PROBABILITY, MEASURE AND MARTINGALES

Michaelmas Term 2025

Lecturer: Harald Oberhauser

Version of December 3, 2025

0 Introduction

These notes accompany my lecture on *Probability, Measure and Martingales* (B8.1). However, I cannot claim authorship of these notes since they are a essentially a modified version of two sets of notes given to me by Jan Obłój and James Martin. I express my deep gratitude to Jan and James for providing me with such excellent lecture notes. Their notes are, in turn, based on notes by Alison Etheridge, Oliver Riordan, and Zhongmin Qian. Naturally, all errors are mine. While the lecturers (as well as the name) of this course have changed over the years, the syllabus did not change significantly. However, there are a few differences in what each lecturer emphasized. My goal is to provide a rigorous foundation in measure theory and probability that still leaves time to develop intuition and see why the theory of stochastic processes is so fascinating; both mathematically and in terms of applications.

The examinable material is summarized in the syllabus and covered in the lectures; nothing less than or more is examinable. Parts in these lecture notes that are not examinable are clearly marked as such.

Please send your comments and corrections to `oberhauser@maths.ox.ac.uk`**. Thank you!**

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:

- R. Durrett, *Probability: theory and examples,* 5th Edition, CUP 2019 (online).
  `The new edition of this classic. Packed with insightful examples and problems.`
- P.-A. Meyer, *Probability and Potentials,* Blaisdell Publishing Company, 1966.
  `This is more extensive than Williams, use for deep-dives.`
- M. Capiński and P. E. Kopp, *Measure, integral and probability*, Springer, 1999.
  `A gentle guided intro to measure theory. Use if you feel lost on our way.`

- S.R.S. Varadhan, *Probability Theory,* Courant Lecture Notes Vol. 7.
  `A classic. Not for the faint-hearted.`
- `... and more. Feel free to ask if you are missing a book, anything from a bedtime read to a real challenge.`

Contents

0.2 Example 1: Simple Symmetric Random Walk

Consider a sequence of independent random variables $(X_n)_{n\geqslant 1}$, all with the same distribution

$$\mathbb{P}(X_n=-1)=\mathbb{P}(X_n=1)=\tfrac{1}{2}.$$

Note that $\mathbb{E}[X_n]=0$ and $\mathrm{Var}(X_n)=\mathbb{E}[X_n^2]=1$. Let $S_0=0$,

$$S_n=\sum_{k=1}^{n}X_k, \quad n\geqslant 1,$$

denote their cumulative sums. This process is known as the *simple symmetric random walk*. Using that the increments have mean 0 we see that

$$\mathbb{E}[S_n|S_{n-1},\dots,S_0]=\mathbb{E}[S_n|S_{n-1}]=S_{n-1}+\mathbb{E}[X_n]=S_{n-1}.$$

This property capture something essential about the stochastic process $(S_n)_n$: the current value of the process is the best prediction we can make about its future value given the whole history of the process. A process with this property is called a *martingale* and we'll give it a proper definition later on (a first step towards this it give rigorous meaning to conditional expectations). For now, ignore the strange name martingale[1] and just note that from this perspective, it is not surprising that martingales show up in many applications. Indeed, martingales form one of the major building blocks in the theory of stochastic processes.

There are many questions one could be interested in if we want to understand the behaviour of the process; one of the first ones is its long-time behaviour. From the weak law of large numbers we know that

$$\frac{S_n}{n}\longrightarrow 0$$

in probability. Later on, we will show that this convergence actually takes place *almost surely*. This is a non-trivial extension: it took mathematicians over 200 years to prove it! You also have seen that the speed of this convergence can be described using the Gaussian distribution, namely

$$\frac{S_n}{\sqrt{n}}\xrightarrow{d}\mathcal{N}(0,1).$$

Put differently, if I run 100 simulations of my SSRW then, for a large n, and I plot $S_n/\sqrt{n}$ then I expect only 2 paths or so to breach the interval $(-2.326, 2.326)$.

So, can we say something more about those two paths? Those rare paths, how do they behave? This is governed by the *law of the iterated logarithm*. It turns out that

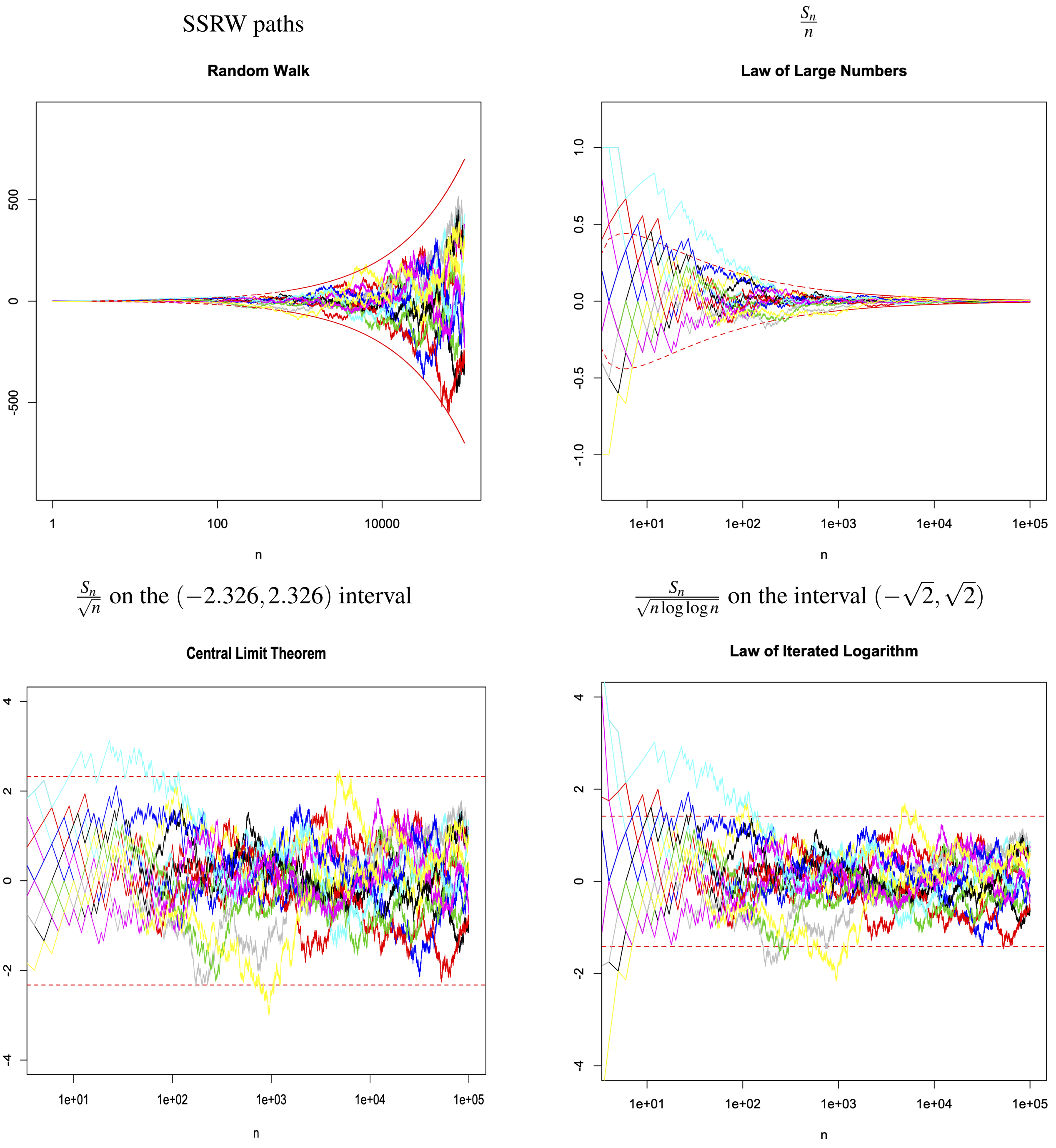
$$\limsup_{n\to\infty}\frac{S_n}{\sqrt{n\log\log n}}=\sqrt{2} \quad \text{and} \quad \liminf_{n\to\infty}\frac{S_n}{\sqrt{n\log\log n}}=-\sqrt{2}, \quad \text{a.s.}$$

See Figure 1 for a visualization. Note that although the process evolves in a completely random fashion, we just managed to gain structural insights about its behaviour and this is just the start of a rich theory.

0.3 Example 2: Mathematical Finance

Suppose $(S_n)_{n\geqslant 0}$ is sequence of random variables modelling the price process of some risky asset, i.e., S_n is the share price at time n. A trader is buying and selling the stock. At time n, they have wealth V_n and decide to

[1] The name martingale comes from some weird historical reasons.



SSRW paths

$\frac{S_n}{n}$

$\frac{S_n}{\sqrt{n}}$ on the $(-2.326, 2.326)$ interval

$\frac{S_n}{\sqrt{n \log\log n}}$ on the interval $(-\sqrt{2}, \sqrt{2})$

Figure 1: Limiting behaviour of a SSRW

buy/sell $H_n = H_n(S_0, S_1, \dots, S_n)$ shares. At time $n+1$, they will have $H_n S_{n+1}$ in shares while their remaining capital/debt grew at rate r:

$$V_{n+1} = H_n S_{n+1} + (V_n - H_n S_n)(1+r) = H_n(S_{n+1} - (1+r)S_n) + V_n(1+r).$$

If we introduce discounted quantities

$$\tilde{V}_n := (1+r)^{-n} V_n, \quad \text{and} \quad \tilde{S}_n := (1+r)^{-n} S_n$$

then the above is re-written as

$$\tilde{V}_{n+1} = H_n(\tilde{S}_{n+1} - \tilde{S}_n) + \tilde{V}_n = \ldots = V_0 + \sum_{t=1}^{n} H_t(\tilde{S}_{t+1} - \tilde{S}_t),$$

an object we will study under the name of *discrete stochastic integral* or a *martingale transform*; again, we'll study this in much more generality later.

Suppose at time $t = 0$ someone wants to purchase from the trader a financial product which, at time $t = N$, will have payoff $f(S_0, S_1, \ldots, S_N)$. What price should the trader set for this product? If they can find a trading strategy H such that $f = V_N$ above, then clearly V_0 is the fair price as it allows the trader to reproduce (hedge) the associated risk fully. But when is this possible and how to find V_0? One example is given by the binomial model.

Proposition 0.1 (Binomial Model pricing)**.** *Suppose there exist two constants u, d such that $0 < 1-d < 1 < 1+r < 1+u$ and $S_{n+1} \in \{(1+u)S_n, (1-d)S_n\}$ a.s., for all $n \geqslant 0$. Then for any f, there exists V_0, H such that $f = V_N$ a.s. In addition, there exists a unique probability measure $\mathbb{Q}$ such that $(\tilde{S}_n)_{n\geqslant 0}$ is a $\mathbb{Q}$-martingale and $V_0 = (1+r)^{-N}\mathbb{E}_{\mathbb{Q}}[f(S_0, \ldots, S_N)]$.*

The take-away message is that although the original process (the stock price) may not be a martingale itself, after a transformation it turns into a martingale. This is useful since it allows to use martingale theory to study the original process. Indeed, this is not special to finance and in many applications there is a martingale lurking somewhere. For our next example, we revisit an example from biology that you probably have encountered in Part A Probability.

0.4 Example 3: The Galton–Watson Branching Process

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 0.2 (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. However this model is applicable to a much wider set of scenarios. You could, for example, see it as a very rudimentary model for spreading a virus, such as Covid-19. Here, each 'generation' lasts maybe 2 weeks and Z_n is the current number of infected individuals. Each of them, independently of the others and in the same manner, then infects further individuals.

In analyzing this process, key roles are played by the expectation $m = \mathbb{E}[X] = \sum_{k=0}^{\infty} kp_k$, 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 0.3. *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*; we'll give a rigorous defintion of conditional expectations later. Notice that the right hand side of (1) is a *random variable*. Now

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

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 0.4. *Let $q = \mathbb{P}[Z_n = 0 \text{ for some } n]$. 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 $m = \mathbb{E}[X] \leqslant 1$, then $q = 1$,*
- *if $m = \mathbb{E}[X] > 1$, then $q < 1$.*

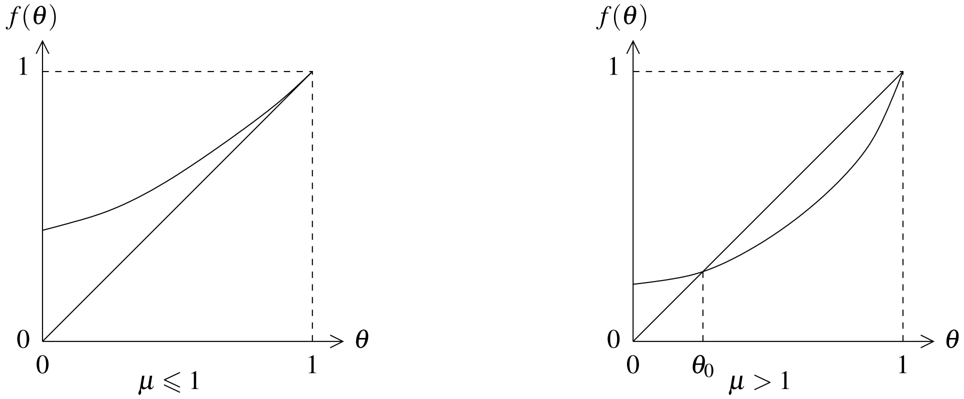
'Proof'

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 $m = f'(1)$:



$\mu \leqslant 1$ $\mu > 1$

In the case $m > 1$, 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 $m > 1$ and $m < 1$, but the case $m = 1$ 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

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

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

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

Then

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

In fact, more is true:

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

Again we encounter 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 $m \leqslant 1$ 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 $\mathscr{P}(\Omega)$ 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 $\mathscr{A} \subseteq \mathscr{P}(\Omega)$ be a collection of subsets of Ω.

1. We say that $\mathscr{A}$ is an *algebra (on Ω)* if $\emptyset \in \mathscr{A}$ and for all $A, B \in \mathscr{A}$, $A^{\mathrm{c}} = \Omega \setminus A \in \mathscr{A}$ and $A \cup B \in \mathscr{A}$.

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

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 $\mathscr{F}$ for a σ-algebra (also called a *σ-field* by some people).

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

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

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

2. *additive* if for all *disjoint* $A, B \in \mathscr{A}$ with $A \cup B \in \mathscr{A}$ (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 $\mathscr{A}$ with $\bigcup_{n=1}^{\infty} A_n \in \mathscr{A}$

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

A measure space is then simply a set Ω equipped with a σ-algebra $\mathscr{F}$ and a countably additive set function μ on $\mathscr{F}$.

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

1. *finite* if $\mu(\Omega) < \infty$,

2. *σ-finite* if there is a sequence $(E_n)_{n \geqslant 1}$ of sets from $\mathscr{F}$ with $\mu(E_n) < \infty$ for all n and $\bigcup_{n=1}^{\infty} E_n = \Omega$,

3. *a probability measure* if $\mu(\Omega) = 1$.

Convention: if μ is a probability measure we call $(\Omega, \mathscr{F}, \mu)$ a *probability space* and often use the notation $\mathbb{P}$ instead of μ to emphasize this.

Any measure μ is also additive and increasing. 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 1.4 (Monotone convergence properties)**.** *Let $(\Omega, \mathscr{F}, \mu)$ be a measure space.*

1. If $(F_n)_{n \geqslant 1}$ is a sequence of sets from $\mathscr{F}$ 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 $\mathscr{F}$ 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 1.5. *Let $\mu : \mathscr{A} \to [0,\infty)$ be an additive set function on an algebra $\mathscr{A}$ taking only finite values. Then μ is countably additive iff for every sequence (A_n) of sets in $\mathscr{A}$ with $A_n \downarrow \emptyset$ we have $\mu(A_n) \to 0$.*

Proof. One implication follows (essentially) from Lemma 1.4; the other is an exercise. □

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

Example 1.6 (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 $(\Omega, \mathscr{P}(\Omega))$ by setting $\mu(A) = \sum_{x\in A} \bar{\mu}(x)$.

Equally, given a measure μ on $(\Omega, \mathscr{P}(\Omega))$ 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 $(\Omega, \mathscr{P}(\Omega))$ 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 1.6). 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 1.12 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 1.7 (Generated σ-algebras)**.** Let $\mathscr{A}$ be a collection of subsets of Ω. Define

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

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

Definition 1.8 (Borel σ-algebra, Borel measure)**.** Let Ω be a topological space with topology (i.e., set of open sets) $\mathscr{T}$. Then *the Borel σ-algebra on* Ω is the σ-algebra generated by the open sets:

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

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

Note that $\mathscr{B}(\Omega)$ 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 1.9 (π-system)**.** Let $\mathscr{I}$ be a collection of subsets of Ω. We say that $\mathscr{I}$ is a *π-system* if $A, B \in \mathscr{I}$ implies $A \cap B \in \mathscr{I}$.

Notice that an algebra is automatically a π-system.

Example 1.10. 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 $\mathscr{B}(\mathbb{R})$, the σ-algebra consisting of all Borel subsets of $\mathbb{R}$ (exercise).

Here's why we care about π-systems.

Theorem 1.11 (Uniqueness of extension)**.** *Let μ_1 and μ_2 be measures on the same measurable space $(\Omega, \mathscr{F})$, and let $\mathscr{I} \subseteq \mathscr{F}$ be a π-system. If $\mu_1(\Omega) = \mu_2(\Omega) < \infty$ and $\mu_1 = \mu_2$ on $\mathscr{I}$, then $\mu_1 = \mu_2$ on $\sigma(\mathscr{I})$.*

We will often apply the theorem to a π-system $\mathscr{I}$ with $\sigma(\mathscr{I}) = \mathscr{F}$, 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 1.11 see (e.g.) Williams, Appendix A.1.

That deals with uniqueness, but what about existence?

Theorem 1.12 (Carathéodory Extension Theorem)**.** *Let Ω be a set and $\mathscr{A}$ an algebra on Ω, and let $\mathscr{F} = \sigma(\mathscr{A})$. Let $\mu_0 : \mathscr{A} \to [0, \infty]$ be a countably additive set function. Then there exists a measure μ on $(\Omega, \mathscr{F})$ such that $\mu = \mu_0$ on $\mathscr{A}$.*

Remark. If $\mu_0(\Omega) < \infty$, then Theorem 1.11 tells us that μ is unique, since an algebra is certainly a π-system.

The Carathéodory Extension Theorem doesn't solve the problem of constructing measures on σ-algebras. However, it reduces it to constructing countably additive set functions on algebras which is a big step toward this goal; we shall see several examples. In particular, we are going to use it to give an essentially complete and practically useful answer on how to characterize probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

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

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

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 $\mathscr{A}$.] 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 $\mathscr{A}$. For details see Appendix A.1 of Williams, or Varadhan and the references therein.

Corollary 1.13. *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 $\mathscr{B}(\mathbb{R})$.*

The proof of Corollary 1.13 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 $\mathscr{M}_{\text{Leb}}$ that contains, but is strictly larger than, $\mathscr{B}(\mathbb{R})$. It turns out (exercise) that $\mathscr{M}_{\text{Leb}}$ consists of all sets that differ from a Borel set on a null set. In this course we shall work with $\mathscr{B}(\mathbb{R})$ rather than $\mathscr{M}_{\text{Leb}}$: the Borel σ-algebra will be 'large enough' for us. (This changes later when studying continuous-time martingales.) An advantage $\mathscr{B}(\mathbb{R})$ 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 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$?)

Definition 1.14. 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 1.4, $F(y_n)=\mu(A_n)\downarrow\mu(A)=F(x)$. 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 $(\mathbb{R},\mathscr{B}(\mathbb{R}))$): 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 $\mathscr{B}(\mathbb{R})$.

Theorem 1.15 (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 $\mathscr{B}(\mathbb{R})$, uniqueness follows by Theorem 1.11. 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 1.12, 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 $\mathscr{I}=\{I_{a,b}:-\infty\leqslant a\leqslant b\leqslant\infty\}$ be the collection of intervals that are open on the left and closed on the right. Let $\mathscr{A}$ be the set of finite disjoint unions of elements of $\mathscr{I}$; then $\mathscr{A}$ is an algebra, and $\sigma(\mathscr{A})=\sigma(\mathscr{I})=\mathscr{B}(\mathbb{R})$.

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

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

for intervals and then extending it to $\mathscr{A}$ by defining it as the sum for disjoint unions from $\mathscr{I}$. 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 $\mathscr{B}(\mathbb{R})$ *provided* that μ_0 is *countably* additive on $\mathscr{A}$. Proving this is slightly tricky. Note that we will have to use right continuity at some point.

First note that by Lemma 1.5, since μ_0 is finite and additive on $\mathscr{A}$, it is *countably* additive if and only if, for any sequence (A_n) of sets from $\mathscr{A}$ 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 $A_1, A_2, \ldots \in \mathscr{A}$ 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

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

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 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 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 $(\Omega,\mathscr{F})$ and $(\Lambda,\mathscr{G})$ be measurable spaces. A function $f:\Omega\to\Lambda$ is *measurable* (with respect to $\mathscr{F},\mathscr{G}$) if

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

Usually $\Lambda=\mathbb{R}$ or $\overline{\mathbb{R}}=[-\infty,\infty]$. In this case we *always* take $\mathscr{G}$ to consist of the Borel sets: $\mathscr{G}=\mathscr{B}(\mathbb{R})$ or $\mathscr{B}(\overline{\mathbb{R}})$, and omit it from the notation. This contrasts with mapping from $\mathbb{R}$, where different σ-algebras are considered in different circumstances (sometimes including $\mathscr{M}_{\mathrm{Leb}}$, 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 $\mathscr{F}$ (and $\mathscr{B}(\mathbb{R})$ or $\mathscr{B}(\overline{\mathbb{R}})$) if and only if $\{x: f(x)\leqslant t\}\in\mathscr{F}$ for every $t\in\mathbb{R}$.*

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

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

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 $(\Omega,\mathscr{F})$ taking values in $\overline{\mathbb{R}}$, and let $h:\mathbb{R}\to\mathbb{R}$ be continuous. Then, whenever they make sense[2], the following are also measurable functions on $(\Omega,\mathscr{F})$:*

$$f_1+f_2,\quad f_1f_2,\quad \max\{f_1,f_2\},\quad \min\{f_1,f_2\},\quad f_1/f_2,\quad h\circ f$$

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

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

Let $(\Omega, \mathscr{F}, \mu)$ 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 $x \mapsto f(x,y)$.

Definition 2.4. A *simple function* ϕ on a measure space $(\Omega, \mathscr{F}, \mu)$ 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 $E_k \in \mathscr{F}$ 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 $(\Omega, \mathscr{F}, \mu)$.

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 $(\Omega, \mathscr{F}, \mu)$ 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 $(\Omega, \mathscr{F}, \mu)$ 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 $(\Omega, \mathscr{F}, \mu)$ 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 $(\Omega, \mathscr{F}, \mu)$ and $A \in \mathscr{F}$ 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 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, 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 $(\Omega_i, \mathscr{F}_i, \mu_i)$ 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 $(\Omega, \mathscr{F}, \mu)$ 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 $(\Omega, \mathscr{F}, \mathbb{P})$ 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 $(\Omega, \mathscr{F}, \mathbb{P})$ 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

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

In particular, $F_X(x) = \mu_X((-\infty, x])$ defines the *distribution function* of X (c.f. Theorem 1.15). 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 $(\Omega, \mathscr{F}, \mu)$. 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 $(\Omega, \mathscr{F}, \mu)$. 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 $(\Omega, \mathscr{F}, \mu)$ 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*

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

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 $\mathscr{F}_i$ 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 $A_1 \in \mathscr{F}_1$ and $A_2 \in \mathscr{F}_2$. The *product σ-algebra* $\mathscr{F} = \mathscr{F}_1 \times \mathscr{F}_2$ is the σ-algebra on Ω generated by the set of all measurable rectangles. (Note that $\mathscr{F}$ is not the Cartesian product of $\mathscr{F}_1$ and $\mathscr{F}_2$.)

Given two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ on $(\Omega_1, \mathscr{F}_1)$ and $(\Omega_2, \mathscr{F}_2)$ respectively, we'd like to define a probability measure on $(\Omega, \mathscr{F})$ 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 $\mathscr{F}$. Note that the set $\mathscr{I}$ of measurable rectangles is a π-system with (by definition) $\sigma(\mathscr{I}) = \mathscr{F}$, so if such a probability measure on $(\Omega, \mathscr{F})$ exists, it is unique by Theorem 1.11.

First, we extend $\mathbb{P}$ to the algebra $\mathscr{A}$ 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 $R_1, \ldots, R_n \in \mathscr{I}$ 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 $\mathscr{F} = \sigma(\mathscr{A})$, we need to check that $\mathbb{P}$ defined by (7) and (8) is actually *countably* additive on $\mathscr{A}$ so that we can apply Carathéodory's Extension Theorem.

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

Proof. For any $A \in \mathscr{A}$ 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 $A_n \in \mathscr{A}$ be disjoint sets with union $A \in \mathscr{A}$, 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 $(\Omega_2, \mathscr{F}_2, \mathbb{P}_2)$) gives $\int f = \sum \int f_n$, i.e., $\mathbb{P}[A] = \sum \mathbb{P}[A_n]$. □

By Carathéodory's Extension Theorem (Theorem 1.12) and Theorem 1.11 we see that $\mathbb{P}$ extends uniquely to a probability measure on $\sigma(\mathscr{A}) = \mathscr{F}$.

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

The definitions extend easily to define the product $\mathscr{F}_1 \times \cdots \times \mathscr{F}_k$ 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 $(\Omega_i, \mathscr{F}_i)_{i \geqslant 1}$ be a sequence of measurable spaces. The *product σ-algebra* $\mathscr{F} = \prod_{i=1}^{\infty} \mathscr{F}_i$ 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 $A_i \in \mathscr{F}_i$, 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 1.12 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 $(\Omega, \mathscr{F}, \mu)$ 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 $(\Omega, \mathscr{F}, \mathbb{P})$ be the product of the probability spaces $(\Omega_i, \mathscr{F}_i, \mathbb{P}_i)$, $i = 1,2$, and let $f(\omega) = f(\omega_1, \omega_2)$ be a measurable function on $(\Omega, \mathscr{F})$. 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 $\mathscr{F}_1$-, $\mathscr{F}_2$-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 $(\Omega, \mathscr{F}, \mathbb{P})$ 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 $A_i \in \mathscr{F}$ 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 $\mathscr{G}_1, \mathscr{G}_2, \dots$ of $\mathscr{F}$ are called independent if whenever $A_i \in \mathscr{G}_i$ and $i_1, i_2, \dots, i_n$ are distinct, then

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

Note that we impose these conditions for finite subsets only, but they then also hold for countable subsets, using Lemma 1.4. 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 $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let X be a real-valued random variable on $(\Omega, \mathscr{F}, \mathbb{P})$. The *$\sigma$-algebra generated by X* is

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

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 $(\Omega, \mathscr{F})$), we have $\sigma(X) \subseteq \mathscr{F}$. 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 $(\Omega, \mathscr{F}, \mathbb{P})$, then

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

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

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

It is easy to check that X is $\mathscr{G}$-measurable if and only if $\{X \leqslant t\} \in \mathscr{G}$ 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 $\sigma(X_1, X_2, \dots)$-measurable if and only if $Y = f(X_1, X_2, \dots)$ for some measurable function f on the countable product of $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with itself.

Definition 2.26 (Independent random variables)**.** Random variables $X_1, X_2, \dots$ are called *independent* if the σ-algebras $\sigma(X_1), \sigma(X_2), \dots$ 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 $(\Omega, \mathscr{F}, \mathbb{P})$ to a measurable space $(\Lambda, \mathscr{G})$ induces a probability measure $\mu_f = \mathbb{P} \circ f^{-1}$ on $(\Lambda, \mathscr{G})$, 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 $(\Omega, \mathscr{F}, \mathbb{P})$ 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 1.11 (uniqueness of extension) to reduce it to something much more manageable.

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

Proof. Fix $G \in \mathscr{G}_0$. The two functions $H \mapsto \mathbb{P}[G \cap H]$ and $H \mapsto \mathbb{P}[G]\mathbb{P}[H]$ define measures on $(\Omega, \mathscr{H})$ (check!) with the same total mass $\mathbb{P}[G]$, and they agree on the π-system $\mathscr{H}_0$. So by Theorem 1.11 they agree on $\sigma(\mathscr{H}_0) = \mathscr{H}$. Hence, for $G \in \mathscr{G}_0$ and $H \in \mathscr{H}$

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

Now fix $H \in \mathscr{H}$ 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 $(\Omega, \mathscr{F}, \mathbb{P})$ is independent* iff *for all $n \geqslant 1$ and all $x_1, \dots x_n \in \mathbb{R}$ (or $\overline{\mathbb{R}}$),*

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

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

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

and

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

Then $\mathscr{T}$ 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 $A \in \mathscr{T}_n$ 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 $\mathscr{T}$ of (X_n) contains only events of probability* 0 *or* 1*. Moreover, any $\mathscr{T}$-measurable random variable is almost surely constant.*

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

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

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

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

For any $A \in \mathscr{A}$, $B \in \mathscr{B}$, 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 $\sigma(\mathscr{A}) = \mathscr{F}_n$ and $\sigma(\mathscr{B}) = \mathscr{T}_n$ are also independent.

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

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

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

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

Now suppose that Y is any (real-valued) $\mathscr{T}$-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, $(\Omega, \mathscr{F}, \mathbb{P})$ will denote a probability space.

Definition 3.1. Let (A_n) be a sequence of sets from $\mathscr{F}$. 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 1.4, $\mathbb{P}[G_n] \downarrow \mathbb{P}[G]$.

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

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

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

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

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

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

□

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 $(\Omega, \mathscr{F}, \mathbb{P})$.

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 $X_n \overset{L^p}{\to} X$) 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 $X_n \overset{d}{\to} X$ or $X_n \overset{\mathscr{L}}{\to} X$) if
$$\lim_{n\to\infty} F_n(x) = F(x)$$
for every $x \in \mathbb{R}$ *at which F is continuous*.

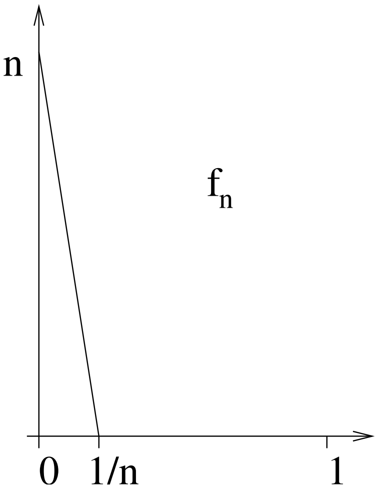
These notions of convergence are all different.

$$\begin{tikzcd}
\textbf{Convergence a.s.} \arrow[r, Rightarrow] & \textbf{Convergence in Probability} \arrow[r, Rightarrow] & \textbf{Convergence in Distribution} \\
& \textbf{Convergence in } L^p \arrow[u, Rightarrow] &
\end{tikzcd}$$

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

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



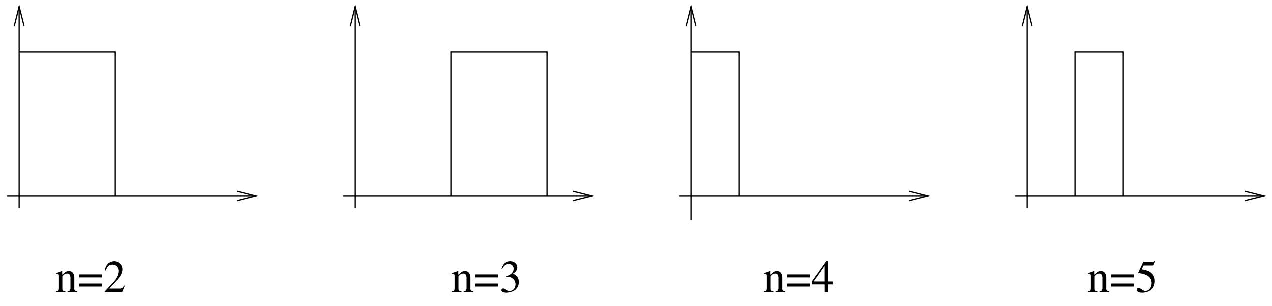
Then $f_n \to 0$ almost everywhere on $[0,1]$ but $f_n \not\to 0$ in L^1. Thinking of each f_n as a random variable, we have $f_n \to 0$ almost surely but $f_n \not\to 0$ in L^1.

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

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

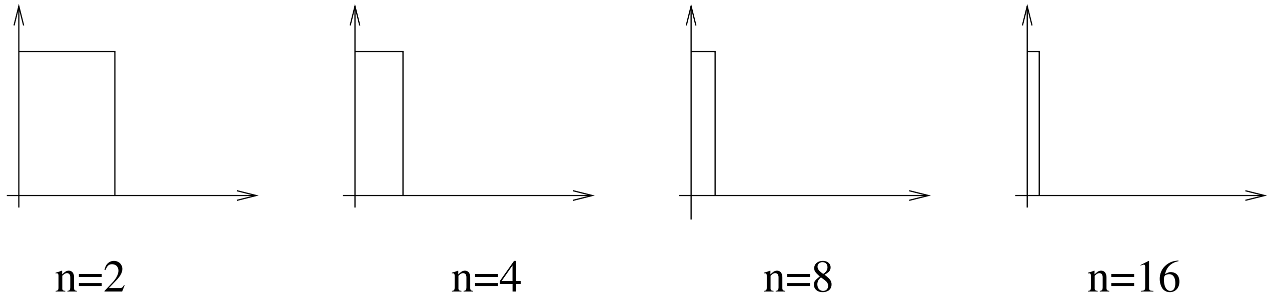
Pictorially we have a 'moving blip' which travels repeatedly across $[0,1]$ getting narrower at each pass.



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 $X_n \xrightarrow{\mathbb{P}} 0$. (Also, $\mathbb{E}[|X_n - 0|] = 1/2^m \to 0$, so $X_n \xrightarrow{L^1} 0$).) 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 $X_1, X_2, \dots$ and X be random variables on $(\Omega, \mathscr{F}, \mathbb{P})$.*

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 $(\Omega, \mathscr{F}, \mathbb{P})$ 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

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

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 $(\Omega, \mathscr{F}, \mathbb{P})$) 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. □

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 $(\Omega, \mathscr{F})$, and define ν by $\nu(A) = \mu(A)/\mu(\Omega)$. Then

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

3.2 $\mathscr{L}^p$ spaces

The set of all random variables on a probability space is a large set. It is useful to structure it by their integrability. This is done with the $\mathscr{L}^p$ spaces that you have already encountered in your integration course.

Definition 3.18. Let $p \geqslant 0$. The space of all random variables X such that $\mathbb{E}[|X|^p] < \infty$ is denoted $\mathscr{L}^p$. In particular, $\mathscr{L}^0$ is the space of all random variables. We also denote $\mathscr{L}^\infty$ the set of all random variables that are bounded.

For $p > 0$ the function $x \to x^p$ is increasing on $\mathbb{R}_+$ so

$$(x+y)^p \leqslant (2 \cdot x \vee y)^p \leqslant 2^p(x^p + y^p), \quad \forall x,y \in \mathbb{R}_+.$$

It follows that $X,Y \in \mathscr{L}^p$ implies $(X+Y) \in \mathscr{L}^p$. Obviously also $\alpha X \in \mathscr{L}^p$ for any $\alpha \in \mathbb{R}$ so *$\mathscr{L}^p$ is a vector space*. For $X \in \mathscr{L}^p$ let us put

$$\|X\|_p := (\mathbb{E}[|X|^p])^{\frac{1}{p}}.$$

Lemma 3.19. *Let $0 \leqslant r \leqslant p$. Suppose $X \in \mathscr{L}^p$. Then $X \in \mathscr{L}^r$ and*

$$\|X\|_r \leqslant \|X\|_p.$$

In particular, convergence in L^p implies convergence in L^r.

Proof. Let $X_k = |X| \wedge k$ which is positive and bounded (and in particular integrable). Applying Jensen's inequality with the convex function $f(x) = x^{p/r}$ on $[0,\infty)$ we get

$$\|X_k\|_r^p = (\mathbb{E}[|X_k|^r])^{p/r} \leqslant \mathbb{E}[|X_k|^p] \leqslant \mathbb{E}[|X|^p] = \|X\|_p^p.$$

Taking limits and invoking the MCT gives the desired inequality. The implications for convergence in $\mathscr{L}^p$ and $\mathscr{L}^r$ is immediate. □

We now derive two crucial inequalities. The Hölder inequality is used in many proofs and Minkowski's inequality shows that $\|\cdot\|_p$ satisfies the triangular inequality.

Theorem 3.20. *Let $p,q > 1$ be such that $\frac{1}{p} + \frac{1}{q} = 1$. Suppose $X,Y \in \mathscr{L}^p$ and $Z \in \mathscr{L}^q$. Then*

$$\textbf{(Hölder's inequality)} \qquad \mathbb{E}[|XZ|] \leqslant \|X\|_p \|Z\|_q,$$
$$\textbf{(Minkowski's inequality)} \qquad \|X+Y\|_p \leqslant \|X\|_p + \|Y\|_p.$$

Proof. Proofs of these inequalities on $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ equipped with the Lebesgue measure were given in Part A Integration. Here we follow Williams and derive these from Jensen's inequality.

If $X = 0$ a.s. then there is nothing to show. Otherwise, define a new probability measure on $(\Omega,\mathscr{F})$ by $\mathbb{Q}(A) = \mathbb{E}[|X|^p \mathbf{1}_A]/\|X\|_p^p$, and a random variable $\bar{Z} := |Z|/|X|^{p-1}\mathbf{1}_{|X|>0}$. Applying Jensen's inequality with $f(x) = x^q$, we have

$$(\mathbb{E}[|XZ|])^q = (\mathbb{E}[\bar{Z}|X|^p])^q = \left(\int Z \,\mathrm{d}\mathbb{Q} \cdot \|X\|_p^p\right)^q \leqslant \int Z^q \,\mathrm{d}\mathbb{Q} \cdot \|X\|_p^{pq} = \mathbb{E}[|Z|^q]\|X\|_p^q,$$

where we used $p+q = pq$. Hölder's inequality follows raising the sides to $1/q$.

For Minkowski's inequality note that $X+Y \in \mathscr{L}^p$ since it is a vector space and let $c = \mathbb{E}[|X+Y|^p]^{1/q} = \||X+Y|^{p-1}\|_q$. Using first the triangular inequality on $\mathbb{R}$, $|x+y| \leqslant |x| + |y|$ and then Hölder's inequality we obtain

$$\mathbb{E}[|X+Y|^p] \leqslant \mathbb{E}[|X| \cdot |X+Y|^{p-1}] + \mathbb{E}[|Y| \cdot |X+Y|^{p-1}] \leqslant \|X\|_p \cdot c + \|Y\|_p \cdot c.$$

Dividing by c gives the desired result since $1 - 1/q = 1/p$. □

The following result is of fundamental importance in functional analysis. We will exploit it for $p = 2$.

Theorem 3.21. *Let $p \geqslant 1$. The vector space $\mathscr{L}^p$ is* complete*, i.e., for any sequence $(X_n)_{n\geqslant 1} \subseteq \mathscr{L}^p$ such that*

$$\sup_{r,s\geqslant n} \|X_s - X_r\|_p \overset{n\to\infty}{\longrightarrow} 0$$

there exists $X \in \mathscr{L}^p$ such that $X_n \to X$ in $\mathscr{L}^p$.

Proof. Pick k_n such that

$$\sup_{r,s\geqslant k_n} \|X_s - X_r\|_p \leqslant 2^{-n}, \quad \text{and in particular } \mathbb{E}[|X_{k_n} - X_{k_{n+1}}|] \leqslant \|X_{k_n} - X_{k_{n+1}}\|_p \leqslant 2^{-n}.$$

Put $Y = \sum_{n\geqslant 1} |X_{k_n} - X_{k_{n+1}}|$. By MCT we have $\mathbb{E}[Y] < \infty$ and in particular $Y < \infty$ a.s. The series being absolutely convergent implies that $\lim_{n\to\infty} X_{k_n}$ exists a.s. We define

$$X(\omega) := \limsup_{n\to\infty} X_{k_n}(\omega), \quad \omega \in \Omega$$

so that X is a random variable and $X_{k_n} \to X$ a.s. For $n \geqslant 1$ and $r > k_n$

$$\mathbb{E}[|X_r - X_{k_m}|^p] = \|X_r - X_{k_m}\|_p^p \leqslant 2^{-np}, \quad m \geqslant n.$$

Taking $m \uparrow \infty$ and using Fatou's lemma gives

$$\mathbb{E}[|X_r - X|^p] \leqslant 2^{-np}.$$

It follows that $X \in \mathscr{L}^p$ and also $X_r \to X$ in $\mathscr{L}^p$, as required. □

Remark. The above shows that $(\mathscr{L}^p, \|.\|_p)$ is almost normed space, the only nuisance is that $\|X\|_p = 0$ implies $X = 0$ a.s. To get rid of this problem, we quotient by the equivalence relation of a.s. equality. This gives us the space L^p – its elements are not random variables anymore but rather equivalence classes relative to a.s. equality. The above shows that $(L^p, \|.\|_p)$ is not only a normed space but also complete. Normed and complete linear spaces are called *Banach spaces* and some you of might take the course on functional analysis in which they are studied in their own and systematic way. However, from the probabilistic point of view, we prefer to work with actual functions, i.e. with random variables and not equivalence classes. One is reason is that when we have a large family $(X_t)_{t\geqslant 0}$ of random variables (i.e. measurable functions), changing each of them on a null set may actually do a lot of harm!

4 Conditional Expectation

Let X be a random variable on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and $\mathscr{G}$ a sub-σ-algebra of $\mathscr{F}$. Does there exist a random variable that is only $\mathscr{G}$-measurable and, in a certain sense, "as close as possible" to X?

There are many situations in which one encounters this question; in particular, when we study martingales we will be interested in the best approximation to the future value of a stochastic process given the σ-algebra generated by the process up to the present moment. However, the importance of conditional expectaton goes well-beyond martingale theory.

Before we study the above question in full generality, let us try to gain some intuition in the case when $\mathscr{G}$ is relatively simple, and we decide to measure closeness by $\|.\|_2$.

Example 4.1. Let $X \in \mathscr{L}^2(\Omega, \mathscr{F}, \mathbb{P})$. Let $B \in \mathscr{F}$ with $\mathbb{P}(B) > 0$ and note that σ-algebra $\sigma(B) = \{\emptyset, \Omega, B, B^c\}$. Consider the optimization problem

$$\inf \mathbb{E}[(X-Y)^2]$$

where the inf is taken over all square-integrable and $\sigma(B)$-measurable random variables. It is easy to check that the infimum is attained by the random variable

$$\omega \mapsto \mathbb{E}[X \mid \sigma(B)](\omega) = \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}(B)}\mathbf{1}_B(\omega) + \frac{\mathbb{E}[X\mathbf{1}_{B^c}]}{\mathbb{P}(B^c)}\mathbf{1}_{B^c}(\omega).$$

We denoted the minimizing random variable with $\mathbb{E}[X \mid \sigma(B)]$ and below we explain why this notation makes sense. Therefore first recall the definition of conditional probability:

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \text{ for } A, B \in \mathscr{F} \text{ with } \mathbb{P}[B] > 0,$$

and note that $\mathbb{P}(\cdot \mid B)$ is a probability measure on $(\Omega, \mathscr{F}, \mathbb{P})$. We can take expectations with respect to this measure and this gives the notion of conditional expectation of a random variable with respect to the set B,

$$\mathbb{E}[X|B] = \int_\Omega X(\omega) d\mathbb{P}(\omega|B) = \frac{\int_B X d\mathbb{P}}{\mathbb{P}(B)} = \frac{\mathbb{E}[X 1_B]}{\mathbb{P}(B)}$$

which you probably have also encountered before in Part A Probability. With this notation we can write

$$E[X \mid \sigma(B)](\omega) = \mathbb{E}[X \mid B]\mathbf{1}_B(\omega) + \mathbb{E}[X \mid B^c]\mathbf{1}_{B^c}(\omega) \tag{11}$$

which hopefully makes the notation $\mathbb{E}[X \mid \sigma(B)]$ more intuitive. Similar explicit formulas for $\mathbb{E}[X \mid \mathscr{G}]$ can be derived if we replace $\sigma(B)$ by another, relatively simple, σ-algebra $\mathscr{G}$ in the minimization problem; for example, if $G_1, \cdots, G_n$ forms a partition of Ω and $\mathbb{P}(G_i) > 0$, then the same argument as before shows that a solution of the minimization problem is given as $\mathbb{E}[X \mid \mathscr{G}] = \sum_{n \geqslant 1} \mathbb{E}[X \mid G_n]\mathbf{1}_{G_n}$ for $\mathscr{G} = \sigma(\{G_1, \ldots, G_n\})$. Finally, we note that

$$\mathbb{E}[\mathbb{E}[X \mid \sigma(B)]\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G] \text{ for all } G \in \sigma(B). \tag{12}$$

which is immediate via the formula (11). Informally, it expresses that the random variable $\mathbb{E}[X \mid \mathscr{G}]$ matches locally the expectation of X; here ”locally” means on the finest level that the σ-algebra $\mathscr{G}$ allows us to look at. This equality will be the key for the general definition when we condition on σ-algebras $\mathscr{G}$ that are not as simple as $\sigma(B)$.

Example 4.2. Let X, Y be random variables on $(\Omega, \mathscr{F}, \mathbb{P})$. We would like a random a random variable Z that is as ”close as possible” to X but only measureable with respect to the $\sigma(Y)$, that is Z must be a function of Y. Motivated by the formula (11) one might be tempted to just set

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

However, in general, this expression does not make sense since $\mathbb{P}(Y=y)$ can be zero. To avoid getting into trouble dividing by zero, we can integrate over $\{Y(\omega) = y\}$ 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

$$\mathbb{E}[Z\mathbf{1}_{Y \in A}] = \mathbb{E}[X\mathbf{1}_{Y \in A}] \tag{13}$$

for each $A \in \mathscr{B}(\mathbb{R})$. At this point we do not have any guarantee for the existence of such a random variable Z but at least equation (13) tells us a non-trivial property of it. (We'll see later and in much greater generality, that equation (13) together with measurability already determines Z uniquely determines up to a nullset). We will soon see that this is a special case of the conditional expectation when we conditioning on the σ-algebra $\sigma(Y)$ that is generated by Y and we'll use the notation $Z = \mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)]$.

This brings us to the general definition of conditional expectation.

Definition 4.3 (Conditional Expectation)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and X an integrable random variable. Let $\mathscr{G} \subseteq \mathscr{F}$ be a σ-algebra. We say that a random variable $\mathbb{E}[X \mid \mathscr{G}]$ is (a version of) the *conditional expectation of X given $\mathscr{G}$* if $\mathbb{E}[X \mid \mathscr{G}]$ is integrable, $\mathscr{G}$-measurable and

$$\mathbb{E}[\mathbb{E}[X \mid \mathscr{G}]\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G] \quad \text{for all } G \in \mathscr{G}. \tag{14}$$

The *conditional probability* is defined as the conditional expectation of an indicator

$$\mathbb{P}(A \mid \mathscr{G}) := \mathbb{E}[\mathbf{1}_A \mid \mathscr{G}] \text{ for } A \in \mathscr{F}.$$

If the σ-algebra $\mathscr{G}$ is generated by a random variable Y, that is $\mathscr{G} = \sigma(Y)$, then we also use the notation $\mathbb{E}[X \mid Y]$ resp. $\mathbb{P}(A \mid Y)$ for $\mathbb{E}[X \mid \mathscr{G}]$ resp. $\mathbb{P}(A \mid \mathscr{G})$. We refer to (14) as the defining relation of conditional expectation.

Above the integrals of X and $\mathbb{E}[X \mid \mathscr{G}]$ over sets $G \in \mathscr{G}$ are the same – this is our "local averaging" property that we first encountered in Example 4.1 – but $\mathbb{E}[X \mid \mathscr{G}]$ is also $\mathscr{G}$-measurable whereas X is $\mathscr{F}$-measurable.

Theorem 4.4 (Existence and uniqueness of conditional expectation)**.** *Let X be an integrable random variable on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and $\mathscr{G} \subseteq \mathscr{F}$ a σ-algebra. The conditional expectation of X given $\mathscr{G}$ exists. It is a.s. unique in the sense that if Z is also a version of the conditional expectation of X given $\mathscr{G}$ then $Z = \mathbb{E}[X \mid \mathscr{G}]$ a.s.*

Proof of Uniqueness. Let Y, Z be two conditional expectations of X given $\mathscr{G}$. Let $G := \{Y > Z\}$ and note that $G \in \mathscr{G}$ as Y, Z are $\mathscr{G}$-measurable. By definition, $\mathbb{E}[Y\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G] = \mathbb{E}[Z\mathbf{1}_G]$ so that $\mathbb{E}[(Y-Z)\mathbf{1}_G] = 0$. But $(Y-Z)\mathbf{1}_G \geqslant 0$ a.s. and hence $(Y-Z)\mathbf{1}_G = 0$ a.s., i.e., $\mathbb{P}(G) = 0$ since $Y - Z > 0$ on G. Swapping Y and Z, we also have $\mathbb{P}(Z > Y) = 0$ and hence $Y = Z$ a.s. □

For the existence of conditional expectation in the general case we will use another important result from measure theory, namely the Radon–Nikodym theorem. Given a measure μ on measurable space $(\Omega, \mathscr{F})$ we can define a new measure ν on $(\Omega, \mathscr{F})$ by taking any measurable $f \geqslant 0$ and setting

$$\nu(A) := \int_A f d\mu. \tag{15}$$

Is there a converse, that is can I write any measure in the form (15)? Well, not every measure can arise in this way, since $\mu(A) = 0$ implies $\nu(A) = 0$. The Radon–Nikodym theorem guarantees that this condition is ready sufficient if we assume that the measures are σ-finite.

Theorem 4.5 (Radon–Nikodym)**.** *Let $(\Omega, \mathscr{F})$ be a measurable space, and let ν and μ be σ-finite measures on $(\Omega, \mathscr{F})$. If*

$$\mu(A) = 0 \textit{ implies } \nu(A) = 0 \quad \textit{for all } A \in \mathscr{F}, \tag{16}$$

then there exists a measurable function $f : \Omega \to [0, \infty)$ such that

$$\nu(A) = \int_A f \, d\mu \quad \textit{for all } A \in \mathscr{F}.$$

The function f is unique up to μ-almost everywhere equality, and is denoted by $f = \frac{d\nu}{d\mu}$.

The requirement (16) is called *absolutely continuity of ν with respect to μ* and the notation is $\nu \ll \mu$; the function f is also called the Radon-Nikodym derivative.

Remark. The Radon-Nikodym theorem has important applications in probability theory besides the existence proof of conditional expectation. For example, if X is a real-valued random variable and the law $\mu_X = \mathbb{P} \circ X^{-1}$ is absolutely continuous with respect to the Lebesgue measure m, $\mu_X \ll m$, then the Radon-Nikodym Theorem implies that X has a density, that is $\mu_X(A) = \int_A f_X(x) m(dx)$ where $f = \frac{d\mu_X}{dm}$ is the Radon-Nikodym derivative. Note that if X is a discete random variable, then μ_X is not absolutely continuous with respect to the Lebesgue measure; indeed X only has a probability mass function and not a density function with respect to the Lebesgue measure).

We omit the proof, which is beyond the scope of the course ((uniqueness of f is essentially the same argument that we saw in the uniqueness proof of conditional expectation; existence is more involved). Instead, let's see how to use this result to deduce the existence of conditional expectation.

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

$$\int_A Z \, \mathrm{d}\mathbb{P} = \int_A X \, \mathrm{d}\mathbb{P}. \tag{17}$$

So, for $A \in \mathscr{G}$, let $\mathbb{Q}[A] = \int_A X \, \mathrm{d}\mathbb{P}$. This defines a finite measure $\mathbb{Q}$ on $(\Omega, \mathscr{G})$. Let $\mathbb{P}|_{\mathscr{G}}$ denote the measure $\mathbb{P}$ restricted to the σ-algebra $\mathscr{G}$. Then $\mathbb{Q} \ll \mathbb{P}|_{\mathscr{G}}$. So applying the Radon–Nikodym Theorem to $\mathbb{Q}$ and $\mathbb{P}_{\mathscr{G}}$ on $(\Omega, \mathscr{G})$, there is a $\mathscr{G}$-measurable function $Z = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}|_{\mathscr{G}}}$ such that (17) holds.[3] 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 $\mathbb{E}[X^+ \mid \mathscr{G}] - \mathbb{E}[X^- \mid \mathscr{G}]$ is $\mathscr{G}$-measurable and, by linearity of the integral, satisfies the defining relation. □

Remark. In Example 4.1, we derived explicit formulas for the conditional expectation when $\mathscr{G}$ was generated by a partition. For other σ-algebra$\mathscr{G}$ it is in general not possible to find explicit formulas for the conditional expectation. However, it may help to observe that for non-negative integrable X, if $\mathscr{I}$ is a π-system that generates $\mathscr{G}$, then it is enough to check the defining relation for $G \in \mathscr{I}$. (To see this, apply Theorem 1.11 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 1.11 for signed measures.)

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

1. $\mathbb{E}[\mathbb{E}[X \mid \mathscr{G}]] = \mathbb{E}[X]$.
2. $\mathbb{E}[aX + bY \mid \mathscr{G}] \overset{\text{a.s.}}{=} a\mathbb{E}[X \mid \mathscr{G}] + b\mathbb{E}[Y \mid \mathscr{G}]$.
3. *If X is $\mathscr{G}$-measurable, then* $\mathbb{E}[X \mid \mathscr{G}] \overset{\text{a.s.}}{=} X$.
4. $\mathbb{E}[c \mid \mathscr{G}] \overset{\text{a.s.}}{=} c$.

[3] 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 $\mathscr{G}$-measurable, then $\int Y \, \mathrm{d}\mathbb{P}|_{\mathscr{G}} = \int Y \, \mathrm{d}\mathbb{P}$. This follows from Remark 2.8.

5. $\mathbb{E}[X \mid \{\emptyset, \Omega\}] = \mathbb{E}[X]$.

6. *If $\sigma(X)$ and $\mathscr{G}$ are independent then $\mathbb{E}[X \mid \mathscr{G}] = \mathbb{E}[X]$ a.s.*

7. *If $X \leqslant Y$ a.s. then $\mathbb{E}[X \mid \mathscr{G}] \leqslant \mathbb{E}[Y \mid \mathscr{G}]$ a.s.*

8. $\big|\mathbb{E}[X \mid \mathscr{G}]\big| \leqslant \mathbb{E}[|X| \mid \mathscr{G}]$ *a.s.*

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

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

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

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

So Z is a version of $\mathbb{E}[aX + bY \mid \mathscr{G}]$, 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 $\mathscr{G}$-measurable and for $G \in \mathscr{G}$

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

so the defining relation holds. □

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

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

Proposition 4.7 (Conditional Convergence Theorems)**.** *Let $X_1, X_2, \ldots$ and X be random variables on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, and let $\mathscr{G} \subseteq \mathscr{F}$ be a σ-algebra.*

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

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

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

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*

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

The proofs all use the defining relation (14) to transfer statements about convergence of the conditional probabilities to our usual convergence theorems and are left as an exercise.

The following results are very useful in manipulating conditional expectations.

Lemma 4.8 ("Taking out what is known")**.** *Let X and Y be random variables on $(\Omega, \mathscr{F}, \mathbb{P})$ with X, Y and XY integrable. Let $\mathscr{G} \subseteq \mathscr{F}$ be a σ-algebra and suppose that Y is $\mathscr{G}$-measurable. Then*

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

Proof. The function $Y\mathbb{E}[X \mid \mathscr{G}]$ is clearly $\mathscr{G}$-measurable, so we must check that it satisfies the defining relation for $\mathbb{E}[XY \mid \mathscr{G}]$. 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 $A \in \mathscr{G}$, then for any $G \in \mathscr{G}$ we have $G \cap A \in \mathscr{G}$ and so by the defining relatio for $\mathbb{E}[X \mid \mathscr{G}]$

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

Now extend by linearity to simple random variables Y. Now suppose that $Y \geqslant 0$ is $\mathscr{G}$-measurable. Then there is a sequence $(Y_n)_{n\geqslant 1}$ of simple $\mathscr{G}$-measurable random variables with $Y_n \uparrow Y$ as $n \to \infty$, it follows that $Y_n X \uparrow YX$ and $Y_n\mathbb{E}[X \mid \mathscr{G}] \uparrow Y\mathbb{E}[X \mid \mathscr{G}]$ 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 4.9 (Tower property of conditional expectations)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, X an integrable random variable and $\mathscr{F}_1$, $\mathscr{F}_2$ σ-algebras with $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \mathscr{F}$. Then*

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

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

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

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

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

□

This extends Part 1 of Proposition 4.6 which (in the light of Part 5) is just the case $\mathscr{F}_1 = \{\emptyset, \Omega\}$.

Jensen's inequality also extends to the conditional setting.

Proposition 4.10 (Conditional Jensen's Inequality)**.** *Suppose that $(\Omega, \mathscr{F}, \mathbb{P})$ 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 $\mathscr{G}$ be a sub σ-algebra of $\mathscr{F}$. If $\mathbb{E}[|f(X)|] < \infty$ then*

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

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

$$\mathbb{E}[f(X) \mid \mathscr{G}] \geqslant \mathbb{E}[a_n X + b_n \mid \mathscr{G}] = a_n \mathbb{E}[X \mid \mathscr{G}] + b_n \quad \text{a.s.} \tag{18}$$

Since the union of a countable collection of null (i.e., probability zero) sets is null we can arrange for (18) to hold simultaneously for all $n \in \mathbb{N}$ except possibly on a null set and so

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

□

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

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

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

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

where $W = Y - \mathbb{E}[X \mid \mathscr{G}]$ and $Z = \mathbb{E}[X \mid \mathscr{G}] - X$. Now Y and $\mathbb{E}[X \mid \mathscr{G}]$ are $\mathscr{G}$-measurable, so W is $\mathscr{G}$ measurable, and using Proposition 4.6 part 1 and Lemma 4.8 we have

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

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

In particular, we can minimize $\mathbb{E}[(Y-X)^2]$ by choosing $Y = \mathbb{E}[X \mid \mathscr{G}]$. In other words, $\mathbb{E}[X \mid \mathscr{G}]$ is the best mean-square approximation of X among all $\mathscr{G}$-measurable random variables. Indeed this is a route to showing that conditional expectations exist without recourse to the Radon–Nikodym Theorem and we give short sketch of this appraoch in the next section.

4.1 Existence of Conditional Expectation via Orthogonal projection in $\mathscr{L}^2$ (not examinable)

We now sketch an alternative existence proof of conditional expectation that has a more gemeotric flavour than the one we saw via the Radon-Nikodym theorem. Neither proof, is better or worse than the other, they just highlight two different ways to think about conditional expectation.

First, recall from Prelims, that if $\mathscr{K} \subset \mathbb{R}^d$ is a linear subspace of $\mathbb{R}^d$, and $x \in \mathbb{R}^d$, then the linear projection of x to $\mathscr{K}$ is the unique vector $y \in \mathscr{K}$ that attains the infimum

$$\inf_{y \in \mathscr{K}} \|x - y\|_{\mathbb{R}^d}.$$

Equivalently, the minimizer y is the unique element in $\mathscr{K}$ such that $(x-y)$ is orthogonal to $\mathscr{K}$, that is

$$\langle x - y, z \rangle_{\mathbb{R}^d} = 0 \quad \forall z \in \mathscr{K}.$$

The minimizer y is also called the orthogonal projection of x to $\mathscr{K}$. Motivated, by this we now replace $\mathbb{R}^d$ by $\mathscr{L}^2(\Omega,\mathscr{F},\mathbb{P})$, consider as subspace $\mathscr{K} = \mathscr{L}^2(\Omega,\mathscr{G},\mathbb{P})$ and study for $X \in \mathscr{L}^2(\Omega,\mathscr{F},\mathbb{P})$ the optimization problem

$$\inf_{Y\in\mathscr{K}} \|X-Y\|_2.$$

If it has a minimizer, this should be our candidate for $\mathbb{E}[X \mid \mathscr{G}]$. The reason we work with $\mathscr{L}^2$ among all the $\mathscr{L}^p$ spaces is that they are special: they carry an inner product,

$$\langle X,Y\rangle := \mathbb{E}[XY] \text{ for } X,Y \in \mathscr{L}^2$$

so that $\|X\|_2 = \sqrt{\langle X,X\rangle}$. In particular, we say that X,Y are *orthogonal* if $\langle X,Y\rangle = 0$. Having access to geometry and the notion of orthogonality helps a lot. In fact, the minimization problem is well-posed for any complete subspace $\mathscr{K}$.

Theorem 4.11. *Let $\mathscr{K}$ be a complete vector subspace of $\mathscr{L}^2(\Omega,\mathscr{F},\mathbb{P})$. For any $X \in \mathscr{L}^2(\Omega,\mathscr{F},\mathbb{P})$ the infimum*

$$\inf_{Z\in\mathscr{K}} \|X-Z\|_2$$

is attained by some $Y \in \mathscr{K}$ and $(X-Y)$ is orthogonal to Z for all $Z \in \mathscr{K}$ with $\|Z\|_2 > 0$.

Proof of Theorem 4.11. Let $(Y_n)_{n\geqslant 1}$ be a sequence which attains the desired infimum, $\|X-Y_n\|_2 \to \Delta$. We argue that the sequence is Cauchy. Using the parallelogram law, we have

$$\|X-Y_r\|_2^2 + \|X-Y_s\|_2^2 = 2\|X-\tfrac{1}{2}(Y_r+Y_s)\|_2^2 + 2\|\tfrac{1}{2}(Y_r-Y_s)\|_2^2.$$

Since $\mathscr{K}$ is a vector space, $\frac{1}{2}(Y_r \pm Y_s) \in \mathscr{K}$ and in particular $\|X-\frac{1}{2}(Y_r+Y_s)\|_2^2 \geqslant \Delta^2$. Optimality of $(Y_n)_{n\geqslant 1}$ readily implies that

$$\sup_{r,s\geqslant n} \|Y_r-Y_s\|_2 \xrightarrow{n\to\infty} 0,$$

i.e., $(Y_n)_{n\geqslant 1}$ is Cauchy. Since $\mathscr{K}$ is complete, there exists $Y \in \mathscr{K}$ with $\|Y_n - Y\|_2 \to 0$ as $n \to \infty$. Minkowski's inequality, see Theorem 3.20, then gives $\|X-Y\|_2 \leqslant \|X-Y_n\|_2 + \|Y-Y_n\|_2$ and taking limits we see that $\|X-Y\|_2 = \Delta$ as required.
Now, let $Z \in \mathscr{K}$ with $\|Z\|_2 > 0$ and note that $(Y+tZ) \in \mathscr{K}$ for all $t \in \mathbb{R}$. Using optimality of Y we have

$$0 \leqslant \|X-(Y+tZ)\|_2^2 - \|X-Y\|_2^2 = t^2\|Z\|_2^2 - 2t\mathbb{E}[Z(X-Y)].$$

Taking $t = \mathbb{E}[Z(X-Y)]/\|Z\|_2^2$, yields

$$0 \leqslant -\frac{\mathbb{E}[Z(X-Y)]^2}{\|Z\|_2^2}$$

which implies that $\mathbb{E}[Z(X-Y)] = 0$ as desired. □

Remark. The above result can be rephrased by saying that any $X \in \mathscr{L}^2$ can be written as $X = Y + (X-Y)$ with $Y \in \mathscr{K}$ and $(X-Y)$ orthogonal to $\mathscr{K}$. Clearly such a decomposition is a.s. unique: if we have two such Y_1, Y_2 then their difference would be both in $\mathscr{K}$ and orthogonal to $\mathscr{K}$ and hence $\mathbb{E}[(Y_1-Y_2)^2] = 0$ so that $Y_1 = Y_2$ a.s. We call Y the (orthogonal) *projection* of X on $\mathscr{K}$.

We now have everything we need to give an alternate proof of the existence of conditional expectations.

Proof of existence in Theorem 4.4. Suppose first that $X \in \mathscr{L}^2(\Omega, \mathscr{F}, \mathbb{P})$ and let $\mathscr{K} = \mathscr{L}^2(\Omega, \mathscr{G}, \mathbb{P})$. Clearly $\mathscr{K}$ is a vector subspace of $\mathscr{L}^2(\Omega, \mathscr{F}, \mathbb{P})$ and is complete by Theorem 3.21. Let Y be the orthogonal projection of X on $\mathscr{K}$ from Theorem 4.11. We will now verify that Y is a version of the conditional expectation of X given $\mathscr{G}$. First Y is $\mathscr{G}$-measurable since $Y \in \mathscr{K}$. Second, for $G \in \mathscr{G}$ note that $\mathbf{1}_G \in \mathscr{K}$ and since $(X - Y)$ is orthogonal to $\mathscr{K}$ we have $\mathbb{E}[(X - Y)\mathbf{1}_G] = 0$ which shows that the defining relationship of conditonal expectation holds.

For $X \in \mathscr{L}^1$, by linearity, it is enough to deal with $X^{\pm}$ separately. Suppose thus that $X \geqslant 0$ and let $X_n = X \wedge n$ which are bounded and in particular in $\mathscr{L}^2$ so that $Y_n = \mathbb{E}[X_n \mid \mathscr{G}]$ exists by the above. From the cMCT we know that $Y := \limsup_{n \to \infty} Y_n$ is a version of $\mathbb{E}[X \mid \mathscr{G}]$. □

5 Filtrations and Stopping Times

The language and tools we have developed so far lend themselves beautifully to describing sequences of random phenomena. These are known as *stochastic processes* and they offer a new level of fun! We will be able to capture their dynamics, their relation to us learning new information, their local properties as well as their long-run behaviour and so much more!

We start with notions relating to information and its evolution. This is captured via σ-algebras and suitable classes of random variables. We work on a fixed probability space $(\Omega, \mathscr{F}, \mathbb{P})$. The measure $\mathbb{P}$ does not play any role here, it's all about sets, functions and their measurability. $\mathbb{P}$ will become important in the next section when we consider the nature of the random evolution.

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

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

Often n is interpreted as time and $\mathscr{F}_n$ encodes everything that can happen by time n. Note in particular that we never forget anything. We usually start at time 0 (the beginning), but not always. We let

$$\mathscr{F}_\infty = \sigma\left(\bigcup_{n\geqslant 0} \mathscr{F}_n\right) \tag{19}$$

be the σ-algebra generated by the filtration. This captures all the information we may acquire but it may be smaller than the abstract $\mathscr{F}$ on our space.

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

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

We may write $\mathbf{X}$ for $(X_n)_{n\geqslant 0}$. If $\mathscr{F}_n$ represents our knowledge at time n, then $\mathbf{X}$ being adapted to $(\mathscr{F}_n)_{n\geqslant 0}$ simply means that X_n is observable at time n. Here is an obvious example of such a filtration.

Definition 5.3 (Natural filtration)**.** The *natural filtration* $(\mathscr{F}_n^X)_{n\geqslant 0}$ associated with a stochastic process $(X_n)_{n\geqslant 0}$ on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is defined by

$$\mathscr{F}_n^X = \sigma(X_0, X_1, \dots, X_n), \quad n \geqslant 0.$$

A stochastic process $\mathbf{X}$ is automatically adapted to the natural filtration it generates. It is also, by definition, the smallest filtration to which $\mathbf{X}$ is adapted.

Often the index n is interpreted as time. This could be days, seconds or years. But it could also be some other, non-uniform, clock ticking. Sometimes more abstract notions of "time" are needed, e.g. n could be the n-th letter in a book, the n-th word etc. In B8.2 you will even consider continuous index sets such as $\{t \geqslant 0\}$ instead of the discrete $\{n : n \geqslant 0\}$ that we consider in B8.1. However, what matters in all these cases is that there is a linear order (which might or might not have a natural interpretation as time).

Whatever the real world interpretation of the index n may be, we shall refer to specific choices of n as *deterministic times*, e.g. $n = 42$, or of if we index by days and X_n could be, e.g., the temperature recorded in Greenwich Observatory at noon on on the first of July, or the Rolls-Royce Holdings plc closing price at the London Stock Exchange.

However, we often use many other, *random* times: the next time I meet you, the first time you see a yeti, the moment the stock price drops by more than 30% from its past maximum. It is clear these are well defined but not known a priori. They are not deterministic but rather of the type 'I know it when it happens'. We shall turn these now into a mathematically precise notion of *stopping times*. Much of the power of martingale methods that we develop later comes from the fact that they work equally well when the index is a stopping time.

Definition 5.4 (Stopping time)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $(\mathscr{F}_n)_{n\geqslant 0}$ a filtration. A random variable τ taking values in $\mathbb{N}\cup\{\infty\} = \{0,1,2,\dots,\infty\}$ is called a *stopping time with respect to* $(\mathscr{F}_n)_{n\geqslant 0}$ if $\{\tau = n\} \in \mathscr{F}_n$ for all n.

So a random time τ is a stopping time if at any point in time n, I can use $\mathscr{F}_n$ to decide if I should stop $\{\tau = n\}$ or not. Because $(\mathscr{F}_n)_{n\geqslant 0}$ is filtration, this is equivalent to $\{\tau \leqslant n\} \in \mathscr{F}_n$ – I stop now or have stopped already – or yet to $\{\tau > n\} \in \mathscr{F}_n$, I decide to continue. You can think of a *stopping time* as a valid strategy for playing a game, investing or gambling. The strategy can rely on the information accrued so far but can not 'peak into the future'. All of the examples listed before the definition have this property.

If the choice of the filtration is unambiguous we shall simply say that τ is a stopping time. Stopping times are sometimes called *optional times*. Note that not all random times are stopping times. If $n = 365$ and τ is the warmest day of the year, then I need $\mathscr{F}_{365}$ to decide when τ actually happens. Likewise, the day in November 2024 on which Rolls Royce is most expensive is not known in advance or when it happens. You need to wait till the end of November to know when it actually occurred. It is not a stopping time.

We now discuss some easy properties of stopping times and first examples. All of this captures the intuition, e.g., it is clear that if I have two valid strategies then I may decide to stop when the first one tells me to, or when both tell me to, i.e., minimum and maximum of stopping times are also stopping times.

Proposition 5.5. *Let $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$ be a filtered probability space and τ, ρ stopping times. Then*

(i) A deterministic time t, $t(\omega) = n$ for all $\omega \in \Omega$ is a stopping time;

(ii) $\tau \wedge \rho$ and $\tau \vee \rho$ are stopping times.

Proof. Exercise □

The following proposition says that the first time an adapted process enters a region is a stopping time. It is also called the first hitting time and provides a canonical example of a stopping time. Indeed, many times will be of this type for some process $\mathbf{X}$. We recall the usual convention that $\inf\emptyset = \infty$.

Proposition 5.6. *Let $\mathbf{X} = (X_n)_{n\geqslant 0}$ be an adapted process on $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$ and $B \in \mathscr{B}(\mathbb{R})$. Then*

$$\mathfrak{h}_B = \inf\{n \geqslant 0 : X_n \in B\},$$

the first hitting time *of B, is a stopping time.*

Proof.

$$\{\mathfrak{h}_B \leqslant n\} = \bigcup_{k=0}^{n} X_k^{-1}(B) \in \mathscr{F}_n.$$

□

We often use stopping times because we are interested in the value our process takes at the stopping that, that is in the random variable X_τ. Since $(X_n)_{n\geqslant 1}$ is a stochastic process that is adapted to $(\mathscr{F}_n)_n$, each X_n is $\mathscr{F}_n$-measurable. But what about X_τ? Clearly, X_τ will be $\mathscr{F}_\infty$-measurable (check yourself!) but in general, we expect that it should be measurable with respect to a smaller σ-algebrathat is determined by the "complexity" of the stopping time τ.

Definition 5.7. Let τ be a stopping time on $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$. We define

$$\mathscr{F}_\tau = \{A \in \mathscr{F}_\infty : A \cap \{\tau = n\} \in \mathscr{F}_n \ \forall n \geqslant 0\}. \tag{20}$$

Note that in the definition we could change $\{\tau = n\}$ to $\{\tau \leqslant n\}$. The following shows that our new notion behaves as we would want it to.

Proposition 5.8. *Let τ, ρ be stopping times on $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$. Then*

(i) $\mathscr{F}_\tau$ defined in (20) is a σ-algebra;

(ii) if $\tau \leqslant \rho$ then $\mathscr{F}_\tau \subseteq \mathscr{F}_\rho$.

Proof. Exercise. □

In particular, combining Propositions 5.5 and 5.8, we have that $(\mathscr{F}_{\tau\wedge n})_{n\geqslant 0}$ is a filtration which is smaller than the original one in the sense that $\mathscr{F}_{\tau\wedge n} \subseteq \mathscr{F}_n$, $n \geqslant 0$.

If $(X_n)_{n\geqslant 0}$ represents our ongoing winning in a game and τ is our stopping strategy then the final win is X_τ. If $\tau < \infty$ then it is a well defined function

$$\Omega \ni \omega \longrightarrow X_\tau(\omega) := X_{\tau(\omega)}(\omega)$$

and is $\mathscr{F}$-measurable since

$$X_\tau^{-1}(B) = \bigcup_{n\geqslant 0} \tau^{-1}(\{n\}) \cap X_n^{-1}(B) \in \mathscr{F}.$$

In fact, X_τ is $\mathscr{F}_\tau$-measurable. We rephrase this introducing the notion of a stopped process.

Proposition 5.9 (Stopped process)**.** *Let $\mathbf{X} = (X_n)_{n\geqslant 0}$ be an adapted process on $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$ and τ a stopping time. Then $\mathbf{X}^\tau = (X_{\tau\wedge n})_{n\geqslant 0}$ is a stochastic process, called the* stopped process. *$\mathbf{X}^\tau$ is adapted to the filtration $(\mathscr{F}_{\tau\wedge n})_{n\geqslant 0}$ and hence also to the filtration $(\mathscr{F}_n)_{n\geqslant 0}$.*

Proof. It suffices to show that if ρ is a finite stopping time then X_ρ is $\mathscr{F}_\rho$-measurable which follows from (20) since

$$\{X_\rho \leqslant x\} \cap \{\rho = n\} = \{X_n \leqslant x\} \cap \{\rho = n\} \in \mathscr{F}_n, \quad \text{for all } n \geqslant 0.$$

□

6 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).

Definition 6.1 (Martingale, submartingales, supermartingale)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $(\mathscr{F}_n)_{n\geqslant 0}$ a filtration. An *integrable, $\mathscr{F}_n$-adapted* stochastic process $(X_n)_{n\geqslant 0}$ is called

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

Viewing the above definition from the point of view of gambling, we may 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 $\mathscr{F}_n$), 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!

Yet another way to view (sub/super) martingales, is by recalling that monotone sequences of real numbers were one of the first things that you studied in your analysis course. Constant sequences are especially simple and together with boundedness, monotone sequences have limits, etc. Such monotone sequences have been popping in many other proofs that you have seen since. (Sub/super) martingales are a probabilisit's version of monotone (or constant) sequences: knowing what happened in the past, "on average" my next element will go up/down/stay the same. From this point of view, it is not suprising that martingales are a building block in the theory of stochastic processes.

Here are some elementary properties.

Proposition 6.2. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space.*

1. *A stochastic process $(X_n)_{n\geqslant 0}$ on $(\Omega, \mathscr{F}, \mathbb{P})$ is a submartingale w.r.t. the filtration $(\mathscr{F}_n)_{n\geqslant 0}$ 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. $(\mathscr{F}_n)_{n\geqslant 0}$ then*
$$\mathbb{E}[X_n] = \mathbb{E}[X_0] \quad \textit{for all } n.$$
3. *If $(X_n)_{n\geqslant 0}$ is a submartingale and $n \geqslant m$ then*
$$\mathbb{E}[X_n \mid \mathscr{F}_m] \geqslant X_m \textit{ a.s.}$$
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 $\mathscr{F}_m$-measurable, we have $\mathbb{E}[X_m \mid \mathscr{F}_m] = X_m$ a.s.

For $n \geqslant m$ we have $\mathscr{F}_m \subseteq \mathscr{F}_n$, so

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

so $\mathbb{E}[X_n \mid \mathscr{F}_m] \geqslant X_m$ 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 6.3 (Natural filtration)**.** The *natural filtration* $(\mathscr{G}_n)_{n\geqslant 0}$ associated with a stochastic process $(X_n)_{n\geqslant 0}$ on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is defined by

$$\mathscr{G}_n = \sigma(X_0, X_1, \dots, X_n), \quad n \geqslant 0.$$

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

Proposition 6.4. *If $(X_n)_{n\geqslant 0}$ is a submartingale w.r.t. some filtration $(\mathscr{F}_n)_{n\geqslant 0}$ on $(\Omega, \mathscr{F}, \mathbb{P})$, then it is also a submartingale with respect to its natural filtration $(\mathscr{G}_n)_{n\geqslant 0}$.*

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

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

□

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 $(\mathscr{F}_n)$, 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 $(\mathscr{F}_n)$.

Example 6.5 (Sums of independent random variables)**.** Suppose that $Y_1, Y_2, \dots$ are independent random variables on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ 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

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

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 6.5 can be replaced by martingale differences (which are not necessarily independent).

Definition 6.6 (Martingale differences)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $(\mathscr{F}_n)_{n\geqslant 0}$ a filtration. A sequence $(Y_n)_{n\geqslant 1}$ of integrable random variables, adapted to the filtration $(\mathscr{F}_n)_{n\geqslant 1}$, is called a *martingale difference sequence* w.r.t. $(\mathscr{F}_n)$ if

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

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

Example 6.7. Let $(\Omega, \mathscr{F}, \mathbb{P})$ 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 6.8. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $(\mathscr{F}_n)_{n\geqslant 0}$ be a filtration. Let X be an integrable random variable (that is $\mathbb{E}[|X|] < \infty$). Then setting

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

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

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

We shall see later that a large class of martingales (called uniformly integrable) can be written in this way. One can think of $(\mathscr{F}_n)_{n\geqslant 0}$ 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 $(\mathscr{F}_n)_{n\geqslant 0}$, and that Y is $\mathscr{F}_0$-measurable. Then $(X_n - Y)_{n\geqslant 0}$ is also a (sub)martingale w.r.t. $(\mathscr{F}_n)$. In particular, if X_0 is $\mathscr{F}_0$-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 6.9. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Suppose that $(X_n)_{n\geqslant 0}$ is a martingale with respect to the filtration $(\mathscr{F}_n)_{n\geqslant 0}$. 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 $(\mathscr{F}_n)_{n\geqslant 0}$.*

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

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

□

Corollary 6.10. *If $(X_n)_{n\geqslant 0}$ is a martingale w.r.t. $(\mathscr{F}_n)_{n\geqslant 0}$ 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. $(\mathscr{F}_n)_{n\geqslant 0}$.*

Definition 6.11 (Predictable process)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $(\mathscr{F}_n)_{n\geqslant 0}$ a filtration. A sequence $(V_n)_{n\geqslant 1}$ of random variables is *predictable* with respect to $(\mathscr{F}_n)_{n\geqslant 0}$ if V_n is $\mathscr{F}_{n-1}$-measurable for all $n \geqslant 1$.

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

Theorem 6.12 (Discrete stochastic integral or martingale transform)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $(\mathscr{F}_n)_{n\geqslant 0}$ a filtration. Let $(Y_n)_{n\geqslant 0}$ be a martingale with respect to $(\mathscr{F}_n)$ with difference sequence $(D_n)_{n\geqslant 1}$. Suppose that $(V_n)_{n\geqslant 1}$ is predictable w.r.t. $(\mathscr{F}_n)$, 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. $(\mathscr{F}_n)$.

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*, and you will learn more about stochastic integrals in B8.2. Here we started with $X_0 = 0$; as far as obtaining a martingale is concerned, it makes no difference if we add some $\mathscr{F}_0$-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 $\mathscr{F}_n$-measurable, so X_n is $\mathscr{F}_n$-measurable. Also,

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

□

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

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

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

Proof. Exercise: imitate the proof of Theorem 6.12. □

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

Example 6.14. Let $(Y_i)_{i\geqslant 1}$ be independent random variables such that $\mathbb{E}[Y_i] = 0$, $\mathrm{Var}(Y_i) = \mathbb{E}[Y_i^2] - \mathbb{E}[Y_i]^2 = \sigma_i^2 < \infty$. 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 $(\mathscr{F}_n)_{n\geqslant 0}$ to be the natural filtration generated by $(Y_n)_{n\geqslant 1}$.

By Example 6.5,

$$X_n = \sum_{i=1}^{n} Y_i$$

is a martingale and so by Proposition **??**, 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*:

$$M_n = X_n^2 - s_n^2, \qquad n \geqslant 0$$

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

Proof. Clearly M_n is $\mathscr{F}_n$-measurable. Then

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

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

This process of 'compensation', whereby we correct a process by something predictable (in this example it was deterministic and the original process was a submartingale) in order to obtain a martingale is a special case of a general result due to Doob.

Theorem 6.15 (Doob's Decomposition Theorem)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $(\mathscr{F}_n)_{n\geqslant 0}$ a filtration. Let $(X_n)_{n\geqslant 0}$ be a sequence of integrable random variables, adapted to $(\mathscr{F}_n)_{n\geqslant 0}$. Then*

1. *$(X_n)_{n\geqslant 0}$ has a* Doob decomposition
$$X_n = X_0 + M_n + A_n \tag{21}$$
where $(M_n)_{n\geqslant 0}$ is a martingale w.r.t. $(\mathscr{F}_n)_{n\geqslant 0}$, $(A_n)_{n\geqslant 1}$ is predictable w.r.t. $(\mathscr{F}_n)$, 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 (21) 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. Since $X_n - X_0 = \sum_{k=1}^n (X_k - X_{k-1})$, we can just use conditional expectation to make each of the terms in the sum predictable, that is we define

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

By setting

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

we get the desired $M_n + A_n = \sum_{k=1}^n (X_k - X_{k-1}) = X_n - X_0$, so (21) holds. The kth summand in A_n is $\mathscr{F}_{k-1}$-measurable, so A_n is $\mathscr{F}_{n-1}$-measurable and (A_n) is predictable w.r.t. $(\mathscr{F}_n)$. Also, since

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

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

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

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

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

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

as shown above. □

Remark (The angle bracket process $\langle M \rangle$)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, $(\mathscr{F}_n)_{n\geqslant 0}$ a filtration and $(M_n)_{n\geqslant 0}$ a martingale with respect to $(\mathscr{F}_n)_{n\geqslant 0}$ with $\mathbb{E}[M_n^2] < \infty$ for each n. (Such a martingale is called an L^2-martingale.) Then by Proposition **??**, $(M_n^2)_{n\geqslant 0}$ is a *sub*martingale. Thus by Theorem 6.15 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 $\mathbb{E}[M_{n+1} \mid \mathscr{F}_n] = M_n$) that

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

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.

6.1 Stopped martingales 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 stopping times. In fact, the 'natural' deterministic times are something of a red herring. It is far better and more useful to think of martingales as living on random time scales. Random, but ones which do not anticipate the future, so ones made up of stopping times.

The following is a simple corollary of Theorem 6.12. It is however so important that it is stated as a theorem!

Theorem 6.16 (Stopped Martingale)**.** *Let* $\mathbf{X}$ *be a martingale on a filtered probability space* $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$ *and* τ *be a finite stopping time. Then* $\mathbf{X}^\tau = (X_{\tau\wedge n} : n \geqslant 0)$ *is a martingale with respect to* $(\mathscr{F}_n)_{n\geqslant 0}$ *and with respect to* $(\mathscr{F}_{\tau\wedge n})_{n\geqslant 0}$.

Proof. Note that $\{\tau \geqslant k\} = \{\tau > k-1\} \in \mathscr{F}_{k-1}$ so that $V_k = \mathbf{1}_{\tau\geqslant k}$, $k \geqslant 1$, is predictable. We have

$$X_0 + \sum_{k=1}^{n} V_k(X_k - X_{k-1}) = X_0 + \sum_{k=1}^{\tau\wedge n}(X_k - X_{k-1}) = X_{\tau\wedge n}$$

and the result follows by Theorem 6.12 and Proposition 6.2. □

More generally, we have fundamental result due to Doob about stopping martingales. We first give a version that applies to bounded stopping times(a stopping time τ is bounded if there is some $N \in \mathbb{N}$ such that $\tau(\omega) \leqslant N$ for all $\omega \in \Omega$.)

Theorem 6.17 (Doob's Optional Stopping Theorem)**.** *Let* $\mathbf{X}$ *be a martingale on a filtered probability space* $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$ *and* τ, ρ *be two bounded stopping times and* $\tau \leqslant \rho$. *Then*

$$\mathbb{E}[X_\rho \mid \mathscr{F}_\tau] = X_\tau \text{ a.s.} \tag{22}$$

and in particular $\mathbb{E}[X_\rho] = \mathbb{E}[X_\tau] = \mathbb{E}[X_0]$.
Similarly, if $\mathbf{X}$ *is a sub- (resp. super-) martingale then* $\mathbb{E}[X_\rho \mid \mathscr{F}_\tau] \geqslant X_\tau$ *(resp.* $\mathbb{E}[X_\rho \mid \mathscr{F}_\tau] \leqslant X_\tau$*) a.s.*

Proof. Consider first the case when $\rho = n$ is a constant. Then (22) follows by simply checking the defining relationship for the conditional expectation. For any $A \in \mathscr{F}_\tau$ we have

$$\mathbb{E}[X_n \mathbf{1}_A] = \sum_{k=0}^{n} \mathbb{E}[X_n \mathbf{1}_A \mathbf{1}_{\tau=k}] = \sum_{k=0}^{n} \mathbb{E}[X_k \mathbf{1}_A \mathbf{1}_{\tau=k}] = \sum_{k=0}^{n} \mathbb{E}[X_\tau \mathbf{1}_A \mathbf{1}_{\tau=k}] = \mathbb{E}[X_\tau \mathbf{1}_A],$$

where the first equality follows since $\tau \leqslant n$ and the second by definition of $\mathscr{F}_\tau$ in (20) and since $\mathbf{X}$ is a martingale. Consider now the general case. The process $Y_n = X_{\rho\wedge n} - X_{\tau\wedge n}$, $n \geqslant 0$, is a martingale as a difference of two martingales, by Theorem 6.16. It follows that:

$$0 = Y_{\tau\wedge n} = \mathbb{E}[Y_n \mid \mathscr{F}_{\tau\wedge n}] = \mathbb{E}[X_{\rho\wedge n} \mid \mathscr{F}_{\tau\wedge n}] - X_{\tau\wedge n} \quad \text{a.s.}$$

where the first equality is by definition, the second follows from the case of a deterministic ρ shown above and the third since $X_{\tau\wedge n}$ is $\mathscr{F}_{\tau\wedge n}$-measurable by Proposition 5.9. It suffices to take n large enough so that $n \geqslant \rho \geqslant \tau$.

The proof for sub-/super- martingales is the same but uses Proposition 6.13 instead of Theorem 6.12. □

We note that the assumption that τ, ρ are bounded is important as the following simple example demonstrates.

Example 6.18. Let $(Y_k)_{k\geqslant 1}$ be i.i.d. random variables with $\mathbb{P}(Y_k = 1) = \mathbb{P}(Y_k = -1) = \frac{1}{2}$. 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 $\tau = \mathfrak{h}_{\{1\}} = \min\{n : M_n = 1\}$, a stopping time by Proposition 5.6. It is easy to show, e.g. using a Borel-Cantelli argument in analogy to Exercise 3.5, that $\tau < \infty$ a.s. and hence $M_\tau = 1$ a.s. But then $\mathbb{E}[M_\tau] = 1 \neq 0 = \mathbb{E}[M_0]$.

The problem in the above example is is that τ is too large. It is finite a.s. but $\mathbb{E}[\tau] = \infty$. Doob's stopping theorem may be extended but requires some further assumptions. Here we give most often invoked extensions.

Theorem 6.19 (Variant of Doob's Optional Stopping Theorem)**.** *Let* $(\Omega, \mathscr{F}, \mathbb{P})$ *be a probability space,* $(\mathscr{F}_n)_{n\geqslant 0}$ *a filtration,* $(M_n)_{n\geqslant 0}$ *a martingale with respect to* $(\mathscr{F}_n)_{n\geqslant 0}$ *and* τ *a stopping time with respect to* $(\mathscr{F}_n)_{n\geqslant 0}$. *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*

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

Then M_τ *is integrable and*

$$\mathbb{E}[M_\tau] = \mathbb{E}[M_0] \tag{23}$$

Proof. By Theorem 6.16 $(M_{n\wedge\tau})_{n\geqslant 0}$ 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

$$\begin{aligned} |M_{n\wedge\tau}| = |M_{n\wedge\tau} - M_{0\wedge\tau}| &\leqslant \sum_{i=1}^{n} |M_{i\wedge\tau} - M_{(i-1)\wedge\tau}| \\ &\leqslant \sum_{i=1}^{\infty} |M_{i\wedge\tau} - M_{(i-1)\wedge\tau}| \\ &= \sum_{i=1}^{\infty} \mathbf{1}_{\tau\geqslant i} |M_i - M_{i-1}|. \end{aligned} \tag{24}$$

Now

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

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 (24) 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 (23) by the appropriate inequality. 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 6.20. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, $(\mathscr{F}_n)_{n\geqslant 0}$ a filtration and τ a stopping time with respect to $(\mathscr{F}_n)_{n\geqslant 0}$. Suppose that there exist $K \in \mathbb{N}$ and $\varepsilon > 0$ such that for all $n \in \mathbb{N}$*

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

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

The proof is an exercise. If $|M_i - M_{i-1}| \leqslant L$ holds a.s. and $\mathbb{E}[\tau] < \infty$, then the third case applies; this is an important case of the Optional Stopping Theorem for applications. We give one such example.

Example 6.21. Suppose that $(\Omega, \mathscr{F}, \mathbb{P})$ 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 $0(1-p_j)+p_ja/p_j=a$, 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 $X_i=0$; she is out if she loses and, 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 $(\mathscr{F}_n)_{n\geqslant 0}$ where $\mathscr{F}_n=\sigma(X_1,\dots,X_n)$. Write τ for the number of random variables to be revealed before we see the required pattern. Let $\varepsilon=p_0^2p_1^2p_2$ and note that $\mathbb{P}(\tau>5)\leqslant(1-\varepsilon)$ and more generally, $\mathbb{P}(\tau>5n)\leqslant(1-\varepsilon)^n$ so that $\mathbb{E}[\tau]=\sum_{n\geqslant 0}\mathbb{P}(\tau\geqslant n)<\infty$. 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 (*ii*) of Theorem 6.19 is satisfied (with this L).

When X_τ is revealed each of the gamblers $1,2,\dots,\tau$ have paid £1 to enter.

- Gambler $\tau-4$ has won £$1/(p_0^2p_1^2p_2)$,
- Gamblers $\tau-3$ and $\tau-2$ have both lost and are out,
- Gambler $\tau-1$ has won £$1/(p_0p_1)$,
- Gambler τ has lost and is out.

Of course, gamblers $\tau+1,\tau+2,\dots$ have not bet at all yet and all gamblers prior to $\tau-4$ have lost and are out.

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

By Theorem 6.19 $\mathbb{E}[M_\tau]=0$, 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.

6.2 Maximal Inequalities

Martingales have to evolve, locally, in a balanced way – in the sense that the conditional expectation of the increment, at any point in time, is zero. This allows us to control the maximum of the process, along its trajectory, using its final value.

Theorem 6.22 (Doob's maximal inequality)**.** *Let $(X_n)_{n\geqslant 0}$ be a submartingale on $(\Omega,\mathscr{F},(\mathscr{F}_n)_{n\geqslant 0},\mathbb{P})$. Then, for $\lambda>0$,*

$$Y_n^\lambda=(X_n-\lambda)\mathbf{1}_{\{\max_{k\leqslant n}X_k\geqslant\lambda\}},\quad n\geqslant 0,$$

is a submartingale. In particular,

$$\lambda\mathbb{P}\big[\max_{k\leqslant n}X_k\geqslant\lambda\big]\leqslant\mathbb{E}[X_n\mathbf{1}_{\{\max_{k\leqslant n}X_k\geqslant\lambda\}}]\leqslant\mathbb{E}[|X_n|].\tag{25}$$

Proof. Let $\tau = \mathfrak{h}_{[\lambda,\infty)} = \inf\{n \geqslant 0 : X_n \geqslant \lambda\}$ and set $V_n = \mathbf{1}_{\{\tau \leqslant n-1\}}$, $n \geqslant 1$. Let $\overline{X}_n := \max_{k\leqslant n} X_k$ and note that $V_n = \mathbf{1}_{\{\overline{X}_{n-1}\geqslant\lambda\}}$. Applying Proposition 6.13 to $-X$ and V we deduce that $(V\circ X)_0 = 0$,

$$(V\circ X)_n = \sum_{k=1}^{n} V_k(X_k - X_{k-1}) = X_{n\vee\tau} - X_\tau = (X_n - X_\tau)\mathbf{1}_{\{\tau\leqslant n\}}, \quad n \geqslant 1,$$

is a submartingale. Further, $X_\tau \geqslant \lambda$ by definition so that $(X_\tau - \lambda)\mathbf{1}_{\{\tau\leqslant n\}}$, $n \geqslant 0$, is an adapted integrable and non-decreasing process and hence a submartingale. This shows that Y^λ is a sum of two submartingales and hence also a submartingale. In particular

$$0 \leqslant \mathbb{E}[(X_0-\lambda)\mathbf{1}_{\{X_0\geqslant\lambda\}}] = \mathbb{E}[Y_0^\lambda] \leqslant \mathbb{E}[Y_n^\lambda] = \mathbb{E}[(X_n-\lambda)\mathbf{1}_{\{\tau\leqslant n\}}] = \mathbb{E}[X_n\mathbf{1}_{\{\overline{X}_n\geqslant\lambda\}}] - \lambda\mathbb{P}(\overline{X}_n \geqslant \lambda).$$

Rearranging we obtain the first required inequality and the second one is trivial. □

Corollary 6.23. *Let $p \geqslant 1$ and $(M_n)_{n\geqslant 0}$ be a martingale on a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$ with $M_n \in \mathscr{L}^p$ for all $n \geqslant 0$. Then, for any $n \geqslant 0$ and $\lambda > 0$*

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

Proof. This follows by applying Theorem 6.22 to $(|M_n|^p)_{n\geqslant 0}$ which is a submartingale by Proposition 6.9. □

The following lemma is an application of Hölder's inequality and allows us to generalize the above Corollary.

Lemma 6.24. *Let X, Y be two positive random variables such that*

$$x\mathbb{P}(X \geqslant x) \leqslant \mathbb{E}[Y\mathbf{1}_{\{X\geqslant x\}}], \quad \forall x > 0.$$

Then for $p > 1$ and $q = p/(p-1)$, we have

$$\|X\|_p \leqslant q\|Y\|_p.$$

Proof. This is only non-trivial if $Y \in \mathscr{L}^p$ so we suppose $\mathbb{E}[Y^p] < \infty$. First use Fubini and the assumption, to show $\mathbb{E}[X^p] \leqslant q\mathbb{E}[X^{p-1}Y]$. Then use Hölder's inequality assuming $X \in \mathscr{L}^p$. In general, use for $X_n = X \wedge n$ and invoke MCT. The details are left as an exercise. □

Theorem 6.25 (Doob's L^p inequality)**.** *Let $p > 1$ and $(X_n)_{n\geqslant 0}$ be a* non-negative *submartingale on $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$ with $X_n \in \mathscr{L}^p$ for all $n \geqslant 0$. Then* $\max_{k\leqslant n} X_k \in \mathscr{L}^p$ *and*

$$\mathbb{E}[X_n^p] \leqslant \mathbb{E}\left[\max_{k\leqslant n} X_k^p\right] \leqslant \left(\frac{p}{p-1}\right)^p \mathbb{E}[X_n^p].$$

Proof. The result follows instantly from Theorem 6.22 and Lemma 6.24. □

Remark. Note that $\max_{k\leqslant n} X_k^p = (\max_{k\leqslant n} X_k)^p$. The above is most often applied with $X_n = |M_n|$ for a martingale M. Note that $p/(p-1) = q$ with $1/p + 1/q = 1$. The above can be rephrased saying that the $\mathscr{L}^p$ norm of the running maximum $\|\max_{k\leqslant n} X_k\|_p$ is comparable with the $\mathscr{L}^p$ norm of the terminal value $\|X_n\|_p$. The assumption $p > 1$ is important. The result is no longer true for $p = 1$.
Note that the stopped process X^n is also a positive submartingale so the values of $\mathbf{X}$ after n are irrelevant, it is enough to have the submartingale defined for $1 \leqslant k \leqslant n$.

We finish the section with a variant of the maximal inequality for supermartingales.

Proposition 6.26. *Let $(X_n)_{n\geqslant 0}$ be a supermartingale on a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$. Then*

$$\lambda \mathbb{P}(\max_{k\leqslant n} |X_k| \geqslant \lambda) \leqslant \mathbb{E}[X_0] + 2\mathbb{E}[X_n^-], \quad \forall \lambda, n \geqslant 0. \tag{26}$$

Proof. Applying Doob's optional sampling theorem to $\mathbf{X}$ and the stopping time $\tau = \min\{k : X_k \geqslant \lambda\} \wedge n$, we obtain

$$\mathbb{E}[X_0] \geqslant \mathbb{E}[X_\tau] \geqslant \lambda \mathbb{P}(\max_{k\leqslant n} X_k \geqslant \lambda) + \mathbb{E}[X_n \mathbf{1}_{\{\max_{k\leqslant n} X_k < \lambda\}}].$$

This leads to

$$\lambda \mathbb{P}(\max_{k\leqslant n} X_k \geqslant \lambda) \leqslant \mathbb{E}[X_0] + \mathbb{E}[X_n^-].$$

On the other hand, the process $(X_n^-)_{n\geqslant 0}$ is a non-negative submartingale so we may apply Theorem 6.22 directly to it giving

$$\lambda \mathbb{P}(\max_{k\leqslant n} X_k^- \geqslant \lambda) \leqslant \mathbb{E}[X_n^-].$$

Combining, we obtain the desired result. □

6.3 The Upcrossing Lemma and Martingale Convergence

We turn now to studying the limiting behaviour of sub-/super- martingales. We start by bounding the number of times these processes can cross an interval of values $[a,b]$. This will allow us to control their oscillations and, in consequence, their limits.

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 6.13 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. Consider the following strategy:

1. You do not invest until the current value X_n 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_n 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. You *can not* outsmart the game.

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 6.27 (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 $(\rho_k)_{k\geqslant 1} = (\rho_k([a,b],\mathbf{x}))_{k\geqslant 1}$ and $(\tau_k)_{k\geqslant 0} = (\tau_k([a,b],\mathbf{x}))_{k\geqslant 0}$ recursively as follows:

Let $\tau_0 = 0$ and for $k \geqslant 1$ let

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

$$\tau_k = \inf\{n \geqslant \rho_k : x_n \geqslant b\},$$

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

Let

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

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

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

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

Lemma 6.28 (Doob's Upcrossing Lemma)**.** *Let $\mathbf{X} = (X_n)_{n\geqslant 0}$ be a supermartingale on a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$ and $a < b$ some 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. ρ_k, τ_k are simply first hitting times *after* previous hitting times. It is an easy induction to check that for $k \geqslant 1$, the random variables $\rho_k = \rho_k([a,b],\mathbf{X})$ and $\tau_k = \tau_k([a,b],\mathbf{X})$ are stopping times. Now set

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

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 $\rho_k < n$ and $\tau_k = \infty$. It encodes our investment strategy above: we hold one unit of stock during an upcrossing or if τ_k is infinite for some k and $n > \rho_k$.

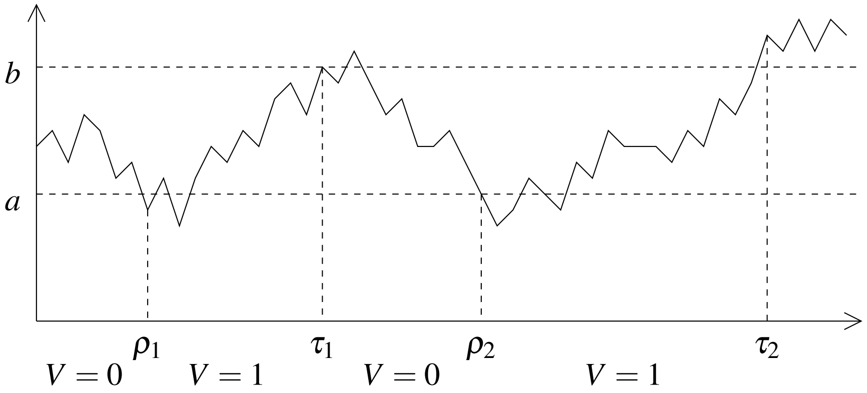


Figure 2: Illustration of the sequence of stopping times introduced in Definition 6.27.

Notice that

$$\{\rho_k < n \leqslant \tau_k\} = \{\rho_k \leqslant n-1\} \cap \{\tau_k \leqslant n-1\}^{\mathrm{c}} \in \mathscr{F}_{n-1}.$$

So $(V_n)_{n\geqslant 1}$ is non-negative and *predictable* so, by Proposition 6.13, $(V\circ X)_n$, $n\geqslant 0$ is a supermartingale. We write $U_n = U_n([a,b],\mathbf{X})$ and compute directly:

$$\begin{aligned}(V\circ X)_n &= \sum_{m=1}^{n} V_m(X_m - X_{m-1}) \\ &= \sum_{k=1}^{U_n}(X_{\tau_k} - X_{\rho_k}) + \mathbf{1}_{\{\rho_{U_n+1}<n\}}(X_n - X_{\rho_{U_n+1}}) &(27)\\ &\geqslant (b-a)U_n - (X_n-a)^-. &(28)\end{aligned}$$

For the last step, note that if indicator function in (27) is non-zero, then $\rho_{U_n+1}<\infty$, so $X_{\rho_{U_n+1}}\leqslant a$. Hence $X_n - X_{\rho_{U_n+1}} \geqslant X_n - a \geqslant -(X_n-a)^-$. Taking expectations in (28),

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

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 6.29. *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 $\mathbf{X}$ 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 6.29 to show that the *random* sequence (X_n) converges. This is our next result.

Definition 6.30. Let (X_n) be a sequence of random variables on a probability space $(\Omega,\mathscr{F},\mathbb{P})$, 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 $\{X_n : n\geqslant 0\}$ of random variables is a bounded subset of $L^p(\Omega,\mathscr{F},\mathbb{P})$: there is some K such that $||X_n||_p \leqslant K$ for all n.

Theorem 6.31 (Doob's Forward Convergence Theorem)**.** *Let $\mathbf{X}$ be a sub- or super- martingale on a filtered probability space $(\Omega,\mathscr{F},(\mathscr{F}_n)_{n\geqslant 0},\mathbb{P})$. If $\mathbf{X}$ 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 6.29 $(X_n)_{n\geqslant 0}$ 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. □

Remark. Warning: the above does *not* say that X_n converge to X in $\mathscr{L}^1$. In particular, it does *not* say that $\mathbb{E}[X_n] \to \mathbb{E}[X]$. This, in general, is false, as Example 6.33 below demonstrates.

Corollary 6.32. *If $(X_n)_{n\geqslant 0}$ is a non-negative supermartingale, then $X_\infty = \lim_{n\to\infty} X_n$ exists a.s. and is integrable.*

Proof. Since $\mathbb{E}[|X_n|] = \mathbb{E}[X_n] \leqslant \mathbb{E}[X_0]$ we may apply Theorem 6.31. □

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 6.33 (Galton–Watson branching process)**.** Recall Definition 0.2: let X be a non-negative integer valued random variable with $0 < m = \mathbb{E}[X] < \infty$. 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

$$Z_{n+1} = \sum_{r=1}^{Z_n} X_{n+1,r} = \sum_{r=1}^{\infty} X_{n+1,r}\mathbf{1}_{\{Z_n \geqslant r\}}$$

so Z_{n+1} is the number of individuals in generation $(n+1)$ of our branching process. Finally, let $M_n = Z_n/m^n$, and let $\mathscr{F}_n = \sigma(\{X_{i,r} : i \leqslant n, r \geqslant 1\})$. By cMCT (which applies since everything is non-negative)

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

and in particular Z_n, M_n are both integrable. Clearly, both are $\mathscr{F}_n$-measurable and $\mathbb{E}[M_{n+1} \mid \mathscr{F}_n] = M_n$ a.s. We conclude that $(M_n)_{n\geqslant 0}$ is a non-negative martingale and, by Corollary 6.32, it converges a.s. to a finite limit M_∞. Does it converge in any other sense?

If $m < 1$ then by the above $(Z_n)_{n\geqslant 0}$ is a non-negative supermartingale and hence also converges a.s. to a finite limit Z_∞. But since $M_n = Z_n/m^n$ converges, we necessarily have $Z_\infty = 0$ a.s. Since Z_n is integer valued it has to be equal to 0 from some point onwards, i.e., $Z_n = 0$ a.s., for $n \geqslant \tau$, where $\tau = \tau(\omega)$ is the extinction time which we conclude has to be finite a.s. Note that $\tau = \inf\{n : Z_n = 0\}$ is a stopping time.

It follows that $M_\infty = 0$ a.s. as well since $M_n = 0$ for $n \geqslant \tau$. However, since (M_n) is a martingale and $M_0 = 1$ we have that $\|M_n\|_1 = \mathbb{E}[M_n] = 1$ does not converge to $\|M_\infty\|_1 = \mathbb{E}[M_\infty] = 0$ which implies that (M_n) does not converge to M_∞ in L^1. (By Lemma 3.19, it also does not converge in L^p for $p > 1$).

What is happening 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 $\mathbb{E}[M_n]$ is constant and does not converge to 0. In the next section we will see that the property that $\{M_n : n \geqslant 0\}$ is missing is so-called uniformly Integrability.

6.4 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 $f_K = |X|\mathbf{1}_{\{|X|>K\}}$, the functions f_K converge to 0 a.s. as $K \to \infty$, and are dominated by the integrable function $|X|$. So by the Dominated Convergence Theorem, $\mathbb{E}[f_K] \to 0$. Uniform integrability demands that this property holds *uniformly* for random variables from some class and UI will allow us to pass from convergence in probability to convergence in $\mathscr{L}^1$ (this will then in particular also deal with a.s. convergence in one go).

Definition 6.34 (Uniform Integrability). A collection $\mathscr{C}$ of random variables is called *uniformly integrable* (UI) if

$$\lim_{K\to\infty} \sup_{X\in\mathscr{C}} \mathbb{E}[|X|\mathbf{1}_{\{|X|>K\}}] = 0.$$

Equivalently: for any $\varepsilon > 0$ there is a K large enough so that $\mathbb{E}[|X|\mathbf{1}_{\{|X|>K\}}] < \varepsilon$ for all $X \in \mathscr{C}$.

Remark. Note that UI property of $\mathscr{C}$ is not affected if we modify its elements on null sets. Consequently, it makes sense to talk about UI of a family of random variables which are only defined a.s. We will use this implicitly in Theorem 6.42 below.

Example 6.35. For $X \in \mathscr{L}^1$ 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 DCT, $\mathbb{E}[f_n] \to 0$. It follows that the singleton family $\{X\}$ is uniformly integrable if and only if X is integrable.

Example 6.36. If $\mathscr{C}$ is a family of random variables with $|X| \leqslant Y$ for all $X \in \mathscr{C}$ and $Y \in \mathscr{L}^1$ then $\mathscr{C}$ is uniformly integrable (this is clear by the previous example). In particular, if we are in the setting of the DCT then UI holds.

From the definition, clearly if $\mathscr{C}$ contains a non-integrable random variable then $\mathscr{C}$ is not UI. But UI of $\mathscr{C}$ is strictly more than just all $X \in \mathscr{C}$ being integrable: we require the convergence $\mathbb{E}[|X|\mathbf{1}_{\{|X|>K\}}] \to 0$, $K \to \infty$, to hold uniformly across $X \in \mathscr{C}$. As easy but very important example is provided by a sequence converging in $\mathscr{L}^1$.

Exercise 6.37. Suppose $X, X_1, X_2, \ldots \in \mathscr{L}^1$ and $\mathbb{E}[|X_n - X|] \to 0$ as $n \to \infty$. Show that $\{X_n : n \geqslant 1\}$ is uniformly integrable.

Remark 6.38. Note that in the definition of UI we can replace $|X|\mathbf{1}_{\{|X|>K\}}$ by a 'comparable' expression such as $(|X| - K)^+$. Their equivalence for the definition follows since

$$0 \leqslant (|X| - K)^+ \leqslant |X|\mathbf{1}_{\{|X|>K\}} \leqslant 2(|X| - K/2)^+.$$

Proposition 6.39. *Let $\mathscr{C}$ be a family of random variables. Then $\mathscr{C}$ is UI if and only if*

$$\sup_{X\in\mathscr{C}} \mathbb{E}[|X|] < \infty \tag{i}$$

and

$$\sup_{A\in\mathscr{F}:\mathbb{P}(A)\leqslant\delta}\ \sup_{X\in\mathscr{C}} \mathbb{E}[|X|\mathbf{1}_A] \overset{\delta\to 0}{\longrightarrow} 0. \tag{ii}$$

Proof. Suppose $\mathscr{C}$ is UI. By definition, there exists K such that $\mathbb{E}[|X|\mathbf{1}_{\{|X|>K\}}] \leqslant 1$, for all $X\in\mathscr{C}$. Thus (i) holds:

$$\mathbb{E}[|X|] = \mathbb{E}\left[|X|\mathbf{1}_{\{|X|\leqslant K\}} + |X|\mathbf{1}_{\{|X|>K\}}\right] \leqslant K + \mathbb{E}\left[|X|\mathbf{1}_{\{|X|>K\}}\right] \leqslant K+1, \quad \forall X\in\mathscr{C}.$$

To see that (ii) holds, fix $\varepsilon>0$ and choose K such that

$$\mathbb{E}[|X|\mathbf{1}_{\{|X|>K\}}] < \tfrac{1}{2}\varepsilon, \quad \forall X\in\mathscr{C}.$$

Set $\delta=\varepsilon/(2K)$ and suppose that $\mathbb{P}(A)<\delta$. Then for any $X\in\mathscr{C}$,

$$\begin{aligned}\mathbb{E}[|X|\mathbf{1}_A] &= \mathbb{E}[|X|\mathbf{1}_A\mathbf{1}_{\{|X|>K\}}] + \mathbb{E}[|X|\mathbf{1}_A\mathbf{1}_{\{|X|\leqslant K\}}]\\ &\leqslant \mathbb{E}[|X|\mathbf{1}_{\{|X|>K\}}] + \mathbb{E}[K\mathbf{1}_A]\\ &\leqslant \tfrac{1}{2}\varepsilon + K\mathbb{P}(A) < \varepsilon,\end{aligned}$$

so that (ii) holds.

For the converse, suppose (i) and (ii) hold. Let $\varepsilon>0$ be given. By (ii) there exists $\delta>0$ such that $\mathbb{P}(A)<\delta$ implies $\mathbb{E}[|X|\mathbf{1}_A]<\varepsilon$ for all $X\in\mathscr{C}$. Let M denote the value of the finite supremum in (i). For K large enough, namely for $K>M/\delta$, by Markov's inequality we have

$$\mathbb{P}(|X|>K) \leqslant \frac{\mathbb{E}[|X|]}{K} \leqslant \frac{M}{K} < \delta, \quad \forall X\in\mathscr{C}.$$

Putting the two together we get the desired result:

$$\mathbb{E}\left[|X|\mathbf{1}_{\{|X|>K\}}\right] < \varepsilon \quad \text{for all } X\in\mathscr{C}.$$

□

Remark. If we impose a minor technical condition on our probability space, namely that it is *atomless*, $\mathbb{P}(\{\omega\})=0$ for all $\omega\in\Omega$, then (ii) on its own implies uniform integrability. So 'morally' (ii) is really equivalent to uniform integrability, and is often the best way of thinking about it.

We start with a variant of the Bounded Convergence Theorem, which is a warm up to the main result.

Lemma 6.40. *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

$$\begin{aligned}\mathbb{E}[|X_n-X|] &= \mathbb{E}\left[|X_n-X|\mathbf{1}_{A_n} + |X_n-X|\mathbf{1}_{A_n^c}\right]\\ &\leqslant \mathbb{E}[|X_n|\mathbf{1}_{A_n}] + \mathbb{E}[|X|\mathbf{1}_{A_n}] + \varepsilon\\ &\leqslant 2\mathbb{E}[K\mathbf{1}_{A_n}] + \varepsilon = 2K\mathbb{P}[A_n] + \varepsilon.\end{aligned} \tag{29}$$

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. □

Naturally if $X_n \to X$ a.s. then the above is a simple corollary to the DCT. Note however that in Example **??** we saw a sequence of $(X_n)_{n\geqslant 1}$ which was uniformly bounded and converged in probability and in L^1 but not almost surely.

The next result extends the previous easy result to the situation when the $(X_n)_{n\geqslant 1}$ are uniformly integrable. In this sense, it provides the converse to Exercise 6.37. It follows that UI is the *right* condition: $X_n \to X$ in L^1 *if and only if* $X_n \to X$ in probability and $\{X_n : n \geqslant 1\}$ is uniformly integrable.

Theorem 6.41 (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):*

(i) the family $\{X_n : n \geqslant 1\}$ is uniformly integrable,

(ii) $X \in \mathscr{L}^1$ and $\mathbb{E}[|X_n - X|] \to 0$ as $n \to \infty$,

(iii) $X \in \mathscr{L}^1$ and $\mathbb{E}[|X_n|] \to \mathbb{E}[|X|] < \infty$ as $n \to \infty$.

Proof. Suppose $\mathscr{C} = \{X_n : n \geqslant 1\}$ is UI. We try to repeat the proof of Lemma 6.40, using the bound (29). 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 6.39, i.e., X is integrable. Now fix $\varepsilon > 0$, and let $A_n = \{|X_n - X| > \varepsilon\}$. As before,

$$\begin{aligned} \mathbb{E}[|X_n - X|] &= \mathbb{E}\left[|X_n - X|\mathbf{1}_{A_n}\right] + \mathbb{E}\left[|X_n - X|\mathbf{1}_{A_n^c}\right] \\ &\leqslant \mathbb{E}\left[|X_n|\mathbf{1}_{A_n}\right] + \mathbb{E}\left[|X|\mathbf{1}_{A_n}\right] + \varepsilon. \end{aligned}$$

Since $X_n \to X$ in probability we have $\mathbb{P}[A_n] \to 0$ as $n \to \infty$, so by Proposition 6.39 (ii)

$$\mathbb{E}[|X_n|\mathbf{1}_{A_n}] \to 0 \quad \text{as } n \to \infty.$$

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 (ii).

$(ii) \Rightarrow (iii)$ follows by $-|X_n - X| \leqslant |X| - |X_n| \leqslant |X - X_n|$ as in the proof of Scheffe's Lemma (Exercise Sheet 2).

It remains to show $(iii) \Rightarrow (i)$. Note that we can not repeat the arguments in the proof of Scheffe's Lemma which relied on a.s. convergence to use the DCT. Instead, we use the bounded convergence result Lemma 6.40. To avoid clutter, let $Y_n = |X_n|$ and $Y = |X|$. Note that $Y_n, Y \geqslant 0$, $Y_n \xrightarrow{\mathbb{P}} Y$. We use Remark 6.38 to establish UI of $\mathscr{C}$.

Since $|(Y_n \wedge K) - (Y \wedge K)| \leqslant |Y_n - Y|$, we have $Y_n \wedge K \xrightarrow{\mathbb{P}} Y \wedge K$ and, by Lemma 6.40, $\mathbb{E}[Y_n \wedge K] \to \mathbb{E}[Y \wedge K]$. Recalling that, by assumption, $\mathbb{E}[Y_n] \to \mathbb{E}[Y]$ this gives

$$\mathbb{E}[(Y_n - K)^+] = \mathbb{E}[Y_n] - \mathbb{E}[Y_n \wedge K] \xrightarrow{n\to\infty} \mathbb{E}[Y] - \mathbb{E}[Y \wedge K] = \mathbb{E}[(Y - K)^+] < \varepsilon,$$

where the last inequality holds for all K large enough since $Y \in \mathscr{L}^1$. 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 now apply the above results in conjunction with what we already know about martingales.

Theorem 6.42. *Let X be an integrable random variable on $(\Omega, \mathscr{F}, \mathbb{P})$ and $\{\mathscr{F}_\alpha : \alpha \in I\}$ a family of σ-algebras with each $\mathscr{F}_\alpha \subseteq \mathscr{F}$. Then the family $\{X_\alpha : \alpha \in I\}$ with*

$$X_\alpha = \mathbb{E}[X \mid \mathscr{F}_\alpha] \quad a.s.$$

is uniformly integrable.

Proof. Since $f(x) = |x|$ is convex, by the conditional form of Jensen's inequality (Proposition 4.10),

$$|X_\alpha| = |\mathbb{E}[X \mid \mathscr{F}_\alpha]| \leqslant \mathbb{E}\big[|X| \mid \mathscr{F}_\alpha\big] \text{ a.s.} \tag{30}$$

and in particular $\mathbb{E}[|X_\alpha|] \leqslant \mathbb{E}[|X|]$. Using (30) and monotonicity of the conditional expectation, Proposition 4.6), we have

$$\mathbb{E}[|X_\alpha|\mathbf{1}_{\{|X_\alpha|>K\}}] \leqslant \mathbb{E}\left[\mathbb{E}[|X| \mid \mathscr{F}_\alpha]\mathbf{1}_{\{|X_\alpha|>K\}}\right] = \mathbb{E}[|X|\mathbf{1}_{\{|X_\alpha|>K\}}], \tag{31}$$

where for the equality we use the measurability of the indicator function to move it inside the conditional expectation and then Proposition 4.6. Since $\{X\}$ is UI, applying Proposition 6.39, for a 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. □

We say that a martingale $\mathbf{M} = (M_n)_{n\geqslant 0}$ is *uniformly integrable* to indicate that the family of random variables $\{M_n : n \geqslant 0\}$ is UI.

Theorem 6.43. *Let $(M_n)_{n\geqslant 0}$ be a martingale on a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$. TFAE*

(i) $\mathbf{M}$ *is uniformly integrable,*

(ii) there is some $\mathscr{F}_\infty$-measurable random variable M_∞ such that $M_n \to M_\infty$ almost surely and in $\mathscr{L}^1$,

(iii) there is an integrable $\mathscr{F}_\infty$-measurable random variable M_∞ such that $M_n = \mathbb{E}[M_\infty \mid \mathscr{F}_n]$ a.s. for all n.

Proof. $(i) \implies (ii)$: $\mathbf{M}$ is UI so in particular, by Proposition 6.39, bounded in $\mathscr{L}^1$ and hence, by Doob's Forward Convergence Theorem (Theorem 6.31) it converges a.s. to some integrable M_∞. Since a.s. convergence implies convergence in probability, $M_n \to M_\infty$ in L^1 by Theorem 6.41. Each M_n is $\mathscr{F}_\infty$-measurable and hence so is M_∞ by Lemma 2.3.

$(ii) \implies (iii)$: Since (M_n) is a martingale, for $m \geqslant n$, we have

$$\mathbb{E}[M_m \mid \mathscr{F}_n] = M_n \quad \text{a.s.},$$

so, by the defining relation (14) for the conditional expectation,

$$\mathbb{E}[M_m\mathbf{1}_A] = \mathbb{E}[M_n\mathbf{1}_A], \quad \text{for all } A \in \mathscr{F}_n.$$

Since

$$\big|\mathbb{E}[M_\infty\mathbf{1}_A] - \mathbb{E}[M_m\mathbf{1}_A]\big| \leqslant \mathbb{E}[|(M_\infty - M_m)\mathbf{1}_A|] \leqslant \mathbb{E}[|M_\infty - M_m|] \to 0,$$

it follows that

$$\mathbb{E}[M_\infty\mathbf{1}_A] = \mathbb{E}[M_n\mathbf{1}_A] \quad \text{for all } A \in \mathscr{F}_n.$$

Since M_n is $\mathscr{F}_n$-measurable, this shows that $M_n = \mathbb{E}[M_\infty \mid \mathscr{F}_n]$ a.s.

$(iii) \implies (i)$ by Theorem 6.42.

The last assertion follows instantly from the Dominated Convergence Theorem and Theorem 6.44 below. □

We now extend the optional sampling theorem as well as the maximal and L^p inequalities to the setting of UI martingales.

Theorem 6.44. *On a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$, let $\mathbf{M}$ be a UI martingale so that $M_n = \mathbb{E}[M_\infty \mid \mathscr{F}_n]$ for some $M_\infty \in \mathscr{L}^1(\Omega, \mathscr{F}_\infty, \mathbb{P})$. Then for any stopping times $\tau \leqslant \rho$*

$$\mathbb{E}[M_\rho \mid \mathscr{F}_\tau] = M_\tau \text{ a.s.} \tag{32}$$

and in particular $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$.
Further, Doob's maximal and L^p inequalities extend to $n = \infty$. Specifically, with $M^*_\infty = \max_{n\geqslant 0} |M_n|$ we have

$$\lambda \mathbb{P}[M^*_\infty \geqslant \lambda] \leqslant \mathbb{E}[|M_\infty| \mathbf{1}_{\{M^*_\infty \geqslant \lambda\}}], \quad \lambda \geqslant 0. \tag{33}$$

Further, if $M_\infty \in \mathscr{L}^p$ for some $p > 1$ then, with $p^{-1} + q^{-1} = 1$,

$$\|M_\infty\|_p \leqslant \|M^*_\infty\|_p \leqslant q\|M_\infty\|_p \tag{34}$$

and $M_n \to M_\infty$ in $\mathscr{L}^p$.

Proof (for those who are interested;the proof of this theorem was not presented in lectures and is not examinable). First note that if τ is bounded, $\tau \leqslant n$ and $\rho = \infty$ then by Theorem 6.17

$$M_\tau = \mathbb{E}[M_n \mid \mathscr{F}_\tau] = \mathbb{E}[\mathbb{E}[M_\infty \mid \mathscr{F}_n] \mid \mathscr{F}_\tau] = \mathbb{E}[M_\infty \mid \mathscr{F}_\tau].$$

It remains the establish the same for any stopping time τ and $\rho = \infty$ as the general case then follows by the tower property.

Let $A \in \mathscr{F}_\tau$ and note that $A \cap \{\tau \leqslant n\}$ is in $\mathscr{F}_n$, by definition of $\mathscr{F}_\tau$, but also in $\mathscr{F}_{\tau\wedge n}$ as is easy to verify. Then

$$\mathbb{E}[M_\infty \mathbf{1}_{A\cap\{\tau<\infty\}}] = \lim_{n\to\infty} \mathbb{E}[M_\infty \mathbf{1}_{A\cap\{\tau\leqslant n\}}] = \lim_{n\to\infty} \mathbb{E}[M_{\tau\wedge n} \mathbf{1}_{A\cap\{\tau\leqslant n\}}] = \mathbb{E}[M_\tau \mathbf{1}_{A\cap\{\tau<\infty\}}],$$

where the first equality follows by the MCT, the second follows since we already have the desired property for bounded stopping times and the last equality is a consequence of Theorem 6.41 thanks to uniform integrability of the family $M_{\tau\wedge n} = \mathbb{E}[M_\infty \mid \mathscr{F}_{\tau\wedge n}]$, $n \geqslant 0$, (by Theorem 6.42) and a.s. convergence $M_{\tau\wedge n}\mathbf{1}_{A\cap\{\tau\leqslant n\}} \to M_\tau \mathbf{1}_A$ (and hence also in probability). Finally, the equality $\mathbb{E}[M_\infty \mathbf{1}_{A\cap\{\tau=\infty\}}] = \mathbb{E}[M_\tau \mathbf{1}_{A\cap\{\tau=\infty\}}]$ is obvious. This establishes (32).

We turn to the two remaining assertions. By conditional Jensen's inequality $(|M_n|)_{0\leqslant n\leqslant\infty}$ is a submartingale. By Doob's maximal inequality, Theorem 6.22, with $M^*_n = \max_{k\leqslant n} |M_k|$, we have

$$\lambda \mathbb{P}\big[M^*_n \geqslant \lambda\big] \leqslant \mathbb{E}[|M_n| \mathbf{1}_{\{M^*_n \geqslant \lambda\}}] \leqslant \mathbb{E}[|M_\infty| \mathbf{1}_{\{M^*_n \geqslant \lambda\}}]$$

since $\{M^*_n \geqslant \lambda\} \in \mathscr{F}_n$ and $\mathbb{E}[|M_\infty| \mid \mathscr{F}_n] \geqslant |M_n|$. Taking the limit in $n \to \infty$, using MCT on the left and DCT on the right, we see that the maximal inequality (33) holds as required. Suppose now that $M_\infty \in \mathscr{L}^p$ for some $p > 1$. Then Doob's L^p inequality (34) follows by Lemma 6.24. It shows in particular that $|M_n|^p \leqslant (M^*_\infty)^p \in \mathscr{L}^1$ and hence $M_n \to M_\infty$ in $\mathscr{L}^p$ by the DCT. □

7 Applications of the Martingale Theory

Martingale theory had profound impact on many areas of mathematics, statistics, and applied fields such as

- Optional stopping and optimal stopping problems (e.g., pricing American options, gambler's ruin with stopping rules)
- Stochastic calculus foundations (defining the Itô integral via martingale properties, see next terms course)
- Doob–Meyer decomposition (used in survival analysis, event history modelling, and point processes)
- Convergence theorems (e.g., proving the Strong Law of Large Numbers via martingale SLLN)
- Filtering theory (Kalman filter and nonlinear filtering derived from martingale orthogonality)
- Likelihood ratio martingales in change-point detection and sequential probability ratio tests
- Branching processes — martingale limits characterize extinction probabilities
- Harmonic functions and potential theory (martingales from Markov processes; optional stopping used to characterize harmonic functions)
- Financial mathematics, in particular Risk-neutral pricing via martingale measures, the Fundamental theorem of asset pricing (no arbitrage equivelent to the existence of martingale measure), and Hedging strategies derived from martingale representation.
- Queueing theory (martingale methods for proving heavy-traffic limits and diffusion approximations)
- Random walks and Brownian motion (e.g., reflection principle derived through martingale arguments)
- Information theory (Azuma–Hoeffding inequality and concentration derived from martingale difference sequences)
- Machine learning:
  - (i) PAC-Bayes and online learning bounds via martingale concentration
  - (ii) Stochastic gradient methods analyzed with martingale convergence
  - (iii) Diffusion models as generative models
- Epidemic modelling where counting processes have martingale compensators
- Geometric probability and percolation theory (martingale methods for revealing processes)

There are many others and above is far from an exhaustive list. On Sheet 4 we are going to see some applications; we conclude the course by a classic result, namely an elegant martingale proof of the the Strong Law of Large Numbers.

7.1 Backwards Martingales and the Strong Law of Large Numbers

So far our martingales were sequences (M_n) of random variables on $(\Omega, \mathscr{F}, \mathbb{P})$ defined for all integers $n \geqslant 0$. 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 $\mathscr{F}_t \subseteq \mathscr{F}$ (information known at time t) and an integrable, $\mathscr{F}_t$-measurable random variable M_t, with $\mathbb{E}[M_{t+1} \mid \mathscr{F}_t] = M_t$ 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. This instantly reminds us of UI martingales in Theorem 6.43 and makes our life easier.

Definition 7.1. Given σ-algebras $(\mathscr{F}_{-n})_{n \geqslant 0}$ with $\mathscr{F}_{-n} \subseteq \mathscr{F}$ and

$$\cdots \subseteq \mathscr{F}_{-(n+1)} \subseteq \mathscr{F}_{-n} \subseteq \cdots \subseteq \mathscr{F}_{-2} \subseteq \mathscr{F}_{-1} \subseteq \mathscr{F}_0,$$

a *backwards martingale* w.r.t. $(\mathscr{F}_{-n})$ is a sequence $(M_{-n})_{n \geqslant 0}$ of integrable random variables, each M_{-n} is $\mathscr{F}_{-n}$-measurable and

$$\mathbb{E}[M_{-n+1} \mid \mathscr{F}_{-n}] = M_{-n} \quad \text{a.s.}$$

for all $n \geqslant 1$.

For any backwards martingale, we have

$$\mathbb{E}[M_0 \mid \mathscr{F}_{-n}] = M_{-n} \quad \text{a.s.}$$

Since M_0 is integrable, it follows from Theorem 6.42 that $(M_{-n})_{n \geqslant 0}$ is *automatically* uniformly integrable.

Doob's Upcrossing Lemma (Lemma 6.28), dealt with martingales on a finite set of time points. We can apply it to $(M_{-m}, M_{-m+1}, \ldots, M_{-1}, M_0)$, to see that if $U_m([a,b], \mathbf{M})$ is the number of upcrossings of $[a,b]$ by the backwards martingale between times $-m$ and 0, then

$$\mathbb{E}[U_m([a,b], \mathbf{M})] \leqslant \frac{\mathbb{E}[(M_0 - a)^-]}{b - a}. \tag{35}$$

Mimicking the proof of Doob's Forward Convergence Theorem (Theorem 6.31), we let $m \to \infty$ and use Monotone Convergence Theorem to conclude that $U([a,b], \mathbf{M}) = U_\infty([a,b], \mathbf{M})$ is integrable and hence finite a.s. Lemma 6.29 then shows that M_{-n} converges a.s. to $M_{-\infty} := \liminf_{n \to \infty} M_{-n}$. Recall that as n increases $\mathscr{F}_{-n}$ decrease, so that $M_{-\infty}$ is $\mathscr{F}_{-n}$-measurable for all $n \geqslant 0$ and hence also measurable with respect to

$$\mathscr{F}_{-\infty} = \bigcap_{k=0}^{\infty} \mathscr{F}_{-k}.$$

Since (M_{-n}) is uniformly integrable, adapting the proof of Theorem 6.43 gives the following result.

Theorem 7.2. *Let $(M_{-n})_{n \geqslant 0}$ be a backwards martingale w.r.t. $(\mathscr{F}_{-n})_{n \geqslant 0}$. Then M_{-n} converges a.s. and in L^1 as $n \to \infty$ to the random variable $M_{-\infty} = \mathbb{E}[M_0 \mid \mathscr{F}_{-\infty}]$.*

Note that we can replace M_0 by any other fixed element of the sequence: $M_{-\infty} = \mathbb{E}[M_{-k} \mid \mathscr{F}_{-\infty}]$ for all $k \geqslant 0$. We now use this result to prove the celebrated Kolmogorov's Strong Law.

Theorem 7.3 (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 m, and set*

$$S_n = \sum_{k=1}^{n} X_k.$$

Then

$$\frac{S_n}{n} \overset{n\to\infty}{\longrightarrow} m \quad \textit{a.s. and in } \mathscr{L}^1.$$

Proof. For $n \geqslant 1$ set

$$\mathscr{F}_{-n} = \sigma(S_n, S_{n+1}, S_{n+2}, \ldots) = \sigma(S_n, X_{n+1}, X_{n+2}, \ldots),$$

noting that $\mathscr{F}_{-n-1} \subseteq \mathscr{F}_{-n}$. Conditioning on $\mathscr{F}_{-n}$ 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,

$$\mathbb{E}[X_1 \mid \mathscr{F}_{-n}] = \mathbb{E}[X_2 \mid \mathscr{F}_{-n}] = \cdots = \mathbb{E}[X_n \mid \mathscr{F}_{-n}]$$

and so they are all equal (a.s.) to their average:

$$\mathbb{E}[X_i \mid \mathscr{F}_{-n}] = \frac{1}{n}\mathbb{E}[X_1 + \cdots + X_n \mid \mathscr{F}_{-n}] = \frac{1}{n}\mathbb{E}[S_n \mid \mathscr{F}_{-n}] = \frac{1}{n}S_n, \quad 1 \leqslant i \leqslant n.$$

Let $M_{-n} = S_n/n$. Then, for $n \geqslant 2$,

$$\mathbb{E}[M_{-n+1} \mid \mathscr{F}_{-n}] = \frac{1}{n-1}\mathbb{E}[S_{n-1} \mid \mathscr{F}_{-n}] = \frac{1}{n-1}\sum_{i=1}^{n-1}\mathbb{E}[X_i \mid \mathscr{F}_{-n}] = \frac{S_n}{n} = M_{-n}.$$

In other words, $(M_{-n})_{n\geqslant 1}$ is a backwards martingale w.r.t. $(\mathscr{F}_{-n})_{n\geqslant 1}$. Thus, by Theorem 7.2, S_n/n converges a.s. and in L^1 to $M_{-\infty} = \mathbb{E}[M_{-1} \mid \mathscr{F}_{-\infty}]$, where $\mathscr{F}_{-\infty} = \bigcap_{k\geqslant 1}\mathscr{F}_{-k}$.

Now by L^1 convergence, $\mathbb{E}[M_{-\infty}] = \lim_{n\to\infty}\mathbb{E}[M_{-n}] = \mathbb{E}[M_{-1}] = \mathbb{E}[S_1] = m$. 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, Theorem 2.31, it is a.s. constant, so $M_{-\infty} = m$ a.s. □

Index