401-3601-00L — PROBABILITY THEORY

Dominik Schröder[*]

CONTENTS

[*] dschroeder@ethz.ch

2 ABSTRACT GENERALITIES: RANDOM VARIABLES

2.3 *Laws of (collections of) random variables*

Definition 2.1 (Law of scalar random variables). Let X be a real-valued random variable X on some probability space $(\Omega, \mathcal{A}, \mathrm{P})$.

(i) We define the *distribution of* X to be the probability measure μ_{X} on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ by setting $\mu_{\mathrm{X}}(\mathrm{B}) = \mathrm{P}(\mathrm{X} \in \mathrm{B})$ for Borel sets B.

(ii) We define the *(cumulative) distribution function of* X to be the function $\mathrm{F}\colon \mathbb{R} \to [0,1]$ such that $\mathrm{F}(x) = \mu_{\mathrm{X}}((-\infty, x]) = \mathrm{P}(\mathrm{X} \leq x)$.

Proof that μ_{X} *is a probability measure.* We check that

- $\mu_{\mathrm{X}}(\mathbb{R}) = \mathrm{P}(\mathrm{X} \in \mathbb{R}) = 1$.
- For a sequence $\mathrm{B}_1, \mathrm{B}_2, \ldots$ of pairwise disjoint Borel sets $\mathrm{B}_i \in \mathcal{B}_{\mathbb{R}}$ the events $(\{\mathrm{X} \in \mathrm{B}_i\})_{i \in \mathbb{N}}$ are clearly pairwise disjoint and therefore

$$\begin{aligned}\sum_{i \in \mathbb{N}} \mu_{\mathrm{X}}(\mathrm{B}_i) = \sum_i \mathrm{P}(\mathrm{X} \in \mathrm{B}_i) &= \mathrm{P}\Big(\bigcup_{i \in \mathbb{N}} \{\mathrm{X} \in \mathrm{B}_i\}\Big) = \mathrm{P}\Big(\mathrm{X} \in \bigcup_{i \in \mathbb{N}} \mathrm{B}_i\Big) \\ &= \mu_{\mathrm{X}}\Big(\bigcup_{i \in \mathbb{N}} \mathrm{B}_i\Big).\end{aligned} \tag{1}$$

to conclude the proof. □

Lemma 2.2 (Change of variables). *For measurable functions* $f\colon \mathbb{R} \to \mathbb{R}$ *it holds that*

$$\int_{\mathbb{R}} f \, \mathrm{d}\mu_{\mathrm{X}} = \mathbf{E} f(\mathrm{X}) = \int_{\Omega} f \circ \mathrm{X} \, \mathrm{dP} \tag{2}$$

whenever $f \geq 0$ or $\int |f| \, \mathrm{d}\mu_{\mathrm{X}} < \infty$.

Proof. First, for $f = \mathbf{1}_{\mathrm{A}}$ for some Borel set $\mathrm{A} \in \mathcal{B}(\mathbb{R})$ we have

$$\int_{\mathbb{R}} \mathbf{1}_{\mathrm{A}} \, \mathrm{d}\mu_{\mathrm{X}} = \mu_{\mathrm{X}}(\mathrm{A}) = \mathrm{P}(\mathrm{X}^{-1}(\mathrm{A})) = \int_{\Omega} \mathbf{1}_{\mathrm{X}^{-1}(\mathrm{A})} \, \mathrm{dP} = \int_{\Omega} \mathbf{1}_{\mathrm{A}} \circ \mathrm{X} \, \mathrm{dP} \tag{3}$$

where $\mathrm{X}^{-1}(\mathrm{A})$ denotes the pre-image of A, and we used that $\mathbf{1}_{\mathrm{X}^{-1}(\mathrm{A})} = \mathbf{1}_{\mathrm{A}} \circ \mathrm{X}$ since $\mathbf{1}_{\mathrm{X}^{-1}(\mathrm{A})}(x) = 1$ if and only if $x \in \mathrm{X}^{-1}(\mathrm{A})$ if and only if $\mathrm{X}(x) \in \mathrm{A}$. By linearity this proves the lemma for any simple function $f = \sum_i \alpha_i \mathbf{1}_{\mathrm{A}_i}$. Monotone convergence now implies the result for $f \geq 0$, and finally if $\int |f| \, \mathrm{d}\mu_{\mathrm{X}} < \infty$ we decompose $f = f_+ - f_-$ and again use linearity. □

Lemma 2.3 (Properties of distribution functions). *Any distribution function* $\mathrm{F}\colon \mathbb{R} \to [0,1]$ *satisfies*

(a) $\mathrm{F}(x)$ *is non-decreasing, i.e.* $a \leq b \Rightarrow \mathrm{F}(a) \leq \mathrm{F}(b)$

(b) $\lim_{x \to -\infty} \mathrm{F}(x) = 0$, $\lim_{x \to \infty} \mathrm{F}(x) = 1$

(c) F *is right-continuous everywhere*

Proof. Using elementary facts about the probability measure μ_X we check:

(a) This follows since $\{X \leq a\} \subseteq \{X \leq b\}$.

(b) We have

$$\lim_{x\to-\infty} F(x) = \lim_{\mathbb{N}\ni n\to\infty} F(-n) = \mu_X\Big(\bigcap_{n\in\mathbb{N}} \{X \leq -n\}\Big) = \mu_X(\emptyset) = 0 \tag{4}$$

and

$$\lim_{x\to\infty} F(x) = \lim_{\mathbb{N}\ni n\to\infty} F(n) = \mu_X\Big(\bigcup_{n\in\mathbb{N}} \{X \leq n\}\Big) = \mu_X(\mathbb{R}) = 1. \tag{5}$$

(c) Since

$$\bigcap_{n\in\mathbb{N}} \{X \in (-\infty, a+1/n]\} = \{X \in (-\infty, a]\} \tag{6}$$

it follows that

$$\lim_{n\to\infty} F(a+1/n) = P\Big(\bigcap_{n\in\mathbb{N}} \{X \leq a+1/n\}\Big) = P(X \leq a) = F(a). \tag{7}$$

□

Proposition 2.4 (Lebesgue-Stieltjes Lemma)**.** *If some function* $F\colon \mathbb{R} \to [0,1]$ *satisfies (a) to (c) of Lemma 2.3, then* F *is the distribution function of some random variable* X*. Moreover* F *is the distribution function of a unique probability measure* μ *on* $(\mathbb{R}, \mathcal{B}_\mathbb{R})$.

Proof. Set $\Omega = (0,1)$, let $\mathcal{B} = \mathcal{B}_{(0,1)}$ be the Borel σ-algebra and $P\colon \mathcal{B} \to [0,1]$ be the Lebesgue measure. We define the random variable $X\colon \Omega \to \mathbb{R}$ by

$$X(\omega) = \sup\{y \mid F(y) < \omega\}. \tag{8}$$

If $\omega > F(x)$, then by right-continuity $\omega > F(x+\epsilon)$ and therefore $X(\omega) \geq x + \epsilon > x$. On the other hand, if $\omega \leq F(x)$, then $X(\omega) \leq x$ and therefore

$$\{\omega \mid X(\omega) \leq x\} = \{\omega \mid \omega \leq F(x)\} \tag{9}$$

Thus $P(X \leq x) = P(\omega \leq F(x)) = F(x)$ and the claim follows.

By setting $\mu := \mu_X$ we have proved the existence of a probability measure. The uniqueness of such a measure has already been observed in Corollary 1.1.19. □

Definition 2.5. We say that two real-valued random variables X, Y are *equal in distribution* and write $X \overset{d}{=} Y$ if $P(X \leq x) = P(Y \leq x)$ for all $x \in \mathbb{R}$.

Definition 2.6.

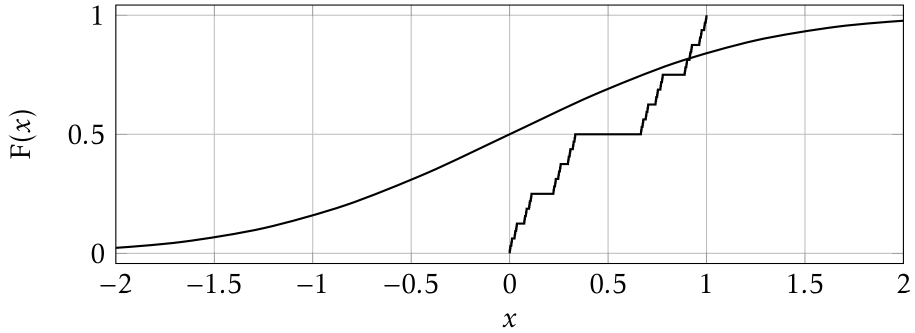

Figure 1: Distribution function for the Gaussian and Cantor distribution

- We say that a probability measure μ on $\mathbb{R}$ has *density function* $f : \mathbb{R} \to [0, \infty)$ if its distribution function F is such that

$$\mathrm{F}(x) = \mu((-\infty, x]) = \int_0^x f(t)\,\mathrm{d}t \tag{10}$$

  for all x.

- We say that a probability measure μ on $\mathbb{R}$ is *discrete* if it is concentrated on some countable set $\mathrm{C} \subset \mathbb{R}$, i.e. $\mu(\mathrm{C}) = 1$.

Example 2.7 (Gaussian distribution). The Gaussian distribution is the distribution with density

$$f(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) \tag{11}$$

The distribution function

$$\mathrm{F}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-t^2/2)\,\mathrm{d}t = \frac{1 + \mathrm{erf}(x/\sqrt{2})}{2} \tag{12}$$

has no closed form expression.

Example 2.8 (Uniform distribution on the Cantor set). Recall that the Cantor set on $[0, 1]$ is constructed by repeatedly removing the middle third of each interval, i.e. in the first step we remove $(1/3, 2/3)$, in the second step we remove $(1/9, 2/9)$ and $(7/9, 8/9)$ etc. We can then define a *continuous* distribution function $\mathrm{F}(x)$ by setting $\mathrm{F}(x) = 1/2$ for $x \in (1/3, 2/3)$ in the first step, $\mathrm{F}(x) = 1/4$ for $x \in (1/9, 2/9) \cup (7/9, 8/9)$ in the second step etc. This distribution has no density as any such density would have to be identically 0 Lebesgue almost everywhere. However, the distribution is also not discrete as every point $x \in [0, 1]$ has measure 0.

3 SEQUENCES, SERIES AND MEANS OF INDEPENDENT RANDOM VARIABLES

3.1 *Existence*

Proposition 3.1. *For any given sequence of distribution functions* $(F_i)_{i\geq 1}$*, it is possible to find a probability space* $(\Omega, \mathcal{A}, P)$ *and a sequence of independent random variables* $(Y_i)_{i\geq 1}$ *defined on this space such that for each* i*, the distribution function of* Y_i *is* F_i*.*

Proof. We have seen in the exercise that if U is a uniformly distributed random variable on $[0,1)$, then $\epsilon_n := \lfloor 2^n U \rfloor$ forms an i.i.d. sequence $(\epsilon_n)_{n\geq 1}$ of $\{0,1\}$-Bernoulli random variables.

Now consider any bijection $\phi \colon \mathbb{N}^2 \to \mathbb{N}$ and define the random variable

$$X_i := \sum_{j\geq 1} \epsilon_{\phi(i,j)} 2^{-j}. \tag{13}$$

We observe that the law of each X_i is the Lebesgue measure since $P(X_i \in [j2^{-k}, (j+1)2^{-k}]) = 2^{-k}$ for each $k \geq 1$ and $j < 2^k$, and the set of such dyadic intervals forms a π-system generating the Lebesgue measure. The sequence $(X_i)_{i\geq 1}$ is independent by construction since each X_i depends on a disjoint set of independent ϵ_n random variables. Finally, we define

$$f_i(x) := \sup\{y \in \mathbb{R} \mid F_i(y) < x\} \tag{14}$$

and set $Y_i := f_i(X_i)$ and recall from Proposition 2.4 that each Y_i is indeed a random variable with distribution function F_i. □

7 CONVERGENCE OF PROBABILITY MEASURES, CHARACTERISTIC FUNCTIONS AND THE CENTRAL LIMIT THEOREM

7.1 *Definition of weak convergence*

Lecture 1
22nd November 2022

We are now going to study a very different type of questions than in the previous chapters. So far, we have mostly been working in a given probability space $(\Omega, \mathcal{A}, P)$ and looking at sequences of random variables defined in this space, and at their convergence when $n \to \infty$. In all cases (almost sure convergence, convergence in probability, L^p convergence), this notion was implying convergence in probability, i.e., that the *realizations* of the random variables X_n and X were likely to be close when n was large.

We are now going instead going to look at sequences of probability measures defined on a given space, and discuss their convergence. While (as we will see in a moment), convergence of probability measures are interesting (and useful to study) on more general metric spaces, we will focus here mostly on sequences of probability measures on $\mathbb{R}$ (we will then briefly discuss probability measures on $\mathbb{R}^d$) endowed with the Borel σ-field.

Definition 7.1 (Weak convergence)**.** We say that a sequence of probability measures $(\mathrm{P}_n)_{n\geq 1}$ on a metric space (E, d) (endowed with its Borel σ-field) converges weakly to a probability measure P (on this same space), if for any continuous bounded function $f\colon \mathrm{E} \to \mathbb{R}$,

$$\lim_{n\to\infty} \int f(x)\,\mathrm{dP}_n(x) = \int f(x)\,\mathrm{dP}(x).$$

Remark 7.2. There will be no notion of strong convergence of probability measures. This notion of convergence will essentially the only one that we will be discussing!

Remark 7.3. We will only be dealing with the cases $\mathrm{E} = \mathbb{R}$ and $\mathrm{E} = \mathbb{R}^d$ in these lectures, even if many of the results that we will be discussing can actually extended to the case of separable complete metric spaces.

Remark 7.4. The use of continuous functions here is quite natural. Indeed, one wants for instance to say that the sequence of Dirac masses at $1/n$ converges weakly to the Dirac mass at 0, and this indeed holds because $f(1/n) \to f(0)$ when f is continuous.

Theorem 7.5 (Portemanteau)**.** *Let* $\mathrm{P}, \mathrm{P}_1, \mathrm{P}_2, \ldots$ *be probability measures on some metric space* (E, d)*. Then the following are equivalent:*

(i) $\mathrm{P}_n \to \mathrm{P}$ *weakly*

(ii) $\liminf_n \mathrm{P}_n(\mathrm{O}) \geq \mathrm{P}(\mathrm{O})$ *for all open* $\mathrm{O} \subseteq \mathrm{E}$

(iii) $\limsup_n \mathrm{P}_n(\mathrm{C}) \leq \mathrm{P}(\mathrm{C})$ *for all closed* $\mathrm{C} \subseteq \mathrm{E}$

(iv) $\lim_n \mathrm{P}_n(\mathrm{A}) = \mathrm{P}(\mathrm{A})$ *for all measurable* $\mathrm{A} \subseteq \mathrm{E}$ *with* $\mathrm{P}(\partial\mathrm{A}) = 0$

Proof. For the implication (i) ⇒ (ii) fix an open set $\mathrm{O} \subseteq \mathrm{E}$. Then for any continuous $0 \leq f \leq \mathbf{1}_\mathrm{O}$ we have $\int f(x)\,\mathrm{dP}(x) = \liminf_n \int f(x)\,\mathrm{dP}_n(x) \leq \liminf_n \mathrm{P}_n(\mathrm{O})$ and by taking $f \uparrow \mathbf{1}_\mathrm{O}$ it follows that $\liminf_n \mathrm{P}_n(\mathrm{O}) \geq \mathrm{P}(\mathrm{O})$ my monotone convergence.

The equivalence in (ii) ⇔ (iii) ⇒ (iv) is trivial by taking complements. For the second part note that for any measurable set $\mathrm{A} \subseteq \mathrm{E}$ by (ii) and (iii) it holds that

$$\begin{aligned}\mathrm{P}(\mathrm{A}^\circ) &\leq \liminf_n \mathrm{P}_n(\mathrm{A}^\circ) \leq \liminf_n \mathrm{P}_n(\mathrm{A}) \leq \limsup_n \mathrm{P}_n(\mathrm{A}) \\ &\leq \limsup_n \mathrm{P}_n(\overline{\mathrm{A}}) \leq \mathrm{P}(\overline{\mathrm{A}})\end{aligned}$$

and therefore $\mathrm{P}(\mathrm{A}) = \lim_n \mathrm{P}_n(\mathrm{A})$ whenever $\mathrm{P}(\partial\mathrm{A}) = 0$.

Finally, for the implication (iv) ⇒ (i) fix a bounded continuous function $|f| \leq \mathrm{C}$. Then for any $\epsilon > 0$ there exists finitely many $f_0 < f_1 < \cdots < f_k$ with $f_0 < -\mathrm{C}, f_k > \mathrm{C}$, $f_{i+1} - f_i \leq \epsilon$ and $\mathrm{P}(\{f = f_i\}) = 0$ for all i (since the set of c's such that $\mathrm{P}(\{f = c\}) > 0$ is at most countable). Then for $\mathrm{A}_i := \{f_{i-1} < f \leq f_i\}$ we have

$\mathrm{P}(\partial \mathrm{A}_i) = 0$ and by (iv) it holds that $\lim_n \mathrm{P}_n(\mathrm{A}_i) = \mathrm{P}(\mathrm{A}_i)$. Thus, by construction it follows that

$$\begin{aligned}\left|\int f(\mathrm{dP} - \mathrm{dP}_n)\right| &\leq 2\epsilon + \left|\int \sum_i c_i \mathbf{1}_{\mathrm{A}_i}(\mathrm{dP} - \mathrm{dP}_n)\right| \\ &= 2\epsilon + \left|\sum_i c_i(\mathrm{P}(\mathrm{A}_i) - \mathrm{P}_n(\mathrm{A}_i))\right| \leq 3\epsilon\end{aligned} \tag{15}$$

for large enough n and therefore (i) follows since ϵ was arbitrary. □

In the next few sections, we will be focusing only on the cases of probability measures on $\mathbb{R}$. We will come back to the case of $\mathbb{R}^d$ in the final section of this chapter.

7.2 *Weak convergence and distribution functions*

Recall that a probability measure P on $\mathbb{R}$ can be characterized by its distribution $\mathrm{F}_{\mathrm{P}}(x) = \mathrm{P}((-\infty, x])$, which is right-continuous, non-decreasing and satisfies $\lim_{-\infty} \mathrm{F} = 0$ and $\lim_{+\infty} \mathrm{F} = 1$, and that conversely, any function F with these properties in the distribution function of some probability measure on $\mathbb{R}$ (Proposition 2.4). Since a point of continuity of F corresponds necessarily to a positive jump from $\mathrm{F}(x-)$ to $\mathrm{F}(x)$ (and the interval $(\mathrm{F}(x-), \mathrm{F}(x))$ contains rational numbers), the set of discontinuity points of F is at most countable.

Example 7.6. Let X be a random variable measure on $\mathbb{R}$ with distribution function F_{X}. Then the law $\mu_{\mathrm{X}+1/n}$ of the shifted random variable $\mathrm{X} + 1/n$ converges weakly to the law μ of X since

$$\int f \, \mathrm{d}\mu_{\mathrm{X}+1/n} = \mathbf{E} f(\mathrm{X} + 1/n) \to \mathbf{E} f(\mathrm{X}) = \int f \, \mathrm{d}\mu_{\mathrm{X}} \tag{16}$$

by dominated convergence. However, the distribution function $\mathrm{F}_{\mathrm{X}+1/n}(x) = \mathrm{P}(\mathrm{X} \leq x - 1/n) = \mathrm{F}_{\mathrm{X}}(x - 1/n) \to \mathrm{F}_{\mathrm{X}}(x-)$ converges to $\mathrm{F}_{\mathrm{X}}(x)$ only at the points of continuity of F_{X}.

Proposition 7.7. *Suppose that* P_n *is a sequence of probability measures on* $\mathbb{R}$*, that* P *is a probability measure on* $\mathbb{R}$*, and let* F_n *(resp.* F*) denote the distribution functions of* P_n *and* P *respectively. Then,* P_n *converges weakly to* P *if and only if for every point of continuity* x *of* F*,* $\lim_{n\to\infty} \mathrm{F}_n(x) = \mathrm{F}(x)$*.*

For the proof of Proposition 7.7 we will make use of the following *coupling result* which proves to be useful in other contexts as well.

Theorem 7.8 (Skorokhod's coupling theorem). *Let* $(\mathrm{F}_n)_{n\geq 1}$ *be a sequence of distribution functions on* $\mathbb{R}$ *which converges to some distribution function* F *at all points of continuity of* F*. Then there exist random variables* $\mathrm{X}, \mathrm{X}_1, \mathrm{X}_2, \ldots$ *on some common probability space* $(\Omega, \mathcal{A}, \mathrm{P})$ *such that* $\mathrm{F}_{\mathrm{X}_n} = \mathrm{F}_n$*,* $\mathrm{F}_{\mathrm{X}} = \mathrm{F}$ *and* $\lim_n \mathrm{X}_n = \mathrm{X}$ P*-almost surely.*

Proof. Take $\Omega = (0,1)$, let $\mathcal{A} = \mathcal{B}_{(0,1)}$ denote the Borel σ-algebra on $(0,1)$ and let P be the Lebesgue measure on $(0,1)$. As in the proof of Proposition 2.4 for $\omega \in (0,1)$ we set $X(\omega) := \sup\{y \mid F(y) < \omega\}$ and $X_n(\omega) := \sup\{y \mid F_n(y) < \omega\}$. Then X is a random variable on $(0,1)$ with distribution function F and X_n is a random variable on $(0,1)$ with distribution function F_n (recall the proof of Proposition 2.4). We claim that for all continuity points ω of X it holds that $X_n(\omega) \to X(\omega)$.

$\liminf_n X_n(\omega) \geq X(\omega)$: Let $x < X(\omega)$ be any continuity point of F. Then $F(x) < \omega$ and also $F_n(x) < \omega$ for sufficiently large n, and therefore $X_n(\omega) \geq x$. Since x was arbitrary the claim follows.

$\limsup_n X_n(\omega) \leq X(\omega)$: Let $x > X(\omega)$ be any continuity point of F. Then $F(x) > \omega$ and for sufficiently large n also $F_n(x) > \omega$, and therefore $X_n(\omega) \leq x$. Since x was arbitrary the claim follows.

Since X is monotone it follows that the exceptional set is at most countable and therefore $X_n \to X$ almost surely. □

Proof of Proposition 7.7. Suppose first that P_n converges weakly to P, and let x be a continuity point of F. Then by Theorem 7.5(iv) it follows that $F_n(x) = P_n((-\infty, x]) \to P((-\infty, x]) = F(x)$ since $P(\partial(-\infty, x]) = P(\{x\}) = F(x) - F(x-) = 0$.

Conversely, let $X, X_1, X_2, \dots$ be the random variables from Theorem 7.8 and let f be a bounded continuous function. Then $f(X_n) \to f(X)$ almost surely and therefore by dominated convergence $\int f \, dP_n = \mathbf{E} f(X_n) \to \mathbf{E} f(X) = \int f \, dP$ as $n \to \infty$. □

7.3 *Weak convergence vs. almost sure convergence of random variables*

Lecture 2
23rd November 2022

We now make some comments on the relation between convergence of random variables and convergence of their laws.

Definition 7.9. When $(X_n)_{n\geq 1}$ is a sequence of random variables such that the law of X_n converges weakly as $n \to \infty$, one sometimes says "X_n converges in distribution".

Remark 7.10. We defined convergence in distribution of random variables in Definition 7.9 as weak convergence of the laws of the random variables, i.e. $X_n \to X$ in distribution if for all bounded continuous f it holds that

$$\lim_n \int f \, dP_{X_n} = \int f \, dP_X. \tag{17}$$

Recalling the change of variables formula from Lemma 2.2 for the law this is equivalent to

$$\lim_n \mathbf{E} f(X_n) = \mathbf{E} f(X) \quad \text{for all bounded continuous } f. \tag{18}$$

On the other hand, due to Proposition 7.7 it is also equivalent to

$$\lim_n P(X_n \le x) = P(X \le x) \quad \text{for all points of continuity of } x \mapsto P(X \le x). \tag{19}$$

Exercise 7.11. If $(X_n)_{n\geq 1}$ is a sequence or random variables which converges in probability to some random variable X, then it converges also in distribution to X.

The coupling result from Theorem 7.8 is very far from a converse of Exercise 7.11. In fact, if $(X_n)_{n\geq 1}$ is a sequence of random variables on the same probability space, such that the law of X_n converges weakly to the law of another random variable X in that space, this does not imply at all that X_n converges in probability or almost surely to some random variable. In many important examples (for instance in the central limit theorem), this will *not* be the case. The simplest counterexample is for instance when $(X_n)_{n\geq 1}$ is a sequence of independent identically distributed random variables (when the law is not a Dirac mass). There is just one special case where such a converse statement holds:

Exercise 7.12. When X_n converges in law to the Dirac mass at some point a, then X_n converges in probability to a (see the final exercise of Exercise sheet 10).

However, Theorem 7.8 is a very useful tool in many situations, and it is often possible to prove that a sequence of random variables converges in law to some random variable:

Theorem 7.13 (Continuous mapping)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a measurable function and let X_n be a sequence of random variables converging in distribution to some random variable X with $P(X \in D(f)) = 0$, where $D(f)$ is the (measurable) set of discontinuities of f. Then $f(X_n)$ converges in distribution to $f(X)$, and if f is bounded, then also $\mathbf{E} f(X_n) \to \mathbf{E} f(X)$.*

Proof. Let $Y, Y_1, Y_2, \ldots$ be the random variables from Theorem 7.8, i.e. X, Y and X_n, Y_n have the same distribution for all n and $Y_n \to Y$ almost surely. Now for an arbitrary bounded continuous function g it follows that $g(f(Y_n)) = (g \circ f)(Y_n) \to g(f(Y))$ almost surely since $D(g \circ f) \subset D(f)$ and therefore by dominated convergence $\mathbf{E} g(f(Y_n)) \to \mathbf{E} g(f(Y))$ and in particular $f(Y_n) \to f(Y)$ in distribution. Finally, if f is bounded then by dominated convergence $\mathbf{E} f(X_n) = \mathbf{E} f(Y_n) \to \mathbf{E} f(Y) = \mathbf{E} f(X)$. □

7.4 *Tightness and compactness*

Let us consider a sequence $(P_n)_{n\geq 1}$ of probability measures in $\mathbb{R}$. The question that we are going to address is under which conditions there exists a subsequence $n_k \to \infty$, such that P_{n_k} converges weakly as $k \to \infty$.

Theorem 7.14 (Helly's selection theorem)**.** *For any sequence of distribution functions $(F_n)_{n\geq 1}$ on $\mathbb{R}$ there exists a subsequence $(F_{n_k})_{k\geq 1}$ and a non-decreasing* *right-continuous* $F\colon \mathbb{R} \to [0,1]$ *such that* $\lim_{k\to\infty} F_{n_k}(x) = F(x)$ *for all continuity points* x *of* F.

Proof. By a diagonal argument we can find a sequence (n_k) along which for all rationals $q \in \mathbb{Q}$ the limit $\tilde{F}(q) := \lim_k F_{n_k}(q)$ exists. Note that $\tilde{F}$ is non-decreasing as the limit of non-decreasing functions. For $x \in \mathbb{R}$ define $F(x) := \inf\{\tilde{F}(q) \mid \mathbb{Q} \ni q > x\}$ which is non-decreasing and right-continuous by construction. We note that for $x \leq p \in \mathbb{Q}$ it holds that $F(x) \leq \tilde{F}(p) \leq F(p)$ and that the last inequality can be strict, and in particular, the restriction of F to $\mathbb{Q}$ does not need to be equal to $\tilde{F}$.

Indeed, let $q_1, q_2, \ldots$ be an enumeration of the rationals, then set $n_k^0 := k$ and for $i \geq 1$ by induction let $(n_k^{(i)})_{k\geq 1}$ be a subsequence of $(n_k^{(i-1)})_{k\geq 1}$ such that $F_{\tilde{n}_k^{(i)}}(q_i)$ converges to some $\tilde{F}(q_i)$. Now the limit $\lim_k F_{n_k^{(k)}}(q) = \tilde{F}(q)$ exists for all $q \in \mathbb{Q}$.

It remains to check that $\lim_k F_{n_k}(x) = F(x)$ for all continuity points x of F. Indeed, for a given $\epsilon > 0$ and continuity point x choose rationals q_1, q_2, q_3 such that $q_1 < q_2 < x < q_3$ with

$$F(x) - \epsilon \leq F(q_1) \leq \tilde{F}(q_2) \leq F(q_2) \leq F(x) \leq \tilde{F}(q_3) \leq F(q_3) \leq F(x) + \epsilon. \qquad (20)$$

Then $\limsup_k F_{n_k}(x) \leq \lim_k F_{n_k}(q_3) = \tilde{F}(q_3) \leq F(x) + \epsilon$ and $\liminf_k F_{n_k}(x) \geq \lim_k F_{n_k}(q_2) = \tilde{F}(q_2) \geq F(x) - \epsilon$. Since ϵ was arbitrary the claim follows. □

Remark 7.15. It is not guaranteed that the subsequential limit in Theorem 7.14 is the distribution function of a probability measure! An example to have in mind is that when P_n is the Dirac mass at n, then the distribution functions F_n converge pointwise to the 0-function, which cannot the be distribution function of a probability measure.

To avoid this "loss of mass to infinity" phenomenon, one is led to the following definition:

Definition 7.16. A family of probability measures $(P_i)_{i\in I}$ on a metric space (E, d) is tight if for every $\epsilon > 0$, one can find a compact set C, such that for all $i \in I$, $P_i(C) \geq 1 - \epsilon$.

In the case a family $(P_i)_{i\in I}$ of probability measures on $\mathbb{R}$, this means that:

Definition 7.17. A family of probability measures $(P_i)_{i\in I}$ on $\mathbb{R}$ is tight if for every $\delta > 0$, one can find $K > 0$, such that for all $i \in I$, $P_i([-K, K]) \geq 1 - \delta$.

We are now ready to state the main result of this section:

Proposition 7.18. *A family of probability measures* $(P_i)_{i\in I}$ *on* $\mathbb{R}$ *is tight if and only if every for every sequence* P_{i_n} *has a weakly convergent subsequence* $P_{i_{n_k}}$.

Remark 7.19. This statement in fact holds also for sequences of probability measure in complete separable metric spaces. It is known as Prokhorov's theorem.

Proof. We first prove the "only if" part. Let P_{i_n} be any sequence of probability measures in $\mathbb{R}$ and let $(F_{i_n})_{n\geq 1}$ be the sequence of distribution functions. By Theorem 7.14 there exists a subsequence $(F_{i_{n_k}})$ and a non-decreasing right-continuous $F\colon \mathbb{R} \to [0,1]$ such that $\lim_{k\to\infty} F_{i_{n_k}}(x) = F(x)$ for all continuity points x of F. Now by tightness there exists K such that $P_{i_n}([-K,K]) \geq \epsilon$ for all n and by increasing K we may assume that ±K are continuity points of F. Thus $F(K) - F(-K) = \lim_k (F_{i_{n_k}}(K) - F_{i_{n_k}}(K)) = \lim_k P_{i_{n_k}}((-K,K]) \geq 1 - \epsilon$ and since ϵ was arbitrary and $F\colon \mathbb{R} \to [0,1]$ it follows that $F(-\infty) = 0$ and $F(\infty) = 1$. By Proposition 2.4 we can thus conclude that F is the distribution function of a unique probability measure P on $\mathbb{R}$. By Proposition 7.7 we have $P_{i_{n_k}} \to P$ weakly, as claimed.

Conversely, if the family is not tight, then for some $\epsilon > 0$ we can construct a sequence $(P_{i_n})_{n\geq 1}$ such that $P_{i_n}([-n,n]) \leq 1-\epsilon$. Suppose that $(P_{i_{n_k}})_{k\geq 1}$ is a weakly convergent subsequence of $(P_{i_n})_{n\geq 1}$ converging to some probability measure P. Choose K such that $P(\{\pm K\}) = 0$ and $P([-K,K]) > 1 - \epsilon$. For large enough k therefore

$$1 - \epsilon \geq P_{i_{n_k}}([-n_k, n_k]) \geq P_{i_{n_k}}([-K,K]) \to P([-K,K]) > 1 - \epsilon \tag{21}$$

which is a contradiction. □

Exercise 7.20. A family $(P_i)_{i\in I}$ of probability measures is tight if there exists a function $\phi\colon \mathbb{R} \to [0,\infty)$ with $\phi(x) \to \infty$ as $|x| \to \infty$ such that $\sup_i \int \phi \, dP_i < \infty$.

7.5 *Characteristic functions*

7.5.1 Definition

A very useful tool to study convergence of probability measures on $\mathbb{R}$ are their characteristic functions:

Definition 7.21. Suppose that P is a probability measure on $\mathbb{R}$. Its characteristic function φ_P is the function from $\mathbb{R}$ into $\mathbb{C}$ defined by

$$\varphi_P(\theta) = \int e^{i\theta x} dP(x).$$

If P is the law of a random variable X, then $\varphi(\theta) = E[e^{i\theta X}]$. We also call this the *characteristic function of* X and then write $\varphi_X = \varphi_P$.

Remark 7.22. One can recognize that (possibly up to normalization by a constant), this is nothing else than the Fourier transform of the measure P.

Here are some obvious properties of φ_P:

- $\varphi(0) = 1$ and $|\varphi(\cdot)| \leq 1$.
- φ is continuous (just use dominated convergence).
- If the law of X and −X are the same, then φ_X is real-valued.
- For all real constant λ, one has $\varphi_{\lambda X}(\theta) = \varphi_X(\lambda\theta)$.

Let us list some examples of characteristic functions that have a nice expression:

Example 7.23 (Gaussian). An important example in what follows is the case of the standard Gaussian distribution, with density $(2\pi)^{-1/2}\exp(-x^2/2)$ on $\mathbb{R}$. What is so special about this distribution is that its characteristic function is $\exp(-\theta^2/2)$. We will comment on how to compute this below.

A bit more generally, when the law of a random variable X is the standard Gaussian distribution, then for all $\sigma > 0$, we call the law of σX a centered Gaussian distribution with variance σ^2, and we denote it by $\mathcal{N}(0, \sigma^2)$. Its characteristic function is then clearly $\exp(-\theta^2\sigma^2/2)$.

Example 7.24 (Poisson). The Poisson distribution with parameter $\lambda > 0$, defined on $\mathbb{N}$ by $P(\{n\}) = e^{-\lambda}\lambda^n/n!$. Its characteristic function is then $\exp(\lambda(e^{i\theta} - 1))$.

Example 7.25 (Cauchy). The Cauchy distribution with density $dx/(\pi(1 + x^2))$ on $\mathbb{R}$. Its characteristic function turns out to be

$$\int_{\mathbb{R}} \frac{e^{i\theta x}}{\pi(1 + x^2)} = \exp(-|\theta|) \tag{22}$$

Computation for Example 7.24. The computation of the characteristic function of the Poisson distribution

$$\sum_{n\geq 0} e^{-\lambda}\frac{\lambda^n}{n!}e^{i\theta n} = e^{-\lambda}\sum_{n\geq 0}\frac{(\lambda e^{i\theta})^n}{n!} = e^{-\lambda}e^{\lambda\exp(i\theta)}. \tag{23}$$

follows directly from the exponential series. □

Computation for Example 7.23. For the standard Gaussian distribution, one first notices that

$$\mathbf{E}[e^{i\theta X}] = (2\pi)^{-1/2}\int_{\mathbb{R}} e^{i\theta x - x^2/2}\,dx = (2\pi)^{-1/2}e^{-\theta^2/2}\int_{\mathbb{R}} e^{-(x-i\theta)^2/2}\,dx, \tag{24}$$

and then using the fact that the contour integral of $\exp(-z^2/2)\,dz$ over the boundary of the rectangle $[-R, R] \times [-\theta, 0]$ is 0 and then letting $R \to \infty$, one gets readily that

$$\int_{\mathbb{R}} e^{-(x-i\theta)^2/2}\,dx = \int_{\mathbb{R}} e^{-x^2/2}dx \tag{25}$$

which allows to conclude. □

Exercise 7.26. Verify that the characteristic function of the Cauchy distribution is given by $e^{-|\theta|}$ as claimed in (22).

7.5.2 Inversion formula

Lecture 3
29th November 2022

An important property is that:

Proposition 7.27. *If two probability measures on $\mathbb{R}$ have the same characteristic function, then they are the same measures.*

In fact, it is possible to explicitly reconstruct the distribution function F of a probability measure out of its characteristic function:

Proposition 7.28 (Inversion formula). *If* F *and* φ *are respectively the distribution function and the characteristic function of a probability measure* P*, then for all* $a < b$*,*

$$\lim_{T\to+\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-ia\theta}-e^{-ib\theta}}{i\theta}\varphi(\theta)\,d\theta = P((a,b)) + \frac{1}{2}P(\{a,b\}) = F^{\#}(b) - F^{\#}(a), \quad (26)$$

where $F^{\#}(x) = (F(x) + F(x-))/2.$ *)*

This second proposition implies indeed the first one – letting $a \to -\infty$ shows that one can recover $\tilde{F}$ from φ, and then, since the discontinuity points of F and $F^{\#}$ are the same, it is easy to recover F at all continuity points, and therefore F – and finally we conclude because F determines P.

Proof of Proposition 7.28. Using Fubini's theorem, $e^{ix} = \cos x + i\sin x$ and the fact that sin, cos are odd and even functions, respectively, we have

$$\begin{aligned}\frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-ia\theta}-e^{-ib\theta}}{i\theta}\varphi(\theta)\,d\theta &= \int_{\mathbb{R}}\int_{-T}^{T} \frac{e^{i(x-a)\theta}-e^{i(x-b)\theta}}{2i\pi\theta}\,d\theta\,dP(x)\\ &= \int_{\mathbb{R}}\Big[\operatorname{sgn}(x-a)S(T|x-a|) - \operatorname{sgn}(x-b)S(T|x-b|)\Big]\,dP(x),\end{aligned}$$

where

$$S(T) := \frac{1}{\pi}\int_0^T \frac{\sin x}{x}\,dx. \quad (27)$$

Here the use of Fubini's theorem is justified since $|e^{-i\theta a} - e^{-i\theta b}| \le \theta(b-a)$. Using the fact[1] that $\lim_{T\to\infty} S(T) = 1/2$ it follows that

$$\lim_{T\to\infty}\Big(\operatorname{sgn}(x-a)S(T|x-a|) - \operatorname{sgn}(x-b)S(T|x-b|)\Big) = \begin{cases}1 & \text{if } x\in(a,b),\\ 0 & \text{if } x\in[a,b]^c,\\ 1/2 & \text{if } x\in\{a,b\}.\end{cases} \quad (28)$$

and the result follows from the dominated convergence theorem. □

7.5.3 Regularity properties

If the characteristic function is integrable, then the inversion formula from Proposition 7.28 is simpler. However, as the next proposition shows, this can only

[1] There are multiple elementary ways to evaluate this so called *Dirichlet integral*. One uses double integration in the form

$$\int_0^r \frac{\sin t}{t}\,dt = \int_0^r\int_0^\infty \frac{\sin t}{e^{st}}\,ds\,dt = \int_0^\infty\int_0^r \frac{\sin t}{e^{st}}\,dt\,ds = \int_0^\infty\left[\frac{1}{1+s^2} - \frac{s\sin r + \cos r}{(1+s^2)e^{rs}}\right]ds = \frac{\pi}{2} + O(r^{-1}),$$

where the exchange of integrals is justified since

$$\int_0^r\int_0^\infty\left|\frac{\sin t}{e^{st}}\right|ds\,dt = \int_0^r\left|\frac{\sin t}{t}\right|dt \le r < \infty.$$

be the case when the measure has a bounded continuous density:

Proposition 7.29. *If* $\int |\varphi(\theta)| d\theta < \infty$*, then the law* P *has a bounded continuous density* f *with respect to the Lebesgue measure on* $\mathbb{R}$ *given by*

$$f(x) = \frac{1}{2\pi} \int e^{-\mathrm{i}\theta x} \varphi(\theta) \, \mathrm{d}\theta \tag{29}$$

Proof. By Proposition 7.28, we have

$$\mathrm{P}((a,b)) + \frac{1}{2}\mathrm{P}(\{a,b\}) \leq \frac{|b-a|}{2\pi} \int_{\mathbb{R}} |\varphi(\theta)| \, \mathrm{d}\theta \tag{30}$$

and therefore it follows that P has no point masses and thus that $\mathrm{F} = \mathrm{F}^{\#}$. We can now write

$$\frac{\mathrm{F}(x+\epsilon) - \mathrm{F}(x)}{\epsilon} = \frac{1}{2\pi} \int \frac{e^{-\mathrm{i}\theta x} - e^{-i\theta(x+\epsilon)}}{\mathrm{i}\epsilon\theta} \varphi(\theta) \, \mathrm{d}\theta \to \frac{1}{2\pi} \int e^{-\mathrm{i}\theta x} \varphi(\theta) \, \mathrm{d}\theta \tag{31}$$

as $\epsilon \downarrow 0$. This concludes the proof since f is continuous by dominated convergence and bounded by definition. □

In Proposition 7.29 we saw that integrable characteristic functions imply that the law has a bounded continuous density. Conversely, the characteristic functions of distributions with density decay at infinity (see Exercise 7.30 below) so that we have established the principle

$$\text{Decay of } \varphi \text{ at infinity} \Leftrightarrow \text{regularity of P.} \tag{32}$$

Exercise 7.30 (Riemann-Lebesgue lemma). If P is a probability measure on $\mathbb{R}$ with measurable density f, then the characteristic function φ_{P} of P satisfies

$$\lim_{\theta \to \pm\infty} \varphi_{\mathrm{P}}(\theta) = 0. \tag{33}$$

Next, we see that smoothness of φ at 0 in turn is related to the decay of P at infinity, or roughly

$$\text{Smoothness of } \varphi \text{ at } 0 \Leftrightarrow \text{decay of P at } \infty. \tag{34}$$

For one implication we, for instance, have the following result:

Lemma 7.31. *Let* P *be a probability distribution on* $\mathbb{R}$ *with characteristic function* φ*. Then for any* $\epsilon > 0$

$$\mathrm{P}([-2/\epsilon, 2/\epsilon]^c) \leq \frac{1}{\epsilon} \int_{-\epsilon}^{\epsilon} (1 - \varphi(t)) \, \mathrm{d}t, \tag{35}$$

where we note that the rhs. is real valued since $\overline{\varphi(t)} = \varphi(-t)$*.*

Example 7.32. The following example shows that Lemma 7.31 asymptotically can be sharp but does not have to be sharp.

- If P is *Cauchy distribution*, then as $\epsilon \to 0$
$$\mathrm{P}([-2/\epsilon, 2/\epsilon]^c) = \frac{\epsilon}{\pi} + \mathrm{O}(\epsilon^3), \tag{36}$$
while
$$\frac{1}{\epsilon}\int_{-\epsilon}^{\epsilon}(1-\varphi(t))\,\mathrm{d}t = \epsilon + \mathrm{O}(\epsilon^2). \tag{37}$$
- If P is a distribution of compact support then the lhs. of (35) eventually is identically 0 while the rhs. is strictly positive for all $\epsilon > 0$ (unless $\mathrm{P} = \delta_0$).

Proof. We have
$$\begin{aligned}\frac{1}{\epsilon}\int_{-\epsilon}^{\epsilon}(1-\varphi(t))\,\mathrm{d}t &= \int_{\mathbb{R}}\frac{1}{\epsilon}\int_{-\epsilon}^{\epsilon}(1-e^{\mathrm{i}tx})\,\mathrm{d}t\,\mathrm{d}\mathrm{P}(x)\\ &= \int_{\mathbb{R}}\Big(2-\frac{e^{\mathrm{i}\epsilon x}-e^{-\mathrm{i}\epsilon x}}{\mathrm{i}\epsilon x}\Big)\mathrm{d}\mathrm{P}(x) = 2\int_{\mathbb{R}}\Big(1-\frac{\sin(\epsilon x)}{\epsilon x}\Big)\mathrm{d}\mathrm{P}(x) \qquad (38)\\ &\geq 2\int_{|x|\geq 2/\epsilon}\Big(1-\frac{1}{|\epsilon x|}\Big)\mathrm{d}\mathrm{P}(x) \geq \mathrm{P}([-2/\epsilon, 2/\epsilon]^c)\end{aligned}$$
using that $\sin u/u \leq 1/|u|$ for $|u| \geq \pi/2 < 2$. □

On the other hand, for distributions with a n finite moments have n-times differentiable characteristic functions:

Exercise 7.33. Let P be a probability distribution on $\mathbb{R}$ with n finite moments, i.e. such that $\int |x|^n\,\mathrm{d}\mathrm{P}(x) < \infty$. Then the characteristic function φ of P is n-times differentiable and satisfies
$$\varphi^{(n)}(t) = \int (\mathrm{i}x)^n e^{\mathrm{i}tx}\,\mathrm{d}\mathrm{P}(x). \tag{39}$$

Near $t = 0$ we may expand the characteristic function as a power series:

Lemma 7.34. *For random variables* X *with n finite moments the characteristic function* $\varphi_{\mathrm{X}}(t) = \mathbf{E}\,e^{\mathrm{i}t\mathrm{X}}$ *has the expansion*
$$\Big|\varphi_{\mathrm{X}}(t) - \sum_{k=0}^{n}\frac{(\mathrm{i}t)^k}{k!}\mathbf{E}\,\mathrm{X}^k\Big| \leq \mathbf{E}\Big[\min\Big\{\frac{|t\mathrm{X}|^{n+1}}{(n+1)!}, \frac{2|t\mathrm{X}|^n}{n!}\Big\}\Big] = o(|t|^n). \tag{40}$$

Proof. The first inequality follows directly from the Taylor expansion
$$\Big|e^{\mathrm{i}x} - \sum_{k=0}^{n}\frac{(\mathrm{i}x)^k}{k!}\Big| \leq \min\Big\{\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!}\Big\} \tag{41}$$
and Jensen's inequality. To see that the expectation is $o(|t|^n)$ as $t \to 0$ note that $\min\{|t||\mathrm{X}|^{n+1}, 2|\mathrm{X}|^n\}$ is a random variable with finite expectation converging pointwise to 0 and apply dominated convergence.

The Taylor expansion (41) can be obtained from repeated integration by parts in the form

$$\begin{aligned} \mathrm{i}^{n+1}\int_0^x \frac{(x-s)^n}{n!}e^{\mathrm{i}s}\,\mathrm{d}s &= -\frac{(\mathrm{i}x)^n}{n!} + \mathrm{i}^n\int_0^x \frac{(x-s)^{n-1}}{(n-1)!}e^{\mathrm{i}s}\,\mathrm{d}s \\ &= \cdots = -\sum_{k=1}^{n}\frac{(\mathrm{i}x)^k}{k!} + \mathrm{i}\int_0^x e^{\mathrm{i}s}\,\mathrm{d}s = -\sum_{k=0}^{n}\frac{(\mathrm{i}x)^k}{k!} + e^{\mathrm{i}x}. \end{aligned} \tag{42}$$

and estimating the error integral by using

$$\frac{x^{n+1}}{(n+1)!} \geq \left|\mathrm{i}\int_0^x \frac{(x-s)^n}{n!}e^{\mathrm{i}s}\,\mathrm{d}s\right| = \left|\int_0^x \frac{(x-s)^{n-1}}{(n-1)!}(e^{\mathrm{i}s}-1)\,\mathrm{d}s\right| \leq \frac{2x^n}{n!}. \qquad \square$$

7.5.4 Characteristic functions and independence

Lecture 4
30th November 2022

An important and simple observation is the following:

Proposition 7.35. *If* $X_1,\ldots,X_n$ *are independent random variables defined on the same probability space, then*

$$\varphi_{X_1+\ldots+X_n}(\theta) = \prod_{j=1}^{n}\varphi_{X_j}(\theta).$$

Proof. This is just due to the fact that the random variables $e^{\mathrm{i}\theta X_1},\ldots,e^{\mathrm{i}\theta X_n}$ are independent (and bounded) random variables with values in $\mathbb{C}$ (one can write $\exp(\mathrm{i}y) = \cos(y) + \mathrm{i}\sin(y)$ and expand the product if one does not feel at ease with the complex multiplications here). □

This can be a very useful tool to actually determine the law of the sum of some independent random variables, in particular in the case of random variables with a density where the computations via density functions can be cumbersome. Basically, if one knows that characteristic functions of $X_1,\ldots,X_n$, then the previous result gives us automatically the characteristic function of $X_1+\cdots+X_n$, and if one happens to recognize that characteristic function as that of a known law, then (since characteristic functions determine the law), one knows the law of $X_1+\cdots+X_n$.

For example: *If* X *and* Y *are two independent centered Gaussian random variables with respective variances* σ_X^2 *and* σ_Y^2, *then* $X+Y$ *is a centered Gaussian random variable with variance* $\sigma_X^2+\sigma_Y^2$. Indeed,

$$\varphi_{X+Y}(\theta) = \varphi_X(\theta)\varphi_Y(\theta) = \exp(-\theta^2(\sigma_X^2+\sigma_Y^2))$$

which is the characteristic function of a centered Gaussian variable with variance $\sigma_X^2+\sigma_Y^2$ and we can conclude noting that the characteristic function characterizes the law.

Similarly, one gets that if X and Y are independent random variables with standard Cauchy distributions, then the law of $(X+Y)/2$ is also a standard

Cauchy distribution.

A similar argument can be used show that if X and Y are independent Poisson random variables with respective parameters λ_Y and λ_Y, then $X+Y$ is a Poisson random variable with parameter $\lambda_X + \lambda_Y$. But this fact could have been derived immediately looking at $P(X+Y=n) = \sum_{j=0}^{n} P(X=j)P(Y=n-j)$.

7.6 *Weak convergence via characteristic functions*

Clearly, since (for all fixed θ), $x \mapsto e^{i\theta x}$ is a bounded continuous function, when P_n converges weakly to P, then for all $\theta \in \mathbb{R}$, one has $\lim_{n\to\infty} \varphi_n(\theta) = \varphi(\theta)$. We will now discuss results in the other direction: Does the convergence of φ_n suffice to obtain weak convergence? The answer is affermative if the sequence of probability measures is tight, or if the limit of characteristic functions is continuous at 0:

Proposition 7.36 (Lévy's theorem for tight sequences). *Let* $(P_n)_{n\geq 1}$ *denote a tight sequence of probability measures on* $\mathbb{R}$*, such that for all* $\theta \in \mathbb{R}$*, the sequence* $\varphi_n(\theta)$ *converges to some number* $\psi(\theta)$*. Then,* ψ *is the characteristic function of a probability measure* P*, and* P_n *does converge weakly to* P.

Corollary 7.37 (Lévy's theorem for sequences with continuous limit). *A sequence* $(P_n)_{n\geq 1}$ *of probability measures whose characteristic functions* ϕ_n *converge pointwise to some function* ψ *continuous at* 0*, is tight. In particular the conclusion of Proposition 7.36 holds true if we assume continuity of* ψ *at* 0 *instead of tightness.*

Example 7.38. Without tightness or continuity of the limit of characteristic functions in 0, the following example shows that the convergence of characteristic functions does not imply weak convergence. For $N_n \sim \mathcal{N}(0,n)$, i.e. a normally distributed random variable of mean 0 and variance n, we have

$$\varphi_{N_n}(\theta) = \exp\Big(-\frac{n\theta^2}{2}\Big) \to \begin{cases} 1, & \text{if } \theta = 0, \\ 0, & \text{otherwise} \end{cases} \tag{43}$$

which is not continuous in 0 and hence cannot be a characteristic function. The sequence N_n does not converge in distribution to any random variable since for each x,

$$P(N_n \leq x) = P(N_1 \leq x/\sqrt{n}) \to \frac{1}{2}. \tag{44}$$

Proof of Proposition 7.36. By Proposition 7.18 we deduce that any subsequence P_{n_k} contains a further subsequence $P_{n_{k_j}}$ converging weakly to some probability measure P. Thus the characteristic functions of $P_{n_{k_j}}$ converge pointwise to the characteristic function of P, and we conclude $\phi_P = \psi$ which in particular is the characteristic function of a probability measure P. This shows that every subsequence of P_n contains a further subsequence converging weakly to the same probability measure P by Proposition 7.27.

Suppose that P_n does not converge weakly to P. Then there exists a bounded continuous function f, some $\epsilon > 0$ and a subsequence P_{n_k} such that $|\int f\,dP -$

$\int f \, \mathrm{d}P_{n_k}| > \epsilon$ for all k. In particular, no further subsequence can converge weakly to P and we have a contradiction. □

Proof of Corollary 7.37. Since $\psi(0) = 1$ and ψ is continuous in 0 we can choose $\delta > 0$ such that

$$\frac{1}{\delta}\int_{-\delta}^{\delta}(1-\psi(t))\,\mathrm{d}t < \epsilon/2. \tag{45}$$

By dominated convergence it follows that

$$\frac{1}{\delta}\int_{-\delta}^{\delta}(1-\varphi_n(t))\,\mathrm{d}t < \epsilon \tag{46}$$

and therefore by Lemma 7.31 that $P_n([-2/\delta, 2/\delta]^c) < \epsilon$ for all $n \geq n_0$ for some finite n_0. By choosing $K \geq 2/\delta$ such that $P_n([-K,K]^c) < \epsilon$ for $n = 1,\dots,n_0 - 1$ tightness follows. □

7.7 *Central limit theorem (CLT)*

Suppose now that $(X_n)_{n\geq 1}$ is a sequence of independent identically distributed random variables that are in L^2 with mean $\mu = \mathbf{E}\,X_1$ and variance $\sigma^2 = \operatorname{Var} X_1$. We assume that $\sigma^2 \neq 0$, i.e. that $P(X_1 = \mu) \neq 1$. We write $S_n := X_1 + \cdots + X_n$.

We know from the law of large numbers that S_n/n tends to μ almost surely when $n \to \infty$. It is a natural question to ask what the actual order of magnitude of $S_n - \mu n$ is when n is large. A first indication that it will be of the order of $\sqrt{n}$ is that

$$E[(S_n - \mu n)^2] = E[(X_1-\mu)^2] + \cdots + E[(X_n - \mu)^2] = n\sigma^2. \tag{47}$$

Theorem 7.39 (Central limit theorem)**.** *Suppose that* $(X_n)_{n\geq 1}$ *is a sequence of independent identically distributed random variables with*

$$\mu := \mathbf{E}\,X_1, \quad \sigma^2 := \operatorname{Var} X_1 = \mathbf{E}(X_1-\mu)^2 = \mathbf{E}\,X_1^2 - \mu^2 < \infty. \tag{48}$$

Then

$$Y_n := \frac{S_n - \mu n}{\sqrt{n}} = \frac{(X_1-\mu)+\cdots+(X_n-\mu)}{\sqrt{n}} \to \mathcal{N}(0,\sigma^2) \tag{49}$$

in distribution as $n \to \infty$.

Proof. Since for a Gaussian random variable $N \sim \mathcal{N}(0,\sigma^2)$ with variance σ^2, the rescaled random variable $N/\sigma \sim \mathcal{N}(0,1)$ is a standard Gaussian, we may divide by σ and without loss of generality assume that $\mu = 0$ and $\sigma = 1$. If φ is the characteristic function of $X_1, X_2, \dots$, then by independence and Lemma 7.34, $Y_n := S_n/\sqrt{n}$ has characteristic function

$$\phi_n(\theta) = \phi\Big(\frac{\theta}{\sqrt{n}}\Big)^n = \Big(1 - \frac{\theta^2}{2n} + o(\theta^2/n)\Big)^n, \quad \text{as } \frac{\theta}{\sqrt{n}} \to 0. \tag{50}$$

Since $|z^n - w^n| \leq n|z-w|(\max\{|z|,|w|\})^{n-1}$ it follows that

$$\phi_n(\theta) = \left(1 - \frac{\theta^2}{2n}\right)^n + o(1) = e^{-\theta^2/2} + o(1), \quad \text{as } n \to \infty. \tag{51}$$

Thus, ϕ_n converges pointwise to the characteristic function of a standard Gaussian distribution computed in Example 7.23. By Corollary 7.37 we deduce that the sequence Y_n of random variables with characteristic function ϕ_n converges in distribution to a standard Gaussian. □

Remark 7.40. We can note that the condition that X is in L^2 actually implies readily that for any $\epsilon > 0$,

$$P(\max(|X_1|,\ldots,|X_n|) \geq \epsilon\sqrt{n}) \leq nP(X_1^2 \geq \epsilon^2 n) \leq \epsilon^{-2}E[X_1^2 1_{X_1^2 \geq \epsilon^2 n}] \to 0$$

as $n \to \infty$ (by dominated convergence). So, one can heuristically say that (with high probability when n is large), none of the individual terms X_j for $j \leq n$ will be of the same order of magnitude than $X_1 + \cdots + X_n$ – which is $\sqrt{n}$. This will contrast the results in Section 7.10.

7.8 *The moment problem*

In the previous sections we saw that convergence of $\mathbf{E}\, e^{itX_n} \to \psi(t)$ implies convergence of X_n in distribution to a random variable with characteristic function ψ, if (i) the sequence of probability measures is tight, or (ii), ψ is continuous in 0. We now suppose that $(X_n)_{n\geq 1}$ is a sequence of random variables with finite moments of all orders, i.e. $\mathbf{E}|X_n|^k \leq C_k$ for all n,k.

If $X_n \to X$ in distribution for some random variable X, then all moments converge,

$$\mathbf{E}\, X_n^k \to \mathbf{E}\, X^k, \quad k \geq 0. \tag{52}$$

Indeed, let Y_n, Y be the random variables from Theorem 7.8 with the same distribution, $Y_n \overset{d}{=} X_n, Y \overset{d}{=} X$ converging $Y_n \to Y$ almost surely. The sequence $(Y_n^k)_{n\geq 1}$ is uniformly integrable since $P(|Y_n|^k \geq C) \leq \mathbf{E}|Y_n|^{2k}/C^{2K} \leq C_{2k}/C^{2K}$ and therefore converges in L^1 to Y^k. In particular $\mathbf{E}\, X_n^k \mathbf{E}\, Y_n^k \to \mathbf{E}\, Y^k = \mathbf{E}\, X^k$ and (52) follows. The following theorem shows that the converse is true whenever the limiting moments determine a *unique distribution*.

Note that without uniqueness assumption such a statement cannot be true. For instance if there exist random variables X, Y *with the same moments but different distribuitons (see Example 7.44 below) and we have* $X_1 \overset{d}{=} X, X_2 \overset{d}{=} Y, X_3 \overset{d}{=} X$ *etc., then* X_n *cannot converge in distribution while* $\mathbf{E}X_n^k = \mathbf{E}Y^k = \mathbf{E}X^k$ *for all n.*

Theorem 7.41 (Second limit theorem; moment continuity theorem)**.** *Suppose that* $(X_n)_{n\geq 1}$ *is a sequence of random variables with finite moments of all orders, i.e.* $\mathbf{E}|X_n|^k \leq C_k$ *for all n, k, and that for each k*

$$\lim_n \mathbf{E}\, X_n^k = \mu_k \tag{53}$$

for some sequence $\mu_1, \mu_2, \ldots$ *of real numbers. If there exists a unique probability measure with moments* $\mu_1, \mu_2, \ldots$*, i.e.* $\int x^k \, dP(x) = \mu_k$*, then the sequence* X_n *converges in distribution to this* P.

Lecture 5
6^{th} December 2022

Proof. The sequence $(\mathrm{P}_{\mathrm{X}_n})_{n\geq 1}$ is tight since $\mathrm{P}(|\mathrm{X}_n| \geq \mathrm{C}) \leq \mathrm{C}^{-2}\,\mathbf{E}\,\mathrm{X}_n^2 \leq \mathrm{C}_2/\mathrm{C}$. Thus due to Proposition 7.18 every subsequence $(n_k)_{k\geq 1}$ has a further subsequence $(n_{k_j})_{j\geq 1}$ along which $\mathrm{P}_{n_{k_j}}$ converges weakly to some P which necessarily has moments μ_k. By the uniqueness assumption this implies that every subsequential limit converges to the same distribution P. If $\mathrm{P}_{\mathrm{X}_n}$ itself would not converge to P_X, then we could find a bounded continuous function f, som $\epsilon > 0$ and a subsequence $(n_k)_{k\geq 1}$ such that $|\int f\,\mathrm{d}(\mathrm{P}_\mathrm{X} - \mathrm{P}_{\mathrm{X}_{n_k}})| \geq \epsilon$ for all k. This subsequence cannot have a further subsequence converging to P_X, which is a contradiction. □

Theorem 7.41 is a useful result when the limiting moments are the moments of a unique probability distribution. Luckily many distributions of interest are uniquely determined by their moments. For instance any distribution with compact support is uniquely determined since on intervals $[-\mathrm{C},\mathrm{C}]$ any continuous bounded function can be uniformly approximated by polynomials. More generally we have the following sufficient condition showing that if the moments do not grow too rapidly, then at most one probability distribution can have these moments.

Theorem 7.42. *Let $\mu_1, \mu_2, \ldots$ be a sequence of real numbers such that*

$$r := \limsup_k \frac{|\mu_{2k}|^{1/2k}}{2k} < \infty. \tag{54}$$

Then at most one probability measure P *exists with moments $\mu_k = \int x^k\,\mathrm{dP}(x)$ for $k \geq 0$.*

Example 7.43. The normal distribution is uniquely determined by its moments. Indeed, by induction[2] we have $\mu_k = (k-1)!! = (k-1)(k-3)\cdots 1$ for even k and $\mu_k = 0$ by symmetry for odd k. Thus

$$\frac{(k-1)!!^{1/k}}{k} \leq \frac{1}{\sqrt{k}} \to 0 \quad \text{as } k \to \infty \tag{56}$$

and the assumption of Theorem 7.42 is satisfied.

Example 7.44. The *log-normal* distribution with density

$$f(x) := \frac{1}{\sqrt{2\pi}x}\exp\Big(-\frac{(\log x)^2}{2}\Big), \quad x > 0 \tag{57}$$

is an example of a distribution not uniquely determined by its moments. The

[2] For even k by integration by parts

$$\int x^k e^{-x^2/2}\,\mathrm{d}x = -\int x^{k-1}\partial_x(e^{-x^2/2})\,\mathrm{d}x = (k-1)\int x^{k-2}e^{-x^2/2}\,\mathrm{d}x = \cdots = (k-1)!!. \tag{55}$$

moments are given by

$$\begin{aligned}\mu_k &:= \int_0^\infty x^k f(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_0^\infty x^{k-1}\exp\Big(-\frac{(\log x)^2}{2}\Big)\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{ky-y^2/2}\,\mathrm{d}y = e^{k^2/2}\end{aligned} \tag{58}$$

which clearly violate (54). Since

$$\int_0^\infty x^k f(x)\sin(2\pi\log x)\,\mathrm{d}x = \frac{e^{k^2/2}}{\sqrt{2\pi}}\int_{\mathbb{R}}\sin(2\pi y)e^{-(y-k)^2/2}\,\mathrm{d}y = 0 \tag{59}$$

for any k by $\sin(2\pi y) = \sin(2\pi(y-k))$ and symmetry, it follows that $g_a(x) := f(x)[1 + a\sin(2\pi\log x)]$ for each $-1 \le a \le 1$ is the density of a probability distribution with the same moments μ_k as the log-normal distribution.

Proof of Theorem 7.42. Let X be a random variable with moments μ_k (should it exist) and define the absolute moments by $\nu_k := \mathbf{E}|X|^k$. By Cauchy-Schwarz it follows that

$$\nu_{2k+1} = \mathbf{E}|X|^k|X|^{k+1} \le \sqrt{\mathbf{E}|X|^{2k}}\sqrt{\mathbf{E}|X|^{2k+2}} = \sqrt{\mu_{2k}\mu_{2k+2}} \tag{60}$$

and therefore

$$\limsup_k \frac{\nu_k^{1/k}}{k} \le r < \infty \tag{61}$$

by assumption.

Due to Exercise 7.33 in the first, Jensen's inequality in the second, and (41) in the third step it follows that

$$\begin{aligned}\left|\phi_X(t+\epsilon) - \sum_{k=0}^{n-1}\frac{\epsilon^k}{k!}\phi_X^{(k)}(t)\right| &= \left|\mathbf{E}\Big[e^{\mathrm{i}(t+\epsilon)X} - \sum_{k=0}^{n-1}\frac{\epsilon^k}{k!}(\mathrm{i}X)^k e^{\mathrm{i}tX}\Big]\right| \\ &\le \mathbf{E}\left|e^{\mathrm{i}tX}\right|\left|e^{\mathrm{i}\epsilon X} - \sum_{k=0}^{n-1}\frac{\epsilon^k}{k!}(\mathrm{i}X)^k\right| \le \frac{|\epsilon|^n\nu_n}{n!} \le \frac{|e\epsilon|^n\nu_n}{n^n}.\end{aligned} \tag{62}$$

For any $|\epsilon| < 1/(er)$ the rhs. of (62) converges to 0, and therefore

$$\phi_X(t+\epsilon) = \sum_{k=0}^{\infty}\frac{\epsilon^k}{k!}\phi_X^{(k)}(t), \qquad |\epsilon| < \frac{1}{er}. \tag{63}$$

Now suppose ϕ_Y is the characteristic function of another random variable Y with the same moments. Since $\phi_X^{(k)}(0) = \phi_Y^{(k)}(0) = \mathrm{i}^k\mu_k$ we conclude $\phi_X(\epsilon) = \phi_Y(\epsilon)$ for all $|\epsilon| < 1/er$ from (63). Using (63) once more for $t = \pm 1/2er$, this time with $\phi_X^{(k)}(t) = \phi_Y^{(k)}(t)$, it follows that $\phi_X(t) = \phi_Y(t)$ for all t with $|t| < 3/2er$. This procedure can be iterated and we conclude $\phi_X = \phi_Y$ on all of $\mathbb{R}$ and the result follows from Proposition 7.27. □

To illustrate the usage of the results above we give a second proof of the CLT, this time in a more general setup for non-identically distributed random variables under the so called *Lindeberg condition*.

Theorem 7.45 (Lindeberg CLT). *Let $(X_{nk})_{1\le k\le n}$ be a triangular array of random variables such that*

(i) $\mathbf{E}\,X_{nk} = 0$ *for each n, k,*

(ii) $X_{n1}, \ldots, X_{nn}$ *are independent and* $\mathbf{E}\,X_{n1}^2 + \cdots + \mathbf{E}\,X_{nn}^2 = 1$ *for each n,*

(iii) for each $\epsilon > 0$

$$\sum_k \mathbf{E}|X_{nk}|^2 \mathbf{1}(|X_{nk}| \ge \epsilon) \to 0 \text{ as } n \to \infty. \tag{64}$$

Then

$$S_n := X_{n1} + \cdots + X_{nn} \to \mathcal{N}(0,1) \tag{65}$$

in distribution as $n \to \infty$.

The moment method is only suitable for random variables with finite moments of all orders. Therefore, the proof of Theorem 7.45 via the moment method requires a truncation step. Instead of (iii) we assume

(iii') There exists a sequence $\epsilon_n \to 0$ such that for each n, k almost surely $|X_{nk}| \le \epsilon_n$.

Proof of Theorem 7.45 under assumption (iii'). We need to check that for each k,

$$\mathbf{E}\,S_n^k \to \begin{cases} (k-1)!!, & k \text{ even} \\ 0, & k \text{ odd} \end{cases} \tag{66}$$

which are the moments of $\mathcal{N}(0,1)$ by Example 7.43. This concludes the proof by Theorem 7.41 and Theorem 7.42. By taking the k-th power of S_n and grouping the terms according to powers of distinct X_{nk}'s we obtain

$$\mathbf{E}\,S_n^k = \sum_{j=1}^{k} \frac{1}{j!} \sum_{m_1+\cdots+m_j=k} \binom{k}{m_1,\ldots,m_j} \sum_{k_1\neq\ldots\neq k_j=1}^{n} (\mathbf{E}\,X_{jk_1}^{m_1})\cdots(\mathbf{E}\,X_{nk_j}^{m_j}) \tag{67}$$

by independence. Here the multinomial coeffient counts the number of ways of grouping k elements into j groups of sizes $m_1, \ldots, m_j$ and the $j!$ compensates for the fact that the unique indices $n_1, \ldots, n_j$ may be assigned to the j groups in $j!$ ways. By mean zero assumption only terms with $\min_i m_i \ge 2$ contribute, and for $m_i > 3$ we use the bound $\mathbf{E}|X_{nk_i}^{m_i}| \le \epsilon_n^{m_i-2}\,\mathbf{E}\,X_{nk_i}^2$, so that due to (ii) terms in (67) with $\max_i m_i \ge 3$ go to zero with ϵ_n as $n \to \infty$. In particular $S_n^k \to 0$ as

$n \to \infty$ for odd k. For even k we obtain

$$\begin{aligned}\mathbf{E}\,S_n^k &= \frac{k!}{(k/2)!2^{k/2}} \sum_{k_1 \neq \ldots \neq k_{k/2}=1}^{n} (\mathbf{E}\,X_{nk_1}^2)\cdots(\mathbf{E}\,X_{nk_{k/2}}^2) + O(\epsilon_n) \\ &= (k-1)!!\Big(\sum_{k=1}^{n} \mathbf{E}\,X_{nk}^2\Big)^{k/2} + O(\epsilon_n) = (k-1)!! + O(\epsilon_n),\end{aligned} \tag{68}$$

where in the second step we dropped the constraint that the $k_1,\ldots,k_{k/2}$ are pairwise disjoint (and estimated those terms with coinciding indices by $\mathbf{E}\,X_{nk_i}^2 \le \epsilon_n^2$. □

Exercise 7.46. The goal of this exercise is to prove Theorem 7.45 from the bounded version with (iii) replaced by (iii').

(i) By a diagonal argument show that (iii) holds for ϵ replaced by ϵ_n for some sequence $\epsilon_n \to 0$.

(ii) Define $\tilde{X}_{nk}^{\le\epsilon_n} := X_{nk}\mathbf{1}(|X_{nk}| \le \epsilon_n)$, $X_{nk}^{\le\epsilon_n} := \tilde{X}_{nk}^{\le\epsilon_n} - \mathbf{E}\,\tilde{X}_{nk}^{\le\epsilon_n}$ and $S_n^{\le\epsilon_n} := X_{n1}^{\le\epsilon_n} + \cdots + X_{nn}^{\le\epsilon_n}$ with standard deviation $\sigma_n^{\le\epsilon_n} := \sqrt{\operatorname{Var} S_n^{\le\epsilon}}$. Show that assumption (iii') is satisfied and therefore $S_n^{\le\epsilon_n}/\sigma_n^{\le\epsilon_n} \to \mathcal{N}(0,1)$ in distribution.

(iii) Show the complementary sum $S_n^{>\epsilon_n} := X_{n1}^{>\epsilon_n} + \cdots + X_{nn}^{>\epsilon_n}$ for $X_{nk}^{>\epsilon_n} := X_{nk} - X_{nk}^{\le\epsilon_n}$ converges to 0 in L^2 and hence in probability.

(iv) Prove that $\sigma_n^{\le\epsilon_n} \to 1$ as $n \to \infty$.

(v) Use Exercise 11.2 to conclude that $S_n \to \mathcal{N}(0,1)$.

7.9 *Stein's method*

Lecture 6
7th December 2022

The basic idea behind Stein's method is the following characterization of the normal distribution.

Lemma 7.47. *It holds that*

$$\mathbf{E}\,f'(N) = \mathbf{E}\,N f(N) \tag{69}$$

for any differentiable function f with $|\mathbf{E}\,f'(N)| < \infty$ and $|\mathbf{E}\,Nf(N)| < \infty$ if and only if $N \sim \mathcal{N}(0,1)$.

Proof. The fact that (69) holds true for the normal distribution follows from integration by parts in the form

$$\begin{aligned}\mathbf{E}\,Nf(N) &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} x f(x) e^{-x^2/2}\,\mathrm{d}x = -\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} f(x)\partial_x(e^{-x^2/2})\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} f'(x)e^{-x^2/2}\,\mathrm{d}x = \mathbf{E}\,f'(N).\end{aligned} \tag{70}$$

On the contrary if (69) holds true for any f then in particular it holds true for moments $f(x) = x^k$ for all $k \in \mathbb{N}$, which implies

$$\mathbf{E}\,\mathrm{N}^{k+1} = k\,\mathbf{E}\,\mathrm{N}^{k-1} = \cdots = \begin{cases} k!!, & k \text{ odd} \\ 0, & k \text{ even,} \end{cases} \tag{71}$$

and therefore $\mathrm{N} \sim \mathcal{N}(0,1)$ by Example 7.43. □

For measurable functions h we are interested in solutions f_h to the differential equation

$$f_h'(x) - x f_h(x) = h(x) - \mathbf{E}\,h(\mathrm{N}), \quad \mathrm{N} \sim \mathcal{N}(0,1). \tag{72}$$

Then, for random variables X it follows that

$$\mathbf{E}\,h(\mathrm{X}) - \mathbf{E}\,h(\mathrm{N}) = \mathbf{E}[f_h'(\mathrm{X}) - \mathrm{X} f_h(\mathrm{X})] \tag{73}$$

and therefore $\mathrm{X}_n \to \mathrm{N}$ in distribution if and only if $\mathbf{E}[f_h'(\mathrm{X}_n) - \mathrm{X}_n f_h(\mathrm{X}_n)] \to 0$ for all bounded continuous h. Note that in order to prove that X_n approaches a normal distribution it is sufficient to compare two expectations only involving X_n rather than comparing $\mathbf{E}\,h(\mathrm{X}_n)$ and $\mathbf{E}\,h(\mathrm{N})$.

Lemma 7.48 (Properties of Stein equations)**.**

The equality of the two expressions follows from the fact that the itnegral over all of $\mathbb{R}$ is zero.

(i) *For any measurable h the function*

$$f_h(x) := e^{x^2/2} \int_{-\infty}^{x} \Big(h(y) - \mathbf{E}\,h(\mathrm{N})\Big) e^{-y^2/2}\,\mathrm{d}y = -e^{x^2/2} \int_{x}^{\infty} \Big(h(y) - \mathbf{E}\,h(\mathrm{N})\Big) e^{-y^2/2}\,\mathrm{d}y$$

is a bounded differentiable solution to (72) *satisfying*

$$\|f_h\|_\infty \le \sqrt{\frac{\pi}{2}} \|h - \mathbf{E}\,h(\mathrm{N})\|_\infty, \qquad \|f_h'\|_\infty \le 2\|h - \mathbf{E}\,h(\mathrm{N})\|_\infty \tag{74}$$

The equality of the two expressions follows from (69)*.*

(ii) *For any h with bounded derivative the function*

$$f_h(x) := -\int_0^1 \mathbf{E}\,\frac{h'(\sqrt{t}x + \sqrt{1-t}\mathrm{N})}{2\sqrt{t}}\,\mathrm{d}t = -\int_0^1 \mathbf{E}\,\frac{\mathrm{N}h(\sqrt{t}x + \sqrt{1-t}\mathrm{N})}{2\sqrt{t(1-t)}}\,\mathrm{d}t \tag{75}$$

is a solution to (72) *satisfying*

$$\|f_h\|_\infty \le \|h'\|_\infty, \quad \|f_h'\|_\infty \le \|h'\|_\infty. \tag{76}$$

Proof.

(i) The fact that f_h solves (72) is a direct consequence from

$$\frac{\mathrm{d}}{\mathrm{d}x}\Big[e^{-x^2/2} f_h(x)\Big] = \Big[h(x) - \mathbf{E}\,h(\mathrm{N})\Big] e^{-x^2/2}. \tag{77}$$

For the first inequality we have

$$\frac{|f_h(x)|}{\|h - \mathbf{E}\,h(\mathrm{N})\|_\infty} \le e^{x^2/2}\int_{|x|}^{\infty} e^{-y^2/2}\,\mathrm{d}y \le \int_0^\infty e^{-y^2/2}\,\mathrm{d}y = \sqrt{\frac{\pi}{2}} \tag{78}$$

because the function $|x| \mapsto e^{x^2/2}\int_{|x|}^\infty e^{-y^2/2}\,\mathrm{d}y$ has negative derivative due to

$$|x|e^{x^2/2}\int_{|x|}^{\infty} e^{-y^2/2}\,\mathrm{d}y \le e^{x^2/2}\int_{|x|}^{\infty} ye^{-y^2/2}\,\mathrm{d}y = 1. \tag{79}$$

For the second inequality we use $f_h'(x) = xf_h(x) + h(x) - \mathbf{E}\,h(\mathrm{N})$ to estimate

$$|f_h'(x)| \le \|h - \mathbf{E}\,h(\mathrm{N})\|_\infty\left(1 + |x|e^{x^2/2}\int_{|x|}^{\infty} e^{-y^2/2}\,\mathrm{d}y\right) \le 2\|h - \mathbf{E}\,h(\mathrm{N})\|_\infty \tag{80}$$

again using (79).

(ii) By differentiating we obtain

$$f_h'(x) - xf(h) = \int_0^1 \mathbf{E}\left(\frac{x}{2\sqrt{t}} - \frac{\mathrm{N}}{2\sqrt{1-t}}\right)h'(\sqrt{t}x + \sqrt{1-t}\mathrm{N})\,\mathrm{d}t = h(x) - \mathbf{E}\,h(\mathrm{N})$$

where the second equality is due to the fundamental theorem of calculus. From the first equality in (75) we obtain $\|f_h\|_\infty \le \|h'\|_\infty$, and from differentiating the second equality similarly $\|f_h'\|_\infty \le \|h'\|_\infty \mathbf{E}|\mathrm{N}| = \sqrt{2/\pi}\|h'\|_\infty \le \|h'\|_\infty$. □

Theorem 7.49 (Berry-Esséen in Wasserstein distance)**.** *Let* $\mathrm{X}_1, \dots, \mathrm{X}_n$ *be independent mean zero random variables with finite third moments such that* $\mathrm{S}_n := \mathrm{X}_1 + \cdots + \mathrm{X}_n$ *has variance* $\operatorname{Var}\mathrm{S}_n = \mathbf{E}\,\mathrm{X}_1^2 + \cdots + \mathbf{E}\,\mathrm{X}_n^2 = 1$ *and let h be continuously differentiable and bounded. Then it holds that*

$$|\mathbf{E}[h(\mathrm{S}_n) - h(\mathrm{N})]| \le \frac{9}{2}\|h'\|_\infty \sum_{k=1}^{n} \mathbf{E}|\mathrm{X}_k|^3 \tag{81}$$

for $\mathrm{N} \sim \mathcal{N}(0,1)$.

Proof. Introduce the notation $\mathrm{S}_n^{(i)} := \mathrm{S}_n - \mathrm{X}_i$ and use the fundamental theorem of calculus to obtain

$$\mathbf{E}\,\mathrm{S}_n f(\mathrm{S}_n) = \sum_i \mathbf{E}\,\mathrm{X}_i[f_h(\mathrm{S}_n^{(i)} + \mathrm{X}_i) - f_h(\mathrm{S}_n^{(i)})] = \sum_i \int_0^1 \mathbf{E}\,\mathrm{X}_i^2 f_h'(\mathrm{S}_n^{(i)} + t\mathrm{X}_i)\,\mathrm{d}t,$$

where we used independence of $\mathrm{X}_i, \mathrm{S}_n^{(i)}$ and $\mathbf{E}\,\mathrm{X}_i = 0$ in the first step. Again using independence and $\mathbf{E}\,X_1^2 + \cdots + \mathbf{E}\,X_n^2 = 1$ we conclude

$$\begin{aligned} \mathbf{E}[f_h'(S_n) - S_n f_h(S_n)] &= \sum_i (\mathbf{E}\,X_i^2)\,\mathbf{E}[f_h'(S_n^{(i)} + X_i) - f_h'(S_n^{(i)})] \\ &\quad + \sum_i \int_0^1 \mathbf{E}\,X_i^2 [f_h'(S_n^{(i)}) - f_h'(S_n^{(i)} + tX_i)]\,\mathrm{d}t. \end{aligned} \tag{82}$$

The most direct way of proceeding now would be to estimate both terms using $\|f_h''\|_\infty \le 2\|h'\|_\infty$ due to [1, Lemma 2.4]. The proof of this fact is somewhat tedious, though, and therefore we instead use (72) to express the differences of f_h' in terms of differences of h and $x \mapsto x f_h(x)$.

Using (72) and $\mathbf{E}|S_n^{(i)}| \le \sqrt{\mathbf{E}(S_n^{(i)})^2} \le 1$ we find for any $t \in [0,1]$

$$\begin{aligned} &\mathbf{E}\Big[\big|f_h'(S_n^{(i)} + tX_i) - f_h'(S_n^{(i)})\big|\Big|X_i\Big] \\ &\le \mathbf{E}\Big[\big|h(S_n^{(i)} + tX_i) - h(S_n^{(i)})\big|\Big|X_i\Big] + \mathbf{E}\Big[\big|(S_n^{(i)} + tX_i) f_h(S_n^{(i)} + tX_i) - S_n^{(i)} f_h(S_n^{(i)})\big|\Big|X_i\Big] \\ &\le t|X_i|\Big(\|h'\| + \|f_h'\|\,\mathbf{E}|S_n^{(i)}| + \|f_h\|\Big) \le 3t|X_i|\|h'\|_\infty. \end{aligned}$$

which together with $\int t\,\mathrm{d}t = 1/2$ and $\mathbf{E}|X_i|\,\mathbf{E}\,X_i^2 \le \mathbf{E}|X_i|^3$ concludes the proof. □

One advantage of Theorem 7.49 compared to the earlier CLTs in Theorems 7.39 and 7.45 is that the error bound is explicit in terms of third moments and depends on the test function h only via $\|h'\|_\infty$. This allows to obtain an estimate in the so called *Wasserstein metric*, and indirectly also in *Kolmogorov metric*.

Definition 7.50. On the space of probability measures on $\mathbb{R}$ we define the *Wasserstein distance* and *Kolmogorov distance* by

$$\begin{aligned} d_W(P,Q) &:= \sup\left\{\left|\int h\,\mathrm{d}(P-Q)\right| \,\middle|\, \|h'\|_\infty \le 1\right\} \\ d_K(P,Q) &:= \sup\left\{|P((-\infty,x]) - Q((-\infty,x])| \,\middle|\, x \in \mathbb{R}\right\}. \end{aligned} \tag{83}$$

Note that in the iid. case $X_k \overset{d}{=} X/\sqrt{n}$ for some random variable X the rhs. is $4.5\,\mathbf{E}|X|^3/\sqrt{n}$. Thus for functions with bounded derivatives and random variables with three finite moments we have a convergence rate of $n^{-1/2}$ in Theorem 7.39.

Thus Theorem 7.49 implies

$$d_W(S_n, N) \le \frac{9}{2} \sum_i \mathbf{E}|X_i|^3 \tag{84}$$

and (see Exercise 7.51 below)

$$d_K(S_n, N) = \sup_x \left| P(S_n \le x) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-x^2/2}\,\mathrm{d}x \right| \le \frac{3}{(2\pi)^{1/4}} \sqrt{\sum_i \mathbf{E}|X_i|^3} \tag{85}$$

There is a long history of results improving the constant C starting from $C = 7.5$ in [2] up to the current record of $C = 0.64$ in [3]

The bound (85) is a weak version of the Berry-Esséen bound

$$d_K(S_n, N) \le C \sum_i \mathbf{E}|X_i|^3 \tag{86}$$

which can be proved using similar methods as Theorem 7.49 by considering solutions $f_z = f_{1_{(-\infty,z]}}$ to equation (72) for indicator functions $h = 1_{(-\infty,z]}$. The somewhat technical argument is omitted here, but can be found in [1, Section 3.4].

Exercise 7.51 (Distance between probability measures).

(i) Prove that both d_W, d_K are metrics.

(ii) Show that convergence in either of the two metrics implies weak convergence, i.e. $d_W(P_n, P) \to 0$ or $d_K(P_n, P \to 0)$ implies $P_n \to P$ weakly.

(iii) Prove that the converse is not true in general, i.e. construct an example where $P_n \to P$ weakly but $d_W(P_n, P) \not\to 0$ and $d_K(P_n, P) \not\to 0$.

(iv) Show that for P with continuous distribution function $F(x) = P((-\infty, x])$ the Kolmogorov distance metrizes the weak topology, i.e. $d_K(P_n, P) \to 0$ if and only if $P_n \to P$ weakly. In particular the classical CLT Theorem 7.39 already implies

$$\sup_x \left| P(Y_n \leq x) - \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-x^2/2\sigma^2}\, dx \right| \to 0. \tag{87}$$

(v) Show that whenever one of the two measures P, Q has bounded density f on $\mathbb{R}$, then

$$d_K(P, Q) \leq \sqrt{2\|f\|_\infty d_W(P, Q)}. \tag{88}$$

Exercise 7.52. The goal of this exercise is to prove the Berry-Esséen theorem, i.e. that in the setting of Theorem 7.49 it holds that

$$|P(S_n \leq x) - F_N(x)| \leq C \sum_{k=1}^{n} \mathbf{E}|X_i|^3, \qquad F_N(x) := P(N \leq x) \tag{89}$$

for C = 16. The proof is based on induction on the number of random variables, so you may assume that (89) is known for S_n replaced by $S_n^{(i)}/\sigma_{(i)}$ for each i, where $S_n^{(i)} = S_n - X_i$ and $\sigma_{(i)}^2 := \mathbf{E}(S_n^{(i)})^2$.

(i) Let $\mu_3 := \sum_i \mathbf{E}|X_i|^3$ and show that $\sigma_{(i)}^2 \geq 1 - \mu_3^{2/3}$, so that without loss of generality we may assume $\mu_3 \leq 1/C$ and $\min_i \sigma_{(i)} := \sigma \geq \sqrt{1 - C^{-2/3}}$.

(ii) For $\epsilon > 0, z \in \mathbb{R}$ define $h_{z,\epsilon}$ to be the continuous piecewise linear function equal to 1 on $(-\infty, z - \epsilon)$ and equal to 0 on $(z + \epsilon, \infty)$. Show that

$$\sup_z |P(S_n \leq z) - F_N(z)| \leq \sup_z |\mathbf{E}\, h_{z,\epsilon}(S_n) - \mathbf{E}\, h_{z,\epsilon}(N)| + \epsilon\sqrt{\frac{2}{\pi}}. \tag{90}$$

(iii) Argue as in the proof of Theorem 7.49 to to show that

$$|\mathbf{E}\, h_{z,\epsilon}(S_n) - \mathbf{E}\, h_{z,\epsilon}(N)| \leq \frac{3}{2}\left(\sqrt{\frac{\pi}{2}} + 2 + \frac{1}{\sigma\sqrt{2\pi}} + \frac{C\mu_3}{\epsilon\sigma^3}\right)\mu_3. \tag{91}$$

Hint. The solution $f_{z,\epsilon}$ to the Stein equation $f'_{z,\epsilon}(x) - xf_{z,\epsilon}(x) = h_{z,\epsilon}(x) - \mathbf{E}\, h_{z,\epsilon}(\mathrm{N})$ can be chosen such that $\|f_{z,\epsilon}\|_\infty \leq \sqrt{\pi/2}$ and $\|f'_{z,\epsilon}\|_\infty \leq 2$. Then estimate

$$\mathbf{E}\Big[\big|h_{z,\epsilon}(\mathrm{S}_n^{(i)} + t\mathrm{X}_i) - h_{z,\epsilon}(\mathrm{S}_n^{(i)})\big|\Big|\mathrm{X}_i\Big] \leq \frac{t|\mathrm{X}_i|}{2\epsilon}\sup_y \mathrm{P}(y - \epsilon < \mathrm{S}_n^{(i)} \leq y + \epsilon). \quad (92)$$

and apply the induction hypothesis to compare the above probability to the probability of $(y-\epsilon)/\sigma_{(i)} < \mathrm{N} < (y+\epsilon)/\sigma_{(i)}$.

(iv) Choose ϵ suitably to show that for $\mathrm{C} = 16$ it holds that

$$\frac{3}{2}\Big(\sqrt{\frac{\pi}{2}} + 2 + \frac{1}{\sigma\sqrt{2\pi}} + \frac{\mathrm{C}\mu_3}{\epsilon\sigma^3}\Big)\mu_3 + \epsilon\sqrt{\frac{2}{\pi}} \leq \mathrm{C}\mu_3 \quad (93)$$

and conclude the proof by induction. *Hint. It might be good to take out a calculator.*

7.10 *Stable distributions*

Lecture 7
13th December 2022

Today we are asking what kind of probability distributions can be the weak limit of sums of independent random variables. For instance if $(\mathrm{X}_n)_{n\geq 1}$ is a sequence of iid. random variables, and $(\alpha_n)_{n\geq 1}$ is a sequence of deterministic numbers $\alpha_n > 0$ such that

$$\frac{\mathrm{X}_1 + \cdots + \mathrm{X}_n}{\alpha_n} \to \mathrm{Y}, \qquad n \to \infty \quad (94)$$

in distribution for some random variable Y. To avoid trivialities we will assume that Y is non-degenerate (i.e. for all $x \in \mathbb{R}$, $\mathrm{P}(\mathrm{Y} = x) < 1$).

A basic observation is that upon "splitting the sum" in (94) we can infer certain additivity properties of Y. For fixed $m \geq 1$ and growing $n \to \infty$ we can evaluate the distribution limit of $\mathrm{Z}_{nm} := (\mathrm{X}_1 + \cdots + \mathrm{X}_{nm})/\alpha_{nm}$ in two ways. On the one hand we have $\mathrm{Z}_{nm} \to \mathrm{Y}$ in distribution, as $n \to \infty$, and on the other hand we also have

$$\frac{\alpha_{nm}}{\alpha_n}\mathrm{Z}_{nm} = \frac{\mathrm{X}_1 + \cdots + \mathrm{X}_n}{\alpha_n} + \cdots + \frac{\mathrm{X}_{(m-1)n+1} + \cdots + \mathrm{X}_{mn}}{\alpha_1} \to \mathrm{Y}_1 + \cdots \mathrm{Y}_m \quad (95)$$

in distribution as $n \to \infty$, where $\mathrm{Y}_1, \ldots, \mathrm{Y}_m$ are independent copies of Y. We now claim that

$$a_m := \lim_{n\to\infty} \frac{\alpha_{nm}}{\alpha_n} \in (0,\infty) \quad \text{exists for each } m. \quad (96)$$

Indeed, let a_m be a subsequential limit $a_m := \lim_k \alpha_{n_k m}/\alpha_{n_k} \in [0,\infty]$ with $0,\infty$ included. If $a_m = 0$, then $\alpha_{n_k m}\mathrm{Z}_{n_k m}/\alpha_{n_k}$ converges in distribution to $0 \cdots \mathrm{Y} = 0$ and $\mathrm{Y}_1 + \cdots + \mathrm{Y}_m$ implying that $\mathrm{Y}_1 + \cdots + \mathrm{Y}_m = 0$ almost surely, contradicting the non-degeneracy of Y. On the other hand if $a_m = \infty$, then for the sequence $\mathrm{Z}_{n_k m}$ converges in distribution to both Y and $0 \cdot (\mathrm{Y}_1 + \cdots + \mathrm{Y}_m) = 0$, again contradicting the non-degeneracy of Y. Finally, if there are two different subsequential limits $a_m \neq a'_m$ then we have $a_m Y \overset{d}{=} Y_1 + \cdots + Y_m \overset{d}{=} a'_m Y$ which again is impossible by non-degeneracy of Y. Thus we have shown that Y is *stable* in the sense of the following definition.

Definition 7.53. A non-degenerate random variable is called (strictly) stable if for each $m \geq 1$ there exists $a_m > 0$ such that

$$\frac{Y_1 + \cdots + Y_m}{a_m} \overset{d}{=} Y, \quad m \geq 1. \tag{97}$$

There is a slightly more general definition sometimes called "stable in the broad sense" or just "stable" which allows for an additive constant, i.e. $(Y_1 + \cdots + Y_m)/a_m - b_m \overset{d}{=} Y$.

Proposition 7.54. *A random variable is stable if and only if it is the weak limit of some iid. sum as in* (94).

Proof. We already saw that any weak limit is necessarily stable. Conversely, if Y is stable we can simply choose $X_n \overset{d}{=} Y$ and $\alpha_n = a_m$ so that for each n, $(X_1 + \cdots + X_n)/\alpha_n \overset{d}{=} Y$. □

We recall that the normal and the Cauchy distribution are *stable* in the sense

$$\begin{aligned} \frac{C_1 + \cdots + C_n}{n} &\overset{d}{=} C \\ \frac{N_1 + \cdots + N_n}{\sqrt{n}} &\overset{d}{=} N \end{aligned} \tag{98}$$

for iid. Cauchy-distributed random variables $C_1, \ldots, C_n, C$ and iid. $\mathcal{N}(0,1)$-distributed random variables $N_1, \ldots, N_n, N$. This property can be proved, for instance, using characteristic functions

$$\begin{aligned} \varphi_{(C_1+\cdots+C_n)/n}(t) &= \prod_{i=1}^{n} \varphi_{C_i}\Big(\frac{t}{n}\Big) = \exp\Big(-\frac{|t|}{n}\Big)^n = \exp(-|t|), \\ \varphi_{(N_1+\cdots+N_n)/\sqrt{n}}(t) &= \prod_{i=1}^{n} \varphi_{N_i}\Big(\frac{t}{\sqrt{n}}\Big) = \exp\Big(-\frac{t^2}{2n}\Big)^n = \exp\Big(-\frac{t^2}{2}\Big). \end{aligned} \tag{99}$$

From now on we consider only symmetric random variables. The theory of stable random variables is well understood also in the general case but for the sake of simplicity we will only consider symmetric random variables.

Theorem 7.55. *For a non-degenerate symmetric random variable* Y *the following conditions are equivalent:*

(a) Y *is stable.*

(b) Y *is stable with norming constant* $a_n = n^\lambda$ *for some* $\lambda \geq 1/2$.

(c) $a_1 Y_1 + a_2 Y_2 \overset{d}{=} (a_1^\alpha + a_2^\alpha)^{1/\alpha} Y$ *for* Y_1, Y_2 *iid. copies and* Y *and any* $a_1, a_2 > 0$ *and some* $\alpha \in (0, 2]$.

(d) *The characteristic function of* Y *is of the form* $\varphi_Y(t) = \exp(-c|t|^\alpha)$ *for some* $\alpha \in (0,2]$ *and some* $c > 0$.

Proof. We first check the difficult implication from (a) to (d). Suppose that Y is symmetric stable with norming constants a_n and characteristic function φ. We shall establish the following properties one by one which will imply (d).

(i) The sequence $(a_n)_{n\geq 1}$ is strictly increasing, i.e. $1 = a_1 < a_2 < a_3 < \cdots$, and satisfies $a_{nm} = a_n a_m$.

(ii) The characteristic function satisfies $\varphi(t) = \varphi(-t) > 0$ for all t.

(iii) The sequence $(a_n)_{n\geq 1}$ is given by $a_n = n^\lambda$ for some $\lambda > 0$.

(iv) The characteristic function is given by $\varphi(t) = e^{-c|t|^\alpha}$ for some $c > 0$ and $\alpha = 1/\lambda$

(v) The constant α satisfies $\alpha \in (0,2]$, and therefore $\lambda \geq 1/2$.

Proof of (i). Note that

$$\varphi(t)^n = \varphi(a_n t) \quad \Leftrightarrow \quad \varphi(t/a_n)^n = \varphi(t) \tag{100}$$

for all n, t. If $a_n = a_m$ for some $n < m$, then (100) implies $\varphi(t)^n = \varphi(t)^m$ for all t, i.e. $\varphi(t) \in \{0, \pm 1\}$, so that by continuity $\varphi \equiv 1$, implying that Y is degenerate. On the other hand, if $a_m < a_n$ for some $m > n$, then $|\varphi(ct)| = |\varphi(t)|^{m/n} \leq |\varphi(t)|$ for all t and $c := a_m/a_n$. But this implies $|\varphi(t)| \geq \lim_k |\varphi(c^k t)| = 1$, and by continuity again $\varphi \equiv 1$. By the same argument we find that

$$\varphi(a_{nm}t) = \varphi(t)^{nm} = \varphi(a_n t)^m = \varphi(a_n a_m t) \quad \Rightarrow \quad a_{nm} = a_n a_m. \tag{101}$$

□

Proof of (ii). We have $\varphi(t) = \varphi(-t) \in \mathbb{R}$ by symmetry and $\varphi(0) = 1$. If φ has a root at t, then $\varphi(t) = 0 = \varphi(t/a_2) = 0$ and in particular φ has roots arbitrarily close to 0 which is impossible. □

Proof of (iii). Fix some $1 < m < n$, then for any p there exists a unique q with $n^p \leq m^q < n^{p+1}$ and therefore $a_n^p \leq a_m^q < a_n^{p+1}$. By taking logarithms we have

$$\frac{p}{p+1}\frac{\log a_n}{\log n} < \frac{p}{q}\frac{\log a_n}{\log m} \leq \frac{\log a_m}{\log m} < \frac{p+1}{q}\frac{\log a_n}{\log m} \leq \frac{p+1}{p}\frac{\log a_n}{\log n}. \tag{102}$$

and therefore by taking $p \to \infty$ it follows that $\log a_n / \log n = \log a_m / \log m = \lambda$ for some $\lambda > 0$. □

Proof of (iv). Applying (100) once more together with (iii) implies

$$\varphi(x^\lambda t) = \varphi(t)^x, \tag{103}$$

for all rational $x > 0$ and by continuity also for all real $x > 0$. In particular, for $t > 0$ it follows that $\varphi(t) = \exp(t^\alpha \log \varphi(1))$ for $t > 0$ and by symmetry

$\varphi(t) = \exp(|t|^\alpha \log \varphi(1))$. Set $c := -\log \varphi(1) \geq 0$ since $0 < \varphi(t) \leq 1$ and observe that $c = 0$ would imply $\varphi \equiv 1$, once again contradicting non-degeneracy. □

Proof of (v). For $\alpha > 2$ we have

$$\begin{aligned} \mathbf{E}\,Y^2 &= 2\,\mathbf{E} \lim_{\epsilon \to 0} \frac{1-\cos(\epsilon Y)}{\epsilon^2} \leq 2 \liminf_{\epsilon \to 0} \mathbf{E}\, \frac{1-\cos(\epsilon Y)}{\epsilon^2} \\ &= 2 \liminf_{\epsilon \to 0} \frac{1-\phi(\epsilon)}{\epsilon^2} = 2 \liminf_{\epsilon \to 0} \frac{c\epsilon^\alpha}{\epsilon^2} = 0 \end{aligned} \tag{104}$$

by Fatou's Lemma and hence $Y = 0$ almost surely. □

The implication from (d) to (c) follows from

$$\varphi_{a_1 Y_1 + a_2 Y_2}(t) = \varphi_Y(a_1 t)\varphi_Y(a_2 t) = \exp\Big(-c|t|^\alpha \Big(a_1^\alpha + a_2^\alpha\Big)\Big) = \varphi_Y\Big((a_1^\alpha + a_2^\alpha)^{1/\alpha} t\Big). \tag{105}$$

Assuming (c) it follows inductively that $Y_1 + \cdots + Y_n \overset{d}{=} n^{1/\alpha} Y$ and therefore Y is stable with norming constant $a_n = n^{1/\alpha}$ with $1/\alpha \geq 1/2$, implying (b) and in particular (a). □

Finally we note that stable random variables exist for any index $\alpha \in (0,2]$. For $\alpha = 2$ this is simply a Gaussian random variable. For $\alpha < 2$ we can construct stable random variables e.g. as weak limits of sums of heavy tailed random variables $X_1, X_2, \ldots$ with law

$$\mathrm{d}P(x) = \alpha \frac{\mathbf{1}(|x| \geq 1)}{2|x|^{1+\alpha}}\, \mathrm{d}x \tag{106}$$

Then the characteristic function is given by

$$1 - \varphi_P(t) = \int_{\mathbb{R}} (1 - e^{itx})\, \mathrm{d}P(x) = \alpha \int_1^\infty \frac{1-\cos(tx)}{x^{1+\alpha}}\, \mathrm{d}x = \alpha |t|^\alpha \int_{|t|}^\infty \frac{1-\cos x}{x^{1+\alpha}}\, \mathrm{d}x. \tag{107}$$

Since $1 - \cos x = x^2/2 + O(x^4)$ as $x \to \infty$ and $\alpha < 2$ the integral is convergent near $x = 0$ and we have

$$\varphi_P(t) = 1 - c|t|^\alpha + o(|t|^\alpha), \quad |t| \to 0. \tag{108}$$

Exercise 7.56. Suppose $X_1, X_2, \ldots$ are iid. with characteristic function (108) for some $\alpha \in (0,2)$. Show that

$$\frac{X_1 + \cdots + X_n}{n^{1/\alpha}} \to Y \tag{109}$$

in distribution, where Y is a stable random variable with index α and characteristic function $\varphi_Y(t) = \exp(-c|t|^\alpha)$.

Remark 7.57. This time, we can note that for any positive y,

$$P(\max(|Y_1|, \ldots, |Y_n|) \leq y n^{1/\alpha}) = P(|Y_1| \leq y n^{1/\alpha})^n \leq (1 - c y^{-\alpha}/n)^n \to \exp(-c y^{-\alpha})$$

for some constant c. In particular, the limsup of these probabilities is strictly smaller than 1. This shows that (with high probability when n is large), the largest of the values $|Y_j|$ for $j \leq n$, will be of the same order of magnitude than $Y_1 + \cdots + Y_n$, which contrasts with the case of the sums of i.i.d. variables that are in L^2 (as discussed in Remark 7.40).

8 LARGE DEVIATIONS

8.1 *Motivation*

Lecture 8
20th December 2022

In the insurance application X_n could be the claim of the n-th client of a insurance company. In order to make a profit the insurance company charges a price of $x > \overline{x}$ exceeding the average claim $\overline{x} = \mathbf{E} X_n$ from each client. The probability of making a loss is then $P(X_1 + \cdots + X_n > xn)$.

Motivated by insurance applications, Cramér was interested in the probability that the empirical mean $\overline{X}_n$ of n independent identically distributed random variables $X_1, \ldots, X_n$ exceeds the true mean $\overline{x} := \mathbf{E} X_i$ by a given amount, say

$$P(\overline{X}_n > x) = P(S_n > nx), \quad \overline{X}_n := \frac{S_n}{n}, \quad S_n := X_1 + \cdots + X_n. \tag{110}$$

A first observation is that by the law of large numbers

$$\lim_n P(\overline{X}_n > x) = 0 \quad \text{for all} \quad x > \overline{x} = \mathbf{E} \overline{X}_n \tag{111}$$

but this does not give any indication on the rate of convergence. In case of finite variance $\operatorname{Var} X_i < \infty$ we find

$$P(\overline{X}_n > x) = P(\overline{X}_n - \overline{x} > x - \overline{x}) \leq \frac{\mathbf{E}(\overline{X}_n - \overline{x})^2}{(x - \overline{x})^2} = \frac{\operatorname{Var} X_1}{n(x - \overline{x})^2}, \tag{112}$$

Without second moments such a bound is not possible. Indeed, take X_i to be symmetric stable of index $\alpha \in (1,2)$ (so that the mean $\overline{x} = 0$ is well defined but the variance is infinite). Then $\overline{X}_n \overset{d}{=} n^{1/\alpha - 1} X_1$ so that $P(\overline{X}_n > x) = P(X_1 < n^{1-1/\alpha} x) \sim n^{1-\alpha} x^{-1}$.

so the convergence rate is at most $1/n$ for each fixed $x > \overline{x}$.

Example 8.1. Take $X_1 \sim \mathcal{N}(\overline{x}, \sigma^2)$, then $\overline{X}_n - \overline{x} \sim \mathcal{N}(0, \sigma^2 n^{-1})$ and therefore

$$P(\overline{X}_n > x) = 1 - \Phi\Big(\sqrt{n}\frac{x - \overline{x}}{\sigma}\Big), \quad \Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, \mathrm{d}t. \tag{113}$$

By estimating

$$\int_x^\infty e^{-t^2/2}\, \mathrm{d}t \leq \frac{1}{x} \int_x^\infty t e^{-t^2/2}\, \mathrm{d}t = \frac{e^{-x^2/2}}{x} \tag{114}$$

Given that the the normal distribution has exponentially small large deviations it might be tempting to use the central limit theorem to compare the empirical mean of an arbitrary distribution with the empirical mean of a normal distribution. However, the error in this comparison is of order $n^{-1/2}$ which is much larger than what we are hoping for.

and

$$\int_x^\infty e^{-t^2/2}\, \mathrm{d}t \geq \int_x^\infty \Big(1 - \frac{3}{t^4}\Big) e^{-t^2/2}\, \mathrm{d}t = \Big(\frac{1}{x} - \frac{1}{x^3}\Big) e^{-x^2/2} \tag{115}$$

for $x > 0$ we conclude that

$$P(\overline{X}_n > x) \sim \frac{\sigma}{\sqrt{n}(x - \overline{x})} \exp\Big(-n\Big(\frac{x - \overline{x}}{\sigma}\Big)^2\Big), \quad x > \overline{x}, \tag{116}$$

so in fact we have exponentially fast convergence for any $x > \overline{x}$.

Example 8.2. If the distribution of X_i has somewhat heavy tails, then exponentially small large devations are not possible. For example, suppose X_i is symmetric with $P(|X_i| > x) \sim \exp(-\sqrt{x})$ for large x. Then

$$P(\overline{X}_n > x) \geq P(X_1 > nx)P(X_2 + \cdots + X_n > 0) \sim \exp(-\sqrt{nx}), \tag{117}$$

so the large deviations are sub-exponential. In particular the existence of moments of all orders is not sufficient for exponentially small large deviations.

8.2 *Cramér's theorem*

In the study of exponentially small large deviations the *moment-* and *cumulant-generating functions* play a central role:

Definition 8.3. For a real valued random variable X we define the moment generating function $M = M_X : \mathbb{R} \to (0, \infty]$ by

$$M(t) := \mathbf{E}\, e^{tX} \tag{118}$$

and the cumulant generating function $\Lambda = \Lambda_X : \mathbb{R} \to (-\infty, \infty]$ by

$$\Lambda(t) := \log M(t) = \log \mathbf{E}\, e^{tX}. \tag{119}$$

Because we allow both functions to take the value $+\infty$, we also introduce the domain

$$D_\Lambda = D_M = \{t \in \mathbb{R} \mid \Lambda(t) < \infty\} = \{t \in \mathbb{R} \mid M(t) < \infty\}. \tag{120}$$

Lemma 8.4 (Properties of M, Λ).

(i) Λ is convex, and in particular D_Λ is always an interval containing 0.

(ii) If $t \in D^o_\Lambda$, then for each $k \in \mathbb{N}$ it holds that

$$\mathbf{E}|X|^k e^{tX} < \infty, \quad M^{(k)}(t) = \mathbf{E}\, X^k e^{tX} \tag{121}$$

and in particular $M, \Lambda \in C^\infty(D^o_\Lambda)$.

(iii) If $\overline{x} = \mathbf{E}\, X \in \mathbb{R}$ exists, then $\Lambda(t) \geq t\overline{x}$ for all t.

Proof.

(i) The convexity of Λ follows from Hölder's inequality in the form

$$M(\alpha s + (1-\alpha)t) = \mathbf{E}\, e^{\alpha s X} e^{(1-\alpha)tX} \leq (\mathbf{E}\, e^{sX})^\alpha (\mathbf{E}\, e^{tX})^{1-\alpha} = M(s)^\alpha M(t)^{1-\alpha} \tag{122}$$

and the claim follows upon taking logarithms.

(ii) Since the exponential grows faster than any polynomial for any $k \in \mathbb{N}, \epsilon > 0$, there exists $C < \infty$ such that $|X|^k \leq e^{\epsilon|X|} + C \leq e^{\epsilon X} + e^{-\epsilon X} + C$. It follows that

$$\mathbf{E}|X|^k e^{tX} \leq C M(t) + M(t+\epsilon) + M(t-\epsilon) < \infty \tag{123}$$

for all $t \in D_\Lambda^o$ by choosing ϵ small enough. For the derivative we argue by induction and compute the difference quotient as

$$M^{(k+1)}(t) = \lim_{h\to 0} \frac{M^{(k)}(t+h) - M^{(k)}(t)}{h} = \lim_{h\to 0} \mathbf{E}\, X^k e^{tX} \frac{e^{hX}-1}{h} = \mathbf{E}\, X^{k+1} e^{tX}, \tag{124}$$

where we used dominated convergence with $(e^{hX}-1)/h = Xe^{h'X}$ for some $|h'| \le |h|$ due to the mean value theorem. The integrability of the majorant $|X|^{k+1} e^{(t+h')X}$ follows from the previous bound for sufficiently small $|h'|$.

(iii) Due to the concavity of the logarithm it follows from Jensen's inequality that

$$\Lambda(t) = \log \mathbf{E}\, e^{tX} \ge \mathbf{E} \log e^{tX} = t\,\mathbf{E}\, X. \qquad \square$$

With Definition 8.3 we have the *exponential Chernoff bounds*

$$\begin{aligned} P(\overline{X}_n \ge x) &= P(e^{tS_n} \ge e^{tnx}) \le \mathbf{E}\, e^{tS_n - ntx} = e^{ntX_1 - ntx} = e^{-n(tx-\Lambda(t))}, \quad t > 0 \\ P(\overline{X}_n \le x) &= P(e^{tS_n} \ge e^{tnx}) \le \mathbf{E}\, e^{tS_n - ntx} = e^{ntX_1 - ntx} = e^{-n(tx-\Lambda(t))}, \quad t < 0. \end{aligned} \tag{125}$$

To get the optimal bound, it is natural to optimize over t. By (iii) above we have $tx - \Lambda(t) \le t\overline{x} - \Lambda(t) \le 0$ whenever $t \le 0$ and $x \ge \overline{x}$ or $t \ge 0$ and $x \le \overline{x}$, and in particular we conclude

$$P(\overline{X}_n \ge x) \le e^{-n\Lambda^*(x)} \quad \text{for } x \ge \overline{x}, \quad \text{and} \quad P(\overline{X}_n \le x) \le e^{-n\Lambda^*(x)}, \quad \text{for } x \le \overline{x}, \tag{126}$$

where

$$\Lambda^*(x) := \sup_{t\in\mathbb{R}} \big(tx - \Lambda(t)\big) \tag{127}$$

is the *Legendre transform* of the function Λ.
Example 8.5.

1. For the normal distribution we have

$$M(t) = \mathbf{E}\, e^{Nt} = \frac{1}{\sqrt{2\pi}} \int e^{tx - x^2/2}\, dx = e^{t^2/2}, \quad \Lambda(t) = \frac{t^2}{2} \tag{128}$$

and therefore to find $\Lambda^*(x)$ we need to solve $x = \Lambda'(t) = t$ resulting in $\Lambda^*(x) = x^2 - x^2/2 = x^2/2$.

2. For the Bernoulli distribution with $P(\{0\}) = P(\{1\}) = 1/2$ we have

$$\Lambda(t) = \log \frac{e^t + 1}{2}. \tag{129}$$

For large $t > 0$ it holds $\Lambda(t) \sim t - \log 2$ while for large negative $t < 0$ it holds that $\Lambda(t) \sim -\log 2$, and therefore $\Lambda^*(x) = \infty$ for $x > 1$ or $x < 0$ and

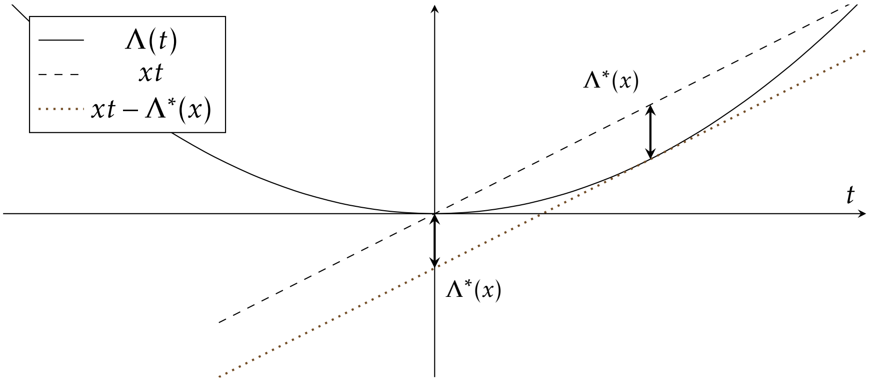

Figure 2: Geometric interpretation of Λ^*: For given x draw the line $t \mapsto tx$ and find the point which is furthest above $\Lambda(t)$, then the distance between the line and the function is $\Lambda^*(x)$. By maximality it follows that the line shifted down by $\Lambda^*(x)$ is tangent to Λ, and therefore $\Lambda^*(x)$ is the distance to 0 of the intercept of the tangent line to Λ with slope x and the vertical axis.

$\Lambda^*(0) = \Lambda^*(1) = \log 2$. For $x \in (0,1)$ to maximize $tx - \Lambda(t)$ we have to solve

$$x = \Lambda'(t) = \frac{e^t}{e^t + 1} \tag{130}$$

resulting in $t = \log x - \log(1-x)$ and

$$\Lambda^*(x) = \begin{cases} \log 2 + x\log x + (1-x)\log(1-x), & x \in [0,1], \\ \infty, & x \in [0,1]^c. \end{cases} \tag{131}$$

3. For the exponential distribution with parameter $\lambda = 1$ we have

$$\Lambda(t) = \log \int_0^\infty e^{x(t-1)}\,\mathrm{d}x = \begin{cases} -\log(1-t), & t < 1, \\ \infty, & t \geq 1. \end{cases} \tag{132}$$

and therefore $\Lambda^*(x) = \infty$ for each $x \leq 0$. For $x > 0$ we have to solve $x = \Lambda'(t)$ resulting in $t = 1 - 1/x$ and therefore

$$\Lambda^*(x) = \begin{cases} x + \log x - 1, & x > 0, \\ \infty, & x < 0. \end{cases} \tag{133}$$

Lemma 8.6 (Properties of Λ^*).

(a) $\Lambda^*(x) \geq 0$ *for all* $x \in \mathbb{R}$

(b) Λ^* *is convex.*

(c) $\Lambda^*(\overline{x}) = 0$ *if* $\overline{x} \in \mathbb{R}$ *exists.*

(d) *If* $\mathrm{D}_\Lambda = \{0\}$, *then* $\Lambda^* \equiv 0$.

Proof. The first statement is obvious since $0x - \Lambda(0) = 0$. The convexity is clear because the pointwise supremum of convex (affine) functions is convex. The fact that $\Lambda^*(\overline{x}) = 0$ follows from (iii) of Lemma 8.4. Finally, if $\Lambda(t) = \infty$ for all $t \neq 0$, then the supremum has to be attained in 0 and equals 0. □

By combining the two bounds in (126) we obtain the upper bound of Cramér's theorem.

Theorem 8.7 (Cramér). *Let* $(\mathrm{X}_n)_{n\geq 1}$ *be a sequence of independent random variables. Then* $\overline{\mathrm{X}}_n := (\mathrm{X}_1 + \cdots + \mathrm{X}_n)/n$ *satisfies the* large deviations principle

$$\limsup_n \frac{\log \mathrm{P}(\overline{\mathrm{X}}_n \in \mathrm{C})}{n} \leq -\inf_{x\in\mathrm{C}} \Lambda^*(x), \quad \liminf_n \frac{\log \mathrm{P}(\overline{\mathrm{X}}_n \in \mathrm{O})}{n} \geq -\inf_{x\in\mathrm{O}} \Lambda^*(x)$$

for all closed sets $\mathrm{C} \subset \mathbb{R}$ *and open sets* $\mathrm{O} \subset \mathbb{R}$.

Proof of the upper bound in Theorem 8.7 in case $\overline{x} \in \mathbb{R}$. If $\inf_{x\in\mathrm{C}} \Lambda^*(x) = 0$ the statement is trivial because the left hand side is non-positive, therefore we may assume that $\Lambda^*(x) > 0$ for all $x \in \mathrm{C}$, so that in particular $\overline{x} \notin \mathrm{C}$ because $\Lambda^*(\overline{x}) = 0$. Because C is closed it follows that $\mathrm{C} \subset (-\infty, \alpha] \cup [\beta, \infty)$ for some $\alpha < \overline{x} < \beta$ with $\alpha, \beta \in \mathrm{C}$ (or either $\alpha = -\infty$, or $\beta = \infty$) and therefore by (126) we have

$$\mathrm{P}(\overline{\mathrm{X}}_n \in \mathrm{C}) \leq \mathrm{P}(\overline{\mathrm{X}}_n \leq \alpha) + \mathrm{P}(\overline{\mathrm{X}}_n \geq \beta) \leq e^{-n\Lambda^*(\alpha)} + e^{-n\Lambda^*(\beta)} \leq 2e^{-n\inf_{x\in\mathrm{C}} \Lambda^*(x)}, \quad (134)$$

where we used (c) in the last inequality. □

Proof of the lower bound in Theorem 8.7. The claimed lower bound follows from

$$\liminf_n \frac{\log \mathrm{P}(|\overline{\mathrm{X}}_n| < \epsilon)}{n} \geq -\Lambda^*(0), \quad \text{for all } \epsilon > 0 \qquad (135)$$

because since O is open, for each $x \in \mathrm{O}$ there exists an open interval $(x - \epsilon_x, x + \epsilon_x) \subset \mathrm{O}$ and therefore

$$\begin{aligned} \liminf_n \frac{\log \mathrm{P}(\overline{\mathrm{X}}_n \in \mathrm{O})}{n} &\geq \sup_{x\in\mathrm{O}} \liminf_n \frac{\log \mathrm{P}(|\overline{\mathrm{X}}_n - x| < \epsilon_x)}{n} \\ &\geq \sup_{x\in\mathrm{O}} (-\Lambda^*_{\mathrm{X}_1 - x}(0)) = -\inf_{x\in\mathrm{O}} \Lambda^*(x). \end{aligned} \qquad (136)$$

Here in the last step we used

$$\Lambda^*_{\mathrm{X}_1 - x}(0) = \sup_{t\in\mathbb{R}} (-\Lambda_{\mathrm{X}_1 - x}(t)) = \sup_{t\in\mathbb{R}} (xt - \Lambda_{\mathrm{X}_1}(t)) = \Lambda^*_{\mathrm{X}_1}(x) \qquad (137)$$

due to $\Lambda_{\mathrm{X}_1 - x}(t) = \log \mathbf{E}\, e^{t\mathrm{X}_1 - tx} = -tx + \mathbf{E}\, e^{t\mathrm{X}_1} = -tx + \Lambda_{\mathrm{X}_1}(t)$. □

Proof of (135). We give the full proof only in case where Λ attains its minimum in some $t_0 \in \mathrm{D}^o_\Lambda$, so that

$$\Lambda^*(0) = \sup_{t\in\mathbb{R}} (-\Lambda(t)) = -\inf_{t\in\mathbb{R}} \Lambda(t) = -\Lambda(t_0), \quad \text{and} \quad \Lambda'(t_0) = 0. \qquad (138)$$

We then consider the "exponentially tilted" independent identically distributed random variables $(Y_n)_{n\geq 1}$ defined such that

$$\mathbf{E}\, f(Y_1) = \mathbf{E}\, f(X_1) e^{t_0 X_1 - \Lambda(t_0)} \tag{139}$$

i.e. with law $dP_{Y_1}(x) = e^{t_0 x - \Lambda(t_0)}\, dP_{X_1}(x)$. The tilted random variables have zero mean and finite absolute first moment since

$$\mathbf{E}\, Y_1 = \frac{\mathbf{E}\, X_1 e^{t_0 X}}{\mathbf{E}\, e^{t_0 X_1}} = \frac{M'(t_0)}{M(t_0)} = \Lambda'(t_0) = 0, \quad \mathbf{E}|Y_1| = \frac{\mathbf{E}|X_1| e^{t_0 X_1}}{\mathbf{E}\, e^{t_0 X_1}} < \infty \tag{140}$$

by Lemma 8.4, and therefore the law of large numbers implies $\overline{Y}_n \to 0$ in probability (or even almost surely). But then it follows that for any $0 < \delta < \epsilon$

$$\begin{aligned} P(|\overline{X}_n| < \epsilon) \geq P(|\overline{X}_n| < \delta) &\geq e^{n[\Lambda(t_0) - \delta|t_0|]}\, \mathbf{E}\, \mathbf{1}(|\overline{X}_n| < \delta) e^{t_0 X_1 - \Lambda(t_0)} \ldots e^{t_0 X_n - \Lambda(t_0)} \\ &= e^{n[\Lambda(t_0) - \delta|t_0|]}\, \mathbf{E}\, \mathbf{1}(|\overline{Y}_n| < \delta) = e^{n[\Lambda(t_0) - \delta|t_0|]} P(|\overline{Y}_n| < \delta) \end{aligned}$$

so that

$$\liminf_n \frac{\log P(|\overline{X}_n| < \epsilon)}{n} \geq \Lambda(t_0) - \delta|t_0| + \liminf_n \log P(|\overline{Y}_n| < \delta) = \Lambda(t_0) - \delta \tag{141}$$

and the claim follows because δ was arbitrary and $\Lambda(t_0) = \Lambda^*(0)$. □

REFERENCES

[1] L. H. Y. Chen, L. Goldstein, and Q.-M. Shao, *Normal approximation by Stein's method*, Probability and its Applications (New York) (Springer, Heidelberg, 2011), pp. xii+405.

[2] C.-G. Esseen, *On the Liapounoff limit of error in the theory of probability*, Ark. Mat. Astr. Fys. **28A**, 19 (1942).

[3] I. S. Tyurin, *On the rate of convergence in Lyapunov's theorem*, Teor. Veroyatn. Primen. **55**, 250–270 (2010).