Probability and Measure

Martin Orr

19th January 2009

0.1 Introduction

This is a set of notes for the Probability and Measure course in Part II of the Cambridge Mathematics Tripos. They were written as my own attempt to understand the course.

These notes are currently only a draft, so many things could be improved. Please send me your comments or corrections. If you can't find something in which is claimed to be in the Appendix, that is because the Appendix has only partially been written yet.

These notes are intended to be used in combination with Prof Rogers' examples sheets (at http://www.statslab.cam.ac.uk/~chris/). However, I have sometimes included results which are on the example sheets and I have made no attempt to ensure that material is presented here in a sensible order for doing the example sheets.

The main sources for these notes were Williams' *Probability with Martingales*, Billingsley's *Probability and Measure* and my lecture notes from Prof Rogers' lectures in Michaelmas 2007. Thanks are also due to my supervisor Dr Tehranchi for helping me to understand much of the material, and to Jacob Davis for comments and corrections.

Responsibility for the material in here, including errors, is entirely my own. Some of my own biases will have shown through. In particular, these are unashamedly the notes of a pure mathematician and I tend to emphasise analysis rather than probability.

1 Sets and measures

> [4] Measure spaces, σ-algebras, π-systems and uniqueness of extension, statement *and proof* of Carathéodory's extension theorem. Construction of Lebesgue measure on $\mathbb{R}$. The Borel σ-algebra of $\mathbb{R}$. Existence of non-measurable subsets of $\mathbb{R}$.

1.1 Probability: A reminder

In probability theory, we work with a *sample space* Ω of possible *outcomes*. An *event* is a subset A of Ω to which we assign a probability $\mathbb{P}(A)$. This should satisfy the following axioms:

1. $0 \leq \mathbb{P}(A) \leq 1$ for all events A
2. $\mathbb{P}(\Omega) = 1$
3. If $\{A_n\}$ is a finite or countable collection of pairwise disjoint events, then $\mathbb{P}(\bigcup_n A_n) = \sum_n \mathbb{P}(A_n)$

If Ω is finite or countable, then we can quite happily let all subsets of Ω be events, and indeed specifying $\mathbb{P}(\{\omega\})$ for each singleton $\{\omega\}$ is sufficient to determine the probability of all subsets by the above axioms.

1.2 Non-measurable sets

When we turn to uncountable Ω things are not so simple. Suppose we try to model the uniform distribution on $\Omega = (0, 1]$. By *uniform distribution* I mean that probabilities are unchanged by translations: for any fixed t, $\mathbb{P}(\{\omega \in \Omega | \omega + t \bmod 1 \in A\}) = \mathbb{P}(A)$. It turns out that (as long as we assume the Axiom of Choice) it is not possible to define a probability for all subsets of $(0, 1]$ which satisfies this condition as well as the above axioms.

Theorem 1 (Vitali's Theorem)**.** *There is no translation-invariant probability defined on all subsets of* $(0, 1]$.

Proof. Suppose we have a translation-invariant probability $\mathbb{P}$ defined on all subsets of $(0, 1]$.

Define an equivalence relation $\sim$ on $(0, 1]$ by $x \sim y$ if $x - y \in \mathbb{Q}$. By the Axiom of Choice, we can form a set A which contains one element of each equivalence class.

For each $q \in \mathbb{Q}$, let $A_q = \{a \in (0, 1] | a + q \bmod 1 \in A\}$. Then the A_q are disjoint, have union $(0, 1]$ and there are countably many. So $\sum_q \mathbb{P}(A_q) = \mathbb{P}((0, 1]) = 1$.

However, $\mathbb{P}(A_q) = \mathbb{P}(A)$ for all $q \in \mathbb{Q}$. So if $\mathbb{P}(A) = 0$ then $\sum_q \mathbb{P}(A_q) = 0$ while if $\mathbb{P}(A) = \epsilon > 0$ then $\sum_q \mathbb{P}(A_q) = \infty$, giving a contradiction. □

Another property we might expect the uniform distribution to have is that $\mathbb{P}((a, b]) = b - a$ for all a, b. It is not too hard to show that this is implied by translation invariance, but does the reverse implication hold? Does there exist a probability on $(0, 1]$ which takes the right value on all intervals? In fact this turns out to lead deep into set theory and be undecidable with the usual axioms. For those who care about such things, the existence of such a probability is equiconsistent with the existence of a measurable cardinal.

1.3 σ-fields

Since we can't satisfactorily assign a probability to all subsets of $(0, 1]$, we restrict our notion of what an event is. Let $\mathcal{F} \subset \wp(\Omega)$ be the set of events. To properly define the $\mathcal{F}$ we will work on, we need to take a much more abstract approach, and write down the properties we would like $\mathcal{F}$ to have, then define $\mathcal{F}$ to be the smallest collection of sets with these properties. This has the benefit that it allows us to develop a more general theory that can be applied to any collection of sets with the required properties, not just subsets of $(0, 1]$.

The main properties we require of $\mathcal{F}$ are that the sets mentioned in the axioms for probability should exist. It also seems natural that if A is an event, not-A should be as well. So the properties are:

1. $\Omega \in \mathcal{F}$.
2. If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$ (complement with respect to Ω).
3. If $\{A_n\}$ is a finite or countable subcollection of $\mathcal{F}$, then $\bigcup_n A_n \in \mathcal{F}$.

Any collection $\mathcal{F} \subset \wp(\Omega)$ with these properties is called a *σ-field*. Note that, by 2 and 3, $\mathcal{F}$ is closed under countable intersections as well.

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

A subset A of Ω is a *measurable set* with respect to a σ-field $\mathcal{F}$ if $A \in \mathcal{F}$. (This is the same as the probabilistic term *event*.)

A *measure* is a function $\mu : \mathcal{F} \to [0, \infty]$ on a σ-field s.t. if $\{A_n\} \subset \mathcal{F}$ is finite or countable and the A_n are pairwise disjoint, then $\mu(\bigcup_n A_n) = \sum_n \mu(A_n)$.

μ is a *probability (measure)* if $\mu(\Omega) = 1$ (note that this agrees with the earlier definition of probability).

μ is a *finite measure* if $\mu(\Omega) < \infty$.

μ is a *σ-finite measure* if there are countably many sets A_n, each with $\mu(A_n) < \infty$, s.t. $\bigcup_n A_n = \Omega$.

A *measure space* is a triple $(\Omega, \mathcal{F}, \mu)$ where Ω is a set, $\mathcal{F}$ is a σ-field on Ω and μ is a measure on $\mathcal{F}$.

1.4 The Borel σ-field

To continue with the scheme outlined in the previous section, we want to define our uniform probability on the smallest σ-field on $(0, 1]$ containing the half-open intervals. What does this mean?

Observe that for any collection of σ-fields on some base set Ω, their intersection is itself a σ-field (it is straightforward to check the definition). So, given any collection C of subsets of Ω, $\bigcap\{\sigma\text{-fields containing } C\}$ is itself a σ-field. It is contained in every σ-field containing C, so it makes sense to call this the smallest σ-field containing C. Note that this really exists because $\wp(\Omega)$ is itself a σ-field, so there is at least one σ-field containing C. (Compare this with defining the closure of a subset S of a topological space as the smallest closed set containing S.)

The smallest σ-field containing C is called the *σ-field generated by C*, written $\sigma(C)$.

For a topological space (X, T), the *Borel σ-field* is $\sigma(T)$, the σ-field generated by the open sets.

Lemma 1. *The σ-field on $(0,1]$ generated by half-open intervals $(a,b]$ is the same as $B((0,1])$.*

Proof. Let S be the set of half-open intervals $(a,b]$ and T the open sets of $(0,1]$.

If $b < 1$, then $(a,b] = \bigcap_n (a, b+1/n) \in B((0,1])$ and if $b = 1$, certainly, $(a,b] \in B((0,1])$. So $S \subset B((0,1])$, so as $B((0,1])$ is a σ-field, $\sigma(S) \subset B((0,1])$.

The open interval $(a,b) = \bigcup_n (a, b-1/n] \in \sigma(S)$, and any interval $(a,1] \in \sigma(S)$. Every open subset of $(0,1]$ can be written as a countable union of open intervals and sets of the form $(a,1]$ (since every connected component of our open set contains a rational). So $T \subset \sigma(S)$, and $B((0,1]) = \sigma(T) \subset \sigma(S)$. □

1.5 Carathéodory Extension Theorem

Now we have our σ-field $B((0,1])$ of events, we have to ensure that there is in fact a suitable measure defined on them. We do this using the Carathéodory Extension Theorem.

In order to state this theorem, we need weaker versions of some of the definitions concerning measures. There are several definitions of families of sets obeying certain closure properties, most of which are unimportant and I have relegated to a table at the end of this section. What we need now is a *field*, with the properties:

1. If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}_0$ (complement with respect to Ω).
2. $\Omega \in \mathcal{F}_0$.
3. If $\{A_n\}$ is a finite subcollection of $\mathcal{F}_0$, then $\bigcup_n A_n \in \mathcal{F}_0$.

A function $\mu_0 : \mathcal{F}_0 \to [0, \infty]$ on a field $\mathcal{F}_0$ is *countably additive* if whenever $\{A_n\} \subset \mathcal{F}_0$ is finite or countable, the A_n are pairwise disjoint and $\bigcup_n A_n \in \mathcal{F}_0$, then $\mu_0(\bigcup_n A_n) = \sum_n \mu_0(A_n)$. (So a measure is a countably additive function on a σ-field.) *Finitely additive* is defined similarly, with the condition only applied to finite subfamilies $\{A_n\}$.

Theorem 2 (Carathéodory Extension Theorem)**.** *If $\mathcal{F}_0$ is a field and $\mu_0 : \mathcal{F}_0 \to [0, \infty]$ is countably additive, then there is a measure μ on $\sigma(\mathcal{F}_0)$ s.t. $\mu(A) = \mu_0(A)$ for all $A \in \mathcal{F}_0$.*

The proof of this theorem is starred in the schedules, and is boring and mostly technical, so it is given in the Appendix. (Note that the unstarred sentence in the schedules "Construction of Lebesgue measure on $\mathbb{R}$" might require the definition of outer measure.)

1.6 Uniqueness of extension

That's all very well; this tells us that, given a countably additive function on a field, there is some measure on the generated σ-field which extends it. But there might be more than one such measure. We want to prove that there is only one. To do this requires defining a couple more types of sets, and some fairly abstract manipulation of them. I find it impossible to visualise the sets worked with here, which are larger than $\mathbb{R}$, but this doesn't really matter: you can just follow the properties defining them and only care about what they actually contain at the beginning and end.

A *π-system* is a family $\mathcal{G}$ of subsets of Ω s.t. if $A, B \in \mathcal{G}$ then $A \cap B \in \mathcal{G}$. Note that any field is a π-system.

A *d-system* is a family $\mathcal{D}$ of subsets of Ω s.t.

1. $\Omega \in \mathcal{D}$.
2. If $A, B \in \mathcal{D}$ and $A \supset B$ then $A \setminus B \in \mathcal{D}$.
3. If $\{A_n\} \subset \mathcal{D}$ is countable and $A_n \uparrow A$ (i.e. $A_n \subset A_{n+1} \forall n$ and $A = \bigcup_n A_n$) then $A \in \mathcal{D}$.

It is easy to check that the intersection of a family of d-systems is a d-system, so for any collection C of subsets of Ω we can define $d(C)$, the *d-system generated by C*, as the smallest d-system containing C, just as with σ-fields.

These are connected to σ-fields by the following two lemmas.

Lemma 2. *A collection $\mathcal{G}$ of subsets of Ω is a σ-field iff it is both a π-system and a d-system.*

Proof. The forwards direction is easy.

Now suppose $\mathcal{G}$ is both a π-system and a d-system.

From the definition of a d-system, $\Omega \in \mathcal{G}$ and for $A \in \mathcal{G}$, $A^c = \Omega \setminus A \in \mathcal{G}$.

Suppose $\{A_n\} \subset \mathcal{G}$ is countable. Then $A_n^c \in \mathcal{G}$ for each n, so as $\mathcal{G}$ is a π-system, we have

$$\left(\bigcup_{n=1}^{N} A_n\right)^c = \bigcap_{n=1}^{N} A_n^c \in \mathcal{G}$$

So $\bigcup_{n=1}^{N} A_n \in \mathcal{G}$ for each N. Hence as $\bigcup_{n=1}^{N} A_n \uparrow \bigcup_n A_n$, we get $\bigcup_n A_n \in \mathcal{G}$.

So $\mathcal{G}$ is a σ-field. □

Lemma 3 (Dynkin's Lemma)**.** *If $\mathcal{I}$ is a π-system, then $d(\mathcal{I}) = \sigma(\mathcal{I})$.*

Proof. Certainly, $\sigma(\mathcal{I})$ is a σ-field, so a d-system, containing $\mathcal{I}$, so $\sigma(\mathcal{I}) \supset d(\mathcal{I})$.

We need to show that $d(\mathcal{I})$ is a π-system. Then by the previous lemma, it is a σ-field, so $d(\mathcal{I}) \supset \sigma(\mathcal{I})$.

To show this, we will construct a subset of $d(\mathcal{I})$ and show that the subset is itself a d-system containing $\mathcal{I}$, so containing, and therefore equal to, $d(\mathcal{I})$. We will need to do this twice.

Let

$$\mathcal{D}_1 = \{A \in d(\mathcal{I})\colon A \cap X \in d(\mathcal{I}) \forall X \in \mathcal{I}\}$$

Certainly $\Omega \in \mathcal{D}_1$.

If $A, B \in \mathcal{D}_1$ and $A \supset B$, then $(A \setminus B) \cap X = (A \cap X) \setminus (B \cap X) \in d(\mathcal{I})$ for all $X \in \mathcal{I}$, so $A \setminus B \in \mathcal{D}_1$.

If $\{A_n\} \subset \mathcal{D}_1$ is countable and $A_n \uparrow A$, then $(A_n \cap X) \uparrow (A \cap X)$ so $A \cap X \in d(\mathcal{I})$ for all $X \in \mathcal{I}$, so $A \in \mathcal{D}_1$.

So $\mathcal{D}_1$ is a d-system. Also, since $\mathcal{I}$ is a π-system, $\mathcal{I} \subset \mathcal{D}_1$.

So $d(\mathcal{I}) \subset \mathcal{D}_1$; and by definition, $\mathcal{D}_1 \subset d(\mathcal{I})$. So $d(\mathcal{I}) = \mathcal{D}_1$.

Now let

$$\mathcal{D}_2 = \{B \in d(\mathcal{I})\colon B \cap Y \in d(\mathcal{I}) \forall Y \in d(\mathcal{I})\}$$

By the same argument as before, $\mathcal{D}_2$ is a d-system.

If $X \in \mathcal{I}$, then $A \cap X \in d(\mathcal{I})$ for all $A \in \mathcal{D}_1 = d(\mathcal{I})$, so $X \in \mathcal{D}_2$.

Hence $\mathcal{I} \subset \mathcal{D}_2$ so $d(\mathcal{I}) \subset \mathcal{D}_2$ so $d(\mathcal{I}) = \mathcal{D}_2$.

Hence $d(\mathcal{I})$ is closed under finite intersections, so is a π-system. □

With those out of the way, the following theorem tells us that if two measures agree on a π-system, then they agree on the σ-field generated by the π-system.

Theorem 3. *If $\mathcal{I}$ is a π-system and μ_1, μ_2 are measures on $\sigma(\mathcal{I})$ with $\mu_1(\Omega) = \mu_2(\Omega) < \infty$ and $\mu_1(A) = \mu_2(A)$ for all $A \in \mathcal{I}$, then $\mu_1 = \mu_2$.*

Proof. Let

$$\mathcal{D} = \{A \in \mathcal{I}\colon \mu_1(A) = \mu_2(A)\}$$

We show that $\mathcal{D}$ is a d-system. We are given that $\mathcal{I} \subset \mathcal{D}$, so then $\mathcal{D} \supset d(\mathcal{I}) = \sigma(\mathcal{I})$.

We are given that $\Omega \in \mathcal{D}$.

If $A, B \in \mathcal{D}$ and $A \supset B$ then

$$\mu_1(A \setminus B) = \mu_1(A) - \mu_1(B) = \mu_2(A) - \mu_2(B) = \mu_2(A \setminus B)$$

so $A \setminus B \in \mathcal{D}$.

If $\{A_n\} \subset \mathcal{D}$ is countable and $A_n \uparrow A$, then let $B_n = A_n \setminus \bigcup_{m=1}^{n-1} A_m$.

Then the B_n are pairwise disjoint, $A_N = \bigcup_{n=1}^{N} B_n$ and $A = \bigcup_{n=1}^{\infty} B_n$ so

$$\mu_j(A) = \sum_{n=1}^{\infty} \mu_j(B_n) = \lim_{N\to\infty} \sum_{n=1}^{N} \mu_j(B_n) = \lim_{N\to\infty} \mu_j(A_N)$$

for $j = 1, 2$, so as $\mu_1(A_N) = \mu_2(A_N)$ for each N, we get $\mu_1(A) = \mu_2(A)$ so $A \in \mathcal{D}$.

So $\mathcal{D}$ is a d-system. □

1.7 Borel measure on $\mathbb{R}$

To apply this theory to $B((0,1])$, let $\mathcal{F}_0$ be the family of finite unions of half-open intervals i.e. sets of the form $\bigcup_{i=1}^n (a_i, b_i]$. This is clearly a field, and $\sigma(\mathcal{F}_0) = B((0,1])$. Every element of $\mathcal{F}_0$ can be written as a finite union of disjoint intervals, and we can define $\mu_0(\bigcup_{i=1}^n (a_i, b_i]) = \sum_{i=1}^n (b_i - a_i)$ if the intervals $(a_i, b_i]$ are disjoint. It is straightforward to check that this is well-defined and finitely additive, but we need to check that it is countably additive.

Lemma 4. *μ_0 is countably additive.*

Proof. We begin by noting an alternative definition of countably additive: a function on a field is countably additive iff it is finitely additive and for any countable sequence $\{F_n\} \subset \mathcal{F}_0$ s.t. $F_n \downarrow \emptyset$ (i.e. $F_n \supset F_{n+1} \forall n$ and $\bigcap_n F_n = \emptyset$), we have $\mu_0(F_n) \downarrow 0$. This is because for any countable collection $\{A_n\} \subset \mathcal{F}_0$ of disjoint sets s.t. $A = \bigcup_n A_n \in \mathcal{F}_0$, we have $F_n = A \setminus \bigcup_{i=1}^n A_i \in \mathcal{F}_0$ with $F_n \downarrow 0$ and $\mu_0(F_n) = \mu_0(A) - \sum_{i=1}^n \mu_0(A_n)$.

So given a sequence $\{F_n\} \subset \mathcal{F}_0$ s.t. $F_n \downarrow 0$ and $\epsilon > 0$, we need to show that $\mu_0(F_n) < \epsilon$ for large enough n.

Each F_n is a finite union of disjoint half-open intervals; we can make these intervals slightly smaller to get $G_n \in \mathcal{F}_0$ such that $\mu_0(F_n \setminus G_n) < \epsilon 2^{-n}$ and $\overline{G}_n \subset F_n$ (the closure of G_n in the topological sense).

$\overline{G}_n$ are closed subsets of the compact space $[0,1]$, and $\overline{G}_n \downarrow \emptyset$ so there is some N for which

$$\bigcap_{n=1}^N \overline{G}_n = \emptyset$$

Now

$$F_N = F_N \setminus \bigcap_{n=1}^N G_n = \bigcup_{n=1}^N (F_N \setminus G_n) \subset \bigcup_{n=1}^N (F_n \setminus G_n)$$

so

$$\mu_0(F_N) \leq \sum_{n=1}^N \mu_0(F_n \setminus G_n) < \sum_{n=1}^N \epsilon 2^{-n} < \epsilon$$

□

The Carathéodory Extension Theorem then tells us that we can extend μ_0 to $\sigma(\mathcal{F}_0) = B((0,1])$. Because $\mathcal{F}_0$ is a π-system, the extension is unique. This gives us the *Borel measure* on $B((0,1])$.

1.8 Lebesgue measure on $\mathbb{R}$: Final details

To get Lebesgue measure, there is one more step (which is rarely important). Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. If $N \subset M \subset \Omega$ and $\mu(M) = 0$, then we would expect that $\mu(N) = 0$. However, N need not be in $\mathcal{F}$. So let the set of *null sets* be

$$\mathcal{N} = \{N \subset \Omega \colon N \subset M \text{ for some } M \text{ with } \mu(M) = 0\}$$

Then define $\mathcal{F}^* = \{A \cup N \colon A \in \mathcal{F}, N \in \mathcal{N}\}$, the *completion of $\mathcal{F}$ with respect to μ*. You can check that $\mathcal{F}^*$ is a σ-field, and there is a unique extension of μ to $\mathcal{F}^*$ (example sheet 1, question 9).

The *Lebesgue σ-field* of $\mathbb{R}$ is defined to be the completion of the Borel σ-field with respect to Borel measure, and *Lebesgue measure* the corresponding measure.

We know that Lebesgue measure satisfies $\mu((a,b]) = b - a$. But we started off looking for a measure satisfying the stronger condition of translation-invariance. You can check that in fact Lebesgue measure is translation-invariant, using example sheet 1, question 10.

Note also that we can define Lebesgue measure on all of $\mathbb{R}$, not just $(0,1]$, by taking a copy on $(n, n+1]$ for each $n \in \mathbb{Z}$, and letting a general set's measure be given by summing the measures of its intersections with each interval $(n, n+1]$.

1.9 lim inf and lim sup

lim inf and lim sup are operations on sequences from elementary analysis. They are not part of the schedules for any IA or IB course, so I will define them now.

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty} \sup_{m\geq n} x_m$$
$$\liminf_{n\to\infty} x_n = \lim_{n\to\infty} \inf_{m\geq n} x_m$$

The values $\sup_{m\geq n} x_m$ are decreasing as n increases, so the lim sup always exists, providing we allow it to take the values $\pm\infty$. These have the properties:

$$\liminf x_n \leq \limsup x_n$$

$\liminf x_n = \limsup x_n$ iff x_n converges, in which case they are equal to $\lim x_n$

Also useful are similar definitions for sequences of sets:

$$\limsup_{n\to\infty} A_n = \bigcap_n \bigcup_{m\geq n} A_n$$
$$\liminf_{n\to\infty} A_n = \bigcup_n \bigcap_{m\geq n} A_n$$

When the A_n are events, $\limsup A_n$ is the event "A_n happens infinitely often" (i.e. for infinitely many n) and $\liminf A_n$ is the event "eventually A_n always happens" (i.e. for all n greater than some n_0).

1.10 Set table

The different types of families of subsets of Ω are shown in the table. Note that a ring of sets is a ring in the algebraic sense, with symmetric difference as the addition operation and intersection as the multiplication. A field of sets is not an algebraic field.

Type of set	Closed under (definition)	Implies closed under
ring	symmetric difference finite intersection	subtraction finite union
σ-ring	symmetric difference countable intersection	subtraction countable union
field (also called algebra)	contains Ω complement finite union	finite intersection symmetric difference subtraction
σ-field (or σ-algebra)	contains Ω complement countable union	countable intersection symmetric difference subtraction
π-system	finite intersection	
d-system	contains Ω subtraction (where $A \supset B$) increasing countable union	complement

2 Random variables

> [2] Lebesgue-Stieltjes measure and probability distribution functions.
> Measurable functions, random variables.

2.1 Measurable functions

Having defined measure spaces, we now define their structure-preserving maps. This definition is essentially the same as the definition of continuous maps from one topological space to another:

For measurable spaces $(S_1, \mathcal{S}_1), (S_2, \mathcal{S}_2)$, a function $f : S_1 \to S_2$ is *measurable* (or properly $\mathcal{S}_1/\mathcal{S}_2$*-measurable*) if $f^{-1}(A) \in \mathcal{S}_1$ for all $A \in \mathcal{S}_2$.

The equivalent probabilistic concept is a *random variable*, defined as a measurable function from a probability space to $(\mathbb{R}, B(\mathbb{R}))$.

Lemma 5. *If $\mathcal{S}_2 = \sigma(\mathcal{T})$ and $f : S_1 \to S_2$ satisfies $f^{-1}(A) \in \mathcal{S}_1$ for all $A \in \mathcal{T}$, then f is measurable.*

Proof. Note that f^{-1} preserves set operations: $f^{-1}(\bigcup_i A_i) = \bigcup_i f^{-1}(A_i)$ etc., so $\mathcal{G} = \{A \subset S_2 \colon f^{-1}(A) \in \mathcal{S}_1\}$ has the same closure properties as $\mathcal{S}_1$, so is a σ-field.

We are given that $\mathcal{G} \supset \mathcal{T}$, so $\mathcal{G} \supset \sigma(\mathcal{T}) = \mathcal{S}_2$, so f is $\mathcal{S}_1/\mathcal{S}_2$-measurable. □

This lemma makes it easier to prove that functions are measurable by telling us two things:

1. If the two spaces are the Borel σ-fields of topological spaces, and f is continuous, then f is measurable. (Apply the lemma with $\mathcal{T} = \{\text{open sets of } S_2\}$.)
2. If $(S_2, \mathcal{S}_2) = (\mathbb{R}, B(\mathbb{R}))$, then we only need to check that $f^{-1}((-\infty, c))$ is measurable for all $c \in \mathbb{R}$, since $B(\mathbb{R}) = \sigma(\{(-\infty, c)\})$.

Measurable functions with range $(\mathbb{R}, B(\mathbb{R}))$ are the most common case, and we should check that these are closed under arithmetic operations.

Lemma 6. *If $f_1, f_2 : (S_1, \mathcal{S}_1) \to (\mathbb{R}, B(\mathbb{R}))$ are measurable, then $f_1 + f_2$, $f_1 f_2$ (pointwise multiplication) and $\lambda f_1 (\lambda \in \mathbb{R})$ are measurable.*

Proof. We will only do $f_1 + f_2$ here. The others are similar.

$$f_1(\omega) + f_2(\omega) < c \text{ iff } f_1(\omega) < q < c - f_2(\omega) \text{ for some } q \in \mathbb{Q}$$

so

$$\begin{aligned}(f_1 + f_2)^{-1}((-\infty, c)) &= \bigcup_{q \in \mathbb{Q}} \{\omega : f_1(\omega) < q \text{ and } f_2(\omega) < c - q\} \\ &= \bigcup_{q \in \mathbb{Q}} \left(f_1^{-1}((-\infty, q)) \cap f_2^{-1}((-\infty, c - q))\right) \in \mathcal{S}_1\end{aligned}$$

□

If $\{f_n\}$ is a countable sequence of measurable functions, then $\inf f_n, \sup f_n, \liminf f_n, \limsup f_n$ are also measurable.

2.2 Distributions and distribution functions

Let $X : \Omega \to \mathbb{R}$ be a random variable. The *distribution* of X is the measure $\mu_X = \mathbb{P} \circ X^{-1}$ on $(\mathbb{R}, B(\mathbb{R}))$. (Probabilists also call this the *law* of X.) You should check that this defines a probability measure.

For any Borel set A, we have $\mu_X(A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega \colon X(\omega) \in A\})$. For convenience, we abbreviate this to $\mathbb{P}(X \in A)$. (Other expressions of the form $\mathbb{P}$(logical statement) should be interpreted similarly.)

The *(cumulative) distribution function* $F_X : \mathbb{R} \to [0, 1]$ of X is

$$F_X(c) = \mathbb{P}(X \le c) = \mathbb{P} \circ X^{-1}((-\infty, c])$$

The intervals $(-\infty, c]$ form a π-system generating $B(\mathbb{R})$, so knowing F_X uniquely determines the distribution of X (by Lemma 3).

Distribution functions have some important properties:

1. F_X is increasing
2. F_X is right-continuous (since $\{\omega : X \le c + 1/n\} \downarrow \{\omega \colon X \le c\}$)
3. $\lim_{c \to -\infty} F_X(c) = 0$ and $\lim_{c \to +\infty} F_X(c) = 1$

Conversely, given a function F with these properties, it is the distribution function of some random variable. We can construct this random variable by taking Lebesgue measure on $(0, 1]$ as our probability space and setting

$$X(\omega) = \inf\{t \colon \omega \le F(t)\}$$

Claim. *This gives us an r.v. X with $F_X = F$.*

Proof. Fix ω. We can find $t > X(\omega)$ and arbitrarily close to $X(\omega)$ s.t. $F(t) \geq \omega$. So by the right-continuity of F, $F(X(\omega)) \geq \omega$.

So if $X(\omega) \leq c$, then as F is increasing, $\omega \leq F(X(\omega)) \leq F(c)$.

Conversely, if $\omega \leq F(c)$, then it is immediate from the definition of X that $X(\omega) \leq c$.

Hence $\{\omega : X \leq c\} = \{\omega : \omega \leq F(c)\}$, which is measurable, so X is measurable and

$$F_X(c) = \mathbb{P}(X \leq c) = \mathbb{P}(0 < \omega \leq F(c)) = F(c) \qquad \square$$

The distribution of X gives us a measure μ_F for which the random variable $I(\omega) = \omega$ has distribution function F.

This is called the *Lebesgue-Stieltjes measure* corresponding to F.

You can also construct the Lebesgue-Stieltjes measure by starting from $\mu_F((a, b]) = F(b) - F(a)$ and following the same procedure as for the construction of Lebesgue measure, using the Carathéodory Extension Theorem. The only bit which requires a bit of extra work is checking that given a finite union of half-open intervals you can fit a compact set inside it while reducing its measure by less than $\epsilon 2^n$, which uses the right-continuity of F.

2.3 Generated σ-fields

For a function $f : \Omega \to (S_2, \mathcal{S}_2)$, the *$\sigma$-field generated by f*, $\sigma(f)$, is defined to be the smallest σ-field on Ω with respect to which f is measurable. This can be defined, as with the σ-field generated by a set, as the intersection of all σ-fields for which f is measurable, but more directly

$$\sigma(f) = \{f^{-1}(A) \colon A \in \mathcal{S}_2\}$$

This is because $\{f^{-1}(A) \colon A \in \mathcal{S}_2\}$ certainly must be contained in a σ-field for f to be measurable, but you can check that it is itself a σ-field.

If X is a random variable on $(\Omega, \mathcal{F})$, then of course $\sigma(X) \subset \mathcal{F}$, but it can be much smaller. For example, if $(\Omega, \mathcal{F}) = ((0, 1], B((0, 1]))$ and

$$X(\omega) = \begin{cases} 0 & \text{if } 0 < \omega \leq \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} < \omega \leq 1 \end{cases}$$

then $\sigma(X)$ is just $\{\emptyset, (0, \frac{1}{2}], (\frac{1}{2}, 1], (0, 1]\}$.

More generally, given a collection $\{f_\lambda \colon \lambda \in \Lambda\}$ of functions $\Omega \to (S_2, \mathcal{S}_2)$, the *$\sigma$-field generated by $\{f_\lambda\}$*, $\sigma(\{f_\lambda\})$, is the smallest σ-field on Ω with respect to which all the f_λ are measurable.

Intuitively, if X_λ are random variables, then $\sigma(X_\lambda)$ is the set of events A for which, if we know the values of all the variables X_λ, then we know whether A occurred or not.

3 Independence

> [2] Independence of events, independence of σ-algebras. The Borel-Cantelli lemmas. Kolmogorov's 0-1 law.
>
> Independence of random variables.

3.1 Definitions

So far, the ideas we have covered are interesting both from a pure mathematical and from a probabilistic perspective, although probability has often provided the easiest motivations and examples. Independence however is a fundamentally probabilistic concept. We work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

A sequence (A_n) of events is *independent* if for any distinct $n_1, \ldots, n_k$, we have

$$\mathbb{P}(A_{n_1} \cap \ldots \cap A_{n_k}) = \prod_{j=1}^{k} \mathbb{P}(A_{n_j})$$

A sequence $(\mathcal{G}_n)$ of sub-σ-fields of $\mathcal{F}$ is *independent* if for any distinct $n_1, \ldots, n_k$ and events $A_j \in \mathcal{G}_{n_j}$, we have

$$\mathbb{P}(A_1 \cap \ldots \cap A_k) = \prod_{j=1}^{k} \mathbb{P}(A_j)$$

The same definition will apply for π-systems (or indeed any families of events). We only really care about σ-fields, but having the definition for π-systems makes it easier to show that σ-fields are independent, using the following lemma. It is proved here only for two π-systems, but can be extended to all sequences (to get it for infinite sequences, use the fact that a sequence of σ-fields/π-systems is independent iff all its finite subsequences are).

A sequence (X_n) of random variables is *independent* if the σ-fields $\sigma(X_n)$ are independent.

Lemma 7. *If $\mathcal{I}, \mathcal{J}$ are independent π-systems with $\sigma(\mathcal{I}) = \mathcal{G}, \sigma(\mathcal{J}) = \mathcal{H}$, then $\mathcal{G}, \mathcal{H}$ are independent.*

Proof. Fix $A \in \mathcal{I}$.

Define measures μ_1, μ_2 on $\mathcal{H}$ by

$$\mu_1(B) = \mathbb{P}(A \cap B),\ \mu_2(B) = \mathbb{P}(A)\mathbb{P}(B)$$

It is easy to check that these are measures. We are given that $\mu_1(B) = \mu_2(B)$ for all $B \in \mathcal{J}$, and of course $\mu_1(\Omega) = \mu_2(\Omega)$, so by Lemma 3, $\mu_1 = \mu_2$.

Applying this for all $A \in \mathcal{I}$, we get $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for all $A \in \mathcal{I}, B \in \mathcal{H}$.

Now we repeat the other way round: fix $B \in \mathcal{H}$ and define measures μ_3, μ_4 on $\mathcal{G}$ by

$$\mu_3(A) = \mathbb{P}(A \cap B),\ \mu_4(A) = \mathbb{P}(A)\mathbb{P}(B)$$

We have just shown that $\mu_3(A) = \mu_4(A)$ for all $A \in \mathcal{I}$, so using Lemma 3 again, we get $\mu_3 = \mu_4$.

Again, this holds for arbitrary B, so $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for all $A \in \mathcal{G}, B \in \mathcal{H}$. □

3.2 The Borel-Cantelli lemmas

The two Borel-Cantelli lemmas are useful tools to prove results about sequences of events. The proof of the first lemma appears very quick and simple, but it is probably the more important of the two. There are many problems where, if you don't quote this lemma, you will re-write its proof from scratch – and it's harder to come up with in the context of a particular problem than as a general statement. Note that the first lemma does not require independence. Indeed it holds not only in probability spaces, but in general measure spaces.

Lemma 8 (First Borel-Cantelli Lemma)**.** *If (A_n) is a sequence of events with $\sum \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\limsup A_n) = 0$.*

Proof.

$$\mathbb{P}(\limsup A_n) = \mathbb{P}(\bigcap_k \bigcup_{n \geq k} A_n) \leq \mathbb{P}(\bigcup_{n \geq k} A_n) \leq \sum_{n \geq k} \mathbb{P}(A_n) \to 0 \text{ as } k \to \infty$$ □

The second lemma does require independence. This is easy to see because for example if A_n is the same event for all n, then $\sum \mathbb{P}(A_n) = \infty$ but $\mathbb{P}(\limsup A_n) \neq 1$. The proof is more involved and will use the following lemma.

Lemma 9. *If (p_n) is a real sequence with $0 \leq p_n \leq 1$ and $\sum p_n = \infty$, then $\prod(1 - p_n) = 0$.*

Proof.

$$1 - x \leq exp(-x)$$

so

$$\prod_{n=1}^{N}(1 - p_n) \leq \prod_{n=1}^{N} exp(-p_n) = exp(-\sum_{n=1}^{N} p_n) \to 0 \text{ since } \sum p_n \to \infty$$ □

Lemma 10 (Second Borel-Cantelli Lemma)**.** *If (A_n) is a sequence of independent events with $\sum \mathbb{P}(A_n) = \infty$, then $\mathbb{P}(\limsup A_n) = 1$.*

Proof. It is usually easier to prove that a probability is zero rather than one, so we consider $(\limsup A_n)^c = \liminf A_n^c$ instead. Note that if (A_n) is independent, then so is (A_n^c).

Let $p_n = \mathbb{P}(A_n)$. For any k,

$$\mathbb{P}(\bigcap_{n \geq k} A_n^c) \leq \mathbb{P}(\bigcap_{n=k}^{m} A_n^c) = \prod_{n=k}^{m} \mathbb{P}(A_n^c) = \prod_{n=k}^{m}(1 - p_n) \to 0 \text{ as } m \to \infty$$

by the previous lemma. So

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

for all k, and taking a countable union,

$$\mathbb{P}(\liminf A_n^c) = \mathbb{P}(\bigcup_k \bigcap_{n\geq k} A_n^c) = 0$$ □

3.3 Tail σ-fields

Given a sequence (X_n) of random variables, let $\mathcal{T}_n = \sigma(X_n, X_{n+1}, \ldots)$.

Then the *tail σ-field* of the sequence is $\mathcal{T} = \bigcap_n \mathcal{T}_n$.

So $\mathcal{T}$ is the σ-field of events which are determined by the values of X_n, but are not affected by any finite number of these values. For example, the event "$\lim X_n$ exists" is in $\mathcal{T}$, and the random variables $\limsup X_n$, $\liminf X_n$ are $\mathcal{T}$-measurable.

Kolmogorov's 0-1 Law tells us that if the X_n are independent, then a tail event has probability either 0 or 1. However, it may not be easy to tell which. (Sometimes, but not always, the Borel-Cantelli lemmas can help.)

A σ-field $\mathcal{F}$ is *trivial* (with respect to a probability measure $\mathbb{P}$) if $\mathbb{P}(A) = 0$ or 1 for all $A \in \mathcal{F}$. The usual way to prove that a σ-field is trivial is to show that it is independent of itself, as then for any $A \in \mathcal{F}$,

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

Theorem 4 (Kolmogorov's 0-1 Law)**.** *The tail σ-field of a sequence of independent random variables is trivial.*

Proof. Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$.

First we show that $\mathcal{F}_n$ and $\mathcal{T}_{n+1}$ are independent (i.e. the first n variables are independent of all the rest).

Let

$$\mathcal{I}_n = \{\{\omega\colon X_1 \leq c_1, \ldots, X_n \leq c_n\}\colon c_1, \ldots, c_n \in \mathbb{R}\},$$

$$\mathcal{J}_n = \{\{\omega\colon X_n \leq c_n, \ldots, X_{n+r} \leq c_{n+r}\}\colon c_n, \ldots, c_{n+r} \in \mathbb{R}, r \in \mathbb{N}\}$$

Then $\mathcal{I}_n, \mathcal{J}_{n+1}$ are π-systems, generate $\mathcal{F}_n, \mathcal{T}_{n+1}$ respectively, and are certainly independent. So by lemma 7, $\mathcal{F}_n$ and $\mathcal{T}_{n+1}$ are independent.

Since $\mathcal{T} \subset \mathcal{T}_{n+1}$, we get that $\mathcal{T}$ is independent of $\mathcal{F}_n$ for all n. This is certainly what we would expect intuitively: tail events are independent of finite subsequences of (X_n).

Now we show that $\mathcal{T}$ is independent of $\mathcal{T}_1 = \sigma(X_1, X_2, \ldots)$.

Because $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ and they are closed under intersection, the union $\mathcal{I}_\infty = \bigcup_n \mathcal{F}_n$ is a π-system. It must be contained in any σ-field on which all X_n are measurable, so $\mathcal{I}_\infty \subset \mathcal{T}_1$; and all X_n are measurable on any σ-field containing $\mathcal{I}_\infty$, so $\sigma(\mathcal{I}_\infty) \supset \mathcal{T}_1$. So $\sigma(\mathcal{I}_\infty) = \mathcal{T}_1$.

If $A \in \mathcal{I}_\infty$ and $B \in \mathcal{T}$, then $A \in \mathcal{F}_n$ for some n and $\mathcal{F}_n$ is independent of $\mathcal{T}$, so $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. So $\mathcal{I}_\infty$ and $\mathcal{T}$ are independent, so $\mathcal{T}_1$ and $\mathcal{T}$ are independent.

Finally $\mathcal{T} \subset \mathcal{T}_1$, so $\mathcal{T}$ is independent of itself. □

It follows that if Y is a tail-measurable random variable, then $\mathbb{P}(Y \leq c) = 0$ or 1 for all $c \in \mathbb{R}$, so Y is almost surely constant.

4 Integration

> [2] Construction of the integral, expectation.
>
> Fatou's lemma, monotone and dominated convergence, differentiation under the integral sign.

4.1 Definition of integration

Let $(S, \mathcal{S}, \mu)$ be a σ-finite measure space and write $\mathrm{m}\mathcal{S}$ for the set of measurable functions $(S, \mathcal{S}) \to (\mathbb{R}, B(\mathbb{R}))$ and $\mathrm{m}\mathcal{S}^+$ for the set of non-negative measurable functions. For $f \in \mathrm{m}\mathcal{S}^+$, we aim to define the *integral* of f with respect to μ. (Allowing f to take negative values

introduces complications we will consider later.) There are a confusingly large number of notations for this:

$$\mu(f) = \int_S f d\mu = \int_S f(x)\mu(dx)$$

In probability, the equivalent concept is expectation. For a random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$, the *expectation* of X is

$$\mathbb{E}(X) = \int_\Omega X d\mathbb{P}$$

We begin with the indicator functions of measurable sets. For these, if A is measurable, then

$$\int I_A d\mu = \mu(A)$$

(In probabilistic terms, this is the statement $\mathbb{E}(I_A) = \mathbb{P}(A)$.)

The integration operator should be linear, so we can extend this to finite combinations of indicators. A function $f : S \to \mathbb{R}$ is *simple* if

$$f = \sum_{k=1}^n \alpha_k I_{A_k} \text{ for some measurable sets } A_k \text{ and constants } \alpha_k \in \mathbb{R}$$

The set of simple functions is denoted SF, and of non-negative simple functions SF^+. For $f \in SF^+$, we define

$$\int \left(\sum_{k=1}^n \alpha_k I_{A_k} \right) d\mu = \sum_{k=1}^n \alpha_k \mu(A_k)$$

We need to check that this is well-defined, because a single simple function will have many different representations of the form $f = \sum_{k=1}^n \alpha_k I_{A_k}$. This is fairly boring and routine. Note that we can always find such a representation with the sets A_k disjoint, as a simple function can take only finitely many values $c_1, \ldots, c_k$; then let $A_k = f^{-1}(\{c_k\})$. Note also that (by writing the functions on each side as suitable sums of indicators) if f_j are simple functions and $\beta_j \geq 0$ then

$$\int \left(\sum_{j=1}^n \beta_j f_j \right) d\mu = \sum_{j=1}^n \beta_j \int f_j d\mu$$

The simple functions are good enough to approximate all measurable functions from below: given $f \in \mathrm{m}\mathcal{S}^+$, let

$$f_n(x) = 2^{-n} \lfloor 2^n \min\{f(x), n\} \rfloor = 2^{-n} \sum_{j=1}^{n2^n} I_{\{x : f(x) \geq j2^{-n}\}}$$

Then f_n is an increasing sequence which converges to f pointwise, and indeed $f_n \to f$ uniformly on any set of the form $\{x \colon f(x) \leq k\}$.

So for general $f \in \mathrm{m}\mathcal{S}^+$, we define

$$\int f d\mu = \sup\{\textstyle\int g d\mu \colon g \leq f, g \in SF^+\}$$

This may be $+\infty$. $f \in \mathrm{m}\mathcal{S}^+$ is *integrable* if $\int f d\mu < \infty$.

4.2 Monotone Convergence Theorem

The key theorem about integrals is the Monotone Convergence Theorem:

Theorem 5 (Monotone Convergence Theorem)**.** *If $f_n \in \mathrm{m}\mathcal{S}^+$ and $f_n \uparrow f$ (pointwise) then*

$$\int f_n d\mu \uparrow \int f d\mu$$

To prove this theorem, we go through a couple of restricted versions.

Lemma 11. *If $f_n \in SF^+, A \in \mathcal{S}$ and $f_n \uparrow I_A$, then $\int f_n d\mu \uparrow \mu(A)$.*

Proof. Certainly $\int f_n d\mu \leq \mu(A)$ for each n and $\int f_n d\mu$ is increasing. We just need to show that $\int f_n d\mu$ becomes arbitrarily close to $\mu(A)$.

Given $\epsilon > 0$, let $A_n = \{\omega \colon f_n(\omega) \geq 1 - \epsilon\}$.

Since $f_n \uparrow I_A$, $A_n \uparrow A$ and so $\mu(A_n) \uparrow \mu(A)$.

Hence we can choose n s.t. $\mu(A_n) \geq (1-\epsilon)\mu(A)$.

Note that, by the definition of A_n, $f_n \geq (1-\epsilon) I_{A_n}$, so

$$\int f_n d\mu \geq (1-\epsilon)\mu(A_n) \geq (1-\epsilon)^2 \mu(A)$$

Since ϵ is arbitrary, we are done. □

Corollary. *If $f_n, f \in SF^+$ and $f_n \uparrow f$, then $\int f_n d\mu \uparrow \int f d\mu$.*

Proof. Write $f = \sum_k \alpha_k I_{A_k}$ where A_k are disjoint measurable sets. Then $f_n I_{A_k} \uparrow \alpha_k I_{A_k}$ for each k, so we can apply the lemma for each k and add up. □

Lemma 12. *If $f_n \in SF^+, f \in \mathrm{m}\mathcal{S}^+$ and $f_n \uparrow f$ then $\int f_n d\mu \uparrow \int f d\mu$.*

Proof. Again $\int f_n d\mu \leq \int f d\mu$ for all n and $\int f_n d\mu$ is increasing, so we just need to check that $\int f_n d\mu$ becomes arbitrarily close to $\int f d\mu$.

Given $\epsilon > 0$, there exists some $g \in SF^+$ s.t. $g \leq f$ and $\int g d\mu > \int f d\mu - \epsilon$ (by the definition of $\int f d\mu$).

Let $g_n = \min\{g, f_n\}$. Now $g_n \uparrow g$ and g_n, g are simple, so by the above corollary we have $\int g_n d\mu \uparrow \int g d\mu$.

Then for large enough n,

$$\int f_n d\mu \geq \int g_n d\mu > \int g d\mu - \epsilon > \int f d\mu - 2\epsilon$$ □

Now we are ready for the main theorem. The proof uses the following lemma from elementary analysis, left as an exercise:

Lemma 13. *If $(a_{nk})_{n,k\in\mathbb{N}}$ are real numbers increasing with both n and k, then*

$$\lim_{n\to\infty} \lim_{k\to\infty} a_{nk} = \lim_{n\to\infty} a_{nn}$$

Theorem 6 (Monotone Convergence Theorem)**.** *If $f_n \in \mathrm{m}\mathcal{S}^+$ and $f_n \uparrow f$ then $\int f_n d\mu \uparrow \int f d\mu$.*

Proof. First note that $f^{-1}((-\infty, c]) = \bigcap_n f_n^{-1}((-\infty, c])$ so f is measurable and $\int f d\mu$ is defined.

For each n, let $(g_{nk})_{k\in\mathbb{N}}$ be a sequence of simple functions s.t. $g_{nk} \uparrow f_n$ as $k \to \infty$.

For each n, k, let $h_{nk} = \max\{g_{mk} : m \leq n\}$. The h_{nk} are certainly increasing in both n and k, and since $g_{nk} \leq h_{nk} \leq f_n$, $h_{nk} \uparrow f_n$.

So for each ω we get

$$\lim_{n\to\infty} h_{nn}(\omega) = \lim_{n\to\infty} \lim_{k\to\infty} h_{nk}(\omega) = \lim_{n\to\infty} f_n(\omega) = f(\omega)$$

So $h_{nn} \uparrow f$. Since h_{nk} is the max of a finite collection of simple functions, h_{nk} is itself simple. So by the earlier lemma, we get

$$\int f d\mu = \lim_{n\to\infty} \int h_{nn} d\mu = \lim_{n\to\infty} \lim_{k\to\infty} \int h_{nk} d\mu = \lim_{n\to\infty} \int f_n d\mu$$ □

4.3 Integration of functions with negative values, and $\mathcal{L}^1$

Functions which can take negative values require a little more care.

Write the positive and negative parts of f as $f^+ = \max\{f, 0\}, f^- = \max\{-f, 0\}$, so that $f = f^+ - f^-$ and $|f| = f^+ + f^-$.

Then $f \in \mathrm{m}\mathcal{S}$ is *integrable* if $\int |f| d\mu < \infty$, and

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu$$

It is necessary to rule out $\int f^+ d\mu = \int f^- d\mu = \infty$ to avoid an undefined $\infty - \infty$, and to rule out cases in which just one of $\int f^+ d\mu, \int f^- d\mu = \infty$ to ensure that the set of integrable functions is closed under addition and subtraction.

Write $\mathcal{L}^1(S, \mathcal{S}, \mu)$ for the set of integrable functions on a σ-finite measure space $(S, \mathcal{S}, \mu)$.

The monotone convergence theorem extends to functions in $\mathcal{L}^1$, as long as $\lim \int f_n d\mu$ is finite.

4.4 Other convergence theorems

There are a number of important consequences of the monotone convergence theorem. Note that the first applies only to nonnegative functions.

Lemma 14 (Fatou's Lemma)**.** *If $f_n \in \mathrm{m}\mathcal{S}^+$ then*

$$\int \liminf f_n d\mu \leq \liminf \int f_n d\mu$$

Proof. Let $g_n = \inf_{m \geq n} f_m$. (This is why we need that $f_n \geq 0$ – it follows that $g_n \geq 0$ too, so $\int g_n$ is defined. If we allowed functions with negative values then g_n need not be integrable.)

Then $g_n \uparrow \liminf f_n$ so by MCT, $\int g_n d\mu \uparrow \int \liminf f_n d\mu$.

Also $\int g_n d\mu \leq \int f_m d\mu$ for all $m \geq n$ so

$$\int g_n d\mu \leq \inf_{m \geq n} \int f_m d\mu$$

And so

$$\int \liminf f_n d\mu = \lim \int g_n d\mu \leq \liminf \int f_m d\mu$$ □

Lemma 15 (Reverse Fatou's Lemma)**.** *If $f_n \in \mathcal{L}^1$ and there exists an integrable function g s.t. $f_n \leq g$ for all n, then*

$$\int \limsup f_n d\mu \geq \limsup \int f_n d\mu$$

Proof. Apply Fatou's lemma to the non-nonnegative functions $g - f_n$. □

Theorem 7 (Dominated Convergence Theorem)**.** *If $f_n \in \mathcal{L}^1$, there exists an integrable g s.t. $|f_n| \leq g$ for all n, and $f_n \to f$ (pointwise) then*

$$\int |f_n - f| d\mu \to 0$$

Proof. Since $f_n \to f$ we have $\limsup |f_n - f| = 0$.

Also, by the triangle inequality,

$$|f_n - f| \leq 2g$$

so we can apply reverse Fatou to get

$$0 = \int \limsup |f_n - f| d\mu \geq \limsup \int |f_n - f| d\mu \geq 0$$

and so

$$\int |f_n - f| d\mu \to 0$$ □

Corollary (Bounded Convergence Theorem)**.** *If X_n are random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, there is a constant K s.t. $|X_n| \leq K$ for all n, and $X_n \to X$ (pointwise) then $\mathbb{E}(|X_n - X|) \to 0$.*

Proof. Because $\mathbb{P}(\Omega)$ is finite, the constant random variable K is integrable. So we can simply apply the dominated convergence theorem. □

4.5 The standard machine

The standard machine refers to a technique for proving general facts about integrals, by working through the stages of the definition: indicator functions (perhaps starting with indicators of special sets e.g. intervals), simple functions, non-negative functions (using monotone convergence), and finally general integrable functions. As an illustration of the technique, we shall prove that integration is linear and that if X, Y are independent random variables, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Theorem 8. *If $f, g \in \mathcal{L}^1(S, \mathcal{S}, \mu)$ then $\int (\alpha f + \beta g) d\mu = \alpha \int f d\mu + \beta \int g d\mu$.*

Proof. We have observed that this fact holds whenever f, g are simple, directly from the definition of integration.

If $f, g \in \mathrm{m}\mathcal{S}^+$ and $\alpha, \beta \geq 0$, then let $f_n \uparrow f, g_n \uparrow f$ be sequences of simple functions. Then $\alpha f_n + \beta g_n \uparrow \alpha f + \beta g$ so by the Monotone Convergence Theorem,

$$\int \alpha f + \beta g d\mu = \lim \int \alpha f_n + \beta g_n d\mu = \lim \left(\alpha \int f_n d\mu + \beta \int g_n d\mu \right) = \alpha \int f d\mu + \beta \int g d\mu$$

Finally, for $f, g \in \mathcal{L}^1$: assume that $\alpha, \beta \geq 0$, by switching the sign of α and f or β and g if necessary.

Let $A = \{x : \alpha f(x) + \beta g(x) \geq 0\}, B = \{x : \alpha f(x) + \beta g(x) \leq 0\}$.

Now $(\alpha f + \beta g)^+ = (\alpha f^+ I_A + \beta g^+ I_A) - (\alpha f^- I_A + \beta g^- I_A)$ and $(\alpha f + \beta g)^+, (\alpha f^- I_A + \beta g^- I_A) \geq 0$ so using what we have just proved,

$$\int (\alpha f + \beta g)^+ d\mu + \int (\alpha f^- I_A + \beta g^- I_A) d\mu = \int (\alpha f^+ I_A + \beta g^+ I_A) d\mu$$

$$\therefore \int (\alpha f + \beta g)^+ d\mu = \alpha \int f^+ I_A d\mu + \beta \int g^+ I_A d\mu - \alpha \int f^- I_A d\mu - \beta \int g^- I_A d\mu$$

Similarly $-(\alpha f + \beta g)^- = (\alpha f^+ I_B + \beta g^+ I_B) - (\alpha f^- I_B + \beta g^- I_B)$, so

$$-\int (\alpha f + \beta g)^- d\mu = \alpha \int f^+ I_B d\mu + \beta \int g^+ I_B d\mu - \alpha \int f^- I_B d\mu - \beta \int g^- I_B d\mu$$

We also have $f^+ = f^+ I_A + f^+ I_B$ and similarly for f^-, g^+, g^-, so putting these together (and using linearity for nonnegative functions again) gives

$$\int (\alpha f + \beta g)^+ d\mu - \int (\alpha f + \beta g)^- d\mu = \alpha \int f^+ d\mu + \beta \int g^+ d\mu - \alpha \int f^- d\mu - \beta \int g^- d\mu$$

which is what we want. □

Theorem 9. *If X, Y are independent r.v.s and $g, h : \mathbb{R} \to \mathbb{R}$ measurable functions s.t. $g(X), h(Y)$ are integrable, then $\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$.*

Proof. Stating the theorem in this form, rather than just $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ allows us to apply the standard machine to g, h.

If g, h are both indicators of measurable sets, say $g = I_A, h = I_B$, then

$$\begin{aligned}
\mathbb{E}(g(X)h(Y)) &= \mathbb{P}(X \in A \text{ and } Y \in B) \\
&= \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \text{ by definition of independence for } X, Y \\
&= \mathbb{E}(g(X))\mathbb{E}(h(Y))
\end{aligned}$$

By linearity, this extends to g, h simple functions.

If $g, h \geq 0$, then take sequences g_n, h_n of simple functions s.t. $g_n \uparrow g, h_n \uparrow h$. Then $g_n(X)h_n(Y) \uparrow g(X)h(Y)$ so by the Monotone Convergence Theorem,

$$\mathbb{E}(g(X)h(Y)) = \lim_{n\to\infty} \mathbb{E}(g_n(X)h_n(Y)) = \lim_{n\to\infty} \mathbb{E}(g_n(X))\mathbb{E}(h_n(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$$

Finally for general g, h, let $f(\omega) = g(X)h(Y)$.

Then $f^+ = g^+(X)h^+(Y) + g^-(X)h^-(Y)$ and $f^- = g^+(X)h^-(Y) + g^-(X)h^+(Y)$ so

$$\begin{aligned}
\mathbb{E}(g(X)h(Y)) = &\ \mathbb{E}(f^+) - \mathbb{E}(f^-) \\
= &\ \mathbb{E}(g^+(X)h^+(Y)) + \mathbb{E}(g^-(X)h^-(Y)) \\
= &\ \mathbb{E}(g^+(X))\mathbb{E}(h^+(Y)) + \mathbb{E}(g^-(X))\mathbb{E}(h^-(Y)) \\
&\ - \mathbb{E}(g^+(X))\mathbb{E}(h^-(Y)) - \mathbb{E}(g^-(X))\mathbb{E}(h^+(Y)) \\
= &\ (\mathbb{E}(g^+(X)) - \mathbb{E}(g^-(X)))(\mathbb{E}(h^+(X)) - \mathbb{E}(h^-(X))) \\
= &\ \mathbb{E}(g(X))\mathbb{E}(h(Y))
\end{aligned}$$

□

4.6 Differentiation under the integral sign

This is fairly unimportant but is explicitly required by the schedules.

Suppose we have a function $f(s,x)$ of two variables, and let $F(x) = \int f(s,x)\mu(ds)$. It would be nice if we could obtain $F'(x)$ by interchanging the order of differentiation and integration, and getting $F'(x) = \int \frac{\partial}{\partial x} f(s,x)\mu(ds)$.

Here we give conditions for this to be valid:

Theorem 10. *Let $f : (S, \mathcal{S}, \mu) \times \mathbb{R} \to \mathbb{R}$ be a function s.t.*

(i) For each $x \in \mathbb{R}$, $s \mapsto f(s,x)$ is a $\mathcal{S}$-measurable function (so $F(x)$ has a chance of existing).

(ii) For each $s \in S$, $x \mapsto f(s,x)$ is differentiable.

(iii) $s \mapsto f(s,x) \in L^1(S, \mathcal{S}, \mu)$ (so $F(0)$ exists and is finite).

(iv) There is a function $g \in L^1(S, \mathcal{S}, \mu)$ s.t. $\left|\frac{f(s,x) - f(s,y)}{x-y}\right| \leq g(s) \forall x \neq y$ (so approximations to $\frac{\partial}{\partial x} f(s,x)$ are dominated by g).

Then $F(x) = \int_S f(s,x)\mu(ds)$ is differentiable, and $F'(x) = \int \frac{\partial}{\partial x} f(s,x)\mu(ds)$.

Proof. Condition (iii) together with (iv) with $y = 0$ ensures that $F(x)$ exists and is finite for all $x \in \mathbb{R}$.

Now for fixed x,

$$\begin{aligned}\frac{F(x+h) - F(x)}{h} &= \int_S \frac{f(s,x+h) - f(s)}{h}\mu(ds) \\ &\to \int_S \frac{\partial}{\partial x} f(s,x)\mu(ds) \text{ as } h \to 0\end{aligned}$$

by dominated convergence. □

5 Modes of convergence

[1] Convergence in measure and convergence almost everywhere.

5.1 Convergence almost everywhere

A sequence of functions $f_n : S \to \mathbb{R}$ converges *μ-almost everywhere* to f if

$$\mu(\{\omega \colon f_n(\omega) \not\to f(\omega)\}) = 0$$

In other words, $f_n \to f$ pointwise except on a set of measure zero. In probability, this is called *almost sure convergence*. So far as we are concerned in measure theory, this is just as good as pointwise convergence; in particular the Monotone and Dominated Convergence Theorems continue to hold if we replace pointwise convergence with almost everywhere convergence (just consider $f_n I_A \to f I_A$ where $A = \{\omega \colon f_n(\omega) \to f(\omega)\}$).

5.2 Convergence in measure

A sequence of functions $f_n : S \to \mathbb{R}$ converges *in μ-measure* to f if, for any $\epsilon, \delta > 0$, there exists N s.t. for all $n \geq N$,

$$\mu(\{\omega \colon |f_n(\omega) - f(\omega)| > \epsilon\}) < \delta$$

In probability, this is called *convergence in probability*. This is a weak definition of convergence and does not imply that $f_n(\omega) \to f(\omega)$ at any point, as the sets $\{\omega : |f_n(\omega) - f(\omega)| > \epsilon\}$ must become small in measure but they can jump back and forward across the entire space.

For example, let

$$\begin{array}{llll} f_1 = I_{(0,1]}, & & & \\ f_2 = 2I_{(0,\frac{1}{2}]}, & & f_3 = 2I_{(\frac{1}{2},1]} & \\ f_4 = 4I_{(0,\frac{1}{4}]}, & f_5 = 4I_{(\frac{1}{4},\frac{1}{2}]}, & f_6 = 4I_{(\frac{1}{2},\frac{3}{4}]}, & f_7 = 4I_{(\frac{3}{4},1]} \end{array}$$

and so on.

For small ϵ we have $\mu(\{\omega : |f_n(\omega)| > \epsilon\}) = 2^{-k}$ when $2^k \leq n < 2^{k+1}$, so $f_n \to 0$ in measure, but $f_n(\omega) \not\to 0$ for any $\omega \in (0,1]$. Also $\int f_n d\mu = 1$ for all n, so $\int f_n d\mu \not\to 0$.

Theorem 11. *If $(S, \mathcal{S}, \mu)$ is a finite measure space and $f_n \to f$ a.e., then $f_n \to f$ in measure.*

Proof. Fix $\epsilon > 0$. Let $A_n = \{\omega\colon |f_m(\omega) - f(\omega)| \leq \epsilon \text{ for all } m \geq n\}$.
Then $A_n \uparrow A$ where $A = \{\omega\colon f_n(\omega) \to f(\omega)\}$, and as $f_n \to f$ a.e., $\mu(A) = \mu(A_n)$.
So $\mu(A_n) \uparrow \mu(S)$.
Since $\mu(S)$ is finite, $\mu(A_n^c) = \mu(S) - \mu(A_n) \downarrow 0$.
Now $\{\omega\colon |f_n(\omega) - f(\omega)| > \epsilon\} \subset A_n^c$, so $f_n \to f$ in measure. □

Theorem 12. *If $f_n \to f$ in measure, then there is a subsequence (f_{n_k}) s.t. $f_{n_k} \to f$ almost everywhere.*

Proof. Say that (f_n) *converges fast in measure* to f if for all $\epsilon > 0$,

$$\sum_n \mu(\{\omega\colon |f_n(\omega) - f(\omega)| > \epsilon\}) < \infty$$

This is clearly a strengthening of the definition of convergence in measure, and it rules out things like the above counter-example because the sets on which it is far from zero get small too slowly. In fact it is a sufficient strengthening that fast convergence in measure implies almost-everywhere convergence (proved below).

It addition it is not too hard to show that if a sequence converges in measure, then it has a subsequence which converges fast in measure: just choose n_k such that, for all $n \geq n_k$,.

$$\mu(\{\omega\colon |f_n(\omega) - f(\omega)| > 2^{-k}\}) < 2^{-k}$$

For any $\epsilon > 0$, choose K s.t. $2^{-K} < \epsilon$, and then

$$\sum_{k=K}^{\infty} \mu(\{\omega\colon |f_{n_k}(\omega) - f(\omega)| > \epsilon\}) < \sum_{k=K}^{\infty} 2^{-k} < \infty$$

To complete the proof, suppose that $f_n \to f$ fast in measure. We show that $f_n \to f$ a.e.
For any $\epsilon > 0$, Borel-Cantelli I gives us

$$\mu(\limsup_{n\to\infty}\{\omega\colon |f_n(\omega) - f(\omega)| > \epsilon\}) = 0$$

Now

$$f_n(\omega) \not\to f(\omega) \Leftrightarrow \exists \epsilon > 0 \text{ s.t. for infinitely many } k, |f_n(\omega) - f(\omega)| > \epsilon$$

$$\therefore \{\omega\colon f_n(\omega) \not\to f(\omega)\} = \bigcup_m \limsup_{n\to\infty}\{\omega\colon |f_n(\omega) - f(\omega)| > 1/m\}$$

$$\therefore \mu(\{\omega\colon f_n(\omega) \not\to f(\omega)\}) \leq \sum_m \mu(\limsup_{n\to\infty}\{\omega\colon |f_n(\omega) - f(\omega)| > 1/m\}) = 0$$ □

5.3 The space L^1

Recall that $\mathcal{L}^1(S, \mathcal{S}, \mu) = \{f \in \mathrm{m}\mathcal{S}\colon \int |f| d\mu < \infty\}$. This is a vector space and $f \mapsto \int f d\mu$ is a linear functional on $\mathcal{L}^1$.

Defining $\|f\| = \int |f| d\mu$ gives something which almost satisfies the definition of a normed space, except that $\|f\| = 0 \not\Rightarrow f = 0$, only that $f = 0$ almost everywhere. Such a space is called a *seminormed space* and we can get a normed space by taking the subspace $W = \{f \in \mathcal{L}^1\colon f = 0 \text{ a.e.}\}$ and letting $L^1(S, \mathcal{S}, \mu)$ be the quotient space $\mathcal{L}^1/W$. You should check that this does give a well-defined normed space.

Now $\|f\| = \int |f| d\mu$ is a norm on L^1, and $f \to \int f d\mu$ is still a well-defined linear functional on L^1.

I don't promise to always be careful about distinguishing between $\mathcal{L}^1$ and L^1.

5.4 Convergence in L^1

We can define convergence in L^1 in the usual way for normed spaces: f_n *converges in* L^1 to f if $\|f_n - f\| \to 0$, or in other words $\int |f_n - f| d\mu \to 0$. In probability, this is called *convergence in mean.*

Note that this is not the same thing as simple convergence of the means or integrals: $\int f_n d\mu \to \int f d\mu$ tells us very little, and the f_n can vary wildly so long as positive and negative

parts of $f_n - f$ cancel out. Considering $|f_n - f|$ prevents this. However, $f_n \to f$ in L^1 does imply that $\int f_n d\mu \to \int f d\mu$ (an analyst would describe this by saying that $\int \cdot\, d\mu$ is a continuous function $L^1 \to \mathbb{R}$).

How does this relate to our earlier types of convergence? Convergence almost everywhere does not imply L^1 convergence – this is just the fact from Analysis II that pointwise convergence does not imply convergence of integrals. The example in Section 5.2 shows that convergence in measure does not imply L^1 convergence, and a slight modification of this example (taking each f_n to be the indicator of an interval, instead of 2^k times an indicator) shows that L^1 convergence does not imply almost everywhere convergence. However we do have:

Theorem 13. *If $f_n \to f$ in L^1, then $f_n \to f$ in measure.*

Proof. Given $\epsilon, \delta > 0$, we can use L^1 convergence to find N s.t. for all $n > N$,

$$\epsilon\delta > \int |f_n - f| d\mu \geq \epsilon\mu(\{\omega \colon |f_n(\omega) - f(\omega)| > \epsilon\})$$

Then

$$n > N \Rightarrow \mu(\{\omega \colon |f_n(\omega) - f(\omega)| > \epsilon\}) < \delta$$ □

Lemma 16 (Scheffé's Lemma)**.** *If $f_n, f \in L^1$, $f_n \geq 0$, $f_n \to f$ a.e., and $\int f_n d\mu \to \int f d\mu$, then $f_n \to f$ in L^1.*

Proof. Consider the positive and negative parts of $f_n - f$ separately.

We have $0 \leq (f_n - f)^- = (f - f_n)^+ \leq f$ a.e. (using $f_n \geq 0$) and $(f_n - f)^- \to 0$ a.e. so by DCT, $\int (f_n - f)^- d\mu \to 0$.

We are given that $\int (f_n - f)^+ d\mu - \int (f_n - f)^- d\mu = \int (f - f_n) d\mu \to 0$, so $\int (f_n - f)^+ d\mu \to 0$.

Now $\int |f - f_n| d\mu = \int (f_n - f)^+ d\mu + \int (f_n - f)^- d\mu \to 0$. □

6 Fubini's theorem

[1] Discussion of product measure and statement of Fubini's theorem.

6.1 Product σ-fields

Let $(S_1, \mathcal{S}_1, \mu_1), (S_2, \mathcal{S}_2, \mu_2)$ be two finite measure spaces.

We define a product σ-field $\mathcal{S}$ on $S = S_1 \times S_2$ in essentially the same way as we would define the product of topological spaces, with continuous functions replaced by measurable: it is the smallest σ-field on which the projections π_1, π_2 are measurable; in other words it is $\sigma(A \times S_2 \colon A \in \mathcal{S}_1, S_1 \times B \colon B \in \mathcal{S}_2)$.

This will certainly contain $\mathcal{I} = \{B_1 \times B_2 \colon B_i \in \mathcal{S}_i\}$, so $\mathcal{S} = \sigma(\mathcal{I})$ (which again is like the definition of product topology as the topology generated by products of open sets).

6.2 Product measures

This is the hard bit in this section.

Let μ_1, μ_2 be finite measures on $(S_1, \mathcal{S}_1), (S_2, \mathcal{S}_2)$.

To define the product measure μ on $(S, \mathcal{S})$, we start with $\mu(B_1 \times B_2) = \mu_1(B_1)\mu_2(B_2)$ on $\mathcal{I}$ (think of our two starting spaces as $[0, 1]$ so S is the unit square, and μ is area). Since $\mathcal{I}$ is a π-system generating $\mathcal{S}$, there is at most one finite measure extending this.

If you were an applied mathematician, you might try to write the product measure of a general set by integrating the measures of "slices" through it:

$$\mu(A) = \int_{S_1} \mu_2(\{s_2 \in S_2 \colon (s_1, s_2) \in A\}) \mu_1(ds_1)$$

or the same thing with 1 and 2 swapped.

We shall do the same, but we have to do some work with d-systems and σ-fields to show that it is valid.

Let $J_{1A}(s_1) = \mu_2(\{s_2 \in S_2 \colon (s_1, s_2) \in A\})$. We need to check that the slices involved are measurable. This is true because all s_1-slices through sets in $\mathcal{I}$ are certainly $\mathcal{S}_2$-measurable, and the family of subsets of S whose s_1-slices are measurable is closed under complements and countable unions, so is a σ-field, so contains $\sigma(\mathcal{I})$.

Lemma 17. *J_{1A} is $\mathcal{S}_1$-measurable for all $A \in \mathcal{S}$.*

Proof. Let $\mathcal{D} = \{A \in \mathcal{S} \colon J_{1A} \text{ is } \mathcal{S}_1\text{-measurable}\}$. We shall show that $\mathcal{D}$ is a d-system.

Certainly $S \in \mathcal{D}$.

If $A, B \in \mathcal{D}$ with $A \supset B$ then $J_{1(A \setminus B)} = J_{1A} - J_{1B}$ is $\mathcal{S}_1$-measurable, so $A \setminus B \in \mathcal{D}$.

If $A_n \uparrow A$ and $A_n \in \mathcal{D}$ then $J_{1A_n} \uparrow J_{1A}$ so J_{1A} is $\mathcal{S}_1$-measurable and $A \in \mathcal{D}$.

So $\mathcal{D}$ is a d-system; and it certainly contains the π-system $\mathcal{I}$, so $\mathcal{D} \supset d(\mathcal{I}) = \sigma(\mathcal{I}) = \mathcal{S}$. □

Hence $\mu(A) = \int_{S_1} J_{1A} d\mu_1$ is well-defined.

It is finitely additive because μ_2 is and because integration is linear, and it is countably additive because if $A = \bigcup_n A_n$ (A_n pairwise disjoint) then

$$\sum_{j=1}^{n} \mu(A_j) = \int_{S_1} \sum_{j=1}^{n} J_{1A_j} d\mu_1 \uparrow \int_{S_1} \sum_{j=1}^{\infty} J_{1A_j} d\mu_1 = \mu(A)$$

by the Monotone Convergence Theorem.

So μ is a measure on $(S, \mathcal{S})$, it takes the right values on $\mathcal{I}$, and it is finite because $\mu(S) = \mu_1(S_1)\mu_2(S_2) < \infty$.

We could alternatively have defined $J_{2A}(s_2) = \mu_1(\{s_1 \in S_1 \colon (s_1, s_2) \in A\})$, $\mu(A) = \int_{S_2} J_{2A} d\mu_2$, and because there is a unique measure taking the right values on $\mathcal{I}$, this gives the same thing.

We have set this all up for finite measures, which is necessary in order to use the uniqueness of extension lemma, but it can now be extended to σ-finite measures μ_1, μ_2 just by working on an increasing sequence of finite subspaces whose union is all of S_1 or S_2.

6.3 Integration on product spaces

The results of Fubini's and Tonelli's theorems tell us that a function f on the product space can be integrated by integrating it in one variable first, then the other, provided that $f \geq 0$ or $|f|$ is μ-integrable. An important consequence is that it doesn't matter which order the integrations are performed in.

The proofs of these are rather easy, now that we have shown how to define the product measure by one-dimensional integrals.

Theorem 14 (Tonelli's Theorem)**.** *If $f \colon S \to \mathbb{R}$ is $\mathcal{S}$-measurable and $f \geq 0$ then*

$$\int_S f d\mu = \int_{S_1} \left(\int_{S_2} f(s_1, s_2) \mu_2(ds_2) \right) \mu_1(ds_1) = \int_{S_2} \left(\int_{S_1} f(s_1, s_2) \mu_1(ds_1) \right) \mu_2(ds_2)$$

Proof. We use the standard machine.

First suppose that f is the indicator function of a $\mathcal{S}$-measurable set A. Then this is just the statement $\mu(A) = \int_{S_1} J_{1A}(s_1) \mu_1(ds_1) = \int_{S_2} J_{2A}(s_2) \mu_2(ds_2)$.

By linearity, the theorem holds for non-negative simple functions on $(S, \mathcal{S})$.

For general $f \in \mathrm{m}\mathcal{S}^+$, take a sequence f_n of nonnegative simple functions s.t. $f_n \uparrow f$.

Then by monotone convergence, $\int_{S_2} f_n(s_1, s_2) \mu_2(ds_2) \uparrow \int_{S_2} f(s_1, s_2) \mu_2(ds_2)$ and so

$$\int_{S_1} \left(\int_{S_2} f_n(s_1, s_2) \mu_2(ds_2) \right) \mu_1(ds_1) \uparrow \int_{S_1} \left(\int_{S_2} f(s_1, s_2) \mu_2(ds_2) \right) \mu_1(ds_1)$$

(And similarly with 1 and 2 swapped.)

Also by monotone convergence, $\int_S f_n d\mu \uparrow \int_S f d\mu$. □

Theorem 15 (Fubini's Theorem)**.** *If $f \in \mathcal{L}^1(S, \mathcal{S}, \mu)$ (i.e. $\int_S |f| d\mu < \infty$) then*

$$\int_S f d\mu = \int_{S_1} \left(\int_{S_2} f(s_1, s_2) \mu_2(ds_2) \right) \mu_1(ds_1) = \int_{S_2} \left(\int_{S_1} f(s_1, s_2) \mu_1(ds_1) \right) \mu_2(ds_2)$$

Proof. Just apply Tonelli's theorem to f^+ and f^-. □

Note that the two theorems can be used together: to show that a general f satisfies the condition $\int_S |f| d\mu < \infty$ in Fubini's theorem, check that $\int_{S_1} \left(\int_{S_2} |f| \mu_2(ds_2) \right) \mu_1(ds_1) < \infty$ and use Tonelli on $|f|$. In Cambridge, it is common to combine the two theorems together and call them both Fubini's theorem.

6.4 Borel-Cantelli from Tonelli

Fubini's and Tonelli's theorems are more powerful than they may at first appear (as indeed is true of a lot of measure theory), because you can choose to apply it to a wide range of measure spaces. As an illustration, we will derive the first Borel-Cantelli lemma from Tonelli.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathbb{N}, \wp(\mathbb{N}), \#)$ be the natural numbers with counting measure (i.e. $\mu(A)$ is simply the number of elements of A; then $\int_{\mathbb{N}} f d\# = \sum_{n=1}^{\infty} f(n)$).

Given a sequence (A_n) of events in $\mathcal{F}$, we define a function $f : \Omega \times \mathbb{N} \to \mathbb{R}$ by $f(\omega, n) = I_{A_n}(\omega)$.

This is nonnegative, so Tonelli's theorem gives us

$$\int_{\Omega\times\mathbb{N}} f d(\mathbb{P}, \mu) = \int_{\mathbb{N}} \int_{\Omega} f d\mathbb{P} d\# = \int_{\mathbb{N}} \mathbb{P}(A_n) \#(dn) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$
$$= \int_{\Omega} \int_{\mathbb{N}} f d\# d\mathbb{P} = \int_{\Omega} \sum_{n=1}^{\infty} I_{A_n} d\mathbb{P} = \mathbb{E}(N)$$

where $N(\omega)$ is the number of values $n \in \mathbb{N}$ s.t. $\omega \in A_n$.

The condition of BCI is that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$. Then we get that $\mathbb{E}(N) < \infty$, so $\mathbb{P}(N = \infty) = 0$. But the event $\{N = \infty\}$ is precisely $\limsup A_n$.

7 Uniform integrability

> [1] Uniform integrability.

7.1 Definition of uniform integrability

I shall only deal with probability spaces here because that is the important case; at least some of this extends to σ-finite measure spaces, but it becomes a bit more complicated.

A family $\{X_\lambda\}$ of random variables is *uniformly integrable* if, for all $\epsilon > 0$, there is some K s.t. for all λ,

$$\int_{|X_\lambda|>K} |X_\lambda| d\mathbb{P} < \epsilon$$

Intuitively this says that the tails of the distributions of all the random variables become small together. (Uniform integrability is related to compactness in some way which I haven't figured out yet, but note that the next theorem looks a bit like the Ascoli-Arzelà theorem). Uniform integrability is defined for any set of random variables, but we usually apply it to sequences.

The following theorem gives an equivalent characterisation which is often more useful.

Theorem 16. *A family $\{X_\lambda\}$ of random variables is uniformly integrable iff both the following conditions*
hold:

(i) The set $\{X_\lambda\}$ is bounded in L^1 (i.e. $\exists M$ s.t. $\forall \lambda\, \mathbb{E}(|X_\lambda|) < M$).

(ii) For every $\epsilon > 0$, there is some $\delta > 0$ s.t. for all λ and all events A with $\mathbb{P}(A) < \delta$, we have

$$\int_A |X_\lambda| d\mathbb{P} < \epsilon$$

Proof. First suppose that $\{X_\lambda\}$ is UI.

For (i), choose K s.t. for all λ,

$$\int_{|X_\lambda|>K} |X_\lambda| d\mathbb{P} < 1$$

Then $\mathbb{E}(|X_\lambda|) = \int_{|X_\lambda|\le K} |X_\lambda| d\mathbb{P} + \int_{|X_\lambda|>K} |X_\lambda| d\mathbb{P} < K + 1$.

For (ii), given $\epsilon > 0$, choose K as in the definition of uniform integrability, and δ s.t. $K\delta < \epsilon$. Then for $\mathbb{P}(A) < \delta$,

$$\int_A |X_\lambda| d\mathbb{P} = \int_{A\cap\{|X_\lambda|\le K\}} |X_\lambda| d\mathbb{P} + \int_{A\cap\{|X_\lambda|>K\}} |X_\lambda| d\mathbb{P} < K\mathbb{P}(A) + \epsilon < 2\epsilon$$

In the other direction, suppose that (i) and (ii) hold.

Given $\epsilon > 0$, choose suitable δ as in (ii), and M as in (i). Set $K = M/\delta$.

Now $M > \mathbb{E}(|X_\lambda|) \geq K\mathbb{P}(|X_\lambda| > K)$ so $\mathbb{P}(|X_\lambda| > K) < \delta$ for all λ. So putting $A = \{|X_\lambda| > K\}$ in (ii) gives exactly what is required. □

7.2 Sufficient conditions for uniform integrability

A single integrable random variable X is UI since the random variables $Z_n = |X|I_{\{|X|>n\}}$ are dominated by X and converge a.s. to 0, so by DCT,

$$\int_{|X|>n} |X| d\mathbb{P} = \mathbb{E}(Z_n) \to 0$$

Any finite set $\{X_1, \ldots, X_n\}$ of integrable r.v.s is UI since for each $\epsilon > 0$, we can find $K_1, \ldots, K_n$ for which $\int_{|X_n|>K_n} |X_n| d\mathbb{P} < \epsilon$, then take $K = \max\{K_1, \ldots, K_n\}$.

If there is some $X \in L^1$ s.t. $|X_\lambda| \leq |X|$ for all λ, then $\{X_\lambda\}$ is UI, since given $\epsilon > 0$, we may choose K s.t. $\int_{|X|>K} |X| d\mathbb{P} < \epsilon$. Then for each λ, $\{\omega \colon |X_\lambda| > K\} \subset \{\omega \colon |X| > K\}$, so

$$\int_{|X_\lambda|>K} |X_\lambda| d\mathbb{P} \leq \int_{|X|>K} |X_\lambda| d\mathbb{P} \leq \int_{|X|>K} |X| d\mathbb{P} < \epsilon$$

If $\{X_\lambda\}$ and $\{Y_{\lambda'}\}$ are UI, then $\{X_\lambda + Y_{\lambda'}\}$ is UI, by the alternative characterisation.

7.3 L^1 convergence and uniform integrability

Theorem 17. *A sequence (X_n) of integrable random variables converges to X in L^1 iff $X_n \to X$ in probability and $\{X_n\}$ is UI.*

Proof. First suppose that $X_n \to X$ in L^1.

Then $X \in L^1$ so $\{X_n - X\}$ UI implies that $\{X_n\}$ is UI, so we may assume wlog that $X = 0$.

We already know that $X_n \to 0$ in probability.

Given $\epsilon > 0$, we can choose N s.t. $n > N \Rightarrow \mathbb{E}(|X_n|) < \epsilon$, so certainly $\int_{|X_n|>K} |X_n| d\mathbb{P} < \epsilon$ for any K.

So we just have to choose K s.t. $\int_{|X_n|>K} |X_n| d\mathbb{P} < \epsilon$ for $1 \leq n \leq N$, which we can do since the finite set $\{X_1, \ldots, X_N\}$ is UI.

Now suppose that $X_n \to X$ in probability and $\{X_n\}$ is UI.

First we show that $X \in L^1$ (which seems to take several lines, but they are quite easy).

Take a subsequence (X_{n_k}) s.t. $X_{n_k} \to X$ almost surely.

Let $Y_{n_k} = \min\{X_{n_k}^+, X^+\}$ so $Y_{n_k} \uparrow X^+$ a.s. and by MCT $\mathbb{E}(Y_{n_k}) \uparrow \mathbb{E}(X^+)$.

$\mathbb{E}(Y_{n_k}) \leq \mathbb{E}(|X_{n_k}|) \leq M$ for some M since $\{X_{n_k}\}$ is UI, so $\mathbb{E}(X^+) \leq M$.

In particular, $\mathbb{E}(X^+) < \infty$ and likewise $\mathbb{E}(X^-) < \infty$ so $X \in L^1$.

Hence $\{X_n\}$ UI implies that $\{X_n - X\}$ is UI, so we may assume wlog that $X = 0$.

Given $\epsilon > 0$, choose δ s.t. $\mathbb{P}(A) < \delta \Rightarrow \int_A |X_n| d\mathbb{P} < \epsilon$.

Then, by convergence in probability, choose N s.t. $n > N \Rightarrow \mathbb{P}(|X_n| > \epsilon) < \delta$.

Now for $n > N$, we have

$$\mathbb{E}(|X_n|) = \int_{|X_n|\leq\epsilon} |X_n| d\mathbb{P} + \int_{|X_n|>\epsilon} |X_n| d\mathbb{P} < \epsilon\mathbb{P}(|X_n| \leq \epsilon) + \epsilon < 2\epsilon$$ □

Because a sequence dominated by an integrable r.v. is UI, this theorem strengthens the Dominated Convergence Theorem to require only convergence in probability rather than almost sure convergence.

8 Inequalities and Banach spaces

[3] Chebyshev's inequality, tail estimates. Jensen's inequality. Completeness of L^p for $1 \leq p \leq \infty$. The Hölder and Minkowski inequalities.

8.1 Chebyshev's inequality

Theorem 18 (Chebyshev's Inequality)**.** *If f is a measurable function $S \to \mathbb{R}$ then*

$$\mu(\{\omega\colon |f(\omega)| \geq t\}) \leq \frac{1}{t^2}\int_S f^2 d\mu$$

Proof.

$$\int_S f^2 d\mu \geq \int_{|f|\geq t} f^2 d\mu \geq t^2 \mu(\{\omega\colon |f(\omega)| \geq t\}) \qquad \square$$

In the probability case where X is a random variable, this becomes $\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq \frac{1}{t^2}\operatorname{Var}(X)$, where $\operatorname{Var}(X)$ has the usual definition of $\mathbb{E}((X - \mathbb{E}X)^2)$.

8.2 Jensen's inequality

Jensen's inequality is an important result concerning the averaging of convex functions. A *convex function* is a function $\varphi : \mathbb{R} \to \mathbb{R}$ s.t. whenever $0 \leq a, b \leq 1$ and $a + b = 1$, $\varphi(ax + by) \leq a\varphi(x) + b\varphi(y)$. It is often easiest to test this using the fact that twice-differentiable function is convex iff $\varphi''(x) \geq 0$ for all x, or just by drawing a graph.

Note that this result holds only in probability spaces; it cannot even be generalised (like most of the results I have given for probability spaces) to finite measure spaces.

Theorem 19 (Jensen's Inequality)**.** *If X is an integrable random variable and φ a measurable convex function $\mathbb{R} \to \mathbb{R}$, then*

$$\varphi(\mathbb{E}(X)) \leq \mathbb{E}(\varphi(X))$$

Proof. Properly we need a condition that $\mathbb{E}(\varphi(X))$ exists.

The key step in the proof is that through every point $(x, \varphi(x))$ in the graph of φ, there is a line which is always below or on the graph. If φ is differentiable at x, this is just the tangent. If not, then we have to work a bit harder.

Given $z \in \mathbb{R}$, define $\psi_z(x) = \dfrac{\varphi(x) - \varphi(z)}{x - z}$ for $x \neq z$.

From the definition of convexity, ψ_z is increasing, and in particular, if $x_1 < z < x_2$, then $\psi_z(x_1) \leq \psi_z(x_2)$.

Set $\alpha_z = \sup_{x<z}\psi_z(x)$ and β_z s.t. the line $y = \alpha_z x + \beta_z$ goes through $(z, \varphi(z))$.

Then $\psi_z(x) \leq \alpha_z$ for all $x < z$ and $\psi_z(x) \geq \alpha_z$ for all $x > z$, so $\alpha_z x + \beta_z \leq \varphi(x)$ for all x.

Now apply this result with $z = \mathbb{E}(X)$.

By the choice of β_z, we have $\varphi(\mathbb{E}(X)) = \varphi(z) = \alpha_z z + \beta_z = \alpha_z \mathbb{E}(X) + \beta_z$.

But also $\alpha_z x + \beta_z \leq \varphi(x)$ for all x, so taking expectations,

$$\varphi(\mathbb{E}(X)) = \alpha_z \mathbb{E}(X) + \beta_z \leq \mathbb{E}(\varphi(X)) \qquad \square$$

8.3 Hölder's inequality

There are many proofs of this important inequality. Analysts like to prove it again and again, once for finite sequences, once for infinite sequences (l^p spaces), once for integrable functions (L^p spaces). The measure theoretic version gives us all of these at once, by taking a suitable underlying measure space (a finite set with counting measure, $\mathbb{N}$ with counting measure, Lebesgue measure). This proof uses Jensen's inequality, but on a specially defined probability measure, so the theorem still holds for general σ-finite measure spaces.

Theorem 20 (Hölder's Inequality)**.** *If $f, g \in \mathrm{m}\mathcal{S}$, $1 < p, q$ and $p^{-1} + q^{-1} = 1$, then*

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

Proof. All the functions we are integrating are all nonnegative, so there is no need to worry about whether the integrals exist (provided we allow the value $+\infty$).

The inequality is obvious if $\int |g|^q d\mu$ is 0 or ∞.

Otherwise, we define a probability measure μ_g on $(S, \mathcal{S})$ by

$$\mu_g(A) = \frac{\int I_A |g|^q d\mu}{\int |g|^q d\mu}$$

(this is countably additive by MCT) so that, applying the standard machine,

$$\int f d\mu_g = \frac{\int f|g|^q d\mu}{\int |g|^q d\mu}$$

The function $\varphi(x) = x^p$ is convex for $x \geq 0$, so applying Jensen to φ and the function $|fg^{1-q}|$ we get

$$\left(\int |fg^{1-q}| d\mu_g\right)^p \leq \int |fg^{1-q}|^p d\mu_g$$

The rest is just simplifying this, using the fact that $p + q - pq = 0$.

$$\left(\frac{\int |f||g|^{1-q}|g|^q d\mu}{\int |g|^q d\mu}\right)^p \leq \frac{\int |f|^p |g|^{(1-q)p}|g|^q d\mu}{\int |g|^q d\mu}$$

$$\therefore \left(\int |f||g| d\mu\right)^p \leq \left(\int |f|^p |g|^{p+q-pq} d\mu\right)\left(\int |g|^q d\mu\right)^{p-1}$$

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

8.4 The space L^p

For $1 < p < \infty$ and $f \in \mathrm{m}\mathcal{S}$, define the L^p *norm* by $\|f\|_p = \left(\int |f|^p d\mu\right)^{1/p}$.

Then let the space of *p-integrable functions* be $\mathcal{L}^p(S, \mathcal{S}, \mu) = \{f \in \mathrm{m}\mathcal{S} \colon \int |f|^p d\mu < \infty\}$.

We need to check that this is a vector space, and in particular that it is closed under addition. This holds because $|f + g|^p \leq (2\max(|f|, |g|))^p \leq 2^p(|f|^p + |g|^p)$.

Now we form a normed space $L^p(S, \mathcal{S}, \mu)$ in the same way as we formed L^1 from $\mathcal{L}^1$: let $W = \{f \in \mathcal{L}^p \colon f = 0 \text{ a.e.}\}$ and $L^p = \mathcal{L}^p/W$.

To show that $\|\cdot\|_p$ is a norm on L^p, we just need to check the triangle inequality. This is given by:

Theorem 21 (Minkowski's Inequality)**.** *If $f, g \in \mathrm{m}\mathcal{S}, 1 < p < \infty$, then $\|f+g\|_p \leq \|f\|_p + \|g\|_p$.*

Proof. If $\int |f + g|^p d\mu = 0$ the inequality is obvious; if it is ∞, then so is the RHS, because $\mathcal{L}^p$ is closed under addition.

Otherwise, set q s.t. $p^{-1} + q^{-1} = 1$. Then by Hölder's inequality,

$$\int |f||f + g|^{p-1} d\mu \leq \left(\int |f|^p d\mu\right)^{1/p}\left(\int |f + g|^{(p-1)q} d\mu\right)^{1/q}$$

so

$$\begin{aligned}\int |f + g|^p d\mu &\leq \int |f||f + g|^{p-1} d\mu + \int |g||f + g|^{p-1} d\mu \\ &\leq \left(\left(\int |f|^p d\mu\right)^{1/p} + \left(\int |g|^p d\mu\right)^{1/p}\right)\left(\int |f + g|^p d\mu\right)^{1/q}\end{aligned}$$

and we simply divide through by $\left(\int |f + g|^p d\mu\right)^{1/q}$. $\square$

8.5 Completeness of L^p

The spaces L^p $(1 \leq p < \infty)$ are in fact Banach spaces (i.e. complete normed spaces).

Theorem 22. *If $1 \leq p < \infty$ and f_n is a Cauchy sequence in L^p then there exists $f \in L^p$ s.t. $f_n \to f$ in L^p.*

Proof. We begin by showing that f_n is "Cauchy in measure".

Given $\epsilon, \delta > 0$, choose N s.t. $m, n > N \Rightarrow \|f_m - f_n\|_p < \delta^{1/p}\epsilon$.

Now $\|f_m - f_n\|_p^p \geq \epsilon^p \mu(\{\omega \colon |f_m(\omega) - f_n(\omega)| > \epsilon\})$ so $m, n > N \Rightarrow \mu(\{\omega \colon |f_m(\omega) - f_n(\omega)| > \epsilon\}) < \delta$.

We earlier had a theorem that if a sequence converges in measure, then a subsequence converges almost everywhere (compare the next part of this proof to the proof of that theorem). In fact, it is sufficient for the sequence to be Cauchy in measure.

Choose a subsequence (f_{n_k}) s.t. $m, n \geq n_k \Rightarrow \mu(\{\omega\colon |f_m(\omega) - f_n(\omega)| > 2^{-k}\}) < 2^{-k}$, so that

$$\sum_k \mu(\{\omega\colon |f_{n_k}(\omega) - f_{n_{k+1}}(\omega)| > 2^{-k}\}) < \infty$$

and so by Borel-Cantelli I, $\mu(A) = 0$ where

$$A = \limsup_{k\to\infty} \{\omega\colon |f_{n_k}(\omega) - f_{n_{k+1}}(\omega)| > 2^{-k}\}$$

Unpacking the definition of A says that if $\omega \in A^c$, then there is some k (depending on ω) s.t. $j \geq k \Rightarrow |f_{n_j}(\omega) - f_{n_{j+1}}(\omega)| \leq 2^{-j}$, and so $i, j \geq k \Rightarrow |f_{n_i}(\omega) - f_{n_j}(\omega)| \leq \sum_{j\geq k} 2^{-j} = 2^{-k+1}$.

Hence for $\omega \in A^c$, the sequence $(f_{n_k}(\omega))$ is Cauchy (in $\mathbb{R}$), so has a limit.

Define $f(\omega) = \limsup f_{n_k}(\omega)$. This is measurable, and if $\omega \in A^c$ then $f(\omega) = \lim f_{n_k}(\omega)$ so $f_{n_k} \to f$ almost everywhere.

We need to check that $f \in L^p$. This follows from Fatou since $|f|^p = \liminf_k |f_{n_k}|^p$ a.e., so $\int |f|^p d\mu \leq \liminf_k \int |f_{n_k}|^p d\mu$ and $\int |f_{n_k}|^p d\mu \leq (\|f_{n_k} - f_{n_1}\|_p + \|f_{n_1}\|_p)^p \leq (1 + \|f_{n_1}\|_p)^p$ for all k.

Finally we show that $\|f_n - f\|_p \to 0$.

Given $\epsilon > 0$, choose k s.t. $m, n \geq n_k \Rightarrow \|f_n - f_m\|_p < \epsilon$.

For $m \geq n_k$, $|f_m - f|^p = \lim_k |f_m - f_{n_k}|^p = \liminf_k |f_m - f_{n_k}|^p$ a.e. so by Fatou,

$$\int |f_m - f|^p d\mu \leq \liminf_{k\to\infty} \int |f_m - f_{n_k}|^p d\mu \leq \epsilon^p$$

So $m \geq n_k \Rightarrow \|f_m - f\|_p \leq \epsilon$. □

8.6 The space L^∞

There is one more space to consider: the space L^∞ of essentially bounded functions.

A function $f \in \mathrm{m}\mathcal{S}$ is *essentially bounded* if there is some K s.t. $\mu(\{\omega\colon |f| > K\}) = 0$.

We define a norm (the *essential supremum*) by $\|f\|_\infty = \inf_{N\in\mathcal{S}, \mu(N)=0} \sup_{\omega\notin N} |f(\omega)|$.

Let $\mathcal{L}^\infty$ be the set of essentially bounded functions $(S, \mathcal{S}) \to \mathbb{R}$, and identify a.e.-equal functions to form the quotient L^∞ as before. $\|\cdot\|_\infty$ is a norm on L^∞; the triangle inequality is a piece of straightforward analysis.

A sequence $f_n \to f$ in L^∞ iff there is a null set N s.t. $f_n \to f$ uniformly on $S\backslash N$ (this uses the fact that a countable union of null sets is null). The relationship between L^∞ convergence and uniform convergence is the same as the relationship between almost everywhere convergence and pointwise convergence.

We can use this to show that L^∞ is complete, by showing that a Cauchy sequence in L^∞ is uniformly Cauchy except on a null set, then using the completeness of the uniform norm.

9 L^2 and conditional expectation

> [1] L^2 as a Hilbert space. Orthogonal projection, relation with conditional probability. Variance and covariance.

9.1 Conditional probability

In elementary probability, for an event B with $\mathbb{P}(B) > 0$, and for any event A, we define the *conditional probability of A given B* as

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

Suppose we perform an experiment with a finite set of possible outcomes, with events $B_1, \ldots, B_n$ corresponding to each outcome (these events are disjoint and cover Ω, and must all have non-zero probability). After gaining this information, we want to know the probability of some other event A; this will be $\mathbb{P}(A|B_j)$ for the appropriate j. Hence our guess as to the probability of A after the experiment is itself a random variable:

$$g(\omega) = \mathbb{P}(A \cap B_j)/\mathbb{P}(B_j) \text{ when } \omega \in B_j$$

Now consider a more sophisticated experiment, where the set of possible outcomes may be infinite and may contain zero-probability events. Let $\mathcal{G}$ be the set of all events for which the experiment tells us whether they occur or not; these form a sub-σ-field of $\mathcal{F}$. (In the above finite case, $\mathcal{G} = \sigma(B_1, \ldots, B_n)$ is the set of all possible unions of the B_j. If the experiment measures some random variable X, then $\mathcal{G} = \sigma(X)$.)

Again, we want a random variable g which gives the probability of A after knowing the results of this experiment.

The value of g must of course depend only on events in $\mathcal{G}$; for example, in the case $\mathcal{G} = \sigma(B_1, \ldots, B_n)$, g is constant on each B_j. This condition is made precise by saying that g must be $\mathcal{G}$-measurable (note that since $\mathcal{G} \subset \mathcal{F}$, this is a stronger condition than being $\mathcal{F}$-measurable).

Secondly, consider $\mathbb{P}(A \cap B)$ for any $B \in \mathcal{G}$. Once we have performed the experiment, we know either that B did not occur, so $A \cap B$ certainly does not occur, or B did occur, in which case $A \cap B$ occurs with conditional probability $g(\omega)$. If we add (integrate) this for all ω, then we should get $\mathbb{P}(A \cap B)$: $\int_B g d\mathbb{P} = \mathbb{P}(A \cap B)$.

So we define a *(version of the) conditional probability of A given $\mathcal{G}$* to be a random variable $\mathbb{P}(A|\mathcal{G})$ s.t.

1. $\mathbb{P}(A|\mathcal{G})$ is $\mathcal{G}$-measurable
2. For any $B \in \mathcal{G}$, $\int_B \mathbb{P}(A|\mathcal{G}) d\mathbb{P} = \mathbb{P}(A \cap B)$

It is not at all obvious that such a random variable exists in general, or that it is unique (hence the need to talk about versions); we shall worry about this later.

9.2 Conditional expectation

If we fix $\mathcal{G}$ and ω, and allow A to vary, then $\mathbb{P}(A|\mathcal{G})(\omega)$ gives a probability measure on $(\Omega, \mathcal{F})$. We can integrate with respect to this measure to get the conditional expectation of a random variable X:

$$\mathbb{E}(X|\mathcal{G})(\omega) = \int_\Omega X d(\mathbb{P}(-|\mathcal{G})(\omega))$$

Now fixing X and considering this as a function of ω, we get a new random variable $\mathbb{E}(X|\mathcal{G})$. This will have the properties (these follow from the corresponding properties for conditional probability, by the standard machine):

1. $\mathbb{E}(X|\mathcal{G})$ is $\mathcal{G}$-measurable
2. For any $B \in \mathcal{G}$, $\int_B \mathbb{E}(X|\mathcal{G}) d\mathbb{P} = \int_B X d\mathbb{P}$

We define a *(version of the) conditional expectation of X given $\mathcal{G}$* to be a random variable with these two properties.

Again it is not clear that such a random variable exists. And it is only unique up to almost sure equality: if g is a version of $\mathbb{E}(X|\mathcal{G})$ and f a $\mathcal{G}$-measurable r.v. which is zero almost surely, then $g + f$ is still a version of $\mathbb{E}(X|\mathcal{G})$.

However, if g_1, g_2 are two versions of $\mathbb{E}(X|\mathcal{G})$, then let $B = \{\omega : g_1(\omega) \geq g_2(\omega) + \epsilon\}$. $B \in \mathcal{G}$ since g_1, g_2 are $\mathcal{G}$-measurable so $\int_B g_1 d\mathbb{P} = \int_B X d\mathbb{P} = \int_B g_2 d\mathbb{P}$ but $\int_B (g_1 - g_2) d\mathbb{P} \geq \epsilon \mathbb{P}(B)$ so $\mathbb{P}(B) = 0$ for any ϵ. Hence $g_1 = g_2$ almost surely.

9.3 L^2 as a Hilbert space

We shall prove the existence of conditional expectation for square-integrable random variables, by exploiting the fact that $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a Hilbert space i.e. its norm is given by an inner product. (Linear Analysis yields more light on why this is the case, specifically that L^2 is self-dual). The inner product is

$$\langle f, g \rangle = \int f g d\mu$$

which is finite for $f, g \in L^2$ by Hölder (the case $p = q = 2$, often called the Schwarz or Cauchy-Schwarz inequality).

In the case of probability, the inner product has an interpretation in terms of covariance: the *covariance* of $X, Y \in L^2$ is $\mathrm{Cov}(X, Y) = \langle X - \mathbb{E}X, Y - \mathbb{E}Y \rangle = \langle X, Y \rangle - \mathbb{E}(X)\mathbb{E}(Y)$.

The *variance* of a random variable is $\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \|X - \mathbb{E}X\|_2^2$, which is finite iff $X \in L^2$.

The benefit of Hilbert space is that we can apply geometrical ideas. In particular there is an idea of orthogonality: $f, g \in L^2$ are *orthogonal* if $\langle f, g\rangle = 0$. We also have the parallelogram law:

$$\|f+g\|^2 + \|f-g\|^2 = 2(\|f\|^2 + \|g\|^2)$$

9.4 Orthogonal projection

An important geometrical property of Hilbert spaces is the following, proved in Linear Analysis:

Theorem 23. *If G is a closed subspace of a Hilbert space H, then there is a unique continuous linear map $\pi : H \to G$ s.t. $\langle f - \pi f, g\rangle = 0$ for all $f \in H, g \in G$. This map has the properties:*

1. *$\|f - \pi f\| \leq \|f - g\|$ for all $g \in G$*
2. *$\pi g = g$ for all $g \in G$*
3. *$\langle \pi f, g\rangle = \langle f, \pi g\rangle$ for all $f, g \in H$*

This map π is called the *orthogonal projection* from H to G. Note that "G a closed subspace" means that G must be closed in the topology on H and must be a subspace of H as a vector space.

Work with $H = L^2(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. The space $G = L^2(\Omega, \mathcal{G}, \mathbb{P})$ is a closed subspace of H (closed because it is complete), so we may let π be the orthogonal projection $H \to G$.

For any $X \in H$, $\pi(X)$ is $\mathcal{G}$-measurable since $\pi(X) \in G$. And for any $B \in \mathcal{G}$, $I_B \in G$ so

$$\int_B \pi(X) d\mathbb{P} = \langle I_B, \pi(X)\rangle = \langle \pi(I_B), X\rangle = \langle I_B, X\rangle = \int_B X d\mathbb{P}$$

Hence $\pi(X)$ is a conditional expectation of X given $\mathcal{G}$. Since we are working with L^2 rather than $\mathcal{L}^2$, this is only defined up to almost sure equality, corresponding to the uniqueness property of conditional expectation.

Intuitively, a projection is an operation into a subspace which discards information lying outside that subspace: for example, consider $\mathbb{R}^n$ with basis $\{e_1, \ldots, e_n\}$. The orthogonal projection onto the subspace generated by $\{e_1, \ldots, e_m\}$ corresponds to throwing away coordinates after the m-th. In the conditional expectation case, π is throwing away information not contained in the σ-field $\mathcal{G}$.

The property $\|X - \mathbb{E}(X|\mathcal{G})\|_2 \leq \|X - Y\|_2$ for all $Y \in G$ shows that for $X \in L^2$, $\mathbb{E}(X|\mathcal{G})$ is the best estimate we can give for X after knowing the information represented by $\mathcal{G}$, if we measure "best" by minimising $\mathbb{E}(|X - Y|^2)$.

10 Ergodicity

[4] The strong law of large numbers, proof for independent random variables with bounded fourth moments. Measure preserving transformations, Bernoulli shifts. Statements *and proofs* of maximal ergodic theorem and Birkhoff's almost everywhere ergodic theorem, proof of the strong law.

10.1 Strong law of large numbers

Theorem 24 (Strong law of large numbers)**.** *If X_n are i.i.d. r.v.s with $\mathbb{E}|X_1| < \infty$ then*

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n} \to \mathbb{E}X_1 \text{ almost surely.}$$

The strong law of large numbers tells us that if we take finite samples from some probability distribution, as we increase the sample size, the sample mean converges almost surely to the expectation of the distribution. This is important because it justifies the intuitive understanding of expectation and many applications in statistics.

The weak law is a strictly weaker result than the strong law: it says that, with the same hypothesis, the sample mean converges in probability to the expectation. It may still be useful because it has a simpler proof than the strong law. Indeed it is non-trivial that $\{\omega \colon \overline{X}_n \to \mathbb{E}X_1\}$

is measurable; this holds because $\overline{X}_n$ is certainly measurable, so $\limsup_n \overline{X}_n, \liminf_n \overline{X}_n$ are measurable, and so $\{\limsup \overline{X}_n = \liminf \overline{X}_n = \mathbb{E}X_1\}$ is measurable. This is not a problem for the weak law, which only talks about events $\{|\overline{X}_n - \mathbb{E}X_1| > \epsilon\}$.

There are many proofs of the strong law, which vary in the conditions they require and techniques they use. We begin with an easy proof, subject to the assumption that the random variables have bounded fourth moments. Note that this proof does not require the variables to be identically distributed; the condition that $\mathbb{E}X_n = 0$ is just for simplicity and can be obtained by translating.

Theorem 25. *If X_n are independent r.v.s with $\mathbb{E}X_n = 0$ and $\mathbb{E}(X_n^4) < c < \infty$ then $\overline{X}_n \to 0$ a.s.*

Proof. Given $\epsilon > 0$, we show that $\mathbb{P}(|\overline{X}_n| > \epsilon \text{ infinitely often}) = 0$. Then we are done because $\overline{X}_n \not\to 0 \Leftrightarrow ((\overline{X}_n > |j^{-1}| \text{ infinitely often}) \text{ for some } j \in \mathbb{N})$, expressing $\overline{X}_n \not\to 0$ as a countable union of probability 0 events.

Let $S_n = X_1 + \ldots + X_n$ and consider $S_n^4 = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n X_i X_j X_k X_l$.

If there is one of i, j, k, l not equal to any of the rest (say i) then by independence, $\mathbb{E}(X_i X_j X_k X_l) = \mathbb{E}(X_i)\mathbb{E}(X_j X_k X_l) = 0$.

So the only non-zero terms in the expansion of $\mathbb{E}(S_n^4)$ have form $\mathbb{E}(X_i^2 X_j^2)$ or $\mathbb{E}(X_i^4)$.

By Hölder, $\mathbb{E}(X_i^2 X_j^2) \leq \left(\mathbb{E}(X_i^4)\right)^{1/2} \left(\mathbb{E}(X_j^4)\right)^{1/2} < c$.

There are $3n(n-1)$ terms $X_i^2 X_j^2$ in S_n^4 and n terms X_i^4, so we get $\mathbb{E}(S_n^4) < 3n^2 c$.

Now $\epsilon^4 n^4 \mathbb{P}(S_n^4 > \epsilon^4 n^4) \leq \mathbb{E}(S_n^4) < 3n^2 c$.

So $\mathbb{P}(|S_n|/n > \epsilon) < 3c\epsilon^{-4} n^{-2}$ and $\sum_n \mathbb{P}(|S_n|/n > \epsilon) < 3c\epsilon^{-4} \sum n^{-2} < \infty$.

Borel-Cantelli I gives us $\mathbb{P}(\limsup\{|\overline{X}_n| > \epsilon\}) = 0$. □

10.2 Measure-preserving transformations

A *measure-preserving transformation* is a measurable function $T : (S, \mathcal{S}, \mu) \to (S, \mathcal{S})$ s.t. $\mu(T^{-1}A) = \mu(A)$ for all $A \in \mathcal{S}$. (Measurability of T ensures that $T^{-1}A$ is always measurable. Note that the definition contains T^{-1} rather than T.)

We can restate this definition in terms of measurable functions on $(S, \mathcal{S})$ using the standard machine: for any $A \in \mathcal{S}$, $\omega \in T^{-1}A \Leftrightarrow T\omega \in A$ so

$$\int I_A d\mu = \int I_{T^{-1}A} d\mu = \int (I_A \circ T) d\mu$$

Linearity and monotone convergence give us $\int f d\mu = \int (f \circ T) d\mu$ for all measurable $f : S \to \mathbb{R}$. The converse is also true: if $\int f d\mu = \int (f \circ T) d\mu$ for all measurable $f : S \to \mathbb{R}$ then f is measure-preserving.

A measurable set A is *T-invariant* if $\mu(A \triangle T^{-1}A) = 0$ ($\triangle$ is symmetric difference). In other words, A is almost the same as $T^{-1}A$, except for a null set. Note that if μ is finite then the T-invariant sets form a σ-field.

10.3 Stationary sequences

An important example of measure-preserving transformations comes from stationary sequences.

A sequence of random variables (X_n) is *stationary* if $(X_n)_{n \geq 1}$ has the same distribution as $(X_{n+1})_{n \geq 1}$.

Examples of stationary sequences are any i.i.d. sequence and the values of an aperiodic positive-recurrent Markov chain with an invariant distribution as the initial distribution.

Instead of considering (X_n) as a sequence of functions $\Omega \to \mathbb{R}$, it is convenient to consider it as a single function taking values in $\mathbb{R}^{\mathbb{N}}$, the space of real-valued sequences. We use the product σ-field $B(\mathbb{R}^{\mathbb{N}})$ on $\mathbb{R}^{\mathbb{N}}$; this is defined as the smallest σ-field on which all the projection maps (picking out one element of the sequence) are measurable; since we are taking a product of only countably many copies of $\mathbb{R}$, this is the same as the σ-field generated by sets of the form $\prod_{n=1}^{\infty} A_n$ where $A_n \in B(\mathbb{R})$. Hence we can define a *random sequence* to be a function $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}^{\mathbb{N}}$ which is measurable w.r.t. $B(\mathbb{R}^{\mathbb{N}})$. Then $X = (X_1, X_2, \ldots)$ is a random sequence iff $X_1, X_2, \ldots$ are random variables.

To obtain a measure-preserving transformation from a stationary sequence, consider the *Bernoulli shift*: the map $S : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ which throws away the first element of the sequence, i.e. $S(x_1, x_2, \ldots) = (x_2, x_3, \ldots)$. Note that this need not correspond to any transformation on the

original probability space Ω (i.e. there need not be any transformation $T : \Omega \to \Omega$ such that $X \circ T = S \circ X$).

Nevertheless, the sequence X induces a measure $\mathbb{P} \circ X^{-1}$ on $(\mathbb{R}^{\mathbb{N}}, B(\mathbb{R}^{\mathbb{N}}))$. We can then consider the Bernoulli shift as a transformation $(\mathbb{R}^{\mathbb{N}}, B(\mathbb{R}^{\mathbb{N}}), \mathbb{P} \circ X^{-1}) \to (\mathbb{R}^{\mathbb{N}}, B(\mathbb{R}^{\mathbb{N}}))$. S is always measurable w.r.t. $B(\mathbb{R}^{\mathbb{N}})$, and it is easy to see that S preserves the measure $\mathbb{P} \circ X^{-1}$ iff the sequence is stationary (simply restating definitions).

Conversely, any measure-preserving transformation T on $(\Omega, \mathcal{F}, \mathbb{P})$ together with a random variable $f : (\Omega, \mathcal{F}) \to (\mathbb{R}, B(\mathbb{R}))$ gives rise to a stationary sequence defined by $X_n(\omega) = f(T^n(\omega))$.

10.4 Ergodic theorems

Birkhoff's (pointwise) ergodic theorem is a generalisation of the strong law of large numbers which applies to general stationary sequences. The maximal ergodic theorem is used as a lemma in the proof, and I believe it has other applications in analysis but I don't know about them. The proofs are in the Appendix.

Theorem 26 (Maximal Ergodic Theorem)**.** *If $T : (S, \mathcal{S}, \mu) \to (S, \mathcal{S}, \mu)$ is measure-preserving, $f \in L^1(S, \mathcal{S}, \mu)$, $S_n(\omega) = \sum_{k=1}^n f(T^{k-1}\omega)$ and $A = \{\omega \colon \sup S_n > 0\}$, then $\int_A f d\mu \geq 0$.*

Theorem 27 (Birkhoff's Ergodic Theorem)**.** *If $T : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega, \mathcal{F}, \mathbb{P})$ is measure-preserving, $\mathcal{I}$ is the invariant σ-field of T, and $f \in L^1$ then*

$$\frac{1}{n} \sum_{k=1}^{n} f(T^{k-1}\omega) \to \mathbb{E}(f|\mathcal{I}) \text{ almost surely.}$$

This can also be viewed from the point of view of dynamics: we take our probability space as the space of states for some dynamical system and let T be the transformation corresponding to advancing by a single unit of time (Liouville's theorem in dynamics tells us that this transformation is measure-preserving). The invariant σ-field $\mathcal{I}$ consists of the components of the state space which a particle (almost surely) does not enter or leave.

Then $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n f(T^{k-1}\omega)$ is the average over time of the value of f for a single particle, and $\mathbb{E}(f|\mathcal{I})$ is the average value of f over all points in a component of the state space, at an instant in time. Birkhoff's ergodic theorem tells us that the time average and space average are the same.

10.5 Ergodic transformations

The $\mathbb{E}(f|\mathcal{I})$ on the right hand side of Birkhoff's ergodic theorem suggests that we should be particularly interested in transformations for which $\mathcal{I}$ is trivial (i.e. $\mathbb{P}(A) = 0$ or 1 $\forall A \in \mathcal{I}$). Since $\mathbb{E}(f|\mathcal{I})$ is a $\mathcal{I}$-measurable r.v., for such a σ-field it will be almost surely constant. So we define:

A measure-preserving transformation T is *ergodic* if its invariant σ-field is trivial.

Note that the value which an a.s. constant r.v. takes with probability 1 must be equal to its expectation. So $\mathbb{E}(f|\mathcal{I}) = \mathbb{E}(\mathbb{E}(f|\mathcal{I})) = \mathbb{E}(f)$ almost surely.

So to prove the strong law of large numbers, all we need to prove is:

Lemma 18. *If $(X_1, X_2, \ldots)$ is a sequence of i.i.d. r.v.s, then the shift operator S is ergodic.*

Proof. As observed above, we are treating S as a measure-preserving transformation on the probability space $(\mathbb{R}^{\mathbb{N}}, B(\mathbb{R}^{\mathbb{N}}), \mathbb{P} \circ X^{-1})$.

Call a measurable set A *strictly invariant* (non-standard terminology) if $S^{-1}A = A$. The strictly invariant sets form a σ-field $\mathcal{J}$.

We shall show that $\mathcal{J}$ is contained in the tail σ-field $\mathcal{T}$ so by Kolmogorov's 0-1 law $\mathcal{J}$ is trivial. We shall also show that an invariant set differs from a strictly invariant set by a null set, so $\mathcal{I}$ is trivial and S is ergodic.

Informally, what we do is observe that $\omega \in A \Leftrightarrow S\omega \in A$ and that whether $S\omega$ is in A can be determined by looking at just $X_2, X_3, \ldots$, so $A \in \sigma(X_2, X_3, \ldots)$.

Doing this formally requires quite a bit of manipulation of σ-fields:

Note that $X_n(S\omega) = X_{n+1}(\omega)$ so $X_n \circ S = X_{n+1}$. So the functions $X_n \circ S$ are all measurable on $\sigma(X_2, X_3, \ldots)$, and $\sigma(X_1 \circ S, X_2 \circ S, \ldots) \subset \sigma(X_2, X_3, \ldots)$.

Let $\mathcal{F}_0 = \{A \subset \Omega\colon S^{-1}A \in \sigma(X_1 \circ S, X_2 \circ S, \ldots)\}$. It is straightforward to check that this is a σ-field. If B is a measurable subset of $\mathbb{R}$, then $S^{-1}(X_n^{-1}(B)) = (X_n \circ S)^{-1}(B) \in \sigma(X_1 \circ S, X_2 \circ S, \ldots)$ so $X_n^{-1}(B) \in \mathcal{F}_0$ for all n. So X_n is $\mathcal{F}_0$-measurable for all n, so $\mathcal{F}_0 \supset \sigma(X_1, X_2, \ldots)$.

Now given $A \in \mathcal{J}$, we have $A \in \sigma(X_1, X_2, \ldots) \subset \mathcal{F}_0$ so $A = S^{-1}A \in \sigma(X_2, X_3, \ldots)$.

Similarly, $A \in \sigma(X_n, X_{n+1}, \ldots)$ for all n, so $A \in \mathcal{T}$.

Given $A \in \mathcal{I}$, we need to check that there is $B \in \mathcal{J}$ s.t. $A \triangle B$ is a null set. There ought to be an easier way than the following, but this is the best I have come up with.

Let $B = \limsup_{n \geq 0} S^{-n}A$. Then $S^{-1}B = \limsup_{n \geq 1} S^{-n}A = B$ so $B \in \mathcal{J}$.

By lots of set chasing,

$$\begin{aligned} A \triangle B = (A \setminus B) \cup (B \setminus A) &= \liminf(A \setminus S^{-n}A) \cup \limsup(S^{-n}A \setminus A) \\ &\subset \liminf(A \triangle S^{-n}A) \cup \limsup(A \triangle S^{-n}A) \\ &= \limsup(A \triangle S^{-n}A) \end{aligned}$$

Now $A \triangle S^{-n}A \subset \bigcup_{k=1}^{n}(S^{-k+1}A \triangle S^{-k}A)$ and as S is measure-preserving,

$$\mu(S^{-k+1}A \triangle S^{-k}A) = \mu(S^{-1}A \triangle A) = 0$$

so $\mu(A \triangle S^{-n}A) = 0$ for all n.

Hence $\mu(A \triangle B) = 0$. □

11 Characteristic functions

> [2] The Fourier transform of a finite measure, characteristic functions, uniqueness and inversion. Weak convergence, statement of Lévy's convergence theorem for characteristic functions.

This discussion of Fourier transforms is of course more rigorous than the treatment in IB Methods, but it is still not entirely satisfactory. For example, there is annoying asymmetry between the domains on which the Fourier transform and the inverse Fourier transform are defined.

11.1 Fourier transforms

The *Fourier transform* of a finite measure μ on $(\mathbb{R}, B(\mathbb{R}))$ is the function $\hat{\mu} : \mathbb{R} \to \mathbb{C}$

$$\hat{\mu}(t) = \int e^{itx}\mu(dx)$$

(Note of course the usual variety of conventions in defining Fourier transforms: it may be defined as the conjugate of this, or with a constant multiplier. Note also that we extend integration from real-valued to complex-valued functions just by adding real and imaginary parts, so long as $\int |f| d\mu < \infty$.)

The Fourier transform defined in applied maths as $\hat{f}(t) = \int_{-\infty}^{\infty} e^{itx} f(x)dx$ is the Fourier transform in this sense of the measure with density function f. (Not all measures have density functions; I will add a definition and discussion of them somewhere eventually.)

In the case of a random variable X, the Fourier transform with respect to its Lebesgue-Stieltjes measure is the *characteristic function* $\varphi_X(t) = \mathbb{E}(e^{itX})$.

Note that $\hat{\mu}(0) = \mu(\mathbb{R})$ and $|\hat{\mu}(t)| \leq \mu(\mathbb{R})$ for all t.

Lemma 19. *The Fourier transform φ of a finite measure μ is uniformly continuous.*

Proof. We have $|e^{itx} - 1| \to 0$ as $t \to 0$, $|e^{itx} - 1| \leq 2$ for all t, and the constant 2 is integrable since μ is finite, so by the Dominated Convergence Theorem,

$$\int |e^{itx} - 1|\mu(dx) \to 0 \text{ as } t \to 0$$

Now given $\epsilon > 0$, we can find $\delta > 0$ s.t. $|h| < \delta \Rightarrow \int |e^{ihx} - 1|\mu(dx) < \epsilon$. For any $t, h \in \mathbb{R}$ with $|h| < \delta$, we have

$$|\hat{\mu}(t+h) - \hat{\mu}(t)| = \left|\int e^{itx}(e^{ihx} - 1)\mu(dx)\right| \leq \int |e^{itx}||e^{ihx} - 1|\mu(dx) = \int |e^{ihx} - 1|\mu(dx) < \epsilon$$

□

11.2 Inverse of Fourier transform

Given a function $\hat{\mu} : \mathbb{R} \to \mathbb{C}$ which is the Fourier transform of some finite measure μ, we would like to recover μ; in particular it will be important to know that there is only one measure with given Fourier transform. (We do not consider here conditions for such a measure to exist; we simply assume that the $\hat{\mu}$ we are given is the transform of some measure.)

Recall the inversion formula from Methods: $f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\hat{f}(t)dt$. But we do not know that the RHS is integrable; or from another point of view, this formula constructs a density f for μ, so can only work if μ has a density function.

To allow us to use this formula, we smooth out μ so that it does have a density. This can be done by convolving with a *mollifier*, a smooth function with properties that allow us to reconstruct the original measure by taking a limit. This convolution of course corresponds to multiplication of Fourier transforms.

A suitable function to use (both this function and its Fourier transform are smooth) is $\exp(-\epsilon t^2/2)$. As $\epsilon \to 0$, this tends to 1 and so $\exp(-\epsilon t^2/2)\hat{\mu}(t) \to \hat{\mu}(t)$. (Another way of looking at this is that we are mixing a small amount of Gaussian noise into the distribution.)

Now we can define

$$f_\epsilon(x) = \frac{1}{2\pi}\int e^{-itx}\exp(-\epsilon t^2/2)\hat{\mu}(t)dt$$

This integral is defined and finite because $e^{-itx}\exp(-\epsilon t^2/2)\hat{\mu(t)}$ is continuous, so measurable, and it is dominated by $\exp(-\epsilon t^2/2)\mu(\mathbb{R})$ which is integrable.

Now $f_\epsilon(x)$ should be an approximate density for μ. In fact we have

$$\begin{aligned} f_\epsilon(x) &= \frac{1}{2\pi}\int e^{-itx}\exp(-\epsilon t^2/2)\left(\int e^{itx'}\mu(dx')\right)dt \\ &= \frac{1}{2\pi}\int\int \exp(it(x'-x) - \epsilon t^2/2)\mu(dx')dt \\ &= \frac{1}{2\pi}\int\int e^{it(x'-x)}\exp(-\epsilon t^2/2)dt\,\mu(dx') \text{ by Fubini} \end{aligned}$$

(The application of Fubini is justified because the integrand is dominated by $\exp(-\epsilon t^2/2)$ and μ is finite, so $\int\int \exp(-\epsilon t^2/2)\mu(dx')dt < \infty$.)

This contains the Fourier transform of a Gaussian (just with different labels), which we know how to compute, so we get

$$f_\epsilon(x) = \frac{1}{\sqrt{2\pi\epsilon}}\int \exp(-(x'-x)^2/2\epsilon)\mu(dx')$$

This is the density of an approximation to μ, so we get the measure by integrating. Note that $g_\epsilon(x,x') = \exp(-(x'-x)^2/2\epsilon)$ is non-negative, so we can apply Tonelli's (a.k.a. Fubini's) theorem without having to check that the integrals involved are finite. Define

$$\begin{aligned} F_\epsilon(a) &= \int_{-\infty}^{a} f_\epsilon(x)dx \\ &= \frac{1}{\sqrt{2\pi\epsilon}}\int_{-\infty}^{a}\int \exp(-(x'-x)^2/2\epsilon)\mu(dx')dx \\ &= \frac{1}{\sqrt{2\pi\epsilon}}\int\int_{-\infty}^{a} \exp(-(x-x')^2/2\epsilon)dx\,\mu(dx') \text{ (Tonelli)} \\ &= \int \Phi\left(\frac{a-x'}{\sqrt{\epsilon}}\right)\mu(dx') \end{aligned}$$

where $\Phi(x) = \int_{-\infty}^{x}\exp(-t^2/2)dt$ is the standard Gaussian distribution function.

Now as $\epsilon \to 0$,

$$\Phi\left(\frac{a-x'}{\sqrt{\epsilon}}\right) \to \begin{cases} 0 & \text{if } a < x' \\ \frac{1}{2} & \text{if } a = x' \\ 1 & \text{if } a > x' \end{cases}$$

and it is dominated by 1, so by DCT we get

$$F_\epsilon(a) \to \int \tfrac{1}{2}I_{\{a=x'\}} + I_{\{a>x'\}}\mu(dx') = \mu((-\infty,a)) + \tfrac{1}{2}\mu(\{a\})$$

This is sufficient to determine μ on the π-system $\{(-\infty,a) : a \in \mathbb{R}\}$, using the fact that $\mu((-\infty,a))$ is increasing and right-continuous, and so to determine μ.

11.3 Weak convergence

The modes of convergence we considered earlier were convergence of functions. Now we consider a form of convergence of measures.

Let $(S, \mathcal{S})$ be a metric space with its Borel σ-field.

A sequence of measures μ_n on $(S, \mathcal{S})$ *converges weakly* to μ if $\int f d\mu_n \to \int f d\mu$ for all bounded, continuous functions $f : S \to \mathbb{R}$. We write this $\mu_n \Rightarrow \mu$.

Note that weak convergence can be considered as a mode of convergence of random variables, by treating it as weak convergence of the corresponding Lebesgue-Stieltjes measure. Regarded in this way, weak convergence of r.v.s is implied by convergence in probability, and so by any of the modes of convergence considered earlier.

There is an alternative characterisation of weak convergence which is often easier to prove:

Theorem 28. *$\mu_n \Rightarrow \mu$ iff $\mu(A) \leq \liminf \mu_n(A)$ for all open sets A.*

Proof. Suppose that $\mu_n \Rightarrow \mu$ and let A be an open set.

Consider the function $h_n(x) = \inf\{1, n|x - y| : y \in A^c\}$, which is a bounded and continuous approximation to I_A. (A^c closed ensures that $\inf\{|x - y| : y \in A^c\} > 0$ for $x \in A$.)

Specifically, $0 \leq h_n \uparrow I_A$.

So for any k we have

$$\begin{aligned} \liminf \mu_n(A) \geq \liminf \textstyle\int h_k d\mu_n = \int h_k d\mu \qquad & \text{(weak convergence)} \\ \uparrow \textstyle\int I_A d\mu = \mu(A) \qquad & \text{(monotone convergence)} \end{aligned}$$

Conversely, suppose that $\mu(A) \leq \liminf \mu_n(A)$ for all open sets A and let f be a bounded, continuous function. Since f is bounded, we can add a constant to it to get $f \geq 0$.

The integral of a measurable function is defined by looking at sums of indicators of measurable sets. Here we only know things about the measure of open sets, but on the other hand f is continuous so can be approximated by sums of indicators of open sets. (Essentially we are computing the Riemann instead of the Lebesgue integral of f.)

Fix $m \in \mathbb{N}$ (higher m means better approximation of f).

For $j \in \mathbb{N}$, let $A_j = \{\omega : f(\omega) > j/m\}$, which is open.

Now $I_{A_j}(\omega) = 1$ iff $1 \leq j < mf(\omega)$ so

$$\sum_{j=1}^{\infty} I_{A_j}(\omega) \leq mf(\omega) \leq \sum_{j=1}^{\infty} I_{A_j}(\omega) + 1 \qquad (\dagger)$$

Integrating (using monotone convergence on the LHS) gives

$$\sum_{j=1}^{\infty} \mu_n(A_j) \leq m \int f d\mu_n$$

so

$$\sum_{j=1}^{\infty} \mu(A_j) \leq \liminf_{n\to\infty} \sum_{j=1}^{\infty} \mu_n(A_j) \leq \liminf_{n\to\infty} m \int f d\mu_n$$

We can also integrate (†) with respect to μ to get

$$m \int f d\mu \leq \sum_{j=1}^{\infty} \mu(A_j) + \mu(S)$$

so combining these,

$$\int f d\mu \leq \liminf_{n\to\infty} \int f d\mu_n + \frac{1}{m}\mu(S)$$

We can apply the same argument to $-f$ to get

$$\limsup_{n\to\infty} \int f d\mu_n - \frac{1}{m}\mu(S) \leq \int f d\mu$$

Since these hold for all $m \in \mathbb{N}$, we have $\lim_{n\to\infty} \int f d\mu_n = \int f d\mu$. □

11.4 Convergence in distribution

A sequence of distribution functions F_n *converges in distribution* to a distribution function F if $F_n(x) \to F(x)$ for all x at which F is continuous.

Theorem 29. *Let F_n, F be distribution functions, with Lebesgue-Stieltjes measures μ_n, μ. Then $F_n \to F$ in distribution iff $\mu_n \Rightarrow \mu$.*

Proof. Suppose $\mu_n \Rightarrow \mu$, and consider $c \in \mathbb{R}$, s.t. F is continuous at c.

We want to show that $F_n(c) \to F(c)$ or in other words that $\mu_n((-\infty, c]) \to \mu((-\infty, c])$.

For any $\delta > 0$, we have $\mu((-\infty, c-\delta)) \leq \liminf \mu_n((-\infty, c-\delta)) \leq \liminf \mu_n((-\infty, c])$.

Since F is continuous at c, $F(c-2\delta) \uparrow F(c)$ as $\delta \to 0$ and $F(c-2\delta) \leq \mu((-\infty, c-\delta)) \leq F(c)$, so $\mu((-\infty, c-\delta)) \uparrow \mu((-\infty, c])$.

Hence $\mu((-\infty, c]) \leq \liminf \mu_n((-\infty, c])$.

Also, $\mu((c, \infty)) \leq \liminf \mu_n((c, \infty))$ so $\mu((-\infty, c]) \geq \limsup \mu_n((-\infty, c])$.

Hence $\lim \mu((-\infty, c]) = \mu((-\infty, c])$.

Now the other way. Suppose that $F_n \to F$, and let A be an open subset of $\mathbb{R}$. We want to show that $\mu(A) \geq \liminf \mu_n(A)$.

All open subsets of $\mathbb{R}$ are a union of countably many disjoint open intervals, so $A = \bigcup_n (a_n, b_n)$.

It is enough to show that $\mu((a, b)) \leq \liminf \mu_n((a, b))$ for a single open interval (a, b), as we can then add these up.

F has only countably many points of discontinuity, so there is a sequence $a_m \downarrow a$ s.t. F is continuous at a_m for all m. Then $\limsup F_n(a) \leq \limsup F_n(a_m) = F(a_m)$ for all m, so by right continuity of F, $\limsup F_n(a) \leq F(a)$.

Similarly, there is a sequence $b_m \uparrow b$ s.t. $b_m < b$ and F is continuous at all b_m. Then $\liminf F_n(b-) \geq \liminf F_n(b_m) = F(b_m)$ for all m, and $F(b_m) \uparrow F(b-)$ so $\liminf F_n(b-) \geq F(b-)$.

Combining these gives

$$\begin{aligned} \mu((a, b)) = F(b-) - F(a) &\leq \liminf F_n(b-) - \limsup F_n(a) \\ &= \liminf (F_n(b-) - F_n(a)) = \liminf \mu_n((a, b)) \end{aligned}$$

Note that this still works if either a or b is infinite. □

11.5 Lévy's convergence theorem

This theorem connects weak convergence and Fourier transforms, and gives one perspective on why weak convergence is an interesting concept (though the reasons for this come much more strongly from functional analysis). It will be a key step in the proof of the central limit theorem. The proof is non-examinable and is included in the Appendix.

Theorem 30 (Lévy's convergence theorem)**.** *$\mu_n \Rightarrow \mu$ iff $\hat{\mu}_n \to \hat{\mu}$ pointwise.*

11.6 Smoothness of Fourier transforms

The next result illustrates the principle that if a distribution has thin tails, then its Fourier transform is smooth at 0. In particular, if $|x|^n$ is integrable w.r.t. μ, then $\hat{\mu}$ is differentiable n times at zero, and its derivatives are given by the moments of μ (up to factors of i). Besides being of general interest, the result is important in the proof of the central limit theorem.

The result essentially comes from the Taylor expansion of $\exp(itx)$, but we need to use dominated convergence to ensure that the integrals converge. Before proving the main result, we will need bounds on the error in this Taylor expansion. Let $R_n(x) = \exp(ix) - \sum_{j=0}^{n} (ix)^j / j!$

Lemma 20. *For $x \in \mathbb{R}$, $|R_n(x)| \leq 2|x|^n/n!$ and $|R_n(x)| \leq |x|^{n+1}/(n+1)!$*

Proof. We have $R_n(x) = \int_0^x i R_{n-1}(y) dy$, so $|R_n(x)| \leq \int_0^x |R_{n-1}(y)| dy$.

For $n = 0$, we have $|R_0(x)| = |\exp(ix) - 1| \leq 2$.

Also $|R_0(x)| = |\int_0^x i \exp(iy) dy| \leq \int_0^x 1 dy = x$.

So the claims hold for $n = 0$, and then for all n by induction. □

Lemma 21. *If $\int |x|^n \mu(dx) < \infty$, then*

$$\hat{\mu}(t) = \sum_{j=0}^{n} \left(\frac{(it)^j}{j!} \int x^j \mu(dx) \right) + o(|t|^n)$$

Proof.

$$\begin{aligned} \frac{1}{|t|^n} \left| \hat{\mu}(t) - \sum_{j=0}^{n} \left(\frac{(it)^j}{j!} \int x^j \mu(dx) \right) \right| &= \left| t^{-n} \int (\exp(itx) - \sum_{j=0}^{n} (itx)^j / j!) \mu(dx) \right| \\ &= \left| t^{-n} \int R_n(tx) \mu(dx) \right| \\ &\leq \int |t^{-n} R_n(tx)| \mu(dx) \end{aligned}$$

By the previous lemma, $|t^{-n} R_n(tx)| \leq |t||x|^{n+1}/(n+1)! \to 0$ as $t \to 0$, and $|t^{-n} R_n(tx)|$ is dominated by $2|x|^n/n!$, which we have assumed to be integrable. So by dominated convergence, $\int |t^{-n} R_n(tx)| \mu(dx) \to 0$ as $t \to 0$. □

In probabilistic language, this becomes $\varphi_X(t) = \sum_{j=0}^{n} \frac{(it)^j}{j!} \mathbb{E}(X^j) + o(|t|^n)$.

12 Gaussian random variables

> [1] Gaussian random variables, the multivariate normal distribution.
> The central limit theorem.

12.1 Normal distribution

The one-dimensional normal (or Gaussian) distribution $N(\mu, \sigma^2)$ is the distribution on $\mathbb{R}$ with distribution function $F(a) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} \exp(-(x-\mu)^2/2\sigma^2) dx$.

A d-dimensional random variable $\mathbf{X} : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^d, B(\mathbb{R}^d))$ has a *multivariate normal distribution* if $\mathbf{a}.\mathbf{X}$ is normally distributed for every vector $\mathbf{a} \in \mathbb{R}^d$.

If we add a constant vector $\mathbf{b}$ to a MVN r.v. $\mathbf{X}$, then $\mathbf{X} + \mathbf{b}$ still has multivariate normal distribution. So given any MVN r.v. $\mathbf{X} = (X_1, \ldots, X_n)$, we can let $\boldsymbol{\mu} = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_n))$ and then $\mathbf{Y} = \mathbf{X} - \boldsymbol{\mu}$ is multivariate normal with mean 0.

In one dimension, we require only one parameter, the variance, to describe a normal distribution with mean 0. In higher dimensions, we require not only the variance of each component but also more parameters to describe the correlations between components.

We use the *covariance matrix* $V_{ij} = \mathbb{E}(Y_i Y_j)$. We can write all these equations at once by saying $V = \mathbb{E}(\mathbf{Y}\mathbf{Y}^T)$.

Note that this matrix is symmetric, and it is positive semidefinite, as for any $\mathbf{a} \in \mathbb{R}^d$,

$$\mathbf{a}^T V \mathbf{a} = \sum_i \sum_j a_i \mathbb{E}(Y_i Y_j) a_j = \mathbb{E}((\sum_i a_i Y_i)^2) \geq 0$$

Since V is symmetric, it can be diagonalised by an orthogonal matrix, say $V = RDR^T$ with R orthogonal and D diagonal.

Let $\mathbf{Z} = R^T \mathbf{Y}$. For any $\mathbf{a} \in \mathbb{R}^d$, $\mathbf{a}.\mathbf{Z} = (R\mathbf{a}).\mathbf{Y}$ so $\mathbf{Z}$ is still MVN; but now $\mathbb{E}(\mathbf{Z}\mathbf{Z}^T) = \mathbb{E}(R^T \mathbf{Y}\mathbf{Y}^T R) = R^T V R = D$, so $\mathbf{Z}$ has diagonal covariance matrix, and its components are uncorrelated.

Now suppose that V is positive definite (and not just positive semidefinite). The diagonal entries of D are all now positive, so have square roots, and there is a diagonal matrix E with positive entries on the diagonal s.t. $E^2 = D$. If we let $\widetilde{\mathbf{Z}} = E^{-1}\mathbf{Z}$ then $\mathbb{E}(\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T) = \mathbb{E}(E^{-1}\mathbf{Z}\mathbf{Z}^T E^{-1}) = E^{-1} D E^{-1} = I$. So all MVN variables with positive definite covariant matrix can be reduced by linear transformations to an MVN variable with covariance matrix I.

We know that the components of $\widetilde{\mathbf{Z}}$ are uncorrelated, because they have zero covariance. In fact, for components of an MVN variable, we have a stronger property: they are independent.

Theorem 31. *If $\widetilde{\mathbf{Z}}$ is a MVN r.v. with $\mathbb{E}\widetilde{\mathbf{Z}} = 0$ and $\mathbb{E}(\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T) = I$, then the components of $\widetilde{\mathbf{Z}}$ are independent $N(0,1)$ random variables.*

Proof. We need to extend the idea of a characteristic function to d-dimensional random variables: if $\mathbf{X} : \Omega \to \mathbb{R}^d$ is a random variable then its characteristic function is $\varphi_{\mathbf{X}} : \mathbb{R}^d \to \mathbb{C}$ given by $\varphi_{\mathbf{X}}(\mathbf{a}) = \mathbb{E}(\exp(i\mathbf{a}.\mathbf{X}))$.

In the case of MVN $\widetilde{\mathbf{Z}}$, $\mathbf{a}.\widetilde{\mathbf{Z}}$ is $N(0,1)$, so from what we know about the characteristic function the one-dimensional normal distribution, $\varphi_{\widetilde{\mathbf{Z}}}(\mathbf{a}) = \mathbb{E}(\exp(i\mathbf{a}.\widetilde{\mathbf{Z}})) = \exp(-(\mathbf{a}.\widetilde{\mathbf{Z}})^2/2) = \exp(-\mathbf{a}^T\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T\mathbf{a}/2) = \exp(-\mathbf{a}^T I\mathbf{a}/2) = \exp(-|\mathbf{a}|^2/2)$.

If $U_1, \ldots, U_d$ are independent $N(0,1)$ variables, then

$$\begin{aligned}\varphi_{\mathbf{U}}(\mathbf{a}) = \mathbb{E}(\exp(i\mathbf{a}.\mathbf{U})) &= \mathbb{E}\left(\prod_j \exp(ia_kU_k)\right) = \prod_j \mathbb{E}(\exp(ia_kU_k)\\ &= \prod_j \varphi_{U_k}(a_k) = \prod_j \exp(-a_k^2/2) = \exp(-|\mathbf{a}|^2/2) = \varphi_{\widetilde{\mathbf{Z}}}(\mathbf{a})\end{aligned}$$

So as distributions are uniquely determined by their characteristic functions, $\widetilde{\mathbf{Z}}$ and $\mathbf{U}$ have the same distribution. □

12.2 Central limit theorem

Theorem 32 (Central Limit Theorem)**.** *Let (X_n) be an i.i.d. sequence with $\mathbb{E}(X_n) = 0, \mathrm{Var}(X_n) = \sigma^2 < \infty$. Let $S_n = X_1 + \ldots + X_n$ and $Z_n = \dfrac{S_n}{\sigma\sqrt{n}}$. Then Z_n converges in distribution to $N(0,1)$.*

Proof. Let φ_n be the characteristic function of Z_n and ψ the characteristic function of the standard normal distribution. We will show that $\varphi_n \to \psi$ pointwise, and then the Lévy convergence theorem and equivalence of weak convergence to convergence in distribution give the result.

Note that $\psi(t) = \exp(-t^2/2)$.

Let φ be the characteristic function of X_1. Then

$$\begin{aligned}\varphi_n(t) = \mathbb{E}(e^{itZ_n}) &= \mathbb{E}(e^{it(X_1+\ldots+X_n)/\sigma\sqrt{n}})\\ &= \mathbb{E}(e^{itX_1/\sigma\sqrt{n}})\ldots\mathbb{E}(e^{itX_n/\sigma\sqrt{n}}) \text{ since } X_1, \ldots, X_n \text{ are independent}\\ &= (\mathbb{E}(e^{itX_1/\sigma\sqrt{n}}))^n \text{ since } X_1, \ldots, X_n \text{ are identically distributed}\\ &= (\varphi(t/\sigma\sqrt{n}))^n\end{aligned}$$

We now seek to estimate $\varphi(t/\sigma\sqrt{n})^n$ using our earlier "Taylor" estimate of the values of characteristic functions.

We have $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^2) = \sigma^2 < \infty$, so $\varphi(t) = 1 - t^2\sigma^2/2 + o(|t|^2)$ as $t \to 0$.

This gives $\varphi_n(t) = (1 - t^2/2n + o(t^2/\sigma^2 n))^n$ as $t/\sigma\sqrt{n} \to 0$ (i.e. as $n \to \infty$).

Taking logs,

$$\begin{aligned}\log\varphi_n(t) &= n\log((1 - t^2/2n + o(t^2/n))\\ &= n(-t^2/2n + o(t^2/n)) \text{ by Taylor's theorem for } \log(1+x)\\ &\to -t^2/2 \text{ as } n \to \infty\end{aligned}$$

So $\varphi_n(t) \to \exp(-t^2/2)$, as required. □

A Ergodic theorems

I did not find these proofs very enlightening, but they are the standard proofs which appear in most books.

Theorem 33 (Maximal Ergodic Theorem)**.** *If $T : (S, \mathcal{S}, \mu) \to (S, \mathcal{S}, \mu)$ is measure-preserving, $f \in L^1(S, \mathcal{S}, \mu)$, $S_n(\omega) = \sum_{k=1}^n f(T^{k-1}\omega)$ and $A = \{\omega\colon \sup S_n > 0\}$, then $\int_A f d\mu \geq 0$.*

Proof. Let $M_n = \max\{0, S_1, \ldots, S_n\}$, and $A_n = \{\omega\colon M_n > 0\}$. Then M_n is increasing, and $A_n \uparrow A$ so $fI_{A_n} \to fI_A$. And $|fI_{A_n}| \leq |f| \in L^1$ so by the Dominated Convergence Theorem,

$$\int_{A_n} f d\mu \to \int_A f d\mu$$

Hence we just need to show $\int_{A_n} f d\mu \geq 0$ for each n.

Consider the shifted sums $S'_n = S_n \circ T = \sum_{k=1}^n f(T^k\omega)$.
Let $M'_n = M_n \circ T = \max\{0, S'_1, \ldots, S'_n\}$ and $A'_n = T^{-1}(A_n) = \{\omega\colon M'_n > 0\}$.
Now $S_n = f + S'_{n-1}$ so if $M_n > 0$ then $M_n = f + \max\{S'_1, \ldots, S'_{n-1}\} \le f + M'_{n-1}$.
Hence $M_n I_{A_n} - M'_{n-1} I_{A_n} \le f I_{A_n}$.
But we can control $M'_{n-1} I_{A_n}$ using $M'_{n-1} I_{A_n} \le M'_n I_{A_n} \le M'_n I_{A'_n} = (M_n I_{A_n}) \circ T$, so

$$\int_{A_n} f \ge \int (M_n - M'_{n-1}) I_{A_n} \ge \int M_n I_{A_n} - \int (M_n I_{A_n}) \circ T = 0$$

(the last equality since T is measure-preserving). □

Theorem 34 (Birkhoff's Ergodic Theorem)**.** *If $T : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Omega, \mathcal{F}, \mathbb{P})$ is measure-preserving, $\mathcal{I}$ is the invariant σ-field of T, and $f \in L^1$ then*

$$\frac{1}{n} \sum_{k=1}^n f(T^{k-1}\omega) \to \mathbb{E}(f|\mathcal{I}) \text{ almost surely.}$$

Proof. Assume wlog that $\mathbb{E}(f|\mathcal{I}) = 0$.
Let $X_n = f(T^{n-1}\omega), S_n = \sum_{k=1}^n X_k$.
Given $\epsilon > 0$, let $A = \{\omega\colon \limsup_n S_n/n > \epsilon\}$. We need to show that $\mathbb{P}(A) = 0$.
Let $f' = (f - \epsilon) I_A$, $X'_n = (X_n - \epsilon) I_A$ and $S'_n = \sum_{k=1}^n X'_n = (S_n - n\epsilon) I_A$.
Now if A occurs then $S_n > n\epsilon$ for some (in fact infinitely many) n, so $\sup S'_n > 0$.
And if $\sup S'_n > 0$, then $I_A = 1$ so A occurs.
So $A = \{\omega\colon \sup S'_n > 0\}$.
Now for any fixed ω, $X_1/n \to 0$ as $n \to \infty$ so $\limsup_n S_n/n = \limsup_n S'_{n-1}/n = \limsup_n S'_n/n$. Hence $A = T^{-1}(A)$ and X'_n is stationary.
Hence by the maximal ergodic theorem,

$$0 \le \int_A X'_1 d\mathbb{P} = \int_A X_1 d\mathbb{P} - \epsilon \mathbb{P}(A)$$

Now $A \in \mathcal{I}$ so the definition of conditional probability gives

$$\int_A X_1 d\mathbb{P} = \int_A \mathbb{E}(X_1|\mathcal{I}) d\mathbb{P} = 0$$

so we get $\epsilon \mathbb{P}(A) \le 0$ giving $\mathbb{P}(A) = 0$. □