PROBABILITY, MEASURE AND MARTINGALES

Michaelmas Term 2026

Lecturer: Harald Oberhauser

Version of November 20, 2026

0 Introduction

These notes accompany my lectures on *Probability, Measure and Martingales* (B8.1). Much of the material is adapted and extended from two sets of notes generously shared with me by Jan Obłój and James Martin. I express my deep gratitude to Jan and James for providing me with such excellent lecture notes; naturally, all errors are mine. Jan's and James' notes are, in turn, heavily based on older lecture notes by Alison Etheridge, Oliver Riordan, and Zhongmin Qian. While the lecturers (as well as the name) of this course have changed over the years, the syllabus did not change significantly. However, there are a few important differences in what each lecturer emphasized. My goal is to provide a rigorous foundation of the theory of stochastic processes in measure theory that still leaves us time to develop intuition and explore why the theory of stochastic processes is so fascinating and powerful; both mathematically and in terms of real world applications.

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

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

0.1 Background

In the last fifty years probability theory has emerged both as a core mathematical discipline, sitting alongside geometry, algebra and analysis, and as a fundamental way of thinking about the world. It provides the rigorous mathematical framework necessary for modelling and understanding the inherent randomness in the world around us. It has become an indispensable tool in many disciplines – physics, neuroscience, genetics, communication networks, finance, statistics, and machine learning to name a few. 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.
After the lecture notes, the nearest thing to a course text is

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

Also highly recommended are:

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

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

Deep Dive

Anything marked as a *Deep Dive* covers material that is non-examinable and not necessary for the course. It is only intended for those who are interested and eager to know a bit more; often indicating links with other courses, parts of mathematics, or historical background. More deep dives may appear as I revise the notes. The depth of deep dives may vary considerably from one dive to another and these can be more opinionated and less objective than the rest of the course.

Remark. Unlike Deep Dives, the content of remarks often adds important context or information to the main text and is examinable.

Contents

0 Introduction **1**
0.1 Background . . . 1
0.2 Example 1: Simple Symmetric Random Walk . . . 4
0.3 Example 2: Mathematical Finance . . . 5
0.4 Example 3: The Galton–Watson Branching Process . . . 6

1 Measure Spaces **9**
1.1 σ-Algebras and Measures . . . 10
1.2 More on Distribution Functions . . . 15

2 Integration **17**
2.1 Measurable functions and the Definition of the Integral . . . 17
2.2 The Case of Probability Measures: Recall of some Terminology . . . 22
2.3 The Convergence Theorems . . . 23
2.4 Product Spaces . . . 24
2.5 Integration on Product Spaces: Fubini and Tonelli . . . 26
2.6 Independence . . . 27
2.7 Tail Events: The Kolmogorov $0-1$ Law . . . 29
2.8 Eventual and Infinitely Often Events: The Borel–Cantelli Lemmas . . . 31

3 Modes of Convergence **33**
3.1 $\mathscr{L}^p$ spaces . . . 37

4 Conditional Expectation **45**
4.1 Existence of Conditional Expectation via Orthogonal Projection in L^2 . . . 52

5 Filtrations and Stopping Times **55**
5.1 Stopping Times . . . 55
5.2 The Stopped Process X_τ . . . 58

6 Martingales **58**
6.1 Basic Properties . . . 60
6.2 Preserving (Sub-)martingale Structure . . . 61

0.2 Example 1: Simple Symmetric Random Walk

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

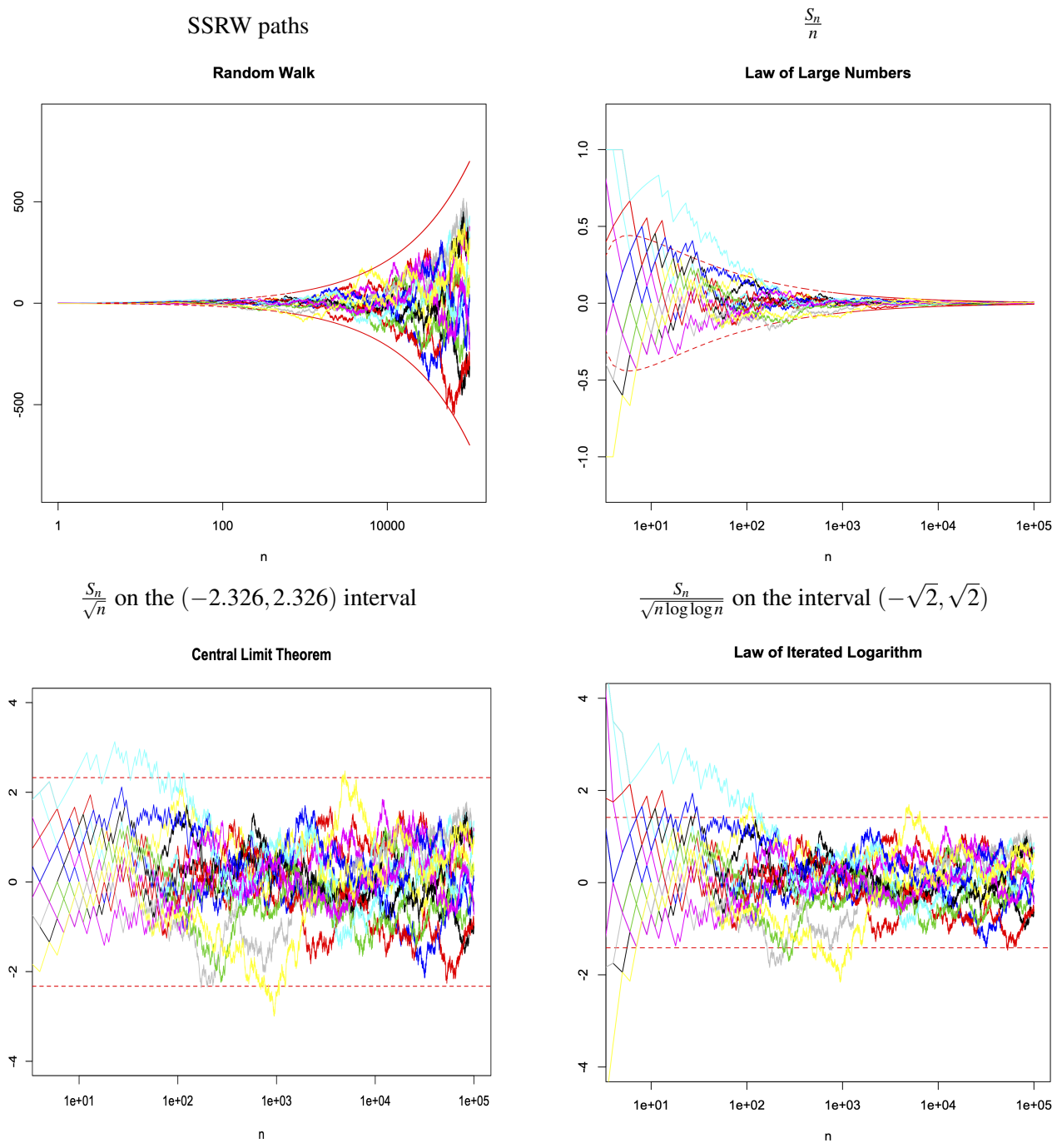


Figure 1: Limiting behaviour of a SSRW

0.3 Example 2: Mathematical Finance

Suppose $(S_n)_{n\geqslant 0}$ is sequence of random variables modelling the price process of some risky asset, i.e., S_n is the share price at time n. A trader is buying and selling the stock. At time n, they have wealth V_n and decide to buy/sell $H_n = H_n(S_0, S_1, \dots, S_n)$ shares. At time $n+1$, they will have $H_n S_{n+1}$ in shares while their remaining

capital/debt grew at rate r:

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

If we introduce discounted quantities

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

then the above is re-written as

$$\tilde{V}_{n+1} = H_n(\tilde{S}_{n+1} - \tilde{S}_n) + \tilde{V}_n = \ldots = V_0 + \sum_{t=0}^{n} H_t(\tilde{S}_{t+1} - \tilde{S}_t), \tag{2}$$

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

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

Proposition 0.1 (Binomial Model pricing. Non-examinable.)**.** *Fix a finite horizon $N \geqslant 1$ and suppose $S_0 > 0$ is deterministic. Assume that $0 < 1-d < 1+r < 1+u$ and that, for $0 \leqslant n < N$, $S_{n+1} \in \{(1+u)S_n, (1-d)S_n\}$ almost surely, with both moves having strictly positive conditional probability given the price history at every node. Then every payoff $f(S_0, \ldots, S_N)$, for $f : \mathbb{R}^{N+1} \to \mathbb{R}$, can be replicated by a self-financing trading strategy with some initial capital V_0. On $\sigma(S_0, \ldots, S_N)$ there is a unique probability measure $\mathbb{Q}$ equivalent to $\mathbb{P}$ under which the discounted price process $(\tilde{S}_n)_{0 \leqslant n \leqslant N}$ is a martingale with respect to its natural filtration, and $V_0 = (1+r)^{-N}\mathbb{E}_{\mathbb{Q}}[f(S_0, \ldots, S_N)]$.*

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

0.4 Example 3: The Galton–Watson Branching Process

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

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

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

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

Definition 0.2 (Galton–Watson branching process)**.** Let $(X_{n,r})_{n,r\geqslant 1}$ be an infinite array of independent identically distributed random variables, each with the same distribution as X, where

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

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 $f(\theta)=\mathbb{E}[\theta^X]=\sum_{k=0}^{\infty}p_k\theta^k$, say for $\theta\in[0,1]$ to avoid convergence issues.

Claim 0.3. *Let $f_n(\theta)=\mathbb{E}[\theta^{Z_n}]$. Then f_n is the n-fold composition of f with itself (where by convention a 0-fold composition is the identity).*

'Proof'

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

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

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

$$\mathbb{E}\left[\theta^{Z_n}\middle|Z_{n-1}\right]=f(\theta)^{Z_{n-1}}. \tag{3}$$

This is our first example of a *conditional expectation*; we'll give a rigorous definition of conditional expectations later. Notice that the right hand side of (3) is a *random variable*. Now

$$\begin{aligned}f_n(\theta)=\mathbb{E}\left[\theta^{Z_n}\right] &= \mathbb{E}\left[\mathbb{E}\left[\theta^{Z_n}\middle|Z_{n-1}\right]\right] \qquad (4)\\ &= \mathbb{E}\left[f(\theta)^{Z_{n-1}}\right]\\ &= f_{n-1}\left(f(\theta)\right),\end{aligned}$$

and the claim follows by induction. □

In (4) we have used what is called the *tower property* of conditional expectations. In this example you can make all this work with the Partition Theorem of Prelims (because the events $\{Z_n=k\}$ form a countable partition of the sample space). In the general theory that follows, we'll see how to replace the Partition Theorem when the sample space is more complicated, for example when considering continuous random variables.

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

Claim 0.4. *Let $q = \mathbb{P}[Z_n = 0 \text{ for some } n]$. Then q is the smallest root in $[0,1]$ of the equation $\theta = f(\theta)$. In particular, assuming $p_1 = \mathbb{P}[X = 1] < 1$,*

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

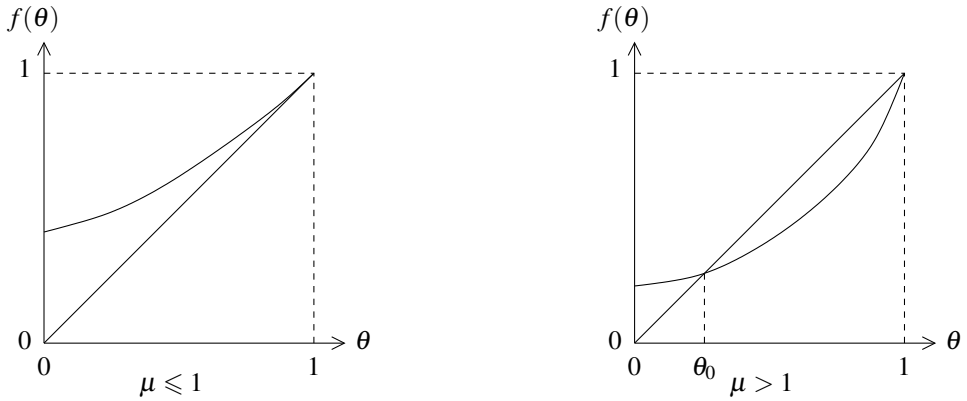
'Proof'

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

$$q = \lim_{n\to\infty} q_n = \lim_{n\to\infty} f_n(0). \tag{5}$$

Since $f_{n+1}(0) = f(f_n(0))$ and f is continuous, (5) 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

$$\mathbb{E}\left[Z_{n+1} \mid Z_n = k\right] = \mathbb{E}\left[X_{n+1,1} + \cdots + X_{n+1,k}\right] = km \quad \text{(linearity of expectation)}. \tag{6}$$

In other words $\mathbb{E}[Z_{n+1} \mid Z_n] = mZ_n$ (another conditional expectation). Now assume $m > 0$ and set

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

Then

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

In fact, more is true:

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

Again we encounter a martingale!

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

$$0 = \mathbb{E}[M_\infty] \neq \lim_{n\to\infty} \mathbb{E}[M_n] = 1. \tag{7}$$

We're going to have to be careful in passing to limits, just as we discovered in Part A Integration. Indeed (7) may remind you of Fatou's Lemma from Part A.

To make the preceding arguments rigorous and develop martingale theory, we need three ingredients. First, a language for events and probability measures; second, a general theory of integration and expectation; and third, a rigorous notion of conditioning.

Deep Dive

Why are martingales such a big deal and form the central part in most first courses about stochastic processes? One answer comes from first year analysis: monotone sequences of real numbers play a big role since they have many nice properties, e.g. boundedness implies existence of a limit. Moreover, these nice properties can be used to understand more general sequences, e.g. by bounding them above and below by monotone sequences. Among monotone sequences, constant sequences are the simplest ones. The defining property of a martingale, (1), is the natural version of a constant sequence for a probabilist since it simply says "given everything we know up to time $n-1$, the next value is *on average* the same as the current value". Similarly, one can replace the equal sign in (1) by a lesser or greater sign and arrive at analogues of monotone increasing and decreasing sequences which we will later introduce in this course as so-called sub- and supermartingales. I hope this gives some intuition why it is promising to look at these types of stochastic processes before trying to understand more complicated ones. We will see, that as opposed to constant sequences in classic analysis which are admittedly fairly boring, martingales already have a rich and interesting structure. So why do we use the term martingale? The reason is purely historical: it originates in eighteenth-century France, where la martingale referred to a gambling strategy. In this strategy, a player repeatedly doubled the stake after each loss, so that an eventual win would recover all previous losses and yield a profit equal to the original stake; note that the wealth evolution can be written as what refer to as martingale transform above in (2). The term martingale itself is of uncertain origin, possibly connected with the Provençal town of Martigues and or also expressions suggesting reckless behaviour. Subsequently, term martingale was then picked in probability theory through the work of Jean Ville in the 1930s and used to ultimately describe the much broader mathematical class of stochastic processes that we now call martingales.

1 Measure Spaces

We begin by revisiting some definitions and results that you encountered in Part A Integration (and, although they were not emphasized there, in Prelims Probability). The idea is that we want to be able to assign a 'mass' or 'size' to subsets of a space in a consistent way. For us, ultimately these subsets will be 'events' or 'collections of outcomes' (subsets of a probability sample space Ω) and the 'mass' will be a probability (giving values in $[0,1]$ to such 'events' which we regard as quantifying of how likely event is to occur) but right this is not important.

Given a space, the first natural attempt is to try to assign 'mass' to all subsets of that space, that is the powerset. Unfortunately, this is already impossible for simple sets like $[0,1]$.

Theorem 1.1 (Non-examinable)**.** *There is no function*

$$m : \mathscr{P}([0,1]) \to [0,\infty]$$

such that:

1. $m([0,1]) = 1$.

2. *If $A_1, A_2, \ldots \subseteq [0,1]$ are pairwise disjoint, then*

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

3. *For every $A \subseteq [0,1]$ and every $t \in \mathbb{R}$ such that $A+t \subseteq [0,1]$,*

$$m(A+t) = m(A), \text{ where } A+t := \{x+t : x \in A\}.$$

The proof is not hard (try it yourself before you look it up[2]) and works by constructing extremely pathological subsets of $[0,1]$for which any attempt to extent the notion of length leads to contradictions. So already for simple set like $[0,1]$ we have to abandon hope to assign 'mass' in a consistent way to each subset. Instead, what are going to do for general spaces, is to specify a collection of subsets that is closed under the natural operations on sets (union, intersection, complement) for which a consistent assignment of 'size' is possible. This naturally leads to the definition of σ-algebras and measures.

Remark. The last paragraph might sound defeatist, but I recommend trying the proof of Theorem 1.1 or similar results. The 'pathological sets' that one needs to construct for such theorems are extremely pathological and require the Axiom of Choice for their construction. For all practical purposes in probability theory, we lose nothing by excluding such 'pathological sets' that prevent a consistent assignment of size.

1.1 σ-Algebras and Measures

Let's first spell out some natural properties of such a 'mass' assignment that we want.

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

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

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

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

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

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

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

The natural domain of additive resp. σ-additive set functions are collections of sets that are closed under complements and (countable) unions. Recall that $\mathscr{P}(\Omega)$ denotes the *power set* of Ω, i.e., the set of all subsets of Ω.

Definition 1.3 (Algebras and σ-algebras)**.** Let Ω be a set and let $\mathscr{A} \subseteq \mathscr{P}(\Omega)$ be a collection of subsets of Ω.

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

[2]Hint: do it first for $[0,1)$ and consider two points equivalent if their difference is rational. Choosing a representative from each equivalence class gives a 'pathological' set

2. We say that $\mathscr{A}$ is a σ-*algebra (on* Ω*)* 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 $\mathscr{A}$, $\bigcup_{n=1}^{\infty} A_n \in \mathscr{A}$.

Remark. Since intersections can be built up from complements and unions, an algebra is closed under *finite* set operations; a σ-algebra is closed under *countable* set operations.

Often we don't bother saying 'on Ω',but note that A^c makes sense only if we know which set Ω we are talking about. We tend to write $\mathscr{F}$ for a σ-algebra (also called a σ-*field* by some people).

Definition 1.4 (Measure spaces)**.** A *measurable space* is a pair $(\Omega, \mathscr{F})$ where $\mathscr{F}$ is a σ-algebra on Ω.
A *measure space* is a triple $(\Omega, \mathscr{F}, \mu)$ where Ω is a set, $\mathscr{F}$ is a σ-algebra on Ω and $\mu : \mathscr{F} \to [0,\infty]$ is a countably additive set function. We call μ is a *measure* on $(\Omega, \mathscr{F})$ and say that μ is

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

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

Any measure μ is also additive and increasing. Recall from Part A Integration that measures also respect monotone limits.
Notation: For a sequence $(F_n)_{n\geqslant 1}$ of sets, $F_n \uparrow F$ means $F_n \subseteq F_{n+1}$ for all n and $\bigcup_{n=1}^{\infty} F_n = F$. Similarly, $G_n \downarrow G$ means $G_n \supseteq G_{n+1}$ for all n and $\bigcap_{n=1}^{\infty} G_n = G$.

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

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

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

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

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

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

What data do we need to specify a measure space? For simple spaces like the discrete measure spaces, we can explicitly write down the values of the measure on small building blocks of the whole space that characterizes the measure.

Example 1.7 (Discrete measure theory)**.** Let Ω be a countable set. A *mass function* on Ω is any function $\bar{\mu} : \Omega \to [0,\infty]$. Given such a $\bar{\mu}$ we can define a measure on $(\Omega, \mathscr{P}(\Omega))$ by setting $\mu(A) = \sum_{x\in A} \bar{\mu}(x)$. Equally, given a measure μ on $(\Omega, \mathscr{P}(\Omega))$ we can define a corresponding mass function by $\bar{\mu}(x) = \mu(\{x\})$. For countable Ω there is a one-to-one correspondence between measures on $(\Omega, \mathscr{P}(\Omega))$ and mass functions; that is the measure μ is determined by its values on a strict (in fact, pretty small) subset of the σ-algebra $\mathscr{P}(\Omega)$.

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 space explicitly. This is because σ-algebras are in general not amenable to an explicit presentation; moreover, we don't want to specify a measure μ by having to explicitly assigning values to $\mu(A)$ for each A in the σ-algebra (there are a lot of such sets!) and then also check that this assignment is consistent with the defining properties of a measure.

Instead 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 small collection of subsets of Ω that act as 'building blocks' for sets in the σ-algebra (as the singletons did in Example 1.7). First we need some more definitions.

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

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

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

An important example that you encountered last year is the σ-algebra generated by all open subset of $\mathbb{R}$; the so-called Borel σ-algebra. Below we give the definition for general topologies.

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

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

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

Note that $\mathscr{B}(\Omega)$ depends not just on the set Ω, but also on the topology on Ω. Usually, this is understood: in particular, when $\Omega = \mathbb{R}$, we mean the usual Euclidean topology on $\mathbb{R}$. A topology is closed under arbitrary unions and finite intersections. However, it need not be closed under complements, so it need not be a σ-algebra. The generated σ-algebra $\mathscr{B}(\Omega) = \sigma(\mathscr{T})$ contains $\mathscr{T}$ and may be strictly larger.

Coming back to our original motivation of finding building blocks that determine the measure, could we use the sets in the topology $\mathscr{T}$ as such building blocks, that is the open set? It is certainly sufficient to determine finite measures (this will be a simple corollary of what we'll do below) but this is far from satisfactory in practice since there are a lot of open sets! Instead we can do much better and we will show below that so-called π-systems provide a convenient closure property that makes the uniqueness argument go through.

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

To motivate the above definition, note that for a measure μ, we must have

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B).$$

Hence, once we specify the right hand side of the above equation, the values of μ on arbitrary unions is also specified. Put differently, finite intersections are a natural algebraic structure that lets a generating family keep track of how its pieces overlap. In fact, you have encountered π-systems before: an algebra is automatically a π-system and coming back to our previous discussion about Borel sets:

Example 1.11. The collection

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

forms a π-system. It is much smaller than the collection of opens sets in $\mathbb{R}$. Moreover, $\sigma(\pi(\mathbb{R})) = \mathscr{B}(\mathbb{R})$ (exercise!).

Example 1.12. Let $\Omega = \{1,2,3,4\}$ and $\mathscr{I} = \{\{1,2\},\{2,3\}\}$. Then $\sigma(\mathscr{I}) = \mathscr{P}(\Omega)$ but is not a π-system. In Sheet 1 you will construct two meaures that agree on $\mathscr{I}$ but not on $\sigma(\mathscr{I})$. This again demonstrates the need for being closed under intersections.

Now that the notion of a π-systems gives us a candidate for what these 'building blocks' are, we are left with two problems:

1. Firstly, we need to know that it is possible to extend the measure that we specify to the whole σ-algebra. This *construction* problem is often handled with *Carathéodory's Extension Theorem* (Theorem 1.14 below).

2. Secondly, we need to guarantee that there is only *one* measure on the σ-algebra that is consistent with our specification.

The second problem is addressed by the following theorem.

Theorem 1.13 (Uniqueness of extension)**.** *Let μ_1 and μ_2 be measures on the same measurable space $(\Omega,\mathscr{F})$, and let $\mathscr{I} \subseteq \mathscr{F}$ be a π-system. If $\mu_1(\Omega) = \mu_2(\Omega) < \infty$ and $\mu_1 = \mu_2$ on $\mathscr{I}$, then $\mu_1 = \mu_2$ on $\sigma(\mathscr{I})$.*

We will often apply the theorem to a π-system $\mathscr{I}$ with $\sigma(\mathscr{I}) = \mathscr{F}$, so the conclusion is that μ_1 and μ_2 agree. A very important special case is that if two *probability* measures on Ω agree on a π-system, then they agree on the σ-algebra generated by that π-system. For a proof of Theorem 1.13 see (e.g.) Williams, Appendix A.1.

That deals with uniqueness, but what about existence? A first step is Carathéodory's Extension Theorem.

Theorem 1.14 (Carathéodory Extension Theorem)**.** *Let Ω be a set and $\mathscr{A}$ an algebra on Ω, and let $\mathscr{F} = \sigma(\mathscr{A})$. Let $\mu_0 : \mathscr{A} \to [0,\infty]$ be a countably additive set function. Then there exists a measure μ on $(\Omega,\mathscr{F})$ such that $\mu = \mu_0$ on $\mathscr{A}$.*

Remark. If $\mu_0(\Omega) < \infty$, then Theorem 1.13 tells us that μ is unique, since an algebra is certainly a π-system.

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

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

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

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

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

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

Corollary 1.15. *There exists a unique Borel measure μ on $\mathbb{R}$ such that for all $a,b \in \mathbb{R}$ with $a < b$, $\mu((a,b]) = b-a$. The measure μ is the Lebesgue measure on $\mathscr{B}(\mathbb{R})$.*

The proof of Corollary 1.15 is an exercise. (The hard part is checking countable additivity on a suitable algebra; we will do a related example in a moment. Note that an extra step is required for uniqueness since $\mu(\Omega) = \infty$.)

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

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

Definition 1.16. Let μ be a Borel probability measure on $\mathbb{R}$. The *distribution function* of μ is the function $F : \mathbb{R} \to \mathbb{R}$ defined by $F(x) = \mu((-\infty,x])$.

Any distribution function F has the following properties:

1. F is (weakly) increasing, i.e., $x < y$ implies $F(x) \leqslant F(y)$,
2. $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$, and
3. F is *right continuous*: $y \downarrow x$ implies $F(y) \to F(x)$.

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

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

Theorem 1.17 (Lebesgue)**.** *Let $F : \mathbb{R} \to \mathbb{R}$ be an increasing, right continuous function with $F(-\infty) = 0$ and $F(\infty) = 1$. Then there is a unique Borel probability measure $\mu = \mu_F$ on $\mathbb{R}$ such that $\mu((-\infty,x]) = F(x)$ for every x. Every Borel probability measure μ on $\mathbb{R}$ arises in this way.*

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

Proof. Suppose for the moment that the existence statement holds. Since $\pi(\mathbb{R}) = \{(-\infty,x] : x \in \mathbb{R}\}$ is a π-system which generates the σ-algebra $\mathscr{B}(\mathbb{R})$, uniqueness follows by Theorem 1.13. Also, to see the final part, let μ be any Borel probability measure on $\mathbb{R}$, and let F be its distribution function. Then F has the properties required for the first part of the theorem, and we obtain a measure μ_F which by uniqueness is the measure μ we started with.

For existence we shall apply Theorem 1.14, so first we need a suitable algebra. For $-\infty \leqslant a \leqslant b < \infty$, let $I_{a,b} = (a,b]$, and set $I_{a,\infty} = (a,\infty)$. So $I_{a,b} = \{x \in \mathbb{R} : a < x \leqslant b\}$. Let $\mathscr{I} = \{I_{a,b} : -\infty \leqslant a \leqslant b \leqslant \infty\}$ be the collection of intervals that are open on the left and closed on the right. Let $\mathscr{A}$ be the set of finite disjoint unions of elements of $\mathscr{I}$; then $\mathscr{A}$ is an algebra, and $\sigma(\mathscr{A}) = \sigma(\mathscr{I}) = \mathscr{B}(\mathbb{R})$.

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

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

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

First note that by Lemma 1.6, 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 $E_{n+1} \subseteq E_n$. 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 $A_n \downarrow \emptyset$. □

1.2 More on Distribution Functions

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

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

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

This measure has distribution function $F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y$ but careful: it is not necessarily true that $F'(x) = f(x)$ for all x since changing f on a nullset does not change F.

Example 1.18.

- 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 $[0,1]$.
- Now let
$$f(x) = \begin{cases} 1, & x \in (0,1),\ x \neq \frac{1}{2}, \\ 2, & x = \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases}$$
and note that its distribution function $F(x) = \int_{-\infty}^{x} f(y)dy$ is the same as above. However, $F'(\frac{1}{2}) = 1$, whereas $f(\frac{1}{2}) = 2$.
- 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.)
- (Non-Examinable) One might be tempted to guess that if the distribution function F is continuous, then it should have a density (with respect to the Lebesgue measure). This is not the case! There are examples of continuous distribution functions F that do not admit a density with respect to Lebesgue measure, for example the devil's staircase, which corresponds (roughly speaking) to the uniform distribution on the Cantor set C. Formally, the devil's staircase is defined as
$$F(x) = \begin{cases} 0, & x \leqslant 0, \\ \sum_{n=1}^{\infty} \frac{a_n}{2^{n+1}}, & x \in C, \quad x = \sum_{n=1}^{\infty} \frac{a_n}{3^n}, \quad a_n \in \{0,2\}, \\ F\big(\sup(C \cap (-\infty, x])\big), & x \in (0,1) \setminus C, \\ 1, & x \geqslant 1. \end{cases}$$
but I invite to draw the distribution function when the sum is truncated after the first terms: you'll see a piecewise constant distribution function that although it is discontinuous approximates the continuous function F. To see that F it is indeed a distribution function, note that
$$F(x) = \mathbb{P}(X \leqslant x) \text{ for } X = \sum_{n \geqslant 1} \frac{2B_n}{3^n}$$
where B_n are iid Bernoulli random variables with parameter 0.5. To see that F is continuous, it suffices to show that $\mathbb{P}(X = x) = 0$ for every $x \in \mathbb{R}$: therefore note that for every n, the first n Bernoulli variables determine one of 2^n disjoint intervals of length 3^{-n} containing X. Each such interval has probability 2^{-n}. Since these intervals have disjoint interiors, a fixed point x belongs to at most two of them. Consequently,
$$\mathbb{P}(X = x) \leqslant 2^{1-n}$$
for every n, and hence $\mathbb{P}(X = x) = 0$. Thus F has no jumps and is continuous. Since X takes values in the Cantor set C and C is a Lebesue null set, the existence of a density f, would imply $1 = \mathbb{P}(X \in C) = \int_C f dx = 0$.

Deep Dive

Realising the existence of a devil's staircase can be scary (like the existence of non-measurable sets) but luckily it is about as bad as it can get! This is the content of the Lebesgue decomposition theorem: any finite Borel measure μ on $\mathbb{R}$ can be written as

$$\mu = \mu_{\text{ac}} + \mu_{\text{sc}} + \mu_{\text{at}},$$

where $\mu_{\text{ac}} \ll dx$ is absolutely continuous with respect to the Lebesgue measure (we will discuss what absolutely continuous means when introduce the Radon-Nikodym theorem), μ_{at} is atomic (a sum of point masses), and μ_{sc} is singular continuous (it has no atoms but is supported on a Lebesgue-null set). In case of the devil's staircase: it is continuous, nondecreasing, constant on every complementary interval of the Cantor set, and has derivative 0 Lebesgue-a.e., yet rises from 0 to 1. Thus its increase is carried entirely by a null set, showing why the singular-continuous component is genuinely distinct from both densities and atoms.

Another important result related to Example 1.18 is the *Lebesgue differentiation Theorem*, which states that for any locally integrable function $f : \mathbb{R} \to \mathbb{R}$,

$$\lim_{r \downarrow 0} \frac{1}{2r} \int_{x-r}^{x+r} |f(y) - f(x)|\, dy = 0$$

for Lebesgue-almost every x. In particular, if

$$F(x) = \int_{-\infty}^{x} f(y)\, dy$$

(with f integrable), then $F'(x) = f(x)$ for Lebesgue-almost every x. More generally, if a finite Borel measure μ has Lebesgue decomposition $\mu = f\, dx + \mu_s$, where μ_s is singular with respect to Lebesgue measure, then its distribution function $F(x) = \mu((-\infty, x])$ satisfies $F'(x) = f(x)$ almost everywhere. Thus, differentiation recovers the density of the absolutely continuous part, while the singular part is invisible to the derivative almost everywhere.

2 Integration

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

2.1 Measurable functions and the Definition of the Integral

Lebesgue integration in a nutshell: the key idea behind Lebesgue integration is to approximate a function by partitioning its *range*, rather than its domain. Recall that Riemann integration partitions the x-axis into intervals and approximates the function on each interval. In contrast, for Lebesgue integration we partition the y-axis into intervals and consider the sets of points where the function takes values in each interval. The key point is that the resulting partitioning depends on the function! Moreover, the sets in the resulting partition need not be intervals, but measurability ensures that we can assign a measure to them. We then approximate the function by a simple function that is constant on each of these sets. Its integral is easy to compute: we multiply each value by the measure of the set on which it is attained and add the contributions. Refining the partition of the range

leads to increasingly accurate approximations.

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

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

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

Remark. Exercise[3]: Explicitly write down the sets I have to add to $\mathscr{B}(\mathbb{R})$ to get $\mathscr{B}(\bar{\mathbb{R}})$?

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

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

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

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

Recall that

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

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

Lemma 2.3. *Let (f_n) be a sequence of measurable functions from $(\Omega,\mathscr{F})$ to $\overline{\mathbb{R}}$. Then*

$$\max\{f_1,f_2\},\quad \min\{f_1,f_2\},\quad \sup_n f_n,\quad \inf_n f_n,\quad \limsup_{n\to\infty}f_n,\quad \liminf_{n\to\infty}f_n$$

are extended-real-valued measurable functions. If f_1,f_2 are real-valued, then f_1+f_2 and f_1f_2 are measurable, as is f_1/f_2 on the measurable set $\{f_2\neq 0\}$ (and any measurable extension of this quotient to Ω). If f is real-valued and $h:\mathbb{R}\to\mathbb{R}$ is continuous, then $h\circ f$ is measurable.

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

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

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

[3]Easy to check that $\mathscr{B}(\bar{\mathbb{R}})=\{A\cup S:A\in\mathscr{B}(\mathbb{R}),S\subset\{-\infty,\infty\}\}$. That is $A\subset\bar{\mathbb{R}}$ is in $\mathscr{B}(\bar{\mathbb{R}})$ if and only if $A\cap\mathbb{R}\in\mathscr{B}(\mathbb{R})$.

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

$$\phi = \sum_{k=1}^{n} a_k \mathbf{1}_{E_k} \tag{8}$$

where each $E_k \in \mathscr{F}$ and each $a_k \in \mathbb{R}$. The *canonical form* of ϕ is the decomposition, unique up to reordering, as in (8) where the numbers a_k are distinct and non-zero and the sets E_k are disjoint and non-empty.

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

Definition 2.5. If ϕ is a non-negative simple function with canonical form (8), then we define the integral of ϕ with respect to μ as

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

Here and in the following we use the convention $0 \cdot \infty = 0$.

This formula then also applies (exercise) whenever ϕ is as in (8), even if this is not the canonical form, as long as we avoid $\infty - \infty$ (for example by taking $a_k \geqslant 0$).

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

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

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

Example 2.7 (Approximating by horizontal strips)**.** Let $f : \Omega \to [0, \infty]$ be measurable. For each $n \geqslant 1$, partition the range of f into intervals of width 2^{-n}, truncating values above n. More precisely, define

$$E_{n,k} = \{\omega \in \Omega : k2^{-n} \leqslant f(\omega) < (k+1)2^{-n}\},$$

for $k = 0, \ldots, n2^n - 1$, and let $E_{n,\infty} = \{\omega : f(\omega) \geqslant n\}$. Then

$$\phi_n = \sum_{k=0}^{n2^n - 1} k2^{-n} \mathbf{1}_{E_{n,k}} + n\mathbf{1}_{E_{n,\infty}}$$

is a non-negative measurable simple function. By construction,

$$0 \leqslant \phi_n \leqslant \phi_{n+1} \leqslant f, \qquad \phi_n(\omega) \longrightarrow f(\omega)$$

for every $\omega \in \Omega$. Thus every non-negative measurable function can be approximated from below by an increasing sequence of simple functions obtained by partitioning its range. Moreover,

$$\int f \, \mathrm{d}\mu = \lim_{n \to \infty} \int \phi_n \, \mathrm{d}\mu.$$

This last identity follows from the monotone convergence theorem (which we recall in Section 2.3 below).

Definition 2.8 (Integral and integrability)**.** Let $f : \Omega \to \overline{\mathbb{R}}$ be measurable, and write

$$f^+ = \max\{f, 0\}, \qquad f^- = \max\{-f, 0\}.$$

We say that the integral of f is *well-defined* if at least one of

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

is finite. In this case, we define

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

which may take values in $[-\infty, \infty]$. We say that f is *integrable* if

$$\int |f| \mathrm{d}\mu < \infty,$$

or equivalently, if both $\int f^+ \mathrm{d}\mu$ and $\int f^- \mathrm{d}\mu$ are finite.

Thus, the integral of a measurable function may be well-defined even if the function is not integrable. The integral is undefined precisely when

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

In particular, the integral of a non-negative measurable function is always well-defined, possibly with value $+\infty$.

There are other possible sequences of steps to defining the integral, giving the same result. The integral has the same basic properties as in the special case of Lebesgue measure, with the same proofs: for example, if f and g are integrable and $a, b \in \mathbb{R}$, then $af + bg$ is integrable and

$$\int (af + bg) \mathrm{d}\mu = a \int f \mathrm{d}\mu + b \int g \mathrm{d}\mu.$$

Moreover, if f and g are non-negative measurable functions, then

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

and for $c \geqslant 0$,

$$\int cf \mathrm{d}\mu = c \int f \mathrm{d}\mu,$$

where we use the convention $0 \cdot \infty = 0$. In proving statements like the above, it is often helpful to start by considering the case $f = 1_E$, then simple functions f, then non-negative measurable f, and finally general measurable f. It also helps to recall that given any measurable $f \geqslant 0$ there are simple functions $f_n \geqslant 0$ with $f_n \uparrow f$.

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, whenever the right-hand side is well-defined,

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

Here $f1_A$ denotes the function equal to f on A and zero outside A, even where f is infinite. In the previous section, we briefly recalled the notion of a probability density function. The concept of a density is not limited to Lebesgue measure and will be useful tool for use when define conditional expectation.

Definition 2.9 (Density with respect to a measure)**.** Let μ and ν be measures on the same measurable space $(\Omega, \mathscr{F})$. We say that a non-negative measurable function $f : \Omega \to [0, \infty]$ is a *density of* μ *with respect to* ν if

$$\mu(A) = \int_A f \mathrm{d}\nu$$

for every $A \in \mathscr{F}$. In this case, we write

$$d\mu = f d\nu.$$

Remark. I want to emphasize that when one talks about the existence of a density, it is with respect to a reference measure. In Example 1.18, the discrete probability measure does not have a density with respect to the Lebesgue measure. However, it would have a density with respect to the discrete measure $\nu = \sum_{n\geqslant 1} \delta_{x_n}$ where δ_x is the usual Dirac measure that assigns 1 to a set if it contains x, and otherwise 0. When talking about measures on $\mathbb{R}$ and a reference measure isn't explicitly mentioned when talking about a density, then it is usually meant with respect to Lebesgue measure.

Example 2.10.

- 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.
- Suppose that μ is a discrete measure with mass p_i at point x_i, where the x_i are distinct and $p_i > 0$ 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 the sum is well-defined in the extended sense, i.e., the sums of its positive and negative parts are not both infinite.
- Suppose that μ has distribution function $F(x) = \int_{-\infty}^{x} g(y)\,\mathrm{d}y$. Then

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

  where the second integral is with respect to Lebesgue measure. This identity holds for non-negative measurable f, and more generally whenever the integrals are well-defined in the extended sense.

Remark 2.11. One final property of integration that is easy to check (exercise) from the definitions is that for $f \geqslant 0$, $\int f \,\mathrm{d}\mu$ is determined by the numbers $\mu(\{x : f(x) \geqslant t\})$ for each $t \geqslant 0$. Hence, for general f, $\int f \,\mathrm{d}\mu$ is determined by the numbers $\mu(\{x : f(x) \geqslant t\})$ for $t \geqslant 0$ and $\mu(\{x : f(x) \leqslant t\})$ for $t \leqslant 0$. Since $\{x : f(x) \geqslant t\}$ is the complement of the union of the sets $\{x : f(x) \leqslant s\}$, $s < t$, on a probability space, say, $\int f \,\mathrm{d}\mu$ is determined by the numbers $\mu(\{x : f(x) \leqslant t\})$ for $t \in \mathbb{R}$. This holds even across probability spaces: if f_i is a measurable function on the probability space $(\Omega_i, \mathscr{F}_i, \mu_i)$ and for every $t \in \mathbb{R}$, $\mu_1(\{x \in \Omega_1 : f_1(x) \leqslant t\}) = \mu_2(\{x \in \Omega_2 : f_2(x) \leqslant t\})$, then $\int f_1 \,\mathrm{d}\mu_1 = \int f_2 \,\mathrm{d}\mu_2$, or both are undefined.

Definition 2.12 (μ-almost everywhere)**.** Let $(\Omega, \mathscr{F}, \mu)$ be a measure space. We say that a property holds *μ-almost everywhere* or *μ-a.e.* if it holds except on a set of μ-measure zero. If μ is a probability measure, we often say *almost surely* or *a.s.* instead of almost everywhere. Thus an event A holds almost surely if $\mathbb{P}[A] = 1$. This does not imply that $A = \Omega$.

An important property of integration is that

$$f = g \quad \mu\text{-a.e.} \quad \Longrightarrow \quad \int f \,\mathrm{d}\mu = \int g \,\mathrm{d}\mu,$$

whenever either integral is well-defined. In fact, one is well-defined if and only if the other is, and in that case, they are equal. Generally speaking, we don't care what happens on sets of measure zero. It is vital to remember that notions of almost everywhere depend on the underlying measure μ.

Deep Dive

The Lebesgue integral is a remarkably powerful notion of integration: compared with Riemann integration, it offers much stronger convergence theorems and is a workhorse of modern probability. It is all that you need for this course (and many others). However, for those who are hungry for more integration theory, there is more! We started out with the general definition of a measurable function but then quickly required the codomain to be $\mathbb{R}$. But much of the remaining arguments make sense when we replace $\mathbb{R}$ by a linear space, such as a Banach space. For example, suppose we have a function $f : [0,1] \to C([0,1])$, so that each $f(t)$ is itself a continuous function. How can we integrate f over t, obtaining another continuous function? This motivates the *Bochner integral*, which extends Lebesgue integration to functions taking values in Banach spaces. The *Pettis integral* goes even a step further. Suppose $f : [0,1] \to B$ takes values in a Banach space B, but is too irregular to be Bochner integrable. It may nevertheless be possible to integrate every scalar-valued function $\ell(f(t))$, where $\ell : B \to \mathbb{R}$ is a continuous linear functional, and find an element $b \in B$ satisfying $\ell(b) = \int_0^1 \ell(f(t))\,dt$ for every such ℓ. This is the idea behind Pettis integration, which allows us to integrate certain functions that cannot be integrated in the Bochner sense. The need for Bochner and Pettis integration can arise naturally; e.g. for a random continuous function $X : [0,1] \to \mathbb{R}$, such as a randomly evolving temperature profile, we might want to take the expectation of the entire function, rather than at each point separately.

Finally, what if we want to integrate *with respect to a random process* rather than time or a fixed measure? For instance, how should we interpret $\int_0^T H_t\,dB_t$, where B_t is Brownian motion and H_t is a trading strategy? In fact, this question arises naturally and in discrete time we have encountered this briefly in (2) and we will return to this later in this course. In B8.2 you will see that taking this from discrete to continuous time leads to *stochastic integrals*, such as the Itô integral.

2.2 The Case of Probability Measures: Recall of some Terminology

The case when the measure is a probability measure is of special importance to us and we will use probabilistic terminology and notation, instead of the more general measure theoretic one.

Definition 2.13 (Random Variable)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. We call a measurable function $X : \Omega \to \mathbb{R}$ a (real-valued) *random variable*. We also say that $(\Omega, \mathscr{F}, \mathbb{P})$ carries the random variable X.

For the above definition, recall our convention that when we deal with a real-valued functions, we mean measurability with respect to $\mathscr{B}(\mathbb{R})$, unless stated otherwise. As we already did in Prelims, we can think of Ω as the sample space of an experiment, and the random variable X as an observable, i.e. something that can be measured. What is the integral of X?

Definition 2.14 (Expectation)**.** Let X be a random variable on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. If $\int X \,\mathrm{d}\mathbb{P}$ is well-defined, we call

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

the *expectation* of X. We say that X is *integrable* if $\mathbb{E}[|X|] < \infty$.

Definition 2.15 (Image Measure and Distribution Function)**.** A random variable X induces a probability measure μ_X on $\mathbb{R}$ via

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

In particular, $F_X(x) = \mu_X((-\infty, x])$ defines the *distribution function* of X (cf. Theorem 1.17).

Since $\{(-\infty, x] : x \in \mathbb{R}\}$ is a π-system, we already know that the distribution function uniquely determines μ_X. From Remark 2.11,

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

This identity holds whenever the expectation is well-defined, including when it takes the value $+\infty$ or $-\infty$.

Remark. When we deal with a random variable X, there are really two probability spaces in play: firstly, the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ that carries X but also the "induced" probability space $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mu_X)$. I want to invite you to think of the domain $(\Omega, \mathscr{F}, \mathbb{P})$ as an abstract mathematical object and in stark contrast, of $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mu_X)$ as the space we have access to via experiments (as in your statistics class). While there are sometimes canonical choices for $(\Omega, \mathscr{F}, \mathbb{P})$ which can be convenient (e.g. the classic dice throw example from your first probability course where it is easy to spell out a discrete set Ω and an explicit σ-algebra), it usually doesn't matter what we choose here, as long as it is rich enough to carry a probability measure $\mathbb{P}$ that is complex enough for our purposes. To sum up, typically we don't explicitly specify the sample space and work directly with μ_X.

Deep Dive

Given a measurable map $T : \Omega \to E$, we can *push forward* a measure μ on Ω to a measure $T_*\mu$ on E, defined by $T_*\mu(A) = \mu(T^{-1}(A))$. Conversely, we can *pull back* a measurable function $f : E \to \mathbb{R}$ to a function $f \circ T$ on Ω. These two operations are closely related through the change-of-variables formula

$$\int_E f\, d(T_*\mu) = \int_\Omega f \circ T\, d\mu,$$

valid for non-negative measurable f, or whenever the integrals are well-defined. In probability, this is precisely what happens when a random variable $X : \Omega \to \mathbb{R}$ pushes forward the probability measure $\mathbb{P}$ to its distribution $\mu_X = X_*\mathbb{P}$. The formula then becomes

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}} f(x)\, \mu_X(dx).$$

This illustrates a fundamental *duality between functions and measures*: integration provides a natural pairing $(f, \mu) \mapsto \int f\, d\mu$. Under suitable assumptions, measures can even be identified with continuous linear functionals on spaces of functions, as formalised by the *Riesz representation theorem*. From this perspective, pushing measures forward and pulling functions back are *dual operations*: the pushforward T_* is the adjoint of the pullback $T^* : f \mapsto f \circ T$, in the sense that

$$\langle f, T_*\mu \rangle = \langle T^* f, \mu \rangle.$$

Thus, *measures move forward along maps, while functions move backward by composition*, and integration connects the two. You might now say: "Ok cute, but so what?". Well it is much more than a mathematical curiosity: this allows us to study how measures evolve by studying how functions transform, and vice versa. For certain questions, one perspective can be much easier than the dual one. This is a fundamental principle that you encounter in probability theory, dynamical systems, and the theory of Markov processes.

2.3 The Convergence Theorems

The following theorems were proved in the Part A integration course for the special case of the Lebesgue integral. Again, the proofs carry over to the more general integral defined here.

Theorem 2.16 (Fatou's Lemma)**.** *Let (f_n) be a sequence of non-negative measurable functions on $(\Omega, \mathscr{F}, \mu)$. Then*

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

Theorem 2.17 (Monotone Convergence Theorem)**.** *Let (f_n) be a sequence of non-negative measurable functions on $(\Omega, \mathscr{F}, \mu)$. Then*

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

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

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

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

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

Theorem 2.18 (Dominated Convergence Theorem)**.** *Let (f_n) be a sequence of measurable functions on $(\Omega, \mathscr{F}, \mu)$ with $f_n \to f$ pointwise. Suppose that for some* **integrable** *function g, $|f_n| \leqslant g$ for all n. Then f is integrable and*

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

Again, convergence almost everywhere is enough.

We will also use the following less standard result.

Lemma 2.19 (Reverse Fatou Lemma)**.** *Let (f_n) be a sequence of non-negative, measurable functions. Assume that there exists an* **integrable** *function g such that $f_n \leqslant g$ for all n. Then*

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

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

2.4 Product Spaces

A natural question is, given two measure spaces, can we construct one "bigger" measure space that contains all the structure of the two measure spaces. This means we have to decide how construct a new [set,σ-algebra,measure] from two [sets,σ-algebras,measures] as building blocks. For the set (the sample space), we can simply take the cartesian product. However, for σ-algebras and measures some care needs to be taken.

Example 2.20. Take $\mathscr{F}_1 = \mathscr{F}_2 = \mathscr{B}(\mathbb{R})$ and consider the two rectangles

$$A_1 \times B_1 = (0,1) \times (0,1), \qquad A_2 \times B_2 = (2,3) \times (2,3).$$

Their union

$$\big((0,1) \times (0,1)\big) \cup \big((2,3) \times (2,3)\big)$$

cannot be written as $A \times B$ for any $A, B \subseteq \mathbb{R}$. Thus the collection

$$\{A \times B : A \in \mathscr{F}_1,\ B \in \mathscr{F}_2\}$$

is not closed under finite unions, so it is not a σ-algebra.

The fix is easy: we consider the smallest σ-algebra that contains all rectangles.

Definition 2.21 (Product σ-algebras)**.** Given two sets Ω_1 and Ω_2, the *Cartesian product* $\Omega = \Omega_1 \times \Omega_2$ is the set of pairs (ω_1, ω_2) with $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$. If $\mathscr{F}_i$ is a σ-algebra on Ω_i, then a *measurable rectangle* in $\Omega = \Omega_1 \times \Omega_2$ is a set of the form $A_1 \times A_2$ with $A_1 \in \mathscr{F}_1$ and $A_2 \in \mathscr{F}_2$. The *product σ-algebra* $\mathscr{F} = \mathscr{F}_1 \times \mathscr{F}_2$ is the σ-algebra on Ω generated by the set of all measurable rectangles. (Note that $\mathscr{F}$ is not the Cartesian product of $\mathscr{F}_1$ and $\mathscr{F}_2$.)

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

$$\mathbb{P}[A_1 \times A_2] = \mathbb{P}_1[A_1]\mathbb{P}_2[A_2] \tag{9}$$

for each measurable rectangle and extending it to the whole of $\mathscr{F}$. Note that the set $\mathscr{I}$ of measurable rectangles is a π-system with (by definition) $\sigma(\mathscr{I}) = \mathscr{F}$, so if such a probability measure on $(\Omega, \mathscr{F})$ exists, it is unique by Theorem 1.13.

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

$$\mathbb{P}[R_1 \cup \cdots \cup R_n] = \sum_{i=1}^{n} \mathbb{P}[R_i] \tag{10}$$

when $R_1, \ldots, R_n \in \mathscr{I}$ are disjoint. We have to check this is indeed well-defined but this follows from the more general Lemma below (in fact, without the Lemma below it is a straightforward but lengthy exercise to check well-definedness).

To check that we can extend $\mathbb{P}$ to the whole of $\mathscr{F} = \sigma(\mathscr{A})$, we need to check that $\mathbb{P}$ defined by (9) and (10) is actually *countably* additive on $\mathscr{A}$ so that we can apply Carathéodory's Extension Theorem.

Lemma 2.22. *The set function $\mathbb{P}$ defined on $\mathscr{A}$ through (9) and* (10) *is* countably *additive on $\mathscr{A}$.*

The proof of this Lemma is part of Sheet 2. Try it yourself before taking inspiration from the sketch below!

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

$$A_{\omega_2} = \{\omega_1 : (\omega_1, \omega_2) \in A\} \subseteq \Omega_1 \text{ and } f(\omega_2) = \mathbb{P}_1[A_{\omega_2}].$$

Convince yourself that $\mathbb{P}(A) = \int f(\omega_2) d\mathbb{P}_2$; then use that A is the disjoint unions of sets. Conclude using that $\mathbb{P}_1$ is countably additive and the Monotone Convergence Theorem. □

By Carathéodory's Extension Theorem (Theorem 1.14) and Theorem 1.13 we see that $\mathbb{P}$ extends uniquely to a probability measure on $\sigma(\mathscr{A}) = \mathscr{F}$.

Definition 2.23 (Product measure)**.** The measure $\mathbb{P}$ defined through (9) is called the *product measure* on $(\Omega, \mathscr{F})$, and denoted $\mathbb{P}_1 \times \mathbb{P}_2$. The probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is the *product probability space* $(\Omega_1 \times \Omega_2, \mathscr{F}_1 \times \mathscr{F}_2, \mathbb{P}_1 \times \mathbb{P}_2)$.

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

Definition 2.24. Let $(\Omega_i, \mathscr{F}_i)_{i \geqslant 1}$ be a sequence of measurable spaces. The *product σ-algebra* $\mathscr{F} = \prod_{i=1}^{\infty} \mathscr{F}_i$ on $\Omega = \prod_{i=1}^{\infty} \Omega_i$ is the σ-algebra generated by all sets of the form $\prod_{i=1}^{n} A_i \times \prod_{i=n+1}^{\infty} \Omega_i$ where $A_i \in \mathscr{F}_i$, i.e., by all finite-dimensional measurable rectangles.

Remark 2.25 (Countable products of probability measures)**.** Given a sequence of probability spaces, one can define a product probability measure on the product σ-algebra with the expected properties. One way to do this is to apply Theorem 1.14 directly as in the proof of Lemma 2.22, but the condition is quite tricky to verify. It also follows by taking a suitable 'limit' of finite products using the Kolmogorov Consistency Theorem. An alternative approach for Borel measures on $\mathbb{R}$ is outlined on the problem sheets.

Although the product-measure construction above was stated for probability measures, it extends to sigma-finite measures (Exercise: convince yourself of this). A familiar example is two-dimensional Lebesgue measure, the product of one-dimensional Lebesgue measure with itself (and similarly in $\mathbb{R}^d$).

2.5 Integration on Product Spaces: Fubini and Tonelli

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

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

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

Theorem 2.26 (Fubini and Tonelli)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be the product of the probability spaces $(\Omega_i, \mathscr{F}_i, \mathbb{P}_i)$, $i = 1,2$, and let $f : \Omega_1 \times \Omega_2 \to \overline{\mathbb{R}}$ be measurable. Suppose that either*

(i) $f \geqslant 0$, or

(ii) f is integrable, i.e., $\int_\Omega |f|\,\mathrm{d}\mathbb{P} < \infty$.

Then

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

In case (i), the section integrals are measurable functions and all integrals are well-defined, possibly with value $+\infty$. In case (ii), it follows from case (i) applied to $|f|$ that the function

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

is measurable and finite for $\mathbb{P}_1$-almost every x. Consequently, the inner integral

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

exists and is finite almost everywhere. Setting it equal to zero on the exceptional measurable null set gives a measurable, integrable function on Ω_1. All three integrals are finite.

There is some inconsistency in the literature regarding what is referred to as Fubini's Theorem and Tonelli's Theorem. Rest assured, the distinction is largely conventional, and different authors package the two results differently. But a common modern convention is to refer to (i) as Tonelli's Thereom (non-negative functions, ∞ as possible value of the double integral), and to (ii) as Fubini's Theorem (integrable functions, finite values of the double integral).

Remark. 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. For example, partition $(0,1]$ into the intervals

$$I_n = (2^{-n}, 2^{-n+1}], \qquad n \geqslant 1,$$

and define $f : (0,1]^2 \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} 2^{2n-1}, & (x,y) \in I_n \times I_n, \\ -2^{2n}, & (x,y) \in I_n \times I_{n+1}, \\ 0, & \text{otherwise.} \end{cases}$$

Is then straightforward to check that both iterated integrals exist and are finite, but they are different! This does not contradict Fubini–Tonelli, since

$$\int_{(0,1]^2} |f(x,y)| \, dx dy = \sum_{n=1}^{\infty} \left(\frac{1}{2} + \frac{1}{2} \right) = \infty.$$

The example illustrates how changing the order of integration can effectively rearrange infinitely many positive and negative contributions.

Remark. By Tonelli's theorem applied to $|f|$, the condition

$$\int_{\Omega_1 \times \Omega_2} |f| \, d\mathbb{P}(x,y) < \infty$$

is equivalent to

$$\int_{\Omega_1} \left(\int_{\Omega_2} |f(x,y)| \, \mathbb{P}_2(dy) \right) \mathbb{P}_1(dx) < \infty,$$

and likewise with x and y interchanged. Thus, in Fubini's theorem, we can equivalently assume that one of the iterated integrals of $|f|$ is finite, rather than assuming integrability on the product space.

2.6 Independence

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

Definition 2.27 (Independence of σ-algebras)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Let I be a finite or countably infinite set. We say that the events $(A_i)_{i \in I}$ where each $A_i \in \mathscr{F}$ are *independent* if for all finite subsets $J \subseteq I$

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

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

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

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

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

Definition 2.28 (σ-algebra generated by a random variable)**.** Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space and let X be a real-valued random variable on $(\Omega,\mathscr{F},\mathbb{P})$. Let $\mathscr{G}\subseteq\mathscr{F}$ be a σ-algebra. We call $\mathscr{G}$ a sub-σ algebra of $\mathscr{F}$ and say that X is $\mathscr{G}$*-measurable* if X is measurable as a function on $(\Omega,\mathscr{G})$. The σ*-algebra generated by* X is

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

More generally, if (X_n) is a finite or infinite sequence of random variables on $(\Omega,\mathscr{F},\mathbb{P})$, then

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

It is easy to check that:

- $\sigma(X)$ is indeed a σ-algebra (see the proof of Proposition 2.2),
- by definition of a random variable (as a measurable function on $(\Omega,\mathscr{F})$), we have $\sigma(X)\subseteq\mathscr{F}$.
- In fact, X is $\mathscr{G}$-measurable if and only if $\sigma(X)\subseteq\mathscr{G}$,
- In particular, $\sigma(X)$ is the smallest σ-algebra $\mathscr{G}$ such that X is $\mathscr{G}$-measurable
- $\sigma(X)=\sigma\big(\{\{X\leqslant t\}:t\in\mathbb{R}\}\big)$, where $\{X\leqslant t\}=\{\omega:X(\omega)\leqslant t\}$ (again, c.f. Proposition 2.2).

Remark. The σ-algebra $\sigma(X)$ represents all the information we could obtain by observing the random variable X. For example, if X is constant, observing it tells us nothing, and $\sigma(X)=\{\emptyset,\Omega\}$. If X takes only two values, then $\sigma(X)$ contains at most four sets, corresponding to whether X takes one value or the other. More generally, if X takes n distinct values, then $\sigma(X)$ consists of the 2^n possible unions of the sets $\{X=x_i\}$. At the other extreme, if X distinguishes all outcomes and its inverse is measurable, then $\sigma(X)=\mathscr{F}$: observing X can reveal everything in $\mathscr{F}$. Thus, the richer the distinctions made by X, the more information is encoded in $\sigma(X)$.

The following result is good for intution and is also useful for many proofs. You have probably seen in it in Part A Integration, but we anyway do it on Sheet 3.

Lemma 2.29. *A random variable Y is $\sigma(X)$-measurable if and only if $Y=f(X)$ for some measurable function $f:\mathbb{R}\to\mathbb{R}$. Similarly, Y is $\sigma(X_1,X_2,\ldots)$-measurable if and only if $Y=f(X_1,X_2,\ldots)$ for some measurable function f on the countable product of $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ with itself.*

We recover the notion of independence of random variables from Prelims and Part A as a special case.

Definition 2.30 (Independence of Random Variables)**.** Random variables $X_1,X_2,\ldots$ are called *independent* if the σ-algebras $\sigma(X_1),\sigma(X_2),\ldots$ are independent.

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

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

Remark. Defining independence in terms of σ-algebras is more general than defining it through joint distributions: it allows us to express independence between arbitrary collections of random variables, or more generally between different sources of information. We see examples of this later in this course; e.g. when we study stochastic processes we can efficently formulate that a future increment is independent of the entire past, using above notion of independence of σ-algebras.

The following result is easy to check.

Lemma 2.31. *Two random variables X and Y on the probability space $(\Omega,\mathscr{F},\mathbb{P})$ are independent if and only if the measure $\mu_{X,Y}$ induced on $\mathbb{R}^2$ by (X,Y) is the product measure $\mu_X \times \mu_Y$, where μ_X and μ_Y are the measures on $\mathbb{R}$ induced by X and Y respectively.*

This Lemma generalizes the result you learned in Prelims and Part A for discrete/continuous random variables – two continuous random variables X and Y are independent if and only if their joint density function can be written as the product of the density function of X and the density function of Y. Note that this is a big generalization, since the assumption of a density is a very strong one.

Of course the conditions of Definition 2.30 would be impossible to check in general – we don't have a nice explicit presentation of the σ-algebras $\sigma(X_i)$. But we can use Theorem 1.13 (uniqueness of extension) to reduce it to something much more manageable.

Theorem 2.32. *Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space. Suppose that $\mathscr{G}$ and $\mathscr{H}$ are sub σ-algebras of $\mathscr{F}$ and that $\mathscr{G}_0$ and $\mathscr{H}_0$ are π-systems with $\sigma(\mathscr{G}_0)=\mathscr{G}$ and $\sigma(\mathscr{H}_0)=\mathscr{H}$. Then $\mathscr{G}$ and $\mathscr{H}$ are independent* iff *$\mathscr{G}_0$ and $\mathscr{H}_0$ are independent, i.e. $\mathbb{P}[G\cap H]=\mathbb{P}[G]\mathbb{P}[H]$ whenever $G\in\mathscr{G}_0$, $H\in\mathscr{H}_0$.*

Proof. Fix $G\in\mathscr{G}_0$. The two functions $H\mapsto\mathbb{P}[G\cap H]$ and $H\mapsto\mathbb{P}[G]\mathbb{P}[H]$ define measures on $(\Omega,\mathscr{H})$ (check!) with the same total mass $\mathbb{P}[G]$, and they agree on the π-system $\mathscr{H}_0$. So by Theorem 1.13 they agree on $\sigma(\mathscr{H}_0)=\mathscr{H}$. Hence, for $G\in\mathscr{G}_0$ and $H\in\mathscr{H}$

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

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

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

Corollary 2.33. *A sequence $(X_n)_{n\geqslant 1}$ of real-valued random variables on $(\Omega,\mathscr{F},\mathbb{P})$ is independent* iff *for all $n\geqslant 1$ and all $x_1,\dots x_n\in\mathbb{R}$ (or $\overline{\mathbb{R}}$),*

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

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

2.7 Tail Events: The Kolmogorov $0-1$ Law

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

Definition 2.34 (Tail σ-algebra)**.** For a sequence of random variables $(X_n)_{n\geqslant 1}$ define

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

and

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

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

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

Theorem 2.35 (Kolmogorov's 0-1 law)**.** *Let (X_n) be a sequence of independent random variables. Then the tail σ-algebra $\mathscr{T}$ of (X_n) contains only events of probability* 0 *or* 1*. Moreover, any $\mathscr{T}$-measurable random variable is almost surely constant.*

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

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

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

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

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

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

and so by Theorem 2.32 the σ-algebras $\sigma(\mathscr{A}) = \mathscr{F}_n$ and $\sigma(\mathscr{B}) = \mathscr{T}_n$ are also independent.

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

Now $\bigcup_{n\geqslant 1} \mathscr{F}_n$ is a π-system (although not in general a σ-algebra) generating the σ-algebra $\mathscr{F}_\infty = \sigma((X_n)_{n\geqslant 1})$. So applying Theorem 2.32 again we see that $\mathscr{F}_\infty$ and $\mathscr{T}$ are independent. But $\mathscr{T} \subseteq \mathscr{F}_\infty$ so that if $A \in \mathscr{T}$

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

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

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

Example 2.36. Let $(X_n)_{n\geqslant 1}$ be a sequence of independent, identically distributed (i.i.d.) random variables and let $S_n = \sum_{k=1}^n X_k$. Consider $L = \limsup_{n\to\infty} S_n/n$. Then L is a tail random variable and so almost surely constant. We'll prove later in the course that, assuming $\mathbb{E}[|X_1|] < \infty$, $L = \mathbb{E}[X_1]$ almost surely.

Deep Dive

We have encountered the following reasoning several times: first establish a property for simple objects (such as indicator functions), then extend it by linearity, and finally use a limiting argument to obtain the property for more general functions. For example, this is the strategy behind the proof of Fubini–Tonelli, where we extend an integration identity from indicators of rectangles to general non-negative measurable functions.

Such arguments can often be streamlined using *Monotone Class Theorems* (MCTs), which provide an abstract mechanism for extending properties from a small generating collection to a much larger class of sets or functions. (Not to be confused with the monotone convergence theorem which unfortunately are also often abbreviated as MCT).

I have not included these theorems in the lectures: when first learning the subject, it is useful to see the underlying arguments explicitly before taking a (albeit elegant) shortcut with an abstract theorem. Further, MCTs require a bit of functional analysis to state them. However, MCTs become a useful tool in advanced stochastic courses, particularly in B8.2 and C8.1. We prove and discuss a classic MCT in Section C of Sheet 1.

2.8 Eventual and Infinitely Often Events: The Borel–Cantelli Lemmas

Throughout, $(\Omega, \mathscr{F}, \mathbb{P})$ will denote a probability space.

Definition 2.37. Let (A_n) be a sequence of sets from $\mathscr{F}$. We define

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

and

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

Lemma 2.38.

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

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

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

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

and hence (taking complements)

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

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

Lemma 2.39 (The First Borel–Cantelli Lemma, BC1)**.** *If* $\sum_{n=1}^{\infty} \mathbb{P}[A_n] < \infty$ *then* $\mathbb{P}[A_n \text{ i.o.}] = 0$.

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

Proof. Let $G_n = \bigcup_{k=n}^{\infty} A_k$. Since $G_n \downarrow \{A_n \text{ i.o.}\}$, we have

$$\begin{aligned}\mathbb{P}(A_n \text{ i.o.}) &= \mathbb{P}\left(\bigcap_{n=1}^{\infty} G_n\right) && \text{(definition of infinitely often)} \\ &= \lim_{n\to\infty} \mathbb{P}(G_n) && \text{(continuity from above)} \\ &\leqslant \lim_{n\to\infty} \sum_{k=n}^{\infty} \mathbb{P}(A_k) && \text{(countable subadditivity)} \\ &= 0 && \text{(tails of a convergent series).}\end{aligned}$$

□

Alternatively, consider $N = \sum_{n=1}^{\infty} \mathbf{1}_{A_n}$, the (random) number of events that hold. Use the Monotone Convergence Theorem to show that $\mathbb{E}[N] = \sum \mathbb{P}[A_n]$, and note that $\mathbb{E}[N] < \infty$ implies $\mathbb{P}[N = \infty] = 0$. A partial converse to BC1 is provided by the second Borel–Cantelli Lemma, but note that we must now assume that the events are *independent*.

Lemma 2.40 (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 $\{A_n^c \text{ eventually}\}$.

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

Hence

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

and

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

□

Example 2.41. 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? How long should we expected to wait to see this word for the first time?

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. The question of expected waiting time is not as straight-forward but we revisit this later in the with the help of a suitable martingale which allows for an elegant, explicit answer. □

We'll see many applications of BC1 and BC2 in what follows and the Sheets.

Example 2.42. Let $(X_n)_{n \geqslant 1}$ be independent exponentially distributed random variables with mean 1 and let $M_n = \max\{X_1, \ldots, X_n\}$. Then

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

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

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

Fix $\varepsilon > 0$. For the upper bound, observe that

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n \geqslant (1+\varepsilon)\log n) = \sum_{n=1}^{\infty} n^{-(1+\varepsilon)} < \infty.$$

By Borel–Cantelli I, almost surely there exists $N < \infty$ such that $X_n < (1+\varepsilon)\log n$ for all $n \geqslant N$. Consequently, for $n \geqslant N$,

$$M_n \leqslant \max\{M_{N-1}, (1+\varepsilon)\log n\},$$

and hence

$$\limsup_{n\to\infty} \frac{M_n}{\log n} \leqslant 1+\varepsilon \quad \text{a.s.}$$

For the lower bound, independence gives

$$\begin{aligned}\mathbb{P}(M_n \leqslant (1-\varepsilon)\log n) &= \left(1 - n^{-(1-\varepsilon)}\right)^n \\ &\leqslant \exp(-n^{\varepsilon}),\end{aligned}$$

where we used $1 - x \leqslant e^{-x}$. Since $\sum_n e^{-n^{\varepsilon}} < \infty$, Borel–Cantelli I implies

$$\liminf_{n\to\infty} \frac{M_n}{\log n} \geqslant 1-\varepsilon \quad \text{a.s.}$$

Letting $\varepsilon \downarrow 0$ along a countable sequence proves

$$\frac{M_n}{\log n} \longrightarrow 1 \quad \text{almost surely.}$$

□

At first sight, it looks as though BC1 and BC2 are not very powerful - they tell us when certain events have probability zero or one. But for many applications, in particular when the events are independent, many interesting events can *only* have probability zero or one, because they are tail events (Kolmogorov's 0/1 Law) and knowing whether it's 0 or 1 describes very different behaviours.

Example 2.43. et $(X_n)_{n\geqslant 1}$ be a sequence of independent, identically distributed (i.i.d.) random variables and let $S_n = \sum_{k=1}^n X_k$. Assume the X_n have mean zero and variance one and set

$$B = \left\{ \limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1 \right\}. \tag{11}$$

Then by Kolmogorov's 0/1-law we have $\mathbb{P}[B] = 0$ or $\mathbb{P}[B] = 1$. It then follows (under the assumption that $\mathbb{E}[|X_n|^{2+\alpha}] < \infty$ for an $\alpha > 0$) by an application of Borel-Cantelli that $\mathbb{P}[B] = 1$. We are not giving the proof, because the proof is a bit subtle and lengthy. Howver, it is a classic example that tell us something very non-trivial about the growth of S_n.

3 Modes of Convergence

You may at this point be feeling a little confused: If (X_n) is a sequence of i.i.d. random variables with mean 0 and variance 1 and $S_n = X_1 + \cdots + X_n$ is a random walk (generalizing the simple symmetric random walk that we encountered in our first lecture) , the Central limit theorem tells us that

$$\mathbb{P}\left[\frac{X_1+\cdots+X_n}{\sqrt{n}} \leqslant a\right] = \mathbb{P}\left[\frac{S_n}{\sqrt{n}} \leqslant a\right] \overset{n\to\infty}{\longrightarrow} \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x. \tag{12}$$

That is, the fluctuations of S_n at a fixed large time n are typically of order $\sqrt{n}$. On the other hand, the Law of the Iterated Logarithm stated in Example 2.43 tells us that, along almost every sample path, fluctuations of the larger order $\sqrt{n\log\log n}$ occur infinitely often. There is no contradiction: the CLT describes the distribution at individual large times, whereas the LIL describes the long-term behaviour along individual sample paths. This demonstrates that convergence for stochastic processes needs a bit more care than convergence of deterministic sequences. The aim of this section understand the different notions of convergence for random variables.

Definition 3.1 (Modes of convergence)**.** Let $X_1, X_2, \dots$ and X be random variables on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

1. We say that (X_n) converges *almost surely* to X (written $X_n \overset{\text{a.s.}}{\to} X$ or $X_n \to X$ a.s.) if
$$\mathbb{P}[X_n \to X] = \mathbb{P}\Big[\big\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\big\}\Big] = 1.$$
2. We say that (X_n) converges to X *in probability* (written $X_n \overset{\mathbb{P}}{\to} X$) if, for every $\varepsilon > 0$,
$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = \lim_{n\to\infty} \mathbb{P}\big[\big\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\big\}\big] = 0.$$
3. Suppose that X and all X_n have finite pth moments for some real number $p > 0$, i.e., $\mathbb{E}[|X|^p], \mathbb{E}[|X_n|^p] < \infty$. We say that X_n converges to X in L^p (or in *pth moment*) (written $X_n \overset{L^p}{\to} X$) if
$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0.$$
4. Let F and F_n denote the distribution functions of X and X_n respectively. We say that X_n converges to X *in distribution* (written $X_n \overset{d}{\to} X$ or $X_n \overset{\mathscr{L}}{\to} X$) if
$$\lim_{n\to\infty} F_n(x) = F(x)$$
for every $x \in \mathbb{R}$ *at which* F *is continuous*.

These notions of convergence capture different ways in which random variables can approach a limit. *Almost sure convergence* means that, for almost every outcome ω, the sequence $X_n(\omega)$ converges to $X(\omega)$, whereas *convergence in probability* only requires that large deviations from X become increasingly unlikely. *Convergence in L^p* controls the size of the error $X_n - X$ through its p-th moment, thus taking into account not just how likely large errors are, but also their magnitude. Finally, *convergence in distribution* is weaker still: it only concerns the distributions of X_n, without requiring the random variables themselves to become close to X. These distinctions are important because rare but large deviations may prevent one type of convergence even when another holds.

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

$\Uparrow$

Convergence in L^p

We want to find relations between these modes of convergence. Let's first look at a few examples.

Example 3.2 (Convergence in distribution does not imply convergence in probability)**.** Let X be a random variable with
$$\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = \frac{1}{2},$$
and define $X_n = -X$ for every $n \geqslant 1$. Since X_n and X have the same distribution, we have $X_n \overset{d}{\to} X$. However,
$$\mathbb{P}(|X_n - X| > 1) = 1$$
for every n, so X_n does not converge to X in probability.

Example 3.3 (Almost sure convergence or convergence in probability does not imply L^p convergence)**.** Fix $p > 0$ and consider the probability space $((0,1), \mathscr{B}((0,1)), m)$ with m denoting the Lebesgue measure. Define

$$X_n(\omega) = n^{1/p} \mathbf{1}_{(0,1/n)}(\omega).$$

Then $X_n(\omega) \to 0$ for every $\omega \in (0,1)$, so $X_n \to 0$ almost surely. Also, for every $\varepsilon > 0$ and all sufficiently large n,

$$\mathbb{P}(|X_n| > \varepsilon) = \frac{1}{n} \to 0.$$

Thus $X_n \to 0$ in probability. However,

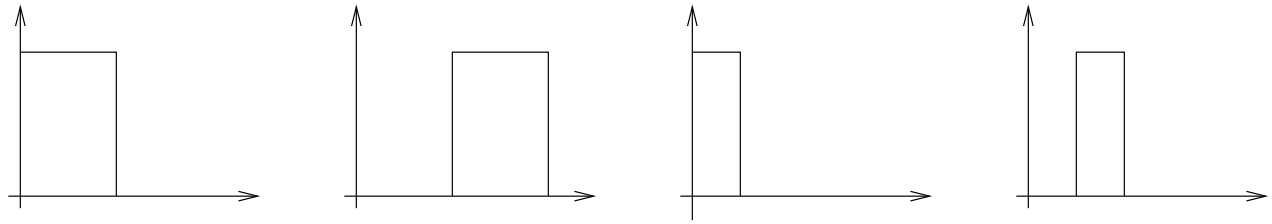
$$\mathbb{E}[|X_n|^p] = n \cdot m((0,1/n)) = 1$$

for every n. Hence X_n does not converge to zero in L^p.

Example 3.4 (Convergence in probability or L^p convergence do not imply a.s. convergence)**.** Take $[0,1]$ 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. and $X_n(\omega) = 0$ i.o, so $X_n \not\to 0$ a.s., but

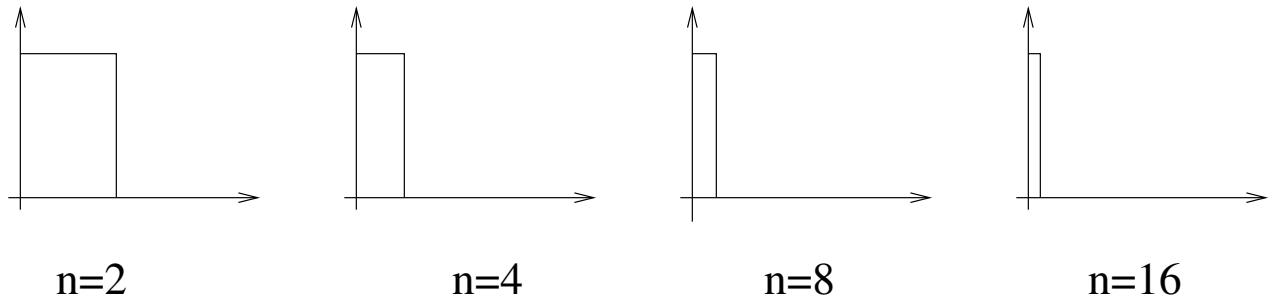
$$\mathbb{P}[X_n \neq 0] = \frac{1}{2^m} \to 0 \quad \text{as } n \to \infty,$$

so $X_n \xrightarrow{\mathbb{P}} 0$. Similarly, for every $p > 0$ we have

$$\mathbb{E}[|X_n|^p] = 2^{-m} \to 0,$$

where $n = 2^m + k$ with $0 \leqslant k < 2^m$. Hence $X_n \to 0$ in L^p for every $p > 0$.

Example 3.5. (Subsequence Convergence) A simple observation related to the above Examples (to which we return later in full generality) is that by switching to a subsequence we can sometimes strengthen the mode of converggence. For $(X_{2^n})_{n \geqslant 1}$, we have



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

It turns out that this is a general phenomenon.

Theorem 3.6 (Convergence in Probability and a.s. Convergence)**.** *Let $X_1, X_2, \dots$ and X be random variables on $(\Omega, \mathscr{F}, \mathbb{P})$.*

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

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

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

1. For every $\varepsilon > 0$,

$$\mathbf{1}_{A_{n,\varepsilon}} \to 0 \quad \text{almost surely.}$$

Since these indicators are bounded by 1, the dominated convergence theorem yields

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{E}\left[\mathbf{1}_{\{|X_n - X| > \varepsilon\}}\right] \longrightarrow 0.$$

Hence $X_n \to X$ in probability.

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

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

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

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

Lemma 3.7 (Markov's inequality)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and X a* non-negative *random variable. Then, for each $\lambda > 0$*

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

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

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

□

Corollary 3.8 (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. By Markov's inequality,

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

□

Example 3.9.

- The most familiar special case is given by taking $\phi(x) = x^2$ on $[0,\infty)$ and applying the result to $Y = |X - \mathbb{E}[X]|$, giving
$$\mathbb{P}\big[|X - \mathbb{E}[X]| \geqslant t\big] \leqslant \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{t^2} = \frac{\mathrm{Var}[X]}{t^2}$$
for $t > 0$.
- Corollary 3.8 is also often applied with $\phi(x) = e^{\theta x}$, $\theta \geqslant 0$, to obtain
$$\mathbb{P}[X \geqslant \lambda] \leqslant e^{-\theta\lambda}\mathbb{E}[e^{\theta X}].$$
This is especially useful since the expectation on the right hand side is the moment-generating function of X which is known explicitly for many distributions. This can be further strengthend by optimizing over θ.
- Let $(X_n)_{n\geqslant 1}$ be i.i.d. random variables (on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$) with mean μ and variance $\sigma^2 < \infty$. Set
$$\overline{X}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$
Then $\overline{X}(n) \to \mu$ in probability as $n \to \infty$. This is known as the weak law of large numbers.

  Proof. Since $\mathbb{E}[\overline{X}(n)] = n^{-1}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mu$ we have by Chebyshev's inequality,
$$\mathbb{P}[|\overline{X}(n) - \mu| > \varepsilon] \leqslant \frac{\mathrm{Var}[\overline{X}(n)]}{\varepsilon^2} = \varepsilon^{-2}n^{-2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \varepsilon^{-2}n^{-2}\sum_{i=1}^{n}\mathrm{Var}[X_i] = \frac{\sigma^2}{\varepsilon^2 n} \to 0.$$
□

The only implication that we still have to show is that convergence in probability implies convergence in distribution and this is a problem on one of the exercise sheets.

3.1 $\mathscr{L}^p$ spaces

Given a probability space $(\Omega, \mathbb{P}, \mathscr{F})$, the set of all random variables carried on this probability space can be a large set. We add some structure by using integrability of random variables.

Definition 3.10 ($\mathscr{L}^p$ spaces)**.** Let $(\Omega, \mathbb{P}, \mathscr{F})$ be a probability space. We define

$$\mathscr{L}^0 = \{X : X \text{ is a real-valued random variable on } (\Omega, \mathscr{F}, \mathbb{P})\}.$$

For $0 < p < \infty$, let

$$\mathscr{L}^p(\Omega, \mathscr{F}, \mathbb{P}) = \{X : X \text{ is a random variable on } (\Omega, \mathscr{F}, \mathbb{P}) \text{ with } \|X\|_p < \infty\} \text{ where } \|X\|_p = \left(\mathbb{E}[|X|^p]\right)^{1/p}.$$

For $p=\infty$ we define

$$\mathscr{L}^\infty(\Omega,\mathscr{F},\mathbb{P})=\{X : X \text{ is a random variable on } (\Omega,\mathscr{F},\mathbb{P}) \text{ such that } \|X\|_\infty<\infty\},$$

where

$$\|X\|_\infty = \operatorname{ess\,sup}|X| = \inf\{M\geqslant 0 : \mathbb{P}(|X|\leqslant M)=1\}.$$

Often we simply write $\mathscr{L}^p$ instead of $\mathscr{L}^p(\Omega,\mathscr{F},\mathbb{P})$.

The set $\mathscr{L}^0$ is clearly a vector space. For $p>0$, note that 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*.

In probabilistic terms, we use the p-th moment to compare random variables and impose some structure on the space of random variables. How are these $\mathscr{L}^p$ spaces related? Since the variance of a random variable is non-negative we get $0\leqslant Var(X)=\mathbb{E}[X^2]-\mathbb{E}[X]^2$. Hence, $\mathscr{L}^1\supseteq\mathscr{L}^2$. Is this a special property of the square function, or part of a more general principle? The key observation is that the function $x\mapsto x^2$ is convex. This leads to **Jensen's inequality**, which turns out to be a fundamental tool for comparing moments and studying L^p spaces.

Definition 3.11 (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 $(0,\infty)$. More generally, if f is twice differentiable, then f is convex iff $f''\geqslant 0$: to see this note that for any $x,y\in I$, the mean value theorem gives

$$f(y)-f(x)=f'(c)(y-x)$$

for c in the open interval between x and y. Using that f' is increasing we see that if $y>x$, then $f'(c)\geqslant f'(x)$; if $y<x$,then $f'(c)\leqslant f'(x)$, but $y-x<0$. Thus, in either case,

$$f(y)\geqslant f(x)+f'(x)(y-x).$$

Equation 3.1 is also called the *supporting-line property* and is equivalent to Definition 3.11: apply the supporting line inequality with $z=(1-t)x+ty$ for $t\in[0,1]$ as the base point, first to x and then to y to get

$$f(x)\geqslant f(z)+f'(z)(x-z) \text{ and } f(y)\geqslant f(z)+f'(z)(y-z). \tag{13}$$

Multiplying the first inequality by $1-t$, the second by t, and adding yields

$$\begin{aligned}(1-t)f(x)+tf(y) &\geqslant f(z)+f'(z)((1-t)x+ty-z)\\ &= f(z)=f((1-t)x+ty),\end{aligned}$$

which is precisely the definition of convexity. To see that convexity implies $f''\geqslant 0$, use convexity to show

$$0\leqslant \frac{f(x+h)-2f(x)+f(x-h)}{h^2}$$

and that the right hand side converges to $f''(x)$ as $h\to 0$.

Theorem 3.12 (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*

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

The expectation on the left is well-defined, possibly equal to $+\infty$.

Note that we make no assumptions on the interval such as boundedness/open/closed etc., i.e. $I = \mathbb{R}$, $I = (0,\infty)$, or $I = (a,b]$ are all fine. Perhaps the nicest proof of Theorem 3.12 rests on the following geometric lemma.

Lemma 3.13 (Supporting-line characterisation of convexity)**.** *Let $f : I \to \mathbb{R}$, where $I \subseteq \mathbb{R}$ is an interval. Then f is convex if and only if, for every interior point m of I, there exists $a \in \mathbb{R}$ such that*

$$f(x) \geqslant f(m) + a(x-m) \qquad \text{for all } x \in I.$$

If f is differentiable at m, then necessarily $a = f'(m)$.

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

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

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

For the reverse implication we argue as before when discussed how twice differentiable functions are convex but replace in (13) f' by a. □

Proof of Theorem 3.12. If $\mathbb{E}[X]$ is not an interior point of I then it is an endpoint, and X must be almost surely constant, so the inequality is trivial. Otherwise, setting $m = \mathbb{E}[X]$ in the previous lemma we have

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

Now take expectations to recover

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

as required. □

Remark. Jensen's inequality only works for probability measures, but often one can exploit it to prove results for finite measures by first normalizing. For example, suppose that μ is a finite measure on $(\Omega, \mathscr{F})$ with $\mu(\Omega) > 0$, and define ν by $\nu(A) = \mu(A)/\mu(\Omega)$. Then

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

A simple application of Jensen's inequality tells us that the $\mathscr{L}^p$ spaces are nested.

Lemma 3.14 (Inclusion of $\mathscr{L}^p$ spaces)**.** *Let $0 < p < q \leqslant \infty$. Then*

$$\|X\|_p \leqslant \|X\|_q \quad \text{for all } X \in \mathscr{L}^q.$$

Hence, $\mathscr{L}^q \subseteq \mathscr{L}^p$ and convergence in L^q implies convergence in L^p.

Proof. First suppose $q < \infty$. Since $p < q$, we have $|X|^p \leqslant 1 + |X|^q$, so $X \in \mathscr{L}^p$. The function $f(x) = -x^{p/q}$ is convex on $[0,\infty)$. Applying Jensen's inequality to $|X|^q$ gives

$$-\mathbb{E}|X|^p \geqslant -(\mathbb{E}|X|^q)^{p/q},$$

and hence $\|X\|_p \leqslant \|X\|_q$.

If $q = \infty$, then $|X| \leqslant \|X\|_\infty$ a.s., so

$$\mathbb{E}|X|^p \leqslant \|X\|_\infty^p,$$

and again $\|X\|_p \leqslant \|X\|_\infty$. In both cases, $\mathscr{L}^q \subseteq \mathscr{L}^p$. □

We have shown that

$$\mathscr{L}^\infty \subseteq \mathscr{L}^q \subseteq \mathscr{L}^p \subseteq \mathscr{L}^0, \qquad 0 < p < q < \infty.$$

Larger values of p impose stronger restrictions on the tails of a random variable: $\mathscr{L}^1$ consists of random variables with finite absolute expectation, $\mathscr{L}^2$ of those with finite second moment, and $\mathscr{L}^\infty$ of those that are essentially bounded. Simple examples show that these inclusions are generally strict.

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

Theorem 3.15. *Let $1 \leqslant p, q \leqslant \infty$ be such that*

$$\frac{1}{p} + \frac{1}{q} = 1$$

with the convetion that $\frac{1}{\infty} = 0$. Suppose $X, Y \in \mathscr{L}^p$ and $Z \in \mathscr{L}^q$. Then

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

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

If $X = 0$ a.s. then there is nothing to show. Similarly, the endpoint cases are immediate: if $p = 1$ and $q = \infty$, then

$$\mathbb{E}[|XZ|] \leqslant \|Z\|_\infty \mathbb{E}[|X|] = \|X\|_1 \|Z\|_\infty.$$

and case $p = \infty$ and $q = 1$ follows by symmetry.

It remains to deal with the case when $1 < p, q < \infty$ and X is not equal to 0 a.s. Here, the idea is to absorb the factor $|X|^p$ into a new probability measure $\mathbb{Q}$, so that Hölder's inequality becomes an instance of Jensen's inequality. Indeed, since $(p-1)q = p$, the choice $\mathbb{Q}(A) = \mathbb{E}[|X|^p \mathbf{1}_A]/\|X\|_p^p$ gives a new probability measure on $(\Omega, \mathscr{F})$ and if we define a random variable $\bar{Z} := \mathbf{1}_{|X|>0}|Z|/|X|^{p-1}$ this transforms the product $\mathbb{E}|XZ|$ into a first moment of $\bar{Z}$ under $\mathbb{Q}$, while its q-th moment involves only $|Z|^q$. More formally: apply Jensen's inequality with $f(x) = x^q$ to the bounded random variable $\bar{Z} \wedge n$ and then letting $n \to \infty$ by MCT, we obtain

$$\left(\int \bar{Z} \, \mathrm{d}\mathbb{Q} \right)^q \leqslant \int \bar{Z}^q \, \mathrm{d}\mathbb{Q}.$$

Hence,

$$\begin{aligned}(\mathbb{E}[|XZ|])^q &= \|X\|_p^{pq}\left(\int \bar{Z}\,\mathrm{d}\mathbb{Q}\right)^q \\ &\leqslant \|X\|_p^{pq}\int \bar{Z}^q\,\mathrm{d}\mathbb{Q} \qquad \text{(Jensen)} \\ &= \|X\|_p^q\mathbb{E}[|Z|^q\mathbf{1}_{|X|>0}] \\ &\leqslant \|X\|_p^q\mathbb{E}[|Z|^q].\end{aligned}$$

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

For Minkowski's inequality, the endpoint cases follow again immediately: for $p=1$, the triangle inequalit follows directly from $|X+Y|\leqslant|X|+|Y|$ by taking expectations; for $p=\infty$, we have

$$|X+Y|\leqslant|X|+|Y|\leqslant\|X\|_\infty+\|Y\|_\infty \qquad \text{a.s.},$$

which gives $\|X+Y\|_\infty\leqslant\|X\|_\infty+\|Y\|_\infty$.

For $1<p,q<\infty$ 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

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

If $c=0$, the inequality is trivial. Otherwise, dividing by c gives the result since $1-1/q=1/p$. □

Remark (Change of measure)**.** The change the measure used in the proof above is an important technique in probability and stochastic analysis. The basic idea is to replace a probability measure by another, suitably chosen probability measure under which a difficult calculation becomes easier. We will return to this when discuss the Radon-Nikodym theorem.

Theorem 3.16. *For every $0<p\leqslant\infty$, every Cauchy sequence (X_n) in $\mathscr{L}^p$ converges to some $X\in\mathscr{L}^p$ in the sense that $\|X_n-X\|_p\to 0$.*

Proof. First let $0<p<\infty$ and put $q=\min\{p,1\}$. Choose a subsequence (X_{k_n}) such that

$$\|X_{k_{n+1}}-X_{k_n}\|_p\leqslant 2^{-n}.$$

Since $q\leqslant p$, Jensen's inequality yields $\mathbb{E}[Y^q]=\mathbb{E}[Y^{p\frac{q}{p}}]\leqslant\mathbb{E}[Y^p]^{\frac{q}{p}}=\|Y\|_p^q$ and applied to $Y=|X_{k_{n+1}}-X_{k_n}|$ we get

$$\sum_{n=1}^{\infty}\mathbb{E}[|X_{k_{n+1}}-X_{k_n}|^q]\leqslant\sum_{n=1}^{\infty}2^{-nq}<\infty.$$

Hence, by MCT we get

$$\mathbb{E}\left[\sum_{n=1}^{\infty}|X_{k_{n+1}}-X_{k_n}|^q\right]=\sum_{n=1}^{\infty}\mathbb{E}[|X_{k_{n+1}}-X_{k_n}|^q]<\infty.$$

Since a nonnegative random variable with finite expectation is finite a.s., we obtain

$$\sum_{n=1}^{\infty}|X_{k_{n+1}}-X_{k_n}|^q<\infty \qquad \text{a.s.}$$

Now, if $\sum_n a_n^q<\infty$ for $a_n\geqslant 0$, then $a_n\to 0$, so $a_n\leqslant 1$ eventually. Since $0<q\leqslant 1$, we have $a_n\leqslant a_n^q$ whenever $a_n\leqslant 1$. Hence $\sum_n a_n<\infty$. Applying this to the successive differences gives

$$\sum_{n=1}^{\infty}|X_{k_{n+1}}-X_{k_n}|<\infty \qquad \text{a.s.}$$

Consequently, the telescoping series

$$X_{k_n} = X_{k_1} + \sum_{j=1}^{n-1} (X_{k_{j+1}} - X_{k_j})$$

converges absolutely a.s., and therefore $X_{k_n} \to X$ a.s. for some random variable X.

By Fatou's lemma,

$$\mathbb{E}[|X_r - X|^p] \leqslant \liminf_{n\to\infty} \mathbb{E}[|X_r - X_{k_n}|^p].$$

The Cauchy property implies that the right-hand side tends to zero as $r \to \infty$. Thus $\|X_r - X\|_p \to 0$, and $X \in \mathscr{L}^p$ since

$$|X|^p \leqslant C_p(|X - X_1|^p + |X_1|^p), \qquad C_p = \max\{1, 2^{p-1}\}.$$

For $p = \infty$, we argue as before: choose a subsequence (X_{k_n}) such that

$$\|X_{k_{n+1}} - X_{k_n}\|_\infty \leqslant 2^{-n}.$$

Outside a single null set, the successive differences are bounded by 2^{-n}, so their telescoping series converges absolutely. Thus $X_{k_n} \to X$ almost surely for some finite random variable X. Since $X_{k_n} \to X$ almost surely, letting $n \to \infty$ yields

$$\|X_r - X\|_\infty \leqslant \varepsilon \qquad (r \geqslant N),$$

and hence $X_r \to X$ in $\mathscr{L}^\infty$. □

Remark. For $1 \leqslant p \leqslant \infty$, $\|.\|_p$ satisfies the triangle inequality and homogenity ($\|cX\|_p = |c|\|X\|_p$). However, $(\mathscr{L}^p, \|.\|_p)$ is not a normed space since $\|X\|_p = 0$ only implies $X = 0$ up to a null set. To get rid of this problem, we simply quotient by the equivalence relation of a.s. equality. Formally, we define for $0 \leqslant p \leqslant \infty$,

$$L^p := \mathscr{L}^p/\sim \text{ where } X \sim Y \text{ iff } \mathbb{P}(X = Y) = 1.$$

Note that the elements of L^p are now equivalence classes,

$$[X] = \{Y \in \mathscr{L}^p : Y = X \text{ a.s.}\}.$$

The results we proved show that for $1 \leqslant p < \infty$, $(L^p, \|.\|_p)$ is a so-called *Banach space*, which is a vector space L^p with a norm $\|.\|_p$ that makes it complete; for $0 < p < 1$, $\|.\|_p$ fails the triangle inequality so it is not a norm, but one can show that (L^p, d_p) is a complete metric space with $d_p([X],[Y]) = \mathbb{E}[|X - Y|^p]$. Similarly, (L^0, d_0) is a complete metric space where $d_0([X],[Y]) = \mathbb{E}[\min(1, |X - Y|)]$. Some you of might take the course on functional analysis in which L^p spaces are studied in their own and systematic way. However, from the probabilistic point of view, we prefer to work with actual functions, i.e. with random variables and not equivalence classes. One is reason is that when we have a large family $(X_t)_{t\geqslant 0}$ of random variables (i.e. measurable functions), changing each of them on a null set may actually do a lot of harm!

Remark. A word of caution: in this course I am concise about the difference between L^p and $\mathscr{L}^p$, resp. the difference between the equivalence class $[X]$ and the random variable X, but this is not always the case in the literature, especially in more advanced books, about stochastic processes, where the difference between an equivalence class and an element of that equivalence class is blurred; e.g. you find statements like $X \in L^p$. This is justified because integration, expectations and L^p norms are unchanged by modifications on null sets But it's good to be aware of this, especially when learning the material so I tried to err on the side of being too pedantic in the lecture notes (except for the Deep Dives).

Deep Dive

The spaces $L^p = L^p(\Omega, \mathscr{F}, \mathbb{P})$ are fundamental not only in probability theory, but also in *functional analysis*, where we study infinite-dimensional vector spaces equipped with notions of distance, continuity and convergence. Our discussion focused on probabilty spaces, but for general measure spaces we again get a Banach space $L^p(\Omega, \mathscr{F}, \mu)$ for $1 \leqslant p \leqslant \infty$; for $0 < p < 1$ the L^p spaces are complete but $\|\cdot\|_p$ fails the triangle inequality so they are not Banach spaces and only have a quasi-norm.

An important special case arises when we replace integration by summation. Consider the measure space $(\mathbb{N}, \mathscr{P}(\mathbb{N}), \mu)$, where μ is the *counting measure*, defined by

$$\mu(A) = \#A, \qquad A \subseteq \mathbb{N}.$$

For a function $f : \mathbb{N} \to \mathbb{R}$, we then have

$$\int_{\mathbb{N}} |f|^p \, d\mu = \sum_{n=1}^{\infty} |f(n)|^p.$$

Thus, identifying a function $f : \mathbb{N} \to \mathbb{R}$ with the sequence $(f(n))_{n \geqslant 1}$, we obtain

$$\ell^p = L^p(\mathbb{N}, \mathscr{P}(\mathbb{N}), \mu), \qquad 1 \leqslant p < \infty.$$

You probably have encounterd these sequence spaces in Prelims and Metric spaces sheets. In particular,

$$\ell^p = \left\{ (x_n)_{n \geqslant 1} : \sum_{n=1}^{\infty} |x_n|^p < \infty \right\}, \qquad \|(x_n)\|_p = \left(\sum_{n=1}^{\infty} |x_n|^p \right)^{1/p}.$$

Since counting measure is infinite, the inclusions between these spaces differ from those on a probability space: for $1 \leqslant p < q < \infty$,

$$\ell^p \subseteq \ell^q, \qquad L^q(\mathbb{P}) \subseteq L^p(\mathbb{P}).$$

For example, $(1/n)_{n \geqslant 1}$ belongs to ℓ^2 but not to ℓ^1. Finally, let me mention that the case $p = 2$ is particularly important since L^2 (resp. ℓ^2) are *Hilbert spaces*: their norms arise from inner products, namely

$$\langle X, Y \rangle = \mathbb{E}[XY],$$

(resp. $\langle x, y \rangle = \sum_{n=1}^{\infty} x_n y_n$ for ℓ^2). Hilbert spaces are even nicer than Banach spaces since notions familiar from Euclidean geometry, such as orthogonality and projections, extend to these (in general infinite-dimensional) spaces. In probability, this provides a geometric interpretation of conditional expectation as an orthogonal projection in L^2, which we shall encounter later in this course.

Another beautiful aspect of L^p spaces is their *duality*. For $1 < p < \infty$, let q be the conjugate exponent, defined by

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then every continuous linear functional $\Lambda : L^p \to \mathbb{R}$ can be represented uniquely in the form

$$\Lambda(X) = \mathbb{E}[XY]$$

for some $Y \in L^q$. Moreover,

$$\|\Lambda\| = \|Y\|_q.$$

In other words, the continuous dual space of L^p can itself be identified with L^q; formally $(L^p)' \simeq L^q$. This is closely related to Hölder's inequality, which guarantees that $X \mapsto \mathbb{E}[XY]$ is indeed a continuous linear functional on L^p whenever $Y \in L^q$. The converse, that every continuous linear functional arises in this way, is a deeper result. The case $p = 2$ is particularly special: L^2 is its own dual, reflecting its Hilbert space structure.

The endpoint cases $p = \infty$ and $q = 1$ behave somewhat differently from $1 < p < \infty$. Although every $Y \in L^1$ defines a continuous linear functional on L^∞ through $X \mapsto \mathbb{E}[XY]$, not every continuous linear functional arises in this way. The reason is that the dual of L^∞ can also contain functionals corresponding to finitely additive measures, whereas integration against an L^1 function always gives a countably additive measure. Thus, in general, L^1 is strictly contained in the dual of L^∞.

Finally, let me reemphasize that L^p spaces are central to modern analysis, including the study of differential equations, Fourier analysis, optimisation and stochastic processes.

Deep Dive

The spaces L^p provide a natural way of measuring the size of a random variable through its moments $\mathbb{E}[|X|^p]$. However, moments are not the only way to quantify the size of a random variable. Depending on the application, we may want to impose different integrability or tail conditions.

A particularly useful generalisation is given by *Orlicz spaces*. The idea is to replace the power function $x \mapsto |x|^p$ by a more general convex function $\Phi : [0,\infty) \to [0,\infty)$ satisfying $\Phi(0) = 0$ and $\Phi(x) \to \infty$ as $x \to \infty$ (such a function is called a Young function). We define

$$L^\Phi = \left\{ X : \mathbb{E}\left[\Phi\left(\frac{|X|}{c}\right)\right] < \infty \text{ for some } c > 0 \right\}.$$

A natural norm on this space is the *Luxemburg norm*:

$$\|X\|_\Phi = \inf\left\{ c > 0 : \mathbb{E}\left[\Phi\left(\frac{|X|}{c}\right)\right] \leqslant 1 \right\}.$$

When $\Phi(x) = x^p$, for $p \geqslant 1$, we recover the usual L^p spaces and their norms. The advantage of Orlicz spaces is that they allow us to impose more specific conditions on the tails of random variables. For example, choosing

$$\Phi(x) = e^x - 1$$

leads to exponential integrability, while

$$\Phi(x) = e^{x^2} - 1$$

is useful for describing random variables with Gaussian-type tail behaviour. Such spaces play an important role in probability theory, concentration inequalities and the study of stochastic processes.

Another possibility is to consider *weighted L^p spaces*. Given a measurable weight $w : \Omega \to [0,\infty]$, we define

$$L^p(w\,d\mathbb{P}) = \{X : \mathbb{E}[w|X|^p] < \infty\},$$

with norm

$$\|X\|_{L^p(w\,d\mathbb{P})} = (\mathbb{E}[w|X|^p])^{1/p}, \qquad p \geqslant 1.$$

The idea is to assign different weights to different regions of the underlying probability space. For example, in a financial application, one might want to penalise errors more heavily in certain market scenarios, or in a

statistical problem one might want to give greater importance to observations from a particular region of the sample space. This is again related to the change of measure idea we encountered in the proof of Hölder's inequality.

There are also other natural ways to put structure on spaces of random variables. For example, L^∞ measures the size of a random variable through its essential supremum, while L^0 is equipped with the metric

$$d(X,Y) = \mathbb{E}[\min\{1, |X-Y|\}],$$

which induces convergence in probability.

The general lesson is that there is no single universally best space of random variables. The choice of function space determines which random variables are admissible, which notion of convergence is natural, and which analytical tools are available.

4 Conditional Expectation

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

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

Before we study the above question in full generality, let us try to gain some intuition in two examples where $\mathscr{G}$ is relatively simple.

Example 4.1. Let $X \in \mathscr{L}^2(\Omega, \mathscr{F}, \mathbb{P})$. Let $B \in \mathscr{F}$ with $\mathbb{P}(B) \in (0,1)$ and note that σ-algebra $\sigma(B) := \sigma(\mathbf{1}_B) = \{\emptyset, \Omega, B, B^c\}$. Consider the optimization problem

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

where the inf is taken over all square-integrable and $\sigma(B)$-measurable random variables. The requirement of $\sigma(B)$-measurability means that the inf is taken over random variables of the form

$$a\mathbf{1}_B + b\mathbf{1}_{B^c} \text{ for } a, b \in \mathbb{R}.$$

Hence,

$$\mathbb{E}[(X-Z)^2] = \mathbb{E}[X^2] - 2a\mathbb{E}[X\mathbf{1}_B] + a^2\mathbb{P}(B) - 2b\mathbb{E}[X\mathbf{1}_{B^c}] + b^2\mathbb{P}(B^c)$$

and minimising separately over a and b shows that the infimum is attained by the random variable

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

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

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

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

$$\mathbb{E}[X|B] = \int_\Omega X(\omega) d\mathbb{P}(\omega|B) = \frac{\int_B X d\mathbb{P}}{\mathbb{P}(B)} = \frac{\mathbb{E}[X1_B]}{\mathbb{P}(B)}$$

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

$$\mathbb{E}[X \mid \sigma(B)](\omega) = \mathbb{E}[X \mid B]\mathbf{1}_B(\omega) + \mathbb{E}[X \mid B^c]\mathbf{1}_{B^c}(\omega) \tag{14}$$

which hopefully makes the notation $\mathbb{E}[X \mid \sigma(B)]$ more intuitive. For later purposes note that the above immediately implies

$$\mathbb{E}[\mathbb{E}[X|\sigma(B)]\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G] \quad \forall G \in \sigma(B) = \{\emptyset, \Omega, B, B^c\}.$$

Similar explicit formulas for $\mathbb{E}[X \mid \mathscr{G}]$ can be derived if we replace $\sigma(B)$ by another, relatively simple, σ-algebra $\mathscr{G}$ in the minimization problem; for example, if $G_1, \cdots, G_n$ forms a partition of Ω and $\mathbb{P}(G_i) > 0$, then the same argument as before shows that a solution of the minimization problem is given as

$$\mathbb{E}[X \mid \mathscr{G}] = \sum_{i=1}^{n} \mathbb{E}[X \mid G_i]\mathbf{1}_{G_i}, \text{ for } \mathscr{G} = \sigma(G_1, \ldots, G_n).$$

Finally, we note that

$$\mathbb{E}[\mathbb{E}[X \mid \sigma(B)]\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G] \text{ for all } G \in \sigma(B). \tag{15}$$

which is immediate via the formula (14). Informally, it expresses that the random variable $\mathbb{E}[X \mid \mathscr{G}]$ matches locally the expectation of X; here "locally" means on the finest level that the σ-algebra $\mathscr{G}$ allows us to look at. This equality will be the key for the general definition when we condition on σ-algebras $\mathscr{G}$ that are not as simple as $\sigma(B)$.

Partitions generate an especially nice subset of σ-algebras. Let's discuss the more general case $\mathscr{G} = \sigma(Y)$ before we give our final definition.

Example 4.2. Let X be an integrable random variable and Y another real-valued random variable on $(\Omega, \mathscr{F}, \mathbb{P})$. Recall that $\sigma(Y)$ is generated by the events $\{Y \in B\}$, where $B \in \mathscr{B}(\mathbb{R})$.

Motivated by the finite-partition formula above, we might try to define

$$\mathbb{E}[X \mid \sigma(Y)] \stackrel{?}{=} \sum_{y \in \mathbb{R}} \mathbb{E}[X \mid Y = y]\mathbf{1}_{\{Y=y\}}, \qquad \mathbb{E}[X \mid Y = y] = \frac{\mathbb{E}[X\mathbf{1}_{\{Y=y\}}]}{\mathbb{P}(Y = y)}. \tag{16}$$

This works when Y takes finitely or countably many values of positive probability, with the sum restricted to those values. In particular, choosing $Y = \mathbf{1}_B$ recovers our previous example, since $\sigma(Y) = \{\emptyset, B, B^c, \Omega\}$ and

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

provided $0 < \mathbb{P}(B) < 1$.

But what if Y has a continuous distribution? In that case, (16) breaks down: firstly, $\mathbb{P}(Y = y) = 0$ for every $y \in \mathbb{R}$, so the conditional expectations on the right-hand side are undefined. Secondly, the proposed sum is uncountable and cannot be interpreted as an ordinary countable sum. Putting rigour aside for a moment, let us multiply the "formula" (16) by $\mathbf{1}_{\{Y \in B\}}$ and take expectations. This suggests that, however we define the elusive $\mathbb{E}[X \mid \sigma(Y)]$, it should be an integrable, $\sigma(Y)$-measurable random variable satisfying

$$\mathbb{E}[\mathbb{E}[X \mid \sigma(Y)]\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G] \qquad \text{for all } G \in \sigma(Y). \tag{17}$$

The key observation is that, unlike (16), this identity remains meaningful even when the events $\{Y = y\}$ have probability zero. We shall soon show that it determines $\mathbb{E}[X \mid \sigma(Y)]$ uniquely up to almost-sure equality, and (with some work) that such a random variable indeed exists.

This brings us to the general definition of conditional expectation.

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

$$\mathbb{E}[\mathbb{E}[X \mid \mathscr{G}]\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G] \quad \text{for all } G \in \mathscr{G}. \tag{18}$$

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

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

If the σ-algebra $\mathscr{G}$ is generated by a random variable Y, that is $\mathscr{G} = \sigma(Y)$, then we also use the notation $\mathbb{E}[X \mid Y]$ resp. $\mathbb{P}(A \mid Y)$ for $\mathbb{E}[X \mid \mathscr{G}]$ resp. $\mathbb{P}(A \mid \mathscr{G})$. We refer to (18) as the *defining relation of conditional expectation*.

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

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

Proof of Uniqueness. Let Z_1, Z_2 be two conditional expectations of X given $\mathscr{G}$. Let $G := \{Z_1 > Z_2\}$ and note that $G \in \mathscr{G}$ as Z_1, Z_2 are $\mathscr{G}$-measurable. B y definition, $\mathbb{E}[Z_1\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G] = \mathbb{E}[Z_2\mathbf{1}_G]$ so that $\mathbb{E}[(Z_1 - Z_2)\mathbf{1}_G] = 0$. But $(Z_1 - Z_2)\mathbf{1}_G \geqslant 0$ a.s. and hence $(Z_1 - Z_2)\mathbf{1}_G = 0$ a.s., i.e., $\mathbb{P}(G) = 0$ since $Z_1 - Z_2 > 0$ on G. Swapping Z_1 and Z_2, we also have $\mathbb{P}(Z_2 > Z_1) = 0$ and hence $Z_1 = Z_2$ a.s. □

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

$$\nu(A) := \int_A f d\mu. \tag{19}$$

Is there a converse, that is can I write any measure in the form (19)? Well, not every measure can arise in this way, since $\mu(A) = 0$ implies $\nu(A) = 0$. The Radon–Nikodym theorem guarantees that this condition is ready sufficient if we assume that the measures are σ-finite.

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

$$\mu(A) = 0 \textit{ implies } \nu(A) = 0 \quad \textit{for all } A \in \mathscr{F}, \tag{20}$$

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

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

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

The requirement (20) is called *absolutely continuity of ν with respect to μ* and the notation is $\nu \ll \mu$; the function f is also called the Radon-Nikodym derivative.

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

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

Proof of Existence. First suppose $X \geqslant 0$ is integrable. Define a measure ν on $(\Omega, \mathscr{G})$ by

$$\nu(G) = \mathbb{E}[X\mathbf{1}_G] = \int_G X \, \mathrm{d}\mathbb{P}, \qquad G \in \mathscr{G}.$$

Since X is integrable, ν is finite. Moreover, $\nu \ll \mathbb{P}|_{\mathscr{G}}$, since $\mathbb{P}(G) = 0$ implies $\nu(G) = 0$. By the Radon–Nikodym theorem, there exists a non-negative $\mathscr{G}$-measurable function Z such that

$$\nu(G) = \int_G Z \, d(\mathbb{P}|_{\mathscr{G}}) = \int_G Z \, \mathrm{d}\mathbb{P} \qquad \text{for all } G \in \mathscr{G}.$$

The last equality holds because Z is $\mathscr{G}$-measurable and $\mathbb{P}|_{\mathscr{G}}$ is the restriction of $\mathbb{P}$ to $\mathscr{G}$. Taking $G = \Omega$ shows that Z is integrable. Thus Z satisfies the defining properties of $\mathbb{E}[X \mid \mathscr{G}]$.

For general integrable X, apply the preceding argument to X^+ and X^-, and define

$$\mathbb{E}[X \mid \mathscr{G}] = \mathbb{E}[X^+ \mid \mathscr{G}] - \mathbb{E}[X^- \mid \mathscr{G}].$$

Both terms are integrable, and their difference satisfies the defining properties by linearity. □

Remark. In Example 4.1, we derived explicit formulas for the conditional expectation when $\mathscr{G}$ was generated by a partition. For other σ-algebra$\mathscr{G}$ it is in general not possible to find explicit formulas for the conditional expectation. However, it may help to observe that for non-negative integrable X, if $\mathscr{I}$ is a π-system that generates $\mathscr{G}$, then it is enough to check the defining relation for $G \in \mathscr{I}$. (To see this, apply Theorem 1.13 to the measures $\mathbb{Q}$ and $\int_A Z \, \mathrm{d}\mathbb{P}$ above; it works also for any integrable X, either considering positive and negative parts separately, or a version of Theorem 1.13 for signed measures.)

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

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

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

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

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

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

8. $\left|\mathbb{E}[X \mid \mathcal{G}]\right| \leqslant \mathbb{E}[|X| \mid \mathcal{G}]$ *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 (**??**). We just do some examples.

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

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

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

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

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

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

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

so the defining relation holds. □

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

We have so far defined conditional expectation only for integrable random variables. But as we can proceed as we did when we discussed integration and extend the definition to certain non-integrable random variables, allowing infinite values.

Definition 4.7 (Extended conditional expectation)**.** Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. For a non-negative random variable X, not necessarily integrable, define

$$\mathbb{E}[X \mid \mathcal{G}] := \lim_{n\to\infty} \mathbb{E}[X \wedge n \mid \mathcal{G}].$$

By monotonicity, the limit exists almost surely and may take the value $+\infty$. Moreover, by MCT, it satisfies

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G] \qquad \text{for all } G \in \mathcal{G},$$

where both sides may be infinite. For a general measurable random variable X, write $X = X^+ - X^-$ and define

$$\mathbb{E}[X \mid \mathcal{G}] := \mathbb{E}[X^+ \mid \mathcal{G}] - \mathbb{E}[X^- \mid \mathcal{G}],$$

provided the two terms are not simultaneously infinite on a set of positive probability. The resulting conditional expectation may take values in $[-\infty,\infty]$.

In particular, this agrees with our original definition whenever X is integrable. Just as for in the case of integration (expectation), the only obstruction is the undefined expression $\infty - \infty$. Unlike the ordinary integral, however, the two conditional expectations may each be infinite on different $\mathscr{G}$-measurable sets, without their difference being undefined.

Example 4.8 (When (conditional) expectation is undefined)**.** Let X be a random variable with

$$\mathbb{P}(X = n) = \mathbb{P}(X = -n) = \frac{3}{\pi^2 n^2}, \qquad n \geqslant 1.$$

Then $\mathbb{E}[X^+] = \mathbb{E}[X^-] = \frac{3}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n} = \infty$, so $\mathbb{E}[X]$ is undefined.

(i) Consider the trivial σ-algebra $\mathscr{G} = \{\emptyset, \Omega\}$. Since conditional expectations with respect to $\mathscr{G}$ are constant,

$$\mathbb{E}[X^+ \mid \mathscr{G}] = \mathbb{E}[X^- \mid \mathscr{G}] = \infty \qquad \text{a.s.}$$

Consequently, $\mathbb{E}[X \mid \mathscr{G}]$ is undefined, since it would involve $\infty - \infty$.

(ii) Now let $\mathscr{G} = \sigma(\{X > 0\})$. Knowing $\mathscr{G}$ tells us the sign of X, and

$$\mathbb{E}[X^+ \mid \mathscr{G}] = \infty \mathbf{1}_{\{X>0\}}, \qquad \mathbb{E}[X^- \mid \mathscr{G}] = \infty \mathbf{1}_{\{X<0\}},$$

where we use the convention $0 \cdot \infty = 0$. Thus

$$\mathbb{E}[X \mid \mathscr{G}] = \begin{cases} +\infty, & X > 0, \\ -\infty, & X < 0. \end{cases}$$

This extended conditional expectation is well-defined, even though $\mathbb{E}[X]$ is not.

The example illustrates that conditioning can separate positive and negative infinite contributions, making an extended conditional expectation well-defined even when the ordinary expectation is undefined.

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

Proposition 4.9 (Conditional Convergence Theorems)**.** *Let $X_1, X_2, \ldots$ and X be random variables on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, and let $\mathscr{G} \subseteq \mathscr{F}$ be a σ-algebra.*

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

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

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

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

The proofs all use the defining relation (18) to transfer statements about convergence of the conditional probabilities to our usual convergence theorems and are left as an exercise.

The following results are very useful in manipulating conditional expectations.

Lemma 4.10 ("Taking out what is known"). *Let X and Y be random variables on $(\Omega,\mathscr{F},\mathbb{P})$ with X, Y and XY integrable. Let $\mathscr{G}\subseteq\mathscr{F}$ be a σ-algebra and suppose that Y is $\mathscr{G}$-measurable. Then*

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

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

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

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

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

Proposition 4.11 (Tower property of conditional expectations). *Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space, X an integrable random variable and $\mathscr{F}_1$, $\mathscr{F}_2$ σ-algebras with $\mathscr{F}_1\subseteq\mathscr{F}_2\subseteq\mathscr{F}$. Then*

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

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

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

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

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

□

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

Jensen's inequality also extends to the conditional setting.

Theorem 4.12 (Conditional Jensen's inequality). *Let X be integrable and let $f:\mathbb{R}\to\mathbb{R}$ be convex, with $f(X)$ integrable. Then*

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

Skecth of Proof (Non-Examinable). We use the following fact (which is not hard to prove but a bit lengthy): for a convex function f on an open interval I, there exist countably many affine functions $x\mapsto a_nx+b_n$ such that

$$f(x)=\sup_{n\geqslant 1}(a_nx+b_n),\qquad x\in I.$$

This follows by choosing supporting lines at a countable dense subset of I such as rational points. The countability is important because the conditional inequalities below hold only almost surely.

For every n, the supporting-line property gives

$$f(X)\geqslant a_nX+b_n.$$

By monotonicity and linearity of conditional expectation,

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

Since there are only countably many n, these inequalities hold simultaneously outside a single null set. Taking the supremum over n yields

$$\begin{aligned}\mathbb{E}[f(X) \mid \mathscr{G}] &\geqslant \sup_{n \geqslant 1} \big(a_n \mathbb{E}[X \mid \mathscr{G}] + b_n\big) \\ &= f(\mathbb{E}[X \mid \mathscr{G}]) \qquad \text{a.s.}\end{aligned}$$

□

Example 4.13. Applying conditional Jensen's inequality to the convex function $f(x) = |x|^p$, for $p \geqslant 1$, gives

$$|\mathbb{E}[X \mid \mathscr{G}]|^p \leqslant \mathbb{E}[|X|^p \mid \mathscr{G}] \qquad \text{a.s.},$$

provided $X \in \mathscr{L}^p$. Taking expectations yields

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

Thus conditional expectation is a contraction on L^p for every $1 \leqslant p < \infty$.

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

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

where $W = Y - \mathbb{E}[X \mid \mathscr{G}]$ and $Z = \mathbb{E}[X \mid \mathscr{G}] - X$. Now Y and $\mathbb{E}[X \mid \mathscr{G}]$ are $\mathscr{G}$-measurable, so W is $\mathscr{G}$ measurable, and using Proposition 4.6 part 1 and Lemma 4.10 we have

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

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

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

4.1 Existence of Conditional Expectation via Orthogonal Projection in L^2

We now sketch an alternative existence proof of conditional expectation that has a more gemeotric flavour than the one we saw via the Radon-Nikodym theorem. Neither proof, is better or worse than the other, they just highlight two different ways to think about conditional expectation; the previous one emphasizes the definigin relation of the conditional expectation and constructs the associated conditional probability measure via Radon-Nikodym. The L^2 proof in this section gives a geometric interpration: the key idea is that the norm $\|.\|_2$ is induced by the inner product

$$\langle X, Y \rangle := \mathbb{E}[XY].$$

That is, $(L^2, \|.\|_2)$ is not just a Banach space, but $(L^2, \langle \cdot, \cdot \rangle)$ is a Hilbert space. This inner product allows to generalize geometric constructions such as orthogonal projections to any linear subspace of $L^2(\Omega, \mathscr{F}, \mathbb{P})$. In

particular, if $\mathscr{G} \subset \mathscr{F}$ is a sub-σ-algebra, the linear subspace of equivalence classes of square integrable $\mathscr{G}$-measureable random variables is such a linear subspace,

$$L^2(\Omega, \mathscr{G}, \mathbb{P}) \subseteq L^2(\Omega, \mathscr{F}, \mathbb{P}).$$

Hence, to find the "best" approximation to an element $[X] \in L^2(\Omega, \mathscr{F}, \mathbb{P})$ we use the orthogonal projection of $[X]$ to $L^2(\Omega, \mathscr{F}, \mathbb{P})$. It is not hard to see that this orthogonal projection gives a us (a version) of the conditional expectation $\mathbb{E}[X|\mathscr{G}]$. Further, a simply monotonicity argument allows to get rid of the restriction $p = 2$.

Since only some of you are familiar with Banach and Hilbert spaces, I spelled out the proof using $\mathscr{L}^2$ instead of L^2 below. It is also part of one of the sheets, so please try it yourself before looking at what is in the remainder of this section. First, recall from Prelims, that if $\mathscr{K} \subset \mathbb{R}^d$ is a linear subspace of $\mathbb{R}^d$, and $x \in \mathbb{R}^d$, then the linear projection of x to $\mathscr{K}$ is the unique vector $y \in \mathscr{K}$ that attains the infimum

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

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

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

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

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

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

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

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

Theorem 4.14. *[Not Examinable] Let $\mathscr{K}$ be a complete vector subspace of $\mathscr{L}^2(\Omega, \mathscr{F}, \mathbb{P})$. For any $X \in \mathscr{L}^2(\Omega, \mathscr{F}, \mathbb{P})$ the infimum*

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

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

Proof of Theorem 4.14. Let $(Y_n)_{n \geqslant 1}$ be a sequence which attains the desired infimum, $\|X - Y_n\|_2 \to \Delta$. We argue that the sequence is Cauchy. Using the parallelogram law, we have

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

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

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

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 3.15, then gives $\|X - Y\|_2 \leqslant \|X - Y_n\|_2 + \|Y - Y_n\|_2$ and taking limits we see that

$\|X-Y\|_2 = \Delta$ as required.
Now, let $Z \in \mathscr{K}$ with $\|Z\|_2 > 0$ and note that $(Y+tZ) \in \mathscr{K}$ for all $t \in \mathbb{R}$. Using optimality of Y we have

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

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

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

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

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

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

Proof (Not examinable) of the existence in Theorem 4.4 via L^2 projection. Suppose first that $X \in \mathscr{L}^2(\Omega, \mathscr{F}, \mathbb{P})$ and let $\mathscr{K} = \mathscr{L}^2(\Omega, \mathscr{G}, \mathbb{P})$. Clearly $\mathscr{K}$ is a vector subspace of $\mathscr{L}^2(\Omega, \mathscr{F}, \mathbb{P})$ and is complete by Theorem 3.16. Let Y be the orthogonal projection of X on $\mathscr{K}$ from Theorem 4.14. We will now verify that Y is a version of the conditional expectation of X given $\mathscr{G}$. First Y is $\mathscr{G}$-measurable since $Y \in \mathscr{K}$. Second, for $G \in \mathscr{G}$ note that $\mathbf{1}_G \in \mathscr{K}$ and since $(X-Y)$ is orthogonal to $\mathscr{K}$ we have $\mathbb{E}[(X-Y)\mathbf{1}_G] = 0$ which shows that the defining relationship of conditional expectation holds.

For $X \in \mathscr{L}^1$, by linearity, it is enough to deal with $X^{\pm}$ separately. Suppose thus that $X \geqslant 0$ and let $X_n = X \wedge n$ which are bounded and in particular in $\mathscr{L}^2$ so that $Y_n = \mathbb{E}[X_n \mid \mathscr{G}]$ exists by the above. Monotonicity gives $Y_n \uparrow Y$ a.s. and the ordinary MCT implies that for every $G \in \mathscr{G}$,

$$\mathbb{E}[Y\mathbf{1}_G] = \lim_n \mathbb{E}[Y_n\mathbf{1}_G] = \lim_n \mathbb{E}[X_n\mathbf{1}_G] = \mathbb{E}[X\mathbf{1}_G].$$

Taking $G = \Omega$ also shows that Y is integrable, hence Y is a version of the conditional expectation. □

5 Filtrations and Stopping Times

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

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

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

$$\mathscr{F}_\infty = \sigma\left(\bigcup_{n\geqslant 0} \mathscr{F}_n\right) \tag{21}$$

be the σ-algebra generated by the filtration.

Often n is interpreted as time and $\mathscr{F}_n$ as the collection of events whose occurrence or non-occurrence can be determined from the information observable up to and including timea n: an event $A \in \mathscr{F}_n$ is one whose occurrence can be determined using observations up to time n. The condition $\mathscr{F}_n \subseteq \mathscr{F}_{n+1}$ expresses that such information can accumulate but is never forgotten. We usually start at time 0 (the beginning), but not always in which case we modify the definitions above in the obvious way. Note that $\mathscr{F}_\infty$ captures all the information we may acquire but it may be strictly smaller than the abstract $\mathscr{F}$.

Definition 5.2 (Adapted stochastic process)**.** A *stochastic process* $(X_n)_{n\geqslant 0}$ is a sequence of random variables defined on $(\Omega, \mathscr{F}, \mathbb{P})$. The process is *integrable* if each X_n is integrable. We say that $(X_n)_{n\geqslant 0}$ is *adapted* to the filtration $(\mathscr{F}_n)_{n\geqslant 0}$ if, for each n, X_n is $\mathscr{F}_n$-measurable.

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

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

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

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.

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

5.1 Stopping Times

Whatever the real world interpretation of the index n may be, we shall refer to specific choices of n as *deterministic times*, e.g. $n = 42$, or of if we index by days and X_n could be, e.g., the temperature recorded in Greenwich

Observatory at noon on on the first of July, or the Rolls-Royce Holdings plc closing price at the London Stock Exchange. However, we often use many other, *random* times: the next time I meet you, the first time you see a yeti, the moment the stock price drops by more than 30% from its past maximum. It is clear these are well defined but not known a priori. They are not deterministic but rather of the type 'I know it when it happens'. We shall turn these now into a mathematically precise notion of *stopping times*. Much of the power of martingale theory comes from the fact that, under suitable assumptions, many results remain valid when deterministic times are replaced by stopping times. This will lead to the optional stopping theorem later in the course.

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

Equivalently, τ is a stopping time if $\{\tau \leqslant n\} \in \mathscr{F}_n$ for every $n \geqslant 0$. Indeed,

$$\{\tau \leqslant n\} = \bigcup_{k=0}^{n} \{\tau = k\}, \qquad \{\tau = n\} = \{\tau \leqslant n\} \setminus \{\tau \leqslant n-1\},$$

where $\{\tau \leqslant -1\} = \emptyset$. If the choice of the filtration is unambiguous we shall simply say that τ is a stopping time. Thus, at time n, the available information suffices to decide whether we have already stopped. A stopping time cannot depend on information that will only become available in the future. 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. Stopping times are sometimes called *optional times*. Note that not all random times are stopping times. If $n = 365$ and τ is the warmest day of the year, then I need $\mathscr{F}_{365}$ to decide when τ actually happens. Likewise, the day in November 2024 on which Rolls Royce is most expensive is not known in advance or when it happens. You need to wait till the end of November to know when it actually occurred. It is not a stopping time.

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

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

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

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

Proof. Exercise □

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

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

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

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

Proof.

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

□

However, not all "natural random times" that are associated with a stochastic process are stopping times: consider a biased random walk $X_0 = 0$ and $X_n = I_1 + \ldots + I_n$ with the increments $I_1, I_2, \ldots$ iid, taking values in $\{-1, 1\}$ and such that $\mathbb{P}(I_1 = 1) \neq \mathbb{P}(I_1 = -1)$. Then

$$L := \sup_n \{n : X_n = 0\}$$

is the last time it visits 0 and since by the LLN $X_n \to \infty$ a.s. we know that $L < \infty$ a.s. However, although L is finite almost sure it is not a stopping time (with respect to the natural filtration)! Why? Because at time n, even if $X_n = 0$, you cannot know whether the walk will return to zero later; formally $\{L = n\} \notin \sigma(X_0, X_1, \ldots, X_n)$. Stopping times formalise what can be decided using present and past information, without access to the future. Let's take this as motivation to make precise y what kind of measurability is needed.

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

$$\mathscr{F}_\tau = \{A \in \mathscr{F}_\infty : A \cap \{\tau = n\} \in \mathscr{F}_n \ \forall n \geqslant 0\}. \tag{22}$$

The restriction $A \in \mathscr{F}_\infty$ is relevant when τ may be infinite: on $\{\tau = \infty\}$, the information available is that of $\mathscr{F}_\infty$ which is in general smaller than $\mathscr{F}$, $\mathscr{F}_\infty \subseteq \mathscr{F}$.

Remark. The σ-algebra $\sigma(\tau)$ contains only the information about *when* we stop, whereas $\mathscr{F}_\tau$ contains all information available *up to the time we stop*. In general,

$$\sigma(\tau) \subseteq \mathscr{F}_\tau,$$

and the inclusion can be strict as Example 5.9 shows.

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

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

(i) $\mathscr{F}_\tau$ defined in (22) *is a σ-algebra;*

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

Proof. Exercise. □

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

Example 5.9 (Information at a stopping time)**.** Let $(S_n)_{n\geqslant 0}$ be a simple symmetric random walk with $S_0 = 0$ and natural filtration $\mathscr{F}_n = \sigma(S_0, \ldots, S_n)$. Consider

$$\tau = \begin{cases} 1, & S_1 = 1, \\ 2, & S_1 = -1. \end{cases}$$

Thus, we stop after the first step if it is upwards, and otherwise take one more step. Write UU, UD, DU, DD for the events corresponding to the four possible sequences of the first two increments. Since $\tau \leqslant 2$, Proposition 5.8 gives $\mathscr{F}_\tau \subseteq \mathscr{F}_2$. By the definition of $\mathscr{F}_\tau$, its atoms are

$$UU \cup UD, \qquad DU, \qquad DD,$$

that is is $\mathscr{F}_\tau$ is the smallest σ-algebra that contains the above three events. Indeed, if the first step is upwards, we stop without observing the second step. Otherwise, we observe both steps. Consequently,

$$\sigma(\tau) = \mathscr{F}_1 \subsetneq \mathscr{F}_\tau \subsetneq \mathscr{F}_2.$$

Thus, $\sigma(\tau)$ records when we stop, whereas $\mathscr{F}_\tau$ records all information available up to that time.

5.2 The Stopped Process X_τ

Given an adapted process $(X_n)_{n\geqslant 0}$ and a finite stopping time τ, we define

$$X_\tau(\omega) = X_{\tau(\omega)}(\omega).$$

This is an $\mathscr{F}_\tau$-measurable random variable. Indeed, for every $x \in \mathbb{R}$ and $k \geqslant 0$,

$$\{X_\tau \leqslant x\} \cap \{\tau = k\} = \{X_k \leqslant x\} \cap \{\tau = k\} \in \mathscr{F}_k.$$

If τ may be infinite, X_τ is only defined once a value X_∞ has been specified on $\{\tau = \infty\}$.

Proposition 5.10 (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 the* stopped process

$$\mathbf{X}^\tau = (X_{\tau\wedge n})_{n\geqslant 0}$$

is adapted to $(\mathscr{F}_{\tau\wedge n})_{n\geqslant 0}$ and hence also to $(\mathscr{F}_n)_{n\geqslant 0}$.

Proof. For each n, $\tau \wedge n$ is a finite stopping time, so $X_{\tau\wedge n}$ is $\mathscr{F}_{\tau\wedge n}$-measurable by the observation above. □

Combining stopping times with expectations will become especially powerful but care is needed! E.g. $\mathbb{E}[X_n] = 0$ does not imply $\mathbb{E}[X_\tau] = 0$ for any stopping time τ as the following example shows.

Example 5.11 (Can you win by doubling your bets? The "Infallible Martingale")**.** Consider a fair coin game in which you win or lose your stake with equal probability. Start by betting £1, and double your stake after every loss, stopping at the first win. Thus, the successive stakes are $1, 2, 4, 8, \ldots$. Let τ be the time of the first win. Then

$$\mathbb{P}(\tau > n) = 2^{-n},$$

so $\tau < \infty$ almost surely. Moreover, at time τ the total profit is

$$X_\tau = 2^{\tau-1} - \sum_{k=0}^{\tau-2} 2^k = 1 \qquad \text{a.s.}$$

Thus, the strategy appears to guarantee a profit of £1 in a fair game! Where is the catch? Although the expected profit after any fixed number of rounds is zero, we have $\mathbb{E}[X_\tau] = 1$. To understand why this does not contradict the fairness of the game, we need the theory of *martingales* and, in particular, the *Optional Stopping Theorem*.

6 Martingales

We now have all the tools necessary to start studying martingales. Here are two motivations to do this

- Monotone sequences of real numbers were one of the first things that you studied in your analysis course. They have many desirable properties are also useful in understanding other sequences. (Sub/super) martingales are simply a probabilist's version of monotone sequences: knowing what happened in the past, "on average," my next element will go up/down/stay the same. From this point of view, it is not suprising that martingales are a building block in the theory of stochastic processes.
- Historically the introduction of martingales into mathematics is due to Ville in 1939, who was inspired by the gambling strategy 'the infallible martingale' from Example 5.11. However, the martingale is clearly not infallible (otherwise I wouldn't be teaching this course). Nailing down why in purely mathematical terms had to await for the further the development of martingale by J.L. Doob in the 1940's. The mathematical term martingale doesn't refer to the gambling *strategy*, but rather models the outcomes of a series of fair games (although, I reemphasize that this is only one application among many).

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

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

Remark. From the definition and linearity of conditional expectation it is clear that if (X_n) and (Y_n) are martingales with respect to the same filtration $(\mathscr{F}_n)$, then so is $(X_n + Y_n)$. But it is easy to find examples where (X_n) is a martingale with respect to its natural filtration, (Y_n) is a martingale with respect to its natural filtration, but $(X_n + Y_n)$ is not a martingale with respect to its natural filtration. So it's not just to be pedantic that we specify a filtration $(\mathscr{F}_n)$.

Deep Dive

The terminology *submartingale* and *supermartingale* is closely related to the notions of *subharmonic* and *superharmonic* functions from analysis and potential theory. There a function $f : \mathbb{Z} \to \mathbb{R}$ is called *harmonic* if

$$f(x) = \frac{1}{2}\big(f(x-1) + f(x+1)\big) \qquad \text{for every } x \in \mathbb{Z}.$$

Thus, the value at each point equals the average of its two neighbours. Similarly, f is *subharmonic* if

$$f(x) \leqslant \frac{1}{2}\big(f(x-1) + f(x+1)\big),$$

and *superharmonic* if the reverse inequality holds. Now consider a simple symmetric random walk $(S_n)_{n\geqslant 0}$ on $\mathbb{Z}$ and observe that

$$\mathbb{E}[f(S_{n+1}) \mid \mathscr{F}_n] = \frac{1}{2}\big(f(S_n - 1) + f(S_n + 1)\big),$$

where $\mathscr{F}_n$ is the natural filtration. Consequently, provided the relevant random variables are integrable,

$$\begin{aligned} f \text{ harmonic} &\implies (f(S_n))_{n\geqslant 0} \text{ is a martingale,} \\ f \text{ subharmonic} &\implies (f(S_n))_{n\geqslant 0} \text{ is a submartingale,} \\ f \text{ superharmonic} &\implies (f(S_n))_{n\geqslant 0} \text{ is a supermartingale.} \end{aligned}$$

For example, $f(x) = x$ is harmonic, whereas $f(x) = x^2$ is subharmonic, since

$$\frac{1}{2}\big((x-1)^2 + (x+1)^2\big) = x^2 + 1.$$

This explains why (S_n) is a martingale but (S_n^2) is a submartingale. In continuous space, the analogous notion of harmonicity is expressed through the Laplacian: a twice differentiable function f is harmonic if $\Delta f = 0$, subharmonic if $\Delta f \geqslant 0$, and superharmonic if $\Delta f \leqslant 0$. For Brownian motion (B_t), the corresponding process $f(B_t)$ is respectively a local martingale, local submartingale, or local supermartingale (under suitable regularity and localisation). This connection between stochastic processes and partial differential equations is a central theme of modern probability theory that will become important in B8.2 and Part C courses.

6.1 Basic Properties

Here are some elementary properties.

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

1. *A stochastic process $(X_n)_{n\geqslant 0}$ on $(\Omega, \mathscr{F}, \mathbb{P})$ is a submartingale w.r.t. the filtration $(\mathscr{F}_n)_{n\geqslant 0}$ if and only if $(-X_n)_{n\geqslant 0}$ is a supermartingale. It is a martingale if and only if it is both a supermartingale and a submartingale.*

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

$$\mathbb{E}[X_n] = \mathbb{E}[X_0] \quad \text{for all } n.$$

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

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

and hence also

$$\mathbb{E}[X_n] \geqslant \mathbb{E}[X_m].$$

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

Proof. 1. is obvious.

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

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

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

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

so $\mathbb{E}[X_n \mid \mathscr{F}_m] \geqslant X_m$ a.s. follows by induction. To deduce that $\mathbb{E}[X_n] \geqslant \mathbb{E}[X_m]$ just take the expectation. □

Whether (X_n) is a martingale or not depends on the filtration under consideration. If none is specified, there is a default, namely the natural filtration of $(X_n)_{n\geqslant 0}$ which is

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

It is the smallest filtration to which (X_n) is adapted.

Proposition 6.3. *If $(X_n)_{n\geqslant 0}$ is a submartingale w.r.t. some filtration $(\mathscr{F}_n)_{n\geqslant 0}$ on $(\Omega, \mathscr{F}, \mathbb{P})$ and $(X_n)_{n\geqslant 0}$ is also adapted to another filtration $(\mathscr{F}'_n)_{n\geqslant 0}$ with $\mathscr{F}'_n \subseteq \mathscr{F}_n$. Then $(X_n)_{n\geqslant 0}$ is also a submartingale with respect to $(\mathscr{F}'_n)_{n\geqslant 0}$. In particular, this holds if $(\mathscr{F}'_n)_{n\geqslant 0}$ is the natural filtration $(\mathscr{G}_n)_{n\geqslant 0}$ of $(X_n)_{n\geqslant 0}$.*

Proof. Integrability is not affected by a change of filtration. Thus, by the tower property,

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

hence $(X_n)_{n\geqslant 0}$ is a submartingale wrt to $(\mathscr{F}'_n)_{n\geqslant 0}$. $(X_n)_{n\geqslant 0}$ is also adapted to its natural filtration $(\mathscr{G}_n)_{n\geqslant 0}$. Since (by definition) $\mathscr{G}_n$ is the smallest σ-algebra with this property, $\mathscr{G}_n \subseteq \mathscr{F}_n$. □

Martingales arise naturally; some classic examples are

1. Let $(Y_n)_{n\geqslant 1}$ be a sequence of independent random variables with $\mathbb{E}[Y_n] = 0$ for each n. For $n \geqslant 0$ let

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

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

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

2. Let $(Z_n)_{n\geqslant 1}$ be a sequence of independent random variables with $\mathbb{E}[Z_n] = 1$ for all n. Define

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

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

3. Let X be an integrable random variable. Then setting

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

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

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

We shall see later that a large class of martingales, called *uniformly integrable martingales*, can be represented in this way. This is especially usefully, since only the random variable X and the filtration $(\mathscr{F}_n)_n$ already completely determine the infinite sequence of random variables $(X_n)_n$.

The situation of the first example arises often so for brevity we introduce the following terminology

Definition 6.4 (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

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

The following two points are straightforward to check and we will use them often:

- $(X_n)_{n\geqslant 0}$ 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. $(\mathscr{F}_n)$.
- 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 $X_0 = 0$.

6.2 Preserving (Sub-)martingale Structure

A natural first question is which functions preserve the martingale structure. Given our previous results about convex functions the following shouldn't be suprising.

Proposition 6.5. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Suppose that $(X_n)_{n\geqslant 0}$ is a martingale with respect to the filtration $(\mathscr{F}_n)_{n\geqslant 0}$. Let f be a* convex *function on $\mathbb{R}$. If $f(X_n)$ is an integrable random variable for each $n \geqslant 0$, then $(f(X_n))_{n\geqslant 0}$ is a* sub*martingale w.r.t $(\mathscr{F}_n)_{n\geqslant 0}$.*

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

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

□

Corollary 6.6. *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. $(\mathscr{F}_n)_{n\geqslant 0}$.*

In gambling, one has to make a bet before seeing the outcome. This motivates the following definitions of predictable process and discrete stochastic integral.

Definition 6.7 (Predictable process)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $(\mathscr{F}_n)_{n\geqslant 0}$ a filtration. A sequence $(V_n)_{n\geqslant 1}$ of random variables is *predictable* with respect to $(\mathscr{F}_n)_{n\geqslant 0}$ if V_n is $\mathscr{F}_{n-1}$-measurable for all $n \geqslant 1$.

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

Theorem 6.8 (Discrete stochastic integral or martingale transform)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $(\mathscr{F}_n)_{n\geqslant 0}$ a filtration. Let $(Y_n)_{n\geqslant 0}$ be a martingale with respect to $(\mathscr{F}_n)$ with difference sequence $(D_n)_{n\geqslant 1}$. Suppose that $(V_n)_{n\geqslant 1}$ is predictable w.r.t. $(\mathscr{F}_n)$, and let*

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

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

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

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

It is a discrete version of the *stochastic integral*, and you will learn more about stochastic integrals in B8.2. Here we started with $X_0 = 0$. We may also add any integrable $\mathscr{F}_0$-measurable random variable Z to all X_n without changing the martingale property. In particular, taking $Z = Y_0$ gives

$$X_n = Y_0 + \sum_{k=1}^{n} V_k D_k.$$

Proof. The process (X_n) is adapted by predictability of (V_n). Moreover, $V_{n+1}D_{n+1} = X_{n+1} - X_n$ is integrable by assumption. Hence

$$\begin{aligned}\mathbb{E}[X_{n+1} - X_n \mid \mathscr{F}_n] &= \mathbb{E}[V_{n+1}D_{n+1} \mid \mathscr{F}_n] \\ &= V_{n+1}\mathbb{E}[D_{n+1} \mid \mathscr{F}_n] = 0 \quad \text{a.s.}\end{aligned}$$

Thus (X_n) is a martingale. □

Predictable sequences appear naturally when one models decision making. In our gambling example, if $(k-1)$ rounds of some gambling game have just been completed, then the strategy for the kth round is to bet V_k; a quantity that can only depend on what is known by time $k-1$. The change in fortune in the kth round is then $V_k D_k$.

Theorem 6.9. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $(\mathscr{F}_n)_{n\geqslant 0}$ a filtration. Let $(Y_n)_{n\geqslant 0}$ be a* supermartingale *with respect to $(\mathscr{F}_n)$ with difference sequence $(D_n)_{n\geqslant 1}$, and $(V_n)_{n\geqslant 1}$ a* non-negative *$(\mathscr{F}_n)$-predictable sequence. Then (modulo integrability)*

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

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

Proof. Exercise: imitate the proof of Theorem 6.8. □