M2 – Sorbonne Université
2024 – 2025

Bayesian nonparametrics

Lecturer

Ismaël Castillo

Contents

1 Introduction **7**

1 Basics of statistics . . . 8

2 Nonparametric models . . . 10

3 Conditioning and Bayes' formula . . . 13

4 Frequentist analysis of Bayesian methods . . . 16

5 A first nonparametric example . . . 17

2 Convergence rates, general principles **21**

1 Setup and objectives . . . 21

2 A first useful lemma . . . 22

3 A generic result, first version . . . 25

4 Testing and entropy . . . 27

5 Extension to the non–iid framework . . . 31

5.1 Gaussian sequence model . . . 32

5.2 Gaussian white noise and nonparametric regression . . . 33

6 A generic result, second version . . . 34

7 A first application: random histograms with given number of jumps . . . 35

7.1 Basic histogram facts . . . 36

7.2 Verifying the conditions of Theorem 3 . . . 36

8 Complements . . . 38

8.1 Refinements of conditions in generic results and extensions . . . 38

8.2 Lower bounds . . . 39

8.3 Entropy of unit ball in $\mathbb{R}^k$ . . . 39

3 Nonparametric Bayes and adaptation **41**

1 Random histograms . . . 42

1.1 White noise model . . . 42

1.2 Density estimation . . . 43

2 Gaussian processes . . . 44

2.1 Definitions and examples . . . 44

2.2 RKHS of a Gaussian process . . . 45

2.3 Fundamental properties via RKHS and concentration function 48
2.4 Pre-concentration theorem 50
2.5 Application: posterior rates with a Brownian motion prior 55
2.6 Application: posterior rates for other Gaussian processes 56
2.7 Adaptive inference using Gaussian processes 57
3 Bayesian methods for high–dimensional models 58
3.1 Introduction 58
3.2 Sparse priors 59
3.3 Posterior convergence for spike–and–slab priors 62
3.4 Applications to inference in high-dimensional models 67
3.5 Complement 68

4 Deep Bayes 75
1 Introduction to neural networks 76
2 Prior distributions on DNNs and posterior concentration 79
2.1 Prior distribution 79
2.2 Posterior contraction 80
2.3 Proof of posterior concentration 80
3 Complement: generic properties of DNNs 82
3.1 Error propagation in a neural network and entropy 82
3.2 Approximation properties for Hölder functions 82
4 Deep Gaussian process priors 85
4.1 Motivation: compositional structures 85
4.2 Deep GPs: definition 87
4.3 Statement for deep GPs 90
5 Complement: Tempered posteriors 90

5 Variational Bayes 95
1 General principles 96
1.1 Divergences 96
1.2 Variational families and optimisation 96
2 A generic result for variational posteriors 98
2.1 Statement 98
2.2 Sufficient conditions 100
2.3 Result for mean-field class 101
2.4 An example of application: the sequence model 101
2.5 A variant in probability 102
3 Proof of the variational theorems 103
4 High-dimensional regression 107

6 Complements 109
1 The Dirichlet process in a nutshell 109
1.1 Priors on distributions 109
1.2 The Dirichlet process 110
2 Mixtures priors for adaptive density estimation 111
2.1 Mixture prior 111
2.2 A bit of context: traditional kernel density estimators 111

2.3 Bayesian mixtures of Gaussians: main result 112
2.4 Structure of the proof 113
2.5 Approximation by continuous mixtures of Gaussians 113
2.6 Approximation by finite mixtures of Gaussians 116
2.7 Classical lemmas: moment matching and Caratheodory 120
2.8 Prior mass condition 121
2.9 End of the proof 123
2.10 Auxiliary lemmas 124

CHAPTER 1

Introduction

Bayesian nonparametrics is a topic at the confluence of statistics, probability and machine learning. As we are going to see, the Bayesian approach is very 'probabilistic' in nature: indeed, its main object of study, the posterior distribution, is a probability measure. This measure will be random, through its dependence on the data. Studying this measure enables one to answer statistical inference questions, such as estimation of unknown parameters, or construction of confidence sets.

The nonparametric Bayesian field is in rapid development: a theory of convergence rates has emerged in the last 20 years, with many mathematical questions still open, in particular regarding: rates for certain distances, for some classes of priors (e.g. based on deep neural networks), uncertainty quantification, high-dimensional models, multiple testing, as well as on computability of posteriors or approximations thereof, to name just a few. The case where the unknown parameter is a function f or a high-dimensional vector θ will interest us most in this course, but there are many other potentially interesting settings, where the unknown quantity is a (possibly high-dimensional) matrix, graph, manifold...

A first example: Bayesians draw unknowns at random. To fix ideas, suppose we observe

$$X_1, \dots, X_n \quad \text{iid}$$

from a distribution P_f of density f on the interval $[0,1]$. This is the so–called density estimation model on the unit interval. One statistical goal in this setting is estimating f. In the Bayesian approach, to be defined more formally in the next pages, the starting point is always to *draw at random* the unknown quantities in the model, here the density function f.

How does one draw a function 'at random'? Probability theory gives a precise meaning to this question: it is enough to put a distribution on spaces of functions and 'draw' from this law. Technically, there are several ways one can do so. Let us give a few examples

1. *random histograms*: for some heights h_k drawn at random, one can set

$$f \sim \sum_{k=1}^{K} h_k \mathbb{1}_{I_k},$$

where $I_1, \dots, I_k$ form a partition (either fixed or random) of $[0,1]$

2. *random expansions on a basis*: for $\{\varphi_k\}$ a basis of $L^2[0,1]$, let us set, for (σ_k) a sequence in ℓ^2,

$$f \sim \sum_{k=1}^{K} \sigma_k \alpha_k \varphi_k,$$

where α_k are, say iid $\mathcal{N}(0,1)$ and K is either fixed (possibly $+\infty$) or itself drawn randomly.

3. *stochastic processes*: Brownian motion $(B_t)_{0\le t\le 1}$ for instance has sample paths in the set of continuous functions and is a special case of Gaussian processes commonly used in machine learning applications.

Coming back to the density estimation setting, one notes that the just mentioned random functions fs cannot be used directly, at least if one wishes to draw a 'density': indeed, the previous samples are not necessarily positive and do not need to integrate to 1. There are various ways to fix this: one can for instance renormalise and set, starting e.g. from Brownian motion (B_t),

$$Z_t = \frac{e^{B_t}}{\int_0^1 e^{B_u} du},$$

whose paths are now by construction (random) densities on $[0,1]$.

Posterior distributions: integrating the information from the data. The probability distribution (called the "prior") chosen on unknown quantities of the statistical model, is *updated* using the information contained in the data at hand through a conditioning operation. We will then get a conditional distribution, which is called *posterior distribution*. The more data we have, the more (hopefully) the posterior will 'learn' and the more 'informative' it will be to do inference on unknown parameters of our model.

The main difference with traditional estimators in classical statistics is that the estimator in the Bayesian approach is a whole (data–dependent) distribution, instead of a point in the parameter space (think of the maximum likelihood estimator). We will see examples below.

1 Basics of statistics

In statistics, the starting point is the data, often a sequence of observations, for instance in form of a numerical sequence $x_1, \dots, x_n$.

Statistical modelling consists in writing $x_i = X_i(\omega)$: data is interpreted as a realisation of random variables $X_1, \dots, X_n$.

Definition 1. A statistical experiment consists in

- a random object X taking values in a set E equipped with a σ–field $\mathcal{E}$.

- a collection of probability measures on $(E, \mathcal{E})$ called the *model*

$$\mathcal{P} = \{P_\theta,\ \theta \in \Theta\},$$

where Θ is a set called *set of parameters.*

Most of the time, X consists of a $n-$tuple $X = X^{(n)} = (X_1, \dots, X_n)$. In this case, the quantities E and $\mathcal{P}$ of the definition above also depend on n.

I.I.D. DATA When $X = X^{(n)} = (X_1, \dots, X_n)$, we will often assume that $P_\theta^{(n)} = P_\theta \otimes \cdots \otimes P_\theta = P_\theta^{\otimes n}$, that is, that the random variables $Y_1, \dots, Y_n$ are independend and identically distributed (i.i.d. in short).

Definition 2. A statistical model $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ is dominated if there exists a positive measure μ on E such that, for any $\theta \in \Theta$, P_θ admits a density p_θ with respect to μ, that is

$$dP_\theta(x) = p_\theta(x) d\mu(x).$$

Note that the measure μ should be the *same* for all $\theta \in \Theta$. And μ is then called *dominating measure.* In what follows, we shall always work with dominated models.

NOTATION. If X is a random variable of distribution Q, we write $X \sim Q$. This means that for any function g integrable with respect to Q, i.e. $g \in L^1(Q)$,

$$E_{X\sim Q}[g(X)] = E_Q[g(X)] = \int_E g(x) dQ(x).$$

If $Y \sim P_\theta$, we often write E_θ for $E_{Y\sim P_\theta}$. For a n iid observations as above, we write E_θ in place of $E_{P_\theta^{\otimes n}}$.

ESSENTIAL EXAMPLES.

$\boxed{1}$ The "FUNDAMENTAL MODEL" is

$$\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}.$$

It is a dominated model, for μ Lebesgue measure on $\mathbb{R}$,

$$dP_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} dx.$$

$\boxed{2}$ The DENSITY ESTIMATION model is

$$\mathcal{P} = \{P_f^{\otimes n},\ f \in \mathcal{F}\},$$

where $\mathcal{F}$ denotes a set of densities on, say, the unit interval $[0, 1]$, or e.g. on $\mathbb{R}$.

The "fundamental model" is the very special case where one restricts to densities of Gaussian distributions of unit variance.

Definition 3. A point estimator $\hat{\theta}(X)$ (or a 'statistic' $S(X)$) in a statistical experiment $(X, \mathcal{P})$ is a measurable function of X, most of the time assumed to take values in the set of parameters Θ.

Frequentist approach. In the frequentist approach, one assumes

$$\exists\, \theta_0 \in \Theta, \quad X \sim P_{\theta_0}$$

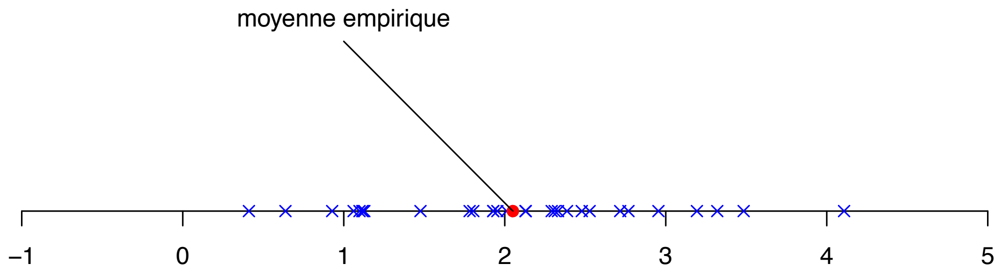
In this setting, θ_0 is called true value of the parameter. Typically, θ_0 is unknown and one tries to "estimate" it (i.e. to approach it) with the help of the data X.

Example (fundamental model). Suppose $X = (X_1, \dots, X_n)$ is generated from the model $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$ with a true $\theta_0 \in \mathbb{R}$:

$$(X_1, \dots, X_n) \sim \mathcal{N}(\theta_0, 1)^{\otimes n}.$$

Figure 1.1 represents $n = 30$ points randomly drawn from $\mathcal{N}(\theta_0, 1)$ and $\theta_0 = 2$. One notes that the sample stays fairly close to the value 2 and that the empirical mean ("moyenne empirique") is very close to 2.

Figure 1.1: Sample of size $n = 30$ from a $\mathcal{N}(\theta_0, 1)$ law.



Main inference questions

$\boxed{1}$ *Estimation.* The goal is to build an estimator $T(X_1, \dots, X_n)$ being close, in a sense to be made more specific (e.g. through a loss functions) of the true value θ_0 of the parameter θ.

$\boxed{2}$ *Confidence intervals/regions.* One wishes to construct $C = C(X_1, \dots, X_n)$ (random) subset of Θ such that $\theta_0 \in C(X_1, \dots, X_n)$ with high probability.

$\boxed{3}$ *Tests.* One wishes to answer by "true" or "false" to a given property of P_θ by constructing a 'test' $\varphi(X_1, \dots, X_n)$ taking values in $\{0, 1\}$.

2 Nonparametric models

Density estimation on $[0, 1]$. One observes $X = (X_1, \dots, X_n)$ with

$$X_i \sim \text{iid} \quad P_f,$$

with P_f the law of density f on $[0,1]$.

Nonparametric Gaussian regression. One observes $Y = (Y_1, \dots, Y_n)$, where, for $1 \le i \le n$,

$$Y_i = f(i/n) + \varepsilon_i,$$

with $f : [0,1] \longrightarrow \mathbb{R}$ and ε_i are iid $\mathcal{N}(0,1)$. That is, the model is

$$\mathcal{P} = \left\{ \bigotimes_{i=1}^{n} \mathcal{N}(f(i/n), 1),\ f \in \mathcal{G} \right\},$$

for $\mathcal{G}$ some set of functions (for instance continuous or Hölder). Let us note that the model is dominated by Lebesgue's measure $\mu^{(n)}$ on $\mathbb{R}^n$: for any $f \in \mathcal{G}$ and φ the Gaussian density,

$$dP_f^{(n)}(y_1, \dots, y_n) = \prod_{i=1}^{n} \varphi(y_i - f(i/n)) d\mu^{(n)}(y_1, \dots, y_n).$$

Gaussian sequence model. Suppose one observes a sequence $X = (X_1, \dots,)$, where, for $k \ge 1$ an integer, and $\theta = (\theta_k)_{k \ge 1}$ a square–integrable sequence,

$$X_k = \theta_k + \frac{\varepsilon_k}{\sqrt{n}}, \tag{1.1}$$

for (ε_k) a sequence of iid standard normal variables. This is a very popular model which can be seen as the 'basis' of nonparametric statistics. Observe that it is obtained by "piling–up" countably many times the elementary model

$$Y = \nu + \xi/\sqrt{n},$$

which can be seen as equivalent to the "fundamental model" $X \sim \mathcal{N}(\nu, 1)^{\otimes n}$ with $\nu \in \mathbb{R}$, through considering the sufficient statistic $Y = \overline{X} \sim \mathcal{N}(\nu, 1/n)$.

The Gaussian sequence model can be written, for $P_{\theta,k} = \mathcal{N}(\theta_k, 1/n)$,

$$\mathcal{P} = \left\{ P_\theta^{(n)} := \bigotimes_{k \ge 1} P_{\theta,k},\ \theta \in \ell^2 \right\}.$$

It can be shown that $P_\theta^{(n)}$ is absolutely continuous (and thus, has a density) with respect to the measure with signal $\theta = 0$ the null vector, with

$$\frac{dP_\theta^{(n)}}{dP_0^{(n)}}(X) = \exp\left\{ n \sum_{k=1}^{\infty} \theta_k X_k - n\|\theta\|_2^2/2 \right\}. \tag{1.2}$$

Truncated Gaussian sequence model. This is the same as before, but one observes only up to $k = n$, that is here $X = (X_1, \dots, X_k)$, with

$$X_k = \theta_k + \varepsilon_k/\sqrt{n}, \qquad 1 \le k \le n.$$

Statistically, for typical 'smoothness' classes the vector θ belongs to, not observing 'frequencies' after $k = n$ is rarely a big problem. Suppose for instance that θ belongs to a Sobolev ball, for $\beta, L > 0$,

$$S_\beta(L) = \left\{ \theta \in \ell^2 \ : \ \sum_{k=1}^{\infty} k^{2\beta} \theta_k^2 \leq L \right\}. \tag{1.3}$$

Then, if the measure of loss of an estimator T of θ is the quadratic loss $\|T - \theta\|_2^2$, the 'bias' incurred for basing T only on the first $(X_i)_{i \leq n}$ and setting $T_i = 0$ for $i > n$ is, if $\theta \in S_\beta(L)$,

$$\sum_{k>n} \theta_k^2 \leq n^{-2\beta} \sum_{k>n} k^{2\beta} \theta_k^2 \leq L n^{-2\beta}.$$

The rate $n^{-2\beta}$ is often much smaller than typical optimal rates in terms of the quadratic risk for smoothness β (often of the type $n^{-2\beta/(2\beta+1)}$).

Gaussian white noise model. For $f \in L^2[0,1]$ one observes $X^{(n)}$ where

$$dX^{(n)}(t) = f(t)dt + \frac{1}{\sqrt{n}} dW(t), \quad t \in [0,1],$$

for $W(t)$ standard Brownian motion on $[0,1]$.

There are two ways to interpret what is observed in this equation. In statistics we will use the second one mostly.

Observation scheme 1: trajectories (mostly used in stochastic process theory). One observes the trajectory

$$X^{(n)}(t) = \int_0^t f(u)du + \frac{1}{\sqrt{n}} W(t), \quad t \in [0,1].$$

Remark. Girsanov's theorem says that the distribution $P_f^{(n)}$ is absolutely continuous with respect to that where $f = 0$ (i.e. the distribution of $t \longrightarrow W(t)/\sqrt{n}$), namely

$$\frac{dP_f^{(n)}}{dP_0^{(n)}}(X) = \exp\left\{ n \int_0^1 f(u)dX(u) - \frac{n}{2} \int_0^1 f(u)^2 du \right\}.$$

Observation scheme 2: signal plus white noise (mostly used in statistics). One observes the Gaussian process $(\mathbb{X}^{(n)}(g),\ g \in L^2[0,1])$, indexed by square–integrable functions g. This means that one has access to the observation of the random variables

$$\mathbb{X}^{(n)}(g) = \int_0^1 g(t)dX^{(n)}(t) = \langle f, g \rangle_2 + \frac{1}{\sqrt{n}} \int_0^1 g(t)dW(t).$$

Note that $\mathbb{X}^{(n)}(g) \sim \mathcal{N}(\langle f, g \rangle_2, \|g\|_2/n)$.

One can also note that the Gaussian sequence model is just a particular case of the second observation scheme for the Gaussian white noise model, where for g's one takes the elements of an orthonormal basis (φ_k) of $L^2[0,1]$. Indeed, in that case one can set $\theta_k := \langle f, \varphi_k \rangle_2$, $X_k := \mathbb{X}^{(n)}(\varphi_k)$ and $\varepsilon_k = \int_0^1 \varphi_k(t)dW(t)$. Note that ε_k has law $\mathcal{N}(0, \|\varphi_k\|_2^2) = \mathcal{N}(0,1)$ and that ε_k's are independent, since

$$E\left[\int_0^1 \varphi_k(t)dW(t) \cdot \int_0^1 \varphi_l(t)dW(t) \right] = \int_0^1 \varphi_k(t)\varphi_l(t)dt = 1\!\!1_{k=l}.$$

3 Conditioning and Bayes' formula

Note. We shall define conditional distributions under a dominated framework, which covers already a huge variety of situations and many examples arising in practice. This enables one to apply Bayes' formula, in possibly infinite–dimensional contexts. A more general definition of conditional distributions is via 'desintegration', in the spirit of Proposition 2 below.

Dominated framework.

Let us consider

- a measurable set E equipped with a σ–field $\mathcal{E}$ and a space F equipped with a σ–field $\mathcal{F}$
- a positive σ-finite measure α on E and a positive σ-finite measure β on F
- a random variable X over E and a random variable Y over F.

Suppose the pair (X, Y) admits a density denoted $f(x, y)$ with respect to $\alpha \otimes \beta$, which we also write, if $P_{X,Y}$ denotes the law of the pair,

$$dP_{X,Y}(x, y) = f(x, y)d\alpha(x)d\beta(y).$$

Marginal distributions and densities

Proposition 1. In the above framework, the individual law of X, called marginal distribution of X, is the law P_X with density given by

$$f_X(x) = \int f(x, y)d\beta(y).$$

Proof.

For every g mesureable and bounded, Fubini's theorem gives

$$\begin{aligned} E[g(X)] &= \int\int g(x)f(x, y)d\alpha(x)d\beta(y) \\ &= \int g(x)\left[\int f(x, y)d\beta(y)\right]d\alpha(x) = \int g(x)f_X(x)d\alpha(x). \end{aligned}$$

Similarly, the marginal distribution of Y is the law P_Y on F whose density with respect to β is given by $f_Y(y) = \int f(x, y)d\alpha(x)$.

Conditional distribution.

Definition 4. The conditional distribution of Y given $X = x$ is the law of density, on F with respect to β, given by, for $f_X(x) > 0$,

$$f_{Y\,|\,X=x}(y) = \frac{f(x,y)}{\int f(x,y)d\beta(y)} = \frac{f(x,y)}{f_X(x)}.$$

We often denote $f(y\,|\,x)$ in place of $f_{Y\,|\,X=x}(y)$ when there is no risk of confusion. By definition, $y \longrightarrow f(y\,|\,x)$ is a density with respect to β, so that $\int f(y\,|\,x)d\beta(y) = 1$.

Definition 5. For real-valued Y, if $E[|Y|] < \infty$, we define the conditional expectation $E[Y\,|\,X]$ by

$$E[Y\,|\,X] = \int yf(y\,|\,X)d\beta(y).$$

More generally, for ϕ measurable with $\phi(Y)$ integrable,

$$E[\phi(Y)\,|\,X] = \int \phi(y)f(y\,|\,X)d\beta(y).$$

Proposition 2. For every measurable $h : E \times F \longrightarrow \mathbb{R}$, provided the variable $h(X,Y)$ is integrable,

$$E[h(X,Y)] = E[E[h(X,Y)\,|\,X]] = \int\int h(x,y)dP_{Y\,|\,X=x}(y)dP_X(x).$$

In particular, under the same conditions, if $h(X,Y) = \varphi(X)\psi(Y)$, for φ, ψ measurable,

$$E[\psi(Y)\varphi(X)] = E[E[\psi(Y)\,|\,X]\varphi(X)].$$

Proof.

$$\begin{aligned} E[h(X,Y)] &= \int\int h(x,y)f(x,y)d\alpha(x)d\beta(y) \\ &= \int\int h(x,y)\frac{f(x,y)}{f_X(x)}f_X(x)d\alpha(x)d\beta(y) \\ &= \int\left[\int h(x,y)dP_{Y\,|\,X=x}(y)\right]f_X(x)d\alpha(x), \end{aligned}$$

using Fubini's theorem for the last identity.

The Bayesian framework. Given a statistical model $\mathcal{P} = \{P_\theta^{(n)},\ \theta \in \Theta\}$ with data $X^{(n)}$, the Bayesian approach consists in, first, choosing a probability distribution Π on Θ, called the prior distribution.

In the following, we suppose we are in the following dominated framework: for μ, ν two sigma–finite measures, suppose

$$\begin{aligned} dP_\theta^{(n)} &= f_\theta^{(n)} d\mu \qquad \forall \theta \in \Theta, \\ d\Pi &= \pi d\nu. \end{aligned}$$

Note that the measure μ has to dominate all measures $P_\theta^{(n)}$, for any possible value of θ.

Second, the Bayesian setting assumes that the distributions for θ and $X^{(n)}$ are specified in such a way that

$$\begin{aligned} \theta &\sim \Pi, \\ X^{(n)} \,|\, \theta &\sim P_\theta^{(n)}. \end{aligned}$$

In this setting, the distribution of $(X^{(n)}, \theta)$ has density $(x^{(n)}, \theta) \longrightarrow f_\theta^{(n)}(x^{(n)})\pi(\theta)$ with respect to $\mu \otimes \nu$. We will always assume (without mentioning it) that this mapping is measurable for suitable choices of σ–fields on the space of X's and θ's, so that the next definition makes sense.

Definition 6. The posterior distribution, denoted $\Pi[\cdot \,|\, X^{(n)}]$, is the conditional distribution $\mathcal{L}(\theta \,|\, X^{(n)})$ of θ given $X^{(n)}$ in the Bayesian setting as above. It is a distribution on Θ, that depends on the data $X^{(n)}$. In the dominated framework as assumed above, it has a density with respect to ν given by Bayes' formula (i.e. the formula for conditional densities given previously)

$$\theta \longrightarrow \frac{f_\theta^{(n)}(X^{(n)})\pi(\theta)}{\int f_\theta^{(n)}(X^{(n)})\pi(\theta) d\nu(\theta)}.$$

Example: fundamental model. Consider the model $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n},\ \theta \in \mathbb{R}\}$. Suppose we take a normal prior $\Pi = \mathcal{N}(0, \sigma^2)$ on θ (with $\sigma^2 > 0$): this will make computations easy.
Exercise. Check that in this setting the posterior $\Pi[\cdot \,|\, X^{(n)}]$ is given by

$$\mathcal{L}(\theta \,|\, X^{(n)}) = \mathcal{N}\left(\frac{n\overline{X}}{n + \sigma^{-2}}, \frac{1}{n + \sigma^{-2}}\right).$$

One advantage of having a distribution (and not only a point estimate such as the MLE, or the posterior mean) is uncertainty quantification.

Definition 7. A credibility region of level (at least) $1 - \alpha$, for $\alpha \in [0, 1]$, is a measurable set $A \subset \Theta$ (typically depending on the data $A = A(X)$), such that

$$\Pi[A \,|\, X] = (\geq) 1 - \alpha.$$

Natural questions are: how does $\Pi[\cdot \mid X^{(n)}]$ behave as $n \to \infty$? Is there convergence? A limiting distribution? Are credibility regions linked in some way to confidence regions?

4 Frequentist analysis of Bayesian methods

Once one adopts the Bayesian approach to build the posterior distribution $\Pi[\cdot \mid X^{(n)}]$, one can study this distribution under the frequentist assumption that the data has actually been generated from a distribution in the model with fixed true parameter θ_0, that is

$$\exists \theta_0 \in \Theta, \quad X^{(n)} \sim P_{\theta_0}^{(n)}. \tag{1.4}$$

Consistency and convergence rate

Definition 8. Under the frequentist framework (1.4), for d a distance on the parameter set Θ, the posterior $\Pi[\cdot \mid X] = \Pi[\cdot \mid X_1, \dots, X_n]$ is

- consistent (for the distance d) at $\theta_0 \in \Theta$ if, for any $\varepsilon > 0$, as $n \to \infty$,

  $$\Pi\left[\, \{\theta \,:\, d(\theta, \theta_0) \le \varepsilon\} \mid X_1, \dots, X_n\right] \to 1,$$

  in probability under $P_{\theta_0}^{(n)}$.

- converges at rate ε_n (for the distance d) at $\theta_0 \in \Theta$ if, as $n \to \infty$,

  $$\Pi\left[\, \{\theta \,:\, d(\theta, \theta_0) \le \varepsilon_n\} \mid X_1, \dots, X_n\right] \to 1,$$

  in probability under P_{θ_0}.

✍ For Z_n a random variable with $0 \le Z_n \le 1$, one has

$$Z_n \xrightarrow{P} 0 \quad \Longleftrightarrow \quad E[Z_n] \to 0 \quad (n \to \infty),$$

and similarly $Z_n \to 1$ in probability iff $E[Z_n] \to 1$ (exercise).

In particular, to show that the posterior converges at rate ε_n for d, it is enough to show that

$$E_{\theta_0} \Pi\left[\, \{\theta \,:\, d(\theta, \theta_0) \le \varepsilon_n\} \mid X\right] \to 1 \quad (n \to \infty),$$

or a similar result with the complementary event and the corresponding expectation going to 0.

Example: fundamental model. In this case $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n}, \ \theta \in \mathbb{R}\}$. Under (1.4), we have $X \sim \mathcal{N}(\theta_0, 1)^{\otimes n}$ for a fixed unknown $\theta_0 \in \mathbb{R}$. By the law of large numbers (LLN), we have $\overline{X} \to \theta_0$ in probability (also almost surely). One can check that for a prior $\Pi = \mathcal{N}(0, 1)$, the posterior distribution $\mathcal{N}(n\overline{X}/(n+1), 1/(n+1))$ is consistent at θ_0 and converges at rate $M_n/\sqrt{n}$, for an arbitrary sequence $M_n \to \infty$ as $n \to \infty$. (exercise)

Limiting shape of the posterior distribution A natural question is whether the posterior $\Pi[\cdot\,|\,X^{(n)}]$ has a limiting shape when n goes to infinity. In the fundamental model with a $\mathcal{N}(0,1)$ prior on θ, one can prove that, for $\|\cdot\|_{TV}$ the total variation distance between probability distributions,

$$\|\Pi[\cdot\,|\,X] - \mathcal{N}(\overline{X}, 1/n)\|_{TV} \longrightarrow 0$$

in probability under $P_{\theta_0}^{(n)}$. This is a very special case of a much more general phenomenon, which can be viewed as a sort of Bayesian central limit theorem (although it deals with the posterior, a quantity in principle fairly more complex than the empirical average in the classical CLT), and called the *Bernstein–von Mises* theorem. In parametric models, under regularity conditions, this result states that

$$\|\Pi[\cdot\,|\,X] - \mathcal{N}(\hat{\theta}^{MLE}, I_{\theta_0}^{-1}/n)\|_{TV} \longrightarrow 0$$

in probability under $P_{\theta_0}^{(n)}$, where I_{θ_0} is the Fisher information matrix and θ^{MLE} the maximum likelihood estimator in the considered model (or any other 'efficient' estimator).

There exist nonparametric versions of this result, but they require some care to be defined, as the limit object is then typically infinite-dimensional.

5 A first nonparametric example

Model. Consider the Gaussian sequence model as above, with $X = (X_1, \ldots)$ and, for $k \geq 1$,

$$X_k = \theta_k + \varepsilon_k/\sqrt{n},$$

that is

$$P_\theta^{(n)} = \bigotimes_{k=1}^{\infty} \mathcal{N}(\theta_k, 1/n).$$

Prior. Suppose as a prior Π on θs one takes, for some $\alpha > 0$,

$$\Pi = \Pi_\alpha = \bigotimes_{k=1}^{\infty} \mathcal{N}(0, \sigma_k^2), \qquad \text{with } \sigma_k^2 := k^{-1-2\alpha}. \tag{1.5}$$

If working with infinite product distributions looks intimidating, one can just consider truncated versions of both model and prior at $k = n$. All what follows can then be computed in finite dimensions, see the exercise below.

Posterior distribution. Bayes' formula gives that the posterior distribution of θ_k given X only depends on X_k and

$$\mathcal{L}(\theta_k\,|\,X_k) \overset{\mathcal{D}}{=} \mathcal{N}\left(\frac{n}{n+\sigma_k^{-2}}X_k, \frac{1}{n+\sigma_k^{-2}}\right).$$

Furthermore, the complete posterior distribution of θ is

$$\Pi[\cdot\,|\,X] = \bigotimes_{k=1}^{\infty} \mathcal{N}\left(\frac{n}{n+\sigma_k^{-2}}X_k, \frac{1}{n+\sigma_k^{-2}}\right).$$

Exercise. Prove this for truncated versions at $k = n$ for model and prior distribution, using Bayes' formula.

The true θ_0. We assume the following smoothness condition, for some $\beta, L > 0$,

$$\theta_0 \in S_\beta(L) := \left\{ \theta \in \ell^2 \ : \ \sum_{k=1}^{\infty} k^{2\beta} \theta_k^2 \leq L \right\}. \tag{1.6}$$

Posterior convergence under θ_0. Considering a frequentist analysis of the posterior with a fixed truth θ_0, it is natural to wonder whether $\Pi[\cdot \,|\, X]$ is consistent at θ_0 and if so at which rate it converges for, say, the $\|\cdot\|_2^2$ loss, given by (setting $\|\cdot\| = \|\cdot\|_2$)

$$\|\theta - \theta'\|^2 = \sum_{k \geq 1} (\theta_k - \theta'_k)^2.$$

Let us consider the posterior mean, defined by

$$\overline{\theta}(X) = \int \theta d\Pi(\theta \,|\, X) = \left(\frac{nX_k}{n + \sigma_k^{-2}} \right).$$

First step: reduction to a mean/variance problem. Using Markov's inequality,

$$\begin{aligned} \Pi[\|\theta - \theta_0\| > \varepsilon_n \,|\, X] &\leq \frac{1}{\varepsilon_n^2} \int \|\theta - \theta_0\|^2 d\Pi(\theta \,|\, X) \\ &\leq \frac{1}{\varepsilon_n^2} \sum_{k \geq 1} \int (\theta_k - \theta_{0,k})^2 d\Pi(\theta \,|\, X). \end{aligned}$$

The "bias–variance decomposition" is (observe that the crossed term is zero because we have centered around the posterior mean)

$$\begin{aligned} \int (\theta_k - \theta_{0,k})^2 d\Pi(\theta \,|\, X) &= \int (\theta_k - \overline{\theta}_k)^2 d\Pi(\theta \,|\, X) + \int (\overline{\theta}_k - \theta_{0,k})^2 d\Pi(\theta \,|\, X) \\ &= \int (\theta_k - \overline{\theta}_k)^2 d\Pi(\theta \,|\, X) + (\overline{\theta}_k - \theta_{0,k})^2. \end{aligned}$$

as the last term does not depend on θ. Note that the first term in the last sum is $\mathrm{Var}(\theta_k \,|\, X_k)$. In order to show that, for some $\varepsilon_n = o(1)$ to be determined,

$$E_{\theta_0} \Pi[\|\theta - \theta_0\| > \varepsilon_n \,|\, X] = o(1),$$

it is enough to study the behaviour of the two terms

$$\begin{aligned} (a) &:= \sum_{k \geq 1} E_{\theta_0} \mathrm{Var}(\theta_k \,|\, X_k) \\ (b) &:= \sum_{k \geq 1} E_{\theta_0} (\overline{\theta}_k - \theta_{0,k})^2. \end{aligned}$$

Study of the terms (a) and (b). For both terms, we distinguish the regimes $\sigma_k^2 < 1/n$ and $\sigma_k^2 \geq 1/n$, or equivalently $k > N_\alpha$ and $k \leq N_\alpha$ respectively, with

$$N_\alpha := \lfloor n^{\frac{1}{1+2\alpha}} \rfloor.$$

We can now use the bounds

$$\begin{aligned}(a) &\le \sum_{k\ge1}\frac{1}{n+\sigma_k^{-2}}\\ &\le \sum_{k\le N_\alpha}\frac{1}{n}+\sum_{k>N_\alpha}\sigma_k^2 \le \frac{N_\alpha}{n}+CN_\alpha^{-2\alpha}\lesssim n^{-\frac{2\alpha}{2\alpha+1}}.\end{aligned}$$

For the second term, by using the explicit expression of $\overline{\theta}_k$, a little computation shows

$$\begin{aligned}E_{\theta_0}(\overline{\theta}_k-\theta_{0,k})^2 &= \frac{\sigma_k^{-4}}{(n+\sigma_k^{-2})^2}\theta_{0,k}^2+\frac{n}{(n+\sigma_k^{-2})^2}\\ &= \quad (I) \quad + \quad (II).\end{aligned}$$

The term (II) is the easiest to bound. Its sum is bounded by

$$\sum_{k\ge1}\frac{n}{n+\sigma_k^{-2}}\frac{1}{n+\sigma_k^{-2}}\le\sum_{k\ge1}\frac{1}{n+\sigma_k^{-2}}\lesssim n^{-\frac{2\alpha}{2\alpha+1}},$$

by the same reasoning as before. The sum of the term (I) is bounded by, with $a\vee b=\max(a,b)$,

$$\begin{aligned}\sum_{k\le N_\alpha}\frac{k^{2+4\alpha}}{n^2}\theta_{0,k}^2+\sum_{k>N_\alpha}\theta_{0,k}^2 &\le n^{-2}\sum_{k\le N_\alpha}k^{2+4\alpha-2\beta}k^{2\beta}\theta_{0,k}^2+\sum_{k>N_\alpha}k^{-2\beta}k^{2\beta}\theta_{0,k}^2\\ &\le n^{-2}\sum_{k\le N_\alpha}N_\alpha^{(2+4\alpha-2\beta)\vee0}k^{2\beta}\theta_{0,k}^2+N_\alpha^{-2\beta}L\\ &\le n^{-2}(N_\alpha^{2+4\alpha-2\beta}\vee1)L+N_\alpha^{-2\beta}L\lesssim(n^{-2}+N_\alpha^{-2\beta})L.\end{aligned}$$

Putting everything together one obtains the following

Theorem 1. In the Gaussian sequence model, consider a Gaussian prior Π_α as in (1.5) for $\alpha>0$. Then for any $\beta, L>0$, there exists $C=C(\alpha,L)$ such that

$$\sup_{\theta_0\in S(\beta,L)} E_{\theta_0}\int\|\theta-\theta_0\|_2^2 d\Pi_\alpha(\theta\,|\,X)\le C\varepsilon_n^2,\qquad \text{with } \varepsilon_n=\varepsilon_n(\alpha,\beta)=n^{-\frac{\alpha\wedge\beta}{2\alpha+1}}.$$

In particular, for any arbitrary sequence $M_n\longrightarrow\infty$ (as slowly as desired), as $n\longrightarrow\infty$,

$$\sup_{\theta_0\in S(\beta,L)} E_{\theta_0}\Pi_\alpha\left[\|\theta-\theta_0\|_2>M_n\varepsilon_n\,|\,X\right]=o(1).$$

Exercise. Using Jensen's inequality deduce from the first display in Theorem 1 that the posterior mean $\overline{\theta}(X)$ verifies, uniformly over $S(\beta,L)$,

$$E_{\theta_0}\|\overline{\theta}(X)-\theta_0\|_2^2\lesssim\varepsilon_n^2.$$

Interpretation and discussion. From the expression of the rate ε_n in Theorem 1 one notes that the fastest rate is obtained for the choice $\alpha = \beta$. This seems coherent: first, it can be checked that a draw from the prior $\Pi = \Pi_\alpha$ in (1.5) belongs to the Sobolev space $S_r = \{\theta = (\theta_k)\ :\ \sum_{k\geq 1} k^{2r}\theta_k^2 < \infty\}$ for any $r < \alpha$ (check that as an exercise), and thus can be seen as a (nearly) α–regular sequence. Now if the true θ_0 is β–regular, then choosing a prior distribution that 'matches' its regularity by setting $\alpha = \beta$ should indeed give good results. This, however, leads to the following question:

> What happens if the regularity parameter β is not known? (so that one cannot set $\alpha = \beta$)

We will see in these lectures that there are natural ways to choose a slightly different prior that leads to *adaptation*, namely to the construction of a posterior distribution that achieves a (near)–optimal rate without being given the knowledge of the regularity parameter β.

Regarding optimality, it can be shown that the rate $\varepsilon_n(\beta,\beta) = n^{-\beta/(2\beta+1)}$ (corresponding to choosing $\alpha = \beta$) is optimal in the minimax sense:

$$\inf_{\hat\theta}\ \sup_{\theta\in S(\beta,L)} \left(E_\theta\|\hat\theta - \theta\|_2^2\right)^{1/2} \asymp n^{-\frac{\beta}{2\beta+1}}.$$

This rate of convergence is a typical optimal rate in nonparametric problems: it is slower than the standard rate $1/\sqrt{n}$ common to (regular) parametric models. The larger β, the closer we are to a parametric rate.

CHAPTER 2

Convergence rates, general principles

1 Setup and objectives

Consider a nonparametric setting $\mathcal{P} = \{P_f^{(n)},\ f \in \mathcal{F}\}$, where f in a function in some class (e.g. square–integrable functions, densities...).

Following a Bayesian approach, we put a prior distribution Π on $(\mathcal{F}, \mathcal{B})$, where $\mathcal{F}$ is equipped with the σ–algebra $\mathcal{B}$,

$$X^{(n)} \,|\, f \sim P_f^{(n)} \tag{2.1}$$

$$f \sim \Pi. \tag{2.2}$$

Bayes' formula gives us an expression of the mass of any measurable set $B \in \mathcal{B}$ under the posterior distribution

$$\Pi[B \,|\, X^{(n)}] = \frac{\int_B p_f^{(n)}(X) d\Pi(f)}{\int p_f^{(n)}(X) d\Pi(f)}. \tag{2.3}$$

✍ Note that $\Pi[B] = 0$ always implies $\Pi[B \,|\, X^{(n)}] = 0$.

In what follows we study the behaviour of $\Pi[\cdot \,|\, X^{(n)}]$ in probability under $P_{f_0}^{(n)}$. We wish to show that, for some ε_n a sequence typically tending to 0 as $n \longrightarrow \infty$, for d a suitable distance over $\mathcal{F}$, as $n \longrightarrow \infty$,

$$E_{f_0} \Pi[d(f, f_0) > \varepsilon_n \,|\, X^{(n)}] = o(1).$$

What will be our target rate ε_n? This will depend on f_0, $\mathcal{F}$ and d. Often, we shall assume that f_0 belongs to some regularity set $S_\beta(L)$ (think of the Sobolev ball from the first chapter) and we will try to take ε_n to be of the order (or as close as possible to) of the minimax rate

$$\bar{\varepsilon}_n = \inf_T \sup_{f \in S_\beta(L)} E_f d(T, f),$$

where the infimum is taken over all possible estimators $T = T(X^{(n)})$ of f. For standard regularity classes and distances, $\bar{\varepsilon}_n$ will often be of the order $C(\beta, L)n^{-\beta/(2\beta+1)}$, possibly up to logarithmic factors.

[Here: Point estimators (if time allows)]

To fix ideas, let us first consider for now the density estimation model on the unit interval $[0,1]$, i.e.

$$P_f^{(n)} = P_f^{\otimes n}, \qquad dP_f(x) = f(x)dx, \ x \in [0,1]. \tag{2.4}$$

In the density model, $X^{(n)} = (X_1, \dots, X_n)$ and Bayes' formula can be written

$$\Pi[B \,|\, X_1, \dots, X_n] = \frac{\int_B \prod_{i=1}^n f(X_i) d\Pi(f)}{\int \prod_{i=1}^n f(X_i) d\Pi(f)} = \frac{\int_B \prod_{i=1}^n \frac{f}{f_0}(X_i) d\Pi(f)}{\int \prod_{i=1}^n \frac{f}{f_0}(X_i) d\Pi(f)}, \tag{2.5}$$

where we use that f_0 does not depend on the integrating variable f.
Technical remark: in order for the study of the ratio in the last display to make sense in probability under P_{f_0}, it will be silently assumed that $P_{f_0}[\int \prod_{i=1}^n f(X_i) d\Pi(f) > 0] = 1$, which will always be the case for the priors we shall consider.

2 A first useful lemma

Definition 1. Let us define, for densities f_0, f on $[0,1]$,

$$K(f_0, f) = \int \log \frac{f_0}{f} f_0$$
$$V(f_0, f) = \int \left(\log \frac{f_0}{f} - K(f_0, f) \right)^2 f_0.$$

and the Kullback–Leibler–type neighborhood

$$B_{KL}(f_0, \varepsilon_n) = \{ f \ : \ K(f_0, f) \le \varepsilon_n^2, \ V(f_0, f) \le \varepsilon_n^2 \}.$$

In the density model, we denote by E_{f_0} the expectation under the law $P_{f_0}^{\otimes n}$ and set $X = X^{(n)}$ for simplicity.

Lemma 1. Let A_n be a measurable set such that, if ε_n verifies $n\varepsilon_n^2 \longrightarrow \infty$,

$$\frac{\Pi[A_n]}{e^{-2n\varepsilon_n^2} \Pi[B_{KL}(f_0, \varepsilon_n)]} = o(1), \tag{2.6}$$

as $n \longrightarrow \infty$. Then we have, as $n \longrightarrow \infty$,

$$E_{f_0} \Pi[A_n \,|\, X] = o(1).$$

This gives a more refined version of the statement $\Pi[B] = 0$ implies $\Pi[B \mid X] = 0$ with 0 replaced by some suitable $o(1)$. The message is that if the prior distribution puts very little prior mass on some (sequence of) set(s), then the posterior distributions puts little mass over such set(s). To prove Lemma 1, we first prove yet another lemma.

Lemma 2. For any probability distribution Π on $\mathcal{F}$, for any $C, \varepsilon > 0$, with $P_{f_0}^{(n)}$ probability at least $1 - 1/(C^2 n\varepsilon^2)$,

$$\int \prod_{i=1}^{n} \frac{f}{f_0}(X_i) d\Pi(f) \geq \Pi[B_{KL}(f_0, \varepsilon)] e^{-(1+C)n\varepsilon^2}. \tag{2.7}$$

Proof of Lemma 1.

[in the proof we assume for simplicity that $f_0 > 0$. If this is not the case, one slightly adapts the proof, see below] As a preliminary remark, note that, since f is by definition a density,

$$E_{f_0}\left[\prod_{i=1}^{n} \frac{f}{f_0}(X_i)\right] = \int \prod_{i=1}^{n} \frac{f}{f_0}(x_i) \prod_{i=1}^{n} f_0(x_i) dx_i = \int \prod_{i=1}^{n} f(x_i) dx_1 \dots dx_n = 1.$$

Bayes' formula as in (2.3) for the set A_n, is $\Pi[A_n \mid X] = N/D$ with $D = \int \prod_{i=1}^{n} \frac{f}{f_0}(X_i) d\Pi(f)$. Lemma 2 implies, on an event E_n with probability at least $1 - (Cn\varepsilon^2)^{-1}$,

$$D \geq \Pi[B_{KL}(f_0, \varepsilon_n)] e^{-(1+C)n\varepsilon_n^2}.$$

Let us now bound N/D from above by

$$\frac{N}{D} \leq \frac{e^{-(1+C)n\varepsilon_n^2}}{\Pi[B_{KL}(f_0, \varepsilon_n)]} \int_{A_n} \prod_{i=1}^{n} \frac{f}{f_0}(X_i) d\Pi(f) 1\!\!1_{E_n} + 1\!\!1_{E_n^c},$$

where the bound for the last term is obtained noting that $N/D = \Pi[A_n \mid X] \leq 1$. Taking expectations (first note $1\!\!1_{E_n} \leq 1$), and invoking first Fubini's theorem and then the preliminary remark,

$$\begin{aligned} E_{f_0} \frac{N}{D} &\leq \frac{e^{-(1+C)n\varepsilon_n^2}}{\Pi[B_{KL}(f_0, \varepsilon_n)]} \int_{A_n} E_{f_0}\left[\prod_{i=1}^{n} \frac{f}{f_0}(X_i)\right] d\Pi(f) + P_{f_0} 1\!\!1_{E_n^c} \\ &\leq \frac{e^{-(1+C)n\varepsilon_n^2}}{\Pi[B_{KL}(f_0, \varepsilon_n)]} \Pi[A_n] + P_{f_0} 1\!\!1_{E_n^c}. \end{aligned}$$

Both terms in the last display go to 0 by assumption and Lemma 2 respectively.

[If f_0 possibly takes the value 0, consider the event $\mathcal{V}_n = \{\exists i : f_0(X_i) = 0\}$ and note $P_{f_0}[\mathcal{V}_n] \leq nP_{f_0}(f_0(X_1) = 0) = n \int 1\!\!1_{f_0(x)=0} f_0(x) dx = 0$. So since $D/N \leq 1$, it is enough to work with $(D/N) 1\!\!1_{\mathcal{V}_n^c}$. As $\mathcal{V}_n^c = \prod_i 1\!\!1_{f_0(X_i)>0}$,

$$E_{f_0}\left[\prod_{i=1}^{n} \frac{f}{f_0}(X_i) 1\!\!1_{f_0(X_i)>0}\right] = \int \prod_{i=1}^{n} \frac{f}{f_0}(x_i) \prod_{i=1}^{n} f_0(x_i) 1\!\!1_{f_0(x_i)>0} dx_i = \int \prod_{i=1}^{n} f(x_i) 1\!\!1_{f_0(x_i)>0} dx_1 \dots dx_n \leq 1,$$

where one uses $1\!\!1_{f_0(x_i)>0} \leq 1$ and the rest of the argument goes through as before.]

Proof of Lemma 2.

Let $B := B_{KL}(f_0, \varepsilon)$ and suppose $\Pi(B) > 0$ (otherwise the result is immediate). Let us denote $\overline{\Pi}(\cdot) = \Pi(\cdot \cap B)/\Pi(B)$. Next let us bound from below

$$\int \prod_{i=1}^{n} \frac{f}{f_0}(X_i) d\Pi(f) \geq \int_B \prod_{i=1}^{n} \frac{f}{f_0}(X_i) d\Pi(f) = \Pi(B) \int \prod_{i=1}^{n} \frac{f}{f_0}(X_i) d\overline{\Pi}(f).$$

As $\overline{\Pi}(\cdot)$ is a probability measure on B, Jensen's inequality applied to the logarithm gives

$$\begin{aligned}\log \int \prod_{i=1}^{n} \frac{f}{f_0}(X_i) d\overline{\Pi}(f) &\geq \sum_{i=1}^{n} \int_B \log \frac{f}{f_0}(X_i) d\overline{\Pi}(f) \\ &= -\sum_{i=1}^{n} \int_B \left[\log \frac{f_0}{f}(X_i) - KL(f_0, f)\right] d\overline{\Pi}(f) - n \int_B KL(f_0, f) d\overline{\Pi}(f) \\ &\geq -\sum_{i=1}^{n} Z_i - n\varepsilon^2,\end{aligned}$$

where we have set $Z_i = \int_B \left[\log \frac{f_0}{f}(X_i) - KL(f_0, f)\right] d\overline{\Pi}(f)$, and used the fact that on B, we have $KL(f_0, f) \leq \varepsilon^2$ by definition. We now use a simple concentration bound on the variables Z_is, which are independent under P_{f_0}. By Tchebychev's inequality

$$P_{f_0}\left[\left|\sum_{i=1}^{n} Z_i\right| > Cn\varepsilon^2\right] \leq \frac{1}{(Cn\varepsilon^2)^2} \mathrm{Var}_{f_0}\left[\sum_{i=1}^{n} Z_i\right].$$

By independence the last term is $n\mathrm{Var}_{f_0} Z_1$ and it is enough to bound

$$\begin{aligned}\mathrm{Var}_{f_0} Z_1 &= E_{f_0}\left[\left(\int_B \left[\log \frac{f_0}{f}(X_i) - KL(f_0, f)\right] d\overline{\Pi}(f)\right)^2\right] \leq E_{f_0}\left[\int_B \left[\log \frac{f_0}{f}(X_i) - KL(f_0, f)\right]^2 d\overline{\Pi}(f)\right] \\ &\leq \int_B V(f_0, f) d\overline{\Pi}(f) \leq \varepsilon^2 \overline{\Pi}(B) = \varepsilon^2,\end{aligned}$$

where we use Jensen's inequality with $t \longrightarrow t^2$ and the fact that $V(f_0, f) \leq \varepsilon^2$ on B. Let us now set

$$\mathcal{B}_n = \{\left|\sum_{i=1}^{n} Z_i\right| \leq Cn\varepsilon^2\}.$$

By combining the previous bounds, we have just proved that $P_{f_0}(\mathcal{B}_n^c) \leq n\mathrm{Var}_{f_0} Z_1/(Cn\varepsilon^2)^2 \leq 1/(C^2 n\varepsilon^2)$. The event $\mathcal{B}_n$ has as desired probability at least $1 - 1/(C^2 n\varepsilon^2)$ and on $\mathcal{B}_n$,

$$\log \int \prod_{i=1}^{n} \frac{f}{f_0}(X_i) d\overline{\Pi}(f) \geq -(C+1)n\varepsilon^2$$

which in turn implies, taking exponentials and renormalising by $\Pi(B)$,

$$\int \prod_{i=1}^{n} \frac{f}{f_0}(X_i) d\overline{\Pi}(f) \geq \Pi(B) e^{-(1+C)n\varepsilon^2}.$$

3 A generic result, first version

Let us start with a brief historical perspective. Doob (1949) showed that posteriors are (nearly) always consistent in a Π–almost sure sense, which is interesting but prior–dependent. Schwartz (1965) proved consistency in the sense of the definition above under some sufficient conditions of existence of certain tests and of enough prior mass around the true f_0. Diaconis and Freedman (1986) exhibited an example of seemingly natural prior whose posterior distribution is not consistent. Ghosal, Ghosh and van der Vaart (2000), Shen and Wasserman (2001) and Ghosal and van der Vaart (2007) gave sufficient conditions for rates of convergence.

We call *test* based on observations X a measurable function $\phi(X)$ taking values in $\{0, 1\}$.

Let us recall that for now we work with the density estimation model $\mathcal{P} = \{P_f^{\otimes n},\ f \in \mathcal{F}\}$. Let Π be a prior distribution on $(\mathcal{F}, \mathcal{B})$. Suppose also that $\mathcal{F}$ is equipped with a distance d (examples will be given below). We denote by $\mathcal{F} \setminus \mathcal{F}_n = \mathcal{F}_n^c$ the complement of $\mathcal{F}_n \subset \mathcal{F}$.

Theorem 1. [GGV, version with tests] Let (ε_n) be a sequence with $n\varepsilon_n^2 \longrightarrow \infty$ as $n \longrightarrow \infty$. Suppose there exist $C, M > 0$ and measurable sets $\mathcal{F}_n \subset \mathcal{F}$ such that

i) there exist tests $\psi_n = \psi_n(X)$ with

$$E_{f_0}\psi_n = o(1), \qquad \sup_{f\in\mathcal{F}_n:\ d(f,f_0)>M\varepsilon_n} E_f(1-\psi_n) \le e^{-(C+4)n\varepsilon_n^2},$$

ii)

$$\Pi[\mathcal{F} \setminus \mathcal{F}_n] \le e^{-n\varepsilon_n^2(C+4)},$$

iii)

$$\Pi[B_{KL}(f_0, \varepsilon_n)] \ge e^{-Cn\varepsilon_n^2}.$$

Then the posterior distribution converges at rate $M\varepsilon_n$ towards f_0: as $n \longrightarrow \infty$,

$$E_{f_0}\Pi[\{f\ :\ d(f, f_0) \ge M\varepsilon_n\}\,|\,X] = o(1).$$

Let us briefly comment on the conditions. Assumption iii) is natural: there should be enough prior mass around the true f_0. Indeed, recall by Lemma 1 above that if the prior mass of a set is too small, its posterior mass will be too: having a too small prior probability of the KL–neighborhood would mean its posterior mass is vanishing, so there could not be convergence at rate ε_n, at least in terms of the 'divergence' defined by the KL–type neighborhood.
Assumption ii) allows to work on a subset $\mathcal{F}_n$, so it gives some flexibility, especially if $\mathcal{F}$ is a 'large' set: indeed, combining ii) with iii)

$$\frac{\Pi[\mathcal{F} \setminus \mathcal{F}_n]}{\Pi[B_{KL}(f_0, \varepsilon_n)]} \le e^{-4n\varepsilon_n^2},$$

which leads to $E_{f_0}\Pi[\mathcal{F} \setminus \mathcal{F}_n\,|\,X] = o(1)$ using Lemma 1.

Assumption i) is so far a little more mysterious. It can be seen more as a ‘meta–condition’, that makes the proof of the result quite quick. We will see below another version of the result, where i) is replaced by another, more interpretable, condition. Let us just note that the distance d in i) is the same as in the result: one needs to find tests with respect to this distance.

Proof.

Since $E_{f_0}\Pi[\mathcal{F}\setminus\mathcal{F}_n\,|\,X]=o(1)$ as noted above, is is enough to prove that $E_{f_0}\Pi[\mathcal{C}_n\,|\,X]=o(1)$, where

$$\mathcal{C}_n=\{f\in\mathcal{F}_n,\ d(f,f_0)\geq M\varepsilon_n\}.$$

Using the tests ψ_n from Assumption i), one decomposes

$$\Pi[\mathcal{C}_n\,|\,X]=\Pi[\mathcal{C}_n\,|\,X]\psi_n+\Pi[\mathcal{C}_n\,|\,X](1-\psi_n).$$

With $\Pi[\mathcal{C}_n\,|\,X]\leq 1$, one gets $E_{f_0}\Pi[\mathcal{C}_n\,|\,X]\psi_n\leq E_{f_0}\psi_n=o(1)$ thanks to i). For the second term, we write, recalling $\psi_n=\psi_n(X_1,\ldots,X_n)=\psi_n(X)$ is a function of the data,

$$\Pi[\mathcal{C}_n\,|\,X](1-\psi_n)=\frac{\int_{\mathcal{C}_n}\prod_{i=1}^n\frac{f}{f_0}(X_i)(1-\psi_n(X)d\Pi(f)}{\int\prod_{i=1}^n\frac{f}{f_0}(X_i)d\Pi(f)}=:\frac{N}{D}.$$

In order to bound the denominator from below, let us introduce the event

$$\mathcal{B}_n=\left\{\int\prod_{i=1}^n\frac{f}{f_0}(X_i)d\Pi(f)\geq\Pi[B_{KL}(f_0,\varepsilon_n)]e^{-2n\varepsilon_n^2}\right\}.$$

Lemma 2 tells us that $P_{f_0}[\mathcal{B}_n]\geq 1-(n\varepsilon_n^2)=1-o(1)$ using $n\varepsilon_n^2\longrightarrow\infty$. Deduce, with $\mathcal{B}_n^c$ the complementary event of $\mathcal{B}_n$,

$$\frac{N}{D}\leq\frac{e^{2n\varepsilon_n^2}}{\Pi[B_{KL}(f_0,\varepsilon_n)]}\int_{\mathcal{C}_n}\prod_{i=1}^n\frac{f}{f_0}(X_i)(1-\psi_n(X)d\Pi(f)+\mathbb{1}_{\mathcal{B}_n^c}.$$

Observe, using a similar argument as in the proof of Lemma 1 (and again modulo adjustment in case f_0 can take the value 0 with = becoming ≤),

$$\begin{aligned}E_{f_0}\left[\prod_{i=1}^n\frac{f}{f_0}(X_i)(1-\psi_n(X)\right]&=\int\prod_{i=1}^n\frac{f}{f_0}(x_i)(1-\psi_n(x_1,\ldots,x_n))\prod_{i=1}^n f_0(x_i)dx_1\cdots dx_n\\&=\int(1-\psi_n(x_1,\ldots,x_n))\prod_{i=1}^n f(x_i)dx_1\cdots dx_n=E_f[1-\psi_n(X)].\end{aligned}$$

By taking expectations and using Fubini’s theorem,

$$\begin{aligned}E_{f_0}\frac{N}{D}&\leq\frac{e^{2n\varepsilon_n^2}}{\Pi[B_{KL}(f_0,\varepsilon_n)]}\int_{\mathcal{C}_n}\prod_{i=1}^n E_{f_0}\left[\frac{f}{f_0}(X_i)(1-\psi_n(X)\right]d\Pi(f)+P_{f_0}[\mathcal{B}_n^c]\\&\leq e^{(C+2)n\varepsilon_n^2}\int_{\mathcal{C}_n}\prod_{i=1}^n E_f\left[(1-\psi_n(X)\right]d\Pi(f)+P_{f_0}[\mathcal{B}_n^c]\\&\leq e^{(C+2)n\varepsilon_n^2}e^{-(C+4)n\varepsilon_n^2}+P_{f_0}[\mathcal{B}_n^c]\leq e^{-2n\varepsilon_n^2}+o(1)=o(1).\qquad\square\end{aligned}$$

Exercise (if time allows)

4 Testing and entropy

In Theorem 1, the testing condition i) requires to be able to test a 'point' f_0 versus the 'complement of a ball' $\{f \in \mathcal{F}_n,\ d(f, f_0) > M\varepsilon_n\}$. The latter set has not a very simple structure (one would prefer a ball for instance instead of a complement!). Let us see how one can simplify this through combining tests of 'point' versus 'ball'.

Testing condition (T). Suppose one can find constants $K > 0$ and $a \in (0, 1)$ such that for any $\varepsilon > 0$, if $f_0, f_1 \in \mathcal{F}$ are such that $d(f_0, f_1) > \varepsilon$, then there exist tests φ_n with

$$E_{f_0}\varphi_n \le e^{-Kn\varepsilon^2} \tag{2.8}$$

$$\sup_{f:\ d(f,f_1)<a\varepsilon} E_f(1-\varphi_n) \le e^{-Kn\varepsilon^2}. \tag{2.9}$$

This condition is in fact always verified for certain distances.

Definition 2. Let P, Q probability distributions dominated by a measure μ, i.e. $dP = p d\mu$ and $dQ = q d\mu$. The L^1–distance is defined as

$$\|P - Q\|_1 = \int |p - q| d\mu$$

and the Hellinger distance as

$$h(P, Q) = \left(\int (\sqrt{p} - \sqrt{q})^2 d\mu \right)^{1/2}.$$

These distances verify the following properties (left as an Exercise)

- $\|P - Q\|_1 \le 2$ and $h(P, Q) \le \sqrt{2}$.
- $\|P - Q\|_1 \le 2h(P, Q)$ [use Cauchy-Schwarz]
- If $\max(p, q) \ge c_0 > 0$ then $h(P, Q) \le C\|P - Q\|_1$ for some $C > 0$.
- Defining the total variation norm (between measures defined on a common σ–field $\mathcal{A}$) as $\|P - Q\|_{TV} = \sup_{A\in\mathcal{A}} |P(A) - Q(A)|$,

$$\|P - Q\|_1 = 2\|P - Q\|_{TV}.$$

Theorem 2. [Le Cam, Birgé] The testing condition (T) is always verified in the density estimation model for d the L^1–distance or the Hellinger distance h.

We prove this result below for the L^1–distance. For the Hellinger distance, we refer to the book by Ghosal and van der Vaart (2017), Proposition D.8.

Definition 3. The ε–covering number of a set $\mathcal{E}$ for the distance d, denoted $N(\varepsilon, \mathcal{E}, d)$, is the minimal number of d–balls of radius ε necessary to cover $\mathcal{E}$.

The entropy of a set measures its 'complexity'/'size'. Let us give a few examples

- If $\mathcal{E} = [0, 1]$ and $d(x, y) = |x - y|$, then $N(\varepsilon, \mathcal{E}, d)$ is of order $1/\varepsilon$.
- If $\mathcal{E}$ is the unit ball in $\mathbb{R}^k$

  $$B(0, 1) = \left\{ \theta \in \mathbb{R}^k,\ \|\theta\|_2^2 := \sum_{i=1}^{k} \theta_i^2 \leq 1 \right\},$$

  then $N(\varepsilon, \mathcal{E}, \|\cdot\|_2)$ is of order ε^{-k}. Note that this number grows exponentially with the dimension k. We prove this below.
- There are many results available for balls in various function spaces (histograms, Sobolev or Hölder balls etc.). Examples will appear in the sequel.

Lemma 3. Suppose that the testing condition (T) holds for a distance d on $\mathcal{F}$ and that, for a sequence of measurable sets $\mathcal{F}_n$, and a sequence (ε_n) with $n\varepsilon_n^2 \geq 1$,

$$\log N(\varepsilon_n, \mathcal{F}_n, d) \leq Dn\varepsilon_n^2.$$

Then for a given $a > 0$ there exists $M = M(c)$ large enough and tests $\psi_n = \psi_n(X)$ such that

$$E_{f_0}\psi_n = o(1), \qquad \sup_{f \in \mathcal{F}_n:\ d(f, f_0) > M\varepsilon_n} E_f(1 - \psi_n) \leq e^{-cn\varepsilon_n^2}.$$

Proof.

Let us consider the set

$$G_n = \{f \in \mathcal{F}_n,\ d(f, f_0) > 4M\varepsilon_n\}$$

and partition it in 'shells' C_j as follows

$$G_n = \bigcup_{j \geq 1} \{f \in \mathcal{F}_n,\ 4Mj\varepsilon_n < d(f, f_0) \leq 4M(j + 1)\varepsilon_n\} = \bigcup_{j \geq 1} C_j.$$