1. Review of 275A material: foundations

We begin by a brief review of the important topics treated in MATH 275A, starting with the foundations and going all the way up to and including independence. Further topics from 275A will be reviewed in the next lecture.

1.1 The Kolmogorov model.

Probability theory has long existed outside mathematics, lacking a proper foundation and axiomatics. This was mended in 1933 by A.N. Kolmogorov, who embedded probability into measure theory. We will follow this approach while noting that it only covers certain uses of probability. Indeed, the Kolmogorov model does not include the probabilistic framework needed to interpret measurements of non-commutative observables in quantum theory. Various interpretations and models of probability also arise in philosophy and logic but we will not go into these as they do not really have much bearing on what we want to do in mathematics.

The Kolmogorov model of probability is phrased in terms of the following concept:

Definition 1.1 *A* probability space *is a triplet* $(\Omega, \mathcal{F}, P)$ *with the following structure:*

- Ω *is a non-empty set*
- $\mathcal{F}$ *is a* σ*-algebra of subsets of* Ω
- P *is a probability measure on* $(\Omega, \mathcal{F})$

The interpretation of Ω is the set of "possible outcomes" of a random experiment. The sets in $\mathcal{F}$ are referred to as *events*; these represent the questions we are allowed to ask about the outcomes. We interpret $P(A)$ as the likelihood, odds, relative chance or, technically, probability that event A occurs.

The requirement that P is a measure means that it is a map $P\colon \mathcal{F} \to [0, \infty]$ which is countably additive on disjoint unions of events and obeys $P(\varnothing) = 0$. To make it into a probability measure we need that $P(\Omega) = 1$, reflecting on Ω containing the totality of all possible outcomes and P being interpreted as the relative chance. The assumption of finite additivity is actually very reasonable given the interpretation of P as a likelihood. Boosting this to countable additivity is mainly a technical requirement that enables powerful tools from mathematical analysis and, in particular, measure theory. Hardly any limit results would be available in probability without countable additivity.

An important concept in intuitive probability is that of a *random variable* which in its basic form refers to a random numerical value that is somehow associated with the random experiment. In the Kolmogorov model, this is realized by a map $X\colon \Omega \mapsto \mathbb{R}$ such that

$$\forall I \subseteq \mathbb{R} \text{ interval}\colon\ X^{-1}(I) \in \mathcal{F} \tag{1.1}$$

One then uses the notation

$$P(X \in B) := P(X^{-1}(B)) \tag{1.2}$$

to denote the probability that "X outputs a result in a set B." While this is *a priori* defined only for B being an interval, a standard argument from measure theory shows that X

obeying (1.1) is automatically *measurable*, which means that $X^{-1}(B) \in \mathcal{F}$ for all sets B in the least σ-algebra containing all the intervals, to be called the Borel σ-algebra $\mathcal{B}(\mathbb{R})$.

Since composition of measurable functions is measurable, with X an $\mathbb{R}$-valued random variable and $f\colon \mathbb{R} \to \mathbb{R}$ a Borel measurable function, also $f(X)$ is random variable. This being said, random variables need not be just real-valued; indeed, the same concept works in more generality. We thus set:

Definition 1.2 *Given a measurable space (S, Σ), an S-valued random variable is a map $X\colon \Omega \to S$ such that $X^{-1}(B) \in \mathcal{F}$ for each $B \in \Sigma$.*

The simplest non-trivial example of the more general setting is $S := \mathbb{R}^n$, in which case we refer to X as a *random vector*, but one can take S to be a set of sequences, functions, operators, measures, etc, provided one can endow these sets with a structure of a measurable space. This is often done with the help of topology, in which case we take Σ to be the Borel sets $\mathcal{B}(S)$ induced by a conveniently chosen topology on S.

An important, albeit mainly conventional, aspect of working in the Kolmogorov model is that, while a probability space is always assumed to be there, its particulars should not be relevant for the mathematical conclusions. Modulo notable exceptions such as theory building, probabilists therefore suppress references to the probability space altogether implying, tacitly, that whatever is chosen should work.

1.2 Measure theoretical tools.

A distinctive feature of the probabilistic approach (compared to, say, that of analysis) is that both the σ-algebra and the probability measure are variable parameters of the problem. This creates a need for efficient tools to construct rather general probability measures on rather general measurable spaces. Our next topic are the statements from measure theory that are typically invoked for this purpose.

Throughout we assume that Ω is a non-empty set and write $\mathcal{P}(\Omega)$ for the powerset of Ω; i.e., the set of all subsets of Ω. We start with:

Definition 1.3 *$\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is an* algebra *if*

- *$\varnothing, \Omega \in \mathcal{A}$, and*
- *$\mathcal{A}$ is closed under finite unions and complements (and thus finite intersections and set differences)*

Writing $\sigma(\mathcal{A})$ for the smallest (subset of $\mathcal{P}(\Omega)$ that is) σ-algebra containing $\mathcal{A}$, a key measure-extension theorem is then:

Theorem 1.4 (Hahn-Kolmogorov) *Let $\mathcal{A} \subseteq \Omega$ be an algebra and assume $\mu\colon \mathcal{A} \to [0, \infty]$ is*

- *finitely additive on $\mathcal{A}$,*
- *countably subadditive on $\mathcal{A}$*

Then μ is the restriction to $\mathcal{A}$ of a measure on $\sigma(\mathcal{A})$. This measure is unique if μ is finite (or even just σ-finite) on $\mathcal{A}$.

Since $\mathcal{A}$ is generally not closed under countable unions, the requirement of countable additivity is limited to disjoint collections of sets from $\mathcal{A}$ whose union lies in $\mathcal{A}$.

The conditions are thus minimal in the sense that they only prevent that P contains a contradiction to additivity or countable additivity already on $\mathcal{A}$.

The Hahn-Kolmogorov applies even when $\mathcal{A}$ is replaced by smaller collections of sets; notably, a so-called semialgebra:

Definition 1.5 *$\mathcal{S} \subseteq \mathcal{P}(\Omega)$ is a* semialgebra *if*

- $\varnothing, \Omega \in \mathcal{S}$
- $\forall A, B \in \mathcal{S}\colon\ A \cap B \in \mathcal{S}$
- $\forall A \in \mathcal{S}\ \exists B_1, \ldots, B_n \in \mathcal{S}$ *disjoint*$\colon\ A^{\mathrm{c}} = \bigcup_{i=1}^{n} B_i$

Since semialgebras are not generally closed even under finite unions, both finite and countable additivity is then restricted to disjoint unions that themselves lie in $\mathcal{S}$.

The salient examples of semi-algebras include the set of half-open intervals

$$\{(a, b]\colon\ a < b\} \tag{1.3}$$

in $\mathbb{R}$ or the set of measurable rectangles

$$\{A \times B\colon\ A, B \in \Sigma\} \tag{1.4}$$

in S^2, for (S, Σ) a measurable space. The point here is that the underlying measure is often defined very canonically on the elements of the semi-algebra, e.g.,

$$\mu\big((a, b]\big) := F(b) - F(a) \tag{1.5}$$

as used heavily in the construction of Radon measures on $\mathbb{R}$, or

$$\mu(A \times B) := \mu_1(A)\mu_2(B), \tag{1.6}$$

as used in the construction of the product measure $\mu := \mu_1 \otimes \mu_2$.

The existence part of the Hahn-Kolmogorov theorem is a consequence of the beautiful Carathéodory Extension Theorem that is a main tool in abstract measure theory. We refer the reader to the relevant textbooks for the statement and proof.

In order to explain where the uniqueness part of the Hahn-Kolmogorov theorem comes from, we introduce the following concepts:

Definition 1.6 *$\mathcal{P} \subseteq \mathcal{P}(\Omega)$ is a* π-system *if* $\forall A, B \in \mathcal{P}\colon\ A \cap B \in \mathcal{P}$

Definition 1.7 *$\mathcal{L} \subseteq \mathcal{P}(\Omega)$ is a* λ-system *if*

- $\varnothing, \Omega \in \mathcal{L}$
- $\forall A, B \in \mathcal{L}\colon\ A \subseteq B \ \Rightarrow\ B \smallsetminus A \in \mathcal{L}$
- $\forall \{A_n\}_{n \geqslant 1} \subseteq \mathcal{L}\colon A_n \uparrow A \ \Rightarrow\ A \in \mathcal{L}$

Examples of π-systems are semi-algebras (and thus algebras and σ-algebras. Examples of λ-systems include any σ-algebra on Ω and, more importantly,

$$\big\{A \in \mathcal{F}\colon\ \mu(A) = \nu(A)\big\} \text{ for } \mu, \nu \text{ finite measures on } (\Omega, \mathcal{F}) \tag{1.7}$$

We then need a beautiful, albeit formal and somewhat unintuitive observation:

Theorem 1.8 (Dynkin's π/λ-theorem) *If $\mathcal{P}$ is π-system and $\mathcal{L}$ is λ-system, then*

$$\mathcal{P} \subseteq \mathcal{L} \ \Rightarrow\ \sigma(\mathcal{P}) \subseteq \mathcal{L} \tag{1.8}$$

Hence we get that if two finite measures agree on a semi-algebra $\mathcal{S}$, which is a π-system, then by (1.7) they agree on a λ-system that includes $\sigma(\mathcal{S})$. Alternative approaches to such questions can be based on the Monotone class theorem for sets but going via Theorem 1.8 is usually more efficient.

1.3 Integration and expectation.

A notable benefit of the Kolmogorov model is the availability of the Lebesgue integration theory. Recall that the Lebesgue integral over a general measure space $(\Omega, \mathcal{F}, \mu)$ is defined in three stages.

- *Integral of simple functions*: For any $n \geqslant 1$, $A_1, \dots, A_n \in \mathcal{F}$ and $a_1, \dots, a_n \in \mathbb{R}$,
$$\varphi = \sum_{i=1}^{n} a_i 1_{A_i} \;\mapsto\; \int \varphi \, \mathrm{d}\mu := \sum_{i=1}^{n} a_i \mu(A_i) \tag{1.9}$$
where (as one has to check using finite additivity of μ) the sum on the right does not depend on the representation of φ.
- *Unsigned integral*: For any $f : \Omega \to [0, \infty]$,
$$\int f \mathrm{d}\mu := \sup\left\{ \int \varphi \, \mathrm{d}\mu : 0 \leqslant \varphi \leqslant f \text{ simple} \right\} \tag{1.10}$$
where the supremum lies in $[0, \infty]$ due to $\varphi := 0$ being simple.
- *Signed integral*: For any $f_+, f_- : \Omega \to [0, \infty]$,
$$f = f^+ - f^- \;\mapsto\; \int f \mathrm{d}\mu := \int f^+ \mathrm{d}\mu - \int f^- \mathrm{d}\mu \tag{1.11}$$
whenever the expression on the right is not of the type $\infty - \infty$.

While these steps assign a numerical value to (most) $f : \mathbb{R} \to [-\infty, \infty]$, one has to restrict f further to get a reasonably behaved object. This is the content of:

Lemma 1.9 *When restricted to the class of Borel measurable functions $f : \Omega \to \mathbb{R}$, the map $f \mapsto \int f \mathrm{d}\mu$ is additive and homogeneous.*

Note that, since infinite value of the integral is allowed, the integral may exist (i.e., is well defined as a signed integral above) and yet be divergent. While we worked with real-valued functions, the Bochner integral provides the corresponding extension for f taking values in a Banach space but this will not be needed in this course.

A great feature of the Lebesgue integration theory (as opposed to other theories) is the behavior of the integral under pointwise limits. We summarize these in:

Theorem 1.10 *Let $(\mathscr{X}, \mathcal{G}, \mu)$ be a measure space with μ not necessarily finite and let $\{f_n\}_{n \geqslant 1}$ and f be real-valued measurable functions on $\mathscr{X}$ such that $f_n \to f$ pointwise and $\int f_n \mathrm{d}\mu$ exists for all $n \geqslant 1$. Then $\int f \mathrm{d}\mu$ exists and*
$$\int f_n \mathrm{d}\mu \underset{n \to \infty}{\longrightarrow} \int f \mathrm{d}\mu \tag{1.12}$$
holds under any of the following three conditions:

- *(Bounded Convergence Theorem) μ is finite and $\{f_n\}_{n\geqslant 1}$ are uniformly bounded,*
- *(Monotone Convergence Theorem) $\{f_n\}_{n\geqslant 1}$ are non-negative and pointwise non-decreasing,*
- *(Dominated Convergence Theorem) there exists a measurable function $g\colon \mathscr{X} \to [0,\infty)$ with $\int g\mathrm{d}\mu < \infty$ and $|f_n| \leqslant g$ for all $n \geqslant 1$.*

Moving back to probability, we introduce the corresponding name and notation for the Lebesgue integral:

Definition 1.11 (Expectation) *Given a real-valued random variable X, its* expectation *is defined by*

$$EX := \int X\mathrm{d}P \tag{1.13}$$

provided the integral exists.

We say that X has finite expectation if $E|X| < \infty$. (This makes sense for all X as an unsigned integral.) A problem with the above definition is its reliance on the probability space whose particulars, as we noted above, should not be of much relevance. For this we introduce:

Definition 1.12 *Let (S,Σ) be a measurable space. Given an S-valued random variable X, its* distribution *is a measure on (S,Σ) such that*

$$\forall B \in \Sigma\colon\ \mu_X(B) := P(X \in B) \tag{1.14}$$

For real-valued X, the distribution μ_X is thus a Borel probability measure on $\mathbb{R}$. This measure is related to the cumulative distribution function (CDF)

$$F(x) := P(X \leqslant x) \tag{1.15}$$

of X via (1.5). Using this measure we then get:

Theorem 1.13 (Change of variables formula) *Let (S,Σ) be a measurable space and X an S-valued random variable. Then for any Borel measurable $f\colon S \to \mathbb{R}$,*

$$Ef(X) = \int f\mathrm{d}\mu_X \tag{1.16}$$

whenever one of (and thus both) integrals exist(s).

The proof of this goes by checking this for simple functions and then invoking the Monotone Class Theorem for functions to extend it to general measurable f. Note that, for real-valued X, the integral is now over $\mathbb{R}$.

We now quickly introduce some standard concepts:

Definition 1.14 *For any $\alpha \in \mathbb{R}$ and any random variable X, we call $E(|X|^\alpha)$ the α-th absolute moment of X. If $X \geqslant 0$ we talk about α-th moment.*

Note that, by Jensen's inequality,

$$\left\{\alpha \in \mathbb{R}\colon E(|X|^\alpha) < \infty\right\} = \text{ interval containing } 0 \tag{1.17}$$

This naturally leads to *L^p-spaces* of random variables defined by

$$L^p = L^p(\Omega, \mathcal{F}, P) := \big\{X\colon \text{random variable } \wedge\ E(|X|^p) < \infty\big\}. \tag{1.18}$$

A convenient fact (due to P being a finite measure) is that $p \mapsto L^p$ is non-increasing under set inclusion (and the spaces are thus nested). We also put forward:

Definition 1.15 *For $X \in L^2$, we define the* variance of X *by*

$$\mathrm{Var}(X) := E\big((X - EX)^2\big) = E(X^2) - (EX)^2 \tag{1.19}$$

While EX is a (possibly poor) way to express where the distribution of X is centered, $\mathrm{Var}(X)$ tells us about how much the distribution spreads around EX. Other important expectations include the characteristic function $\varphi_X(t) := Ee^{itX}$ and the moment generating function $M_X(t) := Ee^{tX}$. We will return to these when these become relevant.

We conclude the discussion of integration by recalling the names labeling some key inequalities: Cauchy, Hölder, Minkowski, Jensen, Markov, Chebyshev. The reader will surely be able to find the statements of these in the relevant literature.

1.4 Independence.

A very important notion in probability is that of independence. Proceeding using the standard route taken in most textbooks, we start by introducing a special case:

Definition 1.16 (Pairwise independence)

$$A, B \in \mathcal{F} \text{ are independent if } P(A \cap B) = P(A)P(B) \tag{1.20}$$

The idea behind the name is, using $P(A|B) := P(A \cap B)/P(B)$ (which assumes that $P(B) > 0$) to denote the conditional probability of A given that B occurs, under independence we have $P(A|B) = P(A)$. The interpretation is that, under independence, no statistical information about A can be inferred from the knowledge that B has occurred. The notion is clearly symmetrical, i.e., if A, B independent then so are B, A, and it extends to complements: A^c, B are independent, A^c, B^c are independent, etc.

The problem with pairwise independence is that it does not make it clear how to extend it to more than two events. This is the content of:

Definition 1.17 (Independence) *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let I be a non-empty set. We then say:*

- Events $\{A_\alpha\}_{\alpha\in I}$ *are independent if*

$$\forall J \subseteq I \text{ finite}: \ P\Big(\bigcap_{\alpha\in J} A_\alpha\Big) = \prod_{\alpha\in J} P(A_\alpha) \tag{1.21}$$

- Random variables $\{X_\alpha\}_{\alpha\in I}$ *are independent if*

$$\forall \{B_\alpha\}_{\alpha\in I} \in \mathcal{B}(\mathbb{R})^I\colon \quad \text{events } \big\{\{X_\alpha \in B_\alpha\}\big\}_{\alpha\in I} \text{ are independent} \tag{1.22}$$

- Families of events $\{\mathcal{C}_\alpha\}_{\alpha\in I}$ *are independent if*

$$\forall \{A_\alpha\}_{\alpha\in I} \in \mathcal{F}^I\colon \Big(\forall \alpha \in I\colon A_\alpha \in \mathcal{C}_\alpha\Big) \Rightarrow \{A_\alpha\}_{\alpha\in I} \text{ independent} \tag{1.23}$$

Note that, in particular, in order to check that three events A, B and C are independent, besides their pairwise independence we also have to check that $P(A \cap B \cap C) = P(A)P(B)P(C)$. There is a standard example that shows that one can have three pairwise independent events that are not independent.

A typical example of $\{\mathcal{C}_\alpha\}_{\alpha \in I}$ in (1.23) is a family of σ-algebras. Note that we only require checking (1.21) for finite J. This is because the expression is then trivially true for J countable and (possibly) meaningless for J uncountable.

Note that if $X_1, \ldots, X_n$ are random variables, then the *random vector*

$$\mathbb{X} := (X_1, \ldots, X_n) \tag{1.24}$$

defines a measurable map $\Omega \to \mathbb{R}^n$ — with measurability on $\mathbb{R}^n$ now with respect to the Borel σ-algebra $\sigma(\mathbb{R}^n) := \sigma(\{O \subseteq \mathbb{R}^n \colon \text{ open}\})$. We then define the *joint distribution* of $(X_1, \ldots, X_n)$ as the distribution of $\mathbb{X}$, i.e.,

$$\mu_{\mathbb{X}}(B) := P\big((X_1, \ldots, X_n) \in B\big) \tag{1.25}$$

Since each X_i is a real-valued random variable, it also induces its distribution μ_{X_i} on $\mathbb{R}$ which, in this context, is referred to as the i-th *marginal* of $\mu_{\mathbb{X}}$. We then have

Lemma 1.18

$$X_1, \ldots, X_n \text{ independent} \;\Leftrightarrow\; \mu_{\mathbb{X}} = \mu_{X_1} \otimes \cdots \otimes \mu_{X_n} \tag{1.26}$$

A convenient tool for integration with respect to product measures is the *Fubini-Tonelli's Theorem* whose general statement we refer the reader to to the literature. For random variables this theorem reads:

Theorem 1.19 (Fubini-Tonelli for random variables) *For $X_1, \ldots, X_n$ independent,*

$$E\big(X_1 \ldots X_n\big) = \prod_{i=1}^{n} EX_i \tag{1.27}$$

when either $\forall i \colon X_i \geqslant 0$ or $\forall i \colon E|X_i| < \infty$.

Similarly as independence is stronger than pairwise independence, the conclusion of the previous theorem implies, but is not equivalent to:

Definition 1.20 *The random variables $X_1, \ldots, X_n \in L^2$ are said to be* uncorrelated *if*

$$\forall i < j \colon\; E(X_i X_j) = (EX_i)(EX_j) \tag{1.28}$$

A very notable exception to this are Gaussian (a.k.a. multivariate normal) random variables for which being uncorrelated does imply independence. We will review this when this becomes relevant.

An important construction from measure theory allows us to construct a measure space that is a product of any finite number of probability spaces. This in particular permits us to put any finite number of random variables as independent on the same probability space. However, the construction is not limited to finite products:

Theorem 1.21 *For any sequence $\{\mu_n\}_{n \geqslant 1}$ of probability measures on $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$ there exists a probability space $(\Omega, \mathcal{F}, P)$ supporting random variables $\{X_n\}_{n \geqslant 1}$ such that*

- $\forall n \geqslant 1$: *μ_n is the distribution of X_n, and*
- *$\{X_n\}_{n\geqslant 1}$ are independent*

Proof (sketch). Let $I := [0,1)$ and consider the probability space $(I, \mathcal{B}(I), \lambda)$ where λ is the Lebesgue measure on I. This space supports a uniform random variable U defined by the identity map. Now define random variables $\{Z_n\}_{n\geqslant 1}$ by

$$Z_n := \lfloor 2^n U \rfloor \bmod 2 \tag{1.29}$$

and check that these are independent Bernoulli (i.e., $0,1$-valued with mean $1/2$). Then define random variables $\{U_n\}_{n\geqslant 1}$ by

$$U_n := \sum_{i\geqslant 1} Z_{2^n i + 2^{n-1}} 2^{-i} \tag{1.30}$$

and check that these are independent copies of U. Finally, for each $n \geqslant 1$ denote

$$F_n(x) := \mu_n\big((-\infty, x]\big) \tag{1.31}$$

and set

$$X_n := F_n^{-1}(U_n) \tag{1.32}$$

where $F_n^{-1}(u) := \inf\{x \in \mathbb{R} \colon u \leqslant F_n(x)\}$, and check that $\{X_n\}_{n\geqslant 1}$ are independent with μ_n being the distribution of X_n. □

The reason why we included the above proof is to demonstrate that there is no need to invoke the tool that is frequently called upon to get this result: the Kolmogorov Extension Theorem. This theorem permits putting general (even non-product) measures on product spaces over general (even uncountable) index sets but there is price to pay: one has to assume that the individual measurable spaces in the product are standard Borel. For product spaces, a different argument based, vaguely, on projection limits allows us to construct a product of full measure spaces (not just a family of random variables on these) regardless of their underlying structure and the cardinality of the index set (thus avoiding the Kolmogorov Extension Theorem altogether).

Further reading: Durrett, Chapter 1 and Appendix A

2. Review of 275A material: convergence theorems

We will now move to the discussion of a number of limit results from introductory probability theory. For most of these we only give their statements referring the reader to textbooks for proofs.

2.1 Weak laws of of large numbers.

We start by the so called Weak Law of Large Numbers (WLLN) whose informal statement goes as far back as G. Cardano (16th century) and whose first proof for the special case of zero-one valued random variables we owe to J. Bernoulli (early 18th century). The following form is largely due to P.L. Chebyshev and A.Y. Khinchin:

Theorem 2.1 (WLLN) *Let $X_1, X_2, \dots$ be i.i.d. and set $S_n := X_1 + \cdots + X_n$. Then the following are equivalent:*

(1) *there exists $\{a_n\}_{n\geqslant 1} \in \mathbb{R}^{\mathbb{N}}$ such that*

$$\forall \epsilon > 0\colon \quad P\left(\left|\frac{S_n}{n} - a_n\right| > \epsilon\right) \underset{n\to\infty}{\longrightarrow} 0 \tag{2.1}$$

(2) $xP(|X_1| > x) \underset{x\to\infty}{\longrightarrow} 0$

If $E|X_1| < \infty$, the above holds with $a_n := EX_1$ for all $n \geqslant 1$.

Note that the statement (2.1) concerns a specific mode of convergence that we record for future use:

Definition 2.2 *Let $\{Y_n\}_{n\geqslant 1}$ and Y be random variables. We say that Y_n converges to Y* in probability*, with notation $Y_n \xrightarrow{P} Y$, if*

$$\forall \epsilon > 0\colon \quad P(|Y_n - Y| > \epsilon) \to 0 \tag{2.2}$$

A short way to state (2.1) is thus by saying $S_n/n - a_n \xrightarrow{P} 0$ and, when $E|X_1| < \infty$, by saying $S_n/n \xrightarrow{P} EX_1$. In either case, the statement implies that the distribution of S_n/n becomes increasingly squeezed, or *concentrated*, at around a deterministic value a_n.

The particular choice when X_1 is an indicator of an event A may serve as a justification of the Kolmogorov model: Indeed, S_n is then the number the event A occurred in n independent samples and S_n/n is the relative frequency of occurrence. The WLLN says that the relative frequency converges to $EX_1 = P(A)$, the probability of event A.

The proof of (2.1) reduces to an elementary application of the Chebyshev inequality when $X_1 \in L^2$. (This is the part that goes back to Chebyshev.) For the general cases we need to first show that, under (2),

$$P(\max_{1\leqslant i\leqslant n} |X_i| > n) \underset{n\to\infty}{\longrightarrow} 0. \tag{2.3}$$

This permits us to truncate the random variables and apply the Chebyshev inequality again. To demonstrate why the tail decay (2) is actually necessary note that, for X_1 Cauchy also S_n/n is Cauchy and so (2.1) fails.

The conditions under which (1) of Theorem 2.1 holds can be generalized multiple ways. Instead of one sequence, we can work with triangular arrays (i.e., a sequence of n independent random variables for each n). Since the proof relies only on a second-moment calculation, one does not even need independence; it suffices that the random variable are uncorrelated.

2.2 Borel-Cantelli lemmas and the Strong law.

As observed by A.N. Kolmogorov based on earlier works of E. Borel and F. Cantelli, the convergence in probability in the WLLN can be augmented using techniques for dealing with infinite families of events. We start with:

Definition 2.3 *Given events $\{A_n\}_{n\geqslant 1}$ we set*

$$\{A_n \text{ i.o.}\} := \bigcap_{n\geqslant 1} \bigcup_{k\geqslant n} A_k \tag{2.4}$$

to denote the event that "A_n occurs infinitely often."

Alternative notations for $\{A_n \text{ i.o.}\}$ are $\limsup_{n\to\infty} A_n$ and $\{\sum_{n\geqslant 1} 1_{A_n} = \infty\}$ with the latter perhaps explaining better the reason for the name. We now use this to give a concise definition of the probabilist's version of almost-everywhere convergence:

Definition 2.4 *Let $\{Y_n\}_{n\geqslant 1}$ and Y be random variables. We say that Y_n tends to Y almost surely, with notation "$Y_n \to Y$ a.s." if*

$$\forall \epsilon > 0\colon \quad P\big(|Y_n - Y| > \epsilon \text{ i.o.}\big) = 0 \tag{2.5}$$

To see that this is really the same as a.e.-convergence, note that

$$\big\{Y_n \to Y \text{ pointwise}\big\} = \bigcap_{m\geqslant 1} \big\{|Y_n - Y| > 2^{-m} \text{ i.o.}\big\}^{\mathrm{c}} \tag{2.6}$$

The event on the right has full probability when (2.5) occurs.

As it turns out, the lack of occurrence of $\{A_n \text{ i.o.}\}$ admits a simple sufficient condition:

Lemma 2.5 (1st Borel-Cantelli)

$$\sum_{n\geqslant 1} P(A_n) < \infty \;\Rightarrow\; P(A_n \text{ i.o.}) = 0 \tag{2.7}$$

For independent events, this can be boosted to full characterization:

Lemma 2.6 (2nd Borel-Cantelli)

$$\sum_{n\geqslant 1} P(A_n) = \infty \;\wedge\; \{A_n\}_{n\geqslant 1} \text{ independent} \;\Rightarrow\; P(A_n \text{ i.o.}) = 1 \tag{2.8}$$

In particular, we get our first zero-one law:

Corollary 2.7 *Let $\{A_n\}_{n\geqslant 1}$ be independent events. Then*

$$P(A_n \text{ i.o.}) \in \{0, 1\}. \tag{2.9}$$

A typical situation demonstrating the use of the above statements comes in an improvement of (2.3) to the form:

Corollary 2.8 *Let $X_1, X_2, \dots$ be i.i.d. random variables. Then for all $A > 0$,*

$$P\big(|X_n| > An \text{ i.o.}\big) = \begin{cases} 0, & \text{if } E|X_1| < \infty \\ 1, & \text{if } E|X_1| = \infty \end{cases} \tag{2.10}$$

In words, this says that a sequence of i.i.d. random variables will grow sublinearly if the first absolute moment is finite and superlinearly (along a subsequence) otherwise. By applying this to powers of random variables, similarly we get that

$$P\big(|X_n| > An^{1/\alpha} \text{ i.o.}\big) = \begin{cases} 0, & \text{if } E\big(|X_1|^\alpha\big) < \infty \\ 1, & \text{if } E\big(|X_1|^\alpha\big) = \infty \end{cases} \tag{2.11}$$

so existence of moments gives us bounds on growth rate of i.i.d. sequences. Note that the bound is qualitative.

With the Borel-Cantelli lemmas at our disposal, one then improve the convergence in probability in the WLLN to almost-sure convergence:

Theorem 2.9 (SLLN) *Let $X_1, X_2, \dots$ be i.i.d. such that EX_1 exists (possibly equal to $\pm\infty$). Set $S_n := X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \underset{n\to\infty}{\longrightarrow} EX_1 \text{ a.s.} \tag{2.12}$$

Without going into details, we note that, for $X_1 \in L^4$, a proof exists that is based on the Chebyshev inequality (for the fourth moment of S_n) and the 1st Borel-Cantelli lemma. (Specifically, assuming $EX_1 = 0$ we estimate $P(|S_n| > \epsilon n)$ by a quantity of order n^{-2} which is summable on n.) This observation already implies the *Borel SLLN* which is a special case of the above for X_1 being an indicator of an event.

If we aim for the statement under the condition $X_1 \in L^1$ (the extension to the case stated above is then achieved by suitable truncation when EX_1 diverges), which was first made by Kolmogorov in 1930s, the argument usually presented (due to N. Etemadi) invokes truncation and the Chebyshev inequality to first extract convergence along exponentially growing subsequences. A renewal-type argument is then invoked (under a restriction to positive sequences) to "fill the gaps" and get full convergence.

Many alternative proofs exist, some of which work in larger generality than just i.i.d. sequences. Indeed, the SLLN is a special case of *de Finetti's Theorem* on a.s. convergence of exchangeable sequences. The SLLN can be proved using *Doob's Martingale Convergence Theorem*. Another theorem that subsumes the SLLN is *Birkhoff's Ergodic Theorem*. While our goal is to present these advanced results sometime later this quarter, in the next subsection we will present a proof that is more or less the same "level" as the SLLN, relying only on the convergence of random series.

We make a few additional observations. The first one (also due to Kolmogorov) demonstrates the necessity of finite first moments:

Lemma 2.10 *Let $X_1, X_2, \dots$ be i.i.d. and $S_n := X_1 + \cdots + X_n$. If $\{S_n/n\}_{n\geqslant 1}$ converges to a finite random variable on a set of positive probability, then $E|X_1| < \infty$.*

Proof. Note that

$$\frac{X_n}{n} = \frac{S_n}{n} - \frac{n-1}{n}\frac{S_{n-1}}{n-1} \tag{2.13}$$

The convergence of S_n/n to a finite (random) limit on a set of positive measure thus implies a similar convergence of X_n/n. By Corollary 2.8 this forces $E|X_1| < \infty$. □

As a corollary of the SLLN we get:

Theorem 2.11 (Glivenko-Cantelli) *Let $X_1, X_2, \dots$ be i.i.d. and, for $n \geqslant 1$ and $t \in \mathbb{R}$, set*

$$F_n(t) := \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leqslant t\}} \tag{2.14}$$

Denote $F(t) := P(X_1 \leqslant t)$. Then

$$\sup_{t\in\mathbb{R}} \big|F_n(t) - F(t)\big| \underset{n\to\infty}{\longrightarrow} 0 \text{ a.s.} \tag{2.15}$$

To put this in words, the empirical CDF generated by the first n terms of the sequence $X_1, X_2, \dots$ converges to the actual CDF uniformly on $\mathbb{R}$. (This mode of convergence corresponds to the Kolmogorov distance on probability measures on $\mathbb{R}$ so Theorem 2.11 in fact says that the empirical law $\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ tends to the distribution of X_1 in the Kolmogorov distance.) As to the proof, the SLLN implies $F_n(t) \to F(t)$ a.s. for each t (with the implicit null set possibly depending on t) so the main issue is to achieve uniformity. Here the monotonicity and right-continuity of F_n and F play a key role.

2.3 Convergence of random series.

Another set of convergence theorems arises while studying convergence of infinite series associated with sequences of independent random variables. A simple application of the Monotone Convergence Theorem shows

$$\sum_{k=1}^{\infty} E|X_k| < \infty \;\Rightarrow\; \sum_{k=1}^{\infty} |X_k| < \infty \text{ a.s.} \tag{2.16}$$

However, we are interested mainly in the conditions in which the series $\sum_{k=1}^{\infty} X_k$ converges conditionally, but not absolutely. A simple result in this vain is:

Theorem 2.12 (Paley-Zygmund) *Let $X_1, X_2, \dots$ be independent, uniformly bounded and symmetric (i.e, $\forall i \geqslant 1\colon X_i \overset{\text{law}}{=} -X_i$). Then*

$$\sum_{i=1}^{\infty} X_i \text{ converges in } \mathbb{R} \text{ a.s.} \;\Leftrightarrow\; \sum_{i\geqslant 1} E(X_i^2) < \infty \tag{2.17}$$

Here we say "in $\mathbb{R}$" to make it unequivocally clear that we do not allow convergence to $\pm\infty$. The proof of sufficiency of the summability condition relies on Kolmogorov's

Maximal Inequality (boundedness is not needed in this step); the reverse direction in turn uses the Paley-Zygmund inequality. General series are handled by:

Theorem 2.13 (Kolmogorov 3-series theorem) *Let $X_1, X_2, \dots$ be independent. Then $\sum_{i=1}^{\infty} X_i$ converges in $\mathbb{R}$ a.s. if and only if $\exists A > 0$ such that, for $Y_i := X_i 1_{|X_i| \leqslant A}$,*

$$\sum_{i \geqslant 1} P(|X_i| > A) < \infty \ \wedge \ \sum_{i \geqslant 1} \mathrm{Var}(Y_i) < \infty \ \wedge \ \sum_{i=1}^{\infty} EY_i \text{ converges in } \mathbb{R} \tag{2.18}$$

We will now use this theorem to give:

Proof of the SLLN in Theorem 2.9. Let $\{X_k\}_{k \geqslant 1}$ be i.i.d. with $E|X_1| < \infty$. For all $k \geqslant 1$, let

$$Z_k := X_k 1_{\{|X_k| \leqslant k\}} \tag{2.19}$$

The 1st Borel-Cantelli lemma shows

$$P\big(Z_n \neq X_n \text{ i.o.}\big) = 0 \tag{2.20}$$

which means that it suffices to prove the claim with $\{Z_k\}_{k \geqslant 1}$ replacing $\{X_k\}_{k \geqslant 1}$. First we show that

$$\sum_{k=1}^{\infty} \frac{Z_k - EZ_k}{k} \text{ converges in } \mathbb{R} \text{ a.s.} \tag{2.21}$$

We will prove this by calling on Theorem 2.13 which requires checking the three-series conditions (2.18). In light of $(Z_k - EZ_k)/k$ being bounded by 2 and having zero mean, all we have to show is that

$$\sum_{k \geqslant 1} \mathrm{Var}\Big(\frac{Z_k}{k}\Big) < \infty \tag{2.22}$$

For this we compute

$$\begin{aligned} \sum_{k \geqslant 1} \mathrm{Var}\Big(\frac{Z_k}{k}\Big) &\leqslant \sum_{k \geqslant 1} \frac{1}{k^2} E(Z_k) = \sum_{k \geqslant 1} \frac{1}{k^2} \int_0^{\infty} 2t P\big(|Z_k| > t\big) \mathrm{d}t \\ &\leqslant \sum_{k \geqslant 1} \frac{1}{k^2} \int_0^{k} 2t P\big(|X_1| > t\big) \mathrm{d}t = \int_0^{\infty} 2t \Big(\sum_{k \geqslant 1} \frac{1}{k^2} 1_{[0,k)}(t) \Big) P\big(|X_1| > t\big) \mathrm{d}t \end{aligned} \tag{2.23}$$

where the last step is by Tonnelli's theorem. The inequality $k + 1 \leqslant 2k$ for $k \geqslant 1$ along with an integral test for series gives

$$\sum_{k \geqslant 1} \frac{1}{k^2} 1_{[0,k)}(t) \leqslant \sum_{k \geqslant \lceil t \rceil} \frac{4}{(k+1)^2} \leqslant 4 \int_t^{\infty} \frac{1}{x^2} \mathrm{d}x \leqslant 4t^{-1} \tag{2.24}$$

Plugging this above then shows

$$\sum_{k \geqslant 1} \mathrm{Var}\Big(\frac{Z_k}{k}\Big) \leqslant 8 \int_0^{\infty} P\big(|X_1| > t\big) \mathrm{d}t = 8E|X_1| < \infty \tag{2.25}$$

thus proving (2.22).

With (2.21) established, we call upon:

Lemma 2.14 (Kronecker) *Let $\{a_n\}_{n\geqslant 1}$ and $\{b_n\}_{n\geqslant 1}$ be $\mathbb{R}$-valued sequences with*

$$\forall n \geqslant 1\colon\ b_{n+1} \geqslant b_n > 0 \quad \wedge \quad \lim_{n\to\infty} b_n = \infty \tag{2.26}$$

Then

$$\sum_{n=1}^{\infty} \frac{a_n}{b_n} \text{ converges in } \mathbb{R} \quad \Rightarrow \quad \frac{1}{b_n}\sum_{k=1}^{n} a_k \underset{n\to\infty}{\longrightarrow} 0 \tag{2.27}$$

Postponing the proof until the end of this proof, from (2.21) we then get

$$\frac{1}{n}\sum_{k=1}^{n}(Z_k - EZ_k) \underset{n\to\infty}{\longrightarrow} 0 \text{ a.s.} \tag{2.28}$$

But the Dominated Convergence Theorem shows

$$EZ_k = E\big(X_1 1_{\{|X_1|\leqslant k\}}\big) \underset{n\to\infty}{\longrightarrow} EX_1 \tag{2.29}$$

and so we conclude

$$\frac{1}{n}\sum_{k=1}^{n} Z_k \underset{n\to\infty}{\longrightarrow} EX_1 \text{ a.s.} \tag{2.30}$$

In light of (2.20), this implies the statement of the SLLN under finite first moment. (The case of $EX_1 = \pm\infty$ is handled by trunction.) □

It remains to give:

Proof of Lemma 2.14. Denote $s_n := \sum_{k=1}^{n} \frac{a_k}{b_k}$ with $s_0 := 0$. Then $a_k = b_k(s_k - s_{k-1})$ and summation by parts shows

$$\begin{aligned}\frac{1}{b_n}\sum_{k=1}^{n} a_k &= \frac{1}{b_n}\sum_{k=1}^{n} b_k(s_k - s_{k-1}) \\ &= s_n + \frac{1}{b_n}\sum_{k=1}^{n-1}(b_k - b_{k+1})s_k = \frac{b_1}{b_n}s_n + \sum_{k=1}^{n-1}\frac{b_k - b_{k+1}}{b_n}(s_k - s_n),\end{aligned} \tag{2.31}$$

where we used that $\sum_{k=1}^{n}(b_k - b_{k+1}) = b_1 - b_n$. Now observe that, since $b_n \to \infty$ while $\{s_n\}_{n\geqslant 1}$ converges and is thus bounded, the first term on the right tends to zero and so do any finite number of terms in the series on the right. For k_0 so large that $|s_k - s_n| < \epsilon$ for $k, n \geqslant k_0$, we in turn get

$$\left|\sum_{k=k_0}^{n-1}\frac{b_k - b_{k+1}}{b_n}(s_k - s_n)\right| \leqslant \epsilon \sum_{k=k_0}^{n-1}\frac{b_{k+1} - b_k}{b_n} \leqslant \epsilon. \tag{2.32}$$

The right-hand side of (2.31) thus tends to zero as $n \to \infty$. □

The above proof of the SLLN has the advantage that it generalizes to:

Theorem 2.15 (Marcinkiewicz-Zygmund SLLN) *Let $p \in (0, 2)$ and let $X_1, X_2, \ldots$ be i.i.d. with $X_1 \in L^p$. If $p \geqslant 1$ assume also $EX_1 = 0$. Then*

$$\frac{1}{n^{1/p}}\sum_{k=1}^{n} X_k \underset{n\to\infty}{\longrightarrow} 0 \text{ a.s.} \tag{2.33}$$

Proof. For $p = 1$ this is the SLLN so we may assume $p \neq 1$. We proceed very much like in the above proof of the SLLN. Denote

$$Z_k := X_k 1_{\{|X_k| \leqslant k^{1/p}\}} \tag{2.34}$$

Then (2.11) shows

$$P\big(Z_n \neq X_n \text{ i.o.}\big) = 0 \tag{2.35}$$

We now claim

$$\sum_{k=1}^{\infty} \frac{Z_k - EZ_k}{k^{1/p}} \text{ converges in } \mathbb{R} \text{ a.s.} \tag{2.36}$$

for which we invoke Theorem 2.13. It again suffices to show

$$\sum_{k \geqslant 1} \operatorname{Var}\Big(\frac{Z_k}{k^{1/p}}\Big) < \infty \tag{2.37}$$

Here the calculation (2.23) gets modified into

$$\begin{aligned} \sum_{k \geqslant 1} \operatorname{Var}\Big(\frac{Z_k}{k^{1/p}}\Big) &\leqslant \sum_{k \geqslant 1} \frac{1}{k^{2/p}} \int_0^{k^{1/p}} 2tP\big(|X_1| > t\big)\mathrm{d}t \\ &= \int_0^{\infty} 2t\Big(\sum_{k \geqslant 1} \frac{1}{k^{2/p}} 1_{[0,k^{1/p})}(t)\Big) P\big(|X_1| > t\big)\mathrm{d}t \end{aligned} \tag{2.38}$$

The inequality $k + 1 \leqslant 2k$ for $k \geqslant 1$ along with an integral test for series give

$$\sum_{k \geqslant 1} \frac{1}{k^{2/p}} 1_{[0,k^{1/p})}(t) \leqslant \sum_{k \geqslant t^p} \frac{2^{2/p}}{(k+1)^{2/p}} \leqslant 2^{2/p} \int_{t^p}^{\infty} \frac{1}{x^{2/p}} \mathrm{d}x \leqslant A(t^p)^{1-2/p} \tag{2.39}$$

where $A := 2^{2/p}(2/p - 1)^{-1}$. Plugging this in (2.38) yields

$$\sum_{k \geqslant 1} \operatorname{Var}\Big(\frac{Z_k}{k^{1/p}}\Big) \leqslant 2A \int_0^{\infty} t^{p-1} P\big(|X_1| > t\big)\mathrm{d}t = 2p^{-1}AE\big(|X_1|^p\big) \tag{2.40}$$

which is finite by our assumptions.

With (2.36) in hand, we now claim that

$$\sum_{k=1}^{\infty} \frac{|EZ_k|}{k^{1/p}} < \infty \tag{2.41}$$

For $p < 1$ we use that $EZ_k = E(X_1 1_{\{|X_1| \leqslant k^{1/p}\}})$ to bound

$$\sum_{k=1}^{\infty} \frac{|EZ_k|}{k^{1/p}} \leqslant E\Big(|X_1| \sum_{k=1}^{\infty} \frac{1}{k^{1/p}} 1_{\{|X_1| \leqslant k^{1/p}\}}\Big) \tag{2.42}$$

and observe that, by an integral test along with the fact that $1/p > 1$,

$$\sum_{k=1}^{\infty} \frac{1}{k^{1/p}} 1_{\{|X_1| \leqslant k^{1/p}\}} \leqslant C(p)\big(|X_1|^p\big)^{1-1/p} = C(p)|X_1|^{p-1} \tag{2.43}$$

for some constant $C(p) < \infty$. For $p > 1$ we in turn use that $EX_1 = 0$ to get

$$EZ_k = E(X_1 1_{\{|X_1| \leqslant k^{1/p}\}}) = E(X_1 1_{\{|X_1| > k^{1/p}\}}) \tag{2.44}$$

which then implies

$$\sum_{k=1}^{\infty} \frac{|EZ_k|}{k^{1/p}} \leqslant E\Big(|X_1| \sum_{k=1}^{\infty} \frac{1}{k^{1/p}} 1_{\{|X_1| > k^{1/p}\}}\Big) \tag{2.45}$$

As $1/p < 1$ we now have

$$\sum_{k=1}^{\infty} \frac{1}{k^{1/p}} 1_{\{|X_1| > k^{1/p}\}} \leqslant C(p)\big(|X_1|^p\big)^{1-1/p} = C(p)|X_1|^{p-1} \tag{2.46}$$

for some constant $C(p) < \infty$. Hence, in both cases we get

$$\sum_{k=1}^{\infty} \frac{|EZ_k|}{k^{1/p}} \leqslant C(p) E\big(|X_1|^p\big) \tag{2.47}$$

which is finite by our assumption.

Since (2.41) implies convergence of $\sum_{k=1}^{\infty} \frac{EZ_k}{k^{1/p}}$, using (2.36) we conclude

$$\sum_{k=1}^{\infty} \frac{Z_k}{k^{1/p}} \text{ converges in } \mathbb{R} \text{ a.s.} \tag{2.48}$$

Lemma 2.14 then gives

$$\frac{1}{n^{1/p}} \sum_{k=1}^{n} Z_k \underset{n\to\infty}{\longrightarrow} 0 \text{ a.s.} \tag{2.49}$$

In light of (2.35) we then get the claim. □

Further reading: Durrett, Chapter 2

3. Weak convergence and Lindeberg's CLT

In this section we still continue reviewing the 275-material; specifically, the part concerning the weak convergence of probability measures on the real line. We proceed by restating the standard version of the Central Limit Theorem and then state and prove its more enhanced form due to Lindeberg.

3.1 Weak convergence on the real line.

Earlier we reviewed the notions of convergence in probability and converence almost surely. Here we will discuss the notion of convergence in distribution, a.k.a., convergence in law or just weak convergence. The latter term should not to be confused with the notion of weak convergence in functional analysis (we will comment on the differences later). We start with a definition:

Definition 3.1 *Let $\{F_n\}_{n\geqslant 1}$ and F are non-decreasing and right continuous functions from $\mathbb{R}$ to $\mathbb{R}$. We say that F_n converges weakly to F, with notation $F_n \xrightarrow{\text{w}} F$, if*

$$\forall x \in D_F\colon\ F_n(x) \underset{n\to\infty}{\longrightarrow} F(x) \tag{3.1}$$

where $D_F := \{x \in \mathbb{R}\colon F \text{ continuous at } x\}$.

We now use the above to give the following derived notions:

Definition 3.2 *For Borel* probability measures $\{\mu_n\}_{n\geqslant 1}$ *and μ on $\mathbb{R}$, we say that μ_n* tends to μ weakly*, with notation $\mu_n \xrightarrow{\text{w}} \mu$, if the associated CDFs defined by $F_n(x) := \mu_n((-\infty, x])$ and $F(x) := \mu((-\infty, x])$ obey $F_n \xrightarrow{\text{w}} F$.*

Definition 3.3 *For $\mathbb{R}$-valued* random variables $\{X_n\}_{n\geqslant 1}$ *and X, we say that X_n* tends to X weakly*, with notation $X_n \xrightarrow{\text{w}} X$, if the distributions $\mu_n(A) := P(X_n \in A)$ and $\mu(A) := P(X \in A)$ associated with these random variables obey $\mu_n \xrightarrow{\text{w}} \mu$.*

The prime motivation for the first definition is the content of the last one: develop a mode of convergence which captures when the statistics of a sequence random variables somehow tends to a limit. Note that $\mu_n \xrightarrow{\text{w}} \mu$ if and only if $\mu((a, b]) \to \mu((a, b])$ for all $a, b \in D_F$. The set D_F is dense since $\mathbb{R} \setminus D_F$ is at most countable.

The exclusion of the discontinuity points is reasonable in light of the following example: Given a strictly decreasing sequence $\{a_n\}_{n\geqslant 1}$ of reals tending to some a, the CDF $F_n(x) = 1_{[a_n,\infty)}$ of random variables $X_n := a_n$ tends to $1_{(a,\infty)}$ which is not a CDF due to lack of right continuity at a. It is till reasonable to declare the limit random variable to be $X := a$, which has CDF $F(x) := 1_{[a,\infty)}$. Since weak convergence excludes the discontinuity points, we have $F_n \xrightarrow{\text{w}} F$ as desired.

It is easy to check that almost-sure convergence implies convergence in probability which then implies convergence in distribution, a.k.a., weak convergence. Using the 1st Borel-Cantelli lemma, convergence in probability in turn implies existence of an a.s.-convergent subsequence. The following gives the corresponding version of the converse for the weak convergence:

Theorem 3.4 (Skorohod representation) *Given Borel probability measures $\{\mu_n\}_{n\geqslant 1}$ and μ_∞ on $\mathbb{R}$, if $\mu_n \xrightarrow{\text{w}} \mu_\infty$, then there exists a probability space $(\Omega, \mathcal{F}, P)$ supporting random variables $\{X_n\}_{n\geqslant 1}$ and X_∞ such that*

- $\forall n \in \{1, \dots, \infty\}$: *$\mu_n$ is the distribution of X_n, and*
- $X_n \to X_\infty$ a.s.

Proof (sketch). Realizing the random variables as $X_n = F_n^{-1}(U)$ and $X_\infty = F_\infty^{-1}(U)$, for U a uniform random variable on $[0,1]$ and F_n^{-1} a suitable inverse of F_n, we then just verify the two claims. □

Note that, since the weak convergence does not require the random variable to be defined on the same probability space, the above is really best one can hope to get. An important consequence of the Skorohod embedding reads:

Corollary 3.5 *$X_n \xrightarrow{\text{w}} X$ implies*

$$\forall f \in C_{\text{b}}(\mathbb{R})\colon \ Ef(X_n) \underset{n\to\infty}{\longrightarrow} Ef(X) \tag{3.2}$$

Highlighting this fact is not a coincidence; indeed, (3.2) actually turns out to be equivalent to $X_n \xrightarrow{\text{w}} X$. (We will prove this later in the form of the *Portmanteau theorem*.)

With the notion of weak convergence settled, one is naturally interested in finding convergent sequences. Here is a tool in this regard, proved with the help of Cantor's diagonal argument and some extension arguments based on monotonicity:

Theorem 3.6 (Helly selection theorem) *Given a sequence $\{F_n\}_{n\geqslant 1}$ of CDFs, there is a strictly increasing sequence $\{n_k\}_{k\geqslant 1}$ of naturals and a non-decreasing right-continuous $F\colon \mathbb{R} \to [0,1]$ such that $F_{n_k} \xrightarrow{\text{w}} F$ (as $k \to \infty$).*

Notice that we do not claim that F is CDF because it may not be. (For instance, take $F_n(x) := 1_{[n,\infty)}(x)$ which tends pointwise to zero.) To prevent this from happening, we need another property:

Definition 3.7 (Tightness) *A sequence of CDFs $\{F_n\}_{n\geqslant 1}$ is* tight *if*

$$\forall \epsilon > 0\, \exists M > 0\colon \quad \sup_{n\geqslant 1} \big[F_n(-M) + 1 - F_n(M)\big] < \epsilon \tag{3.3}$$

Rewriting (3.3) using the probability measures $\{\mu_n\}_{n\geqslant 1}$ associated with the CDFs $\{F_n\}_{n\geqslant 1}$ leads to a perhaps more intuitive expression

$$\forall \epsilon > 0\, \exists M > 0\colon \quad \sup_{n\geqslant 1} \mu_n\big([-M, M]^{\text{c}}\big) < \infty \tag{3.4}$$

Tightness thus ensures that all but a small fraction of total mass of all the measures in the sequence stays confined to the same compact set which in particular means that no mass can escape to infinity in the limit. (This is the form that will define tightness in more general settings.) Returning to our previous line of thought, we now have:

Theorem 3.8 *Given CDFs $\{F_n\}_{n\geqslant 1}$, assume there is a non-decreasing and right-continuous function F such that $F_n \xrightarrow{\text{w}} F$. Then F is a CDF if and only if $\{F_n\}_{n\geqslant 1}$ is tight.*

3.2 Characteristic function.

As noted above, convergence of expectation of bounded continuous functions turns out to be equivalent to weak convergence. Checking all such functions may be tedious but, as it turns out, we only need to focus on a one parameter family; namely, the complex exponentials. This leads to:

Definition 3.9 *Given a real-valued random variable X, its* characteristic function *is*

$$\varphi_X(t) := E\mathrm{e}^{\mathrm{i}tX} := E\cos(tX) + \mathrm{i}E\sin(tX) \tag{3.5}$$

or, by the change of variables formula, $\varphi_X(t) := \int \mathrm{e}^{\mathrm{i}tx}\mu_X(\mathrm{d}x)$.

Notice that φ_X is bounded with $|\varphi_X(t)| \leqslant \varphi_X(0) = 1$, and is uniformly continuous. Moreover, we have the following fact

$$X, Y \text{ independent} \;\Leftrightarrow\; \varphi_{X+Y} = \varphi_X\varphi_Y \tag{3.6}$$

Here "$\Rightarrow$" is quite immediate but for "$\Leftarrow$" we need the conclusion of:

Lemma 3.10 (Inversion formula) *Let φ be the characteristic function associated with a finite Borel measure μ on $\mathbb{R}$. Then*

$$\forall a < b\colon \quad \lim_{T\to\infty}\frac{1}{\pi}\int_{-T}^{T}\varphi(t)\frac{\mathrm{e}^{-\mathrm{i}ta}-\mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t} = \frac{1}{2}\mu\big([a,b]\big) + \frac{1}{2}\mu\big((a,b)\big) \tag{3.7}$$

In particular, μ is uniquely determined by φ.

Proof. Using Fubini-Tonelli and a change of variables we write the integral on the left as

$$\begin{aligned}&\int\left(\int_a^b\left(\int_{-T}^{T}\mathrm{e}^{\mathrm{i}t(x-u)}\mathrm{d}t\right)\mathrm{d}u\right)\mu(\mathrm{d}x)\\ &\qquad = \int\left(\int_a^b\frac{\sin[T(x-u)]}{x-u}\,\mathrm{d}u\right)\mu(\mathrm{d}x) = \int\left(\int_{T(x-b)}^{T(x-a)}\frac{\sin u}{u}\,\mathrm{d}u\right)\mu(\mathrm{d}x)\end{aligned} \tag{3.8}$$

Now recall that $\int_0^T \frac{\sin u}{u}\mathrm{d}u \to \frac{\pi}{2}$ as $T\to\infty$ and note that this gives

$$\int_{T(x-b)}^{T(x-a)}\frac{\sin u}{u}\mathrm{d}u \underset{T\to\infty}{\longrightarrow} \frac{1}{2}1_{(a,b)}(x) + \frac{1}{2}1_{[a,b]}. \tag{3.9}$$

Since this also implies that the integral on the left is bounded uniformly in x, the statement follows using the Bounded Convergence Theorem. □

We now state the key tool that makes working with characteristic functions easy:

Theorem 3.11 (Lévy continuity theorem) *Suppose $\{X_n\}_{n\geqslant 1}$ are random variables such that*

$$\forall t\in\mathbb{R}\colon\ \varphi(t) := \lim_{n\to\infty}\varphi_{X_n}(t) \text{ exists} \tag{3.10}$$

If φ is continuous at $t = 0$, then there exists a random variable X such that

$$\forall t\in\mathbb{R}\colon\ \varphi(t) = E\mathrm{e}^{\mathrm{i}tX} \ \wedge\ X_n \xrightarrow{\mathrm{w}} X. \tag{3.11}$$

Proof. The key role of continuity is to imply tightness. This is based on the following analytic argument. Let Y be a random variable and φ_Y its characteristic function. Then

$$\frac{1}{t}\int_0^t \big[1 - \operatorname{Re}\varphi_Y(s)\big]\,\mathrm{d}u = \frac{1}{t}E\int_0^t \big(1-\cos(uY)\big)\,\mathrm{d}u = E\left(1 - \frac{\sin(tY)}{tY}\right) \tag{3.12}$$

where $\sin(u)/u := 1$ when $u = 0$. Observing that the function under expectation is always non-negative, we restrict the expectation to $|Y| \geqslant 1/t$ and get

$$P\big(|Y| > 1/t\big) \leqslant \frac{c}{t}\int_0^t \big[1 - \operatorname{Re}\varphi_Y(s)\big]\,\mathrm{d}u \tag{3.13}$$

where $c^{-1} := \inf_{u\geqslant 1}(1-\sin(u)/u)$. Returning to our problem, under the continuity of φ, for each $\epsilon > 0$ there exists $t > 0$ such that $|\varphi_{X_n}(t) - 1| < \epsilon$ for all $n \geqslant 1$. But then $P(|X_n| > 1/t) \leqslant c\epsilon$ for all $n \geqslant 1$, implying tightness.

With the laws of $\{X_n\}_{n\geqslant 1}$ tight, Theorem 3.6 ensures that any increasing sequence of naturals contains a subsequence along with the laws converge weakly. But the characteristic function of the limit law is then φ which, by Lemma 3.10, determines the limit law uniquely. It follows that all convergent subsequences converge to the same limit. Hence we get convergence. □

Note that if we recognize φ to be a characteristic function of a random variable, then it is automatically continuous and that random variable is the limiting X. We also note that a similar result holds for the convergence of the moment generating function. This goes by the name *Curtiss theorem*. We refer to homework for details.

As an application of the Lévy Continuity Theorem, we now give:

Theorem 3.12 (Central Limit Theorem) *Let $X_1, X_2, \ldots$ be i.i.d. with $X_1 \in L^2$. Denote $\mu := EX_1$ and $\sigma^2 := \operatorname{Var}(X_1)$ and let $S_n := X_1 + \cdots + X_n$. Then*

$$\frac{S_n - \mu n}{\sqrt{n}} \xrightarrow[n\to\infty]{\mathrm{w}} \mathcal{N}(0,\sigma^2) \tag{3.14}$$

Proof. Replacing X_i by $X_i - \mu$ allows us to assume that $\mu = 0$. By (3.6) we then have

$$E\big(\mathrm{e}^{\mathrm{i}tS_n/\sqrt{n}}\big) = \varphi_{X_1}(t/\sqrt{n})^n \tag{3.15}$$

Under the condition $X_1 \in L^2$ the characteristic function permits Taylor expansion

$$\varphi_{X_1}(t) = 1 + \mathrm{i}0 - \frac{t^2}{2}E(X_1^2) + o(t^2) \tag{3.16}$$

where $o(t^2)$ is a term vanishing upon division by t^2 and taking $t \to 0$. Using the above parametrization, this rewrites as

$$\varphi_{X_1}(t) = 1 - \frac{t^2}{2n}\sigma^2 + o(n^{-1}) \tag{3.17}$$

where $o(n^{-1})$ denote a complex valued sequence that decays faster than $1/n$. Using that for any complex-valued sequence $\{z_n\}_{n\geqslant 1}$ that converges to some $z \in \mathbb{C}$ we have

$(1+z_n/n)^n \to \mathrm{e}^z$ as $n\to\infty$, from (3.15) and (3.17) we conclude

$$\forall t\in\mathbb{R}\colon\ E\big(\mathrm{e}^{\mathrm{i}tS_n/\sqrt{n}}\big) \underset{n\to\infty}{\longrightarrow} \mathrm{e}^{-\frac{1}{2}t^2\sigma^2} \tag{3.18}$$

The right hand side is the characteristic function of $\mathcal{N}(0,\sigma^2)$ which, in particular, means that it is continuous at $t=0$. The claim then follows from Theorem 3.11. □

The history of the above result is very long. A very early version for zero-one valued random variables was posited by de Moivre in 1733 but went largely unnoticed until Laplace provided a first rigorous argument in 1812. A rather general version of the theorem was proved by Lyapounov in 1901 and then ultimately by Lindeberg in 1922. The above elegant proof is due to Lévy (1937). The word "central" was coined by Pólya in 1920 for the central role this result plays in probability and sciences.

3.3 Lindeberg's CLT.

The above CLT is quite elegant but we often need to relax the assumptions further. For instance, there is no need to use the same random variable for each n; their law may change with n. Similarly, there is no need that the random variables are equidistributed; we can allow their laws to change with their index as well.

Of course, implementing these changes requires a different formulation that we owe to Lyapounov (1901) and, ultimately, Lindeberg (1922).

Theorem 3.13 (Lindeberg's CLT) *Let $\{m(n)\}_{n\geqslant 1}$ be a sequence of positive naturals tending to infinity and assume that $\{X_{n,k}\colon n\geqslant 1,\ 1\leqslant k\leqslant m(n)\}$ is a family of random variables such that, for all $n\geqslant 1$,*

- *$X_{n,1},\dots,X_{n,m(n)}$ are independent*
- *$\forall k=1,\dots,m(n)\colon\ E(X_{n,k}^2)<\infty\ \wedge\ EX_{n,k}=0$*

Denote $S_n:=\sum_{k=1}^m(n)X_{n,k}$ and assume $\forall n\geqslant 1\colon \mathrm{Var}(S_n)>0$ and

$$\forall\epsilon>0\colon\ \frac{1}{\mathrm{Var}(S_n)}\sum_{k=1}^{m(n)} E\big(X_{n,k}^2 1_{\{|X_{n,k}|>\epsilon\sqrt{\mathrm{Var}(S_n)}\}}\big) \underset{n\to\infty}{\longrightarrow} 0 \tag{3.19}$$

Then

$$\frac{1}{\sqrt{\mathrm{Var}(S_n)}}\sum_{k=1}^{m(n)} X_{n,k} \underset{n\to\infty}{\overset{\mathrm{w}}{\longrightarrow}} \mathcal{N}(0,1) \tag{3.20}$$

The condition (3.19) is called the *Lindeberg condition*. Note that by taking $X_{n,k}:=X_k$, where $\{X_k\}_{k\geqslant 1}$ are i.i.d. random variable with zero mean and (finite) variance $\sigma^2>0$, and setting $m(n):=n$ we get $\mathrm{Var}(S_n)=n\sigma^2$. The Lindeberg condition (3.19) then follows thanks to

$$E(X_1^2 1_{\{|X_1|>\epsilon\sqrt{n}\}}) \underset{n\to\infty}{\longrightarrow} 0 \tag{3.21}$$

as implied by the Dominated Convergence Theorem. In particular, Theorem 3.13 subsumes Theorem 3.12 as a corollary.

The aforementioned Lyapounov result worked under a different condition than (3.19); namely, under

$$\exists p > 2\colon \ \frac{1}{[\mathrm{Var}(S_n)]^{p/2}} \sum_{k=1}^{m(n)} E(X_{n,k}^p) \underset{n\to\infty}{\longrightarrow} 0 \tag{3.22}$$

This readily implies (3.19) by invoking Chebyshev's inequality. To role of either of these conditions is to ensure that no single random variable contributions a non-trivial proportion of the overal variance. Assuming this, Feller derived the following converse to Lindeberg's result:

Theorem 3.14 (Feller) *Assuming the setting of Theorem 3.13, suppose that*

$$\forall \epsilon > 0\colon \ \max_{1\leqslant k\leqslant m(n)} P\Big(|X_{n,k}| > \epsilon\sqrt{\mathrm{Var}(S_n)}\Big) \underset{n\to\infty}{\longrightarrow} 0 \tag{3.23}$$

Then the weak convergence (3.20) *implies Lindeberg's condition* (3.19).

Lindeberg's condition is thus necessary and sufficient for triangular arrays of random variables satisfying the condition (3.23). For this reason the Lindebeg's result is sometimes referred to as Lindeberg-Feller CLT.

Leaving Feller's theorem to homework, we will now move to the proof of Lindeberg's CLT. Abbreviate

$$\sigma_n := \sqrt{\mathrm{Var}(S_n)} \tag{3.24}$$

and note that Lindeberg's condition implies existence of a positive sequence $\{\epsilon_n\}_{n\geqslant 1}$ tending to zero and satisfying

$$\frac{1}{\epsilon_n^2\sigma_n^2} \sum_{k=1}^{n} E\big(X_{n,k}^2 1_{|X_{n,k}|>\epsilon_n\sigma_n}\big) \underset{n\to\infty}{\longrightarrow} 0 \tag{3.25}$$

We now use these to introduce the truncated variables

$$\overline{X}_{n,k} := X_{n,k} 1_{\{|X_{n,k}|\leqslant\epsilon_n\sigma_n\}} \tag{3.26}$$

and denote

$$\overline{S}_n := \sum_{k=1}^{m(n)} \overline{X}_{n,k} \tag{3.27}$$

Our first item of business is to show that the truncation does not have an effect, including the first and second moments:

Lemma 3.15 *The following hold*

- $P(|S_n - \overline{S}_n| \neq 0) \to 0$
- $\sigma_n^{-1} E\overline{S}_n \to 0$
- $\sigma_n^{-2}\mathrm{Var}(\overline{S}_n) \to 1$

in the limit as $n \to \infty$.

Proof. We have

$$\begin{aligned} P(|S_n - \overline{S}_n| \neq 0) &\leqslant P(\exists k \leqslant n \colon |X_{n,k}| > \epsilon_n \sigma_n) \\ &\leqslant \sum_{k=1}^{n} P(|X_{n,k}| > \epsilon_n \sigma_n) \leqslant \frac{1}{\epsilon_n^2 \sigma_n^2} \sum_{k=1}^{n} E(X_{n,k}^2 1_{\{|X_{n,k}| > \epsilon_n \sigma_n\}}). \end{aligned} \tag{3.28}$$

This tends to zero by (3.25). Similarly, using that $ES_n = 0$,

$$\begin{aligned} \sigma_n^{-1} |E\overline{S}_n| &= \sigma_n^{-1} |E(\overline{S}_n - S_n)| \\ &\leqslant \frac{1}{\sigma_n} \sum_{k=1}^{n} E(|X_{n,k}| 1_{\{|X_{n,k}| > \epsilon_n \sigma_n\}}) \leqslant \frac{1}{\epsilon_n \sigma_n^2} \sum_{k=1}^{n} E(X_{n,k}^2 1_{\{|X_{n,k}| > \epsilon_n \sigma_n\}}) \end{aligned} \tag{3.29}$$

Now apply (3.25) to see that this tends to zero as well.

Finally, denoting $Y_{n,k} := X_{n,k} 1_{\{|X_{n,k}| > \epsilon_n \sigma_n\}}$, independence of $\{\overline{X}_{n,k}\}_{1 \leqslant k \leqslant n}$ for each $n \geqslant 1$ along with $E\overline{X}_{n,k} = -EY_{n,k}$ give

$$\begin{aligned} \mathrm{Var}(\overline{S}_n) - \sigma_n^2 &= \sum_{k=1}^{n} \big[E(\overline{X}_{n,k}^2) - (E\overline{X}_{n,k})^2 - E(X_{n,k}^2) \big] \\ &= - \sum_{k=1}^{n} \big[E(Y_{n,k}^2) + (EY_{n,k})^2 \big] \end{aligned} \tag{3.30}$$

Cauchy-Schwarz gives $(E(Y_{n,k})^2 \leqslant E(Y_{n,k}^2)$ and so

$$\big| \sigma_n^{-2} \mathrm{Var}(\overline{S}_n) - 1 \big| \leqslant \frac{2}{\sigma_n^2} \sum_{k=1}^{n} E(X_{n,k}^2 1_{\{|X_{n,k}| > \epsilon_n \sigma_n\}}) \tag{3.31}$$

By (3.25) this vanishes as $n \to \infty$. □

The upshot of the lemma is that it suffices to prove the result for the truncated variables. As these are no longer centered, we introduce Abbreviate

$$Z_{n,k} := \sigma_n^{-1} (\overline{X}_{n,k} - E\overline{X}_{n,k}) \tag{3.32}$$

and note that $EZ_{n,k} = 0$ and

$$\sum_{k=1}^{n} E(Z_{n,k}^2) \underset{n \to \infty}{\longrightarrow} 1 \tag{3.33}$$

As is readily checked, it suffices to prove weak convergence of

$$\widehat{S}_n := \sum_{k=1}^{m(n)} Z_{n,k} \tag{3.34}$$

to $\mathcal{N}(0,1)$. For this we note that, denoting

$$\varphi_{n,k}(t) := E\mathrm{e}^{\mathrm{i}tZ_{n,k}}, \tag{3.35}$$

we have

$$E\mathrm{e}^{\mathrm{i}t\widehat{S}_n} = \prod_{k=1}^{m(n)} \varphi_{n,k}(t) \tag{3.36}$$

Proceeding as in the proof of Theorem 3.12, we now have to expand the characteristic functions on the right to its second order Taylor polynomial. Due to non-homogeneity, this has to be done under the summation over k:

Lemma 3.16 (Taylor approximation)

$$\forall t \in \mathbb{R}\colon \quad \sum_{k=1}^{m(n)} \left| \varphi_{n,k}(t) - \Big(1 - \frac{t^2}{2} E(Z_{n,k}^2)\Big) \right| \underset{n\to\infty}{\longrightarrow} 0 \tag{3.37}$$

Proof. By Taylor's theorem

$$\varphi_{n,k}(t) = 1 + \mathrm{i}0 - t^2 \int_0^1 u E\big(Z_{n,k}^2 \, \mathrm{e}^{\mathrm{i}t(1-u)Z_{n,k}}\big) \mathrm{d}u \tag{3.38}$$

and hence

$$\varphi_{n,k}(t) - \Big(1 - \frac{t^2}{2} E(Z_{n,k}^2)\Big) = t^2 \int_0^1 u E\big(Z_{n,k}^2 (1 - \mathrm{e}^{\mathrm{i}t(1-u)Z_{n,k}})\big) \mathrm{d}u \tag{3.39}$$

Truncation ensures $|Z_{n,k}| \leqslant 2\epsilon_n$ and so

$$|1 - \mathrm{e}^{\mathrm{i}t(1-u)Z_{n,k}}| = 2\big|\sin\big(t(1-u)Z_{n,k}\big)\big| \leqslant \max_{0\leqslant x\leqslant 2|t|\epsilon_n} 2|\sin(x)| \tag{3.40}$$

which only depends on n. Hence

$$\sum_{k=1}^{m(n)} \left| \varphi_{n,k}(t) - \Big(1 - \frac{t^2}{2} E(Z_{n,k}^2)\Big) \right| \leqslant \frac{1}{2} t^2 \Big(\max_{0\leqslant x\leqslant 2|t|\epsilon_n} 2|\sin(x)| \Big) \sum_{k=1}^{m(n)} E(Z_{n,k}^2). \tag{3.41}$$

The maximum tends to zero as $n \to \infty$. □

In order to apply the above expansion under the product in (3.36), we also need:

Lemma 3.17 *Let $n \geqslant 1$ and $z_1, \ldots, z_n, w_1, \ldots, w_n \in \{z \in \mathbb{C}\colon |z| \leqslant 1\}$. Then*

$$\left| \prod_{i=1}^n z_i - \prod_{i=1}^n w_i \right| \leqslant \sum_{i=1}^n |z_i - w_i| \tag{3.42}$$

Proof. The claim immediate for $n = 1$. For $n > 1$ we have

$$\begin{aligned} \left| \prod_{i=1}^n z_i - \prod_{i=1}^n w_i \right| &\leqslant |z_n| \left| \prod_{i=1}^{n-1} z_i - \prod_{i=1}^{n-1} w_i \right| + |z_n - w_n| \left| \prod_{i=1}^{n-1} w_i \right| \\ &\leqslant |z_n - w_n| + \left| \prod_{i=1}^{n-1} z_i - \prod_{i=1}^{n-1} w_i \right| \end{aligned} \tag{3.43}$$

The bound now follows by induction. □

We are now ready to give:

Proof of Theorem 3.13. As $E(Z_{n,k}^2) \leqslant 4\epsilon_n^2$, we have $|1 - \frac{t^2}{2}E(Z_{n,k}^2)| < 1$ for $n \gg 1$. By Lemmas above,

$$\left| E\mathrm{e}^{\mathrm{i}t\widehat{S}_n} - \prod_{k=1}^{m(n)} \Bigl(1 - \frac{t^2}{2}E(Z_{n,k}^2)\Bigr) \right| \underset{n\to\infty}{\longrightarrow} 0 \tag{3.44}$$

Using that, for $|z| \leqslant 1/2$,

$$\bigl|\log(1+z) - z\bigr| \leqslant \sum_{k\geqslant 2} \frac{|z|^k}{k} \leqslant \sum_{k\geqslant 2} |z|^k \leqslant \frac{|z|^2}{1-|z|} \leqslant 2|z|^2 \tag{3.45}$$

we get that

$$\begin{aligned} &\left| \log \prod_{k=1}^{m(n)} \Bigl(1 - \frac{t^2}{2}E(Z_{n,k}^2)\Bigr) + \frac{t^2}{2} \sum_{k=1}^{m(n)} E(Z_{n,k}^2) \right| \\ &\qquad\qquad \leqslant 2 \sum_{k=1}^{m(n)} \bigl[\tfrac{t^2}{2}E(Z_{n,k}^2)\bigr]^2 \leqslant 2t^4\epsilon_n^2 \sum_{k=1}^{m(n)} E(Z_{n,k}^2) \end{aligned} \tag{3.46}$$

once n is so large that $2t^2\epsilon_n^2 \leqslant 1$. As $\sum_{k=1}^{m(n)} E(Z_{n,k}^2) \to 1$, we conclude

$$\forall t \in \mathbb{R}\colon\ E\mathrm{e}^{\mathrm{i}t\widehat{S}_n} \to \mathrm{e}^{-t^2/2} \tag{3.47}$$

The Lévy continuity theorem (Theorem 3.12) gives $\widehat{S}_n \xrightarrow{\mathrm{w}} \mathcal{N}(0,1)$. □

We note that a different argument bypassing Lemmas 3.16–3.17 above (and also likely closer to the original Lindeberg's argument) will be given in the next section.

3.4 Non-standard CLT.

As an example of the power of Lindeberg's theorem, we give a short proof of CLT under the conditions where the regular CLT does not apply.

Theorem 3.18 *Suppose $X_1, X_2, \ldots$ are i.i.d. with*

$$X_1 \overset{\mathrm{law}}{=} -X_1 \quad \wedge \quad \forall x > 1\colon\ P(|X_1| > x) = x^{-2} \tag{3.48}$$

Set $S_n := X_1 + \cdots + X_n$. Then

$$\frac{S_n}{\sqrt{n\log n}} \underset{n\to\infty}{\xrightarrow{\mathrm{w}}} \mathcal{N}(0,1) \tag{3.49}$$

Proof. We will have to truncate the sequence to bring it to the form to which Theorem 3.13 can be applied. For truncation we will use a sequence $\{a_n\}_{n\geqslant 1}$ of positive reals with specific growth to be determined. Denote $X_{n,k} := X_k 1_{\{|X_k|\leqslant a_n\}}$. The assumed symmetry ensures $EX_{n,k} = 0$ and, assuming also $a_n \geqslant 1$, we have

$$\begin{aligned} \mathrm{Var}(X_{n,k}) = E(X_{n,k}^2) &= \int_0^\infty 2tP\bigl(|X_{n,k}| > t\bigr)\mathrm{d}t \\ &= 1 + \int_1^{a_n} 2t\frac{1}{t^2}\mathrm{d}t = 1 + 2\log(a_n) \end{aligned} \tag{3.50}$$

In order for the truncation to have no effect, it suffices that

$$P\big(\exists k \leqslant n\colon\ |X_k| > a_n\big) \leqslant nP\big(|X_1| > a_n\big) = \frac{n}{a_n^2} \underset{n\to\infty}{\longrightarrow} 0 \tag{3.51}$$

In particular, for this a_n has to grow faster than $\sqrt{n}$. On the other hand, to get Lindeberg's condition, it suffices that $\overline{S}_n := X_{n,1} + \cdots + X_{n,n}$ obeys

$$\frac{a_n}{\sqrt{\mathrm{Var}(\overline{S}_n)}} \underset{n\to\infty}{\longrightarrow} 0. \tag{3.52}$$

Indeed, then $|X_{n,k}| \leqslant a_n < \epsilon\sqrt{\mathrm{Var}(\overline{S}_n)}$ for n sufficiently large which means that Lindeberg's condition holds trivially.

To satisfy the two requirements, set

$$a_n := \sqrt{n \log\log n} \tag{3.53}$$

and observe that this grows faster than $\sqrt{n}$ but slower than $\sqrt{\mathrm{Var}(\overline{S}_n)}$ which by (3.50) is proportional to $\sqrt{n \log n}$. Lindeberg's CLT then gives the result. □

One can of course prove Theorem 3.18 by invoking truncation directly inside the proof of Theorem 3.12 and treating carefully the errors. The main point of Lindeberg's formulation is that it is general enough to adapt to other situations as well; see Durrett's book for its application to counting random permutations or a proof of one direction of Kolmogorov's 3-series theorem.

Further reading: Durrett, Sections 3.1-3.4

4. Quantitative CLTs: Lindeberg and Berry-Esseen

We now move to arguments that permit us to control the convergence to normality in terms of explicit error bounds. Besides practical implications, this will allow us to give new proofs of the qualitative CLTs discussed earlier.

4.1 Lindeberg's method.

The first quantitative bound we present for closeness to normality is due to Lindeberg. We state it as the following theorem:

Theorem 4.1 (Lindeberg's method) *Let $n \geqslant 1$ and let $X_1, \dots, X_n$ be real-valued random variables such that*

- *$X_1, \dots, X_n$ are independent*
- *$\forall k = 1, \dots, n\colon\ X_k \in L^3$*

Let Z be a random variable with the law

$$Z = \mathcal{N}\Big(\sum_{k=1}^{n} EX_k,\ \sum_{k=1}^{n} \mathrm{Var}(X_k) \Big). \tag{4.1}$$

Then for all $f \in C_{\mathrm{b}}^3(\mathbb{R})$,

$$\Big| Ef\Big(\sum_{k=1}^{n} X_k\Big) - Ef(Z) \Big| \leqslant 5 \| f''' \|_\infty \sum_{k=1}^{n} E\big(|X_k|^3\big) \tag{4.2}$$

The phrase "Lindeberg method" actually refers to the main idea of the proof which is to swap $X_1, \dots, X_n$, one by one, for independent normal random variables $Z_1, \dots, Z_n$ that agree with $X_1, \dots, X_n$ in the first two moments; i.e.,

$$\forall k = 1, \dots, n\colon\ Z_k = \mathcal{N}(EX_k, \mathrm{Var}(X_k)) \tag{4.3}$$

This does give the desired bound because $Z \overset{\text{law}}{=} Z_1 + \cdots + Z_n$. The error caused by the swap is quantified using Taylor's theorem as shown in:

Lemma 4.2 *Let $X, Y, Z \in L^3$ be independent with Z having the law*

$$Z = \mathcal{N}\big(EX, \mathrm{Var}(X)\big) \tag{4.4}$$

Then for all $f \in C_{\mathrm{b}}^3(\mathbb{R})$,

$$\big| Ef(Y + X) - Ef(Y + Z) \big| \leqslant 5 \| f''' \|_\infty E\big(|X|^3\big) \tag{4.5}$$

Proof. Taylor's theorem gives

$$f(Y + X) = f(Y) + f'(Y)X + \frac{1}{2} f''(Y) X^2 + \frac{1}{2} \int_0^1 f'''\big(Y + sX\big) X^3 s^2 \mathrm{d}s \tag{4.6}$$

with $f(Y + Z)$ written similarly. Subtracting the two expressions and taking expectation with the help of $EX = EZ$, $\mathrm{Var}(X) = \mathrm{Var}(Z)$ and the assumed independence of X, Y, Z shows

$$\big| Ef(Y + X) - Ef(Y + Z) \big| \leqslant \frac{1}{6} \| f''' \|_\infty \Big(E\big(|X|^3\big) + E\big(|Z|^3\big) \Big). \tag{4.7}$$

In order to express $E(|Z|^3)$, note that $(a+b)^3 \leqslant 8|a|^3 + 8|b|^3$ and apply the scaling properties of normal law to get

$$E\big(|Z|^3\big) \leqslant 8(E|Z|)^3 + 8\mathrm{Var}(Z)^{3/2} E\big(|\mathcal{N}(0,1)|^3\big) \tag{4.8}$$

Using again that $EX = EZ$ and $\mathrm{Var}(X) = \mathrm{Var}(Z)$ while comparing moments via Jensen's inequality shows

$$\begin{aligned} (E|Z|)^3 = (E|X|)^3 &\leqslant E\big(|X|^3\big) \\ \mathrm{Var}(Z)^{3/2} = \mathrm{Var}(X)^{3/2} &\leqslant \big(E(X^2)\big)^{3/2} \leqslant E\big(|X|^3\big) \end{aligned} \tag{4.9}$$

A calculation shows $E(|\mathcal{N}(0,1)|^3) = \frac{4}{\sqrt{2\pi}} \leqslant 2$ and from (4.8) we get $E(|Z|^3) \leqslant 24E(|X|^3)$. Plugging this into (4.7) then yields the claim. (We are not trying to optimize numerical prefactors.) □

We are now ready to give:

Proof of Theorem 4.1. Let $Z_1, \dots, Z_n$ be independent of each other and $X_1, \dots, X_n$ with the law in (4.3). For each $k = 1, \dots, n$, let

$$Y_k := X_1 + \cdots + X_{k-1} + Z_{k+1} + \cdots + Z_n \tag{4.10}$$

Then Lemma 4.2 gives

$$\begin{aligned} \left| Ef\Big(\sum_{k=1}^n X_k\Big) - Ef\Big(\sum_{k=1}^n Z_k\Big) \right| &\leqslant \sum_{k=1}^n \big| Ef(Y_k + X_k) - Ef(Y_k + Z_k) \big| \\ &\leqslant 5\| f''' \|_\infty \sum_{k=1}^n E\big(|X_k|^3\big) \end{aligned} \tag{4.11}$$

Since $Z \overset{\text{law}}{=} Z_1 + \cdots + Z_n$, this is the desired claim. □

As a simple application, we get our first quantitative CLT:

Corollary 4.3 (Quantitative CLT) *Suppose $X_1, X_2, \dots$ are i.i.d. random variables with $X_1 \in L^3$ and $EX_1 = 0$. Let $Z = \mathcal{N}(0, \mathrm{Var}(X_1))$. Then for all $f \in C_b^3(\mathbb{R})$,*

$$\left| Ef\Big(\frac{1}{\sqrt{n}} \sum_{k=1}^n X_k\Big) - Ef(Z) \right| \leqslant \frac{5}{\sqrt{n}} \| f''' \|_\infty E\big(|X_1|^3\big) \tag{4.12}$$

Proof. Apply Lindeberg's bound (4.2) to X_k replaced by $\frac{X_k}{\sqrt{n}}$ and observe that

$$\sum_{k=1}^n E\left(\Big|\frac{X_k}{\sqrt{n}}\Big|^3\right) = \frac{1}{\sqrt{n}} E\big(|X_1|^3\big) \tag{4.13}$$

thanks to $X_1, \dots, X_n$ being equidistributed. □

As another application, we give a somewhat different proof of Lindeberg's CLT:

Proof of Theorem 3.13. Let $X_{n,1}, \dots, X_{n,m(n)}$ be independent for each $n \geqslant 1$ with $X_{n,k} \in L^2$ and $EX_{n,k} = 0$ for each $k = 1, \dots, m(n)$. Without loss of generality, assume that the

random variables are normalized so that

$$\sum_{k=1}^{m(n)} EX_{n,k}^2 = 1 \tag{4.14}$$

Lindeberg's condition then implies existence of $\epsilon_n > 0$ with $\epsilon_n \downarrow 0$ such that

$$\overline{X}_{n,k} := X_{n,k} 1_{\{|X_{n,k}| \leqslant \epsilon_n\}} \tag{4.15}$$

obeys the conclusions of Lemma 3.15. Namely, we get

$$P\big(\exists k \leqslant m(n) \colon\ \overline{X}_{n,k} \neq X_{n,k}\big) \ \underset{n\to\infty}{\longrightarrow}\ 0 \tag{4.16}$$

$$\sum_{k=1}^{m(n)} E\overline{X}_{n,k} \ \underset{n\to\infty}{\longrightarrow}\ 0 \tag{4.17}$$

and

$$\sum_{k=1}^{m(n)} \mathrm{Var}(\overline{X}_{n,k}) \ \underset{n\to\infty}{\longrightarrow}\ 1. \tag{4.18}$$

By (4.16) we then have

$$\forall f \in C_{\mathrm{b}}(\mathbb{R}) \colon \quad Ef\Big(\sum_{k=1}^{m(n)} X_{n,k}\Big) - Ef\Big(\sum_{k=1}^{m(n)} \overline{X}_{n,k}\Big) \ \underset{n\to\infty}{\longrightarrow}\ 0 \tag{4.19}$$

which means that we may focus on $\sum_{k=1}^{m(n)} \overline{X}_{n,k}$ instead of $\sum_{k=1}^{m(n)} X_{n,k}$.

Let Z_n be a random variable with law

$$Z_n = \mathcal{N}\Big(\sum_{k=1}^{m(n)} E\overline{X}_{n,k}, \sum_{k=1}^{m(n)} \mathrm{Var}(\overline{X}_{n,k})\Big) \tag{4.20}$$

and observe that Theorem 4.1 then gives

$$\Big|Ef\Big(\sum_{k=1}^{m(n)} \overline{X}_{n,k}\Big) - Ef(Z_n)\Big| \leqslant 5\|f'''\|_\infty \sum_{k=1}^{m(n)} E\big(|\overline{X}_{n,k}|^3\big) \tag{4.21}$$

for all $f \in C_{\mathrm{b}}^3(\mathbb{R})$. But $|\overline{X}_{n,k}| \leqslant \epsilon_n$ and so

$$\sum_{k=1}^{m(n)} E\big(|\overline{X}_{n,k}|^3\big) \leqslant \epsilon_n \sum_{k=1}^{m(n)} E\big(|\overline{X}_{n,k}|^2\big) \tag{4.22}$$

which tends to zero because $\epsilon_n \downarrow 0$ and sum on the right converges by (4.18). Since (4.17–4.18) implies $Ef(Z_n) \to Ef(Z)$ for $Z = \mathcal{N}(0,1)$, from (4.19), (4.21) and (4.22) we thus get

$$\forall f \in C_{\mathrm{b}}^3(\mathbb{R}) \colon \quad Ef\Big(\sum_{k=1}^{m(n)} X_{n,k}\Big) \ \underset{n\to\infty}{\longrightarrow}\ Ef(Z). \tag{4.23}$$

Applying this to $f(x) = \cos(tx)$ and $f(x) = \sin(tx)$, we conclude that the characteristic function of $\sum_{k=1}^{m(n)} X_{n,k}$ tends to that of Z. The Lévy continuity theorem (Theorem 3.11) then gives the result. □

4.2 Berry-Esseen theorem.

Our second quantitative bound for distance to normality comes in the following theorem that dates back more than 80 years:

Theorem 4.4 (Berry 1941, Esseen 1942) *Let $X_1, \dots, X_n$ be i.i.d. with $X_1 \in L^3$ and $EX_1 = 0$. Then for $Z = \mathcal{N}(0, \mathrm{Var}(X_1))$ and a constant $c \approx 0.4785$,*

$$\left| P\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \leqslant a\Big) - P(Z \leqslant a) \right| \leqslant c \frac{E(|X_1|^3)}{\mathrm{Var}(X_1)^{3/2}} \frac{1}{\sqrt{n}} \tag{4.24}$$

holds for all $a \in \mathbb{R}$.

We remark that the statement gives an estimate on the Kolmogorov distance of two random variables which we define as

$$d_{\mathrm{K}}(X, Y) := \sup_{a \in \mathbb{R}} \big| P(X \leqslant a) - P(Y \leqslant a) \big| \tag{4.25}$$

The fact that this is invariant under scaling both X and Y by the same constant explains why the bound on the right of (4.24) is also invariant under such rescaling.

We also note that the error-rate $n^{-1/2}$ is generally best possible. Indeed, if $X_1, \dots, X_n$ are Bernoulli, the distribution of their normalized sum is piecewise constant with jumps (near the median) of order $1/\sqrt{n}$. (This is checked from the calculation leading to de Moivre-Laplace CLT or by way of the local Central Limit Theorem.) Since the CDF of Z is continuous, the left-hand side is at least order $1/\sqrt{n}$.

The proof is based on manipulations with characteristic functions and convolution with Polya's density; we refer to Durrett's textbook for a full acount of all details. Incidentally, the proof there gives $c := 3$; the above is a result of a long sequence of gradual improvements and is due to Tyurin (2010).

Further reading: Durrett, Section 3.4.4

5. Quantitative CLTs: Stein's method

We proceed to discuss another method for estimating convergence to normality that is due to Charles Stein (with the first article dating back to 1972).

5.1 Characterizing normality.

The presentations of Stein's method typically open up by the following characterization of the standard normal law:

Lemma 5.1 (Stein's lemma) *Let Y be an real-valued random variable. Then the following two properties are equivalent:*

- $Y = \mathcal{N}(0,1)$
- *for all $f \in \mathrm{AC}(\mathbb{R})$ with f' and xf bounded,*

$$Ef'(Y) = E\big(Yf(Y)\big) \tag{5.1}$$

Here $\mathrm{AC}(\mathbb{R})$ is the space of real-valued absolutely-continuous functions on $\mathbb{R}$ and xf is the function defined as $xf(x) := xf(x)$.

One direction of the proof is immediate:

Proof of $\Rightarrow$. Suppose first that f is continuously differentiable with f' and xf bounded. Integration by parts then gives

$$\int yf(y)\mathrm{e}^{-\frac{y^2}{2}}\,\mathrm{d}y = -\int f(y)\frac{\mathrm{d}}{\mathrm{d}y}\mathrm{e}^{-\frac{y^2}{2}}\,\mathrm{d}y = \int f'(y)\mathrm{e}^{-\frac{y^2}{2}}\,\mathrm{d}y. \tag{5.2}$$

The case of general $f \in \mathrm{AC}(\mathbb{R})$ is handled by replacing f by $f_\epsilon(x) := Ef(x + Z_\epsilon)$ where $Z_\epsilon = \mathcal{N}(0,\epsilon)$ and noting that $f_\epsilon \to f$ and $f'_\epsilon \to f'_\epsilon$ pointwise a.e. as $\epsilon \downarrow 0$. The convergence of integrals then follows using the Bounded Convergence Theorem. □

The proof of $\Leftarrow$ requires quite some effort. The main idea comes in:

Lemma 5.2 (Stein's ODE) *Let $Z = \mathcal{N}(0,1)$ and assume $h\colon \mathbb{R} \to \mathbb{R}$ to be measurable with $E|h(Z)| < \infty$. Then for all $x \in \mathbb{R}$,*

$$\begin{aligned} f_h(x) &:= \mathrm{e}^{\frac{x^2}{2}} \int_x^\infty \mathrm{e}^{-\frac{y^2}{2}} \big[Eh(Z) - h(y)\big] \frac{\mathrm{d}y}{\sqrt{2\pi}} \\ &= \mathrm{e}^{\frac{x^2}{2}} \int_{-\infty}^x \mathrm{e}^{-\frac{y^2}{2}} \big[h(y) - Eh(Z)\big] \frac{\mathrm{d}y}{\sqrt{2\pi}} \end{aligned} \tag{5.3}$$

Moreover, $f_h \in AC(\mathbb{R})$ and

$$f'_h(x) - xf_h(x) = h(x) - Eh(Z) \tag{5.4}$$

holds for Lebesgue a.e. $x \in \mathbb{R}$

Proof. Both integrals converge absolutely thanks to our assumption $E|h(Z)| < \infty$. Subtracting the second integral from the first can be written as

$$\int_{-\infty}^\infty \mathrm{e}^{-\frac{y^2}{2}} \big[Eh(Z) - h(y)\big] \frac{\mathrm{d}y}{\sqrt{2\pi}} = Eh(Z) - Eh(Z) = 0. \tag{5.5}$$

Differentiation shows that f_h obeys the ODE (5.4). □

It is easy to check that f_h is the unique solution to the ODE (5.4) subject to $f_h(x) \to 0$ as $x \to -\infty$. The reason why we offer two expressions for f_h is because the first will be used for $x \geqslant 0$ and the second for $x \leqslant 0$.

The formula (5.3) defines a linear map $h \mapsto f_h$ — writing $f_h = \mathcal{A}h$ defines the so-called Stein operator $\mathcal{A}$. Stein's method typically relies on specific functional properties of this map. For Lemma 5.1, the following suffices:

Lemma 5.3 *Let $h \in L^\infty(\mathbb{R})$. Then $f_h \in AC(\mathbb{R})$ and*

$$\|f_h\|_\infty \leqslant 3\|h\|_\infty \ \wedge \ \|xf_h\|_\infty \leqslant 2\|h\|_\infty \ \wedge \ \|f_h'\|_\infty \leqslant 4\|h\|_\infty \tag{5.6}$$

Proof. Let $x > 0$. A change of variables $y \to x + y$ gives

$$f_h(x) = \int_0^\infty \mathrm{e}^{-\frac{y^2}{2} - yx} \big[Eh(Z) - h(y)\big] \frac{\mathrm{d}y}{\sqrt{2\pi}} \tag{5.7}$$

Using $\sqrt{2\pi} \geqslant 1$ to drop $\sqrt{2\pi}$ from the denominator, dominating $|Eh(Z) - h(y)| \leqslant 2\|h\|_\infty$ and bounding the resulting integral by retaining only one term in the exponent gives

$$\big|f_h(x)\big| \leqslant 2\|h\|_\infty \min\Big\{\sqrt{\frac{\pi}{2}}, \frac{1}{x}\Big\} \tag{5.8}$$

A completely analogous (with $1/|x|$) bound is derived for $x < 0$ using the second line in (5.3) instead. Using these bounds and the fact that $2\pi \leqslant 9$ we get $|f_h(x)| \leqslant 3\|h\|_\infty$ and $|xf_h(x)| \leqslant 2\|h\|_\infty$. With the help of Stein's ODE we infer also $|f_h'(x)| \leqslant 4\|h\|_\infty$. □

We remark that the constants in (5.6) are certainly not optimal; the point of these bounds is to show that the map $h \mapsto f_h$ is continuous as a map $L^\infty \to L^\infty$ for f_h itself, its derivative and also its multiplier by x. We are now ready for:

Proof of $\Leftarrow$ in Lemma 5.1. Assume that Y is a random variable such that (5.1) holds for all $f \in \mathrm{AC}(\mathbb{R})$ with f' and xf bounded. Given $a \in \mathbb{R}$ let $h := 1_{(-\infty,a]}$. Lemma 5.3 shows that f_h' and xf_h are bounded and so, by Fubini applied to Stein's ODE,

$$\begin{aligned} P(Y \leqslant a) - P(Z \leqslant a) &= Eh(Y) - Eh(Z) \\ &= E\big[f_h'(Y) - Yf_h(Y)\big] = 0 \end{aligned} \tag{5.9}$$

Since this holds for all $a \in \mathbb{R}$, we get $Y \overset{\text{law}}{=} Z$ as desired. □

5.2 Link to distance on measures.

The characterization of normality in Lemma 5.1 is a mere curiosity whose usefulness is in its own right is rather limited. What should catch our eye is the argument (5.9) before setting the whole quantity to zero. Indeed, the equality links the difference of the CDFs of Y and Z to the expectation of $f_h'(Y) - Yf_h(Y)$, which is only a function of Y. The way we want to think about this is a way to derive, for any non-empty

$$\mathcal{H} \subseteq \big\{h\colon \text{measurable} \ \wedge \ E|h(Z)| < \infty\big\} \tag{5.10}$$

the identity

$$\sup_{h\in\mathcal{H}} \big|Eh(Y) - Eh(Z)\big| = \sup_{h\in\mathcal{H}} \big|E[\, f_h'(Y) - Y f_h(Y)]\big| \tag{5.11}$$

assuming that the right-hand side is meaningful for each $h \in \mathcal{H}$.

A key point of (5.11) that the quantity on the left is a pseudometric (on the space of probability measures) and often a metric. Here are some examples: For the case

$$\mathcal{H} := \big\{1_A \colon A \in \mathcal{B}(\mathbb{R})\big\} \tag{5.12}$$

the supremum on the left of (5.11) defines the so called *total-variational distance*,

$$d_{\mathrm{TV}}(X,Y) := \sup_{A\in\mathcal{B}(\mathbb{R})} \big|P(X\in A) - P(Y\in A)\big| \tag{5.13}$$

(For measures μ and ν, this is sometimes denoted as $\|\mu-\nu\|_{\mathrm{TV}}$.) This distance is often used in discrete probability (where it reduces to half of the ℓ^1-norm of the difference of the probability mass functions), but its use for continuum-distributed random variables is less useful due to poor behavior under perturbations; e.g., $d_{\mathrm{TV}}(X, \epsilon + X) = 1$ whenever $\epsilon \neq 0$.

Reducing (5.12) to indicators of half-infinite intervals,

$$\mathcal{H} := \{1_{(-\infty,a]} \colon a \in \mathbb{R}\} \tag{5.14}$$

the supremum on the left of (5.11) defines the so called *Kolmogorov distance*,

$$d_{\mathrm{K}}(X,Y) := \sup_{a\in\mathbb{R}} \big|P(X\leqslant a) - P(Y\leqslant a)\big| \tag{5.15}$$

We have already encountered the Kolmogorov distance in the statement of the Glivenko-Cantelli SLLN (Theorem 2.11) and (implicitly) the Berry-Esseen theorem (Theorem 4.4). Convergence in the Kolmogorov distance implies weak convergence but the converse fails unless the limit random variable is continuously distributed.

Our last example concerns the space of Lipschitz functions

$$\mathcal{H} := \big\{h \colon \mathrm{AC} \wedge \|h'\|_\infty \leqslant 1\big\} \tag{5.16}$$

with unit Lipschitz norm which gives rise to the notion of *Wasserstein distance*. As for the two distances above, this is really distance on the set of probability measures which we define in full generality as:

Definition 5.4 *Given a metric space $(\mathscr{X}, \rho)$ and $f \colon \mathscr{X} \to \mathbb{R}$, let*

$$\|\, h\|_{\mathrm{Lip}} := \sup_{x\neq y} \frac{|h(y)-h(x)|}{\rho(x,y)} \tag{5.17}$$

Let $\mathcal{M}_1(\mathscr{X})$:= set of Borel probability measures on $\mathscr{X}$. Then

$$d_{\mathrm{W}}(\mu,\nu) := \sup\left\{\left|\int h\mathrm{d}\mu - \int h\mathrm{d}\nu\right| \colon \|h\|_{\mathrm{Lip}} \leqslant 1\right\} \tag{5.18}$$

is the Wasserstein distance of $\mu, \nu \in \mathcal{M}_1(\mathscr{X})$

To justify the use of the word "distance", we state and prove:

Lemma 5.5 *Assuming $\mathscr{X}$ to have finite ρ-diameter, d_{W} is a metric on $\mathcal{M}_1(\mathscr{X})$.*

Proof. Neither the difference of the integrals nor the Lipchtitz norm change if a constant is added to h. This allows us to reduce to h that vanish at some point, The assumption of finite ρ-diameter then ensures that all such Lipschitz functions are bounded by the diameter and so the supremum is finite. The positivity, symmetry and the triangle inequality are then readily shown. It remains to show that

$$d_{\mathrm{W}}(\mu,\nu) = 0 \;\Rightarrow\; \mu = \nu \tag{5.19}$$

Suppose $d_{\mathrm{W}}(\mu,\nu) = 0$. Then $\int h\mathrm{d}\mu = \int h\mathrm{d}\nu$ whenever h is Lipschitz. Given a closed set $C \subseteq \mathscr{X}$, note that $h_n(x) := 1 - \min\{n\rho(x,C), 1\}$ obeys $\|h_n\|_{\mathrm{Lip}} \leqslant n$ and so its integrals under μ and ν are equal. But $h_n \downarrow 1_C$ and the Bounded Convergence Theorem gives $\int h_n \mathrm{d}\mu \to \mu(C)$. It follows that $\mu(C) = \nu(C)$ whenever C is closed. Since the closed sets generate $\mathcal{B}(\mathscr{X})$ and form a π-system, $\mu = \nu$ holds by Dynkin's π/λ-theorem. □

To make our notations consistent with (5.13) and (5.15), we write

$$d_{\mathrm{W}}(X,Y) = d_{\mathrm{W}}(\text{law of } X, \text{ law of } Y) \tag{5.20}$$

whenever X and Y are random variables taking values in the same metric space.

We note that the Wasserstein distance arises naturally in the context of *optimal transport*; indeed, $d_{\mathrm{W}}(\mu,\nu)$ is the minimal cost, as measured by the metric ρ, to "move" or "transport" the mass of μ into the mass of ν. That the minimal transport cost is expressed using (5.18) is the celebrated result called the *Kantorovich duality*.

The Wasserstein distance is more forgiving than the other two distances; indeed, we have $d_{\mathrm{K}}(\delta_x,\delta_y) = 1$ whenever $x \neq y$ yet $d_W(\delta_x,\delta_y) = \rho(x,y) \to 0$ as $y \to x$. The connection improves when one of the measures is continuously distributed:

Lemma 5.6 (Wasserstein bounds Kolmogorov) *Let μ be a Borel probability measure on $\mathbb{R}$ and f a probability density w.r.t. Lebesgue measure λ on $\mathbb{R}$. Then*

$$d_{\mathrm{K}}(\mu, f\lambda) \leqslant \sqrt{2\|f\|_\infty\, d_{\mathrm{W}}(\mu, f\lambda)} \tag{5.21}$$

Proof. Let Y have law μ and Z have law $f\lambda$. Fix $a \in \mathbb{R}$ and, for $\delta > 0$ let $h_\delta(x) = 0$ for $x < a - \delta$, $h_\delta(x) = 1$ for $x \geqslant a$ and h_δ linear on $[a-\delta, a]$. Then $\|h_\delta\|_{\mathrm{Lip}} = \delta^{-1}$ and, writing $h_0(x) := 1_{(-\infty,a]}$,

$$\begin{aligned} P(Y \leqslant a) - P(Z \leqslant a) &= Eh_0(Y) - Eh_0(Z) \\ &\leqslant Eh_\delta(Y) - Eh_\delta(Z) + Eh_\delta(Z) - Eh_0(Z) \\ &\leqslant d_{\mathrm{W}}(Y,Z)\delta^{-1} + \|f\|_\infty \delta/2 \end{aligned} \tag{5.22}$$

The right-hand side is minimizer when $\delta^2 = 2d_{\mathrm{W}}(Y,Z)/\|f\|_\infty$ at which the expression equals the right-hand side of (5.21). The corresponding lower bound is derived similarly and so we omit it. □

5.3 Stein's method and approximate normality.

We will now return to the original line of thought and apply Stein's method to get control of approximate normality. We will work using the Wasserstein distance as that is where the statement is easiest to make and derive:

Theorem 5.7 *Let $X_1, \dots, X_n$ be independent with*

$$\forall k = 1, \dots, n: \quad X_k \in L^4 \ \wedge \ EX_k = 0 \tag{5.23}$$

Assume also

$$\sum_{k=1}^{n} E(X_k^2) = 1 \tag{5.24}$$

and let $Z = \mathcal{N}(0,1)$. Then

$$d_{\mathrm{W}}\Big(\sum_{k=1}^{n} X_k, Z\Big) \leqslant \sum_{k=1}^{n} E\big(|X_k|^3\big) + \sqrt{\sum_{k=1}^{n} \mathrm{Var}(X_k^2)} \tag{5.25}$$

The proof of this theorem will require improvement on the continuity estimates from Lemma 5.3 for the linear map $h \mapsto f_h$ in the situation when $h \in \mathrm{AC}(\mathbb{R})$ and h' is itself bounded. This comes in:

Lemma 5.8 *Suppose $h \in AC(\mathbb{R})$. Then*

$$\|f_h\|_\infty \leqslant 2\|h'\|_\infty \ \wedge \ \|f_h'\|_\infty \leqslant \|h'\|_\infty \ \wedge \ \|f_h''\|_\infty \leqslant 2\|h'\|_\infty \tag{5.26}$$

Proof (ideas). The proof is based on technical estimates for the integrals in (5.3). For instance, Taylor's theorem gives

$$\big|h(y) - Eh(Z)\big| \leqslant \|h'\|_\infty E|y - Z| \tag{5.27}$$

which plugging in the first line of (5.3) shows

$$\big|f_h(x)\big| \leqslant \|h'\|_\infty \int_0^\infty \mathrm{e}^{-\frac{y^2}{2} - yx} E|x + y - Z| \mathrm{d}y \tag{5.28}$$

Bounding (for $x > 0$) $E|x + y - Z| \leqslant x + |y| + E|Z|$, we now check that the integral is bounded uniformly in $x \in [0, \infty)$. (Getting the precise numerical value needs more work.) The proof of other inequalities more tedious; see Lemma 2.4 in "Normal approximation by Stein's method" by L.N.Y. Chen, L. Goldstein and Q.-M.Yao. □

We are now ready for:

Proof of Theorem 5.7. Denote $Y := \sum_{k=1}^n X_k$ and set $Y_k := Y - X_k$. Then $EX_k = 0$ and independence of X_k and Y_k yield

$$\begin{aligned} E\big(Yf(Y)\big) &= \sum_{k=1}^{n} E\big(X_k f(Y)\big) = \sum_{k=1}^{n} E\big(X_k[\, f(Y) - f(Y_k)]\big) \\ &= \sum_{k=1}^{n} E\Big(X_k\big[\, f(Y) - f(Y_k) - (Y - Y_k)f'(Y)\big]\Big) + E\Big(f'(Y)\sum_{k=1}^{n} X_k^2\Big) \end{aligned} \tag{5.29}$$

Using Taylor's theorem then gives

$$\Big|E\big(f'(Y)-Yf(Y)\big)\Big| \leqslant \frac{1}{2}\|f''\|_\infty \sum_{k=1}^n E\big(|X_k|^3\big) + \|f'\|_\infty E\Big|1-\sum_{k=1}^n X_k^2\Big| \tag{5.30}$$

We will bound the term on the extreme right by way of the second moment. Indeed, using the independence of $X_1,\dots,X_n$ we have

$$\begin{aligned} E\bigg(\Big|1-\sum_{k=1}^n X_k^2\Big|^2\bigg) &= 1-2\sum_{k=1}^n E(X_k^2) + \sum_{k,\ell=1}^n E(X_k^2X_\ell^2) \\ &= \sum_{k=1}^n \mathrm{Var}(X_k^2) + \bigg(1-\sum_{k=1}^n E(X_k^2)\bigg)^2 = \sum_{k=1}^n \mathrm{Var}(X_k^2) \end{aligned} \tag{5.31}$$

where we use that $\sum_{k=1}^n E(X_k^2)=1$ in the last step. Invoking Stein's ODE and the regularity bounds (5.26),

$$\begin{aligned} \big|Eh(Y)-Eh(Z)\big| &= \Big|E\big(f_h'(Y)-Yf_h(Y)\big)\Big| \\ &\leqslant \|h'\|_\infty \sum_{k=1}^n E\big(|X_k|^3\big) + \|h'\|_\infty \sqrt{\sum_{k=1}^n \mathrm{Var}(X_k^2)} \end{aligned} \tag{5.32}$$

Optimizing over $h\in AC$ with $\|h'\|_\infty\leqslant 1$ gives the claim □

Applying Theorem 5.7 to i.i.d. random variables $X_1,\dots,X_n$ with $X_1\in L^4$ and $EX_1=0$ turns (5.25) into

$$d_{\mathrm{W}}\bigg(\frac{1}{\sqrt{n}}\sum_{k=1}^n, Z\bigg) \leqslant \frac{1}{\sqrt{n}}\Big[\,E(|X_1|^3) + \sqrt{\mathrm{Var}(X_1^2)}\,\Big] \tag{5.33}$$

and so we again have a normal approximation up errors of order $1/\sqrt{n}$, albeit now in the Wasserstein distance. Using the example of Bernoulli random variables (or, in fact, any integer-valued random variables) shows that also this is optimal since transporting a measure that is supported on $\frac{1}{\sqrt{n}}\mathbb{Z}$ to a measure that is continuously distributed with a density that is uniformly positive over a non-trivial interval requires moving a positive fraction of total mass mass over a distance at least order $1/\sqrt{n}$.

Stein's method is very flexible because it covers other distances (by choosing different $\mathcal{H}$) and even extends to other limit laws (which requires working with different Stein operators $\mathcal{A}$). In particular, one can prove Berry-Esseen type of estimates as well. The argument in the proof of Theorem 5.7 even allows for some amount of dependence; definitely, finite-range dependence. This refers to the situation when each X_k depends on $\{X_j\colon j\in J_k\}$ where the maximal cardinality $\max_{k\leqslant n}|J_k|$ is bounded uniformly in n. We then define $Y_k := \sum_{j\notin J_k} X_j$ and proceed pretty much as before.

Further reading: L.N.Y. Chen, L. Goldstein and Q.-M.Yao, "Normal approximation by Stein's method", Springer 2010

6. Non-gaussian limit laws

Having researched the conditions under which suitably centered and scaled sums of independent random variables tend to a normal random variable, we now move to the situations when the limit random variable is not Gaussian.

6.1 A warm-up example.

As a warm-up, we prove the following result which demonstrates existence of non-gaussian limit laws. The result should be compared with the non-standard CLT in Theorem 3.18 in which the parameter α below takes value 2.

Lemma 6.1 (Symmetric stable convergence) *For each $\alpha \in (0,2)$ there exists a random variable Y_α with characteristic function*

$$E\mathrm{e}^{\mathrm{i}tY_\alpha} = \mathrm{e}^{-|t|^\alpha}, \quad t \in \mathbb{R}, \tag{6.1}$$

such that if $X_1, X_2, \dots$ are i.i.d. with

$$X_1 \overset{\text{law}}{=} -X_1 \quad \wedge \quad \forall x > 1\colon\ P(|X_1| > x) = x^{-\alpha} \tag{6.2}$$

then for $S_n := X_1 + \cdots + X_n$ we get

$$\frac{S_n}{n^{1/\alpha}} \xrightarrow[n\to\infty]{\text{w}} c_\alpha Y_\alpha \tag{6.3}$$

where $c_\alpha := [\alpha \int_0^\infty \frac{1-\cos x}{x^{\alpha+1}}\mathrm{d}x]^{1/\alpha}$.

Proof. We have $E(\mathrm{e}^{\mathrm{i}tS_n}) = [E(\mathrm{e}^{\mathrm{i}tX_1})]^n$ so we need a good representation of characteristic function of X_1:

$$\begin{aligned} E(\mathrm{e}^{\mathrm{i}tX_1}) &= E(\cos(tX_1)) \\ &= 1 - E\big(1-\cos(tX_1)\big) \\ &= 1 - \alpha\int_1^\infty \frac{1-\cos(tx)}{x^{\alpha+1}}\mathrm{d}x \\ &= 1 - |t|^\alpha \underbrace{\int_{|t|}^\infty \alpha\frac{1-\cos(x)}{x^{\alpha+1}}\mathrm{d}x}_{=:I_\alpha(t)} \end{aligned} \tag{6.4}$$

Since $1-\cos(x)$ vanishes proportionally to! x^2, the Monotone Convergence Theorem shows that, for all $\alpha \in (0,2)$, we have $I_\alpha(t) \to I_\alpha(0) \in (0,\infty)$ as $t \to 0$. Hence

$$E(\mathrm{e}^{\mathrm{i}tS_n/n^{1/\alpha}}) = \left(1 - \frac{1}{n}|t|^\alpha I_\alpha(tn^{-1/\alpha})\right)^n \underset{n\to\infty}{\longrightarrow} \mathrm{e}^{-I_\alpha(0)|t|^\alpha} \tag{6.5}$$

The Lévy continuity theorem (Theorem 3.11) now gives that Y_α exists and, noting that $c_\alpha = I_\alpha(0)^{1/\alpha}$, (6.3) holds. □

Lemma 6.1 should be compared with the Marcinkiewicz-Zygmund's SLLN (Theorem 2.15) which ways that $S_n/n^{1/\alpha} \to 0$ a.s. whenever $E(X_1^\alpha) < \infty$. This assumption (and thus the conclusion) fails for the distribution in (6.2), albeit just marginally.

We will later call Y_α a "symmetric stable random variable with index α." Some special cases are easy to identify: Y_1 is a Cauchy random variable, Y_2 is $\mathcal{N}(0,2)$. The random variable $Y_{1/2}$ is called "Lévy distributed" which reflects on the fact that its probability density can be explicitly identified. (All other Y_α have densities but they are inexplicit.)

6.2 Stable convergence.

Having discovered non-gaussian limit laws, we may ask: How general are these? Moreover, we are interested whether similar convergence can be established when the symmetry and pristine power-law tail assumption (6.2) is relaxed. In order to state the corresponding result, we need:

Definition 6.2 $L\colon (0,\infty) \to (0,\infty)$ *is said to be* slowly varying *at* $+\infty$ *if,*

$$\forall x > 0\colon \quad \lim_{t\to\infty} \frac{L(xt)}{L(t)} = 1 \tag{6.6}$$

We remark that functions for which the limit exists for all $x > 0$ (but is not necessarily equal to 1) are called *regularly varying*. Scaling considerations then force the limit to take the form x^α for some exponent α. As is readily checked, every regularly varying function is thus a power times a slowly varying function, so the above is all we need.

Here is our result on non-gaussian limit laws for sums of i.i.d. random variables:

Theorem 6.3 (Stable convergence) *Let $X_1, X_2, \dots$ be i.i.d. random variables such that for some $\alpha \in (0,2)$ and $\theta \in [0,1]$ the following holds:*

(A) $P(|X_1| > x) = x^{-\alpha} L(x)$ *for L slowly varying at $+\infty$*
(B) $P(X_1 > x)/P(|X_1| > x) \underset{x\to\infty}{\longrightarrow} \theta$

Denote $S_n := X_1 + \cdots + X_n$ and set

$$\begin{aligned} a_n &:= \inf\bigl\{x > 0\colon nP(|X_1| > x) \leqslant 1\bigr\} \\ b_n &:= nE\bigl(X_1 1_{\{|X_1| \leqslant a_n\}}\bigr) \end{aligned} \tag{6.7}$$

Then

$$\frac{S_n - b_n}{a_n} \underset{n\to\infty}{\overset{\text{w}}{\longrightarrow}} Y \tag{6.8}$$

where Y is the random variable such that

$$E(\mathrm{e}^{\mathrm{i}tY}) = \exp\left\{ \int \bigl(\mathrm{e}^{\mathrm{i}tx} - 1 - \mathrm{i}tx 1_{[-1,1]}(x)\bigr)\bigl(\theta 1_{\{x>0\}} + (1-\theta)1_{\{x<0\}}\bigr) \frac{\alpha \mathrm{d}x}{|x|^{\alpha+1}} \right\} \tag{6.9}$$

holds for all $t \in \mathbb{R}$.

The parameter α has the same meaning as earlier: it gives us the scaling exponent for the decay of the tail of $|X_1|$. The parameter θ controls how much of the mass of X_1 comes from the positive part vs the negative part. Instead of (A) we could assume that both $x \mapsto P(X_1 > x)$ and $x \mapsto P(X_1 < -x)$ are regularly varying, which implies $P(X_1 > x) = x^{-\alpha_1} L_1(x)$ and $P(X_1 < -x) = x^{-\alpha_2} L_2(x)$ for L_1, L_2 slowly varying. The above would

then hold with the choice $\alpha := \min\{\alpha_1, \alpha_2\}$. (Part (B) would still have to be assumed. We will have $\theta = 1$ if $\alpha_1 < \alpha_2$ and $\theta = 0$ if $\alpha_1 > \alpha_2$.)

In order to justify the definition of a_n, we state and prove:

Lemma 6.4 *Under assumption in Theorem 6.3,*

$$nP(|X_1| > a_n) \leqslant 1 \ \wedge \ nP(|X_1| > a_n) \underset{n\to\infty}{\longrightarrow} 1 \tag{6.10}$$

Moreover, for each $t > 0$,

$$nP(X_1 > ta_n) \underset{n\to\infty}{\longrightarrow} \theta t^{-\alpha} \ \wedge \ nP(X_1 < -ta_n) \underset{n\to\infty}{\longrightarrow} (1-\theta)t^{-\alpha} \tag{6.11}$$

Proof. Let $\epsilon > 0$. The right-continuity of $t \mapsto P(|X_1| > t)$ along with definition of a_n give

$$nP(|X_1| > a_n) \leqslant 1 \leqslant nP(|X_1| > a_n - \epsilon a_n) \tag{6.12}$$

By (A) in Theorem 6.3, the ratio of the right-hand side and the left hand side tends to $(1-\epsilon)^{-\alpha}$ which can be made as close to one as desired. So $nP(|X_1| > a_n) \to 1$. Using this along with (A-B) in Theorem 6.3 then gives also (6.11). □

The formulas (6.10–6.11) explain the reason for defining a_n the way it is defined. Indeed, we get that $P(|X_1| > ta_n)$ is asymptotically of order $1/n$ so n independent attempts for such an event will succeed at least ones with positive probability. More precisely, by (6.10) the probability that at least one of $X_1, \ldots, X_n$ exceeds exceed ϵa_n in absolute value is asymptotically expressed as

$$1 - \big(1 - P(|X_1| > \epsilon a_n)\big)^n \underset{n\to\infty}{\longrightarrow} 1 - \mathrm{e}^{-\epsilon^\alpha} \tag{6.13}$$

Note that the right-hand side tends to one as $\epsilon \downarrow 0$ and so, for ϵ small, this is actually very likely to occur.

In order to work with the tail probabilities $P(|X_1| > ta_n)$, we need more than the mere asymptotics (6.10–6.11). Indeed, as these will invariably appear under integration where the fact that $t \mapsto t^{-\alpha}$ diverges as $t \downarrow 0$ could cause problems, we also derive the following hard upper bound:

Lemma 6.5 *For each $\delta > 0$ there are $c > 0$ and $t_0 < \infty$ such that*

$$\forall x \in (0,1]\, \forall t \geqslant t_0 \colon \ \frac{P(|X_1| > xt)}{P(|X_1| > t)} \leqslant cx^{-\alpha-\delta} \tag{6.14}$$

Proof. By (A) we can find find $t_0 > 0$ so large that

$$\forall t \geqslant t_0 \colon \ \frac{P(|X_1| > t/2)}{P(|X_1| > t)} \leqslant 2^{\alpha+\delta} \tag{6.15}$$

Given $x \in (0,1]$, let $n \in \mathbb{Z}$ be such that $2^{-n} \leqslant x < 2^{-n+1}$. Then for any $t \geqslant t_0$,

$$\frac{P(|X_1| > xt)}{P(|X_1| > t)} \leqslant \frac{P(|X_1| > 2^{-n}t)}{P(|X_1| > t)} = \frac{P(|X_1| > 2^{-n}t)}{P(|X_1| > 2^{-k}t)} \prod_{\ell=0}^{k-1} \frac{P(|X_1| > 2^{-\ell-1}t)}{P(|X_1| > 2^{-\ell}t)} \tag{6.16}$$

where $k := \max\{0 \leqslant \ell \leqslant n \colon 2^{-\ell} t \geqslant t_0\}$. By (6.15), each ratio in the product is at most $2^{\alpha+\delta}$ which bounds the right-hand side by $P(|X_1| > 2t_0)^{-1} 2^{k(\alpha+\delta)}$. Since $2^k \leqslant 2^n \leqslant 2x^{-1}$, the claim follows. □

6.3 Separation in "small" and "large" values.

The proof of Theorem 6.3 is based on the following idea. We use a_n as a "cutoff" scale with "small values" designating those smaller than ϵa_n and "large values" marking those that exceed ϵa_n. We will then show that the sum of the "small values" is asymptotically negligible (as $n \to \infty$ and $\epsilon \downarrow 0$) while the sum of "large values" admits a weak limit as $n \to \infty$ which then also converges as $\epsilon \downarrow 0$. Note that this is quite unlike what happened for CLT, where it was the small values that carried the bulk of the limiting contribution.

In order to start implementing of the above strategy, pick $\epsilon \in (0,1)$ (to be sent to zero later) and, for each $1 \leqslant k \leqslant n$ set

$$\overline{X}_{n,k} := X_k 1_{\{|X_k| \leqslant \epsilon a_n\}}. \tag{6.17}$$

Let

$$\overline{S}_n := \sum_{k=1}^{n} \overline{X}_{n,k} \tag{6.18}$$

We then have:

Lemma 6.6 (Small-value variance) *For each $\delta \in (0, 2-\alpha)$ there is $C < \infty$ such that*

$$\limsup_{n\to\infty} \operatorname{Var}\Big(\frac{\overline{S}_n}{a_n}\Big) \leqslant C\epsilon^{2-\alpha-\delta} \tag{6.19}$$

Proof. By independence we have

$$\begin{aligned} \operatorname{Var}\Big(\frac{\overline{S}_n}{a_n}\Big) &\leqslant nE\Big(\frac{\overline{X}_{n,1}^2}{a_n^2}\Big) \leqslant \frac{n}{a_n^2}\int_0^{\epsilon a_n} 2tP(|X_1| > t)\mathrm{d}t \\ &= n\int_0^{\epsilon} 2xP(|X_1| > xa_n)\mathrm{d}x \leqslant \int_0^{\epsilon} 2cx^{1-\alpha-\delta}\mathrm{d}x \end{aligned} \tag{6.20}$$

where we used

$$P(|X_1| > xa_n) \leqslant cx^{-\alpha-\delta} nP(|X_1| > a_n) \leqslant cx^{-\alpha-\delta} \tag{6.21}$$

as implied by Lemmas 6.5 and 6.4. □

To account for "large values," for each $1 \leqslant k \leqslant n$ set

$$\widehat{X}_{n,k} := X_k 1_{\{|X_k| > \epsilon a_n\}} \tag{6.22}$$

and let

$$\widehat{S}_n := \sum_{k=1}^{n} \widehat{X}_{n,k} \tag{6.23}$$

We then have:

Lemma 6.7 (Large-value limit) *For all $\epsilon > 0$ and all $t \in \mathbb{R}$,*

$$E(\mathrm{e}^{\mathrm{i}t\widehat{S}_n/a_n}) \underset{n\to\infty}{\longrightarrow} \exp\Big\{ \int \big(\mathrm{e}^{\mathrm{i}tx} - 1\big)\big(\theta 1_{\{x>\epsilon\}} + (1-\theta)1_{\{x<-\epsilon\}}\big)\frac{\alpha \mathrm{d}x}{|x|^{\alpha+1}} \Big\} \tag{6.24}$$

In particular, $\widehat{S}_n/a_n$ admits a weak limit

Proof. Define Borel measures $\mu_n^{\epsilon,+}$ and $\mu_n^{\epsilon,-}$ on $\mathbb{R}$ by

$$\mu_n^{\epsilon,\pm}(A) := nP\big(\widehat{X}_{n,1}/a_n \in A \cap \mathbb{R}_\pm\big) \tag{6.25}$$

While these measures are not probabilities, their total mass is bounded. Indeed, we have

$$\mu_n^{\epsilon,+}(\mathbb{R}) + \mu_n^{\epsilon,-}(\mathbb{R}) \leqslant nP(|X_1| > \epsilon a_n) \underset{n\to\infty}{\longrightarrow} \epsilon^{-\alpha}. \tag{6.26}$$

Next observe that, for $x > \epsilon$,

$$\begin{aligned} \mu_n^{\epsilon,+}((-\infty, x]) &= nP(\epsilon a_n < X_1 \leqslant x a_n) \\ &= nP(X_1 > \epsilon a_n) - nP(X_1 > x a_n) \\ &\underset{n\to\infty}{\longrightarrow} \theta\big(\epsilon^{-\alpha} - x^{-\alpha}\big) = \int_{(-\infty,x]} \theta 1_{\{x>\epsilon\}} \frac{\alpha \mathrm{d}x}{|x|^{\alpha+1}} \end{aligned} \tag{6.27}$$

and similarly for $\mu_n^{\epsilon,-}$. It follows that

$$\mu_n^{\epsilon,+} + \mu_n^{\epsilon,-} \underset{n\to\infty}{\overset{\mathrm{w}}{\longrightarrow}} \big(\theta 1_{\{x>\epsilon\}} + (1-\theta)1_{\{x<-\epsilon\}}\big)\frac{\alpha}{|x|^{\alpha+1}}\mathrm{d}x \tag{6.28}$$

where $\mathrm{d}x$ stands for the Lebesgue measure on $\mathbb{R}$.

To prove the claim we write

$$E(\mathrm{e}^{\mathrm{i}t\widehat{S}_n/a_n}) = \left(1 + \frac{nE(\mathrm{e}^{\mathrm{i}t\widehat{X}_{n,1}/a_n} - 1)}{n}\right)^n \tag{6.29}$$

and observe that that, by (6.28) and the fact that $x \mapsto \mathrm{e}^{\mathrm{i}tx} - 1$ is bounded and continuous,

$$\begin{aligned} nE(\mathrm{e}^{\mathrm{i}t\widehat{X}_{n,1}/a_n} - 1) &= \int (\mathrm{e}^{\mathrm{i}tx} - 1)\big(\mu_n^{\epsilon,+}(\mathrm{d}x) + \mu_n^{\epsilon,-}(\mathrm{d}x)\big) \\ &\underset{n\to\infty}{\longrightarrow} \int \big(\mathrm{e}^{\mathrm{i}tx} - 1\big)\big(\theta 1_{\{x>\epsilon\}} + (1-\theta)1_{\{x<-\epsilon\}}\big)\frac{\alpha \mathrm{d}x}{|x|^{\alpha+1}} \end{aligned} \tag{6.30}$$

Plugging this in (6.29), we get the claim. □

To demonstrate the power of these lemmas, we now give a proof of a simplified form of the limit law in the case when $\alpha < 1$:

Corollary 6.8 *Suppose $\alpha \in (0,1)$. Then $S_n/a_n \overset{\mathrm{w}}{\longrightarrow} \widetilde{Y}$ where $\widetilde{Y}$ has the characteristic function*

$$E\mathrm{e}^{\mathrm{i}t\widetilde{Y}} = \exp\Big\{ \int \big(\mathrm{e}^{\mathrm{i}tx} - 1\big)\big(\theta 1_{\{x>0\}} + (1-\theta)1_{\{x<-0\}}\big)\frac{\alpha \mathrm{d}x}{|x|^{\alpha+1}} \Big\} \tag{6.31}$$

In particular, the convergence holds even without centering by b_n.

Proof. The definitions imply

$$S_n = \overline{S}_n + \widehat{S}_n = (\overline{S}_n - E\overline{S}_n) + \widehat{S}_n + E\overline{S}_n \tag{6.32}$$

Lemma 6.6 gives

$$\frac{\overline{S}_n - E\overline{S}_n}{a_n} \xrightarrow[n\to\infty,\,\epsilon\downarrow 0]{L^2} 0 \tag{6.33}$$

while Lemma 6.7 tells us

$$\frac{\widehat{S}_n}{a_n} \xrightarrow[n\to\infty]{\text{w}} \widetilde{Y}_\epsilon \tag{6.34}$$

where $\widetilde{Y}_\epsilon$ is the random variable with the characteristic function on the right of (6.24). Since the integral in that formula converges even after $\epsilon \downarrow 0$, we get

$$\widetilde{Y}_\epsilon \xrightarrow[\epsilon\downarrow 0]{\text{w}} \widetilde{Y} \tag{6.35}$$

where $\widetilde{Y}$ is as in (6.31). We thus need to show

$$\lim_{\epsilon\downarrow 0} \limsup_{n\to\infty} \frac{|E\overline{S}_n|}{a_n} = 0 \tag{6.36}$$

For this we pick $\delta \in (0, 1-\alpha)$ and compute

$$\begin{aligned} \frac{|E\overline{S}_n|}{a_n} &\leqslant \frac{n}{a_n} E\bigl(|\overline{X}_{n,1}|\bigr) = \frac{n}{a_n}\int_0^{\epsilon a_n} P\bigl(|X_1| > t\bigr)\mathrm{d}t \\ &\leqslant \int_0^{\epsilon} nP\bigl(|X_1| > xa_n\bigr)\mathrm{d}x \leqslant c\int_0^{\epsilon} x^{-\alpha-\delta}\mathrm{d}x = \frac{c}{1-\alpha-\delta}\,\epsilon^{1-\alpha-\delta} \end{aligned} \tag{6.37}$$

where we also invoked the bound in Lemma 6.5 and assumed that n is so large that $a_n \geqslant t_0$. This now readily gives (6.36). □

6.4 Proof of stable convergence.

The proof of Corollary 6.8 highlights the issues that are left to be dealt with in order to prove Theorem 6.3. These are

(1) $\widetilde{Y}_\epsilon$ converges weakly as $\epsilon \downarrow 0$ for $\alpha \in (0,1)$ but not fo $\alpha \geqslant 1$ in general because $x \mapsto (\mathrm{e}^{\mathrm{i}tx} - 1)x^{-\alpha-1}$ is not Lebesgue-integrable near 0.
(2) $E\overline{S}_n/a_n$ cannot be expected to converge as $n \to \infty$ for $\alpha \in [1,2)$; check, e.g. $P(X_1 \geqslant x) = x^{-\alpha}$ for $x > 1$.

The centering by $b_n := nE(X_1 1_{\{|X_1|\leqslant a_n\}})$ solves these and gives us a proof that works uniformly for all $\alpha \in (0,2)$. Noting that

$$b_n - E\overline{S}_n = nE\bigl(\widehat{X}_{n,1} 1_{\{|\widehat{X}_{n,1}|\leqslant a_n\}}\bigr) \tag{6.38}$$

all we need to do is to prove:

Lemma 6.9 (Centering term) *For all $\epsilon > 0$,*

$$\frac{n}{a_n} E\bigl(\widehat{X}_{n,1} 1_{\{|\widehat{X}_{n,1}|\leqslant a_n\}}\bigr) \xrightarrow[n\to\infty]{} \int x 1_{[-1,1]}(x)\,\bigl(\theta 1_{\{x>\epsilon\}} + (1-\theta)1_{\{x<-\epsilon\}}\bigr)\frac{\alpha\mathrm{d}x}{|x|^{\alpha+1}} \tag{6.39}$$

Proof. Using notation from Lemma 6.7,

$$\frac{n}{a_n} E\big(\widehat{X}_{n,1} 1_{\{|\widehat{X}_{n,1}| \leqslant a_n\}}\big) = \int x 1_{[-1,1]}(x) \big(\mu_n^{\epsilon,+}(\mathrm{d}x) + \mu_n^{\epsilon,-}(\mathrm{d}x)\big) \tag{6.40}$$

Since $x \mapsto x 1_{[-1,1]}(x)$ is bounded and continuous Lebesgue a.e. the claim follows from a straightforward approximation of $1_{[-1,1]}$ by continuous functions. □

We are now ready for:

Proof of Theorem 6.3. We just put together the facts proved earlier. First

$$\frac{S_n - b_n}{a_n} = \frac{\overline{S}_n - E\overline{S}_n}{a_n} + \frac{1}{a_n}\Big(\widehat{S}_n - nE\big(\widehat{X}_{n,1} 1_{\{|\widehat{X}_{n,1}| \leqslant a_n\}}\big)\Big) \tag{6.41}$$

The above lemmas imply

$$\frac{\overline{S}_n - E\overline{S}_n}{a_n} \xrightarrow[n\to\infty,\, \epsilon \downarrow 0]{P} 0 \tag{6.42}$$

and

$$\frac{1}{a_n}\Big(\widehat{S}_n - nE\big(\widehat{X}_{n,1} 1_{\{|\widehat{X}_{n,1}| \leqslant a_n\}}\big)\Big) \xrightarrow[n\to\infty]{\mathrm{w}} Y_\epsilon \tag{6.43}$$

where Y_ϵ has the characteristic function

$$\exp\left\{ \int \big(\mathrm{e}^{\mathrm{i}tx} - 1 - \mathrm{i}tx 1_{[-1,1]}(x)\big)\big(\theta 1_{\{x>\epsilon\}} + (1-\theta) 1_{\{x<-\epsilon\}}\big) \frac{\alpha \mathrm{d}x}{|x|^{\alpha+1}} \right\} \tag{6.44}$$

Since $x \mapsto \mathrm{e}^{\mathrm{i}tx} - 1 - \mathrm{i}tx 1_{[-1,1]}(x)$ is bounded by a constant times x^2 for x small, this now allows taking $\epsilon \downarrow 0$ for all $\alpha \in (0,2)$. By the Lévy continuity theorem we get $Y_\epsilon \xrightarrow{\mathrm{w}} Y$ with Y having the characteristic function (6.9). □

While the conditions (A-B) in Theorem 6.3 may appear rather special, they are actually necessary. Indeed, assuming that $(S_n - b_n)/a_n$ converges weakly to a random variable then either Y is normal and $X_1 \in L^2$, or conditions (A-B) hold with some $\alpha \in (0,2)$. Proofs of this result (and many other results) can be found in L. Breiman's "Probability" which is an ultimate reference for many facts about sums of independent random variables.

Further reading: Durrett, Section 3.8

7. Stable laws and Convergence of types

In the previous lecture we noted a whole new family of possible limit laws for suitably centered and normalized sums of i.i.d. random variables that lack second moments. In this lecture we will characterize these limit laws in an intrinsic way that will later allow us to show that, along with the normal laws that kick in when the second moment is finite, these are all that one can get.

7.1 More on the limit formula.

Let $\alpha \in (0,2)$. In Theorem 6.3 we showed that suitably centered and normalized sums of certain i.i.d. random variables with heavy, albeit regularly varying, tails that just barely miss the α'th moment tend to a random variable Y with characteristic function

$$E\big(\mathrm{e}^{\mathrm{i}tY}\big) = \exp\left\{ \int \big(\mathrm{e}^{\mathrm{i}tx} - 1 - \mathrm{i}tx1_{[-1,1]}(x)\big)\big(\theta 1_{\{x>0\}} + (1-\theta)1_{\{x<0\}}\big)\frac{\alpha \mathrm{d}x}{|x|^{\alpha+1}} \right\} \tag{7.1}$$

A natural question is: Is there a better (read: non-integral) formula? A calculation that we will not perform answers this affirmatively by producing a different expression

$$E\big(\mathrm{e}^{\mathrm{i}tY}\big) = \begin{cases} \exp\Big\{ \mathrm{i}t\mu - \sigma^\alpha |t|^\alpha \big(1 - \mathrm{i}\beta \mathrm{sign}(t)\tan\big(\frac{\pi\alpha}{2}\big)\big)\Big\} & \text{if } \alpha \neq 1 \\ \exp\Big\{ \mathrm{i}t\mu - \sigma|t|\big(1 - \mathrm{i}\beta \mathrm{sign}(t)\frac{2}{\pi}\log|t|\big)\big)\Big\} & \text{if } \alpha = 1 \end{cases} \tag{7.2}$$

This expression depends on four parameters that have the following names

- α = index of stability
- β = skewness ($\beta := 2\theta - 1$)
- σ = scale
- μ = centering or shift

The notation for μ and σ is done in analogy with the characteristic function

$$\exp\left\{ \mathrm{i}t\mu - \frac{1}{2}\sigma^2 t^2 \right\} \tag{7.3}$$

for the normal law with mean μ and variance σ^2. (Note that, except for the factor of $\frac{1}{2}$, (7.3) arises from (7.2) in the limit as $\alpha \uparrow 2$ because $\tau(\frac{\pi\alpha}{2}) \to 0$ in this case.) The skewness tells us, roughly, how much mass is on the positive side relative to the negative side. The extreme cases $\beta = \pm 1$ are said to be *totally skewed*.

While we did not discuss this in the lecture, let us note some properties of the above random variables. Let $S_\alpha(\sigma,\beta,\mu)$ be the random variable Y with characteristic function (7.2). Assuming $S_\alpha(\sigma_1,\beta_1,\mu_1)$ and $S'_\alpha(\sigma_2,\beta_2,\mu_2)$ to be independent, we then have

$$S_\alpha(\sigma_1,\beta_1,\mu_1) + S'_\alpha(\sigma_2,\beta_2,\mu_2) \overset{\text{law}}{=} S_\alpha(\sigma,\beta,\mu) \tag{7.4}$$

where

$$\sigma^\alpha := \sigma_1^\alpha + \sigma_2^\alpha, \quad \mu := \mu_1 + \mu_2 \quad \text{and} \quad \beta := \frac{\sigma_1^\alpha\beta_1 + \sigma_2^\alpha\beta_2}{\sigma_1^\alpha + \sigma_2^\alpha} \tag{7.5}$$

Moreover,

$$S_\alpha(\sigma,\beta,\mu) \overset{\text{law}}{=} \mu + S_\alpha(\sigma,\beta,0) \tag{7.6}$$

and, for all $c \in \mathbb{R}$,

$$cS_\alpha(\sigma,\beta,\mu) \overset{\text{law}}{=} \begin{cases} S_\alpha(|c|\sigma, \operatorname{sign}(c)\beta, c\mu) & \text{if } \alpha \neq 1 \\ S_\alpha\big(|c|\sigma, \operatorname{sign}(c)\beta, c\mu - \frac{2}{\pi}\beta c \log|c|\big) & \text{if } \alpha = 1 \end{cases} \tag{7.7}$$

In particular,

$$-S_\alpha(\sigma,\beta,\mu) \overset{\text{law}}{=} S_\alpha(\sigma,-\beta,-\mu) \tag{7.8}$$

The fact that the case $\alpha = 1$ picks up an additional additive term under scaling indicates that caution is needed in every use of the parameterization (7.2).

Let us also note some special cases that are worthy of attention. First observe that under vanishing skewness, $S_\alpha(\sigma,0,\mu)$ is a symmetric random variable with characteristic function $\mathrm{e}^{\mathrm{i}t\mu-\sigma^\alpha|t|^\alpha}$. Next note that, for the normal variable

$$S_2(\sigma,\beta,\mu) = \mathcal{N}(\mu,\sqrt{2}\sigma^2) \tag{7.9}$$

skewness is irrelevant. Another special case is the Cauchy distribution,

$$S_1(\sigma,0,\mu) \overset{\text{law}}{=} \mu + \sigma Z \tag{7.10}$$

where Z is the "standard" Cauchy with probability density $z \mapsto \frac{1}{\pi}\frac{1}{1+z^2}$ with respect to the Lebesgue measure. This follows from

$$\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\mathrm{e}^{\mathrm{i}tz}}{1+z^2}\mathrm{d}z = \mathrm{e}^{-|t|} \tag{7.11}$$

as computed using residue calculus. Note that this applies only for $\beta = 0$; the random variable $S_1(\sigma,\beta,\mu)$ is not Cauchy when $\beta \neq 0$.

Finally, the *Lévy distribution* $S_{1/2}(\sigma,1,\mu)$ has probability density

$$x \mapsto \sqrt{\frac{\sigma}{2\pi}}\frac{1}{(x-\mu)^{3/2}}\,\mathrm{e}^{-\frac{\sigma}{2(x-\mu)}}1_{(\mu,\infty)}(x) \tag{7.12}$$

which is also known as the inverse-Gamma distribution.

7.2 Intrinsic characterization.

In spite of all the above developments, a question remains whether others laws besides those we have identified so far can arise as weak limits of centered and scaled sums of i.i.d. random variables. In order to answer this, we note an intrinsic property that any such law has to satisfy:

Definition 7.1 *Y is said to be* stable *if there are i.i.d. random variables $Y_1, Y_2, \ldots$ and $\mathbb{R}$-valued sequences $\{A_n\}_{n\geqslant 1}$ and $\{B_n\}_{n\geqslant 1}$ with $\forall n \geqslant 1\colon\ A_n > 0$ such that*

$$\forall n \geqslant 1\colon\quad Y \overset{\text{law}}{=} \frac{Y_1 + \cdots + Y_n}{A_n} - B_n \tag{7.13}$$

We then state and prove:

Theorem 7.2 *For any random variable Y, the following are equivalent:*

(1) *there are i.i.d. random variables* $X_1, X_2, \dots$ *and* $\mathbb{R}$*-valued sequences* $\{b_n\}_{n\geqslant 1}$ *and* $\{a_n\}_{n\geqslant 1}$ *with* $\forall n \geqslant 1\colon a_n > 0$ *such that*

$$\frac{X_1 + \cdots + X_n}{a_n} - b_n \xrightarrow[n\to\infty]{w} Y \tag{7.14}$$

(2) Y *is stable and the random variable* $Y_1, Y_2, \dots$ *in* (7.13) *can be taken so that* $Y_1 \overset{\text{law}}{=} Y$.

Proof. The part (2) $\Rightarrow$ (1) is immediate from the definition of a stable random variable. Indeed, take $X_1 \overset{\text{law}}{=} Y_1$ and $a_n := A_n$ and $b_n := B_n$ and note that the expression on the left of (7.14) has the law of Y for all $n \geqslant 1$.

The proof of (1) $\Rightarrow$ (2) is more complicated. Assume (1) and denote $S_n = X_1 + \cdots + X_n$. Then (7.14) reads as

$$Z_n := \frac{S_n}{a_n} - b_n \xrightarrow[n\to\infty]{\text{w}} Y \tag{7.15}$$

Now pick $k \geqslant 1$ and note

$$\begin{aligned} Z_{nk} &= \frac{1}{a_{nk}} \sum_{j=1}^{k} (S_{nj} - S_{n(j-1)}) - b_{nk} \\ &= \frac{a_n}{a_{nk}} \sum_{j=1}^{k} \underbrace{\Big(\frac{S_{nj} - S_{n(j-1)}}{a_n} - b_n \Big)}_{=: Z_n^{(j)}} + \Big(\frac{k a_n}{a_{nk}} b_n - b_{nk} \Big) \end{aligned} \tag{7.16}$$

where $Z_n^{(1)}, \dots, Z_n^{(k)}$ are i.i.d. copies of Z_n. Since $Z_n \xrightarrow{\text{w}} Y$, the sum $Z_n^{(1)} + \cdots + Z_n^{(k)}$ tends in law to the sum $Y_1 + \cdots + Y_K$ of i.i.d. copies $Y_1, \dots, Y_k$ of Y. (This is checked at the level of characteristic functions.) The key problem is what happens with the sequences

$$\Big\{ \frac{a_n}{a_{nk}} \Big\}_{n\geqslant 1} \quad \text{and} \quad \Big\{ \frac{k a_n}{a_{nk}} b_n - b_{nk} \Big\}_{n\geqslant 1} \tag{7.17}$$

in this limit. Here we call upon:

Theorem 7.3 (Convergence of types) *Suppose* $X_n \xrightarrow{\text{w}} X$ *with* X *non-degenerate and let* $\alpha_n > 0$ *and* β_n *be real numbers such that* $\alpha_n X_n + \beta_n \xrightarrow{\text{w}} Y$ *for some* Y *non-degenerate. Then there are* $\alpha > 0$ *and* $\beta \in \mathbb{R}$ *such that*

(1) $\alpha_n \to \alpha \ \wedge \ \beta_n \to \beta$

(2) $Y \overset{\text{law}}{=} \alpha X + \beta$

Deferring the proof till later, we now observe that if Y is constant a.s., then it is stable, so we may assume that Y is indeed non-degenerate. Theorem 7.3 then gives

$$\exists A_k > 0 \, \exists B_k \in \mathbb{R}\colon \quad \frac{a_n}{a_{nk}} \xrightarrow[n\to\infty]{} \frac{1}{A_k} \ \wedge \ \frac{k a_n}{a_{nk}} b_n - b_{nk} \xrightarrow[n\to\infty]{} -B_k \tag{7.18}$$

and

$$Y \overset{\text{law}}{=} \frac{Y_1 + \cdots + Y_k}{A_k} - B_k \tag{7.19}$$

where $Y_1, \dots, Y_k$ are i.i.d. with law of Y, proving (2). □

7.3 Proof of Convergence of Types.

The proof of Theorem 7.3 is based on manipulations with characteristic functions. We start by noting:

Lemma 7.4 (Locally uniform convergence of ch.f.'s) *If $X_n \xrightarrow{\text{w}} X$, then*

$$\forall T > 0\colon \quad \sup_{-T \leqslant t \leqslant T} \bigl| E\mathrm{e}^{\mathrm{i}tX_n} - E\mathrm{e}^{\mathrm{i}tX} \bigr| \underset{n\to\infty}{\longrightarrow} 0 \tag{7.20}$$

In short, under weak convergence of r.v.'s, the associated ch.f.'s converge locally uniformly

Proof. By Skorohod representation (Theorem 3.4) we may assume $X_n \to X$ a.s. Using that $|1 - \mathrm{e}^{-\mathrm{i}a}| = 2|\sin(a/2)|$, for each $\epsilon \in (0,1)$ and $t \in [-T, T]$ we then have

$$\begin{aligned} \bigl| E\mathrm{e}^{\mathrm{i}tX_n} - E\mathrm{e}^{\mathrm{i}tX} \bigr| &\leqslant 2E|\sin(t(X_n - X)/2)| \\ &\leqslant 2P\bigl(|X_n - X| > \epsilon/T\bigr) + 2\sin(\epsilon/2) \end{aligned} \tag{7.21}$$

Now take $n \to \infty$ followed by $\epsilon \downarrow 0$. □

The main benefit of the uniformity is that the characteristic functions converge even if t in the characteristic function of X_n is replaced by t_n such that $t_n \to t$. This is quite useful in approximation arguments.

Next we observe three properties whose proof we relegate to homework:

Lemma 7.5 *Denote by $\varphi_X(t) := E(\mathrm{e}^{\mathrm{i}tX})$ the characteristic function of X. Then*

(1) *the statement $\exists \delta > 0 \forall t \in (-\delta, \delta)\colon\ |\varphi_X(t)| = 1$ implies that X is constant a.s.*

(2) *the statement*

$$\exists \delta > 0\, \forall t \in (-\delta, \delta)\colon\ \operatorname{Re} \varphi_{X_n}(t) \to 1 \tag{7.22}$$

implies $X_n \xrightarrow{\text{P}} 0$

(3) *the statement*

$$\exists a \in \mathbb{R} \setminus \{+1, -1\} \forall t \in \mathbb{R}\colon\ |\varphi_X(at)| = |\varphi_X(t)| \tag{7.23}$$

implies X is constant a.s.

We rush to add that other perhaps similar statements about characteristic functions may fail. The most notable is that equality of two characteristic function on an open subinterval of $\mathbb{R}$ does not imply their equality everywhere (and thus equality of the associated laws). For the same reason it does not suffice to check convergence of characteristic function on a subinterval of $\mathbb{R}$, even if that contains the origin.

We are now ready to start:

Proof of Theorem 7.3. Assume $X_n \xrightarrow{\text{w}} X$ and $\alpha_n X_n + \beta_n \xrightarrow{\text{w}} Y$ with X, Y non-degenerate and $\alpha_n > 0$ for each $n \geqslant 1$. Write $\varphi_Z(t)$ for the characteristic function of random variable Z. We then have

$$\varphi_{\alpha_n X_n + \beta_n}(t) = \mathrm{e}^{\mathrm{i}t\beta_n} \varphi_{X_n}(\alpha_n t) \tag{7.24}$$

and so the above convergences give

$$\varphi_{X_n}(t) \to \varphi_X(t) \ \wedge \ \mathrm{e}^{\mathrm{i}t\beta_n}\varphi_{X_n}(\alpha_n t) \to \varphi_Y(t) \tag{7.25}$$

uniformly on compact sets of t. We now prove a series of claims.

Claim 1: $0 < \inf_{n\geqslant 1}\alpha_n \leqslant \sup_{n\geqslant 1}\alpha_n < \infty$

We proceed via argument by contradiction. If $\alpha_{n_k} \to 0$ along some sequence $n_k \to \infty$ then the uniform convergence in the first limit in (7.25) implies $|\varphi_{X_n}(\alpha_{n_k}t)| \to |\varphi_X(0)| = 1$. But then the second limit in (7.25) gives $|\varphi_Y(t)| = 1$ for all $t \in \mathbb{R}$ which by Lemma 7.5 forces Y to be constant a.s., in contradiction with our assumptions.

Similarly, if $\alpha_{n_k} \to \infty$ along some sequence $n_k \to \infty$, then taking t/α_n instead of t in the second limit in (7.25) with the help of uniformity shows $|\varphi_{X_n}(t)| \to 1$ for all t. Using the first limit this again implies X is constant a.s., a contradiction.

Claim 2: $\lim_{n\to\infty}\alpha_n$ exists and belongs to $(0,\infty)$

The boundedness in Claim 1 permits us to consider two subsequential limits α and α' of $\{\alpha_n\}_{n\geqslant 1}$. Claim 1 also ensures that $\alpha, \alpha' > 0$. The uniformity of convergence in the second limit in (7.25) implies

$$\big|\varphi_X(\alpha t)\big| = \big|\varphi_Y(t)\big| = \big|\varphi_X(\alpha' t)\big| \tag{7.26}$$

and, since $\alpha, \alpha' > 0$, we get $|\varphi(at)| = |\varphi(t)|$ for $a := \alpha/\alpha' > 0$ and all $t \in \mathbb{R}$. Lemma 7.5 along with non-degeneracy of X then imply $\alpha = \alpha'$.

Claim 3: $\sup_{n\geqslant 1}|\beta_n| < \infty$

Abbreviate $Y_n := \alpha_n X_n + \beta_n$. For t small enough so that, e.g., $\inf_{n\geqslant 1}|\varphi_n(\alpha_n t)| > \frac{1}{2}$ we have uniform limits

$$\mathrm{e}^{\mathrm{i}t\beta_n} = \frac{\varphi_{Y_n}(t)}{\varphi_{X_n}(\alpha_n t)} \ \underset{n\to\infty}{\longrightarrow} \ \frac{\varphi_Y(t)}{\varphi_X(\alpha t)} \tag{7.27}$$

If $|\beta_{n_k}| \to \infty$ we take this along the sequence $t_k := \pi/\beta_{n_k}$ which obeys $t_k \to 0$. Thanks to uniformity of the convergence, the left-hand side then tends to $\mathrm{e}^{\mathrm{i}\pi} = -1$ yet the right-hand side tends to 1, a contradiction.

Claim 4: $\lim_{n\to\infty}\beta_n$ exists (in $\mathbb{R}$)

Note that Claim 3 permits consideration of subsequential limits β and β' of $\{\beta_n\}_{n\geqslant 1}$. The identity (7.27) then gives

$$\mathrm{e}^{\mathrm{i}t\beta} = \frac{\varphi_Y(t)}{\varphi_X(\alpha t)} = \mathrm{e}^{\mathrm{i}t\beta'} \tag{7.28}$$

for t in an open neighborhood of 0. This means that $\mathrm{e}^{\mathrm{i}t(\beta-\beta')} = 1$ for t small. Taking derivative at $t = 0$ we get $\beta - \beta' = 0$, proving convergence.

Summarizing, we have shown that $\alpha_n \to \alpha > 0$ and $\beta_n \to \beta$ and

$$\forall t \in \mathbb{R}\colon \quad \varphi_Y(t) = \mathrm{e}^{\mathrm{i}t\beta}\varphi_X(\alpha t) \tag{7.29}$$

It follows that $Y \overset{\text{law}}{=} \alpha X + \beta$. □

The upshot of Theorem 7.2 is that, in order to nail the limit laws of centered and normalized sums of i.i.d. sequences, it suffices to characterize the stable laws. This boils down to finding all random variables Y whose characteristic function φ_Y is such that, for each $n \geqslant 1$,

$$\forall t \in \mathbb{R}\colon\ \mathrm{e}^{-\mathrm{i}tB_n}\varphi_Y(t/A_n)^n = \varphi_Y(t) \tag{7.30}$$

holds with some $A_n > 0$ and $B_n \in \mathbb{R}$. Some important facts about φ_Y are relatively easy to derive from this identity; for instance:

Lemma 7.6 *Assume Y to be non-degenerate and such that* (7.30) *holds with some $A_n > 0$ and $B_n \in \mathbb{R}$ for each $n \geqslant 1$. Then there exists $\alpha \in (0, 2]$ such that*

$$\forall n \geqslant 1\colon\ A_n = n^{1/\alpha} \tag{7.31}$$

and

$$\exists \sigma > 0\, \forall t \in \mathbb{R}\colon\ \left|\varphi_Y(t)\right| = \mathrm{e}^{-\sigma^\alpha |t|^\alpha} \tag{7.32}$$

Moreover, if $\alpha = 2$ then Y is normal.

Finding a proof of this claim is instructive and so we relegate it to a homework assignment. That being said, proving a full characterization of stable laws along these lines is a pain. We will instead generalize the notion of stable random variables further and prove a formula for its characteristic function in the next lecture.

Further reading: Durrett, Section 3.8

8. INFINITELY DIVISIBLE LAWS

In this lecture we take the problem of characterizing limit laws associated with sums of random variables a bit further, leading us to the concept of infinitely-divisible laws. We then prove that these laws are characterized by the celebrated Lévy-Khinchin formula.

8.1 Definitions and main theorem.

Recall that Y is stable, or has a stable law, if there exists an i.i.d. sequence $\{Y_i\}_{i\geqslant 1}$ of random variables and two numerical sequences $\{A_n\}_{n\geqslant 1}$ and $\{B_n\}_{n\geqslant 1}$ with $A_n > 0$ for each $n \geqslant 1$ such that,

$$Y \overset{\text{law}}{=} \frac{Y_1 + \cdots + Y_n}{A_n} - B_n \tag{8.1}$$

holds for each $n \geqslant 1$. In Theorem 7.2 we then showed that the stable laws coincides with the set of weak limits of random variables of the form

$$\frac{X_1 + \cdots + X_n}{a_n} - b_n \tag{8.2}$$

for a sequence $\{X_i\}_{i\geqslant 1}$ of i.i.d. random variables and numerical sequences $\{a_n\}_{n\geqslant 1}$ and $\{b_n\}_{n\geqslant 1}$ with a_n positive for all $n \geqslant 1$.

Two remarks are in order. First, the characterization entails that the weak limit of (8.2) actually exists. In particular, one cannot just assume tightness of random variables (8.2) and try to apply the conclusion to subsequential limits. Second, it is important that the *same* sequence $\{X_i\}_{i\geqslant 1}$ is used for each n. It is this requirement that we will now try to relax. Consider the following definition:

Definition 8.1 *A random variable Y is said to be infinitely divisible if for each $n \geqslant 1$ there exist i.i.d. random variables $Y_{n,1}, \ldots, Y_{n,n}$ such that*

$$Y \overset{\text{law}}{=} Y_{n,1} + \cdots + Y_{n,n} \tag{8.3}$$

It is immediate from (8.1) that every stable random variable is infinitely divisible. (This includes normal random variables as a special case.) But there are many random variables that are infinitely divisible and not stable. For instance, since the sum of independent Poisson(λ) and Poisson(μ) is Poisson($\lambda + \mu$), Poisson random variables are infinitely divisible. Other examples include Gamma random variables, log-normal (i.e., exponentials of a normal) random variables, χ^2-distribution, etc.

Here is the analogue of Theorem 7.2 for infinitely-divisible laws:

Theorem 8.2 *A random variable Y is infinitely divisible if and only if there exist random variables $\{X_{n,j}\}_{1\leqslant j\leqslant n}$ such that*

$$\forall n \geqslant 1\colon \ X_{n,1}, \ldots, X_{n,n} \text{ are i.i.d.} \tag{8.4}$$

and

$$X_{n,1} + \cdots + X_{n,n} \xrightarrow[n\to\infty]{\text{w}} Y \tag{8.5}$$

(The law of $X_{n,1}$ can change with n.)

Proof. The proof is similar to that of Theorem 7.2. That each infinitely divisible Y is a limit of sums of i.i.d. random variables follows directly from (8.3), so our main job is to prove the converse. Assume $\{X_{n,j}\}_{1\leqslant j\leqslant n}$ are random variables such that (8.4–8.5) hold. Abbreviate

$$S_n := X_{n,1} + \cdots + X_{n,n} \tag{8.6}$$

For any $k \geqslant 1$ we then have

$$S_{nk} = Z_{n,1} + \cdots + Z_{n,k} \tag{8.7}$$

where

$$Z_{n,j} := \sum_{i=(j-1)n+1}^{jn} X_{n,i} \tag{8.8}$$

Note that $Z_{n,1}, \ldots, Z_{n,k}$ are i.i.d.

While $S_{nk} \xrightarrow{\text{w}} Y$, unlike Theorem 7.2, here we cannot immediately conclude that the random variables $Z_{n,j}$ converge as $n \to \infty$. However, for each $t > 0$ we get

$$P(S_{kn} > t) \geqslant P\bigg(\bigcap_{j=1}^{k}\{Z_{n,j} > t\}\bigg) = P\big(Z_{n,1} > t\big)^k \tag{8.9}$$

The weak convergence of $\{S_n\}_{n\geqslant 1}$ implies tightness which means that, given any $\epsilon > 0$, the quantity on the left is less than ϵ^k once t is sufficiently large, uniformly in $n \geqslant 1$. But then $P(Z_{n,1}^{(k)} > t) < \epsilon$ and, using a similar argument, also $P(Z_{n,1}^{(k)} < t) < \epsilon$ for all $n \geqslant 1$ once t is large. We conclude that $\{Z_{n,1}\}_{n\geqslant 1}$ is tight. The Helly selection theorem (Theorem 3.6) implies existence of a subsequence $n_j \to \infty$ such that $Z_{n_j,1} \xrightarrow{\text{w}} Z$. But then, as shown by an argument based on characteristic functions, (8.2) holds with $Y_{n,1}, \ldots, Y_{n,n}$ i.i.d. copies of Z. It follows that Y is infinitely divisible. □

To demonstrate the difference between Theorems 7.2 and Theorem 8.2 we reiterate that while all stable laws arise from suitably centered and scaled sums of terms in one i.i.d. sequence

$$X_1, X_2, X_3, \ldots \tag{8.10}$$

all infinitely-divisible laws arise as weak limits of the laws of row sums of a triangular array of the form

$$\begin{aligned} &X_{1,1} \\ &X_{2,1}, X_{2,2} \\ &X_{3,1}, X_{3,2}, X_{3,3} \\ &X_{4,1}, X_{4,2}, X_{4,3}, X_{4,4} \end{aligned} \tag{8.11}$$

where the random variables in each row are assumed to be i.i.d. (Different rows may not even be defined on the same probability space.)

With the above settled, we now give an analytic form of the characteristic function of any infinitely divisible law:

Theorem 8.3 (Lévy-Khinchin formula) *A random variable X is infinitely divisible if and only if there exist $\mu \in \mathbb{R}$, $\sigma^2 \in [0,\infty)$ and a finite Borel measure ν on $\mathbb{R}$ with $\nu(\{0\}) = 0$ such that for each $t \in \mathbb{R}$,*

$$E(e^{itX}) = \exp\left\{ it\mu - \frac{\sigma^2}{2}t^2 + \int \Big(e^{itx} - 1 - \frac{itx}{1+x^2} \Big) \frac{1+x^2}{x^2} \nu(dx) \right\} \tag{8.12}$$

In particular, the exponential is a characteristic function for any choice of μ, σ^2 and ν as above. Moreover, the triplet (μ, σ, ν) with above properties is determined uniquely by (8.12).

A couple of remarks are in order. First, the function under the integral is bounded and so the integral exists absolutely. Second, the measure

$$\lambda(dx) := \frac{1+x^2}{x^2} \nu(dx) \tag{8.13}$$

appearing under the integral is standardly called the *Lévy measure*. The conditions on ν are equivalent phrased by requiring that $\lambda(\{0\}) = 0$ and

$$\int \frac{x^2}{1+x^2} \, \lambda(dx) < \infty. \tag{8.14}$$

We will see that the Lévy measure naturally arises in the proof.

Third, the first two terms in the exponent in (8.12) correspond to the characteristic function $e^{i\mu t - \frac{1}{2}\sigma^2 t^2}$ of $\mathcal{N}(\mu, \sigma^2)$ so each infinitely divisible is the sum of an independent normal and a random variable whose characteristic function is exponent of the integral — the "Lévy part." As we will see in the proof of the "if" part of the statement, this part corresponds to a *compound Poisson law*. Incidentally, as we will in the proof of the "only if" part, the term $\frac{1}{2}\sigma^2 t^2$ can be reabsorbed to the integral by adding $\sigma^2 \delta_0$ to ν. This is because the integrand tends to $-t^2/2$ as $x \to 0$.

8.2 Manipulations with characteristic functions.

The proof of the "only if" part of Theorem 8.3 is based mainly on manipulations with characteristic functions and (non-probability) measures on $\mathbb{R}$ although we often make a step back to probability when that is more convenient for the argument. We start by recalling two facts about characteristic functions, the first of which has already been noted in (3.13) as a tool to prove tightness in Lévy's continuity theorem (Theorem 3.11):

Lemma 8.4 *There exists $c \in (0,\infty)$ such that for all $R > 0$ and all random variables Z with characteristic function $\varphi_Z(t) := E(e^{itZ})$,*

$$P(|Z| > R) \leqslant cR \int_0^{1/R} \big[1 - \operatorname{Re} \varphi_Z(t) \big] dt \tag{8.15}$$

Proof. Noting that $\operatorname{Re}\varphi_Z(t) = E(\cos(tZ)$ and $\sin(u)/u \leqslant 1$ holds for all $u \in \mathbb{R}$, we get

$$\begin{aligned}\frac{1}{u}\int_0^u E\big(1-\cos(uZ)\big)\,\mathrm{d}u &= E\Big(1-\frac{\sin(uZ)}{uZ}\Big)\\ &\geqslant E\bigg(1_{\{u|Z|\geqslant 1\}}\Big(1-\frac{\sin(uZ)}{uZ}\Big)\bigg) \geqslant \bigg[\inf_{s\geqslant 1}\Big(1-\frac{\sin s}{s}\Big)\bigg]\,P(u|Z|\geqslant 1)\end{aligned} \tag{8.16}$$

Writing c^{-1} for the square bracket, the claim follows by setting $u := 1/R$. □

We will also need:

Lemma 8.5 *Suppose Z is a random variable with $\operatorname{Re}\varphi_Z(t) = 1 + o(t^2)$ as $t \to 0$; i.e.,*

$$\lim_{t\to 0}\frac{1-\operatorname{Re}\varphi_Z(t)}{t^2} = 0 \tag{8.17}$$

Then $Z = 0$ a.s.

Proof. Using Fatou's lemma we have

$$\lim_{t\to 0}\frac{1-\operatorname{Re}\varphi_Z(t)}{t^2} = \lim_{t\to 0} E\bigg(\frac{1-\cos(tZ)}{t^2}\bigg) \geqslant \frac{1}{2}E(Z^2) \tag{8.18}$$

Since the left-hand side vanishes, we have $E(Z^2) = 0$ which implies $Z = 0$ a.s. □

Suppose now that X is infinitely divisible and, for each $n \geqslant 1$, let $X_{n,1}, \dots, X_{n,n}$ be i.i.d. random variables such that

$$X \overset{\text{law}}{=} X_{n,1} + \cdots + X_{n,n} \tag{8.19}$$

Set

$$\varphi(t) := E\mathrm{e}^{\mathrm{i}tX} \quad \text{and} \quad \varphi_n(t) := E\mathrm{e}^{\mathrm{i}tX_{n,1}} \tag{8.20}$$

and note that (8.19) translates into

$$\forall t \in \mathbb{R}\colon\ \varphi(t) = \varphi_n(t)^n \tag{8.21}$$

We now move to a key technical observation:

Lemma 8.6 *We have $X_{n,1} \xrightarrow[n\to\infty]{P} 0$ and so $\varphi_n(t) \xrightarrow[n\to\infty]{} 1$ locally uniformly in $t \in \mathbb{R}$.*

Proof. Continuity of φ along with $\varphi(0) = 1$ ensure existence of a number $\delta > 0$ such that $\operatorname{Re}\varphi(t) \geqslant 1/2$ for all $t \in [-\delta, \delta]$. We claim that then

$$\forall t \in [-\delta,\delta]\colon \quad \operatorname{Re}\varphi_n(t) > \big|\varphi_n(t)\big|\cos\big(\tfrac{\pi}{2n}\big) \tag{8.22}$$

Indeed, suppose for the sake of contradiction that this fails at some $t_\star \in [-\delta, \delta]$. Using continuity we may assume that $t_\star$ has the smallest absolute among all points with this property. But then, again by continuity, equality holds at $t_\star$. This gives

$$\varphi_n(t_\star) = \big|\varphi_n(t_\star)\big|\,\mathrm{e}^{\pm\mathrm{i}\frac{\pi}{2n}\mathrm{i}} \tag{8.23}$$

But then (8.21) implies $\varphi(t_\star) = |\varphi_n(t_\star)|^n \mathrm{e}^{\pm\frac{\pi}{2}\mathrm{i}} = \pm\mathrm{i}|\varphi_n(t_\star)|^n$ which is purely imaginary, in contradiction with $\operatorname{Re}\varphi(t_\star) \geqslant 1/2$.

With (8.22) in hand, the fact that $|\varphi_n(t)| = |\varphi(t)|^{1/n}$ gives

$$\operatorname{Re}\varphi_n(t) \geqslant 2^{-1/n}\cos\big(\tfrac{\pi}{2n}\big) \underset{n\to\infty}{\longrightarrow} 1 \tag{8.24}$$

for all $t \in [-\delta, \delta]$. Invoking Lemma 7.5(2) we get $X_{n,1} \to 0$ in probability and, by Lemma 7.4, also $\varphi_n(t) \to 1$ locally uniformly in $t \in \mathbb{R}$. □

Next we establish a number of important consequences of the previous lemma. The proof of these involves non-trivial facts from complex analysis.

Corollary 8.7 *The following holds:*

(1) $\varphi(t) \neq 0$ *and* $\varphi_n(t) \neq 0$ *for every* $t \in \mathbb{R}$ *and* $n \geqslant 1$

(2) *there exist continuous functions* $t \mapsto \log\varphi(t)$ *and, for each* $n \geqslant 1$, *also* $t \mapsto \log\varphi_n(t)$ *such that, for all* $n \geqslant 1$ *and* $t \in \mathbb{R}$,

$$\log\varphi(t) = n\log\varphi_n(t) \tag{8.25}$$

and

$$\varphi(t) = \exp\{\log\varphi(t)\} \quad \text{and} \quad \varphi_n(t) = \exp\{\log\varphi_n(t)\} \tag{8.26}$$

where $\exp(z) := \sum_{k\geqslant 1}\frac{1}{k!}z^k$ *is the complex exponential*

(3) *we have*

$$n\big(\varphi_n(t) - 1\big) \underset{n\to\infty}{\longrightarrow} \log\varphi(t) \tag{8.27}$$

locally uniformly in $t \in \mathbb{R}$, *and*

(4) *there exists a continuous function* $t \mapsto \epsilon_n(t)$ *such that for all* $t \in \mathbb{R}$

$$\varphi(t) = \exp\Big\{\big(1 + \epsilon_n(t)\big)n\big(\varphi_n(t) - 1\big)\Big\} \tag{8.28}$$

with $\epsilon_n(t) \to 0$ *as* $n \to \infty$ *locally uniformly in* t.

Proof. (1) If $\varphi(t) = 0$ or $\varphi_m(t) = 0$ for some $t \in \mathbb{R}$ and $m \geqslant 1$, then $\varphi_n(t) = 0$ for all $n \geqslant 1$ contradicting that $\varphi_n(t) \to 1$.

(2) By part (1), the image of $t \mapsto \varphi(t)$ is a continuous curve in $\mathbb{C} \smallsetminus \{0\}$. Let $\Gamma(t)$ be the portion of this curve for the argument ranging from 0 to t and let

$$\log\varphi(t) := \int_{\Gamma(t)}\frac{\mathrm{d}z}{z} \tag{8.29}$$

Here the (Stieltjes or curve) integral exists regardless of whether $\Gamma(t)$ is rectifiable or not thanks to the fact that, by Cauchy's theorem in complex analysis, any piece-wise linear approximation to $\Gamma(t)$ that stays close enough to $\Gamma(t)$ has the same value of the integral. This also guarantees that, at $t > t_0 \geqslant 0$ such that $|\varphi(u) - \varphi(t_0)| < |\varphi(t_0)|$ for all $u \in [t_0, t]$, it suffices to compute the integral by going along any path from $\varphi(t_0)$ to $\varphi(t)$ that does not leave the open disc of radius $|\varphi(t_0)|$ centered at $\varphi(t_0)$.

Writing $\varphi(t) = \varphi(t_0)(1 + r\mathrm{e}^{\mathrm{i}\theta})$ for $r \in [0, 1)$ and $\theta \in [0, 2\pi)$, we choose to go along the linear segment parametrized by $z(u) := \varphi(t_0)(1 + u\mathrm{e}^{\mathrm{i}\theta})$ for $u \in [0, r]$. This gives

$$\log\varphi(t) - \log\varphi(t_0) = \int_0^r \frac{\mathrm{e}^{\mathrm{i}\theta}\mathrm{d}u}{1 + u\mathrm{e}^{\mathrm{i}\theta}} \tag{8.30}$$

Using that $r < 1$, we can expand the function in the integral into a power series in u and integrate term-by-term to get

$$\log \varphi(t) - \log \varphi(t_0) = \sum_{n=0}^{\infty} \frac{(-re^{i\theta})^{n+1}}{n+1} \tag{8.31}$$

The right-hand side is now recognized as the power series representing $z \mapsto \log(1-z)$ whose exponential thus equals $1-z$. Plugging $z := -re^{i\theta}$ yields

$$\exp\{\log \varphi(t)\} = \exp\{\log \varphi(t_0)\}(1 + re^{i\theta}) = \exp\{\log \varphi(t_0)\}\frac{\varphi(t)}{\varphi(t_0)} \tag{8.32}$$

Proceeding along the curve $t \mapsto \varphi(t)$ we thus verify (8.26) for φ. The proof for φ_n is completely analogous.

Writing $\Gamma_n(t)$ for the curve defined by $s \mapsto \varphi_n(s)$ for s ranging from 0 to t, we define $\log \varphi_n(t)$ similarly as (8.29). Using that the substitution $w := z^n$ takes $\Gamma_n(t)$ to $\Gamma(t)$, we have

$$\log \varphi(t) = \int_{\Gamma(t)} \frac{dw}{w} \overset{w:=z^n}{=} \int_{\Gamma_n(t)} \frac{z^{n-1}dz}{z^n} = n \log \varphi_n(t) \tag{8.33}$$

proving (8.25).

(3-4) Let g be the function on $\{z \in \mathbb{C} \colon |z| < 1\}$ defined by $g(0) := 0$ and

$$g(z) := \frac{\log(1+z) - z}{z}, \quad z \neq 0 \tag{8.34}$$

Note that g is holomorphic with $g(z) = O(|z|)$ as $z \to 0$. Define

$$\epsilon_n(t) := \begin{cases} g\big(\varphi_n(t) - 1\big), & \text{if } |\varphi(t) - 1| \leqslant 1/2 \\ \dfrac{\log \varphi_n(t)}{\varphi_n(t) - 1} - 1, & \text{else} \end{cases} \tag{8.35}$$

Since both formulas give the same when $0 < |\varphi(t) - 1| < 1$, the function ϵ_n is continuous with $\epsilon_n(t) = O(|\varphi_n(t) - 1|)$ when $|\varphi_n(t) - 1|$ is small. Expressing the logarithm of $\varphi_n(t)$ from these formulas while invoking (8.25) shows

$$\log \varphi(t) = n \log \varphi_n(t) = \big(1 + \epsilon_n(t)\big)n\big(\varphi_n(t) - 1\big) \tag{8.36}$$

which gives (8.28) with the help of (8.26). Since $\varphi_n(t) \to 0$ and thus also $\epsilon_n(t) \to 0$ locally uniformly in t, we then also get (8.27). □

8.3 Proof of the Lévy-Khinchin formula.

We are now ready to commence the proof of the "only if" part of Theorem 8.3; specifically, the claim that every infinitely divisible law has characteristic function (8.12). Following similar arguments as for stable laws, define a Borel measure λ_n by

$$\lambda_n(A) := nP(X_{n,1} \in A) \tag{8.37}$$

Then

$$n\big(\varphi_n(t) - 1\big) = \int (e^{itx} - 1)\lambda_n(dx) \tag{8.38}$$

Next observe:

Lemma 8.8 *There is $c < \infty$ such that for each $R > 0$,*

$$\limsup_{n\to\infty} \lambda_n\big((-R,R)^{\mathrm{c}}\big) \leqslant cR \int_0^{1/R} \log \frac{1}{|\varphi(t)|}\mathrm{d}t \tag{8.39}$$

Proof. The argument in the proof of Lemma 8.4 gives

$$\lambda_n\big((-R,R)^{\mathrm{c}}\big) \leqslant cR \int_0^{1/R} n\big(1 - \operatorname{Re}\varphi_n(t)\big)\mathrm{d}t \tag{8.40}$$

By Corollary 8.7 the integrand is continuous and tends to

$$-\operatorname{Re}\log\varphi(t) = \log\frac{1}{|\varphi(t)|} \tag{8.41}$$

uniformly on compact sets, and so the integral converges to that in the statement. □

The reader may be surprised by the above lemma for the following reason. The total mass of λ_n is $\lambda_n(\mathbb{R}) = n$ so if the mass outside any compact interval stays bounded, where did the rest of the mass go? The answer is that it has "collapsed" to zero. To control the amount of mass near zero, we also need:

Lemma 8.9 (Mass near zero)

$$\limsup_{n\to\infty} \int_{[-1,1]} x^2\,\lambda_n(\mathrm{d}x) < \infty \tag{8.42}$$

Proof. Noting that $c' := \inf_{|u|\leqslant 1} \frac{1-\cos(u)}{u^2} > 0$ we have

$$\begin{aligned} n\big(1 - \operatorname{Re}\varphi_n(1)\big) &= nE\big(1-\cos(X_{n,1})\big) \\ &\geqslant nE\big(1_{\{|X_{n,1}|\leqslant 1\}}\big(1-\cos(X_{n,1})\big)\big) \\ &\geqslant nE\big(c'X_{n,1}^2 1_{\{|X_{n,1}|\leqslant 1\}}\big) = c'\int_{[-1,1]} x^2\,\lambda_n(\mathrm{d}x) \end{aligned} \tag{8.43}$$

By Corollary 8.7, the left-hand side converges to $\log 1/|\varphi(1)|$. □

We are now ready for:

Proof of "only if" in Theorem 8.3. Define a Borel measure ν_n by

$$\nu_n(\mathrm{d}x) := \frac{x^2}{1+x^2}\,\lambda_n(\mathrm{d}x) \tag{8.44}$$

Lemmas 8.8 and 8.9 show that $\{\nu_n\}_{n\geqslant 1}$ is tight on $\mathbb{R}$ with $\sup_{n\geqslant 1}\nu_n(\mathbb{R}) < \infty$ so, by Helly's selection theorem, there exists a sequence $n_k \to \infty$ and a finite Borel measure $\tilde{\nu}$ on $\mathbb{R}$ such that $\nu_{n_k} \xrightarrow{\mathrm{w}} \tilde{\nu}$. Write

$$\begin{aligned} n\big(\varphi_n(t) - 1\big) &= \int (\mathrm{e}^{\mathrm{i}tx} - 1)\frac{1+x^2}{x^2}\nu_n(\mathrm{d}x) \\ &= \int \Big(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{1+x^2}\Big)\frac{1+x^2}{x^2}\,\nu_n(\mathrm{d}x) + \mathrm{i}t\int \frac{x}{1+x^2}\,\lambda_n(\mathrm{d}x) \end{aligned} \tag{8.45}$$

Noting that the first integrand is bounded and continuous (including $x = 0$ where it is interpreted in a limit sense), the first integral converges to that with respect to $\tilde{\nu}$. Since the left-hand side converges to $\log \varphi(t)$ by Corollary 8.7, the limit in

$$\mu := \lim_{k\to\infty} \int \frac{x}{1+x^2}\, \lambda_{n_k}(\mathrm{d}x) \tag{8.46}$$

must exists and we have

$$\log \varphi(t) = \mathrm{i}t\mu + \int \Bigl(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{1+x^2}\Bigr) \frac{1+x^2}{x^2}\, \tilde{\nu}(\mathrm{d}x) \tag{8.47}$$

Note that $\sigma^2 := \tilde{\nu}(\{0\})$ may be non-zero due to "collapse of mass to 0" in which case we interpret the integrand by its limit value $-\frac{1}{2}t^2$ at $x = 0$. Letting

$$\nu := \tilde{\nu} - \sigma^2 \delta_0 \tag{8.48}$$

gives

$$\log \varphi(t) = \mathrm{i}t\mu - \frac{1}{2}t^2\sigma^2 + \int \Bigl(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{1+x^2}\Bigr) \frac{1+x^2}{x^2} \nu(\mathrm{d}x) \tag{8.49}$$

where $\nu(\{0\}) = 0$. Plugging this in (8.26), we get the Lévy-Khinchin formula. □

In order to finish the proof of Theorem 8.3 we have to prove the converse and also show uniqueness of the triplet (μ, σ^2, ν). This will be done in the next lecture.

***Remark* 8.10** An interesting question that springs to mind is what happens if we only assume subsequential convergence of sums of i.i.d. random variables? This is equivalent to asking: What are all possible laws of Y such that, for a family of random variables

$$\bigl\{X_{n,j} \colon n \geqslant 1,\, j = 1, \dots, k(n)\bigr\} \tag{8.50}$$

with $k(n) \to \infty$ and

$$\forall n \geqslant 1 \colon \;\; X_{n,1}, \dots, X_{n,k(n)} \text{ are i.i.d.} \tag{8.51}$$

we have

$$X_{n,1} + \cdots + X_{n,k(n)} \;\xrightarrow[n\to\infty]{\mathrm{w}}\; Y \tag{8.52}$$

As it turns out, the answer is simple: Y still has to be infinitely divisible, but the proof requires going via the arguments for Theorem 8.3.

Indeed, denoting $\varphi_n(t) := E\mathrm{e}^{\mathrm{i}tX_{n,1}}$, we first note that, as in Lemma 8.6, thanks to $k(n) \to \infty$ we must have

$$X_{n,1} \;\xrightarrow[n\to\infty]{P}\; 0 \text{ and so } \varphi_n(t) \;\xrightarrow[n\to\infty]{}\; 1 \text{ locally uniformly in } t \in \mathbb{R} \tag{8.53}$$

Since $\varphi_n(t)^{k(n)} \to \varphi(t)$, this implies $\varphi(t) \neq 0$ for all $t \in \mathbb{R}$ and so, by the argument in Corollary 8.7, $t \mapsto \log \varphi(t)$ is well defined and

$$k(n)\bigl(\varphi_n(t) - 1\bigr) \;\xrightarrow[n\to\infty]{}\; \log \varphi(t) \tag{8.54}$$

locally uniformly in $t \in \mathbb{R}$. Checking the proof of the "only if" part of Theorem 8.3, this is all what we ever needed there.

8.4 Proof of "if" part of Theorem 8.3.

We now move to the proof of the converse; namely, the statement that a random variable with characteristic function in (8.12) is infinitely divisible. The bulk of the argument boils down to proving that the formula in (8.12) is actually a characteristic function for any choice of the "parameters" μ, σ^2 and ν. The key step for this comes in:

Lemma 8.11 *For all finite Borel measures ν on $\mathbb{R}$, there exists a random variable Z such that*

$$Ee^{itZ} = \left\{ \int \left(e^{itx} - 1 - \frac{itx}{1+x^2} \right) \frac{1+x^2}{x^2} \nu(dx) \right\} \tag{8.55}$$

holds for all $t \in \mathbb{R}$.

Proof. If $\nu = 0$ then we set $Z := 0$ and so let us assume that ν is non-trivial in the sequel. Define a Borel measure λ by

$$\lambda(dx) := \frac{1+x^2}{x^2} \nu(dx) \tag{8.56}$$

and, for $\epsilon > 0$, let

$$\lambda_\epsilon(dx) := 1_{[-\epsilon,\epsilon]^c}(x) \lambda(dx) \tag{8.57}$$

be its restriction to the complement of $(-\epsilon, \epsilon)$. Define also

$$\mu_\epsilon := \int \frac{x}{1+x^2} \, \lambda_\epsilon(dx) \tag{8.58}$$

where the integral exists since $\lambda_\epsilon(\mathbb{R}) < \infty$.

Next, assuming $\epsilon > 0$ so small that $\lambda_\epsilon(\mathbb{R}) > 0$, let $Y_1, Y_2, \ldots$ be i.i.d. random variables with law determined by

$$P(Y_1 \in A) := \frac{\lambda_\epsilon(A)}{\lambda_\epsilon(\mathbb{R})} \tag{8.59}$$

and let N_ϵ be Poisson $(\lambda_\epsilon(\mathbb{R}))$, independent of the Y_i's. Denote

$$Z_\epsilon := -\mu_\epsilon + \sum_{j=1}^{N_\epsilon} Y_j \tag{8.60}$$

In order to compute the characteristic function of Z_ϵ, write $\varphi_\epsilon(t) := E(e^{itY_1})$. Then

$$\begin{aligned} E(e^{itZ_\epsilon}) &= e^{-it\mu_\epsilon} E\Big(\big[\varphi_\epsilon(t)\big]^{N_\epsilon} \Big) \\ &= e^{-it\mu_\epsilon} \sum_{n=0}^{\infty} \frac{\lambda_\epsilon(\mathbb{R})^n}{n!} e^{-\lambda_\epsilon(\mathbb{R})} \varphi_\epsilon(t)^n \\ &= \exp\Big\{ \lambda_\epsilon(\mathbb{R}) \big(\varphi_\epsilon(t) - 1 \big) - it\mu_\epsilon \Big\} \end{aligned} \tag{8.61}$$

Noting that

$$\lambda_\epsilon(\mathbb{R}) \big(\varphi_\epsilon(t) - 1 \big) = \int (e^{itx} - 1) \lambda_\epsilon(dx) \tag{8.62}$$

we conclude that

$$E(\mathrm{e}^{\mathrm{i}tZ_\epsilon}) = \exp\{I_\epsilon(t)\} \tag{8.63}$$

where

$$I_\epsilon(t) := \int_{[-\epsilon,\epsilon]^c} \Big(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{1+x^2}\Big)\,\frac{1+x^2}{x^2}\nu(\mathrm{d}x) \tag{8.64}$$

The integrand is bounded uniformly in $\epsilon > 0$ and so, since $\nu(\{0\}) = 0$, we get

$$I_\epsilon(t) \underset{\epsilon\downarrow 0}{\longrightarrow} I_0(t) := \int_{\{0\}^c} \Big(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{1+x^2}\Big)\frac{1+x^2}{x^2}\nu(\mathrm{d}x) \tag{8.65}$$

by the Bounded Convergence Theorem. The Lévy continuity theorem then implies $Z_\epsilon \xrightarrow{\mathrm{w}} Z'$ where $E\mathrm{e}^{\mathrm{i}tZ'} = \exp\{I_0(t)\}$. Taking $Z := Z' + X$ where $X = \mathcal{N}(0, \nu(\{0\}))$ is independent of Z' gives a random variable satisfying (8.55). □

We are now ready to give:

Proof of "if" part of Theorem 8.3. Let μ, σ^2 and ν be as in the statement. In the previous lemma we constructed a random variable Z with characteristic function (8.55). Adding to Z an independent copy of $\mathcal{N}(\mu, \sigma^2)$ results in a random variable X with characteristic function in (8.12). Note that, given any $k \geqslant 1$, the same construction with μ, σ^2 and ν divided by k produces a random variable Y such that

$$\forall t \in \mathbb{R}\colon\ E\mathrm{e}^{tX} = \big(E\mathrm{e}^{tY}\big)^k \tag{8.66}$$

Since characteristic function determines the law, we get

$$X \overset{\text{law}}{=} Y_1 + \cdots + Y_k \tag{8.67}$$

for $Y_1, \ldots, Y_k$ independent copies of Y. It follows that X is infinitely divisible, proving the "if" part of the statement. □

It remains to prove uniqueness of (μ, σ^2, ν) subject to the condition $\nu(\{0\}) = 0$, which we leave to a homework assignment. The same applies to the proof of the following special case of Theorem 8.3:

Theorem 8.12 (Kolmogorov's theorem) *Suppose $X \in L^2$ and $EX = 0$. Then X is infinitely divisible if and only if there is a finite Borel measure μ such that*

$$E(\mathrm{e}^{\mathrm{i}tX}) = \exp\bigg\{\int \frac{\mathrm{e}^{\mathrm{i}tx} - 1 - \mathrm{i}tx}{x^2}\mu(\mathrm{d}x)\bigg\} \tag{8.68}$$

Moreover, $\mu(\mathbb{R}) = \mathrm{Var}(X)$.

Further reading: Durrett, Section 3.8 and references therein

9. Poisson convergence and processes

The construction underlying the proof of the "if" part of Theorem 8.3 brings us to a connection with Poisson random variables and, specifically, the notion of the Poisson point process. This is the subject we will explore in this lecture.

9.1 Poisson convergence.

As the reader likely knows from undergraduate probability courses, Poisson random variables arise as limits of Binomial random variables. Here is a precise statement of this fact which goes back to S.D. Poisson's work from 1837 (although, per wiki page, A. de Moivre already published it in 1711).

Lemma 9.1 *For each $n \geqslant 1$, let $X_{n,1}, \dots, X_{n,n}$ be i.i.d. $\{0,1\}$-valued random variables with $P(X_{n,i} = 1) = \lambda_n$. Let $S_n := X_{n,1} + \cdots + X_{n,n}$ and suppose that $n\lambda_n \to \lambda \in (0, \infty)$. Then*

$$\forall k \geqslant 0\colon \;\; P(S_n = k) \;\underset{n\to\infty}{\longrightarrow}\; \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} \tag{9.1}$$

Proof. Since S_n is Binomial with parameters n and λ_n, for $0 \leqslant k \leqslant n$ we have

$$\begin{aligned} P(S_n = k) &= \binom{n}{k}\lambda_n^k(1-\lambda_n)^{n-k} \\ &= \frac{1}{k!}\frac{n(n-1)\dots(n-k+1)}{n^k}(n\lambda_n)^k\left(1-\frac{n\lambda_n}{n}\right)^{n-k} \end{aligned} \tag{9.2}$$

For k fixed and $n \to \infty$ we have $(n\lambda_n)^k \to \lambda^k$ and $(1 - \frac{n\lambda_n}{n})^{n-k} \to \mathrm{e}^{-\lambda}$. Since the large fraction tends to 1, the claim follows. □

Definition 9.2 *Let $\lambda \geqslant 0$. A random variable N is said to be Poisson with parameter λ of N takes values in non-negative integers and $P(N = k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$ holds for all $k \geqslant 0$.*

Note that while the convergence (9.1) is phrased for each k, it is actually uniform. Indeed, (9.2) gives

$$\sum_{k\geqslant k_0} P(S_n = k) \leqslant \sum_{k\geqslant k_0}\frac{(n\lambda_n)^k}{k!}\mathrm{e}^{-n\lambda_n} \leqslant \frac{(n\lambda_n)^{k_0}}{k_0!} \leqslant \frac{1}{k_0!}[\sup_{n\geqslant 1} n\lambda_n]^{k_0} \tag{9.3}$$

where we used that $1 - a \leqslant \mathrm{e}^{-a}$ and $k! \geqslant k_0!(k-k_0)!$. Now observe that right-hand side can be made as small as desired by taking k_0 large.

As it turns out, even a stronger estimate holds:

Theorem 9.3 (Le Cam, 1960) *Let $n \geqslant 1$ and let $X_1, \dots, X_n$ be independent $\{0,1\}$-valued random variables. Set $S_n = X_1 + \cdots + X_n$ and denote $p_i := P(X_i = 1)$. Abbreviate $\lambda := p_1 + \cdots + p_n$. Then*

$$\sum_{k\geqslant 0}\left|P(S_n = k) - \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}\right| \leqslant 2\sum_{i=1}^n p_i^2 \leqslant 2\lambda \max_{i=1,\dots,n} p_i \tag{9.4}$$

The proof uses a number of small observations that are of independent interest. We start by noting that the total variational distance between two random variables XS and Y taking values in a measurable space $(\mathscr{X}, \mathcal{G})$ can be written as

$$d_{\mathrm{TV}}(X, Y) = \sup_{A \in \mathcal{G}} \big| P(X \in A) - P(Y \in A) \big| \tag{9.5}$$

For X and Y taking values in $\mathbb{N}$ this simplifies as

$$\begin{aligned} d_{\mathrm{TV}}(X, Y) &= \sup_{A \subseteq \mathbb{N}} \left| \sum_{k \in A} \big[P(X = k) - P(Y = k) \big] \right| \\ &= \sum_{k \in \mathbb{N}} \big(P(X = k) - P(Y = k) \big)^+ = \sum_{k \in \mathbb{N}} \big(P(X = k) - P(Y = k) \big)^- \\ &= \frac{1}{2} \sum_{k \in \mathbb{N}} \big| P(X = k) - P(Y = k) \big| \end{aligned} \tag{9.6}$$

Here we observed that the sum over $k \in A$ can be written as the sum over k where $P(X \in A) - P(Y \in A) \geqslant 0$ and the sum over where $P(X \in A) - P(Y \in A) \leqslant 0$. The latter is negative so dropping one of the two would only increase the result. It follows that an optimal A is $A := \{k \in \mathbb{N} \colon P(X = k) \geqslant P(Y = k)\}$ as well as the complement thereof. This gives the second line. The third line is obtained by noting that $|a| = a^+ + a^-$.

In summary, the total variational distance of two discrete-valued random variables (over the same set of values) is half of the ℓ^1-distance of their probability mass functions. Next we observe:

Lemma 9.4 *For each $\lambda \geqslant 0$, Poisson(λ)-random variable is infinitely divisible. In fact,*

$$\forall \lambda, \mu \geqslant 0 \colon \quad \mathrm{Poisson}(\lambda + \mu) \overset{\text{law}}{=} \mathrm{Poisson}(\lambda) \oplus \mathrm{Poisson}(\mu) \tag{9.7}$$

where "$\oplus$" denotes sum of independent random variables.

Proof. Infinite divisibility is immediate from the distributional convergence in Lemma 9.1 and the argument in (7.16). Alternatively, we just prove (9.7) which is done by writing explicitly the probability mass function of the quantity on the right and reducing it using the Binomial Theorem. □

Let $Y_1, \dots, Y_n$ be independent with $Y_i = \mathrm{Poisson}(p_i)$. Then $Y_1 + \cdots + Y_n$ is Poisson(λ) and, using (9.6) and (9.7), Le Cam's inequality (9.4) can be phrased as

$$d_{\mathrm{TV}}\bigg(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i \bigg) \leqslant \sum_{i=1}^n p_i^2 \tag{9.8}$$

In order to prove this, we first show:

Lemma 9.5 *Suppose $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ are independent (normed-vector-space-valued) random variables. Then*

$$d_{\mathrm{TV}}\bigg(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i \bigg) \leqslant \sum_{i=1}^n d_{\mathrm{TV}}(X_i, Y_i) \tag{9.9}$$

Proof. We will proceed by induction but for that we need to make the following observation. Let X, Y and Z be independent random variables taking values in a linear vector space $\mathscr{X}$ endowed with a σ-algebra $\mathcal{G}$. Then for each $A \in \mathcal{G}$,

$$\begin{aligned} P(X+Z \in A) - P(Y+Z \in A) & \\ &\leqslant \int \mu_Z(\mathrm{d}z) \big[P(X \in -z+A) - P(Y \in -z+A) \big] \\ &\leqslant d_{\mathrm{TV}}(X,Y) \end{aligned} \tag{9.10}$$

where we used independence to write the probabilities as integrals with respect to distribution μ_Z of Z and then applied (9.5) to the integrand. A similar argument bounds the quantity by $-d_{\mathrm{TV}}(X,Y)$ from below and so we get

$$d_{\mathrm{TV}}(X+Z, Y+Z) \leqslant d_{\mathrm{TV}}(X,Y) \tag{9.11}$$

whenever X, Y and Z are independent.

Consider now the setting of the lemma and, for each $j = 0, \dots, n$, denote

$$Z_j := \sum_{i=1}^{j} Y_i + \sum_{i=j+1}^{n} X_i \tag{9.12}$$

Then $\sum_{j=1}^{n} X_j = Z_0$ and $\sum_{j=1}^{n} Y_j = Z_n$ and so, by the triangle inequality for the total-variational distance

$$d_{\mathrm{TV}}\Big(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i \Big) = d_{\mathrm{TV}}(Z_0, Z_n) \leqslant \sum_{j=1}^{n} d_{\mathrm{TV}}\big(Z_{j-1}, Z_j\big) \tag{9.13}$$

But writing $\widetilde{Z}_j := \sum_{i=1}^{j-1} Y_i + \sum_{i=j+1}^{n} X_i$, we have $Z_{j-1} = X_j + \widetilde{Z}_j$ and $Z_j = Y_j + \widetilde{Z}_j$ and so

$$d_{\mathrm{TV}}\big(Z_{j-1}, Z_j\big) = d_{\mathrm{TV}}\big(X_j + \widetilde{Z}_j, Y_j + \widetilde{Z}_j\big) \leqslant d_{\mathrm{TV}}(X_j, Y_j) \tag{9.14}$$

using (9.11) and the fact that $\widetilde{Z}_j$, X_j and Y_j are independent. □

We are now ready to give:

Proof of Theorem 9.3. Using the previous reasoning, our goal is to prove (9.8) which using Lemma 9.5 reduces to the bound on $d_{\mathrm{TV}}(X,Y)$, where X is Bernoulli with parameter p and Y is Poisson with parameter p. Here we observe the following facts:

(1) $P(X=0) = 1-p \leqslant \mathrm{e}^{-p} = P(Y=0)$
(2) $P(X=1) = p \geqslant p\mathrm{e}^{-p} = P(Y=1)$
(3) $P(X=k) = 0 \leqslant P(Y=k)$ for $k \geqslant 2$

It follows that $P(X=k) - P(Y=k)$ is positive only for $k=1$ where

$$P(X=1) - P(Y=1) = p - p\mathrm{e}^{-p} = p(1-\mathrm{e}^{-p}) \leqslant p^2 \tag{9.15}$$

Relying on the middle line in (9.6), this shows $d_{\mathrm{TV}}(X,Y) \leqslant p^2$ proving the claim. □

The second part of Le Cam's inequality (9.4) indicates the conditions under which one expects the approximation by the Poisson random variable to be accurate: $\min_{i=1,\dots,n} p_i$ has to be small. Since $\sum_{i=1}^{n} p_i$ is fixed to λ, this can hardly be achieved without taking $n \to \infty$ and so Poisson convergence arises when (and, in fact, whenever) we count

occurrences of a large number of independent small-probability events. Sometimes this principle is described by the phrase "law of small numbers."

While independence has been crucial in above proofs, there are ways to work with dependent events. However, the corresponding statement requires the notion of conditional probability that we have not covered yet, so we leave it to a later discussion.

***Remark* 9.6** Note that the weak convergence $Y_n \xrightarrow{\text{w}} \text{Poisson}(\lambda)$ is reflected at the level of characteristic functions as follows

$$Ee^{itY_n} \underset{n\to\infty}{\longrightarrow} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} e^{itn} = \exp\bigl\{\lambda(e^{it}-1)\bigr\}, \quad t \in \mathbb{R} \tag{9.16}$$

If only non-negative random variables are involved, we can also phrase it using the Laplace transforms as

$$Ee^{-tY_n} \underset{n\to\infty}{\longrightarrow} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} e^{-tn} = \exp\bigl\{-\lambda(1-e^{-t})\bigr\}, \quad t \geqslant 0 \tag{9.17}$$

Thus, whenever we encounter the expression of the form $(e^{it}-1)$, resp., $(1-e^{-t})$ in the exponent of a characteristic function, resp., a moment generating function, we should suspect a Poisson random variable at play.

9.2 Poisson processes.

Thanks to the "law of small numbers" principle, Poisson random variables play a fundamental role in modeling of physical phenomena. The standard examples include the number of atoms of radioactive material that decayed over a given time period, the number of customers in a queue or the number of rainy days in a place with dry climate. The modeling context usually involves the flow of time which naturally leads to:

Lemma 9.7 *Let $\lambda > 0$ and let $\{T_i\}_{i\geqslant 1}$ be i.i.d. Exponential with parameter λ. (This means $ET_i = \lambda$.) For each $t \geqslant 0$ set*

$$N_t := \max\biggl\{k \geqslant 0 \colon \sum_{i=1}^{k} T_i \leqslant t\biggr\} \tag{9.18}$$

with the proviso $N_t := +\infty$ when $\sum_{i=1}^{\infty} T_i \leqslant t$. Then

$$\forall k \geqslant 1\, \forall 0 = t_0 < t_1 < \cdots < t_k \colon \ \{N_{t_i} - N_{t_{i-1}}\}_{i=1}^{k} \text{ are independent} \tag{9.19}$$

and

$$\forall t \geqslant s \geqslant 0 \colon \ N_t - N_s \overset{\text{law}}{=} \text{Poisson}\bigl(\lambda(t-s)\bigr) \tag{9.20}$$

Moreover, $t \mapsto N_t$ is right-continuous with left-limits a.s.

Proof (hint). We leave the proof of this to a homework assignment subject to the following hint: Discretize the "time axis" to $\{k/n \colon k \geqslant 0\}$ and let $\{X_k\}_{k\geqslant 0}$ be Bernoulli with parameter λ/n. Let $K_1, K_2, \ldots$ enumerate the set $\{k \geqslant 0 \colon X_k = 1\}$ increasingly. Then $K_j - K_{j-1}$ are independent Geometric with parameter λ/n and so $\{n^{-1}(K_j - K_{j-1})\}_{i\geqslant 1}$

tends in distribution to $\{T_i\}_{i\geqslant 1}$. Letting N_t^n be the largest j such that $K_j/n \leqslant t$, the increments of the process $t \mapsto N_t^n$ are independent with $N_t^n - N_s^n$ Binomial with parameter $n(t-s) \pm 1$ and λ/n and, using Lemma 9.1, $N_t^n - N_s^n$ thus tends to Poisson$(\lambda(t-s))$. $\square$

We remark that proofs of the above exist that do not involve discretization but rather rely on the so called memoryless property of the exponential distribution. We find the argument via discretization far more illuminating.

Definition 9.8 *The process $\{N_t\colon t \geqslant 0\}$ with the properties as stated in Lemma 9.7 is called the homogenous rate-λ Poisson process.*

As it turns out the phrase "Poisson process" is actually tied to another object that, in fact, subsumes the one defined above. We go straight to:

Definition 9.9 *Let $(\mathscr{X}, \mathcal{G}, \mu)$ be a measure space. A* Poisson point process with intensity μ *is (to be abbreviated as* PPP(μ)*) is a random measure θ on $(\mathscr{X}, \mathcal{G})$ such that*

(1) $\forall k \in \mathbb{N}\, \forall A_1, \dots, A_k \in \mathcal{G}$:

$$A_1, \dots, A_k \text{ disjoint} \quad \Rightarrow \quad \theta(A_1), \dots, \theta(A_k) \text{ independent} \tag{9.21}$$

(2) $\forall A \in \mathcal{G}$: $\theta(A)$ *is Poisson with parameter $\mu(A)$*

Here Poisson$(\infty) := \infty$ a.s.

Note that a random measure on $(\mathscr{X}, \mathcal{G})$ is a measure-valued random variable which, technically, is a map $\Omega \to \mathcal{M}$, where $(\Omega, \mathcal{F}, P)$ is a probability space and

$$\mathcal{M} := \big\{\theta\colon \text{ measure on } (\mathscr{X}, \mathcal{G})\big\} \tag{9.22}$$

We usually endow this space with the minimal σ-algebra

$$\sigma\Big(\big\{\mu \in \mathcal{M}\colon \mu(A) \in B\big\}\colon A \in \mathcal{G},\, B \in \mathcal{B}(\mathbb{R})\Big) \tag{9.23}$$

that makes $\theta(A)$ measurable (and hence an $\mathbb{R}_+ \cup \{+\infty\}$-valued random variable) for each $A \in \mathcal{G}$. For each $\omega \in \Omega$, the realization $\theta(\omega, \cdot)$ is then a measure in the second coordinate such that $\omega \mapsto \theta(\omega, A)$ is measurable for each $A \in \mathcal{G}$.

The PPP(μ) will hardly be interesting unless there is a good number of finite μ-measure sets. As it turns out, this is all we need to ensure its existence:

Theorem 9.10 *Let μ be a σ-finite. Then* PPP(μ) *exists and is σ-finite a.s.*

The proof will require a tool from the theory of multivariate distributions:

Lemma 9.11 *Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be real-valued random variables such that*

$$\forall t_1, \dots, t_n \in \mathbb{R}\colon \quad E\big(\mathrm{e}^{\mathrm{i}\sum_{j=1}^k t_j X_j}\big) = E\big(\mathrm{e}^{\mathrm{i}\sum_{j=1}^k t_j Y_j}\big) \tag{9.24}$$

Then

$$(X_1, \dots, X_n) \overset{\text{law}}{=} (Y_1, \dots, Y_n) \tag{9.25}$$

in the sense of equality of probability measures on $\mathbb{R}^n$.

Proof. Denote the function on the left of (9.24) by $\varphi_X(t_1,\dots,t_n)$. The multivariate version of the inversion formula in Theorem 3.10 shows that, for any collection of numbers $a_1,\dots,a_n,b_1,\dots,b_n\in\mathbb{R}$ with $a_i<b_i$ for all $i=1,\dots,n$,

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{\pi^n}\int_{[-T,T]^n}\varphi_X(t_1,\dots,t_n)\prod_{j=1}^n\frac{\mathrm{e}^{-\mathrm{i}t_ja_j}-\mathrm{e}^{-\mathrm{i}t_jb_j}}{\mathrm{i}t_j}\mathrm{d}t_1\dots\mathrm{d}t_n\\ =E\left(\prod_{j=1}^n\left[\frac{1}{2}1_{[a_j,b_j]}(X_j)+\frac{1}{2}1_{(a_j,b_j)}(X_j)\right]\right)\end{aligned}\tag{9.26}$$

Taking $a_j\to-\infty$ and restricting each b_j to stay away from the (at most countable) set $\{X_j\in\mathbb{R}\colon P(X_i=x)>0\}$, the right-hand side of (9.26) becomes $P(\bigcap_{j=1}^n\{X_j\leqslant b_j\})$. It follows that φ_X determines the multivariate CDF of the random vector $X=(X_1,\dots,X_n)$ which by Dynkin's π/λ-theorem (Theorem 1.8) determines the distribution of X on $\mathbb{R}^n$. Equality of the characteristic functions thus implies equality in distribution. □

Proof of Theorem 9.10. Assume first that μ is finite. If $\mu(\mathscr{X})=0$ then we can set $\theta:=0$ so let us assume that $\mu(\mathscr{X})>0$. Let

$$N\overset{\text{law}}{=}\text{Poisson}\big(\mu(\mathscr{X})\big)\tag{9.27}$$

and let $X_1,X_2,\dots$ be i.i.d. with

$$P(X_1\in A)=\frac{\mu(A)}{\mu(\mathscr{X})},\quad A\in\mathcal{G}\tag{9.28}$$

with $\{X_i\}_{i\geqslant1}$ independent of N. A probability space $(\Omega,\mathcal{F},P)$ supporting these random variables exists without any additional requirements on their structure.

Let $\theta\colon\mathcal{G}\to\mathbb{N}\cup\{+\infty\}$ be the map defined by

$$\theta(A):=\sum_{i=1}^N1_A(X_i),\quad A\in\mathcal{G}\tag{9.29}$$

Then each realization of θ is an $\mathbb{N}\cup\{+\infty\}$-valued measure on $(\mathscr{X},\mathcal{G})$ and $\theta(A)$ is a random variable on $(\Omega,\mathcal{F},P)$ for each $A\in\mathcal{G}$.

It remains to check that properties (1-2) in Definition 9.9. For this pick $t_1,\dots,t_k\in\mathbb{R}$ and $A_1,\dots,A_k\in\mathcal{G}$ and compute using the fact that N is independent of $X_1,X_2,\dots$ which are themselves independent and equidistributed to X_1 to obtain

$$\begin{aligned}E\big(\mathrm{e}^{\mathrm{i}\sum_{j=1}^kt_j\theta(A_j)}\big)&=E\exp\left\{\mathrm{i}\sum_{i=1}^N\sum_{j=1}^kt_j1_{A_j}(X_i)\right\}\\&=\sum_{n=0}^\infty\frac{\mu(\mathscr{X})^n}{n!}\mathrm{e}^{-\mu(\mathscr{X})}\left[E\exp\left\{\mathrm{i}\sum_{j=1}^kt_j1_{A_j}(X_1)\right\}\right]^n\\&=\exp\left\{\mu(\mathscr{X})\left(E\exp\left\{\mathrm{i}\sum_{j=1}^kt_j1_{A_j}(X_1)\right\}-1\right)\right\}\end{aligned}\tag{9.30}$$

Next we invoke that $A_1, \dots, A_k$ are disjoint with the result

$$\begin{aligned} E\exp\Big\{\mathrm{i}\sum_{j=1}^{k} t_j 1_{A_j}(X_1)\Big\} &= E\Big(\prod_{j=1}^{k} \mathrm{e}^{\mathrm{i}t_j 1_{A_j}(X_1)}\Big) \\ &= E\Big(\prod_{j=1}^{k}\Big(1 + (\mathrm{e}^{\mathrm{i}t_j} - 1)1_{A_j}(X_1)\Big)\Big) \\ &= 1 + \sum_{j=1}^{k}(\mathrm{e}^{\mathrm{i}t_j} - 1)1_{A_j}(X_1) = 1 + \sum_{j=1}^{k}(\mathrm{e}^{\mathrm{i}t_j} - 1)\mu(A_j) \end{aligned} \tag{9.31}$$

Putting these together we conclude

$$E\big(\mathrm{e}^{\mathrm{i}\sum_{j=1}^{k} t_j \theta(A_j)}\big) = \exp\Big\{\sum_{j=1}^{k}(\mathrm{e}^{\mathrm{i}t_j} - 1)\mu(A_j)\Big\} \tag{9.32}$$

Noting that the right-hand side is the multivariate characteristic function of $(Y_1, \dots, Y_k)$ that are independent with Y_j Poisson with parameter $\mu(A_j)$, using Lemma 9.11 we conclude that $\{\theta(A_j)\}_{j=1}^{k}$ satisfy (1-2) above.

Moving to the case when μ is infinite, the assumption that it is σ-finite implies that there are disjoint measurable sets $\{\mathscr{X}_n\}_{n\geqslant 1}$ with $\bigcup_{n\geqslant 1}\mathscr{X}_n = \mathscr{X}$ and $\mu(\mathscr{X}_n) < \infty$ for each $n \geqslant 1$. Let $\{\theta_n\}_{n\geqslant 1}$ be independent with $\theta_n = \mathrm{PPP}(1_{\mathscr{X}_n}\mu)$ for each $n \geqslant 1$. Set

$$\theta := \sum_{n\geqslant 1}\theta_n \tag{9.33}$$

Then for each $A \in \mathcal{G}$, the independence of $\{\theta_n(A)\}_{n\geqslant 1}$ along with Lemma 9.4 and a simple limiting argument ensure

$$\theta(A) = \sum_{n\geqslant 1}\theta_n(A) \overset{\text{law}}{=} \text{Poisson with parameter } \sum_{n\geqslant 1}\mu(A\cap\mathscr{X}_n) = \mu(A) \tag{9.34}$$

(We also used that, by the 2nd Borel-Cantelli lemma, if $\sum_{n\geqslant 1}\lambda_n = \infty$ then the sum of independent Poisson random variables with parameters $\lambda_1, \lambda_2, \dots$ diverges a.s.) It remains to show independence of $\{\theta(A_i)\}_{i=1}^{k}$ for disjoint $A_1, \dots, A_k \in \mathcal{G}$. Here we note that, under such conditions, the whole family $\{\theta_n(A_j)\colon n \geqslant 1, j = 1, \dots, k\}$ of random variables is independent. Then so are $\{\sum_{n\geqslant 1}\theta_n(A_j)\colon j = 1, \dots, k\}$.

As a final note, observe that $\theta(\mathscr{X}_n) = \theta_n(\mathscr{X}) = \text{Poisson}(\mu(\mathscr{X}_k))$ is finite a.s. for each $k \geqslant 1$. It follows that θ is σ-finite a.s. □

We remark that, for μ σ-finite, the presented construction permits us to think of each realization of θ as a collection of random points points (with at most a finite number points possibly falling on top of each other anywhere). This is why we call the resulting random measure a "point" process.

9.3 Some remarks.

We proceed with some remarks. First note that the homogeneous rate-λ Poisson process does fall under the concept of Poisson Point Process as well:

Lemma 9.12 *Let θ be a Poisson Point Process on $\mathbb{R}_+$ with intensity given by the λ-multiple of the Lebesgue measure. Then $\{N_t\colon t \geqslant 0\}$, where*

$$N_t := \theta\big([0,t]\big), \quad t \geqslant 0 \tag{9.35}$$

is a homogeneous rate-λ Poisson process constructed in Lemma 9.7.

This explains the attribute "homogeneous" in Definition 9.8 and also suggest how to define the inhomogenous counterparts: Use (9.35) albeit for $\theta = \mathrm{PPP}(\mu)$ for a general Radon measure μ on $\mathbb{R}_+$. The increments of the resulting process $\{N_t\colon t \geqslant 0\}$ over disjoint (half-open) intervals are still independent and Poisson distributed, but $N_t - N_s$ is now Poisson with parameter $\mu((s,t])$. (Textbooks sometimes define the process by these two properties; the construction then requires work.)

Next we elaborate on the fact that θ is actually a measure. Having a measure naturally suggests consideration of integrals $\int f \mathrm{d}\theta$. Here we note:

Lemma 9.13 *Suppose θ is a $\mathrm{PPP}(\mu)$ defined on a probability space $(\Omega, \mathcal{F}, P)$. Let $f \in L^1(\mu)$. Then $f \in L^1(\theta)$ a.s. and $\int f \mathrm{d}\theta$ admits a version that is a random variable. Moreover, we have*

$$E\int f \mathrm{d}\theta = \int f \mathrm{d}\mu \tag{9.36}$$

and so $f \mapsto \int f \mathrm{d}\theta$ is continuous as a map $L^1(\mu) \to L^1(\Omega, \mathcal{F}, P)$. Finally, for all $f \in L^1(\mu)$,

$$E\big(\mathrm{e}^{\mathrm{i}t\int f \mathrm{d}\theta}\big) = \exp\biggl\{\int (\mathrm{e}^{\mathrm{i}tf(x)} - 1)\mu(\mathrm{d}x)\biggr\} \tag{9.37}$$

holds (with the integral absolutely convergent) for each $t \in \mathbb{R}$.

Note that the law of $\int f \mathrm{d}\theta$ is not Poisson; rather, it is a *mixture* of Poisson random variables. A simple change of variables shows that it falls under:

Definition 9.14 *A real-valued random variable Z is said to be* compound Poisson *if there exists a Borel measure λ on $\mathbb{R}$ with $\int \frac{|z|}{1+|z|}\lambda(\mathrm{d}z) < \infty$ such that*

$$E\big(\mathrm{e}^{\mathrm{i}tZ}\big) = \exp\biggl\{\int (\mathrm{e}^{\mathrm{i}tz} - 1)\lambda(\mathrm{d}z)\biggr\} \tag{9.38}$$

holds for all $t \in \mathbb{R}$.

Using the familiar construction (cf the proof of Lemma 8.11 or Theorem 9.10), assuming $\lambda(\mathbb{R}) \in (0,\infty)$, an alternative (but equivalent) description of a compound random variable goes by taking independent random variables N and i.i.d. $\{X_i\}_{i\geqslant 1}$ with laws specified by

$$N \overset{\text{law}}{=} \mathrm{Poisson}\big(\lambda(\mathbb{R})\big) \quad \text{and} \quad P(X_1 \in \cdot) = \frac{\lambda(\cdot)}{\lambda(\mathbb{R})} \tag{9.39}$$

and setting

$$Z := \sum_{i=1}^{N} X_i \tag{9.40}$$

For $\lambda(\mathbb{R}) = +\infty$ we perform a truncation by introducing

$$\lambda_\epsilon(\mathrm{d}x) := 1_{[-\epsilon,\epsilon]^c}(x)\lambda(\mathrm{d}x) \tag{9.41}$$

which is finite due to $\int \frac{|z|}{1+|z|}\lambda(\mathrm{d}z) < \infty$. Then take $\epsilon \downarrow 0$ with the help of Dominated Convergence. The effect of truncation can alternatively be captured by writing Z as the double sum

$$Z = \sum_{k=1}^{\infty} \sum_{i=1}^{N_k} X_{k,i} \tag{9.42}$$

where $\{N_k\}_{k\geqslant 1}$ and $\{X_{k,i}\}_{k,i\geqslant 1}$ are independent with $N_k \overset{\text{law}}{=} \text{Poisson}(\lambda'_k(\mathbb{R}))$, for λ'_k a measure defined, say, by $\lambda'_k := \lambda_{2^{-k-1}} - \lambda_{2^{-k}}$, and $P(X_{k,i} \in \cdot) = \lambda'_k(\cdot)/\lambda'_k(\mathbb{R})$ when $\lambda'_k(\mathbb{R}) > 0$ and $P(X_{k,i} = 0) = 1$ otherwise. Unless λ is finite, the double sums contains infinitely many non-zero terms but converges absolutely a.s.

As also noted in the proof of Lemma 8.11, a limit argument can be performed even if $\int \frac{|z|}{1+|z|}\lambda(\mathrm{d}z) = \infty$ but $\int \frac{z^2}{1+z^2}\lambda(\mathrm{d}z) < \infty$ holds. What is needed is to introduce a shift by

$$\mu_\epsilon := \int \frac{z}{1+z^2}\,\lambda_\epsilon(\mathrm{d}z) \tag{9.43}$$

Indeed, for Z_ϵ defined via λ_ϵ we then have $Z_\epsilon - \mu_\epsilon \xrightarrow{\text{w}} \widetilde{Z}$ as $\epsilon \downarrow 0$, where $\widetilde{Z}$ is defined by the characteristic function

$$E\big(\mathrm{e}^{\mathrm{i}t\widetilde{Z}}\big) = \exp\biggl\{\int\Big(\mathrm{e}^{\mathrm{i}tz} - 1 - \frac{\mathrm{i}tz}{1+z^2}\Big)\lambda(\mathrm{d}z)\biggr\} \tag{9.44}$$

where the integral converges thanks to the assumption that $\int \frac{z^2}{1+z^2}\lambda(\mathrm{d}z) < \infty$. Comparing this with the Lévy-Khinchin formula, we conclude that the weak closure of shifted compound Poisson random variables are exactly all the infinite divisible laws. (The measure λ is then the Lévy measure.)

The above construction has an interesting consequence: If λ is supported on $\mathbb{R}_+$ then both (9.40) and (9.42) are positive and so a random variable with characteristic function (9.38) with $\operatorname{supp}\lambda \subseteq \mathbb{R}_+$ is necessarily non-negative. The shift by μ_ϵ mucks this up unless $\{\mu_\epsilon\}$ stays bounded. This is why a totally-skewed (to the right) stable random variable is bounded from below if $\alpha < 1$ and unbounded when $\alpha \in [1, 2)$.

We leave the proofs of Lemmas 9.12 and 9.13 to homework.

Further reading: Durrett, Section 3.7

10. CONDITIONAL EXPECTATION

Having discussed at length the limit theory for sums of independent random variables we will now move on to deal with dependent random variables. An important tool in this is conditioning which is what we will focus on in this lecture.

10.1 Definitions and the Radon-Nikodym theorem.

In undergraduate probability we learn that, given two events A and B with $P(B) > 0$, the knowledge that B occurs alters the probability that A occurs to

$$P(A|B) := \frac{P(A \cap B)}{P(B)} \tag{10.1}$$

which we call the *conditional probability of A given B*. The underlying idea behind this formula is that, since probability anyway represents the *relative* chance of something happening compared to all else, knowing that B occurred only restricts both the event in question and the set of "all else" in the ratio.

With (10.1) settled, we then treat $P(\cdot \mid B)$ as a new probability and use it to define conditional expectation of random variable X given B by

$$E(X|B) := \frac{E(X1_B)}{P(B)} \tag{10.2}$$

We will define a similar that instead of conditioning on occurrence of specific events conditions on a whole σ-algebras. The idea is that a σ-algebra represents available information about the problem at hand. The key definition comes in:

Definition 10.1 (Conditional expectation) *Consider a probability space $(\Omega, \mathcal{F}, P)$ and a random variable $X \in L^1(\Omega, \mathcal{F}, P)$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sigma-algebra. A random variable $Y \in L^1(\Omega, \mathcal{F}, P)$ is a* conditional expectation of X given $\mathcal{G}$ *if*

(1) *Y is $\mathcal{G}$-measurable, and*
(2) $\forall A \in \mathcal{G}\colon \;\; E(Y1_A) = E(X1_A)$

hold true.

To see that the new definition does contain the framework described earlier, consider an event $B \in \mathcal{F}$ with $P(B) \in (0,1)$ and let $\mathcal{G} := \{\varnothing, B, B^c, \Omega\}$. Then define

$$Y := \begin{cases} \dfrac{E(X1_B)}{P(B)}, & \text{on } B, \\ \dfrac{E(X1_{B^c})}{P(B^c)}, & \text{on } B^c, \end{cases} \tag{10.3}$$

We then readily check that, since Y is constant on B,

$$E(Y1_B) = \frac{E(X1_B)}{P(B)} E(1_B) = E(X1_B) \tag{10.4}$$

and similarly $E(X1_{B^c}) = E(X1_{B^c})$. Then $E(X1_A) = E(X1_A)$ trivially for $A = \varnothing, \Omega$ and, since Y is also $\mathcal{G}$-measurable, it is thus a conditional expectation of X given $\mathcal{G}$.

Despite its simplicity, the example we just gave is instructive in pointing out that, for $\mathcal{G}$ defined by a partition of Ω, any $\mathcal{G}$-measurable function has to be constant on each element of the partition. In particular, our concept of conditional expectation simultaneously captures the values $E(X|B)$ for all B in the partition that are of non-zero measure. Encoding these into a function, it is then more amenable to tools from analysis.

In order to work with the concept of conditional expectation, it is important to first check whether the imposed conditions identify it uniquely. This comes in:

Lemma 10.2 *If Y and Y' satisfy (1-2), then $Y = Y'$ a.s.*

Proof. Note that $A := \{Y > Y'\} \in \mathcal{G}$ because both Y and Y' are $\mathcal{G}$-measurable by (1), and so it thus $Y - Y'$. Therefore, by (2),

$$E(Y1_{\{Y>Y'\}}) = E(X1_{\{Y>Y'\}}) = E(Y'1_{\{Y>Y'\}}). \tag{10.5}$$

which implies

$$E\big((Y - Y')1_{\{Y>Y'\}}\big) = 0 \tag{10.6}$$

This forces $Y \leqslant Y'$ a.s. Symmetrically, we also get $Y' \leqslant Y$ a.s. proving $Y = Y'$ a.s. □

Since the conditional expectation of X given $\mathcal{G}$ is determined up to a modification on a null set, we will henceforth write $E(X|\mathcal{G})$ to denote any version of this function. However, we first have to address existence:

Proposition 10.3 *For each $X \in L^1$, the conditional expectation $E(X|\mathcal{G})$ exists.*

The proof will be based on the same set of arguments that yield two important results from measure theory: the Radon-Nikodym theorem and the Lebesgue decomposition of measures. We will thus prove all of these together. We start with:

Definition 10.4 *Given two measures μ, ν on a measurable space $(\mathcal{X}, \mathcal{G})$, we say:*

- *μ is* absolutely continuous with respect to *ν, with notation $\mu \ll \nu$, if*

$$\forall A \in \mathcal{G}: \quad \nu(A) = 0 \;\Rightarrow\; \mu(A) = 0 \tag{10.7}$$

- *μ* and *ν* are mutually singular*, with notation $\mu \perp \nu$, if*

$$\exists A \in \mathcal{G}: \;\; \mu(A) = 0 \;\wedge\; \nu(A^c) = 0 \tag{10.8}$$

If $\mu \ll \nu$ and $\nu \ll \mu$ both hold then we say that μ and ν are mutually absolutely continuous or, alternatively, that they are equivalent.

We then have:

Theorem 10.5 (Lebesgue decomposition, Radon-Nikodym Theorem) *Let μ, ν be σ-finite measures on $(\Omega, \mathcal{F})$. Then there are measures μ_1, μ_2 so that*

$$\mu = \mu_1 + \mu_2 \quad \wedge \quad \mu_1 \ll \nu \quad \wedge \quad \mu_2 \perp \nu \tag{10.9}$$

Moreover, there exists an $\mathcal{F}$-measurable, non-negative $f: \Omega \to \mathbb{R}$ such that

$$\forall A \in \mathcal{F}: \quad \mu_1(A) = \int_A f(x)\nu(\mathrm{d}x) \tag{10.10}$$

The part of the statement when $\mu_1 \ll \nu$ implies existence of f such that $\mu_1 = f\nu$ is the *Radon-Nikodym theorem*, with f (sometimes denoted as $\frac{\mathrm{d}\mu_1}{\mathrm{d}\nu}$) called the *Radon-Nikodym derivative* of μ_1 with respect to ν. The fact that we may write each μ as $\mu = f\nu + \mu_2$ with $\mu_2 \perp \nu$ is referred to as the *Lebesgue decomposition*. While these two results are often proved separately (with the Lebesgue decomposition relying on the *Hahn* and *Jordan decompositions* of signed measures), our argument will give both in one shot.

10.2 Proof of Theorem 10.5 and Proposition 10.3.

Our proof of Theorem 10.5 (originally due to J. von Neumann) is based on Hilbert space techniques. Recall that a linear vector space $\mathcal{H}$ with an inner product $(\cdot,\cdot)$ is a Hilbert space if it complete in the norm $\|\cdot\|$ associated with the inner product as $\|f\| = (f,f)^{1/2}$. Examples of Hilbert spaces are ubiquitous in analysis: $\mathbb{R}$, $\mathbb{C}$, $\mathbb{R}^n$, $\ell^2(\mathbb{N})$, $L^2(\mu)$, etc.

A linear map $\varphi\colon \mathcal{H} \to \mathbb{R}$ (assume $\mathcal{H}$ is over reals) is a continuous linear functional if and only if it is bounded in the sense that $\exists C > 0 \forall x \in \mathcal{H}\colon\ |\varphi(x)| \leqslant C\|x\|$. Denoting by $\mathcal{H}^\star$ the space of continuous linear functionals on $\mathcal{H}$, we get:

Lemma 10.6 (Riesz representation) *Let $\mathcal{H}$ be a Hilbert space over the reals and $\varphi \in \mathcal{H}^*$ a continuous linear functional. Then there is $a \in \mathcal{H}$ such that*

$$\forall x \in \mathcal{H}\colon \quad \varphi(x) = (a,x) \tag{10.11}$$

Proof. The continuity of φ implies that $\mathcal{B} := \{x \in \mathcal{H}\colon \varphi(x) = 0\}$ is a closed linear subspace of $\mathcal{H}$. As $\varphi = 0$ is handled by $a := 0$, we may suppose $\varphi \neq 0$. Then there exists $x_0 \in \mathcal{H} \smallsetminus \mathcal{B}$ with $\varphi(x_0) \neq 0$. Next we show existence the orthogonal decomposition $x_0 = \tilde{a} + b$, where $b \in \mathcal{B}$ and $\tilde{a} \perp \mathcal{B}$. Let $\{b_n\}_{n\geqslant 0}$ be a sequence in $\mathcal{B}$ such that

$$\|x_0 - b_n\| \underset{n\to\infty}{\longrightarrow} c := \inf_{b\in\mathcal{B}} \|x_0 - b\| \tag{10.12}$$

Reducing to a subsequence, we may assume that $\|x_0 - b_n\|^2 - c^2 \leqslant 4^{-n-1}$. The Paralellogram Law then shows

$$\begin{aligned} \|b_n - b_m\|^2 &= \|(x_0 - b_m) - (x_0 - b_n)\|^2 \\ &= 2\|x_0 - b_n\|^2 + 2\|x_0 - b_n\|^2 - 4\left\| x_0 - \frac{b_n + b_m}{2}\right\|^2 \\ &\leqslant 2\|x_0 - b_n\|^2 + 2\|x_0 - b_n\|^2 - 4c^2 \leqslant 4^{-\min\{n,m\}} \end{aligned} \tag{10.13}$$

where we used that $\frac{b_n+b_m}{2} \in \mathcal{B}$ by the fact that $\mathcal{B}$ is a linear space. It follows that $\{b_n\}_{n\geqslant 1}$ is Cauchy in $\mathcal{H}$ and, since $\mathcal{B}$ is closed and $\mathcal{H}$ is complete, tends to a limit $b \in \mathcal{B}$.

Next observe that $b_n \to b$ used in (10.12) implies $c = \|x_0 - b\|$. Given any $b' \in \mathcal{B}$ and any $\epsilon \neq 0$, this gives

$$0 \leqslant \|x_0 - b + \epsilon b'\|^2 - \|x_0 - b\|^2 = 2\epsilon(b', x - b) + \epsilon^2\|b'\|^2 \tag{10.14}$$

Dividing by ϵ and taking $\epsilon \downarrow 0$ shows $(b', x-b) \geqslant 0$ and taking $\epsilon \uparrow 0$ gives $(b', x-b) = 0$. It follows that $\tilde{a} := x_0 - b \perp \mathcal{B}$ and $\varphi(\tilde{a}) = \varphi(x_0) \neq 0$. In particular, $\tilde{a} \neq 0$.

To finish the claim, observe that

$$\forall x \in \mathcal{H}: \quad x - \frac{\varphi(x)}{\varphi(\tilde{a})}\tilde{a} \in \mathcal{B} \tag{10.15}$$

as is checked by applying φ and invoking its linearity. Since $\tilde{a} \perp \mathcal{B}$, it follows that

$$\forall x \in \mathcal{H}: \quad (\tilde{a}, x) = \frac{\varphi(x)}{\varphi(\tilde{a})}(\tilde{a}, \tilde{a}) \tag{10.16}$$

Setting $a := \frac{\varphi(\tilde{a})}{(\tilde{a},\tilde{a})}\tilde{a}$, this reduces to (10.11). □

With this fact in hand, we can now give:

Proof of Theorem 10.5. The σ-finiteness assumption permits us to assume that both μ and ν are finite measures. (Otherwise partition the space into parts where this is true.) Consider the Hilbert space $\mathcal{H} := L^2(\Omega, \mathcal{F}, \mu + \nu)$ and define the functional

$$\varphi(g) := \int g \mathrm{d}\mu, \qquad g \in \mathcal{H} \tag{10.17}$$

Since μ is finite, the Cauchy-Schwarz inequality gives

$$\left|\varphi(g)\right| \leqslant \mu(\Omega)^{1/2}\|g\|_{L^2(\mu)} \leqslant C\|g\|_{\mathcal{H}} \tag{10.18}$$

As φ is linear, we get $\varphi \in \mathcal{H}^\star$. The Riesz representation (Lemma 10.6) ensures the existence of $h \in \mathcal{H}$ such that

$$\forall g \in L^2(\Omega, \mathcal{F}, \mu + \nu): \quad \int g \mathrm{d}\mu = \int gh \, \mathrm{d}(\mu + \nu) \tag{10.19}$$

Note that the integral on the right is finite thanks to the Cauchy-Schwarz inequality.

We now make a couple of observations that show that h may be taken $[0,1]$-valued. First, taking $g := 1_{\{h \geqslant 1\}}$ in (10.19) implies $\mu(h \geqslant 1) \geqslant \mu(h \geqslant 1) + \nu(h \geqslant 1)$ which forces

$$\nu(h \geqslant 1) = 0 \tag{10.20}$$

Next, taking $g := 1_{\{h \geqslant 1+\epsilon\}}$ with $\epsilon > 0$ yields $\mu(h \geqslant 1 + \epsilon) \geqslant (1 + \epsilon)\mu(h \geqslant 1 + \epsilon)$ and thus also $\mu(h \geqslant 1 + \epsilon) = 0$ for all $\epsilon > 0$. Letting $\epsilon \downarrow 0$ we get

$$\mu(h > 1) = 0 \tag{10.21}$$

Finally, taking $g := 1_{\{h < -\epsilon\}}$ with $\epsilon > 0$ yields $0 \leqslant \mu(h < -\epsilon) \leqslant -\epsilon(\mu + \nu)(h < -\epsilon)$ implying $(\mu + \nu)(h < -\epsilon) = 0$. Hence we get

$$(\mu + \nu)(h < 0) = 0 \tag{10.22}$$

by letting $\epsilon \downarrow 0$.

With these observations in hand, we now define

$$\mu_1(\mathrm{d}x) := 1_{\{0 \leqslant h(x) < 1\}}\mu(\mathrm{d}x) \tag{10.23}$$

and

$$\mu_2(\mathrm{d}x) := 1_{\{h(x) = 1\}}\mu(\mathrm{d}x) \tag{10.24}$$

Then $\mu_1 + \mu_2 = \mu$ and $\mu_2 \perp \nu$ by the above claims. For the proof of the remaining properties, we rewrite (10.19) as

$$\int g(x)\big[1 - h(x)\big]\mu(\mathrm{d}x) = \int g(x)h(x)\nu(\mathrm{d}x) \tag{10.25}$$

On the left hand side we can now replace μ by μ_1; picking $A \in \mathcal{F}$, setting g to

$$g_\epsilon(x) := 1_{\{0 \leqslant h(x) < 1-\epsilon\}} \frac{1}{1-h(x)} 1_A(x) \tag{10.26}$$

which is bounded and thus in $L^2(\Omega, \mathcal{F}, \mu + \nu)$ for all $\epsilon > 0$, and taking $\epsilon \downarrow 0$ with the help of the Monotone Convergence Theorem then yields

$$\forall A \in \mathcal{F}\colon \quad \mu_1(A) = \int_A \frac{h(x)}{1-h(x)} 1_{\{0 \leqslant h(x) < 1\}}\, \nu(\mathrm{d}x) \tag{10.27}$$

Denoting the integrand as $f(x)$, this is (10.10). Since the integral vanishes when A is ν-null, this also gives $\mu_1 \ll \nu$ as desired. □

We are now also ready to conclude the existence of the conditional expectation:

Proof of Proposition 10.3. Suppose first that $X \geqslant 0$. Define the measure μ on $\mathcal{G}$ by

$$\mu(A) := E(X1_A), \quad A \in \mathcal{G} \tag{10.28}$$

This is a finite measure which is absolutely continuous with respect to P restricted to $\mathcal{G}$. Hence, there is $\mathcal{G}$-measurable $Y \geqslant 0$ such that

$$\forall A \in \mathcal{G}\colon \quad \mu(A) = \int_A Y(\omega)P(\mathrm{d}\omega) = E(Y1_A) \tag{10.29}$$

It follows that Y satisfies the properties (1-2) of conditional expectation. The general integrable X are taken care of by expanding into positive and negative parts and applying the above to each part separately. □

In summary, we showed that, given any sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, to each $X \in L^1$ we can assign a unique $\mathcal{G}$-measurable random variable $E(X|\mathcal{G}) \in L^1$ that integrates to the same number as X on any set $A \in \mathcal{G}$. The interpretation of $E(X|\mathcal{G})$ is the "best estimate" of X given the information contained in $\mathcal{G}$.

Further reading: Durrett, Section 4.1

11. Properties of conditional expectation

Having defined conditional expectation and showed its uniqueness modulo changes on null sets, we now discuss its properties. It will be noted that, in many ways, the conditional expectation behaves just like ordinary expectation.

11.1 Basic properties.

We start with a list of properties that are quite immediate to check:

Lemma 11.1 *Assuming all random variables below are integrable, we have:*

(1) *(triviality on constants)* $E(1|\mathcal{G}) = 1$ *a.s.*
(2) *(positivity)* $X \geqslant 0$ *a.s.* $\Rightarrow E(X|\mathcal{G}) \geqslant 0$ *a.s.*
(3) *(linearity)* $E(\alpha X + \beta Y|\mathcal{G}) = \alpha E(X|\mathcal{G}) + \beta E(Y|\mathcal{G})$ *a.s.*
(4) $E(E(X|\mathcal{G})) = E(X)$.
(5) *If* X *is* $\mathcal{G}$ *measurable, then* $E(X|\mathcal{G}) = X$ *a.s.*

Proof. For (1), (3) and (5) we just check that the suggested forms of conditional expectation satisfies properties in Definition 10.1; the claim then follow from the uniqueness in Lemma 10.2. For (2) we note that $X \geqslant 0$ implies $E(Y1_{\{Y<0\}}) = 0$ for any version Y of $E(X|\mathcal{G})$ and so $Y \geqslant 0$ a.s. Property (4) is checked directly from Definition 10.1(2). □

Note that property (5) is consistent with our interpretation of $E(X|\mathcal{G})$ being the "best estimate" of X given the information available in $\mathcal{G}$ because being $\mathcal{G}$-measurable then just means that the "best estimate" of X is X itself. The following simple observation can be explained in similar vain:

Lemma 11.2 *Suppose that* $\sigma(X)$ *and* $\mathcal{G}$ *are independent. Then*

$$E(X|\mathcal{G}) = EX \qquad \text{a.s.} \tag{11.1}$$

Proof. Constants are universally measurable so condition (1) in the definition of conditional expectation holds. For condition (2) we notice that for any $A \in \mathcal{G}$, the stated independence means

$$E(X1_A) = E(X)E(1_A) = E\big(E(X)1_A\big). \tag{11.2}$$

Hence, $E(X)$ serves as a version of $E(X|\mathcal{G})$. □

Another property is an extension of Definition 10.1(2):

Lemma 11.3 *Let* X *and* Y *be random variables with* $Y \in L^1(\Omega, \mathcal{F}, P)$ *and* $XY \in L^1(\Omega, \mathcal{F}, P)$. *For any* σ*-algebra* $\mathcal{G} \subseteq \mathcal{F}$,

$$X \text{ is } \mathcal{G}\text{-measurable} \quad \Rightarrow \quad E(XY|\mathcal{G}) = XE(Y|\mathcal{G}) \quad \text{a.s.} \tag{11.3}$$

The upshot of this lemma is that anything that is measurable with respect to the conditional σ-algebra can be regarded as constant under the conditional expectation. We leave the proof to a homework exercise.

Next let also examine the dependence of the conditional expectation on the underlying σ-algebra. The following principle is often quite useful:

Lemma 11.4 ("Smaller always wins" principle) *Let $\mathcal{G}_1 \subseteq \mathcal{G}_2$ be sub σ-algebras of $\mathcal{F}$. Then for any $X \in L^1(\Omega, \mathcal{F}, P)$*

$$E\big(E(X|\mathcal{G}_1)\big|\mathcal{G}_2\big) = E(X|\mathcal{G}_1) \qquad \text{a.s.} \tag{11.4}$$

as well as

$$E\big(E(X|\mathcal{G}_2)\big|\mathcal{G}_1\big) = E(X|\mathcal{G}_1) \qquad \text{a.s.} \tag{11.5}$$

In short, iterated conditioning always reduces to that with respect to the smaller sigma algebra.

Proof. For (11.4), note that $E(X|\mathcal{G}_1)$ is $\mathcal{G}_2$-measurable (because it is already $\mathcal{G}_1$-measurable) and the expectation of both sides agree on any event $A \in \mathcal{G}_2$. The proof of uniqueness of the conditional expectation thus applies to show equality a.s.

For (11.5) we note that both sides are automatically $\mathcal{G}_1$-measurable. So picking $A \in \mathcal{G}_1$ and applying condition (2) from the definition of conditional expectation, we get

$$E(X1_A) \underset{A\in\mathcal{G}_2}{=} E\big(E(X|\mathcal{G}_2)1_A\big) \underset{A\in\mathcal{G}_1}{=} E\big(E\big(E(X|\mathcal{G}_2)\big|\mathcal{G}_1\big)1_A\big) \tag{11.6}$$

Again, by the uniqueness argument, this implies the second identity as well. □

A particularly interesting choice of $\mathcal{G}$ is the σ-algebra $\sigma(Y)$ generated by random variable Y. Here we get:

Lemma 11.5 (Doob-Dynkin lemma) *Let $X \in L^1(\Omega, \mathcal{F}, P)$ be $\mathbb{R}$-valued and, given a measurable space (S, Σ), let Y be an S-valued random variable. Then there exists a Borel measurable map $h\colon S \to \mathbb{R}$ such that*

$$E\big(X \,\big|\, \sigma(Y)\big) = h(Y) \text{ a.s.} \tag{11.7}$$

We leave the proof of this lemma to homework.

11.2 Convergence theorems.

As a consequence of above properties we get versions of standard convergence theorems for conditional expectation. We start with:

Lemma 11.6 (Conditional Monotone Convergence Theorem) *Suppose that $\{X_n\}_{n\geqslant 1}$ and X are random variables in L^1. Then*

$$X_n \uparrow X \text{ a.s.} \quad \Rightarrow \quad E(X_n|\mathcal{G}) \uparrow E(X|\mathcal{G}) \text{ a.s.} \tag{11.8}$$

Proof. Subtracting X_1 from X_n and X and applying additivity in Lemma 11.1(3) permits us to assume $X_n \geqslant 0$ for all $n \geqslant 1$. By Lemma 11.1(2), $X_n \uparrow X$ a.s. implies that $E(X_n|\mathcal{G})$ are non-decreasing a.s. Denoting $Y := \liminf_{n\to\infty} E(X_n|\mathcal{G})$, for each $A \in \mathcal{G}$ the Monotone Convergence Theorem along with properties in Definition 10.1 give

$$\begin{aligned} E\big(E(X|\mathcal{G})1_A\big) &= E\big(X1_A\big) \\ &= \lim_{n\to\infty} E\big(X_n 1_A\big) \\ &= \lim_{n\to\infty} E\big(E(X_n|\mathcal{G})1_A\big) = E\big(Y1_A\big) \end{aligned} \tag{11.9}$$

Hence we get

$$\forall A \in \mathcal{G}\colon \quad E\Big(\big[E(X|\mathcal{G}) - Y\big]1_A\Big) = 0 \tag{11.10}$$

Testing this on $A := \{E(X|\mathcal{G}) > Y\}$ and $A := \{E(X|\mathcal{G}) < Y\}$ shows $E(X|\mathcal{G}) = Y$ a.s. □

Another convergence lemma comes in:

Lemma 11.7 (Conditional Fatou's lemma) *Suppose that $\{X_n\}_{n\geqslant 1}$ are non-negative and belong to L^1. Assume also $\liminf_{n\to\infty} X_n \in L^1$. Then*

$$\liminf_{n\to\infty} E(X_n|\mathcal{G}) \geqslant E\big(\liminf_{n\to\infty} X_n\big|\mathcal{G}\big) \text{ a.s.} \tag{11.11}$$

Proof. In light of $X_n \geqslant \inf_{m\geqslant n} X_m$, we have $\inf_{m\geqslant n} X_m \in L^1$ for each $n \geqslant 1$. Lemma 11.1(2) then gives $E(X_n|\mathcal{G}) \geqslant E(\inf_{m\geqslant n} X_m|\mathcal{G})$ and so

$$\liminf_{n\to\infty} E(X_n|\mathcal{G}) \geqslant \liminf_{n\to\infty} E(\inf_{m\geqslant n} X_m|\mathcal{G}) \tag{11.12}$$

Since $\inf_{m\geqslant n} X_m \uparrow \liminf_{n\to\infty} X_n$, the claim follows from Lemma 11.6. □

***Remark* 11.8** Note that, unlike the non-conditional versions of (11.8) and (11.11), the conditional statements assume integrability of the limiting random variables. This can be overcome by defining $E(X|\mathcal{G})$ for any $X \geqslant 0$ as a $\mathcal{G}$-measurable version of

$$Y := \liminf_{n\to\infty} E(X \wedge n|\mathcal{G}) \tag{11.13}$$

which may be infinite on a set of positive or even full measure. (To give an example, let X take values in $\mathbb{N}' := \mathbb{N} \smallsetminus \{0\}$ with $P(X = n) = [n(n+1)]^{-1}$. For the choice $\mathcal{G} := \{\varnothing, \{X \text{ even}\}, \{X \text{ odd}\}, \mathbb{N}'\}$ we get $Y = \infty$.) The argument in (11.9) then shows that $E(X1_A) = E(Y1_A)$ for all $A \in \mathcal{G}$, albeit with both sides possibly infinite.

Noting that the identity gives $E(X1_{\{Y\leqslant a\}}) < \infty$ for each $a > 0$ suggests how to upgrade our previous arguments. For instance, if Y' is another $\mathcal{G}$-measurable non-negative random variable satisfying $E(X1_A) = E(Y1_A)$ for all $A \in \mathcal{G}$, then testing this on $A := \{Y' > Y\} \cap \{Y \wedge Y' \leqslant a\}$ and taking $a \to \infty$ shows $Y' \leqslant Y$ a.s. on $\{Y \wedge Y' < \infty\}$ and, by symmetry, implies $Y = Y'$ a.s. The statement of (11.8) then applies because $Y' := \lim_{n\to\infty} E(X_n|\mathcal{G})$ is a version of Y. Similarly we get (11.11).

We remark that, with conditional Fatou's lemma in place, we can also easily prove:

Lemma 11.9 (Conditional Dominated Convergence Theorem) *Let $\{X_n\}_{n\geqslant 1}$ be random variables such that $\forall n \geqslant 1\colon\ |X_n| \leqslant Y$ for some $Y \in L^1(\Omega, \mathcal{F}, P)$. Given any σ-algebra $\mathcal{G} \subseteq \mathcal{F}$,*

$$X_n \underset{n\to\infty}{\longrightarrow} X \text{ a.s.} \quad \Rightarrow \quad E(X_n|\mathcal{G}) \underset{n\to\infty}{\longrightarrow} E(X|\mathcal{G}) \text{ a.s.} \tag{11.14}$$

Here $E(X|\mathcal{G})$ is well defined because the assumptions imply $X \in L^1(\Omega, \mathcal{F}, P)$.

Proof. Suppose that $X_n \to X$ a.s. with $|X_n| \leqslant Y$ for each $n \geqslant 1$. Since $0 \leqslant X_n + Y \leqslant 2Y$ implies $\liminf_{n\to\infty}(X_n + Y) \in L^1$, Lemma 11.7 gives

$$E(X + Y|\mathcal{G}) \leqslant \liminf_{n\to\infty} E(X_n + Y|\mathcal{G}) \text{ a.s.} \tag{11.15}$$

Working with $Y - X_n$ instead similarly shows

$$E(X - Y|\mathcal{G}) \geqslant \limsup_{n\to\infty} E(X_n - Y|\mathcal{G}) \text{ a.s.} \tag{11.16}$$

where we reversed signs for convenience. Subtracting or adding $E(Y|G)$ on both sides with the help of linearity in Lemma 11.6(3) yields

$$E(X|\mathcal{G}) \leqslant \liminf_{n\to\infty} E(X_n|\mathcal{G}) \leqslant \limsup_{n\to\infty} E(X_n|\mathcal{G}) \leqslant E(X|\mathcal{G}) \text{ a.s.} \tag{11.17}$$

which is the desired conclusion. □

11.3 Inequalities.

The conditional expectation also generalizes a number of inequalities that we know from Lebesgue integration theory. We start with:

Lemma 11.10 *If $X \in L^1(\Omega, \mathcal{F}, P)$ and $\mathcal{G} \subseteq \mathcal{F}$ is a σ-algebra, then*

$$\big|E(X|\mathcal{G})\big| \leqslant E\big(|X|\big|\mathcal{G}\big) \tag{11.18}$$

and so

$$\big\|E(X|\mathcal{G})\big\|_{L^1(\Omega,\mathcal{G},P)} \leqslant \|X\|_{L^1(\Omega,\mathcal{F},P)}. \tag{11.19}$$

In particular, $X \mapsto E(X|\mathcal{G})$ is a continuous linear map — in fact, a contraction — of $L^1(\Omega, \mathcal{F}, P)$ onto $L^1(\Omega, \mathcal{G}, P)$.

Proof. We have $-|X| \leqslant X \leqslant |X|$ and so

$$-E\big(|X|\big|\mathcal{G}\big) \leqslant E(X|\mathcal{G}) \leqslant E\big(|X|\big|\mathcal{G}\big) \qquad \text{a.s.} \tag{11.20}$$

This proves (11.18); taking expectation then yields (11.19). Since $X \mapsto E(X|\mathcal{G})$ is linear a.s. by Lemma 11.6(3), it defines a linear map of $L^1(\Omega, \mathcal{F}, P)$ onto $L^1(\Omega, \mathcal{G}, P)$. By (11.19) the operator norm of this map is less than one so this map is a contraction. The map is onto as it acts as identity on $L^1(\Omega, \mathcal{G}, P)$. □

Lemma 11.11 (Conditional Cauchy-Schwarz inequality) *Suppose $X, Y \in L^2(\Omega, \mathcal{F}, P)$ and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. Then*

$$\big[E(XY|\mathcal{G})\big]^2 \leqslant E(X^2|\mathcal{G})E(Y^2|\mathcal{G}) \qquad \text{a.s.} \tag{11.21}$$

Proof. Note that $XY, X^2, Y^2 \in L^1$ by assumption. This permits us to pick finite versions of the conditional expectations in the statement. By Lemma 11.6(3),

$$E(X^2|\mathcal{G}) - 2\lambda E(XY|\mathcal{G}) + \lambda^2 E(Y^2|\mathcal{G}) \tag{11.22}$$

is a version of $E((X - \lambda Y)^2|\mathcal{G})$, which is non-negative a.s. by Lemma 11.6(2). It follows that there exists $\Omega' \in \mathcal{G}$ with $P(\Omega') = 1$ such that the expression in (11.22) is finite and non-negative for all $\lambda \in \mathbb{Q}$. By continuity, it is then non-negative for all $\lambda \in \mathbb{R}$ on Ω' and so the discriminant of the quadratic polynomial (11.22) is non-negative on Ω'. This is what is stated in (11.21). □

We now upgrade another useful identity from analysis:

Lemma 11.12 (Conditional Jensen's inequality) *Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be convex. Then for any $X \in L^1(\Omega, \mathcal{F}, P)$ such that $\phi(X) \in L^1(\Omega, \mathcal{F}, P)$ and any σ-algebra $\mathcal{G} \subseteq \mathcal{F}$,*

$$E\big(\phi(X)\big|\mathcal{G}\big) \geqslant \phi\big(E(X|\mathcal{G})\big) \qquad \text{a.s.} \tag{11.23}$$

Proof. Since the claim holds trivially if ϕ is linear, we may assume that it is not. Set

$$\mathcal{L} := \bigcap_{x\in\mathbb{R}} \big\{(a,b) \in \mathbb{R}^2\colon \phi(x) \geqslant ax + b\big\} \tag{11.24}$$

We then claim

$$\forall x \in \mathbb{R}\colon \quad \phi(x) = \sup_{(a,b)\in\mathcal{L}\cap\mathbb{Q}^2} (ax + b) \tag{11.25}$$

Indeed, the definition of $\mathcal{L}$ gives "$\geqslant$" and so we need to show "$\leqslant$." Let $x_0 \in \mathbb{R}$. Noting that the convexity of ϕ implies the existence of left and right derivatives at x_0, let $a_0 \in \mathbb{R}$ be any number in the closed interval marked by the two derivatives. For $b := \phi(x_0) - ax_0$, the graph of $x \mapsto ax + b$ is then tangent to the graph of ϕ at x_0 and so $(a, b) \in \mathcal{L}$. Now observe that, for any $b' < b$, we must have $(a', b') \in \mathcal{L}$ whenever $|a' - a|$ is sufficiently small, for otherwise ϕ would have to be linear. This allows us to pick $(a'_n, b'_n) \in \mathcal{L} \cap \mathbb{Q}^2$ such that $(a'_n, b'_n) \to (a, b)$. But then $\phi(x_0) = ax_0 + b = \lim_{n\to\infty}(a'_n x_0 + b'_n)$ proving "$\leqslant$" in (11.25). Hence equality holds as desired.

Since $\phi(X) \geqslant aX + b$ for each $(a, b) \in \mathcal{L}$, and since $\phi(X)$ and X are assumed to be in L^1, we thus have

$$\forall (a,b) \in \mathcal{L}\colon \quad E\big(\phi(X)\big|\mathcal{G}\big) \geqslant aE(X|\mathcal{G}) + b \qquad \text{a.s.} \tag{11.26}$$

Let $\Omega_{a,b}$ be the event where the inequality holds for pair (a, b) and both conditional expectations are finite. Define

$$\Omega' := \bigcap_{(a,b)\in\mathcal{L}\cap\mathbb{Q}^2} \Omega_{a,b} \tag{11.27}$$

and note that, on Ω', the inequality in (11.26) holds for all $(a, b) \in \mathcal{L} \cap \mathbb{Q}^2$. Taking a supremum with the help of (11.25) then proves the inequality in (11.23) on Ω'. Since (11.26) implies $P(\Omega') = 1$, we are done. □

11.4 Behavior in L^p-spaces.

A consequence of the conditional Jensen inequality is a generalization of Lemma 11.10:

Lemma 11.13 *For any $p \in [1, \infty]$, the map $X \mapsto E(X|\mathcal{G})$ is a continuous linear transform of $L^p(\Omega, \mathcal{F}, P)$ onto $L^p(\Omega, \mathcal{G}, P)$. In fact,*

$$\forall X \in L^p(\Omega, \mathcal{F}, P)\colon \quad \big\|E(X|\mathcal{G})\big\|_{L^p(\Omega,\mathcal{G},P)} \leqslant \|X\|_{L^p(\Omega,\mathcal{F},P)} \tag{11.28}$$

and so the map is a contraction.

Proof. Let $p \in [1, \infty)$. By conditional Jensen's inequality we have

$$\Big(E\big(|X|\big|\mathcal{G}\big)\Big)^p \leqslant E\big(|X|^p\big|\mathcal{G}\big) \qquad \text{a.s.} \tag{11.29}$$

Taking expectations and raising both sides to $1/p$ yields the result.

For $p := \infty$ we notice that $|X| \leqslant \|X\|_\infty$ implies

$$\big|E(X|\mathcal{G})\big| \leqslant \|X\|_\infty \qquad \text{a.s.} \tag{11.30}$$

and so the result follows in this case as well. □

For the special case $p = 2$, we even get a geometric representation:

Lemma 11.14 *Let $X \in L^2(\Omega, \mathcal{F}, P)$ and assume $\mathcal{G} \subseteq \mathcal{F}$ is a σ-algebra. Then, using the notion of orthogonality with respect to the canonical inner product,*

$$X - E(X|\mathcal{G}) \;\perp\; L^2(\Omega, \mathcal{G}, P) \tag{11.31}$$

In particular, $E(X|\mathcal{G})$ is the orthogonal projection of X onto $L^2(\Omega, \mathcal{G}, P)$.

Proof. For each $A \in \mathcal{G}$, Definition 10.1(2) gives

$$E\Big(\big[X - E(X|\mathcal{G})\big]\, 1_A\Big) = 0 \tag{11.32}$$

Using linearity, this shows that $X - E(X|\mathcal{G})$ is orthogonal to the set of all simple $\mathcal{G}$-measurable functions. Since these simple functions are dense in $L^2(\Omega, \mathcal{G}, P)$ it is orthogonal to the whole space. □

We remark that this observation reverse-engineers a key step in the proof of Theorem 10.5 where the Radon-Nikodym derivative was constructed exactly as an orthogonal projection. Note that Lemma 11.3 generalizes (11.31) as

$$X \in L^p(\Omega, \mathcal{F}, P) \;\Rightarrow\; X - E(X|\mathcal{G}) \;\perp\; L^q(\Omega, \mathcal{G}, P) \tag{11.33}$$

where q is the Hölder conjugate of p and "$\perp$" is still with respect to the canonical inner product in L^2. However, the concept of orthogonal projection is hard to articulate outside Hilbert spaces.

To wrap up our discussion of conditional inequalities, we also state:

Lemma 11.15 (Conditional Hölder inequality) *Let $p, q \in [1, \infty]$ be such that $1/p + 1/q = 1$. Then for any σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, all $X \in L^p(\Omega, \mathcal{F}, P)$ and all $Y \in L^q(\Omega, \mathcal{F}, P)$,*

$$\big|E(XY|\mathcal{G})\big| \leqslant E\big(|X|^p|\mathcal{G}\big)^{1/p}\, E\big(|Y|^q|\mathcal{G}\big)^{1/q} \qquad \text{a.s.} \tag{11.34}$$

with $E\big(|X|^p|\mathcal{G}\big)^{1/p}$ interpreted as $\|X\|_\infty$ when $p = \infty$.

We leave the proof of Lemma 11.15 to homework.

Further reading: Durrett, Section 4.1

12. Regular conditional probability

The conditional expectation shares so many properties with expectation that one may wonder if it can also be realized as an integral with respect to a measure. This leads to the concepts of conditional distribution and probability that we elaborate on here.

12.1 Conditional distribution.

Consider the setting from the previous sections; namely, a probability space $(\Omega, \mathcal{F}, P)$ with a random variable X and a σ-algebra $\mathcal{G} \subseteq \mathcal{F}$. As is known from undergraduate probability, for $\mathcal{G}$ such that $\mathcal{G} = \sigma(Y)$ for some random variable Y we can often write the conditional expectation of X (or even functions of X) given $\mathcal{G}$ explicitly. For instance, assuming that (X, Y) admits a probability density $\rho\colon \mathbb{R}^2 \to \mathbb{R}_+$, we have

$$E\big(f(X)\,\big|\,\sigma(Y)\big)(\omega) = \frac{\displaystyle\int f(x)\rho(x, Y(\omega))\mathrm{d}x}{\displaystyle\int \rho(z, Y(\omega))\mathrm{d}z} \tag{12.1}$$

whenever $f(X) \in L^1(\Omega, \mathcal{F}, P)$, the denominator is non-vanishing and the numerator is finite. In this case we succeeded in writing the conditional expectation as an integral with respect to a Borel probability measure on $\mathbb{R}$ with probability density

$$\widetilde{\rho}_\omega(x) := \frac{\rho(x, Y(\omega))}{\displaystyle\int \rho(z, Y(\omega))\mathrm{d}z} \tag{12.2}$$

for ω where the integral is positive and, say, $\widetilde{\rho}_\omega(x) := 1_{[0,1]}(x)$ otherwise. (The value in the latter alternative is immaterial as it occurs with probability zero.)

The question is to what extend we can do this when the random variables do not admit a joint density or, more importantly, when $\mathcal{G}$ is not generated by a random variable. This is answered in:

Theorem 12.1 *Let X be a real-valued random variable on a probability space $(\Omega, \mathcal{F}, P)$ and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. Then there exists a map $\mu_X\colon \Omega \times \mathcal{B}(\mathbb{R}) \to [0,1]$ such that*

(1) *for all $\omega \in \Omega$, the map $B \mapsto \mu_X(\omega, B)$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$,*
(2) *for all $B \in \mathcal{B}(\mathbb{R})$, the map $\omega \mapsto \mu_X(\omega, B)$ is $\mathcal{G}$-measurable,*
(3) *for all Borel $f\colon \mathbb{R} \to \mathbb{R}$ with $f(X) \in L^1(\Omega, \mathcal{F}, P)$, the set of $\omega \in \Omega$ where*

$$f \in L^1(\mu_X(\omega, \cdot)) \tag{12.3}$$

and

$$E\big(f(X)\big|\mathcal{G}\big)(\omega) = \int f(x)\mu_X(\omega, \mathrm{d}x) \tag{12.4}$$

is $\mathcal{G}$-measurable and of full P-measure.

Moreover, for any map $\tilde{\mu}_X$ with the above properties, $\bigcap_{A\in\mathcal{B}(\mathcal{R})}\{\omega \in \Omega\colon \mu_X(\omega, A) \neq \tilde{\mu}_X(\omega, A)\}$ is a $\mathcal{G}$-measurable set of full P-measure.

Note that the null set where (12.3–12.4) fail is permitted to depend on f. The integral representation of the conditional expectation suggests the following terminology:

Definition 12.2 *We will call μ_X a version of the* conditional distribution of X given $\mathcal{G}$.

Proof of Theorem 12.1. For each $q \in \mathbb{Q}$, use the Axiom of Countable Choice to pick a version of $E(1_{\{X \leqslant q\}}|\mathcal{G})$. We then use these objects to define the following sets: First, for $q < q'$ rational we let

$$\Omega_{q,q'} := \bigl\{\omega \in \Omega\colon\, E(1_{\{X \leqslant q\}}|\mathcal{G})(\omega) \leqslant E(1_{\{X \leqslant q'\}}|\mathcal{G})(\omega)\bigr\}. \tag{12.5}$$

Next, for $q \in \mathbb{Q}$ we define

$$\Omega_q := \Bigl\{\omega \in \Omega\colon\, E(1_{\{X \leqslant q\}}|\mathcal{G})(\omega) = \inf_{\substack{q' > q \\ q' \in \mathbb{Q}}} E(1_{\{X \leqslant q'\}}|\mathcal{G})(\omega)\Bigr\}. \tag{12.6}$$

In addition, we also define

$$\begin{aligned} \Omega_{+\infty} &:= \Bigl\{\omega \in \Omega\colon\, \sup_{q \in \mathbb{Q}} E(1_{\{X \leqslant q'\}}|\mathcal{G})(\omega) = 1\Bigr\} \\ \Omega_{-\infty} &:= \Bigl\{\omega \in \Omega\colon\, \inf_{q \in \mathbb{Q}} E(1_{\{X \leqslant q'\}}|\mathcal{G})(\omega) = 0\Bigr\} \end{aligned} \tag{12.7}$$

Note that, by the properties of the conditional expectation, all of these are full-measure events in $\mathcal{G}$. So, if we set

$$\Omega' := \Bigl(\bigcap_{\substack{q' > q \\ q' \in \mathbb{Q}}} \Omega_{q,q'}\Bigr) \cap \Bigl(\bigcap_{q \in \mathbb{Q}} \Omega_q\Bigr) \cap \Omega_{+\infty} \cap \Omega_{-\infty}, \tag{12.8}$$

then also $\Omega' \in \mathcal{G}$ with $P(\Omega') = 1$.

We now move to the construction of μ_X. First, for $\omega \in \Omega$ and $q \in \mathbb{Q}$, let

$$F(\omega, q) := \begin{cases} E(1_{\{X \leqslant q\}}|\mathcal{G})(\omega), & \text{if } \omega \in \Omega' \\ 1_{\{q \geqslant 0\}}, & \text{otherwise} \end{cases} \tag{12.9}$$

Then $q \mapsto F(\omega, q)$ is non-decreasing, right-continuous on $\mathbb{Q}$ with limits 1, resp., 0 as $q \to +\infty$, resp., $q \to -\infty$. Setting

$$\widetilde{F}(\omega, x) := \inf_{\substack{q > x \\ q \in \mathbb{Q}}} F(\omega, q) \tag{12.10}$$

we get $x \mapsto \widetilde{F}(\omega, x)$ with exactly the same properties and such that $\widetilde{F}(\omega, q) = F(\omega, q)$ for all $q \in \mathbb{Q}$. In particular, $x \mapsto \widetilde{F}(\omega, x)$ is a CDF and so, by the existence theorems for Lebesgue-Stieltjes measures, there is a Borel probability measure $\mu_X(\omega, \cdot)$ such that

$$\widetilde{F}(\omega, x) = \mu_X\bigl(\omega, (-\infty, x]\bigr), \qquad x \in \mathbb{R} \tag{12.11}$$

It now remains to verify the stated properties.

Property (1) is ensured by the construction. For property (2), denote

$$\mathcal{L} := \bigl\{A \in \mathcal{B}(\mathbb{R})\colon\, \omega \mapsto \mu_X(\omega, A) \text{ is } \mathcal{G}\text{-measurable}\bigr\} \tag{12.12}$$

and check that $\mathcal{L}$ contains $\Omega, \varnothing$ and is closed under finite unions, complements and increasing limits and so it is a σ-algebra. Note also that $\omega \mapsto \widetilde{F}(\omega, x)$ is $\mathcal{G}$-measurable, being the infimum of a countable set of $\mathcal{G}$-measurable functions. In light of (12.11), $\mathcal{L}$ thus contains all sets of the form $(-\infty, q]$ for $q \in \mathbb{Q}$ and so $\mathcal{L} = \mathcal{B}(\mathbb{R})$ thus proving (2).

Condition (3) is verified similarly: First we note that (12.4) holds for f's of the form $f(x) := 1_{\{x \leqslant q\}}$ for $q \in \mathbb{Q}$. The set

$$\mathcal{L}' := \big\{A \in \mathcal{B}(\mathbb{R})\colon \mu_X(\cdot, A) = E(1_A|\mathcal{G}) \text{ a.s.}\big\} \tag{12.13}$$

is again checked to be a σ-algebra and so equality therein holds (a.s.) for all $A \in \mathcal{B}(\mathbb{R})$. This verifies (12.4) for all f of the form $f = 1_A$ with $A \in \mathcal{B}(\mathbb{R})$ and, by additivity, for all simple functions f. Using the Monotone Convergence Theorem, we now extend (12.4) to all non-negative measurable functions $f \in L^1$, thus proving (12.3). The extension to all $f \in L^1$ is performed by additivity.

It remains to prove uniqueness. Given μ_X and $\tilde{\mu}_X$ that both satisfy the stated properties, the representation (12.4) shows that the set

$$\widetilde{\Omega} := \bigcap_{q \in \mathbb{Q}} \Big\{\omega \in \Omega\colon \tilde{\mu}_X\big(\omega, (-\infty, q]\big) = \mu_X\big(\omega, (-\infty, q]\big)\Big\} \tag{12.14}$$

is $\mathcal{G}$-measurable and of full P-measure. Since $\{(-\infty, q]\colon q \in \mathbb{Q}\}$ form a π-system generating $\mathcal{B}(\mathbb{R})$, Dynkin's π/λ-theorem ensures that $\tilde{\mu}_X(\omega, A) = \mu_X(\omega, A)$ hold for all $\omega \in \widetilde{\Omega}$ and all $A \in \mathcal{B}(\mathbb{R})$. Hence $\mu_X(\omega, \cdot) = \tilde{\mu}_X(\omega, \cdot)$ for a.e. $\omega \in \Omega$. □

12.2 Conditional probability and lack thereof.

While the above seems to resolve our original question, the proof made a key use of the fact that X was real-valued. We can thus continue asking whether the same can be done for general-valued random variables and, perhaps more importantly, for *all* random variables simultaneously. This leads to the following concept:

Definition 12.3 *Given a probability space $(\Omega, \mathcal{F}, P)$ and a σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, a* regular conditional probability given $\mathcal{G}$ *is a map $\mu\colon \Omega \times \mathcal{F} \to [0,1]$ such that*

(1) *for all $\omega \in \Omega$, the map $A \mapsto \mu(\omega, A)$ is a probability measure on $(\Omega, \mathcal{F})$,*
(2) *for all $B \in \mathcal{F}$, the map $\omega \mapsto \mu(\omega, B)$ is $\mathcal{G}$-measurable,*
(3) *for all $X \in L^1(\Omega, \mathcal{F}, P)$, the set of $\omega \in \Omega$ such that*

$$X \in L^1(\mu(\omega, \cdot)) \tag{12.15}$$

and

$$E\big(X\big|\mathcal{G}\big)(\omega) = \int X(\omega')\mu(\omega, \mathrm{d}\omega') \tag{12.16}$$

is $\mathcal{G}$-measurable and of full P-measure.

Another name for μ is Markov kernel.

We immediately check:

Lemma 12.4 *Suppose that $\mathcal{F}$ is countably generated. If μ and $\tilde{\mu}$ are two regular conditional probabilities given $\mathcal{G}$, then there exists $\Omega' \in \mathcal{G}$ with $P(\Omega') = 1$ such that $\mu(\omega, A) = \tilde{\mu}(\omega, A)$ holds for all $\omega \in \Omega'$ and all $A \in \mathcal{F}$.*

Proof. For $A \in \mathcal{F}$, let $\Omega_A := \{\omega \in \Omega\colon\ \mu(\omega, A) = \tilde{\mu}(\omega, A)\}$. Then $\Omega_A \in \mathcal{G}$ and, by (12.16), $P(\Omega_A) = 1$. If $\mathcal{P}$ is the countable subset of $\mathcal{F}$ with $\sigma(\mathcal{P}) = \mathcal{F}$, then set $\Omega' := \bigcap_{A\in\mathcal{P}} \Omega_A$ and note that, by Dynkin's π/λ-theorem, $\mu(\omega, \cdot) = \tilde{\mu}(\omega, \cdot)$ for all $\omega \in \Omega'$. Since $\Omega' \in \mathcal{G}$ and $P(\Omega') = 1$, we are done. □

A typical setting when we may need a regular conditional probability is the probability space endowed with a special random variable, say Y, representing an initial value of a process or some other important quantity, and we wish to condition on its value. In other words, we are concerned with the special case $\mathcal{G} = \sigma(Y)$ for a real-valued Y. While this may sound reasonable, even in this case the existence of a map with the stated properties cannot be guaranteed in full generality. Here is a counterexample:

Lemma 12.5 (J. Dieudonne, 1947) *Denote $I := [0, 1]$ and set*

- $\Omega := I \times I$,
- $\mathcal{F} :=$ *Lebesgue measurable subset of* $I \times I$,
- $P :=$ *Lebesgue measure on* $(\Omega, \mathcal{F})$, *and*
- $\mathcal{G} := \{A \times I\colon\ A \in \mathcal{B}(I)\}$.

Then there exists no regular conditional probability given $\mathcal{G}$.

The proof will be based on the following fact:

Theorem 12.6 (G. Vitali 1905) *Let λ denote the Lebesgue measure on $(I, \mathcal{B}(I))$. Assuming the Axiom of Choice, the map $B \mapsto \lambda(B)$ does not extend to a measure on all subsets of I.*

Proof of Lemma 12.5. Continue writing λ for the Lebesgue measure on I and suppose μ is a map with the properties as in Definition 12.3. For each $A, B \in \mathcal{B}(I)$, let

$$\Omega_{A,B} := \Big\{(x, y) \in I \times I\colon\ \mu\big((x, y), A \times B\big) = 1_A(x)\lambda(B)\Big\} \tag{12.17}$$

Since $1_A(x)\lambda(B)$ equals the value of the map $(x, y) \mapsto 1_{A\times I}(x, y)\lambda(B)$, which is demonstrably $\mathcal{G}$-measurable, Definition 12.3(2) gives $\Omega_{A,B} \in \mathcal{G}$. Moreover, for any $C \in \mathcal{B}(I)$, the fact that the Lebesgue measure equals $\lambda \otimes \lambda$ on products of Borel sets gives

$$E\big(1_{C\times I} 1_{A\times I}\lambda(B)\big) = \lambda(C \cap A)\lambda(B) = E\big(1_{C\times I} 1_{A\times B}\big) \tag{12.18}$$

and so $(x, y) \mapsto 1_A(x)\lambda(B)$ is a version of $E(1_{A\times B}|\mathcal{G})$. In light of (12.16) it follows that $P(\Omega_{A,B}) = 1$ for all $A, B \in \mathcal{B}(I)$.

The σ-algebra $\mathcal{B}(I)$ is generated by the π-system $\mathcal{P} := \{[0, q]\colon\ q \in \mathbb{Q} \cap I\}$ which is countable. Denote

$$\Omega' := \bigcap_{A,B\in\mathcal{P}} \Omega_{A,B} \tag{12.19}$$

The standard argument based on Dynkin's π/λ-theorem now shows that equality in (12.17) holds on Ω' for all $A, B \in \mathcal{B}(I)$ simultaneously; i.e.,

$$\forall (x,y) \in \Omega' \, \forall A, B \in \mathcal{B}(I): \quad \mu\big((x,y), A \times B\big) = 1_A(x)\lambda(B) \tag{12.20}$$

Since $P(\Omega') = 1$, we can pick $(x^\star, y^\star) \in \Omega'$ and noting that $\{x^\star\} \in \mathcal{B}(I)$ get

$$\forall B \in \mathcal{B}(I): \quad \mu\big((x^\star, y^\star), \{x^\star\} \times B\big) = \lambda(B) \tag{12.21}$$

But $C \mapsto \mu((x^\star, y^\star), C)$ is a measure on Lebesgue measurable sets and, since $\{x^\star\} \times B$ is Lebesgue-null and thus Lebesgue measurable for any $B \subseteq I$, by (12.21) the map

$$B \mapsto \mu\big((x^\star, y^\star), \{x^\star\} \times B\big) \tag{12.22}$$

extends to a measure on all subsets of I. Since the map coincides with $\lambda(\cdot)$ on Borel sets, this contradicts Vitali's Theorem once we assume the Axiom of Choice. □

12.3 Existence result.

The counterexample above shows that the main obstruction to the existence of the conditional probability: there are too many P-null sets in $\mathcal{F}$. That null sets matter is not surprising because all we are trying to do is to put many Radon-Nikodym derivatives together in a measurable way. A way to get around this is by assuming additional structure on the underlying probability space, as in:

Theorem 12.7 *Suppose that Ω is a Polish space (i.e., a completely metrizable, second-countable topological space) and $\mathcal{F}$ is the associated Borel σ-algebra. Then for all σ-algebras $\mathcal{G} \subseteq \mathcal{F}$, a regular conditional probability given $\mathcal{G}$ exists. Moreover, if $\mathcal{G}$ is countably generated, then*

$$\forall A \in \mathcal{G}: \ \mu(\cdot, A) = 1_A \text{ a.s.} \tag{12.23}$$

(We say that μ is proper *if* (12.23) *holds.)*

While a complete proof can be constructed with no extraneous reference, we will cut the story short by relying on the following result from descriptive set theory:

Theorem 12.8 (Borel isomorphism theorem) *Let $\mathscr{X}$ be a Polish space and $\mathcal{B}(\mathscr{X})$ the associated Borel σ-algebra. Let*

$$\mathscr{Y} := \begin{cases} \{1, \dots, |\mathscr{X}|\}, & \text{if } \mathscr{X} \text{ is finite,} \\ \mathbb{N}, & \text{if } \mathscr{X} \text{ is infinite and countable,} \\ [0,1], & \text{if } \mathscr{X} \text{ is uncountable,} \end{cases} \tag{12.24}$$

which we endow with its natural Borel σ-algebra $\mathcal{B}(\mathscr{Y})$. Then there exists a bi-mesurable bijection $f: \mathscr{X} \to \mathscr{Y}$; i.e., a bijective map such that $f^{-1}(\mathcal{B}(\mathscr{Y})) = \mathcal{B}(\mathscr{X})$ and $f(\mathcal{B}(\mathscr{X})) = \mathcal{B}(\mathscr{Y})$.

We remark that a Polish space endowed with its Borel sets is called a *standard Borel space*. The above theorem states that Polish spaces are either finite, countable or of cardinality of the continuum and two such spaces are equivalent if and only if their cardinality is the same. What matters for us is that, in all three possible cases, every Polish space is equivalent to a Borel subset of $\mathbb{R}$.

Proof of Theorem 12.7. Theorem 12.8 guarantees existence of a random variable X mapping Ω bijectively and bi-mesasurably onto a Borel subset of $\mathbb{R}$. This implies

$$\forall A \in \mathcal{B}(\Omega)\colon\ X(A) \in \mathcal{B}(\mathbb{R}) \ \wedge\ \bigl(\forall \omega \in \Omega\colon\ X(\omega) \in X(A) \ \Leftrightarrow\ \omega \in A\bigr) \tag{12.25}$$

Let μ_X be the conditional distribution of X given $\mathcal{G}$ and set

$$\mu(\omega, A) := \mu_X\bigl(\omega, X(A)\bigr), \quad A \in \mathcal{G} \tag{12.26}$$

This immediately gives that $\omega \mapsto \mu(\omega, A)$ is $\mathcal{G}$-measurable. The fact that X is bijective means that $A \mapsto X(A)$ maps disjoint unions into disjoint unions which along with the first part (12.25) implies that $A \mapsto \mu(\omega, A)$ is a probability measure. Thanks to (12.25) we also get

$$\mu_X\bigl(\cdot, X(A)\bigr) = E\bigl(1_{X \in X(A)} \,\big|\, \mathcal{G}\bigr) = E(1_A|\mathcal{G}) \text{ a.s.} \tag{12.27}$$

which then gives (12.15–12.16) by standard extension arguments. It follows that μ is a regular conditional probability given $\mathcal{G}$.

In order to prove the second part, note that for all $A \in \mathcal{G}$,

$$\Omega_A := \bigl\{\omega \in \Omega\colon\ \mu(\omega, A) = 1_A(\omega)\bigr\} \tag{12.28}$$

is a $\mathcal{G}$-measurable set with $P(\Omega_A) = 1$. If $\mathcal{G}$ is generated by a countable collection $\mathcal{P}$, we can take $\Omega' := \bigcap_{A \in \mathcal{P}} \Omega_A$ and observe that

$$\mathcal{L} := \Bigl\{A \in \mathcal{G}\colon\ \bigl(\forall \omega \in \Omega'\colon\ \mu(\omega, A) = 1_A(\omega)\bigr)\Bigr\} \tag{12.29}$$

is a σ-algebra containing $\mathcal{P}$. Hence $\mathcal{L} = \mathcal{G}$ and $\mu(\cdot, A) = 1_A$ holds for all $A \in \mathcal{G}$ on Ω'. Since $P(\Omega') = 1$, this is (12.23). □

The condition that the underlying space is Polish is usually not too restrictive for most applications. However, what one should be careful about is the fact demonstrated in Lemma 12.5: Working with a completed measure may ruin the existence of conditional probabilities. (This sometimes matters when one discusses stochastic calculus from the point of view of Markov processes.)

12.4 Extension of measures in product spaces.

The existence of conditional probabilities is central in the theory of disintegration of measures and integrals. Standard Borel spaces shows up in other important conclusions in probability, one of which is the *Kolmogorov Extension Theorem*. This theorem says that, given a standard-Borel space $(\mathscr{X}, \mathcal{F})$ and, for each $n \geqslant 1$, a probability measure μ_n on $(\mathscr{X}^n, \mathcal{F}^{\otimes n})$ such that the consistency condition

$$\forall n \geqslant 1\, \forall A \in \mathcal{F}^{\otimes n}\colon \quad \mu_{n+1}(A \times \mathscr{X}) = \mu_n(A) \tag{12.30}$$

holds, there exists a unique probability measure μ on $(\mathscr{X}^{\mathbb{N}}, \mathcal{F}^{\otimes \mathbb{N}})$ that reduces to μ_n on $(\mathscr{X}^n, \mathcal{F}^{\otimes n})$ under the natural embedding of these spaces.

While it is known that such an extension may fail in general, the assumption that $(\mathscr{X}, \mathcal{F})$ is standard Borel is definitely not necessary. Indeed, alternatively it suffices that μ_{n+1} disintegrates with respect to μ_n and that the corresponding conditional measure is very regular. Here is a precise statement:

Theorem 12.9 (Extension of measures in product spaces) *Let $(\mathscr{X}, \mathcal{F})$ be a measurable space. Assume that, for each $n \geqslant 1$, a probability measure μ_n on $(\mathscr{X}^n, \mathcal{F}^{\otimes n})$ is given such that*

$$\forall A \in \mathcal{F}^{\otimes(n+1)}: \quad \mu_{n+1}(A) = \int \mathfrak{q}_n(x, A)\mu_n(\mathrm{d}x) \tag{12.31}$$

where $\mathfrak{q}_n \colon \mathscr{X}^n \times \mathcal{F}^{\otimes(n+1)} \to [0,1]$ is $\mathcal{F}^{\otimes n}$-measurable in the first coordinate, a probability measure on $(\mathscr{X}^{n+1}, \mathcal{F}^{\otimes(n+1)})$ in the second coordinate and obeys

$$\forall x \in \mathscr{X}^n \, \forall A \in \mathcal{F}^{\otimes n} \colon \ \mathfrak{q}_n(x, A \times \mathscr{X}) = 1_A(x) \tag{12.32}$$

Then there exists a probability measure μ on $(\mathscr{X}^{\mathbb{N}}, \mathcal{F}^{\otimes\mathbb{N}})$ that restricts to μ_n on $(\mathscr{X}^n, \mathcal{F}^{\otimes n})$ under the natural embedding of $\mathscr{X}^n$ into $\mathscr{X}^{\mathbb{N}}$.

Proof. We will identify $\mathcal{F}^{\otimes n}$ with the σ-algebra

$$\mathcal{F}_n := \sigma\biggl(\Bigl\{A_1 \times \cdots \times A_n \times \mathscr{X} \times \mathscr{X} \times \ldots \colon A_1, \ldots, A_n \in \mathcal{F}\Bigr\}\biggr) \tag{12.33}$$

on $\mathscr{X}^{\mathbb{N}}$ which permits us to regard μ_n as a measure on $(\mathscr{X}^{\mathbb{N}}, \mathcal{F}_n)$. The kernels $\mathfrak{q}_n$ are then to be interpreted as $\mathcal{F}_n$-measurable maps on $\mathscr{X}^{\mathbb{N}} \times \mathcal{F}_{n+1} \to [0,1]$. This means that $\mathfrak{q}_m(x, \cdot)$ depends only on the first m coordinates of x.

The union $\mathcal{A} := \bigcup_{n\geqslant 1} \mathcal{F}_n$ is an algebra and, since (12.31) implies consistency, there exists a finitely-additive set function $\bar{\mu}$ on $\mathcal{A}$ that restricts to μ_n on $\mathcal{F}_n$. If we can show that $\bar{\mu}$ is countably subadditive, the Hahn-Kolmogorov theorem (Theorem 1.4) implies that $\bar{\mu}$ extends to a measure on $\mathcal{F} := \sigma(\mathcal{A})$. For this it suffices to show that, given any $\{B_n\}_{n\geqslant 1}$ in $\mathcal{A}$,

$$B_n \downarrow \varnothing \ \Rightarrow \ \lim_{n\to\infty} \bar{\mu}(B_n) = 0 \tag{12.34}$$

Relabeling the sets if necessary, we may and will assume that $B_n \in \mathcal{F}^{\otimes n}$ for each $n \geqslant 1$.

We will prove the contrapositive but first we need some preparations. We start by extending $\mathfrak{q}_n$ into a two-parameter quantity $\{\mathfrak{q}_{m,n}\}_{1\leqslant m\leqslant n}$ by setting

$$\mathfrak{q}_{m,m+1}(x, A) := \mathfrak{q}_m(x, A), \quad A \in \mathcal{F}_{m+1} \tag{12.35}$$

and, recursively,

$$\mathfrak{q}_{m,n+1}(x, A) := \int \mathfrak{q}_{n+1}(z, A)\mathfrak{q}_{m,n}(x, \mathrm{d}z), \quad A \in \mathcal{F}_{n+1} \tag{12.36}$$

By induction on m we then check that, for $m < n$,

$$\mathfrak{q}_{m,n}(x, A) = \int \mathfrak{q}_{m+1,n}(z, A)\mathfrak{q}_m(x, \mathrm{d}z), \quad A \in \mathcal{F}_n \tag{12.37}$$

We now define

$$h_{m,n}(x) := \mathfrak{q}_{m,n}(x, B_n) \tag{12.38}$$

and note that (12.32) gives $\mathfrak{q}_m(x, B_n) = 1_{B_n}(x)$ when $n < m$.

We now observe some properties of functions $h_{m,n}$. First note that (12.37) gives

$$h_{m,n}(x) = \int h_{m+1,n}(z)\mathfrak{q}_m(x, \mathrm{d}z) \tag{12.39}$$

On the other hand, (12.36) and (12.32) imply that $h_{m,n} \geqslant h_{m,n+1}$ and so

$$h_m(x) := \lim_{n\to\infty} h_{m,n}(x) \tag{12.40}$$

exists in $[0,1]$. Invoking this in (12.39) with the help of the Bounded Convergence Theorem shows

$$h_m(x) = \int h_{m+1}(z)\mathfrak{q}_m(x,\mathrm{d}z) \tag{12.41}$$

Since (12.31) gives $\int h_{1,n}\mathrm{d}\mu_m = \mu_n(B_n) = \bar{\mu}(B_n)$ we also have

$$\lim_{n\to\infty} \bar{\mu}(B_n) = \int h_1 \mathrm{d}\mu_1 \tag{12.42}$$

by another application of the Bounded Convergence Theorem.

Let us now suppose that $\inf_{n\geqslant 1} \bar{\mu}(B_n) > \delta > 0$ for each $n \geqslant 1$. Then (12.42) gives existence of $x_1 \in \mathscr{X}^{\mathbb{N}}$ such that $h_1(x_1) > \delta$. Next observe that $h_m(x) > \delta$ along with (12.41) dictates

$$\mathfrak{q}_m\Big(x, \big\{z \in \mathscr{X}^{\mathbb{N}} \colon h_{m+1}(z) > \delta\big\}\Big) > 0 \tag{12.43}$$

and, by (12.32), there exists $z \in \mathscr{X}^{\mathbb{N}}$ that agrees with x in the first m coordinates such that $h_{m+1}(z) > \delta$. Using the Axiom of Choice, the above observations identify recursively a sequence $\{x_m\}_{m\geqslant 1}$ such that x_n and x_m agree on the first $\min\{m,n\}$ coordinates for each $m,n \geqslant 1$ and such that $h_m(x_m) > \delta$ for each $m \geqslant 1$.

Taking $\bar{x}$ whose n-th coordinate agrees with the n-coordinate of x_n, the fact that h_n depends only on the first n coordinates of its argument implies $h_m(\bar{x}) > \delta$ for each $m \geqslant 1$. But then also $1_{B_n}(\bar{x}) \geqslant h_n(\bar{x}) > \delta$ and so $\bar{x} \in B_n$ for all $n \geqslant 1$. This shows that $\inf_{n\geqslant 1} \bar{\mu}(B_n) > 0$ implies $\bigcap_{n\geqslant 1} B_n \neq \varnothing$, proving the contrapositive of (12.34). □

Recall that the proof of the Kolmogorov Extension Theorem also proceeds by proving (12.34) by contrapositive, but the non-emptiness of the intersection is inferred using tools from topology; namely, the Cantor intersection property applied to compact subsets of $\mathscr{X}^n$ whose existence is inferred by inner regularity of μ_n.

While condition (12.32), which was key in several steps of the above proof, is quite restrictive it does hold in several interesting cases of prime interest; most notably, when $\{\mu_n\}_{n\geqslant 1}$ are product measures or when they are generated by a Markov chain. The above result thus offers an independent approach to the construction of infinite product measures and general-valued Markov chains. Note, however, that the Axiom of Choice was used inside the proof.

Further reading: Durrett, Section 4.1.3

13. Introduction to martingales

The concept of conditional expectation is useful in analyzing dependent random variables. We will still need that this dependence has a certain structure. In the rest of 275B course, we will explore two specific cases of such structures go under the respective banners "Martingales" and "Markov chains" starting with the former here.

13.1 Definitions and examples.

We begin by introducting concepts that will be useful throughout.

Definition 13.1 (Stochastic process) *A stochastic process (indexed by naturals) or just a process is a sequence of random variables $\{X_n\}_{n\geqslant 0}$ defined on the same probability space. If indexing is clear from context, we may refer to $\{X_n\}_{n\geqslant 0}$ as $\{X_n\}$ or just X.*

Definition 13.2 (Filtration) *Given a probability space $(\Omega, \mathcal{F}, P)$, a sequence $\{\mathcal{F}_n\}_{n\geqslant 0}$ of sub-σ-algebras of $\mathcal{F}$ is said to be a* filtration *if*

$$\forall n \geqslant 0\colon \ \mathcal{F}_n \subseteq \mathcal{F}_{n+1} \tag{13.1}$$

Definition 13.3 (Adapted process) *A process $\{X_n\}_{n\geqslant 0}$ is said to be* adapted to filtration *$\{\mathcal{F}_n\}_{n\geqslant 0}$, or just* adapted*, if X_n is $\mathcal{F}_n$-measurable for each $n \geqslant 0$.*

We should think of $\mathcal{F}_n$ as the information available about our stochastic setting at "time" n. A natural choice of the filtration that makes a process $\{X_k\}_{k\geqslant 0}$ adapted is

$$\mathcal{F}_n := \sigma(X_0, \ldots, X_n) \tag{13.2}$$

As this is also the smallest filtration with this property and is generated by X, we sometimes write it as $\mathcal{F}_n^X$.

Definition 13.4 (Martingale) *Given a filtration $\{\mathcal{F}_n\}_{n\geqslant 0}$ and random variables $\{M_n\}_{n\geqslant 0}$, the pair $\{M_n, \mathcal{F}_n\}_{n\geqslant 0}$ is called a* martingale *if*

(1) *$\{M_n\}_{n\geqslant 0}$ is adapted to $\{\mathcal{F}_n\}_{n\geqslant 0}$ and*
(2) *for all $n \geqslant 0$ we have $M_{n+1} \in L^1$ with*

$$E(M_{n+1}|\mathcal{F}_n) = M_n \quad \text{a.s.} \tag{13.3}$$

If instead the same inequality holds in (13.3) *for all $n \geqslant 0$ we call the sequence $\{M_n, \mathcal{F}_n\}_{n\geqslant 0}$*

- *a submartingale if $E(M_{n+1}|\mathcal{F}_n) \geqslant M_n$ for all $n \geqslant 0$,*
- *a supermartingale if $E(M_{n+1}|\mathcal{F}_n) \leqslant M_n$ for all $n \geqslant 0$.*

A martingale is thus a submartingale and a supermartingale.

A natural way to think of a martingale is the winning in a "fair" game. Indeed, the condition (13.3) can also be written as

$$E(M_{n+1} - M_n|\mathcal{F}_n) = 0 \tag{13.4}$$

so if M_n represents the amount of money a player won (if $M_n \geqslant 0$) or lost (if $M_n < 0$) at time n, then the martingale property means that there is no bias in the game, at least in terms of expectation. (The *Gambler's ruin* problem analysis shows that there is effectively a bias if you play against a more wealthy opponent.)

Here is a remark on notation. To reduce clutter, we will often write $\{\mathcal{F}_n\}$ instead of $\{\mathcal{F}_n\}_{n\geqslant 0}$ and $\{M_n, \mathcal{F}_n\}$ instead of $\{M_n, \mathcal{F}_n\}_{n\geqslant 0}$ unless confusion may arise. We may even suppress the reference to the filtration altogether when it is clear from the context and even write just M for $\{M_n\}$.

Here are some simple examples of martingales:

***Example* 13.5** (Additive martingale) Let $\{X_k\}_{k\geqslant 1}$ be independent with

$$\forall k \geqslant 1\colon\ X_k \in L^1 \ \wedge\ EX_k = 0 \tag{13.5}$$

Set $M_n := X_1 + \cdots + X_n$ with $M_0 := 0$ and let $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$. Then $\{\mathcal{F}_n\}$ is a filtration and M_n is $\mathcal{F}_n$-measurable for each $n \geqslant 0$. Moreover,

$$E(M_{n+1}|\mathcal{F}_n) = M_n + E(X_{n+1}|\mathcal{F}_n) = M_n \quad \text{a.s.} \tag{13.6}$$

because X_{n+1} is independent of $\mathcal{F}_n$ and thus $E(X_{n+1}|\mathcal{F}_n) = E(X_{n+1}) = 0$. Hence $\{M_n, \mathcal{F}_n\}$ is a martingale. Clearly, if instead $\forall k \geqslant 1\colon\ E(X_k) \geqslant 0$, then $\{M_n, \mathcal{F}_n\}$ is a submartingale.

***Example* 13.6** (Multiplicative martingale) Let $X_1, X_2, \ldots$ be independent with

$$\forall k \geqslant 1\colon\ X_k \in L^1 \ \wedge\ X_k \geqslant 0 \ \wedge\ EX_k = 1 \tag{13.7}$$

Set $M_n := \prod_{k=1}^n X_k$ with $M_0 := 1$ and let $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$. Then again M_n is $\mathcal{F}_n$-measurable for each $n \geqslant 0$ and

$$E(M_{n+1}|\mathcal{F}_n) = M_n E(X_{n+1}|\mathcal{F}_n) = M_n \quad \text{a.s.} \tag{13.8}$$

Hence $\{M_n, \mathcal{F}_n\}$ is again a martingale.

Here is another way to generate a martingale:

***Example* 13.7** (Progressive conditioning) Let $X \in L^1$ and let $\{\mathcal{F}_n\}$ be a filtration. Set

$$M_n := E(X|\mathcal{F}_n) \tag{13.9}$$

Then

$$E(M_{n+1}|\mathcal{F}_n) = E\big(E(X|\mathcal{F}_{n+1})\big|\mathcal{F}_n\big) = E(X|\mathcal{F}_n) = M_n \quad \text{a.s.} \tag{13.10}$$

by the "Smaller always wins" principle. Hence $\{M_n, \mathcal{F}_n\}$ is a martingale.

Conditioning on filtration is a way to reveal more information about the stochastic setting. Note that a "finite run" of a martingale is always of the above form; indeed, if $\{M_k, \mathcal{F}_k\}_{k=0}^n$ is a martingale, then

$$\forall k = 0, \ldots, n\colon\ M_k = E(X_n|\mathcal{F}_k) \quad \text{a.s.} \tag{13.11}$$

It is an interesting question whether (and under what conditions) an infinite martingale sequence can be expressed this way. This will be resolved when we introduce the concept of uniform integrability.

13.2 Derived examples.

In order to give our next two examples, we note:

Lemma 13.8 *Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be convex and let $\{M_n\}_{n\geqslant 0}$ be a process such that $\phi(M_n) \in L^1$ for each $n \geqslant 0$. Then*

(1) *$\{M_n, \mathcal{F}_n\}$ a martingale $\Rightarrow$ $\{\phi(M_n), \mathcal{F}_n\}$ a submartingale.*
(2) *$\{M_n, \mathcal{F}_n\}$ a submartingale and ϕ is non-decreasing $\Rightarrow$ $\{\phi(M_n), \mathcal{F}_n\}$ a submartingale.*

Proof. By Jensen's inequality

$$E\big(\phi(M_{n+1})\big|\mathcal{F}_n\big) \geqslant \phi\big(E(M_{n+1}|\mathcal{F}_n)\big) \tag{13.12}$$

The right hand side equals $\phi(M_n)$ when $\{M_n, \mathcal{F}_n\}$ is a martingale. If the latter is only a submartingale, then we still have

$$\phi\big(E(M_{n+1}|\mathcal{F}_n)\big) \geqslant \phi(M_n) \tag{13.13}$$

as soon as ϕ is non-decreasing. □

This lemma stimulates additional examples of interest:

***Example* 13.9** (Variance martingale) Lemma 13.8 says that if $\{M_n, \mathcal{F}_n\}$ is a martingale and $\forall n \geqslant 1\colon M_n \in L^2$, then $\{M_n^2, \mathcal{F}_n\}$ is a submartingale. However, we can say more if M is an additive martingale, i.e., $M_n := X_1 + \cdots + X_n$ for independent $X_1, X_2, \ldots$ with $\forall i \geqslant 1\colon X_i \in L^1 \wedge EX_i = 0$ and $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$. Indeed, assuming $\forall i \geqslant 1\colon X_i \in L^2$, we can compute the "martingale defect" as

$$\begin{aligned} E(M_{n+1}^2|\mathcal{F}_n) &= E\big(M_n^2 + 2M_nX_{n+1} + X_{n+1}^2\big|\,\mathcal{F}_n\big) \\ &= M_n^2 + 2M_nE(X_{n+1}|\mathcal{F}_n) + E(X_{n+1}^2|\mathcal{F}_n) = M_n^2 + E(X_{n+1}^2) \end{aligned} \tag{13.14}$$

This suggest setting

$$\widetilde{M}_n := M_n^2 - \sum_{k=1}^{n} E(X_i^2) \tag{13.15}$$

for which we readily check that $\{\widetilde{M}_n, \mathcal{F}_n\}$ is a martingale.

The calculation in the previous example used that the martingale increments were independent, but pretty much the same can be done even if they are not. We start by introducing a convenient terminology:

Definition 13.10 *We say that $M = \{M_n\}$ is an L^p-martingale if $\forall n \geqslant 0\colon M_n \in L^p$.*

We then have:

Lemma 13.11 *Let $\{M_n, \mathcal{F}_n\}$ be an L^2-martingale. Set*

$$\widetilde{M}_n := M_n^2 - \sum_{k=1}^{n} E\big((M_k - M_{k-1})^2\,\big|\,\mathcal{F}_k\big) \tag{13.16}$$

Then $\{\widetilde{M}_n, \mathcal{F}_n\}$ is a martingale.

Proof. The process $\{\widetilde{M}_n\}$ is clearly adapted to $\{\mathcal{F}_n\}$. Noting that

$$M_{n+1}^2 - M_n^2 = (M_{n+1} - M_n)^2 + 2M_n(M_{n+1} - M_n) \tag{13.17}$$

we get

$$\begin{aligned} E(M_{n+1}^2|\mathcal{F}_n) - M_n^2 &= E\big((M_{n+1}-M_n)^2 \,\big|\, \mathcal{F}_n\big) + 2M_n E(M_{n+1}-M_n|\mathcal{F}_n)_n \\ &= E\big((M_{n+1}-M_n)^2 \,\big|\, \mathcal{F}_n\big) \quad \text{a.s.} \end{aligned} \tag{13.18}$$

where all conditional expectations are defined because $M_n, M_n - M_{n-1} \in L^2$. We now check by induction that this gives $E(\widetilde{M}_{n+1}|\mathcal{F}_n) = \widetilde{M}_n$ a.s. □

We will elaborate on the setting of Lemma 13.11 further when we discuss so called Doob's decomposition. The calculation in the previous proof suggests another way to build a martingale out of a martingale. We need:

Definition 13.12 (Predictable process) *Given a filtration $\{\mathcal{F}_n\}$, a process $\{C_n\}$ is said to be* predictable *if*

$$C_0 := 0 \ \wedge \ \forall n \geqslant 1\colon \ C_n \text{ is } \mathcal{F}_{n-1}\text{-measurable} \tag{13.19}$$

Definition 13.13 *Given two processes $C = \{C_n\}$ and $X := \{X_n\}$, we define $C \cdot X$ to be the process $\{(C \cdot X)_n\}_{n\geqslant 0}$ where*

$$(C \cdot X)_n := \sum_{k=1}^{n} C_k(X_k - X_{k-1}), \quad n \geqslant 1 \tag{13.20}$$

and $(C \cdot X)_0 := 0$. We sometimes refer to $(C \cdot X)_n$ as a discrete stochastic integral.

The term predictable reflects on the fact that the value of C_n is already "known" at "time" $n-1$. A natural interpretation of $C \cdot X$ is the outcome of a *betting strategy* in a game encoded by X. Indeed, if a player bets C_{n+1} dollars in the round where betting one dollar earns her $X_{n+1} - X_n$, her winnings in that round will be

$$C_{n+1}(X_{n+1} - X_n) \tag{13.21}$$

The quantity $(C \cdot X)_n$ is thus the total amount won by time n.

We now observe that, unless the player can predict the future, betting on a fair game results in a fair game:

Lemma 13.14 *Given a filtration $\{\mathcal{F}_n\}$, let $C := \{C_n\}$ be a predictable process and $M := \{M_n\}$ be a process. Assume $\forall n \geqslant 0\colon C_n \in L^\infty$. Then*

$$M \text{ martingale} \quad \Rightarrow \quad C \cdot M \text{ martingale} \tag{13.22}$$

and

$$M \text{ submartingale} \ \wedge \forall n \geqslant 1\colon C_n \geqslant 0 \quad \Rightarrow \quad C \cdot M \text{ submartingale} \tag{13.23}$$

Proof. The assumptions ensure that $(C \cdot M)_n \in L^1$ for all $n \geqslant 1$. The construction gives that $C \cdot M$ is adapted. Using the same calculation as in the previous proof, we thus get

$$\begin{aligned} E\big((C \cdot M)_{n+1} \,\big|\, \mathcal{F}_n\big) &= (C \cdot M)_n + E\big(C_n(M_{n+1} - M_n) \,\big|\, \mathcal{F}_n\big) \\ &= (C \cdot M)_n + C_n E(M_{n+1} - M_n \,|\, \mathcal{F}_n) \end{aligned} \tag{13.24}$$

where we used that C is predictable in the second step. The last term vanishes when M is a martingale and is non-negative if M is a submartingale and $C_n \geqslant 0$. □

We now return back to effects of composing a martingale with convex function:

***Example* 13.15** (Exponential martingale) Another way to turn an additive martingale into an interesting submartingale is by applying the exponential function. Indeed, suppose that $\{X_n\}_{n\geqslant 1}$ are independent and such that $E(\mathrm{e}^{\lambda X_n}) < \infty$ for all $n \geqslant 1$. If also $X_n \in L^1$ and $EX_n = 0$ for all $n \geqslant 1$, then $M_n := X_1 + \cdots + X_n$ is a martingale. Lemma 13.8 then tells us that $\{\mathrm{e}^{\lambda M_n}\}$ is a submartingale.

Thanks to the underlying independence structure we can get more. Indeed,

$$E\big(\mathrm{e}^{\lambda M_{n+1}} \,\big|\, \mathcal{F}_n\big) = E\big(\mathrm{e}^{\lambda M_n}\mathrm{e}^{\lambda X_{n+1}} \,\big|\, \mathcal{F}_n\big) = \mathrm{e}^{\lambda M_n} E\big(\mathrm{e}^{\lambda X_{n+1}}\big) \tag{13.25}$$

where Lemma 11.3 was used to get the second equality, and so setting

$$\widetilde{M}_n := \mathrm{e}^{\lambda M_n} \prod_{k=1}^{n} \varphi_k(\lambda)^{-1} \quad \text{where} \quad \varphi_n(\lambda) := E(\mathrm{e}^{\lambda X_n}) \tag{13.26}$$

we get that $\{\widetilde{M}_n\}$ is a martingale.

13.3 Two "practical" examples.

We will close our introduction by giving two examples that concern "practical" models that are of much interest throughout probability. What makes them interesting for us that they also give rise to natural martingales.

***Example* 13.16** (Galton-Watson branching process) In mid 19th century, there was some interest in developping a theory of family trees — particularly, in connection with royal families. Two statisticians (by standards back then) F. Galton and H.W. Watson (and also independently I.J. Bienaymé) devised a model that nowadays serves as a foundation for such considerations.

In the (Bienaymé)-Galton-Watson model, there are generations indexed by natural-valued "time." Each generation contains a certain number of currently living individuals. The population dynamics is such that, at each time, each individual produces a certain number of off-spring, which is sampled independently from a common law on $\mathbb{N}$ with probability mass function $\{\mathfrak{p}(n)\colon n \geqslant 0\}$, called the *off-spring distribution*. (In particular, if there is no off-spring, the lineage of that individual dies out.) The number of individuals in the next generation is then the sum of all off-spring counts for the current generation.

We will formalize the problem as follows. Consider a family of i.i.d. integer-valued random variables $\{X_{n,k}\colon n, k \geqslant 0\}$ with law determined by

$$P(X_{n,k} = m) = \mathfrak{p}(m), \qquad m \geqslant 0. \tag{13.27}$$

Define, inductively, random variables $\{S_n\}$ as follows: $S_0 := 1$ and

$$S_{n+1} := \begin{cases} 0, & \text{if } S_n = 0, \\ X_{n+1,1} + \cdots + X_{n+1,S_n}, & \text{if } S_n > 0. \end{cases} \tag{13.28}$$

It is easy to verify that the dynamics does what we described above. The additional assumption $S_0 := 1$ means that there is one individual at time zero.

Consider now the filtration $\mathcal{F}_n := \sigma(X_{m,k}\colon 0 \leqslant m \leqslant n,\, k \geqslant 0)$ and let us compute:

$$E(S_{n+1}|\mathcal{F}_n) = \begin{cases} 0, & \text{on } \{S_n = 0\}, \\ [E(X_{1,1})]S_n, & \text{on } \{S_n > 0\}. \end{cases} \tag{13.29}$$

Thus, denoting $\mu := E(X_{1,1})$ we get that $M_n := \mu^{-n} S_n$ defines a martingale. The value $\mu = 1$ is obviously special because $\{S_n\}$ is then itself a martingale.

***Example* 13.17** (Polya's Urn) Our next example concerns an urn problem. Fix integers $r, g \geqslant 1$ and $b \geqslant 1$ and consider an urn that, initially, has r red balls and g green balls in it. Sample a ball from the urn and put it back along with b balls of the same color. (NOTE: We changed parametrization compared to the lecture!) Repeating this step over and over, the question is what is the fraction of the red balls in the urn in the long run.

Let R_n denote the number of red balls in the urn at time n and let G_n denote the corresponding number of green balls. Obviously, $R_n + G_n = r + g + bn$. Now consider the random variable

$$M_n := \frac{R_n}{R_n + G_n} = \frac{R_n}{r + g + bn} \tag{13.30}$$

which denotes the fraction of red balls in the urn.

We will encode the dynamics of the urn using a sequence $U_1, U_2, \ldots$ of i.i.d. uniform random variables on $[0, 1]$ by way of the recursions

$$R_{n+1} := (R_n + b)1_{\{U_{n+1} \leqslant M_n\}} + R_n 1_{\{U_{n+1} > M_n\}}. \tag{13.31}$$

and

$$G_n := r + g + bn - R_n \tag{13.32}$$

subject to the initial value $R_0 := r$ (which gives $G_0 = g$). If set this way, we have:

Lemma 13.18 *$\{M_n, \mathcal{F}_n\}$ is a martingale for the filtration $\mathcal{F}_n := \sigma(U_1, \ldots, U_n)$.*

Proof. Abbreviate $N_n := r + g + bn$. Since R_n and M_n are $\mathcal{F}_n$-measurable, we have

$$\begin{aligned} E(M_{n+1}\,|\mathcal{F}_n) &= \frac{R_n + b}{N_{n+1}} E(1_{\{U_{n+1} \leqslant M_n\}}|\mathcal{F}_n) + \frac{R_n}{N_{n+1}} E(1_{\{U_{n+1} > M_n\}}|\mathcal{F}_n) \\ &= \frac{R_n + b}{N_{n+1}} \frac{R_n}{N_n} + \frac{R_n}{N_{n+1}} \frac{G_n}{N_n} = \frac{R_n}{N_n} \frac{R_n + bn + G_n}{N_{n+1}} = \frac{R_n}{N_n} = M_n \end{aligned} \tag{13.33}$$

implying the claim. □

We will keep returning to the above examples in the the forthcoming lectures.

Further reading: Durrett, Sections 4.2 and 4.3

14. Optional Stopping and Sampling

In this section we will elaborate on the following property of martingales: Suppose that $\{M_n, \mathcal{F}_n\}$ is a martingale. Then

$$E(M_{n+1}) = E\big(E(M_{n+1}|\mathcal{F}_n)\big) = E(M_n) = \underbrace{\dots}_{\text{induction}} = E(M_0) \tag{14.1}$$

and so $E(M_n)$ does not depend on n. (Similarly, for submartingales we get $E(M_n) \geqslant E(M_0)$ and for supermartinagles we get $E(M_n) \leqslant E(M_0)$.) The question to address here is when we have $E(M_T) = E(M_0)$ for a random time T.

14.1 Optional stopping theorem.

Of course, not just any random time will do because we could simply let the sequence run until it becomes larger than M_0 or smaller than M_0 (one of these will occur a.s. for every non-constant sequence). In short, the random time should not be allowed to "look into the future." This leads to the following concept:

Definition 14.1 (Stopping time) *A random variable T taking values in $\mathbb{N} \cup \{+\infty\}$ is a* stopping time for filtration $\{\mathcal{F}_n\}$ *if*

$$\forall n \geqslant 0\colon \quad \{T \leqslant n\} \in \mathcal{F}_n \tag{14.2}$$

We leave it to the reader to check that equivalent definition is obtained when $\{T \leqslant n\}$ is replaced by $\{T > n\}$ or by $\{T = n\}$. Here is one standard example:

Lemma 14.2 (First hitting time of a set) *Let $\{X_n\}$ be an S-valued process adapted to filtration $\{\mathcal{F}_n\}$. Then for all measurable sets $A \subseteq S$ and any stopping time T' for $\{\mathcal{F}_n\}$,*

$$T := \inf\{n \geqslant T'\colon X_n \in A\} \tag{14.3}$$

is a stopping time.

Proof. For each $n \in \mathbb{N}$ we have

$$\{T > n\} = \bigcap_{k=0}^{n} \{X_k \notin A\} \cap \{T' \leqslant k\} \tag{14.4}$$

As X is adapted and T' is a stopping time, $\{X_k \notin A\} \cap \{T' \leqslant k\} \in \mathcal{F}_k \subseteq \mathcal{F}_n$. Hence we get $\{T \leqslant n\} \in \mathcal{F}_n$ for all $n \geqslant 0$. □

Note that a deterministic time is a stopping time, so we can take T' to be one above. The formulation with T' random allow us to iterate this definition. Stopping times are preserved under a number of operations. Here is a list:

Lemma 14.3 *Let T and S be stopping times. Then so are $T \wedge S$, $T \vee S$ and $T + S$.*

Proof. For each $n \in \mathbb{N}$ we have

$$\{T \wedge S \leqslant n\} = \{T \leqslant n\} \cap \{S \leqslant n\} \tag{14.5}$$

and

$$\{T \vee S \leqslant n\} = \{T \leqslant n\} \cup \{S \leqslant n\} \tag{14.6}$$

proving that $T \wedge S$ and $T \vee S$ are stopping times. For the sum we have

$$\{T + S \leqslant n\} = \bigcup_{m=0}^{n} \{T = m\} \cap \{S \leqslant n - m\} \tag{14.7}$$

and note that $\{T \leqslant m\} = \{T \leqslant m\} \smallsetminus \{T \leqslant m-1\} \in \mathcal{F}_m \subseteq \mathcal{F}_n$ when $m \leqslant n$. □

We now return back to the question of whether martingales evaluated at stopping times are still martingales.

Lemma 14.4 (Stopped (sub)martingale) *Suppose $\{M_n, \mathcal{F}_n\}$ is a (sub)martingale and T is a stopping time for $\{\mathcal{F}_n\}$. Then*

$$\forall n \in \mathbb{N}\colon M_{T\wedge n} \in L^1 \quad \wedge \quad \{M_{T\wedge n}, \mathcal{F}_n\} \text{ is a (sub)margingale} \tag{14.8}$$

Proof. Since $T \wedge n$ is $\mathcal{F}_n$-measurable, so is $M_{T\wedge n}$ due to the representation

$$M_{T\wedge n} = \sum_{k=0}^{n} M_k 1_{\{T=k\}} + M_n 1_{\{T>n\}} \tag{14.9}$$

Moreover,

$$|M_{T\wedge n}| \leqslant \sum_{k=0}^{n} |M_k| \tag{14.10}$$

and so $M_{T\wedge n} \in L^1$. For the proof of (sub)martingale property, we need the identities

$$M_{T\wedge n} = M_{T\wedge n} 1_{\{T\leqslant n\}} + M_n 1_{\{T>n\}} \tag{14.11}$$

and

$$M_{T\wedge (n+1)} = M_{T\wedge n} 1_{\{T\leqslant n\}} + M_{n+1} 1_{\{T>n\}} \tag{14.12}$$

For $\{M_n, \mathcal{F}_n\}$ a martingale, the fact that all object on the right of these except M_{n+1} are $\mathcal{F}_n$-measurable shows

$$\begin{aligned} E\big(M_{T\wedge(n+1)}\big|\mathcal{F}_n\big) &= M_{T\wedge n} 1_{\{T\leqslant n\}} + 1_{\{T>n\}} E(M_{n+1}|\mathcal{F}_n) \\ &= M_{T\wedge n} 1_{\{T\leqslant n\}} + M_n 1_{\{T>n\}} = M_{T\wedge n} \end{aligned} \tag{14.13}$$

For submartinagles the second equality is changed into $\geqslant$. □

We are now ready for the first main conclusion:

Theorem 14.5 (Doob's Optional Stopping Theorem) *Let $\{M_n, \mathcal{F}_n\}$ be a martingale and T a stopping time for $\{F_n\}$. Suppose there exists $K \in (0, \infty)$ such that at least one of the conditions*

(1) $T \leqslant K$ a.s.
(2) $T < \infty$ a.s. $\wedge\ \forall n \geqslant 0\colon |M_n| \leqslant K$ a.s.
(3) $ET < \infty \wedge \forall n \geqslant 0\colon |M_{n+1} - M_n| \leqslant K$ a.s.

holds true. Then

$$M_T \in L^1 \quad \wedge \quad E(M_T) = E(M_0) \tag{14.14}$$

Proof. By the previous lemma we have $M_{T\wedge n} \in L^1$ and $E(M_{T\wedge n}) = E(M_0)$. When $T \leqslant K$ a.s., setting $n > K$ makes truncation by n superfluous and so the claim follows.

For (2-3) we will now obtain statements by using an appropriate convergence theorem. In both cases we have $T < \infty$ a.s. and so

$$M_{T\wedge n} \underset{n\to\infty}{\longrightarrow} M_T, \qquad \text{a.s.} \tag{14.15}$$

In (2) we have $|M_{T\wedge n}| \leqslant K$ and so $|M_T| \leqslant K$ as well. The Bounded Convergence Theorem then gives $E(M_{T\wedge n}) \to EM_T$.

In (3) we instead note that

$$M_{T\wedge n} = M_0 + \sum_{k=1}^{T\wedge n} (X_k - X_{k-1}) \tag{14.16}$$

Using that $|X_k - X_{k-1}| \leqslant K$ a.s. this implies

$$\forall n \geqslant 0\colon\ |M_{T\wedge n}| \leqslant |M_0| + K(T \wedge n) \leqslant |M_0| + KT \ \text{ a.s.} \tag{14.17}$$

Since $M_0, T \in L^1$, the Dominated Convergence Theorem yields the claim. □

14.2 Applications to random walks.

To illustrate the above, we will consider the example of a random walk. We start with the simplest instance of all:

Definition 14.6 *The simple (symmetric) random walk started at zero is a process $\{S_n\}$ such that $S_0 := 0$ and $S_n := X_1 + \cdots + X_n$ for $\{X_k\}_{k\geqslant 1}$ i.i.d. $\{+1, -1\}$-valued random variables with $P(X_k = +1) = P(X_k = -1) = 1/2$.*

Clearly, the simple random walk is a $\mathbb{Z}$-valued additive martingale for the filtration $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$. The "walker" strides over the integers by choosing, at each time, a neighbor of the current position uniformly at random.

Our prime interest is concerned with the hitting times of integer level sets,

$$T_a := \inf\{n \geqslant 0\colon S_n = a\}, \qquad a \in \mathbb{Z} \tag{14.18}$$

which are stopping times by Lemma 14.2. Note that

$$\forall a \in \mathbb{Z} \smallsetminus \{0\}\colon \quad ET_a = \infty \tag{14.19}$$

for $ET_a < \infty$ along with $|S_{k+1} - S_k| \leqslant 1$ would enable Theorem 14.5(3) leading to a contradiction because $ES_{T_a} = a \neq 0$ yet $ES_0 = 0$.

In order to study these hitting times better, we will confine the walk to a finite set by considering the stopping time $T_a \wedge T_b$ where $a < 0 < b$. The following is relegated to a homework assignment:

Lemma 14.7 *For all integer $a < 0 < b$ we have $T_a \wedge T_b < \infty$ a.s.*

With this in hand, we now state:

Lemma 14.8 *For all integer $a < 0 < b$,*

$$P\big(T_a < T_b\big) = \frac{b}{b-a} \tag{14.20}$$

In particular,

$$\forall a \in \mathbb{Z}\colon\ T_a < \infty \text{ a.s.} \tag{14.21}$$

In short, the simple random walk visits every integer with probability one.

Proof. Given integers $a < 0 < b$ abbreviate $T := T_a \wedge T_b$. Since $S_{T\wedge n}$ is bounded and, by Lemma 14.7, $T < \infty$ a.s., the argument in the proof for case (2) in Theorem 14.5 implies

$$E(S_T) = E(S_0) = 0. \tag{14.22}$$

Noting that $S_T = a$ on $\{T_a < T_b\}$ and $S_T = b$ when $\{T_b > T_a\}$ along with the fact that $\{T_a < T_b\} \cup \{T_b > T_a\}$ is a disjoint partition of full measure set $\{T < \infty\}$ gives

$$\begin{aligned} 0 = E(S_T) &= aP(T_a < T_b) + bP(T_b < T_a) \\ &= aP(T_a < T_b) + b\big[1 - P(T_a < T_b)\big] \end{aligned} \tag{14.23}$$

Simple algebra now shows (14.20). For the second part we note that $T_b \geqslant b$ and so $T_b \to \infty$ as $b \to \infty$. Using this in (14.20) implies $P(T_a < \infty) = 1$ for all integer $a < 0$. The positive cases follow similarly by taking $a \to -\infty$. (The case of $a = 0$ is trivial.) □

Reiterating, the simple random walk visits every integer with probability one but, with the exception of the starting point, the time this takes has infinite mean. A question arises whether more can be said about the distribution of T_a. This is answered in the next claim whose proof we relegate to a homework assignment.

Lemma 14.9 *For all $a \in \mathbb{Z}$ and all $\lambda \geqslant 0$,*

$$E(\mathrm{e}^{-\lambda T_a}) = \left(\frac{1}{\mathrm{e}^{\lambda} + \sqrt{\mathrm{e}^{2\lambda} - 1}}\right)^{|a|} \tag{14.24}$$

In particular, $E(T_a^s) < \infty$ for $s < 1/2$ and $E(T_a^{1/2}) = \infty$ for all $a \neq 0$.

We remark that the distribution of T_a can be given very explicitly using the so called Reflection Principle to which we will get later in this course.

***Remark* 14.10** Inspired by the proof of Lemma 14.8 we can ask whether there are other functions $\varphi\colon \mathbb{Z} \to \mathbb{R}$ such that $\{\varphi(S_n)\}$ is a martingale for $\{S_n\}$ denoting a simple random walk. Such a function must obey

$$\forall n \in \mathbb{Z}\colon \quad \varphi(n) = \frac{1}{2}\big[\varphi(n-1) + \varphi(n+1)\big] \tag{14.25}$$

which a mean-value property defining so called discrete harmonicity. As is readily checked by rewriting this as $\varphi(n+1) - \varphi(n) = \varphi(n) - \varphi(n-1)$, every function with this property is necessarily linear. This, however, becomes far more interesting when we allow the steps to have a more general distribution (still confined to $\mathbb{Z}$).

While the Optional Stopping Theorem is very useful in general, in most situations it requires additional truncation arguments to get the desired conclusion. Here is one instance of this, drawing on Examples 13.5 and 13.9:

Theorem 14.11 (Wald's equations) *Given i.i.d. $X_1, X_2, \dots$ set $S_n = X_1 + \cdots + X_n$ and let T be a stopping time for the filtration $\mathcal{F}_n := \sigma(X_1, \dots, X_n)$. Then*

(1) *$X_1 \in L^1$ and $T \in L^1$ imply*

$$S_T \in L^1 \ \wedge \ ES_T = (ET)\, EX_1 \tag{14.26}$$

(2) *$X_1 \in L^2$, $EX_1 = 0$ and $T \in L^1$ imply*

$$S_T \in L^2 \ \wedge \ E(S_T^2) = (ET)\, E(X_1^2) \tag{14.27}$$

Note that applying alternative (3) in Theorem 14.5 to martingales $\{S_n - nEX_1\}$ for (1) and $\{S_n^2 - nE(X_1^2)\}$ for (2) is not allowed because their increments are not bounded. The proof of (2) requires Theorem 14.17 from the next subsection and also the concept of uniform integrability two lectures down the road. We leave details to homework.

14.3 Optional sampling theorem.

We will now take our previous discussion to another level and ask whether one can even apply the identity

$$\forall k \leqslant n\colon \quad E(M_n|\mathcal{F}_k) = M_k \text{ a.s.} \tag{14.28}$$

to random times. For this we have to give a meaning to $\mathcal{F}_k$ for k replaced by a random variable. We first observe:

Lemma 14.12 *For any stopping time T for a filtration $\{\mathcal{F}_n\}$ on a probability space $(\Omega, \mathcal{F}, P)$,*

$$\mathcal{F}_T := \big\{A \in \mathcal{F}\colon (\forall n \geqslant 0\colon \ A \cap \{T \leqslant n\} \in \mathcal{F}_n)\big\} \tag{14.29}$$

is a σ-algebra.

Proof. Since intersection of any number of σ-algebras is a σ-algebra, it suffices to show that, given σ-algebras $\mathcal{G}$ and $\mathcal{F}$ on Ω and $B \in \mathcal{G}$,

$$\mathcal{A} := \{A \in \mathcal{F}\colon A \cap B \in \mathcal{G}\} \text{ is a } \sigma\text{-algebra} \tag{14.30}$$

For this we observe that $\varnothing, \Omega \in \mathcal{A}$ and check that $\mathcal{A}$ and is closed under intersections and increasing limits. Since $A^c \cap B = B \smallsetminus (A \cap B) \in \mathcal{G}$ when $B, A \cap B \in \mathcal{G}$, it is also closed under complements and so it is a σ-algebra. □

Definition 14.13 *We call $\mathcal{F}_T$ the σ-algebra of events measurable by stopping time T.*

The reader should not try to think of $\mathcal{F}_T$ as a random σ-algebra. Rather, $\mathcal{F}_T$ is merely a collection of sets associated with function T on a probability space. Still, $\mathcal{F}_T$ behaves very much like the original filtration:

Lemma 14.14 *Let T and S be stopping times for filtration $\{\mathcal{F}_n\}$. Then*

$$\mathcal{F}_{T \wedge S} = \mathcal{F}_T \cap \mathcal{F}_S \tag{14.31}$$

In particular,

$$S \leqslant T \quad \Rightarrow \quad \mathcal{F}_S \subseteq \mathcal{F}_T \tag{14.32}$$

(The inequality $S \leqslant T$ must hold pointwise, not just pointwise a.s.!)

We leave the proof of this to an exercise. Next we need:

Lemma 14.15 *Let $\{X_n\}$ be adapted to $\{\mathcal{F}_n\}$ and T be a finite stopping time for $\{\mathcal{F}_n\}$. Then both T and X_T are $\mathcal{F}_T$-measurable.*

Proof. For all $m, n \in \mathbb{N}$ we have $\{T \leqslant m\} \cap \{T \leqslant n\} \in \mathcal{F}_n$ and so $\{T \leqslant m\} \in \mathcal{F}_T$ for all $m \geqslant 0$, implying measurability of T. For measurability of X pick $B \in \mathcal{B}(\mathbb{R})$ and note

$$\forall n \geqslant 0\colon \ \{X_T \in B\} \cap \{T = n\} = \{X_n \in B\} \cap \{T = n\} \in \mathcal{F}_n \tag{14.33}$$

Hence $\{X_T \in B\} \in \mathcal{F}_T$, implying measurability of X_T. □

In oder to state the next claim, we introduce the following concept:

Definition 14.16 *We say that a martingale $\{M_n, \mathcal{F}_n\}$ admits a* last element *if there exists $M_\infty \in L^1$ such that*

$$\forall n \geqslant 0\colon \ E(M_\infty|\mathcal{F}_n) = M_n \text{ a.s.} \tag{14.34}$$

Note that this means that the martingale is exactly of the kind discussed in Example 13.7. The reason for introducing the last element is that this gives us a way to define

$$M_T := M_\infty \quad \text{on} \quad \{T = \infty\} \tag{14.35}$$

With all the notions in place, we claim:

Theorem 14.17 (Optional sampling) *Let T and S be finite stopping times and $\{M_n\}$ a martingale for filtration $\{\mathcal{F}_n\}$. Assume that $S \leqslant T$ pointwise and that $M_{T\wedge n} \to M_T \in L^1$. Then*

$$E\big(M_T|\mathcal{F}_S\big) = M_S \ \text{ a.s.} \tag{14.36}$$

If $\{M_n\}$ admits a last element, then the above works for all stopping times with $S \leqslant T$.

Proof. Recall that $M_{T\wedge n}$ is a martingale and observe that, given any $A \in \mathcal{F}_S$, we have $A \cap \{S = k\} \in \mathcal{F}_k$ for all $k \geqslant 0$. Assuming $k \leqslant n$, we thus get

$$\begin{aligned} E\big(1_{A\cap\{S=k\}} M_{T\wedge n}\big) &= E\big(1_{A\cap\{S=k\}} E(M_{T\wedge n}|\mathcal{F}_k)\big) \\ &= E\big(1_{A\cap\{S=k\}} M_{T\wedge k}\big) = E\big(1_{A\cap\{S=k\}} M_{S\wedge n}\big) \end{aligned} \tag{14.37}$$

where the last equality uses that $M_{T\wedge k} = M_{S\wedge n}$ on $\{S = k\}$ thanks to $S \leqslant T$ and $k \leqslant n$. Note that $S \leqslant T$ also implies

$$E\big(1_{A\cap\{S>n\}} M_{T\wedge n}\big) = E\big(1_{A\cap\{S>n\}} M_n\big) = E\big(1_{A\cap\{S>n\}} M_{S\wedge n}\big) \tag{14.38}$$

Summing (14.37) over $k = 0, \dots, n$ and adding the result to (14.38) yields

$$\forall A \in \mathcal{F}_S\, \forall n \geqslant 0\colon \ E\big(1_A M_{T\wedge n}\big) = E\big(1_A M_{S\wedge n}\big) \tag{14.39}$$

By Lemma 14.14 we have $\mathcal{F}_{S\wedge n} \subseteq \mathcal{F}_S$ and, by Lemma 14.15, $M_{S\wedge n}$ is thus $\mathcal{F}_S$-measurable. Hence we get

$$\forall n \geqslant 0\colon \ E(M_{T\wedge n}|\mathcal{F}_S) = M_{S\wedge n} \quad \text{a.s.} \tag{14.40}$$

proving the claim for bounded stopping times.

In order to prove the claim for unbounded stopping times, we take $n \to \infty$. Observe that the stated assumption $M_{T\wedge n} \to M_T$ in L^1 implies $M_T \in L^1$ and the continuity of the conditional expectation in L^1-norm then gives

$$E(M_{T\wedge n}|\mathcal{F}_S) \xrightarrow[n\to\infty]{L^1} E(M_T|\mathcal{F}_S) \tag{14.41}$$

From (14.40) it follows that $\{M_{S\wedge n}\}$ converges in L^1 to a random variable Y. Since L^1-convergence implies a.s. convergence along a deterministic subsequence, the fact that $M_{S\wedge n} \to M_S$ as $n \to \infty$ implies $Y = M_S$ a.s. on $\{S < \infty\}$, which proves the claim for finite stopping times. For stopping times taking possibly value $+\infty$ we also need to observe that $S \leqslant T$ gives

$$M_{S\wedge n} = M_{T\wedge n} \xrightarrow[n\to\infty]{L^1} M_T = M_S \quad \text{on } \{S = \infty\} \tag{14.42}$$

and so $Y = M_S$ a.s. in this case as well. □

The same statement holds for submartingales and supermartingales; the equality in (14.36) then becomes the corresponding inequality. Theorem 14.17 is particularly useful when we choose to observe our process only along an increasing sequence $\{T_n\}$ of stopping times because it says that, if our original process was a (sub)martingale, so is the new process as well, albeit now with respect to the filtration $\{\mathcal{F}_{T_n}\}$.

Further reading: Durrett, Sections 4.8 and 4.9

15. MARTINGALE CONVERGENCE THEOREM

One (albeit rather simplistic) way to think of submartingales and supermartingales is as stochastic analogues of monotone sequences of numbers which, we note, converge either to a point in $\mathbb{R}$ or to one of the infinities. Here we demonstrate some validity of this intuition by showing that, under mild moment assumptions on the "dangerous" tail, the same holds for the random analogues.

15.1 Up-crossings and convergence.

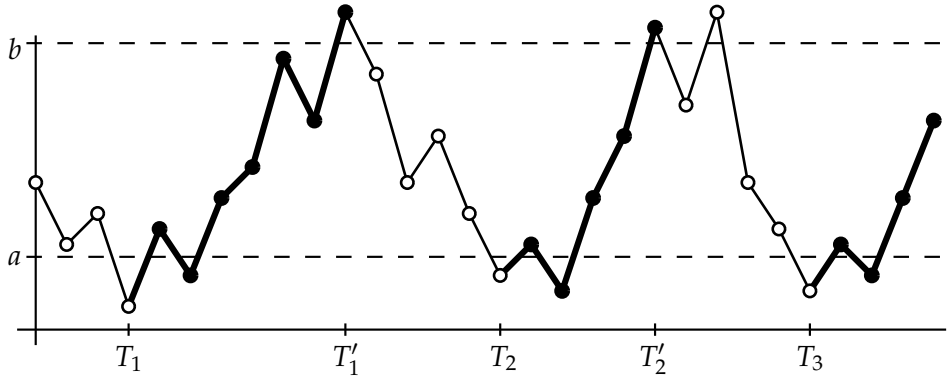
Our task here will be to prove:

Theorem 15.1 (Doob's Martingale Convergence Theorem) *Suppose $\{X_n, \mathcal{F}_n\}$ is a supermartingale with $\sup_{n\geqslant 0} E(X_n^-) < \infty$. Then*

$$X_\infty := \lim_{n\to\infty} X_n \text{ exist a.s. and } X_\infty \in L^1 \tag{15.1}$$

In particular, $X_\infty \in \mathbb{R}$ a.s.

We will actually present two proofs, one here and the other in the next lecture. Key for both of them is control of the oscillation of the sequence $\{X_n\}$ as $n \to \infty$. In our first proof we will achieve that explicitly by estimating the number of up-crossings of an interval. This requires some definitions and notation.



Let $a < b$ be two reals. We will follow the sequence $\{X_n\}$ until the first time it dips below a, marking that first time as T_1. Then we proceed until the sequence first climbs above b, marking that time by T_1'. Then we again wait until the sequence dips below a, marking that time T_2, and then until it climbs above b, marking that time as T_2'; see figure above. This leads to the following sequence of stopping times

$$T_0' < T_1 < T_1' < T_2 < T_2' < \dots \tag{15.2}$$

defined recursively as follows: Set $T_0' := 0$ and for $i \geqslant 1$ let

$$\begin{aligned} T_i &:= \inf\{n > T_{i-1}' \colon X_n < a\} \\ T_i' &:= \inf\{n > T_i \colon X_n > b\} \end{aligned} \tag{15.3}$$

We now write

$$U_n[a,b] := \max\{k \geqslant 0 : T'_k \leqslant n\} \tag{15.4}$$

for the number of (completed) *up-crossings* of $[a,b]$ by time n. Clearly $U_n[a,b] \leqslant n$ and $n \mapsto U_n[a,b]$ is non-decreasing in n. A key ingredient of the proof comes in:

Lemma 15.2 (Upcrossing inequality) *Suppose X is a supermartingale. Then for all $a < b$,*

$$EU_n[a,b] \leqslant \frac{E\big[(X_n - a)^-\big]}{b-a} \tag{15.5}$$

Proof. Define a $\{0,1\}$-valued process $\{C_n\}$ as follows: Set $C_0 := 0$ and, recursively, let

$$C_{n+1} := 1_{\{C_n=1\}}1_{\{X_n\leqslant b\}} + 1_{\{C_n=0\}}1_{\{X_n<a\}} \tag{15.6}$$

It is easy to check that

$$\forall k \geqslant 0\colon\ C_k = 1 \ \Leftrightarrow\ \exists i \geqslant 0\colon k \in (T_i, T'_i] \tag{15.7}$$

The definition (15.6) also ensures that C is predictable. Now consider the process $C \cdot X$. Writing U_n for $U_n[a,b]$, the special form of C then yields

$$(C\cdot X)_n = \sum_{i=1}^{U_n}(X_{T'_i} - X_{T_i}) + 1_{\{T_{U_n+1}<n\}}(X_n - X_{T_{U_n+1}}) \tag{15.8}$$

But $X_{T'_i} - X_{T_i} \geqslant b - a$ while $X_n - X_{T_{U_n+1}} \geqslant X_n - a \geqslant -(X_n - a)^-$ and so

$$(C\cdot X)_n \geqslant (b-a)U_n[a,b] - (X_n - a)^- \tag{15.9}$$

On the other hand, since $C_n \geqslant 0$ and X is a supermartinagle,

$$E((C\cdot X)_n) \leqslant E((C\cdot X)_0) = 0 \tag{15.10}$$

Putting these together we get

$$(b-a)EU_n[a,b] - E\big[(X_n - a)^-\big] \leqslant 0 \tag{15.11}$$

This gives the claim by a simple manipulation. □

We are ready to give:

Proof of Theorem 15.1. The monotonicity of $n \mapsto U_n[a,b]$ permits us to define

$$U_\infty[a,b] := \lim_{n\to\infty} U_n[a,b] \tag{15.12}$$

Thanks to the assumption,

$$\sup_{n\geqslant 0} E\big[(X_n - a)^-\big] \leqslant |a| + \sup_{n\geqslant 0} E(X_n^-) < \infty \tag{15.13}$$

and so, by the Upcrossing Inequality, also $\sup_{n\geqslant 0} EU_n[a,b] < \infty$. Invoking the Monotone Convergence Theorem we then get that $EU_\infty[a,b] < \infty$ and thus

$$\forall a < b\colon \quad U_\infty[a,b] < \infty \text{ a.s.} \tag{15.14}$$

with the null set depending possibly on a and b.

Next we note that

$$\big\{\liminf_{n\to\infty} X_n < a < b < \limsup_{n\to\infty} X_n\big\} \subseteq U_\infty[a,b] \tag{15.15}$$

Hence we get

$$\big\{\liminf_{n\to\infty} X_n < \limsup_{n\to\infty} X_n\big\} \subseteq \bigcup_{\substack{a,b\in\mathbb{Q}\\ a<b}} \big\{U_\infty[a,b] = \infty\big\} \tag{15.16}$$

In particular, from (15.14) we we now get

$$X_n \underset{n\to\infty}{\longrightarrow} X_\infty := \limsup_{n\to\infty} X_n, \qquad \text{a.s.} \tag{15.17}$$

It remains to show that $X_\infty \in L^1$. For that we note that Fatou's Lemma ensures

$$E(X_\infty^-) \leqslant \liminf_{n\to\infty} E(X_n^-) \leqslant \sup_{n\geqslant 0} E(X_n^-) < \infty \tag{15.18}$$

and so the lower tail of X_∞ is integrable. For the upper tail we use $X_n^+ = X_n + X_n^-$ to get

$$E(X_n^+) = E(X_n) + E(X_n^-) \leqslant E(X_0) + \sup_{n\geqslant 0} E(X_n^-) < \infty \tag{15.19}$$

where the middle inequality follows because X is a supermartingale. Invoking Fatou's Lemma again, we get $E(X_\infty^+) < \infty$ as well. □

As we will see later, the Martingale Convergence Theorem has many important consequences. The advantage is that one only needs a uniform bound on the integral of of the tail towards which the expectations go — which means the lower tail for supermartingales and the upper tail for submartingales. For martingales it suffices to check just one, whichever is convenient. A useful special case is:

Corollary 15.3 *A non-negative supermartingale converges a.s. to an L^1-random variable.*

15.2 Some simple applications.

To demonstrate the power of the theorem, we prove convergence some of the martingales we discussed earlier. We begin with that associated with Pólya's urn.

Lemma 15.4 *Let $\{M_n\}$ be the martingale expressing the fraction of red balls in Pólya's urn. Then $M_\infty := \lim_{n\to\infty} M_n$ exists a.s. and $M_\infty \in [0,1]$ a.s.*

Proof. Since $\{M_n\}$ is an $[0,1]$-valued martingale, this follows from Corollary 15.3. □

While having the convergence is nice, we do not learn anything about the distribution of M_∞. This will be further elaborated on when we discuss exchangeability.

Next we move to the Galton-Watson branching process:

Lemma 15.5 *Let $\{\mathfrak{p}(n)\}$ denote the off-spring distribution. Then*

$$\mu := \sum_{k\geqslant 0} k\mathfrak{p}(k) \leqslant 1 \quad \wedge \quad \mathfrak{p}(0) > 0 \quad \Rightarrow \quad P(S_n > 0 \text{ i.o.}) = 0 \tag{15.20}$$

In words, a non-degenerate (sub)critical branching process dies out almost surely.

Proof. Recall that $\{S_n\}$ is an $\mathbb{N}$-valued supermartingale when $\mu \leqslant 1$. By Corollary 15.3 again, $S_\infty := \lim_{n\to\infty} S_n$ exists and is $\mathbb{N}$-valued a.s. Integer valued sequences converge if and only if they are eventually constant and so for each $m \geqslant 1$,

$$\{S_\infty = m\} = \{S_n \neq m \text{ i.o.}\}^{\mathrm{c}} \subseteq \left\{ \bigcap_{k=1}^{m} \{X_{n,k} = 0\} \text{ i.o.}(n) \right\}^{\mathrm{c}} \tag{15.21}$$

As $P(\bigcap_{k=1}^{m}\{X_{n,k} = 0\}) = \mathfrak{p}(0)^m > 0$, the second Borel-Cantelli lemma tells us that the last event occurs with probability zero for any $m \geqslant 1$. Hence, $S_\infty = 0$ a.s. □

There is another martingale associated with Galton-Watson branching process. For this we note:

Lemma 15.6 *Let $\{S_n\}$ be a run of the Galton-Watson process with $S_0 = 1$ a.s. For each $t \geqslant 0$, let $\varphi_n(t) := E\mathrm{e}^{-tS_n}$. Then*

$$\forall n \geqslant 0\colon\ \varphi_{n+1}(t) = \varphi_n \circ \chi(t) \tag{15.22}$$

where

$$\chi(t) := -\log \sum_{k\geqslant 0} \mathfrak{p}(k)\mathrm{e}^{-kt} \tag{15.23}$$

In particular, if $t_\star \geqslant 0$ is such that $\chi(t_\star) = t_\star$, then $\{\mathrm{e}^{-t_\star S_n}\}$ is a martingale.

The lemma actually implies that $\varphi_n(t) = \mathrm{e}^{-\chi^n(t)}$, where χ^n is the n-fold composition of χ with itself. Since χ is checked to be non-decreasing and concave with $\chi'(0) = \mu$ (including $\mu = \infty$), assuming that $\mathfrak{p}(0) > 0$ the iterations tend to zero as $n \to \infty$ for $\mu \leqslant 1$ but converge to the nontrivial intersection of $t \mapsto \chi(t)$ has with the diagonal when $\mu > 1$. In both cases we get the existence of a fixed point $t_\star$ with the above properties for which one can moreover show that

$$\mathrm{e}^{-t_\star} = P(\exists n \geqslant 0\colon S_n = 0) \tag{15.24}$$

We leave the proof of these facts to homework. More advanced application of martingale convergence will be discussed in the next lectures.

Further reading: Durrett, Sections 4.2 and 4.3

16. Applications of martingale convergence

Here we give some applications of martingale convergence that can all be collected under the banner differentiation theorems.

16.1 Connection to Lebesgue differentiation.

We start with a connection to differentiation of integrals. This is a topic that was developed in order to prove that differentiation inverts integration. In Newton/Cauchy/Riemann integral theory is known under the title Fundamental Theorem of Calculus and is the source of known complication with the associated integral. Lebesgue theory resolves this in a very elegant way.

Let μ be a Borel probability measure on $\mathbb{R}^d$. We will partition $\mathbb{R}^d$ into disjoint translates of $[0,2^{-n})^d$. For each $x \in \mathbb{R}^d$ let z be the unique point in $\mathbb{Z}^d$ such that $x \in 2^{-n}z + [0,2^{-n})^d$ and let $B_n(x) := 2^{-n}z + [0,2^n)^d$ denote the term in the partition that contains x. Given $h \in L^1(\mu)$ define

$$h_k(x) := \frac{1}{\mu_n(B_n(x))} \int_{B_n(x)} h \mathrm{d}\mu \tag{16.1}$$

when $\mu_n(B_n(x)) > 0$ and $h_k(x) := 1$ otherwise. We then claim:

Proposition 16.1 *For all $h \in L^1(\mu)$,*

$$h_n \underset{n\to\infty}{\longrightarrow} h \quad \mu\text{-a.e.} \tag{16.2}$$

Proof. Let $h \in L^1(\mu)$ and let X be distributed according to μ. Define

$$\mathcal{F}_n := \sigma\Big(\big\{X \in 2^{-n}z + [0,2^{-n})^d\big\} \colon z \in \mathbb{Z}^d\Big) \tag{16.3}$$

The fact that we partition $\mathbb{R}^d$ into dyadic boxes ensures that $\{\mathcal{F}_n\}$ is a filtration. Since $E(h(X)|\mathcal{F}_n)$ equals the (normalized) average of h against μ on $B_n(X)$, we have

$$E(h(X)|\mathcal{F}_n) = h_n(X) \text{ a.s.} \tag{16.4}$$

But $\{E(h(X)|\mathcal{F}_n)\}$ obeys

$$\big\|E(h(X)|\mathcal{F}_n)\big\|_1 \leqslant \|h(X)\|_1 \tag{16.5}$$

and so Theorem 15.1 gives

$$h_n(X) \underset{n\to\infty}{\longrightarrow} Y_h := \limsup_{n\to\infty} h_n(X) \text{ a.s.} \tag{16.6}$$

Moreover, Fatou's lemma implies

$$\|Y_h\|_1 \leqslant \|h(X)\|_1 \tag{16.7}$$

It remains to identify Y_h in terms of h and X.

We plug into the standard argument from differentiation theory which notes that,

$$h \text{ continuous} \quad \Rightarrow \quad \forall x \in \mathbb{R}^d \colon\ h_n(x) \underset{n\to\infty}{\longrightarrow} h(x) \tag{16.8}$$

In particular, we have $Y_h = h(X)$ a.s. for continuous h. Next note that, given $h \in L^1(\mu)$ and $g \in L^1(\mu)$ continuous we have

$$h_n(X) = g_n(X) + (h-g)_n(X) \tag{16.9}$$

and so

$$Y_h = g(X) + Y_{h-g} \text{ a.s.} \tag{16.10}$$

implying

$$\|Y_h - g(X)\|_1 \leqslant \|h(X) - g(X)\|_1 \tag{16.11}$$

by way of (16.7). Taking a sequence $\{g_k\}_{k\geqslant 1}$ of continuous functions such that $g_k \to h$ in $L^1(\mu)$, we get $Y_h = h(X)$ a.s. This is the desired claim. □

We remark that, while the above is based on partitions into dyadic boxes, its slight enhancement (to allow for $B_n(x) = 2^{-n}(z+z') + [0,2^{-n})$ where $z \in \mathbb{Z}^d$ is a unique element such that $2^n x \in z + [0,1)^d$ and z' is a fixed element of $\mathbb{Z}^d$) implies the standard differentiation theorem from analysis (where $B_n(x)$ is replaced by a box of sidelength 2^{-n} centered at x). The proof of the differentiation theorem in analysis is invariably based on a *maximal inequality* of the form

$$\forall \lambda > 0\colon \quad \mu\Big(\sup_{n\geqslant 1} |f_n| > \lambda\Big) \leqslant \frac{C}{\lambda}\|f\|_{L^1(\mu)} \tag{16.12}$$

which allows for above reduction to continuous functions. Proving the maximal inequality is usually quite technical and requires tricks such as the "Rising-sun lemma" or covering arguments due to Vitali or Besicovich. Our proof shows the relevant maximal inequality (16.7) by way of the Martingale Convergence Theorem.

16.2 Proof of martingale convergence via maximal inequality.

Inspired by our previous example, we can ask whether the Martingale Convergence Theorem can be proved by a maximal inequality as well. This is what we will show next although only for martingales and only those that arise by progressive conditioning of an L^1 random variable.

We start with an inequality that generalizes Kolmogorov's inequality from sums of independent centered random variables:

Lemma 16.2 (Doob's maximal inequality, enhanced version) *Let X be a submartingale with respect to filtration $\{F_n\}$. Then for all $0 \leqslant k \leqslant n$, all $A \in \mathcal{F}_k$ and all $\lambda \in \mathbb{R}$,*

$$\lambda P\bigg(A \cap \Big\{\max_{j=k,\dots,n} X_j > \lambda\Big\}\bigg) \leqslant E\big(X_n 1_{A\cap\{\max_{j=k,\dots,n} X_j > \lambda\}}\big) \tag{16.13}$$

Proof. Denote

$$A_k := A \cap \{X_k > \lambda\} \tag{16.14}$$

and, for $j = k+1, \dots, n$, set

$$A_\ell := A \cap \Big\{\max_{j=k,\dots,\ell-1} X_j \leqslant \lambda\Big\} \cap \{X_\ell > \lambda\} \tag{16.15}$$

Then

$$A \cap \Big\{ \max_{j=k,\dots,n} X_j > \lambda \Big\} = \bigcup_{j=k}^{n} A_j \tag{16.16}$$

with the union on the right disjoint. It follows that

$$P\bigg(A \cap \Big\{ \max_{j=k,\dots,n} X_j > \lambda \Big\} \bigg) = \sum_{j=k}^{n} P(A_j) \tag{16.17}$$

Next observe that, since A_j is $\mathcal{F}_j$-measurable, the submartinagle property of X implies

$$E(X_n 1_{A_j}) = E\big(E(X_n|\mathcal{F}_j)1_{A_j}\big) \geqslant E\big(X_j 1_{A_j}\big) \geqslant \lambda P(A_j) \tag{16.18}$$

for each $j = k, \dots, n$. Combining these we get

$$\begin{aligned} \lambda P\bigg(A \cap \Big\{ \max_{j=k,\dots,n} X_j > \lambda \Big\} \bigg) &= \sum_{j=k}^{n} \lambda P(A_j) \\ &\leqslant \sum_{j=k}^{n} E(X_n 1_{A_j}) = E\big(X_n 1_{A \cap \{\max_{j=k,\dots,n} X_j > \lambda\}}\big) \end{aligned} \tag{16.19}$$

where the last equality follows again from disjointness of the union in (16.16). □

Notice that we claim the inequality even for λ negative. This is because so can be the right-hand side. As we will see, this detail enters quite crucially in:

Proof of Theorem 15.1 for Lévy martingales. Suppose $\{X_n\}$ such that, for some $Y \in L^1$,

$$\forall n \geqslant 0\colon \ X_n = E(Y|\mathcal{F}_n) \text{ a.s.} \tag{16.20}$$

Fix $\beta \in \mathbb{R}$, let $\lambda < \beta$ and fix $A \in \bigcup_{k \geqslant 0} \mathcal{F}_k$. Consider the inequality (16.13) with k so large that $A \in \mathcal{F}_k$. Using (16.20) to replace X_n by Y on the right and taking $n \to \infty$ with the help of upward monotonicity of $n \mapsto \max_{k \leqslant j \leqslant n} X_j$ then gives

$$\lambda P\Big(A \cap \big\{ \sup_{j \geqslant k} X_j > \lambda \big\} \Big) \leqslant E\big(Y 1_{A \cap \{ \sup_{j \geqslant k} X_j > \lambda \}} \big) \tag{16.21}$$

Taking $\lambda \uparrow \beta$ with the help of Dominated Convergence replaces "$> \lambda$" by "$\geqslant \beta$" on both sides. Letting $k \to \infty$ with the help of downward monotonicity of $k \mapsto \sup_{j \geqslant k} X_j$ shows

$$\beta P\Big(A \cap \big\{ \limsup_{n \to \infty} X_n \geqslant \beta \big\} \Big) \leqslant E\big(Y 1_{A \cap \{ \limsup_{n \to \infty} X_n \geqslant \beta \}} \big) \tag{16.22}$$

for any $A \in \bigcup_{n \geqslant 0} \mathcal{F}_n$.

Next recall that, as shown in Carathéodory construction of measure, given an algebra $\mathcal{A}$, a measure μ on $\sigma(\mathcal{A})$ and $\epsilon > 0$, for each $A \in \sigma(\mathcal{A})$ there exists $A' \in \mathcal{A}$ such that $\mu(A \triangle A') < \epsilon$. It follows that the inequality (16.22) will be preserved by the extension of both sides from $\mathcal{A} := \bigcup_{n \geqslant 0} \mathcal{F}_n$ to $\sigma(\mathcal{A})$. In particular, (16.22) holds for all $A \in \mathcal{F}_\infty := \sigma(\bigcup_{n \geqslant 0} \mathcal{F}_n)$.

Given any $\alpha \in \mathbb{R}$, a completely analogous argument with X_n replaced by $-X_n$ and β by $-\alpha$ shows that

$$\alpha P\Big(A \cap \big\{\liminf_{n\to\infty} X_n \geqslant \alpha\big\}\Big) \geqslant E\big(Y 1_{A\cap\{\liminf_{n\to\infty} X_n \geqslant \alpha\}}\big) \tag{16.23}$$

holds for all $A \in \mathcal{F}_\infty$. We now apply (16.22–16.23) to A replaced by

$$A_{\alpha,\beta} := \Big\{\liminf_{n\to\infty} X_n < \alpha < \beta < \limsup_{n\to\infty} X_n\Big\} \tag{16.24}$$

which obviously belongs to $\mathcal{F}_\infty$. This gives

$$\forall \alpha < \beta\colon \quad \beta P(A_{\alpha,\beta}) \leqslant E(Y 1_{A_{\alpha,\beta}}) \leqslant \alpha P(A_{\alpha,\beta}) \tag{16.25}$$

implying

$$\forall \alpha < \beta\colon \quad P(A_{\alpha,\beta}) = 0 \tag{16.26}$$

On the complement of the union of $A_{\alpha,\beta}$ over all $\alpha, \beta \in \mathbb{Q}$ with $\alpha < \beta$, every interval with rational points will be crossed only finitely many times by $\{X_n\}$. Hence we get

$$X_n \underset{n\to\infty}{\longrightarrow} X_\infty := \limsup_{n\to\infty} X_n \text{ a.s.} \tag{16.27}$$

Using (16.20), we have $E|X_\infty| \leqslant E|Y|$ a.s. and so $X_\infty \in L^1$. □

We do not know whether the argument can be adapted to include martingales that are not of the form (16.20). This would matter because, as we will show in Lévy's Forward Theorem (Theorem 17.11) below,

$$E(Y|\mathcal{F}_n) \underset{n\to\infty}{\longrightarrow} E(Y|\mathcal{F}_\infty) \text{ a.s.} \tag{16.28}$$

so if Y is $\mathcal{F}_\infty$-measurable then nothing "new" is discovered in the limit. (Still, the above gives the convergence part of the proof of Theorem 17.11.) The restriction to martingales can be relaxed to sub/supermartingales, provided we invoke Doob's decomposition to be discussed later.

16.3 Lebesgue decomposition revisited.

As another application we observe that the Martingale Convergence Theorem yields the Lebesgue decomposition for probability measures on filtered spaces:

Theorem 16.3 *Given a measurable space $(\Omega, \mathcal{F})$, a filtration $\{\mathcal{F}_n\}$ with $\sigma(\bigcup_{n\geqslant 0} \mathcal{F}_n) = \mathcal{F}$ and two probability measures P and Q, let P_n and Q_n be their respective restrictions to $\mathcal{F}_n$ and assume that $\forall n \geqslant 0\colon Q_n \ll P_n$. Denote $X_n := \frac{\mathrm{d}Q_n}{\mathrm{d}P_n}$. Then*

$$X_n \underset{n\to\infty}{\longrightarrow} X_\infty := \liminf_{n\to\infty} X_n \quad (P+Q)\text{-a.s.} \tag{16.29}$$

with X_∞ in $L^1(\Omega, \mathcal{F}, P)$. Moreover,

$$\forall A \in \mathcal{F}\colon \quad Q(A) = \int_A X_\infty \,\mathrm{d}P + Q\big(A \cap \{X_\infty = \infty\}\big) \tag{16.30}$$

Proof. The setup ensures that, for each $n \geqslant 0$ and each $A \in \mathcal{F}_n$, we have

$$\begin{aligned}\int_A \frac{X_n}{1+X_n}\mathrm{d}(P+Q) &= \int_A \frac{X_n}{1+X_n}\mathrm{d}(P_n+Q_n)\\ &= \int_A X_n \mathrm{d}P_n = Q_n(A) = Q_{n+1}(A) = \int_A X_{n+1}\mathrm{d}P_{n+1} \qquad (16.31)\\ &= \int_A \frac{X_{n+1}}{1+X_{n+1}}\mathrm{d}(P_{n+1}+Q_{n+1}) = \int_A \frac{X_{n+1}}{1+X_{n+1}}\mathrm{d}(P+Q)\end{aligned}$$

Since X_n is $\mathcal{F}_n$-measurable and $Z_n := \frac{X_n}{1+X_n} \in [0,1]$, we conclude that $\{Z_n\}$ is a non-negative bounded martingale with respect to $\{\mathcal{F}_n\}$ on $(\Omega, \mathcal{F}, \frac{1}{2}(P+Q))$.

Theorem 15.1 shows $Z_n \to Z$, $P+Q$-a.s. Since $z \mapsto \frac{z}{1+z}$ is continuous and increasing on $[0,\infty]$ and thus invertible, this implies the convergence (16.29) with

$$Z = \frac{X_\infty}{1+X_\infty} \quad (P+Q)\text{-a.s.} \qquad (16.32)$$

where the right-hand side is interpreted as 1 when $X_\infty = \infty$. Since $\int X_n \mathrm{d}P = 1$, Fatou's lemma shows $X_\infty \in L^1(\Omega, \mathcal{F}, P)$ and thus

$$P(X_\infty = \infty) = 0 \qquad (16.33)$$

Passing to $n \to \infty$ in (16.31) using the Bounded Convergence Theorem in turn gives

$$Q(A) = \int_A \frac{X_\infty}{1+X_\infty}\mathrm{d}(P+Q) \qquad (16.34)$$

for all $A \in \bigcup_{n\geqslant 0}\mathcal{F}_n$. This is an equality between two finite measures and so Dynkin's π/λ-Theorem shows that equality holds for all $A \in \mathcal{F} = \sigma(\bigcup_{n\geqslant 0}\mathcal{F}_n)$.

We now use the same argument as in the proof of the Radon-Nikodym Theorem/Lebesgue decomposition (Theorem 10.5): Integrate both sides of (16.34) against

$$Y := (1+X_\infty)1_{A\cap\{X_\infty \leqslant r\}} \qquad (16.35)$$

and move the integrals with respect to Q to the left hand side. Then pass to $r \to \infty$ with the help of the Monotone Convergence Theorem to get

$$Q\big(A \cap \{X_\infty < \infty\}\big) = \int_A X_\infty \,\mathrm{d}P \qquad (16.36)$$

where we also used (16.33). This now gives the claim. □

We note that this is again a differentiation result, but this time for measures: If $\{\mathcal{F}_n\}$ are generated by nested partitions of Ω that give access to all measurable sets, then the ratios of Q and P on the elements of the partition — which is what the value of $\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}$ will correspond to, converge a.s. both under P and Q. The set where the limit is finite is where Q is absolutely continuous with respect to P. The set where the limit is infinite is the singular part. (The remaining set is null under both P and Q.)

16.4 Kakutani product theorem.

We will now use these ideas further to give necessary and sufficient conditions for absolutely continuous of infinite product measures:

Theorem 16.4 (Kakutani product theorem) *Let $X = \{X_k\}_{k\geqslant 0}$ and $Y = \{Y_k\}_{k\geqslant 0}$ be sequences of independent random variables for which there exist non-negative measurable functions $\{f_k\}_{k\geqslant 0}$ such that*

$$\forall k \geqslant 0\, \forall A \in \mathcal{B}(\mathbb{R})\colon\ P(Y_k \in A) = E\big(f_k(X_k)1_{\{X_k\in A\}}\big) \tag{16.37}$$

(That is, the law of Y_k is absolutely continuous with respect to the law of X_k and f_k is the Radon-Nikodym derivative.) Let μ_X, resp., μ_Y denote the distribution of $\{X_k\}_{k\geqslant 1}$, resp., $\{Y_k\}_{k\geqslant 1}$ on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R})^{\otimes N})$. Then

$$\prod_{k=0}^{\infty} E\big(\, f(X_k)^{1/2}\big) > 0 \quad \Rightarrow \quad \mu_Y \ll \mu_X \tag{16.38}$$

and

$$\prod_{k=0}^{\infty} E\big(\, f(X_k)^{1/2}\big) = 0 \quad \Rightarrow \quad \mu_Y \perp \mu_X \tag{16.39}$$

where infinite product exists because, by Jensen, $E(\, f(X_k)^{1/2}) \leqslant \sqrt{E(f(X_k)} = 1$ for each $k \geqslant 1$. In addition, we have

$$\forall k \geqslant 0\colon\ f_k > 0 \quad \Rightarrow \quad \Big(\mu_Y \ll \mu_X \ \Leftrightarrow\ \mu_X \ll \mu_Y\Big) \tag{16.40}$$

so for positive f_k's the laws μ_X and μ_Y are either equivalent or singular.

Proof. We will realize both random sequences as coordinate projections on $\Omega := \mathbb{R}^{\mathbb{N}}$ with $\mathcal{F} := \mathcal{B}(\mathbb{R})^{\otimes\mathbb{N}}$. Writing $\mathcal{F}_n$ for the events depending only on the coordinates up to n, the Radon-Nikodym derivative of $\mu_Y|_{\mathcal{F}_n}$ with respect to $\mu_Y|_{\mathcal{F}_n}$ is given by

$$M_n := \prod_{k=0}^{n} f_k(X_k) \tag{16.41}$$

Theorem 16.3 then gives

$$M_n \underset{n\to\infty}{\longrightarrow} M_\infty := \liminf_{n\to\infty} M_n \quad (\mu_X + \mu_Y)\text{-a.s.} \tag{16.42}$$

with $M_\infty \in L^1(\mu_X)$ and

$$\mu_Y(A) = \int_A M_\infty \mathrm{d}\mu_X + \mu_Y\big(A \cap \{M_\infty = \infty\}\big) \tag{16.43}$$

for all $A \in \sigma(\bigcup_{n\geqslant 0} \mathcal{F}_n) = \mathcal{F}$.

We are now ready to show (16.39). Indeed, Fatou's lemma along with the product structure of μ_X gives

$$\int M_\infty \mathrm{d}\mu_x \leqslant \liminf_{n\to\infty} \int M_n^{1/2} \mathrm{d}\mu_X = \lim_{n\to\infty} \prod_{k=0}^{n} E\big(\, f(X_k)^{1/2}\big) \tag{16.44}$$

and so vanishing infinite product implies $M_\infty = 0$ μ_X-a.s. From (16.43) we get $\mu_Y \perp \mu_X$.

The proof of (16.7) requires more work. We aim to show that $M_n \to M_\infty$ in $L^1(\mu_X)$. For this we pick $\epsilon > 0$ and use that the product converges and is over quantities in $[0,1]$ to find $n_0 \geqslant 1$ such that

$$\forall n > \ell \geqslant n_0 \colon \quad \prod_{k=\ell+1}^{n} E\big(f(X_k)^{1/2}\big) \geqslant 1 - \epsilon \tag{16.45}$$

With the expectations relative to μ_X, we now observe that

$$\begin{aligned} E|M_n - M_\ell| &= E\big(|\sqrt{M}_n - \sqrt{M}_\ell|(\sqrt{M}_n + \sqrt{M}_\ell)\big) \\ &\leqslant \Big[E\big(|\sqrt{M}_n - \sqrt{M}_\ell|^2\big)E\big(|\sqrt{M}_n + \sqrt{M}_\ell|^2\big)\Big]^{1/2} \end{aligned} \tag{16.46}$$

which with the help of $(|a| + |b|)^2 \leqslant 2a^2 + 2b^2$ and the fact that $E(M_k) = 1$ gives

$$E|M_n - M_\ell| \leqslant 2\Big[E\big(|\sqrt{M}_n - \sqrt{M}_\ell|^2\big)\Big]^{1/2} \tag{16.47}$$

Next we use the explicit form (16.41) to get

$$E\big(|\sqrt{M}_n - \sqrt{M}_\ell|^2\big) = E\left(M_\ell \left| \prod_{k=\ell+1}^{n} f_k(X_k)^{1/2} - 1 \right|^2\right) = E\big(|W^{1/2} - 1|^2\big) \tag{16.48}$$

where

$$W := \prod_{k=\ell+1}^{n} f_k(X_k) \tag{16.49}$$

and the second inequality follows from the independence of $\{X_k\}$. We now use that $E(W) = 1$ to get

$$E\big(|W^{1/2} - 1|^2\big) = E\big(W + 1 - 2W^{1/2}\big) = 2\big[1 - E(W^{1/2})\big] \tag{16.50}$$

Relying on $W^{1/2}$ still having a product structure, the independence of $\{X_k\}$ along with (16.46–16.50) show

$$E|M_n - M_\ell| \leqslant 2\sqrt{2}\left[1 - \prod_{k=\ell+1}^{n} E\big(f(X_k)^{1/2}\big)\right]^{1/2} \tag{16.51}$$

For $n, \ell \geqslant n_0$, (16.45) shows that the right-hand side is less than $2\sqrt{2}\sqrt{\epsilon}$, proving that $\{M_k\}$ is Cauchy in $L^1(\mu_X)$. Hence $M_n \to M_\infty$ in $L^1(\mu_X)$. Using (16.43) with $A := \Omega$ we then get $\mu_Y(M_\infty = \infty) = 0$ and so $\mu_Y \ll \mu_X$.

If all $\{f_k\}$ are strictly positive, then exchanging the roles of X and Y shows

$$\prod_{k=0}^{\infty} E\big(f(Y_k)^{-1/2}\big) > 0 \quad \Rightarrow \quad \mu_Y \ll \mu_X \tag{16.52}$$

where we used that f_k^{-1} is the Radon-Nikodym derivative of the law of X_k with respect to the law of Y_k. But

$$E\big(f(Y_k)^{-1/2}\big) = E\big(f(X_k)^{-1/2} f(X_k)\big) = E\big(f(Y_k)^{1/2}\big) \tag{16.53}$$

and so the premise of (16.52) is the same as that of (16.38). □

To see how the above criteria work, let $\{X_k\}$ be Bernoulli$(1/2)$ and let $\{Y_k\}$ be independent $\{0,1\}$-valued with $P(Y_k = 1) = \frac{1}{2} + \epsilon_k$ for some ϵ_k with $|\epsilon_k| < 1/2$. Note that for the Radon-Nikodym derivative we get

$$f_k(a) = \begin{cases} 1 + 2\epsilon_k, & \text{if } a = 1, \\ 1 - 2\epsilon_k, & \text{if } a = 0, \end{cases} \tag{16.54}$$

and so

$$E\big(f(X_k)^{1/2}\big) = \frac{1}{2}\Big(\sqrt{1+2\epsilon_k} + \sqrt{1-2\epsilon_k}\Big) = 1 - \epsilon_k^2 + o(\epsilon_k^2) \tag{16.55}$$

Theorem 16.4 then says that the laws of X and Y are mutually absolutely continuous if $\sum_{k\geqslant 1} \epsilon_k^2 < \infty$ and singular otherwise.

Further reading: Durrett, Sections 4.2 and 4.3.3

17. Uniformly integrable martingales

Having shown that martingales converge to an integrable random variable a.s., we recall that they obey $EM_n = EM_0$ and ask whether $M_n \to M_\infty$ a.s. implies $EM_\infty = EM_0$. This can be guaranteed in reasonable generality thanks to a concept that we develop here.

17.1 Uniform integrability.

We wish to articulate a condition that guarantees $EX_n \to EX$ from $X_n \to X$ a.s. Recall that, for non-negative X_n, Fatou's lemma gives $\liminf_{n\to\infty} EX_n \geqslant EX$. If the inequality is strict, "mass" somehow got lost to "infinity" from the measures $\mu_n(A) := E(1_A X_n)$ in the limit. Thus, to prevent this from happening, we need to assume that these measures are tight. This is basically the content of:

Definition 17.1 (Uniform integrability) *A family $\{X_\alpha : \alpha \in I\}$ of real-valued random variables is said to be* uniformly integrable *(UI) if*

$$\forall \epsilon > 0\ \exists K \in (0,\infty)\ \forall \alpha \in I: \quad E\bigl(|X_\alpha| 1_{\{|X_\alpha|\geqslant K\}}\bigr) < \epsilon \tag{17.1}$$

That this does indeed imply tightness is shown in:

Lemma 17.2 *Let $\{X_\alpha : \alpha \in I\}$ be UI. Then*

$$\forall \epsilon > 0\ \exists \delta > 0\ \forall A \in \mathcal{F}: \quad P(A) < \delta \ \Rightarrow\ \forall \alpha \in I:\ E(|X_\alpha| 1_A) < \epsilon \tag{17.2}$$

Proof. Note that

$$E(|X_\alpha| 1_A) \leqslant E\bigl(|X_\alpha| 1_{\{|X_\alpha|\geqslant K\}}\bigr) + KP(A) \tag{17.3}$$

so taking K so large that the first expectation is less than $\epsilon/2$, the right-hand side is less than ϵ once $\delta < \epsilon/(2K)$. □

Before we start addressing the initial question, let us give examples of UI families and/or sufficient conditions for uniform integrability.

Lemma 17.3 *If $X \in L^1$ then $\{X\}$ is UI.*

Proof. Dominated Convergence shows $E(|X| 1_{\{|X|\geqslant K\}}) \to 0$ as $K \to \infty$. □

Lemma 17.4 *For all $n \in \mathbb{N}$, if $\mathcal{A}_0, \dots, \mathcal{A}_n \subseteq L^1$ are UI, then so is $\bigcup_{k=0}^n \mathcal{A}_k$. In words, finite unions of UI families and, in particular, finite subsets of L^1 are UI.*

Proof. For each $j = 0, \dots, n$, let K_j be such that $E(|X| 1_{\{|X|\geqslant K_j\}}) < \epsilon$ for all $X \in \mathcal{A}_j$. The constant $K := \max\{K_0, \dots, K_n\}$ then does the same for all $j = 0, \dots, n$ uniformly. □

Next we note UI implies a uniform boundedness in L^1 via the bound

$$\|X_\alpha\|_1 \leqslant K + E\bigl(|X_\alpha| 1_{\{|X_\alpha|\geqslant K\}}\bigr) \leqslant K + \epsilon \tag{17.4}$$

However, simple counterexamples show that infinite sequences with bounded L^1-norm need not be UI. Containment in a ball in L^p for any $p > 1$ nonetheless sufficient:

Lemma 17.5 *Suppose that $\{X_\alpha : \alpha \in I\} \subseteq L^p$ for some $p > 1$. Then*

$$\sup_{\alpha\in I} \|X_\alpha\|_p < \infty \;\Rightarrow\; \{X_\alpha : \alpha \in I\} \text{ is UI} \tag{17.5}$$

Proof. By Chebyshev's inequality

$$E(|X_\alpha| 1_{\{|X_\alpha|\geqslant K\}}) \leqslant \frac{\|X_\alpha\|_p}{K^{p-1}} \leqslant \frac{1}{K^{p-1}} \sup_{\alpha\in I} \|X_\alpha\|_p. \tag{17.6}$$

Now choose K exceeds $\frac{1}{p-1}$-th power of $\epsilon^{-1} \sup_{\alpha\in I} \|X_\alpha\|_p$. □

Another sufficient condition is domination by an L^1-random variable:

Lemma 17.6 *For any random variables $\{X_\alpha : \alpha \in I\}$ and Y,*

$$Y \in L^1 \;\wedge\; \big(\forall \alpha \in I : |X_\alpha| \leqslant Y\big) \;\Rightarrow\; \{X_\alpha : \alpha \in I\} \text{ is UI} \tag{17.7}$$

Proof. Since $x \mapsto x1_{\{x\geqslant K\}}$ is non-decreasing on $(0, \infty)$ we have

$$E(|X_\alpha| 1_{\{|X_\alpha|\geqslant K\}}) \leqslant E(|Y| 1_{\{|Y|\geqslant K\}}) \tag{17.8}$$

for all $\alpha \in I$. Dominated Convergence implies that the right-hand side is less than ϵ once K is sufficiently large. □

Here is one useful way to produce UI families of random variables:

Lemma 17.7 *Let $X \in L^1$. Then $\{E(X|\mathcal{G}) : \mathcal{G} \subseteq \mathcal{F}$ σ-algebra$\}$ is UI.*

Proof. Fix $\epsilon > 0$ and let $\delta > 0$ be such that $P(A) < \delta$ implies $E(|X|1_A) < \epsilon$. (See Lemmas 17.2–17.3.) Next choose $K > 0$ so that $E|X| < K\delta$. Given a sigma algebra $\mathcal{G} \subseteq \mathcal{F}$, abbreviate $Y := E(X|\mathcal{G})$ and note

$$P(|Y| \geqslant K) \leqslant \frac{1}{K} E|Y| \leqslant \frac{1}{K} E|X| < \delta \tag{17.9}$$

where we used that $\|Y\|_1 \leqslant \|X\|_1$ by Lemma 11.10. Then $E(|X|1_{\{|Y|\geqslant K\}}) < \epsilon$ and so

$$E\big(|Y|1_{\{|Y|\geqslant K\}}\big) \leqslant E\big(E(|X||\mathcal{G})1_{\{|Y|\geqslant K\}}\big) = E(|X|1_{\{|Y|\geqslant K\}}) < \epsilon \tag{17.10}$$

where we used Lemma 11.10 to get the first inequality. □

We now return to the original question and show that UI indeed does the job:

Theorem 17.8 *For any sequence $\{X_n\}_{n\geqslant 1}$ of random variables and any random variable X,*

$$X_n \xrightarrow{L^1} X \quad\Leftrightarrow\quad X_n \xrightarrow{P} X \;\wedge\; \{X_n : n \in \mathbb{N}\} \text{ is UI} \tag{17.11}$$

Proof of "⇒". Suppose $X_n \to X$ in L^1. Then

$$P(|X_n - X| > \epsilon) \leqslant \frac{1}{\epsilon} \|X_n - X\|_1 \underset{n\to\infty}{\longrightarrow} 0 \tag{17.12}$$

and so $X_n \to X$ in probability. To get that $\{X_n\}$ is UI, we note that the bound

$$|X_n| \leqslant |X| + |X - X_n| \tag{17.13}$$

implies

$$\begin{aligned} E(|X_n|1_{\{|X_n|\geqslant K\}}) &\leqslant E|X-X_n| + E(|X|1_{\{|X_n|\geqslant K\}}) \\ &\leqslant E|X-X_n| + E(|X|1_{\{|X|\geqslant K/2\}}) + E(|X|1_{\{|X_n-X|\geqslant K/2\}}) \end{aligned} \tag{17.14}$$

Now fix $\epsilon > 0$ and let $\delta > 0$ be such that $P(A) < \delta$ implies $E(|X|1_A) < \epsilon/3$. Then find $n_0 \geqslant 0$ so large that $\forall n \geqslant n_0 \colon \|X - X_n\|_1 < \epsilon/3$. Then use Lemma 17.3 to set

$$K := \max_{j=0,\dots,n_0} \inf\Bigl\{K' \geqslant \epsilon/\delta \colon E(|X_j|1_{\{|X_j|\geqslant K'\}}) < \epsilon\Bigr\} \tag{17.15}$$

Markov's inequality then shows $P(|X_n - X| \geqslant K/2) \leqslant \frac{2}{3}\epsilon/K < \delta$ for $n \geqslant n_0$ and so $E(|X|1_{\{|X_n-X|\geqslant K/2\}}) < \epsilon/3$. Putting these together we get that $E(|X_n|1_{\{|X_n|\geqslant K\}}) < \epsilon$ for all $n \geqslant 0$, showing that the sequence is UI. □

Proof of "⇐". assume that $X_n \to X$ in probability with $\{X_n\}$ UI. Then (**??**) shows $\sup_{n\geqslant 0}\|X_n\|_1$ and Fatou's lemma gives $X \in L^1$. will first prove that $X \in L^1$. For that we note that, by choosing a subsequence $\{n_k\}$, we can ensure $X_{n_k} \to X$ a.s. Fatou's Lemma gives us

$$E(|X|1_{\{|X|\geqslant K\}}) \leqslant \liminf_{k\to\infty} E(|X_{n_k}|1_{\{|X_{n_k}|\geqslant K\}}) \tag{17.16}$$

and so $E(|X|1_{\{|X|\geqslant K\}}) < \epsilon$ for K sufficiently large. A straightforward estimate now shows that, since $\{X_n\}$ is UI and $X \in L^1$, then also $\{X_n - X\}$ is UI. Hence, for any $0 < \epsilon < K < \infty$ we have

$$E|X_n - X| \leqslant \epsilon + KP(\epsilon < |X_n - X| \leqslant K) + E(|X_n - X|1_{\{|X_n-X|\geqslant K\}}) \tag{17.17}$$

The third term on the right can be made less than ϵ by choosing K sufficiently large. Since $X_n \to X$ in probability, the second term actually tends to zero as $n \to \infty$. The L^1-convergence follows. □

17.2 Lévy Forward Theorem.

Our next item of business is to apply the concept of uniform integrability to martingales. Here is a consequence of Theorem 17.8:

Corollary 17.9 *For sub/supermartingales $\{X_n\}$, the following are equivalent:*

(1) *$\{X_n\}$ is UI*
(2) *X_n converges a.s. and in L^1*
(3) *X_n converges in L^1*

Proof. We have (2) ⇒ (3) trivially and (3) ⇒ (1) by Theorem 17.8. To show (1) ⇒ (2), note $\{X_n\}$ UI implies $\sup_{n\geqslant 1} EX_n^{\pm} < \infty$. So $X_n \to X_\infty$ a.s. by the Martingale Convergence Theorem. Uniform integrability then yields L^1 convergence as well. □

Next we observe that once a martingale converges in L^1, it can be represented by conditioning from the limiting random variable.

Lemma 17.10 *Let $\{M_n, \mathcal{F}_n\}$ be a martingale.*

$$M_n \underset{n\to\infty}{\longrightarrow} M_\infty \text{ in } L^1 \quad \Leftrightarrow \quad \forall n \geqslant 0\colon\ M_n = E(M_\infty|\mathcal{F}_\infty) \text{ a.s.} \tag{17.18}$$

Both of these are equivalent to $\{M_n\}$ being UI.

Proof. By L^1-convergence, $E(M_k 1_A) \to E(M_\infty 1_A)$ as $k \to \infty$ for each $A \in \mathcal{F}$. On the other hand, for $A \in \mathcal{F}_n$ we have $E(M_k 1_A) = E(M_n 1_A)$ for each $k \geqslant n$. Hence

$$A \in \mathcal{F}_n \quad \Rightarrow \quad E(M_n 1_A) = E(M_\infty 1_A). \tag{17.19}$$

Since M_n is $\mathcal{F}_n$-measurable, it serves as a version of $E(M_\infty|\mathcal{F}_n)$. That (17.18) is also equivalent to $\{M_n\}$ being UI follows from Corollary 17.9. □

We now take a different perspective and prove:

Theorem 17.11 (Lévy Forward Theorem) *Let $X \in L^1$ and let $\{F_n\}$ be a filtration. Denote*

$$\mathcal{F}_\infty := \sigma\Biggl(\bigcup_{n\geqslant 1} \mathcal{F}_n\Biggr). \tag{17.20}$$

Then

$$E(X|\mathcal{F}_n) \underset{n\to\infty}{\longrightarrow} E(X|\mathcal{F}_\infty) \text{ a.s. and in } L^1 \tag{17.21}$$

holds for any choice of versions of the conditional expectations.

Proof. Splitting X into the positive and negative parts, we may assume $X \geqslant 0$. Given a choice of versions of the conditional expectations denote $X_n := E(X|\mathcal{F}_n)$ and set

$$X_\infty := \limsup_{n\to\infty} X_n. \tag{17.22}$$

Since $\{X_n, \mathcal{F}_n\}$ is a UI martingale, we have $X_n \to X_\infty$ a.s. and in L^1. We thus only have to show that X_∞ is a version of $E(X|\mathcal{F}_\infty)$.

First X_∞ is $\mathcal{F}_\infty$-measurable because X_n is $\mathcal{F}_n$-measurable and thus $\mathcal{F}_\infty$-measurable. To prove the other defining property of conditional expectation, define measures

$$\nu_1(A) := E\bigl(E(X|\mathcal{F}_\infty)1_A\bigr) \quad \text{and} \quad \nu_2(A) := E(X_\infty 1_A) \tag{17.23}$$

on $\mathcal{F}$. We claim that $\nu_1 = \nu_2$ on $\bigcup_{n\geqslant 1} \mathcal{F}_n$. Indeed, $A \in \mathcal{F}_n$ implies $E(X_k 1_A) = E(X_n 1_A)$ once $k \geqslant n$ and thus, by L^1 convergence $X_n \to X_\infty$, also $\nu_2(A) := E(X_\infty 1_A) = E(X_n 1_A)$. On the other hand, $E(E(X|\mathcal{F}_\infty)1_A) = E(X_n 1_A)$ by the "smaller-always-wins" principle and thus $\nu_2(A) = \nu_1(A)$.

The class $\{A \in \mathcal{F}\colon \nu_1(A) = \nu_2(A)\}$ is a λ-system that contains the π-system $\bigcup_{n\geqslant 1} \mathcal{F}_n$. Thus it contains $\mathcal{F}_\infty$ as well. It follows that $X_\infty = E(X|\mathcal{F}_\infty)$ a.s. □

The Lévy Forward Theorem implies a zero-one law:

Corollary 17.12 (Lévy's Zero-One Law) *For any filtration $\{\mathcal{F}_n\}$ and $\mathcal{F}_\infty$ given by* (17.20)*,*

$$\forall A \in \mathcal{F}_\infty\colon \quad E(1_A|\mathcal{F}_n) \underset{n\to\infty}{\longrightarrow} 1_A \text{ a.s.} \tag{17.24}$$

Along similar lines we get a very elegant proof of:

Theorem 17.13 (Kolmogorov's Zero-One Law) *Let $\{Y_k\}_{k\geqslant 0}$ be independent and denote*

$$\mathcal{T}_n := \sigma(Y_{n+1}, Y_{n+2}, \dots) \quad \text{and} \quad \mathcal{T} := \bigcap_{n\geqslant 1} \mathcal{T}_n \tag{17.25}$$

Then P is trivial on $\mathcal{T}$, i.e.,

$$\forall A \in \mathcal{T}\colon \quad P(A) \in \{0,1\}. \tag{17.26}$$

Proof. Denote $\mathcal{F}_n := \sigma(Y_1, \dots, Y_n)$. Then $\mathcal{F}_\infty := \sigma(Y_1, Y_2, \dots)$. Since every $A \in \mathcal{T}$ is independent of every $B \in \mathcal{F}_n$, we have $E(1_A|\mathcal{F}_n) = P(A)$ a.s. But Corollary 17.12 and $\mathcal{T} \subseteq \mathcal{F}_\infty$ gives $P(A) = E(1_A|\mathcal{F}_n) \to 1_A$ a.s. and so $P(A) \in \{0,1\}$ for all $A \in \mathcal{T}$. □

17.3 Lévy Backward Theorem.

Applications naturally lead us to consider the behavior of $E(X|\mathcal{F}_n)$ for decreasing sequences of σ-algebras as well. These are typical examples of *backward martingles* which are just martingales parametrized backwards in time. Here we get:

Theorem 17.14 (Lévy Backward Theorem) *Let $X \in L^1 \in (\Omega, \mathcal{F}, P)$ and let $\{\mathcal{F}_n\}$ be sub-σ-algebras of $\mathcal{F}$ such that $\forall n \geqslant 0\colon\ \mathcal{F}_{n+1} \subseteq \mathcal{F}_n$. Abbreviate*

$$\mathcal{F}_\infty := \bigcap_{n\geqslant 1} \mathcal{F}_n \tag{17.27}$$

Then

$$E(X|\mathcal{F}_n) \underset{n\to\infty}{\longrightarrow} E(X|\mathcal{F}_\infty), \qquad \text{a.s. \& in } L^1 \tag{17.28}$$

holds for any choice of versions of the conditional expectations.

Proof. Abbreviate $X_n := E(X|\mathcal{F}_n)$. We need to temporarily reverse time to establish convergence. Fix an integer $N \geqslant 1$ and denote

$$\mathcal{G}_n := \mathcal{F}_{(N-n)\vee 0} \tag{17.29}$$

and set

$$Y_n := E(X|\mathcal{G}_n) \tag{17.30}$$

Since $\{\mathcal{G}_n\}$ is a filtration, $\{Y_n, \mathcal{G}_n\}$ is a martingale. In particular, if $U_n[a,b]$ denotes the number of completed upcrossings of $[a,b]$ by $\{Y_1, \dots, Y_n\}$, then Doob's Upcrossing Inequality implies

$$EU_n[a,b] \leqslant \frac{EY_N + |a|}{b-a} \leqslant \frac{E|X| + |a|}{b-a} \tag{17.31}$$

Next and let $\widetilde{U}_n[a,b]$ denote the number of up-crossing of $[-b,-a]$ by $\{-X_0, \dots, -X_n\}$. Then

$$\widetilde{U}_n[a,b] \leqslant U_n[a,b] + 1 \tag{17.32}$$

and so $\sup_{n\geqslant 1} E\widetilde{U}_n[a,b] < \infty$. In particular,

$$\widetilde{U}_\infty[a,b] := \lim_{n\to\infty} \widetilde{U}_n[a,b] < \infty \text{ a.s.} \tag{17.33}$$

and so $\{X_k\}$ crosses $[a,b]$ only finite number of times a.s. It follows that

$$X_n \underset{n\to\infty}{\longrightarrow} X_\infty := \limsup_{n\to\infty} X_n, \qquad \text{a.s.} \tag{17.34}$$

The convergence also occurs in L^1 since $\{X_n\}$ is UI.

Our remaining task is to show that X_∞ is a version of $E(X|\mathcal{F}_\infty)$. First note that, thanks to the definition using *limes superior* and the fact that $n \mapsto \mathcal{F}_n$ is decreasing, X_∞ is $\mathcal{F}_n$-measurable for each $n \geqslant 1$. Hence it is also $\mathcal{F}_\infty$-measurable. Next pick $A \in \mathcal{F}_\infty$. Then $A \in \mathcal{F}_n$ for each $n \geqslant 0$ and so we have

$$E(X1_A) = E\big(E(X|\mathcal{F}_n)1_A\big) \underset{n\to\infty}{\longrightarrow} E(X_\infty 1_A) \tag{17.35}$$

by L^1-convergence $X_n \to X_\infty$. Hence $X_\infty = E(X|\mathcal{F}_\infty)$. □

As a bonus, we will give an elegant proof of the SLLN:

Proof of the Strong Law of Large Numbers. Suppose $\{X_n\}$ are i.i.d. with $X_1 \in L^1$. Set $S_n := X_1 + \cdots + X_n$ and let $\mathcal{F}_n := \sigma(S_n, S_{n+1}, \dots)$. We claim that

$$\forall k = 1, \dots, n: \quad E(X_k|\mathcal{F}_n) = E(X_1|\mathcal{F}_n) \tag{17.36}$$

For this pick $m \geqslant 0$ and $B_0, \dots, B_m \in \mathcal{B}(\mathbb{R})$ and set $A := \bigcap_{j=0}^m \{S_{n+j} \in B_j\}$. Denoting the distribution of $(X_1, \dots, X_{n+m})$ as μ, we have

$$\begin{aligned} E(X_k 1_A) &= \int_{\mathbb{R}^{n+m}} x_k \prod_{i=0}^m 1_{\{x_1+\cdots+x_{n+j}\in B_j\}} \mu(\mathrm{d}x_1, \dots, \mathrm{d}x_{n+m}) \\ &= \int_{\mathbb{R}^{n+m}} x_1 \prod_{i=0}^m 1_{\{x_1+\cdots+x_{n+j}\in B_j\}} \mu(\mathrm{d}x_1, \dots, \mathrm{d}x_{n+m}) = E(X_1 1_A) \end{aligned} \tag{17.37}$$

by the fact that μ (being a product measure) is invariant under relabeling of the first and k coordinates. Noting that the above product events form a π-system generating $\mathcal{F}_n$ while $\{A \in \mathcal{F}\colon\ E(X_1 1_A) = E(X_k 1_A)\}$ is a λ-system, Dynkin's π/λ theorem gives that $E(X_1 1_A) = E(X_k 1_A)$ holds for all $A \in \mathcal{F}_n$. Hence we get (17.36) by uniqueness of the conditional expectation.

Summing (17.36) over $k = 1, \dots, n$ yields

$$E(X_1|\mathcal{F}_n) = \frac{S_n}{n} \tag{17.38}$$

and Lévy Backward Theorem (and the fact that $\mathcal{F}_n$ is non-increasing) gives

$$\frac{S_n}{n} \underset{n\to\infty}{\longrightarrow} E(X_1|\mathcal{F}_\infty) \tag{17.39}$$

But $Y := \limsup_{n\to\infty} \frac{S_n}{n}$ is tail measurable and thus a.s. constant, by Kolmogorov's Zero-One Law. Hence $E(X_1|\mathcal{F}_\infty)$ is equal to its expectation, $E(X_1|\mathcal{F}_\infty) = E(X_1)$ a.s. It follows that $\frac{S_n}{n} \to EX_1$ a.s. and in L^1. □

The Lévy Backward Theorem is a great tool whenever we need to condition a family of random variables on the σ-algebra "at infinity." This is useful in the theory of Gibbs measures. We will see another application of this in the next lecture when we discuss exchangeability.

We finish with a remark on a simple corollary of both Lévy's theorems. Observe that one easily adjusts the proofs of Theorem 17.11, resp., Theorem 17.14 to show that for any sequence $\{X_n\}$ (not necessarily adapted)

$$X_n \underset{n\to\infty}{\longrightarrow} X \text{ in } L^1 \quad \Rightarrow \quad E(X_n|\mathcal{F}_n) \underset{n\to\infty}{\longrightarrow} E(X|\mathcal{F}_\infty) \text{ in } L^1 \tag{17.40}$$

However, as is easily checked by constructing a counterexample, this does not generally extend to a.s. convergence even if $X_n \to X$ a.s. and in L^1. The following sufficient condition is useful and easy to prove:

Corollary 17.15 (Dominated convergence for conditional expectations) *Let $\{\mathcal{F}_n\}$ be either increasing or decreasing sequence of σ-algebras and let $\mathcal{F}_\infty$ be either as in* (17.20) *or in* (17.27) *accordingly. Then for any sequence $\{X_n\}_{n\geqslant 0}$ of random variables,*

$$X_n \underset{n\to\infty}{\longrightarrow} X \text{ a.s.} \quad \wedge \quad \sup_{n\geqslant 0}|X_n| \in L^1 \tag{17.41}$$

imply

$$E(X_n|\mathcal{F}_n) \underset{n\to\infty}{\longrightarrow} E(X|\mathcal{F}_\infty) \text{ a.s. and in } L^1 \tag{17.42}$$

Proof. Denote $Z_n := \sup_{k\geqslant n}|X_k - X|$ and $Y := \sup_{n\geqslant 0}|X_n|$. From $X_n \to X$ a.s. we get $Z_n \leqslant 2Y$ a.s. and the assumptions give $Z_n \in L^1$. For each $n \geqslant 0$, Theorem 17.11, resp., Theorem 17.14 gives

$$\sup_{k\geqslant m} E\big(|X_k - X|\,\big|\,\mathcal{F}_n\big) \overset{n\leqslant m}{\leqslant} E(Z_n|\mathcal{F}_m) \underset{m\to\infty}{\longrightarrow} E(Z_n|\mathcal{F}_\infty) \text{ a.s.} \tag{17.43}$$

But $X_n \to X$ a.s. implies $Z_n \downarrow 0$ a.s. and, by $Z_n \leqslant 2Y$, also $Z_n \to 0$ in L^1. Taking $n \to \infty$ on the right thus shows

$$\limsup_{n\to\infty}\Big|E(X_n|\mathcal{F}_n) - E(X|\mathcal{F}_n)\Big| \leqslant \limsup_{n\to\infty} E\big(|X_n - X|\,\big|\,\mathcal{F}_n\big) = 0 \text{ a.s.} \tag{17.44}$$

Since $E(X_n|\mathcal{F}_n) \to E(X|\mathcal{F}_n)$ a.s. by above theorems, we get the claim. □

Further reading: Durrett, Sections 4.6-4.7

18. EXCHANGEABILITY

Our next task is to explore further the property (17.36) that underpinned the backward-martingale based proof of the SLLN. This leads to the concept of exchangeability that has many interesting applications. A key discovery in this angle of research was done by B. de Finetti whose name is now placed (rather generously) on many analogous statements throughout probability.

18.1 Analyzing Pólya's urn.

Let us return to Polya's urn. We will focus on the situation where, initially, there are r red and g green balls and at each time only one additional ball is added to the urn; i.e., $b = 1$. A run of the urn is described by random variables $Z_1, Z_2, \dots$, where

$$Z_k := 1_{\{k\text{-th sampled ball is red}\}} \tag{18.1}$$

Then $R_n = r + Z_1 + \cdots + Z_k$. Let us compute the probability of a given run $Z_1, \dots, Z_n$. First we examine the special case

$$\begin{aligned} P\big(Z_1 = 1, \dots, Z_k =& 1, Z_{k+1} = 0, \dots, Z_n = 0\big) \\ &= \frac{r}{r+g} \cdots \frac{r+k-1}{r+g+k-1} \, \frac{g}{r+g+k} \cdots \frac{g+n-k-1}{r+g+n-1} \\ &= \frac{r(r+1)\dots(r+k-1)\cdot g(g+1)\dots(g+n-k+1)}{(r+g)(r+g+1)\dots(r+g+n-1)} \end{aligned} \tag{18.2}$$

However, looking at the structure of the expression we easily convince ourselves that the same expression will be arrived at for any sequence of k ones and $n-k$ zeros. For all $a_1, \dots, a_n \in \{0,1\}$ with $\sum_{i=1}^n a_i = k$ we will thus have

$$P\big((Z_1, \dots, Z_n) = (a_1, \dots, a_n)\big) = P\Big((Z_1, \dots, Z_n) = (\underbrace{1, \dots, 1}_{k\text{-times}}, \underbrace{0, \dots, 0}_{n-k\text{-times}})\Big) \tag{18.3}$$

It follows that $(Z_1, Z_2, \dots)$ obeys the conditions in the following definition:

Definition 18.1 *A finite permutation π is a bijection $\pi\colon \mathbb{N} \to \mathbb{N}$ such that*

$$\big\{i \in \mathbb{N}\colon \pi(i) \neq i\big\} \text{ is finite} \tag{18.4}$$

We say that the random variables $(X_1, X_2, \dots)$ are exchangeable *if*

$$(X_{\pi(1)}, X_{\pi(2)}, \dots) \overset{\text{law}}{=} (X_1, X_2, \dots) \tag{18.5}$$

holds for any finite permutation π.

A particular consequence of the zero-one nature of the variables is that the computation of probability of a given run of zero's and one's conditioned on the total sum reduces to purely combinatorial argument:

$$P\left((Z_1, \dots, Z_n) = (a_1, \dots, a_n) \,\middle|\, \sum_{i=1}^n Z_i = k\right) = \binom{n}{k}^{-1} \quad \text{if} \quad \sum_{i=1}^n a_i = k \tag{18.6}$$

Reduction to the joint law of a finite number of Z_i's is easily obtained as well: If $1 \leqslant i_1 < i_2 < \cdots < i_m \leqslant n$ and $a_1, \ldots, a_m \in \{0,1\}$ are such that $a := a_1 + \cdots + a_m \leqslant k$, then

$$P\left(Z_{i_1} = a_1, \ldots, Z_{i_m} = a_m \,\middle|\, \sum_{i=1}^{n} Z_i = k \right) = \frac{\binom{n-m}{k-a}}{\binom{n}{k}} \tag{18.7}$$

Now let us consider the limit $n \to \infty$. Abbreviating $S_n := Z_1 + \cdots + Z_n$ and setting $\mathcal{F}_n := \sigma(S_n, S_{n+1}, \ldots)$, the argument in (17.37) with the required "swap" symmetry of the underlying distribution provided by exchangeability shows that

$$\frac{S_n}{n} = E(Z_1 | \mathcal{F}_n) \quad \text{a.s.} \tag{18.8}$$

The Backward Lévy Theorem then gives

$$\frac{S_n}{n} \underset{n\to\infty}{\longrightarrow} U := E(Z_1 | \mathcal{F}_\infty) \quad \text{a.s.} \tag{18.9}$$

A calculation now shows that

$$\frac{\binom{n-m}{k-a}}{\binom{n}{k}} = \frac{\left(\prod_{j=0}^{a-1} (k-j) \right) \left(\prod_{j=0}^{m-a-1} (n-k-j) \right)}{\prod_{j=0}^{m-1} (n-j)} \approx \frac{k^a (n-k)^{m-a}}{n^m} \tag{18.10}$$

With just tiny bit of work we find out that, from $\frac{S_n}{n} \to U$ a.s. we get

$$P\big(Z_{i_1} = a_1, \ldots, Z_{i_m} = a_m \,\big|\, \mathcal{F}_n\big) = \frac{\binom{n-m}{S_n - a}}{\binom{n}{S_n}} \underset{n\to\infty}{\longrightarrow} U^a (1-U)^{m-a} \tag{18.11}$$

for all $m \geqslant 1$ and all $a = 0, \ldots, m$. But the Backward Martingale Theorem implies that the limit of the left-hand side is the conditional probability given $\mathcal{F}_\infty := \bigcap_{n \geqslant 1} \mathcal{F}_n$ and so

$$P\big(Z_{i_1} = a_1, \ldots, Z_{i_m} = a_m \,\big|\, \mathcal{F}_\infty\big) = U^a (1-U)^{m-a} = \prod_{i=1}^{m} U^{a_i} (1-U)^{1-a_i} \tag{18.12}$$

In short, conditional on $\mathcal{F}_\infty$, or even just $\sigma(U)$, the random variable $(Z_1, Z_2, \ldots)$ are Bernoulli(U). As only exchangeability was used throughout, we have proved:

Theorem 18.2 (de Finetti 1931) *Suppose $Z_1, Z_2, \ldots$ are exchangeable with values in $\{0,1\}$. Then there exist a random variable U taking values in $[0,1]$ such that, conditional on U, the random variables $\{Z_i\}$ are Bernoulli(U). In particular, if μ denotes the distribution of U, then*

$$P\big(Z_1 = a_1, \ldots, Z_k = a_k\big) = \int_{[0,1]} \mu(\mathrm{d}u) \prod_{i=1}^{k} u^{a_i} (1-u)^{1-a_i} \tag{18.13}$$

holds for any $k \geqslant 1$ and any $a_1, \ldots, a_k \in \{0,1\}$.

In order to describe the full law, one has to find the distribution μ in (18.13). This of course depends on the particulars of the problem at hand. For the Pólya urn we get:

Lemma 18.3 *Consider the Pólya urn with parameters $r, g \geqslant 1$ and $b = 1$. Then U has distribution*

$$\mu(\mathrm{d}u) = 1_{[0,1]}(u)\frac{1}{B(r,g)}u^{r-1}(1-u)^{g-1}\mathrm{d}u \tag{18.14}$$

where $B(r,g) := \frac{(r+g-1)!}{(r-1)!(g-1)!}$. In particular, U is uniform on $[0,1]$ for $r = 1 = g$.

Proof. Using (18.1–18.2) we get

$$P\left(\sum_{j=1}^{n} Z_j = k\right) = \frac{1}{B(r,g)}\frac{(r+k-1)!(g+n-k-1)!}{(r+g+n-1)!}\binom{n}{k} \tag{18.15}$$

The right-hand side is written further as

$$\frac{1}{B(r,g)}\frac{\left(\prod_{j=1}^{r-1}(k+j)\right)\left(\prod_{j=1}^{g-1}(n-k+j)\right)}{\prod_{j=1}^{r+g-1}(n+j)} \tag{18.16}$$

The numerator and denominator can be estimated with the help of

$$\forall k, \ell \in \mathbb{N}\colon\ (k+1)^{\ell} \leqslant \prod_{j=1}^{\ell}(k+j) \leqslant (k+1+\ell)^{\ell} \leqslant (k+1)^{\ell}\mathrm{e}^{\frac{\ell^2}{k+1}} \tag{18.17}$$

where we used that $1 + s \leqslant \mathrm{e}^s$ for all real s. Noting that there are $r-1+g-1 = r+g-2$ terms in the numerator while $r+g-1$ in the denominator, dividing both numerator and denominator by $(n+1)^{r+g-1}$ shows that

$$P\left(\sum_{j=1}^{n} Z_j = k\right) = \frac{1}{n+1}\frac{1}{B(r,g)}\left(\left(\frac{k+1}{n+1}\right)^{r-1}\left(1-\frac{k+1}{n+1}\right)^{g-1}\right)q_{n,k} \tag{18.18}$$

where

$$\mathrm{e}^{-\frac{(r+g)^2}{n+1}} \leqslant q_{n,k} \leqslant \mathrm{e}^{\frac{r^2}{k+1}+\frac{g^2}{n-k+1}} \tag{18.19}$$

Summing the result for all k with $k \leqslant na$ and using that $u \mapsto u^{r-1}(1-u)^{g-1}$ is continuous to approximate the resulting sum by the Riemann integral gives

$$P\left(\sum_{j=1}^{n} Z_j \leqslant na\right) \underset{n\to\infty}{\longrightarrow} \frac{1}{B(r,g)}\int_0^{a\wedge 1} u^{r-1}(1-u)^{g-1}\mathrm{d}u \tag{18.20}$$

This shows that $\frac{1}{n}\sum_{j=1}^{n} Z_j \xrightarrow{\mathrm{w}} U$ with distribution in (18.14). □

The case $r = 1 = g$ is much simpler to deal with because (18.16) equals $\frac{1}{n+1}$ and so $\sum_{j=1}^{n} Z_j$ is actually uniform on $\{0, 1, \ldots, n\}$ for each $n \geqslant 1$.

18.2 General de Finetti theorem of Hewitt and Savage.

Having resolved the special case of zero-one valued random variables, we can move to general exchangeable families. Such families (not even necessarily indexed by naturals) appear in many parts of science for the simple fact that, by the laws of nature or as a feature of the problem, the random variables of interest do not come with a canonical labeling. Any family of random variables whose labeling is arbitrary (read: can be changed without affecting the result) is necessarily exchangeable.

A formal treatment of the general case will require some notation. Let $X = \{X_k\}_{k\geqslant 0}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$ taking values in (S, Σ). For each $n \in \mathbb{N}$, let

$$\Pi_n := \big\{\pi\colon \text{ finite permutation } \wedge\ \forall i > n\colon \pi(i) = i\big\} \tag{18.21}$$

For any finite permutation π, denote by X_π the sequence $\{X_{\pi(k)}\}_{k\geqslant 0}$. Recall that, by the Doob-Dynkin lemma, every $A \in \sigma(X_1, X_2, \dots)$ can be represented as $A = X^{-1}(B) = \{X \in B\}$ for some $B \in \Sigma^{\otimes\mathbb{N}}$. The phrase "$\{X_k\}_{k\geqslant 0}$ is exchangeable" then means

$$\forall B \in \Sigma^{\otimes\mathbb{N}}\ \forall \pi \in \bigcup_{n\geqslant 1} \Pi_n\colon\ P(X \in B) = P(X_\pi \in B) \tag{18.22}$$

Define

$$\mathcal{E}_n := \Big\{X^{-1}(B)\colon B \in \Sigma^{\otimes\mathbb{N}} \wedge\ \forall \pi \in \Pi_n\colon X_\pi^{-1}(B) = X^{-1}(B)\Big\} \tag{18.23}$$

It is readily checked that $\mathcal{E}_n$ is a σ-algebra and that $\Pi_n \subseteq \Pi_{n+1}$ implies $\mathcal{E}_{n+1} \subseteq \mathcal{E}_n$ for all $n \geqslant 0$. The intersection

$$\mathcal{E} := \bigcap_{n\geqslant 1} \mathcal{E}_n. \tag{18.24}$$

is also a σ-algebra which we refer to as the σ-algebra of *exchangeable events* associated with $\{X_k\}_{k\geqslant 0}$. A key technical tool we already used a few times above is:

Lemma 18.4 *Let $\{X_k\}_{k\geqslant 1}$ be an S-valued exchangeable sequence and let $f\colon S \to \mathbb{R}$ be a Borel measurable function such that $f(X_1) \in L^1$. Then*

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \underset{n\to\infty}{\longrightarrow} E\big(f(X_1)\,\big|\,\mathcal{E}\big) \text{ a.s. and in} L^1 \tag{18.25}$$

where $\mathcal{E}$ is the σ-algebra exchangeable events associated with $\{X_k\}_{k\geqslant 1}$.

Proof. Let $A \in \mathcal{E}_n$. Then $A = X^{-1}(B)$ such that $X_\pi^{-1}(B) = X^{-1}(B)$ for all $\pi \in \Pi_n$. Given $C \in \Sigma$, (18.22) then shows

$$E\big(1_{\{X_1\in C\}}1_A\big) = E\big(1_{\{X_1\in C\}}1_{X^{-1}(B)}\big) = E\big(1_{\{X_{\pi(1)}\in C\}}1_{X_\pi^{-1}(B)}\big) = E\big(1_{\{X_{\pi(1)}\in C\}}1_A\big) \tag{18.26}$$

Using additivity and Dominated Convergence, this gives $E(f(X_1)1_A) = E(f(X_{\pi(1)}1_A)$ for all $\pi \in \Pi_n$ implying

$$\forall k = 1, \dots, n\colon \quad E\big(f(X_k)\,\big|\,\mathcal{E}_n\big) = E\big(f(X_1)\,\big|\,\mathcal{E}_n\big) \tag{18.27}$$

Using that $S_n := \sum_{k=1}^n f(X_k)$ is $\mathcal{E}_n$-measurable, we get $\frac{S_n}{n} = E(f(X_1)|\mathcal{E}_n)$ a.s. The claim then follows from the Backward Lévy Theorem. □

For this setting, we now state:

Theorem 18.5 (de Finetti theorem, expectation version) *Let $\{X_n\}_{n\geqslant 0}$ be exchangeable and taking values in measurable space (S,Σ). Then for all $k \geqslant 0$ and all $B_0,\dots,B_k \in \Sigma$,*

$$E\Big(\prod_{i=0}^{k} 1_{\{X_i\in B_i\}}\,\Big|\,\mathcal{E}\Big) = \prod_{i=0}^{k} E(1_{\{X_1\in B_i\}}|\mathcal{E}) \qquad \text{a.s.} \tag{18.28}$$

Proof. Pick $B_1,\dots,B_k \in \Sigma$ and abbreviate $A_{j,i} := \{X_i \in B_j\}$. Our aim is to show that

$$P\big(A_{1,1}\cap\cdots\cap A_{k,k}\,\big|\,\mathcal{E}\big) = \prod_{j=1}^{k} P(A_{j,1}|\mathcal{E}) \text{ a.s.} \tag{18.29}$$

where we ease the notation by writing conditional probability for expectation of an indicator. For this we pick $n \geqslant k$ and write

$$\begin{aligned} E\bigg(\prod_{j=1}^{k}\Big(\frac{1}{n}\sum_{i=1}^{n} 1_{A_{j,i}}\Big)\,\bigg|\,\mathcal{E}_n\bigg) &= \frac{1}{n^k}\sum_{1\leqslant i_1,\dots,i_k\leqslant n} P\big(A_{1,i_1}\cap\cdots\cap A_{k,i_k}\,\big|\,\mathcal{E}_n\big) \\ &= O\Big(\frac{k^2}{n}\Big) + \frac{1}{n^k}\sum_{\substack{1\leqslant i_1,\dots,i_k\leqslant n\\ \text{distinct}}} P\big(A_{1,i_1}\cap\cdots\cap A_{k,i_k}\,\big|\,\mathcal{E}_n\big) \text{ a.s.} \end{aligned} \tag{18.30}$$

where we noted that the number of terms when at least one pair of indices coincide is at most $n^{k-1}\binom{k}{2}$ which is in turn at most $\frac{k^2}{n}$ times n^k. The key point now is that, for any distinct $i_1,\dots,i_k \leqslant n$,

$$P\big(A_{1,i_1}\cap\cdots\cap A_{k,i_k}\,\big|\,\mathcal{E}_n\big) = P\big(A_{1,1}\cap\cdots\cap A_{k,k}\,\big|\,\mathcal{E}_n\big) \text{ a.s.} \tag{18.31}$$

which is proved by integrating against any event from $\mathcal{E}_n$ and invoking exchangeability of $\{X_k\}$. It thus follows

$$E\bigg(\prod_{j=1}^{k}\Big(\frac{1}{n}\sum_{i=1}^{n} 1_{A_{j,i}}\Big)\,\bigg|\,\mathcal{E}_n\bigg) = O\Big(\frac{k^2}{n}\Big) + \frac{1}{n^k}\frac{n!}{(n-k)!}P\big(A_{1,1}\cap\cdots\cap A_{k,k}\,\big|\,\mathcal{E}_n\big) \tag{18.32}$$

But $\sum_{i=1}^{n} 1_{A_{j,i}}$ is measurable with respect to $\mathcal{E}_n$ and so

$$\begin{aligned} E\bigg(\prod_{j=1}^{k}\Big(\frac{1}{n}\sum_{i=1}^{n} 1_{A_{j,i}}\Big)\,\bigg|\,\mathcal{E}_n\bigg) &= \prod_{j=1}^{k}\Big(\frac{1}{n}\sum_{i=1}^{n} 1_{A_{j,i}}\Big) \\ &= \prod_{j=1}^{k} E\Big(\frac{1}{n}\sum_{i=1}^{n} 1_{A_{j,i}}\,\Big|\,\mathcal{E}_n\Big) = \prod_{j=1}^{k} P\big(A_{j,1}\,\big|\,\mathcal{E}_n\big) \text{ a.s.} \end{aligned} \tag{18.33}$$

where in the last step we used that $P(A_{j,i}|\mathcal{E}_n) = P(A_{j,1}|\mathcal{E}_n)$ a.s. for all $i \leqslant n$ by exchangeability. Summarizing, we have

$$\prod_{j=1}^{k} P\big(A_{j,1}\,\big|\,\mathcal{E}_n\big) = O\Big(\frac{k^2}{n}\Big) + \frac{1}{n^k}\frac{n!}{(n-k)!}P\big(A_{1,1}\cap\cdots\cap A_{k,k}\,\big|\,\mathcal{E}_n\big) \text{ a.s.} \tag{18.34}$$

Taking $n \to \infty$ with the help of the Backward Lévy Theorem, we get (18.29). □

Notice that the above proof makes $(X_1, \ldots, X_k)$ conditioned on $\mathcal{E}_n$ behave as i.i.d. as long as $k = o(\sqrt{n})$ as $n \to \infty$. This is also the best we can hope for in general because the sequence $(X_1, \ldots, X_n)$ could come from conditioning an i.i.d. sequence on its total sum which suppresses fluctuations of that sum that would normally be at least order $\sqrt{n}$. As can be checked, the absence of these fluctuations will start to be visible once order $\sqrt{n}$ of random variables are revealed.

An interesting question is what happens when the random variables $\{X_k\}_{k\geqslant 1}$ are already i.i.d. For this case we get:

Theorem 18.6 (Hewitt-Savage Zero-One Law) *If $\{X_k\}_{k\geqslant 1}$ are i.i.d. then*

$$\forall A \in \mathcal{E}: \quad P(A) \in \{0, 1\} \tag{18.35}$$

Proof. There is a proof that parallels the classical proof of Kolmogorov's Zero-One Law but we rather plug into the argument from the proof of Theorem 18.5. Indeed, thanks to (18.33) and the Strong Law of Large Numbers

$$E\left(\prod_{j=1}^{k} \left(\frac{1}{n} \sum_{i=1}^{n} 1_{A_{j,i}} \right) \middle| \mathcal{E}_n \right) = \prod_{j=1}^{k} \left(\frac{1}{n} \sum_{i=1}^{n} 1_{A_{j,i}} \right) \underset{n\to\infty}{\longrightarrow} \prod_{j=1}^{k} P(A_{j,1}) \text{ a.s.} \tag{18.36}$$

where $A \in \mathcal{E}_n$. Using this in (18.32) gives

$$P\left(\bigcap_{j=1}^{k} A_{j,i} \middle| \mathcal{E} \right) = \prod_{j=1}^{k} P(A_{j,1}) \text{ a.s.} \tag{18.37}$$

Since events of the form $\bigcap_{j=1}^{k} A_{j,i}$ generate $\mathcal{F}_k := \sigma(X_1, \ldots, X_k)$, it follows that $\mathcal{E}$ and $\mathcal{F}_k$ are independent each $k \geqslant 1$. For each $A \in \mathcal{E}$, the Lévy Zero-One Law then shows $1_A = \lim_{k\to\infty} E(1_A|\mathcal{F}_k) = P(A)$ a.s. implying $P(A) \in \{0, 1\}$. □

A standard application of the Hewitt-Savage Zero-One Law is:

Lemma 18.7 *Let $\{X_k\}_{k\geqslant 0}$ be i.i.d. real-valued and set $S_n := X_1 + \cdots + X_n$. Assume that $P(X_1 \neq 0) > 0$. Then exactly one of the following alternatives hold:*

(1) $S_n \to \infty$ a.s.
(2) $S_n \to -\infty$ a.s.
(3) $\liminf_{n\to\infty} S_n = -\infty \ \wedge \ \limsup_{n\to\infty} S_n = +\infty$ a.s.

Proof. All three alternatives are $\mathcal{E}$-measurable and so, by Theorem 18.6, each occurs with probability one or zero. We claim that $P(\liminf_{n\to\infty} S_n \in \mathbb{R}) = 0$. Indeed, if not then this probability is one and $\liminf_{n\to\infty} S_n$ is constant a.s. Denoting that constant by a, we then also have that $\liminf_{n\to\infty}(S_n - X_1) = a$ a.s. But this is impossible because then $a = X_1 + a$ a.s. which contradicts that $P(X_1 \neq 0) > 0$. Similarly we prove that $P(\limsup_{n\to\infty} S_n \in \mathbb{R}) = 0$ thus giving us (3) unless either (1) or (2) occur. □

Another, albeit much simpler, application for sums of i.i.d. random variables is that

$$\forall a \in \mathbb{R}: \ P(S_n = a \text{ i.o.}) \in \{0, 1\} \tag{18.38}$$

again by the fact that $\{S_n = a \text{ i.o.}\}$ is an exchangeable event.

18.3 Extremal decomposition.

While Theorem 18.5 along with its proof contain the main ideas entering Theorem 18.2, its formulation lacks the clarity and beauty of (18.13). This has been addressed by E. Hewitt and L.J. Savage in 1955 who provided a version of (18.13) for exchangeable random variables taking values in a compact Hausdorff space. We will do the same for random variables taking values in a standard Borel space.

Recall that, given a measurable space $(\mathscr{X}, \Sigma)$, we use $\mathcal{M}_1(\mathscr{X})$ to denote the set of probability measures on $(\mathscr{X}, \Sigma)$. We endow $\mathcal{M}_1(\mathscr{X})$ with the minimal σ-algebra

$$\mathcal{S} := \sigma\Big(\big\{\nu \in \mathcal{M}_1(\mathscr{X})\colon \nu(B) \in A\big\}\colon B \in \Sigma,\, A \in \mathcal{B}(\mathbb{R})\Big) \tag{18.39}$$

that makes $\nu \mapsto \nu(B)$ measurable for all $B \in \Sigma$. Note that the Kolmogorov Extension Theorem asserts that the product measure $\nu^{\otimes\mathbb{N}}$ exists for each $\nu \in \mathcal{M}_1(S)$. As is then easily checked using the π/λ-theorem, the map $\nu \mapsto \nu^{\otimes\mathbb{N}}(B)$ is then measurable (as a map from $(\mathcal{M}_1(S), \mathcal{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$) for all $B \in \Sigma^{\otimes\mathbb{N}}$.

Theorem 18.8 (Hewitt and Savage 1955) *Let $\{X_n\}_{n\geqslant 0}$ be exchangeable random variables taking values in a standard Borel space (S, Σ). Then there exists a unique probability measure Q on $(\mathcal{M}_1(S), \mathcal{S})$ such that*

$$\forall B \in \Sigma^{\otimes\mathbb{N}}\colon \quad P\big((X_0, X_1, \dots) \in B\big) = \int \mu^{\otimes\mathbb{N}}(B)\, Q(\mathrm{d}\mu) \tag{18.40}$$

Proof. The fact that (S, Σ) is standard Borel ensures that there exists a conditional distribution of X_0 given $\mathcal{E}$. Explicitly, there exists a map $\mu_{X_0}\colon \Omega \times \Sigma \to [0,1]$ which is an $\mathcal{E}$-measurable function in the first component, a probability measure in the second component and obeys

$$\forall B \in \Sigma\colon \quad \mu_{X_0}(\cdot, B) = E\big(1_{\{X_0 \in B\}} \,\big|\, \mathcal{E}\big) \text{ a.s.} \tag{18.41}$$

In particular, $\omega \mapsto \mu_{X_0}(\omega, \cdot)$ is an $\mathcal{M}_1(S)$-valued random variable in the above sense. Let Q be the distribution of this map on $(\mathcal{M}_1(S), \mathcal{S})$. Taking expectation in (18.28) and invoking the change of variables formula then shows

$$\begin{aligned} P\Big(\bigcap_{j=0}^{k}\{X_j \in B_j\}\Big) &= \int \prod_{j=0}^{k} \mu_{X_0}(\omega, B_j)\, P(\mathrm{d}\omega) \\ &= \int \prod_{j=0}^{k} \mu(B_j)\, Q(\mathrm{d}\mu) \\ &= \int \mu^{\otimes\mathbb{N}}\big(B_0 \times \cdots \times B_n \times \Sigma \times \dots\big) Q(\mathrm{d}\mu) \end{aligned} \tag{18.42}$$

As this is equality between two measures, the standard argument based on Dynkin's π/λ-theorem extends this to (18.40).

In order to prove uniqueness, suppose that Q is a probability measure on $(\mathcal{M}_1(S), \mathcal{S})$ and let $\{X_n\}_{n\geqslant 0}$ be random variables such that (18.40) holds. (We think of these as realized by coordinate projections on $S^{\mathbb{N}}$, so this is a statement about their distribution on $(S^{\mathbb{N}}, \Sigma^{\otimes N})$.) Since product measures are exchangeable, so is $\{X_n\}_{n\geqslant 0}$. Given any $A \in \Sigma$, denote

$$Z_n(A) := \frac{1}{n} \sum_{k=0}^{n-1} 1_{\{X_k \in A\}} \tag{18.43}$$

Writing $E_{\mu^{\otimes\mathbb{N}}}$ for expectation with respect to $\mu^{\otimes\mathbb{N}}$, for any $A_1, \ldots, A_k \in \Sigma$ and any bounded measurable $f\colon \mathbb{R}^k \to \mathbb{R}$, (18.40) gives

$$E\Big(f\big(Z_n(A_1), \ldots, Z_n(A_k)\big)\Big) = \int E_{\mu^{\otimes\mathbb{N}}}\Big(f\big(Z_n(A_1), \ldots, Z_n(A_k)\big)\Big) Q(\mathrm{d}\mu) \tag{18.44}$$

by a simple application of Tonelli's theorem. Our aim now is to take $n \to \infty$ and produce an identity that identifies Q uniquely.

Since $\{X_n\}_{n\geqslant 0}$ are exchangeable under P, Lemma 18.4 shows

$$Z_n(A) \underset{n\to\infty}{\longrightarrow} Z(A) := E(1_{\{X_1 \in A\}} | \mathcal{E}) \quad P\text{-a.s.} \tag{18.45}$$

For $f\colon \mathbb{R}^k \to \mathbb{R}$ bounded and continuous the Bounded Convergence Theorem then gives

$$E\Big(f\big(Z_n(A_1), \ldots, Z_n(A_k)\big)\Big) \underset{n\to\infty}{\longrightarrow} E\Big(f\big(Z(A_1), \ldots, Z(A_k)\big)\Big) \tag{18.46}$$

On the other hand, given any $\mu \in \mathcal{M}_1(S)$, the Strong Law of Large Numbers is in force under the product measure $\mu^{\otimes\mathbb{N}}$ and so we get

$$Z_n(A) \underset{n\to\infty}{\longrightarrow} \mu(A) \quad \mu^{\otimes\mathbb{N}}\text{-a.s.} \tag{18.47}$$

For f as above the Bounded Convergence Theorem turns this into

$$E_{\mu^{\otimes\mathbb{N}}}\Big(f\big(Z_n(A_1), \ldots, Z_n(A_k)\big)\Big) \underset{n\to\infty}{\longrightarrow} f\big(\mu(A_1), \ldots, \mu(A_k)\big) \tag{18.48}$$

Invoking (18.46) and (18.48) along with the Bounded Convergence Theorem for the integral with respect to Q in (18.44) then shows that

$$E\Big(f\big(Z(A_1), \ldots, Z(A_k)\big)\Big) = \int f\big(\mu(A_1), \ldots, \mu(A_k)\big) Q(\mathrm{d}\mu) \tag{18.49}$$

holds for all bounded continuous $f\colon \mathbb{R}^k \to \mathbb{R}$. Taking limits of f to approximate indicators of open sets in $\mathbb{R}^k$ (which form a π-system) and invoking Dynkin's π/λ-theorem then shows that $(\mu(A_1), \ldots, \mu(A_k))$ under Q is equidistributed to $(Z(A_1), \ldots, Z(A_k))$ under P, for any choice of $k \geqslant 1$ and any $A_1, \ldots, A_k \in \Sigma$. By Dynkin's π/λ-theorem again, this determines Q on $(\mathcal{M}_1(S), \mathcal{S})$ uniquely. □

Corollary 18.9 *Let (S, Σ) be a standard Borel space. Then the set of exchangeable probability measures on $(S^{\mathbb{N}}, \Sigma^{\otimes\mathbb{N}})$ is convex with extremal points being exactly the product measures. The formula* (18.40) *gives an extremal decomposition of an exchangeable law.*

Proof. The invariance of exchangeability under convex combinations is immediate, so all we have to prove is extremality of the product laws. Let $\mu \in \mathcal{M}_1(S)$ and there are two exchangeable laws ρ_1, ρ_2 on $(S^{\mathbb{N}}, \Sigma^{\otimes\mathbb{N}})$ and $\alpha \in (0,1)$ such that

$$\mu^{\otimes\mathbb{N}} = \alpha\rho_1 + (1-\alpha)\rho_2 \tag{18.50}$$

The representation (18.40) shows that $\rho_i = \int \nu^{\otimes\mathbb{N}} Q_i(\mathrm{d}\nu)$, $i = 1, 2$, for some probability measures Q_1 and Q_2 on $(\mathcal{M}_1(S), \mathcal{S})$. Hence,

$$\mu^{\otimes\mathbb{N}}(\cdot) = \int \nu^{\otimes\mathbb{N}}(\cdot) Q(\mathrm{d}\nu) \tag{18.51}$$

for $Q := \alpha Q_1 + (1-\alpha)Q_2$. But the left hand side also equals $\int \nu^{\otimes\mathbb{N}} \delta_\mu(\mathrm{d}\nu)$ and so, by the established uniqueness, $Q = \delta_\mu$. As $\alpha \in (0,1)$ implies $Q_1, Q_2 \ll Q$, both Q_1 and Q_2 are concentrated on $\{\mu\}$. Since they are also probability measures, we have $Q_1 = \delta_\mu = Q_2$ proving that $\mu^{\otimes\mathbb{N}}$ is extremal. □

In their paper, E. Hewitt and L.J. Savage call a probability law on a $S^{\mathbb{N}}$ *presentable* if it takes the form (18.40); their theorem then says that every exchangeable law on (product) standard Borel space is presentable. They asked whether the same is true without a topological assumption on S. This was settled negatively in a 1979-paper of L.E. Dubins and D.A. Freedman (*Zeit. War. v. Geb*, vol. 48, pages 115–132) where they give an example of a probability measure on $(I^{\mathbb{N}}, \mathcal{B}(I)^{\otimes\mathbb{N}})$, with $I := [0,1]$, which is exchangeable but not presentable.

Extremal decompositions of the kind (18.40) appear and are useful in other parts of probability. We will encounter one example in ergodic theory; another setting where this shows up is the theory of Gibbs measures.

19. L^p-MARTINGALES AND SOME INEQUALITIES

We will wrap up the subject of martingales by noting some technical facts about martingales that are quite useful in applications. Specifically, we will study convergence of martingales in L^p using maximal inequalities, prove Doob's decomposition and study square-integrable martingales with the help of their bracket process. Finally, we prove a very useful concentration inequality.

19.1 L^p-martingales.

We start by a simple consequence of the inequality we proved in Lemma 16.2 in preparation for an maximum-inequality based proof of martingale convergence:

Corollary 19.1 (Doob's maximal inequality) *Let $\{X_n\}_{n\geqslant 0}$ be a non-negative submartingale with $\sup_{n\geqslant 0}\|X_n\|_1 < \infty$. Set*

$$X_\star := \sup_{n\geqslant 0} X_n \tag{19.1}$$

Then

$$\forall \lambda > 0\colon \quad P(X_\star > \lambda) \leqslant \frac{1}{\lambda}\sup_{n\geqslant 0}\|X_n\|_1 \tag{19.2}$$

In particular, $X_\star < \infty$ a.s.

Proof. Doob's maximal inequality from Lemma 16.2 gives

$$\lambda P\Big(\max_{0\leqslant j\leqslant n} X_j > \lambda\Big) \leqslant E\Big(X_n 1_{\{\max_{0\leqslant j\leqslant n} X_j > \lambda\}}\Big) \tag{19.3}$$

Bounding the expectation by $\|X_n\|_1$, the left hand side is dominated by $\sup_{n\geqslant 0}\|X_n\|_1$ uniformly in n. The claim follows by taking $n\to\infty$. □

The statement (19.2) can be thought of as a version of the "weak-L^1-inequality" from analysis. Note that we similarly have $\|X_\star\| \leqslant \sup_{n\geqslant 0}\|X_n\|_\infty$ and so we have good control of $X_\star$ in L^∞ and (so called) weak-L^1-space. Calling upon the Marcinkiewicz interpolation theory, such estimates often upgrade to a strong-L^p-inequality for any $p\in(1,\infty)$, and this is the case here as well:

Theorem 19.2 (Doob's L^p-inequality) *Let $p\in(1,\infty)$ and let $\{X_n\}_{n\geqslant 0}$ be a non-negative submartinagle such that $\sup_{n\geqslant 0}\|X_n\|_p < \infty$. Define $X_\star$ as in* (19.1)*. Then*

$$\|X_\star\|_p \leqslant \Big(\frac{p}{1-p}\Big)\sup_{n\geqslant 0}\|X_n\|_p \tag{19.4}$$

In particular, $X_\star \in L^p$.

Proof. Abbreviate $Z_n := \max_{0\leqslant j\leqslant n} X_j$. For each $\lambda > 0$, (19.3) reads

$$P(Z_n > \lambda) \leqslant \frac{1}{\lambda}E\big(X_n 1_{\{Z_n>\lambda\}}\big) \tag{19.5}$$

Now multiply both sides by $p\lambda^{p-1}$ and integrate to get

$$\begin{aligned} E(Z_n^p) &= \int_0^\infty p\lambda^{p-1} P(Z_n > \lambda) \mathrm{d}\lambda \\ &\leqslant \int_0^\infty p\lambda^{p-2} E\big(X_n 1_{\{Z_n > \lambda\}}\big) \mathrm{d}\lambda = \frac{p}{p-1} E\big(X_n Z_n^{p-1}\big) \end{aligned} \tag{19.6}$$

where the equalities are obtained by swapping the integral with respect to λ with the expectation using Tonelli's theorem. Letting q be such that $p^{-1} + q^{-1} = 1$, next we apply Hölder's inequality under the expectation to get

$$E(Z_n^p) \leqslant \frac{p}{p-1} \|X_n\|_p \big[E(Z_n^{q(p-1)})\big]^{1/q} \tag{19.7}$$

Noting that $q(p-1) = p$ this reduces to

$$\big\| \max_{k \leqslant n} X_k \big\|_p \leqslant \frac{p}{p-1} \|X_n\|_p \tag{19.8}$$

Bounding the right-hand side by $\sup_{n \geqslant 0} \|X\|_p$, we now take $n \to \infty$ with the help of the Monotone Convergence Theorem on the left to get the desired statement. □

The above has been phrased for non-negative submartingales as this is what makes the proofs easiest to write. Noting that the absolute value of a martingale is a non-negative submartingale, we thus get:

Corollary 19.3 *Let $p > 1$ and let $\{M_n\}$ be a martingale or a non-negative submartingale such that $\sup_{n \geqslant 0} \|M_n\|_p < \infty$. Then $M_n \to M_\infty := \liminf_{n \to \infty} M_n$ a.s. and in L^p.*

Proof. The assumptions give $\sup_{n \geqslant 0} \|M_n\|_1 < \infty$ and so $M_n \to M_\infty$ a.s. by Theorem 15.1. But $|M_n| \leqslant M_\star := \sup_{n \geqslant 0} |M_n|$ with $M_\star \in L^p$ by (19.4) (or $\|M_\star\|_\infty \leqslant \sup_{n \geqslant 0} \|M_n\|_\infty$ when $p = \infty$) and, since also $|M_\infty| \leqslant M_\star$ a.s. and thus $|M_n - M_\infty| \leqslant 2M_\star$ a.s., Dominated Convergence tells us $\|M_n - M_\infty\|_p \to 0$ proving $M_n \to M_\infty$ in L^p. □

Notice that the previous statement blatantly fails if $p = 1$ because $\sup_{n \geqslant 0} \|M_n\| < \infty$ does not imply L^1-convergence. To get $M_n \to M_\infty$ in L^1 one needs that $\{M_n\}$ is UI which is weaker than $M_\star \in L^1$.

19.2 Square integrable martingales.

The case of L^p-martingales with $p = 2$ is special because L^2 is a Hilbert space. This allows us to state the following:

Lemma 19.4 *Let $\{X_n\}$ be an L^2-martingale (i.e., such that $X_n \in L^2$ for all $n \geqslant 0$) with respect to filtration $\{\mathcal{F}_n\}$. Then*

$$\forall k \leqslant \ell \leqslant m \leqslant n\colon \quad E\big((X_n - X_m)(X_\ell - X_k) \,\big|\, \mathcal{F}_k\big) = 0 \text{ a.s.} \tag{19.9}$$

and, in particular, the increments of X are orthogonal. Moreover,

$$\forall k \leqslant n\colon \quad E(X_n^2 | \mathcal{F}_k) = X_k^2 + E\big((X_n - X_k)^2 \,\big|\, \mathcal{F}_k\big) \tag{19.10}$$

and so $E(X_n^2) = E(X_k^2) + E((X_n - X_k)^2)$.

Proof. The first property is immediate by inserting conditioning on $\mathcal{F}_m$ and noting that $E(X_n - X_m|\mathcal{F}_m) = 0$ by the fact that X is a martingale. The second property follows by writing

$$X_n^2 = \big(X_k + (X_n - X_k)\big)^2 = X_k^2 + 2X_k(X_n - X_k) + (X_n - X_k)^2 \tag{19.11}$$

and taking the conditional expectation given $\mathcal{F}_k$ with the help of (19.9). □

The property (19.10) reflects on $\{X_n^2\}$ being a submartingale but has another interesting consequence: Suppose that $\{M_n\}$ is a process such that

$$M_n - M_{n-1} = X_n^2 - X_{n-1}^2 - E\big((X_n - X_{n-1})^2 \,\big|\, \mathcal{F}_{n-1}\big) \tag{19.12}$$

Then (19.10) implies $E(M_n - M_{n-1}|\mathcal{F}_{n-1}) = 0$ and so $\{M_n\}$ is a martingale! Summing both sides and setting $M_0 := X_0^2$ it is easy to check that

$$M_n = X_n^2 - \sum_{k=1}^{n} E\big((X_k - X_{k-1})^2 \,\big|\, \mathcal{F}_{k-1}\big) \tag{19.13}$$

We will introduce a special notation:

Definition 19.5 (Bracket process) *Given an L^2-martingale $\{X_n\}_{n\geqslant 0}$, its* bracket process *$\{\langle X\rangle_n\}_{n\geqslant 0}$ is a stochastic process defined by*

$$\langle X\rangle_n := \sum_{k=1}^{n} E\big((X_k - X_{k-1})^2 \,\big|\, \mathcal{F}_{k-1}\big) \tag{19.14}$$

with $\langle X\rangle_0 := 0$. (All choices of versions of conditional expectations lead to the same random variable, modulo changes on a null set.)

The notation for the process is taken from its counterpart in stochastic analysis where it also referred to as the *quadratic variation* process. The following summarizes the important properties of the bracket process:

Lemma 19.6 *For any L^2-martingale $\{X_n\}_{n\geqslant 0}$, the bracket process $\{\langle X\rangle_n\}_{n\geqslant 0}$ is predictable and a.s. non-decreasing. The process $\{X_n^2 - \langle X_n\rangle\}_{n\geqslant 0}$ is a martingale.*

Proof. The properties follows from the fact that only $\mathcal{F}_{n-1}$-measurable and a.s. nonnegative quantities appear on the right of (19.14). That $\{X_n^2 - \langle X_n\rangle\}_{n\geqslant 0}$ is a martingale reduces to the calculation after (19.12). □

The bracket process is thus a predictable "compensator" that makes the submartingale $\{X_n^2\}$ into a martingale. Such a process actually exists in large generality:

Theorem 19.7 (Doob decomposition) *Let $\{\mathcal{F}_n\}$ be a filtration and $\{X_n\}$ an adapted process with $X_n \in L^1$ for every $n \geqslant 0$. Then there is a (a.s.) unique martingale $\{M_n, \mathcal{F}_n\}$ and an (a.s.) unique predictable process $\{A_n\}$ such that*

$$A_0 := 0 \quad \wedge \quad \forall n \geqslant 0\colon X_n = M_n + A_n \tag{19.15}$$

for each n. Moreover, if $\{X_n, \mathcal{F}_n\}$ is a submartingale, resp., supermartingale then $\{A_n\}$ is nondecreasing, resp., nonincreasing a.s.

Proof. We use the same argument as above. Define $\{A_n\}_{n\geqslant 0}$ by $A_0 := 0$ and, recursively,

$$A_{n+1} := A_n + E(X_{n+1} - X_n|\mathcal{F}_n) \tag{19.16}$$

An induction argument verifies that $\{A_n\}$ is predictable with $A_n \in L^1$ for each $n \geqslant 0$. Setting $M_n := X_n - A_n$ yields an adapted process with

$$E(M_{n+1} - M_n|\mathcal{F}_n) = E(X_{n+1} - X_n|\mathcal{F}_n) - (A_{n+1} - A_n) = 0 \tag{19.17}$$

and so $\{M_n, \mathcal{F}_n\}$ is a martingale. If $\{X_n\}$ is a submartingale, then $A_{n+1} \geqslant A_n$ (and similarly for $\{X_n\}$ being a supermartingale.)

To show uniqueness, assume that $\{M'_n\}$ is a martingale and $\{A'_n\}$ a predictable process such that $A'_0 = 0$ and $X_n = M'_n + A'_n$ for each $n \geqslant 0$. Then $M''_n := M_n - M'_n = A'_n - A_n$ shows that $\{M_n - M'_n\}$ is a predictable martingale. But then $M''_{n+1} = E(M''_{n+1}|\mathcal{F}_n) = M''_n$ a.s. for each $n \geqslant 0$ implying $M''_n = M''_0 = 0$ a.s. for each $n \geqslant 0$. □

For discrete-time processes, the Doob decomposition is merely a convenient rewrite. However, its counterpart for continuous-time submartingales (called "Doob-Mayer decomposition") is highly non-trivial. An important consequence of the decomposition for square integrable martingales is that

$$T \text{ stopping time} \quad \Rightarrow \quad \forall n \geqslant 0\colon\ \langle M_{T\wedge\cdot}\rangle_n = \langle M\rangle_{T\wedge n} \tag{19.18}$$

and that

$$\langle M\rangle_\infty := \lim_{n\to\infty} \langle M\rangle_n \tag{19.19}$$

always exists, albeit possibly taking an infinite value. The random variable $\langle M\rangle_\infty$ can be used to control moments of the maximal function:

Theorem 19.8 *Given an L^2-martingale $\{M_n\}$., let $M_\star$ be its maximal function and $\langle M\rangle$ its bracket process. Assuming also $M_0 = 0$ a.s.,*

$$E\big(M_\star^2\big) \leqslant 4 \sup_{n\geqslant 0} E(M_n^2) \leqslant 4E\big(\langle M\rangle_\infty\big) \tag{19.20}$$

and

$$\forall a \in (0,1)\colon \quad E\big(M_\star^{2a}\big) \leqslant \frac{2-a}{1-a} E\big(\langle M\rangle_\infty^a\big) \tag{19.21}$$

Proof. For (19.20) we observe that Doob's inequality in the form (19.8) gives

$$E\big(\max_{k\leqslant n} |M_n|^2\big) \leqslant 4E(M_n^2) = 4\big(\langle M\rangle_n\big) \tag{19.22}$$

where we used that $\{M_n^2 - \langle M\rangle_n\}$ is a martingale vanishing at $n = 0$. Taking $n \to \infty$ on both sides using the Monotone Convergence Theorem yields (19.20).

For (19.21) pick $\lambda > 0$ and denote

$$T := \inf\big\{n \geqslant 0\colon \langle M\rangle_{n+1} > \lambda^2\big\} \tag{19.23}$$

The predictability of $\langle M\rangle$ ensures that T is a stopping time. Note that then

$$P(T < \infty) = P\big(\langle M\rangle_\infty > \lambda^2\big) \tag{19.24}$$

Using T to define a stopped martingale $\{M_{T\wedge n}\}$, Doob's inequality along with the fact that $\{M^2_{T\wedge n} - \langle M\rangle_{T\wedge n}\}$ is a martingale and $\langle M\rangle_{T\wedge n} \leqslant \lambda^2$ in turn gives

$$\begin{aligned} P\big(\max_{k\leqslant n}|M_{T\wedge k}| > \lambda\big) &\leqslant \frac{1}{\lambda^2}E(M^2_{T\wedge n}) \\ &= \frac{1}{\lambda^2}E\big(\langle M\rangle_{T\wedge n}\big) \leqslant \frac{1}{\lambda^2}E\big(\langle M\rangle_\infty \wedge \lambda^2\big) \end{aligned} \tag{19.25}$$

Hence we get

$$\begin{aligned} P(M_\star > \lambda) &\leqslant P(T<\infty) + P\big(\sup_{k\geqslant 0}|M_{T\wedge k}| > \lambda\big) \\ &\leqslant P\big(\langle M\rangle_\infty > \lambda^2\big) + \frac{1}{\lambda^2}E\big(\langle M\rangle_\infty \wedge \lambda^2\big) \end{aligned} \tag{19.26}$$

We now multiply both sides by $2a\lambda^{2a-1}$ and integrate over $\lambda \in \mathbb{R}_+$. On the left this produces $E(M^{2a}_\star)$. On the right we get two integrals, the first of which is

$$\int_0^\infty 2a\lambda^{2a-1}P\big(\langle M\rangle_\infty > \lambda^2\big)\mathrm{d}\lambda \overset{t=\lambda^2}{=} \int_0^\infty at^{a-1}P\big(\langle M\rangle_\infty > t\big)\mathrm{d}t = E\big(\langle M\rangle^a_\infty\big) \tag{19.27}$$

The second integral is treated as

$$\begin{aligned} \int_0^\infty 2a\lambda^{2a-1}\frac{1}{\lambda^2}E\big(\langle M\rangle_\infty \wedge \lambda^2\big)\mathrm{d}\lambda &= 2a\int_0^\infty \lambda^{2a-3}\left(\int_0^{\lambda^2} P\big(\langle M\rangle_\infty > t\big)\mathrm{d}t\right)\mathrm{d}\lambda \\ &= 2a\int_0^\infty\left(\int_{\sqrt{t}}^\infty \lambda^{2a-3}\mathrm{d}\lambda\right)P\big(\langle M\rangle_\infty > t\big)\mathrm{d}t \\ &= \frac{2a}{2-2a}\int_0^\infty t^{a-1}P\big(\langle M\rangle_\infty > t\big)\mathrm{d}t = \frac{1}{1-a}E\big(\langle M\rangle^a_\infty\big) \end{aligned} \tag{19.28}$$

where we used Tonelli's theorem to exchange the two integrals and then integrated using that $2a-3 < -1$. Putting (19.27–19.28) together we then get (19.21). □

We remark that the inequalities (19.20–19.21) are precursors of *Burkholder-Davis-Gundy inequalities* in stochastic analysis where, for continuous-time process, one can squeeze $E(M^{2a}_\star)$ by upper and lower bounds using $E(\langle M\rangle^a_\infty)$ with universal (positive and finite constants). For discrete time processes, these bounds hold as well albeit with $\langle M\rangle_\infty$ replaced by the *quadratic variation* process,

$$Q_\infty := \sum_{k\geqslant 1}(M_k - M_{k-1})^2 \tag{19.29}$$

which for continuous martingales (indexed by continuous time) can be related to $\langle M\rangle_\infty$ by way of refinements of discrete-time approximations.

19.3 Some applications.

The conclusions for the maximal function in Theorem 19.8 have a number of interesting consequences. We start with:

Corollary 19.9 *Let $\{M_n\}$ be a square integrable martingale and let $\{\langle M\rangle_n\}$ denote its bracket process. Set $M_\infty := \limsup_{n\to\infty} M_n$. Then*

$$|M_\infty| < \infty \quad \wedge \quad M_n \underset{n\to\infty}{\longrightarrow} M_\infty \quad \text{a.s. on } \bigl\{\langle M\rangle_\infty < \infty\bigr\} \tag{19.30}$$

Proof. Fix $\lambda > 0$ and recall the definition (19.23). For the stopped martingale $\{M_{T\wedge n}\}$ we have $\langle M_{T\wedge\cdot}\rangle_n = \langle M\rangle_{T\wedge n} \leqslant \lambda^2$ and so (19.20) gives $(M_{T\wedge\cdot})^\star \in L^1$. It follows that $\{M_{T\wedge n}\}$ is UI and so $\lim_{n\to\infty} M_{T\wedge n}$ exists in $\mathbb{R}$ a.s. But that means that $M_n \to M_\infty$ with $|M_\infty| < \infty$ a.s. on $\{T = \infty\} = \{\langle M\rangle_\infty \leqslant \lambda^2\}$. Taking a union over $\lambda \in \mathbb{N}$, the same holds on $\{\langle M\rangle_\infty < \infty\}$. □

To see why Corollary 19.9 is useful, let us use it to generalize the Borel-Cantelli lemmas (Lemmas 2.5–2.6):

Lemma 19.10 (Lévy's extension of Borel-Cantelli lemmas) *Let $\{A_k\}_{k\geqslant 1}$ be a sequence of events and $\{\mathcal{F}_k\}_{k\geqslant 0}$ a filtration such that $\forall k \geqslant 1\colon A_k \in \mathcal{F}_k$. Then*

$$\sum_{k=1}^\infty 1_{A_k} < \infty \quad \text{a.s. on } \Biggl\{ \sum_{k=1}^\infty P(A_k|\mathcal{F}_{k-1}) < \infty \Biggr\} \tag{19.31}$$

and

$$\sum_{k=1}^\infty 1_{A_k} = \infty \quad \text{a.s. on } \Biggl\{ \sum_{k=1}^\infty P(A_k|\mathcal{F}_{k-1}) = \infty \Biggr\} \tag{19.32}$$

Proof. Denote

$$M_n := \sum_{k=1}^n \bigl[1_{A_k} - P(A_k|\mathcal{F}_{k-1})\bigr] \tag{19.33}$$

Then $\{M_n\}$ is a martingale with bounded increments and, in particular, is a square integrable martingale. For the bracket process we get

$$\langle M\rangle_n = \sum_{k=1}^n \bigl[P(A_k|\mathcal{F}_{k-1}) - P(A_k|\mathcal{F}_{k-1})^2\bigr] \leqslant \sum_{k=1}^n P(A_k|\mathcal{F}_{k-1}) \tag{19.34}$$

It follows that $\langle M\rangle_\infty < \infty$ on $\{\sum_{k\geqslant 1} P(A_k|\mathcal{F}_{k-1}) < \infty\}$. On the latter event $\{M_n\}$ converges and so $\sup_{n\geqslant 1}|M_n| < \infty$ a.s. which implies $\sum_{k\geqslant 1} 1_{A_k} < \infty$ a.s., proving (19.31).

For (19.32), given any $\lambda > 0$ denote

$$N_\lambda := \inf\Biggl\{ n \geqslant 0\colon \sum_{k=1}^n 1_{A_k} > \lambda \Biggr\} \tag{19.35}$$

Since $\{1_{A_k}\}$ is adapted to $\{\mathcal{F}_k\}$, this is a stopping time. Moreover, the fact that the increments of M are bounded by 1 shows $M_{N_\lambda\wedge n} \leqslant \lambda + 1$ a.s. Theorem 15.1 then asserts a.s.-existence and finiteness of $\lim_{n\to\infty} M_{N_\lambda\wedge n}$ a.s. On $\{\sum_{k\geqslant 1} P(A_k|\mathcal{F}_k) = \infty\}$ the limit can be finite only if $N_\lambda < \infty$. Hence we must have $\sum_{k\geqslant 1} 1_{A_k} > \lambda$ for each $\lambda > 0$ a.s. on $\{\sum_{k\geqslant 1} P(A_k|\mathcal{F}_k) = \infty\}$ proving (19.32). □

We can strengthen Corollary 19.9 as follows:

Theorem 19.11 (Strong law for martingales) *Let $\{M_n\}$ be an L^2-martingale. We then have*

$$\frac{M_n}{\langle M\rangle_n} \underset{n\to\infty}{\longrightarrow} 0 \text{ a.s. on } \big\{\langle M\rangle_\infty = \infty\big\} \tag{19.36}$$

Proof. Abbreviate $A_n := \langle M\rangle_n$ and set $X := ((1+A)^{-1}\cdot M)$; i.e.,

$$X_n := \sum_{k=1}^{n} \frac{1}{1+A_k}(M_k - M_{k-1}) \tag{19.37}$$

Since $1 + A_k \geqslant 1$, the summands on the right are in L^1 and X is a martingale. Moreover, using that A is predictable we get

$$\begin{aligned} E\big((X_k - X_{k-1})^2 \,\big|\, \mathcal{F}_{k-1}\big) &= \frac{1}{(1+A_k)^2} E\big((M_k - M_{k-1})^2 \,\big|\, \mathcal{F}_{k-1}\big) \\ &= \frac{A_k - A_{k-1}}{(1+A_k)^2} \leqslant \frac{A_k - A_{k-1})}{(1+A_k)(1+A_{k-1})} \\ &= \frac{1}{1+A_{k-1}} - \frac{1}{1+A_k} \end{aligned} \tag{19.38}$$

Summing the right-hand side on $k \geqslant 1$ gives 1 and so $\langle X\rangle_\infty \leqslant 1$. Corollary 19.9 gives $X_n \to 0$ a.s. But then Kronecker's lemma (Lemma 2.14) implies

$$\frac{M_n - M_0}{1+A_n} \underset{n\to\infty}{\longrightarrow} 0 \text{ a.s. on } \{A_\infty = \infty\} \tag{19.39}$$

which now readily gives the claim. □

Using the previous lemma one can show that the series on the right of (19.32) diverges at exactly the same rate as the series on the right-diverges; namely:

Corollary 19.12 *For events $\{B_k\}$ and filtration $\{\mathcal{F}_k\}$ such that $\forall k \geqslant 1\colon\ B_k \in \mathcal{F}_k$, we have*

$$\frac{\displaystyle\sum_{k=1}^{n} 1_{A_k}}{\displaystyle\sum_{k=1}^{n} P(A_k|\mathcal{F}_{k-1})} \underset{n\to\infty}{\longrightarrow} 1 \quad \text{a.s. on } \left\{\sum_{k=1}^{\infty} P(A_k|\mathcal{F}_{k-1}) = \infty\right\} \tag{19.40}$$

Proof. Assume that $\sum_{k=1}^{\infty} P(A_k|\mathcal{F}_{k-1}) = \infty$ and let M_n be the martingale (19.33). On the event $\{\langle M\rangle_\infty < \infty\}$ we get $M_\star < \infty$ a.s. and so the two sums constituting M_n have to grow at the same rate. On $\{\langle M\rangle_\infty = \infty\}$ we in turn get $M_n/\langle M\rangle_n \to 0$ from Theorem 19.11 and the same argument applies as well. □

We note that for independent events, convergence in probability in (19.40) is readily shown by way of Chebyshev' inequality.

Another consequence of Corollary 19.9 comes in:

Lemma 19.13 *Consider the Galton-Watson process $\{S_n\}_{n\geqslant 0}$ with $S_0 = 1$ and off-spring distribution $\{\mathfrak{p}(n)\}_{n\geqslant 0}$. Suppose that $\mu := \sum_{k\geqslant 0} k\mathfrak{p}(k) > 1$ and $\sum_{k\geqslant 1} k^2\mathfrak{p}(k) < \infty$. Then there*

exists a random variable $Z \in L^2$ with

$$E(Z) = 1 \quad \text{and} \quad E(Z^2) = \frac{1}{\mu(\mu-1)}\Big(\sum_{k\geqslant 1} k^2 \mathfrak{p}(k) - \mu^2\Big) \tag{19.41}$$

such that

$$S_n \mu^{-n} \underset{n\to\infty}{\longrightarrow} Z \text{ a.s. and in } L^2 \tag{19.42}$$

Proof. Write $\sigma^2 := \sum_{k\geqslant 1} k^2\mathfrak{p}(k) - \mu^2$. We know that $M_n := S_n\mu^{-n}$ is a martingale with $EM_0 = 1$. Using that S_k conditioned on $\mathcal{F}_{k-1}$ equals the sum of S_{k-1} independent random variables with variance σ^2, we get

$$\begin{aligned} E\big((M_k - M_{k-1})^2 \,\big|\, \mathcal{F}_{k-1}\big) &= \mathrm{Var}(M_k|\mathcal{F}_{k-1}) \\ &= \mu^{-2k}\mathrm{Var}(S_k|\mathcal{F}_{k-1}) = \sigma^2 S_{k-1}\mu^{-2k} \end{aligned} \tag{19.43}$$

Hence,

$$\langle M\rangle_\infty = \sigma^2\mu^{-2}\sum_{k\geqslant 0}\mu^{-2k}S_k \tag{19.44}$$

Since $E(S_k) = \mu^k$ and $\mu > 1$, the Monotone Convergence Theorem gives $E(\langle M\rangle_\infty) < \infty$ and so $M_\star \in L^2$. Hence, $M_n \to Z := \limsup_{n\to\infty} M_n$ a.s. and in L^2. The claim follows from $E(Z) = \lim_{n\to\infty} E(M_n) = E(M_0) = 1$ and $E(M_n^2) = E(\langle M\rangle_n) \to E(\langle M\rangle_\infty)$. □

An interesting consequence of the fractional-moment bound (19.21) comes in the following addition to the statements in Theorem 14.11:

Corollary 19.14 (A Wald's-type equation) *Let $\{X_k\}_{k\geqslant 1}$ be i.i.d. with $X_1 \in L^2$ and $EX_1 = 0$. Set $S_n = X_1 + \cdots + X_n$ and denote $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$. Then for any stopping time T for the filtration $\{\mathcal{F}_n\}$,*

$$T \in L^{1/2} \quad \Rightarrow \quad S_T \in L^1 \ \wedge\ ES_T = 0 \tag{19.45}$$

Proof. Under the assumptions $\{S_n\}$ is an L^2-martingale with $\langle S\rangle_n = E(X_1^2)n$. It follows that $M_n := S_{T\wedge n}$ is an L^2-martingale with $\langle M\rangle_n = E(X_1^2)(T\wedge n)$ and, in particular, $\langle M\rangle_\infty = E(X_1^2)T$. The condition $E(T^{1/2}) < \infty$ thus gives $E(\langle M\rangle_\infty^{1/2}) < \infty$ which by (19.21) implies $M_\star \in L^1$. Since $T < \infty$ a.s. implies $S_{T\wedge n} \to S_T$ and $|S_T| \leqslant M_\star$ a.s., we get $S_T \in L^1$ and $S_{T\wedge n} \to S_T$ in L^1. The claim follows from $E(S_{T\wedge n}) = ES_0 = 0$. □

19.4 Azuma-Hoeffding inequality.

The above estimates control the moments of martingales. One can do even better if the martingale has bounded increments:

Theorem 19.15 (Azuma-Hoeffding inequality) *Let $\{M_n\}$ be a submartingale with $M_0 = 0$ such that, for some non-random sequence $\{c_k\}$ of non-negative numbers,*

$$\forall k = 1, \ldots, n\colon \quad |M_k - M_{k-1}| \leqslant c_k \text{ a.s.} \tag{19.46}$$

Then

$$\forall \lambda \geqslant 0\colon \quad P\big(M_n > \lambda\big) \leqslant \exp\Big\{-\frac{1}{2}\frac{\lambda^2}{\sum_{k=1}^n c_k^2}\Big\}. \tag{19.47}$$

Proof. The proof is based on the following inequality

$$\forall t \in \mathbb{R}\, \forall y \in [-c, c]: \quad \mathrm{e}^{ty} \leqslant \cosh(tc) + \frac{y}{c}\sinh(tc) \tag{19.48}$$

This follows from the fact that the right-hand side can be written as

$$\cosh(tc) + \frac{y}{c}\sinh(tc) = \mathrm{e}^{tc}\frac{c+y}{2c} + \mathrm{e}^{-tc}\frac{c-y}{2c} \tag{19.49}$$

Under the condition $|y| \leqslant c$, both $\frac{c+y}{2c}$ and $\frac{c-y}{2c}$ are non-negative and they add up to one. So, by the convexity of the exponential, the right-hand side is at most the exponential of

$$tc\frac{c+y}{2c} - tc\frac{c-y}{2c} = ty \tag{19.50}$$

and so we get (19.48).

To see how (19.48) implies the desired claim, consider the random variable e^{tM_n} for any $t \geqslant 0$. Since M_n is bounded, this is integrable and so

$$E\big(\mathrm{e}^{tM_k}\big) = E\big(\mathrm{e}^{tM_{k-1}}E(\mathrm{e}^{t(M_k - M_{k-1})}|\mathcal{F}_k)\big) \tag{19.51}$$

But (19.48) ensures that

$$E(\mathrm{e}^{t(M_k - M_{k-1})}|\mathcal{F}_k) \leqslant \cosh(tc_k) + \frac{1}{c_k}\sinh(tc_k)E(M_k - M_{k-1}|\mathcal{F}_k) \tag{19.52}$$

and this is equal to $\cosh(tc_k)$ for martingales and $\geqslant \cosh(tc_k)$ for submartingales since $t \geqslant 0$. Hence, using induction,

$$\forall t \geqslant 0: \quad E\big(\mathrm{e}^{tM_k}\big) \leqslant \prod_{k=1}^{n} \cosh(tc_k) \tag{19.53}$$

But $x \mapsto \log\cosh(x)$ is symmetric, zero at $x = 0$ and with a decreasing third derivative and so $\log\cosh(x) \leqslant \frac{1}{2}x^2$. It follows that

$$E\big(\mathrm{e}^{tM_k}\big) \leqslant \exp\Big\{\frac{t^2}{2}\sum_{k=1}^{n} c_k^2\Big\}, \qquad t \geqslant 0. \tag{19.54}$$

Now use the exponential Chebyshev inequality to write

$$\forall t \geqslant 0: \quad P\big(M_n > \lambda\big) \leqslant \mathrm{e}^{-t\lambda}E\big(\mathrm{e}^{tM_k}\big) \tag{19.55}$$

Plugging the above bound and minimizing over t shows that the optimal value to use is $t := \lambda(\sum_{k=1}^{n} c_k^2)^{-1}$. Doing so yields the desired inequality. □

Several generalizations exists that permit relaxing the strict boundedness requirement. However, none is such that it would simply permit to drop the boundedness altogether and replace $\sum_{k=1}^{n} c_k^2$ by the expectation of M_n^2.

For submartingales, one gets the bound on the upper tail of M_n. For martingales, the symmetry yields two-sided control:

$$P\big(|M_n| > \lambda\big) \leqslant 2\exp\Big\{-\frac{1}{2}\frac{\lambda^2}{\sum_{k=1}^{n} c_k^2}\Big\}. \tag{19.56}$$

In particular, a typical value of M_n will be at most order $\sqrt{\sum_{k=1}^n c_k^2}$. Bounds of this form belong to the area of *concentration of measure*.

Further reading: Durrett, Sections 4.4 and 4.5