} S_n(f) = \bar{f}$$

where the convergence takes place in $\mathbf{L}^p(X, \mathcal{A}, \mu)$.

Proof. We first observe that, as a consequence of the completeness of $\mathbf{L}^p$, it is enough to prove that the result holds for a dense subspace of functions, say W. Indeed assume that the result holds for every $g \in W$ and that W is dense in $\mathbf{L}^p$. Then given $f \in \mathbf{L}^p$, the sequence of ergodic averages $S_n f$ will be a Cauchy sequence, because for every $\varepsilon > 0$ there is $g \in W$ such that $\|f - g\|_p < \varepsilon$ and hence for all n

$$\|S_n f - S_n g\|_p \leqslant \|f - g\|_p \leqslant \varepsilon$$

Therefore when n is large enough $\|S_n f - \bar{g}\|_p \leqslant 2\varepsilon$. And for any n, m large enough $\|S_n f - S_m f\|_p \leqslant 4\varepsilon$. So $(S_n f)_n$ is a Cauchy sequence in $\mathbf{L}^p$ and thus converges to some limit $\bar{f}$. It is clear that $\bar{f}$ is T-invariant, i.e. $\bar{f} \circ T = \bar{f}$, because

$$S_n f \circ T - S_n f = \frac{1}{n}(f \circ T^n - f)$$

clearly tends to 0.

To conclude the proof, take $W = \mathbf{L}^\infty(X, \mathcal{A}, \mu)$. It is a dense subspace in $\mathbf{L}^p$ for all $p \geqslant 1$ (recall that the vector space spanned by simple functions is dense in any $\mathbf{L}^p$ $p < \infty$). The von Neumann Mean ergodic theorem applies to any $g \in W$ and gives convergence of ergodic averages $S_n g$ towards some T-invariant function $\bar{g}$ in $\mathbf{L}^2$. Note that $\bar{g} \in W$ as well, because $\|S_n g\|_\infty \leqslant \|g\|_\infty$ for all n, so $\|\bar{g}\|_\infty \leqslant \|g\|_\infty$. To see that the convergence $S_n g \to \bar{g}$ holds in $\mathbf{L}^p$ as well, note that if $p \leqslant 2$, $\|\cdot\|_p \leqslant \|\cdot\|_2$, while if $p > 2$, then $\|\cdot\|_p^p \leqslant \|\cdot\|_2^2 \cdot \|\cdot\|_\infty^{p-2}$. □

We will soon prove that ergodic averages actually also converge pointwise assuming only that f belongs to $\mathbf{L}^1$. This is the content of the *pointwise ergodic theorem*. In order to get there, we will first establish a technical result, that goes under the name of *maximal ergodic theorem*.

Theorem 19.3 (Maximal ergodic theorem)**.** *Let $(X, \mathcal{A}, \mu, T)$ be a probability measure preserving system and $f \in \mathbf{L}^1(X, \mathcal{A}, \mu)$. For any $t > 0$ we set*

$$E_t := \{x \in X, \sup_n S_n f(x) > t\}.$$

Then

$$\mu(E_t) \leqslant \frac{1}{t}\|f\|_1.$$

The event E_t is the set of orbits of T whose ergodic averages manage to overshoot t at least once. So the result says that the probability that the ergodic averages ever become larger than t decays at least as $O(1/t)$, where the implied constant in O is simply the $\mathbf{L}^1$-norm of f.

The maximal ergodic theorem can also be understood as follows: say that $f = 1_A$, where $A \in \mathcal{A}$ is a set of small μ-measure. Then the result is a quantified way of saying that except for a set of starting points x of small measure, for all n the average time spent in A up to time n is small. More precisely, if $\mu(A) = \epsilon^2$, setting $t = \epsilon$, we get:

$$\mu(E_\epsilon) \leqslant \epsilon^{-1}\|f\|_1 = \epsilon$$

and E_ϵ is the set of exceptional starting points x whose orbit up to time n can, for certain n's, spend a time larger than ϵn in A.

To prove the maximal ergodic theorem, we will need the:

Lemma 19.4 (the maximal inequality)**.** *Let $f \in \mathbf{L}^1(X, \mathcal{A}, \mu)$ and*

$$f_n = nS_n f = \sum_{i=0}^{n-1} f \circ T^i$$

for $n > 0$. Set also $f_0 = 0$. For each $N \geqslant 0$ let

$$P_N = \{x \in X, \max_{0 \leqslant n \leqslant N} f_n(x) > 0\}.$$

Then

$$\int_{P_N} f d\mu \geqslant 0.$$

Proof. Set $F_N := \max_{0 \leqslant n \leqslant N} f_n$. For all $n \leqslant N$ we have $f_n \leqslant F_N$ so

$$f_{n+1} = f_n \circ T + f \leqslant F_N \circ T + f.$$

If $x \in P_N$, $F_N(x) > 0$, so $F_N(x) = \max_{1 \leqslant n \leqslant N} f_n(x) \leqslant \max_{0 \leqslant n \leqslant N} f_{n+1}(x)$, so

$$F_N(x) \leqslant F_N \circ T(x) + f(x)$$

and integrating over P_N yields:

$$\int_{P_N} F_N d\mu \leqslant \int_{P_N} F_N \circ T d\mu + \int_{P_N} f d\mu.$$

Note that $F_N = 0$ on P_N^c as $f_0 = 0$, and $F_N \geqslant 0$ everywhere, so

$$\int_{P_N} F_N d\mu = \int_X F_N d\mu \leqslant \int_X F_N \circ T d\mu + \int_{P_N} f d\mu,$$

which implies that $\int_{P_N} f d\mu \geqslant 0$ by T-invariance of μ. □

Proof of the Maximal ergodic theorem. Simply apply the maximal inequality to $g = f - t$ and note that

$$E_t(f) = \bigcup_{N \geqslant 1} P_N(g)$$

while $S_n g = S_n f - t$. The maximal inequality implies that

$$\int_{P_N(g)} (f - t) d\mu \geqslant 0,$$

or in other words $t\mu(P_N(g)) \leqslant \int_{P_N(g)} f d\mu \leqslant \int_{P_N(g)} |f| d\mu \leqslant \|f\|_1$. Since $P_N \subset P_{N+1}$, we conclude that

$$t\mu(E_t) \leqslant \|f\|_1.$$

□

Lecture 23

20. The pointwise ergodic theorem

In this last lecture, we present the pointwise ergodic theorem. This is an improvement on the mean ergodic theorem asserting that the ergodic averages converge almost everywhere.

Theorem 20.1 (Pointwise ergodic theorem)**.** *Let $(X, \mathcal{A}, \mu, T)$ be a probability measure preserving system. Let $f \in \mathbf{L}^1(X, \mathcal{A}, \mu)$ and $S_n(f) := \frac{1}{n}\sum_0^{n-1} f \circ T^i$ as before. Then there is a T-invariant function $\bar{f}$ (in fact $\bar{f} = \mathbb{E}(f|\mathcal{I})$, where $\mathcal{I}$ is the σ-algebra of T-invariant subsets) such that*

$$\lim_{n\to+\infty} S_n(f) = \bar{f}$$

where the convergence μ-almost everywhere. If particular, if the system is ergodic, then $S_n(f)$ converges almost everywhere to the constant $\int f d\mu$.

When the system is ergodic and $f = 1_A$ for some $A \in \mathcal{A}$, the theorem says that for μ-almost every starting point x, the time spent inside A by the orbit $\{T^n x\}_{0\leqslant n\leqslant N}$ of x between $n = 0$ and $n = N$, is roughly $\mu(A)N$ as N grows to infinity. In other words, almost every orbit is equidistributed.

Proof. This will follow easily in two steps. First we use the Maximal ergodic theorem to reduce to the case when f is bounded. And in a second step we combine the Maximal ergodic theorem with the $\mathbf{L}^1$ mean ergodic theorem to conclude pointwise convergence in case f is bounded. Note that by considering $f - \bar{f}$ in place of f, we may assume that $\mathbb{E}(f|\mathcal{I}) = 0$.

Step 1: reduction to f bounded. Given $M > 0$, let $f_M := f1_{|f|<M}$ be the truncation of f at height M. We assume that the result holds for f_M for each M. Let $\varepsilon \in (0, 1)$. Let E_M be the subset of those $x \in X$ with $|\mathbb{E}(f_M|\mathcal{I})(x)| > \varepsilon$. Note that $\mu(E_M) \to 0$ as $M \to +\infty$, because $\mathbb{E}(f_M|\mathcal{I})$ converges μ-a.e. to $\mathbb{E}(f|\mathcal{I}) = 0$. Note further that for μ-almost every $x \notin E_M$, if $\limsup_n |S_n f(x)| > 3\varepsilon$, then $\sup_n |S_n(f - f_M)(x)| > 2\varepsilon$, because by assumption $S_n f_M \to \mathbb{E}(f_M|\mathcal{I})$ μ-a.e.

On the other hand, the Maximal ergodic theorem gives:

$$\mu(\{x \in X, \sup_n |S_n(f - f_M)(x)| > 2\varepsilon\}) \leqslant \frac{1}{\varepsilon}\|f - f_M\|_1.$$

Letting M tend to infinity, we thus get:

$$\mu(\{x \in X, \limsup_n |S_n f(x)| > 3\varepsilon\}) \leqslant \mu(E_M) + \frac{1}{\varepsilon}\|f - f_M\|_1,$$

and the right hand side tends to 0 as $M \to +\infty$. This shows that $S_n f \to 0$ μ-a.e. as desired.

Step 2: case when f is bounded. We already know, by the $\mathbf{L}^1$ mean ergodic theorem (Corollary 19.2) that the ergodic averages $S_m f$ converge in $\mathbf{L}^1$. So

$$\lim_{m\to+\infty} \|S_m f\|_1 = 0.$$

Fix a large m and consider a larger $n > m$. Then write:

$$nmS_n(S_m f) = \sum_{0\leqslant i<n, 0\leqslant j<m} f \circ T^{i+j}$$

and observe that in this sum each term $f \circ T^k$ with $m \leqslant k \leqslant n$ appears m times exactly, while the others appear at most m times and there are at most $2m$ other

relevant values of k, so that makes at most $2m^2$ other terms. Since f is bounded, we conclude that:

$$nmS_n(S_mf) = mnS_nf + O(m^2\|f\|_\infty).$$

In particular as $n \to +\infty$ and m stays fixed,

$$\|S_n(S_mf) - S_nf\|_\infty \to 0.$$

So if $x \in X$ is such that $\limsup_n |S_nf(x)| > 2\varepsilon$, then $\limsup_n |S_nS_mf(x)| > 2\varepsilon$ and in particular $\sup_n |S_nS_mf(x)| > 2\varepsilon$. We can thus apply the Maximal ergodic theorem to S_mf and conclude that

$$\mu(\{x \in X, \limsup_n |S_nf(x)| > 2\varepsilon\}) \leqslant \frac{1}{\varepsilon}\|S_mf\|_1.$$

The left hand side is independent of m, so we can let m tend to infinity, and since $\|S_mf\|_1 \to 0$, we conclude that the left hand side is 0. Since ε was arbitrary, this means that $S_nf(x) \to 0$ for μ-a.e. x as desired.

□

We have used the von Neumann Mean ergodic theorem in our proof of the pointwise ergodic theorem. There are other routes that avoid it (see e.g. Norris's notes). Conversely it is an easy exercise to derive the mean ergodic theorem from the pointwise theorem (basically truncating and applying Lebesgue's Dominated Convergence Theorem). The pointwise ergodic theorem was proven by George Birkhoff (who, for the record, scooped out von Neumann by rushing his proof to publication by December 1931, while von Neumann's earlier discovery appeared only in 1932.)

Remark 20.2. Another landmark theorem of real analysis is the Lebesgue differentiation theorem we have alluded to earlier in the course (go to Part II "Analysis of function" in Lent to learn about it). There are striking similarities between the statement and the proof of the pointwise ergodic theorem and the Lebesgue differentiation theorem. Also related and to some extent a generalization of the above is Doob's martingale convergence theorem in probability theory.

A straightforward consequence of the pointwise ergodic theorem is the Strong Law of Large numbers:

Corollary 20.3 (Strong Law of Large Numbers)**.** *Let $(X_n)_{n\geqslant 1}$ be i.i.d. random variables in $\mathbb{R}^d$ with finite first moment (i.e. $\mathbb{E}(\|X_1\|) < \infty$). Then*

$$\frac{1}{n}\sum_1^n X_i \to \mathbb{E}(X_1)$$

almost surely.

"Strong" refers to the fact that the convergence holds almost surely (while a "weak law" gives convergence in probability). This was first proved in this generality by Kolmogorov (in 1930, by a different argument), who also showed that the conclusion holds even if the X_i are not identically distributed, but, say, have the same average and bounded variance. Note that we have already proved the Strong Law under the stronger assumption that there is a finite fourth moment (see Theorem 10.19). As an alternative to the below, the strong law for finite moment of order 1 can also be obtained using a truncation argument by refining the proof method of Theorem 10.19 above, see Williams' lovely book "Probability with Martingales"

Proof. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space on which the random variables are defined and $(X, \mathcal{A}, \mu, T)$ the associated canonical model. Since the $(X_n)_n$'s are i.i.d., $(X, \mathcal{A}, \mu, T)$ is a Bernoulli shift (in particular ergodic). Define the function f on X by $f(x) = x_1(x)$. It is in $\mathbf{L}^1(X, \mathcal{A}, \mu)$ because $\|f\|_1 = \mathbb{E}\|X_1\| < \infty$. The pointwise ergodic theorem implies that

$$S_n f(x) \to \int f d\mu$$

for μ-almost every $x \in X$. But $\mu = \Phi_*(\mathbb{P})$, so $\int f d\mu = \mathbb{E}(X_1)$ and $S_n f(x) = \frac{1}{n}(X_1 + \ldots + X_n)(\omega)$ if $x = \Phi(\omega)$ and $\Phi : \Omega \to X$ is the sample path map. The result follow. □

As a consequence of the ergodic theorem, one may give the following characterization of ergodic measures among invariant ones as extremal points. Let $(X, \mathcal{A})$ be a measurable space and $T : X \to X$ a measurable self map. Let $\mathcal{I}(X)$ be the family of all T-invariant probability measures on $(X, \mathcal{A})$. A measure $\mu \in \mathcal{I}(X)$ is said to be extremal if one cannot find $\mu_1 \neq \mu_2 \in \mathcal{I}(X)$ and $t \in (0, 1)$ such that $\mu = t\mu_1 + (1-t)\mu_2$. We have:

Proposition 20.4. *An invariant measure $\mu \in \mathcal{I}(X)$ is ergodic if and only if it is extremal.*

Proof. If μ is not ergodic, then there is a T-invariant $A \in \mathcal{A}$ with $\mu(A) \in (0, 1)$. Set $\mu_1 = \frac{1}{\mu(A)}\mu_{|A}$ and $\mu_2 = \frac{1}{\mu(A^c)}\mu_{|A^c}$. Then $\mu_1 \neq \mu_2$ are both T-invariant, while $\mu = t\mu_1 + (1-t)\mu_2$ for $t = \mu(A)$. So μ is not extremal.

Conversely, if μ is ergodic, and $\mu = t\mu_1 + (1-t)\mu_2$ for some $\mu_1, \mu_2 \in \mathcal{I}(X)$, then given any $B \in \mathcal{A}$, the pointwise ergodic theorem applied to $(X, \mathcal{A}, \mu)$ implies that for μ-almost every x, and hence for μ_i-almost every x (for both $i = 1, 2$) we have

$$S_n(1_B) \to_{n \to +\infty} \mu(B).$$

By Dominated Convergence, we conclude that $\mu_i(S_n(1_B)) \to \mu(B)$. But $\mu_i(S_n(1_B)) = \mu_i(B)$. And it follows that $\mu_i(B) = \mu(B)$. Hence $\mu_1 = \mu_2$, and μ is extremal. □