# Empirical Processes (MATH-522)

## Lecture 1: Introduction, Basic limit theorems and Important Examples

Myrto Limnios

MATH, Ecole Polytéchnique Fédérale de Lausanne

February 18, 2025

# General Information

- Lectures on Tuesdays 13:00-15:00, exercises with Nikitas Georgakis 15:00-17:00
- Course credits 5ECTS
- 11/03: Course d'Ecole $\Longrightarrow$ No class
- Mid-term exam 15/04 - 1h (before Easter break)
- Lecture and exercise room MA110
- Final written exam mid-June (unknown date)

Organization of the lectures

- Slides + Blackboard lectures
- All material available on Moodle - weekly updated
- Exercise sheets available on Tuesday, and corrections the week after

# Recommended literature

- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, Oxford Academic, 2013.
- Sara van de Geer, *Empirical Processes in M-Estimation.* Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2009.
- Aad W. Vaart , Jon A. Wellner, *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer Series in Statistics (SSS), Springer New York, 1996.

Additional important book on the topic

- Michel Talagrand, *Upper and Lower Bounds for Stochastic Processes. Decomposition Theorems* A Series of Modern Surveys in Mathematics, Springer Nature, 2021.

# What you can expect to learn by the end of the semester

- Formulate the fundamental framework related to empirical processes
- Manipulate probabilistic inequalities for empirical estimators
- Describe the concept of concentration-of-measure phenomenon for empirical processes
- Apply concentration inequalities to derive the performance of statistical learning procedures

# Outline of the semester - up to changes depending on you

1. Introduction: main concepts and important examples
2. Basic concentration inequalities
3. General theory for stochastic processes: measurability, continuity, weak convergence
4. Complexity of classes: measures and relations (bracketing numbers, covering numbers, combinatorial measures)
5. Maximal inequalities and chaining
6. Symmetrization and applications
7. Recent advances in concentration inequalities for processes (fast rates / model selection)

**Applications:**

8. $M$-estimators: rates of convergence and argmax theorem
9. Least-squares estimators
10. $U$-statistics: definitions and concentration results

# Today's outline

# Concentration

Suppose we observe a sequence of independent random variables (r.v.s)

$$X_1, \ldots, X_n$$

with $n \in \mathbb{N}^*$. Then, under some conditions, we have by the Law of Large Numbers

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \underset{n\to\infty}{\longrightarrow} 0$$

We can see this as a very simple function of the data. Define the function

$$h(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

then the LLN proves that the function $h$ valued on the data gets closer to its mean for large sample size $n$.

We will see that it is not restricted to linear functions of the data, and we will try to **quantify** the **fluctuations** of the random function $h$ based on the random sequence $X_1, \ldots, X_n$ around its mean, i.e., of

$$h(X_1, \ldots, X_n) - \mathbb{E}[h(X_1, \ldots, X_n)]$$

as soon as $n$ is *large enough*.

We will focus on how *small* the fluctuations can be depending on how *sensitive* the function $h$ is w.r.t. the coordinates.

The results can be in terms of the sample size, and inherent properties to the underlying distribution of the $X$'s.

## Remark

- Concentration results apply for general function $h$ and we will understand why it holds for those cases.
- The results will mainly be formulated in the form of tail inequalities: we want to state that for any probability level parametrized by $\delta > 0$, then we can explicitly derive the threshold $t_{\delta,n}$ such that

$$\mathbb{P}(|h(X_1, \ldots, X_n) - \mathbb{E}[h(X_1, \ldots, X_n)]| \geq t_{\delta,n}) \leq \delta \ .$$

  We want the fluctuations on the left-hand side (LHS) to be small with high probability (w.h.p.)

However, those results do not necessarily say anything about the expected value $\mathbb{E}[h(X_1, \ldots, X_n)]$...

# Suprema

- Suppose now that we index the r.v. $Z$ by a set $T$ - usually high or infinite dimensional. Then, we call the resulting collection a **stochastic process** $\{Z_t\}_{t \in T}$.
- Suppose that the r.v. $Z$ is centered for all $t \in T$: $\mathbb{E}Z_t = 0$.

A central quantity that is important in many statistical problems is that of

$$\sup_{t \in T} Z_t$$

where $Z_t$ is the stochastic process indexed by a class $T$.

For example, $T$ can be a class of real-valued measurable functions, or a class of sets.

Our main goals are:

1. To understand how to extend convergence properties of r.v.s to **uniform** convergence of r.v.s when indexed by sets $T$
2. To understand why and how we can measure the *size* of the stochastic process $Z_t$ by quantifying

$$\mathbb{E}\left[\sup_{t\in T}|Z_t|\right]$$

## Remark

Notice that the stochastic process $\{Z_t\}_{t\in T}$ can be defined as a function of the data $X_1, \ldots, X_n$.

# Today's outline

# Metric spaces

## Definition

Let $\mathbb{D}$ be a nonempty set. The map $d : \mathbb{D} \times \mathbb{D} \to \mathbb{R}_+$ is said to be a *metric* or *distance function* if it satisfies the following properties, for all $x, x', y \in \mathbb{D}$:

1. $d(x, y) = 0$ iff. $x = y$
2. symmetry: $d(x, y) = d(x, y)$
3. triangle inequality: $d(x, y) \leq d(y, x') + d(x', x)$

It is a *semimetric* does not necessarily satisfy 1. We say that the couple $(\mathbb{D}, d)$ is a *(semi)metric space*.

An open ball center on $x$ and of radius $r > 0$ is the set $\{y, d(x, y) < r\}$. An *open subset* of a metric space is generated by the union of open balls, and it is *closed* iff. it is the complement of an open set.

# Continuity

Let $(\mathbb{D}, d)$ and $(\mathbb{E}, d')$ two metric spaces. We recall the following definitions and properties.

- We say that a sequence $x_n$ *converges* to $x$ iff. $d(x_n, x) \to 0$.
- The *closure* $\bar{A}$ of an open subset $A \subset \mathbb{D}$ is composed of all the limit points of the elements in $A$. It is the smallest closed set containing the elements of $A$.
- The *interior* $\mathring{A}$ is the set of all points in $A$ such that they are contained in an open set $U \subset A$. It is the largest open subset of $A$.
- A function $f : \mathbb{D} \to \mathbb{E}$ is *continuous* at a point $x$ iff. for every sequence $x_n \to x$, $f(x_n) \to f(x)$.
- It is continuous at every point $x$ iff. the inverse image $f^{-1}(U)$, for every open $U \subset \mathbb{E}$ is open in $\mathbb{D}$.

# Dense, compact, separable and bounded spaces

Let $(\mathbb{D}, d)$ a metric space. A subset $A \subset \mathbb{D}$ is said to be:

- *dense* iff. its closure is the whole space $\mathbb{D}$.
- *separable* iff. it has a dense countable subset.
- *compact* iff. it is closed and every sequence in $A$ has a subsequence that convergence.
- is *totally bounded* iff. it can be covered by any finite union of $\varepsilon$-balls for all $\varepsilon > 0$.

A semimetric space is *complete* is every Cauchy sequence has a limit (i.e. $d(x_n, x_m) \to 0$, as $n, m \to \infty$).

# Normed space

## Definition

A *norm* $\|\|\| : \mathbb{D} \to \mathbb{R}_+$ is a map such that, for all $x, y \in \mathbb{D}$, $\alpha \in \mathbb{R}$:

1. $\|x\| = 0$ iff. $x = 0$.
2. $\|\alpha x\| = |\alpha| \|x\|$
3. triangular inequality: $\|x + y\| \leq \|x\| + \|y\|$

A seminorm does not necessarily satisfy 1. We say that the couple $(\mathbb{D}, \|\|\|)$ is a *normed space*.

We can define a metric $d(x, y) = \|x - y\|$.

# Important examples

We recall important normed spaces that we will use throughout the course.

1. **Euclidean spaces.** The Euclidean space $\mathbb{R}^d$, with $d \in \mathbb{N}^*$, equipped with the Euclidean norm
$$\|x\|_2^2 = \sum_{i=1}^{d} x_i^2 \ ,$$
is a normed space. We can, in fact, consider any other equivalent norm, such as $\max_{i \leq d} |x_i|$. The Borel $\sigma$-field $\mathcal{B}(\mathbb{R}^d)$ is generated by the intervals $(-\infty, x]$.
2. **Bounded functions.** Let $T$ be an arbitrary set. We denote by $\ell^\infty(T)$ the class of all bounded real-valued functions $x : T \to \mathbb{R}$. We will endow the space by the *uniform norm* on $T$:
$$\|x\|_T = \sup_{t \in T} |x(t)| \ ,$$
where we define pointwise the sum $(x_1 + x_2)(t) = x_1(t) + x_2(t)$ and product with a scalar $(\alpha x)(t) = \alpha x(t)$, for all $t \in T$.

   The space $\ell^\infty(T)$ contains all functions of finite sup-norm, i.e., such that $\|x\|_T < \infty$.

   Property: It is separable iff. the set $T$ is countable.

3. **Skorohod space.** Suppose that $T = [a, b]$ possibly the extended real line. The space of real-valued functions that are right-continuous with left limits that exist (càdlàg) is the Skorohod space and defined by $D(T, \mathbb{R})$ (or $D(T)$).

   The space $D([a, b])$ is NOT separable (w.r.t the sup-norm).

4. **Space of continuous functions on a compact.** Suppose that $T$ as before. We define $\mathcal{C}(T, \mathbb{R})$ to be the set of all continuous functions $x : [a, b] \to \mathbb{R}$.

   We have that $\mathcal{C}([a, b]) \subset D([a, b]) \subset \ell^\infty([a, b])$. We will always endow these space with the sup-norm 'inherited' from $\ell^\infty([a, b])$.

   The space $\mathcal{C}([a, b])$ is separable (w.r.t the sup-norm).

# $\sigma$-field

We recall in this section key concepts that will be used in all lectures.

## Definition

A collection $\mathcal{A}$ of subsets of a set $\Omega$ is a *$\sigma$-field/algebra in $\Omega$* if:

- $\Omega \in \mathcal{A}$
- If $A \in \mathcal{A}$, then $\Omega \setminus A \in \mathcal{A}$
- $\mathcal{A}$ is stable by countable union: $\cup_{j \geq 1} A_j \in \mathcal{A}$ for all $j \geq 1$ such that $A_j \in \mathcal{A}$.

And, in that case, we say that $\mathcal{A}$ is a *measurable space* (or the pair $(\Omega, \mathcal{A})$).

Let $(B, \mathcal{B})$ a measurable space.

Then, we say that a map $f : \Omega \to B$ is *measurable* if the preimage $f^{-1}(U) = \{x \in \Omega,\ f(x) \in U\}$ is measurable in $\Omega$ for all sets $U \in B$.

## Definition

If $\mathcal{A}$ is a collection of subsets of $\Omega$, not necessarily open, then there exists a smallest $\sigma$-field $\sigma(\mathcal{A})$ in $\Omega$ such that $\mathcal{A} \in \sigma(\mathcal{A})$. We define $\sigma(\mathcal{A})$ the *$\sigma$-field generated by $\mathcal{A}$*.

# Measure

## Definition

Let $\mathcal{A}$ be a $\sigma$-field in a set $\Omega$. The map $\mu : \mathcal{A} \to \bar{\mathbb{R}}$ is a *measure* if:

- $\mu(\emptyset) = 0$
- $\mu(A) \in [0, \infty]$ for all $A \in \mathcal{A}$
- For any disjoint countable sequence $\{A_j\}_{j \geq 1}$ of sets of $\mathcal{A}$, then $\mu\left(\cup_{j \geq 1} A_j\right) = \sum_{j \geq 1} \mu(A_j)$.

We say that $(\Omega, \mathcal{A}, \mu)$ forms a *measure space*.

If, in addition, $\mu(\Omega) = 1$, then $\mu$ is a probability measure.

We will use the notation of $P$ for a probability measure defined on the set $\Omega$ with $\sigma$-field $\mathcal{A}$ and $(\Omega, \mathcal{A}, P)$ forms a *probability space*.

# Borel measurability

Let $(\mathbb{D}, d)$ be a metric space.

- The *Borel* $\sigma$-field of $\mathbb{D}$ is the smallest $\sigma$-field containing all the open sets of $\mathbb{D}$.
- A function is *Borel measurable* relative to two metric spaces if it is measurable w.r.t. their Borel $\sigma$-field.
- A Borel-measurable map $X : \Omega \to \mathbb{D}$ defined on a probability space $(\Omega, \mathcal{A}, P)$ is referred to as a *random element/map* valued in $\mathbb{D}$.

We will use the notation $\mathcal{B}(\mathbb{D})$ to denote the Borel $\sigma$-field of $\mathbb{D}$.

## Remark

For Euclidean spaces, Borel measurability is the usual measurability.

We lastly recall an important result.

## Lemma

*A continuous map between two metric spaces is Borel-measurable.*

## Proof.

Exercise. □

# Modes of convergence of random vectors

Consider a random vector $X = (X^1, \dots, X^d)$ valued in $\mathbb{R}^d$, with $d \in \mathbb{N}^*$, of distribution function $P$. Let $d$ be a Euclidean distance on $\mathbb{R}^d$.

## Definition

A sequence $X_n$ converges to $X$ *almost surely* if

$$\mathbb{P}\left(\lim_{n\to\infty} d(X_n, X) = 0\right) = 1 \ . \tag{1}$$

It implies that $X_n$ converges to $X$ *in probability*, i.e., for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(d(X_n, X) > \varepsilon\right) = 0 \ . \tag{2}$$

Convergence almost surely is denoted by $X_n \xrightarrow[n\to\infty]{a.s.} X$. Convergence in probability by $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ and we will use the notation $X_n = X + o_{\mathbb{P}}(1)$.

# Weak convergence of random vectors

We will focus in this course specifically on *weak* convergence, also known as convergence in law or in distribution.

## Definition

Suppose the random sequence $X_1, X_2, \ldots$ to have a distribution function $F_n$ and p.d. $P_n$. $X_n$ converges in *distribution/weakly/in law* to a r.v. $X$, of d.f. $F$ and drawn from $P$ if, for all points $x \in T$ for which $F$ is continuous,

$$F_n(x) \underset{n\to\infty}{\longrightarrow} F(x) \ .$$

We denote it by $X_n \rightsquigarrow X$ (or $P_n \rightsquigarrow P$).

Weak convergence is inherent to the underlying distributions of the random maps, where the goal is to study the properties of the limit of distribution functions when $n$ tends to infinity (notice as well that we should consider a sequence of probability spaces but we ignore this technicality).

# Portmanteau Theorem

Weak convergence provides a series of equivalent properties referred to as *Portmanteau Theorem* that we state below.

## Theorem (Portmanteau Theorem I)

*Consider a random sequence of vectors $X_1, \ldots, X_n$, of p.d. $P_n$, and $X \sim P$. The following assertions are equivalent.*

1. $X_n \rightsquigarrow X$ *or* $P_n \rightsquigarrow P$
2. $P_n h \underset{n\to\infty}{\longrightarrow} Ph$, *for all* $h \in \mathcal{C}_b(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$
3. $\liminf_{n\infty} P_n(U) \geq P(U)$, *for all open sets* $U \subset \mathbb{R}^d$
4. $\limsup_{n\infty} P_n(F) \geq P(F)$, *for all closed sets* $F \subset \mathbb{R}^d$
5. $P_n(A) \underset{n\to\infty}{\longrightarrow} P(A)$, *for all* $P$*-continuity sets* $A$, *i.e., such that* $P(\partial A) = 0$ *where* $\partial A$ *denotes the boundary of* $A$.

# Continuous Mapping Theorem

Now that we have established the main characterizations, we are able to state the *Continuous Mapping Theorem* (CMT) that is fundamental to any statistical problem.

## Theorem (CMT I)

*Let $C \subseteq \mathbb{R}^d$ be a set such that $\mathbb{P}(X \in C) = 1$.*
*If $X_n \rightsquigarrow X$, then $\Phi(X_n) \rightsquigarrow \Phi(X)$ for any function $\Phi : \mathbb{R}^d \to \mathbb{R}^q$ that is continuous on $C$.*

# Classical limit theorems for random vectors

We recall two fundamental theorems when considering finite-dimensional metric spaces.

## Theorem (Law of Large Numbers)

*Let a sequence $X_1, X_2, \ldots, X_n$, with $n \in \mathbb{N}^*$ be an i.i.d. sequence of r.v.s. If $\mathbb{E}\|X_1\| < \infty$, then the strong LLN states that*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{a.s.} \mathbb{E}X \ .$$

*The weak LLN is satisfied for the convergence in probability.*

## Theorem (Central Limit Theorem)

*Suppose that the r.v.s have finite second moment, i.e., $\mathbb{E}\|X_1\|^2 < \infty$, then*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mathbb{E}X) \rightsquigarrow Z \ ,$$

*where $Z$ is a centered Gaussian r.v. of variance that of the $X$'s.*

# Stochastic symbols

We will often use convenient short expressions defined below for any sequence of r.v.s $X_n$.

- Convergence to zero in probability:

$$X_n = o_P(r_n) \quad \text{iff.} \quad X_n = Y_n r_n, \quad Y_n \xrightarrow[n\to\infty]{\mathbb{P}} 0 \ .$$

- Bounded in probability:

$$X_n = O_P(r_n) \quad \text{iff.} \quad X_n = Y_n r_n, \quad Y_n = O_P(1) \ .$$

The sequence $r_n$ represents the rate of convergence or of the bound. Notice that if the sequences are deterministic, then this notation recovers the classical $o$ and $O$ for sequences.

Some basic rules are:

$$\begin{aligned} o_P(1) + o_P(1) &= o_P(1) \\ o_P(1) + O_P(1) &= O_P(1) \\ o_P(1) O_P(1) &= o_P(1) \\ o_P(r_n) &= r_n o_P(1) \\ O_P(r_n) &= r_n O_P(1) \end{aligned}$$

# Today's outline

# Fundamental characterization of univariate r.v.s

The law of a real-valued random variable (r.v.) $X$ can be characterized by its cumulative distribution function (c.d.f.) defined, for all $t \in \mathbb{R}$, by

$$F(t) = \mathbb{P}\{X \leq t\} \ .$$

However, $F$ is unknown in statistical applications, and we only have access to a dataset of $n$ r.v. $X_1, \ldots, X_n$ of independent and identically distributed (i.i.d.) copies of $X$. A natural empirical counterpart of $F$ is given by:

$$\widehat{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, x]}(X_i) \ , \quad x \in \mathbb{R} \ , \tag{3}$$

where $1_{(-\infty,x]}(t) = \delta_t((-\infty, x])$ is the indicator function for the event $\{t \leq x\}$.

## Remark

The estimator $\widehat{F}_n$ is an unbiased estimator of $F$, has bounded moments, thus converges to $F$ for fixed $x \in \mathbb{R}$ by the Law of Large Numbers (LLN) $\widehat{F}_n(x) \xrightarrow[n\to\infty]{a.s.} F(x)$, for all $x \in \mathbb{R}$.

# Plug-in estimators

- In various applications, one might consider functionals of the c.d.f. $F$, say $\Phi(F) \in \mathbb{R}$, of natural estimator given by $\Phi(\widehat{F}_n)$ known as *plug-in estimator*.
- Suppose that $\Phi$ is continuous on $[0,1]$, then, by the continuous mapping theorem, the plug-in estimator is consistent: for any fixed point $x \in \mathbb{R}$, then

$$\Phi(\widehat{F}_n(x)) \xrightarrow[n\to\infty]{a.s.} \Phi(F(x)), \quad \text{for any fixed point} \quad x \in \mathbb{R} \ .$$

- In many problems, however, we need to establish the limiting distribution in a stronger sense: **uniformly over the sample space**. It is thus natural to require some similar kind of smoothness for the functional $\Phi$.

- **We wish for a Continuous Mapping Theorem (CMT).** If $\Phi$ is 'continuous at $F$' w.r.t. the sup-norm, i.e., for any other c.d.f. $G$, for any $\varepsilon > 0$, there exists $\delta > 0$, such that $\|F - G\|_\infty := \sup_{x\in\mathbb{R}} |F(x) - G(x)| \leq \delta$ implies $\|\Phi(F) - \Phi(G)\|_\infty \leq \varepsilon$.

The following example illustrates that a functional of the data might not be as simple as expected.

# Example 1: quantile functionals

Recall that the quantile at level $\alpha > 0$ for any c.d.f. $F$ can be defined by

$$q_\alpha(F) = \inf\{x \in [0,1],\ F(x) \geq \alpha\} =: F^{-1}(\alpha)\ ,$$

for $\alpha \in [a,b]$, with $0 < a < b < 1$, and of of plug-in estimator given by

$$q_\alpha(\widehat{F}_n) = \inf\{x \in [0,1],\ \widehat{F}_n(x) \geq \alpha\}\ .$$

Unfortunately the quantile functional does not take the form of an empirical process, being nonlinear in the $X_i$'s, but of a plug-in estimator. Nevertheless, it is important to specify how the estimator converges to its mean when $n \to \infty$. Notice that taking $\alpha = 1/2$ yields the median.

In fact, one can prove that, if $F$ is continuous on the interval $[F^{-1}(a) - \varepsilon, F^{-1}(b) - \varepsilon] \subset [0,1]$, for $\varepsilon > 0$, with continuous probability density function $f(x)$ valued in $\mathbb{R}^*$, then, the scaled empirical quantile function $\sqrt{n}(q_\alpha(\widehat{F}_n) - q_\alpha(F))$ is asymptotically equivalent to $-G_n(F^{-1}(\alpha))/f(F^{-1}(\alpha))$ and hence converges weakly to $-G(F^{-1}(\alpha))/f(F^{-1}(\alpha))$. This results from the *functional delta method*.

# Glivenko-Cantelli's Theorem 1.0

Empirical processes theory aims at proving stronger results than pointwise convergence, in particular uniform convergence. A first important result quantifying the distance between $\widehat{F}_n$ and $F$ dates back to the 1930's and is stated in the following fundamental theorem. It is also known as Uniform Law of Large Numbers (ULLN).

## Theorem (Glivenko-Cantelli, 1933)

$$\|\widehat{F}_n - F\|_\infty := \sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F(x)| \xrightarrow[n \to \infty]{a.s.} 0 \ . \tag{4}$$

This means that the sample paths of $\widehat{F}_n$ tend uniformly (in $\mathbb{R}$) to $F$ as the sample size $n$ tends to infinity.

Glivenko-Cantelli's Theorem ensures that any continuous mapping of the empirical c.d.f. is consistent plug-in estimator of its mean.

# Proof

1. Notice that for any fixed $x \in \mathbb{R}$, the r.v. $n\widehat{F}_n(x)$ is Binomial with mean $nF(x)$, hence it is an unbiased estimator.
2. For any fixed $x \in \mathbb{R}$, the strong LLN ensures that $\widehat{F}_n(x) \to F(x)$, and $\widehat{F}_n(x-) \to F(x-)$ a.s. Let $\varepsilon > 0$ be fixed. Consider a finite partition of the extended real line ($[-\infty, \infty]$):

$$-\infty = x_0 < x_1 < \ldots < x_N = \infty \ ,$$

such that $F(x_i-) - F(x_{i-1}) < \varepsilon$, and if the jump exceeds $\varepsilon$, it is considered as a point of the partition. For all $x \in [x_{i-1}, x_i)$, we have that

$$\widehat{F}_n(x) - F(x) \leq \widehat{F}_n(x_i-) - F(x_i-) + \varepsilon$$

$$\widehat{F}_n(x) - F(x) \geq \widehat{F}_n(x_{i-1}) - F(x_{i-1}) - \varepsilon \ .$$

Because both lower and upper bounds CV a.s. to $\varepsilon$, and that the partition is finite, we can deduce that uniformly over the partition they both CV a.s. to $\varepsilon$. Thus $\|\widehat{F}_n - F\|_\infty \leq \varepsilon$.

It is true for all $\varepsilon$ thus $\limsup_{n\infty} \|\widehat{F}_n - F\|_\infty = 0$.

We illustrate a natural application of this theorem.

# Example 2: Goodness-of-fit testing

Suppose we consider a known distribution $F_0$, say Gaussian, and we want to test the hypothesis of whether the random sample $X_1, \dots, X_n$ has been drawn from it. Precisely, at level $\alpha \in (0,1)$, we want to test the null hypothesis:

$$\mathcal{H}_0 : F = F_0, \qquad \text{against the alternative} \quad \mathcal{H}_1 : F \neq F_0 \ . \tag{5}$$

By Glivenko-Cantelli's Theorem, one can easily construct a test statistic by directly measuring the departure from the null in terms of the sup-norm $\|\widehat{F}_n - F_0\|_\infty$, defining the Kolmogorov-Smirnov test ($\times\sqrt{n}$), or by considering a continuous functional $\Phi(\widehat{F}_n) = \int_{\mathbb{R}} (\widehat{F}_n - F_0)^2(x) \mathrm{d}F_0(x)$, known as the Cramér-von Mises test statistic ($\times n$).

In order to reject the null hypothesis $\mathcal{H}_0$, it is required to know the distribution of the test statistic under the null, and this will be provided by the following result.

# Empirical processes

Consider the centered and scaled empirical c.d.f., traditionally termed as *empirical process* indexed on $\mathbb{R}$, (but we will not use this terminology):

$$\sqrt{n}(\widehat{F}_n - F) \ .$$

By the Central Limit Theorem (CLT), for any fixed point $x \in \mathbb{R}$, we have

$$\sqrt{n}(\widehat{F}_n - F)(x) \rightsquigarrow G(x) \ , \tag{6}$$

where $G(x)$ is a centered Gaussian r.v., with variance equal to $F(x)(1 - F(x))$.

A series of improvements is due to Kolmogorov, Donsker and Skorhokod in particular, that extended the CLT uniformly over the state space to empirical c.d.f.s. We present an important version below.

## Reminder

A Gaussian process $\{G(x),\ x \in \mathcal{X}\}$ is a stochastic process, such that for any finite subset $\mathcal{X}_k \subset \mathcal{X}$, the process $\{G(x),\ x \in \mathcal{X}_k\}$ is multivariate normal with continuous sample paths

# Donsker's theorem 1.0

## Theorem (Donsker, 1952)

*Let a sequence of i.i.d. r.v.s $X_1, X_2, \dots$ drawn from the Uniform distribution on $[0,1]$. Then,*

$$\|\sqrt{n}(\widehat{F}_n - F)\|_\infty \rightsquigarrow B(F) \ , \quad in \quad (D(\mathbb{R}),\ \|\cdot\|_\infty) \tag{7}$$

*where $B$ is a standard Brownian bridge process on $[0,1]$, i.e., it is a centered Gaussian process, with covariance function $\mathrm{Cov}(B(s), B(t)) = s \wedge t - st$, for all $s, t \in [0,1]$. The space $(D(\mathbb{R}),\ \|\cdot\|_\infty)$ is composed of all càdlàg functions on $\mathbb{R}$ endowed with the sup-norm $\|\cdot\|_\infty$, also known as the Skorokhod space.*

Consequences:

- For any bounded continuous function $h : D(\mathbb{R}) \to \mathbb{R}$:

$$\mathbb{E}h(\sqrt{n}(\widehat{F}_n - F)) \to \mathbb{E}h(B(F)) \ .$$

- For any continuous function $h : D(\mathbb{R}) \to \mathbb{R}$:

$$h(\sqrt{n}(\widehat{F}_n - F)) \rightsquigarrow h(B(F)) \ .$$

# Empirical averages and consistency

Suppose we want to estimate the true mean $\mu := \mathbb{E}X$ , then one can simply compute the sample average defined by

$$\frac{X_1 + \ldots + X_n}{n} .$$

Also, for any function $h : \mathbb{R} \to \mathbb{R}$, one can also approximate $\mathbb{E}h(X)$ by

$$\frac{h(X_1) + \ldots + h(X_n)}{n} .$$

In fact, we can prove *consistency* of both estimators (LLN), for any **fixed** function $h$. Suppose one has access to a data generating process providing an infinite amount of data points $X_1, X_2, \ldots, X_n, \ldots$. Then, we say that any statistic $T_n : (X_1, \ldots, X_n) \in \mathbb{R}^n \mapsto \mathbb{R}$ estimating the parameter $\theta := \mathbb{E}h(X)$ for instance, is consistent, if

$$T_n(X_1, \ldots, X_n) \underset{n \to \infty}{\longrightarrow} \theta .$$

This is a very important property for statisticians, and is far from being straightforward for more complex summaries than averages.

# Today's outline

# Empirical measure

Suppose now the r.v. $X$ to be valued in a more complicated space than $\mathbb{R}$, e.g., a generic Euclidean space $\mathcal{X} \subseteq \mathbb{R}^d$, with $d \geq 2$. A natural question lies in the generalization of the empirical distribution given above.

An empirical summary that has been studied in the literature is that of the probability measure $P$. Define the *empirical measure* by

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \ , \tag{8}$$

where $\delta_x(A)$ is the Dirac measure of the event $\{X \in A\}$. In particular

$$\begin{aligned} P_n(A) &:= \frac{1}{n} \sum_{i=1}^{n} 1_A(X_i) \\ &= \frac{1}{n} \{\text{number of observations } i \leq n: \quad X_i \in A\} \ , \end{aligned}$$

for any Borel subset $A \subset \mathcal{X}$. We refer to the empirical measure indexed by a collection of subsets $\mathcal{C}$ of $\mathcal{X}$ by

$$\{P_n(C), \quad C \in \mathcal{C}\} \ .$$

# Empirical measure

In some applications, it will be more convenient to consider empirical measures indexed by classes of functions, e.g., for averages. For any function $h : \mathcal{X} \to \mathbb{R}$, consider

$$P_n h := \frac{1}{n} \sum_{i=1}^{n} h(X_i) \ , \tag{9}$$

then the empirical measure indexed by a class of real-valued functions $\mathcal{H}$ is

$$\{P_n h, \quad h \in \mathcal{H}\}$$

considered as the empirical estimator of $Ph := \int_{\mathcal{X}} h(x) P(\mathrm{d}x)$ for a given $h \in \mathcal{H}$.

## Remark

Notice that taking $\mathcal{H}$ to be the collection $\{1_C, \ C \in \mathcal{C}\}$ recovers the first definition. It also recovers the univariate case, i.e., when $\mathcal{X} = \mathbb{R}$, by taking $\mathcal{C} = \{1\{(-\infty, x]\}, \ x \in \mathbb{R}\}$, then the collection of empirical measures indexed by the class of sets $\mathcal{C}$ is the collection of empirical c.d.f.s.

## Remark

Notice that we will often use this notation $(Ph)$ instead of the $\mathbb{E}$ as is helpful to understand the measure we use to integrate, and it can also be used to signify that the function $h$ is random.

# What we care about in empirical process theory

Empirical process theory aims at establishing theoretical guarantees for sequences of estimators uniformly over their index class (of sets or functions).

The first question is how to rigorously define such a supnorm, for which we can ensure some kind of measurability to be able to extend fundamental concepts such as weak convergence and permanence properties.

*What are the necessary conditions on the classes of sets $\mathcal{C}$ and functions $\mathcal{H}$ such that uniform versions of Glivenko-Cantelli and Donsker theorems hold true?*

# Glivenko-Cantelli's Theorem 2.0

We define such classes as being Glivenko-Cantelli below.

## Definition

The collection $\mathcal{C}$ of subsets of $\mathcal{X}$ is a $P$-Glivenko-Cantelli (GC) class of sets, if

$$\|P_n - P\|_{\mathcal{C}} := \sup_{C \in \mathcal{C}} |P_n(C) - P(C)| \xrightarrow[n \to \infty]{\mathbb{P}} 0 \ . \tag{10}$$

## Example

The class of half-closed intervals of $\mathcal{X} = \mathbb{R}$ is GC.

## Definition

The class $\mathcal{H}$ of functions valued in $\mathcal{X}$ is a $P$-GC class if

$$\|P_n - P\|_{\mathcal{H}} := \sup_{h \in \mathcal{H}} |P_n h - P h| \xrightarrow[n \to \infty]{\mathbb{P}} 0 \ , \tag{11}$$

where $Ph = \int h \mathrm{d}P$ is used as an abbreviation for $\mathbb{E}[h(X)]$, with $X$ being drawn from $P$.

We will say that a class is GC when the distribution is implicit. Similarly, we can extend UCL theorem to classes of functions under some conditions.

# Donsker's Theorem 2.0

## Definition

Suppose that

$$\sup_{h \in \mathcal{H}} |h(x) - Ph| < \infty, \quad \forall x \in \mathbb{R} \ ,$$

and that the functions $h$ are square-integrable $Ph^2 < \infty$. Then, the class $\mathcal{H}$ of functions valued in $\mathcal{X}$ is said to be Donsker if:

$$\|\sqrt{n}(P_n - P)\|_{\mathcal{H}} \rightsquigarrow B \ , \tag{12}$$

where $B$ is a Brownian bridge process on $\mathcal{H}$, i.e., a centered Gaussian process of covariance function $\mathrm{Cov}(B(h)B(g)) = Phg - (Ph)(Pg)$, for all $h, g \in \mathcal{H}$.

The convergence is understood in the space $\ell^\infty(\mathcal{H})$ is the set of all bounded real-valued functions endowed with the sup-norm over $\mathcal{H}$.

# What we wish for in statistical learning theory and in applications

- In recent applications related to real data analysis, statisticians might care about other questions than those related to the asymptotic regime of the estimators.
- Suppose one has a fixed sample size from a clinical experience for instance, where the size of the dataset $n$ corresponds to the number of patients. The goal of this experiment can be the effect of a treatment through $d$ indicators (fever, blood pressure, *etc.*) measured after one week of treatment. The statistician might be interested in lower-bounding the maximum value of the treatment effect based on the $n$ patients, with high probability.
- It takes the form of **concentration inequalities**.

Briefly, for any probability $\delta > 0$ we want to find a threshold $t_{\delta,n}$ depending on the class $\mathcal{H}$ such that

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} |Z_n(h) - Z(h)| \geq t_{\delta,n}(\mathcal{H})\right) \leq \delta \ . \tag{13}$$

We will see that the ability to prove such guarantees highly depends on the properties of the class $\mathcal{H}$ (similarly of $\mathcal{C}$).

Once those conditions are met, what can we say about the rate of the threshold $t_{\delta,n}(\mathcal{H})$ such that Eq. (13) holds true?

# Today's outline

# Empirical risk minimization (ERM)

- Let $X$ be a r.v., of p.d. $P$, and defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and valued in $\mathcal{X}$, think of it as a subset of $\mathbb{R}^d$, with $d \in \mathbb{N}^*$, also called a **feature space**.
- Consider a set of parameters $\Theta$, usually subset of $\mathbb{R}^d$ as well.

Based on $X$, we can define a **loss** function as

$$\ell : \Theta \times \mathcal{X} \longrightarrow \mathbb{R}_+ \ .$$

The goal of ERM is to estimate the **best** parameter $\theta^* \in \Theta$ that minimizes the expected loss function, known as the **risk**

$$\theta^* \in \arg\min_{\theta \in \Theta} \underbrace{\mathbb{E}[\ell(\theta, X)]}_{=:\mathcal{R}(\theta)} \ .$$

- In practice, $P$ is unknown, and we only have access to an i.i.d. set of r.v.s drawn from $P$:

$$X_1, \ldots, X_n, \quad n \in \mathbb{N}^*$$

- A simple unbiased empirical estimator based on the data is given by:

$$\mathcal{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i)$$

Thus, we can estimate, based on the data, the **empirical minimizer**

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta} \mathcal{R}_n(\theta) \ .$$

We hope that the empirical minimizer is *close* to the true (**oracle**) optimal solution $\theta^*$, depending on the sample size $n$, both dimensions of the feature space $\mathcal{X}$ and of the parameter space $\Theta$, *etc.*

In fact, we will need to control the uniform fluctuations of this statistical error

$$\sup_{\theta \in \Theta} \ \frac{1}{n} \sum_{i=1}^{n} \{\ell(\theta, X_i) - \mathbb{E}[\ell(\theta, X)]\} \ . \tag{14}$$

# Example 1: binary classification (simple)

- Suppose the $X$ to be the input r.v. valued in $\mathcal{X}$ (e.g. images), and $Y$ to be the output label that is valued in $\{0, 1\}$ (e.g. cats or dogs).
- Consider a collection of classifiers $\mathcal{H} = \{h : \mathcal{X} \to \{0, 1\}, \quad h \text{ measurable}\}$.

The associate **binary loss** is defined by

$$\ell_h : (x, y) \in \mathcal{X} \times \{0, 1\} \mapsto 1\{h(x) \neq y\} \ .$$

It equals to 1 if the predictor $h$ gives the wrong label to the input $x$.

The goal is to find the optimal classifier that minimizes the associated **binary risk**

$$h^* \in \arg\min_{h \in \mathcal{H}} \underbrace{\mathbb{P}(h(X) \neq Y)}_{=:\mathcal{R}(h)} \ .$$

The *true* classifier function known as the Bayes classifier, $h^* : \mathcal{X} \to \{0, 1\}$ can be proved to be a function of the posterior distribution given by

$$\eta(x) = \mathbb{P}(y = 1 | X = x) \ .$$

Based on an i.i.d. random sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ the goal is to learn a classifier $\hat{h}$, that predicts the labels of any new set of observations $X, Y$'s with a smallest empirical risk, defined by

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} 1\{h(X_i) \neq Y_i\} \ .$$

**We will analyze the binary risk in the Exercise session for various probabilistic settings.**

# Example 2: Least-squares estimation (LSE)

- Suppose now that the response variable $Y$ is a real-valued r.v., that we want to model by a covariate $X$, possibly valued in $\mathcal{X}$ as before.

- We can think of the classical regression model

$$Y = h(X) + \varepsilon \ ,$$

where the r.v. $\varepsilon$ typically models noise of standard Gaussian distribution, and is independent of $X$.

The goal of (non)parametric regression is to estimate the **best** function $h$ minimizing a risk functional, typically defined as the **least-squares loss**

$$\mathcal{R}(h) = \mathbb{E}[(Y - h(X))^2] \ .$$

When based on a dataset of i.i.d. r.v.s $(X_1, , Y_1), \ldots, (X_n, Y_n)$, the empirical minimizer is given by

$$\widehat{h} \in \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - h(X_i))^2 \ .$$

## Example (Linear regression)

The simplest class of regressors is that of linear functions $h : (\theta, x) \mapsto \theta^T x$, for all $\theta, x \in \mathbb{R}^d$, and $^T$ denotes the transpose operator.

# Maximum likelihood estimation (MLE)

Consider a collection of distributions indexed by a parameter space $\Theta \subseteq \mathbb{R}^p$, with $p \in \mathbb{N}^*$, of strictly positive densities $\{p_\theta\}_{\theta \in \Theta}$ w.r.t. a common measure $\mu$ ($p_\theta = \mathrm{d}P_\theta/\mathrm{d}\mu$). Suppose the r.v. $X$ is drawn from an unknown distribution of true parameter $\theta^*$, we would like to estimate it using the likelihood ratio as cost function, i.e., define

$$\mathcal{L}_\theta(x) := \log\left(\frac{p_{\theta^*}(x)}{p_\theta(x)}\right) .$$

The true parameter minimizes the associated risk, also known as the *Kullback-Leibler divergence* between the two distributions $p_{\theta^*}$ and $p_\theta$

$$\mathcal{R}(\theta) = \mathbb{E}\left[\log\left(\frac{p_{\theta^*}(x)}{p_\theta(x)}\right)\right] ,$$

where we integrate w.r.t. $P_\theta$. Given a random $n$-i.i.d. sample $X_1, \ldots, X_n$ being copies of $X$, we aim to find instead the optimal empirical parameter $\widehat{\theta}$ that minimizes the empirical risk

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta} \frac{1}{n}\sum_{i=1}^n \log\left(\frac{p_{\theta^*}(X_i)}{p_\theta(X_i)}\right) .$$

Notice that $\widehat{\theta}$ is invariant w.r.t. $\theta^*$, and consider instead

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta} \frac{1}{n}\sum_{i=1}^n \log\left(\frac{1}{p_\theta(X_i)}\right) = \arg\max_{\theta \in \Theta} \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i) .$$

**In fact, all the solution of these problems can be redefined as $M$-estimators.**

# What we saw in this lecture

1. We recalled fundamental definitions and properties for metric spaces, to be able to understand modes of convergence of random sequences of random maps
2. We recalled the continuous mapping Theorem, LLN and CLT
3. We defined empirical process indexed by arbitrary sets
4. We stated the first important extensions of uniform limit theorems
5. We exposed important learning examples in statistics

# What we will see next week

- We will study how to derive probabilistic bounds to quantify the speed of the deviation of averages of r.v.s w.r.t their mean, known as basic *concentration inequalities*.
- We will focus on bounds that have exponential decay, under various assumptions on the moments of the r.v.s
- Importantly, those results allow to assess how averages concentrate around their mean for **fixed** sample size as follows

Let a sequence of centered, independent and integrable r.v.s $X_1, X_2, \ldots, X_n$ of average $Z_n$. We want to provide upper-bounds depending on $n$ of the probability

$$\mathbb{P}\{|Z_n - PZ_n| \geq t\}$$

for all $t > 0$.