Probability Theory Notes

Francesco N. Chotuck

Abstract

This is the Imperial College London postgraduate module MATH70058 Probability Theory, instructed by Dr. Ajay Chandra. The formal name for this class is "Probability Theory".

Contents

Note 0.1. I would like to preface these notes by mentioning that I have not formally studied Measure Theory in a dedicated course. As a result, I will be recapping several concepts from measure theory without providing formal proofs.

1 Measure Theory

Note 1.1. 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.

These notes are based from the Warwick notes, compare with the ICL notes

1.1 Definitions

Remark 1.2. We now fix some notation. Throughout these notes we will use Ω to denote a sample space, and for some set 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. $\{\omega\} \subset \Omega$.

Definition 1.3. Let Ω be a set and $\varnothing \neq \mathcal{A} \subset 2^\Omega$ be a collection of subsets of Ω. We call $\mathcal{A}$ an **algebra** for Ω if it is closed under unions and complements i.e.

- $A \in \mathcal{A} \Rightarrow A^c = \Omega \setminus A \in \mathcal{A}$,
- $A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$.

Corollary 1.4. For an algebra $\mathcal{A}$, we have that $\varnothing, \Omega \in \mathcal{A}$ and for every $A, B \in \mathcal{A}$ then $A \cap B \in \mathcal{A}$.

Definition 1.5. An algebra $\mathcal{A}$ on Ω is called a **σ-algebra** if is closed under countable unions i.e. $A_n \in \mathcal{A}$ for each $n \in \mathbb{N}$ then $\bigcup_{i=1}^\infty A_i \in \mathcal{A}$.

Remark 1.6. σ-algebras are also closed under countable intersections.

Example 1.7. Given Ω, two examples of σ-algebras on Ω are

$$\mathcal{F}_\star = \{\Omega, \emptyset\} \quad \text{and} \quad \mathcal{F}^\star = 2^\Omega.$$

Every other σ-algebra $\mathcal{F}$ on Ω satisfies $\mathcal{F}_\star \subset \mathcal{F} \subset \mathcal{F}^\star$.

Example 1.8

Another example of a σ-algebra: given any partition $\{D_i\}_{i\in I}$ of Ω with I countable, we can define a σ-algebra

$$\sigma(\{D_i\}_{i\in I}) = \left\{ \bigcup_{j\in J} D_j : J \subset I \right\}.$$

Definition 1.9. Let $\mathcal{A}$ be an algebra on Ω. A map $\mu : \mathcal{A} \to [0,\infty] = [0,\infty) \cup \{\infty\}$ is called a **set function**. We say that μ is

1. **increasing** if $\mu(A) \le \mu(B)$ for any $A, B \in \mathcal{A}$ with $A \subset B$.
2. **finitely additive** for any disjoint $A, B \in \mathcal{A}$ we have
$$\mu(A \cup B) = \mu(A) + \mu(B).$$
3. **σ-additive** for any disjoint sets $A_n \in \mathcal{A}$ for $n \in \mathbb{N}$ we have
$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

Remark 1.10. There is no issue with the series appearing in the above definition, which is over non-negative real numbers. Its value is an element of the extended real line $[0,\infty] = [0,\infty) \cup \{\infty\}$ (i.e. the series may well diverge). By convention, we set

- $x + \infty = \infty$,
- $x \times \infty = \infty$, and
- $\infty + \infty = \infty$

for all $x \in [0,\infty]$.

Corollary 1.11. A set function μ being finitely additive implies

- $\mu(\varnothing) = 0$, and
- for any $A, B \in \mathcal{A}$ (i.e. they may not be disjoint) $\mu(A\cup B) = \mu(A)+\mu(B)-\mu(A\cap B)$.

Proof. We prove the statements.

- Consider $\mu(\varnothing \cup \varnothing)$.
- The trick is to write $A = (A \cap B) \cup (A \setminus B), B = (A \cap B) \cup (B \setminus A)$ and $A \cup B = (A \cap B) \cup (A \setminus B) \cup (B \setminus A)$ where all sets in the union are disjoint.

$\square$

Definition 1.12 (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 $\mu : \mathcal{F} \to [0, \infty]$ is a σ-additive function. In this case μ is called a **measure** on $(\Omega, \mathcal{F})$.

Definition 1.13 (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 sets $E_i \in \mathcal{F}$ such that $\mu(E_i) < \infty$ and $\Omega = \bigcup_{j=1}^{\infty} E_j$.

Definition 1.14. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a measure space. A measure $\mathbb{P}$ is a **probability measure** if it is σ-additive and $\mathbb{P}(\Omega) = 1$.

Remark 1.15. We often use the notation $(\Omega, \mathcal{F}, \mathbb{P})$ to emphasise that we have a probability measure.

Definition 1.16. If μ is a σ-additive on an algebra $\mathcal{A} \subset 2^{\Omega}$, we say μ is a **pre-measure** on $\mathcal{A}$.

Definition 1.17. Given a measurable space $(\Omega, \mathcal{F})$, a **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where $\mathbb{P}$ is a probability measure on $\mathcal{F}$. We often call Ω the **sample space** and $\mathcal{F}$ the **σ-algebra of events**. We often say, for $A, B \in \mathcal{F}$,

- $\mathbb{P}(A)$ is the probability of event A.
- $\mathbb{P}(A \cup B)$ is the probability of the event A **or** B occurring.
- $\mathbb{P}(A \cap B)$ is the probability of the event A **and** B occurring.
- $\mathbb{P}(A^c) = \mathbb{P}(\Omega \setminus A)$ is the probability of the event A **not occurring.**

1.2 Basic results

Definition 1.18 (Notation)**.** For a sequence of sets $(A_n)_{n\geq 1}$ we write $A_n \uparrow A$ if $A_n \subseteq A_{n+1}$ for each n and $\bigcup_{n=1}^{\infty} A_n = A$. Similarly, $B_n \downarrow B$ if $B_n \supseteq B_{n+1}$ for all n and $\bigcap_{n=1}^{\infty} B_n = B$.

Lemma 1.19 (Monotone convergence for probability spaces)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

1. If $(A_n)_{n\geq 1}$ is a sequence in $\mathcal{F}$ with $A_n \uparrow A$, then $\mathbb{P}(A_n) \uparrow \mathbb{P}(A)$ as $n \to \infty$.

   Note 1.20. Continuity from below.

2. If $(B_n)_{n\geq 1}$ is a sequence in $\mathcal{F}$ with $B_n \downarrow B$, then $\mathbb{P}(B_n) \downarrow \mathbb{P}(B)$ as $n \to \infty$.

   Note 1.21. Continuity from above.

Remark 1.22. This lemma also holds for measure spaces with a small tweak in the second statement where we require the measure μ to be such that $\mu(B_k) < \infty$ for some $k \in \mathbb{N}$. We illustrate this version of the lemma as the proof is simpler.

3. *Proof.* Note that (1) and (2) are equivalent by using De Morgan's laws to interchange union and intersections with complements. **From [DD19, Theorem 1.1.1 (iii) & (iv)]**. We prove each statement in turn.

   1. We write $D - E$ for $D \cap E^c$, with this notation in mind let $D_n = A_n - A_{n-1}$. Then the D_n are disjoint and have $\bigcup_{m=1}^{\infty} D_m = A, \bigcup_{m=1}^{n} D_m = A_n$ so
   $$\mathbb{P}(A) = \sum_{m=1}^{\infty} \mathbb{P}(D_m) = \lim_{n\to\infty} \sum_{m=1}^{n} \mathbb{P}(D_m) = \lim_{n\to\infty} \mathbb{P}(A_n).$$
   2. $B_k - B_n \uparrow B_k - B$ for some k so the above implies $\mathbb{P}(B_k - B_n) \uparrow \mathbb{P}(B_k - B)$. Since $B_k \supset B$ we have $\mathbb{P}(B_k - B) = \mathbb{P}(B_k) - \mathbb{P}(B)$, and it follows that $\mathbb{P}(B_n) \downarrow \mathbb{P}(B)$.

□

Definition 1.23. Continuity from above with $A = \varnothing$ is called **continuity at** 0.

Lemma 1.24

A finitely additive μ on $(\Omega, \mathcal{F})$ with $\mu(\Omega) = 1$ is a measure if and only if it is continuous at 0.

Proof. We prove each direction.

- Proof of ($\Rightarrow$). We first assume that μ is a measure, then it is σ-additive, and it follows from the above lemma that it is continuous at 0.
- Proof of ($\Leftarrow$). Conversely, we assume that μ is continuous at 0. By the proof of the lemma above, this implies μ satisfies item (1) of the lemma above, that is, it is continuous from below. We now show that μ is σ-additive, so that it is indeed a measure. Let $A_1, A_2, \ldots \in \mathcal{F}$ be disjoint, and set $A = \bigcup_{n=1}^{\infty} A_n$. Let $B_n = \bigcup_{j=1}^{n} A_j$, and note that $B_n \uparrow A$. Then, by continuity from below, we have $\mu(B_n) \uparrow \mu(A)$. However, we also have $\mu(B_n) = \sum_{j=1}^{n} \mu(A_j)$, and so we have $\sum_{j=1}^{\infty} \mu(A_j) = \mu(A)$.

□

Lemma 1.25. 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**) $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu(A_n)$ for $A_1, A_2, \ldots \in \mathcal{F}$.

If μ is also a finite measure ($\mu(\Omega) < \infty$), then (**inclusion-exclusion formula**) for all $A_1, A_2, \ldots, A_n \in \mathcal{F}$,

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

1.3 Borel σ-algebra

Lemma 1.26 (Generating σ-algebras)
Consider any set Ω, and $\mathcal{C}$ collection of subsets of Ω. Then the σ-algebra **generated** by $\mathcal{C}$ is the intersection of all σ-algebras on Ω which contain $\mathcal{C}$, that is

$$\sigma(\mathcal{C}) = \bigcap_{\mathcal{F} \subset 2^\Omega \text{ a } \sigma\text{-algebra with } \mathcal{C} \subset \mathcal{F}} \mathcal{F}$$

Note 1.27. Given a collection $\mathcal{C}$ of subsets of Ω, there exists a unique "smallest" σ-algebra, $\sigma(\mathcal{C})$, containing $\mathcal{C}$. We read $\mathcal{F} \supset \mathcal{C}$, as σ-algebras $\mathcal{F}$ containing $\mathcal{C}$.

Remark 1.28. From this statement we can also deduce that the intersection of an arbitrary collection of σ-algebras is a σ-algebra.

Proof. Since 2^Ω is a σ-algebra containing $\mathcal{C}$, the intersection is non-empty. To prove it is a σ-algebra it suffices to check the axioms for a σ-algebra are satisfied.

- We have $\varnothing \in \mathcal{F}$ for all $\mathcal{F} \supset \mathcal{C}$.
- Suppose $A \in \sigma(\mathcal{C})$ then A is in each $\mathcal{F}$ for all $\mathcal{F} \supset \mathcal{C}$. Since $\mathcal{F}$ is a σ-algebra it follows that $A^c \in \mathcal{F}$ for all $\mathcal{F} \supset \mathcal{C}$ hence, $A^c \in \sigma(\mathcal{C})$.
- Similarly, given a $(A_n)_{n\geq 1}$ all in $\sigma(\mathcal{C})$, it follows that these sets are all in $\mathcal{F}$. Since each $\mathcal{F}$ is a σ-algebra we have that $\bigcup A_n \in \mathcal{F}$ for all $\mathcal{F} \supset \mathcal{C}$, and so $\bigcup A_n \in \sigma(\mathcal{C})$.

$\square$

Definition 1.29. If (X, τ) is a topological space (with $\tau \subset 2^X$ the collection of open sets), then

$$\mathcal{B}(X) = \sigma(\tau)$$

is called the **Borel σ-algebra** of (X, τ).

Note 1.30. Let X be a topological space, set

$$\mathcal{B}(X) = \sigma(\{A \subset X : A \text{ is open}\}).$$

Remark 1.31. The σ-algebra $B(X)$ depends on the topology, not just X, although it does not appear in the notation.

Example 1.32 ($\mathcal{B}(\mathbb{R})$ σ-algebra)
For $X = \mathbb{R}$ we have

$$\begin{aligned}\mathcal{B}(\mathbb{R}) &= \sigma(\{\text{open intervals}\}) = \sigma(\text{closed intervals}) \\ &= \sigma(\{(-\infty, x] : x \in \mathbb{R}\}).\end{aligned}$$

We observe that we can write open interval

$$(a, b) = \bigcup_{n=1}^{\infty} \left(a, b - \frac{1}{n}\right], \quad a < b.$$

Similarly,

$$[a, b] = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, b\right], \quad a < b$$

and

$$\{a\} = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, a\right].$$

Thus, the Borel σ-algebra also contains all sets of the form

$$\{a\}, \quad (a, b), \quad [a, b], \quad (-\infty, b), \quad (-\infty, b], \quad (a, \infty).$$

1.3.1 Product σ-algebras

Definition 1.33 (Product σ-algebra)**.** Given two measurable spaces $(\Omega_1, \mathcal{F}_1)$, $(\Omega_2, \mathcal{F}_2)$, we can obtain the measurable space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ by setting

$$\mathcal{F}_1 \otimes \mathcal{F}_2 = \sigma(\{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}).$$

Lemma 1.34
We have that $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^2)$.

Proof. We prove the double inclusion.

- Proof of $\subseteq$. For this inclusion, it suffices to show that for any $A \in \mathcal{B}(\mathbb{R})$, $B \in \mathcal{B}(\mathbb{R})$, we have $A \times B \in \mathcal{B}(\mathbb{R}^2)$.

  Note that $A \times B = (A \times \mathbb{R}) \cap (\mathbb{R} \times B)$, so it suffices to show that $A \times \mathbb{R}, \mathbb{R} \times B \in \mathcal{B}(\mathbb{R}^2)$. Now we note that

  $$A \times \mathbb{R} \in \sigma(\{C \times \mathbb{R} : C \subseteq \mathbb{R}, C \text{ open}\}) \subseteq \mathcal{B}(\mathbb{R}^2),$$

  where the second inclusion follows from the fact that such $C \subseteq \mathbb{R}$ are open in $\mathbb{R}$. The proof that $\mathbb{R} \times B \in \mathcal{B}(\mathbb{R}^2)$ is similar.

- Proof of $\supseteq$. For this inclusion, it suffices to show that for any $A \subseteq \mathbb{R}^2$ open, we have $A \in \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$. Let $\mathcal{X} \subseteq \mathbb{R}^2$ be the collection of open rectangles $(a, b) \times (c, d)$

with $a, b, c, d \in \mathbb{Q}$. Then $\mathcal{X}$ is countable. We then remark that, by density $\mathbb{Q}^2 \subseteq \mathbb{R}$ and the fact that A is open,

$$A = \bigcup_{R \in \mathcal{X}, R \subseteq A} R.$$

Since each $R \in \mathcal{X}$ belongs to $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$

□

1.3.2 Cylindrical σ-algebra

Note 1.35. This space is used as the basis for constructing probabilistic models of experiments with infinitely many steps.

Definition 1.36. Let $\mathbb{R}^\infty = \{(x_1, x_2, \dots) : x_j \in \mathbb{R}\}$. For each $n \geq 1$, we define $\pi_n : \mathbb{R}^\infty \to \mathbb{R}^n$ by mapping

$$x = (x_n)_{n=1}^\infty \mapsto (x_1, \dots, x_n).$$

We define the collection of **cylinder sets** $\mathcal{C} \subseteq 2^{\mathbb{R}^\infty}$:

$$\mathcal{C} = \{\pi_n^{-1}(A) \subseteq \mathbb{R}^\infty : n \geq 1,\ A \in \mathcal{B}(\mathbb{R}^n)\}.$$

Proposition 1.37

The cylindrical σ-algebra is then given by $\sigma(\mathcal{C})$.

Remark 1.38. $\mathcal{C}$ is not a σ-algebra, but it is an algebra (of sets).

Corollary 1.39

Since $\mathbb{R}$ is a topological space, $\mathbb{R}^\infty$ is a topological space with the product topology. Furthermore,

$$\mathcal{B}(\mathbb{R}^\infty) = \sigma(\mathcal{C}).$$

Example 1.40

Consider the measurable space $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$. Let $r \in \mathbb{R}$, then define the following subsets of $\mathbb{R}^\infty$:

$$A_r = \left\{ x = (x_n)_{n=1}^\infty \in \mathbb{R}^\infty : \limsup_{n\to\infty} x_n = \inf_n \left(\sup_{k\geq n} x_k \right) > r \right\},$$

$$B_r = \left\{ x = (x_n)_{n=1}^\infty \in \mathbb{R}^\infty : \liminf_{n\to\infty} x_n = \sup_n \left(\inf_{k\geq n} x_k \right) > r \right\}.$$

We have that $A_r, B_r \in \mathcal{B}(\mathbb{R}^\infty)$. Since one can take

$$A_r = \bigcap_{n=1}^\infty \bigcup_{k=n}^\infty \{x \in \mathbb{R}^\infty : x_k > r\},$$

and

$$B_r = \bigcup_{n=1}^\infty \bigcap_{k=n}^\infty \{x \in \mathbb{R}^\infty : x_k > r\}.$$

1.4 Lebesgue measure

Theorem 1.41 (Carathéodory's Extension Theorem)

Let $\mathcal{A}$ be an algebra of sets, and suppose $\widetilde{\mu} : \mathcal{A} \to [0, \infty)$ is σ-additive and σ-finite on $\mathcal{A}$. Then there exists a unique extension of $\widetilde{\mu}$ to a σ-finite measure σ-finite measure $\mu : \sigma(\mathcal{A}) \to [0, \infty]$.

Example 1.42 (Lebesgue measure on $(0, 1]$)

Let $\Omega = (0, 1]$ and $\mathcal{A} = \{(a_1, b_1] \cup (a_2, b_2] \cup \cdots \cup (a_r, b_r] : 0 \leq a_1 \leq b_1 \leq a_2 \cdots \leq b_r \leq 1\}$ 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

$$\mu_0(F) = \sum_{k\leq r} (b_k - a_k) \quad \text{for } F \in \mathcal{A}.$$

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 $(0, 1]$.

Note 1.43. We quickly recap Lebesgue measure on $\mathbb{R}$, as an example of completion.

Definition 1.44. Given a measure space $(E, \mathcal{E}, \mu)$, a **null set** is a subset $A \subset E$ such that there exists $B \in \mathcal{E}$ with $B \supset A$ and $\mu(B) = 0$. We often write $\mathcal{N}$ for the collection of such null sets.

Definition 1.45. A measure μ on $(E, \mathcal{E}, \mu)$ is called **complete** if $\mathcal{N} \subset \mathcal{E}$. In this case we call $(E, \mathcal{E}, \mu)$ a complete measure space.

Definition 1.46. Given a measure space $(E, \mathcal{E}, \mu)$, there exists a **completion** of $(E, \mathcal{E}, \mu)$ with respect to μ, which is a complete measure space $(E, \mathcal{E}^*, \mu^*)$ defined as follows. Let $\mathcal{N}$ denote the null sets of $(E, \mathcal{E}, \mu)$.

$$\begin{aligned} \mathcal{E}^* &= \{A \cup B : A \in \mathcal{E}, B \in \mathcal{N}\} \\ &= \sigma(\mathcal{E} \cup \mathcal{N}). \end{aligned}$$

For $A \cup B \in \mathcal{E}^*$ with $A \in \mathcal{E}, B \in \mathcal{N}$, we set $\mu^*(A \cup B) = \mu(A)$. We sometimes just write μ instead of μ^*.

Definition 1.47. Define a measure μ on the algebra $\mathcal{A}$ on $\mathbb{R}$ of finite disjoint unions of intervals $(a, b]$ in $\mathbb{R}$ by

$$\mu\left(\bigcup_{j=1}^{n}(a_j, b_j]\right) = \sum_{j=1}^{n}(b_j - a_j).$$

The completion of $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$ is called the **Lebesgue measure on** $\mathbb{R}$, and is denoted $(\mathbb{R}, \mathcal{M}(\mathbb{R}), \mu)$.

Corollary 1.48

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

2 Random Variables and Integration

2.1 Measurable Maps

Lemma 2.1 (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. $f^{-1}(A \cup B) = f^{-1}(A) \cup f^{-1}(B)$.
3. $f^{-1}(A \cap B) = f^{-1}(A) \cap f^{-1}(B)$.
4. $f^{-1}(A \setminus B) = f^{-1}(A) \setminus f^{-1}(B)$.
5. If $\mathcal{C} \subseteq \mathcal{P}(\Omega')$ then $f^{-1}(\sigma(\mathcal{C})) = \sigma\left(f^{-1}(\mathcal{C})\right)$.

Definition 2.2. 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

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

i.e. $f^{-1}(\mathcal{G}) \subseteq \mathcal{F}$.

Remark 2.3. 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 $m\mathcal{F}$.

Proposition 2.4

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 $f^{-1}(\mathcal{C}) \subseteq \mathcal{F}$.

Lemma 2.5 (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.

Definition 2.6. 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 it is $\mathcal{B}(\Omega), \mathcal{B}(\mathbb{R})$-measurable.

Lemma 2.7. If $f : \Omega \to \mathbb{R}$ is continuous then f is a **Borel function**.

2.2 Random Variables

Definition 2.8. Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. A function $X : \Omega \to \mathbb{R}$ such that it is $\mathcal{F}$-measureable (i.e. $\mathcal{F}, \mathcal{B}(\mathbb{R})$-measurable) is called a **random variable** (r.v.).

Remark 2.9. If X is a r.v. then $\{w \in \omega : X(w) \leq a\} \in \mathcal{F}$ so $\mathbb{P}(\{w \in \omega : X(w) \leq a\}) = \mathbb{P}(X \leq a)$ makes sense.

Definition 2.10 (σ-algebra generated by a r.v.)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X a random variable. Then the σ-algebra generated by X is

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

Note 2.11. More generally, $\sigma(X_i) = \sigma(\{(X_1, \dots, X_i) \in B\} : B \in \mathcal{B}(\mathbb{R}^i))$.

Proof. To show that $\sigma(X)$ is a σ-algebra, we must verify the following three properties:

1. $\sigma(X)$ contains Ω: By definition of a random variable, X is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Since $\mathbb{R} \in \mathcal{B}(\mathbb{R})$, we have
$$X^{-1}(\mathbb{R}) = \Omega \in \sigma(X).$$

2. $\sigma(X)$ is closed under complementation: Suppose $A \in \sigma(X)$, meaning there exists a Borel set $B \in \mathcal{B}(\mathbb{R})$ such that $A = X^{-1}(B)$. Then, the complement of A is given by:
$$A^c = (X^{-1}(B))^c = X^{-1}(B^c).$$
Since B^c is also a Borel set (because $\mathcal{B}(\mathbb{R})$ is a σ-algebra), it follows that $A^c \in \sigma(X)$. Thus, $\sigma(X)$ is closed under complements.

3. $\sigma(X)$ is closed under countable unions: Let $A_n \in \sigma(X)$ for $n \in \mathbb{N}$, so there exist Borel sets $B_n \in \mathcal{B}(\mathbb{R})$ such that $A_n = X^{-1}(B_n)$. Then, we consider the countable union:
$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} X^{-1}(B_n).$$
Since $\mathcal{B}(\mathbb{R})$ is a σ-algebra, it is closed under countable unions, so $\bigcup_{n=1}^{\infty} B_n \in \mathcal{B}(\mathbb{R})$. Hence,
$$X^{-1}\left(\bigcup_{n=1}^{\infty} B_n\right) = \bigcup_{n=1}^{\infty} X^{-1}(B_n) \in \sigma(X).$$

□

Example 2.12

If X is a dicrete random variable then $\sigma(X) = \sigma(\{X = x_1\}, \{X = x_2\}), \ldots$. However, if X is a continuous random variable then $\sigma(X)$ includes sets that look like $\{X \leq a\}, \{X > b\}$ or $\{a < X < b\}$.

Lemma 2.13

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We have that X is a random variable if and only if $\sigma(X) \subset \mathcal{F}$.

Definition 2.14 (σ-algebra generated by a family of r.v.s)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X_\alpha : \alpha \in I)$ a collection of random variables, then
$$\sigma(X_\alpha : \alpha \in I) = \sigma\left(\bigcup_{\alpha \in I} \sigma(X_\alpha)\right).$$

Lemma 2.15. Suppose $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{D})$ for some $\mathcal{D} \subset 2^{\mathbb{R}}$. Then a function $X : \Omega \to \mathbb{R}$ is a random variable if and only if
$$X^{-1}(D) \in \mathcal{F} \quad \text{for all } D \in \mathcal{D}.$$

Proof. We prove each direction in turn.

- Proof of ($\Rightarrow$).
Assume that X is $\mathcal{F}$-measurable. By definition, this means that for every Borel set $B \in \mathcal{B}(\mathbb{R})$, the preimage $X^{-1}(B)$ is in $\mathcal{F}$. Since $\sigma(\mathcal{D}) = \mathcal{B}(\mathbb{R})$, we have $\mathcal{D} \subseteq \mathcal{B}(\mathbb{R})$, which means that $X^{-1}(D) \in \mathcal{F}$ for all $D \in \mathcal{D}$.

- Proof of ($\Leftarrow$).
Suppose that $X^{-1}(D) \in \mathcal{F}$ for all $D \in \mathcal{D}$, and we want to show that X is $\mathcal{F}$-measurable, meaning that $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}(\mathbb{R})$. To do this, we define the collection:
$$\mathcal{E} = \{A \subseteq \mathbb{R} \mid X^{-1}(A) \in \mathcal{F}\}.$$
We claim that $\mathcal{E}$ is a σ-algebra on $\mathbb{R}$, and since $\mathcal{D} \subseteq \mathcal{E}$ and $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{D})$, it will follow that $\mathcal{B}(\mathbb{R}) \subseteq \mathcal{E}$, completing the proof. We now prove our claim.

- Since $X^{-1}(\mathbb{R}) = \Omega$, and we assume $\Omega \in \mathcal{F}$, it follows that $\mathbb{R} \in \mathcal{E}$.
- If $A \in \mathcal{E}$, then by definition $X^{-1}(A) \in \mathcal{F}$. Taking the complement:

$$X^{-1}(A^c) = \Omega \setminus X^{-1}(A),$$

which is in $\mathcal{F}$ since $\mathcal{F}$ is a σ-algebra. Thus, $A^c \in \mathcal{E}$.
- Suppose $A_1, A_2, \dots \in \mathcal{E}$. Then $X^{-1}(A_n) \in \mathcal{F}$ for each n. Since $\mathcal{F}$ is a σ-algebra, we have:

$$X^{-1}\left(\bigcup_{n=1}^{\infty} A_n\right) = \bigcup_{n=1}^{\infty} X^{-1}(A_n) \in \mathcal{F}.$$

Therefore, $\bigcup_{n=1}^{\infty} A_n \in \mathcal{E}$.

Since $\mathcal{E}$ is a σ-algebra and $\mathcal{D} \subseteq \mathcal{E}$, we conclude that the σ-algebra generated by $\mathcal{D}$, which is $\mathcal{B}(\mathbb{R})$, must also be contained in $\mathcal{E}$. That is,

$$\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{D}) \subseteq \mathcal{E}.$$

Thus, $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}(\mathbb{R})$, proving that X is $\mathcal{F}$-measurable.

□

Lemma 2.16

Given random variables $X_1, \dots, X_n, \dots$, the following are also random variables:

1. $X_1 + X_2$, $X_1 - X_2$, $X_1 X_2$, X_1/X_2 (if $X_2 \neq 0$), aX_1 for $a \in \mathbb{R}$.
2. $\max(X_1, X_2)$, $\min(X_1, X_2)$,
3. $\sup_n X_n$, $\inf_n X_n$, when they exist for every $\omega \in \Omega$,
4. $f(X_1)$ for any Borel-measurable function $f : \mathbb{R} \to \mathbb{R}$.

Definition 2.17. A random variable ζ is called **simple** if it only takes finitely many values i.e., there exists $a_1, \dots, a_n \in \mathbb{R}$ and $A_1, \dots, A_n \in \mathcal{F}$ such that

$$\zeta = \sum_{j=1}^{n} a_j \mathbf{1}_{A_j}.$$

Proposition 2.18. We can write $X = X^+ - X^-$ where $X^+ = \max(X, 0)$ and $X^- = -\min(f, 0)$.

Lemma 2.19. Given a random variable $X \geq 0$, there exists a sequence of simple random variables $\zeta_1, \zeta_2, \dots$ such that $\zeta_n \uparrow X$ (i.e., ζ_n increases monotonically and converges to X).

Remark 2.20. When X is not necessarily non-negative, we can write $X = X^+ - X^-$, where $X^+ = \max(X, 0)$ and $X^- = \min(X, 0)$, and apply the lemma separately to X^+ and X^-.

Note 2.21. This lemma formalises the idea that any measurable function can be built up from *stepwise approximations*, much like how digital screens approximate images using pixels.

Proof. We construct a sequence of simple functions that approximate X from below. The idea is to partition Ω based on binned values of X. Specifically, for $j, n \in \mathbb{N}$, define the measurable sets

$$A_{j,n} = \{X \in [j2^{-n}, (j+1)2^{-n})\}.$$

Since X is measurable, it follows that $A_{j,n} \in \mathcal{F}$. Now, define the simple function ζ_n as:

$$\zeta_n = \sum_{j=0}^{2^{2n}} j \cdot 2^{-n} \mathbf{1}_{A_{j,n}}.$$

Clearly, for every $\omega \in \Omega$, $\zeta_n(\omega)$ is monotone increasing and bounded above by $X(\omega)$. Also, for any $n > X(\omega)$, we have

$$X(\omega) \in [\zeta_n(\omega), \zeta_n(\omega) + 2^{-n}).$$

Thus, it follows that

$$|\zeta_n(\omega) - X(\omega)| \to 0 \quad \text{as } n \to \infty.$$

□

2.3 Distribution of Random Variables

Definition 2.22. If X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ into $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then the **distribution** (or **law**) of X is

$$\mathbb{P}_X(A) = \mathbb{P}(X^{-1}(A))$$

for any $A \in \mathcal{B}(\mathbb{R})$.

Remark 2.23. $\mathbb{P}_X$ is often called the push-forward of $\mathbb{P}$ under X.

Corollary 2.24. $\mathbb{P}_X$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Remark 2.25. This notion is generalised by changing the space, but we only offer the probabilistic definition.

2.3.1 Cumulative Distribution functions

Definition 2.26 (Distribution function of random variable)**.** For a real random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ define

$$F_X : \mathbb{R} \to [0,1]$$
$$F_X(x) = \mathbb{P}_X((-\infty, x]) = \mathbb{P}(X^{-1}(-\infty, x]).$$

This function is called a **(cumulative) distribution function**.

Theorem 2.27
The cdf satisfies the following properties:

1. F is non-decreasing, i.e. $x \leq y$ implies $F(x) \leq F(y)$.
2. $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.
3. F is right continuous, i.e. $\lim_{x\to y^+} F(x) = F(y)$.

Proof. We prove that this definition of the cdf satisfies each property.

1. Let $x_1 \leq x_2$. Then:
$$(-\infty, x_1] \subseteq (-\infty, x_2],$$
so by monotonicity of measures:
$$F_X(x_1) = \mathbb{P}_X((-\infty, x_1]) \leq \mathbb{P}_X((-\infty, x_2]) = F_X(x_2).$$
Thus, F_X is non-decreasing.

2. As $x \to \infty$:
$$\lim_{x\to\infty} F_X(x) = \lim_{x\to\infty} \mathbb{P}_X((-\infty, x]) = \mathbb{P}_X\left(\bigcup_{x=1}^{\infty}(-\infty, x]\right) = \mathbb{P}(\mathbb{R}) = 1.$$
As $x \to -\infty$:
$$\lim_{x\to-\infty} F_X(x) = \lim_{x\to-\infty} \mathbb{P}_X((-\infty, x]) = \mathbb{P}_X\left(\bigcap_{x=1}^{\infty}(-\infty, -x]\right) = \mathbb{P}(\varnothing) = 0.$$

3. We show that:
$$\lim_{n\to\infty} F_X\left(x + \frac{1}{n}\right) = F_X(x).$$
Note:
$$F_X\left(x + \frac{1}{n}\right) - F_X(x) = \mathbb{P}(x < X \leq x + \tfrac{1}{n}) = \mathbb{P}(X \in (x, x + \tfrac{1}{n}]).$$
Since $(x, x + \frac{1}{n}]$ is a decreasing sequence of sets as $n \to \infty$, we apply continuity from above:
$$\lim_{n\to\infty} \mathbb{P}(X \in (x, x + \tfrac{1}{n}]) = \mathbb{P}\left(\bigcap_{n=1}^{\infty}(x, x + \tfrac{1}{n}]\right) = \mathbb{P}(\varnothing) = 0.$$
So:
$$\lim_{n\to\infty} F_X\left(x + \tfrac{1}{n}\right) = F_X(x),$$
which proves right-continuity.

□

2.3.2 Discrete distribution functions

Definition 2.28. Discrete distribution functions are of the form

$$F_{\text{disc}}(x) = \sum_{j \in J} p_j \mathbf{1}_{\{x_j \leq x\}}(x)$$

where $p_j > 0$ and $\sum_{j \in J} p_j = 1$, which gives a distribution $\mathbb{P}_{\text{disc}}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ where

$$\mathbb{P}_{\text{disc}}(A) = \sum_{\substack{x \in J \\ x_j \in A}} p_j \quad \text{for } A \in \mathcal{B}(\mathbb{R});$$

(by 'disc' we mean a discrete random variable).

Note 2.29. That is, the distribution function is a step function with steps at each of the points x_i in J which are of "size" given by the probability of the point (**mass function**).

Example 2.30

Example distributions:

- **Uniform** on $\{1, \ldots, N\}$: we use the following data $J = \{1, \ldots, N\}, x_j = j$ and $p_j = 1/N$.
- **Bernoulli** with parameter $p \in [0, 1]$: we use the following data $J = \{0, 1\}, x_0 = 0, x_1 = 1$ and $p_0 = 1 - p, p_1 = p$.
- **Geometric** with parameter $p \in [0, 1]$: we use the following data $J = \mathbb{N}, x_j = j$ and $p_j = (1 - p)^{j-1} p$.
- **Poisson** with parameter $\lambda > 0$: we use the following data $J = \mathbb{N}, x_j = j$ and

$$p_j = \frac{e^{-\lambda} \lambda^j}{j!}.$$

Proposition 2.31. Properties of the discrete distribution.

- F_{disc} is not continuous.
- $F'_{\text{disc}} = 0$ is Lebesgue measurable almost everywhere.

2.3.3 Absolutely continuous distribution functions

Proposition 2.32

Suppose F_X is continuously differentiable, then it is the cdf of some random variable with **probability density function** given by

$$f(x) = F'_X(x).$$

Definition 2.33. **Absolutely continuous distribution functions** are of the form

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

where $f : \mathbb{R} \to [0, \infty)$ is Lebesgue measurable with $\int_{-\infty}^{\infty} f(x)\,dx = 1$. We call f the **probability density function** (pdf) of F (sometimes just density).

Example 2.34

Examples of AC distribution functions.

- **Uniform** on $[a, b]$: $f(x) = \frac{\mathbf{1}\{x\in[a,b]\}}{b-a}$.
- **Normal/Gaussian** with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$:

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

- **Exponential** with rate $\lambda > 0$: $f(x) = \lambda e^{-\lambda x}\mathbf{1}_{\{x\geq 0\}}$.

Proposition 2.35. F_{AC} is absolutely continuous with $F'(x) = f(x)$.

2.3.4 Singular continuous distributions functions

Definition 2.36. We say a measure μ on $(\Omega, \mathcal{F})$ is **concentrated** on a set $A \in \mathcal{F}$ if for every $B \in \mathcal{F}$ disjoint with A we have $\mu(B) = 0$.

Definition 2.37. We say a distribution function F_{SC} is **singular continuous** if

- F_{SC} is continuous,
- there exists $A \in \mathcal{B}(\mathbb{R})$ such A is of Lebesgue measure 0 and $\mathbb{P}_{\text{SC}}$ is concentrated on A.

Remark 2.38. When we imagine distribution functions, we often think of absolute continuous or discrete classes. When writing a proof be sure to consider the SC case.

Proposition 2.39. A continuous distribution function can either be singular or absolutely continuous.

Example 2.40

An example of a SC distribution would be the Cantor staircase.

Theorem 2.41 (Lebesgue decomposition)

Given a distribution function F on $\mathbb{R}$, there exists unique distribution functions $F_{\text{disc}}, F_{\text{AC}}$ and F_{SC} such that

$$F = p_{\text{disc}}F_{\text{disc}} + p_{\text{AC}}F_{\text{AC}} + p_{\text{SC}}F_{\text{SC}}$$

where $p_{\text{disc}}, p_{\text{AC}}, p_{\text{SC}} > 0$ and sum to 1.

Proof. Not examinable. Omitted. □

3 Integration and Expectation

3.1 Definitions

Definition 3.1. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. A function $\varphi : \Omega \to \mathbb{R}$ is called a **simple function** if there exists an $n \in \mathbb{N}$ and $E_1, \dots, E_n \in \mathcal{F}$, $a_1, \dots, a_n \in \mathbb{R}$ such that

$$\varphi = \sum_{k=1}^{n} a_k \mathbf{1}_{E_k}.$$

Note 3.2. That is, a simple function is a measurable step function taking only finitely many steps.

Definition 3.3. If φ is a non-negative simple function as above, then we define the integral of φ w.r.t. μ by

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

Definition 3.4. For a non-negative measurable function f (i.e. f is $\mathcal{F}$-measurable) we define the integral by

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

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

Definition 3.5. We say that a measurable function f (i.e. f is $\mathcal{F}$-measurable) is **integrable** if

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

(note this is well-defined since $|f|$ is $\mathcal{F}$-measurable). 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 $f^+ = \max(f, 0)$ and $f^- = -\min(f, 0)$ are the positive and negative parts of f, respectively.

Definition 3.6 (Notation for integral over a subset)**.** If $A \in \mathcal{F}$ and f is integrable, then we define

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

Maybe example, from Example 4.6 of Warwick notes

Definition 3.7. 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.

Remark 3.8. If $f = g$ a.e. (i.e. $\mu(\{\omega : f(\omega) \neq g(\omega)\}) = \mu(\{f \neq g\}) = 0$), then

$$\int f \, d\mu = \int g \, d\mu.$$

Note 3.9. It means that the set where the property fails is so small that it has no measurable impact. If something happens almost surely, it means it will happen except in an "infinitely rare" scenario.

Theorem 3.10 (Properties of integral)
Let f, g be integrable functions on a measure space $(\Omega, \mathcal{F}, \mu)$ and $\alpha \in \mathbb{R}$.

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

Remark 3.11. Note a little care has to be taken with our notation $\mu(\cdot)$, which has different meanings depending on what type of object the argument is, e.g. $\mu(\mathbf{1}_A) = \mu(A)$: the left-hand side is shorthand for the integral of a function, whereas the right-hand side is simply the measure of a set.

3.2 Key theorems

Definition 3.12. $f_n \to f$ **pointwise** if $f_n(\omega) \to f(\omega)$ for all $\omega \in \Omega$.

Definition 3.13. $f_n \to f$ **a.e.** (or a.s. for probability measure) if

$$\mu(\{\omega \in \Omega : f_n(\omega) \to f(\omega)\}^c) = 0.$$

Theorem 3.14 (Fatou's Lemma)
Let $(f_n)_{n\geq 1}$ be a sequence of non-negative *measurable* functions, then

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

Remark 3.15. Note we are not excluding the case when the integrals are infinite.

Note 3.16. Note that in words the result can be interpreted as 'some mass may escape to infinity.'

Theorem 3.17 (Monotone Convergence Theorem (MCT))
Let $(f_n)_{n\geq 1}$ be a sequence 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

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

Remark 3.18. Note we are not excluding the case when the integrals are infinite.

Corollary 3.19. 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

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

Theorem 3.20 (Dominated Convergence Theorem (DCT))
Let $(f_n)_{n\geq 1}$ be a sequence 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

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

Corollary 3.21 (Reverse Fatou's Lemma)**.** Let $(f_n)_{n\geq 1}$ be a sequence of *measurable* functions, and suppose that there exists an integrable function g such that $f_n \leq g$ for all n. Then

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

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 $\int g \, d\mu < \infty$, so we can cancel the contribution from these terms. $\square$

Product measures

Theorem 3.22 (Measure on product spaces)
Let $(A, \mathcal{A}, \mu_A)$ and $(B, \mathcal{B}, \mu_B)$ be measure spaces. Let $\mathcal{E} = \mathcal{A} \otimes \mathcal{B}$, then there exists a unique measure μ (sometimes denoted as $\mu_A \otimes \mu_B$) on $\mathcal{A} \otimes \mathcal{B}$ such that

$$\mu(S_1 \times S_2) = \mu_A(S_1)\mu_B(S_2).$$

Theorem 3.23 (Fubini's and Tonelli's theorems)
Let $(\Omega, \mathcal{F}, \mathbb{P}) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \otimes \mathbb{P}_2)$ be a product probability space, and let $f(\boldsymbol{w}) = f(w_1, w_2)$ be a measurable function on $(\Omega, \mathcal{F})$, the functions

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

are $\mathcal{F}_1$-$\mathcal{F}_2$ measurable respectively.
Suppose that either

1. f is integrable on Ω, or
2. $f \geq 0$.

Then

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

where in case (2.) values may be ∞.

Maybe examples?

3.3 Expectation

Definition 3.24. If X is random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with associated distribution $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ then we define the **expected value** of X with respect to $\mathbb{P}$ as

$$\mathbb{E}[X] = \int_\Omega X\, d\mathbb{P} = \int X(\omega)\mathbb{P}(d\omega) = \int_{\mathbb{R}} x d\mathbb{P}_X = \int_{\mathbb{R}} x f_X(x)\, dx,$$

where $f_X(x)$ is the density function (if it exists).

Note 3.25. We can abuse notation and think of $d\mathbb{P}_X = f_X(x)\, dx$.

Remark 3.26. The density function exists if it the distribution is not singular.

Corollary 3.27. In the discrete case, we define the expectation to be

$$\mathbb{E}[X] = \sum_x x p_X(x).$$

Definition 3.28. Given $p \geq 1$, a random variable X with $\mathbb{E}\left(|X|^p\right) < \infty$, we say X has finite p-th moment, and we often write $X \in L^p(\Omega)$.

Definition 3.29. Given a random variable X with $\mathbb{E}\left(|X|\right) < \infty$, the **expectation** or **mean** or first moment of X is defined as

$$\mu = \mu_X = \mathbb{E}(X).$$

Definition 3.30. We introduce some notation for the expectation of random variable **on some event** $A \in \mathcal{F}$

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

> **Remark 3.31.** This is not to be confused with the conditional expectation $\mathbb{E}[X \mid A]$ which we will talk about later.

Definition 3.32. If a random variable $X = X^+ - X^-$, we call X **integrable** if $\mathbb{E}(X^+) < \infty$ and $\mathbb{E}(X^-) < \infty$. In this case we can define

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-).$$

Lemma 3.33. 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 $\mathbb{E}[X] > 0$.

Lemma 3.34

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

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(X \in B), \quad B \in \mathcal{B}(\mathbb{R}).$$

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 $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$ and

$$\mathbb{E}[h(X)] = \mathbb{P}_X(h),$$

In particular, if it exists, we have

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

Lemma 3.35. If X is a non-negative random variable, then

$$\mathbb{E}[X] = 0 \iff \mathbb{P}(X > 0) = 0.$$

Proposition 3.36

If F is a distribution function of a random variable X, then

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

3.4 Some useful inequalities

Lemma 3.37 (Markov's inequality)
For any non-negative random variable X and $a > 0$ we have

$$\mathbb{P}(X \geq a) \leq \frac{1}{a}\mathbb{E}(X).$$

Remark 3.38. The lecture notes call this "Chebyshev's inequality".

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

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

Taking expectation on both sides yields,

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

which completes the proof. □

Lemma 3.39 (Chebyshev's inequality)
Suppose X is a random variable, then for any $b > 0$ we have

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq b\right) \leq \frac{\mathrm{Var(X)}}{b^2}.$$

Definition 3.40. A function $f : (a, b) \to \mathbb{R}$ is **convex**, if for any $x, y \in (a, b)$ and $p, q \in [0, 1]$ such that $p + q = 1$ we have

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

Note 3.41. In simple terms, a convex function graph is shaped like a cup $\cup$.

Proposition 3.42. Testing for convexity on an interval I.

- A funcion f is convex if $f''(x) \geq 0$ for all $x \in I$.
- A function f is concave if $f''(x) \leq 0$ for all $x \in I$.
- A function $f(x)$ is convex if and only if $-f(x)$ is concave.

Theorem 3.43 (Jensen's inequality)
Let $f : (a, b) \to \mathbb{R}$ be convex and X a random variable then

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

Proof. Omitted. Not examinable. □

Theorem 3.44 (Cauchy-Schwarz inequality)
If $\mathbb{E}(X^2) < \infty$ and $\mathbb{E}[Y^2] < \infty$ then XY is integrable and

$$|\mathbb{E}(XY)| \leq \mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

Proof. Omitted. Not examinable. □

4 Independence

4.1 Definition of Independence

Note 4.1. The rule of thumb should be familiar, "independence means we can multiply", i.e.

$$\mathbb{P}\left(\bigcap \cdot\right) = \prod \mathbb{P}(\cdot) \quad \text{and} \quad \mathbb{E}\left(\prod \cdot\right) = \prod \mathbb{E}(\cdot).$$

Definition 4.2 (Independent events)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, the events $(A_i)_{i=1}^n \in \mathcal{F}$ are **independent** if

$$\mathbb{P}\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n \mathbb{P}(A_i).$$

Definition 4.3 (Independent random variables)**.** The random variables X and Y are independent if for each pair of Borel sets $A, B \in \mathcal{B}(\mathbb{R})$ we have

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

In general the random variables $(X_i)_{i=1}^n$ are **independent** if for all $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$, we have

$$\mathbb{P}(X_1 \in B_1, \ldots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \cdots \mathbb{P}(X_n \in B_n).$$

Remark 4.4. The sequence of random variables is independent if and only if $\sigma(X_1), \ldots, \sigma(X_n)$ are independent.

Definition 4.5. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, then the sub-σ-algebras $\mathcal{G}_1, \mathcal{G}_2, \ldots, \subset \mathcal{F}$ are called **independent** if, for all $A_j \in \mathcal{G}_j$ the events $(A_j)_j$ are independent i.e.

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

Note 4.6. That is, I pick an event (set) from each subalgebra and compute its probability measure. If it is equal to the product then the subalgebras are independent.

Remark 4.7. This definition extends for σ-algebras too.

We illustrate theorems which will help when checking independence of σ-algebras.

Definition 4.8. A collection of sets $\mathcal{A}$ is called a π**-system** if $A, B \in \mathcal{A}$ implies $A \cap B \in \mathcal{A}$.

Definition 4.9. A collection of sets $\mathcal{B}$ is called a λ**-system** if

- $\Omega \in \mathcal{B}$,
- $A, B \in \mathcal{B}$ and $B \subset A \Rightarrow A \setminus B \in \mathcal{B}$,
- $A_1, \ldots, A_n, \ldots, \in \mathcal{B}$ and $A_n \uparrow A \Rightarrow A \in \mathcal{B}$.

Theorem 4.10 (Dynkin π-λ system)
If $\mathcal{A}$ is a π-system and $\mathcal{B}$ is a λ-system such that $\mathcal{A} \subset \mathcal{B}$, then $\sigma(\mathcal{A}) \subset \mathcal{B}$.

Proof. Not examinable. Omitted. □

Lemma 4.11
Suppose that the collection of π-systems $\mathcal{A}_1, \ldots, \mathcal{A}_n \subset \mathcal{F}$ are independent, then $\sigma(\mathcal{A}_1), \ldots, \sigma(\mathcal{A}_n)$ are independent.

Note 4.12. Independence of $\mathcal{A}_1, \ldots, \mathcal{A}_n$ means that for any $A_j \in \mathcal{A}_j$ (for each j), we have:

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2) \ldots \mathbb{P}(A_n).$$

The goal is to show that this property extends to all elements of the σ-algebras $\sigma(\mathcal{A}_j)$.

Proof. Fix $A_2, \ldots, A_n$ with $A_j \in \mathcal{A}_j$ for $2 \leq j \leq n$. Define $\mathcal{L}$ as the collection of events in $\mathcal{F}$ that are independent of $A_2, \ldots, A_n$, i.e.,

$$\mathcal{L} = \{A \in \mathcal{F} : \mathbb{P}(A \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A)\mathbb{P}(A_2) \ldots \mathbb{P}(A_n)\}.$$

We **claim** that $\mathcal{L}$ is a λ-system. To prove it. Define $F = A_2 \cap \cdots \cap A_n$, then:

1. **Containment of Ω:** Since $\Omega \in \mathcal{A}_1$, we have

$$\mathbb{P}(\Omega \cap F) = \mathbb{P}(\Omega)\mathbb{P}(F) = \mathbb{P}(F),$$

so $\Omega \in \mathcal{L}$.

2. **Closure under differences:** If $A, B \in \mathcal{L}$, then using the definition of independence and properties of probability:

$$\mathbb{P}((A \setminus B) \cap F) = \mathbb{P}(A \cap F) - \mathbb{P}((A \cap B) \cap F).$$

Since $A, B \in \mathcal{L}$, we substitute:

$$\mathbb{P}((A \cap B) \cap F) = \mathbb{P}(A \cap B)\mathbb{P}(F),$$

and thus,

$$\mathbb{P}((A \setminus B) \cap F) = (\mathbb{P}(A) - \mathbb{P}(A \cap B))\mathbb{P}(F) = \mathbb{P}(A \setminus B)\mathbb{P}(F).$$

Therefore, $A \setminus B \in \mathcal{L}$.

3. **Closure under increasing limits:** Suppose $B_1, B_2, \dots \in \mathcal{L}$ with $B_n \uparrow A$, then using monotone convergence:

$$\mathbb{P}((B_n \cap F) \uparrow (A \cap F)).$$

Since $B_n \in \mathcal{L}$,

$$\mathbb{P}(B_n \cap F) = \mathbb{P}(B_n)\mathbb{P}(F).$$

Taking limits,

$$\mathbb{P}(A \cap F) = \lim_{n\to\infty} \mathbb{P}(B_n \cap F) = \lim_{n\to\infty} \mathbb{P}(B_n)\mathbb{P}(F) = \mathbb{P}(A)\mathbb{P}(F).$$

Thus, $A \in \mathcal{L}$.

Since $\mathcal{L}$ satisfies the three properties of a λ-system and contains $\mathcal{A}_1$, by the π-λ theorem, we conclude that:

$$\sigma(\mathcal{A}_1) \subset \mathcal{L}.$$

Since $A_2, \dots, A_n$ were chosen arbitrarily, the same argument applies inductively to show that:

$$\sigma(\mathcal{A}_1), \sigma(\mathcal{A}_2), \dots, \sigma(\mathcal{A}_n) \quad \text{are independent.}$$

□

Definition 4.13 (Notation)**.** We write $\mathcal{B}_b = \mathcal{B}(\mathbb{R})$ for a **bounded** Borel measurable function.

Lemma 4.14

A family of random variables $X_1, \dots, X_n$ are independent if and only if for any bounded measurable functions $f_1, \dots, f_n \in \mathcal{B}_b(\mathbb{R})$, the expectation satisfies:

$$\mathbb{E}\left(f_1(X_1)\cdots f_n(X_n)\right) = \mathbb{E}(f_1(X_1))\cdots\mathbb{E}(f_n(X_n)).$$

Proof. We prove both directions separately.

- Proof of ($\Leftarrow$). To show independence, we choose specific indicator functions:
  - Take $f_i = \mathbf{1}_{B_i}$ for Borel sets $B_1, \dots, B_n$.
  - Then the expectation identity reduces to:

  $$\mathbb{E}[\mathbf{1}_{B_1}(X_1)\cdots\mathbf{1}_{B_n}(X_n)] = \mathbb{E}[\mathbf{1}_{B_1}(X_1)]\cdots\mathbb{E}[\mathbf{1}_{B_n}(X_n)].$$

  - By the definition of expectation, this simplifies to:

  $$\mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1)\cdots\mathbb{P}(X_n \in B_n).$$

  This is precisely the definition of independence.

- Proof of ($\Rightarrow$). The proof proceeds by approximation:

- First, recall that any bounded measurable function f_i can be approximated by simple functions:
$$f_i^{(m)} = \sum_{k=1}^{m} c_k \mathbf{1}_{B_k}$$
for some scalars c_k and Borel sets B_k.
- By linearity of expectation:
$$\mathbb{E}\left(f_1^{(m)}(X_1)\cdots f_n^{(m)}(X_n)\right) = \sum_{k_1,\dots,k_n} c_{k_1}\cdots c_{k_n}\mathbb{E}(\mathbf{1}_{B_{k_1}}(X_1)\cdots\mathbf{1}_{B_{k_n}}(X_n)).$$
- By independence:
$$\mathbb{E}(\mathbf{1}_{B_{k_1}}(X_1)\cdots\mathbf{1}_{B_{k_n}}(X_n)) = \mathbb{E}(\mathbf{1}_{B_{k_1}}(X_1))\cdots\mathbb{E}(\mathbf{1}_{B_{k_n}}(X_n)).$$
- Substituting this back:
$$\mathbb{E}\left(f_1^{(m)}(X_1)\cdots f_n^{(m)}(X_n)\right) = \left(\sum_{k_1} c_{k_1}\mathbb{E}[\mathbf{1}_{B_{k_1}}(X_1)]\right)\cdots\left(\sum_{k_n} c_{k_n}\mathbb{E}[\mathbf{1}_{B_{k_n}}(X_n)]\right).$$
- Since simple functions converge to f_i in expectation,
$$\mathbb{E}\left[f_1(X_1)\cdots f_n(X_n)\right] = \mathbb{E}[f_1(X_1)]\cdots\mathbb{E}[f_n(X_n)].$$

□

Definition 4.15. A sequence of random variables $(X_i)_{i=1}^n$ are **pairwise independent** if for all $i \neq j$ we have X_i and X_j a re independent.

Definition 4.16. Given two random variables, X and Y, with finite second moment, we define the **covariance** of X and Y by
$$\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$
We also write $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$.

> **Note 4.17.** This equality holds because $\mathbb{E}[c] = c$ for a constant c, thus $\mathbb{E}[X\cdot\mathbb{E}[Y]] = \mathbb{E}[X]\mathbb{E}[Y]$, since $\mathbb{E}[Y]$ is now treated as a constant.

Definition 4.18. Therefore, the **variance** is defined as
$$\mathrm{Var(X)} = \mathbb{E}\left[(X-\mathbb{E}[X])^2\right].$$

Definition 4.19. A family of random variables $X_1,\dots,X_n$ with finite second moment are said to be **uncorrelated** if for all $i \neq j$ we have that
$$\mathbb{E}(X_iX_j) = \mathbb{E}(X_i)\mathbb{E}(X_j),$$
that is $\mathrm{Cov}(X_i,X_j) = 0$.

Proposition 4.20

A standard trick.
$$\mathbb{E}[|Z|] \leq \sqrt{\mathbb{E}(Z^2)} = \sqrt{\mathrm{Var}(Z)}$$
when $\mathbb{E}[Z] = 0$.

Proof. By Jensen's inequality. □

4.2 Independence in product spaces

Lemma 4.21

Given random variables $X_1, \ldots, X_n$ the map

$$\begin{aligned}\Omega &\to \mathbb{R}^n \\ \omega &\mapsto \boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))\end{aligned}$$

is $\mathcal{F}$-measurable, that is $\boldsymbol{X}^{-1}(B) \in \mathcal{F}$ for every $B \in \mathcal{B}(\mathbb{R}^n)$.

The map $\boldsymbol{X}$ is called a random vector, and it induces a joint measure $\mathbb{P}_{\boldsymbol{X}}$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ by setting

$$\mathbb{P}_{\boldsymbol{X}} = \mathbb{P}(\boldsymbol{X}^{-1}(B))$$

for $B \in \mathcal{B}(\mathbb{R}^n)$.

Proof. Since each $X_j : \Omega \to \mathbb{R}$ is a random variable, we know that X_j is $\mathcal{F}$-measurable, meaning that for any Borel set $B_j \in \mathcal{B}(\mathbb{R})$, we have:

$$X_j^{-1}(B_j) \in \mathcal{F}.$$

Since the product σ-algebra $\mathcal{B}(\mathbb{R}^n)$ is generated by sets of the form

$$B = B_1 \times B_2 \times \cdots \times B_n,$$

where each $B_j \in \mathcal{B}(\mathbb{R})$, we examine the preimage:

$$\boldsymbol{X}^{-1}(B) = \{\omega \in \Omega \mid \boldsymbol{X}(\omega) \in B_1 \times B_2 \times \cdots \times B_n\}.$$

By definition of $\boldsymbol{X}$,

$$\boldsymbol{X}^{-1}(B) = \bigcap_{j=1}^{n} X_j^{-1}(B_j).$$

Since each $X_j^{-1}(B_j) \in \mathcal{F}$ and $\mathcal{F}$ is a σ-algebra (hence closed under finite intersections), it follows that $\boldsymbol{X}^{-1}(B) \in \mathcal{F}$ for all such rectangles B. Since the collection of these rectangles generates $\mathcal{B}(\mathbb{R}^n)$, by the definition of a σ-algebra, we conclude that

$$\boldsymbol{X}^{-1}(B) \in \mathcal{F} \quad \forall B \in \mathcal{B}(\mathbb{R}^n).$$

Thus, $\boldsymbol{X}$ is $\mathcal{F}$-measurable. □

Lemma 4.22

Given $\boldsymbol{X}$ as above, the following are equivalent.

1. $X_1, \ldots, X_n$ are independent.
2. $\mathbb{P}_{\boldsymbol{X}} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$.
3. Writing $F_{\boldsymbol{X}}$ for the multivariate c.d.f. for $\boldsymbol{X}$ and F_{X_j} for the c.d.f. for X_j

$$F_{\boldsymbol{X}}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n).$$

Proof. We prove the equivalence of the three conditions.

- (1) ⇒ (2):
  By definition, $X_1, \dots, X_n$ are independent if and only if for all Borel sets $B_1, \dots, B_n$,
  $$\mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \cdots \mathbb{P}(X_n \in B_n).$$
  This is precisely the definition of the product measure $\mathbb{P}_{\boldsymbol{X}} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$.

- (2) ⇒ (3):
  By definition, the cumulative distribution function (CDF) is given by
  $$F_{\boldsymbol{X}}(x_1, \dots, x_n) = \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n).$$
  Since the joint measure factorizes as $\mathbb{P}_{\boldsymbol{X}} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$, we obtain
  $$F_{\boldsymbol{X}}(x_1, \dots, x_n) = \mathbb{P}_{X_1}((-\infty, x_1]) \cdots \mathbb{P}_{X_n}((-\infty, x_n]).$$
  Since the marginal CDFs are given by $F_{X_j}(x_j) = \mathbb{P}_{X_j}((-\infty, x_j])$, we conclude:
  $$F_{\boldsymbol{X}}(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n).$$

- (3) ⇒ (1):
  Assume the CDF factorises:
  $$F_{\boldsymbol{X}}(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n).$$
  Then, for any intervals $(-\infty, x_j]$, we have
  $$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \cdots \mathbb{P}(X_n \leq x_n).$$
  Since this holds for all such intervals, the X_j's are independent. □

5 Modes of Convergence

5.1 Definitions

Definition 5.1. 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 almost surely** to a random variable X if and only if
$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

> **Remark 5.2.** 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 $X_n \xrightarrow{\text{a.s.}} X$ a.s. and $X_n \xrightarrow{\text{a.s.}} Y$, then it must be that $X = Y$ a.s.

Definition 5.3. 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 $\forall \varepsilon > 0$,
$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = \lim_{n\to\infty} \mathbb{P}(\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon) = 0.$$

Note 5.4. Think about predicting tomorrow's temperature. Suppose each day's forecast $f_n(x)$ is getting better at approximating the actual temperature $f(x)$. If there are occasional bad predictions, but they affect fewer and fewer locations, then the predictions **converge in measure** (/probability if measure is on probability space) to the true temperature.

Definition 5.5. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For $p \geq 1$, we define the L^p-space as

$$L^p(\Omega) = \{X : \Omega \to \mathbb{R} : \mathbb{E}\left(|X|^p\right) < \infty\}.$$

A random variable X belongs to $L^p(\Omega)$ if its p-th moment is finite.

Note 5.6. The L^p-spaces provide a natural setting for studying functions and random variables under an integrability framework.

Definition 5.7. For $p \geq 1$, we say a sequence of random variables $(X_n)_{n=1}^{\infty} \in L^p(\Omega)$ **converges in** L^p to a random variable X if

$$\lim_{n\to\infty} \mathbb{E}\left(|X_n - X|^p\right) = 0.$$

Lemma 5.8. Convergence in $L^p(\Omega)$ implies convergence in probability.

Proof. By Markov's inequality, for any $p > 0$:

$$\mathbb{P}(|X_n - X| > \varepsilon) \leq \frac{\mathbb{E}(|X_n - X|^p)}{\varepsilon}.$$

Since $\mathbb{E}(|X_n - X|^p) \to 0$ as $n \to \infty$, it follows that:

$$\mathbb{P}(|X_n - X| > \varepsilon) \leq \mathbb{E}(|X_n - X|^p)\varepsilon^{-1} \to 0.$$

Thus, $X_n \to X$ in probability. $\square$

Exam Questions 5.9 (Exercise in lecture notes)
Show that convergence almost surely implies convergence in probability, but that the converse is not true.

Solution. We prove each part separately.

- Fix $\varepsilon > 0$. Since $X_n \to X$ almost surely, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\{\omega : |X_n(\omega) - X(\omega)| \le \varepsilon, \forall n \ge m_\varepsilon(\omega)\}\right) = 1$$
$$\mathbb{P}\left(\lim_{n\to\infty} \bigcap_{n=m}^{\infty} \{\omega : |X_n(\omega) - X(\omega)| \le \varepsilon\}\right) = 1.$$

  Equivalently,

$$\mathbb{P}\left(\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{\omega : |X_n(\omega) - X(\omega)| \le \varepsilon\}\right) = 1,$$

  so taking complements:

$$\mathbb{P}\left(\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} \{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}\right) = 0.$$

  Therefore, for all $\varepsilon > 0$,

$$\lim_{m\to\infty} \mathbb{P}\left(\bigcup_{n=m}^{\infty} \{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}\right) = 0,$$

  which implies:
$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| > \varepsilon\right) = 0.$$

  Thus, almost sure convergence implies convergence in probability.

- Consider the following counterexample. Define a sequence of random variables on $[0, 1]$ equipped with the Lebesgue measure:

$$X_n(\omega) = \begin{cases} 1 & \text{if } \omega \in \left[\frac{j}{2^k}, \frac{j+1}{2^k}\right] \text{ for } j = 0, \dots, 2^k - 1, \\ 0 & \text{otherwise.} \end{cases}$$

  where $n = 2^k + j$ and $k := \lfloor \log_2(n) \rfloor$. We claim: $X_n \to 0$ in probability, but not almost surely. For any $\varepsilon > 0$, since $X_n \in \{0, 1\}$,

$$\mathbb{P}(|X_n| > \varepsilon) = \mathbb{P}(X_n = 1) = \frac{1}{2^k} \to 0 \quad \text{as } n \to \infty,$$

  so $X_n \xrightarrow{\mathbb{P}} 0$. However, for any fixed $\omega \in [0, 1]$, the dyadic intervals $\left[\frac{j}{2^k}, \frac{j+1}{2^k}\right]$ cover $[0, 1]$ as $k \to \infty$, so ω lies in infinitely many such intervals. Hence $X_n(\omega) = 1$ infinitely often, and
$$\mathbb{P}\left(\lim_{n\to\infty} X_n = 0\right) = 0 \ne 1.$$

  Therefore, $X_n \xrightarrow{\mathbb{P}} 0$, but $X_n \not\to 0$ almost surely.

Exam Questions 5.10 (Exercise from lecture notes)
Given a random variable X, show that one can construct a new probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ containing a sequence of independent and identically distributed (iid) random variables $X_1, X_2, \dots$ such that

$$\mathbb{P}_X = \tilde{\mathbb{P}}_{X_j} \quad \text{for all } j.$$

Solution. Define the sample space as $\tilde{\Omega} = \mathbb{R}^{\mathbb{N}}$, the space of all infinite sequences $\tilde{\omega} = (\omega_1, \omega_2, \dots)$ where each ω_j corresponds to an outcome of X_j. Equip this space with the product σ-algebra $\tilde{\mathcal{F}} = \mathcal{B}(\mathbb{R})^{\otimes \mathbb{N}}$, which is the smallest σ-algebra making the coordinate projections $\pi_j : \tilde{\Omega} \to \mathbb{R}$, given by $\pi_j(\tilde{\omega}) = \omega_j$, measurable.

Let $\mathbb{P}_X$ be the probability measure induced by X. Define the infinite product measure $\tilde{\mathbb{P}} = \mathbb{P}_X^{\otimes \mathbb{N}}$ on $(\tilde{\Omega}, \tilde{\mathcal{F}})$. By the definition of the product measure, for any finite subset of indices $\{1, \dots, n\}$ and any Borel sets $B_1, \dots, B_n$, the measure satisfies

$$\tilde{\mathbb{P}}(X_1 \in B_1, \dots, X_n \in B_n) = \mathbb{P}_X(B_1) \cdots \mathbb{P}_X(B_n).$$

This ensures that the coordinate random variables $X_j = \pi_j$ are independent and identically distributed, each satisfying $\mathbb{P}_{X_j} = \mathbb{P}_X$.

5.2 L^2 Weak law of large numbers

Definition 5.11. We say an infinite family of random variables $(X_i : i \in I)$ is **uncorrelated** (resp. **independent**) if for **every** finite $J \subset I$, the subfamily $(X_j : j \in J)$ is uncorrelated (resp. independent).

Lemma 5.12. Given a sequence of random variables $(X_n)_{n \geq 1}$ with finite second moment and uncorrelated, we have

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i).$$

Proof. By the definition of variance, we have:

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathbb{E}\left(\left[\sum_{i=1}^{n} X_i - \mathbb{E}\left(\sum_{i=1}^{n} X_i\right)\right]^2\right).$$

Using the linearity of expectation, we rewrite:

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}(X_i).$$

Thus, substituting this back:

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathbb{E}\left(\left[\sum_{i=1}^{n} (X_i - \mathbb{E}(X_i))\right]^2\right).$$

Expanding the square,

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathbb{E}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}(X_i))^2 + 2\sum_{1\le i<j\le n}(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))\right).$$

Using the linearity of expectation:

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\mathbb{E}\left((X_i - \mathbb{E}(X_i))^2\right) + 2\sum_{1\le i<j\le n}\mathbb{E}\left((X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))\right).$$

By definition, the first sum represents the variances:

$$\sum_{i=1}^{n}\operatorname{Var}(X_i).$$

The second sum represents the covariances between different variables:

$$2\sum_{1\le i<j\le n}\operatorname{Cov}(X_i, X_j).$$

Since the X_i are independent, we have $\operatorname{Cov}(X_i, X_j) = 0$ for all $i \neq j$, which simplifies the expression to:

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\operatorname{Var}(X_i).$$

□

Exam Questions 5.13 (Exercise from lecture notes)
Given a random variable X with finite second moment and any $c \in \mathbb{R}$, one has

$$\operatorname{Var}(cX) = c^2\operatorname{Var}(X).$$

Solution. Using the definition of variance and linearity of expectation,

$$\operatorname{Var}(cX) = \mathbb{E}\left((cX - \mathbb{E}(cX))^2\right) = \mathbb{E}\left((c(X - \mathbb{E}(X)))^2\right).$$

Factoring out c^2 and using the definition of variance,

$$\operatorname{Var}(cX) = c^2\mathbb{E}\left((X - \mathbb{E}(X))^2\right) = c^2\operatorname{Var}(X).$$

Theorem 5.14 (L^2 Weak law of large numbers)
Let $X_1, X_2, \ldots$ be a sequence of identically distributed, uncorrelated random variables with finite second moment. Then, the sample mean

$$\overline{X}_n := \frac{1}{n}\sum_{j=1}^{n} X_j \to \mathbb{E}(X_1) \quad \text{in } L^2 \text{ and in probability.}$$

> **Remark 5.15.** Since the random variables $X_1, X_2, \ldots$ are identically distributed, their means and variances are the same for all j. Therefore, rather than writing $\mathbb{E}(X_j)$ and $\operatorname{Var}(X_j)$ for each j, we simplify the notation by defining $\mathbb{E}(X_1)$ and $\operatorname{Var}(X_1)$, which remain valid for all j due to the identical distribution assumption.

Proof. We first prove that $\overline{X}_n \to \mu = \mathbb{E}(X_1)$ in L^2, meaning that $\mathbb{E}(|\overline{X}_n - \mu|^2) \to 0$ as $n \to \infty$. Using the definition of variance, we compute:

$$\mathbb{E}\left((\overline{X}_n - \mu)^2\right) = \operatorname{Var}(\overline{X}_n).$$

Expanding $\overline{X}_n$ in terms of its definition,

$$\operatorname{Var}(\overline{X}_n) = \operatorname{Var}\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right).$$

Using the variance scaling property $\operatorname{Var}(cY) = c^2 \operatorname{Var}(Y)$,

$$\operatorname{Var}(\overline{X}_n) = \frac{1}{n^2}\operatorname{Var}\left(\sum_{j=1}^{n} X_j\right).$$

Since the X_j are uncorrelated, we use the additivity of variance,

$$\operatorname{Var}\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} \operatorname{Var}(X_j).$$

By the assumption that the X_j are identically distributed, all have the same variance $\operatorname{Var}(X_j) = \operatorname{Var}(X_1)$, so

$$\operatorname{Var}(\overline{X}_n) = \frac{1}{n^2}\sum_{j=1}^{n} \operatorname{Var}(X_1) = \frac{1}{n}\operatorname{Var}(X_1).$$

Since $\operatorname{Var}(X_1)$ is finite by assumption, it follows that

$$\operatorname{Var}(\overline{X}_n) \to 0 \quad \text{as } n \to \infty.$$

Thus, $\overline{X}_n \to \mu$ in L^2, proving convergence in mean square.

Since convergence in L^2 implies convergence in probability, we conclude that $\overline{X}_n \to \mu$ in probability as well.

$\square$

5.2.1 Applications

> **Note 5.16.** We use the WLLN to prove the following theorem.

Theorem 5.17. Let f be a continuous function on $[0, 1]$. Then for every $\varepsilon > 0$, there exists a polynomial g such that

$$\sup_{x \in [0,1]} |f(x) - g(x)| < \varepsilon.$$

Theorem 5.18

Let f be a continuous function on $[0,1]$. Then for every $n \in \mathbb{N}$, the corresponding Bernstein polynomial f_n is defined by

$$f_n(p) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k} p^k (1-p)^{n-k}.$$

These polynomials satisfy the uniform approximation property:

$$\sup_{p\in[0,1]} |f(p) - f_n(p)| \to 0 \quad \text{as } n \to \infty.$$

Proof. Let $\varepsilon > 0$. We aim to show that there exists N such that for all $n \geq N$,

$$\sup_{p\in[0,1]} |f(p) - f_n(p)| < \varepsilon.$$

The proof follows from an application of the weak law of large numbers.

Fix p and consider a sequence of i.i.d. random variables $X_1, X_2, \dots$ such that $\mathbb{P}(X_1 = 1) = p$ and $\mathbb{P}(X_1 = 0) = 1 - p$. Then we have

$$\mathbb{E}(X_j) = p, \quad \operatorname{Var}(X_j) = p(1-p).$$

Define $S_n = \sum_{k=1}^{n} X_k$. Then S_n follows a Binomial distribution,

$$\mathbb{P}(S_n = m) = \binom{n}{m} p^m (1-p)^{n-m}.$$

Taking expectation of $f(S_n/n)$,

$$\mathbb{E}[f(S_n/n)] = \sum_{m=0}^{n} f(m/n)\mathbb{P}(S_n = m) = \sum_{m=0}^{n} f(m/n)\binom{n}{m} p^m (1-p)^{n-m} = f_n(p).$$

Since the weak law of large numbers (WLLN) tells us that $S_n/n \to p$ in probability, we want to estimate $|f(S_n/n) - f(p)|$ uniformly in p. Since f is uniformly continuous on $[0,1]$, there exists $\delta > 0$ such that

$$|x - y| < \delta \Rightarrow |f(x) - f(y)| < \varepsilon/2.$$

Since $S_n/n \to p$ in probability, we bound the probability of deviations using L^2-WLLN:

$$\mathbb{P}(|S_n/n - p| > \delta) \leq \frac{1}{\delta^2}\mathbb{E}[(S_n/n - p)^2] = \frac{1}{\delta^2}\operatorname{Var}(S_n/n) = \frac{1}{2\delta^2 n}.$$

Since $\operatorname{Var}(X_1) = p(1-p)$ is bounded independently of p, choosing

$$N = \delta^{-2}(2\|f\|_\infty + 1)^{-1}$$

ensures that for $n > N$,

$$\begin{aligned} |f_n(p) - f(p)| &= \mathbb{E}|f(S_n/n) - f(p)| \\ &\leq \mathbb{E}|f(S_n/n) - f(p)|1_{\{|S_n/n - p| \leq \delta\}} + \mathbb{E}|f(S_n/n) - f(p)|1_{\{|S_n/n - p| > \delta\}}. \end{aligned}$$

Using uniform continuity,

$$< \varepsilon/2 + (2\|f\|_\infty)\mathbb{P}(|S_n/n - p| > \delta).$$

Since $\mathbb{P}(|S_n/n - p| > \delta) \to 0$, we obtain

$$|f_n(p) - f(p)| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Thus, f_n converges uniformly to f. □

5.2.2 Truncation and WLLN 2

Note 5.19. Truncation is an important tool in probability theory that allows us to extend results that rely on moment conditions to sequences of random variables that may not satisfy those moment conditions. This version of the weak law of large numbers imposes a stronger independence condition but a weaker integrability assumption.

Theorem 5.20 (WLLN with Truncation)**.** Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. Suppose that $b_n > 0$ with $b_n \to \infty$ and define the truncated variables

$$\overline{X}_{n,k} = X_k \mathbf{1}_{\{|X_k| \le b_n\}}.$$

Additionally, assume the following conditions hold:

$$n \cdot \mathbb{P}(|X_1| > b_n) \to 0, \quad \text{and} \quad \frac{n}{b_n^2}\mathbb{E}[\overline{X}_{n,1}^2] \to 0.$$

Then, defining the sum

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

we have that:

$$\frac{1}{b_n}\left(S_n - n\mathbb{E}[\overline{X}_{n,1}]\right) \to 0 \quad \text{in probability.}$$

Proof. Let $\varepsilon > 0$. Our goal is to show that as $n \to \infty$,

$$\mathbb{P}\left(\frac{1}{b_n}|S_n - \mathbb{E}[\overline{S}_n]| > \varepsilon\right) \to 0.$$

Applying the union bound, we obtain the estimate

$$\mathbb{P}\left(\frac{1}{b_n}|S_n - \mathbb{E}[\overline{S}_n]| > \varepsilon\right) \le \mathbb{P}\left(\frac{1}{b_n}|\overline{S}_n - \mathbb{E}[\overline{S}_n]| > \varepsilon\right) + \mathbb{P}(S_n \neq \overline{S}_n).$$

To bound $\mathbb{P}(S_n \neq \overline{S}_n)$, we observe that $S_n \neq \overline{S}_n$ if and only if at least one X_k exceeds the truncation threshold b_n. Using the union bound again, we obtain

$$\mathbb{P}(S_n \neq \overline{S}_n) \le \mathbb{P}\left(\bigcup_{k=1}^{n}\{X_k \neq \overline{X}_{n,k}\}\right) \le n\mathbb{P}(|X_1| > b_n).$$

By assumption, $n\mathbb{P}(|X_1| > b_n) \to 0$, so this term vanishes.

To bound $\mathbb{P}\left(\frac{1}{b_n}|\overline{S}_n - \mathbb{E}[\overline{S}_n]| > \varepsilon\right)$, we apply Chebyshev's inequality:

$$\mathbb{P}\left(\frac{1}{b_n}|\overline{S}_n - \mathbb{E}[\overline{S}_n]| > \varepsilon\right) \le \frac{1}{\varepsilon^2 b_n^2}\mathbb{E}\left[(\overline{S}_n - \mathbb{E}[\overline{S}_n])^2\right].$$

Using $\operatorname{Var}(\overline{S}_n) = n \operatorname{Var}(\overline{X}_{n,1})$, we substitute to get

$$\mathbb{P}\left(\frac{1}{b_n}|\overline{S}_n - \mathbb{E}[\overline{S}_n]| > \varepsilon\right) \le \frac{n}{\varepsilon^2 b_n^2}\operatorname{Var}(\overline{X}_{n,1}).$$

Since $\operatorname{Var}(\overline{X}_{n,1}) \le \mathbb{E}[\overline{X}_{n,1}^2]$, we obtain

$$\mathbb{P}\left(\frac{1}{b_n}|\overline{S}_n - \mathbb{E}[\overline{S}_n]| > \varepsilon\right) \le \frac{n}{\varepsilon^2 b_n^2}\mathbb{E}[\overline{X}_{n,1}^2].$$

By assumption, $\frac{n}{b_n^2}\mathbb{E}[\overline{X}_{n,1}^2] \to 0$, so the probability vanishes.

Since both terms in our bound vanish as $n \to \infty$, we conclude that

$$\mathbb{P}\left(\frac{1}{b_n}|S_n - \mathbb{E}[\overline{S}_n]| > \varepsilon\right) \to 0.$$

$\square$

Theorem 5.21 (WLLN2)
Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}(|X_j|) < \infty$. Define the partial sum

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

Then, the weak law of large numbers states that

$$\frac{1}{n}S_n - \mathbb{E}(X_1) \to 0 \quad \text{in probability.}$$

Proof. We choose the truncation sequence $b_n = n$ and observe that, from the conclusion of the weak law of large numbers (WLLN) with truncation, we have $\mathbb{E}[\overline{X}_{n,1}] \to \mathbb{E}[X_1]$ by the dominated convergence theorem. To complete the proof, we verify the two key assumptions required in the truncated version of WLLN.

To verify that $n\mathbb{P}(|X_1| > n) \to 0$, we use Chebyshev's inequality, which states that for any random variable X_1,

$$\mathbb{P}(|X_1| > n) \le \frac{\mathbb{E}[|X_1|]}{n}.$$

Multiplying by n on both sides gives

$$n\mathbb{P}(|X_1| > n) \le \mathbb{E}[|X_1|\mathbf{1}(|X_1| > n)].$$

Since $\mathbb{E}[|X_1|] < \infty$, the right-hand side vanishes as $n \to \infty$ by dominated convergence. Hence, the first assumption holds.

Next, we verify that $\frac{n}{b_n^2}\mathbb{E}[\overline{X}_{n,1}^2] \to 0$. Since $b_n = n$, this reduces to proving that

$$\frac{1}{n}\mathbb{E}[\overline{X}_{n,1}^2] \to 0.$$

Using integration by parts, we express the expectation as

$$\mathbb{E}[\overline{X}_{n,1}^2] = \int_0^\infty 2y\mathbb{P}(|\overline{X}_{n,1}| > y)dy.$$

Since $\overline{X}_{n,1}$ is truncated at n, we split the integral as

$$\mathbb{E}[\overline{X}_{n,1}^2] = \int_0^n 2y\mathbb{P}(|X_1| > y)dy.$$

Changing variables by setting $ny = u$, we obtain

$$\frac{1}{n}\int_0^1 2n\tilde{y}\mathbb{P}(|X_1| > n\tilde{y})d\tilde{y}.$$

To apply the bounded convergence theorem, we note that by Chebyshev's inequality,

$$2n\tilde{y}\mathbb{P}(|X_1| > n\tilde{y}) \leq 2\mathbb{E}[|X_1|].$$

Since $n\mathbb{P}(|X_1| > n) \to 0$, we conclude that for any fixed $\tilde{y}$, the term $n\tilde{y}\mathbb{P}(|X_1| > n\tilde{y}) \to 0$ as $n \to \infty$. The bounded convergence theorem then ensures that

$$\frac{1}{n}\mathbb{E}[\overline{X}_{n,1}^2] \to 0.$$

Since both assumptions hold, we conclude that the WLLN with truncation implies WLLN2. □

5.3 Borel-Cantelli Lemmas

Definition 5.22 (lim sup/lim inf of events)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(A_n)_n \in \mathcal{F}$, we define the following events:

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

Note 5.23. Intuitively, $\limsup A_n$ captures recurrence, while $\liminf A_n$ describes stabilisation.

Corollary 5.24

By definition $\limsup_n A_n, \liminf_n A_n \in \mathcal{F}$.

Lemma 5.25. We have that

$$\mathbf{1}_{\limsup_n A_n} = \limsup_{n\to\infty} \mathbf{1}_{A_n} \quad \text{and} \quad \mathbf{1}_{\liminf_n A_n} = \liminf_{n\to\infty} \mathbf{1}_{A_n}.$$

Proof. Omitted. $\square$

Exam Questions 5.26

Let X and X_n be random variables. Let $\varepsilon_n > 0$ with $\varepsilon_n \to 0$ and define

$$A_n = \{|X - X_n| \geq \varepsilon_n\}.$$

Then,

$$\{X_n \to X\} \supset \{A_n \text{ i.o.}\}^c.$$

For any $\varepsilon > 0$, define $B_n^\varepsilon = \{|X - X_n| > \varepsilon\}$, and let $B^\varepsilon = \{B_n^\varepsilon \text{ i.o.}\}$. Then,

$$\{X_n \to X\} = \bigcap_{j=1}^{\infty} (B^{\varepsilon_j})^c.$$

Solution. For the first statement, note that if $X_n \to X$, then for every ω, there exists an $N(\omega)$ such that for all $n \geq N$, $|X_n(\omega) - X(\omega)| < \varepsilon_n$. This means that ω belongs to only finitely many A_n, implying it is not in $\{A_n \text{ i.o.}\}$. Taking complements, we obtain the result.

For the second statement, observe that $X_n \to X$ means that for each fixed j, there exists an $N(\omega)$ such that $|X_n - X| \leq \varepsilon_j$ for all $n \geq N$. This means ω does not belong to $B_n^{\varepsilon_j}$ infinitely often, i.e., $\omega \notin B^{\varepsilon_j}$ for all j. Taking the intersection over all j gives the desired result.

Note 5.27. These conditions describe almost sure convergence in terms of event sequences. The first result states that convergence happens whenever large deviations A_n do not occur infinitely often. The second result refines this by showing that convergence is equivalent to the deviations $|X_n - X|$ eventually being smaller than any given ε_j. This formalises how almost sure convergence excludes infinitely large fluctuations and ensures stabilisation over time.

5.3.1 BC1

Lemma 5.28 (Borel-Cantelli 1)

If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$.

Note 5.29. BC1 intuitively tells us that if the total probability mass assigned to a sequence of events is finite, then they cannot keep occurring infinitely often.

Think of each A_n as a rare event. If their total probability sum $\sum \mathbb{P}(A_n)$ is finite, then the events do not have enough probability weight to persist indefinitely.

Proof. Define the random variable

$$N = \sum_{k=1}^{\infty} \mathbf{1}_{A_k},$$

which counts the number of times the events A_k occur. Our goal is to show that $\mathbb{P}(N = \infty) = 0$, meaning that only finitely many of the A_k can occur.

Taking expectations, we apply Fubini's theorem to interchange summation and expectation:

$$\mathbb{E}[N] = \mathbb{E}\sum_{k=1}^{\infty} \mathbf{1}_{A_k} = \sum_{k=1}^{\infty} \mathbb{E}[\mathbf{1}_{A_k}] = \sum_{k=1}^{\infty} \mathbb{P}(A_k).$$

Fubini's theorem justifies this interchange because indicator functions are non-negative and the series $\sum \mathbb{P}(A_k)$ is finite by assumption.

Now, we invoke a measure-theoretic result: if a non-negative random variable has a finite expectation, then it is finite almost surely. The intuition here is that if $N = \infty$ with positive probability, then its expectation would also be infinite. Since we have shown $\mathbb{E}[N]$ is finite, we conclude that $\mathbb{P}(N = \infty) = 0$, meaning $N < \infty$ almost surely.

Since $N = \infty$ corresponds precisely to the event that A_n occurs infinitely often, we obtain

$$\mathbb{P}(A_n \text{ i.o.}) = 0.$$

□

Exam Questions 5.30 (Exercise from lecture notes)

Suppose we are given a sequence of random variables X_n and a random variable X such that, for any $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) < \infty.$$

Then $X_n \to X$ almost surely.

Solution. By the Borel-Cantelli lemma, define the event

$$B^{\varepsilon} = \{|X_n - X| > \varepsilon \text{ infinitely often}\}.$$

Since the given assumption ensures that $\sum \mathbb{P}(B^{\varepsilon}) < \infty$, it follows that B^{ε} has probability zero for all $\varepsilon > 0$. Thus, its complement $\bigcap_{j=1}^{\infty}(B^{\varepsilon_j})^c$ has probability one for any sequence $\varepsilon_j \to 0$, ensuring $X_n \to X$ almost surely.

Exam Questions 5.31
Define a metric on random variables by

$$d(X,Y) = \mathbb{E}\left[\frac{|X-Y|}{1+|X-Y|}\right].$$

Then d metrises convergence in probability, i.e., $X_n \to X$ in probability if and only if $d(X_n, X) \to 0$.

Solution. First, note that $d(X,Y) \geq 0$ and $d(X,Y) = 0$ if and only if $X = Y$ almost surely. Also, $d(X,Y) = d(Y,X)$. To show the triangle inequality, consider

$$\frac{|x-y|}{1+|x-y|} = 1 - \frac{1}{1+|x-y|}.$$

Using the inequality

$$1 - \frac{1}{1+|x-y|} \leq \frac{|x-z|}{1+|x-z|+|z-y|} + \frac{|z-y|}{1+|z-y|},$$

it follows that $d(X,Y) \leq d(X,Z) + d(Z,Y)$, proving that d is a metric.

For convergence, suppose $X_n \to X$ in probability. Given $\varepsilon > 0$, there exists N such that for all $n \geq N$, we have $\mathbb{P}(|X_n - X| > \varepsilon) < \varepsilon$. Then,

$$d(X_n, X) = \mathbb{E}\left[\frac{|X_n - X|}{1+|X_n - X|}\right] \leq 2\varepsilon.$$

Thus, $d(X_n, X) \to 0$.

Conversely, if $d(X_n, X) \to 0$, then for any $\varepsilon \in (0,1]$, there exists N such that for $n \geq N$, $d(X_n, X) < \frac{\varepsilon^2}{1+\varepsilon}$. Applying Chebyshev's inequality,

$$\mathbb{P}(|X_n - X| > \varepsilon) \leq (1+\varepsilon)\varepsilon^{-1}\mathbb{E}\left[\frac{|X_n - X|}{1+|X_n - X|}\right] < \varepsilon(1+\varepsilon) \leq 2\varepsilon,$$

which implies $X_n \to X$ in probability.

Exam Questions 5.32
If $X_n \to X$ almost surely, then for any continuous function f, we have $f(X_n) \to f(X)$ almost surely.

Solution. Since $X_n \to X$ almost surely, there exists a probability-one set where $X_n \to X$ pointwise. Given that f is continuous, we obtain $f(X_n) \to f(X)$ on this set, proving almost sure convergence.

Theorem 5.33

$X_n \to X$ in probability if and only if, for every subsequence n_k, there exists a further subsequence n_{k_j} such that $X_{n_{k_j}} \to X$ almost surely as $j \to \infty$.

Proof. We prove each direction in turn.

- Proof of $\Rightarrow$. Suppose $X_n \to X$ in probability. Given a sequence $\varepsilon_k \downarrow 0$, there exists a subsequence n_k with
$$\mathbb{P}(|X_{n_k} - X| > \varepsilon_k) < 2^{-k}.$$
Summing over k, we obtain
$$\sum_{k=1}^{\infty} \mathbb{P}(|X_{n_k} - X| > \varepsilon_k) < \infty.$$
By Borel-Cantelli, with probability one, $|X_{n_k} - X| \leq \varepsilon_k$ for all but finitely many k, implying $X_{n_k} \to X$ almost surely. The same holds for any subsequence.

- Proof of $(\Leftarrow)$. Suppose that for every subsequence n_k, there eixsts a sub-subsequence n_{k_j} such that $X_{n_{k_j}} \to X$ as $j \to \infty$ almost surely. Then, since convergence almost surely implies convergence in probability, we know that every subsequence of X_n has a sub-subsequence converging in probability to X. Since convergence in probability is metrizable (i.e. it induces a metric), this last fact implies $X_n \to X$ in probability.

$\square$

Corollary 5.34

If $X_n \to X$ in probability, then $f(X_n) \to f(X)$ in probability for any continuous function f. Moreover, if f is bounded, then
$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)].$$

Corollary 5.35. If $X_n \to X$ in probability, and $|X_n|, |X| \leq Y$ with $\mathbb{E}[Y] < \infty$, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.

Proof. To show $f(X_n) \to f(X)$, by the theorem above it suffices to show that for every subsequence n_k, there exists a further subsequence n_{k_j} such that $f(X_{n_{k_j}}) \to f(X)$ almost surely. Since $X_n \to X$ in probability, Theorem 1 ensures that there exists a subsequence X_{n_k} converging to X almost surely. By continuity of f, we then obtain $f(X_{n_k}) \to f(X)$ almost surely, as desired.

For the second statement, we apply the bounded convergence theorem. Given that f is bounded, every subsequence of $\mathbb{E}[f(X_n)]$ has a further subsequence converging to $\mathbb{E}[f(X)]$, ensuring overall convergence. $\square$

Example 5.36

Suppose that $(X_n)_{n=1}^{\infty}$ is a sequence of random variables, and there exists a random variable X such that every subsequence of X_n has a further subsequence that converges to X almost surely. Must it be the case that X_n converges to X almost surely? **False:** Consider a sequence X_n that converges to X in probability but not almost surely.

5.3.2 Strong law of large numbers

Theorem 5.37 (Strong law of large numbers)
Let $X_1, X_2, \ldots$ be independent random variables with $\mathbb{E}[X_j] = \mu$ and $\sup_i \mathbb{E}[X_i^4] < \infty$. Define $S_n = \sum_{i=1}^n X_i$. Then,

$$\frac{S_n}{n} \to \mu \quad \text{almost surely as } n \to \infty.$$

Proof. Without loss of generality, set $\mu = 0$ by considering $X_n' = X_n - \mu$. We aim to show $S_n/n \to 0$ almost surely. To do so, we estimate moments of S_n and apply Borel-Cantelli 1. Second moments are insufficient, so we use fourth moments:

$$\mathbb{E}[S_n^4] = \sum_{1 \le i,j,k,l \le n} \mathbb{E}[X_i X_j X_k X_l].$$

By expanding and using independence,

$$\mathbb{E}[S_n^4] = n \sum_{i=1}^n \mathbb{E}[X_i^4] + \binom{n}{2} \sum_{1 \le i < j \le n} \mathbb{E}[X_i^2]\mathbb{E}[X_j^2].$$

Since $\mathbb{E}[X_i^4] \le C$, this gives

$$\mathbb{E}[S_n^4] \le nC + 6n^2 C \le 7n^2 C.$$

Using Chebyshev's inequality,

$$\mathbb{P}\left(\left| \frac{S_n}{n} \right| > \varepsilon \right) \le \frac{7n^2 C}{n^4 \varepsilon^4} = \frac{7C}{n^2 \varepsilon^4}.$$

Since the right-hand side is summable in n, Borel-Cantelli implies that for any $\varepsilon > 0$,

$$A_\varepsilon = \left\{ \omega \in \Omega : \left| \frac{S_n(\omega)}{n} \right| > \varepsilon \text{ i.o.} \right\}^c$$

has probability 1. Thus, $S_n/n \to 0$ almost surely. $\square$

5.3.3 BC2

Example 5.38

Can we state a converse to the Borel-Cantelli lemma? That is "If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 1$".

Solution. NO! Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega = [0, 1]$ is endowed with the Borel σ-algebra and Lebesgue measure. Define the sequence of events

$$A_n = \left[0, \frac{1}{n}\right].$$

Since the Lebesgue measure of an interval $[0, 1/n]$ is $1/n$, we have

$$\mathbb{P}(A_n) = \frac{1}{n}.$$

Summing over all n, we obtain

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

However, we analyse the probability of infinitely many of these events occurring. The event A_n i.o. corresponds to the set of points that belong to infinitely many of the intervals A_n. This set is precisely $\{0\}$, since any other point in $[0, 1]$ will eventually not be covered by the shrinking intervals $[0, 1/n]$ as $n \to \infty$.

Since a single point has Lebesgue measure zero, we conclude that

$$\mathbb{P}(A_n \text{ i.o.}) = \mathbb{P}(\{0\}) = 0.$$

Thus, even though the series $\sum \mathbb{P}(A_n)$ diverges, A_n does not occur infinitely often with probability one, refuting the naive converse statement of the Borel-Cantelli lemma.

Lemma 5.39 (BC2)

If the events $(A_n)_{n=1}^{\infty}$ are independent and $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then

$$\mathbb{P}(A_n \text{ i.o.}) = 1.$$

Note 5.40. The Second Borel-Cantelli Lemma states that if independent events occur frequently enough (i.e., their probabilities sum to infinity), then they must happen infinitely often with probability one. The key idea is that as $N \to \infty$, the probability of avoiding all future events shrinks exponentially to zero, ensuring that some events must keep occurring indefinitely.

Proof. To show that A_n happens infinitely often with probability one, we consider the tail events:

$$A_n \text{ i.o.} = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m.$$

This means that if we can show $\mathbb{P}(B_n) = 1$ for every n, where $B_n = \bigcup_{m=n}^{\infty} A_m$, then the result follows. Define the truncated sequence $B_{N,n} = \bigcup_{m=n}^{N} A_m$. Since $B_{N,n} \uparrow B_n$ as $N \to \infty$, by continuity of probability it suffices to show $\mathbb{P}(B_{N,n}^c) \to 0$ as $N \to \infty$. Using independence, we compute:

$$\mathbb{P}(B_{N,n}^c) = \mathbb{P}\left(\bigcap_{m=n}^{N} A_m^c\right) = \prod_{m=n}^{N}(1 - \mathbb{P}(A_m)).$$

By the inequality $1 - x \leq e^{-x}$ for $x \geq 0$, we obtain

$$\mathbb{P}(B_{N,n}^c) \leq \prod_{m=n}^{N} e^{-\mathbb{P}(A_m)} = \exp\left(-\sum_{m=n}^{N} \mathbb{P}(A_m)\right).$$

Since $\sum \mathbb{P}(A_m) = \infty$, the exponent diverges to $-\infty$, implying $\mathbb{P}(B_{N,n}^c) \to 0$ as $N \to \infty$, and hence $\mathbb{P}(B_n) = 1$. Thus, $\mathbb{P}(A_n \text{ i.o.}) = 1$, completing the proof.

□

Example 5.41

The first Borel-Cantelli lemma helps prove that sequences of random variables converge almost surely, the second lemma can be used to show that they do *not* converge almost surely under certain conditions.

Theorem 5.42. Let $X_1, X_2, \ldots$ be independent and identically distributed (iid) random variables with $\mathbb{E}[|X_j|] = \infty$. Then:

$$\mathbb{P}(|X_n| > n \text{ i.o.}) = 1.$$

Moreover, defining $S_n = \sum_{j=1}^{n} X_j$, and setting

$$C = \left\{\omega \in \Omega : \lim_{n\to\infty} \frac{S_n}{n} \text{ converges in } \mathbb{R}\right\},$$

we have $\mathbb{P}(C) = 0$.

Proof. For the first part, we use the integral identity:

$$\mathbb{E}[|X_1|] = \int_0^{\infty} \mathbb{P}(|X_1| > y)dy.$$

Since this expectation is infinite, it follows that $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > n) = \infty$. By the Second Borel-Cantelli Lemma, we conclude that $\mathbb{P}(|X_n| > n \text{ i.o.}) = 1$.

For the second part, we show that C is disjoint from the event $\{|X_n| > n \text{ i.o.}\}$, which has probability 1. Fixing $\omega \in C$, we have

$$0 = \lim_{n\to\infty} \frac{S_n(\omega)}{n} = \lim_{n\to\infty}\left(\frac{S_{n+1}(\omega)}{n+1} - \frac{X_{n+1}(\omega)}{n+1}\right).$$

By the algebra of limits, we also obtain

$$\frac{X_{n+1}(\omega)}{n+1} \to 0.$$

This contradicts the fact that $|X_n| > n$ infinitely often. Thus, if $\omega \in C$, then $\omega \notin \{|X_n| > n \text{ i.o.}\}$, implying $\mathbb{P}(C) = 0$. □

6 Tail σ-algebra and $0-1$ laws

6.1 Definitions

Definition 6.1. Given a sequence of random variables $(X_n)_{n=1}^{\infty}$, the **σ-algebra** is defined as

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \ldots).$$

The events in $\mathcal{T}$ are called **tail events**.

Note 6.2. $\mathcal{T}$ captures the idea of the "indefinite future". Roughly speaking $\mathcal{T}$ contains the events which are determined by $(X_n)_{n=1}^{\infty}$, but changing finitely many of the values does not affect if the event holds or not.

Example 6.3

Some examples.

1. For any sequence of events $A_1, A_2, \ldots$ with $A_n \in \sigma(X_n)$, the event $\{A_n \text{ i.o.}\}$ belongs to the tail $\mathcal{T}$.
   Intuition: A tail event is one that remains unchanged if we ignore a finite number of initial terms. Since infinitely many occurrences of A_n depend only on sufficiently large indices, it qualifies as a tail event.

2. If $S_n = \sum_{j=1}^{n} X_j$, then the event $\{S_n \text{ converges}\}$ is in $\mathcal{T}$.
   Intuition: The convergence of S_n depends on the behaviour of the sequence as $n \to \infty$ and not on any finite prefix, making it a tail event.

3. For S_n as above and $A \subset \mathcal{B}(\mathbb{R})$, the event $\{\lim_{n\to\infty} S_n \text{ exists and belongs to } A\}$ is **not** necessarily in $\mathcal{T}$.
   Intuition: The existence of a limit depends on the entire sequence rather than just its tail, so it may not satisfy the definition of a tail event.

4. For S_n as above and A as before, the event defined as $\{\lim_{n\to\infty} S_n/n^{1/2} \text{ exists and belongs to } A\}$ is in $\mathcal{T}$.
   Intuition: Normalised sums like $S_n/n^{1/2}$ often stabilise in the limit under the central limit theorem, making their limiting behaviour dependent only on the tail of the sequence.

Theorem 6.4 (Kolmogorov $0-1$ law)

Suppose $(X_n)_{n=1}^{\infty}$ are independent, if $A \in \mathcal{T}$ then $\mathbb{P}(A) \in \{0, 1\}$.

Proof. We aim to establish that $\mathcal{T}$ is independent of itself, i.e., for any event $A \in \mathcal{T}$,

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

We proceed by proving the independence of certain σ-algebras and using a π-system argument.

- First, we claim that for every $n \in \mathbb{N}_{>0}$, the σ-algebras $\sigma(X_1, \dots, X_n)$ and $\sigma(X_{n+1}, X_{n+2}, \dots)$ are independent. This follows from the fact that $\sigma(X_1, \dots, X_n)$ and $\bigcup_{j=1}^{\infty} \sigma(X_{n+1}, X_{n+2}, \dots)$ are both independent π-systems containing Ω. By the π-system theorem, the σ-algebras they generate are independent.
- Next, we extend this to show that $\sigma(X_1, X_2, \dots)$ and $\mathcal{T}$ are independent. Since $\mathcal{T} \subset \sigma(X_{n+1}, X_{n+2}, \dots)$ for all n, it follows that $\mathcal{T}$ is independent of $\sigma(X_1, X_2, \dots)$. Taking the increasing union over all k, we see that $\bigcup_{k=1}^{\infty} \sigma(X_1, \dots, X_k)$ and $\mathcal{T}$ are independent π-systems, and therefore their generated σ-algebras are also independent.

Since $\mathcal{T} \subset \sigma(X_1, X_2, \dots)$, the independence of $\sigma(X_1, X_2, \dots)$ and $\mathcal{T}$ implies that $\mathcal{T}$ is independent of itself. This completes the proof.

$\square$

6.2 Two series theorem

Theorem 6.5 (Kolmogorov's Maximal Inequality)
Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent random variables with finite second moments such that $\mathbb{E}[X_n] = 0$. Define the partial sums

$$S_n = \sum_{j=1}^{n} X_j.$$

Then, for any $x > 0$ and $n \in \mathbb{N}$,

$$\mathbb{P}\left(\max_{1 \le k \le n} |S_k| \ge x \right) \le \frac{\mathbb{E}[S_n^2]}{x^2}.$$

Note 6.6. The Kolmogorov Maximal Inequality is an improvement over what can be obtained using Chebyshev's inequality due to the inclusion of the maximum on the left-hand side.

Proof. Define the event

$$B = \left\{ \max_{1 \le k \le n} |S_k| \ge x \right\}.$$

To bound $\mathbb{P}(B)$, we decompose it into disjoint events. Let

$$A_k = \{|S_k| \ge x \text{ but } |S_j| < x \text{ for all } j < k\}$$

so that A_k represents the first time S_k exceeds x in absolute value. The sets A_k are disjoint, and we can write

$$B = \bigcup_{k=1}^{n} A_k.$$

Using expectation decomposition,

$$\mathbb{E}[S_n^2] \ge \mathbb{E}[S_n^2 \mathbf{1}_B] = \sum_{k=1}^{n} \mathbb{E}[S_n^2 \mathbf{1}_{A_k}].$$

Since A_k provides information about S_k rather than S_n, we write $S_n = S_k + (S_n - S_k)$ and expand:

$$\mathbb{E}[S_n^2 \mathbf{1}_{A_k}] = \mathbb{E}[S_k^2 \mathbf{1}_{A_k}] + 2\mathbb{E}[S_k(S_n - S_k)\mathbf{1}_{A_k}] + \mathbb{E}[(S_n - S_k)^2 \mathbf{1}_{A_k}].$$

By independence, $S_n - S_k$ is independent of S_k, and since $\mathbf{1}_{A_k}$ is $\sigma(X_1, \dots, X_k)$-measurable,

$$\mathbb{E}[S_k(S_n - S_k)\mathbf{1}_{A_k}] = \mathbb{E}[S_k \mathbf{1}_{A_k}]\mathbb{E}[S_n - S_k] = 0.$$

Thus,

$$\mathbb{E}[S_n^2 \mathbf{1}_{A_k}] = \mathbb{E}[S_k^2 \mathbf{1}_{A_k}].$$

Substituting into the original inequality,

$$\mathbb{E}[S_n^2] \geq \sum_{k=1}^{n} \mathbb{E}[S_k^2 \mathbf{1}_{A_k}].$$

Since $|S_k| \geq x$ on A_k, it follows that $S_k^2 \geq x^2$, yielding

$$\mathbb{E}[S_k^2 \mathbf{1}_{A_k}] \geq x^2 \mathbb{P}(A_k).$$

Summing over k,

$$\mathbb{E}[S_n^2] \geq x^2 \sum_{k=1}^{n} \mathbb{P}(A_k) = x^2 \mathbb{P}(B).$$

Rearranging gives the desired result.

$$\mathbb{P}(B) \leq \frac{\mathbb{E}[S_n^2]}{x^2}.$$

$\square$

Lemma 6.7. Let $X_1 \geq X_2 \geq \cdots \geq 0$ be a non-increasing sequence of non-negative random variables. If $X_n \to 0$ in probability, then $X_n \to 0$ almost surely.

Proof. Since $X_n \to 0$ in probability, we have for every $\varepsilon > 0$,

$$\mathbb{P}(X_n \geq \varepsilon) \to 0 \quad \text{as } n \to \infty.$$

Define the events $A_n = \{X_n \geq \varepsilon\}$. Since X_n is non-increasing, the sequence of events (A_n) is also non-increasing, meaning

$$A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$$

Thus, $\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n) = 0$. This implies that $X_n \to 0$ almost surely, since the probability of the event that X_n remains above ε infinitely often is zero. As ε was arbitrary, we conclude that $X_n \to 0$ almost surely. $\square$

Theorem 6.8 (Two Series theorem)

Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent random variables with finite second moments such that

$$\sum_{n=1}^{\infty} \mathbb{E}[X_n] < \infty, \quad \text{and} \quad \sum_{n=1}^{\infty} \operatorname{Var}(X_n) < \infty.$$

Then, with probability 1, the series $\sum_{n=1}^{\infty} X_n(\omega)$ converges.

Proof. Without loss of generality, assume $\mathbb{E}[X_n] = 0$. Define the partial sums $S_n = \sum_{j=1}^{n} X_j$ and consider the event

$$A = \{\omega \in \Omega : S_n(\omega) \text{ is Cauchy}\}.$$

To prove $\mathbb{P}(A) = 1$, define

$$W_M(\omega) = \sup_{n,m \geq M} |S_n(\omega) - S_m(\omega)|.$$

By Lemma 1, it suffices to show $W_M \to 0$ in probability. Fix $\varepsilon > 0$, then

$$\mathbb{P}(W_M > \varepsilon) \leq \mathbb{P}\left(\sup_{m \geq M} |S_m - S_M| \geq \varepsilon/2\right).$$

Since S_n is a sum of independent random variables, Kolmogorov's Maximal Inequality gives

$$\mathbb{P}\left(\sup_{m \geq M} |S_m - S_M| \geq \varepsilon/2\right) \leq \varepsilon^{-2} \sum_{j=M+1}^{\infty} \mathbb{E}[X_j^2].$$

Since $\sum \mathrm{Var}(X_n) < \infty$, the right-hand side vanishes as $M \to \infty$, implying $W_M \to 0$ in probability. Thus, S_n is Cauchy almost surely, proving the theorem. □

6.3 Kolmogorov's SLLN

Lemma 6.9 (Toeplitz lemma)**.** Let $(a_n)_{n=1}^{\infty}$ be a non-negative sequence such that $b_n = \sum_{j=1}^{n} a_j \to \infty$ as $n \to \infty$. Suppose $x_n \to x$ as $n \to \infty$. Then,

$$\frac{1}{b_n} \sum_{j=1}^{n} a_j x_j \to x \quad \text{as } n \to \infty.$$

Proof. Without loss of generality, assume $x = 0$ by setting $y_n = x_n - x$. For $\varepsilon > 0$, choose N_1 such that $n \geq N_1$ implies $|x_n| < \varepsilon/2$. Since $b_n \to \infty$, there exists $N_2 > N_1$ such that $n \geq N_2$ ensures

$$\frac{1}{b_n} \sum_{j=1}^{N_1} a_j |x_j| < \varepsilon/2.$$

For $n \geq N_2$, we estimate

$$\left| \frac{1}{b_n} \sum_{j=1}^{n} a_j x_j \right| \leq \frac{1}{b_n} \sum_{j=1}^{N_1} a_j |x_j| + \frac{1}{b_n} \sum_{j=N_1+1}^{n} a_j |x_j|.$$

Using $|x_j| < \varepsilon/2$ for $j \geq N_1$,

$$< \varepsilon/2 + (\varepsilon/2) \cdot \frac{b_n - b_{N_1}}{b_n} \leq \varepsilon.$$

Thus, the result follows. □

Lemma 6.10 (Kronecker's lemma)**.** Suppose $b_n \geq 0$ and $b_n \to \infty$ as $n \to \infty$. If $\sum_{n=1}^{\infty} x_n$ converges, then

$$\frac{1}{b_n} \sum_{j=1}^{n} b_j x_j \to 0 \quad \text{as } n \to \infty.$$

Proof. Define $S_n = \sum_{j=1}^n x_j$ and let $b_0 = S_0 = 0$. Applying summation by parts,

$$\sum_{j=1}^n b_j x_j = b_n S_n - b_0 S_0 + \sum_{j=1}^n S_{j-1}(b_j - b_{j-1}).$$

Setting $a_j = b_j - b_{j-1} \geq 0$, we obtain

$$\frac{1}{b_n} \sum_{j=1}^n b_j x_j = S_n - \frac{1}{b_n} \sum_{j=1}^n S_{j-1} a_j.$$

Since S_n converges, $\lim_{n\to\infty} S_n = S$ exists. By Toeplitz's lemma,

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{j=1}^n S_{j-1} a_j = \lim_{n\to\infty} S_{n-1} = S.$$

Thus,

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{j=1}^n b_j x_j = S - S = 0.$$

□

Theorem 6.11 (Second moment SLLN)
Let $(X_n)_{n=1}^\infty$ be a sequence of independent random variables with finite second moments. Suppose that $b_n \to \infty$ as $n \to \infty$ and that

$$\sum_{n=1}^\infty \frac{\operatorname{Var}(X_n)}{b_n^2} < \infty.$$

Then, defining $S_n = \sum_{j=1}^n X_j$, we have

$$\frac{1}{b_n}(S_n - \mathbb{E}[S_n]) \to 0 \quad \text{with probability 1 as } n \to \infty.$$

Proof. We begin by expressing the normalised sum:

$$\frac{1}{b_n}(S_n - \mathbb{E}[S_n]) = \frac{1}{b_n} \sum_{j=1}^n b_j \frac{X_j - \mathbb{E}[X_j]}{b_j}.$$

By the two-series theorem, the series

$$\sum_{j=1}^n \frac{X_j - \mathbb{E}[X_j]}{b_j}$$

converges almost surely, given that

$$\sum_{n=1}^\infty \frac{\operatorname{Var}(X_n)}{b_n^2} < \infty.$$

Since the conditions of Kronecker's Lemma hold, the desired result follows. □

Theorem 6.12 (Kolmogorov's SLLN)
Let $(X_n)_{n=1}^{\infty}$ be a sequence of independent and identically distributed (iid) random variables with finite first moment. Define $\mathbb{E}[X_n] = \mu$ and the partial sum

$$S_n = \sum_{j=1}^{n} X_j.$$

Then,

$$\frac{S_n}{n} \to \mu \quad \text{almost surely as } n \to \infty.$$

Proof. Define the centred sequence $Y_n = X_n - \mu$. Then, $S_n - n\mu = \sum_{j=1}^{n} Y_j$, and we aim to show that

$$\frac{S_n}{n} - \mu = \frac{1}{n}\sum_{j=1}^{n} Y_j \to 0 \quad \text{almost surely.}$$

Applying Kolmogorov's inequality and the Borel-Cantelli lemma, we establish that

$$\sum_{n=1}^{\infty} \frac{\mathrm{Var}(X_n)}{n^2} < \infty.$$

By the Second Moment SLLN, which follows from Kronecker's Lemma, we obtain

$$\frac{1}{n}\sum_{j=1}^{n} Y_j \to 0 \quad \text{almost surely.}$$

□

7 Convergence in distribution

7.1 Definitions

Lemma 7.1 (Stirling's formula). $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ as $n \to \infty$.

Theorem 7.2 (De Moivre-Laplace theorem)
Let $X_1, X_2, \ldots$ be iid random variables with $\mathbb{P}(X_1 = \pm 1) = 1/2$. Define the sum

$$S_n = \sum_{j=1}^{n} X_j.$$

Then, for any $a < b$,

$$\mathbb{P}\left(a \le \frac{S_n}{\sqrt{n}} \le b\right) \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \quad \text{as } n \to \infty.$$

Sketch of proof. The probability mass function is given by

$$\mathbb{P}(S_{2n} = 2k) = \binom{2n}{n+k} 2^{-2n}.$$

Applying Stirling's approximation,

$$\binom{2n}{n+k} \approx \frac{(2n)^{2n}}{(n+k)^{n+k}(n-k)^{n-k}} \cdot \frac{\sqrt{2\pi}2n}{\sqrt{(2\pi)(n+k)(n-k)}},$$

simplifies to

$$\mathbb{P}(S_{2n} = 2k) \approx (\pi n)^{-1/2} e^{-x^2/2}, \quad \text{where } 2k = x\sqrt{2n}.$$

A change of variable $x \leftrightarrow k$ introduces a factor $(n/2)^{1/2}$, yielding the limiting density

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

□

Example 7.3

A natural first attempt to define convergence in distribution might be to say that a sequence of random variables X_n converges to X in distribution if their cumulative distribution functions (CDFs) satisfy

$$F_{X_n}(x) \to F_X(x) \quad \text{for all } x.$$

However, this naive definition fails in many cases.

Let X be a random variable and define

$$X_n = X + \frac{1}{n}.$$

Clearly, $X_n \to X$ almost surely. However, examining the CDFs,

$$F_{X_n}(x) = F_X(x - 1/n) \to F_X(x-), \quad \text{as } n \to \infty,$$

where $F_X(x-)$ denotes the left-hand limit of F_X at x. Thus, the convergence of CDFs does not hold at points of discontinuity.

Definition 7.4. Given a function $f : \mathbb{R} \to \mathbb{R}$, we say that $x \in \mathbb{R}$ is a **point of continuity** of f if f is continuous at x. For a distribution function F, this is equivalent to

$$\lim_{y \uparrow x} F(y) = F(x).$$

We denote $\lim_{y\uparrow x} F(y) := F(x-)$, so the above becomes $F(x) = F(x-)$.

Definition 7.5. Given distribution functions $F_1, F_2, \ldots, F_n, \ldots$ and F, we say that $F_n \Rightarrow F$ (converge weakly) if, for all points of continuity x of F,

$$\lim_{n\to\infty} F_n(x) = F(x).$$

Given random variables $X_1, X_2, \ldots$ and X, which may be defined on different probability spaces, we say that $X_n \Rightarrow X$ (converge weakly or in distribution) if $F_{X_n} \Rightarrow F_X$.

Remark 7.6. The actual definition coincides with the naive definition when F is continuous. If F is not continuous, it can have at most countably many points of discontinuity. This means that if $F_n \Rightarrow F$, then we can still determine F from $\lim_{n\to\infty} F_n(x)$ on the set of x where this limit exists.

Exam Questions 7.7

Let $X \sim \text{Uniform}([-1,1])$ and define $X_n = (-1)^n X$. We show that X_n does not converge in probability to X, but $X_n \Rightarrow X$ in distribution.

Solution. For convergence in probability, we check

$$P(|X_n - X| \geq \varepsilon).$$

Since $|X_n - X| = 2|X|$ for all $X \neq 0$, and $P(X \neq 0) = 1$, we have

$$P(|X_n - X| \geq \varepsilon) = P(2|X| \geq \varepsilon) = P(|X| \geq \varepsilon/2).$$

Since $X \sim \text{Uniform}([-1,1])$, this probability does not tend to 0 as $n \to \infty$, proving that X_n does not converge in probability.

For convergence in distribution, we compare cumulative distribution functions. The CDF of X is

$$F_X(x) = \frac{x+1}{2}, \quad x \in [-1,1].$$

Since X_n has the same distribution as X, we have

$$F_{X_n}(x) = F_X(x).$$

Thus, for all continuity points of F_X,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x),$$

7.2 Quantile function

Exam Questions 7.8

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space where $\Omega = (0,1)$, $\mathcal{F} = \mathcal{B}(0,1)$, and $\mathbb{P}$ is the Lebesgue measure. Given a distribution function F, define

$$X(\omega) = \sup\{y \in \mathbb{R} \mid F(y) < \omega\} =: F^{-1}(\omega).$$

We verify that X has CDF F.

Solution. By definition, for any x,

$$P(X \leq x) = P(F^{-1}(\omega) \leq x).$$

Since $F^{-1}(\omega)$ is the generalized inverse of F, this simplifies to

$$P(X \leq x) = P(\omega \leq F(x)) = F(x),$$

since $\omega \sim \text{Uniform}(0,1)$. Thus, X has the desired distribution.

Definition 7.9. For a distribution function F, the **quantile function** is defined as

$$F^{-1}(\omega) = \sup\{y \in \mathbb{R} : F(y) < \omega\}, \quad \omega \in (0,1).$$

This function is non-decreasing and left-continuous.

Lemma 7.10. Let F^{-1} be the corresponding quantile function. Then:

- F^{-1} is non-decreasing and left-continuous.
- F^{-1} is continuous at ω if and only if

  $$\omega \in \Omega_c = \{\omega \in (0,1) : a(\omega) = b(\omega)\},$$

  where $a(\omega) = \sup\{y \in \mathbb{R} : F(y) < \omega\}$ and $b(\omega) = \inf\{y \in \mathbb{R} : F(y) > \omega\}$.
- The set $(0,1) \setminus \Omega_c$ is at most countable.

Proof. To see that F^{-1} is non-decreasing, we observe that for $\omega < \omega'$, we have

$$F^{-1}(\omega) = \sup\{y \in \mathbb{R} : F(y) < \omega\} \leq \sup\{y \in \mathbb{R} : F(y) < \omega'\} = F^{-1}(\omega').$$

To establish that F^{-1} is left-continuous, we assume that ω_n is an increasing sequence converging to ω, i.e., $\omega_n \uparrow \omega$. Then, we compute

$$\lim_{n\to\infty} F^{-1}(\omega_n) = \sup_n F^{-1}(\omega_n) = \sup_n \{y \in \mathbb{R} : F(y) < \omega_n\}.$$

By properties of the supremum, this can be rewritten as

$$\sup\left(\bigcup_{n=1}^{\infty} \{y \in \mathbb{R} : F(y) \leq \omega_n\}\right).$$

Since $\omega_n \uparrow \omega$, we have

$$\sup\{y \in \mathbb{R} : F(y) \leq \omega\} = F^{-1}(\omega).$$

Thus, $\lim_{n\to\infty} F^{-1}(\omega_n) = F^{-1}(\omega)$, proving left-continuity. To analyse the continuity of F^{-1}, we note that continuity occurs when

$$F^{-1}(\omega+) = F^{-1}(\omega) = F^{-1}(\omega-).$$

It suffices to show that $F^{-1}(\omega+) = F^{-1}(\omega_n) = b(\omega)$. For any decreasing sequence $\omega_n \downarrow \omega$, we compute

$$\lim_{n\to\infty} F^{-1}(\omega_n) = \inf_n F^{-1}(\omega_n) = \inf_n \sup\{y \in \mathbb{R} : F(y) < \omega_n\}.$$

By properties of the infimum,

$$\sup\left(\bigcap_{n=1}^{\infty}\{y \in \mathbb{R} : F(y) < \omega_n\}\right).$$

Since $\omega_n \downarrow \omega$, we conclude

$$\sup\{y \in \mathbb{R} : F(y) \leq \omega\} = \inf\{y \in \mathbb{R} : F(y) \leq \omega\}^c = b(\omega).$$

□

Lemma 7.11 (Weak convergence implies a.s. convergence)
If $F_n \Rightarrow F$, then there exist random variables X_n, X on the same probability space such that $F_{X_n} = F_n$, $F_X = F$, and $X_n \to X$ almost surely.

Proof. Let the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ be given by $\Omega = (0,1)$, $\mathcal{F} = \mathcal{B}(0,1)$, and $\mathbb{P}$ as the Lebesgue measure. We aim to show that we can take X_n and X to be the quantile functions of F_n and F, respectively.

Let Ω_c denote the set of continuity points of the quantile function F^{-1}. Since $\Omega \setminus \Omega_c$ is countable, we have $\omega \in \Omega_c$, which proves the desired result. Fix $\omega \in \Omega_c$.

We first establish the lower bound:

$$\liminf_{n\to\infty} F_n^{-1}(\omega) \geq F^{-1}(\omega).$$

To prove the above, note that for any $y < F^{-1}(\omega)$, we have $F(y) < \omega$ since $\omega \in \Omega_c$ (that is, F is not flat at ω).

Suppose that such a $y < F^{-1}(\omega)$ is also a continuity point of F. Then, as $F_n(y) \to F(y)$, there exists some N_y such that for $n \geq N_y$, we have $F_n(y) < \omega$. This implies that for such y and for all $n \geq N_y$, it holds that

$$F_n^{-1}(\omega) \geq y.$$

This follows from the fact that $F_n^{-1}(\omega)$ is defined as a supremum over a set to which y belongs.

Now, assume for contradiction that the equation does not hold. Then, there exists $\delta > 0$ and a subsequence n_k such that

$$F_{n_k}^{-1}(\omega) < F^{-1}(\omega) - \delta, \quad \text{for all } k.$$

However, since F has a continuity point at some y satisfying $y < F^{-1}(\omega)$ and $y > F^{-1}(\omega) - \delta$, we can find N_y such that for all $n \geq N_y$,

$$F_n^{-1}(\omega) \geq y \geq F^{-1}(\omega) - \delta,$$

which yields a contradiction. By a symmetric argument, where we reverse the inequalities, we obtain:

$$\limsup_{n\to\infty} F_n^{-1}(\omega) \leq F^{-1}(\omega).$$

□

Theorem 7.12

The sequence X_n converges in distribution to X, denoted $X_n \Rightarrow X$, if and only if, for every $g \in C_b(\mathbb{R})$,

$$\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)] \quad \text{as } n \to \infty.$$

Note 7.13. This is an equivalent characterisation of weak convergence, which extends naturally to random variables taking values in spaces beyond $\mathbb{R}$.

Remark 7.14. In what follows, we denote by $C_b(\mathbb{R})$ the space of continuous and bounded functions $g : \mathbb{R} \to \mathbb{R}$, sometimes written simply as C_b.

Proof. We prove each direction in turn.

- Proof of ($\Rightarrow$).
  For any $g \in C_b$, we then have

  $$\lim_{n\to\infty} \mathbb{E}[g(X_n)] = \lim_{n\to\infty} \mathbb{E}[g(Y_n)] = \mathbb{E}[g(Y)] = \mathbb{E}[g(X)].$$

  The first and last equalities follow from the fact that expectation depends only on the distribution function, and the middle step follows from the Bounded Convergence Theorem.

- Proof of ($\Leftarrow$).
  Define, for any $x \in \mathbb{R}$ and $\varepsilon > 0$, the function:

  $$g_{x,\varepsilon}(y) = \begin{cases} 1, & y \leq x, \\ 1 - \frac{y-x}{\varepsilon}, & x < y < x + \varepsilon, \\ 0, & y \geq x + \varepsilon. \end{cases}$$

  Note that $g_{x,\varepsilon}$ is continuous and bounded.
  For any continuity point x of F, we show that $F_{X_n}(x) \to F_X(x)$. Given $\varepsilon > 0$,

  $$\limsup_{n\to\infty} \mathbb{P}(X_n \leq x) = \limsup_{n\to\infty} \mathbb{E}[1_{\{X_n \leq x\}}] \leq \limsup_{n\to\infty} \mathbb{E}[g_{x,\varepsilon}(X_n)].$$

  Using expectation properties,

  $$\mathbb{E}[g_{x,\varepsilon}(X)] = \mathbb{E}[1_{\{X \leq x+\varepsilon\}}] = \mathbb{P}(X \leq x + \varepsilon).$$

By right continuity of F, we obtain

$$\mathbb{P}(X \leq x + \varepsilon) \to \mathbb{P}(X \leq x) \text{ as } \varepsilon \downarrow 0.$$

Thus, we conclude

$$\limsup_{n\to\infty} \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x).$$

A similar argument establishes the reverse inequality for $\liminf$, yielding the desired result:

$$X_n \Rightarrow X.$$

□

Lemma 7.15 (Continuous Mapping Theorem)
Suppose that $X_n \Rightarrow X$, then for any $g \in C(\mathbb{R})$, we have $g(X_n) \Rightarrow g(X)$ as $n \to \infty$.

Note 7.16. This lemma shows that weak convergence is preserved by continuity.

Remark 7.17. We write $C(\mathbb{R}) \supset C_b(\mathbb{R})$ for the space of continuous functions $g : \mathbb{R} \to \mathbb{R}$.

Proof. Let $h \in C_b(\mathbb{R})$, then $h \circ g \in C_b(\mathbb{R})$, and so by Theorem 1 applied to the function $h \circ g$, we obtain

$$\mathbb{E}[h(g(X_n))] \to \mathbb{E}[h(g(X))].$$

It follows, by using the theorem above in the other direction, that $g(X_n) \Rightarrow g(X)$. □

Theorem 7.18 (Helly's Selection Theorem)
Let $(F_n)_{n=1}^{\infty}$ be a sequence of distribution functions. Then there exists a subsequence $(F_{n_k})_{k=1}^{\infty}$ and a right-continuous, increasing function F such that, for every continuity point x, $F_{n_k}(x) \to F(x)$ as $k \to \infty$.

Remark 7.19. The Helly selection theorem does not guarantee that the limiting function F is a distribution function.

Proof. Since each F_n is bounded by 1, we apply a diagonalisation argument, as used in the proof of the Arzelà-Ascoli theorem, to extract a subsequence F_{n_k} such that, for every $q \in \mathbb{Q}$,

$$\lim_{k\to\infty} F_{n_k}(q) =: G(q)$$

exists. The function G is increasing on the rationals. We extend it to the reals by defining

$$F(x) = \inf\{G(q) : q \in \mathbb{Q}, q > x\}.$$

Since G is increasing, this definition ensures that F is also increasing and right-continuous. To confirm right continuity, consider a sequence $x_j \downarrow x$. Then,

$$\lim_{j\to\infty} F(x_j) = \inf_j F(x_j).$$

By the definition of F,

$$\begin{aligned}\inf_j F(x_j) &= \inf_j \inf\{G(q) : q \in \mathbb{Q}, q > x_j\} \\ &= \inf \left\{ \bigcup_j G(q) : q \in \mathbb{Q}, q > x_j \right\} \\ &= \inf\{G(q) : q \in \mathbb{Q}, q > x\} = F(x).\end{aligned}$$

Thus, F is right-continuous. Now, let x be a continuity point of F, and let $\varepsilon > 0$. There exist rationals q_1, q_2, r such that $q_1 < q_2 < x < r$, satisfying

$$F(x) - \varepsilon < F(q_1) \leq F(q_2) \leq F(x) \leq F(r) < F(x) + \varepsilon.$$

Since F is increasing,

$$F(q_2) \leq F(x) \leq F(r) < F(x) + \varepsilon.$$

Since $F_{n_k}(q_2) \to G(q_2)$ and $G(q_2) \geq F(q_1)$, for sufficiently large k,

$$F(x) - \varepsilon < F(q_1) < F_{n_k}(q_2) < F_{n_k}(x) < F_{n_k}(r) < F(x) + \varepsilon.$$

Thus, $F_{n_k}(x) \to F(x)$ □

7.3 Tightness

Example 7.20

Let X be a random variable taking values ± 1 with expectation $\mathbb{E}[X] = 0$. Define the sequence $X_n = nX$. By Helly's selection theorem, there exists a subsequence X_{n_k} and a function F such that the cumulative distribution functions satisfy

$$F_{X_{n_k}}(x) \to F(x) \quad \text{for all continuity points } x \text{ of } F.$$

We claim that F cannot be a distribution function of a probability measure. We argue by contradiction by assuming that F is a distribution function. Let $Y \sim F$. Then, for every compactly supported function $g \in C_b$, we compute the expected value:

$$\mathbb{E}[g(Y)] = \lim_{k\to\infty} \mathbb{E}[g(X_{n_k})] = \lim_{k\to\infty} \left(\frac{1}{2} g(n_k) + \frac{1}{2} g(-n_k) \right) = 0.$$

However, we can construct a sequence of compactly supported functions g_n satisfying $0 \leq g_n \leq 1$ and $g_n \to 1$ pointwise. Applying the bounded convergence theorem, we obtain

$$\lim_{n\to\infty} \mathbb{E}[g_n(Y)] = 1.$$

The issue arises because Helly's selection theorem permits mass to "leak" to $\pm\infty$, preventing F from being a valid probability distribution function. To correct this issue, an additional condition must be imposed on the selection theorem to prevent leakage to infinity.

Definition 7.21. A family $(\mathbb{P}_i : i \in I)$ of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is said to be **tight** if, for every $\varepsilon > 0$, there exists a compact set K_ε such that

$$\sup_{i\in I} \mathbb{P}_i(\mathbb{R} \setminus K_\varepsilon) < \varepsilon.$$

We say a family of distribution functions $(F_i : i \in I)$ is **tight** if the corresponding family of probability measures is tight.

Note 7.22. This definition is equivalent to $\mathbb{P}(X_n \notin K_\varepsilon) < \varepsilon$.

Exam Questions 7.23

A family of distribution functions $(F_i : i \in I)$ is tight if and only if, for every $\varepsilon > 0$, there exists $M_\varepsilon > 0$ such that

$$\sup_{i \in I} (1 - F_i(M_\varepsilon) + F_i(-M_\varepsilon)) < \varepsilon.$$

Solution. We prove each direction in turn.

- Proof of ($\Rightarrow$). By definition, for each $\varepsilon > 0$, there exists a compact set K_ε such that
$$\sup_{i \in I} \mathbb{P}_i(K_\varepsilon^c) < \varepsilon.$$
Since probability measures on $\mathbb{R}$ are tight, we can take $K_\varepsilon = [-M_\varepsilon, M_\varepsilon]$ for some sufficiently large $M_\varepsilon > 0$. Then,
$$\mathbb{P}_i(K_\varepsilon^c) = 1 - F_i(M_\varepsilon) + F_i(-M_\varepsilon),$$
which implies the desired inequality.
- Proof of ($\Leftarrow$). Choosing $K_\varepsilon = [-M_\varepsilon, M_\varepsilon]$, we obtain
$$\mathbb{P}_i(K_\varepsilon^c) = 1 - F_i(M_\varepsilon) + F_i(-M_\varepsilon) < \varepsilon.$$
Thus, the family of probability measures is tight, proving the claim.

Theorem 7.24 (Helly's selection theorem (tight version))

Let $(F_n)_{n=1}^\infty$ be a sequence of distribution functions that is tight. Then, there exists a subsequence $(F_{n_k})_{k=1}^\infty$ and a distribution function F such that $F_{n_k} \Rightarrow F$ as $k \to \infty$.

Proof. Construct the subsequence F_{n_k} and the function F as in the proof of the Helly selection theorem. It remains to show that

$$\lim_{t \to -\infty} F(t) = 0 \quad \text{and} \quad \lim_{t \to \infty} F(t) = 1.$$

Let $\varepsilon > 0$ and choose M_ε such that

$$\sup_n (1 - F_n(M_\varepsilon) + F_n(-M_\varepsilon)) < \varepsilon.$$

Now, consider $x < -M_\varepsilon$, which is a continuity point of F. Then,

$$F(x) = \lim_{k \to \infty} F_{n_k}(x) \leq \limsup_{k \to \infty} F_{n_k}(M_\varepsilon) < \varepsilon.$$

This shows that $\lim_{t\to-\infty} F(t) = 0$, noting that F is increasing and non-negative.

Similarly, we choose $y > M_\varepsilon$, which is a continuity point of F. Since $F_n(y)$ is increasing in n, we obtain

$$\inf_n F_n(y) > 1 - \varepsilon.$$

Thus,

$$F(y) = \lim_{k\to\infty} F_{n_k}(y) \geq \liminf_{k\to\infty} F_{n_k}(M_\varepsilon) > 1 - \varepsilon.$$

□

Theorem 7.25 (Lévy metric)
Given two distribution functions F and G, the Lévy metric is defined as

$$d(F, G) = \inf\{\varepsilon > 0 : F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon \text{ for all } x \in \mathbb{R}\}.$$

Then d is indeed a metric, and it satisfies $d(F_n, F) \to 0$ if and only if $F_n \Rightarrow F$.

Proof. To verify that d is a metric, we check the following properties:

- By definition, $d(F, G) \geq 0$ for all distribution functions F and G.
- If $d(F, G) = 0$, then for every $\varepsilon > 0$, we have
  $$F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon, \quad \forall x \in \mathbb{R}.$$
  Taking $\varepsilon \to 0$, it follows that $F(x) = G(x)$ for all x, so $F = G$.
- The definition of $d(F, G)$ is symmetric in F and G, so $d(F, G) = d(G, F)$.
- Given three distribution functions F, G, H, we need to show that
  $$d(F, H) \leq d(F, G) + d(G, H).$$
  By definition, for any $\varepsilon_1 > d(F, G)$ and $\varepsilon_2 > d(G, H)$, we have
  $$F(x-\varepsilon_1) - \varepsilon_1 \leq G(x) \leq F(x+\varepsilon_1) + \varepsilon_1,$$
  and similarly for G and H. Combining these inequalities gives the required bound.

To show that $d(F_n, F) \to 0$ if and only if $F_n \Rightarrow F$, we note that weak convergence $F_n \Rightarrow F$ means that $F_n(x) \to F(x)$ at all continuity points of F. The definition of the Lévy metric ensures that for sufficiently small ε, the deviations between F_n and F are controlled uniformly, establishing equivalence. □

Corollary 7.26
Let (F_n) be a sequence of tight distribution functions, and assume that there exists a distribution function F such that every weakly convergent subsequence F_{n_k} of F_n satisfies $F_{n_k} \Rightarrow F$. Then it follows that $F_n \Rightarrow F$.

Proof. Suppose, for contradiction, that $F_n \not\Rightarrow F$. Then, there exists a subsequence (F_{n_k}) such that $F_{n_k} \not\Rightarrow F$. Since (F_n) is tight, Helly's selection theorem ensures that (F_{n_k}) has a weakly convergent subsequence $(F_{n_{k_m}})$ that converges to some function G.

By assumption, every weakly convergent subsequence of (F_n) must converge to F, so we must have $G = F$, implying $F_{n_k} \Rightarrow F$, a contradiction. □

8 Limit Theorems

8.1 Characteristic functions

Note 8.1. Characteristic functions are the Fourier transform of probability measures.

Definition 8.2. The **characteristic function** of a random variable X is a function

$$\begin{aligned}\phi_X &: \mathbb{R} \to \mathbb{C} \\ \theta &\mapsto \mathbb{E}\left(e^{i\theta X}\right) \\ &= \int_{\mathbb{R}} e^{i\theta x} dF_X(x) = \int_{\mathbb{R}} e^{i\theta x} f_X(x)\, dx.\end{aligned}$$

Remark 8.3. Since ϕ_X only depends on the distribution function of X, we can also talk about the characteristic function of a probability measure μ on $\mathbb{R}$, which we write ϕ_μ.

Lemma 8.4

Elementary properties of the characteristic function.

1. $\phi_X(0) = 1$.
2. $\phi_X(-\theta) = \overline{\phi_X(\theta)}$.
3. $|\phi_X(\theta)| \leq 1$.
4. ϕ_X is uniformly continuous.
5. If $\mathbb{E}(|X|^n) < \infty$ for some $n \in \mathbb{N}$ then ϕ_X is n-times differentiable with
$$\phi_X^{(n)}(\theta) = \mathbb{E}\left((iX)^n e^{i\theta X}\right) \text{ uniformly continuous.}$$
6. If X and Y are independent and $\alpha \in \mathbb{R}$ then
$$\phi_{aX+Y}(\theta) = \phi_X(\alpha\theta)\phi_Y(\theta).$$

Proof. We prove each statement in turn.

1. $\phi_X(0) = \mathbb{E}[1] = 1$.
2. We use the definition of the characteristic function:
$$\phi_X(-\theta) = \mathbb{E}\left[\exp(-i\theta X)\right] = \mathbb{E}\left[\overline{\exp(i\theta X)}\right] = \overline{\phi_X(\theta)}.$$
3. We use the modulus property of expectations:
$$|\phi_X(\theta)| = |\mathbb{E}\left[\exp(i\theta X)\right]| \leq \mathbb{E}\left[|\exp(i\theta X)|\right] = 1.$$
4. We consider the difference:
$$|\phi_X(\theta+h) - \phi_X(\theta)| = \left|\mathbb{E}\left[\exp(i\theta X)(e^{ihX} - 1)\right]\right| \leq \mathbb{E}\left[\left|e^{ihX} - 1\right|\right].$$
Since $|e^{ihX} - 1| \leq 2$, the right-hand side tends to zero as $h \to 0$ (uniformly in θ) by the bounded convergence theorem.

5. Assume $\mathbb{E}[|X|^n] < \infty$ for some $n \in \mathbb{N}$, and prove by induction for $0 \leq k < n$ that ϕ is k-times differentiable with

$$\phi_X^{(k)}(\theta) = \mathbb{E}\left[(iX)^k \exp(i\theta X)\right].$$

The base case $k = 0$ is immediate. Assuming the result holds for k, we compute

$$\lim_{h\to 0} \frac{\phi_X^{(k)}(\theta + h) - \phi_X^{(k)}(\theta)}{h} = \lim_{h\to 0} \mathbb{E}\left[(iX)^k \exp(i\theta X)\frac{\exp(ihX) - 1}{h}\right].$$

Since

$$\lim_{h\to 0} \frac{\exp(ihX) - 1}{h} = iX,$$

we need to justify interchanging the expectation and limit. Using the dominated convergence theorem, we obtain

$$\left|X^k \exp(i\theta X)\frac{\exp(ihX) - 1}{h}\right| \leq |X|^{k+1}.$$

Since $\mathbb{E}[|X|^{k+1}] < \infty$, we conclude

$$\phi_X^{(k+1)}(\theta) = \mathbb{E}\left[(iX)^{k+1} \exp(i\theta X)\right].$$

6. Using linearity and independence, we get

$$\phi_{\alpha X+Y}(\theta) = \mathbb{E}\left[\exp(i\theta(\alpha X + Y))\right] = \mathbb{E}\left[\exp(i\alpha\theta X)\exp(i\theta Y)\right].$$

By independence,

$$\mathbb{E}\left[\exp(i\alpha\theta X)\exp(i\theta Y)\right] = \mathbb{E}\left[\exp(i\alpha\theta X)\right]\mathbb{E}\left[\exp(i\theta Y)\right].$$

Thus,

$$\phi_{\alpha X+Y}(\theta) = \phi_X(\alpha\theta)\phi_Y(\theta).$$

□

Definition 8.5. The **moment generating function** is defined as

$$M_X(t) = \mathbb{E}(e^{tX})$$

for $t \in \mathbb{R}$. However, this function may or may not be defined, depending on the distribution of X.

Remark 8.6. On the other hand, the characteristic function always exists.

Lemma 8.7

Let X be a Gaussian random variable with mean μ and variance σ^2. Then the characteristic function of X is given by

$$\phi_X(\theta) = \exp\left(i\mu\theta - \frac{1}{2}\sigma^2\theta^2\right) = e^{i\mu\theta - \frac{1}{2}\sigma^2\theta^2}.$$

Proof. We express X as

$$X = \mu + \sigma Z,$$

where Z is a standard normal random variable. By the properties of characteristic functions,

$$\phi_X(\theta) = \exp(i\mu\theta)\phi_Z(\sigma\theta).$$

Thus, it suffices to show that

$$\phi_Z(\theta) = \exp\left(-\frac{1}{2}\theta^2\right).$$

We compute

$$\phi_Z(\theta) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp(i\theta x) \exp\left(-\frac{1}{2}x^2\right) dx.$$

Rewriting the exponent,

$$\exp\left(-\frac{1}{2}(x - i\theta)^2 + \frac{1}{2}\theta^2\right),$$

we recognize a Gaussian integral:

$$\exp\left(\frac{1}{2}\theta^2\right) \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - i\theta)^2\right) dx = \exp\left(-\frac{1}{2}\theta^2\right).$$

To validate the integral property rigorously, we show that ϕ_Z satisfies the differential equation

$$\phi_Z'(\theta) = -\theta\phi_Z(\theta), \quad \phi_Z(0) = 1.$$

This follows because

$$\phi_Z'(\theta) = i\mathbb{E}[Ze^{i\theta Z}] = i\mathbb{E}[Z\cos(\theta Z)] - \mathbb{E}[Z\sin(\theta Z)].$$

Since Z is symmetric and $Z\cos(\theta Z)$ is odd,

$$\mathbb{E}[Z\sin(\theta Z)] = 0.$$

Further computation shows

$$\phi_Z'(\theta) = -\theta\mathbb{E}[\cos(\theta Z)] = -\theta\phi_Z(\theta),$$

which confirms the required form. □

8.1.1 Uniqueness

Lemma 8.8. Let X, Y be independent random variables. Then, we have

$$\mathbb{E}\left[\exp(-itY)\phi_X(Y)\right] = \mathbb{E}\left[\phi_Y(X - t)\right].$$

Proof. First, note that the subscripts X and Y in ϕ_X and ϕ_Y indicate that these characteristic functions are defined using the laws of X and Y, rather than their specific values.

Since X and Y are independent, the joint law of (X, Y) is given by the product measure. We write $\mathbb{E}_X$ for expectation with respect to X, and $\mathbb{E}_Y$ for expectation with respect to Y, so that

$$\mathbb{E} = \mathbb{E}_X\mathbb{E}_Y = \mathbb{E}_Y\mathbb{E}_X.$$

Defining $\phi_1 = \phi_X$ and $\phi_2 = \phi_Y$, we can rewrite the desired identity as

$$\mathbb{E}_Y\left[\exp(-itY)\phi_1(Y)\right] = \mathbb{E}\left[\phi_2(X-t)\right].$$

We now compute

$$\mathbb{E}_Y\left[\exp(-itY)\phi_1(Y)\right] = \mathbb{E}_Y\left[\exp(-itY)\mathbb{E}_X\left[\exp(iYX)\right]\right].$$

By interchanging the order of expectation using Fubini's theorem (which is justified since ϕ_1 and ϕ_2 are both bounded by 1), we obtain

$$= \mathbb{E}_X\mathbb{E}_Y\left[\exp(iY(X-t))\right] = \mathbb{E}_X\left[\phi_2(X-t)\right].$$

□

Theorem 8.9 (Inversion formula)
Let X be a random variable with characteristic function ϕ_X. Then, for any continuity point x of $F_X(x)$, we have

$$F_X(x) = \lim_{a\to\infty}\int_{-\infty}^{x}\left(\frac{1}{2\pi}\int_{-\infty}^{\infty}\exp[-ist]\phi_X(s)\exp\left[-\frac{s^2}{2a^2}\right]ds\right)dt.$$

Proof. Let Z be a standard Gaussian random variable independent of X, and for $a > 0$, define

$$X_a = X + a^{-1}Z.$$

Then, as $a \to \infty$, we have $X_a \to X$ almost surely, which implies $X_a \Rightarrow X$. Therefore, for every continuity point x of F_X,

$$F_X(x) = \lim_{a\to\infty} F_{X_a}(x).$$

We now compute $F_{X_a}(x)$:

$$F_{X_a}(x) = \mathbb{E}_X\mathbb{E}_Z\left[\mathbf{1}\{X + a^{-1}Z \le x\}\right].$$

Rewriting the expectation, we obtain

$$F_{X_a}(x) = \mathbb{E}_X\int_{-\infty}^{a(x-X)}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt.$$

Changing variables, we express it as

$$F_{X_a}(x) = \mathbb{E}_X\int_{-\infty}^{x}\frac{a}{\sqrt{2\pi}}e^{-a^2(t-X)^2/2}dt.$$

Using Fubini's theorem, which is justified since all terms are positive, we obtain

$$F_{X_a}(x) = \int_{-\infty}^{x}\frac{a}{\sqrt{2\pi}}\mathbb{E}_X\left[e^{-a^2(t-X)^2/2}\right]dt.$$

By expressing the expectation in terms of the characteristic function, we get

$$F_{X_a}(x) = \int_{-\infty}^{x}\frac{a}{\sqrt{2\pi}}\mathbb{E}_X\left[\phi_{aZ}(X-t)\right]dt.$$

Using the characteristic function of a Gaussian, we have

$$F_{X_a}(x) = \int_{-\infty}^{x} \frac{a}{\sqrt{2\pi}} \mathbb{E}_Z \left[e^{-itaZ} \phi_X(aZ) \right] dt.$$

Applying the lemma above, we can write

$$F_{X_a}(x) = \int_{-\infty}^{x} \frac{a}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}a} \exp[-its]\phi_X(s) \exp\left[-\frac{s^2}{2a^2}\right] dsdt.$$

□

Theorem 8.10 (Uniqueness of characteristic function)
Given two probability measures μ, ν on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we have

$$\phi_\mu = \phi_\nu \quad \text{if and only if} \quad \mu = \nu.$$

Proof. This result follows from the inversion formula and the fact that a distribution function is uniquely determined by its values at its points of continuity.

To see why, suppose that F_1 and F_2 are two distribution functions that agree on $\mathbb{R} \setminus D$, where D is the set of all points that are discontinuous for either F_1 or F_2. Since every distribution function has at most countably many points of discontinuity, the set D is countable.

For any $x \in \mathbb{R}$, we can choose a sequence $x_n \downarrow x$ with $x_n \in \mathbb{R} \setminus D$. By the right-continuity of distribution functions, we have

$$F_1(x) = \lim_{n\to\infty} F_1(x_n) = \lim_{n\to\infty} F_2(x_n) = F_2(x).$$

Since this holds for all $x \in \mathbb{R}$, we conclude that $F_1 = F_2$, and thus, $\mu = \nu$. □

8.1.2 Results

Theorem 8.11 (Density inversion formula)
Let μ be a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$\int_{\mathbb{R}} |\phi_\mu(\theta)| d\theta < \infty.$$

Then, μ is absolutely continuous with respect to the Lebesgue measure with a bounded, uniformly continuous density f_μ given by

$$f_\mu(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \exp[-i\theta x]\phi_\mu(\theta) d\theta.$$

Remark 8.12. The theorem above is a Fourier inversion formula. If we already knew that

μ had density f_μ, then we would have

$$\phi_\mu(\theta) = \int_{\mathbb{R}} \exp[i\theta x] f_\mu(x)dx.$$

Proof. For any two continuity points $x < y$ of F_μ, we have

$$F_\mu(y) - F_\mu(x) = \lim_{a\to\infty} \int_x^y \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \exp[-ist]\phi_\mu(s) \exp\left[-\frac{s^2}{2a^2}\right] ds \right) dt.$$

Since

$$\left| \exp[-ist]\phi_X(s) \exp\left[-\frac{s^2}{2a^2}\right] \right| \leq |\phi_X(s)|,$$

which is integrable over $(s,t) \in (-\infty, \infty) \times (x, y)$ by assumption, we apply the dominated convergence theorem to interchange the limit and integral:

$$F_\mu(y) - F_\mu(x) = \int_x^y \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \exp[-ist]\phi_X(s)ds \right) dt.$$

Since intervals (x, y) generate the Borel σ-algebra, this density function uniquely determines μ, □

Lemma 8.13

Given any random variable X and $\delta > 0$, we have

$$\mathbb{P}(|X| > 2/\delta) \leq \frac{1}{\delta} \int_{-\delta}^{\delta} (1 - \phi_X(\theta))d\theta.$$

where the right-hand side is a real number.

Note 8.14. This lemma shows how continuity at 0 relates to tightness.

Note 8.15. Note that the realness of the right-hand side follows from the symmetry of the integration domain, as $\phi_X(-\theta) = \overline{\phi_X(\theta)}$.

Proof. We compute

$$\int_{-\delta}^{\delta} (1 - \phi_X(\theta))d\theta = 2\delta - \int_{-\delta}^{\delta} \mathbb{E}[\exp[i\theta X]]d\theta.$$

Rewriting in terms of cosine,

$$= 2\delta - \int_{-\delta}^{\delta} \mathbb{E}[\cos(\theta X)]d\theta.$$

Using expectation and integral exchange,

$$= 2\delta - \mathbb{E} \int_{-\delta}^{\delta} \cos(\theta X)d\theta.$$

Evaluating the integral,

$$= 2\delta - 2\mathbb{E}\left[\frac{\sin(\delta X)}{X}\right].$$

Thus, we obtain

$$\mathbb{P}(|X| > 2/\delta) = \mathbb{E}\left[\mathbf{1}_{|X|>2/\delta}\right] \leq 2\mathbb{E}\left[\left(1 - \frac{1}{\delta|X|}\right)\mathbf{1}_{|X|>2/\delta}\right].$$

Using the bound $(1 - \sin(r)/r) \geq 0$, we conclude

$$\mathbb{P}(|X| > 2/\delta) \leq \frac{1}{\delta}\int_{-\delta}^{\delta}(1 - \phi_X(\theta))d\theta.$$

□

Theorem 8.16 (Lévy's Continuity Theorem)
Let μ_n be a sequence of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$:

1. If $\mu_n \Rightarrow \mu$ as $n \to \infty$, then $\phi_{\mu_n}(\theta) \to \phi_\mu(\theta)$ for every $\theta \in \mathbb{R}$.
2. If $\phi_{\mu_n}(\theta) \to \phi(\theta)$ for every $\theta \in \mathbb{R}$, and ϕ is continuous at 0, then $\phi = \phi_\mu$ for a probability measure μ, and $\mu_n \Rightarrow \mu$ as $n \to \infty$.

Remark 8.17. In the second statement of Lévy's continuity theorem, continuity at 0 of ϕ is crucial. Consider the sequence $X_n = nZ$, where Z is a standard Gaussian random variable. Then,

$$\phi_{X_n}(\theta) = \exp[-n^2\theta^2/2] \to \mathbf{1}\{\theta = 0\}.$$

However, $F_{X_n}(x) \to 1/2$ for every x, meaning X_n does not converge in distribution (they are not tight).

Proof. We prove each statement in turn.

1. For any fixed θ, the function $g(x) = \exp[i\theta x]$ is bounded and continuous. Since $\mu_n \Rightarrow \mu$, we obtain

$$\phi_{\mu_n}(\theta) = \int_{\mathbb{R}} \exp[i\theta x]d\mu_n(x) \to \int_{\mathbb{R}} \exp[i\theta x]d\mu(x) = \phi_\mu(\theta) \quad \text{as } n \to \infty.$$

2. We first note that, since $|\phi_{\mu_n}| \leq 1$ and $\phi_{\mu_n} \to \phi$ pointwise, we must also have $|\phi| \leq 1$.

   Additionally, we have that $\phi(0) = \lim_{n\to\infty} \phi_{\mu_n}(0) = 1$ and that ϕ is continuous at 0. It follows that, for any $\epsilon > 0$, there exists $\delta > 0$ such that $|\phi(\theta) - 1| < \epsilon/4$ for all $|\theta| < \delta$. We then obtain

$$\frac{1}{\delta}\int_{-\delta}^{\delta}(1 - \phi(\theta))d\theta \leq \frac{1}{\delta}\int_{-\delta}^{\delta}|\phi(\theta) - 1|d\theta < \epsilon/2.$$

By the bounded Convergence Theorem, we also have

$$\lim_{n\to\infty} \frac{1}{\delta}\int_{-\delta}^{\delta}(1-\phi_{\mu_n}(\theta))d\theta = \frac{1}{\delta}\int_{-\delta}^{\delta}(1-\phi(\theta))d\theta < \epsilon/2.$$

Combining this convergence with the previous lemma, there exists some N such that for all $n \geq N$,

$$\mu_n\left([-2/\delta, 2/\delta]^c\right) \leq \frac{1}{\delta}\int_{-\delta}^{\delta}(1-\phi_{\mu_n}(\theta))d\theta < \epsilon.$$

This shows that the sequence μ_n is tight. It follows that there exists a probability measure μ and a subsequence μ_{n_k} such that $\mu_{n_k} \Rightarrow \mu$ as $k \to \infty$.

By Part 1 of Lévy's Continuity Theorem, we already have that $\phi_{\mu_{n_k}}(\theta) \to \phi_\mu(\theta)$ as $k \to \infty$ for all $\theta \in \mathbb{R}$. However, since we also have that $\phi_{\mu_{n_k}}(\theta) \to \phi(\theta)$ for all θ, it follows that $\phi_\mu = \phi$.

Since weak convergence is metrisable, every subsequence of μ_n has a further subsequence that weakly converges to a probability measure $\tilde{\mu}$ with $\phi_{\tilde{\mu}} = \phi$. By uniqueness, we must have $\tilde{\mu} = \mu$, forcing the entire sequence μ_n to weakly converge to μ. □

8.2 Central limit theorem

Note 8.18. Before stating and proving the CLT we need to list some lemmas we will use.

Lemma 8.19. For $x \in \mathbb{R}$ and $n \in \mathbb{N}$, we have the inequality

$$\left| e^{ix} - \sum_{m=0}^{n} \frac{(ix)^m}{m!} \right| \leq \min\left(\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right).$$

Proof. By Taylor's theorem with integral remainder, we obtain for any n

$$e^{ix} = \sum_{m=0}^{n} \frac{(ix)^m}{m!} + \frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is} ds.$$

The remainder term on the right-hand side needs to be estimated using the given bounds. The first bound follows from

$$\left| \frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is} ds \right| \leq \frac{|x|^{n+1}}{(n+1)!}.$$

For the second estimate, we apply integration by parts:

$$\frac{i}{n}\int_0^x (x-s)^n e^{is} ds = -\frac{x^n}{n} + \int_0^x (x-s)^{n-1} e^{is} ds.$$

Thus, we rewrite the remainder as

$$\left| \frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is} ds \right| = \left| \frac{i^n}{(n-1)!}\int_0^x (x-s)^{n-1}(e^{is}-1) ds \right|.$$

Using $|e^{is}-1| \leq 2$, we obtain

$$\left| \frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is} ds \right| \leq \frac{2|x|^n}{n!}.$$

Thus, combining both bounds, we conclude the proof. □

Lemma 8.20. If $\mathbb{E}[X^2] < \infty$, then the characteristic function $\phi_X(t)$ satisfies

$$\phi_X(t) = 1 + it\mathbb{E}[X] - \frac{t^2}{2}\mathbb{E}[X^2] + R_X(t),$$

where $R_X(0) = 0$ and $\lim_{t\to 0} R_X(t)/t^2 = 0$.

Proof. We start by defining the remainder term:

$$|R_X(t)| = \left|\phi_X(t) - 1 - it\mathbb{E}[X] + \frac{t^2}{2}\mathbb{E}[X^2]\right|.$$

Expanding e^{itX} using Taylor's series,

$$= \left|\mathbb{E}\left[e^{itX} - 1 - itX + \frac{t^2}{2}X^2\right]\right|.$$

By the Taylor expansion remainder bound (from the lemma above),

$$\leq \mathbb{E}\left[\min\left(|tX|^3, 2|tX|^2\right)\right].$$

Since $|tX|^3 \leq 2|X|^2$ and $\mathbb{E}[|X|^2] < \infty$, by the Dominated Convergence Theorem, we conclude

$$\mathbb{E}[\min(|t||X|^3, 2|X|^2)] \to 0 \quad \text{as } t \to 0.$$

Thus, $\lim_{t\to 0} R_X(t)/t^2 = 0$. □

Lemma 8.21. We have the following:

1. If $z \in \mathbb{C}$ and $|z| \leq 1$, then $|e^z - 1 - z| \leq |z|^2$.
2. For $\theta > 0$ and $z_1, \ldots, z_n, w_1, \ldots, w_n \in \{z \in \mathbb{C} : |z| \leq \theta\}$, we have

$$\left|\prod_{i=1}^n z_i - \prod_{i=1}^n w_i\right| \leq \theta^{n-1}\sum_{i=1}^n |z_i - w_i|.$$

Proof. We prove each statement in turn.

- Trivial.
- We proceed by induction. For $n = 1$, the statement is immediate. For $n = k + 1$, we have

$$\left|\prod_{i=1}^{k+1} z_i - \prod_{i=1}^{k+1} w_i\right| = \left|z_{k+1}\prod_{i=1}^{k} z_i - w_{k+1}\prod_{i=1}^{k} w_i\right|.$$

  Applying the triangle inequality,

$$\leq |z_{k+1}|\left|\prod_{i=1}^{k} z_i - \prod_{i=1}^{k} w_i\right| + |z_{k+1} - w_{k+1}|\left|\prod_{i=1}^{k} w_i\right|.$$

  Bounding $|z_{k+1}| \leq \theta$ and using induction, we obtain

$$\left|\prod_{i=1}^{k+1} z_i - \prod_{i=1}^{k+1} w_i\right| \leq \theta^k\sum_{i=1}^{k} |z_i - w_i| + \theta^k|z_{k+1} - w_{k+1}|.$$

  Summing completes the proof.

□

Lemma 8.22. Let c_n be a complex sequence with $c_n \to c \in \mathbb{C}$ as $n \to \infty$, then

$$(1 + c_n/n)^n \to e^c \quad \text{as } n \to \infty.$$

Lemma 8.23. Define $r = \sup_n |c_n|$. Choose $\theta = e^{r/n}$, so $|(1+c_n/n)|, |e^{c_n/n}| \leq \theta$. Applying the lemma above, we estimate

$$|(1 + c_n/n)^n - e^{c_n}| \leq \theta^{n-1} \sum_{i=1}^{n} |1 + c_n/n - e^{c_n/n}|.$$

Using the lemma above,

$$|1 + c_n/n - e^{c_n/n}| \leq |c_n/n|^2.$$

Since $e^{c_n/n} \to e^c$, continuity implies $(1 + c_n/n)^n \to e^c$.

Theorem 8.24 (Central limit theorem)

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite second moment, satisfying $\mathbb{E}[X_1] = \mu$ and $\mathrm{Var}[X_1] = \sigma^2$. Define the partial sum

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Then, the normalised sum

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}}$$

converges in distribution to a normal distribution:

$$Z_n \Rightarrow N(0, \sigma^2) \quad \text{as } n \to \infty.$$

Note 8.25. Since the random variables are identically distributed we can pick say X_1.

Proof. We prove the theorem using characteristic functions. The goal is to show that the characteristic function $\phi_{Z_n}(\theta)$ of Z_n converges to that of $N(0, \sigma^2)$, which is $e^{-\theta^2\sigma^2/2}$.

Since adding or subtracting a constant does not affect the variance, we assume without loss of generality that $\mu = 0$, so $\mathbb{E}[X_1] = 0$.

From the definition of characteristic functions and using a lemma from above, we have for small t

$$\phi_{X_n}(\theta) = \mathbb{E}[e^{i\theta X_n}] = 1 - \frac{\theta^2}{2}\sigma^2 + R_{X_n}(\theta),$$

where the remainder term $R_{X_n}(\theta)$ satisfies

$$\lim_{\theta \to 0} \frac{R_{X_n}(\theta)}{\theta^2} = 0.$$

Since $S_n = X_1 + X_2 + \cdots + X_n$, the characteristic function of the normalised sum is given by

$$\phi_{Z_n}(\theta) = \phi_{S_n/\sqrt{n}}(\theta) = \phi_{X_1}(\theta/\sqrt{n})^n.$$

Using the expansion for ϕ_{X_n},

$$\phi_{X_1}(\theta/\sqrt{n}) = 1 - \frac{\theta^2}{2n}\sigma^2 + R_{X_1}(\theta/\sqrt{n}),$$

we raise this expression to the nth power:

$$\phi_{Z_n}(\theta) = \left(1 - \frac{\theta^2}{2n}\sigma^2 + R_{X_1}(\theta/\sqrt{n})\right)^n.$$

Define

$$c_n(\theta) = -\frac{\theta^2\sigma^2}{2} + nR_{X_1}(\theta/\sqrt{n}).$$

By a lemma from above, we know that $nR_{X_1}(\theta/\sqrt{n}) \to 0$ as $n \to \infty$, implying that $c_n(\theta) \to -\frac{\theta^2\sigma^2}{2}$. Applying a lemma from above, we conclude

$$\left(1 + \frac{c_n(\theta)}{n}\right)^n \to e^{c_n(\theta)} \to e^{-\theta^2\sigma^2/2}.$$

Thus, we have

$$\phi_{Z_n}(\theta) \to e^{-\theta^2\sigma^2/2},$$

which is the characteristic function of $N(0, \sigma^2)$. By Lévy's Continuity Theorem, we conclude that

$$Z_n \Rightarrow N(0, \sigma^2) \quad \text{as } n \to \infty.$$

□ □

9 Conditional Expectation

9.1 Definitions

Definition 9.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra. Given a random variable X with $\mathbb{E}[|X|] < \infty$, we say a random variable Y is a **conditional expectation of X given $\mathcal{G}$** if

1. Y is $\mathcal{G}$-measurable.
2. For any $A \in \mathcal{G}$, we have $\mathbb{E}[X \cdot \mathbf{1}_A] = \mathbb{E}[Y \cdot \mathbf{1}_A]$.

> **Remark 9.2.** Keep in mind! The conditional expectation is a random variable not a number.

Lemma 9.3. Let X, Y and $\mathcal{G}$ be as in the definition above, then $\mathbb{E}(|Y|) \leq \mathbb{E}(|X|)$.

Proof. Defining $A = \{Y \geq 0\}$, we have $A, A^c \in \mathcal{G}$ and $|Y| = Y \cdot \mathbf{1}_A - Y \cdot \mathbf{1}_{A^c}$. We have that

$$\mathbb{E}(Y \cdot \mathbf{1}_A) = \mathbb{E}(X \cdot \mathbf{1}_A) \leq \mathbb{E}(|X| \cdot \mathbf{1}_A)$$
$$\mathbb{E}(-Y \cdot \mathbf{1}_{A^c}) = \mathbb{E}(-X \cdot \mathbf{1}_{A^c}) \leq \mathbb{E}(|X| \cdot \mathbf{1}_{A^c}).$$

Adding these two inequalities gives the desired result. □

Exam Questions 9.4 (Exercise from lecture notes)
Let X be a random variable such that X and $-X$ have the same distribution, and let $\mathcal{G} = \sigma(|X|)$. For any bounded measurable function $h \in \mathcal{B}_b$, show that

$$Y = \frac{1}{2}(h(X) + h(-X))$$

is a conditional expectation of $h(X)$ given $\mathcal{G}$.

Solution. We need to show that Y is the conditional expectation of $h(X)$ given $\mathcal{G} = \sigma(|X|)$, meaning that:

1. Y is $\mathcal{G}$-measurable.
2. For any $A \in \mathcal{G}$, we have $\mathbb{E}[h(X) \cdot \mathbf{1}_A] = \mathbb{E}[Y \cdot \mathbf{1}_A]$.

We do each step in turn.

1. Since $\mathcal{G} = \sigma(|X|)$, a function of X is $\mathcal{G}$-measurable if it can be written as a function of $|X|$ alone. Notice that:
$$Y = \frac{1}{2}\left(h(X) + h(-X)\right)$$
depends only on $|X|$, since X and $-X$ have the same distribution. Specifically, if we define $g(|X|) = \frac{1}{2}(h(X) + h(-X))$, then $Y = g(|X|)$ is $\mathcal{G}$-measurable.

2. We need to verify that for any $A \in \mathcal{G}$,
$$\mathbb{E}[h(X) \cdot \mathbf{1}_A] = \mathbb{E}[Y \cdot \mathbf{1}_A].$$
Since A is measurable with respect to $\mathcal{G} = \sigma(|X|)$, it follows that if $\omega \in A$, then it must also be in A for both $X(\omega)$ and $-X(\omega)$. Using the symmetry of X, we obtain:
$$\mathbb{E}[h(X) \cdot \mathbf{1}_A] = \mathbb{E}[h(-X) \cdot \mathbf{1}_A].$$
Taking the average of both sides gives:
$$\mathbb{E}[h(X) \cdot \mathbf{1}_A] = \frac{1}{2}\mathbb{E}[h(X) \cdot \mathbf{1}_A] + \frac{1}{2}\mathbb{E}[h(-X) \cdot \mathbf{1}_A].$$
Rewriting this,
$$\mathbb{E}[h(X) \cdot \mathbf{1}_A] = \mathbb{E}\left[\frac{1}{2}(h(X) + h(-X)) \cdot \mathbf{1}_A\right] = \mathbb{E}[Y \cdot \mathbf{1}_A].$$
This confirms that Y satisfies the definition of $\mathbb{E}[h(X) \mid \mathcal{G}]$.

Theorem 9.5 (Uniqueness of conditional expectation)
Let X and $\mathcal{G}$ be as in the definition of conditional expectation, and suppose that Y and Z are both conditional expectations of X given $\mathcal{G}$. Then $Y = Z$ almost, surely. As such we can write $\mathbb{E}(X \mid \mathcal{G})$ for the 'unique' conditional expectation of X given $\mathcal{G}$.

Proof. Let $A_n = \{Y - Z > \frac{1}{n}\}$, then $A_n \in \mathcal{G}$ and we have

$$\mathbb{E}[Y \cdot \mathbf{1}_{A_n}] = \mathbb{E}[X \cdot \mathbf{1}_{A_n}] = \mathbb{E}[Z \cdot \mathbf{1}_{A_n}].$$

So we have $\mathbb{E}[(Y - Z) \cdot \mathbf{1}_{A_n}] = 0$, but on the other hand

$$\mathbb{E}[(Y - Z)\mathbf{1}_{A_n}] \geq \frac{1}{n}\mathbb{P}(A_n),$$

so we must have $\mathbb{P}(A_n) = 0$ for all n. It follows that

$$\mathbb{P}(Y > Z) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = 0,$$

and by symmetry of the argument we also have $\mathbb{P}(Y < Z) = 0$. □

Something more from lecture notes which is below

Example 9.6
Let $\mathcal{G} = \sigma(\mathcal{D})$ for a partition $\mathcal{D} = \{A_1, \dots, A_n\} \subset \mathcal{F}$ of Ω where $\mathbb{P}(A_i) > 0$ for each i. Then for any integrable random variable Y, we have:

$$\mathbb{E}[Y \mid \mathcal{G}] = \sum_{i=1}^{n} \frac{1}{\mathbb{P}(A_i)} \mathbb{E}[Y \mathbf{1}_{A_i}] \mathbf{1}_{A_i}.$$

Solution. Let Z be the right-hand side of the equation above. By uniqueness, we just need to show that:

1. Z is $\mathcal{G}$-measurable.
2. For every $A \in \mathcal{G}$, $\mathbb{E}[Z\mathbf{1}_A] = \mathbb{E}[Y\mathbf{1}_A]$.

For (1), we simply observe that every $\mathbf{1}_{A_i}$ is clearly $\mathcal{G}$-measurable. For (2), we can easily check that for any $1 \leq i \leq n$,

$$\begin{aligned}
\mathbb{E}[Z\mathbf{1}_{A_i}] &= \sum_{i=1}^{n} \frac{1}{\mathbb{P}(A_i)} \mathbb{E}[Y\mathbf{1}_{A_i}] \mathbb{E}[\mathbf{1}_{A_i}\mathbf{1}_{A_i}] \\
&= \frac{1}{\mathbb{P}(A_i)} \mathbb{E}[Y\mathbf{1}_{A_i}] \mathbb{P}(A_i) = \mathbb{E}[Y\mathbf{1}_{A_i}].
\end{aligned}$$

The conclusion then follows since every $A \in \mathcal{G}$ can be written as $A = \bigcup_{i \in I} A_i$ for some subset $I \subset \{1, \dots, n\}$. Therefore,

$$\mathbb{E}[Y\mathbf{1}_A] = \sum_{i \in I} \mathbb{E}[Y\mathbf{1}_{A_i}] = \sum_{i \in I} \mathbb{E}[Z\mathbf{1}_{A_i}] = \mathbb{E}[Z\mathbf{1}_A].$$

Remark 9.7 (Notation)**.** Note that one often writes, for events A of positive measure,

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

We won't use the above notation so often, but note that the above is a number but often $\mathbb{E}[Y \mid \bullet]$ denotes a random variable—be careful.

Remark 9.8 (Notation)**.** Given a collection of random variables $X_1, X_2, \ldots, X_n$ and an integrable random variable Y, we write

$$\mathbb{E}[Y \mid X_1, X_2, \ldots, X_n] = \mathbb{E}[Y \mid \sigma(X_1, X_2, \ldots, X_n)].$$

Exam Questions 9.9

Let $Y, X_1, \ldots, X_n$ be random variables with $Y \in \sigma(X_1, \ldots, X_n)$. Then there exists a measurable function $f : \mathbb{R}^n \to \mathbb{R}$ such that $Y = f(X_1, \ldots, X_n)$. This indicates that the conditional expectation above is really guessing Y using the values of $X_1, \ldots, X_n$.

Solution. We prove the above statement for $Y \geq 0$, the general case follows by writing $Y = Y^+ - Y^-$. We set $\widetilde{X} = (X_1, \ldots, X_n)$, so that $\widetilde{X} : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is a measurable map. It follows that

$$\mathcal{G} := \sigma(X_1, \ldots, X_n) = \{\widetilde{X}^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^n)\}.$$

Note that, for every $A \in \mathcal{G}$, there exists $B \in \mathcal{B}(\mathbb{R}^n)$ such that we can write:

$$\mathbf{1}_A(\omega) = \mathbf{1}_B(\widetilde{X}(\omega)).$$

This verifies the claim when Y is an indicator function. If Y is not an indicator function, we know that there exist simple functions $Y_n \uparrow Y$. It follows that there exist simple functions f_n such that $Y_n = f_n(\widetilde{X})$. In particular, for every $\omega \in \Omega$, we have

$$\lim_{n\to\infty} f_n(\widetilde{X}(\omega)) = Y(\omega),$$

in particular the former limit exists. We would like to take $f = \lim_{n\to\infty} f_n$, but we don't know that $\lim_{n\to\infty} f_n(x)$ exists for every $x \in \mathbb{R}^n$ (we only know this for x in the image of $\widetilde{X}$). However, we can just set:

$$f(x) = \begin{cases} \lim_{n\to\infty} f_n(x) & \text{if the limit exists,} \\ 0 & \text{otherwise.} \end{cases}$$

This function is measurable and satisfies the desired property.

Remark 9.10. Sometimes you'll see the notation $\mathbb{E}[Y \mid X = x]$, denoting a "function of x". Note that $\mathbb{E}[Y \mid X = x]$ is uniquely defined only up to $\mathbb{P}$-almost everywhere equivalence. To relate this notation to what we saw earlier, if we write $\mathbb{E}[Y \mid X] = f(X)$ for some measurable $f : \mathbb{R} \to \mathbb{R}$, we have:

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

9.1.1 Existence

Theorem 9.11 (Radon-Nikodym Derivative)
Let μ and ν be finite measures on a measurable space $(\Omega, \mathcal{F})$. Then there exists a non-negative measurable function $f : \Omega \to \mathbb{R}$ and a set $B \in \mathcal{F}$ such that $\mu(B) = 0$ and, for all $A \in \mathcal{F}$

$$\nu(A) = \int_A f d\mu + \nu(B \cap A).$$

Remark 9.12. When $\nu \ll \mu$ on can take $B = \varnothing$, so $\nu(B \cap A) = 0$.

Proof. Omitted. Not examinable. □

Remark 9.13. The statement of the theorem is examinable.

Theorem 9.14 (Existence of conditional expectation)
Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$ and Y be an integrable random variable. Then there exists a random variable Z such that Z is the conditional expectation of Y given $\mathcal{G}$.

Proof. We prove the claim for $Y \geq 0$, the general case follows by writing $Y = Y^+ - Y^-$, invoking existence for Y^+ and Y^-, and then verifying that

$$\mathbb{E}[Y \mid \mathcal{G}] = \mathbb{E}[Y^+ \mid \mathcal{G}] - \mathbb{E}[Y^- \mid \mathcal{G}]$$

works. Now, since Y is integrable, the function $\nu(A) = \mathbb{E}[Y\mathbf{1}_A]$ defines a finite measure on $(\Omega, \mathcal{G})$. In particular, writing $\widetilde{\mathbb{P}}$ for the restriction of the probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ to $(\Omega, \mathcal{G})$, it follows from the Radon-Nikodym theorem that there exists a $\mathcal{G}$-measurable function f such that, for every $A \in \mathcal{G}$,

$$\mathbb{E}[Y\mathbf{1}_A] = \int_A f \, d\widetilde{\mathbb{P}} = \mathbb{E}[f\mathbf{1}_A].$$

□

9.2 Properties

Remark 9.15. Recall that $\mathbb{E}(X \mid \mathcal{G})$ is a function on Ω, not a number.

Some example from the lecture notes at page 53 merged

Proposition 9.16. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X, Y integrable random variables, $\mathcal{G} \subset \mathcal{F}$ a sub-σ-algebra and $a, b \in \mathbb{R}$. Then, the following properties hold:

1. $\mathbb{E}(\mathbb{E}(X \mid \mathcal{G})) = \mathbb{E}(X)$.
2. If $\mathcal{G} = \{\varnothing, \Omega\}$, then $\mathbb{E}(X \mid \mathcal{G}) = \mathbb{E}(X)$.
3. If X is $\mathcal{G}$-measurable, then $\mathbb{E}(X \mid \mathcal{G}) = X$ almost surely.
4. (Linearity). $\mathbb{E}(aX + bY \mid \mathcal{G}) = a\mathbb{E}(X \mid \mathcal{G}) + b\mathbb{E}(Y \mid \mathcal{G})$ almost surely.
5. (Monotonicity). If $X \leq Y$ almost surely, then $\mathbb{E}(X \mid \mathcal{G}) \leq \mathbb{E}(Y \mid \mathcal{G})$ alsmost surely.
6. ($\triangle$-inequality). $|\mathbb{E}(X \mid \mathcal{G})| \leq \mathbb{E}(|X| \mid \mathcal{G})$ almost surely.

Proof. We prove each statement in turn.

1. Take $\Omega \in \mathcal{G}$

   To continue

2. If $\mathcal{G} = \{\varnothing, \Omega\}$ the $\mathbb{E}(X \mid \mathcal{G})$ is a constant function, so that $\mathbb{E}(\mathbb{E}(X \mid \mathcal{G})) = \mathbb{E}(X)$.
3. This follows from the definition of conditional expectation.
4.

□

Proposition 9.17
Let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra and $X_1, \dots, X_n, \dots$, and X be integrable random variables.

1. **(Monotone Convergence)** If $X_n \uparrow X$ almost surely, then
$$\mathbb{E}[X_n \mid \mathcal{G}] \uparrow \mathbb{E}[X \mid \mathcal{G}] \quad \text{almost surely.}$$
2. **(Fatou's Lemma)** If $X_n \geq 0$ almost surely, then
$$\mathbb{E}[\liminf_n X_n \mid \mathcal{G}] \leq \liminf_n \mathbb{E}[X_n \mid \mathcal{G}] \quad \text{almost surely.}$$
3. **(Dominated Convergence)** If $|X_n| \leq Y$ almost surely for some integrable random variable Y, and $X_n \to X$ almost surely, then
$$\mathbb{E}[X_n \mid \mathcal{G}] \to \mathbb{E}[X \mid \mathcal{G}] \quad \text{almost surely.}$$
4. **(Jensen's Inequality)** If ϕ is a convex function such that $\phi(X)$ is integrable, then
$$\phi(\mathbb{E}[X \mid \mathcal{G}]) \leq \mathbb{E}[\phi(X) \mid \mathcal{G}] \quad \text{almost surely.}$$

Theorem 9.18 (Tower property)
Let $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{F}$ be σ-algebras and Y be an integrable random variable. Then
$$\mathbb{E}\left(\mathbb{E}(Y \mid \mathcal{G}_2) \mid \mathcal{G}_1\right) = \mathbb{E}\left(\mathbb{E}(Y \mid \mathcal{G}_1) \mid \mathcal{G}_2\right) = \mathbb{E}(Y \mid \mathcal{G}_1).$$

Note 9.19. In a repeated conditional expectation the smallest σ-algebra wins.

Proof.

TO do

□

Theorem 9.20 (Factorisation property)
Let X and Y be integrable random variables with XY integrable as well, and suppose $\mathcal{G} \subset \mathcal{F}$ be is a σ-algebra and X is $\mathcal{G}$-measurable. Then

$$\mathbb{E}(XY \mid \mathcal{G}) = X \cdot \mathbb{E}(Y \mid \mathcal{G}).$$

Proof.

TO do

□

10 Martingales

10.1 Definitions

Note 10.1. 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.

Definition 10.2. A **filtration** (of σ-algebras) on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is an increasing sequence of σ-algebras $(\mathcal{F}_n)_{n=1}^{\infty}$ such that

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_n \subset \cdots \subset \mathcal{F}.$$

Note 10.3. 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. $n = 1$, but this is not always the case, so take a little care.

Remark 10.4. We can relabel and use $n = 0$.

Definition 10.5. We sat a sequence of random variables is **adapted** to the filtration $\mathcal{F}_n$ if for each $n \in \mathbb{N}$, the random variable X_n is $\mathcal{F}_n$-measurable.

Definition 10.6. The **natural filtration**, $(\mathcal{G}_n)_{n=1}^{\infty}$, associated with the sequence of random variables $(X_n)_{n=1}^{\infty}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is given by

$$\mathcal{G}_n = \sigma(X_1, X_2, \ldots, X_n) = \sigma\left(\bigcup_{i=1}^{\infty} \sigma(X_i)\right) \quad \text{for each } n \geq 0.$$

Proposition 10.7

Given a sequence of random variables, there is a sort of 'default' filtration we can choose. This is the natural filtration.

Proof. This is a choice we make, but I chose to write as a proposition. □

Definition 10.8. Let $(\mathcal{F}_n)_{n=1}^{\infty}$ be a filtration and $(X_n)_{n=1}^{\infty}$ be a sequence of random variables adapted to the filtration. We say $(X_n)_{n=1}^{\infty}$ is a

- **martingale** for the filtration $(\mathcal{F}_n)_{n=1}^{\infty}$, if for each n the following hold
  - $\mathbb{E}(|X_n|) < \infty$, and
  - $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) = X_n$.
- a **submartingale** if $\mathbb{E}(|X_n|) < \infty$, and $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \geq X_n$.
- a **supermartingale** if $\mathbb{E}(|X_n|) < \infty$, and $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \leq X_n$.

Remark 10.9. We call $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) = X_n$ the martingale property.

Note 10.10. 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 horse's head from rearing up, i.e. it keeps the horse's 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 horse's head is forced down on average.

Proposition 10.11 (Elementary properties)

Suppose $(X_n)_{n=1}^{\infty}$ is a sequence of random variables adapted to a filtration process.

1. $(X_n)_{n=1}^{\infty}$ is a submartingale if and only if $(-X_n)_{n=1}^{\infty}$ is a supermartingale, with respect to the same filtration.
2. $(X_n)_{n=1}^{\infty}$ is martingale if and only if it is submartingale and supermartingale.

Exam Questions 10.12 (Exercise from lecture notes)
Let $(X_n)_{n=1}^{\infty}$ be a martingale for a filtration $(\mathcal{F}_n)_{n=1}^{\infty}$, then for any $n \geq m$, one has

$$\mathbb{E}[X_n \mid \mathcal{F}_m] = X_m.$$

Solution. We prove the statement for all $n > m$ by induction.

- **Base Case:** For $n = m + 1$, we recall the martingale property, which states that for a martingale (X_n), we have

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

  Applying this directly to $n = m$, we obtain:

$$\mathbb{E}[X_{m+1} \mid \mathcal{F}_m] = X_m.$$

- **Inductive Step:** Assume that for some $n > m$, the desired property holds, i.e.,

$$\mathbb{E}[X_n \mid \mathcal{F}_m] = X_m.$$

  We need to show that it also holds for $n + 1$, i.e.,

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_m] = X_m.$$

  Using the tower property, we write:

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_m] = \mathbb{E}[\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \mid \mathcal{F}_m].$$

  Now, by the martingale property, we know that

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = X_n.$$

  Substituting this into the previous equation gives:

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_m] = \mathbb{E}[X_n \mid \mathcal{F}_m].$$

  By the inductive hypothesis, we already assumed that

$$\mathbb{E}[X_n \mid \mathcal{F}_m] = X_m.$$

  Thus, we conclude:

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_m] = X_m.$$

Some lemma with proof in problem sheet 17

Lemma 10.13. Let X be an integrable random variable and $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra such that $\sigma(X)$ and $\mathcal{G}$ are independent (that is, X and $\mathcal{G}$ are independent). Then

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

Proof. Clearly $\mathbb{E}[X]$ is $\mathcal{G}$-measurable since it is a constant. Note that if we write $X = X^+ - X^-$, then both X^+ and X^- are also independent of $\mathcal{G}$. Therefore, by linearity of conditional expectation and expectation, it suffices to prove the claim for $X \geq 0$. By the uniqueness of conditional expectations, it suffices to check that for all $A \in \mathcal{G}$, we have

$$\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}\left[\mathbb{E}[X \mid \mathcal{G}]\mathbf{1}_A\right] = \mathbb{E}[X]\mathbb{P}(A).$$

Now, know that there exists simple functions Y_n, measurable with respect to $\sigma(X)$, such that $Y_n \uparrow X$ as $n \to \infty$. We also have, by monotone convergence for conditional expectations, that $\mathbb{E}[Y_n\mathbf{1}_A] \uparrow \mathbb{E}[X\mathbf{1}_A]$ and, by monotone convergence for expectations, that $\mathbb{E}[Y_n] \uparrow \mathbb{E}[X]$. Therefore, we can finish the proof by observing that for any $B \in \sigma(X)$, one has, by independence,

$$\mathbb{E}[\mathbf{1}_B\mathbf{1}_A] = \mathbb{E}[\mathbf{1}_B]\mathbb{P}(A).$$

It follows that $\mathbb{E}[Y_n\mathbf{1}_A] = \mathbb{E}[Y_n]\mathbb{P}(A)$, which finishes the proof. □

Example 10.14 (Sum of i.i.d. random variables of mean 0)

Let $(X_n)_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with $\mathbb{E}[|X_i|] < \infty$, $\mathbb{E}[X_i] = 0$, and let $(\mathcal{F}_n)_{n=1}^{\infty}$ be the natural filtration associated to $(X_n)_{n=1}^{\infty}$. Then the sequence $(S_n)_{n=1}^{\infty}$ defined by

$$S_n = X_1 + \cdots + X_n$$

is a martingale for the filtration $(\mathcal{F}_n)_{n=1}^{\infty}$. Moreover, suppose that $\mathbb{E}[X_i^2] = \sigma^2 \in [0, \infty)$, then the sequence of random variables $M_n = S_n^2 - n\sigma^2$ is also a martingale for the filtration $(\mathcal{F}_n)_{n=1}^{\infty}$.

We prove each statement in turn.

- For the first statement we first note that $\mathbb{E}[|S_n|] \leq n\mathbb{E}[|X_1|] < \infty$. For the martingale property, we have

$$\mathbb{E}[S_{n+1}|\mathcal{F}_n] = \mathbb{E}[S_n + X_{n+1}|\mathcal{F}_n] = S_n + \mathbb{E}[X_{n+1}|\mathcal{F}_n] = S_n + \mathbb{E}[X_{n+1}] = S_n$$

  In the second equality we used that S_n is $\mathcal{F}_n$-measurable and in the third equality we used that X_{n+1} is independent of $\mathcal{F}_n$.

- For the second statement we note that $\mathbb{E}[|M_n|] = \mathbb{E}[S_n^2 + n\sigma^2] = n\mathbb{E}[X_1^2] + n\sigma^2 < \infty$. To verify the martingale property, we observe that

$$\begin{aligned}
\mathbb{E}[M_{n+1}|\mathcal{F}_n] &= \mathbb{E}[S_{n+1}^2 - (n+1)\sigma^2|\mathcal{F}_n] \\
&= \mathbb{E}[S_n^2 + 2S_nX_{n+1} + X_{n+1}^2 - (n+1)\sigma^2|\mathcal{F}_n] \\
&= S_n^2 - n\sigma^2 + 2S_n\mathbb{E}[X_{n+1}|\mathcal{F}_n] + \mathbb{E}[X_{n+1}^2|\mathcal{F}_n] - \sigma^2 \\
&= S_n^2 - n\sigma^2 + 2S_n\mathbb{E}[X_{n+1}] + \mathbb{E}[X_{n+1}^2] - \sigma^2 \\
&= S_n^2 - n\sigma^2 = M_n
\end{aligned}$$

  In the third equality we used that $\mathbb{E}[S_nX_{n+1}|\mathcal{F}_n] = S_n\mathbb{E}[X_{n+1}|\mathcal{F}_n]$ since S_n is $\mathcal{F}_n$-measurable, and in the fourth equality we used that X_{n+1} is independent of $\mathcal{F}_n$.

Example 10.15 (Product of non-negative independent random variables of mean 1)
Let $(X_n)_{n=1}^{\infty}$ be a sequence of i.i.d. non-negative random variables with $\mathbb{E}[X_i] = 1$, and let $(\mathcal{F}_n)_{n=1}^{\infty}$ be the natural filtration associated to $(X_n)_{n=1}^{\infty}$.
Define a filtration by setting $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Then, the sequence $(M_n)_{n=1}^{\infty}$ defined by

$$M_n = \prod_{j=1}^{n} X_j$$

is a martingale for the filtration $(\mathcal{F}_n)_{n=1}^{\infty}$. We first note that $\mathbb{E}[|M_n|] = \mathbb{E}[\prod_{j=1}^{n} X_j] = 1 < \infty$. To verify the martingale property, we have

$$\mathbb{E}[M_{n+1}|\mathcal{F}_n] = \mathbb{E}\left[\prod_{j=1}^{n+1} X_j \,\middle|\, \mathcal{F}_n\right] = \prod_{j=1}^{n} X_j \, \mathbb{E}[X_{n+1}|\mathcal{F}_n] = \prod_{j=1}^{n} X_j \cdot \underbrace{\mathbb{E}[X_{n+1}]}_{=1} = M_n$$

Example 10.16 (Martingale from Conditional Expectations)
Let X_∞ be a random variable with $\mathbb{E}[|X_\infty|] < \infty$ and let $(\mathcal{F}_n)_{n=1}^{\infty}$ be a filtration. Define the sequence $(X_n)_{n=1}^{\infty}$ by

$$X_n = \mathbb{E}[X_\infty \mid \mathcal{F}_n].$$

Then $(X_n)_{n=1}^{\infty}$ is a martingale for the filtration $(\mathcal{F}_n)_{n=1}^{\infty}$.

Proof. We first note that

$$\mathbb{E}[|X_i|] = \mathbb{E}[\mathbb{E}[|X_\infty| \mid \mathcal{F}_i]] \leq \mathbb{E}[\mathbb{E}[|X_\infty| \mid \mathcal{F}_n]] = \mathbb{E}[|X_\infty|] < \infty.$$

In the first inequality, we used Jensen's inequality for conditional expectations, which states that $\mathbb{E}[| \bullet | \mid \mathcal{G}] \geq |\mathbb{E}[\bullet \mid \mathcal{G}]|$. In the last equality, we used the tower property.
To verify the martingale property, we observe:

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[\mathbb{E}[X_\infty \mid \mathcal{F}_{n+1}] \mid \mathcal{F}_n] = \mathbb{E}[X_\infty \mid \mathcal{F}_n] = X_n.$$

In the first equality, we used the tower property. □

10.2 Martingale transform

Note 10.17. We now describe using martingales a gambling strategy. Think of $(X_n)_{n=0}^{\infty}$ as representing your running score in a game, with $X_0 = 0$ and higher scores being better. Then

$$\begin{aligned} &\mathbb{E}(X_{n+1} - X_n \mid \mathcal{F}_n) = 0 && \text{is a game} \\ &\mathbb{E}(X_{n+1} - X_n \mid \mathcal{F}_n) \leq 0 && \text{is an unfavourable game} \end{aligned}$$

Definition 10.18. Given a filtration, a sequence of random variables $(H_n)_{n=1}^{\infty}$ is called **previsible** if H_n is $\mathcal{F}_{n-1}$ measurable for each n.

Note 10.19. A previsible process $(H_n)_{n=1}^{\infty}$ can be thought as a betting strategy, where H_n is the amount you bet at the beginning of round n before you see the score update.

Definition 10.20. Let $(X_n)_{n=0}^{\infty}$ be adapted to the filtration $(\mathcal{F}_n)_{n=0}^{\infty}$ and let $(H_n)_{n=1}^{\infty}$ be previsible to the same filtration. We define a new sequence

$$(H \bullet X)_n = \sum_{k=1}^{n} H_k(X_k - X_{k-1})$$

which we call the **martingale transform of** (X_n) **by** H_n.

Remark 10.21. This is the discrete version of a stochastic integral.

Exam Questions 10.22

Show that, in the setting above, $(H \bullet X)_n$ is $\mathcal{F}_n$-measurable.

Solution. To show that $(H \bullet X)_n$ is $\mathcal{F}_n$-measurable, we recall that the stochastic integral (or discrete-time martingale transform) is defined as

$$(H \bullet X)_n = \sum_{k=1}^{n} H_k(X_k - X_{k-1}).$$

Since we are given that $(X_n)_{n\geq 0}$ is a super-martingale adapted to the filtration $(\mathcal{F}_n)_{n\geq 0}$, we have that each X_n is $\mathcal{F}_n$-measurable. Moreover, the sequence $(H_n)_{n\geq 1}$ is previsible, meaning that H_n is $\mathcal{F}_{n-1}$-measurable for all $n \geq 1$. To prove $(H \bullet X)_n$ is $\mathcal{F}_n$-measurable, we proceed by induction:

- Base Case ($n = 1$).
$$(H \bullet X)_1 = H_1(X_1 - X_0).$$
  - Since X_1 is $\mathcal{F}_1$-measurable and X_0 is constant, $X_1 - X_0$ remains $\mathcal{F}_1$-measurable.
  - Since H_1 is $\mathcal{F}_0$-measurable (by previsibility), the product $H_1(X_1 - X_0)$ is $\mathcal{F}_1$-measurable.

- Suppose $(H \bullet X)_{n-1}$ is $\mathcal{F}_{n-1}$-measurable. We show that $(H \bullet X)_n$ is $\mathcal{F}_n$-measurable.
$$(H \bullet X)_n = (H \bullet X)_{n-1} + H_n(X_n - X_{n-1}).$$
  - By the induction hypothesis, $(H \bullet X)_{n-1}$ is $\mathcal{F}_{n-1}$-measurable.
  - Since H_n is $\mathcal{F}_{n-1}$-measurable and $X_n - X_{n-1}$ is $\mathcal{F}_n$-measurable (as X_n is adapted to $\mathcal{F}_n$), their product is also $\mathcal{F}_n$-measurable.
  - The sum of an $\mathcal{F}_{n-1}$-measurable variable and an $\mathcal{F}_n$-measurable variable remains $\mathcal{F}_n$-measurable.

Theorem 10.23

Suppose that $(X_n)_{n=0}^{\infty}$ is a super-martingale for the filtration $(\mathcal{F}_n)_{n=0}^{\infty}$ and that $(H_n)_{n=1}^{\infty}$ is a non-negative previsible process with each H_n bounded. Then, the sequence $(H \bullet X)_n$ is a super-martingale for the filtration $(\mathcal{F}_n)_{n=0}^{\infty}$.

Remark 10.24. The analogous result holds for sub-martingales, and the result also holds for martingales even if $(H_n)_{n=1}^{\infty}$ is not necessarily non-negative.

Note 10.25. The result essentially states that a strategy based on a previsible, bounded process cannot systematically turn a super-martingale into a martingale or sub-martingale, reflecting the principle that one cannot use a betting strategy to get ahead in an unfavorable game.

Proof. Since each H_n is bounded and X_n is a super-martingale, we have $\mathbb{E}[|(H \bullet X)_n|] < \infty$. We now show that

$$\mathbb{E}[(H \bullet X)_{n+1}|\mathcal{F}_n] \leq (H \bullet X)_n.$$

By definition of the martingale transform,

$$\mathbb{E}[(H \bullet X)_{n+1}|\mathcal{F}_n] = \mathbb{E}[(H \bullet X)_n + H_{n+1}(X_{n+1} - X_n)|\mathcal{F}_n].$$

Since expectation is linear, we split the terms:

$$= (H \bullet X)_n + \mathbb{E}[H_{n+1}(X_{n+1} - X_n)|\mathcal{F}_n].$$

Since $(X_n)_{n=0}^{\infty}$ is a super-martingale, we know that

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] \leq X_n.$$

Additionally, since H_{n+1} is previsible, we have H_{n+1} being $\mathcal{F}_n$-measurable. Thus, we obtain:

$$\mathbb{E}[H_{n+1}(X_{n+1} - X_n)|\mathcal{F}_n] = H_{n+1}\mathbb{E}[X_{n+1} - X_n|\mathcal{F}_n].$$

By the super-martingale property, $\mathbb{E}[X_{n+1} - X_n|\mathcal{F}_n] \leq 0$, which implies

$$\mathbb{E}[H_{n+1}(X_{n+1} - X_n)|\mathcal{F}_n] \leq H_{n+1}(X_n - X_n) = 0.$$

Thus, we conclude

$$\mathbb{E}[(H \bullet X)_{n+1}|\mathcal{F}_n] \leq (H \bullet X)_n.$$

In the last inequality, we used that $H_{n+1} \geq 0$. However, this argument also applies to sub-martingales. In the case of martingales, we simply have

$$\mathbb{E}[H_{n+1}(X_{n+1} - X_n)|\mathcal{F}_n] = 0$$

without any assumptions on H_{n+1}. $\square$

10.3 Stopping times

Definition 10.26. A $\mathbb{N} \cup \{\infty\}$-valued random variable τ is called a **stopping time** with respect to the filtration $(\mathcal{F}_n)_{n=0}^{\infty}$ if, for each n, the event $\{\tau \leq n\}$ is $\mathcal{F}_n$-measurable.

Note 10.27. We no show that you cannot beat the house by "stopping" when you feel like it.

Lemma 10.28. Let τ be a stopping time with respect to the filtration $(\mathcal{F}_n)_{n=0}^{\infty}$ and let $(X_n)_{n=0}^{\infty}$ be adapted to the same filtration. Then, the stopped process $(X_{\tau\wedge n})_{n=0}^{\infty}$ is adapted to the filtration $(\mathcal{F}_n)_{n=0}^{\infty}$.

Proof. We will write $(X_{\tau\wedge n})_{n=0}^{\infty}$ as a martingale transform. We define, for $n \in \mathbb{N}$, the process $H_n = 1_{\{\tau \geq n\}}$. We claim that H_n is previsible, so it suffices to show that $\{\tau \geq n\}$ is $\mathcal{F}_{n-1}$-measurable. To see this, note that $\{\tau \geq n\} = \{\tau < n\}^c$ and that

$$\{\tau < n\} = \bigcup_{k=0}^{n-1} \{\tau = k\} \in \mathcal{F}_{n-1}.$$

Thus, $\{\tau \geq n\} \in \mathcal{F}_{n-1}$, proving that H_n is previsible. Now we observe that

$$\begin{aligned}(H \bullet X)_n &= \sum_{k=1}^{n} H_k(X_k - X_{k-1}) \\ &= \sum_{k=1}^{n} 1_{\{\tau \geq k\}}(X_k - X_{k-1}) \\ &= \sum_{k=1}^{\tau\wedge n}(X_k - X_{k-1}) \\ &= X_{\tau\wedge n} - X_0.\end{aligned}$$

Since we have that $(H \bullet X)_n$ is $\mathcal{F}_n$-measurable, and X_0 is $\mathcal{F}_0$-measurable, it follows that $X_{\tau\wedge n}$ is $\mathcal{F}_n$-measurable. □

Corollary 10.29. If $(X_n)_{n=0}^{\infty}$ is a super-martingale and τ is a stopping time, both with respect to the filtration $(\mathcal{F}_n)_{n=0}^{\infty}$, then the stopped process $(X_{\tau\wedge n})_{n=0}^{\infty}$ is a super-martingale.

Proof. This follows from setting $H_n = 1_{\{\tau \geq n\}}$ in Theorem 1, then we have $X_{\tau\wedge n} = (H \bullet X)_n + X_0$. However, we know that $(H \bullet X)_n$ is a sub-martingale and the constant process $Z_n = X_0$ for all n is a martingale. Since the difference of a sub-martingale and a martingale is a martingale, the result follows. □

Lemma 10.30. Let $(X_n)_{n=0}^{\infty}$ be a martingale for the filtration $(\mathcal{F}_n)_{n=0}^{\infty}$ and let ϕ be a convex function such that, for any $n \in \mathbb{N}$,

$$\mathbb{E}[|\phi(X_n)|] < \infty.$$

Then $(\phi(X_n))_{n=0}^{\infty}$ is a sub-martingale.

Proof. Suppose that $(X_n)_{n=0}^{\infty}$ is a martingale. Then, by Jensen's inequality for conditional expectations, we have

$$\mathbb{E}[\phi(X_{n+1})|\mathcal{F}_n] \geq \phi(\mathbb{E}[X_{n+1}|\mathcal{F}_n]) = \phi(X_n).$$

When ϕ is increasing and $(X_n)_{n=0}^{\infty}$ is a sub-martingale, we also have

$$\mathbb{E}[\phi(X_{n+1})|\mathcal{F}_n] \geq \phi(\mathbb{E}[X_{n+1}|\mathcal{F}_n]) \geq \phi(X_n).$$

Thus, $(\phi(X_n))_{n=0}^{\infty}$ is a sub-martingale. □

10.4 Upcrossing

Definition 10.31. Let $a, b \in \mathbb{R}$ with $a < b$, and let $(X_n)_{n=0}^{\infty}$ be a sequence of random variables. We define the stopping times:

$$\tau_0 = -1, \quad \tau_{2k-1} = \inf\{n > \tau_{2k-2} : X_n \leq a\}, \quad \tau_{2k} = \inf\{n > \tau_{2k-1} : X_n \geq b\}$$

We define the number of completed upcrossings by time n as:

$$U_n := \sup\{k : \tau_{2k} \leq n\}$$

Note 10.32. This process records alternating times where the process first drops below a, and then later rises above b. Each such pair defines an "upcrossing" from below a to above b.

Note 10.33. U_n counts how many times the process has gone from below a to above b by time n.

Theorem 10.34 (Upcrossing inequality)
Let $(X_n)_{n=0}^{\infty}$ be a submartingale with respect to the filtration $(\mathcal{F}_n)_{n=0}^{\infty}$, and let $a < b$ be real numbers. Then for any $n \in \mathbb{N}$,

$$(b-a)\mathbb{E}[U_n] \leq \mathbb{E}[(X_n - a)^+] - \mathbb{E}[(X_0 - a)^+]$$

Remark 10.35. Where $Y^+ = \max(Y, 0)$.

10.5 Martingale Convergence Theorem

Theorem 10.36
Let $(X_n)_{n=0}^{\infty}$ be a sub-martingale with

$$\sup_{n \in \mathbb{N}} \mathbb{E}[X_n^+] < \infty.$$

Then, there exists an integrable random variable X such that $X_n \to X$ **almost surely** as $n \to \infty$.

Note 10.37. Recall $X_n^+ = \max(X_n, 0)$.

Corollary 10.38
Let $(X_n)_{n=0}^{\infty}$ be a super-martingale with $X_n \geq 0$ almost surely, then there exists an integrable X such that $X_n \to X$ **almost surely** as $n \to \infty$. Moreover, $\mathbb{E}[X] \leq \mathbb{E}[X_0]$.

11 Mastery material

Remark 11.1. This is from [DD19, Section 4.3.3.] and [DD19, Section 4.4.].

11.1 Prelimary definitions

Definition 11.2. We say a measure μ is **absolutely continuous with respect to** ν, and write $\mu \ll \nu$ if $\nu(A) = 0$ implies $\mu(A) = 0$.

Definition 11.3. Two measure μ and ν are said to be **mutually singular** if there is a set A with $\mu(A) = 0$ and $\nu(A^c) = 0$. In this case, we also say μ is **singular with respectively to** ν and write $\mu \perp \nu$.

11.2 Radon-Nikodym derivatives

Remark 11.4. In this section we assume the following. Let μ be a finite measure and ν a probability measure on $(\Omega, \mathcal{F})$. Let $\mathcal{F}_n \uparrow \mathcal{F}$ be σ-algebras. Let μ_n and ν_n be the restrictions of μ and ν to $\mathcal{F}_n$.

Theorem 11.5 (Generalised Radon-Nikodym theorem)

Suppose $\mu_n \ll \nu_n$ for all n. Let $X_n = \dfrac{\mathrm{d}\mu_n}{\mathrm{d}\nu_n}$ and let $X = \limsup X_n$. Then

$$\mu(A) = \int_A X \, d\nu + \mu(A \cap \{X = \infty\}).$$

Note 11.6. We call X_n the Radon-Nikodym derivative.

Note 11.7. What is this theorem saying? The total μ-measure of a measureable set A (i.e. $A \in \mathcal{F}$) can be decomposed in an absolutely continuous part (given by the integral) and singular part.

Remark 11.8. The term $\mu_r(A) = \int_A X \, d\nu$ is measure $\ll \nu$. In the next lemma we show that X_n is martingale, as such using the martingale convergence theorem (of supermartingales) we have that $\nu(X = \infty) = 0$. This is because the expectation $\mathbb{E}_\nu[X] \leq \mathbb{E}_\nu[X_0]$ (i.e. is bounded) therefore, X cannot be infinite otherwise the expectation would be unbounded. It follows then $\mu_s(A) = \mu(A \cap \{X = \infty\})$ is singular with respect ot ν. Therefore, $\mu = \mu_r + \mu_s$ gives the Lebesgue decomposition of μ, and $X_\infty = \dfrac{\mathrm{d}\mu_r}{\mathrm{d}\nu}$ almost surely.

To prove the theorem we need the following lemma.

Lemma 11.9 (Radon-Nikodym derivative is a martingale)

X_n (defined on $(\Omega, \mathcal{F}, \nu)$) is a martingale with respect to $\mathcal{F}_n$.

Proof. We observe that by definition $X_n \in \mathcal{F}_n$. Let $A \in \mathcal{F}_n$. Since $X_n \in \mathcal{F}_n$ and ν_n is the restriction of ν to $\mathcal{F}_n$ we have

$$\int_A X_n \, d\nu = \int_A X_n d\nu_n.$$

Using the definition of X_n and the Lebesgue decomposition we can can write

$$\int_A X_n \, d\nu_n = \mu_n(A) = \mu(A)$$

where the last equality holds since $A \in \mathcal{F}_n$ and μ_n is the restriction of μ to $\mathcal{F}_n$. If $A \in \mathcal{F}_{m-1} \subset \mathcal{F}_m$, using the last result for $n = m$ and $n = m - 1$ gives

$$\int_A X_m \, d\nu = \mu(A) = \int_A X_{m-1} \, d\nu$$

so $\mathbb{E}\left[X_m \mid \mathcal{F}_{m-1}\right] = X_{m-1}$. □

Proof of the theorem

Proof. Suppose X_n is a nonnegative martingale with respect to a filtration $(\mathcal{F}_n)$, and suppose $X_n \to X$ ν-a.s. by Martingale Convergence Theorem.

To prove the Radon–Nikodym derivative, assume without loss of generality that μ is a probability measure. Define:

$$\rho = \frac{\mu + \nu}{2}, \quad \rho_n = \frac{\mu_n + \nu_n}{2},$$

where μ_n and ν_n are the restrictions to $\mathcal{F}_n$. Let:

$$Y_n = \frac{d\mu_n}{d\rho_n}, \quad Z_n = \frac{d\nu_n}{d\rho_n}, \quad \text{so that } Y_n + Z_n = 2.$$

Then $Y_n, Z_n \geq 0$ are bounded martingales, and Exercise A.4.6 shows they converge to:

$$Y = \frac{d\mu}{d\rho}, \quad Z = \frac{d\nu}{d\rho}. \tag{*}$$

Step 1. Prove $\mu(A) = \int_A Y \, d\rho$.
From the definition of Y_n,

$$\mu(A) = \int_A Y_n \, d\rho \to \int_A Y \, d\rho,$$

by bounded convergence. Hence:

$$\mu(A) = \int_A Y \, d\rho \quad \text{for all } A \in \bigcup_n \mathcal{F}_n.$$

The π-λ theorem extends this to all $A \in \mathcal{F}$, completing the first half of (*).
Step 2. Use the martingale ratio to express X.
From Exercises A.4.8 and A.4.9, we know:

$$X_n = \frac{Y_n}{Z_n} \to \frac{Y}{Z} =: X \quad \rho\text{-a.s.}$$

Note that $Y + Z = 2$, so $\rho(Y = 0, Z = 0) = 0$, and we avoid dividing by zero.

Define:

$$W = \frac{1}{Z} \cdot \mathbf{1}_{\{Z>0\}}.$$

Then:

$$1 = ZW + \mathbf{1}_{\{Z=0\}}.$$

Step 3. Show $\mu(A) = \int_A X \, d\nu$.

We compute:

$$\mu(A) = \int_A Y \, d\rho = \int_A YWZ \, d\rho + \int_A \mathbf{1}_{\{Z=0\}} Y \, d\rho \tag{a}$$

From (*), $d\nu = Z \, d\rho$, and:

$$YWZ = X \cdot \mathbf{1}_{\{Z>0\}}, \quad \text{so} \quad YWZ = X \quad \nu\text{-a.s.}$$

Hence:

$$\int_A YWZ \, d\rho = \int_A X \, d\nu. \tag{b}$$

Step 4. Show the second term in (a) is $\int_A \mathbf{1}_{\{X=\infty\}} \, d\mu$.

Note from (*) that $d\mu = Y \, d\rho$, and:

$$\{X = \infty\} = \{Z = 0\} \quad \mu\text{-a.s.}$$

So:

$$\int_A \mathbf{1}_{\{Z=0\}} Y \, d\rho = \int_A \mathbf{1}_{\{X=\infty\}} \, d\mu. \tag{c}$$

Conclusion. Combining (a), (b), and (c), we obtain:

$$\mu(A) = \int_A X \, d\nu + \int_A \mathbf{1}_{\{X=\infty\}} \, d\mu,$$

which proves that $X = \frac{d\mu}{d\nu}$, completing the Radon–Nikodym theorem. □

Example 11.10 (Approximating Radon-Nikodym Derivatives via Partitions)
Suppose $\mathcal{F}_n = \sigma(I_{k,n} : 0 \le k < K_n)$, where for each n, the sets $I_{k,n}$ form a partition of Ω, and each partition refines the previous one.

Assume that $\mu_n \ll \nu_n$, meaning that $\nu(I_{k,n}) = 0$ implies $\mu(I_{k,n}) = 0$. This ensures that the Radon-Nikodym derivative on each partition piece is well-defined.

Define the process:

$$X_n(\omega) = \frac{\mu(I_{k,n})}{\nu(I_{k,n})}, \quad \text{for } \omega \in I_{k,n}.$$

This defines a step function measurable with respect to $\mathcal{F}_n$, and $(X_n)_n$ forms a martingale. Importantly, this process approximates the Radon-Nikodym derivative $\frac{d\mu}{d\nu}$.

Concrete Example: Let $\Omega = [0, 1]$, and define the dyadic intervals:

$$I_{k,n} = \left[k2^{-n}, (k+1)2^{-n}\right), \quad 0 \le k < 2^n,$$

and let ν be Lebesgue measure. If μ is absolutely continuous with density f, then

$$X_n(\omega) = \frac{1}{2^{-n}} \int_{I_{k,n}} f(x)\, dx$$

is the average value of f over $I_{k,n}$, and as $n \to \infty$, $X_n(\omega) \to f(\omega)$ almost everywhere.

11.2.1 Kakutani dichotomy for infinite product measures

Let μ and ν be measures on the sequence space $(\mathbb{R}^n, \mathcal{R}^n)$ (a product σ-algebra) that make the coordiantes $\xi_n(\omega) = \omega_n$ independent. Define:

$$F_n(x) = \mu(\xi_n \le x), \quad \text{and} \quad G_n(x) = \nu(\xi_n \le x)$$

Assume $F_n \ll G_n$, and let $q_n = \frac{dF_n}{dG_n}$ be the Radon-Nikodym derivative. To avoid technical problems, assume $q_n > 0$, G_n-almost surely. Define

$$\mathcal{F}_n = \sigma(\xi_m : m \le n)$$

Let μ_n and ν_n denote the restrictions of μ and ν to $\mathcal{F}_n$, and define:

$$X_n = \frac{d\mu_n}{d\nu_n} = \prod_{m=1}^{n} q_m$$

By the Radon-Nikodym theorem, $X_n \to X$, ν-a.s.

Note 11.11. The sequence (X_n) forms a non-negative martingale under ν, so it converges almost surely and in L^1 to a limit $X = \frac{d\mu}{d\nu}$, when $\mu \ll \nu$.

Since we assumed $q_n > 0$, G_n-a.s., and the sum $\sum_{m=1}^{\infty} \log(q_m) > -\infty$ is a tail event, Kolmogorov's 0-1 law gives:

$$\nu(X = 0) \in \{0, 1\}$$

Hence, by the Radon-Nikodym theorem again:

$$\text{either } \mu \ll \nu \quad \text{or} \quad \mu \perp \nu$$

Note 11.12. This is Kakutani's dichotomy: for infinite product measures built from sequences of absolutely continuous components, either the entire product measure is absolutely continuous, or it is singular—there is no intermediate case.

Theorem 11.13 (Kakutani's criterion)
$\mu \ll \nu$ or $\mu \perp \nu$, according as

$$\prod_{m=1}^{\infty} \int \sqrt{q_m}\, dG_m > 0 \quad \text{or} \quad = 0,$$

where $q_m = \dfrac{\mathrm{d}F_m}{\mathrm{d}G_m}$.

Note 11.14. So:

- If the infinite product is strictly positive, then the measures are absolutely continuous: $\mu \ll \nu$.
- If the product is zero, the measures are mutually singular: $\mu \perp \nu$.

Proof. Jensen's inequality and Lebesgue decomposition imply

$$\left(\int \sqrt{q_m}\, dG_m \right)^2 \le \int q_m\, dG_m = \int dF_m = 1,$$

so the infinite product of these integrals is well-defined and less than or equal to 1. Let

$$X_n = \prod_{m \le n} q_m(\omega)$$

as before, and recall that $X_n \to X$ ν-a.s. If the infinite product is 0, then

$$\int X_n^{1/2}\, d\nu = \prod_{m=1}^{n} \int \sqrt{q_m}\, dG_m \to 0.$$

Fatou's lemma gives

$$\int X^{1/2}\, d\nu \le \liminf_{n\to\infty} \int X_n^{1/2}\, d\nu = 0,$$

so $X = 0$ ν-a.s., and by the Radon-Nikodym theorem, $\mu \perp \nu$.

To prove the converse, let $Y_n = X_n^{1/2}$. Since $\int q_m\, dG_m = 1$, if we use E to denote expectation with respect to ν, then

$$EY_m^2 = EX_m = 1.$$

Now compute the variance:

$$E(Y_{n+k} - Y_n)^2 = E\left((X_{n+k}^{1/2} + X_n^{1/2} - 2X_n^{1/2} X_{n+k}^{1/2})^2 \right) = 2\left(1 - \prod_{m=n+1}^{n+k} \int \sqrt{q_m}\, dG_m \right).$$

Now $|a-b| = |a^{1/2} - b^{1/2}| \cdot (a^{1/2} + b^{1/2})$, so by Cauchy-Schwarz and $(a+b)^2 \le 2a^2 + 2b^2$, we get:

$$\begin{aligned} E|X_{n+k} - X_n| = E(|Y_{n+k} - Y_n|(Y_{n+k} + Y_n)) &\le \left(E(Y_{n+k} - Y_n)^2 E(Y_{n+k} + Y_n)^2\right)^{1/2} \\ &\le \left(4E(Y_{n+k} - Y_n)^2\right)^{1/2}. \end{aligned}$$

From the last two equations, if the infinite product is > 0, then X_n converges to X in $L^1(\nu)$, so $\nu(X = 0) = 0$, and by $\nu(X = 0) \in \{0, 1\}$, $\mu \ll \nu$. The result now follows from the Radon-Nikodym theorem. □

Example 11.15 (Kakutani's criterion for Gaussian product measures)
Let

$$\mu = \bigotimes_{n=1}^{\infty} \mathcal{N}(0,1), \quad \nu = \bigotimes_{n=1}^{\infty} \mathcal{N}(0, \sigma_n^2)$$

We apply Kakutani's criterion to determine whether $\mu \ll \nu$ or $\mu \perp \nu$. Let $F_n = \mathcal{N}(0,1)$, $G_n = \mathcal{N}(0, \sigma_n^2)$, and define

$$q_n(x) = \frac{dF_n}{dG_n}(x) = \sqrt{\sigma_n} \cdot \exp\left(-\frac{x^2}{2} + \frac{x^2}{2\sigma_n^2}\right)$$

Then,

$$\begin{aligned} \int \sqrt{q_n(x)}\, dG_n(x) &= \int \sqrt{q_n(x)} \cdot g_n(x)\, dx \\ &= \int \sqrt{\frac{f_n(x)}{g_n(x)}} \cdot g_n(x)\, dx \\ &= \int \sqrt{f_n(x) g_n(x)}\, dx \\ &= \left(\frac{2\sigma_n}{1 + \sigma_n^2}\right)^{1/2} \end{aligned}$$

Hence, Kakutani's condition becomes

$$\prod_{n=1}^{\infty} \left(\frac{2\sigma_n}{1 + \sigma_n^2}\right)^{1/2}$$

Conclusion:

- If the product converges to a positive number, then $\mu \ll \nu$
- If the product converges to zero, then $\mu \perp \nu$

11.3 Doob's inequality, Convergence in $L^p, p > 1$

Theorem 11.16. If N is a stopping time and X_n is a supermartingale, then $X_{N \wedge n} = X_{\min(N,n)}$ is a supermartingale.

We prove a consequence of the above.

Theorem 11.17 (Optional Stopping Theorem, Bounded Case)**.** If (X_n) is a submartingale and N is a stopping time with $\mathbb{P}(N \leq k) = 1$, then

$$\mathbb{E}[X_0] \leq \mathbb{E}[X_N] \leq \mathbb{E}[X_k].$$

Remark 11.18. Let S_n be a simple random walk starting at $S_0 = 1$, and let $N = \inf\{n : S_n = 0\}$. Then $\mathbb{E}[S_0] = 1 > 0 = \mathbb{E}[S_N]$, so the inequality $\mathbb{E}[X_0] \leq \mathbb{E}[X_N]$ fails if N is unbounded.

Proof. We know that if (X_n) is a submartingale and N is a stopping time, then the stopped process $X_{N \wedge n}$ is also a submartingale. So:

$$\mathbb{E}[X_0] = \mathbb{E}[X_{N \wedge 0}] \leq \mathbb{E}[X_{N \wedge k}] = \mathbb{E}[X_N],$$

giving the left inequality.

Note 11.19. Stopping a submartingale early does not increase expected value. Since $N \wedge 0 = 0$, and $N \wedge k = N$ under the assumption $N \leq k$, we interpolate from time 0 to time N.

To prove the right inequality, define a predictable process:

$$K_n = \mathbf{1}_{\{N<n\}} = \mathbf{1}_{\{N \leq n-1\}}.$$

Then by corollary 10.29, the process $(K \cdot X)_n := \sum_{i=1}^{n} K_i(X_i - X_{i-1}) = X_n - X_{N \wedge n}$ is a submartingale. Hence,

$$\mathbb{E}[X_k - X_N] = \mathbb{E}[(K \cdot X)_k] \geq \mathbb{E}[(K \cdot X)_0] = 0,$$

which implies

$$\mathbb{E}[X_N] \leq \mathbb{E}[X_k].$$

□

Theorem 11.20 (Doob's inequality)

Let (X_m) be a submartingale, define

$$\overline{X}_n = \max_{0 \leq m \leq n} X_m^+, \quad \text{and let } A = \{\overline{X}_n \geq \lambda\}, \quad \lambda > 0.$$

Then

$$\lambda\, \mathbb{P}(A) \leq \mathbb{E}[X_n \mathbf{1}_A] \leq \mathbb{E}[X_n^+].$$

Proof. Let

$$N = \inf\{m : X_m \geq \lambda \text{ or } m = n\}$$

be the first time the process reaches or exceeds λ, or hits time n.

On the event $A = \{\overline{X}_n \geq \lambda\}$, we know $X_N \geq \lambda$, so:

$$\lambda\, \mathbb{P}(A) \leq \mathbb{E}[X_N \mathbf{1}_A] \leq \mathbb{E}[X_n \mathbf{1}_A].$$

The second inequality follows from the theorme above, which gives $\mathbb{E}[X_N] \leq \mathbb{E}[X_n]$ for submartingales and bounded stopping times, and noting that $X_N = X_n$ on the complement A^c, so the expectations match appropriately.

Also, since $X_n \mathbf{1}_A \leq X_n^+$, the final bound follows:

$$\mathbb{E}[X_n \mathbf{1}_A] \leq \mathbb{E}[X_n^+].$$

□

Example 11.21 (Rnadom walks)
Let $S_n = \xi_1 + \cdots + \xi_n$, where the ξ_m are independent and satisfy

$$\mathbb{E}[\xi_m] = 0, \quad \sigma_m^2 = \mathbb{E}[\xi_m^2] < \infty.$$

Then (S_n) is a martingale, and we know the process $X_n = S_n^2$ is a submartingale. Now apply Doob's inequality to $X_n = S_n^2$. Let $\lambda = x^2$, then

$$\mathbb{P}\left(\max_{1 \leq m \leq n} |S_m| \geq x\right) = \mathbb{P}\left(\max_{1 \leq m \leq n} S_m^2 \geq x^2\right) \leq \frac{\mathbb{E}[S_n^2]}{x^2} = x^{-2} \operatorname{Var}(S_n).$$

Integrating in the above example, gives the next result.

Theorem 11.22 (L^p maximal inequality)
If (X_n) is a submartingale, then for $1 < p < \infty$,

$$\mathbb{E}[\overline{X}_n^p] \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}[(X_n^+)^p],$$

where $\overline{X}_n = \max_{0 \leq m \leq n} X_m^+$ is the maximal process.

Consequently, if (Y_n) is a martingale and

$$Y_n^* = \max_{0 \leq m \leq n} |Y_m|,$$

then

$$\mathbb{E}[|Y_n^*|^p] \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}[|Y_n|^p].$$

Proof. We begin by proving the inequality

$$\mathbb{E}[\overline{X}_n^p] \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}[(X_n^+)^p]$$

for a submartingale (X_n), where $1 < p < \infty$. To avoid technical issues, work with the truncated maximum $\overline{X}_n \wedge M$, which will be sent to ∞ at the end.

Since $\{\overline{X}_n \wedge M \geq \lambda\} = \{\overline{X}_n \geq \lambda\} \cap \{\overline{X}_n \leq M\}$, and Doob's maximal inequality gives:

$$\mathbb{P}(\overline{X}_n \geq \lambda) \leq \frac{\mathbb{E}[X_n^+]}{\lambda},$$

we use integration by parts and Fubini's theorem to compute:

$$\mathbb{E}[(\overline{X}_n \wedge M)^p] = \int_0^\infty p\lambda^{p-1}\mathbb{P}(\overline{X}_n \wedge M \geq \lambda)\, d\lambda \leq \int_0^\infty p\lambda^{p-1}\left(\lambda^{-1}\int_{\overline{X}_n \wedge M \geq \lambda} X_n^+ \, d\mathbb{P}\right) d\lambda.$$

Now reverse the order of integration:

$$= \int X_n^+ \int_0^{\overline{X}_n \wedge M} p\lambda^{p-2}\, d\lambda\, d\mathbb{P} = \frac{p}{p-1}\int X_n^+(\overline{X}_n \wedge M)^{p-1}\, d\mathbb{P}.$$

Letting $q = \frac{p}{p-1}$, the Hölder conjugate of p, apply Hölder's inequality:

$$\int X_n^+(\overline{X}_n \wedge M)^{p-1} \leq \left(\int (X_n^+)^p\right)^{1/p}\left(\int (\overline{X}_n \wedge M)^p\right)^{1/q}.$$

Putting this together:

$$\mathbb{E}[(\overline{X}_n \wedge M)^p] \leq \frac{p}{p-1}\left(\mathbb{E}[(X_n^+)^p]\right)^{1/p}\left(\mathbb{E}[(\overline{X}_n \wedge M)^p]\right)^{1/q}.$$

Divide both sides by $\left(\mathbb{E}[(\overline{X}_n \wedge M)^p]\right)^{1/q}$ (which is finite), then raise both sides to the p-th power:

$$\mathbb{E}[(\overline{X}_n \wedge M)^p] \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}[(X_n^+)^p].$$

Finally, apply the Monotone Convergence Theorem as $M \to \infty$ to obtain the desired bound:

$$\mathbb{E}[\overline{X}_n^p] \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}[(X_n^+)^p].$$

□

Example 11.23 (There is no L^1 maximal inequality)

Consider the simple symmetric random walk (S_n) with $S_0 = 1$, and define the stopping time

$$N = \inf\{n : S_n = 0\}, \quad \text{and let} \quad X_n = S_{N\wedge n}.$$

Then X_n is the random walk stopped when it hits zero. By Optional Stopping theorem, we have

$$\mathbb{E}[X_n] = \mathbb{E}[S_{N\wedge n}] = \mathbb{E}[S_0] = 1, \quad \text{for all } n.$$

Now consider the maximum of the stopped process. The tail bound for the maximum:

$$\mathbb{P}\left(\max_m X_m \geq M\right) = \frac{1}{M}.$$

Hence,

$$\mathbb{E}\left[\max_m X_m\right] = \sum_{M=1}^\infty \mathbb{P}\left(\max_m X_m \geq M\right) = \sum_{M=1}^\infty \frac{1}{M} = \infty.$$

By the Monotone Convergence Theorem,

$$\mathbb{E}\left[\max_{m\leq n} X_m\right] \uparrow \infty \quad \text{as } n \to \infty.$$

Theorem 11.24 (L^p convergence theorem)
Let (X_n) be a martingale such that

$$\sup_n \mathbb{E}[|X_n|^p] < \infty \quad \text{for some } p > 1.$$

Then $X_n \to X$ almost surely and in L^p, i.e.,

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

Proof. First, note that Doob's inequality (Theorem 4.4.4) implies

$$\mathbb{E}[\overline{X}_n^p] \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}[|X_n|^p],$$

so the family $(\overline{X}_n)$ is uniformly bounded in L^p.

Also,

$$\mathbb{E}[(X_n^+)^p] \leq (\mathbb{E}[|X_n|])^p \leq \mathbb{E}[|X_n|^p],$$

so the integrability assumption ensures that the martingale convergence theorem applies: $X_n \to X$ almost surely.

Now using monotone convergence:

$$\sup_n |X_n| \in L^p.$$

Since

$$|X_n - X|^p \leq (2 \sup_n |X_n|)^p,$$

and $2\sup_n |X_n| \in L^p$, the dominated convergence theorem yields:

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

□

To treat the next case for when $p = 2$ we need the following results.

Theorem 11.25 (Orthogonality of Martingale Increments)**.** Let (X_n) be a martingale with $\mathbb{E}[X_n^2] < \infty$ for all n. If $m \leq n$ and $Y \in \mathcal{F}_m$ with $\mathbb{E}[Y^2] < \infty$, then:

$$\mathbb{E}[(X_n - X_m)Y] = 0.$$

In particular, for $\ell < m < n$,

$$\mathbb{E}[(X_n - X_m)(X_m - X_\ell)] = 0.$$

Proof. By Cauchy–Schwarz, $\mathbb{E}[|(X_n - X_m)Y|] < \infty$. Using the tower property of conditional expectation and the martingale property:

$$\mathbb{E}[(X_n - X_m)Y] = \mathbb{E}\left[\mathbb{E}[(X_n - X_m)Y \mid \mathcal{F}_m]\right] = \mathbb{E}[Y\,\mathbb{E}[X_n - X_m \mid \mathcal{F}_m]] = \mathbb{E}[Y \cdot 0] = 0. \quad \square$$

□

Theorem 11.26 (Conditional Variance Formula)**.** If (X_n) is a martingale with $\mathbb{E}[X_n^2] < \infty$ for all n, then:

$$\mathbb{E}[(X_n - X_m)^2 \mid \mathcal{F}_m] = \mathbb{E}[X_n^2 \mid \mathcal{F}_m] - X_m^2.$$

Remark 11.27. This is the conditional version of the identity:

$$\mathbb{E}[(X - \mathbb{E}X)^2] = \mathbb{E}[X^2] - (\mathbb{E}X)^2.$$

Proof. Expand the square and apply linearity:

$$\mathbb{E}[(X_n - X_m)^2 \mid \mathcal{F}_m] = \mathbb{E}[X_n^2 - 2X_nX_m + X_m^2 \mid \mathcal{F}_m] = \mathbb{E}[X_n^2 \mid \mathcal{F}_m] - 2X_m\mathbb{E}[X_n \mid \mathcal{F}_m] + X_m^2.$$

Since (X_n) is a martingale:

$$\mathbb{E}[X_n \mid \mathcal{F}_m] = X_m,$$

so:

$$= \mathbb{E}[X_n^2 \mid \mathcal{F}_m] - 2X_m^2 + X_m^2 = \mathbb{E}[X_n^2 \mid \mathcal{F}_m] - X_m^2. \quad \square$$

$\square$

Example 11.28 (Branching process)

We revisit the Galton–Watson branching process (Z_n), where each individual produces offspring independently according to a fixed distribution. Suppose:

$$\mu = \mathbb{E}[\xi_i^m] > 1, \quad \sigma^2 = \mathrm{Var}(\xi_i^m) < \infty.$$

Define the scaled process:

$$X_n = \frac{Z_n}{\mu^n}.$$

Applying the Conditional Variance Formula with $m = n-1$, we obtain:

$$\mathbb{E}[X_n^2 \mid \mathcal{F}_{n-1}] = X_{n-1}^2 + \mathbb{E}[(X_n - X_{n-1})^2 \mid \mathcal{F}_{n-1}].$$

To compute the second term:

$$\mathbb{E}[(X_n - X_{n-1})^2 \mid \mathcal{F}_{n-1}] = \mathbb{E}\left[\left(\frac{Z_n}{\mu^n} - \frac{Z_{n-1}}{\mu^{n-1}}\right)^2 \middle| \mathcal{F}_{n-1}\right] = \mu^{-2n}\mathbb{E}[(Z_n - \mu Z_{n-1})^2 \mid \mathcal{F}_{n-1}].$$

We use the following inequality (Chebyshev's inequality), if $a > 0$ then

$$\mathbb{P}(|X| \geq a \mid \mathcal{F}) \leq \frac{\mathbb{E}[X^2 \mid \mathcal{F}]}{a^2}$$

So, if $Z_{n-1} = k$, then:

$$\mathbb{E}[(Z_n - \mu Z_{n-1})^2 \mid \mathcal{F}_{n-1}] = \mathbb{E}\left[\left(\sum_{i=1}^{k}(\xi_i - \mu)\right)^2 \middle| \mathcal{F}_{n-1}\right] = k\sigma^2 = Z_{n-1}\sigma^2.$$

Putting this together:

$$\mathbb{E}[X_n^2 \mid \mathcal{F}_{n-1}] = X_{n-1}^2 + \frac{Z_{n-1}\sigma^2}{\mu^{2n}} = X_{n-1}^2 + \frac{\sigma^2}{\mu^{n+1}}.$$

Taking expectations yields the recurrence:

$$\mathbb{E}[X_n^2] = \mathbb{E}[X_{n-1}^2] + \frac{\sigma^2}{\mu^{n+1}}.$$

Since $\mathbb{E}[X_0^2] = 1$, iterating gives:

$$\mathbb{E}[X_n^2] = 1 + \sigma^2 \sum_{k=2}^{n+1} \mu^{-k}.$$

Conclusion. The series converges, so $\sup_n \mathbb{E}[X_n^2] < \infty$, and hence (X_n) converges in L^2 to some limit X, with:

$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n] = 1, \quad \text{so } X \neq 0.$$

From Exercise 4.3.11, it follows:

$$\{X > 0\} = \{Z_n > 0 \text{ for all } n\},$$

which expresses survival of the branching process.

Intuition. This example uses martingale techniques to analyse a classical branching process. The scaled process ($X_n = Z_n/\mu^n$) is a martingale and converges in L^2 because the variance of its increments decays rapidly. The non-zero limit indicates positive survival probability—an essential insight in population modelling and extinction theory.

Appendix

Distribution	**PMF**	**CDF**
Bernoulli(p)	$\mathbb{P}(X=1)=p,\ \mathbb{P}(X=0)=1-p$	$F(x)=\begin{cases}0 & x<0\\ 1-p & 0\le x<1\\ 1 & x\ge 1\end{cases}$
Binomial(n,p)	$\mathbb{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$	$\sum_{i=0}^{k}\binom{n}{i}p^i(1-p)^{n-i}$
Discrete Uniform(n)	$\mathbb{P}(X=k)=\frac{1}{n},\ k=1,\dots,n$	$F(k)=\frac{k}{n}$
Geometric(p)	$\mathbb{P}(X=k)=p(1-p)^{k-1},\ k\ge 1$	$F(k)=1-(1-p)^k$
Poisson(λ)	$\mathbb{P}(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}$	$\sum_{i=0}^{k}\frac{\lambda^i e^{-\lambda}}{i!}$

Table 1: PMFs and CDFs of common discrete distributions

Distribution	**PDF**	**CDF**
Normal(μ,σ^2)	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\Phi\left(\frac{x-\mu}{\sigma}\right)$
Exponential(λ)	$f(x)=\lambda e^{-\lambda x},\ x\ge 0$	$F(x)=1-e^{-\lambda x}$
Uniform(a,b)	$f(x)=\frac{1}{b-a},\ x\in[a,b]$	$F(x)=\begin{cases}0 & x<a\\ \frac{x-a}{b-a} & a\le x\le b\\ 1 & x>b\end{cases}$

Table 2: PDFs and CDFs of common continuous distributions

Distribution	**E[X]**	**Var(X)**	**Char. func**	**Support**
Bernoulli(p)	p	$p(1-p)$	$1-p+pe^{it}$	$\{0,1\}$
Binomial(n,p)	np	$np(1-p)$	$(1-p+pe^{it})^n$	$\{0,\dots,n\}$
Discrete Uniform(n)	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$\frac{1}{n}\sum e^{itk}$	$\{1,\dots,n\}$
Geometric(p)	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^{it}}{1-(1-p)e^{it}}$	$\mathbb{N}^+$
Poisson(λ)	λ	λ	$e^{\lambda(e^{it}-1)}$	$\mathbb{N}_0$
Normal(μ,σ^2)	μ	σ^2	$e^{i\mu t-\frac{1}{2}\sigma^2t^2}$	$\mathbb{R}$
Exponential(λ)	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-it}$	$[0,\infty)$
Uniform(a,b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{itb}-e^{ita}}{it(b-a)}$	$[a,b]$

Table 3: Expectation, Variance, Characteristic Function, and Support

References

[DD19] R. Durrett and R. Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.