PROBABILITY, MEASURE AND MARTINGALES

Michaelmas Term 2022

Lecturer: Jan Obłój

Version of December 23, 2022

0 Introduction

These notes accompany my lecture on *Probability, Measure and Martingales* (B8.1). The notes borrow heavily from previous versions by Alison Etheridge, Oliver Riordan and James Martin as well as another set by Zhongmin Qian. I am grateful for them for making their notes available to me. I also want to thank Benjamin Joseph who, as my academic assistant, helped to improve these notes. Finally, in parts, I followed closely the exposition in two wonderful books on the subjects and I want to acknowledge the influence their authors, David Williams and Paul-André Meyer, had on this text. I do not reiterate it elsewhere but I stress it here. Naturally, all errors are mine.

Not having the strict time-limit imposed on a lecture course, the notes tend to go into various (interesting!) digressions and cover additional material which is meant to provide the reader with a "larger and clearer picture". Some parts of the material which are additional and are not covered in the lectures are clearly labeled (as *deep dives*). However, this is not always possible so to know the examinable material you should watch the lectures. I should stress that the material presented in the lectures is examinable – nothing less or more.

These notes are work in progress and are being constantly improved. I am very grateful to all who have helped me to improve them. Your comments, corrections, but also questions during office hours, are precious.

Please send all your comments and corrections to `jan.obloj@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:

- 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.`

- Z. Brzezniak, T. Zastawniak, *Basic stochastic processes: a course through exercises*, Springer, 1999.
  `More elementary than Williams, but a helpful complimentary first reading.`
- R. Durrett, *Probability: theory and examples,* 5th Edition, CUP 2019 (online).
  `The new edition of this classic. Packed with insightful examples and problems.`
- 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.`

0.2 Notation

It is useful to record here some basic notation and conventions used throughout. We let $\mathbb{R}$ denote the real numbers, $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ the extended reals, $\mathbb{Q}$ the rational numbers, $\mathbb{N} = \{1, 2, \ldots\}$ denote strictly positive integers and $\mathbb{Z}$ all integers. Unless specified, we mean non-strict inequalities, i.e., we say "positive" for non-negative, increasing for "non-decreasing" etc. We shall use $|\cdot|$ to denote the natural norm on the usual spaces. In particular, $|A|$ denotes the number of elements for $A \subset \mathbb{N}$ and $|x|$ denotes the Euclidean norm of $x \in \mathbb{R}^d$.

For a set $A \subset \Omega$ we let A^c denote its complement, i.e., $A^c = \{x \in \Omega : x \notin A\}$. Note that for the notion of complement to make sense, we have to specify the larger space of which A is a subset. This should always be clear from the context and will most often be Ω. For two sets $A, B \in \Omega$ we denote their set difference with $A \setminus B = A \cap B^c$ and their symmetric difference with $A \triangle B = (A \cap B^c) \cup (B \cap A^c)$. We shall often work with a subset of points $\omega \in \Omega$ for which a certain property Γ holds and will denote this $\{\omega \in \Omega : \Gamma(\omega)\}$ or simply $\{\Gamma\}$. The most prominent example is '$X(\omega) \in E$', for a given function X and a set E, so that $\{\omega \in \Omega : X(\omega) \in E\}$ will simply be denoted $\{X \in E\}$.

We will often work with collections of subsets, or of functions, and denote these with calligraphic letters $\mathscr{F}, \mathscr{G}, \mathscr{A}$ etc. We will often consider collections closed under certain operations. For example, we say that a collection of sets $\mathscr{F}$ is closed under countable unions if $\bigcup_{n=1}^{\infty} A_n \in \mathscr{F}$ for any sequence of sets $A_n \in \mathscr{F}$, $n \geqslant 1$. Similarly, we would say that a collection of functions $\mathscr{A}$ is closed under pointwise multiplication if the function $fg \in \mathscr{A}$ (defined via $fg(\omega) = f(\omega)g(\omega)$) for any $f, g \in \mathscr{A}$.

We will often consider monotone sequences of sets or functions. 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 $\bigcap_{n=1}^{\infty} G_n = G$. Likewise, $f_n \uparrow f$, for functions on some set Ω, is understood pointwise and means that $f_n(\omega) \leqslant f_{n+1}(\omega)$, $n \geqslant 1$, and $f_n(\omega) \to f(\omega)$ for all $\omega \in \Omega$.

We will denote the operations of min/max with $\wedge/\vee$, i.e., $f \wedge g = \min\{f, g\}$ and $f \vee g = \max\{f, g\}$. We also write $f^+ = f \vee 0$ for the positive part of a function f and $f^- = (-f) \vee 0$ for its negative part.

We use $\mathbf{1}$ to denote the indicator function: $\mathbf{1}_E(\omega)$ is equal to 1 for $\omega \in E$ and 0 elsewhere. If E is defined through the properties of ω we drop the argument, e.g., $\mathbf{1}_{\lfloor 2^n \omega \rfloor \text{ is even}}$ is one on the set of $\omega \in [0, 1]$ for which the integer part of $2^n \omega$ is even and 0 otherwise.

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

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

Deep Dive

Anything marked as a *Deep Dive* covers material outside of the syllabus. It is only intended for those who are

interested and eager to understand things in more depth. It is non-examinable and not necessary for the course. It goes above and beyond the material, often indicating links with other courses and parts of mathematics. Even the eager readers should skip those parts on the first reading. More deep dives may appear as I revise the notes. The depth of deep dives may vary considerably from one dive to another.

Contents

0.3 The Galton–Watson branching process

We begin with an example that illustrates some of the concepts that lie ahead. This example was already introduced in Part A Probability so we don't go into excessive detail.

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

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

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

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

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

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

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

'Proof'

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

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

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

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

This is our first example of a *conditional expectation*, studied in section 6. Notice that the right hand side of (1) is a *random variable*. Now

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

and the claim follows by induction. □

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

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

Claim 0.3. *Let $q = \mathbb{P}[Z_n = 0 \text{ for some } n]$. 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 $q < 1$.*

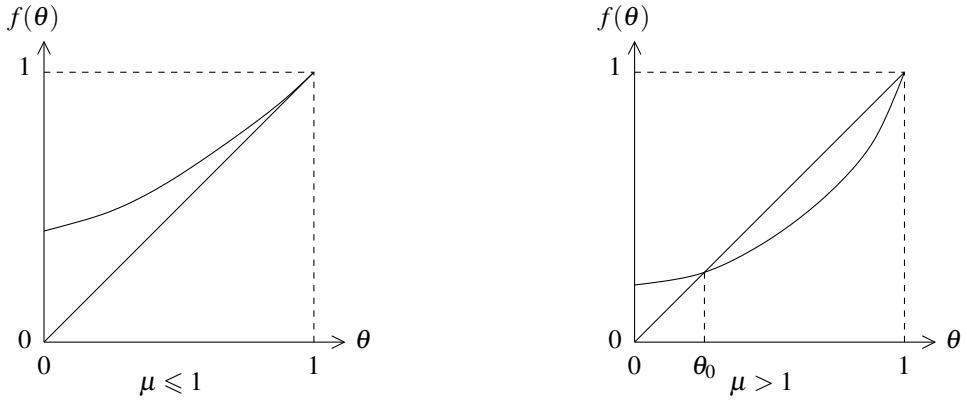
'Proof'

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 $m = f'(1)$:



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:

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

A process $(M_n)_{n\geqslant 0}$ with this property is called a *martingale*. We introduce and study martingales in section 8.

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.

Before we go into theory, let us study the limiting behaviour of processes on one more, more familiar, example.

0.4 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*. Again, it should be intuitively clear that our best prediction of the state at time n, given the history, is S_{n-1} itself as the increment has mean 0:

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

From the weak law of large numbers we know that

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

in probability. In Theorem 9.3, we will show that this convergence actually takes place *almost surely*. This is a non-trivial extension: it took mathematicians over 300 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, see section 9.4, that

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

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

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

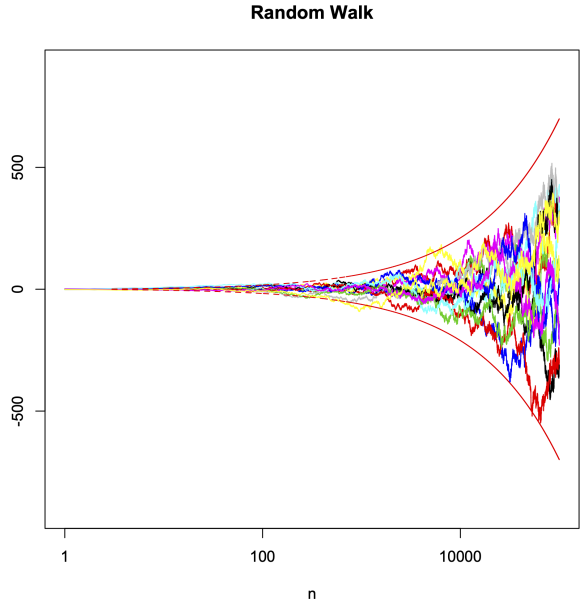
If we introduce discounted quantities

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

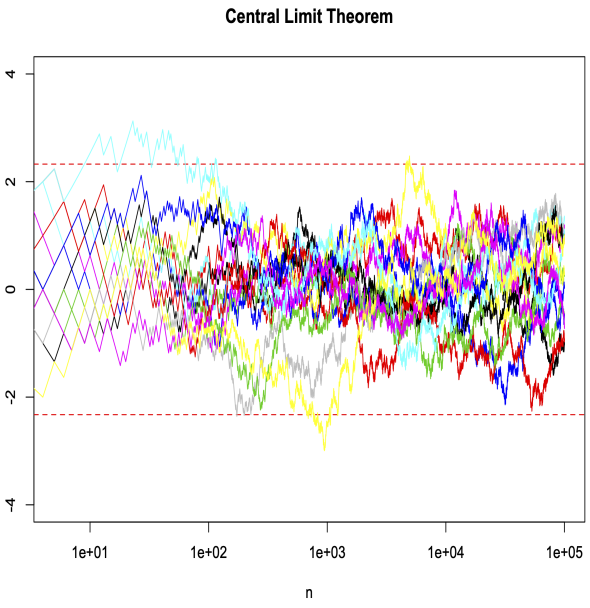
then the above is re-written as

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

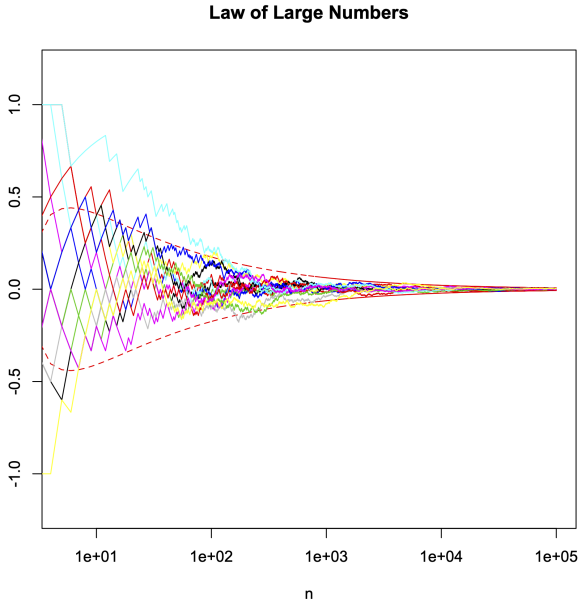
SSRW paths



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



$\frac{S_n}{n}$



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

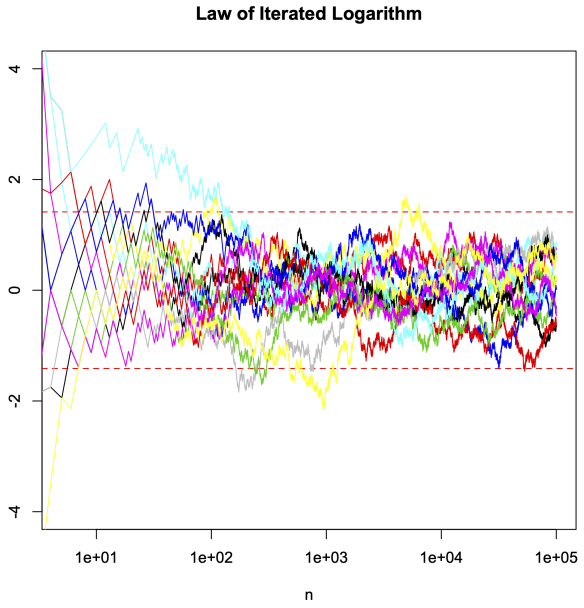


Figure 1: Limiting behaviour of a SSRW

an object we will study under the name of *discrete stochastic integral* or a *martingale transform*, see Theorem 8.12.

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 0.4 (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 $V_0 = (1+r)^{-N}\mathbb{E}_{\mathbb{Q}}[f(S_0, \ldots, S_N)]$.*

1 Measurable sets and functions, a.k.a. events and random variables

Whereof one cannot speak, thereof one must be silent.
The limits of my language mean the limits of my world.
Ludwig Wittgenstein

Our fundamental interest in this course is in endowing a space of outcomes with a measure which describes the relative likelihood of these outcomes and in understanding how this translates into (random) behaviour of functions depending on these outcomes. To achieve this abstract goal we have to invest some time and effort in developing suitable language to speak of sets and functions. This section will appear somewhat arid at the first reading. It may please some readers, those are invited to study it, and its appendix, in detail. Others might be bored by it, those are invited to skim through and then come back when a given notion is needed. You can then study the particular notion knowing that it is actually useful and has its deeper purpose. Nevertheless, an initial reading will equip you with a basic vocabulary without which it is difficult to proceed.

1.1 Events and σ-algebras

For a set Ω, we let $\mathscr{P}(\Omega)$ be 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* if $\emptyset \in \mathscr{A}$ and for all $A, B \in \mathscr{A}$, $A^c = \Omega \setminus A \in \mathscr{A}$ and $A \cup B \in \mathscr{A}$.
2. We say that $\mathscr{A}$ is a σ-*algebra* (or a σ-*field*) if $\emptyset \in \mathscr{A}$, $A \in \mathscr{A}$ implies $A^c \in \mathscr{A}$, 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 a collections of sets which is closed under *finite* set operations. A σ-algebra is a collection of sets which is closed under *countable* set operations. Note that the notions of algebra and σ-algebra are relative to Ω since A^c makes sense only if we specify the "parent" set Ω we have in mind. A σ-algebra will be most often denoted by $\mathscr{F}$.

The couple $(\Omega, \mathscr{F})$, a set with a σ-algebra of its subsets, is called a *measurable space*. We may refer to Ω as the space, or set, of elementary outcomes. The subsets of Ω in $\mathscr{F}$ are called *events*. We may say that an event A occurs to simply indicate A and that two events A and B occur simultaneously to indicate $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$. The collection $\mathscr{F}$ is made up of those sets which are regular enough that we will be able to measure their likelihood, i.e., assign them a probability of happening. While it is helpful to think of Ω as the set of elementary outcomes of some experiments, you should be cautious as many arguments may not be carried out "ω by ω".

Example 1.2. Here are some examples of σ-algebras:

(i) $\{\emptyset, \Omega\}$ is a σ-algebra. It is often referred to as the *trivial* σ*-algebra* and it is the smallest possible σ-algebra since, by definition, $\{\emptyset, \Omega\} \subseteq \mathscr{F}$ for any σ-algebra $\mathscr{F}$.

(ii) The power set $\mathscr{P}(\Omega)$ is a σ-algebra but is usually too large to work with.

(iii) Let $E \subset \Omega$ be any set and $\mathscr{F}$ be a σ-algebra. Then $\{E \cap A : A \in \mathscr{F}\}$ is a σ-algebra. It is sometimes called the *trace* σ*-algebra*.

(iv) The collection of all sets $A \in \mathscr{P}(\Omega)$ such that either A or A^c is countable is a σ-algebra.

(v) For a nontrivial set $A \subseteq \Omega$, i.e., A is neither empty nor the full space, $\sigma(A) := \{\emptyset, \Omega, A, A^c\}$ is a σ-algebra. It just allows us to say if the event A happened or not but nothing else.

The last example above hints at the crucial property, or interpretation, of σ-algebras: they are conveyors of information. They capture the richness, or poorness, of our ability to distinguish between events, to classify elementary outcomes into events. The richer the σ-algebra the better our ability to classify the elements of Ω. To generalise the above example, we need the following property.

Lemma 1.3. *Let I be an index set and $\{\mathscr{F}_i : i \in I\}$ a collection of σ-algebras. Then*

$$\mathscr{F} := \bigcap_{i\in I} \mathscr{F}_i = \{A \subseteq \Omega : A \in \mathscr{F}_i \text{ for all } i \in I\}$$

is a σ-algebra.

Proof. Exercise. □

Definition 1.4. Let $\mathscr{A}$ be a collection of subsets of Ω. The smallest σ-algebra containing all the sets in $\mathscr{A}$ is denoted $\sigma(\mathscr{A})$ and is called the *σ-algebra generated by $\mathscr{A}$*.

Note that Lemma 1.3 ensures that $\sigma(\mathscr{A})$ is well defined and is simply given by the intersection of all the σ-algebras $\mathscr{F}$ such that $\mathscr{A} \subseteq \mathscr{F}$, a non-empty collection since $\mathscr{A} \subseteq \mathscr{P}(\Omega)$. This result allows us instantly to generate many more interesting σ-algebras. We give now two important examples.

Definition 1.5 (Borel σ-algebra)**.** Let E be a topological space with topology (i.e., collection of open sets) $\mathscr{T}$. The σ-algebra generated by the open sets in E is called the *Borel σ-algebra on E* and is denoted $\mathscr{B}(E) = \sigma(\mathscr{T})$.

Example 1.6 (Borel σ-algebra on $\mathbb{R}$)**.** The following collections of sets

- open sets in $\mathbb{R}$,
- open intervals in $\mathbb{R}$,
- $\{(-\infty, a] : a \in \mathbb{R}\}$,
- $\{(-\infty, a) : a \in \mathbb{R}\}$

all generate the same σ-algebra, namely $\mathscr{B}(\mathbb{R})$.

Definition 1.7 (Product space)**.** Let I be an index set and $(\Omega_i, \mathscr{F}_i)_{i\in I}$ a collection of measurable spaces. Let $\Omega = \prod_{i\in I} \Omega_i$ and $\mathscr{F}$ be the σ-algebra generated by cylinder sets $A = \prod_{i\in I} A_i$, where $A_i \in \mathscr{F}_i$ for all $i \in I$ and $A_i = \Omega_i$ except for *finitely many* $i \in I$. The measurable space $(\Omega, \mathscr{F})$ is called the *product space*. The σ-algebra $\mathscr{F}$ is called the *product σ-algebra* and is sometimes denoted $\times_{i\in I} \mathscr{F}_i$.

When $I = \{1, 2\}$, we simply write $\Omega = \Omega_1 \times \Omega_2$ and $\mathscr{F} = \mathscr{F}_1 \times \mathscr{F}_2$. Note that '$\times$' has a different meaning for these 'products': Ω is the Cartesian product of Ω_1 and Ω_2 but $\mathscr{F}$ is **not** the Cartesian product of $\mathscr{F}_1$ and $\mathscr{F}_2$.

It is often the case that the same $\sigma(\mathscr{A})$ may be generated by many different classes of sets $\mathscr{A}$. For example, the product σ-algebra is already generated by sets where $A_i \neq \Omega_i$ for only one coordinate $i \in I$. This is obvious since σ-algebras are closed under finite intersections so we may get the more general *cylinder sets* from these simple ones. Example 1.6 was also an instance of this phenomena. This example in fact extends to higher dimensions, i.e., to products of $\mathbb{R}$. Indeed, each open subset of $\mathbb{R}^n$ is a countable union of open hypercubes (products of open intervals) and hence $\mathscr{B}(\mathbb{R}^d)$ is generated by d-fold products of open intervals. It follows that $\times_{i=1}^d \mathscr{B}(\mathbb{R}) = \mathscr{B}(\mathbb{R}^d)$ and properties of product spaces will allow us to just focus on real-valued objects. While this will carry over to countable product spaces, it may fail for more general index sets.

Here is a familiar example of a product space, already encountered in Part A Probability.

Example 1.8 (Repeated coin tossing)**.** Consider the experiment consisting in repeated coin tossing. Each toss is naturally represented by $(\Omega_{toss}, \mathscr{F}_{toss})$ with $\Omega_{toss} = \{H, T\}$ and

$$\mathscr{F}_{toss} = \sigma(\{H\}) = \sigma(\{T\}) = \{\emptyset, \Omega_{toss}, \{H\}, \{T\}\} = \mathscr{P}(\Omega_{toss}).$$

Repeated coin tossing is then captured by the product space $(\Omega, \mathscr{F}) = (\prod_{n=1}^{\infty} \Omega_n, \times_{n=1}^{\infty} \mathscr{F}_n)$ where each $(\Omega_n, \mathscr{F}_n) = (\Omega_{toss}, \mathscr{F}_{toss})$. Put differently, $\Omega = \{H, T\}^{\mathbb{N}}$ and $\omega = (\omega_1, \omega_2, \ldots) \in \Omega$ encodes the outcomes of successive tosses. The product σ-algebra $\mathscr{F}$ on Ω is generated by events which only depend on the outcomes of finitely many tosses. As observed above, it is in fact generated by the events $A_n = \{\omega \in \Omega : \omega_n = H\}$, i.e., by events which allows us to encode the result of the n^{th} toss, $n \in \mathbb{N}$. It is clear that for our measurable space to describe our experiment we have to have these in $\mathscr{F}$. It turns out we can not have much more: $\mathscr{F}$ is strictly smaller than $\mathscr{P}(\Omega)$ and it may be impossible to understand and codify the likelihood of evens from outside of $\mathscr{F}$. However, $\mathscr{F}$ proves already to be (perhaps surprisingly) rich. In particular the event A that the asymptotic frequency of heads is equal to $\frac{1}{2}$, or more formally

$$A = \left\{ \omega \in \Omega : \frac{|\{k \leqslant n : \omega_k = H\}|}{n} \to \tfrac{1}{2} \right\}$$

is an element in $\mathscr{F}$, see the problem sheet.

Time and again, we will need to establish that a certain property holds for all sets in a given σ-algebra. This might often be tedious and/or difficult to do directly. The following notions and results offer an alternative.

Definition 1.9 (π- and λ- systems)**.**

- A collection of sets $\mathscr{A}$ is called a *π-system* if it is stable under intersections, i.e., $A, B \in \mathscr{A}$ implies $A \cap B \in \mathscr{A}$.
- A collection of sets $\mathscr{M}$ is called a *λ-system* if
  - $\Omega \in \mathscr{M}$,
  - if $A, B \in \mathscr{M}$ with $A \subseteq B$ then $B \setminus A \in \mathscr{M}$,
  - if $\{A_n\}_{n \geqslant 1} \subseteq \mathscr{M}$ with $A_n \subseteq A_{n+1}$ for all $n \geqslant 1$ then $\bigcup_{n \geqslant 1} A_n \in \mathscr{M}$.

Example 1.10. The collection

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

forms a π-system and $\sigma(\pi(\mathbb{R})) = \mathscr{B}(\mathbb{R})$ by Example 1.6 above.

In some sense, the notions of π- and λ- systems split the properties of a σ-algebra into two, as the following lemma demonstrates.

Lemma 1.11. *A collection of sets $\mathscr{F}$ is a σ-algebra if and only if $\mathscr{F}$ is both a π-system and a λ-system.*

Proof. Clearly a σ-algebra is both a π-system and a λ-system so it remains to establish the converse. Let $\mathscr{F}$ be both a π-system and a λ-system. Let $A, B \in \mathscr{F}$. Then, since $\Omega \in \mathscr{F}$, we also have $A^{\mathrm{c}} = \Omega \setminus A \in \mathscr{F}$ and further

$$A \cup B = \Omega \setminus (A^{\mathrm{c}} \cap B^{\mathrm{c}}) \in \mathscr{F}.$$

Finally, let $\{A_n\}_{n \geqslant 1} \subseteq \mathscr{F}$ be a sequence of sets in $\mathscr{F}$. Then

$$\bigcup_{n \geqslant 1} A_n = \bigcup_{n \geqslant 1} \bigcup_{k=1}^{n} A_k \in \mathscr{F}$$

by the properties of λ-sets as the sequence $B_n = \bigcup_{k=1}^{n} A_k$ is increasing. □

While π-system is a universally adopted terminology, λ-systems are also called d-systems, Dynkin classes or monotone classes. The notions of π- and λ- systems may appear rather artificial at first. In fact, they are very useful. So useful that at some point you may start using them implicitly without thinking much about it. This is because quite often the (abstract) collection of sets which satisfy a certain property Γ is a λ-system. At the same time, it is often easy to verify that Γ holds for all sets in a given π-system $\mathscr{A}$. The following (fundemental!) lemma then says that Γ holds on $\mathscr{F} = \sigma(\mathscr{A})$. We shall use it time and again.

Lemma 1.12 ($\pi - \lambda$ systems Lemma)**.** *Let $\mathscr{M}$ be a λ-system and $\mathscr{A}$ be a π-system. Then,*

$$\mathscr{A} \subseteq \mathscr{M} \Longrightarrow \sigma(\mathscr{A}) \subseteq \mathscr{M}.$$

Proof. Let $\lambda(\mathscr{A})$ denote the intersection of all λ-systems containing $\mathscr{A}$. Then, in analogy to Lemma 1.3, $\lambda(\mathscr{A})$ itself is a λ-system, it is the smallest λ-system containing $\mathscr{A}$. In particular, $\lambda(\mathscr{A}) \subseteq \mathscr{M}$. Naturally, a σ-algebra is by definition a λ-system. If we show that $\lambda(\mathscr{A})$ is itself a σ-algebra it will imply that $\lambda(\mathscr{A}) = \sigma(\mathscr{A})$ and the proof will be complete. By Lemma 1.11, it suffices to show that $\lambda(\mathscr{A})$ is a π-system.

Let $\mathscr{C} = \{A \in \lambda(\mathscr{A}) : A \cap C \in \lambda(\mathscr{A}) \ \forall C \in \mathscr{A}\}$. We first show that $\mathscr{C}$ is a λ-system. Clearly, $\Omega \in \mathscr{C}$. Let $A, B \in \mathscr{C}$ with $A \subseteq B$. Then $(B \setminus A) \cap C = B \cap C \setminus A \cap C \in \lambda(\mathscr{A})$ for all $C \in \mathscr{A}$ so that $B \setminus A \in \mathscr{C}$. Finally, if A_n is an increasing sequence in $\mathscr{C}$ and $A = \bigcup_{n \geqslant 1} A_n$ then $A \cap C = \bigcup_{n \geqslant 1} A_n \cap C \in \lambda(\mathscr{A})$ for all $C \in \mathscr{A}$ and hence $A \in \mathscr{C}$. By definition, $\mathscr{C} \subseteq \lambda(\mathscr{A})$ and, since $\mathscr{A}$ is a π-system, also $\mathscr{A} \subseteq \mathscr{C}$. It follows that $\mathscr{C} = \lambda(\mathscr{A})$.

Now let $\mathscr{D} = \{A \in \lambda(\mathscr{A}) : A \cap C \in \lambda(\mathscr{A}) \ \forall C \in \lambda(\mathscr{A})\}$. As above, we can easily show that $\mathscr{D}$ inherits the λ-system structure from $\lambda(\mathscr{A})$. Further, $\mathscr{C} = \lambda(\mathscr{A})$ above implies that $\mathscr{A} \subseteq \mathscr{D}$. Minimality of $\lambda(\mathscr{A})$ again implies that $\mathscr{D} = \lambda(\mathscr{A})$ and hence $\lambda(\mathscr{A})$ is a π-system. □

One of the most important application of the above result will be to assert that if two measures coincide on a π-system then they coincide on the σ-algebra it generates. In particular, a measure on $\mathscr{B}(\mathbb{R})$ is uniquely specified by its distribution function, i.e., its values on $\pi(\mathbb{R})$ in Example 1.10, see 2.16. The π-λ systems lemma will be used in many other contexts, starting from simple exercises like the following one.

Exercise 1.13. Let $\Omega = \Omega_1 \times \Omega_2$ and $\mathscr{F} = \mathscr{F}_1 \times \mathscr{F}_2$ be a product space. Fix $D \in \mathscr{F}$ and denote $D(\omega_1) := \{\omega_2 : (\omega_1, \omega_2) \in D\}$ be its section for a fixed $\omega_1 \in \Omega_1$. Show that $D(\omega_1) \in \mathscr{F}_2$.

1.2 Random variables

So far, we have developed the basic language to speak of sets and collections of sets. We now want to do the same for functions.

Definition 1.14 (Measurable function)**.** Let $(\Omega, \mathscr{F})$ and $(E, \mathscr{E})$ be measurable spaces. A function $f : \Omega \to E$ is said to be *measurable*, or a *random variable*, if

$$f^{-1}(A) = \{\omega \in \Omega : f(\omega) \in A\} \in \mathscr{F} \quad \forall A \in \mathscr{E}.$$

If this is not clear from the context, we shall say more precisely that f is an E-valued random variable and we may specify the σ-algebras $\mathscr{F}, \mathscr{E}$ with respect to which the measurability is taken. The terms *measurable function* and *random variable* are used interchangeably. Similarly, we will use both f and X as our generic notation for a function (one being canonical in analysis and the other in probability) and switch between the two at will. The following is clear:

Proposition 1.15. *Let $(\Omega, \mathscr{F})$, $(E, \mathscr{E})$ and $(H, \mathscr{H})$ be three measurable spaces. Let $f : \Omega \to E$ and $g : E \to H$ be two random variables. Then $g \circ f$ is a random variable from $(\Omega, \mathscr{F})$ to $(H, \mathscr{H})$.*

Proof. For $A \in \mathscr{H}$, $g^{-1}(A) \in \mathscr{E}$ by measurability of g and $(g \circ f)^{-1}(A) = f^{-1}(g^{-1}(A)) \in \mathscr{F}$ by measurability of f. □

Example 1.16. Let $E = \{0,1\}$ and $\mathscr{E} = \mathscr{P}(E)$. A subset $A \subset \Omega$ is an event if and only if its characteristic function $\mathbf{1}_A$ (equal to 1 for $\omega \in A$ and 0 otherwise) is a random variable.

In this way, random variables generalise events. Several notions developed for events can be transcribed to the context of random variables in a straightforward fashion.

Definition 1.17. Let Ω be a set and $(f_i)_{i\in I}$ a collection of functions from Ω to measurable spaces $(E_i, \mathscr{E}_i)_{i\in I}$. The σ-algebra generated by functions $(f_i)_{i\in I}$, denoted $\sigma(f_i : i \in I)$, is the smallest σ-algebra on Ω with respect to which all f_i, $i \in I$, are measurable.

The above is well-posed thanks to Lemma 1.3. Further, it extends Definition 1.4. Indeed, if $\mathscr{A} = \{A_i : i \in I\}$ is a collection of subsets of Ω then $\sigma(\mathscr{A}) = \sigma(\mathbf{1}_{A_i} : i \in I)$. As a way of example, let us specify a bit more the σ-algebra generated by a single random variable.

Lemma 1.18. *Let X be a random variable from $(\Omega, \mathscr{F})$ to $(E, \mathscr{E})$ and suppose $\mathscr{E} = \sigma(\mathscr{A})$. Then*

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

Proof. It is easy to verify that the inverse $A \to X^{-1}(A)$ preserves all the set operations. In particular, $\{X^{-1}(A) : A \in \mathscr{E}\}$ is a σ-algebra. By definition, it is contained in $\sigma(X)$ and by the minimality of the latter, the two are equal. Denote $\sigma(X;\mathscr{A}) = \sigma(X^{-1}(A) : A \in \mathscr{A})$. The inclusion $\sigma(X;\mathscr{A}) \subseteq \sigma(X)$ is clear. For the reverse, let $\mathscr{G} = \{A \subseteq E : X^{-1}(A) \in \sigma(X;\mathscr{A})\}$. We verify easily that $\mathscr{G}$ is a σ-algebra and since $\mathscr{A} \subseteq \mathscr{G}$ we conclude that $\mathscr{E} \subseteq \mathscr{G}$. It follows that $\sigma(X) \subseteq \sigma(X;\mathscr{A})$ and hence we have an equality. □

From Lemma 1.18 and Example 1.6 we have the following simple property.

Corollary 1.19. *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 $t \in \mathbb{R}$.*

Example 1.20. Consider the product space notation from Definition 1.7. Let X_i denote the coordinate mappings, i.e., $X_i : \Omega \to \Omega_i$ is given by $X_i(\omega) = \omega_i$. Then the product σ-algebra is generated by these coordinate mappings, $\mathscr{F} = \times_{i\in I}\mathscr{F}_i = \sigma(X_i : i \in I)$. In particular, all X_i are measurable. On the other hand, if $(E, \mathscr{E})$ is a measurable space and $Y_i : (E, \mathscr{E}) \to (\Omega_i, \mathscr{F}_i)$ are measurable then the mapping $Y : E \to \Omega$ given by $Y = (Y_i : i \in I)$ is measurable (with respect to $\mathscr{F}$).

We give one more simple example of an abstract random variable.

Example 1.21. Let $\mathscr{G} \subseteq \mathscr{F}$. Then the identity mapping of $(\Omega, \mathscr{F})$ onto $(\Omega, \mathscr{G})$ is a random variable.

Example 1.22. Recall the model for repetitive coin tossing described in Example 1.8. It involved a careful choice of Ω which, in an intuitive sense, was minimal for our purposes. If we wanted to expand our experiment and toss a coin and a dice simultaneously we would not be able to do so using Ω. For this reason, it is usually a much better practice to work with a fixed large $(\Omega, \mathscr{F})$ and to encode our experiments using random variables on Ω. For example, we could take $([0,1], \mathscr{B}([0,1]))$ and let $X_n(\omega) = \mathbf{1}_{\lfloor 2^n\omega \rfloor \text{ is even}}$, $n \geqslant 1$, where 0 is even. It is easy to check that X_n is a random variable and $X_n \in \{0,1\}$. We shall see these are just as good a way to express the coin tossing experiment.

Remark. The above example makes it clear that σ-algebra may be thought of as a representation of our information, as already mentioned in the discussion following Example 1.2. Think of a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ as an abstract carrier for randomness. Random variables on Ω represent outcomes of experiments, random things happening. In Example 1.22, $(X_n)_{n\geqslant 1}$ represented successive coin tosses. Then $\mathscr{G}_n = \sigma(X_k : 1 \leqslant k \leqslant n)$ is the σ-algebra corresponding to the information about the first n tosses. It is the smallest σ-algebra which allows us to recognise the outcomes of these tosses. $\mathscr{G} = \sigma(X_n : n \geqslant 1)$ is the σ-algebra generated by all the sequence of tosses but it will typically be much smaller than $\mathscr{F}$, which represents "the ultimate knowledge".

From now on, unless explicitly stated otherwise, we shall consider random variables with values in $E = \mathbb{R}$ or $\overline{\mathbb{R}} = [-\infty, \infty]$. In this case we *always* consider measurability relative to the Borel sets: $\mathscr{E} = \mathscr{B}(\mathbb{R})$ or $\mathscr{B}(\overline{\mathbb{R}})$.

Example 1.23. Let (E,d) be a metric space and let $\mathscr{B}(E)$ be the Borel σ-algebra generated by its open sets. Then the Borel σ-algebra on E is equal to the Baire σ-algebra on E:

$$\mathscr{B}(E) = \sigma(f : E \to \mathbb{R} | f \text{ continuous}).$$

As in Corollary 1.19, for f to be measurable it is enough to check that $f^{-1}(O) \in \mathscr{B}(E)$ for an open interval O and this follows from continuity. In particular, the "$\supseteq$" inclusion follows. For a closed set $F \subseteq E$, let $f_F(x) = d(x,F)$ be the distance of x to F. Then f is continuous and $F = f_F^{-1}(\{0\})$ is an element of the right hand side. This gives the reverse inclusion "$\subseteq$" and hence the equality.

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).

Proposition 1.24. *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 Borel measurable. Then, whenever they make sense*[1]*, the following are also measurable functions on $(\Omega, \mathscr{F})$:*

$$f_1 + f_2, \quad f_1 f_2, \quad \max\{f_1, f_2\}, \quad \min\{f_1, f_2\}, \quad f_1/f_2, \quad h \circ f$$
$$\sup_n f_n, \quad \inf_n f_n, \quad \limsup_{n\to\infty} f_n, \quad \liminf_{n\to\infty} f_n.$$

Definition 1.25. A measurable function f on $(\Omega, \mathscr{F})$ is called a *simple function* if

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

for some $n \geqslant 1$ and where each $E_k \in \mathscr{F}$ and each $a_k \in \mathbb{R}$. The *canonical form* of f 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.

Clearly, a simple function is measurable. Conversely, any measurable function can be obtained as a limit of simple functions. This gives us:

Lemma 1.26. *Let $(\Omega, \mathscr{F})$ be a measurable space. A function $X : \Omega \to \mathbb{R}$ is measurable if and only if it is a limit of simple functions. Further, if f is bounded from below (*resp. *bounded), the limit can be taken to be increasing (*resp. *uniform).*

Proof. That a limit of simple functions is a measurable function follows from Proposition 1.24. Now let X be a random variable and define

$$X_n = \sum_{k \in \mathbb{Z} \cap [-4^n, 4^n]} \frac{k}{2^n} \mathbf{1}_{\frac{k}{2^n} < X \leqslant \frac{k+1}{2^n}}, \quad n \geqslant 1. \tag{7}$$

Let $\Omega_n^+ := \{\omega \in \Omega : X(\omega) \leqslant 2^n\}$, $\Omega_n^- := \{\omega \in \Omega : X(\omega) > -2^n\}$ and $\Omega_n = \Omega_n^- \cap \Omega_n^+$. The result follows by noting that $\sup_{\omega \in \Omega_n} |X_n(\omega) - X(\omega)| \leqslant 2^{-n}$ and $X_n \leqslant X_{n+1}$ on Ω_n^-. □

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

The above remains true for $X : \Omega \to \overline{\mathbb{R}}$ except the sequence may no longer be increasing if X takes the value $-\infty$. The details are left to the reader.

We give a simple example of a result where approximating a general random variable with simple ones is used in the proof. This result also highlights further the information interpretation of a σ-algebra and shows that the abstract measurability definition agrees with a more intuitive one of 'being a function of'.

Theorem 1.27. *Let X be a random variable on $(\Omega, \mathscr{F})$ with values in a measurable space $(E, \mathscr{E})$ and let g be a real-valued random variable on $(\Omega, \mathscr{F})$. Then g is $\sigma(X)$-measurable if and only if $g = h \circ X$ for some real-valued random variable on $(E, \mathscr{E})$.*

Deep Dive

Proof. One direction is clear: $g = h \circ X$ is a real-valued random variable. For the other direction, start with g and suppose it takes at most countably many distinct values $(a_n)_{n \geqslant 1}$. The sets $A_n = g^{-1}(\{a_n\})$ are pairwise disjoint and each is an element of $\sigma(X)$ and hence, by Lemma 1.18, $A_n = X^{-1}(B_n)$ for some $B_n \in \mathscr{E}$. Note that we might have $B_n \cap B_m \neq \emptyset$ but the points in the intersection are not in the range of values of X. Consequently, if we set $C_n := B_n \setminus \bigcup_{k=1}^{n-1} B_k$ then $C_n \in \mathscr{E}$ are pairwise disjoint and $X^{-1}(C_n) = A_n \setminus \bigcup_{k=1}^{n-1} A_k = A_n$. If we put $h = \sum_{n \geqslant 1} a_n \mathbf{1}_{C_n}$ then $g = h \circ X$ as required.

For a general g, let $g_n \uparrow g$ be the sequence of simple random variables converging to g given by Lemma 1.26. By the above, we can write each $g_n = h_n \circ X$. Let $H = \{e \in E : h_n(e) \text{ converges}\}$. Recall that both $\limsup h_n$ and $\liminf h_n$ are measurable and so $H = \{\limsup h_n = \liminf h_n\}$ is measurable. Further, $X(\Omega) \subseteq H$ since $g_n \uparrow g$. It follows that $h(\omega) := (\lim_{n \to \infty} h_n(\omega)) \mathbf{1}_H(\omega)$ is measurable and satisfies $g = h \circ X$. □

A lot of results, e.g., when developing the integration theory, can be shown using a "bare hands method" powered by Lemma 1.18. The schematic is as follows: to establish a "linear" result for all functions in a given class, say for all bounded measurable functions, we proceed in steps:

- first establish the result for indicators of a measurable set, where it usually holds by definition;
- by linearity extend this to all simple functions or all positive simple functions;
- take limits, using a suitable convergence theorem, extend the result to all functions, or all positive functions;
- if needed, write $X = X^+ - X^-$ and use the above to pass from positive to all functions.

Such an approach allows one to see the theory "grow" and demystifies it. It is useful to go through the steps above once in detail but later one can apply these semi-automatically. However, sometimes it is very difficult to use the above bare-hands approach and it becomes necessary to turn to a functional equivalent of Lemma 1.12. This is known as the Monotone Class Theorem. It comes in many variants and flavours and we state just one. It usually gives a quick and elegant proof but may at first appear to be a magic trick of sorts.

Theorem 1.28 (Monotone Class Theorem)**.** *Let $\mathscr{H}$ be a class of bounded functions from Ω to $\mathbb{R}$ satisfying the following conditions:*

(i) $\mathscr{H}$ is a vector space over $\mathbb{R}$,

(ii) the constant function 1 is in $\mathscr{H}$,

(iii) if $(f_n)_{n \geqslant 1} \subseteq \mathscr{H}$ such that $f_n \nearrow f$ for a bounded function f, then $f \in \mathscr{H}$.

If $\mathscr{C} \subseteq \mathscr{H}$ is stable under pointwise multiplication then $\mathscr{H}$ contains all bounded $\sigma(\mathscr{C})$-measurable functions.

Deep Dive

We outline now the proof of the above important result. First, we make the following simple observation.

Lemma. *In the setup of Theorem 1.28, $\mathscr{H}$ is closed under uniform limits.*

Proof. Let f_n be a sequence of functions in $\mathscr{H}$ converging uniformly to some f. Passing to a subsequence, we can assume that $\|f_n - f\|_{\sup} \leqslant 2^{-n}$, where $\|f\|_{\sup} = \sup_{\omega\in\Omega} |f(\omega)|$. Now we can modify the sequence so that it is increasing. Set $g_n = f_n - 2^{1-n}$. Then $g_n - g_{n-1} = f_n - f_{n-1} + 2^{1-n} \geqslant 2^{-n} \geqslant 0$. Also,

$$\|g_n\|_{\sup} = \|f_1 + \sum_{k=2}^{n} f_k - f_{k-1} - 2^{1-n}\|_{\sup} \leqslant \|f_1\|_{\sup} + 3$$

the sequence is uniformly bounded so that its limit is also bounded and hence $\mathscr{H} \ni \lim g_n = \lim f_n = f$. □

Proof of Theorem 1.28 – special case. Consider first the case when $\mathscr{C} = \{\mathbf{1}_A : A \in \mathscr{A}\}$ for a π-system $\mathscr{A}$. Here Theorem 1.28 is a functional equivalent of Lemma 1.12. To see this, simply check that the properties of $\mathscr{H}$ mean that the family of sets $E \subseteq \Omega$ for which $\mathbf{1}_E \in \mathscr{H}$ forms a λ-system. Lemma 1.12 now shows that $\mathbf{1}_E \in \mathscr{H}$ for all $E \in \sigma(\mathscr{A})$ and Lemma 1.26 tells us that any bounded measurable function is a uniform limit of simple functions and hence, by the above lemma, is also in $\mathscr{H}$, as required. □

Proof of Theorem 1.28 – reduction to the special case. We prove the general statement by reducing it to the special case treated above. Note that without any loss of generality we can assume that $1 \in \mathscr{C}$. Let $\mathscr{A}_0$ be the algebra of functions generated by $\mathscr{C}$. Given that $\mathscr{C}$ is already closed under multiplication, $\mathscr{A}_0$ is simply the linear span of $\mathscr{C}$. Let $\mathscr{A}$ be the closure of $\mathscr{A}_0$ under uniform convergence. By the above lemma, $\mathscr{A} \subset \mathscr{H}$ and we check that $\mathscr{A}$ is still an algebra of functions. Take $f \in \mathscr{A}$ and since it is a bounded function we can take a closed interval $R \subseteq \mathbb{R}$ with $f(\omega) \in R$, $\omega \in \Omega$. On R, by the Weierstrass approximation theorem, we can approximate the function $x \to |x|$ uniformly using a sequence of polynomials p_n. Note that $p_n \circ f \in \mathscr{A}$ and hence also its uniform limit $|f|$. It then follows that $\mathscr{A}$ is closed under $\wedge$ and $\vee$ (observe that $f^+ = (|f| + f)/2$ and $f \vee g = f + (g-f)^+$ etc.). Now, for any $f \in \mathscr{A}$ and any $a \in \mathbb{R}$ we have

$$\mathscr{A} \ni n(f-a)^+ \wedge 1 \uparrow \mathbf{1}_{f^{-1}((a,\infty))}$$

and hence the limit is in $\mathscr{H}$, i.e., $\{\mathbf{1}_D : D \in \mathscr{D}\} \subseteq \mathscr{H}$, where $\mathscr{D} = \{f^{-1}((a,\infty)) : f \in \mathscr{A}, a \in \mathbb{R}\}$. Note that $\{f > a\} \cap \{g > b\} = \{(f-a)^+(g-b)^+ > 0\}$ so that $\mathscr{D}$ is a π-system and by Lemma 1.18, $\sigma(\mathscr{D}) = \sigma(f : f \in \mathscr{A})$. This reduces the general result to the special case previously considered. □

Remark. Following the ideas of the proof, one can devise other statements and variants of the Monotone Class Theorem. For example, instead of supposing that $\mathscr{C}$ is stable under multiplication, one can consider cones of non-negative functions stable under taking minimum: $f, g \in \mathscr{C}$ then $af \wedge bg \in \mathscr{C}$ for $a, b \in \mathbb{R}_+$. Then the uniform closure of $\mathscr{A} = \{f - g : f, g \in \mathscr{C}\}$ is a vector space stable under $\wedge, \vee$ and one can show it is also stable under multiplication, first approximating $x \to x^2$ and hence showing that $f^2 \in \mathscr{A}$ for $f \in \mathscr{A}$.

An important common example is the special case of $\mathscr{C} = \{\mathbf{1}_A : A \in \mathscr{A}\}$ for a π-system $\mathscr{A}$. In this case, Theorem 1.28 can be deduced from Lemma 1.12 and Lemma 1.26. Let us give now one application of the above result and use it to highlight this relationship with the π-λ systems lemma.

Lemma 1.29. *Let $(\Omega, \mathscr{F})$ be the product space of two measurable spaces $(\Omega_i, \mathscr{F}_i)$, $i = 1, 2$. If $f : \Omega \to \mathbb{R}$ is measurable then*

- *for each $\omega_1 \in \Omega_1$, $\Omega_2 \ni \omega_2 \to f(\omega_1, \omega_2)$ is $\mathscr{F}_2$-measurable and*

- *for each* $\omega_2 \in \Omega_2$, $\Omega_1 \ni \omega_1 \to f(\omega_1, \omega_2)$ *is* $\mathscr{F}_1$*-measurable.*

The first proof: using the Monotone Class Theorem. Let $\mathscr{H}$ be the class of bounded functions $h : \Omega \to \mathbb{R}$ which satisfy the assertion of the lemma. Clearly $\mathscr{H}$ satisfies the assumptions of the Monotone Class Theorem (Theorem 1.28) and contains the functions $h = \mathbf{1}_{A_1 \times A_2}$ for $A_i \in \mathscr{F}_i$, $i = 1, 2$. These rectangles generate $\mathscr{F}$ and we conclude that $\mathscr{H}$ contains all bounded measurable functions. For an unbounded f, we use the result for $f_n = (f \vee -n) \wedge n$, which is bounded, and use that limits of measurable functions are measurable. □

The second proof: using π*-*λ *systems lemma.* An application of π-λ systems lemma shows that the statement holds for $f = \mathbf{1}_D$ for $D \in \mathscr{F}$, see Exercise 1.13. It thus also holds for simple functions. It remains to apply Lemma 1.26 and note that limits of measurable functions are measurable. □

2 Measures

Now that we have the basic ingredients, we shall start to measure them! In Part A Integration we conceptualised the idea of length (or volume) and saw that there is a good way to construct a measure of length, the Lebesgue measure Leb, which can be assigned in a consistent way to any set in $\mathscr{B}(\mathbb{R})$, or in $\mathscr{M}_{Leb}$ more generally. We want to now take a more abstract view and develop an abstract theory of measuring sets. We formalise the idea of assigning a likelihood or a *probability* to a set and of doing this in a consistent manner.

2.1 Measures and Measurable spaces

Definition 2.1 (Set functions)**.** Let $\mathscr{A}$ be a collection 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 *countably additive*, or σ*-additive*, if for all sequences (A_n) of disjoint sets in $\mathscr{A}$ with $\bigcup_{n=1}^{\infty} A_n \in \mathscr{A}$

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

Recall that a *measurable space* is a pair $(\Omega, \mathscr{F})$ where $\mathscr{F}$ is a σ-algebra on Ω.

Definition 2.2 (Measure space)**.** 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. Then μ is a *measure* on $(\Omega, \mathscr{F})$.

In short, a measure space is a set Ω equipped with a σ-algebra $\mathscr{F}$ and a countably additive set function μ on $\mathscr{F}$. Note that any measure μ is also additive and increasing. Being a measure is relative to the context of the given measurable space hence we say, as above, that μ is a measure on $(\Omega, \mathscr{F})$. However, for simplicity, when the choice of $(\Omega, \mathscr{F})$ is unambiguous, we will often just say that μ is a measure on $\mathscr{F}$ or on Ω. We summarise now some easy properties of measures.

Proposition 2.3. *Let $(\Omega, \mathscr{F}, \mu)$ be a measurable space and $A, B, A_n, B_n \in \mathscr{F}$, $n \geqslant 1$. Then*

(i) $A \cap B = \emptyset \Longrightarrow \mu(A \cup B) = \mu(A) + \mu(B)$ *(additive)*

(ii) $A \subseteq B \Longrightarrow \mu(A) \leqslant \mu(B)$ *(increasing)*

(iii) $\mu(A \cup B) + \mu(A \cap B) = \mu(A) + \mu(B)$

(iv) $A_n \uparrow A$, *then* $\mu(A_n) \uparrow \mu(A)$ *as* $n \to \infty$ *(continuous from below)*

(v) $B_n \downarrow B$, $\mu(B_k) < \infty$ *for some* $k \in \mathbb{N}$, *then* $\mu(B_n) \downarrow \mu(B)$ *as* $n \to \infty$ *(continuous from above)*

(vi) $\mu\left(\bigcup_{n\geqslant 1} A_n\right) \leqslant \sum_{n\geqslant 1} \mu(A_n)$ *(σ-subadditive)*

Proof. The proof is mostly a direct consequence of the defining properties of a measure and is left as an exercise. We just show (iv). Define sets $D_1 := A_1$ and $D_n := A_n \setminus A_{n-1}$ for $n \geqslant 1$ and note these are pairwise disjoint since $A_{n-1} \subseteq A_n$. Further, $A_n = \bigcup_{k\leqslant n} D_k$. It follows that

$$\mu(A) = \mu\left(\bigcup_{n\geqslant 1} A_n\right) = \mu\left(\bigcup_{n\geqslant 1} D_n\right) = \sum_{n\geqslant 1} \mu(D_n) = \lim_{n\to\infty} \sum_{k=1}^{n} \mu(D_k) = \lim_{n\to\infty} \mu(A_n),$$

where the third equality is by countable additivity of μ and the last equality is by finite additivity of μ. □

Note that $\mu(B_k) < \infty$ is essential in (v): for a counter-example take $B_n = (n, \infty) \subseteq \mathbb{R}$ and Lebesgue measure. The following lemma adds a converse to (iv) above and asserts that an additive set function is countably additive if and only if it is continuous from above.

Lemma 2.4. *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 Proposition 2.3; the other is an exercise. □

Definition 2.5 (Types of measure space)**.** Let $(\Omega, \mathscr{F}, \mu)$ be a measure space.

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

Definition 2.6 (Null sets, a.e.)**.** Let $(\Omega, \mathscr{F}, \mu)$ be a measure space. We say that a set A is *null* if $\mu(A) = 0$. We say that a property holds *almost everywhere* (a.e.), or for almost every $\omega \in \Omega$, if it holds outside of a null set.

If $\mathbb{P}$ is a probability measure we typically say that a property holds *almost surely (a.s.)* instead of almost everywhere. For instance, we will say that two events are a.s. equal, $A = B$ a.s., if $\mathbb{P}(A \triangle B) = 0$. Similarly, for two random variables X, Y we say that $X = Y$ a.s., if $\mathbb{P}(X \neq Y) = 0$. If the reference measure is not obvious we shall indicate it explicitly, e.g., by saying μ-null or $\mathbb{P}$-a.s.

The structure of its null sets tells us a lot about a given measure. Intuitively speaking, if two measures have the same null sets, then one is a re-weighted version of the other. If their null sets differ then one can not go from one measure to another – no re-weighting will resurrect zero into a positive number. This intuition will be made precise in Theorem 4.9 but we can already define the relevant concept.

Definition 2.7. Let μ, ν be two measures on a measurable space $(\Omega, \mathscr{F})$. We say that ν *is absolutely continuous with respect to* μ, and write $\nu \ll \mu$, if $\mu(A) = 0$ for some $A \in \mathscr{F}$ implies $\nu(A) = 0$.
We say that μ *and* ν *are equivalent*, and write $\mu \sim \nu$, if $\nu \ll \mu$ and $\mu \ll \nu$.

Let us now specify some easy examples of measures.

Example 2.8. (i) Let $(\Omega, \mathscr{F})$ be a measurable space. The zero function, $\mu(A) = 0$ for all $A \in \mathscr{F}$, defines a measure. Likewise, ν given by $\nu(\emptyset) = 0$, $\nu(A) = +\infty$ for all $\emptyset \neq A \in \mathscr{F}$ also defines a measure. Clearly both are trivial examples and are well defined for any σ-algebra $\mathscr{F}$.

(ii) Let $(\Omega, \mathscr{F})$ be a measurable space and fix $\omega \in \Omega$. Then δ_ω defined via $\delta_\omega(A) = \mathbf{1}_{\omega \in A}$ defines a measure. It is called the *Dirac measure in ω* or the *point mass in ω*.

(iii) On $\mathbb{R}$ consider the σ-algebra $\mathscr{A}$ of sets which are either countable or have a countable complement, see Example 1.2 (iv). Then $\mu(A) = 0$ for countable A and $\mu(A) = 1$ otherwise, $A \in \mathscr{A}$, defines a probability measure on $\mathscr{A}$.

(iv) Let $(\Omega, \mathscr{F})$ be a measurable space. For $A \in \mathscr{F}$, set $\mu(A) = |A|$, the number of elements in A, if A is finite and $\mu(A) = +\infty$ if A is infinite. Then μ is the *counting measure* on Ω.

It is difficult to construct explicitly, in a manner similar to the above, less trivial examples. We shall develop more systematic ways to build measures later. Here, we give one more example which connects our abstract notions with the intuitive counting notions.

Example 2.9 (Discrete measure theory)**.** Let Ω be a countable set. A *mass function* on Ω is any function $p : \Omega \to [0,\infty]$. Given such a p we can define a measure on $(\Omega, \mathscr{P}(\Omega))$ by setting $\mu(A) = \sum_{x\in A} p(x)$. In the notation of Example 2.8 (ii), $\mu = \sum_{x\in\Omega} p(x)\delta_x$.

Conversely, given a measure μ on $(\Omega, \mathscr{P}(\Omega))$ we can define the corresponding mass function by $p(x) = \mu(\{x\})$. Consequently, for a countable Ω, there is a one-to-one correspondence between measures on $(\Omega, \mathscr{P}(\Omega))$ and mass functions on Ω.

Note also, that if μ, ν are two measures with their respective mass functions p, r then $\nu \ll \mu$ if and only if $p(x) = 0$ implies $r(x) = 0$.

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 and work "ω by ω". 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 2.9). 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 2.11 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 is resolved using the π-λ systems Lemma (Lemma 1.12).

Theorem 2.10 (Uniqueness of extension)**.** *Let μ_1 and μ_2 be measures on a measurable space $(\Omega, \mathscr{F})$ and let $\mathscr{A} \subseteq \mathscr{F}$ be a π-system with $\sigma(\mathscr{A}) = \mathscr{F}$. If $\mu_1(\Omega) = \mu_2(\Omega) < \infty$ and $\mu_1 = \mu_2$ on $\mathscr{A}$, then $\mu_1 = \mu_2$.*

Proof. In view of Lemma 1.12 it suffices to verify that $\{A \in \mathscr{F} : \mu_1(A) = \mu_2(A)\}$ is a λ-system, which is left as an exercise. □

Deep Dive

Note that the assumption $\mu_1(\Omega) = \mu_2(\Omega) < \infty$ is an important one. The result, as usual, extends to σ-finite measures with a common sequence of sets K_n with $\mu_1(K_n) = \mu_2(K_n) < \infty$. However, it may fail for infinite measures. Consider, for example, $\mu_1(A)$ is zero or infinity according to whether the set A has no rational points or at least one rational point, and let $\mu_2(A) = \infty$ for all $A \neq \emptyset$. Taking $\mathscr{A}$ the family of open intervals, we have $\mu_1(A) = \mu_2(A)$ for $A \in \mathscr{A}$ but the two measures are not equal.

We can rephrase Theorem 2.10 simply saying that two probability measures which coincide on a π-system also agree on the σ-algebra generated by that π-system. That deals with uniqueness, but what about existence?

Theorem 2.11 (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 2.12. If $\mu_0(\Omega) < \infty$, then Theorem 2.10 tells us that μ is unique, since an algebra is certainly a π-system. This extends to the σ-finite case if we can take $K_n \in \mathscr{A}$ in Definition 2.5. Indeed, we then obtain uniqueness of extension of μ_0 to a measure on $\{A \cap K_n : A \in \mathscr{F}\}$, for $n \geqslant 1$, and hence also on $\mathscr{F}$.

The Carathéodory Extension Theorem doesn't quite solve the problem of constructing measures on σ-algebras – it reduces it to constructing countably additive set functions on algebras; we shall see several examples. The idea of proof of the Carathéodory Extension Theorem is rather simple, even if the details are

tedious. 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.

We comment now on two generic ways to construct measures: through restrictions and by finite sums. Subsequent sections will develop in detail other methods. First, the following is immediate and allows to construct measure spaces by restricting the σ-algebra.

Lemma 2.13. *Let $(\Omega, \mathscr{F}, \mu)$ be a measure space and $\mathscr{G} \subseteq \mathscr{F}$ a σ-algebra. Then $(\Omega, \mathscr{G}, \mu_{|\mathscr{G}})$, where $\mu_{|\mathscr{G}}$ is the restriction of μ to $\mathscr{G}$, is a measure space.*

The reverse direction however is unclear and often untrue: given a measure space $(\Omega, \mathscr{F}, \mu)$ and a larger σ-algebra $\mathscr{H} \supseteq \mathscr{F}$ it may be possible or impossible to extend μ to $\mathscr{H}$ and, if possible, such an extension does not have to be unique. Clearly, Carathéodory Extension Theorem is not useful here since $\sigma(\mathscr{F}) = \mathscr{F}$. Second, sums of measures are measures.

Lemma 2.14. *Let $(\Omega, \mathscr{F})$ be a measurable space and $(\mu_n)_{n \geqslant 1}$ a sequence of probability measures on $\mathscr{F}$. Fix a sequence of positive numbers $(a_n)_{n \geqslant 0}$ with $\sum_{n \geqslant 1} a_n = 1$. Then μ, defined by $\mu(A) = \sum_{n \geqslant 1} a_n \mu_n(A)$ is also a probability measure on $\mathscr{F}$.*

The above lemma follows once we know we can exchange the order of summation in a double (countable) sum of positive numbers. This will in particular follow from (generalised) Fubini's theorem (Theorem 4.24) which we will see later in these lectures.

If μ is a finite measure then $\mathbb{P}(A) := \mu(A)/\mu(\Omega)$ is a probability measure. It is therefore *with no loss of generality* that in the remainder of this course, we shall mostly work with probability measures. We will comment when these results extend to the σ-finite case.

2.2 Conditional probability

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $B \in \mathscr{F}$ a set with $\mathbb{P}(B) > 0$. Define a new measure μ, also denoted $\mathbb{P}(\cdot|B)$ on $\mathscr{F}$ by

$$\mu(A) = \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad A \in \mathscr{F}. \tag{8}$$

Then it is an easy exercise to check that μ is a probability measure on $\mathscr{F}$. Alternatively, we could define μ as a probability measure on $(B, \mathscr{G})$ with $\mathscr{G} = \{A \cap B : A \in \mathscr{F}\}$ by simply putting $\mu(A) = \mathbb{P}(A)/\mathbb{P}(B)$ for $A \in \mathscr{G}$.

The above definition agrees with what you have seen in Prelims and Part A probability courses. Here we will want to get more serious about conditioning. Conditioning should be relative to information one has and we saw earlier that σ-algebra were the natural carriers or descriptions for information content. We would thus like to condition on a σ-algebra. In the example above, we could replace B by its complement B^{c} and obtain a new measure $\mathbb{P}(A|B^{\mathrm{c}})$. Now, for any $\omega \in \Omega$, we have either $\omega \in B$ or $\omega \in B^{\mathrm{c}}$ so it is natural to define

$$\mathbb{P}(A|\sigma(B))(\omega) := \mathbb{P}(A|B)\mathbf{1}_B(\omega) + \mathbb{P}(A|B^{\mathrm{c}})\mathbf{1}_{B^{\mathrm{c}}}(\omega). \tag{9}$$

In this way, for a fixed $\omega \in \Omega$, $\mathbb{P}(\cdot|\sigma(B))(\omega)$ is a probability measure but for a fixed $A \in \mathscr{F}$, $\mathbb{P}(A|\sigma(B))(\cdot)$ is a random variable (taking two values). It is the latter point of view which will prove very powerful and will set probability alive (and apart from analysis) as we will see in §6.

2.3 Measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$

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 $\mathbb{R}$?

Definition 2.15. Let μ be a probability measure on $\mathscr{B}(\mathbb{R})$. The *distribution function* of μ is the function $F_\mu : \mathbb{R} \to \mathbb{R}$ defined by $F_\mu(x) = \mu((-\infty, x])$.

The function F_μ has the following properties:

(i) F_μ is increasing, i.e., $x < y$ implies $F_\mu(x) \leqslant F_\mu(y)$,

(ii) $F_\mu(x) \to 0$ as $x \to -\infty$ and $F_\mu(x) \to 1$ as $x \to \infty$, and

(iii) F_μ is *right continuous*: $y \downarrow x$ implies $F_\mu(y) \to F_\mu(x)$.

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 Proposition 2.3, $F_\mu(y_n) = \mu(A_n) \downarrow \mu(A) = F_\mu(x)$. We often write $F_\mu(-\infty) = 0$ and $F_\mu(\infty) = 1$ as shorthand for the second property.

Any function $F : \mathbb{R} \to \mathbb{R}$ which satisfies the same three properties as F_μ above will be called a *distribution function* on $\mathbb{R}$. 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 2.16 (Lebesgue)**.** *Let $F : \mathbb{R} \to \mathbb{R}$ be a distribution function, i.e., F is 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 $\mathbb{R}$. Before proving this result let us state an immediate corollary.

Corollary 2.17. *There exists a unique Borel measure* $\texttt{Leb}$ *on $\mathbb{R}$ such that for all $a, b \in \mathbb{R}$ with $a < b$, $\texttt{Leb}((a, b]) = b - a$. The measure* $\texttt{Leb}$ *is the Lebesgue measure on $\mathscr{B}(\mathbb{R})$.*

Proof. The statement with $\mathbb{R}$ replaced by $(0, 1]$ follows from Theorem 2.16 with $F(x) = 0$ on $(-\infty, 0]$, $F(x) = x$ on $[0, 1]$ and $F(x) = 1$ on $[1, \infty)$. This gives us the Lebesgue measure $\texttt{Leb}_k$ on any $(k, k+1]$. We set $\texttt{Leb}(A) = \sum_{k \in \mathbb{Z}} \texttt{Leb}_k(A \cap (k, k+1])$ and easily check it defines a measure on $\mathscr{B}(\mathbb{R})$ with the right properties. Uniqueness follows from Remark 2.12. □

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 of $\mathscr{B}(\mathbb{R})$ is that it has a simple definition independent of the measure; recall that which sets are null depends on which measure is being considered.

Proof of Theorem 2.16. 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 2.10. 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 2.11, so first we need a suitable algebra. For $-\infty \leqslant a \leqslant b < \infty$, let $I_{a,b} = (a,b]$, and set $I_{a,\infty} = (a,\infty)$. 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 2.4, 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

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

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

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

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

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 $E_n \supseteq E_{n+1}$. 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 $A_n \downarrow \emptyset$. □

We now have a very rich class of measures to work with. The measures μ described in Theorem 2.16 are sometimes called *Lebesgue–Stieltjes measures*. 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 – as defined in the next section – 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.

2.4 Pushforward (image) measure

So far we saw how to construct measures by specifying their action on a generating algebra of sets. This works in general, as Theorem 2.11 shows, and led to a complete description of probability measures on $\mathbb{R}$. We now introduce a second fundamental way measures can be built: they are transported between spaces via functions.

Definition 2.18. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let X be a random variable from $(\Omega, \mathscr{F})$ to $(E, \mathscr{E})$. Then

$$\mathbb{Q}(A) = \mathbb{P}(X^{-1}(A)), \quad A \in \mathscr{E},$$

defines a measure on $(E, \mathscr{E})$, the *image measure* of μ via X, or the *pushforward measure*. We write $\mathbb{Q} = \mathbb{P} \circ X^{-1}$ and also call it the *law* or the *distribution* of X.

Put differently, to measure a set in E, we transport it back into Ω via X^{-1} and then measure it there using $\mathbb{P}$. It is a matter of a simple exercise to verify that $\mathbb{Q}$ is a measure. This follows since X^{-1} preserves set operations.

Example 2.19. Let X be a real-valued random variable on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Then $\mathbb{P} \circ X^{-1}$ is a probability measure on $\mathbb{R}$, the distribution or the law of the variable X, and we often denote it by μ_X. We have $\mu_X((-\infty, a]) = \mathbb{P}(X \leqslant a) =: F_X(a)$ is the distribution function of X, or of the measure $\mathbb{P} \circ X^{-1}$. Note that μ_X is the Lebesgue-Stieltjes measure associated to F_X through Theorem 2.16.

Let F be a distribution function on $\mathbb{R}$ and μ_F the Lebesgue-Stieltjes measure associated to F through Theorem 2.16. Then the identity mapping on $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mu_F)$, i.e., $X(\omega) = \omega$, is a random variable distributed according to μ_F. The following example gives another, more canonical, way for such a construction.

Example 2.20. Let F be a distribution function on $\mathbb{R}$. Define its right-continuous inverse $F^{-1}(z) = \inf\{y : F(y) > z\}$, which is also known as the *quantile function*. Then a random variable X on $([0,1], \mathscr{B}([0,1]), \mathtt{Leb})$, given by $X(\omega) = F^{-1}(\omega)$ is distributed according to μ_F, $\mu_X = \mu_F$.

To show this, first note that F^{-1} is increasing and hence measurable. Then note that

$$\{\omega : \omega < F(x)\} \subseteq \{\omega : F^{-1}(\omega) \leqslant x\} \subseteq \{\omega : \omega \leqslant F(x)\}$$

and the outer sets both have the same Leb measure $F(x)$. It thus follows that

$$F_X(x) = \mathtt{Leb}(X \leqslant x) = \mathtt{Leb}(F^{-1} \leqslant x) = \mathtt{Leb}(\{\omega : F^{-1}(\omega) \leqslant x\}) = \mathtt{Leb}(\{\omega : \omega < F(x)\}) = F(x).$$

This tells us that we can always construct random variables with a given distribution. For two random variables X, Y, defined possibly on different probability spaces, we shall often write $X \sim Y$ to denote $\mu_X = \mu_Y$, i.e., that X and Y have the same distribution. A lot of properties of random variables will in fact just functions of their distribution and not their particular definition.

Example 2.21 (Marginal measure)**.** Consider a probability measure $\mathbb{P}$ on a product space from Definition 1.7, $(\Omega, \mathscr{F}) = (\prod_{i\in I}\Omega_i, \times_{i\in I}\mathscr{F}_i)$. Let $X_i(\omega) = \omega_i$, $1 \leqslant i \leqslant d$, be random variables given by coordinate projections, see Example 1.20. Then $\mu_i := \mu_{X_i}$ is called the i^{th} marginal measure of μ. Note that μ_i is a probability measure on $(\Omega_i, \mathscr{F}_i)$ and

$$\mu_i(A) = \mu\left(\Omega_1 \times \ldots \Omega_{i-1} \times A \times \Omega_{i+1} \times \ldots \Omega_n\right), \quad A \in \mathscr{F}_i. \tag{10}$$

Note that μ determines its marginals but that the marginal distributions do not determine μ. Indeed, it is easy to construct examples of $\mu \neq \nu$ with the same marginals. One way to do this is to use the method of the next example.

Example 2.22 (Joint distribution)**.** Let X, Y be real-valued random variables on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Then, by Example 1.20, (X, Y) is an $\mathbb{R}^2$-valued random variable. Its distribution, $\mu_{(X,Y)}$ is called the *the joint law* of X and Y. It is easy to verify (and follows instantly from Lemma 2.23 below) that its marginals are given by μ_X and μ_Y, the distributions of X and Y respectively. However the joint law encodes also how the two variables behave jointly, i.e., their (in)dependence.

Let us finally note that the operation of taking the image law is transitive.

Lemma 2.23. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, $(E, \mathscr{E})$ and $(G, \mathscr{G})$ two measurable spaces and $X : \Omega \to E$, $Y : E \to G$ random variables. Then the image measure of μ_X via Y is the image measure of μ via $Y \circ X$.*

Proof. This is instantly seen with a simple drawing. More formally, we have

$$\begin{aligned}\mu_X \circ Y^{-1}(A) &= \mu_X(Y^{-1}(A)) = \mu_X(\{e \in E : Y(e) \in A\}) = \mu(X^{-1}(\{e \in E : Y(e) \in A\})) \\ &= \mu(\{\omega \in \Omega : X(\omega) \in E \text{ such that } Y(X(\omega)) \in A\}) = \mu((Y \circ X)^{-1}(A)) = \mu_{Y\circ X}(A)\end{aligned}$$

as required. □

Deep Dive

Let us comment on some anomalies which may happen when you work with general spaces in relation to Example 2.22 above. Suppose X_1, X_2 are two random variables on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with values in measurable spaces $(E_1, \mathscr{E}_1)$ and $(E_2, \mathscr{E}_2)$ respectively. Then $X = (X_1, X_1)$ *is* a random variable on Ω with values in the product space $(E_1 \times E_2, \mathscr{E}_1 \times \mathscr{E}_2)$ (exercise). However, in general, we can not make sense of $\mathbb{P}(X_1 = X_2)$ as the diagonal does not need to be in the product σ-algebra and hence the set $\{\omega : X_1(\omega) = X_2(\omega)\}$ does not have to measurable.

Suppose now that E_1, E_2 are metrisable topological space endowed with their Borel σ-algebras. We can consider the product topology on $E_1 \times E_2$ and take the Borel σ-algebra it generates, denoted $\mathscr{B}(E_1 \times E_2)$. If further both E_1, E_2 are *separable* (i.e., have a countable dense subset) then $\mathscr{B}(E_1 \times E_2) = \mathscr{B}(E_1) \times \mathscr{B}(E_2)$ and everything works as in the real-valued case. Otherwise however, $\mathscr{B}(E_1 \times E_2)$ (which includes the diagonal)

may be strictly larger than $\mathscr{B}(E_1)\times\mathscr{B}(E_2)$ and the joint law of (X_1,X_2) on $(E_1\times E_2,\mathscr{B}(E_1\times E_2))$ may not exist. Note that our argument for $\mathscr{B}(\mathbb{R}^d)=\times_{i=1}^d\mathscr{B}(\mathbb{R})$ relied on the fact that an open subset of $\mathbb{R}^n$ is a countable union of open hypercubes which uses separability of $\mathbb{R}$.

2.5 Product measure

We saw above how to define new measures via restrictions, summation and images. We now come to taking products of measure. Recall the product space and the product σ-algebra from Definition 1.7.

Theorem 2.24. *Let $(\Omega_i,\mathscr{F}_i,\mathbb{P}_i)$, $i=1,\dots,N$, be probability measures. Then there exists a unique measure $\mathbb{P}$ on the product space $(\Omega,\mathscr{F})=(\prod_{i=1}^N\Omega_i,\times_{i=1}^N\mathscr{F}_i)$ such that*

$$\mathbb{P}(E_1\times\ldots\times E_N)=\mathbb{P}_1(E_1)\cdot\ldots\cdot\mathbb{P}_N(E_N),\quad E_i\in\mathscr{F}_i,1\leqslant i\leqslant N. \tag{11}$$

$\mathbb{P}$ *is called the* product measure *and is also denoted $\bigotimes_{i\leqslant N}\mathbb{P}_i$ or $\mathbb{P}_1\otimes\ldots\otimes\mathbb{P}_N$.*

Proof. We show the statement for $N=2$. The general case then follows by induction since a general N product can be see as product of two spaces: Ω_1 and $\Omega_2\times\ldots\times\Omega_N$.

Suppose $N=2$. A set in $\mathscr{F}$ of the form $A\times B$ for $A\in\mathscr{F}_1,B\in\mathscr{F}_2$ is called a measurable rectangle. These sets form a π-system which, by Definition 1.7, generates $\mathscr{F}$. Let $\mathscr{A}$ denote the collection of finite unions of mutually disjoint measurable rectangles. Then $\mathscr{A}$ is an algebra and we can define a set function $\mathbb{P}$ on $\mathscr{A}$ by

$$\mathbb{P}(A_1\times B_1\cup\ldots\cup A_n\times B_n):=\sum_{i=1}^n\mathbb{P}_1(A_i)\mathbb{P}_2(B_i),\quad A_i\in\mathscr{F}_1,B_i\in\mathscr{F}_2,\quad A_i\times B_i\cap A_j\times B_j=\emptyset,\quad 1\leqslant i,j\leqslant n,i\neq j,$$

for $n\geqslant 1$. Clearly $\mathbb{P}(\emptyset)=0$ and, by Theorem 2.11, it remains to check that $\mathbb{P}$ is countably additive on $\mathscr{A}$. Let $(D_n)_{n\geqslant 1}$ be a sequence of sets in $\mathscr{A}$ with $D_n\downarrow\emptyset$. By Lemma 2.4, it suffices to show that $\lim_{n\to\infty}\mathbb{P}(D_n)=0$.

Each D_n is a finite union of measurable rectangles $A_{n,k}\times B_{n,k}$, $1\leqslant k\leqslant m_n$. If $A_{n,i}\cap A_{n,j}\neq\emptyset$, we may replace these two rectangles by three other rectangles with disjoint first sets, so that with no loss of generality we assume $A_{n\cdot}$ are mutually disjoint. For $\omega_1\in\Omega_1$, let $D_n(\omega_1)=\{\omega_2\in\Omega_2:(\omega_1,\omega_2)\in D_n\}$ so that $D_n(\omega_1)=B_{n,k}$ if $\omega_1\in A_{n,k}$, for some (and hence only one) $1\leqslant k\leqslant m_n$ and $D_n(\omega_2)=\emptyset$ otherwise. In particular, $D_n(\omega_1)\in\mathscr{F}_2$ (this also follows more generally, see Exercise 1.13). Properties of $(D_n)_{n\geqslant 1}$ imply that $D_n(\omega_1)\downarrow\emptyset$ for all $\omega_1\in\Omega_1$. Since $\mathbb{P}_2$ is a probability measure, it follows that if we define a sequence of functions on Ω_1 by $X_n(\omega_1)=\mathbb{P}_2(D_n(\omega_1))$, $n\geqslant 1$, then $X_n\downarrow 0$ pointwise on Ω_1. Note also that X_n is a simple function, constant on any of the sets $A_{n,k}$, and zero otherwise. In particular, for $\varepsilon>0$,

$$X_n^{-1}((\varepsilon,\infty))=\{\omega_1:X_n(\omega_1)>\varepsilon\}=\bigcup_{k\in I_n}A_{n,i},$$

for some subset $I_n\subseteq\{1,\dots,m_n\}$. Again, by properties of $(D_n)_{n\geqslant 1}$, we have $X_n^{-1}((\varepsilon,\infty))\downarrow\emptyset$ and hence the $\mathbb{P}_1$-probability of these sets decreases to zero. This yields

$$\mathbb{P}(D_n)=\sum_{k=1}^{m_n}\mathbb{P}_1(A_{n,k})\mathbb{P}_2(B_{n,k})\leqslant\mathbb{P}_1(X_n>\varepsilon)\mathbb{P}_2(\Omega_2)+\varepsilon\mathbb{P}_1(\Omega_1),$$

where we kept $\mathbb{P}_i(\Omega_i)=1$ terms to make it clear how the inequalities were obtained. Taking limit as $n\to\infty$, gives $\lim_{n\to\infty}\mathbb{P}(D_n)\leqslant\varepsilon$ for any $\varepsilon>0$ and hence $\lim_{n\to\infty}\mathbb{P}(D_n)=0$ as required. □

Remark. Clearly, we could take any finite measures and not only probability measures in the statement of the theorem. Further, through the usual arguments of restricting to subsets, the result also extends to σ-finite measures.

Remark. Note that the marginals, in the sense of Example 2.21, of the product measures $\mathbb{P}$ are given by $\mathbb{P}_i$ and that $\mathbb{P}$ is uniquely specified by its marginals via (11). This is a special property of the product measure and is not true for a general measure μ on the product space, as discussed in Examples 2.21 and 2.22.

3 Independence

There are two notions which really set probability apart and alive: *independence* and *conditional expectation*. Both relate to (degrees of) co-dependence and ways to measure it. We saw a baby example of conditional expectation, namely $\mathbb{P}(A|\sigma(B))(\cdot)$, in §2.2 above. To develop it properly, we will need the theory of integration which is still ahead of us. However, we already have all the tools to talk about independence.

3.1 Definitions and characterisations

Independence, or dependence, is all about information. A given piece of information is relevant if it potentially changes the way we see things. If we do not care about it, then we would say this information is independent of what we have in mind. As σ-algebras describe the information content for us, the notion of independence is best phrased in terms of them.

Definition 3.1. Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space and $(\mathscr{G}_i)_{i\leqslant n}$ a finite collection of σ-algebras, $\mathscr{G}_i\subseteq\mathscr{F}$ for $i\leqslant n$. We say that the σ-algebras $(\mathscr{G}_i)_{i\leqslant n}$ are *independent* if and only if

$$\mathbb{P}(A_1\cap\ldots\cap A_n)=\mathbb{P}(A_1)\cdot\ldots\cdot\mathbb{P}(A_n),\quad \text{for any } A_i\in\mathscr{G}_i, i\leqslant n. \tag{12}$$

For an arbitrary collection $(\mathscr{G}_i)_{i\in I}$ of sub-σ-algebras of $\mathscr{F}$, we say that these σ-algebras are independent if any finite sub-collection of them is.

Example 3.2. The trivial σ-algebra $\{\Omega,\emptyset\}$ is independent of any other σ-algebra. Its information content is null.

Exercise 3.3. Let $(\mathscr{G}_n)_{n\geqslant 1}$ be a sequence of independent σ-algebras. Use continuity of measure from above to show that for any $A_n\in\mathscr{G}_n$, $n\geqslant 1$,

$$\mathbb{P}\left(\bigcap_{n\geqslant 1}A_n\right)=\prod_{n\geqslant 1}\mathbb{P}(A_n).$$

Lemma 3.4. *Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space and $A_1,\ldots,A_n$ some events in $\mathscr{F}$. Then, their generated σ-algebras are independent if and only if*

$$\mathbb{P}\left(\bigcap_{i\in J}A_i\right)=\prod_{i\in J}\mathbb{P}(A_i),\quad \textit{for any } J\subseteq\{1,\ldots,n\}.$$

Proof. We just show the statement for $n=2$. One direction is obvious. For the other recall that $\sigma(A)=\{\Omega,\emptyset,A,A^{\mathrm{c}}\}$ and note that if $\mathbb{P}(A\cap B)=\mathbb{P}(A)\mathbb{P}(B)$ then

$$\mathbb{P}(A\cap B^{\mathrm{c}})=\mathbb{P}(A)-\mathbb{P}(A\cap B)=\mathbb{P}(A)(1-\mathbb{P}(B))=\mathbb{P}(A)\mathbb{P}(B^{\mathrm{c}})$$

and the result follows by symmetry. □

The above simple result also follows from the following much more general one: one does not need to verify (12) for all sets in the σ-algebras but it is enough to verify it for sets in generating π-systems.

Theorem 3.5. *Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space, $(\mathscr{G}_i)_{i\in I}$ an arbitrary collection of σ-algebras, each generated by a π-system $\mathscr{A}_i\subseteq\mathscr{F}$: $\mathscr{G}_i=\sigma(\mathscr{A}_i)$, $i\in I$.Then $(\mathscr{G}_i)_{i\in I}$ are independent if and only if*

$$\mathbb{P}\left(\bigcap_{i\in J}A_i\right)=\prod_{i\in J}\mathbb{P}(A_i)\quad \textit{for any } A_i\in\mathscr{A}_i, i\in J, \textit{ for any finite subset } J\subseteq I. \tag{13}$$

Proof. If $(\mathscr{G}_i)_{i\in I}$ are independent then, by definition, (13) holds. The reverse implication is a simple application of Lemma 1.12 but we give the details nevertheless. Fix a finite subset $J \subset I$ and number its elements $J = \{i_1,\dots,i_n\}$. Let $\mathscr{M}_1$ be the set of $A \in \mathscr{F}$ for which

$$\mathbb{P}(A\cap A_2\cap\ldots\cap A_n) = \mathbb{P}(A)\cdot\mathbb{P}(A_2)\cdot\ldots\cdot\mathbb{P}(A_n) \quad \text{for any } A_l \in \mathscr{A}_{i_l},\ l=2,\dots,n.$$

By assumption, $\mathscr{A}_{i_1} \subseteq \mathscr{M}_1$ and also $\Omega \in \mathscr{M}_1$ by the assumption applied to $J_1 = J\setminus\{i_1\}$. For $A \subseteq B$ both in $\mathscr{M}_1$, we have

$$\begin{aligned}\mathbb{P}((B\setminus A)\cap A_2\cap\ldots\cap A_n) &= \mathbb{P}(B\cap A_2\cap\ldots\cap A_n) - \mathbb{P}(A\cap A_2\cap\ldots\cap A_n)\\ &= (\mathbb{P}(B)-\mathbb{P}(A))\mathbb{P}(A_2)\ldots\mathbb{P}(A_n) = \mathbb{P}(B\setminus A)\mathbb{P}(A_2)\ldots\mathbb{P}(A_n)\end{aligned}$$

so that $B\setminus A \in \mathscr{M}_1$. Finally, for an increasing sequence $B_k \in \mathscr{M}_1$, $B_k \uparrow B$, continuity from below of $\mathbb{P}$, see Proposition 2.3, implies that $B \in \mathscr{M}_1$. We conclude that $\mathscr{M}_1$ is a λ-system and hence, by the π-λ systems Lemma (Lemma 1.12), $\mathscr{G}_{i_1} = \sigma(\mathscr{A}_{i_1}) \subseteq \mathscr{M}_1$. We then proceed by induction. We let $\mathscr{M}_k$ be the $A \in \mathscr{F}$ for which

$$\mathbb{P}(A_1\cap\ldots A_{k-1}\cap A\cap A_{k+1}\ldots\cap A_n) = \mathbb{P}(A_1)\cdot\ldots\cdot\mathbb{P}(A_{k-1})\cdot\mathbb{P}(A)\cdot\mathbb{P}(A_{k+1})\cdot\ldots\cdot\mathbb{P}(A_n),$$

for any $A_l \in \mathscr{G}_{i_l}$, $1 \leqslant l < k$ and $A_l \in \mathscr{A}_{i_l}$, $k < l \leqslant n$. Then, by induction step, $\mathscr{A}_{i_k} \subseteq \mathscr{M}_k$ and, as above, π-λ systems lemma gives $\mathscr{G}_{i_k} \subseteq \mathscr{M}_k$. □

Definition 3.6. Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space and $(X_i)_{i\in I}$ a family of random variables with values in some measurable spaces $(E_i,\mathscr{E}_i)_{i\in I}$. We say that these random variables are *independent* if their generated σ-algebras $(\sigma(X_i))_{i\in I}$ are.

It follows by the definition that $(X_i)_{i\in I}$ are independent if and only if for any finite subset $J \subseteq I$

$$\mathbb{P}(X_i \in A_i \text{ for all } i\in J) = \prod_{i\in J}\mathbb{P}(X_i\in A_i), \quad \text{for any } A_i\in\mathscr{E}_i, i\in J.$$

This can be further rephrased using the nomenclature of product measures.

Theorem 3.7. *Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space and $(X_i)_{i\leqslant n}$ a finite family of random variables with values in some measurable spaces $(E_i,\mathscr{E}_i)_{i\leqslant n}$. These random variables are* independent *in and only if their joint distribution $\mu_{(X_1,\dots,X_n)}$ on the product space $(\prod_{i\leqslant n}E_i, \times_{i\leqslant n}\mathscr{E}_i)$ is the product measure of the marginal distributions μ_{X_i}.*

The above statement extends to an arbitrary family of random variables as independence is defined by considering finite subsets of variables. Note that this theorem generalises the results 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. 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 $\mu_{X_i} = \mu_i$. In particular, the notion of independence is non-vacuous.

Checking independence of random variables from Definition 3.6 or Theorem 3.7 might be difficult. However, when combined with Theorem 3.5, it becomes more manageable! We have the following immediate corollary.

Corollary 3.8. *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 $\overline{\mathbb{R}}$),*

$$\mathbb{P}(X_1\leqslant x_1,\dots,X_n\leqslant x_n) = \mathbb{P}(X_1\leqslant x_1)\ldots\mathbb{P}(X_n\leqslant x_n).$$

Example 3.9. Recall our coin tossing representation in Example 1.22, namely on $([0,1],\mathscr{B}([0,1]))$ we let $X_n(\omega) = \mathbf{1}_{\lfloor 2^n\omega\rfloor \text{ is even}}$, $n \geqslant 1$, where 0 is even. We can now check that $(X_n)_{n\geqslant 1}$ are independent (exercise!). This shows that we built a good model, as different coin tosses ought to be independent, and also that the notion of independence is interesting (and non-vacuous as already observed).

As independence is about information, the following Proposition is obvious from Definition 3.6 and since if $Y = f(X)$ then $\sigma(Y) \subseteq \sigma(X)$.

Proposition 3.10. *Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space and $(X_i)_{i\in I}$ a family of independent random variables with values in some measurable spaces $(E_i,\mathscr{E}_i)_{i\in I}$ and $f_i : E_i \to \mathbb{R}$ be measurable, $i \in I$. Then $(Y_i := f_i(X_i))_{i\in I}$ are independent random variables.*

Deep Dive

Theorem 2.24 extends to countable products and (11) then reads

$$\mathbb{P}\left(E_1 \times \ldots \times E_N \times \prod_{n>N} \Omega_n\right) = \mathbb{P}_1(E_1)\cdot\ldots\cdot\mathbb{P}_N(E_N), \quad \forall N \geqslant 1 \text{ and } E_i \in \mathscr{F}_i, 1 \leqslant i \leqslant N.$$

This is important as it offers a canonical way to build a sequence of independent random variables with given distributions. Indeed, consider $(\Omega_i,\mathscr{F}_i,\mathbb{P}_i) = ([0,1],\mathscr{B}([0,1]),\texttt{Leb})$. On the product space define $X_i(\omega) = \omega_i$, where $\Omega \ni \omega = (\omega_i)_{i\geqslant 1}$. Then $(X_i)_{i\geqslant 1}$ is a sequence of independent identically distributed random variables on the product probability space, each X_i is uniform on $[0,1]$. Given any sequence $(\mu_i)_{i\geqslant 1}$ of probability measures on $\mathbb{R}$ we let $(F_i)_{i\geqslant 1}$ be their respective distribution functions and, as in Example 2.20, we set $Y_i = F_i^{-1}(X_i)$. Then each $Y_i \sim \mu_i$ and, by Proposition 3.10, all $(Y_i)_{i\geqslant 1}$ are independent.

3.2 Kolmogorov's 0-1 Law

We have now the tools to present a beautiful classical result in probability theory concerning 'tail events' associated to sequences of independent random variables.

Definition 3.11 (Tail σ-algebra)**.** For a sequence of random variables $(X_n)_{n\geqslant 1}$ define

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

and

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

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

Exercise 3.12. Check that $\mathscr{T}$ is a σ-algebra.

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 may sound impossible at first, but in fact 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 3.13 (Kolmogorov's 0-1 Law)**.** *Let $(X_n)_{n\geqslant 1}$ be a sequence of independent random variables. Then the tail σ-algebra $\mathscr{T}$ of $(X_n)_{n\geqslant 1}$ contains only events of probability 0 or 1. Moreover, any $\mathscr{T}$-measurable random variable is almost surely constant.*

Proof. Fix $n \geqslant 1$ and let $\mathscr{F}_n = \sigma(X_1, \ldots, X_n)$. Note that $\mathscr{F}_n$ is generated by the π-system of events

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

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

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

For any $A \in \mathscr{A}$, $B \in \mathscr{B}$, 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 3.5 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.

The above was true for all $n \geqslant 1$ and hence $\bigcup_{n\geqslant 1} \mathscr{F}_n$ and $\mathscr{T}$ are also 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 3.5 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 $\{0,1\}$ since $\{Y \leqslant y\} \in \mathscr{T}$. So $\mathbb{P}(Y = c) = 1$ where $c = \inf\{y : F_Y(y) = 1\}$. This extends easily to the extended-real-valued case. □

Example 3.14. 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 $U = \limsup_{n\to\infty} S_n/n$ and $L = \liminf_{n\to\infty} S_n/n$. Then U and L are tail random variables and so almost surely constant. We'll prove later in the course that, $L = U$ is the expectation of X_1, a result known as the Strong Law of Large Numbers.

3.3 The Borel–Cantelli Lemmas

We turn now to second fundamental set of results which assert that certain events have probability one or zero. We work on a fixed probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

Definition 3.15. Let $(A_n)_{n\geqslant 1}$ be a sequence of sets from $\mathscr{F}$. We define

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

and

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

Lemma 3.16.

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

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. □

Lemma 3.17 (Fatou and Reverse Fatou for sets)**.** *Let $(A_n)_{n\geqslant 1}$ be a sequence of sets from $\mathscr{F}$. Then*

$$\mathbb{P}(\liminf_{n\to\infty} A_n) \leqslant \liminf_{n\to\infty} \mathbb{P}(A_n) \quad \textit{and} \quad \mathbb{P}(\limsup_{n\to\infty} A_n) \geqslant \limsup_{n\to\infty} \mathbb{P}(A_n).$$

Proof. Using continuity of $\mathbb{P}$ from above and below, see Proposition 2.3, we have

$$\mathbb{P}(A_n \text{ eventually}) = \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{m\geqslant n} A_m\right) \leqslant \lim_{n\to\infty} \inf_{m\geqslant n} \mathbb{P}(A_m) = \liminf_{n\to\infty} \mathbb{P}(A_n)$$

and hence (taking complements)

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

□

In fact we can say more about the probabilities of these events.

Lemma 3.18 (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 which we will use time and again.

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 Proposition 2.3, $\mathbb{P}(G_n) \downarrow \mathbb{P}(G)$.

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. □

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 3.19 (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 $\{A_n^c \text{ eventually}\}$.

$$\begin{aligned} \mathbb{P}\left[\bigcap_{m\geqslant n} A_m^c\right] &= \prod_{m\geqslant n} (1 - a_m) \quad \text{(by independence, recall Exercise 3.3)} \\ &\leqslant \exp\left(-\sum_{m\geqslant n} a_m\right) = 0. \end{aligned}$$

Hence

$$\mathbb{P}\left(A_n^{\mathrm{c}} \text{ eventually}\right) = \mathbb{P}\left(\bigcup_{n\geqslant 1}\bigcap_{m\geqslant n} A_m^{\mathrm{c}}\right) = \lim_{n\to\infty}\mathbb{P}\left(\bigcap_{m\geqslant n} A_m^{\mathrm{c}}\right) = 0,$$

and

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

□

Exercise 3.20. 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.

Exercise 3.21. Let $(X_n)_{n\geqslant 1}$ be independent exponentially distributed random variables with parameter 1 and let $M_n = \max\{X_1,\ldots,X_n\}$. Then

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

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

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

Fix $0<\varepsilon<1$. Then

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

Thus

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

and so by BC1

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

Since ε was arbitrary, taking a suitable countable union gives

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

The reverse bound is similar: use BC1 to show that

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

□

At first sight, it might look 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.

4 Integration

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.

4.1 Definition and first properties

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:

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

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

Definition 4.1. If f is a non-negative simple function with canonical form (6), then we define the integral of f with respect to μ as

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

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

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

Note that the supremum may be equal to $+\infty$. Recall from Lemma 1.26 that measurability of f is equivalent with f being an increasing limit of simple function. The above definition and this notion of integral can not be extended to non-measurable functions in any meaningful way. Indeed, we know well by now that we can not measure - that is integrate the indicator function - some non-measurable sets! We recall also that one can use a canonical construction to approximate f, see the proof of Lemma 1.26, and the above supremum may be replaced with a limit along such an approximating sequence of simple functions – this is easy to check directly (exercise!) but it will also follow from the more general Theorem 4.6.

One obvious consequence of the above definition is worth pointing out: if $0 \leqslant f \leqslant g$ are two measurable functions then $\int f \,\mathrm{d}\mu \leqslant \int g \,\mathrm{d}\mu$. We sometimes refer to this as the comparison test or comparison principle.

Definition 4.3. We say that a function f on $(\Omega, \mathscr{F}, \mu)$ is *integrable*, and write $f \in \mathscr{L}^1(\Omega, \mathscr{F}, \mu)$, if f is measurable and $\int |f| \,\mathrm{d}\mu < \infty$. 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.

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. Clearly, by comparison, if f is measurable and $|f| \leqslant g$ for some

$g \in \mathscr{L}^1(\Omega, \mathscr{F}, \mu)$ then $f \in \mathscr{L}^1(\Omega, \mathscr{F}, \mu)$. Note that $f = f^+ - f^-$ and $|f| = f^+ + f^-$ so that another important consequence of the above definition is the familiar inequality:

$$\left| \int f \, \mathrm{d}\mu \right| \leqslant \int |f| \, \mathrm{d}\mu. \tag{14}$$

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.

Example 4.4. 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.

Example 4.5. Suppose that μ is a discrete measure with mass p_i at point $x_i \in \mathbb{R}$, 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. This example is very different in nature to the Lebesgue integral above – here integrals are just sums. It is rather pleasing to see that the toolbox we developed covers both cases with a unified language.

Our construction of the integral followed the steps seen in Part A Integration course. Importantly for us, our generalised integral still has all the good properties.

Theorem 4.6 (Monotone Convergence Theorem (MCT))**.** *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, it is easy to see that it is enough to suppose that $f_n \uparrow f$ μ-almost everywhere. An equivalent formulation of the Monotone Convergence Theorem (MCT) considers partial sums: if (f_n) is a sequence of non-negative measurable functions, then

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

Proof. Note that the MCT for $f_n = \mathbf{1}_{A_n}$ is simply the continuity of μ from below, Proposition 2.3 (iv). The general case is deduced from this, see Part A Integration. □

The MCT is a key result from which the rest of the integration theory essentially follows using the 'bare hands method' outlined in the comments following Lemma 1.26: start by considering indicator functions $f = \mathbf{1}_E$, then simple functions f, then non-negative measurable f via Lemma 1.26 and the MCT, and finally general measurable f via $f = f^+ - f^-$. For this reason, MCT is stated here and not in the subsequent section, even if it would also fit there by the virtue of its name.

Exercise 4.7. As a simple warmup exercise, show that 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

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

Exercise 4.8. Use MCT to prove Lemma 3.18.

Solution. Consider $N_n := \sum_{k=1}^n \mathbf{1}_{A_k}$, the (random) number of events A_k that hold for $k \leqslant n$. Then $\int N_n \mathrm{d}\mathbb{P} = \sum_{k=1}^n \mathbb{P}(A_k)$. Since $N_n \uparrow N = N^\infty$, by MCT, we have $\int N \mathrm{d}\mathbb{P} = \sum_{k\geqslant 1} \mathbb{P}[A_k] < \infty$. But $\int N \mathrm{d}\mathbb{P} < \infty$ implies $\mathbb{P}(N = \infty) = 0$, as required. □

4.2 Radon-Nikodym Theorem

The just defined integral offers a canonical way to construct new measures on a given measure space. This was first presented below Theorem 2.16 but can now be made rigorous.

Suppose that $(\Omega, \mathscr{F}, \mu)$ is a measure space and f a positive integrable function. Then

$$\mathscr{F} \ni A \longrightarrow \nu(A) := \int_A f \mathrm{d}\mu = \int f(\omega)\mathbf{1}_A(\omega)\mu(\mathrm{d}\omega)$$

defines a measure. This is easy to verify for a simple function f and follows in general by the MCT (exercise). Note that by definition if A is μ-null then it is also ν-null. We recall the terminology and notation of Definition 2.7 and write $\nu \ll \mu$.

A particularly important special case is when $\int f \mathrm{d}\mu = 1$ so that ν is a probability measure. This is well known to you under the heading of continuous random variables for Prelims or Part A probability. Take $(\Omega, \mathscr{F}, \mu) = (\mathbb{R}, \mathscr{B}(\mathbb{R}), \texttt{Leb})$ and let $F(x) = \int_{-\infty}^x f(y)\mathrm{d}y$. Then $\nu((-\infty, x]) = F(x)$ so that, by Theorem 2.10, $\nu = \mu_F$ is the Lebesgue-Stieltjes measure associated to F by Theorem 2.16. The function $F(x)$ is the *distribution function* corresponding to the probability measure ν.

The following fundamental result tells us that the above construction describes *all* measures ν absolutely continuous w.r.t. μ, $\nu \ll \mu$. We state it for probability measures. An extension to finite measures is immediate and extension to σ-finite measures follows via the usual steps.

Theorem 4.9 (Radon-Nikodym Theorem)**.** *Let μ, ν be two probability measures on a measurable space $(\Omega, \mathscr{F})$. Then $\nu \ll \mu$ if and only if there exists a non-negative random variable f such that*

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

The function f is often denoted $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ and is called the Radon-Nikodym derivative *of ν w.r.t. μ.*
Further, $\nu \sim \mu$ if and only if $f > 0$ μ-a.s. (and then also ν-a.s.) in which case $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} = \frac{1}{f}$.

Exercise 4.10. Recall discrete measure theory on a countable Ω as presented in Example 2.9. Prove Theorem 4.9 in this setting.

Deep Dive

The general proof of the Radon-Nikodym theorem is no joking matter. We *will* prove this result but only much later in the course once we have established a good understanding of martingale convergence. The Radon-Nikodym Theorem is often used to show existence of the conditional expectation so that the whole enterprise may then appear circular. Here, we follow a different path and do *not* use Theorem 4.9 to establish the existence of conditional expectations so there is no appearance of circularity. However, one could also abstain from showing *existence* of the conditional expectation. Instead, one could use its defining properties to define when a family of random variables is a martingale and carry out the whole enterprise this way. Culminate with proving Theorem 4.9 and go back to existence of the basic objects on their own. A motivated

reader is invited to follow through the different logical pathwise to a complete theory.

4.3 Convergence Theorems

The following theorems were proved in Part A for the Lebesgue integral. The proofs essentially rely on the MCT and carry over to the more general integral defined here. We start with the functional versions of Lemma 3.17.

Theorem 4.11 (Fatou's Lemma)**.** *Let (f_n) be a sequence of non-negative measurable functions on $(\Omega, \mathscr{F}, \mu)$. Then*

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

Proof. We write $\liminf f_n$ for $\liminf_{n\to\infty} f_n$. Recall that

$$\liminf f_n = \lim_{k\to\infty} g_k, \quad g_k = \inf_{n\geqslant k} f_n.$$

In particular, for $n \geqslant k$, $f_n \geqslant g_k$ and hence also $\int f_n \, \mathrm{d}\mu \geqslant \int g_k \, \mathrm{d}\mu$. As this holds for all $n \geqslant k$, we have

$$\int g_k \, \mathrm{d}\mu \leqslant \inf_{n\geqslant k} \int f_n \, \mathrm{d}\mu.$$

Since $g_k \uparrow \liminf f_n$, as $k \to \infty$, we apply MCT to obtain the desired inequality:

$$\int \liminf f_n \, \mathrm{d}\mu = \lim_{k\to\infty} \int g_k \, \mathrm{d}\mu \leqslant \liminf_{k\to\infty} \inf_{n\geqslant k} \int f_n \, \mathrm{d}\mu = \liminf_{n\to\infty} \int f_n \, \mathrm{d}\mu.$$

□

Lemma 4.12 (Reverse Fatou's Lemma)**.** *Let (f_n) be a sequence of non-negative measurable functions on $(\Omega, \mathscr{F}, \mu)$. Assume that there exists a function $g \in \mathscr{L}^1(\Omega, \mathscr{F}, \mu)$ 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.) □

The above lemmas gave us inequalities between limits of integrals and the integral of the limit. In most cases however, we are interested in having an equality. This is the subject of the following results. They are all well known and very useful. At the same time however, from a probabilistic point of view, they are not fully satisfactory. We will develop in §5.4 below a finer tool to deal with the issue of convergence of integrals, namely the notion of uniform integrability.

We 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 4.13 (Dominated Convergence Theorem (DCT))**.** *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*

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

Proof. Taking limits we have $0 \leqslant |f| \leqslant g$ so that $f \in \mathscr{L}^1(\Omega, \mathscr{F}, \mu)$ by comparison. Using (14) and applying Lemma 4.12 to $h_n = |f_n - f| \leqslant 2g$, we obtain

$$0 \leqslant \limsup_{n\to\infty} \left| \int f_n \, \mathrm{d}\mu - \int f \, \mathrm{d}\mu \right| \leqslant \limsup_{n\to\infty} \int |f_n - f| \, \mathrm{d}\mu \leqslant \int \limsup_{n\to\infty} |f_n - f| \, \mathrm{d}\mu = \int 0 \, \mathrm{d}\mu = 0.$$

□

Lemma 4.14 (Scheffé). *Suppose that $f_n, f \in \mathscr{L}^1(\Omega, \mathscr{F}, \mu)$ convere pointwise, $f_n \to f$ as $n \to \infty$. Then*

$$\int |f_n - f| \, \mathrm{d}\mu \to 0 \quad \Longleftrightarrow \quad \int |f_n| \, \mathrm{d}\mu \to \int |f| \, \mathrm{d}\mu.$$

Deep Dive

Proof. The "$\Longrightarrow$" implication is trivial since $-|f_n - f| \leqslant |f_n| - |f| \leqslant |f_n - f|$ so we show the reverse. Suppose first that f_n, f are positive and $\int f_n \, \mathrm{d}\mu \to \int f \, \mathrm{d}\mu$. Since $(f_n - f)^- \leqslant f$, DCT gives $\int (f_n - f)^- \, \mathrm{d}\mu \to 0$. For the positive part, we have

$$\int (f_n - f)^+ \, \mathrm{d}\mu = \int_{f_n \geqslant f} (f_n - f) \, \mathrm{d}\mu = \int f_n \, \mathrm{d}\mu - \int f \, \mathrm{d}\mu - \int_{f_n < f} (f_n - f) \, \mathrm{d}\mu.$$

The first term converges to the second by assumption and the last one coverges to zero by the previous argument. Together, we obtain the desired convergence $\int |f_n - f| \, \mathrm{d}\mu \to 0$.

In the general case, we have $\int f^{\pm} \, \mathrm{d}\mu \leqslant \liminf \int f_n^{\pm} \, \mathrm{d}\mu$ by Fatou. By assumption,

$$\int (f^+ + f^-) \, \mathrm{d}\mu = \lim \int (f_n^+ + f_n^-) \, \mathrm{d}\mu$$

so that necessarily the sequences $f_n^+ \to f^+$ and $f_n^- \to f^-$ satisfy the assumption of the Lemma and are positive so the proof above applies and we conclude using $|f_n - f| \leqslant |f_n^+ - f^+| + |f_n^- - f^-|$. □

4.4 Expectation

The notion of image measure developed in §2.4 allows us to see the integral of a function against a measure on one space simply as the integral against the image measure on the image space. We phrase this as a theorem since it is a key result for a lot of computations one has to do.

Theorem 4.15. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability measure, X a random variable with values in a measurable space $(E, \mathscr{E})$ and g a real-valued random variable on $(E, \mathscr{E})$. Let $\mathbb{Q} = \mathbb{P} \circ X^{-1}$ be the image of $\mathbb{P}$ via X. Then g is $\mathbb{Q}$-integrable if and only if $g \circ X$ is $\mathbb{P}$-integrable and then*

$$\int_E g(x) \mathbb{Q}(\mathrm{d}x) = \int_\Omega g(X(\omega)) \mathbb{P}(\mathrm{d}\omega). \tag{15}$$

Proof. (15) holds by definition for $g = \mathbf{1}_A$ an indicator of an event $A \in \mathscr{E}$. By linearity it then holds for any simple function g. For a measurable $g \geqslant 0$, let $g_n \uparrow g$ be a sequence of simple functions increasing to g, say $g_n = \sum_{k \leqslant m_n} a_k \mathbf{1}_{A_k}$ and note that

$$g_n(X(\omega)) = \sum_{k \leqslant m_n} a_k \mathbf{1}_{X(\omega) \in A_k} = \sum_{k \leqslant m_n} a_k \mathbf{1}_{X^{-1}(A_k)}(\omega)$$

are simple functions on Ω, $g_n \circ X \uparrow g \circ X$. MCT then gives the required equality for g with one integral being finite if and only if the other is. The general case follows with $g = g^+ - g^-$ and in particular g is $\mathbb{Q}$-integrable if and only if $g \circ X$ is $\mathbb{P}$-integrable. □

In the reminder of this section, X denotes a random variable defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. We often refer to the integral on Ω with respect to $\mathbb{P}$ as the *expectation*.

Definition 4.16 (Expectation)**.** We say that X admits a first moment, if X is *integrable*, i.e., $X \in \mathscr{L}^1(\Omega, \mathscr{F}, \mathbb{P})$ or

$$\mathbb{E}[|X|] = \int_\Omega |X(\omega)| \mathbb{P}(\mathrm{d}\omega) < \infty.$$

The *expectation* of a random variable X defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is

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

Note that this is well defined and finite if $\mathbb{E}[|X|] < \infty$ but otherwise may be either $+\infty$ or undefined.

Recall that $\mu_X = \mathbb{P} \circ X^{-1}$ denotes the distribution of X. A simple application of Theorem 4.15, with $g(x) = x$, gives

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

In other words, the expectation of X is simply the barycentre of its distribution. As one expects from the barycentre, it is the *optimal* prediction of X using a constant as the following makes precise.

Exercise 4.17. For $X \in \mathscr{L}^2(\Omega, \mathscr{F}, \mathbb{P})$ show that

$$\inf_{c \in \mathbb{R}} \mathbb{E}[(X-c)^2]$$

is attained by $c = \mathbb{E}[X]$. We say that $\mathbb{E}[X]$ is the best constant mean square approximation of X.

Clearly, $\mathbb{E}[X]$ is a property of the distribution of X in the sense that two random variables X, Y, possibly defined on different probability spaces, with $X \sim Y$, have the same expectation. More generally, we have $\mathbb{E}[g(X)] = \int g(x) \mu_X(\mathrm{d}x)$ which is thus determined by μ_X alone, which in turn is determined by its values on a π-system: $\mu_X((-\infty, x]) = \mu(X \leqslant x)$, $x \in \mathbb{R}$. Very often in applications we suppress the sample space Ω and work directly with μ_X.

Definition 4.18 (Variance)**.** Suppose X admits a second moment, i.e., $\mathbb{E}[X^2] < \infty$. Then, the *variance* of X is given by

$$\mathrm{Var}(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

and is also called the *the second centred moment*. The square root of the variance, $\sqrt{\mathrm{Var}(X)}$, is called the *standard deviation* of X.

Note that if we put

$$Y = \frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}}$$

then Y is a random variable with $\mathbb{E}[Y] = 0$ and $\mathrm{Var}(Y) = \mathbb{E}[Y^2] = 1$. We say that Y is the *standardised* version of X: its distribution is that of X but shifted and rescaled to have the first two moments equal to 0 and 1.

Definition 4.19. The n^{th} *standardised moment* of X, if well defined, is given by

$$\mathbb{E}[Y^n] = \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}}\right)^n\right].$$

The third standardised moment is known as *skewness* of X and the fourth one as *kurtosis*.

Note that all the moments defined above are, by Theorem 4.15, determined by the distribution of X.

4.5 Integration on a product space

Recall the definition of product space, Definition 1.7, and the construction of the product measure in Theorem 2.24. The canonical example of a product measure is given by the Lebesgue measure on $\mathbb{R}^2$, or, more generally, 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

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

We now extend this to a general setting.

We fix two probability spaces $(\Omega_i, \mathscr{F}_i, \mathbb{P}_i)$, $i = 1,2$ and let $(\Omega, \mathscr{F}, \mathbb{P})$ denote their product space, i.e., $\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$. Recall from Lemma 1.29 that for a measurable f, the mappings with one coordinate fixed are also measurable (w.r.t. to the appropriate σ-algebra).

Theorem 4.20 (Fubini/Tonelli)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be the product of probability spaces $(\Omega_i, \mathscr{F}_i, \mathbb{P}_i)$, $i = 1,2$, and let $f = f(x,y)$ be a bounded measurable function on $(\Omega, \mathscr{F})$. The functions*

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

are $\mathscr{F}_1$- and $\mathscr{F}_2$-measurable respectively.

Suppose either (i) that f is $\mathbb{P}$-integrable on Ω or (ii) that $f \geqslant 0$. Then

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

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

Remark (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.

Deep Dive

You may recall from Part A Integration that statements about measurability of some functions, e.g., $x \to f(x,y)$, were for a.e. x and not for all x as here. This is because in Part A Integration you worked on the completed σ-algebra of all Lebesgue measurable sets and here we do not complete the σ-algebra by adding the null sets.

Proof. Both statements follow as immediate applications of the Monotone Class Theorem (Theorem 1.28) and we only outline the proof. First we check that the class $\mathscr{H}$ of bounded functions which satisfy the statements satisfies the assumptions in Theorem 1.28. Then we observe that $f = \mathbf{1}_{A_1 \times A_2} \in \mathscr{H}$ for all $A_1 \in \mathscr{F}_1, A_2 \in \mathscr{F}_2$. The statements then hold for all $\mathscr{F}$ measurable bounded functions, including simple functions. The general case follows via the MCT. □

Remark. Note that we used the fact that $\mathbb{P}_i$ are probability measures, or more generally finite measures, when applying the Monotone Class Theorem: we need the integrals of a constant to be bounded! The above arguments can then be extended, in the usual way, to σ-finite measures. But Fubini's theorem may fail for arbitrary measures!

Example 4.21. Let us consider an important example. Let X be a positive random variable on a generic probability space $(\Omega, \mathscr{F}, \mathbb{P})$. We consider the product space $([0,\infty)\times\Omega, \mathscr{B}([0,\infty))\times\mathscr{F}, \mathtt{Leb}\otimes\mathbb{P})$. Consider the area under the graph of $\omega \to X(\omega)$, namely

$$A := \{(x,\omega) : 0 \leqslant x \leqslant X(\omega)\}; \quad f = \mathbf{1}_A.$$

The partial integrals are given by

$$\int_\Omega f(x,\omega)\mathbb{P}(\mathrm{d}\omega) = \mathbb{P}(X \geqslant x) \quad \text{and} \quad \int_{[0,\infty)} f(x,\omega)\,\mathrm{d}x = X(\omega),$$

where $\mathrm{d}x$ denotes $\mathtt{Leb}(\mathrm{d}x)$ in the usual fashion. Fubini gives us

$$(\mathbb{P}\times\mathtt{Leb})(A) = \int_{[0,\infty)} \mathbb{P}(X \geqslant x)\,\mathrm{d}x = \mathbb{E}[X]. \tag{16}$$

Remark. Building on the above example, consider the cornerstone results for functions, e.g., MCT, Fatou's Lemma, and see that they simply correspond to the analogues for sets applied to 'areas under graph'.

Here is a simple corollary of Fubini's theorem which rephrases independence of random variables using expectations.

Corollary 4.22. *Let X,Y be random variables on some probability space $(\Omega,\mathscr{F},\mathbb{P})$. Then X and Y are independent if and only if for any positive measurable functions f,g*

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)].$$

Proof. For the "if" direction, take $f = \mathbf{1}_{(-\infty,r]}, g = \mathbf{1}_{(-\infty,s]}$, $r,s\in\mathbb{R}$, and use Corollary 3.8. For the "only if" direction, by Theorem 3.7, the joint distribution of (X,Y) is the product measure, $\mu_{(X,Y)} = \mu_X \otimes \mu_Y$. The result then follows from Fubini's theorem since, by Theorem 4.15, $\mathbb{E}[f(X)g(Y)] = \int_{\mathbb{R}^2} f(x,y)\mu_{(X,Y)}(\mathrm{d}(x,y))$. □

It is perhaps worth pausing and recalling that you saw the above in Prelims Probability for discrete random variables. It is pleasing to see how much more elegant our language and proofs have become since!

Deep Dive

The statement and applications of Fubini's theorem above pertained only to product measures on Ω. This is perhaps natural in analysis but much less in probability theory where we often consider measures on the product space which are not product measures, i.e., joint distribution of couples of random variables which are *not* independent. It is thus interesting to extend to this context.

Naturally, there are many other measures on Ω. Let us elaborate on other ways to construct such measures and how to integrate against them. We keep the setup akin to Example 4.21 but it is clear things could be written for any product of any two probability spaces.

Definition 4.23. A *probability kernel* on the product space $(\mathbb{R}\times\Omega, \mathscr{B}(\mathbb{R})\times\mathscr{F})$ is a family of probability measures $(\mathbb{P}_x)_{x\in\mathbb{R}}$ on $\mathscr{F}$ such that $\mathbb{R}\ni x\to\mathbb{P}_x(A)$ is measurable for any $A\in\mathscr{F}$.

In words, a probability kernel is a measurable function in one argument and a probability measure in the other. A very special case is given by $\mathbb{P}_x = \mathbb{P}$ is independent of x. This is the case when constructing product measures.

Theorem 4.24 (Generalised Fubini)**.** *Let $(\mathbb{P}_x)_{x\in\mathbb{R}}$ be a probability kernel on $(\mathbb{R}\times\Omega, \mathscr{B}(\mathbb{R})\times\mathscr{F})$ and let μ be a probability measure on $\mathbb{R}$. Then there exists a unique probability measure $\mathbb{Q}$ on $\mathscr{B}(\mathbb{R})\times\mathscr{F}$ such that*

$$\mathbb{Q}(E\times A) = \int_E \mathbb{P}_x(A)\mu(\mathrm{d}x), \quad E\in\mathscr{B}(\mathbb{R}), A\in\mathscr{F}. \tag{17}$$

For a positive mesurable function f on $\mathbb{R}\times\Omega$, the function $x\to\int_\Omega f(x,\omega)\mathbb{P}_x(\mathrm{d}\omega)$ is measurable and

$$\int_{\mathbb{R}\times\Omega} f\,\mathrm{d}\mathbb{Q} = \int_{\mathbb{R}} \mu(\mathrm{d}x)\int_\Omega f(x,\omega)\mathbb{P}_x(\mathrm{d}\omega).$$

The above equation remains true if f is assumed $\mathbb{Q}$-integrable on $\mathbb{R}\times\Omega$ and then the function $\omega\to f(x,\omega)$ is $\mathbb{P}_x$-integrable μ-a.s.

By definition, the first marginal of $\mathbb{Q}$ is μ:

$$\mathbb{Q}(E\times\Omega) = \int_E \mathbb{P}_x(\Omega)\mu(\mathrm{d}x) = \int_E \mu(\mathrm{d}x) = \mu(E), \quad E\in\mathscr{B}(\mathbb{R}).$$

The second marginal, which we call $\mathbb{P}$, results from μ-weighting of the measures $\mathbb{P}_x$, more precisely

$$\mathbb{P}(A) := \mathbb{Q}(\mathbb{R}\times A) = \int_{\mathbb{R}} \mathbb{P}_x(A)\mu(\mathrm{d}x), \quad A\in\mathscr{F}.$$

As we know, this marginal is simply the image law under the projection on the second coordinate. We thus have the following corollary of Theorems 4.24 and 4.15.

Corollary 4.25. *In the setup of Theorem 4.24, let $\mathbb{P}$ be the marginal of $\mathbb{Q}$ on Ω and X be a positive variable on $(\Omega,\mathscr{F})$. Then $x\to\int_\Omega X(\omega)\mathbb{P}_x(\mathrm{d}\omega)$ is measurable and*

$$\int_\Omega X(\omega)\mathbb{P}(\mathrm{d}\omega) = \int_{\mathbb{R}} \mu(\mathrm{d}x)\int_\Omega X(\omega)\mathbb{P}_x(\mathrm{d}\omega).$$

We saw above a rich way to construct measures on the product space and how to integrate against them. In fact, this construction is exhaustive: under mild assumptions on Ω any measure $\mathbb{Q}$ on the product space $\mathbb{R}\times\Omega$ can be *disintegrated* to be in the form (17). This naturally extends to general products $\Omega_1\times\Omega_2$, again under some assumptions.

5 Complements and further results on integration

We stick to the setting of a probability space. All of what follows, with some care given to renormalisation, extends to finite measures. Most results extend to σ-finite measures. Some arguments extend to arbitrary measures. An interested and motivated reader can explore such extensions.

Throughout this section we work on a fixed probability space $(\Omega, \mathscr{F}, \mathbb{P})$. We often drop it from the conventional notation, e.g., the space $\mathscr{L}^p(\Omega, \mathscr{F}, \mathbb{P})$ is simply denoted $\mathscr{L}^p$.

5.1 Modes of convergence

If the X_n in Example 3.14 have mean zero and variance one, then setting

$$B = \left\{ \limsup_{n\to\infty} \frac{S_n}{\sqrt{2n \log\log n}} = 1 \right\}, \tag{18}$$

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

$$\mathbb{P}\left[\frac{X_1 + \cdots + X_n}{\sqrt{n}} \leqslant a \right] = \mathbb{P}\left[\frac{S_n}{\sqrt{n}} \leqslant a \right] \overset{n\to\infty}{\longrightarrow} \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} \right) \mathrm{d}x. \tag{19}$$

This is the Central Limit Theorem without which statistics would be a very different subject. How does it fit with (18)? The results (18) and (19) are giving quite different results about the behaviour of S_n for large n. They correspond to different 'modes of convergence'.

Definition 5.1. 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.

Definition 5.2 (Modes of convergence)**.** Let $X_1, X_2, \ldots$ and X be random variables. We say that X_n converges to X

- *almost surely* (written $X_n \overset{\text{a.s.}}{\to} X$ or $X_n \to X$ a.s.) if
$$\mathbb{P}[X_n \to X] = \mathbb{P}\Big[\big\{ \omega : \lim_{n\to\infty} X_n(\omega) = X(\omega) \big\} \Big] = 1.$$
- *in probability* (written $X_n \overset{\mathbb{P}}{\to} X$) if, for every $\varepsilon > 0$,
$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = \lim_{n\to\infty} \mathbb{P}\big[\big\{ \omega : |X_n(\omega) - X(\omega)| > \varepsilon \big\} \big] = 0.$$
- *in $\mathscr{L}^p$* (or *in L^p*, or *in pth moment*), written $X_n \overset{L^p}{\to} X$, if all $X, X_n \in \mathscr{L}^p$, $n \geqslant 1$ and $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0$.
- *weakly in $\mathscr{L}^1$* (or *in the $\sigma(L^1, L^\infty)$ topology*) if $X_n, X \in \mathscr{L}^1$, $n \geqslant 1$ and
$$\lim_{n\to\infty} \mathbb{E}[X_n Y] = \mathbb{E}[XY], \quad \forall \text{ bounded r.v. } Y.$$
- *in distribution* (or *weakly*) (written $X_n \overset{d}{\to} X$ or $X_n \Rightarrow X$) if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for every $x \in \mathbb{R}$ *at which F_X is continuous* and where F_Y denotes the distribution function of Y.

These notions of convergence are all different. The notion of weak convergence in $\mathscr{L}^1$ will not be used for now. We will come back to it when we discuss *uniform integrability* in §5.4. Note also that the last notion, that of convergence in distribution, is very different to the others: it only depends on the particular sequence of random variables through their distributions. In particular, it makes sense even if all X_n are defined on different probability spaces, unlike all the other notions.

For now we note the following easy relations.

Convergence a.s. $\Longrightarrow$ **Convergence in Probability** $\Longrightarrow$ **Convergence in Distribution**

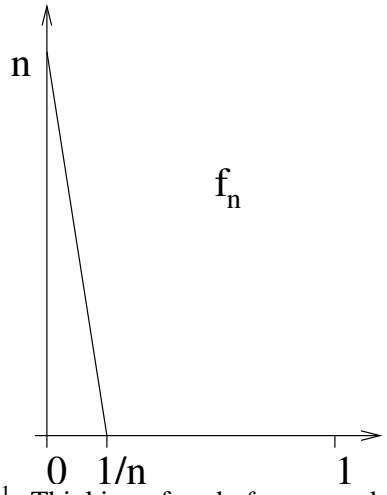
$\Uparrow$

Convergence in L^p

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

Example 5.3 (Convergence a.s. does not imply convergence in L^1)**.** On the probability space $\Omega = [0,1]$ with the Borel σ-algebra and Lebesgue measure, consider the sequence of functions f_n given by

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



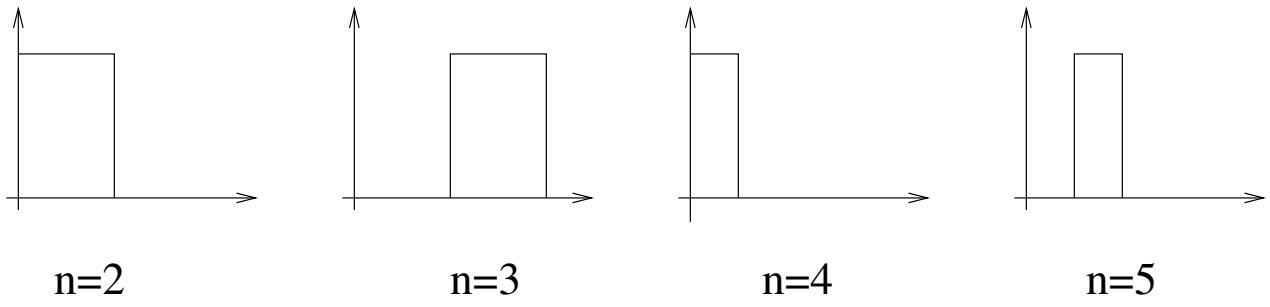
Then $f_n \to 0$ almost everywhere on $[0,1]$ but $f_n \not\to 0$ in L^1. Thinking of each f_n as a random variable, we have $f_n \to 0$ almost surely but $f_n \not\to 0$ in L^1.

Example 5.4 (Convergence in probability does not imply a.s. convergence)**.** To understand what's going on in (18) and (19), 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

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

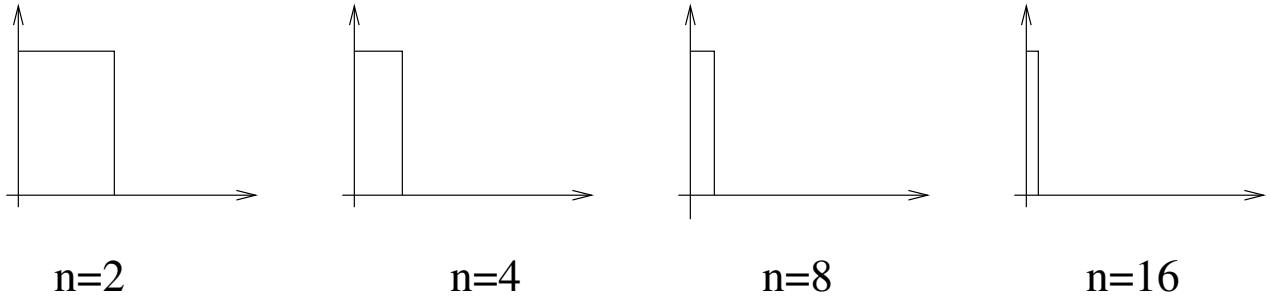
Pictorially we have a 'moving blip' which travels repeatedly across $[0,1]$ getting narrower at each pass.



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 $X_{2^n} \xrightarrow{\text{a.s.}} 0$.

It turns out that this is a general phenomenon.

Theorem 5.5 (Convergence in Probability and a.s. Convergence)**.** *Let $X_1, X_2, \ldots$ and X be random variables.*

(i) If $X_n \xrightarrow{\text{a.s.}} X$ then $X_n \xrightarrow{\mathbb{P}} X$.

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

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

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

(i) Suppose $X_n \xrightarrow{\text{a.s.}} X$. Then for any $\varepsilon > 0$ we have $\mathbb{P}[A_{n,\varepsilon} \text{ i.o.}] = 0$. By Fatou's Lemma for sets (Lemma 3.17, we have

$$0 = \mathbb{P}[A_{n,\varepsilon} \text{ i.o.}] = \mathbb{P}[\limsup_{n\to\infty} A_{n,\varepsilon}] \geqslant \limsup_{n\to\infty} \mathbb{P}[A_{n,\varepsilon}]$$

and in particular $\mathbb{P}[A_{n,\varepsilon}] \to 0$, so $X_n \xrightarrow{\mathbb{P}} X$.

(ii) This is the more interesting direction. Suppose that $X_n \xrightarrow{\mathbb{P}} X$. 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 $X_{n_k} \xrightarrow{\text{a.s.}} X$. □

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.

5.2 Some useful inequalities

We turn now to some inequalities which, in particular, often prove useful in the context discussed above. The first is trivial, but has many applications.

Lemma 5.6 (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. Let $\lambda > 0$. Then, 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 5.7 (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 5.7 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 5.8. *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 5.9 (Weak law of large numbers)**.** *Let $(X_n)_{n\geqslant 1}$ be i.i.d. random variables with mean m and variance $\sigma^2 < \infty$. Set*

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

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

Proof. We have $\mathbb{E}[\overline{X}(n)] = n^{-1}\sum_{i=1}^{n}\mathbb{E}[X_i] = m$ 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,

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

□

Definition 5.10 (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 5.11 (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*

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

Perhaps the nicest proof of Theorem 5.11 rests on the following geometric lemma.

Lemma 5.12. *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 m we can choose $a = f'(m)$) we have that $f(x) \geqslant f(m) + a(x-m)$ for all $x \in I$. □

Proof of Theorem 5.11. 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. □

As a byproduct of the proof, since a convex function is bounded from below by an affine function, $\mathbb{E}[f(X)]$ is well defined, possibly infinite.

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

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

5.3 $\mathscr{L}^p$ spaces

We comment a bit more on the structure and properties of $\mathscr{L}^p$ spaces. Those of you who take the Banach spaces course will see this done in a more systematic and general way. We will encounter Banach spaces, in particular Hilbert spaces, time and again in probability. Those who continue to study martingales in continuous time will use the Riesz representation theorem of elements in the dual space of a given Hilbert space.

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 5.13. *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 5.14. *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*

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

Proof. Proofs of these inequalities on $(\mathbb{R},\mathscr{B}(\mathbb{R}),\texttt{Leb})$ 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 $\mathbb{Q}(A)=\mathbb{E}[|X|^p\mathbf{1}_A]/\|X\|_p^p$, as we did in §4.2, and a random variable $Z:=|Y|/|X|^{p-1}\mathbf{1}_{|X|>0}$. Applying Jensen's inequality with $f(x)=x^q$, we have

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

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 $1-1/q=1/p$. □

Here is a useful application of Hölder's inequality.

Lemma 5.15. *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, in analogy to Example 4.21, 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. □

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

Theorem 5.16. *Let $p\geqslant 1$. The vector space $\mathscr{L}^p$ is* complete*, i.e., for any sequence $(X_n)_{n\geqslant1}\subseteq\mathscr{L}^p$ 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. We proceed in analogy to the proof of (ii) in Theorem 5.5 above. 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 $Y=\sum_{n\geqslant1}|X_{k_n}-X_{k_{n+1}}|$. 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. □

Deep Dive

A *Banach space* is a normed vector space which is complete. The above shows that $\mathscr{L}^p$ is almost a Banach 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. From the function analytic point of view it is a Banach space and a nicer object than $\mathscr{L}^p$. From the probabilistic point of view, we like to work with actual functions. This is, in particular, since when we have a large family $(X_t)_{t\geqslant 0}$ of functions, changing each of them on a null set may actually do a lot of harm!

5.4 Uniform integrability

We come back now to the issue of passing from convergence of random variables to convergence of integrals. Specifically, we are interested in passing from convergence in probability to convergence in $\mathscr{L}^1$ (this will then in particular also deal with a.s. convergence in one go). The right notion which provides an equivalence between the two is given by:

Definition 5.17 (Uniform Integrability)**.** A collection $\mathscr{C}$ of random variables is called *uniformly integrable* (UI) if

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

To put the above into words: 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 6.11 below.

Example 5.18. 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 5.19. 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 5.20. 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 5.21. 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

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

Proposition 5.22. *Let $\mathscr{C}$ be a family of random variables. Then $\mathscr{C}$ is UI if and only if*

(i)

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

(ii) and

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

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:

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

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 5.23. *Let (X_n) be a sequence of random variables with $X_n \to X$ in probability, and suppose that X and all X_n are bounded by the same real number K. Then $X_n \to X$ in L^1.*

Proof. We use an idea which recurs again and again in this context: split by whether the relevant quantity is 'small' or 'large'. Specifically, fix $\varepsilon > 0$. Let A_n be the event $\{|X_n - X| > \varepsilon\}$. Then

$$\begin{aligned} \mathbb{E}[|X_n - X|] &= \mathbb{E}\big[|X_n - X|\mathbf{1}_{A_n} + |X_n - X|\mathbf{1}_{A_n^c}\big] \\ &\leqslant \mathbb{E}[|X_n|\mathbf{1}_{A_n}] + \mathbb{E}[|X|\mathbf{1}_{A_n}] + \varepsilon \\ &\leqslant 2\mathbb{E}[K\mathbf{1}_{A_n}] + \varepsilon = 2K\mathbb{P}[A_n] + \varepsilon. \end{aligned} \tag{20}$$

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 5.4 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 5.20. 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 5.24 (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 5.23, using the bound (20). Since $|X_n| \to |X|$ in probability, by Theorem 5.5 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 5.22, i.e., X is integrable. Now fix $\varepsilon > 0$, and let $A_n = \{|X_n - X| > \varepsilon\}$. As before,

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

Since $X_n \to X$ in probability we have $\mathbb{P}[A_n] \to 0$ as $n \to \infty$, so by Proposition 5.22 (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 Lemma 4.14.

It remains to show $(iii) \Rightarrow (i)$. Note that we can not repeat the arguments in the proof of Lemma 4.14 which relied on a.s. convergence to use the DCT. Instead, we use the bounded convergence result Lemma 5.23. 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 5.21 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 5.23, $\mathbb{E}[Y_n \wedge K] \to \mathbb{E}[Y \wedge K]$. 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. □

5.5 Further results on UI (Deep Dive)

Deep Dive

The following is very helpful in thinking about UI. While Proposition 5.22 makes it clear that just uniform bound on the first moments is not enough for UI, in fact anything more than that already is.

Theorem 5.25 (La Vallée Poussin)**.** *Let $\mathscr{C} \subseteq \mathscr{L}^1$. Then $\mathscr{C}$ is UI if and only if there exists a positive increasing and convex $g : \mathbb{R}_+ \to \mathbb{R}$ such that*

$$\lim_{x\to\infty} \frac{g(x)}{x} = \infty$$

and

$$\sup_{X\in\mathscr{C}} \mathbb{E}[g(|X|)] < \infty.$$

One example of g which we shall meet later on is given by $g(x) = x \log x$.

Proof. TBC □

Let us look again at the Definition of UI. It says that for any $\varepsilon > 0$, we can write each $X \in \mathscr{C}$ as $X = X\mathbf{1}_{\{|X|\leqslant K\}} + X\mathbf{1}_{\{|X|>K\}}$, where the first variable is obviously bounded and the second one is small in $\mathscr{L}^1$. To rephrase, $\mathscr{C}$ is UI if and only if, for any $\varepsilon > 0$, there exists K such that $\mathscr{C}$ is contained in the Minkowski sum

$$\mathscr{C} \subset B_K^\infty + B_\varepsilon^1 := \{Y + Z : Y \in B_K^\infty, Z \in B_\varepsilon^1\},$$

where B_ε^1 is a ball in $\mathscr{L}^1$, $B_\varepsilon^1 = \{Z \in \mathscr{L}^1 : \mathbb{E}[|Z|] \leqslant \varepsilon\}$ and B_K^∞ is a ball in $\mathscr{L}^\infty$ seen as a subset in $\mathscr{L}^1$, $B_K^\infty = \{Y \in \mathscr{L}^1 : |Y(\omega)| \leqslant K \ \forall \omega \in \Omega\}$. Note that the Minkowski sum is a convex set so if it contains $\mathscr{C}$ it also contains its convex hull. It follows that if $\mathscr{C}$ is UI then so is its convex hull. Similarly, if a sequence in $\mathscr{C}$ converges in $\mathscr{L}^1$ to some X then we can also add X to $\mathscr{C}$ without affecting UI. Note also that a union of two UI families $\mathscr{C}, \mathscr{D}$ is still UI and hence so is $\mathscr{C} + \mathscr{D}$ (since $\frac{1}{2}(\mathscr{C} + \mathscr{D})$ is a subset of the convex hull of $\mathscr{C} \cup \mathscr{D}$). All of these properties become natural in light of the following result.

Theorem 5.26 (Dunford–Pettis)**.** *Let $\mathscr{C} \subseteq L^1$. TFAE*

(i) $\mathscr{C}$ is UI

(ii) $\mathscr{C}$ is relatively weakly compact (i.e., in the $\sigma(L^1, L^\infty)$ topology the closure is compact)

(iii) every sequence of elements in $\mathscr{C}$ contains a subsequence converging in $\sigma(L^1, L^\infty)$.

Sketchy sketch of $(i) \Rightarrow (ii)$. From (i) to (ii): consider $\mathbb{Q}(A) := \lim_{\mathfrak{U}} \mathbb{E}[X\mathbf{1}_A]$, where $\mathfrak{U}$ is an ultrafilter on $\mathscr{C}$ and $A \in \mathscr{F}$. Part (i) in Proposition 5.22 shows the limit is well defined, while part (ii), together with Lemma 2.4, shows it is a measure. Using Theorem 4.9 we get $\xi = \frac{d\mathbb{Q}}{d\mathbb{P}}$, in particular $\xi \in \mathscr{L}^1$, and show

that $\lim_{\mathfrak{U}} \mathbb{E}[XY] = \mathbb{E}[\xi Y]$ for any $Y \in \mathscr{L}^\infty$. This is easy for a simple Y and then follows with a universal approximation argument in Lemma 1.26.

The reverse, from (ii) to (i), is more difficult. Equivalence between (ii) and (iii) follows from Eberlein–Smulian theorem, a difficult result which asserts that different types of compactness are equivalent for the weak topology on a Banach space. □

6 Conditional Expectation

From now on, we work on a fixed probability space $(\Omega, \mathscr{F}, \mathbb{P})$. All the random variables are assumed to be defined on $(\Omega, \mathscr{F})$.

As already stated, independence and conditional expectation are the two key notions which set probability alive. We saw the former in §3 and are now about the develop the latter.

6.1 Intuition

Our objective is to capture in a mathematically rigorous way, the intuition that our assessment of probabilities, and hence of behaviour of random variables, should change as a function of our information. In Prelims we did this through the notion of conditional probability. Suppose we consider an event A. Then, in absence of any information, we assess its likelihood as $\mathbb{P}(A)$. However, if someone tells us that an even B actually happens, then we re-assess the chances of A as $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$. Except that this is a post-factum assessment, once we know that B has happened. A more forward thinking approach would be say: suppose you had the information about B, i.e., you shall know if it happens or not, then how would you assess chances of A? We already answered this in §2.2 and the answer was given in (9):

$$\mathbb{E}\left[\mathbf{1}_A \mid \sigma(B)\right](\omega) = \mathbb{P}(A \mid \sigma(B))(\omega) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}\mathbf{1}_B(\omega) + \frac{\mathbb{P}(A \cap B^{\mathrm{c}})}{\mathbb{P}(B^{\mathrm{c}})}\mathbf{1}_{B^{\mathrm{c}}}(\omega).$$

As expected the answer takes one value if B happens and another if B^{c} does. Note that we used expectations notation above, harmless here since $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)$, but more suitable to moving from indicators to more general random variables. For an integrable random variable X we already know from Exercise 4.17 that $\mathbb{E}[X]$ is the single best approximation, in the quadratic sense, to X using a constant. But if we are allowed to use instead a random variable taking two values, one if B happens and another if B^{c} does, then we would conjecture

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

It turns out this answer is correct as the optimality property, known as the mean square approximation, is preserved.

Exercise 6.1. Let X be an integrable random variable and $B \in \mathscr{F}$ with $\mathbb{P}(B) > 0$. For $\alpha, \beta \in \mathbb{R}$ let $Y_{\alpha,\beta} := \alpha\mathbf{1}_B + \beta\mathbf{1}_{B^{\mathrm{c}}}$. Show that

$$\inf_{\alpha,\beta\in\mathbb{R}} \mathbb{E}[(X - Y_{\alpha,\beta})^2]$$

is attained by $Y_{\alpha,\beta} = \mathbb{E}[X \mid \sigma(B)]$ above.

It is also easy to see how the above could generalise to a more detailed information: suppose $(B_n)_{n\geqslant 1}$ is a partition of Ω, i.e., the sequence is pairwise disjoint and $\bigcup_{n\geqslant 1} B_n = \Omega$, and that $\mathbb{P}(B_n) > 0$ for all $n \geqslant 1$. Then

$$\mathbb{E}\left[\mathbf{1}_A \mid \sigma(B_n : n \geqslant 1)\right](\omega) = \mathbb{P}(A \mid \sigma(B_n : n \geqslant 1))(\omega) = \sum_{n\geqslant 1} \frac{\mathbb{P}(A \cap B_n)}{\mathbb{P}(B_n)}\mathbf{1}_{B_n}(\omega)(\omega)$$

or, more generally, for an integrable random variable X,

$$\mathbb{E}\left[X \mid \sigma(B_n : n \geqslant 1)\right](\omega) = \sum_{n\geqslant 1} \frac{\mathbb{E}\left[X\mathbf{1}_{B_n}\right]}{\mathbb{P}(B_n)}\mathbf{1}_{B_n}(\omega) \tag{21}$$

is undoubtedly the right object. Our information is on the levels of B_n's – we are able to tell them apart and hence can reason on each of these instead of the whole of Ω. On each B_n, we just use the old good conditional

probability or averaging of X. The outcome is a random variable, taking possibly countably many different values, which tells us how we shall be evaluating the chances of A happening, or approximating X, depending on our information about B_n's. However, it is not clear how to proceed further as this is where the intuition stops really! If we had an uncountable family, each B_i with $\mathbb{P}(B_i) = 0$, $i \in I$, then we have no apparent way of making sense of the above.

6.2 Definition, existence and uniqueness

If we consider more general types of information, i.e., if we want to condition on a σ-algebra $\mathscr{G} \subset \mathscr{F}$, we can not hope to reason set-by-set or ω-by-ω. Instead we can appeal to the optimal prediction property. Above, one can show that $\mathbb{E}[X \mid \sigma(B_n : n \geqslant 1)]$ minimises the prediction error $\mathbb{E}[(X-Y)^2]$ among all $Y = \sum_{n\geqslant 1} \alpha_n \mathbf{1}_{B_n}$. But this gets a bit tedious if we do it by hand. And it essentially just follows from the fact that on the smallest level of granularity allowed, i.e., on the sets B_n, we use the best constant to approximate X: its expectation on that set. Thus, by definition, we have that the average of $\mathbb{E}[X \mid \sigma(B_n : n \geqslant 1)]$ over any set we 'know' or can distinguish, i.e., any B_n, is the same as average of X. This and the fact that $\mathbb{E}[X \mid \sigma(B_n : n \geqslant 1)]$ has to be $\sigma(B_n : n \geqslant 1)$-measurable leads to the following definition:

Definition 6.2 (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 Y is (a version of) the *conditional expectation of X given $\mathscr{G}$* if Y is integrable, $\mathscr{G}$-measurable and

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

The integrals of X and Y over sets $G \in \mathscr{G}$ are the same – this is our averaging property – but Y is also $\mathscr{G}$ measurable whereas X is $\mathscr{F}$-measurable. The *conditional probability* is defined simply as the conditional expectation of an indicator

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

for $A \in \mathscr{F}$. It is easy to see that when $\mathscr{G} = \sigma(B)$, for an event $B \in \mathscr{F}$, the natural object proposed in (9) satisfies Definition 6.2. More generally, the following result takes care of the first two questions you may want to ask.

Theorem 6.3 (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 and is denoted $\mathbb{E}[X \mid \mathscr{G}]$. It is a.s. unique in the sense that if Z is also 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. □

We will come back to the proof of existence later. Let us reiterate that the conditional expectation satisfies

$$\int_G \mathbb{E}[X \mid \mathscr{G}]\, \mathrm{d}\mathbb{P} = \int_G X \, \mathrm{d}\mathbb{P} \quad \text{for all } G \in \mathscr{G}, \tag{22}$$

i.e., using the expectation notation, $\mathbb{E}[\mathbb{E}[X \mid \mathscr{G}]\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G]$, and we shall call (22) the *defining relation*.

Remark 6.4. If $\mathbb{E}[X] = \mathbb{E}[Y]$ then the DCT shows that the family of sets G for which (22) is true forms a λ-system. A direct application of π-λ systems lemma thus shows that it is enough to verify (22) for $G \in \mathscr{A} \cup \{\Omega\}$ for a π-system $\mathscr{A}$ generating $\mathscr{G}$. While simple, this remark is very useful.

Our first task is to verify that (21) was a correct guess. And, with the above remark, it is enough to check that (22) is satisfied for $G = B_n$. This is trivial, since if Y denotes the random variable on the right hand side of (21) then

$$\mathbb{E}[Y\mathbf{1}_{B_n}] = \frac{\mathbb{E}[X\mathbf{1}_{B_n}]}{\mathbb{P}(B_n)}\mathbb{E}[\mathbf{1}_{B_n}] = \mathbb{E}[X\mathbf{1}_{B_n}].$$

Since the definition of the conditional expectation is so important, let us explain it once again, considering the case $\mathcal{G} = \sigma(\xi)$ for some random variable ξ. In this case, we often simple write $\mathbb{E}[X \mid \xi]$ instead of $\mathbb{E}[X \mid \sigma(\xi)]$. So, $Y = \mathbb{E}[X \mid \xi]$ is supposed to be a random variable which depends only on the value of ξ, in the sense that

$$\text{“}Y(\omega) = \mathbb{E}[X \mid \xi = z] = \mathbb{E}[X\mathbf{1}_{\{\xi=z\}}]/\mathbb{P}[\xi = z]''$$

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

$$\mathbb{E}[Y\mathbf{1}_{\{\xi=z\}}] = \mathbb{E}[X\mathbf{1}_{\{\xi=z\}}].$$

Still, if $\mathbb{P}[\xi = z] = 0$ for every z (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 Y is a function of ξ and

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

for each $A \in \mathcal{B}(\mathbb{R})$. This is exactly what Definition 6.2 says in the case $\mathcal{G} = \sigma(\xi)$. Note that thanks to Theorem 1.27, we can say that $\mathbb{E}[X|\xi] = f(\xi)$ for some measurable function ξ. Thus, intuitively, we have '$f(z) = \mathbb{E}[X|\xi = z]$' except the concept of the conditional expectation actually makes sense of this even if $\mathbb{P}(\xi = z) = 0$ for all $z \in \mathbb{R}$.

In general, it is not the values of ξ that matter, but the 'information' in ξ, coded by the σ-algebra ξ generates, so we define conditional expectation with respect to an arbitrary σ-algebra $\mathcal{G}$. This then covers cases such as conditioning on two random variables at once and much more.

Remark. So far, we defined conditional expectations only when X is integrable. Just as with ordinary expectation, the definitions work without problems if $X \geqslant 0$, allowing $+\infty$ as a possible value. This is (an option) exercise for you to check.

6.3 Important properties

We now turn to basic properties of the 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 6.5. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X and Y integrable random variables, $\mathcal{G} \subseteq \mathcal{F}$ a σ-algebra and a, b, c real numbers. Then*

(i) $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X]$.

(ii) $\mathbb{E}[aX + bY + c \mid \mathcal{G}] \overset{\text{a.s.}}{=} a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}] + c$.

(iii) If X is $\mathcal{G}$-measurable, then $\mathbb{E}[X \mid \mathcal{G}] \overset{\text{a.s.}}{=} X$.

(iv) $\mathbb{E}[c \mid \mathcal{G}] \overset{\text{a.s.}}{=} c$.

(v) $\mathbb{E}[X \mid \{\emptyset, \Omega\}] = \mathbb{E}[X]$.

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

(vii) If $X \leqslant Y$ a.s. then $\mathbb{E}[X \mid \mathcal{G}] \leqslant \mathbb{E}[Y \mid \mathcal{G}]$ a.s. In particular, if $X \geqslant 0$ a.s. then $\mathbb{E}[X \mid \mathcal{G}] \geqslant 0$ a.s.

(viii) $\big|\mathbb{E}[X \mid \mathcal{G}]\big| \leqslant \mathbb{E}[|X| \mid \mathcal{G}]$ a.s.

Proof. The proofs all follow from the requirement that $\mathbb{E}[X \mid \mathcal{G}]$ be $\mathcal{G}$-measurable and the defining relation (22). We just do some examples.

(i) Set $G = \Omega$ in the defining relation.

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

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

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

(v) 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.

(vi) Note that $\mathbb{E}[X]$ is $\mathcal{G}$-measurable and for $G \in \mathcal{G}$

$$\mathbb{E}[\mathbb{E}[X]\mathbf{1}_G] = \mathbb{E}[X]\mathbb{P}[G] = \mathbb{E}[X]\mathbb{E}[\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G],$$

so the defining relation holds, where in the last equality we used independence and Proposition 3.10.

(vii) By linearity it is enough to show the 'in particular' part. Suppose $X \geqslant 0$. If $\mathbb{P}(\mathbb{E}[X \mid \mathcal{G}] < 0) > 0$ then $\mathbb{P}(A) > 0$, where $A = \{\mathbb{E}[X \mid \mathcal{G}] \leqslant -1/n\}$ for some $n > 0$. Since $A \in \mathcal{G}$, by (22), we have

$$0 \leqslant \mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A] \leqslant -\frac{\mathbb{P}(A)}{n} < 0$$

a contradiction. □

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

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

Proposition 6.6 (Conditional Convergence Theorems)**.** *Let $X_1, X_2, \ldots$ and X be integrable random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra.*

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

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

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

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

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

Proof. The proofs all use the defining relation (22) to transfer statements about convergence of the conditional probabilities to our usual convergence theorems. We give details for cMCT and leave the rest as an exercise.

Let $Y_n = \mathbb{E}[X_n \mid \mathcal{G}]$. By Proposition 6.5 (vii) we know that $Y_n \geqslant 0$ a.s. and $A_n = \{Y_n < Y_{n-1}\} \in \mathcal{G}$ and is null, $\mathbb{P}(A_n) = 0$. Let $Y := \limsup_{n\to\infty} Y_n$ and $A = \bigcup_{n\geqslant 2} A_n$. Then $A \in \mathcal{G}$ is a null set, $\mathbb{P}(A) = 0$, Y is $\mathcal{G}$-measurable and outside of A it is an increasing limit of Y_n's. For any $G \in \mathcal{G}$ we have

$$\mathbb{E}[Y\mathbf{1}_G] = \mathbb{E}[Y\mathbf{1}_{G\cap A^c}] \overset{MCT}{=} \lim_{n\to\infty} E[Y_n\mathbf{1}_{G\cap A^c}] \overset{(22)}{=} \lim_{n\to\infty} \mathbb{E}[X_n\mathbf{1}_{G\cap A^c}] \overset{MCT}{=} \mathbb{E}[X\mathbf{1}_{G\cap A^c}] = \mathbb{E}[X\mathbf{1}_G].$$

Taking $G = \Omega$, $\mathbb{E}[Y] = \mathbb{E}[X] < \infty$ and it follows that Y is a version of $\mathbb{E}[X \mid \mathcal{G}]$, as required. □

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

Lemma 6.7. *Let X and Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with X, Y and XY integrable. Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra and suppose that Y is $\mathcal{G}$-measurable. Then*

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

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

First suppose that X and Y are non-negative. If $Y = \mathbf{1}_A$ for some $A \in \mathcal{G}$, then for any $G \in \mathcal{G}$ we have $G \cap A \in \mathcal{G}$ and so by the defining relation (22) for $\mathbb{E}[X \mid \mathcal{G}]$

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

Now extend by linearity to simple positive Ys. Now suppose that $Y \geqslant 0$ is $\mathcal{G}$-measurable. Then there is a sequence $(Y_n)_{n\geqslant 1}$ of simple $\mathcal{G}$-measurable random variables with $Y_n \uparrow Y$ as $n \to \infty$, it follows that $Y_nX \uparrow YX$ and we conclude by cMCT and a.s. uniqueness of the conditional expectation. Finally, for X, Y not necessarily non-negative, write $XY = (X^+ - X^-)(Y^+ - Y^-)$ and use linearity of the integral. □

Proposition 6.8 (Tower property of conditional expectations)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X an integrable random variable and $\mathcal{F}_1$, $\mathcal{F}_2$ σ-algebras with $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}$. Then*

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

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

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

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

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

□

This extends (i) of Proposition 6.5 which (in the light of (v)) is just the case $\mathcal{F}_1 = \{\emptyset, \Omega\}$.

Jensen's inequality, Theorem 5.11, also extends to the conditional setting.

Proposition 6.9 (Conditional Jensen's Inequality)**.** *Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and that X is an integrable random variable taking values in an open interval $I \subseteq \mathbb{R}$. Let $f : I \to \mathbb{R}$ be convex and let $\mathcal{G}$ be a sub σ-algebra of $\mathcal{F}$. If $\mathbb{E}[|f(X)|] < \infty$ then*

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

Proof. A convex function f on I is continuous and can be represented as the supremum over a countable family of affine functions $\{l_n : n \geqslant 1\}$ on I. Indeed, we may simply take l_n to be supporting tangents from Lemma 5.12 over a dense sets of m_n in I. We have

$$l_n(\mathbb{E}[X \mid \mathscr{G}]) = \mathbb{E}[l_n(X) \mid \mathscr{G}] \leqslant \mathbb{E}[f(X)|\mathscr{G}] \quad \text{a.s.}$$

and since a countable union of null sets is null, we may assume that the above holds a.s. for all $n \geqslant 1$ simultaneously. The result follows by taking the supremum in n. □

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

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

A very simple special case of this is the following.

Example 6.10. Suppose that X is a non-trivial non-negative random variable: $X \geqslant 0$ and $\mathbb{P}(X > 0) > 0$. Then

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

Proof. Let $A = \{X > 0\}$ and note that $\mathbb{E}[X\mathbf{1}_{A^c}] = 0$ and $\mathbb{E}[X] = \mathbb{E}[X\mathbf{1}_A]$. In particular

$$\mathbb{E}[X \mid \sigma(A)] = \frac{\mathbb{E}[X]}{\mathbb{P}(A)}\mathbf{1}_A.$$

Using Proposition 6.5 (i) and Proposition 6.9,

$$\mathbb{E}[X^2] = \mathbb{E}\big[\mathbb{E}[X^2 \mid \sigma(A)]\big] \geqslant \mathbb{E}\big[\mathbb{E}[X \mid \sigma(A)]^2\big] = \frac{\mathbb{E}[X]^2}{\mathbb{P}(A)}.$$

Rearranging gives the result. □

Deep Dive

Taking expectations in the conditional Jensen for $f(x) = |x|^p$, $p \geqslant 1$, tells us that for $X \in \mathscr{L}^p$,

$$\|\mathbb{E}[X|\mathscr{G}]\|_p \leqslant \|X\|_p,$$

or in functional analytic terms, $X \to \mathbb{E}[X|\mathscr{G}]$ is a linear operator on L^p with norm $\leqslant 1$. It follows that it is also continuous in the weak topology, i.e., when L^p is endowed with the $\sigma(L^p, L^q)$ topology.

The following provides a very important example of families of uniformly integrable random variables. Indeed, such families will play a key role in the remainder of this course. In the important special case when $(\mathscr{F}_n)$ is a filtration, (X_n) is a martingale, see Example 8.7.

Theorem 6.11. *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*

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

is uniformly integrable.

Proof. Since $f(x) = |x|$ is convex, by the conditional form of Jensen's inequality (Proposition 6.9),

$$|X_\alpha| = |\mathbb{E}[X \mid \mathscr{F}_\alpha]| \leqslant \mathbb{E}\big[|X| \mid \mathscr{F}_\alpha\big] \text{ a.s.} \tag{23}$$

and in particular $\mathbb{E}[|X_\alpha|] \leqslant \mathbb{E}[|X|]$ for all $\alpha \in I$ so that (i) in Proposition 5.22 holds. Also, using (23),

$$\mathbb{E}[|X_\alpha|\mathbf{1}_{\{|X_\alpha|>K\}}] \leqslant \mathbb{E}\left[\mathbb{E}[|X| \mid \mathscr{F}_\alpha]\mathbf{1}_{\{|X_\alpha|>K\}}\right] = \mathbb{E}[|X|\mathbf{1}_{\{|X_\alpha|>K\}}], \tag{24}$$

since we may move the indicator function inside the conditional expectation and then apply the tower law. Since $\{X\}$ is UI, applying Proposition 5.22, 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. □

Finally, we come back to the optimality property discussed in Exercises 4.17 and 6.1. This was our motivating property and it is reassuring to see it holds throughout!

Remark (Conditional Expectation via 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 6.5 (i) and Lemma 6.7 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-term vanishes. The second term only depends on X and the first one is minimised by taking $Y = \mathbb{E}[X \mid \mathscr{G}]$. Thus $\mathbb{E}[(X-Y)^2]$ is minimised taking $Y = \mathbb{E}[X \mid \mathscr{G}]$ or, in other words, $\mathbb{E}[X \mid \mathscr{G}]$ is the best mean-square approximation of X among all $\mathscr{G}$-measurable random variables. We shall now use this property as our starting point to show existence of conditional expectations!

6.4 Orthogonal projection in $\mathscr{L}^2$

We need to develop an abstract equivalent of the well known projection in $\mathbb{R}^d$. We work in $\mathscr{L}^2$. It is (nearly) a Hilbert space and has a natural geometry. From a probabilistic point of view we centre random variables around their mean and consider variance and covariance.

Exercise 6.12. For $X, Y \in \mathscr{L}^2$ let

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])]\mathbb{E}[(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Show that $\mathrm{Cov}(\cdot,\cdot)$ is bilinear on $\mathscr{L}^2$ and that

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y), \quad \text{if } \mathrm{Cov}(X,Y) = 0.$$

When $\mathrm{Cov}(X,Y) = 0$ we say that X and Y are *uncorrelated*. Clearly if X and Y are independent then they are also uncorrelated. Show that the reverse does not need to hold (by means of a counterexample).

From a geometric point of view there is no need to centre things around their mean. We introduce a scalar product

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

Note that this is well defined since by Hölder's inequality, Theorem 5.14, $XY \in \mathscr{L}^1$. We say that X and Y are *orthogonal* if $\langle X,Y\rangle = 0$.

Lemma 6.13 (Pythagoras' theorem)**.** *If $X,Y \in \mathscr{L}^2$ are orthogonal then*

$$\|X+Y\|_2^2 = \|X\|_2^2 + \|Y\|_2^2.$$

Exercise 6.14. Show that $\langle\cdot,\cdot\rangle$ is bilinear on $\mathscr{L}^2$ and use it to establish the parallelogram law

$$\|X\|_2^2 + \|Y\|_2^2 = \tfrac{1}{2}\left(\|X+Y\|_2^2 + \|X-Y\|_2^2\right). \tag{25}$$

Recall from above that *completeness* means that Cauchy sequences converge to elements in the space.

Theorem 6.15. *Let $\mathscr{K}$ be a complete vector subspace of $\mathscr{L}^2$. For any $X \in \mathscr{L}^2$ 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 $Z \in \mathscr{K}$.

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 $\mathscr{K}$.

Example 6.16. Let $\mathscr{K}$ be the vector space of random variables which are a.s. constant. Exercise 4.17 shows that the projection of X on $\mathscr{K}$ is given by $\mathbb{E}[X]$.

Proof of Theorem 6.15. 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 (25), 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

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

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 5.14, 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. □

Proof of existence in Theorem 6.3. 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 5.16. Let Y be the orthogonal projection of X on $\mathscr{K}$ from Theorem 6.15. 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 (22) hold.

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, Proposition 6.6, we know that $Y := \limsup_{n\to\infty} Y_n$ is a version of $\mathbb{E}[X \mid \mathscr{G}]$. □

6.5 Conditional Independence (Deep Dive)

Deep Dive

The notion of conditional independence appears very naturally when we discuss the Markov property. If you recall from Part A Probability, a simply way of saying that $(X_n)_{n\geqslant 0}$ is a Markov chain was to say that the future distribution of the chain only depends on path so far through the present state, or, that future and past are *conditionally independent given present*.

Definition 6.17. Let $\mathscr{F}_1,\mathscr{F}_2,\mathscr{F}_3$ be three sub-σ-algebras of $\mathscr{F}$. We say that $\mathscr{F}_1$ and $\mathscr{F}_3$ are conditionally independent given $\mathscr{F}_2$ if

$$\mathbb{P}(A_1\cap A_3 \mid \mathscr{F}_2)=\mathbb{P}(A_1\mid \mathscr{F}_2)\mathbb{P}(A_3\mid \mathscr{F}_2), \quad \text{a.s.}$$

for all $A_1\in\mathscr{F}_1, A_3\in\mathscr{F}_3$.

It is should be clear, by linearity of conditional expectation (see Proposition 6.5) and the conditional monotone convergence theorem (see Proposition 6.6), that the above is equivalent to saying that

$$\mathbb{E}[X_1X_3\mid \mathscr{F}_2]=\mathbb{E}[X_1\mid \mathscr{F}_2]\mathbb{E}[X_3\mid \mathscr{F}_2] \quad \text{a.s.}$$

for all non-negative random variables X_1 and X_3, respectively $\mathscr{F}_1$- and $\mathscr{F}_3$- measurable. We could also replace non-negativity by integrability of X_1X_3, X_1 and X_3. It is also clear that independence can be recovered by taking the trivial $\mathscr{F}_2=\{\emptyset,\Omega\}$.

Theorem 6.18. *Let $\mathscr{F}_1,\mathscr{F}_2,\mathscr{F}_3$ be three sub-σ-algebras of $\mathscr{F}$, and set $\mathscr{F}_{12}=\sigma(\mathscr{F}_1,\mathscr{F}_2)$. Then $\mathscr{F}_1$ and $\mathscr{F}_3$ are conditionally independent given $\mathscr{F}_2$ if and only if*

$$\mathbb{E}[X_3\mid \mathscr{F}_{12}]=\mathbb{E}[X_3\mid \mathscr{F}_2] \quad \textit{a.s.}$$

for all $\mathscr{F}_3$-measurable integrable random variable X_3.

Proof. To make various equalities clearer, we will refer to different properties in Proposition 6.5 simply by their list numbers $(i),(ii)$ etc. We will refer to the tower property of conditional expectation, Proposition 6.8, as (t) and to the property of "taking out what is known", Lemma 6.7, as (k).
$(\Rightarrow)$ We suppose $\mathscr{F}_1$ and $\mathscr{F}_3$ are conditionally independent given $\mathscr{F}_2$. By definition, $\mathbb{E}[X_3\mid\mathscr{F}_2]$ is $\mathscr{F}_2$-measurable and hence also $\mathscr{F}_{12}$-measurable. To establish the desired equality, we thus need to verify that

$$\mathbb{E}[\mathbb{E}[X_3\mid\mathscr{F}_2]\mathbf{1}_A]=\mathbb{E}[X_3\mathbf{1}_A]$$

for all $A\in\mathscr{F}_{12}$. This holds for $A=\Omega$ by (i). It is then easy to see that the family of sets A for which the above holds is a λ-system and thus it is enough, by Lemma 1.12, to verify it for the π-system of sets $A=A_1\cap A_2$, $A_1\in\mathscr{F}_1, A_1\in\mathscr{F}_2$. We have

$$\mathbb{E}[\mathbb{E}[X_3\mid\mathscr{F}_2]\mathbf{1}_{A_1}\mathbf{1}_{A_2}]\overset{(t)}{=}\mathbb{E}[\mathbb{E}[\mathbb{E}[X_3\mid\mathscr{F}_2]\mathbf{1}_{A_1}\mathbf{1}_{A_2}\mid\mathscr{F}_2]]\overset{(k)}{=}\mathbb{E}[\mathbb{E}[X_3\mid\mathscr{F}_2]\mathbf{1}_{A_2}\mathbb{E}[\mathbf{1}_{A_2}\mid\mathscr{F}_2]] \tag{26}$$

$$=\mathbb{E}[\mathbf{1}_{A_2}\mathbb{E}[X_3\mathbf{1}_{A_1}\mid\mathscr{F}_2]]\overset{(k)}{=}\mathbb{E}[\mathbb{E}[\mathbf{1}_{A_2}X_3\mathbf{1}_{A_1}\mid\mathscr{F}_2]]\overset{(i)}{=}\mathbb{E}[X_3\mathbf{1}_{A_1}\mathbf{1}_{A_2}], \tag{27}$$

as required, and where the third equality followed by the assumed conditional independence.
$(\Leftarrow)$ Suppose now that

$$\mathbb{E}[X_3\mid\mathscr{F}_{12}]=\mathbb{E}[X_3\mid\mathscr{F}_2] \quad \text{a.s.}$$

for all $\mathscr{F}_3$-measurable integrable random variable X_3. Then

$$\mathbb{E}[X_1X_3 \mid \mathscr{F}_2] \overset{(t)}{=} \mathbb{E}[\mathbb{E}[X_1X_3 \mid \mathscr{F}_{12}] \mid \mathscr{F}_2] \overset{(k)}{=} \mathbb{E}[X_1\mathbb{E}[X_3 \mid \mathscr{F}_{12}] \mid \mathscr{F}_2] = \mathbb{E}[X_1\mathbb{E}[X_3 \mid \mathscr{F}_2] \mid \mathscr{F}_2] \overset{(k)}{=} \mathbb{E}[X_1 \mid \mathscr{F}_2]\mathbb{E}[X_3 \mid \mathscr{F}_2],$$

as required and where the third equality followed by assumption. □

7 Filtrations and stopping times

The language and tools we have developed so far lend themselves beautifully to describing random phenomena occurring in time. 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 $(\Omega, \mathscr{F}, \mathbb{P})$. Note however, in analogy to §1, 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 step: when we consider the nature of the random evolution in §8.

Definition 7.1 (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*.

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

$$\mathscr{F}_\infty = \sigma\left(\bigcup_{n\geqslant 0} \mathscr{F}_n\right) \tag{28}$$

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 7.2 (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 7.3 (Natural filtration)**.** The *natural filtration* $(\mathscr{F}^X_n)_{n\geqslant 0}$ associated with a stochastic process $(X_n)_{n\geqslant 0}$ on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is defined by

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

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.

We talked above of the index n as the time. We can think of this as days, seconds or years. But it could also be some other, non-uniform, clock ticking. Whatever the real world interpretation of this clock may be, we shall refer to instances in this clock as *deterministic times*. It is maybe easiest to think of these as days and X_n could be, e.g., the temperature recorded in Greenwich Observatory at noon on this day, or the Rolls-Royce Holdings plc closing price at London Stock Exchange. However, in reality we 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 you when I see you'. 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 index by deterministic times as indexed by stopping times.

Definition 7.4 (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 $\mathbb{N} \cup \{\infty\} = \{0, 1, 2, \ldots, \infty\}$ 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 the current information $\mathscr{F}_n$ 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 2020 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 7.5. *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 7.6. *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.

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

□

The next thing we would like to understand is what information do we have at the moment τ? This is a random time, sometimes it may come early and sometimes very late. But intuitively, since we know it happens when it happens, we should be able to specify the information we have amassed by that time. This is now made precise.

Definition 7.7. Let τ be a stopping time on $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$. The σ-algebra of information at time τ is defined as

$$\mathscr{F}_\tau = \{A \in \mathscr{F}_\infty : A \cap \{\tau = n\} \in \mathscr{F}_n \ \forall n \geqslant 0\}. \tag{29}$$

So an event A is known by time τ if its part learned if $\tau = n$ is normally learned by time n. 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 7.8. *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* (29) *is a* σ*-algebra;*

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

Proof. Exercise. □

In particular, combining Propositions 7.5 and 7.8, 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 7.9 (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* $X^\tau = (X_{\tau\wedge n})_{n\geqslant 0}$ *is a stochastic process, called the* stopped process*.* X^τ *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 Corollary 1.19 and (29) since

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

□

8 Martingales in discrete time

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 – its 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.

8.1 Definitions, examples and first properties

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). Here is the key definition:

Definition 8.1 (Martingale, submartingales, supermartingale)**.** Let $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$ be a filtered probability space. An *integrable, $\mathscr{F}_n$-adapted* stochastic process $(X_n)_{n\geqslant 0}$ is called

(i) a *martingale* if for every $n \geqslant 0$, $\mathbb{E}[X_{n+1} \mid \mathscr{F}_n] = X_n$ a.s.,

(ii) a *submartingale* if for every $n \geqslant 0$, $\mathbb{E}[X_{n+1} \mid \mathscr{F}_n] \geqslant X_n$ a.s.,

(iii) a *supermartingale* if for every $n \geqslant 0$, $\mathbb{E}[X_{n+1} \mid \mathscr{F}_n] \leqslant X_n$ a.s.

If we think of X_n as our accumulated fortune when we make a sequence of bets, then a martingale represents a fair game in the sense that the conditional expectation of $X_{n+1} - X_n$, given our knowledge at the time when we make the $(n+1)^{\text{st}}$ bet (that is $\mathscr{F}_n$), is zero. A submartingale represents a favourable game and a supermartingale an unfavourable game. One could say that these terms are the wrong way round, i.e., they represent the point of view of 'the other player'. However, they are very well established by now, so it's too late to change them!

Here are some elementary properties.

Proposition 8.2. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space.*

(i) 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.

(ii) If $(X_n)_{n\geqslant 0}$ is a submartingale w.r.t. some filtration $(\mathscr{F}_n)_{n\geqslant 0}$ and is adapted to another smaller filtration $(\mathscr{G}_n)_{n\geqslant 0}$, $\mathscr{G}_n \subseteq \mathscr{F}_n$, $n \geqslant 0$, then it is also a submartingale with respect to $(\mathscr{G}_n)_{n\geqslant 0}$. In particular, $\mathbf{X}$ is a submartingale with respect to its natural filtration $(\mathscr{F}^X_n)_{n\geqslant 0}$.

(iii) If $(X_n)_{n\geqslant 0}$ is a submartingale and $n \geqslant m$ then

$$\mathbb{E}[X_n \mid \mathscr{F}_m] \geqslant X_m \text{ a.s.}$$

Proof. (i) is obvious.

For (ii) note that integrability is not affected by a change of filtration. Thus, by the tower property,

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

By definition, $\mathbf{X}$ is adapted to its own natural filtration and it is the smallest such filtration so $\mathscr{F}^X_n \subseteq \mathscr{F}_n$ and the above applies.

(iii). We fix m and prove the result by induction on n. The base case $n = m$ is obvious. For $n > m$ we have $\mathscr{F}_m \subseteq \mathscr{F}_n$ and using the submartingale property

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

so $\mathbb{E}[X_n \mid \mathscr{F}_m] \geqslant X_m$ a.s. follows by induction. □

Of course, part (iii) holds for a supermartingale with the inequalities reversed, and for a martingale with equality instead. Also, taking expectations in (iii), we see that for a submartingale $\mathbf{X}$ we have

$$\mathbb{E}[X_n] \geqslant \mathbb{E}[X_m] \geqslant \mathbb{E}[X_0], \quad n \geqslant m \geqslant 0,$$

with reversed inequalities for supermartingale and equalities for a martingale. Note however that the property $\mathbb{E}[X_{n+1} \mid \mathscr{F}_n] = X_n$ is much stronger than just $\mathbb{E}[X_{n+1}] = \mathbb{E}[X_n]$!

Remark. The collection of all martingales on a fixed filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$ is a vector space: if $(X_n)_{n\geqslant 0}$ and $(Y_n)_{n\geqslant 0}$ are martingales then so is $(aX_n + bY_n)_{n\geqslant 0}$ for any $a, b \in \mathbb{R}$.

Warning. There is a reason why we usually have a filtration in mind. In contrast to the above remark, it is easy (exercise!) 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 8.3 (Sums of independent random variables)**.** Suppose that $Y_1, Y_2, \ldots$ are independent integrable random variables on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and that $\mathbb{E}[Y_n] = 0$ for each n. Let $X_0 = 0$ and

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

Then $(X_n)_{n\geqslant 0}$ is a martingale with respect to the natural filtration given by

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

Indeed, $\mathbf{X}$ is adapted and integrable and

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

Note that we used basic properties of the conditional expectations, notably (iii) and (vi) in Proposition 6.5. These are very useful when dealing with martingales!

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

Definition 8.4 (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. $(\mathscr{F}_n)$. Here are two examples of martingale which are not sums of independent random variables.

Example 8.5. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $(Z_n)_{n\geqslant 1}$ be a sequence of independent integrable 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).

Example 8.6. Suppose that $Y_1, Y_2, \ldots$ are i.i.d. random variables on $(\Omega, \mathscr{F}, \mathbb{P})$ with $\mathbb{E}[\exp(Y_1)] = c < \infty$. Then

$$X_n = \exp(Y_1 + \ldots + Y_n)\, c^{-n}$$

is a martingale with respect to the natural filtration (exercise!).

Example 8.7. Let $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$ be a filtered probability space and X an integrable random variable. Then

$$X_n = \mathbb{E}[X \mid \mathscr{F}_n], \quad n \geqslant 0,$$

is an $(\mathscr{F}_n)_{n\geqslant 0}$-martingale. Indeed, X_n is certainly $\mathscr{F}_n$-measurable and integrable 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 note also that $\mathbf{X}$ is automatically UI by Theorem 6.11 and if $X_n \to X$ in probability then it already converges in $\mathscr{L}^1$ by Theorem 5.24. We shall later see that this is always the case and this convergence characterises such *closed* martingales.

Example 8.8. An integrable adapted process $\mathbf{X}$ which is increasing, $X_n \geqslant X_{n-1}$ a.s., $n \geqslant 1$, is a submartingale.

The above gave a trivial example of a submartingale. We now turn to more interesting examples and 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 8.9. *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 8.10. *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 8.11 (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 8.12 (Discrete stochastic integral or martingale transform)**.** *Let $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$ be a filtered probability space and $(Y_n)_{n\geqslant 0}$ a martingale. Suppose that $(V_n)_{n\geqslant 1}$ is predictable w.r.t. $(\mathscr{F}_n)$, and let $X_0 = 0$ and*

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

If each X_n is integrable then $(X_n)_{n\geqslant 0}$ is a martingale w.r.t. $(\mathscr{F}_n)$.

An important special case when all X_n are automatically integrable is when all V_n are bounded. The sequence $(X_n)_{n\geqslant 0}$ is called a *martingale transform* and is often denoted

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

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

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

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

□

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 k^{th} 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 k^{th} round is then $V_k(Y_k - Y_{k-1})$. More broadly, we will use the above result to retain the martingale property under stopping. This will be fundamental in what follows, see Theorem 8.16.

Proposition 8.13. *Let $(Y_n)_{n\geqslant 0}$ be a* supermartingale *on a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$, $(V_n)_{n\geqslant 1}$ a* non-negative *predictable process and let $X_0 = 0$ and*

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

If X_n is integrable, $n \geqslant 0$, then $\mathbf{X}$ is a supermartingale.

Proof. Exercise: imitate the proof of Theorem 8.12. □

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

Exercise 8.14. Let $(Y_i)_{i\geqslant 1}$ be independent random variables such that $\mathbb{E}[Y_i] = m_i$, $\mathrm{Var}(Y_i) = \sigma_i^2 < \infty$. Let

$$s_n^2 = \sum_{i=1}^n \sigma_i^2 = \mathrm{Var}\left(\sum_{i=1}^n Y_i\right).$$

Take $(\mathscr{F}_n)_{n\geqslant 0}$ to be the natural filtration generated by $(Y_n)_{n\geqslant 1}$. By Example 8.3,

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

is a martingale and so by Proposition 8.9, 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*. Show that

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

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

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

Theorem 8.15 (Doob's Decomposition Theorem)**.** *Let $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$ be a filtered probability space and $\mathbf{X} = (X_n)_{n\geqslant 0}$ an integrable adapted process. Then*

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

$$X_n = X_0 + M_n + A_n \tag{30}$$

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 $M_0 = 0 = A_0$.

(ii) 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

$$\mathbb{P}\left(M_n = \widetilde{M}_n, A_n = \widetilde{A}_n \text{ for all } n \geqslant 0\right) = 1.$$

(iii) $(X_n)_{n\geqslant 0}$ is a submartingale if and only if $(A_n)_{n\geqslant 0}$ in (30) 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. (i). Let

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

and

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

Then $M_n + A_n = \sum_{k=1}^n (X_k - X_{k-1}) = X_n - X_0$, so (30) holds. The k^{th} summand in A_n is $\mathscr{F}_{k-1}$-measurable, so A_n is $\mathscr{F}_{n-1}$-measurable, i.e., A is a predictable process. Also, as $\mathbf{X}$ is integrable so are $(M_n)_{n\geqslant 0}$ and $(A_n)_{n\geqslant 0}$. Finally, since

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

the process $(M_n)_{n\geqslant 0}$ is a martingale.

(ii) 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.

(iii) 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 above proof follows a clear logic and is, all in all, a relatively straightforward exercise. In contrast, the proof of the analogue result for martingales indexed with a continuous time parameter is a delicate affair!

Remark (The angle bracket process $\langle M\rangle$)**.** Let M be a martingale on $(\Omega,\mathscr{F},(\mathscr{F}_n)_{n\geqslant 0},\mathbb{P})$ with $\mathbb{E}[M_n^2]<\infty$ for each n. We then say that M is an L^2-martingale. Naturally, by Proposition 8.9, $(M_n^2)_{n\geqslant 0}$ is a *sub*martingale. Thus by Theorem 8.15 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 its continuous time equivalent, known as the quadratic variation process, will be used extensively in Part B Continuous Martingales and Stochastic Calculus course.

8.2 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 8.12. It is however so important that it is stated as a theorem!

Theorem 8.16 (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 $V_k=\mathbf{1}_{k\leqslant\tau}$, $k\geqslant 1$, 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 8.12 and Proposition 8.2. □

More generally, we have the following fundamental result.

Theorem 8.17 (Doob's Optional Sampling 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,* $\tau \leqslant \rho$*. Then*

$$\mathbb{E}[X_\rho \mid \mathscr{F}_\tau] = X_\tau \text{ a.s.} \tag{31}$$

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 (31) follows by simply checking the defining relationship for the conditional expectation since 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 (29) 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 8.16. 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 7.9. 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 8.13 instead of Theorem 8.12. □

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

Example 8.18. 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 7.6. It is easy to show, e.g., in analogy to Exercise 3.20, 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.

Corollary 8.19 (Variants of Doob's Optional Stopping Theorem)**.** *Let* $(M_n)_{n\geqslant 0}$ *be a martingale on a filtered probability space* $(\Omega, \mathscr{F}, (\mathscr{F}_n)_{n\geqslant 0}, \mathbb{P})$ *and* τ *an a.s. finite stopping time. Then*

$$\mathbb{E}[M_\tau \mathbf{1}_{\tau<\infty}] = \mathbb{E}[M_0]$$

if either of the following two conditions holds:

(i) $\{M_n : n \geqslant 0\}$ *is uniformly integrable;*

(ii) $\mathbb{E}[\tau] < \infty$ *and there exists* $L \in \mathbb{R}$ *such that*

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

Proof. (i) Let $K>0$. The process $(M_n=K)^+$, $n\geqslant 0$ is a submartingale by Proposition 8.9 and hence, by Theorem 8.17, we have $\mathbb{E}[(M_{\tau\wedge n}-K)^+]\leqslant \mathbb{E}[(M_n-K)^+]$. It follows, by Remark 5.21, that the family $(M_{\tau\wedge n}: n\geqslant 0)$ is Uniformly Integrable. We have $M_{\tau\wedge n}\to M_\tau \mathbf{1}_{\tau<\infty}$ a.s., since τ is a.s. finite, and hence also in $\mathscr{L}^1$ by Theorem 5.24. We conclude since, by Theorem 8.17, $\mathbb{E}[M_{\tau\wedge n}]=\mathbb{E}[M_0]$.

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

$$|M_{n\wedge\tau}| = |M_{n\wedge\tau}-M_{0\wedge\tau}| \leqslant \sum_{i=1}^{n}|M_{i\wedge\tau}-M_{(i-1)\wedge\tau}|\leqslant \sum_{i=1}^{\infty}|M_{i\wedge\tau}-M_{(i-1)\wedge\tau}| = \sum_{i=1}^{\infty}\mathbf{1}_{\tau\geqslant i}|M_i-M_{i-1}|. \quad (32)$$

Now

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

The result now follows, as above, by DCT with the function on the right hand side of (32) as the dominating function. □

We stated the Optional Stopping Theorem for martingales, but similar results are available for *sub/super*-martingales – just replace the equality in (31) by the appropriate inequality.

Note that if $|M_i-M_{i-1}|\leqslant L$ always holds, and $\mathbb{E}[\tau]<\infty$, then the second case applies. This is an important case of the Optional Stopping Theorem for applications. We give one such example.

Example 8.20. 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 $0(1-p_j)+p_j a/p_j=a$, the same as the stake.)

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

Gambler i bets £1 that $X_i=0$; 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 $\mathscr{F}_n=\sigma(X_1,\ldots,X_n)$. 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 8.19 is satisfied (with this L).

When X_τ is revealed each of the gamblers $1,2,\ldots,\tau$ 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 8.19 $\mathbb{E}[M_\tau]=0$, 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.

8.3 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 8.21 (Doob's maximal inequality)**.** *Let $(X_n)_{n\geqslant 0}$ be a submartingale on $(\Omega,\mathscr{F},(\mathscr{F}_n)_{n\geqslant0},\mathbb{P})$. 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,

$$\lambda\mathbb{P}\big[\max_{k\leqslant n}X_k\geqslant\lambda\big] \leqslant \mathbb{E}[X_n\mathbf{1}_{\{\max_{k\leqslant n}X_k\geqslant\lambda\}}] \leqslant \mathbb{E}[|X_n|]. \tag{33}$$

Proof. Let $\tau=\mathfrak{h}_{[\lambda,\infty)}=\inf\{n\geqslant0: X_n\geqslant\lambda\}$ and set $V_n=\mathbf{1}_{\{\tau\leqslant n-1\}}$, $n\geqslant1$. 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 8.13 to $-X$ and V we deduce that $(V\circ X)_0=0$,

$$(V\circ X)_n = \sum_{k=1}^n V_k(X_k-X_{k-1}) = X_{n\vee\tau}-X_\tau = (X_n-X_\tau)\mathbf{1}_{\{\tau\leqslant n\}}, \quad n\geqslant1,$$

is a submartingale. Further, $X_\tau\geqslant\lambda$ by definition so that $(X_\tau-\lambda)\mathbf{1}_{\{\tau\leqslant n\}}$, $n\geqslant0$, 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 8.22. *Let $p\geqslant1$ and $(M_n)_{n\geqslant0}$ be a martingale on a filtered probability space $(\Omega,\mathscr{F},(\mathscr{F}_n)_{n\geqslant0},\mathbb{P})$ with $M_n\in\mathscr{L}^p$ for all $n\geqslant0$. Then, for any $n\geqslant0$ 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 8.21 to $(|M_n|^p)_{n\geqslant 0}$ which is a submartingale by Proposition 8.9. □

Theorem 8.23 (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 8.21 and Lemma 5.15. □

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 8.24. *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*

$$\lambda\mathbb{P}(\max_{k\leqslant n}|X_k|\geqslant\lambda)\leqslant\mathbb{E}[X_0]+2\mathbb{E}[X_n^-],\quad\forall\lambda,n\geqslant 0. \tag{34}$$

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 8.21 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. □

8.4 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 8.13 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 8.25 (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 8.26 (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 $n>\rho_k$.
Notice that

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

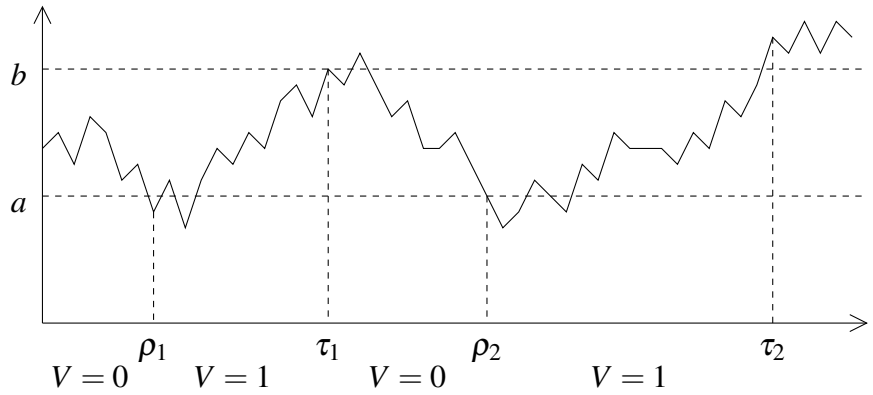


Figure 2: Illustration of the sequence of stopping times introduced in Definition 8.25.

So $(V_n)_{n\geqslant 1}$ is non-negative and *predictable* so, by Proposition 8.13, $(V\circ X)_n$, $n\geqslant 0$ is a supermartingale. We write $U_n = U_n([a,b],\mathbf{X})$ and compute directly:

$$\begin{aligned}(V\circ X)_n &= \sum_{k=1}^{n} V_k(X_k - X_{k-1}) \\ &= \sum_{i=1}^{U_n}(X_{\tau_i} - X_{\rho_i}) + \mathbf{1}_{\{\rho_{U_n+1}<n\}}(X_n - X_{\rho_{U_n+1}}) \qquad (35)\\ &\geqslant (b-a)U_n - (X_n - a)^-. \qquad (36)\end{aligned}$$

For the last step, note that if indicator function in (35) 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 (36),

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

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 8.27. *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 8.27 to show that the *random* sequence (X_n) converges. This is our next result.

Definition 8.28. 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 8.29 (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 8.27 $(X_n)_{n\geqslant 0}$ 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 8.31 below demonstrates.

Corollary 8.30. *If $(X_n)_{n\geqslant 0}$ is a non-negative supermartingale, then $X_\infty = \lim_{n\to\infty} X_n$ exists a.s.*

Proof. Since $\mathbb{E}[|X_n|] = \mathbb{E}[X_n] \leqslant \mathbb{E}[X_0]$ we may apply Theorem 8.29. □

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 8.31 (Galton–Watson branching process)**.** Recall Definition 0.1: 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 8.30, 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 $n \geqslant \tau$. In particular, M_n does *not* converge to M_∞ in $\mathscr{L}^1$ by Lemma 4.14, and hence also not in any other $\mathscr{L}^p$ for $p > 1$ by Lemma 5.13.

What is happening for our subcritical branching process is that although for large n, M_n is very likely to be zero, if it is *not* zero then it is very *big* with sufficiently high probability that $\mathbb{E}[M_n]$ is constant and does not converge to 0. This mirrors what we saw with sequences in Example 5.3. Finally note that, by Theorem 5.24, we can also conclude that $\{M_n : n \geqslant 0\}$ is *not* Uniformly Integrable.

8.5 Uniformly integrable martingales

We have done most of the work in §5.4. It remains to use it in conjunction with what we already know about martingales. 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 8.32. *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) **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.

Further, under these conditions, if $M_\infty \in \mathscr{L}^p$ for $p > 1$ then the convergence $M_n \to M_\infty$ also holds in $\mathscr{L}^p$.

Proof. $(i) \implies (ii)$: **M** is UI so in particular, by Proposition 5.22, bounded in $\mathscr{L}^1$ and hence, by Doob's Forward Convergence Theorem (Theorem 8.29) 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 5.24. Each M_n is $\mathscr{F}_\infty$-measurable and hence so is M_∞ by Proposition 1.24.

$(ii) \implies (iii)$: 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 (22) 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

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

it follows that

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

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 6.11.

The last assertion follows instantly from the Dominated Convergence Theorem and Theorem 8.33 below. □

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

Theorem 8.33. *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 $\tau \leqslant \rho$*

$$\mathbb{E}[M_\rho \mid \mathscr{F}_\tau] = M_\tau \text{ a.s.} \tag{37}$$

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

$$\lambda \mathbb{P}[M_\infty^* \geqslant \lambda] \leqslant \mathbb{E}[|M_\infty| \mathbf{1}_{\{M_\infty^* \geqslant \lambda\}}], \quad \lambda \geqslant 0. \tag{38}$$

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

$$\|M_\infty\|_p \leqslant \|M_\infty^*\|_p \leqslant q \|M_\infty\|_p \tag{39}$$

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

Deep Dive

Proof. First note that if τ is bounded, $\tau \leqslant n$ and $\rho = \infty$ then by Theorem 8.17

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

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 5.24 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 6.11) 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 (37).

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 8.21, with $M_n^* = \max_{k\leqslant n} |M_k|$, we have

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

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 (38) holds as required. Suppose now that $M_\infty \in \mathscr{L}^p$ for some $p > 1$. Then Doob's L^p inequality (39) follows by Lemma 5.15. 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. □

9 Some applications of the martingale theory

9.1 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 8.32 and makes our life easier.

Definition 9.1. 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 6.11 that $(M_{-n})_{n \geqslant 0}$ is *automatically* uniformly integrable.

Doob's Upcrossing Lemma (Lemma 8.26), 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

$$\mathbb{E}[U_m([a,b], \mathbf{M})] \leqslant \frac{\mathbb{E}[(M_0 - a)^-]}{b-a}. \tag{40}$$

Mimicking the proof of Doob's Forward Convergence Theorem (Theorem 8.29), 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 8.27 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 8.32 gives the following result.

Theorem 9.2. *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 9.3 (Kolmogorov's Strong Law of Large Numbers)**.** *Let $(X_n)_{n\geqslant 1}$ be a sequence of i.i.d. random variables each of which is integrable and has mean m, and set*

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

Then

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

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 9.2, 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 3.13) it is a.s. constant, so $M_{-\infty} = m$ a.s. □

Deep Dive

9.2 Exchangeability and the ballot theorem

The material in §9.2 is not part of the "examinable syllabus". You won't be asked to reproduce these results directly. However, just like many of the problem sheet questions, the methods help to develop your intuition for the ideas of the course.

In our proof of the Strong Law of Large Numbers we used symmetry in a key way. There it followed from independence of our random variables, but in general a weaker condition suffices.

Definition 9.4 (Exchangeability)**.** The random variables $X_1, \ldots, X_n$ are said to be *exchangeable* if the vector $(X_{i_1}, \ldots, X_{i_n})$ has the same probability distribution for every permutation $i_1, \ldots, i_n$ of $1, \ldots, n$.

Example 9.5. Let $X_1, \ldots, X_n$ be the results of n successive samples *without replacement* from a pool of at least n values (some of which may be the same). Then the random variables $X_1, \ldots, X_n$ are exchangeable but *not* independent.

It turns out that we can use the construction in the proof of the Strong Law of Large Numbers to manufacture a finite martingale from a finite collection of exchangeable random variables. Suppose that $X_1, \ldots, X_n$

are exchangeable and integrable, and set $S_j = \sum_{i=1}^{j} X_i$. Let

$$Z_j = \mathbb{E}[X_1 \mid \sigma(S_{n+1-j}, \ldots, S_{n-1}, S_n)], \quad j = 1, 2, \ldots n.$$

Note that Z_j is defined by conditioning on the last j sums; since we condition on more as j increases, $(Z_j)_{j=1}^{n}$ is certainly a martingale. Now

$$\begin{aligned} S_{n+1-j} &= \mathbb{E}[S_{n+1-j} \mid \sigma(S_{n+1-j}, \ldots, S_n)] \\ &= \sum_{i=1}^{n+j-1} \mathbb{E}[X_i \mid \sigma(S_{n+1-j}, \ldots, S_n)] \\ &= (n+1-j)\mathbb{E}[X_1 \mid \sigma(S_{n+1-j}, \ldots, S_n)] \quad \text{(by exchangeability)} \\ &= (n+1-j)Z_j, \end{aligned}$$

so $Z_j = S_{n+1-j}/(n+1-j)$.

Definition 9.6. The martingale

$$Z_j = \frac{S_{n+1-j}}{n+1-j}, \quad j = 1, 2, \ldots, n,$$

is sometimes called a *Doob backward martingale*.

Example 9.7 (The ballot problem)**.** In an election between candidates A and B, candidate A receives n votes and candidate B receives m votes, where $n > m$. Assuming that in the count of votes all orderings are equally likely, what is the probability that A is always ahead of B during the count?

Solution:

Let $X_i = 1$ if the ith vote counted is for A and -1 if the ith vote counted is for B, and let $S_k = \sum_{i=1}^{k} X_i$. Because all orderings of the $n+m$ votes are equally likely, $X_1, \ldots, X_{n+m}$ are exchangeable, so

$$Z_j = \frac{S_{n+m+1-j}}{n+m+1-j}, \quad j = 1, 2, \ldots, n+m,$$

is a Doob backward martingale.

Because

$$Z_1 = \frac{S_{n+m}}{n+m} = \frac{n-m}{n+m},$$

the mean of this martingale is $(n-m)/(n+m)$.

Because $n > m$, either (i) A is always ahead in the count, or (ii) there is a tie at some point. Case (ii) happens if and only if some $S_j = 0$, i.e., if and only if some $Z_j = 0$.

Define the bounded stopping time τ by

$$\tau = \min\{j \geqslant 1 : Z_j = 0 \text{ or } j = n+m\}.$$

In case (i), $Z_\tau = Z_{n+m} = X_1 = 1$. (If A is always ahead, he must receive the first vote.) Clearly, in case (ii), $Z_\tau = 0$, so

$$Z_\tau = \begin{cases} 1 & \text{if } A \text{ is always ahead,} \\ 0 & \text{otherwise.} \end{cases}$$

By Theorem 8.17, $\mathbb{E}[Z_\tau] = (n-m)/(n+m)$ and so

$$\mathbb{P}[A \text{ is always ahead}] = \frac{n-m}{n+m}.$$

□

9.3 Azuma-Hoeffding inequality and concentration of Lipschitz functions

The material in §9.3 is not part of the "examinable syllabus". You won't be asked to reproduce any of these results directly. However, the methods involved are very good illustrations of ideas from earlier in the course: particularly the Doob martingale ideas involved in Theorem 9.12 and its applications.

By applying Markov's inequality to the moment generating function, we can get better bounds than we get from the mean and variance alone.

Lemma 9.8. *(i) Let Y be a random variable with mean 0, taking values in $[-c,c]$. Then*

$$\mathbb{E}[e^{\theta Y}] \leqslant \exp\left(\frac{1}{2}\theta^2 c^2\right).$$

(ii) Let $\mathscr{G}$ be a σ-algebra, and Y be a random variable with $\mathbb{E}[Y|\mathscr{G}] = 0$ a.s. and $Y \in [-c,c]$ a.s. Then

$$\mathbb{E}[e^{\theta Y} \mid \mathscr{G}] \leqslant \exp\left(\frac{1}{2}\theta^2 c^2\right) \text{ a.s.}$$

Proof. Let $f(y) = e^{\theta y}$. Since f is convex,

$$f(y) \leqslant \frac{c-y}{2c} f(-c) + \frac{c+y}{2c} f(c)$$

for all $y \in [-c,c]$. Then taking expectations,

$$\begin{aligned}\mathbb{E}[f(Y)] &\leqslant \mathbb{E}\left[\frac{c-Y}{2c} f(-c) + \frac{c+Y}{2c} f(c)\right] \\ &= \frac{1}{2} f(-c) + \frac{1}{2} f(c) \\ &= \frac{e^{-\theta c} + e^{\theta c}}{2}.\end{aligned}$$

Now, comparing Taylor expansions term by term,

$$\frac{e^{-\theta c} + e^{\theta c}}{2} = \sum_{n=0}^{\infty} \frac{(\theta c)^{2n}}{(2n)!} \leqslant \sum_{n=0}^{\infty} \frac{(\theta c)^{2n}}{2^n n!} = \exp\left(\frac{1}{2}\theta^2 c^2\right).$$

giving part (i).

For the conditional version of the statement, consider any $G \in \mathscr{G}$ with $\mathbb{P}[G] > 0$. Then $\mathbb{E}[Y \mathbf{1}_G] = 0$, so $\mathbb{E}[Y \mid G] = 0$. Applying part (i) with probability measure $\mathbb{P}[. \mid G]$, we obtain $\mathbb{E}[e^{\theta Y} \mid G] \leqslant \exp\left(\frac{1}{2}\theta^2 c^2\right)$.

Now consider the G-measurable set $G := \{\omega : \mathbb{E}[e^{\theta Y}|\mathscr{G}](\omega) > \exp\left(\frac{1}{2}\theta^2 c^2\right)\}$. If this set has positive probability, it contradicts the previous paragraph. So indeed $\mathbb{E}[e^{\theta Y} \mid \mathscr{G}] \leqslant \exp\left(\frac{1}{2}\theta^2 c^2\right)$ a.s. as required. □

Lemma 9.9. *Suppose M is a martingale with $M_0 = 0$ and $|M_n - M_{n-1}| \leqslant c$ a.s. for all n. Then*

$$\mathbb{E}\left[e^{\theta M_n}\right] \leqslant \exp\left(\frac{1}{2}\theta^2 c^2 n\right).$$

Proof. Let $W_n = e^{\theta M_n}$, so that W_n is non-negative and $W_n = W_{n-1}e^{\theta(M_n - M_{n-1})}$.

Then applying Lemma 9.8(ii) with $Y = M_n - M_{n-1}$ and $\mathscr{G} = \mathscr{F}_{n=1}$,

$$\begin{aligned}\mathbb{E}(W_n \mid \mathscr{F}_{n-1}) &= W_{n-1}\mathbb{E}\left[e^{\theta(M_n - M_{n-1})} \mid \mathscr{F}_{n-1}\right] \\ &\leqslant W_{n-1}\exp\left(\frac{1}{2}\theta^2 c^2\right) \text{ a.s.}\end{aligned}$$

Taking expectations we obtain $\mathbb{E}[W_n] \leqslant \exp\left(\frac{1}{2}\theta^2 c^2\right)\mathbb{E}[W_{n-1}]$ and the result follows by induction.

□

Theorem 9.10 (Simple version of the Azuma-Hoeffding inequality)**.** *Suppose M is a martingale with $M_0 = 0$ and $|M_n - M_{n-1}| \leqslant c$ a.s. for all n. Then*

$$\mathbb{P}(M_n \geqslant a) \leqslant \exp\left(-\frac{1}{2}\frac{a^2}{c^2 n}\right),$$

and

$$\mathbb{P}(|M_n| \geqslant a) \leqslant 2\exp\left(-\frac{1}{2}\frac{a^2}{c^2 n}\right).$$

Proof.

$$\begin{aligned}\mathbb{P}(M_n \geqslant a) &\leqslant \mathbb{P}\left(e^{\theta M_n} \leqslant e^{\theta a}\right) \\ &\leqslant e^{-\theta a}\exp\left(\frac{1}{2}\theta^2 c^2\right)\end{aligned}$$

using Markov's inequality. Now we are free to optimise over θ. The RHS is minimised when $\theta = a/(c^2 n)$, giving the required bound.

The same argument applies replacing M by the martingale $-M$. Summing the two bounds then gives the bound for $|M|$. □

We now introduce the idea of *discrete Lipschitz functions*.

Definition 9.11. Let h be a function of n variables. The function h is said to be c-Lipschitz, where $c > 0$, if changing the value of any one coordinate causes the value of h to change by at most c. That is, whenever $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ differ in at most one coordinate, then $|h(\mathbf{x}) - h(\mathbf{y})| \leqslant c$.

Theorem 9.12 (Concentration of discrete Lipschitz functions)**.** *Suppose h is a c-Lipschitz function, and $X_1, \ldots, X_n$ are independent random variables. Then*

$$\mathbb{P}\left(|h(X_1, \ldots, X_n) - \mathbb{E}[h(X_1, \ldots, X_n)]| \geqslant a\right) \leqslant 2\exp\left(-\frac{1}{2}\frac{a^2}{c^2 n}\right).$$

Proof. The proof is based on the idea of the Doob martingale. We reveal information about the underlying random variables $X_1,\dots,X_n$ one step at a time, gradually acquiring a more precise idea of the value $h(X_1,\dots,X_n)$.

For $0 \leqslant k \leqslant n$, let $\mathscr{F}_k = \sigma(X_1,\dots,X_k)$, and let

$$M_k = \mathbb{E}[h(X_1,\dots,X_n) \mid \mathscr{F}_k] - \mathbb{E}[h(X_1,\dots,X_n)].$$

Then $M_0 = 0$, and $M_n = h(X_1,\dots,X_n) - \mathbb{E}[h(X_1,\dots,X_n)]$.

We claim $|M_{k+1} - M_k| \leqslant c$ a.s. To show this, let $\widehat{X}_{k+1}$ be a random variable with the same distribution as X_{k+1}, which is independent of $X_1,\dots,X_n$.

Then

$$\begin{aligned}
&\mathbb{E}[h(X_1,\dots,X_k,X_{k+1},\dots,X_n) \mid \mathscr{F}_k] \\
&= \mathbb{E}[h(X_1,\dots,X_k,\widehat{X}_{k+1},\dots,X_n) \mid \mathscr{F}_k] \\
&= \mathbb{E}[h(X_1,\dots,X_k,\widehat{X}_{k+1},\dots,X_n) \mid \mathscr{F}_{k+1}].
\end{aligned}$$

This gives

$$M_{k+1} - M_k = \mathbb{E}[h(X_1,\dots,X_k,\widehat{X}_{k+1},\dots,X_n) - h(X_1,\dots,X_k,X_{k+1},\dots,X_n) \mid \mathscr{F}_{k+1}].$$

But the difference between the two values of h inside the conditional expectation on the RHS is in $[-c,c]$, so we obtain $|M_{k+1} - M_k| \leqslant c$ a.s. as required. Now the required estimate for M_n follows from the Azuma-Hoeffding bound (Theorem 9.10). □

The examples below of the application of Theorem 9.12 show that martingale methods can be applied to problems far away from what one might think of as "stochastic process theory".

Example 9.13 (Longest common subsequence)**.** Let $\mathbf{X} = (X_1, X_2, \dots, X_m)$ and $\mathbf{Y} = (Y_1, Y_2, \dots, Y_m)$ be two independent sequences, each with independent entries.

Let L_m be the length of the longest sequence which is a subsequence (not necessarily consecutive) of both sequences.

For example, if $m = 12$ and $\mathbf{X}$ ="CAGGGTAGTAAG" and $\mathbf{Y}$ ="CGTGTGAAAACT" then both $\mathbf{X}$ and $\mathbf{Y}$ contain the substring "CGGTAAA", and $L_m = 7$.

Changing a single entry can't change the length of the longest common subsequence by more than 1. We can apply Theorem 9.12 with $n = 2m$ and $c = 1$, to get

$$\mathbb{P}(|L_m - \mathbb{E}[L_m]| \geqslant a) \leqslant 2\exp\left(-\frac{a^2}{4m}\right).$$

We obtain that for large m, "typical fluctuations" of L_m around its mean are on the scale at most $\sqrt{m}$.

Note that we didn't require the sequences $\mathbf{X}$ and $\mathbf{Y}$ to have the same distribution, or for the entries of each sequence to be identically distributed.

As suggested by the choice of strings above, longest common subsequence problems arise for example in computational biology, involving the comparison of DNA strings (which evolve via mutation, insertion or deletion of individual nucleotides).

Example 9.14 (Minimum-length matching)**.** Suppose there are m red points in the box $[0,1]^2 \subset \mathbb{R}^2$, with positions $R_1,\dots,R_m$, and m blue points with positions $B_1,\dots,B_m$.

Let X be the length of the minimal-length *matching*, which joins pairs consisting of one blue and one red point. That is,

$$X_m = \min \sum_{k=1}^{m} \|R_k - B_{i_k}\|,$$

where the minimum is taken over all permutations $i_1, i_2, \ldots, i_m$ of $1, 2, \ldots, m$, and $\|r-b\|$ denotes Euclidean distance between r and b.

Alternatively let Y be the length of the minimal-length *alternating tour*, a path which visits all $2m$ points, alternating between red and blue, and returning to its starting point:

$$Y_m = \min \left\{ \sum_{k=1}^{m} \|R_{i_k} - B_{j_k}\| + \sum_{k=1}^{m-1} \|B_{j_k} - R_{i_{k+1}}\| + \|B_{j_m} - R_{i_1}\| \right\},$$

where now the minimum is over all pairs of permutations $i_1, i_2, \ldots, i_m$ and $j_1, j_2, \ldots, j_m$ of $1, 2, \ldots, m$.

Moving a single point cannot change X_m by more than $\sqrt{2}$, and cannot change Y_m by more than $2\sqrt{2}$. If the positions of the points are independent, then applying Theorem 9.12 with $n = 2m$ and the appropriate value of c, we obtain

$$\mathbb{P}(|X_m - \mathbb{E}[X_m]| \geqslant a) \leqslant 2\exp\left(-\frac{a^2}{8m}\right)$$
$$\mathbb{P}(|Y_m - \mathbb{E}[Y_m]| \geqslant a) \leqslant 2\exp\left(-\frac{a^2}{32m}\right).$$

Again this gives concentration of X_m and Y_m around their means on the scale of $\sqrt{m}$. This may be a poor bound; for example if all the points are i.i.d. uniform on the box $[0,1]^2$, then in fact the means themselves grow like $\sqrt{m}$ as $m \to \infty$. However, we didn't assume identical distribution. For example we might have red points uniform on the left half $[0, 1/2] \times [0,1]$, and blue points uniform on the right half $[1/2, 1] \times [0,1]$, in which case the means grow linearly in m, and the $O(\sqrt{m})$ fluctuation bound is more interesting.

Example 9.15 (Chromatic number of a random graph)**.** The Erdös-Rényi random graph model $G(N,p)$ consists of a graph with N vertices, in which each edge (out of the $\binom{N}{2}$ possible edges) appears independently with probability p. If $p = 1/2$, then the graph is uniformly distributed over all possible graphs with N vertices.

The *chromatic number* $\chi(G)$ of a graph G is the minimal number of colours needed to colour the vertices of G so that any two adjacent vertices have different colours.

Consider applying Theorem 9.12 to the chromatic number $\chi(G)$ of a random graph $G \sim G(N, 1/2)$. We could write $\chi(G)$ as a function of $\binom{N}{2}$ independent Bernoulli random variables, each one encoding the presence or absence of a given edge. Adding or removing a single edge cannot change the chromatic number by more than 1. This would give us a fluctuation bound on $\chi(G)$ on the order of N as $N \to \infty$. However, for large N this is an extremely poor, in fact trivial, result, since $\chi(G)$ itself is known to be on the order of $N/\log(N)$.

We can do much better. For $2 \leqslant k \leqslant N$, let X_k consist of a collection of $k-1$ Bernoulli random variables, encoding the presence or absence of the $k-1$ edges $\{1,k\}, \{2,k\}, \ldots, \{k-1,k\}$. It's still the case that $X_2, \ldots, X_N$ are independent. All the information in X_k concerns edges that intersect the vertex k; changing the status of any subset of these edges can only change the chromatic number by at most 1 (consider recolouring vertex k as necessary). The Doob martingale from the proof of Theorem 9.12 involves revealing information about the graph vertex by vertex, rather than edge by edge, and is called the *vertex exposure martingale*.

Applying the theorem with $n = N-1$ and $c = 1$, we obtain

$$\mathbb{P}(|\chi(G) - \mathbb{E}[\chi(G)]| \geqslant a) \leqslant 2\exp\left(-\frac{a^2}{2(N-1)}\right),$$

giving a concentration bound on the scale of $\sqrt{N}$ for large N.

9.4 The Law of the Iterated Logarithm

9.5 Likelihood Ratio and Statistics

9.6 Radon-Nikodym Theorem

Index