ST318
Probability Theory

Paul Chleboun
Last updated February 15, 2023
Department of Statistics
University of Warwick

ST318
Probability Theory

Equivalent of 30 lectures Term 2 2023

Prerequisites: ST342 Mathematics of Random Events or MA359 Measure Theory.

Commitment: 3 lectures of 1 hour per week and one tutorial every fortnight (starting in Week 2). This module runs in Term 2. In 2021/22 this was lectured in hybrid format, for more details of the module structure and all the accompanying material see the ST318 Moodle page. If you would like to request previous recordings please email me.

Content

- Independence and zero-one laws
- Modes of convergence for sequences of random variables
- Limit theorems: Law of Large Numbers (LLN) and Central Limit Theorems (CLT)
- Conditioning and discrete-time martingales

Aims

To give the student a rigorous presentation of some fundamental results in measure theoretic probability and an introduction to the theory of discrete time martingales. In so doing it aims to provide a firm basis for advanced work on probability and its applications.

Objectives

At the end of the course the student will: Understand the ideas relating to independence and zero-one laws and be able to apply these ideas in simple contexts. Understand the different modes of convergence for sequences of random variables and the relationship between these different modes. Be able to state and prove the Central Limit Theorem via the method of characteristic functions and understand how this result can be applied. Understand some basic results on discrete-time martingales including the martingale convergence theorem and optional stopping theorem, and show how these results can be used to obtain various characteristics of simple random walks.

Assessment: 100% by 2-hour Summer examination.

Course Material: Lecture notes, example sheets, and other module material are to be found at the Module Resources page (Moodle page).

Reading: The subject of this course, and the topics we choose to cover, are rather classic in the sense that there are *many* good books that will cover the same material - often with sections with very similar names. I strongly encourage students to read around and find a presentation they like. Much of this material is available in the library and online.

- D. Williams, *Probability with Martingales*, Cambridge Mathematical Textbooks (1991)
- A. Klenke, *Probability Theory*, Universitext Springer (2008)
- R. Durrett, *Probability Theory and Examples,* Cambridge Series in Statistical and Probabilistic Mathematics (2010)
- O. Kallenberg, *Foundations of Modern Probability*, Springer (2002)

Contents

Chapter 1

Introduction and Motivation

Reading: ST342 Mathematics of Random Events or MA359 Measure Theory.
Cohn, D.L, Measure Theory, Second Edition, Birkhauser (2013) [E-book in Library]
Further reading: D. Williams, Chapter 0

These will remain a work in progress, first started January 2020, they were largely written during Term 2 2019/2020. They were reshaped for "blended learning" in Term 2 of 2020/2021, and are under further development since Term 2 of 2021/2022. Some of the material is based on 2014 lecture notes by Wilfrid Kendall, with many thanks! Thanks also to Nikolaos Constantinou for contributions to the notes in 2020/21 and Shiva Mahesh in 2021/22.

Directed reading will always be included in these notes where they are otherwise incomplete and for further reading, which is strongly encouraged.

Comments and corrections are more than welcome, in fact I am more than happy for people to contribute at least to correcting typos, if not adding details, as we go (the notes are currently on Overleaf). Please feel free to contact me at paul.i.chleboun@warwick.ac.uk if you want to contribute.

The first two chapters, marked by $*$, should largely be revision of prerequisite material and will therefore be covered relatively quickly. If the pace of these first few chapters feels too fast it is definitely worth revising the prerequisite material. More details and links to your previous modules will be included in the notes. Please note that your relevant background may differ significantly from other students on the module (for example based on pre-requisite of ST342 or MA359). This means that you will have to choose *your own emphasis* for these initial chapters.

Lecture One Intended Learning Outcomes:

- Be familiar with the main concepts/names that will appear in the module on a conceptual level.
- Understand some of the reasons why we would like to develop a rigorous approach to probability based on measure theory.

1.1 Background

Probability is a relatively young but now central aspect of mathematics. 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. Probability theory arose from the need to quantify and understand uncertainty. Historically development was driven by very varied scenarios, such as gamblers who needed to grasp the "odds" of certain outcomes with given (limited) information, and experimental scientists who wanted to build detailed models based on imperfect (noisy) observations of the universe. At the start of the twentieth century, developments in quantum mechanics suggested that uncertainty was a fundamental property of the universe we live in, not just a reflection of human ignorance. Furthermore, developments in the understanding of the behaviour of non-linear dynamical systems showed that even deterministic systems could behave in very "random" ways. Probability theory is now an indispensable tool across the physical and social sciences, from physics to neuroscience, from genetics to 'big data' and, of course, in mathematical finance.

In short, the aim of this course is introduce some of the key tools and concepts which are essential to the understanding of modern probability theory, and would be reasonable assumed knowledge of a mathematician who has studied some probability by the end of their undergraduate program.

1.2 A Sneak Peek

We will start by explaining the terms used in the module objectives and giving a little more motivation. During this course the will make the concepts discussed below rigorous.

Independence: One of the central ideas in probability theory is *independence*. Intuitively, two events are independent if they have no influence on each other. The notation of independence of finitely many events, or random variables (such as throwing two dice), has been covered in previous modules.

What about describing independence of a large, possibly infinite, number of random variables (such as 3 Cartesian coordinates locating a random object in $\mathbb{R}^3$)? Or even independence or lack of independence of more exotic objects, such as rotations in $\mathbb{R}^3$, or stochastic processes.

It turns out that independence and the lack of independence can raise subtle issues. This is one of the main concepts by which Probability Theory supplements Measure Theory and adds extra structure.

Exercise 1.1. 100 prisoners problems. The director of a prison offers 100 death row prisoners, who are numbered from 1 to 100, a last chance. A room contains a cupboard with 100 drawers. The director randomly puts one prisoner's number in each closed drawer. The prisoners enter the room, one after another. Each prisoner may open and look into 50 drawers in any order. The drawers are closed again afterwards. If, during this search, every prisoner finds his number in one of the drawers, all prisoners are pardoned. If even one prisoner does not find his number, all prisoners are executed. Before the first prisoner enters the room, the prisoners may discuss strategy — but may not communicate once the first prisoner enters to look in the drawers. Can you find a strategy that has a better than 30% chance of success? If the prisoners behave *independently* the chances of survival do not look good!

Zero-one laws: Roughly speaking, any event that depends on a sequence of random variables $(X_n)_{n\geq 1}$, but changing finitely many of these random values does not affect whether the event occurs or not, is called a *tail* event. It turns out that this class of events still includes many interesting things, for example many events involving limits have the

required properties. Sometimes probability questions can have surprisingly simple answers, and the family of tail events fall into this class. Kolmogorov's zero-one law states that, if they are independent, these tail events are trivial in the sense that they have probability either 0 or 1. There are more advanced examples of 0-1-laws of this type, we will come across some of these when studying martingales.

Example 1.2. Take $X_1, X_2, \ldots$, independent and identically distributed *Bernoulli* random variables (values 0 or 1), with $\mathbb{P}[X_i = 1] = 1/2$. Suppose I win $£a_i$ when $X_i = 1$, and lose $£a_i$ when $X_i = 0$.

What can I say about the long-run profit/loss $Y_n = a_1(2X_1 - 1) + \ldots + a_n(2X_n - 1)$ for large n? It turns out that *either* Y_n converges to a limit with probability one, *or* it does not with probability one (no half-way houses!). Note, the distribution of the random variable $Z_i = 2X_i - 1$, which takes values -1 or 1, is sometimes called the *Rademacher distribution* and obviously extends to other parameter values analogously to the Bernoulli distribution.

Actually one can show that convergence happens if and only if $\sum a_i^2 < \infty$. Consider $a_i = 1/i$, then Y_n still converges with probability 1 even though $\sum_i a_i$ diverges.

Modes of convergence: Let $X_1, X_2, \ldots$, be independent random variables with common mean μ and finite variance σ^2. Then the Weak Law of Large numbers states that the sample mean *converges in probability:*

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad \text{in probability as } n \to \infty.$$

In fact, a stronger statement holds, the sequence (Y_n) converges to μ with probability 1. This type of convergence is called *almost sure convergence*, and analogue of the statement above is called the Strong Law of Large Numbers.

Convergence can be thought of in terms of approximation, that is eventually some penalty (or metric) becomes small. Convergence in probability fails to penalize very large differences that occur with small probability, which is one reason it is often useful to consider *convergence in* L^p. $Y_n \to Z$ in L^p if $\mathbb{E}[|Y_n - Z|^p] \to 0$. This idea of convergence is also fundamental in functional analysis, and is therefore used frequently.

In some situations we just want the probabilities of events to be close, such as in the Central Limit Theorem. In this situation we often need the concept of *weak convergence*.

Different modes of convergence can tell very different things about the limiting behaviour of sequences of random variables. For example, comparing the Strong Law of Large Numbers with the law of iterated logarithm and the Central Limit Theorem, each of which tells us different things about the sum of i.i.d. random variables.

Conditional expectation: Probability can be thought of as quantifying ignorance. In this, new information decreases ignorance and we can update our probabilities. That is, if there is partial information on the outcome of a random experiment, then the probabilities of events may change.

Example 1.3. Given a sequence $X_1, X_2, \ldots, X_n$ (for example, cumulative winnings/losses at a casino), how can we predict the value of X_{n+1}? If the we choose to minimise the expected penalty $|M_{n+1} - X_{n+1}|^2$ (i.e. L^2 distance), then the "best" prediction is what is called the conditional expectation

$$M_{n+1} = \mathbb{E}[X_{n+1} \mid X_1, \ldots, X_n]\,.$$

Notice that this prediction is still random, since it depends on the outcomes $X_1, X_2, \ldots, X_n$. Conventionally in order to complete this procedure I have to use different methods depending on whether I work with discrete or continuous random variables. Can these concepts be unified? The answer is of course yes, and in making this rigorous we will introduce a concept of conditional expectation (and hence conditional probability) which generalises all the definitions you will have come across before into one very elegant package.

We need a concept of conditional expectation that allows us to condition on random variables (i.e. for the expectation to remain 'random'), and in some cases the probability these take specific values may be zero (for example when we consider continuous random variables we must be able to cope with condition on events of probability zero), all in a way that is consistent with our intuition. We get around this by using σ-algebras to represent the 'information' given by a random variable. It turns out that in this way there is an extremely general and powerful concept of conditional expectation (which at first sight may appear very abstract).

Martingales: These are one of the most important concepts in modern probability theory, simply put they formalise the notion of a 'fair game'; the expected value of the random variable (for example winnings in a game) on the $(n+1)^{\text{th}}$ go, given I know what has happened up to the n^{th} go, in a fair game, is simply the value at the n^{th} go (on average you would not expect to win or lose in one step). They crop up in very many contexts and have revolutionised the approach to many areas of probability.

Examples include the position of a simple symmetric random walk, or more generally the sum of independent mean zero random variables. Another important example of a martingale is the successive conditional probabilities of a future event as more and more information is released. In words (it will be worth revisiting this example later to clarify so don't be afraid if this sounds a little odd or even tautological), the average probability an event occurs given all the information up to time n, if we average with only the information up to time $n-1$, will be the same as the average probability the event occurs given all the information up to time $n-1$.

Although we will focus here on the case where the random variables are indexed by a countable set, the concept of martingales extends to arbitrary ordered sets, such as $(X_t)_{t\in\mathbb{R}_+}$. We stick to the discrete setting to reduce the amount of technicalities, but this is a very rich topic and central to many areas of probability, not least financial mathematics.

1.3 The Galton-Watson branching process

By way of further motivation we will we take a quick look at a prototypical branching process. Such processes frequently arise in applications (such as biology) and are in general a very rich source of theory. This particular example is sufficiently rich to demonstrate some of the power of the theory we will develop in this module. I would suggest you watch the video before reading this chapter in detail. For more information see D. Williams, Prob. with Martingales Chapter 0.

The Galton-Watson branching process was popularised by Galton and Watson in the 1870's, although it had been studied earlier by Bienaymé in the 1840's. Galton and Watson were preoccupied by the demise of English aristocratic family names and wanted to know "What is the probability that a family name dies out by the 'ordinary law of chances'?".

It has now been widely applied as a simple model of evolution of population size and it has proved to be very powerful.

Suppose that $(X_r^{(n)})_{n,r\geq 1}$ is a (doubly) infinite array of independent identically distributed random variables, each with the same distribution as some random variable X, where

$$\mathbb{P}(X=k)=p_k\,,\quad k=0,1,2,\ldots$$

We will interpret $X_r^{(n)}$ as the number of off-spring of the r^{th} individual in the $(n-1)^{th}$ generation of some population who will be born into the n^{th} generation. Then the total population size Z_n at generation n is given by

1. $Z_0=1$,
2. $Z_n=X_1^{(n)}+X_2^{(n)}+\ldots+X_{Z_{n-1}}^{(n)}$ for $n\geq 1$, where we interpret the sum as zero in the case that $Z_{n-1}=0$.

The process $(Z_n)_{n\geq 1}$ defines a *Galton-Watson branching process* (started from a single ancestor) with offspring distribution X. In the application considered by Galton and Watson the random variable Z_n models the number of male descendants of a single male ancestor after n generations.

Remark 1.4. Notice that we have already glossed over one of the first challenges. That is, we need to be able to define what it means to have a countable collection of i.i.d. random variables. Suppose that X can take value 0 or 1 then there is a surjection from the set of all sequences $(x_r^{(1)})_{r\in\mathbb{N}}$ to the unit interval $[0,1]$ (think of the dyadic expansion), so indeed our sample space (of all possible outcomes) must be uncountable. If one assumes the Axiom of Choice (which we will), then one can prove that it is impossible to assign to all subsets of the sample space a probability that still satisfies all the obvious requirements that we would want. So we need to know which 'events' we can assign probabilities too. More of this later in Chapter 2, also for more details see Chapter 0.5 in D. Williams *Probability with Martingales* and Chapter 1.1 in P. Billingsley *Probability and Measure*.

We will assume that

$$\mathbb{P}(X=0)=p_0>0\,,$$

and that the expectation of the offspring distribution $\mu=\mathbb{E}[X]=\sum_{k=0}^{\infty}kp_k$ is finite. In the analysis a key role will be played by the *probability generating function* $G\colon[0,1]\to[0,1]$ of X given by $G(s)=\sum_{k=0}^{\infty}p_ks^k$.

The first main observation allows us to calculate the probability generating function of Z_n in from that of X. You will probably have come across this result in some form before for the sum of i.i.d. random variables.

Claim 1.5. *Let $G_n(s)=\mathbb{E}[s^{Z_n}]$, then G_n is given by the n-fold composition of G with itself, i.e.*

$$G_n(s)=\mathbb{E}[s^{Z_n}]=G_{n-1}(G(s))=\underbrace{G(G(\ldots G}_{n\text{-times}}(s)\ldots))=G(G_{n-1}(s))\,.$$

"Proof". We will use - for now just intuitively - the idea of conditional expectation as a random variable. This will be made precise when we treat the rigorous definition of the

conditional expectation later in the module. Furthermore, we will use the *Tower Property of Conditional Expectation* to calculate the expected value of some random variable U by first fixing some other random variable V (to 'reduce the randomness') and averaging over everything else and then subsequently averaging over V;

$$\mathbb{E}[U] = \mathbb{E}[\mathbb{E}[U \mid V]] \,.$$

To prove the claim we will proceed by induction. first note that $G_0(s) = s$. Now assume that for $n \geq 1$ we have $G_{n-1} = G \circ \ldots \circ G$ is the $(n-1)$-fold composition of G with itself. Then to compute G_n we first condition on the size of the population at generation $(n-1)$,

$$\mathbb{E}[s^{Z_n} \mid Z_{n-1} = k] = \mathbb{E}[s^{X_1^{(n)} + \ldots + X_k^{(n)}} \mid Z_{n-1} = k] \,.$$

Now we use the independence, since the Z_{n-1} only depend on the values of $X_r^{(m)}$ for $m \leq n-1$ and all the $(X_r^{(n)})_{n,r \geq 1}$ are independent and identically distributed it follows that

$$\mathbb{E}[s^{Z_n} \mid Z_{n-1} = k] = \mathbb{E}[s^{X_1^{(n)} + \ldots + X_k^{(n)}}] = G(s)^k \,,$$

where intuitively the first equality follows from the independence because conditioning on Z_{n-1} can not change the distribution of $X_1^{(n)}, \ldots, X_k^{(n)}$ and the second inequality follows because the expectation of a product of independent random variables is the product of the expectations ("*independence means multiply*"). From this, and a hefty dose of intuition we may write

$$\mathbb{E}[s^{Z_n} \mid Z_{n-1}] = G(s)^{Z_{n-1}} \,,$$

which is still a random variable (a function of Z_{n-1}). Now by the tower property, taking expectation of this random variable

$$G_n(s) = \mathbb{E}[s^{Z_n}] = \mathbb{E}[\mathbb{E}[s^{Z_n} \mid Z_{n-1}] = \mathbb{E}[G(s)^{Z_{n-1}}] = G_{n-1}(G(s)) \,,$$

and the claim follows by induction. □

Remark 1.6. In the case above we do not need the full power of the *tower property* of conditional expectations, you can make everything work using the total law of expectation you will have seen in a first or second year module, because the events $\{Z_{n-1} = k\}$ form a countable partition of the sample space. In this sense you can view the tower property as a generalisation of the Partition Theorem for expectations (some times called the law of total expectation), which will also hold in much more complicated settings including for continuous random variables.

Galton and Watson wanted to know the *extinction probability* of the branching process, the probability that the population dies out, i.e. the probability that $Z_n = 0$ for some n. Notice that if $P(X = 0) = p_0 = 0$ then every individual has at least one child and so the population can never die out. On the other hand if $P(X = 0) > 0$ then the population dies out in generation one with probability $p_0 > 0$ and hence has a positive probability to die out eventually which is lower bounded by p_0. In fact we can make much more precise statements about the probability for the population to die out.

Claim 1.7. *Let $\pi = \mathbb{P}(Z_n = 0 \text{ for some } n)$, then π is the smallest root in $[0,1]$ of the equation $\pi = G(\pi)$. In particular, assuming $p_1 = \mathbb{P}(X = 1) < 1$,*

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

"Proof". Let $\pi_n = \mathbb{P}(Z_n = 0) = G_n(0)$ where the last equality follows from the definition of the probability generating function. Since $\{Z_n = 0\} \subseteq \{Z_{n+1} = 0\}$ we see that π_n is increasing in n (and bounded above by one), therefore, *intuitively*

$$\pi = \lim_{n\to\infty} \pi_n = \lim_{n\to\infty} G_n(0).$$

Since $G_{n+1}(0) = G(G_n(0))$ and G is continuous, the above limit implies that π satisfies $\pi = G(\pi)$.

Now observe that G is convex (i.e. $G'' \geq 0$), and $G(1) = 1$, so only two things can happen in terms of the fixed points of G, which depend on the value of $\mu = \mathbb{E}[X] = G'(1)$: Either there is a single fixed point if $\mu \leq 1$ and exactly two roots if $\mu > 1$ (see the picture in lectures). In the case that $\mu > 1$, to see that π is the smallest root, s_0, of $x = G(x)$, note that G is an increasing function and $0 = \pi_0 \leq s_0$. It follows by induction that $\pi_n \leq s_0$ for all n and so $\pi \leq s_0$. □

Remark 1.8. Notice that the intuition that allowed us to state that $\pi = \lim_{n\to\infty} \pi_n$ can be misleading. In fact, for this to hold requires us to have defined a reasonable probability space (probability triple) and identified that the events we are interested in belong to the class of events that we can assign probabilities to. To see one way things can go wrong consider the following approach to trying to define a 'uniform probability' measure on the natural numbers $\mathbb{N} = \{1, 2, \ldots\}$, lets call it ρ. For such a measure we would want $\rho(\{k : k \text{ is divisible by } 2\}) = 1/2$ and similar for numbers divisible by 3 their 'density' in the natural numbers is $1/3$. Let $\mathcal{C}$ be the class of subsets $C \subseteq \mathbb{N}$ such that the 'density'

$$\rho(C) = \lim_{m\to\infty} \frac{|\{k : 1 \leq k \leq m,\ k \in C\}|}{m} \quad \text{exists.}$$

Let $C_n = \{1, 2, \ldots, n\}$, then $C_n \in \mathcal{C}$ and $C_n \subseteq C_{n+1}$. Also $\bigcup_n C_n = \mathbb{N}$. However, $\rho(C_n) = 0$ for each n, but $\rho(\mathbb{N}) = 1$, so it is not true that $\rho(\mathbb{N}) = \lim_{n\to\infty} \rho(C_n)$ and hence our intuition that led to $\pi = \lim_{n\to\infty} \pi_n$ fails in this case. It will turn out that the triple $(\mathbb{N}, \mathcal{C}, \rho)$ does not satisfy the conditions we will require of a probability triple. **Can you identify which property is missing?**

We have identified the extinction probability, but this is just one statistics that we might be interested in. For example, we might want to know about the way in which the size of the population grows or declines. Consider, applying the same argument as we have already,

$$\mathbb{E}[Z_{n+1} \mid Z_n = k] = \mathbb{E}[X_1^{(n+1)} + \ldots + X_k^{(n+1)}] = k\mu \quad \text{(by linearity of expectation)}\,.$$

So as before we are able to write a conditional expectation as a random variable, namely $\mathbb{E}[Z_{n+1} \mid Z_n] = \mu Z_n$. If we now define $M_n = Z_n/\mu^n$, then

$$\mathbb{E}[M_{n+1} \mid M_n] = M_n\,,$$

and in fact more is true (simply from the independence structure), we have

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

A process satisfying this condition is called a *martingale*.

Since the martingale we have defined above is non-negative ($M_n \geq 0$ for each n) it turns out that the *Martingale Convergence Theorem* implies that with probability one (almost surely) there exists a limiting random M_∞ such that

$$M_\infty = \lim_{n\to\infty} M_n \,.$$

Note that if $M_\infty > 0$ then the statement $Z_n/\mu^n \to M_\infty$ (almost surely) describes the asymptotic exponential growth of the population (See *D. Williams* for more details on identifying the distribution of this limiting random variable). However, if $\mu \leq 1$ then we have seen that the population goes extinct with probability one, i.e. $M_\infty = 0$ almost surely, so

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

This is a nice example to keep in mind of a case where the inequality seen in *Fatou's Lemma* is actually strict.

Chapter 2

Basic Measure Theory*

Reading: ST342/MA359
Further reading: D. Williams, Chapter 1 and A. Klenke, Section 1.1
Also: P. Billingsley, Probability and Measure, Sections 1,2,3

We begin by recalling some definitions that you will have encountered before. The idea of measure theory is that we want to assign a 'mass' or 'size' to relevant subsets of a space in a consistent way. In particular, in probability theory, these subsets will be 'events' or 'collections of outcomes' (which are subsets of a sample space), and the 'mass' is the probability - i.e. a measure of how likely that event is to occur. For example, we perform an experiment with random outcomes, such as tossing a coin or rolling a die, since the outcome depends on unknown circumstances we wish to quantify the chances of a given outcome rather than compute exactly what will happen.

We now fix some notation. Throughout these notes we will use Ω to denote a sample space, and for some set A we write $\mathcal{P}(A) = 2^A$ for the power set of A, i.e. the set of all subsets of A.

Since we are interested in probability, we will use some language which reflects this, a point $\omega \in \Omega$ is sometimes called an outcome or a realisation. We call subsets $A \subset \Omega$ events. It is important to remain conscious of the difference between a realisation (sample point) $\omega \in \Omega$ and the singleton event that ω occurred, i.e. $\{\omega\} \subset \Omega$).

Example 2.1. Consider a finite state space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ (or more generally a countable state space), you could have in mind sequences of coin-tosses like $\Omega = \{HH, HT, TH, HH\}$ (then the event at least one of two coins landed heads up would be $\{HH, HT, TH\}$).

We could take the probability of any subset of Ω to be the sum over the probabilities of each outcome contained in the set, i.e. for any $A \in \mathcal{P}(A)$

$$\mathbb{P}(A) = \sum_{\omega_i \in A} \mathbb{P}(\{\omega_i\})\,. \tag{2.1}$$

In this way we can assign a probability to any subset of Ω (any element of $\mathcal{P}(\Omega)$). However if Ω is uncountable then we can't expect to be able to assign a probability (measure) to every element of $\mathcal{P}(\Omega)$, and Eq. (2.1) becomes unworkable.

It is not the case that for arbitrary Ω all subsets of Ω can be assigned a measure in a reasonable way. In previous courses you have probably seen constructions of non-Lebesgue measurable subsets of $\mathbb{R}$. The following example also serves to indicate that we can't hope to be able to measure every subset of pretty 'simple' looking spaces.

Example 2.2. Suppose we pick a point uniformly at random on the surface of the unit sphere. Can we define the probability that this point belongs to any subset of the surface? The answer is no! There exist disjoint subsets of the surface which each 'look like' copies of the surface of the unit sphere. This is contained in the Banach–Tarski paradox (a quick google and read of Wikipedia will provide more information). Any reasonable definition of area would require that the measure of the union of disjoint things is the sum of the measure of the parts, in light of the Banach–Tarski paradox this can not hold for all subsets.

In light of the brief discussion above it is clear that we need to limit ourselves to classes of subsets that are both big enough to be useful and small enough not to cause problems with defining the probability of events.

Definition 2.3 (Algebras and σ-algebras)**.** Let Ω be a set and $\mathcal{A} \subset \mathcal{P}(\Omega)$ be a collection of subsets of Ω.

1. $\mathcal{A}$ is called an *algebra* (on Ω) if $\Omega \in \mathcal{A}$ and
   - $A \in \mathcal{A} \implies A^c = \Omega \setminus A \in \mathcal{A}$,
   - $A, B \in \mathcal{A} \implies A \cup B \in \mathcal{A}$
2. $\mathcal{A}$ is called a σ-algebra (on Ω) if $\mathcal{A}$ is an algebra and for all sequences $(A_n)_{n\geq 1}$ in $\mathcal{A}$ (i.e. $A_n \in \mathcal{A}$ for each $n \in \mathbb{N}$) then $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Since intersections can be built up from complements and unions an algebra is closed under *finite* set operations (see the exercises below). The difference between an algebra and a σ-algebra is that the latter is closed under not just finite but also *countable* set operations. The sets $A_1, A_2, \ldots$ in 2. above may or may not overlap, some could be empty or Ω itself etc., but it is important that the property holds only for collections of countably many sets in $\mathcal{A}$. When the sample space is clear from context we will not refer to it explicitly.

Exercise 2.4. Suppose that $\mathcal{A}$ is a σ-algebra on Ω, and $A_n \in \mathcal{A}$ for each $n \in \mathbb{N}$, show that

- $\emptyset \in \mathcal{A}$,
- $\cap_{i=1}^{\infty} A_i \in \mathcal{A}$.

The following result will be important in a moment when we discuss generating σ-algebras.

Exercise 2.5. Show that the intersection of an arbitrary collection of σ-algebras is again a σ-algebra.

Definition 2.6 (Set functions)**.** Let $\mathcal{A} \subset \mathcal{P}(\Omega)$ be *any* collection of subsets of Ω containing $\emptyset$. A *set function* on $\mathcal{A}$ is a function $\mu : \mathcal{A} \to [0, \infty]$ (i.e. a non-negative function on the collection of sets). We say that μ is

1. *increasing* if $\mu(A) \leq \mu(B)$ for any $A, B \in \mathcal{A}$ with $A \subset B$.
2. *additive* if $\mu(\emptyset) = 0$ and for all disjoint $A, B \in \mathcal{A}$ with $A \cup B \in \mathcal{A}$

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

3. *σ-additive* if $\mu(\emptyset) = 0$ and for all sequences $(A_n)_{n\geq 1}$ of disjoint sets in $\mathcal{A}$ with $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$

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

Note that in 2. and 3. above we need to specify that the union is an element of the collection since the collection of sets need not be a σ-algebra.

Definition 2.7 (Measure space)**.**

1. A *measurable space* is a pair $(\Omega, \mathcal{F})$ where Ω is a set and $\mathcal{F}$ is a σ-algebra on Ω.
2. A *measure space* is a triple $(\Omega, \mathcal{F}, \mu)$ where Ω is a set, $\mathcal{F}$ is a σ-algebra on Ω and $\mu\colon \mathcal{F} \to [0, \infty]$ is a σ-additive set function. In this case μ is called a *measure* on $(\Omega, \mathcal{F})$.

Observe (check) that if $(\Omega, \mathcal{F}, \mu)$ is a measure space then μ is also additive and increasing. During this module we will mostly be interested in a specific type of measure space.

Definition 2.8 (Types of measure space)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.

1. μ is called *finite* if $\mu(\Omega) < \infty$.
2. μ is called *σ-finite* if there exists a sequence of sets $(E_n)_{n\geq 1}$ in $\mathcal{F}$ such that $\mu(E_n) < \infty$ for each $n \in \mathbb{N}$, and $\bigcup_{n=1}^{\infty} E_n = \Omega$.
3. μ is called a *probability measure* if $\mu(\Omega) = 1$, and then $(\Omega, \mathcal{F}, \mu)$ is called a *probability space* and we often use the notation $(\Omega, \mathcal{F}, \mathbb{P})$ to emphasise that we have a probability measure.

We now recall an important (fundamental) consequence of the definitions above, which is that measures respect monotone limits, a continuity result. Recall this was important in our analysis of a motivating example, namely the Galton-Watson Process, in Lecture 1. This result is often important for making intuitively obvious things rigorous. First we need some notation. *This was contained in Lecture 3 of MA359 and Chapter 2.2 in ST342.*

Notation: For a sequence of sets $(F_n)_{n\geq 1}$ we write $F_n \nearrow F$ if $F_n \subseteq F_{n+1}$ for each n and $\bigcup_{n=1}^{\infty} F_n = F$. Similarly, $G_n \searrow G$ if $G_n \supseteq G_{n+1}$ for all n and $\bigcap_{n=1}^{\infty} G_n = G$.

Lemma 2.9 (Monotone convergence for measures)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.*

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

Notice that the extra restriction that $\mu(G_k) < \infty$ for some k, in the second statement is important. A canonical example of what can go wrong here is given by considering the Lebesgue measure of the sequence of sets given by $H_n = (n, \infty)$.

The following partial converse is sometimes useful (in particular for applying the Carathéodory Extension Theorem later to construct the Lebesgue measure)

Lemma 2.10. *Let $\mathcal{A}$ be and algebra and $\mu\colon \mathcal{A} \to [0,\infty)$ be an additive (and finite) set function on $\mathcal{A}$. Then μ is σ-additive if and only if (iff) for every sequence $(A_n)_{n\geq 1}$ in $\mathcal{A}$ with $A_n \searrow \emptyset$ we have $\mu(A_n) \to 0$ as $n \to \infty$.*

Proof. One direction follows from the monotone convergence theorem for measures, the other is an exercise. □

We now give some common terminology and simple consequences of the the definition of measures. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. We call a set N *null* if $N \in \mathcal{F}$ (N is measurable) and $\mu(N) = 0$. We say a statement S holds μ-almost surely (abbreviated μ-a.e.) if $F = \{\omega \in \Omega : S(w) \text{ is false } \} \in \mathcal{F}$ and F is null. If μ is a probability measure then we say an event A occurs, or statement S holds, μ-almost surely (μ-a.s.) if $\mu(A) = 1$, or $\mu(\{\omega \in \Omega : S(w) \text{ is true } \}) = 1$ (this is simply the special case of μ-a.s. when μ is also a probability measure).

Recall a measure is by definition σ-additive on *disjoint* sets. The following inequalities hold for general measurable sets.

Lemma 2.11. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, then*

- *(subadditive) $\mu(A \cup B) \leq \mu(A) + \mu(B)$ for $A, B \in \mathcal{F}$,*
- *(σ-subadditive) $\mu\Big(\bigcup_{n=1}^{\infty} A_n\Big) \leq \sum_{n=1}^{\infty} \mu(A_n)$ for $A_1, A_2, \ldots \in \mathcal{F}$.*

If μ is also a finite measure ($\mu(\Omega) < \infty$), then

- *$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$ for $A, B \in \mathcal{F}$,*
- *(inclusion-exclusion formula) for all $A_1, A_2, \ldots, A_n \in \mathcal{F}$*

$$\mu\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mu(A_i) - \sum_{i<j\leq n}\sum \mu(A_i \cap A_j) + \\ + \sum_{i<j<k\leq n}\sum\sum \mu(A_i \cap A_j \cap A_k) - \ldots + (-1)^{n-1}\mu(A_1 \cap A_2 \cap \ldots \cap A_n)\,.$$

Proof. See Example Sheet 1. □

When applying subadditivity (or countable version) in probability arguments, such bounds are often called "union bounds". Note that the restriction to finite measures in the second set of results is only because we don't want to claim that statements like $\infty = \infty - \infty$.

Example 2.12. We go back to Example 2.1. Suppose $\Omega = \{\omega_1, \omega_2, \ldots\}$ is a countable set. Then given a mass function (i.e. a non-negative function on Ω), $\bar{\mu}\colon \Omega \to [0,\infty]$, we can define a measure on the measurable space $(\Omega, \mathcal{P}(\Omega))$ by

$$\mu(A) = \sum_{\omega \in A} \bar{\mu}(\omega)\,.$$

You should check that this 'works' in the sense that we have defined a measure space. Conversely, given a measure μ on $(\Omega, \mathcal{P}(\Omega))$ we can define a mass function by

$$\bar{\mu}(x) = \mu(\{x\})\,.$$

Again check this works!

This example shows that there is a one-to-one correspondence between measures on $\mathcal{P}(\Omega)$ and mass functions on Ω. Discrete spaces provide a nice "toy" version of the general (measure) theory, which are helpful for building some useful intuition (see the example sheets and also think about them and cook up your own examples of σ-algebras etc.). However, focusing on discrete spaces we miss many important subtleties of measure theory. The discrete setting is essentially the only context in which we can define everything (σ-algebras and measures) so explicitly. In general σ-algebras are not amenable to explicit presentation. We would therefore like to be able to write down an explicit definition of a measure on some sufficiently large collection of "nice" sets, with good properties. Then we would need a concept to extend these nice sets to a σ-algebra (see σ-algebra generated by a collection below). Subsequently we would like to know that there exists a consistent extension of the measure we defined for nice sets to the entire σ-algebra (this is typically handled by the Carathéodory Extension Theorem), and further that this extension is uniquely determined by the specification we gave (typically using Dynkin's π-system Uniqueness Lemma below). In summary, the standard strategy is:

1. Define a set function that looks like the measure we want on some collection of "nice" sets.
2. Extend this collection of sets to a σ-algebra (Generated σ-algebras).
3. Show that the set function can be extended to a measure (Carathèodory Theorem).
4. Show the extension is unique (Dynkin's π-system Uniqueness Lemma).

We start with generated σ-algebras. *These were covered in Section 2.1 of ST343 and Lecture 2 'smallest sigma algebra' in MA359*

Definition 2.13 (Generated σ-algebras)**.** Let $\mathcal{A}$ be a collection of subsets of Ω (i.e. $\mathcal{A} \subseteq \mathcal{P}(\Omega)$), then the σ-algebra generated by $\mathcal{A}$ is given by

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

You should check that the the definition really defines a σ-algebra and that it is the intersection of all σ-algebras containing $\mathcal{A}$ (this is an exercise on the first sheet). Check that the intersection of σ-algebras is a σ-algebra, but in general the union of σ-algebras is not a σ-algebra. For example, if $A, B \in \mathcal{F}$ then $\sigma(\{A\}) = \{\emptyset, \Omega, A, A^c\}$, and $\sigma(\{B\})$ are σ-algebras but if $A \neq B$ then $\sigma(\{A\}) \cup \sigma(\{B\})$ is not, and in particular $\sigma(\{A, B\}) \neq \sigma(\{A\}) \cup \sigma(\{B\})$.

Remark 2.14. The following three statements hold directly from the definition (we will use them frequently so convince yourself they all hold):

1. $\mathcal{A} \subseteq \sigma(\mathcal{A})$.
2. If $\mathcal{A}_1 \subseteq \mathcal{A}_2$, then $\sigma(\mathcal{A}_1) \subseteq \sigma(\mathcal{A}_2)$.
3. $\mathcal{A}$ is a σ-algebra if and only if $\sigma(\mathcal{A}) = \mathcal{A}$.

Every topological space (e.g. the real line, plane, sphere, ...) has a σ-algebra which naturally relates to the topology, this is called the Borel σ-algebra.

Definition 2.15 (Borel σ-algebra)**.** Let Ω be a topological space with topology (set of open sets) T. The Borel σ-algebra on Ω is

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

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

Remark 2.16. The σ-algebra $\mathcal{B}(\Omega)$ depends on the topology, not just Ω, although it doesn't appear in the notation. This isn't usually an issue since the topology is normally clear from context, e.g. if $\Omega = \mathbb{R}^n$ then we will take the Euclidean topology.

Being closed under finite intersections is such an important property we give it a name.

Definition 2.17 (π-system)**.** Let $\mathcal{I}$ be a collection of subsets of Ω, then $\mathcal{I}$ is called a π-system if

$$A, B \in \mathcal{I} \implies A \cap B \in \mathcal{I}\,.$$

Remark 2.18. An algebra is necessarily a π-system.

Example 2.19. It turns out (check) $\pi(\mathbb{R}) = \{(-\infty, x] : x \in \mathbb{R}\}$ is a π-system, and

$$\sigma\big(\pi(\mathbb{R})\big) = \mathcal{B}(\mathbb{R}) = \sigma\big(\{(a, b) : a, b \in \mathbb{R}\}\big)\,. \tag{2.2}$$

Almost all "useful" (for us) subsets of $\mathbb{R}$ are contained in $\mathcal{B}(\mathbb{R})$, but a generic element of $\mathcal{B}(\mathbb{R})$ can be very complicated. However, it turns out that almost everything we need to know is contained in 2.2.

We now observe why π-systems are so useful. *This was covered in Chapter 2.3 of ST342 and Lecture 9 of MA359.*

Lemma 2.20 (Dynkin's Uniqueness Lemma for π-systems)**.** *Let μ_1,μ_2, be measures on a measurable space $(\Omega, \mathcal{F})$ and let $\mathcal{I} \subseteq \mathcal{F}$ be a π-system. If $\mu_1(\Omega) = \mu_2(\Omega) < \infty$, and $\mu_1 = \mu_2$ on $\mathcal{I}$, then $\mu_1 = \mu_2$ on $\sigma(\mathcal{I})$. In particular if $\mathcal{F} = \sigma(\mathcal{I})$ then $\mu_1 = \mu_2$.*

Proof. The proof relies on Dynkin's π-λ Theorem (also sometimes simply called Dynkin's Lemma, and sometimes λ is replaced by d in the name) which requires yet another class of sets called a λ(or a d)-system. The theorem states that the σ-algebra generated by a π-system is equal to the λ-system generated by the same π-system. Since we won't use these concepts again we won't cover the details here, but the interested reader can check Appendix A.1 of D. Williams, or Sections 1.1-1.3 of A. Klenke. □

The result can be generalised to σ-finite measures with a little extra work, this is useful when applying it for example to the Lebesgue measure on $\mathbb{R}^d$.

The lemma above covers the uniqueness problem for finite measures, but what about existence? *The following parts were certainly covered in ST342, if you took MA359 and are interested in more details you can see 'C. S. Kubrusly. Essentials of measure theory. Springer, Cham, 2015', which is available online through the library.*

Theorem 2.21 (Carathéodory's Extension Theorem)**.** *Let Ω be a set, $\mathcal{A} \subset \mathcal{P}(\Omega)$ an algebra on Ω, and $\mathcal{F} = \sigma(\mathcal{A})$. If $\mu_0 \colon \mathcal{A} \to [0, \infty]$ is a* σ-additive set function*, then there exists a measure μ on $(\Omega, \mathcal{F})$ such that $\mu = \mu_0$ on $\mathcal{A}$. Furthermore, if $\mu_0(\Omega) < \infty$ then the extension is unique.*

Remark 2.22. Note that the 'furthermore' part is simply a consequence of the previous uniqueness lemma since an algebra must be a π-system.

This theorem reduces the problem of constructing a measure on $(\Omega, \mathcal{F})$ to constructing a countably additive set function on an algebra that generates $\mathcal{F}$. The tricky point is normally to show σ-additivity.

The proof of Carathéodory's Extension Theorem follows much the same argument as the construction of the Lebesgue measure that you may have seen before. It starts by defining an outer measure μ^* on all subsets of Ω in terms of the size of the smallest (in the sense of μ_0) coverings by sets in $\mathcal{A}$. Then call a set B 'measurable' if it satisfies

$$\mu^*(B \cap E) + \mu^*(B^c \cap E) = \mu^*(E) \quad \forall\ E \subset \Omega\,,$$

i.e. if B 'splits every set properly'. If $\mu_0(\Omega) < \infty$ then the condition above reduces to $\mu^*(B) + \mu^*(B^c) = \mu_0(\Omega)$, which intuitively says that it is possible to cover B 'efficiently' with sets in $\mathcal{A}$. Then one needs to check that the set of 'measurable' sets really forms a σ-algebra, and that μ^* is a σ-additive set function on this σ-algebra that extends μ_0. For details see Appendix A.1 of D. Williams.

Example 2.23 (Lebesgue measure on $(0,1]$)**.** Let $\Omega = (0,1]$ and $\mathcal{A} = \{(a_1, b_1] \cup (a_2, b_2] \cup \ldots \cup (a_r, b_r] \,:\, 0 \le a_1 \le b_1 \le a_2 \ldots \le a_r \le b_r \le 1\}$ be the collection of all subsets of $(0,1]$ that can be written as a finite union of disjoint (half open) intervals. It turns out that $\mathcal{A}$ is an algebra and $\sigma(\mathcal{A}) = \mathcal{B}(0,1]$. Let $\mu_0(F) = \sum_{k \le r}(b_k - a_k)$ for $F \in \mathcal{A}$. Then μ_0 is well defined and σ-additive on $\mathcal{A}$. It is relatively straightforward to show that μ_0 is additive, the tricky part is to show that μ_0 is countably additive on $\mathcal{A}$ which requires a compactness argument. It follows from Carathéodory's Extension Theorem that there exists a unique measure on $((0,1], \mathcal{B}(0,1])$ that extends μ_0. This is the Lebesgue measure on $(0,1]$. A bit more effort is needed to cover uniqueness if we want to use the same argument on the whole of $\mathbb{R}$. For details see for example A. Klenke Section 1.1-1.3.

Corollary 2.24. *There exists a unique Borel measure μ on $\mathbb{R}$ such that for all $a < b \in \mathbb{R}$ we have $\mu((a,b]) = b - a$. The measure μ is called the Lebesgue measure.*

Remark 2.25. You have probably seen the set of Lebesgue measurable sets, often denoted $\mathcal{M}_{\text{Leb}}$, in place of $\mathcal{B}(\mathbb{R})$ previously. It turns out that that any set in $\mathcal{M}_{\text{Leb}}$ differs from a set in $\mathcal{B}(\mathbb{R})$ at most on a set of measure zero. That is $\mathcal{M}_{\text{Leb}}$ is the completion of $\mathcal{B}(\mathbb{R})$ by null sets i.e. $\mathcal{M}_{\text{Leb}} = \sigma(\mathcal{B}(\mathbb{R}) \cup \mathcal{N})$ where $\mathcal{N}$ is the class of subsets Borel null sets.

Recall that in discrete example (Example 2.12) there was a one-to-one correspondence between measures on the natural σ-algebra and mass functions on Ω. It turns out that we can say something at least in the same spirit for probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. *This part may be new to people who took MA359.*

Definition 2.26 (Distribution function of $\mathbb{P}$)**.** Let $\mathbb{P}$ be a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The *distribution function* of $\mathbb{P}$ is $F\colon \mathbb{R} \to [0, \infty]$ defined by $F(x) = \mathbb{P}\big((-\infty, x]\big)$.

We now observe some important properties of the distribution function defined above.

Lemma 2.27 (Properties of distribution functions)**.** *The function F in Definition 2.26 satisfies the following properties;*

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

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

3. *F is right continuous, i.e. $y \searrow x$ implies $F(y) \to F(x)$.*

Proof. 1. follows immediately from the observation that a measure is increasing. 2. follows immediately from monotone convergence for measures. To prove 3. use monotone convergence for measures again to show

$$F\Big(x + \frac{1}{n}\Big) = \mathbb{P}\Big(\Big(-\infty, x + \frac{1}{n}\Big]\Big) \searrow \mathbb{P}\big((-\infty, x]\big) \quad \text{as } n \to \infty\,,$$

then the result follows from 1., i.e. since F is monotone. □

Since probability measures are finite, Carathéodory's Extension Theorem together with Dynkin's Uniqueness Lemma, gives us a one-to-one correspondence between probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and distribution functions $F\colon \Omega \to [0,1]$.

Theorem 2.28 (Lebesgue)**.** *Let $F\colon \mathbb{R} \to [0,1]$ be a (weakly) increasing, right continuous function with $F(-\infty) = 0$ and $F(\infty) = 1$. Then there exists a unique Borel probability measure $\mathbb{P}_F$ such that $\mathbb{P}_F\big((-\infty, x]\big) = F(x)$. Every Borel probability measure on $\mathbb{R}$ (i.e. on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$) arises in this way.*

Definition 2.29. In light of Lemma 2.27 and Theorem 2.28, we call any function $F\colon \mathbb{R} \to [0,1]$ satisfying properties 1.-3. of Lemma 2.27 a *distribution function.*

Remark 2.30. A Borel probability measure which is induced by a distribution function F via $\mu_F((a,b]) = F(b) - F(a)$ for each $a < b \in \mathbb{R}$ is often call a *Lebesgue-Stieltjes measure.*

Sketch proof of Theorem 2.28. (Non-examinable, though you should be bale to follow and use similar arguments). Suppose F is a distribution function on $\mathbb{R}$. For existence we will apply Carathéodory's Extension Theorem. The construction follows the same argument as for the Lebesgue measure on $(0,1]$, c.f. Example 2.23. Just as in that example, we consider the algebra, $\mathcal{A}$, of sets consisting of all subsets of $\mathbb{R}$ that can be written as a finite disjoint union of half open intervals (but now we may have include intervals of the form $(-\infty, b]$ and (a, ∞)). Then this algebra generates $\mathcal{B}(\mathbb{R})$. On this algebra define the measure $\mu_0\left(\bigcup_{n=1}^r (a_n, b_n]\right) = \sum_{n=1}^r F(b_n) - F(a_n)$. It is relatively simple (check) to show that μ_0 is finitely additive. Then it remains to show that μ_0 is σ-additive on the algebra $\mathcal{A}$ (this is the heart of the proof). Note that we have not yet used right continuity of F, in fact it will be important for proving this countable additivity.

We need to show that μ_0 on $\mathcal{A}$ is σ-additive. By Lemma 2.10 it is sufficient to show that for any sequence (A_n) in $\mathcal{A}$ with $A_n \searrow \emptyset$ we have $\mu_0(A_n) \to 0$ as $n \to \infty$. Suppose for contradiction there exists a decreasing sequence H_n in $\mathcal{A}$ with $H_n \searrow \emptyset$, but there exists an $\varepsilon > 0$ such that $\mu_0(H_n) \geq 2\epsilon$ for all n (i.e. $\mu_0(H_n) \not\to 0$). We will construct a sequence of non empty decreasing compact sets contained in the (H_n), then by a standard compactness argument the intersection (which is a subset of $\lim H_n$) must be non-empty, which gives the desired contradiction.

First we approximate H_n with $B_n = H_n \cap (-\ell, \ell] \in \mathcal{A}$. Observe that

$$\mu_0(H_n \setminus B_n) \leq \mu_0\big((-\infty, -\ell] \cup (\ell, \infty)\big) = F(-\ell) + 1 - F(\ell) \to 0 \quad \text{as } \ell \to \infty\,.$$

Then, by finite additivity, for ℓ sufficiently large $\mu_0(B_n) \geq \varepsilon$ for all n (since $H_n = B_n \sqcup (H_n \setminus B_n)$ and $\mu_0(H_n) \geq 2\varepsilon$).

Since B_n is of the form $\bigcup_{i=1}^{k_n}(a_{n,i}, b_{n,i}]$ we can always find a family $\bar{a}_{n,i}$ such that $a_{n,i} < \bar{a}_{n,i} < b_{n,i}$ and $C_n = \bigcup_{i=1}^{k_n}(\bar{a}_{n,i}, b_{n,i}]$ approximate B_n in the sense

$$\mu_0(B_n \setminus C_n) < 2^{-n}\varepsilon \quad \text{for each } n.$$

Then the closure, given by including each of the points $\bar{a}_{n,i}$, satisfies $\overline{C}_n \subset B_n$ (note the $\overline{C}_n$ are closed and bounded hence compact). Then, since (B_n) is decreasing

$$\begin{aligned}\mu_0\left(\bigcap_{i=1}^{n} C_i\right) &= \mu_0(B_n) - \mu_0\left(B_n \setminus \bigcap_{i=1}^{n} C_i\right) \geq \mu_0(B_n) - \mu_0\left(\bigcup_{i=1}^{n}(B_i \setminus C_i)\right) \\ &\geq 2\varepsilon - \sum_{i=1}^{n} \varepsilon 2^{-i} > \varepsilon\,,\end{aligned}$$

so $\bigcap_{i=1}^{n} \overline{C}_i \supset \bigcap_{i=1}^{n} C_i$ is non-empty, and by the Heine-Borel theorem also compact, for every n. It follows that $\bigcap_{i=1}^{\infty} \overline{C}_i \neq \emptyset$ (otherwise we can construct an open covering of $\overline{C}_1$ with no finite sub-covering which would contradict compactness). Finally, observe that this implies $\bigcap H_n \neq \emptyset$, as required.

To complete the proof, observe that each Borel probability measure induces a distribution function by Lemma 2.27, and that the measure μ_F induced by the above construction must be unique by Dynkin's Uniqueness Lemma 2.20. □

Finally, another result which is often useful is that the measure of any measurable sets can be approximated by the measure of a set in a generating algebra (recall algebras do not have to be closed under countable set operations). Intuitively this is reasonably clear and makes precise the idea that you can approximate the measure of any set constructed from a countable collection of sets in terms of an arbitrarily large finite collection of sets. Recall the *symmetric difference* of two sets A and B is defined by $A \triangle B = (A \setminus B) \cup (B \setminus A)$, which is the set of all points in exactly one of A or B but not both (i.e. the set of points who's "included-ness" between A and B differs).

Theorem 2.31 (Approximation theorem for probability measures)**.** *Suppose $(\Omega, \mathcal{F}, \mu)$ is a probability triple and $\mathcal{A} \subset \mathcal{F}$ is an algebra. For every $B \in \sigma(\mathcal{A})$ and $\varepsilon > 0$, there exists $A \in \mathcal{A}$ such that $\mu(A \triangle B) < \varepsilon$. In particular the measure of A is within ε of the measure of B.*

Again this theorem can be extended to the σ-finite setting.

Proof. Proofs can be found for for example in the ST342 notes or *A. Klenke* or most other books on Measure Theory. There is a proof that uses a nice application of Dynkin's π-λ Theorem. □

Chapter 3

Random Variables and Integration*

Reading: ST342/MA359
Further reading: D. Williams, Chapter 2 & 3 and A. Klenke, Section 1.4 & 1.5

In general it is a central task of mathematics to study morphisms between objects, i.e. structure preserving maps. For topological spaces, these are continuous maps (you will have seen in metric spaces/topology courses), and for measurable spaces these are measurable maps.

An extremely important example of such maps in probability theory is when the domain is given by the sample space of some probability space, Ω and the co-domain is the real numbers, $\mathbb{R}$ or the extended reals $\overline{\mathbb{R}} = [-\infty, \infty]$. Such measurable maps are called (real valued or extended real valued) random variables. It is often the case that the probability space itself may be very difficult to understand, and indeed events $A \subseteq \Omega$ are not observed directly. We may think of random variables as observations of a random experiment (something we can quantify given an outcome of some complicated experiment). The probabilities of the random observations is then given in terms of the distribution of the corresponding random variable, which is the image measure of the underlying probability measure under the measurable map. These ideas should make more sense by the end of this Chapter, but it is worth trying to keep the big picture in mind if you ever feel lost in a sea of definitions and technicalities.

Again, this chapter should be a refresher of material covered in the prerequisite modules. However, in particular if you took MA359, you may not have come across some of the language (e.g. random variables) and the emphasis may be very different.

Example 3.1 (Coin tossing)**.** We would like to be able to describe on a single probability space an experiment in which we toss a coin an arbitrary number of times and view the outcomes. Let $\Omega = \{H, T\}^{\mathbb{N}}$, so that a realisation is a sequence $\omega = (\omega_1, \omega_2, \ldots) \in \Omega$, such that $w_n \in \{H, T\}$ for each $n \in \mathbb{N}$. At the very least we would like to be able to say if the event $\{$the n^{th} coin was head$\}$ occurs or not in a given experiment, i.e. $\{\omega : \omega_n = H\}$ should be measurable. We therefore let $\mathcal{F} = \sigma\big(\{\omega : \omega_n = W\} : n \in \mathbb{N},\ W \in \{H, T\}\}\big)$. It turns out that $\mathcal{F} \neq \mathcal{P}(\Omega)$ but $\mathcal{F}$ is "big enough", for example the truth set of the statement "the proportion of heads in the first n tosses" converges to $1/2$ as $n \to \infty$ is measurable, in the notation we have developed in the previous chapter

$$F = \left\{\omega \in \Omega : \frac{\#(k \leq n : \omega_k = H)}{n} \to \frac{1}{2}\right\} \in \mathcal{F}\,.$$

Note that (heuristically) we can never "see" an entire outcome of the infinite sequence, but we can observe the number of heads in the first n tosses for each n. We will make this idea of random variables as things we can observe or measure given a random outcome rigorous in the following. We will return to this example later once we have covered some definitions and basic results related to random variables.

Let Ω and Ω' be two sets and $f : \Omega \to \Omega'$. For $A \subseteq \Omega'$ we define the pre-image of A under f by $f^{-1}(A) = \{\omega \in \Omega : f(\omega) \in A\}$. For a collection of sets $\mathcal{A}' \subseteq \mathcal{P}(\Omega')$ we will use the notation $f^{-1}(\mathcal{A}')$ to denote the collection of all the preimages of sets contained in $\mathcal{A}'$, defined by $f^{-1}(\mathcal{A}') = \{f^{-1}(A') : A' \in \mathcal{A}'\} \subseteq \mathcal{P}(\Omega)$.

Remark 3.2. In general, for $A \subseteq \Omega'$ and $B \subseteq \Omega$, we may have $h(h^{-1}(A)) \neq A$ and $h^{-1}(h(B)) \neq B$.

Exercise 3.3. Find examples where the (non)equality's hold in the previous remark. Try to find at least one function for each of the four cases. Hint: It is sufficient to consider $\Omega = \Omega' = [0, 1]$.

3.1 Measurable Maps

The contents of this section were largely in MA359 around week 4 and covered in ST342 in Chapter 4.

Lemma 3.4 (Elementary properties of the pre-image)**.** *Let $f : \Omega \to \Omega'$.*

1. *The map f^{-1} preserves all set operations:*

$$f^{-1}\left(\bigcup_\alpha A_\alpha\right) = \bigcup_\alpha f^{-1}(A_\alpha), \quad f^{-1}(A^c) = \left(f^{-1}(A)\right)^c .$$

2. *If $\mathcal{C} \subseteq \mathcal{P}(\Omega')$ then $f^{-1}(\sigma(\mathcal{C})) = \sigma\left(f^{-1}(\mathcal{C})\right)$.*

Proof. You have almost certainly come across this result before (e.g. the prerequisites or earlier). The first result, 1., is left as an exercise but is just application of the definitions with no deeper ideas being used.

For 2., First suppose we show that the RHS is contained in the the LHS. You should check that $f^{-1}(\sigma(\mathcal{C}))$ is a σ-algebra (you may use the result of part 1.). Since $\mathcal{C} \subseteq \sigma(\mathcal{C})$ (by definition), it follows that $f^{-1}(\mathcal{C}) \subseteq f^{-1}(\sigma(\mathcal{C}))$, so $\sigma\big(f^{-1}(\mathcal{C})\big) \subseteq \sigma\big(f^{-1}(\sigma(\mathcal{C}))\big)$ but $\sigma\big(f^{-1}(\sigma(\mathcal{C}))\big) = f^{-1}(\sigma(\mathcal{C}))$ by our first observation.

It remains to show that the LHS is contained in the RHS. You should check that $\mathcal{G} = \{B : f^{-1}(B) \in \sigma\big(f^{-1}(\mathcal{C})\big)\}$ is a σ-algebra (again this follows from part 1.). Suppose $C \in \mathcal{C}$ then $f^{-1}(C) \in f^{-1}(\mathcal{C}) \subseteq \sigma\big(f^{-1}(\mathcal{C})\big)$ so $\mathcal{C} \subseteq \mathcal{G}$. It follows that $\sigma(\mathcal{C}) \subseteq \sigma(\mathcal{G}) = \mathcal{G}$, and hence $f^{-1}\big(\sigma(\mathcal{C})\big) \subseteq f^{-1}\big(\mathcal{G}\big) \subseteq \sigma\big(f^{-1}(\mathcal{C})\big)$. □

Definition 3.5 (Measurable functions)**.** Suppose $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ are measurable spaces, then $f : \Omega \to \Lambda$ is called *measurable* (with respect to $\mathcal{F}, \mathcal{G}$) if

$$G \in \mathcal{G} \quad \Longrightarrow \quad f^{-1}(G) \in \mathcal{F} ,$$

i.e. $f^{-1}(\mathcal{G}) \subseteq \mathcal{F}$. In words, if the preimage of every measurable set is measurable.

Notation: In the special case that $(\Lambda, \mathcal{G})$ in the previous definition is given by $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ then f is called $\mathcal{F}$-measurable (i.e. we drop the reference to $\mathcal{B}(\mathbb{R})$ when it is clear from context). The class of such functions on $(\Omega, \mathcal{F})$ is written $m\mathcal{F}$.

It is not normally possible to check that a function is measurable by checking the preimage of all measurable sets (most σ-algebras are too large). The following proposition is therefore extremely useful for checking if a function is measurable. It states that it is sufficient to check for measurability on any class of sets that generates the full σ-algebra.

Proposition 3.6. *Suppose $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ are measurable spaces, and $f : \Omega \to \Lambda$. If $\mathcal{C} \subseteq \mathcal{G}$ and $\sigma(\mathcal{C}) = \mathcal{G}$ then f is $\mathcal{F}, \mathcal{G}$-measurable if and only if $f^{-1}(\mathcal{C}) \subseteq \mathcal{F}$.*

Proof. Fix $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ measurable spaces, $f : \Omega \to \Lambda$, and $\mathcal{C} \subseteq \mathcal{G}$ such that $\sigma(\mathcal{C}) = \mathcal{G}$. First suppose $f^{-1}(\mathcal{C}) \subseteq \mathcal{F}$. Let $\mathcal{A} = \{A \in \mathcal{G} : f^{-1}(A) \in \mathcal{F}\} \subseteq \mathcal{G}$, then just as in the proof of the previous lemma it turns out that $\mathcal{A}$ is a σ-algebra. We want to show that $\mathcal{A} = \mathcal{G}$. If $C \in \mathcal{C}$ then $C \in \mathcal{G}$ (since $\mathcal{G} = \sigma(\mathcal{C})$) and $f^{-1}(C) \in \mathcal{F}$ (since $f^{-1}(\mathcal{C}) \subseteq \mathcal{F}$) so $C \in \mathcal{A}$, i.e. $\mathcal{C} \subseteq \mathcal{A}$. Since $\mathcal{A}$ is a σ-algebra it follows that $\mathcal{G} = \sigma(\mathcal{C}) \subseteq \mathcal{A}$ as required.

Now suppose f is $\mathcal{F}, \mathcal{G}$-measurable, then $f^{-1}(\mathcal{C}) \subseteq f^{-1}(\mathcal{G}) \subseteq \mathcal{F}$. The first inclusion follows from $\mathcal{C} \subseteq \sigma(\mathcal{C}) = \mathcal{G}$, and the second since f is $\mathcal{F}, \mathcal{G}$-measurable. □

Recall, we write $\overline{\mathbb{R}}$ for the extended reals, and $\{f \leq t\}$ as a short way of writing $\{\omega \in \Omega : f(\omega) \leq t\}$.

Corollary 3.7. *Suppose $(\Omega, \mathcal{F})$ is measurable space, then $f \colon \Omega \to \mathbb{R}$ or $f \colon \Omega \to \overline{\mathbb{R}} = [-\infty, \infty]$ is in $m\mathcal{F}$ if an only if $\{\omega \in \Omega : f(\omega) \leq t\} \in \mathcal{F}$ for each $t \in \mathbb{R}$.*

Remark 3.8. By the same reasoning as the above Corollary, to show that $f \in m\mathcal{F}$ it is sufficient to check $f^{-1}\big((a, b)\big) \in \mathcal{F}$ for each $a < b \in \mathbb{R}$, or $f^{-1}\big([a, \infty)\big) \in \mathcal{F}$ for each $a \in \mathbb{R}$, etc. . These collections of sets generate $\mathcal{B}(\mathbb{R})$.

Remark 3.9. It is worth keeping in mind that real-valued functions on Ω generalise subsets of Ω in a natural way, with the indicator function $\mathbb{1}_A$ corresponding to the subset A. Check that $\mathbb{1}_A$ is a measurable function if and only if A is a measurable set, i.e. $A \in \mathcal{F}$.

Lemma 3.10. *$m\mathcal{F}$ is an algebra over $\mathbb{R}$, i.e. if $\lambda \in \mathbb{R}$ and $f_1, f_2 \in m\mathcal{F}$ then*

$$f_1 + f_2 \in m\mathcal{F}\,, \quad f_1 f_2 \in m\mathcal{F}\,, \quad \text{and} \quad \lambda f_1 \in m\mathcal{F}\,.$$

Proof. (As way of example we will prove $f_1 + f_2 \in m\mathcal{F}$). Fix $a \in \mathbb{R}$ and $f_1, f_2 \in m\mathcal{F}$, we will show $\{f_1 + f_2 > a\} \in \mathcal{F}$ which is sufficient by the remark above. It is clear that

$$\{f_1 + f_2 \geq a\} \supseteq \bigcup_{q \in \mathbb{Q}} \Big(\{f_1 > q\} \cap \{f_2 > a - q\} \Big)\,.$$

Notice that if $f_1 + f_2 > a$ then $f_1 > a - f_2$ and hence there exist some $q \in \mathbb{Q}$ such that $f_1 > q > a - f_2$, and hence

$$\{f_1 + f_2 \geq a\} = \bigcup_{q \in \mathbb{Q}} \Big(\{f_1 > q\} \cap \{f_2 > a - q\} \Big)\,.$$

Then $\{f_1 + f_2 \geq a\} \in \mathcal{F}$, since we have written $\{f_1 + f_2 \geq a\}$ as a countable union of measurable sets, and $\mathcal{F}$ is a σ-algebra. □

Lemma 3.11 (Composition Lemma)**.** *Suppose* $(\Omega_1, \mathcal{F}_1)$*,* $(\Omega_2, \mathcal{F}_2)$ *and* $(\Omega_3, \mathcal{F}_3)$ *are measurable spaces and* $f : \Omega_1 \to \Omega_2$*,* $g : \Omega_2 \to \Omega_3$ *are measurable (w.r.t* $\mathcal{F}_1, \mathcal{F}_2$ *and* $\mathcal{F}_2, \mathcal{F}_3$ *respectively), then* $g \circ f : \Omega_1 \to \Omega_3$ *is* $\mathcal{F}_1, \mathcal{F}_3$*-measurable.*

Sketch proof. (By picture)

$$\Omega_1 \xrightarrow{f} \Omega_2 \xrightarrow{g} \Omega_3 \,,$$

$$\mathcal{F}_1 \xleftarrow{f^{-1}} \mathcal{F}_2 \xleftarrow{g^{-1}} \mathcal{F}_3 \,.$$

Notation: Recall the following notation for the limsup and liminf of a real sequence $(x_n)_{n\geq 1}$.

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

Then $\lim_{\to\infty} x_n$ exists in $\overline{\mathbb{R}}$ if $\liminf x_n = \limsup x_n$. Note, it is convenient to consider the extended reals (two point compactification of $\mathbb{R}$) so that the lim inf and lim sup always exists.

Aside on $\overline{\mathbb{R}}$: [If you have come across the extended reals before this brief summary can safely be skipped]. We define $\overline{\mathbb{R}} = [-\infty, \infty] = \mathbb{R} \cup \{-\infty, \infty\}$. We consider the usual topology on $\overline{\mathbb{R}}$ given by declaring sets of the form (a, b), $[-\infty, a)$, $(a, \infty]$ open for each $a, b \in \mathbb{R}$, and any union of theses. For the sake of this discussion we denote this topology by $\overline{T}$ and the usual topology on $\mathbb{R}$ by T (usual we don't have need to refer to the topology directly like this). In this sense it is often called the two point compactification of $\mathbb{R}$. This topology is metrisable as follows. Define $\phi\colon [-1, 1] \to \mathbb{R}$ by

$$\phi(x) = \begin{cases} \tan(\pi x/2) & \text{if } x \in (-1, 1) \,. \\ -\infty & \text{if } x = -1 \,, \\ \infty & \text{if } x = 1 \,, \end{cases}$$

then ϕ is bijective, and $d(x, y) = |\phi^{-1}(x) - \phi^{-1}(y)|$ for $x, y \in \overline{\mathbb{R}}$ defines a metric on $\overline{\mathbb{R}}$. It turns out the ϕ and ϕ^{-1} are continuous with respect to to $d(\cdot, \cdot)$. Hence $\overline{\mathbb{R}}$ is topologically isomorphic to $[-1, 1]$.

It turns out that $\overline{T}|_{\mathbb{R}} := \{G \cap \mathbb{R} \,:\, G \in \overline{T}\} = T$ and hence $\mathcal{B}(\overline{\mathbb{R}})|_{\mathbb{R}} = \mathcal{B}(\mathbb{R})$. Consequently, any measurable real value random variable can be considered in an obvious way as a measurable extended real valued random variable. Thus $\overline{\mathbb{R}}$ is really an extension of the real line, and the inclusion map from $\mathbb{R}$ to $\overline{\mathbb{R}}$ is measurable.

Finally, we note that the the intervals $[\infty, a)$ for $a \in \overline{\mathbb{R}}$ generate the Borel σ-algebra $\mathcal{B}(\overline{\mathbb{R}})$ (similarly for $(a, \infty]$). Hence, by Proposition 3.6, if $X^{-1}[-\infty, a)$ is measurable for each $a \in \overline{\mathbb{R}}$ then X is measurable (similarly for $(a, \infty]$).

We define $a + \infty = \infty + a = \infty$ for $a \in [0, \infty]$. Also $a \cdot \infty = \infty \cdot a = \infty$ if $a > 0$ and $0 \cdot \infty = \infty \cdot 0 = 0$. This way commutative, associative and distributive laws hold in $[0, \infty]$ without any restriction. If X is an extended real valued random variable then $\mathbb{E}[X] = \infty$ if $\mathbb{P}(X = \infty) > 0$.

Lemma 3.12 (Measurability of inf, liminfs etc.)**.** *Suppose* $(\Omega, \mathcal{F})$ *is a measurable space and* $(f_n)_{n\geq 1}$ *a sequence in* $m\mathcal{F}$*. Then the following functions belong to* $m\mathcal{F}$*,*

$$\inf f_n \,, \quad \liminf f_n \,, \quad \limsup f_n \,,$$

and $\{\omega \in \Omega : \lim f_n(\omega) \in \mathbb{R}\} \in \mathcal{F}$.

Proof. Again it is sufficient to check measurablility for the preimage of any collection of sets that generate $\mathcal{B}(\overline{\mathbb{R}})$.

($\inf f_n$) $\{\omega : \inf f_n(\omega) \geq a\} = \bigcap_{n\geq 1}\{\omega : f_n(\omega) \geq a\} \in \mathcal{F}$ for each $a \in \overline{\mathbb{R}}$, since $\inf f_n(\omega) \geq a$ if and only if for all $n \in \mathbb{N}$, $f_n \geq a$.

($\liminf f_n$) Let $L_n(\omega) = \inf_{m\geq n} f_m(\omega)$, then $L_n(\omega)$ is an increasing for each $\omega \in \Omega$, so $L(\omega) = \lim_{n\to\infty} L_n(\omega) = \liminf f_n(\omega) = \sup_n L_n(\omega)$. By the previous part we know that for each $n \in \mathbb{N}$ we have $L_n \in m\mathcal{F}$, so $\{L \leq a\} = \bigcap_n \{L_n \leq a\} \in \mathcal{F}$. The proof that $\limsup f_n$ is measurable follows the same strategy.

We finish by showing that $\{\omega : \lim f_n(\omega) \in \mathbb{R}\}$ is measurable by writing it as a finite intersection of measurable sets,

$$\{\omega : \lim f_n(\omega) \in \mathbb{R}\} = \{\omega : \limsup f_n(\omega) < \infty\} \cap \{\omega : \liminf f_n(\omega) > -\infty\} \cap \\ \cap \{\omega : \limsup f_n(\omega) - \liminf f_n(\omega) = 0\}$$

the first two sets on the RHS are measurable by the previous part of the proof, and the final set is measurable by Lemma 3.10 (since $\{0\} \in \mathcal{B}(\mathbb{R})$). □

Beyond all the measure structure, σ-algebras, etc. we have been discussing so far, we will also need a various concepts of closeness. Most immediately obvious with the set-up so far is provided by topological structure. Recall $(\Omega, \mathcal{T})$ is called a topological space if $\mathcal{T}$ is closed under finite intersections and arbitrary unions, and $\emptyset, \Omega \in \mathcal{T}$. The sets in $\mathcal{T}$ are called the open sets, and $f : \Omega \to \Omega'$ is continuous if the preimage of any open set is open.

Definition 3.13 (Borel functions)**.** Let $(\Omega, \mathcal{T})$ be a topological space, and $(\Omega, \mathcal{B}(\Omega))$ be the Borel measure space on Ω. Then $f : \Omega \to \mathbb{R}$ is called *Borel* if $f \in m\mathcal{B}(\Omega)$, i.e. it is $\mathcal{B}(\Omega), \mathcal{B}(\mathbb{R})$-measurable.

Lemma 3.14. *If $f : \Omega \to \mathbb{R}$ is continuous then f is a Borel function.*

Proof. Let $\mathcal{C} = \{\text{open subsets of } \mathbb{R}\}$ and use Proposition 3.6. □

Exercise 3.15 (The converse is not in general true)**.** Find a function which is Borel but not continuous.

3.2 Random Objects and Random Variables

Since we are interested in probability, we give some things special names in the case that we have a probability space. ***If you took MA359 then these names might be new.***

Definition 3.16 (Random objects and random variables)**.** Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $(\Lambda, \mathcal{G})$ a measurable space. A function $X : \Omega \to \Lambda$ is called a *random object* if X is $\mathcal{F}, \mathcal{G}$-measurable. In the case $\Lambda = \mathbb{R}$ and $\mathcal{G} = \mathcal{B}(\mathbb{R})$ then $X \in m\mathcal{F}$ is called a *random variable* (r.v.).

Remark 3.17. If X is a r.v. then $\{w \in \Omega : X(\omega) \leq a\} \in \mathcal{F}$ so $\mathbb{P}(\{\omega : X(\omega) \leq a\}) = \mathbb{P}(X \leq a)$ makes sense.

Now that we have some more results and language we return to the coin tossing example.

Example 3.18 (Back to coin tossing)**.** Recall $\Omega = \{H,T\}^{\mathbb{N}}$, $\omega = (\omega_1, \omega_2, \ldots)$ such that $\omega_n \in \{H,T\}$ for each $n \in \mathbb{N}$, and $\mathcal{F} = \sigma\big(\{\omega \in \Omega \,:\, \omega_n = W\} \,:\, n \in \mathbb{N},\ W \in \{H,T\}\big)$. Let $X_n = \mathbb{1}_{\{\omega_n = H\}}$ (you should check this is a r.v.) and $S_n = \sum_{i=1}^n X_i$, i.e. S_n is the number of heads in the first n tosses (sum of a finite number of r.v.s is a r.v.). Let $A_p = \{\omega \,:\, \frac{\#\text{heads in n tosses}}{n} \to p\} = \{\omega \,:\, \frac{1}{n}S_n \to p\}$, then

$$A_p = \{\limsup \frac{S_n}{n} = p\} \cap \{\liminf \frac{S_n}{n} = p\} \in \mathcal{F}\,,$$

by Lemma 3.12. We will see (and prove) later in the module that if $\mathbb{P}(X_n = 1) = p$ then $\mathbb{P}(A_p) = 1$ (the Strong Law of Large Numbers).

It turns out, in the example above, that the choice of $\mathcal{F}$ is actually the smallest σ-algebra such that X_n is a measurable function for each n. We will write $\mathcal{F} = \sigma(X_n \,:\, n \in \mathbb{N})$.

Definition 3.19 (σ-algebra generated by a r.v.)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X \in m\mathcal{F}$ a random variable. Then the σ-algebra generated by X is

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

Remark 3.20. $\sigma(X)$ is in fact a σ-algebra (you should check that it follows immediately from Lemma 3.4). By definition, since $X \in m\mathcal{F}$ we must have $\sigma(X) \subseteq \mathcal{F}$. Moreover $\sigma(X) = \sigma(\big\{\{X \le t\} \,:\, t \in \mathbb{R}\big\})$ by Corollary 3.7.

Definition 3.21 (σ-algebra generated by a family of r.v.s)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X_\alpha \,:\, \alpha \in \mathcal{I})$ a collection of random variables, then

$$\sigma(X_\alpha \,:\, \alpha \in \mathcal{I}) = \sigma\Big(\bigcup_{\alpha \in \mathcal{I}} \sigma(X_\alpha)\Big) = \sigma(\{X_\alpha \le t\} \,:\, \alpha \in \mathcal{I},\ t \in \mathbb{R}\})\,.$$

It is also going to be very important later to have a concept of stronger forms of measurablility. Since X is a random variable we know it must be measurable with respect to the σ-algebra chosen on the domain (normally called $\mathcal{F}$), however it may be more than that, it could be measurable with respect to some sub-σ-algebra contained inside this one (in fact this is very often the case).

Definition 3.22. If X is a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subseteq \mathcal{F}$ is a σ-algebra (sub-σ-algebra) then X is called *$\mathcal{G}$-measurable* if X is measurable on $(\Omega, \mathcal{G})$, i.e. $X^{-1}(\mathcal{B}(\mathbb{R})) \subseteq \mathcal{G}$.

Note that by definition $\sigma(X)$ is the smallest σ-algebra such that X is $\sigma(X)$-measurable. Also, X is $\mathcal{G}$-measurable if and only if $\{X \le t\} \in \mathcal{G}$ for each $t \in \mathbb{R}$. A random variable Y is $\sigma(X)$-measurable if and only if $Y = f(X)$, for some measurable $f : \mathbb{R} \to \mathbb{R}$.

These concepts (related σ-algebras generated by random variables and sub-σ-algebras, the importance of measurable with respect to some σ-algebra etc.) will keep coming up throughout the course. They should become more familiar and comfortable as the course goes on. We introduce them early so that you can think about them, but not panic about them. With experience you should come to connect σ-algebras with "information" in a certain way that will become clearer with some examples, and also much more precise when we start to look at conditional expectations.

As we have seen previously, thinking about simple discrete examples (and making up your own) can be helpful to build intuition, although care must be taken that subtleties are missing in the discrete setting.

Example 3.23 (Geometric random walk)**.** Let $\Omega = \{H, T\}^3$ and $\mathcal{F} = \mathcal{P}(\Omega)$ (think about this example also if we took $\{H, T\}^{\mathbb{N}}$ as a sample space then the space is no-longer countable and there are some more subtleties). Fix $u > 1$ and $0 < d < 1$. Let $S_0 = 1$ and

$$S_k = \begin{cases} uS_{k-1} & \text{if } \omega_k = H\,, \\ dS_{k-1} & \text{if } \omega_k = T\,, \end{cases}$$

so $S_2 \in \{u^2, ud, d^2\}$. Then $\sigma(S_2)$ contains the sets:

- $A_{HH} = \{HHT, HHH\} = S_2^{-1}(\{u^2\}) = \{\omega : S_2(\omega) = u^2\} = \{S_2 = u^2\} = \{\omega_1 = H,\ \omega_2 = H\}$ where we have just list some fairly standard notation.
- $A_{TT} = \{TTT, TTH\} = \{S_2 = d^2\}$.
- $A_{TH} \cup A_{HT} = \{S_2 = ud\} = \{HTT, THT, HTH, THT\}$.
- All the complements and unions of these sets.
- $\emptyset$ and Ω.

Notice that this is not all of $\mathcal{F}$.

Suppose $(Z_i : i \in \mathcal{I})$ is a family of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. If someone gives you the observed values of the entire collection $(Z_i : i \in \mathcal{I})$, then you can say whether each event $F \in \sigma(Z_i : i \in \mathcal{I})$ occurred or not. Think about the example above, and suppose you are told e.g. $S_2(\omega) = u^2$, then A_{HH} occurred, A_{TT} did not etc.. Also, if you know exactly which of the events in $\sigma(Z_i : i \in \mathcal{I})$ have occurred and which haven't then you know the value of each of the random variables in $(Z_i : i \in \mathcal{I})$, without observing an entire outcome $\omega \in \Omega$. We will come back to this idea of σ-algebras as information again later.

The probabilities of possible random observations will be described in terms of the distribution of the corresponding random object X, which is the image measure of $\mathbb{P}$ under X.

Definition 3.24 (Distribution or Law of a random variable)**.** If X is a random object on $(\Omega, \mathcal{F}, \mathbb{P})$ into $(\Lambda, \mathcal{G})$, then the *distribution* or *law* of X is

$$\mathcal{L}_X = \mathbb{P} \circ X^{-1}\,.$$

Exercise 3.25. Show that $\mathcal{L}_X$ is a probability measure on $(\Lambda, \mathcal{G})$.

The most important special case for us will be when X is a random variable, i.e. a real valued Borel measurable function, written $X \in m\mathcal{F}$ (that is $(\Lambda, \mathcal{G}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Recall that probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (Borel probaility measures) are in 1-to-1 correspondence with distribution functions $F\colon \mathbb{R} \to [0, 1]$ (Theorem 2.28).

Definition 3.26 (Distribution function of random variable)**.** For a real valued random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$

$$F_X(x) = \mathbb{P}(X \le x) = \mathbb{P}\big(X^{-1}(-\infty, x]\big) = \mathcal{L}_X\big((-\infty, x]\big)$$

is called the *distribution function of* X.

Definition 3.27 (Discrete distribution)**.** A probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called *discrete* if there exists $C \subset \mathbb{R}$ countable such that $\mu(C) = 1$. Similarly a random variable is called discrete if there exists $C \subset \mathbb{R}$ countable such that $\mathcal{L}_X(C) = 1$.

Notice that in this case

$$F_\mu(x) = \mu\big((-\infty, x]\big) = \sum_{\substack{x_i \in C \\ x_i \leq x}} \mu(\{x_i\})\,, \text{ or}$$

$$F_X(x) = \sum_{\substack{x_i \in C \\ x_i \leq x}} \mathbb{P}(X = x_i)\,,$$

i.e. the distribution function is a step function with steps at each of the points x_i in C which are of "size" given by the probability of the point (mass function). This step function could be very complicated, for example jump at every point in $\mathbb{Q}$ (the distribution function is not necessarily simple in the sense we will define soon and you will have seen in a prerequisite measure theory module).

Suppose F_X is continuously differentiable, then it is the cumulative distribution function of some random variable with probability density function (p.d.f.) given by

$$f(x) = F_X'(x)\,.$$

Actually we can generalise this class somewhat (a similar result holds even if F is not differentiable on a set of (Lebesgue) measure zero).

Definition 3.28 (probability density function (p.d.f.) of an absolutely continuous r.v.)**.** A random variable X with law $\mathcal{L}_X$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is said to be absolutely continuous (with respect to the Lebsegue measure) if there exists an integrable function $f : \mathbb{R} \to [0, \infty)$ such that

$$F_X(x) = \int_{-\infty}^{x} f(t)dt\,,$$

and f is called the probability distribution function (p.d.f.) of X.

Note that F_X is by definition continuous in this case and $F_X'(x) = f(x)$ a.e. (i.e. except on a null set with respect to the Lebesgue measure). *However* F_X may be continuous but not come from a density function, for example the Devil's Stair Case (sometimes called the Cantor distribution) corresponding (roughly speaking) to the uniform measure on the Cantor set.

Definition 3.29 (Singular random variable)**.** A random variable with law $\mathcal{L}_X$ is called singular (with respect to the Lebesgue measure) if F_X is continuous but there exists an $A \in \mathcal{B}(\mathbb{R})$ such that $\mathcal{L}_X(A) = 1$ but the Lebesgue measure of A is zero.

Remark 3.30. A continuous distribution function can be either singular or absolutely continuous.

Remark 3.31. It turns out that any distribution function can be written as a (unique) convex combination of these three cases, i.e.

$$F = \alpha F_{\mathrm{d}} + \beta F_{\mathrm{s}} + \gamma F_{\mathrm{ac}} \quad \alpha, \beta, \gamma \geq 0, \quad \alpha + \beta + \gamma = 1\,.$$

Example 3.32 (Discrete random variables)**.** Let F be a right continuous, non-decreasing, step-function such that $F(-\infty) = 0$ and $F(\infty) = 1$, and that it has jumps at $C = \{x_1, x_2, \ldots\}$ each of size $m(x_i) > 0$, for $i \in \mathbb{N}$. Then F is a distribution function. It turns out C is then at most countable (you can check - there are at most n steps of size at least $1/n$). Then there exists a random variable X such that $\mathbb{P}(X \in C) = 1$ and $\mathbb{P}(X = x_i) = m(x_i)$.

As a quick practice you should draw/write down the distribution function of a Bernoulli p random variable.

Example 3.33 (Absolutely continuous)**.** Let $\mu \in \mathbb{R}$, $\sigma^2 > 0$, and X be a random variable with distribution function

$$\mathbb{P}(X \leq x) = F_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \mathrm{d}t \quad \text{for } x \in \mathbb{R}$$

then X is said to have normal distribution, or a Gaussian random variable.

In this case X is an absolutely continuous random variable (its law is absolutely continuous with respect to the Lebesgue measure) and its probability density function is given by

$$f_X(x) = F_X'(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Example 3.34 (Singular)**.** The "prototypical" example of a singular distribution is given, roughly speaking, by taking the uniform measure on the middle third Cantor set. The associated distribution function is sometimes called the *Devil's staircase* or the *Cantor function*. The distribution function is continuous, but there does not exist a probability density function associated with the distribution function. You may have seen this example before, more information is certainly available with some diligent googling.

Example 3.35 (Mixture)**.** Let $0 < a < b$ and $0 < q < r < 1$, and let F be a distribution function given by (draw a picture),

$$F(x) = \begin{cases} 0 & \text{if } x < 0\,, \\ q & \text{if } x \in [0, a)\,, \\ \frac{r-q}{b-a}x + \frac{qb-ar}{b-a} & \text{if } x \in [a, b)\,, \\ 1 & \text{if } b \leq x\,. \end{cases}$$

Then $F(x) = \alpha F_d(x) + (1-\alpha)F_{\mathrm{ac}}(x)$ where

$$F_d(x) = \begin{cases} 0 & \text{if } x < 0\,, \\ \frac{q}{1-(r-q)} & \text{if } x \in [0, b)\,, \\ 1 & \text{if } b \leq x\,. \end{cases} \qquad F_{\mathrm{ac}}(x) = \begin{cases} 0 & \text{if } x < a\,, \\ \frac{x-a}{b-a} & \text{if } x \in [a, b)\,, \\ 1 & \text{if } b \leq x\,. \end{cases}$$

and $\alpha = 1 - r + q$.

Chapter 4

Integration and Expectation

4.1 Revision of integration*

Reading: ST342/MA359
Further reading: D. Williams, Chapter 5 and A. Klenke, Chapter 4

In this section we revise some of background on the Lebesgue integral which you have will have seen in the prerequisite modules. *See Chapter 5 of the ST342 notes and MA359 Weeks 4-6.* This section will be purposefully brief, please consult the relevant chapters of the suggested text, or your material from prerequisite modules, if any of this seems unfamiliar.

Throughout this section $(\Omega, \mathcal{F}, \mu)$ will be a measure space. We want to define (where possible), the notation of the integral of a function $f\colon \Omega \to \overline{\mathbb{R}} = [-\infty, \infty]$. We will use various notation for the integral, all synonymous, i.e.

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

The dummy variable (x above) is sometimes useful, for example when integrating functions of several variables (see Fubini's theorem below).

Definition 4.1 (Simple functions)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. A function $\varphi\colon \Omega \to \mathbb{R}$ is called a *simple function* if there exists an $n \in \mathbb{N}$ and $E_1, \ldots, E_n \in \mathcal{F}$, $a_1, \ldots, a_n \in \mathbb{R}$ such that

$$\varphi = \sum_{k=1}^{n} a_k \mathbb{1}_{E_k} \,. \tag{4.1}$$

The (unique) canonical form of φ is the unique decomposition of this form where all the a_k are distinct and non-zero, and the sets E_k are disjoint and all non-empty.

That is, a simple function is a measurable step function taking only finitely many steps. You will certainly have seen the following definitions for the special case of the Lebesgue measure before, but everything you have done extends without significant change to a general measure space $(\Omega, \mathcal{F}, \mu)$.

Definition 4.2. If φ is a non-negative simple function with canonical form (4.1), then we define the integral of φ w.r.t. μ by

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

Definition 4.3. For a non-negative measurable function f (i.e. $f \in (m\mathcal{F})^+$) we define the integral by

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

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

Definition 4.4. We say that a measurable function f (i.e. $f \in m\mathcal{F}$) is *integrable* if $\int |f| \, \mathrm{d}\mu < \infty$ (note this is well defined since $|f| \in (m\mathcal{F})^+$). If f is integrable, then its integral is defined by

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

where $f^+ = f \vee 0$ and $f^- = -(f \wedge 0)$ are the positive and negative parts of f respectively.

The generalised integral defined in this way shares the same properties as the Lebesgue integral you will have seen. For example, it is linear, if f and g are both integrable, and $c \in \mathbb{R}$, then

$$\int f + cg \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu + c \int g \, \mathrm{d}\mu \, .$$

Definition 4.5 (Notation for integral over a subset)**.** If $A \in \mathcal{F}$ and f integrable, then we define

$$\int_A f \, \mathrm{d}\mu = \int f \, \mathbb{1}_A \, \mathrm{d}\mu = \int_\Omega f(x) \, \mathbb{1}_A(x) \, \mathrm{d}\mu(x) \, .$$

Example 4.6.

- If μ is the Lebesgue measure on $(\mathbb{R}, \mathcal{B})$ then we just redefined the Lebesgue integral.
- Suppose that μ is discrete (see Example 3.32), i.e. μ has a mass function that assigns mass p_i to point x_i for some countable collection $\{x_i\}_{i \in \mathcal{I}}$ of points in $\mathbb{R}$. Then you can check that

  $$\int f \, \mathrm{d}\mu = \sum_{i \in \mathcal{I}} f(x_i) \, p_i \, ,$$

  i.e. the integral generalises sums.
- Assume that μ is a probability measure on $(\mathbb{R}, \mathcal{B})$ with distribution function $F(x) = \int_{-\infty}^{x} f_\mu(y) \, \mathrm{d}y$, i.e. μ is absolutely continuous with respect to the Lebesgue measure and has a probability density function f_μ. Then

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

  We will return to these things when we look at expectations (the expectation of a random variable is just it's integral with respect to the associated probability measure).

There is general proof strategy when dealing with integrals, which can also be applied to showing rigorously that the statements in the previous example hold. In *D. Williams* (1991) it is referred to as '*the standard machine*'. It goes like this:

1. Prove the result holds for indicator functions of measurable sets $f = \mathbb{1}_E$ for $E \in \mathcal{F}$,
2. then use linearity to extend this to simple functions f,
3. next use monotone convergence (see below) to extend this to any non-negative f by approximation with simple functions,
4. finally for a generic $f \in m\mathcal{F}$ consider the positive and negative parts separately and use linearity again.

Definition 4.7. A property is said to hold μ-almost everywhere (abbreviated μ-a.e.) if it holds except on a set of μ measure zero. if μ is a probability measure we say *μ-almost surely*, abbreviated μ-a.s., or when the measure is clear from context simply a.s..

Remark 4.8. If $f = g$ a.e. (i.e. $\mu(\{\omega : f(\omega) \neq g(\omega)\}) = \mu(\{f \neq g\}) = 0$), then $\int f \,\mathrm{d}\mu = \int g \,\mathrm{d}\mu$.

I collect here a bunch of properties of the integral that you've probably seen before, at least for the Lebesgue integral, in last terms module. They all translate directly to the more general setting.

Theorem 4.9 (Properties of integral)**.** *Let f, g be integrable functions on a measure space $(\Omega, \mathcal{F}, \mu)$ and $\alpha \in \overline{\mathbb{R}}$.*

- *(Monotonicity) If $f \leq g$ μ-a.e., then $\int f \,\mathrm{d}\mu \leq \int g \,\mathrm{d}\mu$.*
- *(Triangle inequality) $|\int f \,\mathrm{d}\mu| \leq \int |f| \,\mathrm{d}\mu$.*
- *(Linearity) $\int (\alpha f + g) \,\mathrm{d}\mu = \alpha \int f \,\mathrm{d}\mu + \int g \,\mathrm{d}\mu$.*
- *(Unitary) $\int \mathbb{1}_A \,\mathrm{d}\mu = \mu(A)$.*

Proof. The proofs are part of your previous module or relatively straight forward. For more details see for example A. Klenke, Probability Theory. □

Note a little care has to be take with our notation $\mu(\cdot)$ has different meanings depending what type of object the argument is, e.g. $\mu(\mathbb{1}_A) = \mu(A)$ the left-hand side is short hand for the integral of a function whereas the right-hand side is simply the measure of a set.

Recall, $f_n \to f$ point wise if $f_n(\omega) \to f(\omega)$ for all $\omega \in \Omega$ and $f_n \to f$ a.e (or a.s. for prob measure) if $\mu(\{\omega \in \Omega : f_n(\omega) \to f(\omega)\}^c\} = 0$. We now state (without proof) the important convergence theorems (again see pre-requisite material).

Theorem 4.10 (Fatou's Lemma)**.** *Let $(f_n)_{n\geq 1}$ be a sequences of* non-negative *measurable functions, then*

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

(Note we are not excluding the case when the integrals are infinite).

Note that in words the result can be interpreted as 'some mass may escape to infinity.'

Exercise 4.11. You should construct memorable examples of when the inequality in Fatou's Lemma is strict, and a non-trivial example when there is equality (this will help to ensure you never forget which way round the inequality goes). You should try this for a probability measure, for example Lebesgue measure on $[0,1]$, and for on measure that is not finite, e.g. Lebesgue on $[0,\infty)$. This can be down with very simple sequences of functions. If you're luck there might be some pictures in the video lectures.

Theorem 4.12 (Monotone Convergence Theorem (MCT))**.** *Let $(f_n)_{n\geq 1}$ be a sequences of* non-negative *measurable functions, such that for all $n \in \mathbb{N}$ we have $f_n \leq f_{n+1}$ and $f_n \to f$ a.s., then*

$$\int f_n \,\mathrm{d}\mu \to \int f \,\mathrm{d}\mu\,, \quad \textit{i.e.} \quad \lim_{n\to\infty} \int f_n \,\mathrm{d}\mu = \int \lim_{n\to\infty} f_n \,\mathrm{d}\mu\,.$$

(Note we are not excluding the case when the integrals are infinite).

In analogy with previous notation, if $f_n \leq f_{n+1}$ and $f_n \to f$ then we write $f_n \nearrow f$.

Note, as a corollary to the MCT we get the following result for series which is often useful. If $(f_n)_{n\geq 1}$ is a sequence of non-negative measurable functions, then

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

Theorem 4.13 (Dominated Convergence Theorem (DCT))**.** *Let $(f_n)_{n\geq 1}$ be a sequences of measurable functions such that $f_n \to f$ a.e., and suppose that there exists an* integrable *function g such that $|f_n| \leq g$ for all n (i.e. g dominates the sequence $|f_n|$). Then f is integrable and*

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

We will also use the following partial converse to Fatou's Lemma occasionally, in fact it is a Corollary of Fatou's Lemma.

Corollary 4.14 (Reverse Fatou's Lemma)**.** *Let $(f_n)_{n\geq 1}$ be a sequences of measurable functions, and suppose that there exists an* integrable *function g such that $f_n \leq g$ for all n. Then*

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

Proof. Apply Fatou's Lemma to $h_n = g - f_n$, so it is important that $f_n \leq g$ so that h_n is non-negative, and $\int g \,\mathrm{d}\mu < \infty$ so we can cancel the contribution from these terms. □

4.2 Integration and Expectation

Reading: D. Williams, Chapter 6 and A. Klenke, Chapter 5

We recall some notation and cover some extra, useful, results of integration (expectation) that we might not have mentioned so far. I will try to stick to capital Latin letters (e.g. X) for random variables, recall this is just a measurable functions in $m\mathcal{F}$ (i.e. functions which are $\mathcal{B},\mathcal{F}$-measurable). Occasionally they may be things like f.

Notation: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If X is a $\mathcal{F}$-measurable and $\int |X| \, \mathrm{d}\mathbb{P} < \infty$ then we say that X is integrable and write $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, more generally we say $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ for some $1 \leq p < \infty$ if $\int |X|^p \, \mathrm{d}\mathbb{P} < \infty$. You may well have seen $\mathcal{L}^p$ and L^p spaces before in a measure theory module. We will return to $\mathcal{L}^p$ and L^p spaces later in a little more detail.

Remark 4.15. The triangle inequality hods for integrals, in particular if μ is a measure and f a measurable function then $|\mu(f)| \leq \mu(|f|)$. For the special case of a probability measure and random variables this is equivalent to $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$. In the more general setting of L^p spaces the triangle inequality is often called the Minkoswki inequality.

Recall (from any previous probability modules), if X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, with associated law $\mathcal{L}_X$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ then we define the expected value of X with respect to $\mathbb{P}$ simply as the integral of X with respect to the measure $\mathbb{P}$,

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

You will have used this many times before, though you may never have thought about it in such abstract terms until now. This general set-up covers discrete random variables (sums), as well as (absolutely) continuous random variables, singular random variables and mixtures of all three. The final equality is stated precisely in Lemma 4.18 below, but again is something you have very likely used many times if you have ever actually calculated an expected value. We first state some elementary properties of the expectation.

Remark 4.16 (Notation). We introduce some notation for the expectation of a random variable *on* some event $A \in \mathcal{F}$

$$\mathbb{E}[X \, ; \, A] = \mathbb{E}[X \mathbb{1}_A] = \int_A X \, \mathrm{d}\mathbb{P} \, . \tag{4.2}$$

This is not to be confused with the conditional expectation, for which we also have to appropriately re-weight the measure. Again, if you have ever calculated a conditional expectation you have probably used the following equation before although the notation here may be new; if $\mathbb{P}(A) > 0$ then

$$\mathbb{E}[X \mid A] = \frac{\mathbb{E}[X \, ; \, A]}{\mathbb{P}(A)} \, .$$

We will return to the concept of conditional expectations in much more generality later (we will give a more abstract definition which is truelly a generalisation of this idea).

Lemma 4.17. *If X and Y are integrable functions on $(\Omega, \mathcal{F}, \mathbb{P})$, then*

1. $X \geq 0 \implies \mathbb{E}[X] \geq 0$.
2. $X \geq Y \implies \mathbb{E}[X] \geq \mathbb{E}[Y]$.
3. $X \geq 0$ *and* $\mathbb{P}(X > 0) > 0$ *then* $\mathbb{E}[X] > 0$.

Proof. See Exercise Sheet 5 (monotonicity was discussed more generally in the previous section, which covers 1. and 2.). □

Lemma 4.18. *Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with law $\mathcal{L}_X$, i.e.*

$$\mathcal{L}_X(B) = \mathbb{P}(X \in B), \quad B \in \mathcal{B}(\mathbb{R}).$$

Suppose $h : \mathbb{R} \to \mathbb{R}$ is a Borel measurable function. Then, $h(X)$ is integrable on $(\Omega, \mathcal{F}, \mathbb{P})$ if and only if h is integrable on $(\mathbb{R}, \mathcal{B}, \mathcal{L}_X)$ and

$$\mathbb{E}[h(X)] = \mathcal{L}_X(h)\,, \tag{4.3}$$

In particular, if it exists, we have

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \,\mathrm{d}\mathcal{L}_X(x)\,.$$

Remark 4.19. The previous lemma is actually a special case of the change of variables formula: If μ is a measure on Ω, f a measurable function $f\colon \Omega \to \Lambda$, and g a Borel measurable function $g\colon \Lambda \to \mathbb{R}$, then

$$\int g \circ f \,\mathrm{d}\mu = \int g \,\mathrm{d}(\mu \circ f^{-1})\,.$$

Remark 4.20. Before giving a proof, there is quite a lot of new notation hidden in (4.3), so it's worth expanding that equality and making sure you are happy with all the notation below. Note there is no new content in the equation given in this remark, just recalling the notation we have introduced so far, with one new but hopefully intuitively obvious piece of notation in the middle namely $\mathbb{P}(X \in \mathrm{d}x)$. In the setting of the Lemma, $X : \Omega \to \mathbb{R}$ is an $\mathcal{F}, \mathcal{B}(\mathbb{R})$-measurable function and $h \circ X : \Omega \to \mathbb{R}$ is as well (the composition of measurable function is measurable), the law of X is just a name for the image measure $\mathbb{P} \circ X^{-1}$ of $\mathbb{P}$ under X and (4.3) states

$$\mathbb{E}[h(X)] = \int h \circ X \,\mathrm{d}\mathbb{P} = \int h \,\mathrm{d}(\mathbb{P} \circ X^{-1}) = \int_{\mathbb{R}} h(x)\mathbb{P}(X \in \mathrm{d}x) = \int h \,\mathrm{d}\mathcal{L}_X = \mathcal{L}_X(h)\,,$$

Proof. The proof of the lemma follows the 'standard machine'. First we consider $h = \mathbb{1}_B$, for $B \in \mathcal{B}$, and by linearity of the integral, we can extend the result to simple functions. Then, approximate a generic element of $(m\mathcal{B})^+$ by simple functions and apply the MCT. For general $h \in m\mathcal{B}$, consider its positive and negative parts, h^+ and h^- and use linearity of the integral.

For the first step (which is most of the content of the proof), fix $B \in \mathcal{B}$ and consider $h = \mathbb{1}_B$, then

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} \mathbb{1}_B(X(\omega)) \,\mathrm{d}\mathbb{P}(\omega) = \int_{\mathbb{R}} \mathbb{1}_{\{X \in B\}} \,\mathrm{d}\mathbb{P} = \mathbb{P}(X \in B) = \int_B \mathcal{L}_X(\,\mathrm{d}x) = \mathcal{L}_X(\mathbb{1}_B).$$

[*More details will be contained in the video lecture.*] □

Remark 4.21. We have already discussed the consequences of this result in the special case of absolutely continuous, or purely discrete, law (distributions) $\mathcal{L}_X$ in Example 4.6.

Lemma 4.22. *If X is a non-negative random variable then*

$$\mathbb{E}[X] = 0 \quad \Longleftrightarrow \quad \mathbb{P}[X > 0] = 0\,.$$

Proof. "⇐": Assume $X = 0$ a.s., and let $N = \{X > 0\}$, then $X \leq \infty \mathbb{1}_N$ (here we are using extended real valued functions and you might have to take my word for it that this makes sense). By Property two of Lemma 4.17, and MCT, we infer

$$0 \leq \mathbb{E}[X] \leq \mathbb{E}[\infty \mathbb{1}_N] = \lim_{n\to\infty} \mathbb{E}[n\mathbb{1}_N] = \lim_{n\to\infty} n\mathbb{P}(N) = 0$$

"⇒": Consider the increasing sequence of events $A_n \in \mathcal{F}$ given by,

$$A_n = \left\{ X \geq \frac{1}{n} \right\}, \quad n \in \mathbb{N},$$

and observe that $\{X > 0\} = \bigcup_n A_n$ (i.e. $A_n \nearrow N$). Now, by Property two of Lemma 4.17 it follows that

$$0 = \mathbb{E}[X] \geq \mathbb{E}[X\mathbb{1}_{A_n}] \geq \mathbb{E}[(1/n)\mathbb{1}_{A_n}] = \frac{1}{n}\mathbb{P}(A_n),$$

i.e. $\mathbb{P}(A_n) = 0$ for each n, so N is a countable union of null sets, and hence null.
[*Note:* We will hear "countable union of null sets, and hence null" a lot! Please spend a few minutes now making sure you are really happy with why this is true and what it means.] □

Remark 4.23. I'm sure you all want to see even more notation at this point... if F is a distribution function of a random variable X then you might also see

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x)\, \mathrm{d}F(x).$$

4.3 Some Useful Inequalities

Reading: D. Williams, Chapter 6 and A. Klenke, Chapter 5

The following inequalities are used frequently and important to keep in mind when bounding probabilities (more generally expectations). You will have certainly seen Markov's inequality last term, in *Week 6 of MA359 and Chapter 5 of ST342*, and Chebyshev's inequality is an immediate corollary. Jensen's inequality was in *Chapter 5 of ST342* but maybe new to MA359 people, and finally the Cauchy-Schwarz inequality you will have presumably seen before in several modules and holds in general for inner product spaces (a little more on this later when we look at L^2) - it turns out to be extremely useful.

Lemma 4.24 (Markov's inequality)**.** *Let $X \in (m\mathcal{F})^+$ be a random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, for any $\lambda \geq 0$,*

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

Proof. First, notice that, $\forall \omega \in \Omega$,

$$X(\omega) \geq \lambda \mathbb{1}_{\{X \geq \lambda\}}(\omega) = \begin{cases} 0 & \text{if } X(\omega) < \lambda, \\ \lambda & \text{if } X(\omega) \geq \lambda. \end{cases}$$

Taking expectation on both sides yields,

$$\mathbb{E}[X] \geq \mathbb{E}[\lambda \mathbb{1}_{\{X \geq \lambda\}}] = \lambda \mathbb{P}(X \geq \lambda),$$

which complete the proof. □

Corollary 4.25 (General Chebyshev's inequality)**.** *Let X be a random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in some $A \in \mathcal{B}$, and $\phi : A \to [0, \infty)$ be an increasing Borel-measurable function. Then, for any $\lambda > 0$ with $\phi(\lambda) > 0$,*

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

In the special case of $\phi(x) = x^2$ and $Y = |X - \mathbb{E}[X]|$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \lambda) = \mathbb{P}(Y \geq \lambda) \leq \frac{\mathbb{E}[Y^2]}{\lambda^2} = \frac{\mathrm{Var}(X)}{\lambda^2},$$

Proof. Using the fact that ϕ is increasing,

$$\text{if } X \geq \lambda, \text{ then } \phi(X) \geq \phi(\lambda) \implies \mathbb{P}(X \geq \lambda) \leq \mathbb{P}\big(\phi(X) \geq \phi(\lambda)\big),$$

we now apply Markov's inequality to $\phi(X)$ which is non-negative by assumption on ϕ. More explicitly,

$$\mathbb{E}[\phi(X)] \geq \mathbb{E}[\mathbb{1}_{\{\phi(X) \geq \phi(\lambda)\}} \phi(X)] \geq \phi(\lambda) \mathbb{P}\big(\phi(X) \geq \phi(\lambda)\big) \geq \phi(\lambda) \mathbb{P}(X \geq \lambda).$$

□

Another fairly common application of the above version of Chebyshev's inequality is when $\phi(x) = e^{\theta x}$ for some $\theta \geq 0$. In this case we get

$$\mathbb{P}(X \geq \lambda) \leq e^{-\theta\lambda} \mathbb{E}(e^{\theta X}) .$$

This is a common 'trick' in Large Deviation Theory, the next step is often to optimize the inequality over the parameter θ.

As a direct consequence of Markov's inequality we get our first version (or prototype) Weak Law of Large Numbers.

Corollary 4.26 (Prototype Weak Law of Large Numbers)**.** *Let $X_1, ..., X_n$, $n \in \mathbb{N}$, be i.i.d. random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with mean $\mathbb{E}[X_i] = \mu$ and finite variance $Var(X_i) = \sigma^2 < \infty$, $i = 1, ..., n$. Set $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Then, $\forall \epsilon > 0$,*

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \xrightarrow{n \to \infty} 0.$$

Proof. First, note that the expectation and variance of $\bar{X}_n$ can simply be computed by,

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \mu,$$

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2} \mathrm{Var}(\sum_{i=1}^n X_i) = \frac{1}{n^2} n \mathrm{Var}(X_1) = \frac{\sigma^2}{n}.$$

Now, using Chebyshev's inequality in the special case of variance for $\bar{X}_n$, we get that,

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \leq \frac{1}{\epsilon^2} \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n\epsilon^2} \to 0 \quad \text{as } n \to \infty .$$

□

Jensen's inequality is another quite common inequality used in practice, but before stating it, let us recall the definition of a convex functions. I would strongly encourage people to draw a picture of a convex function and a chord, and think about the convex combination of two points as being a type of 'average' then (as long as you can remember what a convex function looks like) it is almost impossible to forget Jensen's inequality. If you have trouble remembering what a convex function looks like remember it means that the 'area' above the graph is convex (opposite of concave - and you can remember what a concave set looks like because it has the word 'cave' in it). The "area above the curve" is often called the *epigraph* of the function, and saying a function is convex is equivalent to saying that the epigraph is a convex subset of $\mathbb{R}^2$.

Definition 4.27 (Convex function)**.** Let $G \subset \mathbb{R}$ be an open interval. Then, a function $f : G \to \mathbb{R}$ is convex, if for any $x, y \in G$ and $p, q \in [0, 1]$ such that $p + q = 1$,

$$f(px + qy) \leq pf(x) + qf(y).$$

Remark 4.28. It can be shown that if $f : G \to \mathbb{R}$ is convex, then f is continuous on G (it is important here that G is open, otherwise it could be discontinuous at the end points). In addition, if f is twice differentiable, then f is convex, if and only if $f''(x) \geq 0$ for each $x \in G$.

Theorem 4.29 (Jensen's inequality)**.** *Let $G \subseteq R$ be an open interval, $f : G \to \mathbb{R}$ be convex and $X \in \mathcal{L}^1(\mathbb{P})$ be a G-valued random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then,*

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

Proof. (Non-examinable) Given $x \in G$, define the function of difference quotients,

$$g_x(y) := \frac{f(y) - f(x)}{y - x}, \quad \forall y \in G \setminus \{x\},$$

which is increasing, by convexity of f. It is now clear (check!) that f is continuous, and for each $x \in G$, the left-sided and right-sided derivatives exist, which are given by,

$$D^- f(x) := \lim_{y \uparrow x} g_x(y) = \sup\{g_x(y) : y < x\} \quad \text{and} \quad D^+ f(x) := \lim_{y \downarrow x} g_x(y) = \inf\{g_x(y) : y > x\},$$

where both are increasing and $D^- f(x) \leq D^+ f(x)$. Then, for every $x \in G$ and $t \in [D^- f(x), D^+ f(x)]$, we have,

$$f(y) \geq t(y - x) + f(x), \quad \forall y \in G \setminus \{x\}.$$

Now, for $y = X,\ x = \mathbb{E}[X] \in G$, take expectation on both sides to get the result, that is,

$$\mathbb{E}[f(X)] \geq \mathbb{E}[t(X - \mathbb{E}[X])] + \mathbb{E}[f(\mathbb{E}[X])] = f(\mathbb{E}[X]).$$

□

Theorem 4.30 (Cauchy-Schwarz Inequality)**.** *If $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$ then XY is integrable and*

$$|\mathbb{E}(XY)| \leq \mathbb{E}(|XY|) \leq \mathbb{E}(X^2)^{1/2}\mathbb{E}(Y^2)^{1/2}. \tag{4.4}$$

Proof. Probably seen in other modules for other inner product spaces - make sure you can translate the notation across. □

Chapter 5

Independence

Reading: D. Williams, Chapter 4 and A. Klenke, Chapter 2
Further reading: R. Durrett, Probability Theory and Examples, Section 2.1
A. Klenke, Chapter 14

This is the point that probability theory begins to diverge from something purely about measure theory. That is, the measure theory of previous chapters is a linear theory that can't (on it's own) describe the dependence structure of events or random variables. The concept of dependencies (and independence) plays a central rôle in probability theory. The idea of independence is intricately linked with the concept of product measures, and this is where we will start our discussion. Intuitively, two "things" are independent if they have no influence on each other. Knowing what happens to one of the things tells us nothing extra about what happens to the other.

We now give definitions that capture all the concepts of independence you have encountered before in probability in one consistent way, starting with constructing product measures.

5.1 Product spaces

See Week 9 of MA359 or Chapter 7 of ST342 for more details. If ou are interested in this topic and it's connections with probability there are much more details in *A. Klenke, Chapter 14.*

Definition 5.1 (Product σ-algebra)**.** Given two sets Ω_1 and Ω_2 , the *Cartesian product* is given by

$$\Omega = \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) \,:\, \omega_1 \in \Omega_1,\ \omega_2 \in \Omega_2\}\,.$$

If $\mathcal{F}_i$ is a σ-algebra on Ω_i for $i = 1, 2$, then a *measurable rectangle* in Ω is a set of the form $A_1 \times A_2$ with $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$. The *product σ-algebra* $\mathcal{F} = \mathcal{F}_1 * \mathcal{F}_2$ is the σ-algebra generated by measurable rectangles. (Note $\mathcal{F}$ is **not** the Cartesian product of $\mathcal{F}_1$ and $\mathcal{F}_2$).

More generally if $(\Omega_k, \mathcal{F}_k)$, $k = 1, \ldots, n$ are measurable spaces then

$$\mathcal{F}_1 * \mathcal{F}_2 * \ldots * \mathcal{F}_n = \sigma(\{A_1 \times A_2 \times \ldots \times A_n \,:\, A_i \in \mathcal{F}_i,\ i = 1, 2, \ldots, n\})\,.$$

For some intuition on the remark that the product σ-algebra is not the Cartesian product one can think about the Borel measurable structure on $\mathbb{R}^2$. Note that the open (Euclidean) ball is measurable but it can not be written as a measurable rectangle (i.e. an element of the Cartesian product of the two σ-algebras).

Alternatively the product σ-algebra can be defined as the smallest σ-algebra such that all the coordinate maps (or canonical projects) are measurable. This is very similar to the concept of the product topology, which is the coarsest topology with respect to which each of the coordinate maps are continuous.

Given probability spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ we would like to define a probability space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 * \mathcal{F}_2, \mathbb{P}_1 \otimes \mathbb{P}_2)$ by setting

$$\mathbb{P}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2) \quad \text{for each measurable rectangle} \quad A_1 \times A_2 \,. \tag{5.1}$$

Note that the set $\mathcal{I}$ of measurable rectangles is a π-system (the intersection of two rectangles is another rectangle - check!), and $\sigma(\mathcal{I}) = \mathcal{F}_1 * \mathcal{F}_2$ by definition. So by Dynkin's uniqueness lemma (Lemma 2.20) if $\mathbb{P}_1 \otimes \mathbb{P}_2$ exists it must be unique.

To construct $\mathbb{P}_1 \otimes \mathbb{P}_2$ using 5.1 we apply Carathéodory's Extension Theorem (Theorem 2.21). Let $\mathcal{A}$ be the algebra of all finite disjoint unions of measurable rectangles and define an additive set function by

$$\mathbb{P}(R_1 \cup R_2 \cup \ldots \cup R_n) = \sum_{i=1}^{n} \mathbb{P}(R_i) \tag{5.2}$$

for any finite disjoint union of rectangles, where each of the terms in the sum on the right hand side is given by Eq. 5.1. It is tedious (but nothing deep) to show that in fact this gives a well-defined and σ-additive set function.

Lemma 5.2. *The set function $\mathbb{P}$ defined on $\mathcal{A}$ above is σ-additive on $\mathcal{A}$.*

Proof. Follow the standard strategy discussed before Definition 2.3 (this is an exercises for the keen reader). □

Now, by Carathéodory's Extension Theorem 2.21, we know that we know that $\mathbb{P}$ extends uniquely to a probability measure on the product space. Also, it is important that all this machinery still works in the σ-finite setting (not just the setting of finite measures), so we can construct the product of standard Lebesgue measures on $\mathbb{R}$.

Definition 5.3. Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ be probability spaces, then $(\Omega_1 \times \Omega_2, \mathcal{F}_1 * \mathcal{F}_2, \mathbb{P}_1 \otimes \mathbb{P}_2)$ is called the *product probability space.*

The above definition extends immediately to finite product spaces. Note that it is not true in general that $\mathbb{P}_1 \times \mathbb{P}_2 = \mathbb{P}_2 \times \mathbb{P}_1$, for example if $\Omega_1 \neq \Omega_2$ then the two measures are not even defined on the same space. Actually the construction above can be extended (with some effort!) to countable product spaces. We summarise this briefly below, since it is important for example when discussing a countable sequence of independent random variables. For more details see *A. Klenke, Section 14.3* and *D. Willaims, Section 8.7.*

Definition 5.4. Let $(\Omega_i, \mathcal{F}_i)_{i \geq 1}$ be a countable sequence of measurable spaces. Then the product σ-algebra $\mathcal{F} = *_{i=1}^{\infty} \mathcal{F}_i$ on $\Omega = \otimes_{i=1}^{\infty} \Omega_i$ is the σ-algebra generated by all sets of the form $\prod_{i=1}^{n} A_i \times \prod_{i=n+1}^{\infty} \Omega_i$ where $A_i \in \mathcal{F}_i$, i.e. it is generated by all finite-dimensional measurable rectangles.

Proposition 5.5 (Countable product measure)**.** *Let $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)_{i \geq 1}$ be a countable sequence of probability spaces. Then there exists a measure $\mathbb{P}$ on $(\omega, \mathcal{F})$, called the* product measure*, which is uniquely determined by*

$$\mathbb{P}\left(\otimes_{i=1}^{n} A_i \times \otimes_{i=n+1}^{\infty} \Omega_i\right) = \prod_{k=1}^{n} \mathbb{P}_k(A_k)\,,$$

where $A_i \in \mathcal{F}_i$ for $i = 1, \ldots, n$ and $n \in \mathbb{N}$.

Proof. (Non-examinable) The proof is beyond the scope of this course. One way is to apply Caraéodoty's Extension Theorem 2.21 directly as in the proof of the finite case, but the conditions required are tricky to verify. It also follows from a suitable limit of finite products using Kolmogorov Consistency Theorem (see A. Klenke). An alternative approach for Borel measures on $\mathbb{R}$ is outlined in D. Williams. □

Probably the product space which will be most familiar to you is the Lebesgue measure on $\mathbb{R}^2$ (with the Borel measurable sets), or more generally $\mathbb{R}^d$. You are probably used to calculating areas of measurable sets and integrating functions over $\mathbb{R}^d$. To actually do calculations it is convenient to be able to proceed in states, in the following sense. Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is measurable (and integrable) with respect to the Lebesgue measure on $\mathbb{R}^2$ then

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

This result (Fubini's Theorem) applies also to more general product measures.

Theorem 5.6 (Fubini's and Tonelli's Theorems)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P}) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 * \mathcal{F}_2, \mathbb{P}_1 \otimes \mathbb{P}_2)$ be a product probability space, and let $f(\omega) = f(\omega_1, \omega_2)$ be a measurable function on $(\Omega, \mathcal{F})$, then the functions*

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

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

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

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

where in case (ii) values may be ∞.

We give some language around product spaces and random variables (objects) and define random vectors (recall Definition 3.16).

Lemma 5.7 (Vectors of random objects)**.** *Let $(\Omega, \mathcal{F})$ and $(\Lambda_i, \mathcal{G}_i)$ for $i = 1, \ldots, n$ all be measurable spaces. The maps $X_i \colon \Omega \to \Lambda_i$, for $i = 1, \ldots, n$, are all random objects if and only if the vector*

$$Z = (X_1, X_2, \ldots, X_n) \colon \Omega \to \Lambda_1 \times \Lambda_2 \times \ldots \times \Lambda_n$$

is a random object in $(\Lambda_1 \times \Lambda_2 \times \ldots \times \Lambda_n, \mathcal{G}_1 * \mathcal{G}_2 * \ldots * \mathcal{G}_n)$

Proof. First suppose that Z is a random object. For each i define $\Pi_i \colon \Lambda_1 \times \Lambda_2 \times \ldots \times \Lambda_n \to \Lambda_i$ to be the canonical projection, so $X_i = \Pi_i \circ Z$. Then, by the composition lemma, it is sufficient to show that the projection is a random object (i.e. a measurable map), this is left as an exercises.

Now, suppose that each of the X_i's is $\mathcal{F}$,$\mathcal{G}_i$-measurable. We must show that Z is $\mathcal{F}$,$\mathcal{G}_1 * \ldots * \mathcal{G}_n$-measurable. By Proposition 3.6, it is sufficient to check that the preimage

of any measurable rectangle is measurable. Fix $A = A_1 \times A_2 \times \ldots \times A_n \in \mathcal{G}_1 * \mathcal{G}_2 * \ldots * \mathcal{G}_n$, then

$$\begin{aligned} Z^{-1}(A) &= \{\omega \in \Omega : Z(\omega) \in A\} \\ &= \{\omega : X_1(\omega) \in A_1, X_2(\omega) \in A_2, \ldots X_n(\omega) \in A_n\} \\ &= \{\omega : \omega \in X_1^{-1}(A_1), \omega \in X_2^{-1}(A_2), \ldots \omega \in X_n^{-1}(A_n)\} \\ &= X_1^{-1}(A_1) \cap X_2^{-1}(A_2) \cap \ldots \cap X_n^{-1}(A_n)\,. \end{aligned}$$

Since X_i's is $\mathcal{F}$,$\mathcal{G}_i$-measurable it follows that $X_i^{-1}(A_i) \in \mathcal{F}$ for each i, and hence $Z^{-1}(A) \in \mathcal{F}$ as required. □

In the setting of the previous lemma the distribution (law) of the random object Z is called the *joint distribution* of $X_1, X_2, \ldots, X_n$.

Definition 5.8 (Joint distribution)**.** Suppose the maps $X_i \colon \Omega \to \Lambda_i$, for $i = 1, \ldots, n$ are all random objects, then the distribution (law) of the random object $Z = (X_1, X_2, \ldots, X_n)$ is called the *joint distribution* of $X_1, X_2, \ldots, X_n$.

Given the marginal distributions of each of the X_i's above, there is still a lot of freedom in the joint distribution of the random object Z above. This joint distribution captures the 'dependencies' between the X_i's. If the distribution of Z is given by a product measure defined by the product of the marginals, then the X_i's are independent under the joint distribution. That is product measures give us an explicit way to construct independent random variables. We now give a more precises discussion of the idea of independence.

5.2 General Definition of Independence

Independence is a characteristic property of probability theory. We will see now how the measure theoretic definition corresponds with the notions of independence you will have seen before, and how it is intricately linked with product measures. The rule of thumb should be familiar, "independence means we can multiply", i.e. $\mathbb{P}(\bigcap \cdot) = \prod \mathbb{P}(\cdot)$, and $\mathbb{E}(\prod \cdot) = \prod \mathbb{E}(\cdot)$.

Recall from first year probability modules, that formally we say two events $A, B \in \mathcal{F}$ are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. More generally you will have defined independence for any finite collection of measurable sets $(A_i)_{i \in I}$ by saying they are independent if for any finite subset $J \subset I$

$$\mathbb{P}\Big(\bigcap_{j \in J} A_j\Big) = \prod_{j \in J} \mathbb{P}(A_j)\,.$$

You have also probably seen the following consequence of this definition (although perhaps not stated like this). If I is some index set and $(A_i)_{i \in I}$ are independent, and $B_i^0 = A_i$, $B_i^1 = A_i^c$ for each $i \in I$, then for any $\alpha \in \{0,1\}^I$ the family $(B_i^{\alpha_i})_{i \in I}$ is also independent (i.e. arbitrary combinations of the sets A_i and their complements are also independent).

We now generalise the concept of independence to classes of events (this is genuinely a generalisation of the definitions you have seen before).

Definition 5.9. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, then sub-σ-algebras $\mathcal{G}_1, \mathcal{G}_2, \ldots \subset \mathcal{F}$ are called *independent* if whenever $A_j \in \mathcal{G}_j$ and $i_1, i_2, \ldots, i_n$ are distinct then

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_n}) = \prod_{k=1}^{n} \mathbb{P}(A_{i_k})\,.$$

Remark 5.10. Note that to meet the definition we only need the condition to hold for finite subsets, but then it must also hold for countable collections by applying Monotone Convergence for measures Lemma 2.9. As a consequence of the definition it also holds if we take complements of some or all of the A_j's.

Definition 5.11 (Independent random variables)**.** The random variables $X_1, X_2, \ldots$ are called *independent* if $\sigma(X_1), \sigma(X_2), \ldots$ are independent σ-algebras.

In more familiar language we see that X and Y are independent if for each pair of Borel sets $A, B \in \mathcal{B}(\mathbb{R})$ we have

$$\mathbb{P}(X \in A,\, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)\,,$$

since $X^{-1}(A) \in \sigma(X)$ and $Y^{-1}(B) \in \sigma(Y)$, recall $\mathbb{P}(Y \in B) = \mathbb{P}(Y^{-1}(B))$.

Definition 5.12 (Independent events)**.** The events $E_1, E_2, \ldots$ are called *independent* if the σ-algebras $\mathcal{E}_1, \mathcal{E}_2, \ldots$, are independent, where $\mathcal{E}_i = \sigma(\{E_i\}) = \{\emptyset, \Omega, E_i, E_i^c\}$ for $i = 1, 2, \ldots$.

Note that events $E_1, E_2, \ldots$ are independent if and only if the random variables $\mathbb{1}_{E_1}, \mathbb{1}_{E_2}, \ldots$ are independent. Also, the more familiar definition of independence of events $E_1, E_2, \ldots$, i.e. that for any $n \in \mathbb{N}$ and whenever $i_1, i_2, \ldots, i_n$ are distinct then

$$\mathbb{P}(E_{i_1} \cap E_{i_2} \cap \ldots \cap E_{i_n}) = \prod_{k=1}^{n} \mathbb{P}(E_{i_k})\,,$$

is an immediate consequence of the definition above. Furthermore they are actually equivalent (Check. You will have probably done the necessary steps, such as showing that this condition implies that it holds also under complements, as an exercises before).

Of course the conditions in the definitions above appear impossible to check since σ-algebras in general don't have nice explicit presentations. However the Dynkin's Uniqueness Lemma for π-systems (Lemma 2.20) allows us to reduce it to something much more manageable.

Lemma 5.13. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, suppose $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ are sub-σ-algebras, and $\mathcal{G}_0, \mathcal{H}_0$ are π-systems generating them, i.e. $\sigma(\mathcal{G}_0) = \mathcal{G}$ and $\sigma(\mathcal{H}_0) = \mathcal{H}$. Then $\mathcal{G}$ and $\mathcal{H}$ are independent if and only if $\mathcal{G}_0$ and $\mathcal{H}_0$ are independent, i.e.*

$$\mathbb{P}(G \cap H) = \mathbb{P}(G)\mathbb{P}(H) \quad \textit{whenever} \;\; G \in \mathcal{G}_0,\; H \in \mathcal{H}_0\,.$$

Proof. If $\mathcal{G}$ and $\mathcal{H}$ are independent then it follows directly from the definition (and inclusion) that $\mathcal{G}_0$ and $\mathcal{H}_0$ are. Suppose that $\mathcal{G}_0$ and $\mathcal{H}_0$ are independent. Fix $G \in \mathcal{G}_0$, then define two maps as follows, for each $H \in \mathcal{H}$

$$H \mapsto \mathbb{P}(G \cap H) \quad H \mapsto \mathbb{P}(G)\mathbb{P}(H)\,.$$

The two maps define measure on $(\Omega, \mathcal{H})$ (you should check this). Note these are not necessarily probability measure since $\mathbb{P}(G)$ may be less than one, but they are finite, with total mass $\mathbb{P}(G)$. By assumption (since $\mathcal{G}_0$ and $\mathcal{H}_0$ are independent) these two measures agree on the π-system $\mathcal{H}_0$. It follows from Dynkin's uniqueness lemma (Lemma 2.20) that they must agree on $\sigma(\mathcal{H}_0) = \mathcal{H}$, hence for $G \in \mathcal{G}_0$ and $H \in \mathcal{H}$

$$\mathbb{P}(G \cap H) = \mathbb{P}(G)\mathbb{P}(H)\,.$$

To complete the proof apply the same argument for fixed $H \in \mathcal{H}$ to the maps $G \mapsto \mathbb{P}(G \cap H)$ and $G \mapsto \mathbb{P}(G)\mathbb{P}(H)$. □

Let X and Y be two random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that for each $x, y \in \mathbb{R}$

$$\mathbb{P}(X \leq x,\, Y \leq y) = \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y)\,,$$

then it follows that the π-systems $\big\{\{X \leq x\} : x \in \mathbb{R}\big\}$ and $\big\{\{Y \leq y\} : y \in \mathbb{R}\big\}$ are independent, so by the previous lemma X and Y are.

Corollary 5.14. *A sequence $(X_n)_{n\geq 1}$ of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ are independent if and only if for each $n \in \mathbb{N}$, and $x_1, x_2, \ldots, x_n \in \overline{\mathbb{R}}$*

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x_i)\,.$$

We can now make our previous discussion that connected independence with product measures more precise. Recall, if X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ then the image measure $\mathcal{L}_X = \mathbb{P} \circ X^{-1}$ is a Borel probability measure on $\mathbb{R}$ (see Definition 3.24). That is for each $B \in \mathcal{B}(\mathbb{R})$ we define $\mathcal{L}_X(B) = \mathbb{P}\big(X^{-1}(B)\big)$, called the law (or distribution) of X, and $F_X(x) = \mathcal{L}_X(-\infty, x]$ is the distribution function of X.

Lemma 5.15. *Let X and Y be random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Then (X, Y) is a random vector, and X and Y are independent if and only if the joint distribution is a product measure, i.e.*

$$\mathcal{L}_{(X,Y)} = \mathcal{L}_X \otimes \mathcal{L}_Y\,.$$

Product spaces, therefore, allow us to construct independent random variables, with specified marginals, on a single probability space.

Example 5.16. Suppose X is a random variable on $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$ and Y on $(\Omega, \mathcal{F}_Y, \mathbb{P}_Y)$, with distribution functions F_X and F_Y respectively. Let

$$(\Omega, \mathcal{F}, \mathbb{P}) = (\Omega_X \times \Omega_Y, \mathcal{F}_X * \mathcal{F}_Y, \mathbb{P}_X \times \mathbb{P}_Y)\,,$$

and for $\omega = (\omega_x, \omega_y) \in \Omega$ let

$$\tilde{X}(\omega) = X(\omega_x) \quad \tilde{Y}(\omega) = Y(\omega_y)\,,$$

then $\tilde{X}, \tilde{Y}$ are independent on $(\Omega, \mathcal{F}, \mathbb{P})$, and have the specified marginals, i.e.

$$\mathcal{L}_{\tilde{X}} = \mathcal{L}_X \quad \text{and} \quad \mathcal{L}_{\tilde{Y}} = \mathcal{L}_Y\,,$$

and $\mathcal{L}_{(\tilde{X},\tilde{Y})} = \mathcal{L}_X \otimes \mathcal{L}_Y$.

5.3 Tail σ-algebras and 0-1 laws

We now turn to some of the most beautiful results in probability concerning 'tail events' and the rôle of independence. First we have to define what we mean by tail events, we do this in the most general setting of σ-algebras, but often we have in mind a sequence of random variables $X_1, X_2, \ldots$ and events that only depend on the tail of the sequence.

Definition 5.17 (Tail σ-algebra)**.** For a sequence $(\mathcal{F}_n)_{n\geq 1}$ of σ-algebras define

$$\mathcal{T}_n = \sigma\left(\bigcup_{k\geq n} \mathcal{F}_k\right) = \sigma(\mathcal{F}_n, \mathcal{F}_{n+1}, \ldots)\,,$$

and

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

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

As a special case of the definition, above which we will use often, suppose $(X_n)_{n\geq 1}$ is a sequence of random variables, then $\mathcal{T} = \bigcap_n \sigma(X_n, X_{n+1}, \ldots)$ is called the *tail σ-algebra* of the sequence $(X_n)_{n\geq 1}$.

$\mathcal{T}$ captures the idea of the "indefinite future". If we think of $\mathcal{F}_n$ as being the collection of events which are determined by the n^{th} experiment in some series of experiments, then $\mathcal{T}$ is the ensemble of events determined by the 'tail' run of the experiments, arbitrarily far in the future. Roughly speaking $\mathcal{T}$ contains the events which are determined by the sequence $(X_n)_{n\geq 1}$, but changing finitely many of the values does not affect if the event holds or not. It turns out, although this sounds very restrictive so you might guess that $\mathcal{T}$ is just $\{\emptyset, \Omega\}$, actually this is not true, many interesting events belong to the tail σ-algebra.

Example 5.18. Let $(X_n)_{n\geq 1}$ be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. The tail σ-algebra of $\mathcal{F}_n = \sigma(X_n)$ contains, for example,

- $F_1 = \{\lim_{n\to\infty} X_n \text{ exists } \} = \{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \text{ exists } \}$,
- $F_2 = \{\sum_k X_k \text{ converges } \}$,
- $F_3 = \left\{ \lim \frac{1}{n} \sum_{k=1}^{n} X_k \text{ exists } \right\}$,

also $\xi = \limsup_n \frac{1}{n} \sum_{k=1}^{n} X_k$ is a $\mathcal{T}$-measurable random variable. We cover the first one, F_1, and the last point, $\xi \in m\mathcal{T}$, in some detail, the others follow a similar argument and are left as exercises.

Why is $F_1 \in \mathcal{T}$? Whether the limit $\lim_{n\to\infty} X_n$ exists or not is determined entirely by X_n,X_{n+1}, X_{n+2},... for each n, hence $F_1 \in \mathcal{T}_n$ for each n, and hence $F_1 \in \mathcal{T}$. More precisely, let $Y_n = \sup_{m\geq n} X_m$, then following the proof of Lemma 3.12 we have Y_n is $\mathcal{T}_n$-measurable. Hence $\limsup_n X_n$ is $\mathcal{T}_n$-measurable for each $n \in \mathbb{N}$. Now, following the last part of the proof of Lemma 3.12, we have that $\{\omega : \lim_n X_n(\omega) \text{ exists } \} \in \mathcal{T}_n$ for each $n \in \mathbb{N}$, and hence it is contained in $\mathcal{T}$ as required.

Why is $\xi \in m\mathcal{T}$? Fix $N \in \mathbb{N}$ and observe that

$$\xi = \limsup_n \frac{1}{n}\sum_{k=1}^{n} X_k = \limsup_n \frac{1}{n}\sum_{k=N}^{n} X_k\,,$$

since $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{N-1} X_k = 0$. Hence, by Lemma 3.10 followed by Lemma 3.12, $\xi \in m\mathcal{T}_N$ for each $N \in \mathbb{N}$, it follows that ξ must be measurable with respect to $\mathcal{T}$.

As an example of an event which is not in the tail σ-algebra, $\{\sum_{n=1}^{\infty} X_n = 0\} \notin \mathcal{T}$ if X_1 is random, since it depends on the value of X_1.

The following very elegant result demonstrates that sometimes probability questions have surprisingly simple answers. The next result very useful (as we will see later in the notes), there are very many nice applications of this, not least to *percolation* which you may have seen if you took MA3H2: Markov Processes and Percolation Theory last term.

Theorem 5.19 (Kolmogorov's zero-one law)**.** *Let $(\mathcal{F}_n)_{n\geq 1}$ be a sequence of* independent *σ-algebras on probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then the tail σ-algebra $\mathcal{T}$ is* trivial*, meaning*

1. *$F \in \mathcal{T}$ implies that $\mathbb{P}(F) \in \{0, 1\}$, and*
2. *any $\mathcal{T}$-measurable random variable is almost surely constant, i.e. $\xi \in m\mathcal{T}$ implies there exists $c \in \overline{\mathbb{R}} = [-\infty, \infty]$ such that $\mathbb{P}(\xi \equiv c) = 1$.*

Proof. (An alternative proof, using the Approximation Theorem for measures (Theorem 2.31), which we have not really discussed much but is never the less worth looking into if you are interested, can be found in *A. Klenke, Probability Theory, Theorem 2.37*).

For part 1. note that, if we can show that $\mathcal{T}$ is independent of $\mathcal{T}$ (independent of itself), then we are done by definition, since for $A \in \mathcal{T}$ we would have

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2\,,$$

for which the only solutions are 0 or 1. This will be our goal.

Fix $n \in \mathbb{N}$, let $\mathcal{H}_n = \sigma(\mathcal{F}_1, \ldots, \mathcal{F}_n)$, and recall $\mathcal{T}_n = \sigma\big(\bigcup_{k\geq n} \mathcal{F}_k\big)$.
Claim 1: $\mathcal{H}_n$ is independent of $\mathcal{T}_{n+1}$.
We now prove this first Claim 1. Let

$$\mathcal{H}_n^0 = \Big\{ \bigcap_{j=1}^{n} A_j \,:\, A_j \in \mathcal{F}_j \text{ for } j \in \{1, \ldots, n\} \Big\}\,,$$

then $\mathcal{H}_n^0$ is a π-system (check!) and $\sigma(\mathcal{H}_n^0) = \mathcal{H}_n$ (to check this last fact observe that $\bigcup_{j=1}^{n} \mathcal{F}_j \subseteq \mathcal{H}_n^0$ and if $A_j \in \mathcal{F}_j$ for each j then $\bigcap_{j=1}^{n} A_j \in \sigma\big(\bigcup_{j=1}^{n} \mathcal{F}_j\big)$). Similarly, let $I_n = \{n+1, n+2, \ldots\}$ and

$$\mathcal{T}_{n+1}^0 = \Big\{ \bigcap_{j\in I_n} A_j \,:\, A_j \in \mathcal{F}_j \text{ for } j \in I_n \text{ and } |\{j \,:\, A_j \neq \Omega\}| < \infty \Big\}\,,$$

is a π-system that generates $\mathcal{T}_{n+1}$, by the same argument as for $\mathcal{H}_n^0$ (the extra condition in the specification of $\mathcal{T}_{n+1}^0$ ensures we only have finite intersections in the π-system). Since $(\mathcal{F}_n)_{n\geq 1}$ are independent, it follows immediately that $\mathcal{H}_n^0$ and $\mathcal{T}_{n+1}^0$ are independent, and hence (by Lemma 5.13) that $\mathcal{H}_n$ and $\mathcal{T}_{n+1}$ are independent.

Claim 2: $\mathcal{H}_n$ is independent of $\mathcal{T}$, by Claim 1, since $\mathcal{T} \subseteq \mathcal{T}_{n+1}$.
Claim 3: $\mathcal{H}_\infty = \bigcup_{n\geq 1} \mathcal{H}_n$ is a π-system.
This follows from the fact that $\mathcal{H}_n \subseteq \mathcal{H}_{n+1}$ (check), so if you pick two elements in the union above then they both belong to $\mathcal{H}_m$ for some m. **Caution:** it is not in general true that a union of π-systems is a π-system, you should find some simple counter examples, you might find an example that also shows a union of σ-algebras is not necessarily a σ-algebra.

Now, putting together the three claims 2 and 3, we have $\sigma(\mathcal{H}_\infty) = \sigma(\mathcal{F}_1, \mathcal{F}_2, \ldots)$ and $\mathcal{H}_\infty$ is independent of $\mathcal{T}$, hence $\sigma(\mathcal{H}_\infty)$ is independent of $\mathcal{T}$. Finally observe that $\mathcal{T} \subseteq \sigma(\mathcal{F}_1, \mathcal{F}_2, \ldots) = \sigma(\mathcal{H}_\infty)$, and hence $\mathcal{T}$ is independent of $\mathcal{T}$ as required.

Proof of 2.: Suppose Y is $\mathcal{T}$-measurable and $\mathbb{R}$-valued, then $Y^{-1}\big((-\infty, x]\big) \in \mathcal{T}$, and so by Part 1 we have $F_Y(x) = \mathbb{P}(Y \leq x) \in \{0, 1\}$ for each $x \in \mathbb{R}$. Let $c = \inf\{y \in \mathbb{R} : F_Y(y) = 1\}$, where we take $\inf \mathbb{R} = -\infty$ and $\inf \emptyset = \infty$. If $c = -\infty$ then it is clear that $\mathbb{P}(Y = -\infty) = 1$ and if $c = \infty$ it is clear that $\mathbb{P}(Y = \infty) = 1$. Since F_Y is (weakly) increasing and right-continuous we have $\mathbb{P}(Y = c) = 1$. $\square$

5.4 Borel-Cantelli Lemmas

This was covered (possibly briefly) in Chapter 8 of ST342 but maybe new to people who did MA359.

It can be easy to apply Kolmogorov's 0-1 law, but it is often more tricky to know when the probability is zero or one. Some important tools in this regard, but also in their own right, are the Borel-Cantelli Lemmas.

We start by recalling some notation for limits of sets. Drawing some pictures might be helpful to to get your head around the following definitions the first time.

Definition 5.20 (lim sup/lim inf of events)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(A_n)_n$ be a sequence in $\mathcal{F}$. Then:

$$\begin{aligned}\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m\geq n} A_m &= \{\omega \in \Omega : \omega \in A_m \text{ for infinitely many } m\} \\ &= \{A_m \text{ occurs infinitely often.}\} = \{A_m \ i.o.\}\,. \\ \liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{m\geq n} A_m &= \{\omega \in \Omega : \exists\, m_0(\omega) \in \mathbb{N} \text{ s.t. } \omega \in A_m \text{ for all } m \geq m_0(\omega)\} \\ &= \{A_m \text{ keeps occurring eventually.}\} = \{A_m \ ev.\}\,.\end{aligned}$$

Recall, in terms of the notation used in MCT for measures (Lemma 2.9), we have $\bigcup_{m\geq n} A_m \searrow \limsup A_n$ and $\bigcap_{m\geq n} A_m \nearrow \liminf A_n$ as $n \to \infty$.

Remark 5.21. It is relatively straightforward to check that $\{A_m \ ev.\}^c = \{A_m^c \ i.o.\}$. Also note that $\{A_m \ i.o.\}$ (or ev.) can be considered as tail events with respect to the sequence of σ-algebras given by $\mathcal{F}_n = \sigma(A_n) = \{\emptyset, \Omega, A_n, A_n^c\}$, since $\{A_m \ i.o.\} = \bigcap_n G_n$ where $G_n = \bigcup_{m\geq n} A_m \in \sigma(A_n, A_{n+1}, \ldots) = \mathcal{T}_n$.

Lemma 5.22.

$$\mathbb{1}_{\limsup A_n} = \limsup_{n\to\infty} \mathbb{1}_{A_n}\,, \quad \textit{and} \quad \mathbb{1}_{\liminf A_n} = \liminf_{n\to\infty} \mathbb{1}_{A_n}\,.$$

Proof. Observe

$$\mathbb{1}_{\cup_n E_n}(\omega) = \left\{\begin{matrix} 1 & \text{if } \omega \in \cup_n E_n \\ 0 & \text{otherwise.}\end{matrix}\right\} = \sup_n \mathbb{1}_{E_n}(\omega)\,,$$

and similarly we have $\mathbb{1}_{\cap_n E_n} = \inf_n \mathbb{1}_{E_n}$. The proof of the first part of the lemma follows by applying the two equality's one after the other and noting that $\limsup_{n\to\infty} \mathbb{1}_{A_n} = \inf_{n\geq 0} \sup_{m\geq n} \mathbb{1}_{A_m}$ by definition of the limit supremum. The second part of the lemma follows analogously. □

Corollary 5.23.

$$\begin{aligned}\mathbb{P}(A_n \ ev.) = \mathbb{P}(\liminf A_n) &\leq \liminf_{n\to\infty} \mathbb{P}(A_n)\,, \\ \mathbb{P}(A_n \ i.o.) &\geq \limsup_{n\to\infty} \mathbb{P}(A_n)\,.\end{aligned}$$

Proof. Apply Fatou's lemma to $\displaystyle\int \liminf_{n\to\infty} \mathbb{1}_{A_n}\, \mathrm{d}\mathbb{P}$, and for the second line take complements (see Sheet 4). □

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

Lemma 5.24 (Borel-Cantelli Lemmas)**.** *Let $(A_n)_{n\geq 1}$ be a sequence of events on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then,*

(BC1) If $\sum_n \mathbb{P}(A_n) < \infty$ then $\mathbb{P}(\limsup A_n) = \mathbb{P}(A_n \text{ i.o.}) = 0$,

(BC2) If $(A_n)_{n\geq 1}$ are also independent *then $\sum_n \mathbb{P}(A_n) = \infty$ implies $\mathbb{P}(\limsup A_n) = \mathbb{P}(A_n \text{ i.o.}) = 1$ (i.e. the converse of (BC1) holds under independence).*

Notice (BC1) makes no assumption about independence, it is a very powerful result.

Proof. For (BC1) we assume that $\sum_n \mathbb{P}(A_n) < \infty$. Let $G_n = \bigcup_{m\geq n} A_m$, which is a decreasing sequence of sets such that $G_n \searrow G$ where $G = \limsup_{n\to\infty} A_n$. Fix $k \in \mathbb{N}$ and observe

$$\mathbb{P}(\limsup_{n\to\infty} A_n) = \mathbb{P}\big(\bigcap_{n\geq 1} G_n\big) \leq \mathbb{P}(G_k) \leq \sum_{n\geq k} \mathbb{P}(A_n)\,,$$

where the final inequality holds by σ-subadditivity. By assumption the right hand side converges to zero as $k \to \infty$, which concludes the proof of (BC1).

For (BC2) we assume that $(A_n)_{n\geq 1}$ are *independent* and $\sum_n \mathbb{P}(A_n) = \infty$. The main idea of the proof is to take complements and use the standard bound $1 - x \leq e^{-x}$ for $x \in \mathbb{R}$. Note, by the remark above that $\mathbb{P}(A_n \text{ i.o.}) = 1$ if and only if $\mathbb{P}(A_n^c \text{ ev.}) = 0$.

Fix $m, r \in \mathbb{N}$, then by independence

$$\mathbb{P}\big(\bigcap_{m\leq n} A_n^c\big) \leq \mathbb{P}\big(\bigcap_{m\leq n\leq r} A_n^c\big) = \prod_{m\leq n\leq r} \mathbb{P}(A_n^c) = \prod_{m\leq n\leq r} \big(1 - \mathbb{P}(A_n)\big)\,,$$

(the first inequality holds since $\mathbb{P}$ is an increasing set function - here onwards we take such bounds for granted without further comment). Applying the bound $1 - x \leq e^{-x}$, we find

$$\mathbb{P}\big(\bigcap_{m\leq n} A_n^c\big) \leq e^{-\sum_{m\leq n\leq r} \mathbb{P}(A_n)} \to 0 \quad \text{as } r \to \infty\,,$$

where convergence follows by assumption on $\sum_n \mathbb{P}(A_n)$. It follows that, since $\{A_n^c \text{ ev.}\}$ is a countable union of null sets, that $\mathbb{P}(A_n \text{ i.o.}) = 1 - \mathbb{P}(A_n^c \text{ ev.}) = 1$, as required.

Note the proof could be simplified (i.e. we can remove the extra limit in r), by Remark 5.10 and careful use of extended real valued functions. □

Example 5.25 (Monkey at a typewriter)**.** Imagine we put a capuchin at a typewriter, that we happen to have dug out of the loft (the typewriter not the monkey). The monkey happens to be particularly entertained by pressing keys but equal preference for each key, i.e. it presses keys one at a time uniformly at random and independently of previous presses. Will this fascinated capuchin ever type ABRACADABRA at least once? If it happens to be an immortal capuchin, will it type ABRACADABRA infinitely often?

Solution Yes! for each $k \in \mathbb{N}$ let

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

By the model assumptions $\mathbb{P}(A_k) = (1/26)^{11}$ (assuming our typewriter has exactly 26 keys representing the alphabet, so that when the capuchin is playing with it, he pounds

on the alphabetical keys uniformly at random), and the A_k's are independent. Thus $\sum_{k\geq 1}\mathbb{P}(A_k)=\infty$, so by BC2 $\{A_k \text{ } i.o.\}$ with probability one, and $\{A_k \text{ } i.o.\} \subset \{\text{types ABRACADABRA infinitely often}\} \subset \{\text{types ABRACADABRA}\}$. How long do we typically wait before we first observe the monkey type ABRACADABRA? It turns out we can answer this once we have looked at martingales (in particular as a neat consequence of the Optional Stopping Theorem).

Example 5.26 (Random walks)**.** Consider three independent simple symmetric random walks on $\mathbb{Z}$ each started from 0, called $(X_n^{(1)})_{n\geq 1}, (X_n^{(2)})_{n\geq 1}, (X_n^{(3)})_{n\geq 1}$. Consider the event that all three random walks coincide at the origin at time $2n$ (the random walks are periodic and can only be at 0 at even times since they start from 0 and always takes a single step either 'up' or 'down'). Let $A_n = \{X_{2n}^{(1)} = X_{2n}^{(2)} = X_{2n}^{(3)} = 0\}$ for each $n \in \mathbb{N}$, then by independence $\mathbb{P}(A_n) = \mathbb{P}(X_{2n}^{(1)} = 0)^3$. To compute $\mathbb{P}(X_{2n}^{(1)} = 0)$ we follow an argument you have probably seen in first year, that is each fixed path of length $2n$ occurs with probability 2^{-2n} by construction, and a path starts from 0 and returns to 0 at time $2n$ if and only if it contains exactly n 'up' moves and n 'down' moves. Hence,

$$\mathbb{P}(A_n) = \mathbb{P}(X_{2n}^{(1)} = 0)^3 = \left(\binom{2n}{n} 2^{-2n}\right)^3 .$$

Applying Stirling's formula to the right hand side we find

$$\mathbb{P}(A_n) \leq C \cdot \frac{1}{n^{3/2}} ,$$

where $C = \pi^{-3/2}$ (the keen reader can check that the bound holds for $n \geq 1$ with this constant) and hence $\sum_n \mathbb{P}(A_n) < \infty$, therefore by BC1, $\mathbb{P}(A_n \text{ } i.o.) = 0$. The three walks will not simultaneously return to zero infinitely often (c.f. the simple random walk is transient in dimensions three and higher).

Example 5.27 (Independent exponential random variables)**.** Let $(X_n)_{n\geq 1}$ be a sequence of independent, mean 1, exponential random variables, i.e. $X_i \sim \text{Exp}(1)$ for each $i \in \mathbb{N}$ so

$$\mathbb{P}(X_n \leq x) = 1 - e^{-x} , \quad \forall\, x \geq 0 .$$

For which value of $\alpha > 0$ is $X_n > \alpha \log n$ i.o.? Further, if $L = \limsup \frac{X_n}{\log n}$ then can we show that $\mathbb{P}(L = 1) = 1$?

Answer: For $\alpha > 0$ we have $\mathbb{P}(X_n > \alpha \log n) = e^{-\alpha \log n} = n^{-\alpha}$ for each $n \in \mathbb{N}$. So by BC1 and BC2 (considering sums over these events) we have

$$\mathbb{P}(X_n > \alpha \log n \text{ } i.o.) = \begin{cases} 0 & \text{if } \alpha > 1 , \\ 1 & \text{if } \alpha \leq 1 . \end{cases} \tag{5.3}$$

For the second part, let $L_n = \sup_{m\geq n} \frac{X_m}{\log m}$ so $L = \lim_{n\to\infty} L_n$ (and (L_n) is an decreasing sequence). Then since L_n is $\sigma(X_n, X_{n+1}, \ldots)$-measurable we have L is $\mathcal{T}$-measurable, where $\mathcal{T}$ is the tail σ-algebra of $\sigma(X_1), \sigma(X_2), \ldots$ So Kolmogorov's 0-1 law (Theorem 5.19) implies that L is almost surely a constant.

To determine the value of L we can make careful application of Eq. (5.3). Observe[1] that $\{L \geq 1\} \supseteq \{L_n \geq 1 \text{ } i.o.\} \supseteq \{X_n \geq \log n \text{ } i.o.\}$ (you should spend a little time

[1] You should check that for a sequence of random variables $(Y_n)_{n\geq 1}$ we have $\{Y_n \geq a \text{ } i.o.\} \subseteq \{\limsup Y_n \geq a\}$. We apply this with $Y_n = L_n = \sup_{m\geq n} \frac{X_m}{\log m}$. Note that the opposite inclusion does not hold in general, but it in fact holds here since L_n is a decreasing sequence.

convincing yourself that these inclusions hold - caution; none of these hold in general the other way round). It now follows from Eq. (5.3) that $\mathbb{P}(L \geq 1) = 1$ (we gave a somewhat longer proof of this fact in lectures). Hence it remains to show that $\{L > 1\}$ is null. To this end, we will show that $\{L > 1\}$ is a countable union of null sets and hence null. Fix $k \in \mathbb{N}$, and observe

$$\left\{L > 1 + \frac{2}{k}\right\} \subseteq \left\{L \geq 1 + \frac{2}{k}\right\} \subseteq \bigcap_{m \geq 1} \left\{L_n > 1 + \frac{2}{k} - \frac{1}{m} \; i.o.\right\} \subseteq \left\{L_n > 1 + \frac{1}{k} \; i.o.\right\},$$

where the last set inclusion followed by considering the case $m = k$. By definition $\left\{L_n > 1 + \frac{1}{k} \; i.o.\right\} = \left\{X_n > \left(1 + \frac{1}{k}\right) \log n \; i.o.\right\}$, and by Eq. (5.3) the later is a null event. The conclusion follows since

$$\{L > 1\} = \bigcup_k \left\{L > 1 + \frac{1}{k}\right\} = \bigcup_k \left\{L > 1 + \frac{2}{k}\right\},$$

and a countable union of null sets is null.

For yet another (neat) application of the BC lemmas see Exercises Sheet 6 Q6.3 on the law of iterated logarithms. Suppose $(X_n)_{n \geq 1}$ is a sequence of independent identically distributed random variables, with mean zero and variance one and define $S_n = \sum_{k=1}^n X_k$ for each $n \in \mathbb{N}$. Then by Kolmogorov's 0-1 law

$$\mathbb{P}\Big(\limsup_{n \to \infty} \frac{S_n}{\sqrt{2n \log \log n}} = 1\Big) \in \{0, 1\}.$$

In fact this event has probability one. This is called the law of iterated logarithms. Varadhan gave a proof of this result, under the slightly stronger moment assumption that there is some $\alpha > 0$ such that $\mathbb{E}(|X_n|^{2+\alpha}) < \infty$, by a delicate application of the the Borel-Cantelli lemmas. On the problem sheet you will prove something slightly weaker.

You might be a little suprised by the result above when you compare it to the central limit theorem that you will have learn before, which states that (under the same conditions)

$$\mathbb{P}\Big(\frac{1}{\sqrt{n}} S_n \leq a\Big) \xrightarrow{n \to \infty} \Phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, \mathrm{d}x \,.$$

These two results say very different things about about the behaviours of S_n for large n, the statements use different notions of convergence on different scales. We will study and compare many different modes of convergence in the next chapter. It turns out that the law of iterated logarithms is about almost sure convergence, whereas the CLT is about convergence in distribution.

Chapter 6

Modes of Convergence

Reading: A. Klenke, Chapter 4-7.
Modes of convergence; D. Williams, Chapter A13.

6.1 Defining Modes of Convergence

There are many reasons that we might be interested in different types of convergence of random variables. Firstly, it is often the case that different modes of convergence are easier or harder to study depending on the situation we are in. It is then important to understand how the different modes relate to each other. Also, as we will see in some examples, different modes of convergence can tell us very different things about the 'limits' of sequences of random variables.

Given a general measure space $(\Omega, \mathcal{F}, \mu)$, the main types of convergence for a sequence of measurable functions $(f_n)_{n\geq 1}$ are convergence in measure (μ), almost everywhere (a.e.) and in L^p, which you should have seen before in a Measure Theory type module. In the setting of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $(X_n)_{n\geq 1}$ a sequence of ($\mathbb{R}$-valued) random variables, we define the analogue modes of convergence, see some new notions and study the relationships between them.

I have checked that the following three definitions (excluding convergence in distribution) were discussed in some detail in the prerequisites, in MA359 Week 7 and 8, and ST342 Chapter 8. We will come back to convergence in distribution later when we look at convergence of probability measures under weak convergence.

Definition 6.1 (Converge almost surely ($\mathbb{P}$-a.s.))**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X_n)_{n\in\mathbb{N}}$ be random variables. We say that $(X_n)_{n\in\mathbb{N}}$ converges to X $\mathbb{P}$-a.s., if and only if

$$\mathbb{P}(X_n \xrightarrow{n\to\infty} X) = \mathbb{P}(\{\omega \in \Omega : X_n(\omega) \xrightarrow{n\to\infty} X(\omega)\}) = 1.$$

Then, we write $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} X$ or $X_n \to X$ $\mathbb{P}$-a.s. (when the probability measure is clear from context we will drop it to reduce notation).

Observe that almost sure convergence is just a special case of convergence almost everywhere (*a.e.*) when the measure has total mass one.

Remark 6.2. Note that, almost sure limits are unique up to almost sure equality. If $(X_n)_{n\in\mathbb{N}}$, X and Y are random variables on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} X$ and $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} Y$, then it must be that $X = Y$ $\mathbb{P}$-a.s.

Our prototype weak law of large numbers (Corollary 4.26) is a motivating application of convergence in probability. It states that under suitable conditions, the empirical mean of the given sequence of random variables converges in probability to the theoretical mean. Convergence in probability is just the name we give to convergence in measure when the measure in question is a probability a measure.

Definition 6.3 (Convergence in probability)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X_n)_{n\geq 1}$ a sequence of random variables. We say that $(X_n)_{n\geq 1}$ converges to X in probability if $\forall \epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = \lim_{n\to\infty} \mathbb{P}(\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon) = 0.$$

Then, we write $X_n \xrightarrow{\mathbb{P}} X$ or $X_n \to X$ in *prob.*

You have probably seen the following definition before, we recall it here for completeness

Definition 6.4 (Convergence in distribution)**.** A sequence of random variables $(X_n)_{n\geq 1}$, with respective distribution functions $(F_n)_{n\geq 1}$, is said to converge *in distribution* to a random variable X, with distribution function F if

$$F_n(x) \to F(x) \quad \text{for every point of continuity of } F.$$

It turns out that we can generalise the concept of convergence in to probability measures. This more general concept is called *weak convergence*. We will prove later that this is really a generalisation.

Definition 6.5 (Weak Convergence)**.** Let $(\mu_n)_{n\in\mathbb{N}}, \mu$ be probability measures in $\mathcal{M}_1(\mathbb{R})$ (the set of probability measures on $\mathbb{R}$). We say that $(\mu)_{n\in\mathbb{N}}$ converges to μ weakly, if

$$\mu_n(f) = \int_{\mathbb{R}} f \, \mathrm{d}\mu_n \xrightarrow{n\to\infty} \int_{\mathbb{R}} f \, \mathrm{d}\mu = \mu(f), \quad \forall f \in C_b(\mathbb{R}).$$

Then, we write $\mu_n \xrightarrow{w} \mu$.

Definition 6.6 (Convergence in $\mathcal{L}^p$ (p-norm) or p^{th}-moment)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(X_n)_{n\in\mathbb{N}}$ a sequence of random variables, and X a random variables with $X_n, X \in \mathcal{L}^p(\mathbb{P})$ for some $p \in [1, \infty)$ (that is $\mathbb{E}[|X|^p], \mathbb{E}[|X_n|^p] < \infty$). We say that $(X_n)_{n\geq 1}$ converges to X in $\mathcal{L}^p$ (or in p^{th}-moment) if

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = \lim_{n\to\infty} \int_\Omega |X_n - X|^p \, \mathrm{d}\mathbb{P} = 0.$$

Then, we write $X_n \xrightarrow{\mathcal{L}^p} X$.

The above result can be more neatly expressed in terms of the $\mathcal{L}^p$ semi-norm. You may have seen this concept before in a measure theory module. For $X \in \mathcal{L}^p$ we define $\|X\|_p = \mathbb{E}(|X|^p)^{\frac{1}{p}}$ (obviously this is a special case of a more general definition of $\mathcal{L}^p$ spaces for more general measures - not necessarily probability measures, by replacing the expectation by the integral).

Example 6.7 (Convergence almost surely does not imply convergence in $\mathcal{L}^p$)**.** Let $\Omega = [0,1]$, and $\mathcal{F} = \mathcal{B}([0,1])$ and let $\mathbb{P}$ be the standard Lebesgue measure on $[0,1]$. Consider the sequence of random variables

$$X_n(\omega) = \begin{cases} n(1-n\omega) & \text{if } \omega \in [0, 1/n]\,, \\ 0 & \text{otherwise.} \end{cases}$$

it is worth drawing a picture here (the pictures were drawn in lectures so you can checkout lecture capture). Then for each $\omega \in (0,1]$ we have $X_n(\omega) \to 0$ so $X_n \to 0$ a.s., however $\mathbb{E}[X_n] = 1/2$ for each $n \in \mathbb{N}$ and so X_n does not converge to 0 in $\mathcal{L}^1$.

In this case the random variable takes very large values with small probability. Convergence almost surely only identifies the behaviour outside events with vanishing small probability. Whereas convergence in $\mathcal{L}^1$ also depends on the size of the random variable itself, even on small events. Roughly speaking the p in $\mathcal{L}^p$ controls the relative weighting given to the 'size of the difference' $|X_n - X|$ to the probability of events. As $p \to \infty$ we recover the usual 'sup-norm' (and convergence).

Example 6.8 (Convergence in probability does not imply convergence almost surely)**.** For each $n \in \mathbb{N}$ there exists a unique pair $(m,k) \in \mathbb{N}_0$ such that $n = 2^m + k$ with $0 \le k < 2^m$. Let

$$X_n(\omega) = \mathbb{1}_{\left[\frac{k}{2^m}, \frac{k+1}{2^m}\right]}(\omega)\,,$$

i.e. we can picture a moving 'blip' (again it is best to draw a picture - which will be on lecture capture anyway). The blip moves 'from left to right' and each time it crosses $[0,1]$ it halves in size. For $\omega \in (0,1)$ we have $X_n(\omega) = 1$ infinitely often, that is for each $N_0 \in \mathbb{N}$ there exists $N > N_0$ such that $X_N(\omega) = 1$. Hence X_n does not converge to 0 almost surely. However $\mathbb{P}(X_n \neq 0) = \frac{1}{2^m} \to 0$ as $n \to \infty$, i.e. $X_n \xrightarrow{\mathbb{P}} 0$. On the other hand $X_{2^n} \to 0$ *a.s.*, it turns out this observation holds more generally, see Theorem 6.11 below. Also $\mathbb{E}[|X_n|] = \frac{1}{2^m} \to 0$ so $X_n \xrightarrow{\mathcal{L}^1} 0$.

6.2 Relationships between Modes of Convergence

Reading: D. Williams, Chapter 6 and A. Klenke, Chapter 6

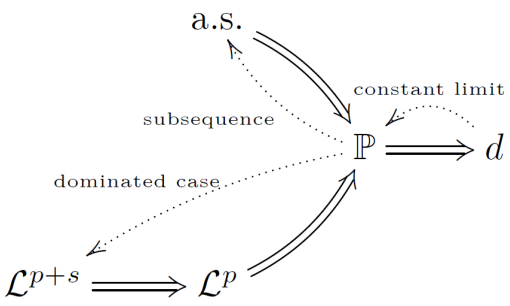

Figure 6.1: Summary of the relationship between modes of convergence.

The first relationship is useful for translating almost sure convergence into something that looks like convergence in probability.

Lemma 6.9. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X_n)_{n\geq 1}, X$ be random variables. Then the following are equivalent (TFAE):*

(a) $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} X$.

(b) $\forall \epsilon > 0,\ \lim_{n\to\infty} \mathbb{P}(\sup_{m\geq n} |X_m - X| > \epsilon) = 0$.

(c) $\forall \epsilon > 0,\ \mathbb{P}(|X_n - X| > \epsilon \ i.o.) = 0$.

Proof. See Exercise Sheet 5. □

The following lemma states that convergence in probability fast enough implies convergence $\mathbb{P}$-a.s.

Lemma 6.10. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X_n)_{n\in\mathbb{N}}$ be random variables such that $X_n \xrightarrow{\mathbb{P}} X$. If for each $\varepsilon > 0$ we have*

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \epsilon) < \infty, \tag{6.1}$$

then $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} X$.

Proof. This is an immediate consequence of (BC1). Assuming that (6.1) holds, we can apply (BC1) to get,

$$\mathbb{P}(|X_n - X| > \epsilon \ i.o.) = 0.$$

Thus, by the previous lemma, it follows that, $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} X$. □

At this point, your intuition may tell you that almost sure convergence is a stronger notion than convergence in probability. Lemma 6.9 (c) states that if $X_n \to X$ *a.s.* then for $\mathbb{P}$-almost-every ω we have $X_n(\omega)$ is ϵ close to $X(\omega)$ eventually (i.e. this event always occurs for sufficiently large n), for each $\varepsilon > 0$. On the other hand, under convergence in probability, $X_n(\omega)$ might be away from $X(\omega)$ (by at least ϵ) for ω in a set A_n with vanishing, but positive, probability (as $n \to \infty$), so 'occurrence eventually' can break down. However, convergence in probability does imply that there exists some subsequence $(X_{n_k})_{k\in\mathbb{N}}$ for which convergence almost surely is guaranteed. This is made rigorous in the following theorem.

Theorem 6.11 (Convergence in $\mathbb{P}$ and $\mathbb{P}$-a.s.)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X_n)_{n\in\mathbb{N}}, X$ be random variables.*

1. $X_n \xrightarrow{\mathbb{P}\text{-}a.s.} X \implies X_n \xrightarrow{\mathbb{P}} X$.

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

Proof. 1. Assume $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} X$. Fix $\epsilon > 0$, for $n \in \mathbb{N}$ let

$$A_{n,\epsilon} := \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\} = \{|X_n - X| > \epsilon\}.$$

By Lemma 6.9 we have $\mathbb{P}(A_{n,\varepsilon}\ i.o.) = 0$. Applying inverse Fatou's Lemma to $\mathbb{1}_{A_{n,\varepsilon}}$, i.e. Fatous Lemma on $\mathbb{1}_{A^c_{n,\varepsilon}}$, and recalling that $\mathbb{1}_{\limsup A_{n,\varepsilon}} = \limsup \mathbb{1}_{A_{n,\varepsilon}}$, we have

$$0 = \mathbb{P}(A_{n,\varepsilon}\ i.o.) = \mathbb{P}(\limsup_{n\to\infty} A_{n,\varepsilon}) \geq \limsup_{n\to\infty} \mathbb{P}(A_{n,\varepsilon}) \geq 0\,.$$

Hence $\mathbb{P}(A_{n,\varepsilon}) \to 0$ as $n \to \infty$ as required.

2. Suppose $X_n \xrightarrow{\mathbb{P}} X$, then for each $k \in \mathbb{N}$,

$$\mathbb{P}\Big(|X_n - X| > \frac{1}{k}\Big) = \mathbb{P}(A_{n,1/k}) \to 0 \quad \text{as } n \to \infty.$$

In particular, for each $k \in \mathbb{N}$ there exists some $N_k \in \mathbb{N}$ such that,

$$\mathbb{P}(A_{n_k,1/k}) < \frac{1}{k^2}, \quad \text{for all } n_k \geq N_k.$$

Let $B_k = A_{n_k,1/k}$, then

$$\sum_{k=1}^{\infty} \mathbb{P}(B_k) \leq \sum_{k=1}^{\infty} \frac{1}{k^2} < \infty,$$

and so by (BC1) it follows that $\mathbb{P}(B_n\ i.o.) = 0$ and hence by Lemma 6.9 we have $X_{n_k} \xrightarrow{\mathbb{P}\text{-a.s.}} X$ as $k \to \infty$.

□

Lemma 6.12. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X_n)_{n\geq 1}, X$ be random variables. Then, $X_n \xrightarrow{\mathbb{P}} X$ if and only if every subsequence of $(X_n)_{n\geq 1}$ has a further subsequence that converges to X $\mathbb{P}$-a.s.*

Proof. The 'only if' part follows from the previous theorem by first taking a subsequence of (X_n) which clearly still converges to X in probability. For the other direction see Exercise Sheet 5. □

We now turn our attention to convergence in $\mathcal{L}^p$.

Lemma 6.13. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X_n)_{n\geq 1}, X$ be random variables. If $1 \leq p \leq r < \infty$ and $X \in \mathcal{L}^r(\mathbb{P})$ then $X \in \mathcal{L}^p(\mathbb{P})$, and $\|X\|_p \leq \|X\|_r$. In particular,*

$$X_n \xrightarrow{\mathcal{L}^r} X \implies X_n \xrightarrow{\mathcal{L}^p} X.$$

Proof. The proof relies on a truncation and Jensen's inequality, this is a simple but effective trick that is used quite often. Fix $1 \leq p \leq r < \infty$ and suppose $X \in \mathcal{L}^r(\mathbb{P})$. Consider the random variables $(Y_n)_{n\geq 1}$ given by,

$$Y_n = |\min\{X, n\}|^p = |X \wedge n|^p, \quad \text{for } \ n \in \mathbb{N}.$$

Then Y_n is bounded for each $n \in \mathbb{N}$, and so $\mathbb{E}[|Y_n^{r/p}|] < \infty$, i.e. $Y_n^{r/p} \in \mathcal{L}^1(\mathbb{P})$. We can apply Jensen's inequality to the function $f(x) = x^{r/p}$, which is convex on $(0, \infty)$, that is,

$$\mathbb{E}[Y_n]^{r/p} \leq \mathbb{E}[Y_n^{r/p}] = \mathbb{E}[|X \wedge n|^r] \leq \mathbb{E}[|X|^r] < \infty. \tag{6.2}$$

Applying MCT, $\mathbb{E}[Y_n] \to \mathbb{E}[|X|^p]$, so by the Ineq. (6.2) we have

$$\mathbb{E}[|X|^p]^{r/p} \leq \mathbb{E}[|X|^r] \quad \text{so} \quad ||X||_p \leq ||X||_r.$$

□

Convergence in $\mathcal{L}^p$ is also stronger than convergence in probability. This follows directly from Chebyshev's inequality.

Lemma 6.14 (Convergence in $\mathcal{L}^p$ implies convergence in probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X_n)_{n\geq 1}, X$ be random variables such that $X_n, X \in \mathcal{L}^p$, for some $p \geq 1$, and $X_n \xrightarrow{\mathcal{L}^p} X$. Then, $X_n \xrightarrow{\mathbb{P}} X$.*

Proof. Since $X_n \xrightarrow{\mathcal{L}^p} X$ for some $p \geq 1$,

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

By Chebyshev's inequality, applied with $\phi(x) = |x|^p$, it follows that for any $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) \leq \frac{1}{\epsilon^p}\mathbb{E}[|X_n - X|^p] \to 0 \quad \text{as} \quad n \to \infty\,,$$

as required. □

Remark 6.15. We drew a nice picture in lectures to summarise some of the results of this section, i.e. which modes of convergence are weaker in general (and and showed by way of examples that these were the only implications that held in general). This picture is worth keeping in mind even though it is not yet in these notes!

6.3 Quick discussion of $\mathcal{L}^p$ and L^p spaces

This may be a topic you have covered in more detail in other analysis/measure theory modules and we we will not dwell on it here, although I encourage you to build the bridges between these modules yourself.

Since, for $1 \leq p < \infty$ and $a, b \in [0, \infty)$ we have $(a+b)^p \leq 2^p(a \vee b)^p \leq 2^p(a^p + b^p)$ it follows that $\mathcal{L}^p$ is a vector space. It is not a normed-vector space under $\|\cdot\|_p$, since this only defines a semi-norm. That is

$$\|X\|_p = 0 \quad \text{if and only if} \quad X = 0 \ a.s.\,,$$

but there may be many function which are zero almost surely (i.e. the right hand side is not unique). The 'standard' solution to this is to define an equivalence relation

$$X \sim Y \quad \text{if and only if } X = Y \quad a.s.$$

and define L^p as '$\mathcal{L}^p$ quotiented out by this equivalence relation', that is the space of equivalence classes, i.e. $L^p = \mathcal{L}^p/\mathcal{N}$ where $\mathcal{N} = \{f \in \mathcal{L}^p \,:\, f = 0 \ a.s.\}$. Then L^p indeed becomes a normed vector space. The next result show that it is in fact a Banach space (a complete normed vector space).

Lemma 6.16 ($\mathcal{L}^p$ is complete)**.** *If $1 \leq p < \infty$ and $(X_n)_{n\geq 1}$ is sequence in $\mathcal{L}^p$, then there exists a random variable $X \in \mathcal{L}^p$ such that $X_n \xrightarrow{\mathcal{L}^p} X$ as $n \to \infty$ if and only if*

$$\mathbb{E}|X_n - X_m|^p \to 0 \quad as \quad n, m \to \infty\,,$$

i.e. $(X_n)_{n\geq 1}$ is a Cauchy sequence in $\mathcal{L}^p$.

Proof. (Non-examinable) See *D. Williams* (1991) Section 6.10. □

The space $\mathcal{L}^2$ deserves some special attention, since it is also an inner product space (that is L^2 is a Hilbert space). We define the inner product of $X, Y \in \mathcal{L}^2$ by $\langle X, Y\rangle = \mathbb{E}(XY)$ (you can check that this indeed forms a scalar product or bi-linear form). Hence the usual Cauchy-Schwarz inequality holds. If $X, Y \in \mathcal{L}^2$ then $XY \in \mathcal{L}^1$ and

$$|\mathbb{E}(XY)| \leq \mathbb{E}(|XY|) \leq \|X\|_2 \|Y\|_2\,, \tag{6.3}$$

(the first inequality is just the triangle inequality). In this context we may prove the Cauchy-Schwarz (C-S) inequality following the standard approach (examine the determinant of $\mathbb{E}[(X + \lambda Y)^2]$), combine with the truncation trick we have seen before and MCT. *We should add the C-S inequality* (6.3) *to our list of useful inequalities, it is used frequently and can be extremely useful.* Spotting how and when to apply it effectively is a matter of developing experience.

If, for $X \in \mathcal{L}^p$ we define $\widetilde{X} = X - \mathbb{E}(X)$, then $\mathrm{Cov}(X, Y) = \langle \widetilde{X}, \widetilde{Y}\rangle$ and $\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \langle \widetilde{X}, \widetilde{X}\rangle$. There is much more to be said on the topic of $\mathcal{L}^p$ and L^p spaces, and in particular on the geometry of $\mathcal{L}^2$ and L^2, see for example Sections 6.8-6.10 in *D. Williams* and Chapter 7 in *A. Klenke*. We will return in some small part to this topic later when we look at conditional expectations.

6.4 Weak Convergence

In this section we examine another form of convergence. You are probably already familiar with the concept of *convergence in distribution* for random variables. It turns out that this is actually a special case of a much more general framework of convergence of probability measures, that of *weak convergence*. This form of convergence is related to the idea of probabilities of events being close. The name *weak convergence* is a little unfortunate, since in analysis outside of probability theory, that name actually means something different, and the concept we define here is more closely related to weak-$\star$ convergence. What probabilists call *weak convergence* is often called *narrow convergence* in analysis outside of probability theory.

We will restrict our attention here largely to probability measures on the Borel measure space $(\mathbb{R}, \mathcal{B})$. Some of our results will rely on the fact that the underlying space is $\mathbb{R}$, with the Euclidean metric, in particular that this is a Polish space, i.e. induced by a complete, separable, metric space. However, we could take a general complete and separable metric space (S, d) and associated Borel measurable structure $(S, \mathcal{S})$ and most of what we say still holds. As mentioned above, since weak convergence generalises convergence in distribution the real reason for preferring the new formulation in this chapter is that it makes sense for general metric spaces, for example function spaces - this is exactly the setting for powerful results such as Donsker's theorem that say the law of scaled random walks coverage to that of Brownian motion (the Wiener measure). For a much more general treatment of weak convergence see 'Convergence of Probability Measures', by *Billingsley* or Chapter 3 in 'Markov Processes' by *Ethier and Kurtz*. The set of probability measures on this space are denoted $\mathcal{M}_1(\mathbb{R})$ (the subscript indicates that the total mass is 1). The set of bounded and continuous function from $\mathbb{R}$ to $\mathbb{R}$ is denoted $C_b(\mathbb{R})$.

Definition 6.17 (Weak Convergence)**.** Let $(\mu_n)_{n\in\mathbb{N}}, \mu$ be probability measures in $\mathcal{M}_1(\mathbb{R})$. We say that $(\mu)_{n\in\mathbb{N}}$ converges to μ weakly, if

$$\mu_n(f) = \int_{\mathbb{R}} f \, \mathrm{d}\mu_n \xrightarrow{n\to\infty} \int_{\mathbb{R}} f \, \mathrm{d}\mu = \mu(f), \quad \forall f \in C_b(\mathbb{R}).$$

Then, we write $\mu_n \xrightarrow{w} \mu$.

Note that weak convergence is defined in terms of the expectation (integral) of bounded continuous functions $f \in C_b(\mathbb{R})$. We have already discussed that any continuous function is measurable, so the integral 'makes sense'. Further, since the measures are finite (in particular they are probability measures so $\mu(\mathbb{R}) = 1$), and the functions are bounded, the integrals involved are all finite.

Remark 6.18. The weak limit is unique (this relies on the fact that $\mathbb{R}$ is Polish).

We know from from previous results that there is one-to-one correspondence between probability measures in $\mathcal{M}_1(\mathbb{R})$ and distribution functions (recall from Theorem 2.28 that the distribution function associated with measure μ is given by $F(x) = \mu((-\infty, x])$, for each $x \in \mathbb{R}$). This gives rise to an entirely natural definition of weak convergence of distribution functions.

Definition 6.19 (Weak convergence of distribution functions)**.** Let $(\mu_n)_{n\geq 1}, \mu$ be probability measures in $\mathcal{M}_1(\mathbb{R})$, and $(F_n)_{n\geq 1}, F$ be the associated distribution functions on $\mathbb{R}$, respectively. We say that $(F_n)_{n\geq 1}$ converges to F weakly, if $\mu_n \xrightarrow{w} \mu$. Then, we write $F_n \xrightarrow{w} F$.

We would like to use the concept of weak convergence to define an associated convergence of random variables, we do this in the natural way as follows. Recall that the law (or distribution) of a real valued random variable X on probability space $(\Omega, \mathcal{F}, \mathbb{P})$, denoted by $\mathcal{L}_X$, is an element of $\mathcal{M}_1(\mathbb{R})$ given by the image measure $\mathbb{P} \circ X^{-1}$ (i.e. $\mathbb{P} \circ X^{-1}(B) = \mathbb{P}(X \in B)$). As a reminder we summarise some of the notation used so far. If F_n is the distribution function of a random variable X_n, with associated law $\mu_n = \mathcal{L}_{X_n}$ then

$$\mu_n(h) = \int_{\mathbb{R}} h(x)\,\mathrm{d}F_n(x) = \int_{\mathbb{R}} h(x)\,\mathrm{d}\mathcal{L}_{X_n}(x) = \mathbb{E}[h(X_n)]\,.$$

Definition 6.20 (Weak convergence of random variables)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X_n)_{n\geq 1}, X$ be random variables, with $(F_{X_n})_{n\geq 1}, F$ their corresponding distribution functions, respectively. We say that $(X_n)_{n\geq 1}$ converges to X weakly if $F_{X_n} \xrightarrow{w} F$ (i.e. their respective laws converge weakly). Then, we write $X_n \xrightarrow{w} X$.

Remark 6.21. weak convergence of random variables makes sense even if the random variables $X_1, X_2, \ldots, X$ are all defined on different probability spaces. Their law (distribution) is always a Borel measure on $\mathbb{R}$. This is not the case for convergence almost surely or convergence in measure, which both rely on the random variables all being defined on the same probability space. We return to this point later in Theorem 6.28.

Recall Definition 6.4 of Convergence in distribution for random variables. It turns out that weak convergence of random variables (on $\mathbb{R}$) corresponds exactly to convergence in distribution. Although weak convergence generalise well, convergence in distribution does not.

Theorem 6.22 (Weak convergence and convergence in distribution)**.** *Let $(F_n)_{n\geq 1}$ be a sequence of distribution functions on $\mathbb{R}$, and F a distribution function on $\mathbb{R}$. Let μ_n, μ be the corresponding probability measures. Then $F_n \xrightarrow{d} F$, if and only if $\mu_n \xrightarrow{w} \mu$.*

Proof. Suppose $\mu_n \xrightarrow{w} \mu$, then for every $f \in C_b(\mathbb{R})$ we have $\mu_n(f) \to \mu(f)$ as $n \to \infty$. So to show convergence in distribution we will approximate the indicator functions, who's expectation gives the distribution function, by a continuous function and take limits. Notice $F_n(x) = \mu_n(-\infty, x] = \mu_n\big(\mathbb{1}_{(-\infty,x]}\big)$ (again it is a little unfortunate that our notation $\mu(\cdot)$ has different meanings depending on whether the argument is a measurable set or a measurable function - but you should try to get used to this). Pictures can be helpful for this proof, see the lectures.

Fix $x \in \mathbb{R}$ a continuity point of F, and $\delta > 0$. Define $h_x \in C_b(\mathbb{R})$ by

$$h_x(y) = \begin{cases} 1 & \text{if } y \leq x\,, \\ 1 - \frac{y-x}{\delta} & \text{if } y \in (x, x+\delta)\,, \\ 0 & \text{if } y \geq x + \delta\,. \end{cases}$$

Then by assumption $\mu_n(h_x) \to \mu(h_x)$ as $n \to \infty$, and by construction of h_x we have

$$F_n(x) \leq \mu_n(h_x) \quad \text{and} \quad \mu(h_x) \leq F(x+\delta)\,,$$

which implies that

$$\limsup_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} \mu_n(h_x) = \mu(h_x) \leq F(x+\delta)\,.$$

Now take the limit $\delta \to 0$ and use continuity of F at x to get $\limsup_{n\to\infty} F_n(x) \le F(x)$.

We use the same trick to get the desired upper bound on $F(x)$. That is, define $g_x \in C_b(\mathbb{R})$ by

$$g_x(y) = \begin{cases} 1 & \text{if } y \le x - \delta\,, \\ 1 - \frac{y-(x-\delta)}{\delta} & \text{if } y \in (x-\delta, x)\,, \\ 0 & \text{if } y \ge x\,. \end{cases}$$

Then by the same argument,

$$\liminf_{n\to\infty} F_n(x) \ge \liminf_{n\to\infty} \mu_n(g_x) = \mu(g_x) \ge F(x-\delta)\,.$$

Now take the limit $\delta \to 0$ and use continuity of F at x to get $\liminf_{n\to\infty} F_n(x) \ge F(x)$. It follows that $F_n \xrightarrow{d} F$.

The reverse implication (in fact both ways) is a consequence of the more general *Portmanteau Theorem*, which is Theorem 6.31 below. □

Our new characterisation of convergence in distribution from weak convergence can sometimes be simpler to check than convergence of the distribution function.

Theorem 6.23 (Continuous mapping theorem)**.** *Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is continuous and $X_n \xrightarrow{w} X$ then $\phi(X_n) \xrightarrow{w} \phi(X)$.*

Proof. Suppose $f\colon \mathbb{R} \to \mathbb{R}$ is bounded and continuous, then $h = f \circ \phi$ is also bounded and continuous. Therefore

$$\mathbb{E}[f(\phi(X_n))] = \mathbb{E}[h(X_n)] \to \mathbb{E}[h(X)] = \mathbb{E}[f(\phi(X))]\,.$$

□

In fact the continuous mapping theorem can be extend to the case when ϕ has discontinuous, so long as ϕ is continuous almost everywhere.

Lemma 6.24. *Let $(X_n)_{n\geq 1}$ be a sequence of random variables and X a random variable, all on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then;*

(i) $X_n \to X$ in probability implies $X_n \xrightarrow{w} X$.

(ii) $X_n \to X$ a.s. implies $X_n \xrightarrow{w} X$.

Proof. (ii) follows from (i) and the fact that $X_n \to X$ a.s. implies $X_n \to X$ in probability (see Theorem 6.11).

Suppose $X_n \xrightarrow{\mathbb{P}} X$, and let F_n and F be the distribution functions of X_n and X respectively. Fix $\varepsilon > 0$ and x a continuity point of F, we need to show that $F_n(x) \to F(x)$ as $n \to \infty$. To this end we apply the law of total probability

$$\begin{aligned} F_n(x) = \mathbb{P}(X_n \leq x) &= \mathbb{P}(X_n \leq x\,;\, X \leq x + \varepsilon) + \mathbb{P}(X_n \leq x\,;\, X > x + \varepsilon) \\ &\leq F(x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon)\,, \end{aligned}$$

since on the event $\{X_n \leq x\} \cap \{X > x + \varepsilon\}$ we have $X > x + \varepsilon \geq X_n + \varepsilon$ so $|X_n - X| > \varepsilon$. Taking $\varepsilon \to 0$, and using convergence in probability, we have $\limsup_{n\to\infty} F_n(x) \leq F(x)$.

We prove the opposite bounded similarly. That is,

$$\begin{aligned} F(x - \varepsilon) = \mathbb{P}(X \leq x - \varepsilon) &= \mathbb{P}(X \leq x - \varepsilon\,;\, X_n \leq x) + \mathbb{P}(X \leq x - \varepsilon\,;\, X_n > x) \\ &\leq F_n(x) + \mathbb{P}(|X_n - X| > \varepsilon)\,, \end{aligned}$$

and applying convergence in probability again, we have $\liminf F_n(x) \geq F(x - \varepsilon) \to F(x)$ as $\varepsilon \to 0$. □

Exercise 6.25. Prove (ii) above using BCT (Theorem 6.35).

The converse to (i) does not hold in general, unless the limiting random variable is (almost surely) constant.

Lemma 6.26. *Suppose $(X_n)_{n\geq 1}$ is a sequence of random variables all on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $a \in \mathbb{R}$. If $X_n \xrightarrow{w} a$ (i.e. in distribution) then $X_n \xrightarrow{\mathbb{P}} a$.*

Proof. The proof is left as an exercises on the problem sheets. □

Example 6.27 ($X_n \xrightarrow{w} X$ does not imply $X_n \xrightarrow{\mathbb{P}} X$)**.** Suppose $X \sim \text{Ber}\left(\frac{1}{2}\right)$ and $X_n = X$ for each $n \in \mathbb{N}$. Then $X_n \xrightarrow{w} X$, in fact they are all equal in distribution. Let $Y = 1 - X$, then Y and X have the same law (distribution function) by symmetry, so $X_n \xrightarrow{w} Y$ (again they are all equal in distribution), *however* $|X_n(\omega) - Y(\omega)| = 1$ for each $\omega \in \Omega$ and so they do not convergence almost surely or in probability. The issues is that convergence in probability and almost surely depend on the probability space's that support the random variables, but weak convergence of the their laws (convergence in distribution) does not care.

The following theorem provides a partial converse.

Theorem 6.28 (Skorokhod Representation Theorem)**.** *Let $(F_n)_{n\geq 1}$ be a sequence of distribution functions on $\mathbb{R}$, and F a distribution function on $\mathbb{R}$. Suppose $F_n \xrightarrow{d} F$, then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ carrying a sequence of random variables $(X_n)_{n\geq 1}$ and a random variable X such that $F_n = F_{X_n}$, $F = F_X$ and $X_n \to X$ a.s. as $n \to \infty$.*

Proof. See D. Williams Section 17.3 for details. Filling in the details in the case F_n and F all strictly increasing and continuous is left as an exercises following the argument given below. Fix $\varepsilon > 0$

- F_n^{-1} and F^{-1} exists and are continuous.
- Given a Uniform$(0, 1)$ random variable U the distribution of the random variables $F_n^{-1}(U)$ and $F^{-1}(U)$ are equal to X_n and X respectively.
- $F_n(F^{-1}(x \pm \varepsilon)) \to F(F^{-1}(x \pm \varepsilon)) = x \pm \varepsilon$ as $n \to \infty$, so $F_n(F^{-1}(x \pm \varepsilon))$ is smaller/larger than x for n sufficiently large.
- Then $F^{-1}(x - \varepsilon) < F_n^{-1}(x) < F^{-1}(x + \varepsilon)$ for n large. Finally use continuity to complete the argument.

□

Extending now from $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ but still focusing on weak convergence of probability measures on $(S, \mathcal{B}(S))$ when S is a Polish space, we look at some more general properties of weak convergence. First we give a little bit of background. We first see that every event can be estimated above and below by a closed and open set respectively. We suppose S is a complete (recall; this means every Cauchy sequences converges), separable (it has a countable dense subset) metric space with respect to metric $d(\cdot, \cdot)$. Since it is Polish such a metric exists. Further, for $A \subset S$, we let $d(\omega, A) = \inf_{x\in A} d(\omega, x)$.

Theorem 6.29. *Every probability measure $\mathbb{P}$ on $(S, \mathcal{B}(S))$ is* regular *meaning that for each $A \in \mathcal{B}(S)$ and each $\varepsilon > 0$ there exists a closed set F and an open set G such that $F \subseteq A \subseteq G$ and $\mathbb{P}(G \setminus F) < \varepsilon$.*

Proof. See also the proof of the approximation theorem (Theorem 2.31). Recall $\mathcal{B}(S)$ is generated by closed subsets of S. Also, if A is closed then we can approximate it from above by an open set $G = \{\omega \in S : d(\omega, A) < \delta\}$ then taking δ small and using continuity of $\mathbb{P}$ from above (Lemma 2.9) the required property holds for all closed elements of $\mathcal{B}(S)$. The result now follows if we can show that sets satisfying the required property form a σ-algebra (since $\mathcal{B}(S)$ is the smallest σ-algebra containing all closed sets). We can show that the class is closed under countable unions by taking suitable sequences $F_n \subseteq A_n \subseteq G_n$

and $\mathbb{P}(G_n \setminus F_n) < \varepsilon/2^{n+1}$ (check). It is more straightforward to check that it is closed under complements, which completes the proof. □

The above result tells us that the probability measure $\mathbb{P}$ is completely determined by its value on closed sets (or equally on open sets). Actually it is completely determined by the value of the integrals $\mathbb{P}(f) = \int f \, \mathrm{d}\mathbb{P}$ on bounded and continuous functions. Note we have multiple probability measures in the next theorem so I fall back on Greek letters.

Theorem 6.30. *Two probability measures μ and ν on $(S, \mathcal{B}(S))$ coincide if*

$$\mu(f) = \nu(f)\,, \quad \forall f \in C_b(S)\,.$$

Proof. Fix F a closed set, we can estimate $\mathbb{1}_F$ by a continuous function (c.f. the proof of Theorem 6.22), specifically for a fixed $\varepsilon > 0$ let $f(\omega) = (1 - d(\omega, F)/\varepsilon)^+$, then

$$\mathbb{1}_F(\omega) \leq f(\omega) \leq \mathbb{1}_{F^\varepsilon}(\omega)\,,$$

where $F^\varepsilon = \{\omega \in S : d(\omega, F) < \varepsilon\}$. Then, by the same argument as in the proof of Theorem 6.22, $\mu(F) \leq \mu(f) = \nu(f) \leq \nu(F^\varepsilon)$. Hence, taking $\varepsilon \searrow 0$ and using continuity from above we have $\mu(F) \leq \nu(F)$. The result follows by symmetry (interchanging μ and ν). □

We now turn to equivalent characterisations of weak convergence which can be very useful. We need to introduce some language first. For $A \subset S$ we let ∂A be the boundary of A, i.e. the closure of A set-minus the interior of A. Equivalently it is exactly the set of points which are limits of some sequence inside A and a sequence which is outside A. Given a probability measure $\mathbb{P}$ on S, we call a measurable set A a $\mathbb{P}$-*continuity set* if $\mathbb{P}(\partial A) = 0$. Notice then that the sets $(-\infty, x]$ are continuity sets for the law induced by a distribution function F on $\mathbb{R}$ if and only if x is a continuity point of F, so the following theorem contains the direction of implication we proved in Theorem 6.22.

Theorem 6.31 (Portmanteau Theorem)**.** *Let $(\mathbb{P}_n)_{n\geq 1}$ be a sequence of probability measures and $\mathbb{P}$ a probability measure all on $(S, \mathcal{B}(S))$. TFAE*

1. $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$.
2. $\mathbb{P}_n(f) \to \mathbb{P}(f)$ *for all bounded, uniformly continuous,* f.
3. $\limsup_n \mathbb{P}_n(F) \leq \mathbb{P}(F)$ *for all closed sets* F.
4. $\liminf_n \mathbb{P}_n(G) \geq \mathbb{P}(G)$ *for all open sets* G.
5. $\mathbb{P}_n(A) \to \mathbb{P}(A)$ *for all* $\mathbb{P}$*-continuity sets* A.

Sketch proof. 1. $\implies$ 2. is relatively straightforward. 2. $\implies$ 3. follows by observing that the continuous approximation in the previous proof is in fact uniformly continuous.

3. and 4. can be shown to be equivalent by taking complements. Together 3. and 4. can be shown to imply 5. by considering the interior and closure of a $\mathbb{P}$-continuity set.

Finally, 5. can be show to imply 1. by fixing a bounded continuous f which is without loss of generality bounded by 0 and 1, then using the fact we proved in the exercises sheets that $\mathbb{P}(f) = \int_0^1 \mathbb{P}(f > x)\,\mathrm{d}x$ (and similarly for $\mathbb{P}_n$) then by applying condition 5., together with the bounded convergence theorem we arrive at 1.. □

Due to results like the previous two theorems, it is possible to work equivalently with the probabilities of events, $\mathbb{P}(A)$, or with the expected values of bounded continuous functions $\mathbb{P}(f)$. In a sense the two 'views' are dual to each other. So we may therefore pick which ever one is most convenient at the time - we already did this earlier in the section, for example proving the Continuous Mapping Theorem.

Finally we take a look at a condition that, for Polish spaces, is equivalent to a class (collection) of probability measures being relatively compact (with respect to weak convergence), i.e. weak limit points (probability measures) exist, although they are not necessarily in the class. The condition is called *tightness* and it is in a sense the analogue of *uniform integrability* that we study in more detail in the next section. This condition is normally applied to sequences of probability measures, so we know if they are tight then there exists limit points, then in order to identify if there is a unique limit we may rely on weaker/different types of convergence. Just as for uniform integrability the condition will prevent 'mass running off to infinity' - i.e. it is a 'compactness condition'.

Definition 6.32 (Tightness)**.** A family of probability measures $\{\mu_\alpha\}_{\alpha\in\mathcal{I}}$, where $\mathcal{I}$ is some index set, is called *tight* if for any $\varepsilon > 0$ there exists a compact set $K \subset S$ such that $\sup_{\alpha\in\mathcal{I}} \mu_\alpha(S \setminus K) < \varepsilon$,

The definition is perhaps easier to parse in the case of measures induced by random variables on $\mathbb{R}$, i.e. the collection $\{X_n\}_{n\geq 1}$ is tight if and only if

$$\limsup_n \mathbb{P}(|X_n| > r) \to 0 \quad \text{as } r \to \infty\,.$$

(You might want to check this).

Then, as already advertised, we have the following equivalence. This is sometimes called the *Fundamental Theorem of Tightness*. **Note**: For relative compactness the simplest example you will have seen before is a bounded subset of $\mathbb{R}$, every sequence of points in such a set has a further subsequence which converges - maybe to a point outside the set if it is not closed.

Theorem 6.33 (Prokhorov's Theorem)**.** *The collection $\{\mu_\alpha\}_{\alpha\in\mathcal{I}}$ of probability measure on $(S, \mathcal{B}(S))$ is tight if and only if it is weakly relatively sequential compact, i.e. for every subsequence there is a further subsequence which converges (to a probability measure not necessarily in the class $\{\mu_\alpha\}_{\alpha\in\mathcal{I}}$).*

Note: *(feel free to ignore)* In a metric space, compactness and sequential compactness are the same thing. It turns out that the topology of weak convergence of probability measures is metrisable if S is seperable (Lévy-Prokhorov metric) and hence we can drop the 'sequentially' in the theorem above.

Proof. For a proof see for example Billingsley, *Convergence of Probability Measures*, Chapter 1 Section 5, or A. Klenke, *Probability Theory*, Section 13.3. □

As a useful corollary for applying tightness we have the following.

Corollary 6.34. *If the sequence of probably measures $(\mathbb{P}_n)_{n\geq 1}$ on $(S, \mathcal{B}(S))$ is tight, and every subsequence which is actually converging happens to converge to $\mathbb{P}$ then $\mathbb{P}_n \xrightarrow{w} \mathbb{P}$.*

Proof. Suppose $(\mathbb{P}_n)_{n\geq 1}$ is tight. Suppose for contradiction that $\mathbb{P}_n$ is not converging weakly to $\mathbb{P}$. In this case (negating the definition of weak convergence) there exists a $f \in C_b(S)$, a subsequence (n_i) and an $\varepsilon > 0$ such that $|\mathbb{P}_{n_i}(f) - \mathbb{P}(f)| > \varepsilon$ for all i. However, by Theorem 6.33 there is a further subsequence of (n_i) which converges weakly, and by assumption the weak-limit must be $\mathbb{P}$, which contradicts the distance always being bigger than ϵ. □

As a way of demonstrating a fairly typical tightness argument, we shall finally prove the other direction in Theorem 6.22, i.e. that convergence of distribution functions (at continuity points) implies weak convergence of the laws.

Completing the proof of Theorem 6.22. Using the notation from Theorem 6.22, suppose $F_n \xrightarrow{d} F$, we want to show that $\mu_n \xrightarrow{w} \mu$. We break this down into steps which follow a typically pattern for this type of argument. We will: (1) Show that the conditions imply $(\mu_n)_{n\geq 1}$ is tight. (2) Apply the converse result which we already proved. (3) Combine (2) with an already established uniqueness result and then these combined with the Corollary above (using tightness) to conclude that the entire sequence must converge weakly.

(1) If $F_n \to F$ at continuity points then

$$\begin{aligned}\limsup_n \mathbb{P}(|X_n| > r) &= \limsup_n \mathbb{P}(X_n < -r \cup X_n > r) \\ &\leq \limsup_n \left(F_n(-r) + 1 - F_n(r)\right) \\ &\leq F(-r) + 1 - F(r) \to 0 \quad \text{as} \quad r \to \infty\,.\end{aligned}$$

(2) Now $\mu_{n_i} \xrightarrow{w} \nu$ implies $F_{n_i} \xrightarrow{d} F_\nu$ (we already proved this converse).

(3) Suppose (n_i) is a subsequence along which $(\mathbb{P}_{n_i})$ convergence. Then by (2) and assumption on the limit of the F_n we have $\mu_{n_i} \xrightarrow{w} \nu$ implies $F_{n_i} \xrightarrow{d} F_\nu = F$. The final equality implies $\nu = \mu$ by uniqueness of the distribution function. To compelte the proof we apply the previous corollary. □

For random variables, that is Borel measures on $\mathbb{R}$, there are a simpler version of these results that you can prove in more 'hands on' ways just using "limsupery." I will put some of these on example sheets (guided a little) so you can test that you have a practical understanding of these things. The full abstract stuff in this section that relies on topology/metric space stuff you might not have seen is not examinable.

6.5 Uniform Integrability

To borrow the punch line from just the next page, the point of this section is to study a criterion that connects convergence in probability with $\mathcal{L}^1$ convergence. It turns out that uniform integrability is the 'right' condition to connect convergence in probability and convergence in $\mathcal{L}^1$, in particular $X_n \to X$ in $\mathcal{L}^1$ if and only if $X_n \to X$ in probability and $(X_n)_{n\geq 1}$ is uniformly integrable. That is "uniform integrability is necessary and sufficient for passing limits under expectations".

As a 'warm-up' to looking at uniform integrebility we re-state the Dominated Convergence Theorem 4.13, but with weaker assumptions on the form of convergence (although slightly stronger assumption on the domination). Recall the DCT states that if a sequence of random variables is dominated (in absolute value) by an integrable random variable, and it converges almost surely, then we can pass limits through the integral (expectation). As an immediate corollary we can show that if a sequence of random variables converges almost surely, and is dominated (in absolute value) by some real number, then the sequence converges in $\mathcal{L}^1$ (you should be able to prove this using DCT - keep in mind that since $\mathbb{P}(\Omega) = 1$ the constant function is integrable). Now we go a little further and replace the assumption of almost sure convergence with the weaker assumption of convergence in probability.

Theorem 6.35 (Bounded Convergence Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X_n)_{n\geq 1}, X$ be random variables. If $X_n \xrightarrow{\mathbb{P}} X$ and $(X_n)_{n\geq 1}$ are bounded, i.e. there exists $K > 0$ such that for any $n \in \mathbb{N}$, and $\omega \in \Omega$ we have $|X_n(\omega)| \leq K$, then $X_n \xrightarrow{\mathcal{L}^1} X$ as $n \to \infty$.*

Proof. The heart of this proof is used fairly often. We will split into a small set (w.r.t. $\mathbb{P}$) where the random variable of interest is large and a large set where it is small. First, we show that X is also bounded by K almost surely, i.e. $\mathbb{P}(|X| \leq K) = 1$. Fix $m \in \mathbb{N}$, then by assumption on X_n we have

$$\mathbb{P}(|X| > K + 1/m) \leq \mathbb{P}(|X - X_n| > 1/m) \xrightarrow{n\to\infty} 0,$$

where we use the triangle inequality for $|X| \leq |X - X_n| + |X_n|$. Hence, since the countable union of null sets is null, we get

$$\mathbb{P}(|X| > K) = \mathbb{P}(\bigcup_{m=1}^{\infty} \{|X| > K + 1/m\}) = 0\,.$$

Now fix $\varepsilon > 0$, since $X_n \xrightarrow{\mathbb{P}} X$, there exists $n_\epsilon \in \mathbb{N}$ such that

$$\mathbb{P}(|X_n - X| > \epsilon/3) < \frac{\epsilon}{3K}, \quad \text{for all} \quad n \geq n_\epsilon.$$

Hence, it follows that

$$\begin{aligned}
\mathbb{E}[|X_n - X|] &= \mathbb{E}[|X_n - X|\mathbb{1}_{\{|X_n - X| > \epsilon/3\}}] + \mathbb{E}[|X_n - X|\mathbb{1}_{\{|X_n - X| \leq \epsilon/3\}}] \\
&\leq 2K\,\mathbb{P}(|X_n - X| > \epsilon/3) + \frac{\epsilon}{3} \\
&< 2K\,\frac{\epsilon}{3K} + \frac{\epsilon}{3} = \epsilon,
\end{aligned}$$

so that indeed, $\mathbb{E}[|X_n - X|] \xrightarrow{n\to\infty} 0$, i.e. $X_n \xrightarrow{\mathcal{L}^1} X$. □

We have seen in Lemma 6.14 that convergence in $\mathcal{L}^p$ implies convergence in probability, but in Example 6.7 (together with Theorem 6.11 part 1.) we observed that the reverse is not in general true. However, under certain circumstances convergence in probability does imply convergence in $\mathcal{L}^1$. It turns out that it is sufficient to know that limit points of the sequence exists (in the $\mathcal{L}^1$ sense), i.e. if the sequence of random variables is relatively sequentially compact in $\mathcal{L}^1$ and they converge in probability they must converge in $\mathcal{L}^1$ (to the same thing - hopefully by now this type of argument is a little familiar). To this end, we introduce a condition called *uniform integrability* which is just strong enough to prohibit the behaviour in Example 6.7, and characterises the requirement of relative sequential compactness. We will see in Vitali's Convergence Theorem (Theorem 6.44 below) that indeed uniform integrability is the 'right' condition to connect convergence in probability and convergence in $\mathcal{L}^1$, in particular $X_n \to X$ in $\mathcal{L}^1$ if and only if $X_n \to X$ in probability and $(X_n)_{n\geq 1}$ is uniformly integrable. That is "uniform integrability is necessary and sufficient for passing limits under expectations". It will turn out that this concept is also of importance later when we examine Conditional Expectation and martingales, in particular because in that context it is often relatively simple to verify the uniform integrability condition.

We fist express integrability of a random variable in terms of a necessary and sufficient condition which we can then insist hold *uniformly* for a collection of random variables. The condition "prevents mass from running off to infinity" (c.f. intuition behind compactness).

Lemma 6.36. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X a random variable. Then, $X \in \mathcal{L}^1(\mathbb{P})$, if and only if*

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

Equivalently, if for any $\varepsilon > 0$ there exists a $K \in [0,\infty)$ such that $\mathbb{E}(|X|\,;\,|X| > K) < \varepsilon$.

Proof. First, suppose that $X \in \mathcal{L}^1(\mathbb{P})$. Consider the random variables X_n given by,

$$X_n = |X|\mathbb{1}_{\{|X|>n\}}, \quad \text{for each} \quad n \in \mathbb{N}.$$

By construction $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} 0$ as $n \to \infty$, and $|X_n| \leq |X| \in \mathcal{L}^1$, for each $n \in \mathbb{N}$. Hence, by the DCT, we can deduce that,

$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[\lim_{n\to\infty} X_n] = 0\,,$$

as required. Conversely, suppose that there exists $K \in \mathbb{N}$ such that $\mathbb{E}[|X|\mathbb{1}_{\{|X|>K\}}] < 1$. By law of total probability, this in turn yields,

$$\mathbb{E}[|X|] = \mathbb{E}[|X|\mathbb{1}_{\{|X|>K\}}] + \mathbb{E}[|X|\mathbb{1}_{\{|X|\leq K\}}] \leq 1 + K < \infty,$$

so that $X \in \mathcal{L}^1(\mathbb{P})$ as required. □

Uniform integrability demands that the condition above holds *uniformly* for all random variables in the same class (typically in application the class will be a sequence $(X_n)_{n\geq 1}$ and we require that the integrability condition is uniform in n).

Definition 6.37 (Uniformly Integrable class)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We say that a collection of random variables $\mathcal{C}$ is uniformly integrable (UI), if for all $\epsilon > 0$, there exists $K > 0$ such that

$$\mathbb{E}[|X|\mathbb{1}_{\{|X|>K\}}] < \epsilon, \quad \text{for all } X \in \mathcal{C},$$

(note we have the same K for each X). Equivalently,

$$\lim_{K\to\infty} \sup_{X\in\mathcal{C}} \mathbb{E}(|X|\mathbb{1}_{\{|X|>K\}}) = \lim_{K\to\infty} \sup_{X\in\mathcal{C}} \mathbb{E}(|X|\,;\,|X|>K) = 0.$$

Remark 6.38. Note that a uniform integrable (UI) class $\mathcal{C}$ of random variables is in $\mathcal{L}^1(\mathbb{P})$, i.e. $X \in \mathcal{L}^1(\mathbb{P})$, for all $X \in \mathcal{C}$, which follows from Lemma 6.36. But $\mathcal{C} \subseteq \mathcal{L}^1(\mathbb{P})$ does not necessarily imply that $\mathcal{C}$ that is UI. The underlying idea is that, unlike the case of a single random variable, the tails of random variables in $\mathcal{C}$ could, in some limiting sense, escape to infinity (again it is maybe helpful to have in mind Example 6.7 to picture what we are trying to control).

Remark 6.39. When the measure is not finite then we have to be a little more careful about the definition of uniform integrability, and there are indeed many equivalent forms of this condition. For more details see Section 6.2 in *A. Klenke*.

Example 6.40. If $\mathcal{C} = \{X\}$, then $\mathcal{C}$ is UI if and only if X is integrable. The same statement holds for any finite collection, by the same argument.

If $\mathcal{C} = \{X_n\}_{n\geq 1}$, which will often be the case in applications, then $\mathcal{C}$ is UI if and only if $\lim_{K\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}(|X_n|\,;\,|X_n|>K) = 0$.

Checking uniform integrability for a class of random variables straight from the definition might be too involved, sometimes it is more convenient to check the following sufficient conditions.

Proposition 6.41. *A collection $\mathcal{C}$ is uniformly integrable if and only if $\mathbb{E}[(|X|-K)^+] \to 0$ as $K \to \infty$, uniformly in $X \in \mathcal{C}$ (where $Y^+ = Y\mathbb{1}_{Y\geq 0}$ is the positive part).*

Proof. ($\Rightarrow$) Observe that $0 \leq (|X| - K)^+ \leq |X|\mathbb{1}_{\{|X|>K\}}$.

($\Leftarrow$) Observe that $0 \leq |X|\mathbb{1}_{\{|X|>2K\}} \leq 2(|X| - K)^+$.

□

Lemma 6.42 (Two sufficient conditions for UI)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{C}$ a class of random variables. Then $\mathcal{C}$ is uniformly integrable if either of the following holds:*

(a) The random variables in $\mathcal{C}$ are uniformly bounded in $\mathcal{L}^p$, for some $p > 1$, that is $||X||_p \leq A$ for some $A > 0$ and for all $X \in \mathcal{C}$.

(b) The random variables in $\mathcal{C}$ are uniformly dominated by some random variable $Y \in \mathcal{L}^1(\mathbb{P})$, that is $\mathbb{E}[|Y|] < \infty$ and $|X| \leq |Y|$ a.s., for all $X \in \mathcal{C}$.

Proof. (a) Suppose $||X||_p \leq A$ for some $A > 0$ and for all $X \in \mathcal{C}$. Given $X \in \mathcal{C}$, suppose that $|X| > K$, for some $K > 0$. Then, we have that,

$$|X|^p = |X|^{p-1}|X| > K^{p-1}|X|.$$

Thus, it follows that

$$\int_\Omega |X| \mathbb{1}_{\{|X|>K\}} \, \mathrm{d}\mathbb{P} < \int_\Omega \frac{|X|^p}{K^{p-1}} \, \mathrm{d}\mathbb{P} = \frac{1}{K^{p-1}} \mathbb{E}[|X|^p] \leq \frac{1}{K^{p-1}} A^p \xrightarrow{K\to\infty} 0, \quad \forall X \in \mathcal{C}.$$

(b) Since $|X| \leq Y$ a.s., for each $X \in \mathcal{C}$, then

$$\mathbb{1}_{\{|X|>K\}} \leq \mathbb{1}_{\{|Y|>K\}} \text{ a.s.}, \quad \forall X \in \mathcal{C}\,.$$

Now, integrating yields,

$$\int_\Omega |X| \mathbb{1}_{\{|X|>K\}} \, \mathrm{d}\mathbb{P} \leq \int_\Omega |Y| \mathbb{1}_{\{|Y|>K\}} \, \mathrm{d}\mathbb{P} \xrightarrow{K\to\infty} 0, \quad \forall X \in \mathcal{C}.\,,$$

by Lemma 6.36.

□

Proposition 6.43. *Let $\mathcal{C} = (X_\alpha)_{\alpha\in I}$ be a uniformly integrable family of random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, the following hold:*

(a) $\sup_{\alpha\in I} \mathbb{E}|X_\alpha| < \infty$,

(b) $\sup_{\alpha\in I} \mathbb{P}(|X_\alpha| > K) \to 0$ as $K \to \infty$,

(c) $\sup_{\alpha\in I} \mathbb{E}[|X_\alpha| \mathbb{1}_A] \to 0$ as $\mathbb{P}(A) \to 0$,
i.e. $\forall \epsilon > 0,\ \exists \delta > 0\ \forall \alpha \in I : \mathbb{P}(A) < \delta \implies \mathbb{E}[|X_\alpha| \mathbb{1}_A] < \epsilon$.

Conversely, if either (a) and (c), or (b) and (c), hold then $\mathcal{C}$ is uniformly integrable.

Proof. We first prove (a). By definition of uniform integrability, there exists a $K > 0$ such that for all α

$$\mathbb{E}[|X_\alpha| \,;\, |X_\alpha| > K] \leq 1\,.$$

Then, for all α,

$$\mathbb{E}[|X_\alpha|] = \mathbb{E}\big[|X_\alpha| \mathbb{1}_{\{X_\alpha|\leq K\}} + |X_\alpha| \mathbb{1}_{\{|X_\alpha|>K\}}\big] \leq K + 1 < \infty\,.$$

Now, (a) implies (b) by applying Markov's inequality

$$\mathbb{P}[|X_\alpha| > K] \leq \frac{1}{K}\mathbb{E}[|X_\alpha|] \leq \frac{1}{K}\sup_\beta \mathbb{E}[|X_\beta|] \to 0 \quad \text{as} \quad K \to \infty, \text{ uniformly in } \alpha\,.$$

By law of total probability, for any $\alpha \in I$ and $A \in \mathcal{F}$ with $\mathbb{P}(A) \to 0$, we get that,

$$\begin{aligned}\mathbb{E}[|X_\alpha|\mathbb{1}_A] &= \mathbb{E}[|X_\alpha|\mathbb{1}_{A\cap\{|X_\alpha|>K\}}] + \mathbb{E}[|X_\alpha|\mathbb{1}_{A\cap\{|X_\alpha|\leq K\}}] \\ &\leq \mathbb{E}[|X_\alpha|\mathbb{1}_{\{|X_\alpha|>K\}}] + K\mathbb{P}(A\cap\{|X_\alpha|\leq K\}) \\ &\leq \mathbb{E}[|X_\alpha|\mathbb{1}_{\{|X_\alpha|>K\}}] + K\mathbb{P}(A) \\ &\to 0, \quad \text{as } \mathbb{P}(A)\to 0,\ K\to\infty.\end{aligned}$$

For the converse, since (a) implies (b) , it is sufficient to check that (b) and (c) imply uniform integrability. Fix $\varepsilon > 0$, by (c) there exists a $\delta > 0$ such that $\mathbb{P}(A) < \delta$ implies $\mathbb{E}[|X_\alpha|\,;\, A] < \varepsilon$. Also, by (b) we know that there exists a K such that $\mathbb{P}(|X_\alpha| > K) < \delta$ for all α. Putting these two facts together we have $\mathbb{E}\big[|X_\alpha|\,;\, |X_\alpha| > K\big] < \varepsilon$ for all α. □

If the measure $\mathbb{P}$ is atomless (i.e. for any countable $C \subset \mathbb{R}$ we have $\mathbb{P}(C) = 0$), then (c) on it's own implies uniform integrability. Hence in the 'continuous setting' (c) is often the best way of thinking about it. If $\mathbb{P}(\{\omega\}) > 0$ for some x and $X_n(\omega) \to \infty$ then (*c*) could still hold for each sequence of sets which have vanishing mass under $\mathbb{P}$ but (a) doesn't hold and hence the collection is not U.I. .

Recall, convergence in $\mathcal{L}^p$ implies convergence in probability, but not the reverse, essentially because of the possible contribution to the mean of rare but very large values. The next theorem is the main result of this section, and shows that the uniform integrability is the right condition to connect these concepts.

Theorem 6.44 (Vitali's Convergence Theorem)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of integrable random variables which converge in probability to a random variable X. Then the following are equivalent (TFAE)*

1. *the family $\{X_n\}_{n\geq 1}$ is uniformly integrable,*
2. *$X_n \xrightarrow{\mathcal{L}^1} X$, i.e. $\mathbb{E}[|X_n - X|] \to 0$ as $n \to \infty$,*
3. *$\mathbb{E}|X_n| \to \mathbb{E}|X| < \infty$ as $n \to \infty$.*

Proof. Fix $X_1, X_2, \ldots \in \mathcal{L}^1$ and assume that $X_n \xrightarrow{\mathbb{P}} X$.

We will first show that 1 implies 2. Suppose $\{X_n\}_{n\geq 1}$ are uniformly integrable. We will try to repeat the proof of Theorem 6.35. After splitting the expectation, the fact that the integral of $|X_n|$ on a small set tends to zero follows from Proposition 6.43 (c), the fact that the expectation of $|X_n - X|$ is small on the large set where $|X_n - X|$ is straightforward, so it remains to show that X is integrable and hence the expectation of $|X|$ on a small set is small. So we first show that X is integrable, then complete the proof following this argument.

Since $X_n \xrightarrow{\mathbb{P}} X$ it follows from the triangle inequality that $|X_n| \xrightarrow{\mathbb{P}} |X|$, so by Theorem 6.11 (2) there exists a subsequence $(X_{n_k})_{k\geq 1}$ such that $X_{n_k} \to X$ a.s. as $k \to \infty$, hence Fatou's Lemma gives

$$\mathbb{E}[|X|] \leq \liminf_{k\to\infty} \mathbb{E}[|X_{n_k}|] \leq \sup_n \mathbb{E}[|X_n|] < \infty\,,$$

where the last inequality follows from Proposition 6.43 (a). Hence $X \in \mathcal{L}^1(\mathbb{P})$. Now we need to complete the argument sketched above. Fix $\varepsilon > 0$ and let $A_n = \{|X_n - X| > \epsilon\}$, then

$$\mathbb{E}[|X_n - X|] = \mathbb{E}[|X_n - X|\,;\, A_n] + \mathbb{E}[|X_n - X|\,;\, A_n^c] \tag{6.4}$$
$$\leq \mathbb{E}[|X_n|\,;\, A_n] + \mathbb{E}[|X|\,;\, A_n] + \varepsilon\,. \tag{6.5}$$

Since $X_n \xrightarrow{\mathbb{P}} X$ we have $\mathbb{P}(A_n) \to 0$ as $n \to \infty$. Further $\{X_n\}_{n\geq 1}$ is uniformly integrable and so is $\{X\}$, since $X \in \mathcal{L}^1(\mathbb{P})$, hence by Proposition 6.43 (c),

$$\mathbb{E}[|X_n|\,;\, A_n] \to 0 \quad \text{and} \quad \mathbb{E}[|X|\,;\, A_n] \to 0 \quad \text{as} \quad n \to \infty\,.$$

So, for n sufficiently large we have $\mathbb{E}[|X_n - X|] \leq 2\varepsilon$ which completes the proof of 1. $\implies$ 2..

Notice that 2. implies 3. by the triangle inequality (check).

It remains to show 3. implies 1. To simplify notation a little let $Y_n = |X_n|$ and $Y = |X|$. By assumption $X_n \xrightarrow{\mathbb{P}} X$, so by the triangle inequality $Y_n \xrightarrow{\mathbb{P}} Y$. Also, by 3., $\mathbb{E}[Y_n] \to \mathbb{E}[Y] < \infty$. We will use the characterisation of UI given in Proposition 6.41.

Fix $\epsilon > 0$. Since $\{Y\}$ is uniformly integrable, there exists $K > 0$ such that $\mathbb{E}[(Y - K)^+] < \epsilon$. To help notation again we introduce some short hand for the truncation at K of a non-negative random variable (just for this proof). For any non-negative random variable Z, define

$$Z^{(K)} = Z \wedge K = \begin{cases} Z & Z \leq K\,, \\ K & Z > K\,. \end{cases}$$

Notice (picture in lectures) that

$$(Z - K)^+ = Z - Z^{(K)}\,. \tag{6.6}$$

We want to show that $\mathbb{E}[(Y_n - K')^+]$ is small for K' sufficiently large, uniformly in n.

Since $|(a\wedge K) - (b\wedge K)| \leq |a - b|$ we have $Y_n^{(K)} \xrightarrow{\mathbb{P}} Y^{(K)}$. So by the bounded convergence theorem (Theorem 6.35), $\mathbb{E}[Y_n^{(K)}] \to \mathbb{E}[Y^{(K)}]$. Now, using identity (6.6) and assumption $\mathbb{E}[Y_n] \to \mathbb{E}[Y]$,

$$\mathbb{E}[(Y_n - K)^+] = \mathbb{E}[Y_n] - \mathbb{E}[Y_n^{(K)}] \to \mathbb{E}[Y] - \mathbb{E}[Y^{(K)}] = \mathbb{E}[(Y - K)^+] < \epsilon\,. \tag{6.7}$$

Hence, there exists some $n_0 \in \mathbb{N}$ such that for each $n \geq n_0$ we have $\mathbb{E}[(Y_n - K)^+] \leq 2\epsilon$. So we only have to control what happens for finitely many n. Any finite collection of uniformly integrable random variables is uniformly integrable so we may choose K' such that

$$\mathbb{E}[(Y_n - K')^+] \leq 2\epsilon\,, \quad \text{for all } n \in \mathbb{N}, \tag{6.8}$$

as required.

□

The following is an immediate consequence of 1. and 2. above.

Corollary 6.45. *Suppose $(X_n)_{n\geq 1}$ be a sequence of random variables that are integrable, and X also integrable (i.e. $X \in \mathcal{L}^1$). Then $X_n \xrightarrow{\mathcal{L}^1} X$ if and only if;*

(i) $X_n \to X$ in probability, and

(ii) the family $\{X_n\}$ is uniformly integrable.

Chapter 7

Limit Theorems

Reading (For characteristic functions and CLT):
D. Williams, Chapter 17 and 18 and A. Klenke, Chapter 15

In this Chapter we will look at characteristic functions and how they may be used to prove a central limit theorem (CLT), as well as stating and proving a more general weak law of large numbers (WLLN), and finally turning to Etemadi's proof of the strong law of large numbers (SLLN).

7.1 Characteristic Functions

The theory of characteristic functions forms an important tool for investigating probability distributions. Characteristic functions form a special case of the more general Fourier transform, in particular they are the Fourier transforms of probability measures. The theory around this topic is extremely rich, here we will restrict to a very brief account of the theory, with a focus on being able to prove a CLT for real-valued random variables.

The characteristic function is related to the *moment generating function* (or Laplace transform) which we have already seen in the introduction, see Claim 1.5,

$$G_X(s) = \mathbb{E}[e^{sX}] \quad \text{for all } x \text{ such that the expectation is finite.}$$

It turns out that the moment generating function (Laplace transform if we take $s \in (-\infty, 0]$) uniquely characterises any probability measure on $[0, \infty)$, that is the distribution of any non-negative *bounded* random variable. We will see below that the characteristic function uniquely characterises any probability measure on $\mathbb{R}$ (without requiring that the support is bounded). We will start with definitions, basic properties and the typical workflow when using characteristic functions.

Definition 7.1 (Characteristic function of a random variable)**.** Let X be a random variable, F_X its distribution function and μ_X its law. The *characteristic function* (CF) of X is the map $\varphi_X \colon \mathbb{R} \to \mathbb{C}$, defined by

$$\varphi_X(\theta) = \mathbb{E}[e^{i\theta X}] = \mathbb{E}[\cos(\theta X)] + i\mathbb{E}[\sin(\theta X)] = \int_{\mathbb{R}} e^{i\theta x} \, \mathrm{d}F_X(x) = \int_{\mathbb{R}} e^{i\theta x} \mu_x(\, \mathrm{d}x) \, .$$

It is important that although the co-domain of φ_X is $\mathbb{C}$, the domain is $\mathbb{R}$, i.e. the function is only defined for real values of θ.

Note, we do not divide by $(2\pi)^{1/2}$ as is often the case in Fourier analysis (to normalise things). The second equality in the definition above is just a result of Euler's formula $e^{ix} = \cos(x) + i\sin(x)$ for $x \in \mathbb{R}$.

Remark 7.2. Since $|e^{i\theta X(\omega)}| = 1$ for each $\omega \in \Omega$ and $\mathcal{L}_X(\mathbb{R}) = 1$ it follows that $\mathcal{L}_X(|e^{i\theta X}|) = 1 < \infty$, i.e. the characteristic function $\varphi_X(\theta)$ is finite for all values of $\theta \in \mathbb{R}$. This is an important difference between the characteristic function and the moment generating function $\mathbb{E}[e^{tX}]$, for $t \in \mathbb{R}$, of a random variable X.

Lemma 7.3 (Elementary properties of CFs)**.** *Let $\varphi = \varphi_X$ for some random variable X, then:*

(i) $\varphi(0) = 1$ *(clear since $\mathbb{E}e^{i0X} = \mathbb{E}1 = 1$).*

(ii) $|\varphi(\theta)| \leq 1$ *for all $\theta \in \mathbb{R}$.*

(iii) $\theta \mapsto \varphi(\theta)$ *is continuous on $\mathbb{R}$.*

(iv) $\varphi_{-X}(\theta) = \varphi_X(-\theta) = \overline{\varphi_X(\theta)}$ *for all $\theta \in \mathbb{R}$.*

(v) $\varphi_{aX+b}(\theta) = e^{i\theta b}\varphi_X(a\theta)$ *for all $a, b, \theta \in \mathbb{R}$.*

(vi) If $\mathbb{E}(|X|^n) < \infty$ then the n^{th}-derivative satisfies $\varphi_X^{(n)}(0) = i^n\mathbb{E}[X^n]$ (note it is possible that $\varphi'(0)$ exists even if $\mathbb{E}|X| = \infty$).

Proof. On exercises sheet. □

Characteristic functions can be particularly useful for studying random variables, here is a list of just some of their applications;

- To prove Central Limit Theorems (CLTs) and analogous results.
- To calculate the distribution of limiting random variables.
- To obtain estimates on tail probabilities via saddle point approximations (Laplace method).
- To prove results like; If X and Y are independent and the sum $X + Y$ is normally distributed then both X and Y must have normal distribution.

We will focus on the first two points above.

We now summarise in words three key theoretical results which are worth always keeping in mind.

1. If X and Y are independent random variables then

   $$\forall\ \theta \in \mathbb{R} \quad \text{we have} \quad \varphi_{X+Y}(\theta) = \varphi_X(\theta)\varphi_Y(\theta)\,.$$

   (Note this just involves applying the rule 'independence means multiply' $\mathbb{E}[e^{i\theta(X+Y)}] = \mathbb{E}[e^{i\theta X}e^{i\theta Y}] = \mathbb{E}[e^{i\theta X}]\mathbb{E}[e^{i\theta Y}]$.)

2. There is a 1-to-1 correspondence between measures on $(\mathbb{R}, \mathcal{B})$ and characteristic functions. This is stated precisely in Lévy's inversion formula (below), which shows explicitly how to reconstruct the distribution function from the CF.

3. Weak convergence corresponds exactly to convergence of the associated CFs. This is stated precisely in Lévy's convergence theorem below.

The "tidiness" of Lévy's inversion formula and convergence theorem is compromised a little by the presence of atoms in the distribution. If for some $c \in \mathbb{R}$ we have $\mathbb{P}(X = c) > 0$ then we say that the law of X, $\mathcal{L}_X$, has an *atom* at c. Observe that the number of atoms is at most countable, since there are at most n atoms of mass at least $1/n$. Also, a distribution (law) that is made up of only atoms is a discrete distribution. In order to reduce notation a bit we use $F(a-) = \lim_{x \nearrow a} F(x)$ (i.e. the left limit of F at a). Recall that F is right continuous so $\lim_{x \searrow a} F(x) = F(a)$, and $\mathbb{P}(X = c) = \mathcal{L}_X(\{c\}) = \mathbb{P}(X \leq c) - \mathbb{P}(X < c) = F(c) - F(c-)$. That is the probability 'mass' on the point c under $\mathcal{L}_X$ is given by the size of the 'jump' in the distribution function at the point c.

Theorem 7.4 (Lévy's inversion formula)**.** *Let φ be the characteristic function of a random variable X, with law μ and distribution function F. Then for $a < b \in \mathbb{R}$*

$$\begin{aligned}\lim_{T \nearrow \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-i\theta a} - e^{-i\theta b}}{i\theta} \varphi(\theta) \, \mathrm{d}\theta &= \frac{1}{2}\mu(\{a\}) + \mu((a,b)) + \frac{1}{2}\mu(\{b\}) \\ &= \frac{1}{2}(F(b) + F(b-)) - \frac{1}{2}(F(a) + F(a-)) \, .\end{aligned}$$

Moreover if $\int_{\mathbb{R}} |\varphi(\theta)| \, \mathrm{d}\theta < \infty$ then X has a continuous density function (with respect to the Lebesgue measure on $\mathbb{R}$) given by $f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-i\theta x} \varphi(\theta) \, \mathrm{d}\theta$.

Note the duality, in the latter case, with

$$\varphi(\theta) = \mathbb{E}[e^{i\theta X}] = \int_{\mathbb{R}} e^{i\theta x} f(x) \, \mathrm{d}x \, ,$$

which may be familiar from Fourier analysis.

Proof. See Williams 1991. □

It follows from inversion formula that two random variables X and Y have the same distribution (law) if and only if they have the same characteristic functions. We now turn to convergence.

Theorem 7.5 (Lévy's convergence theorem)**.** *Let $(F_n)_{n \geq 1}$ be a sequence of distribution functions with associated characteristic functions $(\varphi_n)_{n \geq 1}$. Suppose;*

(i) $g(\theta) = \lim_{n \to \infty} \varphi_n(\theta)$ exists for all $\theta \in \mathbb{R}$, and

(ii) $g(\cdot)$ is continuous at 0.

Then $g = \varphi$ is the characteristic function of some distribution F and $F_n \xrightarrow{w} F$ as $n \to \infty$.

Proof. See Williams 1991. The proof typically uses a tightness argument based on the tail estimate

$$\mathbb{P}(|X| \geq r) \leq \frac{r}{2} \int_{-2/r}^{2/r} (1 - \varphi_X(t)) \, \mathrm{d}t \, ,$$

which implies that the sequence of distributions corresponding to F_n are tight. Hence, there is a weakly converging subsequence, and weak convergence implies convergence of the characteristic functions (by definition of weak convergence). A proof by contradiction then shows the limit point must be unique. This exemplifies a typical 'tightness' type argument, that is first show tightness, then use a previously established converse convergence result (in this case weak convergence implies convergence of the CFs) and finally some previously established uniqueness result (in this case Lévy's inversion formula). □

7.2 The Central Limit Theorem

The Central Limit Theorem is one of the greatest results of mathematics. It is extremely general and extraordinarily useful a computational tool. We now prove the central limit theorem as a consequence of Lévy's Convergence Theorem. To do this, we will need Taylor expansion estimates of characteristic functions. For instance, let X be a random variable with law $\mathcal{L}_X$ and associated distribution F_X, where $\mathbb{E}[|X|^k] < \infty$, for some $k \in \mathbb{N}$. Recall also that $f(\theta) = o(g(\theta))$ means that $f(\theta)/g(\theta) \to 0$, as $\theta \to 0$. Then the characteristic function of X, φ_X, satisfies

$$\varphi_X(\theta) = \mathbb{E}\big[e^{i\theta X}\big] = \mathbb{E}\Big[\sum_{n=0}^{\infty} \frac{(i\theta X)^n}{n!}\Big] = \sum_{n=0}^{k} \frac{\mathbb{E}[X^n]}{n!}(i\theta)^n + o(\theta^k), \quad \text{as } \theta \to 0. \tag{7.1}$$

In fact, proving Eq. (7.1) is the 'real work' of the proof. It follows from the Taylor's Theorem (in-particular bounds on the remainder) together with the dominated convergence theorem. The details are in the lemma below, a similar proof made a little more general proves that we can exchange the order of the expectation and the differentiation in Lemma 7.3 (vi).

Lemma 7.6 (Taylor estimate)**.** *If X is a zero-mean random variable in $\mathcal{L}^2$, i.e. $\mathbb{E}[X] = 0$ and $\sigma^2 = \mathrm{Var}[X] < \infty$, then*

$$\varphi_X(\theta) = 1 - \frac{1}{2}\sigma^2\theta^2 + o(\theta^2) \quad \textit{as } \theta \to 0\,.$$

Proof. Note, we are trying to show that

$$\left|\varphi_X(\theta) - \left(1 - \frac{1}{2}\sigma^2\theta^2\right)\right| = \left|\mathbb{E}[e^{i\theta X}] - \left(1 - \frac{1}{2}\sigma^2\theta^2\right)\right| = \left|\mathbb{E}\left[e^{i\theta X} - \sum_{k=0}^{2} \frac{(i\theta X)^n}{k!}\right]\right|$$

goes to zero faster than θ^2 as $\theta \to 0$ (where the first equality is the definition of the CF and the second is linearity of expectation together with the assumptions of the lemma). We begin with a 'standard' Taylor approximation which can be derived by observed that all the derivatives of e^{ix} have modulus one. Fix $x \in \mathbb{R}$ and define the remainders

$$R_n(x) = e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!}\,.$$

[We will evaluate the remainders at $x = \theta X(\omega)$ and $n = 2$]. Notice, $R_0(x) = e^{ix} - 1$ and, by differentiating the expression above with respect to x, that

$$R_0(x) = \int_0^x ie^{iy}\,\mathrm{d}y\,, \quad \text{and} \quad R_n(x) = \int_0^x iR_{n-1}(y)\,\mathrm{d}y\,, \quad \text{for } n \geq 1\,. \tag{7.2}$$

So, $|R_0(x)| = |e^{ix} - 1| \leq 2$ by the triangle inequality, and $R_0(x) \leq |x|$ by pulling the absolute value inside the integral above (triangle inequality for integral), hence $R_0(x) \leq \min(2, |x|)$. Now applying (7.2) inductively we find

$$|R_n(x)| \leq \min\left(\frac{2|x|^n}{n!}, \frac{|x|^{n+1}}{(n+1)!}\right)\,. \tag{7.3}$$

Applying this bound in the case $n = 2$ we have,

$$\left| e^{i\theta x} - \left(1 + i\theta x - \frac{1}{2}\theta^2 x^2\right)\right| = |R_2(\theta x)| \leq \theta^2 \min\left(x^2, |\theta|\frac{|x|^3}{6}\right) .$$

Now suppose X has mean zero and variance $\sigma < \infty$, then by linearity of the expectation

$$\left|\varphi_X(\theta) - \left(1 - \frac{1}{2}\sigma^2\theta^2\right)\right| = \left|\mathbb{E}\left[e^{i\theta X} - \sum_{k=0}^{2}\frac{(i\theta X)^n}{k!}\right]\right| = |\mathbb{E}[R_2(\theta X)]| \leq \mathbb{E}[|R_2(\theta X)|] .$$

Finally, dividing by θ^2 and applying (7.3), by monotonicity of the expectation

$$\frac{1}{\theta^2}\mathbb{E}[|R_2(\theta X)|] \leq \mathbb{E}\left[\min\left(X^2, |\theta|\frac{|X|^3}{6}\right)\right] ,$$

Then, since the integrand on the right is dominated by X^2 and $\mathbb{E}[X^2] < \infty$, so we can apply dominated convergence and pass the limit $\theta \to 0$ through the expectation to get the desired result (i.e. that $\mathbb{E}[|R_2(\theta X)|]$ is little-o of θ^2). The point of having the min over both terms is we use the first one to get the domination requirement in DCT and the second one still has a theta in front, so point-wise, taking theta to zero, the term under the expectation is tending to zero. □

Now, combining this approximation with Lévy's convergence theorem we can prove the CLT.

Theorem 7.7 (Central Limit Theorem)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $\mathbb{E}[X_n] = 0$ and $Var(X_n) = \sigma^2 < \infty$ for each $n \in \mathbb{N}$. For $n \in \mathbb{N}$ let $S_n := X_1 + ... + X_n$ and $G_n = \frac{S_n}{\sigma\sqrt{n}}$. Then $G_n \to \mathcal{N}(0, 1)$ in distribution, i.e. for any $x \in \mathbb{R}$*

$$\mathbb{P}(G_n \leq x) \xrightarrow{n\to\infty} \Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{y^2}{2}}\, \mathrm{d}y .$$

Proof. Fix $\theta \in \mathbb{R}$. Then, by Lemma 7.3 (v), independence of $X_1, ..., X_n$ and (7.1), we have

$$\varphi_{G_n}(\theta) = \varphi_{S_n}\Big(\frac{\theta}{\sigma\sqrt{n}}\Big) = \Big[\varphi_{X_1}\Big(\frac{\theta}{\sigma\sqrt{n}}\Big)\Big]^n = \Big(1 - \frac{1}{2}\frac{\theta^2}{n} + o(\frac{\theta^2}{\sigma^2 n})\Big)^n, \quad \text{as } n \to \infty ,$$

where the last equality follows from Lemma 7.6. If we let

$$z_n = -\frac{1}{2}\frac{\theta^2}{n} + o(\frac{\theta^2}{\sigma^2 n}) ,$$

then using the fact that $|\log(1 + z) - z| \leq |z|^2$ for $z \in \mathbb{C}$ with $|z| \leq 1/2$[1] we observe that

$$\frac{1}{n}\log\varphi_{G_n}(\theta) = \log\Big(1 + z_n\Big) = z_n + o(z_n) = z_n + o(\frac{\theta^2}{n}) .$$

[1]Note this inequality is essentially just $\log(1 + x) \approx x$ for x small, which you should be familiar with. This more quantitative version on $\mathbb{C}$ follows from integration by parts using principal values for the log. This little bit of complex analysis is not assumed knowledge for the module, so take this approximation for granted if necessary.

Hence

$$\log \varphi_{G_n}(\theta) = -\frac{\theta^2}{2} + n\, o\Big(\frac{\theta^2}{n}\Big) \to -\frac{\theta^2}{2} \quad \text{as} \quad n \to \infty\,.$$

It follows that $\varphi_{G_n}(\theta) \to e^{-\frac{\theta^2}{2}}$ as $n \to \infty$. Finaly, observe that $\theta \mapsto e^{-\frac{\theta^2}{2}}$ is the characteristic function of a $N(0,1)$ random variable. The result now follows from Lévy's Convergence Theorem (Theorem 7.5). □

Using characteristic functions we arrive at a more standard expression of the Weak Law of Large Numbers (WLLN) than our prototype that we already proved, that only assumes finite first moments.

Theorem 7.8 (Weak Law of Large Numbers)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of independent identically distributed random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $X_1 \in \mathcal{L}^1(\mathbb{P})$ and $\mathbb{E}[X_1] = \mu < \infty$. Let $S_n = \sum_{i=1}^n X_i$, then*

$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu. \quad \textit{as} \quad n \to \infty\,.$$

Remark 7.9. We have already seen a special case of this theorem, when $\mathrm{Var}(X_i) < \infty$ using Chebyshev's inequality.

Proof. We know from previous results, Lemma 6.26, that convergence in distribution to a constant implies convergence in probability. We will therefore apply Lévy's Convergence theorem to show the appropriate convergence in distribution (weak convergence).

Fix $\theta \in \mathbb{R}$. Similar to the proof of the Central Limit Theorem (CLT), we apply the Taylor's expansion again

$$\varphi_{\frac{S_n}{n}}(\theta) = \Big[\varphi_{X_1}\Big(\frac{\theta}{n}\Big)\Big]^n = \Big(1 + i\mu\frac{\theta}{n} + o\Big(\frac{\theta}{n}\Big)\Big)^n, \quad \text{as } n \to \infty.$$

Similar to the proof of the CLT, we have

$$\log \varphi_{\frac{S_n}{n}}(\theta) := \mu\theta i + n\, o\Big(\frac{\theta}{n}\Big) \to \mu\theta i, \quad \text{as} \quad n \to \infty\,,$$

and hence

$$\varphi_{\frac{S_n}{n}}(\theta) = \Big(1 + \frac{x_n}{n}\Big)^n \to e^{i\theta\mu} \quad \text{as} \quad n \to \infty\,.$$

Finaly, observe that $\theta \mapsto e^{i\theta\mu}$ is the characteristic function of a constant random variable μ. The result follows by Lévy's Convergence Theorem. □

Remark 7.10. Actually the convergence in Theorem 7.8 holds almost surely. There are nice proofs of this fact using (reverse)martingales, and also one we can follow with the machinery already developed (Etemadi's proof) which only requires pairwise independence. We will see this in the next lecture.

A typical workflow for using characteristic functions:

- Given a problem involving sums of independent random variables.
- Convert into a problem involving products of characteristic functions.

- Solve using tools and results from algebra and analysis to find the characteristic function of the limiting object.
- Make conclusion about the (distribution of the) desired random variable using Lévy's Convergence theorem.

Lévy's Inversion Formula allows us to identify the distribution of a random variable if we recognise the characteristic function. To this end, we record the characteristic functions of some known discrete and continuous distributions in the following table. It is a good exercise to try and derive these, using the associated pmf/pdf.

Distribution	**pmf/pdf**	**Support**	**CF**
Constant a	1	$\{a\}$	$e^{i\theta a}$
$Ber(p)$	$p^k(1-p)^{1-k}$	$\{0,1\}$	$(1-p)+pe^{i\theta}$
$Bin(n,p)$	$\binom{n}{k}p^k(1-p)^{n-k}$	$\{0,...,n\}$	$((1-p)+pe^{i\theta})^n$
$Poi(\lambda)$	$\frac{e^{-\lambda}\lambda^k}{k!}$	$\mathbb{N}\cup\{0\}$	$e^{\lambda(e^{i\theta}-1)}$
$N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{-(x-\mu)^2}{\sigma^2}}$	$\mathbb{R}$	$e^{i\theta\mu-\frac{1}{2}\theta^2\sigma^2}$
$U[a,b]$	$\frac{1}{b-a}$	$[a,b]$	$\frac{e^{i\theta b}-e^{i\theta a}}{i\theta(b-a)}$
$Exp(\lambda)$	$\lambda e^{-\lambda x}$	$[0,\infty)$	$\frac{\lambda}{\lambda-i\theta}$
Double Exponential	$\frac{1}{2}e^{-\lvert x\rvert}$	$\mathbb{R}$	$\frac{1}{1+\theta^2}$
Cauchy	$\frac{1}{\pi(1+x^2)}$	$\mathbb{R}$	$e^{-\lvert\theta\rvert}$
Triangular	$1-\lvert x\rvert$	$[-1,1]$	$2(\frac{1-\cos(\theta)}{\theta^2})$
Anon	$\frac{1-\cos(x)}{\pi x^2}$	$\mathbb{R}$	$(1-\lvert\theta\rvert)\mathbb{1}_{[-1,1]}(\theta)$

Remark 7.11. There is a duality between the pdfs and characteristics functions of the Double Exponential and Cauchy distributions, as well as the Triangular distribution and the Anon. distribution. This demonstrates the duality between the density function and the characteristic function in the case that a random variable has a continuous pdf (see the final remark in our statement of Lévy's Inversion Formula).

7.3 The Strong Law of Large Numbers

In this section we will prove a result which is stronger than Theorem 7.8, namely we improve that theorem by showing convergence holds almost surely. We will see a "neater" proof later once we have studied discrete time martingales.

Theorem 7.12 (Etemadi's Strong Law of Large Numbers)**.** *Suppose $X_1, X_2, \ldots \in \mathcal{L}^1(\mathbb{P})$ are a sequence of pairwise independent and identically distributed random variables with mean $\mathbb{E}[X_1] = \mu$, and $S_n = \sum_{i=1}^n X_i$, then*

$$\frac{S_n}{n} \to \mu \quad a.s..$$

We start with a lemma that shows that it is sufficient to consider truncated versions of the random variables $(X_n)_{n\geq}$, the details of the proof are left as an exercises on this weeks problem sheet.

Lemma 7.13. *For $n \in \mathbb{N}$ define $Y_n = X_n \mathbb{1}_{\{|X_n| \leq n\}}$ and $T_n = Y_1 + \ldots + Y_n$. The sequence $(X_n)_{n\in\mathbb{N}}$ satisfies the strong law of large numbers above if $T_n/n \to \mu$ a.s. as $n \to \infty$.*

Proof. It follows from Exercises Q3.2 that $\sum_n \mathbb{P}(|X_n| > n) \leq \mathbb{E}[|X_n|] + 1 < \infty$, where the last equality follows by assumption (note, since the sum starts at $n = 1$ a standard integral approximation of the sum means we can drop the extra $+1$ here). So by the first Borel-Cantelli lemma (Lemma 5.24) we have $\mathbb{P}(X_n = Y_n \text{ e.v. }) = 1$ (See this weeks exercises sheet to check the details). Hence there exists some $n_0(\omega)$ such that $X_n(\omega) = Y_n(\omega)$ for all $n \geq n_0$ and so

$$\left| \frac{S_n}{n} - \frac{T_n}{n} \right| \to 0 \quad \text{a.s.}\,,$$

again see the exercises for the details. □

Proof of Theorem 7.12. As usual, by considering $X_n = X_n^+ - X_n^-$ it is enough to consider the case of $X_n \geq 0$. By the previous lemma it is sufficient to show that $T_n/n \to \mu$ almost surely. The first key step then is, with the aim of applying the first Borel-Cantelli lemma again, to use Chebyshev's inequality to control how far T_n/n is from it's mean. Fixing parameters so that everything works out nicely later; fix $\varepsilon > 0$, $\alpha = 1 + \varepsilon > 1$ and $k_n = \lfloor \alpha^n \rfloor$ (note that by construction $k_n \to \infty$ as $n \to \infty$). Then by Chebyshev's inequality

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|T_{k_n} - \mathbb{E}[T_{k_n}]| > \varepsilon k_n\right) \leq \varepsilon^{-2} \sum_{n=1}^{\infty} \mathrm{Var}[T_{k_n}] k_n^{-2}\,.$$

Now, using the fact that the X_i's and hence Y_i's are *pairwise* independent, we can rewrite the variance on the right hand side as the sum of variances (the "cross terms" which are covariances are all zero). Hence

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|T_{k_n} - \mathbb{E}[T_{k_n}]| > \varepsilon k_n\right) \leq \varepsilon^{-2} \sum_{n=1}^{\infty} k_n^{-2} \sum_{m=1}^{k_n} \mathrm{Var}[Y_m] = \varepsilon^{-2} \sum_{m=1}^{\infty} \mathrm{Var}[Y_m] \sum_{n\,:\,k_n \geq m} k_n^{-2}\,,$$

where on the right hand side we exchanged the order of the sums by Tonelli's Theorem (everything is non-negative). To control the sum on the final right hand side of the

inequality above notice that $\alpha^n \geq k_n = \lfloor \alpha^n \rfloor \geq \alpha^n/2$ for $n \geq 1$, then comparing to a geometric sum

$$\sum_{n\,:\,k_n \geq m} k_n^{-2} \leq 4 \sum_{n\,:\,\alpha^n \geq m} \alpha^{-2n} = 4 \sum_{n \geq n_0} \alpha^{-2n} = \frac{4\alpha^{-2n_0}}{1-\alpha^{-2}} \leq \frac{4\alpha^2}{\alpha^2 - 1} m^{-2}$$

where $n_0 = \lceil \log m / \log \alpha \rceil$. Combining the inequalities above, and using the fact that $\mathrm{Var}[Y_m] \leq \mathbb{E}[Y^2]$

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|T_{k_n} - \mathbb{E}[T_{k_n}]| > \varepsilon k_n \right) \leq \varepsilon^{-2} \frac{4\alpha^2}{\alpha^2 - 1} \sum_{m=1}^{\infty} \mathbb{E}[Y_m^2]\, m^{-2} \,. \tag{7.4}$$

it remains to control the sum on the right hand side above. We do this by finding an upper bound in terms of the expectation of X_1 (recall the definition of $Y_n = X_n \mathbb{1}_{\{|X_n| \leq n\}}$), again using Tonelli's theorem

$$\begin{aligned} \sum_{m=1}^{\infty} \mathbb{E}[Y_m^2]\, m^{-2} &= \sum_{m=1}^{\infty} \mathbb{E}[X_1^2 \,;\, X_1 \leq m]\, m^{-2} \\ &= \mathbb{E}\left[X_1^2 \sum_{m \geq |X_1|}^{\infty} m^{-2} \right] \leq 1 + \mathbb{E}\left[X_1^2 \int_{|X_1|}^{\infty} x^{-2}\, \mathrm{d}x \right] \\ &= 1 + \mathbb{E}[X_1] < \infty \,, \end{aligned} \tag{7.5}$$

where the inequality follows from the 'standard' integral approximation of the sum and we have used the fact that X_1 is non-negative. Combining (7.4) and (7.5) and applying the first Borel-Cantelli theorem we have shown that $|T_{k_n} - \mathbb{E}[T_{k_n}]| > \varepsilon k_n$ for only finitely many n. By the monotone convergence theorem $\mathbb{E}[Y_n] \to \mathbb{E}[X_1]$, and hence $\mathbb{E}[T_{k_n}]/k_n \to \mathbb{E}[X_1] = \mu$. It follows that

$$\frac{T_{k_n}}{k_n} \to \mu \quad \text{a.s.} \quad \text{as } n \to \infty \,.$$

To complete the proof we have to 'fill in' the intermediate values (between the k_n's). This is a fairly standard sandwiching 'trick', for $m \in \{k_n, \ldots, k_{n+1}\}$ we have

$$\frac{1}{(1+2\varepsilon)} \frac{T_{k_n}}{k_n} \leq \frac{T_{k_n}}{k_{n+1}} \leq \frac{T_m}{m} \leq \frac{T_{k_{n+1}}}{k_n} \leq (1+2\varepsilon) \frac{T_{k_{n+1}}}{k_{n+1}}$$

where we used $k_n \geq \alpha^n - 1 \geq (\lfloor \alpha^{n+1} \rfloor / \alpha) - 1 = k_{n+1}/\alpha - 1$ and $\varepsilon k_n > 1 + \varepsilon$ for n sufficiently large, hence $k_{n+1} \geq k_n \geq k_{n+1}/(1+2\varepsilon)$. Since $\varepsilon > 0$ was arbitrary the result follows. □

Chapter 8

Conditional Expectation

Reading: D. Williams, Chapter 9 and A. Klenke, Chapter 8
Further reading: R. Durrett, Probability Theory and Examples, Chapter 5

Probability can be thought of as a measure of ignorance. If there is partial information on the outcome of some random experiment it will decrease our ignorance and change our assessment of probabilities. The concepts of conditional probability and expectation formalise this. We will start by recapping briefly some forms of conditioning you will be familiar with, and then we set up the technology to generalise these into one unifying framework. The definitions and results in this chapter will be central to the next one, that is for the theory of martingales, but much more generally, conditional expectation plays a central rôle in modern probability theory.

8.1 Definitions

Recall from first year probability, if we have a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $B \in \mathcal{F}$ is and event with $\mathbb{P}(B) > 0$, then we can define the conditional probability of an event $A \in \mathcal{F}$, on the event B, by

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}\,. \tag{8.1}$$

This defines a new probability measure (on the same measurable space as the original) but this one gives mass one to the event B and only positive measure to events that have non-empty intersection with B (you might want to check this). We will now introduce an extension of this by 'correctly' defining the conditional expectation. Recall that for an event $A \in \mathcal{F}$ we have $\mathbb{P}[A] = \mathbb{E}[\mathbb{1}_A]$, so that expectations generalise probabilities. So, from this point on, we will prefer to deal with expectations than probabilities of events. The main difficulty that we would like to address with the definition above is that when conditioning on random variables, in many cases, the probability of a random variable taking a specific value is zero, and we can't really handle this case systematically with the definition above. It turns out that we may get round this problem by using generated σ-algebras to represent the 'information' given by a random variable.

We will first develop some intuition from the familiar set up. Firstly, we may immediately extend the definition in Eq. (8.1) to expectations. Suppose we have some integrable random variable X, i.e. $X \in \mathcal{L}^1(\mathbb{P})$, then since $X\mathbb{1}_B \in \mathcal{L}^1(\mathbb{P})$ we have $X \in \mathcal{L}^1(\mathbb{P}[\cdot \mid B])$. Hence we can define the conditional expectation of X with respect to $\mathbb{P}[\cdot \mid B]$ in the

obvious way. That is

$$\mathbb{E}[X \mid B] = \int_\Omega X(\omega)\mathbb{P}[\,\mathrm{d}\omega \mid B] = \int_\Omega \frac{X(\omega)\mathbb{1}_B}{\mathbb{P}[B]}\mathbb{P}[\,\mathrm{d}\omega] = \frac{\mathbb{E}[X\mathbb{1}_B]}{\mathbb{P}[B]}\,. \tag{8.2}$$

Notice that this is consistent with Eq. (8.1) in the sense that $\mathbb{P}(A \mid B) = \mathbb{E}[\mathbb{1}_A \mid B]$. It might take a little bit longer to convince yourself that the expression on the right hand side of (8.2) is really the expected value of X with respect to the measure $\mathbb{P}(\cdot \mid B)$ induced by (8.1). The most obvious limitation of this procedure is that we can not condition on events of measure zero in a way that is consistent with our intuition. For example consider a Uniform$[0,1]$ random variable X and random variables $Y_1, Y_2, \ldots, Y_n$ all defined on the same probability space, such that the distributions of the Y's are independent Bernoulli(x) given $X = x$. Then with the above set up we can condition on events of the type $\mathbb{P}[\cdot \mid X \in [a,b]]$ for $a < b$, but how about $\mathbb{P}[Y_1 = Y_2 = \ldots = Y_n \mid X = x]$? Intuitively the later should be x^n, but, using the technology so far, we can not express this without taking appropriate limits.

We will work a bit harder to motivate the general definition of the conditional expectation, because the first time students come across it it can be quite mystifying. As we have seen before, it is often helpful to develop intuition in the discrete setting, though we lose important measure theoretic subtleties that arise more generally. Let X and Y be discrete random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (see Definition 3.27), i.e. there exist countable collections $\{x_n\}_{n\geq 1}$ and $\{y_n\}_{n\geq 1}$ such that

$$\mathbb{P}(X = x_i),\ \mathbb{P}(Y = y_j) > 0, \quad i,j \in \mathbb{N} \quad \text{and} \quad \sum_i \mathbb{P}(X = x_i) = \sum_i \mathbb{P}(Y = y_i) = 1$$

It is clear that the events $\{Y = y_j\}$ partition Ω. Informally, reporting the value(s) of Y might provide some information about X. So, we want to define the conditional expectation of X given Y, denoted $\mathbb{E}[X \mid Y]$. Notice that this depends on the value Y takes, so intuitively will also be a random variable (a measurable function of Y). In this simple setting we can use the formalism discussed so far to make this rigorous. That is we can define the random variable $\mathbb{E}[X \mid Y] : \Omega \to \mathbb{R}$ by

$$Z(\omega) = \mathbb{E}[X|Y](\omega) = \mathbb{E}[X \mid \{Y = y_i\}] \quad \text{if} \quad Y(\omega) = y_j.$$

Notice that Z is a random variable that is determined by the value of Y for each $\omega \in \Omega$, so that Z is $\sigma(Y)$-measurable (this will be very important later). Following Eq (8.2) we should have

$$Z(\omega) = \mathbb{E}[X \mid \{Y = y\}] = \frac{\mathbb{E}[X\mathbb{1}_{\{Y=y\}}]}{\mathbb{P}[Y = y]} \quad \text{if} \quad Y(\omega) = y.$$

It turns out that to avoid dividing by zero it is better to integrate over the set $\{Y = y\}$. The random variable Z is constant on each set in the partition $\{Y = y_j\}_{j\in\mathbb{N}}$, and for each $\omega \in \{Y = y_j\}$ it takes the value $Z(\omega) = \mathbb{E}[X \mid \{Y = y_i\}]$. It follows that

$$\begin{aligned}
\mathbb{E}[Z\,;\{Y = y_i\}] &= \int Z\mathbb{1}_{\{Y=y_i\}}\,\mathrm{d}\mathbb{P} = \int_{\{Y=y_i\}} Z\,\mathrm{d}\mathbb{P} = \mathbb{E}[X \mid \{Y = y_i\}]\mathbb{P}(Y = y_i) \\
&= \sum_{j=1}^n x_j\,\mathbb{P}(X = x_j \mid Y = y_i)\mathbb{P}(Y = y_i) = \sum_{j=1}^n x_j\mathbb{P}(X = x_j, Y = y_i) \\
&= \int_{\{Y=y_j\}} X\,\mathrm{d}\mathbb{P} = \mathbb{E}[X\,;\{Y = y_j\}],
\end{aligned}$$

where we used elementary theory on conditional expectation, in that

$$\mathbb{E}[X \mid Y = y_j] = \sum_{j=1}^{n} x_j \mathbb{P}(X = x_j \mid Y = y_j)\,,$$

which also follows directly from the set-up we just gave. Let $G_j = \{Y = y_j\}$ for each $j \in \mathbb{N}$. The result above can be summarised more compactly as $\mathbb{E}[Z\mathbb{1}_{G_j}] = \mathbb{E}[X\mathbb{1}_{G_j}]$ (or equivalently $\mathbb{E}[Z\,;\,G_j] = \mathbb{E}[X\,;\,G_j]$). If $\{Y = y_j\}$ has probability zero then this relationship just states $0 = 0$, which still isn't very useful, but we are headed in a good direction.

Since the G_j's partition Ω, we have that any $G \in \sigma(Y)$ is a union of disjoint G_j's (this is a special feature of the discrete setting - or in the 'σ-algebra language', of σ-algebras generated by finite partitions). Therefore $\mathbb{1}_G$ is a sum of $\mathbb{1}_{G_j}$'s, so by dominated convergence and linearity of the expectation, we have

$$\mathbb{E}[Z\,;\,G] = \mathbb{E}[Z\mathbb{1}_G] = \int_G Z\,\mathrm{d}\mathbb{P} = \int_G X\,\mathrm{d}\mathbb{P} = \mathbb{E}[X\mathbb{1}_G] = \mathbb{E}[X\,;\,G].$$

It turns out that the two properties we have discussed, that is (1); Z is $\sigma(Y)$-measurable, and (2); $\mathbb{E}[Z\,;\,G] = \mathbb{E}[X\,;\,G]$ for all $G \in \sigma(Y)$, are essentially sufficient to define the conditional expectation much more generally. Note that X itself is only $\sigma(Y)$-measurable if it is a function of Y, and therefore it makes sense that in this case $\mathbb{E}[X \mid Y] = X$, because we know the value of X if we are told Y. Otherwise, the random variable Z is a 'best approximation' of X given only the information contained in Y.

For more on the geometric interpretation you might also want to check out *Conditional Expectations without tears* by Philipp Düren, which is currently available online.

Theorem 8.1 (Kolmogorov 1933)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ (i.e. X is an integrable random variable, $\mathbb{E}[|X|] < \infty$). Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra. Then there exists a random variable Z on $(\Omega, \mathcal{F}, \mathbb{P})$ such that*

(i) $\mathbb{E}|Z| < \infty$ *(i.e. $Z \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$),*

(ii) Z *is $\mathcal{G}$-measurable,*

(iii) $\int_G Z\,\mathrm{d}\mathbb{P} = \int_G X\,\mathrm{d}\mathbb{P}$ *for all $G \in \mathcal{G}$ (i.e. $\mathbb{E}[Z\,;\,G] = \mathbb{E}[X\,;\,G]$).*

Moreover, if $\tilde{Z}$ is another random variable with these properties then

$$Z = \tilde{Z} \text{ a.s., that is } \mathbb{P}(Z = \tilde{Z}) = 1\,.$$

Definition 8.2. If Z satisfies properties (i)-(iii) of Theorem 8.1, then it is called a *version of the conditional expectation* $\mathbb{E}[X \mid \mathcal{G}]$ of X given $\mathcal{G}$, and we write $Z = \mathbb{E}[X \mid \mathcal{G}]$ a.s..

We will often refer to property (*iii*) of Theorem 8.1 as the *defining relation*.

Remark 8.3 (More intuition). For $\omega \in \Omega$ the value of $\mathbb{E}[X \mid \mathcal{G}](\omega)$ does not 'know', or at least it doesn't matter, which exact ω occurred. However, we do know the value of $Y(\omega)$ for each $\mathcal{G}$-measurable random variable Y, equivalently we know exactly which events in $\mathcal{G}$ occurred (we 'know the information' in $\mathcal{G}$), and we average out over everything else. $\mathbb{E}[X \mid \mathcal{G}]$ is "less random" than X ("more constant" on Ω).

Remark 8.4 (Other constructions). There are other, equivalent, constructions of the conditional expectation. In particular it can be defined as a certain $\mathcal{L}^2$ projection. That is for $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$, we can define $\mathbb{E}[X \mid \mathcal{G}]$, as the projection of X onto the linear subspace spanned by $\mathcal{G}$-measurable functions, with respect to he inner product $\langle X, Y \rangle = \mathbb{E}[XY]$ that was discussed briefly in Section 6.3. So $\mathbb{E}[X \mid \mathcal{G}]$ is the 'best approximation' to X in the subspace of $\mathcal{L}^2$ spanned by $\mathcal{G}$-measurable functions. Some work has to be done then to extend this definition to $\mathcal{L}^1$. The advantage of this approach is that the proof of existence of the conditional expectation does not require the Radon-Nikodym Theorem, it is then possible to prove the Radon-Nikodym Theorem using a nice application of martingale convergence. We do not have time to discuss this more here, but see any of the suggested literature if you are interested in more details, in-particular this is the approach take in D. Williams *Probability with Martingales.*

Crucially, a version of the conditional expectation is not stated as a definition, since existence is not immediate. Next lecture, we will first look at some examples and then we establish almost sure uniqueness, finally we look at the more tricky problem of existence.

We discussed some intuition behind the conditional expectation here in the video lectures, and in particular how we can think about $\mathbb{E}[X \mid \mathcal{G}]$ as the conditional expectation of X given the information in the σ-algebra $\mathcal{G}$.

Example 8.5 (Trivial examples)**.** We list some trivial examples:

- IF $\mathcal{G} = \{\emptyset, \Omega\}$ then $\mathcal{G}$ contains no information. If Z is $\mathcal{G}$-measurable then it must be almost surely constant (this is worth checking - it is just 'definition chasing'), so $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$ almost surely. Keep in mind that the left hand side of this equality is a random variable, the right hand side is typically a number - but in this case the almost sure equality is between the random variable on the left and the constant function which is everywhere equal to $\mathbb{E}[X]$.
- In the other extreme, if $\mathcal{G} = \mathcal{F}$ then we have 'complete' information, since any random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ is by definition $\mathcal{F}$ measurable then $\mathbb{E}[X \mid \mathcal{F}] = X$ almost surely.
- If $B \in \mathcal{F}$ is an event with $\mathbb{P}(B) > 0$ then $\sigma(B) = \{\emptyset, B, B^c, \Omega\}$. If Z is $\sigma(B)$-measurable then it must be constant on B (again check). If we let α be the value of Z on B then

$$\mathbb{E}[Z\,;\,B] = \mathbb{E}[X\,;\,B] \quad \text{i.e.} \quad \mathbb{E}[X \mid \sigma(B)](\omega) = \alpha = \frac{\mathbb{E}[X\mathbb{1}_B]}{\mathbb{P}(B)} \ \text{ if } \omega \in B\,.$$

  We have see this more generally for any countable partition of Ω, into sets in $\mathcal{F}$, in the introduction to this chapter, when we were discussing conditioning on discrete random variables.

Remark 8.6 (Notation)*.* Recall that any function which is $\sigma(Y)$-measurable can be written as a function of Y (perhaps not in a nice explicit way). We will often write $\mathbb{E}[X \mid Y]$ for $\mathbb{E}[X \mid \sigma(Y)]$ and $\mathbb{E}[X \mid Y_1, Y_2 \ldots, Y_n]$ for $\mathbb{E}[X \mid \sigma(Y_1, Y_2, \ldots, Y_n)]$.

Example 8.7 (Binomial and Poisson distribution)**.** Suppose $N \sim \text{Poisson}(\lambda)$, and $X \sim \text{Bin}(N, p)$, with joint distribution described by

$$\mathbb{P}(X = k \mid N = n) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } \ 0 \le k \le n\,,$$

so $\mathbb{E}[X \mid N = n] = np$. It follows that $\mathbb{E}[X \mid N] = Np$ a.s., since $\mathbb{E}[Np] = \lambda p < \infty$ and Np is $\sigma(N)$-measurable, and finally

$$\mathbb{E}[Np\,;\,\{N = n\}] = np\,\mathbb{P}(N = n) = \mathbb{E}[X\,;\,\{N = n\}]\,,$$

so Np is a version of $\mathbb{E}[X \mid \sigma(N)]$.

Example 8.8. I would encourage you to make up your own examples, in particular it is useful to continue developing intuition on discrete spaces, for example $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{H, T\}$.

Example 8.9 (Condition $\mathbb{E}$ for p.d.fs, see also Q9.5)**.** If X and Y are random variables with joint probability density function $f_{X,Y}(x, y)$ (i.e. the joint law has a density with respect to the Lebesgue measure on $\mathbb{R}^2$), and $\mathbb{E}[X] < \infty$, then

$$\mathbb{E}[X \mid \sigma(Y)] = \int_{\mathbb{R}} x f_{X|Y}(x \mid Y)\,\mathrm{d}x\,,$$

where

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)\,\mathrm{d}x$$

$$f_{X|Y}(x \mid Y=y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} & \text{if } f_Y(y) \neq 0\,, \\ 0 & \text{otherwise.} \end{cases} \quad \text{and}$$

$$f_{X|Y}(x \mid Y)(\omega) = f_{X|Y}(y \mid Y=y) \quad \text{if } \; Y(\omega) = y\,.$$

Proof of almost sure uniqueness in Theorem 8.1 (of the conditional expectation). Suppose that Z_1 and Z_2 are both versions of the conditional expectation, $\mathbb{E}[X|\mathcal{G}]$, of X given $\mathcal{G}$. Since $Z_1, Z_2 \in m\mathcal{G}$, then $Z_1 - Z_2 \in m\mathcal{G}$, so that $A_\epsilon := \{Z_1 - Z_2 > \epsilon\} \in \mathcal{G}$, for each $\epsilon > 0$. By property *(iii)*, we have that,

$$\mathbb{E}[X\mathbb{1}_{A_\varepsilon}] = \mathbb{E}[Z_1\mathbb{1}_{A_\varepsilon}] = \mathbb{E}[Z_2\mathbb{1}_{A_\varepsilon}]\,, \quad \text{so} \quad 0 = \mathbb{E}[Z_1 - Z_2; A_\epsilon] \geq \epsilon\mathbb{P}(A_\epsilon),$$

were the final inequality follows from the fact that $Z_1 - Z_2 > \varepsilon$ of A_ε. This implies that $\mathbb{P}(A_\epsilon) = 0$. By symmetry, we can see that $\mathbb{P}(Z_2 - Z_1 > \varepsilon) = 0$, and hence $\mathbb{P}(|Z_1 - Z_2| > \epsilon) = 0$. Therefore, by continuity from above (monotone convergence for measures Lemma 2.9), it follows that

$$\mathbb{P}(Z_1 = Z_2) = \mathbb{P}(\bigcap_{n=1}^{\infty}\{|Z_1 - Z_2| \leq 1/n\}) = \lim_{n\to\infty}\mathbb{P}(|Z_1 - Z_2| \leq 1/n) = 1,$$

so that $Z_1 = Z_2$ $\mathbb{P}$-a.s. as required. □

For the proof of existence of the conditional expectation in the general cases we will use another important result from measure theory, which generalises the idea of a probability density function (pdf) that you are familiar with. We will come to the precise statement in a moment.

Observe that if $Z \geq 0$ is measurable on $(\Omega, \mathcal{F}, \mathbb{P})$, then we can define a new measure $\mathbb{Q}$ on $(\Omega, \mathcal{F})$ by $\mathbb{Q}(A) = \int_A Z\,\mathrm{d}\mathbb{P} = \mathbb{E}[Z\,;\,A]$. We will spell out how this relates the conditional expectation explicitly for the case of conditioning on a σ-algebra generated by some finite partition. Consider a finite partition $\{G_1, ..., G_m\}$, (for example $G_j = \{Y = y_j\}$ for $j = 1, ..., m$). Let $\mathcal{G} = \sigma(G_1, ..., G_m)$ $(= \sigma(Y))$. We know that $Z \in m\mathcal{G}$, so it is constant on each G_j, thus

$$Z(\omega) = \sum_{j=1}^{m} z_j\mathbb{1}_{G_j}(\omega), \quad \forall\omega \in \Omega,$$

for some real numbers $z_1, ..., z_m$. Also, any $G \in \mathcal{G}$ is a union of disjoint G_j's, so by linearity of expectation, we get that,

$$\begin{aligned}\mathbb{E}[Z;G] &= \mathbb{E}[\sum_{j=1}^{m} z_j\mathbb{1}_{G_j}\sum_{s=1}^{r}\mathbb{1}_{G_{k_s}}] = \sum_{j=1}^{m}\sum_{s=1}^{r} z_j\mathbb{E}[\mathbb{1}_{G_j}\mathbb{1}_{G_{k_s}}] \\ &= \sum_{s=1}^{r} z_{k_s}\mathbb{E}[\mathbb{1}_{G_{k_s}}] = \sum_{s=1}^{r} z_{k_s}\mathbb{P}(G_{k_s}).\end{aligned}$$

Taking $z_j = \frac{\mathbb{E}[X;G_j]}{\mathbb{P}(G_j)}$, $j = 1, ..., m$, yields property *(iii)*. That is $\mathbb{Q}(G) = \mathbb{E}[Z\,;\,G]$ is a re-weighting of the measure $\mathbb{P}$ by the 'value's' of the function $Z = \mathbb{E}[X \mid \mathcal{G}]$.

Is it true that the converse also holds in general? That is, if $\mathbb{Q}$ and $\mathbb{P}$ are measures on $(\Omega, \mathcal{F})$, then there exists a function $Z \geq 0$ such that $\mathbb{Q}(A) = \int_A Z \, \mathrm{d}\mathbb{P}$? **No**, not every measure can arise in this way, since the equality implies that if $\mathbb{P}(A) = 0$ then necessarily $\mathbb{Q}(A) = 0$. However, under this condition, then the Radon-Nikodym theorem does give a converse. This result is of more general importance than just to the conditional expectation that we see here, since, as stated earlier, this generalises the concept of a density function that you have seen before for pdf's (which described absolutely continuous probability measures on $\mathbb{R}$ by 're-weighting' the standard Lebesgue measure).

Definition 8.10 (Absolute continuity)**.** Let $\mathbb{P}$ and $\mathbb{Q}$ be measures on the same measure space $(\Omega, \mathcal{F})$. The measure $\mathbb{Q}$ is said to be *absolutely continuous* with respect to $\mathbb{P}$, written $\mathbb{Q} \ll \mathbb{P}$ if

$$\mathbb{P}(A) = 0 \quad \text{implies} \quad \mathbb{Q}(A) = 0 \quad \forall\, A \in \mathcal{F}\,.$$

Theorem 8.11 (The Radon-Nikodym Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and suppose that $\mathbb{Q}$ is a finite measure on $(\Omega, \mathcal{F})$ with $\mathbb{Q} \ll \mathbb{P}$. Then there exists a function $Z \in (m\mathcal{F})^+$ such that*

$$\mathbb{Q}(A) = \int_A Z \, \mathrm{d}\mathbb{P} = \mathbb{E}[Z\,;\, A] \quad \textit{for each } A \in \mathcal{F}\,.$$

Moreover, Z is $\mathbb{P}$-a.s. unique. The function Z is called the Radon-Nikodym derivative *of $\mathbb{Q}$ with respect to $\mathbb{P}$ (or the density of $\mathbb{Q}$ with respect to $\mathbb{P}$), and is written*

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

We omit the proof, which is beyond the scope of the course, and instead focus on how it can be used to prove existence of the conditional expectation.

Proof of existence. Suppose $X \geq 0$ (if not use linearity and $X = X^+ - X^-$). We want to find a $\mathcal{G}$-measurable function $Z \in \mathcal{L}^1(\mathbb{P})$ such that for each $A \in \mathcal{G}$

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

To this end, we define the set function $\mathbb{Q}\colon \mathcal{G} \to [0, \infty)$ by

$$\mathbb{Q}(A) := \mathbb{E}[X; A], \quad \forall\, A \in \mathcal{G}.$$

Showing that $\mathbb{Q}$ is a measure on $(\Omega, \mathcal{G})$ is Q9.1 on Sheet 9. Since $\mathbb{P}(A) = 0$ implies $\mathbb{Q}(A) = 0$, for each $A \in \mathcal{G}$, means that $\mathbb{Q} \ll \mathbb{P}|_{\mathcal{G}}$, where $\mathbb{P}|_{\mathcal{G}}$ denotes the restriction of the measure $\mathbb{P}$ to $(\Omega, \mathcal{G})$. So, by the Radon-Nikodym theorem (Theorem 8.11), there exists a $\mathcal{G}$-measurable random variable $Z \in \mathcal{L}^1(\mathbb{P}|_{\mathcal{G}})$, such that

$$\mathbb{E}[X; A] = \mathbb{Q}(A) = \int_A Z \, \mathrm{d}\mathbb{P}|_{\mathcal{G}} = \mathbb{E}[Z; A], \quad \forall\, A \in \mathcal{G}.$$

Since $Z \geq 0$ and $\int_\Omega Z \, \mathrm{d}\mathbb{P} = \int_\Omega X \, \mathrm{d}\mathbb{P} < \infty$ we have $Z \in \mathcal{L}^1(\mathbb{P})$.

Note there was a small subtlety in this argument that we glossed over a little, that is we used if Y is $\mathcal{G}$-measurable, then $\int Y \, \mathrm{d}\mathbb{P}|_{\mathcal{G}} = \int Y \, \mathrm{d}\mathbb{P}$. □

In contrast to our discrete examples in this section, it is much harder (or impossible) to write out $\mathbb{E}[X \mid \mathcal{G}]$ explicitly if $\mathcal{G}$ is not generated by a countable partition. It follows from the proof of existence above, that if $\mathcal{I}$ is a π-system that generates $\mathcal{G}$, then it is enough to check the defining relation for all $G \in \mathcal{I}$ (by applying Dynkin's Uniqueness Theorem 2.20 to $\mathbb{Q}$).

Just as with the standard expectation, if $X \geq 0$ then it is possible to extend the definition of the conditional expectation to the case that $X \notin \mathcal{L}^1$ by allowing ∞ as a possible value (extended real valued). The only tricky point to check is uniqueness. Therefore the most important defining relations of the conditional expectation from Theorem 8.1 are (ii) and (iii).

8.2 Properties

In this section, we list most properties of conditional expectation that one needs to know well in order to understand and apply the conditional expectation. More intuition will come with practice for this particular concept, but hopefully the following basic properties also help to clarify how the definition of conditional expectation behaves the way you would expect.

Most of the following properties are obvious, we will separate off some particularly powerful ones. Recall that $\mathbb{E}[X \mid \mathcal{G}]$ is a function on Ω, not a number, and defined up to almost sure equivalence.

Proposition 8.12. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $X, Y \in \mathcal{L}^1(\mathbb{P})$ random variables, $\mathcal{G}, \mathcal{H}$ sub-σ-algebras of $\mathcal{F}$, and $a, b, c \in \mathbb{R}$. Then, the following properties hold:*

1. $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X]$.
2. *If $\mathcal{G} = \{\emptyset, \Omega\}$, then $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$.*
3. *If $X \in m\mathcal{G}$, then $\mathbb{E}[X \mid \mathcal{G}] = X$ a.s.. In the special case of $X \equiv c$ then $\mathbb{E}[c \mid \mathcal{G}] = c$ a.s..*
4. *(Linearity): $\mathbb{E}[aX + bY \mid \mathcal{G}] = a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}]$ a.s..*
5. *(Role of Independence): If $\mathcal{H}$ is independent of $\sigma(\sigma(X), \mathcal{G})$, then $\mathbb{E}[X \mid \sigma(\mathcal{H}, \mathcal{G})] = \mathbb{E}[X \mid \mathcal{G}]$ a.s.. In the special case of $\mathcal{G}$ being independent of $\sigma(X)$, then $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$ a.s..*
6. *(Positivity): If $X \geq 0$, then $\mathbb{E}[X \mid \mathcal{G}] \geq 0$ a.s..*
7. *(Monotonicity): If $Y \in \mathcal{L}^1(\mathbb{P})$ is a random variable such that $X \leq Y$, then*

   $$\mathbb{E}[X \mid \mathcal{G}] \leq \mathbb{E}[Y|\mathcal{G}] \text{ a.s..}$$

8. *(Δ-inequality): $|\mathbb{E}(X \mid \mathcal{G})| \leq \mathbb{E}(|X| \mid \mathcal{G})$ a.s.*

Proof. Most of these properties follow almost immediately from definition:

1. Simply take $\Omega \in \mathcal{G}$ in (iii) of the defining relations, then

   $$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\,;\,\Omega] = \mathbb{E}[X\,;\,\Omega] = \mathbb{E}[X]\,.$$

2. If $\mathcal{G} = \{\emptyset, \Omega\}$ the $\mathbb{E}[X \mid \mathcal{G}]$ is a constant function, so that $\mathbb{E}[X \mid \mathcal{G}] \equiv \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X]$ a.s..

3. This follows immediately from Theorem 8.1, since $X \in m\mathcal{G}$ and $X \in \mathcal{L}^1$ so X is a version of the conditional expectation of X given $\mathcal{G}$.

4. Let $Z_1 = a\,\mathbb{E}[X \mid \mathcal{G}]$ and $Z_2 = b\,\mathbb{E}[Y \mid \mathcal{G}]$. Since both Z_1 and Z_2 are $\mathcal{G}$-measurable, so is $Z = Z_1 + Z_2$. We just need to check the main defining relation (i.e. *(iii)* in Theorem 8.1). Fix $G \in \mathcal{G}$, then by linearity of the expectation

   $$\begin{aligned}\int_G Z\,\mathrm{d}\mathbb{P} = \mathbb{E}[Z_1 + Z_2\,;\,G] &= a\int_G \mathbb{E}[X \mid \mathcal{G}]\,\mathrm{d}\mathbb{P} + b\int_G \mathbb{E}[Y \mid \mathcal{G}]\,\mathrm{d}\mathbb{P} \\ &= a\int_G X\,\mathrm{d}\mathbb{P} + b\int_G Y\,\mathrm{d}\mathbb{P} = \int_G aX + bY\,\mathrm{d}\mathbb{P}\,,\end{aligned}$$

where the second line follows by Theorem 8.1 (*iii*) for $\mathbb{E}[X \mid \mathcal{G}]$ and $\mathbb{E}[Y \mid \mathcal{G}]$. Hence $a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}]$ is a version of $\mathbb{E}[aX + bY \mid \mathcal{G}]$.

5. We only prove the special case. We need to show that, if $\sigma(X)$ and $\mathcal{G}$ are independent, then $\mathbb{E}[X]$ (which is a constant function) is a version of the conditional expectation of X given $\mathcal{G}$. Fix $G \in \mathcal{G}$, by assumption X is independent of $\mathbb{1}_G$, hence

$$\mathbb{E}[\mathbb{E}[X]\,;\, G] = \mathbb{E}[X]\mathbb{P}(G) = \mathbb{E}[X]\,\mathbb{E}[\mathbb{1}_G] = \mathbb{E}[X\mathbb{1}_G] = \mathbb{E}[X\,;\, G],$$

so $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$ a.s.. (Exercises: check the details, i.e. if $\mathcal{G}$ and $\mathcal{H}$ are independent, $X \in \mathcal{G}$ and $Y \in \mathcal{H}$ then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ - try applying the 'standard machine').

6. Note $\{\mathbb{E}[X|\mathcal{G}] < 0\} \in \mathcal{G}$, suppose by contradiction that $\mathbb{P}(\mathbb{E}[X|\mathcal{G}] < 0) > 0$. Then, by property *(iii)*, we have

$$0 > \mathbb{E}[\mathbb{E}[X|\mathcal{G}]; \{\mathbb{E}[X|\mathcal{G}] < 0\}] = \mathbb{E}[X; \{\mathbb{E}[X|\mathcal{G}] < 0\}],$$

which contradicts the fact that $X \geq 0$ a.s..

7. Let $A = \{\mathbb{E}(X \mid \mathcal{G}) > \mathbb{E}(Y \mid \mathcal{G})\} \in \mathcal{G}$. Since $X \leq Y$ we have $\mathbb{E}[(Y - X)\mathbb{1}_A] \geq 0$ and hence $\mathbb{P}(A) = 0$.

8. This follows from linearity and monotonicity.

□

Proposition 8.13 (Conditional convergence theorems)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of random variables and X a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra.*

1. ***cMCT:*** *If $X_n \geq 0$ for all n and $X_n \nearrow X$, then $\mathbb{E}[X_n|\mathcal{G}] \nearrow \mathbb{E}[X \mid \mathcal{G}]$ as $n \to \infty$.*
2. ***cFatou:*** *If $X_n \geq 0$ for all $n \geq 1$ then*

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

3. ***cDCT:*** *If Y is an integrable random variable, $|X_n| \leq Y$ for all n, and $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} X$ then*

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

(Unlike DCT, the limiting conditional expectation is a random variable, so here we need to specify almost sure convergence.)

The proofs all use the defining relation, (*iii*) of Theorem 8.1, to transfer statements about convergence properties of the standard expectation to conditional expectations. We do just one of the proofs by way of examples, the others are left as exercises.

Proof. For each $n \in \mathbb{N}$ let $Y_n = \mathbb{E}[X_n \mid \mathcal{G}]$. Since $X_{n+1} \geq X_n \geq 0$, by monotonicity of the conditional expectation, we have $0 \leq Y_n \nearrow \sup_n Y_n = Y \in m\mathcal{G}$ (where measurability follows from Lemma 3.12). We need to show that Y is a version of $\mathbb{E}[X \mid \mathcal{G}]$. Fix $G \in \mathcal{G}$,

then $X_n \mathbb{1}_G \nearrow X \mathbb{1}_G$, and so by MCT $\mathbb{E}[X_n\,;\,G] \nearrow \mathbb{E}[X\,;\,G]$. Similarly $\mathbb{E}[Y_n\,;\,G] \nearrow \mathbb{E}[Y\,;\,G]$. Now applying the defining relation (iii) to Y_n we have

$$\mathbb{E}[Y_n\,;\,G] = \mathbb{E}[X_n\,;\,G]\,.$$

Finally taking limits

$$\mathbb{E}[Y\,;\,G] = \lim_{n\to\infty} \mathbb{E}[Y_n\,;\,G] = \lim_{n\to\infty} \mathbb{E}[X_n\,;\,G] = \mathbb{E}[X\,;\,G]\,,$$

as required. □

The next two results are extremely useful for manipulating conditional expectations. The first one makes rigorous the idea that if $Y \in m\mathcal{G}$ and we condition on $\mathcal{G}$, i.e. we know which events in $\mathcal{G}$ have and haven't occurred, then we know the value of Y - so it can be 'taken out' of the average.

Lemma 8.14 (Taking out what is known)**.** *Let X and Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with X, Y, and XY all integrable. If $\mathcal{G} \subseteq \mathcal{F}$ is a σ-algebra and Y is $\mathcal{G}$-measurable then $\mathbb{E}[YX \mid \mathcal{G}] = Y\mathbb{E}[X \mid \mathcal{G}]$ a.s..*

Proof. The function $Y\mathbb{E}[X \mid \mathcal{G}]$ is clearly $\mathcal{G}$-measurable, so we must check that it satisfies the defining relation (iii) for $\mathbb{E}[YX \mid \mathcal{G}]$. We show this using the standard machine. Suppose X is non-negative and $Y = \mathbb{1}_A$ for some $A \in \mathcal{G}$. Then, for any $G \in \mathcal{G}$ we have $G \cap A \in \mathcal{G}$ and so by the defining relation for $\mathbb{E}[X \mid \mathcal{G}]$ we have

$$\mathbb{E}[Y\mathbb{E}[X \mid \mathcal{G}]\,;\,G] = \mathbb{E}[\mathbb{1}_A \mathbb{E}[X \mid \mathcal{G}]\,;\,G] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]\,;\,A\cap G] = \mathbb{E}[X\,;\,A\cap G] = \mathbb{E}[YX\,;\,G],$$

so that indeed $Y\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[YX|\mathcal{G}]$ a.s..

We extend to simple functions by using linearity, and then MCT to extend to general non-negative $\mathcal{G}$-measurable Y. Finally for general X and Y decompose into the positive and negative part and use linearity again. □

Proposition 8.15 (Tower property)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X an integrable random variable and $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}$ all σ-algebras. Then*

$$\mathbb{E}\big[\mathbb{E}[X \mid \mathcal{F}_2] \mid \mathcal{F}_1\big] = \mathbb{E}[X \mid \mathcal{F}_1] \text{ a.s.}$$

Proof. Since the left hand side and right hand side are both clearly $\mathcal{F}_1$-measurable it only remains to check the defining relation for $\mathbb{E}[X \mid \mathcal{F}_1]$. Fix $G \in \mathcal{F}_1 \subseteq \mathcal{F}_2$, and apply the defining relation twice,

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

□

Jensen's inequality also extends to the conditional setting.

Proposition 8.16 (Conditional Jensen's inequality)**.** *Suppose taht $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, $X \in \mathcal{L}^1$, and $I \subseteq \mathbb{R}$ is an open interval satisfying $\mathbb{P}(X \in I) = 1$. If $f : I \to \mathbb{R}$ is a convex function such that $f(X) \in \mathcal{L}^1(\mathbb{P})$, and $\mathcal{G} \subseteq \mathcal{F}$ is a σ-algebra, then*

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

In particular, $\|\mathbb{E}(X \mid \mathcal{G})\|_p \leq \|X\|_p$ for each $p \in [1, \infty)$.

Proof. Since f is convex, it is the supremum of countably many affine functions, i.e. there exists a countable sequence $(a_n, b_n)_{n\in\mathbb{N}}$ of points in $\mathbb{R}^2$ such that

$$f(x) = \sup_{n\in\mathbb{N}}(a_n x + b_n), \quad \forall x \in I.$$

Notice that $f(X) \geq a_n X + b_n$ for each $n \in \mathbb{N}$, so by conditional monotonicity we have $\mathbb{E}[f(X) \mid \mathcal{G}] \geq a_n \mathbb{E}[X \mid \mathcal{G}] + b_n$ a.s. for all $n \in \mathbb{N}$. Taking the supremum on the right hand side yields

$$\mathbb{E}[f(X) \mid \mathcal{G}] \geq \sup_{n\in\mathbb{N}}(a_n \mathbb{E}[X \mid \mathcal{G}] + b_n) = f(\mathbb{E}[X \mid \mathcal{G}]) \text{ a.s.}.$$

□

Example 8.17 (Second moment method)**.** Suppose X is a non-negative random variable, i.e. $X \geq 0$, then

$$\mathbb{P}(X > 0) \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}\,.$$

There are simpler ways of proving this result, be we will show that it is also a consequence of conditional Jensen's inequality. Recall, for a random variable Y we write $\mathbb{E}[X \mid Y]$ as short for $\mathbb{E}[X \mid \sigma(Y)]$. Let $A = \{X > 0\}$, then

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

Similarly, and by applying Conditional Jensen's,

$$\mathbb{E}[X^2] = \mathbb{E}\big[\mathbb{E}[X^2 \mid \mathbb{1}_A]\big] \geq \mathbb{E}\big[\mathbb{E}[X \mid \mathbb{1}_A]^2\big] = \mathbb{E}[X \mid A]^2\mathbb{P}(A)\,.$$

Putting these two together gives the desired result.

Chapter 9

Martingales

Reading: D. Williams, Chapter 10 and A. Klenke, Chapter 9
Further reading: R. Durrett, Probability Theory and Examples, Section 5.2

Martingales, and associated sub- and super-martingales, form some of the most important concepts in modern probability theory. A martingale formalises the idea of a 'fair game', that is your expected fortune in the future is always the same as your current fortune. A martingale makes no other assumptions on the process, the 'games' played at each consecutive time do not have to be identically distributed or independent - only fair.

The name *martingale* was popularised in the context of modern probability by J. L. Doob in the 1940's. However, the terms introduction to mathematics goes back to Ville in 1939. Ville was inspired by the gambling strategy called 'the infallible martingale.' The strategy is to keep doubling your stake, on say a fair coin toss, until you win - and then leave the game with a profit as soon as you win one round. The name is rather miss-leading, as we will see later the strategy is in fact not infallible. The origin of the name *martingale* in 'the infallible martingale' is not completely clear. One argument goes that it is inspired by the name for the part of a horses harness that stops it from lifting it's head. This is at least useful to keep in mind, especially since the name given to a process which on average is decreasing is a *supermartingale*, and a *submartingale* increases on average. These terms may seem backward at first sight, but if you think of them in terms of the strap on the horses harness, the supermartingale is 'overachieving' in that the head goes down, a 'submartingale' under achieves - the head may rise up.

Much of the usefulness of the theory of martingales is due to a few key powerful and general results. The first of these that we will see (the optional sampling theorem) says roughly "you can't make any money betting on a martingale" (slightly more precisely the (sub/super)martingale property still holds at any stopping time). The second one (submartingale convergence) heuristically says that submartingales are the stochastic analogue of non-decreasing sequences; if they are bounded then they must converge. There are also a wide range of maximum inequalities, and we will also touch on how martingales play a basics role in stochastic calculus.

There are very many applications of the theory we develop in this chapter, generally in studying stochastic processes, but not least in mathematical finance. To borrow from A. Klenke *"the momentousness of the following concept will become manifest only gradually."*

To avoid too many technicalities, we will focus on discrete-time theory, but many of the results extend almost immediately to continuous time, and certainly much of the intuition extends directly.

9.1 Definitions and basic results

We start with some technical definitions and setup required for defining stochastic processes rather generally.

Definition 9.1 (Filtration)**.** A *filtration* (of σ-algebras) on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is an increasing sequence of σ-algebras, $(\mathcal{F}_n)_{n\geq 0}$, such that $\mathcal{F}_n \subseteq \mathcal{F}$ and $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ for each $n \in \mathbb{N}_0$. Moreover, in this case, $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$ is called a filtered probability space (or filtered space for short).

Usually n is interpreted as a time parameter, and $\mathcal{F}_n$ as the knowledge (or information) accumulated by time n. Note that in this interpretation we never forget anything, knowledge is accumulated. We typically start from 'time zero' i.e. $n = 0$, but this is not always the case, so take a little care.

Definition 9.2 (Adapted stochastic process)**.** A (discrete-time) *stochastic process* is a sequence, $(X_n)_{n\geq 0}$, of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A stochastic process is called *integrable* if for each $n \in \mathbb{N}_0$ we have $X_n \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. We say that a stochastic process $(X_n)_{n\geq 0}$ is *adapted* to the filtration $(\mathcal{F}_n)_{n\geq 0}$ if for each $n \in \mathbb{N}_0$, the random variable X_n is $\mathcal{F}_n$-measurable.

The notion of a martingale we see soon will depend on the filtration chosen, the filtration does matter! On the other hand, given some sequence of random variables $(X_n)_{n\geq 1}$, there is a sort of 'default' filtration we can choose. The natural filtration associated with $(X_n)_{n\geq 1}$ is the smallest filtration to which the sequence is adapted.

Definition 9.3 (Natural filtration)**.** The *natural filtration*, $(\mathcal{G}_n)_{n\geq 0}$, associated with the stochastic process $(X_n)_{n\geq 0}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is given by

$$\mathcal{G}_n = \sigma(X_0, X_1, \ldots, X_n) = \sigma\Big(\bigcup_{i=0}^{n} \sigma(X_i)\Big) \quad \text{for each } n \geq 0\,.$$

A stochastic process is automatically adapted to its own natural filtration. Throughout this section we will keep returning to the following example.

Example 9.4 (SSRW)**.** As a running example we will keep in mind the symmetric simple random walk. Here symmetric means it has equal chance to go up or down, and simple means it changes by ± 1 each time step. The walk simply makes i.i.d. jumps up or down by ± 1 at each step. We define the SSRW by $X_0 = 0$ and

$$\mathbb{P}(X_{n+1} = X_n \pm 1 \mid X_0, X_1, \ldots, X_n) = \frac{1}{2}\,,$$

with the natural filtration $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$. In this case X_n is given by the partial sum of i.i.d random variables Y_i such that $\mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i = -1) = 1/2$, so that $X_n = Y_1 + Y_2 + \ldots + Y_n$ for $n \geq 1$. Then $\{X_1 = k\}, \{X_n = k\}, \{X_1 = i_1, \ldots, X_n = i_n\} \in \mathcal{F}_n$ for $n \geq 1$, but $\{X_{n+1} = k\} \notin \mathcal{F}_n$.

In the example above, for $m < n$, by the properties of conditional expectation discussed in the last lecture; namely linearity of conditional expectation, taking out what is known, and independence, we have $\mathbb{E}[X_n \mid \mathcal{F}_m] = X_m + \sum_{i=m+1}^{n} \mathbb{E}[Y_i] = X_m$. In the following definition is (due to J. L. Doob) we use this equation to define the concept of a martingale without insisting on the independence and identically distributed nature of the increments that we have in these examples.

Definition 9.5 (Martingale, (sub/super)martingale)**.** Let $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathbb{P})$ be a filtered space. An integrable, $(\mathcal{F}_n)$-adapted, stochastic process $(X_n)_{n\geq 0}$ is

1. a *martingale* (with respect to $(\mathcal{F}_n)$) if $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) = X_n$ a.s. for each $n \in \mathbb{N}_0$,
2. a *submartingale* if $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \geq X_n$ a.s. for each $n \in \mathbb{N}_0$,
3. a *supermartingale* if $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \leq X_n$ a.s. for each $n \in \mathbb{N}_0$.

If we think of X_n as your 'fortune accumulated by time n' when making some series of bets, then a martingale is a *fair game*, in the sense that $\mathbb{E}(X_{n+1} - X_n \mid \mathcal{F}_n) = 0$ a.s.. A submartingale is a favourable game to you, and a supermartingale is an unfavourable game to you. This may sound like the names 'sub' and 'super' are the wrong way round, but this is how they have stuck and it is now universally accepted. As mentioned in the introduction to this chapter, a good way to remember is to think back to the idea of a martingale being something that prevents the horses head from rearing up, i.e. it keeps the horses head 'in check'. A submartingale may not achieve its goal on average, the horse wants to rear-up and indeed it is able. A supermartingale might over-achieve, the horses head is forced down on average. The name convention also relates to sub and super harmonic functions which you may have seen in an analysis course. Indeed, a function which is subharmonic with respect to the transition kernel of a Markov process indeed gives rise to a submartingale - this leads nicely into the rather deep connections between Markov process and potential theory (both discrete and in the continuum), but that is beyond the scope of this module.

There are two important facts about martingales that we will examine in more detail. The first is that you can't make money betting on them (see Theorem 9.30 below), and the second relates to convergence. In particular, to borrow from *R. Durrett*, Probability Theory and Examples, submartingales are in a sense the stochastic analogue of non-decreasing sequences, that is if they are bounded (if $\sup_n \mathbb{E}X^+ < \infty$) then they converge almost surely. See the section below on Martingale Convergence. First we look at some elementary properties that follow from the definitions...

Proposition 9.6 (Elementary properties)**.** *Suppose $(X_n)_{n\geq 0}$ is an adapted stochastic process on a filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathbb{P})$.*

1. *$(X_n)_{n\geq 0}$ is a submartingale with respect to $(\mathcal{F}_n)$ if and only if $(-X_n)$ is a supermartingale with respect to $(\mathcal{F}_n)$.*
2. *$(X_n)_{n\geq 0}$ is a martingale if and only if it is a submartingale and a supermartingale.*
3. *(Martingales are constant on average) If $(X_n)_{n\geq 0}$ is a martingale, then $\mathbb{E}(X_n) = \mathbb{E}(X_0)$ for each $n \in \mathbb{N}_0$.*
4. *If $(X_n)_{n\geq 0}$ is a submartingale and $n \geq m$, then*

$$\mathbb{E}(X_n \mid \mathcal{F}_m) \geq X_m \text{ a.s.}$$

and

$$\mathbb{E}(X_n) \geq \mathbb{E}(X_m)\,.$$

The same conclusion holds in the case of a supermartingale with the inequalities reversed.

Proof. 1. an 2. are relatively straightforward and left as short exercises. 3. is a special case of 4.. It remains to prove 4.

Fix $m \geq 0$, we shall proceed by induction on n. The base case is $n = m$, then since X_m is $\mathcal{F}_m$-measurable (since the process is adapted to $(\mathcal{F}_n)_{n\geq 0}$) we have $\mathbb{E}(X_m \mid \mathcal{F}_m) = X_m$ a.s., so the conclusion holds. For $n \geq m$ we have $\mathcal{F}_m \subseteq \mathcal{F}_n$, so by the tower property of expectation

$$\mathbb{E}(X_{n+1} \mid \mathcal{F}_m) = \mathbb{E}\big(\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \mid \mathcal{F}_m\big) \geq \mathbb{E}(X_n \mid \mathcal{F}_m) \geq X_m \,,$$

where the first inequality holds a.s. since (X_n) is a submartingale, and the second inequality holds a.s. by the inductive hypothesis.

To deduce the final part simply take expectations (over Ω) in the inequality above. □

Insisting on being clear about the filtration may seem fussy at first, but actually it really matters. Clearly the following statement about sub-martingales holds equally well for (super)martingales.

Proposition 9.7. *Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a filtered space, and $(\mathcal{F}_n)$ and $(\mathcal{F}'_n)$ are filtration's such that $\mathcal{F}'_n \subseteq \mathcal{F}_n$ for each n. If $(X_n)_{n\geq 0}$ is adapted to $(\mathcal{F}'_n)$ and is a submartingale with respect to $(\mathcal{F}_n)$, then $(X_n)_{n\geq 0}$ is a submartingale with respect to the smaller filtration $(\mathcal{F}'_n)_{n\geq 0}$. In particular if $(X_n)_{n\geq 0}$ is a submartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$ then it is always a submartingale with respect to its natural filtration $(\mathcal{G}_n)_{n\geq 0}$.*

Proof. By the tower property of the expectation

$$\mathbb{E}(X_{n+1} \mid \mathcal{F}'_n) = \mathbb{E}(\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \mid \mathcal{F}'_n) \geq \mathbb{E}(X_n \mid \mathcal{F}'_n) = X_n \text{ a.s.} \,,$$

where in the inequality we used the fact that $(X_n)_{n\geq 0}$ is a submartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$ and the final equality we used X_n is $\mathcal{F}'_n$-measurable. □

On the other hand, you should keep in mind the following **warning** that indicates how the filtration is really important. It is clear that if $(X_n)_{n\geq 0}$ and $(Y_n)_{n\geq 0}$ are both martingales with respect to *the same* filtration $(\mathcal{F}_n)_{n\geq 0}$, then the sum-process $(X_n+Y_n)_{n\geq 0}$ is also a martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$ (just by linearity of conditional expectation). **However**, it is relatively simple to cook up examples where $(X_n)_{n\geq 0}$ is a martingale with respect to its natural filtration, $(Y_n)_{n\geq 0}$ is a martingale with respect to its natural filtration, but $(X_n + Y_n)_{n\geq 0}$ is not a martingale with respect to its natural filtration. For example take two simple random walks, $(X_n)_{n\geq 0}$, and $(Y_n)_{n\geq 0}$, defined on the same probability space, generate by a sequence of independent fair 'coin tosses' $\omega_1, \omega_2, \ldots$ (taking values ± 1), such that

$$\begin{aligned} X_n &= \omega_1 + \omega_2 + \ldots + \omega_n \,, \\ Y_n &= \omega_2 + \omega_1 + \ldots \omega_n \,. \end{aligned}$$

Then $\mathbb{E}(X_2 \mid X_1) = X_1$ and $\mathbb{E}(Y_2 \mid Y_1) = Y_1$, but $\mathbb{E}(X_2 + Y_2 \mid X_1 + Y_1) \neq X_1 + Y_1$ (check!).

We now generalise slightly the example of the SSRW, by dropping the 'simple' assumption - that is the steps are independent mean zero, but not necessarily ± 1. This is still a martingale.

Example 9.8 (Sum of independent integrable random variables of mean 0)**.** Let $(Y_n)_{n\geq 0}$ be a sequence of independent integrable random variables with $\mathbb{E}[Y_n] = 0$, for each $n \geq 0$. Then we define the stochastic process $(S_n)_{n\geq 0}$ on a filtered space $(\Omega, \mathcal{F}, (\mathcal{G}_n)_{n\geq 0}, \mathbb{P})$, where $\mathcal{G}_n = \sigma(Y_0, \ldots, Y_n)$, by

$$S_n = \sum_{i=0}^{n} Y_i, \quad \forall n \geq 0\,.$$

Then (S_n) is a martingale with respect to $(\mathcal{G}_n)_{n\geq 0}$. To show this, observe that $(S_n)_{n\geq 0}$ is adapted and integrable with respect to $(\mathcal{G}_n)_{n\geq 0}$, and indeed for any $n \geq 0$ we have

$$\mathbb{E}[S_{n+1} \mid \mathcal{G}_n] = \mathbb{E}\Big[\sum_{i=0}^{n+1} Y_i \mid \mathcal{G}_n\Big] = \mathbb{E}\Big[\sum_{i=0}^{n} Y_i \mid \mathcal{G}_n\Big] + \mathbb{E}[Y_{n+1} \mid \mathcal{G}_n] = \mathbb{E}[S_n \mid \mathcal{G}_n] + \mathbb{E}[Y_{n+1}] = S_n \text{ a.s.},$$

where we used the *linearity* and *role of independence* properties of the conditional expectation, and the fact that $S_n \in m\mathcal{G}_n$. Therefore, $(S_n)_{n\geq 0}$ is indeed a martingale with respect to $(\mathcal{G}_n)_{n\geq 0}$.

In a sense martingales generalise the notion of sums of independent random variables with mean zero. The independent random variables Y_i in the last example can be replaced by martingale differences, which are not necessarily independent.

Definition 9.9 (Martingale Difference)**.** Let $(Y_n)_{n\geq 0}$ be a process on a filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$, which is adapted and integrable with respect to $(\mathcal{F}_n)_{n\geq 0}$. Then, $(Y_n)_{n\geq 0}$ is called a martingale difference, if

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

It is relatively straightforward to check that $(X_n)_{n\geq 0}$ is a martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$ if and only if X_0 is integrable $\mathcal{F}_0$-measurable, and $Y_n = X_n - X_{n-1}$ forms a martingale difference sequences with respect to $(\mathcal{F}_n)_{n\geq 0}$ (you should check this fact).

The following is a case for which the martingale is not a sum of independent random variables.

Example 9.10 (Product of non-negative independent random variables of mean 1)**.** Let $(Y_n)_{n\geq 0}$ be a sequence of independent integrable random variables with $\mathbb{E}[Y_n] = 1$, for each $n \geq 0$. Then we define the stochastic process $(M_n)_{n\geq 0}$ on a filtered space $(\Omega, \mathcal{F}, (\mathcal{G}_n)_{n\geq 0}, \mathbb{P})$, where $\mathcal{G}_n = \sigma(Y_0, \ldots, Y_n)$, by $Y_0 = 1$ and

$$M_n = \prod_{i=1}^{n} Y_i, \quad \forall n \geq 1\,.$$

Then (M_n) is a martingale with respect to $(\mathcal{G}_n)_{n\geq 0}$. Clearly, $(M_n)_{n\geq 0}$ is adapted and integrable with respect to $(\mathcal{G}_n)_{n\geq 0}$, and for any $n \geq 0$ we have that

$$\mathbb{E}[M_{n+1} \mid \mathcal{G}_n] = \mathbb{E}\Big[\prod_{i=0}^{n+1} Y_i \mid \mathcal{G}_n\Big] = \prod_{i=0}^{n} Y_i\, \mathbb{E}[Y_{n+1} \mid \mathcal{G}_n] = M_n\, \mathbb{E}[Y_{n+1}] = M_n \text{ a.s.},$$

where we used 'Taking out what is known' and role of independence properties of conditional expectation. Therefore, $(M_n)_{n\geq 0}$ is indeed a martingale with respect to $(\mathcal{G}_n)_{n\geq 0}$.

You have almost certainly come across Markov processes before, the following example provides a clear cases of a process which is a martingale but not a Markov process.

Example 9.11. Suppose $X_0 \in \mathcal{L}^1(\mathbb{P})$ with $\mathbb{E}[X_0] = 0$, and let $(Y_n)_{n\geq 0}$ be a sequence of independent integrable random variables with $\mathbb{E}[Y_n] = 0$ for each $n \geq 0$ (also independent of X_0). Then

$$X_{n+1} = X_n + Y_n X_0$$

defines a martingale with respect to the natural filtration $\mathcal{F}_n = \sigma(X_0, X_1 \ldots, X_n)$ (check).

The following example can be thought of as accumulating information about a random variable.

Example 9.12. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $(\mathcal{F}_n)$ be a filtration. Suppose X is integrable, then

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

defines a martingale $(X_n)_{n\geq 0}$ with respect to $(\mathcal{F}_n)_{n\geq 0}$. By definition X_n is $\mathcal{F}_n$-measurable, and by the tower property

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

In the previous example we can think about $\mathcal{F}_0, \mathcal{F}_1, \ldots$ as representing unfolding information about the random variable X. We shall see later that many martingales (in fact all uniformly integrable martingales) can be written in the form of the previous example with X replaced by an almost sure limit of the X_n's.

We now cover some important ways of obtaining (sub/super)martingales from other martingales. Firstly suppose $(X_n)_{n\geq 0}$ is a martingale with respect to a filtration $(\mathcal{F}_n)_{n\geq 0}$, and $Y \in m\mathcal{F}_0$, then $(X_n - Y)_{n\geq 0}$ is also a martingale with respect to the same filtration. This is often useful, so that in many situations we can assume without loss of generality that $(X_n)_{n\geq 0}$ is started from $X_0 = 0$.

Proposition 9.13. *Suppose $(X_n)_{n\geq 0}$ is a martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$ and f is a convex function on $\mathbb{R}$. If $f(X_n) \in \mathcal{L}^1$ for each $n \geq 0$ then $(f(X_n))_{n\geq 0}$ is a submartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$.*

Proof. Follows from Conditional Jensen's inequality. Firstly, since $X_n \in m\mathcal{F}_n$ so is $f(X_n) \in m\mathcal{F}_n$, and therefore $(f(X_n))$ is adapted. Now by Jensen's inequality and the martingale property of (X_n),

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

□

It follows immediately that if $(X_n)_{n\geq 0}$ is a martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$ then the following are all submartingales with respect to the same filtration (subject to integrability); $(e^{X_n})_{n\geq 0}$, $(e^{-X_n^2})_{n\geq 0}$ and if $p \geq 1$ then $(|X_n|^p)_{n\geq 0}$.

Similarly to the result above:

Proposition 9.14. *Suppose $(X_n)_{n\geq 0}$ is a submartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$ and f is an* increasing *convex function on $\mathbb{R}$. If $f(X_n) \in \mathcal{L}^1$ for each $n \geq 0$ then $(f(X_n))_{n\geq 0}$ is a submartingale with respect to $(\mathcal{F}_n)_{n\geq 0}$.*

Proof. The proof is the same as the previous proposition. □

Consequently if $(X_n)_{n\geq 0}$ is a submartingale, then $(X_n - a)^+$ is a submartingale. Also, if $(X_n)_{n\geq 0}$ is a supermartingale, then $(X_n \wedge a)$ is a supermartingale.

Exercise 9.15. Give an example of a submartingale (X_n) such that (X_n^2) is a supermartingale. **Hint:** the sequence does not have to be truely random.

We call a process $(V_n)_{n\geq 1}$ a *predictable process* or a *previsible process* if we know the value at of the process 'one step ahead', i.e. we know V_n given the information in the filtration at time $(n-1)$. A good example to have in mind is how much a gambler choose to bet on the n^{th} game in some sequence, they must decide on how much to bet on the n^{th} turn of the game *before* the outcome of the n^{th} turn is revealed, just given the information they have gathered in the first $(n-1)$ turns.

Definition 9.16 (Predictable process)**.** Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 1}, \mathbb{P})$ be a filtered space. A sequence of $(V_n)_{n\geq 1}$ of random variables is called *predictable* (or previsible), with respect to $(\mathcal{F}_n)_{n\geq 0}$, if $V_n \in m\mathcal{F}_{n-1}$ for all $n \geq 1$.

With the interpretation above, thinking of a predictable process $(V_n)_{n\geq 1}$ as a gamblers stake in some sequences of games, it is natural to then think about how much money this gambler would end up making by playing this strategy. Let $(X_n)_{n\geq 0}$ be the net amount of money you would have won by time n if you bet one pound each time, i.e. $X_n - X_{n-1}$ is the net winnings per unit stake in the n^{th} game. Now if you bet according to the strategy (V_n) then your winnings on the n^{th} game will be $V_n(X_n - X_{n-1})$, and your net winnings by time n will be $\sum_{k=1}^n V_k(X_k - X_{k-1})$. This expression defines the martingale transform of (X_n) by (V_n), this is also the discrete analogue of the *stochastic integral* $\int V \, \mathrm{d}X$. Stochastic calculus in general is one of the greatest achievements of modern probability theory.

Theorem 9.17 (Discrete stochastic integral or martingale transform)**.** *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$ be a filtered probability space. Let $(X_n)_{n\geq 0}$ be a (super)martingale with respect to $(\mathcal{F}_n)$ and suppose that $(V_n)_{n\geq 1}$ is a (respectively non-negative) predictable process with respect to $(\mathcal{F}_n)$, and let*

$$(V \bullet X)_n = \sum_{k=1}^{n} V_k(X_k - X_{k-1}) \, .$$

Then, assuming it is integrable, $\big((V \bullet X)_n\big)_{n\geq 0}$ is a (super)martingale with respect to $(\mathcal{F}_n)$. We call $\big((V \bullet X)_n\big)_{n\geq 0}$ the martingale transform *of (X_n) by (V_n).*

Proof. We prove the martingale case (the super martingale case follows the same argument). For $k \leq n$, we have $V_k, X_k, X_{k-1} \in m\mathcal{F}_n$ so $(V \bullet X)_n$ is $\mathcal{F}_n$-measurable. Also, by linearity of the conditional expectation and taking out what is known

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

□

The results says that you can't "beat the system." Consider the "infallible martingale" mentioned in the introduction to this chapter, that is suppose $\xi_n = X_n - X_{n-1}$ is a series of fair coin tosses (that is ± 1 with probability $1/2$) and let $V_1 = 1$ and for $n \geq 2$ if $\xi_{n-1} = 1$ then $V_n = 2V_{n-1}$ and if $\xi_{n-1} = -1$ then $V_n = 1$. Then we double our bet each time we loose, until we win - and hence when we win we regain our losses, so when we first win our winnings will be 1. It looks like it provides us with a "sure thing", but the previous theorem tells us that in fact (unless we have unlimited wealth) we can not gain on average.

2022: A 'compensation example' is covered in the live lecture to motivate the following result

In order to be able to easily extend results of the following two sections from martingales to sub and supermartingales, we will first take a very quick look at the Doob's decomposition theorem. The continuous time analogue of this result (the Doob-Meyer decomposition) is a deep result which is fundamental in stochastic-integral theory.

Theorem 9.18 (Doob's Decomposition Theorem)**.** *Let $(X_n)_{n\geq 0}$ be an adapted integrable process on a filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$. Then $(X_n)_{n\geq 0}$ has a Doob decomposition*

$$X_n = X_0 + M_n + A_n\,,$$

where $(M_n)_{n\geq 0}$ is a martingale (w.r.t $(\mathcal{F}_n)_{n\geq 0}$), and $(A_n)_{n\geq 0}$ is predictable (w.r.t $(\mathcal{F}_n)_{n\geq 0}$), and $M_0 = A_0 = 0$. The Doob decomposition is essentially unique (up to almost sure equivalence). Also, $(X_n)_{n\geq 0}$ is a submartingale if an only if $(A_n)_{n\geq 0}$ is increasing ($A_{n+1} \geq A_n$ a.s.).

Remark 9.19. It follows simply by switching signs that $(X_n)_{n\geq 0}$ is a supermartingale if an only if $(A_n)_{n\geq 0}$ is decreasing.

Proof. Let

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

and

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

Then by construction $M_n + A_n = X_n - X_0$ so $X_n = X_0 + M_n + A_n$. The k^{th} summand in the definition of A_n is $\mathcal{F}_{k-1}$ measurable and so A_n is $\mathcal{F}_{n-1}$ measurable and therefore (A_n) is predictable. Also,

$$\mathbb{E}[M_n - M_{n-1} \mid \mathcal{F}_{n-1}] = \mathbb{E}[X_n - \mathbb{E}\big[X_n \mid \mathcal{F}_{n-1}\big] \mid \mathcal{F}_{n-1}\big] = 0$$

so (M_n) is a martingale.

It remains to prove that the decomposition is essentially unique. Assume $X_0 + M' + A'$ is another Doob decomposition, and hence $M_n - M'_n = A_n - A'_n$ for each n. Since the process $(A_n - A'_n)$ is predictable and $(M_n - M'_n)$ is a martingale they must both be predictable martingales. Any predictable martingale is almost surely constant (exercises).

For the submartingale part, simply observe $\mathbb{E}[X_{n+1} - X_n \mid \mathcal{F}_n] = A_{n+1} - A_n$ a.s., this completes the proof. □

The following definition goes beyond the scope of this module, but it turns out it provides an extremely powerful tool for studying square integrable martingales.

Definition 9.20 (The angle-bracket process)**.** Let $(X_n)_{n\geq 0}$ be a square integrable (i.e. $\mathbb{E}[|X_n|^2] < \infty$ for each n or equivalently $X_n \in \mathcal{L}^2$) martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$. Then $\mathbb{E}[X_n^2]$ is a submartingale by Proposition 9.13. The essentially unique increasing predictable process (A_n) for which $(X_n^2 - A_n)_{n\geq 0}$ is a martingale, and $A_0 = 0$, is called the *square variation* of $(X_n)_{n\geq 0}$ and is denoted by $(\langle X\rangle_n)_{n\geq 0}$.

Note that, by definition, $\mathbb{E}[X_n^2] = \mathbb{E}[X_0^2] + \mathbb{E}[A_n]$ and since X_n is a martingale $A_{n+1} - A_n = \mathbb{E}[X_{n+1}^2 - X_n^2 \mid \mathcal{F}_n] = \mathbb{E}[(X_{n+1} - X_n)^2 \mid \mathcal{F}_n]$ (check this calculation), so the increments of A_n are exactly the conditional variances of the martingale difference sequence (the increments of the martingale).

9.2 Stopping times

We saw in the previous section that the martingale property is 'stable' under transformation by fairly general predictable processes. Much of the power of martingales comes from the fact that the martingales property is preserved if we stop the process at suitable random times (another form of stability).

You may have seen the concept of a stopping time before in the context of other modules on stochastic processes. Intuitively a *stopping time* is a random time that we can recognise when it arrives, without looking into the future. Standard examples are given in terms of stocks, or option prices. For example the first time a stock falls by 3% in a single day, or the first time that a stock reaches some fixed value, are both examples of stopping times. However, the time in March that a stock reaches its highest value is not a stopping time. In this case we have to wait until the end of March to know for sure when the random time occurred, at which point the time may have already passed (at the time that the time occurred you did not know it unless you could see into the future). We now make this concept precise on a filtered space.

Definition 9.21 (Stopping time)**.** Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$ be a filtered space. A map $\tau : \Omega \to \mathbb{N}_0 \cup \{\infty\}$ is called a stopping time, with respect to $(\mathcal{F}_n)_{n\geq 0}$, if

$$\{\tau \leq n\} \in \mathcal{F}_n, \quad \text{for each } n \geq 0.$$

Equivalently (in discrete time) if

$$\{\tau = n\} \in \mathcal{F}_n, \quad \text{for each } n \geq 0.$$

Note that equivalence between the two requirements in Definition 9.21 simply follows by the monotonincity of the filtration $(\mathcal{F}_n)_{n\geq 0}$ (check). Stopping times are sometimes called *optional times*.

Example 9.22. The canonical example of a stopping time is the first time that a process $(X_n)_{n\geq 0}$ reaches some measurable set $A \in \mathcal{B}(\mathbb{R})$, that is

$$\tau_A = \inf\{n \geq 0 : X_n \in A\}\,,$$

is a stopping time with respect to the natural filtration of (X_n). It is intuitively clear that τ_A is a stopping time since we can determine by observations up to time t whether the process has entered A, i.e. the event $\{\tau_A \leq t\}$. More formally $\{X_s \in A\} \in \mathcal{F}_s \subseteq \mathcal{F}_t$, for each $s \leq t$, by definition of the natural filtration. Hence,

$$\{\tau_A \leq t\} = \bigcup_{s=0}^{t} \{X_s \in A\} \in \mathcal{F}_t\,.$$

Similarly the k^{th} return-time, defined inductively by

$$\tau_A^{(1)} = \inf\{n \geq 1 \,:\, X_n \in A\}\,,$$
$$\tau_A^{(k)} = \inf\{n > \tau_A^{(k-1)} : X_n \in A\}, \quad k \in \mathbb{N}\,,$$

is also a stopping time. On the other hand, the last visit time to A, given by $\tilde{\tau}_A = \sup\{n \geq 0 \,:\, X_n \in A\}$ is not in general a stopping time, you have to look into the future to know if it has occurred.

Recall the notation $n \wedge \tau = \min\{n, \tau\}$.

Lemma 9.23. *Suppose σ and τ are stopping times, then*

(i) $\sigma \vee \tau$ and $\sigma \wedge \tau$ are both stopping times.

(ii) $\sigma + \tau$ is also a stopping time.

Note that, for $s \geq 0$ the time $\tau - s$ is not in general a stopping time.

Proof. Exercises. □

For a stopping time τ we define the *stopped process* $(X_{n\wedge\tau})_{n\geq 0}$ by

$$X_{n\wedge\tau} = \begin{cases} X_n, & n \leq \tau \\ X_\tau, & n > \tau \end{cases}.$$

Remark 9.24. Note that if $(X_n)_{n\geq 0}$ is an integrable process, then for each $n \geq 0$

$$\mathbb{E}[|X_{n\wedge\tau}|] \leq \mathbb{E}\big[\max_{m\leq n}|X_m|\big] \leq \mathbb{E}[|X_0|] + \ldots \mathbb{E}[|X_n|] < \infty\,,$$

so the stopped process is also integrable.

Definition 9.25 (Pre-τ σ-algebra)**.** Let τ be a stopping time on a filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$. The pre-$\tau$ σ-algebra, denoted by $\mathcal{F}_\tau$, is given by

$$\mathcal{F}_\tau = \{A \in \mathcal{F} :\ \forall n \geq 0\,,\ A \cap \{\tau \leq n\} \in \mathcal{F}_n\}.$$

By construction, $\mathcal{F}_\tau$ is indeed a σ-algebra on Ω (we have to use the fact that τ is a stopping time to prove this fact (check)). Intuitively, $\mathcal{F}_\tau$ contains all the information accumulated up to time τ, where τ maybe random. If τ is deterministic, i.e. $\tau = t$ for some $t \geq 0$, we observe that $\mathcal{F}_\tau = \mathcal{F}_t$, which follows clearly from the definition (and is reassuring in terms of consistency). It is also relatively straightforward to check that τ is $\mathcal{F}_\tau$-measurable. We give two more properties of the pre-τ σ-algebra below, which are both stated with respect to a fixed filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathbb{P})$.

Lemma 9.26. *If σ and τ are stopping times and $\sigma \leq \tau$, then $\mathcal{F}_\sigma \subseteq \mathcal{F}_\tau$.*

Proof. Fix $A \in \mathcal{F}_\sigma$ and $n \geq 0$. Then, since $\{\tau \leq n\} \subseteq \{\sigma \leq n\}$ (by assumption $\sigma \leq \tau$), we have

$$A \cap \{\tau \leq n\} = (A \cap \{\sigma \leq n\}) \cap \{\tau \leq n\} \in \mathcal{F}_n\,,$$

because $A \cap \{\sigma \leq n\} \in \mathcal{F}_n$ (since $A \in \mathcal{F}_\sigma$) and $\{\tau \leq n\} \in \mathcal{F}_n$ (since τ is a stopping time). □

Lemma 9.27. *If $(X_n)_{n\geq 0}$ is adapted and τ is a stopping time which is almost surely finite, then X_τ is $\mathcal{F}_\tau$-measurable.*

Proof. It is sufficient to check that $\{X_\tau \leq x\} \in \mathcal{F}_\tau$ for each $x \in \mathbb{R}$ (why? Check). Fix $x \in \mathbb{R}$ and $n \geq 0$. The

$$\{X_\tau \leq x\} \cap \{\tau \leq n\} = \bigcup_{i=0}^{n} \big(\{\tau = i\} \cap \{X_i \leq x\}\big) \in \mathcal{F}_n\,,$$

because $\{X_i \leq x\} \in \mathcal{F}_i$ (since $(X_n)_{n\geq 0}$ is adapted to $(\mathcal{F}_n)$), $\{\tau = i\} \in \mathcal{F}_i$ (since τ is a stopping time) and $\mathcal{F}_i \subseteq \mathcal{F}_n$ for each $i \leq n$. □

9.3 The Optional Stopping Theorem

Reading: D. Williams, Chapter 10 and A. Klenke, Chapter 10
Further reading: R. Durrett, Probability Theory and Examples, Section 5.2

The following proposition, sometimes called 'Optional stopping,' states that a stopped martingale is still a martingale. It is in fact a special case of the more general results in the next section, see Theorem 9.30. It also can be shown to follow directly from Theorem 9.17 (see D. Williams Chapter 10)

sort out referencing

Proposition 9.28. *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$ be a filtered space, $(X_n)_{n\geq 0}$ a (sub/super)martingale and τ a stopping time (both with respect to $(\mathcal{F}_n)_{n\geq 0}$). Then the stopped process, $(X_{n\wedge\tau})_{n\geq 0}$, is also a (sub/super)martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$.*

Proof. We only prove the case that (X_n) is a submartingale, the others follow by considering $-X$. Note that the integrability condition for the stopped process is satisfied by Remark 9.24.

Let (X_n) be a submartingale and fix $n \geq 0$. First observe that by Lemma 9.27 we have $X_{n\wedge\tau} \in m\mathcal{F}_{n\wedge\tau}$, and by Lemma 9.26 $\mathcal{F}_{n\wedge\tau} \subseteq \mathcal{F}_n$, hence $(X_{\tau\wedge n})_{n\geq 0}$ is adapted to $(\mathcal{F}_n)_{n\geq 0}$. Then since $(X_{\tau\wedge n})_{n\geq 0}$ is adapted to $(\mathcal{F}_n)_{n\geq 0}$ and $\{\tau > n-1\} \in \mathcal{F}_{n-1}$, we have

$$\begin{aligned}
\mathbb{E}[X_{\tau\wedge n} \mid \mathcal{F}_{n-1}] - X_{\tau\wedge(n-1)} &= \mathbb{E}[X_{\tau\wedge n} - X_{\tau\wedge(n-1)} \mid \mathcal{F}_{n-1}] \\
&= \mathbb{E}[(X_n - X_{n-1})\mathbb{1}_{\{\tau > n-1\}} \mid \mathcal{F}_{n-1}] \\
\text{(taking out what's known)} \quad &= \mathbb{1}_{\{\tau>n-1\}}\mathbb{E}[X_n - X_{n-1} \mid \mathcal{F}_{n-1}] \\
\text{(since } (X_n)_{n\geq 0} \text{ is a submart)} \quad &\geq 0\,,
\end{aligned}$$

where we used that $X_{\tau\wedge n} - X_{\tau\wedge(n-1)} = 0$ on the event $\{\tau \leq n-1\}$. □

The previous proposition, together with the elementary properties of martingales (Proposition 9.6), tells us that if $(M_n)_{n\geq 0}$ is a martingale and τ is a stopping time then $(M_{\tau\wedge n})_{n\geq 0}$ is also a martingale and hence $\mathbb{E}[M_{\tau\wedge n}] = \mathbb{E}[M_0]$. We would like to know if martingale properties like this hold at the stopping time itself. That is, can we take the limit as $n \to \infty$ and have a result like $\mathbb{E}[M_\tau] = \lim_{n\to\infty} \mathbb{E}[M_{\tau\wedge n}] = \mathbb{E}[M_0]$. It turns out that this is *not* true in general. However it is true under some reasonable assumptions. Let's first look at what can go wrong.

Example 9.29 (Back to the SSRW)**.** Let $(Y_k)_{k\geq 1}$ be a sequence of i.i.d. random variables with $\mathbb{P}(Y_k = 1) = \mathbb{P}(Y_k = -1) = 1/2$ [Asside: this type of random variable is obvious closely related to a Bernoulli random variable, in fact Y_k is equal to $2B_k - 1$ in distribution if $B_k \sim \text{Ber}(1/2)$. The (Y_k) are sometimes called Rademacher random variables]. Let $M_0 = 0$ and $M_n = \sum_{i=1}^n Y_i$ for $n \geq 1$. Then (M_n) is the position of a SSRW stated at the origin, after n steps. In particular $(M_n)_{n\geq 0}$ is a martingale, hence $\mathbb{E}[M_n] = 0$ for each n.

Now let $\tau = \min\{n \geq 0 : M_n = 1\}$ which is defined almost surely, $\mathbb{P}(\tau < \infty) = 1$ since the SSRW on $\mathbb{Z}$ is recurrent. It is clear that τ is a stopping time (it is a hitting time) and by definition $M_\tau = 1$. However $\mathbb{E}[M_\tau] = 1 \neq 0 = \mathbb{E}[M_0]$.

What went wrong? It turns out that τ is 'too big' - i.e. $\mathbb{E}[\tau] = \infty$ (the SSRW on $\mathbb{Z}$ is *null* recurrent).

It turns out that if we impose suitable boundedness assumptions then we can avoid the problem we saw in the previous example, so that $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$. This is part of the celebrated Optional Stopping Theorem (OST). There are many variants of this result and its precise statement.

Theorem 9.30 (Doob's Optional Stopping Theorem)**.** *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$ be a filtered space, $(X_n)_{n\geq 0}$ a supermartingale and τ, σ stopping times with $\sigma \leq \tau$. Suppose any of the following conditions holds:*

(i) τ and σ are bounded, i.e. there exists $N \in \mathbb{N}$ such that $0 \leq \sigma(\omega) \leq \tau(\omega) \leq N$ for all $\omega \in \Omega$.

(ii) τ, σ are almost surely finite and $(X_n)_{n\geq 0}$ is uniformly bounded, i.e. there is some $K > 0$ such that $|X_n(\omega)| \leq K$ for every $n \geq 0$ and $\omega \in \Omega$.

(iii) $\mathbb{E}[\tau], \mathbb{E}[\sigma] < \infty$, and the increments of $(X_n)_{n\geq 0}$ are uniformly bounded, i.e. there is some $K > 0$ such that $|X_n - X_{n-1}| \leq K$, for all $n \geq 0$.

Then X_τ is integrable and

$$\mathbb{E}[X_\tau \mid \mathcal{F}_\sigma] \leq X_\sigma \text{ a.s.}$$

Remark 9.31. We often apply the theorem above with $\sigma = 0$. In this case, if $(X_n)_{n\geq 0}$ is a supermartingale and τ a stopping times that satisfy one of *(i)-(iii)*, then by taking expectation on both sides of the inequality $\mathbb{E}[X_\tau|\mathcal{F}_0] \leq X_0$ a.s., we have

$$\mathbb{E}[X_\tau] \leq \mathbb{E}[X_0]\,.$$

In the special case that $(X_n)_{n\geq 0}$ is a martingale then, if any of $(i) - (iii)$ hold, X_τ is integrable and

$$\mathbb{E}[X_\tau|\mathcal{F}_\sigma] = X_\sigma \text{ a.s.} \quad \text{and} \quad \mathbb{E}[X_\tau] = \mathbb{E}[X_0]\,.$$

Corollary 9.32 (OST for UI martingales)**.** *Suppose $(X_n)_{n\geq 1}$ in the previous theorem is also uniformly integrable, then we only need that the stopping times are almost surely finite for the conclusion to hold.*

In lectures we proved the Optional Stopping Theorem for bounded stopping times only. You can find the other cases in the suggested literature. Before we prove Theorem 9.30 we first prove the following lemma which will be important in the proof.

Lemma 9.33. *Let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ be a random variable and $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$. Then, $\mathbb{E}[X|\mathcal{G}] \geq 0$ a.s., if and only if $\mathbb{E}[X; A] \geq 0$ for each $A \in \mathcal{G}$.*

Proof. Suppose $\mathbb{E}[X \mid \mathcal{G}] \geq 0$ a.s., and fix $A \in \mathcal{G}$, then by the defining relation of conditional expectation

$$\mathbb{E}[X\,;\, A] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]\,;\, A] \geq 0\,.$$

Conversely, suppose that $\mathbb{E}[X; A] \geq 0$ for each $A \in \mathcal{G}$, and for each $n \in \mathbb{N}$ define $B_n := \{\mathbb{E}[X|\mathcal{G}] < -1/n\} \in \mathcal{G}$. Then,

$$0 \leq \mathbb{E}[X\,;\, B_n] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]\,;\, B_n] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]\mathbb{1}_{B_n}] \leq -\frac{1}{n}\,\mathbb{P}(B_n) \leq 0,$$

and hence $\mathbb{P}(B_n) = 0$ for each $n \geq 1$. Since the countable union of null sets is null we conclude that $\mathbb{P}\big(\mathbb{E}[X \mid \mathcal{G}] < 0\big) = 0$, as required. □

Proof of Theorem 9.30 for Bounded Stopping Times. We show this for the supermartingale case (the others follow directly). Firstly note that, since $(X_n)_{n\geq 0}$ is supermartingale it is integrable, and $\tau \leq N$, hence $\mathbb{E}[|X_\tau|] = \mathbb{E}\Big[\big|\sum_{i=0}^N X_i \mathbb{1}_{\{\tau=i\}}\big|\Big] < \infty$, thus the conditional expectation exists. Since X_σ is $\mathcal{F}_\sigma$-measurable, it suffices to show that, $\mathbb{E}[X_\sigma - X_\tau \mid \mathcal{F}_\sigma] \geq 0$ a.s.. In particular, by Lemma 9.33, we equivalently show that $\mathbb{E}[X_\sigma - X_\tau\,;\, A] \geq 0$ for each $A \in \mathcal{F}_\sigma$.

Fix $A \in \mathcal{F}_\sigma$, then by expressing $X_s - X_\tau$ as a telescoping sum

$$\begin{aligned}\mathbb{E}[X_\sigma - X_\tau\,;\, A] &= \mathbb{E}\Big[\sum_{j=1}^N (X_{j-1} - X_j)\,;\, A \cap \{\sigma < j \leq \tau\}\Big] \\ &= \mathbb{E}\Big[\sum_{j=1}^N (X_{j-1} - X_j)\,;\, A \cap \{\sigma < j\} \cap \{j \leq \tau\}\Big] \\ &= \sum_{j=1}^N \mathbb{E}\big[(X_{j-1} - X_j)\,;\, A \cap \{\sigma \leq j-1\} \cap \{\tau > j-1\}\big] \geq 0,\end{aligned}$$

where we used the fact $A \cap \{\sigma \leq j-1\} \in \mathcal{F}_{j-1}$ (by definition of $\mathcal{F}_\sigma$) and $\{\tau > j-1\} \in \mathcal{F}_{j-1}$ (since τ is a stopping times), so by the supermartingale property of $(X_n)_{n\geq 0}$ we have

$$\mathbb{E}\big[(X_{j-1} - X_j)\,;\, A \cap \{\sigma \leq j-1\} \cap \{\tau > j-1\}\big] \geq 0 \quad \text{for each } j \geq 1\,.$$

This completes the proof in case (i).

As sketch proof of the remaining two cases: (ii) Using that the associated stopped processes $(X_{n\wedge\tau})_{n\geq 0}$, $(X_{n\wedge\sigma})_{n\geq 0}$ are bounded and integrable. Given $A \in \mathcal{F}_\sigma$, by dominated convergence we have

Add some more details here

$$\mathbb{E}[X_\sigma - X_\tau\,;\, A] = \lim_{n\to\infty} \mathbb{E}[X_{n\wedge\sigma} - X_{n\wedge\tau}\,;\, A] \geq 0.$$

For (iii), similarly as in (i), express $X_{n\wedge\sigma} - X_{n\wedge\tau}$ as a telescoping sum to get

$$|X_{n\wedge\sigma} - X_{n\wedge\tau}| = \Big|\sum_{j=(n\wedge\sigma)+1}^{n\wedge\tau} X_{j-1} - X_j\Big| \leq K\tau$$

which shows that $X_{n\wedge\sigma} - X_{n\wedge\tau}$ is uniformly dominated and integrable. Hence, by the DCT, given $A \in \mathcal{F}_\sigma$, it follows that,

$$\mathbb{E}[X_\sigma - X_\tau\,;\, A] = \lim_{n\to\infty} \mathbb{E}[X_{n\wedge\sigma} - X_{n\wedge\tau}\,;\, A] \geq 0.$$

□

Sketch proof of Corollary 9.32. We assume without proof the following result: If the family $\{X_n\}_{n\geq 0}$ is uniformly integrable then so is the family $\{X_\sigma : \sigma \text{ a stopping time}\}$. Then the results follows observing:

(i) The result holds for $\sigma \wedge n \leq \tau \wedge n$ for each $n \in \mathbb{N}$.

(ii) Hence $\mathbb{E}[X_{\sigma\wedge n}\,;\, A] = \mathbb{E}[X_{\tau\wedge n}\,;\, A]$.

(iii) Now apply Theorem 6.45: $X_{\sigma\wedge n} \to X_\sigma$ a.s. and $X_{\tau\wedge n} \to X_\tau$ a.s., therefore by uniform integrability $\mathbb{E}[X_\sigma\,;\, A] = \mathbb{E}[X_\tau\,;\, A]$ for any $A \in \mathcal{F}_{\sigma\wedge m}$ (some $m \in \mathbb{N}$).

(iv) Finally, extend to any A in F_σ by a π-system argument.

□

Example 9.34 (Back to SSRW)**.** Recall our previous construction of the simple symmetric random walk on $\mathbb{Z}$, i.e. $M_n = \sum_{i=1}^n Y_i$ where $(Y_i)_{i\geq 1}$ are i.i.d. $\mathbb{P}(Y_i = \pm 1) = 1/2$, and $M_0 = 0$. Let $\tau = \inf\{n \geq 0 : M_n \in \{-a, b\}\}$. Then the stopped process $(M_{n\wedge\tau})_{n\geq 0}$ is bounded and hence uniformly integrable. Also, $\tau < \infty$ a.s., since a run of $a + b$ ones occurs with positive probability in the sequence $(Y_i)_{i\geq 1}$, so we will eventually see such a sequence with probability one. After such a sequence M_n will certainly have either hit $-a$ or b. So, by Corollary 9.32, or Theorem 9.30 (ii), $\mathbb{E}[M_\tau] = \mathbb{E}[M_0] = 0$. Also

$$\mathbb{E}[M_\tau] = b\mathbb{P}(M_\tau = b) - a\mathbb{P}(M_\tau = -a) = b\big(1 - \mathbb{P}(M_\tau = -a)\big) - a\mathbb{P}(M_\tau = -a)\,.$$

We can now solve $\mathbb{E}[M_\tau] = 0$ to find

$$\mathbb{P}(M_\tau = -a) = \frac{b}{a+b}\,.$$

This give us a very simple way to calculate the probability of leaving an interval on a given boundary using the Optional Stopping Theorem.

Example 9.35 (How long before we see ABRACADABRA?)**.** Returning to our monkey we left at the type writer, Example 5.25. We know that the monkey will eventually write ABRACADABRA, in fact infinitely often, but how long do you typically have to wait before you observe this sequence? It turns out that we can cook up an ingenious martingale and apply the Optional Stopping Theorem in order to answer this. [This might be easier to understand if you try to draw a picture of what's going on - se the lectures].

We consider a slightly simpler (to compute) problem, which is analogous, and hopefully it is clear how you could perform the calculation for the problem above. Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space carrying i.i.d. discrete random variables $(X_i)_{i\geq 1}$ with $\mathbb{P}(X_i = j) = p_j > 0$ for each $j \in \mathbb{N}_0$. What is the expected number of random variables we must observe before the consecutive sequence $0, 1, 2, 0, 1$ occurs?

We consider a (imaginary) casino offering fair bets, i.e. the expected gain from each bet is exactly zero. In particular, a gambler betting $£a$ on the outcome of the next random variable being j will lose all their stake of $£a$ with probability $(1 - p_j)$, and they will win $£a/p_j$ with probability p_j (so that the expected amount of money they get back after seeing the j^{th} random variable is the amount they bet $£a$).

Now, imagine a sequence of gamblers betting at this casino, each with an initial fortune of $£1$. Gambler i bets $£1$ that $X_i = 0$, if she wins she bets her entire fortune of $£1/p_0$ that $X_{i+1} = 1$, if she wins she again bets her fortune of $£1/(p_0p_1)$ that $X_{i+2} = 2$, if she wins that bet 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 stops playing having amassed a fortune of $£1/(p_0^2p_1^2p_2)$. If any gambler loses at any point then she will have lost her entire fortune, and goes home empty handed.

Notice that, by the $(i+4)^{th}$ draw, gambler i will have left the casino having lost their initial fortune $£1$ unless between i and $i+4$ the draws were exactly the pattern $0, 1, 2, 0, 1$, in which case they leave with $£1/(p_0^2p_1^2p_2)$.

Let M_n be the casino's winnings after n games, i.e. when X_n has just been revealed, stated from $M_0 = 0$. Then (since the casino is fair) $(M_n)_{n\geq 0}$ is a mean zero martingale with respect to the filtration given by $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Let τ be the number of random variables revealed before we see the desired pattern. Then $\mathbb{E}[\tau] < \infty$ (check). Since at most five people are betting at one time (everyone who started more than 5 turns earlier

have already left the casino), we have $|M_{n+1} - M_n| \leq 5/(p_0^2 p_1^2 p_2)$, so condition (*iii*) of Theorem 9.30 is satisfied.

After X_τ is revealed, gamblers $1, 2, \ldots, \tau - 5$ have each lost £1 (the casino has gained this many pounds).

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

[Again - a picture might help]. Hence

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

By Theorem 9.30 $\mathbb{E}[M_\tau] = 0$, so taking expectations we have

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

Of course, the same trick can be used to calculate the expected time until any fixed finite pattern occurs in a sequence of i.i.d. random variables.

Example 9.36. If we combine all our machine so far, we can use the square variation process (angle-bracket process of Definition 9.20), to calculate the expected hitting time in Example 9.34. Recall $(M_n)_{n\geq 1}$ is a simple symmetric random walk started from the origin and $\tau = \inf\{n \geq 0 : M_n \in \{-a, b\}\}$. Recall from Definition 9.20 and the construction of the predictable process in the proof of Theorem 9.18, so

$$\begin{aligned}
\langle M\rangle_n &= \sum_{k=1}^{n} \left(\mathbb{E}[M_k^2 \mid \mathcal{F}_{k-1}] - M_{k-1}^2\right) \\
&= \sum_{k=1}^{n} \left(\mathbb{E}[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}] - 2M_{k-1}^2 + 2\mathbb{E}[M_{k-1}M_k \mid \mathcal{F}_{k-1}]\right) \\
\text{(taking out what's known)} &= \sum_{k=1}^{n} \left(\mathbb{E}[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}] - 2M_{k-1}^2 + 2M_{k-1}\mathbb{E}[M_k \mid \mathcal{F}_{k-1}]\right) \\
\text{((M_n) is a martingale)} &= \sum_{k=1}^{n} \mathbb{E}[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}]\,.
\end{aligned}$$

Now, observe that the increments of the SSRW are ± 1, so their square is always 1. Hence, $\langle M\rangle_n = n$, and (again by Theorem 9.18) we have $(M_n^2 - n)_{n\geq 0}$ is a martingale. By Proposition 9.28 (a stopped martingale is a martingale)

$$\mathbb{E}[M_{\tau\wedge n}^2 - (\tau \wedge n)] = 0 \quad \text{for all } n \geq 0\,.$$

Finally, applying the monotone convergence theorem gives

$$\mathbb{E}[\tau] = \mathbb{E}[M_\tau^2] = a^2\mathbb{P}[M_\tau = -a] + b^2\mathbb{P}[M_\tau = b] = a \cdot b\,.$$

Observe that this also gives us the hitting time $\tau_{\{b\}}$ of $\{b\}$ by monotone convergence, since $\tau \nearrow \tau_{\{b\}}$ as $a \to \infty$, so $\mathbb{E}[\tau_{\{b\}}] = \lim_{a\to\infty} \mathbb{E}[\tau] = \infty$ (a fact we have already used when $b = 1$).

9.4 Martingale Convergence

Reading: D. Williams, Chapter 11 and A. Klenke, Chapter 11

Under weak conditions, namely non-negativity or uniform integrability, martingales converge almost surely. Also, under weaker assumptions than we have seen so far in the more general setting, this almost sure convergence also comes with convergence in L^p.

We will first examine two main tools in martingale convergence, Doob's maximal (or submartingale) inequality, and secondly Doob's up-crossing lemma.

Recall Markov's inequality, which states

$$\mathbb{P}(|X| \geq \lambda) \leq \frac{\mathbb{E}|X|}{\lambda}\,.$$

It turns out that martingales satisfy a similar, but more powerful inequality, which bounds the maximum of the process.

Theorem 9.37 (Doob's Maximal Inequality, or Doob's Submartingale Inequality)**.** *Let $(X_n)_{n\geq 0}$ be a submartingale, then for fixed $\lambda \geq 0$ and $N \geq 0$*

$$\lambda\,\mathbb{P}\big(\sup_{n\leq N} X_n \geq \lambda\big) \leq \mathbb{E}\big[X_N\,;\,\sup_{n\leq N} X_n \geq \lambda\big] \leq \mathbb{E}[|X_N|\mathbb{1}_{\{\sup_{n\leq N} X_n \geq \lambda\}}] \leq \mathbb{E}[|X_N|]\,.$$

Proof. Let $\tau = \inf\{n \geq 0 : X_n \geq \lambda\} \wedge N$. Then, since the stopped process is a submartingale (see Proposition 9.28), applying the Optional Stopping Theorem for bounded stopping times (Theorem 9.30), $\mathbb{E}[X_N \mid \mathcal{F}_\tau] \geq X_\tau$. So by taking expectation, and applying the law of total expectation

$$\begin{aligned}\mathbb{E}[X_N] \geq \mathbb{E}[X_\tau] &= \mathbb{E}[X_\tau\,;\,\sup_{n\leq N} X_n \geq \lambda] + \mathbb{E}[X_\tau\,;\,\sup_{n\leq N} X_n < \lambda] \\ &\geq \lambda\mathbb{P}\big(\sup_{n\leq N} X_n \geq \lambda\big) + \mathbb{E}[X_N\,;\,\sup_{n\leq N} X_n < \lambda]\,,\end{aligned}$$

since $\tau = N$ on the event $\{\sup_{n\leq N} X_n < \lambda\}$. Subtracting $\mathbb{E}[X_\tau\,;\,\sup_{n\leq N} X_n < \lambda]$ completes the proof. □

Actually, L^p versions exists.

Corollary 9.38 (Doob's L^p-inequality)**.** *Let $(X_n)_{n\geq 0}$ be a martingale or a positive submartingale, then for $N \geq 0$, $\lambda \geq 0$ and $p \geq 1$*

$$\lambda^p\,\mathbb{P}\Big(\sup_{n\leq N} |X_n| \geq \lambda\Big) \leq \mathbb{E}[|X_N|^p]\,.$$

Proof. Note that $f(x) = |x|^p$ is convex for $p \geq 1$, so by Theorem 9.13, $(|X_n|^p)_{n\geq 0}$ is a submartingale. Then, by Theorem 9.37 above, $\lambda^p\,\mathbb{P}\big(\sup_{n\leq N} |X_n|^p \geq \lambda^p\big) \leq \mathbb{E}[|X_N|^p]$ as required. □

Those of you who have seen the reflection principle for the simple symmetric random walk might want to compare the result with Doob's Maximal Inequality above. By drawing a picture and 'counting paths' you should be able to remind yourself of the reflection principle, which states if $X_n = \sum_{k=1}^n Y_k$ for $Y_k \in \{-1, 0, 1\}$ i.i.d. and symmetric, then

$$\mathbb{P}(\sup_{m\leq N} X_m \geq a) = 2\mathbb{P}(X_N \geq a) - \mathbb{P}(X_N = a) \leq \mathbb{E}[X_N]/a\,,$$

where the last inequality follows from Markov's inequality. In a sense Theorem 9.37 can be interpreted as a "grown-up" version of the reflection principle.

Next lecture we come to the Upcrossing Lemma and Martingale Convergence Theorem.

We now move on to up-crossings. **The picture here is extremely useful - see lectures.** Let $(X_n)_{n\geq 0}$ be an integrable process on $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathbb{P})$. Let $a < b \in \mathbb{R}$. Imagine (X_n) is the price of some stock. Take the following trading strategy on the value of (X_n)

(i) Don't buy unless the value drops below a, in which case you think it is undervalued.

(ii) Keep your shares in (X_n) until it gets above b, at which time you think that it is overvalued - take the current value out as cash.

(iii) Return to (i).

Some observations immediate if we take the above strategy;

- No matter how clever this strategy may look, if $(X_n)_{n\geq 0}$ is a supermartingale then the Optional Stopping Theorem tells us that on average we *can't gain* (since the times we start and stop are stopping times).
- Each time the process makes an up-crossing from at or below a, to at or above b, we take away a profit of at least $(b-a)$ (at least since it could make jumps - again check the picture).
- From the previous point, if we can bound the maximal profit we expect, then dividing by $(b-a)$ gives an upper bound on the maximal up-crossings.
- Finally, and we will come to this point rigorously very soon, if the number of up-crossings for all $a < b$ is finite, then the price must converge.

The above trading strategy is suggestive of a strategy to prove convergence of (super)martingales. Now let's try to make things rigorous.

Definition 9.39 (Up-crossings)**.** Let $\mathbf{x} = (x_n)_{n\geq 0}$ be a sequence of real numbers and $a, b \in \mathbb{R}$ with $a < b$. Define $\tau_0^{(\mathbf{x})} = 0$, and inductively for $k \geq 1$,

$$\sigma_k^{(\mathbf{x})} = \inf\{n \geq \tau_{k-1}^{(\mathbf{x})} : x_n \leq a\} \text{ and } \tau_k^{(\mathbf{x})} = \inf\{n \geq \sigma_k^{(\mathbf{x})} : x_n \geq b\}.$$

The number of up-crossings of the interval $[a, b]$ made by the sequence $\mathbf{x}$ by time $N \in \mathbb{N}$, denoted by $U_N^{(\mathbf{x})}[a, b]$, is

$$U_N^{(\mathbf{x})}[a, b] = \max\{k \geq 0 : \tau_k^{(\mathbf{x})} < N\}.$$

As $N \to \infty$, we have $0 \leq U_N^{(\mathbf{x})}[a, b] \nearrow U^{(\mathbf{x})}[a, b] = \max\{k \geq 0 : \tau_k^{(\mathbf{x})} < \infty\}$.

The following lemma together with Doob's Upcrossing Lemma are the main tools to be used to prove convergence of $\mathcal{L}^1$-bounded (super)martingales.

Lemma 9.40. *A sequence $\mathbf{x} = (x_n)_{n\geq 0}$ of real numbers converges in the extended reals $\bar{\mathbb{R}} = [-\infty, \infty]$, if and only if $U^{(\mathbf{x})}[a, b] < \infty$ for each $a, b \in \mathbb{Q}$ with $a < b$.*

Proof. Note that $\mathbf{x}$ converges if and only if $\liminf x_n = \limsup x_n$. We proceed by considering the contrapositive.

Assume that there exist $a, b \in \mathbb{Q}$ with $a < b$ such that $U^{(\mathbf{x})}[a,b] = \infty$. Then, since $\mathbf{x}$ is infinitely often below a and above b, we have

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

hence $\mathbf{x}$ does not converge.

Conversely suppose that $\mathbf{x}$ does not converge, i.e. $\liminf_{n\to\infty} x_n < \limsup_{n\to\infty} x_n$. Since $\mathbb{Q}$ is dense in $\mathbb{R}$, there exists $a, b \in \mathbb{Q}$ with $a < b$ such that,

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

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

Lemma 9.41 (Doob's Upcrossing Lemma)**.** *Let $\mathbf{X} = (X_n)_{n\ge 0}$ be a (super)martingale on some filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\ge 0}, \mathbb{P})$ and $a, b \in \mathbb{R}$ with $a < b$. Then, given $N \in \mathbb{N}$,*

$$\mathbb{E}[U_N^{(\mathbf{X})}[a,b]] \le \mathbb{E}[(a - X_N)^+]/(b-a)\,,$$

where $(a - X_N)^+ = (a - X_N)\mathbb{1}_{\{X_N \le a\}} \ge 0$ is the non-negative part of $(a - X_N)$.

Proof. Fix $N \in \mathbb{N}$, and set $\tau'_k = N \wedge \tau_k^{(\mathbf{X})}$, and $\sigma'_k = N \wedge \sigma_k^{(\mathbf{X})}$, for each $k \ge 1$. Notice that $U_N^{(\mathbf{X})}[a,b] = \max\{k \ge 0 : \tau'_k < N\}$. By construction, $\tau'_k \ge \sigma'_k$ are bounded stopping times, so by Doob's Optional Stopping theorem (Theorem 9.30)

$$\mathbb{E}\big[X_{\tau'_k} \mid \mathcal{F}_{\sigma'_k}\big] \le X_{\sigma'_k} \quad \text{a.s.,} \quad \text{for each } \ k \ge 1.$$

By definition of σ_k and σ'_k we have

$$\begin{aligned} 0 \ge (X_{\sigma'_k} - a)\mathbb{1}_{\{\sigma'_k < N\}} &\ge \mathbb{E}[(X_{\tau'_k} - a)\mathbb{1}_{\{\sigma'_k < N\}} \mid \mathcal{F}_{\sigma'_k}] \\ &= \mathbb{E}[(X_{\tau'_k} - a)\mathbb{1}_{\{\sigma'_k < \tau'_k < N\}} + (X_{\tau'_k} - a)\mathbb{1}_{\{\sigma'_k < N = \tau'_k\}} \mid \mathcal{F}_{\sigma'_k}] \\ &\ge \mathbb{E}[(b-a)\mathbb{1}_{\{\sigma'_k < \tau'_k < N\}} + (X_N - a)\mathbb{1}_{\{\sigma'_k < N = \tau'_k\}} \mid \mathcal{F}_{\sigma'_k}], \quad \text{a.s.}\,, \end{aligned}$$

where in the second line we used $\tau'_k < N$ or $\tau'_k = N$, and in the third $X_{\tau'_k} \ge b$ on the event $\{\tau'_k < N\}$. Rearranging and taking expectation on both sides yields

$$\mathbb{E}[(a - X_N)\mathbb{1}_{\{\sigma'_k < N = \tau'_k\}}] \ge (b-a)\mathbb{P}(\sigma'_k < \tau'_k < N) = (b-a)\mathbb{P}(U_N^{(\mathbf{X})} \ge k),$$

where we used the fact that $\{\sigma'_k < \tau'_k < N\} = \{U_N^{(\mathbf{X})}[a,b] \ge k\}$ (convince yourself that this holds). We bound above the left hand side as

$$\begin{aligned} \mathbb{E}[(a - X_N)\,;\, \{\sigma'_k < N = \tau'_k\}] &\le \mathbb{E}[(a - X_N)^+\,;\, \{\sigma'_k < N = \tau'_k\}] \\ &\le \mathbb{E}[(a - X_N)^+\,;\, \{U_N^{(\mathbf{X})}[a,b] = k-1\}]\,, \end{aligned}$$

where the last inequality follows from $\{\sigma'_k < N = \tau'_k\} = \{U_N^{(\mathbf{X})}[a,b] = k-1\}$, since the k^{th} up-crossing has begun, since $\sigma'_k < N$, but has not finished before N, since $\tau'_k = N$. Putting everything together gives,

$$\mathbb{E}\big[(a - X_N)^+\,;\, \{U_N^{(\mathbf{X})}[a,b] = k-1\}\big] \ge (b-a)\mathbb{P}\big(U_N^{(\mathbf{X})}[a,b] \ge k\big).$$

Summing over $k \ge 1$ and applying $\mathbb{E}[Z] = \sum_{k\ge 1} \mathbb{P}(Z \ge k)$ for non-negative integer random variables (see Q3.4(a) on Exercise Sheet 3), it follows that

$$\mathbb{E}\big[(a - X_N)^+\big] \ge (b-a)\,\mathbb{E}\big[U_N^{(\mathbf{X})}[a,b]\big].$$

□

A supermartingale $(X_n)_{n\geq 0}$ is just a random real sequence. By Doob's Upcrossing Lemma we can bound the expected number of up-crossings for each $a < b$. If we can use this bound to say that the number of upcrossings of any rational interval is bounded in expectation then the number of upcrossings must be finite (almost surely), since otherwise the expectation would be infinite. We now makes this argument precise to get almost sure convergence for $(X_n)_{n\geq 0}$, if the sequence is bounded in $\mathcal{L}^1$.

Theorem 9.42 (Doob's Forward Convergence Theorem)**.** *Let $(X_n)_{n\geq 0}$ be a (super/sub)martingale on a filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$, that is bounded in $\mathcal{L}^1(\mathbb{P})$, i.e. $\sup_{n\geq 0}\mathbb{E}[|X_n|] < \infty$. Then, $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} X_\infty \in L^1(\Omega, \mathcal{F}_\infty, \mathbb{P})$ (recall $\mathcal{F}_\infty := \sigma(\mathcal{F}_n : n \geq 0)$).*

Proof. Suppose that $\mathbf{X} = (X_n)_{n\geq 0}$ is a supermartingale (consider $(-X_n)_{n\geq 0}$ to cover the submartingale case). Fix $N \in \mathbb{N}$ and $a, b \in \mathbb{Q}$ with $a < b$. Then, by Doob's Upcrossing Lemma, we have

$$\mathbb{E}\big[U_N^{(\mathbf{X})}[a,b]\big] \leq \frac{\mathbb{E}\big[(a - X_N)^+\big]}{b-a} \leq \frac{\mathbb{E}[|X_N| + |a|]}{b-a} \leq \frac{|a| + \sup_{n\geq 0}\mathbb{E}[|X_n|]}{b-a} < \infty\,,$$

where the second inequality follows from the standard triangle-inequality. By construction, $0 \leq U_N^{(\mathbf{X})}[a,b] \nearrow U^{(\mathbf{X})}[a,b]$, so by the monotone convergence theorem

$$\mathbb{E}\big[U^{(\mathbf{X})}[a,b]\big] = \lim_{N\to\infty}\mathbb{E}\big[U_N^{(\mathbf{X})}[a,b]\big] \leq \frac{|a| + \sup_{n\geq 0}\mathbb{E}[|X_n|]}{b-a} < \infty,$$

which implies that $\mathbb{P}\big(U^{(\mathbf{X})}[a,b] = \infty\big) = 0$ [since if a random variable is infinite with positive probability it must have infinite mean]. Now, since a countable union of null sets is null, and $\mathbb{Q}$ is countable,

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

Equivalently,

$$\mathbb{P}(\text{for any } a, b \in \mathbb{Q},\ a < b \text{ it holds } U^{(\mathbf{X})}[a,b] < \infty) = 1\,.$$

Hence, by Lemma 9.40, $\mathbb{P}(X_n \to X_\infty) = 1$ where $X_\infty := \liminf_{n\to\infty} X_n$, which is in $m\mathcal{F}_\infty$ by Lemma 3.12. It remains to show that $X_\infty \in \mathcal{L}^1(\mathbb{P})$. Since $|X_n| \to |X_\infty|$ almost surely, then Fatou's lemma gives

$$\mathbb{E}[|X_\infty|] = \mathbb{E}[\liminf |X_n|] \leq \liminf_{n\to\infty}\mathbb{E}[|X_n|] \leq \sup_{n\geq 0}\mathbb{E}[|X_n|] < \infty\,,$$

where the final inequality holds by assumption on $(X_n)_{n\geq 0}$. □

Corollary 9.43. *Let $(X_n)_{n\geq 0}$ be a non-negative supermartingale on some filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$. Then, $X_n \xrightarrow{\mathbb{P}\text{-a.s.}} X_\infty \in m\mathcal{F}_\infty$ for some random variable $X_\infty \in \mathcal{L}^1(\mathbb{P})$, where $\mathcal{F}_\infty := \sigma(\mathcal{F}_n : n \geq 0)$.*

Proof. Combining the supermartingale property and non-negativity of $(X_n)_{n\geq 0}$ yields,

$$\mathbb{E}[|X_n|] = \mathbb{E}[X_n] \leq \mathbb{E}[X_0] < \infty, \quad \forall n \geq 0,$$

i.e. $\sup_{n\geq 0}\mathbb{E}[|X_n|] < \infty$. The result now follows from Doob's Forward Convergence Theorem. □

Remark 9.44. The same holds for any lower bound, not necessarily zero, by considering a constant shift. Also a similar result holds for submartingales bounded above (by considering $(-X_n)_{n\geq 0}$ above).

Example 9.45 (re-re-return to SSRW)**.** Recall Example 9.29. We let $(M_n)_{n\geq 0}$ be a simple symmetric random walk on $\mathbb{Z}$, started from $M_0 = 0$. Let $\tau = \inf\{n \geq 0 : M_n = 1\}$, then τ is a stopping time (everything is with respect to the natural filtration of (M_n)). So $(M_{\tau\wedge n})_{n\geq 0}$ is a martingale, hence a submartingale, by Lemma 9.28, and by definition it is bounded above by 1. Therefore (by the preceding remark), $(M_{\tau\wedge n})_{n\geq 0}$ converges almost surely, with slight abuse of notation $M_{\tau\wedge n} \to X_\infty$. Also, by Corollary 9.43 (and the Remark 9.44), the almost sure limit, X_∞, is integrable and hence almost surely finite. Then we must have $X_\infty = 1$ almost surely, otherwise there is some ω such that $M_{\tau\wedge\infty}(\omega) = k \neq 1$ but if $M_{n\wedge\tau}(\omega) = k$ then $M_{\tau\wedge(n+1)} = k \pm 1$ which contradicts convergence. $M_{\tau\wedge n} \to 1$ almost surely as $n \to \infty$. Does $(M_{n\wedge\tau})$ converge in any other sense?

Although we have convergence almost surely to X_∞, so $\mathbb{E}[X_\infty] = 1$, we know from Proposition 9.28 that $\mathbb{E}[M_{\tau\wedge n}] = \mathbb{E}[M_0] = 0$, hence $M_{n\wedge\tau}$ does not converge in $\mathcal{L}^1$ (in mean). The problem is one which by now we are familiar with, there is a small chance of $M_{n\wedge\tau}$ getting very large ('mass runs off to infinity'). As we have seen before, a natural condition to consider now is uniform integrability (UI). We will take a look at UI martingales soon.

Example 9.46 (The Galton-Watson branching process)**.** We now take a look at a somewhat more interesting example of the behaviour above. Recall the construction of the Galton-Watson process in Section 1.3. Let X be a non-negative integer valued random variable with $0 < \mu = \mathbb{E}[X] < \infty$, and $\{X_r^{(n)} : n, r \geq 1\}$ an array of i.i.d. random variables each with the same distribution as X. Let $Z_0 = 1$ and define inductively the n^{th} generation size to be

$$Z_n = \sum_{r=1}^{Z_{n-1}} X_r^{(n)} \quad \text{for } n \geq 1\,.$$

Finally, following the logic in Section 1.3, we let $M_n = Z_n/\mu^n$ and let $(\mathcal{F}_n)_{n\geq 0}$ be the natural filtration of $(Z_n)_{n\geq 0}$.

We would like to show that $(M_n)_{n\geq 0}$ is a martingale. Since the process is discrete we know that $\mathcal{F}_n$ is given by all the possible finite or countable unions of sets in some countable partition of Ω, i.e. the union of sets that look like $\{Z_1 = z_1, Z_2 = z_2, \dots Z_n = z_n\}$. In the discrete setting, using this idea, we showed in the Section 8.1 that the conditional expectation in terms of σ-algebras was consistent with the 'basic' definition you have used before, so using standard arguments you could show that $(M_n)_{n\geq 0}$ is a martingale, it would be a little painful though. We would now like to use the power of our more general definition to make this precises without having to fall back on the 'old' definition of conditional expectation and do many calculations.

For convenience we will let

$$\mathcal{F}_n = \sigma(\{X_r^{(j)}\,;\, j \leq n, r \geq 1\})\,.$$

First note that $Z_n = \sum_{r=1}^{\infty} \mathbb{1}_{\{r\leq Z_{n-1}\}} X_r^{(n)}$ (the usual trick for dealing with sums with a

random number of terms). Then by cMCT (see Proposition 8.13) we have

$$\begin{aligned}
\mathbb{E}[Z_n \mid \mathcal{F}_{n-1}] &= \mathbb{E}\left[\lim_{N\to\infty} \sum_{r=1}^{N} \mathbb{1}_{\{r\le Z_{n-1}\}} X_r^{(n)} \mid \mathcal{F}_{n-1}\right] \\
\text{(cMCT)} \quad &= \sum_{r=1}^{\infty} \mathbb{E}[\mathbb{1}_{\{r\le Z_{n-1}\}} X_r^{(n)} \mid \mathcal{F}_{n-1}] \\
\text{(taking out what's known)} &= \sum_{r=1}^{\infty} \mathbb{1}_{\{Z_{n-1}\ge r\}} \mathbb{E}[X_r^{(n)} \mid \mathcal{F}_{n-1}] \\
\text{(independence)} &= \sum_{r=1}^{\infty} \mathbb{1}_{\{Z_{n-1}\ge r\}} \mathbb{E}[X_r^{(n)}] \\
\text{(identically distributed)} &= \sum_{r=1}^{\infty} \mathbb{1}_{\{Z_{n-1}\ge r\}} \mu = Z_{n-1}\,\mu\,,
\end{aligned}$$

and hence $\mathbb{E}[M_n \mid \mathcal{F}_{n-1}] = M_{n-1}$.

Since $(M_n)_{n\ge 0}$ is a non-negative martingale (and hence a supermartingale), by Corollary 9.43 we know that $(M_n)_{n\ge 0}$ converges almost surely to a finite limit M_∞. Does in converge in any other sense? By the argument in Section 1.3 the only candidate for the almost sure limit when $\mu < 1$ is the constant random variable 0. To prove this rigorously you could, for example, use the Bounded Convergence Theorem and consider the Laplace transform, for $\lambda > 0$ (recall the definition of G_n from Section 1.3)

$$\mathbb{E}[e^{-\lambda M_\infty}] = \lim_{n\to\infty} \mathbb{E}[e^{-\lambda M_n}] = \lim_{n\to\infty} \mathbb{E}[e^{-(\lambda/\mu^n) Z_n}] = \lim_{n\to\infty} G_n(e^{-\lambda/\mu^n}) = 1\,,$$

since $e^{-\lambda M} \le 1$ (pointwise on Ω) for any non-negative random variable M. The distribution of any non-negative random variable is uniquely determined by it's Laplace transform, and the only random variable who's Laplace transform is identically 1 is zero. Alternatively you could use "convergence in probability fast enough implies almost sure convergence" (Lemma 6.10). Notice $Z_n \ge 1$ on the event $Z_n > 0$ (since the process takes values in $\mathbb{N}$), so

$$\mathbb{P}(Z_n > 0) \le \mathbb{E}[Z_n\,;\, Z_n > 0] = \mathbb{E}[Z_n] = \mu^n\,,$$

and the right hand side is summable if $\mu < 1$. Therefore $\mathbb{E}[M_\infty] = 0$ but $\mathbb{E}[M_n] = 1$ for each n (since it is a martingale), so again we do not have convergence in $\mathcal{L}^1$ - and hence in $\mathcal{L}^p$ for any $p \ge 1$

Again, the way to prohibit the bad behaviour observed in the previous two examples is to insist on uniform integrability. First we will consider the benefits of $\mathcal{L}^2$ boundedness.

9.5 Martingales bounded in $\mathcal{L}^2$

Reading: D. Williams, Chapter 12 and A. Klenke, Chapter 11

In practice it is often relatively straightforward to check that a martingale is bounded in $\mathcal{L}^2$. We have seen that $\mathcal{L}^2$ forms a nice space, L^2 is a Hilbert space, and the associated inner products are associated with covariances and (as a special case) variances of random variables. So, when it works, one of the easiest ways to show that a martingale is bounded in $\mathcal{L}^1$, and also to ensure that it converges in $\mathcal{L}^1$, is to prove that it is bounded in $\mathcal{L}^2$.

We start by showing why boundedness in $\mathcal{L}^2$ of a martingale is often easy to check. Suppose for now that $(M_n)_{n\geq 0}$ is square-integral, i.e. $\mathbb{E}[M_n^2] < \infty$ for each n. Notice, *this is not the same* as saying that (M_n) is bounded in $\mathcal{L}^2$, since the sequence $\mathbb{E}[M_n^2]$ could diverge - we do not require that $\sup_n \mathbb{E}[M_n^2] < \infty$ which is what we mean by $(M_n)_{n\geq 0}$ is bounded in $\mathcal{L}^2$. Firstly we observe that martingale increments have zero co-variance, i.e. they are orthogonal in $\mathcal{L}^2$, fix $k > j \geq 0$ and for convenience let $M_{-1} = 0$, then

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

This allows us to write down a 'Pythagorean formula' (c.f. $\|a+b\|^2 = \|a\|^2 + \|b\|^2$ if a and b are orthogonal)

$$\begin{aligned}
\mathbb{E}[M_n^2] &= \mathbb{E}\left[\left(\sum_{k=0}^{n}(M_k - M_{k-1})\right)^2\right] \quad \text{(telescopic sum)} \\
\text{(expanding)} &= \sum_{k=0}^{n}\mathbb{E}[(M_k - M_{k-1})^2] + 2\sum_{n\geq k>j\geq 0}\mathbb{E}[(M_k - M_{k-1})(M_j - M_{j-1})] \\
\text{(zero covariance)} &= \mathbb{E}[M_0^2] + \sum_{k=1}^{n}\mathbb{E}[(M_k - M_{k-1})^2] \qquad (9.1)
\end{aligned}$$

Similarly, by the same argument but 'starting from' M_n, for $m > n \geq 0$ we have

$$\mathbb{E}[(M_m - M_n)^2] = \sum_{k=n+1}^{m}\mathbb{E}[(M_k - M_{k-1})^2]\,. \qquad (9.2)$$

So we can check if a martingale is bounded in $\mathcal{L}^2$ by looking at the variance of the interments.

Lemma 9.47. *Let $(M_n)_{n\geq 0}$ be a martingale, then it is bounded in $\mathcal{L}^2$ if and only if*

$$\mathbb{E}[M_0^2] < \infty \quad \textit{and} \quad \sum_{k=1}^{\infty}\mathbb{E}[(M_k - M_{k-1})^2] < \infty \qquad (9.3)$$

Proof. First assume (9.3) holds, then we can check that $\mathbb{E}[M_n^2]$ is finite for each n by induction and considering the partial sums of $\sum_{k=1}^{\infty}\mathbb{E}[(M_k - M_{k-1})^2]$. Then the result follows from (9.1). The other direction follows immediately from (9.1). □

We now see that Doob's Martingale Convergence Theorem gives us convergence in $\mathcal{L}^2$, and hence $\mathcal{L}^1$, when we consider martingales bounded in $\mathcal{L}^2$.

Theorem 9.48 ($\mathcal{L}^2$ martingale convergence)**.** *Let $(M_n)_{n\geq 0}$ be a martingale on $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$. Suppose that $(M_n)_{n\geq 0}$ is bounded in $\mathcal{L}^2$ (i.e. $\sup_n \mathbb{E}[M_n^2] < \infty$), then there exists a random variable $M_\infty \in \mathcal{L}^2$ such that*

$$M_n \to M_\infty \quad \textit{almost surely and in } \mathcal{L}^2 .$$

Proof. By Jensen's inequality,

$$\mathbb{E}[|M_n|]^2 \leq \mathbb{E}[M_n^2] \quad \text{hence} \quad \sup_n \mathbb{E}[|M_n|] < \infty .$$

Hence we can apply Doob's Forward Convergence Theorem (Theorem 9.42), so $M_n \xrightarrow{\mathbb{P}-\text{a.s.}} M_\infty \in \mathcal{L}^1$. It remains to prove that the convergence also holds in $\mathcal{L}^2$ (which implies convergence in $\mathcal{L}^1$). We use the Pythagorean formula 9.2

$$\mathbb{E}[(M_{n+k} - M_n)^2] = \sum_{j=n+1}^{n+k} \mathbb{E}[(M_j - M_{j-1})^2] , \tag{9.4}$$

and so by Fatous Lemma

$$\begin{aligned} \mathbb{E}[(M_\infty - M_n)^2] &= \mathbb{E}[\liminf_{k\to\infty} (M_{n+k} - M_n)^2] \\ \text{(Fatou's Lemma)} \quad &\leq \liminf_{k\to\infty} \mathbb{E}[(M_{n+k} - M_n)^2] \\ \text{(Using (9.4))} \quad &= \sum_{j=n+1}^{\infty} \mathbb{E}[(M_j - M_{j-1})^2] . \end{aligned}$$

Finally, by Lemma 9.47, since the sum in (9.3) is finite the tail of the sum must be null, so

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

as required. □

Example 9.49 (Sums of zero-mean independent variables in $\mathcal{L}^2$)**.** Let $(Y_k)_{k\geq 1}$ be a sequence of independent random variables with mean zero, $\mathbb{E}[Y_k] = 0$ and variance $\sigma_k^2 = \text{Var}(Y_k) < \infty$ for each k. Let $M_0 = 0$ and $M_n = \sum_{k=1}^n Y_k$, then $(M_n)_{n\geq 0}$ is a martingale since the sum of mean zero independent random variables is always a martingale (check!).

Claim 1: If $\sum_{k\geq 1} \sigma_k^2 < \infty$ then $(M_n)_{n\geq 0}$ converges almost surely. To prove this we observe that

$$\mathbb{E}[(M_k - M_{k-1})^2] = \mathbb{E}[Y_k^2] = \sigma_k^2 ,$$

so applying Equation (9.1),

$$\mathbb{E}[M_n^2] = \sum_{k=1}^n \sigma_k^2 =: A_n$$

If $\sum_k \sigma_k^2 < \infty$ then $(M_n)_{n\geq 0}$ is bounded in $\mathcal{L}^2$ and hence converges almost surely by Theorem 9.48 (compare the definition of the predictable process A_n when defining the angle bracket process).

Claim 2: If $(Y_k)_{k\geq 1}$ are bounded, i.e. there exists a $K > 0$ such that $|Y_k(\omega)| \leq K$ for each $k \geq 1$ and $\omega \in \Omega$, then $(M_n)_{n\geq 0}$ converges almost surely implies $\sum_{k\geq 1} \sigma_k^2 < \infty$. To prove this second claim we follow an argument that should be familiar from Theorem 9.18 and Definition 9.20. Assume $(M_n)_{n\geq 0}$ converges almost surely . First, by independence of the Y_i's, we have, almost surely,

$$\mathbb{E}[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}] = \mathbb{E}[Y_k^2 \mid \mathcal{F}_{k-1}] = \mathbb{E}[Y_k^2] = \sigma_k^2 = A_k - A_{k-1}\,.$$

Then, since M_{k-1} is $\mathcal{F}_{k-1}$ measurable,

$$\begin{aligned}\sigma_k^2 &= \mathbb{E}[M_k^2 \mid \mathcal{F}_{k-1}] - 2M_{k-1}\mathbb{E}[M_k \mid \mathcal{F}_{k-1}] + M_{k-1}^2 \\ &= \mathbb{E}[M_k^2 \mid \mathcal{F}_{k-1}] - M_{k-1}^2\,.\end{aligned}$$

It follows, by rearranging, that $N_n = M_n^2 - A_n$ is a martingale (in fact we have just constructed the angle-bracket process, or square variation, of $(M_n)_{n\geq 0}$ again in this special case.

Now fix $c \in (0, \infty)$ and let $\tau_c = \inf\{r \geq 0 : |M_r| > c\}$. Since (N_n) is a martingale so is the stopped process $N_{n\wedge\tau_c}$, and hence for each $n \geq 0$

$$\mathbb{E}[N_{n\wedge\tau_c}] = \mathbb{E}[M_{n\wedge\tau_c}^2] - \mathbb{E}[A_{n\wedge\tau_c}] = 0 \quad \text{i.e.} \quad \mathbb{E}[A_{n\wedge\tau_c}] = \mathbb{E}[M_{n\wedge\tau_c}^2]\,.$$

However $|M_{\tau_c} - M_{\tau_c - 1}| = |Y_{\tau_c}| \leq K$ if τ_c is finite (otherwise it is zero since $(M_n)_{\geq 0}$ converges almost surely), so

$$|M_{n\wedge\tau_c}| \leq \left.\begin{cases} c & \text{if } n < \tau_c \\ |M_{\tau_c} - M_{\tau_c - 1}| + |M_{\tau_c - 1}| & \text{if } n \geq \tau_c \end{cases}\right\} \leq K + c\,.$$

Hence, by monotonicity of the expectation, we have

$$\mathbb{E}[A_{n\wedge\tau_c}] \leq (K + c)^2 \quad \text{for each } n \geq 1\,. \tag{9.5}$$

Since $(M_n)_{n\geq 0}$ convergences almost surely, the partial sums must be almost surely bounded, and hence there must be some c such that $\mathbb{P}(\tau_c = \infty) > 0$. It now follows from (9.5) and MCT that $A_\infty = \sum_{k=1}^{\infty} \sigma_k^2 < \infty$.

The argument above can be extended to general martingales $(M_n)_{n\geq 0}$ replacing the condition on $\sum_{k\geq 1} \sigma_k^2$ with finiteness of $\langle M \rangle_\infty$.

9.6 Uniformly integrable martingales

Reading: D. Williams, Chapter 14 and R. Durrett, Section 5.5

We now look at what happens when we combine uniform integrability that we saw in Section 6.5 with the martingale property. This can give rise to new proofs of the Kolmogorov 0-1 Law and the Strong Law of Large Numbers. The latter of these two we will definitely look at.

Let $(M_n)_{n\geq 0}$ be a martingale on a filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$, the $(M_n)_{n\geq 0}$ is called a uniformly integrable martingale if the collection $\{M_n\}_{n\geq 0}$ is also unifrmly integrable.

Our first result is to show that one of the basic construction of a martingale (unfolding information) gives something that is automatically uniformly integrable.

Theorem 9.50. *Let X be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and $(\mathcal{F}_n)_{n\geq 0}$ a filtration. Then the process $(X_n)_{n\geq 0}$ with*

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

is a uniformly integrable martingale.

Proof. One strategy would be to use a 'standard machine' type argument. Instead we will use the conditional form of Jensen's inequality (proposition 8.16). Since $f(x) = |x|$ is convex, then by Jensen's inequality,

$$|X_n| = |\mathbb{E}[X \mid \mathcal{F}_n]| \leq \mathbb{E}[|X| \mid \mathcal{F}_n] \quad \text{a.s.}$$

taking average's we have $\mathbb{E}[|X_n|] \leq \mathbb{E}[|X|]$. Also,

$$\begin{aligned} \mathbb{E}[|X_n|\mathbb{1}_{\{X_n>K\}}] &\leq \mathbb{E}[\mathbb{E}[|X| \mid \mathcal{F}_n]\mathbb{1}_{\{X_n>K\}}] \quad (\text{since } |X_n| \leq \mathbb{E}[|X| \mid \mathcal{F}_n]) \\ (\text{taking out what's known}) &\leq \mathbb{E}[\mathbb{E}[|X|\mathbb{1}_{\{X_n>K\}} \mid \mathcal{F}_n]] \\ (\text{tower property}) &\leq \mathbb{E}[|X|\mathbb{1}_{\{X_n>K\}}]\,. \end{aligned}$$

Now, fix $\varepsilon > 0$, since $\{X\}$ on its own forms a uniformly integral family, by Proposition 6.43 (c), we know that there exists a $\delta > 0$ such that $\mathbb{P}[A] < \delta$ implies $\mathbb{E}[|X|\mathbb{1}_A] < \varepsilon$. By Markov's inequality

$$\mathbb{P}[|X_n| \geq K] \leq \frac{\mathbb{E}[|X_n|]}{K} \leq \frac{\mathbb{E}[|X|]}{K}\,,$$

so letting $K = 2\mathbb{E}[|X|]/\delta < \infty$, it follows that $\mathbb{E}[|X_n|\mathbb{1}_{\{X_n>K\}]} < \varepsilon$ uniformly in n as required. □

Since a uniformly bounded family is also bounded in $\mathcal{L}^1$ it follows from Doob's forward Convergence Theorem 9.42 that the limit of a uniformly integrable martingale exists almost surely. Then, since almost sure convergence implies convergence in probability, convergence must also hold in $\mathcal{L}^1$ by Corollary 6.45. It turns out that the converse also holds, and that the limit has the form you might guess from the previous theorem. We collect these results precisely in the next theorem. We already have almost all the ingredients of the proof.

Theorem 9.51. *Let $(M_n)_{n\geq 0}$ be a martingale on filtered space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathbb{P})$, the following are equivalent*

1. *$(M_n)_{n\geq 0}$ is uniformly integrable,*
2. *there is some M_∞ such that $M_n \to M_\infty$ almost surely and in $\mathcal{L}^1$,*
3. *there is an integrable M_∞ such that $M_n = \mathbb{E}[M_\infty \mid \mathcal{F}_n]$ a.s. for all n.*

Remark 9.52. Notice that the first two in the theorem above are also equivalent for sub and supermartingales following the same proof as below.

Proof. 1. $\implies$ 2.: If $(M_n)_{n\geq 0}$ is uniformly integrable then it is certainly bounded in $\mathcal{L}^1$ so it converges in $\mathcal{L}^1$ to some $M_\infty \in \mathcal{L}^1$ by Doob's Forward Convergence Theorem (Theorem 9.42). Since almost sure convergence implies convergence in probability, $M_n \to M_\infty$ in $\mathcal{L}^1$ by Corollary 6.45.

2. $\implies$ 3.: Fix $N \in \mathbb{N}$, since $(M_n)_{n\geq 0}$ is a martingale, for $r \geq N$ we have

$$\mathbb{E}[M_r \mid \mathcal{F}_N] = M_N \quad \text{a.s.}$$

so (recall the defining relation in the definition of conditional expectation) for and $A \in \mathcal{F}_N$ we have

$$\mathbb{E}[M_r \mathbb{1}_A] = \mathbb{E}[M_N \mathbb{1}_A]\,. \tag{9.6}$$

By assumption M_∞ exists and is the $\mathcal{L}^1$ limit of (M_n), hence

$$|\mathbb{E}[M_\infty \mathbb{1}_A] - \mathbb{E}[M_r \mathbb{1}_A]| \leq \mathbb{E}[|(M_\infty - M_r)|\mathbb{1}_A] \leq \mathbb{E}[|(M_\infty - M_r)|] \to 0 \quad \text{as } r \to \infty\,.$$

Now taking the limit as $r \to \infty$ in Equation (9.6) above we have

$$\mathbb{E}[M_\infty \mathbb{1}_A] = \mathbb{E}[M_N \mathbb{1}_A] \quad \text{for all } A \in \mathcal{F}_N\,.$$

Since we also know M_N is $\mathcal{F}_N$-measurable it follows that $M_N = \mathbb{E}[M_\infty \mid \mathcal{F}_N]$ a.s. by the definition of conditional expectation.

3. $\implies$ 1. by Theorem 9.50 above. □

Recall $\mathcal{F}_\infty = \sigma(\bigcup_{n\geq 0} \mathcal{F}_n)$, and notice that $\bigcup_{n\geq 0} \mathcal{F}_n$ is an algebra (hence a π-system) but not necessarily a σ-algebra (hence the need to take the σ-algebra it generates).

Theorem 9.53 (Lévy's 'Upward' Theorem)**.** *Let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and $(\mathcal{F}_n)_{n\geq 0}$ be a filtration. Define $M_n = \mathbb{E}[X \mid \mathcal{F}_n]$ a.s. then $(M_n)_{n\geq 0}$ is a uniformly integrable martingale and*

$$M_n \to \mathbb{E}[X \mid \mathcal{F}_\infty]$$

almost surely and in $\mathcal{L}^1$.

Proof. In light of Theorem 9.50 and Theorem 9.51 we know that $(M_n)_{n\geq 0}$ is a UI martingale that converges a.s. and in $\mathcal{L}^1$ to some M_∞ and $M_n = \mathbb{E}[M_\infty \mid \mathcal{F}_n]$. It only remains to prove that $M_\infty = \mathbb{E}[X \mid \mathcal{F}_\infty]$. Without loss of generality we assume the $X \geq 0$.

[We have used a similar proof technique before]. Define two measures on $(\Omega, \mathcal{F})$ by

$$Q_1(F) = \mathbb{E}[X\,;\, F] \quad \text{and} \quad Q_2 = \mathbb{E}[M_\infty\,;\, F] \quad \text{for } F \in \mathcal{F}_\infty\,.$$

Since

$$\mathbb{E}[X \mid \mathcal{F}_n] = M_n = \mathbb{E}[M_\infty \mid \mathcal{F}_n]$$

then by the defining relation of conditional probability we have for each $F \in \mathcal{F}_n$

$$Q_1(F) = \mathbb{E}[X\,;\,F] = \mathbb{E}[M_\infty\,;\,F] = Q_2(F).$$

Thus, Q_1 and Q_2 agree on the π-system $\bigcup \mathcal{F}_n$ and hence they agree on $\mathcal{F}_\infty$. Furthermore, since $M_\infty(\omega) = \limsup M_n(\omega)$ we know M_∞ is $\mathcal{F}_\infty$-measurable, so by the definition of conditional expectation we have $M_\infty = \mathbb{E}[X \mid \mathcal{F}_\infty]$ a.s., as required. □

We get the following 0-1 law now as an immediate consequence.

Theorem 9.54 (Lévy's 0-1 Law)**.** *If $(\mathcal{F}_n)_{n\geq 0}$ is a filtration and $\mathcal{F}_\infty = \sigma(\bigcup_{n\geq 0} \mathcal{F}_n)$ then for $A \in \mathcal{F}_\infty$ we have $\mathbb{E}[\mathbb{1}_A \mid \mathcal{F}_n] \to \mathbb{1}_A$.*

This result looks fairly 'innocent', but notice that it is very useful, in particular the first part of Kolmogorovs 0-1 law (Theorem 5.19) is an immediate consequence: If $(\mathcal{F}_n)_{n\geq 0}$ are independent and $A \in \mathcal{T}$, then for each n we have A is independent of $\mathcal{F}_n$, so $\mathbb{E}[1_A \mid \mathcal{F}_n] = \mathbb{E}[1_A] = \mathbb{P}(A)$ and the right hand side here is independent on n. Taking $n \to \infty$ we have $\mathbb{P}(A) = \mathbb{1}_A$ a.s., so the probability of A must be zero or one.

We finish this brief section on uniformly integrable martingales with a version of hte Optional Stopping Theorem for uniformly integrable martingales, which applies for *any* stopping time. We already stated this result as Corollary 9.32. *The proof of the following theorem is non-examinable.*

Theorem 9.55. *Let (M_n) be a uniformly integrable martingale, and τ a stopping time. Then M_τ is integrable and $\mathbb{E}[M_\tau] = \mathbb{E}[M_0]$.*

Proof. Since $f(x) = |x|$ is convex $(|M_n|)_{n\geq 0}$ is a submartingale by Proposition 9.13. Then, by Lemma 9.28 and Proposition 6.43 (a), for any n

$$\mathbb{E}[|M_{n\wedge\tau}|] \leq \mathbb{E}[|M_n|] \leq \sup_k \mathbb{E}[|M_k|] < \infty$$

(the first inequality above is left as an exercises for the very keen). It follows that $(M_{n\wedge\tau})_{n\geq 0}$ is a martingale bounded in $\mathcal{L}^1$, it therefore converges almost surely to M_τ which must be integrable.

We want to show that $(M_{n\wedge\tau})_{n\geq 0}$ is in fact uniformly integrable. Fix $K \geq 0$,

$$\mathbb{E}[|M_{n\wedge\tau}\mathbb{1}_{|M_{n\wedge\tau}|>K}] \leq \mathbb{E}[|M_\tau|\mathbb{1}_{|M_\tau|>K}] + \mathbb{E}[|M_n|\mathbb{1}_{M_n>K}]\,,$$

since $M_{n\wedge\tau}$ is either M_n or M_τ. Now the right hand side goes to 0 as $K \to \infty$, uniformly in n, since (for the first term) M_τ is integrable and (for the second term) (M_n) is uniformly integrable. So $(M_{n\wedge\tau})_{n\geq 0}$ is uniformly integrable. Now by Theorem 9.51 $M_{n\wedge\tau} \to M_\tau$ in $\mathcal{L}^1$, so $\mathbb{E}[M_\tau] = \lim_{n\to\infty} \mathbb{E}[M_{n\wedge\tau}] = \lim_{n\to\infty} \mathbb{E}[M_0] = \mathbb{E}[M_0]$ as required.

□

9.7 Backwards martingales and the strong law of large numbers

So far we have defined martingales indexed by non-negative numbers, $(M_n)_{n\geq 0}$, and we thought about the index as 'discrete time'. However, the definition makes just as much sense for indices in any discrete 'interval'. The conditions we required were that the martingale is adapted to some filtration - which is an increasing sequence of σ-algebras, and for each t we have $\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] = M_t$ a.s.

A *backward martingale* (sometimes called reversed) is a martingale which is indexed by non-positive integers, i.e. $I = \{n \in \mathbb{Z} : n \leq 0\}$. We often choose to write $(M_{-n})_{n\geq 0}$ instead of $(M_n)_{n\leq 0}$. Note that a backward martingales *ends* at time 0. Roughly speaking, information is collected from some prehistoric past, until the present day at time zero.

Definition 9.56 (Backward martingale)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Given a sequence of σ-algebras $(\mathcal{F}_{-n})_{n\geq 0}$ with $\mathcal{F}_{-n} \in \mathcal{F}$ and increasing, i.e.

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

a *backward martingale* w.r.t. $(\mathcal{F}_{-n})_{n\geq 0}$ is a sequence $(M_{-n})_{n\geq 0}$ of integrable random variables such that M_{-n} is $\mathcal{F}_{-n}$ measurable and

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

The martingale property implies that the expected value at time zero given the information up to time $-n$ is M_{-n};

$$\mathbb{E}[M_0 \mid \mathcal{F}_{-n}] = M_{-n} \quad \text{a.s. for } n \geq 1.$$

Then by the same argument as in the proof of Theorem 9.50 (in fact the theorem holds for any sequence of increasing σ-algebras), we since M_0 is integrable we know that $(M_{-n})_{n\geq 0}$ is uniformly integrable. So any backward martingale is automatically uniformly integrable.

If we consider the sequence $(M_{-m}, M_{-m+1}, \ldots, M_{-1}, M_0)$, this is just $m+1$ steps of a 'normal' martingale started from initial random starting point M_{-m}. Hence, we may apply Doob's Upcrossing Lemma (Lemma 9.41). If $U_m[a,b]$ is the number of upcrossings the backward martingale between $-m$ and 0, then

$$\mathbb{E}[U_m[a,b]] \leq \frac{\mathbb{E}[(a - M_0)^+]}{b-a}\,.$$

Letting $n \to \infty$ and using monotone convergence theorem, we have $\mathbb{E}[U_\infty] < \infty$. Following the same argument then as in the proof of Doob's Forward Convergence Theorem (Theorem 9.42) show that M_{-n} converges almost surely as $n \to \infty$ to some random variable $M_{-\infty}$ (for definiteness, like in the proof of Theorem 9.53, let $M_{-\infty} = \liminf_{n\to\infty} M_{-n}$ since we know the limit exists almost everywhere). Since we have shown that a backward martingale is automatically uniformly integrable, we also know by Theorem 9.51 that convergence also holds in $\mathcal{L}^1$. We have proved the following theorem.

Theorem 9.57. *Let $(M_{-n})_{n\geq 0}$ be a backward martingale, then $M_{-\infty} = \lim_{n\to\infty} M_{-n}$ exists almost surely and convergence holds also in $\mathcal{L}^1$.*

The next result identifies the limit above. Notice that as n increases the sequence $\mathcal{F}_{-n}$ decreases, so in order to define the 'limiting' family $\mathcal{F}_{-\infty}$ we need to take the intersection. Recall that the intersection of σ-algebras is automatically a σ-algebra.

Theorem 9.58. *In the setting of the previous theorem, if* $M_{-\infty} = \lim_{n\to\infty} M_{-n}$ *and* $\mathcal{F}_{-\infty} = \cap_{n\geq 0}\mathcal{F}_{-n}$*, then* $M_{-\infty} = \mathbb{E}[M_0 \mid \mathcal{F}_{-\infty}]$ *a.s.* .

Proof. The random variable $M_{-\infty}$ is $\mathcal{F}_{-k}$ measurable for each $k \geq 0$ since M_{-n} is $\mathcal{F}_{-k}$ measurable for each $n \geq k$, so $M_{-\infty}$ must be $\mathcal{F}_{-\infty}$-measurable. Then, following the same strategy as in the proof of Theorem 9.53, for $A \in \mathcal{F}_{-\infty} \subseteq \mathcal{F}_{-n}$

$$\mathbb{E}[M_{-n}\,;\, A] = \mathbb{E}[M_0\,;\, A]\,,$$

and, taking the limit as $n \to \infty$, we get

$$\mathbb{E}[M_{-\infty}\,;\, A] = \mathbb{E}[M_0\,;\, A] \quad \text{for all } A \in \mathcal{F}_{-\infty}\,,$$

which completes the proof. □

Even though convergence theory for backwards martingales looks relatively straightforward, it turns out there are some very nice applications. We now use the results of this subsection to give a proof of the celebrated Kolmogorov Strong Law.

Theorem 9.59 (Kolmogorov's Strong Law of Large Numbers)**.** *Let* $(Y_n)_{n\geq 1}$ *be a sequence of i.i.d. random variables with* $\mathbb{E}[Y_i] = \mu < \infty$*, and let*

$$S_n = \sum_{i=1}^{n} Y_i\,.$$

Then

$$\frac{1}{n}S_n \to \mu \quad \textit{a.s. and in } \mathcal{L}^1 \textit{ as } n \to \infty.$$

Proof. For $n \geq 1$ let

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

and observe that $\mathcal{F}_{-n-1} \subseteq \mathcal{F}_{-n}$. By time $-n$ we know the value of S_n and all the subsequent partial sums, but not that value of the random variables $Y_1, Y_2, \ldots, Y_{n-1}$. We now use the symmetry between $Y_1, Y_2, \ldots, Y_n$ which is not affected by conditioning on $\mathcal{F}_{-n}$, since none of the $S_n, S_{n+1}, \ldots$ change if we permute $Y_1, \ldots, Y_n$. So,

$$\mathbb{E}[Y_1 \mid \mathcal{F}_{-n}] = \mathbb{E}[Y_2 \mid \mathcal{F}_{-n}] = \ldots = \mathbb{E}[Y_n \mid \mathcal{F}_{-n}]\,,$$

then by linearity of the conditional expectation, for $i = 1, 2, \ldots, 2$,

$$\mathbb{E}[Y_i \mid \mathcal{F}_{-n}] = \frac{1}{n}\mathbb{E}[Y_1 + \ldots + Y_n \mid \mathcal{F}_{-n}] = \frac{1}{n}\mathbb{E}[S_n \mid \mathcal{F}_{-n}] = \frac{1}{n}S_n\,,$$

where the final inequality follows from the fact that S_n is $\mathcal{F}_{-n}$-measurable. Now, for $n \geq 1$, let $M_{-n} = S_n/n$, then for $n \geq 2$

$$\mathbb{E}[M_{-n+1} \mid \mathcal{F}_{-n}] = \mathbb{E}[S_{n-1}/(n-1) \mid \mathcal{F}_{-n}] = \frac{1}{n-1}\sum_{i=1}^{n-1}\mathbb{E}[Y_i \mid \mathcal{F}_{-n}] = \frac{1}{n}S_n = M_{-n}\,.$$

This shows that $(M_{-n})_{n\geq 1}$ is a backwards martingale with respect to $(\mathcal{F}_{-n})_{n\geq 1}$. Thus, by Theorem 9.57, we know $M_{-n} = S_n/n$ converges almost surely and in $\mathcal{L}^1$, and by Theorem 9.58 the limit is given by $M_{-\infty} = \mathbb{E}[M_{-1} \mid \mathcal{F}_{-\infty}]$, where $\mathcal{F}_{-\infty} = \bigcap_{k\geq 1}\mathcal{F}_{-k}$.

Finally, by $\mathcal{L}^1$ convergence $\mathbb{E}[M_{-\infty}] = \lim_{n\to\infty}\mathbb{E}[M_{-n}] = \mathbb{E}[M_{-1}] = \mathbb{E}[S_1] = \mathbb{E}[Y_1] = \mu$. Also, $M_\infty = \liminf S_n/n$ is a tail random variable with respect to the natural filtration of $(Y_n)_{n\geq 1}$, which is an i.i.d. sequence, so by Kolmogorov's 0-1 law the random variable must be almost surely constant. Since it's mean is μ it must be almost surely equal to μ. □

9.8 Exchangeability [Non examinable 2021]

Reading: R. Durrett, Section 5.6, A. Klenke Chapter 12

This is a very quick introduction and taster of exchangeability, please see the literature above for further details and more references.

The use of symmetry was key to the proof of the strong law of large numbers that we gave at the end of the previous section. In that case, it followed from independence of the Y_i's, but actually such symmetry follows from a weaker condition - namely *exchangeability*. In many situations the order in which data arrives does not matter. Mathematically, we say that a countable (or finite) sequence of random variables is exchangeable if the joint distribution does not change under permuting finitely many of the random variables. It turns out, a fundamental structural theorem about exchangeable sequences, called De Finetti's Theorem, sates that it is possible to sample exchangeable sequences in nice two-step way. If $(X_1, X_2, \ldots)$ is an exchangeable sequence of random variables $X_i \in E$, then to sample we may first pick a random distribution μ on E (according to some probability measure on distributions of E), and then pick a sequence of i.i.d. random variables each with distribution μ.

Definition 9.60. A countable sequence of random variables $(X_1, X_2, \ldots)$ is called *exchangeable* if for any $n \in \mathbb{N}$ and any permutation $\pi \in S_n$, the laws of $(X_1, X_2, \ldots, X_n)$ and $(X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)})$ are the same.

Example 9.61. If $(Y_1, Y_2, \ldots)$ are i.i.d. then they are exchangeable.

Example 9.62. Let $(X_1, \ldots, X_n)$ be the results of n successive samples without replacement from a pool of at least n values (some of which may be the same). The the random variables $X_1, \ldots, X_n$ and exchangeable, but *not* independent.

It turns out that we can use the construction in the previous proof of the Strong Law of Large Numbers to construct a finite martingale from a collection of exchangeable random variables. Suppose $X_1, X_2, \ldots, X_N$ are exchangeable and integrable, then let $S_n = \sum_{i=1}^n X_i$, and let

$$Z_j = \mathbb{E}[X_1 \mid \sigma(S_{N-j+1}, S_{N-j+2} \ldots, S_{N-1}, S_N)] \quad \text{for } j = 1, 2, \ldots, N\,.$$

So Z_j is X_1 conditioned on the last j terms in the sum. Since Z_j is an integrable random variable, namely X_1, conditioned on information which is increasing in j, we have already seen that this must define a (finite) martingale. Now we show that this construction corresponds to what we did in the proof of the SLLN,

$$\begin{aligned} S_{N+1-j} = \mathbb{E}[S_{N+1-j} \mid \sigma(S_{N+1-j}, \ldots, S_N)] &= \sum_{i=1}^{N+j-1} \mathbb{E}[X_i \mid \sigma(S_{N+1-j}, \ldots, S_N)] \\ \text{(by exchangeability)} \quad &= (N+1-j)\mathbb{E}[X_1 \mid \sigma(S_{N+1-j}, \ldots, S_N)] \\ &= (N+1-j)Z_j\,, \end{aligned}$$

so $Z_j = S_{N+1-j}/(N+1-j)$. The martingale $(Z_j)_{1 \le j \le N}$ is sometimes called the Doob backward martingale.

Example 9.63 (The ballot problem)**.** In an election between two candidates, called A and B, candidate A receives n gets and candidate B receives m, where $n > m$. Assume the every order of votes is equally likely, what is the probability that A is always ahead of B throughout the counting of the votes?

Let $X_i = 1$ if the ith vote counted is for candidate A and -1 if it is for B, and let $S_k = \sum_{i=1}^{k} X_i$. For convenience let $N = n + m$ be the total number of votes. Since all the orderings of the N vote are assumed to be equally likely, the sequence of random variables $X_1, \ldots, X_N$ is exchangeable, so

$$Z_j = \frac{S_{N+1-j}}{N+1-j}\,, \quad \text{for } j = 1, 2, \ldots, N = n + m\,,$$

is a Doob backward martingale with respect to $\mathcal{F}_j = \sigma(S_{N+1-j}, \ldots, S_N)$ [Equivalently $M_{-j} = S_j/j$ is a backward martingale with respect to $\mathcal{F}'_{-j} = \sigma(S_j, \ldots, S_N)$, this explains the name].

Since we assume that $n > m$, either candidate A is always ahead of B, or there is a tie (equal votes) at some point during the counting. We can define the bounded stopping time

$$\tau = \inf\{j \geq 1 \,:\, Z_j = 0 \text{ or } j = N\}\,.$$

On the event $\{A \text{ is always ahead}\}$ we have $Z_\tau = Z_N = X_1 = 1$, since if A is always ahead then they must have won the first vote. On the event $\{A \text{ is always ahead}\}^c$ we have $Z_\tau = 0$ from the definition of the hitting time, so

$$\mathbb{E}[Z_\tau] = \mathbb{P}(\{A \text{ is always ahead}\})\,.$$

But we also know, either by the Optional Stopping Theorem (Theorem 9.30) or by Proposition 9.28, that

$$\mathbb{E}[Z_\tau] = \mathbb{E}[Z_1] = \frac{S_N}{N} = \frac{n-m}{n+m}\,.$$

Hence $\mathbb{P}(\{A \text{ is always ahead}\}) = (n-m)/(n+m)$.

We close this short section with a very brief discussion of de Finetti's Theorem. It turns out that we can prove de Finetti's Theorem by following a similar argument to the one used above and in the proof of the Strong Law of Large Numbers.

Suppose that $(X_n)_{n\geq 0}$ is a stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$ where each X_i takes values in a Polish space E. Let $\mathcal{E}_n$ be the σ-algebra generated by the events that are invariant under permutations of the first n random variables $X_1, \ldots, X_n$. We define the exchangeable σ-algebra, $\mathcal{E}$, by $\mathcal{E} = \bigcap_n \mathcal{E}_n$.

Theorem 9.64 (de Finetti's Theorem)**.** *If $(X_n)_{n\geq 0}$ is exchangeable then $(X_n)_{n\geq 0}$ are i.i.d. given (conditioned on) $\mathcal{E}$.*

idea of the proof. For a proof using the symmatrised average of bounded measurable real functions, and applying the convergence theorem for backward martingales, see R. Durrett, Section 5.6, or A. Klenke Chapter 12. □

When the X_i take values in 'nice' spaces then there is a regular conditional distirbution for $(X_1, X_2, \ldots)$ given $\mathcal{E}$, and then (as advertised at the begning of this section) the sequence can be represented by a mixture of i.i.d. sequence (i.e. first pick a random measure for each marginal and then pick an i.i.d sequence according to this measure). The siplest case of this and easiest to express is when $X_i \in \{0, 1\}$. In this case:

Theorem 9.65 (de Finetti's for $\{0,1\}$ r.v.s)**.** *Suppose $(X_n)_{n\geq 0}$ is an exchangeable sequence of random variables that each take values in $\{0,1\}$, then there exists a probability distribution μ on $[0,1]$ (parameter space of Bernoulli measures) such that*

$$\mathbb{P}(X_1 = 1, \ldots, X_k = 1, X_{k+1} = 0, \ldots, X_n = 0) = \int_{[0,1]} \theta^k (1-\theta)^{n-k} \, \mathrm{d}\mu(\theta) \, .$$

The right hand side above expresses the fact that the law of the X_i's is a convex combination (mixture) of identical Bernoulli measures.