M2 – Sorbonne Université
2023 – 2024

Bayesian nonparametrics

Lecturer

Ismaël Castillo

Contents

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 $X = X^{(n)} = (X_1, \ldots, X_n)$. In this case, the quantities E and $\mathcal{P}$ of the definition above also depend on n.

I.I.D. DATA When $X = X^{(n)} = (X_1, \ldots, X_n)$, 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 $Y_1, \ldots, Y_n$ 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

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

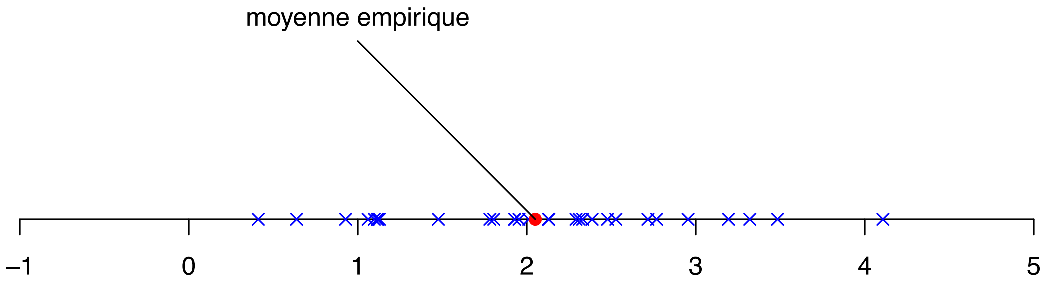
In this setting, θ_0 is called true value of the parameter. Typically, θ_0 is unknown and one tries to "estimate" it (i.e. to approach it) with the help of the data X.

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 $\theta = (\theta_k)_{k\ge1}$ 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 $P_{\theta,k} = \mathcal{N}(\theta_k, 1/n)$,

$$\mathcal{P} = \left\{ P_\theta^{(n)} := \bigotimes_{k\ge1} P_{\theta,k},\ \theta \in \ell^2 \right\}.$$

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

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

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

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

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 $\Pi[\cdot \,|\, X^{(n)}]$ behave as $n \longrightarrow \infty$? 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 $\Pi[\cdot \,|\, X^{(n)}]$, 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

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

Consistency and convergence rate

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

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

$$\Pi\left[\,\{\theta \,:\, d(\theta, \theta_0) \leq \varepsilon\} \,|\, X_1, \ldots, X_n\right] \longrightarrow 1,$$

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

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

$$\Pi\left[\,\{\theta \,:\, d(\theta, \theta_0) \leq \varepsilon_n\} \,|\, X_1, \ldots, X_n\right] \longrightarrow 1,$$

  in probability under P_{θ_0}.

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

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

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

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

$$E_{\theta_0} \Pi\left[\,\{\theta \,:\, d(\theta, \theta_0) \leq \varepsilon_n\} \,|\, X\right] \longrightarrow 1 \quad (n \longrightarrow \infty),$$

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 $\overline{X} \longrightarrow \theta_0$ 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 $M_n \longrightarrow \infty$ as $n \longrightarrow \infty$. (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

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

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

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

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

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

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

$$
\begin{aligned}
\sum_{k\leq N_\alpha} \frac{k^{2+4\alpha}}{n^2}\theta_{0,k}^2 + \sum_{k>N_\alpha} \theta_{0,k}^2 &\leq n^{-2} \sum_{k\leq N_\alpha} k^{2+4\alpha-2\beta} k^{2\beta} \theta_{0,k}^2 + \sum_{k>N_\alpha} k^{-2\beta} k^{2\beta} \theta_{0,k}^2 \\
&\leq n^{-2} \sum_{k\leq N_\alpha} N_\alpha^{(2+4\alpha-2\beta)\vee 0} k^{2\beta} \theta_{0,k}^2 + N_\alpha^{-2\beta} L \\
&\leq n^{-2}(N_\alpha^{2+4\alpha-2\beta} \vee 1)L + N_\alpha^{-2\beta} L \lesssim (n^{-2} + N_\alpha^{-2\beta})L.
\end{aligned}
$$

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

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

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, $X^{(n)} = (X_1, \ldots, X_n)$ and Bayes' formula can be written

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

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 $n \longrightarrow \infty$,

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

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,

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

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 $\mathcal{V}_n^c = \prod_i 1\!\!1_{f_0(X_i)>0}$,

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

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

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

Using the tests ψ_n from Assumption i), one decomposes

$$\Pi[C_n \,|\, X] = \Pi[C_n \,|\, X]\psi_n + \Pi[C_n \,|\, X](1-\psi_n).$$

With $\Pi[C_n \,|\, X] \leq 1$, one gets $E_{f_0}\Pi[C_n \,|\, X]\psi_n \leq E_{f_0}\psi_n = o(1)$ 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,

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

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 $\mathcal{B}_n$,

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

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,

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

□

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. $dP = pd\mu$ and $dQ = qd\mu$. 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

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

Now let us cover each shell C_j by balls.

- Let $\varepsilon = jM\varepsilon_n$ and consider a minimal covering of C_j by balls B_{ij} of radius $a\varepsilon$, for $a \in (0,1)$ the constant appearing in condition (T): by definition of the covering number, the number of these balls is $N(a\varepsilon, C_j, d)$.
- Let us denote by g_{ij} the centers of the balls of the previous covering. Since B_{ij} must intersect C_j (otherwise it could be removed from the covering which would then not be minimal), we have, as $a \in (0,1)$,

$$d(f_0, g_{ij}) \geq 4Mj\varepsilon_n - 2a\varepsilon = 4Mj\varepsilon_n - 2ajM\varepsilon_n \geq 2Mj\varepsilon_n > \varepsilon.$$

  So, for each g_{ij} there exists a test φ_{ij} satisfying the properties given by condition (T).
- On the other hand, we also have for any $j \geq 1$, if $M \geq a^{-1}$,

$$\begin{aligned} N(a\varepsilon, C_j, d) = N(ajM\varepsilon_n, C_j, d) &\leq N(ajM\varepsilon_n, \mathcal{F}_n, d) \\ &\leq N(\varepsilon_n, \mathcal{F}_n, d). \end{aligned}$$

- Let us now combine the just–contructed tests φ_{ij} by setting

$$\psi := \sup_{i,j\geq 1} \varphi_{ij}.$$

Let us now verify that the test ψ satisfies the desired properties. First, recalling $\varepsilon = jM\varepsilon_n$,

$$\begin{aligned} E_{f_0}\psi \leq E_{f_0}\left(\sum_{i,j}\varphi_{ij}\right) &\leq \sum_{j\geq 1}\sum_{i} E_{f_0}\varphi_{ij} \leq \sum_{j\geq 1} N(\varepsilon_n, \mathcal{F}_n, d)e^{-Kn\varepsilon^2} \\ &\leq \sum_{j\geq 1} N(\varepsilon_n, \mathcal{F}_n, d)e^{-KnM^2\varepsilon_n^2 j^2} \leq N(\varepsilon_n, \mathcal{F}_n, d)\frac{e^{-KnM^2\varepsilon_n^2}}{1-e^{-KnM^2\varepsilon_n^2}} \\ &\leq Ce^{cn\varepsilon_n^2 - KnM^2\varepsilon_n^2} \end{aligned}$$

which is $o(1)$ if $c < KM^2/2$ say. On the other hand, uniformly for $f \in \mathcal{F}_n$ such that $d(f, f_0) > 4M\varepsilon_n$,

$$\begin{aligned} E_f(1-\psi) \leq \sup_{j,i}\sup_{f\in B_{ij}} E_f(1-\psi) &\leq \sup_{j,i}\sup_{f\in B_{ij}} E_f(1-\varphi_{ij}) \\ &\leq \sup_{j,i} e^{-Kn(jM\varepsilon_n)^2} \leq e^{-Kn(M\varepsilon_n)^2} \leq e^{-cn\varepsilon_n^2} \end{aligned}$$

as soon as $KM^2 > c$, which concludes the proof.

Proof of Theorem 2 for $d = \|\cdot\|_1$.

Let f_0, f_1 two densities with $\|f_0 - f_1\|_1 > \varepsilon$. Let $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ denote the empirical measure associated to $X_1, \ldots X_n$ and for a measurable set $B \subset [0,1]$, let

$$\mathbb{P}_n(B) = \frac{1}{n}\sum_{i=1} \mathbb{1}_{X_i\in B}.$$

Let A denote the set, abbreviated as $A = \{f_0 < f_1\}$,

$$A = \{x : f_0(x) < f_1(x)\}$$

Let us define the test

$$\varphi_n = 1\!\!1 \left\{ \mathbb{P}_n(A) > P_{f_0}(A) + \frac{\|f_0 - f_1\|_1}{3} \right\}.$$

The term $E_{f_0}\varphi_n$, also called type I–error of the test, is bounded by

$$\begin{aligned} E_{f_0}\varphi_n &= P_{f_0}\left[\sum_{i=1}^n (1\!\!1_{X_i \in A} - P_{f_0}(A)) > \|f_0 - f_1\|_1/3 \right] \\ &\leq \exp\{-Cn\|f_0 - f_1\|_1^2\} \leq e^{-Cn\varepsilon^2}, \end{aligned}$$

where we use Hoeffding's inequality Lemma 4 and $\|f_0 - f_1\|_1 > \varepsilon$.

Let us now consider the term $E_f(1-\varphi_n)$, also called type II–error of the test, for fs in the ball $\{f : \|f - f_1\| < a\varepsilon\}$ with $a = 1/5$.

$$E_f(1-\varphi_n) = P_{f_0}\left[\mathbb{P}_n(A) - P_f(A) \leq P_{f_0}(A) - P_f(A) + \|f_0 - f_1\|_1/3\right]$$

We now claim that with f chosen as above the last display, the term $P_{f_0}(A) - P_f(A)$ is at most $-D\|f_0 - f_1\|_1$ for suitably large $D > 0$.

$$P_{f_0}(A) - P_f(A) = P_{f_0}(A) - P_{f_1}(A) + P_{f_1}(A) - P_f(A) =: (i) + (ii).$$

The choice of A ensures $(i) = -\|f_0 - f_1\|_1/2$ by Lemma 5, which also implies $|(ii)| \leq \|P_{f_1} - P_f\|_{TV} = \|f_1 - f\|_1/2 \leq a\varepsilon/2 = \varepsilon/10 \leq \|f_0 - f_1\|_1/10$. So $(i) + (ii) \leq -(2/5)\|f_1 - f_0\|_1$. As $-2/5 + 1/3 = -1/15$, one obtains

$$E_f(1-\varphi_n) \leq P_{f_0}\left[\mathbb{P}_n(A) - P_f(A) \leq -\|f_1 - f_0\|_1/15\right] \leq e^{-c'n\|f_1 - f_0\|_1^2} \leq e^{-c'n\varepsilon^2},$$

invoking Hoeffding's inequality (Lemma 4) again with $c' > 0$ a suitably small constant.

Lemma 4. [Hoeffding's inequality] Let Z_i be independent random variables with $a_i \leq Z_i \leq b_i$ for reals a_i, b_i and $1 \leq i \leq n$. Then

$$P\left[\sum_{i=1}^n Z_i > t\right] \leq \exp\left\{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}.$$

Proof.

See, e.g. Boucheron, Lugosi and Massart's book.

Lemma 5. [Total variation distance] Let P, Q be two probability measures defined on a joint $\sigma-$field $\mathcal{A}$ and dominated by μ, that is $dP = pd\mu$, $dQ = qd\mu$. The total variation distance $\|P-Q\|_{TV} = \sup_{A\in\mathcal{A}} |P(A) - Q(A)|$ verifies

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

Also, the supremum defining the total variation distance is attained for $A = \{x \ : \ q(x) < p(x)\}$.

Proof.

Let A denote the set $A = \{x \ : \ q(x) < p(x)\}$ and A^c its complement. Then

$$\begin{aligned}\|p-q\|_1 &= \int_A (p-q)d\mu + \int_{A^c}(q-p)d\mu - \int_{p=q}(q-p)d\mu \\ &= P(A) - Q(A) + Q(A^c) - P(A^c) = 2(P(A) - Q(A)) = 2\int_A (p-q)d\mu.\end{aligned}$$

By symmetry, one also has $\|p-q\|_1 = 2\int_{q>p}(q-p)d\mu$. On the other hand, for any $B \in \mathcal{A}$,

$$P(B) - Q(B) = \int 1\!\!1_B(p-q)d\mu \le \int_{p>q} 1\!\!1_B(p-q)d\mu = \int_A (p-q)d\mu = \|p-q\|_1/2.$$

By symmetry, $Q(B) - P(B) \le \int_{q>p}(q-p)d\mu = \|p-q\|_1/2$. Combining all these facts gives the result.

5 Extension to the non–iid framework

Let us now consider the general setting $\mathcal{P} = \{P_\theta^{(n)}, \ \theta \in \Theta\}$ where Θ is a separable metric space equipped with a (semi–)metric d (one allows that $d = d_n$ depends on n) and we have the domination assumption

$$dP_\theta^{(n)}(x) = p_\theta^{(n)}(x)d\mu^{(n)}(x),$$

for dominating measures $\mu^{(n)}$. We have seen previously that several canonical nonparametric settings fall into this framework with $\theta = f$ the unknown function in the model. As an example, recall the fixed–design nonparametric regression model

$$X_i = f(i/n) + \varepsilon_i, \quad 1 \le i \le n,$$

where ε_i are iid $\mathcal{N}(0,1)$. In this case $\theta = f$ and

$$P_f^{(n)} = \bigotimes_{i=1}^n \mathcal{N}(f(t_i), 1),$$

and $\mu^{(n)}$ is the Lebesgue measure on $\mathbb{R}^n$. This model has independent but not identically distributed observations (given f).

All results from Sections 2 and 3 extend to this more general setting, up to the following adjustments: in Bayes' formula, one replaces $\prod_{i=1}^n f(X_i)$ by $p_\theta^{(n)}(X^{(n)})$, so that the posterior mass of set A is, writing $X = X^{(n)}$ as a shorthand,

$$\Pi[A\,|\,X] = \frac{\int_A p_\theta^{(n)}(X)d\Pi(\theta)}{\int p_\theta^{(n)}(X)d\Pi(\theta)} = \frac{\int_A \frac{p_\theta^{(n)}}{p_{\theta_0}^{(n)}}(X)d\Pi(\theta)}{\int \frac{p_\theta^{(n)}}{p_{\theta_0}^{(n)}}(X)d\Pi(\theta)}.$$

The KL–type neighborhood is re–defined, for $\varepsilon > 0$, as

$$B_n(\theta_0, \varepsilon) = \left\{\theta\ :\ KL(P_{\theta_0}^{(n)}, P_\theta^{(n)}) \le n\varepsilon^2, V(P_{\theta_0}^{(n)}, P_\theta^{(n)}) \le n\varepsilon^2\right\}. \tag{2.10}$$

We denote now by E_θ the expectation under $P_\theta^{(n)}$ and similarly for E_{θ_0}, omitting the dependence on n in the notation for simplicity.

Exercice Verify that Lemmas 1, 2, 3 easily extend to this more general setting upon doing the above adjustements.

5.1 Gaussian sequence model

Let us now see an example of verification of the testing condition (T) in the non–iid setting: recall the Gaussian sequence model where $X = (X_k)_{k\ge1}$ is observed and, for ε_k iid $\mathcal{N}(0,1)$ and θ_k a given sequence in ℓ^2,

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

For sequences $a = (a_k)$ and $b = (b_k)$ let us write, provided the corresponding series converge

$$\langle a, b\rangle = \sum_{k\ge1} a_k b_k, \qquad \|a\|^2 = \sum_{k\ge1} a_k^2.$$

Lemma 6. Let $\theta, \theta_1, \theta_0$ be squared–integrable sequences, and let for $r = \|\theta_0 - \theta_1\|/4$,

$$B(x_1, r) = \{\theta\ :\ \|\theta - \theta_1\| \le r\}.$$

The test $\varphi_n = 1\!\!1\{2\langle\theta_1 - \theta_0, X\rangle > \|\theta_1\|^2 - \|\theta_0\|^2\}$ verifies, for $\bar\Phi(u) = P(\mathcal{N}(0,1) > u)$,

$$E_{\theta_0}\varphi_n \le \bar\Phi(\sqrt{n}\|\theta_1 - \theta_0\|/2),$$
$$\sup_{\theta\in B(\theta_1,r)} E_\theta(1-\varphi_n) \le \bar\Phi(\sqrt{n}\|\theta_1 - \theta_0\|/4)$$

In particular, condition (T) is verified.

Proof.

For the first inequality, one works under E_{θ_0}

$$P_{\theta_0}[2\langle \theta_1 - \theta_0, X\rangle > \|\theta_1\|^2 - \|\theta_0\|^2] = P[2\langle \theta_1 - \theta_0, \theta_0\rangle + 2\langle \theta_1 - \theta_0, \varepsilon/\sqrt{n}\rangle > \|\theta_1\|^2 - \|\theta_0\|^2]$$
$$= P[2\langle \theta_1 - \theta_0, \varepsilon\rangle > \sqrt{n}\|\theta_0 - \theta_1\|^2] = P[\mathcal{N}(0,1) > \sqrt{n}\|\theta_0 - \theta_1\|/2],$$

where the last line uses that $\langle \theta_1 - \theta_0, \varepsilon\rangle$ is a random variable of distribution $\mathcal{N}(0, \|\theta_1 - \theta_0\|^2)$, which gives the first inequality. For the second, one works this time under E_θ, for $\theta \in B(\theta_1, r)$.

$$P_\theta[2\langle \theta_1 - \theta_0, X\rangle \le \|\theta_1\|^2 - \|\theta_0\|^2] = P[2\langle \theta_1 - \theta_0, \theta\rangle + 2\langle \theta_1 - \theta_0, \varepsilon/\sqrt{n}\rangle \le \|\theta_1\|^2 - \|\theta_0\|^2].$$

We now write $\langle \theta_1 - \theta_0, \theta\rangle = \langle \theta_1 - \theta_0, \theta - \theta_0\rangle + \langle \theta_1 - \theta_0, \theta_0\rangle$. By writing $\theta = \theta_1 + rv$ with $\|v\| \le 1$,

$$\langle \theta_1 - \theta_0, \theta - \theta_0\rangle = \|\theta_1 - \theta_0\|^2 + r\langle \theta_1 - \theta_0, v\rangle \ge \|\theta_1 - \theta_0\|^2 - r\|\theta_1 - \theta_0\| \ge \frac{3}{4}\|\theta_1 - \theta_0\|^2,$$

where the last line uses Cauchy–Schwarz' inequality and $\|v\| \le 1$. Rearranging the probability at stake,

$$E_\theta[1 - \varphi_n] \le P\left[2\langle \theta_1 - \theta_0, \varepsilon/\sqrt{n}\rangle \le \|\theta_1 - \theta_0\|^2 - \frac{3}{2}\|\theta_1 - \theta_0\|^2\right]$$
$$\le P[\mathcal{N}(0, \|\theta_1 - \theta_0\|^2) \le -\sqrt{n}\|\theta_1 - \theta_0\|^2/4] = \bar{\Phi}(\sqrt{n}\|\theta_1 - \theta_0\|/4)$$

as desired. Property (T) now immediately follows for $\|\theta_1 - \theta_0\| > \varepsilon$ by using the standard inequality $\bar{\Phi}(u) \le e^{-u^2/2}$ for $u > 0$.

In the sequence model, we have, adapting the formula (1.2) given above for $\theta_0 = 0$,

$$\frac{dP_\theta^{(n)}}{dP_{\theta_0}^{(n)}}(X) = e^{n\langle X, \theta - \theta_0\rangle - \frac{n}{2}\|\theta\|^2 + \frac{n}{2}\|\theta_0\|^2}.$$

One deduces $K(P_{\theta_0}^{(n)}, P_\theta^{(n)}) = n\|\theta_0 - \theta\|^2/2$ and $V(P_{\theta_0}^{(n)}, P_\theta^{(n)}) = n\|\theta_0 - \theta\|^2$, so that

$$B_n(\theta_0, \varepsilon_n) = \left\{\theta \in \ell^2 \ :\ \|\theta - \theta_0\|^2 \le \varepsilon_n^2\right\}$$

(both inclusions hold, despite the 1/2 factor above). In this case, the KL–type neighborhood is just a ball for the L^2–norm.

5.2 Gaussian white noise and nonparametric regression

In the Gaussian white noise model, a similar proof as in the sequence model above shows that the test, with $\|\cdot\|_2$ denoting the L^2–norm on functions,

$$\varphi_n = 1\!\!1\{2\int_0^1 (f_1 - f_0)(t)dX^{(n)}(t) > \|f_1\|^2 - \|f_0\|^2\}\}$$

verifies the conclusions of Lemma 6. Also, $B_n(f_0, \varepsilon_n) = \{f \ :\ \|f - f_0\|_2 \le \varepsilon_n\}$ is again an L^2–ball.

In the nonparametric regression with fixed design, similar properties hold as in the sequence model, upon replacing $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ by

$$\langle f, g\rangle = \frac{1}{n}\sum_{i=1}^n f(t_i)g(t_i), \quad \|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f(t_i)^2.$$

Exercise. Establish an analogue of Lemma 6 in this case.

6 A generic result, second version

In the general setting $\mathcal{P} = \{P_\theta^{(n)},\ \theta \in \Theta\}$ with $dP_\theta^{(n)}(x) = p_\theta^{(n)}(x)d\mu^{(n)}(x)$ as in the previous Section, let us formulate a result generalising Theorem 1 (and that in particular applies to density estimation as well). Recall that E_θ is a shorthand for the expectation under $P_\theta^{(n)}$ and the testing condition, now formulated in the general setting:

Testing condition (T). Suppose one can find constants $K > 0$ and $a \in (0,1)$ such that for any $\varepsilon > 0$, if $\theta_0, \theta_1 \in \Theta$ are such that $d(\theta_0, \theta_1) > \varepsilon$, then there exist tests φ_n with

$$E_{\theta_0}\varphi_n \le e^{-Kn\varepsilon^2}$$
$$\sup_{\theta:\ d(\theta,\theta_1)<a\varepsilon} E_\theta(1-\varphi_n) \le e^{-Kn\varepsilon^2}.$$

Theorem 3. [GGV, entropy version] Let $(\bar{\varepsilon}_n, \underline{\varepsilon}_n)$ be sequences with $n(\bar{\varepsilon}_n^2 \vee \underline{\varepsilon}_n^2) \longrightarrow \infty$ as $n \longrightarrow \infty$. Suppose d is a distance on Θ such that the testing condition (T) holds with constants $a, K > 0$. Assume there exist $C, D > 0$ and measurable sets $\Theta_n \subset \Theta$ such that, for B_n as in (2.10),

i) $\log N(\bar{\varepsilon}_n, \Theta_n, d) \le Dn\bar{\varepsilon}_n^2$,

ii) $\Pi[\Theta \setminus \Theta_n] \le e^{-n\underline{\varepsilon}_n^2(C+4)}$,

iii) $\Pi[B_n(\theta_0, \underline{\varepsilon}_n)] \ge e^{-Cn\underline{\varepsilon}_n^2}$.

Set $\varepsilon_n = \bar{\varepsilon}_n \vee \underline{\varepsilon}_n$. Then for $M = M(a,K,C,D)$ large enough, the posterior distribution converges at rate $M\varepsilon_n$ towards f_0: as $n \longrightarrow \infty$,

$$E_{\theta_0}\Pi[\{\theta\ :\ \ d(\theta,\theta_0) \ge M\varepsilon_n\}\,|\,X] = o(1).$$

Proof.

We start by noting that, for given $n \ge 1$, the maps

$$\varepsilon \longrightarrow \log N(\varepsilon, \Theta_n, d), \qquad \varepsilon \longrightarrow n\varepsilon^2$$

are respectively non–increasing and increasing: for the first, note that if $\varepsilon' > \varepsilon$, a covering of Θ_n with ε–balls gives rise to a covering with ε'–balls using the same centers. Combining this monotonicity property with the entropy condition i), one now can apply Lemma 3 with $\varepsilon_n = \bar{\varepsilon}_n \vee \underline{\varepsilon}_n$. Indeed, the entropy condition required is also valid with the slower rate ε_n which gives the existence of tests ψ_n with

$$E_{\theta_0}\psi_n = o(1), \qquad \sup_{\theta\in\Theta_n:\ d(\theta,\theta_0)>M\varepsilon_n} E_\theta(1-\psi_n) \le e^{-cn\varepsilon_n^2},$$

that is, the first condition of Theorem 1. Now one proceeds as in the proof of Theorem 1 (now in the more general non–i.i.d. setting): first by combining ii) and iii) one obtains $E_{\theta_0}\Pi[\Theta \setminus \Theta_n \,|\, X] = o(1)$ by using the (generalised version of) Lemma 1, that requires $n\underline{\varepsilon}_n^2 \longrightarrow \infty$. We further introduce the set

$$C_n \,:\, \{\theta \in \Theta_n,\ d(\theta, \theta_0) \geq M\varepsilon_n\}.$$

The prior mass condition iii) is automatically verified if one replaces $\underline{\varepsilon}_n$ by ε_n: indeed by doing so the prior mass does not decrease and the exponential term decreases. Now one can follow line by line the proof of Theorem 1, only making the adjustements for the present general setting as explained in Section 5, which concludes the proof.

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

Histogram prior on $[0,1]$ *with deterministic number of jumps.* Let $K = K_n$ be an integer, a number of 'jumps' – to be chosen later –, and let us subdivide $[0,1]$ in K equally spaced intervals: for $I_k = [(k-1)/K, k/K)$, let us set

$$f = \sum_{k=1}^{K} h_k 1\!\!1_{I_k}, \qquad (h_1, \ldots, h_k) \sim P_\psi^{\otimes K}, \tag{2.11}$$

where P_ψ is the common distribution of the (random) histogram heights. For simplicity in what follows we take $P_\psi = \text{Lap}(1)$ the standard Laplace distribution, which has density $x \longrightarrow e^{-|x|}/2$ on $\mathbb{R}$, although many other choices are possible.

Statistical model. Let us consider one of the canonical nonparametric models: it turns out the simplest to verify the conditions of Theorem 3 is the Gaussian white noise model, but the proof is quite easily adapted for the regression and density models. We shall come back to the density model later. Recall that in the white noise model one observes $X = X^{(n)}$ with $dX(t) = f(t)dt + dW(t)/\sqrt{n}$.

Bayesian setting. We put as prior on $f \in L^2[0,1]$ a histogram prior Π defined as in (2.11), which combined with the law of $X \,|\, f$ in the white noise model gives a posterior distribution $\Pi[\cdot \,|\, X]$.

Frequentist study of $\Pi[\cdot \,|\, X]$ *and regularity condition on* f_0. To study the frequentist behaviour of the posterior, as usual we need to impose some regularity conditions on f_0, which will then typically influence the expression of the convergence rate one obtains. For $\alpha \leq 1$ define a Hölder–ball $C^\alpha(L) = \{g \,:\, [0,1] \longrightarrow \mathbb{R},\ \forall x, y \in [0,1],\ |g(x) - g(y)| \leq L|x-y|^\alpha\}$. We assume that the true f_0 belongs to $\mathcal{F} = \mathcal{F}(\alpha, L, M)$, for $L, M > 0$ and $\alpha \leq 1$, with

$$\mathcal{F} = \{f \,:\, [0,1] \longrightarrow \mathbb{R} \,:\, f \in C^\alpha(L),\ \|f\|_\infty \leq M\}.$$

Posterior convergence rate. The specific form of the model will actually not matter much for Theorem 3: it is enough to know we can apply it, since the white noise model verifies the testing condition (T) with $d = \|\cdot\|_2$ as noted earlier. So it is enough to verify the conditions i), ii), iii) of Theorem 3 with this distance and suitably chosen sets $\mathcal{F}_n$ (we construct them below). If we can do so, we will obtain a convergence rate for the posterior distribution $\Pi[\cdot \,|\, X]$ in terms of $d = \|\cdot\|_2$. For simplicity in this section we denote $\|\cdot\| = \|\cdot\|_2$ the L^2–norm on $[0,1]$.

7.1 Basic histogram facts

Let $\mathcal{V}_K = \mathrm{Vect}_{L^2}(\mathbb{1}_{I_1}, \dots, \mathbb{1}_{I_K})$ denote the subspace of $L^2 = L^2[0,1]$ spanned by histograms over the partition (I_k). For a sequence $(u_1, \dots, u_K) \in \mathbb{R}^K$, let us denote $\|\cdot\|_K$ the euclidean norm in $\mathbb{R}^K$

$$\|u\|_K^2 = \sum_{k=1}^K u_k^2.$$

Fact 1. The orthogonal projection of $f \in L^2$ onto $\mathcal{V}_K$ is

$$f^{[K]} = \sum_{k=1}^K \overline{f}_k \mathbb{1}_{I_k}, \qquad \text{with } \overline{f}_k = K\int_{I_k} f.$$

Let us denote $\overline{f} := (\overline{f}_1, \dots, \overline{f}_K)$. For any $f \in L^2$,

$$\|f^{[K]}\|^2 = \frac{1}{K}\sum_{k=1}^K \overline{f}_k^2 = \frac{1}{K}\|\overline{f}\|_K^2.$$

This, up to a factor K^{-1}, the L^2–norm of $f^{[K]}$ coincides with the $\|\cdot\|_K$–norm of the sequence $\overline{f}$.

Fact 2. Let $f_0 \in C^\alpha(L)$ with $\alpha \in (0,1]$. Then

$$\|f_0 - f_0^{[K]}\|_\infty \le LK^{-\alpha}.$$

Indeed, by the mean–value theorem $\overline{f}_{0,k} = K\int_{I_k} f_0 = f_0(c_k)$, for a $c_k \in I_k$. For $t \in I_k$, we have $|f_0(t) - f_0^{[K]}(t)| = |f_0(t) - f_0(c)| \le L|t-c|^\alpha \le LK^{-\alpha}$. This gives the result by making k range from 1 to K.

7.2 Verifying the conditions of Theorem 3

Let us choose some *sieve* sets $\mathcal{F}_n$ as follows

$$\mathcal{F}_n = \left\{ f \in \mathcal{V}_K \;:\; f = \sum_{k=1}^K h_k \mathbb{1}_{I_k}, \;\; (h_1, \dots, h_K) \in \mathcal{H}_n \right\},$$

where $\mathcal{H}_n$ is the set of sequences $h = (h_k)_{1 \le k \le K}$ defined as

$$\mathcal{H}_n = \left\{ h = (h_k), \;\; \max_{1 \le k \le K} |h_k| \le n \right\}.$$

The upper bound on the heights turns helpful to verify the entropy condition.

Entropy condition i). Since any $f \in \mathcal{V}_K$ is equivalently characterised by its height sequence h and $\|f\|^2 = K^{-1}\|h\|_K^2$, it is enough to cover the set of sequences $\mathcal{H}_n$. More precisely, by using Fact 1 above,

$$N(\varepsilon, \mathcal{F}_n, \|\cdot\|) = N(\sqrt{K}\varepsilon, \mathcal{H}_n, \|\cdot\|_K).$$

Now note that $\|h\|_K^2 \le K \max_{1 \le k \le K} h_k^2 \le Kn^2$ for any $h \in \mathcal{H}_n$, so that $\mathcal{H}_n \subset B_{\mathbb{R}^K}(0, \sqrt{K}n)$.

Lemma 7 gives that for $3M/\delta \ge 1$, we have $N(\delta, B_{\mathbb{R}^K}(0,M), \|\cdot\|_K) \le (3M/\delta)^K$. This implies, for $\varepsilon \le 1$,

$$N(\varepsilon, \mathcal{F}_n, \|\cdot\|) \le N(\sqrt{K}\varepsilon, B_{\mathbb{R}^K}(0, \sqrt{K}n), \|\cdot\|_K) \le (3n/\varepsilon)^K.$$

In order to fulfill i), one obtains the condition $K\log(3n/\bar{\varepsilon}_n) \le Dn\bar{\varepsilon}_n^2$.

Sieve condition ii). By definition of $\mathcal{F}_n$, using that $P[|\mathrm{Lap}(1)| > n] = e^{-n}$,

$$\Pi[\mathcal{F}_n^c] \le \Pi[\exists k \in \{1,\dots,K\} \;:\; |h_k| > n] \le Ke^{-n} \le \exp\{\log K - n\}.$$

In order to fulfill i), one obtains the condition $\log K - n \le -n\underline{\varepsilon}_n^2(C+4)$. This is always satisfied for large enough n provided $K = K_n$ is chosen so that $K_n = o(n)$.

Prior mass condition iii). Recall that in the white noise model $B_n(f_0,\varepsilon)$ is just the L^2–ball $\{f \;:\; \|f - f_0\| < \varepsilon\}$. Pythagoras theorem gives $\|f - f_0\|^2 = \|f - f_0^{[K]}\|^2 + \|f_0 - f_0^{[K]}\|^2$. So for any $\eta > 0$

$$\Pi[\|f - f_0\| < \eta] = \Pi[\|f - f_0^{[K]}\|^2 < \eta^2 - \|f_0 - f_0^{[K]}\|^2].$$

By Fact 2 above, $\|f_0 - f_0^{[K]}\| \le \|f_0 - f_0^{[K]}\|_\infty \le LK^{-\alpha}$. This means that provided K is chosen large enough in terms of η (the condition involving the rate is given below), one can always make sure that $\eta^2 - \|f_0 - f_0^{[K]}\|^2 \le \eta^2/2$. It is thus enough to consider, for $\varepsilon > 0$,

$$\begin{aligned}\Pi[\|f - f_0^{[K]}\| < \varepsilon] &= \Pi\left[K^{-1}\sum_{k=1}^{K}(\bar{f}_k - \bar{f}_{0,k})^2 \le \varepsilon^2\right] = \Pi\left[K^{-1}\sum_{k=1}^{K}(h_k - \bar{f}_{0,k})^2 \le \varepsilon^2\right]\\ &\ge \Pi\left[\bigcap_{k=1}^{K}\{|h_k - \bar{f}_{0,k}| \le \varepsilon\right] \ge \prod_{k=1}^{K}\Pi\left[|h_k - \bar{f}_{0,k}| \le \varepsilon\right],\end{aligned}$$

using the independence of the heights h_k under the considered prior distribution. To further bound from below the last display, note that $\Pi\left[|h_k - \bar{f}_{0,k}| \le \varepsilon\right]$ is the probability that a standard Laplace variable belongs to a certain interval of length 2ε. Since $|\bar{f}_{0,k}| \le \|f_0\|_\infty \le M$ by assumption, this interval is included in $[-M-\varepsilon, M+\varepsilon] \subset [-2M, 2M]$ if $\varepsilon \le 1$. On the latter interval, the standard Laplace density is at least $e^{-2M}/2$. Deduce that

$$\Pi\left[|h_k - \bar{f}_{0,k}| \le \varepsilon\right] \ge 2\varepsilon \cdot e^{-2M}/2 = \varepsilon e^{-2M},$$

so that $\Pi[\|f - f_0^{[K]}\| < \varepsilon] \ge \varepsilon^K \exp\{-2MK\}$. Putting all the previous bounds together, one sees that if $LK^{-\alpha} \le \underline{\varepsilon}_n/2$, then

$$\Pi[\|f - f_0\|_2 < \underline{\varepsilon}_n] \ge \Pi[\|f - f_0^{[K]}\| < \underline{\varepsilon}_n/2] \ge (\underline{\varepsilon}_n/2)^K \exp\{-2MK\}.$$

So the prior mass condition is verified if

$$\begin{aligned}LK^{-\alpha} &\le \underline{\varepsilon}_n/2\\ 2MK + K\log(2/\underline{\varepsilon}_n) &\le Cn\underline{\varepsilon}_n^2.\end{aligned}$$

It is now easy to verify that conditions i) up to iii) are satisfied for the choices

$$\underline{\varepsilon}_n \asymp \bar{\varepsilon}_n \asymp \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}, \qquad K \asymp \left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}}.$$

We then get the following result.

Theorem 4. In the Gaussian white noise model, suppose the true f_0 belongs to the class $\mathcal{F}(\alpha, L, M)$. Let Π be a random histogram prior as above with a number of jumps equal to

$$K \asymp \left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}}.$$

Then for any $\alpha \in (0, 1]$ and $m > 0$ a large enough constant,

$$E_{f_0}\Pi[\|f - f_0\|_2 \leq m\varepsilon_n \,|\, X] \longrightarrow 1, \qquad \varepsilon_n \asymp \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}.$$

8 Complements

8.1 Refinements of conditions in generic results and extensions

The aim when stating the previous two GGV theorems was to give simple – yet general and already fairly broadly applicable – statements and proofs. These results can be in turn refined in a number of ways. Many refinements are described in the book by Ghosal and van der Vaart (2017) (referred to as GV–book in the sequel). we only briefly mention a few

1. *Coupling numerator and denominator when studying Bayes' formula.* The previous formulations of the GGV theorem treat denominator and numerator separately. This could be suboptimal, especially in situations where the parameter space is large/unbounded: we will see an example in the Chapter on high–dimensional models.

2. *Other notions of entropy.* It is already clear from the proof of Lemma 3 that upper-bounds are possibly generous there, and indeed one can provide more precise conditions. Instead of looking at a 'global' entropy, one can also look at a more 'local' version of the entropy.

Fractional posteriors. A popular generalisation (in particular in machine learning and PAC–Bayes theory) of the posterior distribution is the so–called *α–posterior*, where given a prior Π on θ, for $\alpha > 0$ one defines the distribution, for every measurable B,

$$\Pi_\alpha[B \,|\, X] = \frac{\int_B L_{n,\alpha}(\theta)\Pi(d\theta)}{\int L_{n,\alpha}(\theta)\Pi(d\theta)}, \quad L_{n,\alpha}(\theta) = \left[p_\theta^{(n)}(X)\right]^\alpha.$$

Typically one chooses $0 < \alpha < 1$, which tempers the influence of the data in the obtained distribution. A technical advantage of working with an α–posterior is that in some situations (and for certain loss functions) convergence rate results can be obtained under prior–mass conditions only, without requiring entropy–type bounds. This was already noted in a paper by Tong Zhang (2006) and recently re–explored in the work by Bhattacharya et al. (2020). Another advantage is that it is more robust to model–misspecification. A drawback is that $L_{n,\alpha}(X)$ is not a likelihood anymore, so the original Bayesian interpretation is lost: optimality properties related to the use of the likelihood may then be lost. Typically, statistical efficiency (e.g. optimal variance) could be lost and 'credible' set from the α–posterior will also often be larger for $\alpha < 1$ than in the posterior case $\alpha = 1$.

8.2 Lower bounds

Definition 4. For d a distance on the parameter set Θ, we say that ζ_n is a lower bound for the posterior $\Pi[\cdot \,|\, X] = \Pi[\cdot \,|\, X_1, \dots, X_n]$ contraction rate, in terms of the distance d, if

$$\Pi\,[\,\{\theta\ :\ \ d(\theta,\theta_0) \leq \zeta_n\}\,|\,X_1,\dots,X_n] \longrightarrow 0,$$

in probability under P_{θ_0}.

The interpretation is that if one looks with a magnifying glass 'too close' to a given point θ_0 then asymptotically there is no posterior mass.

Exercice. Suppose $\Theta \subset \mathbb{R}$ and that the model $\mathcal{P} = \{P_\theta^{\otimes n},\ \theta \in \Theta\}$ is *"regular"* in that KL–type balls of radius ε are 'comparable' to intervals of size ε i.e. there exists a constant $c > 0$ such that for any $\varepsilon > 0$ and $\theta_0 \in \Theta$,

$$B_{KL}(\theta_0,\varepsilon) \supset \{\theta\ :\ \ |\theta-\theta_0| \leq c\varepsilon\},$$

where B_{KL} has the same definition as for density estimation but with f_0, f replaced by p_{θ_0}, p_θ (check that this is true e.g. in the fundamental model $\{\mathcal{N}(\theta,1)^{\otimes n},\ \theta \in \mathbb{R}\}$.) Prove that for any sequence (m_n) going to 0,

$$E_{\theta_0}\Pi[|\theta-\theta_0| \leq \frac{m_n}{\sqrt{n}}\,|\,X] \longrightarrow 0,$$

that is, $m_n/\sqrt{n}$ is a lower bound for the posterior contraction rate.

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

Let $B_{\mathbb{R}^k}(y,R) = \{x \in \mathbb{R}^k\ :\ \ \|x-y\| \leq R\}$ denote the euclidian ball of center $y \in \mathbb{R}^k$ and radius $R \geq 0$, for $\|x\|^2 = \sum_{i=1}^k x_i^2$ the standard euclidian norm.

Lemma 7. [Entropy of unit ball in $\mathbb{R}^k$] For any $\delta > 0$, for any $M > 0$, the covering number of $B_{\mathbb{R}^k}(0,M)$ with respect to the euclidean norm verifies, with $a \vee b = \max(a,b)$,

$$N(\delta, B_{\mathbb{R}^k}(0,M), \|\cdot\|) \leq \left(1 \vee \frac{3M}{\delta}\right)^k.$$

Proof.

If $\delta \geq M$, the result is clear: one ball suffices to cover, so one assumes $\delta < M$. Since by applying an homothecy of ratio M, norms are multiplied by M, we have $N(\delta, B_{\mathbb{R}^k}(0,M), \|\cdot\|) = N(1, B_{\mathbb{R}^k}(0,M/\delta), \|\cdot\|)$. So it is enough to consider the case $M = 1$ with $\delta < 1$ (up to setting $\delta' = \delta/M$). Let

$$N\ := N(\delta, B, \|\cdot\|), \quad \text{with } B := B_{\mathbb{R}^k}(0,1).$$

Let $N' = N_s(\delta)$ denote the *maximal* number of points of B separated by at least δ for $\|\cdot\|$. Consider a collection of such points $(x_i,\ 1 \leq i \leq N')$ and note that the collection of balls $B(x_i, \delta) = B_{\mathbb{R}^k}(x_i, \delta)$ must cover B, otherwise one could find a point y separated from all the x_i's by at least δ, contradicting maximality. On the other hand, since x_i's are in B,

$$B(x_1, \frac{\delta}{2}) \bigcup \cdots \bigcup B(x_{N'}, \frac{\delta}{2}) \subset B(0, 1 + \frac{\delta}{2}).$$

Also, the balls $B(x_i, \frac{\delta}{2})$ are disjoint by definition of δ−separation. Denoting by $\mathscr{V}(A)$ the volume of a measurable subset A of $\mathbb{R}^k$ with respect to Lebesgue measure, one deduces

$$\sum_{i=1}^{N'} \mathscr{V}(B(x_i, \delta/2)) \leq \mathscr{V}(B(0, 1 + \delta/2)).$$

A change of variables gives $\mathscr{V}(B(0, r)) = r^k \mathscr{V}(B(0, 1))$ for $r > 0$. Since $\mathscr{V}(B(x_i, \delta/2)) = \mathscr{V}(B(0, \delta/2))$, one obtains

$$N'(\delta/2)^k \mathscr{V}(B(0, 1)) \leq \left(1 + \frac{\delta}{2}\right)^k \mathscr{V}(B(0, 1)).$$

One concludes that $N \leq N' \leq (\frac{2+\delta}{\delta})^k \leq (3/\delta)^k$ as announced.

CHAPTER 3

Nonparametric Bayes and adaptation

We explore different classes of prior distributions that lead to 'adaptive' behaviour in nonparametric problems, where adaptive often means that the posterior automatically adapts to the unknown smoothness parameter of the true function.

1 Random histograms

1.1 White noise model

In the setting of the Gaussian white noise model, let us recall the definition of the histogram prior on $[0,1]$ with deterministic number of jumps: given K_n an integer, a number of 'jumps' we subdivided $[0,1]$ in K equally spaced intervals: for $I_k = [(k-1)/K, k/K)$, and set

$$\textit{[Fixed K prior]} \qquad f = \sum_{k=1}^{K} h_k \mathbb{1}_{I_k}, \qquad (h_1, \dots, h_K) \sim P_\psi^{\otimes K},$$

where P_ψ is the common distribution of the (random) histogram heights for instance standard Laplace distribution. For this prior we denoted for simplicity h_k the weights, although $h_{k,K}$ would be more explicit: we use this notation below.

A natural way to attempt making the prior *adaptive* with respect to the unknown regularity of f_0 is to take K itself random by setting $\Pi = \Pi_H$

$$K \sim \pi_K(\cdot), \qquad \text{with } \pi_K(k) \propto e^{-k \log k},$$
$$f \,|\, K \sim \mathcal{L}\left(f = \sum_{k=1}^{K} h_{k,K} \mathbb{1}_{I_k}, \quad (h_{1,K}, \dots, h_{k,K}) \sim P_\psi^{\otimes K} \right).$$

This is called a hierarchical Bayes approach. We note that other choices of the prior on K are possible. For instance, one could take $\pi_K(k) \propto e^{-k}$. This would lead to a similar result, but with a slightly different log–factor in the rate (check this as an exercise after having done the proof below for the former prior).

Theorem 1. In the Gaussian white noise model, suppose the true f_0 belongs to the class $\mathcal{F}(\alpha, L, M)$ for some $\alpha \in (0,1]$. Then for $\Pi = \Pi_H$ the prior with random K as above and $m > 0$ a large enough constant,

$$E_{f_0} \Pi[\|f - f_0\|_2 \le m\varepsilon_n \,|\, X] \longrightarrow 1, \qquad \varepsilon_n \asymp \left(\frac{\log n}{n} \right)^{\frac{\alpha}{2\alpha+1}}.$$

Proof.

One defines a sieve as, with $K_n = Dk_n$ for D large enough to be chosen, and $k_n = (n/\log n)^{1/(2\alpha+1)}$,

$$\mathcal{F}_n = \bigcup_{k=1}^{K_n} \left\{ f = \sum_{j=1}^{k} h_{j,k} \mathbb{1}_{I_j}, \ \max_j |h_{j,k}| \le n \right\} = \bigcup_{k=1}^{K_n} \mathcal{F}_{n,k}.$$

The entropy condition is easily verified using $\|u\|_2 \le \sqrt{k} \max_j |u_j|$ and (see previous Chapter)

$N(\varepsilon, \mathcal{F}_{n,k}, \|\cdot\|_2) \le (3n\sqrt{k}/\varepsilon)^k$, so that

$$N(\varepsilon, \mathcal{F}_n, \|\cdot\|_2) \le \sum_{k=1}^{K_n} (3nK_n/\varepsilon)^k \lesssim (nK_n/\varepsilon)^{k_n+1}$$

which gives $\log N(\bar{\varepsilon}_n, \mathcal{F}_n, \|\cdot\|_2) \lesssim K_n \log(n/\bar{\varepsilon}_n) + K_n \log K_n$.
Also, the complement of the sieve has small prior mass as

$$\begin{aligned}\Pi[\mathcal{F}_n^c] &\le \Pi(K > K_n) + \sum_{k=1}^{K_n} \Pi[\mathcal{F}_n^c \,|\, K = k]\Pi[K = k] \\ &\lesssim e^{-K_n \log K_n} + \sum_{k=1}^{K_n} k e^{-n}\Pi[K = k] \lesssim e^{-K_n \log K_n} + e^{-n}E[K] \lesssim e^{-K_n \log K_n} + e^{-n},\end{aligned}$$

where we use that K has finite expectation. Finally, for the prior mass condition,

$$\begin{aligned}\Pi[\|f - f_0\|_2 \le \underline{\varepsilon}_n] &\ge \Pi[\{\|f - f_0\|_2 \le \underline{\varepsilon}_n\} \cap \{K = k_n\}] \\ &= \Pi[\|f - f_0\|_2 \le \underline{\varepsilon}_n \,|\, K = k_n]\Pi[K = k_n]\end{aligned}$$

and we can now used the bound for fixed $K = k_n$ used in the previous chapter, which gives, provided $Lk_n^{-\alpha} \le \underline{\varepsilon}_n/2$, that $\Pi[\|f - f_0\|_2 < \underline{\varepsilon}_n \,|\, K = k_n] \ge (\underline{\varepsilon}_n/2)^{k_n} \exp\{-2Mk_n\}$.

Putting everything together, we see that we need: $K_n \log(nK_n/\bar{\varepsilon}_n) \lesssim n\bar{\varepsilon}_n^2$, and $Lk_n^{-\alpha} \le \underline{\varepsilon}_n/2$ as well as $k_n \log(2/\underline{\varepsilon}_n) + 2Mk_n \lesssim n\underline{\varepsilon}_n^2$. This is satisfied for $\bar{\varepsilon}_n \asymp \underline{\varepsilon}_n \asymp \varepsilon_n$ as in the statement of the result (choose first $\underline{\varepsilon}_n$ with a large enough constant to verify prior mass, then D large enough to verify the second condition and finally the constant in front of $\underline{\varepsilon}_n$ to be large enough).

1.2 Density estimation

Theorem 1 admits a counterpart in density estimation: in order to apply the GGV theorem, we have to work with a testing distance: for instance $d = h$ (or $d = \|\cdot\|_1$). The prior has to be modified too: one needs to take a histogram prior on *densities* on $[0, 1]$. To do so, we wish that the random histogram drawn from the prior is positive and we should have $\int_0^1 f = 1$. A possible fixed K–prior is as follows, for $a_1, \dots, a_K$ positive,

$$\textit{[Fixed K prior]} \qquad f = \sum_{k=1}^{K} g_k K \mathbb{1}_{I_k}, \qquad (g_1, \dots, g_K) \sim \text{Dir}(a_1, \dots, a_K),$$

where $\text{Dir}(a_1, \dots, a_K)$ is the Dirichlet distribution, which samples K–tuples directly from the simplex in dimension K. Then, a random K prior can be obtained as in the previous example, by setting e.g. $\pi_K(k) \propto e^{-k \log k}$.

One can then formulate a result similar to Theorem 1 we briefly sketch the differences: in the class of true f_0's, one assumes that the true density f_0 is bounded away from 0. In the proof, one has to work a little to relate the KL–neighborhood to an L^2 neighborhood, since those are not equal in the density model (see the handwritten notes for details). Finally, the prior mass condition is handled via a specific lemma on the Dirichlet distribution.

2 Gaussian processes

We study below Gaussian processes seen as prior distributions on functions. We explore some of the properties of a given Gaussian process in term of its 'geometry' interpreted as the shape of (the unit ball of) its Reproducing Kernel Hilbert Space (RKHS), that we define below. We explain how this influences the associated posterior convergence rates in typical nonparametric settings.

2.1 Definitions and examples

We now define two notions: the one of Gaussian process, interpreted as a collection of normal random variables, and the one of (Banach–valued) Gaussian random variable. In standard settings such as within separable Banach spaces, both notions are essentially equivalent, as we briefly discuss below.

Definition 1. A Gaussian process $W = (W_t)_{t\in T}$ is a stochastic process (i.e. a collection of random variables) indexed by the set T such that for any $t_1, \dots, t_k \in T$ and any $k \geq 1$, the vector $(W_{t_1}, \dots, W_{t_k})$ is a Gaussian random vector.

In the sequel to fix ideas we take $T = [0, 1]$ as index set. Recall that a Gaussian vector is characterised by its mean and variance–covariance matrix. So, a Gaussian process (we also write GP) W, if it exists (we assume so), must be characterised by the quantities

$$\mu(t) = E[W_t]$$
$$K(s, t) = \text{Cov}(W_s, W_t) = E[(W_s - EW_s)(W_t - EW_t)].$$

Those are called respectively *mean function* and *covariance* (operator).

In the sequel we restrict for simplicity to *centered* Gaussian processes, i.e. we take $\mu(t) = 0$ for all t. The map $(s, t) \longrightarrow K(s, t)$ is symmetric and *definite–positive* in that for any finite collection $t_1, \dots, t_k \in [0, 1]$, the matrix $(K(t_i, t_j))$ is definite positive. This follows immediately from the expression of $K(\cdot, \cdot)$.

Let us insist again that the definition above gives the finite–dimensional distributions (also called FIDIs) $(W_{t_1}, \dots, W_{t_k})$ but we shall not construct a process $W(t) = W_t$ that verifies this (it is possible to do so).

The map $t \longrightarrow W(t)$ is called trajectory (or realisation) of W. It can then happen that the process admits a version (that is, there exists $(Z_t)_{t\in T}$ with $P[Z_t = W_t] = 1$ for any $t \in T$, which imples the FIDIs are the same) whose trajectories are continuous, that is $t \longrightarrow Z_t(\omega)$ is continuous. Then the image of the map $\omega \longrightarrow (Z_t(\omega))_t$ is included in $C^0[0, 1]$. More generally, W may have a version that has trajectories in a *separable Banach space* $\mathbb{B}$, such as $(C^0[0, 1], \|\cdot\|_\infty)$ above.

Definition 2. A $\mathbb{B}$–valued Gaussian random variable is a map $W : \Omega \longrightarrow \mathbb{B}$, measurable for the Borel σ–field of $\mathbb{B}$, such that for any b^* in the dual space $\mathbb{B}^*$, the real variable b^*W is Gaussian.

Note that if Y is a Gaussian variable in $\mathbb{B}$ according to this definition, then if $\mathcal{T} \subset \mathbb{B}^*$, the collection $(b^*Y,\ b^* \in \mathcal{T})$ is a Gaussian process indexed by $\mathcal{T}$ (this is because: a random vector is Gaussian iff any finite linear combination of its coordinates is Gaussian, and: $\mathbb{B}^*$ is a linear space). For example, if $\mathbb{B} = C^0[0,1]$ as above, and $\mathcal{T}$ is the set of linear maps $b_t : f \longrightarrow f(t)$ for $f \in \mathbb{B}$ (they are continuous, therefore in $\mathbb{B}^*$), then the collection of variables $(W(\omega)(t),\ t \in [0,1])$ is a Gaussian process on $[0,1]$ with trajectories in $\mathbb{B}$.

Under a measurability condition, we have that if (W_t) is a Gaussian process with trajectories in $\mathbb{B}$, then it is also a Gaussian random variable in $\mathbb{B}$.

Lemma 1. Si $(W_t)_{t\in T}$ has a version that admits trajectories in a separable Banach space $\mathbb{B}$, and if for any $w \in \mathbb{B}$, the quantity $\|W - w\|_{\mathbb{B}}$ is a random variable (i.e. is a measurable quantity as function of ω), then W is a Gaussian random variable in $\mathbb{B}$.

We omit the proof: it is based on the fact that when $\mathbb{B}$ is separable the Borel σ–field is generated by balls. In the sequel, since we generally work with separable Banach spaces, we will use interchangeably both concepts.

Definition 3. For W taking values in the separable Banach space $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$, we call small ball probability the quantity, for $\varepsilon > 0$,

$$P[\|W\|_{\mathbb{B}} < \varepsilon] =: \exp(-\varphi_0(\varepsilon)).$$

Sometimes $\varphi_0(\varepsilon) = -\log P[\|W\|_{\mathbb{B}} < \varepsilon]$ is called small–ball term.

2.2 RKHS of a Gaussian process

Let $(W_t)_{t\in T}$ be a Gaussian process. Consider the space

$$C_W = \overline{\text{Vect}\{W(t),\ t \in T\}}^{L^2} = \overline{\left\{\sum_{i=1}^{p} \alpha_i W(t_i),\ t_i \in T, p \geq 1\right\}}^{L^2},$$

where $\overline{\mathcal{A}}^{L^2}$ stands for the completion of given set $\mathcal{A}$ of random variables (defined on a common probability space Ω) in $L^2(\Omega)$. This space is called the *first order chaos* associated to W.

Definition 4. The RKHS of a centered Gaussian process $W = (W_t)_{t\in T}$ is the set

$$\mathbb{H} = \{g_H : T \longrightarrow \mathbb{R},\ \ g_H(t) = E[W_t H],\ H \in C_W\}.$$

This set is equipped with the inner product $\langle\cdot,\cdot\rangle_{\mathbb{H}}$ given by, for $H_1, H_2 \in \mathbb{H}$,

$$\langle g_{H_1}, g_{H_2}\rangle_{\mathbb{H}} = E[H_1 H_2].$$

The space $\mathbb{H}$ is a Hilbert space called Reproducing Kernel Hilbert Space (RKHS) of W.

Note that $\mathbb{H}$ is indeed a Hilbert space since the map $H \longrightarrow g_H$, mapping $\mathcal{C}_W$ into $\mathbb{H}$, is by definition an isometry, and $\mathcal{C}_W$ is a Hilbert space as closed sub–space of L^2.

We now list a number of useful properties

- If $H = W_s$, then $g_H(t) = E(W_t W_s) = K(s,t)$ so $g_H(\cdot) = K(s,\cdot)$ in this case.
- If $H = \sum_{i=1}^p a_i W_{s_i}$, then similarly $g_H(\cdot) = \sum_{i=1}^p K(s_i,\cdot)$.
- For $H \in \mathcal{C}_W$ arbitrary, we have

  $$g_H(t) = E[W_t H] = \langle K(t,\cdot), g_H(\cdot)\rangle_{\mathbb{H}}$$

  This is sometimes called the reproducing formula, as it expresses the value of any element of the RKHS at a point as an inner product involving the function itself.
- Paralleling the definition of $\mathcal{C}_W$ and thanks to the isometry $H \longrightarrow g_H$, we have

  $$\overline{\left\{\sum_{i=1}^p \alpha_i K(s_i,\cdot),\ s_i \in T, p \geq 1\right\}}^{\mathbb{H}} = \mathbb{H}.$$

- In case W has trajectories in a separable Banach space $\mathbb{B}$, it can be proved that $\mathbb{H}$ identifies with a subspace of $\mathbb{B}$. We will see this is indeed the case in the examples investigated below.

Let $\mathcal{S}_W$ denote the space $\mathrm{Vect}\{W(t),\ t \in T\}$, so that $\mathcal{C}_W = \overline{\mathcal{S}_W}^{L^2}$.

Example 1 (Gaussian vector in $\mathbb{R}^k$). Let $W = (W_1,\dots,W_k)^T$ be a (column) Gaussian vector. It is a Gaussian process indexed by $T = \{1,\dots,k\}$. Here we will identify a vector in $\mathbb{R}^k$ and a function $T = \{1,\dots,k\} \longrightarrow \mathbb{R}$.

If $W \sim \mathcal{N}(0,\Sigma)$ with an invertible covariance matrix Σ then, for any $u, v \in \mathbb{R}^k$,

$$(\mathbb{H}, \|\cdot\|_{\mathbb{H}}) = (\mathbb{R}^k, \|\cdot\|_{\mathbb{H}}), \qquad \langle u, v\rangle_{\mathbb{H}} = u^T\Sigma^{-1}v. \tag{3.1}$$

This means that the RKHS of a Gaussian vector in $\mathbb{R}^k$ coincides with the ambient space, but with a twisted geometry that is given by the inner product as above. Let us check this: for $H = \sum_{j=1}^k a_j W_j$ in $\mathcal{S}_W$ and $a = (a_1,\dots,a_k)$, $W = (W_1,\dots,W_k)$,

$$g_H(i) = E[HW_i] = \sum_{j=1}^k a_j E[W_j W_i] = \sum_{j=1}^k a_j \Sigma_{i,j} = (\Sigma a)_i,$$

so that g_H can be identified with the vector Σa. In particular, as Σ is invertible, any vector $v \in \mathbb{R}^k$ is in $\mathbb{H}$, as $v = \Sigma a_v$ with $a_v = \Sigma^{-1}v$. Now $v = E[WW^T]a_v = E[WW^T a_v]$. For $u, v \in \mathbb{R}^k$,

$$\langle u, v\rangle_{\mathbb{H}} = E[W^T a_u W^T a_v] = a_u^T E[WW^T]a_v = a_u^T \Sigma a_v = u^T\Sigma^{-1}v,$$

using that for any real number $x^T = x$ (here $x = W^T a_u$), which proves (3.1).

Example 2 (Random series). More generally, if one now considers

$$W_t = \sum_{j=1}^{\infty} \sigma_j \zeta_j e_j(t), \qquad t \in [0,1],$$

with $(\sigma_j) \in \ell^2$, ζ_j independent $\mathcal{N}(0,1)$ variables and $\{e_j(\cdot)\}$ an orthonormal basis of $L^2[0,1]$, it can be shown that

$$\mathbb{H} = \left\{ \sum_{j=1}^{\infty} \lambda_j e_j(t), \quad (\lambda_j) \text{ such that } \sum_{j=1}^{\infty} \sigma_j^{-2} \lambda_j^2 < \infty \right\},$$

equipped with the inner product

$$\langle \sum_{j=1}^{\infty} \lambda_j e_j, \sum_{j=1}^{\infty} \mu_j e_j \rangle_{\mathbb{H}} = \sum_{j=1}^{\infty} \sigma_j^{-2} \lambda_j \mu_j.$$

Example 3 (Brownian motion). Consider standard Brownian motion $(B_t)_{t\in[0,1]}$.

Lemma 2. The RKHS of Brownian motion on $[0,1]$ is

$$\mathbb{H} = \left\{ \int_0^{\cdot} g(u)du, \quad g \in L^2[0,1] \right\},$$

equipped with the inner product

$$\langle \int_0^{\cdot} g_1(u)du, \int_0^{\cdot} g_2(u)du \rangle_{\mathbb{H}} = \int_0^1 g_1(u)g_2(u)du.$$

Remark. The last inner product can equivalently been written, observing that elements of $\mathbb{H}$ are differentiable almost everywhere, $\langle h_1, h_2 \rangle_{\mathbb{H}} = \int_0^1 h_1' h_2'$. Also, note that for any $h \in \mathbb{H}$ as above, we have $h(0) = 0$. One can 'release' Brownian motion at zero and instead consider the process $Z_t = B_t + N$, for N a standard normal variable independent of (B_t). Using a similar proof as for Lemma 2, one can show that the RKHS $\mathbb{H}_Z$ of this process is $\mathbb{H} = \left\{ c + \int_0^{\cdot} g(u)du, \; g \in L^2[0,1], c \in \mathbb{R} \right\}$ with inner–product $\langle f_1, f_2 \rangle_{\mathbb{H}_Z} = f_1(0)f_2(0) + \int_0^1 f_1' f_2'$.

Proof of Lemma 2.

Let us denote $\mathcal{E} = \left\{ \int_0^{\cdot} g(u)du, \quad g \in L^2[0,1] \right\}$. If $Y = \sum_{i=1}^p a_i B_{t_i}$ the corresponding $g_Y \in \mathbb{H}$ can be written $\int_0^t h(u)du$ with h the step function $h = \sum_{i=1}^p a_i \mathbb{1}_{[0,t_i]}$. This is because

$$E[YB_t] = \sum_{i=1}^p a_i E[B_{t_i} B_t] = \sum_{i=1}^p a_i (t_i \wedge t) = \int_0^t (\sum_{i=1}^p a_i \mathbb{1}_{[0,t_i]})(u)du.$$

Also, for $Y = \sum_{i=1}^p a_i B_{t_i}$ and $Z = \sum_{j=1}^q b_j B_{t_j'}$ two elements of $\mathcal{S}_B$, with corresponding $g_Y(\cdot) =$

$\sum_{i=1}^{p} a_i(t_i \wedge \cdot)$ and $g_Z(\cdot) = \sum_{j=1}^{q} b_i(t'_j \wedge \cdot)$, by definition of the RKHS, $\langle g_Y, g_Z \rangle_{\mathbb{H}}$ equals

$$\langle \sum_{i=1}^{p} a_i(t_i \wedge \cdot), \sum_{j=1}^{q} b_i(t'_j \wedge \cdot) \rangle_{\mathbb{H}} = E[YZ] = \sum_{i,j} a_i b_j (t_i \wedge t'_j) = \int_0^1 (\sum_{i=1}^{p} a_i \mathbb{1}_{[0,t_i]})(\sum_{j=1}^{q} a_i b_j \mathbb{1}_{[0,t'_j]}). \tag{3.2}$$

The above also shows that for any g_k step function, the function $\int_0^{\cdot} g_k$ belongs to $\mathbb{H}$: it suffices to use $\mathbb{1}_{[a,b]} = \mathbb{1}_{[0,b]} - \mathbb{1}_{[0,a]}$ for any indicator involved in the definition of the step function.

We now check $\mathcal{E} \subset \mathbb{H}$. Let $h = \int_0^{\cdot} g$ for $g \in L^2[0,1]$. As any square–integrable function is a limit in L^2 of a sequence of step functions, consider $\int_0^{\cdot} g_k =: h_k \in \mathbb{H}$ for g_k step function with $\|g_k - g\|_2 \longrightarrow 0$. Now (h_k) is a Cauchy sequence in $\mathbb{H}$: indeed by (3.2) we have $\|h_p - h_q\|_{\mathbb{H}}^2 = \|g_p - g_q\|_2^2$ and (g_q) is Cauchy since converging in L^2. Deduce that (h_k) converges in $\mathbb{H}$ to $\eta \in \mathbb{H}$. To check that $h = \eta$ we use the reproducing formula

$$\eta(t) = \langle K(\cdot, t), \eta \rangle_{\mathbb{H}} = \lim_k \langle K(\cdot, t), h_k \rangle_{\mathbb{H}} = \lim_k h_k(t) = h(t),$$

by continuity of the $\mathbb{H}$–inner product and since (h_k) converges pointwise to h (as $\|g_k - g\|_2 \longrightarrow 0$).

Let now $\gamma \in \mathbb{H}$. By definition $\gamma(t) = EB_tH$, where $H \in \mathcal{C}_B = \overline{\mathcal{S}_B}$ and we have $H_k \longrightarrow H$ in L^2 for some $H_k \in \mathcal{S}_B$. From what we have seen above EB_tH_k can be written $\int_0^t h_k = g_{H_k}$ for h_k a step function, and we have $\gamma(t) = \lim_k EB_tH_k = \lim_k \int_0^t h_k$ (reproducing formula). Since g_{H_k} converges to γ in $\mathbb{H}$ (as follows from the isometry $\mathcal{C}_B \longrightarrow \mathbb{H}$), the sequence (g_{H_k}) is Cauchy in $\mathbb{H}$. As $\|g_{H_p} - g_{H_q}\|_{\mathbb{H}}^2 = \|h_p - h_q\|_2^2$, we deduce that (h_k) is Cauchy in L^2 which implies $\|h_k - h\|_2 \longrightarrow 0$ for some $h \in L^2$, so that $\int_0^t h_k \longrightarrow \int_0^t h$. This means $\gamma(\cdot) = \int_0^{\cdot} h$ which shows $\mathbb{H} \subset \mathcal{E}$ and concludes the proof.

2.3 Fundamental properties via RKHS and concentration function

Cameron–Martin formula

Theorem 2. [Cameron–Martin] For a Gaussian random variable W taking values in $\mathbb{B}$ separable Banch space and $h \in \mathbb{B}$, the distributions P_{W+h} and P_W of $W + h$ and W are mutually absolutely continuous if and only if $h \in \mathbb{H}$. Let us further define the map

$$U : \mathbb{H} \longrightarrow \mathcal{C}_W$$
$$h = g_H \longrightarrow H.$$

Then for any $h \in \mathbb{H}$,

$$\frac{dP_{W+h}}{dP_W}(W) = \exp\left\{ Uh - \frac{\|h\|_{\mathbb{H}}^2}{2} \right\}.$$

Remark. For Brownian motion, this in fact coincides with Girsanov's formula.

Ball probabilities and concentration function

Using Cameron–Martin formula above, it is possible to show that if h belongs to $\mathbb{H}$, then for any $\varepsilon > 0$,

$$P[\|W - h\|_{\mathbb{B}} < \varepsilon] \geq e^{-\|h\|_{\mathbb{H}}^2/2} P[\|W\|_{\mathbb{B}} < \varepsilon]. \tag{3.3}$$

If h does not belong to $\mathbb{H}$, some control of the probability on the left-hand side is possible via the concentration function that we define now.

Definition 5. Let W be a Gaussian random variable taking its values in $\mathbb{B}$ separable Banach space, with RKHS $\mathbb{H}$. Let w belong to $\overline{\mathbb{H}}^{\mathbb{B}}$, the closure in $\mathbb{B}$ (with respect to the norm of $\mathbb{B}$) of $\mathbb{H}$. For any $\varepsilon > 0$, define

$$\begin{aligned} \varphi_w(\varepsilon) &= \inf_{h \in \mathbb{H},\ \|h-w\|_{\mathbb{B}} < \varepsilon} \frac{1}{2}\|h\|_{\mathbb{H}}^2 - \log P[\|W\|_{\mathbb{B}} < \varepsilon] \\ &=: \ \varphi_w^A(\varepsilon) + \varphi_0(\varepsilon) \end{aligned}$$

The function $\varphi_w(\cdot)$ is called the concentration function of the process W.

The concentration function is the sum of the small–ball term and of an approximation term, which measures the ability of elements of $\mathbb{H}$ to approximate a given w. Note that for $w \in \overline{\mathbb{H}}^{\mathbb{B}}$ and a given $\varepsilon > 0$, the approximation term is always finite. If $w \in \overline{\mathbb{H}}^{\mathbb{B}}$ but $w \notin \mathbb{H}$, then $\varphi_w^A(\varepsilon) \longrightarrow +\infty$ as $\varepsilon \longrightarrow 0$ (otherwise by extracting a subsequence the norm $\|h\|_{\mathbb{H}}$ would be finite).

Theorem 3. Let W be a Gaussian random variable taking its values in $\mathbb{B}$ separable Banach space, with RKHS $\mathbb{H}$. Suppose $w \in \overline{\mathbb{H}}^{\mathbb{B}}$. Then for any $\varepsilon > 0$,

$$e^{-\varphi_w(\varepsilon/2)} \leq P(\|W - w\|_{\mathbb{B}} < \varepsilon) \leq e^{-\varphi_w(\varepsilon)}.$$

This result is particularly useful for Bayesian nonparametric arguments, as, if the $\|\cdot\|_{\mathbb{B}}$–norm can be related to the KL–type divergence defining the KL–type neighborhood $B_{KL}(f_0, \varepsilon)$, then the above result in particular can provide a lower bound on the prior mass term $\Pi[B_{KL}(f_0, \varepsilon_n)]$ appearing in the third condition of the GGV theorem.

Borell's inequality

Let $\mathbb{B}_1$ and $\mathbb{H}_1$ respectively denote the unit ball of $\mathbb{B}$ and of $\mathbb{H}$.

By definition $P[W \in \varepsilon\mathbb{B}_1] = P[\|W\|_{\mathbb{B}} < \varepsilon] = e^{-\varphi_0(\varepsilon)}$. Borell's inequality generalises this result. In words, it says that for large M, slightly enlarging $M\mathbb{H}_1$ by adding elements of norm (in $\mathbb{B}$) of at most ε, the resulting set captures most of the mass of W.

Theorem 4. [Borell] For W a Gaussian random variables taking values in $\mathbb{B}$ separable Banch space, for any $\varepsilon > 0$ and any $M > 0$, for $\Phi(u) = P(\mathcal{N}(0,1) \leq u)$,

$$P[W \in \varepsilon \mathbb{B}_1 + M\mathbb{H}_1] \geq \Phi(\Phi^{-1}(e^{-\varphi_0(\varepsilon)}) + M).$$

2.4 Pre-concentration theorem

Below we use the following standard inequality on the inverse of the Gaussian distribution function $\Phi(u) = \int_{-\infty}^{u}(e^{-u^2/2}/\sqrt{2\pi})du$: for any $0 < y < 1/2$,

$$0 > \Phi^{-1}(y) \geq -\sqrt{\frac{5}{2}\log(1/y)}.$$

Theorem 5. [van der Vaart and van Zanten 2008] Let W be a Gaussian random variable taking values in $\mathbb{B}$ separable Banach space, with RKHS $\mathbb{H}$. Let $w_0 \in \overline{\mathbb{H}}^{\mathbb{B}}$ and let $\varepsilon_n > 0$ be such that

$$\varphi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2. \tag{3.4}$$

Then for any $C > 1$ with $Cn\varepsilon_n^2 > \log 2$, there exists $B_n \subset \mathbb{B}$ measurable sets such that

$$\begin{aligned} (i) &\quad \log N(3\varepsilon_n, B_n, \|\cdot\|_{\mathbb{B}}) \leq 6Cn\varepsilon_n^2 \\ (ii) &\quad P[W \notin B_n] \leq e^{-Cn\varepsilon_n^2} \\ (iii) &\quad P[\|W - w_0\|_{\mathbb{B}} < 2\varepsilon_n] \geq e^{-n\varepsilon_n^2}. \end{aligned}$$

One notes that the result of this theorem curiously ressembles the assumptions of the GGV theorem of the first chapter...

Proof.

The inequality (iii) is a consequence of the theorem seen just before on probability of balls for Gaussian processes and their link to the concentration function:

$$P[\|W - w_0\|_{\mathbb{B}} < 2\varepsilon_n] \geq e^{-\varphi_{w_0}(\varepsilon_n)},$$

which combined with (3.4) leads to (iii).

In order to prove (ii), we define

$$B_n = \varepsilon_n \mathbb{B}_1 + M_n \mathbb{H}_1,$$

where M_n is to be chosen. By Borell's inequality,

$$P[W \notin B_n] \leq 1 - \Phi(\Phi^{-1}(e^{-\varphi_0(\varepsilon_n)}) + M_n).$$

By definition of the concentration function as a sum of two nonnegative terms, $\varphi_0(\varepsilon_n) \le \varphi_{w_0}(\varepsilon_n) \le n\varepsilon_n^2$ using (3.4). Let us set, for some $C > 1$,

$$M_n = -2\Phi^{-1}(e^{-Cn\varepsilon_n^2}).$$

Then we have, by monotonicity of Φ^{-1} and definition of M_n,

$$\Phi^{-1}(e^{-\varphi_0(\varepsilon_n)}) \ge \Phi^{-1}(e^{-n\varepsilon_n^2}) \ge -M_n/2.$$

Inserting this back into the previous upper-bound on $P[W \notin B_n]$ leads to

$$P[W \notin B_n] \le 1 - \Phi(M_n/2) = 1 - \Phi(-2\Phi^{-1}(e^{-Cn\varepsilon_n^2})) = e^{-Cn\varepsilon_n^2},$$

using that $\Phi(-x) = 1 - \Phi(x)$ for any real x, so (ii) is established.

It now remains to check (i). Let $h_1, \dots, h_N$ be elements of $M_n\mathbb{H}_1$ separated by at least $2\varepsilon_n$ in terms of the $\|\cdot\|_{\mathbb{B}}$ norm, and suppose this set of points is *maximal* (in the sense that N is the maximal number of $2\varepsilon_n$–separated points in $M_n\mathbb{H}_1$; the argument below shows that N is necessarily finite). The balls $h_1 + \varepsilon_n\mathbb{B}_1, \dots, h_N + \varepsilon_n\mathbb{B}_1$ are disjoint since the h_i's are $2\varepsilon_n$–separated. This implies

$$1 \ge P\left[W \in \bigcup_j (h_j + \varepsilon_n\mathbb{B}_1)\right] = \sum_{j=1}^N P[W \in h_j + \varepsilon_n\mathbb{B}_1].$$

Applying Proposition 1, since h_j's belong to $\mathbb{H}$, and using $\|h_j\|_{\mathbb{H}} \le M_n$, one gets

$$P[W \in h_j + \varepsilon_n\mathbb{B}_1] \ge e^{-\|h_j\|_{\mathbb{H}}^2/2} P[W \in \varepsilon_n\mathbb{B}_1] \ge e^{-M_n^2/2-\varphi_0(\varepsilon_n)}.$$

Inserting this into the previous inequality leads to

$$1 \ge Ne^{-M_n^2/2-\varphi_0(\varepsilon_n)},$$

from which one sees in particular that N must be finite. Deduce

$$N(2\varepsilon_n, M_n\mathbb{H}_1, \|\cdot\|_{\mathbb{B}}) \le N \le e^{M_n^2/2+\varphi_0(\varepsilon_n)}.$$

This implies

$$N(3\varepsilon_n, \varepsilon_n\mathbb{B}_1 + M_n\mathbb{H}_1, \|\cdot\|_{\mathbb{B}}) \le e^{M_n^2/2+\varphi_0(\varepsilon_n)}.$$

Let us now apply the standard inequality on Φ^{-1} recalled above (using $Cn\varepsilon_n^2 > \log 2$)

$$M_n = -2\Phi^{-1}(e^{-Cn\varepsilon_n^2}) \le 2\sqrt{\frac{5}{2}\log(e^{Cn\varepsilon_n^2})}.$$

Combining with the previous inequality on N, one obtains

$$N(3\varepsilon_n, B_n, \|\cdot\|_{\mathbb{B}}) \le e^{5Cn\varepsilon_n^2+\varphi_0(\varepsilon_n)} \le e^{6Cn\varepsilon_n^2},$$

using once again (3.4), which leads to (i) and concludes the proof.

Application: Gaussian white noise model

The Gaussian white noise model is

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

Recall that in this model, tests verifying condition (T) exist for the $\|\cdot\|_2$–norm, and that the neighborhood B_{KL} of the GGV theorem is just the L^2 ball $\{f : \|f - f_0\|_2 < \varepsilon_n\}$.

Theorem 6. Let $X^{(n)}$ be observations from the Gaussian white noise model.
Let Π be a prior distribution on $f \in L^2[0,1]$, defined as the distribution of a centered Gaussian random variable in $\mathbb{B} = L^2$, with RKHS $\mathbb{H}$.
Suppose the true $f_0 \in \overline{\mathbb{H}}^{\mathbb{B}}$ and let ε_n be such that

$$\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2,$$

where φ_{f_0} is the concentration function of W in $\mathbb{B} = L^2$.
Then for M large enough, as $n \longrightarrow \infty$,

$$E_{f_0}\Pi[\|f - f_0\|_2 > M\varepsilon_n \,|\, X^{(n)}] \longrightarrow 0.$$

Proof.

It is enough to note that the conclusion of Theorem 5 matches exactly the conditions of the GGV Theorem, noting that $d = \|\cdot\|_2$ and that the neighborhood B_{KL} of the GGV theorem is the L^2 ball $\{f : \|f - f_0\|_2 < \varepsilon_n\}$. The Theorem thus follows from the GGV theorem (up to setting $\varepsilon_n' = 2\varepsilon_n$ and noting that $C > 1$ can be taken arbitrarily large).

Application: Density estimation

In the density estimation model on $[0,1]$,

$$X^{(n)} = (X_1, \dots, X_n) \sim P_f^{\otimes n},$$

where P_f is the distribution of density f on $[0,1]$. For the next result, we work in $\mathbb{B} = C^0[0,1]$ space of continuous functions on $[0,1]$, equipped with the supremum norm $\|\cdot\|_\infty$. The next result implicitly assumes that $\log f_0$ is well–defined, that is, that f_0 is bounded from below.

Theorem 7. Let $X^{(n)}$ be observations from the density estimation model.
Let Π be a prior distribution on $f \in C^0[0,1] = \mathbb{B}$, defined as the distribution of

$$t \longrightarrow \frac{e^{W_t}}{\int_0^1 e^{W_u}du}, \tag{3.5}$$

where $(W_t, t \in [0,1])$ is a centered Gaussian process with continuous sample paths, with RKHS $\mathbb{H}$.
Let $w_0 := \log f_0$. Suppose $w_0 \in \overline{\mathbb{H}}^{\mathbb{B}}$. Suppose, for some $\varepsilon_n > 0$, we have

$$\varphi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2,$$

where φ_{f_0} is the concentration function of W in $\mathbb{B} = C^0[0,1]$.
Then for M large enough, as $n \longrightarrow \infty$,

$$E_{f_0}\Pi[h(f, f_0) > M\varepsilon_n \,|\, X^{(n)}] \longrightarrow 0.$$

Proof.

One can apply Theorem 5 to the function w_0: there exist sets B_n such that the conclusions (i)–(ii)–(iii) of that Theorem are satisfied.
Our goal is to verify the conditions of the GGV theorem with the Hellinger distance $d = h$.
For such B_n, let us set

$$\mathcal{F}_n := \left\{ f = \frac{e^w}{\int_0^1 e^{w(u)}du}, \quad \text{for } w \in B_n \right\}.$$

By (ii), we have $\Pi[\mathcal{F}_n^c] = P_W[\mathbb{B} \setminus B_n] \leq e^{-Cn\varepsilon_n^2}$, so the second condition of GGV is satisfied (we denote by P_W the distribution of the Gaussian process at the level of w's, while Π is the induced distribution at the level of densities f).

In order to verify the entropy and prior mass conditions of the GGV theorem, one needs to link the distance on w's to the distance on densities. This is done in Lemma 3.

From the first inequality in Lemma 3, one deduces that a covering of B_n by $3\varepsilon_n$–balls using the $\|\cdot\|_\infty$–metric induces a covering of $\mathcal{F}_n$ by $3\varepsilon_n e^{3\varepsilon_n/2}$–balls for the Hellinger distance h. For $\varepsilon_n \longrightarrow 0$ and large n, this implies

$$\log N(4\varepsilon_n, \mathcal{F}_n, h) \leq \log N(3\varepsilon_n, B_n, \|\cdot\|_\infty),$$

which means using (i) of Theorem 5 that the entropy condition of the GGV theorem is satisfied.

The second and third inequalities in Lemma 3 imply, for a large enough constant $K > 0$,

$$\Pi[B_{KL}(f_0, K\varepsilon_n)] \geq P_W[\|W - w_0\|_\infty \leq 2\varepsilon_n],$$

which is larger than $e^{-n\varepsilon_n^2}$ using (iii) of Theorem 5, which shows the prior mass condition is satisfied.

The result now follows from the GGV theorem.

Lemma 3. For any bounded functions v, w, if one denotes $p_v = e^v / \int_0^1 e^{v(u)} du$,

$$\begin{aligned} h(p_v, p_w) &\le \|v - w\|_\infty e^{\|v-w\|_\infty/2} \\ K(p_v, p_w) &\lesssim \|v - w\|_\infty^2 (1 + \|v - w\|_\infty) e^{\|v-w\|_\infty} \\ V(p_v, p_w) &\lesssim \|v - w\|_\infty^2 (1 + \|v - w\|_\infty)^2 e^{\|v-w\|_\infty}. \end{aligned}$$

Proof.

We prove the first inequality. For the second and third, we refer to the paper of van der Vaart of van Zanten (2008) (or to the book of Ghosal and van der Vaart (2017)).

$$\begin{aligned} h(p_v, p_w) &= \| \frac{e^{v/2}}{\|e^{v/2}\|_2} - \frac{e^{w/2}}{\|e^{w/2}\|_2} \|_2 \\ &= \| \frac{e^{w/2} - e^{v/2}}{\|e^{w/2}\|_2} + e^{v/2} \left(\frac{1}{\|e^{w/2}\|_2} - \frac{1}{\|e^{v/2}\|_2} \right) \|_2 \\ &\le 2 \frac{\|e^{w/2} - e^{v/2}\|_2}{\|e^{w/2}\|_2}. \end{aligned}$$

One can also bound from above

$$\begin{aligned} |e^{v/2} - e^{w/2}| &= e^{w/2} |e^{v/2 - w/2} - 1| \\ &\le e^{w/2} \| \frac{v - w}{2} \|_\infty e^{\|v-w\|_\infty/2}, \end{aligned}$$

where one uses the inequality $|e^x - 1| \le |x| e^{|x|}$, valid for all $x > 0$. Combining the previous bounds, one deduces that

$$h(p_v, p_w)^2 \le \frac{\int e^{w + \|v-w\|_\infty} \|v - w\|_\infty^2}{\int e^w},$$

which is no more than $\|v - w\|_\infty^2 e^{\|v-w\|_\infty}$, as requested.

Take-away message

The main take-away message from Theorem 5 and its applications in Thms 6 and 7 is that, when a Gaussian process is used as prior distribution (and provided the $\|\cdot\|_{\mathbb{B}}$–norm is easily related to the testing distance d, KL and V), the rate of convergence of the posterior distribution is essentially determined by solving the equation $\varphi_{w_0}(\varepsilon_n) \lesssim n\varepsilon_n^2$, where $w_0 = f_0$ in the white noise model (respectively $w_0 = \log f_0$ in density estimation). This message actually remains the same for many statistical models (classification, regression etc.) as well, see Chapter 11 of the book Ghosal and van der Vaart (2017).

We now see how the equation is solved in practice: we see the example of Brownian motion in details with proof, and give hints on the general picture for more general Gaussian processes, as well as on how to achieve statistical adaptation to unknown regularities.

2.5 Application: posterior rates with a Brownian motion prior

Let us consider the case of Brownian motion $W_t = B_t$ in the setting $(\mathbb{B}, \|\cdot\|_{\mathbb{B}}) = (C^0[0,1], \|\cdot\|_\infty)$ (so, with the density estimation application in mind; but the results are essentially identical in the L^2–setting).

As we mentioned earlier, the small ball probability of Brownian motion is well–known from the probability literature: one can show (we admit it), as $\varepsilon \longrightarrow 0$,

$$\varphi_0(\varepsilon) = -\log P[\|B\|_\infty < \varepsilon] \asymp \varepsilon^{-2}.$$

It remains to study the approximation term in the concentration function. This is done in the following Lemma. Recall that we have seen in the last lecture that the RKHS of Brownian motion on $[0,1]$ is $\{\int_0^\cdot g(u)du,\ g \in L^2[0,1]\}$, equipped with the Hilbert norm $\|\int_0^\cdot g\|_{\mathbb{H}}^2 = \|g\|_2^2$.

Lemma 4. Let $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ be the RKHS of Brownian motion.
Suppose $w_0 \in C^\beta[0,1]$, for some $\beta \in (0,1]$ and $w_0(0) = 0$.
Then

$$\inf_{h\in\mathbb{H}:\ \|h-w_0\|_\infty<\varepsilon} \|h\|_{\mathbb{H}}^2 \lesssim \varepsilon^{\frac{2\beta-2}{\beta}}.$$

We note that if $\beta \geq 1$, the result also holds with a constant replacing the power of ε, as then w_0 belongs to $\mathbb{H}$.

Proof.

We define a sequence $h \in \mathbb{H}$ that approximates w_0. The idea is to use a convolution. To do so, one notes that w_0 can be extended to $\mathbb{R}$ while keeping the Hölder-type property $|w_0(x) - w_0(y)| \lesssim |x-y|^\beta$. It suffices to prolongate w_0 by a constant outside of $[0,1]$.

Let $\phi_\sigma(u) = \phi(u/\sigma)/\sigma$, for $\sigma > 0$, and $\phi(u) = e^{-u^2/2}/\sqrt{2\pi}$ the Gaussian density. Let

$$h_\sigma(t) := (\phi_\sigma * w_0)(t) - (\phi_\sigma * w_0)(0),$$

with $\phi_\sigma * w_0(t) = \int_{\mathbb{R}} \phi_\sigma(t-u)w_0(u)du$.

Note that $h_\sigma(0) = 0$ and h_σ is a C^∞ map (because it is a convolution by a smooth function), so h_σ belongs to $\mathbb{H}$. We now evaluate

$$\begin{aligned}|\phi_\sigma * w_0(t) - w_0(t)| &= |\int \phi_\sigma(u)(w_0(t-u) - w_0(t))du \\ &\lesssim \int \phi_\sigma(u)|u|^\beta du \lesssim \sigma^\beta \int |v|^\beta \phi(v)dv \lesssim \sigma^\beta.\end{aligned}$$

Since $w_0(0) = 0$, we get a similar bound for $\phi_\sigma * w_0(0)$ by setting $t = 0$ in the previous inequality. This shows $\|h_\sigma - w_0\|_\infty \lesssim \sigma^\beta$.

On the other hand, $\|h_\sigma\|_{\mathbb{H}}^2 = \int_0^1 (h_\sigma)'(t)^2 dt$, where

$$\begin{aligned}|(h_\sigma)'(t)| &= |\int w_0(t-u)\frac{1}{\sigma^2}\phi'(u/\sigma)du| \\ &= |\int (w_0(t-u) - w_0(t))\frac{1}{\sigma^2}\phi'(u/\sigma)du| \quad (\text{as } \int \phi' = 0) \\ &\lesssim \sigma^{-2}\int |u|^\beta |\phi'(u/\sigma)|du \lesssim \sigma^{\beta-1}.\end{aligned}$$

The result follows by taking $\sigma \asymp \varepsilon^{1/\beta}$.

By gathering the small ball probability estimate and Lemma 4, one gets, with $a \vee b = \max(a, b)$,

$$\varphi_{w_0}(\varepsilon_n) \le \varepsilon_n^{-2} + C \vee \varepsilon_n^{(2\beta-2)/\beta}.$$

By equating this rate to $n\varepsilon_n^2$, one obtains, with $a \wedge b = \min(a, b)$,

$$\varepsilon_n \asymp n^{-1/4} \vee n^{-\beta/2} = n^{-\left\{\frac{1}{4}\wedge\frac{\beta}{2}\right\}}.$$

By using Theorem 7 in the density estimation model, with a normalised Brownian motion prior, one obtains that for any true density f_0 such that $f_0 = e^{w_0}$ with $w_0(0) = 0$, the posterior contraction rate is ε_n as above. To remove the condition $w_0(0) = 0$, one can consider 'Brownian motion released at zero', see below.

The rate is the fastest if $\beta = 1/2$, for which $\varepsilon_n \asymp n^{-1/4}$. When $\beta < 1/2$, the rate is $\varepsilon_n \asymp n^{-\beta/2}$: the approximation term (the 'bias') dominates in the contribution from the concentration function. When $\beta \ge 1/2$, the small ball probability term (analog of the 'variance') dominates.

It can be shown that the above rate cannot be improved for Brownian motion: it is the best one that one can get with this prior. From the minimax perspective, the rate ε_n above matches the minimax rate for estimating C^β functions, that is $n^{-\beta/(2\beta+1)}$ if and only if $\beta = 1/2$.

It follows also from the proof of Lemma 4 that $\overline{\mathbb{H}}^{\mathbb{B}}$ is the set of continuous functions f such that $f(0) = 0$ (one uses the proof for w_0 continuous and $w_0(0) = 0$, replacing the Hölder condition by absolute continuity of w_0). That is, almost all of $\mathbb{B}$ except for the restriction $f(0) = 0$. One can show that to obtain all of $\mathbb{H}$, it suffices to consider 'Brownian motion released at zero'

$$Z_t = B_t + Y,$$

with Y an $\mathcal{N}(0, 1)$ variable independent of (B_t). The RKHS of (Z_t) can be shown to be $H = \{c + \int_0^{\cdot} g(u)du,\ g \in L^2[0, 1], c \in \mathbb{R}\}$, for which $\overline{\mathbb{H}}^{\mathbb{B}} = \mathbb{B}$.

2.6 Application: posterior rates for other Gaussian processes

One can show a similar result as in the previous subsection for the Riemann-Liouville process $W_t = R_t^\alpha$. It can be checked that the obtained rate in this case is (for most α, or otherwise up to a log factor)

$$\varepsilon_n \asymp n^{-\frac{\alpha\wedge\beta}{2\alpha+1}}.$$

This rate is the same as the one we obtained for the first nonparametric example in the first lecture! Again, the rate is the optimal one (from the minimax perspective) if $\alpha = \beta$, but sub-optimal otherwise.

Another example of Gaussian process leading to this rate for $\mathbb{B} = L^2[0,1]$ is the random series prior

$$W_t = \sum_{j=1}^{\infty} \sigma_j \zeta_j e_j(t),$$

for $\sigma_j = j^{-1/2-\alpha}$, ζ_j a sequence of iid $\mathcal{N}(0,1)$ variables, and (e_j) an orthonormal basis of $L^2[0,1]$.

There are many other examples of Gaussian processes, for which the above theory can be applied: 'squared-expontial', Matern ... We refer to the book by Ghosal and van der Vaart (2017) for more information.

2.7 Adaptive inference using Gaussian processes

From the previous paragraphs, comparing the obtained posterior convergence rate with the minimax one $n^{-\beta/(2\beta+1)}$, it appears that the Gaussian process prior gives a posterior distribution that converges at optimal rate if its 'regularity' ($\alpha = 1/2$ for Brownian motion, $\alpha > 0$ for the Riemann Liouville process) matches the regularity β of the true function or density to be estimated.

This is encouraging, but in practice β is typically *unknown*. It turns out, that, similar to what we saw for random histograms (for regularities limited to $\beta \in [0,1]$ though), adding a single extra random variable to the Gaussian process allows the posterior to be adaptive: in a 2009 paper, van der Vaart and van Zanten proved that the following prior on functions f leads to adaptation (that is, the posterior automatically converges at optimal rate $n^{-\beta/(2\beta+1)}$ up to a logarithmic factor, without using a priori the knowledge of β):

1. A is a positive random variable with Gamma distribution
2. Given A, and a certain Gaussian process Z, one considers the random function

$$f : t \longrightarrow Z_{At}.$$

The proof of adaptation uses similar arguments as that of Theorem 5, but one needs a more refined argument to make the dependence on A explicit. For more details, we refer to the paper van der Vaart and van Zanten (2009).

3 Bayesian methods for high–dimensional models

We introduce so–called high–dimensional statistical models, where the number of parameters can be equal or larger to the number of observations. Inference is then typically made possible in these models by appealing to a sparsity *assumption, which is often realistic in applications. The main difficulty is that the sparsity pattern is not known in advance. Classes of parsimonious prior distributions are introduced, including the famous spike–and–slab prior. We then give two results of concentration of posteriors in sparse settings, each corresponding to a different method of proof.*

3.1 Introduction

Since the 2000's practical applications where the number of unknown parameters is 'large', even possibly much larger than the number of observations, have become commonplace. Although it may seem paradoxical at first to be able to solve or even say something in such 'difficult' settings, a key pattern that has emerged in the study of these models is that of *sparsity*. Namely, although the number of parameters is very large, possibly only a few are really significant.

Sparsity

One very commonly assumed form of sparsity is the following: the true θ_0 belongs to the *nearly-black* class

$$\ell_0[s] = \{\theta \in \mathbb{R}^n \,:\, \#\{i \,:\, \theta_i \neq 0\} \leq s\} \tag{3.6}$$

for $0 \leq s \leq n$, where # stands for the cardinality of a finite set. This means that only s out of n coordinates of θ are nonzero (but we do not know which ones), and typically it is assumed that $s = o(n)$, so only a very small number of coordinates of θ have 'signal', that is are nonzero. In the sequel we assume $s_n \longrightarrow \infty$ and $s_n = o(n)$ as $n \longrightarrow \infty$.

Some high-dimensional sparse models

The simplest high-dimensional model is given by the normal sequence model:

$$X_i = \theta_i + \epsilon_i, \quad i = 1, \dots, n, \tag{3.7}$$

where ε_i are i.i.d. $\mathcal{N}(0,1)$, the parameter set Θ for $\theta = (\theta_1, \dots, \theta_n)$ is $\mathbb{R}^n$ but θ is assumed to be sparse in the sense that it belongs to one of the sets $\ell_0[s]$ for some $0 \leq s \leq n$.

It is known that the optimal convergence rate in model (3.7) over $\ell_0[s]$, in terms of asymptotic minimax risk for the squared error loss, is $2s\log(n/s)$. That is, if $\|\theta\|^2 = \sum_{i=1}^n \theta_i^2$ is the (squared)–L^2 norm,

$$\inf_T \sup_{\theta_0 \in \ell_0[s]} E_{\theta_0}[\|T(X) - \theta\|^2] = 2(1 + o(1))s\log(n/s),$$

where $T = T(X)$ is an estimator of θ based on the observation of $X = (X_1, \dots, X_n)$.

This model is a special case of the high-dimensional Gaussian linear regression model

$$Y = X\theta + \varepsilon, \tag{3.8}$$

where $\theta \in \mathbb{R}^p$, the noise vector ε follows a $\mathcal{N}(0, \sigma^2 I_n)$ distribution and X is a $n \times p$ matrix with real coefficients. The 'high-dimensional case' corresponds to $n \leq p$, possibly $n = o(p)$. In that case, one typically assumes some form of sparsity such as $\theta \in \ell_0[s]$, for some $s = o(n)$.

In the sequel, for simplicity of presentation we focus on the simpler sequence model (3.7), although most the obtained results can be transferred to the regression model (3.8) by assuming appropriate conditions on the design matrix X.

The need for prior modelling

In the case where Θ is a subset of $\mathbb{R}^n$, the simplest prior that comes to mind is $\Pi = \otimes_{i=1}^n G$, making the coordinates of θ independent of distribution G on $\mathbb{R}$. However, from the point of view of the posterior distribution, this unstructured prior is often not suitable. Consider for instance model (3.7) and let us endow θ with the a product of Laplace (double-exponential) priors

$$\Pi_\lambda = \bigotimes_{i=1}^n \mathrm{Lap}(\lambda/2), \qquad \lambda > 0.$$

For this choice, the posterior mode (that is, the mode of the posterior density) is [exercise: check it]

$$\hat{\theta}_\lambda^L = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \left[\|X - \theta\|_2^2 + \lambda \|\theta\|_1 \right].$$

This is nothing but the classical LASSO estimator. In the special case of model (3.7), for the choice $\lambda = \lambda^* \asymp \sqrt{\log n}$, the LASSO achieves the minimax rate

$$\sup_{\theta \in \ell_0[s]} E_\theta [\|\hat{\theta}_{\lambda^*}^L - \theta\|^2] \lesssim s \log n,$$

up to the form of the log factor. However, if the true $\theta_0 = 0$, for small $\delta > 0$, one can show that (we do not prove it here)

$$E_0 \Pi_{\lambda^*} \left[\|\theta\|^2 \leq \delta \frac{n}{\log n} \,|\, Y \right] \longrightarrow 0.$$

This means that the "LASSO–posterior distribution" $\Pi_{\lambda^*}[\cdot \,|\, X]$ is suboptimal over sparse classes $\ell_0[s]$ for $s \ll n/\log^2 n$. The intuition behind this result is that, although its mode is the LASSO and is thus sparse, the LASSO–posterior as a probability distribution is not sparse. A sample from $\Pi_{\lambda^*}[\cdot \,|\, X]$ almost surely sets no coordinate of θ to 0. From the Bayesian perspective, this means that one needs to take structural assumptions such as sparsity into account when proposing a prior distribution.

3.2 Sparse priors

Spike–and–slab: priors with many exact zeroes

Spike–and–slab priors. For $w \in [0, 1]$ and Γ a distribution on $\mathbb{R}$, the prior

$$\Pi_\alpha = \Pi_{\alpha,\Gamma} = \bigotimes_{i=1}^n (1 - \alpha)\delta_0 + \alpha\Gamma, \tag{3.9}$$

where δ_0 is the Dirac mass at 0, is called spike and slab (SAS) with parameter $\alpha \in [0,1]$ and slab distribution Γ. To inforce sparsity, one may choose a deterministic α: the choice $\alpha = 1/n$ is standard and implies that under the prior, the expected number of nonzero coefficients is of the order of a constant. To obtain an improved data fit, options performing better in practice include an empirical Bayes choice $\hat{\alpha}$ of α (more on that below), or hierarchical Bayes, where α is itself given a prior, for instance a Beta distribution, e.g. $\text{Beta}(1, n+1)$.

Subset–selection priors. For π_n a prior on the set $\{0,1,2,\dots,n\}$ and $\mathcal{S}_k$ the collection of all subsets of $\{1,\dots,n\}$ of size k, let Π be constructed as

$$k \sim \pi_n, \qquad S \,|\, k \sim \text{Unif}(\mathcal{S}_k), \qquad \theta \,|\, S \sim \bigotimes_{i\in S} \Gamma \otimes \bigotimes_{i\notin S} \delta_0. \tag{3.10}$$

The spike–and–slab prior (3.9) is a particular case where π_n is the binomial $\text{Bin}(n,\alpha)$ distribution. Through the prior π_n, it is possible to chose dimensional priors that 'penalise' more large dimensions than the binomial, for instance the complexity prior $\pi(k) \propto \exp(-ak\log(bn/k))$.

Continuous shrinkage priors

While subset selection priors are particularly appealing in view of their naturally built-in model selection, one may instead use prior distributions that do not put any coefficient exactly to 0 but instead draw either very small values or intermediate/strong ones. A way to do so is to replace the Dirac mass in (3.9) by an absolutely continuous distribution with density having a high or infinite density at zero.

Spike and slab LASSO. This prior replaces δ_0, Γ by two Laplace distributions $\text{Lap}(\lambda_0), \text{Lap}(\lambda_1)$ with λ_0 large, typically going to ∞ with n to enforce (near–)sparsity and λ_1 a constant, that is

$$\Pi_\alpha = \Pi_{\alpha,\Gamma} = \bigotimes_{i=1}^{n} (1-\alpha)\text{Lap}(\lambda_0) + \alpha\text{Lap}(\lambda_1),$$

where α, as above for the spike and slab prior, to be chosen.

Horseshoe prior. Leaving finite mixtures, one may also consider continuous mixtures: a popular choice is the *horseshoe* prior of Carvalho, Polson and Scott (2010), which is a continuous scale–mixture of Gaussians: given a parameter $\tau > 0$ to be chosen, this horseshoe draws coordinates θ_i independently as

$$\theta_i \,|\, \lambda_i \sim \mathcal{N}(0, \lambda_i^2), \qquad \lambda_i \sim C^+(0,\tau), \tag{3.11}$$

where $C^+(0,\tau)$ is a half-Cauchy distribution with scale τ. This leads to a marginal density h in θ that satisfies, with $K = 1/\sqrt{2\pi^3}$,

$$\frac{K}{2\tau}\log\left(1 + \frac{4\tau^2}{\theta^2}\right) \le h(\theta) \le \frac{K}{\tau}\log\left(1 + \frac{2\tau^2}{\theta^2}\right).$$

This density has a pole at 0 and Cauchy–tails. One may also consider different priors on the scales λ_i. This is considered in the paper Salomond, Schmidt-Hieber and van der Pas (2016).

Back to spike–and–slab: on the choice of α

All previously introduced sparse prior distribution have a tuning parameter: α for the spike–and–slab (SAS) prior, the parameter(s) of the prior on dimension π_n for the subset selection prioe, the parameter τ for the horseshoe. The performance of the posterior distributions associated to these priors in model (3.7) are quite sensitive the choice of these parameters, as we will see below. We now present some possibilities to chose them, focusing on the case of α for the SAS prior (the discussion being quite similar in the other cases).

Deterministic value of α.

1. Choice $\alpha = 1/n$. Under this prior, the expected (prior) number of nonzero coefficients is $n * 1/n = 1$. We will prove below that this choice already leads to a nearly-optimal convergence rate for the associated posterior distribution. However, one can find estimators with a significantly better behaviour, in particular for n not so large. The priors with $\alpha = 1/n^b$, $b \geq 1$, behave similarly.

2. 'Oracle choice' $\alpha = s_n/n$. Under this prior, the expected (prior) number of nonzero coefficients is $n * s_n/n = s_n$, that is, the 'correct' one. We will prove below that this choice leads to an optimal convergence rate for the associated posterior distribution. However, it requires the knowledge of s_n, which is typically unknown in practice

Empirical Bayes. One possibility is to replace α by an ad-hoc estimator of the number of non-zero coordinates of θ, for instance by keeping only coordinates above the expected noise level

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} 1\!\!1\{|X_i| > \sqrt{2\log n}\}.$$

However this choice may be too *conservative* in that signals below the universal $\sqrt{2\log n}$ threshold may not be detected.

The marginal maximum likelihood empirical Bayes approach (MMLE) consists in forming a likelihood in terms of the parameter of interest (here α) by integrating out the parameter θ. In model (3.7), for a spike–and–slab prior and fixing the distribution Γ with density γ, we have $\theta \,|\, \alpha \sim \Pi_\alpha$ and $X_i \,|\, \theta \sim P_\theta$ independent normals. Then the Bayesian distribution of X given α has density $\int \prod_i p_\theta(X_i) d\Pi_\alpha(\theta)$ [Exercise: check it]. The maximisation of the corresponding 'likelihood' leads to, with $g = \gamma * \phi$,

$$\hat{\alpha} = \text{argmax} \prod_{i=1}^{p} \left((1-\alpha)\phi(X_i) + \alpha g(X_i)\right).$$

The plug-in posterior $\Pi_{\hat{\alpha}}[\cdot \,|\, X]$ has been advocated and studied in George and Foster (2000) and Johnstone and Silverman (2004) among others.

Hierarchical Bayes. In this approach, one draws α at random

$$\theta \,|\, \alpha \sim \Pi_\alpha$$
$$\alpha \sim \pi_H,$$

where π_H is some distribution on $[0,1]$. It can be checked that for Π_α the SAS prior of parameter α, taking $\pi_H = \text{Beta}(1, n+1)$ leads to a posterior distribution contracting at the optimal rate.

3.3 Posterior convergence for spike–and–slab priors

Posterior distribution with fixed α

We consider the following choices

$$\Gamma = \begin{cases} & \text{Lap}(1) \\ or & \\ & \text{Cauchy}(1) \end{cases}$$

where Lap(λ) denotes the Laplace (double exponential) distribution with parameter λ and Cauchy(1) the standard Cauchy distribution. Different choices of parameters and prior distributions are possible but for clarity of exposition we stick to these common distributions. In the sequel γ denotes the density of Γ with respect to the Lebesgue measure.

By Bayes' formula the posterior distribution under (3.7) with fixed $\alpha \in [0,1]$ is

$$\Pi_\alpha[\cdot \,|\, X] \sim \bigotimes_{i=1}^{n} (1 - a(X_i))\delta_0 + a(X_i)G_{X_i}(\cdot), \tag{3.12}$$

where, denoting by ϕ the standard normal density and $g(x) = \phi * \Gamma(x) = \int \phi(x-u)d\Gamma(u)$ the convolution of ϕ and Γ at point $x \in \mathbb{R}$, the posterior weight $a(X_i)$ is given by, for any i,

$$a(X_i) = a_\alpha(X_i) = \frac{\alpha g(X_i)}{(1-\alpha)\phi(X_i) + \alpha g(X_i)}. \tag{3.13}$$

The distribution G_{X_i} has density

$$\gamma_{X_i}(\cdot) := \frac{\phi(X_i - \cdot)\gamma(\cdot)}{g(X_i)} \tag{3.14}$$

with respect to Lebesgue measure on $\mathbb{R}$. The behaviour of the posterior distribution $\Pi_\alpha[\cdot \,|\, X]$ heavily depends on the choices of the smoothing parameters α and γ. It turns out that some aspects of this distribution are thresholding-type estimators, as established in Johnstone and Silverman (*Annals of Statistics*, 2004).

Posterior median and threshold $t(\alpha)$. The posterior median $\hat\theta_\alpha^{med}(X_i)$ of the ith coordinate has a thresholding property: there exists $t(\alpha) > 0$ such that $\hat\theta_\alpha^{med}(X_i) = 0$ if and only if $|X_i| \leq t(\alpha)$, see Johnstone and Silverman (2004). A default choice can be $\alpha = 1/n$; one can check that this leads to a posterior median behaving similarly as a hard thresholding estimator with threshold $\sqrt{2\log n}$. One can improve on this default choice by taking a well-chosen data-dependent α.

A generic posterior convergence result

Let us work with the subset-selection prior (3.15).

$$k \sim \pi_n, \qquad S \,|\, k \sim \text{Unif}(\mathcal{S}_k), \qquad \theta \,|\, S \sim \bigotimes_{i \in S} \Gamma \otimes \bigotimes_{i \notin S} \delta_0. \tag{3.15}$$

Let us denote, for $\theta \in \mathbb{R}^n$, by S_θ its support

$$S_\theta = \{i \,:\, \theta_i \neq 0\},$$

that is the indices of its nonzero coordinates. We denote $S_0 = S_{\theta_0}$.

Definition 6. We say that a prior on dimension π_n as in (3.15) has *exponential decrease* if there exists a constant $D < 1$ such that, for any $k \geq 1$,

$$\pi_n(k) \leq D\pi_n(k-1).$$

Examples. The prior $\pi_n(k) \propto e^{-k}$ and the prior $\pi_n(k) \propto e^{-ak\log(bn/k)}$ for $a > 0, b > 1 + e$, both have exponential decrease [Exercise]. For binomial priors on dimension, $\pi_n = \mathrm{Bin}(n, \alpha)$, it can be checked that they also verify the exponential decrease if $n\alpha \lesssim s_n$, which is verified for the choices $\alpha = 1/n$ and $\alpha = s_n/n$. It can also be verified, see Castillo and van der Vaart (2012), that the beta-binomial prior

$$\begin{aligned} k\,|\,\alpha &\sim \mathrm{Bin}(n, \alpha) \\ \alpha &\sim \mathrm{Beta}(1, n+1) \end{aligned}$$

verifies the exponential decrease property.

Theorem 8. Take a prior Π as in (3.15) with $\Gamma = \mathrm{Lap}(1)$ and suppose, for $S_0 = S_{\theta_0}$, for some constant $d > 0$,

$$\Pi(S_0) \geq e^{-ds_n\log n}.$$

Assume that the prior on dimension π_n verifies the exponential decrease as in Definition 6. Then for M large enough, as $n \longrightarrow \infty$,

$$\sup_{\theta_0\in\ell_0[s_n]} E_{\theta_0}\Pi[\|\theta-\theta_0\|^2 > Ms_n\log n\,|\,X] = o(1).$$

Corollary. It is not hard to check that all the examples of priors mentioned just above the statement of Theorem 8 verify, combined with the uniform prior on subsets as in (3.15), the condition on $\Pi(S_0)$ from the Theorem. So they all lead to a posterior convergence rate at least of the order $s_n\log n$. By a more precise argument, one can in fact prove (see Castillo and van der Vaart (2012)) that both priors on dimension $\pi_n(k) \propto e^{-k}$ and the Beta-binomial prior as above both lead to a posterior convergence rate of $Ms_n\log(n/s_n)$ for large M (so, with the precise logarithmic) factor. This shows that both priors achieve the optimal rate in an adaptive way. Note that we also 'recover' (although the results are not exactly equivalent as noted above), the posterior contraction rate obtained in Theorem 9 for the binomial priors on dimension – that lead to the SAS prior construction and vice-versa – with $\alpha = 1/n$ or $\alpha = s_n/n$ (up to the form of the log factor for the latter).

Proof of Theorem 8

Proof.

We follow ideas similar to that of the proof of the GGV theorem, but do not do the testing/entropy part explicitly. For any $C \subset \mathbb{R}^n$ measurable, Bayes' formula can be written

$$\Pi[C\,|\,X] = \frac{\int_C \exp\{-\|X-\theta\|^2/2\}d\Pi(\theta)}{\int \exp\{-\|X-\theta\|^2/2\}d\Pi(\theta)} = \frac{\int_C \exp\{\Delta(\theta,X)\}d\Pi(\theta)}{\int \exp\{\Delta(\theta,X)\}d\Pi(\theta)},$$

where we have set $\Delta(\theta,X) = -\|\theta-\theta_0\|^2/2 + \langle \theta-\theta_0, X-\theta_0\rangle$. [One may remark that the parameter set of sparse vectors is unbounded, so one cannot hope for a uniform bound from below of the denominator independent of how 'large' the coordinates of θ_0 are. In fact, the bound below features an L^1 norm $\|\theta_0\|_1$; fortunately, this can be compensated from a similar term appearing in the bound for the numerator as seen below]

Now let $D := \int \exp\{\Delta(\theta,X)\}d\Pi(\theta)$ be the denominator on the last term of the display on Bayes' formula above, and $N = \int_C \exp\{\Delta(\theta,X)\}d\Pi(\theta)$ the numerator, with C to be chosen below. We start by a bound on D.

Bound on the denominator D. Let us set, for $r_n \longrightarrow \infty$ to be chosen,

$$B = \{\theta\ :\ \|\theta-\theta_0\| \le r_n\}.$$

By restricting the integral on the denominator D to the set B,

$$D \ge \Pi(B)\int e^{\Delta(\theta,X)}d\bar{\Pi}(\theta),$$

where $\bar{\Pi}$ is the probability distribution $\bar{\Pi}(\cdot) = \Pi(\cdot \cap B)/\Pi(B)$. Let us now apply Jensen's inequality with the exponential map to obtain

$$\int_B e^{\Delta(\theta,X)}d\bar{\Pi}(\theta) \ge \exp\left\{\int_B \Delta(\theta,X)d\Pi(\theta)\right\}.$$

Noting that, on B, we have $-\|\theta-\theta_0\|^2/2 \ge r_n^2/2$, and setting

$$Z := \langle \int_B (\theta-\theta_0)d\bar{\Pi}(\theta), X-\theta_0\rangle = \int_B \langle \theta-\theta_0, X-\theta_0\rangle d\bar{\Pi}(\theta)$$

by linearity, one gets

$$D \ge \Pi(B)e^{-\frac{r_n^2}{2}+Z}.$$

Now under P_{θ_0}, we have $X-\theta_0 = \varepsilon \sim \mathcal{N}(0,I_n)$, with I_n the identity matrix in dimension n. In particular,

$$\langle \theta-\theta_0, X-\theta_0\rangle \sim \mathcal{N}(0,\|\theta-\theta_0\|^2).$$

So, $E_{\theta_0}Z = 0$ (Fubini), and by Jensen's inequality again this time applied with $x \longrightarrow x^2$, and Fubini,

$$\begin{aligned} E_{\theta_0}Z^2 &\le \int_B E_{\theta_0}\langle \theta-\theta_0, X-\theta_0\rangle^2 d\bar{\Pi}(\theta) \\ &= \int_B \|\theta-\theta_0\|^2 d\bar{\Pi}(\theta) \le r_n^2, \end{aligned}$$

using the definition of B. By Markov's inequality, $P_{\theta_0}[|Z| > r_n^2] \leq r_n^{-4} E_{\theta_0} Z^2 \leq r_n^{-2}$. One deduces that on an event $\mathcal{A}$ of P_{θ_0}–probability at least $1 - (1/r_n^2)$, we have

$$D \geq \Pi(B) e^{-3r_n^2/2}.$$

Let us now focus on $\Pi(B)$. Note that an equivalent way of writing the prior Π is as follows

$$\Pi = \sum_S Q(S)\Pi_S, \quad \Pi_S := \bigotimes_{i=1}^n \Pi_{S,i},$$

where $\Pi_{S,i} = \text{Lap}(1)$ if $i \in S$ and $\Pi_{S,i} = \delta_0$ otherwise.

Denoting by $S_0 := S_{\theta_0}$ the support of θ_0, and $\theta_S = (\theta_i, i \in S)$ for $S \subset \{1, \dots, n\}$, and for B as above, with $\|u\|_1 = \sum_{i=1}^n |u_i|$,

$$\begin{aligned}
\Pi(B) &\geq Q(S_0)\Pi_{S_0}[\|\theta - \theta_0\| \leq r_n] \geq Q(S_0)\Pi_{S_0}[\|\theta - \theta_0\|_1 \leq r_n] \\
&\geq Q(S_0) \int_{\|\theta-\theta_0\|_1 \leq r_n} \prod_{i \in S_0} \frac{1}{2} e^{-|\theta_i|} d\theta_i \\
&\geq Q(S_0) 2^{-s_n} \int_{\|\theta_{S_0} - \theta_0\|_1 \leq r_n} e^{-\|\theta_{S_0} - \theta_0\|_1 - \|\theta_0\|_1} d\theta_{S_0} \\
&\geq Q(S_0) 2^{-s_n} e^{-\|\theta_0\|_1} \int_{\|\theta_{S_0}\|_1 \leq r_n} e^{-\|\theta_{S_0}\|_1} d\theta_{S_0},
\end{aligned}$$

where we have used that $\|u\|_2 \leq \|u\|_1$ for $u \in \mathbb{R}^d$, $d \geq 1$ and invariance by translation of Lebesgue's measure [note: one can also use a lower bound involving volumes of balls]. Next, as $\{|\theta_i| \leq r_n/s_n, \ i \in S_0\} \subset \{\|\theta_{S_0}\|_1 \leq r_n\}$, with $s_0 = |S_0| \leq s_n$,

$$\int_{\|\theta_{S_0}\|_1 \leq r_n} e^{-\|\theta_{S_0}\|_1} d\theta_{S_0} \geq \left(\int_{-r_n/s_n}^{r_n/s_n} e^{-|u|} du \right)^{s_0} \gtrsim (Ce^{-r_n/s_n})^{s_0} \gtrsim e^{-r_n}.$$

From the previous bounds one deduces that, on the event $\mathcal{A}$,

$$D \gtrsim Q(S_0) e^{-3r_n^2/2 - r_n - Cs_n} e^{-\|\theta_0\|_1}.$$

Bound on the numerator N. Recall that $N := \int_C \exp\{\Delta(\theta, X)\} d\Pi(\theta)$, and we define C now as $C := C_1 \cap C_2$, with, for $K, M > 0$, with $|A|$ the cardinality of A,

$$C_1 := \{\theta : |S_\theta| \leq Ks_n\}, \quad C_2 := \{\theta : \|\theta - \theta_0\| > Mr_n\}.$$

Let us also define the event

$$\mathcal{A}_N = \left\{ \max_{1 \leq i \leq n} |\varepsilon_i| \leq \sqrt{2 \log n} \right\}.$$

A union bound gives, using the standard bound $\bar{\Phi}(x) \leq \phi(x)/x$, for $x > 0$,

$$P(\mathcal{A}_N^c) \leq 2n\bar{\Phi}(\sqrt{2 \log n}) = o(1)$$

Also, noting that if $\theta \in C_1$ we have $|S_{\theta-\theta_0}| \leq (K+1)s_n$, for any such θ on the event $\mathcal{A}_N$,

$$\begin{aligned} |\langle \theta - \theta_0, X - \theta_0 \rangle| &\leq \|\theta - \theta_0\|_1 \max_{1\leq i\leq n} |X_i - \theta_{0,i}| \\ &\leq |S_{\theta-\theta_0}| \|\theta - \theta_0\| \max_{1\leq i\leq n} |\varepsilon_i| \\ &\leq \sqrt{2(K+1)s_n \log n} \|\theta - \theta_0\|, \end{aligned}$$

where the second line uses Cauchy-Schwarz inequality. Now using the inequality $2ab \leq \delta^{-1}a^2 + \delta b^2$, one finds that, on $\mathcal{A}_N$ and for large enough $c_2 > 0$,

$$|\langle \theta - \theta_0, X - \theta_0 \rangle| \leq c_2 s_n \log n + \frac{\|\theta - \theta_0\|^2}{4}.$$

Inserting this into the definition of N leads to, on the event $\mathcal{A}_N$,

$$\begin{aligned} N &\leq \int_C e^{-\|\theta-\theta_0\|^2/4 + c_2 s_n \log n} d\Pi(\theta) \\ &\leq e^{c_2 s_n \log n} \sum_S Q(S) \int_{C_S} e^{-\|\theta_S - \theta_0\|^2/4} \frac{e^{-\|\theta_S\|_1}}{2^{|S|}} d\theta_S, \end{aligned} \tag{3.16}$$

where $C_S = \{\theta : S_\theta = S\} \cap C$, using that by definition the prior on the selected subset is product Laplace. By the triangle inequality,

$$-\|\theta_S\|_1 \leq -\|\theta_{0,S}\|_1 + \|\theta_S - \theta_{0,S}\|_1. \tag{3.17}$$

One bounds each term on the right hand side. Cauchy-Schwarz implies, on C,

$$\|\theta_S - \theta_{0,S}\|_1 \leq \sqrt{K s_n} \|\theta_S - \theta_0\|.$$

On the other hand, we have

$$-\|\theta_{0,S}\|_1 \leq -\|\theta_0\|_1 + \|\theta_{0,S_0 \cap S^c}\|_1 \leq -\|\theta_0\|_1 + \sqrt{s_n} \|\theta_S - \theta_0\|,$$

as indeed, using again Cauchy-Schwarz,

$$\begin{aligned} \|\theta_{0,S_0\cap S^c}\|_1 &\leq |S_0 \cap S^c|^{1/2} \|\theta_{0,S_0\cap S^c}\| \\ &\leq \sqrt{s_n} \|\theta_S - \theta_0\|. \end{aligned}$$

Inserting back the obtained bounds in (3.17) one gets

$$-\|\theta_S\|_1 \leq -\|\theta_0\|_1 + (\sqrt{K} + 1)\sqrt{s_n} \|\theta_S - \theta_0\|.$$

The last term is bounded by $K' s_n + \|\theta_S - \theta_0\|^2/8$ for large enough K'. Now the integral in (3.16) is bounded from above by

$$e^{-\|\theta_0\|_1 - M^2 r_n^2/16 + K' s_n} \int_{C_S} e^{-\|\theta_S - \theta_0\|^2/16} d\theta_S.$$

By definition, C_S contains only vectors of support at most Ks_n, so that

$$\int_{C_S} e^{-\|\theta_S-\theta_0\|^2/16} d\theta_S \le \int_{\mathbb{R}^{|S|}} e^{-\|\theta_S\|^2/16} d\theta_S \le C^{Ks_n}.$$

This gives the following bound on the numerator, on the event $\mathcal{A}_N$,

$$N \le e^{c_2 s_n \log n - M^2 r_n^2/16 + C' s_n} e^{-\|\theta_0\|_1}.$$

Putting the bounds together. Gathering the previous bounds gives, on $\mathcal{A} \cap \mathcal{A}_N$,

$$\frac{N}{D} \le e^{3r_n^2/2 + r_n + c_3 s_n \log n - M^2 r_n^2/16},$$

where one uses the assumption on $Q(S_0)$. By choosing $r_n^2 = s_n \log n$ and taking M large enough, one deduces, on $\mathcal{A} \cap \mathcal{A}_N$, that

$$N/D \le e^{-M^2 r_n^2/32} = o(1),$$

which implies that $E_{\theta_0}\Pi[C_1 \cap \{\theta : \|\theta - \theta_0\| > Mr_n\} \,|\, X] = o(1)$. Combining this with Lemma 5 for K large enough, one obtains that $E_{\theta_0}\Pi[\|\theta - \theta_0\| > Mr_n \,|\, X] = o(1)$, which concludes the proof of Theorem 8.

Lemma 5. Suppose the prior on dimension π_n in (3.15) verifies the exponential decrease property from Definition 6 and that Γ has a centered density with finite second moment. Then, for $S_\theta = \{i : \theta_i \ne 0\}$ the support of the vector θ, for K large enough, as $n \longrightarrow \infty$,

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0}\Pi[\theta : |S_\theta| > Ks_n \,|\, X] = o(1).$$

Proof.

See Proposition 4.1 and Lemma 4.1 in Castillo and van der Vaart (2012).

3.4 Applications to inference in high-dimensional models

We briefly mention a few applications of the previous theory. Of course, each specific applications requires specific extra work, but the results we have seen so far an important first step towards them in the Bayesian setting. Much research remains to be done in this very active area.

1. *Estimation and prediction of sparse vectors.* In the previous sections, we have seen how to estimate θ in the sparse Gaussian sequence model using a Bayesian approach with sparse priors. Analogous results can be obtained in the high-dimensional regression model under assumptions on X, either on estimation of θ there, or on estimation of $X\theta$ (the latter typically requires less assumptions and is called the *prediction* task).

2. *Confidence sets for θ or for coordinates of θ.* One may try to use regions that get high posterior probability (so–called credible sets) to get confidence sets having specific coverage and small-

est possible diameter. This is possible only to a certain extent, and one may have to either make assumptions on the sparsity parameter, or to allow for larger regions, to be able to solve the problem. Such questions are considered, for instance, in Szabo, van der Pas and van der Vaart (2017), and Castillo and Szabo (2020).

3. *Variable selection and Multiple testing.* In applications such as genomics, one very important practical question is to select a subset of coordinates that contain 'signal', with a certain control of the number of false positive (and possibly also false negative). In the sparse sequence model, this can be done through the famous Benjamini-Hochberg procedure (BH procedure), but also in a Bayesian way via, for instance, an empirical Bayes posterior, as suggested by Efron (2007), and investigated from the frequentist perspective in Castillo and Roquain (2020).

3.5 Complement

Note: this section was not covered in the class and can be skipped.

The following results help understanding the role of α in the previous result: they quantify the behaviour of the posterior when α is fixed.

Convergence of posterior for fixed α

Expected posterior L^2–squared risk. For a fixed weight α, the posterior distribution of θ is given by (3.12). On each coordinate, the mixing weight $a(X_i)$ is given by (3.13) and the density of the non-zero component γ_{X_i} by (3.14). In the sequel we will obtain bounds on the following quantity, already for a given $\alpha \in [0,1]$,

$$\int \|\theta - \theta_0\|^2 d\Pi_\alpha(\theta \,|\, X) = \sum_{i=1}^n \int (\theta_i - \theta_{0,i})^2 d\Pi_\alpha(\theta_i \,|\, X_i).$$

To do so, we study $r_2(\alpha, \mu, x) := \int (u-\mu)^2 d\pi_\alpha(u \,|\, x)$, where $\pi_\alpha(\cdot \,|\, x) \sim (1 - a(x))\delta_0 + a(x)\gamma_x(\cdot)$ using (3.12). By definition

$$r_2(\alpha, \mu, x) = (1 - a(x))\mu^2 + a(x) \int (u-\mu)^2 \gamma_x(u) du.$$

This quantity is controlled by $a(x)$ and the term involving γ_x. From the formula for the posterior weight $a(x)$ in (3.13), bounding the denominator from below by one of its two components, and using $a(x) \le 1$ yields, for any real x and $\alpha \in [0,1]$,

$$\alpha \frac{g}{g \vee \phi}(x) \le a(x) \le 1 \wedge \frac{\alpha}{1-\alpha} \frac{g}{\phi}(x). \tag{3.18}$$

Theorem 9. In the sparse sequence model, let us consider a spike–and–slab (SAS) prior with fixed $\alpha \in (0,1)$. Let Γ be either the standard Laplace or Cauchy distribution. Then there exist $\alpha_0 > 0$, $N_0 \ge 2$ and $C > 0$ such that for any $\alpha \le \alpha_0$ and $n \ge N_0$,

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_\alpha(X) \le Cn\alpha\sqrt{\log(1/\alpha)} + Cs_n(1 + \log(1/\alpha)).$$

As a particular case, one obtains, for n large enough,

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_\alpha(X) \leq \begin{cases} Cs_n \log n & \text{if } \alpha = 1/n, \\ Cs_n \log(n/s_n) & \text{if } \alpha = s_n/n. \end{cases}$$

In words, Theorem 9 controls the average L^2–squared norm between a posterior draw and θ_0, in expectation under E_{θ_0}, for a sparse θ_0. This result implies, using Markov's inequality, that if $r_n(\alpha)$ is the rate obtained in Theorem 9, we have, for any $M_n \to \infty$, as $n \to \infty$,

$$E_{\theta_0} \Pi[\|\theta - \theta_0\|_2^2 > M_n r_n(\alpha) \,|\, X] \longrightarrow 0.$$

Note that for $\alpha = 1/n$, or more generally $\alpha = n^{-b}$ with $b \geq 1$, the rate given by the first display in the statement of Theorem 9 is given by $Cs_n \log n$, that is the minimax rate, up to replacing $\log n/s_n$ by $\log n$. Note that the rate is only significantly different from $Cs_n \log(n/s_n)$ only in the case where $\log(n/s_n) = o(\log n)$ which happens in the almost dense case where s_n is $o(n)$ but n/s_n grows very slowly, e.g. logarithmically.

If one takes the 'oracle' choice $\alpha = s_n/n$, which means that one knows beforehand s_n, the obtained rate is the optimal one $Cs_n \log(n/s_n)$ up to a constant. If α goes significantly above s_n/n in that $\alpha \gg (s_n/n)\sqrt{\log n}$, the obtained rate is of larger order than $s_n \log(n/s_n)$, so is suboptimal.

A result for the integrated posterior distance is in general not equivalent to a posterior convergence result such as the ones we have seen so far. But it implies such a result up to an *arbitrary* diverging sequence result. A nice aspect of the result obtained in Theorem 9 is that it implies convergence for the posterior mean $\bar{\theta}(X) = \int \theta d\Pi_\alpha(\theta \,|\, X)$. Indeed, using Jensen's inequality for the convex map $u \longrightarrow \|u\|^2$,

$$E_{\theta_0} \|\bar{\theta}(X) - \theta_0\|^2 \leq E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_\alpha(X).$$

The proof of Theorem 9 is presented in Section 3.5 and is based on a direct analysis of the posterior distribution via the explicit expressions (3.12)–(3.13). It is also possible to analyse the empirical Bayes $\Pi_{\hat{\alpha}}[\cdot \,|\, X]$ posterior using this approach. To do so, one needs to carry out a detailed analysis of $\hat{\alpha}(X)$, which is done in the paper by Johnstone and Silverman (2004).

The above discussion shows that the behaviour of the posterior $\Pi_\alpha[\cdot \,|\, X]$ is quite sensitive to the choice of α. Even though the default choice $\alpha = 1/n$ gives an almost optimal rate, it is desirable both for theory and practice to develop methods that choose α in an automatic way, 'adapting' to the unknown sparsity level s_n. In fact, such methods give also much better results in practice. In Section 3.3, we develop a result based on a method of proof similar to that of the GGV theorem that enables one to prove such adaptive results on the basis of qualitative assumptions.

Proof of Theorem 9

The thresholds $\zeta(\alpha)$ and $\tau(\alpha)$. From Lemma 9 below, we know that g/ϕ, and therefore $\beta = g/\phi - 1$, is a strictly increasing function on $\mathbb{R}^+$. It is also continuous, so given α, a pseudo-threshold $\zeta = \zeta(\alpha)$

can be defined by

$$\beta(\zeta) = \frac{1}{\alpha}. \tag{3.19}$$

Further one can also define $\tau(\alpha)$ as the solution in x of

$$\Omega(x, \alpha) := \frac{a(x)}{1 - a(x)} = \frac{\alpha}{1-\alpha}\frac{g}{\phi}(x) = 1.$$

Equivalently, $a(\tau(\alpha)) = 1/2$. Also, $\beta(\tau(\alpha)) = \alpha^{-1} - 2$ so

$$\tau(\alpha) \leq \zeta(\alpha). \tag{3.20}$$

We prove in Lemma 7 below that $\zeta(\alpha)$ is of order $\sqrt{\log(1/\alpha)}$.

Proof of Theorem 9. Recall the notation $r_2(\alpha, \mu, x) = \int (u-\mu)^2 d\pi_\alpha(u \mid x)$. By definition

$$\begin{aligned}
&\int \|\theta - \theta_0\|^2 d\Pi_\alpha(X) \\
&= \sum_{i:\ \theta_{0,i}=0} \int (\theta_i - \theta_{0,i})^2 d\Pi_\alpha(X) + \sum_{i:\ \theta_{0,i}\neq 0} \int (\theta_i - \theta_{0,i})^2 d\Pi_\alpha(X) \\
&= \sum_{i:\ \theta_{0,i}=0} r_2(\alpha, 0, X_i) + \sum_{i:\ \theta_{0,i}\neq 0} r_2(\alpha, \theta_{0,i}, X_i).
\end{aligned}$$

Theorem 9 is proved by taking the expectation and invoking Lemma 6:

$$\begin{aligned}
&E_{\theta_0} \int \|\theta - \theta_0\|^2 d\Pi_\alpha(X) \\
&\qquad \lesssim \sum_{i:\ \theta_{0,i}=0} \tau(\alpha)\alpha + \sum_{i:\ \theta_{0,i}\neq 0} (1 + \tau(\alpha)^2).
\end{aligned}$$

The proof is complete by noting that the first sum has at most n terms and the second at most s_n since $\theta_0 \in \ell_0[s_n]$, and using the bounds $\tau(\alpha) \leq \zeta(\alpha) \leq \sqrt{\log(1/\alpha)}$ from Lemma 7.

Lemma 6. Let γ be the Cauchy or Laplace density. For any x and $\alpha \in [0, 1/2]$,

$$\begin{aligned}
r_2(\alpha, 0, x) &\leq C\Big[1 \wedge \frac{\alpha}{1-\alpha}\frac{g}{\phi}(x)\Big](1 + x^2) \\
r_2(\alpha, \mu, x) &\leq (1 - a(x))\mu^2 + Ca(x)((x-\mu)^2 + 1).
\end{aligned}$$

Let γ be the Cauchy density. For any real x and small enough α,

$$\begin{aligned}
E_0 r_2(\alpha, 0, x) &\leq C\tau(\alpha)\alpha \\
E_\mu r_2(\alpha, \mu, x) &\leq C(1 + \tau(\alpha)^2).
\end{aligned}$$

Proof of Lemma 6.

First one proves the first two bounds. To do so, we derive moment bounds on γ_x. Since $\gamma_x(\cdot)$ is a density function, we have for any x, $\int \gamma_x(u)du = 1$. This implies $(\log g)'(x) = \int (u-x)\gamma_x(u)du = \int u\gamma_x(u)du - x$. It can be checked, see Johnstone and Silverman (2004) p. 1623, that $\int u\gamma_x(u)du =: \tilde{m}_1(x)$ is a shrinkage rule, that is $0 \le \tilde{m}_1(x) \le x$ for $x \ge 0$, so by symmetry, for any real x,

$$|\int u\gamma_x(u)du| \le |x|.$$

Writing $u^2 = (u-x)^2 + 2x(u-x) + x^2$ and noting that $\int (u-x)^2\gamma_x(u)du = g''(x)/g(x) + 1$,

$$\int u^2\gamma_x(u)du = \frac{g''}{g}(x) + 1 + 2x\frac{g'}{g}(x) + x^2.$$

Note that for γ Laplace or Cauchy, we have $|\gamma'| \le c_1\gamma$ and $|\gamma''| \le c_2\gamma$. This leads to

$$|g'(x)| = |\int \gamma'(x-u)\phi(u)du| \le c_1 \int \gamma(x-u)\phi(u)du = c_1 g(x)$$

and similarly $|g''| \le c_2 g$, so that $\int u^2\gamma_x(u)du \le C(1+x^2)$ which gives the first bound using (3.18). Also, for any real μ,

$$\int (u-\mu)^2\gamma_x(u)du = (x-\mu)^2 + \frac{g''}{g}(x) + 1 + 2(x-\mu)\frac{g'}{g}(x).$$

Now using again $g'/g \le c_1$ and $g''/g \le c_2$ leads to

$$\int (u-\mu)^2\gamma_x(u)du \le C(1 + (x-\mu)^2).$$

By using the expression of $r_2(\alpha,\mu,x)$, this yields the second bound of the lemma.

We now turn to the bounds in expectation. For a zero signal $\mu = 0$, one notes that $x = \tau(\alpha)$ is the value at which both terms in the minimum in the first inequality of the lemma are equal. So

$$E_0 r_2(\alpha,0,x) \lesssim \int 1\!\!1_{|x|\le\tau(\alpha)} \frac{\alpha}{1-\alpha}\frac{g}{\phi}(x)\phi(x)(1+x^2)dx + \int 1\!\!1_{|x|>\tau(\alpha)}(1+x^2)\phi(x)dx.$$

By Lemma 9, g has same tails as γ, so for both γ Laplace or Cauchy, g is such that $x \longrightarrow (1+x^2)g(x)$ is bounded, so one gets, with $\alpha \le 1/2$,

$$\begin{aligned} E_0 r_2(\alpha,0,x) &\lesssim \alpha \int 1\!\!1_{|x|\le\tau(\alpha)}dx + \tau(\alpha)\phi(\tau(\alpha)) + \phi(\tau(\alpha))/\tau(\alpha) \\ &\lesssim \tau(\alpha)\alpha + \tau(\alpha)\phi(\tau(\alpha)) \lesssim \tau(\alpha)\alpha + \tau(\alpha)\alpha g(\tau(\alpha)) \lesssim \tau(\alpha)\alpha. \end{aligned}$$

Turning to the last bound of the lemma, we distinguish two cases. Set for the remaining of the proof $T := \tau(\alpha)$ for simplicity of notation. The first case is $|\mu| \le 4T$, for which

$$E_\mu r_2(\alpha,\mu,x) \le \mu^2 + C \le C_1(1+T^2).$$

The second case is $|\mu| > 4T$. We bound the expectation of each term in the second bound of the lemma (that for $r_2(\alpha, \mu, x)$) separately. First, $E[a(x)(1 + (x - \mu)^2)] \leq C$. It thus suffices to bound $\mu^2 E_\mu[1 - a(x)]$. To do so, one uses the bound (3.21) and starts by noting that, if $Z \sim \mathcal{N}(0, 1)$,

$$E[1\!\!1_{|Z+\mu|\leq T}] \leq P[|Z| \geq |\mu| - T] \leq P[|Z| \geq |\mu|/2].$$

This implies, with $\bar{\Phi}(u) = \int_u^\infty \phi(t)dt \leq \phi(u)/u$ for $u > 0$,

$$E_\mu[\mu^2 1\!\!1_{|x|\leq T}] \leq C_2|\mu|\phi(|\mu|) \leq C_3.$$

If $A = \{x,\ |x - \mu| \leq |\mu|/2\}$ and A^c denotes its complement,

$$\sqrt{2\pi} E_\mu[e^{-\frac{1}{2}(|x|-T)^2}] \leq \int_{A^c} e^{-\frac{1}{2}(x-\mu)^2} dx + \int_A e^{-\frac{1}{2}(|x|-T)^2} dx.$$

The first term in the last sum is bounded above by $2\bar{\Phi}(|\mu|/2)$. The second term, as $A \subset \{x,\ |x| \geq |\mu|/2\}$, is bounded above by $2\bar{\Phi}(|\mu|/4)$. This implies, in the case $|\mu| > 4T$, that

$$E_\mu r_2(\alpha, \mu, x) \leq C_4 + 4\mu^2\bar{\Phi}(|\mu|/4) + 5 \leq C.$$

The last bound of the lemma follows by combining the previous bounds in the two cases.

Bound on threshold $\zeta(\alpha)$.

Lemma 7. For any $\alpha \leq \alpha_0$ with α_0 small enough, and $\zeta(\alpha)$ solution of the equation $\beta(\zeta(\alpha)) = \alpha^{-1}$, we have for C large enough,

$$\zeta(\alpha) \leq C\sqrt{\log(1/\alpha)}.$$

Proof.

Since $\beta = g/\phi - 1$ is strictly increasing by Lemma 9, the equation has a unique solution, and $\zeta(\cdot)$ must be strictly decreasing (as $\alpha \longrightarrow \alpha^{-1}$ is).

Also, g and γ have same tails so $g/\phi(y) - 1 \geq ce^{y^2/4} - 1 \geq de^{y^2/4}$ for $y > 0$ for small enough $c, d > 0$, recalling that γ is either the Laplace or Cauchy density. Writing the previous inequality for $y = \zeta(\alpha)$ leads to

$$\zeta(\alpha) \leq \sqrt{4\log(1/(d\alpha))},$$

which gives the result by taking $\alpha \leq \alpha_0$ small enough.

Bound on posterior weight $1 - a(x)$ in terms of $\tau(\alpha)$.

Lemma 8. Let $\tau(\alpha)$ be the unique solution of the equation $\alpha g(x) = (1 - \alpha)\phi(x)$. Then for $a(x)$ as in (3.13),

$$1 - a(x) \leq 1 1\!\!1_{|x|\leq\tau(\alpha)} + e^{-\frac{1}{2}(|x|-\tau(\alpha))^2} 1\!\!1_{|x|>\tau(\alpha)}. \tag{3.21}$$

Proof.

See Johnstone and Silverman (2009) p. 1623.

Miscellaneous: properties of g and g/ϕ.

Lemma 9. For Γ the standard Laplace or Cauchy density, let us recall the notation $g = \phi * \Gamma$, the convolution of ϕ and Γ. We have the following properties.

1. The function $x \longrightarrow g(x)$ is decreasing on $[0, \infty)$.
2. The function $x \longrightarrow g(x)$ has same tails as $x \longrightarrow \gamma(x)$.
3. The map $x \longrightarrow \frac{g}{\phi}(x)$ is strictly increasing on $[0, +\infty)$.

Proof.

These properties are proved in Johnstone and Silverman (2004). Points 2. and 3. are a part of Lemma 1 there; Point 1. follows from the bounds around Eq. (55) there.

CHAPTER 4

Deep Bayes

In this chapter we consider two directions for Bayesian deep methods. First, we consider random deep neural networks as prior distributions for estimating smooth functions. We recall terminology and basic concepts for such networks and define a prior distribution on parcimonious deep networks. We then show that corresponding posterior distributions converge at near–optimal rate around Hölder functions of arbitrary regularities.

Second, we introduce deep Gaussian process priors and sketch a few ideas of why they are able to adapt to compositional structures.

1 Introduction to neural networks

Let us introduce some vocabulary. A neural network is a structure with an input layer, a number of hidden layers ('couches' in french) and an output layer. Each hidden layer has a number of units called *neurons*. An input is taken by the network in the form of a vector $x = (x_1, \dots, x_d)'$ in $\mathbb{R}^d$ for some $d \geq 1$. The input layer just contains d units: each one passes the input coordinate x_i unchanged to the neurons of the first hidden layer. A network with one hidden layer is depicted in Figure 4.1. Then x is modified at the level of the first hidden layer in a way described below. Each layer is linked to the next one by arrows between neurons (or between units of input or output layer and a neuron).

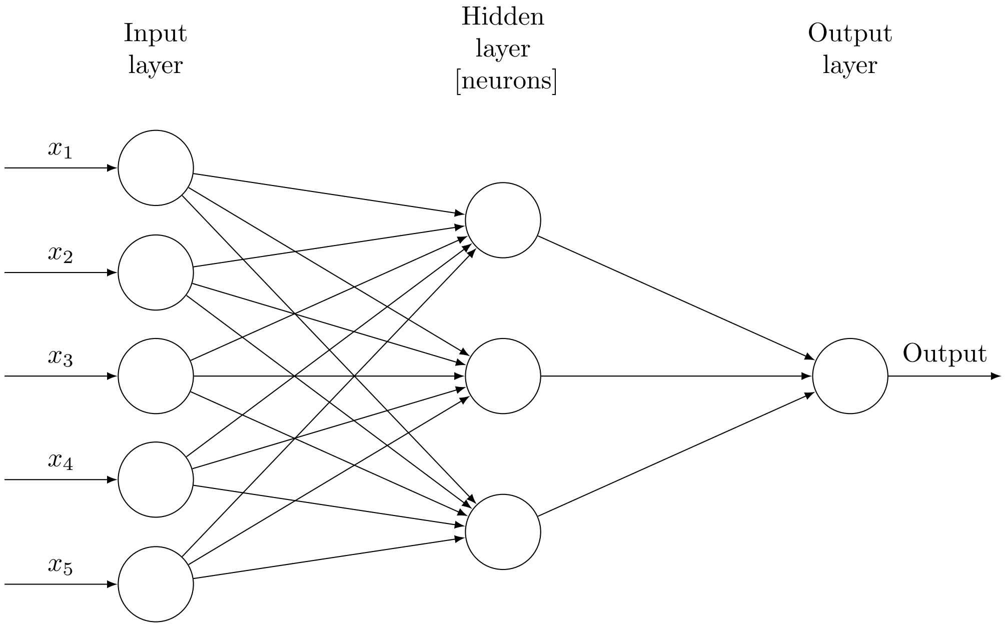


Figure 4.1: Structure of neural network with one hidden layer

Activation function. The action of a given neuron will be specified using an activation function, a function $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$. In the sequel, we consider the ReLU activation (ReLU stands for Rectified Linear Unit) given by

$$\sigma(x) = x \vee 0 = x_+. \tag{4.1}$$

To encode the action of all neurons in a given layer input dimension $r \geq 1$, we define the multidimensional shifted activation function as, given $v = (v_1, \dots, v_r)'$ a vector of shifts,

$$\sigma_v y = \sigma_v(y) := \begin{bmatrix} \sigma(y_1 - v_1) \\ \sigma(y_2 - v_2) \\ \vdots \\ \sigma(y_r - v_r) \end{bmatrix} \tag{4.2}$$

Action of one layer. Each arrow is given a *weight* that multiplies the input of the arrow. Let us consider a given layer with q neurons and a vector of biases $v = (v_1, \dots, v_q)'$. At the level of each neuron of the layer the incoming numbers from each arrow are summed, a bias relative to the neuron is applied and the result is finally passed through the activation function σ. If all the weights are aggregated in a matrix, say W, with dimension $p \times q$, where p is the input dimension and q the number of neurons of the considered layer, this means that the vector output of the considered layer is $\sigma_v Wx$. The action of the first neuron of the first layer is illustrated in Figure 4.2.

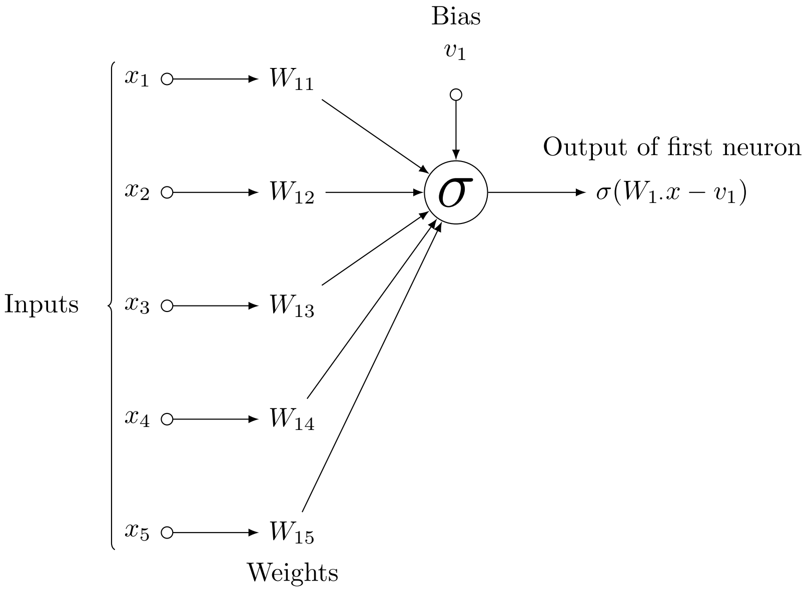


Figure 4.2: Operations at the level of the first neuron and output

Depth of the neural network. The previous operations are repeated over successive layers, each new layer taking as input the output of the previous one (this of course requires that dimensions match). The total number of layers is denoted by L and called network depth. A network with two layers is depicted in Figure 4.3: the layers that are inbetween input and output layers are called ‘hidden’ layers.

Width vector. The number of neurons of a layer is called width. The width vector $\mathbf{p} = (p_0, p_1, \dots, p_L, p_{L+1})$ with $p_0 = d$ (input dimension) and $p_{L+1} = 1$ collects all widths.

Network architecture. The pair $(L, \mathbf{p})$ defines a network architecture. The network parameters are the entries of the matrices $(W_j)_{0 \le j \le L}$ and shifts vectors $(v_j)_{1 \le j \le L}$ of the successive layers. The total number of parameters is

$$T = \sum_{l=0}^{L} p_l p_{l+1} + \sum_{l=1}^{L} p_l. \tag{4.3}$$

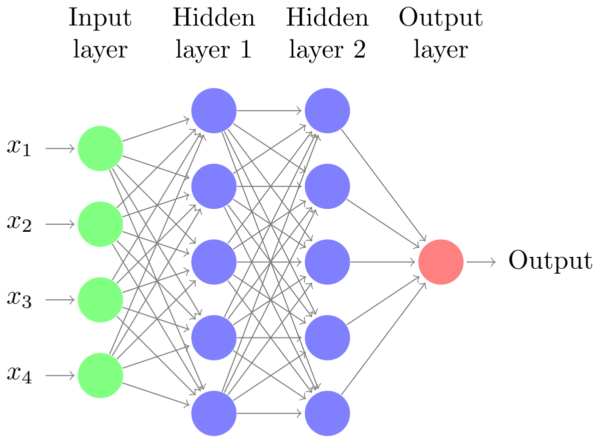


Figure 4.3: Neural network with 2 hidden layers

For a matrix $A = (a_{ij})$, let $\|A\|_\infty$ denote the (entrywise) maximum of the $|a_{ij}|$s, and similarly for $\|v\|_\infty$ for a vector v.

Definition 1. [Global network] Denoting by $(W_j)_{0\le j\le L}$ and $(v_j)_{1\le j\le L}$ the successive weight matrices and bias vectors, the global network operates as a map $x \longrightarrow f(x)$ from $\mathbb{R}^d$ to $\mathbb{R}$, where

$$f(x) = W_L \sigma_{v_L} W_{L-1} \sigma_{v_{L-1}} \cdots \sigma_{v_1} W_0 x, \qquad (\square)$$

where σ_v is defined in (4.2) and the activation σ is the ReLU function (4.1).

For a vector or matrix B, let $\|B\|_0$ denote the number of nonzero coefficients of B. In the sequel, DNN stands for ‘Deep Neural Network’, where *deep* means that the depth is not ‘small’ (e.g. 1 or 2) but rather a number possibly allowed to go to infinity (slowly, perhaps) with the number of observations. A network with just a few layers (e.g. 1 or 2) is sometimes referred to as *shallow* network.

Definition 2. [DNN classes] Given a network architecture $(L, \mathbf{p})$, we denote, setting $v_0 = 0$,

$$\mathcal{F}(L, \mathbf{p}) = \left\{ f \ \text{ as in } (\square), \ \text{ for some } (W_j)_{0\le j\le L}, (v_j)_{1\le j\le L}, \quad \max_{0\le j\le L} \left(\|W_j\|_\infty \vee \|v_j\|_\infty \right) \le 1 \right\},$$

and we also set, for $s > 0$ a sparsity parameter,

$$\mathcal{F}(L, \mathbf{p}, s) = \left\{ f \in \mathcal{F}(L, \mathbf{p}), \quad \sum_{j=0}^{L} (\|W_j\|_0 + \|v_j\|_0) \le s \right\}.$$

We assume here that the network parameters are bounded in absolute value by 1. Another positive constant could be used. The rationale behind this choice is that in practice most of the time network parameters are initialized using bounded parameters. It is also typically observed that even after training the parameters of the network remain quite close in range to initial parameters. From the theoretical point of view, some approximation results are quite easily reachable if network parameters are allowed to be very large. But in order to be closer to practical applications where parameters typically remain bounded, we restrict ourselves to that case, and we will see below that this does not prevent us to obtain good inference properties.

The regularity class of functions we consider in these notes is the classical class of Hölder functions in dimension d over, say, the unit interval.

$$\mathcal{C}_d^{\beta}([0,1]^d, K) = \left\{ f \, : \, [0,1]^d \longrightarrow \mathbb{R} \, : \sum_{\boldsymbol{\alpha}: \, \sum_{i=1}^{r} \alpha_i < \beta} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + \sum_{\boldsymbol{\alpha}: \, \|\boldsymbol{\alpha}\|_1 = \lfloor \beta \rfloor} \sup_{x,y \in D, \, x \neq y} \frac{|\partial^{\boldsymbol{\alpha}} f(x) - \partial^{\boldsymbol{\alpha}} f(y)|}{\|x - y\|_\infty^{\beta - \lfloor \beta \rfloor}} \leq K \right\}.$$

Classes of composite functions could be considered as well, we comment on that again below.

2 Prior distributions on DNNs and posterior concentration

2.1 Prior distribution

Let us define an n−dependent prior distribution Π as follows; the choices of parameters are motivated by the theoretical properties described in the next section. Set

$$L = \log^2 n, \qquad p_1 = p_2 = \cdots = p_L = n. \tag{4.4}$$

The total number of parameters T in a network of $\mathcal{F}(L, \mathbf{p})$ is then, using (4.3), of order Ln^2 (taking n large enough so that $d \leq n$). Let $\mathcal{T}$ denote the set of all parameters, that is, elements of vectors $(v_j)_{1 \leq j \leq L}$ and matrices $(W_j)_{0 \leq j \leq L}$, written in some given order; the order does not matter. We have $\mathrm{Card}(\mathcal{T}) = T$ and we denote $\mathcal{T} = \{\theta_t, \ 1 \leq t \leq T\}$.

Definition 3. A sparse prior on DNNs

1. Draw a sparsity parameter s in $\{0, 1, \dots, T\}$ according to a distribution π_s defined by

$$\pi_s(k) \propto e^{-k \log k}.$$

2. Given s, draw a subset $S \subset \{0, 1, \dots, T\}$ of cardinality s uniformly at random among all possible such subsets.

3. Given S, set, for uniform variables independent across coefficients,

$$\theta_t \sim \begin{cases} \delta_0 & \text{if } t \notin S \\ \mathrm{Unif}[-1, 1] & \text{if } t \in S \end{cases}$$

By definition, the prior Π puts mass on the set, for T the total number of possibly active parameters

$$\Pi\left[\mathcal{F}(L, \mathbf{p}, T)\right] = 1.$$

2.2 Posterior contraction

To make for the simplest statement, we consider the Gaussian white noise model setting, with a prior on f defined as above. Analogous statements can easily be written in density estimation or nonparametric fixed design regression, similar to what we did in Chapter 2.

Theorem 1. [Posterior convergence rate for Hölder functions] Let $f_0 \in \mathcal{C}_d^\beta([0,1]^d, K)$ a function of regularity $\beta > 0$. Define the prior Π on the function f from the white noise model as in Definition 3. Then there exists $M > 0$ large enough such that

$$E_{f_0}\Pi[\|f - f_0\|_2 > M\varepsilon_n \,|\, X] = o(1),$$

as $n \longrightarrow \infty$, where $\varepsilon_n = (\log n)^3 n^{-2\beta/(2\beta+d)}$.

Theorem 1 shows that a DNN with logarithmic depth and polynomial width (both in terms of n) are good models for approximating β–smooth functions, for any $\beta > 0$. The output (□) of a network with ReLU activation function is necessarily a piecewise affine function: it is interesting to note that while such a network is not highly smooth (the 'highest' regularity one can hope for is Lipschitz, but not C^2 for instance, except in the trivial case where the output is a linear function), it still enables one to recover optimal minimax rate, up to log factors, for any arbitrarily high smoothness level $\beta > 0$.

It can be shown that DNNs also adapt nearly optimally to other regularity structures, for instance in case the true f_0 is a composite function $f_0 = g_k \circ g_{k-1} \circ \cdots \circ g_1$, DNNs also yield near–optimal rates: we refer to the paper Schmidt–Hieber (2020), from which some of the key lemmas below are borrowed, for more details on this.

2.3 Proof of posterior concentration

We check the three conditions of the GGV theorem successively, with here d equal to the $\|\cdot\|_2$–distance on $[0, 1]$, for which appropriate tests exists in the white noise model.

Sieve. For $\bar{s}$ an integer to be chosen below, let $\mathcal{F}_n := \mathcal{F}(L, \mathbf{p}, \bar{s})$. By definition of the prior,

$$\Pi[\mathcal{F}_n^c] \leq \pi_s(s > \bar{s}) \lesssim e^{-\bar{s}\log\bar{s}}.$$

The sieve condition of the GGV Theorem is then fulfilled if $\bar{s}\log\bar{s} \gtrsim n\varepsilon_n^2$.

Entropy. Using the entropy Lemma 2, with $\|\cdot\|_2 \leq \|\cdot\|_\infty$, for any $\delta > 0$,

$$\log N(\delta, \mathcal{F}_n, \|\cdot\|_2) \leq \log N(\delta, \mathcal{F}(L, \mathbf{p}, \bar{s}), \|\cdot\|_\infty) \leq (\bar{s}+1)\log\left(\frac{2(L+1)V^2}{\delta}\right),$$

where $V \leq (n+1)^{L+2} \leq e^{C\log^3(n)}$ for a large enough constant C, so that for any vanishing sequence $(\bar{\varepsilon}_n)$ with $\bar{\varepsilon}_n \geq 1/\sqrt{n}$,

$$\log N(\bar{\varepsilon}_n, \mathcal{F}_n, \|\cdot\|_\infty) \lesssim \bar{s}(\log n)^3.$$

For the entropy condition of the GGV Theorem to be fulfilled, we need $\bar{s}(\log n)^3 \lesssim n\bar{\varepsilon}_n^2$.

Prior mass condition. For $\tilde{f}_0$ a suitable DNN–approximation of f_0 (to be defined below),

$$\Pi[\|f - f_0\|_2 \leq \underline{\varepsilon_n}] \geq \Pi[\|f_0 - \tilde{f}_0\|_2 + \|f - \tilde{f}_0\|_2 \leq \underline{\varepsilon_n}] \geq \Pi[\|f_0 - \tilde{f}_0\|_\infty + \|f - \tilde{f}_0\|_\infty \leq \underline{\varepsilon_n}].$$

Let us apply Theorem 2 to f_0 and the choices, for A_0 large enough constant (depending on d, K, β),

$$N = n^{\frac{d}{2\beta+d}}, \qquad m = A_0 \log n.$$

There exists $\tilde{f}_0$ in $\mathcal{F}(L', (d, \Lambda, \ldots, \Lambda, 1), s_0)$ that approximates f_0 as in Theorem 2, with depth $L' \lesssim m \lesssim \log n$ and sparsity $s_0' \lesssim (\log n)N$. Up to adding at the end of the network a subnetwork that equals the identity over $L - L' \approx \log^2 n$ layers (note $x = (x \vee 0) - (-x) \vee 0$ which uses one layer, two units and 6 parameters, out of which 4 are non–zero; this is called 'synchronisation'), one can suppose that $\tilde{f}_0 \in \mathcal{F}(L, (d, \Lambda, \ldots, \Lambda, 1), s_0') \subset \mathcal{F}(L, (d, n, \ldots, n, 1), s_0)$ (this is called 'enlarging'), where $s_0 \lesssim s_0'$. More precisely, we append to the subnetwork above encoding $\tilde{f}_0$ and with 1-dimensional output a network of length of order $L - L'$ and width 2, and then artificially add neurons with connexions that have all zero coefficients to form the desired architecture with width n and length L. We add in total of the order $\log^2 n$ non–zero parameters, so the overall sparsity s_0 verifies $s_0 \lesssim s_0'$.

Let $\theta_0 = (\theta_0)_{t\in\mathcal{T}}$ denote the collection of parameters of the network encoding $\tilde{f}_0$. By definition, only s_0 are non–zero. Let S_0 denote the subset of the index set $\mathcal{T}$ corresponding to these non–zero parameters. Similarly, for a draw f from the prior let $S(f)$ denote the set of indices of its non–zero parameters.

By Theorem 2, we know that, for large enough n (depending on β, K, d),

$$\|f_0 - \tilde{f}_0\|_\infty \lesssim \frac{N}{2^m} + N^{-\frac{\beta}{d}} \lesssim n^{-\Delta} + n^{-\frac{\beta}{2\beta+d}},$$

where Δ can be made arbitrarily large for large A_0, so the last display is smaller than $\underline{\varepsilon_n}/2$ as long as $n^{-\frac{\beta}{2\beta+d}} \lesssim \underline{\varepsilon_n}$. If this holds we then have

$$\Pi[\|f - f_0\|_2 \leq \underline{\varepsilon_n}] \geq \Pi[\|f - \tilde{f}_0\|_\infty \leq \underline{\varepsilon_n}/2] \geq \Pi[\|f - \tilde{f}_0\|_\infty \leq \underline{\varepsilon_n}/2 \,|\, S(f) = S_0]\Pi[S(f) = S_0]$$

On the event that $S(f) = S_0$, the corresponding networks encoding f and $\tilde{f}_0$ have same index sets for their non–zero coefficients, so we may now use the 'error propagation' Lemma 1 to obtain

$$\begin{aligned}\Pi[\|f - \tilde{f}_0\|_\infty \leq \underline{\varepsilon_n}/2 \,|\, S(f) = S_0] &\geq \Pi\left[\forall\, t \in S_0,\ \ |\theta_t(f) - \theta_t(\tilde{f}_0)| \leq \frac{\underline{\varepsilon_n}}{2V(L+1)}\right] \\ &\geq \prod_{t\in S_0} \frac{\underline{\varepsilon_n}}{V(L+1)} = \left(\frac{\underline{\varepsilon_n}}{2V(L+1)}\right)^{s_0} \geq e^{-Cs_0\log^3 n},\end{aligned}$$

for a large enough $C > 0$, using that $\log V \lesssim \log^3 n$. On the other hand, again for large $C > 0$,

$$\Pi[S(f) = S_0] = \frac{1}{\binom{T}{s_0}} e^{-s_0\log(s_0)} \geq e^{-s_0\log(Te/s_0) - s_0\log s_0} \geq e^{-Cs_0\log n}.$$

Deduce that $\Pi[\|f - f_0\|_2 \leq \underline{\varepsilon_n}] \geq e^{-n\underline{\varepsilon}_n^2}$ provided one chooses

$$n\underline{\varepsilon}_n^2 \gtrsim s_0 \log^3 n.$$

Conclusion. Let us choose, for suitably large $A_2 > 0$,

$$\underline{\varepsilon_n}^2 = A_2 \max(n^{-\frac{2\beta}{2\beta+d}}, s_0 \log^3 n/n) = A_2(\log n)^4 N/n = A_2(\log n)^4 n^{-\frac{\beta}{2\beta+d}}.$$

Then the conditions on ε_n for the prior mass are satisfied and the condition for the sieve is $n\underline{\varepsilon}_n^2 \lesssim \bar{s}\log \bar{s}$ which holds for the choice $\bar{s} = A_3 s_0 \log^2 n$ for large enough $A_3 > 0$. Finally, from the sieve condition one gets the condition

$$\bar{\varepsilon}_n^2 \gtrsim \frac{\bar{s}\log^3 n}{n} \gtrsim (\log n)^6 n^{-\frac{2\beta}{2\beta+d}}.$$

The proof of the Theorem is complete since $\varepsilon_n = \bar{\varepsilon}_n \vee \underline{\varepsilon}_n = \bar{\varepsilon}_n$ as desired.

3 Complement: generic properties of DNNs

3.1 Error propagation in a neural network and entropy

Considering the class of functions $\mathcal{F}(L, \mathbf{p}, s)$, let us denote

$$V := \prod_{l=0}^{L+1}(p_l + 1). \tag{4.5}$$

Lemma 1. Let f, f^* be two functions in $\mathcal{F}(L, \mathbf{p}, s)$ with matrix parameters W_k, W_k^* and shift vectors v_k, v_k^* for $k = 0, 1, \ldots, L+1$. Suppose that every individual parameter of f (i.e. elements of matrices W_k or bias vectors v_k) is at most $\varepsilon > 0$ away from the corresponding parameter of f^*. Then for V as in (4.5),

$$\|f - f^*\|_\infty \leq \varepsilon V(L+1).$$

Lemma 2. For V as in (4.5) and any $\delta > 0$,

$$\log N(\delta, \mathcal{F}(L, \mathbf{p}, s), \|\cdot\|_\infty) \leq (s+1)\log\left(\frac{2(L+1)V^2}{\delta}\right).$$

In particular if $L \lesssim \log n$ and $p_l \leq n$ for all l, we have $\log N(\delta, \mathcal{F}(L, \mathbf{p}, s), \|\cdot\|_\infty) \lesssim s\log^2(n)\log(1/\delta)$.

3.2 Approximation properties for Hölder functions

Theorem 2. [Approximation of smooth functions by DNNs] Let $f \in \mathcal{C}_d^\beta([0,1]^d, K)$ a function of regularity $\beta > 0$. Let $m, N \geq 1$ be two integers. There exists a network, with $\Lambda := 6(d + \lceil \beta \rceil)N$,

$$\tilde{f} \in \mathcal{F}(L, (d, \Lambda, \dots, \Lambda, 1), s)$$

with depth and sparsity verifying, for $C_0 = 1 + \log_2\lceil (d \wedge \beta) \rceil$, $c_0 = 141(d + \beta + 1)^{3+d}$,

$$L = 8 + C_0(m+5), \qquad s \leq c(m+6)N,$$

such that, for $c_1 = (2K+1)(1 + d^2 + \beta^2)6^d$ and $c_2 = K3^\beta$, and $N \geq (\beta+1)^d \vee (K+1)e^d$,

$$\|\tilde{f} - f\|_\infty \leq c_1 \frac{N}{2^m} + c_2 N^{-\frac{\beta}{d}}.$$

[Sketch of proof] The general idea of the proof is as follows: there are two main steps. The first is not specific to DNNs and is that any β–Hölder function can be well–approximated locally, using Taylor expansions, by a polynomial of order $\lfloor \beta \rfloor$: one can approximate f_0 by a piecewise polynomial function, with a quality of approximation that depends on β. The second idea, where the choice of activation function σ comes in, is that it is possible to approximate quickly, in one dimension, the monomial $x \longrightarrow x^2$ using a ReLU network. From there one then shows that ReLU networks suitably approximate $x \longrightarrow x^p$ for $p \geq 2$; one can also check that the argument extends to dimensions $d \geq 2$ for approximating general monomials. From monomials one can easily approximate polynomials by combining networks, and now one can connect to the first part of the argument, by constructing a network that approximates the piecewise polynomial function mentioned above, that itself approximates f_0.

Lemma 3. [Approximating $x(1-x)$ with piecewise affine functions]
Let $T^1 : [0,1] \longrightarrow [0, 1/4]$ and more generally $T^k : [0, 2^{-2(k-1)}] \longrightarrow [0, 2^{-2k}]$, $k \geq 1$, be the maps

$$T^1(x) = \frac{x}{2} \wedge \left(\frac{1}{2} - \frac{x}{2}\right), \qquad T^k(x) = \frac{x}{2} \wedge \left(\frac{1}{2^{2k-1}} - \frac{x}{2}\right).$$

Let us set $R^k := T^k \circ T^{k-1} \circ \cdots \circ T^1$, for $k \geq 1$. Then for any $m \geq 1$,

$$\left| x(1-x) - \sum_{k=1}^{m} R^k(x) \right| \leq 4^{-m}.$$

Lemma 4. [Approximating $(x, y) \longrightarrow xy$ by a DNN] Let $m \geq 1$. There exist a DNN that we denote $\mathrm{Mult}_m(x, y)$ with

$$\mathrm{Mult}_m \in \mathcal{F}(m + 4, (2, 6, \cdots, 6, 2, 2, 2, 1)),$$

such that for any $x, y \in [0, 1]$ it holds $\mathrm{Mult}_m(x, y) \in [0, 1]$, $\mathrm{Mult}_m(0, y) = \mathrm{Mult}_m(x, 0) = 0$ and

$$|\mathrm{Mult}_m(x, y) - xy| \leq 4^{-m}.$$

In order to approximate a function $f \in C_d^\beta([0, 1]^d, K)$, we define a grid of $[0, 1]^d$ as

$$D(M) = \left\{ x_l = \left(\frac{l_j}{M} \right)_{j=1,\dots,d}, \quad l = (l_1, \dots, l_d) \in \{0, 1, \dots, M\}^d \right\}.$$

Around a given point $\boldsymbol{a} \in [0, 1]^d$, the function f can be approximated by its Taylor polynomial: in dimension d its expression is, for $\boldsymbol{a} = (a_1, \dots, a_d)$,

$$P_{\boldsymbol{a}}^\beta(x) := \sum_{0 \leq |\alpha| < \beta} (\partial^\alpha f)(\boldsymbol{a}) \frac{(x - \boldsymbol{a})^\alpha}{\alpha!}.$$

Taylor's expansion with Lagrange remainder gives, for any $f \in C_d^\beta([0, 1]^d, K)$,

$$|f(x) - P_{\boldsymbol{a}}^\beta(x)| \leq K \|x - \boldsymbol{a}\|_\infty^\beta. \tag{4.6}$$

Define, again for any $f \in C_d^\beta([0, 1]^d, K)$ and $x = (x_1, \dots, x_d)$,

$$P^\beta f(x) := \sum_{x_l \in D(M)} (P_{x_l}^\beta f)(x) \prod_{j=1}^d (1 - M|x_j - x_{l,j}|)_+. \tag{4.7}$$

Inside the hypercubes defined by consecutive gridpoints, $P^\beta f(x)$ is a polynomial, so the overall function $P^\beta f$ is piecewise–polynomial.

Lemma 5. [Approximation of f by a piecewise–polynomial function] For any $f \in C_d^\beta([0, 1]^d, K)$, define $P^\beta f$ as in (4.7). Then

$$\|f - P^\beta f\|_\infty \leq K M^{-\beta}.$$

Proof. One notes that the terms of the sum in the definition (4.7) are nonzero only at a given x for x_l such that $\|x - x_l\|_\infty \leq 1/M$, otherwise the product in (4.7) is zero. Combine this with the fact that

$$\sum_{x_l = (l_1/M, \dots, l_d/M)} \prod_{j=1}^d (1 - M|x_j - x_{l,j}|)_+ = \prod_{j=1}^d \sum_{l=0}^M (1 - M|x_j - l/M|)_+ = 11$$

(these functions form a 'partition of unity') and Taylor's approximation (4.6) to obtain the result.

4 Deep Gaussian process priors

4.1 Motivation: compositional structures

Here we will state results in the so-called random design regression model (but we could state analog results in Gaussian white noise as in the previous section).
Consider observing i.i.d. pairs $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n)$ with

$$Y_i = f_0(X_i) + \varepsilon_i, \qquad 1 \le i \le n, \tag{4.8}$$

where X_i are $[0,1]^d$–valued random variables (also called *design points*) and ε_i are independent standard normal $\mathcal{N}(0,1)$ variables, and independent of the X_i's, and $f_0 : [0,1]^d \longrightarrow \mathbb{R}$ an unknown function.

Typical statistical goals in this setting are

- estimating the unknown regression function f_0 from the observations
- finding estimates that behave (near–)"optimally" with respect to some criterion (e.g. minimax) over natural classes of parameters.

Let $\hat{f}(\cdot) = \hat{f}_n(Z_1, \dots, Z_n)(\cdot)$ be an estimator of f.
The *prediction* risk in the setting of model (4.8) is defined as follows. Let T be a 'synthetic' data point, that is a variable independent of the X_i's and generated from the distribution of X_1. Let

$$R(\hat{f}, f_0) = E\left[\left(\hat{f}(T) - f_0(T)\right)^2\right] = E\left[\left(\hat{f}(Z_1, \dots, Z_n)(T) - f_0(T)\right)^2\right]. \tag{4.9}$$

Discovering a hidden 'structure'. The 'raw' regression data collected by the statistician takes the form, in the setting model (4.8), of n vectors of size $d+1$: the n pairs (X_i^T, Y_i) with $X_i \in [0,1]^d$ and Y_i a real, with the dimension d possibly large (think for instance of e.g. $d = 10$ or 20). The unknown regression function $f_0(x_1, \dots, x_d)$ depends on d of variables, and we have seen that if d is larger than a few units this may lead to a slow uniform convergence rate of the form $n^{-2\beta/(2\beta+d)}$ for the prediction risk. It is often the case though that the problem is effectively of smaller dimension than d. We give a number of frequently encountered examples

1. f_0 in fact depends on just one variable (but we do not know it a priori), for instance

   $$f_0(x_1, \dots, x_d) = g(x_1),$$

   for some $g : [0,1] \longrightarrow \mathbb{R}$. In this case it seems reasonable to expect a rate $n^{-2\beta/(2\beta+1)}$, since the f_0 effectively depends on 1 variable only. More generally, f_0 may depend on a small number $t \le d$ of variables, although we do not know a priori which ones, e.g.

   $$f_0(x_1, \dots, x_d) = g(x_2, x_3, x_d),$$

   in which case the effective dimension should be 3, so we expect a rate $n^{-2\beta/(2\beta+3)}$.

2. In the preceding example, the function effectively depends on a small number of the original variables x_i, but it could depend on few variables only after transformation of the variables, for instance

   $$f_0(x_1, \dots, x_d) = g(x_1 + x_2 + \cdots + x_d).$$

   In this case $f_0(x_1, \dots, x_d) = g(x')$ only depends on 'one' variable $x' = x_1 + \cdots + x_d$, so one expect a rate $n^{-2\beta/(2\beta+1)}$.

3. *Additive models.* It may be possible to write f_0 in an additive form
$$f_0(x_1, \dots, x_d) = \sum_{i=1}^{d} f_i(x_i),$$
for some functions $f_1, \dots, f_d$ depending on one variable only. If all functions f_i are at least β–Hölder, one expects a rate $d \cdot n^{-2\beta/(2\beta+1)}$ that is $n^{-2\beta/(2\beta+1)}$ if d is a fixed constant.

4. *Generalised additive models.* It may be possible to write f_0 in the form
$$f_0(x_1, \dots, x_d) = h\left(\sum_{i=1}^{d} f_i(x_i)\right),$$
for some real-valued functions $f_1, \dots, f_d$ (that are, as before, say all β–Hölder) and an unknown real 'link' function h that is γ–Hölder. One expects the rate to depend on β, γ, but not (too much) on the dimension d.

Class of compositions. In all the settings of the previous paragraph, one may note that the original function f_0 can be written as a composition of functions
$$f_0 = g_q \circ \cdots \circ g_1 \circ g_0,$$
for some integer $q \geq 1$. For instance, in the case of additive models one can set $g_0(x_1, \dots, x_d) = (f_1(x_1), \dots, f_d(x_d))$ (note that g_0 is then $\mathbb{R}^d$–valued) and $g_1(y_1, \dots, y_d) = y_1 + \cdots + y_d$. For each of the examples in the above list, *if one knew beforehand* that f_0 is in one class of the other, one could certainly develop a specific estimation method using the special structure at hand. In practice, however, it would be desirable to have a method that is able to automatically 'learn the structure'. We are going to see that this is achieved by deep ReLU estimators.

Let us introduce the class, for $d = (d_0, \dots, d_{q+1}), t = (t_0, \dots, t_q), \beta = (\beta_0, \dots, \beta_q)$,
$$\mathcal{G}(q, d, t, \beta, K) = \Big\{ f = g_q \circ \cdots \circ g_0 \ : \quad g_i = (g_{ij})_j \ : \ [a_i, b_i]^{d_i} \longrightarrow [a_{i+1}, b_{i+1}]^{d_{i+1}},$$
$$g_{ij} \in C_{t_i}^{\beta_i}([a_i, b_i]^{t_i}, K), \quad |a_i|, |b_i| \leq K \Big\}, \tag{4.10}$$
where we denoted $C_{t_i}^{\beta_i}$ for the Hölder ball over t_i variables to insist on the fact that these functions depend on t_i variables only (at most). The coefficients t_i can be interpreted as the maximal number of variables each function g_{ij} is allowed to depend on. In particular, this number is always at most d_i, but may actually be much smaller. Let us note that the decomposition of f_0 as a composition is typically not unique, but this is not of concern us here because we are interested in estimation of f_0 itself only.

Compositional classes are quite rich and contain many interesting functions having a low dimensional "effective dimensionality". There are quite popular for the analysis of deep learning algorithms. In particular, a recent work by Johannes Schmidt–Hieber (2020) shows that deep ReLU neural networks can get near optimal rates over such classes (one still assumes that some parameters of the classes are known). We will see below that deep Gaussian processes possess analog properties (and are even fully adaptive to smoothness and structure). Let us first give an example and state what the optimal minimax rate over these classes is.

Example. In ambient dimension 5, consider the function

$$f(x_1, x_2, x_3, x_4, x_5) = h_1(h_{01}(x_1, x_3, x_4), h_{02}(x_1, x_4, x_5), h_{03}(x_5)).$$

Then h_0 takes as input 5 coordinates (so $d_0 = 5$) and takes its values in $\mathbb{R}^3$ (hence $d_1 = 3$) so has three coordinate functions $h_{01}, h_{0,2}, h_{0,3}$, which themselves depend on only (at most) 3 variables, so that here $d_0 = 3$. Since h_1 has three coordinates and (in general) depends on each of these, we have $d_1 = t_1 = 3$. Finally, the final output of the regression is always a real number in this chapter so $d_2 = 1$.

Note that for $f_0 = g_1 \circ g_0$ with $d_1 = d_0 = t_1 = t_0 = 1$ and $\beta_0, \beta_1 \leq 1$, it follows from the definition of the Hölder class that f_0 has regularity $\beta_0\beta_1$, so that one expects a convergence rate of order $n^{-\frac{\beta_0\beta_1}{1+2\beta_0\beta_1}}$. It turns out that the actual (or 'effective') regularity depends on whether $\beta_i \leq 1$ or not. Let us define the following new 'regularity' parameter

$$\beta_i^* = \beta_i \prod_{\ell=i+1}^{q} (\beta_\ell \wedge 1). \tag{4.11}$$

Convergence result for compositions. Given d, t, β as before, let us define the rate

$$\varepsilon_n^* = \max_{0 \leq i \leq q} \left\{ n^{-\frac{\beta_i^*}{2\beta_i^* + t_i}} \right\}. \tag{4.12}$$

Example. For $d_0 = d_1 = t_0 = t_1 = q = 1$ and $f = g_1 \circ g_0$ with $\beta_1, \beta_0 \leq 1$, we have $\beta_0^* = \beta_0(\beta_1 \wedge 1) = \beta_0\beta_1$ and $\beta_1^* = \beta_1$, and the rate ε_n^* equals, since $\beta_0\beta_1 \leq \beta_1$,

$$\max\left(n^{-\frac{\beta_1}{2\beta_1+1}}, n^{-\frac{\beta_0\beta_1}{2\beta_0\beta_1+1}}\right) = n^{-\frac{\beta_0\beta_1}{2\beta_0\beta_1+1}},$$

which gives the rate announced above for this example. One may check that the formula (4.12) also gives the expected rate in the other examples above.

Theorem 3. [Minimax optimality for compositions] Consider the regression model (4.8), where the X_is are drawn from a distribution with density on $[0, 1]^d$ which is bounded from above and below by positive constants. For arbitrary $\beta > 0$, integer q and vector of integers d, t, suppose $t_i \leq \min(d_0, \dots, d_{i-1})$ for all i. Then for large enough K,

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{G}(q,d,t,\beta,K)} R(\hat{f}, f_0) \geq c\varepsilon_n^{*2},$$

where the infimum is taken over all possible estimators $\hat{f}$ of f in model (4.8).

4.2 Deep GPs: definition

Definition 4. [Deep Gaussian process]. A deep Gaussian process (deep GP or simply DGP) is a composition of Gaussian processes: for some integer $q \geq 2$, it is a stochastic process defined as

$$Z(t) = W_q \circ \cdots \circ W_1(t), \qquad t \in [0,1]^d,$$

where $W_i : \mathbb{R}^{d_i} \longrightarrow \mathbb{R}^{d_{i+1}}$, with (d_i) some integers and $d_1 = d$, $d_{q+1} = 1$.

Remark. Often, one restricts the range of the GPs in the composition defining a deep GP so that the successive GPs take values in a same compact subset, e.g. one sets $W_i' = (-M) \vee (W_i \wedge M)$ for some given $M > 0$ and

$$Z' = W_q' \circ \cdots \circ W_1'.$$

The idea to take a deep GP as a prior is to make the prior more flexible (by adding 'more randomness'): it seems then likely that such a prior will approximate well compositions of functions – which enable to approximate quite complex objects, as we will see below –. As such a deep GP as in Definition 5 is not yet flexible enough to do adapt well to arbitrary compositional structure and smoothness. First, it seems natural to draw the 'depth' q randomly in the prior, but also, in order not to 'overfit', to select randomly at each level which variables the process W_i depends on, in particular if one believes that there is a low dimensional compositional structure to which the true function f_0 we are trying to recover belongs. This motivates the following more general definition.

Definition 5. [Hierarchical DGP]. A hierarchical deep Gaussian process (HdGP) is defined as

$$\begin{array}{rl}
q & \sim \Pi_q \\
d_1, \dots, d_q \mid q & \sim \Pi_d[\cdot|q] \\
A_{ij} \mid q, d_1, \dots, d_q & \overset{\text{i.i.d.}}{\sim} \pi_\tau^{\otimes d_i} \\
g_{ij} \mid q, d_1, \dots, d_q, A_{ij} & \overset{\text{i.i.d.}}{\sim} W^{A_{ij}} \\
f \mid q, d_1, \dots, d_q, g_{ij} & = \Psi(g_q) \circ \cdots \circ \Psi(g_0),
\end{array}$$

where the $(g_{ij})_j$ are the coordinate functions of g_i (which takes values in $\mathbb{R}^{d_{i+1}}$) and

- Π_q and $\Pi_d[\cdot \mid q]$ are priors on integers,
- π_τ is a prior on scale parameters $A_{ij} > 0$,
- one denotes $W^{A_{ij}}(u) = W(A_{ij} u)$,
- $\Psi(x) = (-M) \vee (x \wedge M)$ for some $M > 0$.

Deep horseshoe GP. Let us consider the following prior choices: for the dimension q, one takes a prior with exponential decrease $\Pi_q(k) \propto e^{-q}$ and similarly for (d_i), one takes an exponentially decreasing prior for each d_i independently. The coordinates functions g_{ij} of the function g_i in the composition are given GP priors (given A_{ij}): they are taken to be centered GPs with squared–exponential covariance function, i.e. $E[W_x W_y] = \exp(-\|x-y\|^2)$, with $\|\cdot\|$ the euclidian norm on $\mathbb{R}^{d_i}$. It now suffices to specify the prior on the A_ijs. We take them independent with a *horseshoe distribution* with parameter $\tau > 0$ fixed (e.g. $\tau = 1$), where the horseshoe prior is defined as follows.

Definition 6. [Horseshoe prior]. The horseshoe prior with parameter $\tau > 0$ is the distribution on $\mathbb{R}$ of the variable X_τ defined as

$$\begin{aligned} \lambda &\sim C^+(0,1) \\ X_\tau \mid \lambda &\sim \mathcal{N}^+(0, \tau^2\lambda^2) \end{aligned}$$

where C^+ and $\mathcal{N}^+$ are half–Cauchy and half–Normal distributions (i.e. if Y is Cauchy then C^+ is the distribution of $Y \mid Y > 0$ and similarly for normal). It can be checked that the density π_τ of the horseshoe prior verifies

$$\frac{1}{(2\pi)^{3/2}\tau}\log\left(1+\frac{4\tau^2}{t^2}\right) < \pi_\tau(t) < \frac{1}{\sqrt{2\pi^3}\tau}\log\left(1+\frac{\tau^2}{t^2}\right).$$

In particular, the horseshoe density has a pole at zero and Cauchy tails.

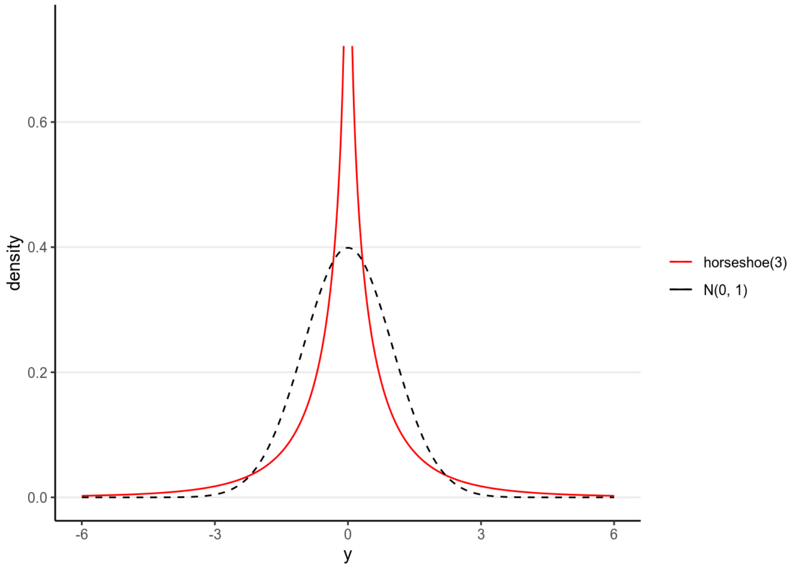


Figure 4.4: Symmetrised horseshoe prior density with parameter $\tau = 3$

The idea of the above choice of Horseshoe deep GP prior is as follows: the density π_τ puts quite a lot of mass near zero and in the tails (i.e. the probability of drawing large values is quite high due to the Cauchy tails). Very small values of A_{ij} allow to “freeze” the corresponding coordinate: it is as if the prior is almost constant on this coordinate. On the other hand, large values of A_{ij} enable one to “unsmooth” the very smooth paths of squared–exponential GPs and thus to adapt to smoothness as well (this type of adaptation to smoothness was already known in the literature since the paper van der Vaart and van Zanten (2009), as noted in Chapter 3).

4.3 Statement for deep GPs

For a given model with observations Z and likelihood $p_f(Z)$, define a fractional posterior or "α–posterior", for $0 < \alpha < 1$, by, for any measurable set B,

$$\Pi^\alpha[B \mid Z] = \frac{\int_B p_f(Z)^\alpha d\Pi(f)}{\int p_f(Z)^\alpha d\Pi(f)}.$$

For $\alpha = 1$, this is the standard Bayes formula. For $\alpha < 1$, one 'reweights' the likelihood so that the data has a bit less importance compared to the prior. The advantage of working with fractional posteriors is that, remarkably, one can show that there is α–posterior contraction under a prior mass condition only, see below.

Theorem 4. [Deep Horseshoe GP (idea)] Consider the α–posterior for $\alpha \in (0,1)$ and suppose one observes data in the random design regression model, with a deep horseshoe GP prior on f (as defined in the previous section). Then the corresponding α–posterior distribution contracts at the optimal rate (up to logarithmic factors) if the true f_0 belongs to a compositional class as defined above.

We will not give a proof here, but just give the idea: thanks to Theorem 5 below, since one works with the α–posterior, it is enough to verify the prior mass condition. This is done using relatively similar tools as for a simple GP, but it is somewhat more involved due to the successive steps involved in the compositions; again, the concentration functions of the successive GPs play a key role. Finally, one relates the α–Rényi divergence to the target quadratic distance for the regression model (which is quite easy for Gaussian noise as assumed here; see also below for a general link to the L^1–distance).

5 Complement: Tempered posteriors

Notation. Let $\mathcal{P} = \{P_\eta^{(n)},\ \eta \in S\}$ dominated model $[dP_\eta^{(n)} = p_\eta^{(n)} d\mu]$ with observations Y^n.

α–Rényi divergence between densities f and g [wrt μ]: for $\alpha \in (0,1)$,

$$D_\alpha(f,g) = -\frac{1}{1-\alpha}\log\left(\int f^\alpha g^{1-\alpha} d\mu\right)$$

KL–type neighborhood of $\eta_0 \in S$. Recall the definition, for $\varepsilon > 0$,

$$B_n(p_{\eta_0}^{(n)},\varepsilon) = B_n(\eta_0,\varepsilon) = \left\{\eta \in S\ :\ K(p_{\eta_0}^{(n)}, p_\eta^{(n)}) \le n\varepsilon^2,\ V(p_{\eta_0}^{(n)}, p_\eta^{(n)}) \le n\varepsilon^2\right\}$$

with $K(f,g) = \int f\log(f/g)d\mu,\ \ V(f,g) = \int f\,(\log(f/g) - K(f,g))^2\, d\mu$.

Theorem 5. For any non negative sequence ε_n and $0 < \alpha_n < 1$ such that $n\alpha_n\varepsilon_n^2 \longrightarrow \infty$ and

$$\Pi(B_n(\eta_0,\varepsilon_n)) \ge e^{-n\alpha_n\varepsilon_n^2}, \tag{4.13}$$

there exists $C > 0$ such that as $n \longrightarrow \infty$, for $P_0 = P^n_{\eta_0}$,

$$\Pi_{\alpha_n}\left(\eta\ :\ \frac{1}{n}D_{\alpha_n}(p^n_\eta, p^n_{\eta_0}) \geq C\frac{\alpha_n\varepsilon_n^2}{1-\alpha_n}|Y^n\right) = o_{P_0}(1).$$

Remarks. Note that in Theorem 5, we allow for α_n to possibly go to 0, but it is of course valid as a particular case for $\alpha_n = \alpha$ fixed in $(0, 1)$. In the limiting case $\alpha_n = 0$, one simply obtains the prior distribution itself (so, the data Y^n plays no role), while if one lets $\alpha_n \longrightarrow \infty$ one gets close to "maximum likelihood" (we keep this as an intuition, as of course this notion here would need to be properly defined).

As written the result is in terms of the normalised divergence $D_{\alpha_n}((p^n_\eta, p^n_{\eta_0})/n$ which still depends both on α_n and n. However, the following classical inequality enables to more easily interpret the result

$$D_{\alpha_n}((p^n_\eta, p^n_{\eta_0}) \geq n\alpha_n\|p^{(n)}_\eta - p^{(n)}_{\eta_0}\|_1^2/2.$$

Therefore Theorem 5 automatically implies convergence of the posterior in terms of the *squared–L^1* distance at rate $\varepsilon_n^2/(1-\alpha_n)$, for any ε_n that verifies the stated prior mass condition. Note the α_n inside the exponential in the prior mass condition: this makes it quite different from the typical GGV condition in the regime when α_n tends to 0. More precisely, one then typically obtains a rate similar to the one obtained from GGV, but with n replaced by $n' = n\alpha_n$ (precisely due to this extra α_n factor in the prior mass condition). For instance, nonparametric squared rates $n^{-2\beta/(2\beta+d)}$ typically become $(n\alpha_n)^{-2\beta/(2\beta+d)}$. This only changes the constant if α_n is bounded away from 0, but otherwise the rate is typically slower.

Finally, note that the obtained rate blows up as $\alpha_n \longrightarrow 1$. This is quite expected, as it is known that the entropy condition is in a sense necessary for the GGV theorem to hold: a counter-example of Barron, Schervish and Wasserman (1999) indeed shows that the true posterior ($\alpha_n = 1$) can be inconsistent under a prior mass condition only.

Proof.

By Lemma 6, on a subset C_n of P_0-probability at least $1 - \frac{1}{n\varepsilon_n^2}$, for any measurable set $A \subset S$,

$$\begin{aligned} E_0\Pi_{\alpha_n}(A|Y^n) = E_0\frac{\int_A \frac{p^n_\eta(Y^n)^{\alpha_n}}{p^n_{\eta_0}(Y^n)^{\alpha_n}}d\Pi(\eta)}{\int \frac{p^n_\eta(Y^n)^{\alpha_n}}{p^n_{\eta_0}(Y^n)^{\alpha_n}}d\Pi(\eta)} &\leq E_0\frac{\int_A \frac{p^n_\eta(Y^n)^{\alpha_n}}{p^n_{\eta_0}(Y^n)^{\alpha_n}}d\Pi(\eta)}{\Pi(B_n(\eta_0,\varepsilon_n))e^{-2\alpha_n n\varepsilon_n^2}}1_{C_n} + P_0(C_n^c) \\ &= \frac{\int_A\int p^n_\eta(x)^{\alpha_n}p^n_{\eta_0}(x)^{1-\alpha_n}d\mu(x)d\Pi(\eta)}{\Pi(B_n(\eta_0,\varepsilon_n))e^{-2\alpha_n n\varepsilon_n^2}} + o(1), \end{aligned} \tag{4.14}$$

where the last equality follows from Fubini's theorem. Set

$$\begin{aligned} A_n &:= \left\{\eta,\ \int p_\eta^n(x)^{\alpha_n} p_{\eta_0}^n(x)^{1-\alpha_n}\, d\mu(x) \leq e^{-4n\alpha_n\varepsilon_n^2}\right\} \\ &= \left\{\eta,\ -\frac{1}{n(1-\alpha_n)}\log(\int p_\eta^n(x)^{\alpha_n} p_{\eta_0}^n(x)^{1-\alpha_n}\, d\mu(x)) \geq 4\frac{\alpha_n\varepsilon_n^2}{1-\alpha_n}\right\} \\ &= \left\{\eta,\ \frac{1}{n}D_{\alpha_n}(p_\eta^n, p_{\eta_0}^n) \geq 4\frac{\alpha_n\varepsilon_n^2}{1-\alpha_n}\right\}. \end{aligned}$$

Substituting A_n into the second-last display and using the prior mass condition (4.13) yields

$$E_0\Pi_{\alpha_n}(A_n|Y^n) \leq \frac{\int_{A_n} e^{-4n\alpha_n\varepsilon_n^2}d\Pi(\eta)}{\Pi(B_n(\eta_0,\varepsilon_n))e^{-2\alpha_n n\varepsilon_n^2}} + o(1) \leq e^{-n\alpha_n\varepsilon_n^2} + o(1) = o(1),$$

since $n\alpha_n\varepsilon_n^2 \longrightarrow \infty$.

Lemma 6. For any distribution Π on S, any $C, \varepsilon > 0$ and $0 < \alpha \leq 1$, with P_0-probability at least $1 - \frac{1}{C^2 n\varepsilon^2}$, we have

$$\int_S \frac{p_\eta^n(Y^n)^\alpha}{p_{\eta_0}^n(Y^n)^\alpha} d\Pi(\eta) \geq \Pi(B_n(\eta_0,\varepsilon))e^{-\alpha(C+1)n\varepsilon^2}.$$

Proof.

Suppose $\Pi(B_n(\eta_0,\varepsilon)) > 0$ (otherwise the result is immediate), and denote by $\bar{\Pi} = \frac{\Pi(\cdot\cap B_n(\eta_0,\varepsilon))}{\Pi(B_n(\eta_0,\varepsilon))}$ the normalized prior to $B_n(\eta_0,\varepsilon)$. Now let us bound from below

$$\int_S \frac{p_\eta^n(Y^n)^\alpha}{p_{\eta_0}^n(Y^n)^\alpha} d\Pi(\eta) \geq \int_{B_n(\eta_0,\varepsilon)} \frac{p_\eta^n(Y^n)^\alpha}{p_{\eta_0}^n(Y^n)^\alpha} d\Pi(\eta) = \Pi(B_n(\eta_0,\varepsilon)) \int \frac{p_\eta^n(Y^n)^\alpha}{p_{\eta_0}^n(Y^n)^\alpha} d\bar{\Pi}(\eta). \tag{4.15}$$

Since $\bar{\Pi}$ is a probability measure on S, Jensen's inequality applied to the logarithm gives,

$$\log\left(\int \frac{p_\eta^n(Y^n)^\alpha}{p_{\eta_0}^n(Y^n)^\alpha} d\bar{\Pi}(\eta)\right) \geq \alpha \int \log\left(\frac{p_\eta^n(Y^n)}{p_{\eta_0}^n(Y^n)}\right) d\bar{\Pi}(\eta).$$

Consider now the random variable $Z := \int \log\left(\frac{p_\eta^n(Y^n)}{p_{\eta_0}^n(Y^n)}\right) d\bar{\Pi}(\eta)$. Then

$$\begin{aligned} E_0|Z| \leq \int_{B_n(\eta_0,\varepsilon)} E_0\left|\log\left(\frac{p_\eta^n(Y^n)}{p_{\eta_0}^n(Y^n)}\right)\right| d\bar{\Pi}(\eta) &= \int_{B_n(\eta_0,\varepsilon)} \int \left|\log\left(\frac{p_\eta^n(x)}{p_{\eta_0}^n(x)}\right)\right| p_{\eta_0}^n(x)d\mu^n(x)d\bar{\Pi}(\eta)) \\ &\leq n\varepsilon^2 + 1. \end{aligned}$$

Thus Z is integrable and using Fubini's theorem,

$$E_0 Z = \int_{B_n(\eta_0,\varepsilon)} \int \log\left(\frac{p^n_\eta(x)}{p^n_{\eta_0}(x)}\right) p^n_{\eta_0}(x) d\mu^n(x) d\bar{\Pi}(\eta) = \int_{B_n(\eta_0,\varepsilon)} -K(p^n_{\eta_0}, p^n_\eta) d\bar{\Pi}(\eta) \geq -n\varepsilon^2.$$

Turning to the variance,

$$\begin{aligned} \mathrm{Var}_0(Z) = \mathrm{Var}_0(-Z) &= E_0\left(\int \log\left(\frac{p^n_{\eta_0}(Y^n)}{p^n_\eta(Y^n)}\right) d\bar{\Pi}(\eta) - \int_{B_n(\eta_0,\varepsilon)} K(p^n_{\eta_0}, p^n_\eta) d\bar{\Pi}(\eta)\right)^2 \\ &= E_0\left(\int \log\left(\frac{p^n_{\eta_0}(Y^n)}{p^n_\eta(Y^n)}\right) - K(p^n_{\eta_0}, p^n_\eta) d\bar{\Pi}(\eta)\right)^2 \\ &\leq \int_{B_n(\eta_0,\varepsilon)} E_0\left(\log\left(\frac{p^n_{\eta_0}(Y^n)}{p^n_\eta(Y^n)}\right) - K(p^n_{\eta_0}, p^n_\eta)\right)^2 d\bar{\Pi}(\eta) \leq n\varepsilon^2 \end{aligned}$$

using that $\bar{\Pi}$ is supported on $B_n(\eta_0, \varepsilon)$. By Chebychev's inequality, $P_0(|Z - E(Z)| \geq Cn\varepsilon^2) \leq \frac{1}{Cn\varepsilon^2}$. Thus, on the event $\{|Z - E(Z)| \leq Cn\varepsilon^2\}$, which has a probability at least $1 - \frac{1}{Cn\varepsilon^2}$,

$$\log\left(\int \frac{p^n_\eta(Y^n)^\alpha}{p^n_{\eta_0}(Y^n)^\alpha} d\bar{\Pi}(\eta)\right) \geq \alpha(Z - EZ + EZ) \geq -\alpha(C+1)n\varepsilon^2.$$

Substituting this bound into (4.15) then gives the result.

CHAPTER 5

Variational Bayes

In this chapter we consider a popular approach to simulation from approximations *of posterior distributions. Choosing a best approximation of the posterior distribution from a given class of distribution is an optimisation problem, finding a (approximate) solution of which is often relatively fast for simple classes of distributions, such as* mean-field *classes. We explain the main idea and give theoretical backup: we give general conditions under which the variational posterior is shown to converge at the same rate as the posterior distribution itself.*

In this chapter, we consider the use of variational approximations to posterior distributions. The three complementary works by Alquier and Ridgway (AoS 2020), Yang, Bhattacharya adnd Pati (AoS 2020) and Zhang and Gao (AoS 2020) provide generic results and conditions under which approximations of posterior distributions in certain variational classes converge at (at least) the same rate as the posterior distribution itself. The case of tempered posterior distributions is also considered. These results apply already to a variety of models and priors, including many non-parametric or latent variable models. Here we follow mostly the presentation of Zhang and Gao (2020). The case of high-dimensional models needs a separate treatment, and is considered in Ray and Szabo (JASA 2022): we present it briefly.

1 General principles

In variational methods, one wishes to find a best (or close to best) approximation of a given *target* distribution (in the framework of these lectures it will be the *posterior* distribution) within a given class of simple distributions. The approximation will be quantified in terms of a measure of distance (or divergence) between distributions.

1.1 Divergences

The ρ–Rényi divergence between probability measures P and Q is defined as, for $\rho > 0$ and $\rho \neq 1$,

$$D_\rho(P,Q) = \frac{1}{\rho - 1} \log \int \left(\frac{dP}{dQ}\right)^{\rho-1} dP,$$

if P is absolutely continuous with respect to Q, and $+\infty$ otherwise. In the first case, and if P, Q have densities p, q with respect to μ, we have

$$D_\rho(P,Q) = \frac{1}{\rho - 1} \log \int p^\rho q^{1-\rho} d\mu.$$

If $\rho = 1$, one similarly defines, for $P \ll Q$ (otherwise we set it to $+\infty$ as above),

$$D_1 = K(P,Q) = \int \log\left(\frac{dP}{dQ}\right) dP$$

the Kullback–Leibler divergence between P and Q.

The following facts are classical, see for example the review paper by van Erven and Harremoës (IEEE 2010): $\rho \longrightarrow D_\rho(P,Q)$ is an increasing function; as $\rho \longrightarrow 1$, $D_\rho \longrightarrow D_1$. Also, $D_{1/2}, D_2$ are related respectively to the squared–Hellinger distance h^2 and the χ^2 divergence in the sense that

$$D_{1/2} = -2\log(1 - h^2/2), \qquad D_2 = \log(1 + \chi^2).$$

1.2 Variational families and optimisation

Definition 1. Let $\mathcal{S}$ be a family of distributions. The variational posterior with respect to the family $\mathcal{S}$ is the miminiser of the KL-divergence between any element of $\mathcal{S}$ and the posterior distribution. That is,

$$\hat{Q} = \operatorname*{argmin}_{Q \in \mathcal{S}} K(Q, \Pi(\cdot \,|\, X)). \tag{5.1}$$

Often, an exact solution to (5.1) is not available, but an approximation is; then the results that follow also hold for this approximation as long as the latter is close enough to an exact solution, if it exists, of (5.1).

If the class $\mathcal{S}$ is very large, it may even contain the true posterior in which case one would have $\hat{Q} = \Pi(\cdot \,|\, X))$. Of course, the purpose is to choose a class $\mathcal{S}$ sufficiently simple so that the optimisation problem (5.1) is simpler to solve compared to direct sampling from the posterior. For direct sampling from (an approximation of) $\Pi(\cdot \,|\, X)$, unless the posterior is available in closed form (which is rarely the case), one generally resorts to a general method such as MCMC (Monte Carlo Markov Chain). However in high dimensions or in problems with latent variables the MCMC method may be slow to converge. In such cases, variational approximations of the posterior are very popular in practice. The idea is to choose a class both sufficiently rich to approach the true posterior reasonably well, but at the same time sufficiently simple so that (5.1) is fast to solve numerically. In other words, there is a *trade–off* between good approximation properties and computability.

We will not focus much more here on this trade-off, but give two examples of popular classes below. Before this, we note that a nice property of the optimisation problem (5.1) is that the normalising constant in the expression of the posterior density from Bayes' formula (i.e. the denominator) vanishes when one optimises in $Q \in \mathcal{S}$. Indeed, writing $dQ = qd\mu$ and, using Bayes' formula, $d\Pi(\theta \,|\, X) = p_\theta(X)\pi(\theta)/\int p_\theta(X)\pi(\theta)d\mu$, and noting that $D_X = \int p_\theta(X)\pi(\theta)d\mu$ depends only on X but not on Q or θ,

$$\begin{aligned} K(Q, \Pi(\cdot \,|\, X)) &= \int \log\left(\frac{q(\theta)}{p_\theta(X)\pi(\theta)/D_X}\right) q(\theta)d\mu \\ &= \int \log\left(\frac{q(\theta)}{p_\theta(X)\pi(\theta)}\right) q(\theta)d\mu + \log D_X, \end{aligned}$$

and the last term is independent of Q so it is enough to minimise the first term. In particular, when solving the variational problem, there is no need to compute D_X, which often can be delicate or at least time-consuming. The previous identity is sometimes rewritten

$$\log D_X = K(Q, \Pi(\cdot \,|\, X)) + \int \log\left(\frac{p_\theta(X)\pi(\theta)}{q(\theta)}\right) q(\theta)d\mu.$$

The term $\log D_X$ does not depend on Q and is called *evidence* (it is the logarithm of the density of X in the Bayesian model). Since the KL–divergence is nonnegative, we have

$$\log D_X \geq \int \log\left(\frac{p_\theta(X)\pi(\theta)}{q(\theta)}\right) q(\theta)d\mu.$$

The term $\int \log\left(p_\theta(X)\pi(\theta)/q(\theta)\right) q(\theta)d\mu$ is called the *Evidence Lower BOund* (ELBO). Minimizing the KL–divergence $K(Q, \Pi(\cdot \,|\, X))$ is equivalent to maximising the ELBO.

Definition 2. [Mean–field classes] Suppose the parameter $\theta \in \Theta$ can be written $\theta = (\theta_1, \theta_2, \dots, \theta_m)$ with m an integer, or $m = +\infty$. The mean–field variational class $\mathcal{S}_{MF}$ is the class of distributions

$$\mathcal{S}_{MF} = \left\{ Q \ : \ dQ(\theta) = \prod_{j=1}^{m} dQ_j(\theta_j) \right\}.$$

That is, $\mathcal{S}_{ML}$ consists of product measures only. As a special subcase, one may consider specific distributions for the Q_j. Let $\mathcal{G} = \{\mathcal{N}(\mu, \sigma^2),\ \mu \in \mathbb{R}, \sigma \geq 0\}$ be the set of 1–dimensional Gaussian distributions. The Gaussian mean field class is

$$\mathcal{S}_{GMF} = \left\{ Q \ : \ dQ(\theta) = \prod_{j=1}^{m} dQ_j(\theta_j), \quad Q_j \in \mathcal{G} \ (\forall j) \right\}.$$

The idea of the mean–field class is to ignore dependencies in the posterior distribution and to approximate it by a distribution of product form. Of course, some information is then typically lost in this process: for instance, for $\theta \in \mathbb{R}^2$, a Gaussian distribution $\mathcal{N}(\theta, \Sigma)$ with Σ a *non-diagonal* 2×2 covariance matrix cannot be perfectly approximated by a product of 1–dimensional Gaussians. One may think though that the loss is 'of the order of a multiplicative factor in the variance', so maybe not huge.

From the implementation point of view, the problem of minimising the KL–divergence $K(Q, \Pi(\cdot \mid X))$, ot equivalently maximising the ELBO, is an *optimisation* problem. There is a vast literature on algorithms performing this task; a famous algorithm is CAVI, for Coordinate Ascent Variational Inference. We refer to the review paper by Blei et al. (2017) for more on the algorithmical aspect. The main idea is that in complex models this optimisation problem can often be much faster than the task of sampling from the posterior. From the theoretical perspective, in order to validate this approach, it is then natural to ask whether the VB–posterior $\hat{Q}$ is also a 'good estimator' of the true parameter θ_0, i.e. if $\hat{Q}$ contracts at a rate ε_n towards θ_0, if possible (at least) as fast as the contraction rate of the original posterior distribution. We give generic sufficient conditions for this along with a few examples in the next sections.

2 A generic result for variational posteriors

2.1 Statement

Consider a statistical model $\mathcal{P} = \{P_\theta^{(n)},\ \theta \in \Theta\}$ as before in the course, dominated by $\mu^{(n)} = \mu$, where Θ is a parameter set (e.g. space of functions).

Let $L(\cdot, \cdot)$ be a loss function between probability measures such that $L(P, Q) \geq 0$ for any such measures P, Q. Examples of losses include $L = nh^2$, i.e. n times the Hellinger squared distance, or also, if θ is a sequence, $L(P_\theta^{(n)}, P_{\theta'}^{(n)}) = n\|\theta - \theta'\|^2$. Note the specific normalisation chosen with a multiplicative factor n: this is related to the fact that a typical example is the one of product measures $P_\theta^{(n)} = P_\theta^{\otimes n}$ for which typical divergences such as the KL scale with n (recall that $K(P_\theta^{\otimes n}, P_{\theta'}^{\otimes n}) = nK(P_\theta, P_{\theta'})$).

Generic conditions. Consider the following conditions

(T) For any $\varepsilon > \varepsilon_n$, there exists $\Theta_n(\varepsilon)$ measurable subsets of Θ and φ_n test functions such that

$$E_{\theta_0}\varphi_n + \sup_{\theta\in\Theta_n(\varepsilon),\ L(P_\theta^{(n)},P_{\theta_0}^{(n)})>C_1 n\varepsilon^2} E_\theta(1-\varphi_n) \le e^{-Cn\varepsilon^2}.$$

(S) For any $\varepsilon > \varepsilon_n$,

$$\Pi(\Theta_n(\varepsilon)^c) \le e^{-Cn\varepsilon^2}.$$

(P) There exists $\rho > 1$ such that

$$\Pi\left(D_\rho(P_{\theta_0}^{(n)}, P_\theta^{(n)}) \le C_3 n\varepsilon_n^2\right) \ge e^{-C_2 n\varepsilon_n^2}.$$

These conditions are almost identical to the ones used before in the lectures. There are two differences. First, (T) and (S) are required to hold for any $\varepsilon > \varepsilon_n$. It is generally not too difficult to find a sequence of sets $\Theta_n(\varepsilon)$ indexed by ε verifying (T) for any $\varepsilon > \varepsilon_n$ (and not just for $\varepsilon = \varepsilon_n$). Second, the KL-neighborhood used before is replaced by a D_ρ–neighborhood where $\rho > 1$. This is useful in that it enables one to obtain posterior masses of complements of neighborhoods that decrease exponentially fast to 0 (instead of just polynomially – recall we used simply Tchebychev's inequality in proving the GGV theorem – here we get rather an exponential-type inequality).
See Lemma 4, where it is shown that under such slightly strengthened assumptions compared to the GGV Theorem, the original posterior converges at rate ε_n and with an exponential decrease to 0.

In what follows, for a given function f, we use the notation $Qf = \int f dQ$.

Theorem 1. [Convergence rate for $\hat{Q}$] Let (ε_n) be a sequence such that $n\varepsilon_n^2 \ge 1$. Let Π be a prior distribution on Θ. Consider $\hat{Q}$ the variational Bayes approximation (5.1) to the posterior distribution $\Pi[\cdot\,|\,X]$ with variational class $\mathcal{S}$ and set

$$\gamma_n^2 = \frac{1}{n}\inf_{Q\in\mathcal{S}} E_{\theta_0}K(Q,\Pi(\cdot\,|\,X)). \tag{5.2}$$

Suppose the generic conditions (T), (S), (P) are verified with rate ε_n, loss function L, positive constants C_1, C_2, C_3, $C > C_2 + C_3 + 2$ and $\rho > 1$. Then there exists $M = M(C_1, C, \rho)$ such that

$$E_{\theta_0}\hat{Q}L(P_\theta^{(n)}, P_{\theta_0}^{(n)}) \le Mn(\varepsilon_n^2 + \gamma_n^2).$$

The interpretation is as follows: ε_n is the convergence rate of the original posterior distribution in terms of the loss L, while γ_n is the contribution arising from considering the variational approximation. Note that γ_n^2 is defined in terms of the posterior and is this still implicit. In the next lines, we bound it from above by a more universal quantity.

2.2 Sufficient conditions

Lemma 1. The rate γ_n defined in (5.2) verifies

$$\gamma_n^2 \leq \frac{1}{n} \inf_{Q \in \mathcal{S}} \left[K(Q, \Pi) + QK(P_{\theta_0}^{(n)}, P_\theta^{(n)}) \right]. \tag{5.3}$$

Proof.

Denote as shorthand $\Pi_X = \Pi(\cdot \,|\, X)$ and $P_\Pi^{(n)} = \int P_\theta^{(n)} d\Pi(\theta)$ so that the denominator in Bayes' formula is $p_\Pi^{(n)} = \int p_\theta^{(n)} d\Pi(\theta)$ the marginal density of X in the Bayesian setting. Bayes' formula writes $d\Pi_X = p_\theta^{(n)} d\Pi / p_\Pi^{(n)}$. Then $K(Q, P_X) = \int \log(dQ/d\Pi_X) dQ$, and

$$\log \frac{dQ}{d\Pi_X} = \log \frac{dQ}{d\Pi} + \log \frac{p_\Pi^{(n)}}{p_\theta^{(n)}}.$$

Deduce that $K(Q, P_X)$ can be further written as

$$E_{\theta_0} K(Q, \Pi_X) = \int \log \frac{dQ}{d\Pi} dQ + E_{\theta_0} \int \log \frac{p_\Pi^{(n)}}{p_\theta^{(n)}} dQ$$

and, using Fubini's theorem and $K(P, Q) \geq 0$,

$$\begin{aligned} E_{\theta_0} \int \log \frac{p_\Pi^{(n)}}{p_\theta^{(n)}} dQ &= Q \int \log \frac{dP_\Pi^{(n)}}{dP_\theta^{(n)}} dP_{\theta_0}^{(n)} \\ &= Q \left[\int \log \frac{dP_{\theta_0}^{(n)}}{dP_\theta^{(n)}} dP_{\theta_0}^{(n)} + \int \log \frac{dP_\Pi^{(n)}}{dP_{\theta_0}^{(n)}} dP_{\theta_0}^{(n)} \right] \\ &= Q \left[K(P_{\theta_0}^{(n)}, P_\theta^{(n)}) - K(P_{\theta_0}^{(n)}, P_\Pi^{(n)}) \right] \leq QK(P_{\theta_0}^{(n)}, P_\theta^{(n)}). \end{aligned}$$

The lemma follows by using the definition of γ_n and taking the infimum over $Q \in \mathcal{S}$.

By combining Lemma 1 and Theorem 1, we obtain the following, which is just another way to write the previous Lemma: to bound γ_n, it suffices, for $Q_1 \in \mathcal{S}$, to bound $K(Q_1, \Pi) + Q_1 K(P_{\theta_0}^{(n)}, P_\theta^{(n)})$.

Corollary 1. Under the conditions of Theorem 1, suppose further that, for

$$\mathcal{E} = \left\{ Q \;:\; \text{Supp}(Q) \subset \{\theta \;:\; K(P_{\theta_0}^{(n)}, P_\theta^{(n)}) \leq C_2 n \varepsilon_n^2\} \right\},$$

it holds

$$\inf_{Q \in \mathcal{S} \cap \mathcal{E}} K(Q, \Pi) \leq C_1 n \varepsilon_n^2. \tag{5.4}$$

Then for a large enough constant M',

$$E_{\theta_0}\hat{Q}L(P_\theta^{(n)}, P_{\theta_0}^{(n)}) \leq M' n\varepsilon_n^2.$$

2.3 Result for mean-field class

Theorem 2. [Convergence rate for $\hat{Q}$, mean-field case] Under the conditions of Theorem 1, suppose that one can find a distribution $\tilde{Q}$ in the mean-field class $\mathcal{S}_{MF}$ and a subset

$$\mathcal{A}_m := \bigotimes_{j=1}^{m} \tilde{\Theta}_j \subset \Theta$$

of product form which verifies

$$(A) \qquad \mathcal{A}_m \subset \left\{ \theta : \ K(P_{\theta_0}^{(n)}, P_\theta^{(n)}) \leq C_1 n\varepsilon_n^2, \ \log\left(\frac{d\tilde{Q}}{d\Pi}(\theta)\right) \leq C_2 n\varepsilon_n^2 \right\},$$

$$(B) \qquad \tilde{Q}(\mathcal{A}_m) \geq e^{-C_3 n\varepsilon_n^2}.$$

Then the term γ_n in (5.2) with $\mathcal{S} = \mathcal{S}_{MF}$ verifies

$$\gamma_n^2 \leq (C_1 + C_2 + C_3)\varepsilon_n^2.$$

In particular, the conclusion of Theorem 1 holds with rate ε_n^2.

Note that in case the prior itself belongs to $\mathcal{S}_{MF}$, one can take $\tilde{Q} = \Pi$, which simplifies the conditions even further. Also, condition (B) in the Theorem can be interpreted as asking a prior mass condition which is “coherent with the structure of the variational class”.

2.4 An example of application: the sequence model

Consider the Gaussian sequence model $X_j = \theta_j + \xi_j/\sqrt{n}$ with ξ_i independent $\mathcal{N}(0,1)$ variables. Suppose the true $\theta_0 = (\theta_{0,1}, \theta_{0,2}, \ldots,)$ belongs to a Sobolev ball

$$\theta_0 \in \{\theta : \sum_{j\geq 1} j^{2\beta}\theta_j^2 \leq L^2\}.$$

Also set $L(P_\theta^{(n)}, P_{\theta'}^{(n)}) = n\|\theta - \theta'\|^2$.

Consider a sieve prior Π defined hierarchically: sample k from a distribution π on integers; then given k sample $\theta_1, \ldots, \theta_k$ independently with density f_j on coordinate j; set $\theta_j = 0$ for all $j > k$.
Let us consider the variational posterior $\hat{Q}$ using the mean-field class

$$\hat{Q} = \operatorname*{argmin}_{Q\in\mathcal{S}_{MF}} K(Q, \Pi(\cdot \,|\, X)),$$

where $\mathcal{S}_{MF}$ is as in Definition 2. The next result shows that the variational posterior $\hat{Q}$ is adaptive to smoothness and reaches an optimal contraction rate in ℓ^2 up to a logarithmic term.

Theorem 3. In the Gaussian sequence model, suppose θ_0 and Π are as described above, for some $\beta, L > 0$. Take as π the prior on integer such that $\pi(k) \propto e^{-\tau k}$ and take f_j to be the standard normal density for any coordinate j that is nonzero under the prior. Then

$$E_{\theta_0}\hat{Q}\|\theta-\theta_0\|^2 \lesssim \left(\frac{\log n}{n}\right)^{\frac{2\beta}{2\beta+1}}.$$

Proof.

Define k_n to be the integer part of $(n/\log n)^{1/(2\beta+1)}$. One first checks that the prior and model verify the conditions of Theorem 1. The prior mass condition is easy to verify by including an ℓ^∞–neighborhood within the KL–neighborhood (which here equals an ℓ^2–neighborhood). One takes as sieve set the set of sequences whose first k_n coefficients are arbitrary reals and coefficients from index k_n+1 are zero; then one only needs to check the 'strengthened' testing condition: since this does not directly relate to the variational posterior we omit the details, but note that this follows by 'local entropy' arguments as in Ghosal, Ghosh, van der Vaart (2000), Section 7.

We now apply Theorem 2. Let us set $\mathcal{A} = \tilde{\Theta}_1 \times \cdots \times \tilde{\Theta}_j \times \cdots$, with

$$\tilde{\Theta}_j = \begin{cases} \left[\theta_{0,j} - 1/\sqrt{n}, \theta_{0,j} + 1/\sqrt{n}\right], & \text{if } j \le k_n, \\ \{0\}, & \text{if } j > k_n. \end{cases}$$

Let us also define the mean-field distributions $\tilde{Q}(k)$ for any $k \ge 1$ as

$$\tilde{Q}(k) = \bigotimes_{j\ge1} \tilde{Q}_j, \qquad \tilde{Q}_j = \begin{cases} f_j(x)dx, & \text{if } j \le k, \\ \delta_0, & \text{if } j > k. \end{cases}$$

We set $\tilde{Q} = \tilde{Q}(k_n)$. One verifies the condition of Theorem 2. That $\tilde{Q}$ puts enough mass on $\mathcal{A}$ is straightforward: using that $\theta_{0,j}$'s are all bounded by L, one gets $\tilde{Q}(\mathcal{A}) \ge (C/\sqrt{n})^{k_n} \ge e^{-Ck_n \log n}$. Recalling that in the sequence model $K(P_{\theta_0}^{(n)}, P_\theta^{(n)}) = n\|\theta-\theta_0\|^2$, it follows from the definition of $\mathcal{A}$ that $\theta \in \mathcal{A}$ implies $\|\theta-\theta_0\|^2 \le k_n/n + \sum_{j>k_n}\theta_{0,j}^2 \lesssim n\varepsilon_n^2$ if ε_n^2 is the rate in the statement.

Finally, one notes that if $\theta \in \mathcal{A}$, the prior Π equals the mixture $\sum_{j\ge1}\pi(j)Q(j)$, so $d\tilde{Q}/d\Pi = 1/\pi(k_n)$ and then $\log(1/\pi(k_n)) \lesssim k_n \lesssim n\varepsilon_n^2$ by definition of π, as required.

3 Proof of the generic theorem

Useful lemmas and their proofs

The proofs of Theorems 1 and 2 are quite direct applications of the combination of the next Lemmas. Lemma 4 makes the conclusion of Theorem 1 more precise by assuming slightly stronger conditions.

Lemma 2. Let $f \geq 0$ and let P, Q be two probability measures. Then

$$\int f\,dQ \leq K(Q,P) + \log \int e^{f(x)}\,dP(x).$$

Proof.

One writes, with the notation $\int e^f\,dP = Pe^f$,

$$\begin{aligned} K(Q,P) + \log \int e^{f(x)}\,dP(x) &= \int \log\left(\frac{dQ}{dP}\cdot Pe^f\right) dQ \\ &= \int \log\left(\frac{dQ}{e^f\,dP}\cdot e^f \cdot Pe^f\right) dQ = \int \log\left(\frac{dQ}{dP'}\right) dQ + \int f\,dQ, \end{aligned}$$

with $dP' = e^f\,dP/(Pe^f)$. The results follows by using $\int \log\left(dQ/dP'\right) dQ = K(Q,P') \geq 0$.

Lemma 3.

$$E_{\theta_0}\hat{Q}L(P_\theta^{(n)}, P_{\theta_0}^{(n)}) \leq \inf_{a>0} \frac{1}{a}\left[\inf_{Q\in\mathcal{S}} E_{\theta_0}K(Q,\Pi(\cdot\,|\,X)) + \log E_{\theta_0}\int e^{aL(P_\theta^{(n)},P_{\theta_0}^{(n)})}\,d\Pi(\theta\,|\,X)\right].$$

Proof.

One applies Lemma 2 with $Q = \hat{Q}$, $P = \Pi[\cdot\,|\,X]$ and $f(\theta) = aL(P_\theta^{(n)}, P_{\theta_0}^{(n)})$ for a given $a > 0$. Deduce, writing $P_\theta = P_\theta^{(n)}$ as shorthand,

$$a\hat{Q}L(P_\theta, P_{\theta_0}) \leq K(\hat{Q},\Pi[\cdot\,|\,X]) + \log\left[\Pi[\cdot\,|\,X]\left\{e^{aL(P_\theta,P_{\theta_0})}\right\}\right].$$

First, one takes the expectation under E_{θ_0} and uses Jensen's inequality with the logarithm. Upon noting that $K(\hat{Q},\Pi[\cdot\,|\,X]) \leq K(Q,\Pi[\cdot\,|\,X])$ for any $Q \in \mathcal{S}$ by definition, the result follows by dividing by a, taking the infimum over such Q's, followed by the infimum over $a > 0$.

Lemma 4. Suppose conditions (T), (S), (P) hold. Then for $\lambda = \rho - 1$ and for any $\varepsilon > \varepsilon_n$,

$$E_{\theta_0}\Pi(L(P_\theta^{(n)}, P_{\theta_0}^{(n)}) > C_1 n\varepsilon^2\,|\,X) \leq e^{-Cn\varepsilon^2} + e^{-\lambda n\varepsilon^2} + 2e^{-n\varepsilon^2}.$$

Proof.

Writing $P_\theta = P_\theta^{(n)}$ as shorthand, let us set, with $\rho := 1 + \lambda$,

$$U_n = \left\{ \theta : L(P_\theta, P_{\theta_0}) > C_1 n\varepsilon^2 \right\}, \quad K_n = \left\{ \theta : D_\rho(P_{\theta_0}, P_\theta) \leq C_3 n\varepsilon^2 \right\},$$

and, for $\tilde{\Pi} = \Pi|_{K_n} = \Pi(\cdot \cap K_n)/\Pi(K_n)$,

$$A_n = \left\{ \int \frac{dP_\theta}{dP_{\theta_0}}(X) d\tilde{\Pi}(\theta) > e^{-(C_3+1)n\varepsilon^2} \right\}.$$

Using the tests φ_n in ($\underline{T}$), the quantity at stake for the Lemma is bounded by

$$E_{\theta_0}\Pi(U_n \,|\, X) \leq E_{\theta_0}\left[\Pi(U_n \,|\, X)(1-\varphi_n)1_{A_n}\right] + P_{\theta_0}A_n^c + E_{\theta_0}\varphi_n.$$

The last term is bounded by $e^{-Cn\varepsilon^2}$ by ($\underline{T}$). Using Markov's inequality,

$$\begin{aligned} P_{\theta_0}A_n^c &\leq P_{\theta_0}\left[\left\{ \int \frac{dP_\theta}{dP_{\theta_0}}(X) d\tilde{\Pi}(\theta) \right\}^{-\lambda} > e^{\lambda(C_3+1)n\varepsilon^2}\right] \\ &\leq e^{-\lambda(C_3+1)n\varepsilon^2} E_{\theta_0}\left\{ \int \frac{dP_\theta}{dP_{\theta_0}}(X) d\tilde{\Pi}(\theta) \right\}^{-\lambda} \\ &\leq e^{-\lambda(C_3+1)n\varepsilon^2} E_{\theta_0}\left\{ \int \left(\frac{dP_\theta}{dP_{\theta_0}}(X) \right)^{-\lambda} d\tilde{\Pi}(\theta) \right\}, \end{aligned}$$

where the last line uses Jensen's inequality and convexity of $x \longrightarrow x^{-\lambda}$ on $\mathbb{R}^+$. Fubini's theorem implies

$$E_{\theta_0}\left\{ \int \left(\frac{dP_\theta}{dP_{\theta_0}}(X) \right)^{-\lambda} d\tilde{\Pi}(\theta) \right\} = \int\int \frac{(dP_{\theta_0})^{\lambda+1}}{(dP_\theta)^\lambda} d\tilde{\Pi}(\theta) = \int e^{\lambda D_{1+\lambda}(P_{\theta_0}, P_\theta)} d\tilde{\Pi}(\theta) \leq e^{\lambda C_3 n\varepsilon^2},$$

using the definition of K_n, on which $\tilde{\Pi}$ is supported. Putting the previous inequalities together gives $P_{\theta_0}A_n^c \leq e^{-\lambda n\varepsilon^2}$. It remains to bound $\Pi(U_n \,|\, X)(1-\varphi_n)$ on the event A_n. By definition, on A_n,

$$\int \frac{dP_\theta}{dP_{\theta_0}}(X) d\Pi(\theta) \geq \Pi[K_n] \int \frac{dP_\theta}{dP_{\theta_0}}(X) d\tilde{\Pi}(\theta) \geq \Pi[K_n] e^{-(C_3+1)n\varepsilon^2}.$$

We have $\Pi[K_n] \geq e^{-C_2 n\varepsilon_n^2} \geq e^{-C_2 n\varepsilon^2}$ for $\varepsilon > \varepsilon_n$ using ($\underline{P}$). The last display is thus bounded from below by $e^{-(C_2+C_3+1)n\varepsilon^2}$. Using this fact, one can bound the denominator of Bayes' formula (written with dP_θ/dP_{θ_0}) from below to get

$$\begin{aligned} E_{\theta_0}\left[\Pi(U_n \,|\, X)(1-\varphi_n)1_{A_n}\right] &\leq e^{(C_2+C_3+1)n\varepsilon^2} E_{\theta_0}\left[\int_{U_n} \frac{dP_\theta}{dP_{\theta_0}}(X)(1-\varphi_n) d\Pi(\theta)\right] \\ &\leq e^{(C_2+C_3+1)n\varepsilon^2} \int_{U_n} P_\theta(1-\varphi_n) d\Pi(\theta). \end{aligned}$$

Let us further bound from above, using $(\underline{T})$, $(\underline{S})$,

$$\int_{U_n} P_\theta(1-\varphi_n)d\Pi(\theta) \le \Pi\left[\Theta_n(\varepsilon)^c\right] + \int_{U_n\cap\Theta_n(\varepsilon)} P_\theta(1-\varphi_n)d\Pi(\theta) \le e^{-Cn\varepsilon^2} + e^{-Cn\varepsilon^2}.$$

Putting the previous inequalities together yields, provided $C > C_2 + C_3 + 2$,

$$E_{\theta_0}\left[\Pi(U_n \,|\, X)(1-\varphi_n)1_{A_n}\right] \le 2e^{(C_2+C_3+1)n\varepsilon^2 - Cn\varepsilon^2} \le 2e^{-Cn\varepsilon^2}.$$

Combining all previous bounds gives the result.

Lemma 5. Suppose the random variable X verifies

$$P(X \ge t) \le c_1 e^{-c_2 t} \qquad \text{for any } t \ge t_0 > 0.$$

Then for any $a \in (0, c_2/2]$,

$$Ee^{aX} \le e^{at_0} + c_1.$$

Proof.

Using the formula $EY \le M + \int_M^\infty P[Y \ge y]dy$ for $Y = e^{aX}$ and the assumption,

$$E[e^{aX}] \le M + \frac{c_1}{c_2 - a} aM^{1-(c_2/a)}.$$

Setting $M = e^{at_0}$ and using $a \le c_2/2$, the former is bounded by $M + c_1(a/a)e^{a-c_2} \le M + c_1$.

Proof of the main results

Proof of Theorem 1.

From Lemma 4, one deduces that for any $t \ge t_0 = C_1 n\varepsilon_n^2$,

$$E_{\theta_0}\Pi[L(P_\theta, P_{\theta_0}) > t \,|\, X] \le C_1 e^{-C_2 t}.$$

One then uses Lemma 5 to deduce, for small a, that

$$E_{\theta_0}\left\{\Pi[\cdot \,|\, X]\left[e^{aL(P_\theta, P_{\theta_0})}\right]\right\} \le 4 + e^{aC_1 n\varepsilon_n^2}$$

for small a. The result now follows from an application of Lemma 3.

Proof of Theorem 2.

Invoking Lemma 1, it is enough to find $Q \in \mathcal{S}_{MF}$ such that

$$QK(P_{\theta_0}, P_\theta) + K(Q, \Pi) \le (C_1 + C_2 + C_3)\varepsilon_n^2.$$

> Define $Q = \bigotimes_{j=1}^m Q_j$, with $Q_j = \tilde{Q}_j|_{\tilde{\Theta}_j}$ the restriction of $\tilde{Q}_j$ to $\tilde{\Theta}_j$, both defined in the statement of the lemma. By definition $Q \in \mathcal{S}_{MF}$ and $\mathrm{Supp}(Q) \subset \bigotimes_{j=1}^m \tilde{\Theta}_j$.

4 High-dimensional regression

Consider the high-dimensional regression model

$$Y = X\theta + \varepsilon,$$

where the notation is as in the corresponding earlier chapter. Here for simplicity we focus on one example of design matrix, namely we assume it has independent Gaussian entries

$$X_{ij} \sim \mathcal{N}(0,1) \quad \text{iid.} \tag{5.5}$$

Let us consider a subset-selection prior on θ (with here p instead of n)

$$k \sim \pi_p, \qquad S \mid k \sim \mathrm{Unif}(\mathcal{S}_k), \qquad \theta \mid S \sim \bigotimes_{i \in S} \Gamma \otimes \bigotimes_{i \notin S} \delta_0, \tag{5.6}$$

with here $\Gamma = \mathrm{Lap}(\lambda)$ a Laplace distribution with parameter λ. Suppose the following slightly faster than exponential decrease: there exist constants $A_1, \ldots, A_4 > 0$ with

$$A_1 p^{-A_3} \pi_p(s-1) \leq \pi_p(s) \leq A_2 p^{-A_4} \pi_p(s-1), \tag{5.7}$$

for $s = 1, \ldots, p$. This condition is satisfied for instance for the following hierarchical Bayes version of the spike and slab prior, for some fixed $u > 1$ and $\lambda > 0$,

$$\alpha \sim \mathrm{Beta}(1, p^u)$$
$$\theta = (\theta_i)_{1 \leq i \leq p} \mid \alpha \sim \bigotimes_{i=1}^p (1-\alpha)\delta_0 + \alpha \mathrm{Lap}(\lambda),$$

As a variational class, let us consider the *mean-field* spike and slab class

$$\mathcal{P}_{MF} = \left\{ P_{\mu,\sigma,\gamma} = \bigotimes_{i=1}^p (1-\gamma_i)\delta_0 + \gamma_i \mathcal{N}(\mu_i, \sigma_i^2), \quad \mu_i \in \mathbb{R}, \sigma_i \in \mathbb{R}^+, \gamma_i \in [0,1] \right\}. \tag{5.8}$$

Define the corresponding variational Bayes posterior distribution

$$\tilde{\Pi} = \underset{P_{\mu,\sigma,\gamma} \in \mathcal{P}_{MF}}{\mathrm{argmin}} \; K(P_{\mu,\sigma_\gamma}, \Pi(\cdot \mid Y)). \tag{5.9}$$

By taking the mean-field class (5.8) in this context, one enforces substantial independence in the variational posterior distribution, with a much reduced complexity in terms of models: there are only p inclusion variables in (5.8) instead of 2^p models the posterior puts mass on. Note also that while one may choose a Gaussian distribution for slabs from the variational class, it is important to keep a Laplace slab in the prior itself (otherwise one may face over-shrinkage).

Theorem 4. Let the prior Π be a subset-selection prior as in (5.6) that satisfies (5.7) with slab Γ in (5.6) a standard Laplace variable. Suppose also that $s_n = o(\sqrt{n/\log n})$. Then, on an event of overwhelming probability under the law of X as in (5.5), for M_n going to infinity arbitrarily slowly,

$$\sup_{\theta_0 \in \ell_0[s_n]} E_{\theta_0} \tilde{\Pi}\left(\theta : \|\theta - \theta_0\|_2 > M_n \sqrt{\frac{s_n \log p}{n}} \right) = o(1).$$

Theorem 4 is a special case of Theorem 1 in Ray and Szabo (2022), obtained by using the conditions on X assumed therein are verified for the design (5.5) with overwhelming probability. On the other hand, under the same conditions on X and for the same prior on Π, it follows from Castillo, Schmidt-Hieber and van der Vaart (AoS 2015), Theorem 2, that the original posterior distribution $\Pi[\cdot \mid Y]$ converges towards θ_0 in $\|\cdot\|_2$ norm at the same rate $\sqrt{s_n \log p/n}$ which can be shown to be (near)-optimal in this setting.

This shows that the variational Bayes approximation $\tilde{\Pi}[\cdot \mid Y]$ converges at the same rate as the original posterior $\Pi[\cdot \mid Y]$. An advantage here of the VB-posterior is that this approximation is quite fast to compute, while sampling from the original posterior typically requires the use of MCMC algorithms that scale significantly slower in terms of dimension.

For more details on the proposed variational algorithm (used to solve the optimisation problem, i.e. finding the best approximant in the considered mean-field class) we refer to Ray and Szabo (2022); the method is implemented in the R package sparsevb, which covers both linear and logistic regression.