Probability, Measure and Martingales

Michaelmas Term 2019, 16 lectures

Lecturer: James Martin

Version of 8 December 2019

0 Introduction

These notes may be updated and revised as term progresses. They are based on notes by Alison Etheridge and Oliver Riordan.

Comments and corrections are welcome, to `martin@stats.ox.ac.uk`.

0.1 Background

In the last fifty years probability theory has emerged both as a core mathematical discipline, sitting alongside geometry, algebra and analysis, and as a fundamental way of thinking about the world. It provides the rigorous mathematical framework necessary for modelling and understanding the inherent randomness in the world around us. It has become an indispensable tool in many disciplines – from physics to neuroscience, from genetics to communication networks, and, of course, in mathematical finance. Equally, probabilistic approaches have gained importance in mathematics itself, from number theory to partial differential equations.

Our aim in this course is to introduce some of the key tools that allow us to unlock this mathematical framework. We build on the measure theory that we learned in Part A Integration and develop the mathematical foundations essential for more advanced courses in analysis and probability. We'll then introduce the powerful concept of martingales and explore just a few of their remarkable properties.
The nearest thing to a course text is

- David Williams, *Probability with Martingales,* CUP.

Also highly recommended are:

- S.R.S. Varadhan, *Probability Theory,* Courant Lecture Notes Vol. 7.
- R. Durrett, *Probability: theory and examples,* 4th Edition, CUP 2010.
- A. Gut, *Probability: a graduate course*, Springer 2005.

Comment on notation: for probability and expectation, the type of brackets used has no significance – some people use one, some the other, and some whichever is clearest in a given case. So $\mathbb{E}[X]$, $\mathbb{E}(X)$ and $\mathbb{E}X$ all mean the same thing.

What is here called a σ-algebra is sometimes called a σ-field. Our default notation $(\Omega, \mathcal{F}, \mu)$ for a measure space differs from that of Williams, who writes (S, Σ, μ).

0.2 Course Synopsis

Review of σ-algebras, measure spaces. Uniqueness of extension of π-systems and Carathéodory's Extension Theorem, monotone-convergence properties of measures, limsup and liminf of a sequence of events, Fatou's Lemma, reverse Fatou Lemma, first Borel-Cantelli Lemma.

Random variables and their distribution functions, σ-algebras generated by a collection of random variables. Product spaces. Independence of events, random variables and σ-algebras, π-systems criterion for independence, second Borel-Cantelli Lemma. The tail σ-algebra, Kolomogorov's 0-1 Law. Convergence in measure and convergence almost everywhere.

Integration and expectation, review of elementary properties of the integral and L^p spaces [from Part A Integration for the Lebesgue measure on $\mathbb{R}$]. Scheffé's Lemma, Jensen's inequality. The Radon-Nikodym Theorem [without proof]. Existence and uniqueness of conditional expectation, elementary properties. Relationship to orthogonal projection in L^2.

Filtrations, martingales, stopping times, discrete stochastic integrals, Doob's Optional-Stopping Theorem, Doob's Upcrossing Lemma and "Forward" Convergence Theorem, martingales bounded in L^2, Doob decomposition, Doob's submartingale inequalities.

Uniform integrability and L^1 convergence, backwards martingales and Kolmogorov's Strong Law of Large Numbers.

Examples and applications.

0.3 The Galton–Watson branching process

We begin with an example that illustrates some of the concepts that lie ahead.

In spite of earlier work by Bienaymé, the Galton–Watson branching process is attributed to the great polymath Sir Francis Galton and the Revd Henry Watson. Like many Victorians, Galton was worried about the demise of English family names. He posed a question in the Educational Times of 1873. He wrote

> The decay of the families of men who have occupied conspicuous positions in past times has been a subject of frequent remark, and has given rise to various conjectures. The instances are very numerous in which surnames that were once common have become scarce or wholly disappeared. The tendency is universal, and, in explanation of it, the conclusion has hastily been drawn that a rise in physical comfort and intellectual capacity is necessarily accompanied by a diminution in 'fertility'...

He went on to ask "What is the probability that a name dies out by the 'ordinary law of chances'?"

Watson sent a solution which they published jointly the following year. The first step was to distill the problem into a workable mathematical model; that model, formulated by Watson, is what we now call the Galton–Watson branching process. Let's state it formally:

Definition 0.1 (Galton–Watson branching process)**.** Let $(X_{n,r})_{n,r\geqslant 1}$ be an infinite array of independent identically distributed random variables, each with the same distribution as X, where

$$\mathbb{P}[X = k] = p_k, \qquad k = 0, 1, 2, \ldots$$

The sequence $(Z_n)_{n\geqslant 0}$ of random variables defined by

1. $Z_0 = 1$,
2. $Z_n = X_{n,1} + \cdots + X_{n,Z_{n-1}}$ for $n \geqslant 1$

is the *Galton–Watson branching process* (started from a single ancestor) with *offspring distribution* X.

In the original setting, the random variable Z_n models the number of male descendants of a single male ancestor after n generations.

In analyzing this process, key roles are played by the expectation $\mu = \mathbb{E}[X] = \sum_{k=0}^{\infty} kp_k$, which we shall assume to be finite, and by the *probability generating function* $f = f_X$ of X, defined by $f(\theta) = \mathbb{E}[\theta^X] = \sum_{k=0}^{\infty} p_k\theta^k$.

Claim 0.2. *Let $f_n(\theta) = \mathbb{E}[\theta^{Z_n}]$. Then f_n is the n-fold composition of f with itself (where by convention a 0-fold composition is the identity).*

'Proof'

We proceed by induction. First note that $f_0(\theta) = \theta$, so f_0 is the identity. Assume that $n \geqslant 1$ and $f_{n-1} = f \circ \cdots \circ f$ is the $(n-1)$-fold composition of f with itself. To compute f_n, first note that

$$\begin{aligned}
\mathbb{E}\left[\theta^{Z_n} \middle| Z_{n-1} = k\right] &= \mathbb{E}\left[\theta^{X_{n,1}+\cdots+X_{n,k}}\right] \\
&= \mathbb{E}\left[\theta^{X_{n,1}}\right] \cdots \mathbb{E}\left[\theta^{X_{n,k}}\right] \quad \text{(independence)} \\
&= f(\theta)^k,
\end{aligned}$$

(since each $X_{n,i}$ has the same distribution as X). Hence

$$\mathbb{E}\left[\theta^{Z_n} \middle| Z_{n-1}\right] = f(\theta)^{Z_{n-1}}. \tag{1}$$

This is our first example of a *conditional expectation*. Notice that the right hand side of (1) is a *random variable*. Now

$$\begin{aligned} f_n(\theta) = \mathbb{E}\left[\theta^{Z_n}\right] &= \mathbb{E}\left[\mathbb{E}\left[\theta^{Z_n}\middle|\, Z_{n-1}\right]\right] \\ &= \mathbb{E}\left[f(\theta)^{Z_{n-1}}\right] \\ &= f_{n-1}\left(f(\theta)\right), \end{aligned} \tag{2}$$

and the claim follows by induction. □

In (2) we have used what is called the *tower property* of conditional expectations. In this example you can make all this work with the Partition Theorem of Prelims (because the events $\{Z_n = k\}$ form a countable partition of the sample space). In the general theory that follows, we'll see how to replace the Partition Theorem when the sample space is more complicated, for example when considering continuous random variables.

Watson wanted to establish the *extinction probability* of the branching process, i.e., the probability that $Z_n = 0$ for some n.

Claim 0.3. *Let $q = \mathbb{P}[Z_n = 0$ for some $n]$. Then q is the smallest root in $[0,1]$ of the equation $\theta = f(\theta)$. In particular, assuming $p_1 = \mathbb{P}[X = 1] < 1$,*

- *if $\mu = \mathbb{E}[X] \leqslant 1$, then $q = 1$,*
- *if $\mu = \mathbb{E}[X] > 1$, then $q < 1$.*

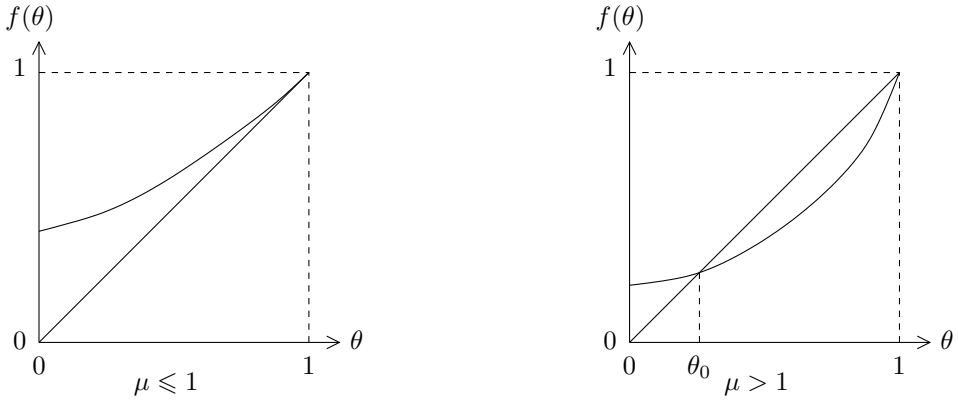
'Proof'

Let $q_n = \mathbb{P}[Z_n = 0] = f_n(0)$. Since $\{Z_n = 0\} \subseteq \{Z_{n+1} = 0\}$ we see that q_n is an increasing function of n and, intuitively,

$$q = \lim_{n\to\infty} q_n = \lim_{n\to\infty} f_n(0). \tag{3}$$

Since $f_{n+1}(0) = f(f_n(0))$ and f is continuous, (3) implies that q satisfies $q = f(q)$.

Now observe that f is convex (i.e., $f'' \geqslant 0$) and $f(1) = 1$, so only two things can happen, depending upon the value of $\mu = f'(1)$:



In the case $\mu > 1$, to see that q must be the *smaller* root θ_0, note that f is increasing, and $0 = q_0 \leqslant \theta_0$. It follows by induction that $q_n \leqslant \theta_0$ for all n, so $q \leqslant \theta_0$. □

It's not hard to guess the result above for $\mu > 1$ and $\mu < 1$, but the case $\mu = 1$ is far from obvious.

The extinction probability is only one statistic that we might care about. For example, we might ask whether we can say anything about the way in which the population grows or declines. Consider

$$\mathbb{E}\left[Z_{n+1} \mid Z_n = k\right] = \mathbb{E}\left[X_{n+1,1} + \cdots + X_{n+1,k}\right] = k\mu \quad \text{(linearity of expectation)}. \tag{4}$$

In other words $\mathbb{E}[Z_{n+1} \mid Z_n] = \mu Z_n$ (another conditional expectation). Now write

$$M_n = \frac{Z_n}{\mu^n}.$$

Then

$$\mathbb{E}\left[M_{n+1} \mid M_n\right] = M_n.$$

In fact, more is true:

$$\mathbb{E}\left[M_{n+1} \mid M_0, M_1, \ldots, M_n\right] = M_n.$$

A process $(M_n)_{n\geqslant 0}$ with this property is called a *martingale*.

It is natural to ask whether M_n has a limit as $n \to \infty$ and, if so, can we say anything about that limit? We're going to develop the tools to answer these questions, but for now, notice that for $\mu \leqslant 1$ we have 'proved' that $M_\infty = \lim_{n\to\infty} M_n = 0$ with probability one, so

$$0 = \mathbb{E}[M_\infty] \neq \lim_{n\to\infty} \mathbb{E}[M_n] = 1. \tag{5}$$

We're going to have to be careful in passing to limits, just as we discovered in Part A Integration. Indeed (5) may remind you of Fatou's Lemma from Part A.

One of the main aims of this course is to provide the tools needed to make arguments such as that presented above precise. Other key aims are to make sense of, and study, martingales in more general contexts. This involves defining conditional expectation when conditioning on a continuous random variable.

1 Measure spaces

We begin by recalling some definitions that you encountered in Part A Integration (and, although they were not emphasized there, in Prelims Probability). The idea is that we want to be able to assign a 'mass' or 'size' to subsets of a space in a consistent way. In particular, for us these subsets will be 'events' or 'collections of outcomes' (subsets of a probability sample space Ω) and the 'mass' will be a probability (a measure of how likely that event is to occur).

Recall that $\mathcal{P}(\Omega)$ denotes the *power set* of Ω, i.e., the set of all subsets of Ω.

Definition 1.1 (Algebras and σ-algebras)**.** Let Ω be a set and let $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ be a collection of subsets of Ω.

1. We say that $\mathcal{A}$ is an *algebra (on Ω)* if $\emptyset \in \mathcal{A}$ and for all $A, B \in \mathcal{A}$, $A^{\mathrm{c}} = \Omega \setminus A \in \mathcal{A}$ and $A \cup B \in \mathcal{A}$.
2. We say that $\mathcal{A}$ is a *σ-algebra (on Ω)* if $\emptyset \in \mathcal{A}$, $A \in \mathcal{A}$ implies $A^{\mathrm{c}} \in \mathcal{A}$, and for all sequences $(A_n)_{n\geqslant 1}$ of elements of $\mathcal{A}$, $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.

Since intersections can be built up from complements and unions, an algebra is closed under *finite* set operations; a σ-algebra is closed under *countable* set operations. Often we don't bother saying 'on Ω', but note that A^{c} makes sense only if we know which set Ω we are talking about. We tend to write $\mathcal{F}$ for a σ-algebra (also called a *σ-field* by some people).

Definition 1.2 (Set functions)**.** Let $\mathcal{A}$ be *any* set of subsets of Ω containing the empty set $\emptyset$. A *set function* on $\mathcal{A}$ is a function $\mu : \mathcal{A} \to [0, \infty]$ with $\mu(\emptyset) = 0$. We say that μ is

1. *increasing* if for all $A, B \in \mathcal{A}$ with $A \subseteq B$,

$$\mu(A) \leqslant \mu(B),$$

2. *additive* if for all *disjoint* $A, B \in \mathcal{A}$ with $A \cup B \in \mathcal{A}$ (note that we must specify this in general)

$$\mu(A \cup B) = \mu(A) + \mu(B),$$

3. *countably additive*, or *σ-additive*, if for all sequences (A_n) of disjoint sets in $\mathcal{A}$ with $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$

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

Definition 1.3 (Measure space)**.** A *measurable space* is a pair $(\Omega, \mathcal{F})$ where $\mathcal{F}$ is a σ-algebra on Ω. A *measure space* is a triple $(\Omega, \mathcal{F}, \mu)$ where Ω is a set, $\mathcal{F}$ is a σ-algebra on Ω and $\mu : \mathcal{F} \to [0, \infty]$ is a countably additive set function. Then μ is a *measure* on $(\Omega, \mathcal{F})$.

In short, a measure space is a set Ω equipped with a σ-algebra $\mathcal{F}$ and a countably additive set function μ on $\mathcal{F}$. Note that any measure μ is also additive and increasing.

Definition 1.4 (Types of measure space)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.

1. We say that μ is *finite* if $\mu(\Omega) < \infty$.
2. If there is a sequence $(E_n)_{n \geqslant 1}$ of sets from $\mathcal{F}$ with $\mu(E_n) < \infty$ for all n and $\bigcup_{n=1}^{\infty} E_n = \Omega$, then μ is said to be *σ-finite*.
3. In the special case when $\mu(\Omega) = 1$, we say that μ is a *probability measure* and $(\Omega, \mathcal{F}, \mu)$ is a *probability space*; we often use the notation $(\Omega, \mathcal{F}, \mathbb{P})$ to emphasize this.

Recall from Part A Integration that measures also respect monotone limits.

Notation: For a sequence $(F_n)_{n \geqslant 1}$ of sets, $F_n \uparrow F$ means $F_n \subseteq F_{n+1}$ for all n and $\bigcup_{n=1}^{\infty} F_n = F$. Similarly, $G_n \downarrow G$ means $G_n \supseteq G_{n+1}$ for all n and $\bigcap_{n=1}^{\infty} G_n = G$.

Lemma 1.5 (Monotone convergence properties)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.*

1. *If $(F_n)_{n \geqslant 1}$ is a sequence of sets from $\mathcal{F}$ with $F_n \uparrow F$, then $\mu(F_n) \uparrow \mu(F)$ as $n \to \infty$,*
2. *If $(G_n)_{n \geqslant 1}$ is a sequence of sets from $\mathcal{F}$ with $G_n \downarrow G$, and $\mu(G_k) < \infty$ for some $k \in \mathbb{N}$, then $\mu(G_n) \downarrow \mu(G)$ as $n \to \infty$.*

Proof. See Part A Integration (or Exercise). □

Note that $\mu(G_k) < \infty$ is essential in (ii): for example take $G_n = (n, \infty) \subseteq \mathbb{R}$ and Lebesgue measure. The following partial converse is sometimes useful.

Lemma 1.6. *Let $\mu : \mathcal{A} \to [0, \infty)$ be an additive set function on an algebra $\mathcal{A}$ taking only finite values. Then μ is countably additive iff for every sequence (A_n) of sets in $\mathcal{A}$ with $A_n \downarrow \emptyset$ we have $\mu(A_n) \to 0$.*

Proof. One implication follows (essentially) from Lemma 1.5; the other is an exercise. □

There are lots of measure spaces out there, several of which you are already familiar with.

Example 1.7 (Discrete measure theory)**.** Let Ω be a countable set. A *mass function* on Ω is any function $\bar{\mu} : \Omega \to [0, \infty]$. Given such a $\bar{\mu}$ we can define a measure on $(\Omega, \mathcal{P}(\Omega))$ by setting $\mu(A) = \sum_{x \in A} \bar{\mu}(x)$.

Equally, given a measure μ on $(\Omega, \mathcal{P}(\Omega))$ we can define a corresponding mass function by $\bar{\mu}(x) = \mu(\{x\})$. For countable Ω there is a one-to-one correspondence between measures on $(\Omega, \mathcal{P}(\Omega))$ and mass functions.

These discrete measure spaces provide a 'toy' version of the general theory, but in general they are not enough. Discrete measure theory is essentially the only context in which one can define the measure explicitly. This is because σ-algebras are not in general amenable to an explicit presentation, and it is *not* in general the case that for an arbitrary set Ω all subsets of Ω can be assigned a measure – recall from Part A Integration the construction of a non-Lebesgue measurable subset of $\mathbb{R}$. Instead one shows the existence of a measure defined on a 'large enough' collection of sets, with the properties we want. To do this, we follow a variant of the approach you saw in Part A; the idea is to specify the values to be taken by the measure on a smaller class of subsets of Ω that 'generate' the σ-algebra (as the singletons did in Example 1.7). This leads to two problems. First we need to know that it is possible to extend the measure that we specify to the whole σ-algebra. This *construction* problem is often handled with *Carathéodory's Extension Theorem* (Theorem 1.13 below). The second problem is to know that there is only *one* measure on the σ-algebra that is consistent with our specification. This *uniqueness* problem can often be resolved through a corollary of Dynkin's π-system Lemma that we state below. First we need some more definitions.

Definition 1.8 (Generated σ-algebras)**.** Let $\mathcal{A}$ be a collection of subsets of Ω. Define

$$\sigma(\mathcal{A}) = \{A \subseteq \Omega : A \in \mathcal{F} \text{ for all } \sigma\text{-algebras } \mathcal{F} \text{ on } \Omega \text{ containing } \mathcal{A}\}.$$

Then $\sigma(\mathcal{A})$ is a σ-algebra (exercise) which is called *the σ-algebra generated by $\mathcal{A}$*. It is the smallest σ-algebra containing $\mathcal{A}$: if $\mathcal{F} \supseteq \mathcal{A}$ is a σ-algebra then $\mathcal{F} \supseteq \sigma(\mathcal{A})$.

Definition 1.9 (Borel σ-algebra, Borel measure)**.** Let Ω be a topological space with topology (i.e., set of open sets) $\mathcal{T}$. Then *the Borel σ-algebra on Ω* is the σ-algebra generated by the open sets:

$$\mathcal{B}(\Omega) = \sigma(\mathcal{T}).$$

A measure μ on $(\Omega, \mathcal{B}(\Omega))$ is called a *Borel measure* on Ω.

Note that $\mathcal{B}(\Omega)$ depends not just on the set Ω, but also on the topology on Ω. Usually, this is understood: in particular, when $\Omega = \mathbb{R}$, we mean the usual Euclidean topology on $\mathbb{R}$.

Definition 1.10 (π-system)**.** Let $\mathcal{I}$ be a collection of subsets of Ω. We say that $\mathcal{I}$ is a *π-system* if $A, B \in \mathcal{I}$ implies $A \cap B \in \mathcal{I}$.

Notice that an algebra is automatically a π-system.

Example 1.11. The collection

$$\pi(\mathbb{R}) = \{(-\infty, x] : x \in \mathbb{R}\}$$

forms a π-system and $\sigma(\pi(\mathbb{R}))$, the σ-algebra generated by $\pi(\mathbb{R})$, is $\mathcal{B}(\mathbb{R})$, the σ-algebra consisting of all Borel subsets of $\mathbb{R}$ (exercise).

Here's why we care about π-systems.

Theorem 1.12 (Uniqueness of extension)**.** *Let μ_1 and μ_2 be measures on the same measurable space $(\Omega, \mathcal{F})$, and let $\mathcal{I} \subseteq \mathcal{F}$ be a π-system. If $\mu_1(\Omega) = \mu_2(\Omega) < \infty$ and $\mu_1 = \mu_2$ on $\mathcal{I}$, then $\mu_1 = \mu_2$ on $\sigma(\mathcal{I})$.*

We will often apply the theorem to a π-system $\mathcal{I}$ with $\sigma(\mathcal{I}) = \mathcal{F}$, so the conclusion is that μ_1 and μ_2 agree. A very important special case is that if two *probability* measures on Ω agree on a π-system, then they agree on the σ-algebra generated by that π-system.

For a proof of Theorem 1.12 see (e.g.) Williams, Appendix A.1.

That deals with uniqueness, but what about existence?

Theorem 1.13 (Carathéodory Extension Theorem)**.** *Let Ω be a set and $\mathcal{A}$ an algebra on Ω, and let $\mathcal{F} = \sigma(\mathcal{A})$. Let $\mu_0 : \mathcal{A} \to [0, \infty]$ be a countably additive set function. Then there exists a measure μ on $(\Omega, \mathcal{F})$ such that $\mu = \mu_0$ on $\mathcal{A}$.*

Remark. If $\mu_0(\Omega) < \infty$, then Theorem 1.12 tells us that μ is unique, since an algebra is certainly a π-system.

The Carathéodory Extension Theorem doesn't quite solve the problem of constructing measures on σ-algebras – it reduces it to constructing countably additive set functions on algebras; we shall see several examples.

The proof of the Carathéodory Extension Theorem is not examinable. Here are some of the ideas; this is much the same as the proof of the existence of Lebesgue measure in Part A Integration (which was also non-examinable). First one defines the *outer measure* $\mu^*(B)$ of *any* $B \subseteq \Omega$ by

$$\mu^*(B) = \inf\Big\{\sum_{j=1}^{\infty} \mu_0(A_j) : A_j \in \mathcal{A},\ \bigcup_{j=1}^{\infty} A_j \supseteq B\Big\}.$$

Then define a set B to be *measurable* if for all sets E,

$$\mu^*(E) = \mu^*(E \cap B) + \mu^*(E \cap B^{\mathrm{c}}).$$

[Alternatively, if $\mu_0(\Omega)$ is finite, then one can define B to be measurable if $\mu^*(B) + \mu^*(B^{\mathrm{c}}) = \mu_0(\Omega)$; this more intuitive definition expresses that it is possible to cover B and B^{c} 'efficiently' with sets from $\mathcal{A}$.] One must check that μ^* defines a countably additive set function on the collection of measurable sets extending μ_0, and that the measurable sets form a σ-algebra that contains $\mathcal{A}$. For details see Appendix A.1 of Williams, or Varadhan and the references therein.

Corollary 1.14. *There exists a unique Borel measure μ on $\mathbb{R}$ such that for all $a, b \in \mathbb{R}$ with $a < b$, $\mu((a, b]) = b - a$. The measure μ is the Lebesgue measure on $\mathcal{B}(\mathbb{R})$.*

The proof of Corollary 1.14 is an exercise. (The hard part is checking countable additivity on a suitable algebra; we will do a related example in a moment. Note that an extra step is required for uniqueness since $\mu(\Omega) = \infty$.)

Remark. In Part A Integration, the Lebesgue measure was defined on a σ-algebra $\mathcal{M}_{\mathrm{Leb}}$ that contains, but is strictly larger than, $\mathcal{B}(\mathbb{R})$. It turns out (exercise) that $\mathcal{M}_{\mathrm{Leb}}$ consists of all sets that differ from a Borel set on a null set. In this course we shall work with $\mathcal{B}(\mathbb{R})$ rather than $\mathcal{M}_{\mathrm{Leb}}$: the Borel σ-algebra will be 'large enough' for us. (This changes later when studying continuous-time martingales.) An advantage $\mathcal{B}(\mathbb{R})$ is that it has a simple definition independent of the measure; recall that which sets are null depends on which measure is being considered.

Recall that in our 'toy example' of discrete measure theory there was a one-to-one correspondence between measures and mass functions. Can we say anything similar for Borel measures on $\mathbb{R}$? (I.e., measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$?)

Definition 1.15. Let μ be a Borel probability measure on $\mathbb{R}$. The *distribution function* of μ is the function $F : \mathbb{R} \to \mathbb{R}$ defined by $F(x) = \mu((-\infty, x])$.

Any distribution function F has the following properties:

1. F is (weakly) increasing, i.e., $x < y$ implies $F(x) \leqslant F(y)$,

2. $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$, and

3. F is *right continuous*: $y \downarrow x$ implies $F(y) \to F(x)$.

To see the last, suppose that $y_n \downarrow x$ and let $A_n = (-\infty, y_n]$. Then $A_n \downarrow A = (-\infty, x]$. Thus, by Lemma 1.5, $F(y_n) = \mu(A_n) \downarrow \mu(A) = F(x)$. We often write $F(-\infty) = 0$ and $F(\infty) = 1$ as shorthand for the second property.

Using the Carathéodory Extension Theorem, we can construct *all* Borel probability measures on $\mathbb{R}$ (i.e., probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$): there is one for each distribution function. Since finite measures can all be obtained from probability measures (by multiplying by a constant), this characterizes *all* finite measures on $\mathcal{B}(\mathbb{R})$.

Theorem 1.16 (Lebesgue)**.** *Let $F : \mathbb{R} \to \mathbb{R}$ be an increasing, right continuous function with $F(-\infty) = 0$ and $F(\infty) = 1$. Then there is a unique Borel probability measure $\mu = \mu_F$ on $\mathbb{R}$ such that $\mu((-\infty, x]) = F(x)$ for every x. Every Borel probability measure μ on $\mathbb{R}$ arises in this way.*

In other words, there is a 1-1 correspondence between distribution functions and Borel probability measures on $\mathbb{R}$.

Proof. Suppose for the moment that the existence statement holds. Since $\pi(\mathbb{R}) = \{(-\infty, x] : x \in \mathbb{R}\}$ is a π-system which generates the σ-algebra $\mathcal{B}(\mathbb{R})$, uniqueness follows by Theorem 1.12. Also, to see the final part, let μ be any Borel probability measure on $\mathbb{R}$, and let F be its distribution function. Then F has the properties required for the first part of the theorem, and we obtain a measure μ_F which by uniqueness is the measure μ we started with.

For existence we shall apply Theorem 1.13, so first we need a suitable algebra. For $-\infty \leqslant a \leqslant b < \infty$, let $I_{a,b} = (a, b]$, and set $I_{a,\infty} = (a, \infty)$. So $I_{a,b} = \{x \in \mathbb{R} : a < x \leqslant b\}$. Let $\mathcal{I} = \{I_{a,b} : -\infty \leqslant a \leqslant b \leqslant \infty\}$ be the collection of intervals that are open on the left and closed on the right. Let $\mathcal{A}$ be the set of finite disjoint unions of elements of $\mathcal{I}$; then $\mathcal{A}$ is an algebra, and $\sigma(\mathcal{A}) = \sigma(\mathcal{I}) = \mathcal{B}(\mathbb{R})$.

We can define a set function μ_0 on $\mathcal{A}$ by setting

$$\mu_0(I_{a,b}) = F(b) - F(a)$$

for intervals and then extending it to $\mathcal{A}$ by defining it as the sum for disjoint unions from $\mathcal{I}$. It is an easy exercise to show that μ_0 is well defined and *finitely* additive. Carathéodory's Extension Theorem tells us that μ_0 extends to a probability measure on $\mathcal{B}(\mathbb{R})$ *provided* that μ_0 is *countably* additive on $\mathcal{A}$. Proving this is slightly tricky. Note that we will have to use right continuity at some point.

First note that by Lemma 1.6, since μ_0 is finite and additive on $\mathcal{A}$, it is *countably* additive if and only if, for any sequence (A_n) of sets from $\mathcal{A}$ with $A_n \downarrow \emptyset$, $\mu_0(A_n) \downarrow 0$.

Suppose that F has the stated properties but, for a contradiction, that there exist $A_1, A_2, \ldots \in \mathcal{A}$ with $A_n \downarrow \emptyset$ but $\mu_0(A_n) \not\to 0$. Since $\mu_0(A_n)$ is a decreasing sequence, there is some $\delta > 0$ (namely, $\lim \mu_0(A_n)$) such that $\mu_0(A_n) \geqslant \delta$ for all n. We look for a descending sequence of *compact* sets; since if all the sets in such a sequence are non-empty, so is their intersection.

Step 1: Replace A_n by $B_n = A_n \cap (-l, l]$. Since

$$\mu_0(A_n \setminus B_n) \leqslant \mu_0\big((-\infty, l] \cup (l, \infty)\big) = F(-l) + 1 - F(l),$$

if we take l large enough then we have $\mu_0(B_n) \geqslant \delta/2$ for all n.

Step 2: Suppose that $B_n = \bigcup_{i=1}^{k_n} I_{a_{n,i}, b_{n,i}}$. Let $C_n = \bigcup_{i=1}^{k_n} I_{\tilde{a}_{n,i}, b_{n,i}}$ where $a_{n,i} < \tilde{a}_{n,i} < b_{n,i}$ and we use right continuity of F to do this in such a way that

$$\mu_0(B_n \setminus C_n) < \frac{\delta}{2^{n+2}} \quad \text{for each } n.$$

Let $\overline{C}_n$ be the closure of C_n (obtained by adding the points $\tilde{a}_{n,i}$ to C_n).

Step 3: The sequence (C_n) need not be decreasing, so set $D_n = \bigcap_{i=1}^n C_i$, and $E_n = \bigcap_{i=1}^n \overline{C}_i$. Since

$$\mu_0(D_n) \geqslant \mu_0(B_n) - \sum_{i=1}^n \mu_0(B_i \backslash C_i) \geqslant \frac{\delta}{2} - \sum_{i=1}^n \frac{\delta}{2^{i+2}} \geqslant \frac{\delta}{4},$$

D_n is non-empty. Thus $E_n \supseteq D_n$ is non-empty.

Each E_n is closed and bounded, and so compact. Also, each E_n is non-empty, and $E_n \subseteq E_{n+1}$. Hence, by a basic result from topology, there is some x such that $x \in E_n$ for all n. Since $E_n \subseteq \overline{C}_n \subseteq B_n \subseteq A_n$, we have $x \in A_n$ for all n, contradicting $A_n \downarrow \emptyset$. □

The function $F(x)$ is the *distribution function* corresponding to the probability measure μ. In the case when F is continuously differentiable, say, it is precisely the cumulative distribution function of a continuous random variable with probability density function $f(x) = F'(x)$ that we encountered in Prelims.

More generally, if $f(x) \geqslant 0$ is measurable and (Lebesgue) integrable with $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$, then we can use f as a density function to construct a measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by setting

$$\mu(A) = \int_A f(x)\,\mathrm{d}x.$$

This measure has distribution function $F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y$. (It is not necessarily true that $F'(x) = f(x)$ for all x, but this will hold for almost all x.) For example, taking $f(x) = 1$ on $(0,1)$, or on $[0,1]$, and $f(x) = 0$ otherwise, we obtain the distribution function F with $F(x) = 0$, $x < 0$, $F(x) = x$, $0 \leqslant x \leqslant 1$ and $F(x) = 1$ for $x > 1$, corresponding to the uniform distribution on $[0,1]$.

For a very different example, if $x_1, x_2, \ldots$ is a sequence of points (for example the non-negative integers), and we have probabilities $p_n > 0$ at these points with $\sum_n p_n = 1$, then for the discrete probability measure

$$\mu(A) = \sum_{n\,:\,x_n \in A} p_n,$$

we have the distribution function

$$F(x) = \sum_{n\,:\,x_n \leqslant x} p_n,$$

which increases by jumps, the jump at x_n being of height p_n. (The picture can be complicated though, for example if there is a jump at every rational.)

There are examples of continuous distribution functions F that don't come from any density f, e.g., the Devil's staircase, corresponding (roughly speaking) to the uniform distribution on the Cantor set.

The measures μ we have just described are sometimes called *Lebesgue–Stieltjes measures*. We'll return to them a little later.

We now have a very rich class of measures to work with. In Part A Integration, you saw a theory of integration based on Lebesgue measure. It is natural to ask whether we can develop an analogous theory for other measures. The answer is 'yes', and in fact almost all the work was done in Part A; the proofs used there carry over to any measure. It is left as a (useful) exercise to check that. Here we just state the key definitions and results.

2 Integration

2.1 Measurable functions and the definition of the integral

Definition 2.1 (Measurable function)**.** Let $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ be measurable spaces. A function $f : \Omega \to \Lambda$ is *measurable* (with respect to $\mathcal{F}$, $\mathcal{G}$) if

$$A \in \mathcal{G} \implies f^{-1}(A) \in \mathcal{F}.$$

Usually $\Lambda = \mathbb{R}$ or $\overline{\mathbb{R}} = [-\infty, \infty]$. In this case we *always* take $\mathcal{G}$ to consist of the Borel sets: $\mathcal{G} = \mathcal{B}(\mathbb{R})$ or $\mathcal{B}(\overline{\mathbb{R}})$, and omit it from the notation. This contrasts with mapping from $\mathbb{R}$, where different σ-algebras are considered in different circumstances (sometimes including $\mathcal{M}_{\text{Leb}}$, though not in this course).

Proposition 2.2. *A function $f : \Omega \to \mathbb{R}$ or $f : \Omega \to \overline{\mathbb{R}}$ is measurable with respect to $\mathcal{F}$ (and $\mathcal{B}(\mathbb{R})$ or $\mathcal{B}(\overline{\mathbb{R}})$) if and only if $\{x : f(x) \leqslant t\} \in \mathcal{F}$ for every $t \in \mathbb{R}$.*

Proof. For $f : \Omega \to \mathbb{R}$ this was proved in Integration; the key points are that $\{A \subseteq \mathbb{R} : f^{-1}(A) \in \mathcal{F}\}$ is a σ-algebra, and that $\mathcal{B}(\mathbb{R})$ is generated by $\{(-\infty, t] : t \in \mathbb{R}\}$. The proof for $f : \Omega \to \overline{\mathbb{R}}$ is the same: $\mathcal{B}(\overline{\mathbb{R}})$ is generated by $\{[-\infty, t] : t \in \mathbb{R}\} \subset \mathcal{P}(\overline{\mathbb{R}})$. □

Unless otherwise stated, measurable functions map to $\overline{\mathbb{R}}$ with the Borel σ-algebra. Thus a measurable function on $(\Omega, \mathcal{F})$ means a function $\Omega \to \overline{\mathbb{R}}$ that is $(\mathcal{F}, \mathcal{B}(\overline{\mathbb{R}}))$-measurable.

Remark. It is worth bearing in mind that (real-valued) functions on Ω generalise subsets of Ω in a natural way, with the function 1_A corresponding to the subset A. As a sanity check, note that 1_A is a measurable function if and only if A is a measurable set, i.e., $A \in \mathcal{F}$.

Recall that

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty} \sup_{m \geqslant n} x_m \quad \text{and} \quad \liminf_{n\to\infty} x_n = \lim_{n\to\infty} \inf_{m \geqslant n} x_m.$$

The following result was proved in Part A (in some cases only for functions taking finite values, but the extension is no problem).

Lemma 2.3. *Let (f_n) be a sequence of measurable functions on $(\Omega, \mathcal{F})$ taking values in $\overline{\mathbb{R}}$, and let $h : \mathbb{R} \to \mathbb{R}$ be continuous. Then, whenever they make sense*[1]*, the following are also measurable functions on $(\Omega, \mathcal{F})$:*

$$f_1 + f_2, \quad f_1 f_2, \quad \max\{f_1, f_2\}, \quad \min\{f_1, f_2\}, \quad f_1/f_2, \quad h \circ f$$

$$\sup_n f_n, \quad \inf_n f_n, \quad \limsup_{n\to\infty} f_n, \quad \liminf_{n\to\infty} f_n.$$

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Given a measurable function $f : \Omega \to \overline{\mathbb{R}}$, we want to define, where possible, the integral of f with respect to μ. There are many variants of the notation, such as:

$$\int f \, \mathrm{d}\mu = \int_\Omega f \, \mathrm{d}\mu = \mu(f) = \int_{x\in\Omega} f(x) \, \mathrm{d}\mu(x) = \int f(x) \mu(\mathrm{d}x)$$

and so on. The dummy variable (here x) is sometimes needed when, for example, we have a function $f(x, y)$ of two variables, and with y fixed are integrating the function $f(\cdot, y)$ given by $x \mapsto f(x, y)$.

[1] For example, $\infty - \infty$ is not defined.

Definition 2.4. A *simple function* ϕ on a measure space $(\Omega, \mathcal{F}, \mu)$ is a function $\phi : \Omega \to \mathbb{R}$ that may be written as a finite sum

$$\phi = \sum_{k=1}^{n} a_k \mathbf{1}_{E_k} \tag{6}$$

where each $E_k \in \mathcal{F}$ and each $a_k \in \mathbb{R}$. The *canonical form* of ϕ is the unique decomposition as in (6) where the numbers a_k are distinct and non-zero and the sets E_k are disjoint and non-empty.

The following definitions and results were given in Part A Integration in the special case of Lebesgue measure. But they extend with no change to a general measure space $(\Omega, \mathcal{F}, \mu)$.

Definition 2.5. If ϕ is a non-negative simple function with canonical form (6), then we define the integral of ϕ with respect to μ as

$$\int \phi \, \mathrm{d}\mu = \sum_{k=1}^{n} a_k \mu(E_k).$$

This formula then also applies (exercise) whenever ϕ is as in (6), even if this is not the canonical form, as long as we avoid $\infty - \infty$ (for example by taking $a_k \geqslant 0$).

Definition 2.6. For a non-negative measurable function f on $(\Omega, \mathcal{F}, \mu)$ we define the integral

$$\int f \, \mathrm{d}\mu = \sup\left\{ \int \phi \, \mathrm{d}\mu \ : \ \phi \text{ simple}, \ 0 \leqslant \phi \leqslant f \right\}.$$

Note that the supremum may be equal to $+\infty$.

Definition 2.7. We say that a measurable function f on $(\Omega, \mathcal{F}, \mu)$ is *integrable* if $\int |f| \, \mathrm{d}\mu < \infty$. If f is integrable, its integral is defined to be

$$\int f \, \mathrm{d}\mu = \int f^+ \, \mathrm{d}\mu - \int f^- \, \mathrm{d}\mu,$$

where $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$ are the positive and negative parts of f.

Note that $f = f^+ - f^-$. A very important point is that if f is measurable, then $\int f \, \mathrm{d}\mu$ is defined either if f is non-negative (when ∞ is a possible value) or if f is integrable.

There are other possible sequences of steps to defining the integral, giving the same result. This generalized integral has the same basic properties as in the special case of Lebesgue measure, with the same proofs. For example, if f and g are measurable functions on $(\Omega, \mathcal{F}, \mu)$ that are either both non-negative or both integrable, and $c \in \mathbb{R}$, then

$$\int (f + g) \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu + \int g \, \mathrm{d}\mu, \qquad \int cf \, \mathrm{d}\mu = c \int f \, \mathrm{d}\mu.$$

We have defined integrals only over the whole space. This is all we need – if f is a measurable function on $(\Omega, \mathcal{F}, \mu)$ and $A \in \mathcal{F}$ then we define

$$\int_A f \, \mathrm{d}\mu = \int f 1_A \, \mathrm{d}\mu,$$

i.e., we integrate (over the whole space) the function that agrees with f on A and is 0 outside A.

If μ is the Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then we have just redefined the Lebesgue integral as in Part A. For a very different example, suppose that μ is a discrete measure with mass p_i at point x_i, for a (finite or countably infinite) sequence $x_1, x_2, \ldots$. Then you can check that

$$\int f \, \mathrm{d}\mu = \sum_i f(x_i) p_i,$$

whenever $f \geqslant 0$ (where $+\infty$ is allowed as the answer) or the sum converges absolutely. For another example, suppose that μ has distribution function $F(x) = \int_{-\infty}^{x} g(y) \, \mathrm{d}y$. Then

$$\int f \, \mathrm{d}\mu = \int f(x) g(x) \, \mathrm{d}x,$$

where the second integral is with respect to Lebesgue measure. In proving statements like this it is often helpful to start by considering the case $f = 1_E$, then simple functions f, then non-negative measurable f, and finally general measurable f. It also helps to recall that given any measurable $f \geqslant 0$ there are simple functions $f_n \geqslant 0$ with $f_n \uparrow f$.

Remark 2.8. One final property of integration that is easy to check (exercise) from the definitions is that for $f \geqslant 0$, $\int f \, \mathrm{d}\mu$ is determined by the numbers $\mu(\{x : f(x) \geqslant t\})$ for each $t \geqslant 0$. Hence, for general f, $\int f \, \mathrm{d}\mu$ is determined by the numbers $\mu(\{x : f(x) \geqslant t\})$ for $t \geqslant 0$ and $\mu(\{x : f(x) \leqslant t\})$ for $t \leqslant 0$. Since $\{x : f(x) \geqslant t\}$ is the complement of the union of the sets $\{x : f(x) \leqslant s\}$, $s < t$, on a probability space, say, $\int f \, \mathrm{d}\mu$ is determined by the numbers $\mu(\{x : f(x) \leqslant t\})$ for $t \in \mathbb{R}$. This holds even across probability spaces: if f_i is a measurable function on the probability space $(\Omega_i, \mathcal{F}_i, \mu_i)$ and for every $t \in \mathbb{R}$, $\mu_1(\{x \in \Omega_1 : f_1(x) \leqslant t\}) = \mu_2(\{x \in \Omega_2 : f_2(x) \leqslant t\})$, then $\int f_1 \, \mathrm{d}\mu_1 = \int f_2 \, \mathrm{d}\mu_2$, or both are undefined.

Definition 2.9 (μ-almost everywhere)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. We say that a property holds *μ-almost everywhere* or *μ-a.e.* if it holds except on a set of μ-measure zero. If μ is a probability measure, we often say *almost surely* or *a.s.* instead of almost everywhere. Thus an event A holds almost surely if $\mathbb{P}[A] = 1$. This does not imply that $A = \Omega$.

An important property of integration is that

$$f = g \ \ \mu\text{-almost everywhere} \implies \int f \, \mathrm{d}\mu = \int g \, \mathrm{d}\mu.$$

Generally speaking, we don't care what happens on sets of measure zero. It is vital to remember that notions of almost everywhere depend on the underlying measure μ.

The measurable functions that are going to interest us most in what follows are random variables.

Definition 2.10 (Random Variable)**.** In the special case when $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, we call a measurable function $X : \Omega \to \mathbb{R}$ a (real-valued) *random variable*.

Sometimes we consider $X : \Omega \to \overline{\mathbb{R}}$ instead.

As we already did in Prelims, we can think of Ω as the sample space of an experiment, and the random variable X as an observable, i.e. something that can be measured. What is the integral of X?

Definition 2.11 (Expectation)**.** The *expectation* of a random variable X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ is

$$\mathbb{E}[X] = \int X \, \mathrm{d}\mathbb{P} = \int_\Omega X(\omega) \, \mathrm{d}\mathbb{P}(\omega).$$

A random variable X induces a probability measure μ_X on $\mathbb{R}$ via

$$\mu_X(A) = \mathbb{P}[X^{-1}(A)] \quad \text{for } A \in \mathcal{B}(\mathbb{R}).$$

In particular, $F_X(x) = \mu_X((-\infty, x])$ defines the *distribution function* of X (c.f. Theorem 1.16). Since $\{(-\infty, x] : x \in \mathbb{R}\}$ is a π-system, we see that the distribution function uniquely determines μ_X. From Remark 2.8,

$$\mathbb{E}[X] = \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega) = \int_{\mathbb{R}} x\,\mathrm{d}\mu_X(x).$$

Very often in applications we suppress the sample space Ω and work directly with μ_X.

2.2 The Convergence Theorems

The following theorems were proved in Part A for Lebesgue integral. Again the proofs carry over to the more general integral defined here.

Theorem 2.12 (Fatou's Lemma)**.** *Let (f_n) be a sequence of non-negative measurable functions on $(\Omega, \mathcal{F}, \mu)$. Then*

$$\int \liminf_{n\to\infty} f_n\,\mathrm{d}\mu \leqslant \liminf_{n\to\infty} \int f_n\,\mathrm{d}\mu.$$

Theorem 2.13 (Monotone Convergence Theorem)**.** *Let (f_n) be a sequence of non-negative measurable functions on $(\Omega, \mathcal{F}, \mu)$. Then*

$$f_n \uparrow f \implies \int f_n\,\mathrm{d}\mu \uparrow \int f\,\mathrm{d}\mu.$$

Note that we are not excluding $\int f\,\mathrm{d}\mu = \infty$ here. Also, we could just as well write $f_n \uparrow f$ μ-almost everywhere.

Equivalently, considering partial sums, the Monotone Convergence Theorem says that if (f_n) is a sequence of non-negative measurable functions, then

$$\int \sum_{n=1}^{\infty} f_n\,\mathrm{d}\mu = \sum_{n=1}^{\infty} \int f_n\,\mathrm{d}\mu.$$

Recall that (f_n) converges *pointwise* to f if, for every $x \in \Omega$, we have $f_n(x) \to f(x)$ as $n \to \infty$.

Theorem 2.14 (Dominated Convergence Theorem)**.** *Let (f_n) be a sequence of measurable functions on $(\Omega, \mathcal{F}, \mu)$ with $f_n \to f$ pointwise. Suppose that for some* **integrable** *function g, $|f_n| \leqslant g$ for all n. Then f is integrable and*

$$\int f_n\,\mathrm{d}\mu \to \int f\,\mathrm{d}\mu \quad \text{as } n \to \infty.$$

Again, convergence almost everywhere is enough.

We will also use the following less standard result.

Lemma 2.15 (Reverse Fatou Lemma)**.** *Let (f_n) be a sequence of measurable functions. Assume that there exists an* **integrable** *function g such that $f_n \leqslant g$ for all n. Then*

$$\int \limsup_{n\to\infty} f_n\,\mathrm{d}\mu \geqslant \limsup_{n\to\infty} \int f_n\,\mathrm{d}\mu.$$

Proof. Apply Fatou to $h_n = g - f_n$. (Note that $\int g\,\mathrm{d}\mu < \infty$ is *needed*.) □

2.3 Product Spaces and Independence

Definition 2.16 (Product σ-algebras)**.** Given two sets Ω_1 and Ω_2, the *Cartesian product* $\Omega = \Omega_1 \times \Omega_2$ is the set of pairs (ω_1, ω_2) with $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$.

If $\mathcal{F}_i$ is a σ-algebra on Ω_i, then a *measurable rectangle* in $\Omega = \Omega_1 \times \Omega_2$ is a set of the form $A_1 \times A_2$ with $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$. The *product σ-algebra* $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$ is the σ-algebra on Ω generated by the set of all measurable rectangles. (Note that $\mathcal{F}$ is not the Cartesian product of $\mathcal{F}_1$ and $\mathcal{F}_2$.)

Given two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ on $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ respectively, we'd like to define a probability measure on $(\Omega, \mathcal{F})$ by setting

$$\mathbb{P}[A_1 \times A_2] = \mathbb{P}_1[A_1]\mathbb{P}_2[A_2] \tag{7}$$

for each measurable rectangle and extending it to the whole of $\mathcal{F}$. Note that the set $\mathcal{I}$ of measurable rectangles is a π-system with (by definition) $\sigma(\mathcal{I}) = \mathcal{F}$, so if such a probability measure on $(\Omega, \mathcal{F})$ exists, it is unique by Theorem 1.12.

First, we extend $\mathbb{P}$ to the algebra $\mathcal{A}$ consisting of all finite disjoint unions of measurable rectangles by setting

$$\mathbb{P}[R_1 \cup \cdots \cup R_n] = \sum_{i=1}^{n} \mathbb{P}[R_i] \tag{8}$$

when $R_1, \ldots, R_n \in \mathcal{I}$ are disjoint. It is a tedious, but straightforward, exercise to check that this is well-defined. (This also follows from the (proof of) the next lemma.)

To check that we can extend $\mathbb{P}$ to the whole of $\mathcal{F} = \sigma(\mathcal{A})$, we need to check that $\mathbb{P}$ defined by (7) and (8) is actually *countably* additive on $\mathcal{A}$ so that we can apply Carathéodory's Extension Theorem.

Lemma 2.17. *The set function $\mathbb{P}$ defined on $\mathcal{A}$ through (7) and* (8) *is* countably *additive on $\mathcal{A}$.*

Proof. For any $A \in \mathcal{A}$ and $\omega_2 \in \Omega_2$, define the *section*

$$A_{\omega_2} = \{\omega_1 : (\omega_1, \omega_2) \in A\} \subseteq \Omega_1,$$

and let $f(\omega_2) = \mathbb{P}_1[A_{\omega_2}]$. Then f is a simple function on Ω_2 (consider first the case $A = A_1 \times A_2$), and

$$\mathbb{P}[A] = \int f(\omega_2)\, \mathrm{d}\mathbb{P}_2.$$

Now let $A_n \in \mathcal{A}$ be disjoint sets with union $A \in \mathcal{A}$, let $A_{n,\omega_2} = \{\omega_1 : (\omega_1, \omega_2) \in A_n\}$, and define $f_n(\omega_2) = \mathbb{P}_1[A_{n,\omega_2}]$, so (as above) $\mathbb{P}[A_n] = \int f_n \, \mathrm{d}\mathbb{P}_2$.

For each $\omega_2 \in \Omega_2$, the sets A_{n,ω_2} are disjoint, with union A_{ω_2}. Hence (since $\mathbb{P}_1$ is a measure),

$$\mathbb{P}_1[A_{\omega_2}] = \sum_{n=1}^{\infty} \mathbb{P}_1[A_{n,\omega_2}],$$

i.e., $f = \sum_{n=1}^{\infty} f_n$. Since the f_n are non-negative, the Monotone Convergence Theorem (applied on $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$) gives $\int f = \sum \int f_n$, i.e., $\mathbb{P}[A] = \sum \mathbb{P}[A_n]$. □

By Carathéodory's Extension Theorem (Theorem 1.13) and Theorem 1.12 we see that $\mathbb{P}$ extends uniquely to a probability measure on $\sigma(\mathcal{A}) = \mathcal{F}$.

Definition 2.18 (Product measure)**.** The measure $\mathbb{P}$ defined through (7) is called the *product measure* on $(\Omega, \mathcal{F})$, and denoted $\mathbb{P}_1 \times \mathbb{P}_2$. The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is the *product probability space* $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, \mathbb{P}_1 \times \mathbb{P}_2)$.

The definitions extend easily to define the product $\mathcal{F}_1 \times \cdots \times \mathcal{F}_k$ of k σ-algebras, and the product $\mathbb{P}_1 \times \cdots \times \mathbb{P}_k$ of k probability measures. These product operations behave as you expect: for example, $(\mathbb{P}_1 \times \mathbb{P}_2) \times \mathbb{P}_3 = \mathbb{P}_1 \times \mathbb{P}_2 \times \mathbb{P}_3$. [It is not true in general that $\mathbb{P}_1 \times \mathbb{P}_2 = \mathbb{P}_2 \times \mathbb{P}_1$. If $\Omega_1 \neq \Omega_2$ then these measures aren't even defined on the same space.]

Definition 2.19. Let $(\Omega_i, \mathcal{F}_i)_{i \geqslant 1}$ be a sequence of measurable spaces. The *product σ-algebra* $\mathcal{F} = \prod_{i=1}^{\infty} \mathcal{F}_i$ on $\Omega = \prod_{i=1}^{\infty} \Omega_i$ is the σ-algebra generated by all sets of the form $\prod_{i=1}^{n} A_i \times \prod_{i=n+1}^{\infty} \Omega_i$ where $A_i \in \mathcal{F}_i$, i.e., by all finite-dimensional measurable rectangles.

Remark 2.20 (Countable products of probability measures)**.** Given a sequence of probability spaces, one can define a product probability measure on the product σ-algebra with the expected properties. One way to do this is to apply Theorem 1.13 directly as in the proof of Lemma 2.17, but the condition is quite tricky to verify. It also follows by taking a suitable 'limit' of finite products using the Kolmogorov Consistency Theorem. An alternative approach for Borel measures on $\mathbb{R}$ is outlined on the problem sheets.

The most familiar example of a product measure is, of course, Lebesgue measure on $\mathbb{R}^2$, or, more generally, by extending the above in the obvious way on $\mathbb{R}^d$.

Our integration theory was valid for any measure space $(\Omega, \mathcal{F}, \mu)$ on which μ is a countably additive measure. But as we already know for $\mathbb{R}^2$, in order to calculate the integral of a function of two variables it is convenient to be able to proceed in stages and calculate the repeated integral. So if f is integrable with respect to Lebesgue measure on $\mathbb{R}^2$ then we know that

$$\int_{\mathbb{R}^2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int \left(\int f(x,y)\,\mathrm{d}x \right) \mathrm{d}y = \int \left(\int f(x,y)\,\mathrm{d}y \right) \mathrm{d}x.$$

This result (Fubini's Theorem) applies just as well to the product of general probability measures:

Theorem 2.21 (Fubini + Tonelli)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the product of the probability spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, $i = 1, 2$, and let $f(\omega) = f(\omega_1, \omega_2)$ be a measurable function on $(\Omega, \mathcal{F})$. The functions*

$$x \mapsto \int_{\Omega_2} f(x,y)\,\mathrm{d}\mathbb{P}_2(y), \;\; y \mapsto \int_{\Omega_1} f(x,y)\,\mathrm{d}\mathbb{P}_1(x)$$

are $\mathcal{F}_1$-, $\mathcal{F}_2$-measurable respectively.

Suppose either (i) that f is integrable on Ω or (ii) that $f \geqslant 0$. Then

$$\int_{\Omega} f\,\mathrm{d}\mathbb{P} = \int_{\Omega_2} \left(\int_{\Omega_1} f(x,y)\,\mathrm{d}\mathbb{P}_1(x) \right) \mathrm{d}\mathbb{P}_2(y) = \int_{\Omega_1} \left(\int_{\Omega_2} f(x,y)\,\mathrm{d}\mathbb{P}_2(y) \right) \mathrm{d}\mathbb{P}_1(x),$$

where in case (ii) the common value may be ∞.

Warning: Just as we saw for functions on $\mathbb{R}^2$ in Part A Integration, for f to be integrable we require that $\int |f|\,\mathrm{d}\mathbb{P} < \infty$. If we drop the assumption that f must be integrable or non-negative, then it is not hard to cook up examples where both repeated integrals exist but their values are different.

One of the central ideas in probability theory is *independence* and this is intricately linked with product measure. Intuitively, two events are independent if they have no influence on each other. Knowing that one has happened tells us nothing about the chance that the other has happened. More formally:

Definition 2.22 (Independence)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let I be a finite or countably infinite set. We say that the events $(A_i)_{i\in I}$ where each $A_i \in \mathcal{F}$ are *independent* if for all finite subsets $J \subseteq I$

$$\mathbb{P}\left[\bigcap_{i\in J} A_i\right] = \prod_{i\in J} \mathbb{P}[A_i].$$

Sub σ-algebras $\mathcal{G}_1, \mathcal{G}_2, \ldots$ of $\mathcal{F}$ are called independent if whenever $A_i \in \mathcal{G}_i$ and $i_1, i_2, \ldots, i_n$ are distinct, then

$$\mathbb{P}[A_{i_1} \cap \ldots \cap A_{i_n}] = \prod_{k=1}^{n} \mathbb{P}[A_{i_k}].$$

Note that we impose these conditions for finite subsets only, but they then also hold for countable subsets, using Lemma 1.5. They also hold after complementing some or all of the A_i (exercise).

How does this fit in with our notion of independence from Prelims? We need to relate random variables to σ-algebras.

Definition 2.23 (σ-algebra generated by a random variable)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X be a real-valued random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. The *$\sigma$-algebra generated by X* is

$$\sigma(X) = \{X^{-1}(A) : A \in \mathcal{B}(\mathbb{R})\}.$$

It is easy to check that $\sigma(X)$ is indeed a σ-algebra (see the proof of Proposition 2.2), and by definition of a random variable (as a measurable function on $(\Omega, \mathcal{F})$), we have $\sigma(X) \subseteq \mathcal{F}$. Moreover,

$$\sigma(X) = \sigma\big(\{\{X \leqslant t\} : t \in \mathbb{R}\}\big),$$

where $\{X \leqslant t\} = \{\omega : X(\omega) \leqslant t\}$ (again, c.f. Proposition 2.2).

Definition 2.24 (σ-algebra generated by a sequence of random variables)**.** More generally, if (X_n) is a finite or infinite sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, then

$$\sigma(X_1, X_2, \ldots) = \sigma\left(\bigcup_n \sigma(X_n)\right) = \sigma\big(\{\{X_n \leqslant t\} : n \geqslant 1, t \in \mathbb{R}\}\big).$$

Definition 2.25. Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. Then X is called *$\mathcal{G}$-measurable* if X is measurable as a function on $(\Omega, \mathcal{G})$.

In other words, X is $\mathcal{G}$-measurable if and only if $\sigma(X) \subseteq \mathcal{G}$. Thus $\sigma(X)$ is the smallest σ-algebra with respect to which X is measurable.

It is easy to check that X is $\mathcal{G}$-measurable if and only if $\{X \leqslant t\} \in \mathcal{G}$ for every $t \in \mathbb{R}$.

To understand what these definitions mean, note that a random variable Y is $\sigma(X)$-measurable if and only if $Y = f(X)$ for some measurable function $f : \mathbb{R} \to \mathbb{R}$. Similarly, Y is $\sigma(X_1, X_2, \ldots)$-measurable if and only if $Y = f(X_1, X_2, \ldots)$ for some measurable function f on the countable product of $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with itself.

Definition 2.26 (Independent random variables)**.** Random variables $X_1, X_2, \ldots$ are called *independent* if the σ-algebras $\sigma(X_1), \sigma(X_2), \ldots$ are independent.

If we write this in more familiar language we see that X and Y are independent if for each pair A, B of Borel subsets of $\mathbb{R}$

$$\mathbb{P}[X \in A, Y \in B] = \mathbb{P}[X \in A]\mathbb{P}[Y \in B].$$

Any measurable function f from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a measurable space $(\Lambda, \mathcal{G})$ induces a probability measure $\mu_f = \mathbb{P} \circ f^{-1}$ on $(\Lambda, \mathcal{G})$, defined by $\mu_f(A) = \mathbb{P}[f \in A] = \mathbb{P}[f^{-1}(A)]$. The following result is easy to check.

Lemma 2.27. *Two random variables X and Y on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are independent if and only if the measure $\mu_{X,Y}$ induced on $\mathbb{R}^2$ by (X, Y) is the product measure $\mu_X \times \mu_Y$, where μ_X and μ_Y are the measures on $\mathbb{R}$ induced by X and Y respectively.*

This generalizes the result you learned in Prelims and Part A for discrete/continuous random variables – two continuous random variables X and Y are independent if and only if their joint density function can be written as the product of the density function of X and the density function of Y.

Of course the conditions of Definition 2.26 would be impossible to check in general – we don't have a nice explicit presentation of the σ-algebras $\sigma(X_i)$. But we can use Theorem 1.12 (uniqueness of extension) to reduce it to something much more manageable.

Theorem 2.28. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose that $\mathcal{G}$ and $\mathcal{H}$ are sub σ-algebras of $\mathcal{F}$ and that $\mathcal{G}_0$ and $\mathcal{H}_0$ are π-systems with $\sigma(\mathcal{G}_0) = \mathcal{G}$ and $\sigma(\mathcal{H}_0) = \mathcal{H}$. Then $\mathcal{G}$ and $\mathcal{H}$ are independent* iff *$\mathcal{G}_0$ and $\mathcal{H}_0$ are independent, i.e. $\mathbb{P}[G \cap H] = \mathbb{P}[G]\mathbb{P}[H]$ whenever $G \in \mathcal{G}_0$, $H \in \mathcal{H}_0$.*

Proof. Fix $G \in \mathcal{G}_0$. The two functions $H \mapsto \mathbb{P}[G \cap H]$ and $H \mapsto \mathbb{P}[G]\mathbb{P}[H]$ define measures on $(\Omega, \mathcal{H})$ (check!) with the same total mass $\mathbb{P}[G]$, and they agree on the π-system $\mathcal{H}_0$. So by Theorem 1.12 they agree on $\sigma(\mathcal{H}_0) = \mathcal{H}$. Hence, for $G \in \mathcal{G}_0$ and $H \in \mathcal{H}$

$$\mathbb{P}[G \cap H] = \mathbb{P}[G]\mathbb{P}[H].$$

Now fix $H \in \mathcal{H}$ and repeat the argument with the two measures $G \mapsto \mathbb{P}[G \cap H]$ and $G \mapsto \mathbb{P}[G]\mathbb{P}[H]$. □

This extends easily to n σ-algebras and hence (since independence can be defined considering finitely many at a time) to a sequence of σ-algebras.

Corollary 2.29. *A sequence $(X_n)_{n \geqslant 1}$ of real-valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ is independent* iff *for all $n \geqslant 1$ and all $x_1, \ldots x_n \in \mathbb{R}$ (or $\overline{\mathbb{R}}$),*

$$\mathbb{P}[X_1 \leqslant x_1, \ldots, X_n \leqslant x_n] = \mathbb{P}[X_1 \leqslant x_1] \ldots \mathbb{P}[X_n \leqslant x_n].$$

The existence of countable product spaces tells us that, given Borel probability measures $\mu_1, \mu_2, \ldots$ on $\mathbb{R}$, there is a probability space on which there are *independent* random variables $X_1, X_2, \ldots$ with $\mu_{X_i} = \mu_i$.

We finish this section with one of the most beautiful results in probability theory, concerning 'tail events' associated to sequences of independent random variables.

Definition 2.30 (Tail σ-algebra)**.** For a sequence of random variables $(X_n)_{n \geqslant 1}$ define

$$\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2} \ldots)$$

and

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \mathcal{T}_n.$$

Then $\mathcal{T}$ is called the *tail σ-algebra* of the sequence $(X_n)_{n \geqslant 1}$.

Roughly speaking, any event A such that (a) whether A holds is determined by the sequence (X_n) but (b) changing finitely many of these values does not affect whether A holds is in the tail σ-algebra. These conditions sound impossible, but many events involving limits have these properties. For example, it is easy to check that $A = \{(X_n) \text{ converges}\}$ is a tail event: just check that $A \in \mathcal{T}_n$ for each n.

Theorem 2.31 (Kolmogorov's 0-1 law)**.** *Let (X_n) be a sequence of independent random variables. Then the tail σ-algebra $\mathcal{T}$ of (X_n) contains only events of probability 0 or 1. Moreover, any $\mathcal{T}$-measurable random variable is almost surely constant.*

Proof. Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Note that $\mathcal{F}_n$ is generated by the π-system of events

$$\mathcal{A} = \left\{ \{X_1 \leqslant x_1, \ldots, X_n \leqslant x_n\} : x_1, \ldots, x_n \in \overline{\mathbb{R}} \right\}$$

and $\mathcal{T}_n$ is generated by the π-system of events

$$\mathcal{B} = \left\{ \{X_{n+1} \leqslant x_{n+1}, \ldots, X_{n+k} \leqslant x_{n+k}\} : k \geqslant 1, x_{n+1}, \ldots, x_{n+k} \in \overline{\mathbb{R}} \right\}.$$

For any $A \in \mathcal{A}$, $B \in \mathcal{B}$, by the independence of the random variables (X_n), we have

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$$

and so by Theorem 2.28 the σ-algebras $\sigma(\mathcal{A}) = \mathcal{F}_n$ and $\sigma(\mathcal{B}) = \mathcal{T}_n$ are also independent.

Since $\mathcal{T} \subseteq \mathcal{T}_n$ we conclude that $\mathcal{F}_n$ and $\mathcal{T}$ are also independent. Hence $\bigcup_{n\geqslant 1} \mathcal{F}_n$ and $\mathcal{T}$ are independent.

Now $\bigcup_{n\geqslant 1} \mathcal{F}_n$ is a π-system (although not in general a σ-algebra) generating the σ-algebra $\mathcal{F}_\infty = \sigma((X_n)_{n\geqslant 1})$. So applying Theorem 2.28 again we see that $\mathcal{F}_\infty$ and $\mathcal{T}$ are independent. But $\mathcal{T} \subseteq \mathcal{F}_\infty$ so that if $A \in \mathcal{T}$

$$\mathbb{P}[A] = \mathbb{P}[A \cap A] = \mathbb{P}[A]^2$$

and so $\mathbb{P}[A] = 0$ or $\mathbb{P}[A] = 1$.

Now suppose that Y is any (real-valued) $\mathcal{T}$-measurable random variable. Then its distribution function $F_Y(y) = \mathbb{P}[Y \leqslant y]$ is increasing, right continuous and takes only values in $\{0, 1\}$. So $\mathbb{P}[Y = c] = 1$ where $c = \inf\{y : F_Y(y) = 1\}$. This extends easily to the extended-real-valued case. □

Example 2.32. Let $(X_n)_{n\geqslant 1}$ be a sequence of independent, identically distributed (i.i.d.) random variables and let $S_n = \sum_{k=1}^n X_k$. Consider $L = \limsup_{n\to\infty} S_n/n$. Then L is a tail random variable and so almost surely constant. We'll prove later in the course that, under weak assumptions, $L = \mathbb{E}[X_1]$ almost surely.

3 Modes of convergence

3.1 The Borel–Cantelli Lemmas

We'll return to independence, or more importantly lack of it, in the next section, but first we look at some ramifications of our theory of integration for probability theory. Throughout, $(\Omega, \mathcal{F}, \mathbb{P})$ will denote a probability space.

Definition 3.1. Let (A_n) be a sequence of sets from $\mathcal{F}$. We define

$$\begin{aligned} \limsup_{n\to\infty} A_n &= \bigcap_{n=1}^{\infty} \bigcup_{m\geqslant n} A_m \\ &= \{\omega \in \Omega : \omega \in A_m \text{ for infinitely many } m\} \\ &= \{A_m \text{ occurs infinitely often}\} \\ &= \{A_m \text{ i.o.}\} \end{aligned}$$

and

$$\begin{aligned}\liminf_{n\to\infty} A_n &= \bigcup_{n=1}^{\infty}\bigcap_{m\geqslant n} A_m \\ &= \{\omega\in\Omega:\ \exists m_0(\omega) \text{ such that } \omega\in A_m \text{ for all } m\geqslant m_0(\omega)\} \\ &= \{A_m \text{ eventually}\} \\ &= \{A_m^{\mathrm{c}} \text{ infinitely often}\}^{\mathrm{c}}.\end{aligned}$$

Lemma 3.2.

$$\mathbf{1}_{\limsup_{n\to\infty} A_n} = \limsup_{n\to\infty} \mathbf{1}_{A_n}, \quad \mathbf{1}_{\liminf_{n\to\infty} A_n} = \liminf_{n\to\infty} \mathbf{1}_{A_n}.$$

Proof. Note that $\mathbf{1}_{\bigcup_n A_n} = \sup_n \mathbf{1}_{A_n}$ and $\mathbf{1}_{\bigcap_n A_n} = \inf_n \mathbf{1}_{A_n}$, and apply these twice. □

If we apply Fatou's Lemma to the functions $\mathbf{1}_{A_n}$, we see that

$$\mathbb{P}[A_n \text{ eventually}] \leqslant \liminf_{n\to\infty} \mathbb{P}[A_n]$$

and hence (taking complements)

$$\mathbb{P}[A_n \text{ i.o.}] \geqslant \limsup_{n\to\infty} \mathbb{P}[A_n].$$

These are not surprising, and easy to prove directly. In fact we can say more about the probabilities of these events.

Lemma 3.3 (The First Borel–Cantelli Lemma, BC1)**.** *If* $\sum_{n=1}^{\infty} \mathbb{P}[A_n] < \infty$ *then* $\mathbb{P}[A_n \text{ i.o.}] = 0$.

Remark. Notice that we are making no assumptions about independence here. This is a very powerful result.

Proof. Let $G_n = \bigcup_{m\geqslant n} A_m$. Then

$$\mathbb{P}[G_n] \leqslant \sum_{m=n}^{\infty} \mathbb{P}[A_m]$$

and $G_n \downarrow G = \limsup_{n\to\infty} A_n$, so by Lemma 1.5, $\mathbb{P}[G_n] \downarrow \mathbb{P}[G]$.

Since $\sum_{n=1}^{\infty} \mathbb{P}[A_n] < \infty$, we have that

$$\sum_{m=n}^{\infty} \mathbb{P}[A_m] \to 0 \quad \text{as } n\to\infty,$$

and so

$$\mathbb{P}\big[\limsup_{n\to\infty} A_n\big] = \lim_{n\to\infty} \mathbb{P}[G_n] = 0$$

as required. □

Alternatively, consider $N = \sum_{n=1}^{\infty} \mathbf{1}_{A_n}$, the (random) number of events that hold. Use the Monotone Convergence Theorem to show that $\mathbb{E}[N] = \sum \mathbb{P}[A_n]$, and note that $\mathbb{E}[N] < \infty$ implies $\mathbb{P}[N = \infty] = 0$.

A partial converse to BC1 is provided by the second Borel–Cantelli Lemma, but note that we must now assume that the events are *independent*.

Lemma 3.4 (The Second Borel–Cantelli Lemma, BC2)**.** *Let* (A_n) *be a sequence of* independent *events. If* $\sum_{n=1}^{\infty} \mathbb{P}[A_n] = \infty$ *then* $\mathbb{P}[A_n \text{ i.o.}] = 1$.

Proof. Set $a_m = \mathbb{P}[A_m]$ and note that $1-a \leqslant e^{-a}$. We consider the complementary event $\{A_n^{\mathrm{c}} \text{ eventually}\}$.

$$\begin{aligned}\mathbb{P}\left[\bigcap_{m\geqslant n} A_m^{\mathrm{c}}\right] &= \prod_{m\geqslant n}(1-a_m) \quad \text{(by independence)} \\ &\leqslant \exp\left(-\sum_{m\geqslant n} a_m\right) = 0.\end{aligned}$$

Hence

$$\mathbb{P}\left[A_n^{\mathrm{c}} \text{ eventually}\right] = \mathbb{P}\left[\bigcup_{n\in\mathbb{N}}\bigcap_{m\geqslant n} A_m^{\mathrm{c}}\right] \leqslant \sum_{n=1}^{\infty}\mathbb{P}\left[\bigcap_{m\geqslant n} A_m^{\mathrm{c}}\right] = 0,$$

and

$$\mathbb{P}[A_n \text{ i.o.}] = 1 - \mathbb{P}[A_n^{\mathrm{c}} \text{ eventually}] = 1.$$

□

Example 3.5. A monkey is provided with a typewriter. At each time step it has probability 1/26 of typing any of the 26 letters independently of other times. What is the probability that it will type ABRACADABRA at least once? infinitely often?

Solution. We can consider the events

$$A_k = \{\text{ABRACADABRA is typed between times } 11k+1 \text{ and } 11(k+1)\}$$

for each k. The events are independent and $\mathbb{P}[A_k] = (1/26)^{11} > 0$. So $\sum_{k=1}^{\infty}\mathbb{P}[A_k] = \infty$. Thus BC2 says that with probability 1, A_k happens infinitely often. □

Later in the course, with the help of a suitable martingale, we'll be able to work out how long we must wait, on average, before we see patterns appearing in the outcomes of a series of independent experiments.

We'll see many applications of BC1 and BC2 in what follows. Before developing more machinery, here is one more.

Example 3.6. Let $(X_n)_{n\geqslant 1}$ be independent exponentially distributed random variables with mean 1 and let $M_n = \max\{X_1, \ldots, X_n\}$. Then

$$\mathbb{P}\left[\lim_{n\to\infty}\frac{M_n}{\log n} = 1\right] = 1.$$

Proof. First recall that if X is an exponential random variable with parameter 1 then

$$\mathbb{P}[X \leqslant x] = \begin{cases} 0 & x < 0, \\ 1 - e^{-x} & x \geqslant 0. \end{cases}$$

Fix $0 < \varepsilon < 1$. Then

$$\begin{aligned}\mathbb{P}[M_n \leqslant (1-\varepsilon)\log n] &= \mathbb{P}\left[\bigcap_{i=1}^{n}\{X_i \leqslant (1-\varepsilon)\log n\}\right] \\ &= \prod_{i=1}^{n}\mathbb{P}\left[X_i \leqslant (1-\varepsilon)\log n\right] \quad \text{(independence)} \\ &= \left(1 - \frac{1}{n^{1-\varepsilon}}\right)^n \leqslant \exp(-n^{\varepsilon}).\end{aligned}$$

Thus

$$\sum_{n=1}^{\infty} \mathbb{P}[M_n \leqslant (1-\varepsilon)\log n] < \infty$$

and so by BC1

$$\mathbb{P}[M_n \leqslant (1-\varepsilon)\log n \text{ i.o.}] = 0.$$

Since ε was arbitrary, taking a suitable countable union gives

$$\mathbb{P}\left[\liminf_{n\to\infty} \frac{M_n}{\log n} < 1\right] = 0.$$

The reverse bound is similar: use BC1 to show that

$$\mathbb{P}\left[M_n \geqslant (1+\varepsilon)\log n \text{ i.o.}\right] = \mathbb{P}\left[X_n \geqslant (1+\varepsilon)\log n \text{ i.o.}\right] = 0.$$

□

At first sight, it looks as though BC1 and BC2 are not very powerful - they tell us when certain events have probability zero or one. But for many applications, in particular when the events are independent, many interesting events can *only* have probability zero or one, because they are tail events.

If the X_n in Example 2.32 have mean zero and variance one, then setting

$$B = \left\{\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1\right\}, \tag{9}$$

then by Kolmogorov's 0/1-law we have $\mathbb{P}[B] = 0$ or $\mathbb{P}[B] = 1$. In fact $\mathbb{P}[B] = 1$. This is called the law of the iterated logarithm. Under the slightly stronger assumption that $\exists \alpha > 0$ such that $\mathbb{E}[|X_n|^{2+\alpha}] < \infty$, Varadhan proves this by a (delicate) application of Borel–Cantelli.

You may at this point be feeling a little confused. In Prelims Statistics or Part A Probability (or possibly even at school) you learned that if (X_n) is a sequence of i.i.d. random variables with mean 0 and variance 1 then

$$\mathbb{P}\left[\frac{X_1 + \cdots + X_n}{\sqrt{n}} \leqslant a\right] = \mathbb{P}\left[\frac{S_n}{\sqrt{n}} \leqslant a\right] \overset{n\to\infty}{\longrightarrow} \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x. \tag{10}$$

This is the Central Limit Theorem without which statistics would be a very different subject. How does it fit with (9)? The results (9) and (10) are giving quite different results about the behaviour of S_n for large n. They correspond to different 'modes of convergence'.

Definition 3.7 (Modes of convergence)**.** Let $X_1, X_2, \ldots$ and X be random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

1. We say that (X_n) converges *almost surely* to X (written $X_n \overset{\text{a.s.}}{\to} X$ or $X_n \to X$ a.s.) if

$$\mathbb{P}[X_n \to X] = \mathbb{P}\Big[\big\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\big\}\Big] = 1.$$

2. We say that (X_n) converges to X *in probability* (written $X_n \overset{\mathbb{P}}{\to} X$) if, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = \lim_{n\to\infty} \mathbb{P}\big[\big\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\big\}\big] = 0.$$

3. Suppose that X and all X_n have finite pth moments for some real number $p > 0$, i.e., $\mathbb{E}[|X|^p], \mathbb{E}[|X_n|^p] < \infty$. We say that X_n converges to X in L^p (or in *pth moment*) (written $X_n \xrightarrow{L^p} X$) if

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0.$$

4. Let F and F_n denote the distribution functions of X and X_n respectively. We say that X_n converges to X *in distribution* (written $X_n \xrightarrow{d} X$ or $X_n \xrightarrow{\mathcal{L}} X$) if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for every $x \in \mathbb{R}$ *at which* F *is continuous.*

These notions of convergence are all different.

Convergence a.s. $\implies$ Convergence in Probability $\implies$ Convergence in Distribution

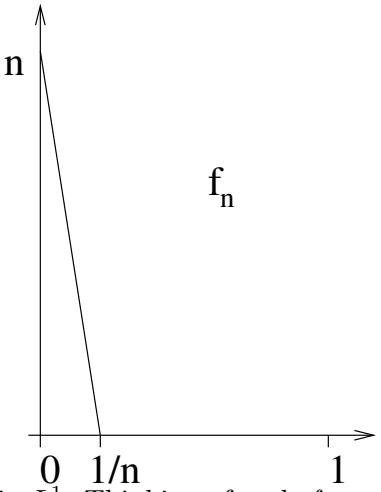
$$\Uparrow$$

Convergence in L^p

The notions of convergence almost surely and convergence in L^p were discussed (for Lebesgue measure, rather than for arbitrary probability measures as here) in Part A Integration.

Example 3.8. On the probability space $\Omega = [0, 1]$ with the Borel σ-algebra and Lebesgue measure, consider the sequence of functions f_n given by

$$f_n(x) = \begin{cases} n(1 - nx) & 0 \leqslant x \leqslant 1/n, \\ 0 & \text{otherwise.} \end{cases}$$



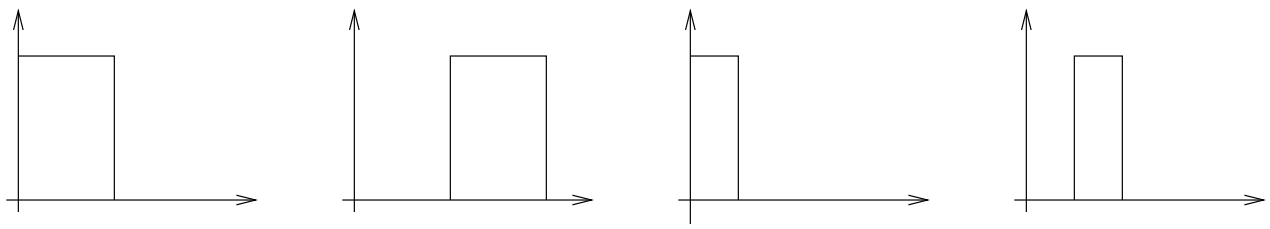
Then $f_n \to 0$ almost everywhere on $[0, 1]$ but $f_n \not\to 0$ in L^1. Thinking of each f_n as a random variable, we have $f_n \to 0$ almost surely but $f_n \not\to 0$ in L^1.

Example 3.9 (Convergence in probability does not imply a.s. convergence)**.** To understand what's going on in (9) and (10), let's stick with $[0, 1]$ with the Borel sets and Lebesgue measure as our probability space. We define $(X_n)_{n\geqslant 1}$ as follows:

for each n there is a unique pair of integers (m, k) such that $n = 2^m + k$ and $0 \leqslant k < 2^m$. We set

$$X_n(\omega) = \mathbf{1}_{[k/2^m,(k+1)/2^m)}(\omega).$$

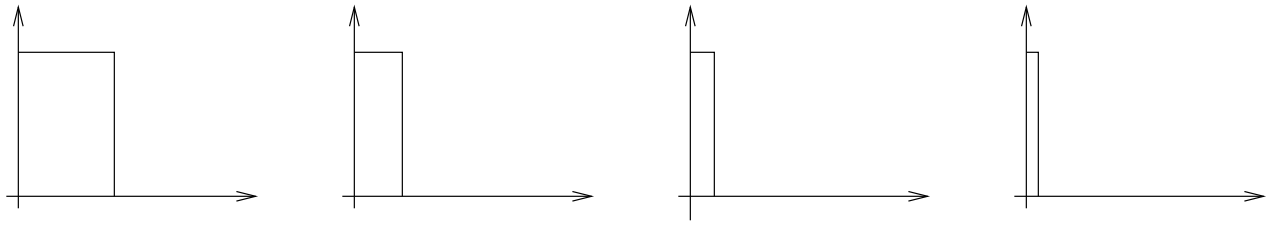
Pictorially we have a 'moving blip' which travels repeatedly across $[0, 1]$ getting narrower at each pass.



For fixed $\omega \in (0,1)$, $X_n(\omega) = 1$ i.o., so $X_n \not\to 0$ a.s., but

$$\mathbb{P}[X_n \neq 0] = \frac{1}{2^m} \to 0 \quad \text{as } n \to \infty,$$

so $X_n \overset{\mathbb{P}}{\to} 0$. (Also, $\mathbb{E}[|X_n - 0|] = 1/2^m \to 0$, so $X_n \overset{L^1}{\to} 0$).) On the other hand, if we look at the $(X_{2^n})_{n \geqslant 1}$, we have



and we see that $X_{2^n} \overset{\text{a.s.}}{\to} 0$.

It turns out that this is a general phenomenon.

Theorem 3.10 (Convergence in Probability and a.s. Convergence)**.** *Let $X_1, X_2, \ldots$ and X be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$.*

1. *If $X_n \overset{\text{a.s.}}{\to} X$ then $X_n \overset{\mathbb{P}}{\to} X$.*

2. *If $X_n \overset{\mathbb{P}}{\to} X$, then there exists a subsequence $(X_{n_k})_{k \geqslant 1}$ such that $X_{n_k} \overset{\text{a.s.}}{\to} X$ as $k \to \infty$.*

Proof. For $\varepsilon > 0$ and $n \in \mathbb{N}$ let

$$A_{n,\varepsilon} = \{|X_n - X| > \varepsilon\}.$$

1. Suppose $X_n \overset{\text{a.s.}}{\to} X$. Then for any $\varepsilon > 0$ we have $\mathbb{P}[A_{n,\varepsilon} \text{ i.o.}] = 0$. However, applying Fatou's Lemma to $\mathbf{1}_{A^c_{n,\varepsilon}}$, we have

$$\mathbb{P}[A_{n,\varepsilon} \text{ i.o.}] = \mathbb{P}[\limsup_{n\to\infty} A_{n,\varepsilon}] \geqslant \limsup_{n\to\infty} \mathbb{P}[A_{n,\varepsilon}].$$

Hence $\mathbb{P}[A_{n,\varepsilon}] \to 0$, so $X_n \overset{\mathbb{P}}{\to} X$.

2. This is the more interesting direction. Suppose that $X_n \overset{\mathbb{P}}{\to} X$. Then for each $k \geqslant 1$ we have $\mathbb{P}[A_{n,1/k}] \to 0$, so there is some n_k such that $\mathbb{P}[A_{n_k,1/k}] < 1/k^2$ and $n_k > n_{k-1}$ for $k \geqslant 2$. Setting $B_k = A_{n_k,1/k}$, we have

$$\sum_{k=1}^{\infty} \mathbb{P}[B_k] \leqslant \sum_{k=1}^{\infty} k^{-2} < \infty.$$

Hence, by BC1, $\mathbb{P}[B_k \text{ i.o.}] = 0$. But if only finitely many B_k hold, then certainly $X_{n_k} \to X$, so $X_{n_k} \overset{\text{a.s.}}{\to} X$. □

The First Borel–Cantelli Lemma provides a very powerful tool for proving almost sure convergence of a sequence of random variables. Its successful application often rests on being able to find good bounds on the random variables X_n. We end this section with some inequalities that are often helpful in this context. The first is trivial, but has many applications.

Lemma 3.11 (Markov's inequality)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X a* non-negative *random variable. Then, for each $\lambda > 0$*

$$\mathbb{P}[X \geqslant \lambda] \leqslant \frac{1}{\lambda}\mathbb{E}[X].$$

Proof. For each $\omega \in \Omega$ we have $X(\omega) \geqslant \lambda \mathbf{1}_{\{X \geqslant \lambda\}}(\omega)$. Hence,

$$\mathbb{E}[X] \geqslant \mathbb{E}[\lambda \mathbf{1}_{\{X \geqslant \lambda\}}] = \lambda \mathbb{P}[X \geqslant \lambda].$$

□

Corollary 3.12 (General Chebyshev's Inequality)**.** *Let X be a random variable taking values in a (measurable) set $A \subseteq \mathbb{R}$, and let $\phi : A \to [0, \infty]$ be an increasing, measurable function. Then for any $\lambda \in A$ with $\phi(\lambda) < \infty$ we have*

$$\mathbb{P}[X \geqslant \lambda] \leqslant \frac{\mathbb{E}[\phi(X)]}{\phi(\lambda)}.$$

Proof. We have

$$\begin{aligned} \mathbb{P}[X \geqslant \lambda] &\leqslant \mathbb{P}[\phi(X) \geqslant \phi(\lambda)] \\ &\leqslant \frac{1}{\phi(\lambda)}\mathbb{E}[\phi(X)], \end{aligned}$$

by Markov's inequality. □

The most familiar special case is given by taking $\phi(x) = x^2$ on $[0, \infty)$ and applying the result to $Y = |X - \mathbb{E}[X]|$, giving

$$\mathbb{P}\big[|X - \mathbb{E}[X]| \geqslant t\big] \leqslant \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{t^2} = \frac{\operatorname{Var}[X]}{t^2}$$

for $t > 0$.

Corollary 3.12 is also often applied with $\phi(x) = e^{\theta x}$, $\theta \geqslant 0$, to obtain

$$\mathbb{P}[X \geqslant \lambda] \leqslant e^{-\theta\lambda}\mathbb{E}[e^{\theta X}].$$

The next step is often to optimize over θ.

Corollary 3.13. *For $p > 0$, convergence in L^p implies convergence in probability.*

Proof. Recall that $X_n \to X$ in L^p if $\mathbb{E}[|X_n - X|^p] \to 0$ as $n \to \infty$. Now

$$\mathbb{P}[|X_n - X| > \varepsilon] = \mathbb{P}[|X_n - X|^p > \varepsilon^p] \leqslant \frac{1}{\varepsilon^p}\mathbb{E}[|X_n - X|^p] \to 0.$$

□

The next corollary is a reminder of a result you have seen in Prelims. It is called the 'weak law' because the notion of convergence is a weak one.

Corollary 3.14 (Weak law of large numbers)**.** *Let $(X_n)_{n\geqslant 1}$ be i.i.d. random variables (on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$) with mean μ and variance $\sigma^2 < \infty$. Set*

$$\overline{X}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then $\overline{X}(n) \to \mu$ in probability as $n \to \infty$.

Proof. We have $\mathbb{E}[\overline{X}(n)] = n^{-1}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mu$ and, since the X_n are independent,

$$\mathrm{Var}[\overline{X}(n)] = n^{-2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = n^{-2}\sum_{i=1}^{n}\mathrm{Var}[X_i] = \sigma^2/n.$$

Hence, by Chebyshev's inequality,

$$\mathbb{P}[|\overline{X}(n) - \mu| > \varepsilon] \leqslant \frac{\mathrm{Var}[\overline{X}(n)]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n} \to 0.$$

□

Definition 3.15 (Convex function)**.** Let $I \subseteq \mathbb{R}$ be a (bounded or unbouded) interval. A function $f : I \to \mathbb{R}$ is *convex* if for all $x, y \in I$ and $t \in [0, 1]$,

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

Important examples of convex functions include x^2, e^x, e^{-x} and $|x|$ on $\mathbb{R}$, and $1/x$ on $(0, \infty)$. Note that a twice differentiable function f is convex if and only if $f''(x) \geqslant 0$ for all x.

Theorem 3.16 (Jensen's inequality)**.** *Let $f : I \to \mathbb{R}$ be a convex function on an interval $I \subseteq \mathbb{R}$. If X is an integrable random variable taking values in I then*

$$\mathbb{E}[f(X)] \geqslant f(\mathbb{E}[X]).$$

(We assume the expectation of $f(X)$ exists also; this is usually no problem since f is often non-negative.)

Perhaps the nicest proof of Theorem 3.16 rests on the following geometric lemma.

Lemma 3.17. *Suppose that $f : I \to \mathbb{R}$ is convex and let m be an interior point of I. Then there exists $a \in \mathbb{R}$ such that $f(x) \geqslant f(m) + a(x - m)$ for all $x \in I$.*

Proof. Let m be an interior point of I. For any $x < m$ and $y > m$ with $x, y \in I$, by convexity we have

$$f(m) \leqslant \frac{y-m}{y-x}f(x) + \frac{m-x}{y-x}f(y).$$

Rearranging (or, better, drawing a picture), this is equivalent to

$$\frac{f(m) - f(x)}{m - x} \leqslant \frac{f(y) - f(m)}{y - m}.$$

It follows that

$$\sup_{x<m} \frac{f(m)-f(x)}{m-x} \leqslant \inf_{y>m} \frac{f(y)-f(m)}{y-m},$$

so choosing a so that

$$\sup_{x<m} \frac{f(m)-f(x)}{m-x} \leqslant a \leqslant \inf_{y>m} \frac{f(y)-f(m)}{y-m}$$

(if f is differentiable at x we can choose $a = f'(x)$) we have that $f(x) \geqslant f(m) + a(x-m)$ for all $x \in I$. □

Proof of Theorem 3.16. If $\mathbb{E}[X]$ is not an interior point of I then it is an endpoint, and X must be almost surely constant, so the inequality is trivial. Otherwise, setting $m = \mathbb{E}[X]$ in the previous lemma we have

$$f(X) \geqslant f(\mathbb{E}[X]) + a(X - \mathbb{E}[X]).$$

Now take expectations to recover

$$\mathbb{E}[f(X)] \geqslant f(\mathbb{E}[X])$$

as required. □

Corollary 3.18. *On a probability space, for any $0 < r < p$, convergence in L^p implies convergence in L^r.*

Proof. The function $f(x) = x^{p/r}$ on $[0,\infty)$ is convex. Suppose $X_n \xrightarrow{L^p} X$. Then

$$\big(\mathbb{E}[|X_n - X|^r]\big)^{p/r} \leqslant \mathbb{E}\big[(|X_n - X|^r)^{p/r}\big] = \mathbb{E}[|X_n - X|^p] \to 0,$$

so $X_n \xrightarrow{L^r} X$. □

Remark. Jensen's inequality only works for probability measures, but often one can exploit it to prove results for finite measures by first normalizing. For example, suppose that μ is a finite measure on $(\Omega, \mathcal{F})$, and define ν by $\nu(A) = \mu(A)/\mu(\Omega)$. Then

$$\begin{aligned} \int |f|^3 \,\mathrm{d}\mu &= \mu(\Omega) \int |f|^3 \,\mathrm{d}\nu \\ &\geqslant \mu(\Omega) \left| \int f \,\mathrm{d}\nu \right|^3 \\ &= \mu(\Omega)^{-2} \left| \int f \,\mathrm{d}\mu \right|^3. \end{aligned}$$

4 Conditional Expectation

Probability is a measure of ignorance. When new information decreases that ignorance we change our probabilities. We formalized this in Prelims through the definition of conditional probability: for a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $A, B \in \mathcal{F}$ with $\mathbb{P}[B] > 0$,

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

We want now to introduce an extension of this which lies at the heart of martingale theory: the notion of conditional expectation. The main difficulty is that we will want to condition on a random variable,

and in many cases, the probability of this taking any specific value will be 0. We get around this by using σ-algebras to represent the 'information' given by random variables.

Here, then, is the definition; we discuss existence, uniqueness and the meaning of this definition below.

Definition 4.1 (Conditional Expectation)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X be an integrable random variable (that is one for which $\mathbb{E}[|X|] < \infty$). Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. The *conditional expectation* $\mathbb{E}[X \mid \mathcal{G}]$ is any $\mathcal{G}$-measurable, integrable random variable Z such that

$$\int_G Z \, \mathrm{d}\mathbb{P} = \int_G X \, \mathrm{d}\mathbb{P} \quad \text{for all } G \in \mathcal{G}.$$

The integrals of X and Z over sets $G \in \mathcal{G}$ are the same, but X is $\mathcal{F}$-measurable whereas Z is $\mathcal{G}$-measurable. The conditional expectation (assuming it exists – we'll show this later) satisfies

$$\int_G \mathbb{E}[X \mid \mathcal{G}] \, \mathrm{d}\mathbb{P} = \int_G X \, \mathrm{d}\mathbb{P} \quad \text{for all } G \in \mathcal{G} \tag{11}$$

and we shall call (11) the *defining relation*.

Just as the probability of an event is a special case of expectation (corresponding to integrating an indicator function rather than a general measurable function), so *conditional probability* is a special case of *conditional expectation*. The conditional probability $\mathbb{P}[A \mid \mathcal{G}]$ is any $\mathcal{G}$-measurable random variable such that

$$\int_G \mathbb{P}[A \mid \mathcal{G}] \, \mathrm{d}\mathbb{P} = \mathbb{P}[A \cap G] \quad \text{for all } G \in \mathcal{G}. \tag{12}$$

This is the same as taking $X = \mathbf{1}_A$ in (11).

Let's see how this fits with our understanding from Prelims. Suppose that X is a discrete random variable taking values $\{x_n\}_{n \in \mathbb{N}}$. Then the events $\{X = x_n\}$ form a *partition* of Ω (that is Ω is the disjoint union of these events.) Let[2]

$$\mathbb{P}[A \mid X] = \mathbb{P}[A \mid \sigma(X)] = \sum_{n=1}^{\infty} \mathbb{P}[A \mid X = x_n] \mathbf{1}_{\{X = x_n\}},$$

which means that for a given $\omega \in \Omega$

$$\mathbb{P}[A \mid \sigma(X)](\omega) = \begin{cases} \mathbb{P}[A \mid X = x_1], & \text{if } X(\omega) = x_1, \\ \mathbb{P}[A \mid X = x_2], & \text{if } X(\omega) = x_2, \\ \quad \cdots & \cdots \\ \mathbb{P}[A \mid X = x_n], & \text{if } X(\omega) = x_n \\ \quad \cdots & \cdots \end{cases}$$

To see that this satisfies (12), write $G_n = \{X = x_n\}$. Then any $G \in \sigma(X)$ is a (necessarily countable) union of these sets (the advantage of with working with discrete random variables again – the σ-algebra

[2]We are assuming here that $\mathbb{P}[X = x_n] > 0$ for each n. In fact, in this context it does not matter how we define $\mathbb{P}[A \mid X = x_n]$ when $\mathbb{P}[X = x_n] = 0$, since there are countably many x_n, so the union of the corresponding events $\{X = x_n\}$ has probability 0.

is easy). So $G = \bigcup_{n\in S} G_n$ for some $S \subseteq \mathbb{N}$, and thus (using monotone convergence in the first step)

$$\begin{aligned}\int_G \left(\sum_{n=1}^{\infty} \mathbb{P}[A \mid X = x_n]\mathbf{1}_{\{X=x_n\}}\right) \mathrm{d}\mathbb{P} &= \sum_{n=1}^{\infty} \int_G \mathbb{P}[A \mid X = x_n]\mathbf{1}_{\{X=x_n\}}\,\mathrm{d}\mathbb{P} \\ &= \sum_{n\in S} \mathbb{P}[A \mid X = x_n]\mathbb{P}[X = x_n] \\ &= \sum_{n\in S} \mathbb{P}[A \cap \{X = x_n\}] \\ &= \mathbb{P}[A \cap G].\end{aligned}$$

This would have worked equally well for any other *countable* partition in place of $\{\{X = x_n\}\}_{n\in\mathbb{N}}$. So, more generally, let $\{G_n\}_{n\in\mathbb{N}}$ be a partition of Ω, let $\mathcal{G}$ be the corresponding σ-algebra (consisting of all unions of sets G_n), and let $\mathbb{E}[X \mid G_n]$ be the conditional expectation relative to the conditional measure $\mathbb{P}[\cdot \mid G_n]$. In other words,

$$\mathbb{E}[X \mid G_n] = \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}[\omega \mid G_n] = \frac{\int_{G_n} X\,\mathrm{d}\mathbb{P}}{\mathbb{P}[G_n]} = \frac{\mathbb{E}[X\mathbf{1}_{G_n}]}{\mathbb{P}[G_n]}.$$

(Note that just like $\mathbb{P}[A \mid G_n]$, $\mathbb{E}[X \mid G_n]$ is a number, not a random variable; conditioning on an *event* gives a number, conditioning on a *random variable* or on a σ-algebra gives a random variable.) We claim that

$$\mathbb{E}[X \mid \mathcal{G}] = \sum_{n=1}^{\infty} \mathbb{E}[X \mid G_n]\mathbf{1}_{G_n}, \tag{13}$$

or, spelled out,

$$\mathbb{E}[X \mid \mathcal{G}](\omega) = \begin{cases} \mathbb{E}[X \mid G_1] & \text{if } \omega \in G_1, \\ \mathbb{E}[X \mid G_2] & \text{if } \omega \in G_2, \\ \quad\cdots & \quad\cdots \\ \mathbb{E}[X \mid G_n] & \text{if } \omega \in G_n, \\ \quad\cdots & \quad\cdots \end{cases}$$

So $\mathbb{E}[X \mid \mathcal{G}]$ is constant on each set G_n, where it takes the value $\mathbb{E}[X \mid G_n]$. To check this, let Z be the right-hand side of (13). Certainly Z is $\mathcal{G}$-measurable; we must show that it satisfies the defining relation. We write the calculation a different way this time even though it is essentially the same as that for conditional probability above.

On G_n, Z takes the constant value $\mathbb{E}[X \mid G_n]$, so

$$\int_{G_n} Z\,\mathrm{d}\mathbb{P} = \mathbb{E}[X \mid G_n]\mathbb{P}[G_n] = \int_{G_n} X\,\mathrm{d}\mathbb{P}, \tag{14}$$

i.e., the defining relation holds for the set G_n. Now any set $G \in \mathcal{G}$ is a countable union $G = \bigcup_{n\in S} G_n$, and the defining relation for G follows by summing[3] over $n \in S$; to see this it may help to rewrite (14) as

$$\mathbb{E}[\mathbf{1}_{G_n} Z] = \mathbb{E}[\mathbf{1}_{G_n} X].$$

At this point, the definition (11) is hopefully starting to make more sense. Since the definition is so important, let us explain it once again, considering the case $\mathcal{G} = \sigma(Y)$. We would like to define a random variable $Z = \mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X \mid Y]$ so that Z depends only on the value of Y and such that

$$Z(\omega) = \mathbb{E}[X \mid Y = y] = \mathbb{E}[X\mathbf{1}_{Y=y}]/\mathbb{P}[Y = y]$$

[3] Of course, we need to be careful with this infinite sum; if X and Z are non-negative we can use monotone convergence. Otherwise either consider positive and negative parts, or use dominated convergence.

when $Y(\omega) = y$. To avoid getting into trouble dividing by zero, we can integrate over $\{Y = y\}$ to express this as

$$\mathbb{E}[Z\mathbf{1}_{Y=y}] = \mathbb{E}[X\mathbf{1}_{Y=y}].$$

Still, if $\mathbb{P}[Y = y] = 0$ for every y (as will often be the case), this condition simply says $0 = 0$. So, just as we did when we failed to express the basic axioms for probability in terms of the probabilities of individual values, we pass to *sets* of values, and in particular Borel sets. So instead we insist that Z is a function of Y and

$$\mathbb{E}[Z\mathbf{1}_{Y\in A}] = \mathbb{E}[X\mathbf{1}_{Y\in A}]$$

for each $A \in \mathcal{B}(\mathbb{R})$. This is exactly what Definition 4.1 says in the case $\mathcal{G} = \sigma(Y)$. In general, it is not the values of Y that matter, but the 'information' in Y, coded by the σ-algebra Y generates, so we define conditional expectation with respect to an arbitrary σ-algebra $\mathcal{G}$. This then covers cases such as conditioning on two random variables at once.

So far we have proved that conditional expectations exist for sub σ-algebras $\mathcal{G}$ generated by partitions. Before proving existence in the general case we show that we have (a.s.) uniqueness.

Proposition 4.2 (Almost sure uniqueness of conditional expectation)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X an integrable random variable and $\mathcal{G} \subseteq \mathcal{F}$ a σ-algebra. If Y and Z are two $\mathcal{G}$-measurable random variables that both satisfy the defining relation (11), then $\mathbb{P}[Y \neq Z] = 0$. That is, Y and Z agree up to a null set.*

Proof. Since Y and Z are both $\mathcal{G}$-measurable,

$$G_1 = \{\omega : Y(\omega) > Z(\omega)\} \in \mathcal{G},$$

so using linearity of the integral and then the defining relation,

$$\int_{G_1} (Y - Z)\,\mathrm{d}\mathbb{P} = \int_{G_1} Y\,\mathrm{d}\mathbb{P} - \int_{G_1} Z\,\mathrm{d}\mathbb{P} = \int_{G_1} X\,\mathrm{d}\mathbb{P} - \int_{G_1} X\,\mathrm{d}\mathbb{P} = 0.$$

Since $Y - Z > 0$ on G_1, from basic properties of integration this implies $\mathbb{P}[G_1] = 0$.

Similarly, $\mathbb{P}[Z > Y] = 0$, which completes the proof. □

In the light of this, we sometimes consider the following equivalence relation.

Definition 4.3 (Equivalence class of a random variable)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X a random variable. The *equivalence class* of X is the collection of random variables that differ from X only on a null set.

For existence of conditional expectation in the general case we will use another important result from measure theory, which we come to in a moment. First observe that if $Z \geqslant 0$ is measurable on $(\Omega, \mathcal{F}, \mathbb{P})$, then we can define a new measure $\mathbb{Q}$ on $(\Omega, \mathcal{F})$ by $\mathbb{Q}[A] = \int_A Z\,\mathrm{d}\mathbb{P}$. Is there a converse? Well, not every measure can arise in this way, since $\mathbb{P}[A] = 0$ implies $\mathbb{Q}[A] = 0$. However, under this condition, the Radon–Nikodym theorem does give a converse.

Definition 4.4. Let $\mathbb{P}$ and $\mathbb{Q}$ be measures on the same measurable space $(\Omega, \mathcal{F})$. The measure $\mathbb{Q}$ is *absolutely continuous* with respect to $\mathbb{P}$, written $\mathbb{Q} \ll \mathbb{P}$, if

$$\mathbb{P}[A] = 0 \implies \mathbb{Q}[A] = 0 \qquad \forall A \in \mathcal{F}.$$

Theorem 4.5 (The Radon–Nikodym Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and suppose that $\mathbb{Q}$ is a finite measure on $(\Omega, \mathcal{F})$ with $\mathbb{Q} \ll \mathbb{P}$. Then there exists an $\mathcal{F}$-measurable function $Z : \Omega \to [0, \infty)$ such that*

$$\mathbb{Q}[A] = \int_A Z \, \mathrm{d}\mathbb{P} \qquad \text{for all } A \in \mathcal{F}.$$

Moreover, Z is unique up to equality $\mathbb{P}$-a.s. It is written

$$Z = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}$$

and is called the Radon–Nikodym derivative *of $\mathbb{Q}$ with respect to $\mathbb{P}$.*

We omit the proof, which is beyond the scope of the course. Instead, let's see how to use this result to deduce the existence of conditional expectation.

Theorem 4.6 (Existence of conditional expectation)**.** *Let X be an integrable random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $\mathcal{G} \subseteq \mathcal{F}$ a σ-algebra. Then there exists a unique equivalence class of $\mathcal{G}$-measurable random variables for which the defining relation (11) holds.*

Proof. We have already dealt with uniqueness in Proposition 4.2, so we just need existence.

Suppose first that X is non-negative. We want to find an integrable $\mathcal{G}$-measurable Z such that, for all $A \in \mathcal{G}$,

$$\int_A Z \, \mathrm{d}\mathbb{P} = \int_A X \, \mathrm{d}\mathbb{P}. \tag{15}$$

So, for $A \in \mathcal{G}$, let $\mathbb{Q}[A] = \int_A X \, \mathrm{d}\mathbb{P}$. This defines a finite measure $\mathbb{Q}$ on $(\Omega, \mathcal{G})$. Let $\mathbb{P}|_{\mathcal{G}}$ denote the measure $\mathbb{P}$ restricted to the σ-algebra $\mathcal{G}$. Then $\mathbb{Q} \ll \mathbb{P}|_{\mathcal{G}}$. So applying the Radon–Nikodym Theorem to $\mathbb{Q}$ and $\mathbb{P}_{\mathcal{G}}$ on $(\Omega, \mathcal{G})$, there is a $\mathcal{G}$-measurable function $Z = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}|_{\mathcal{G}}}$ such that (15) holds.[4] Certainly Z is integrable, since $Z \geqslant 0$ and $\int Z \, \mathrm{d}\mathbb{P} = \int X \, \mathrm{d}\mathbb{P} < \infty$.

For the general case, write $X = X^+ - X^-$ where X^+ and X^- are the positive and negative parts of X. Then $\mathbb{E}[X^+ \mid \mathcal{G}] - \mathbb{E}[X^- \mid \mathcal{G}]$ is $\mathcal{G}$-measurable and, by linearity of the integral, satisfies the defining relation. □

So far, we defined conditional expectations only when X is integrable. Just as with ordinary expectation, the definitions work without problems if $X \geqslant 0$, allowing $+\infty$ as a possible value; this is an (optional) exercise – you have to be a little careful with uniqueness.

It is much harder to write out $\mathbb{E}[X \mid \mathcal{G}]$ explicitly when $\mathcal{G}$ is not generated by a partition. It may help to observe that for non-negative integrable X, if $\mathcal{I}$ is a π-system that generates $\mathcal{G}$, then it is enough to check the defining relation for $G \in \mathcal{I}$. (To see this, apply Theorem 1.12 to the measures $\mathbb{Q}$ and $\int_A Z \, \mathrm{d}\mathbb{P}$ above; it works also for any integrable X, either considering positive and negative parts separately, or a version of Theorem 1.12 for signed measures.)

If $\mathcal{G} = \sigma(Y)$ for some random variable Y on $(\Omega, \mathcal{F}, \mathbb{P})$, then any $\mathcal{G}$-measurable function can, in principle, be written as a function of Y. We saw an example of this with our branching process in §0.3. If Z_n was the number of descendants of a single ancestor after n generations, then

$$\mathbb{E}[Z_{n+1} \mid \sigma(Z_n)] = \mu Z_n$$

where μ is the expected number of offspring of a single individual. In general, of course, the relationship can be much more complicated.

[4] There is a small subtlety here: we are using twice (with $Y = \mathbf{1}_A X$ and $Y = \mathbf{1}_A Z$) that if Y is $\mathcal{G}$-measurable, then $\int Y \, \mathrm{d}\mathbb{P}|_{\mathcal{G}} = \int Y \, \mathrm{d}\mathbb{P}$. This follows from Remark 2.8.

Let's turn to some elementary properties of conditional expectation. Most of the following are obvious. Always remember that whereas expectation is a number, conditional expectation is a *function* on Ω and, since conditional expectation is only defined up to equivalence (i.e., up to equality almost surely) we have to qualify many of our statements with the caveat 'a.s.'.

Proposition 4.7. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X and Y integrable random variables, $\mathcal{G} \subseteq \mathcal{F}$ a σ-algebra and a, b, c real numbers. Then*

1. $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X]$.
2. $\mathbb{E}[aX + bY \mid \mathcal{G}] \overset{\text{a.s.}}{=} a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}]$.
3. *If X is $\mathcal{G}$-measurable, then* $\mathbb{E}[X \mid \mathcal{G}] \overset{\text{a.s.}}{=} X$.
4. $\mathbb{E}[c \mid \mathcal{G}] \overset{\text{a.s.}}{=} c$.
5. $\mathbb{E}[X \mid \{\emptyset, \Omega\}] = \mathbb{E}[X]$.
6. *If $\sigma(X)$ and $\mathcal{G}$ are independent then* $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$ *a.s.*
7. *If $X \leqslant Y$ a.s. then* $\mathbb{E}[X \mid \mathcal{G}] \leqslant \mathbb{E}[Y \mid \mathcal{G}]$ *a.s.*
8. $\big|\mathbb{E}[X \mid \mathcal{G}]\big| \leqslant \mathbb{E}[|X| \mid \mathcal{G}]$ *a.s.*

Proof. The proofs all follow from the requirement that $\mathbb{E}[X \mid \mathcal{G}]$ be $\mathcal{G}$-measurable and the defining relation (11). We just do some examples.

1. Set $G = \Omega$ in the defining relation.

2. Clearly $Z = a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}]$ is $\mathcal{G}$-measurable, so we just have to check the defining relation. But for $G \in \mathcal{G}$,

$$\begin{aligned}\int_G Z \,\mathrm{d}\mathbb{P} = \int_G \big(a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}]\big)\,\mathrm{d}\mathbb{P} &= a\int_G \mathbb{E}[X \mid \mathcal{G}]\,\mathrm{d}\mathbb{P} + b\int_G \mathbb{E}[Y \mid \mathcal{G}]\,\mathrm{d}\mathbb{P} \\ &= a\int_G X \,\mathrm{d}\mathbb{P} + b\int_G Y \,\mathrm{d}\mathbb{P} \\ &= \int_G (aX + bY)\,\mathrm{d}\mathbb{P}.\end{aligned}$$

So Z is a version of $\mathbb{E}[aX + bY \mid \mathcal{G}]$, and equality a.s. follows from uniqueness.

5. The sub σ-algebra is just $\{\emptyset, \Omega\}$ and so $\mathbb{E}[X \mid \{\emptyset, \Omega\}]$ (in order to be measurable with respect to $\{\emptyset, \Omega\}$) must be constant. Now integrate over Ω to identify that constant.

6. Note that $\mathbb{E}[X]$ is $\mathcal{G}$-measurable and for $G \in \mathcal{G}$

$$\begin{aligned}\int_G \mathbb{E}[X]\,\mathrm{d}\mathbb{P} &= \mathbb{E}[X]\mathbb{P}[G] = \mathbb{E}[X]\mathbb{E}[\mathbf{1}_G] \\ &= \mathbb{E}[X\mathbf{1}_G] \quad \text{(by independence – see Problem Sheet 3)} \\ &= \int X\mathbf{1}_G \,\mathrm{d}\mathbb{P} = \int_G X \,\mathrm{d}\mathbb{P},\end{aligned}$$

so the defining relation holds. □

Notice that 6 is intuitively clear. If X is independent of $\mathcal{G}$, then telling me about events in $\mathcal{G}$ tells me nothing about X and so my assessment of its expectation does not change. On the other hand, for 3, if X is $\mathcal{G}$-measurable, then telling me about events in $\mathcal{G}$ actually tells me the value of X.

The conditional counterparts of our convergence theorems of integration also hold good.

Proposition 4.8 (Conditional Convergence Theorems)**.** *Let $X_1, X_2, \dots$ and X be random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra.*

1. **cMON:** *If $X_n \geqslant 0$ for all n and $X_n \uparrow X$ as $n \to \infty$, then $\mathbb{E}[X_n \mid \mathcal{G}] \uparrow \mathbb{E}[X \mid \mathcal{G}]$ a.s. as $n \to \infty$.*

2. **cFatou:** *If $X_n \geqslant 0$ for all n then*

$$\mathbb{E}[\liminf_{n\to\infty} X_n \mid \mathcal{G}] \leqslant \liminf_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{G}] \quad a.s.$$

3. **cDOM:** *If Y is an integrable random variable, $|X_n| \leqslant Y$ for all n and $X_n \overset{\text{a.s.}}{\to} X$, then*

$$\mathbb{E}[X_n \mid \mathcal{G}] \overset{\text{a.s.}}{\to} \mathbb{E}[X \mid \mathcal{G}] \quad \text{as } n \to \infty.$$

The proofs all use the defining relation (11) to transfer statements about convergence of the conditional probabilities to our usual convergence theorems and are left as an exercise.

The following two results are incredibly useful in manipulating conditional expectations. The first is sometimes referred to as 'taking out what is known'.

Lemma 4.9. *Let X and Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with X, Y and XY integrable. Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra and suppose that Y is $\mathcal{G}$-measurable. Then*

$$\mathbb{E}[XY \mid \mathcal{G}] \overset{\text{a.s.}}{=} Y\mathbb{E}[X \mid \mathcal{G}].$$

Proof. The function $Y\mathbb{E}[X \mid \mathcal{G}]$ is clearly $\mathcal{G}$-measurable, so we must check that it satisfies the defining relation for $\mathbb{E}[XY \mid \mathcal{G}]$. We do this by a standard sequence of steps.

First suppose that X and Y are non-negative. If $Y = \mathbf{1}_A$ for some $A \in \mathcal{G}$, then for any $G \in \mathcal{G}$ we have $G \cap A \in \mathcal{G}$ and so by the defining relation (11) for $\mathbb{E}[X \mid \mathcal{G}]$

$$\int_G Y\mathbb{E}[X \mid \mathcal{G}]\,\mathrm{d}\mathbb{P} = \int_{G\cap A} \mathbb{E}[X \mid \mathcal{G}]\,\mathrm{d}\mathbb{P} = \int_{G\cap A} X\,\mathrm{d}\mathbb{P} = \int_G YX\,\mathrm{d}\mathbb{P}.$$

Now extend by linearity to simple random variables Y. Now suppose that $Y \geqslant 0$ is $\mathcal{G}$-measurable. Then there is a sequence $(Y_n)_{n\geqslant 1}$ of simple $\mathcal{G}$-measurable random variables with $Y_n \uparrow Y$ as $n \to \infty$, it follows that $Y_nX \uparrow YX$ and $Y_n\mathbb{E}[X \mid \mathcal{G}] \uparrow Y\mathbb{E}[X \mid \mathcal{G}]$ from which we deduce the result by the Monotone Convergence Theorem. Finally, for X, Y not necessarily non-negative, write $XY = (X^+ - X^-)(Y^+ - Y^-)$ and use linearity of the integral. □

Proposition 4.10 (Tower property of conditional expectations)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X an integrable random variable and $\mathcal{F}_1$, $\mathcal{F}_2$ σ-algebras with $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}$. Then*

$$\mathbb{E}\Big[\mathbb{E}[X \mid \mathcal{F}_2] \;\Big|\; \mathcal{F}_1\Big] = \mathbb{E}[X \mid \mathcal{F}_1] \quad a.s.$$

In other words, writing $X_i = \mathbb{E}[X \mid \mathcal{F}_i]$,

$$\mathbb{E}[X_2 \mid \mathcal{F}_1] = X_1 \quad a.s.$$

Proof. The left-hand side is certainly $\mathcal{F}_1$-measurable, so we need to check the defining relation for $\mathbb{E}[X \mid \mathcal{F}_1]$. Let $G \in \mathcal{F}_1$, noting that $G \in \mathcal{F}_2$. Applying the defining relation twice

$$\int_G \mathbb{E}\Big[\mathbb{E}[X \mid \mathcal{F}_2] \;\Big|\; \mathcal{F}_1\Big]\,\mathrm{d}\mathbb{P} = \int_G \mathbb{E}[X \mid \mathcal{F}_2]\,\mathrm{d}\mathbb{P} = \int_G X\,\mathrm{d}\mathbb{P}.$$

□

This extends Part 1 of Proposition 4.7 which (in the light of Part 5) is just the case $\mathcal{F}_1 = \{\emptyset, \Omega\}$.

Jensen's inequality also extends to the conditional setting.

Proposition 4.11 (Conditional Jensen's Inequality)**.** *Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and that X is an integrable random variable taking values in an interval $I \subseteq \mathbb{R}$. Let $f : I \to \mathbb{R}$ be convex and let $\mathcal{G}$ be a sub σ-algebra of $\mathcal{F}$. If $\mathbb{E}[|f(X)|] < \infty$ then*

$$\mathbb{E}[f(X) \mid \mathcal{G}] \geqslant f\left(\mathbb{E}[X \mid \mathcal{G}]\right) \quad a.s.$$

Sketch proof; not examinable. We shall take I to be an open interval so that we don't have to worry about endpoints. In general the endpoints cause an inconvenience rather than a real problem.

Recall from our proof of Jensen's inequality that if f is convex, then for m in the interior of I (i.e., now for all $m \in I$) we can find at least one straight line touching f from below at $x = m$; i.e., we can find $a, b \in \mathbb{R}$ with $f(x) \geqslant ax + b$ for all $x \in I$, with equality at m.

Consider the set of all functions $g(x)$ of the form $g(x) = ax + b$ with $g(x) \leqslant f(x)$ for all $x \in I$. Then we can check that f is the pointwise supremum of this set of functions. Also f is continuous. With a little work, it follows that we can find a *countable* set of linear functions such that $f(x) = \sup_n\{a_n x + b_n\}$.

Now for our random variable X, since $f(X) \geqslant a_n X + b_n$ we have

$$\mathbb{E}[f(X) \mid \mathcal{G}] \geqslant \mathbb{E}[a_n X + b_n \mid \mathcal{G}] = a_n \mathbb{E}[X \mid \mathcal{G}] + b_n \quad \text{a.s.} \tag{16}$$

Since the union of a countable collection of null (i.e., probability zero) sets is null we can arrange for (16) to hold simultaneously for all $n \in \mathbb{N}$ except possibly on a null set and so

$$\begin{aligned} \mathbb{E}[f(X) \mid \mathcal{G}] &\geqslant \sup_n\{a_n \mathbb{E}[X \mid \mathcal{G}] + b_n\} \quad \text{a.s.} \\ &= f\left(\mathbb{E}[X \mid \mathcal{G}]\right) \quad \text{a.s.} \end{aligned}$$

□

An important special case is $f(x) = x^p$ for $p > 1$. In particular, for $p = 2$

$$\mathbb{E}[X^2 \mid \mathcal{G}] \geqslant \mathbb{E}[X \mid \mathcal{G}]^2 \quad \text{a.s.}$$

A very simple special case of this is the following.

Example 4.12. Suppose that X is a non-negative random variable. Then

$$\mathbb{P}[X > 0] \geqslant \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

Proof. Although this is not the simplest way to see it, we use the conditional form of Jensen's inequality. Recall that for a random variable Y, $\mathbb{E}[X \mid Y]$ is short for $\mathbb{E}[X \mid \sigma(Y)]$. Let $A = \{X > 0\}$. Then

$$\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid \mathbf{1}_A]\big] = \mathbb{P}[A]\mathbb{E}[X \mid A] + \mathbb{P}[A^{\mathrm{c}}]\mathbb{E}[X \mid A^{\mathrm{c}}] = \mathbb{P}[A]\mathbb{E}[X \mid A].$$

Similarly, and using Proposition 4.11,

$$\mathbb{E}[X^2] = \mathbb{E}\big[\mathbb{E}[X^2 \mid \mathbf{1}_A]\big] \geqslant \mathbb{E}\big[\mathbb{E}[X \mid \mathbf{1}_A]^2\big] = \mathbb{P}[A]\mathbb{E}[X \mid A]^2.$$

Combining and rearranging gives the result. □

There is another interesting characterization of $\mathbb{E}[X \mid \mathcal{G}]$.

Remark (Conditional Expectation and Mean Square Approximation)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X, Y square integrable random variables. Let $\mathcal{G}$ be a sub σ-algebra of $\mathcal{F}$ and suppose that Y is $\mathcal{G}$-measurable. Then

$$\begin{aligned}\mathbb{E}[(Y-X)^2] &= \mathbb{E}\Big[\big(Y - \mathbb{E}[X \mid \mathcal{G}] + \mathbb{E}[X \mid \mathcal{G}] - X\big)^2\Big] \\ &= \mathbb{E}\big[(Y - \mathbb{E}[X \mid \mathcal{G}])^2\big] + \mathbb{E}\big[(\mathbb{E}[X \mid \mathcal{G}] - X)^2\big] + 2\mathbb{E}[WZ]\end{aligned}$$

where $W = Y - \mathbb{E}[X \mid \mathcal{G}]$ and $Z = \mathbb{E}[X \mid \mathcal{G}] - X$. Now Y and $\mathbb{E}[X \mid \mathcal{G}]$ are $\mathcal{G}$-measurable, so W is $\mathcal{G}$ measurable, and using Proposition 4.7 part 1 and Lemma 4.9 we have

$$\mathbb{E}[WZ] = \mathbb{E}\big[\mathbb{E}[WZ \mid \mathcal{G}]\big] = \mathbb{E}\big[W\mathbb{E}[Z \mid \mathcal{G}]\big].$$

But $\mathbb{E}\big[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{G}\big] = \mathbb{E}[X \mid \mathcal{G}]$, so $\mathbb{E}[Z \mid \mathcal{G}] = 0$. Hence $\mathbb{E}[WZ] = 0$, i.e., the cross-terms vanish.

In particular, we can minimize $\mathbb{E}[(Y-X)^2]$ by choosing $Y = \mathbb{E}[X \mid \mathcal{G}]$. In other words, $\mathbb{E}[X \mid \mathcal{G}]$ is the best mean-square approximation of X among all $\mathcal{G}$-measurable random variables.

If you have already done Hilbert space theory then $\mathbb{E}[X \mid \mathcal{G}]$ is the orthogonal projection of $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ onto the closed subspace $L^2(\Omega, \mathcal{G}, \mathbb{P})$. Indeed this is a route to showing that conditional expectations exist without recourse to the Radon–Nikodym Theorem.

5 Martingales

Much of modern probability theory derived from two sources: the mathematics of measure and gambling. (The latter perhaps explains why it took so long for probability theory to become a respectable part of mathematics.) Although the term 'martingale' has many meanings outside mathematics – it is the name given to a strap attached to a fencer's épée, it's a strut under the bowsprit of a sailing ship and it is part of a horse's harness that prevents the horse from throwing its head back – it's introduction to mathematics, by Ville in 1939, was inspired by the gambling strategy 'the infallible martingale'. This is a strategy for making a sure profit on games such as roulette in which one makes a sequence of bets. The strategy is to stake £1 (on, say, black or red at roulette) and keep doubling the stake until that number wins. When it does, all previous losses and more are recouped and you leave the table with a profit. It doesn't matter how unfavourable the odds are, only that a winning play comes up eventually. But the martingale is not infallible. Nailing down why in purely mathematical terms had to await the development of martingales in the mathematical sense by J.L. Doob in the 1940's. Doob originally called them 'processes with property E', but in his famous book on stochastic processes he reverted to the term 'martingale' and he later attributed much of the success of martingale theory to the name.

The mathematical term martingale doesn't refer to the gambling *strategy*, but rather models the outcomes of a series of fair games (although as we shall see this is only one application).

We begin with some terminology.

Definition 5.1 (Filtration)**.** A *filtration* on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a sequence $(\mathcal{F}_n)_{n\geqslant 0}$ of σ-algebras $\mathcal{F}_n \subseteq \mathcal{F}$ such that for all n, $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$.

We then call $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geqslant 0}, \mathbb{P})$ a *filtered probability space*.

Usually n is interpreted as time and $\mathcal{F}_n$ represents knowledge accumulated by time n (we never forget anything). We usually start at time 0 (the beginning), but not always.

Definition 5.2 (Adapted stochastic process)**.** A *stochastic process* $(X_n)_{n\geqslant 0}$ is simply a sequence of random variables defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The process is *integrable* if each X_n is integrable.

We say that $(X_n)_{n\geqslant 0}$ is *adapted* to the filtration $(\mathcal{F}_n)_{n\geqslant 0}$ if, for each n, X_n is $\mathcal{F}_n$-measurable.

Definition 5.3 (Martingale, submartingales, supermartingale)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_n)_{n\geqslant 0}$ a filtration. An *integrable, $\mathcal{F}_n$-adapted* stochastic process $(X_n)_{n\geqslant 0}$ is called

1. a *martingale* if for every $n \geqslant 0$, $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = X_n$ a.s.,
2. a *submartingale* if for every $n \geqslant 0$, $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \geqslant X_n$ a.s.,
3. a *supermartingale* if for every $n \geqslant 0$, $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \leqslant X_n$ a.s.

If we think of X_n as our accumulated fortune when we make a sequence of bets, then a martingale represents a fair game in the sense that the conditional expectation of $X_{n+1} - X_n$, given our knowledge at the time when we make the $(n+1)$st bet (that is $\mathcal{F}_n$), is zero. A submartingale represents a favourable game and a supermartingale an unfavourable game. It could be argued that these terms are the wrong way round, but they are very well established, so even if so, it's too late to change this!

Here are some elementary properties.

Proposition 5.4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.*

1. *A stochastic process $(X_n)_{n\geqslant 0}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is a submartingale w.r.t. the filtration $(\mathcal{F}_n)_{n\geqslant 0}$ if and only if $(-X_n)_{n\geqslant 0}$ is a supermartingale. It is a martingale if and only if it is both a supermartingale and a submartingale.*
2. *If $(X_n)_{n\geqslant 0}$ is a martingale w.r.t. $(\mathcal{F}_n)_{n\geqslant 0}$ then*
$$\mathbb{E}[X_n] = \mathbb{E}[X_0] \quad \text{for all } n.$$
3. *If $(X_n)_{n\geqslant 0}$ is a submartingale and $n \geqslant m$ then*
$$\mathbb{E}[X_n \mid \mathcal{F}_m] \geqslant X_m \text{ a.s.}$$
and
$$\mathbb{E}[X_n] \geqslant \mathbb{E}[X_m].$$

Of course, part 3 holds for a supermartingale with the inequalities reversed, and for a martingale with equality instead.

Proof. 1. is obvious.

2. Is a special case of (the martingale version of) 3.

3. Fix m; we prove the result by induction on n. The base case is $n = m$ where, since X_m is $\mathcal{F}_m$-measurable, we have $\mathbb{E}[X_m \mid \mathcal{F}_m] = X_m$ a.s.

For $n \geqslant m$ we have $\mathcal{F}_m \subseteq \mathcal{F}_n$, so
$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_m] = \mathbb{E}\big[\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \mid \mathcal{F}_m\big] \geqslant \mathbb{E}[X_n \mid \mathcal{F}_m] \text{ a.s.},$$
so $\mathbb{E}[X_n \mid \mathcal{F}_m] \geqslant X_m$ a.s. follows by induction. To deduce that $\mathbb{E}[X_n] \geqslant \mathbb{E}[X_m]$ just take the expectation. □

Note that whether (X_n) is a martingale or not depends on the filtration under consideration. If none is specified, there is a default.

Definition 5.5 (Natural filtration)**.** The *natural filtration* $(\mathcal{G}_n)_{n\geqslant 0}$ associated with a stochastic process $(X_n)_{n\geqslant 0}$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is defined by
$$\mathcal{G}_n = \sigma(X_0, X_1, \ldots, X_n), \quad n \geqslant 0.$$

A stochastic process is automatically adapted to the natural filtration associated with it.

Proposition 5.6. *If $(X_n)_{n\geqslant 0}$ is a submartingale w.r.t. some filtration $(\mathcal{F}_n)_{n\geqslant 0}$ on $(\Omega, \mathcal{F}, \mathbb{P})$, then it is also a submartingale with respect to its natural filtration $(\mathcal{G}_n)_{n\geqslant 0}$.*

Proof. $(X_n)_{n\geqslant 0}$ is certainly adapted to its natural filtration $(\mathcal{G}_n)_{n\geqslant 0}$. For each n, since $\mathcal{F}_0, \ldots, \mathcal{F}_{n-1} \subseteq \mathcal{F}_n$, all of $X_0, \ldots, X_n$ are $\mathcal{F}_n$-measurable. Since (by definition) $\mathcal{G}_n$ is the smallest σ-algebra with this property, $\mathcal{G}_n \subseteq \mathcal{F}_n$. Thus, by the tower property,

$$\mathbb{E}[X_{n+1} \mid \mathcal{G}_n] = \mathbb{E}\big[\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \mid \mathcal{G}_n\big] \geqslant \mathbb{E}[X_n \mid \mathcal{G}_n] = X_n \text{ a.s.}$$

□

Warning: There is a reason why we usually have a filtration in mind. It's clear that if (X_n) and (Y_n) are martingales with respect to the same filtration $(\mathcal{F}_n)$, then so is $(X_n + Y_n)$. But it is easy to find examples where (X_n) is a martingale with respect to its natural filtration, (Y_n) is a martingale with respect to its natural filtration, but $(X_n + Y_n)$ is not a martingale with respect to its natural filtration. So it's not just to be fussy that we specify a filtration $(\mathcal{F}_n)$.

Example 5.7 (Sums of independent random variables)**.** Suppose that $Y_1, Y_2, \ldots$ are independent random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and that $\mathbb{E}[Y_n] = 0$ for each n. For $n \geqslant 0$ let

$$X_n = \sum_{k=1}^{n} Y_k,$$

so in particular $X_0 = 0$. Then $(X_n)_{n\geqslant 0}$ is a martingale with respect to the natural filtration given by

$$\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n) = \sigma(Y_1, \ldots, Y_n).$$

In this sense martingales generalize the notion of sums of independent random variables with mean zero. The independent random variables $(Y_i)_{i\geqslant 1}$ of Example 5.7 can be replaced by martingale differences (which are not necessarily independent).

Definition 5.8 (Martingale differences)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_n)_{n\geqslant 0}$ a filtration. A sequence $(Y_n)_{n\geqslant 1}$ of integrable random variables, adapted to the filtration $(\mathcal{F}_n)_{n\geqslant 1}$, is called a *martingale difference sequence* w.r.t. $(\mathcal{F}_n)$ if

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] = 0 \quad \text{a.s.} \quad \text{for all } n \geqslant 0.$$

It is easy to check that $(X_n)_{n\geqslant 0}$ is a martingale w.r.t. $(\mathcal{F}_n)_{n\geqslant 0}$ if and only if X_0 is integrable and $\mathcal{F}_0$-measurable, and $(X_n - X_{n-1})_{n\geqslant 1}$ is a martingale difference sequence w.r.t. $(\mathcal{F}_n)$.

Example 5.9. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(Z_n)_{n\geqslant 1}$ be a sequence of independent random variables with $\mathbb{E}[Z_n] = 1$ for all n. Define

$$X_n = \prod_{i=1}^{n} Z_i \quad \text{for } n \geqslant 0,$$

so $X_0 = 1$. Then $(X_n)_{n\geqslant 0}$ is a martingale w.r.t. its natural filtration. (Exercise).

This is an example where the martingale is (obviously) not a sum of independent random variables.

Example 5.10. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(\mathcal{F}_n)_{n\geqslant 0}$ be a filtration. Let X be an integrable random variable (that is $\mathbb{E}[|X|] < \infty$). Then setting

$$X_n = \mathbb{E}[X \mid \mathcal{F}_n]$$

defines a martingale $(X_n)_{n\geqslant 0}$ w.r.t. $(\mathcal{F}_n)_{n\geqslant 0}$. Indeed, X_n is certainly $\mathcal{F}_n$-measurable, and by the tower property of conditional expectation,

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{F}_{n+1}] \mid \mathcal{F}_n] = \mathbb{E}[X \mid \mathcal{F}_n] = X_n \quad \text{a.s.}$$

We shall see later that a large class of martingales (called uniformly integrable) can be written in this way. One can think of $(\mathcal{F}_n)_{n\geqslant 0}$ as representing unfolding information about X, and we'll see that under suitable assumptions $X_n \to X$ a.s. as $n \to \infty$.

We now turn to ways of obtaining (sub/super)martingales from other martingales. The first way is trivial: suppose that $(X_n)_{n\geqslant 0}$ is a (sub)martingale with respect to $(\mathcal{F}_n)_{n\geqslant 0}$, and that Y is $\mathcal{F}_0$-measurable. Then $(X_n - Y)_{n\geqslant 0}$ is also a (sub)martingale w.r.t. $(\mathcal{F}_n)$. In particular, if X_0 is $\mathcal{F}_0$-measurable, then $(X_n)_{n\geqslant 0}$ is a martingale if and only if $(X_n - X_0)_{n\geqslant 0}$ is a martingale. This is often useful, as in many contexts it allows us to assume without loss of generality that $X_0 = 0$.

Proposition 5.11. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose that $(X_n)_{n\geqslant 0}$ is a martingale with respect to the filtration $(\mathcal{F}_n)_{n\geqslant 0}$. Let f be a* convex *function on $\mathbb{R}$. If $f(X_n)$ is an integrable random variable for each $n \geqslant 0$, then $(f(X_n))_{n\geqslant 0}$ is a* sub*martingale w.r.t $(\mathcal{F}_n)_{n\geqslant 0}$.*

Proof. Since X_n is $\mathcal{F}_n$-measurable, so is $f(X_n)$. By Jensen's inequality for conditional expectations and the martingale property of (X_n),

$$\mathbb{E}[f(X_{n+1}) \mid \mathcal{F}_n] \geqslant f\big(\mathbb{E}[X_{n+1} \mid \mathcal{F}_n]\big) = f(X_n) \quad \text{a.s.}$$

□

Corollary 5.12. *If $(X_n)_{n\geqslant 0}$ is a martingale w.r.t. $(\mathcal{F}_n)_{n\geqslant 0}$ and $K \in \mathbb{R}$ then (subject to integrability) $(|X_n|)_{n\geqslant 0}$, $(X_n^2)_{n\geqslant 0}$, $(e^{X_n})_{n\geqslant 0}$, $(e^{-X_n})_{n\geqslant 0}$, $(\max(X_n, K))_{n\geqslant 0}$ are all submartingales w.r.t. $(\mathcal{F}_n)_{n\geqslant 0}$.*

Definition 5.13 (Predictable process)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_n)_{n\geqslant 0}$ a filtration. A sequence $(V_n)_{n\geqslant 1}$ of random variables is *predictable* with respect to $(\mathcal{F}_n)_{n\geqslant 0}$ if V_n is $\mathcal{F}_{n-1}$-measurable for all $n \geqslant 1$.

In other words, the value of V_n is known 'one step in advance.'

Theorem 5.14 (Discrete stochastic integral or martingale transform)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_n)_{n\geqslant 0}$ a filtration. Let $(Y_n)_{n\geqslant 0}$ be a martingale with respect to $(\mathcal{F}_n)$ with difference sequence $(D_n)_{n\geqslant 1}$. Suppose that $(V_n)_{n\geqslant 1}$ is predictable w.r.t. $(\mathcal{F}_n)$, and let*

$$X_n = \sum_{k=1}^{n} V_k D_k = \sum_{k=1}^{n} V_k (Y_k - Y_{k-1}).$$

Then, assuming each X_n is integrable, $(X_n)_{n\geqslant 0}$ is a martingale w.r.t. $(\mathcal{F}_n)$.

The sequence $(X_n)_{n\geqslant 0}$ is called a *martingale transform*, and is often denoted

$$((V \circ Y)_n)_{n\geqslant 0}.$$

It is a discrete version of the stochastic integral. Here we started with $X_0 = 0$; as far as obtaining a martingale is concerned, it makes no difference if we add some $\mathcal{F}_0$-measurable random variable Z to all X_n; sometimes we take $Z = Y_0$, so $X_n = Y_0 + \sum_{k=1}^{n} V_k D_k$.

Proof. For $k \leqslant n$, D_k and V_k are $\mathcal{F}_n$-measurable, so X_n is $\mathcal{F}_n$-measurable. Also,

$$\begin{aligned}\mathbb{E}[X_{n+1} - X_n \mid \mathcal{F}_n] &\overset{\text{a.s.}}{=} \mathbb{E}[D_{n+1}V_{n+1} \mid \mathcal{F}_n] \\ &\overset{\text{a.s.}}{=} V_{n+1}\mathbb{E}[D_{n+1} \mid \mathcal{F}_n] \quad \text{(taking out what is known)} \\ &= 0 \quad \text{a.s}\end{aligned}$$

□

Typical examples of predictable sequences appear in gambling or finance contexts where they might constitute strategies for future action. The strategy is then based on the current state of affairs. If, for example, $(k-1)$ rounds of some gambling game have just been completed, then the strategy for the kth round is to bet V_k; a quantity that can only depend on what is known by time $k-1$. The change in fortune in the kth round is then $V_k D_k$.

Another situation is when $V_k = 1$ as long as some special event has not yet happened and $V_k = 0$ thereafter. That is the game goes on until the event occurs. This is called a *stopped* martingale – a topic we'll return to in due course.

Theorem 5.15. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_n)_{n\geqslant 0}$ a filtration. Let $(Y_n)_{n\geqslant 0}$ be a* supermartingale *with respect to $(\mathcal{F}_n)$ with difference sequence $(D_n)_{n\geqslant 1}$, and $(V_n)_{n\geqslant 1}$ a* non-negative *$(\mathcal{F}_n)$-predictable sequence. Then (modulo integrability)*

$$X_n = \sum_{k=1}^{n} V_k D_k$$

defines a supermartingale w.r.t. $(\mathcal{F}_n)$.

Proof. Exercise: imitate the proof of Theorem 5.14. □

There are more examples on the problem sheet. Here is one last one.

Example 5.16. Let $(Y_i)_{i\geqslant 1}$ be independent random variables such that $\mathbb{E}[Y_i] = \mu_i$, $\mathrm{Var}(Y_i) = \mathbb{E}[Y_i^2] - \mathbb{E}[Y_i]^2 = \sigma_i^2 < \infty$. Let

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2.$$

(That is $s_n^2 = \mathrm{Var}(\sum_{i=1}^n Y_i)$ by independence.) Take $(\mathcal{F}_n)_{n\geqslant 0}$ to be the natural filtration generated by $(Y_n)_{n\geqslant 1}$.

By Example 5.7,

$$X_n = \sum_{i=1}^{n} (Y_i - \mu_i)$$

is a martingale and so by Proposition 5.11, since $f(x) = x^2$ is a convex function, $(X_n^2)_{n\geqslant 0}$ is a submartingale. But we can recover a martingale from it by *compensation*:

$$M_n = \left(\sum_{i=1}^{n} (Y_i - \mu_i)\right)^2 - s_n^2, \qquad n \geqslant 0$$

is a *martingale* with respect to $(\mathcal{F}_n)_{n\geqslant 0}$.

Proof. Clearly M_n is $\mathcal{F}_n$-measurable. By considering the sequence $\widetilde{Y}_i = Y_i - \mu_i$ of independent mean zero random variables if necessary, we see that w.l.o.g. we may assume $\mu_i = 0$ for all i. Then

$$\begin{aligned}
\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} Y_i + Y_{n+1}\right)^2 - s_{n+1}^2 \;\middle|\; \mathcal{F}_n\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} Y_i\right)^2 + 2Y_{n+1}\sum_{i=1}^{n} Y_i + Y_{n+1}^2 - s_{n+1}^2 \;\middle|\; \mathcal{F}_n\right] \\
&= \left(\sum_{i=1}^{n} Y_i\right)^2 + 2\sum_{i=1}^{n} Y_i\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] + \mathbb{E}[Y_{n+1}^2 \mid \mathcal{F}_n] - s_n^2 - \sigma_{n+1}^2 \quad \text{a.s.} \\
&= M_n
\end{aligned}$$

since, by independence, $\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[Y_{n+1}] = 0$ a.s. and $\mathbb{E}[Y_{n+1}^2 \mid \mathcal{F}_n] = \mathbb{E}[Y_{n+1}^2] = \sigma_{n+1}^2$. □

This process of 'compensation', whereby we correct a process by something predictable (in this example it was deterministic) in order to obtain a martingale reflects a general result due to Doob.

Theorem 5.17 (Doob's Decomposition Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_n)_{n\geqslant 0}$ a filtration. Let $(X_n)_{n\geqslant 0}$ be a sequence of integrable random variables, adapted to $(\mathcal{F}_n)_{n\geqslant 0}$. Then*

1. *$(X_n)_{n\geqslant 0}$ has a* Doob decomposition

$$X_n = X_0 + M_n + A_n \tag{17}$$

where $(M_n)_{n\geqslant 0}$ is a martingale w.r.t. $(\mathcal{F}_n)_{n\geqslant 0}$, $(A_n)_{n\geqslant 1}$ is predictable w.r.t. $(\mathcal{F}_n)$, and $M_0 = 0 = A_0$.

2. *Doob decompositions are essentially unique: if $X_n = X_0 + \widetilde{M}_n + \widetilde{A}_n$ is another Doob decomposition of $(X_n)_{n\geqslant 0}$ then*

$$\mathbb{P}[M_n = \widetilde{M}_n,\ A_n = \widetilde{A}_n \text{ for all } n] = 1.$$

3. *$(X_n)_{n\geqslant 0}$ is a submartingale if and only if $(A_n)_{n\geqslant 0}$ in (17) is an increasing process (i.e., $A_{n+1} \geqslant A_n$ a.s. for all n) and a supermartingale if and only if $(A_n)_{n\geqslant 0}$ is a decreasing process.*

Proof.

1. Let

$$A_n = \sum_{k=1}^{n} \mathbb{E}[X_k - X_{k-1} \mid \mathcal{F}_{k-1}] = \sum_{k=1}^{n} \big(\mathbb{E}[X_k \mid \mathcal{F}_{k-1}] - X_{k-1}\big)$$

and

$$M_n = \sum_{k=1}^{n} \big(X_k - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]\big).$$

Then $M_n + A_n = \sum_{k=1}^{n}(X_k - X_{k-1}) = X_n - X_0$, so (17) holds. The kth summand in A_n is $\mathcal{F}_{k-1}$-measurable, so A_n is $\mathcal{F}_{n-1}$-measurable and (A_n) is predictable w.r.t. $(\mathcal{F}_n)$. Also, since

$$\mathbb{E}[M_{n+1} - M_n \mid \mathcal{F}_n] = \mathbb{E}\big[X_{n+1} - \mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \;\big|\; \mathcal{F}_n\big] = 0,$$

the process $(M_n)_{n\geqslant 0}$ is a martingale w.r.t. $(\mathcal{F}_n)$.

2. For uniqueness, note that in any Doob decomposition, by predictability we have

$$\begin{aligned} A_{n+1} - A_n &= \mathbb{E}[A_{n+1} - A_n \mid \mathcal{F}_n] \\ &= \mathbb{E}[(X_{n+1} - X_n) - (M_{n+1} - M_n) \mid \mathcal{F}_n] \\ &= \mathbb{E}[X_{n+1} - X_n \mid \mathcal{F}_n] \quad \text{a.s.}, \end{aligned}$$

which combined with $A_0 = 0$ proves uniqueness of (A_n). Since $M_n = X_n - X_0 - A_n$, uniqueness of (M_n) follows.

3. Just note that

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] - X_n = \mathbb{E}[X_{n+1} - X_n \mid \mathcal{F}_n] = A_{n+1} - A_n \quad \text{a.s.}$$

as shown above. □

Remark (The angle bracket process $\langle M \rangle$)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\mathcal{F}_n)_{n \geqslant 0}$ a filtration and $(M_n)_{n \geqslant 0}$ a martingale with respect to $(\mathcal{F}_n)_{n \geqslant 0}$ with $\mathbb{E}[M_n^2] < \infty$ for each n. (Such a martingale is called an L^2-martingale.) Then by Proposition 5.11, $(M_n^2)_{n \geqslant 0}$ is a *sub*martingale. Thus by Theorem 5.17 it has a Doob decomposition (which is essentially unique),

$$M_n^2 = M_0^2 + N_n + A_n$$

where $(N_n)_{n \geqslant 0}$ is a martingale and $(A_n)_{n \geqslant 0}$ is an increasing predictable process. The process $(A_n)_{n \geqslant 0}$ is often denoted by $(\langle M \rangle_n)_{n \geqslant 0}$.

Note that $\mathbb{E}[M_n^2] = \mathbb{E}[M_0^2] + \mathbb{E}[A_n]$ and (since $\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = M_n$) that

$$A_{n+1} - A_n = \mathbb{E}[M_{n+1}^2 - M_n^2 \mid \mathcal{F}_n] = \mathbb{E}[(M_{n+1} - M_n)^2 \mid \mathcal{F}_n].$$

That is, the increments of A_n are the conditional variances of our martingale difference sequence. It turns out that $(\langle M \rangle_n)_{n \geqslant 0}$ is an extremely powerful tool with which to study $(M_n)_{n \geqslant 0}$. It is beyond our scope here, but see for example Neveu 1975, Discrete Parameter Martingales.

6 Stopping Times and Stopping Theorems

Much of the power of martingale methods, as we shall see, comes from the fact that (under suitable boundedness assumptions) the martingale property is preserved if we 'stop' the process at certain random times. Such times are called 'stopping times' (or sometimes 'optional times').

Intuitively, stopping times are times that we can recognize when they arrive, like the first time heads comes up in a series of coin tosses or the first time the FTSE 100 index takes a 3% fall in a single day. They are times which can be recognized without reference to the future. Something like 'the day in December when the FTSE 100 reaches its maximum' is *not* a stopping time - we must wait until the *end* of December to determine the maximum, and by then, in general, the time has passed.

Stopping times can be used for strategies of investing and other forms of gambling. We recognize them when they arrive and can make decisions based on them (for example to stop playing).

Definition 6.1 (Stopping time)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_n)_{n \geqslant 0}$ a filtration. A random variable τ taking values in $\mathbb{N} \cup \{\infty\} = \{0, 1, 2, \ldots, \infty\}$ is called a *stopping time with respect to* $(\mathcal{F}_n)_{n \geqslant 0}$ if $\{\tau = n\} \in \mathcal{F}_n$ for all n. Stopping times are sometimes called *optional times*.

Equivalently, τ is a stopping time if and only if $\{\tau > n\} \in \mathcal{F}_n$ for all n.

Warning: Some authors assume $\mathbb{P}[\tau < \infty] = 1$.

We write $n \wedge \tau$ for the smaller of n and τ, i.e., for $\min\{n, \tau\}$. If $(X_n)_{n\geqslant 0}$ is a stochastic process, then $(X_{n\wedge\tau})_{n\geqslant 0}$ is the process *stopped at time* τ: we have $X_{n\wedge\tau} = X_n$ if $n \leqslant \tau$ and $X_{n\wedge\tau} = X_\tau$ if $n \geqslant \tau$. Note that $\tau = \infty$ corresponds to never stopping.

Lemma 6.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\mathcal{F}_n)_{n\geqslant 0}$ a filtration, $(M_n)_{n\geqslant 0}$ a (sub/super)martingale with respect to $(\mathcal{F}_n)_{n\geqslant 0}$ and τ a stopping time with respect to $(\mathcal{F}_n)_{n\geqslant 0}$. Then $(M_{n\wedge\tau})_{n\geqslant 0}$ is also a (sub/super)martingale with respect to $(\mathcal{F}_n)_{n\geqslant 0}$.*

Proof. Note that $X_{n\wedge\tau} - X_{(n-1)\wedge\tau} = \mathbf{1}_{\tau\geqslant n}(X_n - X_{n-1})$. Let $V_n = \mathbf{1}_{\tau\geqslant n}$. Since $\{\tau \geqslant n\} = \{\tau \leqslant n-1\}^{\mathrm{c}} \in \mathcal{F}_{n-1}$, the random variable V_n is $\mathcal{F}_{n-1}$-measurable, so (V_n) is predictable. Since V_n is non-negative (and bounded, so there are no problems with integrability) the conditions of Theorem 5.14/Theorem 5.15 are satisfied. □

This lemma tells us that if (M_n) is a martingale and τ is a stopping time, then $\mathbb{E}[M_{n\wedge\tau}] = \mathbb{E}[M_0]$. Can we let $n \to \infty$ to obtain $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$? The answer is 'no'.

Example 6.3. Let $(Y_k)_{k\geqslant 1}$ be i.i.d. random variables with $\mathbb{P}[Y_k = 1] = \mathbb{P}[Y_k = -1] = 1/2$. Set $M_n = \sum_{k=1}^n Y_k$. Thus M_n is the position of a simple random walk started from the origin after n steps. In particular, $(M_n)_{n\geqslant 0}$ is a martingale and $\mathbb{E}[M_n] = 0$ for all n.

Now let $\tau = \min\{n : M_n = 1\}$, which is defined a.s. It is clear that τ is a stopping time and evidently $M_\tau = 1$. But then $\mathbb{E}[M_\tau] = 1 \neq 0 = \mathbb{E}[M_0]$.

The problem is that τ is too large – $\mathbb{E}[\tau] = \infty$. It turns out that if we impose suitable boundedness assumptions then we will have $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$ and that is the celebrated Optional Stopping Theorem. There are many variants of this result.

Theorem 6.4 (Doob's Optional Stopping Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\mathcal{F}_n)_{n\geqslant 0}$ a filtration, $(M_n)_{n\geqslant 0}$ a martingale with respect to $(\mathcal{F}_n)_{n\geqslant 0}$ and τ a stopping time with respect to $(\mathcal{F}_n)_{n\geqslant 0}$. Suppose any of the following conditions holds:*

1. *τ is bounded, i.e., there is some $N \in \mathbb{N}$ such that $\tau(\omega) \leqslant N$ for all $\omega \in \Omega$.*
2. *τ is a.s. finite and $(M_n)_{n\geqslant 0}$ is uniformly bounded, i.e., there is some $K \in \mathbb{R}$ such that $|M_n(\omega)| \leqslant K$ for every $n \in \mathbb{N}$ and every $\omega \in \Omega$.*
3. *$\mathbb{E}[\tau] < \infty$ and there exists $L \in \mathbb{R}$ such that*

$$\mathbb{E}\big[|M_{n+1} - M_n| \;\big|\; \mathcal{F}_n\big] \leqslant L, \qquad \text{a.s. for all } n.$$

Then M_τ is integrable and

$$\mathbb{E}[M_\tau] = \mathbb{E}[M_0] \tag{18}$$

Proof. By Lemma 6.2, $(M_{n\wedge\tau})_{n\geqslant 0}$ is a martingale, so for each n, $\mathbb{E}[M_{n\wedge\tau}] = \mathbb{E}[M_{0\wedge\tau}] = \mathbb{E}[M_0]$.

1. Since $\tau \leqslant N$ always holds, we have $M_\tau = M_{N\wedge\tau}$, so we are done by the comment above.

2. Because $\tau < \infty$, $\lim_{n\to\infty} M_{n\wedge\tau} = M_\tau$ a.s. and since $(M_n)_{n\geqslant 0}$ is bounded we may apply the Dominated Convergence Theorem with dominating function $g(\omega) \equiv K$ to deduce the result.

3. Replacing M_n by $M_n - M_0$, we assume without loss of generality that $M_0 = 0$. Then

$$\begin{aligned} |M_{n\wedge\tau}| = |M_{n\wedge\tau} - M_{0\wedge\tau}| &\leqslant \sum_{i=1}^{n} |M_{i\wedge\tau} - M_{(i-1)\wedge\tau}| \\ &\leqslant \sum_{i=1}^{\infty} |M_{i\wedge\tau} - M_{(i-1)\wedge\tau}| \\ &= \sum_{i=1}^{\infty} \mathbf{1}_{\tau\geqslant i}|M_i - M_{i-1}|. \end{aligned} \tag{19}$$

Now

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbf{1}_{\tau\geqslant i}|M_i - M_{i-1}|\right] &= \sum_{i=1}^{\infty} \mathbb{E}\big[\mathbf{1}_{\tau\geqslant i}|M_i - M_{i-1}|\big] \quad \text{(by monotone convergence)} \\ &= \sum_{i=1}^{\infty} \mathbb{E}\left[\ \mathbb{E}\big[\mathbf{1}_{\tau\geqslant i}|M_i - M_{i-1}| \ \big|\ \mathcal{F}_{i-1}\big]\ \right] \quad \text{(tower property)} \\ &= \sum_{i=1}^{\infty} \mathbb{E}\left[\ \mathbf{1}_{\tau\geqslant i}\mathbb{E}\big[|M_i - M_{i-1}| \ \big|\ \mathcal{F}_{i-1}\big]\ \right] \quad \text{(since } \{\tau \geqslant i\} \in \mathcal{F}_{i-1}\text{)} \\ &\leqslant L\sum_{i=1}^{\infty} \mathbb{E}[\mathbf{1}_{\tau\geqslant i}] \\ &= L\sum_{i=1}^{\infty} \mathbb{P}[\tau \geqslant i] = L\mathbb{E}[\tau] < \infty. \end{aligned}$$

Moreover, $\tau < \infty$ a.s. and so $M_{n\wedge\tau} \to M_\tau$ a.s. as $n \to \infty$ and so by the Dominated Convergence Theorem with the function on the right hand side of (19) as dominating function, we have the result. □

We stated the Optional Stopping Theorem for martingales, but similar results are available for *sub/super*-martingales – just replace the equality in (18) by the appropriate inequality.

Note that if $|M_i - M_{i-1}| \leqslant L$ always holds, and $\mathbb{E}[\tau] < \infty$, then the third case applies; this is perhaps the most important case of the Optional Stopping Theorem for applications.

In order to make use of condition 3, we need to be able to check when $\mathbb{E}[\tau] < \infty$. The following lemma provides a useful test.

Lemma 6.5. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\mathcal{F}_n)_{n\geqslant 0}$ a filtration and τ a stopping time with respect to $(\mathcal{F}_n)_{n\geqslant 0}$. Suppose that there exist $K \in \mathbb{N}$ and $\varepsilon > 0$ such that for all $n \in \mathbb{N}$*

$$\mathbb{P}[\tau \leqslant n + K \mid \mathcal{F}_n] \geqslant \varepsilon \quad \text{a.s.}$$

Then $\mathbb{E}[\tau] < \infty$.

The proof is an exercise.

Let's look at an application of Theorem 6.4.

Example 6.6. Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $(X_i)_{i\geqslant 1}$ are i.i.d. random variables with $\mathbb{P}[X_i = j] = p_j > 0$ for each $j = 0, 1, 2, \ldots$. What is the expected number of random variables that must be observed before the subsequence $0, 1, 2, 0, 1$ occurs?

Solution

Consider a casino offering fair bets, where the expected gain from each bet is zero. In particular, a gambler betting £a on the outcome of the next random variable being a j will lose with probability $1 - p_j$ and will win £a/p_j with probability p_j. (Her expected pay-out is $0(1 - p_j) + p_j a/p_j = a$, the same as the stake.)

Imagine a sequence of gamblers betting at the casino, each with an initial fortune of £1.

Gambler i bets £1 that $X_i = 0$; if she wins, she bets her entire fortune of £$1/p_0$ that $X_{i+1} = 1$; if she wins again she bets her fortune of £$1/(p_0p_1)$ that $X_{i+2} = 2$; if she wins that bet, then she bets £$1/(p_0p_1p_2)$ that $X_{i+3} = 0$; if she wins that bet then she bets her total fortune of £$1/(p_0^2p_1p_2)$ that $X_{i+4} = 1$; if she wins she quits with a fortune of £$1/(p_0^2p_1^2p_2)$.

Let M_n be the casino's winnings after n games (so when X_n has just been revealed). Then $(M_n)_{n\geqslant 0}$ is a mean zero martingale w.r.t. the filtration $(\mathcal{F}_n)_{n\geqslant 0}$ where $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. If we write τ for the number of random variables to be revealed before we see the required pattern, then by Lemma 6.5, $\mathbb{E}[\tau] < \infty$. Since at most 5 people bet at any one time, $|M_{n+1} - M_n|$ is bounded by a constant (say $L = 5/(p_0^2p_1^2p_2)$), so Condition 3 of Theorem 6.4 is satisfied (with this L).

After X_τ has been revealed, gamblers $1, 2, \ldots, \tau - 5$ have each lost £1.

- Gambler $\tau - 4$ has gained £$1/(p_0^2p_1^2p_2) - 1$,
- Gamblers $\tau - 3$ and $\tau - 2$ have each lost £1,
- Gambler $\tau - 1$ has gained £$1/(p_0p_1) - 1$,
- Gambler τ has lost £1.

Of course, gamblers $\tau + 1, \tau + 2, \ldots$ have not bet at all yet. Thus

$$M_\tau = \tau - \frac{1}{p_0^2p_1^2p_2} - \frac{1}{p_0p_1}.$$

By Theorem 6.4 $\mathbb{E}[M_\tau] = 0$, so taking expectations,

$$\mathbb{E}[\tau] = \frac{1}{p_0^2p_1^2p_2} + \frac{1}{p_0p_1}.$$

□

The same trick can be used to calculate the expected time until any specified (finite) pattern occurs in i.i.d. data.

Before finishing this section, we'll use a stopping-time idea to give a tail bound for (sub)martingales.

According to Markov's inequality, if X is a random variable and $\lambda > 0$, then

$$\mathbb{P}[|X| \geqslant \lambda] \leqslant \frac{\mathbb{E}[|X|]}{\lambda}.$$

Martingales satisfy a similar, but much more powerful inequality, which bounds the *maximum* of the process.

Lemma 6.7. *Let X be a submartingale, and let $N \in \mathbb{N}$. Let τ be a stopping time. Then $\mathbb{E}[X_{\tau\wedge N}] \leqslant \mathbb{E}[X_N]$.*

Proof. Let $Y_n = X_n - X_{\tau\wedge n} = \sum_{k=1}^n V_k(X_k - X_{k-1}$, where $V_k = \mathbf{1}_{\tau<k}$. Then V is a non-negative, bounded, and predictable process, so by Theorem 5.15, Y is a submartingale. Then $0 = \mathbb{E}[Y_0] \leqslant \mathbb{E}[Y_N] = \mathbb{E}[X_N - X_{\tau\wedge N}]$. □

Theorem 6.8 (A form of Doob's maximal inequality)**.** *Let $(X_n)_{n\geqslant 0}$ be a non-negative submartingale (w.r.t. a filtration $(\mathcal{F}_n)_{n\geqslant 0}$). Then for fixed $N \geqslant 0$ and $\lambda > 0$,*

$$\mathbb{P}\big[\max_{n\leqslant N} X_n \geqslant \lambda\big] \leqslant \frac{\mathbb{E}[X_N]}{\lambda}.$$

Proof. Let $\tau = \inf\{n : X_n \geqslant \lambda\}$. Then τ is a stopping time, and $\{\max_{n\leqslant N} X_n \geqslant \lambda\} = \{X_{\tau\wedge N} \geqslant \lambda\}$. Thus using Lemma 6.7,

$$\mathbb{P}\Big[\max_{n\leqslant N} X_n \geqslant \lambda\Big] = \mathbb{P}[X_{\tau\wedge N} \geqslant \lambda] \leqslant \frac{1}{\lambda}\mathbb{E}[X_{\tau\wedge N}] \leqslant \frac{1}{\lambda}\mathbb{E}[X_N]$$

as required. □

Corollary 6.9. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_n)_{n\geqslant 0}$ a filtration. If $(M_n)_{n\geqslant 0}$ is a martingale with respect to $(\mathcal{F}_n)_{n\geqslant 0}$ then for $N \geqslant 0$ and $\lambda > 0$,*

$$\mathbb{P}\left[\max_{n\leqslant N} |M_n| \geqslant \lambda\right] \leqslant \frac{\mathbb{E}[|M_N|]}{\lambda},$$

and, assuming $\mathbb{E}[M_n^2] < \infty$ for each n,

$$\mathbb{P}\left[\max_{n\leqslant N} |M_n| \geqslant \lambda\right] \leqslant \frac{\mathbb{E}[M_N^2]}{\lambda^2}.$$

Proof. For the first statement, note that since $f(x) = |x|$ is convex, $(|M_n|)_{n\geqslant 0}$ is a submartingale. For the second, use $f(x) = x^2$; by the integrability assumption and Proposition 5.11, (M_n^2) is a submartingale. Hence,

$$\mathbb{P}\left[\max_{n\leqslant N} |M_n| \geqslant \lambda\right] = \mathbb{P}\left[\max_{n\leqslant N} M_n^2 \geqslant \lambda^2\right] \leqslant \frac{\mathbb{E}[M_N^2]}{\lambda^2}.$$

□

Of course, this works using $|M_n|^p$ for any $p \geqslant 1$, or, more generally, any non-negative *increasing* convex function of $|M_n|$.

7 The Upcrossing Lemma and Martingale Convergence

Let $(X_n)_{n\geqslant 0}$ be an integrable random process, for example modelling the value of an asset. Suppose that $(V_n)_{n\geqslant 1}$ is a predictable process representing an investment strategy based on that asset. The result of Theorem 5.15 tells us that if $(X_n)_{n\geqslant 0}$ is a supermartingale and our strategy $(V_n)_{n\geqslant 1}$ only allows us to hold non-negative amounts of the asset, then our fortune is also a supermartingale. (This is of course bad news, our expected fortune goes down.)

Consider the following strategy:

1. You do not invest until the value of X goes below some level a (representing what you consider to be a bottom price), in which case you buy a share.

2. You keep your share until X gets above some level b (a value you consider to be overpriced) in which case you sell your share and you return to the first step.

Three remarks:

1. However clever this strategy may seem, if $(X_n)_{n\geqslant 0}$ is a supermartingale and you stop playing at some bounded stopping time, then in expectation your losses will at least equal your winnings.

2. Your 'winnings', i.e., profit from shares actually sold, are at least $(b-a)$ times the number of times the process went up from a to b. (They can be greater, since the price can 'jump over' a and b.)

3. If you stop, owning a share, at a time n when the value is below the price at which you bought, then (selling out) you lose an amount which is at most $(X_n - a)^-$: you bought at or below a.

Combining these remarks, if $(X_n)_{n\geqslant 0}$ is a supermartingale we should be able to bound (from above) the expected number of times the stock price rises from a to b by $\mathbb{E}[(X_n-a)^-]/(b-a)$. This is precisely what Doob's upcrossing inequality will tell us. To make it precise, we need some notation.

Definition 7.1 (Upcrossings)**.** If $\mathbf{x} = (x_n)_{n\geqslant 0}$ is a sequence of real numbers and $a < b$ are fixed, define two integer-valued sequences $(S_k)_{k\geqslant 1} = (S_k([a,b],\mathbf{x}))_{k\geqslant 1}$ and $(T_k)_{k\geqslant 0} = (T_k([a,b],\mathbf{x}))_{k\geqslant 0}$ recursively as follows:

Let $T_0 = 0$ and for $k \geqslant 1$ let

$$S_k = \inf\{n \geqslant T_{k-1} : x_n \leqslant a\},$$

$$T_k = \inf\{n \geqslant S_k : x_n \geqslant b\},$$

with the usual convention that $\inf \emptyset = \infty$.

Let

$$U_n([a,b],\mathbf{x}) = \max\{k \geqslant 0 : T_k \leqslant n\}$$

be the number of upcrossings of $[a,b]$ by $\mathbf{x}$ by time n and let

$$U([a,b],\mathbf{x}) = \sup_n U_n([a,b],\mathbf{x}) = \sup\{k \geqslant 0 : T_k < \infty\}$$

be the total number of upcrossings of $[a,b]$ by $\mathbf{x}$.

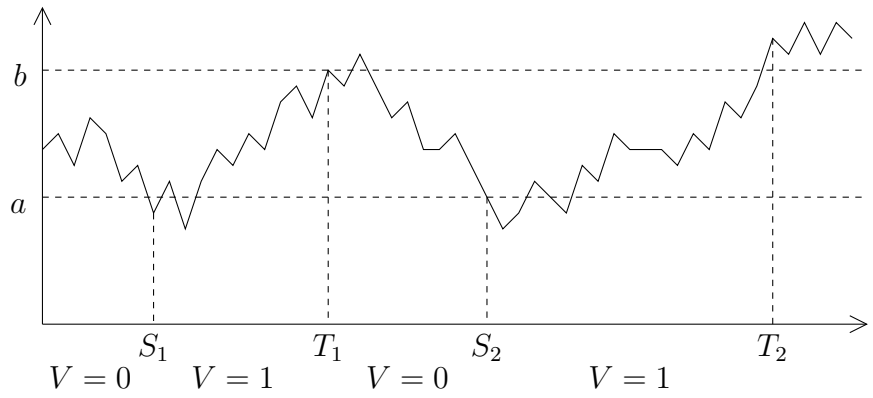
Lemma 7.2 (Doob's upcrossing lemma)**.** *Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, $(\mathcal{F}_n)_{n\geqslant 0}$ a filtration and $\mathbf{X} = (X_n)_{n\geqslant 0}$ a supermartingale w.r.t. $(\mathcal{F}_n)_{n\geqslant 0}$. Let $a < b$ be fixed real numbers. Then for every $n \geqslant 0$,*

$$\mathbb{E}[U_n([a,b],\mathbf{X})] \leqslant \frac{\mathbb{E}[(X_n - a)^-]}{b-a}.$$

Proof. It is an easy induction to check that for $k \geqslant 1$, the random variables $S_k = S_k([a,b],\mathbf{X})$ and $T_k = T_k([a,b],\mathbf{X})$ are stopping times. Now set

$$V_n = \sum_{k\geqslant 1} \mathbf{1}_{\{S_k < n \leqslant T_k\}}.$$

Notice that V_n only takes the values 0 and 1. It is 1 at time n if $\mathbf{X}$ is in the process of making an upcrossing from a to b or if $S_k < n$ and $T_k = \infty$. It is precisely the predictable process in our investment strategy above: we hold one unit of stock during an upcrossing or if T_k is infinite for some k and $n > S_k$.



Notice that

$$\{S_k < n \leqslant T_k\} = \{S_k \leqslant n-1\} \cap \{T_k \leqslant n-1\}^{\mathrm{c}} \in \mathcal{F}_{n-1}.$$

So $(V_n)_{n\geqslant 1}$ is *predictable* (recall Definition 5.13). Now just as in Theorem 5.14 we construct the discrete stochastic integral

$$\begin{aligned}
(V \circ X)_n &= \sum_{k=1}^{n} V_k(X_k - X_{k-1}) \\
&= \sum_{i=1}^{U_n} (X_{T_i} - X_{S_i}) + \mathbf{1}_{\{S_{U_n+1}<n\}}(X_n - X_{S_{U_n+1}}) \qquad (20) \\
&\geqslant (b-a)U_n - (X_n - a)^-. \qquad (21)
\end{aligned}$$

For the last step, note that if indicator function in (20) is non-zero, then $S_{U_n+1} < \infty$, so $X_{S_{U_n+1}} \leqslant a$. Hence $X_n - X_{S_{U_n+1}} \geqslant X - a \geqslant -(X-a)^-$.

Since $(V_n)_{n\geqslant 1}$ is bounded, non-negative and predictable and $(X_n)_{n\geqslant 0}$ is a supermartingale, by Theorem 5.15 $((V \circ X)_n)_{n\geqslant 0}$ is a supermartingale. So taking expectations in (21),

$$0 = \mathbb{E}[(V \circ X)_0] \geqslant \mathbb{E}[(V \circ X)_n] \geqslant (b-a)\mathbb{E}[U_n] - \mathbb{E}[(X_n - a)^-]$$

and rearranging gives the result. □

One way to show that a sequence of real numbers converges as $n \to \infty$ is to show that it doesn't oscillate too wildly; this can be expressed in terms of upcrossings as follows.

Lemma 7.3. *A real sequence $\mathbf{x} = (x_n)$ converges to a limit in $[-\infty, \infty]$ if and only if $U([a,b], \mathbf{x}) < \infty$ for all $a, b \in \mathbb{Q}$ with $a < b$.*

Proof. From the definitions/basic analysis, $\mathbf{x}$ converges if and only if $\liminf x_n = \limsup x_n$.

(i) If $U([a,b], \mathbf{x}) = \infty$, then

$$\liminf_{n\to\infty} x_n \leqslant a < b \leqslant \limsup_{n\to\infty} x_n$$

and so $\mathbf{x}$ does not converge.

(ii) If $\mathbf{x}$ does not converge, then we can choose rationals a and b with

$$\liminf_{n\to\infty} x_n < a < b < \limsup_{n\to\infty} x_n,$$

and then $U([a,b], \mathbf{x}) = \infty$. □

A supermartingale (X_n) is just a random sequence; by Doob's Upcrossing Lemma we can bound the expected number of upcrossings of $[a, b]$ that it makes for any $a < b$ and so our hope is that we can combine this with Lemma 7.3 to show that the *random* sequence (X_n) converges. This is our next result.

Definition 7.4. Let (X_n) be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $p \geqslant 1$. We say that (X_n) is *bounded in L^p* if

$$\sup_n \mathbb{E}[|X_n|^p] < \infty.$$

Note that the condition says exactly that the set $\{X_n\}$ of random variables is a bounded subset of $L^p(\Omega, \mathcal{F}, \mathbb{P})$: there is some K such that $||X_n||_p \leqslant K$ for all n.

Theorem 7.5 (Doob's Forward Convergence Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_n)_{n\geqslant 0}$ a filtration. Suppose that $(X_n)_{n\geqslant 0}$ is a sub- or supermartingale w.r.t $(\mathcal{F}_n)_{n\geqslant 0}$ that is bounded in L^1. Then $(X_n)_{n\geqslant 0}$ converges a.s to a limit X_∞, and X_∞ is integrable.*

Proof. Considering $(-X_n)$ if necessary, we may suppose without loss of generality that $\mathbf{X} = (X_n)$ is a supermartingale.

Fix rationals $a < b$. Then by Doob's Upcrossing Lemma

$$\mathbb{E}[U_n([a, b], \mathbf{X})] \leqslant \frac{\mathbb{E}[(X_n - a)^-]}{b - a} \leqslant \frac{\mathbb{E}[|X_n|] + |a|}{b - a}.$$

Since $U_n(\cdots) \uparrow U(\cdots)$ as $n \to \infty$, by the Monotone Convergence Theorem

$$\mathbb{E}[U([a, b], \mathbf{X})] = \lim_{n\to\infty} \mathbb{E}[U_n([a, b], \mathbf{X})] \leqslant \frac{\sup_n \mathbb{E}[|X_n|] + |a|}{b - a} < \infty.$$

Hence $\mathbb{P}[U([a, b], \mathbf{X}) = \infty] = 0$. Since $\mathbb{Q}$ is countable, it follows that

$$\mathbb{P}\Big[\exists\, a, b \in \mathbb{Q},\ a < b,\ \text{s.t. } U([a, b], \mathbf{X}) = \infty\Big] = 0.$$

So by Lemma 7.3 $(X_n)_{n\geqslant 0}$ converges a.s. to some X_∞. (Specifically, we may take $X_\infty = \liminf X_n$, which is always defined, and measurable.) It remains to check that X_∞ is integrable. Since $|X_n| \to |X_\infty|$ a.s., Fatou's Lemma gives

$$\mathbb{E}[|X_\infty|] = \mathbb{E}\big[\liminf_{n\to\infty} |X_n|\big] \leqslant \liminf_{n\to\infty} \mathbb{E}[|X_n|] \leqslant \sup_n \mathbb{E}[|X_n|],$$

which is finite by assumption. □

Corollary 7.6. *If $(X_n)_{n\geqslant 0}$ is a non-negative supermartingale, then $X_\infty = \lim_{n\to\infty} X_n$ exists a.s.*

Proof. Since $\mathbb{E}[|X_n|] = \mathbb{E}[X_n] \leqslant \mathbb{E}[X_0]$ we may apply Theorem 7.5. □

Of course, the result holds for any supermartingale bounded below by a constant, and for any submartingale bounded above by a constant. The classic example of a non-negative supermartingale is your bankroll if you bet in a (realistic) casino, where all bets are at unfavourable (or, unrealistically, neutral) odds, and you can't bet more than you have. Here is another example.

Example 7.7 (Galton–Watson branching process)**.** Recall Definition 0.1: let X be a non-negative integer valued random variable with $0 < \mu = \mathbb{E}[X] < \infty$. Let $(X_{n,r})_{n,r\geqslant 1}$ be an array of i.i.d. random variables with the same distribution as X. Set $Z_0 = 1$ and

$$Z_{n+1} = \sum_{r=1}^{Z_n} X_{n+1,r},$$

so Z_n is the number of individuals in generation n of our branching process. Finally, let $M_n = Z_n/\mu^n$, and let $(\mathcal{F}_n)$ be the natural filtration associated to the sequence (Z_n) (or (M_n)). Then

$$\begin{aligned}
\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] &= \frac{1}{\mu^{n+1}}\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] \\
&= \frac{1}{\mu^{n+1}}\mathbb{E}\big[X_{n+1,1} + \cdots + X_{n+1,Z_n} \;\big|\; \mathcal{F}_n\big] \\
&= \frac{Z_n}{\mu^{n+1}}\mathbb{E}[X] = \frac{Z_n}{\mu^n} = M_n,
\end{aligned}$$

so $(M_n)_{n\geqslant 0}$ is a martingale w.r.t. $(\mathcal{F}_n)$.

[The above derivation is valid but somewhat informal. Since the process is discrete (so $\mathcal{F}_n$ corresponds to a countable partition of Ω), there is no problem making it precise using the Prelims definition of conditional expectation, which we have shown is the same as the general definition in this discrete case. As practice using the general definition of conditional expectation, here is a derivation using what we know about conditional expectation, without going back to the Prelims version. Here we take

$$\mathcal{F}_n = \sigma(\{X_{i,j} : i \leqslant n\})$$

since this turns out to be more convenient.

First note that $Z_{n+1} = \sum_{i=1}^{\infty} \mathbf{1}_{\{Z_n \geqslant i\}} X_{n+1,i}$; this is the standard way to handle a sum with a random number of terms. Thus, by cMON (which applies since everything is non-negative)

$$\begin{aligned}
\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] &= \sum_{i=1}^{\infty} \mathbb{E}[\mathbf{1}_{\{Z_n \geqslant i\}} X_{n+1,i} \mid \mathcal{F}_n] \text{ a.s.} \\
&= \sum_{i=1}^{\infty} \mathbf{1}_{\{Z_n \geqslant i\}} \mathbb{E}[X_{n+1,i} \mid \mathcal{F}_n] \text{ a.s.} \quad \text{(taking out what is known)} \\
&= \sum_{i=1}^{\infty} \mathbf{1}_{\{Z_n \geqslant i\}} \mathbb{E}[X_{n+1,i}] \text{ a.s.} \quad \text{(independence)} \\
&= \sum_{i=1}^{\infty} \mathbf{1}_{\{Z_n \geqslant i\}} \mu = Z_n \mu,
\end{aligned}$$

which gives $\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = M_n$ a.s.]

Since $(M_n)_{n\geqslant 0}$ is a non-negative martingale, by Corollary 7.6 we see that $(M_n)_{n\geqslant 0}$ converges a.s. to a finite limit M_∞. Does it converge in any other senses?

Recall that for a random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ (i.e., a measurable function) and a real number $p \geqslant 1$, the L^p norm of X is

$$||X||_p = \left(\int |X|^p \, \mathrm{d}\mathbb{P}\right)^{1/p} = \mathbb{E}\big[|X|^p\big]^{1/p},$$

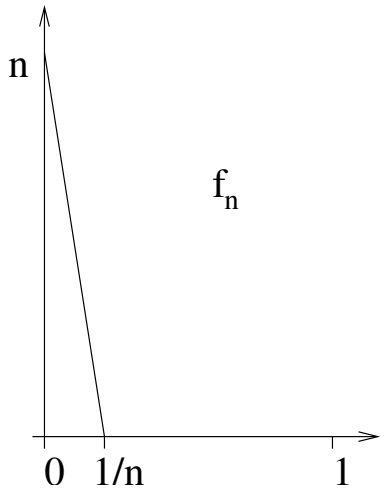
and that $L^p = L^p(\Omega, \mathcal{F}, \mathbb{P})$ denotes the set of all random variables X with $||X||_p$ finite, i.e., with $\mathbb{E}[|X|^p]$ finite. Recall (from Part A integration) that L^p is a vector space: the key point is that

$$||X + Y||_p \leqslant ||X||_p + ||Y||_p, \tag{22}$$

so the space L^p is closed under addition. Recall also that $X_n \to X$ in L^p if the sequence (X_n) converges to X in the space L^p, i.e., if $||X_n - X||_p \to 0$, or equivalently, if $\mathbb{E}[|X_n - X|^p] \to 0$, and that this convergence condition is stronger for larger p (see Corollary 3.18). Finally, note that if $X_n \to X$ in L^p then $||X_n||_p \to ||X||_p$, or, equivalently, $\mathbb{E}[|X_n|^p] \to \mathbb{E}[|X|^p]$; this follows from the triangle inequality (22). Of course, the reverse implication does not hold.

As we saw in §0.3, if $\mu \leqslant 1$ then $M_n \to 0$ with probability one. [We can now prove this in a different, rather clean way: by Corollary 7.6 the non-negative *super*martingale (Z_n) converges, i.e., is eventually constant. It is not hard to check that the constant can only be 0.] Hence $M_\infty = 0$ a.s., and $\mathbb{E}[M_\infty] = 0$, even though $\mathbb{E}[M_n] = 1$ for all n. It follows that M_n does not converge to M_∞ in L^1, or in L^p for any $p \geqslant 1$.

Convergence in L^1 will require a stronger condition. What is happening for our subcritical branching process is that although for large n, M_n is very likely to be zero, if it is *not* zero then it is very *big* with sufficiently high probability that $\mathbb{E}[M_n] \not\to 0$. This mirrors what we saw in Part A Integration with sequences like



for which we have a strict inequality in Fatou's Lemma. In §8 we will introduce a condition called 'uniform integrability' which is just enough to prohibit this sort of behaviour. First we consider another sort of boundedness.

7.1 Martingales bounded in L^2

Suppose that $(M_n)_{n\geqslant 0}$ is a square-integrable martingale, i.e., that $\mathbb{E}[M_n^2] < \infty$ for all n – we are *not* assuming that (M_n) is bounded in L^2. Adopting for the moment the ugly convention that $M_{-1} = 0$, for $k > j \geqslant 0$ we have

$$\begin{aligned}
\mathbb{E}[(M_k - M_{k-1})(M_j - M_{j-1})] &= \mathbb{E}\left[\mathbb{E}\left[(M_k - M_{k-1})(M_j - M_{j-1}) \mid \mathcal{F}_{k-1}\right]\right] \quad \text{(tower property)} \\
&= \mathbb{E}\left[(M_j - M_{j-1})\mathbb{E}[M_k - M_{k-1} \mid \mathcal{F}_{k-1}]\right] \quad \text{(taking out what is known)} \\
&= 0. \quad \text{(martingale property)}
\end{aligned}$$

This allows us to obtain a 'Pythagoras rule':

$$\begin{aligned}\mathbb{E}[M_n^2] &= \mathbb{E}\left[\left(\sum_{k=0}^{n}(M_k - M_{k-1})\right)^2\right] \\ &= \sum_{k=0}^{n}\mathbb{E}[(M_k - M_{k-1})^2] + 2\sum_{n\geqslant k>j\geqslant 0}\mathbb{E}[(M_k - M_{k-1})(M_j - M_{j-1})] \\ &= \mathbb{E}[M_0^2] + \sum_{k=1}^{n}\mathbb{E}[(M_k - M_{k-1})^2]. \end{aligned} \tag{23}$$

Similarly (using again only that the cross terms in the expanded sum have zero expectation), for $m > n \geqslant 0$ we have

$$\mathbb{E}[(M_m - M_n)^2] = \sum_{k=n+1}^{m}\mathbb{E}[(M_k - M_{k-1})^2]. \tag{24}$$

Lemma 7.8. *Let $(M_n)_{n\geqslant 0}$ be a martingale. Then $(M_n)_{n\geqslant 0}$ is bounded in L^2 if and only if*

$$\mathbb{E}[M_0^2] < \infty \quad \textit{and} \quad \sum_{k\geqslant 1}\mathbb{E}[(M_k - M_{k-1})^2] < \infty. \tag{25}$$

Proof. This follows easily from (23). (Starting from assumption (25), before we can apply (23) we must first check that each $\mathbb{E}[M_n^2]$ is finite. But this follows from (25) and the fact that L^2 is closed under addition.) □

Theorem 7.9. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\mathcal{F}_n)_{n\geqslant 0}$ a filtration and $(M_n)_{n\geqslant 0}$ a martingale w.r.t. $(\mathcal{F}_n)_{n\geqslant 0}$ that is bounded in L^2. Then there is a random variable M_∞ such that $M_n \to M_\infty$ a.s. and*

$$\lim_{n\to\infty}\mathbb{E}[(M_n - M_\infty)^2] = 0,$$

that is $M_n \to M_\infty$ in L^2.

Proof. From Jensen's inequality (since $f(x) = x^2$ is convex)

$$\mathbb{E}[|M_n|]^2 \leqslant \mathbb{E}[M_n^2],$$

so since (M_n) is bounded in L^2, it is bounded in L^1. Hence Doob's Forward Convergence Theorem shows that $M_\infty = \lim_{n\to\infty} M_n$ exists a.s. To check convergence in L^2 we use Pythagoras: from (24),

$$\mathbb{E}[(M_{n+k} - M_n)^2] = \sum_{j=n+1}^{n+k}\mathbb{E}[(M_j - M_{j-1})^2], \tag{26}$$

and so by Fatou's Lemma

$$\begin{aligned}\mathbb{E}[(M_\infty - M_n)^2] &= \mathbb{E}\Big[\liminf_{k\to\infty}(M_{n+k} - M_n)^2\Big] \\ &\leqslant \liminf_{k\to\infty}\mathbb{E}[(M_{n+k} - M_n)^2] \\ &= \sum_{j=n+1}^{\infty}\mathbb{E}[(M_j - M_{j-1})^2]. \quad \text{(using (26))}\end{aligned}$$

The final bound is the tail of a sum that, by Lemma 7.8, is convergent. Hence, as $n \to \infty$, this bound tends to 0. □

For L^2 convergence we have a very nice result: a sequence can only converge in L^2 (or in any normed space) if it is bounded in L^2 (in that space), so we have shown that a martingale converges in L^2 exactly when we could hope that it might.

Notice that martingales that are bounded in L^2 form a strict subset of those that are bounded in L^1 (those for which we proved Doob's Forwards Convergence Theorem). And convergence in L^2 implies convergence in L^1, so for these martingales we don't have the difficulty we had with our branching process example. L^2-boundedness is often relatively straightforward to check, so is convenient, but it is a stronger condition than we *need* for L^1-convergence.

8 Uniform Integrability

If X is an integrable random variable (that is $\mathbb{E}[|X|] < \infty$), then the decreasing function $\mathbb{E}[|X|\mathbf{1}_{\{|X|>K\}}]$ tends to 0 as $K \to \infty$. Indeed, setting $f_n = |X|\mathbf{1}_{\{|X|>n\}}$, the functions f_n converge to 0 a.s., and are dominated by the integrable function $|X|$. So by the Dominated Convergence Theorem, $\mathbb{E}[f_n] \to 0$. Uniform integrability demands that this property holds *uniformly* for random variables from some class.

Definition 8.1 (Uniform Integrability)**.** A collection $\mathcal{C}$ of random variables is called *uniformly integrable* if for every $\varepsilon > 0$ there exists a K such that

$$\mathbb{E}[|X|\mathbf{1}_{\{|X|>K\}}] < \varepsilon \qquad \text{for all } X \in \mathcal{C}.$$

Note that, unsurprisingly, the singleton family $\{X\}$ is uniformly integrable if and only if X is integrable. Uniform integrability says essentially that the 'upper tail' of the integrals tends to zero uniformly; the following formulation is sometimes more convenient.

Proposition 8.2. *A collection $\mathcal{C}$ is uniformly integrable if and only if $\mathbb{E}[(|X| - K)^+] \to 0$ as $K \to \infty$, uniformly in $X \in \mathcal{C}$, i.e., for every $\varepsilon > 0$ there exists a K such that $\mathbb{E}[(|X| - K)^+] < \varepsilon$ for all $X \in \mathcal{C}$.*

Proof. The forward implication is immediate since $0 \leqslant (|X| - K)^+ \leqslant |X|\mathbf{1}_{\{|X|>K\}}$. For the reverse, note that

$$|X|\mathbf{1}_{\{|X|>2K\}} \leqslant 2(|X| - K)^+.$$

□

There are two reasons why uniform integrability is important:

1. For sequences that converge in probability (or a.s.), uniform integrability is necessary and sufficient for passing to the limit under an expectation,
2. it is often easy to verify in the context of martingale theory.

Property 1 should be sufficient to guarantee that uniform integrability is interesting, but in fact uniform integrability is not often used in analysis where it is usually simpler to use the Monotone or Dominated Convergence Theorem. It more commonly used in probability, and one reason is 2 above.

Proposition 8.3. *Suppose that $\{X_\alpha, \alpha \in I\}$ is a uniformly integrable family of random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then*

1.

$$\sup_\alpha \mathbb{E}[|X_\alpha|] < \infty,$$

2.

$$\mathbb{P}[|X_\alpha| > K] \to 0 \quad \textit{as } K \to \infty, \text{ uniformly } \textit{in } \alpha,$$

i.e., $\forall \varepsilon > 0 \quad \exists K \quad \forall \alpha \in I : \mathbb{P}[|X_\alpha| > K] < \varepsilon.$

3.

$$\mathbb{E}[|X_\alpha| \mathbf{1}_A] \to 0 \quad \textit{as } \mathbb{P}[A] \to 0, \text{ uniformly } \textit{in } \alpha.$$

i.e., $\forall \varepsilon > 0 \quad \exists \delta > 0 \quad \forall \alpha \in I : \mathbb{P}[A] < \delta \implies \mathbb{E}[|X_\alpha| \mathbf{1}_A] < \varepsilon.$

Conversely, either 1 and *3 or 2* and *3 implies uniform integrability.*

Proof. 1. By definition of uniform integrability, there exists K such that for all α

$$\mathbb{E}[|X_\alpha| \mathbf{1}_{\{|X_\alpha| > K\}}] \leqslant 1.$$

Then for all α

$$\mathbb{E}[|X_\alpha|] = \mathbb{E}\left[|X_\alpha| \mathbf{1}_{\{|X_\alpha| \leqslant K\}} + |X_\alpha| \mathbf{1}_{\{|X_\alpha| > K\}}\right] \leqslant K + \mathbb{E}\left[|X_\alpha| \mathbf{1}_{\{|X_\alpha| > K\}}\right] \leqslant K + 1.$$

Now 1 implies 2 since

$$\begin{aligned} \mathbb{P}[|X_\alpha| > K] &\leqslant \frac{1}{K} \mathbb{E}[|X_\alpha|] \quad \text{(Markov)} \\ &\leqslant \frac{1}{K} \sup_\beta \mathbb{E}[|X_\beta|] \end{aligned}$$

and the final bound, which evidently tends to zero as $K \to \infty$, is independent of α.

To see 3, fix $\varepsilon > 0$ and choose K such that

$$\mathbb{E}[|X_\alpha| \mathbf{1}_{\{|X_\alpha| > K\}}] < \frac{\varepsilon}{2} \quad \text{for all } \alpha.$$

Set $\delta = \varepsilon/(2K)$ and suppose that $\mathbb{P}[A] < \delta$. Then for any α,

$$\begin{aligned} \mathbb{E}[|X_\alpha| \mathbf{1}_A] &= \mathbb{E}[|X_\alpha| \mathbf{1}_A \mathbf{1}_{\{|X_\alpha| > K\}}] + \mathbb{E}[|X_\alpha| \mathbf{1}_A \mathbf{1}_{\{|X_\alpha| \leqslant K\}}] \\ &\leqslant \mathbb{E}[|X_\alpha| \mathbf{1}_{\{|X_\alpha| > K\}}] + \mathbb{E}[K \mathbf{1}_A] \\ &\leqslant \frac{\varepsilon}{2} + K \mathbb{P}[A] \\ &< \varepsilon. \end{aligned}$$

For the converse, since 1 implies 2, it is enough to check that 2 and 3 imply uniform integrability. Let $\varepsilon > 0$ be given. By 3 there exists $\delta > 0$ such that $\mathbb{P}[A] < \delta$ implies $\mathbb{E}[|X_\alpha| \mathbf{1}_A] < \varepsilon$ for all α. By 2 there is a K such that $\mathbb{P}[|X_\alpha| > K] < \delta$ for all α. But then

$$\mathbb{E}\left[|X_\alpha| \mathbf{1}_{\{|X_\alpha| > K\}}\right] < \varepsilon \quad \text{for all } \alpha.$$

□

[If we impose a very minor technical condition on our probability space, namely that it is *atomless*, then 3 on its own implies uniform integrability. So 'morally' 3 is really equivalent to uniform integrability, and is often the best way of thinking about it.]

Recall that for a sequence of random variables (X_n) on $(\Omega, \mathcal{F}, \mathbb{P})$ we say that $X_n \to X$ in L^1 if

$$\mathbb{E}[|X_n - X|] \to 0 \quad \text{as } n \to \infty.$$

We say that $X_n \to X$ in probability if, for every $\varepsilon > 0$,

$$\mathbb{P}[|X_n - X| \geqslant \varepsilon] \to 0$$

as $n \to \infty$. Finally, $X_n \to X$ a.s. if $\mathbb{P}[X_n \to X] = 1$. Recall that either of convergence a.s. or in L^1 implies convergence in probability, but not the reverse, essentially because of the possible contribution of rare but very large values.

The following is a variant of the Bounded Convergence Theorem that you may or may not have seen in Part A integration; for us, it is just a warm up to the next result.

Lemma 8.4. *Let (X_n) be a sequence of random variables with $X_n \to X$ in probability, and suppose that X and all X_n are bounded by the same real number K. Then $X_n \to X$ in L^1.*

Proof. We use an idea which recurs again and again in this context: split by whether the relevant quantity is 'small' or 'large'. Specifically, fix $\varepsilon > 0$. Let A_n be the event $\{|X_n - X| > \varepsilon\}$. Then

$$\begin{aligned} \mathbb{E}[|X_n - X|] &= \mathbb{E}\big[|X_n - X|\mathbf{1}_{A_n} + |X_n - X|\mathbf{1}_{A_n^c}\big] \\ &\leqslant \mathbb{E}[|X_n|\mathbf{1}_{A_n}] + \mathbb{E}[|X|\mathbf{1}_{A_n}] + \varepsilon \\ &\leqslant 2\mathbb{E}[K\mathbf{1}_{A_n}] + \varepsilon = 2K\mathbb{P}[A_n] + \varepsilon. \end{aligned} \tag{27}$$

Since X_n converges to X in probability, $\mathbb{P}[A_n] \to 0$, so the bound above is at most 2ε if n is large enough, and $\mathbb{E}[|X_n - X|] \to 0$ as required. □

The next result extends this to the situation when the (X_n) are uniformly integrable. This is the *right* condition: $X_n \to X$ in L^1 *if and only if* $X_n \to X$ in probability and (X_n) is uniformly integrable.

Theorem 8.5 (Vitali's Convergence Theorem)**.** *Let (X_n) be a sequence of integrable random variables which converges in probability to a random variable X. TFAE (The Following Are Equivalent):*

1. *the family $\{X_n\}$ is uniformly integrable,*
2. *$\mathbb{E}[|X_n - X|] \to 0$ as $n \to \infty$,*
3. *$\mathbb{E}[|X_n|] \to \mathbb{E}[|X|] < \infty$ as $n \to \infty$.*

Proof. Suppose 1 holds. We try to repeat the proof of Lemma 8.4, using the bound (27). Proposition 8.3 will tell us that the first term tends to 0: surprisingly, the hardest part is to deal with the second term, which requires us to show that X is integrable.

Since $|X_n| \to |X|$ in probability, by Theorem 3.10 there exists a subsequence $(X_{n_k})_{k \geqslant 1}$ that converges to X a.s. Fatou's Lemma gives

$$\mathbb{E}[|X|] \leqslant \liminf_{k \to \infty} \mathbb{E}[|X_{n_k}|] \leqslant \sup_n \mathbb{E}[|X_n|],$$

which is finite by Proposition 8.3. Thus X is integrable. Now fix $\varepsilon > 0$, and let A_n be the event $\{|X_n - X| > \varepsilon\}$. Then, as before,

$$\begin{aligned} \mathbb{E}[|X_n - X|] &= \mathbb{E}\big[|X_n - X|\mathbf{1}_{A_n}\big] + \mathbb{E}\big[|X_n - X|\mathbf{1}_{A_n^c}\big] \\ &\leqslant \mathbb{E}\big[|X_n|\mathbf{1}_{A_n}\big] + \mathbb{E}\big[|X|\mathbf{1}_{A_n}\big] + \varepsilon. \end{aligned}$$

Since $X_n \to X$ in probability we have $\mathbb{P}[A_n] \to 0$ as $n \to \infty$, so by uniform integrability and Proposition 8.3,

$$\mathbb{E}[|X_n|\mathbf{1}_{A_n}] \to 0 \quad \text{as } n \to \infty.$$

Similarly, since $\{X\}$ is uniformly integrable,

$$\mathbb{E}[|X|\mathbf{1}_{A_n}] \to 0 \quad \text{as } n \to \infty.$$

Hence $\mathbb{E}[|X_n - X|] \leqslant 2\varepsilon$ for n large enough. Since $\varepsilon > 0$ was arbitrary this proves 2.

For 2 implies 3 we are implicitly assuming that $X_n - X$ is integrable, so X is integrable. Then, as noted earlier, our assumption $||X_n - X||_1 \to 0$ implies that $||X_n||_1 \to ||X||_1$.

It remains to show that 3 implies 1. To avoid clutter, let $Y_n = |X_n|$ and $Y = |X|$, noting that $Y_n, Y \geqslant 0$, $Y_n \xrightarrow{\mathbb{P}} Y$, and by assumption $\mathbb{E}[Y_n] \to \mathbb{E}[Y] < \infty$.

Let $\varepsilon > 0$ be given. Then, since $\{Y\}$ (or $\{X\}$ – it makes no difference) is uniformly integrable, there is some K such that

$$\mathbb{E}[(Y - K)^+] < \varepsilon.$$

For any random variable Z, define

$$Z \wedge K = \begin{cases} Z & Z \leqslant K, \\ K & Z > K, \end{cases}$$

noting that

$$(Z - K)^+ = Z - (Z \wedge K). \tag{28}$$

Since $|(a \wedge K) - (b \wedge K)| \leqslant |a - b|$, we have $Y_n \wedge K \xrightarrow{\mathbb{P}} Y \wedge K$, so by Lemma 8.4 $\mathbb{E}[Y_n \wedge K] \to \mathbb{E}[Y \wedge K]$. Since $\mathbb{E}[Y_n] \to \mathbb{E}[Y]$, using (28) for Y_n and Y it follows that

$$\mathbb{E}[(Y_n - K)^+] \to \mathbb{E}[(Y - K)^+] < \varepsilon.$$

Hence there is an n_0 such that for $n \geqslant n_0$,

$$\mathbb{E}[(|X_n| - K)^+] = \mathbb{E}[(Y_n - K)^+] < 2\varepsilon.$$

There are only finitely many $n < n_0$, so there exists $K' \geqslant K$ such that such that

$$\mathbb{E}[(|X_n| - K')^+] < 2\varepsilon$$

for *all* n, as required. □

We will use the next result to show that one of the basic constructions of a martingale gives something that is automatically uniformly integrable.

Theorem 8.6. *Let X be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and $\{\mathcal{F}_\alpha : \alpha \in I\}$ a family of σ-algebras with $\mathcal{F}_\alpha \subseteq \mathcal{F}$. Then the family $\{X_\alpha : \alpha \in I\}$ with*

$$X_\alpha = \mathbb{E}[X \mid \mathcal{F}_\alpha]$$

is uniformly integrable.

An important special case is when $(\mathcal{F}_n)$ is a filtration, in which case (X_n) is a martingale; see Example 5.10.

Proof. [5] Since $f(x) = |x|$ is convex, by the conditional form of Jensen's inequality (Proposition 4.11),

$$|X_\alpha| = |\mathbb{E}[X \mid \mathcal{F}_\alpha]| \leqslant \mathbb{E}\big[|X| \mid \mathcal{F}_\alpha\big] \text{ a.s.} \tag{29}$$

[5] An alternative strategy is to deal with the case $X \geqslant 0$ first, and handle the general case by writing $X = X^+ - X^-$.

which certainly implies that

$$\mathbb{E}[|X_\alpha|] \leqslant \mathbb{E}[|X|].$$

Also, using (29),

$$\mathbb{E}[|X_\alpha|\mathbf{1}_{\{|X_\alpha|>K\}}] \leqslant \mathbb{E}\left[\mathbb{E}[|X| \mid \mathcal{F}_\alpha]\mathbf{1}_{\{|X_\alpha|>K\}}\right] = \mathbb{E}[|X|\mathbf{1}_{\{|X_\alpha|>K\}}], \tag{30}$$

since we may move the indicator function inside the conditional expectation and then apply the tower law.

Now the single integrable random variable X forms on its own a uniformly integrable family and so by Proposition 8.3 given $\varepsilon > 0$ we can find $\delta > 0$ such that $\mathbb{P}[A] < \delta$ implies $\mathbb{E}[|X|\mathbf{1}_A] < \varepsilon$. Since

$$\mathbb{P}[|X_\alpha| \geqslant K] \leqslant \frac{\mathbb{E}[|X_\alpha|]}{K} \leqslant \frac{\mathbb{E}[|X|]}{K},$$

setting $K = 2\mathbb{E}[|X|]/\delta < \infty$, it follows that $\mathbb{E}[|X_\alpha|\mathbf{1}_{\{|X_\alpha|>K\}}] < \varepsilon$ for every α. □

Theorem 8.7. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\mathcal{F}_n)_{n\geqslant 0}$ a filtration and $(M_n)_{n\geqslant 0}$ a martingale w.r.t. $(\mathcal{F}_n)_{n\geqslant 0}$. TFAE*

1. *$(M_n)_{n\geqslant 0}$ is uniformly integrable,*
2. *there is some M_∞ such that $M_n \to M_\infty$ almost surely and in L^1,*
3. *there is an integrable M_∞ such that $M_n = \mathbb{E}[M_\infty \mid \mathcal{F}_n]$ a.s. for all n.*

Proof. We already have most of the ingredients.

1 $\implies$ 2: If $(M_n)_{n\geqslant 0}$ is uniformly integrable then it is bounded in L^1 and so by Doob's Forward Convergence Theorem (Theorem 7.5) it converges a.s. to some M_∞. Since a.s. convergence implies convergence in probability, $M_n \to M_\infty$ in L^1 by Theorem 8.5.

2 $\implies$ 3: Since (M_n) is a martingale, for $m \geqslant n$ we have

$$\mathbb{E}[M_m \mid \mathcal{F}_n] = M_n \quad \text{a.s.},$$

so for any $A \in \mathcal{F}_n$ we have[6]

$$\mathbb{E}[M_m\mathbf{1}_A] = \mathbb{E}[M_n\mathbf{1}_A].$$

Since

$$\left|\mathbb{E}[M_\infty\mathbf{1}_A] - \mathbb{E}[M_m\mathbf{1}_A]\right| \leqslant \mathbb{E}[|(M_\infty - M_m)\mathbf{1}_A|] \leqslant \mathbb{E}[|M_\infty - M_m|] \to 0,$$

it follows that

$$\mathbb{E}[M_\infty\mathbf{1}_A] = \mathbb{E}[M_n\mathbf{1}_A] \quad \text{for all } A \in \mathcal{F}_n.$$

Since M_n is $\mathcal{F}_n$-measurable, this shows that $M_n = \mathbb{E}[M_\infty \mid \mathcal{F}_n]$ a.s. by definition of conditional expectation.

3 $\implies$ 1 by Theorem 8.6. □

Finally we record a version of the Optional Stopping Theorem 6.4 for uniformly integrable martingales, which applies for *any* stopping time.

Theorem 8.8. *Let M be a uniformly integrable martingale. Let τ be a stopping time (we allow τ to take the value ∞ with positive probability). Then M_τ is integrable and $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$.*

[6] Recall that $\mathbb{E}[X\mathbf{1}_A]$ and $\int_A X\,\mathrm{d}\mathbb{P}$ mean the same thing.

Proof. The function $x \mapsto |x|$ is convex, so $(|M_n|)$ is a submartingale by Proposition 5.11. Then Lemma 6.7 gives that for any n,

$$\mathbb{E}[|M_{\tau\wedge n}|] \leqslant \mathbb{E}[|M_n|] \leqslant \sup_k \mathbb{E}[|M_k|] < \infty,$$

so $(M_{\tau\wedge n})$ is a martingale bounded in $\mathcal{L}^1$. It converges to M_τ, which by Theorem 7.5 is integrable.

We want to show that $(M_{\tau\wedge n})$ is in fact uniformly integrable. For any K,

$$\mathbb{E}[|M_{\tau\wedge n}|\mathbf{1}_{|M_{\tau\wedge n}|>K}] \leqslant \mathbb{E}[|M_\tau|\mathbf{1}_{|M_\tau|>K}] + \mathbb{E}[|M_n|\mathbf{1}_{|M_n|>K},$$

since $M_{\tau\wedge n}$ is equal to either M_τ or M_n. Now the RHS goes to 0 as $K \to \infty$, uniformly in n, since (for the first term) M_τ is integrable, and (for the second term) (M_n) is uniformly integrable. So indeed $(M_{\tau\wedge n})$ is uniformly integrable.

Hence in fact (Theorem 8.7) $M_{\tau\wedge n} \to M_\tau$ in $\mathcal{L}^1$. Then we have $\mathbb{E}[M_\tau] = \lim_{n\to\infty} \mathbb{E}[M_{\tau\wedge n}] = \lim_{n\to\infty} \mathbb{E}[M_0] = \mathbb{E}[M_0]$ as required. □

9 Backwards Martingales and the Strong Law of Large Numbers

So far our martingales were sequences (M_t) of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ defined for all integers $t \geqslant 0$. But in fact the definition makes just as good sense for any 'interval' I of integers. The conditions are that for every $t \in I$ we have a σ-algebra $\mathcal{F}_t \subseteq \mathcal{F}$ (information known at time t) and an integrable, $\mathcal{F}_t$-measurable random variable M_t, with $\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] = M_t$ a.s. Note that we already implicitly considered the finite case $I = \{0, 1, 2, \ldots, N\}$.

Backwards martingales are martingales for which time is indexed by $I = \{t \in \mathbb{Z} : t \leqslant 0\}$. The main difficulty is deciding whether to write $(M_n)_{n\leqslant 0}$ or $(M_{-n})_{n\geqslant 0}$. From now on we write the latter. Note that a backwards martingale *ends* at time 0.

Definition 9.1. Given σ-algebras $(\mathcal{F}_{-n})_{n\geqslant 0}$ with $\mathcal{F}_{-n} \subseteq \mathcal{F}$ and

$$\cdots \subseteq \mathcal{F}_{-(n+1)} \subseteq \mathcal{F}_{-n} \subseteq \cdots \subseteq \mathcal{F}_{-2} \subseteq \mathcal{F}_{-1} \subseteq \mathcal{F}_0,$$

a *backwards martingale* w.r.t. $(\mathcal{F}_{-n})$ is a sequence $(M_{-n})_{n\geqslant 0}$ of integrable random variables with M_{-n} $\mathcal{F}_{-n}$-measurable and

$$\mathbb{E}[M_{-n+1} \mid \mathcal{F}_{-n}] = M_{-n} \quad \text{a.s.}$$

for all $n \geqslant 1$.

For any backwards martingale, we have

$$\mathbb{E}[M_0 \mid \mathcal{F}_{-n}] = M_{-n} \quad \text{a.s.}$$

Since M_0 is integrable, it follows from Theorem 8.6 that $(M_{-n})_{n\geqslant 0}$ is *automatically* uniformly integrable.

Doob's Upcrossing Lemma, a result about *finite* martingales, shows that if $U_m([a,b], \mathbf{M})$ is the number of upcrossings of $[a,b]$ by a backwards martingale between times $-m$ and 0, then

$$\mathbb{E}[U_m([a,b], \mathbf{M})] \leqslant \frac{\mathbb{E}[(M_0 - a)^-]}{b-a}. \tag{31}$$

(Simply consider the finite martingale $(M_{-m}, M_{-m+1}, \ldots, M_{-1}, M_0)$.) A minor variant of the proof of Doob's Forward Convergence Theorem (Theorem 7.5) then shows that as $n \to \infty$, M_{-n} converges a.s. to a random limit $M_{-\infty}$. (For definiteness, say $M_{-\infty} = \liminf_{n\to\infty} M_{-n}$.) Let

$$\mathcal{F}_{-\infty} = \bigcap_{k=0}^{\infty} \mathcal{F}_{-k},$$

noting that as k increases, the σ-algebras *decrease*. The limit $M_{-\infty}$ is $\mathcal{F}_{-k}$-measurable for every k (since M_{-n} is for all $n \geqslant k$), so $M_{-\infty}$ is $\mathcal{F}_{-\infty}$-measurable. Since (M_{-n}) is uniformly integrable, adapting the proof of Theorem 8.7 gives the following result.

Theorem 9.2. *Let $(M_{-n})_{n\geqslant 0}$ be a backwards martingale w.r.t. $(\mathcal{F}_{-n})_{n\geqslant 0}$. Then M_{-n} converges a.s. and in L^1 as $n \to \infty$ to the random variable $M_{-\infty} = \mathbb{E}[M_0 \mid \mathcal{F}_{-\infty}]$.*

We now use this result to prove the celebrated Kolmogorov Strong Law.

Theorem 9.3 (Kolmogorov's Strong Law of Large Numbers)**.** *Let $(X_n)_{n\geqslant 1}$ be a sequence of i.i.d. random variables each of which is integrable and has mean μ, and set*

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

Then

$$\frac{S_n}{n} \to \mu \text{ a.s. and in } L^1 \text{ as } n \to \infty.$$

Proof. For $n \geqslant 1$ set

$$\mathcal{F}_{-n} = \sigma(S_n, S_{n+1}, S_{n+2}, \ldots) = \sigma(S_n, X_{n+1}, X_{n+2}, \ldots),$$

noting that $\mathcal{F}_{-n-1} \subseteq \mathcal{F}_{-n}$. Conditioning on $\mathcal{F}_{-n}$ preserves the symmetry between $X_1, \ldots, X_n$, since none of $S_n, S_{n+1}, \ldots$ is affected by permuting $X_1, \ldots, X_n$. Hence,

$$\mathbb{E}[X_1 \mid \mathcal{F}_{-n}] = \mathbb{E}[X_2 \mid \mathcal{F}_{-n}] = \cdots = \mathbb{E}[X_n \mid \mathcal{F}_{-n}]$$

and so they are all equal (a.s.) to their average

$$\frac{1}{n}\mathbb{E}[X_1 + \cdots + X_n \mid \mathcal{F}_{-n}] = \frac{1}{n}\mathbb{E}[S_n \mid \mathcal{F}_{-n}] = \frac{1}{n}S_n.$$

Let $M_{-n} = S_n/n$. Then, for $n \geqslant 2$,

$$\mathbb{E}[M_{-n+1} \mid \mathcal{F}_{-n}] = \mathbb{E}[S_{n-1}/(n-1) \mid \mathcal{F}_{-n}] = \frac{1}{n-1}\sum_{i=1}^{n-1} \mathbb{E}[X_i \mid \mathcal{F}_{-n}] = S_n/n = M_n.$$

In other words, $(M_{-n})_{n\geqslant 1}$ is a backwards martingale w.r.t. $(\mathcal{F}_{-n})_{n\geqslant 1}$. Thus, by Theorem 9.2, S_n/n converges a.s. and in L^1 to $M_{-\infty} = \mathbb{E}[M_{-1} \mid \mathcal{F}_{-\infty}]$, where $\mathcal{F}_{-\infty} = \bigcap_{k\geqslant 1} \mathcal{F}_{-k}$.

Now by L^1 convergence, $\mathbb{E}[M_{-\infty}] = \lim_{n\to\infty} \mathbb{E}[M_{-n}] = \mathbb{E}[M_{-1}] = \mathbb{E}[S_1] = \mu$. In terms of the random variables $X_1, X_2, \ldots$, the limit $M_{-\infty} = \liminf S_n/n$ is a tail random variable, so by Kolmogorov's 0-1 law it is a.s. constant, so $M_{-\infty} = \mu$ a.s. □

9.1 Exchangeability and the ballot theorem (not covered in lectures)

The material in Section 9.1 is not part of the "examinable syllabus". You won't be asked to reproduce these results directly. However, just like many of the problem sheet questions, the methods help to develop your intuition for the ideas of the course.

In our proof of the Strong Law of Large Numbers we used symmetry in a key way. There it followed from independence of our random variables, but in general a weaker condition suffices.

Definition 9.4 (Exchangeability)**.** The random variables $X_1, \ldots, X_n$ are said to be *exchangeable* if the vector $(X_{i_1}, \ldots, X_{i_n})$ has the same probability distribution for every permutation $i_1, \ldots, i_n$ of $1, \ldots, n$.

Example 9.5. Let $X_1, \ldots, X_n$ be the results of n successive samples *without replacement* from a pool of at least n values (some of which may be the same). Then the random variables $X_1, \ldots, X_n$ are exchangeable but *not* independent.

It turns out that we can use the construction in the proof of the Strong Law of Large Numbers to manufacture a finite martingale from a finite collection of exchangeable random variables. Suppose that $X_1, \ldots, X_n$ are exchangeable and integrable, and set $S_j = \sum_{i=1}^{j} X_i$. Let

$$Z_j = \mathbb{E}[X_1 \mid \sigma(S_{n+1-j}, \ldots, S_{n-1}, S_n)], \quad j = 1, 2, \ldots n.$$

Note that Z_j is defined by conditioning on the last j sums; since we condition on more as j increases, $(Z_j)_{j=1}^n$ is certainly a martingale. Now

$$\begin{aligned} S_{n+1-j} &= \mathbb{E}[S_{n+1-j} \mid \sigma(S_{n+1-j}, \ldots, S_n)] \\ &= \sum_{i=1}^{n+j-1} \mathbb{E}[X_i \mid \sigma(S_{n+1-j}, \ldots, S_n)] \\ &= (n+1-j)\mathbb{E}[X_1 \mid \sigma(S_{n+1-j}, \ldots, S_n)] \quad \text{(by exchangeability)} \\ &= (n+1-j)Z_j, \end{aligned}$$

so $Z_j = S_{n+1-j}/(n+1-j)$.

Definition 9.6. The martingale

$$Z_j = \frac{S_{n+1-j}}{n+1-j}, \quad j = 1, 2, \ldots, n,$$

is sometimes called a *Doob backward martingale*.

Example 9.7 (The ballot problem)**.** In an election between candidates A and B, candidate A receives n votes and candidate B receives m votes, where $n > m$. Assuming that in the count of votes all orderings are equally likely, what is the probability that A is always ahead of B during the count?

Solution:

Let $X_i = 1$ if the ith vote counted is for A and -1 if the ith vote counted is for B, and let $S_k = \sum_{i=1}^{k} X_i$. Because all orderings of the $n+m$ votes are equally likely, $X_1, \ldots, X_{n+m}$ are exchangeable, so

$$Z_j = \frac{S_{n+m+1-j}}{n+m+1-j}, \quad j = 1, 2, \ldots, n+m,$$

is a Doob backward martingale.

Because

$$Z_1 = \frac{S_{n+m}}{n+m} = \frac{n-m}{n+m},$$

the mean of this martingale is $(n-m)/(n+m)$.

Because $n > m$, either (i) A is always ahead in the count, or (ii) there is a tie at some point. Case (ii) happens if and only if some $S_j = 0$, i.e., if and only if some $Z_j = 0$.

Define the bounded stopping time τ by

$$\tau = \min\{j \geqslant 1 : Z_j = 0 \text{ or } j = n+m\}.$$

In case (i), $Z_\tau = Z_{n+m} = X_1 = 1$. (If A is always ahead, he must receive the first vote.) Clearly, in case (ii), $Z_\tau = 0$, so

$$Z_\tau = \begin{cases} 1 & \text{if } A \text{ is always ahead,} \\ 0 & \text{otherwise.} \end{cases}$$

By Lemma 6.2 (or Theorem 6.4), $\mathbb{E}[Z_\tau] = (n-m)/(n+m)$ and so

$$\mathbb{P}[A \text{ is always ahead}] = \frac{n-m}{n+m}.$$

□

10 Azuma-Hoeffding inequality and concentration of Lipschitz functions

The material in Section 10 is not part of the "examinable syllabus". You won't be asked to reproduce any of these results directly. However, the methods involved are very good illustrations of ideas from earlier in the course: particularly the Doob martingale ideas involved in Theorem 10.5 and its applications.

By applying Markov's inequality to the moment generating function, we can get better bounds than we get from the mean and variance alone.

Lemma 10.1. *(i) Let Y be a random variable with mean 0, taking values in $[-c, c]$. Then*

$$\mathbb{E}[e^{\theta Y}] \leqslant \exp\left(\frac{1}{2}\theta^2 c^2\right).$$

(ii) Let $\mathcal{G}$ be a σ-algebra, and Y be a random variable with $\mathbb{E}[Y|\mathcal{G}] = 0$ a.s. and $Y \in [-c, c]$ a.s. Then

$$\mathbb{E}[e^{\theta Y} \mid \mathcal{G}] \leqslant \exp\left(\frac{1}{2}\theta^2 c^2\right) \text{ a.s.}$$

Proof. Let $f(y) = e^{\theta y}$. Since f is convex,

$$f(y) \leqslant \frac{c-y}{2c} f(-c) + \frac{c+y}{2c} f(c)$$

for all $y \in [-c, c]$. Then taking expectations,

$$\begin{aligned} \mathbb{E}[f(Y)] &\leqslant \mathbb{E}\left[\frac{c-Y}{2c} f(-c) + \frac{c+Y}{2c} f(c)\right] \\ &= \frac{1}{2} f(-c) + \frac{1}{2} f(c) \\ &= \frac{e^{-\theta c} + e^{\theta c}}{2}. \end{aligned}$$

Now, comparing Taylor expansions term by term,

$$\frac{e^{-\theta c} + e^{\theta c}}{2} = \sum_{n=0}^{\infty} \frac{(\theta c)^{2n}}{(2n)!} \leqslant \sum_{n=0}^{\infty} \frac{(\theta c)^{2n}}{2^n n!} = \exp\left(\frac{1}{2}\theta^2 c^2\right).$$

giving part (i).

For the conditional version of the statement, consider any $G \in \mathcal{G}$ with $\mathbb{P}[G] > 0$. Then $\mathbb{E}[Y\mathbf{1}_G] = 0$, so $\mathbb{E}[Y \mid G] = 0$. Applying part (i) with probability measure $\mathbb{P}[. \mid G]$, we obtain $\mathbb{E}[e^{\theta Y} \mid G] \leqslant \exp\left(\frac{1}{2}\theta^2 c^2\right)$.

Now consider the G-measurable set $G := \{\omega : \mathbb{E}[e^{\theta Y}|\mathcal{G}](\omega) > \exp\left(\frac{1}{2}\theta^2 c^2\right)\}$. If this set has positive probability, it contradicts the previous paragraph. So indeed $\mathbb{E}[e^{\theta Y} \mid \mathcal{G}] \leqslant \exp\left(\frac{1}{2}\theta^2 c^2\right)$ a.s. as required. □

Lemma 10.2. *Suppose M is a martingale with $M_0 = 0$ and $|M_n - M_{n-1}| \leqslant c$ a.s. for all n. Then*

$$\mathbb{E}\left[e^{\theta M_n}\right] \leqslant \exp\left(\frac{1}{2}\theta^2 c^2 n\right).$$

Proof. Let $W_n = e^{\theta M_n}$, so that W_n is non-negative and $W_n = W_{n-1}e^{\theta(M_n - M_{n-1})}$.

Then applying Lemma 10.1(ii) with $Y = M_n - M_{n-1}$ and $\mathcal{G} = \mathcal{F}_{n=1}$,

$$\begin{aligned}\mathbb{E}(W_n \mid \mathcal{F}_{n-1}) &= W_{n-1}\mathbb{E}\left[e^{\theta(M_n - M_{n-1})} \mid \mathcal{F}_{n-1}\right] \\ &\leqslant W_{n-1}\exp\left(\frac{1}{2}\theta^2 c^2\right) \text{ a.s.}\end{aligned}$$

Taking expectations we obtain $\mathbb{E}[W_n] \leqslant \exp\left(\frac{1}{2}\theta^2 c^2\right)\mathbb{E}[W_{n-1}]$ and the result follows by induction. □

Theorem 10.3 (Simple version of the Azuma-Hoeffding inequality)**.** *Suppose M is a martingale with $M_0 = 0$ and $|M_n - M_{n-1}| \leqslant c$ a.s. for all n. Then*

$$\mathbb{P}(M_n \geqslant a) \leqslant \exp\left(-\frac{1}{2}\frac{a^2}{c^2 n}\right),$$

and

$$\mathbb{P}(|M_n| \geqslant a) \leqslant 2\exp\left(-\frac{1}{2}\frac{a^2}{c^2 n}\right).$$

Proof.

$$\begin{aligned}\mathbb{P}(M_n \geqslant a) &\leqslant \mathbb{P}\left(e^{\theta M_n} \leqslant e^{\theta a}\right) \\ &\leqslant e^{-\theta a}\exp\left(\frac{1}{2}\theta^2 c^2\right)\end{aligned}$$

using Markov's inequality. Now we are free to optimise over θ. The RHS is minimised when $\theta = a/(c^2 n)$, giving the required bound.

The same argument applies replacing M by the martingale $-M$. Summing the two bounds then gives the bound for $|M|$. □

We now introduce the idea of *discrete Lipschitz functions.*

Definition 10.4. Let h be a function of n variables. The function h is said to be c-Lipschitz, where $c > 0$, if changing the value of any one coordinate causes the value of h to change by at most c. That is, whenever $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ differ in at most one coordinate, then $|h(\mathbf{x}) - h(\mathbf{y})| \leqslant c$.

Theorem 10.5 (Concentration of discrete Lipschitz functions)**.** *Suppose h is a c-Lipschitz function, and $X_1, \ldots, X_n$ are independent random variables. Then*

$$\mathbb{P}\left(|h(X_1, \ldots, X_n) - \mathbb{E}[h(X_1, \ldots, X_n)]| \geqslant a\right) \leqslant 2\exp\left(-\frac{1}{2}\frac{a^2}{c^2 n}\right).$$

Proof. The proof is based on the idea of the Doob martingale. We reveal information about the underlying random variables $X_1, \ldots, X_n$ one step at a time, gradually acquiring a more precise idea of the value $h(X_1, \ldots, X_n)$.

For $0 \leqslant k \leqslant n$, let $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, and let

$$M_k = \mathbb{E}[h(X_1, \ldots, X_n) \mid \mathcal{F}_k] - \mathbb{E}[h(X_1, \ldots, X_n)].$$

Then $M_0 = 0$, and $M_n = h(X_1, \ldots, X_n) - \mathbb{E}[h(X_1, \ldots, X_n)$.

We claim $|M_{k+1} - M_k| \leqslant c$ a.s. To show this, let $\widehat{X}_{k+1}$ be a random variable with the same distribution as X_{k+1}, which is independent of $X_1, \ldots, X_n$.

Then

$$\begin{aligned} M_k &= \mathbb{E}[h(X_1, \ldots, X_k, \widehat{X}_{k+1}, \ldots, X_n) \mid \mathcal{F}_k] \\ &= \mathbb{E}[h(X_1, \ldots, X_k, \widehat{X}_{k+1}, \ldots, X_n) \mid \mathcal{F}_k] \\ &= \mathbb{E}[h(X_1, \ldots, X_k, \widehat{X}_{k+1}, \ldots, X_n) \mid \mathcal{F}_{k+1}]. \end{aligned}$$

This gives

$$M_{k+1} - M_k = \mathbb{E}[h(X_1, \ldots, X_k, \widehat{X}_{k+1}, \ldots, X_n) - h(X_1, \ldots, X_k, X_{k+1}, \ldots, X_n) \mid \mathcal{F}_{k+1}].$$

But the difference between the two values of h inside the conditional expectation on the RHS is in $[-c, c]$, so we obtain $|M_{k+1} - M_k| \leqslant c$ a.s. as required. Now the required estimate for M_n follows from the Azuma-Hoeffding bound (Theorem 10.3). □

The examples below of the application of Theorem 10.5 show that martingale methods can be applied to problems far away from what one might think of as "stochastic process theory".

Example 10.6 (Longest common subsequence)**.** Let $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$ be two independent sequences, each with independent entries.

Let L_m be the length of the longest sequence which is a subsequence (not necessarily consecutive) of both sequences.

For example, if $m = 12$ and $\mathbf{X}$ ="CAGGGTAGTAAG" and $\mathbf{Y}$ ="CGTGTGAAAACT" then both $\mathbf{X}$ and $\mathbf{Y}$ contain the substring "CGGTAAA", and $L_m = 6$.

Changing a single entry can't change the length of the longest common subsequence by more than 1. We can apply Theorem 10.5 with $n = 2m$ and $c = 1$, to get

$$\mathbb{P}(|L_m - \mathbb{E}[L_m]| \geqslant a) \leqslant 2\exp\left(-\frac{a^2}{4m}\right).$$

We obtain that for large m, "typical fluctuations" of L_m around its mean are on the scale at most $\sqrt{m}$.

Note that we didn't require the sequences $\mathbf{X}$ and $\mathbf{Y}$ to have the same distribution, or for the entries of each sequence to be identically distributed.

As suggested by the choice of strings above, longest common subsequence problems arise for example in computational biology, involving the comparison of DNA strings (which evolve via mutation, insertion or deletion of individual nucleotides).

Example 10.7 (Minimum-length matching)**.** Suppose there are m red points in the box $[0,1]^2 \subset \mathbb{R}^2$, with positions $R_1, \ldots, R_m$, and m blue points with positions $B_1, \ldots, B_m$.

Let X be the length of the minimal-length *matching*, which joins pairs consisting of one blue and one red point. That is,

$$X_m = \min \sum_{k=1}^{m} \|R_k - B_{i_k}\|,$$

where the minimum is taken over all permutations $i_1, i_2, \ldots, i_m$ of $1, 2, \ldots, m$, and $\|r - b\|$ denotes Euclidean distance between r and b.

Alternatively let Y be the length of the minimal-length *alternating tour*, a path which visits all $2m$ points, alternating between red and blue, and returning to its starting point:

$$Y_m = \min \left\{ \sum_{k=1}^{m} \|R_{i_k} - B_{j_k}\| + \sum_{k=1}^{m-1} \|B_{j_k} - R_{i_{k+1}}\| + \|B_{j_m} - R_{i_1}\| \right\},$$

where now the minimum is over all pairs of permutations $i_1, i_2, \ldots, i_m$ and $j_1, j_2, \ldots, j_m$ of $1, 2, \ldots, m$.

Moving a single point cannot change X_m by more than $\sqrt{2}$, and cannot change Y_m by more than $2\sqrt{2}$. If the positions of the points are independent, then applying Theorem 10.5 with $n = 2m$ and the appropriate value of c, we obtain

$$\mathbb{P}(|X_m - \mathbb{E}[X_m]| \geqslant a) \leqslant 2 \exp\left(-\frac{a^2}{8m}\right)$$
$$\mathbb{P}(|Y_m - \mathbb{E}[Y_m]| \geqslant a) \leqslant 2 \exp\left(-\frac{a^2}{32m}\right).$$

Again this gives concentration of X_m and Y_m around their means on the scale of $\sqrt{m}$. This may be a poor bound; for example if all the points are i.i.d. uniform on the box $[0,1]^2$, then in fact the means themselves grow like $\sqrt{m}$ as $m \to \infty$. However, we didn't assume identical distribution. For example we might have red points uniform on the left half $[0, 1/2] \times [0, 1]$, and blue points uniform on the right half $[1/2, 1] \times [0, 1]$, in which case the means grow linearly in m, and the $O(\sqrt{m})$ fluctuation bound is more interesting.

Example 10.8 (Chromatic number of a random graph)**.** The Erdös-Rényi random graph model $G(N, p)$ consists of a graph with N vertices, in which each edge (out of the $\binom{N}{2}$ possible edges) appears independently with probability p. If $p = 1/2$, then the graph is uniformly distributed over all possible graphs with N vertices.

The *chromatic number* $\chi(G)$ of a graph G is the minimal number of colours needed to colour the vertices of G so that any two adjacent vertices have different colours.

Consider applying Theorem 10.5 to the chromatic number $\chi(G)$ of a random graph $G \sim G(N, 1/2)$. We could write $\chi(G)$ as a function of $\binom{N}{2}$ independent Bernoulli random variables, each one encoding the presence or absence of a given edge. Adding or removing a single edge cannot change the chromatic number by more than 1. This would give us a fluctuation bound on $\chi(G)$ on the order of N as $N \to \infty$. However, for large N this is an extremely poor, in fact trivial, result, since $\chi(G)$ itself is known to be on the order of $N/\log(N)$.

We can do much better. For $2 \leqslant k \leqslant N$, let X_k consist of a collection of $k-1$ Bernoulli random variables, encoding the presence or absence of the $k-1$ edges $\{1,k\}, \{2,k\}, \ldots, \{k-1,k\}$. It's still the case that $X_2, \ldots, X_N$ are independent. All the information in X_k concerns edges that intersect the vertex k; changing the status of any subset of these edges can only change the chromatic number by at most 1 (consider recolouring vertex k as necessary). The Doob martingale from the proof of

Theorem 10.5 involves revealing information about the graph vertex by vertex, rather than edge by edge, and is called the *vertex exposure martingale*. Applying the theorem with $n = N - 1$ and $c = 1$, we obtain

$$\mathbb{P}\left(|\chi(G) - \mathbb{E}[\chi(G)]| \geqslant a\right) \leqslant 2\exp\left(-\frac{a^2}{2(N-1)}\right),$$

giving a concentration bound on the scale of $\sqrt{N}$ for large N.