Mathematical Methods in Economics
Lecture Notes

Thomas Bourany*

The University of Chicago

Inspired by lecture notes by Kai Hao Yang and Yu-Ting Chiang

August 31, 2020

[*] *thomasbourany@uchicago.edu*, thomasbourany.github.io, I thank the previous lecturers of this math camp – Kai Hao Yang, Kai-Wei Hsu and Yu-Ting Chiang – for generously sharing their material. I am also grateful to my past professors of mathematics in Université Pierre and Marie Curie (UPMC) – Sorbonne in Paris who gave inspiration to part of these notes

1 Introduction and foreword

2 Prerequisites:

2.1 Basics of calculus and matrix algebra

2.1.1 Continuity, Derivative and basic multivariate calculus

2.1.2 Vector, matrices and inner-products

2.2 Basis of topology and analysis

2.2.1 Distance and metrics spaces

2.2.2 Norms and Normed vector space

2.2.3 Banach spaces and Hilbert spaces

2.2.4 Remark : Finite vs. Infinite dimension

2.3 Linear Algebra

2.3.1 Eigenvalues and Eigenvectors

2.3.2 Matrix decomposition

2.3.3 A brief excursion on linear operators

2.4 Calculus and Optimization

2.4.1 Properties of functions

2.4.2 Jacobian and differentiability

2.4.3 Existence and uniqueness of optimizers

We consider an optimization problem of the form $(\mathcal{P})$:

$$\inf_{x \in X} f(x)$$

where X is an abstract space, that we can consider to be $X = \mathbb{R}^n$ in the following.

Proposition 2.1.
If (X, d) is a compact metric space and f is continuous function, then :
there exists a maximum and a minimum. Said differently, f reaches its boundaries, i.e.

$$\exists\, x^\star \in X, \;\; \textit{such that } f(x^\star) = \inf_{x \in X} f(x) \quad \textit{or} \quad f(x^\star) = \sup_{x \in X} f(x)$$

In this case the infimum or supremum (that is a set with a unique element) is called minimum of maximum $f(x^\star) = \inf_{x \in X} f(x) = \min_{x \in X} f(x)$ and similarly for maximum.

Proposition 2.2.
If (X, d) is a compact metric space and f is lower semi continuous function, then :
there exists a minimum (i.e. the infimum is reached, i.e. $(\mathcal{P})$ has a solution)

$$\exists\, x^\star \in X, \;\; \textit{such that } f(x^\star) = \inf_{x \in X} f(x) = \min_{x \in X} f(x)$$

Theorem 2.1.
If (X, d) is a reflexive Banach space with an non-empty subset $Y \subset X$ and $Y \neq \emptyset$, and if

- *the function $f : Y \to \mathbb{R}$ is a convex and lower-semi continuous*
- *the set C is convex*
- *either C is bounded or f is coercive ($f(x) \to \infty$ when $||x|| \to \infty$)*

then, with these 5 conditions, there exists a minimum (i.e. the infimum is reached, i.e. $(\mathcal{P})$ has a solution) on the set C.

$$\exists\, x^\star \in C, \;\; \textit{such that } f(x^\star) = \inf_{x \in C} f(x) = \min_{x \in C} f(x)$$

Moreover, if the function is strictly convex, then the minimum is unique *Note:* This is a very important/strong theorem of optimization because the assumption are the weakest (compactness is usually really/too strong and replaced here by closed, convex, bounded set in a reflexive Banach space, very often met in practice).

2.4.4 Unconstrained optimization and first order condition

Definition 2.1.
Let $f : X \to \mathbb{R}$ be a function, f is differentiable in $x \in X$ if there exists a linear continuous map $DJ(x) \in \mathcal{L}(X, \mathbb{R})$ such that

$$\lim_{||h|| \to 0} \frac{|f(x+h) - f(x) - Df(x) \cdot h|}{||h||} = 0$$

when $DJ(x)$ exists it is unique, and we call it differential or Frechet differential
Note:

- $f : X \to \mathbb{R}$ and if f is derivable (standard case) then it is differentiable and $Df(x) \cdot h = f'(x)h, \forall h \in \mathbb{R}$.

- In the first-order Taylor expansion in the point x_0, we write $f(x) = f(x_0) + Df(x_0) \cdot (x - x_0) + o(||x - x_0||)$, when $o(h)$ is the Laudau's o notation as : $\lim_{h \to 0} \frac{o(h)}{h} = 0$

Theorem 2.2.
Let $(X, ||\cdot||)$ be a normed vector space, and $\mathcal{O}$ an open set of X and $f : \mathcal{O} \to \mathbb{R}$ a differentiable function, then,

$$\begin{aligned} &\textit{If} \qquad x^\star \in X \quad \textit{such that} \quad f(x_0) = \min_{x \in \mathcal{O}} f(x) \\ \textit{Then we have} \quad & Df(x^\star) = 0 \end{aligned}$$

This first-order condition is a necessary condition (i.e. a consequence) for optimality.

Note: It is *not sufficient* (yet), since even if $x^\star$ respects the FOC, it can be max or saddle point.

Theorem 2.3.
Let $(X, ||\cdot||)$ be a normed vector space, and $\mathcal{C}$ an open set of X and $f : \mathcal{C} \to \mathbb{R}$ a differentiable function. If f is convex, then the FOC is also sufficient, i.e.,

$$\begin{aligned} &\textit{If} \qquad Df(x^\star) = 0 \qquad \textit{or} \qquad Df(x^\star) \cdot (x - x_0) \geq 0 \qquad \forall x \in \mathcal{C} \\ \textit{Then we have} \qquad & x^\star \in X \quad \textit{such that} \quad f(x^\star) = \min_{x \in \mathcal{C}} f(x) \end{aligned}$$

2.4.5 Convex duality

(...)

2.4.6 Constrained optimization and Kuhn-Tucker theorem

Equality constraints

Now, let us suppose that the set $\mathcal{C}$ in theorem 2.3 is defined by a equality constraint function $\mathcal{C} = \{x \in X, \text{s.t. } g(x) = 0\}$. As a result the problem $\mathcal{P}$ becomes :

$$\inf_{x \in \mathcal{C}} f(x) = \inf_{\substack{s.t. \\ g(x)=0}} f(x)$$

Theorem 2.4 (Necessity).
Let $(X, || \cdot ||)$ be a normed vector space, and f and g, $f : X \to \mathbb{R}$, $g : X \to \mathbb{R}$, two functions which are both continuous and with continuous derivative (i.e. $f, g \in \mathcal{C}^1$), if, $x^\star \in X$ such that

$$f(x_0) = \min_{x \in \mathcal{C}} f(x) = \min_{s.t.\ g(x)=0} f(x)$$

(and also $Df(x^\star) \neq 0$) then there exists a Lagrange multiplier $\lambda \in \mathbb{R}$, such that :

$$Df(x^\star) = \lambda Dg(x^\star) \tag{1}$$

Notes:

- This is a necessary condition. Again, the FOC is not sufficient for determining optimality.
- This optimality condition generalizes when there are M constraints, if $(Dg_1, \ldots, Dg_M)$ are linearly independent.
- The value $\lambda \in \mathbb{R}$ is the shadow value of the constraint $g(x) = 0$: when relaxing the constraint, we can have $\widetilde{x} = x^\star + \varepsilon$, with the two first-order approximations :

$$\begin{cases} g(\widetilde{x}) & \approx g(x^\star) + Dg(x^\star) \cdot \varepsilon \\ f(\widetilde{x}) & \approx g(x^\star) + Df(x^\star) \cdot \varepsilon \end{cases}$$

  what would the marginal change of f for this change of x? It would be:

$$\frac{\frac{f(\widetilde{x}) - f(x^\star)}{\varepsilon}}{\frac{g(\widetilde{x}) - g(x^\star)}{\varepsilon}} \approx \frac{Df(x^\star)}{Dg(x^\star)} = \lambda$$

- If you define the "Lagrangian" function:

$$\mathcal{L}(x, \lambda) = f(x) + \lambda g(x)$$

  one can show that the first order condition is equivalent to find the saddle point of the Lagrangian function
- Here, the sign of the Lagrange multiplier doesn't matter: λ would be strictly positive if the unconstrained problem would make $g(x) > 0$ and conversely $\lambda < 0$ if the unconstrained problem makes $g(x) < 0$. The sign of the constraint will matter in th KKT theorem.

Theorem 2.5 (Sufficiency).

Given the assumptions of the previous theorem, if in addition we assume that f and g are convex, *then the optimality conditions are also* sufficient

Inequality constraints and KKT

Now, let us suppose that constraints are multiple inequality functions $\mathcal{C} = \{x \in X, \text{s.t. } g_1(x), \dots, g_M \leq 0, \}$. As a result the problem $\mathcal{P}$ becomes :

$$\inf_{x \in \mathcal{C}} f(x) = \inf_{\substack{s.t.\ \forall i=1\dots M \\ g_i(x)=0}} f(x)$$

Theorem 2.6 (Karush-Kuhn-Tucker, Necessity).

Let $(X, ||\cdot||)$ be a normed vector space, and f and multiple constraint g_i, $f : X \to \mathbb{R}$, $g_i : X \to \mathbb{R}$, $\forall i = 1, \dots, M$, functions which are both continuous and with continuous derivative. We introduce the "Lagrangian" $\mathcal{L}$ function associated to this problem:

$$\mathcal{L}(x, \lambda_1, \dots, \lambda_M) = f(x) + \sum_{i=1}^{M} \lambda_i \, g_i(x) \qquad \forall (x, \lambda_i) \in X \times \mathbb{R}_+ \quad \forall 1 \leq i \leq M$$

The optimality condition of the solution x^* is a saddle point of this Lagrangian function, under the condition that the constraints are "qualified"[1]*: Under all the previous hypothesis, the four following conditions are* necessary *for optimality. Formally, if x^* is a global minimum, then the four conditions are satisfied:*

1. *Stationarity:*
$$Df(x^*) + \sum_{i=1}^{M} \lambda_i \, Dg_i(x^*) = 0$$
(Equivalent to the "saddle point conditions" on the Lagrangian: $\frac{\partial \mathcal{L}}{\partial x}(x, \lambda) = 0$, $\frac{\partial \mathcal{L}}{\partial \lambda_i}(x, \lambda) = 0 \quad \forall\, 1 \leq i \leq M$)

2. *Primal feasibility (simply, constraints should be satisfied):* $\quad g_i(x^*) \leq 0 \quad$ *for* $\quad 1 \leq i \leq M$

3. *Dual feasibility:* $\qquad \lambda_i \geq 0 \qquad \forall 1 \leq i \leq M$

4. *Complementarity:* $\qquad \sum_{i=1}^{M} \lambda_i \, g_i(x^*) = 0$
(If the constraint is binding at optimum (i.e. $g(x^*) = 0$) then the Lagrange multiplier is strictly positive (again, it stands for the "shadow value" of relaxing the constraint) and conversely)

[1] The constraints are "qualified" when $\forall 1 \leq i \leq M$*, the derivative of the constraint function $F_i'(u^*)$ should be negative (or equal to zero if F_i are affine). These conditions are sometimes called "Slater condition" in case of convex constraint functions, and "Mangasarian-Fromovitz constraint qualification" in the general case (where there are also equality constraint, which is not the case here). The main idea of qualification (very important for the proof of the "necessary condition" of KKT theorem) is that you can look in the neighborhood of the local minimum to find the optimality condition (after some "linearization" along the lines defined by gradients).*

Theorem 2.7 (Karush-Kuhn-Tucker, sufficiency).
Given the assumptions of the previous theorem, and under the additional assumption that the objective function f and the constraints $g_1, \ldots, g_M$ are convex, *then these four conditions are also* sufficient.
Said differently, if x^* satisfy the four conditions, then x^* is global minimum.

Note:

- Again, be careful to check for convexity when using if for sufficiency! (something economists rarely do!)
- Similarly as above, the Lagrange multiplier is the shadow value of relaxing constraint, for example λ is the "marginal value of income", when the constraint g is a budget constraint.
- However, this time the Lagrange multiplier has a positive sign, because the inequality constraint is directional (on one side of the constraint it binds, but not on the other)

2.4.7 Numerical optimization methods

Gradient descent, Newton methods, Solution of linear and non-linear system of equation

3 Probability theory

Probability is about "measuring" the frequency of events happening. Since its mathematical formalization in 1933 by A. Kolmogorov, it has borrowed a lot from measure theory, introduced as a theory of integration by H. Lebesgue in 1904. Sadly, it is really abstract as a first exposure to probability, but I will try to use only the most important properties in the probability theory setting.

3.1 Foreword: from measure theory to probability theory

In a nutshell, Lebesgue theory of integration was groundbreaking because it was able to prove properties of integrals without requiring any conditions on the function (or very mild condition: the function just needs to be "measurable", which happens very (very) often!).
To reach this, it required to define function $f : X \to \mathbb{R}$ is a new way: we don't need to consider all the point of the set/space $x \in X$ but only "almost everywhere", i.e. everywhere except on a countable number of points. This countable set of points doesn't matter because it has "measure zero".

The measure (or distribution) μ is an extension of measuring interval sizes. For example, on the following picture, the two functions are equal almost everywhere and hence the integral (with respect to a measure μ) of $f(x)$ on $[a, b]$ are the same on the LHS and RHS, even after changing 3 points of the function: this is because the "measure" μ of these 3 points is zero.

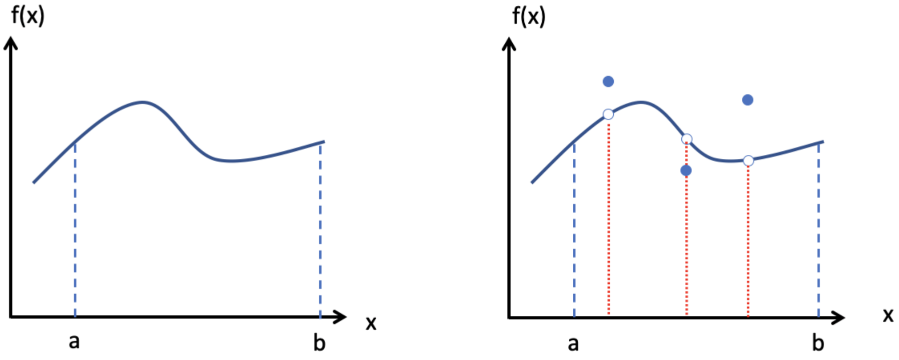

Figure 1: **Two functions equal almost everywhere**

The benefit of this is a great flexibility for integration. Despite a long (and a bit tedious) construction of the Lebesgue integral, the main difference with the the Riemann integral is displayed in the following picture: the subdivisions in the Lebesgue integral are made with respect to the function on the $y-$axis (instead of the x-axis for the Riemann integral).

The main results of this procedure are the convergence theorems: Monotone convergence theorem, Fatou's lemma and Dominated convergence theorem (more on that below and in A. Shaikh's class). The main advantage of these theorems is to switch the limit and integral signs, and thus to eliminate many of the pathological cases when a limit of integrable function

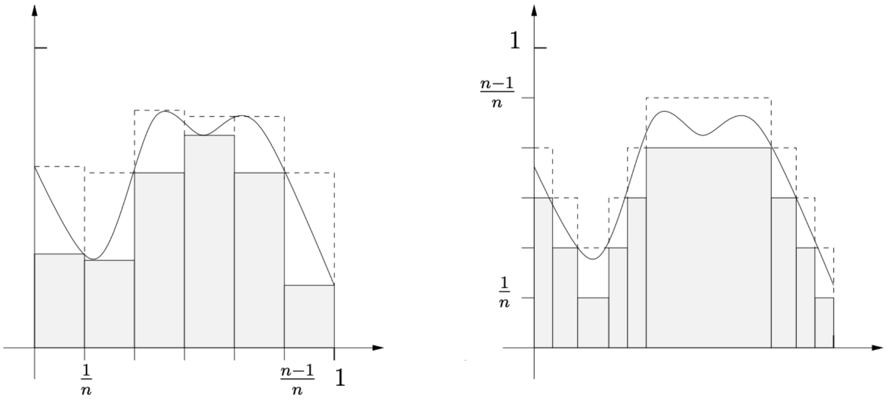

Figure 2: **Difference between the constructions of Riemann (LHS) and Lebesgue (RHS) integrals**

f_n isn't Riemann-integrable (but is very well Lebesgue integrable thanks to these convergence theorems.).

For $f(x) = \lim_{n\to\infty} f_n(x)$, converging pointwisely in $x \in X$ (almost everywhere), under conditions on monotonicity of positive function f_n ($0 \leq f_n \leq f_{n+1}$) or domination of integrable functions $|f_n| \leq |g|$, $\forall n$, then we have:

$$\int_X f(x)d\mu(x) = \lim_{n\to\infty} \int_X f_n(x)d\mu(x)$$

where the formalism of this integral will be make clear below. This is in a couples of lines the main gist of measure theory.

Kolmogorov used this formalism for probability. Considering a space of "states-of-the-world" $\omega \in \Omega$, random variables are functions $X : \Omega \to \mathbb{R}$, $X(\omega) \in \mathbb{R}$ that are defined almost-everywhere : in probability we call this "almost-surely". We consider distributions – or "laws" of probability $\mathbb{P}(\cdot)$ – as our "measures" of interest: for an event $A \subset \mathbb{R}$

$$P_X(A) = \mathbb{P}(X(\omega) \in A) := \mathbb{P}(\omega \in \Omega | X(\omega) \in A) = \int_\Omega \mathbb{1}\{\omega \in X^{-1}(A)\} d\mathbb{P}(\omega)$$

This definition will be made clear below! In particular, if two random variables X and Y have the same distribution $P_X(A) = P_Y(A)$ "almost surely" – that is "everywhere" expect on a set of probability (i.e. measure) null $\mathbb{P}(X \neq Y) = 0$ – we consider them to be the same (almost surely!). We can hence use all the artillerie of measure theory, in particular for convergence of sequences of functions. In probability, we will focus of convergence of random variables, which will be useful for proving the famous and important convergence theorems like Law of Large Numbers and Central limit theorem.

3.2 Basics: Random space, Random variables, Moments

Now, let us define many measure-theoric objects that appears in all concepts and properties of random variables and distributions.

Definition 3.1.
The couple $(\Omega, \mathscr{F}, \mathbb{P})$ *is a probability space, where* Ω *the sample space, i.e. set of all possible outcomes/"states-of-the-world", is attached to a collection* $\mathscr{F}$ *of sets (parts of* Ω*) – this* $\mathscr{F}$ *includes all the potential events – and a measure of probability* $\mathbb{P}$ *– over these sets* $A \in \mathscr{F}$*.*

Definition 3.2.
A σ*-algebra* $\mathscr{F}$ *over the set/space* Ω *is a family of sets, such that :*

(i) $\Omega \in \mathscr{F}$

(ii) If $A \in \mathscr{F}$ *Then* $\Rightarrow A^c \in \mathscr{F}$

(iii) $A_n \in \mathscr{F}, \forall n \Rightarrow \cup_{n\geq 1} A_n \in \mathscr{F}$

It is intuitively the set of all information available. If an event/outcome A *is not in* $\mathscr{F}$*, this means it can not happen.*

Example 3.1.
Consider the sample set of a dice with 3 outcomes $\{L, M, H\}$ *(or a financial that has low, median, and high values at a given date). The set* $\Omega^d = \{L, M, H\}$*. Hence, thanks to properties (i) and (ii), the* $\sigma-$*algebra generated by this set is*

$$\mathcal{F}^d = \big\{\emptyset, \{L\}, \{M, H\}, \{M\}, \{L, H\}, \{H\}, \{L, M\}, \{L, M, H\}\big\}$$

Example 3.2.
Consider the more abstract but ubiquitous example of Borel. The sample space is $\mathbb{R}$ *and we consider all the open intervals* $A = (a, b) \subseteq \mathbb{R}, \forall\, a, b \in \mathbb{R}$*. The Borel* $\sigma-$*algebra* $\mathscr{B}_{\mathbb{R}}$ *is defined as "the* $\sigma-$*algebra generated by the collection of open sets, i.e. the smallest* $\sigma-$*algebra associated to* $\mathbb{R}$ *that contains all the open sets. More precisely, this collection of sets contains all the open sets* A_i*, as well as their complement* A_i^c *and their countable union* $\cup_i A_i$*. This Borel measurable space* $(\mathbb{R}, \mathscr{B}_{\mathbb{R}})$ *with its Borel* $\sigma-$*algebra makes the bridge between standard real analysis and measure/probability theory.*

Definition 3.3.
A probability measure $\mathbb{P}$ *is a map* $\mathbb{P} : \Omega \to [0, \infty]$ *such that*

(i) $\mathbb{P}(\emptyset) = 0$

(ii) For all sequences of events $(A_n)_n$ *of measurable sets, which are* disjoints two-by-two *i.e.* $\mathbb{P}(A_i \cap A_j) = 0, \forall i, j$*, then we have* $\mathbb{P}(\cup_n A_n) = \sum_n \mathbb{P}(A_n)$*. This is called* σ*-additivity*

(iii) The measure is a finite measure with total mass 1: $\mathbb{P}(\Omega) = 1$*. This is specific to probability measure (but not general measure that can have infinite mass).*

Note: When we "associate" a sample space and σ-algebra with a measure, it implies that all the events have a probability $\mathbb{P}$, (i.e. you can "measure" how frequent the outcome will be). Moreover, the rules of σ-algebra imply that if you can measure $\mathbb{P}(A)$ or $\mathbb{P}(A_n)$, you can also measure $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ or $\mathbb{P}(\cup_n A_n)$ $(\leq \sum_n \mathbb{P}(A_n))$

Now the following concept is one of most important here:

Definition 3.4.

Consider two measurable spaces $(\Omega, \mathscr{F})$ and $(E, \mathscr{E})$.

- *A function or application $f : \Omega \to E$ is measurable if*

$$\forall B \in \mathscr{E},\ \exists\, A = f^{-1}(B) \in \mathscr{F}$$

- *A random variable $X : \Omega \to E$ is a measurable* function *from the set of possible outcomes Ω to a set $(E, \mathscr{E})$*

More intuitively, a random variable is a measurable function because each value/outcome of the random variable is associated with an event included in $\mathcal{F}$. If an outcome $\widetilde{B}$ is not *associated with an event (i.e. $\nexists \widetilde{A} = X^{-1}(\widetilde{B})$ in the $\sigma-$algebra, then you don't know what can happen/what has happened. Because of that, in particular, you can't compute the probabilities of the events of the random variables to have happened.*

Example 3.3.

Reconsider the example of the dice: $\Omega = \{L, M, H\}$ and $(\Omega, \mathscr{F}_\Omega)$ and a random variable X_1 such that $X_1(L) = -1, X_1(M) = 0, X_1(H) = +1$.

Now consider the second case where you have two such dices thrown simultaneously (and independently): $\widetilde{\Omega} = \{LL, LM, LH, ML, MM, MH, HL, HM, HH\}$. We have the measurable space $(\widetilde{\Omega}, \mathscr{F}_{\widetilde{\Omega}})$ associated with this and we consider a second random variable $X_2 = \frac{X_1 + X_{1'}}{2}$ (hence $X_2(MH) = \frac{0+1}{2} = 0.5$ and $X_2(LL) = -1$ for example). In the following picture, we have that the random variable X_1 is measurable on the LHS for the space $(\Omega, \mathscr{F}_\Omega)$, but X_2 is not *measurable on the RHS on the same space $(\Omega, \mathscr{F}_\Omega)$ (but it is for $(\widetilde{\Omega}, \mathscr{F}_{\widetilde{\Omega}})$!).*

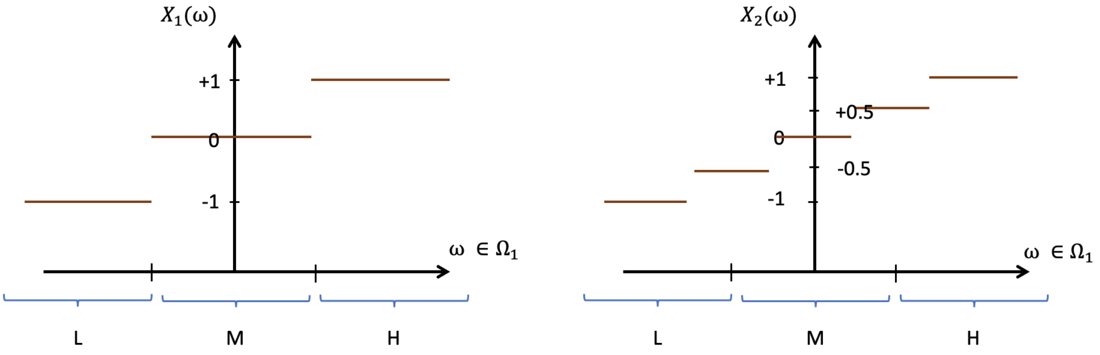

Figure 3: **Measurability (or not!) of the random variable X_1 and X_2 w.r.t.** $(\Omega, \mathscr{F}_\Omega)$

Note: In practice, we do not focus so much on Ω (except for some definition of stochastic processes, c.f. comment below and in L. Hansen's lectures on this topic).

An adjacent concept (a bit hinted in the example above) is the $\sigma-$algebra generated by a random variable, as we see in the next definition:

Definition 3.5.
Let $X : \Omega \to E$ be a random variable with values in a measurable space $(E, \mathscr{E})$. The $\sigma-$algebra generated by X, denoted $\sigma(X)$ is defined as the smallest $\sigma-$algebra on Ω that makes X measurable (on Ω), i.e.

$$\sigma(X) := \{A := X^{-1}(B), B \in \mathscr{E}\}$$

Note: More generally, let $(X_i, i \in I)$ any family (or sequence) or random variables, X_i with values in $(E_i, \mathscr{E}_i)$ then

$$\sigma\left(X_i, i \in I\right) := \sigma\left(X_i^{-1}\left(B_i\right) : B_i \in \mathscr{E}_i, i \in I\right)$$

Proposition 3.1.
Let X a random variable with values in $(E, \mathscr{E})$. Let Y a real random variable. Then Y is $\sigma(X)$-mesurable if and only if $Y = f(X)$ for a measurable function (i.e. deterministic function) $f : E \to \mathbb{R}$.
Note: This implies that any Y that includes the same "relevant information" as X, i.e. Y is measurable w.r.t. $\sigma(X)$, (but not more information than that!!), implies that there is a deterministic mapping between X and Y (indeed the necessity Y is $\sigma(X)$-mesurable $\Rightarrow$ $Y = f(X)$ is the one trickier to prove).

Now, let us consider measure of probability for random variables:

Definition 3.6.
Let $(\Omega, \mathscr{F})$ be a measurable space and a random variable $X : \Omega \to E$ (where E can be $\mathbb{R}$, in the case of "real random variables" (r.r.v) for example). We call law (or distribution) of the random variable X the measure P_X given, for all event $A \in \mathscr{F}$, by:

$$P_X(B) = \mathbb{P}(X \in B) = \mathbb{P}\big(\omega \in \Omega \ s.t.\ X(\omega) \in B\big) = \mathbb{P}\big(\omega \in A \ with \ \ A = X^{-1}(B)\big), \ \forall A \in \mathscr{F}$$

Note: The measure P_X is the "image measure" of $\mathbb{P}$ via the application X. For real random variable, it is quite common to consider the Borel measurable space $(\mathbb{R}, \mathscr{B})$, with a standard measure (called "Lebesgue measure" λ) to

From this law, if the random variable is real (maps into $\mathbb{R}$), we can compute the usual things, like the expectation of this random variable, i.e. integral of the function with respect its probability measure.

Expectation and integration and related concepts

Definition 3.7.
Let $(\Omega, \mathscr{F})$ be a measurable space and a random variable $X : \Omega \to E$ (where E can be $\mathbb{R}$, we define the mathematical expectation as :

$$\mathbb{E}(X) := \int_{\Omega} X(\omega)\mathbb{P}(\mathrm{d}\omega)$$

The condition for this expectation to be appropriately defined is to assume that $\mathbb{E}(|X|) < \infty$, where $\mathbb{E}(|X|)$ is defined in the same way. This condition is called "integrability" of the random-variable /function $X : \Omega \to E$, or in other words, we say that X admit a first moment.
Note: [-7mm]

- We can extend this definition to the case of random vectors $X := (X_1, \cdots, X_d)$, which is "simply" a random variable with values in $\mathbb{R}^d$,, by taking $\mathbb{E}(X) := (\mathbb{E}(X_1), \cdots, \mathbb{E}(X_d))$ (provided that all X_i admit a first moment, for the expectations $\mathbb{E}(X_i)$ to be well defined.
- All the usual result on integrals, like homogeneity, linearity, monotonicity are valid for the Lebesgue integral as much as for the usual (Riemann) integral.

Theorem 3.1 (Transfer theorem (?))**.**
Let X be a random variable in $(E, \mathscr{E})$. Then P_X the probability law of X is the unique measure on $(E, \mathscr{E})$ such that

$$\mathbb{E}[f(X)] = \int_E f(x) P_X(\mathrm{d}x)$$

for every measurable (i.e. deterministic) function $f : E \to \mathbb{R}_+$
Note: As a result of this theorem, the expectation of a real random variable write:

$$\mathbb{E}(X) = \int_{\Omega} X(\omega)\mathbb{P}(d\omega) = \int_{\mathbb{R}} x \, P_X(dx)$$

Definition 3.8 (Notation)**.**
Many mathematician and economists are a bit handwavy on the notation of measures. Usually for an abstract measure μ on the set E, they define integral the following way:

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

Similarly in probability, for a measure of probability, we have interchangeably

$$\begin{aligned}\mathbb{E}[X] = \int_{\Omega} X(\omega)\mathbb{P}(d\omega) &= \int_{\Omega} X d\mathbb{P} \\ &= \int_E x P_X(dx) = \int_E x dP_X = \int_E x dF \\ &= \int_{\mathbb{R}} x \, f(x) \, dx\end{aligned}$$

where the 2nd line holds because of Transfer's theorem, and $X \sim F$ where $F(x)$ is the c.d.f of X and the last line holding only *if the r.v. X has a p.d.f. $f(x)$ more on that below).*

Definition 3.9.
In a general way, we can define higher order moment! Let $(\Omega, \mathscr{F})$ be a measurable space and a random variable $X : \Omega \to E$, the $n-th$ order moment is defined as

$$\mathbb{E}[X^n] = \int_E x^n P_X(\mathrm{d}x)$$

and the standard variance, skewness and kurtosis defined in the first equalities (the second equality being results one can prove easily as an exercise), given that $\mathbb{E}(X) = \mu < \infty$

$$\mathbb{V}ar(X) := \mathbb{E}\Big[\big(X-\mu\big)^2\Big] = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

$$Skew(X) := \mathbb{E}\Big[\big(\frac{X-\mu}{\sigma}\big)^3\Big] = \frac{\mathbb{E}[(X-\mu)^3]}{\mathbb{E}[(X-\mu)^2]^{3/2}} = \frac{\mu_3}{\sigma^3}$$

$$Kurt(X) := \mathbb{E}\Big[\big(\frac{X-\mu}{\sigma}\big)^4\Big] = \frac{\mathbb{E}[(X-\mu)^4]}{\mathbb{E}[(X-\mu)^2]^2} = \frac{\mu_4}{\sigma^4}$$

Proposition 3.2.
We have that the Existence of higher moments imply existence of lower moments. Let X be a random variable. Then,

$$\mathbb{E}\big[|X|^k\big] < \infty \qquad \Rightarrow \mathbb{E}[|X_n|^j] < \infty, \forall k \geq j \geq 1$$

Note: This can be proved easily with Hölder inequality, with one of the two functions being $= 1$ a.e. and because the total mass for measures of probability is one. This can also be proved using Jensen's inequality (see below)

Comprehension questions:

A couple of (hopefully easy) questions to see if you understood the material above:

- What is the $\sigma-$algebra associated with two coins tossed sequentially? Is the random variable $H_n = \{\text{number of heads}\}$ measurable with respect to the entire sample set? Is it measurable with respect to the $\sigma-$algebra generated by the random variable $T_1 = \{\text{the first toss is a tail}\}$
- What is the law, i.e. probability measure associated with the random variable H_n.
- Consider a random variable following a standard Normal distribution $X \sim \mathcal{N}(0, 1)$. What could be a $\sigma-$algebra associated with this random variable.
- Consider the Lebesgue measure that $\lambda(dx) = dx$ (the usual thing for 101-integration!) what is the image measure of the Lebesgue measure w.r.t. the Normal distribution $X \sim \mathcal{N}(0, 1)$. Is that a measure of probability?
- Consider a Poisson distribution Y (check wikipedia if needed :p), what is the image measure of the Lebesgue measure, w.r.t Y
- Consider the same Normal distribution X, and $Y = X^2$. Can we say that X is measurable with respect to the σ-algebra generated by Y? If yes why? If not why not?
- Consider a sequence of random variable $X_1, \dots, X_n$ i.i.d. Is X_n measurable w.r.t. $\sigma(X_1)$? And what about $\sigma(X_1, \dots, X_{n-1})$? And what about $\sigma(X_1, \dots, X_n)$?

3.3 Convergence theorems

In the following we will consider $(X_n)_{n\leq 0}$ a sequence of random variables – i.e., and we will need to analyze the convergence toward a limit. The question of the nature of convergence is at the heart of statistics (to attest the quality of estimators and C.I. as covered extensively by A. Shaikh in Metrics 1). There exists 4 main modes of convergences:

- Convergence "Almost-surely" ("the probability of converging is one")
- Convergence in mean (or L^p) ("the difference fades out in norm L^p/moment of order p")
- Convergence in probability ("the probability of diverging tends towards zero")
- Convergence in distribution ("the law/c.d.f. tends towards another law/c.d.f.)

We will cover them in turn, but beforehand, we will makes sense of the main theorem of convergence of sequence of functions that one encounter in measure theory, as explained in the foreword of section 4.1. above.

From the construction of Lebesgue integral to convergence theorems

Definition 3.10 (Terminology).
We say that a property is true "almost-surely" (or a.s.) or $\mathbb{P}-$almost everywhere, (or $\mathbb{P}-$a.e.), if it is valid $\forall\omega\in\Omega$ except for a set of null probability. *Note:* For example the two random variables X and Y are equal almost surely (or simply $X=Y$, *a.s.*) if $\mathbb{P}\big(\omega\, s.t.\ X(\omega)\neq Y(\omega)\big) = 0$

(i) Given a measurable space $(\Omega,\mathscr{F},\mathbb{P})$ and a random variable (function) $X:\Omega\to E$, Lebesgue's integral was build by considering positive "step function" (or "simple functions"), i.e. that can be written as :

$$X(\omega)=\sum_{i=1}^{n}\alpha_i\mathbb{1}\{\omega\in A_i\}\qquad \omega\in\Omega$$

where $\alpha_i<\alpha_{i+1},\forall i$ and $A_i=X^{-1}(\{\alpha_i\})\in\mathscr{F}$, and hence the integral can be easily written as :

$$\int_\Omega X(\omega)\mathbb{P}(d\omega):=\sum_{i=1}^{n}\alpha_i\mathbb{P}(A_i)\ \in[0,\infty]$$

(ii) The second stage was to extend this to *positive* functions that have a step-functions as their lower bound, and the integral is defined as the supremum over all potential step-functions that bound it below:

$$\int_\Omega X(\omega)\mathbb{P}(d\omega)=\sup\Big\{\int_\Omega\widetilde{X}(\omega)\mathbb{P}(d\omega)\ \&\ \widetilde{X}\leq X,\ \&\ \widetilde{X}\text{ step-function (r.v)}\Big\}$$

That is where the important theorem of monotone convergence appears:

Theorem 3.2 (Monotone convergence theorem of Beppo-Levi).
Let $\{X_n\}_n$ a sequence of positive *and* increasing *random variables, i.e. such that $X_n(\omega) \leq X_{n+1}(\omega)$ and let X its almost-sure pointwise limit, i.e. for almost all points $\omega \in \Omega$ (every ω except a set with null probability) such that :*

$$X(\omega) = \lim_{n\to\infty} \uparrow X_n(\omega)$$

Then we have the integral of the limit as the limit of the integral:

$$\int_\Omega X(\omega)\mathbb{P}(d\omega) = \lim_{n\to\infty} \int_\Omega X_n(\omega)\mathbb{P}(d\omega)$$

Corollary 3.1.
A consequence is to be able to switch integral and sum sign (since a sum can always be written as a particular sequence $Y_n = \sum_{i=1}^n X_i$) for positive *(!) random variables.*

$$\mathbb{E}\Big[\sum_i X_i\Big] = \int_\Omega \sum_i X_i(\omega)\mathbb{P}(d\omega) = \sum_i \int_\Omega X_i(\omega)\mathbb{P}(d\omega) = \sum_i \mathbb{E}\Big[\sum_i X_i\Big]$$

Corollary 3.2.
Another consequence, very obvious but used a lot *in economics, is the following, for every* positive *random variable.*

- $\int_\Omega X(\omega)\mathbb{P}(d\omega) < \infty \;\Rightarrow\; X < \infty$ *almost surely*
- $\int_\Omega X(\omega)\mathbb{P}(d\omega) = 0 \;\Rightarrow\; X = 0$ *almost surely*

Note: The proof of the first property requires the Markov inequality (a *must* to know if you ever do statistics! even if unrelated with convergence theorems).

Proposition 3.3 (Markov-Chebyshev's inequality).
Let $X : \Omega \to \mathbb{R}_+$ a positive random variable. Then, for any constant $c > 0$:

$$\mathbb{P}\big(\{\omega \in \Omega : X(\omega) \geq c\}\big) = \frac{\mathbb{E}[X]}{c}$$

The proof holds in one picture (easy to memorize as well).

The second big theorem of measure theory is the Fatou's lemma

Theorem 3.3.
Let X_n a sequence of positive *random variables, then*

$$\int_\Omega \Big(\liminf_{n\to\infty} X_n(\omega)\Big)\,\mathbb{P}(d\omega) \leq \liminf_{n\to\infty} \int_\Omega X_n(\omega)\mathbb{P}(d\omega)$$

Note: Again, Fatou's lemma is more a corollary (why?) of the monotone convergence theorem with clever use of definitions of limit inferior (check out this definition!). But it's used a lot

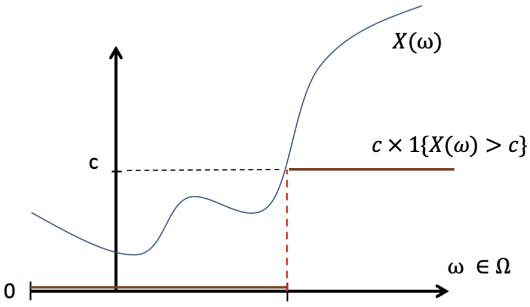

Figure 4: **Markov inequality in one picture**

in analysis and probability theory to provide upper bounds of integrals and estimators for examples.

(iii) We provided properties for positive function/random variables. The third stage is to extend that to function of both sign. In particular, we really want to avoid to end up with results of the type $\int f d\mu = +\infty - \infty =(?)$, giving indeterminacy. The important concept here is the one of integrability :

Definition 3.11.
Let $X : \Omega \to [-\infty, +\infty]$ a random variable (hence measurable). We say that X is integrable, or admit a first moment, w.r.t. $\mathbb{P}$, if

$$\mathbb{E}\big[|X|\big] = \int_\Omega |X| d\mathbb{P} < \infty$$

In this case, we define the integral of any random variable (not only positive!) by :

$$\int_\Omega X(\omega)\mathbb{P}(d\omega) = \int_\Omega X^+(\omega)\mathbb{P}(d\omega) - \int_\Omega X^-(\omega)\mathbb{P}(d\omega) \in \mathbb{R}$$

where $X^+ = \max\{X, 0\}$ and $X^- = -\min\{X, 0\} = \max\{-X, 0\}$ *Note:* We denote by $L^1(\Omega, \mathscr{F}, \mathbb{P})$ (or simply L^1 if there is no ambiguity) the space of all the random variable (or function) that are $\mathbb{P}$-integrable, i.e. that admit a first moment. Note that in this definition again use the fact that random variables are defined almost-everywhere/almost-surely.

Theorem 3.4 (Change of variable and integrability)**.**
Let $\Phi : (E, \mathscr{F}_E) \to (F, \mathscr{F}_F)$ a measurable (i.e. deterministic) function, P_Y is the image-measure of P_X w.r.t. Φ, in the sense that $\forall B \in \mathscr{F}_F$, $P_Y(B) = P_X(\Phi^{-1}(B)), \forall B$.
P_Y is also called pushforward measure of P_X by Φ and also denoted $P_Y = P_X \circ \Phi^{-1}$ or $P_Y = \Phi\sharp P_X$ (a bit as if we would define $Y = \Phi(X)$).
Now, for every measurable function $f : F \to [-\infty, \infty]$, we have

$$\int_E (f \circ \Phi) dP_X := \int_E f(\Phi(x))P_X(dx) = \int_F f(y)P_Y(dy) = \int_F f dP_Y$$

where the equality in the middle holds if one of the two integrals is well-defined (i.e. the function $f(X)$ is integrable, i.e. $f(X) \in L^1$). *Note:* That is quite an abstract definition of a change of variable with a measure-theory angle.

Now, we have covered enough definition to consider the most important theorem of measure theory and probability: the Lebesgue dominated convergence theorem.

Theorem 3.5 (Lebesgue's dominated convergence theorem).
Let $\{X_n\}_n$ a sequence of random variables in $L^1(\Omega, \mathscr{F}, \mathbb{P})$ (i.e. $\mathbb{E}[|X_n|] < \infty$, $\forall n$ and let X its limit for almost all points $\omega \in \Omega$ (i.e. every ω except those with null probability) such that :

$$X(\omega) = \lim_{n\to\infty} \uparrow X_n(\omega)$$

and if X_n is dominated – i.e. there exists an other integrable random variable $Y \in L^1$ such that almost surely we have $|X_n| \leq |Y|$, $\forall n \geq 0$. Then we have that X is integrable (i.e. $X \in L^1$) and the integral of the limit as the limit of the integral:

$$\int_\Omega X(\omega)\mathbb{P}(d\omega) = \lim_{n\to\infty} \int_\Omega X_n(\omega)\mathbb{P}(d\omega)$$

Note:

- In the proof, the slightly different propriety actually shown is the following:

$$\lim_{n\to\infty} \int_\Omega |X_n(\omega) - X(\omega)|\mathbb{P}(d\omega) = 0$$

  This is the definition of L^1-convergence, that we'll define below!

- The boundedness by an integrable random variable is important because there are a lot of case where the integral is finite for each n but the limit is not, as for these 3 types of examples where the functions/random variable is converging pointwisely to the vanishing function $X(\omega) = 0$ but its integral is not.

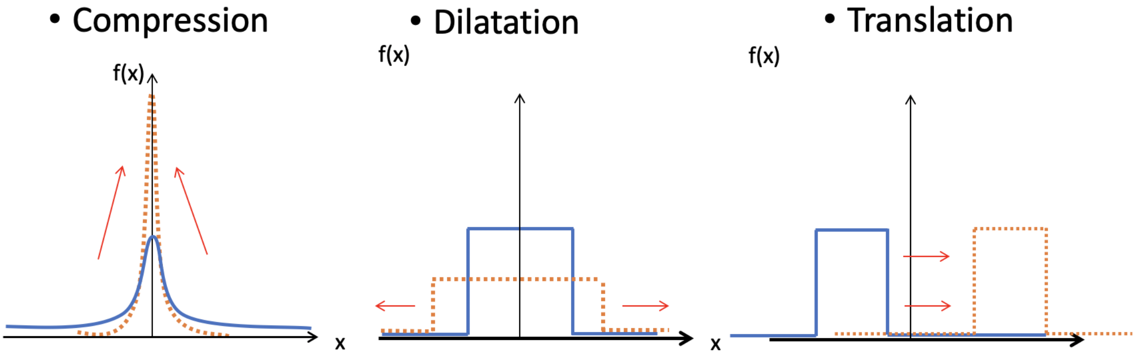

Figure 5: **Counterexamples to the dominated CV thm, because of lack of domination**

Convergence theorem for sequences of random variables

Definition 3.12.
A sequence of random variables $(X_n)_{n\geq 0}$ converges "Almost-surely" *toward X if there exists an event A with proba one ($\mathbb{P}(A) = 1$) where, $\forall\omega \in A, \lim_{n\to\infty} X_n(\omega) = X(\omega)$ Said differently,*

$$\mathbb{P}\Big(\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big) = 1$$

Note:

- Intuitively After some fluctuations of the sequence, we are (almost-) sure that X_n won't fall too far from X
- This type of convergence is the assumption we used in the condition of the Monotone convergence and the Dominated convergence theorem, where $X_n(\omega) \to_n X(\omega)$ almost-surely pointwisely.

Example 3.4.
Let X_n be a sequence of Normal random variable of law $\mathcal{N}(0,1)$. Let $S_n = X_1 + \cdots + X_n$, which then follow $S_n \sim \mathcal{N}(0,n)$. By Markov inequality, we have that, for all $\varepsilon > 0$:

$$\mathbb{P}(|S_n| > n\varepsilon) = \mathbb{P}(|S_n|^3 > n^3\varepsilon^3) \leq \frac{\mathbb{E}[|S_n|^3]}{\varepsilon^3 n^3} = \frac{\mathbb{E}[|S_n|^3]}{\varepsilon^2 n^{3/2}}$$

We have that $\sum_{k=1}^{\infty} \mathbb{P}(|S_n| > n\varepsilon) < \infty$. Thanks to this condition, we now use a theorem (not covered too much in this course) called Borel-Cantelli's theorem, that allow to claims that :

$$\mathbb{P}\Big(\limsup_{n\to\infty}\{|S_n| > n\varepsilon\}\Big) = 0$$

This last equality is the result of Borel Cantelli. This implies that, by definition of limits, we have that $\exists A \in \mathscr{F}$, with $\mathbb{P}(A) = 1$, such that

$$\forall\, \omega \in A,\ \ \exists\, n_0 = n_0(\omega,\varepsilon) < \infty,\ \ \textit{such that}\ |S_n| \leq n\varepsilon,\ \ \forall\, n \geq n_0$$

For all $\varepsilon > 0$, we have as a result:

$$\mathbb{P}\Big(\omega : \limsup_{n\to\infty} \frac{|S_n|}{n} \leq \varepsilon\Big) = 1$$

This implies that $\limsup_{n\to\infty} \frac{|S_n|}{n} = 0$, *a.s., and* $\lim_{n\to\infty} \frac{S_n}{n} = 0$

Let us take a little detour via Borel-Cantelli's lemma. Let us define the main object and state the result

Definition 3.13.

Let $(\Omega, \mathscr{F}, \mathbb{P})$ a measured space. Let $\{A_n\}_n$ a sequence of events and $B_n = \bigcup_{k\geq n} A_k$, is weakly decreasing. We define

$$A = \limsup_{n\to\infty} A_n := \bigcap_{n\geq 1} \bigcup_{k\geq n} A_k = \bigcap_{n\geq 1} B_n = \{\omega \text{ s.t. } \omega \in A_n \text{ for an infinity of } n)\}$$

All these terms are simply different notations for the same thing. A is also an event in $A \in \mathscr{F}$. This represents the set of events/states-of-the-world ω which belong to an infinity of events A_n. Also $\mathbb{1}_A(\omega) = \limsup_{n\to\infty} \mathbb{1}_{A_n}(\omega)$, justifying the notation. For these states-of-the-world, the events A_n occurs infinitely many times. Using the rules of complementarity, we also have:

$$\liminf_{n\to\infty} A_n = \bigcup_{n\geq 1} \bigcap_{k\geq n} A_k = \{\omega \text{ s.t. } \omega \in A_n \text{ for only finitely many} n)\}$$

Theorem 3.6 (Borel-Cantelli's lemma).

Let $\{A_n\}_{n\geqslant 1}$ a sequence of events (i) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then

$$\mathbb{P}\big(A_n \text{ infinitely many}\big) = 0$$

(ii) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, and if the events $\{A_n\}_{n\geqslant 1}$ are independant (i.e. $\forall\, n, A_1, \ldots, A_n$ are independent), then

$$\mathbb{P}(A_n \text{ infinitely many }) = 1$$

Note:

- *This theorem is quite abstract and i.m.o. not so useful for the core sequence in economics. But it is fundamental for probability theory and for almost sure convergence, including the proof of law of large number.*
- *In applications for almost-sure convergence, we often use the following version of part (i): there exists an event B with $\mathbb{P}(B) = 1$ (hence an almost sure event) such that for all $\omega \in B$ we can find $n_0 = n_0(\omega) < \infty$ such that $\omega \in A_n^c$ when $n \geqslant n_0$. Typically A_n could be an event of the type $A_n = \{|X_n - X| > \varepsilon\}$ to show the a.s. convergence of $X_n \to X$*

Definition 3.14 (Convergence in Probability).

A sequence of random variables $(X_n)_{n\geq 0}$ converges "in probability" toward X if, for all $\varepsilon > 0$

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

Note: Intuitively the probability that the sequence X_n falls far away from X is decreasing in n (but it can potentially be strictly positive)

Example 3.5.

Let X_n be a sequence of random variable, such that $\mathbb{E}[X_n] \to a \in \mathbb{R}$ and $\mathbb{V}ar(X_n) \to 0$, then again by Markov inequality

$$\mathbb{P}(|X_n - a| > \varepsilon) = \mathbb{P}(|X_n - a|^2 > \varepsilon^2) \leq \frac{\mathbb{E}[|X_n - a|^2]}{\varepsilon^2} = \frac{\mathbb{V}ar(X_n) + \left(|\mathbb{E}(X_n) - a|^2\right)}{\varepsilon^2} \to_{n\to\infty} 0$$

Hence X_n converges in probability to the constant a

Example 3.6 (Difference convergence *a.s.* and in probability)**.**

Consider exponential distribution, with intensity λ (recall, the higher the intensity the lowest the value of X, in expectation: $\mathbb{E}[X] = \frac{1}{\lambda}$).

First, consider $X_n \sim \mathscr{E}(\lambda = n)$. It is not difficult to show that

$$X_n \underset{p.s.}{\longrightarrow} X = 0$$

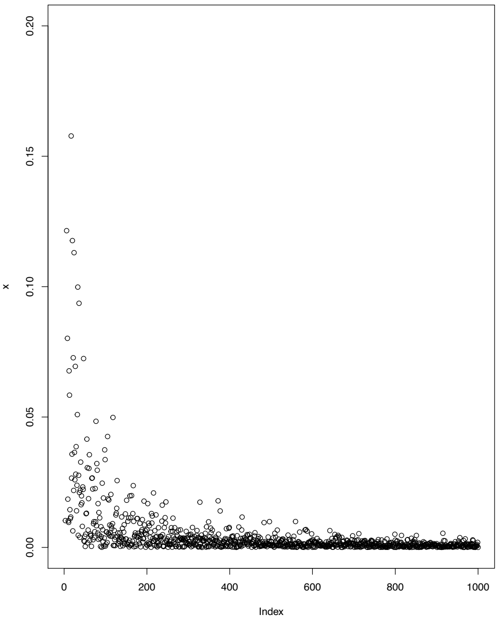

Figure 6: $X_n \sim \mathscr{E}(\lambda = n)$ **converges to** $X = 0, a.s.$

We see well that, for any given (fixed) ε, $\exists N \geq 1$ after which $\mathbb{P}(|X_n - 0| > \varepsilon) = 0$, $\forall\, n \geq N$, hence the sequence converges almost surely.

Second, consider $\widetilde{X}_n \sim \mathscr{E}(\lambda = \log(n))$, where the intensity diverges more slowly. Again, it is not really difficult to show that :

$$\widetilde{X}_n \underset{\mathbb{P}}{\to} \widetilde{X} = 0 \qquad \textit{and} \qquad \widetilde{X}_n \nrightarrow_{p.s.} 0$$

<u>*Note:*</u> All the usual results on limits, like unicity, monotonicity, linearity, homogeneity, are valid for the almost-sure and in-probability convergence.

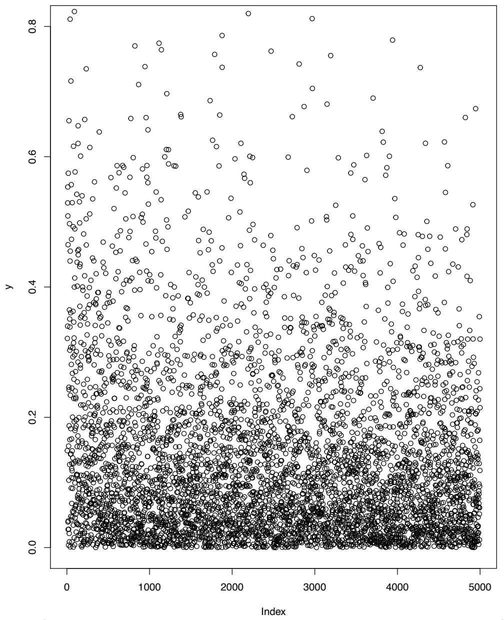

Figure 7: $\widetilde{X}_n \sim \mathscr{E}(\lambda = \log(n))$ **converges to** $\widetilde{X} = 0$**, in proba, but not almost surely**

The next theorem is showing the link between these two modes of convergences

Theorem 3.7 (CV *a.s.* $\Rightarrow$ CV in $\mathbb{P}$).

- *If* $X_n \xrightarrow[n\to\infty]{a.s.} X$ *almost surely, then the convergence also occurs in probability* $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$
- *If* $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ *in probability, then there is a subsequence* $X_{N(n)}$ *that converges almost surely* $X_n \xrightarrow[n\to\infty]{a.s.} X$.

Example 3.7 (CV in proba $\nRightarrow$ CV *a.s.*).
Consider the space $\Omega = [0, 1]$ *with* $\mathscr{B}_{[0,1]}$ *and the Lebesgue measure. Consider for all n,* k_n *is such that* $2^{k_n - 1} \leq n < 2^{k_n}$ *and consider the sequence* $X_n(\omega) = \mathbb{1}_{\left(\frac{n-1}{2^k}, \frac{n}{2^k}\right]}(\omega)$*, with the first few elements such that:*

$$\begin{array}{lll} X_1(\omega) := \mathbb{1}_{\left(0,\frac{1}{2}\right]}(\omega) & X_2(\omega) := \mathbb{1}_{\left(\frac{1}{2},1\right]}(\omega) & \\ X_3(\omega) := \mathbb{1}_{\left(\frac{1}{2},\frac{3}{4}\right]}(\omega) & X_4(\omega) := \mathbb{1}_{\left(\frac{3}{4},1\right]}(\omega) & \ldots \\ X_5(\omega) := \mathbb{1}_{\left(\frac{1}{2},\frac{5}{8}\right]}(\omega) & X_6(\omega) := \mathbb{1}_{\left(\frac{5}{8},\frac{6}{8}\right]}(\omega) & \ldots \end{array}$$

Then X_n *does not convergence almost surely (since for any* $\omega \in (0, 1]$ *and* $N \in \mathbb{N}$ *there exist* $m, n \geq N$ *such that* $X_n(\omega) = 1$ *and* $X_m(\omega) = 0$ *(we have that* $\limsup_n X_n = 1$*). On the other hand, since*

$$\mathbb{P}(|X_n| > 0) \to 0 \quad \text{as } n \to \infty$$

it follows easily that X_n *converges in probability to* 0

Moreover, it is easy to find a subsequence of X_n*, for example with* $N(n) = 2^n$*, which converges almost-surely.*

Definition 3.15 (Convergence in Norm L^p).
A sequence of random variables $(X_n)_{n\geq 0}$ converges "in mean p" *or in norm $L^p(\Omega, \mathcal{F}, \mathbb{P})$ toward X if*

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

Note:

- By Hölder inequality, if $X_n \to X$ in norm L^p, and if $q \in [1,p]$, then $X_n \to X$ in norm L^q as well. In other words, the higher p the stronger the convergence.
- If $X_n \to X$ in norm L^p, then $|X_n| \to |X|$ in L^p, since by reverse triangle inequality we have $\Big||X_n| - |X|\Big| \leq |X_n - X|$
- If $X_n \to X$ in norm L^p, then $\mathbb{E}[X_n^p] \to \mathbb{E}[X^p]$

Again a theorem following the link between these two modes of convergence.

Theorem 3.8 (CV $L^p \Rightarrow$ CV in $\mathbb{P}$). • *If $X_n \xrightarrow[n\to\infty]{L^p} X$ in norm L^p, then the convergence also occurs in probability $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$*

Note:

- The proof of the first point is simply to use the Markov inequality.
- The reciprocal is false as shown in the next example. However, it works in the case of dominated random variables, c.f. the next theorem.

Example 3.8 (CV in $\mathbb{P} \nRightarrow$ CV L^p).
Let $\{X_n\}_{n\geq 3}$ a sequence of real random variables, such that $\mathbb{P}(X_n = n) = \frac{1}{\ln n}$ and $\mathbb{P}(X_n = 0) = 1 - \frac{1}{\ln n}\cdot$ For all tout $\varepsilon > 0$, we have:

$$\mathbb{P}(|X_n| > \varepsilon) \leqslant\leq \frac{1}{\ln n} \to 0$$

therefore, on the one hand, $X_n \to 0$ in probability. On the other hand, we have:

$$\mathbb{E}(|X_n|^p) = \frac{n^p}{\ln n} \to \infty$$

So X_n doesn't converge toward 0 in L^p.

We already claimed with the dominated convergence theorem implies that CV *a.s.* implies CV in norm L^1. Actually we can weaken the assumption and start with a sequence of random variables converging in probability instead of almost-surely.

Theorem 3.9 (Weaker Dominated convergence theorem and Fatou's lemma). • *If $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ in probability, and if $\{X_n\}_n$ is dominated $|X_n| \leq Y$, $\forall\, n$ and Y is integrable $\mathbb{E}[Y] < \infty$, then $X_n \to X$ in L^p*

- *If $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ in probability, and if $X_n \geq 0, a.s.$, then$\mathbb{E}(X) \leq \liminf_{n\to\infty} \mathbb{E}(X_n)$*

Now that we have introduced all these definitions of convergence, we can finally state the most important theorem of this sections.

Theorem 3.10 (Law of Large Numbers).

Let $\{X_n\}_n$ a sequence of variable independent and identically distributed. If $\mathbb{E}(|X|) < \infty$, and if $\mathbb{E}(X) = \mu$, then:

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu$$

- *This convergence is **almost sure** (strong law of large numbers)*
- *This converges also **in probability** (weak law of large numbers)*

Note: The proof of this theorem is long and technical. However, by strengthening the assumption, with X_n admitting a 4^{th} order moment $\mathbb{E}(|X|^4) < \infty$, we can prove it easily with Markov inequality. First assume that $\mu = 0$ (if not, we can always define $\widetilde{X}_n = X_n - \mu$. As a result $\mathbb{E}[S_n] = 0, \forall n$. For all $\varepsilon > 0$

$$\mathbb{P}[|S_n| > \varepsilon] = \mathbb{P}[|S_n|^4 > \varepsilon^4] \leq \frac{\mathbb{E}[|S_n|^4]}{\varepsilon^4}$$

By tediously developing the sum $S_n^4 = \left(\sum_{k=1}^n X_k\right)$ and using the fact that the random variables are independent, such that $\mathbb{E}[X_n X_{n'}] = \mathbb{E}[X_n]\mathbb{E}[X_{n'}]$ and $\mathbb{E}[X_n] = 0$, all the terms at the first power drops out. We end up with

$$\begin{aligned}\mathbb{E}[S_n] &= \frac{1}{n^4}\Big[n\mathbb{E}[X_n^4] + 3n(n-1)\mathbb{E}[X_i^2 X_j^2]\\ &= \frac{\mu_4}{n^3} + \frac{3\sigma^4}{n^2}\end{aligned}$$

We now have the convergence of the series, making the use of Borel-Cantelli possible:

$$\mathbb{P}[\underbrace{|S_n| > \varepsilon}_{A_n^\varepsilon}] \leq \frac{1}{\varepsilon^4}\Big(\frac{\mu_4}{n^3} + \frac{3\sigma^4}{n^2}\Big) \qquad \sum_{n=1}^{\infty} \mathbb{P}(A_n^\varepsilon) < \infty$$

As a result, only finitely many A_n^ε occurs. We can find a threshold n_0 such that $\{\omega, s.t. |S_n| < \varepsilon\}$ is almost-sure $\forall\, n \geq n_0$, justifying the convergence almost-surely of the sequence.

You can find a lot of textbooks/on the web different version of the proof of the law of large number (with 2nd order moment, simpler, or only first moment, more difficult).

Convergence in distribution

This mode of convergence is slightly different than the 3 modes considered above. In convergence almost-sure, in probability or in norms L^p, we focused on the sequence of random variables, i.e. $\{X_n\}$, i.e. sequence of functions. In the convergence in distributions, we focus on the contrary on the convergence of a sequence of laws! (i.e. measures $\mu_{X_i} = P_{X_1}, P_{X_2} \cdots \to P_X$). This is much weaker!

Definition 3.16 (CV in distribution)**.**
A sequence of random variables $\{X_n\}_{n\geq 0}$ converges "in law or in distribution toward X if, for all continuous and bounded functions f

$$\lim_{n\to\infty} \mathbb{E}\big(\varphi(X_n)\big) = \mathbb{E}\big(\varphi(X)\big)$$

It is denoted $X_n \xrightarrow[\mathcal{D}]{n\to\infty} X$ or $X_n \xrightarrow[\mathcal{L}]{n\to\infty} X$. And alternative definition is one in which we make the distribution appear clearly : X_n converges in law if :

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

for every point x where $F(x)$ is continuous, with F_n and F the c.d.f. of X_n and X respectively.
Note:

- The sequences may not need to be defined on the same space, i.e. we can con$X_n : \Omega \to E_n$.
- If all the r.v. are defined on the same space, we can replace some of the X_n by other Y_n, provided that they are the same law $P_{X_n} = P_{Y_n}$!.
- The next proposition show an equivalence with another formulation. In many proofs of convergence in distribution
- In functional analysis, the convergence in distribution is called the weak convergence of measure[2]. There are many more measures converging weakly than there is functions converging in probability (or $a.s.$ or in L^p).

Proposition 3.4.
The sequence $\{X_n\}_{n\geq 0}$ converges in distribution if and only if, for all functions $f \in \mathcal{C}_c$, space of continuous function with compact support, we have

$$\lim_{n\to\infty} \mathbb{E}\big(\varphi(X_n)\big) = \mathbb{E}\big(\varphi(X)\big)$$

[2] The main idea is that we consider the convergence of measure μ_n (and not functions as in other modes of convergence), where we average against any "nice" test function φ (continuous and bounded or continuous with compact support)

$$\int \varphi \, d\mu_n \to \int \varphi \, d\mu$$

Note: These are the 3 first points of equivalence of portmanteau lemma (which can include many). An a list of 3 others that specify the

Theorem 3.11 (portmanteau lemma)**.**
We provides several equivalent definitions of convergence in distribution. Although these definitions are less intuitive, they are used to prove a number of statistical theorems. The convergence in distribution of $X_n \xrightarrow[n\to\infty]{\mathcal{D}} X$ *if and only if any of the following statements are true:*

- $\mathbb{P}(X_n \le x) \to \mathbb{P}(X \le x)$ *for all continuity points of* $x \mapsto \mathbb{P}(X \le x)$
- $\mathbb{E}\big[\varphi(X_n)\big] \to \mathbb{E}\big[\varphi(X)\big]$*, for all bounded continuous function's* f
- $\mathbb{E}\big[\varphi(X_n)\big] \to \mathbb{E}\big[\varphi(X)\big]$*, for all bounded, Lipschitz function's* f
- $\liminf \mathbb{E}\big[\varphi(X_n)\big] \ge \mathbb{E}\big[\varphi(X)\big]$ *for all nonnegative, continuous functions* f
- $\liminf \mathbb{P}(X_n \in G) \ge \mathbb{P}(X \in G)$ *for every open set* G
- $\limsup \mathbb{P}(X_n \in F) \le \mathbb{P}(X \in F)$ *for every closed set* F
- $\mathbb{P}(X_n \in B) \to \mathbb{P}(X \in B)$ *for all continuity sets* B *of random variable* X*;*
- $\limsup \mathbb{E}\big[\varphi(X_n) \le \mathbb{E}\big[\varphi(X)\big]$ *for every upper semi-continuous function* f *bounded above*
- $\liminf \mathbb{E}\big[\varphi(X_n)\big] \ge \mathbb{E}\big[\varphi(X)\big]$ *for every lower semi-continuous function* f *bounded below.*

Theorem 3.12 (CV in $\mathbb{P} \Rightarrow$ CV in $\mathcal{D}$)**.**
Let (X_n) *a sequence of random variables converging in probability to* X *then* (X_n) *converges in law / in distribution* X *:*

$$X_n \xrightarrow[n\to\infty]{\mathbb{P}} X \qquad \Longrightarrow \qquad X_n \xrightarrow[n\to\infty]{\mathcal{D}} X$$

The reciprocal

$$X_n \xrightarrow[\mathcal{D}]{n\to\infty} a \qquad \Longrightarrow \qquad X_n \xrightarrow[n\to\infty]{\mathbb{P}} a$$

Saying that (X_n) *converges in law to the constant* a *implies that the distribution/measure of* X_n *converges toward the Dirac measure at the point* a*:*

$$\delta_a(x) = \begin{cases} +\infty & x = a \\ 0 & x = 0 \end{cases} \qquad \text{and} \qquad \int_{-\infty}^{\infty} d\delta_a = 1 \qquad \int_{-\infty}^{\infty} x d\delta_a(x) = a$$

or, said differently:

$$\mathbb{E}\left[\varphi\left(X_n\right)\right] \xrightarrow[n\to\infty]{} \mathbb{E}[\varphi(a)] = \varphi(a)$$

Example 3.9 (CV in $\mathbb{P} \overset{\Rightarrow}{\not\Leftarrow}$ CV in $\mathcal{D}$)**.**
Examples and comments to see the link between these two notions.

- *Degenerate logistic regression: Consider a random variable following the logistic distribution:*

$$F_{X_n}(x) = \frac{\exp(nx)}{1+\exp(nx)} \qquad x \in \mathbb{R}$$

Then as $n \to \infty$ we have the limit c.d.f.:

$$F_X(x) = \begin{cases} 0 & \textit{if} \quad x < 0 \\ \frac{1}{2} & \textit{if} \quad x = 0 \\ 1 & \textit{if} \quad x > 0 \end{cases}$$

This is not exactly a c.d.f. as it is not right continuous at $x = 0$ (a defining properties of c.d.f.). However, as $x = 0$ is not a continuity of $F_X(x)$, we don't need to consider it in the definition of distribution. Moreover, it is clear that we have convergence in probability

$$\mathbb{P}[|X_n| < \varepsilon] = \frac{\exp(nx)}{1 + \exp(nx)} - \frac{\exp(-nx)}{1 + \exp(-nx)} \to 1 \quad \textit{as} \qquad n \to \infty$$

Hence we have that the limiting distribution is degenerate at $X = 0$ $X_n \xrightarrow[n\to\infty]{\mathcal{D}} X$ where $\mathbb{P}[X = 0] = 1$, or $X = 0$ almost surely, or the measure of X is a Dirac at zero: $P_X(x) = \delta_0(x)$

- *More generally, random variables that converge to a discrete random variable on $\{x_1, \dots, x_n\}$ have their probability distribution (or c.d.f.) converges toward the Dirac measure (measure with mass points) on $\{x_1, \dots, x_n\}$., and their c.d.f. converges towards the step function $F_X(x) = \sum_i \alpha_i \mathbb{1}[x_i, x_{i+1})(x)$*
- *It is quite easy to see why convergence in distribution is the weakest notion of convergence and doesn't imply others, for example in probability. Take simply a sequence of copies of a random variable: $X_n = X, \forall n$ and suppose $X \sim \mathcal{N}(0, 1)$. By symmetry of the Gaussian, we have that $\widetilde{X} = -X \sim \mathcal{N}(0, 1)$ as well. As a result:*

$$X_n = X \xrightarrow[n\to\infty]{\mathcal{D}} \widetilde{X} = -X$$

but of course, we don't have convergence in probability (as indeed $|X - \widetilde{X}| = 2X$ is strictly positive almost-surely!). To avoid the confusion with convergence of random variables, we replace the limit directly by its distribution:

$$X_n \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, 1)$$

Theorem 3.13 (Continuous mapping theorem).
Let (X_n) be a sequence of random variable and X another random variable, and g a function continuous everywhere on the set of discontinuity D_g. If $\mathbb{P}(X \in D_g) = 0$ (i.e. g is continuous P_X-almost everywhere, given the underlying distribution of X), then the sequence $g(X_n)$ inherit the mode of convergence of X_n, toward $g(X)$ $(g\,(X_n))$ herite du mode de convergence de la suite (X_n) :

1. $X_n \xrightarrow[n\to\infty]{a.s.} X \quad \Longrightarrow \quad g(X_n) \xrightarrow[n\to\infty]{a.s.} g(X)$
2. $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X \quad \Longrightarrow \quad g(X_n) \xrightarrow[n\to\infty]{\mathbb{P}} g(X)$
3. $X_n \xrightarrow[n\to\infty]{\mathcal{D}} X \quad \Longrightarrow \quad g(X_n) \xrightarrow[n\to\infty]{\mathcal{D}} g(X)$

What matters is not that g is continuous everywhere, but is continuous where g where X have some chance of falling, what we emphasize with condition $\mathbb{P}(X \in D_g) = 0$.

Note:

- This is one of the most important theorem in statistics, to evaluate the consistency of estimators, as countless proofs in A. Shaikh's class use it.
- Note however that convergence in distribution of $\{X_n\}_n$ to X and Y_n to Y does in general "not" imply convergence in distribution of $X_n + Y_n \rightarrow X + Y$ or of $X_n Y_n \rightarrow XY$
- The reason for that is that (X_n, Y_n) do not converge to (X, Y), *jointly*, preventing a potential convergence. The next theorem makes that clear and is a great generalization of the law of large number in the case of sequence of random vectors.

Theorem 3.14 (Marginals and joint convergence)**.** • *Let (X_n) be a sequence of random variable in $\mathbb{R}^k$ and X another random variable in $\mathbb{R}^k$. Let $X_{n,j}$ denote the j-th element of sequence X_n. Then,*

$$X_{n,j} \xrightarrow[n\to\infty]{\mathbb{P}} X_j \qquad \Longrightarrow \qquad X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$$

• *Convergences in marginal distributions does not imply convergence in joint distribution. To see this, consider*

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{matrix} 1 & (-1)^n \\ (-1)^n & 1 \end{matrix} \right)$$

Note that $X_n \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,1)$ and $Y_n \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,1)$ However, the joint density does not ever converge as it "flips" from being perfectly positive and negative correlated between X_n and Y_n

Note: Associated with the continuous mapping, we can easily have the convergence of estimators. An easy consequence is the Slutsky's lemma.

Corollary 3.3 (Slutsky's theorem)**.**

If $\{X_n\}_n$ converges in distribution to X and Y_n converges in probability to a constant $c \in \mathbb{R}^k$ then

$$\begin{aligned} X_n + Y_n &\xrightarrow[n\to\infty]{\mathcal{D}} X + c \\ X_n Y_n &\xrightarrow[n\to\infty]{\mathcal{D}} Xc \\ X_n / Y_n &\xrightarrow[n\to\infty]{\mathcal{D}} X/c \end{aligned}$$

Definition 3.17 (Characteristic function - Fourier transform)**.**

The characteristic function is given by the following mapping $\phi_X : \mathbb{R}^k \rightarrow \mathbb{C}$ the set of complex number:

$$\phi_X(t) = \mathbb{E}\left[e^{i\langle t,X\rangle}\right] = \int_{\mathbb{R}^k} e^{i\langle t,x\rangle} P_X(dx)$$

Said differently, the characteristic function is a rescaled version of the Fourier transform. One can use all the results from Fourier analysis to compute it.

Note:

- We always have $\phi_X(0) = 1$ (integral of the distribution sums to one) and $|\phi_X(t)| \leq 1$ alway lives in the unit disk. Moreover, the characteristic function is absolutely continuous w.r.t. Lebesgue measure.
- In particular, if X and Y are independent, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$
- It can be used to compute moments $\mathbb{E}[X^n] = \phi_X^{(n)}(0)/i^n$
- As its name says, this function characterize the law/distribution in the sense that two random variables X, Y have the same law i.f.f. they have the same characteristic function $\phi_X(t) = \phi_Y(t)$
- An important example is the Normal distribution: it is special since the Fourier transform of $\mathcal{N}(\mu, \sigma^2)$ is $\phi_X(t) = \exp(i\mu t - \frac{\sigma^2 t^2}{2})$ (which is a rescaled version of ... a Gaussian density!)
- This is fundamental for the proof of the central limit theorem

Now, we do a very quick detour to what mathematician call Laplace transform, and what economists use a lot in Gaussian log-linear models (linear models where the error terms/shocks follow Normal distribution and all the variables are expressed in logs).

Definition 3.18 (Laplace transform and example of expectation of log-normal)**.**
The characteristic function is given by the following mapping $\mathcal{L}_X(t) : \mathbb{R}^k \to \mathbb{R}$:

$$\mathcal{L}_X(t) = \mathbb{E}[e^{-\langle t, X\rangle}\Big] == \int_{\mathbb{R}^k} e^{-\langle t, x\rangle} P_X(dx)$$

More specifically it looks similar to the Characteristic function, but with the imaginary sign raised to power 2 (indeed $i^2 = -1$ by definition).

- An important example is again the Normal distribution: the Laplace transform of $\mathcal{N}(\mu, \sigma^2)$ is $\mathcal{L}_X(t) = \exp(\mu t + \frac{\sigma^2 t^2}{2})$ (which is again a rescaled version of ... a Gaussian density!)

Theorem 3.15 (Lévy's continuity theorem)**.**
The sequence $\{X_n\}_n$ converges in distribution to X if and only if the sequence of corresponding characteristic functions ϕ_n converges pointwise to the characteristic function ϕ of X, i.e.

$$\forall\ t \in\ \mathbb{R} \qquad \phi_{X_n}(t) \xrightarrow{n\to\infty} \phi_X(t)$$

We finally arrive to one of the most important result of statistics.

Theorem 3.16 (Central limit theorem)**.**
Let $(X_n)_{n\geq 1}$ a sequence of real random variables, i.i.d.*, with moments of second order $\mathbb{E}(X^2) < \infty$, and noting $S_n = \sum_{i=1}^{n} X_i$ and $\sigma^2 = Var(X)$, then:*

$$\lim_{n\to\infty} \sqrt{n}(\frac{S_n}{n} - \mu) \sim \mathcal{N}(0, \sigma^2)$$

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

This convergence is in law*, and that intuitively implies that any sum of r.v. falls "normally" around its mean μ, with a variance σ^2 and at a speed of convergence $\sqrt{n}$.*

Recap

In the following figure, I display the different modes of convergences and the links between them.

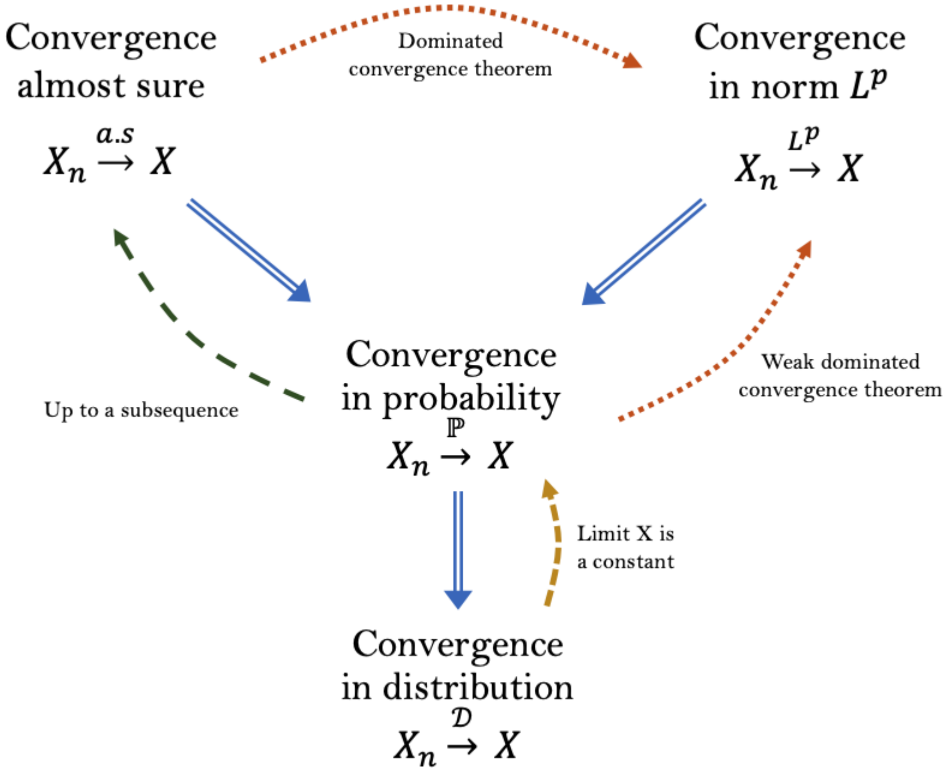
Convergence almost sure
$X_n \overset{a.s}{\rightarrow} X$
Dominated convergence theorem
Convergence in norm L^p
$X_n \overset{L^p}{\rightarrow} X$
Up to a subsequence
Convergence in probability
$X_n \overset{\mathbb{P}}{\rightarrow} X$
Weak dominated convergence theorem
Limit X is a constant
Convergence in distribution
$X_n \overset{\mathcal{D}}{\rightarrow} X$

Figure 8: **Modes of convergences – summary**

Comprehension questions:

A couple of (hopefully easy) questions to see if you understood the material above:

- Let $Y_n \sim \mathscr{E}(\lambda = n^2)$ and $Z_n \sim \mathcal{N}(\alpha^n, 1)$, where $\alpha < 1$ is a constant parameter. Justify carefully, with the help of some convergence theorem of measure theory, why

$$(i) \qquad \mathbb{E}\Big[\sum_{k=0}^{\infty} Y_k\Big] = \sum_{k=0}^{\infty} \mathbb{E}\Big[Y_k\Big] \qquad\qquad (ii) \qquad \mathbb{E}\Big[\sum_{k=0}^{\infty} Z_k\Big] = \sum_{k=0}^{\infty} \mathbb{E}\Big[Z_k\Big]$$

  and find these two values.
- Provide a careful (but easy) proof of the Markov inequality.
- Let two positive random variables X and Y, i.i.d. (independent and identically distributed). Can you find an example where X is integrable and the random variable $Z = \frac{X-Y}{2}$ is not ? If yes, why/which one? If not, why not?
- Find three types of counterexamples (c.f. the note) for the theorem 3.5 where removing the domination prevent the L^1-convergence.
- Prove the claims of example 3.6 about the respective convergences almost-surely and in probability of X_n and $\widetilde{X}_n$.
-
- Prove the 3rd remark of definition 3.15.
- Prove the proposition theorem 3.8 using the Markov inequality