Mathematical Methods in Economics
Lecture Notes

Thomas Bourany*
The University of Chicago

September 9, 2021

*thomasbourany@uchicago.edu, thomasbourany.github.io, I thank the previous lecturers of this math camp and TA of core courses – Kai Hao Yang, Kai-Wei Hsu and Yu-Ting Chiang as well as Takuma Habu and Agustin Gutierriez – for generously sharing their material as . 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

Contents

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 2.3.
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 2.4.
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,

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

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 2.5.
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.,

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

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 2.5 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 2.6 (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 2.7 (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 $\mathcal{C} = \{x \in X, \text{s.t. } g_1(x), \ldots, g_M \leq 0, \}$. As a result the problem $\mathcal{P}$ becomes :

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

Theorem 2.8 (Karush-Kuhn-Tucker, Necessity)**.**
Let $(X, ||\cdot||)$ be a normed vector space, and f and multiple constraint g_i, $f : X \to \mathbb{R}$, $g_i : X \to \mathbb{R}$, $\forall i = 1, \ldots, M$, functions which are both continuous and with continuous derivative. We introduce the "Lagrangian" $\mathcal{L}$ function associated to this problem:

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

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):* $g_i(x^*) \leq 0 \quad$ *for* $\quad 1 \leq i \leq M$

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

4. *Complementarity:* $\sum_{i=1}^{M} \lambda_i \, g_i(x^*) = 0$
(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 2.9 (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 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 only needs to be "measurable", which happens very (very) often!).
To reach this, it is required to define a function $f : X \to \mathbb{R}$ in a new way: we do not need to consider all the points 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 does not matter because it has "measure zero".

The measure (or distribution) μ is an extension of measuring interval sizes. For example, in 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.

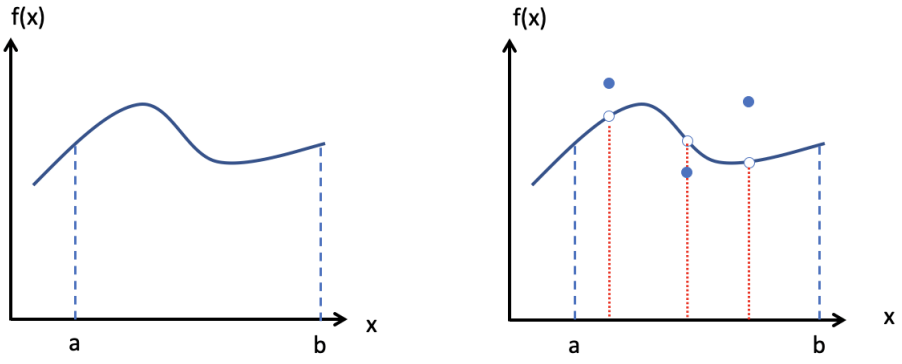

Figure 1: **Two functions equal almost everywhere**

The benefit of this is the great flexibility of integration. Despite a long (and a bit tedious) construction of the Lebesgue integral, the main difference with 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 eliminating many pathological cases when a limit of integrable function f_n

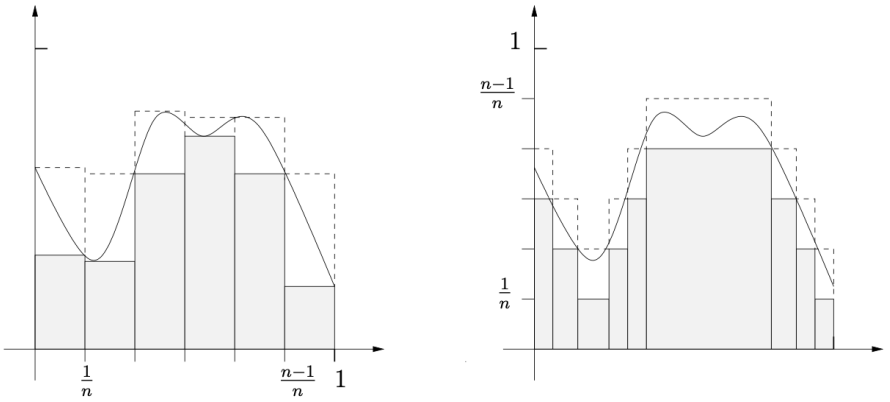

Figure 2: **Difference between the constructions of Riemann (LHS) and Lebesgue (RHS) integrals**

isn't Riemann-integrable (but it 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!). Hence, we can use all the artillerie of measure theory, in particular the 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 appear 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, i.e. the events associated with the potential results. Because of that, you can not compute the probabilities of the random variables outcome.

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 $(\widetilde{\Omega}, \mathscr{F}_{\widetilde{\Omega}})$!).
Note: In practice, we do not focus too much on Ω (except for some definitions 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 following 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.

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

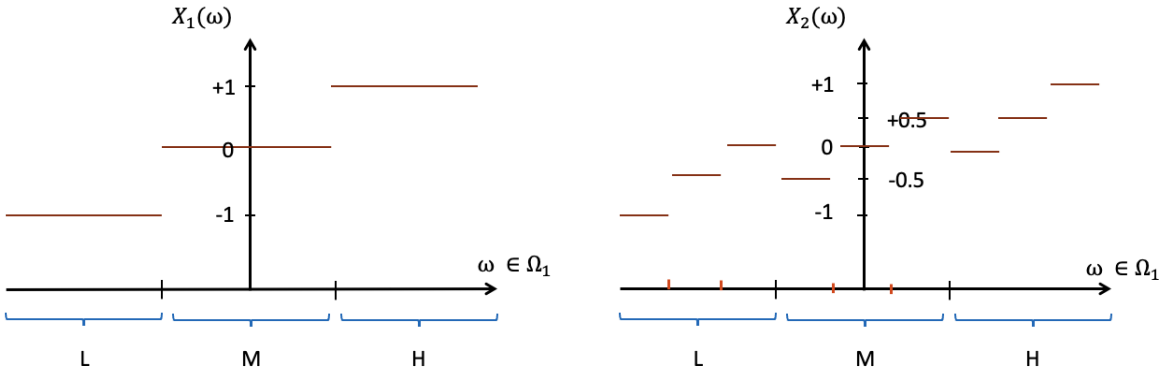

Figure 3: **Measurability (or not!) of the random variable X_1 and X_2 w.r.t.** $(\Omega, \mathscr{F}_\Omega)$

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: As a result, if Y includes the same "relevant information" as X – i.e. Y is measurable w.r.t. $\sigma(X)$ but doesn't have more information than that!! – this implies that there exists a deterministic mapping between X and Y. However, this necessity Y is $\sigma(X)$-mesurable $\Rightarrow$ $Y = f(X)$ is the one trickier to prove.

Now, let us consider the probability measure for random variables: that's what we call their "law":

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:

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

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 admits 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 results on integrals – like homogeneity, linearity, monotonicity, etc. – are valid for the Lebesgue integral as much as for the usual (Riemann) integral.

Theorem 3.2 (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 mathematicians and economists are a bit handwavy on the notation of measures. Usually, for an abstract measure μ on the set E, we 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 3.3.
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)

A last but essential theorem of measure theory links the measures of probability we saw above and the usual density of

Theorem 3.4 (Radon-Nikodym Theorem)**.**
Let $(E, \mathscr{E})$ be a measurable space and let μ and v be two measures on $(E, \mathscr{E})$ with $\mu(E) < \infty$ and $\nu(E) < \infty$ such that μ is absolutely continuous with respect to ν. Then, there exists $f : E \to \mathbb{R}$ that is $\mathscr{E}$-measurable such that

$$\mu(B) = \int_B f d\nu, \qquad \forall B \in \mathscr{E}$$

Furthermore, if there exists $f : E \to \mathbb{R}$ and $g : E \to \mathbb{R}$ such that this equation holds, then $f \equiv g$ ν-almost everywhere.
Note: The Radon-Nikodym Theorem says that if $\mu \ll \nu$ then there exists an essentially unique function f such that the measure μ can be represented by the integral of f with respect to ν.

We call this function f the Radon-Nikodym derivative of μ with respect to ν often denoted by $f \equiv \frac{d\mu}{d\nu}$ As such the equation above can be written alternatively, $\forall B \in \mathscr{E}$ as

$$\mu(B) = \int_B \mathbb{1} d\mu = \int_B \frac{d\mu}{d\nu} d\nu = \int_B f(\cdot) d\nu$$

As a corollary, we can define the density function.

Corollary 3.5.
Let us consider for example the probability measure $\mu = P_X \in \Delta(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that is absolutely continuous with respect to the Lebesgue measure. Then, there exists an essentially (almost-everywhere) unique $f : \mathbb{R} \to \mathbb{R}$ such that

$$\mu(B) = P_X(B) := \int_B f d\lambda, \forall B \in \mathscr{E}$$

The function f in the corollary above is precisely the density function of the probability distribution P_X, as the Lebesgue integral of the function on any measurable set B is exactly the probability of B Notice that, for any such probability measure P_X

$$F_X(x) := F_{P_X}(x) = \int_{-\infty}^{x} f(z) dz, \forall x \in \mathbb{R}$$

Together with the Fundamental Theorem of Calculus, we have that $F' \equiv f$ Lebesgue-almost everywhere.

Note that if P_X is not absolutely continuous, then the statement above is not true i.e. not all CDF has corresponding density functions. For example, the Cantor function is counterexample.

Although not all the probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ are absolutely continuous with respect to the Lebesgue measure, all of the measures can be decomposed so that part of it is absolutely continuous. (c.f. notes and Lebesgue decomposition theorem).

3.3 Additional results of measure theory

More is coming: Some more insight on the construction of the Lebesgue integral (outer measure, definition of the sum of unions, countable additivity, simple approximation theorem, extension with sum and difference (provided integrability), extension to the limit), Borel Cantelli (?), Fubini's theorem, Lebesgue decomposition

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 $X_1, \ldots, X_n$ i.i.d. Is X_n measurable w.r.t. $\sigma(X_1)$? And what about $\sigma(X_1, \ldots, X_{n-1})$? And what about $\sigma(X_1, \ldots, X_n)$?

3.4 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 3.6 (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 3.7.
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 3.8.
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 3.9 (Markov-Chebyshev's inequality)**.**
Let $X : \Omega \to \mathbb{R}_+$ a positive random variable. Then, for any constant $c > 0$:

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

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

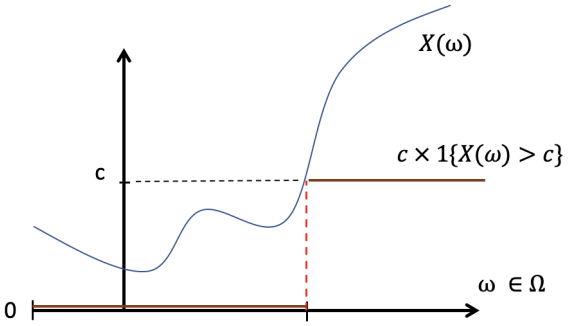

Figure 4: **Markov inequality in one picture**

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

Theorem 3.10.
Let X_n a sequence of positive *random variables, then*

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

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 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-surely (or $\mathbb{P}$-a.e.).

Theorem 3.11 (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 3.12 (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 :

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

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.

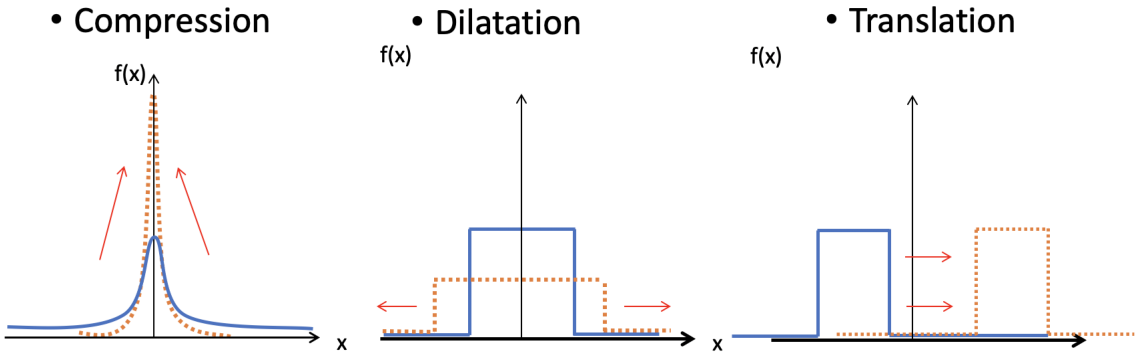

Figure 5: **Counterexamples to the dominated CV thm, because of lack of domination**

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* $\varepsilon > 0$*:*

$$\mathbb{P}(|S_n| > n\varepsilon) = \mathbb{P}(|S_n|^3 > n^3\varepsilon^3) \leq \frac{\mathbb{E}[|S_n|^3]}{\varepsilon^3 n^3} = \frac{\mathbb{E}[|X_1|^3]}{\varepsilon^2 n^{3/2}}$$

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

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

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:

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

Theorem 3.13 (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

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

(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

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

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*

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

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*

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

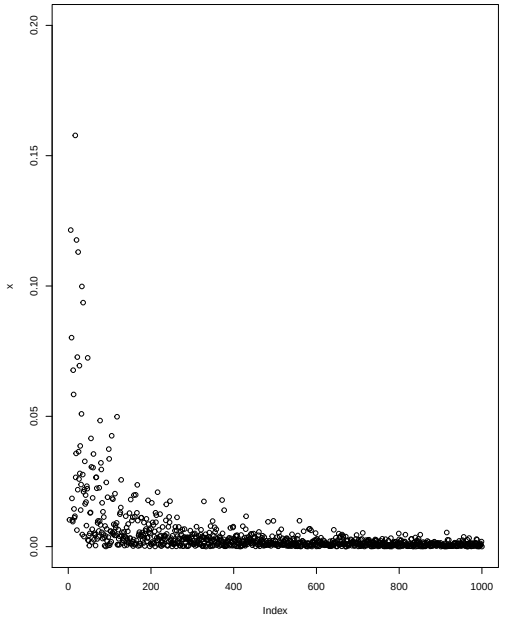

Figure 6: $X_n \sim \mathscr{E}(\lambda = n)$ **converges to** $X = 0, a.s.$

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 :*

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

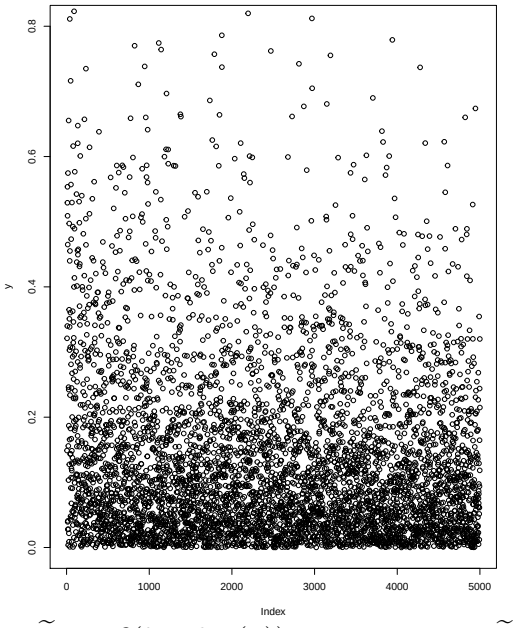

Figure 7: $\widetilde{X}_n \sim \mathscr{E}(\lambda = \log(n))$ **converges to** $\widetilde{X} = 0$**, in proba, but not almost surely**

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

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

Theorem 3.14 (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 $\Omega = [0, \frac{1}{2}]$ *with* $\mathscr{B}_{[0,\frac{1}{2}]}$ *and the Lebesgue measure. Consider for all* n*,* k_n *is such that* $2^{k_n} < n \leq 2^{k_n+1}$ *and consider the sequence* $X_n(\omega) = \mathbb{1}_{\left(\frac{n-2^k-1}{2^{k+1}}, \frac{n-2^k}{2^{k+1}}\right]}(\omega)$*, with the first few elements such that:*

$$X_2(\omega) := \mathbb{1}_{\left(0,\frac{1}{2}\right]}(\omega)$$

$$X_3(\omega) := \mathbb{1}_{\left(0,\frac{1}{4}\right]}(\omega) \qquad X_4(\omega) := \mathbb{1}_{\left(\frac{1}{4},\frac{1}{2}\right]}(\omega) \qquad X_4(\omega) := \mathbb{1}_{\left(\frac{1}{4},\frac{1}{2}\right]}(\omega)$$

$$X_5(\omega) := \mathbb{1}_{\left(0,\frac{1}{8}\right]}(\omega) \qquad X_6(\omega) := \mathbb{1}_{\left(\frac{1}{8},\frac{2}{8}\right]}(\omega) \qquad X_7(\omega) := \mathbb{1}_{\left(\frac{2}{8},\frac{3}{8}\right]}(\omega) \ldots$$

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

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

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 3.15 (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 $\mathbb{P} \not\Rightarrow$ 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 $\mathbb{P}(X_n = 0) = 1 - \frac{1}{\ln n}$. For all tout $\varepsilon > 0$, we have:

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

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

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

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 actually weaken the assumption.

Theorem 3.16 (Weaker Dominated convergence theorem and Fatou's lemma)**.**
These theorems hold for sequences of r.v. converging in probability instead of almost-surely:

- *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* $\mathbb{E}(X) \leq \liminf_{n\to\infty} \mathbb{E}(X_n)$

In the dominated convergence, we assumed that the sequence of random variables is dominated, which is enough to get rid of the pathological cases of compressions/dilatation/translation. However, convergence almost-sure and domination can be very strong. The aim of Vilati's theorem below is to relax these assumptions and choosing the weakest conditions sufficient (so weak that they are in fact necessary) – which are: convergence in probability, uniform integrability and tightness – to obtain the convergence in L^p as a necessary and sufficient condition. Let us introduce these two new notions first.

Definition 3.16 (Tightness)**.**
A sequence of random variables $\{X_n : n \geq 1\}$ *is tight if, for any* $\varepsilon > 0$ *there exists a (finite) constant* $B > 0$ *such that*

$$\inf_n \mathbb{P}\left(|X_n| \leq B\right) \geq 1 - \varepsilon$$

Equivalently, $\{X_n\}_n$ *is tight if, for any* $\varepsilon > 0$*, there exists a finite constant* $M_\varepsilon < \infty$ *and* $n_\varepsilon \in \mathbb{N}$ *such that*

$$\mathbb{P}\left(|X_n| > M_\varepsilon\right) < \varepsilon, \ \forall n \geq n_\varepsilon$$

Note:

- This is called "boundedness in probability" or uniform tightness. It is a condition to prevent the collection of measure P_{X_n} to "escape to infinity" (i.e. to avoid the translation cases in the examples above).
- A finite sequence of random variables is always tight.

Definition 3.17 (Uniform integrability)**.**
A collection of random variable $\{X_n\}_n$ *is said to be* uniformly integrable *if for any* $\varepsilon > 0$*, there exists* $\delta_\varepsilon > 0$ *such that whenever* $A \subseteq \omega$ *is measurable* $\mathbb{P}(A) < \delta$

$$\int_A |X(\omega)|\mathbb{P}(d\omega) < \varepsilon$$

for all $X_n \in \{X_n\}_n$ *(i.e. for all* n*).*
A more probabilistic definition (but equivalent) replace the set A *by* $\{|X| \geq K\}$*: the sequence* $\{X_n\}_n$ *is* uniformly integrable *if* $\forall \varepsilon > 0$*, there exists* $0 < K < \infty$ *(with* $K = K_\varepsilon$*) such that, for all* $X_n \in \{X_n\}_n$*. i.e. for all* n*:*

$$\int_\Omega |X(\omega)|\mathbb{1}\{|X(\omega)| \geq K\}\mathbb{P}(d\omega) = \mathbb{E}\Big[|X|\mathbb{1}\{|X| \geq K\}\Big] < \varepsilon$$

Note:

- Uniform integrability is a really important notion for convergence of martingales (however not really covered in the core sequence)
- A finite sequence of integrable random variables is always uniform integrable.

Thanks to these two notions, let us cover the sufficient and necessary conditions for the Dominated CV theorem

Theorem 3.17 (Vitali's theorem)**.**
Let $(X_n)_n \subseteq L^p(\Omega, \mathscr{F}, \mathbb{P})$, $X \in L^p(\Omega, \mathscr{F}, \mathbb{P})$ *with* $1 \leq p < \infty$.
Then, $X_n \to X$ *in* L^p if and only if *we have:*

(i) X_n *converge in probability to* X
(ii) $\{X_n\}$ *is tight and/or* $\mathbb{P}(\Omega) < \infty$ *(trivial for proba. measures, but not other measures).*
(iii) $\{X_n\}_n$ *is uniform integrable*

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

Theorem 3.18 (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 3.18 (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 φ*

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

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 consider $X_n : \Omega \to E_n$.
- 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. There are many more measures converging weakly than there is functions converging in probability (or *a.s.* or in L^p). 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)

$$\int \varphi(x)\,\mu_n(dx) \to \int \varphi(x)\,\mu(dx)$$

Proposition 3.19.
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

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

Note: These are several equivalence statements listed in the portmanteau lemma.

Theorem 3.20 (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:

- $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ *for all continuity points of* $x \mapsto \mathbb{P}(X \leq x)$
- $\mathbb{E}\big[\varphi(X_n)\big] \to \mathbb{E}\big[\varphi(X)\big]$, *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
- $\liminf \mathbb{E}\big[\varphi(X_n)\big] \geq \mathbb{E}\big[\varphi(X)\big]$ *for all nonnegative, continuous functions* f
- $\liminf \mathbb{P}(X_n \in G) \geq \mathbb{P}(X \in G)$ *for every open set* G
- $\limsup \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$ *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*;*
- $\limsup \mathbb{E}\big[\varphi(X_n) \leq \mathbb{E}\big[\varphi(X)\big]$ *for every upper semi-continuous function* f *bounded above*
- $\liminf \mathbb{E}\big[\varphi(X_n)\big] \geq \mathbb{E}\big[\varphi(X)\big]$ *for every lower semi-continuous function* φ *bounded below.*

Theorem 3.21 (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

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

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.:

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

This is not exactly a c.d.f. as it is not right continuous at $x = 0$ (a defining property 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

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

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: $P_X(x) = \delta_0(x)$. As a result, the convergence in distribution toward a constant and the convergence in probability are equivalent here, as implied by the last theorem.

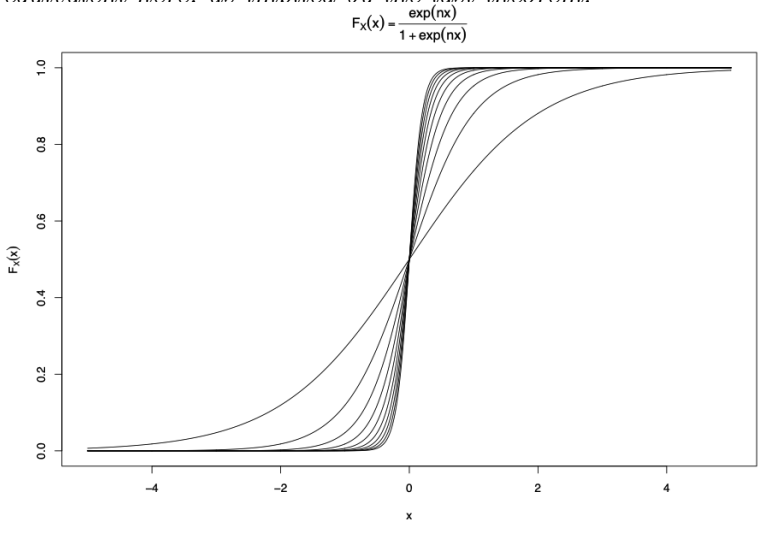

Figure 8: $\widetilde{X}_n$ **converges to** $\widetilde{X} = 0$**, in distribution, i.e.** $P_{X_n} \xrightarrow[n\to\infty]{\text{weak-*}} \delta_0$

- *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:

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

Here are some propositions that are covered in Metrics 1 (when talking about τ-consistency) that link the tightness and convergence in distribution.

Proposition 3.22.
Let $\{X_n\}_n$ sequence of random variables.

- *If $X_n \xrightarrow[n\to\infty]{\mathcal{D}} X$, then X_n is tight. (or more precisely $\{P_{X_n}\}$ is tight)*
- *Tightness is not a sufficient condition*
- *Prokhorov's theorem: If the sequence $\{X_n\}$ is tight, then there exists a subsequence $N(n), \forall n$ such that $X_n \xrightarrow[n\to\infty]{\mathcal{D}} X$.*

Theorem 3.23 (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. $X_n \xrightarrow[n\to\infty]{a.s.} X \implies g(X_n) \xrightarrow[n\to\infty]{a.s.} g(X)$
2. $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X \implies g(X_n) \xrightarrow[n\to\infty]{\mathbb{P}} g(X)$
3. $X_n \xrightarrow[n\to\infty]{\mathcal{D}} X \implies g(X_n) \xrightarrow[n\to\infty]{\mathcal{D}} g(X)$

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 $X_n + Y_n \to X + Y$ or of $X_n Y_n \to XY$
- 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 3.24 (Marginals and joint convergence)**.**
Let $X_n = (X_{n,1}, \dots, X_{n,k})$ vectors of random variables (random vectors).

- *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 3.25 (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

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

Note: This is simply an application of the joint convergence in probability and the continuous mapping theorem.

Definition 3.19 (Characteristic function - Fourier transform)**.**
The characteristic function is given by the following mapping $\phi_X : \mathbb{R}^k \to \mathbb{C}$ 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 $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$. This is simpler than using the usual method: Indeed, for a test function $g(\cdot)$ we can compute the sum as follow:

$$\mathbb{E}[g(X+Y)] = \int\int g(x+y)f_X(x)f_Y(y)dxdy \overset{z:=x+y}{=} \int\int g(z)f_X(z-y)f_Y(y)dzdy = \int g(z)\int f_X(z-y)f_Y(y)$$

  Hence we see that $f_{X+Y}(z) = \int f_X(z-y)f_Y(y)dz \equiv (f_X \star f_Y)(z)$ (the distribution of the sum is the convolution of the distributions!). That's consistent with the fact that the Fourier transform of the convolution is the product of the Fourier transforms!
- The characteristic function can be used to compute moments $\mathbb{E}[X^n] = \phi_X^{(n)}(0)/i^n$
- As its name says, this function characterizes 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 log-normal / linear models (linear models where the error terms/shocks follow Gaussian distribution and all the variables are expressed in logs).

Definition 3.20 (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 3.26 (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 3.27 (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:*

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

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

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* $\sqrt{n}$*.*

Note: The following picture show that the CLT, like the LLN requires the finiteness of first (and second) moments! Cauchy Distribution is the typical example of a distribution with an infinite mean (if it even makes sense as we can't even define the expectation when the first moment doesn't exist!)

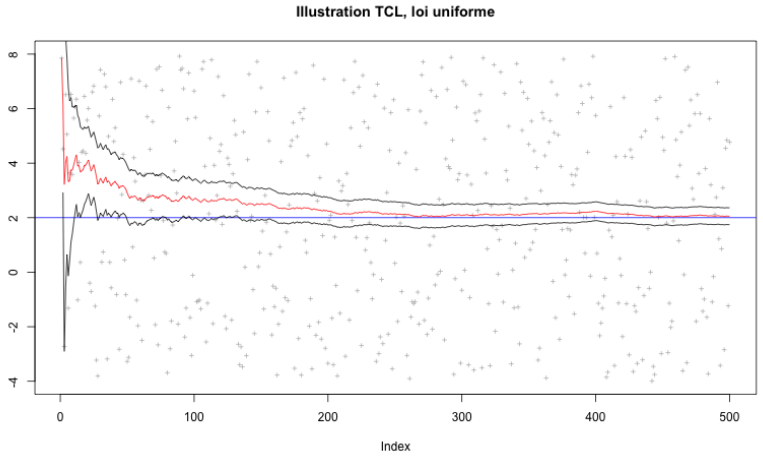

Figure 9: Example of convergence in law in the CLT

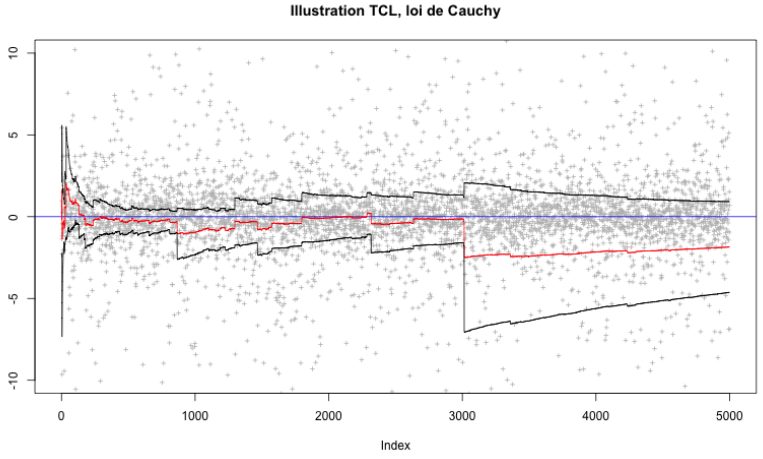

Figure 10: Example of non-convergence in law (no first moment) in the CLT

Glivenko Cantelli

We have all the tools to cover this theorem (from V. Glivenko, and F Cantelli in 1933) slightly advanced for the core sequence but very important for econometric theory (Vapnik-Chervonenkis theory in machine learning and M-estimators in econometrics).

Assume that $\{X_n\}_n$ is a sequence of independent and identically-distributed random variables in $\mathbb{R}$ with common cumulative distribution function $F(x)$. The empirical distribution function for $X_1, \ldots, X_n$ is defined by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i,\infty)}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\} = \frac{1}{n}\#\{1 \leq i \leq n \mid X_i \leq x\}$$

where $\mathbb{1}_C$ is the indicator function of the set C. Said differently, this is the cumulative sum of the histogram of sample $\{X_n\}_n$. First notice that, for every (fixed) x, $F_n(x)$ is a sequence of random variables which converges to $F(x)$ almost surely by the strong law of large numbers, that is, F_n converges to F pointwise. Glivenko-Cantelli theorem strengthens this result by considering the function F as a whole (for all x) using the supremum-norm.

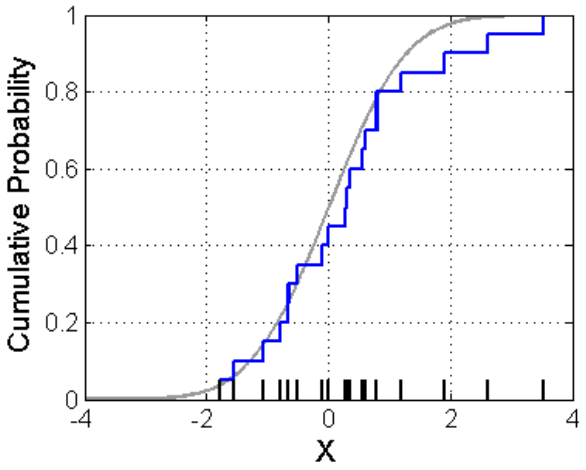

Figure 11: Glivenko-Cantelli: An empirical c.d.f and the limiting theoretical c.d.f

Theorem 3.28 (Glivenko Cantelli)**.**
Consider the empirical distribution F_n of the elements of sequence $\{X_n\}_n$, the function F_n converges uniformly to F:

$$\|F_n - F\|_\infty = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \longrightarrow 0 \qquad \textit{almost surely}$$

Note:

- One can generalize it to empirical measure indexed by sets $C \in \mathcal{C}$.

$$P_n(C) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_C(X_i), \qquad C \in \mathcal{C}$$

- If X_n is a stationary ergodic process (c.f. definition below in section 4), then $F_n(x)$ converges almost surely to $F(x) = \mathbb{E}\left(\mathbb{1}_{\{X_1 \leq x\}}\right)$. The Glivenko-Cantelli theorem gives a stronger mode of convergence than this in the iid case. Moreover, this has the same intuition as an ergodic theorem (more on this below). The average law over time (the empirical distribution) converges to the law over space at the limit.

CLT and Confidence interval

In the following, we will derive confidence regions for a test.

Definition 3.21 (Confidence set/region)**.**
A confidence set/region of level $1-\alpha$ *for* $\mu = \mathbb{E}[X]$*, denoted* $C_n = C_n(X_1, X_2, \ldots, X_n)$*, is a set such that the probability that the true mean is contained in the set is greater than* $1-\alpha, \alpha \in (0,1)$ *i.e.*

$$P(\mu \in C_n) \geq 1 - \alpha$$

For example, if $\alpha = 5\%$, then C_n gives us the interval in which the the probability that C_n contains the true mean is 95%. Here we will rely on the asymptotic properties of the CLT. In the next example we will show that we can also obtain non-asymptotic (but usually wider), confidence set using Markov's Inequality.

Example 3.10 (Confidence region using CLT))**.**
$X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ *Bernoulli* $\mathcal{B}(q)$ *where* $q \in (0,1)$*. Let* α *be given. We wish to construct a confidence region for* $\mu = \mathbb{E}[X] = q$ *at level* $1-\alpha$ *Recall that the WLLN tells us that :*

$$\bar{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} \mu = q$$

since $\mathbb{V}ar(X_n) =: \sigma^2 = q(1-q)$ *a natural candidate for* σ^2 *is*

$$s_n^2 = \bar{X}_n\left(1 - \bar{X}_n\right)$$

Thus, we can write s_n^2 *as a function* g*, parameterized as* $s_n^2 = g\left(\bar{X}_n\right)$ *with* $g(a) = a(1-a)$*. Since* g *is of course continuous, by the Continuous Mapping Theorem:*

$$s_n^2 \xrightarrow[n\to\infty]{\mathbb{P}} \sigma^2$$

since $\sigma^2 > 0$*, by Slutsky's Lemma,*

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu(P)\right)}{\sqrt{\bar{X}_n\left(1 - \bar{X}_n\right)}} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,1)$$

Defining, for $z_x = \Phi(x)$ the quantile of the standard normal distribution:

$$c_n := z_{1-\frac{\alpha}{2}} \frac{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}{\sqrt{n}}$$

and hence our confidence interval:

$$C_n := \left[\bar{X}_n - c_n, \bar{X}_n + c_n\right]$$

Now that we constructed it, let us check that it is indeed a confidence region, i.e. show that $\mathbb{P}\left(\mu \in C_n\right) \to 1-\alpha$:

$$\begin{aligned}
\mathbb{P}\left(\mu \in C_n\right) &= \mathbb{P}\big(\left[\bar{X}_n - c_n \leq \mu \geq \bar{X}_n + c_n\right) \\
&= \mathbb{P}\big(|\bar{X}_n - \mu| \leq c_n\big) = \mathbb{P}\left(|\bar{X}_n - \mu| \leq z_{1-\frac{\alpha}{2}} \frac{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(\frac{\sqrt{n}|\bar{X}_n - \mu|}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}} \leq z_{1-\frac{\alpha}{2}}\right) \\
&\to_{n\to\infty} \mathbb{P}\big(|z| \leq z_{1-\frac{\alpha}{2}}\big) \qquad\qquad CV \text{ in } \mathcal{D} \\
&= 1 - (\frac{\alpha}{2} + \frac{\alpha}{2}) = 1-\alpha
\end{aligned}$$

We can write confident regions in the following equivalent way:

$$C_n := \left\{ x \in \mathbb{R} : \frac{\sqrt{n}\left|\bar{X}_n - x\right|}{\sqrt{\bar{X}_n\left(1-\bar{X}_n\right)}} \leq z_{1-\frac{n}{2}} \right\}$$

The probability $\mathbb{P}\left(\mu(P) \in C_n\right)$ above is called the coverage probability. The actual coverage probability based on data may be poor in finite samples, in particular, when q is close to 0 or 1. Notice that confidence region using Markov's inequality does not have this problem (although the region is wider).

Example 3.11 (Confidence sets with Markov's inequality)**.**
Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli $\mathcal{B}(q)$ where $q \in (0,1)$. Let α be given. We wish to construct a confidence region for $\mu = \mathbb{E}[X] = q$ at level $1-\alpha$, for $\alpha \in (0,1)$. We can use the Markov's inequality to construct this set (later, we will use the Central Limit Theorem).

Letting $q = 2$, we obtain:

$$\begin{aligned}\mathbb{P}\left(\left|\bar{X}_n - \mu\right| > \varepsilon\right) &\leq \frac{\mathbb{E}\left[\left(\bar{X}_n - \mu\right)^2\right]}{\varepsilon^2} \\ &= \frac{\mathbb{V}ar\left[\bar{X}_n\right]}{\varepsilon^2} = \frac{\frac{1}{n^2}\mathbb{V}ar\left[\sum_{i=1}^n X\right]}{\varepsilon^2} = \frac{\frac{1}{n^2}\left(\sum_{i=1}^n \mathbb{V}ar[X] + 2\sum_{i\neq j}\mathbb{C}ov[X_i, X_j]\right)}{\varepsilon^2} \\ &= \frac{1}{n^2}\frac{n\mathbb{V}ar(X)}{\varepsilon^2} = \frac{q(1-q)}{n\varepsilon^2}\end{aligned}$$

where we used the fact that X_i 's are identically distributed (i.e. $\mathbb{V}ar[X_i] = \mathbb{V}ar[X]$ for all i) and independently distributed (i.e. $\mathbb{C}ov[X_i, X_j] = 0$ for all $i \neq j$)

Recap

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

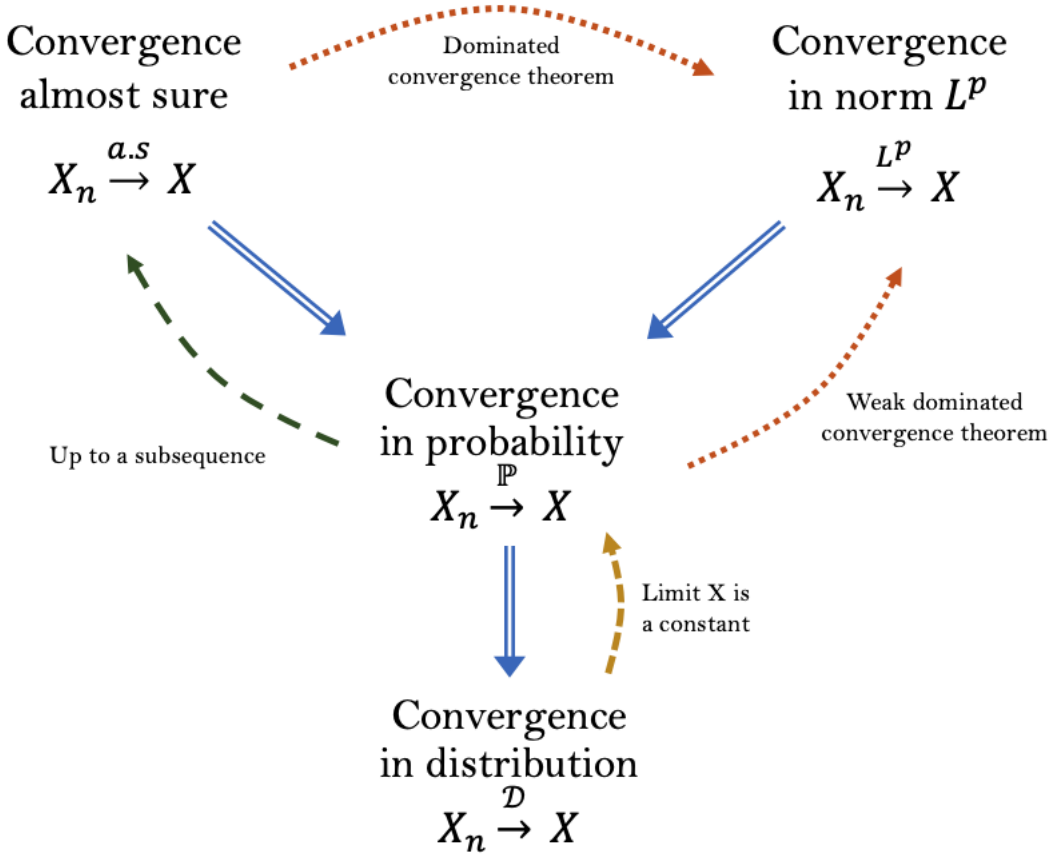

Figure 12: **Modes of convergences – summary**

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 3.12 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 3.15 using the Markov inequality

3.5 Additional topics in probability and statistics

Independence

Definition 3.22.
We can (re-)define independence using our measure theory formalism.

- *Let $\mathscr{G}_1, \cdots, \mathscr{G}_n$ be sub-σ-algebra of $\mathscr{F}$. We say they are* independent *if :*

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \times \cdots \times \mathbb{P}(A_n), \qquad \forall A_i \in \mathscr{G}_i \qquad 1 \leq i \leq n$$

- *Let $X_1, \cdots, X_n$ random variables with values in $(E_1, \mathscr{E}_1), \cdots, (E_n, \mathscr{E}_n)$, respectively. We say that $X_1, \cdots, X_n$ are independent if $\sigma(X_1), \cdots, \sigma(X_n)$ are independent. This is equivalent to*

$$\mathbb{P}(X_1 \in B_1, \cdots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \times \cdots \times \mathbb{P}(X_n \in B_n), \quad \forall B_i \in \mathscr{E}_i \qquad 1 \leq i \leq n$$

  Indeed we just need to recall that $\sigma(X_i) = \left\{X_i^{-1}(B) : B \in \mathscr{E}_i\right\}$.
- *We say that event $A_1, \cdots, A_n$ are independent if the random variables $\mathbb{1}_{A_1}, \cdots, \mathbb{1}_{A_n}$ are, which gives use the usual definition from our high-school class (remember $\sigma(\mathbb{1}A) = \{\emptyset, A, A^c, \Omega\}$)*

Consequences:
If $\mathscr{G}_1, \cdots, \mathscr{G}_n$ independent sub-σ-algebra, and if for all i X_i is a random measure that is $\mathscr{G}_i$-measurable, then $X_1, \cdots, X_n$ are independent. Moreover, given $f_i : (E_i, \mathscr{E}_i) \to (\widetilde{E}_i, \widetilde{\mathscr{E}})$ for $1 \leq i \leq n$ are measurable functions, then $f_1(X_1), \cdots, f_n(X_n)$ are also independent random variables.

Theorem 3.29.
Let $X_1, \cdots, X_n$ are independent random variables, with values in $(E_1, \mathscr{C}_1), \cdots, (E_n, \mathscr{E}_n)$ respectively. The following conditions are equivalent:

(i) $X_1, \cdots, X_n$ *are independent*
(ii) $P_{(X_1, \cdots, X_n)} = P_{X_1} \otimes \cdots \otimes P_{X_n}$
(iii) $\mathbb{E}[f_1(X_1) \cdots f_n(X_n)] = \mathbb{E}[f_1(X_1)] \cdots \mathbb{E}[f_n(X_n)]$ *for all positive measurable function (or bounded) f_i on $(E_i, \mathscr{E}_i)$*

Note: The last implication $(ii) \Rightarrow$ is simply Fubini's theorem, which is also good to have in our toolbox.

$$\begin{aligned}
\mathbb{E}[f_1(X_1) \cdots f_n(X_n)] &= \int_{E_1 \times \cdots \times E_n} f_1(x_1) \cdots f_n(x_n) \, P_{X_1}(dx_1) \cdots P_{X_n}(dx_n) \\
&= \left(\int_{E_1} f_1(x_1) \, P_{X_1}(\mathrm{d}x_1)\right) \cdots \left(\int_{E_n} f_n(x_n) \, P_{X_n}(\mathrm{d}x_n)\right) \\
&= \mathbb{E}[f_1(X_1)] \cdots \mathbb{E}[f_n(X_n)]
\end{aligned}$$

Corollary 3.30.
Let $X_1, \cdots, X_n$ be independent real random variables, such that X_i admit a density denoted f_{X_i}. Then $(X_1, \ldots, X_n)$ admit a density denoted

$$f_{(X_1,\cdots X_n)}(x_1, \cdots, x_n) := f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

Corollary 3.31.
X_1 and X_2 are independent real random variables in L^2 (i.e. admit a 2nd order moment). Then

$$\mathbb{C}ov(X_1, X_2) := \mathbb{E}\Big[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2]\Big] = 0$$

Note:

- The converse in not true!!
- The only random variables that are independent if and only if they are uncorrelated are the Gaussian vectors (Tamdam! $\rightarrow$ perfect transition!).

Gaussian Vectors

Let $\mathbf{X} = (X_1, \ldots, X_K)$ a vector following the multivariate normal distribution : it is said to be "non-degenerate" when the symmetric covariance matrix $\mathbf{\Sigma}$ is positive definite. In this case, let us recall the density:

$$f_{\mathbf{X}}(x_1, \ldots, x_K) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)}{\sqrt{(2\pi)^K \det(\mathbf{\Sigma})}}$$

where x $\in \mathbb{R}^k$-dimensional column vector and $\det(\mathbf{\Sigma})$ the determinant of $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbb{V}\mathrm{ar}(X_1) & \mathbb{C}\mathrm{ov}(X_1, X_2) & \ldots & \mathbb{C}\mathrm{ov}(X_1, X_n) \\ \vdots & \ldots & \ldots & \vdots \\ \mathbb{C}\mathrm{ov}(X_n, X_1) & \ldots & \ldots & \mathbb{V}\mathrm{ar}(X_n) \end{pmatrix}$$

Note: We can rewrite any Gaussian vector as a linear combination:

$$\mathbf{X} = A\mathbf{Z} + \mu$$

where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_K)$ and A is a matrix that decompose Σ such that $\Sigma = AA^T$

Bivariate case:

In the 2 -dimensional nonsingular case ($k = \text{rank}(\Sigma) = 2$), the probability density function of a vector (X, Y) is:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right)$$

where ρ is the correlation between X and Y and where $\sigma_X > 0$ and $\sigma_Y > 0$ are the respective standard deviation of X and Y. In this case,

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

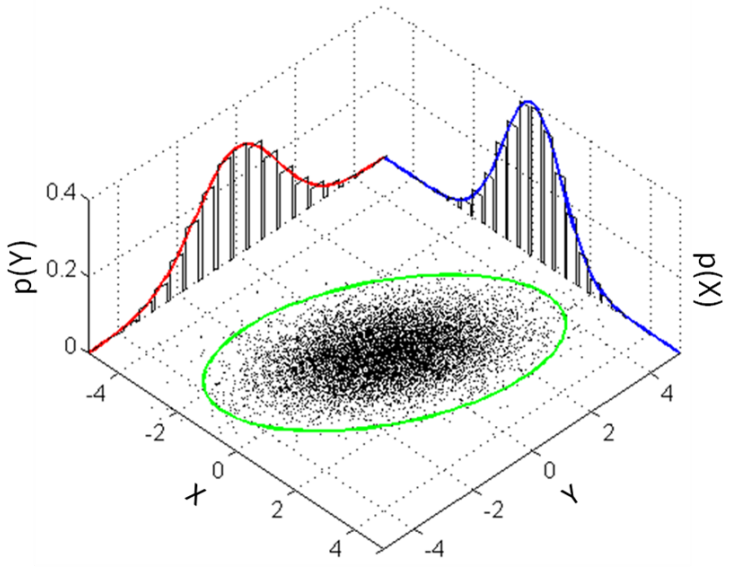

Figure 13: **Joint normal distributions, with $\rho > 0$ and the marginals on the axis**

Definition 3.23 (Gaussian vector/Gaussian process)**.**
A vector $\mathbf{X} = (X_1, \ldots, X_K)$ *or even any collection* $\{X_n\}_n$ *is a Gaussian vector/Gaussian process, if all the linear combination of its elements follows a Gaussian distribution:*

$$Y = \sum_{i=1}^{K} \lambda_i X_i \sim \mathcal{N}(\mu, \sigma^2)$$

Consequences:

- If $\mathbf{X}$ is Gaussian then $\Rightarrow$ $X_i, \forall\, i$ are Gaussian
- The converse is not true: let $X \sim \mathcal{N}(0,1)$ and $\varepsilon \sim$ Rademacher (i.e. $\varepsilon = +1$ or -1 each with probability $1/2$. Now consider the vector $\mathbf{X} = (X, \varepsilon X)$. The sum of the two elements is $= 0$ with probability $1/2$, surely not a Gaussian!

- These collections of non-gaussian vectors of normal r.v. generally sums to "mixture models" (check it out!)
- The converse is true if the elements $X_i, \forall\, i$ are independent. This is because the the sum of independent Normal random variables is always normal.

Theorem 3.32.
Let $\mathbf{X} = (X_1, \ldots, X_K)$ be a Gaussian vector. The random variable $X_1, \ldots, X_K$ are independent if and only if their covariance (taken two-by-two) is null ; or equivalently, when their variance covariance matrix is diagonal.
Note: Careful: it is not enough to say that two *individual* gaussian variables are independent iff they are uncorrelated! Again take the counter example after the definition! The "couple" needs to be Gaussian vector/collection ! (which isn't obvious!)

Other tricks

Here is a list of other useful information (for the core sequence) about Normal distribution:

- Two sided truncation: Let $X \sim \mathcal{N}(\mu, \sigma^2)$ Let $\alpha = (a - \mu)/\sigma$ and $\beta = (b - \mu)/\sigma$. Then:

$$\mathbb{E}(X \mid a < X < b) = \mu + \sigma \frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}$$

and

$$\mathbb{V}\text{ar}(X \mid a < X < b) = \sigma^2 \left[1 + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} - \left(\frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}\right)^2\right]$$

- Moreover, when the truncated has one tail (respective truncated on the lower tail (LHS) or on the upper tail (RHS), the expectation rewrite :

$$\mathbb{E}(X \mid a < X) = \mu + \sigma \frac{\phi(\alpha)}{1 - \Phi(\alpha)} \qquad \text{or } \mathbb{E}(X \mid X < b) = \mu - \sigma \frac{\phi(\beta)}{\Phi(\beta)}$$

  You will use these Normal distribution in Metrics 3 for the Heckman correction.
- Slightly different topic: the ratio above (RHS)

$$h(x) = \frac{\phi(x)}{\Phi(x)}$$

  is called the hazard rate (i.e. the pdf over the cdf). This is defined as :

$$h(x) = \lim_{\delta \to 0} \frac{1}{\delta} \mathbb{P}[x < X \leq x + \delta \mid X > x]$$

  Moreover, $\frac{\phi(x)}{1-\Phi(x)}$ is the inverse-hazard rate (useful for Price theory 3 for Monopoly screening problems). Also $h(x)$ is the inverse of the Mills ratio: $m(x) = \frac{\Phi(x)}{\phi(x)}$. For the normal distribution, this ratio converges to x (i.e. $\frac{\phi(x)}{\Phi(x)} \sim x$, when $x \to \infty$. For the uniform $\mathcal{U}([a, b])$ this ratio is $\frac{x-a}{b-a}$

- Coming back on Normal distribution, check out the Characteristic function and the Laplace transform (expectation of the log-normal), defined above!

Comprehension questions (In progress)

- Find a example of Normally distributed r.v. that are uncorrelated but not independent.
- Work out the very easy proof of corollary 3.31
-
- Find a counterexample where two random variables are uncorrelated but not independent

3.6 Conditional expectation

In this section we are interested in knowing the expectation of a random variable, conditional on the information contained in a (sub)-σ-algebra $\mathcal{G}$/another random variable. *Careful!* Conditional expectation is a mathematical object that tends to be greatly misunderstood by economists (and students in mathematics).

Example 3.12 (Motivating example)**.**
Suppose two random variables associated with dices: $Y, Z : \Omega \to \{1, \dots, 6\}$ *are independant random variables with the same distribution over these numbers* $\mathbb{P}(Y = i) = \mathbb{P}(Z = i) = 1/6, \forall\, i$. *The expectation is obviously* $\mathbb{E}[Y] = \mathbb{E}[Z] = 7/2$. *Now let us consider the sum* $X := Y + Z$, *with linearity of expectation* $\mathbb{E}[X] = \mathbb{E}[Y] + \mathbb{E}[Z] = 7$.
Now suppose that the result of the 2nd dice is known (Z is determined), and we want to compute the mean of X conditional/knowing Z, i.e. $\mathbb{E}[X|\textit{knowing } Z]$. *We can say naively (but rightly so:)*

$$\mathbb{E}[X|\textit{knowing } Z] = \mathbb{E}[Y|\textit{knowing } Z] + \mathbb{E}[Z|\textit{knowing } Z]$$

- *As Y is independent (and hence doesn't depend on) the value of Y, it doesn't change the usual result and we just have* $\mathbb{E}[Y|\textit{knowing } Z] = \mathbb{E}[Y] = 7/2$.
- *Moreover,* $\mathbb{E}[Z|\textit{knowing } Z]$ *isn't less obvious: it should be* $\mathbb{E}[Z|\textit{knowing } Z] = Z$.

As a result, after linearity of expectation (conditional or not), we hence have :

$$\mathbb{E}[X|\textit{knowing } Z] = \mathbb{E}[X|\sigma(Z)] = Z + 7/2$$

where $\sigma(Z)$ *is the* σ*-algebra generated by the information provided by Z (i.e. the set of events that have happened and could happened).*

Definition 3.24 (Conditional expectation w.r.t to a σ-algebra)**.**
Let $(\Omega, \mathscr{F}, \mathbb{P})$ *a measure space, and* $\mathscr{G} \subset \mathscr{F}$ *a sub-*σ*-algebra, we can define the* ***conditional expectation*** *of* Y *with respect to* $\mathscr{G}$, *denoted* $\mathbb{E}(Y|\mathscr{G})$. *We define it as any* variable m_Y *checking the two following conditions:*

- *(i)* m_Y *is* $\mathscr{G}$*-measurable*
- *(ii)* $\forall A \in \mathscr{G}$, $\mathbb{E}(Y\, \mathbb{1}_A) = \mathbb{E}(m_Y\, \mathbb{1}_A)$

Therefore $\mathbb{E}(Y|\mathscr{G})$ *is a random variable ! It is* not *a number !*
Note: The two defining properties can be intuitively translated as follow:

(i) For all the different values of m_Y (and thus $\mathbb{E}(Y|\mathscr{G})$), there exists a corresponding event in $\mathscr{G}$ (if there is not such event, then Y is not $\mathscr{G}$-measurable)

(ii) Along each of the events A in the information set $\mathscr{G}$, the value of Y (and thus $\mathbb{E}(X|\mathscr{G})$) is the same as the averaged value of X.

Theorem 3.33 (Existence and uniqueness)**.**
For all random variable Y, and any σ-algebra $\mathscr{G}$, there exists a random variable $m_Y = \mathbb{E}[Y|\mathscr{G}]$. Moreover, the conditional expectation is defined uniquely almost-surely, i.e. if there are two random variable $m_Y = \mathbb{E}[Y|\mathscr{G}]$ and $\widetilde{m}_Y = \mathbb{E}[Y|\mathscr{G}]$, then

$$\mathbb{P}[m_Y = \widetilde{m}_Y] = 1$$

Note: The existence relies of the Theorem of Radon Nikodym.

Conditional expectation - Important properties

- If Y is $\mathscr{G}$-measurable, then $\mathbb{E}(Y|\mathscr{G}) = Y$ Y is a $\mathscr{G}$-measurable random variable, (that implies that Y cannot have more information than $\mathscr{G}$), therefore it implies that the information contained in Y is redundant with the information of $\mathscr{G}$ and Y averaged on all events of $\mathscr{G}$ equal Y.
- If Y is independent of $\mathscr{G}$, then $\mathbb{E}(Y|\mathscr{G}) = \mathbb{E}(Y)$

  Independence is here the opposite of measurability: no info contained in $\mathscr{G}$ can inform us on the value of W, therefore averaging W on each event of $\mathscr{G}$ ends up as the same thing as averaging over the whole space, i.e. $\mathbb{E}(Y)$ (which is the only value of the random variable $\mathbb{E}(Y|\mathscr{G})$).
- (Law of iterated expectations) $\mathscr{H} \subset \mathscr{G}$ (both sub-σ-algebra)

$$\Rightarrow \qquad \mathbb{E}(\mathbb{E}(Y|\mathscr{G})|\mathscr{H}) = \mathbb{E}(Y|\mathscr{H})$$

  If the info in $\mathscr{H}$ is smaller than the info in $\mathscr{G}$ (which is smaller than the info in $\mathscr{F}$), then averaging w.r.t. $\mathscr{H}$ ends up taking only the smaller set of info available (and it doesn't change anything if you have a more (or less) refined variable inside the sign $\mathbb{E}(\cdot|\mathscr{H})$

  As a result, $\mathbb{E}(\mathbb{E}(Y|\mathscr{G})) = \mathbb{E}(Y)$
- "Extreme" conditional expectation. Let $\mathscr{F}$ the largest σ-algebra, and $\mathscr{T} = \{\emptyset, \Omega\}$ the (most) trivial σ-algebra. We have :

$$\mathbb{E}[Y|\mathscr{F}] = Y \qquad\qquad \mathbb{E}[Y|\mathscr{T}] = \mathbb{E}[Y]$$

  First, any random variable is always measurable w.r.t. the largest $\mathscr{F}$, so we can take it out of the expectation because all the information is already provided. Second, any random variable is always non-measurable w.r.t the trivial σ-algebra $\mathscr{T}$ (except constant!) Hence the conditional expectation is a constant, i.e. $\mathbb{E}[Y]$

Conditional expectation w.r.t. a random variable

Definition 3.25.
Let Y, X two random variables $X : \Omega \to E$ and $Y : \Omega \to \widetilde{E}$. We define the conditional expectation of Y with respect to X, denoted $\mathbb{E}(Y|X)$ as conditional expectation w.r.t. $\mathscr{G} = \sigma(X)$, i.e. $\mathbb{E}(Y|X) = \mathbb{E}(Y|\sigma(X))$. Again this expectation is a random var. depending on X!

Basically, we can just rewrite the above properties given by the σ-algebra $\sigma(X)$.

Proposition 3.34 (Properties)**.**
Let $m_Y(X) = \mathbb{E}(Y|X) = \mathbb{E}[Y|\sigma(X)]$

1. *We can write* $\mathbb{E}(Y|X) = g(X)$
2. *If* Y *is* $\sigma(X)$*-measurable / i.e. if* $Y = h(X)$ *(cf definition of measurability above), then*

$$\mathbb{E}(Y|X) = \mathbb{E}[h(X)|\sigma(X)] = h(X)$$

3. *Similarly, with* $Y = h(X)$ *and for all random variable* Z*, we have*

$$\mathbb{E}(YZ|X) = \mathbb{E}[h(X)Z|\sigma(X)] = h(X)\mathbb{E}[Z|\sigma(X)]$$

4. *If* Y *is independent of* X*, i.e. independent of* $\sigma(X)$*, then* $\mathbb{E}[Y|X] = \mathbb{E}[Y]$
5. *Similarly, if* Y *is mean-independent of* X*, i.e. its conditional expectation is a constant and equal its unconditional expectation, i.e.* $\mathbb{E}[Y|\sigma(X)] = c = \mathbb{E}[Y]$*, then :*

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[c] = c = \mathbb{E}[Y]$$

   (note that independence implies mean-independence (c.f. 4th point!) but the converse is not true! Mean independence is a condition about a particular moment.
6. *Law of iterated expectation: Let* X_0 *be a* $\sigma(X)$*-measurable function, but* X *is not* $\sigma(X_0)$*-measurable – i.e.* X *embeds more information than* X_0*, then we have:*

$$\mathbb{E}\Big[\mathbb{E}(Y|X)\big|X_0\Big] = \mathbb{E}\big[Y\big|X_0\big]$$

Theorem 3.35 (Conditional Expectation as Orthogonal projection)**.**
Let Y *a random variable that has a 2nd moment* $\mathbb{E}[Y] < \infty$*. Then* $m_Y(X) = \mathbb{E}[Y|X]$ *is the minimizer of the loss function:*

$$m_Y(X) \in \operatorname{argmin} \mathbb{E}\Big[\big(Y - m(X)\big)^2\Big]$$

Note that instead of considering $\sigma(X)$ *and* $m(X)$*, we can condition* $\mathscr{G}$ *and search for a random variable* $m(\cdot)$ *that satisfy the* $\mathscr{G}$*-measurability constraint.*

Proposition 3.36 (Discrete conditioning and link with Bayes rule)**.**
Recall Bayes rule : For two events A *and* B*, the conditional probability is given by the Bayes rules, which can be extended with law of total probabilities*

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\sum_i \mathbb{P}(A|B_i)\mathbb{P}(B_i)}$$

Now, let's come back on conditional expectation, with $\mathbb{E}[Y|X] = g(X)$ *and* $Y : \Omega \to \widetilde{E}$*. Let* X *be a* discrete *random variable on* $\{x_1, \dots, x_N\}$*. Then we can characterize the function* $g(\cdot)$

with :

$$g(x) = \frac{\mathbb{E}[Y\mathbb{1}\{X = x\}]}{\mathbb{P}(X = x)}$$

We can denote this function, and rewrite it with conditional probability measure:

$$g(x) = \mathbb{E}[Y|X = x] = \int_{\widetilde{E}} yP_{Y|X=x}(dy)$$

where this measure is defined as an image measure of the law of Y conditioning on $\{X = x\}$, i.e.

$$P_{Y|X=x}(B) = \mathbb{P}(\omega\, s.t. \omega \in A \,\&\, A = Y^{-1}(B)|X = x) = P_Y(B|X = x) = \frac{\mathbb{P}(\{dy\} \cap \{X = x\})}{\mathbb{P}(\{X = x\})}$$

where the end of the line is an abuse of notations (but to make it look more similar to the Bayes rule).

Proposition 3.37 (Continuous conditioning)**.**
Let (X, Y) a couple of random variables and let $f_{(X,Y)}(x, y)$ its density. Let $f_Y(y)$ the marginal density of Y. We "define" (informally) the "conditional density" :

$$f_{Y|X=x}(y) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}$$

As a result, we can write the conditional expectation

$$\mathbb{E}[Y|X = x] = \int_{\widetilde{E}} yP_{Y|X=x}(dy) = \int_{\widetilde{E}} f_{Y|X=x}(y)y(dy) = g(x)$$

and again $\mathbb{E}[Y|X] = g(X)$ is the random variable-conditional expectation.
We can make the notation $f_{Y|X=x}(y)$

$$\begin{aligned} f_{Y|X=x}(y) &= \lim_{\mathrm{d}y\to 0,\mathrm{d}x\to 0} \frac{\mathbb{P}(Y \in [y, y + \mathrm{d}y], X \in [x, x + \mathrm{d}x])/(\mathrm{d}y\mathrm{d}x)}{\mathbb{P}(X \in [x, x + \mathrm{d}x])/\mathrm{d}x} \\ &= \lim_{\mathrm{d}y\to 0,\mathrm{d}x\to 0} \mathbb{P}(Y \in [y, y + \mathrm{d}y] \mid X \in [x, x + \mathrm{d}x])/\mathrm{d}y \end{aligned}$$

That's why $f_{Y|X=x}$ is often called conditional density of Y given $X = x$ (sometimes written $f_{Y|X=x}(y) = f(y|X = x) = f(y|x)$ by economists). Similarly, the notation $\mathbb{E}(Y \mid X = x)$ is frequent to denote this function $g(x)$ (and in this case this conditional expectation is a number)
Note: This result is due to a clever application of Fubini's theorem.

Example 3.13 (Bayesian statistics)**.**
Let θ be a (vector of) parameter and let X a r.v. with law parametrized by θ, in the sense that its pdf is $f(x|\theta)$. Consider a sample $\{X_1, \ldots, X_n\}$ with some value $\{x_1, \ldots, x_n\}$. We define the likelihood as :

$$\mathcal{L}(\theta|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

We say that Bayesian statistics "invert" the relationship between the parameter θ and the data $\{x_1, \dots, x_n\}$ (compared to the frequentist approach) in the same way as Bayes rule invert the relation between A and B (c.f. proposition 3.36).
In this context the parameter has a distribution that is informed by the data, and depending on the likelihood of the parameter $\mathcal{L}(\theta|x)$. The "posterior" distribution is given by the Bayes rule (with continuous data and values), given a prior – the ex-ante distribution of θ before knowing the data. In the case of $n = 1$

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\widetilde{\theta})\pi(\widetilde{\theta})d\widetilde{\theta}}$$

with $f(x) = \int_\Theta f(x|\widetilde{\theta})\pi(\widetilde{\theta})d\widetilde{\theta}$ uses the law of total probabilities. Now considering the entire sample:

$$\pi(\theta|x_1, \dots, x_n) = \frac{\mathcal{L}(\theta|x_1, \dots, x_n)\pi(\theta)}{\int_\Theta \mathcal{L}(\theta|x_1, \dots, x_n)\pi(\widetilde{\theta})d\widetilde{\theta}}$$

Example 3.14 (Gaussian vector)**.**
Let (X, Y) be a Gaussian vector in $\mathbb{R}^{K+M}$

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0}_K \\ \mathbf{0}_M \end{pmatrix} \begin{bmatrix} S_{XX'} S_{XY'} \\ S_{YX'} S_{YY'} \end{bmatrix}\right)$$

Question: Can one compute the conditional distribution of $Y|X$?
Rules of measurability told us that $\mathbb{E}[Y|X]$ should be a function X.
Theorem: *The nice properties of Gaussian vector yield a linear relation, with a matrix A:*

$$Y|X \sim \mathcal{N}(AX, S_{YY'|X})$$

where A and $S_{YY'|X}$ are given by:

$$\begin{cases} A &= S_{YX'} S_{XX'}^{-1} \\ S_{YY'|X} &= S_{YY'} - S_{YX'} S_{XX'}^{-1} S_{XY'} = S_{YY'} - A S_{XX'} A' \end{cases}$$

Hence the conditional expectation is :

$$\mathbb{E}(Y|X) = S_{YX'} S_{XX'}^{-1} X$$

Let us uncover the simpler case (but very frequent in the core sequence):

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}\right)$$

Hence the previous theorem yields : $A = \alpha = \frac{\rho\sigma_X\sigma_Y}{\sigma_X^2} = \frac{\rho\sigma_Y}{\sigma_X}$

$$Y|X \sim \mathcal{N}(\rho\frac{\sigma_Y}{\sigma_X} X, \Sigma_{Y|X})$$

with $\Sigma_{Y|X} = \sigma_Y^2 - \rho^2\left(\frac{\sigma_Y}{\sigma_X}\right)^2\sigma_X^2 = \sigma_Y^2(1-\rho^2)$ *You discount more the variance if the two variables are more correlated (less independent!).*

Two other important theorems for optimization

Theorem 3.38 (Jensen's inequality)**.**
Let $X : (\Omega, \mathscr{F}, \mathbb{P}) \to \mathbb{R}$ *be a real random variable and* ϕ *be a real* convex *function. Suppose* X *and* $\varphi(X)$ *are integrable (i.e.* $\mathbb{E}(|X|) < \infty$ *and* $\mathbb{E}(|\phi(X)|) < \infty$*). Then:*

$$\phi(\mathbb{E}(X)) \leq \mathbb{E}(\phi(X))$$

This also holds also for conditional expectations, for any $\mathscr{G} \subset \mathscr{F}$ *sub-*σ*-algebra:*

$$\phi(\mathbb{E}(X|\mathscr{G})) \leq \mathbb{E}(\phi(X|\mathscr{G}))$$

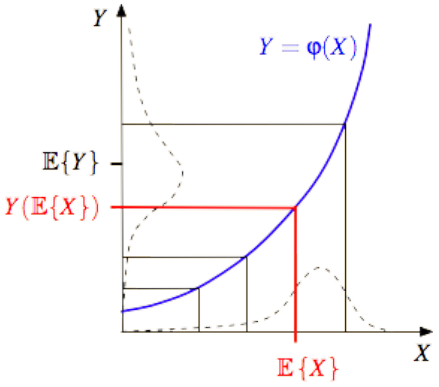

Figure 14: Illustration of the Jensen's inequality

Theorem 3.39 (Interchanging differentation and expectation)**.**
On $(\Omega, \mathscr{F}, \mathbb{P})$*, and* I *an interval in*$\mathbb{R}$*, let define* $\varphi : I \times \Omega \to \mathbb{R}$ *be a measurable function. If it satisfies:*

1. *For every* $x \in I$*, the random variable* $\varphi(x, \cdot)$ *is integrable,*
2. $\frac{\partial\varphi(x,\omega)}{\partial x}$ *exists at every* $x \in I$
3. *There exists* Y *an integrable random variable such that,* $\forall x \in I : \left|\frac{\partial\varphi(x,\omega)}{\partial x}\right| \leq Y(\omega)$

Then, the function $\Phi(x) = \mathbb{E}(\varphi(x, \cdot))$ *is well defined and differentiable at every* $x \in I$*, with:*

$$\Phi'(x) = \mathbb{E}\left(\frac{\partial\varphi(x, \cdot)}{\partial x}\right)$$

3.7 Numerics : Monte-Carlo based methods

4 Statistics

Let us start with a toy example : suppose we flip a coin n times, and measure the results $(x_1, \dots, x_n)$ of 1 for heads and 0 for tail (we can imagine an econ example where a worker is looking for job and is being hired or not. We can "consider" the values – the sample – $(x_1, \dots, x_n)$ as the realizations of the random variables $X_1, \dots, X_n$ independent and identically distributed (*i.i.d*) with a Bernoulli distribution with parameter $\theta \in (0, 1)$ (the probability of head (or being hired): we write $(X_1, \dots, X_n) \overset{\text{iid}}{\sim} \mathcal{B}(\theta)$. Starting from this sample we want to measure θ. That's the most basic question of statistical inference.

What is the difference between Probability and Statistics now? In probability theory, θ is given and known, so the probability law of $X = X_i$, i.e. P_X is also known and we can answer questions like "what is the law of $S_n = \sum_i X_i$?, what is the limit $\lim_{n\to\infty} S_n/n$?, etc. In statistical inference, it is the reverse. You start from the data $(x_1, \dots, x_n)$ and try to deduce characteristics of the law of X i.e. P_X. Many of the tools are the same, the definition of random variables (as function on the sample space), asymptotic theorems like LLN and CLT, classical inequalities (Markov, etc.), properties of the most standard probability distribution (Normal, exponential, student, etc.) that we briefly introduced above (note that the previous section is far from complete compared to a full semester class on probability).

4.1 Statistical models

Statistical inference usually includes two steps : the first relates to modelling, something us economists are good at, given the theoretical roots of the discipline! It consists of formalizing a real phenomenon with a mathematical structure, typically a probability distribution P_X that can be unknown, but belongs to a collection of parametrized distributions $(P_\theta)_{\theta\in\Theta}$ that is specified by the modeller.

Given that first step, the second step consists of the inference properly defined: given the family $(P_\theta)_{\theta\in\Theta}$ and the observation $\mathbf{X} = (X_1, \dots, X_n)$, we look for the best information on the model parameters, i.e. the law P_X. Recall that $\mathbf{X}$ is a stochastic object (random variable, vector, process) with value in a measurable space $(E, \mathscr{E})$. Its law is defined on all sets $A \in \mathscr{E}$ such as:

$$P_X(A) = \mathbb{P}(\mathbf{X} \in A) = \mathbb{P}(\{\omega \in \Omega : \mathbf{X}(\omega) \in A\})$$

i.e. the probability (i.e. the "measure"/size of the sample space) that $\mathbf{X}$ falls in this set A.

Definition 4.1 (Statistical experience)**.**
A statistical experience is the data given by (i) the random object $\mathbf{X}$ *with values in the space* $(E, \mathscr{E})$*, and (ii) the family of probability distribution* $(P_\theta)_{\theta\in\Theta}$ *that is assumed to contain the law* $P_{\mathbf{X}}$*, that is call the statistical model for the law* $\mathbf{X}$*.*

Note: The law $P_{\mathbf{X}}$ of random object $\mathbf{X}$ is called Data Generating Process, that is assumed to follow a particular form (the model we assume!)

In this definition, the fundamental assumption is that there exists a value $\theta \in \Theta$ such that $P_{\mathbf{X}} = P_\theta$. The parameter is unknown a priori, but the space Θ is known.

Definition 4.2 (Parametric model)**.**
If the space Θ of model parameters of $(P_\theta)_{\theta\in\Theta}$ is a space contained in $\mathbb{R}^k$ for a given $k \in \mathbb{N}$ we define it as a parametric model. If not, it is non-parametric.

Example 4.1.
Consider three basic example of labor economics:

- *In a given population, the wages of people is modeled with a normal distribution, with mean and variance that are unknown. We estimate them with a sample of n persons taken randomly in the population. We consider $E = \mathbb{R}^n$ with the Borel $\sigma-$algebra : $\mathscr{E} = \mathcal{B}(\mathbb{R}^n)$. The random object is the n-tuple $\mathbf{X} = (w_1, \ldots, w_n)$ with w_i i.i.d. with a Normal distribution $\mathcal{N}(\mu, \sigma^2)$. In this case $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R}\times\mathbb{R}^\star_+$. The parametrized family is hence*

$$(P_\theta)_{\theta\in\Theta} = \left(\mathcal{N}(\mu, \sigma^2)^{\otimes n}\right)_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}^\star_+}$$

  Suppose the law is not assumed to be Normal or not assumed to follow any known distribution, on, say, $E = [0, \bar{w}]$. In this case Θ is the set of probability distribution on $[0, \bar{w}]$, which is clearly an infinite dimensional space. Usually, since this set is too big (poor Stata), we put assumptions on the set of distributions, for example regularity properties.
- *Consider that the wages depends on observable characteristics (ages, diplomas, location of residence): now $\mathbf{X} = (w_1, \ldots, w_n)$ has the law modeled as w_i i.i.d. with a Normal distribution $\mathcal{N}(X_i\beta, \sigma^2)$, where X is a m-vector of observables (which are also random) and β a vector of unknown parameters. This is synonymous for the regression equation*

$$w_i = X_i\beta + \varepsilon \qquad\qquad \varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

  That's the linear Gaussian model (c.f. next subsection). In this case $\theta = (\beta, \sigma^2)$ and $\Theta = \mathbb{R}^m\times\mathbb{R}^\star_+$. Note that the observables are given and treated as "constant" in the sense that their distribution P_{X_i} are not assumed to follow any law a priori. Apologies for using twice the letter X (one for the outcome/dependent variable – following my formalism in the definition – and one for the independent variables – following the standard formalism in econometrics. Note that the probability law P_{w_i} is itself a random variable that is a function of observables X!
- *In the case where the data on wages are biased – since the workers have "selected themselves" into employment due to intrinsic characteristics – the Heckman correction model assumes a method with two steps:*
  First, the probability of working is specified with a Probit regression of the form:

$$\mathbb{P}(D = 1|Z) = \Phi(Z^T\gamma)$$

where Z is a vector of observables – i.e. explanatory characteristics of working – $D = 1$ indicates employment, γ is an unknown vector of parameters and $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution. Note that the conditional probability is itself a random variable that is a function of observables Z! Said differently, the Probit model can be reformulated as a latent variable model, with an auxiliary variable of the form:

$$D^\star = Z^T\gamma + u \qquad\qquad u \overset{iid}{\sim} \mathcal{N}(0,1)$$

The employment status D can be viewed as an indicator for whether this latent variable is positive:

$$D = \begin{cases} 1 & Y^\star = Z^T\gamma + u > 0 \\ 0 & otherwise \end{cases}$$

Second, the wage equation model is specified as:

$$w_i^\star = X_i^T\beta + \varepsilon \qquad\qquad \epsilon \overset{iid}{\sim} \mathcal{N}(0,\sigma_\varepsilon)$$

where $w^\star$ denotes an underlying wage offer, which may not be observed if the respondent does not work. The conditional expectation of wages given the person works is then

$$w_i = \mathbb{E}[w|X, D=1] = X_i^T\beta + \mathbb{E}[\epsilon|X, D=1]$$
$$w_i = \mathbb{E}[w|X, D=1] = X_i^T\beta + \rho\sigma_\varepsilon\lambda(Z^T\gamma)$$

where $\lambda(Z^T\gamma)$ is the inverse Mills ratio evaluated at $Z^T\gamma$, and ρ is the correlation between unobserved determinants of the propensity to work u and unobserved determinants of the wage offer ϵ.
This equation demonstrates Heckman's insight that sample selection can be viewed as a form of omitted-variables bias.
Exercise: *given the 2-steps nested structure we didn't write the stastistical model for P_{w_i} but you should do it* :D

Note: Every statistical model is an approximation of reality (oh really?). In the case where we assume that wages follow a Normal distribution, there is an inherent inconsistency with the fact that wages are usually not negative. This could be that the model is not well suited for wages. However, this is not really a good argument since Normal distribution have a super low probability of taking extreme value: the probability that $X \sim \mathcal{N}(0,1)$ is outside $[-8,8]$ is around 10^{-15} (even with a billion data points, the chance of being below -8 is one in a million!).

Another super important point (especially in the applied micro group of UChicago) relates to the relationship between model parameters and the Data generating process, and this defined without ambiguity. That's the next definition:

Definition 4.3 (Identifiability)**.**

The statistical model $(P_\theta)_{\theta\in\Theta}$ on $(E,\mathscr{E})$ is identifiable if the application $\theta \mapsto P_\theta$ is injective, i.e. if two parameters $\theta_1 \neq \theta_2$ can not correspond to the same law $P_{\theta_1} = P_{\theta_2}$.
Note: Be careful with non-identified models! or you can get caught in a Norwegian storm during seminars (your Nobel prize won't save you!)

Definition 4.4 (Statistics and estimators)**.**
A statistic $T(X)$ is a measurable function of the random object $\mathbf{X}$ (and eventually of other known parameters) that does not depend on θ. An estimand is the target value, taken for the entire population, of the parameter θ as function of the probability $P_{\mathbf{X}}$. An estimator of θ is a particular statistic $\hat{\theta} = \hat{\theta}(\mathbf{X})$ that is aimed at approximating the estimand of θ.
Note: Let $(X_1,\dots,X_n) \sim P_X$, the estimand for the first moment is denoted $\mu(P)$ in the class of A. Shaikh. The estimator is the sample mean $\bar{X}_n = \sum_{i=1}^n$ and the estimate is the value this estimator takes for a particular sample.

Definition 4.5 (Loss and decomposition biais-variance)**.**
Given an statistical experience such that $\Theta \subseteq \mathbb{R}$, the mean squared error (or the quadratic loss or quadratic risk) of the estimator $\hat{\theta}$ is defined for all θ (the true parameter) as :

$$\mathcal{R}(\hat{\theta},\theta) = \mathbb{E}\big[(\hat{\theta}(\mathbf{X}) - \theta)^2\big] = \int_E \hat{\theta}(\mathbf{x}-\theta)^2 P_\theta(\mathbf{dx})$$

We can decompose this loss between biais and variance:

Proposition 4.1 (Biais-variance tradeoff)**.**

$$\mathcal{R}(\hat{\theta},\theta) = \Big(\mathbb{E}[(\hat{\theta}] - \theta\Big)^2 + \mathbb{E}\Big[\big(\hat{\theta} - \mathbb{E}[\hat{\theta}]\big)^2\Big] =: B(\hat{\theta})^2 + \mathbb{V}ar(\hat{\theta})$$

The term $B(\hat{\theta})$ is the biais of the estimator $\hat{\theta}$. If it is nul, the estimator is unbiased.
Note:

- The biais is the mean error of the estimator and the variance measure fluctuations of this estimator around its average value. An estimator is good if both biais and variance are both low, but there is often a tradeoff!
- The decomposition can be generalized in larger dimension.

Other notions such as consistency of estimators, confidence interval and hypothesis testing are introduced in A. Shaikh's class or any other Metrics 1 class, and not developed here for the sake of time.

4.2 Linear regressions

The principle of regressions is modelling the relation between the dependent variable Y and the independant variables $\mathbf{X} = [X_1, \dots, X_m]'$, also called explanatory variables.

$$Y = \mathcal{G}(\mathbf{X}) = \mathcal{G}(X_1, \dots, X_m)$$

We have a sample of n observations Y_i and n vectors of m dimensions X_i and the aim is to find the function $\mathcal{G}$. The easiest way is when $\mathcal{G}$ is linear, in which case the function is exactly approximated with a set of coefficients $\beta = [\beta_1, \dots, \beta_m]$. In practice of course, the model is assumed to follow a linear approximation or there are measurement errors, we can rewrite the model as:

$$Y = \mathbf{X}'\beta + \epsilon = \beta_1 X_1 + \cdots + \beta_m X_m + \epsilon$$

where ϵ is assumed to be normal with unknown variance σ^2. That's the basic definition of linear regression. We aim at estimating β and σ^2 using statistical inference.

Note: Y and $\mathbf{X}'$ are random variables/m-vectors. For the sample of data, we can stack them in vector/matrix form: $\mathbb{Y} = [Y_1, \dots, Y_n]'$ is $n\times 1$ vector, and $\mathbb{X} = [\mathbf{X}_1, \dots, \mathbf{X}_n]'$ is a $n\times m$ matrix.

$$\mathbb{Y} = \mathbb{X}\beta + \boldsymbol{\epsilon}$$

The assumption of this linear model are : (i) $\text{rank}(X) = m$, i.e. there is no couple of observables that are colinear (for example both the age in month and in years!), and (ii) the error terms ϵ_i are <u>*i.i.d*</u> with $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}_n$ and $\mathbb{V}\text{ar}(\boldsymbol{\epsilon}) = \sigma^2\mathbb{I}_n$, i.e. errors are centered and homoskedasticity.

Definition 4.6 (OLS)**.**
The ordinary least square (OLS) estimator $\hat{\beta}$ is defined as :

$$\hat{\beta} = \underset{\alpha\in\mathbb{R}^p}{\text{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{m} \alpha_j X_{ij}\right)^2 = \underset{\alpha\in\mathbb{R}^p}{\text{argmin}} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i^T\alpha\right)^2 = \underset{\alpha\in\mathbb{R}^p}{\text{argmin}} \left\|\mathbb{Y} - \mathbb{X}\alpha\right\|^2$$

<u>*Note:*</u> There is a geometric interpretation of $\hat{\beta}$: the matrix $\mathbb{X} = [X_1, \dots, X_m]$ of the plane of observables is formed with m-column vector in $\mathbb{R}^n$. The subspace of $\mathbb{R}^n$ generated by these m vectors is called the linear span (spanned space), denoted

$$\mathcal{M}_X = \text{Im}(\mathbb{X}) = \text{span}(\mathbb{X}) = \text{span}(X_1, \dots, X_m)$$

It is of dimension m and every vector of this space (by definiton of the linear span) has the form: $\mathbb{X}\alpha = \alpha_1 X_1 + \cdots + \alpha_m X_m$. Hence, the vector Y is the sum of a element $\mathbb{X}\beta$, that's the closest to Y in the sense of the Euclidian distance (square of the difference, summed on the m dimensions).

This element is unique (the Euclidian distance is convex :p) and is by definition the projection of Y on $\mathcal{M}_X$. This projection is denoted $\hat{\mathbb{Y}} = \mathcal{P}_X \mathbb{Y}$ where $\mathcal{P}_X$ is the orthogonal projection matrix on $\mathcal{M}_X$. Of course, we can also write it as $\hat{\mathbb{Y}} = \mathbb{X}\hat{\beta}$ where $\hat{\beta}$ is the OLS estimator.

As a result, the orthogonal space of $\mathcal{M}_X$ denoted $\mathcal{M}_X^{\perp}$ is called the space of residuals. This space is of dimension (because it's an orthogonal complement) is

$$\dim(\mathcal{M}_X^{\perp}) = \dim(\mathbb{R}^n) - \dim(\mathcal{M}_X) = n - m$$

which is the degree of freedom of our linear model!

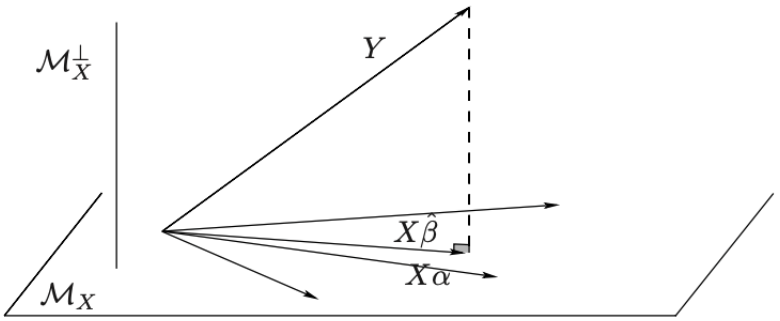

Figure 15: Representation of $X'\hat{\beta}$ in the space of observables

Proposition 4.2 (OLS estimator, formula)**.**
The estimator $\hat{\beta}$ has the expression

$$\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}$$

and the matrix $\mathcal{P}_X$ of orthogonal projection on $\mathcal{M}_X$ is written

$$\mathcal{P}_X = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

<u>Note:</u> The idea of the proof is super simple: one version consists of minimizing the mean squared error $\left\|\mathbb{Y} - \mathbb{X}\alpha\right\|^2$, implying to take the First order condition (thanks to convexity!) and using the fact that $\mathbb{X}'\mathbb{X}$ is invertible, thanks to assumption *(i)* above. Another version relates to the definition of the orthogonal projection on the space $\mathcal{M}_X$

<u>Note:</u> Again, there is much more to say on OLS: unbiaisedness, formula for the variance, estimator for σ, confidence intervals, student law of the estimator of unknown variance, student test for significance of one varia le and Fisher test for significance of all/a group of variables, Gauss-Markov theorem – OLS is the best (in the sense of minimizing the variance) unbiased estimator – consistency and limit distributions, etc. All that is covered in A. Shaikh's class or any other metric class.

Note: Sometimes, linearity is not satisfying, so one might want to take (known) non-linear functions of the linear combination $\mathbb{X}\beta$, and perform compositions to improve the fit. That's the object of a whole branch of machine learning!

4.3 Maximum Likelihood

Let us start with the toy example of coinÂ flipping (or job applications). How to choose the value of the parameter θ – the proba of a head/of being hired – given the data we observe? The idea is to choose the "best" parameter θ to have the maximum chance to observe the realizations $X_i = x_i$ that are in our dataset, given our assumption on the statistical model $(P_\theta)_{\theta\in\Theta}$.

We hence study the likelihood function $\mathcal{L}(\theta|\mathbf{x})$

Definition 4.7 (Likelihood function)**.**

$$\mathcal{L}(\theta|x_1, x_2, \dots, x_n) = P_\theta(x_1, x_2, \dots, x_n) = \begin{cases} \prod_{i=1}^n \mathbb{P}(X_i = x_i) & \text{if } X_i \text{ is a discrete r.v.} \\ \prod_{i=1}^n f(x_i|\theta) & \text{if } X_i \text{ is a continuous r.v.} \end{cases}$$

where the second equality follows from the usual assumption that our data are i.i.d.*, so that we can take the product. The log-likelihood is the logarithm of the likelihood (oh really?):*

$$\ell(\theta|x_1, x_2, \dots, x_n) = \log \mathcal{L}(\theta|x_1, x_2, \dots, x_n)$$

Note:

- The last equality follows from a standard assumption that our class of statistical model $(P_\theta)_{\theta\in\Theta}$ admit a density w.r.t. the Lebesgue measure μ – in this case we say that the statistical model is *dominated* by the Lebesgue measure. In particular the density (p.d.f) follows from the definition of Radon Nikodym (c.f. **??** 3.4) $f(x|\theta) = \frac{dP_\theta}{d\mu}$. In the case where the random variable is assumed to be discrete (as in the case of our Bernoulli coin flip/job search, this is replaced by the probability mass function $p_\theta(x) = \mathbb{P}(X_i = x_i)$]
- We already see that the likelihood function already operates the inversion we talked about in the introduction: statistics goes from the data to the parameter θ, while probability take the parameter as given and obtain the (law/simulation/properties of the) random variables X_i. This inversion logic will be even stronger for Bayesian statistics.

Definition 4.8 (MLE)**.**
In statistical model $(P_\theta)_{\theta\in\Theta}$, we call Maximum likelihood estimator $\hat{\theta}$ the parameter that maximize the likelihood function (it's in the name no?), i.e. such that

$$\mathcal{L}(\hat{\theta}|x_1, \dots, x_n) = \max_{\theta\in\Theta} \mathcal{L}(\hat{\theta}|x_1, \dots, x_n) \Leftrightarrow \max_{\theta\in\Theta} \mathcal{L}(\hat{\theta}|x_1, \dots, x_n) \Leftrightarrow \hat{\theta} = \operatorname*{argmax}_{\theta\in\Theta} \mathcal{L}(\hat{\theta}|x_1, x_2, \dots, x_n)$$

Note:

- Here we put a "max" instead of a "sup" because we implicitely assume that the maximum exists.
- In the last equivalence we put a unique argmax (instead of a multi-elements set) because we assume the conditions for uniqueness (what are they? check the section on optimization in the 1st part of the mathcamp!).
- Moreover, if the log-likelihood is differentiable ($\mathcal{C}^1$) and if the maximum is not obtained on the boundaries of Θ, a necessary condition to find the MLE is the first-order condition:

$$\nabla_\theta \ell(\theta|\mathbf{x})\big|_{\theta=\hat{\theta}} = 0$$

  This is sometimes called the likelihood equation /system of equations. Note that a root of the likelihood may *not* be the MLE! (one of the roots is, though). We need restrictive conditions on the model for this necessary condition to become sufficient!

We will see a common example of "conditional" Maximum likelihood, where we consider a model where $(X_i, Y_i) \overset{[}{i}.i.d.]{\sim} P$ some law, and we assume the parametric model for the conditional law $Y_i|X_i \sim P_\theta$, and P_X is the (independent) law of X.

Example 4.2 (OLS and MLE)**.**
The conditional likelihood is defined as

$$\mathcal{L}(\hat{\theta}|\mathbf{y}, \mathbf{x}) = \prod_{i=1}^{n} f(y_i|\theta, x_i)$$

Let us assume the conditional law is of linear form:

$$Y_i = X_i\beta + \varepsilon \qquad\qquad \varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

The conditional log-likelihood is (do the calculation as an exercise!) for the n-vector of observations $\mathbf{y}$ *and the* $n \times m$ *matrix* $\mathbf{x}$ *– (note that* $\mathbf{x}$ *is the new notation for* $\mathbb{X}$ *from the previous notation, i.e. a* $n \times m$ *matrix!)*

$$\ell(\beta, \sigma^2|\mathbf{y}, \mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2$$

Note that the maximization w.r.t. β *does not depend on* σ*. The FOC gives (that's both a necessary* and *sufficient condition (why?)*

$$\nabla_\beta \ell(\beta, \sigma|\mathbf{y}, \mathbf{x}) = -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i'(y_i - x_i\beta) = \mathbf{0}_{p\times 1}$$

Which gives (tadam!) the same formula for the estimator $\hat{\beta}$ than the OLS estimator:

$$\hat{\beta} = \big(\sum_{i=1}^{n}(x_i' x_i)\big)^{-1} \sum_{i=1}^{n} x_i' y_i \qquad \Rightarrow \qquad \hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$$

Moreover, for the variance, the FOC gives the MLE for σ^2:

$$\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2$$

which is the standard empirical variance for the error terms.

4.4 Generalized Methods of moments

c.f. TA session

5 Stochastic process and stochastic calculus

Definition 5.1 (Filtration)**.**
Given a probability space $(\Omega, \mathscr{F}, P)$, we define a ***filtration*** *$(\mathscr{F}_t)_{t\geq 0}$ as an increasing sequence of sub-σ-algebras.*
$\mathscr{F}_0 \subset \mathscr{F}_{t_1} \subset \mathscr{F}_{t_2} \subset \cdots \subset \mathscr{F}$
Note: In econ or finance, we often call $(\mathscr{F}_t)_{t\geq 0}$ the *information set*, as the knowledge of what can happen (i.e. the set of events that can be measured) grows over time.

Definition 5.2 (Stochastic process)**.**
A stochastic process is a sequence of random variables X_t indexed (and ordered) by their time $t \in T$.
We can pose $(\mathscr{F}_t)_{t\geq 0} \equiv \sigma(X_s : 0 \leq s \leq t)$, which is a filtration generated by the stochastic process (or canonical filtration).
Note: t is an index of time: it can be countable ($t \in \mathbb{N}$) and the time is discrete, or it can be uncountable ($t \in \mathbb{R}$) and the time is continuous. With the use of stochastic calculus, most models in finance are in continous-time.

Definition 5.3 (Adaptability and Predictabillity)**.**
A stochastic process is adapted *w.r.t. $(\mathscr{F}_t)_{t\geq 0}$, if $\forall t$, X_t is $\mathscr{F}_t$-measurable.*
A stochastic process is said to be predicable*, if $\forall t \in \mathbb{N}$, X_t is $\mathscr{F}_{t-1}$-measurable.*
Note:

- If X_t is not $\mathscr{F}_t$-measurable, it often means that X_t contains more (or different) information than $\mathscr{F}_t$
- If X_t is $\mathscr{F}_{t-1}$-measurable, then the knowledge of X_t can be predicted by the information in $\mathscr{F}_{t-1}$ (i.e. predictability)
- It implies that if $(X_t)_t$ is adapted, the knowledge of X_t does not give you *more* information than the information set $\mathscr{F}_t$ (in particular you can't predict the future).
- A stochastic process is always adapted to its canonical filtration.

Example 5.1.
Any sequence of random variables can be a stochastic process.

- *A sequence of deterministic variables (constant accross Ω), such as $X_t = t$ is a stochastic process, but quite boring. Informally, because there is no randomness, it is not even "stochastic".*
- *A sequence of random variables $(X_t)_{t\geq 0}$ which are all following the same law (for example $X_t \sim \mathcal{N}(0,1)$) is also a stochastic process, but not some much interesting neither. Indeed, informally, there is no dynamics (always the same law), but this represents the baseline for "stationary processes".*

$\Rightarrow$ *Researchers in probability are looking for processes that "behave well", i.e. which have some structure and probability law that vary or have constant properties over time and that are simple to study.*

- *Two "simple" processes are i)* martingales*, and ii)* Markov process

Definition 5.4 (Link with the conditional expectation)**.**
In economics, the conditional expectation w.r.t. a σ-algebra from a filtration $(\mathscr{F}_t)_{t\geq 0}$ is a crucial tool. It is denoted compactly by the operator $\mathbb{E}_t$

$$\mathbb{E}_t(X) \equiv \mathbb{E}(X|\mathscr{F}_t)$$

Moreover, the Law of iterated expectations rewrites :

$$\mathscr{F}_t \subset \mathscr{F}_{t+1} \textit{(both sub-}\sigma\textit{-algebra)} \qquad \Rightarrow \qquad \mathbb{E}(\mathbb{E}(X|\mathscr{F}_{t+1})|\mathscr{F}_t) = \mathbb{E}(X|\mathscr{F}_t)$$

or in short: $\mathbb{E}_t(\mathbb{E}_{t+1}(X)) = \mathbb{E}_t(X)$
Note: $\mathbb{E}_t(X)$ is ***not*** a number, but rather a function of the different shocks present in $\mathscr{F}_t$ (in economics: TFP shocks – aggregate or idiosyncratic – or policy shocks)

Example 5.2.
For a stochastic processes evolving over time:

- *If X_t is adapted, then $\mathbb{E}_t(X_t) = X_t$*
- *If ε_t is idiosyncratic, i.i.d., mean zero and not predictable, then $\mathbb{E}_t(\varepsilon_{t+1}) = \mathbb{E}(\varepsilon_{t+1}) = 0$.*
- *If X_t is adapted, but not Y_t, $\mathbb{E}_t(X_t\, Y_t) = X_t\ \mathbb{E}_t(Y_t)$*

Example 5.3 (Additional example)**.**
Two examples with graphs:

- *AR(1) process:*

$$X_t = \rho X_{t-1} + \varepsilon$$

with $\varepsilon \sim \mathcal{N}(0,1)$. We hence have a structure of dependence and some randomness. The graph shows this example with two different values of ρ but the same path of exogenous shocks.

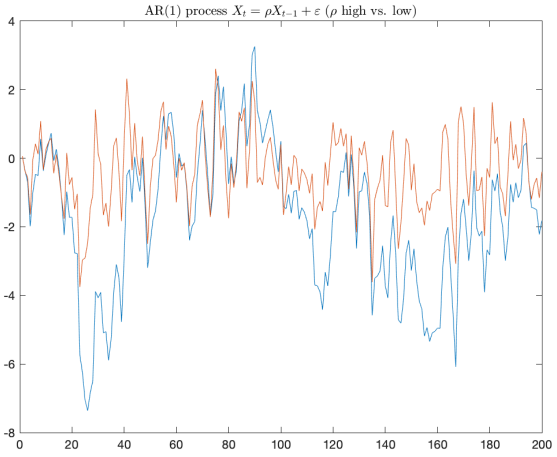

Figure 16: **Example 1 in 5.3.**

- *Stochastic process "closed" by the random variable $Y = \widetilde{X}_T$, where $\{\widetilde{X}_t\}_t$ is itself another stochastic process (we take a random walk here for simplicity).*

$$X_t = \mathbb{E}[Y|\mathscr{F}_t] = \mathbb{E}[X_T|\mathscr{F}_t]$$

If $\widetilde{X}_t = \widetilde{X}_{t-1} + \varepsilon_t$ *a random walk, then we observe that*

$$X_t = \mathbb{E}[X_T|\mathscr{F}_t] = \widetilde{X}_t \qquad \forall t < T \qquad \text{and} \qquad X_s = \mathbb{E}[X_T|\mathscr{F}_s] = \widetilde{X}_T \qquad \forall s \geq T$$

As displayed on the following graph:

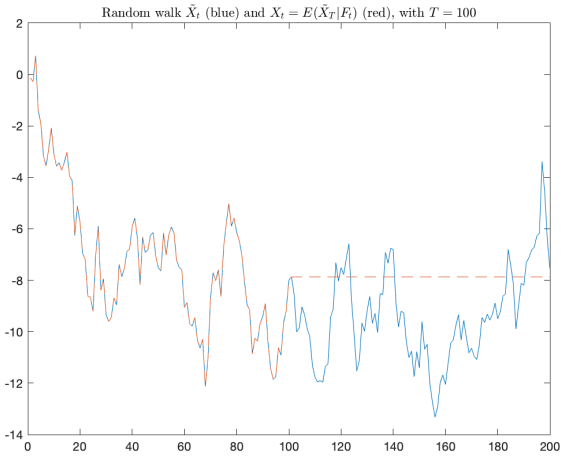

Figure 17: **Example 2 in 5.3**

Stationary and Ergodicity

We finish this section by adding two important definitions for macroeconometrics.

Definition 5.5 (Stationarity)**.**
We say that a stochastic process $\{X_t\}_t$ *is stationary, if it the law/joint distribution of a subset of these random variables* $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ *is the same as the joint distribution, translated in time at a date* τ*. More precisely, the*

$$P_{(X_{t_1},\ldots,X_{t_n})}(x_1,\ldots,x_n) = P_{(X_{t_1+\tau},\ldots,X_{t_n+\tau})}(x_1,\ldots,x_n) \qquad \forall \tau \in \mathbb{R} \qquad \text{and} \forall\ t_1,\ldots,t_n \in \mathbb{R}$$

We sometimes call such processes "measure preserving" stochastic process, because the translation doesn't affect the joint law. This measure-theoric definition yields: $\forall B \in \mathscr{E}$*, we have* $\mathbb{P}(X_t^{-1}(B)) = \mathbb{P}(X_{t+\tau}^{-1}(B))$*. We say that the stochastic process is measure-preserving if* $A = \{\omega \in \Omega \ s.t. X_t \in B\}$ *has the same measure by translation of time* $t' = t + \tau, \forall \tau \in \mathbb{R}$
Note:

- This is the strongest sense of stationary. A weaker sense is to have a constant moment over this translation.
- A natural consequence is simply that the law of a stationary process is the same and doesn't not depend on time t, i.e. with $n = 1$, $P_{X_t} = P_{X_{t+\tau}}\ \forall \tau \in \mathbb{R}$

Next, we need a simple /yet general definition for invariant set.

Definition 5.6 (Invariant sets)**.**
Let $(\Omega, \mathscr{F}, \mathbb{P})$ *a probability space. We say that an event* $A \in \mathscr{F}$ *is invariant w.r.t to the stochastic process* $\{X_t\}_t$ *if for all* $\omega \in A$*, then* $X_t(\omega) \in B \implies X_{t+1}(\omega) \in B$*.*
A measure theory-type of definition says that a set $A \in \mathscr{F}$ *is invariant if* $A := X_{t+1}^{-1}(B) = X_t^{-1}(B)$

Note:

- Said differently, an invariant set is a set where the stochastic process is "trapped": once it is inside this set, it can't go out of it!
- Of course Ω and $\emptyset$ are invariant events.
- The definition here focuses on A, but in applications, the invariant set is often implicitely the one denoted B here!
- We denote (as in Hansen's class) $\mathcal{I}$ the set of all invariant events. This collection $\mathcal{I}$ is a σ-algebra.

We say that a stochastic process is ergodic – in the physical sense – if over (long) periods of time, the time share spent by a system/stochastic process in some regions of the state space equals the probability distribution over this space (i.e. is proportional to its volume). Mathematically, we can make this statement more precise:

Definition 5.7.
Let $(\Omega, \mathscr{F}, \mathbb{P})$ probability space. A stochastic process $\{X_t\}_t$ is P_X-ergodic (or P_X is an ergodic measure), then for $A = X_{t+\tau}^{-1}(B) \subseteq X_t^{-1}(B)$, $\forall \tau$ (i.e. A is invariant) we have $P_X(B) = 0$ or $P_X(B) = 1$ *Note:* In other words there are no invariant strict-subsets (almost-surely).

Theorem 5.1 (Birkhoff Law of Large number)**.**
Let $\{X_t\}_t$ be a stationary/measure-preserving process, we have that the average over time converges to the average over space (i.e. the mean $\mathbb{E}[f(X)|\mathcal{I}]$):

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t(\omega)) \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[f(X)|\mathcal{I}]$$

for all measurable function $f : E \to F$, and $\mathcal{I}$ is the set (σ-algebra) of all invariant events and where the convergence is both almost sure (and hence in probability) and in L^2.

Moreover, if the process is ergodic (measure-preserving and all its invariant events have probability 0 or 1) then $\mathbb{E}[X|\mathcal{I}] = \mathbb{E}[X] < \infty$. And hence we have a Law of Large number, without requiring i.i.d.*!!*

$$\frac{1}{T}\sum_{t=1}^{T} X_t(\omega) \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[X]$$

5.1 Markov chains

Definition 5.8 (Transition matrix)**.**
*We now consider **Markov chains** – a simple example of stochastic process – which are* finite *– i.e. that happen on a finite number of "states".*

- *S the state space is a finite space, with n elements $\{x_1, \dots, x_n\}$*

*The **Transition function**, or transition matrix we can denote $P_t \equiv p_t(x, y)$, is a function $P_t : S \times S \to [0, 1]$ such that*

(i) Each element of $P_t(\cdot, \cdot)$ is non-negative

(ii) $\sum_{y \in S} p_t(x, y) = 1, \forall x \in S$

This means the rows of the matrix some to one.

Note:

- S can be a real value (consumption level, growth rate) or anything else (high or low "states of the world": $h, l \in S$, employed/unemployed, etc.).
- It is easy to see that if P is a transition matrix, then it k-th power $\tilde{P} = P^k$ is also a transition matrix.
- Moreover, since all the rows sum to 1, we have the matrix equation $\mathbb{1}_n = P\mathbb{1}_n$.
- If the transition does not depend on time $p_t(\cdot,\cdot) \equiv p(\cdot,\cdot)$, we say that the Markov chain is *homogeneous*. If it is not, we say it is inhomogeneous. For simplicity and time, we only consider homogeneous Markov chain in the following (except the next definition that is the most general).

Definition 5.9 (Transition function – general definition)**.**
A transition function is a function $p : \mathcal{T} \times S \times \mathcal{G} \to [0,1]$ s.t.

(a) for each $t \in \mathcal{T}, x \in \mathcal{S}$, $p(t,x,.)$ is a probability measure on $(S,\mathcal{G})$;
(b) for each $t \in \mathcal{T}, A \in \mathcal{G}$, $p(t,.,A)$ is a $\mathcal{G}$-measurable function;
(c) (Chapman-Kolmogorov) $\forall s,t \in \mathcal{T}$, $x \in S$ and $A \in \mathcal{G}$,

$$p(t+s,x,A) = \int_S p(s,y,A)p(t,x,dy)$$

Note: For discrete time Markov processes where $\mathcal{T} = \mathbb{N}$, it is enough to specify $p(t,x,A)$ for $t = 1$ and the rest follows from (c). In this case, we write $p(x,A)$.

Definition 5.10 (Markov property)**.**
We say that a discrete-time/discrete-state stochastic process $\{X_t\}_t$ respect the Markov "memoryless" property if:

$$\mathbb{P}(X_{t+1} = y \,|\, X_0, X_1, \ldots, X_t) = \mathbb{P}(X_{t+1} = y \,|\, X_t)$$

A general definition (for uncountable state-space and time-space): A stochastic process $\{X_t\}_t$ if $\forall A \subset S$ and $s < t$:

$$\mathbb{P}(X_t \in A|\mathcal{F}_s) = \mathbb{P}(X_t \in A|X_s)$$

Note:

- In general, c.f. the definition above, $\mathbb{P}(X_{t+1} = y \,|\, X_0, X_1, \ldots, X_t)$ is a measurable function of $X_0, X_1, \ldots, X_t$, not only X_t
- In other words, to forecast the distribution of X_{t+1} on S, the only information need is the current state X_t.
- This is quite general, given that you can consider very large state-space, c.f. example 4 below.

Definition 5.11 (Markov Chain)**.**
A Markov chain *X_t is a sequence of S-valued random variables, with transition matrix P, if, for all $t \geq 0$, and for all $y \in S$ we have:*

$$\mathbb{P}(X_{t+1} = y \,|\, X_0, X_1, \ldots, X_t) = p(X_n, y)$$

Note: Therefore, it satisfies the *Markov property.*

Example 5.4.

A set of examples:

- Example 1: *A worker can be either (i) unemployed or (ii) employed*
  - *When unemployed, he finds a job at rate α*
  - *When employed, he loses its job with probability β*
- *Therefore, the transition matrix writes:*

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

  Note: *Question (exercises?)*

  - *What is the average duration of unemployment?*
  - *Over the long-run, what fraction of time does a worker find herself unemployed?*
  - *Conditional on employment, what is the probability of becoming unemployed at least once over the next 12 months?*

- Example 2: *Hamilton (2005) used the US employment data, and determined the frequency of: (i) Normal growth (ii) Mild recession (iii) Severe recession.*
  The stochastic matrix is estimated such as:

$$P = \begin{pmatrix} 0.971 & 0.029 & 0 \\ 0.145 & 0.778 & 0.077 \\ 0 & 0.508 & 0.492 \end{pmatrix}$$

  It says that, when US are in a severe recession, there is a 50.8 probability to face a mild recession next month, and no chance at all to come back to normal growth.

- Example 3: *Random walk:*

$$\begin{aligned} X_{t+1} &= X_t + \varepsilon_t \\ &= X_0 + \sum_{i=0}^{t} \varepsilon_i \qquad \forall t \geq 0 \end{aligned}$$

  where ε_t are i.i.d. random walk s.t. $\varepsilon_t \sim \mathscr{P}$ (any distribution) with probability mass function ψ – recall that S is finite and thus countable.
  Question: What would be the transition matrix? (Problem set 1!)

- Example 4: *Markov chain may depend on a finite set of event/random variables in the past. Indeed, consider an AR(2) process:*

$$z_t = \rho_1 z_{t-1} + \rho_2 z_{t-2} + \varepsilon_t$$

  Thus, $\{z_t\}_{t=0}^{\infty}$ is not a Markov process. However, it could be written as:

$$\begin{pmatrix} z_t \\ z_{t-1} \end{pmatrix} = \begin{pmatrix} \rho_1 & \rho_2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} z_{t-1} \\ z_{t-2} \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \varepsilon_t$$

Let $\tilde{z}_t = \begin{pmatrix} z_t \\ z_{t-1} \end{pmatrix}$. *Clearly,* $\{\tilde{z}_t\}_{t=0}^{\infty}$ *is a Markov process. The cost is to increasing the state-space, which is not always computationally feasible.*

Recursive formulation and probability distribution of Markov chain

We defined the Markov chain $\{X_t\}_t$. We are now interested by its law, or probability distribution $\mathbb{P}(X_t = x)$ for every x. Recall that we start from the transition kernel.

$$\mathbb{P}(X_{t+1} = x_{t+1} \,|\, X_t = x_t) = p(x_t, x_{t+1})$$

Therefore, knowing the initial state, one can iterate over the transition matrix:

$$\mathbb{P}(X_t = x_t \,|\, X_0 = x_0) = [P^t](x_0, x_t)$$

In other words, initial conditions and transition matrix are the only determinant of the path of X_t.

To know the "marginal distribution" at time t. Let us proceed in a recursive way (over time):

- Knowing the distribution at time $X_t \sim P_X$ (with p.m.f., probability mass function π) and the transition matrix $P \equiv p(x_t, x_{t+1})$, what can we say about the probability of X_{t+1}?
- The solution lies in the *law of total probabilities*:

$$\mathbb{P}(X_{t+1} = x_{t+1}) = \sum_{x \in S} \mathbb{P}(X_{t+1} = x_{t+1} \,|\, X_t = x) \cdot \mathbb{P}(X_t = x)$$

$$\pi_{t+1}(y) = \sum_{x \in S} p(x, y)\, \pi_t(x)$$

  where the second line use the formalism of the transition matrix and the distribution of present states π_t (expressed as a p.m.f., $\pi_t = \Big(\mathbb{P}(X_t = x_1), \ldots, \mathbb{P}(X_t = x_n)\Big)$.

We express the p.m.f. π as a n-values *rows vector* (of probabilities), the n equations become matrices as:

$$\pi_{t+1} = \pi_t\, P$$

This equation represents the *law of motion* of the distribution – something found in countless economics problems. This is a simple version – discrete time and discrete state space – of the famous Kolmogorov equation (that originally describes the evolution of a distribution of particles moving with a stochastic process).

Definition 5.12 (Distribution dynamics: Kolmogorov forward in discrete time)**.**
Consider ψ_t *the distribution a Markov process at time* t*, i.e. the probability measure of* X_t*. Given a transition kernel* $p(x, y)$ *over* $S \times S$

$$\psi_{t+1}(y) = \int_S p(x, y) \psi_t(dx)$$

where P *is thought as an operator (more on that in the section 4.4.)* Coming back on discrete states, similarly, we can derive the Multi-step transition probabilities by iterating it recursively over time.

Indeed, if $\pi_{t+1} = \pi_t P$, therefore, we generalize it:

$$\pi_t = \pi_0 P^t$$
$$\pi_{t+m} = \pi_t P^m$$

Finally, we can also compute these change in distributions when starting from a given state x, for example if it is common knowledge that $X_0 = x$, the initial probability distribution is a Dirac mass point $P_X(\cdot) = \delta_x(\cdot)$ and hence $\pi_t = \Big(0, \dots, 0, \underbrace{1}_{\uparrow x}, 0, \dots, 0\Big)$. Hence the distribution π_t rewrites, again a row vector as :

$$\pi_t = \Big(0, \dots, 0, 1, 0, \dots, 0\Big) P^t = \big[P^t\big](x, \cdots)$$

Example 5.5.

Exercise: *Using the transition matrix on recessions seen before, and considering the today's state as unknown, (you only know the distribution-vector π_t), what is the probability to be in a mild or severe recession in 6 months? Answer:*

$$\mathbb{P}(\textit{recession}) = \pi_{t+6} \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \pi_t \cdot P^6 \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

Conditional expectation of functions of a Markov chain

We are now interested in computing conditional expectation of the type $\mathbb{E}[f(X_{t+1})|X_t]$, which formally writes as

$$\mathbb{E}[f(X_{t+1})|X_t] = \mathbb{E}[f(X_{t+1})|\sigma(X_t)] = \mathbb{E}[f(X_{t+1})|\mathscr{F}_t]$$

Again, this is a random variable $\mathscr{F}_t$-measurable.
Given a *columns vector*:

$$f(X_t) \equiv \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$$

As a result,

$$\mathbb{E}[f(X_{t+1})|X_t] = \sum_{y \in S} f(y) \mathbb{P}(X_{t+1} = y|X_t)$$

Given $\mathbb{P}(X_{t+1} = y|X_t) = P(\cdot, y)$ a column vector, given each present states X_t, we have using the transition matrix :

$$\mathbb{E}[f(X_{t+1})|X_t] = \sum_{y \in S} f(y) p(X_t, y) = Pf$$

by remember that we are talking about a random variable, hence a column vector! As a result, for each of these rows at given x we have:

$$\mathbb{E}[f(X_{t+1})|X_t = x] = \sum_{y \in S} f(y) p(x, y) = P(x, \cdot) f(\cdot) = \big[Pf\big](x)$$

Hence, we can realize that starting at $X_t = x$, we can get a conditional expectation as a number (but otherwise it is not a number, it's a random variable).

$$\cdots \text{———} \cdots$$

For discrete time/discrete space - Markov chains, the formalism you should remember is:

- Functions are expressed in column vector, e.g. $f(X_t) = \{f(x)\}_x \equiv \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$
- Measures/distributions are expressed in row vector $\pi_t = \{\mathbb{P}(X_t = x)\}_x \equiv (\pi_t(x_1) \ \dots \ \pi_t(x_n))$

Let us move on to two of the most important theorems for Markov Chains. But beforehand we introduce the relevant notions:

Definition 5.13 (Irreducibility)**.**
Let $\{X_t\}_t$ a Markov chain on the state space S

- *Two states $x_a, x_b \in S \times S$ are said to communicate with each other if there exist positive integers m and n such that:*
$$P^m(x_a, x_b) > 0 \qquad \text{and} \qquad P^n(x_b, x_a) > 0$$

- *The stochastic/transition matrix is said **irreducible** if all states communicate, i.e. x_a and x_b in $S \times S$ can communicate.*

Note: Question: is the "growth/recession regime matrix" irreducible? If yes why, if no why not?

Definition 5.14 (Aperiodicity)**.**
Let $\{X_t\}_t$ a Markov chain on the state space S

- *The period of a state x_o is the greatest common divisor of the set of integers defined by:*
$$D(x_o) \equiv \{j \geq 1 : P^j(x_o, x_o) > 0\}$$

- *A stochastic matrix is said* aperiodic *if the period of every state is 1, or* periodic *otherwise.*

Note: Example: if $D(x) = \{3, 6, 9, \dots\}$, the period is 3. Question: what is the period of :

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

when $\alpha = 1$ and/or $\beta = 1$?

Stationary distribution:

Definition 5.15.
Some distributions are invariant under the transition matrix. We call these distribution stationary*, if the distribution $\pi^\star$ on S is such that:*
$$\pi^\star = \pi^\star P$$

Note:

- Obviously, an immediate consequence is : $\pi^\star = \pi^\star P^t \quad \forall t$
- Therefore if the random variable X_0 has a stationary distribution, then X_t also have this same distribution.

- Again, we can give a more general definition

Definition 5.16 (Stationary distribution)**.**
The stationary distribution *associated with a Markov process X_t is a probability measure π over $(S, \mathcal{G})$ such that*

$$\int_S p(t, x, y)\, \pi(dx) = \pi(y), \ \forall t$$

Example 5.6.
Let take again the simple example given by the transition of growth regimes in the US economy. We compute the stationary distribution by iterating on the Law of Motion of the distribution (i.e. the Kolmogorov Forward equation).

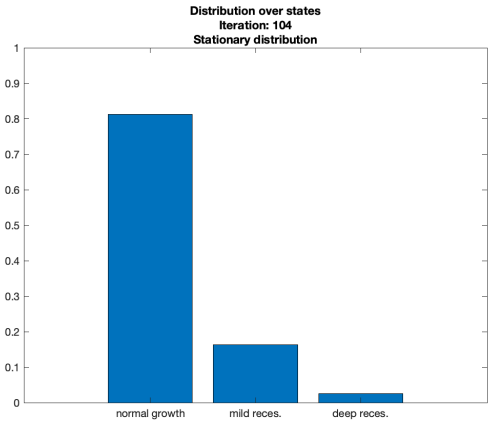

Figure 18: **Stationary distribution**

Starting from an initial distribution $\pi_0 = (0, 1/2, 1/2)$, we converge fast (100 iterations) to the limit.

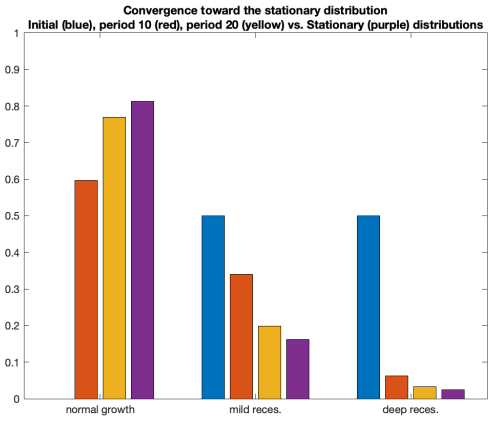

Figure 19: **Convergence toward the stationary distribution**

Theorem 5.2 (Existence of stationary distribution)**.**
Every stochastic matrix P has at least one stationary distribution. $\forall P, \exists \pi^\star$, s.t. $\pi^\star = \pi^\star P$
Note:

- Here, the assumption that S is a finite set is a key one.
- The proof of this theorem lie in the *Brouwer fixed point theorem*

- If P is the identity matrix, then all distributions are stationary

Questions:
(i) Is this stationary distribution unique? and
(ii) How fast does the stochastic process converges to its stationary distribution?
The answers to these natural questions are in the next two theorem:

Theorem 5.3.
If the stochastic matrix P is irreducible *and* aperiodic*, then :*

1. *P has exactly one stationary distribution $\pi^\star$*
2. *For any initial distribution π_0, we have $\|\pi_0 P^t - \pi^\star\| \to 0$ when $t \to \infty$*

Note:

- A stochastic matrix satisfying the conditions of the theorem is sometimes called *uniformly ergodic*
- Note that part 1 of the theorem requires only irreducibility, whereas part 2 requires both irreducibility and aperiodicity
- One easy sufficient condition for aperiodicity and irreducibility is that every element of P is strictly positive (Exercise?)

Ergodicity

Proposition 5.4 (Ergodicity for Markov chains)**.**
The definition above is quite abstract but here is a sufficient condition for Markov chains.
A Markov chain is said to be ergodic *if for $\pi^\star$ stationary distribution, if $f(\cdot)$ is solution of this "eigenvalue equation"*

$$\mathbb{E}[f(X_{t+1})|X_t] = f(X_t)$$

then it is simply the constant function $\pi^\star$-almost surely (i.e. constant wherever the stationary distribution is strictly positive). To visualize why we are talking about an eigenvalue problem, rewrite in matrix form:

$$Pf = f$$

Ergodicity implies that this f is constant $f(x_1) = \cdots = f(x_n) = \alpha$, $\pi^\star$ almost-surely.

Theorem 5.5 (Ergodic theorem for Markov Chains)**.**
Under irreducibility, an important result:

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\{X_t = x\} \xrightarrow[n\to\infty]{a.s.} \pi^\star(x)$$

Note that the convergence is almost sure *and it does not depend on the initial distribution π_0 of X_0.*

Note:

- The result tells us that the fraction of time the chain spends at state x converges to $\pi^\star(x)$ as time goes to infinity
- This convergence theorem is a special case of a **Law of large numbers** result for Markov chains

5.2 Martingales

Definition 5.17 (Martingale)**.**
In discrete-time, we define M_t as a ***martingale*** *(resp. super-martingale, sub-martingale), w.r.t. a filtration $(\mathscr{F}_t)_{t\geq 0}$, a stochastic process verifying:*

1. *$(M_t)_t$ is adapted*
2. *$\forall t, \mathbb{E}\big(|M_t|\big) < \infty$*
3. *$\forall t, \mathbb{E}(M_{t+1}|\mathscr{F}_t) = M_t$*
   (resp. $\mathbb{E}(M_{t+1}|\mathscr{F}_t) \leq M_t$ and $\mathbb{E}(M_{t+1}|\mathscr{F}_t) \geq M_t$)

Note:

- Intuitively, the mean of a martingale M_t is constant over time, while decreasing for a super-martingale and increasing for a submartingale.
- We see this by applying the law of iterated expectations to $\mathbb{E}[M_{t+1}] = \mathbb{E}[\mathbb{E}(M_{t+1}|\mathscr{F}_t)] \lesseqgtr \mathbb{E}[M_t]$

Proposition 5.6 (Doob-Meyer decomposition)**.**
Let $(X_t)_{t\geq 0}$ a sequence of real random variables $(\mathscr{F}_t)_{t\geq 0}$-adapted and integrable. There exists a unique pair of stochastic processes $(M_t)_{t\geq 0}, (V_t)_{t\geq 0}$ such that:

(i) $X_t = X_0 + M_t + V_t, n \geq 0$
(ii) $(M_t)_{t\geq 0}$ *is a $(\mathscr{F}_t)_{t\geq 0}$-martingale.*
(iii) $(V_t)_{t\geq 0}$ *is $(\mathscr{F}_t)_{t\geq 0}$-predictable process with $V_0 = 0$*

Proof: Uniqueness. Let $(M_t)_{t\geq 0}$ and $(V_t)_{t\geq 0}$, satisfying the three points. By (i) we have $X_{t+1} - X_t = M_{t+1} - M_t + V_{t+1} - V_t$. By taking the conditional expectation knowing $\mathscr{F}_t$, we find: $\mathbb{E}[X_{t+1} - X_t \mid \mathscr{F}_t] = V_{t+1} - V_t$. B $V_0 = 0$, and we conclude that $V_t = \sum_{k=0}^{t-1} \mathbb{E}[X_{k+1} - X_k \mid \mathscr{F}_k]$, and then necessarily, $M_t = X_t - X_0 - V_t$.

Existence. We pose $V_0 = 0$ and $V_t = \sum_{k=0}^{t-1} \mathbb{E}[X_{k+1} - X_k \mid \mathscr{F}_k]$ and then $M_t = X_t - X_0 - V_t$, both for $t \geq 1$. The points (i) and (iii) are obviously verified. For (ii), we remarque that $M_{t+1} - M_t = (X_{t+1} - X_t) - (V_{t+1} - V_t) = (X_{t+1} - X_t) - \mathbb{E}[X_{t+1} - X_t \mid \mathscr{F}_t]$, and hence $\mathbb{E}[M_{t+1} - M_t \mid \mathscr{F}_t] = 0$.

Stopping times

Stopping times are really useful for "stopping problem", for example American options, Entry-Exit problems of firms, Time of adjustment (change in prices or inventory, etc.), run to the bank, etc.

Definition 5.18 (Stopping time)**.**
Let $\{\mathscr{F}_t\}_{t\geq 0}$ a filtration. A random variable $T : \mathbb{N} \cup \{\infty\} \to \infty$ is a $\{\mathscr{F}_t\}_t$ stopping time if $\forall\, s > 0$, we have that $\{T \leq s\} \in \mathscr{F}$.
We can therefore pose $\mathscr{F}_T = \{A \in \mathscr{F}_\infty,\ s.t.\ A \cap \{T < s\} \in \mathscr{F}_s, \forall\, s \in \mathbb{R}\}$.

Example 5.7.

- *Let $T = t_n \in \mathbb{N}$ (or$\mathbb{R}$) a deterministic time. It is a stopping time, and $\mathscr{F}_T = \mathscr{F}_{t_n}$*
- *If $\{X_t\}_t$ is another stochastic process adapted. Then for all $B \in \mathscr{B}_\mathbb{R}$, we define $T_B = \inf\{n \geq 0,\ s.t.\ X_t \in B\}$. It is a $\mathscr{F}-$stopping time,*

Proposition 5.7.
Let $\{X_t\}_t$ be a martingale (or resp. supermartingale, or submartingale) on $\{\mathscr{F}_t\}_t$, and let T be a stopping time on $\{\mathscr{F}_t\}_t$. Then, the "stopped process" $(X_{t\wedge T})_{t\geq 0}$ is also a martingale (or resp. super-martingale, or submartingale) on $\{\mathscr{F}_t\}_t$ *Note:* The second example 5.3. in figure 13 also works with T a stopping time and indeed it is a martingale.

Proposition 5.8.
Let $\{X_t\}_t$ be a martingale on $\{\mathscr{F}_t\}_t$ and let T be a stopping time on $\{\mathscr{F}_t\}_t$. We suppose that $T < \infty$ almost-surely and there exists a real random variable Z such that $\forall t \in \mathbb{N}$ (or $\mathbb{R}$) we have $|X_{t\wedge T}| \leq Z$. Then X_T is integrable and $\mathbb{E}[X_0] = \mathbb{E}[X_T]$.

We arrive at the main point of this section:

Theorem 5.9 (Doob Stopping time theorem)**.**
Let $\{X_t\}_t$ be a martingale (or resp. supermartingale, or submartingale) on $\{\mathscr{F}_t\}_t$, X_∞ it limit almost-sure, and let $S \leq T$ be two stopping times on $\{\mathscr{F}_t\}_t$. Suppose either that X_t be uniformly integrable that or S and T are bounded*, we have*

- *X_T is $\mathscr{F}_T$ measurable and integrable and $X_T = \mathbb{E}[X_\infty|\mathscr{F}_T]$ and consequently $\mathbb{E}[X_T] = \mathbb{E}[X]$*
- *$\{X_{t\wedge T}\}_t$ is a martingale uniform integrable, which converges almost-surely and in L^1 toward X_T*
- *$\mathbb{E}[X_T|\mathscr{F}_S] = X_S$*

5.3 Continuous time stochastic processes

In this section, we will cover two main classes of stochastic process in continuous time: first diffusion processes, based on Brownian motion – known to be smooth and regular, and second Markovian Jump processes, that are simply a generalization of discrete-time discrete-state-space Markov chains.

Brownian motion

Let us start with the Brownian Motion, the *"continuous-time"* stochastic process which is the closest to a random walk.

Definition 5.19.
We define as a ***Brownian motion*** *the continuous process B_t valued in $\mathbb{R}$ such that:*

1. *The function $t \mapsto B_t(\omega)$ is continuous on $\mathbb{R}_+$*
2. *For all $0 \leq s < t$, the increment $B_t - B_s$ is independent of $\sigma(B_u, u \leq s)$*
3. *For all $t \geq s \geq 0$, $B_t - B_s$ follows the normal distribution $\mathcal{N}(0, (t-s)\sigma^2)$*

Note:

- The Brownian motion is "standard" if $B_0 = 0$ and $\sigma = 1$.
- Here, the Brownian motion is a martingale
- It is used to model any "small" shock in a continuous-time finance/macro models.
- *By a theorem (Donsker theorem), it is possible to show that a "normal-shock"-random-walk converges in law toward a Brownian motion, when time increment tends to zero.*
- *Notation:* we call $dB_t = \lim_{dt \to 0} B_{t+dt} - B_t$ the increment of the Brownian motion when time goes to zero.

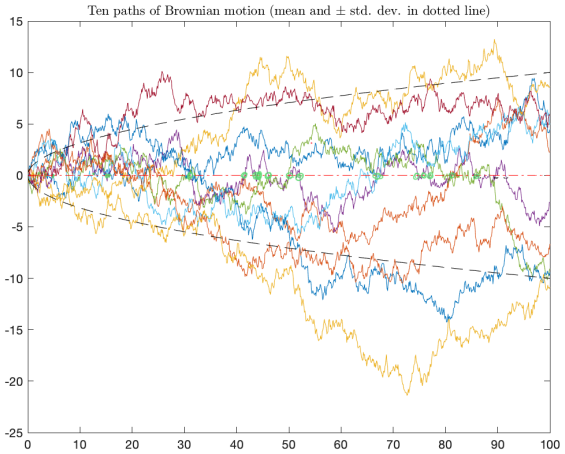

Figure 20: **Sample paths of the Brownian Motion**

Proposition 5.10 (Combo!)**.**
The Brownian motion $\{B_t\}_t$ is the continuous-time stochastic process that is all :

1. *A continuous process*
2. *A Markov process*
3. *A martingale*

Note: The other processes that are together (2) and (3) are Levy processes, which are a mixture of diffusion processes and Jump processes.

Let us introduce some properties for the Brownian motion:

Proposition 5.11 (Properties)**.**
The Brownian motion has the following features:

- *A continuous process $\{B_t\}_t$ is a Brownian motion (BM) if and only if* (i) *it is a Gaussian process (c.f. section above) and* (ii) *it verifies:*

$$\mathbb{E}[B_t] = 0 \qquad \mathbb{E}[B_t B_s] = \min\{s,t\}, \qquad \forall\ t,s \geq 0$$

This is due to the fact that Gaussian processes are entirely characterized by their mean and variance. Some consequences are implied:

- *Translation : $\forall\, \tau > 0\ \ \{B_{t+\tau} - B_\tau\}_{t\geq 0}$ is a BM*
- *Scaling : $\forall\, \alpha > 0\ \ \{\frac{1}{\sqrt{\alpha}} B_{\alpha t}\}_{t\geq 0}$ is a BM*
- *Time inversion $\{tB_{1/t}\}_{t\geq 0}$ is a BM*
- *Time reversal $\forall\, \tau > 0,\ \ \{B_r - B_{r-t}\}_{t\geq 0}$ is a BM*

Slightly different (and unrelated) properties:

- *Non-Hölder paths : $\limsup_{t\to 0} \frac{B_t}{\sqrt{t}} = +\infty$*
  Said differently, for small t, B_t goes slower to zero than $\sqrt{t}$.
- *A Brownian path $t \mapsto B_t(\omega)$ pass by zero infinitely many times (we say 0 is recurrent)*
- *A Brownian path is nowhere differentiable*

Proposition 5.12 (Brownian-based Martingales)**.**
If $\{B_t\}_t$ is a Brownian motion (BM), the following 3 processes are martingales:

$$(i) \qquad B_t, \qquad (ii) \qquad B_t^2 - t \qquad (iii) \qquad e^{\lambda B_t - \frac{\lambda^2}{2} t}$$

Stochastic integral and diffusion processes

Let us first define the stochastic integral very briefly. A two-line summary of the Ito's integral is the following. For a $\mathscr{F}_t$-predictable process (right-continuous and adapted) σ_t such that $\int_0^\infty \sigma_s^2 ds < \infty$, an "Ito integral process" is a stochastic process M_t defined as the limit – with a partition Π_n with mesh size going to zero:

$$M_t = \int_0^t \sigma_s dB_s = \int_0^t = \lim_{n\to\infty} \sum_{(t_i,t_{i+1})\in\Pi_n} \sigma_s \cdot (B_{t_{i+1}} - B_{t_i})$$

This integration is a stochastic counterpart of Riemann-Stieltjes integral.

A longer summary follows. As the Lebesgue integral, it takes a couple of steps to construct Ito's stochastic integral:

- We start with step processes $\{X_t\}_t$, with a collection $\{X_{t_i}\}_{t_i}$, $\mathscr{F}_{t_i}$-measurable, of the form:

$$X_t = \sum_{i=0}^{k} X_{t_i} \mathbb{1}_{[t_i,t_{i+1}]}(t)$$

This allow us to define a first "step" of the integral:

$$M_t = \int_0^t X_s dB_s := \sum_{i=0}^{k} X_{t_i}\big(B_{t_{i+1}\wedge t} - B_{t_i \wedge t}\big)$$

- We then do an extension (i.e. the limit, a bit like in the 2nd step of the construction of Lebesgue integral), for all process that are adapted and continuous and bounded $\{X_t\}_t$

$$M_t = \int_0^t X_s dB_s$$

- We can then define the integral with respect to any semi-martingale that are bounded in L^2.

$$\widetilde{M_t} = \int_0^t Y_s dX_s$$

Definition 5.20 (Ito process)**.**
We define an Ito process as a stochastic process which the sum, for two processes $\{\mu_t\}_t$ and $\{\sigma_t\}_t$ adapted and continuous of integrals

$$X_t = X_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dB_s$$

Or written as a differential form :

$$dX_t = \mu_t dt + \sigma_t dB_t$$

Note: (Important) This definition takes the processes $\{\mu_t\}_t$ and $\{\sigma_t\}_t$ as given. In the following we will see processes where $\{\mu_t\}_t$ and $\{\sigma_t\}_t$ are *functions* of X_t (and potentially other variables: controls, General equilibrium feedback $\bar{X}_t = \mathbb{E}[X_t]$ or other moments – which is called the "mean field" impact). These processes – called diffusions – are actually the *solutions* of stochastic differential equations (SDE), provided these solutions exists ! (which is not obvious at all).

Definition 5.21 (Quadratic variation)**.**
Let M_t a continuous martingale.

- *We define the* quadratic variation *of $\langle M\rangle_t$ the unique adapted and increasing process such that $M_t^2 - \langle M\rangle_t$ is a continuous martingale.*
  In differential form, if $dX_t = \mu_t^X dt + \sigma_t^X dB_t$, then

$$d\langle M\rangle_t = (\sigma_t^X)^2 dt$$

- *Let us generalize to the multidimensional case: Let X_t, Y_t be Ito processes*

$$dX_t = \mu_t^X dt + \sigma_t^X dB_t \qquad\qquad dY_t = \mu_t^Y dt + \sigma_t^Y dB_t$$

  Their quadratic covariation process $\langle X, Y\rangle_t$ is given by:

$$d\langle X, Y\rangle_t = \sigma_t^X \sigma_t^Y dt$$

Note: Hence, intuitively, the quadratic covariation process $\langle X\rangle_t$ capture the "variance" and $\langle X, Y\rangle_t$ the

"covariance" of the Brownian part of two Ito processes at time t. Note that in general, $\langle X, Y\rangle_t$ will be a random variable instead of a constant as the simple case above.

Moreover, after this construction, we obtain that :

Proposition 5.13 (Integral and martingales)**.**
The following processes are martingales:

$$M_t = \int_0^t X_s dB_s \qquad \Rightarrow \qquad \mathbb{E}[M_t] = \mathbb{E}\Big(\int_0^t X_s dB_s\Big) = 0$$

$$Q_t = M_t^2 - \int_0^t X_s^2 ds = M_t^2 - \langle M\rangle_t \qquad \Rightarrow \quad \mathbb{E}[Q_t] = 0 \qquad \Rightarrow \mathbb{E}\Big[\big(\int_0^t X_s dB_s\big)^2\Big] = \mathbb{E}[\int_0^t X_s^2 ds]$$

However, before we state the main result of this section, the Ito's lemma, which is very important to compute the function of diffusion process and is used a lot in Econ-Finance.

Theorem 5.14 (Ito's lemma)**.**
For any X_t Itô process:

$$dX_t = b_t\, dt + \sigma_t\, dB_t$$

with b_t and σ_t continuous and adapted processes and any $\mathcal{C}^{1,2}([0,t]\times\mathbb{R})$ scalar function $f(t,x)$ of two real variables t and x, one has:

$$df(t,X_t) = \frac{\partial f_{(t,X_t)}}{\partial t}dt + \frac{\partial f_{(t,X_t)}}{\partial x}dX_t + \frac{1}{2}\frac{\partial^2 f_{(t,X_t)}}{\partial x^2}d\langle X\rangle_t$$

or

$$df(t,X_t) = \left(\frac{\partial f}{\partial t} + b_t\frac{\partial f}{\partial x} + \frac{\sigma_t^2}{2}\frac{\partial^2 f}{\partial x^2}\right)dt + \sigma_t\frac{\partial f}{\partial x}\,dB_t$$

Theorem 5.15 (Ito's lemma, multidimensional version)**.**
For k-dimensions vector-valued processes $\mathbf{X}_t = (X_t^1, X_t^2, \ldots, X_t^k)$ and given d-dimensions source of Brownian noise $\mathbf{B}_t = (B_t^1, \ldots, B_t^d)$, the process :

$$d\mathbf{X}_t = \boldsymbol{b}_t\, dt + \sigma_t\, d\mathbf{B}_t$$

where $\mathbf{b}_t$ and σ_t is are respectively $k\times 1$ and $k\times d$ vector/matrices of continuous and adapted processes, the Itô formula rewrites:

$$\begin{aligned} df(t,\mathbf{X}_t) &= \frac{\partial f}{\partial t}(t,X_t)\,dt + \sum_{i=1}^k \frac{\partial f}{\partial x_i}(t,X_t)dX_t^i + \frac{1}{2}\sum_{i,j=1}^k \frac{\partial^2 f}{\partial x_i \partial x_j}(t,X_t)d\langle X^i, X^j\rangle_t \\ &= \partial_t f\, dt + \nabla_x f\cdot d\mathbf{X}_t + \frac{1}{2}Tr\left(\sigma_t\sigma_t^T D_{xx}^2 f\right)\, dt, \\ &= \left\{\partial_t f + \nabla_x f\cdot \boldsymbol{b}_t + \frac{1}{2}Tr\left(\sigma_t\sigma_t^T D_{xx}^2 f\right)\right\}dt + \nabla_x f\cdot\sigma_t\, d\mathbf{B}_t \end{aligned}$$

Note: The proof of the Ito's lemma relies on the discretization procedure for the construction of the stochastic integral, explained (but not detailed...) at the beginning of this section

Proposition 5.16 (Laplace transform and exponential of martingales)**.**
The Ito processes $\mathbf{X} = \{X_t\}_t$ *is a Gaussian process. In particular, its law is entirely determined by its characteristic functions (Fourier transform) and/or its Laplace transform :*

$$\forall \lambda \in \mathbb{C}^K \qquad \mathbb{E}\left[e^{\lambda \cdot \mathbf{X}}\right] = e^{\lambda \cdot \mathbb{E}(\mathbf{X}_t) + \lambda^T \mathbb{V}ar(\mathbf{X}_t)\lambda}$$

As a result, let $\{X_t\}_t$ *a continuous sub/super- martingale (or simply an Ito process), then there exists a continuous martingale* $\{\mathscr{E}(X_t)\}_t$

$$\mathscr{E}(X)_t := \exp\left(X_t - \frac{1}{2}\langle X\rangle_t\right)$$

This is called the "exponential" of the martingale X_t *(or sometimes the "Doleans-Dade" exponential).*

The following the main artillery to prove all the asset pricing formulas initiated by the Black-Scholes model. However, it is a very deep result on martingales.

Theorem 5.17 (Girsanov theorem)**.**
Let us consider $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_t, \mathbb{P})$ *an Ito processes* $\{X_t\}_t$ *and consider the exponential* $\mathscr{E}(X)_t$ *given in the result proposition. If* $\mathscr{E}(X)_t$ *is a martingale (i.e. the drift of the Ito process is null), then we can define a new measure* $\mathbb{Q}$ *such that the Radon-Nikodym derivative can be expressed as a exponential of* X_t*:*

$$\frac{d\mathbb{Q}}{d\mathbb{P}}\Big|_{\mathscr{F}_t} = \mathscr{E}(X)_t$$

Moreover, if Y_t *is a continuous martingale under* $\mathbb{P}$*, then the process* $\widetilde{Y}$

$$\widetilde{Y}_t = Y_t - \langle X\rangle_t$$

is also a martingale under $\mathbb{Q}$*.*

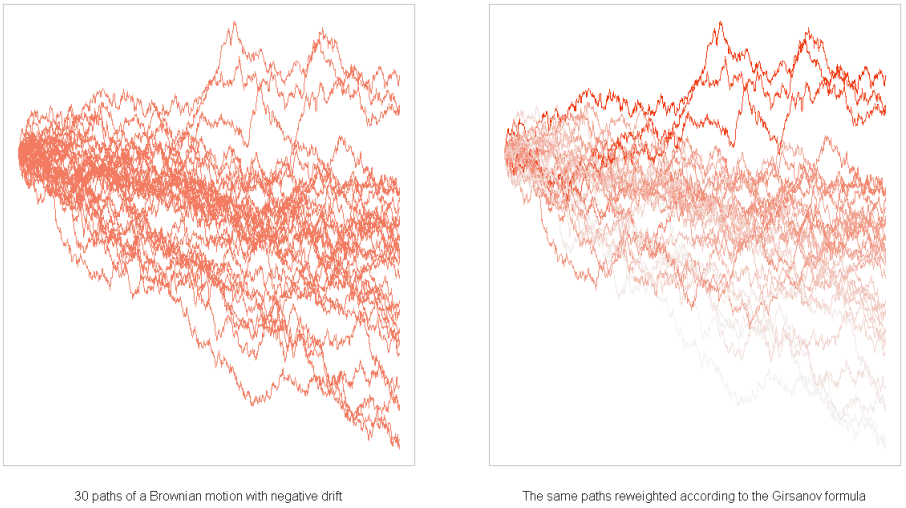

Figure 21: **Girsanov theorem with** $Y_t = B_t$

Definition 5.22 (Stochastic differential equation)**.**
Let us consider $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_t, \mathbb{P})$ *and a Brownian noise* $\mathbf{B}_t = (B_t^1, \dots, B_t^d)$*. We call the following equation a stochastic differential equation.*

$$dX_t = b(t, X_t)dt + \sigma_t(t, X_t)dB_t$$

Note:

- This is simply the generalization of $\dot{x}_t = \frac{dx_t}{dt} = b(t, x_t)$ for x_t
- There are different notions of existence and uniqueness of solution (weak or strong existence for example) and this is the object of an entire field in mathematics called "stochastic analysis"

Example 5.8 (Resolution using Girsanov)**.**
Take the SDE on $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_t, \mathbb{P})$*, with* $b(\cdot)$ *a bounded function:*

$$dX_t = b(t, X_t)dt + \sigma dB_t$$

First, X_t *is a Brownian motion on the space* $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_t, \mathbb{Q})$*, where we define the measure* $\mathbb{Q}$

$$\mathbb{Q} = \dots$$

The rest of this example is forthcoming

Kolmogorov Forward equation and Feynman Kac theorem

Suppose we consider N particles X_t^i, $i = 1, \dots, N$ subject to shocks given by a SDE:

$$\begin{cases} dX_t^i = b(t, X_t^i)dt + \sigma dB_t^i \\ X_{t_0}^i = Y^i \end{cases}$$

with Y^i and B_t^i i.i.d. (independence is key!).

From the evolution of these particles when $N \to \infty$, we look for their measure/law of the process: $m(t, x) = P_{X_t}(x)$, and one can obtain the Kolmogorov Forward or Fokker-Planck equation:

$$\begin{cases} \partial_t m(t, x) - div\big(b(t, x)\, m(t, x)\big) + \frac{\sigma^2}{2} D_{xx}^2\big(m(t, x)\big) = 0 \\ m(0, x) = m_0(x) \end{cases}$$

To obtain this more formally, derive the Itô's formula for test function $\varphi \in \mathcal{C}_c^\infty$ on X_t, take the expectation and derive the 'adjoint' operators on m (which is a more elaborate way to think to integration by part)

Note: On adjoints, recall that : $(b\nabla\cdot)^\star \equiv -\operatorname{div}(b\cdot)$ and $(\sigma\sigma^T\Delta\cdot)^* \equiv D^2(\sigma\sigma^T\cdot)$

Link with Feynman-Kac

The Feynman-Kac theorem is giving us a conceptual link between expectation of a process and its local dynamics given by a Kolmogorov Backward equation (which is a PDE).

If $w(t, x)$ is a $\mathcal{C}^{1,2}$ function and has bounded derivative, $\nabla_x v \in L^\infty$, and is solution of :

$$\begin{cases} \partial_t w(t, x) + b \cdot\ \nabla_x w(t, x) + \frac{1}{2} Tr\big(\sigma\sigma^T D_{xx}^2 w(t, x)\big) = 0 \\ w(T, x) = g(x) \end{cases}$$

Then, the Feynman Kac formula gives us the form of the solution:

$$w(t,x) = \mathbb{E}_{t_0}\Big(g(X_T^{t_0,x})\Big)$$

where X_T is the solution of the SDE:

$$\begin{cases} dX_t^x = b(t,X_t^x)dt + \sigma(t,X_t^x)dB_t \\ X_{t_0}^x = x \qquad\qquad (t_0,x_0) \in [0,T]\times\mathbb{R}^d \end{cases}$$

The above PDE is called Feynman-Kac equation or "Kolmogorov Backward equation"

The Feynman-Kac/Kolmogorov Backward equation is

$$\begin{cases} \partial_t w(t,x) + b\cdot\nabla_x w(t,x) + \frac{1}{2}Tr\big(\sigma\sigma^T D_{xx}^2 w(t,x)\big) = 0 \\ v(T,x) = g(x) \end{cases}$$

When one "return the time" (and call $w \equiv p$), one finds the following "Kolmogorov Forward equation"

$$\begin{cases} -\partial_t p(t,x) - div\big(b\,p(t,x)\big) + \frac{1}{2}D_{xx}^2\big(\sigma\sigma^T p(t,x)\big) = 0 \\ p(0,x) = p_0(x) \end{cases}$$

More formally, this equation is the "adjoint" equation of the KBE.

Let us give a general formula for the Feynman Kac.

Theorem 5.18 (Feynman-Kac)**.**
Consider the function

$$v(t_0,x_0) = \mathbb{E}_{t_0}\Big[\int_{t_0}^T e^{-\int_{t_0}^s r(u,X_u)du} f(s,X_s)ds \;+\; e^{-\int_{t_0}^T r(u,X_u)du} g(X_T)\Big] \qquad \forall (t,x)\in[0,T]\times\mathbb{R}^d$$

Supposing that X follows the SDE:

$$\begin{cases} dX_t = b(t,X_t)dt + \sigma(t,X_t)dB_t \\ X_{t_0} = x_0 \qquad\qquad (t_0,x_0)\in[0,T]\times\mathbb{R}^d \end{cases}$$

The Feynman-Kac formula tells us that v is solution to the PDE:

$$\begin{cases} r(t,x)\,v(t,x) - \partial_t v(t,x) - \nabla_x v(t,x)\cdot b - \frac{1}{2}Tr\big(\sigma\sigma^T D_{xx}^2 v(t,x)\big) = f(t,x) \\ v(T,.) = g \end{cases}$$

Moreover, if $w(t,x)$ is $\mathcal{C}^{1,2}$ and has bounded derivative, then $w(t,x) = v(t,x)$, i.e. admits the representation above.

Intuitions: a function v of X subject to a diffusion can be represented by the expected future value g, adding running gain f and discounting r. It is used a lot in finance to compute option prices (Black-Scholes). Moreover, one can compute w using Monte-Carlo methods for instance.

5.4 Continuous-time Markov processes

We consider continuous time processes that are "right continuous with left limits", or *càdlàg*[2] processes (or functions, think to c.d.f.!). This is defined as $\forall\ \omega \in \Omega$ and $\forall\ t \geq 0$ there exists ε such that $X_s(\omega) = X_t(\omega)$ for all $s \in [t, t+\varepsilon]$. This allows to have consistent definition of jump processes while keeping fundamental properties like adaptability. Recall that a (time-homogeneous) continuous time Markov chain X_t is can be described by a transition function P defined in definition 5.9.

We will start with jump process on discrete state-space, where the main example are Poisson jump processes, because turning to more general Markov process.

Jump process in discrete state-spaces

In the following, consider $\{X_t\}_t$ a cadlag process.

Definition 5.23 (Jump times)**.**
We introduce the times of jump $J_0, J_1, J_2, \dots$ *:*

$$J_0 = 0, \quad J_{n+1}(\omega) = \inf\left\{t \geq J_n(\omega) : X_t(\omega) \neq X_{J_n}(\omega)\right\}, \quad n = 1, 2, \dots, \qquad \textit{inf}\,\emptyset = \infty$$

and the time spent between the jumps:

$$S_n(\omega) = \begin{cases} J_n(\omega) - J_{n-1}(\omega) & \textit{if } J_{n-1}(\omega) < \infty \\ \infty & \textit{is } J_{n-1}(\omega) = \infty \end{cases}$$

The discrete time process $Y_n = X_{J_n}$ *is said to be the jump chain of* $\{X_t\}_{t\geq 0}$ Again assume that the state space $S = \{x_1, \dots, x_i, \dots\}$ is discrete (i.e. countable or finite). Again the Markov process is described by a special matrix:

Definition 5.24.
Intensity matrix An intensity matrix (also called the transition rate matrix) called Q *that has the following properties:*

- $0 \leq -q_{i,i} < \infty$*, for all* $x_i \in S$
- $q_{i,j} \geq 0$ *pour tout* $i \neq j, x_i, x_j \in S$
- $\sum_{x_j \in S} q_{i,j} = 0$ *for all* $x_i \in S$

Note: For Q given, we introduce a matrix $P = p_{k,j}, \forall\, x_i, x_j \in S$ which is the Markov transition matrix associated with the intensity matrix above. Intuitively it describes the transition (hence in discrete time) of the continuous time Markov-process, *conditional on jumping*, i.e. the transition matrix of $Y_n = X_{J_n}$ the jump chain is :

$$p_{i,j} = \begin{cases} q_{i,j}/\left(-q_{i,t}\right) & \text{if } j \neq i,\ q_{i,t} \neq 0 \\ 0 & \text{if } \ j \neq i,\ q_{i,i} = 0 \end{cases}$$

$$\pi_{i,i} = \begin{cases} 0 & \text{if } q_{i,i} \neq 0 \\ 1 & \text{if } q_{i,i} = 0 \end{cases}$$

In the following we denote $-q_{i,i} = q_i$.

[2] From French *continue à droite et limite à gauche*, no joke it's called like that in English!

Definition 5.25 (Markov process)**.**
A process $\{X_t\}_{t\geq 0}$ in continuous time is a Markov process if it is described by an intensity matrix Q and if its jump chain is a Markov chain on the same state space S, and the times spent between jumps $S_1, \ldots, S_n$ has the conditional law of independent random variable of exponential law of parameters $q(Y_0), \ldots, q(Y_{n-1})$ respectively conditional on fixed values $Y_0, \ldots, Y_{n-1}$.

Let us do a brief recap detour:

Definition 5.26 (Exponential law)**.**
A random variable T follows an exponential law of parameter λ (or $\mathscr{E}(\lambda)$ if for all $t > 0$

$$\mathbb{P}(T > t) = e^{-\lambda t}$$

The p.d.f of the exponential distribution is $f_T(t) = \lambda e^{-\lambda t} \mathbb{1}_{t>0}$ and $\mathbb{E}T = \frac{1}{\lambda}$

Example 5.9.
The matrix:

$$Q = \begin{pmatrix} -2 & 1 & 1 \\ 3 & -7 & 4 \\ 0 & 0 & 0 \end{pmatrix}$$

for the jump process generates the jump chain that follows a Markov chain with intensity:

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 3/7 & 0 & 4/7 \\ 0 & 0 & 1 \end{pmatrix}$$

A jump process will jump faster if its intensity q_i is more negative. The Markov jump process with intensity Q is associated with the jump chain $\{Y_n\}_{n\geq 0}$ which is Markov chain with transition matrix P. When a trajectory is fixed e.g. $Y_0 = 1, Y_1 = 2, Y_3 = 1, Y_4 = 3$, then S_1 the time lapsed between 1 and 2 is the exponential law of intensity 2. S_2 the time lapsed between 2 and 3 is the exponential law of intensity 7 and similarly $S_3 \sim \mathscr{E}(2)$. Moreover, S_1, S_2, S_3 are independent. Moreover, the last /third state is fully absorbing (when you reach this state, the intensity of jumping is null, so you never jump again).

Proposition 5.19 (Constructions)**.**
There are different way to construct a Markov process in practice (and numerically).

(i) Construct a Markov chain $\{Y_n\}_n$ of transition matrix P and then to use $T_1, T_2, \ldots$ independent r.v. of parameters 1. Let $S_n = T_n / q(Y_{n-1})$ – using rescaling properties of exponentials and

$$X_t = \begin{cases} Y_n & \text{if } S_1 + \cdots + S_n \leq t < S_1 + \cdots + S_n + S_{n+1} \\ \infty & \text{if not} \end{cases}$$

(ii) This second equivalent construction is recursive. Start from $X_0 = Y_0$ use $T_1, T_2, \ldots$ independent r.v.

of parameters 1. *By recursion, we have, if* $Y_n = i$ *we pose :*

$$\begin{aligned} S_{n+1}^j &= T_{n+1}^j / q_{i,j} \text{ for } j \neq i \\ S_{n+1} &= \inf_{j \neq i} S_{n+1}^j \\ Y_{n+1} &= \begin{cases} j & \text{if } S_{n+1}^j = S_{n+1} < \infty \\ i & \text{if } S_{n+1} = \infty \end{cases} \end{aligned}$$

Then, under conditions $Y_n = i$, S_{n+1}^j *follow independent exponential law of parameters* $q_{i,j}$.

Example 5.10 (Poisson process)**.**
We consider a Poisson process: it is a jump process on $\mathbb{N}$ *where all the jump are upward of one unit. Here we consider the Poisson process of parameter* $\lambda > 0$, *i.e. it is the counting process associated to the jump time process* $\{T_n\}_{n\geq 1}$ *where the random variables* T_n *are called time of jumps, and are defined as:*

$$\forall n \geq 1, \quad T_n - T_{n-1} = S_n, \qquad \text{with } T_0 = 0$$

with $\{S_n\}_{n\geq 1}$ *a sequence of i.i.d. of exponential law* $\lambda > 0$.

Note that the intensity matrix has the i*-th row* $(\dots, 0, -\lambda, \lambda, 0, \dots)$ *where the negative term is on the* i*-th column.*

For all $t \geq 0$, *we define:*

$$N_t = \sum_{n\geq 0} \mathbb{1}_{T_n \leq t},$$

and in the next graph we simulate the trajectories $\{N_t\}_{t\in[0,T]}$ *for a finite horizon* $T > 0$.
Note that the parameter is quite high: $\lambda = 20$, *so a trajectory jumps on average every* $\mathbb{E}[S_t] = \frac{1}{\lambda} = 0.05$ *units of time.*

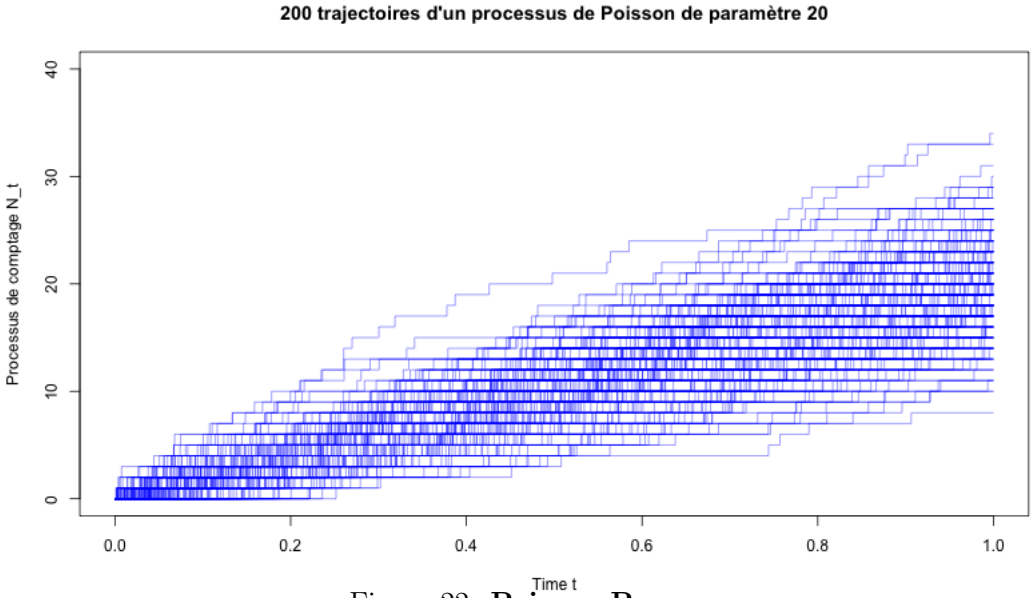

Figure 22: **Poisson Process**

Theorem 5.20 (Markov Property)**.**
Let $\{X_t\}_{t\geq 0}$ *a Markov process of intensity matrix* Q *and let* $s > 0$*. Under condition* $\{X_s = i\}$*,* $(X_{t+s})_{t\geq 0}$ *is a Markov process with intensity matrix* Q *independent of* $\sigma\left(X_r, r \leq s\right)$ The proof of this property relies on the important of property of exponential laws :

Theorem 5.21 (Memoryless properties)**.**
Let $T : \Omega \in (0, \infty]$ *follows an exponential law if and only if*

$$\mathbb{P}(T > s + t \mid T > s) = \mathbb{P}(T > t) \quad \forall s, t > 0$$

We now find again the dynamics of the distribution using the properties of the exponential (in particular the first order approximation).

Theorem 5.22.
The three following properties are equivalent:

- *The process* $\{X_t\}_t$ *is a Markov process in continuous time with intensity* Q
- *For all* t *and* $h > 0$*, conditional on* $X_t = x_i$*,* X_{t+h} *is independent of* $\sigma(X_s, s \leq t)$ *and for all* x_j*, when* $h \downarrow 0$ *uniformly for all* $t \geq 0$*:*

$$\mathbb{P}(X_{t+h} = x_j | X_t = x_i) = \delta_{i,j} + q_{i,j}h + o(h)$$

- *For all* $n = 0, 1, 2, \ldots,$ *and* $t_0 \leq t_1 \leq \cdots \leq t_{n+1}$ *times and* $x_{i,0}, x_{i,1}, \ldots, x_{i,n+1}$ *the states, as well as a distribution over states* $\pi(t, x)$ *starting at* $\pi(0, x)$ *(intuitively a row vector for all* t*), we have :*

$$\mathbb{P}(X_{t_n+1} = x_{i,n+1} | X_{t_0} = x_{i,0}, \ldots, X_{t_n} = x_{i,n}) = \mathbb{P}(X_{t_n+1} = x_{i,n+1} | X_{t_n} = x_{i,n})$$

and we have the Kolmogorov forward equation:

$$\frac{d\pi(t, y)}{dt} = \int_S \pi(t, x)Q(x, y)dx$$

or in matrix form:

$$\frac{d\pi(t)}{dt} = \pi(t)Q$$

General Markov processes

A (time-homogeneous) continuous time Markov chain X_t is can be described by a transition function P defined in definition 5.9. Recall, given the time homogeneous Transition function $p : \mathcal{T}\times S\times\mathcal{G}$, we denote the transition function over the interval of time s:

$$P_s = [p(s, x, dy)]_{x,dy}$$

where $p_{s,i,j} = \mathbb{P}(X_{t+s} \in dy|X_t = i)$. Note that we assume that the process is time-homogenous in the sense that $p_{s,i,j}$ doesn't on time t but only on the interval of time. By the Chapman Kolmogorov property, it is immediate that $P_{t+s} = P_tP_s$ and $P_0 = \mathbb{I}$. We will also assume that P_t is right-continuous so $\lim_{h\to 0^+} P_h = \mathbb{I}$.

Definition 5.27.
The infinitesimal generator of a continuous-time Markov process X_t is an operator $\mathcal{A}$ such that:

$$\mathcal{A}f(x) = \lim_{h\to 0} \frac{1}{h}\mathbb{E}[f(X_{t+h}) - f(X_t)|X_t = x]$$

Note:

- In the cases of pure jump process as in the previous section, the Infinitesimal generators are nothing else that the intensity matrix. However, in uncountable spaces the intensity matrix becomes an operator, i.e. the generalization of matrix in infinite dimension.
- For intuitions, assume $|S| < \infty$ for Markov chains, any function $f : S \to R$ is simply a column-vector $[f(x_j)]_j$ and:

  $$\mathbb{E}[f(X_t)|X_0 = x_i] = e_iP^t[f(x_j)]$$

  where $e_i = (0, .., 1, .., 0)$ is a row vector with 1 at the ith component and zero otherwise (c.f. discussion above in the discrete time case). Then, subtract $f(X_t) = f(x_i)$ on both side,

  $$\mathbb{E}[f(X_{t+h}) - f(X_t)|X_t = x_i] = e_iP_h[f(x_j)]_j - f(x_i)$$

  and, in matrix form:

  $$\Big[\mathbb{E}[f(X_{t+h}) - f(X_t)|X_t = x_i]\Big]_i = \big[P_h\ [f(x_j)]\big]_i - [f(x_j)]_j = (P_h - \mathbb{I})[f(x_j)]_j$$

  In this finite state case, define $\mathcal{A} : \mathbb{R}^N \to \mathbb{R}^N$ a matrix such that

  $$\mathcal{A} \coloneqq \lim_{h\to 0+} \frac{P_h - \mathbb{I}}{h},$$

  then $\mathcal{A}$ is the infinitesimal generator of the Markov process:

  $$\mathcal{A}\ f(x_i) = \lim_{h\to 0+} \frac{1}{h}(P_h - \mathbb{I})f(x_i) = \lim_{h\to 0} \frac{1}{h}\mathbb{E}[f(X_{t+h}) - f(X_t)|X_t = x_i]$$

- We ignore the problem of the existence of this limit and proceed as if this linear operator (simply a "matrix" with infinite entries) $\mathcal{A}$ exists.

Infinitesimal generators and Kolmogorov equations

With an infinitesimal generator, we can describe the transition function P_t in a concise way. Consider the derivative of P_t at time t:

$$\frac{dP_t}{dt} = \lim_{h \to 0+} \frac{P_{t+h} - P_t}{h}$$

We can either factor P_t out on the left or on the right:

$$\frac{dP_t}{dt} = \mathcal{A}P_t, \quad \frac{dP_t}{dt} = P_t\mathcal{A}$$

These differential equations correspond to Kolmogorov's backward and forward equations respectively.

The solution of the Kolmogorov equations are given by

$$P_t = Ce^{\mathcal{A}t}$$

where given matrix B, the matrix exponential is defined as $e^B \coloneqq \sum_{j=0}^{\infty} \frac{B^j}{j!}$. Moreover, $P_0 = I$ implies $C = I$.

Following the same logic as in discrete time and discrete space, we compute the dynamics of the distribution π and the conditional expectation of a function f

Let $X_t \sim \pi_t$, the distribution of X_{t+s} is given by:

$$\pi_t P^s = \pi_{t+s} \implies d\pi_t = \pi_t \mathcal{A} dt$$

Let $V(x_i, T) = f(x_i)$ and $V(x_i, t) = \mathbb{E}[f(X_T)|X_t = x_i]$, then we have the Kolmogorov backward equation:

$$V_t = P^{T-t}V_T \quad \implies \quad -dV_t = \mathcal{A}V_t dt$$

These system of differential equations, together with initial conditions π_0 and V_T, are the Kolmogorov forward and backward equations (or Fokker-Planck and Feynman-Kac equations using):

$$d\pi_t = \pi_t \ \mathcal{A} \ dt, \ \pi_0 \text{ given}$$
$$-dV_t = \mathcal{A} \ V_t \ dt, \ V_T \text{ given}$$

The stationary distribution π of the process $\{X_t\}$ is again the distribution that stabilize over time. Then $\pi_t = \pi P_t = \pi, \forall t$ and satisfies

$$0 = \frac{d\pi_t}{dt} = \pi\mathcal{A}$$