B10a: Martingales through measure theory

Alison Etheridge

0 Introduction

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 indispensible tool in many disciplines - from physics to neuroscience, from genetics to communication networks, and, of course, in mathematical finance. Equally, probabilistic approaches have gained importance in mathematics itself, from number theory to partial differential equations.

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

The nearest thing to a course text is

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

Also highly recommended are:

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

0.2 The Galton-Watson branching process

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

If you did Part A Probability then you'll have already come across the Galton-Watson branching process. In spite of earlier work by Bienaymé, it is attributed to the great polymath Sir Frances 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 and that model, formulated by Watson, is what we now call the Galton-Watson branching process. Let's state it formally:

Definition 0.1 (Galton-Watson branching process)**.** *Let $\{X_r^{(m)}; m, r \in \mathbb{N}\}$ be a doubly infinite sequence of independent identically distributed random variables, each with the same distribution as X, where*

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

and $\mu = \sum_{k=0}^{\infty} k p_k < \infty$. Write $f(\theta) = \sum_{k=0}^{\infty} p_k \theta^k$ for the probability generating function of X. Then the sequence $\{Z_n\}_{n \in \mathbb{N}}$ of random variables defined by

1. $Z_0 = 1$,
2. $Z_{n+1} = X_1^{(n+1)} + \cdots + X_{Z_n}^{(n+1)}$,

is the Galton-Watson branching branching process (started from a single ancestor) with offspring generating function f.

The random variable Z_n models the number of male descendants of a single male ancestor after n generations.

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

'Proof'

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

$$\begin{aligned} \mathbb{E}\left[\theta^{Z_{n+1}} \middle| Z_n = k\right] &= \mathbb{E}\left[\theta^{X_1^{(n+1)} + \cdots + X_k^{(n+1)}}\right] \\ &= \mathbb{E}\left[\theta^{X_1^{(n+1)}}\right] \cdots \mathbb{E}\left[\theta^{X_k^{(n+1)}}\right] \quad \text{(independence)} \\ &= f(\theta)^k, \end{aligned}$$

(since $X_i^{(n+1)}$ has the same distribution as X). Hence

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

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

$$\begin{aligned} \mathbb{E}\left[\theta^{Z_{n+1}}\right] &= \mathbb{E}\left[\mathbb{E}\left[\theta^{Z_{n+1}} \middle| Z_n\right]\right] \qquad (2) \\ &= \mathbb{E}\left[f(\theta)^{Z_n}\right] \\ &= f_n\left(f(\theta)\right) \quad \text{(inductive hypothesis).} \end{aligned}$$

□

In (2) we have used what is called the *tower property* of conditional expectataions. In this example you can make all this work with the Partition Theorem of mods (because the events $\{Z_n = k\}$ partition our space). In the general theory that follows, we'll see how to replace the Partition Theorem when the sample space is not so nice.

Watson wanted to establish the *extinction probability* of the branching process, that is the value of $\mathbb{P}[Z_n = 0 \text{ for some } n]$.

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

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

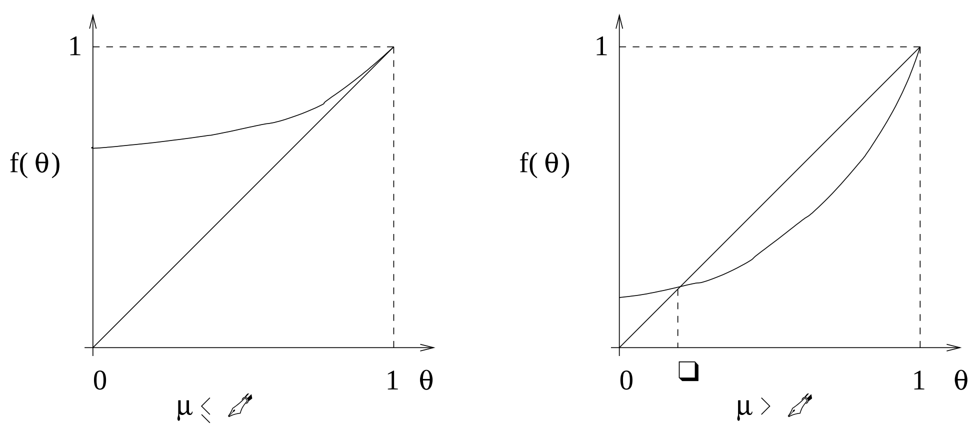
'Proof'

Writing $q_n = \mathbb{P}[Z_n = 0]$, since $\{Z_n = 0\} \subseteq \{Z_{n+1} = 0\}$ we see that q_n is an increasing function of n and, intuitively,

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

A *proof* will need the Monotone Convergence Theorem. Since $f_{n+1}(0) = f(f_n(0))$, evidently, granted (3), q solves $q = f(q)$.

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



To see that q must be the *smaller* root, note that the sequence q_n satisfies $f(q_n) \geq q_n$ and so $\{q_n\}_{n\in\mathbb{N}}$ is monotone increasing and bounded above by q (since for $\theta > q$ we have $f(\theta) < \theta$) and by AOL the limit must be q. □

In fact, Watson didn't spot that the extinction probability was given by the smaller root and concluded that the population (name in this case) would always die out. Galton came up with a more plausible explanation. 'Prominent names' meant people like politicians. To finance their political ambitions men married heiresses and these women were genetically predisposed to only bear female children and so the family name was not passed on.

In spite of this inauspicious start, branching processes play a very important rôle in probabilistic modelling.

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

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

$$\mathbb{E}\left[Z_{n+1} \middle| Z_n = k\right] = \mathbb{E}\left[X_1^{(n+1)} + \cdots + X_k^{(n+1)}\right] = k\mu \quad \text{(linearity of expectation)}. \tag{4}$$

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

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

then

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

In fact, more is true.

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

The process $\{M_n\}_{n\in\mathbb{N}}$ is our first example of a *martingale*.

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

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

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

1 Measure spaces

We begin by recalling some definitions that we encountered in Part A Integration (and, less explicitly, in Mods Probability). The idea was that we wanted to be able to assign a 'mass' or 'size' to subsets of a space in a consistent way. In particular, for us these subsets will be 'events' or 'collections of outcomes' (subsets of a probability sample space Ω) and the 'mass' will be a probability (a measure of how likely that event is to occur).

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

1. *We say that $\mathcal{F}$ is an* algebra *if $\emptyset \in \mathcal{F}$ and for all $A, B \in \mathcal{F}$, $A^c \in \mathcal{F}$ and $A \cup B \in \mathcal{F}$,*
2. *We say that $\mathcal{F}$ is a σ-*algebra *if $\emptyset \in \mathcal{F}$ and for all sequences $\{A_n\}_{n\in\mathbb{N}}$ of elements of $\mathcal{F}$, $A_1^c \in \mathcal{F}$ and $\cup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.*

An algebra is closed under *finite* set operations whereas a σ-algebra is closed under *countable* set operations.

Definition 1.2 (Measure space)**.** *We say that $(\Omega, \mathcal{F}, \mu)$ is a* measure space *if Ω is a set, $\mathcal{F}$ is a σ-algebra of subsets of Ω and $\mu : \mathcal{F} \to [0, \infty]$ satisfies $\mu(\emptyset) = 0$ and for any sequence $\{A_n\}_{n\in\mathbb{N}}$ of* disjoint *elements of $\mathcal{F}$,*

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n). \tag{6}$$

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

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

Example 1.3 (Discrete measure theory). *Let Ω be a countable set and $\mathcal{F}$ the* power set *of Ω (that is the set of* all *subsets of Ω). A* mass function *is any function $\bar{\mu} : \Omega \to [0, \infty]$. We can then define a measure on Ω by $\mu(\{x\}) = \bar{\mu}(x)$ and extend to arbitrary subsets of Ω using Property (6).*

Equally given a measure on Ω we can define a mass function. So there is a one-to-one correspondence between measures on a countable set Ω and mass functions.

These discrete measure spaces provide a 'toy' version of the general theory but in general they are not enough. Discrete measure theory is essentially the only context in which one can define the measure explicitly. This is because σ-algebras are not in general amenable to an explicit presentation and it is *not* in general the case that for an arbitrary set Ω all subsets of Ω can be assigned a measure - recall from Part A Integration that we constructed a non-Lebesgue measurable subset of $\mathbb{R}$. Instead one specifies the values to be taken by the measure on a smaller class of subsets of Ω that 'generate' the σ-algebra (as the singletons did in Example 1.3). This leads to two problems. First one needs 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.* The second problem is to know that there is only *one* measure on the σ-algebra that is consistent with our specification. This uniqueness problem can often be resolved through a corollary of Dynkin's π-system Lemma that we state below. First we need more definitions.

Definition 1.4 (Generated σ-algebras). *Let $\mathcal{A}$ be a collection of subsets of Ω. Define*

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

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

Example 1.5 (Borel σ-algebra, Borel measure, Radon measure). *Let Ω be a topological space with topology (that is open sets) $\mathcal{T}$. Then* the Borel σ-algebra of Ω *is the σ-algebra generated by the open sets,*

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

A measure μ on $(\Omega, \mathcal{B}(\Omega))$ is called a Borel measure. *If also $\mu(K) < \infty$ for every compact set $K \subseteq \Omega$ then μ is called a* Radon measure.

Definition 1.6 (π-system). *Let $\mathcal{I}$ be a collection of subsets of Ω. We say that $\mathcal{I}$ is a* π-system *if $\emptyset \in \mathcal{I}$ and for all $A, B \in \mathcal{I}$, $A \cap B \in \mathcal{I}$.*

Notice that an algebra is automatically a π-system.

Example 1.7. *The collection*

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

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

Here's why we care about π-systems.

Theorem 1.8 (Uniqueness of extension). *Let Ω be a set and let $\mathcal{I}$ be a π-system on Ω. Let $\mathcal{F} = \sigma(\mathcal{I})$ be the σ-algebra generated by $\mathcal{I}$. Suppose that μ_1, μ_2 are measures on $(\Omega, \mathcal{F})$ such that $\mu_1(\Omega) = \mu_2(\Omega) < \infty$ and $\mu_1 = \mu_2$ on $\mathcal{I}$. Then $\mu_1 = \mu_2$ on $\mathcal{F}$.*

In particular, if two probability measures on Ω agree on a π-system, then they agree on the σ-algebra generated by that π-system.

Exercise 1.9. *Find an example where uniqueness fails if $\mathcal{I}$ is not a π-system on $\Omega = \{1, 2, 3, 4\}$.*

That deals with uniqueness, but what about existence?

Definition 1.10 (Set functions). *Let $\mathcal{A}$ be* any *set of subsets of Ω containing the emptyset $\emptyset$. A* set function *is a function $\mu : \mathcal{A} \to [0, \infty]$ with $\mu(\emptyset) = 0$. Let μ be a set function. We say that μ is*

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

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

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

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

3. countably additive *if for all sequences of disjoint sets $\{A_n\}_{n\in\mathbb{N}}$ in $\mathcal{A}$ with $\cup_{n\in\mathbb{N}} A_n \in \mathcal{A}$*

$$\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

In this language, a measure space is a set Ω equipped with a σ-algebra $\mathcal{F}$ and a countably additive set function on $\mathcal{F}$.

An immediate consequence of σ-additivity of measures is the following useful lemma.

Notation: For a sequence of sets $\{F_n\}_{n\in\mathbb{N}}$, $F_n \uparrow F$ means $F_n \subseteq F_{n+1}$ for all n and $\cup_{n\in\mathbb{N}} F_n = F$. Similarly, $G_n \downarrow G$ means $G_n \supseteq G_{n+1}$ for all n and $\cap_{n\in\mathbb{N}} G_n = G$.

Lemma 1.11 (Monotone convergence properties). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.*

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

Proof

1. Let $H_1 = F_1$, $H_n = F_n \backslash F_{n-1}$, $n \geq 2$. Then $\{H_n\}_{n\in\mathbb{N}}$ are disjoint and

$$\begin{aligned}
\mu(F_n) &= \mu(H_1 \cup \cdots \cup H_n) \\
&= \sum_{k=1}^{n} \mu(H_k) \quad \text{(additivity)} \\
&\uparrow \sum_{k=1}^{\infty} \mu(H_k) \quad \text{(positivity)} \\
&= \mu\left(\bigcup_{k=1}^{\infty} H_k\right) \quad (\sigma\text{-additivity}) \\
&= \mu(F).
\end{aligned}$$

2. follows on taking $F_n = G_k \backslash G_{k+n}$. □

Note that $\mu(G_k) < \infty$ is essential (for example take $G_n = (n, \infty) \subseteq \mathbb{R}$ and Lebesgue measure).

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

Remark 1.13. *If $\mu_0(\Omega) < \infty$, then Theorem 1.8 tells us that μ is unique since an algebra is certainly a π-system.*

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

The proof of this result is an exercise. (The tricky bit is that Theorem 1.8 requires $\mu(\Omega) < \infty$ and so you must work a little harder.)

The proof of the Carathéordory Extension Theorem proceeds in much the same way as our (sketch) proof of the existence of Lebesgue measure in Part A Integration. First one defines an *outer measure* μ^* by

$$\mu^*(A) = \inf\{\sum_j \mu_0(A_j) : A \subseteq \cup_{j\in\mathbb{N}} A_j, A_j \in \mathcal{A}\}$$

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

$$\mu^*(A) = \mu^*(A \cap E) + \mu^*(A \cap E^c).$$

One must check that μ^* then defines a countably additive set function on the collection of measurable sets and that the measurable sets form a σ-algebra that contains $\mathcal{A}$. For more details see Varadhan and the references therein.

The Carathéodory Extension Theorem doesn't quite solve the problem of constructing measures on σ-algebras - it reduces it to constructing countably additive set functions on algebras. The following theorem is very useful for constructing probability measures on Borel subsets of $\mathbb{R}$. First we need some notation.

Notation:

For $-\infty \le a < b < \infty$, let $I_{a,b} = (a, b]$ and set $I_{a,\infty} = (a, \infty)$

Let $\mathcal{I} = \{I_{a,b} : -\infty \le a < b \le \infty\}$. That is $\mathcal{I}$ is the collection of intervals that are open on the left and closed on the right.

Now suppose that $F : \mathbb{R} \to [0, 1]$ is a non-decreasing function with $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. Recall that F is said to be *right continuous* if for each $x \in \mathbb{R}$ $F(x) = \lim_{y\downarrow x} F(y)$. Given such an F we can define a *finitely* additive probability measure on the algebra $\mathcal{A}$ consisting of the emptyset and finite disjoint unions of intervals from $\mathcal{I}$ by setting

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

for intervals and then extending it to $\mathcal{A}$ by defining it as the sum for disjoint unions from $\mathcal{I}$.

Notice that the Borel σ-field $\mathcal{B}(\mathbb{R})$ is the σ-field generated by $\mathcal{A}$. So Carathéodory's Extension Theorem tells us that this extends to a probability measure on $\mathcal{B}$ *provided* that μ is *countably* additive on $\mathcal{A}$.

Theorem 1.15 (Lebesgue)**.** *μ is countably additive on $\mathcal{A}$ if and only if $F(x)$ is a right continuous function of x. Therefore for each right continuous non-decreasing function $F(x)$ with $F(-\infty) = 0$ and $F(\infty) = 1$ there is a unique probability measure μ on the Borel sets of the line such that $F(x) = \mu(I_{-\infty,x})$. Conversely, every countably additive probability measure μ on $\mathcal{B}(\mathbb{R})$ comes from some F. The correspondence is one-to-one.*

Sketch of key points of proof

We'll see essentially this result again in a slightly different guise later in the course. Rather than give a detailed proof here, let's see where right continuity comes into it.

First note that by the monotone convergence properties of Lemma 1.11 (as you prove on problem sheet 1) the σ-additivity of μ on $\mathcal{A}$ is equivalent to saying that for any sequence $\{A_n\}_{n\in\mathbb{N}}$ of sets from $\mathcal{A}$ with $A_n \downarrow \emptyset$, $\mu(A_n) \downarrow 0$.

If μ is σ-additive, then right continuity of F is immediate since this implies

$$\mu(I_{xy}) = F(y) - F(x) \downarrow 0 \quad \text{as } y \downarrow x.$$

The other way round is a bit more work. Suppose that F is right continuous but, for a contradiction, that there exist $\{A_n\}_{n\in\mathbb{N}}$ from $\mathcal{A}$ with $\mu(A_n) \geq \delta > 0$ and $A_n \downarrow \emptyset$.

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

$$|\mu(A_n) - \mu(B_n)| \leq 1 - F(l) + F(-l),$$

we may do this in such a way that $\mu(B_n) \geq \delta/2 > 0$.

Step 2: Supose that $B_n = \cup_{i=1}^{k_n} I_{a_{n_i}, b_{n_i}}$. Replace B_n by $C_n = \cup_{i=1}^{k_n} I_{\tilde{a}_{n_i}, b_{n_i}}$ 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

$$\mu(B_n \backslash C_n) < \frac{\delta}{10 \cdot 2^n} \quad \text{for each } n.$$

Step 3: Set $D_n = \overline{C}_n$, the closure of C_n (obtained by adding the points $\tilde{a}_{n_i}$ to C_n). Set $E_n = \cap_{i=1}^{n} D_{n_i}$ and $F_n = \cap_{i=1}^{n} C_i$. Then

$$F_n \subseteq E_n \subseteq A_n.$$

So $E_n \downarrow \emptyset$ (since $A_n \downarrow \emptyset$). But

$$\mu(F_n) \geq \mu(B_n) - \sum_i \mu(B_i \backslash C_i) = \frac{\delta}{2} - \sum_i \frac{\delta}{10 \cdot 2^i} = \frac{2\delta}{5}$$

and so F_n and hence D_n is non-empty. The decreasing limit of a sequence of *closed* sets cannot be empty and so we have the desired contradiction. □

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

If $x_1, x_2, \ldots$ is a sequence of points and we have probabilities p_n at these points (for example $x_1, x_2, \ldots$ could be the non-negative integers), then for the discrete measure

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

we have the distribution function

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

which only increases by jumps, the jump at x_n being of height p_n.

There are examples of continuous F that don't come from any density (recall the Devil's staircase of Part A Integration).

The measure μ is sometimes called a Lebesgue-Stieltjes measure. We'll return to it a little later.

We now have a very rich class of measures to work with. In Part A Integration, we developed a theory based on Lebesgue measure. It is natural to ask whether we can develop an analogous theory for other measures. The answer is 'yes' and it is gratifying that we already did the work in Part A. The proofs that we used there will 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 Integration

2.1 Definition of the integral

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

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

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. We suppose that $[-\infty, \infty]$ is endowed with the Borel sets $\mathcal{B}(\overline{\mathbb{R}})$. We want to define, where possible, for measurable functions $f : \Omega \to [-\infty, \infty]$, the integral of f with respect to μ,

$$\mu(f) = \int f d\mu = \int_{x \in \Omega} f(x)\mu(dx).$$

Unless otherwise stated, measurable functions map to $\overline{\mathbb{R}}$ with the Borel σ-algebra.

Recall that

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

Theorem 2.2. *Let* $\{f_n\}_{n\in\mathbb{N}}$ *be a sequence of measurable functions with respect to* $\mathcal{F}, \mathcal{B})$*. Then the following are also measurable:*

$$\max_{n\leq k} f_n, \quad \min_{n\leq k} f_n, \quad \sup_{n\in\mathbb{N}} f_n, \quad \inf_{n\in\mathbb{N}} f_n, \quad \limsup_{n\to\infty} f_n, \quad \liminf_{n\to\infty} f_n.$$

Definition 2.3. *A* simple function *is a finite sum*

$$\phi(x) = \sum_{k=1}^{N} a_k \mathbf{1}_{E_k}(x) \tag{7}$$

where each E_k *is a measurable set of finite measure and the* a_k *are constants.*

The *canonical form* of a simple function ϕ is the unique decomposition as in (7) where the numbers a_k are distinct and the sets E_k are disjoint.

Definition 2.4. *If* ϕ *is a simple function with canonical form*

$$\phi(x) = \sum_{k=1}^{M} c_k \mathbf{1}_{F_k}(x)$$

then we define the integral of ϕ *with respect to* μ *as*

$$\int \phi(x)\mu(dx) = \sum_{k=1}^{M} c_k \mu(F_k).$$

Definition 2.5. *For a non-negative measurable function* f *on* $(\Omega, \mathcal{F}, \mu)$ *we define the integral*

$$\mu(f) = \sup\big\{\mu(g) : g \text{ simple}, g \leq f\big\}.$$

Definition 2.6. *We say that a measurable function* f *on* $(\Omega, \mathcal{F}, \mu)$ *is integrable if* $\mu(|f|) < \infty$ *and then we set*

$$\mu(f) = \mu(f^+) - \mu(f^-).$$

Definition 2.7 (μ-almost everywhere)**.** *Let* $(\Omega, \mathcal{F}, \mu)$ *be a measure space. We say that a property holds* μ*-almost everywhere if it holds except on a set of* μ*-measure zero. If* μ *is a probability measure, we often say* almost surely *instead of almost everywhere.*

Note: It is vital to remember that notions of almost everywhere depend on the underlying measure μ.

2.2 The Convergence Theorems

Theorem 2.8 (Fatou's Lemma)**.** *Let $\{f_n\}_{n\geq 1}$ be a sequence of non-negative measurable functions on $(\Omega, \mathcal{F}, \mu)$. Then*

$$\mu(\liminf_{n\to\infty} f_n) \leq \liminf_{n\to\infty} \mu(f_n).$$

Corollary 2.9 (Reverse Fatou Lemma)**.** *Let $\{f_n\}_{n\geq 1}$ be a sequence of non-negative integrable functions. Assume that there exists a measurable function $g \geq 0$ such that $\mu(g) < \infty$ and $f_n \leq g$ for all $n \in \mathbb{N}$. Then*

$$\mu(\limsup_{n\to\infty} f_n) \geq \limsup_{n\to\infty} \mu(f_n).$$

Proof

Apply Fatou to $\{g - f_n\}_{n\geq 1}$. (Note that $\mu(g) < \infty$ is *needed*.) $\square$

Theorem 2.10 (Monotone Convergence Theorem)**.** *Let $\{f_n\}_{n\geq 1}$ be a sequence of non-negative measurable functions. Then*

$$f_n \uparrow f \quad \Longrightarrow \quad \mu(f_n) \uparrow \mu(f).$$

(Note that we are not excluding $\mu(f) = \infty$ here.)

Theorem 2.11 (Dominated Convergence Theorem)**.** *Let $\{f_n\}_{n\geq 1}$ be a sequence of integrable functions on $(\Omega, \mathcal{F}, \mu)$ with $f_n(x) \to f(x)$ as $n \to \infty$ for each $x \in \Omega$. (We say that f_n converges* pointwise *to f.) Suppose that for some integrable function g, $|f_n| \leq g$ for all n. Then f is integrable and*

$$\mu(f_n) \to \mu(f) \quad \textit{as } n \to \infty.$$

A useful lemma that you prove on the problem sheet is the following.

Lemma 2.12 (Scheffé's Lemma)**.** *Let $\{f_n\}_{n\geq 1}$ be a sequence of non-negative integrable functions on $(\Omega, \mathcal{F}, \mu)$ and suppose that $f_n(x) \to f(x)$ for μ-almost every $x \in \Omega$ (written $f_n \to f$ a.e.). Then*

$$\mu(|f_n - f|) \stackrel{n\to\infty}{\longrightarrow} 0 \quad \text{iff} \quad \mu(f_n) \stackrel{n\to\infty}{\longrightarrow} \mu(f).$$

The corollaries of the MCT and DCT for series also extend to this general setting.

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

Definition 2.13 (Random Variable)**.** *In the special case when $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, we'll call a measurable map $X : \Omega \to \mathbb{R}$ a* random variable.

In the language of mods, Ω is the sample space of an experiment and the random variable X is a measurement of the outcome of the experiment. We can think of X as inducing a probability measure on $\mathbb{R}$ via

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

and, in particular, $F_X(x) = \mu_X((-\infty, x])$ defines the *distribution function* of X (c.f. Theorem 1.15). Since $\{(-\infty, x] : x \in \mathbb{R}\}$ is a π-system, we see that the distribution function uniquely determines μ_X. In this notation,

$$\int X(\omega)\mathbb{P}(d\omega) = \int x\mu_X(dx) \equiv \mathbb{E}[X].$$

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

In fact this idea of using a measurable function to map a measure on one space onto a measure on another is more general. Let $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ be measurable spaces and let μ be a measure on $\mathcal{F}$.

Then any measurable (with respect to $(\mathcal{F}, \mathcal{G})$) function $f : \Omega \to \Lambda$ induces an *image measure* $\nu = \mu \circ f^{-1}$ on $\mathcal{G}$ given by

$$\nu(A) = \mu\left(f^{-1}(A)\right).$$

On the problem sheet you'll use this to construct the measure μ of Theorem 1.15 from Lebesgue measure on $[0, 1]$.

2.3 Product Spaces and Independence

Because we want to be able to discuss more than one random variable at a time, we need the notion of product spaces.

Definition 2.14 (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 Ω_1 and Ω_2 come with σ-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ respectively, then we can define a natural σ-algebra $\mathcal{F}$ on Ω as the σ-algebra generated by sets of the form $A_1 \times A_2$ with $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$. This σ-algebra will be called the product σ-algebra.

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

$$\mathbb{P}[A_1 \times A_2] = \mathbb{P}_1[A_1] \times \mathbb{P}_2[A_2] \tag{8}$$

and extending it to the whole of $\mathcal{F}$.

Evidently it can be extended to the algebra $\mathcal{A}$ of sets that are finite disjoint unions of measurable rectangles as the obvious sum. It is a tedious, but straightforward, exercise to check that this is well-defined.

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

Lemma 2.15. *The finitely additive set function $\mathbb{P}$, defined on $\mathcal{A}$ through (8) is* countably *additive on $\mathcal{A}$.*

Proof

Recall that countable additivity is equivalent to checking that for any sequence of measurable sets with $A_n \downarrow \emptyset$, $\mathbb{P}[A_n] \downarrow 0$.

For any $A \in \mathcal{A}$, define the section

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

Then $\mathbb{P}_1[A_{\omega_2}]$ is a measurable function of ω_2 (in fact it is a simple function - exercise) and

$$\mathbb{P}[A] = \int_{\Omega_2} \mathbb{P}_1[A_{\omega_2}] d\mathbb{P}_2.$$

Now let $A_n \in \mathcal{A}$ be a sequence of sets with $A_n \downarrow \emptyset$. Then

$$A_{n,\omega_2} = \{\omega_1 : (\omega_1, \omega_2) \in A_n\}$$

satisfies $A_{n,\omega_2} \downarrow \emptyset$ for each $\omega_2 \in \Omega_2$. Since $\mathbb{P}_1$ is countably additive, $\mathbb{P}_1[A_{n,\omega_2}] \to 0$ for each $\omega_2 \in \Omega_2$ and since $0 \leq \mathbb{P}_1[A_{n,\omega_2}] \leq 1$ for $n \geq 1$ it follows from the DCT (with dominating function $g \equiv 1$) that

$$\mathbb{P}[A_n] = \int \mathbb{P}_1[A_{n,\omega_2}] d\mathbb{P}_2 \to 0.$$

So $\mathbb{P}$ is countably additive as required. $\square$

By an application of the Carathéodory Extension Theorem we see that $\mathbb{P}$ extends uniquely to a countably additive set function on $\sigma(\mathcal{A}) = \mathcal{F}$.

Definition 2.16 (Product measure)**.** *The measure $\mathbb{P}$ defined through (8) is called the* product measure *on $(\Omega, \mathcal{F})$.*

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

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

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

What is the analogous result here?

Theorem 2.17 (Fubini's Theorem)**.** *Let $f(\omega) = f(\omega_1, \omega_2)$ be a measurable function on $(\Omega, \mathcal{F})$. Then f can be considered as a function of ω_2 for each fixed ω_1 or the other way around. The functions $g_{\omega_1}(\cdot)$ on Ω_2 and $h_{\omega_2}(\omega_1)$ on Ω_1 defined by*

$$g_{\omega_1}(\omega_2) = h_{\omega_2}(\omega_1) = f(\omega_1, \omega_2)$$

are measurable for each ω_1 and ω_2.

If f is integrable with respect to the product measure $\mathbb{P}$ then the function $g_{\omega_1}(\cdot)$ is integrable with respect to $\mathbb{P}_2$ for $\mathbb{P}_1$-almost every ω_1 and the function $h_{\omega_2}(\cdot)$ is integrable with respect to $\mathbb{P}_1$ for $\mathbb{P}_2$-almost every ω_2. Their integrals

$$G(\omega_1) = \int_{\Omega_2} g_{\omega_1}(\omega_2) d\mathbb{P}_2$$

and

$$H(\omega_2) = \int_{\Omega_1} h_{\omega_2}(\omega_1) d\mathbb{P}_1$$

are measurable, finite almost everywhere, and integrable with respect to $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively. Finally,

$$\int_{\Omega} f(\omega_1, \omega_2) d\mathbb{P} = \int_{\Omega_1} G(\omega_1) d\mathbb{P}_1 = \int_{\Omega_2} H(\omega_2) d\mathbb{P}_2.$$

Conversely, for a non-negative *function, if either $\int G d\mathbb{P}_1$ or $\int H d\mathbb{P}_2$ is finite then so is the other and f is integrable with integral equal to either of the repeated integrals.*

Warning: Just as we saw for functions on $\mathbb{R}^2$ in Part A Integration, for f to be integrable we require that $|f|$ is integrable. If we drop the assumption of non-negative in the last part then the result is false and it is not hard to cook up examples where both repeated integrals exist but f is *not* integrable.

The proof of Fubini's Theorem is *not* examinable. It follows a standard pattern that Williams calls *the standard machine*:

- Check the result for $f = \mathbf{1}_A$ where $A \in \mathcal{F}$.
- Extend by linearity to non-negative simple functions.

- Pass to increasing limits using the MCT.
- Take positive and negative parts.

However, in this case, it turns out that what one might hope would be the easy bit - checking the result for indicator functions of measurable sets - is highly non-trivial. It relies on a result called the Monotone Class Theorem. We include this here, but it is not examinable.

Definition 2.18 (Monotone Class). *A family of subsets $\mathcal{M}$ of Ω is called a* monotone class *if it is stable under countable unions and countable intersections.*

Theorem 2.19 (The Monotone Class Theorem). *The smallest monotone class containing an algebra $\mathcal{A}$ is the σ-algebra generated by $\mathcal{A}$.*

The point is that if we have a result that we know to be valid on an algebra $\mathcal{A}$ and we can check that the sets for which the result holds form a monotone class, then necessarily the result holds on $\sigma(\mathcal{A})$.

To illustrate, here's the difficult bit of the proof of Fubini's Theorem.

Corollary 2.20 (to Lemma 2.15). *For and $A \in \mathcal{F}$, if we denote by A_{ω_1} and A_{ω_2} the respective sections*

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

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

then the functions $\mathbb{P}_1[A_{\omega_2}]$ and $\mathbb{P}_2[A_{\omega_1}]$ are measurable and

$$\mathbb{P}[A] = \int \mathbb{P}_1[A_{\omega_2}] d\mathbb{P}_2 = \int \mathbb{P}_2[A_{\omega_1}] d\mathbb{P}_1.$$

In particular, for a measurable set A, $\mathbb{P}[A] = 0$ iff *for $\mathbb{P}_1$-almost all ω_1 the sections A_{ω_1} have $\mathbb{P}_2$-measure zero or equivalently for $\mathbb{P}_2$-almost every ω_2, the sections A_{ω_2} have $\mathbb{P}_1$-measure zero.*

Proof

The assertion clearly works for a rectangle of the form $A_1 \times A_2$ with $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$. It also follows by simple addition to sets in $\mathcal{A}$. By the MCT, the class of sets for which the assertion is valid form a monotone class, and since it contains $\mathcal{A}$ it also contains $\sigma(\mathcal{A}) = \mathcal{F}$. □

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

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

Sub σ-algebras $\mathcal{G}_1, \mathcal{G}_2, \ldots$ of $\mathcal{F}$ are called independent if whenever $G_i \in \mathcal{G}_i$ $(i \in \mathbb{N})$ and $i_1, i_2, \ldots, i_n$ are distinct

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

How does this fit in with our notion of independence from mods?

Definition 2.22 (σ-algebra generated by a random variable). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X be a real-valued random variable on $\Omega, \mathcal{F}, \mathbb{P})$ (that is a measurable function from $\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then*

$$\begin{aligned}\sigma(X) &= \sigma\left(\{\{\omega \in \Omega : X(\omega) \in A\}; A \in \mathcal{B}(\mathbb{R})\}\right) \\ &= \sigma\left(\left\{X^{-1}(A) : A \in \mathcal{B}(\mathbb{R})\right\}\right).\end{aligned}$$

It is the smallest sub σ-algebra of $\mathcal{F}$ with respect to which X is a measurable function.

Definition 2.23 (Independent random variables). *Random variables $X_1, X_2, \ldots$ are called* independent *if the σ-algebras $\sigma(X_1), \sigma(X_2), \ldots$ are independent.*

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

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

From this it is easy to check the following result.

Lemma 2.24. *Two random variables X and Y on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are independent* iff *the measure μ_{XY} 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 generalises the result you learned in mods and part A for discrete/continuous random variables - two continuous random variables X and Y are independent if and only if their joint density function can be written as the product of the density function of X and the density function of Y.

Of course the conditions of Definition 2.23 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 our result of Theorem 1.8 (uniqueness of extension) to reduce it to something much more manageable.

Theorem 2.25. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose that $\mathcal{G}$ and $\mathcal{H}$ are sub σ-algebras of $\mathcal{F}$ and that $\mathcal{G}_0$ ad $\mathcal{H}_0$ are π-systems with*

$$\sigma(\mathcal{G}_0) = \mathcal{G}, \quad \text{and} \quad \sigma(\mathcal{H}_0) = \mathcal{H}.$$

Then $\mathcal{G}$ and $\mathcal{H}$ are independent iff *$\mathcal{G}_0$ and $\mathcal{H}_0$ are independent, i.e. $\mathbb{P}[G \cap H] = \mathbb{P}[G]\mathbb{P}[H]$ whenever $G \in \mathcal{G}_0$, $H \in \mathcal{H}_0$.*

Proof

The two measures $H \mapsto \mathbb{P}[G \cap H]$ and $H \mapsto \mathbb{P}[G]\mathbb{P}[H]$ on $(\Omega, \mathcal{H})$ have the same total mass $\mathbb{P}[G]$ and they agree on the π-system $\mathcal{H}_0$. So by Theorem 1.8 they agree on $\sigma(\mathcal{H}_0) = \mathcal{H}$. Hence, for $G \in \mathcal{G}_0$ and $H \in \mathcal{H}$

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

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

Corollary 2.26. *A sequence $\{X_n\}_{n\in\mathbb{N}}$ of real-valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ is independent* iff *for all $x_1, \ldots x_n \in \mathbb{R}$ and $n \in \mathbb{N}$*

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

Remark 2.27 (Kolmogorov's Consistency Theorem)**.** *Notice that our notion of independence only really makes sense if our random variables are defined on a common probability space. For a countable sequence of random variables with specified distributions it is not completely clear that there exists such a probability space.*

By extending our construction of product measure in the obvious way, we can construct a measure on $\mathbb{R}^n$ for every n which is the joint distribution of the first n of the random variables. Let us denote it by $\mathbb{P}_n$. The $\mathbb{P}_n$'s are consistent *in the sense that if we project from $\mathbb{R}^{(n+1)}$ to $\mathbb{R}^n$ in the obvious way, then $\mathbb{P}_{n+1}$ projects to $\mathbb{P}_n$. Such a family is called a* consistent *family of finite dimensional distributions.*

Now look at the space $\Omega = \mathbb{R}^\infty$ of all real sequences $\omega = \{x_n : n \geq 1\}$ with the σ-algebra $\mathcal{F}$, generated by the algebra $\mathcal{A}$ of finite dimensional cylinder sets*, that is sets of the form $B = \{\omega : (x_1, \ldots, x_n) \in A\}$ where A varies over the Borel sets in $\mathbb{R}^n$ and n varies over positive integers.*

The Kolmogorov Consistency Theorem tells us that given a consistent family of finite dimensional distributions $\mathbb{P}_n$, there exists a unique $\mathbb{P}$ on $(\Omega, \mathcal{F})$ such that for every n, under the natural projection $\pi_n(\omega) = (x_1, \ldots, x_n)$, the induced measure $\mathbb{P}\pi_n^{-1} = \mathbb{P}_n$ on $\mathbb{R}^n$.

3 Tail events and modes of convergence

3.1 The Borel-Cantelli Lemmas

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

First we're going to look at Fatou's Lemma in this setting, but in the special case where the functions f_n are indicator functions of measurable sets. Recall that Fatou's Lemma says that for a sequence of non-negative measurable functions $\{f_n\}_{n\in\mathbb{N}}$

$$\mathbb{P}[\liminf_{n\to\infty} f_n] \leq \liminf_{n\to\infty} \mathbb{P}[f_n]$$

and the reverse Fatou lemma says that if $0 \leq f_n \leq g$ for some g with $\mathbb{P}[g] < \infty$ then

$$\mathbb{P}[\limsup_{n\to\infty} f_n] \geq \limsup_{n\to\infty} \mathbb{P}[f_n].$$

So what does $\liminf_{n\to\infty} f_n$ or $\limsup_{n\to\infty} f_n$ look like if the f_n's are indicator functions of sets?

Let $f_n = \mathbf{1}_{A_n}$ for some $A_n \in \mathcal{F}$.

$$\begin{aligned}
\limsup_{n\to\infty} f_n &= \lim_{n\to\infty} \sup_{m\geq n} f_m \\
&= \lim_{n\to\infty} \sup_{m\geq n} \mathbf{1}_{A_m} \\
&= \lim_{n\to\infty} \mathbf{1}_{\cup_{m\geq n} A_m} \\
&= \mathbf{1}_{\cap_{n\in\mathbb{N}}\cup_{m\geq n} A_m}.
\end{aligned}$$

This motivates the following definition.

Definition 3.1. *Let $\{A_n\}_{n\in\mathbb{N}}$ be a sequence of sets from $\mathcal{F}$. We define*

$$\begin{aligned}
\limsup_{n\to\infty} A_n &= \bigcap_{n\in\mathbb{N}} \cup_{m\geq n} A_m \\
&= \{A_m \textit{ occurs infinitely often}\} \\
&= \{\omega \in \Omega : \omega \in A_m \textit{ for infinitely many } m\}.
\end{aligned}$$

$$
\begin{aligned}
\liminf_{n\to\infty} A_n &= \bigcup_{n\in\mathbb{N}} \cap_{m\geq n} A_m \\
&= \{\omega\in\Omega : \textit{ for some } m(\omega),\ \omega\in A_m \textit{ for all } m\geq m(\omega)\} \\
&= \{\text{'}A_m \textit{ eventually'}\} \\
&= \{A_m^c \textit{ infinitely often}\}^c.
\end{aligned}
$$

Lemma 3.2.

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

The proof is an exercise (based on our calculations above).

In this terminology the Fatou and reverse Fatou Lemmas say

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

and

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

(both of which make good intuitive sense). In fact we can say more about the probabilities of these events.

Lemma 3.3 (The First Borel-Cantelli Lemma, BC1)**.** *Suppose that*

$$
\sum_{n\in\mathbb{N}} \mathbb{P}[A_n] < \infty,
$$

then

$$
\mathbb{P}[A_n\, i.o.] = 0.
$$

Remark 3.4. *Notice that we are making no assumptions about independence here. This is a very powerful result.*

Proof of BC1

Let $G_n = \cup_{m\geq n} A_m$. Then

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

and $G_n \downarrow G = \limsup_{n\to\infty} A_n$, so by Lemma 1.11 (monotone convergence properties), $\mathbb{P}[G_n] \downarrow \mathbb{P}[G]$.

On the other hand, since $\sum_{n\in\mathbb{N}} \mathbb{P}[A_n] < \infty$, we have that

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

and so

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

as required. □

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

Lemma 3.5 (The Second Borel-Cantelli Lemma, BC2)**.** *Assume that* $\{A_n\}_{n\in\mathbb{N}}$ *are* independent *events. If*

$$\sum_{n\in\mathbb{N}} \mathbb{P}[A_n] = \infty$$

then

$$\mathbb{P}[A_n\, i.o.] = 1.$$

Proof

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

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

Hence using Lemma 1.11 (monotone convergence properties) again

$$\begin{aligned}\mathbb{P}\left[\bigcup_{n\in\mathbb{N}}\bigcap_{m\ge n} A_m^c\right] &= \lim_{N\to\infty}\mathbb{P}\left[\bigcup_{i=1}^{N}\bigcap_{m\ge n} A_m^c\right]\\ &\le \lim_{N\to\infty}\sum_{n=1}^{N}\mathbb{P}\left[\bigcap_{m\ge n} A_m^c\right] = 0.\end{aligned}$$

Thus

$$\mathbb{P}\left[\liminf_{n\to\infty} A_n^c\right] = \mathbb{P}\left[\bigcup_{n\in\mathbb{N}}\bigcap_{m\ge n} A_m^c\right] = 0.$$

and since

$$\left(\liminf_{n\to\infty} A_n^c\right)^c = \limsup_{n\to\infty} A_n$$

we have

$$\mathbb{P}[\limsup_{n\to\infty} A_n] = \mathbb{P}[A_n\, i.o.] = 1$$

as required. □

Example 3.6. *A monkey is provided with a typewriter. At each time step it has probability* 1/26 *of typing any of the 26 letters independently of other times. What is the probability that it will type ABRACADABRA at least once? infinitely often?*

Solution

We can consider the events

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

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

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

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

Example 3.7. *Let $\{X_n\}_{n\geq 1}$ be independent exponentially distributed random variables with mean 1 and let $M_n = \max\{X_1, \dots, X_n\}$. Then*

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

Remark 3.8. *In fact with a little more work we can show that*

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

See, for example, S.C. Port, Theoretical Probability for Applications, Wiley 1993, Example 4.7.2 p.560.

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

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

Fix $0 < \epsilon < 1$. Then

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

Thus

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

and so by BC1

$$\mathbb{P}[M_n \leq (1-\epsilon)\log n \; i.o.] = 0.$$

Since ϵ was arbitrary this gives

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

□

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 in fact for many applications, in particular when the events are independent, these are events that can *only* have probability zero or one. This is because they are examples of what are known as 'tail' events.

Recall from Definition 2.22 that the σ-algebra generated by a random variable X is the smallest sub σ-algebra of $\mathcal{F}$ with respect to which X is measurable. That is

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

Definition 3.9 (Tail σ-algebra)**.** *For a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ define*

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

and

$$\mathcal{T} = \bigcap_{n\in\mathbb{N}} \mathcal{T}_n.$$

Then $\mathcal{T}$ is called the tail σ-algebra *of the sequence $\{X_n\}_{n\in\mathbb{N}}$.*

We can think of the tail σ-algebra as containing events describing the limiting behaviour of the sequence as $n \to \infty$.

Theorem 3.10 (Kolmogorov's 0-1 law)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of independent random variables. Then the tail σ-algebra $\mathcal{T}$ of $\{X_n\}_{n\in\mathbb{N}}$ contains only events of probability 0 or 1. Moreover, any $\mathcal{T}$-measurable random variable is almost surely constant.*

Proof

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

$$\mathcal{A} = \{\{X_1 \le x_1, \dots, X_n \le x_n\} : x_1, \dots, x_n \in \mathbb{R}\}$$

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

$$\mathcal{B} = \{\{X_{n+1} \le x_{n+1}, \dots, X_{n+k} \le x_{n+k}\} : x_{n+1}, \dots, x_{n+k} \in \mathbb{R}, k \in \mathbb{N}\}.$$

For any $A \in \mathcal{A}$, $B \in \mathcal{B}$, by the independence of the random variables $\{X_n\}_{n\in\mathbb{N}}$, we have

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

and so by Theorem 2.25 the σ-algebras $\sigma(\mathcal{A}) = \mathcal{F}_n$ and $\sigma(\mathcal{B}) = \mathcal{T}_n$ are also independent.

Since $\mathcal{T} \subseteq \mathcal{T}_n$ we conclude that $\mathcal{F}_n$ and $\mathcal{T}$ are also independent.

Now $\cup_{n\in\mathbb{N}}\mathcal{F}_n$ is a π-system (although not in general a σ-algebra) generating the σ-algebra $\mathcal{F}_\infty = \sigma(\{X_n\}_{n\in\mathbb{N}})$. So applying Theorem 2.25 again we see that $\mathcal{F}_\infty$ and $\mathcal{T}$ are independent. But $\mathcal{T} \subseteq \mathcal{F}_\infty$ so that if $A \in \mathcal{T}$

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

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

Now suppose that Y is any $\mathcal{T}$-measurable random variable. Then $F_Y(y) = \mathbb{P}[Y \le y]$ is right continuous and takes only values in $\{0, 1\}$. So $\mathbb{P}[Y = c] = 1$ where $c = \inf\{y : F_Y(y) = 1\}$. □

Example 3.11. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed (i.i.d.) random variables and let $Z_n = \sum_{k=1}^n X_k$. Consider $L = \limsup_{n\to\infty} Z_n/n$. Then L is a tail random variable and so almost surely constant. We'll prove later in the course (Theorem 9.2) that, under weak assumptions, $L = \mathbb{E}[X_1]$ almost surely.*

If the X_n in the last example have mean zero one, then setting

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

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

You may at this point be feeling a little confused. In mods probability/statistics (possibly even at school) you learned that if $\{X_n\}_{n\in\mathbb{N}}$ is a sequence of i.i.d. random variables with mean zero and variance one then

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

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

Definition 3.12 (Modes of convergence)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X a given random variable and $\{X_n\}_{n\in\mathbb{N}}$ a sequence of random variables.*

1. *We say that $\{X_n\}_{n\in\mathbb{N}}$ converges* almost surely *to X (written $X_n \stackrel{a.s.}{\rightarrow} X$) if*
$$\mathbb{P}[\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)] = 1.$$

2. *We say that $\{X_n\}_{n\in\mathbb{N}}$ converges to X* in probability *(written $X_n \stackrel{\mathbb{P}}{\rightarrow} X$) if, given $\epsilon > 0$,*
$$\lim_{n\to\infty} \mathbb{P}[\omega : |X_n(\omega) - X(\omega)| > \epsilon] = 0.$$

3. *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 $X_n \stackrel{d}{\rightarrow} X$ or $X_n \stackrel{\mathcal{L}}{\rightarrow} X$) if*
$$\lim_{n\to\infty} F_n(x) = F(x).$$

4. *Suppose that X and X_n have finite rth moment for some $r > 0$ (that is $\mathbb{E}[|X|^r] < \infty$). We say that X_n converges to X in L^r (written $X_n \stackrel{L^r}{\rightarrow} X$) if*
$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^r] = 0.$$

These notions of convergence are all different.

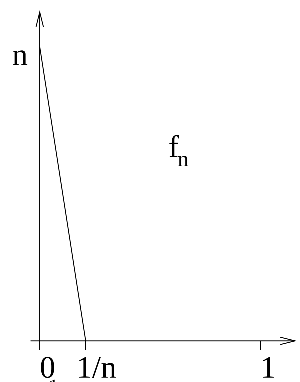
$$\text{Convergence a.s.} \implies \text{Convergence in Probability} \implies \text{Convergence in Distribution}$$

$$\Uparrow$$

$$\text{Convergence in } L^r$$

We already saw some different modes of convergence in Part A Integration. For example, for the sequence f_n given by:

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

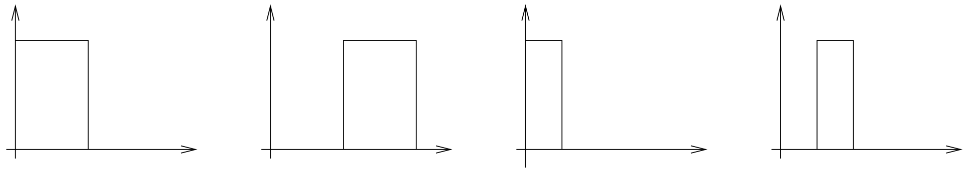


we have $f_n \to 0$ almost everywhere on $[0,1]$ but $f_n \not\to 0$ in L^1. If we think of $([0,1], \mathcal{B}([0,1]), Leb.)$ as a probability space that immediately gives us an example here of a.s convergence but not L^1 convergence.

Example 3.13 (Convergence in probability does not imply a.s. convergence)**.** *To understand what's going on in (9) and (10), let's stick with $[0,1]$ with the Borel sets and Lebesgue measure as our probability space. We define $\{X_n\}_{n\in\mathbb{N}}$ as follows:*

for each n there is a unique pair of integers (m,k) such that $n = 2^m + k$. We set
$$X_n(\omega) = \mathbf{1}_{[2^{-m}k, 2^{-m}(k+1))}(\omega).$$

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



n=2 n=3 n=4 n=5

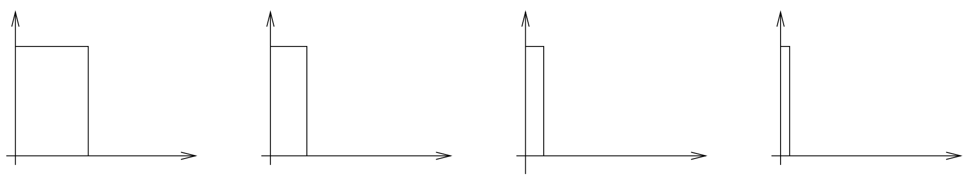
For fixed ω, $X_n(\omega) = 1\, i.o.$, *but*

$$\mathbb{P}[\omega : X_n(\omega) = 0] = 1 - \frac{1}{2^m} \to 0 \qquad as\ n \to \infty.$$

So

$$X_n \xrightarrow{\mathbb{P}} 0, \qquad but \qquad X_n \not\to 0\, a.s.\ as\ n \to \infty.$$

On the other hand, if we look at the $\{X_{2^n}\}_{n\in\mathbb{N}}$, *we have*



n=2 n=4 n=8 n=16

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

It turns out that this is a general phenomenon.

Theorem 3.14 (Convergence in Probability and a.s. Convergence)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space,* X *a given random variable and* $\{X_n\}_{n\in\mathbb{N}}$ *a sequence of random variables.*

1. *If* $X_n \xrightarrow{a.s.} X$ *then* $X_n \xrightarrow{\mathbb{P}} X$.
2. *If* $X_n \xrightarrow{\mathbb{P}} X$, *then there exists a subsequence* $\{X_{n_k}\}_{k\in\mathbb{N}}$ *such that* $X_{n_k} \xrightarrow{a.s.} X$.

Proof

Write $Y_n = X_n - X$ and

$$A_{n,\epsilon} = \{|Y_n| > \epsilon\}.$$

1. By the reverse Fatou Lemma (since $\mathbb{P}$ is a probability measure we certainly have the boundedness assumption)

$$\mathbb{P}[A_{n,\epsilon}\, i.o] = \mathbb{P}[\limsup_{n\to\infty} A_{n,\epsilon}] \geq \limsup_{n\to\infty} \mathbb{P}[A_{n,\epsilon}]$$

and since $X_n \xrightarrow{a.s.} X$, $\mathbb{P}[A_{n.\epsilon}\, i.o.] = 0$ and so certainly $\lim_{n\to\infty} \mathbb{P}[A_{n,\epsilon}] = 0$ as required.

2. This is the more interesting direction. Since $X_n \xrightarrow{\mathbb{P}} X$, given $\epsilon > 0$ there exists $N(\epsilon)$ such that

$$\mathbb{P}[A_{n,\epsilon}] < \epsilon, \qquad \forall n \geq N(\epsilon).$$

Choosing $\epsilon = 1/k^2$ and writing $n_k = N(1/k^2)$ for $k \in \mathbb{N}$,

$$\sum_{k\in\mathbb{N}} \mathbb{P}[A_{n_k,1/k^2}] < \infty$$

and so by BC1, $\mathbb{P}[A_{n_k,1/k^2}\ i.o.] = 0$. That is

$$\mathbb{P}[\limsup_{n\to\infty} A_{n_k,1/k^2}] = 0$$

which, in turn, says $X_{n_k} \stackrel{a.s.}{\to} X$ as required. □

The First Borel-Cantelli Lemma provides a very powerful tool for proving almost sure convergence of a sequence of random variables. It's successful application often rests on being able to find good bounds on the random variables $\{X_n\}_{n\in\mathbb{N}}$. We end this section with some inequalities that are often helpful in this context. The first is trivial, but has many applications.

Lemma 3.15 (Chebyshev's inequality)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X a* non-negative *random variable, then for each $\lambda > 0$*

$$\mathbb{P}[X \geq \lambda] \leq \frac{1}{\lambda}\mathbb{E}[X].$$

More generally, let Y be any random variable (not necessarily non-negative) and let $\phi : \mathbb{R} \to [0, \infty]$ be non-decreasing and measurable. Then for any $\lambda \in \mathbb{R}$,

$$\begin{aligned} \mathbb{P}[Y \geq \lambda] &= \mathbb{P}[\phi(Y) \geq \phi(\lambda)] \\ &\leq \frac{1}{\phi(\lambda)}\mathbb{E}[\phi(Y)]. \end{aligned}$$

The second inequality is often applied with $\phi(x) = e^{\theta x}$ to obtain

$$\mathbb{P}[Y \geq \lambda] \leq e^{-\theta\lambda}\mathbb{E}[e^{\theta Y}]$$

and then optimised over θ.

For the next inequality we recall

Definition 3.16 (Convex function)**.** *Let $I \subseteq \mathbb{R}$ be an interval. A function $c : I \to \mathbb{R}$ is* convex *if for all $x, y \in I$ and $t \in [0, 1]$,*

$$c\left(tx + (1-t)y\right) \leq tc(x) + (1-t)c(y).$$

Theorem 3.17 (Jensen's inequality)**.** *Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and X an integrable random variable taking values in I. Let $c : I \to \mathbb{R}$ be convex. Then*

$$\mathbb{E}[c(X)] \geq c\left(\mathbb{E}[X]\right).$$

Important examples of convex functions include x^2, e^x, $1/x$. To check that a twice continuously differentiable function is convex, it suffices to check that $c''(x) > 0$ for all x.

The proof of Theorem 3.17 rests on the following lemma.

Lemma 3.18. *Suppose that $c : I \to \mathbb{R}$ is convex and let m be an interior point of I. Then there exits $a, b \in \mathbb{R}$ such that $c(x) \geq ax + b$ for all x with equality at $x = m$.*

Proof

For $x < m < y$, by convexity,

$$c(m) \leq \frac{(m-x)}{(y-x)}c(y) + \frac{(y-m)}{(y-x)}c(x).$$

Rearranging,

$$\frac{c(m)-c(x)}{m-x} \leq \frac{c(y)-c(m)}{y-m}.$$

So for an interior point m, since the left hand side does not depend on y and the right hand side does not depend on x,

$$\sup_{x<m} \frac{c(m)-c(x)}{m-x} \leq \inf_{y>m} \frac{c(y)-c(m)}{y-m}$$

and choosing a so that

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

we have that $c(x) \geq c(m) + a(x-m)$ for all $x \in I$. $\square$

Proof of Theorem 3.17

Since $\mathbb{E}[X]$ is certainly an interior point of I (other than in the trivial case X is almost surely constant), set $m = \mathbb{E}[X]$ in the previous lemma and we have

$$c(X) \geq c\left(\mathbb{E}[X]\right) + a(X - \mathbb{E}[X]).$$

Now take expectations to recover

$$\mathbb{E}[c(X)] \geq c\left(\mathbb{E}[X]\right)$$

as required. $\square$

4 Conditional Expectation

Probability is a measure of ignorance. When new information decreases that ignorance we change our probabilities. We formalised this in mods through Bayes' rule. For a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $A, B \in \mathcal{F}$

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

We want now to introduce an extension of this which lies at the heart of martingale theory: the notion of conditional expectation. First a preliminary definition:

Definition 4.1 (Equivalence class of a random variable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X a random variable. The* equivalence class *of X is the collection of random variables that differ from X only on a null set.*

Definition 4.2 (Conditional Expectation)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X be an integrable random variable (that is one for which $\mathbb{E}[|X|] < \infty$). Let $\mathcal{G}$ be a sub σ-algebra of $\mathcal{F}$. The* conditional expectation $\mathbb{E}[X|\mathcal{G}]$ *is any $\mathcal{G}$-measurable, integrable random variable Z in the equivalence class of random variables such that*

$$\int_\Lambda Z d\mathbb{P} = \int_\Lambda X d\mathbb{P} \quad \text{for any } \Lambda \in \mathcal{G}.$$

The integrals of X and Z over sets $\Lambda \in \mathcal{G}$ are the same, but X is $\mathcal{F}$-measurable whereas Z is $\mathcal{G}$-measurable. The conditional expectation satisfies

$$\int_\Lambda \mathbb{E}[X|\mathcal{G}] d\mathbb{P} = \int_\Lambda X d\mathbb{P} \quad \text{for any } \Lambda \in \mathcal{G} \tag{11}$$

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

Just as probability of an event is a special case of expectation (corresponding to integrating an indicator function rather than a general measurable function), so *conditional probability* is a special case of *conditional expectation.* In that case (11) becomes

$$\int_\Lambda \mathbb{P}[A|\mathcal{G}]d\mathbb{P} = \mathbb{P}[A \cap \Lambda] \quad \text{for any } \Lambda \in \mathcal{G}. \tag{12}$$

Let's see how this fits with our understanding from mods. Suppose that X is a discrete random variable taking values $\{x_n\}_{n\in\mathbb{N}}$. Then the events $\{X = x_n\}$ are a *partition* of Ω (that is Ω is a disjoint union of these events.) So, by the *Partition Theorem* of mods,

$$\begin{aligned}
\mathbb{P}[A] &= \mathbb{P}\left[\bigcup_{n\in\mathbb{N}} (A \cap \{X = x_n\})\right] \\
&= \sum_{n\in\mathbb{N}} \mathbb{P}[A \cap \{X = x_n\}] \\
&= \sum_{n\in\mathbb{N}} \mathbb{P}[A|X = x_n]\mathbb{P}[X = x_n].
\end{aligned}$$

Now we 'randomize' - replace $\mathbb{P}[X = x_n]$ by $\mathbf{1}_{\{X=x_n\}}$ and we write

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

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

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

To see that this coincides with (12), notice that if $\Lambda \in \sigma(X)$ then it can be expressed as a union of sets of the form $\{X = x_n\}$ (the advantage with working with discrete random variables again - the σ-algebra is easy) and for such Λ

$$\begin{aligned}
\int_\Lambda \left(\sum_{n=1}^{\infty} \mathbb{P}[A|X = x_n]\mathbf{1}_{\{X=x_n\}}\right) d\mathbb{P} &= \sum_{n:x_n\in\Lambda} \mathbb{P}[A|X = x_n]\mathbb{P}[X = x_n] \\
&= \sum_{n:x_n\in\Lambda} \mathbb{P}[A \cap \{X = x_n\}] \\
&= \mathbb{P}[A \cap \Lambda].
\end{aligned}$$

This would have worked equally well for any other partition in place of $\{\{X = x_n\}\}_{n\in\mathbb{N}}$. So more generally, let $\{\Lambda_n\}_{n\in\mathbb{N}}$ be a partition of Ω and let $\mathbb{E}[X|\Lambda_n]$ be the conditional expectation relative to the conditional measure $\mathbb{P}[\cdot|\Lambda_n]$ so that

$$\mathbb{E}[X|\Lambda_n] = \int_\Omega X(\omega)d\mathbb{P}[\omega|\Lambda_n] = \frac{\int_{\Lambda_n} X d\mathbb{P}}{\mathbb{P}[\Lambda_n]}.$$

Then for $\Lambda = \sum_{j\in J} \Lambda_j \in \mathcal{G}$ we obtain (using the disjointness of the Λ_j's)

$$\begin{aligned}
\int_\Lambda \left(\sum_{n=1}^\infty \mathbb{E}[X|\Lambda_n]\mathbf{1}_{\Lambda_n}\right) d\mathbb{P} &= \sum_{j\in J}\sum_{n=1}^\infty \int_{\Lambda_j} \mathbb{E}[X|\Lambda_n]\mathbf{1}_{\Lambda_n}\mathbf{1}_{\Lambda_j} d\mathbb{P} \text{ (the summand is zero if } n \neq j) \\
&= \sum_{j\in J} \int_{\Lambda_j} \mathbb{E}[X|\Lambda_j] d\mathbb{P} \\
&= \sum_{j\in J} \mathbb{E}[X|\Lambda_j]\mathbb{P}[\Lambda_j] \\
&= \sum_{j\in J} \frac{\int_{\Lambda_j} X d\mathbb{P}}{\mathbb{P}[\Lambda_j]}\mathbb{P}[\Lambda_j] \\
&= \sum_{j\in J} \int_{\Lambda_j} X d\mathbb{P} \\
&= \int_{\cup_j \Lambda_j} X d\mathbb{P} = \int_\Lambda X d\mathbb{P}.
\end{aligned}$$

So in this case

$$\mathbb{E}[X|\mathcal{G}] = \sum_{n=1}^\infty \mathbb{E}[X|\Lambda_n]\mathbf{1}_{\Lambda_n} \quad a.s.,$$

or, spelled out, that

$$\mathbb{E}[X|\mathcal{G}] = \begin{cases} \mathbb{E}[X|\Lambda_1] & \text{if } \omega \in \Lambda_1, \\ \mathbb{E}[X|\Lambda_2] & \text{if } \omega \in \Lambda_2, \\ \cdots & \cdots \\ \mathbb{E}[X|\Lambda_n] & \text{if } \omega \in \Lambda_n, \\ \cdots & \cdots \end{cases}$$

So $\mathbb{E}[X|\mathcal{G}]$ is constant on each set Λ_i (where it takes the value $\mathbb{E}[X|\Lambda_i]$).

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

Proposition 4.3 (Almost sure uniqueness of conditional expectation). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X an integrable random variable and $\mathcal{G}$ a sub σ-algebra of $\mathcal{F}$. If Y and Z are two $\mathcal{G}$-measurable random variables that both satisfy the defining relation (11), then $\mathbb{P}[Y \neq Z] = 0$. That is Y and Z agree up to a null set.*

Proof

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

$$\Lambda_1 = \{\omega : Y(\omega) < Z(\omega)\} \in \mathcal{G},$$

so using the defining relation

$$\int_{\Lambda_1} (Y - Z) d\mathbb{P} = 0$$

which implies $\mathbb{P}[\Lambda_1] = 0$.

Similarly,

$$\Lambda_2 = \{\omega : Y(\omega) > Z(\omega)\} \in \mathcal{G}$$

and

$$\int_{\Lambda_2} (Y - Z)d\mathbb{P} = 0$$

gives $\mathbb{P}[\Lambda_2] = 0$ which completes the proof. $\square$

For existence in the general case we will use another important result from measure theory. First a definition.

Definition 4.4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathbb{Q}$ be a finite measure on $(\Omega, \mathcal{F})$. The measure $\mathbb{Q}$ is* absolutely continuous *with respect to* $\mathbb{P}$ iff

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

We write $\mathbb{Q} \ll \mathbb{P}$.

Theorem 4.5 (The Radon-Nikodym Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and suppose that $\mathbb{Q}$ is a finite measure that is absolutely continuous with respect to $\mathbb{P}$. Then there exists an $\mathcal{F}$-measurable random variable Z with finite mean such that*

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

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

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

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

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

Proof

Let $\mathbb{P}|_{\mathcal{G}}$ denote the measure $\mathbb{P}$ restricted to the sub σ-algebra $\mathcal{G}$. Set

$$\mathbb{Q}[A] = \int_A X d\mathbb{P} \qquad \text{for } A \in \mathcal{G}.$$

Then $\mathbb{Q} \ll \mathbb{P}|_{\mathcal{G}}$ and so the Radon-Nikodym Theorem applies to $\mathbb{Q}$, $\mathbb{P}|_{\mathcal{G}}$ on $(\Omega, \mathcal{G})$. Write

$$\mathbb{E}[X|\mathcal{G}] = \frac{d\mathbb{Q}}{d\mathbb{P}|_{\mathcal{G}}}.$$

$\square$

It is much harder to write out $\mathbb{E}[X|\mathcal{G}]$ explicitly when $\mathcal{G}$ is not generated by a partition. But note that if $\mathcal{G} = \sigma(Y)$ for some random variable Y on $(\Omega, \mathcal{F}, \mathbb{P})$, then any $\mathcal{G}$-measurable function can, in principle, be written as a function of Y. We saw an example of this with our branching process in §0.2. If Z_n was the number of descendants of a single ancestor after n generations, then

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

where μ is the expected number of offspring of a single individual.

In general, of course, the relationship can be much more complicated.

Exercise 4.7. *Roll a fair die until we get a six. Let Y be the total number of rolls and X the number of 1's. Show that*

$$\mathbb{E}[X|Y] = \frac{1}{5}(Y-1) \quad \textit{and} \quad \mathbb{E}[X^2|Y] = \frac{1}{25}(Y^2+2Y-3).$$

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 $(\Omega, \mathcal{G})$ and, since conditional expectation is only defined up to equivalence (so up to null sets) we have to qualify many of our statements with the caveat 'a.s.'.

Proposition 4.8. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X and Y integrable random variables, $\mathcal{G} \subseteq \mathcal{F}$ a sub σ-algebra and a, b, c real numbers. Then*

1. $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X]$.
2. $\mathbb{E}[aX + bY|\mathcal{G}] \overset{a.s.}{=} a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}]$.
3. *If X is $\mathcal{G}$-measurable, then* $\mathbb{E}[X|\mathcal{G}] \overset{a.s.}{=} X$.
4. $\mathbb{E}[c|\mathcal{G}] \overset{a.s.}{=} c$.
5. $\mathbb{E}[X|\{\emptyset, \Omega\}] = \mathbb{E}[X]$.
6. *If $X \leq Y$ a.s. then $\mathbb{E}[X|\mathcal{G}] \leq \mathbb{E}[Y|\mathcal{G}]$ a.s.*
7. $|\mathbb{E}[X|\mathcal{G}]| \leq \mathbb{E}[|X||\mathcal{G}]$.
8. *If X is independent of $\mathcal{G}$ then $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$ a.s.*

Proof

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

1. Set $\Lambda = \mathbb{R}$ in the defining relation.

2.

$$\begin{aligned}
\int_\Lambda \mathbb{E}[aX + bY|\mathcal{G}]d\mathbb{P} &= \int_\Lambda (aX + bY)d\mathbb{P} \\
&= a\int_\Lambda X d\mathbb{P} + b\int_\Lambda Y d\mathbb{P} \quad \text{(linearity of the integral)} \\
&= a\int_\Lambda \mathbb{E}[X|\mathcal{G}]d\mathbb{P} + b\int_\Lambda \mathbb{E}[Y|\mathcal{G}]d\mathbb{P} \\
&= \int_\Lambda (a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}])d\mathbb{P},
\end{aligned}$$

where the last line again follows by linearity of the integral. And if two $\mathcal{G}$-measurable functions agree on integration over any $\mathcal{G}$-measurable set then they are $\mathbb{P}$-a.s equal.

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

Jumping to 8. Note that $\mathbb{E}[X]$ is $\mathcal{G}$-measurable and for $\Lambda \in \mathcal{G}$

$$\begin{aligned}
\int_\Lambda \mathbb{E}[X]d\mathbb{P} &= \mathbb{E}[X]\mathbb{P}[\Lambda] = \mathbb{E}[X]\mathbb{E}[\mathbf{1}_\Lambda] \\
&= \mathbb{E}[X\mathbf{1}_\Lambda] \quad \text{(by independence)} \\
&= \int X\mathbf{1}_\Lambda d\mathbb{P} = \int_\Lambda X d\mathbb{P},
\end{aligned}$$

so the defining relation holds. $\square$

Notice that 8 is intuitively clear. If X is independent of $\mathcal{G}$, then telling me about events in $\mathcal{G}$ tells me nothing about X and so my assessment of its expectation does not change. On the other hand for 3, if X is $\mathcal{G}$-measurable, then telling me about events in $\mathcal{G}$ actually tells me the value of X.

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

Proposition 4.9 (Conditional Convergence Theorems)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\{X_n\}_{n\in\mathbb{N}}$ a sequence of integrable random variables, X a random variable and $\mathcal{G}$ a sub σ-algebra of $\mathcal{F}$.*

1. **cMON:** *If $X_n \uparrow X$ as $n \to \infty$, then $\mathbb{E}[X_n|\mathcal{G}] \uparrow \mathbb{E}[X|\mathcal{G}]$ a.s. as $n \to \infty$.*
2. **cFatou:** *If $\{X_n\}_{n\in\mathbb{N}}$ are non-negative then*
$$\mathbb{E}[\liminf_{n\to\infty} X_n|\mathcal{G}] \leq \liminf_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}]\, a.s.$$
3. *If $\{X_n\}_{n\in\mathbb{N}}$ are non-negative and $X_n \leq Z$ for all n where Z is an integrable random variable then*
$$\mathbb{E}[\limsup_{n\to\infty} X_n|\mathcal{G}] \geq \limsup_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}]\, a.s.$$
4. **cDOM:** *If Y is an integrable random variable and $|X_n| \leq Y$ for all n and $X_n \stackrel{a.s.}{\to} X$ then*
$$\mathbb{E}[X_n|\mathcal{G}] \stackrel{a.s.}{\to} \mathbb{E}[X|\mathcal{G}] \qquad as\ n \to \infty.$$

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

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

Proposition 4.10. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X, Y integrable random variables. Let $\mathcal{G}$ be a sub σ-algebra of $\mathcal{F}$ and suppose that Y is $\mathcal{G}$-measurable. Then*
$$\mathbb{E}[XY|\mathcal{G}] \stackrel{a.s.}{=} Y\mathbb{E}[X|\mathcal{G}].$$

Proof

We use the 'standard machine'.

First suppose that X and Y are non-negative. If $Y = \mathbf{1}_A$ for some $A \in \mathcal{G}$, then for any $\Lambda \in \mathcal{G}$ we have $\Lambda \cap A \in \mathcal{G}$ and so by the defining relation (11)
$$\int_\Lambda Y\mathbb{E}[X|\mathcal{G}]d\mathbb{P} = \int_{\Lambda\cap A} \mathbb{E}[X|\mathcal{G}]d\mathbb{P} = \int_{\Lambda\cap A} X d\mathbb{P} = \int_\Lambda YX d\mathbb{P}.$$
Now extend by linearity to simple random variables Y. Next if $\{Y_n\}_{n\geq 1}$ are simple random variables with $Y_n \uparrow Y$ as $n \to \infty$, it follows that $Y_n X \uparrow YX$ and $Y_n\mathbb{E}[X|\mathcal{G}] \uparrow Y\mathbb{E}[X|\mathcal{G}]$ from which we deduce the result by the MCT. Finally, for X, Y not necessarily non-negative, write $XY = (X^+ - X^-)(Y^+ - Y^-)$ and use linearity of the integral. $\square$

Proposition 4.11 (Tower property of conditional expectations)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X an integrable random variable and $\mathcal{F}_1$, $\mathcal{F}_2$ sub σ-algebras of $\mathcal{F}$ with $\mathcal{F}_1 \subseteq \mathcal{F}_2$. Then*
$$\mathbb{E}\left[\mathbb{E}[X|\mathcal{F}_2]|\,\mathcal{F}_1\right] = \mathbb{E}[X|\mathcal{F}_1] \quad a.s.$$
In other words, writing $X_i = \mathbb{E}[X|\mathcal{F}_i]$,
$$X_1 = \mathbb{E}[X_2|\mathcal{F}_1] \quad a.s.$$

This extends Part 5 of Proposition 4.8 which dealt with the case $\mathcal{F}_1 = \{\emptyset, \Omega\}$.

The usefulness of this result mirrors that of the 'partition theorem' of mods probability theory. If sets $A_1, \ldots, A_n$ partition Ω then for any random variable X,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X|A_i]\mathbb{P}[A_i].$$

Proof of Proposition 4.11

Choose $\Lambda \in \mathcal{F}_1$ and observe that automatically $\Lambda \in \mathcal{F}_2$. Applying the defining relation (three times) gives

$$\int_\Lambda \mathbb{E}\left[\mathbb{E}[X|\mathcal{F}_2]|\,\mathcal{F}_1\right] d\mathbb{P} = \int_\Lambda \mathbb{E}[X|\mathcal{F}_2]d\mathbb{P} = \int_\Lambda X d\mathbb{P} = \int_\Lambda \mathbb{E}[X|\mathcal{F}_1]d\mathbb{P}.$$

□

Jensen's inequality also extends to the conditional case.

Proposition 4.12 (Conditional Jensen's Inequality)**.** *Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and that X is an integrable random variable taking values in an open interval $I \subseteq \mathbb{R}$. Let $c : I \to \mathbb{R}$ be convex and let $\mathcal{G}$ be a sub σ-algebra of $\mathcal{F}$. If $\mathbb{E}[|c(X)|] < \infty$ then*

$$\mathbb{E}[c(X)|\mathcal{G}] \geq c\left(\mathbb{E}[X|\mathcal{G}]\right) \quad a.s.$$

Proof

Recall from our proof of Jensen's inequality that if c is convex, then for $x < m < y \in I$

$$\frac{c(m) - c(x)}{m - x} \leq \frac{c(y) - c(m)}{y - m}. \tag{13}$$

Letting $x \uparrow m$ and writing

$$(D_- c)(q) = \lim_{r \uparrow q} \frac{c(q) - c(r)}{q - r}$$

for the left derivative of c at the point q, we see that

$$c(y) \geq \sup_{m \in I} \left\{(D_- c)(m)(y - m) + c(m)\right\}$$

and in fact we have equality (by setting $y = m$ on the right hand side).

(Existence of $(D_-(c)$ follows from (13) which also automatically guarantees continuity of c).

In particular, there exists a pair of sequences $\{a_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}}$ of real numbers such that

$$c(x) = \sup_n \{a_n x + b_n\} \quad \text{ for } x \in I.$$

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

$$\mathbb{E}[c(X)|\mathcal{G}] \geq a_n \mathbb{E}[X|\mathcal{G}] + b_n \quad a.s. \tag{14}$$

Since the union of a countable union of null sets is null (Part A Integration) we can arrange for (14) to hold simultaneously for all $n \in \mathbb{N}$ except possibly on a null set and so

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

□

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

$$\mathbb{E}[X^2|\mathcal{G}] \geq \mathbb{E}[X|\mathcal{G}]^2.$$

This leads to another interesting characteristaion of $\mathbb{E}[X|\mathcal{G}]$.

Remark 4.13 (Conditional Expectation and Mean Square Approximation)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X, Y square integrable random variables. Let $\mathcal{G}$ be a sub σ-algebra of $\mathcal{F}$ and suppose that Y is $\mathcal{G}$-measurable. Then*

$$\begin{aligned} \mathbb{E}[(Y-X)^2] &= \mathbb{E}\left[\{Y - \mathbb{E}[X|\mathcal{G}] + \mathbb{E}[X|\mathcal{G}] - X\}^2\right] \\ &= \mathbb{E}[(Y-\mathbb{E}[X|\mathcal{G}])^2] + \mathbb{E}[(\mathbb{E}[X|\mathcal{G}]-X)^2] + 2\mathbb{E}[(Y-\mathbb{E}[X|\mathcal{G}])(\mathbb{E}[X|\mathcal{G}]-X)]. \end{aligned}$$

Now Y is $\mathcal{G}$-measurable and so, using Proposition 4.8 part 1 and Proposition 4.10 we have

$$\begin{aligned} \mathbb{E}[(Y-\mathbb{E}[X|\mathcal{G}])(\mathbb{E}[X|\mathcal{G}]-X)] &= \mathbb{E}\left[\mathbb{E}[(Y-\mathbb{E}[X|\mathcal{G}])(\mathbb{E}[X|\mathcal{G}]-X)|\mathcal{G}\right] \\ &= \mathbb{E}\left[(Y-\mathbb{E}[X|\mathcal{G}])\left(\mathbb{E}[\mathbb{E}[X|\mathcal{G}]-X|\mathcal{G}]\right)\right] = 0, \end{aligned}$$

and so the cross-terms vanish.

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

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

We are now, finally, in a position to introduce martingales.

5 Martingales

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

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

We begin with some terminology.

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

Usually n is interpreted as time and $\mathcal{F}_n$ represents knowledge accumulated by time n (we never forget anything).

Definition 5.2 (Adapted stochastic process)**.** *A* stochastic process, *$\{X_n\}_{n\geq 0}$ is a collection of random variables defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.*

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

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

1. *a* martingale *if* $\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n$ *a.s. for* $n \geq 0$,

2. *a* submartingale *if* $\mathbb{E}[X_{n+1}|\mathcal{F}_n] \geq X_n$ *a.s. for* $n \geq 0$,

3. *a* supermartingale *if* $\mathbb{E}[X_{n+1}|\mathcal{F}_n] \leq X_n$ *a.s. for* $n \geq 0$.

If we think of X_n as our accumulated fortune when we make a sequence of bets, then a martingale represents a fair game in the sense that the conditional expectation of $X_{n+1} - X_n$, given our knowledge at the time when we make the $(n+1)$st bet (that is $\mathcal{F}_n$), is zero. A submartingale represents a favourable game and a supermartingale an unfavourable game.

Note that the concept of martingale makes *no sense* unless we specify the filtration. Very often, if a filtration is not specified, it is implicitly assumed that the *natural filtration* is intended.

Definition 5.4 (Natural filtration)**.** *The* natural filtration *associated with a stochastic process* $\{X_n\}_{n\geq 0}$ *on the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is defined by*

$$\mathcal{F}_n = \sigma(X_0, X_1, \dots, X_n), \quad n \geq 0.$$

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

Here are some elementary properties.

Proposition 5.5. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space.*

1. *A stochastic process* $\{X_n\}_{n\geq 0}$ *on* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a submartingale w.r.t. the filtration* $\{\mathcal{F}_n\}_{n\geq 0}$ *if and only if* $\{-X_n\}_{n\geq 0}$ *is a supermartingale. It is a martingale if and only if it is both a martinagle and a submartingale.*

2. *If* $\{X_n\}_{n\geq 0}$ *is a martingale w.r.t.* $\{\mathcal{F}_n\}_{n\geq 0}$ *then*

$$\mathbb{E}[X_n] = \mathbb{E}[X_0] \quad \textit{for all } n.$$

If $\{X_n\}_{n\geq 0}$ *is a submartingale and* $m < n$ *then*

$$X_m \leq \mathbb{E}[X_n|\mathcal{F}_m] \; a.s.$$

and

$$\mathbb{E}[X_m] \leq \mathbb{E}[X_n].$$

If $\{X_n\}_{n\geq 0}$ *is a supermartingale and* $m < n$ *then*

$$X_m \geq \mathbb{E}[X_n|\mathcal{F}_m] \; a.s.$$

and

$$\mathbb{E}[X_m] \geq \mathbb{E}[X_n].$$

3. *If* $\{X_n\}_{n\geq 0}$ *is a submartingale w.r.t. some filtration* $\{\mathcal{F}_n\}_{n\geq 0}$*, then it is also a submartingale with respect to the natural filtration* $\mathcal{G}_n = \sigma(X_0, \dots, X_n)$.

Proof

1 is obvious.

2. We prove the result when $\{X_n\}_{n\geq 0}$ is a submartingale w.r.t $\{\mathcal{F}_n\}_{n\geq 0}$.

The result is true for $n = m+1$ by definition. Suppose that it is true for $n = m+k$ for some $k \in \mathbb{N}$. Then

$$X_{m+k} \leq \mathbb{E}[X_{m+k+1}|\mathcal{F}_{m+k}]$$

(by definition) and so (by the inductive hypothesis)

$$X_m \leq \mathbb{E}\left[\mathbb{E}[X_{m+k+1}|\mathcal{F}_{m+k}]\right|\mathcal{F}_m]$$

and since $\mathcal{F}_m \subseteq \mathcal{F}_{m+k}$, the tower property gives

$$X_m \leq \mathbb{E}[X_{m+k+1}|\mathcal{F}_m]i \quad a.s.$$

and the result follows by induction. For the second conclusion, take expectations.

3. $\{X_n\}_{n\geq 0}$ is adapted to its natural filtration $\{\mathcal{G}_n\}_{n\geq 0}$ and since (by definition) $\mathcal{G}_n$ is the smallest σ-algebra with respect to which $\{X_0, \ldots, X_n\}$ are all measurable, $\mathcal{G}_n \subseteq \mathcal{F}_n$. Thus, by the tower property,

$$\mathbb{E}[X_n|\mathcal{G}_m] = \mathbb{E}\left[\mathbb{E}[X_n|\mathcal{F}_m]\right|\mathcal{G}_m] \geq \mathbb{E}[X_m|\mathcal{G}_m] = X_m.$$

□

Proposition 5.6. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose that $\{X_n\}_{n\geq 0}$ is a martingale with respect to the filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Let c be a* convex *function on $\mathbb{R}$. If $c(X_n)$ is an integrable random variable for each $n \geq 0$, then $\{c(X_n)\}_{n\geq 0}$ is a* sub*martingale w.r.t $\{\mathcal{F}_n\}_{n\geq 0}$.*

Proof

By Jensen's inequality for conditional expectations

$$\begin{aligned} c(X_m) &= c\left(\mathbb{E}[X_n|\mathcal{F}_m]\right) \quad \text{(martingale property)} \\ &\leq \mathbb{E}[c(X_n)|\mathcal{F}_m] \quad \text{(Jensen's inequality).} \end{aligned}$$

□

Corollary 5.7. *If $\{X_n\}_{n\geq 0}$ is a martingale w.r.t. $\{\mathcal{F}_n\}_{n\geq 0}$ then (subject to integrability) $\{|X_n|\}_{n\geq 0}$, $\{X_n^2\}_{n\geq 0}$, $\{e^{X_n}\}_{n\geq 0}$, $\{e^{-X_n}\}_{n\geq 0}$, $\{\max(X_n, K)\}_{n\geq 0}$ are all submartingales w.r.t. $\{\mathcal{F}_n\}_{n\geq 0}$.*

Example 5.8 (Sums of independent random variables)**.** *Suppose that $Y_1, Y_2, \ldots$ are independent random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and that $\mathbb{E}[Y_n] = 0$ for each n. Define*

$$X_n = \sum_{k=1}^{n} Y_k, \qquad X_0 = 0.$$

Then $\{X_n\}_{n\geq 0}$ is a martingale with respect to the natural σ-algebra

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

In this sense martingales generalise the notion of sums of independent random variables with mean zero. The independent random variables $\{Y_i\}_{i\in\mathbb{N}}$ of Example 5.8 can be replaced by martingale differences (which are not necessarily independent).

Definition 5.9 (Martingale differences)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* $\{\mathcal{F}_n\}_{n\geq 0}$ *a filtration. A sequence* $\{U_n\}_{n\in\mathbb{N}}$ *of integrable random variables, adapted to the filtration* $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ *is called a* martingale difference sequence *if*

$$\mathbb{E}[U_{n+1}|\mathcal{F}_n] = 0 \quad a.s. \quad \text{for all } n \geq 0.$$

Example 5.10. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and let* $\{X_n\}_{n\in\mathbb{N}}$ *be a sequence of independent, non-negative random variables with* $\mathbb{E}[X_n] = 1$ *for all* n*. Define*

$$M_0 = 1, \qquad M_n = \prod_{i=1}^{n} X_i \quad \text{for } n \geq 1.$$

Let $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ *be the natural filtration associated with* $\{M_n\}_{n\in\mathbb{N}}$*. Then* $\{M_n\}_{n\in\mathbb{N}}$ *is a martingale. (Exercise).*

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

Example 5.11. *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and let* $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ *be a filtration. Let* X *be an integrable random variable (that is* $\mathbb{E}[|X|] < \infty$*). Then setting*

$$X_n = \mathbb{E}[X|\mathcal{F}_n], \quad n \geq 1,$$

$\{X_n\}_{n\in\mathbb{N}}$ *is a martingale w.r.t.* $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$*. This is an easy consequence of the tower property. Indeed*

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

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

Definition 5.12 (Predictable process)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ *a filtration. A sequence* $\{U_n\}_{n\in\mathbb{N}}$ *of random variables is* predictable *with respect to* $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ *if* U_n *is measurable with respect to* $\mathcal{F}_{n-1}$ *for all* $n \geq 1$*.*

Example 5.13 (Discrete stochastic integral or martingale transform)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ *a filtration. Let* $\{Y_n\}_{n\in\mathbb{N}}$ *be a martingale with difference sequence* $\{U_n\}_{n\geq 1}$*. Suppose that* $\{v_n\}_{n\geq 1}$ *is a* predictable *sequence. Set*

$$X_0 = 0, \qquad X_n = \sum_{k=1}^{n} U_k v_k \quad \text{for } n \geq 1.$$

The sequence $\{X_n\}_{n\geq 1}$ *is called a* martingale transform *and is itself a martingale. It is a discrete version of the stochastic integral. To check the martingale property:*

$$\begin{aligned}
\mathbb{E}[X_{n+1}|\mathcal{F}_n] &= \mathbb{E}\left[\sum_{k=1}^{n} U_k v_k \Big| \mathcal{F}_n\right] + \mathbb{E}[U_{n+1}v_{n+1}|\mathcal{F}_n] \\
&= X_n + v_{n+1}\mathbb{E}[U_{n+1}|\mathcal{F}_n] \quad \textit{(using Proposition 4.10)} \\
&= X_n.
\end{aligned}$$

Typical examples of predictable sequences appear in gambling or finance contexts where they might constitute strategies for future action. The strategy is then based on the current state of affairs. If, for example, $(k-1)$ rounds of some gambling game have just been completed, then the strategy for the kth round is $v_k \in \mathcal{F}_{k-1}$. The change in fortune in the kth round is then $U_k v_k$.

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

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

Example 5.14. *Let $\{Y_i\}_{i\in\mathbb{N}}$ be independent random variables such that $\mathbb{E}[Y_i] = \mu_i$, $\mathtt{var}(Y_i) = \mathbb{E}[Y_i^2] - \mathbb{E}[Y_i]^2 = \sigma_i^2$. Let*

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

(That is $s_n^2 = \mathtt{var}(\sum_{i=1}^n Y_i)$ by independence.) Take $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ to be the natural filtration generated by $\{Y_n\}_{n\in\mathbb{N}}$.

It is easy to check that

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

is a martingale (just by modifying Example 5.8) and so by Proposition 5.6, since $c(x) = x^2$ is a convex functions, $\{X_n^2\}_{n\in\mathbb{N}}$ is a submartingale. But we can recover a martingale from it by compensation*:*

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

is a martingale *with respect to $\{\mathcal{F}_n\}_{n\geq 1}$.*

Proof

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

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

since $\mathbb{E}[Y_{n+1}|\mathcal{F}_n] = 0$ and $\mathbb{E}[Y_{n+1}^2|\mathcal{F}_{n+1}] = \sigma_{n+1}^2$. □

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

Theorem 5.15 (Doob's Decomposition Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ a filtration. Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of integrable random variables, adapted to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$. Then*

1. *$\{X_n\}_{n\in\mathbb{N}}$ has a* Doob decomposition

$$X_n = X_0 + M_n + A_n \tag{15}$$

where $\{M_n\}_{n\in\mathbb{N}}$ is a martingale and $\{A_n\}_{n\in\mathbb{N}}$ is a predictable process and $M_0 = 0 = A_0$. Moreover, if $X_n = X_0 + \widetilde{M}_n + \widetilde{A}_n$ is another Doob decompositon of $\{X_n\}_{n\in\mathbb{N}}$ then

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

2. *$\{X_n\}_{n\in\mathbb{N}}$ is a supermartingale if and only if $\{A_n\}_{n\in\mathbb{N}}$ in (15) is a decreasing process and a submartingale if and only if $\{A_n\}_{n\in\mathbb{N}}$ is an increasing process.*

Proof

Let

$$M_n = \sum_{k=1}^{n} \left(X_k - \mathbb{E}[X_k|\mathcal{F}_{k-1}]\right), \quad \text{and } A_n = X_n - M_n, \quad n \geq 1.$$

Then, since

$$\mathbb{E}[X_k - \mathbb{E}[X_k|\mathcal{F}_{k-1}]|\mathcal{F}_{k-1}] = 0,$$

the process $\{M_n\}_{n\in\mathbb{N}}$ is a martingale. We must check that $\{A_n\}_{n\in\mathbb{N}}$ is *predictable*. But

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

which is $\mathcal{F}_{n-1}$-measurable.

That establishes *existence* of a decomposition. For uniqueness, suppose that $X_n = X_0 + \widetilde{M}_n + \widetilde{A}_n$. Then by predictability,

$$\begin{aligned}
\widetilde{A}_{n+1} - \widetilde{A}_n &= \mathbb{E}[\widetilde{A}_{n+1} - \widetilde{A}_n|\mathcal{F}_n] \\
&= \mathbb{E}[(X_{n=1} - X_n) - (\widetilde{M}_{n+1} - \widetilde{M}_n)|\mathcal{F}_n] \\
&= \mathbb{E}[X_{n+1}|\mathcal{F}_n] - X_n \\
&= A_{n+1} - A_n,
\end{aligned}$$

which combined with $A_0 = 0 = \widetilde{A}_0$ proves uniqueness of the increasing process and therefore, since

$$M_n = X_n - X_0 - A_n = X_n - X_0 - \widetilde{A}_n = \widetilde{M}_n,$$

also proves uniqueness of the martingale.

2 is obvious. □

Remark 5.16 (The angle bracket process $\langle M\rangle$)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ a filtration and $\{M_n\}_{n\in\mathbb{N}}$ a martingale with respect to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ with $\mathbb{E}[M_n^2] < \infty$ for each n. (Such a martingale is called an L^2-martingale.) Then by Proposition 5.6, $\{M_n^2\}_{n\in\mathbb{N}}$ is a* sub*martingale. Thus by Theorem 5.15 it has a Doob decomposition (which is essentially unique),*

$$M_n^2 = N_n + A_n$$

where $\{N_n\}_{n\in\mathbb{N}}$ is a martingale and $\{A_n\}_{n\in\mathbb{N}}$ is an increasing predictable process. The process $\{A_n\}_{n\in\mathbb{N}}$ is often denoted by $\{\langle M\rangle_n\}_{n\in\mathbb{N}}$.

Note that $\mathbb{E}[M_n^2] = \mathbb{E}[A_n]$ and that

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

It turns out that $\{\langle M\rangle_n\}_{n\in\mathbb{N}}$ is an extremely powerful tool with which to study $\{M_n\}_{n\in\mathbb{N}}$. It is beyond our scope here, but see for example Neveu 1975, Discrete Parameter Martingales.

6 Stopping Times and Stopping Theorems

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

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

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

Definition 6.1 (Stopping time)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ a filtration. A positive integer-valued (possibly infinite) random variable τ is called a* stopping time with respect to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ *if $\{\tau = n\} \in \mathcal{F}_n$ for all n or, equivalently, if $\{\tau > n\} \in \mathcal{F}_n$ for all n. Stopping times are sometimes called* optional times.

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

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

Proof

Take $v_n = \mathbf{1}_{\tau>n-1}$ in Example 5.13. $\square$

This Lemma tells us that $\mathbb{E}[M_{n\wedge\tau}] = \mathbb{E}[M_0]$, but can we let $n \to \infty$ to obtain $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$? The answer is 'no'.

Example 6.3. *Let $\{Y_i\}_{i\in\mathbb{N}}$ be i.i.d. random variables with $\mathbb{P}[Y_i = 1] = \mathbb{P}[Y_i = -1] = 1/2$. Set $X_n = \sum_{k=1}^{n} Y_k$, $n \geq 1$. That is X_n is the position of a simple random walk started from the origin after n steps. In particular, $\{X_n\}_{n\in\mathbb{N}}$ is a martingale and so $\mathbb{E}[X_n] = 0$ for all n.*

Now let $\tau = \min\{n : X_n = 1\}$. It is clear that τ is a stopping time and evidently $X_\tau = 1$. But then $\mathbb{E}[X_\tau] = 1 \neq 0 = \mathbb{E}[X_0]$.

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

Theorem 6.4 (Doob's Optional Stopping Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ a filtration, $\{M_n\}_{n\in\mathbb{N}}$ a martingale with respect to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ and τ a stopping time with respect to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$. Then M_τ is integrable and*

$$\mathbb{E}[M_\tau] = \mathbb{E}[M_0] \tag{16}$$

in each of the following situations:

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

$$\mathbb{E}\left[|M_{n+1} - M_n| \,|\, \mathcal{F}_n\right] < M, \qquad \textit{for all } n.$$

Proof

1. Because $\tau < \infty$, $\lim_{n\to\infty} M_{n\wedge\tau} = M_\tau$ *a.s.* and since $\{M_n\}_{n\in\mathbb{N}}$ is bounded we may apply the Bounded Convergence Theorem (that is the DCT with dominating function $g(\omega) \equiv K$) to deduce the result.

2. Take $n = N$ in Lemma 6.2.

3. Note first that setting $M_{0\wedge\tau} = 0$,

$$\begin{aligned} |M_{n\wedge\tau}| &= \left|\sum_{i=1}^{n}(M_{i\wedge\tau} - M_{(i-1)\wedge\tau})\right| \\ &\leq \sum_{i=1}^{n}|M_{i\wedge\tau} - M_{(i-1)\wedge\tau}| \\ &\leq \sum_{i=1}^{\infty}\mathbf{1}_{\tau\geq i}|M_{i\wedge\tau} - M_{(i-1)\wedge\tau}|. \end{aligned} \tag{17}$$

Now

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{\infty}\mathbf{1}_{\tau\geq i}|M_{i\wedge\tau} - M_{(i-1)\wedge\tau}|\right] &= \sum_{i=1}^{\infty}\mathbb{E}[\mathbf{1}_{\tau\geq i}|M_{i\wedge\tau} - M_{(i-1)\wedge\tau}|] \quad \text{(by MCT)} \\ &= \sum_{i=1}^{\infty}\mathbb{E}\left[\mathbb{E}[\mathbf{1}_{\tau\geq i}|M_{i\wedge\tau} - M_{(i-1)\wedge\tau}|\big|\,\mathcal{F}_{i-1}]\right] \quad \text{(tower property)} \\ &= \sum_{i=1}^{\infty}\mathbb{E}\left[\mathbf{1}_{\tau\geq i}\mathbb{E}[|M_{i\wedge\tau} - M_{(i-1)\wedge\tau}|\big|\,\mathcal{F}_{i-1}]\right] \quad \text{(since } \{\tau\geq i\}\in\mathcal{F}_{i-1}) \\ &\leq M\sum_{i=1}^{\infty}\mathbb{E}[\mathbf{1}_{\tau\geq i}] \\ &= M\sum_{i=1}^{\infty}\mathbb{P}[\tau\geq i] = M\mathbb{E}[\tau] < \infty. \end{aligned}$$

Moreover, $\tau < \infty$ *a.s.* and so $M_{n\wedge\tau} \to M_\tau$ *a.s.* as $n \to \infty$ and so by the DCT with the function on the right hand side of (17) as dominating function, we have the result. □

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

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

$$\mathbb{P}[\tau < n + N|\,\mathcal{F}_n] \geq \epsilon\, a.s.$$

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

The proof is an exercise.

Let's look at some applications of Theorem 6.4.

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

Solution

Consider a fair gambling casino - that is one in which the expected gain from each bet is zero. In particular, a gambler betting £a on the outcome of the next bet being a j will lose with probability $1-p_j$ and will win £a/p_j with probbaility p_j. (Her expected fortune after the game is then $0(1-p_j)+p_j a_j/p_j = a$, the same as before the game.)

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

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

Let M_n be the casino's winnings after n games (so when X_n has just been revealed). Then $\{M_n\}_{n\in\mathbb{N}}$ is a mean zero martingale, adapted to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ where $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. If we write τ for the number of random variables to be revealed before we see the required pattern, then by Lemma 6.5, $\mathbb{E}[\tau] < \infty$. Condition 3 of Theorem 6.4 is satisfied (for example take $M = 4/(p_0^2p_1^2p_2)$).

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

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

Thus

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

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

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

□

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

We stated the Optional Stopping Theorem for martingales, but similar results are available for *super*martingales - just replace the equality in (16) by an inequality. We also have the following useful analogue of Lemma 6.2.

Lemma 6.7. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ a filtration, $\{X_n\}_{n\in\mathbb{N}}$ a submartingale with respect to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ and τ a stopping time (finite or infinite) with respect to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$. Then $\{X_{n\wedge\tau}\}_{n\in\mathbb{N}}$ is also a submartingale with respect to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$.*

Proof

Let $\Lambda \in \mathcal{F}_n$.

$$\begin{aligned}\int_\Lambda \mathbb{E}[X_{(n+1)\wedge\tau} | \mathcal{F}_n] d\mathbb{P} &= \int_\Lambda X_{(n+1)\wedge\tau} d\mathbb{P} \\ &= \int_{\Lambda\cap\{\tau\leq n\}} X_{(n+1)\wedge\tau} d\mathbb{P} + \int_{\Lambda\cap\{\tau>n\}} X_{n+1} d\mathbb{P}.\end{aligned}$$

Now $\Lambda \cap \{\tau > n\} \in \mathcal{F}_n$ so by the submartingale inequality the right hand side is

$$\geq \int_{\Lambda\cap\{\tau\leq n\}} X_{n\wedge\tau} d\mathbb{P} + \int_{\Lambda\cap\{\tau>n\}} X_n d\mathbb{P} = \int_\Lambda X_{n\wedge\tau} d\mathbb{P}.$$

Thus

$$\int_\Lambda X_{n\wedge\tau} d\mathbb{P} \leq \int_\Lambda \mathbb{E}[X_{(n+1)\wedge\tau}|\mathcal{F}_n] d\mathbb{P}$$

for each $\Lambda \in \mathcal{F}_n$ and since $X_{n\wedge\tau}$ is $\mathcal{F}_n$-measurable this implies

$$X_{n\wedge\tau} \leq \mathbb{E}[X_{(n+1)\wedge\tau}|\mathcal{F}_n] \quad a.s.$$

□

Let's just record one more result for submartingales.

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

$$\mathbb{P}[|X| \geq \lambda] \leq \frac{\mathbb{E}[|X|]}{\lambda}.$$

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

Theorem 6.8 (A maximal inequality)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a positive submartingale (adapted to a filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$). Then for each fixed $N \in \mathbb{N}$,*

$$\mathbb{P}[\max_{n\leq N} X_n \geq \lambda] \leq \frac{\mathbb{E}[X_N]}{\lambda}.$$

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

$$\mathbb{P}\left[max_{n\leq N}|M_n| \geq \lambda\right] \leq \frac{\mathbb{E}[|M_N|]}{\lambda}.$$

Proof

Since $c(x) = |x|$ is convex, set $X_n = |M_n|$ and $\{X_n\}_{n\in\mathbb{N}}$ is a submartingale. □

Proof of Theorem 6.8

First let τ be a stopping time with $\mathbb{P}[\tau \leq N] = 1$. Then

$$\mathbb{E}[X_\tau] = \sum_{k=1}^N \mathbb{E}[X_\tau \mathbf{1}_{\{\tau=k\}}] = \sum_{k=1}^N \mathbb{E}[X_k \mathbf{1}_{\{\tau=k\}}].$$

Now

$$\begin{aligned}
\mathbb{E}[X_N \mathbf{1}_{\{\tau=k\}}] &= \mathbb{E}[\mathbb{E}[X_N \mathbf{1}_{\{\tau=k\}}|\mathcal{F}_k]] \\
&= \mathbb{E}[\mathbf{1}_{\{\tau=k\}}\mathbb{E}[X_N|\mathcal{F}_k]] \\
&\geq \mathbb{E}[\mathbf{1}_{\{\tau=k\}}X_k] \quad \text{since } \{X_n\}_{n\in\mathbb{N}} \text{ a submartingale}
\end{aligned}$$

and summing over k,

$$\mathbb{E}[X_N] = \sum_{k=1}^N \mathbb{E}[X_N \mathbf{1}_{\{\tau=k\}}] \geq \sum_{k=1}^N \mathbb{E}[X_k \mathbf{1}_{\{\tau=k\}}] = \mathbb{E}[X_\tau].$$

Now define τ by

$$\tau = \inf\{n \leq N : X_n \geq \lambda\}$$

if the set is non-empty and $\tau = N$ otherwise. Then $\{\max_{n\leq N} X_n \geq \lambda\} = \{X_\tau \geq \lambda\}$ and

$$\mathbb{P}[\max_{n\leq N} X_n \geq \lambda] = \mathbb{P}[X_\tau \geq \lambda] \leq \frac{1}{\lambda}\mathbb{E}[X_\tau] \leq \frac{1}{\lambda}\mathbb{E}[X_N]$$

as required. □

7 The Upcrossing Lemma and Martingale Convergence

Let $\{X_n\}_{n\in\mathbb{N}}$ be an integrable random process, for example modelling the value of the stockmarket. Consider the following strategy:

1. You do not invest until the value of X goes below some level a (representing what you consider to be a bottom price), in which case you buy a share.
2. You keep your share until X gets above some level b (a value you consider to be overpriced) in which case you sell your share and you return to the first step.

Three remarks:

1. However clever this strategy may seem, if $\{X_n\}_{n\in\mathbb{N}}$ is a supermartingale and you stop playing at some bounded stopping time then your losses will be greater than your winnings.
2. Your earnings are bounded above by $(b-a)$ times the number of times the process went up from a to b.
3. If you stop at a time n when the value is below the price at which you bought, then you make a loss which is bounded above by $(X_n - a)^-$.

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

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

Let $T_0(x) = 0$ and for $k \geq 0$ let

$$S_{k+1}(x) = \inf\{n \geq T_k(x) : x_n < a\},$$

$$T_{k+1}(x) = \inf\{n \geq S_{k+1}(x) : x_n > b\},$$

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

Let

$$N_n([a,b], x) = \sup\{k > 0 : T_k(x) \leq n\}$$

be the number of upcrossings of $[a,b]$ by x before time n and let

$$N([a,b], x) = \sup\{k > 0 : T_k(x) < \infty\}$$

be the total number of upcrossings of $[a,b]$ by x.

Lemma 7.2 (Doob's upcrossing lemma)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ a filtration and $\{X_n\}_{n\in\mathbb{N}}$ a supermartingale adapted to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$. Let $a < b$ be fixed real numbers. Then for every $n \geq 0$,*

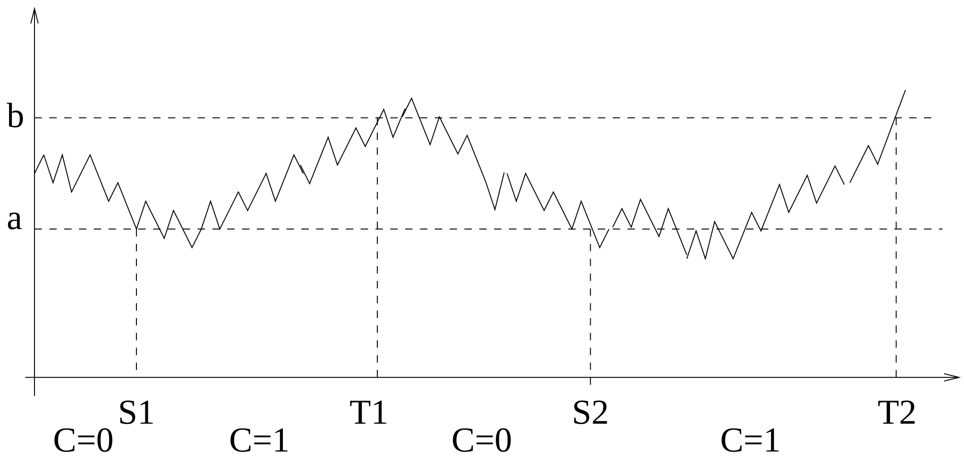
$$\mathbb{E}[N_n([a,b], X)] \leq \frac{\mathbb{E}[(X_n - a)^-]}{(b-a)}.$$

Proof

It is an easy induction to check that $S_k = S_k(X)$, $k \geq 1$, and $T_k = T_k(X)$, $k \geq 1$ are stopping times. Now set

$$C_n = \sum_{k\geq 1} \mathbf{1}_{\{S_k < n \leq T_k\}}.$$

Notice that C_n only takes the values 0 and 1. It is 1 at time n if X is in the process of making an upcrossing from a to b or if $T_K = \infty$ for some K and $n > S_K$.



Notice that

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

So $\{C_n\}_{n\in\mathbb{N}}$ is *predictable* (recall Definition 5.12). Now just as in Example 5.13 we construct the discrete stochastic integral

$$\begin{aligned}
(C \circ X)_n &= \sum_{k=1}^{n} C_k (X_k - X_{k-1}) \\
&= \sum_{i=1}^{N_n} (X_{T_i} - X_{S_i}) + (X_n - X_{S_{N_n+1}}) \mathbf{1}_{\{S_{N_n+1} \leq n\}} \\
&\geq (b-a) N_n + (X_n - a) \mathbf{1}_{\{X_n \leq a\}} \\
&\geq (b-a) N_n - (X_n - a)^-. \qquad (18)
\end{aligned}$$

Now since $\{C_n\}_{n\in\mathbb{N}}$ is bounded, non-negative and predictable and $\{X_n\}_{n\in\mathbb{N}}$ is a supermartingale we can deduce exactly as in Example 5.13 that $\{(C \circ X)_n\}_{n\in\mathbb{N}}$ is also a supermartingale. So finally, taking expectations in (18),

$$0 = \mathbb{E}[(C \circ X)_0] \geq \mathbb{E}[(C \circ X)_n] \geq (b-a)\mathbb{E}[N_n] - \mathbb{E}[(X_n - a)^-]$$

and rearranging gives the result. □

Now one way to show that a sequence of real numbers converges as $n \to \infty$ is to show that it doesn't oscillate too wildly and that can be expressed in terms of upcrossings. The following lemma makes this precise.

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

Proof

Because we need to be able to take a, b rational in our next theorem, we'll show that $\{x_n\}_{n\in\mathbb{N}}$ does *not* converge if and only if there exist rationals $a<b$ such that $N([a,b],x)=\infty$.

(i) If $N([a,b],x)=\infty$, then

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

and so $\{x_n\}_{n\in\mathbb{N}}$ does not converge.

(ii) If $\{x_n\}_{n\in\mathbb{N}}$ does not converge, then

$$\liminf_{n\to\infty} x_n < \limsup_{n\to\infty} x_n$$

and so choose two rationals $a<b$ in between. □

Now a supermartingale $\{X_n\}_{n\in\mathbb{N}}$ is just a random sequence and by Doob's Upcrossing Lemma we can bound the number of upcrossings of $[a,b]$ that it makes for any $a<b$ and so our hope is that we can combine this with Lemma 7.3 to show that the *random* sequence $\{X_n\}_{n\in\mathbb{N}}$ converges. This is our next result.

Theorem 7.4 (Doob's Forward Convergence Theorem)**.** *Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space and $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ a filtration. Suppose that $\{X_n\}_{n\in\mathbb{N}}$ is a supermartingale adapted to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ with the property that*

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

(we say that $\{X_n\}_{n\in\mathbb{N}}$ is bounded in L^1), then $\{X_n\}_{n\in\mathbb{N}}$ converges with probability one to an a.s. finite limit X_∞.

Proof

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

$$\mathbb{E}[N_n([a,b],X)] \le \frac{\mathbb{E}[(X_n-a)^-]}{(b-a)} \le \frac{\mathbb{E}[|X_n|+a]}{(b-a)}$$

and so by the MCT,

$$\mathbb{E}[N([a,b],X)] \le \frac{\sup_n \mathbb{E}[|X_n|+a]}{(b-a)}$$

which implies that $N([a,b],X)<\infty\,a.s.$. That is $\mathbb{P}[N([a,b],X)<\infty]=1$ for any $a<b\in\mathbb{Q}$. Hence

$$\mathbb{P}\Big[\bigcap_{a<b\in\mathbb{Q}}\{N([a,b],X)<\infty\}\Big]=1$$

(since a countable union of null sets is null). So $\{X_n\}_{n\in\mathbb{N}}$ converges a.s. to some X_∞. It remains to check that X_∞ is finite a.s. But Fatou's Lemma gives

$$\mathbb{E}[|X_\infty|] \le \liminf_{n\to\infty}\mathbb{E}[|X_n|]$$

and the right hand side is finite by assumption. Hence $|X_\infty|<\infty\,a.s.$. □

Corollary 7.5. *If $\{X_n\}_{n\in\mathbb{N}}$ is a non-negative supermartingale, then $X_\infty=\lim_{n\to\infty}X_n$ exists a.s.*

Proof

Since $\mathbb{E}[|X_n|] = \mathbb{E}[X_n] \leq \mathbb{E}[X_0]$ we may apply Theorem 7.4. $\square$

Example 7.6. *Recall our branching process of Definition 0.1. We defined* $Z_0 = 1$ *and*

$$Z_{n+1} = X_1^{(n+1)} + \cdots + X_{Z_n}^{(n+1)}$$

where the random variables $X_i^{(n+1)}$ *are i.i.d. all with law* $\mathbb{P}[X = k] = p_k$ *for suitable constants* p_k*. We also wrote* $\mu = \sum_{k=1}^{\infty} kp_k = \mathbb{E}[X]$ *and* $M_n = Z_n/\mu^n$*.*

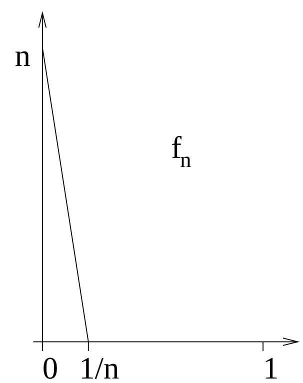
Let $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ *be the natural filtration. Then*

$$\begin{aligned}
\mathbb{E}[M_{n+1}|\mathcal{F}_n] &= \mathbb{E}\left[\left.\frac{Z_{n+1}}{\mu^{n+1}}\right|\mathcal{F}_n\right] \\
&= \mathbb{E}\left[\left.\frac{1}{\mu^{n+1}}(X_1^{(n+1)} + \cdots + X_{Z_n}^{(n+1)}\right|\mathcal{F}_n\right] \\
&= \frac{Z_n}{\mu^{n+1}}\mathbb{E}[X] = \frac{Z_n}{\mu^n} = M_n.
\end{aligned}$$

So $\{M_n\}_{n\in\mathbb{N}}$ *is a non-negative martingale and by Doob's Forwards Convergence Theorem we see that* $\{M_n\}_{n\in\mathbb{N}}$ *converges a.s. to a finite limit.* However, *as we saw in §0.2, if* $\mu \leq 1$ *then* $M_n \to 0$ *with probability one even though* $\mathbb{E}[M_0] = M_0 = 1$*. So we have convergence a.s. but not 'in* L^1*'. That is*

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

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



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

7.1 Martingales bounded in L^2

The assumption that $\sup_n \mathbb{E}[|X_n|] < \infty$ in Doob's Forwards Convergence Theorem is not always easy to check, so sometimes it is convenient to work with *uniformly square integrable martingales.*

Definition 7.7 (Martingales bounded in L^p). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ a filtration and $\{M_n\}_{n\in\mathbb{N}}$ a martingale adapted to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$. We say that $\{M_n\}_{n\in\mathbb{N}}$ is* bounded in L^p *(or for $p=2$ 'uniformly square integrable') if*

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

Suppose that $\{M_n\}_{n\in\mathbb{N}}$ is a uniformly square integrable martingale and that $k > j \geq 0$. We adopt the convention that $M_{-1} = 0$. Then

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

This allows us to obtain a 'Pythagoras rule':

$$\begin{aligned}
\mathbb{E}[M_n^2] &= \mathbb{E}\left[\left(\sum_{k=1}^n (M_k - M_{k-1})\right)^2\right] \\
&= \mathbb{E}[M_0^2] + \sum_{k=1}^n \mathbb{E}[(M_k - M_{k-1})^2] + 2\sum_{n\geq k>j\geq 1} \mathbb{E}[(M_k - M_{k-1})(M_j - M_{j-1})] \\
&= \mathbb{E}[M_0^2] + \sum_{k=1}^n \mathbb{E}[(M_k - M_{k-1})^2]. \qquad (19)
\end{aligned}$$

Theorem 7.8. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ a filtration and $\{M_n\}_{n\in\mathbb{N}}$ a martingale adapted to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$. The martingale $\{M_n\}_{n\in\mathbb{N}}$ is uniformly square integrable if and only if*

$$\sum_{k\geq 1} \mathbb{E}[(M_k - M_{k-1})^2] < \infty \qquad (20)$$

and in this case $M_n \to M_\infty$ a.s. and

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

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

Proof

From (19) it is clear that (20) is equivalent to uniform square integrability.

Now suppose (20) holds. From Jensen's inequality (since $c(x) = x^2$ is convex)

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

and so Doob's Forward Convergence Theorem shows that $M_\infty = \lim_{n\to\infty} M_n$ exists a.s. To check convergence in L^2 we use Pythagoras again,

$$\mathbb{E}[(M_{n+k} - M_n)^2] = \sum_{j=n+1}^{n+k} \mathbb{E}[(M_j - M_{j-1})^2], \qquad (21)$$

and so by Fatou's Lemma

$$\begin{aligned}
\mathbb{E}[(M_\infty - M_n)^2] &= \mathbb{E}[\liminf_{k\to\infty} (M_{n+k} - M_n)^2] \\
&\leq \liminf_{k\to\infty} \mathbb{E}[(M_{n+k} - M_n)^2] \\
&= \sum_{j\geq n+1} \mathbb{E}[(M_j - M_{j-1})^2] \quad \text{(using (21))}
\end{aligned}$$

and since

$$\sum_{j=1}^{\infty} \mathbb{E}[(M_j - M_{j-1})^2] < \infty$$

(which is (20)) the right hand side tends to zero as $n \to \infty$. That is

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

as required. □

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

8 Uniform Integrability

If X is an integrable random variable (that is $\mathbb{E}[|X|] < \infty$) and Λ_n is a sequence of sets with $\mathbb{P}[\Lambda_n] \to 0$, then $\mathbb{E}[|X\mathbf{1}_{\Lambda_n}|] \to 0$ as $n \to \infty$. (This is a consequence of the DCT since $|X|$ dominates $|X\mathbf{1}_{\Lambda_n}|$ and $|X\mathbf{1}_{\Lambda_n}| \to 0\, a.s.$) Uniform integrability demands that this type of property holds *uniformly* for random variables from some class.

Definition 8.1 (Uniform Integrability). *A class $\mathcal{C}$ of random variables is called* uniformly integrable *if given $\epsilon > 0$ there exists $K \in (0, \infty)$ such that*

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

There are two reasons why this definition is important:

1. Uniform integrability is necessary and sufficient for passing to the limit under an expectation,

2. it is often easy to verify in the context of martingale theory.

Property 1 should be sufficient to guarantee that uniform integrability is interesting, but in fact uniform integrability is not often used in analysis where it is usually simpler to use the MCT or the DCT. It is only taken seriously in probability and that is because of 2.

Proposition 8.2. *Suppose that $\{X_\alpha, \alpha \in I\}$ is a uniformly integrable family of random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then*

1.
$$\sup_\alpha \mathbb{E}[|X_\alpha|] < \infty,$$

2.
$$\mathbb{P}[|X_\alpha| > N] \to 0 \quad \textit{as } N \to \infty, \text{ uniformly } \textit{in } \alpha.$$

3.
$$\mathbb{E}[|X_\alpha|\mathbf{1}_\Lambda] \to 0 \quad \textit{as } \mathbb{P}[\Lambda] \to 0, \text{ uniformly } \textit{in } \alpha.$$

Conversely, either 1 and *3 or 2* and *3 implies uniform integrability.*

Proof

1 By definition of uniform integrability, there exists N_0 such that for all α

$$\mathbb{E}[|X_\alpha|\mathbf{1}_{\{|X_\alpha|>N_0\}}] \leq 1.$$

Then for all α

$$\mathbb{E}[|X_\alpha|] = \mathbb{E}\left[|X_\alpha|\mathbf{1}_{\{|X_\alpha|\leq N_0\}} + |X_\alpha|\mathbf{1}_{\{|X_\alpha|>N_0\}}\right] \leq N_0\mathbb{P}[|X_\alpha| \leq N_0] + 1 \leq N_0 + 1.$$

Now 1 implies 2 since

$$\begin{aligned}
\mathbb{P}[|X_\alpha| > N] &\leq \frac{1}{N}\mathbb{E}[|X_\alpha|] \quad \text{(Chebyshev)} \\
&\leq \frac{1}{N}\sup_\alpha \mathbb{E}[|X_\alpha|]
\end{aligned}$$

and the bound on the right, which evidently tends to zero as $N \to \infty$, is independent of α.

To see 3, fix $\epsilon > 0$ and choose N_ϵ such that

$$\mathbb{E}[|X_\alpha|\mathbf{1}_{\{|X_\alpha|>N_\epsilon\}}] < \frac{\epsilon}{2} \quad \text{for all } \alpha.$$

Then choose $\delta = \epsilon/(2N_\epsilon)$ and suppose $\mathbb{P}[\Lambda] < \delta$, then

$$\begin{aligned}
\mathbb{E}[|X_\alpha|\mathbf{1}_\Lambda] &= \mathbb{E}[|X_\alpha|\mathbf{1}_{\Lambda\cap\{|X_\alpha|\leq N_\epsilon\}} + |X_\alpha|\mathbf{1}_{\Lambda\cap\{|X_\alpha|>N_\epsilon\}}] \\
&\leq N_\epsilon\mathbb{E}[\mathbf{1}_\Lambda] + \mathbb{E}[|X_\alpha|\mathbf{1}_{\Lambda\cap\{|X_\alpha|>N_\epsilon\}}] \\
&\leq N_\epsilon\mathbb{P}[\Lambda] + \frac{\epsilon}{2} \\
&< \epsilon \quad \text{independent of } \alpha.
\end{aligned}$$

Thus, given $\epsilon > 0$, choosing δ in this way, $\mathbb{P}[\Lambda] < \delta$ implies $\mathbb{E}[|X_\alpha|\mathbf{1}_\Lambda] < \epsilon$ as required.

For the converse, since 1 implies 2, it is enough to check that 2 and 3 imply uniform integrability. Choose $\epsilon > 0$, by 3 there exists $\delta > 0$ such that $\mathbb{P}[\Lambda] < \delta$ implies $\mathbb{E}[|X_\alpha|\mathbf{1}_\Lambda] < \epsilon$ for all α. Then since by 2 there is N_0 such that $N \geq N_0$ implies $\mathbb{P}[|X_\alpha| > N] < \delta$ for all α, we have that for all $N \geq N_0$

$$\mathbb{E}\left[|X_\alpha|\mathbf{1}_{\{|X_\alpha|>N\}}\right] < \epsilon \quad \text{for all } \alpha.$$

□

We've already said that uniform integrability is a necessary and sufficient condition for going to the limit under the integral. Let's state this a bit more formally. Recall that for a family of random variables $\{X_n\}_{n\in\mathbb{N}}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ we say that $X_n \to X$ in L^1 if

$$\mathbb{E}[|X_n - X|] \to 0 \quad \text{as } n \to \infty.$$

We say that $X_n \to X$ in probability if given $\epsilon > 0$

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

It's easy to check that if $X_n \to X$ in L^1, then for any set Λ

$$\mathbb{E}\left[|X_n\mathbf{1}_\Lambda - X\mathbf{1}_\Lambda|\right] \to 0 \quad \text{as } n \to \infty.$$

Theorem 8.3 (Uniform Integrability and L^1 convergence)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of integrable random variables which converge in probability to a random variable X. TFAE*

1. $\{X_n\}_{n\in\mathbb{N}}$ *is uniformly integrable,*

2. $\mathbb{E}[|X_n - X|] \to 0$ *as* $n \to \infty$,

3. $\mathbb{E}[|X_n|] \to \mathbb{E}[|X|]$ *as* $n \to \infty$.

Proof

Since $|X_n| \to |X|$ in probability, by Theorem 3.14 there exists a subsequence $\{n_k\}_{k\in\mathbb{N}}$ such that $\{X_{n_k}\}_{k\in\mathbb{N}}$ converges a.s.. Fatou's Lemma gives

$$\mathbb{E}[|X|] \le \liminf_{n\to\infty} \mathbb{E}[|X_{n_k}|]$$

and by Proposition 8.2 1 the right hand side is bounded. Thus X is integrable. Now let $\epsilon > 0$.

$$\begin{aligned}\mathbb{E}[|X - X_n|] &= \mathbb{E}[|X - X_n|\mathbf{1}_{\{|X-X_n|\le\epsilon/3\}} + |X - X_n|\mathbf{1}_{\{|X-X_n|>\epsilon/3\}}] \\ &\le \frac{\epsilon}{3} + \mathbb{E}\left[|X|\mathbf{1}_{\{|X-X_n|>\epsilon/3\}} + |X_n|\mathbf{1}_{\{|X-X_n|>\epsilon/3\}}\right].\end{aligned}$$

Now since $X_n \to X$ in probability we have $\mathbb{P}[|X - X_n| > \epsilon/3] \to 0$ as $n \to \infty$ and so using the integrability of X and the DCT gives

$$\mathbb{E}[|X|\mathbf{1}_{\{|X-X_n|>\epsilon/3\}}] \to 0 \quad \text{as } n \to \infty.$$

Since by Proposition 8.2 3, uniform integrability implies

$$\mathbb{E}[|X_n|\mathbf{1}_{\{|X-X_n|>\epsilon/3\}}] \to 0 \quad \text{as } n \to \infty,$$

we have

$$\mathbb{E}[|X_n - X|] \to 0 \quad \text{as } n \to \infty$$

as required.

2 implies 3 is obvious, so let's prove that 3 implies 1.

Suppose then thet $\mathbb{E}[|X_n|] \to \mathbb{E}[|X|]$ as $n \to \infty$. For any real number M such that $\mathbb{P}[X = M] = 0$, we define

$$X^{(M)} = \begin{cases} X & |X| \le M, \\ 0 & |X| > M, \end{cases}$$

with a parallel definition for $X_n^{(M)}$.

Then for any such M,

$$\begin{aligned}\mathbb{E}[|X_n|\mathbf{1}_{\{|X_n|>M\}}] &= \mathbb{E}[|X_n|] - \mathbb{E}[|X_n^{(M)}|] \\ &\to \mathbb{E}[|X|] - \mathbb{E}[|X^{(M)}|]\end{aligned}$$

as $n \to \infty$ since $\mathbb{E}[|X_n|] \to \mathbb{E}[|X|]$ by hypothesis and using the bounded convergence theorem.

Now let $\epsilon > 0$. Choose M_ϵ large enough that

$$\mathbb{E}[|X|] - \mathbb{E}[|X^{(M_\epsilon)}|] < \frac{\epsilon}{3}$$

(and $\mathbb{P}[|X| = M_\epsilon] = 0$ as above). There exists N_0 such that for $n \ge N_0$ we have both

$$|\mathbb{E}[|X_n|] - \mathbb{E}[|X|]| < \frac{\epsilon}{3}$$

and

$$\left|\mathbb{E}[|X_n^{(M_\epsilon)}|] - \mathbb{E}[|X^{(M_\epsilon)}|]\right| < \frac{\epsilon}{3}.$$

Then for all $n \geq N_0$,

$$\begin{aligned}
\mathbb{E}[|X_n|\mathbf{1}_{\{|X_n|>M_\epsilon\}}] &= \mathbb{E}[|X_n|] - \mathbb{E}[|X_n^{(M_\epsilon)}|] \\
&\leq \left|\mathbb{E}[|X|] - \mathbb{E}[|X^{(M_\epsilon)}|]\right| + |\mathbb{E}[|X_n|] - \mathbb{E}[|X|]| + \left|\mathbb{E}[|X_n^{(M_\epsilon)}|] - \mathbb{E}[|X^{(M_\epsilon)}|]\right| \\
&< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon
\end{aligned}$$

There are only finitely many $n < N_0$, so there exists $M'_\epsilon \geq M_\epsilon$ such that

$$\mathbb{E}\left[|X_n|\mathbf{1}_{\{|X_n|>M'_\epsilon\}}\right] < \epsilon$$

for *all* n, which is 1. □

The second reason that uniform integrability is important is that it is readily verified for martingales and submartingales.

Theorem 8.4. *Let $\{\mathcal{F}_\alpha\}_{\alpha\in I}$ be a family of sub σ-fields of $\mathcal{F}$ and let X be an integrable random variable. Define*

$$X_\alpha = \mathbb{E}[X|\mathcal{F}_\alpha].$$

Then $\{X_\alpha, \alpha \in I\}$ is uniformly integrable.

Proof

Let $Y = \mathbb{E}[X|\mathcal{F}_\alpha]$. By the conditional Jensen inequality (Proposition 4.12), since $c(x) = |x|$ is convex,

$$|Y| = |\mathbb{E}[X|\mathcal{F}_\alpha]| \leq \mathbb{E}\left[|X|\,|\,\mathcal{F}_\alpha\right] \; a.s.$$

and so

$$\mathbb{E}[|Y|\mathbf{1}_{\{|Y|\geq K\}}] \leq \mathbb{E}\left[\mathbb{E}[|X|\,|\,\mathcal{F}_\alpha]\mathbf{1}_{\{|Y|\geq K\}}\right] = \mathbb{E}[|X|\mathbf{1}_{\{|Y|\geq K\}}]. \tag{22}$$

Now the single integrable random variable X forms on its own a uniformly integrable class and so by Proposition 8.2 given $\epsilon > 0$ we can find $\delta > 0$ such that $\mathbb{P}[\Lambda] < \delta$ implies $\mathbb{E}[|X|\mathbf{1}_\Lambda] < \epsilon$. So to show that (22) is less than ϵ for sufficiently large K we must check that $\mathbb{P}[|Y| \geq K]$ converges to zero (uniformly in α) as $K \to \infty$. But

$$\begin{aligned}
\mathbb{P}[|Y| \geq K] &\leq \frac{\mathbb{E}[|Y|]}{K} \quad \text{(Chebyshev)} \\
&= \frac{\mathbb{E}[|X|]}{K} \quad \text{(tower property)}
\end{aligned}$$

so for any α, taking $K > \mathbb{E}[|X|]/\delta$ implies $\mathbb{P}[|Y| \geq K] < \delta$ and hence the right hand side of (22) is less than ϵ as required. □

Theorem 8.5. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ a filtration and $\{M_n\}_{n\in\mathbb{N}}$ a martingale adapted to $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$. TFAE*

1. *$\{M_n\}_{n\in\mathbb{N}}$ is uniformly integrable,*
2. *M_n converges a.s. and in $L^1(\Omega, \mathcal{F}, \mathbb{P})$ to a limit M_∞,*
3. *there exists $Z \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ such that $M_n = \mathbb{E}[Z|\mathcal{F}_n]$, $n \geq 0$*

Proof

We already have most of the ingredients.

1 $\implies$ 2: If $\{M_n\}_{n\in\mathbb{N}}$ is uniformly integrable then it is bounded in L^1 and so by Doob's Forward Convergence Theorem it converges a.s.. Theorem 8.3 and uniform integrability ensure convergence in L^1 to a limit which we denote M_∞.

2 $\implies$ 3: Choose $Z = M_\infty$. For all $m, n \in \mathbb{N}$ with $m \geq n$ and all $F \in \mathcal{F}_n$

$$\mathbb{E}[X_m \mathbf{1}_F] = \mathbb{E}[X_n \mathbf{1}_F]$$

(combining $X_n = \mathbb{E}[X_m|\mathcal{F}_n]\,a.s.$ and the tower property) and letting $m \to \infty$, the L^1 convergence gives

$$\mathbb{E}[X_\infty \mathbf{1}_F] = \mathbb{E}[X_n \mathbf{1}_F] \quad \text{for all } F \in \mathcal{F}_n$$

which gives $X_n = \mathbb{E}[X_\infty|\mathcal{F}_n]\,a.s.$ by definition of conditional expectation.

3 $\implies$ 1 follows directly from Theorem 8.4. $\square$

9 Backwards Martingales and the Strong Law of Large Numbers

Backwards martingales are martingales for which time is indexed by the negative integers $\mathbb{Z}$.

Given sub σ-algebras

$$\mathcal{G}_{-\infty} = \bigcap_{k\in\mathbb{N}} \mathcal{G}_{-k} \subseteq \cdots \subseteq \mathcal{G}_{-(n+1)} \subseteq \mathcal{G}_{-n} \subseteq \cdots \subseteq \mathcal{G}_0$$

an integrable $\{\mathcal{G}_{-n}\}_{n\in\mathbb{N}}$-adapted process $\{M_{-n}\}_{n\in\mathbb{N}}$ is a backwards martingale if

$$\mathbb{E}[M_{-n+1}|\mathcal{G}_{-n}] = M_{-n} \quad a.s.$$

Backwards martingales are automatically uniformly integrable by Theorem 8.4 since M_0 is integrable and

$$M_{-n} = \mathbb{E}[M_0|\mathcal{G}_{-n}].$$

We can easily adapt Doob's Upcrossing Lemma to prove that if $N_m([a,b], M)$ is the number of upcrossings of $[a,b]$ by a backwards martingale between times $-m$ and 0, then

$$(b-a)\mathbb{E}[N_m([a,b],M)] \leq \mathbb{E}[(M_0 - a)^-]. \tag{23}$$

To see this consider the *forwards* martingale $\{M_{-m+k}\}$ for $0 \leq k \leq m$. As $m \to \infty$, $N_m([a,b],M)$ converges a.s. to the total number of upcrossings of $[a,b]$ by $\{M_{-n}\}_{n\in\mathbb{N}}$ and since the bound in (23) is uniform in m, we conclude, exactly as in Doob's Forward Convergence Theorem, that M_{-n} converges a.s. to a $\mathcal{G}_{-\infty}$-measurable random variable $M_{-\infty}$ as $-n \to -\infty$. As remarked above, our backwards martingale is automatically uniformly integrable and so we have:

Theorem 9.1. *Let $\{M_{-n}\}_{n\in\mathbb{N}}$ be a backwards martingale. Then M_{-n} converges a.s. and in L^1 as $-n \to -\infty$ to the random variable $M_{-\infty} = \mathbb{E}[M_0|\mathcal{G}_{-\infty}]$.*

We now use this result to prove the celebrated Kolmogorov Strong Law.

Theorem 9.2 (Kolmogorov's SLLN)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables with $\mathbb{E}[|X_k|] < \infty$ for each k. Write $\mu = \mathbb{E}[X_k]$ and define*

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

Then
$$\frac{S_n}{n} \to \mu \, a.s. \text{ and in } L^1 \text{ as } n \to \infty.$$

Proof

Define
$$\mathcal{G}_{-n} = \sigma(S_n, S_{n+1}, S_{n+2}, \ldots)$$
and
$$\mathcal{G}_{-\infty} = \cap_{k \in \mathbb{N}} \mathcal{G}_{-k}.$$
Note that $\mathcal{G}_{-n} = \sigma(S_n, X_{n+1}, X_{n+2}, \ldots)$ and that by independence of the X_i's, $\mathcal{G}_{-(n+1)}$ is independent of the sub σ-algebra $\sigma(X_1, S_n)$ (which is a sub σ-algebra of $\sigma(X_1, X_2, \ldots, X_n)$). Thus
$$\mathbb{E}[X_1 | \mathcal{G}_{-n}] = \mathbb{E}[X_1 | \sigma(S_n, \sigma(X_{n+1}, X_{n+2}, \ldots))] = \mathbb{E}[X_1 | S_n].$$
But by symmetry (recall that $\{X_n\}_{n \in \mathbb{N}}$ is an i.i.d. sequence)
$$\mathbb{E}[X_1 | S_n] = \mathbb{E}[X_2 | S_n] = \cdots = \mathbb{E}[X_n | S_n]$$
and
$$\frac{1}{n} \mathbb{E}[X_1 + \cdots + X_n | S_n] = \frac{1}{n} S_n.$$
Therefore, for all $n \in \mathbb{N}$
$$\mathbb{E}[X_1 | \mathcal{G}_{-n}] = \cdots = \mathbb{E}[X_n | \mathcal{G}_{-n}] = \frac{S_n}{n}$$
so that
$$\begin{aligned} \mathbb{E}[S_n | \mathcal{G}_{-(n+1)}] &= \mathbb{E}[S_{n+1} | \mathcal{G}_{-(n+1)}] - \mathbb{E}[X_{n+1} | \mathcal{G}_{-(n+1)}] \\ &= S_{n+1} - \frac{S_{n+1}}{n+1} = \frac{n}{n+1} S_{n+1}. \end{aligned}$$
Hence, setting $M_{-n} = S_n/n$, $\{M_{-n}\}_{n \in \mathbb{N}}$ is a backwards martingale with respect to its natural filtration. Thus S_n/n converges a.s. and in L^1 to a limit which is a.s. constant by Kolmogorov's 0-1 law and so it must equal its mean value,
$$\begin{aligned} \mathbb{E}[\lim_{-n \to -\infty} \mathbb{E}[X_1 | \mathcal{G}_{-n}]] &= \lim_{-n \to -\infty} \mathbb{E}[\mathbb{E}[X_1 | \mathcal{G}_{-n}]] \\ &= \lim_{n \to \infty} \mathbb{E}[\frac{S_n}{n}] = \mu. \end{aligned}$$
□

Note that the result is not completely obvious. For example, if the X_i are normally distributed with mean zero and variance 1, S_n/n converges to zero a.s., but $S_n/\sqrt{n}$ does not converge and indeed $\limsup S_n/\sqrt{2n \log \log n} = 1$.

Index

L^1-bounded, 42
L^2-martingale, 35, 43
π-system, 5
σ-algebra, 4
  Borel, 5
  generated by a collection of sets, 5
  generated by a random variable, 14
  independence of collection, 13
  product, 11
  tail, 18

absolutely continuous, 26
adapted process, 30
algebra, 4
almost everywhere, 9
almost sure convergence, 20
almost surely, 9
angle bracket process, 35

backwards martingale, 49
Bayes' rule, 23
BC1, 16
BC2, 17
Borel σ-algebra, 5
Borel measure, 5
Borel-Cantelli Lemma, 16
branching process, 2, 43

Chebyshev's inequality, 22
compensation, 34
conditional convergence theorems, 28
conditional expectation, 2, 23
  a.s. uniqueness, 25
  and mean square approximation, 30
  defining relation, 24
  existence, 26
  taking out what is known, 28
  tower property, 28
conditional Jensen's inequality, 29
consistent family, 15
convergence
  almost surely, 20, 21
  in L^r, 20
  in distribution, 20
  in probability, 20, 21
convex function, 22
cylinder sets, 15

defining relation (conditional expectation), 24
discrete measure theory, 5
discrete stochastic integral, 33
distribution function, 8, 10
Dominated Convergence Theorem, 10
Doob's forward convergence theorem, 42

extinction probability, 2

Fatou's Lemma, 10
  reverse, 10
filtration, 30
  natural, 31
Fubini's Theorem, 12

Galton-Watson branching process, 2, 43

image measure, 11
independent
  σ-algebras, 13
  events, 13
  random variables, 14
integrable function, 9

Jensen's inequality, 22
  conditional version, 29

Kolmogorov 0-1 Law, 19
Kolmogorov's Consistency Theorem, 15

law of the iterated logarithm, 19
Lebesgue-Stieltjes measure, 8
liminf, 9
  sets, 15
limsup, 9
  sets, 15

martingale, 4, 31
  backwards, 49
martingale convergence theorem, 42
martingale difference, 33
martingale transform, 33
maximal inequality, 39
measurable function, 9
measure space, 4

monotone convergence properties, 6
modes of convergence, 20
monotone class, 13
Monotone Class Theorem, 13
monotone convergence properties, 6
Monotone Convergence Theorem, 10

natural filtration, 31

Optional Stopping Theorem, 36
optional time, 36

partition, 24
Partition Theorem, 24, 29
predictable process, 33
product σ-algebra, 11
product measure, 12

Radon measure, 5
Radon-Nikodym derivative, 26
Radon-Nikodym Theorem, 26
random variable, 10
equivalence class, 23
independent, 14
reverse Fatou Lemma, 10

Scheffé's Lemma, 10
set function, 6
simple function, 9
canonical form, 9
standard machine, 12
stopped martingale, 34
stopping time, 36
Strong law of large numbers, 49
submartingale, 31
supermartingale, 31

tail σ-algebra, 18
taking out what is known, 28
tower property, 2, 28

uniform integrability, 45
and L^1 convergence, 46
uniqueness of extension, 5
upcrossing, 40
upcrossing lemma, 41