Probability Theory

Daniele Zago

October 17, 2021

Contents

Lecture 1: Convergence and limit theorems

2021-10-14

References Gut (2009), first portion of the course

Email: stefano.pagliarani9@unibo.it

The course will be focussed on the stochastic processes portion of probability theory, after a brief reminder of limit theorems, conditional probability, and measure theory.

1.1 Convergence of random variables

Convergence of random variables is a little bit trickier than just real numbers.

Notation: AC is the set of absolutely continuous probability measures wrt the Lebesgue measure.

- *Absolute continuity*: if $\mu \in \mathrm{AC}$ is absolutely continuous, we write

$$\mu(dx) = f(x)dx$$

- *Integration in measure spaces:* Let $X \sim \mu$, then by a theorem we have

$$\mathbb{E}\big[f(X)\big] = \int_{\mathbb{R}^d} f(x)\mu(dx), \tag{1}$$

and we can differentiate between two types of distribution:

a) μ discrete $\implies \mathbb{E}\big[X\big] = \sum_n xp(x)$

b) $\mu \in \mathrm{AC} \implies \mathbb{E}\big[X\big] = \int_{\mathbb{R}^d} x \cdot f(x)dx$

Example (Intuition of convergence)

Consider $\mu_n = \mathrm{Unif}_{[0,\frac{1}{n}]}$ for $n \in \mathbb{N}$, and it is absolutely continuous w.r.t. Lebesgue measure. This means that it admits a probability density which is defined by

$$\mu_n(dx) = \left(\begin{cases} n & \text{if } x \in [0,\frac{1}{n}] \\ 0 & \text{if } x \notin [0,\frac{1}{n}] \end{cases}\right) dx$$

It is intuitive to think that the measure is converging to a spike in zero, i.e.

$$\mu_n \xrightarrow{n\to\infty} \delta_0,$$

where δ_x denotes the Dirac delta distribution centered in x, such that $\delta_x(\{x\}) = 1$. We need to mathematically characterize this type of convergence in a more formal way than by intuition.

Maybe it could be that for any Borel set $A \subseteq \mathscr{B}(\mathbb{R})$,

$$\mu_n(A) \xrightarrow{n\to\infty} \delta_0(A),$$

but unfortunately this is wrong since we can see that, for $A = \{0\}$ and for all $n \in \mathbb{N}$:

$$\mu_n(\{0\}) = 0 \neq 1 = \delta_0(\{0\}).$$

So we can either throw out the idea that the uniform converges to a Dirac delta, or change the definition of convergence to accommodate for the behaviour in Figure 1.

Moreover, assume now that $X_n \sim \mu_n$ such that $\mu_n \xrightarrow{n\to\infty} \delta_0$, what can we say about the properties of X_n? In general (as we will see afterwards), this depends on the specific type of convergence that we assume.

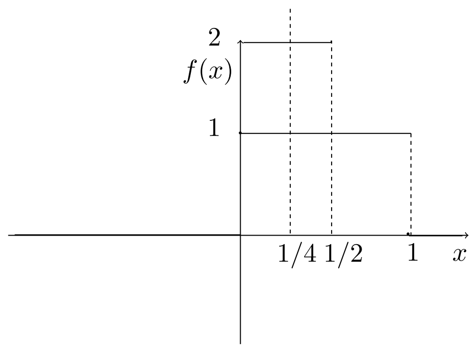

Figure 1: Convergence of the sequence of uniform distributions to the Dirac measure in zero.

Def. (Convergence in distribution)
Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of distributions on $(\mathbb{R}^d, \mathscr{B})$. We say that μ_n *converges in distribution* to another distribution μ,

$$\mu_n \xrightarrow{d} \mu,$$

if, for any possible choice of *test function* $f \in C_b(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d} f(x)\mu_n(dx) \xrightarrow{n\to\infty} \int_{\mathbb{R}^d} f(x)\mu(dx).$$

This convergence is in the sense of standard real analysis.

Notation: $C_b(\mathbb{R}^d)$ is the set of continuous bounded functions

Remark

All test functions f define a measure when integrated wrt to $\mu_n(dx)$, and when all said measures are equal to those obtained by integrating against another distribution μ, then we obtain the convergence in distribution.

Example (Uniform distribution)

Consider $\mu_n = \text{Unif}_{[0,\frac{1}{n}]}$ and $\mu = \delta_0$, take any function $f \in C_b(\mathbb{R})$ and compute

$$\begin{aligned}\int_{\mathbb{R}} f(x)\mu_n(dx) &= \int_0^{\frac{1}{n}} f(x) \cdot n \cdot dx \\ &= n \cdot \underbrace{\int_{[0,\frac{1}{n}]} f(x)dx}_{\approx \frac{1}{n} \cdot f(0)} \\ &\xrightarrow{n\to\infty} f(0).\end{aligned}$$

The last equality holds since f is continuous, and by the mean value theorem we can approximate it by the left extrema. However, by definition of the abstract integral wrt the Dirac delta function we have that

$$f(0) = \int_{\mathbb{R}} f(x)\delta_0(dx),$$

which proves that $\mu_n \xrightarrow{d} \mu$.

Remark

If $A \in \mathscr{B}(\mathbb{R}^d)$ is an event and μ is a distribution, then

$$\mu(A) = \int_{\mathbb{R}^d} \mathbb{1}_A(x)dx,$$

where $\mathbb{1}_A$ is the indicator function such that

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

Had we used $f \notin C_b(\mathbb{R}^d)$ instead, then we could have chosen $f = \mathbb{1}_{\{0\}}$ and convergence in distribution would not have been satisfied. The example below shows another case in which another type of convergence is useful in order to characterize a common-sense behaviour of random variables.

Example (Sequence of Dirac functions)

Consider $\mu_n = \delta_{1/n}$ and $\mu = \delta_0$, then it is clear that this is a discrete measure that in some intuitive sense converges to zero. If we choose $f(x) = \mathbb{1}_{\{0\}}$, then we find that

$$\int_{\mathbb{R}} f(x)\mu_n(dx) = \int_{\mathbb{R}} \mathbb{1}_{\{0\}}(x)\delta_{\frac{1}{n}}(dx) = \mathbb{1}_{\{0\}}(1/n) = 0 \quad \forall n,$$

and therefore does not converges to δ_0.

Recall: A random variable is such that the event $(X_n \in A) \in \mathcal{F}_n$, which means that the function is measurable.

Def. (Weak convergence of random variables)
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables, $X_n : (\Omega_n, \mathcal{F}_n, \mathbb{P}_n) \longrightarrow (\mathbb{R}^d, \mathscr{B})$. Let now X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Then, we say that X_n *converges weakly/in distribution/in law*, $X_n \xrightarrow{d} X$, if their measures are such that

$$\mu_{X_n} \xrightarrow{d} \mu_X.$$

Remark

By the definition of expected value in Equation (1), a family of random variables $(X_n)_{n\in\mathbb{N}}$ is such that, for any $f \in C_b(\mathbb{R}^d)$

$$X_n \xrightarrow{d} X \iff \mathbb{E}\big[f(X_n)\big] \xrightarrow{n\to\infty} \mathbb{E}\big[f(X)\big].$$

This is however the weakest type of convergence out of all those that we will consider, since in other cases the probability spaces might be different.

Def. (Stronger definitions of convergence)
$(X_n)_{n\in\mathbb{N}}$ sequence of random variables and X a r.v., all defined on the same probability space

$$X_n, X : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (\mathbb{R}^d, \mathscr{B}).$$

Then we say that

a) If $X_n, X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ for $p \geq 1$, where

$$L^p = \left\{\text{r.v. on } (\Omega, \mathcal{F}, \mathbb{P}) \text{ such that } \mathbb{E}\big[|X|^p\big] < \infty\right\}.$$

then $X_n \xrightarrow{L^p} X$ if

$$\|X_n - X\|_{L^p} \xrightarrow{n\to\infty} 0,$$

where $\|X\|_{L^p} = \mathbb{E}\big[|X|^p\big]^{\frac{1}{p}}$.

b) X_n *converges in probability* to X, $X_n \xrightarrow{P} X$ if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(|X_n - X| \geq \varepsilon\big) = 0.$$

c) X_n *converges almost surely* to X, $X_n \xrightarrow{\text{a.s.}} X$ if

$$\mathbb{P}\Big(\lim_{n\to\infty} X_n = X\Big) = 1,$$

where the event inside $\mathbb{P}$ is in the sense of real analysis,

$$\left\{w \in \Omega : X_n(\omega) \xrightarrow{n\to\infty} X(\omega)\right\},$$

which can be proven to be a measurable set and therefore a valid event.

Remark

The L^p norm of the difference induces a *distance between functions* in the sense of functional analysis.

Example (Difference in interpretation)
Consider a Bernoulli game where we equally bet on an outcome ± 1. The second type of convergence does not tell us that almost surely our gain will converge to zero, but rather that we can set a small tolerance and find some n such that our gain will be smaller than that.

The following inequality is a basic tool for probability, which will be useful later on.

Thm. 1 (Markov's inequality)
Let X be a r.v. and $\lambda > 0$, then

$$\mathbb{P}\big(|X| > \lambda\big) \leq \frac{\mathbb{E}\big[|X|^p\big]}{\lambda^p}, \quad p \geq 0.$$

Proof.
If $\mathbb{E}\big[|X|^p\big] = \infty$, then there is nothing to prove. If instead $\mathbb{E}\big[|X|^p\big] < \infty$, then since $\mathbb{1}_A$ is either 1 or 0 we have

$$\begin{aligned}\mathbb{E}\big[|X|^p\big] &\geq \mathbb{E}\big[|X|^p \cdot \mathbb{1}_{|X|>\lambda}\big] \\ &\geq \mathbb{E}\big[\lambda^p \cdot \mathbb{1}_{|X|>\lambda}\big] && \text{(since } |X| \geq \lambda\text{)} \\ &= \lambda^p \cdot \mathbb{P}\big(|X| > \lambda\big).\end{aligned}$$

□

Corollary (Chebyshev's inequality)
By choosing $p = 2$ and considering the random variable $X - \mathbb{E}[X]$, Markov's inequality states that

$$\mathbb{P}\big[|X - \mathbb{E}[X]| > \lambda\big] \leq \frac{\mathbb{E}\big[|X - \mathbb{E}[X]]|^2}{\lambda^2} = \frac{\mathbb{V}[X]}{\lambda^2}.$$

Thm. 2
Under the according assumptions for X_n, X we have the following set of implications:

1. $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{d} X$.
2. $X_n \xrightarrow{P} X \implies$ *there is a subsequence X_{k_n} such that $X_{k_n} \xrightarrow{a.s.} X$.*
3. $X_n \xrightarrow{d} X \implies X_n \xrightarrow{P} X$ *iff $\mu_X = \delta_{x_0}$*
4. $X_n \xrightarrow{L^1} X \implies X_n \xrightarrow{P} X$
5. $X_n \xrightarrow{P} X \implies X_n \xrightarrow{L^1} X$ *iff $|X_n| \leq Y \in L^p$*

Proof.

1. $\boxed{\text{a.s.} \implies \text{p}}$: $\mathbb{P}\big(|X_n - X| \geq \varepsilon\big) = \mathbb{E}\big[\mathbb{1}_{|X_n - X| \geq \varepsilon}\big]$ and the indicator function converges to zero as $n \to \infty$ by assumption. Since $\mathbb{1}_A$ is bounded, by the dominated convergence theorem the integral (expectation) also converges to zero.

4. $\boxed{L^p \implies p}$: Follows as a consequence of Markov's property, since we can majorize the probability by the expected value

$$\mathbb{P}\big(|X_n - X| \geq \varepsilon\big) \overset{\text{Thm.1}}{\leq} \frac{\mathbb{E}\big[|X_n - X|^p\big]}{\varepsilon^p} = \frac{\|X_n - X\|_{L^p}^p}{\varepsilon^p} \xrightarrow{n\to\infty} 0.$$

where the convergence to 0 is a consequence of the L^p convergence assumption.

□

Example (A.s. does not imply L^p)

Let $m \in \mathbb{R}$ and $X_n = n^m \mathbb{1}_{[0,\frac{1}{n}]}$ on the probability space $\big([0,1], \mathscr{B}([0,1]), \lambda_{[0,1]}\big) \to \mathbb{R}$, and let's try to establish some convergence for the random variable X_n.

- If $\omega > 0$, then we can find some $\bar{n}$ such that X_n is equal to zero:

$$X_n(\omega) = n^m \mathbb{1}_{[0,\frac{1}{n}]}(\omega) \xrightarrow{n\to\infty} 0.$$

- If $\omega = 0$, then

$$X_n(0) = n^m \xrightarrow{n\to\infty} +\infty, \quad \text{for } m > 0,$$

however the event $\{0\}$ has null probability since we have a uniform distribution on $\big[0, \frac{1}{n}\big]$ at all steps of the limit, and as such we have

$$\mathbb{P}_{\mu_n}\big(\{0\}\big) = 0 \quad \text{for all } n \in \mathbb{N}.$$

Therefore, the set of limit elements for absolute convergence is

$$\Big\{\omega \in \Omega : X_n(\omega) \xrightarrow{n\to\infty} X(\omega)\Big\} = \Omega \setminus \{0\}\,.$$

Since $\mathbb{P}\big(\lim_{n\to\infty} X_n = X\big) = \mathbb{P}\big(\Omega \setminus \{0\}\big) = 1$, we have that

$$X_n \xrightarrow{\text{a.s.}} X \equiv 0 \qquad (\implies X \xrightarrow{P} X).$$

On the other hand for L^p convergence we have that

$$\begin{aligned}\mathbb{E}\big[|X_n - X|^p\big] &= \mathbb{E}\big[|X_n|^p\big] \\ &= \int_{[0,1]} n^{mp} \cdot \mathbb{1}_{[0,\frac{1}{n}]}(x)dx \\ &= n^{mp} \cdot \frac{1}{n} \\ &= n^{mp-1}.\end{aligned}$$

We conclude that $X_n \xrightarrow{L^p} X \iff mp - 1 < 0 \iff m < 1/p$, but we always have almost-sure convergence for any $m > 0$.

Example (Gaussian distribution)

Consider $\mathcal{N}_{\mu,\sigma^2} = \varphi_{\mu,\sigma^2}(x)dx$, with

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}.$$

Consider now a sequence of real numbers $\mu_n \to \mu$ and a sequence of real numbers $\sigma_n \to 0$.

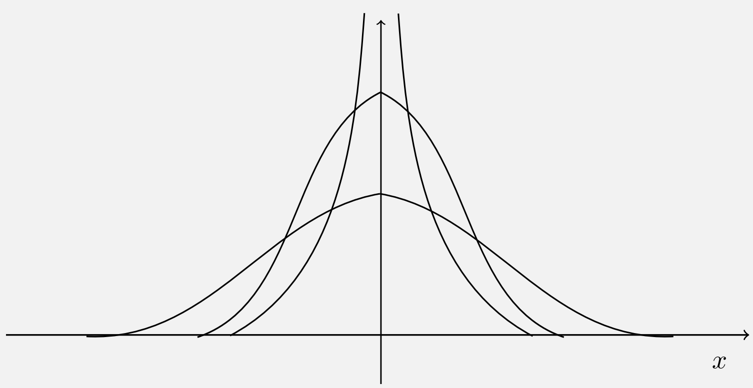

Figure 2: Convergence of the normal distribution to the Dirac delta function.

So we can expect that $\mathcal{N}_{\mu_n,\sigma_n} \xrightarrow{d} \delta_\mu$. As an exercise, prove this convergence (use a simple change of variables).

However, for the Gaussian case we can prove something stronger: if $X_n \sim \mathcal{N}_{\mu_n,\sigma_n}$ and $X \equiv \mu$ we can prove convergence in L^2. Using the triangle inequality, we can write

$$\mathbb{E}\big[|X_n - \mu|^2\big] \leq \mathbb{E}\big[|X_n - \mu_n|^2 + \underbrace{|\mu_n - \mu|^2}_{\to 0}\big],$$

and since $\mathbb{E}\big[|X_n - \mu_n|^2\big] = \mathbb{V}\big[X_n\big] = \sigma_n^2 \xrightarrow{n\to\infty} 0$, we also have L^2 convergence.

Exercise: prove that $\mathcal{N}_{\mu_n,\sigma_n} \xrightarrow{d} \delta_\mu$ if $\mu_n \to \mu$ and $\sigma_n \to 0$.

Proof.
Consider any test function $f \in \mathrm{C}_b(\mathbb{R})$, then we have that

$$\begin{aligned}
\int_{\mathbb{R}} f(x)\mathcal{N}_{\mu_n,\sigma_n}(dx) &= \int_{\mathbb{R}} f(x)\varphi\left(\frac{x-\mu_n}{\sigma_n}\right)dx && \text{(abs. continuity)} \\
&= \int_{\mathbb{R}} f\left(\sigma_n y + \mu_n\right)\varphi(y)dy && \text{(change of var.)}.
\end{aligned}$$

Since both f and φ are bounded the function $t \mapsto f(t)\varphi(t)$ is bounded by $g(t) = \max_{t'} f(t') \cdot \varphi(t)$, which is Lebesgue integrable and the dominated convergence theorem can be therefore applied to obtain the following equivalence

$$\lim_{n\to\infty} \int_{\mathbb{R}} f\left(\sigma_n y + \mu_n\right)\varphi(y)dy = \int_{\mathbb{R}} \lim_{n\to\infty} f(\sigma_n y + \mu_n)\varphi(y)dy = f(\mu)\int_{\mathbb{R}} \varphi(y)dy = f(\mu).$$

Therefore we have convergence in distribution to δ_μ by definition of the abstract integral wrt the Dirac measure.

$\square$

Def. (C.d.f. of a distribution)
Given a distribution μ on $\mathbb{R}$, the cdf of μ is the function $F_\mu : \mathbb{R} \longrightarrow [0,1]$ defined by

$$F_\mu(x) = \mu\big((-\infty, x]\big).$$

Remark

Among all known properties such as monotonicity, boundedness, etc, the most important for what follows is the property of *right-continuity*.

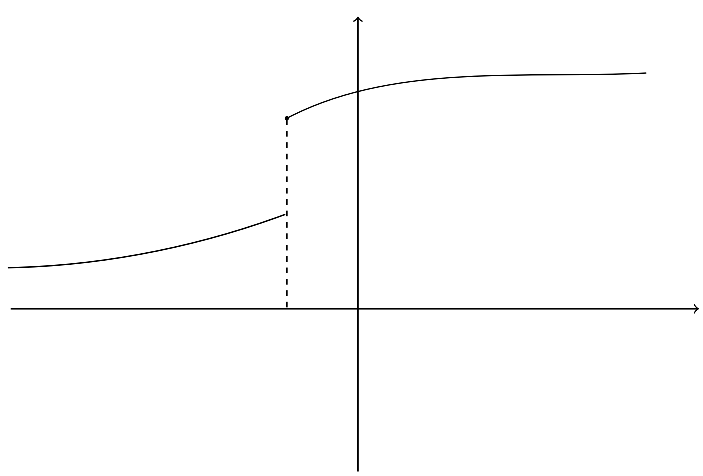

Figure 3: Right-continuity of the cumulative distribution function.

Def. (Cumulative distribution function)
Let X be a real-valued random variable, then the *cumulative distribution function* (CDF) of X is the function $F_X : \mathbb{R} \longrightarrow [0,1]$ defined by

$$F_X(x) = F_{\mu_X}(x) = \mathbb{P}\big(X \leq x\big)$$

Since the property of convergence in distribution is quite hard to prove for any bounded test function f, we want to characterize this property with respect to something else in order to make it easier to check it.

Example (Cdf of a uniform distribution)
Let $\mu_n = \mathrm{Unif}_{[0,\frac{1}{n}]}$, then the cdf is

$$F_n(x) = \begin{cases} 0 & \text{if } x < 0 \\ nx & \text{if } 0 < x < \frac{1}{n} \\ 1 & \text{if } x \geq \frac{1}{n} \end{cases}$$

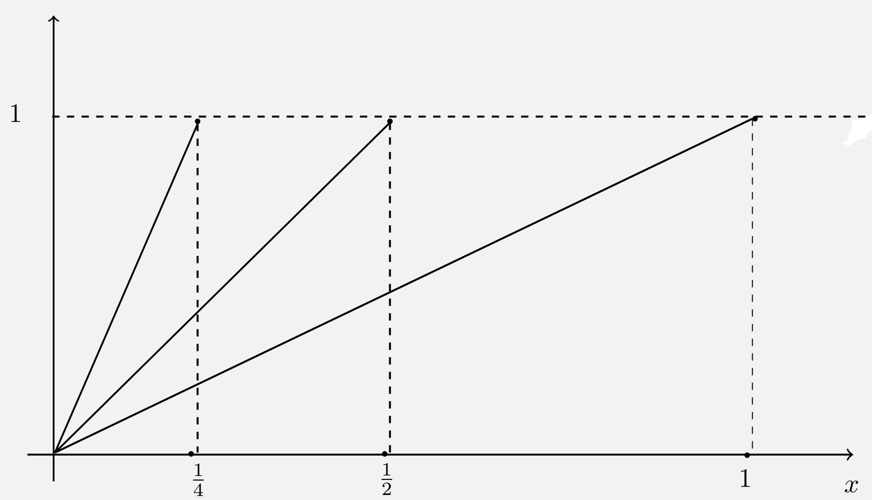

Figure 4: Convergence of the cdf of the uniform distribution to the unit step function.

The Dirac delta measure has a very simple cdf given by the unit step function,

$$F(x) = \mathbb{1}_{[0,\infty)}(x),$$

and in this example we have convergence of $F_n(x) \to F(x)$ in all points $x \in \mathbb{R}$ except for $x = 0$, since $F_n(0) = 0$ for all $n \in \mathbb{N}$.

Thm. 3 (Characterization of $\xrightarrow{d}$ using the cdf)
Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of distributions and μ be a distribution, then we have that

$$\mu_n \xrightarrow{d} \mu \iff F_{\mu_n}(x) \xrightarrow{n\to\infty} F_\mu(x),$$

for all x that are points of continuity of F_μ.

Proof.
No.

□

Remark

There can also be convergence in points of discontinuity, but it is not guaranteed in general.

Example (of convergence in the points of discontinuity)
$\mu_n = \delta_{-\frac{1}{n}}$, then it is clear that in this case also $\mu_n \to \delta_0$, and continuity is guaranteed for all points $x > 0$. However, in this case the cdf is such that

$$F_{\mu_n}(0) = F_{\delta_{-\frac{1}{n}}}(0) = 1 \quad \text{for all } n \in \mathbb{N},$$

therefore $\lim\limits_{n\to\infty} F_{\mu_n}(0) = 1$ and convergence is satisfied both in the points of continuity as well as in the point of discontinuity of F.

Let us now discuss another important function when dealing with real-valued random variables, which also allows a convenient characterization of $\xrightarrow{d}$.

Def. (Characteristic function of a distribution)
Let μ be a distribution, then we say that the *characteristic function* (CHF) of μ is the function $\varphi : \mathbb{R}^d \longrightarrow \mathbb{R}$ defined by

$$\varphi(\eta) = \int_{\mathbb{R}^d} e^{i\langle \eta, x\rangle} \mu(dx).$$

Def. (Characteristic function of a random variable)
Let X be a random variable with distribution μ on $\mathbb{R}^d$, then the *characteristic function of X* is the function $\varphi : \mathbb{R}^d \longrightarrow \mathbb{R}$ defined by

$$\varphi_X(\eta) = \varphi_{\mu_X}(\eta) = \mathbb{E}\big[e^{i\langle X, \eta\rangle}\big].$$

Remark

If $\mu \in \mathrm{AC}$ has density f, then we can write it exactly as a Lebesgue integral and it equals to a scaled and "slowed" version of the Fourier transform,

$$\varphi(\eta) = \int_{\mathbb{R}^d} e^{i\langle \eta, x\rangle} f(x)dx.$$

Thm. 4 (Lévy, characterization of $\xrightarrow{d}$ using the CHF)
Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of distributions and μ be a distribution, then

a) $\mu_n \xrightarrow{d} \mu \implies \varphi_n(\eta) \xrightarrow{n\to\infty} \varphi(\eta)$ *for any* $\eta \in \mathbb{R}^d$.

b) $\varphi \xrightarrow{n\to\infty} \varphi$ *everywhere, with* φ *continuous in* $\eta = 0$, *then* φ *is a CHF of a distribution* μ *and* $\mu_n \xrightarrow{d} \mu$.

Remarks

CHF's have some interesting properties, most notably

1. $\varphi(0) = 1$ since $\mathbb{E}\big[e^{i\langle 0,x\rangle}\big] = \mathbb{E}\big[1\big] = 1$.
2. φ_X is continuous in $\nu = 0$, which we can check by the limiting procedure

$$\lim_{\eta\to 0} \varphi_X(\eta) \overset{?}{=} \varphi_X(0) = 1.$$

Since $e^{i\vartheta} = \cos\vartheta + i\sin\vartheta$ is always equal in norm to 1 (Euler's formula), we can apply the dominated convergence theorem

$$\lim_{\eta\to 0} \mathbb{E}\big[e^{i\langle X,\eta\rangle}\big] \overset{\text{DCT}}{=} \mathbb{E}\big[\lim_{\eta\to 0} e^{i\langle X,\eta\rangle}\big] = \mathbb{E}\big[1\big] = 1.$$

1.2 Limit theorems

Notation: If $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables, we define the partial sums and partial means by

$$S_n = X_1 + X_2 + \ldots + X_n,$$
$$M_n = S_n/n.$$

Thm. 5 (Law of large numbers)
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables in $L^1(\Omega,\mathbb{P})$ that are i.i.d with mean $\mathbb{E}\big[X_n\big] = \mu$, then

- *(Weak L.L.N.)* $M_n \xrightarrow{d} \mu$
- *(Strong L.L.N.)* $M_n \xrightarrow{a.s.} \mu$

Proof.
We only prove the weak form since the strong one is very difficult. However, even for the weak form we would have to prove Lévy's theorem, which is also quite difficult.

□

Remark

Had we also assumed that $X_n \in L^2(\Omega, \mathbb{P})$ with $\mathbb{V}\big[X_n\big] = \sigma^2$, then this would've become a one-line proof since

$$\mathbb{P}\big(|M_n - \mu| > \varepsilon\big) \leq \frac{\mathbb{E}\big[\overbrace{|M_n - \mu|^2}^{L^2\text{converg.}}\big]}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n\to\infty} 0.$$

Using this, we have convergence in L^2 which implies $\xrightarrow{P}$ and $\xrightarrow{d}$. These inequalities are useful as a very basic estimate of the speed of convergence for Monte Carlo simulations and confidence regions, in order to provide error bounds. However, proper estimates are more refined and will be discussed later on.

References

Gut, A. (2009). *An Intermediate Course in Probability.* 2° edizione. Dordrecht : New York, NY: Springer Nature.