Asymptotic Statistics and Dependence

Lionel Truquet
email: lionel.truquet@ensai.fr

Contents

Chapter 0

Introduction

0.1 Why asymptotic theory ?

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with common probability distribution $P \in \mathcal{P}$ where $\mathcal{P}$ is a subset of the set of probability measures. Suppose we want to estimate a parameter $\theta = \theta(P)$ of this distribution. To this end, we would like to find an estimator $\hat{\theta}_n = T_n(X_1, \ldots, X_n)$ of θ. Here T_n is a measurable mapping.

To assess accuracy of $\hat{\theta}_n$, the two following questions are natural.

- Do we have convergence (in probability, almost surely, in quadratic mean...) of $\hat{\theta}_n$ to θ as $n \to \infty$?
- Can we exhibit a convergence rate? For instance, which kind of sequence $(r_n)_n$ of positive real numbers, diverging to infinity, entails that

$$\limsup_{n\to\infty} r_n \mathbb{E}\left(\left|\hat{\theta}_n - \theta\right|^2\right) < \infty?$$

  And can we exhibit a non-degenerate limiting distribution for $\sqrt{r_n}\left(\hat{\theta}_n - \theta\right)$, which is useful for constructing confidence intervals and statistical tests?

For studying these problems, general limit theorems are available and many results exist for quite sophisticated statistical models. On the other hand, asymptotic theory is not suitable for evaluating the quality of the estimator for a fixed value of n. Non-asymptotic statistics can then be useful. However, getting accurate non-asymptotic results often requires to work with simpler models, especially if our aim is to obtain sharp constants. In this sense, both theory are complementary.

The aim of this course is to present some general classes of statistical models for which a nice asymptotic theory can be obtained. In the following sections, we introduce some examples of estimators which will be studied in the next chapters.

0.2 Parametric M-estimators

A parametric estimator is obtained as a solution of a minimization problem

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i),$$

for some $\Theta \subset \mathbb{R}^d$ and $m_\theta : \mathbb{R}^k \to \mathbb{R}$ is a measurable mapping for every $\theta \in \Theta$. Note that we implicitly assume that such an argmin exists. When it is not unique for some $\omega \in \Omega$, we assume that $\hat{\theta}_n(\omega)$ is one of the possible argmin.

We provide some specific examples below.

1. Maximum likelihood estimators (MLE) corresponds to the case $m_\theta(x) = -\log p_\theta(x)$ where $\mathcal{P} = \{p_\theta \cdot \mu : \theta \in \Theta\}$, μ is a measure of reference (e.g. counting measure on $\mathbb{N}$, Lebesgue measure on $\mathbb{R}^k$) and $\nu_\theta := p_\theta \cdot \mu$ denotes the probability measure defined by $\nu_\theta(A) = \int_A p_\theta d\mu$ for any $A \in \mathcal{B}(\mathbb{R}^k)$, the Borel sigma-field of $\mathbb{R}^k$. For instance,

   - the exponential distribution corresponds to $p_\theta(x) = \theta \exp(-\theta x)$ for $\theta > 0$ and μ is the Lebesgue measure on $\mathbb{R}_+$. A generalization is given by the gamma distribution with parameters $\theta_1, \theta_2 > 0$, for which $p_\theta(x) = x^{\theta_1 - 1} \theta_2^{\theta_1} e^{-\theta_2 x} / \Gamma(\theta_1)$. Here $\Gamma(z) = \int_0^\infty x^{z-1} \exp(-x) dx$ for $z > 0$.
   - The Poisson distribution with parameter $\theta > 0$ has the probability density $p_\theta : x \mapsto e^{-\theta} \theta^x / x!$ with respect to the counting measure on $\mathbb{N}$.
   - The case $p_\theta(x) = \exp\left(\phi(\theta)^T S(x) - Z(\theta)\right)$ for some measurable mapping $S : \mathbb{R}^k \to \mathbb{R}^\ell$ and mappings $\phi : \Theta \to \mathbb{R}^\ell$, $Z : \Theta \to \mathbb{R}$ with $\phi(\theta)^T$ denoting the transpose of the column vector $\phi(\theta)$ corresponds to the exponential family which contains the two previous examples as special cases as well many other such that the multivariate Gaussian distributions).

2. Regression estimators with $X_i = (Y_i, Z_i) \in \mathbb{R} \times \mathbb{R}^p$ satisfying $Y_i = r_\theta(Z_i) + \varepsilon_i$ with $(Z_1, \varepsilon_1), \ldots, (Z_n, \varepsilon_n)$ i.i.d. such that $\mathbb{E}(\varepsilon_i | Z_i) = 0$. It is often assumed that $r_\theta(Z_1)$ and ε_1 are square integrable and the least squares estimator (LSE) of θ is defined by

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - r_\theta(X_i)\right)^2.$$

   Linear regression corresponds to the case $r_\theta(Z_i) = Z_i^T \theta$ where θ is vector of $\mathbb{R}^p$.

   When the probability distribution P of the pair (Z_i, ε_i) is not assumed to be an element of a parametric family, the model is usually called semi-parametric (i.e. the probability distribution of the observations can be described by a finite-dimensional parameter θ and an infinite-dimensional parameter P). However, in our context, it will be possible to estimate θ independently from P.

Regression models can be extended to dependent data $(Y_t, Z_t)_{1\leq t\leq n}$ where t denotes the time. Sometimes $Z_t = Y_{t-1}$ and the model is said to be autoregressive. Autoregressive models are widely used in many areas, e.g. for analyzing prizes dynamics in finance, the evolution of temperatures, species dynamics in ecology...

3. A binary regression model with Y_i taking values in $\{0,1\}$ and X_i taking values in $\mathbb{R}^d$ can be obtained setting

$$\mathbb{P}\left(Y_i = 1 | Z_i = z\right) = F\left(z^T\theta\right),$$

where F is a cumutative distribution function and $\theta \in \mathbb{R}^d$ is an unknown parameter. When $F(u) = \frac{\exp(u)}{1+\exp(u)}$, we call this model a logistic regression model and when

$$F(u) = \int_{-\infty}^{u} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx,$$

we call it a probit regression model. Many practical applications can be considered. For instance, in epidemiology, Y_i takes the value 1 if the patient i has a given disease (e.g. cancer) and Z_i is a vector containing some information for the patient (e.g. age, weight, smoking or not...). It is possible to consider conditional likelihood estimation from the conditional distribution of Y_i given $Z_i = z$, which is given by

$$p_\theta\left(y|z\right) = F\left(z^T\theta\right)^y \left(1 - F\left(z^T\theta\right)\right)^{1-y}, \quad (y,z) \in \{0,1\} \times \mathbb{R}^d.$$

The condition Maximum Likelihood Estimator (MLE) is given by

$$\begin{aligned}
\hat{\theta}_n &= \underset{\theta\in\mathbb{R}^d}{\operatorname{argmax}} \prod_{i=1}^{n} p_\theta\left(Y_i|Z_i\right) \\
&= \underset{\theta\in\mathbb{R}^d}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} \left\{Y_i \log\left(F\left(Z_i^T\theta\right)\right) + \left(1-Y_i\right)\log\left(1 - F\left(Z_i^T\theta\right)\right)\right\}.
\end{aligned}$$

0.3 Measurability of M-estimators

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $\hat{\theta}_n = \operatorname{argmin}_{\theta\in\Theta} M_n(\theta)$ where for any $\theta \in \Theta$, $M_n(\theta)$ is a random variable. A crucial example concerns the case $M_n(\theta) = S_n\left(\theta, X_1, \ldots, X_n\right)$ where S_n is a measurable real-valued mapping on a suitable product space and $X_1, \ldots, X_n$ are random variables taking values the same measurable space (typically $\mathbb{R}^k$) . The case

$$S_n\left(\theta, X_1, \ldots, X_n\right) = \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i),$$

is of special interest.

We start with a very simple result ensuring measurability of M-estimators. The uniqueness assumption is not easy to check and the compactness of Θ is a limitation. We will consider a more general result just after but without giving a proof.

Proposition 1. *If Θ is a compact subset of $\mathbb{R}^d$, the mapping $\theta \mapsto M_n(\theta)$ is a.s. continuous over Θ (i.e. there exists $\Omega' \in \mathcal{A}$ such that $\mathbb{P}(\Omega') = 1$ and for all $\omega \in \Omega'$, the mapping $\theta \mapsto M_n(\theta)_\omega$ is continuous over Θ) and there exist $\hat{\theta}_n : \Omega \to \Theta$ such that for all $\omega \in \Omega'$ and all $\theta \in \Theta \setminus \left\{\hat{\theta}_n(\omega)\right\}$, $M_n\left(\hat{\theta}_n\right)_\omega < M_n(\theta)_\omega$. If $\theta_1 \in \Theta$, we set $\hat{\theta}_n(\omega) = \theta_1$ for $\omega \in \Omega \setminus \Omega'$. Then $\hat{\theta}_n$ is measurable.*

Proof of Proposition 1. Let A be an open subset of $\mathbb{R}^d$. It is enough to show that $\left\{\hat{\theta}_n \in A\right\}$ is a measurable set which is true as soon as $\left\{\hat{\theta}_n \in A\right\} \cap \Omega'$ is a measurable set. Then

$$\left\{\hat{\theta}_n \in A\right\} \cap \Omega' = \left\{\min_{\theta \in \Theta \setminus A} M_n(\theta) > \min_{\theta \in \Theta} M_n(\theta)\right\} \cap \Omega'.$$

Indeed, the set $\Theta \setminus A$ is a compact subset of $\mathbb{R}^d$ (as the intersection between a compact set and a closed set) and if $\omega \in \Omega'$, the continuous mapping $\theta \mapsto M_n(\theta)_\omega$ reaches its minimum over $\Theta \setminus A$. By definition of $\hat{\theta}_n$, the minimal value is larger than $\min_\theta M_n(\theta)_\omega = M_n\left(\hat{\theta}_n\right)_\omega$. Additionally, for any compact set K included in Θ, $\min_{\theta \in K} M_n(\theta)$ is a random variable. Indeed, one can write $\min_{\theta \in K} M_n(\theta) = \inf_{\theta \in \widetilde{K}} M_n(\theta)$ where $\widetilde{K}$ is finite or infinite or numerable subset of K. For instance, one can set

$$\widetilde{K} = \left\{x_i^{(k)}, 1 \leq i \leq p_k, k \in \mathbb{N}^*\right\}, \quad K \subset \cup_{i=1}^{p_k} B\left(x_i^{(k)}, \frac{1}{k}\right),$$

where the $x_i^{(k)}$'s are suitable points in K. Finally, we have shown that the set

$$\left\{\min_{\theta \in \Theta \setminus A} M_n(\theta) > \min_{\theta \in \Theta} M_n(\theta)\right\} \cap \Omega'$$

is a measurable set which leads to the result.□

We next give a more general result which is applicable in a quite general framework, provided that the random function is continuous with respect to the parameter of interest. A proof of the following result can be found in Niemiro (1992), Corollary 1.

Theorem 1. *Suppose that $M_n(\theta) = S_n(\theta, X_1, \ldots, X_n)$ with $\theta \mapsto S_n(\theta, X_1, \ldots, X_n)$ continuous a.s. on Θ, $(x_1, \ldots, x_n) \mapsto S_n(\theta, x_1, \ldots, x_n)$ measurable for any $\theta \in \Theta$ and*

$$\Gamma(X_1, \ldots, X_n) := \left\{\theta' \in \Theta : S_n(\theta', X_1, \ldots, X_n) = \inf_{\theta \in \Theta} S_n(\theta, X_1, \ldots, X_n)\right\}$$

is almost surely non empty. Then there exists

$$\hat{\theta} = \Delta_n(X_1, \ldots, X_n) = \operatorname*{argmin}_{\theta \in \Theta} M_n(\theta) \in \Gamma(X_1, \ldots, X_n)$$

with Λ_n measurable.

To apply the previous result, one can simply check the condition on $\Gamma(X_1, \ldots, X_n)$ by considering the behavior of $M_n(\theta)$ when $\|\theta\| \to \infty$.

0.4 Kernel density estimation

We now give a classical example in non-parametric estimation. Here $\theta(P)$ is simply the probability density of the probability measure P. The parameter space Θ is now a subset of the family of probability densities with respect to the Lebesgue measure λ_k on $\mathbb{R}^k$. We do not want to make a parametric assumption on Θ, but only suitable regularity conditions (e.g. continuity, differentiability...). The idea for estimating the common probability density f of some identically distributed random variables $X_1, \ldots, X_n$ taking values in $\mathbb{R}^k$, is to use the properties of convolution products. Let us consider another probability density $K : \mathbb{R}^k \to \mathbb{R}_+$ that will be called a kernel and for some $h > 0$, let us define $K_h(x) = h^{-k}K(x/h)$ for $x \in \mathbb{R}^k$. Note that K_h is also a probability density. We know that the convolution product $K_h * f$ defined by

$$K_h * f(x) = \int_{\mathbb{R}^k} K_h(x-y)f(y)\lambda_k(dy), \quad x \in \mathbb{R}^k$$

approximates f (for instance in $\mathbb{L}^1$) when $h \to 0$. Since f is unknown, we use the empirical distribution $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and define

$$\hat{f}_h(x) = \int_{\mathbb{R}^k} K_h(x-y)\mathbb{P}_n(dy) = \frac{1}{n}\sum_{i=1}^n K_h\left(x - X_i\right).$$

We then see that $\mathbb{E}\hat{f}_h(x) = K_h * f(x)$. The bias for estimating $f(x)$ is defined by $K_h * f(x) - f(x)$ and we have to choose h as small as possible to decrease the bias. However the variance of $\hat{f}_h(x)$ generally increases when h becomes small. Later in this course, we will see the the variance is of order $(nh^k)^{-1}$. Then a suitable choice $h = h_n$ has to be made in practice.

Standard examples of kernels are the Gaussian kernel $K(x) = (2\pi)^{-k/2}\exp\left(-\frac{\|x\|^2}{2}\right)$ or the indicator kernel $K(x) = 2^{-k}\mathbb{1}_{\|x\|_\infty \leq 1}$ where $\|x\|_\infty = \max_{1\leq i\leq d}|x_i|$. Note that for this second kernel, $\hat{f}_h(x)$ denotes the proportion of observations inside the ball $B_\infty(x,h) = \left\{y \in \mathbb{R}^k : \|y-x\|_\infty \leq h\right\}$ divided by the volume of the ball. This gives another intuition for using such estimator for the probability density and explains why the hyper-parameter h is called the "bandwidth".

0.5 Non-parametric regression estimation

In this section we assume that $Y_i = r(Z_i) + \varepsilon_i$, $1 \leq i \leq n$, with $r(Z_1)$ and ε_1 integrable and $\mathbb{E}\left[\varepsilon_i | Z_i\right] = 0$ a.s. We only observe $X_i = (Z_i, Y_i)$ for $1 \leq i \leq n$ and we do not assume that $r : \mathbb{R}^d \to \mathbb{R}$ is contained in a predetermined parametric family of functions. A standard estimator for r is the Nadaraya-Watson estimator with

$$\hat{r}_h(z) = \frac{n^{-1}\sum_{i=1}^n Y_i K_h\left(z - Z_i\right)}{n^{-1}\sum_{i=1}^n K_h\left(z - Z_i\right)}, \quad z \in \mathbb{R}^d,$$

where $h > 0$ is a bandwidth and K is a kernel. Note that the denominator of $\hat{r}_h(z)$ is precisely an estimator of the density f_Z of the random vector Z at point z. If we assume

that the pair (Y_1, Z_1) has a density $f_{Y,Z}$ with respect to the Lebesgue measure on $\mathbb{R} \times \mathbb{R}^d$, one can note that

$$r(z) = \mathbb{E}[Y_1 | Z_1 = z] = \frac{\int_{\mathbb{R}} y f_{Y,Z}(y,z)dy}{f_Z(z)}.$$

Then

$$\mathbb{E}[Y_1 K_h(z - Z_1)] = \int_{\mathbb{R}} y \int_{\mathbb{R}^k} K_h(z - z') f_{Y,Z}(y, z')dz'dy \approx \int y f_{Y,Z}(y,z)dy = f_Z(z)r(z),$$

which justifies the use of such estimator for estimating $r(z)$.

There exist other methods based on the same idea of local averaging.

For instance, the k nearest neighbors (kNN) estimator of $r(z)$ is defined by

$$\hat{r}(z) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}_{\left\{\|z - Z_i\| \leq \hat{\tau}_{n,k}(z)\right\}} Y_i, \quad \hat{\tau}_{n,k}(z) = \inf \left\{ \tau \geq 0 : \sum_{i=1}^{n} \mathbb{1}_{\{\|z - Z_i\| \leq \tau\}} \geq k \right\}.$$

Note that $\hat{\tau}_{n,k}(z)$ corresponds to the kth smallest value of $\|z - Z_i\|$, $1 \leq i \leq n$. We then simply average the values of Y_i for which Z_i is among the kNN of z in the sample.

Note that this estimator depends on k and of a norm. For the norm, one can take the Euclidean norm but not only. We remind that all the norms are equivalent on $\mathbb{R}^d$. For k, the intuition is that a small value of k will lead to a small bias but to a large variance (we localize a lot the average) while a large value of k will produce a large bias and a small variance (the average is over a large number of variables and we do not localize sufficiently). This hyper-parameter k plays the same role as the bandwidth for the Nadaraya-Watson estimator. Note that $\hat{r}(z)$ is similar to this estimator with the indicator kernel and a random bandwidth.

Let us mention that both estimators can be used when Y_i takes values in $\{0, 1\}$ (we use the term classification instead of regression) and produce estimators of $r(z) = \mathbb{P}(Y_1 = 1 | Z_1 = z)$.

For classifying a new observation Z_{n+1} (for which the label Y_{n+1} is not known) with the nearest-neighbor approach, one simply predict 1 if $\hat{r}(Z_{n+1}) \geq 1/2$ (i.e. if there is a majority of 1 in the kNN of Z_{n+1}) and 0 otherwise.

Chapter 1

Complements on stochastic convergence

1.1 Reminders

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. On the space $\mathbb{R}^k$ endowed with its Borel sigma-field $\mathcal{B}\left(\mathbb{R}^k\right)$, we denote by $\|\cdot\|$ an arbitrary norm (using the same notation whatever the value of k). A sequence of random variables $(Y_n)_{n\in\mathbb{N}}$ taking values in $\mathbb{R}^k$ converges

1. almost surely (a.s.) to Y if $\exists\ \widetilde{\Omega} \in \mathcal{A}$ with $\mathbb{P}\left(\widetilde{\Omega}\right) = 1$ and $\forall \omega \in \widetilde{\Omega}$, $\lim_{n\to\infty} Y_n(\omega) = Y(\omega)$,

2. in probability to Y if $\forall\ \epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(\|Y_n - Y\| > \epsilon\right) = 0$,

3. in distribution (or weakly, or in law) to Y if for all mapping $h : \mathbb{R}^k \to \mathbb{R}$ continuous and bounded, $\lim_{n\to\infty} \mathbb{E}\left[h(Y_n)\right] = \mathbb{E}\left[h(Y)\right]$.

We use the respective notations $Y_n \overset{a.s.}{\to} Y$, $Y_n \overset{p}{\to} Y$ and $Y_n \hookrightarrow Y$ for the a.s. convergence, convergence in probability and convergence in distribution.

The following result ensures the stability of the three convergence properties after composition with a continuous mapping.

Theorem 2 (Continuous mapping theorem)**.** *Let Y_n and Y be some random vectors taking values in $\mathbb{R}^k$ such that $Y_n \overset{a.s.}{\to} Y$ (resp. $Y_n \overset{p}{\to} Y$, $Y_n \hookrightarrow Y$) and $f : \mathbb{R}^k \to \mathbb{R}^\ell$ a mapping, continuous at any point of $\mathcal{C} \in \mathcal{B}\left(\mathbb{R}^k\right)$ such that $\mathbb{P}\left(Y \in \mathcal{C}\right) = 1$. Then $f(Y_n) \overset{a.s.}{\to} f(Y)$ (resp. $f(Y_n) \overset{p}{\to} f(Y)$, $f(Y_n) \hookrightarrow f(Y)$).*

Proof. It is obvious for the almost sure convergence. For the convergence in probability, let $\epsilon > 0$ and k be a positive integer. Set

$$B_k = \left\{ x \in \mathcal{C} : \exists\, y \in \mathbb{R}^k \text{ s.t. } |x - y| < \frac{1}{k} \text{ and } |f(x) - f(y)| > \epsilon \right\}.$$

We have $\cap_{k\geq 1} B_k = \emptyset$ by continuity of f and $B_{k+1} \subset B_k$. Then

$$\{\|f(Y_n) - f(Y)\| > \epsilon\} \subset \left\{\|Y_n - Y\| \geq \frac{1}{k}\right\} \cup \{Y \in B_k\} \cup \{Y \notin \mathcal{C}\}.$$

Indeed if $Y(\omega) \in \mathcal{C}$ and $\|f(Y_n(\omega) - f(Y(\omega))\| > \epsilon$, we have either $Y(\omega) \in B_k$ or $\|Y_n(\omega) - Y(\omega)\| \geq \frac{1}{k}$. We then get

$$\mathbb{P}\left(\|f(Y_n) - f(Y)\| > \epsilon\right) \leq \mathbb{P}\left(\|Y_n - Y\| \geq \frac{1}{k}\right) + \mathbb{P}\left(Y \in B_k\right) + \mathbb{P}\left(Y \notin \mathcal{C}\right).$$

Since $\mathbb{P}\left(Y \notin \mathcal{C}\right) = 0$ and $Y_n \xrightarrow{p} Y$, we have

$$\overline{\lim_n}\, \mathbb{P}\left(\|f(Y_n) - f(Y)\| > \epsilon\right) \leq \mathbb{P}(Y \in B_k).$$

Since $\lim_k \mathbb{P}(Y \in B_k) = 0$ by the continuity properties of the measure, we get the result.

For the convergence in distribution, the proof will be given after the next result.

1.2 Portmanteau lemma

In what follows, for a Borel set A of $\mathbb{R}^k$, we denote by ∂A the boundary of the set A. It is defined by $\partial A = \overline{A} \setminus \mathring{A}$ where $\overline{A}$ is the closure of A (that is the smallest closed set containing A) and $\mathring{A}$ is the interior of A (that is the largest open set included in A).

Lemma 1 (Portmanteau lemma)**.** *The following assertions are equivalent.*

1. $Y_n \hookrightarrow Y$.
2. *For every mapping* $f : \mathbb{R}^k \to \mathbb{R}$ *Lipschitz and bounded,* $\lim_n \mathbb{E}\left[f(Y_n)\right] = \mathbb{E}\left[f(Y)\right]$.
3. *If* F *is a closed set,* $\overline{\lim}_n \mathbb{P}\left(Y_n \in F\right) \leq \mathbb{P}\left(Y \in F\right)$.
4. *If* O *is an open set,* $\underline{\lim}_n \mathbb{P}\left(Y_n \in O\right) \geq \mathbb{P}\left(Y \in O\right)$.
5. *If* $A \in \mathcal{B}\left(\mathbb{R}^k\right)$ *is a continuity set for* $\mathbb{P}_Y$*, i.e.* $\mathbb{P}\left(Y \in \partial A\right) = 0$*, then* $\lim_n \mathbb{P}\left(Y_n \in A\right) = \mathbb{P}\left(Y \in A\right)$.

Proof. 1. $\Rightarrow$ 2. follows from the fact that a Lipschitz function is also continuous.

Let us show that 2. $\Rightarrow$ 3. For $\epsilon > 0$, let $f_\epsilon(y) = \left(1 - \frac{d(y,F)}{\epsilon}\right)_+$ (where $x_+ = \max(x, 0)$). We remind that the distance $d(y, F) = \inf_{f \in F} \|y - f\|$ is always attained for some $f_0 \in F$. It is automatic to check that $\mathbb{1}_F(y) \leq f_\epsilon(y)$, $\lim_{\epsilon \to 0} f_\epsilon(y) = 0$ if $y \notin F$ and $f_\epsilon(y) = 1$ if $y \in F$. Moreover for $y, y' \in \mathbb{R}^k$,

$$|f_\epsilon(y) - f_\epsilon(y')| \leq \frac{\|y - y'\|}{\epsilon}.$$

The mapping f_ϵ is Lipschitz and bounded and then $\lim_n \mathbb{E}[f_\epsilon(Y_n)] = \mathbb{E}[f_\epsilon(Y)]$. We then get

$$\overline{\lim_n} \mathbb{P}(Y_n \in F) \leq \overline{\lim_n} \mathbb{E}[f_\epsilon(Y_n)] = \mathbb{E}[f_\epsilon(Y)].$$

We conclude by letting $\epsilon \to 0$, using the dominated convergence theorem which leads to $\lim_{\epsilon \to 0} \mathbb{E}[f_\epsilon(Y)] = \mathbb{P}(Y \in F)$.

3. $\Leftrightarrow$ 4. It is obvious since the complement of an open set (resp. a closed set) is a closed set (resp. an open set) and for any real-valued sequence $(x_n)_n$, $\overline{\lim}_n(-x_n) = -\underline{\lim}_n x_n$.

3. + 4. $\Rightarrow$ 5. We note that

$$\mathbb{P}\left(Y \in \overline{A}\right) \geq \overline{\lim_n} \mathbb{P}\left(Y_n \in \overline{A}\right) \geq \overline{\lim_n} \mathbb{P}(Y_n \in A) \geq \underline{\lim_n} \mathbb{P}(Y_n \in A) \geq \underline{\lim_n} \mathbb{P}\left(Y_n \in \mathring{A}\right) \geq \mathbb{P}\left(Y \in \mathring{A}\right).$$

Since the continuity property ensures that $\mathbb{P}\left(Y \in \overline{A}\right) = \mathbb{P}\left(Y \in \mathring{A}\right) = \mathbb{P}(Y \in A)$, we get

$$\overline{\lim_n} \mathbb{P}(Y_n \in A) = \underline{\lim_n} \mathbb{P}(Y_n \in A) = \mathbb{P}(Y \in A),$$

which shows the result.

5. $\Rightarrow$ 1. Let $f : \mathbb{R}^k \to \mathbb{R}$ be a continuous and bounded mapping. Without loss of generality, we will assume that $0 < f < 1$ (otherwise one can always replace f by $\alpha f + \beta$ with $(\alpha, \beta) \in \mathbb{R}^2$ to get this property). We use the formula

$$\mathbb{E}[f(Y_n)] = \int_0^1 \mathbb{P}(f(Y_n) > t)\, dt = \int_0^1 \mathbb{P}\left(Y_n \in f^{-1}((t, \infty))\right) dt.$$

By continuity of f, we know that $f^{-1}((t, \infty))$ is an open set (as the reciprocal image of an open set by a continuous mapping) and $f^{-1}([t, \infty))$ is a closed set (as the reciprocal image of a closed set by a continuous mapping). Then

$$\overline{f^{-1}((t, \infty))} \subset f^{-1}([t, \infty)).$$

We deduce that

$$\partial f^{-1}((t, \infty)) \subset f^{-1}([t, \infty)) \setminus f^{-1}((t, \infty)) = \{f = t\}.$$

We know that $A = \{t \in \mathbb{R} : \mathbb{P}(f(Y) = t) > 0\}$ is finite or infinite but numerable. Indeed $A = \cup_{n \geq 1} A_n$ with $A_n = \{t \in \mathbb{R} : \mathbb{P}(f(Y) = t) \geq 1/n\}$ and A_n is necessarily finite (otherwise $\mathbb{P}$ cannot be a probability measure). We conclude that for all $t \notin A$, we have

$$\mathbb{P}\left(Y_n \in f^{-1}(t, \infty)\right) \to \mathbb{P}\left(Y \in f^{-1}(t, \infty)\right).$$

From the dominated convergence theorem, we conclude that $\lim_n \mathbb{E}[f(Y_n)] = \mathbb{E}[f(Y)]$.$\square$

End of the proof of the continuous mapping theorem. Here, we assume that $Y_n \hookrightarrow Y$. We use the point 3. of the portmanteau lemma. Let F be a closed set. We have

$$\{f(Y_n) \in F\} = \left\{Y_n \in f^{-1}(F)\right\} \subset \left\{Y_n \in \overline{f^{-1}(F)}\right\}.$$

Moreover, we have the inclusion $\overline{f^{-1}(F)} \subset f^{-1}(F) \cup \mathcal{C}^c$. Indeed, if $\lim_n y_n = y$ with $f(y_n) \in F$, either $y \in \mathcal{C}$ and then $\lim_n f(y_n) = f(y)$ is in F because F is closed or $y \notin \mathcal{C}$.

We then conclude that

$$\begin{aligned}
\overline{\lim_n} \mathbb{P}\left(f(Y_n) \in F\right) &\leq \overline{\lim_n} \mathbb{P}\left(Y_n \in \overline{f^{-1}(F)}\right) \\
&\leq \mathbb{P}\left(Y \in \overline{f^{-1}(F)}\right) \\
&\leq \mathbb{P}\left(Y \in f^{-1}(F)\right) + \mathbb{P}\left(Y \notin \mathcal{C}\right) = \mathbb{P}\left(f(Y) \in F\right).
\end{aligned}$$

The second inequality follows from an application of point 3. of the Portmanteau lemma to the sequence $(Y_n)_n$ (direct sense). We then conclude that $f(Y_n) \hookrightarrow f(Y)$ from point 3. of the Portmanteau lemma (reciprocal sense).□

1.3 Slutsky's lemma

Theorem 3. *Let c be vector of $\mathbb{R}^k$ and $(Y_n)_n$ and $(Z_n)_n$ be two sequences of random vectors taking values in $\mathbb{R}^k$.*

1. *We have $Y_n \xrightarrow{p} c$ if and only if $Y_n \hookrightarrow c$.*
2. *If $Y_n \hookrightarrow Y$ and $\|Y_n - Z_n\| \xrightarrow{p} 0$, then $Z_n \hookrightarrow Y$.*
3. *If $Y_n \hookrightarrow Y$ and $Z_n \xrightarrow{p} c$, then $(Y_n, Z_n) \hookrightarrow (Y, c)$.*

Note. Point 3. of the previous theorem is often called Slutsky's lemma. An example of application is the estimation of an unknown parameter in the expression of a weakly converging sequence. For instance, let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}(X_1) = m$ and $\mathrm{Var}(X_1) = \sigma^2 \in (0, \infty)$. Set $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. From the central limit theorem, we have $\frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - m\right) \hookrightarrow \mathcal{N}(0,1)$. Let $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$. We know that $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ (an even a.s. from the strong law of large numbers). From Slutsky's lemma and the continuous mapping theorem, we get $\frac{\sqrt{n}}{\hat{\sigma}}\left(\overline{X}_n - m\right) \hookrightarrow \mathcal{N}(0,1)$. One can then deduce a confidence interval of a given asymptotic level for the mean of the distribution when the variance is unknown.

Proof

1. Note first that the convergence in probability always entails the convergence in distribution. Suppose then that $Y_n \hookrightarrow c$. For any $\epsilon > 0$, we get from point 3. of the Portmanteau lemma,

$$\overline{\lim_n} \mathbb{P}\left(Y_n \notin B(c, \epsilon)\right) \leq \mathbb{P}\left(c \notin B(c, \epsilon)\right) = 0.$$

This shows that $Y_n \xrightarrow{p} c$.

2. Let $f : \mathbb{R}^k \to \mathbb{R}$ be a Lipschitz and bounded mapping, with Lipschitz constant $L > 0$. Set $\|f\|_\infty = \sup_{x \in \mathbb{R}^k} |f(x)|$. For any $\delta > 0$, we have

$$\begin{aligned} |\mathbb{E}[f(Y_n) - f(Z_n)]| &\leq 2\|f\|_\infty \mathbb{P}(\|Y_n - Z_n\| > \delta) + \sup_{\|y-z\| \leq \delta} |f(x) - f(y)| \\ &\leq 2\|f\|_\infty \mathbb{P}(\|Y_n - Z_n\| > \delta) + L\delta. \end{aligned}$$

We get $\overline{\lim}_n |\mathbb{E}[f(Y_n) - f(Z_n)]| \leq L\delta$. Since $\delta > 0$ can be arbitrarily small, the result follows from the point 2. of the Portmanteau lemma.

3. We use the previous point since $(Y_n, c) \hookrightarrow (Y, c)$ (it can be checked using the general definition of convergence in distribution) and $\|(Y_n, Z_n) - (Y_n, c)\| \xrightarrow{p} 0$.

1.4 Stochastic o and O

Definition 1. *The sequence $(Y_n)_n$ is said to be bounded in probability if for any $\epsilon > 0$, there exists $M > 0$ such that $\sup_{n \in \mathbb{N}} \mathbb{P}(\|Y_n\| > M) \leq \epsilon$. We then note $Y_n = O_{\mathbb{P}}(1)$.*

Notes

1. If there exists a constant $L > 0$ such that $\|Y_n\| \leq L$ a.s. (bounded sequence), then $Y_n = O_{\mathbb{P}}(1)$.
2. If $Y_n = Y$ for any n, then $Y_n = O_{\mathbb{P}}(1)$.
3. $Y_n = O_{\mathbb{P}}(1)$ if and only if for all $\epsilon > 0$, there exists $M > 0$ such that $\overline{\lim}_n \mathbb{P}(\|Y_n\| > M) \leq \epsilon$.

The following result is sometimes used to show convergence in distribution.

Theorem 4 (Prokorov). *1. If $Y_n \hookrightarrow Y$, then $Y_n = O_{\mathbb{P}}(1)$.*

2. If $Y_n = O_{\mathbb{P}}(1)$ then there exists a subsequence $\left(Y_{n_j}\right)_j$ converging in distribution.

Proof

1. Let $\epsilon > 0$ and $M > 0$ such that $\mathbb{P}(\|Y\| \geq M) \leq \epsilon$ (note that $\lim_{M \to \infty} \mathbb{P}(\|Y\| \geq M) = 0$). Using the point 3. of the Portmanteau lemma, we have $\overline{\lim}_n \mathbb{P}(\|Y_n\| \geq M) \leq \mathbb{P}(\|Y\| \geq M) \leq \epsilon$.
2. The second point is much more difficult to get. See for instance Van der Vaart (2000), pp. $8 - 9$ or Billingsley (2013), Theorem 5.1.

Notation. The convergence in probability $Y_n \xrightarrow{p} 0$ is also denoted by $Y_n = o_{\mathbb{P}}(1)$. Of course $Y_n = o_{\mathbb{P}}(1) \Rightarrow Y_n = O_{\mathbb{P}}(1)$.

Rules of calculus. One can show that the following rules are valid. $o_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$, $O_{\mathbb{P}}(1) + O_{\mathbb{P}}(1) = O_{\mathbb{P}}(1)$, $o_{\mathbb{P}}(1)O_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$, $(1 + o_{\mathbb{P}}(1))^{-1} = O_{\mathbb{P}}(1)$.

Comparison of random sequences. For a sequence $(R_n)_n$ of real-valued random variables, we say that $Y_n = o_{\mathbb{P}}(R_n)$ if $Y_n = R_n Z_n$ with $Z_n = o_{\mathbb{P}}(1)$ and $Y_n = O_{\mathbb{P}}(R_n)$ if $Y_n = R_n Z_n$ with $Z_n = O_{\mathbb{P}}(1)$. We obtain the following new rules of calculus $o_{\mathbb{P}}(R_n) = R_n o_{\mathbb{P}}(1)$, $O_{\mathbb{P}}(R_n) = R_n O_{\mathbb{P}}(1)$ and $o_{\mathbb{P}}\left(O_{\mathbb{P}}(1)\right) = o_{\mathbb{P}}(1)$.

Lemma 2. *Let $R : \mathbb{R}^k \to \mathbb{R}$ be a measurable mapping such that $R(0) = 0$ and $Y_n = o_{\mathbb{P}}(1)$ taking values in $\mathbb{R}^k$. For any $p > 0$,*

1. *if $R(h) = o\left(\|h\|^p\right)$ when $h \to 0$ then $R(Y_n) = o_{\mathbb{P}}\left(\|Y_n\|^p\right)$,*
2. *if $R(h) = O\left(\|h\|^p\right)$ when $h \to 0$ then $R(Y_n) = O_{\mathbb{P}}\left(\|Y_n\|^p\right)$.*

Proof. Let $g : \mathbb{R}^k \to \mathbb{R}$ be the mapping defined by $g(0) = 0$ and $g(h) = R(h)/\|h\|^p$ if $h \neq 0$. Note that under the assumption of point 1., g is continuous at 0. The equality $R(Y_n) = \|Y_n\|^p g(Y_n)$ is valid in both cases.

1. The continuous mapping theorem ensures that $g(Y_n) = o_{\mathbb{P}}(1)$ if $Y_n = o_{\mathbb{P}}(1)$.
2. In the second case, we use the bound

$$\mathbb{P}\left(|g(Y_n)| > M\right) \leq \mathbb{P}\left(\|Y_n\| > K\right) + \mathbb{P}\left(\|Y_n\| \leq K, |g(Y_n)| > M\right).$$

   By the assumption on R, one can take K small enough and M large enough such that the second probability is equal to 0. Then $\overline{\lim}_n \mathbb{P}\left(|g(Y_n)| > M\right) \leq \overline{\lim}_n \mathbb{P}\left(\|Y_n\| > K\right) = 0$ which shows that $g(Y_n) = O_{\mathbb{P}}(1)$.□

1.5 $\delta-$method

Let O be an open subset of $\mathbb{R}^k$.

Theorem 5. *Let $\phi : O \to \mathbb{R}^m$ be a differentiable mapping at point $\theta \in O$ and $T_n : \Omega \to O$ a random vector such that $r_n(T_n - \theta) \hookrightarrow T$ with $r_n \to \infty$. Then*

$$r_n\left(\phi(T_n) - \phi(\theta)\right) \hookrightarrow J_\phi(\theta) \cdot T$$

where $J_\phi(\theta) = \left(\frac{\partial \phi_i}{\partial x_j}(\theta)\right)_{\substack{1 \leq i \leq m \\ 1 \leq j \leq k}}$.

Proof. Using a Taylor expansion at order 1, we have

$$\phi(\theta+h)=\phi(\theta)+J_\phi(\theta)h+o\left(\|h\|\right).$$

Since $r_n(T_n-\theta)=O_{\mathbb{P}}(1)$, we get $T_n-\theta=\frac{1}{r_n}O_{\mathbb{P}}(1)=o_{\mathbb{P}}(1)$ and $T_n\xrightarrow{p}\theta$. From the previous lemma, we get $\phi(T_n)=\phi(\theta)+J_\phi(\theta)\cdot(T_n-\theta)+o_{\mathbb{P}}\left(\|T_n-\theta\|\right)$. We then get

$$r_n\left(\phi(T_n)-\phi(\theta)\right)=J_\phi(\theta)r_n(T_n-\theta)+r_n\|T_n-\theta\|o_{\mathbb{P}}(1).$$

Since $r_n\|T_n-\theta\|=O_{\mathbb{P}}(1)$ (from our assumptions), we obtain

$$r_n\left(\phi(T_n)-\phi(\theta)\right)=J_\phi(\theta)r_n(T_n-\theta)+o_{\mathbb{P}}(1)$$

and the result follows from the continuous mapping theorem and Slutsky's lemma.□

Example. Consider

$$S^2=\frac{1}{n}\sum_{i=1}^n\left(X_i-\overline{X}_n\right)^2=\frac{1}{n}\sum_{i=1}^n X_i^2-\overline{X}_n^2=\overline{X^2}_n-\overline{X}_n^2.$$

Set $\mu_i=\mathbb{E}X_1^i$ for any positive integer i. Suppose that $\mu_1=0$, set $\sigma^2=\mu_2-\mu_1^2=\mu_2$ and $\phi(x,y)=y-x^2$. Then $J_\phi(x,y)=(-2x,1)$. Setting $T_n=\left(\overline{X}_n,\overline{X^2}_n\right)$, we have

$$\sqrt{n}\left(T_n-(0,\mu_2)\right)\hookrightarrow\mathcal{N}_2\left((0,0),\begin{pmatrix}\mu_2 & \mu_3\\ \mu_3 & \mu_4-\mu_2^2\end{pmatrix}\right).$$

We deduce that

$$\sqrt{n}\left(S^2-\sigma^2\right)\hookrightarrow\mathcal{N}\left(0,\mu_4-\mu_2^2\right).$$

If $\mu_1\neq 0$, one can replace μ_i by $\mathbb{E}\left[(X_1-\mu_1)^i\right]$ to get a similar result.

Note. If the first derivative of ϕ vanishes, it is still possible to study a higher-order Taylor expansion to get the asymptotic distribution of $\phi(T_n)$. For instance, using a Taylor expansion of order 2, one can show that if $\mathbb{E}(X_1)=0$ and $\mathbb{E}\left(X_1^2\right)=1$,

$$-2n\left(\cos\left(\overline{X}_n\right)-1\right)\hookrightarrow\chi^2(1).$$

Chapter 2

Examples of convergence of estimators

2.1 M-estimators

An $M-$estimator $\hat{\theta}_n$ is a minimizer of a random function $\theta \mapsto M_n(\theta)$ that can be computed using realizations of n random variables or random vectors $X_1, \ldots, X_n$. More precisely,

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} M_n(\theta),$$

where Θ is the set of possible parameters. In this paragraph, we will only consider finite-dimensional parameter spaces, i.e. $\Theta \subset \mathbb{R}^d$ for some $d \geq 1$. Extension to more general metric spaces is possible. For simplicity, we will always assume that an infimum or a minimizer of a random function is measurable. A more thorough discussion of measurability problems has been given in the introduction chapter.

In this section, we denote by $\|\cdot\|$ an arbitrary norm on $\mathbb{R}^d$ of $\mathbb{R}^k$. $B(x, \epsilon)$ will denote the corresponding open ball of center x and radius ϵ, i.e. $B(x, \epsilon) = \big\{y \in \mathbb{R}^d : \|y - x\| < \epsilon\big\}$.

2.1.1 Consistency of $M-$estimators

The first result is very simple to state and already ensures weak consistency of a sequence of M-estimators (i.e. convergence in probability to the minimizer of a limit criterion).

Theorem 6. *Assume that there exists a non random mapping $M : \Theta \to \mathbb{R}$ such that*

1. $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| = o_{\mathbb{P}}(1)$.
2. *For all $\epsilon > 0$, $\sup_{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon} M(\theta) > M(\theta_0)$.*

Then $\hat{\theta}_n \xrightarrow{p} \theta_0$ (weak consistency).

Note. The first assumption of this theorem is an assumption of uniform convergence which is often used for studying consistency of M-estimators. The second assumption is an assumption of "good" separation of $M(\theta_0) = \inf_{\theta \in \Theta} M(\theta)$.

Proof. Let $\epsilon > 0$. From our second assumption, there exists $\eta > 0$ such that $\|\theta - \theta_0\| > \epsilon \Rightarrow M(\theta) - M(\theta_0) > \eta$. Moreover we have $M_n\left(\hat{\theta}_n\right) \leq M_n(\theta_0)$ and

$$M\left(\hat{\theta}_n\right) - M(\theta_0) = M\left(\hat{\theta}_n\right) - M_n\left(\hat{\theta}_n\right) + M_n\left(\hat{\theta}_n\right) - M_n(\theta_0) + M_n(\theta_0) - M(\theta_0) \leq 2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)|.$$

We then get

$$\mathbb{P}\left(\|\hat{\theta}_n - \theta_0\| > \epsilon\right) \leq \mathbb{P}\left(M\left(\hat{\theta}_n\right) \geq M(\theta_0) + \eta\right) \leq \mathbb{P}\left(2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| > \eta\right).$$

We conclude using the first assumption.□

Note. If the first assumption of the previous theorem is replaced by $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \overset{a.s.}{\to} 0$, then we have strong consistency, i.e. $\hat{\theta}_n \overset{a.s.}{\to} \theta_0$. This a consequence of the inclusion

$$\left\{\|\hat{\theta}_n - \theta_0\| > \epsilon\right\} \subset \left\{2 \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| > \eta\right\}.$$

Details are omitted.

We now are interested in checking the uniform convergence property in the special case where $M_n(\theta) = \frac{1}{n} \sum_{i=1}^n m_\theta(X_i)$ when Θ is compact. The following result is an example of uniform law of large numbers.

Lemma 3. *Suppose that Θ is compact, $X_1, \ldots, X_n$ i.i.d., $\theta \mapsto m_\theta(x)$ continuous for $\mathbb{P}_{X_1}-$almost all x and $\mathbb{E}\left[\sup_{\theta \in \Theta} |m_\theta(X_1)|\right] < \infty$. Then $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \overset{a.s.}{\to} 0$.*

Proof. For $\delta > 0$ and $x \in \mathbb{R}^k$, set

$$w_\Delta(x) = \sup\left\{|m_{\theta+h}(x) - m_\theta(x)| : \theta, \theta + h \in \Theta, \quad \|h\| \leq \Delta\right\}.$$

$w_\Delta(x)$ is the modulus of continuity of $\theta \mapsto m_\theta(x)$ at point x. Using the dominated convergence, we have $\lim_{\Delta \to 0} \mathbb{E}\left[w_\Delta(x)\right] = 0$. Since Θ is compact, one can find $\theta_1, \ldots, \theta_\ell \in \Theta$ such that $\Theta \subset \cup_{i=1}^\ell B\left(\theta_i, \Delta\right)$. Let $\theta \in B\left(\theta_i, \Delta\right)$ for some $i = 1, \ldots, \ell$. Then

$$|M_n(\theta) - M_n(\theta_i)| \leq \frac{1}{n} \sum_{i=1}^n w_\Delta(X_i) \overset{\Delta \to 0}{\to} \mathbb{E}\left[w_\Delta(X_1)\right] \text{ a.s.}$$

using the law of large numbers. We get

$$\begin{aligned} |M_n(\theta) - M(\theta)| &\leq |M_n(\theta) - M_n(\theta_i)| + |M_n(\theta_i) - M(\theta_i)| + |M(\theta_i) - M(\theta_i)| \\ &\leq \frac{1}{n} \sum_{i=1}^n w_\Delta(X_i) + \max_{1 \leq i \leq \ell} |M_n(\theta_i) - M(\theta_i)| + \mathbb{E}\left[w_\Delta(X_1)\right]. \end{aligned}$$

Using the law of large numbers (note that the maximum in the right-hand side only depends on a finite number of points), we then get

$$\overline{\lim_n} |M_n(\theta) - M(\theta)| \leq 2\mathbb{E}\left[w_\Delta(X_1)\right] \text{ a.s.}$$

We conclude by letting $\Delta \to 0$.□

Note. In the compact case, strong consistency can be reduced to a deterministic problem: the convergence of a sequence of minimizers when we have uniform convergence of the objective functions. Let $f_n : \Theta \to \mathbb{R}$ be a continuous mapping defined on a compact set Θ and such that $\|f_n - f\|_\Theta := \sup_{\theta\in\Theta} |f_n(\theta) - f(\theta)| \overset{n\to\infty}{\to} 0$. Suppose that $f(\theta_0) < f(\theta)$ for all $\theta \neq \theta_0$. Then the uniform convergence ensures that f is continuous and the compactness assumption guarantees that $\theta_n = \arg\min_\theta f_n(\theta)$, $n \geq 1$, has a subsequence $(\theta_{\phi_n})_{n\geq 1}$ converging to some point $\theta^* \in \Theta$. Moreover,

$$|f_{\phi_n}(\theta_{\phi_n}) - f(\theta^*)| \leq \|f_{\phi_n} - f\|_\Theta + |f(\theta_{\phi_n}) - f(\theta^*)| \overset{n\to\infty}{\to} 0.$$

Since $f_{\phi_n}(\theta_{\phi_n}) \leq f_{\phi_n}(\theta_0)$, by letting $n \to \infty$, $f(\theta^*) \leq f(\theta_0)$. We deduce that $\theta^* = \theta_0$. One can deduce that $\lim_{n\to\infty} \theta_n = \theta_0$ (a sequence in a compact set converges if and only if it has a unique cluster point).

The previous discussion leads to the following result.

Theorem 7. *Let $X_1, \ldots, X_n$ be i.i.d. random vectors. Suppose that*

1. *Θ is compact,*
2. *for $\mathbb{P}_{X_1}-$almost all x, the mapping $\theta \mapsto m_\theta(x)$ is continuous on Θ and $\mathbb{E}[\sup_{\theta\in\Theta} |m_\theta(X_1)|] < \infty$,*
3. *$\theta \mapsto M(\theta)$ is minimized only in θ_0.*

Then $\hat{\theta}_n \overset{a.s.}{\to} \theta_0$.

Note. The second assumption ensures the continuity of the mapping $\theta \mapsto M(\theta)$ (using the theorem of continuity under the sign integral).

2.1.2 Application to maximum likelihood estimators (MLE)

When $\mathbb{P}_{X_1} = p_{\theta_0} \cdot \mu$, the maximum likelihood estimator can be seen as a M-estimator corresponding to $m_\theta(x) = -\log \frac{p_\theta(x)}{p_{\theta_0}(x)}$. Note that the division by p_{θ_0} is considered only for theoretical reasons, in practice $\theta \mapsto -\log p_\theta(x)$ is used. The aim of this part is to give a sufficient (and necessary) for the third assumption needed for applying Theorem 7.

We first introduce an important quantity for measuring the closeness of two probability measures $P = p \cdot \mu$ and $Q = q \cdot \mu$. The Kullback-Leibler divergence is between P and Q is defined by

$$K(P, Q) = \int_{p>0} p \log(p/q) d\mu \text{ if } \mu(\{q = 0, p > 0\}) = 0,$$

otherwise we set $K(P, Q) = \infty$.

Lemma 4. *We have $K(P, Q) \geq d_H(P, Q)^2 := \int \left(\sqrt{p} - \sqrt{q}\right)^2 d\mu$.*

Proof. We assume that $\mu(\{q=0, p>0\}) = 0$ otherwise $K(P,Q) = \infty$ and the result is obvious. We first note that

$$\int_{p>0} p\,(\log(p/q))_- \, d\mu = \int_{p>0} q\left(\frac{p}{q}\log\frac{p}{q}\right)_- d\mu < \infty.$$

Indeed the negative part of $x \mapsto x\log(x)$ is bounded. If $\int_{p>0} p\left(\log\frac{p}{q}\right)_+ d\mu = \infty$, one can deduce that $K(P,Q) \geq d_H(P,Q)^2$. Otherwise, using the inequality $\log(x) \leq 2\left(\sqrt{x}-1\right)$ for all $x \geq 0$, we get

$$\begin{aligned}
\int_{p>0} p\log\frac{p}{q}d\mu &= -\int_{p>0} p\log\frac{q}{p}d\mu \\
&\geq -2\int_{p>0} p\left(\sqrt{q/p}-1\right)d\mu \\
&= -2\int \sqrt{pq}d\mu + 2 \\
&= d_H(P,Q)^2.\square
\end{aligned}$$

Notes. From the previous lemma, one can notice that $K \geq 0$ and $K(P,Q) = 0$ if and only if $p = q$. However K is not symmetric and does not satisfy the triangular inequality. K is then not a distance. However, d_H is a distance called Hellinger distance. From the previous lemma, a small divergence entails proximity between the two probability measures. We defer the reader to Tsybakov (2004) for some additional properties of the Kullback-Leibler divergence as well as some comparisons with other metrics between probability measures.

Proposition 2. *Suppose that for any $\theta \in \Theta$, $\mu(\{p_\theta = 0\}) = 0$ and for any $\theta \neq \theta_0$, $\mu(\{p_\theta \neq p_{\theta_0}\}) > 0$. Then M has a unique minimizer at point θ_0.*

Proof. Observe that $M(\theta) = \mathbb{E}[m_\theta(X_1)] = K(P_{\theta_0}, P_\theta)$ and from the previous lemma, we have $M(\theta) \geq 0 = M(\theta_0)$ and $M(\theta) = 0 = M(\theta_0)$ implies $d_H(P_\theta, P_{\theta_0}) = 0$ which in turn implies that $p_\theta = p_{\theta_0}$ $\mu-$almost everywhere.$\square$

Note. Proposition 2 with Theorem 7 can be used to prove consistency of MLE when the state space is compact. Note that the assumptions of Proposition 2, which guaranty identification of the parameter, are quite weak. Additionally to the existence of a common support for all the densities, it is simply necessary to assume that two different parameters do not lead to the same probability density (up to some set of null measure for μ).

2.1.3 A more general result in the $\sigma-$compact case

It is possible to relax the compactness assumption which is crucial in Theorem 7.

Theorem 8. *Suppose that $\Theta = \cup_{k\geq 0}\mathring{\Theta}_k$ where Θ_k is a compact subset of $\mathbb{R}^d$ and $\Theta_k \subset \Theta_{k+1}$. Suppose furthermore that the following conditions are satisfied.*

1. *The mappings M_n and M are a.s. continuous on Θ.*
2. *θ_0 is the unique minimizer of M.*
3. *For any $k \in \mathbb{N}$, $\lim_{n\to\infty} \sup_{\theta\in\Theta_k} |M_n(\theta) - M(\theta)| = 0$ a.s.*
4. *There exists $\hat{\theta}_n = \arg\min_{\theta\in\Theta} M_n(\theta)$ such that for all $\omega \in \Omega$, there exists a compact $K = K(\omega)$ such that $\hat{\theta}_n(\omega) \in K$ for all $n \geq 1$.*

Then $\hat{\theta}_n \overset{a.s.}{\to} \theta_0$.

Proof. We start by noticing that any compact subset K of Θ is automatically included in a compact set Θ_k for some integer k. If not, one can always find a sequence $(x_k)_k$ in K such that $x_k \notin \Theta_k$. But there then exists a cluster point x of the sequence in K and since $x \in \mathring{\Theta}_\ell$ for some integer ℓ, one can conclude that $x_{k'} \in \mathring{\Theta}_\ell \subset \Theta_{k'}$ for large k' which is a contradiction.

Now, let $\omega \in \Omega$ and $K(\omega)$ compact such that $\hat{\theta}_n(\omega) \in K(\omega)$ for all $n \geq 1$. Since $K(\omega) \subset \Theta_{k(\omega)}$ for some integer $k(\omega)$ and $\theta_0 \in \Theta_j$ for another integer j, we use the uniform convergence of $\theta \mapsto M_n(\theta)_\omega$ to M on the set $\Theta_{\max(j,k(\omega))}$ as well as the deterministic argument presented before the statement of Theorem 7 to conclude.□

The example of geometric median Let $\Theta = \mathbb{R}^d$ for $d \geq 2$ and $X_1, \dots, X_n$ i.i.d. and taking values in Θ. Suppose that $\mathbb{E}\|X_1\| < \infty$ where $\|\cdot\|$ denotes the Euclidean norm. Let us define

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} M_n(\theta), \quad M_n(\theta) = \frac{1}{n}\sum_{i=1}^n \|X_i - \theta\|.$$

It is easy to check the inequalities $|M_n(\theta) - M_n(\theta')| \leq \|\theta - \theta'\|$ and $|M(\theta) - M(\theta')| \leq \|\theta - \theta'\|$. Moreover, the uniform convergence of M_n is valid on any compact subset of Θ.

Next one can note that an argmin $\hat{\theta}_n$ always exists because $\lim_{\|\theta\|\to\infty} M_n(\theta) = \infty$. For a given $\theta_* \in \Theta$, any argmin satisfies the inequalities

$$\|\hat{\theta}_n - \theta_0\| \leq \frac{1}{n}\sum_{i=1}^n \|X_i - \hat{\theta}_n\| + \frac{1}{n}\sum_{i=1}^n \|X_i - \theta_*\| \leq \frac{2}{n}\sum_{i=1}^n \|X_i - \theta_*\| \to 2\mathbb{E}\|X_1 - \theta_*\| < \infty.$$

This shows the condition 4. of Theorem 8 and conditions 1. and 3. are also satisfied.

It remains to check condition 2. A median, i.e. $\theta_0 = \arg\min_{\theta\in\Theta} M(\theta)$, always exists. This is a consequence of the continuity of M and

$$\arg\min_{\theta\in\Theta} M(\theta) = \arg\min_{\theta:\|\theta\|\leq 2\mathbb{E}\|X_1\|} M(\theta).$$

To show uniqueness of θ_0, we will assume that the support of the measure $\mathbb{P}_{X_1}$ is not included in a line $\mathcal{D}$, i.e. $\mathbb{P}(X_1 \in \Theta \setminus \mathcal{D}) > 0$. If θ_1 and θ_2 are two medians, let $\mathcal{D}$ be the line joining

these two distinct points. Let $\lambda \in (0,1)$. If $x \notin \mathcal{D}$, we have $\|x-(1-\lambda)\theta_1-\lambda\theta_2\| < (1-\lambda)\|x-\theta_1\|+\lambda\|x-\theta_2\|$ and the inequality is large if $x \in \mathcal{D}$. We conclude that

$$M\left((1-\lambda)\theta_1+\lambda\theta_2\right) < (1-\lambda)M(\theta_1)+\lambda M(\theta_2) = M(\theta_1),$$

which contradicts the definition of θ_1. Then condition 2. of Theorem 8 is also satisfied.

2.1.4 Z-estimators

We call $\hat{\theta}_n$ a Z-estimator if $Z_n\left(\hat{\theta}_n\right) = 0$ or more generally $Z_n\left(\hat{\theta}_n\right) = o_{\mathbb{P}}(1)$, where

$$Z_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} z_\theta(X_i), \quad \theta \in \Theta.$$

This estimator is meaningful when the target $\theta_0 \in \Theta$ satisfies $\mathbb{E}\left[z_{\theta_0}(X_1)\right] = 0$.

Examples

1. When $z_\theta(x) = \dot{m}_\theta(x)$ (notation for the gradient of $\theta \mapsto m_\theta(x)$), a Z-estimator is an example of M-estimator since we simply want to vanish the gradient of the objective function M_n for finding $\hat{\theta}_n = \arg\min_{\theta\in\Theta} M_n(\theta)$.

2. It can happen that Z_n is not the derivative of a differentiable mapping. For instance, if $z_\theta = \text{ sign } (x-\theta)$ where $\text{sign}(u) = \mathbb{1}_{u>0} - \mathbb{1}_{u<0}$, $\hat{\theta}_n$ is called the median. Alternatively, a median can be defined from the M-estimator such that $m_\theta(x) = |x-\theta|$. Both estimators enjoy similar properties.

3. We next give a Z-estimator based on the idea of instrumental variable in Econometrics. Suppose that
$$Y_i = \theta_{0,1} + \theta_{0,2}X_i + \theta_{0,3}Z_i + \varepsilon_i, \quad 1 \le i \le n,$$
where ε_i is independent from (X_i, Z_i) and $\mathbb{E}(\varepsilon_i) = \mathbb{E}(Z_i) = 0$. But only (X_i, Y_i) is observed. For instance, Y_i can represent the income of an individual, X_i the number of years of education and Z_i the qualities of the individual. If X_i and Z_i are not independent, $\mathbb{E}(Y_i|X_i) \neq \theta_{0,1} + \theta_{0,2}X_i$ and the least-squares method does not apply. The idea is to find an "instrument" uncorrelated with (Z_i, ε_i), for instance the salary of the parents. We then get the two following equalities
$$\mathbb{E}(Y_i) = \theta_{0,1} + \theta_{0,2}\mathbb{E}(X_i), \quad \mathbb{E}(Y_iW_i) = \theta_{0,1}\mathbb{E}(W_i) + \theta_{0,2}\mathbb{E}(X_iW_i).$$
We then set
$$z_\theta(y,x,w) = \left(y-\theta_1-\theta_2 x, y-\theta_1 w-\theta_2 xw\right).$$
If the covariance between W_1 and X_1 is different from 0, the determinant of the matrix $\begin{pmatrix} 1 & \mathbb{E}(X_1) \\ \mathbb{E}(W_1) & \mathbb{E}(W_1X_1) \end{pmatrix}$ is different form 0 and one can identify θ_0.

We next give a result analogue to Theorem 6 for Z-estimators. The proof is similar and then omitted. Theorem 7 can be also stated for Z-estimators.

Theorem 9. *Suppose that*

1. $\sup_{\theta\in\Theta} \|Z_n(\theta) - Z(\theta)\| = o_{\mathbb{P}}(1)$,
2. *For all* $\epsilon > 0$, $\inf_{\theta\in\Theta:\|\theta-\theta_0\|>\epsilon} \|Z(\theta)\| > \|Z(\theta_0)\| = 0$.

Then any sequence of Z-estimators is weakly consistent.

2.1.5 Asymptotic normality for M-estimation

In this subsection, we consider $M_n(\theta) = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$. For $x \in \mathbb{R}^k$, the gradient vector and the Hessian matrix of the mapping $\theta \mapsto m_\theta(x)$ at point θ will be denoted by $\dot{m}_\theta(x)$ and $\ddot{m}_\theta(x)$ respectively.

Theorem 10. *We suppose that the following assumptions hold true.*

1. Θ *is a compact subset of* $\mathbb{R}^d$ *and* $\theta_0 \in \mathring{\Theta}$.
2. *The point* θ_0 *is the unique minimizer of the mapping* $\theta \mapsto \mathbb{E}[m_\theta(X_1)]$ *and* $\sup_{\theta\in\Theta} |m_\theta(X_1)|$ *is integrable.*
3. *For all* x, $\theta \mapsto m_\theta(x)$ *is two times continuously differentiable and there exists a neighborhood* $\mathcal{V}_{\theta_0}$ *of* θ_0 *such that* $\sup_{\theta\in\Theta} \|\ddot{m}_{\theta_0}(X_1)\|$*is integrable.*
4. $\dot{m}_{\theta_0}(X_1)$ *is square integrable and* $W_{\theta_0} := \mathbb{E}[\ddot{m}_{\theta_0}(X_1)]$ *is invertible.*

Then

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \hookrightarrow \mathcal{N}_d\left(0, W_{\theta_0}^{-1} V_{\theta_0} W_{\theta_0}^{-1}\right),$$

where $V_{\theta_0} = \mathbb{E}\left[\dot{m}_{\theta_0}(X_1)\dot{m}_{\theta_0}(X_1)^T\right]$.

Note. Observe that the assumptions of Theorem 10 guaranty that $\mathbb{E}[\dot{m}_{\theta_0}(X_1)]$ is the gradient at point θ_0 of the mapping $\theta \mapsto \mathbb{E}[m_\theta(X_1)]$ and then $V_{\theta_0} = \mathbb{E}\left[\dot{m}_{\theta_0}(X_1)\dot{m}_{\theta_0}(X_1)^T\right]$.

Proof of Theorem 10 The assumptions of Theorem 10 contain that of Theorem 8. Then $\hat{\theta}_n \overset{a.s.}{\to} \theta_0$.

The idea is to make a Taylor expansion of the following form

$$\begin{aligned} 0 \approx \dot{M}_n\left(\hat{\theta}_n\right) &= \dot{M}_n(\theta_0) + \ddot{M}_n(\theta_0) \cdot \left(\hat{\theta}_n - \theta_0\right) + o_{\mathbb{P}}\left(1/\sqrt{n}\right) \\ &= \dot{M}_n(\theta_0) + W_{\theta_0} \cdot \left(\hat{\theta}_n - \theta_0\right) + o_{\mathbb{P}}\left(1/\sqrt{n}\right). \end{aligned}$$

From the central limit theorem, we have $\sqrt{n}\dot{M}_n(\theta_0) \hookrightarrow \mathcal{N}_d\left(0, V_{\theta_0}\right)$. We will then deduce that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = -W_{\theta_0}^{-1}\sqrt{n}\dot{M}_n(\theta_0) + o_{\mathbb{P}}(1) \tag{2.1}$$

and the result will follow from Slutsky's lemma. To prove (2.1), we start by noticing that for $\mathbb{P}-$almost all ω, there exists $n_0(\omega)$ such that for $n \geq n_0(\omega)$, $\dot{M}_n\left(\hat{\theta}_n(\omega)\right)_\omega = 0$ because $\lim_{n\to\infty}\hat{\theta}_n(\omega) = \theta_0$ and $\hat{\theta}_n(\omega) \in \mathring{\Theta}$ if n is large enough.

We deduce that $\dot{M}_n\left(\hat{\theta}_n\right) = o_{\mathbb{P}}\left(1/\sqrt{n}\right)$ (and even $o_{\mathbb{P}}(r_n)$ for any $r_n \to 0$).

The Taylor-Lagrange formula at order 1 gives for $1 \leq i \leq d$,

$$\frac{\partial M_n}{\partial\theta_i}\left(\hat{\theta}_n\right) = \frac{\partial M_n}{\partial\theta_i}(\theta_0) + \sum_{j=1}^{d}\frac{\partial^2 M_n}{\partial\theta_i\partial\theta_j}\left(\widetilde{\theta}_n^{(i)}\right)\cdot\left(\hat{\theta}_{n,j} - \theta_{0,j}\right)$$

for some $\widetilde{\theta}_n^{(i)} \in [\theta_0, \hat{\theta}_n]$. Using vectors, we conclude that

$$\dot{M}_n\left(\hat{\theta}_n\right) = \dot{M}_n(\theta_0) + S_n\left(\hat{\theta}_n - \theta_0\right), \quad S_n = \left(\frac{\partial^2 M_n}{\partial\theta_i\partial\theta_j}\left(\widetilde{\theta}_n^{(i)}\right)\right)_{1\leq i,j\leq d}. \tag{2.2}$$

From the third assumption, which guarantees a uniform law of large numbers on $\mathcal{V}_{\theta_0}$ and the almost sure convergence of $\hat{\theta}_n$, one can easily show that $S_n = W_{\theta_0} + o_{\mathbb{P}}(1)$. Now (2.1) can be obtained from (2.2) and the invertibility of W_{θ_0} if we show that

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) = O_{\mathbb{P}}(1). \tag{2.3}$$

To show (2.3), we start by noticing that there exist $\eta, c > 0$ such that if H is a matrix of size $d \times d$ such that $\|H\| \leq \eta$ then $\left\|\left(W_{\theta_0} + H\right)^{-1}\right\| \leq c$. We then get

$$\begin{aligned}\mathbb{P}\left(\sqrt{n}\|\hat{\theta}_n - \theta_0\| > M\right) &\leq \mathbb{P}\left(\|S_n - W_{\theta_0}\| > \eta\right) + \mathbb{P}\left(\|S_n - W_{\theta_0}\| \leq \eta, \sqrt{n}\|\hat{\theta}_n - \theta_0\| > M\right)\\ &= p_{1,n} + p_{2,n}.\end{aligned}$$

We already know that $\lim_{n\to\infty} p_{1,n} = 0$. Moreover,

$$p_{2,n} \leq \mathbb{P}\left(\|S_n - W_{\theta_0}\| \leq \eta, \sqrt{n}\|S_n\left(\hat{\theta}_n - \theta_0\right)\| > M/c\right) \leq \mathbb{P}\left(\sqrt{n}\|\dot{M}_n(\theta_0) + o_{\mathbb{P}}(1/\sqrt{n})\| > M/c\right).$$

But $\sqrt{n}\|\dot{M}_n(\theta_0) + o_{\mathbb{P}}(1/\sqrt{n})\| = O_{\mathbb{P}}(1)$ and then $\sup_{n\geq 1} p_{2,n}$ can be made arbitrarily small if M is large enough. This shows (2.3) and completes the proof.□

We mention without proof another result for asymptotic normality that does require $\theta \mapsto m_\theta(x)$ to be differentiable in a neighborhood of θ_0 but transfers this smoothness to its expectation M. Contrarily to the previous one, this result can be applied to the median, i.e. $m_\theta(x) = |x - \theta|$. See https://perso.univ-rennes1.fr/bernard.delyon/param.pdf, Theorem 15, p. 45.

Theorem 11. *Suppose that the following assumptions hold true.*

1. Θ *is a compact of* $\mathbb{R}^d$ *and* $\theta_0 \in \mathring{\Theta}$.
2. *There exists a measurable mapping* $N : \mathbb{R}^k \to \mathbb{R}_+$ *such that* $|m_\theta(x) - m_{\theta'}(x)| \leq N(x)\|\theta - \theta'\|$ *with* $\mathbb{E}\left[N(X_1)^2\right] < \infty$.
3. $\theta \mapsto m_\theta(X_1)$ *is a.s. differentiable at point* θ_0 *and* $\dot{m}_{\theta_0}(X_1)$ *is square integrable.*
4. *The mapping* M *is two times continuously differentiable with* θ_0 *as unique minimizer.*

Then,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \hookrightarrow \mathcal{N}_d\left(0, \ddot{M}(\theta_0)^{-1}\mathbb{E}\left[\dot{m}_{\theta_0}(X_1)\dot{m}_{\theta_0}(X_1)^T\right]\ddot{M}(\theta_0)^{-1}\right).$$

2.1.6 Maximum likelihood estimators

Here $m_\theta(x) = -\log p_\theta(x)$. Under the assumptions of Theorem 10 or Theorem 11, we have the weak convergence

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \hookrightarrow \mathcal{N}_d\left(0, \ddot{M}(\theta_0)^{-1}\mathbb{E}\left[\dot{m}_{\theta_0}(X_1)\dot{m}_{\theta_0}(X_1)^T\right]\ddot{M}(\theta_0)^{-1}\right).$$

The quantity

$$I\left(\theta_0\right) =: \mathbb{E}\left[\dot{m}_{\theta_0}(X_1)\dot{m}_{\theta_0}(X_1)^T\right] = \mathbb{E}\left[\frac{\dot{p}_{\theta_0}(X_1)\dot{p}_{\theta_0}(X_1)^T}{p_{\theta_0}(X_1)^2}\right]$$

is called Fisher information (at point θ_0).

Lemma 5. *Suppose that* $\dot{p}_{\theta_0}(X_1)/p_{\theta_0}(X_1)$ *is square integrable. If there exists a neighborhood* $\mathcal{V}_{\theta_0}$ *of* θ_0 *such that the mapping* $\theta \mapsto p_\theta(x)$ *is two times continuously differentiable on* $\mathcal{V}_{\theta_0}$ *(*$\mu-$*almost everywhere) and* $\int \sup_{\theta\in\mathcal{V}_{\theta_0}} \|\ddot{p}_{\theta_0}(x)\|\mu(dx) < \infty$, *then*

$$I(\theta_0) = \mathbb{E}\left[\ddot{m}_{\theta_0}(X_1)\right] = \mathbb{E}\left[-\frac{\ddot{p}_{\theta_0}(X_1)}{p_{\theta_0}(X_1)} + \frac{\dot{p}_{\theta_0}(X_1)\dot{p}_{\theta_0}(X_1)^T}{p_{\theta_0}(X_1)^2}\right].$$

Proof. To show the result, it is sufficient to prove that

$$\mathbb{E}\left[\frac{\ddot{p}_{\theta_0}(X_1)}{p_{\theta_0}(X_1)}\right] = \int \ddot{p}_{\theta_0}(x)\mu(dx) = 0. \tag{2.4}$$

To this end, we apply the theorem of derivation under the sign integral. For the first integrability assumption, we know that $\mathbb{E}\left[\|\dot{p}_{\theta_0}(X_1)\|/p_{\theta_0}(X_1)\right] = \int \|\dot{p}_{\theta_0}(x)\|\mu(dx) < \infty$. Moreover

$$\sup_{\theta\in\mathcal{V}_{\theta_0}} \|\dot{p}_\theta(x)\| \leq \|\dot{p}_{\theta_0}(x)\| + \sup_{\theta\in\mathcal{V}_{\theta_0}} \|\ddot{p}_{\theta_0}(x)\| \times \left|\mathcal{V}_{\theta_0}\right|,$$

where $|\mathcal{V}_{\theta_0}|$ denotes the diameter of $\mathcal{V}_{\theta_0}$. We then get $\int \sup_{\theta\in\mathcal{V}_{\theta_0}} \|\dot{p}_\theta(x)\mu(dx) < \infty$. Finally, the theorem of derivation applies to $\theta \mapsto \int p_\theta(x)\mu(dx) = 1$ and (2.4) is valid.□

Note. If $\hat{\theta}_{n,MLE}$ denotes the MLE, under appropriate conditions, we get

$$\sqrt{n}\left(\hat{\theta}_{n,MLE}-\theta_0\right)\hookrightarrow\mathcal{N}_d\left(0,I(\theta_0)^{-1}\right).$$

It is possible to show that $I(\theta_0)^{-1}$ is the smallest asymptotic variance among the M-estimators for which the assumptions of Theorem 10 are valid (in fact, optimality of the MLE holds true under more general assumptions but this is outside the scope of this course). We first introduce a non-total order relation on the set of symmetric non-negative definite matrices of size $d\times d$,

$$A\preceq B\Leftrightarrow x^TAx\leq x^TBx,\quad x\in\mathbb{R}^d.$$

Note first that under our regularity assumptions

$$\int\dot{m}_{\theta_0}(x)p_{\theta_0}(x)\mu(dx)=0,\quad \theta_0\mathring{\Theta}.$$

Taking the derivative with respect to $\theta_0\in\mathring{\Theta}$ in the previous equality

$$\int\ddot{m}_{\theta_0}(x)p_{\theta_0}(x)\mu(dx)+\int\dot{m}_{\theta_0}(x)\dot{p}_{\theta_0}(x)^T\mu(dx)=0.$$

From this identity, if $x,y\in\mathbb{R}^d$, we get

$$\begin{aligned}x^T\mathbb{E}\left[\ddot{m}_{\theta_0}(X_1)\right]y &= -\mathbb{E}\left[x^T\dot{m}_{\theta_0}(X_1)\frac{\dot{p}_{\theta_0}(X_1)^T}{p_{\theta_0}(X_1)}y\right]\\ &\leq \sqrt{\mathbb{E}\left[x^T\dot{m}_{\theta_0}(X_1)\dot{m}_{\theta_0}(X_1)^Tx\right]\cdot\mathbb{E}\left[y^T\frac{\dot{p}_{\theta_0}(X_1)\dot{p}_{\theta_0}(X_1)^T}{p_{\theta_0}(X_1)^2}y\right]},\end{aligned}$$

where we applied the Cauchy-Schwarz inequality. Setting $x=W_{\theta_0}^{-1}z$ and $y=I(\theta_0)^{-1}z$ for some $z\in\mathbb{R}^d$, we get

$$z^TI(\theta_0)^{-1}z\leq\sqrt{z^TW_{\theta_0}^{-1}V_{\theta_0}W_{\theta_0}^{-1}z}\times\sqrt{z^TI(\theta_0)^{-1}z}$$

and then

$$z^TI(\theta_0)^{-1}z\leq z^TW_{\theta_0}^{-1}V_{\theta_0}W_{\theta_0}^{-1}z.$$

In particular for any $z\in\mathbb{R}^d$, the asymptotic variance of the linear combination $z^T\hat{\theta}_{n,MLE}$ is smaller than $z^T\hat{\theta}_n$ where $\hat{\theta}_n$ is another $M-$estimator. This justifies the asympotic optimality property of the MLE under suitable regularity conditions.

2.1.7 Model selection. Akaike information criterion

Usually, for M-estimators, we have several natural submodels that we want to select. For instance, in the case of a regression model with p predictors available, the true model could write as

$$Y_i=\sum_{j\in\mathcal{M}_0}\theta_{0,j}X_{j,i}+\varepsilon_i,\quad 1\leq i\leq n,$$

where $\mathcal{M}_0$ is a subset of $\{1,\dots,p\}$. In this case, we may want to select the good subset of predictors $\mathcal{M}_0$. In the general case of M-estimators, we have a finite collection of models $\mathcal{M}$ and we have a finite number of estimators $\hat{\theta}_{n,\mathcal{M}} = \arg\min_{\theta\in\Theta_{\mathcal{M}}} M_n(\theta)$ where $\Theta_{\mathcal{M}}$ denotes the parameter space corresponding to a submodel $\mathcal{M}$. Set $\theta_{0,\mathcal{M}} = \arg\min_{\theta\in\Theta_{0,\mathcal{M}}} M(\theta)$ and $\mathcal{M}_0$ such that $\theta_{0,\mathcal{M}_0} = \arg\min_{\mathcal{M}} M\left(\theta_{0,\mathcal{M}}\right)$. Note that in the case of nested models, i.e. $\mathcal{M}_0 \subset \mathcal{M}$, $\theta_{0,\mathcal{M}}$ can be identified to $\theta_{0,\mathcal{M}_0}$ and this vector will be simply denoted by θ_0.

A first idea to estimate $\mathcal{M}_0$ would be to minimize $\mathcal{M} \mapsto M_n\left(\hat{\theta}_{n,\mathcal{M}}\right)$ but unfortunately the selected model $\mathcal{M}$ is generally much larger than $\mathcal{M}_0$ (overfitting problem). It could be then more interesting to minimize $\mathcal{M} \mapsto M\left(\hat{\theta}_{n,\mathcal{M}}\right)$ but M is unknown.

In the rest of the discussion, suppose that $\mathcal{M} \supset \mathcal{M}_0$ and for simplicity write $\hat{\theta}_n$ instead of $\hat{\theta}_{n,\mathcal{M}}$. We remind that under some assumptions, if $\mathcal{M} \supset \mathcal{M}_0$,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \hookrightarrow \mathcal{N}_{|\mathcal{M}|}\left(0, W_{\theta_0,\mathcal{M}}^{-1} V_{\theta_0,\mathcal{M}} W_{\theta_0,\mathcal{M}}^{-1}\right),$$

where $|\mathcal{M}|$ denotes the number of free parameters in model $\mathcal{M}$. Under some regularity conditions, we have

$$\begin{aligned}
M\left(\hat{\theta}_n\right) &= M(\theta_0) + \dot{M}(\theta_0)\left(\hat{\theta}_n - \theta_0\right) + \frac{1}{2}\left(\hat{\theta}_n - \theta_0\right)^T W_{\theta_0,\mathcal{M}}\left(\hat{\theta}_n - \theta_0\right) + o_{\mathbb{P}}(1/n) \\
&= M(\theta_0) + \frac{1}{2}\left(\hat{\theta}_n - \theta_0\right)^T W_{\theta_0,\mathcal{M}}\left(\hat{\theta}_n - \theta_0\right) + o_{\mathbb{P}}(1/n), \\
M_n\left(\hat{\theta}_n\right) &= M_n(\theta_0) + \dot{M}_n(\theta_0)\left(\hat{\theta}_n - \theta_0\right) + \frac{1}{2}\left(\hat{\theta}_n - \theta_0\right)^T W_{\theta_0,\mathcal{M}}\left(\hat{\theta}_n - \theta_0\right) + o_{\mathbb{P}}(1/n), \\
\dot{M}_n(\theta_0) &= \dot{M}_n\left(\hat{\theta}_n\right) - \ddot{M}_n(\theta_0)\left(\hat{\theta}_n - \theta_0\right) + o_{\mathbb{P}}\left(1/\sqrt{n}\right).
\end{aligned}$$

We then get

$$M_n\left(\hat{\theta}_n\right) = M_n\left(\theta_0\right) - \frac{1}{2}\left(\hat{\theta}_n - \theta_0\right)^T W_{\theta_0,\mathcal{M}}\left(\hat{\theta}_n - \theta_0\right) + o_{\mathbb{P}}(1/n)$$

and finally

$$M\left(\hat{\theta}_n\right) - \tau_n = M_n\left(\hat{\theta}_n\right) + \left(\hat{\theta}_n - \theta_0\right)^T W_{\theta_0,\mathcal{M}}\left(\hat{\theta}_n - \theta_0\right) + o_{\mathbb{P}}(1/n), \tag{2.5}$$

with $\tau_n = M(\theta_0) - M_n(\theta_0)$ not depending on $\mathcal{M} \supset \mathcal{M}_0$. Since

$$\arg\min_{\mathcal{M}} M\left(\hat{\theta}_{n,\mathcal{M}}\right) = \arg\min_{\mathcal{M}}\left\{M\left(\hat{\theta}_{n,\mathcal{M}}\right) - \tau_n\right\},$$

we see from (2.5) that $M_n\left(\hat{\theta}_{n,\mathcal{M}}\right)$ under estimate $M\left(\hat{\theta}_{n,\mathcal{M}}\right) - \tau_n$. In AIC criterion, we replace the correcting term $\left(\hat{\theta}_n - \theta_0\right)^T W_{\theta_0,\mathcal{M}}\left(\hat{\theta}_n - \theta_0\right)$ by its expectation and we let $n \to \infty$. Using the Gaussian limiting distribution, we get

$$\mathbb{E}\left(\hat{\theta}_n - \theta_0\right)^T W_{\theta_0,\mathcal{M}}\left(\hat{\theta}_n - \theta_0\right) \approx \frac{1}{n}\ \mathrm{Tr}\ \left(V_{\theta_0,\mathcal{M}} W_{\theta_0,\mathcal{M}}^{-1}\right).$$

For the MLE

$$\text{Tr}\ \left(V_{\theta_0,\mathcal{M}} W_{\theta_0,\mathcal{M}}^{-1}\right) = |\mathcal{M}|\,.$$

In this case, we define

$$\hat{\mathcal{M}} = \arg\min_{\mathcal{M}} \left\{ M_n\left(\hat{\theta}_{n,\mathcal{M}}\right) + \frac{|\mathcal{M}|}{n} \right\}\ \text{[AIC criterion]}.$$

In the case of a general M-estimator, it is necessary to estimate $V_{\theta_0,\mathcal{M}}$ and $W_{\theta_0,\mathcal{M}}$.

2.1.8 Additional results for convex objective functions

We now consider the case of convex objective functions $\theta \mapsto M_n(\theta)$. In this case, asymptotic results are easier to state.

We first start with a useful result showing that pointwise convergence of convex functions entails uniform convergence on compact subsets. The following technical lemma is given without proof. See Tyrrell Rockafellar (1970), Theorem 10.8.

Lemma 6. *Let U be an open and convex subset of $\mathbb{R}^d$ and $(f_n)_{n\in\mathbb{N}}$ a sequence of convex functions from U to $\mathbb{R}$. If there exist a convex function $f : U \to \mathbb{R}$ and a dense subset $D \subset U$ such that $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in D$, then $(f_n)_{n\in\mathbb{N}}$ converges to f uniformly on any compact subset of U.*

Corollary 1. *Suppose that all the assumptions of Lemma 6 are valid. Additionally, suppose that f has a unique minimizer $x^* \in U$. Then if n is large enough, f_n is lower-bounded, reaches its minimal value and the sequence of argmin converges to x^*.*

Proof. Let $\epsilon > 0$ such that $K := \overline{B}\,(x^*, \epsilon) \subset U$, where $\overline{B}\,(x^*, \epsilon) = \{x \in U : \|x - x^*\| \leq \epsilon\}$. From Lemma 6, we have $r_n := \sup_{x\in K} |f_n(x) - f(x)| \overset{n\to\infty}{\to} 0$. We remind that a convex function defined on an open subset of $\mathbb{R}^d$ is always continuous. Since the boundary of K, ∂K, is compact (as a closed subset of K compact), we have $\delta := \inf_{x\in\partial K} (f(x) - f(x^*)) > 0$. This is due to the fact that $f(x) > f(x^*)$ for any $x \in \partial K$ and to the continuity of f. We are now going to show that $\inf_U f_n = \inf_K f_n$ is n is large enough. Since K is compact and f_n continuous, we will have $\inf_K f_n$ is reached at a point $x_n \in K$. Let $y \in U \setminus K$. There exists $\lambda \in (0, 1)$ such that $x = \lambda x^* + (1 - \lambda)y \in \partial K$. Then the following inequalities hold true.

$$f_n(x^*) \leq f(x^*) + r_n \leq f(x) + r_n - \delta \leq f_n(x) + 2r_n - \delta \leq \lambda f_n(x^*) + (1 - \lambda) f_n(y) + 2r_n - \delta.$$

If n is large enough, $2r_n - \delta < 0$ and then the previous inequalities yield to $f_n(x^*) < f_n(y)$. We conclude that $\inf_U f_n = \inf_K f_n = f_n(x_n)$ for some $x_n \in K$ and $\|x_n - x^*\| \leq \epsilon$ if n is large enough.□

We now go back to the problem of consistency of M-estimators.

Theorem 12. *Let Θ be an open and convex of $\mathbb{R}^d$. Suppose that the three following assumptions hold true.*

1. *The sequence $(M_n)_n$ is a sequence of convex random functions converging point by point on a dense subset of Θ to a deterministic convex function M a.s. (resp. in probability).*

2. *There exists a sequence $\left(\hat{\theta}_n\right)_{n\geq 1}$ of near-argmin of $(M_n)_{n\geq 1}$, in the sense that $\hat{\theta}_n - \inf_\Theta M_n \overset{n\to\infty}{\to} 0$ a.s. (resp. in probability).*

3. *θ_0 is the unique minimizer of M.*

Then $\hat{\theta}_n$ converges a.s. (resp. in probability) to θ_0.

Notes

1. If the sequence of convex functions $(M_n)_{n\geq 1}$ converges pointwise to function $M : \Theta \to \mathbb{R}$, then M is automatically convex. This is true in the deteministic case and then automatic for a.s. convergence. For the convergence in probability, use the a.s. convergence along a subsequence to conclude the convexity of the limit.

2. The existence of a near-argmin can be useful in some examples when Θ is unbounded. See below for the logistic regression.

Proof of Theorem 12. The almost sure convergence is a consequence of Corollary 1. In particular, taking an arbitrary dense subset D of Θ, it can be shown that for $\mathbb{P}-$almost all $\omega \in \Omega$, for all $\theta \in D$, $\lim_{n\to\infty} M_n(\theta)_\omega = M(\theta)$.

For the convergence in probability, we remind that $Z_n \overset{p}{\to} Z$ if and only if for any subsequence of $(Z_n)_n$, there exists a subsubsequence converging to Z a.s. (remind that convergence in probability entails almost sure convergence of a subsequence). Let $\overline{M}_n = M_{\phi(n)}$ a subsequence of M_n and $(s_j)_{j\geq 1}$ dense in Θ. We have $\overline{M}_n(s_j) \overset{p}{\to} M(s_j)$ for all $j \geq 1$. Let ℓ be a positive integer. There then exists an integer n_ℓ such that for all $1 \leq j \leq \ell$,

$$\mathbb{P}\left(\left|\overline{M}_{n_\ell}(s_j) - M(s_j)\right| > 1/\ell\right) \leq 2^{-\ell}.$$

One can assume that $n_\ell \leq n_{\ell+1}$. We deduce that for all $\epsilon > 0$ and $j \geq 1$,

$$\sum_{\ell=1}^{\infty} \mathbb{P}\left(\left|\overline{M}_{n_\ell}(s_j) - M(s_j)\right| > \epsilon\right) < \infty.$$

From the Borel-Cantelli lemma, we have for any $j \geq 1$, $\overline{M}_{n_\ell}(s_j) \overset{a.s.}{\to} M(s_j)$ a.s. Since the set of point of a.s. convergence is numerable, one can deduce that for $\mathbb{P}-$almost all $\omega \in \Omega$, $\lim_{n\to\infty} \overline{M}_{n_\ell}(s_j)_\omega = M(s_j)$ for any $j \geq 1$. We then deduce form Corollary 1 that $\hat{\theta}_{\phi(n_\ell)} \overset{a.s.}{\to} \theta_0$. This concludes the proof.□

Examples

1. The geometric median. Let $\hat{\theta}_n = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \|X_i - \theta\|$, $\Theta = \mathbb{R}^d$. Here M_n and M are convex. We recover the convergence of $\hat{\theta}_n$ to θ_0 obtained in the previous section, provided that θ_0 is the unique argmin of $\theta \mapsto M(\theta) = \mathbb{E}(\|X_1 - \theta\|)$.

2. Logistic regression. Let $(X_i, Y_i) \in \mathbb{R}^d \times \{0,1\}$, $1 \le i \le n$, some i.i.d. pairs of random variables such that $\mathbb{P}(Y_i = 1|X_i = x) = F\left(x^T\theta_0\right)$ with $F(z) = (1+\exp(-z))^{-1}$. Here $\Theta = \mathbb{R}^d$. One can include an intercept term in the linear combination $X_i^T\theta_0$, (i.e. the first component of X_i is equal to 1). Let

$$g_{x,y}(\theta_0) = \mathbb{P}(Y_i = 1|X_i = x) = F\left(x^T\theta_0\right)^y \left(1 - F\left(x^T\theta_0\right)\right)^{1-y}.$$

We set

$$m_\theta(Y_i, X_i) = -\log g_{X_i,Y_i}(\theta) = -Y_i X_i^T\theta + \log\left(1 + \exp(X_i^T\theta)\right).$$

The corresponding M-estimator corresponds to the conditional MLE because it is based on the maximization of the logarithm of the conditional density of $(Y_1, \ldots, Y_n)$ given $(X_1, \ldots, X_n)$. Note that

$$\ddot{m}_\theta(y, x) = \frac{xx^T \exp(x^T\theta)}{1 + \exp(x^T\theta)}.$$

We deduce that $\ddot{M}_n(\theta)$ is a positive semi-definite matrix. Hence the convexity of M_n. To show that M is well defined, we only have to check integrability of $m_\theta(Y_1, X_1)$. This is satisfied as soon as $\mathbb{E}(\|X_1\|) < \infty$.

We now check under which condition θ_0 is the unique argmin of M. To this end, one can check that

$$M(\theta) - M(\theta_0) = \int KL\left(\mathcal{B}\left(F(\theta^T x), F(\theta_0^T x)\right)\right) d\mathbb{P}_{X_1}(x),$$

where $KL\left(\mathcal{B}\left(F(\theta^T x), F(\theta_0^T x)\right)\right)$ denotes the Kullback-Leibler divergence between the Bernoulli distributions with respective parameters $F(\theta^T x)$ and $F(\theta_0^T x)$. Hence we get $M(\theta) \ge M(\theta_0)$ and the equality only holds if $F(\theta^T x) = F(\theta_0^T x)$ for $\mathbb{P}_{X_1}-$almost all x. Since F is one-to-one, we get $M(\theta) = M(\theta_0)$ if and only if $\theta^T X_1 = \theta_0^T X_1$ a.s. We conclude that θ_0 is the unique minimizer of M if and only if the components of X_1 are linearly independent.

The most difficult problem is to check the second assumption of Theorem 12. Existence of the MLE does not hold when data coming from $Y_i = 1$ and $Y_i = 0$ are separated by an hyperplane of $\mathbb{R}^d$. See Albert and Anderson (1984) for a precise statement. However, for a fixed ω, Corollary 1 ensures that $\theta \mapsto M_n(\theta)_\omega$ has a minimizer $\hat{\theta}_n(\omega)$ if n is large enough. A sequence of near-argmin then exists.

We now consider asymptotic normality of minimizers of convex criteria. We first consider a simple result which is not difficult to prove. A much more general result will be given without proof at the end of the subsection. The results below is formulated for $M_n(\theta) = \sum_{i=1}^n m_\theta(X_i)$, without the $1/n$ normalization.

Theorem 13. *Let M_n be a convex random mapping defined on $\mathbb{R}^d$ with $\hat{\theta}_n$ as near-argmin. Suppose that for any $z \in \mathbb{R}^d$,*

$$M_n\left(\theta_0 + z/\sqrt{n}\right) - M_n(\theta_0) = \frac{1}{2} z^T V z + U_n^T z + E_n + r_n(z),$$

with V symmetric, positive definite, non random and $r_n(z) = o_{\mathbb{P}}(1)$, $U_n = O_{\mathbb{P}}(1)$.

Then

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) &= \arg\min_{z \in \mathbb{R}^d} \left\{ \frac{1}{2} z^T V z + U_n^T z + E_n \right\} + o_{\mathbb{P}}(1) \\ &= -V^{-1} U_n + o_{\mathbb{P}}(1).\end{aligned}$$

Moreover if $U_n \hookrightarrow U$ then $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \hookrightarrow -V^{-1}U$.

Proof of Theorem 13 Let

$$D_n(z) = M_n\left(\theta_0 + z/\sqrt{n}\right) - M_n(\theta_0), \quad \overline{D}_n(z) = \frac{1}{2} z^T V z + U_n^T z + E_n.$$

The mapping D_n (resp. $\overline{D}_n$) is convex and has a minimizer $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ (resp. $Z_n := -V^{-1}U_n$). The mapping $z \mapsto D_n(z) - U_n^T z - E_n$ is also convex and converges pointwise to $z \mapsto z^T V z$ in probability. From Lemma 6 and a subsequence argument, one can deduce that the convergence is uniform on compact sets. As a consequence, $r_n = D_n - \overline{D}_n$ converges in probability to 0, uniformly on compact sets. More precisely, for any compact subset K of $\mathbb{R}^d$, $\sup_{z \in K} |r_n(z)| = o_{\mathbb{P}}(1)$.

In what follows, we set $K_n = \overline{B}\left(Z_n, \varepsilon\right)$, $R_n = \sup_{z \in K_n} \left|D_n(z) - \overline{D}_n(z)\right|$ and

$\Delta_n = \inf_{z \in \partial K_n} \left\{\overline{D}_n(z) - \overline{D}_n\left(Z_n\right)\right\}$, where ∂K_n denotes the boundary of K_n. Observe that K_n is a random compact subset of $\mathbb{R}^d$ because the center of the ball is a random variable. Note also that $\Delta_n > 0$ a.s. The proof will be based on the following lemma.

Lemma 7. *Let $\omega \in \Omega$. If $\Delta_n(\omega) > 2R_n(\omega)$ and $D_n(y)_\omega < \inf_z D_n(z)_\omega + \Delta_n(\omega) - 2R_n(\omega)$, then $y \in K_n(\omega)$.*

Proof of Lemma 7 If $y \notin K_n(\omega)$, there exists $x \in \partial K_n(\omega)$ such that $x = \lambda Z_n(\omega) + (1-\lambda)y$ for some $\lambda \in (0,1)$. We then get the following upper-bounds.

$$\begin{aligned} D_n\left(Z_n(\omega)\right)_\omega &\leq \overline{D}_n\left(Z_n(\omega)\right)_\omega + R_n(\omega) \\ &\leq \overline{D}_n(x)_\omega + R_n(\omega) - \Delta_n(\omega) \\ &\leq D_n(x)_\omega + 2R_n(\omega) - \Delta_n(\omega) \\ &\leq \lambda D_n\left(Z_n(\omega)\right)_\omega + (1-\lambda) D_n(y)_\omega + 2R_n(\omega) - \Delta_n(\omega). \end{aligned}$$

Using our assumptions, we then get

$$D_n\left(Z_n(\omega)\right)_\omega \leq D_n(y)_\omega + \frac{2R_n(\omega) - \Delta_n(\omega)}{1-\lambda} < D_n(y)_\omega + 2R_n(\omega) - \Delta_n(\omega) < D_n\left(Z_n(\omega)\right)_\omega.$$

This yields to a contradiction.□

End of the proof of Theorem 13. From Lemma 7, we get the following inclusion

$$A_n := \{\Delta_n > 2R_n\} \cap \left\{ D_n\left(\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)\right) < \inf D_n + \Delta_n - 2R_n \right\} \subset \left\{ \|\sqrt{n}(\hat{\theta}_n - \theta_0) - Z_n\| \leq \varepsilon \right\}.$$

It is enough to prove that $\lim_{n\to\infty} \mathbb{P}(A_n) = 1$; this will prove that $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) - Z_n = o_{\mathbb{P}}(1)$ and the conclusion of the theorem will follow from Slutsky's lemma. Now let $h \in \mathbb{R}^d$ such that $\|h\| = \varepsilon$. We have

$$\overline{D}_n\left(Z_n + h\right) - \overline{D}_n(Z_n) = \frac{1}{2} h^T V h.$$

We conclude that

$$\Delta_n = \inf_{\|h\|=\varepsilon} \frac{1}{2} h^T V h = \frac{1}{2}\lambda_- \varepsilon^2,$$

where $\lambda_- > 0$ is the smallest eigenvalue of V. Next for $\kappa > 0$, we have

$$\mathbb{P}\left(R_n > \kappa\right) \leq \mathbb{P}\left(\|Z_n\| > M\right) + \mathbb{P}\left(\|Z_n\| \leq M, R_n > \kappa\right) := \alpha_n + \beta_n.$$

Since $Z_n = O_{\mathbb{P}}(1)$, one can choose M large enough so that $\sup_{n\geq 1} \alpha_n$ is arbitrarily small. For such a M,

$$\beta_n \leq \mathbb{P}\left(\sup_{\|z\| \leq M+\varepsilon} \left| D_n(z) - \overline{D}_n(z) \right| > \kappa \right) \overset{n\to\infty}{\to} 0.$$

We conclude that $R_n = o_{\mathbb{P}}(1)$ and then $\mathbb{P}\left(\Delta_n \leq 2R_n\right) \overset{n\to\infty}{\to} 0$. Finally, since $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$is a near argmin of D_n,

$$\mathbb{P}\left(D_n\left(\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)\right) \geq \inf D_n + \Delta_n - 2R_n \right) \overset{n\to\infty}{\to} 0$$

and we automatically get $\mathbb{P}(A_n^c) \overset{n\to\infty}{\to} 0$, which concludes the proof.□

Examples

1. Theorem 13 applies to logistic regression. It is simply necessary to make a Taylor expansion at order 2. If $\mathbb{E}\|X_1\|^2 < \infty$ and the coordinates of X_1 are linearly independent random variables, one can set $V = \mathbb{E}\left[\ddot{m}_{\theta_0}(X_1)\right]$ and $U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{m}_{\theta_0}(X_i)$.

2. Let us apply the result to the median in the univariate case, i.e. $m_\theta(x) = |x - \theta|$. We assume here that $X_1, \ldots, X_n$ are i.i.d. with a density f which is continuous and positive at point $\theta_0 := \inf\{x \in \mathbb{R} : F(x) \geq 1/2\}$, where F is the cumulative distribution function corresponding to f. In this case $\theta_0 = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}\left[|X_1 - \theta|\right]$ is the unique minimizer and satisfies $\mathbb{P}(X_1 \leq \theta_0) = \mathbb{P}\left(X \geq \theta_0\right) = 1/2$. See http://www.stat.yale.edu/~pollard/Papers/convex.pdf for additional results and examples. We have

$$m_{\theta_0+t}(x) - m_{\theta_0}(x) = D(x)t + R(x,t),$$

with $D(x) = \mathbb{1}_{x \leq \theta_0} - \mathbb{1}_{x > \theta_0}$ and

$$R(x,t) = \begin{cases} 2(t + \theta_0 - x)\mathbb{1}_{\theta_0 < x \leq \theta_0 + t} \text{ if } t > 0, \\ 2(y - t - \theta_0)\mathbb{1}_{\theta_0 + t < y \leq \theta_0} \text{ if } t < 0 \end{cases}$$

Of course $R(x,0) = 0$. We then get

$$\begin{aligned} M_n\left(\theta_0 + z/\sqrt{n}\right) - M_n(\theta_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D(X_i) z + \sum_{i=1}^{n} \mathbb{E}\left[R(X_i, z/\sqrt{n})\right] \\ &+ \sum_{i=1}^{n} \left\{R(X_i, z/\sqrt{n}) - \mathbb{E}\left[R(X_i, z/\sqrt{n})\right]\right\}. \end{aligned}$$

One can show that $\mathbb{E}\left[R(X_i,t)\right] = f(\theta_0)t^2 + o(t^2)$ and $\mathbb{E}\left[R(X_i,t)^2\right] = \frac{4}{3}|t|^3 f(\theta_0) + o(|t|^3)$. This shows that $r_n(z) = \sum_{i=1}^{n} \left\{R(X_i, z/\sqrt{n}) - \mathbb{E}\left[R(X_i, z/\sqrt{n})\right]\right\}$ satisfies $\mathbb{E}[r_n(z)^2] = o(1)$ and then $r_n(z) = o_{\mathbb{P}}(1)$ for any $z \in \mathbb{R}$. Moreover

$$\sum_{i=1}^{n} \mathbb{E}\left[R(X_i, z/\sqrt{n})\right] = f(\theta_0)z^2 + o(1).$$

An application of Theorem 13 yields to $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \hookrightarrow \mathcal{N}\left(0, \frac{1}{4f(\theta_0)^2}\right)$.

Finally, let us mention a more general result showing that minimizers of convex random functions always converge in distribution provided that the finite-dimensional distributions converges to that of a random convex function possessing a unique minimizer. The proof of the following theorem, which can be found under a more general framework in Kato (2009), is based on a representation theorem which allows to derive convergence in distribution from a.s. convergence. See http://www.stat.yale.edu/~pollard/Books/Iowa/Iowa-notes.pdf, Theorem (9.4) for a statement of the representation theorem.

Theorem 14. *Suppose that $z \mapsto g_n(z)$ are random convex functions defined on $\mathbb{R}^d$ such that $\hat{z}_n = \arg\min_{z \in \mathbb{R}^d} g_n(z)$ and $z \mapsto g_\infty(z)$ is another random convex function with a unique argmin z_∞. Then if for any $z_1, \ldots, z_k \in \mathbb{R}^d$,*

$$\left(g_n(z_1), \ldots, g_n(z_k)\right) \hookrightarrow \left(g(z_1), \ldots, g(z_k)\right),$$

we have $\hat{z}_n \hookrightarrow z_\infty$.

Typically, one can apply Theorem 14 to some criterion of type $g_n(z) = M_n\left(\theta_0 + z/\sqrt{n}\right) - M_n\left(\theta_0\right)$, but it is now not required to get a quadratic expansion as formulated in the statement of Theorem 13. We will apply this result in the next section.

2.2 An example of penalized regression method. LASSO type estimators

In this section, we consider a regression model of the form

$$Y_i = x_i^T \theta_0 + \varepsilon_i, \quad 1 \leq i \leq n,$$

with $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. with $\mathbb{E}(\varepsilon_1) = 0$ and $\mathbb{E}(\varepsilon_1^2) = \sigma^2 < \infty$. We consider a non-random design here, i.e. $x_1, \ldots, x_n$ are deterministic. For some $\gamma \geq 1$, we set

$$\hat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^d} L_n(\theta), \quad L_n(\theta) = \sum_{i=1}^n \left(Y_i - x_i^T \theta\right)^2 + \lambda_n \sum_{j=1}^d |\theta_j|^\gamma,$$

where $\lambda_n > 0$ is a hyperparameter selected by the user.

One can show that finding solutions of this penalized regression problem is equivalent to minimize $\theta \mapsto \sum_{i=1}^n \left(Y_i - x_i^T \theta\right)^2$ under the constraint $\|\theta\|_\gamma \leq R_n$ with a one-to-one correspondence between λ_n and R_n. Here $\|\theta\|_\gamma = \left(\sum_{j=1}^d |\theta_j|^\gamma\right)^{1/\gamma}$. The most popular choices are $\gamma = 1$ (LASSO regression) and $\gamma = 2$ (ridge regression) and are useful respectively if many components of the true θ_0 vanish or when the covariates are collinear.

We will investigate the asymptotic properties of penalized regression estimators when d is fixed and $n \to \infty$. Fu and Knight (2000) investigated these properties also for the case $\gamma \in (0, 1)$, though the arguments are no more based on convexity. We will use following assumption.

A1 There exists a positive-definite matrix C such that $C_n := \frac{1}{n} \sum_{i=1}^n x_i x_i^T \overset{n \to \infty}{\to} C$.

A2 We have $\max_{1 \leq i \leq n} \frac{\|x_i\|}{\sqrt{n}} = o(1)$.

Note. When the design is a realization of $X_1, \ldots, X_n$ i.i.d., **A1-A2** are satisfied if $\mathbb{E}\left[\|X_1\|^2\right] < \infty$ and $\mathbb{E}\left(X_1 X_1^T\right)$ is positive definite (or equivalently the coordinates of X_1 are linearly independent). Indeed **A1** is a consequence of the law of large numbers. Moreover

$$\begin{aligned} \frac{\|X_i\|^2}{n} &\leq \frac{\|X_i\|^2 \mathbb{1}_{\|X_i\| \leq M}}{n} + \frac{\|X_i\|^2 \mathbb{1}_{\|X_i\| > M}}{n} \\ &\leq \frac{M^2}{n} + \frac{1}{n} \sum_{j=1}^n \|X_j\|^2 \mathbb{1}_{\|X_j\| > M}. \end{aligned}$$

From the law of large numbers, we then get $\overline{\lim}_n \sum_{i=1}^n \frac{\|X_i\|^2}{n} \leq \mathbb{E}\left[\|X_1\|^2 \mathbb{1}_{\|X_1\| > M}\right]$ which goes to 0 as $M \to \infty$. In such a case, working with a deterministic design is equivalent to work with the conditional distribution of $(Y_1, \ldots, Y_n)$ given $X_1 = x_1, \ldots, X_n = x_n$. The advantage of working with a non-random design is the level of generality, since the sequence $(x_i)_{i \geq 1}$ is not required to be the realization of a sequence of i.i.d. random variables.

For consistency, we have the following result.

Theorem 15. *Suppose that Assumptions* **A1-A2** *hold true. If* $\lambda_n/n \to \lambda_0$, *then* $\hat{\theta}_n \xrightarrow{p} \arg\min_{\theta\in\mathbb{R}^d} L(\theta)$ *with*

$$L(\theta) = (\theta - \theta_0)^T C (\theta - \theta_0) + \lambda_0 \sum_{j=1}^{d} |\theta_j|^\gamma.$$

In particular, when $\lambda_0 = 0$ *(i.e.* $\lambda_n = o(n)$*), then* $\arg\min_{\theta\in\mathbb{R}^d} M(\theta) = \theta_0$.

We then conclude that $\lambda_n = o(n)$ is a necessary and sufficient condition to ensure consistency of the penalized regression estimator.

Proof of Theorem 15. The convex mapping L_n/n converges pointwise in probability to $L + \sigma^2$. Moreover L is strictly convex and $\lim_{\|\theta\|\to\infty} L(\theta) = \infty$ (since C is positive definite), then it has a unique minimizer. The result is then a consequence of Theorem 12. Note that a minimizer of L_n always exists because $\lim_{\|\theta\|\to\infty} L_n(\theta) = \infty$. □

For the asymptotic normality, we have the following result.

Theorem 16. *Suppose that Assumptions* **A1-A2** *holds true and that* $\lambda_n/\sqrt{n} \overset{n\to\infty}{\to} \lambda_0 \geq 0$.

1. *If* $\gamma > 1$, *then*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \hookrightarrow \mathcal{N}_d\left(-C^{-1}\frac{\lambda_0\gamma}{2}(sign(\theta_{0,j})|\theta_{0,j}|^{\gamma-1})_{1\leq j\leq d}, \sigma^2 C^{-1}\right).$$

2. *If* $\gamma = 1$, $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \hookrightarrow \arg\min_{z\in\mathbb{R}^d} V(z)$ *where*

$$V(z) = -2z^T W + z^T C z + \lambda_0 \sum_{j=1}^{d} \left\{ z_j sign(\theta_{0,j} \mathbb{1}_{\theta_{0,j}\neq 0} + |z_j| \mathbb{1}_{\theta_{0,j}=0} \right\}$$

and W *follows the distribution* $\mathcal{N}_d(0, \sigma^2 C)$.

Note. The asymptotic distribution of the ordinary least-squares estimator (i.e. with $\lambda_n = 0$) is $\mathcal{N}_d(0, \sigma^2 C^{-1})$. Indeed, we have

$$\hat{\theta}_n = \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - X_i^T\theta\right)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i Y_i$$

and

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i \varepsilon_i.$$

As stated in the beginning of the proof of Theorem 16, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i \hookrightarrow \mathcal{N}_d\left(0,\sigma^2 C\right)$$

and from **A1** and Slutsky's lemma, we deduce that $\sqrt{n}\left(\hat{\theta}_n-\theta_0\right)\hookrightarrow\mathcal{N}_d\left(0,\sigma^2C^{-1}\right)$. This is also the asymptotic distribution for penalized regression estimators when $\lambda_0=0$. However, the case $\lambda_0>0$ is interesting for getting additional properties. Let us consider the case $\gamma=1$. One can show that when $d-r$ coefficients vanish for the true model, then the asymptotic distribution of the penalized regression estimator put a positive mass to 0 for the corresponding coordinates. Set $\beta=\theta_0$. Without loss of generality, assume that $\beta_{r+1}=\cdots=\beta_d=0$ and $\beta_i\neq 0$ for $1\leq i\leq r$ (otherwise one can always index the variables accordingly). Set also $E=(C_{i,j})_{1\leq i,j\leq r}$, $F=(C_{i,j})_{r+1\leq i\leq d,1\leq j\leq r}$, $s(\beta)=(sign(\beta_j))_{1\leq j\leq r}$, W_1 and z_1 the first r components of W and z and W_2,z_2 their last $d-r$ components. One can show that z with $z_2=0$ is a solution for minimizing V if and only if the inequalities $-\frac{\lambda_0}{2}\mathbb{1}\leq Fz_1-W_2\leq\frac{\lambda_0}{2}\mathbb{1}$ hold true component by component, $\mathbb{1}$ being the vector of $\mathbb{R}^d$ with all coordinates equal to 1, and $z_1=E^{-1}\left(W_1-\frac{\lambda_0}{2}s(\beta)\right)$. This clearly happens with a positive probability. A more interesting property would be to show that the LASSO estimator recovers asymptotically the zero coefficients. Under some conditions, this property is true for a fixed p but also in a high-dimensional framework when p grows with n. See Zhao and Yu (2006).

Proof of Theorem 16. We have

$$\begin{aligned} L_n\left(\theta_0+z/\sqrt{n}\right)-L_n(\theta_0) &= z^TC_nz-\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i x_i^Tz \\ &+ \lambda_n\sum_{j=1}^{d}\left\{\left|\theta_j+\frac{z}{\sqrt{n}}\right|^\gamma-|\theta_j|^\gamma\right\}. \end{aligned}$$

Moreover, $z^TC_nz=z^TCz+o_{\mathbb{P}}(1)$ and $\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_ix_i\hookrightarrow\mathcal{N}_d\left(0,4\sigma^2C\right)$. For the second assertion, note that we have a sum of independent but not identically distributed random variables but one can use the central limit theorem given by Theorem 30 in Chapter 4. Indeed, setting $Y_{n,i}=x_i\varepsilon_i/\sqrt{n}$, we have $\sum_{i=1}^{n}\mathrm{Var}\left(Y_{n,i}\right)=\frac{\sigma^2}{n}\sum_{i=1}^{n}x_ix_i^T\overset{n\to}{\to}V:=\sigma^2C$. Moreover, the second assumption of Theorem 30 is satisfied, since $\|Y_{n,i}\|\leq c_n|\varepsilon_i|$ with $c_n=\max_{1\leq i\leq n}\|x_i\|/\sqrt{n}$ and for $\epsilon>0$,

$$\sum_{i=1}^{n}\mathbb{E}\left[\|Y_{n,i}\|^2\mathbb{1}_{\|Y_{n,i}\|>\epsilon}\right]\leq\frac{1}{n}\sum_{i=1}^{n}x_ix_i^T\mathbb{E}\left[\epsilon_1^2\mathbb{1}_{|\varepsilon_1|>\epsilon/c_n}\right],$$

with goes to 0 as $n\to\infty$, using **A1-A2** and the square integrability of ε_1.

1. Suppose first that $\gamma>1$. We have

$$\lambda_n\sum_{j=1}^{d}\left\{\left|\theta_{0,j}+\frac{z}{\sqrt{n}}\right|^\gamma-|\theta_{0,j}|^\gamma\right\}\overset{n\to\infty}{\to}\gamma\lambda_0\sum_{j=1}^{d}z_j|\theta_{0,j}|^{\gamma-1}sign(\theta_{0,j}).$$

The result then follows from Theorem 13, with $V = 2C$ and

$$U_n = -\frac{2}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i x_i + \gamma \lambda_0 \sum_{j=1}^{d} z_j |\theta_{0,j}|^{\gamma-1} sign(\theta_{0,j}).$$

2. Suppose next that $\gamma = 1$. We have

$$\lambda_n \sum_{j=1}^{d} \left\{ \left| \theta_{0,j} + \frac{z}{\sqrt{n}} \right| - |\theta_{0,j}| \right\} \overset{n\to\infty}{\to} \lambda_0 \sum_{j=1}^{d} \left\{ z_j sign(\theta_{0,j}) \mathbb{1}_{\theta_{0,j}\neq 0} + |z_j| \mathbb{1}_{\theta_{0,j}=0} \right\}.$$

The assumptions of Theorem 13 are satisfied only when $\lambda_0 = 0$ and we obtain a $\mathcal{N}_d\left(0, \sigma^2 C^{-1}\right)$ asymptotic distribution. If $\lambda_0 > 0$, one can use Theorem 14 with $g_n(z) = L_n\left(\theta_0 + z/\sqrt{n}\right) - L_n\left(\theta_0\right)$. Note that V is strictly convex and $\lim_{\|z\|\to\infty} V(z) = \infty$; there then exists a unique minimizer.□

2.3 Kernel density estimation

In this section, we go back to the problem of kernel density estimation. Let us assume that $X_1, \ldots, X_n$ are i.i.d. random vectors, taking values in $\mathbb{R}^k$ and for which $\mathbb{P}_{X_1} = f \cdot \lambda_k$ where λ_k is the Lebesgue measure on $\mathbb{R}^k$ and $f \in \mathcal{F}$ where $\mathcal{F}$ is a subset of the set of probability densities on $\mathbb{R}^k$. When $\mathcal{F} = \left\{p_\theta : \theta \in \Theta \subset \mathbb{R}^d\right\}$, we face to a parametric problem and maximum likelihood estimation as in the previous section can be studied. However, when we do not want to assume that $\mathcal{F}$ is a parametric family of probability densities, kernel density estimation is one of the classical method to estimate the unknown density f. Let $K : \mathbb{R}^k \to \mathbb{R}_+$ be a probability density and set $K_h(u) = h^{-k} K(u/h)$. As discussed in Chapter 0, one can define a natural estimator called kernel density estimator (KDE),

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \quad x \in \mathbb{R}^k.$$

There is a tuning parameter h to choose for computing the estimator. The additional parameter, the kernel K, plays a more minor rule in KDE convergence, only a few regularity properties are required for this kernel to compute the convergence rate of $\hat{f}_h$ to f. The optimal choice of h will depend on n, i.e. $h = h_n$ with $\lim_{n\to\infty} h_n = 0$ but the convergence to 0 should be not too fast. To get a better intuition about the properties of KDE, let us assume that $k = 1$ and $K(u)\frac{1}{2}\mathbb{1}_{[-1,1]}(u)$. In this case

$$\hat{f}_h(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x-h\leq X_i\leq x+h}}{2h}.$$

We then simply count the proportion of observations in the interval $[x-h, x+h]$, which estimates the probability $\mathbb{P}\left(x - h \leq X_1 \leq x + h\right)$ and divide this proportion by the length of

the interval. Intuitively this should converge to $f(x)$ if f is continuous at point x. However, if $h_n \searrow 0$ to fast with n, the KDE will exhibit too much variability. In contrast if h_n is too large, the KDE will be too flat. There is then a tradeoff for the choice of this tuning parameter. Figure 2.1 illustrates this problem when K is the Gaussian kernel, i.e. $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$.

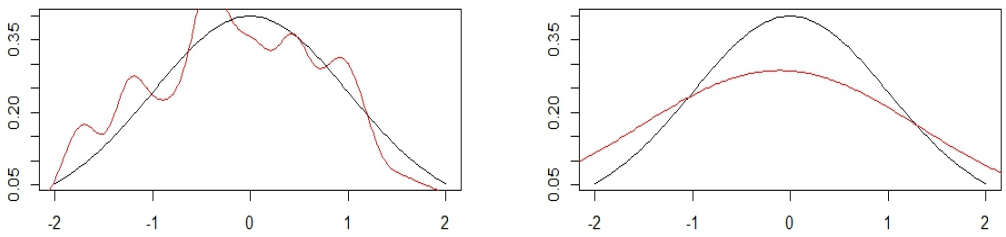

Figure 2.1: KDE (in red) for $n = 100$ standard Gaussian observations with $h = 0.1$ (left) and $h = 1$ (right)

2.3.1 Upper-bound for the integrated mean square error

The integrated mean squared error (MSE) is defined by

$$\text{IMSE } (f) = \int_{\mathbb{R}^k} \left(\hat{f}_h(x) - f(x)\right)^2 dx.$$

Setting $B_h(x) = \mathbb{E}\hat{f}_h(x) - f(x)$, the bias of the estimator at point x, we have the usual bias/variance decomposition

$$\text{IMSE } (f) = \int_{\mathbb{R}^k} B_h(x)^2 dx + \int_{\mathbb{R}^k} \text{Var } \left(\hat{f}_h(x)\right) dx.$$

Let us assume that every probability density f in $\mathcal{F}$ is two times continuously differentiable and let us compute the bias $B_h(x) = \mathbb{E}\hat{f}_h(x) - f(x)$. Using a Taylor expansion at order 2, we have

$$\begin{aligned} B_h(x) &= \int_{\mathbb{R}^k} K_h(x-y) f(y) dy - f(x) \\ &= \int_{\mathbb{R}^k} K(u) \left(f(x-hu) - f(x)\right) du \\ &= \int_{\mathbb{R}^k} K(u) \left[-h\nabla f(x)u + h^2 \int_0^1 (1-t) u^T \nabla^2 f(x-thu) u dt\right] du. \end{aligned}$$

In what follows, we denote by $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^k$. For a square matrix M of size $k \times k$, we also denote by $\|M\|$ the corresponding operator norm of M, i.e. $\|M\| = \sup_{\|x\|\leq 1} \|Mx\|$ (it coincides with the square of the spectral radius of the matrix $M^T M$). From the Cauchy-Schwarz inequality, we have $u^T M u \leq \|u\|^2 \|M\|$ for every vector u of $\mathbb{R}^k$. Assuming that $\int_{\mathbb{R}^k} uK(u)du = 0$ (it is the case when the kernel K is symmetric), the first term vanishes and applying the Cauchy-Schwarz inequality, we get

$$|B_h(x)|^2 \leq h^4 \int_{\mathbb{R}^k} \int_0^1 K(u)\|u\|^2 du \times \int_{\mathbb{R}^k} \int_0^1 (1-t)^2 K(u)\|u\|^2 \|\nabla^2 f\left(x - thu\right)\|^2 du dt.$$

We then get

$$\int_{\mathbb{R}^k} |B_h(x)|^2 \, dx \leq \frac{h^4}{3} \left(\int_{\mathbb{R}^k} \|u\|^2 K(u) du \right)^2 \int_{\mathbb{R}^k} \|\nabla^2 f(x)\|^2 dx.$$

Moreover, for the variance part, we have

$$\begin{aligned}
\int_{\mathbb{R}^k} \text{Var}\left(\hat{f}_h(x)\right) dx &= \frac{1}{n} \int_{\mathbb{R}^k} \text{Var}\left(K_h(x - X_1)\right) dx \\
&\leq \frac{1}{n} \int_{\mathbb{R}^k} \mathbb{E}\left(K_h(x - X_1)^2\right) dx \\
&= \frac{1}{nh^k} \int_{\mathbb{R}^k} \int_{\mathbb{R}^k} K(u)^2 f(x - hu) du dx \\
&= \frac{\int_{\mathbb{R}^k} K^2(u) du}{nh^k}.
\end{aligned}$$

We then get the following result.

Theorem 17. *Suppose that f is twice continuously differentiable on $\mathbb{R}^k$ with $\int_{\mathbb{R}^k} \|\nabla^2 f(x)\|^2 dx < \infty$. Suppose furthermore that $\int_{\mathbb{R}^k} uK(u)du = 0$, $\int_{\mathbb{R}^k} \|u\|^2 K(u)du < \infty$ and $\int_{\mathbb{R}^k} K^2(u)du < \infty$. There then exists a constant $C_{K,f} > 0$ such that*

$$\textit{IMSE}\,(f) \leq C_{K,f} \left(h^4 + \frac{1}{nh^k} \right).$$

Notes

1. If we optimize the upper-bound in $h > 0$, we find that $h_n \sim n^{-\frac{1}{4+k}}$ gives the best rate of convergence. In this case, the convergence rate of $\hat{f}_h$ is $n^{\frac{2}{4+k}}$ (considering the square root of the IMSE). One can note that the rate of convergence is slower than the standard $\sqrt{n}-$rate obtained in parametric estimation. However the space $\mathcal{F}$ of probability densities is of infinite dimension here and free of any parametric assumption that can be quite misleading in practice.

2. Suppose that $\mathcal{F}_M$ denotes the subset of probability densities $f : \mathbb{R}^k \to \mathbb{R}$ two-times continuously differentiable with $\int_{\mathbb{R}^k} \|\nabla^2 f(x)\| dx \leq M$. Then it can be shown that there exists a positive real number C_M such that for any density estimator $\hat{f}$ of f,

$$\inf_{f \in \mathcal{F}_M} \mathbb{E} \int_{\mathbb{R}^k} \left(\hat{f}(x) - f(x) \right)^2 dx \geq C_M n^{-\frac{4}{4+k}}.$$

   This shows that KDE are rate optimal.

3. See Van der Vaart (2000), Section 24.2 and Section 24.3 for a proof of the previous lower bound as well as an improvement of the convergence rate when f can be assumed to be $m-$times continuously differentiable with $m > 2$.

4. In practice h has to be selected from the sample, otherwise we only know that the optimal choice is of the form $h = Cn^{-\frac{1}{4+k}}$ with an unknown constant $C > 0$. There exist many "data-driven" procedures for bandwidth selection. One of the most popular is cross-validation. See in particular Hall (1983) and Stone (1984) for asymptotic results. Other selection methods are possible. See Goldenshluger and Lepski (2011) for an approach based on an estimation of the bias of KDE. A discussion about bandwidth selection methods is outside the scope of this course.

Chapter 3

An introduction to empirical process theory

This chapter is a short and partial introduction to empirical process theory. Most of the elements are taken from Van der Vaart (2000), Chapter 19. More complete references are Vaart and Wellner (2023) or Dudley (2014). The lecture notes available at `http://www.stat.columbia.edu/~bodhi/Talks/Emp-Proc-Lecture-Notes.pdf` are accessible and provide a list of interesting applications. In particular, Section 3.4.2 is taken from these notes.

3.1 Uniform weak convergence of random functions

Let $X_1, \ldots, X_n$ be some i.i.d. random vectors taking valued in $\mathbb{R}^k$ and with common probability distribution P. Let $\mathbb{P}_n$ be the empirical measure associated to a sample $X_1, \ldots, X_n$ taking values in $\mathbb{R}^k$, i.e. $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. For a measurable mapping $f : \mathbb{R}^k \to \mathbb{R}$, set $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$ and $Pf = \mathbb{E}[f(X_1)]$. If $\mathcal{F}$ denotes a set of measurable functions $f : \mathbb{R}^k \to \mathbb{R}$ (called a class of functions in what follows), $\{\mathbb{P}_n f : f \in \mathcal{F}\}$ is called an empirical process.

In this chapter, our aim will be two-fold.

1. First we are interested in the almost sure converge of $\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - Pf|$ to 0. When this convergence occurs, we will say that that the class $\mathcal{F}$ is $P-$Glivenko-Cantelli. It is clear that if $\mathcal{F}$ contains a finite number of elements, then it is $P-$Glivenko-Cantelli as soon as $P|f| < \infty$ for all $f \in \mathcal{F}$. This is a simple consequence of the strong law of large numbers. In Chapter 2, we have seen that the class $\{f_\theta : \theta \in \Theta\}$ is P-Glivenko-Cantelli as soon as Θ is a compact subset of $\mathbb{R}^d$, with $\theta \mapsto f_\theta(x)$ continuous on Θ for all $x \in \mathbb{R}^k$ and $x \mapsto \sup_{\theta\in\Theta} |f_\theta(x)|$ is $P-$integrable. When $f(x) = \mathbb{1}_{x\leq t}$, setting $\mathbb{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i\leq t}$ and $F(t) = P((-\infty, t])$, it is also widely known that

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} |\mathbb{F}_n(t) - F(t)| = 0 \text{ a.s.}$$

showing the class of half-line intervals is $P-$Glivenko Cantelli for any P. In this

chapter, we will derive more general conditions to get uniform convergence from a measure of complexity of the class $\mathcal{F}$.

2. We will also study the weak convergence of the random element $\{\mathbb{G}_n f := \sqrt{n}\,(\mathbb{P}_n f - \mathbb{P} f) : f \in \mathcal{F}\}$ as an element of $\ell^\infty(\mathcal{F})$. Here $\ell^\infty(\mathcal{F})$ denotes the set of mappings $g : \mathcal{F} \mapsto \mathbb{R}$ such that $\|g\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |g(f)| < \infty$. This will ensure that for every continuous and bounded function $h : \ell^\infty(\mathcal{F}) \to \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}\left[h\left(\mathbb{G}_n\right)\right] = \mathbb{E}\left[h\left(\mathbb{G}\right)\right] \tag{3.1}$$

for a random element $\mathbb{G}$ taking values in $\ell^\infty(\mathcal{F})$. For example, using a suitable function h (e.g. $h = \widetilde{h}\left(\|g\|_{\mathcal{F}}\right)$ with $\widetilde{h} : \mathbb{R} \mapsto \mathbb{R}$ continuous and bounded), this convergence will ensure the weak convergence of $\|\mathbb{G}_n\|_{\mathcal{F}}$ to $\|\mathbb{G}\|_{\mathcal{F}}$. Such a convergence will be interesting for non-parametric testing for instance.

When $f_1, \ldots, f_\ell \in \mathcal{F}$ are such that $\mathbb{P} f_i^2 < \infty$ for $1 \le i \le \ell$, the multivariate central limit theorem ensures that

$$\left(\mathbb{G}_n f_1, \ldots, \mathbb{G}_n f_\ell\right) \hookrightarrow \mathcal{N}_\ell\left(0, \Sigma\right),$$

where the covariance matrix of the limiting Gaussian vector is defined by

$$\Sigma(i,j) = \mathrm{Cov}\left(f_i(X_1), f_j(X_1)\right).$$

Then a good candidate for $\mathbb{G}$ is a Gaussian process, i.e. a process $\{\mathbb{G} f : f \in \mathcal{F}\}$ for which any finite-dimensional marginal vector $(\mathbb{G} f_1, \ldots, \mathbb{G} f_\ell)$ is a Gaussian vector, with mean 0 and covariance matrix $(P f_i f_j - \mathbb{P} f_i \cdot P f_j)_{1\le i,j\le \ell}$. But this finite-dimensional convergence property only ensures (3.1) for some specific functions h, i.e. $h(\mathbb{G}_n) = g\left(\mathbb{G}_n f_1, \ldots, \mathbb{G}_n f_\ell\right)$ for a continuous and bounded function $g : \mathbb{R}^\ell \to \mathbb{R}$. When $\mathcal{F}$ is not numerable, this is not sufficient to ensure convergence (3.1) for an arbitrary continuous and bounded function h.

To get an intuition on why convergence for finite-dimensional distributions is not sufficient for convergence in a uniform sense, let us consider the following elementary example. Consider $\ell^\infty([0,1])$ and the Dirac masses δ_{x_n} where $x_n : [0,1] \to \mathbb{R}$ is defined by $x_n(1/n) = 1$ and $x_n(t) = 0$ if $t \in [0,1] \setminus \{1/n\}$. Then δ_{x_n} convergences weakly to δ_0 for finite-dimensional distributions but since $\|x_n\|_{[0,1]} = 1$, one cannot expect convergence to 0 for the uniform topology. The same problem holds in empirical processes theory.

When $\mathbb{G}_n$ converges in distribution to $\mathbb{G}$, we will say the class $\mathcal{F}$ is $P-$Donsker.

3.1.1 Outer probabilities and expectations

There are often some problems of measurability of $\mathbb{G}_n$ taken as a random element in $\ell^\infty(\mathcal{F})$. For instance, take the simple example of the indicator functions $\mathcal{F} = \left\{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}\right\}$ and consider $\Omega = \mathbb{R}$, $X_1(\omega) = \omega$ and

$$\{\mathbb{P}_1 f : f \in \mathcal{F}\} = \{\mathbb{1}_{X_1 \le t} : t \in \mathbb{R}\}.$$

Here $\ell^\infty(\mathcal{F})$ can be identified to $\ell^\infty(\mathbb{R})$. Let S be a subset of $\mathbb{R}$ which is not a Borel set and $G = \cup_{s\in S} B_s$, where

$$B_s = \left\{ g \in \ell^\infty(\mathbb{R}) : \sup_{u\in\mathbb{R}} |g(u) - \mathbf{1}_{u\geq s}| < 1/2 \right\}.$$

Then G is an open set and then a Borel subset of $\ell^\infty(\mathbb{R})$. However

$$\{\omega \in \Omega : \mathbb{1}_{\omega\leq \cdot} \in G\} = S$$

because $\sup_{u\in\mathbb{R}} |\mathbb{1}_{\omega\leq u} - \mathbb{1}_{s\leq u}| = \mathbb{1}_{s\neq\omega}$. Then $\mathbb{P}_1$ is not measurable as an element of $\ell^\infty(\mathcal{F})$.

To circumvent measurability problems, we will consider an extension of the weak convergence notion using outer probability measures. This extension is presented to get a rigorous presentation of the results, **it can be skipped for a first reading**.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, (G, d) a metric space (for example $G = \ell^\infty(\mathcal{F})$ and $d(g, g') = \|g - g'\|_{\mathcal{F}}$), $X : \Omega \to G$ a random element and $f : G \to \mathbb{R}$ a mapping, we define

$$\mathbb{E}^*[f(X)] = \inf\{\mathbb{E}(U) : U : \Omega \to \mathbb{R} \text{ measurable }, U \geq f(X), \mathbb{E}(U) \text{ exists }\}.$$

The terminology $\mathbb{E}(U)$ exists means $\mathbb{E}(U_+) < \infty$ or $\mathbb{E}(U_-) < \infty$ where $U_+ = \max(U, 0)$ and $U_- = \max(-U, 0)$. We also have the following definition of an outer probability of a subset $B \subset \Omega$,

$$\mathbb{P}^*(B) = \inf\{\mathbb{P}(A) : A \in \mathcal{A}, B \subset A\}.$$

One can show that $\mathbb{P}^*(X \in C) = \mathbb{E}^*[\mathbb{1}_{X\in C}]$ for a subset C of G. See Vaart and Wellner (2023), chapter 1.2 for the main properties of outer probabilities and expectations. There also exists a notion of inner probability of a subset $B \subset \Omega$,

$$\mathbb{P}_*(B) = \sup\{\mathbb{P}(A) : A \in \mathcal{A}, A \subset B\}.$$

The notions of convergence are now as follows. The limit $X : \Omega \to G$ will be always assumed to be measurable in what follows.

Definition 2. *Let $X_n : \Omega \to G$, $n \geq 1$, be a sequence of random elements and $X : \Omega \to G$ a measurable mapping.*

1. *We say that $(X_n)_n$ converges in probability to X if for every $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}^*(d(X_n, X) > \epsilon) = 0$.*

2. *We say that $(X_n)_n$ converges weakly to X if for every continuous and bounded mapping $h : G \to \mathbb{R}$, $\lim_{n\to\infty} \mathbb{E}^*[h(X_n)] = \mathbb{E}[h(X)]$.*

3. *We say that $(X_n)_n$ converges a.s. to X if $d(X_n, X) \leq \Delta_n$ with Δ_n measurable and $\lim_{n\to\infty} \Delta_n = 0$ a.s.*

We then extend Portmanteau lemma in an arbitrary metric space using outer probabilities. The proof is very similar to the lemma proved in the first chapter of the course and is omitted.

Lemma 8. *Let $X_n : \Omega \to G$, $n \geq 1$, be a sequence of random elements and $X : \Omega \to G$ a measurable mapping. The following statements are equivalent.*

1. *For every continuous and bounded mapping $h : G \to \mathbb{R}$, $\lim_{n\to\infty} \mathbb{E}^* [h(X_n)] = \mathbb{E} [h(X)]$.*
2. *For every Lipschitz and bounded mapping $h : G \to \mathbb{R}$, $\lim_{n\to\infty} \mathbb{E}^* [h(X_n)] = \mathbb{E} [h(X)]$.*
3. *For every open subset O of G, $\liminf \mathbb{P}_* (X_n \in G) \geq \mathbb{P} (X \in G)$.*
4. *For every closed subset F of G, $\limsup \mathbb{P}^* (X_n \in F) \leq \mathbb{P} (X \in F)$.*
5. *For every Borel subset B of G such that $\mathbb{P} (X \in \delta B) = 0$, $\lim_{n\to\infty} \mathbb{P}^* (X_n \in B) = \mathbb{P} (X \in B)$.*

Analogues of Slutsky's lemma and the continuous mapping theorem follow similarly. **In the rest of the chapter, you can forgot the $\mathbb{P}^*$ of $\mathbb{P}_*$ notations and consider that they are simply equal to $\mathbb{P}$ (even if it is not true in theory).**

3.1.2 A criterion for convergence in distribution

Since convergence of finite-dimensional distributions is not sufficient for convergence in distribution in the space $\ell^\infty (\mathcal{F})$, we introduce an additional condition which guarantees this weak convergence. In the next result T denotes an arbitrary set.

Theorem 18. *A sequence $Z_n : \Omega \to \ell^\infty(T)$ converges weakly to a tight measurable random element $Z : \Omega \to \ell^\infty(T)$ if the two following conditions are satisfied.*

1. *For $t_1, \ldots, t_k \in T$, $(Z_n(t_1), \ldots, Z_n(t_k))$ converges in distribution in $\mathbb{R}^k$.*
2. *For every $\epsilon, \eta > 0$, there exist a partition $T_1, \ldots, T_k$ of T such that*

$$\limsup_{n\to\infty} \mathbb{P}^* \left(\max_{1\leq j\leq k} \sup_{s,t\in T_j} |Z_n(t) - Z_n(s)| \geq \epsilon \right) \leq \eta.$$

Notes

1. The notion of tight random element will be defined in the Appendix section 3.5.
2. The second assumption 2. in Theorem 18 plays the rule of an asymptotic equicontinuity condition. It means that one can always find a suitable partition of the index set T in such a way that the maximal increment of Z_n in the elements of this partition can be arbitrary small in probability.
3. The proof of Theorem 18 requires additional notions that are discussed in Section 3.5.

3.2 Bracketing numbers, entropy and uniform limit theorems

Let $X_1, \ldots, X_n$ be some i.i.d. random vectors taking values in $\mathbb{R}^k$. We denote by P their common probability distribution. For any real number $r \geq 1$, we denote by $L_r(P)$ the set of measurable mapping $f : \mathbb{R}^k \to \mathbb{R}$ such that $\int |f|^r dP < \infty$.

We first introduce the entropy with bracketing which is useful to measure the complexity of a class of measurable functions $\mathcal{F}$. For two measurable functions u and v from $\mathbb{R}^k$ to $\mathbb{R}$ such that $u \leq v$, the set of all functions $f : \mathbb{R}^k \to \mathbb{R}$ such that $u \leq f \leq v$ is called a bracket and is denoted by $[u, v]$. For $\varepsilon > 0$, an $\varepsilon-$bracket $[u, v]$ in $L_r(P)$ is a bracket such that $P(v-u)^r \leq \varepsilon^r$. The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, L_r(P))$ is the minimal number of $\varepsilon-$brackets needed to cover $\mathcal{F}$. Of course, the bracketing number increases when ε decreases. Note the the functions u and v need not to be elements of $\mathcal{F}$ (but they have to be in $L_r(P)$). The entropy with bracketing is defined as the logarithm of the bracketing number.

The next result guarantees that a finite entropy implies uniform convergence.

Theorem 19. *Suppose that $N_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$ for all $\varepsilon > 0$. Then $\mathcal{F}$ is $P-$Glivenko-Cantelli.*

Note. Of course if $N_{[]}(\varepsilon, \mathcal{F}, L_r(P)) < \infty$ for some $r > 1$, then $N_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$.

Proof of Theorem 19. Let $\varepsilon > 0$ and $[u_\ell, v_\ell]$, $1 \leq \ell \leq k$, some $\varepsilon-$brackets covering $\mathcal{F}$. This means that $\mathcal{F} \subset \cup_{\ell=1}^k [u_\ell, v_\ell]$ and $P(v_\ell - u_\ell) \leq \varepsilon$. Set $g_{1,n} = \max_{1 \leq \ell \leq k} |\mathbb{P}_n u_\ell - P u_\ell|$ and $g_{2,n} = \max_{1 \leq \ell \leq k} |\mathbb{P}_n v_\ell - P v_\ell|$. For $f \in [u_\ell, v_\ell]$, we have the following inequalities.

$$-\varepsilon - g_{1,n} \leq \mathbb{P}_n u_\ell - P u_\ell + P u_\ell - P v_\ell \leq \mathbb{P}_n f - P f \leq \mathbb{P}_n v_\ell - P v_\ell + P v_\ell - P u_\ell \leq g_{2,n} + \varepsilon.$$

This yields to

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - P f| \leq \max(g_{1,n}, g_{2,n}) + \varepsilon.$$

We then get

$$\overline{\lim}_n \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - P f| \leq \varepsilon.$$

This concludes the proof.□

We now turn out to a result which guarantees a uniform central limit theorem. Finiteness of the entropy is not sufficient for this. We require the root of the entropy to be integrable for $r = 2$. We then define

$$J_{[]}(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L_2(P))} d\varepsilon.$$

Note that finiteness of $J_{[]}(\delta, \mathcal{F}, L_2(P))$ for a particular value of δ entails finiteness of $J_{[]}(\delta', \mathcal{F}, L_2(P))$ for any $\delta' > 0$. In this case, the entropy is always finite for $r = 2$ and then for $r = 1$; the class $\mathcal{F}$ is $P-$Glivenko-Cantelli.

The following theorem is proved in Section 3.5. It is based on the criterion for weak convergence in $\ell^\infty(\mathcal{F})$, Theorem 4 and on a control of the expectation of $\|\mathbb{G}_n\|_{\mathcal{F}}$, given in Lemma 10.

Theorem 20 (Donsker). *Suppose that* $J_{[]}(1, \mathcal{F}, L_2(P)) < \infty$. *Then* $\mathcal{F}$ *is* $P-$*Donsker.*

Note. There also exists another standard notion of entropy, the entropy based on uniform covering numbers, which leads to interesting Glivenko-Cantelli or Donsker classes. See Van der Vaart (2000), Chapter 19 with an introduction to the specific case of VC classes of functions, widely encountered in empirical processes theory are which are defined through combinatorial properties. We will note discuss this alternative entropy notion in this course.

3.2.1 A few examples

Parametric classes. We revisit the example $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ where Θ is a compact subset of $\mathbb{R}^d$, $\theta \mapsto f_\theta(x)$ is continuous over Θ for all x and $F := \sup_{\theta\in\Theta} |f_\theta|$ is integrable with respect to P.

Let $\theta^* \in \Theta$ and B_δ be an open ball with center θ^* and radius δ. Set $u^\delta = \inf_{\theta\in B_\delta} f_\theta$ and $v^\delta = \sup_{\theta\in B_\delta} f_\theta$. From the dominated convergence theorem, we get

$$\lim_{\delta\searrow 0} P\left(v^\delta - u^\delta\right) = P\left(\lim_{\delta\searrow 0}\left(v^\delta - u^\delta\right)\right) = 0.$$

There then exists $\delta = \delta(\theta^*, \varepsilon)$ such that $P\left(v^\delta - u^\delta\right) \leq \varepsilon$. If $\Theta \subset \cup_{i=1}^k B(\theta_i, \delta(\theta_i, \varepsilon))$, then $\mathcal{F} \subset \cup_{i=1}^k \left[u_i^{\delta_i}, v_i^{\delta_i}\right]$ and $N_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$. However, we have no control on the size of the bracketing numbers.

Suppose now that there exists a measurable function $m : \mathbb{R}^k \to \mathbb{R}$ such that $Pm < \infty$ and

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq m(x)\|\theta_1 - \theta_2\|.$$

If $\Theta \subset \cup_{\ell=1}^k B(\theta_\ell, \varepsilon)$, let $u_\ell = f_{\theta_\ell} - \varepsilon m$ and $v_\ell = f_{\theta_\ell} + \varepsilon m$. For $\theta \in B(\theta_\ell, \varepsilon)$, note that $f \in [u_\ell, v_\ell]$ and the brackets cover $\mathcal{F}$. Moreover, $P(v_\ell - u_\ell) \leq 2\varepsilon Pm$. Then $N_{[]}(2\varepsilon Pm, \mathcal{F}, L_1(P)) \leq k$. To get the minimal value of k, we will suppose that $\|\cdot\|$ corresponds to the infinite norm. If it is not the case, there always exists $L > 0$ s.t. $\|\cdot\| \leq L\|\cdot\|_\infty$ and replacing m by Lm, our assumptions are satisfied for the infinite norm. For the infinite norm, if $\mathrm{diam}(\Theta)$ denotes the diameter of Θ and $\varepsilon < \mathrm{diam}\,(\Theta)$, the number of open balls of radius ε covering $\mathcal{F}$ is bounded by $2\left(\mathrm{diam}\,(\Theta)/\varepsilon\right)^d$ and one can obtain a covering of Θ with centers in Θ if we multiply the radius by 2. We then get the bound

$$N_{[]}(2\varepsilon Pm, \mathcal{F}, L_1(P)) \leq K\left(\mathrm{diam}\,(\Theta)/\varepsilon\right)^d,$$

where the constant K only depends on d and Θ. Then $\mathcal{F}$ is $P-$Glivenko-Cantelli. A similar analysis can be conducted with brackets in $L_2(P)$, as soon as Pm^2 and Pf_θ^2 are finite for $\theta \in \Theta$. The bracketing numbers are still bounded by ε^{-d}, up to a constant and $\mathcal{F}$ is also a $P-$Donsker class.

Distribution function. Let $\mathcal{F} = \left\{\mathbb{1}_{(-\infty,t]} : t \in \mathbb{R}\right\}$. Here

$$\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - Pf| = \sup_{t\in\mathbb{R}} |\mathbb{F}_n(t) - F(t)| .$$

Let us introduce the brackets $[\mathbb{1}_{(-\infty,t_{i-1}]}, \mathbb{1}_{(-\infty,t_i)}]$ with $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ chosen such that $F(t_i^-) - F(t_{i-1}) \leq \varepsilon$ for $1 \leq i \leq k$. For $s \in \mathbb{R}$, we set $F(s^-) = \lim_{x\nearrow s} F(x)$. To this end, one can use the generalized inverse of the cumulative distribution function,

$$F^{-1}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t\}, \quad t \in (0,1)$$

and set $t_i = F^{-1}(i\varepsilon)(\varepsilon)$. Indeed, we have the inequalities $F\left(F^{-1}(t)\right) \geq s$ and $F\left(F^{-1}(t^-)\right) \leq t$ for any $t \in (0,1)$. We have

$$N_{[]}\left(\varepsilon, \mathcal{F}, L_1(P)\right) \leq [1/\varepsilon] + 1 \leq 2/\varepsilon$$

and $\mathcal{F}$ is $P-$Glivenko-Cantelli. Since

$$P\left(\left|\mathbb{1}_{(-\infty,t_i)} - \mathbb{1}_{(-\infty,t_{i-1}]}\right|^2\right) = F\left(t_i^-\right) - F\left(t_{i-1}\right) \leq \varepsilon = \sqrt{\varepsilon}^2.$$

we get $N_{[]}\left(\varepsilon, \mathcal{F}, L_2(P)\right) \leq 4/\varepsilon^2$ and $\mathcal{F}$ is also $P-$Donsker. To summarize these important results, we give them as a corollary.

Corollary 2. *We have* $\sup_{t\in\mathbb{R}} |\mathbb{F}_n(t) - F(t)| \overset{a.s.}{\to} 0$. *Moreover* $\sqrt{n}\left(\mathbb{F}_n - F\right)$, *as a random element in* $\ell^\infty(\mathcal{F})$, *converges in distribution to a zero mean Gaussian process* $\mathbb{G}_F$ *with covariance* $\mathbb{E}\mathbb{G}_F(s)\mathbb{G}_F(t) = F\left(\min(s,t)\right) - F(s)F(t)$ *for* $s, t \in \mathbb{R}$.

When P is the uniform distribution over $[0,1]$, we have $\mathbb{E}\mathbb{G}_F(s)\mathbb{G}_F(t) = \min(s,t) - st$ for $0 \leq s, t \leq 1$. The corresponding Gaussian process has the same probability distribution as the Gaussian process $\{U_t := B_t - tB_1 : 0 \leq t \leq 1\}$, where $\{B_t : t \geq 0\}$ is a Gaussian process called Brownian motion, that is a centered Gaussian process with covariance $\mathbb{E}B_tB_s = \min(s,t)$, $s, t \geq 0$. The process $\{U_t : 0 \leq t \leq 1\}$ is called Brownian bridge. For a general distribution function F, one can check that $\mathbb{G}_F$ has the same probability distribution as the process $\left\{U_{F(t)} : 0 \leq t \leq 1\right\}$ which is called $F-$Brownian bridge.

A "bigger" class of functions. Donsker classes can be obtained as soon as the entropy $\log N_{[]}\left(\varepsilon, \mathcal{F}, L_2(P)\right)$ can be bounded by $C\varepsilon^{-2+\delta}$ with some $\delta > 0$. The previous classes of functions were small, because the entropy was of order $\log(1/\varepsilon)$. Consider the class of Lipschitz functions $\mathcal{F} = \{f : [0,1] \to [0,1] : |f(x) - f(y)| \leq |x - y|\}$. Let $\varepsilon > 0$ and $a_i = i\varepsilon$, for $i \in \mathbb{Z}$. Setting $A_i = (a_{i-1}, a_i] \cap [0,1]$ for $1 \leq i \leq k$ where k is the first integer greater than $1/\varepsilon$, we consider some functions of the form $u = \sum_{i=1}^k a_{\ell_i} \mathbb{1}_{A_i}$ where $\ell_i \in \mathbb{Z}$ for $i = 1, \ldots, k$. Let $f \in \mathcal{F}$ and set s_i the integer part of $f(a_{i-1})/\varepsilon$. If $x \in A_i$, we have

$$\varepsilon(s_i - 1) \leq f(a_{i-1}) - \varepsilon \leq f(x) \leq f(a_{i-1}) + \varepsilon \leq \varepsilon(s_i + 2).$$

Moreover, the Lipschitz properties of f guaranty that $s_i - 2 \leq s_{i+1} \leq s_i + 2$. We deduce that f is an element of a bracket $[u, v]$ with $u = \sum_{i=1}^k a_{\ell_i} \mathbb{1}_{A_i}$ and $v = \sum_{i=1}^k a_{\ell_i+3} \mathbb{1}_{A_i}$ with $\ell_{i+1} \in \{\ell_i - 2, \ell_i - 1, \ell_i, \ell_i + 1, \ell_i + 2\}$. The number of such brackets, which have size controlled by 3ε for the infinite norm, is smaller the $\frac{1}{\varepsilon}5^{1/\varepsilon}$ (up to a constant). This is smaller that $\exp(C/\varepsilon)$ for a suitable $C > 0$. This class is then $P-$ Donsker for any P.

Additional examples can be found in Vaart and Wellner (2023) and Van der Vaart (2000), Chapter 19.

3.3 Maximal inequalities

In this section, our aim is to control the expectation $\mathbb{E}^*\|\mathbb{G}_n\|_{\mathcal{F}}$. We start with a useful exponential inequality.

Proposition 3 (Bernstein inequality)**.** *Let $Y_1, \ldots, Y_n$ be some i.i.d. random variables, centered and bounded by M. Set $v = \mathbb{E}(Y_1^2)$ and $S_n = \sum_{i=1}^n Y_i$. For any $x > 0$, we have*

$$\mathbb{P}(S_n \geq x) \leq \exp\left(-\frac{x^2}{2(vn + Mx)}\right).$$

Proof. For any $\lambda > 0$, we get from the Markov inequality,

$$\mathbb{P}(S_n \geq x) \leq e^{-\lambda x}\mathbb{E}\left(e^{\lambda S_n}\right) = e^{-\lambda x}\left[\mathbb{E}\left(e^{\lambda Y_1}\right)\right]^n.$$

Using our notations, we will use the following bound. For any integer $k \geq 2$,

$$\mathbb{E}\left(Y_1^k\right) \leq vM^{k-2}.$$

Now let $\lambda < M^{-1}$. Using the Taylor expansion of the exponential function and the fact that $\mathbb{E}(X_1) = 0$, we deduce the following upper-bounds.

$$\begin{aligned}
\mathbb{E}\left(e^{\lambda Y_1}\right) &= \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}\left(Y_1^k\right) \\
&\leq 1 + \sum_{k\geq 2}(\lambda M)^k \frac{v}{2M^2} \\
&= 1 + \frac{v\lambda^2}{1-\lambda M}.
\end{aligned}$$

Using the inequality $(1+x)^n \leq \exp(nx)$, we get

$$\mathbb{P}(S_n \geq x) \leq e^{-\lambda x}\left(1 + \frac{v\lambda^2}{1-\lambda M}\right) \leq e^{-\lambda x + \frac{nv\lambda^2}{2(1-\lambda M)}}.$$

We next minimize the mapping

$$f : \lambda \mapsto -\lambda x + \frac{nv\lambda^2}{2(1-\lambda M)}.$$

The derivative vanishes at points

$$\lambda_{\pm} = \frac{1}{M}\left[1 \pm \frac{1}{\sqrt{1 + \frac{2mx}{nv}}}\right].$$

Only the root λ_- is smaller than $1/M$. Moreover, using the inequality $\sqrt{1+x} \leq 1 + \frac{1}{2}x$ for $x \geq 0$, we get

$$\lambda_- \leq \widetilde{\lambda} := \frac{1}{M}\left[1 - \frac{1}{1 + \frac{Mx}{nv}}\right].$$

Taking $\lambda = \widetilde{\lambda}$ instead of λ_-, we get $\exp(f(\lambda)) = -\frac{x^2}{2(nv+Mx)}$, which ends the proof.□

For the deviation of the empirical process, we deduce the following result. For a measurable mapping $g : \mathbb{R}^k \to \mathbb{R}$, we denote by $\|g\|_\infty$ the infinite norm of g.

Corollary 3. *Let f be a bounded measurable function. We have for any $x > 0$,*

$$\mathbb{P}(|\mathbb{G}_n f| \geq x) \leq 2\exp\left(-\frac{x^2}{2\left(Pf^2 + 2x\|f\|_\infty/\sqrt{n}\right)}\right).$$

Proof of Corollary 3. Use the bounds

$$\mathbb{P}(|\mathbb{G}_n f| \geq x) \leq \mathbb{P}(\mathbb{G}_n f \geq x) + \mathbb{P}(\mathbb{G}_n(-f) \geq x)$$

and apply the Bernstein inequality to $Y_i = \frac{g(X_i) - Pg}{\sqrt{n}}$ for $g = \pm f$ which is bounded by $2\|f\|_\infty/\sqrt{n}$.□

We next use the previous result to bound the expectation of the suprema of empirical processes. We first consider

$$\|G_n\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |G_n f|$$

when the family $\mathcal{F}$ is finite, i.e. $|\mathcal{F}| < \infty$.

Lemma 9. *Let $\mathcal{F}$ be a finite class of measurable and bounded functions. There then exists $C > 0$, not depending on $\mathcal{F}$, n and P such that*

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}} \leq C\left\{\frac{\max_{f\in\mathcal{F}}\|f\|_\infty}{\sqrt{n}}\log(1 + |\mathcal{F}|) + \max_{f\in\mathcal{F}}\sqrt{Pf^2}\sqrt{\log(1+|\mathcal{F}|)}\right\}.$$

Proof of Lemma 9. Let $f \in \mathcal{F}$ and set $a = 4\|f\|_\infty/\sqrt{n}$ and $b = 2Pf^2$. Define $A_f = \mathbb{G}_n f \mathbb{1}_{|\mathbb{G}_n f| > b/a}$ and $B_f = \mathbb{G}_n f \mathbb{1}_{|\mathbb{G}_n f| \leq b/a}$. From Corollary 3, we get for $x > 0$,

$$\begin{aligned}
\mathbb{P}(|A_f| > x) &\leq \mathbb{P}(|\mathbb{G}_n f| > \max(x, b/a)) \\
&\leq 2\exp\left(-\frac{\max(x,b/a)^2}{b + a\max(x,b/a)}\right) \\
&\leq 2\exp\left(-\frac{\max(x,b/a)}{2a}\right) \leq 2\exp\left(-\frac{x}{2a}\right)
\end{aligned}$$

and

$$\mathbb{P}(|B_f| > x) \leq \mathbb{P}(|\mathbb{G}_n(f)| > x)\,\mathbb{1}_{x \leq b/a} \leq 2\exp\left(-\frac{x^2}{b+ax}\right)\mathbb{1}_{x\leq b/a} \leq 2\exp\left(-\frac{x^2}{2b}\right).$$

Next setting for $p = 1, 2$, $\phi_p(x) = \exp(x^p) - 1$, we deduce that

$$\mathbb{E}\phi_1\left(\frac{|A_f|}{4a}\right) = \mathbb{E}\int_0^{|A_f|/4a} e^x dx = \int_0^\infty \mathbb{P}\left(|A_f| > 4xa\right) e^x dx \leq 1$$

and similarly $\mathbb{E}\phi_2\left(\frac{|B_f|}{\sqrt{6b}}\right) \leq 1$. Since ϕ_p is convex and non-negative, we get from Jensen's inequality,

$$\phi_1\left(\mathbb{E}\max_{f\in\mathcal{F}}\frac{|A_f}{4a}\right) \leq \mathbb{E}\phi_1\left(\max_{f\in\mathcal{F}}\frac{|A_f}{4a}\right) \leq \mathbb{E}\sum_{f\in\mathcal{F}}\phi_1\left(\frac{|A_f|}{4a}\right) \leq |\mathcal{F}|\,.$$

Similarly,

$$\phi_2\left(\mathbb{E}\max_{f\in\mathcal{F}}\frac{|B_f}{\sqrt{6b}}\right) \leq |\mathcal{F}|.$$

The proof follows by applying the triangular inequality

$$\begin{aligned}\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}} &\leq \mathbb{E}\sup_{f\in\mathcal{F}}|A_f| + \mathbb{E}\sup_{f\in\mathcal{F}}|B_f| \\ &\leq 4a\mathbb{E}\sup_{f\in\mathcal{F}}\frac{|A_f|}{4a} + \sqrt{6b}\mathbb{E}\max_{f\in\mathcal{F}}\frac{|B_f|}{\sqrt{6b}}.\end{aligned}$$

and the inverse of the mapping ϕ_p to the previous inequalities.□

We now consider an arbitrary class $\mathcal{F}$ possessing an envelope function, i.e. there exists a function $F : \mathbb{R}^k \to \mathbb{R}$ such that $\sup_{f\in\mathcal{F}}|f(x)| \leq F(x)$ for $x \in \mathbb{R}^d$. It is possible as soon a $\sup_{f\in\mathcal{F}}|f(x)| < \infty$ for any $x \in \mathbb{R}^k$.

Lemma 10. *Let $\mathcal{F}$ be a class of measurable functions $f : \mathbb{R}^k \to \mathbb{R}$ with envelope F and such that $\sup_{f\in\mathcal{F}} Pf^2 \leq \delta^2$. Set $a(\delta) = \delta/\sqrt{\log N_{[]}\left(\delta, \mathcal{F}, L_2(P)\right)}$. There then exists $\widetilde{C} > 0$, not depending on $\mathcal{F}$, n and P, such that*

$$\mathbb{E}^*\|\mathbb{G}_n\|_{\mathcal{F}} \leq \widetilde{C}\left\{J_{[]}\left(\delta, \mathcal{F}, L_2(P)\right) + \sqrt{n}P^*F\mathbb{1}_{F>\sqrt{n}a(\delta)}\right\}.$$

Proof of Lemma 10. For technical reasons, we assume that $\delta \leq 1/8$. If we prove the lemma with such a δ, one can always apply the bound to the pair $(\delta, \mathcal{F}/\alpha)$ (for a given $\alpha > 1$) to get the bound for the pair $(\alpha\delta, \mathcal{F})$.

Since $|\mathbb{G}_n f| \leq \sqrt{n}(\mathbb{P}_n F + PF)$, we have

$$\mathbb{E}^*\sup_{f\in\mathcal{F}}|\mathbb{G}_n f|\mathbb{1}_{F>\sqrt{n}a(\delta)} \leq 2\sqrt{n}PF\mathbb{1}_{F>\sqrt{n}a(\delta)}.$$

It then remains to bound $p_n := \mathbb{E}^*\sup_{f\in\mathcal{F}}|\mathbb{G}_n f|\mathbb{1}_{F\leq\sqrt{n}a(\delta)}$. If $\mathcal{F}_n = \left\{f\mathbb{1}_{F\leq\sqrt{n}a(\delta)}\right\}$, we have $N_{[]}\left(\delta, \mathcal{F}_n, L_2(P)\right) \leq N_{[]}\left(\delta, \mathcal{F}, L_2(P)\right)$. For simplicity, we will identify $\mathcal{F}_n$ and $\mathcal{F}$ and assume that all the elements of $\mathcal{F}$ are bounded by $\sqrt{n}a(\delta)$.

Next, again for technical reasons that will appear latter in the proof, we fix a positive integer q_0 such that $4\delta \leq 2^{-q_0} \leq 8\delta$. This is possible because $\delta \leq 1/8$. We have the lower bound

$$\begin{aligned} J_{[]}\left(\delta, \mathcal{F}, L_2(P)\right) &\geq \sum_{q \geq q_0+3} \int_{2^{-q-1}}^{2^{-q}} \sqrt{N_{[]}\left(s, \mathcal{F}, L_2(P)\right)} ds & (3.2) \\ &\geq \frac{1}{2} \sum_{q \geq q_0+3} 2^{-q} \sqrt{\log N_q} & (3.3) \\ &\geq c \sum_{q \geq q_0} 2^{-q} \sqrt{\log N_q}, & (3.4) \end{aligned}$$

where $N_q = N_{[]}\left(2^{-q}, \mathcal{F}, L_2(P)\right)$ for any non-negative integer q and c is a universal constant. We have used that $N_q \leq N_{q+1}$. Now if $I_{q,i} := [u_{q,i}, v_{q,i}]$, $1 \leq \ell \leq N_q$, is a covering of $\mathcal{F}$ such that $P(v_{q,i} - u_{q,i})^2 < 2^{-2q}$, we set $\Delta_{qi} = v_{q,i} - u_{q,i}$. Replacing $I_{q,i}$ by $\mathcal{F}_{q1} = I_{q,1}$ and $\mathcal{F}_{qi} = I_{q,i} \setminus \cup_{j=1}^{i-1} I_{q,j}$ for $i = 2, \ldots, N_q$. We then get a partition $\mathcal{P}_q = \{\mathcal{F}_{qi} : 1 \leq i \leq N_q\}$ of $\mathcal{F}$ for all $q \geq q_0$.

Without loss of generality, we assume that the partitions are nested, i.e. we will assume that $\mathcal{P}_{q+1}$ is a refinement of $\mathcal{P}_q$ which means that for $i = 1, \ldots, N_{q+1}$, there exists $j \in \{1, \ldots, N_q\}$ such that $\mathcal{F}_{(q+1)i} \subset \mathcal{F}_{qj}$. If it is not the case, at each stage $q \geq 0$, one can replace $\mathcal{F}_{qi}$ by all the intersections of the form $\mathcal{F}_{qi} \cap \mathcal{F}_{(q-1)j}$ for all possible values of j. This operation will give a partition of cardinal at most $\overline{N}_q = N_{q_0} \cdots N_q$ at stage q, instead of N_q. However, we note that

$$\sum_{q \geq q_0} 2^{-q} \sqrt{\log \overline{N}_q} \leq \sum_{q=q_0}^{\infty} 2^{-q} \sum_{p=q_0}^{q} \sqrt{\log N_p} = \sum_{p=q_0}^{\infty} \sqrt{\log N_p} 2^{-p} = 2 \sum_{p \geq q_0} 2^{-p} \sqrt{\log N_p}.$$

We then conclude that there exists a sequence of nested partitions $\mathcal{P}_q = \left\{\mathcal{F}_{qi} : 1 \leq i \leq \overline{N}_q\right\}$, $q \geq q_0$, such that for a universal constant $\overline{c}$,

$$J_{[]}\left(\delta, \mathcal{F}, L_2(P)\right) \geq \overline{c} \sum_{q \geq q_0} 2^{-q} \sqrt{\log \overline{N}_q}, \qquad \sup_{f,g \in \mathcal{F}_{qi}} |f - g| \leq \Delta_{qi}, \quad P\Delta_{qi}^2 < 2^{-2q}. \tag{3.5}$$

It now remains to bound $\mathbb{E}^* \|\mathbb{G}_n\|_{\mathcal{F}}$ by $\sum_{q \geq q_0} 2^{-q} \sqrt{\log \overline{N}_q}$ up to a constant. To this end for any $q \geq q_0$ and $1 \leq i \leq \overline{N}_q$, we consider an arbitrary element $f_{qi} \in \mathcal{F}_{qi}$ and we set

$$\pi_q f = f_{qi}, \quad \Delta_q f = \Delta_{qi} \text{ if } f \in \mathcal{F}_{qi}.$$

The principle will be to use an argument called *chaining*. In order to apply Lemma 9, we will introduce the differences $D_{q+1} f := \pi_{q+1} f - \pi_q f$ for approximating f. Note that $\{D_q f : f \in \mathcal{F}\}$ is a finite set but $|D_{q+1} f| \leq \Delta_q f$ and $\Delta_q f$ is not necessarily a bounded function. This is why we will only consider the differences $D_{q+1} f(x)$ for x in the event

$$A_q f = \left\{\Delta_{q_0} f \leq \sqrt{n} a_{q_0}, \ldots, \Delta_q f \leq \sqrt{n} a_q\right\},$$

where $a_q = 2^{-q}/\sqrt{\log \overline{N}_{q+1}}$ are chosen to get the desired upper bound from Lemma 9. We will then get

$$f(x) - \pi_{q_0} f(x) = \sum_{q \geq q_0+1} D_q f \mathbb{1}_{A_{q-1}f} + f(x) - \pi_{q_1(x)} f(x),$$

where $q_1(x)$ is the first index p (possibly infinite) for which $x \in B_p f = A_{p-1} f \cap \{\Delta_p f > \sqrt{n} a_p\}$. But note that either $x \in \cap_{q \geq q_0} A_q f$ and then $x \notin B_q f$ for all $q \geq q_0 + 1$ or there exists a unique integer $p = q_1(x)$ such that $x \in B_p f$ and $x \notin B_q f$ for $q \neq p$ and $x \notin A_q f$ for $q \geq p$. Note also that from our choice of q_0, we have $2a(\delta) \leq a_{q_0}$ and then $x \in A_{q_0} f$.

This allows to write the decomposition

$$f - \pi_{q_0} f = \sum_{q \geq q_0+1} D_q f \mathbb{1}_{A_{q-1}f} + \sum_{q \geq q_0} (f - \pi_q f) \mathbb{1}_{B_q f}.$$

We then get

$$\begin{aligned} \|\mathbb{G}_n\|_{\mathcal{F}} &\leq \sup_{f \in \mathcal{F}} |\mathbb{G}_n \pi_{q_0} f| + \sum_{q \geq q_0+1} \sup_{f \in \mathcal{F}} \left|\mathbb{G}_n D_q f \mathbb{1}_{A_{q-1}f}\right| + \sum_{q \geq q_0+1} \sup_{f \in \mathcal{F}} \left|\mathbb{G}_n \left(f - \pi_q f\right) \mathbb{1}_{B_q f}\right| \\ &:= U_1 + U_2 + U_3. \end{aligned}$$

- For U_1, we apply Lemma 9, noticing that $|\pi_{q_0} f| \leq \sqrt{n} a_{q_0}$ and $P\left(|\pi_{q_0} f|^2\right) < \delta$ by assumption. We get

$$U_1 \leq C \left\{ a_{q_0} \log(1 + \overline{N}_{q_0}) + \delta \sqrt{\log(1 + \overline{N}_{q_0})} \right\}.$$

  Since $\delta \leq 2^{-q_0-2}$ and using the definition of a_q, the right-hand side in the previous inequality can be clearly bounded by $\sum_{q \geq q_0} 2^{-q} \sqrt{\log N_q}$ up to a universal constant.

- For the second term U_2, we note that $|D_q f| \leq \Delta_{q-1} \leq \sqrt{n} a_{q-1}$ on the set $A_{q-1} f$ and $P\left(|D_q f|^2\right) \leq 2^{-q+1}$ by the definition of our nested partitions. Moreover there are at most $\overline{N}_q$ functions $D_q f$ and at most $\overline{N}_{q-1}$ indicator functions $\mathbb{1}_{A_q f}$. The number of functions is then bounded here by $\overline{N}_q \times \overline{N}_{q-1} \leq \overline{N}_q^2$. Lemma 9 leads to

$$U_2 \leq C \left\{ \sum_{q \geq q_0+1} a_{q-1} \log(1 + \overline{N}_q^2) + \sum_{q \geq q_0+1} 2^{-q+1} \sqrt{\log(1 + \overline{N}_q^2)} \right\}.$$

  Once again, U_2 can be bounded by $\sum_{q \geq q_0} 2^{-q} \sqrt{\log \overline{N}_q}$ up to a universal constant.

- Finally, we bound U_3. Since our partitions are nested, we have $|f - \pi_q f| \leq \Delta_q f \leq \Delta_{q-1} f$ which is bounded by $\sqrt{n} a_{q-1}$ on the event $B_q f$. Moreover $P(f - \pi_q f)^2 \leq P(\Delta_{q-1} f)^2 \leq 2^{-2(q-1)}$ and the number of functions in the supremum is at most $\overline{N}_q^2$, as in the previous case, we obtain the same bound as for U_2, which completes the proof.□

3.4 Two applications of empirical process theory

3.4.1 Goodness-of-Fit Statistics

Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in $\mathbb{R}$. Our aim is to test if the data are generated from a probability distribution P contained in a specific set of probability measures. In what follows, for a function $f : \mathbb{R} \to \mathbb{R}$, we denote by $\|f\|_\infty := \sup_{t\in\mathbb{R}} |f(t)|$ its infinite norm.

We start with the case of a single probability measure. More precisely, our aim is to test H_0: $P = P_0$ versus H_1: $P \neq P_0$ where P_0 is a prescribed probability measure. We denote by F_0 the cumulative distribution function (cdf) of X_1. Two popular statistics are $S_1 := \sqrt{n}\|\mathbb{F}_n - F_0\|_\infty$ (Kolmogorov-Smirnov) and $S_2 = n\int (\mathbb{F}_n - F_0)^2 dF_0$ (Cramér-von Mises).

Theorem 21. *We have $S_1 \hookrightarrow \|\mathbb{G}_{F_0}\|_\infty$ and $S_2 \hookrightarrow \int \mathbb{G}_{F_0}^2 dF_0$ where $\mathbb{G}_{F_0}$ is a Gaussian process, with mean 0 and covariance $Cov(\mathbb{G}_{F_0}(s), \mathbb{G}_{F_0}(t)) = F_0(\min(s,t)) - F_0(s)F_0(t)$.*

Proof of Theorem 21. The two mappings $z \to \|z\|_\infty$ and $z \mapsto \int z^2 dF_0$, defined on $\ell^\infty(\mathbb{R})$ are continuous and the result follows from Corollary 2 and the continuous mapping theorem.□

Notes

1. Suppose that F_0 is continuous. In this case, one can show that the limiting distributions of the two statistics S_1 and S_2 do not depend on F_0. Here are two arguments.

   - We have the representation $\mathbb{G}_F = P_F$ where $\{U_t : t \in [0,1]\}$ is a Brownian bridge. In this case $\|\mathbb{G}_{F_0}\|_\infty = \|U\|_\infty$ does not depend on F_0. It is also possible to show that $\int_{-\infty}^{\infty} U_{F_0(x)} dF_0(x) = \int_0^1 U_t dt$. This equality is clear when F_0 is continuously differentiable (in this case $x \mapsto \int_0^{F_0(x)} U_t dt$ is a primitive of $x \mapsto U_{F_0(x)} F_0'(x)$), but it can be also generalized to any continuous cdf F_0.
   - One can also show directly that the distribution of S_1 and S_2 do not depend on F_0. To this end, one can use the generalized inverse of the cdf F_0, i.e.

   $$F_0^{-1}(u) = \inf\{x \in \mathbb{R} : F_0(x) \geq u\}, \quad u \in (0,1).$$

   and the representation $X_i = F_0^{-1}(U_i)$ where $U_1, \ldots, U_n$ are i.i.d. random variables uniformly distributed over $[0,1]$. The fundamental equivalence

   $$F_0^{-1}(u) \leq x \Leftrightarrow u \leq F_0(x)$$

   can be used. We then get $\mathbb{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{U_i \leq F_0(x)}$ and the distribution of $S_1 = \|\mathbb{G}_n\|_\infty$ or $S_2 = \int \mathbb{G}_n^2 dF_0$ are the same as for uniformly distributed random variables. One can then simulate approximately the quantiles of S_1 and

S_2 (this requires the simulations of several samples of n variables uniformly distributed, to use them to compute several realizations of S_i $(i = 1, 2)$ and then the associated empirical distribution).

2. When F_0 is continuous, the limiting distributions of S_1 and S_2 are respectively $\|U\|_\infty$ and $\int_0^1 U_t dt$, where $\{U_t : 0 \leq t \leq 1\}$ is a Brownian bridge. The probability distributions of these two random variables are tabulated. Additionally, it is possible to derive the following expression.

$$\mathbb{P}\left(\|U\|_\infty > x\right) = 2\sum_{j=1}^{\infty}(-1)^{j+1}\exp\left(-2j^2x^2\right), \quad x > 0.$$

For testing H_0 versus H_1, we reject the null hypothesis at level α for large values of S_1 (or S_2) using the quantile of order $1-\alpha$ obtained either from these limiting distributions or from the simulation procedure given in the previous point.

For adequation tests, it is often more relevant to test adequacy with respect to a family of probability distributions $\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta\}$, for instance a parametric distribution such as the Gaussian, $P_\theta = \mathcal{N}(\theta_1, \theta_2)$ for $\theta = (\theta_1, \theta_2) \in \mathbb{R} \times \mathbb{R}_+^*$. Here we suppose that Θ is a subset of $\mathbb{R}^d$. Suppose that we want to test H_0: $P \in \mathcal{P}_\Theta$ vs H_1: $P \notin \mathcal{P}_\Theta$. If we have an estimator $\hat{\theta}_n$ for the true parameter θ_0, under H_0, we have the following decomposition

$$\mathbb{P}_n - P_{\hat{\theta}} = \mathbb{P}_n - P_{\theta_0} - \left(P_{\hat{\theta}_n} - P_{\theta_0}\right) \approx \mathbb{P}_n - P_{\theta_0} - \dot{P}_{\theta_0}\left(\hat{\theta}_n - \theta_0\right),$$

where $\dot{P}_{\theta_0}$ denotes the derivative of $\theta \mapsto P_\theta$ at point θ_0 (for a topology on the set of probability measures to precise). We then observe two kinds of fluctuation, one for the empirical process and another one coming from the estimation error. To derive the asymptotic distributions of the previous statistics in this context, we have to study the limiting behavior of $\sqrt{n}\left(\mathbb{P}_n - P_{\hat{\theta}}\right)$. The two following assumptions will be used when H_0 is considered to be valid.

H1 There exists a measurable mapping $\psi_{\theta_0} : \mathbb{R} \to \mathbb{R}^d$ with $\mathbb{E}\left[\psi_{\theta_0}(X_1)\right] = 0$ and $\mathbb{E}\left[\|\psi_{\theta_0}(X_1)\|^2\right] < \infty$ and such that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\theta_0}(X_i) + o_{\mathbb{P}}(1).$$

H2 The mapping $\theta \mapsto P_\theta$ is differentiable at point θ_0, as an application from Θ to $\ell^\infty(\mathcal{F})$, where $\mathcal{F}$ is a class of $P_{\theta_0}-$square integrable functions and such that the assumptions of Theorem 18 are satisfied for $Z_n = \left(\mathbb{G}_n f\right)_{f\in\mathcal{F}}$ and $T = \mathcal{F}$. We denote by $\dot{P}_{\theta_0}$ this derivative. Here $\mathbb{G}_n f = \sqrt{n}\left(\mathbb{P}_n f - P_{\theta_0} f\right)$.

Theorem 22. *Suppose that Assumptions* **H1-H2** *are valid and the hypothesis H_0 holds true. Then*

$$\sqrt{n}\left(\mathbb{P}_n - P_{\hat{\theta}_n}\right) \hookrightarrow \left(\mathbb{G}_{P_{\theta_0}} f - \mathbb{G}_{P_{\theta_0}} \psi_{\theta_0}^T \dot{P}_{\theta_0} f\right)_{f\in\mathcal{F}}$$

in $\ell^\infty(\mathcal{F})$.

Notes

1. Under H_0, we note that the asymptotic distribution obtained Theorem 18 depends on the parametric model as well as on the estimator used.

2. Assumption **H1** is valid for MLE for instance, under the regularity assumptions discussed in the previous chapter.

3. The assumption that $(Z_n, T)) = (\mathbb{G}_n, \mathcal{F})$ satisfies the assumptions of Theorem 18 can be weakened in $\mathcal{F}$ is a Donsker class. As shown in the proof of Theorem 22, we have to use convergence of the empirical process for the class $\mathcal{G} = \mathcal{F} \cup \{\psi_{\theta_0}\}$ and it is possible to show that the union of two Donsker classes is still Donsker. See Section 2.10.2 in Vaart and Wellner (2023). However, it is more direct to show that $\mathcal{G}$ still satisfies the two assumptions of Theorem 18, provided that $\mathcal{F}$ also satisifies them. When these two assumptions are satisfied for a family of square integrable functions, it is straightforward to show that this family is a Donsker class.

4. When $\mathcal{F}$ denotes the class of indicator functions, one can obtain convergence for the corresponding Kolmogorov-Smirnov type statistics $S_{1,\Theta} := \sqrt{n}\|\mathbb{F}_n - F_{\hat{\theta}}\|_\infty$. The existence of a derivative with a uniform convergence in **H2** has to be shown model by model. Of course if $P_\theta f = \int f p_\theta d\mu$, the natural candidate for $\dot{P}_{\theta_0}$ is the mapping $f \mapsto \int f \dot{p}_{\theta_0} d\mu$.

Proof of Theorem 22. From **H2**, we have

$$\|P_{\theta_0+h} - P_{\theta_0} - h^T \dot{P}_{\theta_0}\|_{\mathcal{F}} = o\left(\|h\|\right).$$

Using **H1**, this yields to

$$\sqrt{n}\|P_{\hat{\theta}_n} - P_{\theta_0} - (\hat{\theta}_n - \theta_0)^T \dot{P}_{\theta_0}\|_{\mathcal{F}} = \sqrt{n}\|\hat{\theta}_n - \theta_0\| o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1).$$

We then get

$$\begin{aligned}
\sqrt{n}\left(\mathbb{P}_n - P_{\hat{\theta}_n}\right) &= \sqrt{n}\left(\mathbb{P}_n - P_{\theta_0}\right) - \sqrt{n}\left(P_{\hat{\theta}_n} - P_{\theta_0}\right) \\
&= \sqrt{n}\left(\mathbb{P}_n - P_{\theta_0}\right) - \sqrt{n}\left(\hat{\theta}_n - \theta_0\right)^T \dot{P}_{\theta_0} + o_{\mathbb{P}}(1) \\
&= \sqrt{n}\left(\mathbb{P}_n - P_{\theta_0}\right) - \sqrt{n}\mathbb{P}_n \psi_{\theta_0}^T \dot{P}_{\theta_0} + o_{\mathbb{P}}(1).
\end{aligned}$$

Setting $\mathcal{G} = \mathcal{F} \cup \{\psi_{\theta_0}\}$ and

$$\phi_{\theta_0}\left(\sqrt{n}\left(\mathbb{P}_n g - P_{\theta_0} g\right)_{g\in\mathcal{G}}\right) = \sqrt{n}\left(\mathbb{P}_n - P_{\theta_0}\right)_{f\in\mathcal{F}} - \sqrt{n}\mathbb{P}_n \psi_{\theta_0}^T \dot{P}_{\theta_0},$$

the mapping ϕ_{θ_0} is continuous and $\mathcal{G}$ also satisfies the assumptions of Theorem 18 (for the second assumption, just complete the partition of $\mathcal{F}$ by the singleton $\{\psi_{\theta_0}\}$). $\mathcal{G}$ is then $P_{\theta_0}-$Donsker. We then conclude using the continuous mapping theorem and Slutsky's lemma.□

3.4.2 High-dimensional regression

Here we consider the problem of prediction of a random variable Y given some predictors $X = \left(X^{(1)}, \ldots, X^{(p_n)}\right)^T$ when $p_n \to \infty$ as $n \to \infty$. Then X depends on n but for simplicity, we omit this dependence. Suppose that we have a sample $\{(X_i, Y_i) : 1 \leq i \leq n\}$ with i.i.d. random vectors distributed as (X, Y). Our aim is to study LASSO type estimators, i.e.

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta_n} M_n(\theta), \quad M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - X_i^T \theta\right)^2,$$

where $\Theta_n = \{\theta \in \mathbb{R}^{p_n} : \|\theta\|_1 := \sum_{i=1}^{p_n} |\theta_i| \leq R_n\}$ for some values of $R_n > 0$. We assume that p_n is equivalent to Cn^α at infinity for some positive α and C.

Our aim is to consider the theoretical risk $\theta \mapsto \overline{M}_n(\theta) := \mathbb{E}\left[(Y - X^T\theta)^2\right]$ and to find conditions under which

$$\overline{M}_n\left(\hat{\theta}_n\right) - \overline{M}_n(\theta_n) = o_{\mathbb{P}}(1), \tag{3.6}$$

when $\theta_n = \arg\min_{\theta \in \Theta_n} \overline{M}_n(\theta)$. When (3.6) is valid, we say that $\hat{\theta}_n$ is Θ_n-persistent. It means that the theoretical risk evaluated at $\hat{\theta}_n$ is asymptotically closed to the optimal risk.

Theorem 23. *For $Z_i = (Y_i, X_i)$, $1 \leq i \leq n$, set $F_n(Z_i) = \max_{1 \leq j,k \leq p_n+1} |Z_{j,i}Z_{k,i} - \mathbb{E}(Z_{j,i}Z_{k,i})|$ satisfies $M := \sup_{n \geq 1} \mathbb{E}[F_n(Z_1)^2] < \infty$. Then for $R_n = o\left(\left(\frac{n}{\log n}\right)^{1/4}\right)$, $\hat{\theta}_n$ is Θ_n-persistent.*

Note. We do not assume the existence a correctly specified linear model for this result. In particular, the Z_i's can be any i.i.d. random vectors of dimension $p_n + 1$ satisfying the required moment assumptions.

Proof of Theorem 23. We have the inequalities

$$\begin{aligned} 0 &\leq \overline{M}_n\left(\hat{\theta}_n\right) - \overline{M}_n(\theta_n) \\ &= \overline{M}_n\left(\hat{\theta}_n\right) - M_n\left(\hat{\theta}_n\right) + M_n\left(\hat{\theta}_n\right) - M_n(\theta_n) + M_n(\theta_n) - \overline{M}_n(\theta_n) \\ &\leq 2 \sup_{\theta \in \Theta_n} \left|M_n(\theta) - \overline{M}_n(\theta)\right|. \end{aligned}$$

Setting

$$\Sigma_n = \left(\frac{1}{n} \sum_{i=1}^{n} Z_{j,i}Z_{k,i}\right)_{1 \leq j,k \leq p_n+1}, \quad \overline{\Sigma}_n = \left(\mathbb{E}(Z_{j,1}Z_{k,1})\right)_{1 \leq j,k \leq p_n+1}$$

and $\gamma = \begin{pmatrix} -1 \\ \theta \end{pmatrix}$, we have

$$\begin{aligned} \left|M_n(\theta) - \overline{M}_n(\theta)\right| &= \left|\gamma^T\left(\Sigma_n - \overline{\Sigma}_n\right)\gamma\right| \\ &\leq \|\Sigma_n - \overline{\Sigma}_n\|_\infty \|\gamma\|_1^2 \\ &\leq \|\Sigma_n - \overline{\Sigma}_n\|_\infty (1 + R_n)^2, \end{aligned}$$

where for any matrix A, $\|A\|_\infty$ denotes the maximum of the absolute values of the entries of A. Using Markov's inequality, we get for $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\overline{M}_n\left(\hat{\theta}_n\right) - \overline{M}_n(\theta_n) > \epsilon\right) &\leq \mathbb{P}\left(2(R_n+1)^2\|\Sigma_n - \overline{\Sigma}_n\|_\infty > \epsilon\right) \\ &\leq \frac{2(R_n+1)^2}{\epsilon}\mathbb{E}\left[\|\Sigma_n - \overline{\Sigma}_n\|_\infty\right].\end{aligned}$$

Now take $\mathcal{F} = \mathcal{F}_n = \{f_{j,k} : 0 \leq j,k \leq p_n\}$ with

$$f_{j,k}(z) = z_j z_k - \mathbb{E}\left(Z_{j,1}Z_{k,1}\right), \quad z \in \mathbb{R}^{p_n+1}.$$

Let $\mathbb{G}_n$ be the empirical process for $Z_1, \ldots, Z_n$ and the class of function $\mathcal{F}$ with enveloppe F_n. From Lemma 10, we have for the choice $\delta^2 = PF_n^2$,

$$\begin{aligned}\mathbb{E}\left[\sqrt{n}\|\Sigma_n - \overline{\Sigma}_n\|_\infty\right] &= \mathbb{E}\left[\|\mathbb{G}_n\|_{\mathcal{F}}\right] \\ &\leq \widetilde{C}\left\{J_{[]}\left(\delta, \mathcal{F}, L_2(P)\right) + \sqrt{n}PF_n\mathbb{1}_{F_n > \sqrt{n}a(\delta)}\right\} \\ &\leq \widetilde{C}\left\{J_{[]}\left(\delta, \mathcal{F}, L_2(P)\right) + a(\delta)^{-1}PF_n^2\right\} \\ &\leq \widetilde{C}\left\{J_{[]}\left(\sqrt{PF_n^2}, \mathcal{F}, L_2(P)\right) + \sqrt{\log N_{[]}\left(\sqrt{PF_n^2}, \mathcal{F}, L_2(P)\right)}\sqrt{PF_n^2}\right\} \\ &\leq 2\widetilde{C}J_{[]}\left(\sqrt{PF_n^2}, \mathcal{F}, L_2(P)\right).\end{aligned}$$

Since $\mathcal{F}_n$ is finite, we have $N_{[]}\left(s, \mathcal{F}_n, L_2(P)\right) \leq (1+p_n)^2$ for any $s > 0$. Since $\sup_{n\geq 1} PF_n^2 < \infty$, we deduce that $\mathbb{E}\left[\|\Sigma_n - \overline{\Sigma}_n\|_\infty\right]$ can be bounded, up to a positive constant, by $\sqrt{\log p_n}/\sqrt{n}$ which is negligible with respect to $(R_n+1)^2$. The proof is now complete.□

3.5 Appendix

3.5.1 Some complements in Topology and in measure theory

In this part, T denotes an arbitrary set.

Definition 3. *A mapping $\rho : T \times T \to \mathbb{R}_+$ is said to be a semimetric if*

1. *For every $(x,y) \in T \times T$, $\rho(x,y) = \rho(y,x)$.*
2. *For every $x \in T$, $\rho(x,x) = 0$.*
3. *For every $(x,y,z) \in T^3$,*
$$\rho(x,y) \leq \rho(x,z) + \rho(z,y).$$

When $\rho(x,y) = 0 \Rightarrow x = y$, a semimetric is called a metric and (T,ρ) a metric space. Otherwise, (T,ρ) is called a semimetric space.

Definition 4. *Let (T,ρ) be a semimetric space. For every $(x,r)\in T\times[0,\infty)$, $B_\rho(x,r)=\{y\in T:\rho(x,y)<r\}$ is called an open ball. A space T with a semimetric ρ is said to be totally bounded if for any $\epsilon>0$, T can be covered by finitely many open balls of radius ϵ. A subset $A\subset T$ is said totally bounded if for any $\epsilon>0$, A can be covered by finitely many open balls of radius ϵ.*

There are some important links between the notions of totally bounded, completeness and compactness. If (T,ρ) is a metric space, than it is compact if and only if it is complete and totally bounded. Moreover if (T,ρ) is a complete metric space, then $A\subset T$ is totally bounded if and only if its closure $\overline{A}$ is compact. The following space will be important in what follows.

Definition 5. *Let (T,ρ) a semimetric space. The space of uniformly continuous functions $f:T\to\mathbb{R}$ is denoted by $UC(T,\rho)$. We recall that*

$$f\in UC\,(T,\rho)\Longleftrightarrow\lim_{\delta\to 0}\sup_{(x,y)\in T^2:\rho(x,y)<\delta}|f(x)-f(y)|=0.$$

Definition-Proposition 1. *Let (T,ρ) be a semimetric space. We denote by $\ell^\infty(T)$ the space of bounded functions $f:T\to\mathbb{R}$ equipped with the uniform norm $\|f\|_T=\sup_{x\in T}|f(x)|$. If (T,ρ) is also totally bounded, we have $UC(T,\rho)\subset\ell^\infty(T)$. Moreover the space $UC(T,\rho)$, equipped with the uniform norm, is separable and complete.*

By separable, we mean that there exists a sequence $(g_n)_{n\in\mathbb{N}}$ of elements of $\mathrm{UC}(T,\rho)$ such that for any $f\in\mathrm{UC}\,(T,\rho)$, $\inf_{n\geq 0}\|f-g_n\|_T=0$. A totally bounded metric space is separable. The space $\ell^\infty(T)$ is not separable when T is not a finite set.

Proof Let $f\in\mathrm{UC}\,(T,\rho)$ and fix $\delta>0$ such that $|f(x)-f(y)|\leq 1$ when $\rho(x,y)\leq\delta$. Fix also $x_1,\ldots,x_k\in T$ such that $T\subset\cup_{i=1}^k B_\rho(x_i,\delta)$. Then if $\rho(x,x_i)<\delta$, $|f(x)-f(x_i)|\leq 1$. We deduce that

$$\|f\|_T\leq 1+\max_{1\leq i\leq k}|f(x_i)|$$

and then $f\in\ell^\infty(T)$.

Completeness of the space $\mathrm{UC}(T;\rho)$ follows from standard arguments already used for proving completeness of spaces of continuous functions.

Separability will not be proved here. It follows from the Stone-Weierstrass theorem with a proof analogue to prove separability of the space of continuous real-valued functions defined on a compact metric space. □

Definition 6. *A probability measure P on a metric space (G,d) is said to be tight if for any $\epsilon>0$, there exists a compact subset K_ϵ of G such that $P\,(G\setminus K_\epsilon)\leq\epsilon$.*

The previous definition means that a tight probability measure has a mass that concentrates on compact subsets. If $G=\mathbb{R}^k$ and d is a distance defined by an arbitrary norm on G, than every probability measure P is tight. This is a consequence of the

compactness property of closed balls. Indeed if $\overline{B}_d(0,r) = \{y \in G : d(y,0) \leq r\}$, we have $\lim_{r\to\infty} P\left(G \setminus \overline{B}_d(0,r)\right) = 0$ from the lower-continuity property of the measure (alternatively the dominated convergence theorem) and one can choose a compact subset $K_\epsilon = \overline{B}_d(0,r_\epsilon)$ for r_ϵ large enough. One can extend the tightness property (which is not automatic in spaces of infinite dimensions) to more general spaces.

Proposition 4. *If (G,d) is a separable and complete metric space, then every measure P is tight.*

Proof of Proposition 4 Let $\epsilon > 0$ and k be a positive integer. From separability, if $(g_i)_{i\in\mathbb{N}}$ is a dense subset of G and $A_{k,i} = B_d(g_i, 1/k)$, then $G = \cup_{i\in\mathbb{N}} A_{k,i}$. Let now $n_k = n_k(\epsilon)$ be an integer such that $P\left(\cup_{i=0}^{n_k} A_{k,i}\right) > 1 - \epsilon/2^k$. If K_ϵ denotes the closure of $\cap_{k\geq 1} \cup_{i=0}^{n_k} A_{k,i}$, which is a totally bounded set, we obtain a compact subset of G. Moreover

$$P\left(G \setminus K_\epsilon\right) \leq \sum_{k\geq 1} P\left(G \setminus \cup_{i=0}^{n_k} A_{k,i}\right) \leq \sum_{k\geq 1} \epsilon/2^k = \epsilon$$

and the proof is complete.□

3.5.2 Kolmogorov's extension theorem

We will need the following important theorem about the existence of a random process defined from a family of finite-dimensional distributions. Let $E = \mathbb{R}^{\mathbb{N}^*}$ be the set of real-valued sequences indexed by the set of positive integers. On E, we consider the sigma-field $\mathcal{E}$ generated by the cylinder set, i.e. the set C of the form

$$C = \{(x_n)_{n\geq 1} \in E : x_1 \in A_1, \ldots, x_k \in A_k\}$$

for a positive integer k and some Borel subsets of $\mathbb{R}$, $A_1, \ldots, A_k$. A proof of the following result can be found in Durrett (2019), Theorem $A.3.1$.

Theorem 24 (Kolmogorov's extension theorem)**.** *Assume that for each $n \geq 1$, π_n is a probability measure on $\mathbb{R}^n$ and such that for $A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R})$,*

$$\pi_n\left(A_1 \times \cdots \times A_n\right) = \pi_{n+1}\left(A_1 \times \cdots \times A_n \times \mathbb{R}\right) \tag{3.7}$$

There then exists a unique probability measure π on $(E,\mathcal{E})$ such that for every integer $n \geq 1$ and $A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R})$,

$$\pi\left((x_k)_{k\geq 1} \in E : x_1 \in A_1, \ldots, x_n \in A_n\right) = \pi_n\left(A_1 \times \cdots \times A_n\right).$$

To construct a random element taking values in E, one can use the canonical construction. We simply set $\Omega = E$, $\mathcal{A} = \mathcal{E}$ and $X_n(\omega) = \omega_n$ (the coordinate mapping) for every $n \geq 1$. One can note that the cylinder sigma-field $\mathcal{E}$ is the smallest sigma-field making the coordinate mappings measurable. The first application of this result is a rigorous construction of a sequence of i.i.d. random variables with marginal probability distribution μ (in this case, $\pi_n = \mu^{\otimes n}$, the product of measures).

3.5.3 Proof of Theorem 18

The aim of the proof is to construct a limiting process Z taking values un $\mathrm{UC}(T,\rho)$ for a suitable semimetric ρ making T totally bounded. From Proposition 3.7 and Proposition 4, the probability distribution of this process will be tight, as required in the statement of Theorem 18. To this end, we will first construct this process Z on a numerable set of indices T_0 and extend it to T by continuity. To construct T_0, for each positive integer m, we consider a partition $T_1^m,\dots,T_{k_m}^m$ such that the condition 2 is satisfied with $\eta=\epsilon=1/2^m$. Note that if $S_1,\dots,S_\ell$ is a refinement of a partition of a partition $T_1,\dots,T_k$, then

$$\max_{1\le j\le \ell}\sup_{s,t\in S_j}|Z_n(t)-Z_n(s)|\le \max_{1\le j\le k}\sup_{s,t\in T_j}|Z_n(t)-Z_n(s)|$$

and one can assume that T^{m+1} is a refinement of T^m (Intersect each T_j^{m+1} with all the T_i^m's). Next, take an arbitrary point t_j^m in T_j^m and set $T_0=\left\{t_j^m : 1\le j\le k_m, m\ge 1\right\}$. Note that T_0 can be enumerated with a sequence $\{s_i : i\ge 1\}$. For a positive integer k, let π_k be the limiting distribution of $(Z_n(s_1),\dots,Z_n(s_k))$. By Kolmogorov's extension theorem, there exists a stochastic process $(Z(t))_{t\in T_0}$ compatible with the π_k's. By the portmanteau lemma,

$$\mathbb{P}\left(\max_j\max_{s,t\in T_j^m\cap T_0}|Z(t)-Z(s)|>2^{-m}\right)\le \liminf_n \mathbb{P}\left(\max_j\max_{s,t\in T_j^m\cap T_0}|Z_n(t)-Z_n(s)|>2^{-m}\right)\le 2^{-m}.$$

Next, we define the metric ρ on T by

$$\rho(s,t)=\sum_{m\ge 1}2^{-m}\rho_m(s,t),\quad \rho_m(s,t)=\min_{1\le j\le k_m}\mathbb{1}_{(s,t)\notin T_j^m\times T_j^m}.$$

That is $\rho_m(s,t)=1$ if s and t are not in the same element of the partition $\left\{T_1^m,\dots,T_{k_m}^m\right\}$ and 0 otherwise and $\rho(s,t)=\sum_{m\ge m_0}2^{-m}$ where m_0 is the first integer for which s,t are located in two different elements of the partition T^m. ρ is a semimetric on T. Since the diameter of T_j^m is $\sum_{j>m}2^{-j}=2^{-m}$, T is totally bounded for ρ. Note also that T_0 is a $\rho-$dense subset of T. Moreover if $\rho(s,t)<2^{-m}$, s and t are necessarily both located in the same element T_j^m. We conclude that

$$\mathbb{P}\left(\max_{s,tT_0,\rho(s,t)<2^{-m}}|Z(t)-Z(s)|>2^{-m}\right)\le 2^{-m}.$$

From the Borel-Cantelli lemma, we conclude that for almost every ω, if m is large enough and $\rho(s,t)<2^{-m}$, then $|Z(t)_\omega-Z(s)_\omega|\le 2^{-m}$. This proves that the paths of $\{Z(t):t\in T_0\}$ are in $\mathrm{UC}(T_0,\rho)$. By the extension theorem of uniformly continuous functions on dense subsets, one can define $Z(t)$ for $t\in T$ and the paths of $\{Z(t):t\in T\}$ are still in $\mathrm{UC}(T,\rho)$.

To end the proof, we define $p_m:T\to T$ by $p_m(t)=t_j^m$ when $t\in T_j^m$. By the assumption 1., we have $Z_n\circ p_m\to Z\circ p_m$ in distribution in $\ell^\infty(T)$. This is because $p_m(T)$ is a finite set (as an exercise, check the required convergence carefully). Moreover $Z\circ p_m\to Z$ a.s. in $\ell^\infty(T)$ as $m\to\infty$. Indeed, by uniform continuity of the paths, $\|Z\circ p_m-Z\|_T=\sup_{t\in T_0}|Z\circ p_m(t)-Z(t)|$ is measurable and converges to 0. Almost sure convergence also

entails convergence in distribution. We then get

$$\begin{aligned} & \mathbb{E}^* h\left(Z_n\right)-\mathbb{E} h\left(Z\right) \\ = \; & \mathbb{E}^* h\left(Z_n\right)-\mathbb{E} h\left(Z_n \circ p_m\right)+\mathbb{E} h\left(Z_n \circ p_m\right)-\mathbb{E}\left(Z \circ p_m\right)+\mathbb{E} h\left(Z \circ p_m\right)-\mathbb{E} h\left(Z\right) \\ = \; & \mathbb{E}^* h\left(Z_n\right)-\mathbb{E} h\left(Z_n \circ p_m\right)+o(1). \end{aligned}$$

Moreover if $h : \ell^\infty(T) \to \mathbb{R}$ is $L-$Lipschitz and bounded, we have

$$\left|\mathbb{E} h\left(Z_n\right)-\mathbb{E} h\left(Z_n \circ p_m\right)\right| \leq L 2^{-m}+2\|h\|_\infty \mathbb{P}\left(\left\|Z_n-Z_n \circ p_m\right\|_T>2^{-m}\right)$$

and from our assumptions,

$$\limsup_n \mathbb{P}\left(\left\|Z_n-Z_n \circ p_m\right\|_T>2^{-m}\right)=\lim \sum_n \mathbb{P}\left(\max_j \sup_{s,t n T_j^m}\left\|Z_n(t)-Z_n(s)\right\|_T>2^{-m}\right) \leq 2^{-m}.$$

Since m can be arbitrarily big, this proves the weak convergence.□

3.5.4 Proof of Theorem 20.

To apply Theorem 18, we introduce the class $\mathcal{G}=\{f-g:(f, g) \in \mathcal{F}^2\}$. Note that if $\mathcal{F} \subset \cup_{\ell=1}^k[u_\ell, v_\ell]$, then $\mathcal{G} \subset \cup_{\ell=1}^k \cup_{\ell'=1}^k\left[u_\ell-v_{\ell'}, v_\ell-u_{\ell'}\right]$, meaning that

$$N_{[]}\left(\varepsilon, \mathcal{F}, L_2(P)\right)^2 \geq N_{[]}\left(2 \varepsilon, \mathcal{G}, L_2(P)\right),$$

which implies that $J_{[]}\left(1, \mathcal{G}, L_2(P)\right)<\infty$. Now let $\delta, \eta, \varepsilon>0$ and set $k=N_{[]}\left(\delta, \mathcal{F}, L_2(P)\right)$. Setting $\mathcal{F}_1=\left[u_1, v_1\right] \cap \mathcal{F}$ and for $2 \leq \ell \leq k$,

$$\mathcal{F}_\ell=\left(\left[u_\ell, v_\ell\right] \setminus \cup_{i=1}^{\ell-1}\left[u_i, v_i\right]\right) \cap \mathcal{F}.$$

Then $\left\{\mathcal{F}_1, \ldots, \mathcal{F}_k\right\}$ forms a partition of $\mathcal{F}$ and the diameter of each element of the partition is controlled by δ, for the $L_2(P)$ norm. Using Lemma 10, we get

$$\begin{aligned} \varepsilon \mathbb{P}^*\left(\max_{1 \leq j \leq k} \sup_{f, g \in \mathcal{F}_j}\left|\mathbb{G}_n(f-g)\right|>\varepsilon\right) & \leq \mathbb{E}^*\left[\max_{1 \leq j \leq k} \sup_{f, g \in \mathcal{F}_j}\left|\mathbb{G}_n(f-g)\right|\right] \\ & \leq \widetilde{C}\left\{J_{[]}\left(\delta, \mathcal{G}, L_2(P)\right)+\sqrt{n} P F \mathbb{1}_{F>a(\delta) \sqrt{n}}\right\} \\ & \leq \widetilde{C}\left\{J_{[]}\left(\delta, \mathcal{G}, L_2(P)\right)+a(\delta)^{-1} P F^2 \mathbb{1}_{F>a(\delta) \sqrt{n}}\right\}. \end{aligned}$$

Here we take

$$F=\sup_{f, g \in \mathcal{F}}|f-g| \leq 2 \sup_{f \in \mathcal{F}}|f| \leq 2 \max_{1 \leq \ell \leq k}\left\{u_\ell, v_\ell\right\}$$

which is square integrable. Choosing $\delta>0$ sufficiently small in such a way $J_{[]}\left(\delta, \mathcal{G}, L_2(P)\right) \leq \eta \varepsilon$, the second assumption of Lemma 10 is satisfied, since the second term in the last upper bound goes to 0 with n. This shows the validity of the second assumption of Theorem 18. The first assumption of finite-dimensional convergence is automatic.□

Chapter 4

Introduction to asymptotic theory for stationary sequences

The aim of this chapter is to introduce some basic notions useful for studying statistical inference of some parameters when the sample $X_1, \ldots, X_n$ is composed of identically distributed but not necessarily independent random variables. This situation arises in analyzing time series, i.e. a collection of random variables measuring the same phenomenon at different time points and for which past values will have an influence on the present or future values. We will restrict to random sequences called " stationary", a stronger notion than identically distributed random variables and which means that the finite-dimensional distributions of the sequence are invariant under time shift. For instance $X_1, X_2, \ldots, X_n$ have the same distribution, but $(X_1, X_2), (X_2, X_3), \ldots, (X_{n-1}, X_n)$ also have the same distribution, $(X_1, X_2, X_3), (X_2, X_3, X_4), \ldots$ are also identically distributed and so on.

Such stationarity notion is mostly meaningful when the time points are equidistant. Some examples concern the evolution of daily temperatures, daily stock prices, monthly unemployment... Stationarity is a very restrictive notion since for many applications, data exhibit a seasonal behavior, a trend evolution (e.g. on average, daily temperatures are not the same in summer or winter and also increased during the last 50 past years) or a random walk behavior (i.e. the $X_t - X_{t-1}$'s form a stationary sequence). There exist many techniques to transform the original data in a new sequence which will be approximately stationary (such transformations depend on the context). We will not study this step and assume here it is reasonable to model the observations with a stationary sequence.

Studying asymptotic statistics in this context requires to first generalize the law of large number and the central limit theorems when the data are not independent. This will be our primary goal. We will also present simple stationary models for which the theory applies.

4.1 Stationary processes indexed by $\mathbb{Z}$ and Bernoulli shifts

For defining stationary sequences of random variables, it is often more convenient to formulate the theory for double-sided sequences, i.e. a sequence of random variables $(X_t)_{t\in\mathbb{Z}}$

indexed by the set of positive and negative integers. This commodity will appear more clearly when we will study autoregressive processes. We first generalize Kolmogorov's extension theorem in this setup. The following result is a simple extension of Theorem 7 in Chapter 3. Let $E = \mathbb{R}^d$. On $E^{\mathbb{Z}}$ (the set of sequences indexed by $\mathbb{Z}$ and taking values in E), we consider the sigma-field $\mathcal{C}$ generated by the cylinder sets

$$C = \left\{(x_t)_{t\in\mathbb{Z}} \in E^{\mathbb{Z}} : x_{-n} \in A_{-n}, \ldots, x_n \in A_n\right\},$$

for any $n \in \mathbb{N}$ and Borel subsets $A_{-n}, A_{-n+1}, \ldots, A_n$ of E.

Theorem 25 (Kolmogorov's extension theorem)**.** *For each $n \in \mathbb{N}$, let μ_n be a probability measure on E^{2n+1}. We assume that for any $A_j \in \mathcal{B}(E)$, $-n \leq j \leq n$,*

$$\mu_{n+1}\left(E \times A_{-n} \times A_{-n+1} \times \cdots A_n \times E\right) = \mu_n\left(A_{-n} \times A_{-n+1} \times \cdots A_n\right).$$

There then exists a unique probability measure $\mu : \mathcal{C} \to [0,1]$ such that for any $n \in \mathbb{N}$ and $A_{-n}, A_{-n+1}, \ldots, A_n \in \mathcal{B}(E)$,

$$\mu\left(\left\{(x_t)_{t\in\mathbb{Z}} \in E^{\mathbb{Z}} : x_i \in A_i, -n \leq i \leq n\right\}\right) = \mu_n\left(A_{-n} \times A_{-n+1} \times \cdots A_n\right).$$

Starting from a family of probability measures μ_n, $n \in \mathbb{N}$, satisfying the compatibility conditions of the previous theorem, one can construct a compatible canonical stochastic process $(X_t)_{t\in\mathbb{Z}}$, defined on $\Omega = E^{\mathbb{Z}}$ by $X_t(\omega) = \omega_t$ for $t \in \mathbb{Z}$ and $\omega = (\omega_t)_{t\in\mathbb{Z}} \in E^{\mathbb{Z}}$.

Examples

1. When $\mu_n = \nu^{\otimes 2n+1}$ (the product measure on E^{2n+1} with marginal ν), one can then define a sequence $(X_t)_{t\in\mathbb{Z}}$ of i.i.d. random variables with common distribution ν.

2. Suppose that we have a homogeneous Markov chain $(X_t)_{t\in\mathbb{N}}$ on a numerable subset of E, with transition matrix P and such that the distribution of X_0 is an invariant measure. Then one can also construct a double-sided sequence $(X_t)_{t\in\mathbb{Z}}$ such that the distribution of X_t is π of any $t \in \mathbb{Z}$ and the conditional distribution of $X_t | X_{t-1}, X_{t-2}, \ldots$ is given by P (and then only depends on X_{t-1}). Indeed, let

   $$\mu_n\left(\{(x_{-n}, \ldots, x_n)\}\right) = \pi(x_{-n})P(x_{-n}, x_{-n+1}) \cdots P(x_{n-1}, x_n), \quad (x_{-n}, \ldots, x_n) \in E^{2n+1}.$$

   Using the equation $\pi P = \pi$, one can show that

   $$\mu_n\left(\{(x_{-n}, \ldots, x_n)\}\right) = \sum_{x_{-n-1}, x_{n+1} \in E} \mu_{n+1}\left(\{(x_{-n-1}, \ldots, x_{n+1})\}\right).$$

   Then Theorem 25 applies.

The next notion is the concept of stationarity.

Definition 7. *A process $(X_t)_{t\in\mathbb{Z}}$ is said to be strictly stationary (or simply stationary) if for any positive integer h, the probability distribution of $(X_{t+h})_{t\in\mathbb{Z}}$ coincides with that of $(X_t)_{t\in\mathbb{Z}}$.*

In the previous definition, the distribution of a process is thought as a probability measure on the space $E^{\mathbb{Z}}$ endowed with the cylinder sigma-field $\mathcal{C}$. Let us mention that $\mathcal{C}$ is often denoted by $\mathcal{B}(E)^{\otimes\mathbb{Z}}$.

The following result is a consequence of the monotone class theorem.

Proposition 5. *A process $(X_t)_{t\in\mathbb{Z}}$ taking values on E is stationary if and only if for any $k\in\mathbb{N}$ and $t\in\mathbb{Z}$, the distribution of the vectors $(X_t,\dots,X_{t+k})$ and $(X_0,\dots,X_k)$ are the same.*

An important operator used to formulate some properties of stationary processes is the shift operator $\tau : E^{\mathbb{Z}}\to E^{\mathbb{Z}}$ defined by $\tau\omega=(\omega_{t+1})_{t\in\mathbb{Z}}$. τ is invertible with inverse $\tau^{-1} : E^{\mathbb{Z}}\to E^{\mathbb{Z}}$ defined by $\tau^{-1}\omega=(\omega_{t-1})_{t\in\mathbb{Z}}$. If t is positive integer, we denote by τ^t the composition $\underbrace{\tau\circ\cdots\circ\tau}_{t}$ and if t is a negative integer, $\tau^t=\underbrace{\tau^{-1}\circ\cdots\circ\tau^{-1}}_{-t}$.

Proposition 6. *A process $X:=(X_t)_{t\in\mathbb{Z}}$ is stationary if and only if τX and X have the same probability distribution.*

Proof of Proposition 6. The direct sense follows from the definition of stationarity. For the reciprocal sense, if τX has the same distribution as X and h is a positive integer, we have $\tau^h X=\tau^{h-1}\circ\tau X$ and from the measurability of $\tau^{h-1} : E^{\mathbb{Z}}\to E^{\mathbb{Z}}$, we deduce that $\tau^h X$ has the same distribution as $\tau^{h-1}X$. Iterating this, we conclude that $\tau^h X$ as the same distribution as X, as required.□

Let $H : E^{\mathbb{Z}}\to E':=\mathbb{R}^{k'}$ be a measurable mapping and $\varepsilon:=(\varepsilon_t)_{t\in\mathbb{Z}}$ a sequence of i.i.d. random variables taking values in E. We then define a new process $X_t=H\left((\varepsilon_{t+j})_{j\in\mathbb{Z}}\right)$. This kind of process is called a Bernoulli shift. In general, the mapping H is defined on a measurable subset F of $E^{\mathbb{Z}}$ such that $\mathbb{P}(\varepsilon\in F)=1$. One can extend H on $E^{\mathbb{Z}}$ by $H(x)=0$ if $x\in E^{\mathbb{Z}}\setminus F$.

Proposition 7. *A Bernoulli shift is a stationary process.*

Proof of Proposition 7. Since $X_t=H\left(\tau^t\varepsilon\right)$ and denoting again by τ the shift operator on $(E')^{\mathbb{Z}^d}$, we have $\tau X=\left(H\left(\tau^{t+1}\varepsilon\right)\right)_{t\in\mathbb{Z}}$. The result is then a consequence of the stationarity property of ε.□

Examples of Bernoulli shifts

1. Let us study the recursive equations $X_t=aX_{t-1}+\varepsilon_t$ when $|a|<1$ and $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a sequence of i.i.d. and integrable random variables. It is easy to show that the series $X_t=\sum_{j=0}^{\infty}a^j\varepsilon_{t-j}$ is normally convergent in $\mathbb{L}^1$ (and then a.s. absolutely convergent)

and that $(X_t)_{t\in\mathbb{Z}}$ is a solution of these recursive equations. It is also a Bernoulli shift, setting

$$H\left((s_j)_{j\in\mathbb{Z}}\right) = \sum_{j\geq 0} a^j s_{-j},$$

which is defined on the subset F of $E^{\mathbb{Z}}$ for which the previous series is absolutely converging. One can also show that there is only one stationary solution. Indeed, if $(X'_t)_{t\in\mathbb{Z}}$ is another stationary solution, setting $D_t = X_t - X'_t$, we have $D_t = a^n D_{t-n}$ for any $n \in \mathbb{N}$. But $a_n|D_{t-n}| \leq a^n|X_{t-n}| + a^n|X'_{t-n}|$ and the two terms on right-hand side of this inequality are both converging to 0 in probability when $n \to \infty$. We deduce that $D_t = 0$ a.s. and then $X_t = X'_t$ a.s.

2. One can extend the previous model to

$$X_t = \sum_{j=1}^{p} a_j \varepsilon_{t-j} + \varepsilon_t, \quad t \in \mathbb{Z}, \tag{4.1}$$

with the same assumptions for ε but now with the assumption that the roots of the polynomial $\mathcal{P}(z) = 1 - \sum_{j=1}^{p} a_j z^j$ are outside the unit disc of the set of complex numbers $\mathbb{C}$. It is easy to show that we have a one-to-one correspondence between the stationary process $(X_t)_{t\in\mathbb{Z}}$ solutions of (4.1) and the stationary solutions of the multivariate recursions $Y_t = AY_{t-1} + \widetilde{\varepsilon}_t$ where

$$A = \begin{pmatrix} a_1 & \cdots & a_{p-1} & a_p \\ & I_{p-1} & & 0_{p-1,1} \end{pmatrix}, \quad \widetilde{\varepsilon}_t = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

where I_{p-1} denotes the identity matrix of size $p-1$ and $0_{p-1,1}$ the column vector with $p-1$ components all equal to 0. The matrix A is often called companion matrix. A well known result about companion matrices states that the spectral radius $\rho(A)$ of the matrix A is less than 1 if and only if the polynomial $\mathcal{P}$ has all its roots outside the unit disc. More precisely, one can show that the characteristic polynomial of A equals to $\mathcal{Q}(z) = (-1)^p \left(z^p - \sum_{j=1}^{p} a_j z^{p-j}\right)$. Moreover, the Gelfand formula guarantees that for any matrix norm $\|\cdot\|$, for instance the norm 2 associated to the Euclidean norm $\|\cdot\|_2$ on $\mathbb{R}^p$, i.e.

$$\|A\| = \sup_{x\in\mathbb{R}^p : \|x\|_2 = 1} \|Ax\|_2 = \sqrt{\rho(A^T A)},$$

$\lim_{n\to\infty} \|A^n\|^{1/n} = \rho(A)$. When $\rho(A) < 1$, we then deduce that there exists a positive integer n_0 such that $\kappa := \|A^{n_0}\| < 1$. It is easy to deduce that the series $Y_t = \sum_{j=0}^{\infty} A^j \widetilde{\varepsilon}_{t-j}$ is normally convergent in $\mathbb{L}^1$ and is the unique stationary solution of the multivariate recursions. Note that if $\mathbb{E}(\varepsilon_1^2) < \infty$, the series also converges in $\mathbb{L}^2$ and the solution has a finite second moment.

3. GARCH models are widely used in financial econometrics to model the dynamic of stock prices or currency exchange rates. See Francq and Zakoian (2019) for a broad introduction to these kind of models. In what follows, we consider a sequence $(\varepsilon_t)_{t\in\mathbb{Z}}$ of i.i.d. random variables with mean 0 and variance 1. We set

$$X_t = \varepsilon_t \sigma_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \alpha_2 \sigma_{t-1}^2, \tag{4.2}$$

where $\alpha_0 > 0$ and $\alpha_1, \alpha_2 \geq 0$. When σ_t and ε_t are independent with $\sigma_t \in \mathcal{F}_{t-1} := \sigma\left(X_{t-j} : j \geq 1\right)$, then σ_t^2 is simply the conditional variance of X_t given $\mathcal{F}_{t-1}$, i.e.

$$\text{Var}\ \left(X_t|\mathcal{F}_{t-1}\right) := \mathbb{E}\left(X_t^2|\mathcal{F}_{t-1}\right) - \mathbb{E}^2\left(X_t|\mathcal{F}_{t-1}\right) = \sigma_t^2.$$

Our aim here is to construct a stationary solution $\left((X_t, \sigma_t)\right)_{t\in\mathbb{Z}}$ of (4.2) which is also non-anticipative (i.e. $\sigma_t \in \sigma\left(\varepsilon_{t-j} : j \geq 1\right)$). To this end, we assume that $\alpha_1 + \alpha_2 < 1$. We first note that any solution of (4.2) satisfies $\sigma_t^2 = \alpha_0 + a_{t-1}\sigma_{t-1}^2$ where $a_{t-1} = \alpha_1 \varepsilon_{t-1}^2 + \alpha_2$. Under our assumptions, we have $\mathbb{E}a_1 < 1$ and one can show that

$$\sigma_t^2 = \alpha_0 \left[1 + \sum_{j=1}^{\infty} a_{t-1} \cdots a_{t-j}\right], \quad t \in \mathbb{Z}$$

is a random series converging in $\mathbb{L}^1$ and a.s. and solution of the recursions. It is of course stationary and $X_t = \varepsilon_t \sigma_t$ is stationary solution of (4.2). One can also show that the previous solution is the unique stationary solution of (4.2).

Though is is possible in theory to use two-sided Bernoulli shifts, such as $X_t = \sum_{j\in\mathbb{Z}} a_j \varepsilon_{t+j}$, many interesting examples are one-sided, i.e.

$$H\left((s_t)_{t\in\mathbb{Z}}\right) = G\left(s_t, s_{t-1}, \ldots\right)$$

for a measurable mapping G. A general construction of one-sided Bernoulli shifts uses contraction properties of iterative systems, as illustrated in the aforementioned examples. The following result, which can be found in Wu and Shao (2004), extends this setup. The proof is left as an exercise.

Theorem 26. *Let E be a Borel subset of $\mathbb{R}^k$, $(G, \mathcal{G})$ be a measurable space, $F : E \times G \to E$ a measurable mapping and $(\varepsilon_t)_{t\in\mathbb{Z}}$ be a sequence of i.i.d. random variables taking values in G. Setting $f_t(x) = F(x, \varepsilon_t)$ for $(t, x) \in \mathbb{Z} \times E$, and $f_s^t(x) = f_t \circ f_{t-1} \circ \cdots \circ f_s(x)$ for $s \leq t$, we assume that the following conditions hold true for some $p \geq 1$ and a norm $\|\cdot\|$ on $\mathbb{R}^k$.*

1. *For all $x \in E$, $f_t(x) \in \mathbb{L}^p$.*
2. *There exists a positive integer m and two positive real numbers L and κ, with $\kappa < 1$ and such that*

$$\|f_t(x) - f_t(y)\|_p \leq L\|x - y\|,$$
$$\|f_{t+1}^{t+m}(x) - f_{t+1}^{t+m}(y)\|_p \leq \kappa\|x - y\|,$$

where for a random variables Z taking values in E, $\|Z\|_p = \mathbb{E}^{1/p}\left[\|Z\|^p\right]$. The sequence $\left(f_{t-n}^t(x)\right)_{n\geq 1}$ has an almost sure limit denoted by $f_{-\infty}^t$ and not depending on x. There then exists a unique stationary solution $(X_t)_{t\in\mathbb{Z}}$ such that $X_t \in \sigma\left(\varepsilon_{t-j} : j \geq 0\right)$ and $X_t = F\left(X_{t-1}, \varepsilon_t\right)$ a.s. We have $X_t = f_{-\infty}^t$ a.s. Moreover $\mathbb{E}\left[\|X_1\|^p\right] < \infty$.

Sketch of the proof. For an arbitrary $x \in E$, the sequence $\left(f_{t-n}^t(x)\right)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{L}^p$ and has then a limit denoted by $X_t(x)$. The a.s. convergence to this limit also holds true. From the second assumption, we have $X_t(x) = X_t(x')$ for any $x' \in E$. Uniqueness of this non-anticipative stationary solution (which has a Bernoulli shift representation) follows from the contraction condition in the second assumption.□

A note on Markov chains on $\mathbb{R}^k$. A sequence $(X_t)_{t\in\mathbb{N}}$ of $E-$valued random variables is called a homogeneous Markov chain if for any integer $t \geq 1$,

$$\mathbb{P}\left(X_t \in A | X_{t-1} = x_{t-1}, \ldots, X_0 \in x_0\right) = \mathbb{P}\left(X_1 \in A | X_0 = x_{t-1}\right),$$

for any $A \in \mathcal{B}(E)$ and $x_0, \ldots, x_{t-1} \in E$. The mapping $(x, A) \mapsto K(x, A) := \mathbb{P}\left(X_1 \in A | X_0 = x\right)$ is called a Markov kernel (or a transition kernel). A Markov kernel $K : E \times \mathcal{B}(E) \to [0, 1]$ is simply a mapping such that for any $A \in \mathcal{B}(E)$, $x \mapsto K(x, A)$ is measurable and for any $x \in E$, $A \mapsto K(x, A)$ defines a probability measure. An integral $\int_E f(y)K(x, dy)$ can be defined using such a probability measure.

Under the assumptions of Theorem 26, for any $x \in E$, the sequence $(f_1^t(x))_{t\geq 0}$ is a Markov chain, starting at 0 (using the convention $f_1^0(x) = x$). Its transition kernel is given by $K(x, A) = \mathbb{P}\left(f_1(x) \in A\right)$.

A probability measure μ on E is said to be invariant if $\mu(A) = \mu K(A) := \int_E \mu(dy)K(y, A)$. The distribution μ of X_0 given in the previous theorem is an invariant probability measure (it corresponds to the marginal distributions of a stationary sequence). Now define by induction the transition kernels K^n by

$$K^n(y, A) = \int_E K(y, dz)K^{n-1}(z, A), \quad n \geq 2.$$

It is easy to check that $K^n(y, A) = \mathbb{P}\left(f_1^n(y) \in A\right)$. By induction, we get $\mu K^n = \mu$. If $h : E \to \mathbb{R}$ is a continuous and bounded function, we have

$$\int_E h(y)K^n(x, dy) = \mathbb{E}\left[h \circ f_1^n(x)\right] = \mathbb{E}\left[h \circ f_{-n}^{-1}(x)\right] \overset{n\to\infty}{\to} \mathbb{E}\left[h(X_{-1})\right] = \int_E h d\mu.$$

We then conclude that for any $x \in E$, $f_1^n(x) \hookrightarrow \mu$, which means that the distribution of the Markov chain converges to μ whatever the initial state x. From the dominated convergence theorem, this is also true for any initial probability measure ν, since

$$\int_E \int_E h(y)K^n(x, dy)\nu(dx) = \int_E h(y)\nu K^n(dy)$$

and νK^n is the probability distribution of X_n when X_0 is generated from ν. As a consequence, the invariant probability measure of the chain is unique.

4.2 Ergodic theory for stationary processes indexed by $\mathbb{Z}$

For generalizing the strong law of large numbers to stationary sequences, we directly face to some pathological problems. For instance, when $X_t = X_0$ for all $t \in \mathbb{Z}$, then $\frac{1}{n}\sum_{t=1}^{n} X_t = X_0$ which does not coincide with $\mathbb{E}(X_0)$ except is X_0 is a.s. constant. Moreover, for a Markov chain on $E = \{0, 1\}$ with transition $P = I_2$ the identity matrix, the probability $\pi = (1/2, 1/2)$ is invariant. However $\frac{1}{n}\sum_{t=1}^{n} X_t$ is equal to 0 on the set $\{X_0 = 0\}$ which has probability $1/2$. The a.s. limit of these partial sums cannot be the expectation of π which is $1/2$. We then see that some problems can occur when there exist some non trivial "invariant" sets, i.e. sets A with probability in $(0, 1)$ and for which $\{X_0 \in A\} = \{X_t \in A\}$ for any $t \geq 1$.

Ergodic theory is a branch of mathematics which studies the properties of some mappings $\tau : G \to G$ that are invariant under a probability measure μ on a measurable space $(G, \mathcal{G})$, i.e. $\mu(\{g \in G : \tau g \in A\}) = \mu(A)$ for any $A \in \mathcal{G}$. In our framework, $G = E^{\mathbb{Z}}$, τ is the shift operator already defined in the previous section and $\mu = \mathbb{P}_X$, the distribution of a stationary process $X = (X_t)_{t\in\mathbb{Z}}$.

Before giving the generalization of the law of large numbers, we introduce the following definition.

Definition 8. *Let $(G, \mathcal{G}, \mu)$ be a probability space and $\tau : G \to G$ a measurable mapping.*

1. *We say that τ preserves the measure μ if $\mu(\tau^{-1}A) = \mu(A)$ for any $A \in \mathcal{G}$. Here $\tau^{-1}A = \{x \in G : \tau x \in A\}$.*
2. *A measurable subset $I \in \mathcal{G}$ is said to be invariant if $\tau^{-1}I = I$.*
3. *τ is said to be ergodic for μ if any invariant subset I is trivial, i.e. it has measure 0 or 1.*

A generalization of the strong law of large numbers can be obtained by inspecting the limiting behavior of $x \mapsto S_n f(x) := \frac{1}{n}\sum_{t=1}^{n} f(\tau^t x)$. However, if I is a non trivial invariant subset, the choice $f = \mathbb{1}_I$ leads to $S_n f = 1$ which cannot converge to $\mu(I) = \int f d\mu$. A proof of the following important result can be found in Petersen (1989), Chapter 2.

Theorem 27 (Birkhoff's ergodic theorem)**.** *Suppose that τ is ergodic for μ. Then is $\int |f| d\mu < \infty$, we have*

$$\lim_{n\to\infty} S_n f = \int f d\mu \ \ a.s.$$

When τ is the shift operator on $G = E^{\mathbb{Z}}$ and $\mu = \mathbb{P}_X$, we say that the stationary process X is ergodic if τ is ergodic for μ. In this context, we obtain the following result.

Corollary 4. *Suppose that $X = (X_t)_{t\in\mathbb{Z}}$ is a stationary and ergodic process taking values in E and $f : E^{\mathbb{Z}} \to \mathbb{R}$ is a measurable function such that $\mathbb{E}[|f(X)|] < \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} f\left((X_{t+j})_{j\in\mathbb{Z}}\right) = \mathbb{E}f(X), \quad a.s.$$

Under the assumptions of Corollary 4, we note that if $g : E \to \mathbb{R}$ is $\mathbb{P}_{X_0}$ integrable, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} g(X_t) = \mathbb{E}g(X_0), \quad \text{a.s.}$$

However, one can consider partial sums for much more complicated functions f, for instance depending of infinitely many coordinates. To show that we truly obtained an extension of the law of large numbers, the following result will be needed.

Proposition 8. *A sequence $(X_t)_{t\in\mathbb{Z}}$ of i.i.d. random variables is ergodic.*

Proof of Proposition 8. Let I be an invariant subset of $E^{\mathbb{Z}}$. Then $\mathbb{P}_X(I) = \mathbb{P}(X \in I) = \mathbb{P}_X\left(\tau^{-n} I \cap I\right)$. We are going to show that for any A and B in the cylinder sigma-field $\mathcal{C}$,

$$\lim_{n\to\infty} \mathbb{P}\left(X \in \tau^{-n} A, X \in B\right) = \mathbb{P}\left(X \in A\right) \mathbb{P}\left(X \in B\right). \tag{4.3}$$

Applying (4.3) to $A = B = I$, we will deduce that $\mathbb{P}_X(I) = \mathbb{P}_X(I)^2$, meaning that $\mathbb{P}_X(I) \in \{0, 1\}$, as required.

To prove (4.3), suppose first that A and B are finite unions of cylinder sets. In this case $\{X \in A\} \in \sigma\left(X_s : s \in U\right)$ and $\{X \in B\} \in \sigma\left(X_s : s \in V\right)$ for two finite subsets U and V of $\mathbb{Z}$. But since $\{\tau^n X \in A\} \in \sigma\left(X_{s+n} : s \in U\right)$, this event is independent of $\{X \in B\}$ when n is large enough. Then for such n,

$$\mathbb{P}\left(X \in \tau^{-n} A, X \in B\right) = \mathbb{P}\left(X \in \tau^{-n} A\right) \mathbb{P}\left(X \in B\right) = \mathbb{P}\left(X \in A\right) \mathbb{P}\left(X \in B\right).$$

For the general case, one can note that the finite unions of cylinder sets form an algebra (that is a set of subsets of $E^{\mathbb{Z}}$, containing the empty set, stable by finite union and stable by taking complements) which generates the cylinder sigma-field. A general result in measure theory ensures that if $\mathcal{A}$ is a sigma-field, μ a probability measure on $\mathcal{A}$ and $\widetilde{\mathcal{A}}$ is an algebra generating $\mathcal{A}$, then for any $\varepsilon > 0$ and $A \in \mathcal{A}$, there exists $A_\varepsilon \in \widetilde{\mathcal{A}}$ such that

$$\mu\left(A \Delta A_\varepsilon\right) \leq \varepsilon, \quad A \Delta A_\varepsilon = \left(A \setminus A_\varepsilon\right) \cup \left(A_\varepsilon \setminus A\right).$$

Applying this result to $\mathcal{A} = \mathcal{C}$ the cylinder sigma-field, $\mu = \mathbb{P}_X$, and $\widetilde{\mathcal{A}}$ the set of finite unions of cylinder sets, we get

$$\begin{aligned}
& \left|\mathbb{P}_X\left(\tau^{-n} A \cap B\right) - \mathbb{P}_X\left(\tau^{-n} A_\varepsilon \cap B_\varepsilon\right)\right| \\
\leq\ & \mathbb{P}_X\left(\left(\tau^{-n} A \cap B\right) \Delta \left(\tau^{-n} A_\varepsilon \cap B_\varepsilon\right)\right) \\
\leq\ & \mathbb{P}_X\left(\tau^{-n} A \Delta \tau^{-n} A_\varepsilon\right) + \mathbb{P}_X\left(B \Delta B_\varepsilon\right) \\
=\ & \mathbb{P}_X\left(A \Delta A_\varepsilon\right) + \mathbb{P}_X\left(B \Delta B_\varepsilon\right) \\
\leq\ & 2\varepsilon.
\end{aligned}$$

Note that the previous equality is a consequence of stationarity. Since $\varepsilon > 0$ is arbitrary, it is easy to conclude that (4.3), already valid for the pair $(A_\varepsilon, B_\varepsilon)$, extends to the pair (A, B). This concludes the proof. □

4.3 Semiparametric M-estimation for autoregressive processes

4.3.1 Estimation of a conditional mean

Let the model

$$X_t = f_{\theta_0}(X_{t-1}) + \varepsilon_t, \quad t \in \mathbb{Z}, \tag{4.4}$$

where the ε_t's are i.i.d. with mean 0 and finite variance and $f_{\theta_0} : \mathbb{R} \to \mathbb{R}$ is a measurable mapping depending on a parameter $\theta_0 \in \Theta$. A natural estimator can be obtained by minimizing

$$\theta \mapsto M_n(\theta) := \frac{1}{n} \sum_{t=2}^{n} \left(X_t - f_\theta(X_{t-1})\right)^2.$$

The corresponding estimator $\hat{\theta}_n$ is called non-linear least squares estimator. Existence of a stationary and ergodic solution for (4.4) can be obtained from Theorem 26, as soon as there exists $\kappa \in (0,1)$ such that

$$|f_{\theta_0}(x) - f_{\theta_0}(y)| \leq \kappa |x - y|, \quad (x,y) \in \mathbb{R}^2.$$

Moreover, when Θ is compact, $\theta \mapsto f_\theta(x)$ is continuous for all x and $\sup_{\theta\in\Theta} |f_\theta(X_1)|$ is integrable, Birkhoff's ergodic theorem is sufficient to ensure strong consistency of $\hat{\theta}_n$. Indeed one can apply Theorem 2 of Chapter 2. The reason is that, except some regularity and integrability conditions, only the pointwise law of large numbers was necessary to obtain consistency in the i.i.d. setting. We have know extended this law to dependent data and all the other arguments needed to prove Theorem 2 in Chapter 2 are not restricted to independent observations. The single assumption to check is the third one. We note that by independence between X_{t-1} and ε_t,

$$M(\theta) = \mathbb{E}\left(\varepsilon_1^2\right) + \mathbb{E}\left[\left(f_\theta(X_1) - f_{\theta_0}(X_1)\right)^2\right]$$

and $M(\theta) \geq M(\theta_0)$. Moreover $M(\theta) = M(\theta_0)$ if and only if $f_\theta = f_{\theta_0}$ $\mu-$a.s., where μ is the probability distribution of X_1. Then the third assumption follows as soon as $\mu\left(\{f_\theta \neq f_{\theta_0}\}\right) > 0$ for $\theta \neq \theta_0$.

Let us know investigate the case of a linear autoregressive process of order p (4.1), which is often denoted AR(p), and for which the least squares estimator has en explicit form. We assume here that $\mathbb{E}\left(\varepsilon_1^2\right) < \infty$ and the root of $\mathcal{P}(z) = 1 - \sum_{j=1}^{p} a_j z^j$ are outside the unit disc. This ensures that the unique stationary solution has a finite second moment. Our aim is to estimate $\theta_0 = (a_1, \ldots, a_p)^T$ when $\mathbb{E}\left(\varepsilon_1^2\right) < \infty$. The least squares estimator is given by

$$\hat{\theta}_n = \arg\min_{\theta\in\mathbb{R}^p} \frac{1}{n} \sum_{t=p+1}^{n} \left(X_t - \mathcal{X}_{t-1}^T \theta\right)^2 = \left(\frac{1}{n} \sum_{t=p+1}^{n} \mathcal{X}_{t-1} \mathcal{X}_{t-1}^T\right)^{-1} \frac{1}{n} \sum_{t=p+1}^{n} \mathcal{X}_{t-1} X_t.$$

Here $\mathcal{X}_{t-1}$ denotes the column vector with entries $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$. From the ergodic theorem, we have

$$\frac{1}{n}\sum_{t=p+1}^{n} \mathcal{X}_{t-1}X_t \overset{a.s.}{\to} \mathbb{E}\left[\mathcal{X}_0 X_1\right] = \mathbb{E}\left[\mathcal{X}_0\mathcal{X}_0^T\right]\theta_0.$$

To get strong consistency, we only have to check that $\Gamma := \mathbb{E}\left[\mathcal{X}_0\mathcal{X}_0^T\right]$ is invertible. Using the ergodic theorem and the continuity of the inverse of a matrix, this will ensure that

$$\left(\frac{1}{n}\sum_{t=p+1}^{n} \mathcal{X}_{t-1}\mathcal{X}_{t-1}^T\right)^{-1} \overset{a.s.}{\to} \Gamma^{-1}$$

and then strong consistency. If Γ is not invertible, there exists $u \in \mathbb{R}^p \setminus \{0\}$ such that

$$u^T\Gamma u = \mathbb{E}\left[\left(u^T\mathcal{X}_0\right)^2\right] = 0.$$

We then conclude that $u^T\widetilde{X}_0 = 0$ a.s. and one variable, for instance X_0, writes a.s. as a linear combination of $X_{-1}, \ldots, X_{-p+1}$. But this, due to the model form, this would mean that ε_0 writes as a linear combination of $X_{-1}, \ldots, X_{-p}$. By independence, this is impossible when ε_1 is not constant a.s. Then $u^T\Gamma u = 0$ is not possible unless $u = 0$.

4.3.2 Estimation of a conditional variance

Let the model

$$X_t = \varepsilon_t\sigma_t, \quad \sigma_t^2 = \theta_{0,1} + \theta_{0,2}X_{t-1}^2,$$

where $\theta_{0,1}$ and $\theta_{0,2}$ are unknown non-negative real numbers and the ε_t's are i.i.d. with mean 0 and variance 1. We have already seen that $\theta_{0,2} < 1$ is a necessary and sufficient condition for existence of a stationary solution. Setting $V_t(\theta) = \theta_1 + \theta_2 X_{t-1}^2$ for $t \in \mathbb{Z}$, a first idea would be to use a least squares estimator with

$$M_n(\theta) = \frac{1}{n}\sum_{t=2}^{n} m_\theta(X_{t-1}, X_t), \quad m_\theta(X_{t-1}, X_t) = \left(X_t - V_t(\theta)\right)^2.$$

This is natural since $\mathbb{E}\left[X_t^2|X_{t-1}\right] = V_t(\theta_0)$. However, such estimator requires the existence of the fourth moment for consistency. One can show that existence of the fourth moment induces a supplementary restriction on $\theta_{0,2}$. This is why, we prefer another M-estimator called Gaussian Quasi-Maximum Likelihood Estimator (QMLE). The principle is to compute the density of $(X_2, \ldots, X_n)$ conditionally on X_1, assuming that ε_t follows a standard Gaussian distribution. The conditional density of X_t given X_{t-1} is given by

$$f\left(x_t|X_{t-1}\right) = \frac{1}{\sqrt{2\pi V_t(\theta_0)}}\exp\left(-\frac{x_t^2}{2V_t(\theta_0)}\right).$$

Then the conditional density of $(X_2, \ldots, X_n)$ given X_1 and evaluated at $(X_2, \ldots, X_n)$ is given by

$$L_n(\theta_0) = \prod_{t=2}^{n} \frac{1}{\sqrt{2\pi V_t(\theta_0)}} \exp\left(-\frac{X_t^2}{2V_t(\theta_0)}\right).$$

Maximizing $\theta \mapsto L_n(\theta)$ is equivalent to minimize

$$\theta \in M_n(\theta) := \frac{1}{n}\sum_{t=2}^{n} m_\theta(X_{t-1}, X_t), \quad m_\theta(X_{t-1}, X_t) = \frac{X_t^2}{V_t(\theta)} + \log V_t(\theta).$$

To obtain a compact parameter space, one can set

$$\Theta = [0, M] \times [0, 1/M] \text{ for some } M > 1.$$

When ε_t is Gaussian, the QMLE is simply called conditional likelihood estimator (since it is based on the conditional density). The main interest is that when ε_t is not necessarily Gaussian, the method still works and can be used to get a consistent estimator (which explains the terminology Quasi Likelihood). This is justified by the equality

$$M(\theta) - M(\theta_0) = \mathbb{E}\left[\frac{V_t(\theta_0)}{V_t(\theta)} - \log\frac{V_t(\theta_0)}{V_t(\theta)} - 1\right]$$

and the inequality $x - \log(x) - 1 \geq 0$ for $x > 0$ with equality if and only if $x = 1$. We then get $M(\theta) \geq M(\theta_0)$ and $M(\theta) = M(\theta_0)$ if and only if $V_t(\theta) = V_t(\theta_0)$ a.s. When the distribution of ε_t is not concentrated on $\{-1, 1\}$, it is possible to show that necessarily $\theta = \theta_0$. All the other assumptions of Theorem 2 in Chapter 2 are verified. In particular, we have automatically $\mathbb{E}\left[\sup_{\theta\in\Theta} |m_\theta(X_0, X_1)|\right] < \infty$ here, since $\mathbb{E}X_t^2 < \infty$.

4.3.3 What about asympotic normality?

For semiparametric-models of the previous types, one can proceed as in the i.i.d. case. For the non-linear least squares estimator and the Gaussian QMLE, the quantity $\dot{M}_n(\theta_0)$ which gives the asymptotic distribution of the M-estimator are given respectively by

$$\dot{M}_n(\theta_0) = -\frac{2}{n}\sum_{t=2}^{n} (X_t - f_{\theta_0}(X_{t-1}))\, \dot{f}_{\theta_0}(X_{t-1})$$

and

$$\dot{M}_n(\theta_0) = -\frac{1}{n}\sum_{t=2}^{n} \left\{\frac{X_t^2 \dot{V}_t(\theta_0)^2}{V_t(\theta_0)} - \frac{\dot{V}_t(\theta_0)}{V_t(\theta_0)}\right\}.$$

One can observe that in both cases, $\mathbb{E}\left[\dot{m}_{\theta_0}(X_{t-1}, X_t)|\mathcal{F}_{t-1}\right] = 0$ a.s. where $\mathcal{F}_{t-1} = \sigma\left(X_s : s \leq t-1\right)$. Then the partial sums $\sum_{t=2}^{n} \dot{m}_{\theta_0}(X_{t-1}, X_t)$ form a martingale. We then need a central limit theorem for this kind of martingales, written as a partial sum of stationary and ergodic sequences.

This kind of situation is classical in conditional models. It corresponds to the situation where $\theta \mapsto \mathbb{E}\left[m_{\theta_0}(X_0, X_1)|\mathcal{F}_0\right]$ is minimized at $\theta = \theta_0$. Inverting derivative and conditional expectation yields to the martingale property discussed above.

4.4 Central limit theorems for martingales

In this section, we consider some partial sums of the form $S_n = \sum_{i=1}^{k_n} X_{n,j}$ where

$\{X_{n,j} : 1 \leq j \leq k_n, n \geq 1\}$ is a "triangular array" of random variables. Additionally, we assume that $(X_{n,j})_{1 \leq j \leq k_n}$ is a martingale difference, meaning that for each positive integer n, there exists a filtration $(\mathcal{F}_{n,j})_{0 \leq j \leq k_n}$ such that for $1 \leq j \leq k_n$, $X_{n,j}$ is integrable and measurable with respect to $\mathcal{F}_{n,j}$ and $\mathbb{E}[X_{n,j}|\mathcal{F}_{n,j-1}] = 0$ a.s.

For instance, in the context of the previous paragraph, $X_{n,j} = Y_j/\sqrt{n}$, where $(Y_j)_{j \in Z}$ is a stationary and ergodic sequence of integrable random variables and $\mathbb{E}[Y_j|\mathcal{F}_{j-1}] = 0$ for $\mathcal{F}_{n,j} = \mathcal{F}_j = \sigma(Y_t : t \leq j)$, we are interested by a central limit theorem for S_n.

The principle is to study convergence of the characteristic function $\phi_n(t) = \mathbb{E}[\exp(itS_n)]$ using a subtle factorization of the complex exponential. In what follows, the notation $|\cdot|$ is used for both the absolute value of a real number or the modulus of a complex number. We follow the approach of McLeish (1974) for proving martingale central limit theorems.

Lemma 11. *There exists a mapping $r : \mathbb{R} \to \mathbb{C}$ such that*

$$\exp(ix) = (1 + ix) \exp\left(-\frac{x^2}{2} + r(x)\right)$$

and $|r(x)| \leq |x|^2$ for any $-1 < x < 1$.

Proof of Lemma 11. The series $\log(1 + z) = \sum_{k \geq 1} (-1)^{k+1} \frac{z^k}{k}$ converges for $|z| < 1$ and it is known that $\exp(\log(1 + z)) = 1 + z$. Now, take $z = ix$ for some $x \in (-1, 1)$. We have

$$\log(1 + ix) = ix + \frac{x^2}{2} + x^3 r(x), \quad r(x) = \sum_{k=0}^{\infty} (-1)^{k+1} \frac{i^{k+1} x^k}{k + 3}.$$

We then get the required factorization by taking the exponential function in the previous equality. Finally,

$$r(x) = r_1(x) + i r_2(x), \quad r_1(x) = \sum_{p \geq 0} (-1)^{p+1} \frac{x^{2p+1}}{2p + 4}, \quad r_2(x) = \sum_{p \geq 0} (-1)^{p+1} \frac{x^{2p}}{2p + 3}.$$

Suppose that $x > 0$, the argument will be the same if $x = -y < 0$. Due to the presence of alternating series, we have $-1/4 \leq r_1(x) \leq 0$ and $-1/3 \leq r_2(x) \leq 0$ from which we conclude that $|r(x)| \leq 1$.□

The idea is then to use the decomposition $\exp(itS_n) = T_n U_n$ where

$$T_n = \prod_{j=1}^{k_n} (1 + itX_{n,j}), \quad U_n = \exp\left(\frac{-t^2}{2} \sum_{j=1}^{k_n} X_{n,j}^2 + \sum_{j=1}^{k_n} r(tX_{n,j})\right).$$

For martingale differences, we have

$$\mathbb{E}[1 + itX_{n,j}|\mathcal{F}_{n,j-1}] = 1$$

and we then get $\mathbb{E}(T_n) = 1$. Moreover, if $\sum_{j=1}^{k_n} X_{n,j}^2 \xrightarrow{P} \sigma^2$, one can hope that $\mathbb{E}(T_n U_n) \to \exp\left(-\frac{t^2\sigma^2}{2}\right)$, which suggests that $S_n \hookrightarrow \mathcal{N}(0, \sigma^2)$. Note that when $X_{n,j} = X_j/\sqrt{n}$ with $(X_j)_{j\in\mathbb{Z}}$ stationary, ergodic and square-integrable, we have $\sigma^2 = \mathbb{E}(X_1^2)$.

However, converge of the expectation of the product requires a specific attention. In particular, the following result is helpful. The concept of uniform integrability will be needed. We recall that a sequence of random variables $(T_n)_{n\geq 1}$ is uniformly integrable if for any $\epsilon > 0$, one can find $M > 0$ sufficiently large such that

$$\sup_{n\geq 1} \mathbb{E}\left[|T_n| \mathbb{1}_{|T_n|\geq M}\right] \leq \epsilon.$$

If for all n, $T_n = T$ with T integrable then $(T_n)_{n\geq 1}$ is uniformly integrable. Moreover, the sum of two uniformly integrable sequences is still uniformly integrable. Let n_0 be a positive integer. If $\mathbb{E}[|T_n|] < \infty$ for all $n \geq 1$ and for any $\epsilon > 0$, one can find $M > 0$ such that

$$\sup_{n\geq n_0} \mathbb{E}\left[|T_n| \mathbb{1}_{|T_n|\geq M}\right] \leq \epsilon,$$

then $(T_n)_{n\geq 1}$ is also uniformly integrable. Indeed, one can always increase M to also get

$$\max_{1\leq n\leq n_0-1} \mathbb{E}\left[|T_n| \mathbb{1}_{|T_n|> M}\right] \leq \epsilon.$$

Finally, if additionally $T_n \xrightarrow{P} 0$, it is easy to show that $\lim_{n\to\infty} \mathbb{E}(T_n) = 0$.

Lemma 12. *Let $(T_n)_{n\geq 1}$ and $(U_n)_{n\geq 1}$ be two sequences of random variables such that for some real number a,*

1. *$U_n \xrightarrow{P} a$,*
2. *$(T_n)_{n\geq 1}$ is uniformly integrable,*
3. *$(T_n U_n)_{n\geq 1}$ is uniformly integrable,*
4. *$\lim_{n\to\infty} \mathbb{E}(T_n) = 1$.*

Then $\lim_{n\to\infty} \mathbb{E}[T_n U_n] = 1$.

Proof of Lemma 12 Since $T_n U_n = T_n(U_n - a) + T_n a$, we simply have to show that $\lim_{n\to\infty} \mathbb{E}[T_n(U_n - a)] = 0$. From the second and the third assumption, $(T_n(U_n - a))_{n\geq 1}$ is uniformly integrable, as a sum of two uniformly integrable sequences. It is then sufficient to show that this sequence converges to 0 in probability. Let $\varepsilon > 0$. We have

$$\begin{aligned}
\mathbb{P}(|T_n(U_n - a)| > \varepsilon) &\leq \mathbb{P}(|T_n| > M) + \mathbb{P}(|T_n| \leq M, M|T_n - a| > \varepsilon) \\
&\leq \frac{1}{M}\mathbb{E}\left[|T_n| \mathbb{1}_{|T_n|>M}\right] + \mathbb{P}(|U_n - a| > \varepsilon/M).
\end{aligned}$$

If M is large, from uniform integrability, the first term in the last upper-bound can be made arbitrarily small, uniformly over n, smaller than δ for a given $\delta > 0$. For such M, the second term converges to 0. Then

$$\overline{\lim}\, \mathbb{P}\left(|T_n(U_n - a)| > \varepsilon\right) \leq \delta$$

which shows the result.□

Lemma 13. *Let $\{X_{n,j} : 1 \leq j \leq k_n, n \geq 1\}$ be an array of random variables. Let the decomposition $S_n = T_n U_n$ with*

$$T_n = \prod_{1 \leq j \leq k_n} (1 + itX_{n,j}), \quad U_n = \exp\left(-\frac{t^2}{2}\sum_{j=1}^{k_n} X_{n,j}^2 + \sum_{j=1}^{k_n} r(tX_{n,j})\right).$$

Suppose that the following assumptions hold true.

1. $\lim_{n\to\infty} \mathbb{E}(T_n) = 1$.
2. $(T_n)_{n\geq 1}$ *is uniformly integrable.*
3. $\sum_{j=1}^{k_n} X_{n,j}^2 \xrightarrow{P} 1$.
4. $\max_{1\leq j\leq k_n} |X_{n,j}| \xrightarrow{P} 0$.

Then $S_n \hookrightarrow \mathcal{N}(0,1)$.

Proof of Lemma 13. We first show that $R_n = \sum_{j=1}^{k_n} r\left(tX_{n,j}\right) = o_P(1)$. Let $\varepsilon > 0$. On the set $A_n := \{\max_{1\leq j\leq k_n} |X_{n,j}| < 1\}$, we have

$$|R_n| \leq \sum_{j=1}^{k_n} |t|^3 \cdot |X_{n,j}|^3 \leq |t|^3 \max_{1\leq j\leq k_n} |X_{n,j}| \sum_{j=1}^{k_n} X_{n,j}^2 = o_P(1),$$

where we used the third and the fourth assumption. Moreover, from the third assumption, $\lim_{n\to\infty} \mathbb{P}(\Omega \setminus A_n) = 0$. We then conclude that $R_n = o_P(1)$, which leads to $U_n \xrightarrow{P} a := \exp\left(-\frac{t^2\sigma^2}{2}\right)$. We also have $|T_n U_n| = 1$ which is uniformly integrable. The result is then a consequence of Lemma 12.□

We now get one of the two main result of this section.

Theorem 28. *Let $\{X_{n,j} : 1 \leq j \leq k_n, n \geq 1\}$ be a triangular array of martingale differences such that*

1. $\lim_{n\to\infty} \mathbb{E}\left[\max_{1\leq j\leq k_n} |X_{n,j}|\right] = 0$,
2. $\sum_{j=1}^{k_n} X_{n,j}^2 \xrightarrow{P} 1$.

Then $S_n \hookrightarrow \mathcal{N}(0,1)$.

Proof of Theorem 28. Set $Z_{n,1} = X_{n,1}$ and for $2 \leq j \leq k_n$, $Z_{n,j} = X_{n,j}\mathbb{1}_{\sum_{r=1}^{j-1} X_{n,r}^2 \leq 2}$. One can observe that $\{Z_{n,j} : 1 \leq j \leq k_n, n \geq 1\}$ is a triangular array of martingale differences with the same filtration. Moreover,

$$\mathbb{P}\left(\cup_{r=1}^{k_n}\{X_{n,r} \neq Z_{n,r}\}\right) \leq \mathbb{P}\left(\sum_{r=1}^{k_n} X_{n,r}^2 > 2\right) \to 0. \tag{4.5}$$

From (4.5), it is enough to show that $\sum_{j=1}^{k_n} Z_{n,j} \hookrightarrow \mathcal{N}(0,1)$. This will follow from an application of Lemma 13. Indeed, if $T_n = \prod_{j=1}^{k_n}(1 + itZ_{n,j})$, we have $\mathbb{E}(T_n) = 1$. Let

$$J = \inf\left\{j \geq 1 : \sum_{r=1}^{j} X_{n,r}^2 > 2\right\} \wedge k_n,$$

where $a \wedge b = \min(a,b)$. We have

$$\begin{aligned}
|T_n| = \prod_{j=1}^{k_n}\sqrt{1+t^2Z_{n,j}^2} &= \prod_{r=1}^{J-1}\sqrt{1+t^2Z_{n,r}^2}\sqrt{1+t^2Z_{n,J}^2} \\
&\leq \exp\left(\frac{t^2}{2}\sum_{r=1}^{J-1} Z_{n,r}^2\right) \times (1 + |t| \cdot |Z_{n,J}|) \\
&\leq \exp(t^2) \times \left(1 + |t| \max_{1\leq j\leq k_n} |X_{n,j}|\right).
\end{aligned}$$

From the first assumption of the theorem, $\max_{1\leq j\leq k_n} |X_{n,j}|$ is uniformly integrable and so is T_n. The other assumptions of Lemma 13 are verified and we conclude that $S_n \hookrightarrow \mathcal{N}(0,1)$ from the convergence of characteristic functions.□

For stationary and ergodic martingale differences, we get the following important result.

Theorem 29. *Let $(X_j)_{j\in\mathbb{Z}}$ a square integrable stationary and ergodic sequence such that $\mathbb{E}[X_j|\mathcal{F}_{j-1}] = 0$ a.s., where $(\mathcal{F}_j)_{j\in\mathbb{Z}}$ is a filtration such X_j is $\mathcal{F}_j-$measurable for all $j \in \mathbb{Z}$.*

Then $S_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} X_j \hookrightarrow \mathcal{N}(0,\sigma^2)$, with $\sigma^2 = Var(X_1)$.

Proof of Theorem 29. First suppose that $\sigma^2 = 0$, then $X_j = 0$ a.s. for all $j \in \mathbb{Z}$ and then $S_n = 0$ a.s. which converges in distribution to $\mathcal{N}(0,0)$, that is the Dirac mass at point 0. Suppose now that $\sigma^2 > 0$. Setting $X_{n,j} = X_j/\sqrt{n\sigma^2}$, we will apply Theorem 28. From ergodicity and square integrability of $(X_j)_{j\in\mathbb{Z}}$, the second assumption of Theorem 28 is verified. To check the first assumption, we use the Cauchy-Schwarz inequality as well as

truncation. For a given $M > 0$, we have

$$\begin{aligned}
\frac{1}{\sqrt{n}}\mathbb{E}\left[\max_{1\le j\le n}|X_j|\right] &\le \sqrt{\mathbb{E}\left[\max_{1\le j\le n}\frac{|X_j|^2}{n}\right]} \\
&\le \sqrt{\frac{M}{n}} + \sqrt{\mathbb{E}\left[\frac{\max_{1\le j\le n}|X_j|^2\mathbb{1}_{|X_j|>M}}{n}\right]} \\
&\le \sqrt{\frac{M}{n}} + \sqrt{\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[|X_j|^2\mathbb{1}_{|X_j|>M}\right]} \\
&= \sqrt{\frac{M}{n}} + \sqrt{\mathbb{E}\left[|X_1|^2\mathbb{1}_{|X_1|>M}\right]}.
\end{aligned}$$

We then get

$$\overline{\lim_n}\,\mathbb{E}\left[\max_{1\le j\le n}|X_{n,j}|\right] \le \sigma^{-1}\mathbb{E}\left[|X_1|\mathbb{1}_{|X_1|>M}\right]$$

and we obtain the desired condition by letting $M \to \infty$. The result is then a consequence of Theorem 28.□

Next, we deduce a multivariate version of Theorem 29.

Corollary 5. *Let $(X_j)_{j\in\mathbb{Z}}$ a square integrable stationary and ergodic sequence, taking values in $\mathbb{R}^k$ and such that $\mathbb{E}[X_j|\mathcal{F}_{j-1}] = 0$ a.s., where $(\mathcal{F}_j)_{j\in\mathbb{Z}}$ is a filtration such X_j is $\mathcal{F}_j-$measurable for all $j \in \mathbb{Z}$.*

Then $S_n = \frac{1}{\sqrt{n}}\sum_{j=1}^n X_j \hookrightarrow \mathcal{N}_k(0,\Sigma)$, with $\Sigma = Var(X_1)$.

Proof of Corollary 5. From Lemma 15, it is easily seen that weak convergence of S_n to $\mathcal{N}_k(0,\Sigma)$is equivalent to weak convergence of $u^T S_n$ to $\mathcal{N}\left(0, u^T\Sigma u\right)$, for any vector $u \in \mathbb{R}^k$. But $u^T S_n = \frac{1}{\sqrt{n}}\sum_{j=1}^n u^T X_j$ and $\left(Z_j = u^T X_j\right)_{j\in\mathbb{Z}}$ is a martingale difference satisfying the assumptions of Theorem 29 with $\sigma^2 = u^T\Sigma u$. We then get the result.□

Example. We go back to the linear autoregressive process AR(p),

$$X_t = \sum_{j=1}^{p} a_{0,j}X_{t-j} + \varepsilon_t, \quad t \in \mathbb{Z},$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a sequence of i.i.d. random variables with mean 0 but now with finite positive variance v. We assume that the roots of the polynomial $\mathcal{P}$ defined by $\mathcal{P}(z) = 1 - \sum_{j=1}^p a_{0,j}z^j$ are outside the unit disc. In this case, we have already seen there is a unique stationary solution which writes as $X_t = \sum_{j\ge 0} s_j\varepsilon_{t-j}$ with $s_0 = 1$ and $(s_j)_{j\ge 0}$ has a geometric decay. In particular, the series is also converging normally in $\mathbb{L}^2$ and X_t is square integrable.

Our aim here is to estimate $\theta_0 = (a_{0,1}, \ldots, a_{0,p})^T$. As seen previously, the least squares estimator is defined by

$$\hat{\theta}_n = \left(\frac{1}{n} \sum_{t=p+1}^{n} \mathcal{X}_{t-1} \mathcal{X}_{t-1}^T \right)^{-1} \frac{1}{n} \sum_{t=p+1}^{n} \mathcal{X}_{t-1} X_t.$$

We have

$$\sqrt{n} \left(\hat{\theta}_n - \theta_0 \right) = \left(\frac{1}{n} \sum_{t=p+1}^{n} \mathcal{X}_{t-1} \mathcal{X}_{t-1}^T \right)^{-1} \frac{1}{\sqrt{n}} \sum_{t=p+1}^{n} \mathcal{X}_{t-1} \varepsilon_t.$$

Using Corollary 5, it is easily seen that $\frac{1}{\sqrt{n}} \sum_{t=p+1}^{n} \mathcal{X}_{t-1} \varepsilon_t \hookrightarrow \mathcal{N}_p(0, v\Gamma)$ where $\Gamma = \mathbb{E}\left[X_1 X_1^T\right]$. Slutsky's lemma ensures that $\sqrt{n} \left(\hat{\theta}_n - \theta_0 \right) \hookrightarrow \mathcal{N}_p(0, v\Gamma^{-1})$.

From Theorem 28, we also deduce the following convergence result for triangular arrays of independent random variables. The following result was used in Chapter 2.

Theorem 30. *Let $\{Y_{n,i} : 1 \leq i \leq k_n, n \geq 1\}$ be a triangular array of centered random variables, taking values in $\mathbb{R}^k$ and such that $Y_{n,1}, \ldots, Y_{n,k_n}$ are mutually independent. Suppose that the two following assumptions are fulfilled.*

1. *There exists a symmetric and semi-definite positive matrix V of size $k \times k$ such that $\lim_{n\to\infty} \sum_{i=1}^{k_n} Var\,(Y_{n,i}) = V$.*
2. *For any $\varepsilon > 0$, $\lim_{n\to\infty} \sum_{i=1}^{k_n} \mathbb{E}\left[\|Y_{n,i}\|^2 \mathbb{1}_{\|Y_{n,i}\|^2 > \varepsilon} \right] = 0$.*

Then $S_n := \sum_{i=1}^{k_n} Y_{n,i} \hookrightarrow \mathcal{N}_k(0, V)$.

Proof of Theorem 30. We start with the case $k = 1$ and $V = 1$. For any $\varepsilon > 0$, using the decomposition $Y_{n,i}^2 = Y_{n,i}^2 \mathbb{1}_{|Y_{n,i}| \leq \varepsilon} + Y_{n,i}^2 \mathbb{1}_{|Y_{n,i}| > \varepsilon}$ and bounding the maximum by the sum, we get the bound

$$\mathbb{E}\left[\max_{1 \leq i \leq k_n} Y_{n,i}^2 \right] \leq \varepsilon + \sum_{i=1}^{k_n} \mathbb{E}\left[Y_{n,i}^2 \mathbb{1}_{|Y_{n,i}| > \varepsilon} \right].$$

From the second assumption of the theorem, we conclude that $\max_{1 \leq i \leq k_n} |Y_{n,i}|$ converges to 0 in $\mathbb{L}^2$ and then in $\mathbb{L}^1$. The first assumption of Theorem 28 is checked.

To check the second one, we set $Z_{n,i} = Y_{n,i}^2 \mathbb{1}_{|Y_{n,i}| \leq \varepsilon}$ and $W_{n,i} = Y_{n,i}^2 \mathbb{1}_{|Y_{n,i}| > \varepsilon}$. From our second assumption $\lim_{n\to\infty} \sum_{i=1}^{k_n} W_{n,i} = 0$ in $\mathbb{L}^1$ and then in probability. Then

$$\sum_{i=1}^{k_n} Y_{n,i}^2 = \sum_{i=1}^{k_n} \left[Z_{n,i} - \mathbb{E} Z_{n,i} \right] + \sum_{i=1}^{k_n} \mathbb{E} Z_{n,i} + o_{\mathbb{P}}(1).$$

Since $\sum_{i=1}^{k_n} \mathbb{E} Z_{n,i} = \sum_{i=1}^{k_n} \mathbb{E} Y_{n,i}^2 - \sum_{i=1}^{k_n} \mathbb{E} W_{n,i} = 1 + o(1)$ and

$$\mathrm{Var}\left(\sum_{i=1}^{k_n} Z_{n,i} \right) = \sum_{i=1}^{k_n} \mathrm{Var}\,(Z_{n,i}) \leq \sum_{i=1}^{k_n} \mathbb{E} Z_{n,i}^2 \leq \epsilon^2 \sum_{i=1}^{k_n} \mathbb{E} Z_{n,i} = \epsilon^2 (1 + o(1)),$$

we conclude that $\sum_{i=1}^{k_n} Y_{n,i}^2$ goes to 1 in $\mathbb{L}^1$ and then in probability. From Theorem 28, we get the result in this case.

For the general case, we use Lemma 15. Let $u \in \mathbb{R}^k$. If $u^T V u = 0$, it is easy to show that $\sum_{i=1}^{k_n} u^T Y_{n,i}$ goes to 0 in $\mathbb{L}^2$ and then in distribution. Now if $u^T V u \neq 0$, we set $X_{n,i} = \frac{u^T Y_{n,i}}{\sqrt{u^T V u}}$. Applying the result for $k = 1$ and $V = 1$, we get $\sum_{i=1}^{k_n} X_{n,i} \hookrightarrow \mathcal{N}(0,1)$. This means that $\sum_{i=1}^{k_n} u^T Y_{n,i} \hookrightarrow \mathcal{N}\left(0, u^T V u\right)$. This completes the proof. □

4.5 A more general central limit theorem for stationary sequences

Suppose that $(X_t)_{t\in\mathbb{Z}}$ is a general stationary and ergodic sequence. We have already seen in the previous sections that for many semi-parametric conditional models, convergence in distribution of M-estimators of finite-dimensional parameters simply requires a central limit theorem for martingale differences. However, some simple semi-parametric estimation problems are excluded from this framework. This is for instance the case for estimating the population mean $\theta_0 = \mathbb{E}(X_1)$ for which a natural estimator is the empirical mean $\overline{X}_n = \frac{1}{n}\sum_{j=1}^n X_j$. However $(X_j)_{j\in\mathbb{Z}}$ might not be a martingale difference here (for instance AR processes are not martingale differences in general).

In general, stationarity and ergodicity are not enough to extend the Central Limit Theorem. There exist numerous stochastic dependence measures based on covariances type inequalities and which can be used to extend the CLT. See for instance the book Dedecker et al. (2007) for a survey of many existing dependence measures developed in this sense. In this course, we simply give a CLT based on a projective criterion which allows to deduce the convergence from the CLT for martingale differences. The proof of the following result is taken from Billingsley (2013), Theorem 19.1. In what follows, for a random variable Y, we set $\|Y\|_2 = \sqrt{\mathbb{E}\left(Y^2\right)}$.

Theorem 31. *Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary and ergodic sequence such that $\mathbb{E}X_0^2 < \infty$ and $\mathbb{E}(X_0) = 0$. Suppose that*

$$\sum_{n=1}^{\infty} \|\mathbb{E}\left(X_n|\mathcal{F}_0\right)\|_2 < \infty, \tag{4.6}$$

with $\mathcal{F}_0 = \sigma\left(X_t : t \leq 0\right)$.

Then the series $\sigma^2 = Var\left(X_0\right) + 2\sum_{n=1}^{\infty} Cov(X_0, X_n)$ is absolutely converging. Moreover,

$$S_n := \frac{1}{\sqrt{n}}\sum_{t=1}^{n} X_t \hookrightarrow \mathcal{N}\left(0, \sigma^2\right).$$

Notes

1. If $m := \mathbb{E}(X_1) \neq 0$, one can try to apply the result to $\overline{X}_j = X_j - m$. Note that condition (4.6) already implies that $\mathbb{E}(X_1) = 0$. Indeed,

$$\mathbb{E}(X_1) = \mathbb{E}(X_n) = \mathbb{E}\left[\mathbb{E}\left(X_n|\mathcal{F}_0\right)\right] \overset{n\to\infty}{\to} 0,$$

because (4.6) ensures the convergence of $\mathbb{E}\left(X_n|\mathcal{F}_0\right)$ to 0 in $\mathbb{L}^2$ and then in $\mathbb{L}^1$.

2. In some sense, condition (4.6) ensures that $\mathbb{E}\left(X_n\mathcal{F}_0\right)$ is close to $\mathbb{E}(X_n)$ sufficiently fast. Note that for i.i.d. integrable random variables or more generally martingale differences, this conditional expectation is equal to 0.

3. As for martingale differences, one can easily obtain a multivariate version of Theorem 31. This is left as an exercise.

4. Checking (4.6) for Bernoulli shifts $X_t = H\left(\varepsilon_t, \varepsilon_{t-1}, \ldots\right)$ can be done as follows. Suppose that $\mathbb{E}(X_1) = 0$ (otherwise center the variables). Let $\left(\varepsilon'_t\right)_{t\in\mathbb{Z}}$ a sequence of i.i.d. random variables independent of $(\varepsilon_t)_{t\in\mathbb{Z}}$ and with the same distribution. For a positive integer n, let

$$\delta_n = \|X_n - X'_n\|_2, \quad X'_n := H\left(\varepsilon_n, \ldots, \varepsilon_1, \varepsilon'_{-1}, \varepsilon_{-2}, \ldots\right).$$

Suppose that

$$\sum_{n\geq 1} \delta_n < \infty. \tag{4.7}$$

Since X'_n is independent of $\mathcal{F}_0$, we have $\mathbb{E}\left(X_n|\mathcal{F}_0\right) = \mathbb{E}\left(X_n - X'_n|\mathcal{F}_0\right)$. Moreover,

$$\|\mathbb{E}\left(X_n|\mathcal{F}_0\right)\|_2 = \|\mathbb{E}\left(X_n - X'_n|\mathcal{F}_0\right)\|_2 \leq \|X_n - X'_n\|_2 = \delta_n,$$

where we used Jensen's inequality for conditional expectations. Then (4.6) is valid.

The stochastic recursions of Theorem 26 satisfy (4.7) when $p \geq 2$. In this case, one can show that $\delta_n \leq C\kappa^n$ for some $C > 0$ and $\kappa \in (0, 1)$.

Proof of Theorem 31. The absolute convergence of series of autocovariances follows from (4.6) and the bound

$$|\mathbb{E}(X_0X_n)| = |\mathbb{E}\left(X_0\mathbb{E}\left(X_n|\mathcal{F}_0\right)\right)| \leq \|X_0\|_2 \cdot \|\mathbb{E}\left(X_n\mathcal{F}_0\right)\|_2.$$

For $h \in \mathbb{Z}$, set $\gamma(h) = \mathrm{Cov}(X_0, X_h) = \mathbb{E}(X_0X_h)$. Note that from stationarity, $\mathrm{Cov}(X_t, X_s) = \gamma(t-s)$ for any $s, t \in \mathbb{Z}$. We get

$$\mathbb{E}\left[S_n^2\right] = \frac{1}{n}\sum_{s,t=1}^{n} \mathrm{Cov}(X_t, X_s) = \frac{1}{n}\sum_{s,t=1}^{n} \gamma(t-s) = \frac{1}{n}\sum_{h=-n+1}^{n-1} (n - |h|)\,\gamma(h),$$

which yields to

$$\mathbb{E}\left[S_n^2\right] = \gamma(0) + 2\sum_{h\geq 1} \mathbb{1}_{h\leq n-1}\left(1 - \frac{|h|}{n}\right)\gamma(h) \to \gamma(0) + 2\sum_{h=1}^{\infty}\gamma(h),$$

as $n \to \infty$, using the dominated convergence theorem.

For $k \in \mathbb{Z}$, set

$$Z_k = X_k - \mathbb{E}\left(X_k|\mathcal{F}_{k-1}\right) + \sum_{i\geq 1}\left\{\mathbb{E}\left(X_{k+i}|\mathcal{F}_k\right) - \mathbb{E}\left(X_{k+i}|\mathcal{F}_{k-1}\right)\right\}.$$

Note that the series in the definition of Z_k converges in $\mathbb{L}^2$, using (4.6) and Lemma 14 below. Moreover $\mathbb{E}\left[Z_k|\mathcal{F}_{k-1}\right] = 0$ a.s. and Lemma 14 and (4.6) ensure that $(Z_k)_{k\in\mathbb{Z}}$ is a stationary and ergodic sequence of martingale differences, adapted to the filtration $(\mathcal{F}_k)_{k\in\mathbb{Z}}$. Moreover,

$$Z_k = X_k + \Delta_k - \Delta_{k-1}, \quad \Delta_k = \sum_{i=1}^{\infty}\mathbb{E}\left[X_{k+i}|\mathcal{F}_k\right]$$

and

$$\sum_{k=1}^{n} Z_k = \sum_{k=1}^{n} X_k + \Delta_n - \Delta_0,$$

with $\|\Delta_n\|_2 = \|\Delta_0\|_2 < \infty$. We then conclude that

$$\|S_n - \frac{1}{\sqrt{n}}\sum_{k=1}^{n} Z_k\|_2 = o(1).$$

From Theorem 5, we know that $\frac{1}{\sqrt{n}}\sum_{k=1}^{n} Z_k \hookrightarrow \mathcal{N}\left(0, \mathrm{Var}\left(Z_1\right)\right)$. We then conclude that S_n has the same limit. But since $\lim_{n\to\infty}\mathbb{E}\left(S_n^2\right) = \sigma^2$, necessarily, $\mathrm{Var}\left(Z_1\right) = \sigma^2$. The proof of the theorem is now complete.□

4.6 Appendix

Lemma 14. *Suppose that $(X_t)_{t\in\mathbb{Z}}$ is a stationary process such that $\mathbb{E}\left[|X_1|\right] < \infty$. For any $i \geq 1$, there exists a measurable mapping $g_i : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ such that for all $k \in \mathbb{Z}$, $\mathbb{E}\left[X_{k+i}|\mathcal{F}_k\right] = g_i\left(X_k, X_{k-1}, \ldots\right)$. Additionally, if $(X_t)_{t\in\mathbb{Z}}$ is ergodic, then any process of the form $Y_t = g\left(X_t, X_{t-1}, \ldots\right)$ where $g : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ is a measurable mapping, is also stationary and ergodic.*

Proof of Lemma 14. From Doob's theorem, there exists a measurable mapping $g_i : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ such that $\mathbb{E}\left[X_i|\mathcal{F}_0\right] = g_i\left(X_0, X_{-1}, \ldots\right)$ a.s. Moreover if $H : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ is a measurable and bounded mapping, we have, setting $H_k = H\left(X_k, X_{k-1}, \ldots\right)$,

$$\mathbb{E}\left[g_i\left(X_k, X_{k-1}, \ldots\right) H_k\right] = \mathbb{E}\left[g_i\left(X_0, X_{-1}, \ldots\right) H_0\right] = \mathbb{E}\left[X_i H_0\right] = \mathbb{E}\left[X_{k+i} H_k\right],$$

where we used stationary properties of the process $(X_t)_{t\in\mathbb{Z}}$. From the characterization of the conditional expectation, we get the first part of the lemma. The invariance of the stationarity and ergodicity properties after composition with the measurable mapping g is straightforward and left as an exercise.□

Lemma 15 (Cramér-Wold device)**.** *A sequence of random vectors $(X_n)_{n\in\mathbb{N}}$, taking values in $\mathbb{R}^k$, converges in distribution to a random vector X if and only if*

$$\forall \quad u \in \mathbb{R}^k n, \quad u^T X_n \hookrightarrow u^T X.$$

Proof of Lemma 15. This is a consequence of the equivalence between convergence in distribution and convergence of characteristic functions.□

Bibliography

Adelin Albert and John A Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

Jérôme Dedecker, Paul Doukhan, Gabriel Lang, León R José Rafael, Sana Louhichi, Clémentine Prieur, Jérôme Dedecker, Paul Doukhan, Gabriel Lang, León R José Rafael, et al. *Weak dependence*. Springer, 2007.

Richard M Dudley. *Uniform central limit theorems*, volume 142. Cambridge university press, 2014.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Christian Francq and Jean-Michel Zakoian. *GARCH models: structure, statistical inference and financial applications*. John Wiley & Sons, 2019.

Wenjiang Fu and Keith Knight. Asymptotics for lasso-type estimators. *The Annals of statistics*, 28(5):1356–1378, 2000.

Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. 2011.

Peter Hall. Large sample optimality of least squares cross-validation in density estimation. *The Annals of Statistics*, pages 1156–1174, 1983.

Kengo Kato. Asymptotics for argmin processes: Convexity arguments. *Journal of Multivariate Analysis*, 100(8):1816–1829, 2009.

Donald L McLeish. Dependent central limit theorems and invariance principles. *the Annals of Probability*, 2(4):620–628, 1974.

Wojciech Niemiro. Asymptotics for m-estimators defined by convex minimization. *The Annals of Statistics*, pages 1514–1533, 1992.

Karl E Petersen. *Ergodic theory*, volume 2. Cambridge university press, 1989.

Charles J Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, pages 1285–1297, 1984.

Alexandre B Tsybakov. Introduction to nonparametric estimation, 2009. *URL https://doi.org/10.1007/b13794. Revised and extended from the*, 9(10), 2004.

R Tyrrell Rockafellar. Convex analysis. *Princeton mathematical series*, 28, 1970.

AW van der Vaart and Jon A Wellner. Empirical processes. In *Weak Convergence and Empirical Processes: With Applications to Statistics*, pages 127–384. Springer, 2023.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wei Biao Wu and Xiaofeng Shao. Limit theorems for iterated random functions. *Journal of Applied Probability*, 41(2):425–436, 2004.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.