GRADUATE PROBABILITY

NIKOS ZYGOURAS

ABSTRACT. These are notes on the Graduate Probability course taught at Warwick during 2021-2024. These are really meant to be notes, in the sense that they are not polished and may often be casually written. Students should also refer to the suggested textbooks for more details, if needed, and also for further examples and exercises. Hopefully, in the course of time these notes will take more shape. In the meanwhile, read these carefully and pick any typos and mistakes (and let me know).

CONTENTS

Date: January 22, 2026.

1. Measure Theory

1.1. The basics. We start with the very basics...

Definition 1.1. *A finitely additive probability is a non-negative set function such that*

1. $\mathrm{P}(A) \geq 0$ *for all* $A \in \mathcal{B}$ *for some class of set* $\mathcal{B}$ *of a set* Ω*,*
2. $\mathrm{P}(\Omega) = 1$ *and* $\mathrm{P}(\emptyset) = 0$*,*
3. *for* $A, B \in \mathcal{B}$ *with* $A \cap B = \emptyset$*, then* $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$*.*

Note that condition (3) in the above definition implies that for $A \in \mathcal{B}$ *and* A^c *its complement, then* $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$*.*

Definition 1.2 (algebras and $\sigma-$algebras). *A family of sets* $\mathcal{B}$ *is called a field if:*

1. *for* $A, B \in \mathcal{B}$ *then* $A \cup B \in \mathcal{B}$*,*
2. *for* $A \in \mathcal{B}$ *then the its complement* $A^c \in \mathcal{B}$*,*
3. $\emptyset \in \mathcal{B}$*.*

A field $\mathcal{B}$ *is a* σ*-field if for any sequence* $(A_n)_{n\geq 1}$ *in* $\mathcal{B}$*, then* $\cup_{n\geq 1} A_n \in \mathcal{B}$ *and* $\cap_{n\geq 1} A_n \in \mathcal{B}$ *.*

Note: in some books, eg in [V] the terms *"field"* and *"σ-field"* are used instead of the terms *" algebra"* and *"σ-algebra"*.

Definition 1.3. *A set function* $\mathrm{P} \colon \mathcal{B} \to \mathbb{R}$ *is a countably additive probability if it is a finitely additive probability that, additionally, satisfies*

$$\mathrm{P}\big(\cup_{n\geq 1} A_n \big) = \sum_{n\geq 1} \mathrm{P}(A_n) \qquad \textit{for } (A_n)_{n\geq 1} \textit{ a family of pairwise disjoint sets}$$

Definition 1.4. *For an algebra* $\mathcal{B}$ *the* σ*-algebra generated by* $\mathcal{B}$*, denoted by* $\sigma(\mathcal{B})$ *is the smallest* σ*-algebra that contains* $\mathcal{B}$*.*

Definition 1.5. *A monotone class is an algebra, which is closed under monotone limits, i.e. if*

$$\textit{if} \quad \mathcal{B} \ni A_n \downarrow, \quad \textit{meaning } A_{n+1} \subset A_n, \quad \textit{then } \cap_n A_n \in \mathcal{B},$$

or

$$\textit{if} \quad \mathcal{B} \ni A_n \uparrow, \quad \textit{meaning } A_n \subset A_{n+1}, \quad \textit{then } \cup_n A_n \in \mathcal{B}.$$

We can now state the first (but still basic) proposition:

Proposition 1.6. *A finitely additive probability measure* $\mathrm{P}(\cdot)$ *defined on a* σ*-algebra* $\mathcal{B}$ *is countable additive if an only if*

$$\mathrm{P}(A) = \lim_{n\to\infty} \mathrm{P}(A_n),$$

for any sequence of monotone sets A_n *and* A *defined as*

$$A := \lim_{n\to\infty} A_n := \begin{cases} \cap_{n\geq 1} A_n, & \textit{if } A_n \downarrow, \\ \\ \cup_{n\geq 1} A_n, & \textit{if } A_n \uparrow, \end{cases}$$

Proof. (1) Assume countable additivity and an increasing sequence of set $(A_n)_{n\geq 1}$. Then

$$\begin{aligned} \mathrm{P}\big(\cup_n A_n\big) &= \mathrm{P}\big(\cup_n \{A_n \setminus A_{n-1}\}\big) \\ &= \sum_n \mathrm{P}\big(A_n \setminus A_{n-1}\big) && \text{[by countable additivity]}, \\ &= \lim_{N\to\infty} \sum_n^N \mathrm{P}\big(A_n \setminus A_{n-1}\big) \\ &= \lim_{N\to\infty} \mathrm{P}(A_n) && \text{[by simple additivity]}. \end{aligned}$$

(2) Assume monotonicity, that is, for any $A_n \uparrow$ it holds $\mathrm{P}\big(\cup_n A_n\big) = \lim_{N\to\infty}\mathrm{P}(A_n)$. Then for a sequence $(A_n)_{n\geq 1}$ of mutually disjoint sets we have:

$$\begin{aligned} \mathrm{P}\big(\cup_n A_n\big) &= \mathrm{P}\big(\cup_{N\geq 1}\cup_{n\geq 1} A_n\big) \\ &= \lim_{N\to\infty} \mathrm{P}(\cup_{n=1}^N A_n) && \text{[by monotonicity]} \\ &= \lim_{N\to\infty} \sum_{n=1}^N \mathrm{P}(A_n) \\ &= \sum_{n=1}^\infty \mathrm{P}(A_n). \end{aligned}$$

□

Exercise 1. *Show that for* P *to be a countably additive probability, it suffices to have that* $\mathrm{P}(A_n)\downarrow 0$ *for any sequence of sets* $A_n \downarrow \emptyset$.

Proposition 1.7. *Given any family* $\mathcal{F}$ *of subsets of* Ω*, there is a unique* σ*-field* $\sigma(\mathcal{F})$*, which is the smallest* σ*-field containing* $\mathcal{F}$.

Proof. Let

$$\mathcal{A} := \Big\{\Sigma\colon \Sigma\supset\mathcal{F} \text{ and is a } \sigma\text{-field}\Big\}$$

We then have:

- $\mathcal{A}\neq\emptyset$. This is because $\mathcal{A}$ contains the power set $\mathcal{P}(\Omega)$.
- define $\sigma(\mathcal{F}) := \bigcap_{\Sigma\in\mathcal{A}}\Sigma$. This is a σ-algebra because if $(A_n)_{n\geq 1}\subset\sigma(\mathcal{F})$, then $A_n\in\Sigma$ for all $n\geq 1$ and $\Sigma\in\mathcal{A}$, which means that $\cup_{n\geq 1}A_n\in\Sigma$ for all $\Sigma\in\mathcal{A}$, which then means that $\cup_{n\geq 1}A_n\in\cap_{\Sigma\in\mathcal{A}}\Sigma=\sigma(\mathcal{F})$.
- $\sigma(\mathcal{F})$ is the smallest σ field that contains $\mathcal{F}$ since it is obtained as the intersection of all σ-fields containing $\mathcal{F}$.

□

So far we have discussed about (countably-additive) probability measures but we haven't proved that such objects exist. It is now the time to do so and we will construct the Lebesque probability measure. The difficulty in constructing countably-additive probability measures is to show that they are well defined on the corresponding σ-algebra. To achieve this task we will be making use of the Caratheodory theorem:

Theorem 1.8. *Any countably additive probability measure on a algebra* $\mathcal{F}$ *has a* **unique** *extension as a countably additive probability measure on the* σ*-algebra* $\sigma(\mathcal{F})$.

A proof of this theorem can be found in all classical books of measure theory or probability, see for example [V], Theorem 1.1.

Let us now give a recipe for the construction of (countably-additive) probability measures:

Step 1. Consider any non-decreasing, right continuous function $F\colon \mathbb{R} \to [0,1]$, such that $F(-\infty) = 0$ and $F(\infty) = 1$

Step 2. Define the **Borel σ-algebra** as the σ-algebra which is generated by all finite unions of intervals $(a,b]$, with $-\infty \le a < b \le \infty$. Check that this is a field.

Step 3. Define $\mathrm{P}\big((a,b]\big) := F(b) - F(a)$.

The above recipe gives all countable additive probability measures:

Theorem 1.9. *For every non-decreasing, right-continuous, real function F with $F(-\infty) = 0$ and $F(\infty) = 1$ there exists a unique countably-additive probability measure on the Borel σ-algebra $\mathcal{B}$. Conversely, every countably-additive probability measure on $\mathbb{R}$ comes from such a function.*

Proof. Let us prove the first direction. It will suffice to show that for any sequence of sets A_n in the field of finite unions of left-open, right-closed interval (the algebra that generates the Borel)

$$\text{if} \quad A_n \downarrow \emptyset \quad \text{then it holds that } \mathrm{P}(A_n) \downarrow 0. \tag{1.1}$$

Then we will be having a countably additive probability measure on an algebra and we can use Caratheordy's theorem to extend it as a countably additive probability on the σ-algebra generated by this algebra.

Assume that (1.1) is not valid. Then there will be a sequence $A_n = \cup_{j=1}^{k_n}(a_j,b_j]$ such that $\mathrm{P}(A_n) \ge \delta$ for all $n \ge 1$ and a $\delta > 0$.

We can assume that all $A_n \subset [-L,L]$ for some L large, since for large L the probability outside $[-L,L]$ can be made smaller than $\delta/2$ (by the right-continuity of F and that $F(-\infty) = 0$ and $F(+\infty) = 1$). So we could consider the sets $A'_n := A_n \cap [-L,L]$, instead.

We can also assume that (A_n) is a decreasing family of sets (why?). By right-continuity, we can choose a'_j such that $(a_j,b_j] \subset (a'_n,b_n]$ and $\mathrm{P}\big((a_j,b_j]\big) \approx \mathrm{P}\big((a'_j,b_j]\big)$. You can formalise this, if you want as

$$\mathrm{P}\big(\cup_{j=1}^{k_n}(a_j,b_j]\big) - \mathrm{P}\big(\cup_{j=1}^{k_n}(a'_j,b_j]\big) \le \frac{1}{100}\delta,$$

with the factor $1/100$ being rather arbitrary. The main thing is that $\mathrm{P}\big(\cup_{j=1}^{k_n}(a'_j,b_j]\big)$ remains uniformly bounded away from 0, say larger that $\delta' > 0$..

Now enlarge a bit the set A_n so that it is closed, to the set $A'_n := \cup_{j=1}^{k_n}[a'_j,b_j]$. Then all A'_n are compact (because they are closed and bounded – this is why we restricted in large intervals $[-L,L]$). We also have that $\mathrm{P}(A'_n) \ge \delta'$, which means that all the sets A'_n are nonempty. But then we have a nonempty sequence of compact sets, which (by assumption) $A'_n \subset A_n \downarrow \emptyset$. This is a contradiction by basic topological facts.

Let us prove, now, that opposite directions. This is easier, since we can define

$$F(x) := \mathrm{P}\big((-\infty,x]\big),$$

and the desired properties of F follow from the properties of the probability measure. For the right-continuity you would need to use Theorem 1.6. The other properties are more trivial. ☐

1.2. DYNKIN'S $\pi - \lambda$ THEOREM.

We will now present a very interesting theorem. Its interest lies on the demonstration of how useful abstraction can be in proving statements that might seem to complicated by checking case-by-case. We start with a definition.

Definition 1.10 (Dynkin system or monotone class.). *A Dynkin system or monotone class is a family of subsets of Ω such that*

- $\Omega \in \mathcal{D}$,
- *if* $A, B \in \mathcal{D}$ *with* $A \subset B$, *then* $B \setminus A \in \mathcal{D}$
- *if* $A_n \uparrow$ *then* $\cup_n A_n \in \mathcal{D}$.

Definition 1.11. *A π system is any collection $\mathcal{D}$ of subsets of Ω, which is closed under finite intersections, i.e. if $A, B \in \mathcal{D}$ then $A \cap B \in \mathcal{D}$.*

Dynkin's theorem is the following:

Theorem 1.12 **($\pi - \lambda$ theorem).** *If $\mathcal{P}$ is a π-system and $\mathcal{D}$ is a Dynkin class such that $\mathcal{P} \subset \mathcal{D}$, then $\sigma(\mathcal{P}) \subset \mathcal{D}$. In particular, a π-system, which is also a Dynkin class is a $\sigma-$algebra.*

For the proof of this theorem we refer to [D], Theorem (2.1) in the Appendix. We would like to demonstrate the use of this abstract theorem by proving the following:

Proposition 1.13. *The Lebesque measure on the Borels is the unique translation invariant measure, meaning the only measure such that $\lambda((a,b]) = b - a$.*

Proof. For simplicity and to stay within probability we will prove the statement on $[0,1]$, instead of $\mathbb{R}$. Let λ be the Lebesque measure and μ some other measure such that $\mu((a,b]) = b - a$. Let

$$\mathcal{D} : \big\{ A \in \mathcal{B} \colon \mu(A) = \lambda(A) \ \big\}.$$

We want to show that $\mathcal{D} = \mathcal{B}$, which would then imply that the now measure are identical as they would agree on all Borel sets. We will use the $\pi - \lambda$ theorem: Let $\mathcal{I}$ be the collection of all intervals of the form $(a,b], (a,b), [a,b), [a,b]$ for $-\infty \leq a < b \leq \infty$. We have that:

- $\mathcal{I}$ is closed under finite intersections,
- $\sigma(\mathcal{I}) = \mathcal{B}$,
- $\mathcal{I} \subset \mathcal{D}$.

The first two bullets are definitions and the last follows from the assumptions of the measures. Also $\mathcal{D}$ is a Dynkin system - this follows from the monotonicity of the measure Theorem 1.6. So we have that $\mathcal{I} \subset \mathcal{D}$, $\mathcal{I}$ is a π-system, $\mathcal{D}$ is a Dynkin class. Therefore, by the $\pi - \lambda$ Theorem, it follows that $\sigma(\mathcal{I}) \subset \mathcal{D}$. But also $\sigma(I)$ is the Borel σ-algebra and so we are done. ☐

Exercise 2. *Show that one cannot construct a Lebesque measure on the rationals $\mathbb{Q}$. That is a measure such that* $\mathrm{P}([a,b] \cap \mathbb{Q}) = b - a$. **Hint:** *actually this is not an exercise on $\pi - \lambda$ theorem, rather on the basic properties of measures.*

1.3. INTEGRATION AND MODES OF CONVERGENCE.

In this section we will define the Lebesque integration, starting from measurable functions and random variables, state the definition of convergence and expose the fundamental integration limit theorems that we will be freely using. So this is a basic but very fundamental section. Let us start with the definition:

Definition 1.14 **(Measurable functions & random variables).** *Let (Ω, Σ) be a measurable space, that is a space (set) Ω with a σ-algebra Σ. A measurable function or random variable on (Ω, Σ) is*

$$\begin{aligned} f \colon \Omega \to \mathbb{R}, \qquad & \textit{in measure theory or,} \\ X \colon \Omega \to \mathbb{R}, \qquad & \textit{in probability,} \end{aligned}$$

is a function such that $f^{-1}(B) \in \Sigma$ or $X^{-1}(B) \in \Sigma$ for all $B \in \mathcal{B}(\mathbb{R})$ Borel sets on $\mathbb{R}$.

Let us look at some basic examples, which will be the building blocks towards building more complicated ones. Most theorems in measure theory will start by checking their validity on these building blocks and then derive the statement for general functions via a limiting procedure. The basic measurable functions are:

- **Indicator functions.** These are the functions / variables:

$$\mathbb{1}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \ni A, \end{cases}$$

for any set $A \in \Sigma$.

- **Simple functions.** These are functions / variables of the form
$$f(\omega)=\sum_j c_j \mathbb{1}_{A_j}(\omega),$$
where the sum is finite, c_j real numbers and $A_j \in \Sigma$.

Exercise 3. (very important !) *Prove that any bounded measurable function is a uniform approximation of simple functions.*

We can now define the Lebesque integral.

Theorem 1.15 (Lebesque integration). *Let* $(\Omega, \Sigma, \mathrm{P})$ *be a probability space. That is, a measurable space* (Ω, Σ) *with a probability measure* P. *We define:*

$$1)\qquad \int \mathbb{1}_A(\omega)\mathrm{dP} := \mathrm{P}(A),$$

$$2)\qquad \int \sum_{\text{finite}} c_j \mathbb{1}_{A_j}(\omega)\,\mathrm{dP} = \sum_{\text{finite}} c_j \mathrm{P}(A_j).$$

Finally, for general measurable function f *(or random variable) we find a sequence of simple functions* $(f_n)_{n\geq 1}$ *such that*
$$\sup_\omega |f(\omega)-f_n(\omega)| \xrightarrow[n\to\infty]{} 0,$$
and then we define
$$\int f(\omega)\,\mathrm{dP} := \lim_n \int f_n(\omega)\,\mathrm{dP},$$
where the righ-hand side limit can be defined as a limit of Cauchy sequence.

1.4. MODES OF CONVERGENCE.

Definition 1.16 (almost sure convergence). *Let* $(\Omega, \Sigma, \mathrm{P})$ *be a probability space. A sequence of measurable functions* $(f_n)_{n\geq 1}$ *or random variables* $(X_n)_{n\geq 1}$ *is said to convergence almost surely (a.s.) to* f *and* X, *respectively if*
$$\mathrm{P}\big(\omega\colon f_n(\omega)\to f(\omega)\,\big)=1 \qquad \textit{or} \qquad \mathrm{P}\big(\omega\colon X_n(\omega)\to X(\omega)\,\big)=1.$$

A useful, variant formulation of the above is to say that for any $\varepsilon>0$:
$$\mathrm{P}\big(\omega\colon\, \big|f_n(\omega)-f(\omega)\big|>\varepsilon\,, i.o.\big)=0, \qquad \text{or} \qquad \mathrm{P}\big(\,\big|X_n(\omega)-X(\omega)\big|>\varepsilon\,, i.o.\big)=0$$
where i.o. stands for **"infinitely often"**. It will be useful to quantify the *infinitely often* notion in order to be able to compute. To this end, for a sequence of events (sets) $(A_n)_{n\geq 1}$ we will identify the event
$$(A_n)_{n\geq 1} \text{ happen } \textit{infinitely often} \text{ as the set } A=\cap_{n\geq 1}\cup_{m\geq n} A_n.$$
We will also denote the event $(A_n)_{n\geq 1}$ *happen infinitely often* by $\limsup_{n\to\infty} A_n$.

There is also the dual event of **"eventually always"**, which is $\cup_{n\geq 1}\cap_{m\geq n} A_m$ and will be denoted by $\liminf_n A_n$.

A very central Lemma that is the standard approach to proving a.s. convergence is the Borel-Cantelli. We state here the first Borel-Cantelli lemma, which is very elementary. Later we will also state the second Borel-Cantelli, which requires indepence.

Lemma 1.17 (First Borel-Cantelli lemma). *Let* $(A_n)_{n\geq 1}$ *be a sequence of sets (events) in a probability space* $(\Omega, \Sigma, \mathrm{P})$. *We have*
$$\textit{if} \quad \sum_{n\geq 1}\mathrm{P}(A_n)<\infty \quad \textit{then} \quad \mathrm{P}\big(A_n \textit{ i.o.}\big)=0.$$

Proof. We have

$$\begin{aligned}
\mathrm{P}\big(A_n \ i.o.\big) &= \mathrm{P}\big(\cap_{n\ge1}\cup_{m\ge n}A_m\big) \\
&= \lim_{n\to\infty}\mathrm{P}\big(\cup_{m\ge n}A_m\big) && \text{[by monotonicity]} \\
&\le \limsup_{n\to\infty}\sum_{m\ge n}\mathrm{P}(A_m) \\
&= 0 && \text{[by the convergence of the series].}
\end{aligned}$$

☐

Remark 1.18. The typical way we will be using it if we want to prove that $f_n \xrightarrow[n\to\infty]{a.s.} f$ will be to consider the events $A_n := \{\omega\colon |f_n(\omega)-f(\omega)|>\varepsilon\}$ and show that $\sum_{n\ge1}\mathrm{P}(A_n)<\infty$. However, most often this simple approach will not work (the naive series will not be summable) and we will need to devise smart tricks.

Definition 1.19 (Convergence in probability). *Let $(\Omega,\Sigma,\mathrm{P})$ be a probability space. A sequence of measurable functions $(f_n)_{n\ge1}$ or random variables $(X_n)_{n\ge1}$ is said to convergence in probability to f and X, respectively if for every $\varepsilon>0$,*

$$\mathrm{P}\big(\omega\colon\big|f_n(\omega)-f(\omega)\big|>\varepsilon\big)\xrightarrow[n\to\infty]{}0. \qquad \textit{or} \qquad \mathrm{P}\big(\big|X_n-X\big|>\varepsilon\big)\xrightarrow[n\to\infty]{}0.$$

a.s. convergence impliesconvergence in probability:

Proposition 1.20. *If $f_n \xrightarrow[n\to\infty]{a.s} f$ then $f_n \xrightarrow[n\to\infty]{\mathrm{P}} f$.*

Proof. We will play with the set theoretic formulation of *a.s.* convergence. For every $\varepsilon>0$ we have:

$$\begin{aligned}
\mathrm{P}\big(\omega\colon|f_n(\omega)-f(\omega)|>\varepsilon\big) &\le \mathrm{P}\big(\cup_{m\ge n}\big\{|f_n(\omega)-f(\omega)|>\varepsilon\big\}\big) \\
&\xrightarrow[n\to\infty]{}\mathrm{P}\big(\cap_{n\ge1}\cup_{m\ge n}\big\{|f_n(\omega)-f(\omega)|>\varepsilon\big\}\big) && \text{[by monotonicity]} \\
&= \mathrm{P}\big(|f_n(\omega)-f(\omega)|>\varepsilon,\ i.o.\big) && \text{[by definition of } i.o.\text{]} \\
&= 0 && \text{[by definition of } a.s. \text{ convergence].}
\end{aligned}$$

☐

Exercise 4. *Find an example of a sequence (f_n) which converges in probability but not a.s.*

Exercise 5. *Let a sequence $(f_n)_{n\ge1}$. Show that $f_n \xrightarrow[n\to\infty]{\mathrm{P}} f$ if and only if every subsequence f_{n_j} has a further subsequence $f_{n_{j_k}}$ which converges a.s..*

Exercise 6. *If a sequence of random variables $(X_n)_{n\ge1}$ converges to a random variable X in probability and f is a continuous function, show that $f(X_n)$ converges to $f(X)$ in probability.*

1.5. INTEGRAL CONVERGENCE THEOREMS.

In this section we expose the fundamental convergence theorems for integral that we will be freely using.

Theorem 1.21 (Bounded Convergence Theorem). *Let $(f_n)_{n\ge1}$ be a sequence of measurable functions on a probability space $(\Omega,\Sigma,\mathrm{P})$ such that for all n, it holds that $|f_n|\le M$ for some finite $M>0$. We then have:*

$$\textit{if} \qquad f_n \xrightarrow[n\to\infty]{\mathrm{P}} f \qquad \textit{then} \qquad \int f_n\,\mathrm{dP}\xrightarrow[n\to\infty]{}\int f\,\mathrm{dP}\ .$$

Proof. For any $\varepsilon > 0$, have

$$\begin{aligned}\Big|\int f_n \,\mathrm{dP} - \int f \,\mathrm{dP}\Big| &\le \int \big|\, f_n - f \,\big| \,\mathrm{dP} \\ &= \int_{\{|\, f_n - f| \le \varepsilon\}} \big|\, f_n - f \,\big| \,\mathrm{dP} + M\mathrm{P}\Big(|\, f_n - f \,| > \varepsilon\Big) \\ &\le \varepsilon + M\mathrm{P}\Big(\Big| f_n - f \Big| > \varepsilon\Big),\end{aligned}$$

and the second term converges to 0 as $n \to \infty$ by the assumption of convergence in probability. ☐

Theorem 1.22 (Fatou's lemma). *Let $f_n \ge 0$ be a sequence of nonnegative, measurable functions on a probability space $(\Omega, \Sigma, \mathrm{P})$, such that $f_n \xrightarrow[n\to\infty]{\mathrm{P}} f$. Then*

$$\int f \mathrm{dP} \le \liminf_{n\to\infty} \int f_n \,\mathrm{dP}.$$

Proof. We want to use the bounded convergence theorem. To this end, consider a bounded g such that $0 \le g \le f$, for example $g := f \mathbb{1}_{f \le M}$ for $M > 0$. Consider, also, the sequence $h_n := f_n \wedge g$. This sequence is uniformly bounded. By the bounded convergence theorem we have

$$\int g \,\mathrm{dP} = \int f \wedge g \,\mathrm{dP} = \lim_{n\to\infty} \int f_n \wedge g \,\mathrm{dP} \le \liminf_{n\to\infty} \int f_n \,\mathrm{dP}$$

Since the above inequality is valid for arbitrary bounded $g \le f$, eg. $g = f \mathbb{1}_{\{f \le M\}}$, it will also hold that

$$\int f \mathrm{dP} \le \liminf_{n\to\infty} \int f_n \,\mathrm{dP},$$

(e.g. let $M \to \infty$). ☐

Theorem 1.23 (Dominated convergence). *Let $(f_n)_{n\ge}$ be a sequence of measurable functions on a probability space $(\Omega, \Sigma, \mathrm{P})$ such that $f_n \xrightarrow[n\to\infty]{\mathrm{P}} f$. Moreover, assume that for a.e. ω it holds that $|f_n(\omega)| \le g(\omega)$ for $g \in L^1(\mathrm{P})$. Then*

$$\lim_{n\to\infty} \int f_n \,\mathrm{dP} = \int f \,\mathrm{dP}.$$

Proof. We will reduce this to Fatou's lemma. We first note that $f_n + g \ge 0$ and $-f_n + g \ge 0$ and then by Fatou's we have that

$$\int (f + g) \,\mathrm{dP} \le \liminf_{n\to\infty} \int (f_n + g) \,\mathrm{dP},$$

which implies, since $g \in L^1(\mathrm{P})$ that

$$\int f \,\mathrm{dP} \le \liminf_{n\to\infty} \int f_n \,\mathrm{dP}.$$

In a similar way we have:

$$\int (-f + g) \,\mathrm{dP} \le \liminf_{n\to\infty} \int (-f_n + g) \,\mathrm{dP} \qquad \Longrightarrow \qquad \int f \ge \limsup_{n\to\infty} \int f_n \,\mathrm{dP},$$

and the result follows. ☐

We close with a reminder

Theorem 1.24 (Jensen's inequality). *If $\phi\colon \mathbb{R} \to \mathbb{R}$ is a convex function, then*

$$\phi\Big(\int f \mathrm{dP}\Big) \le \int \phi(f) \mathrm{dP}.$$

Exercise 7. *Use the fact (which you should check) that for real x it holds*

$$\frac{x^p}{p} = \sup_y \{xy - \frac{y^q}{q}\},$$

to show that for f, g nonnegative functions:

$$\int fg \, \mathrm{dP} \leq \Big(\int f^p \, \mathrm{dP} \Big)^{1/p} \Big(\int g^q \, \mathrm{dP} \Big)^{1/q},$$

for $\frac{1}{p} + \frac{1}{q} = 1$. **Hint:** *Write fg as $(\lambda f)(g/\lambda)$ for a real parameter λ and eventually optimise over λ.*

1.6. PRODUCT SPACES AND MEASURES.

The notion of product spaces and product measure that we will define here will axiomatise the notion of independence. We know from basic probability that two events A, B are independent if $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$. However, the notation used kind of hides quite a bit of structure. In fact the symbol P in the LHS and the symbols P in the RHS don't quite have the same meaning. The formally correct way is that the LHS is a short (and arguably more intuitive) way to write $\mathrm{P} \times \mathrm{P}\big((A \times \mathbb{R}) \cap (\mathbb{R} \times B)\big) = \mathrm{P} \times \mathrm{P}(A \times B)$. We will define all these in this section.

Definition 1.25 (Product σ-algebras). *Let (Ω_1, Σ_1) and (Ω_2, Σ_2) be two measurable spaces. We define the Cartesian product $\Omega := \Omega_1 \times \Omega_2$. The product σ-algebra $\Sigma_1 \times \Sigma_2$ is defined the $\sigma\Big(\cup_{\mathit{finite}} \, \mathit{rectangles} \Big) = \sigma\Big(\cup_{n=1}^{k} A_n \times B_n, : A_n \in \Sigma_1 \, , \, B_n \in \Sigma_2 \, , \, k \, \mathit{finite} \Big)$.*

Given two probability measures $\mathrm{P}_1, \mathrm{P}_2$ on (Ω_1, Σ_1) and (Ω_2, Σ_2), respectively, we will next define the **product measure** $\mathrm{P}_1 \times \mathrm{P}_2$ on $(\Omega_1 \times \Omega_2, \Sigma_1 \times \Sigma_2)$ as follows:

- for rectangles $A_1 \times A_2$ we define $\Big(\mathrm{P}_1 \times \mathrm{P}_2\Big)(A_1 \times A_2) := \mathrm{P}_1(A_1)\mathrm{P}_2(A_2)$.
- for $E = \cup_i A_i \times B_i$ a finite union of disjoint rectangles, we define $\Big(\mathrm{P}_1 \times \mathrm{P}_2\Big)(E) := \sum_i \mathrm{P}_1(A_i)\mathrm{P}_2(A_2)$.

Exercise 8. *Show that the definition in the second bullet does not depend on the representation of the set E as a union of disjoint rectangles.*

We next need to lift the definition of $\mathrm{P}_1 \times \mathrm{P}_2$ to the product σ-algebra. To do this we will use Caratheodory's theorem and so need to prove:

Lemma 1.26. *The product measure $\mathrm{P}_1 \times \mathrm{P}_2$ is a countably additive measure on the collection of all finite unions of rectangles.*

Proof. We need to show that if $(E_n)_{n \geq 1}$ is a decreasing sequence of sets in $\mathcal{F}$ with $\cap_n E_n = \emptyset$, then $\mathrm{P}_1 \times \mathrm{P}_2(E_n) \downarrow 0$. To this end, let us define *sections*: For any $\omega_2 \in \Omega_2$ we define

$$E_{\omega_2} := \big\{ \omega_1 \in \Omega_1 \colon (\omega_1, \omega_2) \in E \big\}.$$

This set belongs to Σ_1. We also have that $\omega_2 \to \mathrm{P}_1(E_{\omega_2})$ is a simple function, measurable with respect to Σ_2. As such, we can define the integral

$$\int_{\Omega_2} \mathrm{P}_1(E_{\omega_2}) \mathrm{P}_2(\mathrm{d}\omega_2), \tag{1.2}$$

which actually equals $\mathrm{P}_1 \times \mathrm{P}_2(E)$ (why ?) Similarly we define the sections E_{n,ω_2}. The assumption $E_n \downarrow \emptyset$ implies that $E_{n,\omega_2} \downarrow \emptyset$ for all $\omega_2 \in \Omega_2$ and so $\mathrm{P}_1(E_{n,\omega_2}) \downarrow 0$. By bounded convergence theorem, we then have that

$$\Big(\mathrm{P}_1 \times \mathrm{P}_2\Big)(E_n) = \int_{\Omega_2} \mathrm{P}_1(E_{n,\omega_2}) \, \mathrm{P}_2(\mathrm{d}\omega_2) \xrightarrow[n \to \infty]{} 0.$$

$\square$

Equation (1.2) can be extended for all measurable sets in the product σ-algebra, into a formula called **disintegration formula**:

Proposition 1.27. *$(\Omega_1, \Sigma_1, \mathrm{P}_1)$ and $(\Omega_2, \Sigma_2, \mathrm{P}_2)$ be two probability spaces and $(\Omega_1 \times \Omega_2, \Sigma_1 \times \Sigma_2, \mathrm{P}_1 \times \mathrm{P}_2)$ the product probability space. For any $E \in \Sigma_1 \times \Sigma_2$ we define the sections:*

$$E_{\omega_1} := \big\{\omega_2 \in \Omega_2 \colon (\omega_1, \omega_2) \in E\big\} \quad \text{and} \quad E_{\omega_2} := \big\{\omega_1 \in \Omega_1 \colon (\omega_1, \omega_2) \in E\big\}.$$

Then the functions $P_1(A_{\omega_2})$ and $P_2(A_{\omega_1})$ are measurable and

$$\Big(\mathrm{P}_1 \times \mathrm{P}_2\Big)(E) = \int_{\Omega_2} \mathrm{P}_1(E_{\omega_2})\, \mathrm{P}_2(\mathrm{d}\omega_2) = \int_{\Omega_1} \mathrm{P}_2(E_{\omega_1})\, \mathrm{P}_1(\mathrm{d}\omega_1). \tag{1.3}$$

Proof. The proof is another nice application of the $\pi - \lambda$ theorem. We let $\mathcal{D}$ be the collection of all sets for which (1.3) is true. We note that the field (i.e. closed under finite intersections) $\mathcal{F}$ of all finite unions of rectangles belongs to $\mathcal{D}$ by (1.2), i.e. $\mathcal{F}$ is a π-system.

Let us check that $\mathcal{D}$ is a monotone class:

- $\Omega := \Omega_1 \times \Omega_2 \ni \mathcal{D}$. This is becauce
$$\Big(\mathrm{P}_1 \times \mathrm{P}_2\Big)(\Omega) = 1 = \int_{\Omega_2} \mathrm{P}_1(\Omega_{\omega_2})\mathrm{P}_2(\mathrm{d}\omega_2),$$
since for every ω_2 we have that $\Omega_{\omega_2} = \Omega_1$ and $\mathrm{P}_1(\Omega_1) = 1$ as well as $\int_{\Omega_2} 1\mathrm{P}_2(\mathrm{d}\omega_2) = 1$.
- if $A, B \in \mathcal{D}$ with $A \subset B$ we have that $B \setminus A \in \mathcal{D}$ (**exercise**: check this.)
- If $A_n \in \mathcal{D}$ increasing, then $\cup A_n \in \mathcal{D}$. **Exercise**: check this using the monotonicity of measure and the bounded convergence theorem.

Thus, by the $\pi - \lambda$ theorem we have that $\sigma(\mathcal{F}) \subset \mathcal{D}$ but since $\sigma(\mathcal{F})$ is the product σ-algebra then (1.3) holds for all measurable sets E.

We have checked (1.3) before actually checking the measurability of the functions involved in the integrals. To do so we follow the same steps as in the above paragraph. We need to let $\mathcal{D}'$ be the collection of all $E \in \Omega_1 \times \Omega_2$ for which $\omega_1 \to \mathrm{P}_2(E_{\omega_1})$ and $\omega_2 \to \mathrm{P}_1(E_{\omega_2})$ are measurable functions. We know that this class contains all finite unions of rectangles and we need to check that $\mathcal{D}$ is a monotone class. For the latter, we just need to know that limits of measurable functions (in particular monotone limits) is a measurable function, which is a basic measure-theoretic principle. □

The above proposition can be extended to the **Fubini Theorem**, which says the under certain assumption (integrability or positivity) on a function $f(\omega_1, \omega_2)$ the order of integrations does not matter. In particular, we have that if $f(\omega_1, \omega_2)$ is an integrable function with respect to $\mathrm{P}_1 \times \mathrm{P}_2$ or just nonnegative, then

$$\begin{aligned}\int_{\Omega_1 \times \Omega_2} f(\omega_1, \mathrm{d}\omega_2)\Big(\mathrm{P}_1 \times \mathrm{P}_2\Big)(\mathrm{d}\omega_1, \omega_2) &= \int_{\Omega_1} \Big(\int_{\Omega_2} f(\omega_1, \omega_2)\, \mathrm{P}_2(\mathrm{d}\omega_2)\Big)\mathrm{P}_1(\mathrm{d}\omega_1) \\ &= \int_{\Omega_2} \Big(\int_{\Omega_1} f(\omega_1, \omega_2)\, \mathrm{P}_1(\mathrm{d}\omega_1)\Big)\mathrm{P}_2(\mathrm{d}\omega_2)\end{aligned}$$

The proof would go via the standard procedure: first you would do it for indicator functions, using the previous proposition, then by linearity for simple functions and then you would use the approximation scheme of measurable, bounded functions via simple functions to deduce the general case with the help of the integral convergence theorems. We refer to [V], Theorem 1.12 for the detailed statement and proof.

We close with

Definition 1.28 (Independence). *Two random variables variables X_1, X_2 on probability spaces $(\Omega_1, \Sigma_1, \mathrm{P}_1)$ and $(\Omega_2, \Sigma_2, \mathrm{P}_2)$ are independent if their joint law (X_1, X_2) on $(\Omega_1 \times \Omega_2, \Sigma_1 \times \Sigma_2)$ is the product measure $\mathrm{P}_1 \times \mathrm{P}_2$.*

1.7. DISTRIBUTIONS AND EXPECTATION.

In this susection we will introduce some basic notation that we should be comfortable with as we will be using it freely throughout the rest of the notes.

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and X be a random variable on the space (this means a measurable function but we will not be repeating so from now on). We can induce a probability measure (we will also be calling it **distribution**) α on $(\mathbb{R},\mathcal{B})$ where $\mathcal{B}$ is the Borel σ-algebra as :

$$\alpha\big((-\infty,x]\big) := \mathrm{P}\big(\omega\colon X(\omega)\le x\big) = \mathrm{P}\big(X\le x\big)$$

In this case we may use the notation $\alpha=\mathrm{P}X^{-1}$, the pullback measure in measure theory notation.

Definition 1.29 (Expectation). *The expectation of a random variable X is defined as*

$$\mathrm{E}\big[X\big] := \int_\Omega X(\omega)\mathrm{P}(\omega). \tag{1.4}$$

In the above notation we can also write

$$\mathrm{E}\big[X\big] = \int_{\mathbb{R}} x\alpha(\mathrm{d}x),$$

which essentially comes from (1.4) by the change of variables $x:=X(\omega)$. Generalising (1.4) we have that for any real function g:

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

In particular, the k-moment of a random variable X is $\mathrm{E}[X^k]$.

Finally, for two variables X,Y, we define their covariance as

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

2. WEAK CONVERGENCE

2.1. CHARACTERISTIC FUNCTIONS. The notion of characteristic functions will be central to the foundations of *weak convergence*. For a random variable X on a probability space $(\Omega,\mathcal{F},\mathrm{P})$ we denote by $\alpha=\mathrm{P}X^{-1}$ its distribution. We then have

Definition 2.1. *The characteristic function of X is*

$$\phi(t) := \mathrm{E}\big[e^{\mathrm{i}tX}\big] = \int_\Omega e^{\mathrm{i}tX(\omega)}\mathrm{P}(\mathrm{d}\omega) = \int_{\mathbb{R}} e^{\mathrm{i}tx}\alpha(\mathrm{d}x),$$

where $\mathrm{i}:=\sqrt{-1}$.

The significance of the characteristic function, as we will see, is that it characterises the distribution of the corresponding random variable. This will be the subject of Theorem 2.4. Before let us record some general properties of the characteristic function:

Proposition 2.2. *The characteristic function of any probability distribution satisfies:*

1. $|\phi(t)|\le 1$ *for any* $t\in\mathbb{R}$ *with quality if* $t=0$,
2. $\phi(\cdot)$ *is uniformly continuous,*
3. $\phi(\cdot)$ *is a positive functions, i.e. for any* $\xi_1,\dots,\xi\in\mathbb{C}$ *and* $t_1,\dots,t_n\in\mathbb{R}$ *it holds that*

$$\sum_{i,j=1}^n \phi(t_i-t_j)\,\xi_i\,\overline{\xi_j}\ge 0.$$

Proof. 1. We have that

$$|\phi(t)| = \Big|\int_{\mathbb{R}} e^{\mathrm{i}tx}\alpha(\mathrm{d}x)\Big| \le \int_{\mathbb{R}} \big|e^{\mathrm{i}tx}\big|\alpha(\mathrm{d}x) = \int_{\mathbb{R}}\alpha(\mathrm{d}x)=1.$$

2. We have

$$|\phi(t)-\phi(s)| = \Big|\int_{\mathbb{R}}\big(e^{\mathrm{i}tx}-e^{\mathrm{i}sx}\big)\alpha(\mathrm{d}x)\Big| \le \int_{\mathbb{R}}\big|e^{\mathrm{i}(t-s)x}-1\big|\alpha(\mathrm{d}x).$$

Since $\lim_{t-s\to 0}\big|e^{\mathrm{i}(t-s)x}-1\big|=0$ we have the claim by the bounded convergence theorem and the convergence is uniform as $\big|e^{\mathrm{i}(t-s)x}-1\big|$ depends only on $t-s$.

3. We have

$$\sum_{i,j=1}^{n}\phi(t_i-t_j)\,\xi_i\,\overline{\xi_j}=\sum_{i,j=1}^{n}\int_{\mathbb{R}}e^{\mathrm{i}(t_i-t_j)x}\,\xi_i\overline{\xi}_j\,\alpha(\mathrm{d}x)=\sum_{i,j=1}^{n}\int_{\mathbb{R}}e^{\mathrm{i}t_ix}\,\xi_i\overline{e^{\mathrm{i}t_ix}\xi_j}\,\alpha(\mathrm{d}x)$$
$$=\int_{\mathbb{R}}\big|\sum_{i,=1}^{n}e^{\mathrm{i}t_ix}\xi_i\big|^2\,\alpha(\mathrm{d}x)\geq 0.$$

□

The following theorem provides a characterisation of characteristic functions. It says that the properties listed in the previous proposition (with the statement about continuity a bit relaxed) characterise characteristic functions. We refer to [V] for the proof of the Theorem.

Theorem 2.3 (Bochner). *Any positive definite function, which is continuous at* 0 *with* $\phi(0)=1$ *is a characteristic function of a probability distribution on* $\mathbb{R}$.

The next proposition states the characteristic functions determine the distribution and provides an inversion formula. This is closely related to inverting the Fourier transform.

Theorem 2.4. *Characteristic functions define uniquely a probability distribution. In particular, if* α *is a probability distribution on* $\mathbb{R}$ *with characteristic function* ϕ *and cumulative distribution function (c.d.f.)* F*, then at every continuity points* a,b *of the c.d.f. it holds that*

$$F(b)-F(a)=\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-\mathrm{i}tb}-e^{-\mathrm{i}ta}}{-\mathrm{i}t}\phi(t)\mathrm{d}t. \tag{2.1}$$

Exercise 9. *(i) Show that if two c.d.f.'s agree at continuity points, then they are identical.*

(ii) Extend formula (2.1) when a *or* b *are not continuity points.*

Remark 2.5. Before proving the formula, let us explain how to guess it. First, assume that $\alpha(\mathrm{d}x)$ has a density with respect to Lebesque, that is that it can be written as $\alpha(\mathrm{d}x)=f(x)\,\mathrm{d}x$ for a suitable function f. Then

$$\phi(t)=\int_{\mathbb{R}}e^{\mathrm{i}tx}\alpha(\mathrm{d}x)=\int_{\mathbb{R}}e^{\mathrm{i}tx}f(x)\,\mathrm{d}x,$$

and by Fourier inversion we would have that

$$f(x)=\frac{1}{2\pi}\int_{\mathbb{R}}e^{-\mathrm{i}tx}\phi(t)\,\mathrm{d}t.$$

We also have that $f(x)=F'(x)$ and so

$$F(b)-F(b)=\int_a^b f(x)\,\mathrm{d}x=\int_a^b\int_{\mathbb{R}}e^{-\mathrm{i}tx}\phi(t)\,\mathrm{d}t\,\mathrm{d}x=\frac{1}{2\pi}\int_{\mathbb{R}}\int_a^b e^{-\mathrm{i}tx}\phi(t)\,\mathrm{d}x\,\mathrm{d}t$$
$$=\frac{1}{2\pi}\int_{\mathbb{R}}\frac{e^{-\mathrm{i}tb}-e^{-\mathrm{i}ta}}{-\mathrm{i}t}\phi(t)\,\mathrm{d}t.$$

One thing that needs care is the interchange of integrals. Moreover, α may not have a density. However, knowing what the formula should look like, we can try to prove taking the steps backwards.

Proof of Theorem 2.4. We have

$$\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-\mathrm{i}tb}-e^{-\mathrm{i}ta}}{-\mathrm{i}t}\phi(t)\mathrm{d}t=\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-\mathrm{i}tb}-e^{-\mathrm{i}ta}}{-\mathrm{i}t}\int_{\mathbb{R}}e^{\mathrm{i}tx}\alpha(\mathrm{d}x)\mathrm{d}t.$$

At this occasion we can use Fubini to interchange the integral (before taking the limit $T\to\infty$) as on $[-T,T]\times\mathbb{R}$ the function $\frac{e^{-\mathrm{i}tb}-e^{-\mathrm{i}ta}}{-\mathrm{i}t}e^{\mathrm{i}tx}$ is bounded and integrable with respect $\mathrm{d}t\,\alpha(\mathrm{d}x)$. Doing so, we have that the above is equal to

$$\lim_{T\to\infty}\int_{\mathbb{R}}\alpha(\mathrm{d}x)\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-\mathrm{i}t(b-x)}-e^{-\mathrm{i}t(a-x)}}{-\mathrm{i}t}\mathrm{d}t=\lim_{T\to\infty}\int_{\mathbb{R}}\alpha(\mathrm{d}x)\frac{1}{2\pi}\int_{-T}^{T}\frac{\sin\big(t(x-a)\big)-\sin\big(t(x-b)\big)}{t}\mathrm{d}t, \tag{2.2}$$

where we used that $e^{\mathrm{i}tx}=\cos(tx)+\mathrm{i}\sin(tx)$ and cosine part has integral 0 due to parity. We can now use dominated convergence to pass the limit $\lim_{T\to\infty}$ inside the integral (check why !) and the fact that

$$\frac{1}{\pi}\lim_{T\to\infty}\int_0^T\frac{\sin(tz)}{t}\,\mathrm{d}t=\begin{cases}1, & \text{if} \quad z>0,\\ -1, & \text{if} \quad z<0,\\ 0, & \text{if} \quad z=0\end{cases}$$

to get that (2.2) equals

$$\frac{1}{2}\int_{\mathbb{R}}\big[sgn(x-a)-sgn(x-b)\big]\,\alpha(\mathrm{d}x)$$

and if a,b are continuity points, it equals

$$\int_{\mathbb{R}}\mathbb{1}_{(a,b]}(x)\,\alpha(\mathrm{d}x)=\int_{\mathbb{R}}\big(\mathbb{1}_{(-\infty,b]}(x)-\mathbb{1}_{(-\infty,a]}(x)\big)\,\alpha(\mathrm{d}x)=F(b)-F(a).$$

$\square$

Exercise 10. *Show that if* α *is a probability distribution such that* $\int_{\mathbb{R}}|x|\alpha(\mathrm{d}x)<\infty$, *then the characteristic function* ϕ *of* α *is continuously differentiable and* $\phi'(0)=i\int_{\mathbb{R}}x\alpha(\mathrm{d}x)=i\mathbb{E}\big[X\big]$.

Exercise 11. *Look at the example of distributions - characteristic functions in Chapter 2 of* [V]. *Prove the formulas of the characteristic functions for the: Poisson, geometric, uniform, gamma, two sided exponential and Cauchy distributions.*

2.2. WEAK CONVERGENCE OF MEASURES: DEFINITIONS & CRITERIA. We will now introduce the notion of weak convergence (of probability distributions or equivalently of distributions or random variables), which will be a central notion and a basis for many of our limit theorems (in distribution). Let us start with the definition:

Definition 2.6. *Let* $(\alpha_n)_{n\geq 1}$ *be a sequence of probability distributions on* $\mathbb{R}$ *with cumulative distribution functions* F_n. *We say that* α_n *converges weakly to a probability distribution* α *with cumulative distribution function* F *and denote it by* $\alpha_n \implies \alpha$ *or* $F_n \implies F$ *if*

$$\alpha_n(I)\xrightarrow[n\to\infty]{}\alpha(I),\qquad \textit{for all } I=[a,b] \textit{ with } a,b \textit{ continuity points of } F.$$

Equivalently, $\alpha_n \implies \alpha$ *if* $F_n(x)\xrightarrow[n\to\infty]{}F$ *at all continuity points of* F.

Exercise 12. *Show that* x *is a continuity point of a c.d.f.* F *corresponding to a probability distribution* α *if and only if* $\alpha(\{x\})=0$.

We can transcribe the above definition into the setting of random variables:

Definition 2.7. *Let* $(X_n)_{n\geq 1}$ *be a sequence of random variables with distribution* α_n *and* X *a random variable with distribution* α. *We say that* X_n *converges in distribution to* X *and write* $X_n \implies X$ *if* $\alpha_n \implies \alpha$.

We will see that

$$X_n\xrightarrow[n\to\infty]{a.s.}X \text{ implies } X_n\xrightarrow[n\to\infty]{\mathrm{P}}X \text{ implies } X_n \implies X\,.$$

Remark 2.8. Since $F(x) = \mathrm{P}\big((-\infty, x]\big)$, convergence in distribution can be translated to

$$\mathrm{P}\big(X_n \le x\big) \xrightarrow[n\to\infty]{} \mathrm{P}\big(X \le x\big), \qquad \text{for every } x \text{ which is a continuity point of the limiting c.d.f. } F.$$

We have already seen the first implication. For the second we will first need to develop some handy criteria for weak convergence. The opposite implications do not always hold. Can you think of examples ?

However, we have the following interesting theorem, which says that if one a sequence of converging distributions, one can find random variables with those distributions, which converge *a.s.*:

Theorem 2.9 (Skorokhod representation). *Let S be a complete, metric space (think of just $\mathbb{R}$ – if you want to see the general statement, refer to* [D]*, Chapter 2, Theorem (2.1)) and a sequence of probability distributions $(\alpha_n)_{n\ge1}$ such that $\alpha_n \implies \alpha$. Then we can define a sequence of random variables $(Y_n)_{n\ge1}$ on $[0,1]$ with distributions α_n and a random variable Y, again on $[0,1]$ with distribution α, such that $Y_n \xrightarrow[n\to\infty]{a.s.} Y$.*

Proof. We will prove it for $S = \mathbb{R}$ for the general proof see [D], Chapter 2, Theorem (2.1). Consider U a uniform random variable on $[0,1]$. The probability space here $([0,1], \mathcal{B}([0,1]), \mathrm{d}x)$ where $\mathrm{d}x$ is the Lebeque measure. For any $x \in [0,1]$ we have $U(x) = x$.

Let F_n the c.d.f.'s corresponding to the distributions α_n, i.e. $F_n(x) = \alpha_n\big((-\infty, x]\big)$. Let $F_n^{-1}(y) := \inf\{x\colon F_n(x) \ge y\}$ and similarly for F. Define the random variables $Y_n := F_n^{-1}(U)$ and $Y := F^{-1}(U)$. These are random variables on $[0,1]$ and for $x \in [0,1]$ (here x plays the more common role of ω) we have $Y_n(x) := F_n^{-1}(U(x)) = F_n^{-1}(x)$ and similarly for Y.

We have that Y_n has distribution α_n. Indeed,

$$\mathrm{P}(Y_n \le a) = \mathrm{P}(F_n^{-1}(U) \le a) = \mathrm{P}(U \le F_n(a)) = F_n(a),$$

where in the second equality we used the monotonicity of F_n and in the last that U is a uniform random variable.

We have the fact (check) that $F_n(x) \to F(x)$ for all continuity points of F if and only if $F_n^{-1}(x) \to F^{-1}(x)$ for all continuity points of F^{-1}. Moreover, F^{-1} has only countable discontinuity points as it is monotone. Therefore, a.e. point of $[0,1]$ is a continuity point of F^{-1} and on these points we will have $Y_n(x) := F_n^{-1}(x) \xrightarrow[n\to\infty]{} F^{-1}(x) =: Y(x)$. □

Let us next provide some more quantitative criteria for weak convergence.

Theorem 2.10 (Lévy-Crámer). *Let $\{(\alpha_n)_{n\ge1}, \alpha\}$ be a family of probability distributions with characteristic functions $(\phi_n)_{n\ge1}$ and ϕ, respectively. The following are equivalent:*

1. *$\alpha_n \implies \alpha$,*
2. *for every bounded, continuous f on $\mathbb{R}$, denoted $f \in C_b(\mathbb{R})$, we have that $\int_{\mathbb{R}} f \mathrm{d}\alpha_n \xrightarrow[n\to\infty]{} \int_{\mathbb{R}} f \mathrm{d}\alpha$,*
3. *for every $t \in \mathbb{R}$, we have that $\phi_n(t) \xrightarrow[n\to\infty]{} \phi(t)$.*

Proof. (1) $\implies$ (2). We will do a Riemann approximation. For this we need first to restrict to a compact interval $[a,b]$. We can do so since

$$\int_{[a,b]^c} f\, \mathrm{d}\alpha_n \le \|f\|_\infty \int_{[a,b]^c} \mathrm{d}\alpha_n = \|f\|_\infty \big(F_n(a) + 1 - F_n(b)\big)$$

and we can take a, b large enough (negative and positive, respectively) so that the RHS is less that ε.

We next partition $[a,b] = \cup_{i=1}^N [a_{i-1}, a_i]$ with $a_0 = a$ and $a_N = b$ with $\sup_i |a_i - a_{i-1}|$ small enough so that $\sup_{x\in[a_{i-1},a_i]} |f(x) - f(a_i)| \le \delta$ for arbitrarily small δ. This can be done since f is continuous and $[a,b]$ compact, so restricted to $[a,b]$ it is uniformly continuous. We also choose $(a_i)_{i\ge0}$ to be continuity points of F, so that

$$\sum_{i=1}^N f(a_{i-1})\, \alpha_n\big((a_{i-1}, a_i]\big) = \sum_{i=1}^N f(a_{i-1})\, \big(F_n(a_i) - F_n(a_{i-1})\big) \xrightarrow[n\to\infty]{} \sum_{i=1}^N f(a_{i-1})\, \big(F(a_i) - F(a_{i-1})\big)$$

We then have

$$\begin{aligned}\Big|\int_{[a,b]} f\mathrm{d}\alpha_n - \sum_{i=1}^{N} f(a_{i-1})\big(F_n(a_i)-F_n(a_{i-1})\big)\Big| &= \Big|\int_{[a,b]} f\mathrm{d}\alpha_n - \sum_{i=1}^{N} f(a_{i-1})\int_{a_{i-1}}^{a_i}\mathrm{d}\alpha_n\Big|\\ &\le \sum_{i=1}^{n}\int_{a_{i-1}}^{a_i}|f(x)-f(a_{i-1})|\mathrm{d}\alpha_n\\ &\le \delta.\end{aligned}$$

We can do the same estimate with F_n replaced by F and α_n with α_n. Comparing the two we then obtain

$$\Big|\int_{[a,b]} f\mathrm{d}\alpha_n - \int_{[a,b]} f\mathrm{d}\alpha\Big| \le \delta,$$

for arbitrarily small δ. All the above complete the claim.

(2) $\implies$ (3). This is easy since we can take $f(x)=\cos(tx)$ and $f(x)=\sin(tx)$ and then combine them to $f(x)=e^{\mathrm{i}tx}$.

(3) $\implies$ (1). This is more complicated and follows from the more general theorem that we state and prove next. ☐

Theorem 2.11. *Let (ϕ_n) be characteristic functions corresponding to probability distributions (α_n). If $\phi_n(t)$ converges to $\phi(t)$ for every $t\in\mathbb{R}$ and $\phi(t)$ is continuous at $t=0$, then $\phi(t)$ is a characteristic function of a probability distribution α and $\alpha_n \implies \alpha$.*

Proof. If $\phi(t) := \lim_{n\to\infty}\phi_n(t)$ is continuous at $t=0$, then by Bochner's theorem $\phi(t)$ is a characteristic function; notice that positive definiteness is preserved by pointwise limits. So we just need to check that if α is the distribution function corresponding to ϕ, then $\alpha_n \implies \alpha$. Equivalently, we need to show that $F_n(x) \xrightarrow[n\to\infty]{} F(x)$ for every x which is a continuity point of F, where F_n, F are the cdf's of α_n and α, respectively. Let us do so.

Step 1. First, let $(r_j)_{j\ge1}$ be an enumeration of the rational numbers. For any j, the sequence $(F_n(r_j))_{n\ge1}$ is a bounded sequence (since $0\le F\le1$) and so there is a subsequence $(\hat{n}_k^j)$ such that $F_{\hat{n}_k^j}(r_j)$ converges as $k\to\infty$ to a number, which we denote by $G(r_j)$.

We next want to show that we can pick a subsequence (n'_k) such that $F_{n'_k(r)}$ converges to $G(r)$ for all rationals r. We do so by a *diagonal sequence argument*:

- for rational r_1 consider the subsequence $n_k^{(1)}$ for which $F_{n_k^{(1)}}(r_1)\to G(r_1)$,
- for rational r_2 consider the subsequence $n_k^{(2)}$ of $n_k^{(1)}$ for which $F_{n_k^{(2)}}(r_1)\to G(r_2)$,
- in general, for rational r_i consider the subsequence $n_k^{(j)}$ of $n_k^{(j-1)}$ for which $F_{n_k^{(j)}}(r_j)\to G(r_j)$.

Consider now the sequence $(n_j^{(j)})_{\ge1}$, i.e. the j element of all above subsequences (we consider the j element so that to guarantee that this sequence actually goes to infinity). Then for every rational r, we have that $F_{n_j^{(j)}}(r)$ will converge to $G(r)$.

Step 2. We next want to extend G as a function on $\mathbb{R}$. We do this by defining

$$G(x) := \inf_{\mathbb{Q}\ni r>x} G(r), \qquad \text{for any } x\in\mathbb{R}.$$

G has the properties of a cdf:

- **Monotonicity:** if $x_1<x_2$ (assume they are not rationals) then $\{r\in\mathbb{Q}\colon r>x_1\}\supset\{r\in\mathbb{Q}\colon r>x_2\}$ and so $G(x_1):=\inf_{r\in\mathbb{Q}\colon r>x_1}G(r)\le\inf_{r\in\mathbb{Q}\colon r>x_2}=:G(x_2)$. **Exercise:** check the case when either x_1 or x_2 are rationals.
- **Right-continuity:** Assume $x_n\downarrow x$. Pick rationals $r_n\in[x_{n+1},x_n]$. We now that

$$G(r_n)\ge G(x_{n+1})\ge G(r_{n+1})$$

. By the definition of $G(x) := \inf_{\mathbb{Q}\ni r>x} G(r)$, we have that $G(r_n) \to G(x)$ and so by the above sandwiching we will also have that $G(x_n) \to G(x)$.

- G **is a probability cdf**: that is we need to check that $G(+\infty) = 1$ and $G(-\infty) = 0$. This will introduce the very important concept of **tightness** – meaning that there is no "mass" escaping to infinity. We will formalise this notion after we finish the proof for the purposes of which we will use the following useful inequality:

Lemma 2.12. *Let F be the cdf of a probability distribution corresponding to a characteristic function ϕ. For any $T > 0$ we have:*

$$1 - F\Big(\frac{2}{T}\Big) + F\Big(-\frac{2}{T}\Big) \le 2\Big(1 - \frac{1}{2T}\int_{-T}^{T} \phi(t)\,\mathrm{d}t\Big). \tag{2.3}$$

We will defer the proof of this inequality until after the proof of this Theorem. For the moment let us just use inequality (2.3) to show that G is a probability cdf as follows:

Employ (2.3) for all F_{n_k}:

$$1 - F_{n_k}\Big(\frac{2}{T}\Big) + F_{n_k}\Big(-\frac{2}{T}\Big) \le 2\Big(1 - \frac{1}{2T}\int_{-T}^{T} \phi_{n_k}(t)\,\mathrm{d}t\Big),$$

and for T rational. Passing to the limit $n_k \to \infty$ (assuming that $2/T$ and $-2/T$ are continuity points of G) this will also imply the inequality

$$1 - G\Big(\frac{2}{T}\Big) + G\Big(-\frac{2}{T}\Big) \le 2\Big(1 - \frac{1}{2T}\int_{-T}^{T} \phi(t)\,\mathrm{d}t\Big).$$

Now we pass to the limit $T \downarrow 0$. The left-hand side will converge to $1 - G(\infty) - G(-\infty)$, while the right-hand side will converge to 0 because of the assumption of continuity of ϕ at $t = 0$ and we are done with this step.

Step 3. Having constructed the limiting G we now want to show that $F_n \implies G$. Remember that we have only shown that there is a subsequence n_k such that $F_{n_k}(r) \to G(r)$ on rationals.

We start by showing that for every continuity point x of G and for (n_k) the aforementioned subsequence, we have that $F_{n_k}(x) \to G(x)$. To this end, let r be any rational greater than x. We have that $F_{n_k}(x) \le F_{n_k}(r)$ and then $\limsup_{n_k} F_{n_k}(x) \le \limsup_{n_k} F_{n_k}(r) = G(r)$ and letting $r \downarrow x$, we have that $G(r) \downarrow G(x)$ and so $\limsup_{n_k} F_{n_k}(x) \le G(x)$. Similarly, we can obtain that $\liminf_{n_k} F_{n_k}(x) \ge G(x)$, *assuming that x is a continuity point.*

To show that the whole sequence converges to G we follow this argument: Suppose it doesn't. Then there will be a subsequence, which doesn't converge to G. However, we have shown that every (sub)sequence has a (further) subsequence, say (m_k) that converges i.e. $F_{m_k} \implies \widetilde{G}$. Then by assumption (in the first equality of each line below) and by weak convergence (in the second equalities) we have:

$$\begin{aligned} \phi(t) &= \lim \int e^{\mathrm{i}tx}\mathrm{d}\alpha_{m_k} = \lim \int e^{\mathrm{i}tx}\mathrm{d}\widetilde{\alpha}, \qquad \text{and} \\ \phi(t) &= \lim \int e^{\mathrm{i}tx}\mathrm{d}\alpha_{n_k} = \lim \int e^{\mathrm{i}tx}\mathrm{d}\alpha, \end{aligned}$$

but then $\int e^{\mathrm{i}tx}\mathrm{d}\widetilde{\alpha} = \int e^{\mathrm{i}tx}\mathrm{d}\alpha$ and we know that characteristic functions determine the distribution, so $\widetilde{\alpha} = \alpha$. This completes the proof. ☐

We now go back to prove (2.3):

Proof of (2.3)**.** We have:

$$\begin{aligned}
\frac{1}{2T}\int_{-T}^{T}\phi(t)\,\mathrm{d}t &= \frac{1}{2T}\int_{-T}^{T}\Big(\int e^{\mathrm{i}tx}\alpha(\mathrm{d}x)\Big)\,\mathrm{d}t \\
&= \int\Big(\frac{1}{2T}\int_{-T}^{T}e^{\mathrm{i}tx}\,\mathrm{d}t\Big)\alpha(\mathrm{d}x) && \text{[we can use Fubini]} \\
&= \int\frac{\sin(Tx)}{Tx}\,\alpha(\mathrm{d}x) \\
&\leq \int\Big|\frac{\sin(Tx)}{Tx}\Big|\,\alpha(\mathrm{d}x) \\
&\leq \int_{|x|\leq\ell}\Big|\frac{\sin(Tx)}{Tx}\Big|\,\alpha(\mathrm{d}x)+\int_{|x|>\ell}\Big|\frac{\sin(Tx)}{Tx}\Big|\,\alpha(\mathrm{d}x) && \text{[for arbitrary } \ell] \\
&\leq \alpha\big(x\colon |x|\leq\ell\big)+\frac{1}{T\ell}\alpha\big(x\colon |x|>\ell\big)
\end{aligned}$$

where in the last we bounded each integral by using, respectively, that $|\sin x/x|\leq 1$ and $|\sin x/x|\leq 1/|x|$. We can then deduce that

$$\begin{aligned}
1-\frac{1}{2T}\int_{-T}^{T}\phi(t)\,\mathrm{d}t \;&\geq\; 1-\alpha\big(x\colon |x|\leq\ell\big)-\frac{1}{T\ell}\alpha\big(x\colon |x|>\ell\big) \\
&=\; \Big(1-\frac{1}{T\ell}\Big)\alpha\big(x\colon |x|>\ell\big) \\
&=\; \Big(1-\frac{1}{T\ell}\Big)\big(1-F(\ell)-F(-\ell)\big) \\
&=\; \frac{1}{2}\big(1-F(\ell)-F(-\ell)\big) && \text{[choosing } \ell:=2/T].
\end{aligned}$$

□

Let us now formalise the important notion of **tightness**:

Definition 2.13. *A family of probability measure $\mathcal{A}$ is called* **totally bounded** *or (more commonly)* **tight** *if anysequence $(\alpha_n)\subset\mathcal{A}$ has a subsequence that converges weakly.*

We can also state a criterion for tightness:

Proposition 2.14. *A family $\mathcal{A}$ of probability distributions is tight if and only if either one of the following two conditions is satisfied:*

$$\begin{aligned}
&(1) \qquad \lim_{\ell\to\infty}\sup_{\alpha\in\mathcal{A}}\alpha(x\colon |x|\geq\ell)=0, \\
&(2) \qquad \lim_{h\downarrow 0}\sup_{\alpha\in\mathcal{A}}\sup_{|t|\leq h}|1-\phi_\alpha(t)|=0.
\end{aligned}$$

Proof. By the proof of Theorem 2.11 we have that every sequence of probability distribution (α) has always a subsequence which converges weakly. However, the question is whether the limiting distribution α_* is actually a probability distribution, i.e. $\alpha_*(\mathbb{R})=1$.

Step 1: upwards direction. We know that (2) $\implies$ (1). Let us now prove that condition (1) implies that $\alpha_*(\mathbb{R})=1$. Condition (1) translates to:

$$\forall\varepsilon>0,\exists\,\ell_\varepsilon \quad \text{such that} \quad \sup_{\alpha\in\mathcal{A}}\alpha(|x|\geq\ell_\varepsilon)<\varepsilon,$$

which will also imply that $\alpha_*(|x|\geq\ell_\varepsilon)<\varepsilon$ since α_* is a limit of elements in $\mathcal{A}$. But this equivalent to $\lim_{\ell\to\infty}\alpha_*(|x|\geq\ell)=0$ or that $\lim_{\ell\to\infty}\alpha_*(|x|<\ell)=1\Leftrightarrow\alpha_*(\mathbb{R})=1$.

Step 2: downwards direction. Assume that $\mathcal{A}$ is tight, which means that every sequence of elements of $\mathcal{A}$ has a convergence subsequence, but (1) is not valid. This would then imply that there exists a sequence $(\alpha_n) \subset \mathcal{A}$ such that

$$\limsup_{\ell\to\infty}\limsup_{n}\alpha_n(|x|\geq \ell) > \varepsilon, \qquad \text{for some } \varepsilon>0,$$

We can assume that $\alpha_n \implies \alpha_*$ for some probability distribution α_* and then the above would imply that

$$\limsup_{\ell\to\infty}\alpha_*(|x|\geq \ell) > \varepsilon,$$

but then this would contradict the fact that α_* is a probability distribution.

Finally, we will show that (1) $\implies$ (2): We have

$$\begin{aligned}|1-\phi_\alpha(t)| &\leq \int |1-e^{\mathrm{i}tx}|\alpha(\mathrm{d}x)\\ &= \Big(\int_{|x|>\ell}+\int_{|x|\leq \ell}\Big)|1-e^{\mathrm{i}tx}|\alpha(\mathrm{d}x)\\ &\leq 2\alpha(|x|\geq \ell)+t\int_{|x|\leq\ell}|x|\alpha(\mathrm{d}x)\\ &\leq 2\alpha(|x|\geq \ell)+t\ell,\end{aligned}$$

for arbitrary $\ell>0$ and so

$$\begin{aligned}\limsup_{h\downarrow 0}\sup_{\alpha\in\mathcal{A}}\sup_{|t|\leq h}|1-\phi_\alpha(t)| &\leq \limsup_{h\downarrow 0}\sup_{\alpha\in\mathcal{A}}\big(2\alpha(|x|\geq \ell)+h\ell\big)\\ &= 2\sup_{\alpha\in\mathcal{A}}\alpha(|x|\geq\ell)\xrightarrow[\ell\to\infty]{}0,\end{aligned}$$

by assumption (1). □

Let us close this section with a reminder of all the three modes of convergence and their relations:

- **Almost sure (a.s.) convergence.** Random variables (X_n) are said to converge a.s. to X and we write $X_n \xrightarrow[n\to\infty]{a.s.} X$, if for every $\varepsilon>0$,
$$\mathrm{P}(\limsup_{n\to\infty}|X_n-X|>\varepsilon)\equiv \mathrm{P}(\{|X_n-X|>\varepsilon,\ i.o.\})=0$$
- **Convergence in Probability.** Random variables (X_n) are said to converge to X in probability and we write, $X_n\xrightarrow[n\to\infty]{\mathrm{P}} X$, if for every $\varepsilon>0$,
$$\mathrm{P}(|X_n-X|>\varepsilon)\xrightarrow[n\to\infty]{}0$$
- **Convergence in Distribution.** Random variables (X_n) are said to converge to X in distribution and we write, $X_n\xrightarrow[n\to\infty]{d}X$ or $X_n\implies X$ if the corresponding distritions $\alpha_n=\mathrm{P}X_n^{-1}$ converge weakly to $\alpha=\mathrm{P}X^{-1}$. Equivalently if the characteristic functions $\phi_n(t):=\mathrm{E}\big[e^{\mathrm{i}tX_n}\big]$ converge for every $t\in\mathbb{R}$ to the characteristic function $\phi(t):=\mathrm{E}\big[e^{\mathrm{i}tX}\big]$.

We have the relations:

a.s. converence $\implies$ convergence in probability $\implies$ convergence in distribution

The opposite assertions do not in generally .

Exercise 13. *1. Prove the convergence in probability implies convergence in distribution*

2. Prove that if (X_n) converge in probability to an a.s. constant random variable X, then (X_n) also converge to X in probability.

3. LAWS OF LARGE NUMBERS

In this section we will prove the standard Laws of Large Numbers (LLN). There are two kinds: the Weak LLN, which is stated as convergence in probability and the Strong LLN which is stated as an a.s. convergence.

3.1. Weak Law of Large Numbers. We start with the Weak LLN, which we first prove under a suboptimal condition. We will be using a lot this basic lemma:

Lemma 3.1 (Chebyshev's inequality). *Let* X *be a nonnegative random variable and* f *a nonnegative, nondecreasing function. Then*

$$\mathrm{P}(X \geq a) \leq \frac{1}{f(a)}\mathrm{E}[f(X)\mathbb{1}_{X\geq a}] \leq \frac{1}{f(a)}\mathrm{E}[f(X)].$$

Proof. We just need to use the inequality $\mathbb{1}_{X\geq a} \leq \mathbb{1}_{X\geq a}\frac{f(X)}{f(a)}$ and also write $\mathrm{P}(X \geq a) = \mathrm{E}[\mathbb{1}_{X\geq a}]$. ☐

Theorem 3.2. *Let* $(X_n)_{n\geq 1}$ *be a family of independent, identically distributed random variables (***i.i.d.***), such that* $\mathrm{E}[X_i] = \mu$ *and* $\mathrm{E}[X_i^2] < \infty$*. We then have that*

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{P} \mu.$$

Proof. First, we can assume that $\mu = 0$, otherwise we consider, instead, the random variables $X_i' := X_i - \mu$. Next, we denote $S_n := X_1 + \cdots + X_n$. Then by Chebyshev we have that

$$\begin{aligned}\mathrm{P}\big(|S_n| \geq n\varepsilon\big) &\leq \frac{1}{n^2\varepsilon^2}\mathrm{E}\big[|S_n|^2\big] \\ &= \frac{1}{n^2\varepsilon^2}\Big\{\sum_{i=1}^{n}\mathrm{E}[X_i^2] + 2\sum_{i<j}\mathrm{E}[X_iX_j]\Big\} \\ &= \frac{1}{n\varepsilon^2}\mathrm{E}[X_1^2] \xrightarrow[n\to\infty]{} 0,\end{aligned}$$

where in the last equality we used the independence and the assumption that X_i are mean 0, to drop the cross term and the identically distributed assumption. ☐

We will next remove the assumption on second moments and state the standard wLLN, with the more natural condition of finite first moment. More precisely, we have

Theorem 3.3. *Let* $(X_n)_{n\geq 1}$ *be a family of independent, identically distributed random variables (***i.i.d.***), such that* $\mathrm{E}[X_i] = \mu$ *and* $\mathrm{E}[|X_i|] < \infty$*. We then have that*

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{P} \mu.$$

Proof. The trick is to reduce this theorem to the previous theorem. We will achieve this via the standard trick of **truncation**. First,, we again assume that $\mu = 0$ and then choose a truncation level C and define the random variables

$$X_i^{\leq C} := X_i\mathbb{1}_{|X_i|\leq C} \qquad \text{and} \qquad X_i^{>C} := X_i\mathbb{1}_{|X_i|>C}.$$

We clearly have $X_i = X_i^{\leq C} + X_i^{>C}$ and we write $S_n^{\leq C} = \sum_{i=1}^{n} X_i^{\leq C}$ and $S_n^{>C} = \sum_{i=1}^{n} X_i^{>C}$ These random variables might not be mean 0 any more but

$$0 = \mathrm{E}[X_i] = \mathrm{E}[X_i\mathbb{1}_{|X_i|\leq C}] + \mathrm{E}[X_i\mathbb{1}_{|X_i|>C}] = \mathrm{E}[X_i^{\leq C}] + \mathrm{E}[X_i^{>C}].$$

So we can write

$$S_n = \Big(S_n^{\leq C} - n\mathrm{E}[X_i^{\leq C}]\Big) + \Big(S_n^{<C} - n\mathrm{E}[X_i^{>C}]\Big),$$

and so

$$\begin{aligned}\mathrm{P}\big(|S_n| \geq n\varepsilon\big) &\leq \mathrm{P}\Big(\,|S_n^{\leq C} - n\mathrm{E}[X_i^{\leq C}]| + |S_n^{>C} - n\mathrm{E}[X_i^{>C}| \geq n\varepsilon\,\Big) \\ &\leq \mathrm{P}\Big(\,|S_n^{\leq C} - n\mathrm{E}[X_i^{\leq C}]| \geq \frac{n\varepsilon}{2}\,\Big) + \mathrm{P}\Big(\,|S_n^{>C} - n\mathrm{E}[X_i^{>C}| \geq \frac{n\varepsilon}{2}\,\Big).\end{aligned}$$

The first term goes to 0 as $n \to \infty$ by the previous theorem as the variables X_i^C are bounded by C and, thus, have second moments. We bound the second term by simple Chebyshev as

$$\mathrm{P}\Big(\,|S_n^{>C} - n\mathrm{E}[X_i^{>C}| \geq \frac{n\varepsilon}{2}\,\Big) \leq \frac{2}{n\varepsilon}\mathrm{E}\Big[\,|S_n^{>C} - n\mathrm{E}[X_i^{>C}|\,\Big] \leq \frac{4}{n\varepsilon}n\mathrm{E}\Big[\,|X_1^{>C}\Big] = \frac{4}{\varepsilon}4\mathrm{E}\Big[\,|X_1^{>C}\Big],$$

but the last term goes to 0, uniformly in n, as $C \to \infty$ by the dominated convergence theorem. ☐

We can also give a quick proof of the wLLN using characteristic functions:

Proof of wLLN via characteristic functions. In this proof there is not much simplification to assume mean 0. Let

$$\phi_{S_n}(t) := \mathrm{E}\big[e^{\mathrm{i}tS_n/n}\big] = \mathrm{E}\big[\prod_{i=1}^n e^{\mathrm{i}tX_i/n}\big] = \prod_{i=1}^n \mathrm{E}\big[e^{\mathrm{i}tX_i/n}\big] = \Big(\mathrm{E}\big[e^{\mathrm{i}tX_i/n}\big]\Big)^n = \phi_{X_1}\Big(\frac{t}{n}\Big)^n.$$

We will next Taylor expand the characteristic function $\phi_{X_1}\big(\frac{t}{n}\big)$ around 0 and write the above as

$$\phi_{X_1}\Big(\frac{t}{n}\Big)^n = \Big(1 + \phi'(0)\frac{t}{n} + o(\frac{1}{n})\Big)^n \to e^{\phi'(0)t}, \qquad \text{as } n \to \infty.$$

Using the fact that $\phi'(0) = i\mu$ we have that the limit of the characteristic functions is $e^{\mathrm{i}\mu t}$ which is the characteristic function of δ_μ, the Dirac distribution at μ or the characteristic function of the constant random variable μ. We can now use the result of Exercise 13 to conclude the convergence in probability. ☐

3.2. Strong Law of Large Numbers. The statement of the Strong LLN is that for i.i.d. sequence (X_n) with mean μ and finite first moment $\mathrm{E}[|X_i|] < \infty$, we have the a.s. convergence of the sample mean to the statistical mean:

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{a.s.} \mu.$$

For a warm-up, let us first prove this statement with the extra assumption of finite fourth moment.

Theorem 3.4. *Assume (X_n) is an i.i.d. sequence with mean μ and $\mathrm{E}[X_i^4] < \infty$, then*

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{a.s.} \mu.$$

Proof. Assume that $\mu = 0$. With the view of using the first Borel-Cantelli lemma we estimate:

$$\begin{aligned}\mathrm{P}\big(|S_n| > \varepsilon n\big) &\le \frac{1}{\varepsilon^4 n^4}\mathrm{E}\big[S_n^4\big] = \frac{1}{\varepsilon^4 n^4}\mathrm{E}\big[(X_1 + \cdots + X_n)^4\big] \\ &= \frac{1}{\varepsilon^4 n^4}\Big\{\sum_{i=1}^n \mathrm{E}[X_i^4] + \sum_{i,j,k\,\text{different}} \mathrm{E}[X_i^3 X_j X_k] + \sum_{i\neq j} \mathrm{E}[X_i^2 X_j^2]\Big\},\end{aligned}$$

and using the independence and the i.i.d. property we have that the above equals

$$\frac{1}{\varepsilon^4 n^4}\Big\{n\mathrm{E}[X_1^4] + 3n(n-1)\sum_{i\neq j}\mathrm{E}[X_1^2]^2\Big\} \le \frac{C}{n^2},$$

So we have that $\sum_{n\ge 1} \mathrm{P}(|S_n| > \varepsilon) < \infty$ and the result follows by Borel-Cantelli. ☐

We will next move to the real stuff ! There are several proofs of the strong LLN under only finite first moment. We will present Etemadi's proof as it contains a number of interesting ideas, which can be applied in various other context. In the proof we will skip a few of the technical details, which can be found in Durrett's book [D].

Before we start we need the following very useful representation of the mean of a random variable:

Lemma 3.5. *For a nonnegative random variable X we have*

$$\mathrm{E}[X] = \int_0^\infty \mathrm{P}(X \ge \ell)\,\mathrm{d}\ell. \tag{3.1}$$

Proof. We have the trivial (but clever) identity $X = \int_0^\infty \mathbb{1}_{X\ge\ell}\,\mathrm{d}\ell$ which implies that

$$\mathrm{E}[X] = E\big[\int_0^\infty \mathbb{1}_{X\ge\ell}\,\mathrm{d}\ell\big] = \int_0^\infty E\big[\mathbb{1}_{X\ge\ell}\big]\mathrm{d}\ell = \int_0^\infty \mathrm{P}(X \ge \ell)\,\mathrm{d}\ell.$$

☐

We can now prove:

Theorem 3.6. *Assume (X_n) is an i.i.d. sequence with mean μ and* $\mathrm{E}[X_i^4] < \infty$, *then*

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{a.s.} \mu.$$

Proof. Step 1. (truncation) Let

$$Y_k := X_k \mathbb{1}_{|X_k| \le k}.$$

Note that the truncation used here is not constant but the truncation level employed increases with k. Let also

$$T_n = Y_1 + \cdots + Y_n \qquad \text{and} \qquad S_n = X_1 + \cdots + X_n.$$

Let us check that it is enough to prove the sLLN for T_n, i.e. that $T_n \xrightarrow[n\to\infty]{a.s.} \mu$. Indeed,

$$\begin{aligned}\sum_k \mathrm{P}(|X_k| \ge k) &\le \int_0^\infty \mathrm{P}(|X_k| \ge \ell)\,\mathrm{d}\ell \\ &= \int_0^\infty \mathrm{P}(|X_1| \ge \ell)\,\mathrm{d}\ell && \text{[by i.i.d.]} \\ &= \mathrm{E}[|X_1|] < \infty && \text{[by (3.1) and assumption].}\end{aligned}$$

Therefore, by Borel-Cantelli $\mathrm{P}(|X_k| \ge k,\ i.o.) = 0$, which means that a.s. for all large enough n we have $X_n = Y_n$. We can also unravel this and write that for almost every ω, we have that for all large enough n $X_n(\omega) = Y_n(\omega)$. This implies that for a.e. ω there exists a random constant $R(\omega) < \infty$ such that

$$|S_n(\omega) - T_n(\omega)| < R(\omega) \implies \frac{1}{n}|S_n(\omega) - T_n(\omega)| < \frac{1}{n}R(\omega) \to 0.$$

Step 2. (Borel-Cantelli along subsequences) Ideally we would like to prove that $\sum_{n\ge1} \mathrm{P}(|\frac{S_n}{n} - \mu| \ge \varepsilon) < \infty$ (or the analogous statement for the truncated variable sum T_k) but this cannot be achieved – we sum too many terms having a poor a priori bound. The trick will be to prove the convergence along a subesequence and the sandwich all in between terms. To do the sandwiching we will need some positivity in order to compare. The trick here is

Ertemadi's idea: *We can assume that X_n's are nonnegative because otherwise we can split into the positive and negative parts $X_n = X_n^+ - X_n^-$ and since (X_n^+) and (X_n^-) are i.i.d. sequences with finite first moment we could use the sLLN for each one individually and combine to obtain the sLLN for (X_n).*

We next choose the appropriate **subsequence**:

$$k_n := \lceil a^n \rceil, \qquad \text{for arbitrary } a{>}1.$$

Eventually we will be taking $a \downarrow 1$.

Let us now estimate the Borel-Cantelli terms along this subsequence:

$$\begin{aligned}\sum_{n\ge1} \mathrm{P}\big(|\,T_{k_n} - \mathrm{E}[T_{k_n}]\,| \ge \varepsilon k_n\big) &\le \sum_{n\ge1} \frac{1}{\varepsilon^2 k_n^2} \mathrm{E}\big[|\,T_{k_n} - \mathrm{E}[T_{k_n}]\,|^2\big] \\ &= \sum_{n\ge1} \frac{1}{\varepsilon^2 k_n^2} \mathbb{V}\mathrm{ar}\left(T_{k_n}\right) \\ &= \sum_{n\ge1} \frac{1}{\varepsilon^2 k_n^2} \sum_{m=1}^{k_n} \mathbb{V}\mathrm{ar}\left(Y_m\right),\end{aligned}$$

where in the last we used that the variance of the sum of independent variable is the sum of the variances. Note that Y_m are not i.i.d. (because they have different truncation levels), so more work is needed to bound the last sum. To this end, we write it as

$$\sum_{n\geq 1}\frac{1}{\varepsilon^2 k_n^2}\sum_{m=1}^{k_n}\mathbb{V}\mathrm{ar}\left(Y_m\right)=\frac{1}{\varepsilon^2}\sum_{m=1}^{\infty}\mathbb{V}\mathrm{ar}\left(Y_m\right)\sum_{n\geq 1}\frac{1}{k_n^2}\mathbb{1}_{k_n\geq m}$$
$$\leq\frac{1}{\varepsilon^2}\sum_{m=1}^{\infty}\mathbb{V}\mathrm{ar}\left(Y_m\right)\frac{1}{m^2}\frac{C}{1-a^{-2}},$$

for some constant C, where we estimated the sum over n, using the explicit expression $k_n=\lceil a^n\rceil$ (check).

So it remains to check that

$$\sum_{m\geq 1}\frac{\mathbb{V}\mathrm{ar}(Y_m)}{m^2}<\infty.$$

Prove (or check in [D]) that $\sum_{m\geq 1}\frac{\mathbb{V}\mathrm{ar}(Y_m)}{m^2}\leq C\mathrm{E}[|X_1|]$.

So in this step we have proven that

$$\mathrm{P}\big(|\,T_{k_n}-\mathrm{E}[T_{k_n}]\,|\geq\varepsilon k_n,\ i.o.\big)=0,$$

which means that

$$\frac{T_{k_n}-\mathrm{E}[T_{k_n}]}{k_n}\xrightarrow[n\to\infty]{a.s}0.$$

We note that

$$\frac{\mathrm{E}[T_{k_n}]}{k_n}\xrightarrow[n\to\infty]{}\mu.$$

because by dominated convergence we have $\mathrm{E}[Y_k]=\mathrm{E}[X_k\mathbb{1}_{X_k\leq k}]=\mathrm{E}[X_1\mathbb{1}_{X_1\leq k}]\to\mathrm{E}[X_1]$ as $k\to\infty$ and combine this with Césaro mean. So we have proven that

$$\frac{T_{k_n}}{k_n}\xrightarrow[n\to\infty]{a.s}\mu.$$

We are now ready for the final step, which is to show that this limit holds not just for the subsequence but for all n.

Step 3. (Sandwiching) For every m, find n such that $k_n\leq m\leq k_{n+1}$. Using the positivity of the summands and the monotonicity of k_n, we have that

$$\frac{T_{k_n}}{k_{n+1}}\leq\frac{T_m}{m}\leq\frac{T_{k_{n+1}}}{k_n}$$

and noting that $k_{n+1}/k_n\approx a$ we have that

$$\frac{1}{a}\frac{T_{k_n}}{k_n}\leq\frac{T_m}{m}\leq a\frac{T_{k_{n+1}}}{k_{n+1}},$$

and the result follows by first taking the limit $n\to\infty$ and then $a\downarrow 1$. □

Exercise 14. *Show that the sLLN does not hold if* $\mathrm{E}[|X_1|]=\infty$.

Exercise 15 (A taste of renewal theory). *Let* $(X_n)_{n\geq 1}$ *a family of positive, i.i.d. random variables with mean* μ, $S_n:=X_1+\cdots+X_n$ *and* $N_t:=\sup\{n\colon S_n\leq t\}$. *Show that*

$$\frac{N_t}{t}\to\frac{1}{\mu},\qquad a.s.\ for\ t\to\infty.$$

3.3. Kolmogorov's 0-1 Law.

The Kolmorogorov's 0-1 Law gives an explanation why the limit in the sLLN's is a constant. Basically it says that if $(X_n)_{n\geq 1}$ is an i.i.d. sequence, then any event that does not depend on any finite number of the random variables can only have probability 0 or 1. To formulate this we first need the definition a **tail σ-field**:

Definition 3.7. *Let $(X_n)_{n\geq 1}$ an i.i.d. family. Define $\mathcal{B}^n := \sigma\big(X_j : j \geq n\big)$. The tail σ-field of the sequence $(X_n)_{n\geq 1}$ is defined as $\mathcal{B}^\infty := \cap_{n\geq 1}\mathcal{B}^n$.*

Some examples of events that belong to the tail σ-field are

- $\big\{\omega\colon \limsup_n X_n = 1\big\}$,
- $\big\{\omega\colon \lim_n X_n \text{ exists}\big\}$,
- $\big\{\omega\colon \sup_n |X_n| < \infty\,\big\}$
- $\big\{\omega\colon \lim_n \dfrac{X_1 + \cdots + X_n}{n} \in [a,b]\big\}$

On the other hand the event $\big\{\omega\colon \sup_n |X_n| = 1\,\big\}$, is not a tail event as it depends on all values of the sequence.

We can now state and prove the theorem

Theorem 3.8. *If $A \in \mathcal{B}^\infty$, then* $\mathrm{P}(A) \in \{0,1\}$.

Proof. We will show that any $A \in \mathcal{B}^\infty$ is independent of itself, thus $\mathrm{P}(A) = \mathrm{P}(A\cap A) = \mathrm{P}(A)^2$, which implies the conclusion.

Let us prove the claim. We have that

$$A \in \mathcal{B}^\infty \subset \mathcal{B}^{n+1} = \sigma\big(X_j : j \geq n+1\big),$$

but the latter is independent of $\sigma\big(X_j : j \leq n\big)$ and since this holds for every n, we have that A is independent of $\cup_n \sigma\big(X_j : j \geq n\big)$. We denote the latter by $\mathcal{F}$. This somehow should imply the conclusion since A is by default an event in $\sigma\big(X_j : j \geq 1\big)$. We only need to check that being independent from $\mathcal{F} := \cup_n \sigma\big(X_j : j \geq n\big)$ implies that it is independent of $\sigma\big(X_j : j \geq 1\big)$. To this end, we will use again Dynkin's theorem: Define

$$\mathcal{A} := \big\{B\colon \mathrm{P}(A\cap B) = \mathrm{P}(a)\mathrm{P}(B)\big\}.$$

By the discussion above $\mathcal{A}$ contains $\mathcal{F}$. Check that $\mathcal{A}$ is a monotone class and that $\mathcal{F}$ is a field. Therefore, by Dynkin's theorem, $\mathcal{A}$ contains $\sigma(\mathcal{F}) = \sigma\big(X_j : j \geq 1\big)$, which means that $A \in \mathcal{A}$. □

Corollary 3.9. *If $(X_n)_{n\geq 1}$ is an i.i.d. sequence, then*

$$\lim_n \frac{X_1 + \cdots + X_n}{n}$$

has to be a.s. a constant.

4. Central Limit Theorem

In this section we will present the central limit theorem (CLT). Informally this should be thought of as follows: Suppose $(X_n)_{n\geq 1}$ is an i.i.d. sequence with mean μ and variance σ^2. Then we have the asumptotic

$$X_1 + \cdots + X_n \approx n\mu + \sqrt{n} \times \text{Gaussian}.$$

We will present three approaches to establish the CLT.

4.1. The method of characteristic functions.

Let us first formally state the CLT:

Theorem 4.1. *Suppose $(X_n)_{n\geq 1}$ is an i.i.d. sequence with mean μ and variance σ^2 and $S_n = X_1 + \cdots + X_n$, Then*

$$\frac{S_n - n\mu}{\sqrt{n}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\sigma^2),$$

where $\mathcal{N}(0,\sigma^2)$ is a Gaussian (Normal) random variable with mean 0 and variance σ^2. The above limit is to be interpreted as

$$\mathrm{P}\Big(\frac{S_n - n\mu}{\sqrt{n}} > x\Big) \xrightarrow[n\to\infty]{} \int_x^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} \mathrm{d}y.$$

The first proof will be via characteristic functions:

Proof. We start by noting that the characteristic function of the Gaussian is

$$\int_{\mathbb{R}} e^{\mathrm{i}tx} e^{-\frac{x^2}{2\sigma^2}} \frac{\mathrm{d}x}{\sqrt{2\pi\sigma^2}} = e^{-\frac{t^2\sigma^2}{2}}.$$

Assume, as usual that $\mu = 0$ and compute the asymptotics of the characteristic function

$$\begin{aligned}\phi_n(t) = \mathrm{E}\Big[e^{\mathrm{i}t\frac{S_n}{\sqrt{n}}}\Big] &= \Big(\mathrm{E}\Big[e^{\mathrm{i}t\frac{X_1}{\sqrt{n}}}\Big]\Big)^n = \phi_{X_1}\big(\frac{t}{\sqrt{n}}\big)^n \\ &= \Big\{1 + \frac{t}{\sqrt{n}}\phi'_{X_1}(0) + \frac{t^2}{2n}\phi''_{X_i}(0)^2 + o(n^{-1})\Big\}^n \\ &= \Big\{1 - \frac{t^2}{2n}\sigma^2 + o(n^{-1})\Big\}^n \\ &\xrightarrow[n\to\infty]{} e^{-\frac{t^2}{2}\sigma^2}.\end{aligned}$$

where we used the Taylor expansion up to second order. □

Let us now present the first generalisation of the CLT. The following, called Lindeberg's condition (not to be confused with Lindeberg's method) says that the random variables do not need to be i.i.d. The CLT would still hold if essentially no random variable contributes significantly to the total variance.

Theorem 4.2 (Lindeberg's CLT). *Assume $(X_n)_{n\ge1}$ are i.i.d., with mean zero and finite variance $(\sigma_n^2)_{n\ge1}$. Let*

$$s_n^2 := \mathbb{V}\mathrm{ar}\,\big(X_1 = \cdots + X_n\big) =: \sigma_1^2 + \cdots \sigma_n^2.$$

If

$$\frac{1}{s_n^2}\sum_{j=1}^n \mathbf{E}\Big[X_j^2\,\mathbb{1}_{|X_j|\ge\varepsilon s_n}\Big] \xrightarrow[n\to\infty]{} 0, \qquad \forall\varepsilon > 0,$$

then the CLT holds and

$$\frac{X_1 + \cdots + X_n}{s_n} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

Proof. We refer to [V], Chapter 3 for the proof. □

Exercise 16. *(CLT without 2^{nd} moment.) Let $(X_n)_{n\ge1}$ i.i.d. with $\mathrm{P}(X_1 > x) = \mathrm{P}(X_1 < -x)$ and $\mathrm{P}(|X_1| > x) = x^{-2}$ for $x \ge 1$. Then*

$$\frac{S_n}{\sqrt{n\log n}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

4.2. Lindeberg Principle. In this subsection we will demonstrate the Lindeberg principle, also called the Lindeberg method, through proving Lindeberg's CLT. However, the principle has much wider applicability. The idea is that the convergence of

$$\frac{X_1 + \cdots + X_n}{s_n}$$

where s_n is as in Lindeberg's CLT, to a normal random variable is trivial if X_i are independent Gaussian variables with mean zero and variance σ_i^2. Denote the latter by Z_i. Then one is left with comparing the distance

$$\mathrm{E}\Big[f\big(\frac{X_1 + \cdots + X_n}{s_n}\big)\Big] - \mathrm{E}\Big[f\big(\frac{Z_1 + \cdots + Z_n}{s_n}\big)\Big]$$

for any $f \in C_b(\mathbb{R})$. If this difference converges to 0 as $n \to \infty$, since $\mathrm{E}\big[f\big(\frac{Z_1+\cdots+Z_n}{s_n}\big)\big] = \mathrm{E}\big[f(Z)\big]$ for $Z \sim \mathcal{N}(0,1)$, then CLT for $(X_i)_{i\geq 1}$ follows. In order to control this difference we will be switching from Z_i's to X_i's one variable at a time. In other words we will telescope

$$\sum_{j=1}^{n} \mathrm{E}\Big[f\Big(\frac{Z_1+\cdots+Z_{j-1}+X_j+\cdots X_n}{s_n}\Big)\Big] - \mathrm{E}\Big[f\Big(\frac{Z_1+\cdots+Z_j+X_{j+1}+\cdots X_n}{s_n}\Big)\Big]$$

so that there is only one discrepancy at each summand. The idea is that if changing just one variable to another has sufficiently negligible effect, then the total difference should be small. In such a situation we often say that each individual random variable has *negligible influence*. The notion of *influence* plays an important role in studies of CLTs, noise sensitivity etc. We refer to [Z] for further discussion on this as well as more general Lindeberg principles and application.

We now give the proof of Lindeberg's CLT via the Lindeberg method. The proof can be summarised by the points:

- Taylor expansion
- matching first and second moment and control on some higher moment
- each individual variable has small influence. This is quantified by Lindeberg's condition.

Proof of Lindeberg's CLT via Lindeberg method. Let $f \in C_b^3(\mathbb{R})$. Eventually we will need to go down to $C_b(\mathbb{R})$ but this can be done by standard approximation. Define $\omega_{n,i} = X_i/s_n$ and denote

$$f_n(\omega_{n,1},\ldots,\omega_{n,n}) := f\big(\omega_{n,1}+\cdots+\omega_{n,n}\big). \tag{4.1}$$

We will also consider the i.i.d. sequence of normal variables $\xi_1, \xi_2, \ldots$ and consider $\xi_{n,i} := n^{-1/2}\xi_i$ for $i = 1, 2, \ldots$. By the definition of weak convergence, it suffices to show that

$$\mathrm{E}\big[f_n(\omega_{n,1},\ldots,\omega_{n,n})\big] \xrightarrow[n\to\infty]{} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x)\, e^{-\frac{x^2}{2}}\,\mathrm{d}x,$$

and since this limit is trivially valid for $\mathrm{E}\big[f_n(\xi_{n,1},\ldots,\xi_{n,n})\big]$, it suffices to show that

$$\Big|\mathrm{E}\big[f_n(\omega_{n,1},\ldots,\omega_{n,n})\big] - \mathrm{E}\big[f_n(\xi_{n,1},\ldots,\xi_{n,n})\big]\Big| \xrightarrow[n\to\infty]{} 0. \tag{4.2}$$

The perturbation argument alluded to in the above remarks will be done through a telescoping argument, where we will successively change the array $(\omega_{n,1},\ldots,\omega_{n,n})$ one by one, until we change all the array to $(\xi_{n,1},\ldots,\xi_{n,n})$. In this way, we can bound the left hand side of (4.2) by

$$\begin{aligned}\sum_{i=1}^{n} \Big| \,\mathrm{E}\big[f_n(\,\xi_{n,1}\,,\ldots,\,\xi_{n,i-1}\,,\,\xi_{n,i}\,,\,\omega_{n,i+1}\,,\ldots,\,\omega_{n,n}\,)\big] \;-\\ -\; \mathrm{E}\big[f_n(\,\xi_{n,1}\,,\ldots,\,\xi_{n,i-1}\,,\,\omega_{n,i}\,,\,\omega_{n,i+1}\,,\ldots,\,\omega_{n,n}\,)\big]\Big|,\end{aligned} \tag{4.3}$$

where we notice that in the above difference there is only a discrepancy at the i^{th} coordinate. We will Taylor expand in that coordinate. For this, let us introduce, for a sequence $x = (x_1,\ldots,x_n)$, the function

$$h^x_{n,i}(y) := f_n(x_1,\ldots,x_{i-1},y,x_{i+1},\ldots,x_n).$$

The Taylor expansion is as follows:

$$h^x_{n,i}(y) = h^x_{n,i}(0) + \Big(\partial_y h^x_{n,i}(0)\Big)\, y + \frac{1}{2}\Big(\partial_y^2 h^x_{n,i}\Big)(0)\, y^2 + R^x_{n,i}(y), \tag{4.4}$$

where the remainder term has the expression

$$R^x_{n,i}(y) = \frac{1}{2}\int_0^y \Big(\partial_y^3 h^x_{n,i}(t)\Big)\,(y-t)^2\mathrm{d}t, \tag{4.5}$$

and the following two bounds hold:

$$\big|\,R^x_{n,i}(y)\,\big| \leq \frac{1}{6}\|\,\partial_y^3 h^x_{n,i}\,\|_\infty\, |y|^3 = \frac{1}{6}\|f'''\|_\infty |y|^3 \tag{4.6}$$

$$\big|\,R^x_{n,i}(y)\,\big| \leq \|\,\partial_y^2 h^x_{n,i}\,\|_\infty\, y^2 = \|f''\|_\infty y^2. \tag{4.7}$$

The first bound follows by bounding $\partial_y^3 h^x_{n,i}$ in (4.5) by its supremum norm, while for the second bound we first perform an integration by parts and write the remainder as

$$R^x_{n,i}(y) = -\frac{1}{2}\partial_y^2 h^x_{n,i}(0)y^2 + \int_0^y \partial_y^2 h^x_{n,i}(t)(y-t)\,\mathrm{d}t,$$

and then bound the $\partial_y^2 h^x_{n,i}$ by its supremum norm. Let us introduce the notation

$$[\xi,\omega]_i := (\xi_{n,1},\dots,\xi_{n,i-1},\omega_{n,i+1},\dots,\omega_{n,n}),$$

then each difference (4.3) writes as

$$\begin{aligned} h^{[\xi,\omega]_i}_{n,i}(\xi_i) - h^{[\xi,\omega]_i}_{n,i}(\omega_i) &= \Big\{ h^{[\xi,\omega]_i}_{n,i}(0) + \Big(\partial_y h^{[\xi,\omega]_i}_{n,i}(0)\Big)\,\xi_{n,i} + \frac{1}{2}\Big(\partial_y^2 h^{[\xi,\omega]_i}_{n,i}\Big)(0)\,\xi^2_{n,i} + R^x_{n,i}(\xi_{n,i}) \Big\} \\ &\quad - \Big\{ h^{[\xi,\omega]_i}_{n,i}(0) + \Big(\partial_y h^{[\xi,\omega]_i}_{n,i}(0)\Big)\,\omega_{n,i} + \frac{1}{2}\Big(\partial_y^2 h^{[\xi,\omega]_i}_{n,i}\Big)(0)\,\omega^2_{n,i} + R^x_{n,i}(\omega_{n,i}) \Big\} \end{aligned} \tag{4.8}$$

Taking expectation and using the *independence between the variables* we express

$$\mathrm{E}\big[\big(\partial_y h^{[\xi,\omega]_i}_{n,i}(0)\big)\,\xi_{n,i}\big] = \mathrm{E}\big[\big(\partial_y h^{[\xi,\omega]_i}_{n,i}(0)\big)\big]\,\mathrm{E}\big[\xi_{n,i}\big],$$

and similarly for the rest of the terms in (4.8). Using the assumption that $\xi_{n,i}$'s and $\omega_{n,i}$'s have matching first and second moments, we have

$$\mathrm{E}\Big[h^{[\xi,\omega]_i}_{n,i}(\xi_i)\Big] - \mathrm{E}\Big[h^{[\xi,\omega]_i}_{n,i}(\omega_i)\Big] = \mathrm{E}\Big[R^{[\xi,\omega]_i}_{n,i}(\omega_{n,i})\Big] - \mathrm{E}\Big[R^{[\xi,\omega]_i}_{n,i}(\xi_{n,i})\Big]$$

So (4.3) is bounded by

$$\sum_{i=1}^n \mathrm{E}\Big[\Big|\,R^{[\xi,\omega]_i}_{n,i}(\omega_{n,i})\,\Big|\Big] + \sum_{i=1}^n \mathrm{E}\Big[\Big|\,R^{[\xi,\omega]_i}_{n,i}(\xi_{n,i})\,\Big|\Big].$$

We will estimate the first term, the second one being identical. For this, we denote by $\mathsf{C}_f := \max\{\|f''\|_\infty, \|f'''\|_\infty\}$ and we have by estimates (4.6), (4.7) that

$$\begin{aligned} \sum_{i=1}^n \mathrm{E}\Big[\Big|\,R^{[\xi,\omega]_i}_{n,i}(\omega_{n,i})\,\Big|\Big] &\le \mathsf{C}_f \sum_{i=1}^n \mathrm{E}\Big[\min\{\omega^2_{n,i}, \tfrac{1}{6}|\omega_{n,i}|^3\}\Big] \\ &= \mathsf{C}_f \sum_{i=1}^n \mathrm{E}\Big[\min\{\omega^2_{n,i}, \tfrac{1}{6}|\omega_{n,i}|^3\}\,;\,|\omega_{n,i}|\ge\varepsilon\Big] \\ &\qquad + \mathsf{C}_f \sum_{i=1}^n \mathrm{E}\Big[\min\{\omega^2_{n,i}, \tfrac{1}{6}|\omega_{n,i}|^3\}\,;\,|\omega_{n,i}|<\varepsilon\Big] \\ &\le \mathsf{C}_f \sum_{i=1}^n \mathrm{E}\Big[\omega^2_{n,i};\,|\omega_{n,i}|\ge\varepsilon\Big] + \tfrac{\varepsilon}{6}\mathsf{C}_f \sum_{i=1}^n \mathrm{E}\Big[|\omega_{n,i}|^2\Big], \end{aligned}$$

and the first term converges to zero by the Lindeberg assumption, while the second can be made arbitrarily small by choosing ε small enough. ☐

4.3. STEIN'S METHOD.

Stein's method is another powerful method to prove CLTs (there is also a generalisation of it for Poisson convergence). Here I will follow the lecture notes of Chatterjee [C].

The key idea. Stein's method is based on the following elementary observation:

Lemma 4.3 (Stein's Lemma). *Let $Z \sim \mathcal{N}(0,1)$ and $f\colon \mathbb{R}\to\mathbb{R}$ an absolutely continuous functions such that $\mathrm{E}[|f'(Z)|] < \infty$. Then*

$$\mathrm{E}\big[Zf(Z)\big] = \mathrm{E}\big[f'(Z)\big].$$

Proof. The proof is integration by parts. We refer to [C], Section 3. ☐

Let us now describe the idea behind Stein's method. Let W be a generic random variable and $Z \sim \mathcal{N}(0,1)$. We want to now "how far" is the distribution of W from the normal. Motivated by weak convergence, we would like to control:

$$\sup_{g \in \mathcal{D}} \big| \mathrm{E}\big[g(W)\big] - \mathrm{E}\big[g(Z)\big] \big|,$$

for $\mathcal{D}$ a certain class of functions.Weak convergence would require that $\mathcal{D}$ is the family of bounded continuous functions. However, this might be more difficult to control and we might need to resort to subclasses. We will discuss these and the metrics they are associated with later on. For the moment let us continue with Stein's approach: Given g, suppose we can find $f \in \mathcal{D}'$, with $\mathcal{D}'$ another class of functions, such that

$$g(x) - \mathrm{E}[g(Z)] = f'(x) - xf(x), \qquad \text{for all } x.$$

Then we would also have, by setting $x = W$ and taking expectation with respect to W:

$$\mathrm{E}g(W) - \mathrm{E}g(Z) = \mathrm{E}\big[f'(W) - Wf(W)\big] \implies \sup_{g \in \mathcal{D}} \big| \mathrm{E}\big[g(W)\big] - \mathrm{E}\big[g(Z)\big] \big| \le \sup_{f \in \mathcal{D}'} \big| \mathrm{E}\big[f'(W) - Wf(W)\big] \big|.$$

If, now, W was normal, then the RHS would be 0 (in fact, both sides would be 0 but the emphasis is on the RHS). If, now, we had a sequence $(W_n)_{n \ge 1}$ such that such that $\sup_{f \in \mathcal{D}'} \big| \mathrm{E}\big[f'(W_n) - Wf(W_n)\big] \big| \to 0$, then the above inequality would also imply that $\sup_{g \in \mathcal{D}} \big| \mathrm{E}\big[g(W_n)\big] - \mathrm{E}\big[g(Z)\big] \big| \to 0$, which means that W_n would converge (in a suitable sense dictated by the family $\mathcal{D}$ to a normal.

Let us formalise this idea. We first need to determine distributional distances, which would be more suitable for Stein's method.

Distributional distances. Let μ, ν be probability distributions on $\mathbb{R}$. We have the following distances:

- **Kolmogorov distance.** This is very much associated to weak convergence. It is defined as
$$d_{\text{Kolm.}}(\mu, \nu) := \sup\Big\{ \int f \mathrm{d}\mu - \int f \mathrm{d}\nu \colon f \in C_b(\mathbb{R}) \Big\}.$$
- **Total Variation distance.** This is defined as
$$\mathrm{TV}(\mu, \nu) := \sup\Big\{ \mu(A) - \nu(A) \colon A \in \mathcal{B}(\mathbb{R}) \Big\}.$$
The total variation distance admits also a very useful "coupling" formula:
$$\mathrm{TV}(\mu, \nu) := \sup\Big\{ \mathbb{P}(X \neq Y) \colon \mathbb{P} \sim (X, Y), X \sim \mu,\, Y \sim \nu \Big\},$$
where the supremum is over all possible joint distribution of variables (X, Y) on $\mathbb{R} \times \mathbb{R}$ such that the marginal distribution of X is μ and the marginal distribution of Y is ν.
- **Wasserstein distance.** This is defined as
$$\begin{aligned} \mathrm{Wass}(\mu, \nu) &:= \sup\Big\{ \int f \mathrm{d}\mu - \int f \mathrm{d}\nu \colon f \text{ is 1-Lipschitz} \Big\} \\ &= \sup\Big\{ \mathbb{E}|X - Y| \colon X \sim \mu,\, Y \sim \nu \Big\}. \end{aligned}$$

Exercise 17. *Show that the above distances are stronger than the weak convergence. That is, if* $\mu_n \xrightarrow[n \to \infty]{D} \nu$ *in one of the above distances* D*, then* $\mu_n \implies \nu$.

Exercise 18. *Show that the total variation distance is too strong through the following example: Let* $X_i = \pm 1$ *i.i.d. random variables and* $Z \sim \mathcal{N}(0,1)$*. Then* $TV\Big(\frac{S_n}{\sqrt{n}}, Z\Big) = 1$*, while* $\frac{S_n}{\sqrt{n}} \implies Z$.

Stein's method via an example. We will demonstrate Stein's method via the following example:

Proposition 4.4. *Let* $(X_n)_{n \ge 1}$ *be i.i.d. variable with mean* 0 *and variance* 1 *and* $\mathrm{E}[|X_i|^3] < \infty$*. Then*

$$\mathrm{Wass}\Big(\frac{S_n}{\sqrt{n}}, Z\Big) \le \frac{3}{n^{3/2}} \sum_{i=1}^{n} \mathrm{E}[|X_i|^3] = \frac{3}{\sqrt{n}} \mathrm{E}[|X_1|^3].$$

The above theorem is obviously not optimal in terms of condition for a CLT. However, the third moment assumption is needed to obtain an estimate on the rate of convergence. Obtaining control on the rate of convergence on the CLT is the content of the Berry-Esseen Theorem. The above proposition is a bit weaker than the Berry-Esseen theorem as it expresses the rate in terms of the Wasserstein distance instead of the Kolmogorov distance. One can also obtain the optimal rate of convergence via Stein's method but we will not show it here; see the discussion in [C], Section 5.

To prove Proposition 4.4 we need some lemmas:

Lemma 4.5. *For any $g\colon \mathbb{R}\to\mathbb{R}$ bounded, there exists an absolutely continuous f such that*

$$f'(x) - xf(x) = g(x) - \mathrm{E}\big[g(Z)\big] \qquad \text{with}$$
$$\|f\|_\infty \le \frac{\pi}{2}\|g - \mathrm{E}\big[g(Z)\|_\infty \quad \text{and} \quad \|f'\|_\infty \le 2\|g - \mathrm{E}\big[g(Z)\|_\infty. \tag{4.9}$$

If g is Lipschitz (but not necessarily bounded), then it also holds that

$$\|f\|_\infty \le \|g'\|_\infty, \quad \|f'\|_\infty \le \sqrt{\frac{2}{\pi}}\|g'\|_\infty, \quad \|f''\|_\infty \le 2\|g'\|_\infty, \tag{4.10}$$

Proof. We will refer to [C], Lecture 4 for the details. Here we will just record two expression for the solution f. The first one is by simply solving the ODE and reads as

$$f(x) = e^{\frac{x^2}{2}}\int_{-\infty}^{x}\big(g(y) - \mathrm{E}[g(Z)]\big)\,\mathrm{d}y.$$

The second is more unusual and interesting. It bears the important idea of *Gaussian interpolation.* The solution reads as

$$f(x) = -\int_0^1 \frac{1}{2\sqrt{t(1-t)}}\mathrm{E}\big[Z\,g\big(\sqrt{t}x + \sqrt{1-t}Z\big)\big]\,\mathrm{d}t.$$

Having these two expressions at hand the estimates of the proposition follow via standard calculus. We refer to [C] for details. ☐

Let us now prove Proposition 4.4.

Proof of Proposition 4.4. Recall the Wasserstein distance. We then have

$$\mathrm{Wass}\Big(\tfrac{S_n}{\sqrt{n}}, Z\Big) = \sup\Big\{\mathrm{E}\big[g\big(\tfrac{S_n}{\sqrt{n}}\big)\big] - \mathrm{E}[g(Z)]\colon g \text{ is 1-Lipschitz}\Big\}$$

By Lemma 4.5 we have that there exists an absolutely continuous f satisfying condition (4.10). Let's denote the class of functions which satisfy (4.10) by $\mathcal{D}'$. We then have that

$$g\big(\tfrac{S_n}{\sqrt{n}}\big) - \mathrm{E}[g(Z)] = f'\big(\tfrac{S_n}{\sqrt{n}}\big) - \tfrac{S_n}{\sqrt{n}}f\big(\tfrac{S_n}{\sqrt{n}}\big)$$

and so

$$\mathrm{Wass}\Big(\tfrac{S_n}{\sqrt{n}}, Z\Big) \le \sup\Big\{\mathrm{E}\Big[f'\big(\tfrac{S_n}{\sqrt{n}}\big) - \tfrac{S_n}{\sqrt{n}}f\big(\tfrac{S_n}{\sqrt{n}}\big)\Big]\colon f\in\mathcal{D}'\Big\}. \tag{4.11}$$

We will now estimate

$$\mathrm{E}\Big[f'\big(\tfrac{S_n}{\sqrt{n}}\big) - \tfrac{S_n}{\sqrt{n}}f\big(\tfrac{S_n}{\sqrt{n}}\big)\Big].$$

First we denote by

$$S_n^i := \sum_{1\le j\le n,\, j\neq i} X_j,$$

and write

$$\begin{aligned}
\mathrm{E}\Big[\tfrac{S_n}{\sqrt{n}}f\big(\tfrac{S_n}{\sqrt{n}}\big)\Big] &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{E}\Big[X_i f\big(\tfrac{S_n}{\sqrt{n}}\big)\Big]\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{E}\Big[X_i\,\Big(f\big(\tfrac{S_n}{\sqrt{n}}\big) - f\big(\tfrac{S_n^i}{\sqrt{n}}\big)\Big)\Big],
\end{aligned}$$

where in the second equality we used the independence between X_i and S_n^i and the fact that $\mathrm{E}[X_i] = 0$. We will next Taylor expand ! To keep notation short we introduce

$$W_n := \frac{S_n}{\sqrt{n}} \qquad \text{and} \qquad W_n^i := \frac{S_n^i}{\sqrt{n}}.$$

We now proceed:

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathrm{E}\Big[X_i\big(f(W_n) - f(W_n^i)\big)\Big] &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathrm{E}\Big[X_i\Big(f(W_n) - f(W_n^i) - (W_n - W_n^i) f'(W_n^i)\Big)\Big] \\ &\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathrm{E}\Big[(W_n - W_n^i) f'(W_n^i))\Big] \end{aligned} \tag{4.12}$$

Let us control the first term:

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathrm{E}\Big[|X_i| \, |f(W_n) - f(W_n^i) - (W_n - W_n^i) f'(W_n^i)|\Big] &\le \frac{1}{2\sqrt{n}} \sum_{i=1}^n \mathrm{E}\Big[\, |X_i| \, (W_n^i - W_n)^2\Big] \, \|f''\|_\infty \\ &= \frac{1}{2n^{3/2}} \sum_{i=1}^n \mathrm{E}\Big[\, |X_i|^3\Big] \, \|f''\|_\infty \\ &= \frac{1}{2n^{1/2}} \mathrm{E}\Big[\, |X_1|^3\Big] \, \|f''\|_\infty \end{aligned} \tag{4.13}$$

The second term is:

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathrm{E}\Big[X_i \, (W_n - W_n^i) \, f'(W_n^i)\Big] &= \frac{1}{n} \sum_{i=1}^n \mathrm{E}\Big[X_i^2 \, f'(W_n^i)\Big] \\ &= \frac{1}{n} \sum_{i=1}^n \mathrm{E}\Big[X_i^2\Big] \, \mathrm{E}\Big[f'(W_n^i)\Big] \\ &= \frac{1}{n} \sum_{i=1}^n \mathrm{E}\Big[X_i^2\Big] \, \mathrm{E}\Big[f'(W_n^i) - f'(W_n)\Big] + \frac{1}{n} \sum_{i=1}^n \mathrm{E}\Big[f'(W_n)\Big] \\ &= \frac{1}{n} \sum_{i=1}^n \mathrm{E}\Big[X_i^2\Big] \, \mathrm{E}\Big[f'(W_n^i) - f'(W_n)\Big] + \mathrm{E}\Big[f'(W_n)\Big], \end{aligned} \tag{4.14}$$

where in the last we used that $\mathrm{E}\Big[X_i^2\Big] = 1$. The first in (4.14) is control by Taylor expansion as:

$$\frac{1}{n} \sum_{i=1}^n \mathrm{E}\Big[|f'(W_n^i) - f'(W_n)|\Big] \le \frac{\|f''\|_\infty}{n} \sum_{i=1}^n \mathrm{E}\Big[|W_n^i - W_n|\Big] = \frac{\|f''\|_\infty}{n^{3/2}} \sum_{i=1}^n \mathrm{E}\Big[|X_i|\Big] = \frac{\|f''\|_\infty}{n^{1/2}} \tag{4.15}$$

Using (4.12), (4.14), (4.15) as well as bounds (4.10), we have that

$$\Big|\mathrm{E}\Big[f'\big(\tfrac{S_n}{\sqrt{n}}\big) - \tfrac{S_n}{\sqrt{n}} f\big(\tfrac{S_n}{\sqrt{n}}\big)\Big]\Big| \le \frac{3}{n^{1/2}} E\Big[|X_i|^3\Big]$$

□

Exercise 19. *(*) Consider a graph G with n vertices, where each edge is added with probability p and omitted with probability $1 - p$ independently of all other edges. Let T_n be the number of triangles in the graph. Use Stein's method to show a CLT for T_n.*

5. The Local Limit Theorem

6. Stable and infinitely divisible laws

6.1. Poisson Convergence.

We will introduce the Poisson convergence. We will start by drawing a contrast with the CLT. In short,

the CLT regime is when all random variables are small while
Poisson convergence holds holds when occasionally a random variable takes a large value.

To make this contrast more clear, let us start with the following computation:

Proposition 6.1. *Let $(X_n)_{n\geq 1}$ be i.i.d. variable with mean 0 and variance 1 and $X_{n,i} := \frac{1}{\sqrt{N}}X_i$. Then*

$$M_n := \max_{i\leq n} |X_{n,i}| \xrightarrow[n\to\infty]{\mathrm{P}} 0.$$

Proof. Compute

$$\begin{aligned}\mathrm{P}\big(M_n<\varepsilon\big) &= \mathrm{P}\Big(|X_{n,i}|<\varepsilon,\ i=1,...,n\Big) = \mathrm{P}\Big(|X_{n,1}|<\varepsilon\Big)^n = \Big(1-\mathrm{P}\Big(|X_1|\geq\varepsilon\sqrt{n}\Big)\Big)^n\\ &\geq \Big(1-\frac{1}{n\varepsilon^2}\mathrm{E}\Big(X_1^2\,\mathbb{1}_{|X_1|\geq\varepsilon\sqrt{n}}\Big)\Big)^n\\ &\approx \exp\Big(\frac{1}{\varepsilon^2}\mathrm{E}\Big(X_1^2\,\mathbb{1}_{|X_1|\geq\varepsilon\sqrt{n}}\Big)\Big)\\ &\to 1,\end{aligned}$$

by dominated convergence and existence of second moments. □

In the above proposition we where in the CLT regime. On the other hand, we have:

Proposition 6.2. *Let $(X_{n,i})_{1\leq i\leq n})$ independent with $\mathrm{P}(X_{n,i}=0)=p_n$ and $\mathrm{P}(X_{n,i}=1)=1-p_n$ with $p_n\to 0$ and $np_n\to\lambda>0$ when $n\to\infty$. Denote $M_n:=\max_{i\leq n}X_{n,i}$. Then $\mathrm{P}\big(M_n=0\big)\to e^{-\lambda}$, as $n\to\infty$.*

Proof.

$$\mathrm{P}\big(M_n=0\big)=\mathrm{P}\big(X_{n,i}=0,\,i=1,...,n\big)=\mathrm{P}\big(X_{n,1}=0\big)^n=(1-p_n)^n\approx e^{-np_n}\to e^{-\lambda}$$

□

This raises the question what is the limiting distribution of $S_n = X_{n,1}+\cdots+X_{n,n}$ when $X_{n,i}$ take values 0 or 1 with probabilities p_n and $1-p_n$ respectively. We have:

Proposition 6.3. *Let $(X_{n,i})_{1\leq i\leq n})$ independent with $\mathrm{P}(X_{n,i}=0)=p_n$ and $\mathrm{P}(X_{n,i}=1)=1-p_n$ with $p_n\to 0$ and $np_n\to\lambda>0$ when $n\to\infty$. Then S_n converges weakly to a Poisson distribution with parameter λ. That is, to a random variable X with $\mathrm{P}(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}$, for $k=0,1,...$*

Proof. Compute the characteristic function:

$$\mathrm{E}\big[e^{\mathrm{i}tS_n}\big]=\mathrm{E}\big[e^{\mathrm{i}tX_{n,1}}\big]^n=\big(1-p_n+e^{\mathrm{i}t}p_n\big)^n=\big(1+\big(e^{\mathrm{i}t}-1\big)p_n\big)^n\approx e^{np_n(e^{\mathrm{i}t}-1)}\to e^{\lambda(e^{\mathrm{i}t}-1)},$$

since $np_n\to\lambda$ as $n\to\infty$. Taylor expanding $e^{\mathrm{i}t\lambda}$ we get that

$$e^{np_n(e^{\mathrm{i}t}-1)}=e^{-\lambda}\sum_{k\geq 0}e^{\mathrm{i}kt}\frac{\lambda^k}{k!}=\sum_{k\geq 0}e^{\mathrm{i}kt}\,\mathrm{Poiss}_\lambda(X=k),$$

that is equal to the characteristic function of a Poisson variable with parameter λ. □

Let us now start generalising the Poisson convergence. The first step is to consider jump distribution which can take values different that 1. In particular consider

$$\mathrm{P}(X_{n,1}=x_j)=p_{n,j}, \quad \text{with } \ x_j\neq 0 \ \text{ for } j=1,...,k,$$
$$\text{and } \ \mathrm{P}(X_{n,1}=0)=1-\sum_{j=1}^{k} p_{n,j}. \tag{6.1}$$

Then a similar computation as above shows that

$$\mathrm{E}\big[e^{\mathrm{i}tS_n}\big] \xrightarrow[n\to\infty]{} \exp\Big(-\sum_{j=1}^{k}(e^{\mathrm{i}tx_j}-1)\lambda_j\Big)=\exp\Big(-\int(e^{\mathrm{i}tx}-1)\mu(\mathrm{d}x)\Big), \tag{6.2}$$

where $\mu(\mathrm{d}x)=\sum_{j=1}^{k}\lambda_j\delta_{x_j}(\mathrm{d}x)$. This will be the **jump measure** of the process. We also note that

$$\exp\Big(-\sum_{j=1}^{k}(e^{\mathrm{i}tx_j}-1)\lambda_j\Big)=\prod_{j=1}^{k}\exp\Big(-(e^{\mathrm{i}tx_j}-1)\lambda_j\Big),$$

that is, the limiting distribution is a sum of independent Poisson variables with intensity λ_j and jump variable x_j.

6.2. INFINITELY DIVISIBLE LAWS. In this section we will classify the possible limits of sums of independent random variables $S_n=X_{n,1}+\cdots+X_{n,n}$ and we will see that this are mixtures of a Gaussian distribution and a jump process. The latter can be thought of as a Poisson jump process only that jumps can be infinitesimally small. We will eventually show that the limiting characteristic function of S_n takes the form

$$\exp\Big(-\mathrm{i}\mu t-\frac{\sigma^2t^2}{2}+\int_{\mathbb{R}}\big(e^{-\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2}\big)\mu(\mathrm{d}x)\Big). \tag{6.3}$$

The first two terms correspond to the characteristic function of a Gaussian distribution. The integral part generalises (6.2) and it corresponds to the jump process. The measure $\mu(\mathrm{d}x)$ will be the jump measure. In the case of (6.2) this was a *pure jump* measure, that is the jumps were having positive size. In the general case the jumps could be infinitesimally small.

Law that arise as limits of sums of independent random variables have a particular structure, which goes under the name *infinitely divisible laws.* Let us give the formal definition

Definition 6.4. *A distribution α is called* **infinitely divisible** *if any random variable X with distribution α can be written as a sum of n i.i.d. variables, for any n. That is, for any n, we can write $X=X_{n,1}+\cdots+X_{n,n}$ with $(X_{n,i})_{i=1,\dots,n}$ i.i.d.*

We have

Proposition 6.5. *Let $(X_{n,i})1\le i\le n$ be i.i.d. random variables for any n. Then any possible limit of $S_n=X_{n,1}+\cdot+X_{n,n}$ is infinitely divisible. Conversely, any infinitely divisible law arises as such a limit.*

Proof. Assume that X is infinitely divisible. Then for any n we can find i.i.d. random variables $(X_{n,i})_{i\le n}$ such that

$$X=X_{n,1}+\cdots+X_{n,n}=:S_n.$$

So it trivially follows that $S_n\implies X$.

Conversely, assume that $S_n\implies X$. Let us show that X is infinitely divisible. Let us start by writing

$$S_{2n}=\big(X_{2n,1}+\cdots+X_{2n,n}\big)+\big(X_{2n,n+1}+\cdots+X_{2n,2n}\big)=:Y_n+Y_n'.$$

Obviously Y_n and Y_n' are independent and have the same distribution. If S_{2n} converges, we would like to conclude that Y_n and Y_n' also converge. Let us show this.

First, let us show that Y_n and Y_n' are tight. We have, for any $y>0$:

$$\mathrm{P}\big(Y_n>y\big)^2=\mathrm{P}\big(Y_n>y,Y_n'>y\big)\le\mathrm{P}\big(S_{2n}>y\big)$$

Then

$$\limsup_{n\to\infty} \mathrm{P}\big(Y_n > y\big)^2 \le \limsup_{n\to\infty} \mathrm{P}\big(S_{2n} > y\big) = \mathrm{P}\big(X > y\big),$$

and so

$$\lim_{y\to\infty} \limsup_{n\to\infty} \mathrm{P}\big(Y_n > y\big)^2 = \lim_{y\to\infty} \mathrm{P}\big(X > y\big) = 0.$$

Similarly, we have that

$$\lim_{y\to\infty} \limsup_{n\to\infty} \mathrm{P}\big(Y_n < -y\big)^2 = \lim_{y\to\infty} \mathrm{P}\big(X < -y\big) = 0.$$

This shows that, indeed, the sequence Y_n and Y'_n is tight, which then implies that there exist subsequece n' such that $Y_{n'} \implies Y$ and $Y'_{n'} \implies Y'$. In turn this implies that $X = Y + Y'$ for Y, Y' independent and identically distributed.

We can next do the same procedure with S_{nk} for any $k \in \mathbb{N}$ and conclude that for any k we can write $X = Y_1 + \cdots + Y_k$ with $Y_1, ..., Y_k$ i.i.d. That is, X is infinitely divisible. ☐

We will gradually move towards establishing expression (6.3) as the general form of any characteristic function. A preparatory step is the following proposition

Proposition 6.6. *If $S_n := X_{n,1} + \cdots + X_{n,n}$ is a sum of i.i.d. random variables and $S_n \implies X$ for some random variable X, then $X_{n,1} \implies 0$.*

Exercise 20. *Show that this proposition is not in contradiction with Proposition 6.3.*

Proof. Let ϕ_{S_n} be the characteristic function of S_n, ϕ_X the characteristic function of X and ϕ_n the characteristic function of $X_{n,1}$. Since $S_n \implies X$, we have that $\phi_{S_n}(t) \to \phi_X(t)$ for all t. It turns out that this convergence is uniform on compact sets. This needs some justification but let us accept it for the moment. This would then imply that there exists a neighbourhood N of 0, such that

$$\mathrm{Re}\big(\phi_{S_n}(t)\big) \ge \delta > 0, \qquad \forall t \in N \text{ and } n \ge 1. \tag{6.4}$$

Moreover, we have

$$\phi_{S_n}(t) = \phi_n(t)^n,$$

which implies that

$$|\phi_n(t)| = |\phi_{S_n}(t)|^{1/n} \xrightarrow[n\to\infty]{} 1 \qquad \text{and} \qquad \mathrm{Arg}\big(\phi_n(t)\big) = \frac{1}{n}\mathrm{Arg}\big(\phi_{S_n}(t)\big) \xrightarrow[n\to\infty]{} 0$$

where the limits follow from (6.4). This means that $\phi_n(t) \xrightarrow[n\to\infty]{} 1$ for any $t \in N$, from which we conclude that $\phi_n(t) \xrightarrow[n\to\infty]{} 1$ for any t (by the following exercise), which implies that $X_{n,1} \implies 0$. ☐

Exercise 21. *If ϕ is a characteristic function, show that*

$$1 - \mathrm{Re}\Big(\phi(2t)\Big) \le 4\Big(1 - \mathrm{Re}\Big(\phi(t)\Big)\Big).$$

We will next move towards proving that the characteristic function of any infinitely divisible distribution has the form (6.3). Let us start with the heuristics:

Heuristics: Let X be an infinitely divisible random variable. Then for any n we can find i.i.d. random variables such that $X = X_{n,1} + \cdots + X_{n,n}$. Let ϕ_X and ϕ_n be the characteristic functions of X and $X_{n,1}$, respectively. The

$$\log \phi_X(t) = n \log \phi_n(t) = n \log \big(1 + \phi_n(t) - 1\big) = n(\phi_n(t) - 1)\,(1 + \varepsilon_n(t)),$$

where we use the previous proposition that $\phi_n(t)$ converges to 1 uniformly on compact sets and so $\varepsilon_n(t) \to 0$ uniformly on compact sets. Continuing, we can write

$$\log \phi_X(t) = \int_{\mathbb{R}} (e^{\mathrm{i}tx} - 1) n\alpha_n(\mathrm{d}x) \cdot (1 + \varepsilon_n(t)) = \sum_{i=1}^{k} \int_{B_i} (e^{\mathrm{i}tx} - 1) n\alpha_n(\mathrm{d}x) \cdot (1 + \varepsilon_n(t)),$$

where we have chosen $(B_i)_{i=1}^k$ to be a disjoint partition of $\mathbb{R}$. Suppose we can restrict to compact subset of $\mathbb{R}$ and then take $B_i = [x_i, x_i + \delta)$ for small enough δ. Then

$$\log \phi_X(t) \approx \sum_i (e^{\mathrm{i}tx_i} - 1)\, n\alpha_n(B_i) \qquad \text{or} \qquad \phi_X(t) \approx \exp\Big(\sum_i (e^{\mathrm{i}tx_i} - 1)\, n\alpha_n(B_i)\Big)$$

which has the form of a superposition of independent Poisson jumps with rate $\lambda_i := n\alpha_n(B_i)$.

But there are a few issues to address:

- $n\alpha_n(\mathbb{R}) \to \infty$ as $n \to \infty$. So what is actually the meaning of the limiting distribution of $n\alpha_n$?
- We said that the general infinitely divisible law can have a Gaussian component. But we don't see this in the above heuristics.

Spoiler with regards to the second bullet : the Gaussian component will arise as a point mass at zero (i.e. "no jump component") $n\alpha_n \implies \delta_0$.

Let us also note the following

$$n\alpha_n(B) = n\mathrm{E}\big[\mathbb{1}_{X_{n,1} \in B}\big] = \mathrm{E}\big[\sum_{i=1}^{n} \mathbb{1}_{X_{n,i} \in B}\big] = \mathrm{E}\big[\sharp i \colon X_{n,i} \in B\big],$$

that is the expected number of jumps whose size is in the range of B.

To make the above heuristics rigorous we will need the following two lemmas:

Lemma 6.7. *Let $\mu_n := n\alpha_n$. Then for any $a > 0$:*

$$\limsup_n \mu_n\big([-a, a]^c\big) \le Ca \int_0^{1/a} \big|\mathrm{Re} \log \phi_X(t)\big| \, \mathrm{d}t.$$

Proof. This is similar to the proof of inequality (2.3). Do it as an exercise ☐

Lemma 6.8. *Let $\mu_n := n\alpha_n$. Then for any $a > 0$:*

$$\limsup_n \int_{-1}^{1} x^2 \mu_n(\mathrm{d}x) < \infty.$$

Proof. We have that

$$n\big(1 - \mathrm{Re}(\phi_n(1))\big) = \int_{\mathbb{R}} (1 - \cos(x))\, \mu_n(\mathrm{d}x) \ge \int_{[-1,1]} (1 - \cos(x))\, \mu_n(\mathrm{d}x) \ge c \int_{[-1,1]} x^2\, \mu_n(\mathrm{d}x),$$

which then implies that

$$\limsup_n \int_{[-1,1]} x^2\, \mu_n(\mathrm{d}x) \le c^{-1} \mathrm{Re}\big(\log \phi_X(1)\big) < \infty.$$

☐

Let us assume these two Lemmas and redo the above computation rigorously. We use the notation $\mu_n = n\alpha_n$.

$$\begin{aligned}
\log\phi_X(t) &= \lim_{n\to\infty}\int_{\mathbb{R}}(e^{\mathrm{i}tx}-1)n\alpha_n(\mathrm{d}x)\cdot(1+\varepsilon_n(t))\\
&= \lim_{n\to\infty}\int_{\mathbb{R}}(e^{\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2})\mu_n(\mathrm{d}x)+\mathrm{i}t\lim_{n\to\infty}\int_{\mathbb{R}}\frac{x}{1+x^2}\mu_n(\mathrm{d}x)\\
&= \lim_{n\to\infty}\int_{\mathbb{R}}(e^{\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2})\frac{1+x^2}{x^2}\frac{x^2}{1+x^2}\mu_n(\mathrm{d}x)+\mathrm{i}t\lim_{n\to\infty}\int_{\mathbb{R}}\frac{x}{1+x^2}\mu_n(\mathrm{d}x)\\
&=: \lim_{n\to\infty}a_n\int_{\mathbb{R}}(e^{\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2})\frac{1+x^2}{x^2}G_n(\mathrm{d}x)+\mathrm{i}t\lim_{n\to\infty}\int_{\mathbb{R}}\frac{x}{1+x^2}\mu_n(\mathrm{d}x),
\end{aligned}$$

where we have defined

$$G_n(\mathrm{d}x) := \frac{1}{a_n}\frac{x^2}{1+x^2}\mu_n(\mathrm{d}x), \quad \text{and} \quad a_n = \int_{\mathbb{R}}\frac{x^2}{1+x^2}\mu_n(\mathrm{d}x).$$

By Lemmas 6.7 and 6.8 we have that $\sup_n a_n < \infty$ and so that G_n is a well defined probability measure. Lemma 6.7 also shows that the family $(G_n)_{n\geq}$ is tight and so along a subsequence n' we have that $G_{n'} \Longrightarrow G$ for some probability distribution G. Moreover since (a_n) is a bounded real sequence, we can assume that it converges to some a along the same subsequence (n').

Since $(e^{\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2})\frac{1+x^2}{x^2}$ is a bounded sequence, tightness will imply that along (n'), we have that

$$a_n\int_{\mathbb{R}}(e^{\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2})\frac{1+x^2}{x^2}G_n(\mathrm{d}x)\xrightarrow[n\to\infty]{} a\int_{\mathbb{R}}(e^{\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2})\frac{1+x^2}{x^2}G(\mathrm{d}x)$$

Therefore,

$$\log\phi_X(t) = a\int_{\mathbb{R}}(e^{\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2})\frac{1+x^2}{x^2}G(\mathrm{d}x)+\mathrm{i}t\lim_{n'\to\infty}\int_{\mathbb{R}}\frac{x}{1+x^2}\mu_n(\mathrm{d}x)$$

which implies that the last term above must also have a limit, which we call $\mathrm{i}t\beta$. This shows the representation

$$\log\phi_X(t) = a\int_{\mathbb{R}}(e^{\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2})\frac{1+x^2}{x^2}G(\mathrm{d}x)+\mathrm{i}t\beta$$

for the characteristic function of an infinitely divisible distribution. To recover the Gaussian component, we note that the limiting distribution $G(\mathrm{d}x)$ might have an atom $G(\{0\})$ at 0 (this would be the continuous, i.e. “no-jump” part of the distribution). Since the value of $(e^{\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2})\frac{1+x^2}{x^2}$ at 0 is $-\frac{t^2}{2}$ we have

$$\log\phi_X(t) = \mathrm{i}\beta t-\frac{aG(\{0\})t^2}{2}+a\int_{\mathbb{R}\setminus\{0\}}(e^{\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2})\frac{1+x^2}{x^2}G(\mathrm{d}x).$$

We have shown that the characteristic function of an infinitely divisible distribution is written in the form (6.3). We will now show that any such representation is the characteristic function of an infinitely divisible distribution:

Since

$$\exp\Big(-\mathrm{i}\mu t-\frac{\sigma^2t^2}{2}+\int_{\mathbb{R}}\big(e^{-\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2}\big)\mu(\mathrm{d}x)\Big) \tag{6.5}$$

$$= \exp\Big(-\mathrm{i}\mu t-\frac{\sigma^2t^2}{2}\Big)\exp\Big(\int_{\mathbb{R}}\big(e^{-\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2}\big)\mu(\mathrm{d}x)\Big). \tag{6.6}$$

and the first exponential corresponds to the characteristic function of a Gaussian, it suffices to show the statement for the second exponential. To this end, take M large and partition $[-M,M]=\sum_j[x_j,x_j+\varepsilon)$

for ε small enough and approximate

$$\exp\Big(\int_{\mathbb{R}}\big(e^{-\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2}\big)\mu(\mathrm{d}x)\Big)\approx\exp\Big(\sum_j\int_{I_j}\big(e^{-\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2}\big)\mu(\mathrm{d}x)\Big) \tag{6.7}$$

$$\approx\exp\Big(\sum_j\big(e^{-\mathrm{i}tx_j}-1\big)\mu(I_j)-\mathrm{i}t\sum_j\frac{x_j}{1+x_j^2}\mu(I_j)\Big) \tag{6.8}$$

considering the second term as a constant $\beta=\sum_j\frac{x_j}{1+x_j^2}\mu(I_j)$, the second term can be considered as a constant shift. On the other hand,

$$\exp\Big(\sum_j\big(e^{-\mathrm{i}tx_j}-1\big)\mu(I_j)\Big) \tag{6.9}$$

can be viewed as the characteristic function of a sum $S_{n(\varepsilon)}:=X_{n(\varepsilon),1}+\cdots+X_{n(\varepsilon),n(\varepsilon)}$ of a Poisson process with jump $(x_j)_{j\geq1}$ and rates $(\mu(I_j))_{j\geq1}$. By Proposition 6.5, the limiting distribution of such sums is an infinitely divisible distribution.

We can summarise all the above in the following theorem:

Theorem 6.9. *Let ϕ_X be the characteristic function of an infinitely divisible distribution. Then there exists a unique finite measure $\gamma(\mathrm{d}x)$, with possibly mass at the origin, and a constant β such that*

$$\log\phi_X(t)=\mathrm{i}t\beta+\int\varphi(x,t)\gamma(\mathrm{d}x),$$

with $\varphi(x,t)$ jointly continuous, bounded in x and bounded for t in compact intervals. In particular,

$$\varphi(x,t)=\Big(e^{-\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2}\Big)\frac{x^2}{1+x^2}.$$

If α_n is the distribution of random variable $X_{n,1}$, we can infer from the above theorem conditions for the limit of $S_n=X_{n,1}+\cdots+X_{n,n}$. If

$$\gamma_n(\mathrm{d}x):=\frac{x^2}{1+x^2}n\alpha_n(\mathrm{d}x)\quad\text{and}\quad\beta_n:=\int\frac{x}{1+x^2}n\alpha_n(\mathrm{d}x)$$

then the limit of S_n is determined by the limits $\gamma_n\implies\gamma$ and $\beta_n\to\beta$ for a measure γ and a constant β. Let us look an application of this fact:

Proposition 6.10. *Let $S_n=X_{n,1}+\cdots+X_{n,n}$ for $(X_{n,i})_{i\leq n}$ i.i.d. with distribution α_n. Denote $M_n:=\max_{1\leq i\leq n}X_{n,i}$. Then $S_n\implies\mathcal{N}(0,1)$ if and only if $M_n\implies0$.*

Proof. By the above conclusion, we look at $\gamma_n(\mathrm{d}x):=\frac{x^2}{1+x^2}n\alpha_n(\mathrm{d}x)$. The fact that $S_n\implies\mathcal{N}(0,1)$ will then be equivalent to $\gamma_n\implies\delta_0$, which then implies that for any $\varepsilon>0$ we have that

$$\gamma_n\big([-\varepsilon,\varepsilon]^c\big):=\int_{|x|\geq\varepsilon}\frac{x^2}{1+x^2}n\alpha_n(\mathrm{d}x)\xrightarrow[n\to\infty]{}0,$$

which is equivalent (use the bound $\frac{\varepsilon^2}{1+\varepsilon^2}\leq\frac{x^2}{1+x^2}\leq1$ for $|x|\geq\varepsilon$) to

$$n\alpha_n([-\varepsilon,\varepsilon]^c)\xrightarrow[n\to\infty]{}0,$$

but the left-hand side is $n\mathrm{P}(|X_{n,1}|>\varepsilon)$ and we have

$$\mathrm{P}\big(M_n<\varepsilon\big)=\mathrm{P}\big(|X_{n,i}|<\varepsilon,\ \forall i\leq n\big)=\Big(1-\mathrm{P}\big(|X_{n,1}|>\varepsilon\big)\Big)^n$$

$$\approx\exp\Big(n\mathrm{P}\big(|X_{n,1}|>\varepsilon\big)\Big)=\exp\Big(n\alpha_n([-\varepsilon,\varepsilon]^c)\Big)\xrightarrow[n\to\infty]{}1,$$

which is the desired result. ☐

Exercise 22. *Show that the Gamma distribution* $\mathrm{Gamma}(a)$ *i.e. the distribution with density* $\frac{1}{\Gamma(a)}x^{a-1}e^{-x}$ *is infinitely divisible (Hint: if* $X_i \sim \mathrm{Gamma}(a_i)$ *then* $X_1+\dots+X_n \sim \mathrm{Gamma}(\sum_{i=1}^n a_i)$*) and compute its Lévy-Khinchine represenation.*

6.3. Stable Laws. Let us start with a question:

> Let $(X_n)_{n\geq 1}$ be i.i.d. What is the class of all possible limit laws of (normed) sums
> $$\frac{X_1+\dots+X_n}{A_n}-B_n, \quad \text{for suitable normalisation } A_n, B_n?$$

The answer will be the class of **stable laws**, for which the Lévy-Khintchine represenation takes a particular form. We should highlight that our interest will be in the case that the random variables (X_n) do not have second moments, i.e. they are out of the scope of the CLT.

Let us with a definition:

Definition 6.11. *A random variable X is said to have a* **stable law** *if for any i.i.d. random variables $X_1, ..., X_k$ with distribution identical to X, there exist numbers a_k, b_k such that*

$$X_1+\dots+X_k \overset{d}{=} a_k X + b_k$$

The Gaussian distribution is a stable law by choosing $a_k = 1, b_k = 0$. The following theorem classifies stable laws. We will not provide the proof as it is quite technical but it can be found in [B].

Proposition 6.12. *X is the limit of normed sums if and only if is a stable law. Moreover, in this case, X will either be Gaussian or it characteristic function will have the form*

$$\phi_X(t) = \exp\Big(\mathrm{i}t\beta + m_1\int_0^\infty \big(e^{\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2}\big)\frac{\mathrm{d}x}{x^{1+\alpha}} + m_2\int_{-\infty}^0 \big(e^{\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2}\big)\frac{\mathrm{d}x}{|x|^{1+\alpha}}\Big)$$

for an exponent $\alpha \in (0,2)$.

The next proposition identifies the domain of attraction of Stable Laws. The proof can also be found in [B].

Proposition 6.13. *Let $(X_n)_{n\geq 1}$ i.i.d. random variables with mean 0. Then $S_n = X_1+\dots+X_n$ is in the domain of attraction of a Stable Law with parameter $\alpha \in (0,2)$, denoted* $\mathrm{Stable}(\alpha)$*, if and only if there exist $M_+, M_- \geq 0$ with $M_+ + M_- > 0$ such that*

$$(1) \qquad \lim_{y\to\infty}\frac{\mathrm{P}(X<-y)}{P(X>y)} = \frac{M_-}{M_+},$$

$$(2) \qquad M_+>0 \implies \lim_{y\to\infty}\frac{\mathrm{P}(X>\xi y)}{P(X>y)} = \xi^{-\alpha}, \quad \text{for all } \xi>0$$

$$(3) \qquad M_->0 \implies \lim_{y\to\infty}\frac{\mathrm{P}(X<-\xi y)}{P(X<y)} = \xi^{-\alpha}, \quad \text{for all } \xi>0.$$

In this case we have

$$\frac{X_1+\dots+X_n}{n^{1/\alpha}} \xrightarrow[n\to\infty]{d} \mathrm{Stable}(\alpha).$$

Conditions (2), (3) in the above theorem say that the random variables are "heavy tailed", that is, $\mathrm{P}(X>y)\sim y^{-\alpha}$ for $y\to\infty$ and analogously for $y\to-\infty$.

Exercise 23. *Find an explicit example of a distribution X where you can directly prove the convergence of*

$$\frac{X_1+\dots+X_n}{n^{1/\alpha}} \xrightarrow[n\to\infty]{d} \mathrm{Stable}(\alpha),$$

and determine the limiting characteristic function.

7. LARGE DEVIATIONS

A very good reference for large deviations is the book by den Hollander [dH] Large deviations should be thought as a principle of the theory of probability. We are already familiar with the two basic ones, the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT).

The clasical LLN says that if $(X_i)_{i\geq 1}$ is a sequence of i.i.d. variables, on a probability space $(\Omega, \mathcal{F}, P)$, with $E[X_1] = \mu$, $E[|X_1|] < \infty$, then

$$\frac{X_1 + \cdots + X_n}{n} \to \mu, \quad P - a.s.$$

The clasical CLT says that, under the additional assumption that the variance exists, say equals 1, then

$$P\left(\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}} > x\right) \cong \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy.$$

That is it tells us what is the asymptotic probability that the sum $S_n = X_1 + \cdots + X_n$ has a typical deviation from its mean $n\mu$ of order $\sqrt{n}$.

Large deviations study the asymptotic probability of a (large) deviation of S_n from the mean $n\mu$ of order n. This can be summarised by

$$P\left(X_1 + \cdots + X_n - n\mu \cong n\, x\right) \cong e^{-nI(x)}.$$

The function $I(\cdot)$ that appears is called the *rate function* and governs the asymptotics of such probabilities.

Example Let $(X_i)_{i\geq 1}$, such that $P(X_i = 1) = P(X_i = 0) = 1/2$. Then for all $a > 1/2$,

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(S_n \geq na\right) = -I(a),$$

with

$$I(z) = \begin{cases} \log 2 + z\log z + (1-z)\log(1-z) & , \quad z \in (0,1), \\ \infty & , \quad z \notin (0,1) \end{cases}$$

Proof. It is easy to verify the case that $a > 1$, so we will restrict to the case $1/2 < a < 1$. The proof goes by a direct combinatorial computation.

$$P(S_n \geq na) = \sum_{k\geq na} \binom{n}{k} \frac{1}{2^n}.$$

Use know the easy and useful fact that, for every sequence of positive numbers $(a_n)_{n\geq 1}$,

$$\lim_n \frac{1}{n} \log \sum_{1\leq i\leq n} a_n = \lim_n \frac{1}{n} \log \max_{1\leq i\leq n} a_n, \tag{7.1}$$

to get that

$$\lim_n \frac{1}{n} \log P(S_n \geq na) = \lim_n \frac{1}{n} \log \binom{n}{na} \frac{1}{2^n}$$

Use now Stirling's formula $n! \sim \sqrt{2\pi n}\, n^n e^{-n}$ to get that $\binom{n}{na} \sim \frac{1}{\sqrt{2\pi\, n}} a^{-na}(1-a)^{-n(1-a)}$. The result now follows. ☐

Knowing the large deviations is also very important when we want to evaluate asymptotically exponential integrals. Consider for example the case of a sequence of measures $\mu_n(dx)$. You can think of the case $P\left(\frac{S_n}{n} \in dx\right)$. Supose we want to evaluate the integrals

$$\frac{1}{n} \log \int e^{n\vartheta x} \mu_n(dx).$$

Think again of $\frac{1}{n} \log E[e^{\vartheta S_n}]$. If $\mu_n(dx) \cong e^{-nI(x)}$, then by the Laplace asymptotics, we have that the integral is asymptotically equivallent to

$$\frac{1}{n} \log \int e^{n\vartheta x - nI(x)} dx \cong \sup_x \left(\vartheta x - I(x)\right).$$

Large deviation theory was founded to large extent by Varadhan who also gave the following formal definition of a Large Deviation Principle (LDP):

Definition 7.1. *Let X be a Polish space (complete, separable, metric space). We say that the sequence of measures μ_n satisfies a LDP with rate function I, if for every Borel set Γ*

$$-\inf_{x\in\Gamma^o} I(x) \le \liminf_n \frac{1}{n}\log\mu_n(\Gamma) \le \limsup_n \frac{1}{n}\log\mu_n(\Gamma) \le -\inf_{x\in\bar\Gamma} I(x) \tag{7.2}$$

We want the rate function $I : X \to [0,\infty)$ to be lower semicontinuous. Often I is called good rate function if the level sets $\{x : I(x) \le L\}$ are compact.

REMARK: 1. For the moment think of $\mu_n(\Gamma)$ as $P\left(\frac{S_n}{n}\in\Gamma\right)$.

2. If $\inf_{x\in\Gamma^o} I(x) = \inf_{x\in\bar\Gamma} I(x)$, then $\lim_n \frac{1}{n}\log\mu_n(\Gamma) = -\inf_{x\in\Gamma} I(x)$

3. We need the distinction between $\Gamma^o, \Gamma, \bar\Gamma$ to deal with pathological cases e.g. when μ_n are non-atomic, i.e. $\mu_n(\{x\}) = 0$. In this case (7.2) cannot hold for $\Gamma = \{x\}$, without considering Γ^o. This formulation takes also into account the possibility of concentration of measure on the boundaries of Γ.

4. Notice that the reason of the presence of the inf in the formulation lies is relation (7.1). a way to see it formally is that $\lim\frac{1}{n}\log\mu_n(\Gamma)" = "\lim\frac{1}{n}\log\sum_{x\in\Gamma}\mu_n(x)$
$" = "\lim\frac{1}{n}\log\max_{x\in\Gamma}\mu_n(x)$

7.1. CRAMER'S THEOREM.

This is the starting example of large deviation theory.

Theorem 7.2. *Let $(X_i)_{i\ge1}$ a sequence of i.i.d variables. Assume that $\phi(t) := E[e^{tX_1}]$ is finite for every $t\in\mathbb{R}$. Then the measures $\mu_n(\cdot) := P\left(\frac{S_n}{n}\in\cdot\right)$ satisfy a LDP with rate function $I(x) = \sup_{t\in\mathbb{R}}\{xt - \log\phi(t)\}$.*

Before the proof let us remark on the rate function.

Definition 7.3. *For any real function ϕ, we define the Legendre transform $\phi^*(x) := \sup_{t\in\mathbb{R}}\{xt - \log\phi(t)\}$. This definiton also generalise to many dimensions (even infinite), by $\phi^*(x) := \sup_{t\in\mathbb{R}^d}\{<x,t> - \log\phi(t)\}$.*

Lemma 7.4. *Let $\phi(t) := E[e^{tX_1}]$ and $\phi^*(t)$ its Legendre transform. Then*

1. $t\to\log\phi(t)$ is convex.

2. ϕ^* is nonnegative, convex and lower semicontinuous.

3. If $\mu = E[X_1]$, then $\phi^*(\mu) = 0$.

4. For $x\ge\mu$, $x\to\phi^*(x)$ is nondecreasing, and for $x\le\mu$ it is nonincreasing.

5. For $x\ge\mu$, $\phi^*(x) = \sup_{t\ge0}\{xt - \log\phi(t)\}$, and for $x\le\mu$, $\phi^*(x) = \sup_{t\le0}\{xt - \log\phi(t)\}$.

Proof. 1. and 2. are trivial (for 1. just use Jensen's inequality).

3. By Jensen's inequality we have that $\log E[e^{tX_1}] \ge tE[X_1] = t\mu$, for every t. So $\phi^*(\mu) = \sup_t\{\mu t - \log\phi(t)\} \le 0$, which implies the result by the nonnegativity of ϕ^*.

4. It follows from the convexity of ϕ^* and the fact that $\phi^*(\mu) = 0$.

5. Notice that the slope of $\log\phi(t)$ at the origin is equal to $E[X_1]$. The result now follows by the convexity of $\log\phi(t)$.

□

We are now ready to prove Crámer's Theorem.

Proof. Without loss of generality we will assume that $E[X_1] = 0$.

THE UPPER BOUND. The upper bound is an optimization over a family of exponential Chebyshev's inequalities. Let us first bound the quantity, let $t\ge0$

$$\begin{aligned} P\left(\frac{S_n}{n}\ge x\right) &\le e^{-n\,tx}E\left[e^{tS_n}\right] \\ &= e^{-n\,tx}\left(E[e^{t\,X_1}]\right)^n = \exp\{-n(tx - \log\phi(t))\}. \end{aligned}$$

Since this bound is true for all $t \geq 0$ we have that

$$\begin{aligned} P\left(\frac{S_n}{n} \geq x\right) &\leq \exp\{-n \sup_{t\geq 0}(tx - \log\phi(t))\} \\ &= \exp\{-n\phi^*(x)\}. \end{aligned}$$

where the last equality is due to part 5. of the previous lemma. To conclude the upper bound notice that $\phi^*(x) = \inf_{y\geq x} I(y)$, by part 4. of the previous lemma.

To pass from the event $\{\frac{S_n}{n} \geq x\}$ to a general event $\{\frac{S_n}{n} \in \Gamma\}$ we do

$$\begin{aligned} P\left(\frac{S_n}{n} \in \Gamma\right) &\leq P\left(\frac{S_n}{n} \in \bar{\Gamma}\right) \\ &\leq P\left(\frac{S_n}{n} \in \bar{\Gamma} \cap [0,\infty)\right) + P\left(\frac{S_n}{n} \in \bar{\Gamma} \cap (-\infty,0]\right) \\ &\leq e^{-n\phi^*(x_+)} + e^{-n\phi^*(x_-)}, \end{aligned}$$

where $x_+ = \inf\{x \in \bar{\Gamma} \cap [0,\infty)\}$ and analogously for x_-. We then conclude, using (7.1) that $\limsup_n \frac{1}{n} \log P\left(\frac{S_n}{n} \in \Gamma\right) \leq -\phi^*(x_+) \wedge \phi^*(x_-)$, and by part 4. of the previous lemma it is equal to $-\inf_{x\in\bar{\Gamma}} \phi^*(x)$.

THE LOWER BOUND. It is clear that it is enough to prove that for every x and $\varepsilon > 0$, we have $\liminf_n \frac{1}{n} \log P\left(\frac{S_n}{n} \in (x-\varepsilon, x+\varepsilon)\right) \geq -\phi^*(x)$.

The idea (which is very important) to prove this is the following: Remember that $E[X_1] = 0$, so by the LLN $S_n/n \to 0$, $P\, a.s.$. This is the *typical* event, i.e. $P\left(\frac{S_n}{n} \in (-\varepsilon,\varepsilon)\right) \to 1$. We want though to compute the probability of the *atypical* event $P\left(\frac{S_n}{n} \in (x-\varepsilon, x+\varepsilon)\right)$. To do this we will introduce a new measure $\hat{P}$, such that $\hat{E}[X_1] = x$ and thus $\hat{P}\left(\frac{S_n}{n} \in (-\varepsilon,\varepsilon)\right) \to 1$. In other words the atypical event becomes under the new measure typical. The asymptotic probability that we are after will be captured by this change of measure.

Let's fix the ideas. Let $\alpha(dy)$ the distribution function related to P. Define the new measure $\hat{P}$ by

$$\hat{\alpha}(dy) = \frac{e^{\tau y}}{\phi(\tau)} \alpha(dy).$$

The value of τ will be chosen as the unique value (if it exists), for which $\hat{E}[X_1] = \int y\hat{\alpha}(dy) = x$. Notice that

$$\int y\hat{\alpha}(dy) = (\log\phi(\tau))'.$$

So we are looking for a value of τ for which $(\log\phi(\tau))' = x$. Such a τ exists if the $\sup_t (tx - \log\phi(t))$ is achieved. Suppose that this is the case (we will deal with the case that this is not the case separately), then $\phi^*(x) = \tau x - \log\phi(\tau)$. Now compute, let $\delta < \varepsilon$ and also suppose that $\tau \geq 0$ (the case $\tau \leq 0$ is handled similarly).

$$\begin{aligned} P\left(\frac{S_n}{n} \in (x-\delta, x+\delta)\right) &= \int_{\{\frac{S_n}{n} \in (x-\delta,x+\delta)\}} \alpha(dy_1)\cdots\alpha(dy_n) \\ &= \int_{\{\frac{S_n}{n} \in (x-\delta,x+\delta)\}} e^{-\tau(y_1+\cdots+y_n)}\, \phi(\tau)^n\, \hat{\alpha}(dy_1)\cdots\hat{\alpha}(dy_n) \\ &\geq e^{-n\tau(x+\delta)}\phi(\tau)^n \int_{\{\frac{S_n}{n} \in (x-\delta,x+\delta)\}} \hat{\alpha}(dy_1)\cdots\hat{\alpha}(dy_n) \\ &= e^{-n(x\tau+\delta\tau-\log\phi(\tau))}\hat{P}\left(\frac{S_n}{n} \in (x-\delta, x+\delta)\right). \end{aligned}$$

Finally we have (we will use the fact that $\hat{P}\left(\frac{S_n}{n} \in (x-\delta, x+\delta)\right) \to 1$, as $n \to \infty$)

$$\begin{aligned}
\liminf_n \frac{1}{n} \log P\left(\frac{S_n}{n} \in (x-\varepsilon, x+\varepsilon)\right) \geq \\
&\geq \liminf_n \frac{1}{n} \log P\left(\frac{S_n}{n} \in (x-\delta, x+\delta)\right) \\
&= -(x\tau + \delta\tau - \log\phi(\tau)) + \liminf_n \frac{1}{n} \log \hat{P}\left(\frac{S_n}{n} \in (x-\delta, x+\delta)\right) \\
&= -(x\tau + \delta\tau - \log\phi(\tau)) \\
&= -\phi^*(x) - \delta\tau.
\end{aligned}$$

Now, we just need to let $\delta \to 0$.

There only remains the case that the $\sup_t\{xt - \log\phi(t)\}$ is not achieved. In this case because of the convexity of $\log\phi(t)$, and the fact that $\phi(t) < \infty$, for every t, there must be a sequence $t_n \to +\infty$ (assume $x > 0$, similarly for $x < 0$), such that

$$\begin{aligned}
\phi^*(x) &= \lim_{t_n \to \infty} \left(xt_n - \log\phi(t_n)\right) \\
&= -\lim_{t_n \to \infty} \log \int e^{t_n(y-x)} \alpha(dy) \\
&= -\lim_{t_n \to \infty} \log \int_{y \geq x} e^{t_n(y-x)} \alpha(dy).
\end{aligned}$$

If now $F\left((x,\infty)\right) > 0$, then by the monotone convergence the last integral converges to $+\infty$, and this will imply that $\phi^*(x) = -\infty$, which is false by the positivity of ϕ^*. Thus we get that $\alpha(\{x\}) = e^{-\phi^*(x)}$. Then

$$\begin{aligned}
&\liminf_n \frac{1}{n} \log P\left(\frac{S_n}{n} \in (x-\varepsilon, x+\varepsilon)\right) \\
&\geq \liminf_n \frac{1}{n} \log P\left(X_1 = \cdots X_n = x\right) = \log P\left(X_1 = x\right) = -\phi^*(x)
\end{aligned}$$

This completes the proof of Crámer's theorem. $\square$

REMARK: Crámer's theorem holds also without the assumption that $\phi(t) < \infty$, for every t. In fact it holds even in the case that $D_\phi := \{t\colon \phi(t) < \infty\} = \{0\}$, although in this case the rate function ϕ^* might be trivial. To prove it in this case one proves it first for the measures $\nu_n^M(\cdot) := P\left(\frac{S_n}{n} \in \cdot \big| |X_i| < M\right)$, which reduces to the case that we considered and then passes to the limit $M \to \infty$.

8. LAW OF ITERATED LOGARITHM

The Law of Iterated Logarithm (LIL) provides the third general limit phenomenon for sums of i.i.d. random variables. The other two are (1) the Law of Large Numbers (LLN) which says that $\frac{S_n}{n} \xrightarrow{a.s.} \mu$, as $n \to \infty$ for $S_n = X_1 + \cdots + X_n$ with (X_i) i.i.d. with mean μ and (2) the Central Limit Theorem (CLT), which says that $\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$, where σ is the standard deviation of X_1. The Law of Iterated Logarithm interpolates between the above two limit theorems:

Theorem 8.1. *Let $S_n = X_1 + \cdots + X_n$ with (X_i) i.i.d. with mean 0 and variance 1. Then a.s. it holds that*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n \log\log n}} = \sqrt{2} \qquad \text{and} \qquad \liminf_{n\to\infty} \frac{S_n}{\sqrt{n \log\log n}} = -\sqrt{2}.$$

Exercise 24. *Extend the statement of the LLN to the case of i.i.d. random variables with general; mean μ and variance σ^2.*

The proof of the LLN uses two *maximal inequalities*:

Lemma 8.2 (Kolmogorov's maximal inequality). *Let* $(X_n)_{n\geq 1}$ *be independent random variables with mean* 0 *and variances* σ_n^2. *Let* $M_n := \sup_{i\leq n}|S_i|$. *Then for* $\ell > 0$ *we have that*

$$\mathrm{P}\big(T_n \geq \ell\big) \leq \frac{1}{\ell^2}\,\mathbb{V}\mathrm{ar}(S_n)$$

Proof. The idea is to look at the sequence (process – S_n will be called *random walk*) $(|S_n|)_{n\geq 1}$ and decompose according to the first time this hits or goes above level ℓ. In particular, denote (what in stochastic processes will be called *stopping times*

$$\tau_\ell := \inf\{n\colon |S_n| \geq \ell\}.$$

We then have that

$$\{T_n \geq \ell\} = \bigcup_{k=1}^{n}\{\tau_\ell = k\} = \bigcup_{k=1}^{n}\big\{|S_1| < \ell, ..., |S_{k-1}| < \ell, |S_k| \geq \ell\big\}$$

with the sets $\{\tau_\ell = k\}$ being disjoint. Then we have

$$\begin{aligned}
\mathrm{P}\big(T_n \geq \ell\big) &= \sum_{k=1}^{n}\mathrm{P}(\tau_\ell = k)\\
&\leq \sum_{k=1}^{n}\frac{1}{\ell^2}\mathrm{E}\big[S_k^2;\ \tau_\ell = k\big] && \text{[since on } \{\tau_\ell = k\} \text{ we have that } |S_k| \geq \ell\ \text{]}\\
&\leq \sum_{k=1}^{n}\frac{1}{\ell^2}\mathrm{E}\big[S_k^2 + (S_n - S_k)^2;\ \tau_\ell = k\big]\\
&= \sum_{k=1}^{n}\frac{1}{\ell^2}\mathrm{E}\big[S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2;\ \tau_\ell = k\big]\\
&= \sum_{k=1}^{n}\frac{1}{\ell^2}\mathrm{E}\big[S_n^2;\ \tau_\ell = k\big]\\
&= \frac{1}{\ell^2}\mathrm{E}\big[S_n^2; T_n \geq \ell\big] \leq \frac{1}{\ell^2}\mathrm{E}\big[S_n^2\big]
\end{aligned}$$

In the fourth line we used crucially the independence: $S_k\mathbb{1}_{\tau_\ell = k}$ is measurable with respect to the variables $X_1, ..., X_k$, while $S_n - S_k = X_{k+1} + \cdots + X_n$ and so they are independent, which leads to $\mathrm{E}[S_k(S_n - S_k)\,;\,\tau_\ell = k] = \mathrm{E}[S_k\,;\,\tau_\ell = k]\mathrm{E}[(S_n - S_k)] = 0$. In the last line we used again the disjointness of the sets $\{\tau_\ell = k\}$. □

We will see that Kolmogorov's inequality is a special case of Doob's maximal inequality in martingales. The second inequality that we need is Lévy's inequality, which shows how to control M_n from the probabilities of the tails.

Lemma 8.3. *Let* $(X_n)_{n\geq 1}$ *be i.i.d. such that for* $0 < \delta < 1$ *it holds that*

$$\sup_{i\leq n}\mathrm{P}\big(|X_i + \cdots + X_n| \geq \tfrac{\ell}{2}\big) < \delta.$$

Then

$$\mathrm{P}\big(M_n \geq \ell\big) \leq \frac{1}{1-\delta}\mathrm{P}\big(|S_n|| \geq \tfrac{\ell}{2}\big).$$

Proof. Again, we will decompose according to the hitting time of level ℓ. We start witih

$$\begin{aligned} \mathrm{P}\big(M_n \geq \ell\,,\, |S_n| < \tfrac{\ell}{2}\big) &= \sum_{k=1}^{n} \mathrm{P}\big(\{\tau_\ell = k\} \cap \{|S_n| < \tfrac{\ell}{2}\}\big) \\ &\leq \sum_{k=1}^{n} \mathrm{P}\big(\{\tau_\ell = k\} \cap \{|S_n - S_k| \geq \tfrac{\ell}{2}\}\big) \\ &= \sum_{k=1}^{n} \mathrm{P}\big(\{\tau_\ell = k\} \cap \{|S_{k+1} + \cdots + S_n| \geq \tfrac{\ell}{2}\}\big) \\ &= \sum_{k=1}^{n} \mathrm{P}\big(\tau_\ell = k\big)\mathrm{P}\big(|S_{k+1} + \cdots + S_n| \geq \tfrac{\ell}{2}\big) \\ &\leq \delta \sum_{k=1}^{n} \mathrm{P}\big(\tau_\ell = k\big) \\ &= \delta \mathrm{P}\big(M_n \geq \ell\big) \end{aligned}$$

where in the second line we used the fact that if $|S_k| \geq \ell$ and $|S_n| < \ell/2$, then we must have $|S_n - S_k| \geq \ell/2$. In the fourth line we used the independence of the events $\{\tau_\ell = k\}$ and $\{|S_{k+1} + \cdots + S_n| \geq \frac{\ell}{2}\}$.

We can now conclude by

$$\begin{aligned} \mathrm{P}\big(M_n \geq \ell\big) &= \mathrm{P}\big(M_n \geq \ell\,,\, |S_n|| < \tfrac{\ell}{2}\big) + \mathrm{P}\big(M_n \geq \ell\,,\, |S_n|| \geq \tfrac{\ell}{2}\big) \\ &\leq \mathrm{P}\big(M_n \geq \ell\,,\, |S_n|| < \tfrac{\ell}{2}\big) + \mathrm{P}\big(|S_n|| \geq \tfrac{\ell}{2}\big) \\ &\leq \delta \mathrm{P}\big(M_n \geq \ell\big) + \mathrm{P}\big(|S_n|| \geq \tfrac{\ell}{2}\big), \end{aligned}$$

from which the inequality follows. ☐

Proof of the LIL. For the proof of the LIL we refer to [V], Chapter 3. ☐

Exercise 25. *Show that the interval* $[-\sqrt{2}, \sqrt{2}]$ *is a.s. identical to all limit points of the sequence* $\frac{S_n}{\sqrt{n \log\log n}}$.

9. MARTINGALES

9.1. CONDITIONAL PROBABILITIES AND EXPECTATIONS.

We will make a quick and rather informal review on conditional expectations. More details can be found in several books and sources but a nice one is in [B].

We start with the elementary definition of conditional probability:

if A, B are sets / events, then the conditional probability of A given B is

$$\mathrm{P}(A \,|\, B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}.$$

Let us now move to conditioning with respect to random variables. If X is a random variable taking discrete values, $\{x_1, ..., x_k\}$, we have that

$$\mathrm{P}(A \,|\, X = x_i) = \frac{\mathrm{P}(A,\, X = x_i)}{\mathrm{P}(X = x_i)}.$$

Subtleties start when X are continuous variables and so the event $\{X = x\}$ might be of measure 0. Informally, we would like to write

$$\mathrm{P}(A \,|\, X = x) = \frac{\mathrm{P}(A,\, X = x)}{\mathrm{P}(X = x)}.$$

but then we have the issues of $0/0$. One way could be to define the above via limits

$$\mathrm{P}(A \,|\, X = x) = \lim_{h \to 0} \frac{\mathrm{P}(A,\, X \in (x - h, x + h))}{\mathrm{P}(X \in (x - h, x + h)}.$$

This is still a cumbersome approach. However, even though the above are not formally correct, they usually gives the correct guidance on how to actually write down the conditional probabilities. An alternative approach is to go through Radon-Nikodym derivatives: For a given set A, define the measure $\hat{Q}(B) := \mathrm{P}(A,\, X \in B)$ which is absolutely continuous with respect to the measure $\hat{\mathrm{P}}(B) := \mathrm{P}(X \in B)$. Then the Radon-Nikodym derivative $\frac{\mathrm{d}\hat{Q}}{\mathrm{d}\hat{\mathrm{P}}}(x)$ exists and

$$\hat{Q}(A) := \mathrm{P}(A\,,\, X \in B) = \int_B \frac{\mathrm{d}\hat{Q}}{\mathrm{d}\hat{\mathrm{P}}}(x)\,\hat{\mathrm{P}}(\mathrm{d}x) = \int_B \frac{\mathrm{d}\hat{Q}}{\mathrm{d}\hat{\mathrm{P}}}(x)\,\mathrm{P}(X \in \mathrm{d}x),$$

and so we can define $\mathrm{P}(A \,|\, X = x) := \frac{\mathrm{d}\hat{Q}}{\mathrm{d}\hat{\mathrm{P}}}(x)$. We can also have

Definition 9.1. *Let X be a random variable and A an event in the probability space $(\Omega, \mathcal{F}, \mathrm{P})$, Then the conditional probability* $\mathrm{P}(A \,|\, X)$ *is defined are any real random variable of* Ω*, measurable with respect to* $\sigma(X)$ *such that for any B Borel on $\mathbb{R}$ we have*

$$\mathrm{P}(A\,,\, X \in B) = \int_{\{X \in B\}} \mathrm{P}(A \,|\, X)\,\mathrm{dP}$$

A couple of problems with these approach are : (1) that this definition depends on the set A and (2) that it is not practical. For practical ways to compute conditional probability it is better to resort to the informal ways we exposed above – for example see the exercise below. For a more robust way we will introduce later on the concept of *regular conditional probability.* Before doing so let us also introduce (informally) the concept of **condition expectation**.

Having two random variables X, Y we would like to define the conditional expectation

$$\mathrm{E}[Y \,|\, X].$$

The natural way would be to imitate the standard definition of expectation and write (informally)

$$\mathrm{E}[Y \,|\, X = x] = \int_{\mathbb{R}} y\,\mathrm{P}(Y = y \,|\, X = x)\,\mathrm{d}y = \frac{1}{\mathrm{P}(X = x)} \int_{\mathbb{R}} y\,\mathrm{P}(Y = y\,,\, X = x)\,\mathrm{d}y$$

which writing it in terms of densities

$$\mathrm{E}[Y \,|\, X = x] = \frac{1}{f_Y(y)} \int_{\mathbb{R}} y f_{X,Y}(x, y)\,\mathrm{d}y$$

where $f_{X,Y}(x, y)$ denotes the joint density of (X, Y) and $f_Y(y)$ the marginal density of Y.

Although this a natural, intuitive "definition" (and in reality this a a good guiding way) it is informal as we deal with 0/0. A more formal approach is as we did above. Namely, define the measure $\hat{Q}(B) : \mathrm{E}[Y,\, X \in B]$ and $\hat{P}(B) := \mathrm{P}(X \in B)$ from which we have that $\hat{Q} << \hat{P}$, where "$<<$" means "absolute continuity". Then, defining the Radon-Nikodym derivative $\frac{\mathrm{d}\hat{Q}}{\mathrm{d}\hat{P}}$ we have for any set B that

$$\mathrm{E}[Y\,,\, X \in B] = \int_B \frac{\mathrm{d}\hat{Q}}{\mathrm{d}\hat{P}}(x)\,\hat{P}(\mathrm{d}x) =: \int_B \mathrm{E}[Y \,|\, X = x]\,\hat{P}(\mathrm{d}x).$$

To motivate what's coming, let us write the left hand side as

$$\mathrm{E}[Y \mathbb{1}_{\{X \in B\}}] = \int_{X \in B} Y\,\mathrm{d}P,$$

and then the above equality takes the form

$$\int_{X \in B} Y\,\mathrm{d}P \int_{X \in B} \mathrm{E}[Y \,|\, X]\,\mathrm{d}P$$

and so we could define $\mathrm{E}[Y \,|\, X]$ as the $\sigma(X)$ measurable random variable such that for any $B \in \mathcal{B}(\mathbb{R})$ it holds that

$$\int_{X \in B} Y\,\mathrm{d}P = \int_{X \in B} \mathrm{E}[Y \,|\, X]\,\mathrm{d}P.$$

More generally, we would like to upgrade the notion of conditional probability / expectation so that we can condition over an arbitrary σ-field $\mathcal{D}$ (in the discussion so far, when we were conditioning with respect to X, we were really conditioning with respect to the information generated by the random variable X, which is the σ-algebra $\sigma(X)$.

Motivated by the above discussion, given a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and a sub-field $\mathcal{D} \subset \mathcal{F}$, we can define the *conditional expectation of* Y *given* $\mathcal{D}$, denoted by $\mathrm{E}[Y \,|\, \mathcal{D}]$ as the $\sigma(\mathcal{D})$-measurable function such that for any $D \in \mathcal{D}$ we have

$$\int_D \mathrm{E}[Y \,|\, \mathcal{D}] \, \mathrm{dP} = \int_D Y \, \mathrm{dP}.$$

Eventually, we would like to treat conditional expectations are real expectation, eg be able to write

$$\mathrm{E}[Y \,|\, \mathcal{D}] = \int_{\mathbb{R}} y \, \mathrm{P}(Y \in \mathrm{d}y \,|\, \mathcal{D}),$$

and the subtlety is that we need to consider $\mathrm{P}(\cdot \,|\, \mathcal{D})$ as a real probability, i.e. for any disjoint sets $(A_n)_{n\geq}$ to be able to write

$$\mathrm{P}(\cup_{n\geq 1} A_n \,|\, \mathcal{D}) = \sum_{n\geq 1} \mathrm{P}(A_n \,|\, \mathcal{D}).$$

The subtlety is that for any A, $\mathrm{P}(A \,|\, \mathcal{D})$ is defined a.s. i.e. on a set $\Omega' \subset \Omega$ of measure 1. But if we do so for every (uncountable many) different set(s) A, then sets of measure 0 may pile up to something of non-zero measure. The fact that we can circumvent this subtlety is a deep result, which leads to the definition of *regular conditional probability*:

Definition 9.2 (Regular Conditional Probability). *Given a probability space* $(\Omega, \mathcal{F}, \mathrm{P})$ *and a sub-field* $\mathcal{D} \subset \mathcal{F}$ *is a measure* $\mathrm{P}^*(\cdot \,|\, \mathcal{D})$ *such that*

1. *for any* $A \in \mathcal{F}$ *fixed,* $\omega \to \mathrm{P}^*(A \,|\, \mathcal{D})$ *is a* $\mathcal{D}$ *measurable random variable.*
2. *for any* $\omega \in \Omega$ *fixed,* $A \to \mathrm{P}^*(A \,|\, \mathcal{D})$ *is a probability on* $\mathcal{F}$

The fact that a regular conditional probability exists is a deep theorem which we will not prove. However, we assume it and manipulate conditional expectations as if they were expectation ! The following proposition summarises the main properties:

Proposition 9.3. *Given a probability space* $(\Omega, \mathcal{F}, \mathrm{P})$*, a sub-field* $\mathcal{D} \subset \mathcal{F}$ *and random variables* $Y, Y_1, Y_2, ...$ *such that* $\mathrm{E}[|Y|], \mathrm{E}[|Y_1|], \mathrm{E}[|Y_2|], ... < \infty$ *we have*

1. *For* $a, b \in \mathbb{R}$ *we have that*
$$\mathrm{E}[aY_1 + bY_2 \,|\, \mathcal{D}] = a\mathrm{E}[Y_1 \,|\, \mathcal{D}] + bE[Y_2 \,|\, \mathcal{D}], \qquad a.s.,$$
2. *if* $Y \geq 0$*, then* $\mathrm{E}[Y \,|\, \mathcal{D}] \geq 0$*,*
3. *for two subfields* $\mathcal{D} \subset \mathcal{E} \subset \mathcal{F}$ *we have*
$$\mathrm{E}\Big[\mathrm{E}\big[Y \mid \mathcal{E}\big] \,|\, \mathcal{D}\Big] = \mathrm{E}\big[Y \mid \mathcal{D}\big], \qquad a.s..$$
4. *if* Y *is independent of* $\mathcal{D}$*, then* $\mathrm{E}\big[Y \mid \mathcal{D}\big] = \mathrm{E}[Y]$*, a.s.,*
5. *if* Y *is measurable with respect to* $\mathcal{D}$*, then* $\mathrm{E}\big[Y \mid \mathcal{D}\big] = Y$*, a.s.,*
6. *if* $Y_n \uparrow Y$*, a.s., then* $\mathrm{E}\big[Y_n \mid \mathcal{D}\big] \uparrow \mathrm{E}[Y \,|\, \mathcal{D}]$*, a.s.*

9.2. MARTINGALES AND MARTINGALE INEQUALITIES.

We can now define the important notion of Martingales. These are in a sense "stochastic constants" and they provide a powerful tool in stochastic processes, with powerful inequalities, a.s. convergence theorems and links to Markov processes, analysis and PDEs. We will give a glimpse of these. We begin with the definition:

Definition 9.4 (Martingales). *Let* $(\Omega, \mathcal{F}, \mathrm{P})$ *be a probability space with an increasing sequence of* σ*-algebras* $\mathcal{F}_1 \subset \mathcal{F}_2 \cdots \subset \mathcal{F}_n \subset \cdots$ *A sequence of random variables* X_n*, which are* $\mathcal{F}_n$ *measurable is called a martingale if*

$$\mathrm{E}[X_{n+1} \,|\, \mathcal{F}_n] = X_n, \qquad \textit{a.s. for every } n.$$

$(X_n)_{n\geq 1}$ is called a sub-martingale if a.s. $X_n \leq \mathrm{E}[X_{n+1} \,|\, \mathcal{F}_n]$ for every n and it is called a super-martingale if if a.s. $X_n \geq \mathrm{E}[X_{n+1} \,|\, \mathcal{F}_n]$ for every n.

A useful proposition is the following:

Proposition 9.5. *If (X_n) is a martingale and ϕ a convex function, then $(\phi(X_n))$ is a sub-martingale.*

Proof. This uses the power of viewing conditional expectation as simple expectations: we can just use Jensen's inequality:

$$\mathrm{E}\big[\phi(X_{n+1}) \,|\, \mathcal{F}_n\,\big] \geq \phi\Big(\mathrm{E}\big[X_{n+1} \,|\, \mathcal{F}_n\,\big]\Big) = \phi(X_n),$$

where the equality is just the martingale property. ☐

The next proposition justifies the idea of martingales being stochastic constants

Proposition 9.6. *For a martingale (X_n) we have that*

1. *for $n \geq m$ we have that $\mathrm{E}\big[\mathrm{E}[X_n \,|\, \mathcal{F}_m]\big] = X_m$, a.s.*
2. *for every n we have that $\mathrm{E}[X_n] = \mathrm{E}[X_0]$*

Proof. (1) follows from the martingale property and property (3) of Proposition 9.3. (2) follows from $\mathrm{E}[X_0] = \mathrm{E}\big[\mathrm{E}[X_n \,|\, \mathcal{F}_0]\big] = \mathrm{E}[X_n]$, where first step we used propery (1) and took expectations on both sides and in the second step we used property (3) of Proposition 9.3. ☐

The standard example of a martingale is the Random Walk: if $(\xi_n)_{n\geq 1}$ are independent random variables with mean zero, then $X_n = \xi_1 + \cdots + \xi_n$ is a Martingale. Check it ! If additionally $(\xi_n)_{n\geq 1}$ are i.i.d., mean zero and variance one, then $X_n^2 - n$ is a martingale. Check !

We are now ready to prove the first martingale inequality:

Theorem 9.7 (Doob's maximal inequality). *If $(X_n)_{n\geq 1}$ is a sub-martingale with respect to a filtration $(\mathcal{F}_n)_{n\geq 1}$, we have that*

$$\begin{aligned} \mathrm{P}\Big(\sup_{i\leq n} |X_j| \geq \ell\Big) &\leq \frac{1}{\ell}\mathrm{E}\Big[|X_n|\,;\,\big\{\sup_{i\leq n}|X_j| \geq \ell\big\}\Big] \\ &\leq \frac{1}{\ell}\mathrm{E}\big[|X_n|\big] \end{aligned} \tag{9.1}$$

Proof. The proof is similar to the of Kolmogorov's inequality, only that the Martingale property will provide a shortcut and a sharper inequality.

Define

$$\tau_\ell := \inf\{j \leq n \colon |X_j| \geq \ell\}.$$

We then have

$$\begin{aligned} \mathrm{P}\Big(\sup_{i\leq n}|X_j| \geq \ell\Big) &= \mathrm{P}(\tau_\ell \leq n) = \sum_{j=1}^n \mathrm{P}(\tau_\ell = j) \\ &\leq \frac{1}{\ell}\sum_{j=1}^n \mathrm{E}(|X_j|\,;\,\tau_\ell = j) \\ &\leq \frac{1}{\ell}\sum_{j=1}^n \mathrm{E}(|X_n|\,;\,\tau_\ell = j) \qquad \text{[by the martingale property - we will explain in more detail]} \\ &\leq \frac{1}{\ell}\mathrm{E}[|X_n|\,;\,\tau_\ell \leq n] \qquad \text{[by disjointness]} \\ &\leq \frac{1}{\ell}\mathrm{E}[|X_n|] \end{aligned}$$

The third line is explained as follows: Since (X_n) is a martingale, $|X_n|$ is a sub-martingale. We also have

$$\begin{aligned}
\mathrm{E}\Big[|X_n|\,;\,\tau_\ell=j\Big] &= \mathrm{E}\Big[\mathrm{E}\big[|X_n|\,;\,\tau_\ell=j\,|\mathcal{F}_j\,\big]\Big] && \text{[by the tower property]}\\
&= \mathrm{E}\Big[\mathrm{E}\big[|X_n|\,|\mathcal{F}_j\,\big]\mathbb{1}_{\tau_\ell=j}\Big] && \text{[because } \mathbb{1}_{\tau_\ell=j} \text{ is measurable with respect to } \mathcal{F}_j]\\
&\geq \mathrm{E}\Big[[|X_j|\mathbb{1}_{\tau_\ell=j}\Big] && \text{[by the sub-martingale property of } X_n]\\
&= \mathrm{E}\Big[|X_j|\,;\,\tau_\ell=j\Big].
\end{aligned}$$

□

Theorem 9.8 (Doob's L^p inequality). *Let $(X_n)_{n\geq1}$ be a martingale with respect to a filtration $(\mathcal{F}_n)_{n\geq1}$ and $S:=\sup_{j\leq n}|X_j|$. Then for $p>1$ we have that*

$$\mathrm{E}[S^p]\leq\Big(\frac{p}{p-1}\Big)^p\mathrm{E}[|X_n|^p].$$

Proof. We have

$$\begin{aligned}
\mathrm{E}[S^p] &= \int_0^\infty p\,\ell^{p-1}\mathrm{P}(S\geq\ell)\,\mathrm{d}\ell\\
&\leq p\int_0^\infty \ell^{p-1}\,\frac{1}{\ell}\,\mathrm{E}\big[|X_n|^p\mathbb{1}_{\{S\geq\ell\}}\big]\,\mathrm{d}\ell && \text{[by (9.1)]}\\
&= p\mathrm{E}\Big[|X_n|\int_0^\infty \ell^{p-2}\mathbb{1}_{\{S\geq\ell\}}\,\mathrm{d}\ell\Big]\\
&= p\mathrm{E}\Big[|X_n|\int_0^S \ell^{p-2}\mathrm{d}\ell\Big]\\
&= \frac{p}{p-1}\mathrm{E}\Big[|X_n|\,S^{p-1}\Big]\\
&\leq \frac{p}{p-1}\mathrm{E}\Big[|X_n|^p\Big]^{1/p}\mathrm{E}\Big[\,S^p\Big]^{1/q} && \text{[by Hölder with } \frac{1}{p}+\frac{1}{q}=1],
\end{aligned}$$

and rearranging the terms we obtain the result. At this step, a thorough reader might object that the division is rearrangement is not well defined as we don't know if $\mathrm{E}\Big[\,S^p\Big]$ is finite or not. So, to be secure, we repeat the above argument for $S_M:=S\wedge M$, for some $M>0$ that eventually will be taken to infinity. So let's repeat the above argument:

$$\begin{aligned}
\mathrm{E}[S_M^p] &= \int_0^\infty p\,\ell^{p-1}\mathrm{P}(S_M\geq\ell)\,\mathrm{d}\ell\\
&\leq p\int_0^\infty \ell^{p-1}\,\frac{1}{\ell}\,\mathrm{E}\big[|X_n|^p\mathbb{1}_{\{S_M\geq\ell\}}\big]\,\mathrm{d}\ell && \text{[by (9.1)]}
\end{aligned}$$

Here we need to notice that $\{S\wedge M\geq\ell\}$ is either equal to $\emptyset$ if $M<\ell$ or $\{S\geq\ell\}$ if $M>\ell$, therefore (9.1) is also valid for S_M. The rest of the steps are then the same as above and at the last step we can now divide without any problem by $\mathrm{E}\Big[\,S_M^p\Big]$ before taking the limit $M\to\infty$ via the use of monotone convergence theorem.

□

Exercise 26. *Show that the above inequality cannot extend to L^1 by considering the martingale (check) $(X_n)_{n\geq1}$ on $\omega\in[0,1]$ which is defined as $X_n(\omega)=2^n\mathbb{1}_{[0,2^{-n}]}(\omega)$.*

Remark 9.9. *We saw that the maximal inequality is not valid in* L^1*. However, there is a so-called* $L\log L$ *inequality (see* [RY], *Exercise (1.16)):*

$$\mathrm{E}[S] \leq C\Big(1 + \mathrm{E}\big[|X_n| \log^+ |X_n|\big]\Big),$$

where, again, $S := \sup_{j\leq n} |X_j|$*. Inequalities of this sort have found interesting applications in various stochastic contexts such branching processes, multiplicative functionals, turbulence and directed polymer models.*

9.3. MARTINGALE CONVERGENCE THEOREM.

There is a number of martingale convergence theorems. We will aim to just prove that martingales under certain conditions converge a.s. Towards this we will need some preparation, which will also require the introduction of a number of important concepts such as **stopping times** and their associated σ-algebras and **optional stopping theorems.** Let us start with stopping times.

Definition 9.10. *Let* $(\mathcal{F}_n)_{n\geq 1}$ *be a filtration. A random variable* τ *with values in* $\mathbb{N}\cup\{\infty\}$ *is called a stopping time if for every* $n\in\mathbb{N}$ *we have that* $\{\omega\colon \tau(\omega)\leq n\}\in\mathcal{F}_n$*.*

The requirement of the stopping time being integer valued is just because we are discrete; stopping times can take real values. Some examples of stopping times are

- $\tau(\omega) = k$ a.s.
- if $(X_n)_{n\geq 1}$ is a stochastic process and $\mathcal{F}_n := \sigma(X_i\colon i\leq n)$, then $\tau_\ell := \inf_{n\colon X_n\geq \ell}$ is a stopping time.
- if τ is a stopping time and f nondecreasing, integer valued function, such that $f(n)\geq n$, then $\tau' := f(\tau)$ is also a stopping time.
- if τ_1, τ_2 are stopping times, then $\max(\tau_1,\tau_2)$ and $\min(\tau_1,\tau_2)$ are stopping times.

What is not a stopping time ? Well, times that depend on the future such as $\tau_\ell = \sup_n\{X_n\colon X_n\geq \ell\}$.

A important notion is the **stopped** $\boldsymbol{\sigma}$**-algebra**, that is the information that is generated by observing a stochastic process until a stopping (random) time. We have

Definition 9.11. *Let* $(\Omega,\mathcal{F},\mathrm{P})$ *be a probability space,* $(\mathcal{F}_n)_{n\geq 1}$ *be a filtration and* τ *be a stopping time with respect to that filtration. Then*

$$\mathcal{F}_\tau := \Big\{A\in\mathcal{F}\colon A\cap\{\tau\leq n\}\in\mathcal{F}_n \textit{ for every } n\Big\}$$

Exercise 27. *If* τ_1, τ_2 *are stopping times such that a.s.* $\tau_1\leq\tau_2$*, then* $\mathcal{F}_{\tau_1}\subset\mathcal{F}_{\tau_2}$

Exercise 28. *If* τ *is a stopping time, then* τ *is* $\mathcal{F}_\tau$ *measurable.*

We are now ready to state Doob's Optional Stopping Theorem, which says that the martingale property can be extended to stopping σ-algebras:

Theorem 9.12 (Doob's Optional Stopping Theorem). *Let* $(\Omega,\mathcal{F},\mathrm{P})$ *be a probability space,* $(\mathcal{F}_n)_{n\geq 1}$ *be a filtration,* $(X_n)_{n\geq}$ *be a martingale with respect to this filtration and* τ_1,τ_2 *be a stopping times, which are a.s. bounded and such that* $\tau_1\leq\tau_2$ *a.s. Then*

$$\mathrm{E}\big[X_{\tau_2}\,|\,\mathcal{F}_{\tau_1}\big] = X_{\tau_1} \qquad a.s.$$

Proof. Assume $\tau_1\leq\tau_2\leq M$, hence $\mathcal{F}_{\tau_1}\subset\mathcal{F}_{\tau_2}\subset\mathcal{F}_M$. It suffices to show that For any n and any stopping time $\tau\leq n$, we have that

$$\mathrm{E}\big[X_n\,|\,\mathcal{F}_\tau\big] = X_\tau \qquad a.s. \tag{9.2}$$

Indeed, if so then we have (all the below equalities are to be interpreted as a.s.)

$$\mathrm{E}\big[X_M\,|\,\mathcal{F}_{\tau_2}\big] = X_{\tau_2}$$

and we can take conditional expectation with respect to $\mathcal{F}_{\tau_1}$ and use the tower property:

$$\mathrm{E}\Big[\mathrm{E}\big[X_M \,|\, \mathcal{F}_{\tau_2}\big] \,|\, \mathcal{F}_{\tau_1}\Big] = \mathrm{E}\big[X_{\tau_2} \,|\, \mathcal{F}_{\tau_1}\big] \implies \mathrm{E}\Big[X_M \,|\, \mathcal{F}_{\tau_1}\Big] = \mathrm{E}\big[X_{\tau_2} \,|\, \mathcal{F}_{\tau_1}\big] \implies X_{\tau_1} = \mathrm{E}\big[X_{\tau_2} \,|\, \mathcal{F}_{\tau_1}\big].$$

So let us show (9.2). To this end, we will use the formal definition of conditional expectation. Thus, for any $A \in \mathcal{F}_\tau$ we need to check

$$\mathrm{E}\big[X_n; A\big] = \mathrm{E}\big[X_\tau; A\big].$$

We decompose he right-hand side as

$$\begin{aligned}\mathrm{E}\big[X_\tau; A\big] &= \sum_{k=1}^n \mathrm{E}\big[X_\tau; A \cap \{\tau = k\}\big]\\ &= \sum_{k=1}^n \mathrm{E}\big[X_k; A \cap \{\tau = k\}\big],\end{aligned}$$

since $A \cap \{\tau = k\} \in \mathcal{F}_k$ by the stopping time property, we can use the martingale property to write the above as

$$\begin{aligned}\sum_{k=1}^n \mathrm{E}\big[X_k; A \cap \{\tau = k\}\big] &= \sum_{k=1}^n \mathrm{E}\big[X_n; A \cap \{\tau = k\}\big]\\ &= \mathrm{E}\big[X_n; A \cap \{\tau \le n\}\big]\\ &= \mathrm{E}\big[X_n; A\big],\end{aligned}$$

with the last equality since a.s. $\tau \le n$. ☐

Boundedness is important: Let $X_n = \xi_1 + \cdots + \xi_n$, with (ξ_n) i.i.d., Bernoulli ±1 with probabilities 1/2 and $\tau := \inf\{n \colon X_n = 1\}$. By convention $X_0 = 0$. Then τ is a.s. finite but not bounded, $X_\tau = 1$ a.s. and so $\mathrm{E}[X_\tau] = 1$ while $X_0 = 0$. We can, however, extend the Optional Stopping Theorem if we have some integrability. In particular, if $S := \sup_{k \le \tau_2} |X_k|$ is integrable. Then we can establish (9.2) by defining the bounded stopping times $\tau_k := \tau \wedge k \le k$, for arbitrary k for which we have that

$$\mathrm{E}\big[X_n \,|\, \mathcal{F}_{\tau_k}\big] = X_{\tau_k} \qquad a.s.$$

But since a.s. $\tau_k \uparrow \tau$, we can use the monotone convergence theorem to pass to the limit and get that

$$\mathrm{E}\big[X_n \,|\, \mathcal{F}_\tau\big] = X_\tau \qquad a.s.$$

Actually, in order to rigorously pass to this limit one needs to go through the formal definition of conditional expectations and would also need that $\sigma\big(\cup_n \mathcal{F}_{\tau_n}\big) = \mathcal{F}_\tau$. Show the last statement and complete the rest of the steps.

Exercise 29. *Extend the optional stopping theorem to super- and sub-martingales.*

We are now ready to prove **Doob's up-crossing inequality**. First, let us define the up-crossings: Let $a < b$ and $(X_n)_{n \ge 0}$ be a stochastic process. Define the stopping times

$$\begin{aligned}\tau_1 &:= \inf\{j \ge 0 \colon X_j \le a\} \wedge n,\\ \tau_2 &:= \inf\{j \ge \tau_1 \colon X_j \ge b\} \wedge n,\\ &\cdots\\ \tau_{2k-1} &:= \inf\{j \ge \tau_{2k-2} \colon X_j \le a\} \wedge n,\\ \tau_{2k} &:= \inf\{j \ge \tau_{2k-1} \colon X_j \ge b\} \wedge n,\\ &\cdots\end{aligned}$$

We then define the number of up-crossings of $[a, b]$ as the number of times that (X_n) has gone from below a to above b:

$$U_n(a, b) := \max\{k \colon \tau_{2k} < n\}.$$

We can now prove:

Theorem 9.13 (Doob's up-crossing inequality). *Let $(X_n)_{n\geq 0}$ be a martingale with respect to a filtration $(\mathcal{F}_n)$ and $U_n(a,b)$ the number of up-crossings of the interval $[a,b]$. Then*

$$\mathrm{E}\big[U_n(a,b)\big] \leq \frac{1}{b-a}\mathrm{E}\big[(a-X_n)_+\big] \leq \frac{1}{b-a}\Big(a+\mathrm{E}[|X_n|]\Big).$$

Proof. Since we are in discrete time, notice that $\tau_{k+1}\geq \tau_k+1$ and so $\tau_n = n$. Define, next,

$$D_n := \sum_{k=1}^{2\lceil \frac{n}{2}\rceil} \big(X_{\tau_{2k}} - X_{\tau_{2k-1}}\big).$$

Note that some of the terms above might be 0. We have that

$$D_n \geq (b-a)U_n(a,b) + R_n, \tag{9.3}$$

where the remainder R_n has to do with whether there is an "incomplete up-crossing" at the end or not. By this we mean the following : Let $\bar{k} := \max\{k\colon \tau_k < n\}$. If $\bar{k}$ is even,

$$D_n = \sum_{k=1}^{\bar{k}/2} \big(X_{\tau_{2k}} - X_{\tau_{2k-1}}\big) \geq (b-a)U_n(a,b),$$

since for all $k\leq \frac{\bar{k}}{2}$ we have that $X_{\tau_{2k}} - X_{\tau_{2k-1}} \geq b-a$. On the other hand, if $\bar{k}$ is odd, then this means that $\tau_{\bar{k}+1} = n$ and

$$\begin{aligned} D_n &= \sum_{k=1}^{\lfloor \bar{k}/2\rfloor} \big(X_{\tau_{2k}} - X_{\tau_{2k-1}}\big) + \big(X_{\tau_{\bar{k}+1}} - X_{\tau_{\bar{k}}}\big) \\ &\quad + \sum_{k=1}^{\lfloor \bar{k}/2\rfloor} \big(X_{\tau_{2k}} - X_{\tau_{2k-1}}\big) + \big(X_n - X_{\tau_{\bar{k}}}\big) \\ &\geq (b-a)U_n(a,b) + \big(X_n - a\big). \end{aligned}$$

The stopping times (τ_k) are bounded and so we can use the optional stopping theorem to get that $\mathrm{E}[\tau_{2k}] = E[\tau_{2k-1}]$ and so $\mathrm{E}[D_n]=0$ (this may sound a bit weird as you expect $X_{\tau_{2k}}$ to be $\geq b$ and $X_{\tau_{2k-1}}$ to be $\leq a$! What's the explanation ?) So taking expectation in (9.3) we have that

$$0 \geq (b-a)\,\mathrm{E}\big[U_n(a,b)\big] + \mathrm{E}\big[\,R_n\,\big] \implies \mathrm{E}\big[U_n(a,b)\big] \leq -\frac{1}{b-a}\,\mathrm{E}\big[\,R_n\,\big] \leq \frac{1}{b-a}\mathrm{E}\big[(a-X_n)_+\big] \tag{9.4}$$

where in the last inequality we use the above considerations on the remained R_n. ☐

Remark 9.14. *The up-crossing inequality can be generalised to the case where (X_n) is a super-martingale. Indeed, in this case we would have that $\mathrm{E}\big[D_n\,\big]\leq 0$ and so inequality (9.4) remains valid. If (X_n) is a sub-martingale, then $(-X_n)$ is a super-martingale and since $|-X_n| = |X_n|$ the up-crossing inequality is also valid for sub-martingales.*

The up-crossing inequality can be used to prove a.s. convergence theorems for (sub- / sup-) martingales.

Corollary 9.15. *Let $(X_n)_{n\geq 1}$ be an L^1-bounded martingale, i.e. $\sup_{n\geq 1}\mathrm{E}[|X_n|]<\infty$. Then X_n convergences a.s.*

Proof. By the up-crossing inequality, then number of up-crossings of any interval $[a,b]$ of an L^1-bounded martingale has finite expectation, which then means that it is a.s. finite. Assume that X_n doesn't converge. This would then mean that with positive probability, either there will be an interval that will be crossed infinitely often or that $|X_n|\to\infty$. The first case cannot happen as we already said that the up-crossings of any interval are a.s. finite. The second case cannot happen because it would then mean that for any ℓ

$$0 < \mathrm{P}\big(\lim_{n\to\infty}|X_n|\geq \ell\big) \leq \liminf_{n\to\infty}\mathrm{P}(|X_n|\geq \ell) \leq \liminf_{n\to\infty}\frac{1}{\ell}\mathrm{E}[|X_n|],$$

which by the uniform boundedness $\sup_n \mathrm{E}[|X_n|] < \infty$ and by taking ℓ large enough, leads to contradiction. In the first inequality we used Fatou's lemma and in the second Markov's inequality. ☐

Corollary 9.16. *If $(X_n)_{n\geq 1}$ is a non-negative super-martingale, then it converges a.s.*

Proof. If $X_n \geq 0$, then $|X_n| = X_n$ and then the up-crossing inequality writes

$$\mathrm{E}\big[U_n(a,b)\big] \leq \frac{1}{b-a}\Big(a + \mathrm{E}[X_n]\Big),$$

while by the super-martingale property the RHS is bounded by $\frac{1}{b-a}(a + \mathrm{E}[X_1])$, which then leads to the conclusion of a.s. finite number of up-crossings. At this point we can follow the steps of the previous corollary. ☐

Exercise 30. *State the analogous to the last corollary for sub-martingales.*

Exercise 31. *What would you need to have an L^p convergence for martingales with $p \in (1,\infty)$.*

9.4. MARTINGALES AND MARKOV CHAINS.

REFERENCES

[B] L. Breiman, Probability, *Classics in Applied Mathematics, vol. 7, Society for Industrial and Applied Mathematics (SIAM)*, Philadelphia, PA, 1992.
[C] S. Chatterjee, Stein's methods and applications, `https://souravchatterjee.su.domains/AllLectures.pdf`
[dH] F. Hollander, Large deviations *American Mathematical Soc.*, (2000)
[D] R. Durrett, Probability: theory and examples *Cambridge university press.* Vol. 49, (2019).
[RY] D. Revuz, M. Yor, Continuous martingales and Brownian motion, *Springer Science & Business Media.*, Vol. 293 (2013).
[V] S.R.S. Varadhan, Probability theory, *American Mathematical Soc.*, No. 7, 2001
[Z] N. Zygouras, Discrete Stochastic Analysis, `https://warwick.ac.uk/fac/sci/maths/people/staff/zygouras/research_work/discrete_stochastc_analysis2.pdf`

DEPARTMENT OF STATISTICS, UNIVERSITY OF WARWICK, COVENTRY CV4 7AL, UK
Email address: `N.Zygouras@warwick.ac.uk`