Measure theory and probability notes

Angus Rush

June 24, 2018

Contents

1 Measures and σ-algebras

1.1 σ-algebras and monotone classes

1.1.1 σ-algebras

Definition 1 (algebra of sets)**.** A collection of subsets of Ω is called an algebra if it is closed under binary unions, binary intersections, and complements.

Definition 2 (σ-algebra, measurable set)**.** A σ-algebra on Ω is a collection $\mathcal{F}$ of subsets of Ω such that

1. $\Omega \in \mathcal{F}$
2. $X \in \mathcal{F} \implies X^c \in \mathcal{F}$
3. $\mathcal{F}$ is closed under countable unions.

We call the elements of $\mathcal{F}$ measurable.

That is, a σ-algebra is an algebra of sets which is closed under *countable*, rather than finite, unions and intersections. This is a general theme: 'σ-' is often shorthand for something like 'countable'. (Another example is so-called σ-finiteness).

Note 3. Note that σ-algebras are often referred to as 'σ-fields'. We use the more standard terminology of 'σ-algebra' in these notes.

The theory of σ-algebras is similar to that of topologies; there are two differences in their definitions:

- σ-algebras are closed under the taking of complements, while topologies are not.
- Topologies are closed under *arbitrary* unions and *finite* intersections; σ-algebras are closed under *countable* unions and (as we will see below) *countable* intersections.

Lemma 4. *Any σ-algebra $\mathcal{F}$ is closed under countable intersections. That is, if $A_i \in \mathcal{F}$ for $i \in \mathbb{N}$, then*

$$\bigcap_{i\in\mathbb{N}} A_n \in \mathcal{F}.$$

Proof. Since σ-algebras are closed under complements, it suffices to show that

$$\left(\bigcap_{i\in\mathbb{N}} A_n\right)^c \in \mathcal{F}.$$

By DeMorgan's law,

$$\left(\bigcap_{i\in\mathbb{N}} A_n\right)^c = \bigcup_{i\in\mathbb{N}} A_n^c;$$

each $A_n^c \in \mathcal{F}$ since σ-algebras are closed under the taking of complements, and their union is since σ-algebras are closed under countable unions. □

The notion of a topology naturally gives rise to the notion of a topological space, which is simply a set which carries a topology. There is an equivalent notion for σ-algebras.

Definition 5 (measurable space)**.** A measurable space is a pair $(\Omega, \mathcal{F})$, where Ω is a set and $\mathcal{F}$ is a σ-algebra on Ω.

Just as in topology, it is often either impractical or impossible to list explicitly all the sets in a σ-algebra, and one often has to resort to specifying them by other means. The easiest, most straightforward way of doing this is by specifying some sets which one wants to ensure are measurable, and then taking the smallest σ-algebra which contains them. This is possible because of the following lemma.

Lemma 6. *The intersection of* any *collection $\mathcal{F}_i$ of σ-algebras on a universal set Ω is a σ-algebra.*

Proof. Let $S_1, S_2, \ldots \in \bigcap_i \mathcal{F}_i$. Then $S_1^c \in \bigcap_i \mathcal{F}_i$ since it is in each $\mathcal{F}_i$ (by definition). Similarly, $\bigcup_i S_i$ and $\bigcap_i S_i$ are in $\bigcap_i \mathcal{F}_i$. □

Definition 7 (σ-algebra generated by a collection of subsets)**.** Let $S \subset 2^\Omega$ be any collection of subsets of a set Ω. The σ-algebra generated by S is defined to be the intersection of all σ-algebras containing S.

Definition 8 (product σ-algebra)**.** Let Ω_1 and Ω_2 be two (universal) sets and $\Omega = \Omega_1 \times \Omega_2$ be their Cartesian product. The σ-algebra generated by the rectangles $A_1 \times A_2$, where $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$ is called the product σ-algebra, and is denoted $\mathcal{F}_1 \times \mathcal{F}_2$.

This may not immediately seem like the natural definition. However, it is the right one. In fact, this definition is natural for the same reason the analogous definition for topologies is natural: it satisfies the universal property for products.

Theorem 9. *Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be measurable spaces. Then the product measurable space*

$$(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$$

satisfies the following universal property: For any other measurable space $(\Xi, \mathcal{G})$ and any measurable maps $f\colon \Xi \to \Omega_1$ and $g\colon \Xi \to \Omega_2$, there exists a unique map $h\colon \Xi \to \Omega_1 \times \Omega_2$ which makes the following diagram commute.

$$\begin{tikzcd} & \Xi \arrow[dl, "f"'] \arrow[d, "h"] \arrow[dr, "g"] & \\ \Omega_1 & \Omega_1 \times \Omega_2 \arrow[l, "\pi_1"] \arrow[r, "\pi_2"'] & \Omega_2 \end{tikzcd}$$

Theorem 10. *The product $\mathcal{F}_1 \times \mathcal{F}_2$ is generated by 'cylinders' of the form $\Omega_1 \times A_j$ and $A_i \times \Omega_2$, $A_j \in \mathcal{F}_2$, $A_i \in \mathcal{F}_1$.*

Proof. Certainly, cylinders of this form are in particular rectangles, so the σ-algebra generated by cylinders must be contained in the σ-algebra generated by rectangles.

If we can show that the σ-algebra containing rectangles must contain the σ-algebra generated by cylinders, we will be done. But this is trivial, because any rectangle $A_1 \times A_2$ can be expressed as the intersection of the rectangles

$$(A_1 \times \Omega_2) \cap (\Omega_1 \times A_2).$$

so the σ-algebra containing cylinders contains all rectangles, hence the σ-algebra containing them. □

1.1.2 Monotone classes

Definition 11 (monotone class)**.** A collection of subsets $\mathcal{G}$ is a monotone class if it satisfies the following two conditions.

1. If A_n is an increasing countable sequence of sets in $\mathcal{G}$, i.e.

$$A_1 \subset A_2 \subset \cdots \subset A_n \subset \cdots,$$

then

$$\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{G}.$$

2. If A_n is an decreasing countable sequence of sets in $\mathcal{G}$, i.e.

$$A_1 \supset A_2 \supset \cdots \supset A_n \supset \cdots,$$

then

$$\bigcap_{n \in \mathbb{N}} A_n \in \mathcal{G}.$$

Lemma 12. *Every σ-algebra is a monotone class.*

Proof. Sigma-algebras are closed under *all* countable unions and intersections, hence they are closed under increasing unions and decreasing intersections. □

Definition 13 (monotone class generated by a set of subsets)**.** Let $\mathcal{S}$ be a set of subsets of Ω. Then the monotone class generated by $\mathcal{S}$ is the intersection of all monotone classes containing $\mathcal{S}$.

Monotone classes are not terribly interesting objects in their own right. Then why are we studying them? Because they distill the properties that an algebra of subsets must have in order to be a σ-algebra. That is, if one is studying a family of sets which is known to be an algebra, and one wants to check that these subsets are even a σ-algebra, then checking that *all* the properties of a σ-algebra are satisfied is more trouble than is necessary. Instead, one needs only check that $\mathcal{A}$ is a monotone class; this is enough to guarantee that $\mathcal{A}$ is a σ-algebra. This is one way of understanding the content of the monotone class theorem.

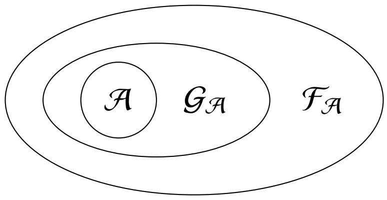
Theorem 14 (monotone class theorem)**.** *Let $\mathcal{A}$ be an algebra (not necessarily a σ-algebra!) of subsets of Ω. Then the monotone class generated by $\mathcal{A}$ is equal to the σ-algebra generated by $\mathcal{A}$.*

Proof. Denote by

- $\mathcal{G}_{\mathcal{A}}$ the monotone class generated by $\mathcal{A}$, and by
- $\mathcal{F}_{\mathcal{A}}$ the sigma-algebra generated by $\mathcal{A}$.

We need to show that $\mathcal{G}_{\mathcal{A}} = \mathcal{F}_{\mathcal{A}}$

Since every σ-algebra is a monotone class, the smallest σ-algebra containing $\mathcal{A}$ is also a monotone class which contains $\mathcal{A}$. However, there may be smaller monotone classes which contain $\mathcal{A}$, so the smallest monotone class containing $\mathcal{A}$ may be smaller than the smallest σ-algebra containing $\mathcal{A}$. Hence, $\mathcal{G}_{\mathcal{A}} \subset \mathcal{F}_{A}$.



Now the hard part: to show that $\mathcal{F}_{\mathcal{A}} \subset \mathcal{G}_{\mathcal{A}}$. To do this we need to show that $\mathcal{G}_{\mathcal{A}}$ is a σ-algebra, i.e. contains Ω and is closed under countable unions and complements. If this is true then it is a σ-algebra containing $\mathscr{A}$, hence certainly contains the smallest σ-algebra containing $\mathscr{A}$.

Our work is made much easier by the following observation: we will be done if we can show that $\mathcal{G}_{\mathcal{A}}$ is an algebra (it is *not* necessary to show that $\mathcal{G}_{\mathcal{A}}$ is a σ-algebra!). That is, if we can show closure under finite unions and intersections, countable unions and intersections will take care of themselves.

Why is this? Note that to show that $\mathcal{G}_{\mathcal{A}}$ is closed under, for example, countable unions, we would have to show that for any $\{A_i\}_{i\in\mathbb{N}} \subset \mathcal{G}_{\mathcal{A}}$,

$$\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{G}_{\mathcal{A}}.$$

We have the equality

$$\bigcup_{i\in\mathbb{N}} A_i = \bigcup_{i\in\mathbb{N}} \left(\bigcup_{j=1}^{i} A_j \right).$$

The outside union is increasing and monotone, and hence in $\mathcal{G}_{\mathcal{A}}$ because $\mathcal{G}_{\mathcal{A}}$ is a monotone class; we only need to show that each of the inside unions are in $\mathcal{G}_{\mathcal{A}}$. But these are only finite unions!

So we need only show that $\mathcal{G}_{\mathcal{A}}$ is an algebra, i.e. closed under binary unions, binary intersections, and complements. In each case, our strategy will be the same: for each property we want $\mathcal{G}_{\mathcal{A}}$ to have, we will show that the set of all sets which have this property contains $\mathcal{A}$ and is a monotone class, hence contains $\mathcal{G}_{\mathcal{A}}$. This will imply that all of the sets of $\mathcal{G}_{\mathcal{A}}$ have this property.

- **Complements:** Consider the set of all subsets of Ω whose complemements are in $\mathcal{G}_{\mathcal{A}}$. Let's call it C.

  $$C = \left\{ C \subset \Omega \,\middle|\, C^c \in \mathcal{G}_{\mathcal{A}} \right\}$$

  Certainly, $\mathcal{A}$ is in C since $\mathcal{A}$ is an algebra and $\mathcal{A} \subset \mathcal{G}_{\mathcal{A}}$.

  We can say more about C: it is a monotone class! To see this, we need to show that for a sequence $\{A_i\}_{i\in\mathbb{N}} \subset C$,

- if $A_1 \subset A_2 \subset \cdots$, then
$$\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{C}.$$
In this case, we have $A_1^c \supseteq A_2^c \supseteq \cdots$, and
$$\bigcup_{i\in\mathbb{N}} A_i = \left(\bigcap_{i\in\mathbb{N}} A_i^c\right)^c.$$
All the A_i^c are in $\mathcal{G}_{\mathcal{A}}$ by definition of $\mathcal{C}$, and their countable monotone intersection is in $\mathcal{G}_{\mathcal{A}}$ since $\mathcal{G}_{\mathcal{A}}$ is a monotone class. Since $\mathcal{C}$ is the set of all sets whose complements are in $\mathcal{G}_{\mathcal{A}}$ and $\bigcap_{i\in\mathbb{N}} A_i^c$ is in $\mathcal{G}_A$, its complement $\bigcup_{i\in\mathbb{N}} A_i$ is in $\mathcal{G}_{\mathcal{A}}$.
- If $A_1 \supseteq A_2 \supseteq \cdots$, then
$$\bigcap_{i\in\mathbb{N}} A_i \in \mathcal{C}.$$
The proof is identical, but with inclusions reversed.

We give only sketches of the proofs of the other two conditions, as they are similar to the first case.

- **Finite unions:** By induction, we only need to prove closure under binary unions. So fix an arbitrary set $A \in \mathcal{A}$ and let $\mathcal{U}$ be the set of all subsets of Ω such that the union of A and B lands in $\mathcal{G}_{\mathcal{A}}$:
$$\mathcal{U} = \left\{B \subset \Omega \,\middle|\, A \cup B \in \mathcal{G}_{\mathcal{A}}\right\}.$$
We omit the proof that this is indeed a monotone class.
- **Intersections:** As before, it suffices to show that
$$\mathcal{I} = \left\{B \subset \Omega \,\middle|\, A \cap B \in \mathcal{G}_{\mathcal{A}}\right\}.$$
is a monotone class. We omit the proof of this.

□

1.2 Some definitions from topology

In the previous definitions, we have seen parallels between measure theory and topology. In what follows, one of our main goals will be to investigate the ways in which measure theory and topology interact. To do this, we will need definitions and results from topology in earnest. However, this is not a topology course, and we will not have time to go into topology, nor can we assume a working knowledge of it. In this section we give some working definitions, so we can get on with measure theory.

Definition 15 (open set)**.** A subset O of the real line is open if it is a union of open intervals.

Definition 16 (closed set)**.** A subset C of the real line is closed if its complement is open.

Theorem 17. *If a subset $O \subset \mathbb{R}$ is open, then it can be expressed as a countable union of disjoint open intervals.*

Definition 18 (Borel σ-algebra)**.** Let (X, τ) be a topological space. The σ-algebra generated by the open sets τ is called the Borel σ-algebra.

1.3 Measures and measure spaces

1.3.1 Abstract measures

Definition 19 (abstract measure)**.** Let $(\Omega, \mathcal{F})$ be a measurable space. A measure on $(\Omega, \mathcal{F})$ is a function $\mu\colon \mathcal{F} \to \mathbb{R}^{\geq 0}$ such that for any pairwise disjoint sequence $A_i \in \mathcal{F}$, we have

$$\mu\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mu(A_i).$$

The above condition is known as *countable additivity.*

Note 20. In general, one also has the condition $\mu(\emptyset) = 0$ to ensure that $\mu(\emptyset) \neq \infty$, but we do not demand this in this class for some reason.

Example 21. Consider a general measure space $(\Omega, \mathcal{F})$, and let $a \in \Omega$. Then the function

$$\delta_a\colon \mathcal{F} \to \mathbb{R}^+; \qquad A \mapsto \begin{cases} 1, & a \in A \\ 0, & a \notin A \end{cases}$$

is a measure, called the *Dirac measure.*

Definition 22 (measure space)**.** A measure space is a triple $(\Omega, \mathcal{F}, \mu)$, where $(\Omega, \mathcal{F})$ is a measurable space (Definition 5) and μ is a measure on $(\Omega, \mathcal{F})$.

Example 23. With $\Omega = \mathbb{N}$, $\mathcal{F} = 2^\Omega$, and $\mu(A) = |A|$, $(\Omega, \mathcal{F}, \mu)$ is a measure space.

1.3.2 Lebesgue measure

The Lebesgue measure is a special case of the more general notion of measure defined above, which takes place on the real numbers $\mathbb{R}$.

Definition 24 (outer measure)**.** For any subset $S \subset \mathbb{R}$, define the outer measure μ^* of S to be

$$\mu^*(S) = \inf\left(\sum_{i\in I} \ell(I_i)\right),$$

where the infimum is taken over all coverings of S by countably many intervals I_i.

The idea is this: we imagine trying to cover our set S by sets whose measure we 'know,' namely intervals. It may be helpful to picture papier-mâché. Since each covering of our set by intervals contains our set, the measure of our set should certainly not be greater than the measure of the cover. Thus, the total length of all the intervals comprising such a cover should give us an upper bound for the measure of the set. We define the *outer measure* to be the total length of the smalles covering, or more precisely the supremum of the lengths of all such coverings.

Definition 25 (null set)**.** A null set is a set N for which

$$\mu^*(N) = 0.$$

Here is a more explicit way of phrasing this defintion: A set N is null if for any $\varepsilon > 0$ there exists a countable covering of N by intervals of total length $\leq \varepsilon$.

Definition 26 (measurable set)**.** A set $E \subset \mathbb{R}$ is measurable if for any set $A \in \mathbb{R}$,

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

The set of all measurable sets is denoted $\mathcal{M}$.

This condition is also known as the *Carathéodory criterion.*

Note 27. Roughly speaking, the reasoning behind the definition of a measurable set is as follows. A set E is measurable if it is regular enough that it can act as a screen in the following way: for *any* set A, the outer measure of A is equal to the outer measure of the part of A which is hidden by the screen plus the outer measure part of A which is visible.

There is also a more high-brow way of viewing the above definition, which can be seen in, for example, [1]. In this approach, measurable sets in the above sense are shown to be the limit points of the collection of so-called 'elementary sets,' i.e. sets which are expressible as a finite union of intervals.

Practically speaking, encountering a non-measurable set in the wild is extraordinarily rare, and their construction generally involves the invocation of such arcanities as the axiom of choice. The following is a quick sketch of the construction of a non-measurable set.

Example 28. Consider the group action on the circle S^1 by the group G of rational rotations, i.e. rotations by an angle

$$2\pi \cdot q, \qquad q \in \mathbb{Q} \cap [0, 2\pi).$$

The group G is countable; in fact it is isomorphic to $\mathbb{Q}/\mathbb{Z}$. Thus, the orbit of any point on the circle is countable, and the group action partitions S^1 into uncountably many disjoint orbits X_i. From this partition, we can create our non-measurable set as follows.

Using the axiom of choice, pick one representative x_i from each orbit X_i. This gives us an uncountable subset $Z = \bigcup_i \{x_i\} \subset S^1$. Acting on Z with each element of G in turn gives us all of S^1, so the images of Z under rational rotation form a countable partition of S^1. Since these images are obtained from each other by rotation, any rotationally invariant measure on S^1 would assign them the same measure.

Assigning any positive measure to Z would not do since countable additivity would imply that the measure of S^1 was infinite, but they also cannot be of measure zero because then by countable additivity, S^1 would have measure zero. Hence, Z cannot be measurable for any rotationally invariant measure on S^1.

Definition 29 (Lebesgue measure)**.** For any Lebesgue measurable set $M \in \mathcal{M}$, we write

$$m(M) = \mu^*(M),$$

and call $m(M)$ the Lebesgue measure of M.

Theorem 30.

1. *Any null set N is neasurable and $m(N) = 0$.*
2. *Any interval $\langle a, b\rangle$ is measureable and the measure is $b - a$.*

Proof.

1. If N is a null set, then there is a cover I_n of N by intervals such that
$$\sum_{n\in\mathbb{N}} \ell(I_n) = 0.$$
For any set $A \subset \mathbb{R}$, the intervals also cover $N \cap A$ and $N \cap A^c$, so both of these are null sets. But then
$$m^*(N \cap A) + m^*(N \cap A^c) = 0 = m(N),$$
so N is measurable.
2. The interval $I = [a, b]$ covers itself, so outer measure of I is at most $b - a$. To see that it cannot be any less, let A_i be a covering of I by intervals whose total length is less than $b - a$. Expanding the ith interval A_i by $\frac{\varepsilon}{2^i}$, we can replace it with an open interval while still covering I. This gives us an open cover of a compact set, which has a finite sub cover $\tilde{A}_i$ $i = 1, 2, \dots n$.

□

Theorem 31. *The Lebesgue measurable subsets $\mathcal{M}$ form a σ-algebra.*

Proof.

1. For any set $A \subset \mathbb{R}$, $A \cap \mathbb{R} = A$ and $A \cap \mathbb{R}^c = \emptyset$, so
$$m(A \cap \mathbb{R}) = m(A).$$
2. Suppose $E \in \mathcal{M}$. Then
$$m(E) = m(E \cap A) + m(E \cap A^c)$$
for any subset $A \subset \mathbb{R}$.

□

Corollary 32. *Any full set is measurable.*

Consider the definition of a measurable set: E is measurable if, for every set A,
$$m^*(A) = m^*(A \cap E) + m^*(A \cap E^c).$$
It is easy to show that any element of the Borel σ-algebra is measurable. There are elements of the Borel σ-algebra which have measure zero. But *any* subset of such a set will be a null set, and hence Lebesgue measurable. This is an indication of the fact that there are more Lebesgue measurable sets than there are sets in the Borel σ-algebra.

Definition 33 (completion of σ-algebra)**.** The completion of a σ-algebra $\mathcal{G}$ w.r.t. a measure μ is the smallest σ-algebra containing $\mathcal{G}$ such that if $G \in \mathcal{G}$ and $\mu(G) = 0$ then $N \in \mathcal{G}$ for all $N \subset G$.

Theorem 34. *The σ-algebra of Lebesgue-measurable subsets of $\mathbb{R}$ is the completion of the Borel σ-algebra w.r.t. the Lebesgue measure.*

2 Measure spaces in probability

2.1 Probability spaces

Measure theory was originally developed to deal with the problem of finding a consistent way to assign 'volumes' to subsets of a set. It was soon realized by Kolmogorov that the notion of a measure provided the correct way to assign probabilities to different outcomes in a consistent way.

Definition 35 (probability space)**.** A probability space is a measure space $(\Omega, \mathcal{F}, P)$ such that $P(\Omega) = 1$.

Example 36. The measure space $\big([0,1], \mathcal{M}_{[0,1]}, m\big)$ is a probability space since $m([0,1]) = 1$.

Example 37. Consider the measure space $(\mathbb{R}, \mathcal{M}, m)$, where $\mathcal{M}$ is the σ-algebra of Lebesgue measurable subsets of $\mathbb{R}$ and m is the Lebesgue measure. Pick some $B \in \mathcal{M}$.

Let

$$\mathcal{M}_B = \left\{ A \cap B \,\middle|\, A \in \mathcal{M} \right\}.$$

Then $\mathcal{M}_B$ is a σ-algebra on B. If, for $X \in \mathcal{M}_B$ we define $P(X) = \frac{m(X)}{m(B)}$, then P is a probability measure, so $(B, \mathcal{M}_B, P)$ is a probability space.

Example 38. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $B \in \mathcal{F}$. Then for any set $A \in \mathcal{F}$ we define the *conditional probability* $P(A|B)$ as follows:

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

In fact, $P(\,\cdot\,|B)$ is also a probability measure on $(\Omega, \mathcal{F})$.

2.2 Independent σ-algebras

Definition 39 (independent σ-algebras)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. Two σ-algebras $\mathcal{F}_1, \mathcal{F}_2 \subset \mathcal{F}$ are independent if for any $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$,

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2).$$

Definition 40 (independent events)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. The events A_1, A_2, $\ldots$, $A_n \in \mathcal{F}$ are independent if for all $K \subset \{1, \ldots, n\}$,

$$P\left(\bigcap_{k \in K} A_k\right) = \prod_{k \in K} P(A_k).$$

Example 41. Suppose A and B are independent events. Then the σ-algebras they generate,

$$\mathcal{F}_A = \{\emptyset, A, A^c, \Omega\} \qquad \text{and} \qquad \mathcal{F}_B = \{\emptyset, B, B^c, \Omega\}$$

are independent since

$$\begin{aligned}
P(A \cap B) &= P(A) \cdot P(B) && \text{(by definition)} \\
P(A \cap B^c) &= P(A) \cdot (1 - P(B)) && (m(A \cap B) + m(A \cap B^c) = m(A)) \\
P(A^c \cap B) &= (1 - P(A)) \cdot P(B) && \text{(by symmetry)} \\
P(A^c \cap B^c) &= (1 - P(A)) \cdot (1 - P(B)) && (m(A \cap B^c) + m(A^c \cap B^c) = m(A^c))
\end{aligned}$$

and

$$P(A^c) = 1 - P(A) \qquad \text{and} \qquad P(B^c) = 1 - P(B).$$

Independence of events is a bit nasty in the following way. Suppose we have three events A, B, and C. If A and B are independent, A and C are independent, and B and C are independent, it does not follow that A, B, and C are independent. In other words, pairwise independence does not imply independence.

This is why we have to check all possible subsets K of $\{1, \ldots, n\}$: we want events to be independent only if they are pairwise independent, triple-wise independent, etc.

2.3 Measurable functions

Certain unifying concepts lurk in the shadows behind almost every major event in modern mathematics. Possibly the most prevalent of these is the idea of a structure-preserving map. The idea is this: we have two objects A and B which have some structure (say, a topology, a binary operation, etc.), and we want to talk about maps $f\colon A \longrightarrow B$ which preserve this structure. In the case where our structures are measurable spaces (Definition 5), we call the structure-preserving maps *measurable.*

Definition 42 (measurable map)**.** Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be measurable spaces. A map $\Omega \longrightarrow \Omega'$ is measurable if for any $M \in \mathcal{F}'$, the set $f^{-1}(M) \in \mathcal{F}$.

That is, f is measurable if the preimage of every measurable set is measurable.

Lemma 43. *Let $(\Omega, \mathcal{F})$ be a measure space, let f be a measurable function on $(\Omega, \mathcal{F})$, and let $\mathcal{F} \subset \mathcal{G}$. Then f is $\mathcal{G}$-measurable.*

Proof. Let $\mathcal{B}$ be a measurable set in the codomain of f. Then since f is $\mathcal{F}$-measurable, $f^{-1}(B) \in \mathcal{F}$. But since $\mathcal{F} \subset \mathcal{G}$, $f^{-1}(B) \in \mathcal{G}$, so f is also $\mathcal{G}$-measurable. □

For the purposes of this course, we will mainly be interested in maps between (Lebesgue measurable) subsets of $\mathbb{R}$. In this case we have the following definition.

Definition 44 (Lebesgue measurable function)**.** Let $E \in \mathcal{M}$. A function $f\colon E \longrightarrow \mathbb{R}$ is Lebesgue measurable if the preimage of every Lebesgue measurable subset is Lebesgue is measurable.

Theorem 45. *The following are equivalent:*

1. *The map f is Lebesgue measurable*
2. *The preimage under f of every Borel set is Lebesgue measurable*
3. *The preimage under f of every open set is Lebesgue measurable*
4. *The preimage under f of every closed set is Lebesgue measurable*
5. *the preimave under f of every interval is Lebesgue measurable.*
6. *The preimage under f of each set of the form $(-\infty, A)$ is Lebesgue measurable.*

The set of functions $E \to \mathbb{R}$ forms an $\mathbb{R}$-algebra with addition and multiplication defined pointwise. It turns out that the Lebesgue measurable functions form a subalgebra.

Theorem 46. *If (Ω, m) is a measure space, f and g are measurable functions $\Omega \to \mathbb{R}$, and $\lambda \in \mathbb{R}$, then the following functions are also measurable.*

- $f + g$
- fg
- λf

Proof.

1. We need to show that $(f+g)^{-1}[(-\infty, A)]$ is Lebesgue measurable for every $A \in \mathbb{R}$, i.e. that the set
$$\left\{x \in \mathbb{R} \,\middle|\, f(x) + g(x) < A\right\}$$
is Lebesgue measurable.

   $f(x) + g(x) < A$ if and only there exists a rational number between $f(x) + g(x)$ and A, or equivalently between $g(x)$ and $A - f(x)$; call it r. Then $g(x) < r$ and $r < A - f(x)$, i.e. $f(x) > A - r$. Moreover, these inequalities are satisfied if and only if r is between $g(x)$ and $A - f(x)$. Thus,
$$\{x \colon f(x) + g(x) < A\} = \bigcup_{q \in \mathbb{R}} \left(f^{-1}[(-\infty, r)] \cap g^{-1}[(-\infty, A - r)]\right).$$

   The RHS is measurable because it is built from countable unions and intersections of sets which are measurable because f and g are.

2. If we can show that for any measurable function f the function f^2 is measurable, then we will be done since
$$f \cdot g = \frac{1}{2}\left((f+g)^2 - (f-g)^2\right)$$

   We need to show that the set
$$\left\{x \in \mathbb{R} \,\middle|\, f(x)^2 < A\right\}$$
is measurable for any A. But if $A \geq 0$ this is the same as the set
$$\left\{x \in \mathbb{R} \,\middle|\, f(x) < \sqrt{A} \text{ or } f(x) > -\sqrt{A}\right\},$$
and if $A < 0$, this is the empty set, which is trivially measurable.

Okay, so suppose $A \geq 0$. Then

$$(f^2)^{-1}[(-\infty, A)] = f^{-1}(-\sqrt{A}, \sqrt{A}),$$

which is the preimage of an interval and hence measurable.

3. $(\lambda f)^{-1}([a, b]) = f^{-1}\left(\left[\frac{a}{\lambda}, \frac{b}{\lambda}\right]\right)$.

□

Definition 47 (limit supremum, limit infimum)**.** Let $\{z_n\}_{n=1}^{\infty}$ be a sequence of real numbers. Then

1. $\limsup_{n\to\infty} z_n = \lim_{N\to\infty}\left(\sup_{n\geq N} z_n\right) = \inf_{N\geq 1}\left(\sup_{n\geq N} z_n\right)$;
2. $\liminf_{n\to\infty} z_n = \lim_{N\to\infty}\left(\inf_{n\geq N} z_n\right) = \sup_{N\geq 1}\left(\inf_{n\leq N} z_n\right)$;

Note 48. Here are more user-friendly 'definitions' of the limit supremum and limit infimum of a sequence z_i: $\limsup_{n\to\infty} z_n$ is the largest number that any subsequence of z_n can converge to; $\liminf_{n\to\infty}$ is the smallest.

Lemma 49. *Let a_n be a sequence of numbers. Then*

- $\limsup_{n\to\infty}(-a_n) = -\liminf_{n\to\infty} a_n$*, and*
- $\liminf_{n\to\infty}(-a_n) = -\limsup_{n\to\infty} a_n$.

Proof. We prove the first equality. The proof of the second is identical.

$$\begin{aligned}
\limsup_{n\to\infty}(-a_n) &= \lim_{N\to\infty}\left(\sup_{n\geq N}(-a_n)\right) \\
&= \lim_{N\to\infty}\left(-\inf_{n\geq N} a_n\right) \\
&= -\lim_{N\to\infty}\left(\inf_{n\geq N} a_n\right) \\
&= -\liminf_{n\to\infty} a_n.
\end{aligned}$$

□

Theorem 50. *Let $\{f_n\}$ be a sequence of measurable functions. Then*

1. $f_M(x) = \max_{n\leq k} f_n(x)$
2. $f_m(x) = \min_{n\leq k} f_n(x)$
3. $f_s(x) = \sup_{n\in\mathbb{N}} f_n(x)$
4. $f_i(x) = \inf_{n\in\mathbb{N}} f_n(x)$
5. $f_{ls}(x) = \liminf_{n\in\mathbb{N}} f_n(x)$
6. $f_{li}(x) = \limsup_{n\in\mathbb{N}} f_n(x)$

are all measurable.

Corollary 51. *If a sequence of measurable functions converges pointwise then the function to which the sequence converges is measurable.*

Theorem 52. *Let $(\Omega, \mathcal{M})$ be a measurable space. Let $E \in \mathcal{M}$, $f\colon E \to \mathbb{R}$ be measurable, and $g\colon E \to \mathbb{R}$ be a function which is equal to f almost everywhere. Then g is measurable.*

Proof. The difference $g - f$ is equal to zero almost everywhere. Consider the preimage

$$\mathrm{preim}_{g-f}\left((-\infty, a]\right),$$

Since $g - f$ is zero almost everywhere, this is a null set if $a < 0$, and a full set if $a \geq 0$. Both null sets and full sets are measurable, so $g - f$ is measurable. But then $(g - f) + f = g$ is measurable. □

Corollary 53. *Suppose that a sequence of measurable functions f_n converge to a function f almost everywhere. Then f is measurable.*

Proof. Denote the set on which f_n converges to f by S. Then $f_n \cdot \mathbb{I}_S$ converges everywhere to $f \cdot \mathbb{I}_S$. But f and $f \cdot \mathbb{I}_S$ agree almost everywhere, so f agrees with a measurable function almost everywhere, hence is itself measurable. □

2.4 Random variables

Measurable functions on probability spaces have a special name.

Definition 54 (random variable)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. A random variable X is a measurable function $\Omega \to \mathbb{R}$.

As we saw in Section 2.1, probability spaces $(\Omega, \mathcal{F}, P)$ encode the idea of a space of events and the probabilities that they occur. A random variable should be thought of as associating a number to each event.

Example 55. Consider the probability space $(\Omega, \mathcal{M}, P)$ where

$$\Omega = \{1, 2, 3, 4, 5, 6\}; \qquad \mathcal{M} = 2^{\Omega}; \qquad P(A) = \frac{|A|}{6}.$$

This is the probability space that encodes the notion of a fair die roll.

An example of a random variable is

$$X\colon \Omega \to \mathbb{R}; \qquad X(\omega) = \begin{cases} 0, & \omega \text{ even} \\ 1, & \omega \text{ odd.} \end{cases}$$

A possibly simpler example is

$$Y(\omega) = \omega,$$

which assigns to each possible outcome the value on the die.

Theorem 56. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, X a random variable. The family of sets*

$$\left\{S \in \mathcal{F} \,\middle|\, S = X^{-1}(B) \text{ for some } B \in \mathcal{B}\right\} \equiv X^{-1}(\mathcal{B}),$$

where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$, is a σ-algebra.

Proof. We need to show that (1) $\Omega \in X^{-1}(\mathcal{B})$, (2) $X^{-1}(\mathcal{B})$ is closed under complements, and (3) $X^{-1}(\mathcal{B})$ is closed under countable unions.

1. $\Omega \in \mathcal{F}$ and $\Omega = X^{-1}(\mathbb{R})$.
2. Let $A \in X^{-1}(\mathcal{B})$. Then $A = X^{-1}(B)$ for some $B \in \mathcal{B}$, so $A^c = X^{-1}(B^c)$.
3. Let $A_i \in X^{-1}(\mathcal{B})$, $i \in \mathbb{N}$. Then for each i, $A_i = X^{-1}(B_i)$ for some Borel set $B_i \in \mathcal{B}$.

   Claim:

   $$\bigcup_{i\in\mathbb{N}} A_i = X^{-1}\left(\bigcup_{i\in\mathbb{N}} B_i\right).$$

   If $x \in \bigcup_{i\in\mathbb{N}} A_i$, then x is in the preimage of at least one of the B_i, say B_n. Therefore it is in the preimage of any set which contains that B_n, such as $X^{-1}\left(\bigcup_{i\in\mathbb{N}} B_i\right)$.

   Now, suppose that $x \in X^{-1}\left(\bigcup_{i\in\mathbb{N}} B_i\right)$. Then its image under X is in $\bigcup_{i\in\mathbb{N}} B_i$, i.e. it is in at least one of the A_i, say A_m. Then it is in the union of all the A_m.

   Then we are done, since $\bigcup_{i\in\mathbb{N}} B_i \in \mathcal{B}$.

□

Definition 57 (sigma-algebra generated by a random variable)**.** For $(\Omega, \mathcal{F}, P)$ a probability space and $X : \Omega \to \mathbb{R}$ a random variable, the <u>σ-algebra generated by X</u> is the σ-algebra

$$X^{-1}(\mathcal{B})$$

from Theorem 56.

Example 58. Let $(\Omega, \mathcal{F})$ be a measure space, $f : \Omega \to \mathbb{R}$ the constant function

$$f(x) = a.$$

Then the σ-algebra generated by f is the trivial σ-algebra $\{\emptyset, \Omega\}$.

Example 59. Let $\Omega = [0, 1]$. Consider the measurable space $(\Omega, \mathcal{B}_{[0,1]})$.

Let X be a random variable defined by

$$X(\omega) = \begin{cases} a, & \omega \in [0, \frac{1}{4}) \cup [\frac{1}{2}, \frac{3}{4}) \\ b, & \omega \in [\frac{1}{4}, \frac{3}{4}) \cup [\frac{3}{4}, 1] \end{cases},$$

with $a \neq b$.

For any borel set $B \in \mathcal{B}$, $X^{-1}(B)$ can take one of four values.

- $X^{-1}(B) = \emptyset$ if $a \notin B$ and $b \notin B$.
- $X^{-1}(B) = [0, \frac{1}{4}) \cup [\frac{1}{2}, 1)$, if $a \in B$ and $b \notin B$.

- $X^{-1}(B) = [\frac{1}{4}, \frac{1}{2}) \cup [\frac{3}{4}, 1]$, if $a \notin B$ and $b \in B$
- $X^{-1}(B) = \Omega$ if $a \in B$ and $b \in B$.

Therefore, the σ-algebra $\mathcal{F}_X$ generated by X is given by

$$\mathcal{F}_X = \left\{\emptyset, \left[0, \frac{1}{4}\right) \cup \left[\frac{1}{2}, \frac{3}{4}\right), \left[\frac{1}{4}, \frac{1}{2}\right) \cup \left[\frac{3}{4}, 1\right], \Omega\right\}.$$

Here is an interpretation of the above results: *The size of the σ-algebra generated by a random variable tells you how much information the measurement of that random variable gives you.*

Example 60.

- If X is the constant random variable $X(\omega) = a$, then $X^{-1}(\mathcal{B}) = \{\emptyset, \Omega\}$. This is the smallest σ-algebra there is. Morally speaking this happens, because measuring the value of a constant function tells you nothing.
- The σ-algebra generated by the identity function on $(\mathbb{R}, \mathcal{B})$ is equal to $\mathcal{B}$. This is the biggest σ-algebra one could get. Morally, this happens because if you know the value of the identity function where you're standing, you know where you are. The same is true of any injection.
- The σ-algebra generated by the random variable

  $$X\colon [0,1] \longrightarrow \mathbb{R}; \qquad \omega \mapsto \begin{cases} 0, & \omega \in [0, \frac{1}{2}) \\ 1, & \omega \in [\frac{1}{2}, 1] \end{cases}$$

  is

  $$\mathcal{F}_X = \left\{\emptyset, [0, \frac{1}{2}), \left[\frac{1}{2}, 1\right], \Omega\right\}$$

  which reflects the fact that a measurement of X can tell you which side of $[0, 1]$ you're on, but never more than that.
- The sets in the σ-algebra generated by an even function contain subsets of the form $\{x, -x\}$, but never x without $-x$. This reflects the fact that if you measure the value of an even function, you don't know which side of the origin you're on.

Definition 61 (independent random variables)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. Random variables X_1, X_2, …, X_n are *independent* if the σ-algebras generated by them are mutually independent, i.e. if for any $B_1, B_2, \dots, B_n \in \mathcal{B}$,

$$P\left(\bigcap_{i=1}^{n} X^{-1}(B_i)\right) = \prod_{i=1}^{n} P(X_i^{-1}(B_i)).$$

Here is an interpretation of the above results: *Two random variables are independent if measuring the value of one tells you nothing about measuring the value of the other.* To see this a bit more clearly, consider the case of two independent random variables X and Y. We say that X and Y are independent if for each $A, B \in \mathcal{B}$,

$$P\left(X^{-1}(A) \cap Y^{-1}(B)\right) = P\left(X^{-1}(A)\right) \cdot P\left(Y^{-1}(B)\right).$$

Sets of the form $X^{-1}(A)$ are exactly those which belong to $\mathcal{F}_X$, so we could equally say: X and Y are independent if for any events $C \in \mathcal{F}_X$ and $D \in \mathcal{F}_D$,

$$P(C \cap D) = P(C) \cdot P(D).$$

We can re-arrange this to say that X and Y are independent if and only if

$$P(C|D) = P(C) \qquad \text{and} \qquad P(D|C) = P(D) \qquad \text{for any } C \in \mathcal{F}_X, D \in \mathcal{F}_Y.$$

Note that the definition is symmetric: we say that X and Y are independent, not X is independent of Y.

Example 62. Consider the probability space $\big([0,1], \mathcal{M}_{[0,1]}, m\big)$, and the random variables

$$X(\omega) = \begin{cases} 0, & \omega \in [0, \frac{1}{2}) \\ 1, & \omega \in \left[\frac{1}{2}, 1\right] \end{cases}, \qquad Y(\omega) = \begin{cases} 0, & \omega \in [0, \frac{1}{4}) \cup [\frac{1}{2}, \frac{3}{4}) \\ 1 & \omega \in [\frac{1}{4}, \frac{1}{2}) \cup \left[\frac{3}{4}, 1\right] \end{cases}.$$

It is easy to check that the σ-algebras generated by X and Y are independent. This is intuitively because of the following:

A priori each value of Y is equally likely since

$$m\left(\left[0, \frac{1}{4}\right) \cup \left[\frac{1}{2}, \frac{3}{4}\right)\right) = m\left(\left[\frac{1}{4}, \frac{1}{2}\right) \cup \left[\frac{3}{4}, 1\right)\right) = \frac{1}{2}.$$

If we measure, say, $X = 0$, then the probability of either outcome is still $\frac{1}{2}$, and similarly $X = 1$, $Y = 0$, and $Y = 1$.

Example 63. Suppose we have two random variables X and Y such that

$$\mathcal{F}_X = \mathcal{F}_Y.$$

These random variables are not *necessarily* dependent, but their structure is constrained. For any $A \in \mathcal{F}_X$, we have

$$P(A) = P(A \cap A) = P(A) \cdot P(A),$$

i.e. $P(A) = 0$ or $P(A) = 1$ for all $A \in \mathcal{F}_X$.

2.5 Probability distributions

Definition 64 (probability distribution of a random variable)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X\colon \Omega \to \mathbb{R}$ be a random variable. The *probability distribution defined by* X, denoted P_X, is the measure on $(\mathbb{R}, \mathcal{B})$ defined by

$$P_X(B) = P(X^{-1}(B)).$$

That is, $P_X\colon \mathcal{B} \to \mathbb{R}$.

The value of a probability distribution $P_X(B)$ of a random variable X tells you the probability that the outcome of a measurement of the value of X lies somewhere in a set B.

Theorem 65. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X\colon \Omega \to \mathbb{R}$ be a random variable. Then P_X is a probability measure, i.e. $(\mathbb{R}, \mathcal{B}, P_X)$ is a probability space.*

Proof. Non-negativity is obvious, and $P_X(\mathbb{R}) = P(X^{-1}(\mathbb{R})) = P(\Omega) = 1$, so we need only show that if $\{B_i\}$ is a countable, pairwise disjoint collection of Borel sets, then

$$P_X\left(\bigcup_i B_i\right) = \sum_i P_X(B_i).$$

By definition,

$$P_X\left(\bigcup_i B_i\right) = P\left(X^{-1}\left(\bigcup_i B_i\right)\right).$$

Since the B_i are pairwise disjoint, so are $X^{-1}(B_i)$. The result follows since P is countably additive. □

Example 66. Recall the function $X(\omega)$ from Example 59:

$$X(\omega) = \begin{cases} a, & \omega \in [0, \frac{1}{4}) \cup [\frac{1}{2}, \frac{3}{4}) \\ b, & \omega \in [\frac{1}{4}, \frac{1}{2}) \cup [\frac{3}{4}, 1] \end{cases}, \qquad a \neq b.$$

The probability distribution generated by X, denoted P_X, has the following form. For any Borel measurable set B, $P_X(B) =$

- $P(\emptyset) = 0$ if $a \notin B$ and $b \notin B$.
- $P([0, \frac{1}{4}) \cup [\frac{1}{2}, \frac{3}{4})) = \frac{1}{2}$, if $a \in B$ and $b \notin B$.
- $P([\frac{1}{4}, \frac{1}{2}) \cup [\frac{3}{4}, 1) = \frac{1}{2}$, if $a \notin B$ and $b \in B$
- $P(\Omega) = 1$ if $a \in B$ and $b \in B$.

Thus, we could equally write

$$P_X = \frac{1}{2}\left(\delta_a + \delta_b\right),$$

where $\delta_{(\cdot)}$ is the Dirac measure from Example 21.

Example 67. Consider the probability space $\big([0,1], \mathcal{M}_{[0,1]}, m\big)$ and the random variable

$$X(\omega) = \begin{cases} 5\omega, & \omega \in \left[0, \frac{1}{3}\right] \\ 4, & \omega \in \left(\frac{1}{3}, \frac{2}{3}\right) \\ 5(1-\omega) & \omega \in \left[\frac{2}{3}, 1\right] \end{cases}.$$

First, we calculate P_X. The preimage of any set A which is contained entirely in $\left[0, \frac{5}{3}\right]$ will be mapped to two places: $A/5$ and $1 - A/5$. Each of these is shrunk by a factor of 5, and there are two of them, so $P_X(A) = \frac{2}{5}m(A)$.

If A contains 4 and does not have a component in $\left[0, \frac{5}{3}\right]$, then $X^{-1}(A)$ will contain the interval $\left(\frac{1}{3}, \frac{2}{3}\right)$.

Thus,

$$P_X(B) = \frac{1}{3}\delta_4(B) + \frac{2}{5}m\left(B \cap \left[0, \frac{5}{3}\right]\right).$$

This makes sense intuitively: on $[0, \frac{1}{3}]$, X stretches out sets by a factor of 5 so preimages will be shrunk by $\frac{1}{5}$.

Definition 68 (cumulative distribution function)**.** For any random variable X on a probability space $(\Omega, \mathcal{F}, P)$, the cumulative distribution function, or CDF, of X, denoted $F_X(t)$, is the real function

$$F_X(t) = P_X((-\infty, t]).$$

3 Integration

3.1 Integration

The preliminary results of Lebesgue integration theory take rather a different form to those from Riemann integration. The most obvious difference is that many results from Lebesgue integration theory are proved only for non-negative measurable functions, and then extended to all measurable functions via positive decomposition. (See, for example, Definition 86 for the definition of positive decomposition.)

3.1.1 Simple functions

Definition 69 (simple function)**.** Let $E \subset \mathbb{R}$. A non-negative function $\varphi\colon E \longrightarrow \mathbb{R}$ is simple if it is measurable and takes on only finitely many values $a_1, a_2, \ldots, a_n$. Equivalently, a function φ which takes on finitely many values a_i is simple if each preimage $\varphi^{-1}(\{a_i\})$ is measurable.

Definition 70 (lebesgue integral of a simple function)**.** Let $(\mathbb{R}, \mathcal{M}, m)$ be a measure space, let $E \subset \mathcal{M}$, and let $\varphi\colon E \longrightarrow \mathbb{R}$ be a simple function. Let $\varphi^{-1}(\{a_i\}) = A_i$. The Lebesgue integral of φ over E is defined by

$$\int_E \varphi \, dm = \sum_i a_i \cdot m(A_i \cap E).$$

Note 71. We explicitly allow the case in which $\int_E \varphi \, dm = \infty$. This is in stark contrast to the usual treatment, and will come back to bite us in the rear end when we define the Lebesgue integral for general measurable functions in Definition 86; see Note 87 for more details.

Example 72. Recall the function

$$f\colon \mathbb{R} \longrightarrow \mathbb{R}; \qquad f(x) = \begin{cases} 1, & x \in \mathbb{Q} \\ 0, & x \notin \mathbb{Q} \end{cases}.$$

The function f is measurable since the preimage of any (measurable) set is either null or full and hence measurable, and simple since it takes on only finitely many values (0 and 1). Its Lebesgue integral is

$$\int_{\mathbb{R}} f \, dm = 0 \cdot m(\mathbb{R} \setminus \mathbb{Q}) + 1 \cdot m(\mathbb{Q}) = 0.$$

This function is not Riemann integrable!

3.1.2 Integration of non-negative measurable functions

Definition 73 (Legesgue integral of a non-negative measurable function)**.** Let $E \in \mathcal{M}$,

$$f : E \longrightarrow \mathbb{R}$$

be a measurable function. The Lebesgue integral of f over E is

$$\sup_{\varphi \in \Phi(f,E)} \int_E \varphi \, dm,$$

where $\Phi(f, E) = \{\varphi \colon E \longrightarrow \mathbb{R} \text{ simple and } 0 \leq \varphi(x) \leq f(x) \text{ for all } x \in E\}$.

Example 74. Consider the measure space $(\Omega = \mathbb{N}, \mathcal{F} = 2^{\mathbb{N}}, \mu = |\cdot|)$ from Example 23. Any non-negative function $f : \mathbb{N} \longrightarrow \mathbb{R}$ is measurable, and the integral

$$\int_\Omega f \, d\mu = \sum_{n=1}^{\infty} f(n).$$

That is, the Lebesgue integral over $(\Omega = \mathbb{N}, \mathcal{F} = 2^{\mathbb{N}}, \mu = |\cdot|)$ is simply the sum of the series $f(n)$.

Note 75. Again, we allow $\int_E f \, dm = \infty$.

Lemma 76 (basic properties of the lebesgue integral for non-negative functions)**.**

1. *If $A \in \mathcal{M}$ and $f \leq g$ on A, then*

$$\int_A f \, dm \leq \int_A g \, dm.$$

2. *If $B \subset A$ are two measurable sets, then*

$$\int_B f \, dm \leq \int_A f \, dm.$$

3. *For any non-negative number a,*

$$\int af \, dm = a \int f \, dm.$$

4. *If A is null, then*

$$\int_A f \, dm = 0.$$

5. *If A, $B \in \mathcal{M}$ and $A \cap B = \emptyset$, then*

$$\int_{A \cup B} f \, dm = \int_A f \, dm + \int_B f \, dm.$$

Theorem 77. *Suppose f is a non-negative measurable function. Then*

$$\int_{\mathbb{R}} f \, dm = 0 \iff f \overset{\text{a.e.}}{=} 0.$$

Proof. Suppose $f = 0$ almost everywhere. Then any $\varphi \in \Phi(f, E)$ is zero almost everywhere, so

$$\int_{\mathbb{R}} \varphi \, dm = 0 \qquad \text{and} \qquad \sup_{\varphi \in \Phi(f,\mathbb{R})} \int_{\mathbb{R}} \varphi \, dm = 0.$$

Now suppose that $\int_{\mathbb{R}} f \, dm = 0$. Consider the set

$$E = \left\{ x \in \mathbb{R} \,\middle|\, f(x) > 0 \right\}.$$

We will show that

$$m(E) = 0.$$

Consider the sets

$$E_n = \left\{ x \in \mathbb{R} \,\middle|\, f(x) \geq 1/n \right\}, \qquad n \in \mathbb{N}.$$

Clearly, $E = \bigcup_{n \in \mathbb{N}} E_n$.

The function

$$\varphi(x) = \frac{1}{n} \mathbb{I}_{E_n}(x)$$

is simple since f is measurable, and $\varphi \in \Phi(f, \mathbb{R})$ since $0 \leq \varphi(x) \leq f(x)$.

If $m(E_n) > 0$ for any n, then

$$\int_{\mathbb{R}} f \, dm \geq \int_{\mathbb{R}} \varphi \, dm \geq \frac{1}{n} m(E_n) > 0,$$

a contradiction, so $m(E_n) = 0$ for all n.

Since m is a measure (hence countably subadditive),

$$m(E) = m\left(\bigcup_{n \in \mathbb{N}} E_n \right) \leq \sum_{n \in \mathbb{N}} m(E_n) = 0.$$

□

Corollary 78. *Any statement about the integrals of functions f, g which holds when $f = g$ also holds when $f \overset{\text{a.e.}}{=} g$.*

Theorem 79 (Fatou's lemma)**.** *Let f_n be a sequence of non-negative integrable functions. Then*

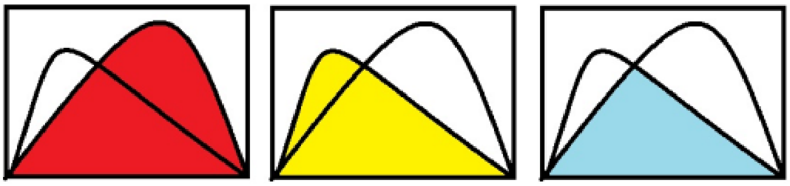
$$\liminf_{n \to \infty} \int_E f_n \, dm \geq \int_E \left(\liminf_{n \to \infty} f_n \right) dm.$$

Having searched far and wide for a nice intuitive understanding of Fatou's lemma, all I can say is that as far as I know, it is mainly a technical lemma with little intuition behind it. However, the following related fact puts it in perspective.

In the case that we consider only two non-negative measurable functions f_1 and f_2, Fatou's lemma can be understood intuitively as the following statement:

$$\min \left(\int_E f_1 \, dm, \int_E f_2 \, dm \right) \geq \int_E \min(f_1, f_2) \, dm.$$

This becomes obvious after, for example, looking at the following picture:



Example 80. The inequality in Fatou's lemma is often strict. Consider the series of functions

$$f_n(x) = n \cdot \mathbb{I}_{(0,1/n]}(x).$$

The integral of any of these functions over $\mathbb{R}$ is 1, so the LHS of the inequality in the statement of Fatou's lemma is 1. However, the functions f_n converge to 0 pointwise, hence the RHS is zero.

Note 81. Fatou's lemma is our first attempt at an understanding of when the integral is continuous as a function $\mathcal{L}^1(E) \to \mathbb{R}$, and in what way.

Theorem 82 (monotone convergence theorem)**.** *Let f_n be a sequence of non-negative measurable functions defined on a measurable set $E \in \mathcal{M}$ which increase monotonically to a function $f\colon E \to \mathbb{R}$.*[1] *Then*

$$\lim_{n\to\infty} \int_E f_n \, dm = \int_E f \, dm.$$

Proof. By part 1 of Lemma 76, we have that

$$\int_E f_n \, dm \leq \int_E f \, dm \qquad \text{for all } n.$$

Hence,

$$\limsup_{n\to\infty} \int_E f_n \, dm \leq \int_E f \, dm.$$

Fatou's lemma (Theorem 79) tells us that

$$\liminf_{n\to\infty} \int_E f \, dm \geq \int_E \left(\liminf_{n\to\infty} f_n\right) dm = \int_E f \, dm,$$

where the last equality is because $f_n(x)$ increase monotonically in n.

Hence, we have the string of inequalities

$$\int_E f \, dm \leq \liminf_{n\to\infty} \int_E f_n \, dm \leq \limsup_{n\to\infty} \int_E f_n \, dm \leq \int_E f \, dm.$$

The limit infimum is always less than or equal to the limit and the limit supremum is always greater than or equal to the limit, so if the limit infimum is equal to the limit supremum, they are both equal the limit. Hence

$$\lim_{n\to\infty} \int_E f_n \, dm = \int_E f \, dm.$$

$\square$

[1] We are guaranteed by Theorem 50 that f is measurable.

Note 83. The monotone convergence theorem gives a sufficient condition for the inequality in Fatou's lemma to be an equality: if the f_n are monotonically nondecreasing.

Example 84. One easy consequence of the MCT is that we can pull infinte sums out of any Lebesgue integral if all the terms are non-negative. That is, for any sequence of non-negative measurable functions f_n,

$$\begin{aligned}\int_E \left(\sum_{n\in\mathbb{N}} f_n\right) dm &= \int_E \lim_{N\to\infty} \sum_{n\le N} f_n \, dm \\ &= \lim_{N\to\infty} \int_E \sum_{n\le N} f_n \, dm && \text{(by MCT)} \\ &= \lim_{N\to\infty} \sum_{n\le N} \int_E f_n \, dm && \text{(by Lemma 76, prop. 5)} \\ &= \sum_{n\in\mathbb{N}} \int_E f_n \, dm.\end{aligned}$$

Theorem 85. *For any non-negative measurable function f on $E \in \mathcal{M}$, there exists a sequence φ_n of non-negative simple functions such that $\varphi_n \uparrow f$.*

Proof. The easiest way to prove something exists is by writing it down, so here we go.

$$\varphi_n = \sum_{k=0}^{2^{2n}} \frac{k}{2^n} \mathbb{I}_{f^{-1}([\frac{k}{2^n}, \frac{k+1}{2^n}))}.$$

□

3.1.3 Integration of general functions

Definition 86 (integrable function, Lebesgue integral)**.** Let $E \in \mathcal{M}$ and let f be a measurable function, and define the functions f^+ and f^- as follows.

$$f^+(x) = \max(f(x), 0); \qquad f^-(x) = \max(-f(x), 0).$$

Then $f(x) = f^+(x) + f^-(x)$ for all $x \in E$, and both f^+ and f^- are non-negative.

Then, if $\int_E f^+ \, dm$ and $\int_E f^- \, dm$ both exist and are finite, we say that f is integrable, and define the Lebesgue integral of f to be

$$\int_E f \, dm = \int_E f^+ \, dm - \int_E f^- \, dm.$$

Note 87. Definitionally speaking, we have backed ourselves into a pernicious little corner. For a *non-negative* measurable function f, the integral $\int_E f \, dm$ as defined in Definition 73 is allowed to be infinite. However, if we view the *same function* f as a general measurable function, i.e. forget that it is non-negative, then it is by Definition 86 not integrable since

$$\int_E f^+ \, dm = \int_E f \, dm = \infty,$$

which we do not allow.

Example 88. Consider again the measure space $(\Omega = \mathbb{N}, \mathcal{F} = 2^{\mathbb{N}}, \mu = |\cdot|)$ from Example 23 and Example 74. Any function $f\colon \mathbb{N} \to \mathbb{R}$ is measurable, and the integral

$$\int_\Omega f\, d\mu = \sum_{n=1}^{\infty} f(n).$$

That is, the Lebesgue integral over $(\Omega = \mathbb{N}, \mathcal{F} = 2^{\mathbb{N}}, \mu = |\cdot|)$ is simply the sum of the series $f(n)$.

Theorem 89 (basic properties of the Lebesgue integral)**.**

1. *If $E \in \mathcal{M}$ and $f \leq g$ are two integrable functions, then*

$$\int_E f\, dm \leq \int_E g\, dm.$$

2. *For any functions f, g which are integrable over $E \in \mathcal{M}$, the sum $f + g$ is also integrable.*
3. *If f is integrable over $E \in \mathcal{M}$ and $c \in \mathbb{R}$, then cf is integrable over E and*

$$\int_E cf\, dm = c \int_E f\, dm.$$

4. *If*

$$\int_A f\, dm \leq \int_A g\, dm$$

for all $A \in \mathcal{M}$, then $f \leq g$ almost everywhere.

5. *An integrable function is almost everywhere finite.*
6. *For an integrable function f and $A \in \mathcal{M}$,*

$$m(A) \cdot \inf_{x\in A} f(x) \leq \int_A f\, dm \leq m(A) \cdot \sup_{x \in A} f(x).$$

7. $\left|\int_D f\, dm\right| \leq \int_E |f|\, dm.$
8. *If $f \geq 0$ and $\int_E f\, dm = 0$, then $f = 0$ almost everywhere on E.*
9. *If $f = g$ almost everywhere on $E \in \mathcal{M}$ and f is integrable then g is integrable and*

$$\int_E f\, dm = \int_E g\, dm.$$

10. *If $A \cap B = \emptyset$ are two disjoint sets from $\mathcal{M}$ and f is integrable on A and B, then f is integrable on $A \cup B$ and*

$$\int_{A\cup B} f\, dm = \int_A f\, dm + \int_B f\, dm.$$

Theorem 90. *Let f be a non-negative measurable function. Then the map*

$$\mu\colon \mathcal{M} \to \mathbb{R}; \qquad A \mapsto \int_A f\, dm$$

is a measure.

Proof. Clearly, $\mu(A) \geq 0$ for all $A \in \mathcal{M}$. It remains only to show that for $A_i \in \mathcal{M}$ pairwise disjoint,

$$\mu\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mu(A_i).$$

Since the A_i are pairwise disjoint, we can write

$$\mathbb{I}_{\bigcup_{i=1}^{\infty} A_i} = \sum_{i=1}^{\infty} \mathbb{I}_{A_i}.$$

Then

$$\begin{aligned}
\mu\left(\bigcup_{i=1}^{\infty} A_i\right) &= \int_{\bigcup_{i=1}^{\infty} A_i} f\, dm \\
&= \int_{\mathbb{R}} \mathbb{I}_{\bigcup_{i=1}^{\infty} A_i} f\, dm \\
&= \int_{\mathbb{R}} \sum_{i=1}^{\infty} \mathbb{I}_{A_i} f\, dm \\
&= \sum_{i=1}^{\infty} \int_{\mathbb{R}} \mathbb{I}_{A_i} f\, dm && \text{(By Example 84)} \\
&= \sum_{i=1}^{\infty} \int_{A_i} f\, dm \\
&= \sum_{i=1}^{\infty} \mu(A_i).
\end{aligned}$$

□

Example 91. Let

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

Then $A \mapsto \int_A f\, dm$ is a probability measure on $\mathbb{R}$ since

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{x^2}{2}}\, dm = 1.$$

Theorem 92 (dominated convergence theorem)**.** *Let $E \in \mathcal{M}$, and let g be integrable over E. Let f_n be a sequence of measurable functions such that $|f_n| \leq g$ almost everywhere on E for all $n \geq 1$.*

If

$$f \overset{\text{a.e.}}{=} \lim_{n\to\infty} f_n,$$

then f is integrable over E and

$$\lim_{n\to\infty} \int_E f_n\, dm = \int_E f\, dm.$$

Proof. First, suppose $f_n \geq 0$ for all n.

We aim to show that

$$\limsup_{n\to\infty} \int_E f\, dm \leq \int_E f\, dm \leq \liminf_{n\to\infty} \int_E f_n\, dm.$$

If we can do this, we will be done since $\limsup \geq \liminf$.

The second inequality follows immediately from Fatou's lemma (Theorem 79) since by assumption the f_n converge to f so

$$\liminf_{n\to\infty} f_n = \lim_{n\to\infty} f_n = f.$$

For the first inequality, we apply Fatou's lemma to the functions $g - f_n$. We find

$$\liminf_{n\to\infty} \int_E (g - f_n)\, dm \geq \int_E (g - f)\, dm = \int_E g\, dm - \int_E f\, dm.$$

But

$$\begin{aligned}
\liminf_{n\to\infty} \int_E (g - f_n)\, dm &= \liminf_{n\to\infty} \left(\int_E g\, dm - \int_E f_n\, dm \right) \\
&= \int_E g\, dm + \liminf_{n\to\infty} \int_E (-f_n)\, dm \\
&= \int_E g\, dm - \limsup_{n\to\infty} \int_E f_n\, dm. && \text{(By Lemma 49.)}
\end{aligned}$$

Putting these results together, we find

$$\int_E g\, dm - \limsup_{n\to\infty} \int_E f_n\, dm \geq \int_E g\, dm - \int_E f\, dm,$$

so

$$\limsup_{n\to\infty} \int_E f_n\, dm \leq \int_E f\, dm.$$

Thus, we are done in the case $f_n \geq 0$ for all n.

Now we relax the requirement that the f_n be non-negative. In this case, we have

$$|f_n| \leq g \iff -g(x) \leq f_n(x) \leq g(x),$$

so

$$0 \leq f_n(x) + g(x) \leq 2g(x).$$

Now we can apply the above result to the non-negative series $f_n + g$ to find

$$\begin{aligned}
\lim_{n\to\infty} \int_E (f_n + g)\, dm &= \int_E (f + g)\, dm \\
\lim_{n\to\infty} \int_E f_n\, dm &= \int_E f\, dm.
\end{aligned}$$

□

Note 93. Morally, this theorem shows that the function

$$\int_E (\,\cdot\,)\, dm \colon \mathcal{L}^1(E) \to \mathbb{R}$$

is *almost* continuous. That is, we are allowed to pull the limit inside the integral as long as the integrands are always bounded in absolute value by some measurable function.

Example 94. Let $E = [1, \infty)$, and let $f_n(x) = \frac{\sqrt{x}}{1+nx^3}$. Since

$$|f_n(x)| \leq \frac{\sqrt{x}}{x^3} \qquad \text{for all } n,$$

we have

$$\lim_{n\to\infty} \int_E f_n\, dm = \int \lim_{n\to\infty} f_n\, dm = 0.$$

Theorem 95 (Beppo-Levi theorem)**.** *Let f_n be a sequence of measurable functions, let $E \in \mathcal{M}$, and suppose that*

$$\sum_{k=1}^{\infty} \int_E |f_k|\, dm < \infty.$$

Then the series

$$\sum_{k=1}^{\infty} f_k(x)$$

converges almost everywhere, its sum is integrable, and

$$\int_E \left(\sum_{k=1}^{\infty} f_k(x) \right) dm = \sum_{k=1}^{\infty} \int_E f_k\, dm.$$

Proof. Let

$$\varphi(x) = \sum_{k=1}^{\infty} \int |f_k|\, dm.$$

Since each summand is non-negative, $\varphi(x) > 0$ for all x.

By definition,

$$\int \varphi(x)\, dm = \int \sum_{k=1}^{\infty} |f_k|\, dm.$$

By Example 84, this is equal to

$$\sum_{k=1}^{\infty} \int |f_k|\, dm < \infty.$$

Hence the integrand of

$$\int \sum_{k=1}^{\infty} |f_k|\, dm$$

is integrable, so by Lemma 89, part 5 is almost everywhere finite.

Therefore for almost every x the series

$$\sum_{k=1}^{\infty} f_k(x)$$

converges absolutely, hence converges. Let

$$f(x) = \sum_{k=1}^{\infty} f_k(x).$$

Now consider the partial sums.

$$\begin{aligned}
\left|\sum_{k=1}^{n} f_k(x)\right| &\leq \sum_{k=1}^{n} |f_k(x)| && \text{(triangle inequality)} \\
&\leq \sum_{k=1}^{\infty} |f_k(x)| \\
&= \varphi(x).
\end{aligned}$$

Now

$$\begin{aligned}
\int f\, dm &= \int \lim_{n\to\infty} \sum_{k=1}^{n} f_k\, dm \\
&= \lim_{n\to\infty} \int \sum_{k=1}^{n} f_k\, dm && \left(\substack{\text{by the MCT,} \\ \text{Theorem 82}}\right) \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} \int f_k\, dm \\
&= \sum_{n=1}^{\infty} \int f_k\, dm.
\end{aligned}$$

□

Example 96. Using the Beppo-Levi theorem, calculate

$$I = \int_0^{\infty} \frac{x}{e^x - 1}\, dx.$$

Solution: Let $E = [0, \infty)$.

First, notice that we can algebraically manipulate the integrand slightly:

$$\frac{x}{e^x - 1} = x\frac{e^{-x}}{1 - e^{-x}}.$$

Since for $x \geq 0$ we always have the inequality $|e^{-x}| \leq 1$, we can expand in a geometric series

$$= \sum_{n=1}^{\infty} xe^{-nx}.$$

Let $f_n(x) = xe^{-nx}$. Then

$$\int_E f_n \, dm = \frac{1}{n^2},$$

so

$$\sum_{n=1}^{\infty} \int_E f_n \, dm = \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

A priori, we do not know that what we have calculated is actually the correct answer. It could be that

$$\int_E \sum_{n=1}^{\infty} f_n \, dm \neq \sum_{n=1}^{\infty} \int_E f_n \, dm,$$

since we are not generically allowed to pull infinite sums out of integrals. However, the Beppo-Levi theorem tells us that we *do* have this equality. More explicitly, since

$$\sum_{n=1}^{\infty} \int_E |f_n| \, dm < \infty,$$

the Beppo-Levi theorem guarantees us the following.

1. The series
$$\sum_{n=1}^{\infty} f_k$$
converges almost everywhere.
2. The function to which it conveges, i.e. $\frac{x}{e^x-1}$, is integrable.
3. We have the equality
$$\int_E \sum_{n=1}^{\infty} f_n \, dm = \sum_{n=1}^{\infty} \int_E f_n \, dm.$$

3.1.4 Integration with respect to general measures

Theorem 97. *Let $X \colon \Omega \to \mathbb{R}$ be a random variable (Definition 54) on a probability space $(\Omega, \mathcal{F}, P)$. Let P_X be the probability distribution defined by X (Definition 64). Let g be a measurable function $\mathbb{R} \to \mathbb{R}$. Then*

$$\int_\Omega g(X(\omega)) \, d\omega = \int_{\mathbb{R}} g(x) \, dP_X(x).$$

Both integrals are well-defined (or not) simultaneously.

Theorem 97 is a Lebesguian reflection of the following result in Riemann integration: Suppose $\Omega = [0, 1]$, and suppose X is a differentiable function $E \to \mathbb{R}$ with increasing, differentiable inverse $\omega(x)$. Then

$$\int_\Omega g(X(\omega)) \, d\omega = \int_{X(\Omega)} g(x) \frac{d\omega(x)}{dx} \, dx.$$

Note that the integral is now over the set $X(E) \subset \mathbb{R}$. Thus, if the random variable X satisfies the conditions listed above, we can write

$$P_X(A) = \int_A dP_X = \int_A \left(\frac{d\omega(x)}{dx} \right) \mathbb{I}_{X(\Omega)}(x) \, dx.$$

Example 98. Consider the measure space $(\mathbb{R}, \mathcal{M}, m)$, and let $X(\omega) = a$, a constant. Then for any $A \in \mathcal{M}$,

$$P_X(A) = P(X^{-1}(A)) = \begin{cases} 1, & a \in A \\ 0, & a \notin A \end{cases} = \delta_a(A).$$

We can check that both sides of the equality in Theorem 97 are indeed equal:

$$\int_\Omega g(X(\omega))\, dP(\omega) = \int_\Omega g(a)\, dP(\omega) = g(a) \stackrel{!}{=} \int_\Omega g(x) d\delta_a(x).$$

Example 99. Consider the random variable on $\left(\Omega = [0,1], \mathcal{M}_{[0,1]}, P = m\right)$ from Example 67 defined by

$$X(\omega) = \begin{cases} 5\omega, & \omega \in \left[0, \frac{1}{3}\right] \\ 4, & \omega \in \left(\frac{1}{3}, \frac{2}{3}\right) \\ 5(1-\omega) & \omega \in \left[\frac{2}{3}, 1\right] \end{cases}.$$

Recall from that example that the probability distribution generated by X is

$$P_X(B) = \frac{1}{3}\delta_4(B) + \frac{2}{5} m \left(B \cap \left[0, \frac{5}{3}\right]\right).$$

Let g be the identity function on $\mathbb{R}$. Then Theorem 97 tells us that

$$\int_\Omega X\, dP = \int_\mathbb{R} x\, dP_X.$$

Let us check this explicitly. The left-hand side is

$$\int_{\left[0, \frac{1}{3}\right]} 5\omega\, dm(\omega) + \int_{\left(\frac{1}{2}, \frac{3}{2}\right)} 4\, dm(\omega) + \int_{\left[\frac{2}{3}, 1\right]} 5(1-\omega)\, dm(\omega) = \frac{5}{18} + \frac{4}{3} + \frac{5}{18} = \frac{17}{9}.$$

The right-hand side is

$$\begin{aligned} \int_\mathbb{R} \omega\, d\left(\frac{1}{3}\delta_4(B) + \frac{2}{5} m \left(B \cap \left[0, \frac{5}{3}\right]\right)\right)(\omega) &= \frac{1}{3}\int_\mathbb{R} \omega\, d\delta_4(\omega) + \frac{2}{5}\int_\mathbb{R} \omega\, d\left(m\left(\mathbb{R} \cap \left[0, \frac{5}{3}\right]\right)\right)(\omega) \\ &= \frac{4}{3} + \frac{2}{5}\int_0^{\frac{5}{3}} \omega\, dm \\ &= \frac{4}{3} + \frac{2}{5} \cdot \frac{1}{2} \cdot \left(\frac{5}{3}\right)^2 = \frac{17}{9}. \end{aligned}$$

Definition 100 (mathematical expectation)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $X \colon \Omega \to \mathbb{R}$ be a random variable. The mathematical expectation of X is

$$\mathbb{E}(X) = \int_\Omega X\, dP = \int_\mathbb{R} x\, dP_X(x),$$

where the second equality follows from Theorem 97.

Note 101. The mathematical expectation is also commonly known as the *expectation value* or the *expected value.*

Example 102. Consider any probability space $(\Omega, \mathcal{F}, P)$. For the identically constant random variable $X(\omega) = a$ the expectation value $\mathbb{E}(X)$ is given by

$$\int_\Omega X\, dP = \int_\Omega a\, dP = a \cdot P(\Omega) = a.$$

By Theorem 97, we could also have calculated

$$\int_\mathbb{R} x\, dP_X = \int_\mathbb{R} x\, d\delta_a(x) = a.$$

Example 103. We have already seen, in Example 99, that the mathematical expectation of the random variable X defined by

$$X(\omega) = \begin{cases} 5\omega, & \omega \in \left[0, \frac{1}{3}\right] \\ 4, & \omega \in \left(\frac{1}{3}, \frac{2}{3}\right) \\ 5(1-\omega) & \omega \in \left[\frac{2}{3}, 1\right] \end{cases}$$

is $\frac{17}{9}$.

Theorem 104. *Let $\mathcal{P} = (\Omega, \mathcal{F}, P)$ be a probability space and X and Y be random variables on $\mathcal{P}$. If X and Y are independent (Definition 61), then*

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

Example 105. Consider the random variables X and Y from Example 62 defined by

$$X(\omega) = \begin{cases} 0, & \omega \in [0, \frac{1}{2}) \\ 1, & \omega \in \left[\frac{1}{2}, 1\right] \end{cases}, \qquad Y(\omega) = \begin{cases} 0, & \omega \in [0, \frac{1}{4}) \cup [\frac{1}{2}, \frac{3}{4}) \\ 1 & \omega \in [\frac{1}{4}, \frac{1}{2}) \cup \left[\frac{3}{4}, 1\right] \end{cases}.$$

As we saw in that example, these random variables are independent. We calculate:

$$\mathbb{E}(X) = \frac{1}{2}; \qquad \mathbb{E}(Y) = \frac{1}{2}; \qquad \mathbb{E}(XY) = \frac{1}{4} = \mathbb{E}(X) \cdot \mathbb{E}(Y).$$

3.2 Spaces of integrable functions

3.2.1 L^p spaces

Definition 106 (normed space)**.** Let V be a vector space. A function $\|\cdot\| \colon V \to \mathbb{R}$ is called a norm if

1. $\|x\| \geq 0$ for all $x \in V$.
2. $\|x\| = 0$ if and only if $x = 0$.
3. $\|cx\| = |c| \cdot \|x\|$.
4. $\|x + y\| \leq \|x\| + \|y\|$.

Definition 107 (L^p norm)**.** Let $E \in \mathcal{M}$ be a measurable subset of $\mathbb{R}$. Let $(E, \mathcal{M}|_E, m)$ be a measure space. For any measurable function $f\colon E \to \mathbb{R}$, let

$$\|f\|_p = \left(\int_E |f|^p \ dm\right)^{\frac{1}{p}}.$$

It is not hard to show that the function $\|\cdot\|_p$ is a norm, called the L^p-norm.

Definition 108 (L^p space)**.** Denote by $\mathcal{L}^p(E)$ the set

$$\mathcal{L}^p(E) = \left\{f\colon E \to \mathbb{R} \,\middle|\, f \text{ measurable and } \|f\|_p < \infty\right\}.$$

Let $f, g \in \mathcal{L}^p(E)$, and write

$$f \sim g \iff f \overset{\text{a.e.}}{=} g.$$

Denote

$$L^p(E) = \mathcal{L}^p(E)/\sim .$$

Then $L^p(E)$ is a normed space (Definition 106).

3.2.2 Types of convergence

Definition 109 (convergence in measure)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $\mu(\Omega) \leq \infty$, that is, the measure μ is finite. A sequence of real measurable functions f_n converges in measure to a measurable function f if for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mu\left(\{\omega\colon\ |f_n(\omega) - f(\omega)| > \varepsilon\}\right) = 0.$$

If μ is a probability (i.e. $\mu(\Omega) = 1$), then we say that $f_n \to f$ *in probability.*

That is, we pick ε first, then take the limit as $n \to \infty$.

Example 110. With the probability space $\mathcal{P} = ([0,1], \mathcal{M}_{[0,1]}, m)$, consider the random variables

$$f_n(x) = \begin{cases} 1, & x \in \left[0, \frac{1}{n}\right] \\ 0, & \text{otherwise.} \end{cases}$$

Then $f_n \to 0$ in probability since

$$m\left(\{\omega\colon\ |f_n(x)| < \varepsilon\}\right) = \begin{cases} \frac{1}{n}, & \varepsilon < 1 \\ 0, & \varepsilon \geq 1 \end{cases},$$

so $\lim_{n\to\infty} m\left(\{\omega\colon\ |f_n(x)| < \varepsilon\}\right) = 0$.

Definition 111 (Types of convergence)**.** Here is an (incomplete) list of the various types of convergence one often encounters.

1. *Point-wise convergence:* $f_n \to f$ pointwise if for all x,

$$\lim_{n\to\infty} |f_n(x) - f(x)| = 0.$$

2. *Uniform convergence:* $f_n \to f$ uniformly if

$$\lim_{n\to\infty}\left[\sup_{x\in E}|f_n(x)-f(x)|\right]=0.$$

3. *Convergence almost everywhere:* $f_n \to f$ almost everywhere if

$$m\left(\left\{x\colon \lim_{n\to\infty} f_n(x)\neq f(x)\right\}\right)=0.$$

4. *Convergence in measure:* $f_n \to f$ in measure if for any $\varepsilon > 0$,

$$\lim_{n\to\infty} m\left(\{x\colon |f_n(x)-f(x)|>\varepsilon\}\right)=0.$$

5. *Convergence in L^p norm:* $f_n \to f$ if

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

Note 112. Uniform convergence is the strongest sort of convergence listed above, in the sense that we have the following implications

- Uniform convergence $\Rightarrow$ convergence almost everywhere $\Rightarrow$ convergence in measure.
- Uniform convergence $\Rightarrow$ convergence in L^2 norm $\Rightarrow$ convergence in L^1 norm $\Rightarrow$ convergence in measure.

3.3 Chebyshev inequality

Theorem 113 (Chebyshev inequality)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let*

- *$Y > 0$ be a random variable on Ω,*
- *$\varepsilon > 0$, and*
- *$p > 0$.*

Then

$$P(\{\omega\colon Y(\omega)\geq\varepsilon\})\leq\frac{\mathbb{E}(Y^p)}{\varepsilon^p}.$$

Proof.

$$\begin{aligned}\mathbb{E}(Y^p) &= \int_\Omega Y^p\, dP\\ &\geq \int_{\{\omega\colon Y(\omega)\geq\varepsilon\}} Y^p\, dP\\ &\geq \int_{\{\omega\colon Y(\omega)\geq\varepsilon\}} \varepsilon^p\, dP\\ &\geq \varepsilon^p P(\{\omega\colon Y(\omega)\geq\varepsilon\}).\end{aligned}$$

□

Note 114. We can trivially generalize the above formula to the case in which the measure is not a probability measure. That is, if $(\Omega, \mathcal{F}, \mu)$ is a measure space and Y is a non-negative measurable map $X \to \mathbb{R}$, then identical manipulations show that

$$\mu(\{\omega\colon Y(\omega) \geq \varepsilon\}) \leq \frac{\int_\Omega Y^p \, d\mu}{\varepsilon^p}.$$

Example 115. Here is an application of the Chebyshev inequality to probability theory. Suppose one is given a random variable X on a probability space $(\Omega, \mathcal{F}, P)$ with well-defined mean

$$m_X = \mathbb{E}(X)$$

and variance

$$\sigma_X^2 = \mathbb{E}\left((X - m_X)^2\right).$$

Then the probability that a measurement of X yields a result more than k standard deviations away from m_X is

$$P(\{\omega\colon |X(\omega) - m_X| \geq k\sigma_X\}).$$

According to Chebyshev's inequality,

$$P(\{\omega\colon |X(\omega) - m_X| \geq k\sigma_X\}) \leq \frac{\mathbb{E}((X - m_X)^p)^p}{(k\sigma_X)}$$

for any $p \in \mathbb{R}$. But with $p = 2$, $\mathbb{E}((X - m_X)^2) = \sigma_X^2$, so

$$P(\{\omega\colon |X(\omega) - m_X| \geq k\sigma_X\}) \leq \frac{1}{k^2}.$$

Example 116. Here is a practical example.

> **Problem:** Find a lower bound for the probability that the average number of 'heads' in 300 tosses of a fair coin differs from $\frac{1}{2}$ by less than 0.05.
>
> **Solution:** The probability distribution for the number of heads observed in 300 coin tosses is binomial with $n = 300$ and $p = 0.5$. Let $X \sim \mathrm{Bin}(300, 0.5)$.
>
> The mean of X is $m_X = 300 \times 0.5 = 150$, and the variance is $\sigma_X^2 = 300 \times 0.5^2 = 75$. We want to know the probability that the number of heads differs from 300 by less than $300 \times 0.05 = 15$, which is
>
> $$\frac{15}{\sqrt{75}} = \sqrt{3}$$
>
> standard deviations away from m_X. By Example 115, we see that the probability that we are *more* than $\sqrt{3}$ standard deviations away from m_X is no more than $\frac{1}{3}$, so the probability that we are *less* than $\sqrt{3}$ standard deviations away from m_X no less than $\frac{2}{3}$.

Corollary 117. *Let $E \in \mathcal{M}$ be a measurable subset of $\mathbb{R}$ with $m(E) < \infty$, and let $(E, \mathcal{M}_E, m)$ be the restriction of the Lebesgue measure m to E.*

Then if $\|f_n \to f\|_1$, then $f_n \to f$ in measure.

Proof. We need to show that, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \int_\Omega |f_n - f|\, dm = 0 \implies \lim_{n\to\infty} m(\{x\colon |f_n(x) - f(x)| \geq \varepsilon\}) = 0.$$

Pick any $\varepsilon > 0$. By the Chebyshev inequality with $p = 1$,

$$\frac{\int_E |f_n(x) - f(x)|\, dm(x)}{\varepsilon} \geq \mu(\{x\colon |f_n(x) - f(x)| > \varepsilon\}).$$

Taking the limit of both sides with ε fixed, we find

$$\lim_{n\to\infty} \mu(\{x\colon |f_n(x) - f(x)| > \varepsilon\}) = \frac{1}{\varepsilon} \lim_{n\to\infty} \|f_n - f\|_1 = 0.$$

□

3.4 Product measures and the Riesz Representation Theorem

Definition 118 (standard rectangle)**.** A standard rectangle in $\mathbb{R}^2$ is the cartesian product of intervals

$$R = I_1 \times I_2,$$

which may be closed, open, half-open, etc.

Recall Definition 8 of the product σ-algebra. Roughly speaking, it is the σ-algebra generated by rectangles.

The question is how to extend the definition of a product of measurable spaces to include measure spaces. The categorical product (see Theorem 9) will not do since the canonical projections in general are not measure-preserving.

The trick is to mimic $2d$ volume integrals. Recall: in a standard vector calculus course, one shows that one can use Fubini's theorem to swap from a bona fide $2d$ volume integral to two iterated single integrals, which one knows how to compute. One can then define the area of a region as the integral of the indicator function of that region.

Definition 119 (section)**.** Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be measurable spaces, and let $A \subset \Omega_1 \times \Omega_2$. The sections of A are defined as follows:

- For $\omega_2 \in \Omega_2$, let
$$A_{\omega_2} = \{\omega_1 \in \Omega_1 \colon (\omega_1, \omega_2) \in A\}.$$
- For $\omega_1 \in \Omega_1$, let
$$A_{\omega_1} = \{\omega_2 \in \Omega_2 \colon (\omega_1, \omega_2) \in A\}.$$

Definition 120 (product measure)**.** Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be measure spaces. The product measure space is defined as follows.

- The underlying set is the Cartesian product $\Omega_1 \times \Omega_2$.

- The σ-algebra is the product σ-algebra $\mathcal{F}_1 \times \mathcal{F}_2$ (Definition 8).
- The product measure $\mu = \mu_1 \times \mu_2$ is defined via

$$\mu(A) = \int_{\Omega_1} \mu_2(A_{\omega_1})\, d\mu_1 = \int_{\Omega_2} \mu_1(A_{\omega_2})\, d\mu_2.$$

Note 121. The two above integrals are indeed equal. We don't prove this.

3.4.1 Fubini's theorem for Lebesgue integrals

Theorem 122 (Fubini's theorem)**.** *Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_1, \mathcal{F}_1, \mu_1)$ be measure spaces. Let $P = \mu_1 \times \mu_2$ be the product measure (Definition 120).*

Let $f \in L^1(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, P)$. Then the functions

$$\omega_1 \mapsto \int_{\Omega_2} f(\omega_1, \omega_2)\, d\mu_2(\omega_2)$$

and

$$\omega_2 \mapsto \int_{\Omega_1} f(\omega_1, \omega_2)\, d\mu_1(\omega_1)$$

are in $L^1(\Omega_1)$ and $L^1(\Omega_2)$ respectively, and

$$\int_{\Omega_1 \times \Omega_2} f\, dP = \int_{\Omega_1} \left(\int_{\Omega_2} f\, d\mu_2 \right) d\mu_1 = \int_{\Omega_2} \left(\int_{\Omega_1} f\, d\mu_1 \right) d\mu_2.$$

Example 123. The condition $f \in L^1(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, \mu_1 \times \mu_2)$ is an important one. Consider the probability space

$$\left([0,1]^2, \mathcal{M}^2_{[0,1]}, m^2 \right)$$

and the random variable

$$f(x,y) = \begin{cases} \frac{1}{x^2}, & y < x \\ -\frac{1}{y^2}, & x \leq y \end{cases}.$$

It is routine to check that $f \notin L^1\left([0,1]^2, \mathcal{M}^2_{[0,1]}, m^2\right)$. Therefore, we should not expect Fubini's theorem to hold. Indeed, it does not.

3.4.2 Joint distributions

Definition 124 (joint density)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $X, Y \colon \Omega \to \mathbb{R}$ be two random variables. We can consider them together as one function

$$(X, Y) \colon \Omega \to \mathbb{R}^2; \qquad \omega \mapsto (X(\omega), Y(\omega)).$$

The probability distribution defined by (X, Y) is $P_{(X,Y)}$, given by

$$P_{(X,Y)}(B) = P\left((X,Y) \in B\right), \qquad B \subset \mathbb{R}^2.$$

If we can find a function $f_{(X,Y)}(x, y)$ such that

$$P_{(X,Y)}(B) = \int_B f_{(X,Y)}(x, y)\, dm_2(x, y),$$

then we say that X and Y have a joint density, or that *the distribution* $P_{(X,Y)}$ *has density* $f_{(X,Y)}$.

Note 125. For any rectangle $R = I_1 \times I_2$, the probability $P_{(X,Y)}(R)$ tells you the probability that $X \in I_1$ and $Y \in I_2$.

The joint distribution has the following interpretation: the probability distribution $f_{(X,Y)}$ evaluated at a point (x_0, y_0) gives the probability *density* that $X = x_0$ and $Y = y_0$.

This means that integrating $f_{(X,Y)}$ over the whole Y-axis gives the probability density for X.

Example 126. Let X and Y be random variables $([0,1]^2, \mathcal{M}_{[0,1]^2}, m_2)$, and consider the random variable (X, Y) defined by

$$(X, Y)\colon [0,1]^2 \to \mathbb{R}^2; \qquad (\omega_1, \omega_2) \mapsto (X(\omega_1, \omega_2), Y(\omega_1, \omega_2)).$$

For any $B \subset \mathbb{R}^2$, let us calculate $P_{(X,Y)}(B)$. Intuitively, this is the probabililty that a random point from $[0,1]^2$ will land in B.

Here is, roughly, how we will do this. We will take each point of $[0,1]^2$ in turn, tally up the ones that land in B, and discard the ones that don't. Of course, what we really mean is that we take the volume of the part of $[0,1]^2$ which lands in B.

Certainly, the probability that the point will land *anywhere* in $\mathbb{R}^2$ is 1. This is because

$$P_{(X,Y)}(\Omega) = \int_{[0,1]^2} \mathbb{I}_{\mathbb{R}^2}(X(\omega_1, \omega_2), Y(\omega_1, \omega_2))\, dm_2(\omega_1, \omega_2) = 1,$$

since m_2 is a probability distribution.

If we want to count only those points $(\omega_1, \omega_2) \in [0,1]^2$ which land in B, we can just change the scope of the indicator function:

$$P_{(X,Y)}(B) = \int_{[0,1]}^2 \mathbb{I}_B(X(\omega_1), Y(\omega_2))\, dm_2(\omega_1, \omega_2).$$

Roughly speaking, we are counting up the points in $[0,1]^2$ which land inside B and ignoring those which don't land inside B. Think of Monte Carlo integration.

This is not necessarily an integral we can evaluate. Suppose, however, that we can invert the relationship between (ω_1, ω_2) and (X, Y); that is, we can write

$$\omega_1 = \omega_1(x, y); \qquad \omega_2 = \omega_2(x, y),$$

where (x, y) are taken from the image of $[0,1]^2$ under (X, Y).

Then

$$\int_{[0,1]^2} \mathbb{I}_B(X(\omega_1, \omega_2), Y(\omega_1, \omega_2)))\, dm_2(\omega_1, \omega_2) = \int_{(X,Y)([0,1]^2)} \mathbb{I}_B(x, y) \left|\det\left(\frac{\partial(\omega_1, \omega_2)}{\partial(x, y)}\right)\right| dm_2(x, y).$$

In this case, the joint density is

$$j_{(X,Y)}(x, y) = \left|\frac{\partial(\omega_1, \omega_2)}{\partial(x, y)}(x, y)\right|.$$

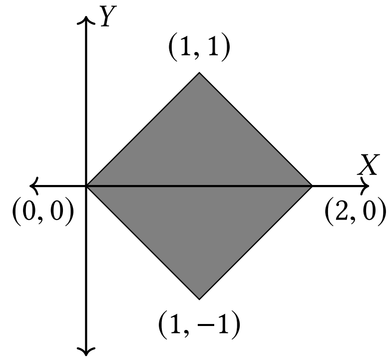
Example 127. Here is a concrete example of the above scenario.

Problem: Let $\Omega = [0,1]^2$, $\mathcal{F} = \mathcal{M}$, and $P = m_2$. Let $X = \sqrt{\omega_1} + \omega_2$, $Y = \sqrt{\omega_1} - \omega_2$.

1. What is the range of the random vector (X, Y)? Sketch the graph.
2. Find the joint density $f_{X,Y}$.
3. Calculate $P_{(X,Y)}(B)$ for $B = [0,1] \times [-1, 0]$.

Solution:

1. We have
$$X + Y = 2\sqrt{\omega_1}, \quad \text{and} \quad X - Y = 2\omega_2.$$
Thus, $0 \leq X + Y \leq 2$ and $0 \leq X - Y \leq 2$, and the range of (X, Y) looks like the following.



2. The joint density is given by the absolute value of the Jacobian determinant
$$\det\left(\frac{\partial(\omega_1, \omega_2)}{\partial(x, y)}\right)$$
with
$$\omega_1(x,y) = \left(\frac{x+y}{2}\right)^2 \quad \text{and} \quad \omega_2(x,y) = \frac{x-y}{2}.$$
We find
$$\begin{vmatrix} \frac{\partial\omega_1}{\partial x} & \frac{\partial\omega_1}{\partial y} \\ \frac{\partial\omega_2}{\partial x} & \frac{\partial\omega_2}{\partial y} \end{vmatrix} = \begin{vmatrix} \frac{x+y}{2} & \frac{x+y}{2} \\ \frac{x}{2} & -\frac{y}{2} \end{vmatrix} = -\frac{1}{4}(x+y)^2,$$
so
$$f_{(X,Y)}(x,y) = \begin{cases} \frac{1}{4}(x+y)^2, & (x,y) \in (X,Y)(\Omega) \\ 0, & \text{otherwise.} \end{cases}$$

Since the function $f_{(X,Y)}$ is zero outside the range of (X, Y), we have to choose the bounds of our integrals to accommodate this. We have
$$P_{(X,Y)}(B) = \int_0^1 dx \int_{-x}^0 dy\, \frac{(x+y)^2}{4} = \frac{1}{48}.$$

Theorem 128. *The random variables X and Y are independent if and only if*
$$P_{(X,Y)} = P_X \times P_Y.$$

3.4.3 Riesz representation theorem

Definition 129 (space of continuous bounded functions). For any topological space (X, τ), $C(X)$ denotes the space of bounded continuous functions $X \to \mathbb{R}$. This is a real vector space.

Theorem 130 (Riesz representation theorem). *Let $[a, b] \subset \mathbb{R}$, and let $\Lambda\colon C([a, b]) \to R$ be a non-negative linear functional such that*

$$\Lambda(1) = 1.$$

Then there is a unique probability measure μ on the Borel σ-algebra $\mathcal{B}_{[a,b]}$ such that

$$\Lambda(f) = \int_{[a,b]} f\, d\mu \qquad \text{for all } f \in C([a, b]).$$

Example 131. Consider the linear functional

$$\Lambda\colon C([0, 1]) \to \mathbb{R}; \qquad f \mapsto \frac{1}{2}\left(\int_{[0,1]} f\, dm + \sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^n f\left(\left(\frac{1}{3}\right)^n\right)\right).$$

The functional Λ is normalized since

$$\Lambda(1) = \frac{1}{2}\left(\int_{[0,1]} 1\, dm + \sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^n\right) = \frac{1}{2}(1 + 1) = 1.$$

It is equivalent to integration with respect to the measure

$$\frac{1}{2}\left(m + \sum_{n=1}^{\infty} \frac{1}{2^n}\delta_{\left(\frac{1}{3}\right)^n}\right).$$

3.5 The Radon-Nikodym theorem

Definition 132 (absolutely continuous). Let $(\Omega, \mathcal{F})$ be a measurable space, and let μ and ν be measures on $(\Omega, \mathcal{F})$. We say that ν is absolutely continuous with respect to μ, and write $\nu \ll \mu$, if

$$\mu(A) = 0 \implies \nu(A) = 0$$

for all $A \in \mathcal{F}$.

Note 133. Here is a way of remembering which way things go: the $\ll$ sorts of looks like the head of an double arrow, and it points the same way as the implication arrow. That is,

$$\nu(A) = 0 \impliedby \mu(A) = 0 \text{ for all } A \in \mathcal{F}; \qquad \nu \ll \mu.$$

You know the thing is pointing the right way if it looks like a ‘less than’. No idea why.

Example 134. Let ν be the measure defined on $([0,1], \mathcal{M})$ by

$$\nu(A) = \int_A x^2 \, dm.$$

Then $\nu \ll m$, since

$$m(A) = 0 \implies \int_A x^2 \, dm = 0,$$

and $m \ll \nu$ since

$$\int_A x^2 \, dm = 0 \implies m(A) = 0.$$

Example 135. Let μ be the measure on $([0,1], \mathcal{M}_{[0,1]})$ defined by

$$\mu = m + \delta_{\frac{1}{2}}.$$

Then if $\mu(A) = 0$, $m(A) = 0$ and $\delta_{\frac{1}{2}}(A) = 0$, so $m \ll \mu$.

However, $m\left(\left\{\frac{1}{2}\right\}\right) = 0$ but $\mu\left(\left\{\frac{1}{2}\right\}\right) = 1$, so $\mu \not\ll m$.

Theorem 136 (Radon-Nikodym)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space, and let μ and ν be measures on $(\Omega, \mathcal{F})$. Then $\nu \ll \mu$ if and only if there exists a non-negative measurable function $h\colon \Omega \to \mathbb{R}$ such that*

$$\nu(A) = \int_A h \, d\mu$$

for all $A \in \mathcal{F}$. We write

$$h = \frac{d\nu}{d\mu},$$

and call h the Radon-Nikodym derivative of ν with respect to μ.

Example 137. Consider the measure space $(\mathbb{R}, \mathcal{M}, m)$, and consider a probability measure P on $\mathbb{R}$. One can describe P by its probability distribution function

$$F(x) = P(X \leq x).$$

Corollary 138. *Let $(\Omega, \mathcal{F})$ be a measurable space, and $\nu \ll \mu$ two measures. Then, for any measurable function $g\colon \Omega \to \mathbb{R}$ and any $A \in \mathcal{F}$,*

$$\int_A g \, d\nu = \int_A gh \, d\mu,$$

where $h = \frac{d\nu}{d\mu}$ is the Radon-Nikodym derivative, see Theorem 136.

Proof. If we can prove the theorem in the case that g is non-negative, then we can extend to the real case by splitting g into positive and negative parts via positive decomposition (see, for example, Definition 86 for an example of this). This is what we will do.

We can always find a sequece of non-negative simple functions (Definition 69) φ_n which increase monotonically to g pointwise, i.e.

$$\varphi_n \uparrow g.$$

Therefore, by the monotone convergence theorem (Theorem 82),

$$\lim_{n\to\infty}\int_E \varphi_n\,d\nu = \int \lim_{n\to\infty} \varphi_n\,d\nu = \int_E g\,d\nu.$$

This means that we have reduced the proof to one involving simple functions, which are much better behaved.

Since simple functions only take on finitely many values, we can always write

$$\varphi_n = \sum_i a_{i,n}\mathbb{I}_{E_{i,n}},$$

where $a_{i,n}$ is the ith value that φ_n takes, and $E_{i,n}$ is the set on which $\varphi_{i,n}$ takes the value $a_{i,n}$. Then integration is simple: we have

$$\begin{aligned}
\int_A \varphi_n\,d\nu &= \int_A \left(\sum_i a_{i,n}\mathbb{I}_{E_{i,n}}\right)\,d\nu \\
&= \sum_i \int_E a_{i,n}\mathbb{I}_{E_{i,n}}\,d\nu \\
&= \sum_i a_{i,n}\nu(E_{i,n}\cap E) \\
&= \sum_i a_{i,n}\int_{E_{i,n}\cap E} d\nu \\
&= \sum_i a_{i,n}\int_{E_{i,n}\cap E} h\,d\mu \\
&= \sum_i \int_E h\,\mathbb{I}_{E_{i,n}} a_{i,n}\,d\mu \\
&= \int_E h\,\varphi_n\,d\mu.
\end{aligned}$$

Now, since $\varphi_n \uparrow g$ and g and h are both non-negative measurable functions, we must have that $\varphi_n h \uparrow gh$, hence by the monotone convergence theorem, we have

$$\lim_{n\to\infty}\int_E \varphi_n h\,d\mu = \int_E \lim_{n\to\infty}\varphi_n h\,d\mu = \int_E gh\,d\mu.$$

□

4 Applications to probability theory

4.1 Conditional expectation

For any random variable X, the sigma algebra generated by X, denoted $\mathcal{F}_X$ (see Definition 57) encodes a lot of information about X. For example, suppose one is to make a measurement of a physical quantity, and models the outcome by the random variable X. The accuracy of the measurement device controls the size of $\mathcal{F}_X$. If X can only take on 2 values (for example, a particle being either on the left or the right of a container) then the σ-algebra $\mathcal{F}_X$ contains only four elements, and takes on the schematic form $\{\emptyset, \text{left}, \text{right}, \Omega\}$. Increasing the resolution of the experiment (by, for example, further subdividing the box), increases the size of the σ-algebra; this reflects the fact that the measurement gave us more information.

Sometimes, the value that some random variable takes gives us information about the values that another random variable is likely to take. For example, if we make a measurement of the location of a particle at time t and learn that it is on the left side of the box, and wait a short enough amount of time ε, the particle is more likely to still be on the left side of the box at time $t + \varepsilon$. This increase in information is formalized by the conditional expectation.

In principle, we can compute the contitional expectation no matter what the outcome of our previous experiment was. For example, given that the particle was in $X = 0, 1$ side of the box a second ago, the expected value of the result of the future measurement of which side the particle is on can be expressed as a function of 0, 1. Thus, we can view the expected value of the future measurement (now a number between 0 and 1) as a function of the possible outcomes of the previous measurement. The conditional probability can also accomodate a situation in which we don't know which value the measurement took, but only that it took one of several values.

This sort of behavior: 'any outcome or any combination thereof' is exactly what the σ-algebra $\mathcal{F}_X$ does! Therefore, the expected value of a random variable X given some other random variable Y, denoted

$$\mathbb{E}(X|\mathcal{F}_Y),$$

should be given by a function $\Omega \to \mathbb{R}$ which is measurable with respect to $\mathcal{F}_Y$. The measurability with respect to $\mathcal{F}_X$ is exactly the condition that says that the conditional expectation is a function only of the outcome of the previous experiment, and does not contain any additional information.

What does a conditional expectation really tell us? If a previous measurement narrows down the state of our system to a subset G of its state space, the conditional expectation value is the expected value of X given that the system is somewhere in G. Mathematically, this means that for any element of the set of all possible amounts of information that X can give us, i.e. for any $G \in \mathcal{F}_X$, we should have

$$\int_G \mathbb{E}(X|\mathcal{F}_Y)\, dP = \int_G X\, dP.$$

This is all formalized in the following definition.

Definition 139 (conditional expectation)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let X be a random variable on Ω. Let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra. The conditional expectation of X relative to $\mathcal{G}$, denoted $\mathbb{E}(X|\mathcal{G})$, is a random variable on Ω defined by the following characteristics.

1. $\mathbb{E}(X|\mathcal{G})$ is $\mathcal{G}$-measurable, and
2. $\int_G \mathbb{E}(X|\mathcal{G})\, dP = \int_G X\, dP$ for all $G \in \mathcal{G}$.

Note 140. If the σ-algebra with respect to which we are taking the conditional expectation is derived from a random variable X, we will often write $\mathbb{E}(Y|X)$ instead of $\mathbb{E}(Y|\mathcal{F}_X)$.

Example 141. The result of a roll of a fair die is modelled by a random variable X. Let

$$Z = \begin{cases} 0, & X \text{ even} \\ 1, & X \text{ odd}. \end{cases}$$

Here, we are working with the probability space $(\Omega, \mathcal{F}, P)$, where

$$\Omega = \{1, 2, 3, 4, 5, 6\}; \qquad \mathcal{F} \text{ is the discrete } \sigma\text{-algebra on } \Omega; \qquad P(A) = \frac{|A|}{6}.$$

The σ-algebra $\mathcal{F}_Z$ is

$$\mathcal{F}_Z = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}.$$

Let us calculate $\mathbb{E}(X|Z)$. The conditions of Definition 139 tell us the following:

1. Since $\mathbb{E}(X|Z)$ must be measurable with respect to $\mathcal{F}_Z$, it must take the form

$$\mathbb{E}(X|Z)(\omega) = \begin{cases} a, & \omega \in \{1, 3, 5\} \\ b, & \omega \in \{2, 4, 6\}. \end{cases}$$

2. We must have

$$\int_{\{1,3,5\}} \mathbb{E}(X|Z)\, dP = \int_{\{1,3,5\}} X\, dP$$

and

$$\int_{\{2,4,6\}} \mathbb{E}(X|Z)\, dP = \int_{\{2,4,6\}} X\, dP$$

These integrals are really sums in disguise (see Example 88); we have

$$\begin{aligned} \int_{\{1,3,5\}} \mathbb{E}(X|Z)\, dP &= \int_{\{1,3,5\}} X\, dP \\ \frac{1}{6} \sum_{\omega \in \{1,3,5\}} \mathbb{E}(X|Z)(\omega) &= \frac{1}{6} \sum_{\omega \in \{1,3,5\}} X(\omega) \\ \frac{1}{2} \cdot a &= \frac{3}{2}, \end{aligned}$$

so $a = 3$. An almost identical calculation tells us that $b = 4$. Thus,

$$\mathbb{E}(X|Z) = \begin{cases} 3, & \omega \in \{1, 3, 5\} \\ 4, & \omega \in \{2, 4, 6\}. \end{cases}$$

Theorem 142. *The conditional expectation always exists and is unique up to agreement almost everywhere. That is, for $(\Omega, \mathcal{F}, P)$ a probability space, X a random variable on Ω and $\mathcal{G}$ a subalgebra of $\mathcal{F}$, the conditional expectation $\mathbb{E}(X|G)$ is uniquely defined almost everywhere.*

Proof. First, we show existence. Perform positive decomposition on X, so

$$X = X^+ - X^-.$$

Define measures ν^+ and ν^- on $(\Omega, \mathcal{G})$ as follows:

$$\nu^\pm(G) = \int_G X^\pm \, dP$$

for any $G \in \mathcal{G}$. Also define μ to be the restriction of P to $\mathcal{G}$.

If $G_0 \in \mathcal{G}$ is null with respect to P, then

$$\nu^\pm(G_0) = \int_{G_0} X^\pm \, d\mu = \int_{G_0} X^\pm \, dP = 0,$$

so $\nu^\pm \ll \mu$. Thus, the Radon-Nikodym theorem (Theorem 136) tells us that there exist non-negative measurable functions $h^\pm$ such that

$$\nu^\pm(B) = \int_B h^\pm \, d\mu \quad \text{for all } B \in \mathcal{B}.$$

Now,

$$\begin{aligned} \nu^+(G) - \nu^-(G) &= \int_G X^+ \, d\mu - \int_G X^- \, d\mu \\ &= \int_G X \, d\mu \\ &= \int_G X \, dP, \end{aligned}$$

but

$$\nu^+(G) - \nu^{-1}(G) = \int_G \left(h^+ - h^-\right) dP,$$

so for any $G \in \mathcal{G}$,

$$\int_B X \, dP = \int_B h^+ - h^- \, dP.$$

□

Lemma 143 (properties of the conditional expectation)**.**

1. $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$.
2. *If X is $\mathcal{G}$-measurable, then $\mathbb{E}(X|\mathcal{G}) = X$.*
3. *If X is independent of $\mathcal{G}$, then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$*
4. $\mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G})$.
5. *If X_n, $n = 1, 2, \ldots$ are non-negative random variables increasing to X almost everywhere, then $\mathbb{E}(X_n|\mathcal{G}) = \mathbb{E}(X|\mathcal{G})$ almost everywhere.*

6. *If Y is $\mathcal{G}$-measurable and XY is integrable, then $\mathbb{E}(XY|\mathcal{G}) = Y \cdot \mathbb{E}(X|G)$.*
7. *If $\mathcal{H} \subset \mathcal{G}$, then $\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(X|\mathcal{H})$.*

Example 144. On the probability space $\left([-1,1], \mathcal{B}_{[-1,1]}, \frac{1}{2}m\right)$, consider the random variable $X(\omega) = \omega^2$. The σ-algebra generated by X is given by

$$\left\{A \cup (-A) \,\middle|\, A \in \mathcal{B}_{[-0,1]}\right\}.$$

For any random variable Y on $[-1,1]$, the conditional expectation $\mathbb{E}(Y|X)$ is given by

$$\mathbb{E}(Y|X)(\omega) = \frac{Y(\omega) - Y(-\omega)}{2}.$$

4.2 Random processes

Notation 145. Let $\mathcal{F}_i$, $i = 1, 2, \ldots, n$, be σ-algebras on a set Ω. Denote by

$$\bigwedge_i \mathcal{F}_i = \mathcal{F}_1 \wedge \mathcal{F}_2 \wedge \cdots \wedge \mathcal{F}_n$$

the σ-algebra generated by the set $\bigcup_i \mathcal{F}_i$, i.e. the smallest σ-algebra which contains all the $\mathcal{F}_i$.

Definition 146 (discrete-time random process). A discrete-time random process (or *DTRP*) with real values on a probability space $(\Omega, \mathcal{F}, P)$ is a function

$$X_t(\omega) : \{1, 2, \ldots, n\} \times \Omega \to \mathbb{R}$$

such that for each $t_0 \in \{1, 2, \ldots, n\}$, $\mathcal{F}_{t_0}(\omega)$ is a random variable on $(\Omega, \mathcal{F}, P)$. For a fixed $\omega \in \Omega$, the sequence

$$\{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\}$$

is called the sample path (or *trajectory*) or ω.

Definition 147 (filtration, adapted process, natural). A collection $\{\mathcal{F}_t\}$ of σ-algebras on a probability space $(\Omega, \mathcal{F}, P)$ is called a filtration if $\mathcal{F}_t \subset \mathcal{F}_{t+1} \subset \mathcal{F}$ for all $t \in \{1, 2, \ldots, n\}$. A process X is adapted to a filtration $\{\mathcal{F}_t\}$ if X_t is $\mathcal{F}_t$-measurable for all t. The quadruple $(\Omega, \mathcal{F}, \{\mathcal{F}_i\}, P)$ is called a filtered probability space. The natural filtration is defined by

$$\mathcal{F}_t = \bigwedge_{i=1}^{t} \mathcal{F}_{X_i}.$$

Definition 148 (filtered probability space). A filtered probability space is a quadruple

$$(\Omega, \mathcal{F}, \{\mathcal{F}_t, P\}),$$

where $(\Omega, \mathcal{F}, P\})$ is a probability space and $\{\mathcal{F}_t\}$ is a filtration.

Lemma 149. *Here is a collection of results that will be useful.*

1. *Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ be a filtered probability space, and let X be a DTRP which is adapted to $\{\mathcal{F}_i\}$. Then X_t is $\mathcal{F}_s$ measurable for all $s \geq t$.*

Proof.

1. Since $\mathcal{F}_t \subset \mathcal{F}_s$ for all $s \geq t$, by Lemma 43, X_t is $\mathcal{F}_s$ measurable for all $s \geq t$.

□

Definition 150 (martingale, submartingale, supermartingale)**.** A real-valued process X with $\mathbb{E}(|X_t|) < \infty$ for all t and adapted to a filtration $\{\mathcal{F}_t\}$ is called an $\mathcal{F}_t$-martingale if $\mathbb{E}(X_{t+s}|\mathcal{F}_t) \overset{\text{a.e.}}{=} X_t$ for all s, t. It is called a submartingale if $\mathbb{E}(X_{t+s}|\mathcal{F}_t) \overset{\text{a.e.}}{\geq} X_t$, and a supermartingale if $\mathbb{E}(X_{t+s}|\mathcal{F}_t) \overset{\text{a.e.}}{\leq} X_t$.

Note 151. For me, this terminology is confusing. Naïvely, a *submartingale* would be a process for which the average value decreased over time, and a *supermartingale* a process for which the average value increased; this is not the case. Oh well.

Corollary 152. *Let X be an an adapted, real-valued process on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ with $\mathbb{E}(|X_t|) < \infty$ for all $t \in \{1, 2, \dots\}$. Suppose that for all $t \in \{1, 2, \dots\}$, $\mathbb{E}(X_{t+1}|\mathcal{F}_t) \overset{\text{a.e.}}{=} X_t$. Then X is martingale.*

Proof. Follows trivially by induction. More concretely,

$$X_t = \mathbb{E}(X_{t+1}|\mathcal{F}_t) = \mathbb{E}(\mathbb{E}(X_{t+2}|\mathcal{F}_{t+1})|\mathcal{F}_t) = \cdots = \mathbb{E}(\mathbb{E}(\cdots \mathbb{E}(X_{t+s}|\mathcal{F}_{t+s-1})|\mathcal{F}_{t+s-2}) \cdots |\mathcal{F}_t).$$

Property 7 of Lemma 143 tells us that if $\mathcal{H} \subset \mathcal{G}$, then $\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(X|\mathcal{H})$. Since $\{\mathcal{F}_t\}$ is a filtration, $\mathcal{F}_{t+s} \in \mathcal{F}_t$ for all t and s. Thus, the above expression collapses to

$$\mathbb{E}(X_{t+s}|\mathcal{F}_t),$$

so $\mathbb{E}(X_{t+s}|\mathcal{F}_t) = X_t$, and X_t is martingale. □

Example 153. Consider a gambler betting on the flip of a coin. If the coin comes up heads, the gambler wins \$1. If the coin comes up tails, the gambler loses \$1. Consider the wealth W of the gambler after the t^{th} toss. This is a DTRP.

- If the coin is fair, W is a martingale.
- If the coin lands on heads more often than tails, so the gambler wins on average, then W is a submartingale.
- If the coin lands on tails more often than heads, W is a supermartingale.

Example 154. Imagine a 1-dimensional box $[0, 1]$ filled with a gas. At any time

$$t \in \{0, 1, \dots, n\},$$

the position of a particular gas particle can be modelled by a random variable X_t. This is a DTRP with filtration

$$\mathcal{F}_t = \mathcal{M}_{[0,1]} \qquad \text{for all } t.$$

Bibliography

[1] Jun Tanaka and Peter F. McLoughlin. *A Realization of Measurable Sets as Limit Points.* arXiv:0712.2270 [math.FA]