Bayesian nonparametrics, convergence and limiting shape of posterior distributions

Ismael Castillo

Habilitation à diriger des recherches

Bayésien non-paramétrique, convergence et forme limite de lois a posteriori

Ismaël Castillo

Le 18 novembre 2014

Rapporteurs :

Fabrice Gamboa	Université Toulouse 3
Subhashis Ghoshal	North Carolina State University
Dominique Picard	Université Paris 7

Jury :

Lucien Birgé	Université Paris 6
Stéphane Boucheron	Université Paris 7
Fabrice Gamboa	Université Toulouse 3
Elisabeth Gassiat	Université Paris-Sud
Dominique Picard	Université Paris 7
Judith Rousseau	Université Paris-Dauphine

Remerciements

Je suis très reconnaissant à Dominique Picard, Fabrice Gamboa et Subhashis Ghosal d'avoir accepté d'être rapporteurs de ce mémoire.

C'est un plaisir de remercier ici Elisabeth Gassiat, qui m'a initié à la recherche, pour tous ses conseils et pour l'intérêt qu'elle a toujours porté à mes travaux.

Je remercie aussi Lucien Birgé, Stéphane Boucheron et Judith Rousseau de me faire l'honneur de participer au jury.

Ce mémoire doit beaucoup à mon séjour d'il y a quelques années à Amsterdam en tant que Postdoc. Je voudrais remercier très chaleureusement Aad van der Vaart pour cette opportunité, ainsi que pour l'autonomie qu'il m'a donnée, me permettant de développer des idées à mon rythme, et pour toutes les discussions et ses conseils pendant et après mon séjour.

C'est un plaisir de remercier ici mes co-auteurs Aad, Catherine, Céline, Dominique, Eric, Gérard, Jean-Michel, Johannes, Judith et Richard. J'ai beaucoup appris auprès de vous, et espère que nous pourrons collaborer très bientôt sur de nouveaux projets. Plusieurs de nos travaux ont pu voir le jour grâce à des séjours de recherche à l'étranger, et je remercie notamment Aad pour les invitations à Amsterdam et Leiden, Eric pour toutes nos discussions à Delft, et Richard pour un passionant séjour de deux mois à Cambridge.

J'ai été très bien accueilli au LPMA, et tiens à remercier particulièrement Stéphane pour avoir guidé mes premiers pas –et plusieurs des suivants– au laboratoire, Dominique et Gérard pour nos discussions toujours source d'inspiration, ainsi que Lucien et Sacha pour les questions et discussions autour d'exposés notamment. Merci également à Arnak, Catherine, Dominique, Richard N., Richard S. et Sacha pour tous vos conseils relatifs à l'enseignement.

Merci aussi à mes collègues et amis du LPMA (et du voisin LSTA), d'Amsterdam (VU et UvA), de Cambridge (merci aussi au Queens' college pour son accueil distingué) et de Leiden.

Last but not least, merci à mes parents, et à Egle, pour son soutien constant.

Résumé

Ce document de synthèse présente mes travaux en statistique bayésienne non-paramétrique. Il s'agit d'étudier le comportement de la mesure a posteriori, une mesure aléatoire qui dépend des observations, pour des modèles statistiques de dimension très grande ou infinie, tels que les modèles non-paramétriques et semi-paramétriques.

Si l'origine de l'approche bayésienne remonte à Thomas Bayes [7] au XVIII^{e} siècle, l'utilisation des méthodes bayésiennes dans des modèles statistiques complexes s'est spectaculairement développée depuis le milieu des années 1990. Une raison est le développement d'algorithmes efficaces de simulation suivant une approximation de la loi a posteriori. En parallèle, le besoin de mesurer la qualité de la convergence de la loi a posteriori a conduit au développement progressif d'outils théoriques d'étude de la convergence de ces lois depuis une quinzaine d'années. C'est cette approche de mesure fréquentiste des performances des procédures bayésiennes que je suis, et que nous suivons avec mes co-auteurs, dans les travaux décrits ci-après.

J'ai choisi pour cette présentation un plan thématique.

Dans le CHAPITRE 1, nous définissons une notion simple de borne inférieure [P4] pour la vitesse de convergence de la loi a posteriori. Ce concept permet notamment de montrer que certaines vitesses de convergence ne peuvent être améliorées, ainsi que de donner des conditions qualitatives sur les lois a priori. Nous illustrons cette définition par des exemples étudiés dans [P9] et [P15].

Dans le CHAPITRE 2, nous présentons des résultats de bornes supérieures pour la convergence a posteriori, dans différents contextes statistiques : un cadre géométrique [P11] et des modèles parcimonieux [P9, P15]. Nous proposons également une approche [P12] pour étudier la vitesse de convergence a posteriori en norme infinie.

Dans le CHAPITRE 3, nous considérons la forme limite des lois a posteriori. Cette forme limite est étudiée tout d'abord dans un cadre semi-paramétrique ; nous présentons un résultat général pour les lois a priori gaussiennes [P7], ainsi que des exemples, notamment celui étudié dans [P8], et une généralisation étudiée dans [P13]. Enfin, nous proposons une théorie des formes limites a posteriori non-paramétriques d'après [P10, P14].

Table des matières

Introduction

Overview

This manuscript presents a synthesis of my research work over the last few years. It discusses my contributions to the study of the Bayes posterior distribution in statistical models with many or infinitely many parameters, such as nonparametric and semiparametric models.

- Chapter 1 – Lower bounds for posterior rates. We define a simple notion of lower bound rate [P4], which serves several purposes, notably checking whether some rates are sharp, and finding necessary conditions on qualitative aspects of priors. Some examples from [P9], [P15] are presented.
- Chapter 2 – Upper bounds for posterior rates. We describe several rate-results obtained in different contexts, such as rates in geometric frameworks [P11] and rates in sparse models [P9, P15]. We then propose an approach [P12] for studying posterior supremum-norm convergence rates.
- Chapter 3 – Limiting shape of posterior ditributions. The asymptotic shape of posterior distributions is investigated, first from the semiparametric perspective: a result for Gaussian priors is presented [P7], as well as examples [P8] and some generalisations [P13]. Next a nonparametric limiting shape theory is proposed [P10, P14].

The chosen outline does not present the results in chronological order. I started my research on Bayesian procedures via semiparametric models and Bernstein–von Mises theorems. It turned out that to solve certain questions in the semiparametric context, several difficulties linked rather to nonparametrics and convergence rates had to be overcome. For instance, the lower bound concept presented in Chapter 1 is a side-product of investigations on using Gaussian priors in a semiparametric translation model, which is quite unexpected. This concept was itself later of use to prove that posterior rates presented in Chapter 2 are sharp for some combinations of model and prior. Also, another fruitful and apparently unrelated application of studies around Bernstein–von Mises results from Chapter 3 is presented in Chapter 2. To define a notion of convergence to an infinite dimensional limiting object, convergence rates in some unusual loss functions were of use, and the method we introduce has later led to new methods in terms of other distances of interest.

I do not present here the results following my PhD thesis [P1, P2, P3, P5, P6], which have semiparametrics as common denominator. Of course, these have had and have an important influence on my research, notably via a semiparametric 'perspective'.

The following short introduction is not ment as a comprehensive overview of the background, nor as an exhaustive bibliographical account of the state-of-the-art results. It is simply aimed at presenting in a (hopefully) gentle way the main tools used in the sequel.

Statistical models

Let $\mathcal{X}^{(n)}$ be a metric space equipped with a σ-algebra $\mathcal{A}^{(n)}$, where n is an integer corresponding to the amount of information, for instance the number of available observations. A *statistical experiment* is a collection of probability measures $\{P_\eta^{(n)}\}$ indexed by a parameter η in some measurable space $\mathcal{H}$ to be specified. We use the generic notation $X^{(n)}$ to denote observations from this experiment. From now on $n \geq 1$ is a given integer, which we may let tend to ∞.

Nonparametric models

In all following examples, the statistical model is indexed by, either an infinite dimensional parameter, typically a real-valued function over some space denoted by f, or a *high-dimensional* parameter, denoted by θ or β. In the latter case, the model is parametric for each fixed n but the dimension of the parameter can increase with n. The quantities f, θ, β are unknown and the goal of the statistician is to say something about them after having observed data from the model. Many statistical questions can be classified as belonging to one of the following trilogy: *estimation*, *testing* and *confidence sets*. Here we will be mostly interested in estimation and confidence sets, although testing sometimes plays an important role, notably in the proofs.

Gaussian white noise model. Let $L^2 := L^2([0,1])$ be the space of square integrable functions on $[0,1]$. For $f \in L^2$, dW standard white noise, consider observing

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

Two equivalent data-generating mechanisms are: observing the path $X^{(n)}(x) = \int_0^x f(t)dt + n^{-1/2}W(x)$, where W is standard Brownian motion on $[0,1]$; or, given a collection of orthonormal functions $\{\varphi_k,\ k \geq 1\}$ in L^2 forming a basis of L^2, observing the collection

$$X_k = f_k + \frac{1}{\sqrt{n}}\xi_k, \qquad k \geq 1, \tag{2}$$

where $f_k = \int_0^1 \varphi_k f$ and $\{\xi_k\}_{k\geq 1}$ are independent $\mathcal{N}(0,1)$ variables. This last version of the model is often also called the (infinite) *Gaussian sequence model*.

Density estimation. On the unit interval, the density model consists of observing independent identically distributed (i.i.d.) data

$$X_1, \ldots, X_n \quad i.i.d. \ \sim f, \tag{3}$$

with f a density function on the interval $[0,1]$. In this case the common law of the X_is is the distribution P_f of density f with respect to Lebesgue measure on $[0,1]$. Then $P_f^{(n)}$ is the product measure $\otimes_{i=1}^n P_f$ on $[0,1]^n$. For models with i.i.d. data (and those only) for simplicity in the sequel we denote $P_f^n := P_f^{(n)}$.

Geometric spaces. It is of interest to generalise the previous models to the case where data 'sits' on a geometrical object, say a compact metric space $\mathcal{M}$. One may think of a torus, a sphere, a manifold, or maybe even a discrete structure such as a tree, a graph etc.

The white noise model becomes, on a compact metric space $\mathcal{M}$

$$dX^{(n)}(x) = f(x)dx + \frac{1}{\sqrt{n}} dZ(x), \quad x \in \mathcal{M}, \tag{4}$$

where f is square-integrable on $\mathcal{M}$ and Z is a white noise on $\mathcal{M}$.

The density estimation model on a manifold $\mathcal{M}$ consists in observing

$$X_1, \ldots, X_n \quad i.i.d. \ \sim f, \tag{5}$$

where X_i are $\mathcal{M}$-valued random variables with positive density function f on $\mathcal{M}$.

NEEDLES AND STRAW IN A HAYSTACK. Suppose that we observe

$$X_i = \theta_i + \varepsilon_i, \quad i = 1, \dots, n, \tag{6}$$

for independent standard normal random variables ε_i and an unknown vector of means $\theta = (\theta_1, \dots, \theta_n)$. Suppose θ is *sparse* in that it belongs to the class of *nearly black vectors*

$$\ell_0[p_n] = \{\theta \in \mathbb{R}^n : \ \mathrm{Card}\{i : \theta_i \neq 0\} \leq p_n\}\,. \tag{7}$$

Here p_n is a given number, which in theoretical investigations is typically assumed to be $o(n)$, as $n \to \infty$. Sparsity may also mean that many means are small, but possibly not exactly zero.

HIGH-DIMENSIONAL LINEAR REGRESSION. Consider estimation of a parameter $\beta \in \mathbb{R}^p$ in the linear regression model

$$Y = X\beta + \varepsilon \tag{8}$$

where X is a given, deterministic $(n \times p)$ matrix, and ϵ is an n-variate standard normal vector. As for the previous model, we are interested in the *sparse* setup, where $n \leq p$, and possibly $n \ll p$, and most of the coefficients β_i of the parameter vector are zero, or close to zero. Model (6) is a special case with X the identity matrix of size n. Model (8) shares some features with this special case, but is different in that it must take account of the noninvertibility of X and its interplay with the sparsity assumption, and does not allow a factorization of the model along the coordinate axes.

There are of course further links between all the above models. For instance, the study of the sparse Gaussian sequence model (6) is related to the study of certain sparse nonparametric classes, namely sparse Besov spaces. We shall discuss further similarities along the way.

Semiparametric models

Slightly informally and broadly speaking, one may define semiparametric models as those models where the parameter of interest is a (often, but not necessarily) finite-dimensional *aspect* of the parameter η of the model, where η is typically infinite-dimensional. We give two first examples.

SEPARATED SEMIPARAMETRIC MODELS. A model $\{P_\eta^{(n)},\ \eta = (\theta, f),\ \theta \in \Theta,\ f \in \mathcal{F}\}$, where Θ is a subset of $\mathbb{R}^k$ for given $k \geq 1$ and $\mathcal{F}$ a nonparametric set, is called *separated* semiparametric model. The full parameter η is a pair (θ, f), with θ called parameter *of interest* and f *nuisance* parameter. Despite the terminology, this does not exclude f to be of interest too.

For instance, the following is called shift or *translation model*: one observes sample paths of the process $X^{(n)}$ such that, for W standard Brownian motion,

$$dX^{(n)}(t) = f(t - \theta)dt + \frac{1}{\sqrt{n}}dW(t), \quad t \in [-1/2, 1/2], \tag{9}$$

where the unknown function f is *symmetric* (that is $f(-x) = f(x)$ for all x), smooth and say 1-periodic, and the unknown parameter of interest θ is the center of symmetry of the *signal* $f(\cdot - \theta)$. This model has a very specific property: estimation of θ can be done as efficiently as in the parametric case where f would be known, at least asymptotically. This is called a model without loss of information. This and other models with loss of information, such as the famous Cox's proportional hazards model, will be more formally introduced in Chapter 3.

FUNCTIONALS. More generally, given a model $\{P_\eta^{(n)},\ \eta \in \mathcal{H}\}$, one may be interested in estimating a function $\psi(\eta)$ of the parameter η, for some function ψ on $\mathcal{H}$. For instance, in the density model (3), one may consider the estimation of linear functionals of the density $\psi(f) = \int_0^1 a(u)f(u)du$, for $a(\cdot)$ a bounded measurable function on $[0, 1]$.

The Bayesian paradigm

The Bayesian approach. Given a statistical model $\{P_\eta^{(n)},\ \eta \in \mathcal{H}\}$ and observations $X^{(n)}$, a Bayesian defines a model by attributing a probability distribution to the pair $(X^{(n)}, \eta)$. The distribution $P_\eta^{(n)}$ is viewed as the conditional law $X^{(n)} \,|\, \eta$. A distribution on $(X^{(n)}, \eta)$ is now fully specified by attributing a law to the variable η, called *a priori*, or *prior* distribution. The estimator of η, in the Bayesian sense, is then the conditional distribution $\eta \,|\, X^{(n)}$, called *a posteriori* or *posterior*. This is a data-dependent probability *measure* on $\mathcal{H}$. The prior law is often denoted $\Pi(\cdot)$ or simply Π, and the corresponding posterior $\Pi(\cdot \,|\, X^{(n)})$.

Informally, in order to estimate a given parameter η, one first makes it random by giving oneself a prior distribution. Next one updates this a priori knowledge by conditioning on the observed data, obtaining the posterior distribution. Of course, many choices of prior are in principle possible, and one can expect this choice to have an important impact on how the posterior distribution looks like. As the number of observations grows however, one may expect that the influence of the prior becomes less and less eventually. We shall see through all three next Chapters that in high dimensional models this typically cannot be achieved without special care.

Frequentist analysis of posterior distributions. Slightly anticipating the convergence Section below, a natural way to assess convergence properties of $\Pi(\cdot \,|\, X^{(n)})$ is to make the frequentist assumption that the data actually *is* generated from $P_{\eta_0}^{(n)}$, for some 'true' η_0 in $\mathcal{H}$.

Why Bayesian estimators ? Often, priors have a natural probabilistic interpretation and insights from the construction of various stochastic processes in probability theory can be helpful. Additional 'smoothing' parameter may themselves get a prior, thus leading to natural constructions of priors via hierarchies.

Also, as the posterior is a measure, it has both a 'location' and a 'spread'. Quantifying this spread naturally leads to defining so-called credible sets (we define them below), which under some conditions can be shown to be confidence sets. Hence in principle the Bayesian paradigm can help the statistician in solving both 'estimation' and 'confidence set construction' tasks simultaneously. Of course proving that the previous steps are legitimate is not always an easy task, especially in high dimensional models.

There are other attractive aspects of the Bayesian approach that we do not discuss here: for instance the fact that there are natural priors corresponding to exchangeable data, as developed among others by the Italian school after de Finetti.

From the practical perspective, implementation methods of posterior distributions based on e.g. Markov Chain Monte Carlo techniques have been very much developed since the mid-90's. This in turn leads to the need of developing theoretical tools to determine the convergence properties of the corresponding estimators.

Nonparametric priors, examples

Maybe the most natural idea to build a prior on a nonparametric object such as a function is to decompose the object into simple, finite-dimensional, 'pieces'. Next put a prior distribution on each piece and finally 'combine' the pieces together to form a prior on the whole object. Let us give some examples for functions on the interval $[0, 1]$.

If $\eta = f$ is an element of $L^2[0,1]$, one may first decompose f into its coefficients $\{f_k\}$ onto an orthonormal basis $\{\varphi_k\}$ of $L^2[0,1]$, such as the Fourier basis, a wavelet basis etc. Next, draw real-valued independent variables as prior on each coefficient. One constraint appears: one has to choose the individual laws so that the so-formed function f almost surely belongs to L^2. This can be easily accommodated by taking the coordinate variances going to 0 fast enough. This leads us to set

$$f(\cdot) = \sum_{k=1}^{\infty} \sigma_k A_k \varphi_k(\cdot), \tag{10}$$

where $\{A_k\}$ is a sample from a centered distribution with finite second moment and $\{\sigma_k\}$ is a deterministic sequence in ℓ^2. This gives ample room for choosing sequences $\{\sigma_k\}$ and the common law of the $\{A_k\}$. And, anticipating slightly, the variety of behaviours of the corresponding posterior distributions in such simple models as white noise (1) is already quite broad.

GAUSSIAN PROCESS PRIORS. Specialising the previous construction to Gaussian distributions for the law of A_k, one obtains particular instances of Gaussian processes taking values in $L^2[0,1]$.

Another way of building a, say centered, Gaussian process prior (Z_t) on the interval $[0,1]$ is via a covariance kernel $K(s,t) = \mathbb{E}(Z_s Z_t), (s,t) \in [0,1]^2$. The choice $K(s,t) = s \wedge t$ gives Brownian motion. The choice $K(s,t) = e^{-(s-t)^2}$ corresponds to the so-called squared-exponential Gaussian process, whose paths can be seen to be much smoother than those of Brownian motion.

Starting from Brownian motion, one can define a new Gaussian process by integrating it a fractional number $(\alpha - 1/2)$ of times. This leads to the so-called Riemann-Liouville process of parameter $\alpha > 0$

$$R_t^\alpha = \int_0^t (t-s)^{\alpha-1/2} dW(s), \quad t \in [0,1], \tag{11}$$

where W is standard Brownian motion. One further defines a *Riemann-Liouville type process* (RL-type process) as, for $\underline{\alpha}$ the largest integer smaller than α,

$$X_t^\alpha = R_t^\alpha + \sum_{k=0}^{\underline{\alpha}+1} Z_k t^k, \quad t \in [0,1], \tag{12}$$

where $Z_0, \ldots, Z_{\underline{\alpha}+1}, R_t$ are independent, Z_i is standard normal and R_t^α is the Riemann-Liouville process of parameter α. If $\alpha = 1/2$ then R_t^α is simply standard Brownian motion and if $\{\alpha\} = 1/2$, with $\{\alpha\} \in [0,1)$ the fractional part of α, then R_t^α is a k-fold integrated Brownian motion. The reason for adding the polynomial part to form the RL-type process X_t^α is that the support in $\mathcal{C}^0[0,1]$ of X_t^α is the whole space $\mathcal{C}^0[0,1]$, see [100], Section 4.

Yet another, slightly more abstract, way of building a Gaussian prior is by defining it as a Gaussian measure on a separable Banach space $\mathbb{B}$ (e.g. $L^2[0,1]$, $\mathcal{C}^0[0,1]$ etc.) with a norm denoted $\|\cdot\|_{\mathbb{B}}$ or simply $\|\cdot\|$ if no confusion can arise. It can be shown that, in general, this construction coincides with the one starting from a covariance kernel as above. We refer to [102] for a comprehensive review.

PRIORS ON DENSITY FUNCTIONS. Now consider the question of building a prior distribution on a density f on the interval $[0,1]$. A difficulty is the presence of two constraints on f, that is $f \geq 0$ and $\int_0^1 f = 1$, which prevents the direct use of a prior such as (10). We briefly present some approaches. Although arguably not the first to have been considered historically, a simple and natural approach consists in applying a transformation to a given function on $[0,1]$ to make it a density. Leonard (1978) [76] and Lenk [75] considered the use of an exponential link function. Given a, say, continuous function w on $[0,1]$, consider the mapping $w \to p_w$ defined by

$$p_w(s) = \frac{e^{w(s)}}{\int_0^1 e^{w(u)} du}, \quad s \in [0,1]. \tag{13}$$

Now any prior on continuous functions, such as a random series expansion (10) or a Gaussian process prior on $[0,1]$ as before, gives rise to a prior on densities by taking the image measure under the transform (13).

A different yet perhaps more 'canonical' approach is to build the random density directly via the construction of a random probability measure on $[0,1]$, absolutely continuous with respect to Lebesgue measure. This connects this question to the central topic of construction of random measures. A landmark progress in that area was the construction of the Dirichlet process by Ferguson (1973) [40]. In terms of density estimation however, samples from the Dirichlet process cannot be used directly since the corresponding random measure is discrete. However, the Dirichlet

process turns out to be a particular case of some more general random structures: so called tail-free processes, which where introduced by Freedman (1963) [42] and Fabius (1964) [39]. For well-chosen parameters, the so-obtained random probability measures have a density. This way one obtains as particular case the Pólya tree processes [81], [67].

Other ways to build random densities include random histograms, that we shall consider as an example in Chapter 2 and 3, random kernel mixtures such as Bernstein polynomials [86], Beta mixtures [90], location scale mixtures [59] etc.

PRIORS IN SEMIPARAMETRIC MODELS. In a separated semiparametric model $\{\mathcal{P}_{\theta,f}\}$, a natural way to build a prior on the pair (θ, f) is simply via a product prior $\pi_\theta \otimes \pi_f$ on each coordinate.

Convergence of the posterior distribution

CONSISTENCY. Suppose the data $X^{(n)}$ we observe is effectively generated from one fixed element $P_{\eta_0}^{(n)}$ of the collection of distributions $\mathcal{P} = \{P_\eta^{(n)},\ \eta \in \mathcal{H}\}$. In this case η_0 is called the 'true' η. It is then natural to expect that the posterior distribution $\Pi[\cdot \,|\, X^{(n)}]$, a random probability measure, concentrates around the true η_0 as the number n increases.

In what follows the space $\mathcal{H}$ of parameters is a (first countable) topological space, and Π a prior on the Borel σ-algebra $\mathcal{T}$ of $\mathcal{H}$.

Definition 0.1 *The posterior distribution $\Pi[\cdot \,|\, X^{(n)}]$ is said to be (weakly) consistent at $\eta_0 \in \mathcal{H}$ if, for every neighborhood $\mathcal{V}$ of η_0, the posterior mass of its complement tends to 0 in $P_{\eta_0}^{(n)}$-probability, as $n \to \infty$. That is,*

$$\Pi[\mathcal{V}^c \,|\, X^{(n)}] \to^{P_{\eta_0}^{(n)}} 0, \quad (n \to \infty). \tag{14}$$

A famous result by Doob (1948) states that under surprisingly mild conditions, the posterior distribution is consistent for Π-almost all values of the parameter η, see [22, 36]. However, this is consistency only *from the prior's perspective* and does not guarantee consistency at a *fixed, given* η_0, which is required by the previous definition. A pioneer contribution in that direction was the paper by Schwartz [92], who gave sufficient conditions for consistency in the i.i.d. case in terms of 1) the amount of mass the prior distribution puts on a fixed Kullback-Leibler type neighborhood of the true distribution η_0

$$B_{KL}(\eta_0, \varepsilon) := \Big\{\eta, \quad \int \log\left(p_{\eta_0}/p_\eta\right) p_{\eta_0} d\mu \le \varepsilon, \quad \int \log^2\left(p_{\eta_0}/p_\eta\right) p_{\eta_0} \le \varepsilon\Big\}, \tag{15}$$

where $\varepsilon > 0$, and 2) the existence of exponential tests of the simple null hypothesis $H_0 = \{\eta = \eta_0\}$ versus a composite hypothesis of the type $H_1 = \{\eta \in \mathcal{V}^c\}$. Schwartz' results were later extended to cover other examples of priors, as in [6] and [47].

In the seventies, focus seems to have partly shifted to other aspects of Bayes estimation, with important developments on new priors for e.g. density estimation, with the introduction of classes of random probability measures such as the Dirichlet process [40], tail-free processes and Pólya trees. Nevertheless, we mention the fundamental work of Le Cam (1973) [70], which, while focusing on parametric models, set the ground for future works on rates, with the use of empirical processes techniques such as peeling combined with statistical tests with exponential power.

More surprises were to come, as it was realised that, although the Bayesian approach in principle provides a huge freedom through the number of possible prior distributions one can imagine to build, 'most priors', however, 'do not work', see e.g. Freedman (1963) [42]. 'Most' is, however, in a topological sense of sets with large measure. Maybe more problematic is the fact that even innocent looking priors may yield inconsistency. In a famous contribution Diaconis and Freedman (1986) [32] give an example of natural –yet inconsistent– prior in the semiparametric symmetric location problem. Although it illustrates a somewhat different phenomenon –semiparametric bias rather than inconsistency– we obtain another at first sight surprising result in the world of rates of convergence in Chapter 3, where two nearly identical nonparametric priors are shown to lead

to completely different convergence rates for the parameter of interest.

CONVERGENCE RATE. Let now the parameter space $\mathcal{H}$ be equipped with a metric d and let ε_n be a sequence going to 0 as $n \to \infty$.

Definition 0.2 *The posterior distribution is said to converge at rate at least ε_n towards η_0 in terms of a distance d on $\mathcal{H}$ if, for some $M > 0$,*

$$\Pi[\eta : \ d(\eta, \eta_0) \leq M\varepsilon_n \,|\, X^{(n)}] \to^{P^{(n)}_{\eta_0}} 1, \quad (n \to \infty). \tag{16}$$

Some comments are in order. In some cases, one may replace the fixed constant M by an arbitrary sequence $M = M_n \to \infty$. This is typically necessary when $\mathcal{H}$ is finite dimensional and the statistical model $\mathcal{P}$ is smooth, see below. For the nonparametric setting where η is infinite dimensional, this is often not necessary, as the measure will typically concentrate on the boundaries of balls, see Chapter 1. We also note that sometimes, depending on the chosen normalisation for the convergence rate, one may have rates going to ∞.

Pioneering general rates of convergence results for posterior distributions were obtained at the end of the nineties by Ghosal, Ghosh and van der Vaart [48] and Shen and Wasserman [95]. Qualitative, model-free assumptions are given that guarantee posterior convergence at a certain rate in i.i.d. observations models, in terms of specific distances. A precise statement following [48] is given below. Such results were later extended to non-i.i.d. models by Ghosal and van der Vaart [49], in such different contexts as regression, Markov chain, time series data etc.

POINT ESTIMATORS. If the posterior distribution converges at rate $\varepsilon_n \downarrow 0$ towards η_0 as in Definition 0.2, then there exists a point estimator $\hat{\eta}_n$ such that $d(\hat{\eta}_n, \eta_0)/\varepsilon_n$ is bounded in probability under $P^{(n)}_{\eta_0}$. An example of such a point estimator $\hat{\eta}_n$ is the center of the smallest ball that contains at least half of the posterior mass.

BAYES FORMULA. It is now time to state the Reverend's celebrated formula, that appeared posthumously in a 1763's essay [7] in a specific context with a binomial likelihood. It is in fact valid much more generally assuming some measurability and a dominated likelihood assumption. It consists of an expression of the posterior distribution, the condional law of η given the data in the Bayesian framework, in terms of a ratio of integrated likelihoods.

In all what follows, we assume that the considered statistical experiment is *dominated*. There exists a σ-finite measure $\mu^{(n)}$ such that $P^{(n)}_\eta$ is for any η in $\mathcal{H}$ absolutely continuous with respect to $\mu^{(n)}$, with corresponding density denoted $p^{(n)}_\eta$. We further assume that, for $\mathcal{T}$ a σ-field on $\mathcal{H}$, the mapping $(x^{(n)}, \eta) \to p^{(n)}_\eta(x^{(n)})$ is jointly measurable relative to $\mathcal{A}^{(n)} \otimes \mathcal{T}$.

For any measurable set B in $\mathcal{T}$, Bayes' theorem states that

$$\Pi(B \,|\, X^{(n)}) = \frac{\int_B p^{(n)}_\eta(X^{(n)}) d\Pi(\eta)}{\int p^{(n)}_\eta(X^{(n)}) d\Pi(\eta)}. \tag{17}$$

All our results are in the above setting where Bayes' formula is valid. Nevertheless, we note that in some relevant situations for Bayesian nonparametrics, Bayes' formula may not apply. A prototypical example is the i.i.d. sampling setting with a Dirichlet process prior on the distribution function F. More generally, completely random measures [64] form a class of priors involving Dirac masses at random locations and Bayes' formula may not apply.

A THEOREM. We follow [48] and for simplicity state the result for i.i.d. observations. That is, $\mathcal{P} = \{P^n_\eta, \ \eta \in \mathcal{H}\}$, with $P^n_\eta = \otimes^n_{i=1} P_\eta$ and $dP_\eta = p_\eta d\mu$. Let $d = h$ be the Hellinger distance between densities

$$h(\eta_1, \eta_2) := h(P_{\eta_1}, P_{\eta_2}) = \left(\int (p_{\eta_1} - p_{\eta_2})^2 d\mu \right)^{1/2}. \tag{18}$$

For a subset $\mathcal{G} \subset \mathcal{H}$, let $N(\varepsilon, \mathcal{G}, d)$ denote the ε-covering number of $\mathcal{G}$ with respect to d, that is the minimal number of d-balls of radius ε needed to cover $\mathcal{G}$.

Theorem 0.1 *If there exist $\mathcal{H}_n \subset \mathcal{H}$ and $c > 0$ such that, with B_{KL} given in (15),*

$$\log N(\varepsilon_n, \mathcal{H}_n, h) \leq n\varepsilon_n^2 \qquad \textit{entropy}$$

$$\Pi(\mathcal{H} \backslash \mathcal{H}_n) \leq e^{-(c+4)n\varepsilon_n^2} \qquad \textit{remaining mass}$$

$$\Pi(B_{KL}(\eta_0, \varepsilon_n)) \geq e^{-cn\varepsilon_n^2}, \qquad \textit{prior mass}$$

then for $M > 0$ large enough, as $n \to \infty$,

$$\Pi(\eta : \ h(\eta, \eta_0) \leq M\varepsilon_n \,|\, X^{(n)}) \to^{P_{\eta_0}^n} 1.$$

Let us briefly comment on the result. One can exclude a part $\mathcal{H}_n^c$ of the parameter set $\mathcal{H}$ provided its prior mass is sufficiently small. Next the prior should put enough mass on a Kullback-Leibler type neighborhood of the true η_0. The neighborhood shrinks to 0 at the target rate. Finally, the first condition ensures the set $\mathcal{H}_n$ is not 'too large'. It can be replaced by a testing condition.

Testing is an essential ingredient of the proof of Theorem 0.1. The idea is first to construct a test of a point η_0 versus a generic Hellinger-ball centered at another point η_1, with errors of first and second kind bounded exponentially in terms of $nh^2(P_{\eta_0}, P_{\eta_1})$. This is always possible for the Hellinger distance due to general results of Birgé (1984) [15] and Le Cam (1986) [71] for testing convex sets, see also [16] for a recent perspective. Next, to build a test of a point versus the *complement* of the ball $\{\eta : \ h(\eta, \eta_0) \leq M\varepsilon_n\}$, one may partition this complement into shells. The shells are themselves covered by balls for which one uses the individual tests point–versus–ball. Finally, one combines the previous tests into a single one. The entropy condition enables to control the overall error of the resulting test. On the other hand, the prior mass condition enables a control from below of the denominator in Bayes' formula. Finally, the remaining mass condition provides some extra flexibility.

Uniformity in the results. Although often not explicitly written to simplify the notation, the previous rate result often holds uniformly over a given class $\mathcal{B}$ of parameters η, such as for instance a Sobolev ball. That is, one typically has $\sup_{\eta_0 \in \mathcal{B}} E_{\eta_0}^{(n)} \Pi(h(\eta, \eta_0) > M\varepsilon_n \,|\, X^{(n)}) \to 0$ (also, since the posterior takes values in $[0, 1]$, its convergence in probability or in expectation are equivalent).

Target rate. Definition 0.2 defines 'a' rate rather than 'the' rate. Generally, for a given prior Π, one looks for ε_n 'as small as possible' such that convergence to 0 in probability still holds. Often, we shall try to find classes of priors so that the corresponding posterior converges at a rate ε_n that is 'optimal' in some sense. A typical benchmark is the minimax rate corresponding to point estimators for a given class of parameters η and a specific loss function specified via a distance d. Making a formal link with minimaxity may sometimes require additional assumptions. As noted above, if ε_n-convergence of the posterior occurs, there is a point estimator converging at rate ε_n in probability. Possibly under additional conditions, this may be strenghthened to a convergence in expectation, uniform over the considered class, thus leading to a minimax point estimator. Another way is to prove a result for the posterior mean, which again may require some extra work and/or assumptions, typically that the posterior mass outside balls of radius ε goes to 0 fast enough with ε. In the present work, and as in [48, 49] and [95], we will typically be content with a (and, though often not written for simplicity, uniform) posterior mass result as in Theorem 0.1.

Extensions, related results. The results hold more generally for distances such that certain tests exists, and for non-i.i.d. data as well, as investigated in Ghosal and van der Vaart [49]. As above, a main tool of the proof consists in building tests of the true parameter versus balls of alternatives. This is possible in a large variety of frameworks, thanks notably to the testing results established by Le Cam [70, 71] and Birgé [13, 14]. It is also possible to apply the result in semiparametric models to get consistency and some (nonparametric-type) rate for the parameter of interest.

Theorem 0.1 is formulated in an asymptotic way, as $n \to \infty$. The proof and statement can often be made non-asymptotic: typically the difference between the expectation in Theorem 0.1 and 1 is less than $Ce^{-cn\varepsilon_n^2}$.

Related results were simultaneously obtained by Shen and Wasserman [95], where the testing/entropy condition is formulated in terms of a control of likelihood ratios, and where a variety of other interesting examples is also studied. In the special case of density estimation, it was also noted by Walker [103] and Walker, Lijoi and Prünster [104] that alternative arguments based on martingale methods under somewhat adapted assumptions lead to the conclusion of Theorem 0.1.

Some methods or estimators are close in spirit to the posterior distribution. For instance, so called pseudo-posterior distributions correspond to formula (17) where the likelihood term is raised to some power $\delta > 0$. Varying this 'temperature' parameter enables to give more weight to data or prior respectively. This is related to families of (pseudo-)posteriors considered in the PAC-Bayesian literature, where the likelihood term in Bayes-formula may be further replaced by an exponentiated negative empirical risk. This is a very important and active area of research, in particular with connections to machine learning. We refer to Catoni (2004) [27] for a theoretical treatment and to Alquier (2013) [2] for an overview of recent developments.

Posterior rates for Gaussian process priors

Theorem 0.1 is based on qualitative assumptions. Its consequences are far-reaching: even in situations where closed-form expressions are far from being available, it may give sharp rate results. This fact was illustrated in a striking way through the results by van der Vaart and van Zanten for Gaussian process priors in [99, 100]. Their results can be summarised as follows: for Gaussian process priors, the prior mass condition can be verified using the so-called concentration function of the Gaussian process, itself linked to its *small ball probability*; on the other hand, Gaussian concentration of measure [74] provides natural candidate sets $\mathcal{H}_n$ that verify the two first conditions of Theorem 0.1. These results serve as important building block for some of our results, and we describe them now in some more detail.

Let the prior be constructed as the law $\mathbb{P}$ of Z, a centered and tight measurable random map in the Banach space $(\mathbb{B}, \|\cdot\|)$. Let $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ be the Reproducing Kernel Hilbert Space (RKHS, see [102]) of Z. We will generally assume that f_0 belongs to the support of the prior in $\mathbb{B}$, which for Gaussian process priors is nothing but the closure of $\mathbb{H}$ in $\mathbb{B}$.

For Gaussian priors an upper-bound for the concentration rate of the posterior distribution can often be obtained in a simple way from the so-called *concentration function* of the Gaussian process. This quantity is defined as follows. For any $\varepsilon > 0$, let

$$\varphi_{f_0}(\varepsilon) = \inf_{h \in \mathbb{H} : \|h - f_0\| < \varepsilon} \|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|Z\| < \varepsilon) \tag{19}$$

Assume that the norm $\|\cdot\|$ on $\mathbb{B}$ is comparable to a metric d appropriate to the statistical problem (often, d is a distance for which certain tests exists, which allows to apply the theory presented in [48]; for instance, in i.i.d. settings, one may choose Hellinger's distance). Here "comparable" means that the ball $\{f \in \mathcal{F},\ \|f - f_0\| \leq \varepsilon_n\}$ should be included in the ball for d around f_0 of radius $c\varepsilon_n$ and also in the Kullback-Leibler neighborhood $B_{KL}(f_0, c\varepsilon_n)$ defined above, for some $c > 0$. Then van der Vaart and van Zanten [100] prove that if $\varepsilon_n \to 0$ satisfies

$$\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2, \tag{20}$$

then the posterior contracts at the rate ε_n for the distance d, in that for large enough $M > 0$, $\Pi(f :\ d(f, f_0) \leq M\varepsilon_n \mid X^{(n)}) \to 1$ in $P_{f_0}^{(n)}$-probability as $n \to \infty$.

These results mean that for Gaussian priors an upper-bound on the posterior rate is obtained as soon as the next two quantities are controlled

$$\varphi_{f_0}^A(\varepsilon) = \inf_{h \in \mathbb{H} : \|h - f_0\| < \varepsilon} \|h\|_{\mathbb{H}}^2, \qquad \varphi^B(\varepsilon) = -\log \mathbb{P}(\|Z\| < \varepsilon). \tag{21}$$

The first term measures how well elements in the RKHS $\mathbb{H}$ of the Gaussian process can approximate the true function in $\mathbb{B}$. Note in particular that if f_0 happens to be in $\mathbb{H}$, this term simply remains bounded. The second term, which does not depend on f_0, is the so-called *small ball probability* of the Gaussian process. Intuitively it can be understood as a measure of 'complexity' of the process. Small ball probabilities have been studied in many papers in the probability literature and precise equivalents as $\varepsilon \to 0$ of $\varphi^B(\varepsilon)$ are available for many classes of Gaussian processes, see for instance [77]. Yet at first sight it is not obvious to see why the concentration function φ_{f_0} should appear in the study of posterior rates. This will be explained in Chapter 1, see Lemma 1.4.

Let us now give an example. In density estimation, if the prior is the law induced by p_W, with W Brownian motion and $w \to p_w$ as in (13), the rate ε_n can be shown to depend on the Hölder regularity β of the true f_0 as follows. If $\beta \geq 1/2$, then ε_n can be chosen equal to $n^{-1/4}$, whereas if $\beta < 1/2$ the rate ε_n must be in $n^{-\beta/2}$ to satisfy (20). Thus, up to constants, the rate is optimal in the minimax sense if $\beta = 1/2$. However, for all other values of β, the obtained (upper-bound-)rate is below the minimax rate which is $n^{-\beta/(2\beta+1)}$. At this point it natural to ask whether the fastest possible rate of concentration for Brownian motion is really the one we just described or if in fact the posterior contracts faster. We study this question in Chapter 1.

The Laplace-Bernstein-von Mises phenomenon

Consider an i.i.d. parametric model $\mathcal{P} = \{P_\theta^n,\ \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$, $k \geq 1$. The next result considers *smooth* models, for which the notions of *Fisher information* $\mathcal{I}_{\theta_0}$ at an interior point $\theta_0 \in \Theta$, and of *efficient estimator*, are well-defined. We refer to van der Vaart (1998) [98] for precise definitions. We shall not elaborate on these here, but only note a few points useful to get some intuition. Smoothness of the model implies that estimation is possible at rate $1/\sqrt{n}$ and that the model at θ_0 is 'asymptotically equivalent' to a Gaussian shift experiment $\{N(h, \mathcal{I}_{\theta_0}^{-1}),\ h \in \mathbb{R}\}$, under rescaling by a factor $\sqrt{n}$. An *efficient* estimator $\hat{\theta}_n$ then converges in distribution, as $n \to \infty$,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \mathcal{I}_{\theta_0}^{-1}).$$

The phenomenon in the next Theorem was observed by Laplace [66], and later studied by Bernstein, von Mises [83] and Le Cam [68, 69] among others. Its first message is that posterior distributions are, in smooth parametric models, asymptotically normal. Also, it exhibits a striking equivalence: the Bayes posterior distribution asymptotically looks like the efficient frequentist limiting distribution, centered at an efficient estimator.

Let π be a prior distribution on Θ and let $\|P(\cdot) - Q(\cdot)\|$ denote the *total variation distance* between the probability measures P and Q.

Theorem 0.2 ([98], Theorem 10.1) *In a parametric model $\mathcal{P}$ as above, assume that*

$$\forall \varepsilon > 0,\ \exists\, \varphi_n \text{ test}, \qquad P_{\theta_0}^n \varphi_n \to 0, \quad \sup_{|\theta - \theta_0| \geq \varepsilon} P_\theta^n (1 - \varphi_n) \to 0 \qquad \textit{Testing}$$

$$\textit{The model is } \text{smooth} \textit{ at } \theta_0 \textit{ and } \mathcal{I}_{\theta_0} > 0 \qquad \textit{Regularity}$$

$$\textit{The prior } \pi \textit{ has a continuous positive density at } \theta_0. \qquad \textit{Prior}$$

Then, as $n \to \infty$, under $P_{\theta_0}^n$, it holds, for $\hat{\theta}$ an efficient estimator of θ_0,

$$\left\| \pi(\cdot \,|\, X^{(n)}) \ - \ N\left(\hat{\theta}_n, \frac{\mathcal{I}_{\theta_0}^{-1}}{n}\right)(\cdot) \right\| \longrightarrow 0, \tag{22}$$

This result is called the Bernstein-von Mises theorem (hereafter BvM). Let us now discuss one important consequence of it. A measurable random set $C_n = C_n(X^{(n)})$ is a *confidence set* for the parameter θ of level $1 - \alpha$ asymptotically if for any θ, it holds $P_\theta^n(\theta \in C_n) \to 1 - \alpha$ as $n \to \infty$. A *credible set* of level $1 - \alpha$ for the posterior distribution $\pi(\cdot \,|\, X^{(n)})$ is a measurable set $\mathcal{C}_n$ such that

$\pi(\mathcal{C}_n \,|\, X^{(n)}) = 1 - \alpha$.

BvM and confident credible sets. Let us consider the random interval having as endpoints the 2.5% and 97.5% percentiles of the posterior distribution $\pi(\cdot \,|\, X^{(n)})$. It is the interval $[A_n, B_n]$ such that

$$\pi((-\infty, A_n) \,|\, X^{(n)}) = 0.025, \qquad \pi((B_n, +\infty) \,|\, X^{(n)}) = 0.025.$$

Note that $[A_n, B_n]$ is accessible in practice as soon as simulation from the posterior is feasible. Now, some simple calculations reveal that the conclusion (22) of the BVM theorem implies that, if q_α denotes the standard normal quantile at level α,

$$[A_n, B_n] = \left[\hat{\theta}_n + \frac{q_{\alpha/2}}{\sqrt{n}\mathcal{I}_{\theta_0}^{1/2}} + o_{P_{\theta_0}^n}(n^{-1/2}) \;,\; \hat{\theta}_n + \frac{q_{1-\alpha/2}}{\sqrt{n}\mathcal{I}_{\theta_0}^{1/2}} + o_{P_{\theta_0}^n}(n^{-1/2}) \right]. \tag{23}$$

Simple verifications reveal that the latter interval contains θ_0 with probability 95% as $n \to \infty$. So $[A_n, B_n]$ is asymptotically a 95%-confidence interval in the frequentist sense. In particular, Bayes and frequentist credible regions asymptotically coincide. An advantage of the Bayes approach is that, to build $[A_n, B_n]$, estimation of $\mathcal{I}_{\theta_0}$ is not required.

CHAPTER 1

Lower bounds for posterior rates

We introduce a concept of lower bound for posterior rates following [P4]. We illustrate its use through a variety of examples, among which some are from [P4], [P9], [P15]. This Chapter also serves as a way to introduce simple versions of several statistical models whose study will be pursued in the following Chapters.

In the introduction, we have reviewed the notion of posterior convergence rate. Roughly speaking, 'the' rate in terms of a distance d is generally thought of as an ε_n as small as possible such that the posterior probability of the d-ball centered at the true f_0 and of radius ε_n still tends to 1 in probability. The tools proposed e.g. in Ghosal, Ghosh and van der Vaart [48] and Shen and Wasserman [95] often allow to obtain an *upper bound* for the posterior rate corresponding to a given prior distribution Π. However, the result does not say whether one could actually do better, that is obtaining an even faster rate. In a variety of situations, the notion of lower bound below enables to prove that the obtained rate is sharp, possibly up to slowly varying factors. Also, the concept is often useful in practical terms, even independently of upper bounds, in that it may allow to investigate *necessary* conditions for the prior to converge at a given target rate.

1.1 A definition and a first result

In this Chapter unless otherwise stated we consider an experiment $(\mathcal{X}^{(n)}, \mathcal{A}^{(n)}, P_f^{(n)})$ with $f \in \mathcal{H}$ an unknown function in a space $\mathcal{H}$ to be specified, and the measurability and domination assumptions stated above.

Definition. Let d be a distance on the parameter space $\mathcal{H}$. Let ζ_n is an arbitrary sequence indexed by n, most of the time going to 0 with n. Given a prior Π on $\mathcal{H}$, in [P4] we propose the following definition.

Definition 1.1 *The rate ζ_n be a* lower bound *for the concentration rate of the posterior distribution $\Pi(\cdot|X^{(n)})$ in terms of the distance d if, as $n \to \infty$,*

$$\Pi(f : \ d(f, f_0) \leq \zeta_n \mid X^{(n)}) \to^{P_{f_0}^{(n)}} 0. \tag{1.1}$$

The definition mainly means that ζ_n is too fast for the posterior measure to capture mass in the ball of radius ζ_n around f_0. Note also that Definition 1.1 is not the mere negation of Definition

0.2 but a stronger requirement. Not only not all the mass belongs to a ball of radius ζ_n around f_0, but asymptotically no mass at all belongs to it.

The definition at first looks surprising, since it seems as if the posterior would not be allowed to actually be close to the true f_0 ! Also, if both an upper-bound as in (16) and a lower bound as in (1.1) hold simulatenously and for a common distance d, it means that the posterior distribution stays in a shell $\{f : C^{-1}\varepsilon_n \leq d(f, f_0) \leq C\varepsilon_n\}$, as if there would not nearly be enough space for the mass to stay ... This actually should not be surprising in high or infinite dimensional contexts, and is related to the concentration of measure phenomenon. Indeed, we shall see in the sequel some examples where this shell-behaviour is typical, for infinite-dimensional f's.

On the other hand, the definition is still meaningful for finite dimensional models, where one can deduce from the BvM phenomenon, if it holds at standard parametric rate $1/\sqrt{n}$ as in (0.2), that all posterior mass concentrates on the euclidian shell $\{\theta : M_n^{-1}/\sqrt{n} \leq \|\theta - \theta_0\| \leq M_n/\sqrt{n}\}$ asymptotically, for an arbitrary $M_n \to \infty$. Note in that case the factor M_n *going to* ∞ as opposed to the *fixed* constant C in the former paragraph.

A POSSIBLE STRATEGY. A simple, yet, as it turns out, quite effective method to prove a lower bound rate appeals to a Lemma due to Barron [5]. Recall the notation B_{KL} from (15).

Lemma 1.1 *Let α_n be a sequence such that $n\alpha_n^2 \to \infty$. Let B_n be a measurable set in $\mathcal{H}$. Suppose*

$$\Pi(B_n)/\Pi(B_{KL}(f_0, \alpha_n)) \leq e^{-2n\alpha_n^2}. \tag{1.2}$$

Then $\Pi(B_n \mid X^{(n)}) \to 0$ in $P_{f_0}^{(n)}$-probability as $n \to \infty$.

So, if an event has an exponentially small *prior* mass compared to the prior mass of a KL-neighborhood of the true η_0, the posterior mass of such an event is also small. Note that this Lemma is already put to good use in establishing upper bounds for posterior rates since it enables to show that one can restrict the study to *sieve*-type sets $\mathcal{H}_n$ provided their complement have sufficiently small prior mass.

This immediately yields the next key Lemma.

Lemma 1.2 *Let α_n be as in Lemma 1.1 and suppose (1.2) holds true for $B_n = \{f : d(f, f_0) \leq \zeta_n\}$, for an arbitrary sequence ζ_n and d some distance on $\mathcal{H}$. Then as $n \to \infty$,*

$$\Pi[f : d(f, f_0) \leq \zeta_n \,|\, X^{(n)}] \to^{P_{f_0}^{(n)}} 0.$$

Controlling a ratio of prior masses of balls (in some cases the KL-type neighborhood can be shown to contain a ball for some distance, otherwise one may still informally see the KL-neighborhood as being of 'ball-type', though the KL-divergence is not a distance) is enough to obtain a lower bound result. We show how this principle applies to a variety of examples.

1.2 Example: Sparsity

Consider the Gaussian sequence model with a sparsity assumption (6), that is, 'finding needles among straw in a haystack', and suppose one wants to do estimation in this model in a Bayesian way. How does one choose the prior distribution ?

Before all, let us see how Lemma 1.2 particularises to the sequence model.

Lemma 1.3 *In model (6), let Π be a prior on θ. We have $P_{\theta_0}^{(n)}\Pi\big(\theta : \|\theta - \theta_0\| < s_n \,|\, X\big) \to 0$, for any s_n for which there exist r_n such that*

$$\frac{\Pi\big(\theta : \|\theta - \theta_0\| < s_n\big)}{\Pi\big(\theta : \|\theta - \theta_0\| < r_n\big)} = o(e^{-r_n^2}).$$

The following motivates the introduction of the so-called 'sparse' or 'Spike and Slab' prior, whose properties we study in detail in Chapter 2. For simplicity of exposition we restrict to the sparse sequence model (6) but similar remarks apply in the linear regression model (8).

Discarding straw, Dirac mass priors and the 'Spike'.

A simple prior Π_λ in the sparse sequence model samples the coordinates of θ independently

$$\Pi_\lambda \sim \otimes_{i=1}^n \mathcal{E}(\lambda), \qquad \lambda > 0, \tag{1.3}$$

where $\mathcal{E}(\lambda)$ denotes the Laplace (double-exponential) prior with scale parameter λ^{-1}.

Proposition 1.1 *Consider the sequence model (6) and suppose the true θ_0 is 0. Let Π_λ be defined by (1.3). Let $\lambda = \lambda_n$ such that $\sqrt{n}/\lambda_n \to \infty$. There exists $\delta > 0$ such that, as $n \to \infty$,*

$$\Pi_{\lambda_n}\Big(\theta : \|\theta\|_2 \leq \delta\sqrt{n}\Big(\frac{1}{\lambda_n} \wedge 1\Big) \,|\, Y\Big) \to^{P^{(n)}_{\theta_0=0}} 0.$$

The minimax rate for estimating a vector in the nearly-black class $\ell_0[p_n]$ defined in (7), in terms of the squared-Euclidian loss, is $2p_n \log p_n(1 + o(1))$ as $n \to \infty$, see [34]. The above result from [P15] shows that the prior (1.3) performs much worse than this when the true θ_0 vector is 0, except perhaps if the parameter λ_n is very large, nearly of the order $\sqrt{n}$. But if λ_n is very large, say larger than some even small power of n, one can show – we do not explicitly state this here – that the convergence rate is slow for sparse vectors with large non-zero entries.

Note that the prior (1.3) does not take into account the information that the true vector is *sparse*, that is that most of its coordinates are zero. This suggests a simple modification of (1.3) where at each coordinate, the prior is allowed to pick either a '0' with some probability $1 - \alpha$, or some other non-zero value with probability α, which may be denoted

$$\Pi_{\alpha,g} \sim \otimes_{i=1}^n (1 - \alpha)\delta_0 + \alpha g, \tag{1.4}$$

where each coordinate is a mixture of a Dirac mass at 0 and of a continuous distribution of density g on $\mathbb{R}$, for instance a Laplace density as above. We will show in Chapter 2 that this prior has nice properties, provided α is well chosen (possibly random of course !)

At this point, the reader familiar with sparsity may wonder: but is not prior (1.3) good in that for well-chosen λ it is should be related to the LASSO ?

Discussing the Bayesian interpretation of the LASSO.

The LASSO estimator [97] for θ in the sequence model may be defined as, with $X = (X_1, \ldots, X_n)$,

$$\hat{\theta}^{LASSO}_\lambda = \arg\min_{\theta \in \mathbb{R}^n} \big[\|X - \theta\|_2^2 + 2\lambda\|\theta\|_1\big]. \tag{1.5}$$

This is the posterior mode, or 'maximum a posteriori estimator', corresponding to the prior (1.3). In the sequence model $\hat{\theta}^{LASSO}_\lambda$ can be written explicitly as a soft thresholding estimator.

The LASSO has many desirable properties: it is computationally tractable; it automatically leads to sparse solutions, and, with the standard choice $\lambda = (c \log n)^{1/2}$, it attains the minimax rate over nearly-black classes. Proposition 1.1 shows that for this choice of λ the *full posterior* distribution corresponding to the LASSO puts no mass on balls of radius of the order $\sqrt{n}/(\log n)^{1/2}$, which is substantially bigger than the minimax rate $(s \log n)^{1/2}$, except for extremely dense signals.

Therefore, the full posterior and its mode have in this case completely different behaviours. The sparsity inherent to the LASSO comes from taking the maximum. On the other hand, the corresponding posterior measure itself is nearly nowhere sparse. It faces two conflicting demands: the scaling parameter λ in the Laplace prior must be large in order to shrink coefficients θ_i to zero, but at the same time reasonable so that the Laplace prior can model the nonzero coordinates.

That these conflicting demands do not affect the good behaviour of the LASSO estimators is due to the special geometric, sparsity-inducing form of the posterior mode.

More generally, especially in high-dimensional settings, the posterior measure may behave differently from some of its 'aspects' such as mode, mean, median etc. Another example is given in Chapter 2, for which the full posterior behaves optimally, as opposed to its mean.

FINDING NEEDLES, ENOUGH STRENGTH VIA HEAVY TAILS: THE 'SLAB'.

Consider now a prior of the type (1.4). It is natural to ask whether any density g supported on the whole real line leads to a satisfactory rate, or whether some further conditions arise.

The next result from [P9] shows that product priors with marginal densities proportional to $y \mapsto e^{-|y|^\alpha}$ for some $\alpha > 1$ lead to suboptimal contraction rates for large true vectors θ_0. In Chapter 2, we prove that $\alpha \leq 1$ is compatible with optimal rates.

The theorem applies in particular to the normal distribution. For this prior a problem (only) arises if the parameter vector $\theta_0 =: \theta_0^n$ (recall that in model (6) everything depends on n, including θ, that has n coordinates) has squared-norm larger than the optimal rate:

$$\|\theta_0\|_2^2 = \|\theta_0^n\|_2^2 \gg p_n \log(n/p_n).$$

The posterior then puts no mass on balls of radius a multiple of $\|\theta_0^n\|_2$ around the true parameter. For "small" θ_0^n no problem occurs, because shrinkage to the origin is desirable in that case.

Product priors with marginal density proportional to $y \mapsto e^{-|y|^a}$ give behaviour as the Gaussian prior for every $\alpha \geq 2$. For $\alpha \in (1,2)$ the result is slightly more complicated and involves

$$\rho_{0,\alpha}^n = \left(\frac{\|\theta_0^n\|_\alpha^\alpha}{\|\theta_0^n\|_2^2} \wedge 1 \right) \|\theta_0^n\|_\alpha \, p_n^{1/2-1/\alpha}, \tag{1.6}$$

where $\|\theta\|_\alpha^\alpha = \sum_{i=1}^n |\alpha_i|^\alpha$ is the usual L^α-norm on $\mathbb{R}^n$.

Proposition 1.2 (Necessity of heavy tails) *In model (6), let the prior be as in (1.4), with g a density proportional to $y \mapsto e^{-|y|^\alpha}$ and the prior π_n on dimension satisfies $\pi_n(p_n) \geq \exp(-cp_n \log(n/p_n))$.*

(i) If $\alpha \geq 2$ and $\|\theta_0\|_2^2/\big(p_n \log(n/p_n)\big) \to \infty$, then for sufficiently small $\eta > 0$, as $n \to \infty$,

$$\Pi\big(\, \theta : \|\theta - \theta_0\|_2 \leq \eta \, \|\theta_0\|_2 \,|\, X^{(n)}\big) \to^{P_{\theta_0}^{(n)}} 0.$$

(ii) If $1 < \alpha < 2$ and $(\rho_{0,\alpha}^n)^2/(p_n \log(n/p_n)) \to \infty$, then for sufficiently small $\eta > 0$, as $n \to \infty$,

$$\Pi\big(\, \theta : \|\theta - \theta_0\|_2 \leq \eta \, \rho_{0,\alpha}^n \,|\, X^{(n)}\big) \to^{P_{\theta_0}^{(n)}} 0.$$

Proposition 1.2 shows problematic behaviour of the posterior distribution for signals with large norm $\|\theta_0\|_2$. Instead of using fixed priors on the coordinates, we could make them depend on the sample size, for instance Gaussian priors with variance $v_n \to \infty$, or uniform priors on intervals $[-K_n, K_n]$ with $K_n \to \infty$. Such priors will push the "problematic boundary" towards infinity, but the same reasoning as for the theorem will show that shrinkage remains for (very) large θ_0.

The above results show that g_S needs to have heavy tails. One can also check that the fact that the amount of mass $\pi_n(p_n)$ at the true dimension is large enough is a necessary condition.

We have examined necessary conditions for sparse priors to converge (uniformly) at optimal rate. Of course spike-and-slab priors with Dirac masses at 0 are not the only possible option. Other Bayesian alternatives are e.g. the horseshoe [26] and nonparametric empirical Bayes [57].

1.3 Example: Gaussian process priors

Prior mass and Concentration function. As briefly mentioned in the introduction, for Gaussian priors there are natural fairly tight bounds for the prior mass of certain neighborhoods, which involve the concentration function φ of the Gaussian prior as defined in (19).

Lemma 1.4 *Let Z be a Gaussian process in $(\mathbb{B}, \|\cdot\|)$ with associated RKHS $\mathbb{H}$. Assume that f_0 belongs to the support of Z in $\mathbb{B}$. Then for any $\varepsilon > 0$,*

$$\varphi_{f_0}(\varepsilon) \leq -\log \mathbb{P}(\|Z - f_0\| < \varepsilon) \leq \varphi_{f_0}(\varepsilon/2).$$

The intuition behind the result is as follows: recall from (21) that φ is the sum of two terms, the small ball probability and an approximation term. If f_0 belongs to the RKHS $\mathbb{H}$ of the prior, one can directly apply the famous Cameron-Martin-Girsanov formula, which asserts that shifting a Gaussian measure by elements of $\mathbb{H}$, absolute continuity of measures remains, so one is left with a probability around the zero-function, that is the small ball probability. If f_0 does not belong to $\mathbb{H}$, an extra approximation term of it by elements of $\mathbb{H}$ arises, which results into the second term in (21).

A consequence of the next Lemma is that φ_{f_0} admits an inverse $\varphi_{f_0}^{-1}$.

Lemma 1.5 *Let Z be a non-degenerate centered Gaussian process in $(\mathbb{B}, \|\cdot\|)$. For any f_0 in $\mathbb{B}$, the associated concentration function $\varepsilon \to \varphi_{f_0}(\varepsilon)$ is strictly decreasing and convex on $(0, \infty)$. In particular, it is continuous on $(0, \infty)$.*

Result in terms of the concentration function. Combining Lemmas 1.2 and 1.4 one gets

Theorem 1.1 *Let Z be a Gaussian process with associated distribution Π on the space $\mathcal{H} = (\mathbb{B}, \|\cdot\|)$. Let the data $X^{(n)}$ be generated according to $P_{f_0}^{(n)}$ and assume that f_0 belongs to the support of Π in $\mathbb{B}$. Let $\alpha_n \to 0$ such that $n\alpha_n^2 \to \infty$ and $\Pi(B_{KL}(f_0, \alpha_n)) \geq e^{-cn\alpha_n^2}$, for some $c > 0$. Suppose that $\zeta_n \to 0$ is such that*

$$\varphi_{f_0}(\zeta_n) \geq (2 + c)n\alpha_n^2. \tag{1.7}$$

Then, as $n \to \infty$, we have, in $P_{f_0}^{(n)}$-probability,

$$\Pi(\|f - f_0\| \leq \zeta_n \mid X^{(n)}) \to 0.$$

While for obtaining upper-bounds results, Condition (20) from van der Vaart and van Zanten [100] requires to bound the concentration function f_0 *from above*, the previous result states that it is enough to bound φ_{f_0} *from below* to obtain a lower-bound posterior rate. Further, the result suggests that it may sometimes be possible to precisely determine the rate of the posterior by having the same (up to constants) upper- and lower-bound rate.

Let us discuss this last point informally first. The condition on the neighborhood B_{KL} in Theorem 1.1 is typically satisfied for $\alpha_n = \varepsilon_n$, an upper-bound rate. If neighborhoods B_{KL} contain $\|\cdot\|$-balls, then [100] show that the last condition is true when $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$. So if $\alpha_n = \varepsilon_n$, Condition (1.7) requires $\varphi_{f_0}(\zeta_n) \gtrsim n\varepsilon_n^2$. This means that if $\varphi_{f_0}(ct)$ compares up to constants to $c^\delta \varphi_{f_0}(t)$ for some real δ, then one can take $\zeta_n \approx \varepsilon_n$ and a precise rate of convergence ε_n^* for the posterior is obtained by solving

$$\varphi_{f_0}(\varepsilon_n^*) \approx n{\varepsilon_n^*}^2$$

and $\zeta_n \approx \varepsilon_n \approx \varepsilon_n^*$. This is of course somewhat informal, since it requires in particular that $\varphi_{f_0}(c\varepsilon_n)$ compares to $\varphi_{f_0}(\varepsilon_n)$. However this can be shown to be the case for a variety of functions f_0 and Gaussian process priors.

Illustration: Gaussian white noise model. First, one can specialise Theorem 1.1 to the white noise model: the next result is a consequence of Theorem 3.4 in [100] for the upper-bound and of Theorem 1.1 for the lower bound.

Theorem 1.2 *Suppose the data is generated according to (1). Let the prior Π be a Gaussian process prior on L^2. Suppose f_0 belongs to the support of Π in L^2. Let ε_n and ζ_n be such that*

$$\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2 \quad \text{and} \quad \zeta_n \leq \varphi_{f_0}^{-1}(9n\varepsilon_n^2).$$

Then for M large enough, as $n \to \infty$,

$$\Pi(\zeta_n \leq \|f - f_0\|_2 \leq M\varepsilon_n \mid X^{(n)}) \to^{P_{f_0}^{(n)}} 1.$$

Let us consider series priors such as (10), now with Gaussian distributions on the coefficients,

$$\Pi = \Pi_\alpha : \quad f(\cdot) \sim \sum_{k=1}^{\infty} k^{-\frac{1}{2}-\alpha} \alpha_k \varepsilon_k(\cdot), \tag{1.8}$$

where α_k are independent standard normal random variables, $\{\varepsilon_k\}$ is a (smooth enough) orthonormal basis of L^2 and $\alpha > 0$. The choice $\sigma_k = k^{-1/2-\alpha}$ for the standard deviations on the coefficients is especially natural if f_0 happens to belong to a Sobolev ball, which we shall assume below. We note that the prior (1.8) stands at the 'boundary' of an α-Sobolev space. A realisation of the prior (1.8) does not belong to a Sobolev space of order α (because the harmonic series diverges) but almost surely belongs to Sobolev spaces of order $\gamma < \alpha$. Such an object is seen as having 'regularity' or 'smoothness' α. Zhao [108] proved upper-bound rates for the posterior-mean for these priors, and discusses spaces the prior and posterior belong to. Her result is revisited by Shen and Wasserman [95] who prove a result for the full posterior. Also, Belitser and Ghosal [8] consider putting a prior on the parameter α and derive adaptive upper-bound posterior rates when the regularity parameter α belongs to a grid.

The prior $\Pi = \Pi_\alpha$ from (1.8) defines a random element in $\mathbb{B} = L^2[0,1]$. Theorem 1.2 tells us that upper and lower-bound rates can be determined via inequalities on the concentration function φ_{f_0} of the prior Π_α at a given f_0. The RKHS $\mathbb{H}^\alpha$ of Π_α in $\mathbb{B}$ can be shown to be the space

$$\mathbb{H}^\alpha = \Big\{ \sum_{k\geq 1} h_k \sigma_k \varepsilon_k,\ (h_k)_{k\geq 1} \in \ell^2 \Big\}, \quad \| \sum_{k\geq 1} h_k \sigma_k \varepsilon_k \|_{\mathbb{H}_\alpha}^2 = \sum_{k\geq 1} h_k^2$$

see van der Vaart and van Zanten [102], Theorem 4.2. The support of the prior in L^2 is L^2 itself, using that for Gaussian priors the closure of the RKHS in the $\mathbb{B}$-norm coincides with the support of the prior in $\mathbb{B}$. For the prior Π_α, the small ball probability is a well-studied quantity, see Kuelbs and Li [65], who show that as $\varepsilon \to 0$,

$$-\log \Pi_\alpha(\|f\|_2 < \varepsilon) \asymp \varepsilon^{-1/\alpha}.$$

Also, one can study the approximation term in φ_{f_0} using that $\mathbb{H}^\alpha$ is actually itself a Sobolev space.

The Sobolev ball $\mathcal{F}_{\beta,L}$ of order $\beta > 0$ and radius $L > 0$ is, for a smooth enough basis $\{\varepsilon_k\}$, the set $\mathcal{F}_{\beta,L} = \{f \in L^2,\ \sum_{k\geq 1} k^{2\beta} \langle f, \varepsilon_k \rangle_2^2 \leq L^2\}$. Denote, for positive α, β,

$$r_n^{\alpha,\beta} = n^{-(\alpha\wedge\beta)/(2\alpha+1)}. \tag{1.9}$$

The next statement reveals that the posterior corresponding to the prior Π_α achieves the rate $r_n^{\alpha,\beta}$ and that this rate cannot be improved in general. The upper-bounds essentially follow from the general results in [100], although the specific application to the priors Π_α (with a proof using general principles; [8] also get upper-bound rates but using explicit expressions) is new. Yet, the main point of the result is to show that the rate is sharp via providing a lower-bound rate.

Proposition 1.3 *Let the prior $\Pi = \Pi_\alpha$ be defined by (1.8), for some $\alpha > 0$. Let f_0 be in $\mathcal{F}_{\beta,L}$, for some $\beta, L > 0$. Let the rate $r_n^{\alpha,\beta}$ be defined by (1.9). Then, the upper-bound rate ε_n in Theorem 1.2 can be chosen such that $\varepsilon_n \lesssim r_n^{\alpha,\beta}$.*

- *If $\alpha \leq \beta$, then the lower-bound rate ζ_n in Theorem 1.2 can be taken such that $\zeta_n \gtrsim r_n^{\alpha,\beta}$ and, for M large enough,*

$$\Pi(M^{-1} r_n^{\alpha,\beta} \leq \|f - f_0\|_2 \leq M r_n^{\alpha,\beta} \mid X^{(n)}) \to^{P_{f_0}^{(n)}} 1.$$

- *If $\beta < \alpha$, there exists f_0 in $\mathcal{F}_{\beta,L}$ such that, for $p > 1 + \beta/2$ and M large enough,*

$$\Pi(r_n^{\alpha,\beta} \log^{-p} n \leq \|f - f_0\|_2 \leq M r_n^{\alpha,\beta} \mid X^{(n)}) \to^{P_{f_0}^{(n)}} 1.$$

There are two regimes. If the prior is as smooth or 'rougher' than the true function, that is $\alpha \leq \beta$, a case called *undersmoothing*, then the shell behaviour mentioned above occurs and the posterior rate remains determined, equal to $n^{-\alpha/(2\alpha+1)}$ up to constants. One can see from the proof of the Proposition that this corresponds to the case where the small ball probability part of the concentration function φ_{f_0} dominates over the RKHS-approximation part. If $\alpha > \beta$, more information on the true function f_0, in terms of the behaviour of the sequence of coefficients $\langle f_0, \varepsilon_k \rangle_2$, is needed to evaluate the RKHS-approximation term. This term may behave differently depending on f_0. For f_0's so that a control *from below* of the basis-coefficients is available, one may obtain a lower bound for the approximation term in terms of a power of n^{-1} and in turn a corresponding lower bound rate result. Such special f_0's are those 'at the boundary' of the Sobolev ball $\mathcal{F}_{\beta,L}$. This is how the specific "worst-case" function f_0 appearing in the second part of the statement of Proposition 1.3 is constructed.

Note that the result of Proposition 1.3 is stated in a pointwise fashion, that is for one single f_0. As is typically the case for rate theorems, as noted in the Introduction, here in the case $\alpha \leq \beta$ one can check that the result holds uniformly over the Sobolev ball $\mathcal{F}_{\beta,L}$. In the case $\alpha \leq \beta$, Proposition 1.3 provides an existence result, for which it is natural to ask whether one can avoid the log-factor in the lower bound. The answer is yes if one allows sequences of functions: it can be checked that there exists a sequence $f_{0,n}$ in $\mathcal{F}_{\beta,L}$, where the function $f_{0,n}$ has only one properly chosen non-zero Fourier coefficient, such that, for M large enough, $\Pi(r_n^{\alpha,\beta}/M \leq \|f - f_{0,n}\|_2 \mid X^{(n)})$ tends to 1 in probability.

Finally, we note that Shen and Wasserman [95] obtained interesting partial lower-bound type results for the prior Π_α in their Theorem 6, which states that the negation of the upper-bound definition holds true for balls of small enough radius. As we have noted above, this is weaker than the lower-bound concept (1.1), since it does not exclude that some posterior mass remains close to the true f_0.

Illustration: Density estimation.

Consider observations $X^{(n)} = (X_1, \ldots, X_n)$ from the density model (3). So now f is a *density* on $[0,1]$. Suppose the true f_0 is a continuous, *positive* density and denote $w_0 = \log f_0$, so that $f_0 = e^{w_0}$.

We consider the prior on densities arising from, first, taking a Gaussian process prior on $[0,1]$, and next applying the exponential-type transformation (13) to make it a density. To simplify the presentation we take here Brownian motion W as a starting prior. Extensions to other priors are mentioned below. Then the quantity

$$p_W(t) = \frac{e^{W(t)}}{\int_0^1 e^{W(u)} du}, \tag{1.10}$$

with W standard Brownian motion, defines a random density on $[0,1]$. The corresponding (non-Gaussian) prior on the set of densities is denoted by Π_{p_w}. Brownian motion induces a Gaussian measure on L^2, but for the proof of the next result it is actually more convenient to view it within the Banach space $\mathcal{C}^0[0,1]$ of continuous functions on $[0,1]$ equipped with the supremum norm.

Theorem 1.3 (Case of Brownian motion $\alpha = 1/2$) *Suppose that $w_0 = \log f_0$ belongs to the Hölder class $\mathcal{C}^\beta[0,1]$ for some $\beta > 0$ and let the prior on densities be the distribution Π_{p_w} of p_W, where W is standard Brownian motion. Let φ_{w_0} denote the concentration function associated to W and w_0. Then there exist finite constants $C_1, C_2 > 0$ such that, if ε_n and ζ_n are such that*

$$\varphi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2 \quad \text{and} \quad \zeta_n \leq C_1 \varphi_{w_0}^{-1}(C_2 n\varepsilon_n^2),$$

then for M large enough, as $n \to \infty$, in $P_{f_0}^n$-probability,

$$\begin{aligned} \Pi_{p_w}(h(f, f_0) \leq M\varepsilon_n \mid X^{(n)}) &\to 1, \\ \Pi_{p_w}(\|f - f_0\|_\infty \geq \zeta_n \mid X^{(n)}) &\to 1. \end{aligned}$$

Moreover, one can choose $\varepsilon_n \lesssim n^{-1/4}$. Finally, in the case that $\beta \geq 1/2$, one can take $\zeta_n \gtrsim n^{-1/4}$.

Theorem 1.3 states results for density estimation, similar to those for white noise above, for a natural prior of 'regularity' $\alpha = 1/2$ (more precisely, Brownian motion paths almost surely belong to the Hölder space $\mathcal{C}^\beta$, any $\beta < 1/2$; the paths do belong to a well-chosen Besov space with regularity index $1/2$, almost surely), with a few differences. Before discussing these, let us note that Theorem 1.3 essentially extends to arbitrary α's by replacing Brownian motion by Riemann-Liouville type priors, see [P4], Theorem 3. Also, we note *en passant* that the latter result also extends upper-bound results derived in [100] in the case $\alpha = \beta$ to arbitrary pairs (α, β), which will be of use in Chapter 3.

While the upper-bound rate in Theorem 1.3 is in terms of Hellinger's distance, the lower bounds are in terms of the uniform norm. To obtain the lower bounds, the uniform norm is in a way the simplest distance to work with since KL-neighborhoods appearing in Theorem 1.1 can conveniently be related to sets of the form $B_n = \{f,\ \|f - f_0\|_\infty \leq \zeta_n\}$. For upper-bounds, Hellinger's distance is a rather natural choice since it is a natural testing distance for i.i.d. data in view of the theory of [48]. It would certainly be interesting to get a result in term of a common distance. We only note that already the question of obtaining upper-bounds for other distances than Hellinger is non-trivial, we will say more on this in Chapters 2-3.

1.4 Other examples

Squared-exponential Gaussian processes. A Gaussian prior distribution on smooth functions often encountered in applications is the centered Gaussian process $(Z_t)_{[0,1]}$ with covariance function $K(s,t) = \mathbb{E}(Z_s Z_t)$ given by

$$K(s,t) = e^{-(s-t)^2}, \qquad (s,t) \in [0,1]^2. \tag{1.11}$$

However, in this simple form it turns out that the corresponding prior has an undesirable behaviour even if the true function f_0 belongs to some natural classes such as Hölder or Sobolev. More precisely, van der Vaart and van Zanten [101] prove the following in the fixed regression model (we state it for simplicity in the Gaussian white noise model, the proof being the same).

Theorem 1.4 ([101]) *Consider the white noise model (1) with prior Π on f given by (1.11). There exists a function f_0, regular of order β in the Sobolev sense, such that, for some $l > 0$,*

$$\Pi(f :\ \|f - f_0\|_2 \leq (\log n)^{-l} \,|\, X^{(n)}) \to^{P_{f_0}^{(n)}} 0.$$

The result in [101] is even more precise: *any* function f_0 whose Fourier transform decreases slowly enough cannot lead to a rate faster than the logarithmic rate of Theorem 1.4.

This can be explained as follows: the sample paths of the squared exponential Gaussian process (1.11) can be shown to be extremely smooth (analytical). If the true f_0 happens to be a little bit 'non-smooth', then the prior will have difficulties detecting it; similar to the result for regularity

α above, a strong mismatch between regularity of the Gaussian prior and regularity of the true leads to slow rates. So without further modification the prior (1.11) is too rigid. A possible fix to this issue is discussed in Chapter 2.

Consequences for Gaussian processes.

Gaussian process are fine priors but, if the goal is nonparametric estimation, are a little too 'rigid': as the results of Proposition 1.3 and Theorem 1.3, 1.4 show, such priors lead to the optimal minimax rate only if their regularity matches that of the true f_0. If there is a mismatch in regularity, then the posterior rate may be suboptimal. Nevertheless, these priors turn out to be very useful building blocks within more complex priors leading to *adaptation* to the unknown regularity β, as we shall see at the beginning of Chapter 2.

Lower-bounds arguments are also useful in investigations on *adaptive* priors, see Theorem 2.4 below and, for a different problem, the paper [9] on anisotropic classes.

1.5 Discussion and perspectives

The notion of lower bound is a natural counterpart to the notion of upper-bound rate. One first obvious use is in proving that a rate obtained by a result such as Theorem 0.1 cannot be improved. Another useful aspect is in finding the 'boundaries' of possible classes of priors for a given problem: that is, finding *necessary* conditions for priors to attain a target rate.

The last sentence is reminiscent of the notion of *maxiset*, introduced by Kerkyacharian and Picard (2002) [60]. For a given procedure (point estimator), one looks for maximal sets of parameters on which the procedure attains a given rate. To our knowledge, maxisets have been investigated only for point estimators so far. We would find it interesting to 1) define a notion of maxiset for 'posterior measure estimators', 2) investigate the corresponding sets. The previous lower bound concept could be particularly useful in proving that a given parameter does not belong to a certain maxiset.

CHAPTER 2

Upper bounds for posterior rates

We present contributions to rates of convergence for posterior distributions. First, we focus on a natural method to obtain convergence on geometric spaces following [P11], based on random solutions of the heat equation. Convergence properties of the posterior are studied and optimality is discussed. Next, we examine Bayesian procedures in sparse settings as in [P9, P15]. We introduce families of priors which contain as special cases Bayesian analogues of thresholding. The linear regression model is studied, both from the point of view of recovery of the true parameter and prediction of the unknown mean vector. Finally we propose a programme [P12] to derive posterior rates of convergence for strong measures of distance not covered by available general theorems. We show how this programme can be put to use via some examples.

2.1 Posterior convergence on geometric spaces

On the real line. A particularly popular prior distribution on functions defined on $[0,1]$ (or $[0,1]^d$) is the squared-exponential Gaussian process (1.11). It is routinely used in machine learning, up to one, small in apparence, modification: one or several constant multiplicative factors are typically added to (1.11). These extra '*hyper-parameters*' are often tuned in practice by empirical Bayes methods. Van der Vaart and van Zanten proposed an elegant way to perform this tuning in a fully Bayes way, while at the same time enabling to obtain fast rates of convergence (Theorem 1.4 shows this is not possible for the basic version (1.11) of the prior). It turns out that it is enough to randomly *rescale* the prior. Consider, if $(Z_x)_{x\in[0,1]}$ is the squared-exponential Gaussian process (1.11),

$$\Pi \ \sim \ \ x \to Z_{Ax}, \tag{2.1}$$

where A is an independent random variable with a standard Gamma distribution. To make a prior on densities, it suffices to apply the exponential transformation $w \to p_w := e^w / \int_0^1 e^w$ as in (1.10).

Theorem 2.1 ([99]) *In the density model (3), let Π_{p_w} be the prior on f defined as the image measure of Π in (2.1) through the exponential transform p_w. Suppose $\log f_0$ belongs to $\mathcal{C}^\beta[0,1]$, $\beta > 0$. Then for M large enough, with h the Hellinger distance,*

$$\Pi_{p_w}\left[f : \ h(f,f_0) \le M\varepsilon_n \,|\, X^{(n)}\right] \to^{P^n_{f_0}} 1,$$

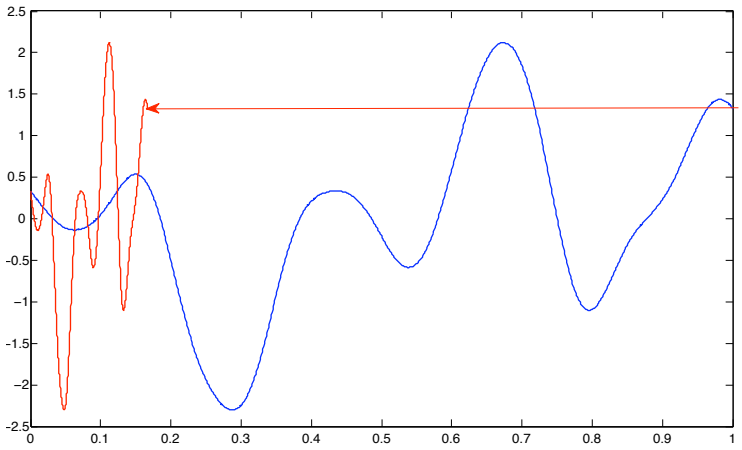

Figure 2.1: Random rescaling of paths.

where $\varepsilon_n = n^{-\frac{\beta}{2\beta+1}} (\log n)^{\frac{4\beta+1}{4\beta+2}}$.

So, the randomly rescaled Gaussian process achieves the minimax rate of convergence over Hölder classes, up to a logarithmic factor. Also, the result is *adaptive*, since the regularity of f_0 is not used in the prior construction. The result is striking, because a single shrinked path as in Figure 2.1 has the same regularity as the original, supersmooth, prior. Nevertheless, the 'extra randomness' introduced via the rescaling variable A is sufficient to make the whole prior flexible enough to lead to adaptation (up to a logarithmic factor). A natural question is: is it possible to obtain similar results for data sitting on a geometrical object, such as a sphere ? Indeed, rescaling of paths does not seem obvious (and is not !) when some geometry is involved ...

Geometric spaces. Consider a geometric framework such as the white noise model (4) or density estimation (5) on a compact metric space $\mathcal{M}$ with metric ρ and equipped with a Borel measure μ. There is a simple reason why the squared-exponential kernel cannot be used in such a context. Although (1.11) admits the immediate generalisation, for ρ the metric on $\mathcal{M}$,

$$\kappa_\rho(s,t) = e^{-\rho(s,t)^2}, \qquad (s,t) \in \mathcal{M}^2, \tag{2.2}$$

it can be shown that this function is *not* positive definite in general already for the simplest examples such as $\mathcal{M}$ taken to be the sphere in $\mathbb{R}^k$, $k \geq 2$. Yet, we shall see below that (1.11) admits a natural generalisation to this context, but it is not as simple as (2.2).

For simplicity we present in a slightly informal style the construction in [P11], giving pointers to the paper for details when appropriate. We take the case of the sphere $\mathcal{M} = \mathbb{S}^2$ as recurrent illustration. Let $B(x,r)$ denote the ball of center x and radius r for the metric ρ on $\mathcal{M}$. Suppose that $\mathcal{M}$ verifies the so-called *Ahlfors property*: there exist positive c_1, c_2, d such that

$$\text{for all } x \in \mathcal{M}, \text{ for all } \ 0 < r \leq 1, \quad c_1 r^d \leq |B(x,r)| \leq c_2 r^d. \tag{2.3}$$

In the case of the sphere $\mathbb{S}^2$, $d = 2$. More generally d in the sequel can be thought of as the 'dimension' of $\mathcal{M}$ although in general d could be non-integer.

Operator L, Laplacian and heat kernel. The starting point is a self-adjoint positive operator L on functions on $\mathcal{M}$ (more precisely on a domain D dense in $L^2(\mathcal{M})$, the space of square integrable functions with respect to the measure μ). When defined, *minus* the *Laplacian* on $\mathcal{M}$, that is $L = -\Delta_{\mathcal{M}}$ is typically appropriate. Suppose L admits a discrete spectrum with finite dimension spectral spaces $\mathcal{H}_k = \text{Vect}\{(e_k^l),\ 1 \leq l \leq \dim(\mathcal{H}_k)\}$ and that its eigenfunctions e_k^l are continuous functions on $\mathcal{M}$. The numbering is chosen so that the eigenvalues are ordered in an

increasing order. Also, in all this section, sums over k and l range over $1 \leq k \leq \dim(\mathcal{H}_k)$ and $1 \leq l < \infty$. Under some conditions, the following series converges to a continuous function

$$P_t(x,y) := \sum_k e^{-t\lambda_k} \sum_l e_k^l(x) e_k^l(y), \tag{2.4}$$

on $\mathcal{M} \times \mathcal{M}$, called the *heat kernel*. Let us justify this terminology in an informal way when $L = -\Delta_{\mathcal{M}}$. By informally differentiating under the series sign, we see that $P_t(\cdot, y)$ for any fixed y is a solution in g of the heat equation

$$\frac{\partial g}{\partial t} = (-L)g = \Delta_{\mathcal{M}} g. \tag{2.5}$$

For a more formal characterisation of P_t, in particular the connection to semi-groups, see [P11] Section 2.4 and references therein.

The orthonormal basis of $L^2(\mathcal{M})$ generated by $\{e_k^l\}$ can be interpreted as a *harmonic analysis* over $\mathcal{M}$. In the case of the sphere $\mathcal{M} = \mathbb{S}^2$, the eigenvectors of the Laplacian $\Delta_{\mathbb{S}^2}$ are well-known: these are the *spherical harmonics*, which have explicit expressions in terms of homogeneous polynomials of three variables, see [P11], Section 3.

Decoupling time and space. Let us note the presence of the indexing variable t, the 'time', in (2.4). For the squared-exponential kernel on the real-line, the prior can be made more flexible by stretching the path along the 'x'-axis, that is the *space* domain. Since in general there is no natural analogue of stretching on a geometric object, a natural idea is to stretch *time* instead. Indeed, our procedure puts a prior on time as we describe below. Let us now discuss a further property of P_t.

Estimates for the heat kernel. The following Gaussian-like estimates of the heat kernel P_t are satisfied in a surprisingly large variety of situations, in particular on all compact manifolds without boundary, see e.g. Grigor'yan [53], for instance on the sphere. We assume them to hold: suppose that there exist $C_1, C_2 > 0, c_1, c_2 > 0$, such that, for all $t \in]0,1[$, and any $x, y \in \mathcal{M}$,

$$\frac{C_2 e^{-\frac{c_2 \rho^2(x,y)}{t}}}{|B(x,\sqrt{t})|^{1/2} |B(y,\sqrt{t})|^{1/2}} \leq P_t(x,y) \leq \frac{C_1 e^{-\frac{c_1 \rho^2(x,y)}{t}}}{|B(x,\sqrt{t})|^{1/2} |B(y,\sqrt{t})|^{1/2}}, \tag{2.6}$$

where $|B(x,r)|$ denotes the volume of the ball $B(x,r)$.

Geometric prior. A prior on functions from $\mathcal{M}$ to $\mathbb{R}$ is constructed hierarchically as follows.

First, generate a collection of independent standard normal variables $\{X_k^l\}$ with indexes $k \geq 0$ and $1 \leq l \leq dim(\mathcal{H}_{\lambda_k})$. Set, for $x \in \mathcal{M}$ and any $t \in (0,1]$,

$$W^t(x) = \sum_k \sum_l e^{-\lambda_k t/2} X_k^l e_k^l(x). \tag{2.7}$$

This process is centered and has covariance kernel precisely P_t, as follows by direct computation,

$$\mathbb{E}(W^t(x) W^t(y)) = P_t(x,y).$$

Second, draw a positive random variable T according to a density g on $(0,1]$. This variable can be interpreted as a random scaling, or random 'time'. It turns out that convenient choices of g are deeply connected to the geometry of $\mathcal{M}$. We choose the density g of T such that, for a positive constant q, with d defined in (2.3),

$$g(t) = e^{-t^{-d/2} \log^q(1/t)}, \quad t \in (0,1]. \tag{2.8}$$

We show below that the choice $q = 1 + d/2$ leads to sharp rates.

The full (non-Gaussian) prior we consider is W^T, where T is random with density g. That is,

$$W^T(x) = \sum_k \sum_l e^{-\lambda_k T/2} X_k^l e_k^l(x), \tag{2.9}$$

and we define Π as the prior on functions on $\mathcal{M}$ induced by W^T.

Does the prior (2.8) relate to the square-exponential GP ? So far there does not seem to be a direct connection between our construction and that of [99]. However, such a connection becomes apparent when taking another look at equation (2.6). We see that the covariance kernel of W^t for a given t very closely relates to $e^{-c\rho^2(x,y)/t}$, which however is not itself in general a covariance kernel as noted above. In this sense, the heat kernel *is* the natural generalisation of the squared-exponential kernel $e^{-C(x-y)^2}$ to geometric spaces.

Sketch of required arguments. To obtain convergence rates corresponding to the prior Π and derive Theorem 2.3 below, we use the general rate Theorem 0.1. As we have explained in the Introduction, for Gaussian processes a rate is obtained by solving in ε_n the equation $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$, with φ the concentration function of the process. Here Π is not Gaussian, but conditionally on a given value of T, say $T = t$, the prior induced by W^t is Gaussian by construction. So, an important step in the proof is the study of the concentration function φ_{f_0} of W^t at the true function f_0, which involves an approximation term as well as the small ball probability of W^t.

The approximation part of φ requires some regularity condition on f_0; it turns out that it is particularly natural to work with a scale of Besov spaces, which may precisely be defined in terms of quality of approximation. Define first the 'low frequency' functions from the eigenspaces $\mathcal{H}_\lambda$ as

$$\Sigma_t = \bigoplus_{\lambda \leq \sqrt{t}} \mathcal{H}_\lambda.$$

Next, let $\mathcal{E}_t(f)_p := \inf_{g \in \Sigma_t} \|f - g\|_p$ denote the best approximation of $f \in L^p = L^p(\mathcal{M})$ from Σ_t. Then the Besov space $B^s_{pq}(\mathcal{M})$ is defined as

$$B^s_{pq}(\mathcal{M}) := \{f \in L^p, \quad \|f\|_{A^s_{pq}} := \|f\|_p + \Big(\sum_{j \geq 0} \big(2^{sj} \mathcal{E}_{2^j}(f)_p\big)^q \Big)^{1/q} < \infty\}. \tag{2.10}$$

Assuming a $B^s_{2,\infty}(\mathcal{M})$-regularity in the white noise case and a $B^s_{\infty,\infty}(\mathcal{M})$-regularity in the density estimation case enables a control of the approximation part.

The study of the small ball probability of the process W^t is more delicate, especially since we look for sharp rates. We achieve this by using the general very precise link existing for Gaussian processes between small ball probability and entropy of the RKHS, as established in [65]. For this, we need first the expression of the RKHS say $\mathbb{H}^t$ of W_t.

The prior W^t and its RKHS $\mathbb{H}^t$. For any $t > 0$, it follows from the expression of W^t that

$$\mathbb{H}^t = \Big\{ h = \sum_k \sum_l a_k^l e^{-\lambda_k t/2} e_k^l, \qquad \sum_{k,l} |a_k^l|^2 < \infty \Big\}, \tag{2.11}$$

equipped with the inner product

$$\langle \sum_k \sum_l a_k^l e^{-\lambda_k t/2} e_k^l \ , \ \sum_k \sum_l b_k^l e^{-\lambda_k t/2} e_k^l \ \rangle_{\mathbb{H}^t} = \sum_k \sum_l a_k^l b_k^l.$$

Let us further denote $\mathbb{H}^t_1$ the unit ball of $\mathbb{H}^t$.

Key estimates. The next result is a sharp entropy estimate of the RKHS unit ball $\mathbb{H}^t_1$, uniform in a range of time parameters t. The statement brings together *geometry* via the covering number $N(\epsilon, \mathcal{M}, \rho)$ of the space $\mathcal{M}$, *probability* via the RKHS of the process W^t and *approximation*, via the entropy of $\mathbb{H}^t_1$, denoted $H(\cdot, \mathbb{H}^t_1, D) = \log N(\cdot, \mathbb{H}^t_1, D)$, for a given distance D on $\mathbb{H}$.

Theorem 2.2 *Suppose the space $\mathcal{M}$, the operator L and its eigenfunctions e_k^l verify the properties listed above. For $t > 0$, let $\mathbb{H}^t$ be defined by (2.11). Let us fix $a > 0, \nu > 0$. There exists $\epsilon_0 > 0$ such that for ϵ, t with $\epsilon^\nu \leq at$ and $0 < \epsilon \leq \epsilon_0$,*

$$H(\epsilon, \mathbb{H}_1^t, \|\cdot\|_2) \asymp H(\epsilon, \mathbb{H}_1^t, \|\cdot\|_\infty) \asymp N(\delta(t,\epsilon), \mathcal{M}, \rho) \cdot \log\frac{1}{\epsilon}, \quad \text{with} \quad \frac{1}{\delta(t,\epsilon)} := \sqrt{\frac{1}{t}\log\frac{1}{\epsilon}}.$$

Under the assumption (2.3) that balls have a polynomially increasing volume in terms of their radius, the covering number $N(\eta, \mathcal{M}, \rho)$ of $\mathcal{M}$ is shown to be $N(\eta, \mathcal{M}, \rho) \asymp \eta^{-d}$, which yields the estimate $\delta(t,\epsilon)^{-d}\log(1/\epsilon)$ for the entropy in Theorem 2.2. From this one can deduce an estimate of the same order $-\log\mathbb{P}(\|W^t\|_2 < \epsilon) \asymp -\log\mathbb{P}(\|W^t\|_\infty < \epsilon) \asymp t^{-d/2}\log^{1+d/2}(1/\epsilon)$ for the small ball probabilities, both in terms of the L^2- and L^∞-norms.

Convergence rate for the geometric prior. The following theorem states a result for the white noise and density estimation problems on $\mathcal{M}$. In the first case, the prior is directly the law on $L^2(\mathcal{M})$ induced by W^T in (2.9). In density estimation, the prior is the image measure of the law of W^T viewed as a random element of $\mathcal{C}^0(\mathcal{M})$ under the exponential transformation $w \to p_w^{[\mathcal{M}]} = e^w / \int_{\mathcal{M}} e^w$ on $\mathcal{M}$. Recall the definition of the Besov spaces from (2.10).

Theorem 2.3 *Let the set $\mathcal{M}$ and the operator L satisfy the properties listed above. Consider the white noise model (4) on $\mathcal{M}$. Suppose that f_0 is in the Besov space $B^s_{2,\infty}(\mathcal{M})$ with $s > 0$ and that the prior Π on f is W^T given by (2.9). Let $q = 1 + d/2$ in (2.8). Set $\varepsilon_n = (\log n/n)^{2s/(2s+d)}$. For M large enough, as $n \to \infty$,*

$$\Pi(\|f - f_0\|_2 \geq M\varepsilon_n \mid X^{(n)}) \to^{P_{f_0}^{(n)}} 0.$$

Consider the density model (5) on $\mathcal{M}$. Suppose that $\log f_0$ is in the Besov space $B^s_{\infty,\infty}(\mathcal{M})$ with $s > 0$ and that the prior Π on f is $p_{W^T}^{[\mathcal{M}]}$ with W^T as in (2.9). With q, ε_n as before and h the Hellinger distance between densities on $\mathcal{M}$, for M large enough, as $n \to \infty$,

$$\Pi(h(f, f_0) \geq M\varepsilon_n \mid X^{(n)}) \to^{P_{f_0}^n} 0.$$

Again, uniformity in the results can be obtained on balls of the considered Besov spaces.

The rate is sharp. The rate ε_n in Theorem 2.3 contains an additional logarithmic term with respect to the minimax rates [38] of estimation on $\mathcal{M}$. It is natural to ask whether the posterior rate indeed includes such a term. Again, the set $\mathcal{M}$ and the operator L are as before.

Theorem 2.4 *Consider the white noise model (4) on $\mathcal{M}$. Let $\varepsilon_n = (\log n/n)^{s/(2d+s)}$ for $s > 0$ and let the prior Π on f be the law induced by W^T, see (2.9), with $q > 0$ in (2.8). Then there exist f_0 in the unit ball of $B^s_{2,\infty}(\mathcal{M})$ and a constant $c > 0$ such that*

$$\Pi(\|f - f_0\|_2 \leq c(\log n)^{0\vee(q-1-\frac{d}{2})}\varepsilon_n \mid X^{(n)}) \to^{P_{f_0}^{(n)}} 0.$$

This result shows that the posterior rate obtained in Theorem 2.3 cannot be improved upon over $B^s_{2,\infty}(\mathcal{M})$, so the logarithmic factor is necessary. Also, it reveals that if q in (2.8) exceeds $1 + d/2$, then the convergence rate becomes slower.

We now move to a quite different model, but where the general approach of the rate Theorem 0.1 will still be of use.

2.2 Needles and straw in a haystack with sparse priors

In this Section we present results from [P9] for the sparse sequence model (6). The model is at the basis of many other statistical frameworks, and the results can find applications in other

contexts. One main achievement is in the understanding of posterior rates for one important class of prior distributions modelling sparsity, namely priors involving Dirac masses at 0. The results also suggest a kind of 'dictionary' between some thresholding rules [35, 54] and their fully Bayesian counterparts, which are posterior measures. In particular, we present a natural way of choosing the thresholding constant. The results also serve as guideline for the study of the linear regression model in the next Section. In the sequence model here, we shall use the approach of the general rate Theorem 0.1, although the scheme of proofs needs to be adapted if one looks for sharp rates. We prove results for estimators without 'explicit' expressions (meaning there is one, but is a complicated ratio of integrals). We denote by $E_{\theta_0}^{(n)}$ the expectation under $P_{\theta_0}^{(n)}$.

Model and prior. Let us recall the sparse sequence model (6)

$$X_i = \theta_i + \varepsilon_i, \qquad i = 1, \dots, n,$$

where the true $\theta_0 = (\theta_{0,1}, \dots, \theta_{0,n})$ has at most p_n non-zero coefficients. Due to its importance, this model has been studied by many authors, and we only cite a few key papers such as [35] (thresholding), [18, 52] (model selection and penalisation), [1] (false discovery rate approach) and [57, 58] (empirical Bayes). Close connections with, in particular, [58], will appear below.

We consider a prior $\Pi = \Pi_n$ on $\mathbb{R}^n$ constructed in three steps:

(P1) A *dimension* k is chosen according to a prior probability measure π_n on the set $\{0, 1, 2, \dots, n\}$.

(P2) Given k a subset $S \subset \{1, \dots, n\}$ of size $|S| = k$ is chosen uniformly at random from the $\binom{n}{k}$ subsets of size k.

(P3) Given (k, S) a vector $\theta_S = (\theta_i : i \in S)$ is chosen from a probability distribution with Lebesgue density g_S on $\mathbb{R}^S$ and this is extended to $\theta \in \mathbb{R}^n$ by setting the remaining coordinates θ_{S^c} equal to 0.

Suppose g_S is an independent product of $|S|$ times a same given density g,

$$g_S = \bigotimes_S g \tag{2.12}$$

We also assume that $\pi_n(k) > 0$ for any k. Note that the prior is characterised by the pair (π_n, g).

Posterior. Given the prior Π, one can form the posterior distribution, and use Bayes' formula to write it as a fairly complicated ratio of sum of integrals. We do not discuss further the expression here, which may be found as Equation (2.1) in [P9].

We now examin conditions on the pair (π_n, g) that appear naturally.

Prior π_n, the Spike, and posterior dimension.

We say a prior π_n on dimension has *exponential decrease* if, for some $C > 0$ and $D < 1$,

$$\pi_n(k) \le D\pi_n(k-1), \quad k > Cp_n. \tag{2.13}$$

If the condition is also satisfied with $C = 0$, we speak of *strict* exponential decrease.

Theorem 2.5 (Dimension) *If π_n has exponential decrease (2.13) and g_S is a product of $|S|$ copies of a univariate density g, with mean zero and finite second moment, then there exists $M > 0$ such that, as $p_n, n \to \infty$,*

$$\sup_{\theta_0 \in \ell_0[p_n]} E_{\theta_0}^{(n)} \Pi_n\big(\theta : \ |S_\theta| > Mp_n \,|\, X^{(n)}\big) \to 0.$$

For reasonable priors, we may hope that the posterior distribution spreads mass in the p_n-dimensional subspace that supports a true mean vector $\theta_0 \in \ell_0[p_n]$. The theorem shows that

the posterior distribution "overshoots" this space by subspaces of dimension at most a multiple of p_n. Because the overshoot can have a random direction, this does not mean that the posterior distribution concentrates overall on a fixed Mp_n-dimensional subspace. The theorem shows that it concentrates along Mp_n-dimensional coordinate planes, but its support will be far from convex.

PRIOR g, THE SLAB, AND HEAVY TAILS.

We further assume that g in (2.12) can be written $g = e^h$, for a function $h : \mathbb{R} \to \mathbb{R}$ satisfying

$$\big|h(x) - h(y)\big| \lesssim 1 + |x - y|, \qquad \forall x, y \in \mathbb{R}. \tag{2.14}$$

This covers all densities e^h with a uniformly Lipshitz function h, such as the Laplace and Student densities. It also covers other smooth densities with polynomial tails, and densities of the form $c_\alpha e^{-|x|^\alpha}$ for some $\alpha \in (0, 1]$. On the other hand the standard normal density is ruled out. Indeed, as we know from the lower bound result of Proposition 1.2, for $g(x)$ proportional to $e^{-|x|^\alpha}$ and $\alpha > 1$, the minimax rate cannot be expected uniformly over the whole class $\ell_0[p_n]$.

Example 2.1 (Independent Dirac mixtures, or Bayesian α-thresholding) Consider the prior as in (1.4)

$$\Pi_{\alpha,g} \sim \otimes_{i=1}^n (1 - \alpha)\delta_0 + \alpha g$$

This construction induces a prior π_n on dimension equal to the *binomial* law with parameters n and α. It has exponential decrease (2.13) if $\alpha \lesssim p_n/n$. Furthermore, the nonzero coordinates are distributed according to the product of copies of g. Thus this prior fits in our set-up (P1)-(P3) above. This prior is considered in George and Foster (2000) [44] and Johnstone and Silverman (2004) [58], in combination with a Gaussian or a heavy tailed density g, respectively.

For a fixed α the coordinates θ_i are independent, under both the prior and the posterior distribution. Furthermore, the posterior distribution of θ_i depends on X_i only.

The prior $\Pi_{\alpha,g}$ has connections with thresholding rules. As noted in Chapter 1, if $\alpha = 1$ and g is the Laplace law, the posterior mode is a soft-thresholding rule. If $\alpha \in (0, 1)$, some (but not all) aspects of the posterior distribution put some coordinates to 0. This is the case for the posterior *median*, as considered in [58], where the posterior median is shown to be a thresholding rule: a given coordinate equals 0 if and only if the value of the corresponding observation drops in absolute value below a certain level. On the contrary, the posterior *mean* is not a thresholding rule: all its coordinates are typically nonzero. This has some nontrivial consequences, see below.

Now at this point the question, also central for thresholding procedures arises: how does one choose α ? One possibility is to set $\alpha = \alpha_n = 1/n$, see below the next Theorem for a discussion. The authors of [58] propose, on the top of using the coordinatewise posterior median for estimating θ, to set the weight parameter α by a thresholded *empirical Bayes method*. The parameter is chosen equal to the maximum likelihood estimator of α based on the marginal distribution of X in the Bayesian set-up (i.e. with θ integrated out but with fixed α) subject to the constraint that the resulting posterior median (after plugging in $\hat{\alpha}$) given an observation in the interval $[-(2\log n)^{1/2}, (2\log n)^{1/2}]$ is zero. The authors show that the resulting point estimator works remarkably well, in a minimax sense, for various metrics and sparsity classes.

Example 2.2 (Binomial and Beta-binomial priors, or full Bayesian thresholding) The binomial (n, α_n) distribution for π_n gives an expected dimension of $n\alpha_n$. In the sparse setting a small value of α_n is therefore natural. If the sparsity parameter p_n were known, we could consider the choice $\alpha_n = p_n/n$; we shall refer to the corresponding law as *oracle binomial prior*.

A natural Bayesian strategy is to view the unknown "sparsity" parameter α as a hyperparameter and put a prior on it. The classical choice is the Beta prior, leading to the hierarchical scheme $\alpha \sim \text{Beta}(\kappa,)$ and $k \,|\, \alpha \sim \text{Binomial}(n, \alpha)$. This yields a mixture of binomials as a prior π_n on the dimension of the model. The independence of the coordinates θ_i is then lost.

For $\kappa = 1$ and $= n + 1$ we obtain $\pi_n(k) \propto \binom{2n-k}{n}$. Then $\pi_n(k)/\pi_n(k-1) = (n - k + 1)/(2n - k + 1)$, showing (strict) exponential decrease (2.13), with $D = 1/2$.

Example 2.3 (Complexity prior) For positive constants a, b, let us set

$$\pi_n(k) \propto e^{-ak\log(bn/k)}, \tag{2.15}$$

where $\propto$ stands for 'proportional to'. Because $e^{k\log(n/k)} \le \binom{n}{k} \le e^{k\log(ne/k)}$, this prior is inversely proportional to the number of models of size k, a quantity that can be viewed as the *model complexity* for a given dimension k. Thus this prior appears particularly suited to the purpose of "downweighting the complexity". Forgetting about the extra component g_S of the prior, we can also consider it an analogue of the penalty "$2k\log(n/k)$" used in model selection in this context by (e.g.) Birgé and Massart in [18]. Every particular model with support S of size $|S| = k$ then receives prior probability bounded below and above by expressions of the type $e^{-a_1 k\log(b_1 n/k)}$.

RECOVERY. We next obtain posterior convergence rates for a variety of measures of loss. We show below that the conditions of the Theorem are satisfied for all three examples above.

The ℓ^q metric for $0 < q \le 2$, is defined (without qth-root) by

$$d_q(\theta, \theta') = \sum_{i=1}^{n} |\theta_i - \theta_i'|^q. \tag{2.16}$$

For $q < 2$ this "metric" is more sensitive to small variations in the coordinates than the square Euclidean metric, which is d_2.

Theorem 2.6 (Recovery) *In the sparse sequence model (6) with prior $\Pi \equiv (\pi_n, g)$, if π_n has exponential decrease (2.13) and g in (2.12) has mean zero, finite second moment and can be written $g = e^h$ with h satisfying (2.14), then for any $q \in (0, 2]$, for r_n satisfying*

$$r_n^2 \ge \{p_n \log(n/p_n)\} \vee \log \frac{1}{\pi_n(p_n)}, \tag{2.17}$$

and sufficiently large M, as $p_n, n \to \infty$ such that $p_n/n \to 0$,

$$\sup_{\theta_0 \in \ell_0[p_n]} E_{\theta_0}^{(n)} \Pi\big(\theta : d_q(\theta, \theta_0) > M r_n^q p_n^{1-q/2} \,|\, X^{(n)}\big) \to 0.$$

The minimax rate over $\ell_0[p_n]$ for d_q is known to be of the order, see e.g. [34],

$$r_{n,q}^* = p_n \log^{q/2}(n/p_n). \tag{2.18}$$

For $q = 2$ the theorem refers to the square Euclidean distance d_2, and asserts that the posterior distribution contracts at the rate r_n^2, uniformly over $\ell_0[p_n]$. The first inequality in (2.17) says that this rate is, of course, not faster than the minimax rate $r_{n,2}^* \asymp p_n \log(n/p_n)$ above. The second shows that it is also limited by the amount of prior mass $\pi_n(p_n)$ put on the true dimension. If this satisfies, for some $c > 0$,

$$\pi_n(p_n) \gtrsim \exp\big(-c p_n \log(n/p_n)\big). \tag{2.19}$$

then $\log(1/\pi_n(p_n)) \lesssim r_{n,2}^*$ and the rate r_n^2 is equal to the minimax rate.

For $q \in (0, 2)$ we can make similar remarks. The minimax rate $r_{n,q}^*$ over $\ell_0[p_n]$ for d_q is given in (2.18). Because

$$(r_{n,2}^*)^{q/2} p_n^{1-q/2} = r_{n,q}^*,$$

the theorem shows contraction of the posterior distribution relative to d_q at the minimax rate $r_{n,q}^*$ over $\ell_0[p_n]$ under the same conditions that it gives the minimax rate $r_{n.2}^*$ for d_2: (2.19) suffices.

THE EXAMPLES. The binomial prior of Example 2.1 has exponential decrease (2.13) if $\alpha = \alpha_n \lesssim p_n/n$. The oracle binomial prior $\alpha_n \asymp p_n/n$ is at the upper end of this range, and also satisfies (2.19), and thus yields the minimax rate of contraction. The choice $\alpha_n = 1/n$ yields $\log \pi_n(p_n)$

of the order $-p_n \log p_n$, and hence attains the minimax rate if p_n is of the order n^a, $a < 1$; for larger p_n it may miss the minimax rate by a logarithmic factor. The full Bayesian thresholding of Example 2.2 with $\kappa = 1$, $\lambda = n + 1$ can be shown to verify $\pi_n(p_n) \gtrsim e^{-p_n(1+o(1))}$ if $p_n/n \to 0$, and hence (2.19) is satisfied. Finally, the complexity prior of Example 2.3, verifies exponential decrease for large enough b, as well as (2.19), so yields optimal rates as well.

THE MEAN/MEDIAN PHENOMENON. When $0 < q < 1$ the result of Theorem 2.6 is surprising at first when compared to the finding in [58] that the posterior *median*, or more generally so-called "strict-thresholding rules", attain the convergence rate $r^*_{n,q}$, but the posterior *mean* converges at a *strictly slower* rate (even when $\theta_0 = 0$; see [58], Section 10). By the preceding theorem the *full* posterior distribution *does* contract at the optimal rate $r^*_{n,q}$, for any $0 < q < 2$.

This is another striking illustration of the fact that in general, the *full* posterior measure and its aspects (mean, mode, median etc.) may have *completely different* behaviours.

The slower convergence of the posterior mean relative to the contraction of the full posterior distribution is made possible by the fact that d_q-balls have astroid-type shapes for $0 < q < 1$, and differ significantly from their convex hull if n is large. The posterior mean, which is in the convex hull of the support of the posterior, can therefore be significantly farther in d_q-distance from θ_0 than the bulk of the distribution. By Theorem 2.5 only few coordinates outside the support of θ_0 are given non-zero values by the posterior. However, the corresponding indices are random and *on average* spread over $\{1, 2, \ldots, n\}$, which makes that the posterior mean at a fixed coordinate is typically non-zero. Adding up all small errors in ℓ^q typically gives a much higher total sum for $q < 1$ than for $q \geq 1$. In contrast the posterior median does not suffer from this averaging effect.

The posterior measure thus provides a unifying point of view on the considered objects. In this perspective for $0 < q < 1$ the posterior mean is a bad representation of the full posterior measure.

RECOVERY FOR COMPLEXITY PRIORS. For the complexity priors of Example 2.3, the following theorem gives a more precise result on the contraction of the posterior measure in terms of the Euclidian distance $\|\cdot\|_2 = d_2^{1/2}$.

Theorem 2.7 (Recovery, complexity priors) *In model (6) with prior $\Pi \equiv (\pi_n, g)$, suppose that π_n is given by (2.15) with $a \geq 1$ and $b \geq e^{7+2c_1}$, and that g is as in Theorem 2.6. Then, for r_n satisfying (2.17), for any $1 \leq p_n \leq n$ and $r \geq 1$,*

$$\sup_{\theta_0 \in \ell_0[p_n]} E^{(n)}_{\theta_0} \Pi\big(\theta : \|\theta - \theta_0\|_2 > 45 r_n + 10 r \,|\, X\big) \lesssim e^{-r^2/10}.$$

The result provides a fast decrease to 0 of the posterior mass outside a euclidian ball of a constant times the minimax rate. From this result –for brevity we refrain from writing down the full statements and refer to [P9]– one can

- deduce that several point estimators obtained from the posterior converge at optimal rates. For instance, the posterior mean corresponding to any prior as in Theorem 2.7 converges at minimax rate in terms of Euclidian loss. Also, we prove in [P9] that the posterior *coordinatewise median* is minimax with respect to all d_q-losses $0 < q \leq 2$.

- derive results for so-called *weak-classes* $m_s[p_n]$, for which sparsity is defined as corresponding to a certain decrease to 0 of the ordered coefficients, instead of most coefficients being exactly 0 as for the $\ell_0[p_n]$ class. Priors as in Theorem 2.7 are minimax for such classes.

- extend the results of Theorem 2.7 to priors allowing some dependence between the coordinates, with g_S not necessarily an independent product of g's, see [P9], Example 2.6, where 'weakly-mixing' priors are considered.

ALGORITHM AND SIMULATIONS. For a general prior from (P1)-(P3) above and under the assumption that g_S is a coordinatewise product as in (2.12), we provide in [P9], Section 3, a polynomial time algorithm for simulating from some aspects of the posterior distribution, such as posterior

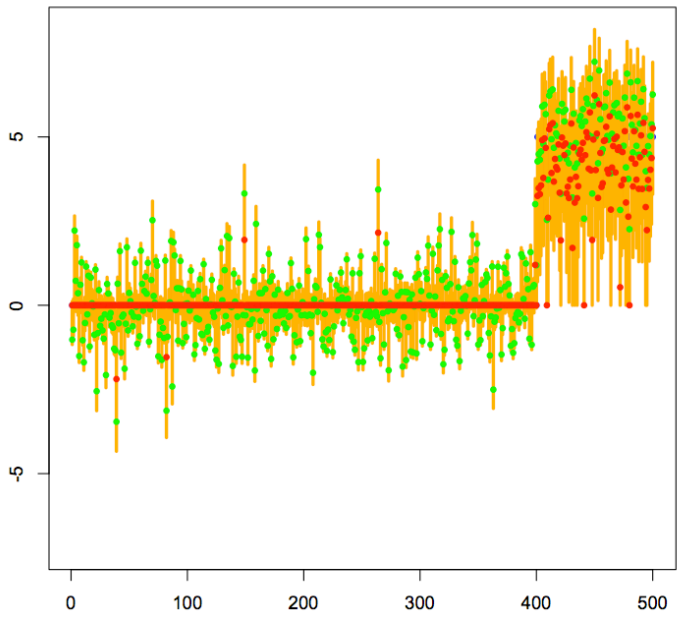

Figure 2.2: Marginal posterior medians (red dots) and marginal credible intervals (orange) for the parameters $\theta_1, \ldots, \theta_n$ for a single data vector $X_1, \ldots, X_n$ simulated according to the model (6) with $\theta = (0, 0, \ldots, 0, 5, \ldots, 5)$, where $n = 500$ and the last $p_n = 100$ coordinates are nonzero. The data points are indicated by green dots. The prior g is standard Laplace and $\pi_n(k) \propto \binom{2n-k}{n}$.

mean, posterior coordinatewise median, posterior number of selected models etc. The algorithm is based on a polynomial multiplication idea: using the product structure of the likelihood combined with the specific form of the mixture priors we consider, one can recursively compute the terms appearing in Bayes' formula by identifying them to coefficients of a polynomial. The latter can be computed at fairly low computational cost. The estimates of θ we propose, based on the posterior coordinatewise median for complexity-type priors, are competitive with the best results obtained by using the 'R' package `EbayesThresh` of Johnstone and Silverman [58]. A detailed comparison with discussion can be found in [P9], Section 3.

We present in Figure 2.2 a typical simulation output for a prior satisfying the conditions of Theorem 2.6 and corresponding to the 'full Bayesian thresholding' of Example 2.2. The posterior coordinatewise median (red dots) correctly sets to 0 most of the true zero coefficients while at the same time performing quite well on the part where true coefficients are separated from 0.

FURTHER RESULTS. So far we have mainly discussed sparse rates of convergence for posterior distributions for recovery of the parameter θ. Clearly, and as suggested by Figure 2.2, there are further interesting questions. For instance, the model selection question: if there is enough signal strength on some given coordinates, does the posterior selected correctly picks up at least those coordinates ? Can one say something about a possible limit in distribution of the posterior ? Also, slightly more on the technical side, pushing further than minimaxity, one may be interested in oracle-type results for the posterior. Some of these topics are considered in the following Section, under slightly more specific conditions for the prior distribution.

2.3 Sparse Bayesian linear regression

MODEL. Consider estimation of the parameter $\beta \in \mathbb{R}^p$ in the linear regression model (8)

$$Y = X\beta + \epsilon,$$

where X is a given, deterministic $(n \times p)$ matrix, and ϵ is an n-variate standard normal vector.

RECOVERY, PREDICTION, MODEL RECOVERY. As far as estimation is concerned, there are three main questions one may ask. It is important to distinguish between them, as they typically require assumptions of different natures. *Recovery* is concerned with estimation of β in the above model. When $n < p$ the parameter β is typically not identifiable, but if sparsity of β is assumed, it may become identifiable under some non-trivial assumptions on the design matrix X. Recovery can also be seen as an inverse problem with operator X. *Prediction* is concerned with estimation of the mean parameter $X\beta$. One may hope that this requires less stringent assumptions, as one does not have to 'invert the inverse problem'. Finally, *model recovery*, or model selection, consists in finding, when possible, the set of indexes for which a nonzero coefficient is present. This task requires the most stringent assumptions: indeed, already in the sparse sequence model (6), finding the support S_θ of the true θ is only possible if the nonzero coefficients of θ are well-enough separated from 0. The type of practical problem at hand may determine which type of assumptions on X are reasonable, and in turn which ones of these estimation questions one can solve.

Due to its central importance in applications, there is a large growing body of literature on the sparse regression model, and some important contributions in a non-Bayesian framework include Donoho et al. (2006) [33], Candès and Tao (2007) [25], Bickel et al. (2009) [11], as well as [3, 4, 24, 61, 87, 106, 107]. We refer to the book by Bühlmann and van de Geer (2011) [23] for an overview and further references. Some important Bayesian contributions include [20, 43, 44, 56, 84, 93, 105]. To our knowledge though little is known so far concerning posterior convergence rates.

So, pursuing the approach used for the sequence model in the last Section, we now consider a Bayesian approach for model (8) based on priors that set a selection of coefficients β_i a priori to zero; equivalently, priors that distribute their mass over models that use only a (small) selection of the columns of X. Bayes's formula gives a posterior distribution as before, and we denote by $E_{\beta^0} = E_{\beta^0}^{(n,p)}$ the expectation under $P_{\beta_0}^{(n,p)}$.

NOTATION. For a vector $\beta \in \mathbb{R}^p$ and a set $S \subset \{1, 2, \ldots, p\}$, let β_S be the vector $(\beta_i)_{i \in S} \in \mathbb{R}^S$, and $|S|$ the cardinality of S. The *support* of the parameter β is the set $S_\beta = \{i : \beta_i \neq 0\}$. The support of the true β^0 is denoted S_0, with cardinality $s_0 := |S_0|$. Moreover, we write $s = |S|$ if there is no ambiguity to which set S is referred to. We let $X_{\cdot,i}$ be the ith column of the design matrix X, and

$$\|X\| = \max_{i=1,\ldots,p} \|X_{\cdot,i}\|_2 = \max_{i=1,\ldots,p} (X^t X)_{i,i}^{1/2}. \tag{2.20}$$

PRIOR. We consider a specific sub-class of the priors $\Pi \equiv (\pi_n, g)$ from Section 2.2. More precisely, we construct Π again via (P1)-(P3), with g as in (2.12), but this time assuming that

- g is the Laplace density on $\mathbb{R}$ with parameter λ

$$g(\beta) = (\lambda/2)e^{-\lambda|\beta|}, \qquad \beta \in \mathbb{R},\ \lambda > 0. \tag{2.21}$$

- the prior on dimension π_n is given by, for constants $a, c > 0$,

$$\pi_p(s) \propto c^{-s} p^{-as}, \qquad s = 0, 1, \ldots, p. \tag{2.22}$$

That is, we restrict to priors on coefficients that are Laplace distributed with parameter λ. Assumptions on λ itself are discussed next. The prior π_p on dimension is essentially the same as the complexity prior (2.15), the only minor difference being that a factor $\log p/s$ is replaced by $\log p$.

This only possibly affects logarithmic factors in the rate so here we work with the simpler (2.22).

THE TUNING PARAMETER λ. We allow the inverse scale parameter λ to change with p, within the range, with $\|X\|$ defined in (2.20),

$$\frac{\|X\|}{p} \leq \lambda \leq 2\overline{\lambda}, \qquad \text{where } \overline{\lambda} := 2\|X\|\sqrt{\log p}. \tag{2.23}$$

So $\lambda/\|X\|$ is allowed to be a constant, or any quantity between p^{-1} and $2(\log p)^{1/2}$. This broad range is in stark contrast to the usual choice of the smoothing parameter in the LASSO, which must be chosen proportional to $\|X\|(\log p)^{1/2}$ and corresponds to the upper bound in (2.23).

This can be explained as follows: the standard LASSO must have $\lambda/\|X\|$ tending to infinity in order to be sufficiently good at identifying noise as 'zero' (and indeed $\|X\|(\log p)^{1/2}$ corresponds to a 'noise level' under which it is desirable to threshold). This has the undesirable effect that the LASSO will slightly shrink the nonzero coefficients of β towards 0. While this extra 'bias' towards 0 does not harm its rate of convergence, it may be interesting in practice to have some correction for it. What happens is that the LASSO must perform both model selection and good estimation of non-zero coefficient at once, and hence the choice of λ is essentially forced up to a constant.

The prior Π naturally allows for some extra flexibility, via its two components: the prior π_p takes care of the model dimension part, while the Laplace prior densities model the nonzero coordinates. Large values of λ would shrink the nonzero coordinates to zero, which is clearly undesirable. Thus it is natural to assume $\lambda \ll \overline{\lambda}$, and fixed values of $\lambda/\|X\|$, and even values decreasing to zero – making the prior non-informative – should be well-adapted to the problem. Also, small values of λ permit a distributional approximation to the posterior distribution centered at unbiased estimators.

Example 2.4 (Sequence model) *This corresponds to $X = I$ and $n = p$ in the present setting, whence $\|X\| = 1$. Condition (2.23) reduces to $p^{-1} \leq \lambda \leq 4\sqrt{\log p}$. Fixed values of λ, as considered in the previous Section 2.2 are easily included. As there is only one observation per parameter, it may not be unreasonable to consider $\lambda \to 0$, in order to create noninformative priors for the nonzero coefficients. This is allowed easily also.*

Example 2.5 (Response model) *If every row of the regression equation $Y = X\beta + \epsilon$ refers to measurement of an instance of a fixed relationship between an input vector $X_{i,\cdot} \in \mathbb{R}^p$ and the corresponding output Y_i, then the entry $X_{i,j}$ of X is the value of individual i on the jth covariable. It is then reasonable to think of these entries as being sampled from some fixed distribution, independent of n and p, in which case $\|X\|$ will (typically) be of the order $\sqrt{n}$. Condition (2.23) reduces to $\sqrt{n}/p \leq \lambda \leq 4\sqrt{n}\sqrt{\log p}$. Fixed values of λ, as before are included provided $p \gtrsim \sqrt{n}$.*

RECOVERY, CONDITIONS ON THE DESIGN MATRIX.

The parameter β in the model (8) is not estimable without conditions on the regression matrix when $p > n$. If β is known to be sparse, then 'local invertibility' of the Gram matrix X^tX is sufficient for estimability, even in the case $p > n$. We make this precise in the following definitions, which are variants on definitions in the literature, slightly adapted to suit to our Bayesian setup. We refer to the book by Bühlmann and van de Geer [23] for an overview of possible conditions. We follow their terminology in the sequel up to small adaptations.

Definition 2.1 (Compatibility) *The compatibility number of model $S \subset \{1, \ldots, p\}$ is*

$$\phi(S) := \inf\Big\{ \frac{\|X\beta\|_2 |S|^{1/2}}{\|X\|\,\|\beta_S\|_1} : \|\beta_{S^c}\|_1 \leq 7\|\beta_S\|_1,\ \beta_S \neq 0 \Big\}.$$

The compatibility number compares the ℓ^2-norm of the predictive vector $X\beta$ to the ℓ^1-norm of the parameter β. A model S is considered 'compatible' if $\phi(S) > 0$. It then satisfies the nontrivial inequality $\|X\beta\|_2 |S|^{1/2} \geq \phi(S)\|X\|\,\|\beta_S\|_1$. We shall see that true vectors β^0 with compatible

support S_{β^0} can be recovered from the data, uniformly in a lower bound on the size of their compatibility numbers. The number 7 has no particular interest and is used for simplicity.

The compatibility number involves the full vectors β (also their coordinates outside of S) and allows to reduce the recovery problem to sparse vectors. The next definition concerns sparse vectors only. Unlike the compatibility number it is uniform in vectors up to a given dimension.

Definition 2.2 *The compatibility number in vectors of dimension s is defined as*

$$\overline{\phi}(s) := \inf\Big\{ \frac{\|X\beta\|_2 |S_\beta|^{1/2}}{\|X\|\,\|\beta\|_1} : \quad 0 \neq |S_\beta| \leq s \Big\}$$

The smallest scaled singular value of dimension s is defined as

$$\widetilde{\phi}(s) := \inf\Big\{ \frac{\|X\beta\|_2}{\|X\|\,\|\beta\|_2} : \quad 0 \neq |S_\beta| \leq s \Big\}. \tag{2.24}$$

For recovery we shall impose that these numbers for s equal to (a multiple of) the dimension of the true parameter vector are bounded away from zero. Since $\|\beta\|_1 \leq |S_\beta|^{1/2}\|\beta\|_2$ by the Cauchy-Schwarz inequality, it follows that $\widetilde{\phi}(s) \leq \overline{\phi}(s)$, for any $s > 0$. The stronger assumptions on the design matrix imposed through $\widetilde{\phi}(s)$ will be used for recovery with respect to the ℓ^2-norm, whereas the numbers $\overline{\phi}(s)$ suffice for ℓ^1-reconstruction. The 'scaled' in Definition 2.2 refers to the scaling of the matrix X by division by the maximum column length $\|X\|$; if the latter is unity, then $\widetilde{\phi}(s)$ is just the smallest singular value of a submatrix of X of dimension s.

The final and strongest invertibility condition is in terms of 'mutual coherence' of the regression matrix, which is the maximum correlation between its columns. Equivalently, it is the ratio between entries on the diagonal of X^tX and off-diagonal entries.

Definition 2.3 *The mutual coherence number is*

$$\mathrm{mc}(X) = \max_{1\leq i\neq j\leq p} \frac{|\langle X_{\cdot,i}, X_{\cdot,j}\rangle|}{\|X_{\cdot,i}\|_2\,\|X_{\cdot,j}\|_2}.$$

We also say that X satisfies the *'(K,s) mutual coherence condition'* if $\mathrm{mc}(X)$ is bounded above by $(Ks)^{-1}$, in which case reconstruction is typically possible for true vectors β of dimension up to s. Conditions of this type have been used by many authors, following Donoho, Elad and Temlyakov (2006) [33], who coined the name. Below we use a version of the condition to obtain rates of contraction of the posterior distribution with respect to the maximum norm, similarly as in the study of [79] of the LASSO and the Dantzig estimator under the maximum norm.

For extensive discussion of the preceding and various other conditions we refer to Section 6.13 of [23]. To see that the compatibility indices are well behaved in interesting examples we note the following. In the sequence model (6) the regression matrix X is the identity, and hence the compatibility numbers are 1 and the mutual coherence number is zero, which is the optimal situation. In the response setting of Example 2.5 it is reasonable to assume that the entries of X are i.i.d. random variables. Then the mutual coherence number is with high probability bounded by a multiple of $(n/\log n)^{-1/2}$. Models up to nearly dimension $\sqrt{n}$ can then be identified from the data.

The following lemma shows that control of the mutual coherence numbers imply control of the compatibility numbers and sparse singular values. Notice that $\widetilde{\phi}(1) = \overline{\phi}(1) = \min_i \|X_{\cdot,i}\|_2/\|X\|$, as follows by evaluating the infimum in Definition 2.2 with β equal to unit vectors.

Lemma 2.1 *We have $\phi(S)^2 \geq \overline{\phi}(1)^2 - 15|S|\,\mathrm{mc}(X)$ and $\overline{\phi}(s)^2 \geq \widetilde{\phi}(s)^2 \geq \overline{\phi}(1)^2 - s\,\mathrm{mc}(X)$.*

We are ready to state the main results on the regression model. For simplicity all results are stated in limit form, for $p, n \to \infty$. We omit to recall '$p, n \to \infty$' in the following statements.

DIMENSION. We start by a result on dimension reduction, analogous to the one of Theorem 2.5.

Theorem 2.8 *In the regression model (8), let the prior $\Pi \equiv (\lambda, \pi_p)$ on β be chosen according to (2.21)-(2.22) with λ satisfying (2.23). Recall that a is the constant in (2.22). Then, with $s_0 = |S_{\beta^0}|$ and for any $M > 2$,*

$$\sup_{\beta^0} E_{\beta^0} \Pi\Big(\beta : |S_\beta| > s_0 + \tfrac{M}{a}\big(1 + \tfrac{16}{\phi(S_0)^2}\tfrac{\lambda}{\overline{\lambda}}\big)s_0 \,\big|\, Y\Big) \to 0.$$

RECOVERY. The second theorem concerns the ability of the posterior distribution to recover the true parameters from the data. It gives rates of contraction of the posterior distribution both regarding *prediction error* $\|X\beta - X\beta^0\|_2$ and regarding the parameter β relative to the ℓ^1- and ℓ^2- and ℓ^∞-distances. Besides on the dimensionality the rate depends on compatibility. Set, for 'a' the constant in (2.22),

$$\begin{aligned} \overline{\psi}(S) &= \overline{\phi}\Big(\big(2 + \tfrac{3}{a} + \tfrac{33}{\phi(S)^2}\tfrac{\lambda}{\overline{\lambda}}\big)|S|\Big), \\ \widetilde{\psi}(S) &= \widetilde{\phi}\Big(\big(2 + \tfrac{3}{a} + \tfrac{33}{\phi(S)^2}\tfrac{\lambda}{\overline{\lambda}}\big)|S|\Big). \end{aligned} \tag{2.25}$$

In the interesting case that $\lambda \ll \overline{\lambda}$, these numbers are bounded below by $\overline{\phi}((2 + \frac{4}{a})|S_\beta|)$ and $\widetilde{\phi}((2 + \frac{4}{a})|S_\beta|)$ asymptotically if $\phi(S_\beta)$ is bounded away from zero. Thus the following theorem gives rates of recovery that are uniform in true vectors β such that $\phi(S_\beta)$ and $\overline{\phi}((2 + \frac{4}{a})|S_\beta|)$ or $\widetilde{\phi}((2 + \frac{4}{a})|S_\beta|)$ are bounded away from zero.

Theorem 2.9 (Recovery) *In the regression model (8), let the prior $\Pi \equiv (\lambda, \pi_p)$ on β be chosen according to (2.21)-(2.22) with λ satisfying (2.23). Then for sufficiently large M, with $S_0 = S_{\beta^0}$,*

$$\begin{aligned} \sup_{\beta^0} E_{\beta^0} \Pi\big(\beta : \|X(\beta - \beta^0)\|_2 > \tfrac{M}{\overline{\psi}(S_0)} \frac{\sqrt{|S_0| \log p}}{\phi(S_0)} \,\big|\, Y\big) &\to 0, \\ \sup_{\beta^0} E_{\beta^0} \Pi\big(\beta : \|\beta - \beta^0\|_1 > \tfrac{M}{\overline{\psi}(S_0)^2} \frac{|S_0|\sqrt{\log p}}{\|X\|\phi(S_0)^2} \,\big|\, Y\big) &\to 0, \\ \sup_{\beta^0} E_{\beta^0} \Pi\big(\beta : \|\beta - \beta^0\|_2 > \tfrac{M}{\widetilde{\psi}(S_0)^2} \frac{\sqrt{|S_0| \log p}}{\|X\|\phi(S_0)} \,\big|\, Y\big) &\to 0. \end{aligned}$$

Furthermore, for every $c_0 > 0$, any $d_0 < c_0^2(1 + 2/a)^{-1}/8$ for a the constant in (2.22), and s_n with $\lambda s_n \sqrt{\log p}/\|X\| \to 0$, for sufficiently large M,

$$\sup_{\substack{\beta^0 : \phi(S_0) \geq c_0, \widetilde{\psi}(S_0) \geq c_0 \\ |S_0| \leq s_n, |S_0| \leq d_0 \,\mathrm{mc}(X)^{-1}}} E_{\beta^0} \Pi\big(\beta : \|\beta - \beta^0\|_\infty > M \tfrac{\sqrt{\log p}}{\|X\|} \,\big|\, Y\big) \to 0.$$

The conditions to obtain the rates $|S_0|\sqrt{\log p}/\|X\|$, $\sqrt{|S_0| \log p}$ and $\sqrt{\log p}$ for the $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty$ norms respectively in Theorem 2.9 are increasingly strong. They are in line with the conditions required for the LASSO to achieve the corresponding rates, see [23].

MODEL SELECTION. If a coefficient of the true β is large enough, one can show that it gets automatically selected by our procedure. How 'large' it should be depends on the conditions satisfied by the matrix X. If $\phi(S_0)$, $\widetilde{\psi}(S_0)$ are bounded from below, the detection threshold is of order $\sqrt{|S_0| \log p}/\|X\|$. Under a mutual coherence condition on X, the smaller threshold $\sqrt{\log p}/\|X\|$ can also be achieved, see [P15], Theorem 5.

Under some additional conditions on the growth of s_n and on λ, one can further show that S *only* consists of the coordinates which are above the threshold, that is, that the posterior distribution *consistently selects* the correct model. Let

$$\widetilde{\mathrm{B}} = \widetilde{\mathrm{B}}(M) = \Big\{\beta : \min_{i \in S_\beta} |\beta_i| \geq \tfrac{M}{\widetilde{\psi}(S)^2} \frac{\sqrt{|S_\beta| \log p}}{\|X\|\phi(S_\beta)}\Big\}. \tag{2.26}$$

Theorem 2.10 (Consistent model selection) *Let the conditions of Theorem 2.9 be satisfied and* $\widetilde{\mathrm{B}}$ *be defined in (2.26). Suppose that* $a > 1$ *in (2.22) and that* $s_n \leq p^A$ *for some* $A < a - 1$*, and* $s_n \lambda \sqrt{\log p}/\|X\| \to 0$*. Then, for* M *large enough, for every* $c_0 > 0$*,*

$$\inf_{\substack{\beta^0 \in \widetilde{\mathrm{B}}: \phi(S_0) \geq c_0 \\ |S_0| \leq s_n, \widetilde{\psi}(S_0) \geq c_0}} E_{\beta^0} \Pi\big(\beta : S_\beta = S_{\beta^0} \,|\, Y\big) \to 1.$$

Prediction. The vector $X\beta$ is the mean vector of the observation Y in (8), and one might guess that this is estimable without identifiability conditions on the regression matrix X. Next we show that the posterior distribution based on a prior of type Π above can indeed solve this *prediction problem* at (nearly) optimal rates under no condition on the design matrix X. The best results are achieved by taking a heavy tailed distribution on g. Suppose g is of the form

$$g(x) \propto \frac{\lambda}{1 + |\lambda x|^\mu}, \qquad x \in \mathbb{R}, \quad \lambda > 0, \ \mu > 3. \tag{2.27}$$

Theorem 2.11 *If* π_p *satisfies (2.22) with* $a \geq 1$*, and* g *is of the form (2.27) with* $\lambda = \|X\|$ *and* $\mu > 3$*, then for sufficiently large* M*,*

$$\sup_{\beta^0} E_{\beta^0} \Pi\big(\beta \in \mathbb{R}^p : \|X\beta - X\beta^0\|_2^2 > M \rho_n(\beta^0) \,|\, Y\big) \to 0,$$

for $\rho_n(\beta) = |S_\beta| \log p \ \vee \sum_{i \in S_\beta} \log\big(1 + \|X\|^\mu |\beta_i|^\mu\big)$.

For simplicity we have stated the result in an asymptotic fashion, but a more general oracle-type inequality result for the posterior holds as well, see [P15] Theorem 10. A similar result was obtained Dalalyan and Tsybakov [31] for point-estimators in a PAC-Bayesian framework. The rate $|S_\beta| \log p$ for the squared euclidian distance is achieved uniformly over most β's except those with some very large coefficients for which the logarithmic term in ρ_n above may become dominant. It is also possible to consider other choices of g in (2.27), such as the Laplace density considered above, but then the second term in the definition of $\rho_n(\beta)$ becomes larger (the heavier the tails, the smallest $\rho_n(\beta)$, see [P15], Theorem 10).

One may further ask whether it is possible to obtain a prediction result without condition on X uniformly in *all* βs (so, without the possible extra log term in $\rho_n(\beta)$ above). We prove in [P15], Theorem 11 that this is possible by constructing an improper prior directly on $X\beta$.

Distributional approximation. Although we shall not present this aspect in details here, we note that in the small λ regime, we also derive in [P15] a distributional approximation for the posterior distribution in the form of a BvM-type theorem.

We conclude by a few words on practical implementation. It is generally believed that simulating for posterior distributions corresponding to priors with mixtures of point masses at 0 is computationally intensive in the regression model. Although not quite as computationally fast as the LASSO, we note that recent years have seen some progresses in this direction for algorithms simulating from approximations of the posterior or aspects of it. Recent contributions on the subject include [20, 89, 91].

2.4 Supremum norm posterior rates

Motivation. The general rate Theorem 0.1 is well-suited for use in combination with specific distances: either the Hellinger distance between probabilities, or more generally distances in which some tests at exponential rate are available. Although this covers already quite an important number of applications, sometimes one has a specific distance at hand, for which building tests as

mentioned may not be straightforward. A typical example is the supremum norm on functions. Giné and Nickl (2011) [50] have shown that in density estimation it is possible to build tests with exponential-type decrease of errors using sharp concentration inequalities for certain estimators. Then one is able to apply the general rate Theorem 0.1 but the corresponding posterior rates are above the minimax rate by a polynomial factor in n for Hölder classes.

Obtaining posterior rates in L^p norms $p > 2$ is an interesting question in itself, but we mention that the question also finds applications for instance in the study of remainder terms appearing in the study of semiparametric functionals: [P13] considers some examples. On the other hand, the problem is well-studied from the frequentist minimax perspective and rates are known since a long time: a brief bibliographic review is made in [P12]. Simple frequentist estimators achieving the rate are wavelet estimators based on thresholding [54]. So, it is natural to think that sparse priors in the spirit of Section 2.2 should work. Indeed, in a recent preprint [55], Hoffmann, Rousseau and Schmidt-Hieber prove that adaptive rates in white noise are achieved by such priors. The authors also derive a number of interesting results for white noise and suggest an abstract construction for density estimation. Yet, one may think that other methods are possible, and that sparsity-inducing priors per se should not be a necessary requirement.

Here, our goal is to provide a methodology to handle the question for a given prior, not necessarily sparsity inducing (also, to our knowledge no minimax posterior sup-norm rate in density estimation were known before [P12]). In particular, Theorem 2.13 below shows that exponentially transformed Gaussian processes similar to those considered in [50] *do* achieve the minimax sup-norm convergence rate, so the method enables to achieve sharp rates, which seems not to be always the case with the testing approach. In fact, we had to build a somewhat related argument without testing in our paper [P10], see Theorem 3.7 in Chapter 3 below.

The semiparametric perspective and multiscale. The main idea comes from a multiscale analysis of f combined to a connection to semiparametrics. For a localised wavelet basis ψ_{lk}, see below, the supremum norm of a function f can be related to maxima of its wavelet coefficients $\langle f, \psi_{lk}\rangle_2$. But

$$f \to \langle f, \psi_{lk}\rangle_2$$

can be seen as a *semiparametric functional*. It is precisely one of the goals of Chapter 3 to study posterior behaviour and shape for *fixed* functionals. Here the problem is somewhat different from the semiparametric questions considered in Chapter 3 since one needs to control many functionals *simultaneously*. However, the fact that this is feasible, at least for some range of indexes l, leads to the results stated below.

Localised bases and wavelets. Wavelet basis are particularly suited here and we refer to Härdle, Kerkyacharian, Picard and Tsybakov [54] for an introduction to wavelet bases constructions and applications of wavelets in statistics.

The Haar basis on $[0,1]$ is defined by $\varphi^H(x) = 1$, $\psi^H(x) := \psi^H_{0,0}(x) = -\mathbb{1}_{[0,1/2]}(x) + \mathbb{1}_{(1/2,1]}(x)$ and $\psi^H_{l,k}(x) = 2^{l/2}\psi(2^l x - k)$, for any integer l and $0 \le k \le 2^l - 1$. The supports of Haar wavelets form dyadic partitions of $[0,1]$, corresponding to intervals $I^l_k := (k2^{-l}, (k+1)2^{-l}]$ for $k > 0$, and where the interval is closed to the left when $k = 0$. One drawback of the Haar basis is that it has non-smooth basis elements.

For wavelets on an interval, an alternative is the boundary corrected basis of Cohen, Daubechies, Vial [29], which we will refer to as CDV basis. The CDV basis enables a treatment on compact intervals and at the same time can be chosen sufficiently smooth. We adopt a double indexing as for the Haar basis and denote the CDV basis by $\{\psi_{lk}\}$.

The key localisation property shared by both bases is: for some universal $C > 0$ and any $l \ge 0$,

$$\| \sum_{k=0}^{2^l-1} |\psi_{lk}| \|_\infty \le C 2^{l/2}. \tag{2.28}$$

The basis $\{\psi_{lk}\}$ can be chosen smooth enough so that it characterises Besov spaces $B^s_{\infty,\infty}[0,1]$,

up to a given arbitrary level say S, in terms of wavelet coefficients. That is, for $s \leq S$, we have $g \in B^s_{\infty,\infty}[0,1]$ if and only if

$$\|g\|_{\infty,\infty,s} := \sup_{l\geq 0,\ 0\leq k\leq 2^l-1} 2^{l(\frac{1}{2}+s)} |\langle g, \psi_{lk}\rangle_2| < \infty. \tag{2.29}$$

We recall that $B^s_{\infty,\infty}$ coincides with the Hölder space $\mathcal{C}^s$ when s is not an integer and otherwise the inclusion $\mathcal{C}^s \subset B^s_{\infty,\infty}$ holds. If the Haar-wavelet is considered, the fact that f_0 is in $\mathcal{C}^s$, $0 < s \leq 1$, implies that the supremum in (2.29) with $\psi_{lk} = \psi^H_{lk}$ is finite.

Rates and notation. For any $\alpha > 0$ and any $n \geq 1$, denote by $\bar{\varepsilon}_{n,\alpha}$ and $\varepsilon^*_{n,\alpha}$ the rates

$$\bar{\varepsilon}_{n,\alpha} := n^{-\frac{\alpha}{2\alpha+1}}, \qquad \varepsilon^*_{n,\alpha} := \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}. \tag{2.30}$$

Let us also set, omitting the dependence in α in the notation,

$$h_n = \left(\frac{n}{\log n}\right)^{-\frac{1}{2\alpha+1}}, \qquad L_n = \lfloor \log_2(1/h_n) \rfloor. \tag{2.31}$$

Gaussian white noise model. In the white noise model (1), suppose one wants to estimate the function f with respect to the $\|\cdot\|_\infty$-loss from a Bayesian perspective. We already have natural candidate priors on functions via the prior given by (10), which assigns independent priors to the coordinates of f onto a basis.

Prior, white noise case. Consider the prior Π on functions on $[0,1]$ induced by, for a doubly-indexed orthonormal basis $\{\psi_{lk}\}$ of $L^2[0,1]$ satisfying (2.28),

$$f = \sum_{l=-1}^{\infty} \sum_{k=0}^{2^l-1} \sigma_l \zeta_{lk} \psi_{lk}, \tag{2.32}$$

where $\{\zeta_{lk}\}$ are independent indentically distributed variables with density φ with respect to Lebesgue measure on $[0,1]$, and where $\{\sigma_l\}$ is a sequence of parameters depending on l only to be specified below. Further assume that φ is strictly positive on $[-1,1]$ and that it satisfies

$$\exists\, b_1, b_2, c_1, c_2, \delta > 0,\ \forall x:\ |x| \geq 1, \quad c_1 e^{-b_1|x|^{1+\delta}} \leq \varphi(x) \leq c_2 e^{-b_2|x|^{1+\delta}}. \tag{2.33}$$

Consider a scaling σ_l for the prior equal to, for δ the constant in (2.33),

$$\sigma_l = \frac{2^{-l(\frac{1}{2}+\alpha)}}{(l+1)^\mu}, \qquad \mu = \frac{1}{1+\delta}. \tag{2.34}$$

Possible choices for φ cover several commonly used classes of prior distributions, such as exponential power distributions. The precise tuning of σ_l with the logarithmic-type term in $(l+1)^\mu$ seems necessary to get sharp rates in the next Theorem (that is, without extra logarithmic terms). Also, under (2.33) the density φ has exponential moments. This is mostly for simplicity, as moment-generating functions can then be used in the proofs. We do not think this is an essential restriction though, and results for heavier tailed priors may presumably be obtained by adapting the proof.

Statement, white noise. The next result is for fixed regularity $\alpha > 0$.

Theorem 2.12 *Let $X^{(n)}$ be observations from the white noise model (1). Suppose f_0 belongs to $B^\alpha_{\infty,\infty}[0,1]$, for some $\alpha > 0$. Let the prior Π be a product prior defined through φ and σ_l satisfying (2.33), (2.34). Then there exists $M > 0$ such that for $\varepsilon^*_{n,\alpha}$ defined by (2.30),*

$$E^n_{f_0} \int \|f - f_0\|_\infty d\Pi(f \,|\, X^{(n)}) \leq M\varepsilon^*_{n,\alpha}.$$

Theorem 2.12 can be seen as a generalisation to non-conjugate priors of Theorem 1 in [50]. Possible choices for φ cover several commonly used classes of prior distributions, such as so-called exponential power (EP) distributions. Several other choices of priors distributions are possible, up sometimes to some adaptations, see [P12].

DENSITY ESTIMATION. Consider the density estimation model (3) on the interval $[0,1]$. Let $\mathcal{F}$ be the set of densities f on $[0,1]$ which are bounded away from 0 and ∞. Let $\mathcal{F}_{\rho,D} = \{f,\ 0 < \rho \leq f \leq D < \infty,\ \int_0^1 f = 1\}$. For the following results we assume that the true function f_0 belongs to $\mathcal{F}_0 := \mathcal{F}(\rho_0, D_0)$, for some $0 < \rho_0 \leq D_0 < \infty$.

EXAMPLE: LOG-DENSITIES. Let us define a prior Π on log-densities via the exponential transformation (13). Given a sufficiently smooth CDV-wavelet basis $\{\psi_{lk}\}$, consider the prior induced on densities f by, for L_n defined in (2.31),

$$T(x) = \sum_{l=0}^{L_n} \sum_{k=0}^{2^l-1} \sigma_l \alpha_{lk} \psi_{lk}(x) \tag{2.35}$$

$$f(x) = \exp\left\{T(x) - c(T)\right\}, \quad c(T) = \log \int_0^1 e^{T(x)} dx, \tag{2.36}$$

where α_{lk} are i.i.d. random variables of density φ with respect to Lebesgue measure and σ_{lk} some positive reals. We consider the choices $\varphi(x) = \varphi_G(x) = e^{-x^2/2}/\sqrt{2\pi}$ the Gaussian density and $\varphi(x) = \varphi_H(x)$, where φ_H is any density such that its logarithm $\log \varphi_H$ is Lipschitz on $\mathbb{R}$. We refer to this last case as the 'Log-Lipschitz case'. For instance, the α_{lk}'s can be Laplace-distributed or have heavier tails, such as, for a given $0 \leq \tau < 1$ and $x \in \mathbb{R}$, and c_τ a normalising constant,

$$\varphi_{H,\tau}(x) = c_\tau \exp\{-(1+|x|)^{1-\tau}\}. \tag{2.37}$$

Suppose the prior parameters σ_l satisfy, for some $\alpha > 1/2$ and $0 < r \leq \alpha - \frac{1}{4}$,

$$\sigma_l \geq 2^{-l(\alpha+\frac{1}{2})} \ \text{(Log-Lipschitz case)}, \quad \sigma_l = 2^{-l(\frac{1}{2}+r)} \ \text{(Gaussian-case)}. \tag{2.38}$$

Typically, see below, such priors f in (2.36) under $\varphi = \varphi_G$ or φ_H and (2.38) attain the rate $\bar{\varepsilon}_{n,\alpha}$ in (2.30) in terms of Hellinger loss, up to logarithmic terms. For some $\nu > 0$, suppose

$$\Pi[f:\ h(f,f_0) > (\log n)^\nu \bar{\varepsilon}_{n,\alpha} \,|\, X^{(n)}] \to^{P^n_{f_0}} 0. \tag{2.39}$$

Theorem 2.13 *Consider observations $X^{(n)}$ in the density model (3). Suppose $\log f_0$ belongs to $\mathcal{C}^\alpha[0,1]$, with $\alpha \geq 1$. Let Π be the prior on $\mathcal{F}$ defined by (2.36), with $\varphi = \varphi_G$ or φ_H. Suppose that σ_l satisfy (2.38) and that (2.39) holds. Then, for $\alpha > 1$ and $\varepsilon^*_{n,\alpha}$ defined by (2.30), any $M_n \to \infty$, it holds, as $n \to \infty$,*

$$\Pi[f:\ \|f - f_0\|_\infty > M_n \varepsilon^*_{n,\alpha} \,|\, X^{(n)}] \to^{P^n_{f_0}} 0.$$

In the case $\alpha = 1$, the same holds with $\varepsilon^*_{n,\alpha}$ replaced by $(\log n)^\eta \varepsilon^*_{n,\alpha}$, for some $\eta > 0$.

Theorem 2.13 implies that log-density priors for many natural priors on the coefficients achieve the optimal minimax rate of estimation over Hölder spaces under sup-norm loss, as soon as the regularity is at least 1. We note that obtaining this result requires quite some work, especially for log-density priors: indeed, for those a natural semiparametric analysis is based on $f \to \langle \log f, \psi_{lk} \rangle_2$. These functionals can be shown to relate to $f \to \langle f, \zeta_{lk} \rangle_2$, with $\zeta_{lk} = \psi_{lk}/f_0$. Note in particular that $\{\zeta_{lk}\}$ do not themselves form a wavelet basis in general. That a uniform control of these previous linear functionals holds up to the desired cut-off level for the considered priors is true but non-trivial.

Let us give some examples of prior distributions satisfying the assumptions of Theorem 2.13. In the Gaussian case, any sequence of the type $\sigma_l = 2^{-l(\frac{1}{2}+\gamma)}$ with $0 < \gamma \leq \alpha - 1/4$ satisfies both

(2.38) and (2.39). In the Log-Lipschitz case, the choice $\varphi = \varphi_{H,\tau}$ in (2.37) with any $0 \leq \tau < 1$ combined with $\sigma_l = 2^{-l\alpha}$ satisfies (2.38)-(2.39). Both claims follow from minor adaptations of Theorem 4.5 in [100] and Theorem 2.1 in [88] respectively.

EXAMPLE: RANDOM HISTOGRAMS. Associated to the regular dyadic partition of $[0,1]$ at level $L \in \mathbb{N}^*$, given by $I_0^L = [0, 2^{-L}]$ and $I_k^L = (k2^{-L}, (k+1)2^{-L}]$ for $k = 1, \ldots, 2^L - 1$, is a natural notion of histogram

$$\mathcal{H}_L = \{h \in L^\infty[0,1], \quad h(x) = \sum_{k=0}^{2^L-1} h_k \mathbb{1}_{I_k^L}(x), \qquad h_k \in \mathbb{R},\ k = 0, \ldots, 2^L - 1\}$$

the set of all histograms with 2^L regular bins on $[0,1]$. Let $\mathcal{S}_L = \{\omega \in [0,1]^{2^L}; \sum_{k=0}^{2^L-1} \omega_k = 1\}$ be the unit simplex in $\mathbb{R}^{2^L}$. Let $\mathcal{H}_L^1$ be the subset of $\mathcal{H}_L$ consisting of histograms which are *densities* on $[0,1]$. Let $\mathcal{H}^1$ be the set of all histograms which are densities on $[0,1]$.

A simple way to specify a prior on $\mathcal{H}_L^1$ is to set $L = L_n$ deterministic and to fix a distribution for $\omega_L := (\omega_0, \ldots, \omega_{2^L-1})$. Set $L = L_n$ as defined in (2.31). Choose some fixed constants $a, c_1, c_2 > 0$ and let

$$L = L_n, \qquad \omega_L \sim \mathcal{D}(\alpha_0, \ldots, \alpha_{2^L-1}), \quad c_1 2^{-La} \leq \alpha_k \leq c_2, \tag{2.40}$$

for any admissible index k, where $\mathcal{D}$ denotes the Dirichlet distribution on the simplex S_L.

Theorem 2.14 *Let $f_0 \in \mathcal{F}_0$ and suppose f_0 belongs to $\mathcal{C}^\alpha[0,1]$, where $1/2 < \alpha \leq 1$. Let Π be the prior on $\mathcal{H}^1 \subset \mathcal{F}$ defined by (2.40). Then, for $\varepsilon_{n,\alpha}^*$ defined by (2.30) and any $M_n \to \infty$ it holds, as $n \to \infty$,*

$$\Pi[f : \ \|f - f_0\|_\infty > M_n \varepsilon_{n,\alpha}^* \,|\, X^{(n)}] \to^{P_{f_0}^n} 0.$$

According to Theorem 2.14, random dyadic histograms achieve the minimax rate in sup-norm over Hölder balls. Condition (2.40) is quite mild. For instance, the uniform choice $\alpha_0 = \cdots = \alpha_{2^L-1} = 1$ is allowed, as well as a variety of others, for instance one can take $\alpha_k = \alpha_{k,L_n}$ to originate from a measure $A = A_{L_n}$ on the interval $[0,1]$, of finite total mass $\bar{A}_{L_n} := A([0,1])$. By this we mean $\alpha_k = A(I_k^{L_n})$. If $A/\bar{A}_{L_n}$ has say a fixed continuous and positive density a with respect to Lebesgue measure on $[0,1]$, then (2.40) is satisfied as soon as there exists a $\delta > 0$ with $2^{-\delta L_n} \lesssim \bar{A}_{L_n} \lesssim 2^{L_n}$.

FURTHER EXAMPLES. The scope of the described technique is not restricted to the previous examples. In particular, in a forthcoming work we prove that a class of Pólya trees density priors achieves supremum norm minimax rates for density estimation.

2.5 Perspectives

We have presented several posterior convergence rate results in nonparametric frameworks. Some relate to families of prior distributions such as heat kernel Gaussian process priors, prior for sparse objects etc. Other families of prior distributions are of particular interest. One example we have in mind is the class of mixtures. As briefly mentioned in the Introduction, mixture of kernels for instance arise naturally as priors on density functions. Such priors appear to be very flexible: for instance, Judith Rousseau [90] has shown that certain random mixtures of Beta kernels achieve Hellinger-posterior convergence rate that are adaptive over Hölder classes, up to a logarithmic factor, without restrictions due to the order of the kernel. It would be interesting to develop a broader theory for mixture priors, especially for more complex models, and possibly in terms of different loss functions. Also, in the present manuscript, we mostly focus on priors on functions and vectors, but some families of priors on complex objects such are matrices or graphs are presently very much developped in probability theory or in machine learning, and it seems natural to ask about their statistical properties, in particular their posterior convergence rates.

The question of obtaining posterior rates with respect to strong measures of loss such as the supremum norm appears to be quite challenging in general: the programme presented above to study those relates it to the question of obtaining Bernstein-von Mises results (uniformly) for functionals, as we discuss in Chapter 3. Also, in this case, the role of the prior on high frequencies may be even more important than for other losses. Posterior adaptation with respect to strong losses further rises a variety of interesting questions, and investigation of some of these is under current investigation.

Finally, refinements of the notion of posterior rates are very desirable. Two directions appear naturally: providing oracle-type results for rates in the spirit of oracle inequalities (though not explicitly presented here we obtain some results in this direction for sparse priors in [P15]); developping nonasymptotic results, ressembling for instance the one of Theorem 2.7: [17] is a recent interesting step in this direction.

CHAPTER 3

Limiting shape of posterior distributions

We derive Bernstein-von Mises (BvM) theorems in infinite-dimensional settings. First, we present an approach to the semiparametric BvM theorem for separated models based on [P7]. Some examples are then presented in details. We briefly mention some extensions [P13] and a counterexample [P8]. Second, we define a notion of nonparametric BvM in nonparametric models [P10, P14]. From it we derive several applications, such as Bayesian Donsker theorems in nonconjugate settings and the construction of nonparametric confident credible sets.

3.1 Semiparametric BvM for separated models

It is natural to ask whether the BvM Theorem 0.2 admits a counterpart in semiparametric models. Such a result is very desirable, as it implies asymptotic normality of the marginal posterior distribution for the parameter of interest, and immediately yields asymptotic confidence sets.

Few semiparametric BvM results are available in the literature. Kim and Lee [63] and Kim [62] obtain BvM results in the proportional hazards for the cumulative hazard function and the parameter in Cox' model, with a specific class of Lévy processes as priors and using that this class is partly conjugate to the model. Shen [94] states quite general results, but a few of his conditions are fairly implicit and may not be easy to check in practice. Bickel and Kleijn [10] provide a set of conditions for i.i.d. data, assuming $\sqrt{n}$-consistency for the posterior of the parameter of interest. Rivoirard and Rousseau [88] obtain a general semiparametric BvM result in the density model for linear functionals of the density. We first present the results obtained in [P7], and in the next Section we briefly discuss [P13], which is a natural continuation of [P7] and [88].

NOTATION. Let us consider a sequence of statistical experiments $(\mathcal{X}^{(n)}, \mathcal{G}^{(n)}, P_\eta^{(n)},\ \eta \in \mathcal{E})$, with observations $X^{(n)}$, where $\mathcal{E}$ is a parameter set of the form $\Theta \times \mathcal{F}$ with Θ an interval of $\mathbb{R}$ – the results extend without much effort to $\mathbb{R}^k$, but we keep $k = 1$ for simplicity – and $\mathcal{F}$ a subset of a separable Banach space. The true value of the parameter is denoted by $\eta_0 = (\theta_0, f_0)$ and is assumed to be an interior point of $\mathcal{E}$. We assume that the measures $P_\eta^{(n)}$ admit densities $p_\eta^{(n)}$ with respect to a σ-finite measure $\mu^{(n)}$. The log-likelihood is denoted $\ell_n(\eta)$,

$$\ell_n(\eta) = \log p_\eta^{(n)}, \quad \text{and} \quad \Lambda_n(\eta) = \ell_n(\eta) - \ell_n(\eta_0).$$

The space $\mathcal{E} = \Theta \times \mathcal{F}$ is equipped with a product σ-field $\mathcal{T} \otimes \mathcal{B}$ and we assume that $(x, \eta) \to p_\eta^{(n)}(x)$ is jointly measurable. Note that the data does not have to be i.i.d.

PRIOR, CONDITION (P). We put a prior Π on the pair (θ, f), of the form $\Pi = \pi_\theta \otimes \pi_f$. For π_θ we choose any probability measure on Θ having a density λ with respect to Lebesgue measure on Θ, with λ *positive* and *continuous* at the point θ_0.

BAYES' FORMULA AND NEIGHBORHOODS. The neighborhoods B_{KL} from (15) have natural analogues in the semiparametric, possibly non-i.i.d., context. For any $\varepsilon > 0$, define

$$B_{KL,n}(\eta_0, \varepsilon) = \{\eta \in \mathcal{E} : K(P_{\eta_0}^{(n)}, P_\eta^{(n)}) \leq n\varepsilon^2, \quad V(P_{\eta_0}^{(n)}, P_\eta^{(n)}) \leq n\varepsilon^2\}. \tag{3.1}$$

It is also useful to introduce as a technical tool $\Pi^{\theta=\theta_0}(\cdot \,|\, X^{(n)})$, the posterior distribution in the model where θ is known to be equal to θ_0 and one takes π_f as prior on f. By Bayes' theorem, for any $B \in \mathcal{B}$,

$$\Pi^{\theta=\theta_0}(B|X^{(n)}) = \frac{\int_B p_{\theta_0,f}^{(n)}(X^{(n)}) d\pi_f(f)}{\int p_{\theta_0,f}^{(n)}(X^{(n)}) d\pi_f(f)}.$$

We also define a neighborhood of f_0 in $\mathcal{F}$ restricted to the case $\theta = \theta_0$ as

$$B_{KL,n}^{\theta=\theta_0}(f_0, \varepsilon) = \{f \in \mathcal{F} : K(P_{\eta_0}^{(n)}, P_{\theta_0,f}^{(n)}) \leq n\varepsilon^2, \quad V(P_{\eta_0}^{(n)}, P_{\theta_0,f}^{(n)}) \leq n\varepsilon^2\}. \tag{3.2}$$

A SPECIFIC SEMIPARAMETRIC FRAMEWORK. A natural way to study efficiency in a semiparametric model is to study estimation along a maximal collection of 1-dimensional paths locally around the true parameter, as explained for instance in [98], Chap. 25, where the i.i.d. case is considered. The paper by McNeney and Wellner (2000) [82] develops similar tools in non-i.i.d. situations. Likelihood ratios along paths may then for instance be well approximated by the likelihood ratios for a Gaussian shift experiment, which leads to the notion of local asymptotic normality (LAN). The approach we follow is closely related, and assumes for simplicity that linear approximations to a given true (θ_0, f_0) belong to the model. We describe this setting precisely now.

Given a true $\eta_0 = (\theta_0, f_0)$ in $\mathcal{E}$, for any $\eta = (\theta, f)$ in $\mathcal{E}$ (possibly restricted to a subset of $\mathcal{E}$, possibly close enough to η_0), let us assume that the pair $(\theta - \theta_0, f - f_0)$ can be embedded in a product Hilbert space of the form $\mathcal{V}_{\eta_0} = \mathbb{R} \times \mathcal{G}_{\eta_0}$ equipped with an inner-product $\langle\ ,\ \rangle_L$ with associated norm $\|\cdot\|_L$. Locally around the true parameter, we shall compare the log-likelihood differences to a quadratic term plus a stochastic term. We set

$$R_n(\theta, f) = \Lambda_n(\theta, f) + n\|\theta - \theta_0, f - f_0\|_L^2/2 - \sqrt{n} W_n(\theta - \theta_0, f - f_0), \tag{3.3}$$

where $\|(h, a)\|_L$ is denoted $\|h, a\|_L$ for simplicity, and where

- $\Lambda_n(\theta, f) = \ell_n(\theta, f) - \ell_n(\theta_0, f_0)$ is the difference of log-likelihoods between the points (θ, f) and (θ_0, f_0).
- $\{W_n(v),\ v \in \mathcal{V}_{\eta_0}\}$ is a collection of random variables, measurable with respect to the observations $X^{(n)}$ and satisfying the following properties
  - ◇ For any $v_1, \ldots, v_d$ in $\mathcal{V}_{\eta_0}$, the d-tuple $(W_n(v_1), \ldots, W_n(v_d))$ converges in distribution to the d-dimensional centered Gaussian distribution with covariance structure given by the matrix $(\langle v_i, v_j \rangle_L)_{1 \leq i,j \leq d}$.
  - ◇ The map $v \to W_n(v)$ is linear.

Below we will further assume that one has a form of uniform control of the R_n's over (sieved) shrinking neighborhoods of the true η_0, see assumptions **(N)** and $(\boldsymbol{\mathcal{N}})$ below.

The inner-product and the stochastic term introduced above are often identified from LAN-type expansions. For instance, one might be in a situation where the model is LAN with linear paths (see e.g. [82]) in that for each $v = (s, g) \in \mathcal{V}_{\eta_0}$, as $n \to \infty$,

$$\Lambda_n(\theta_0 + s/\sqrt{n}, f_0 + g/\sqrt{n}) = -\|s, g\|_L^2/2 + W_n(s, g) + o_{P_{\eta_0}^{(n)}}(1), \tag{3.4}$$

where $\|\cdot\|_L$, W_n and $\mathcal{V}_{\eta_0}$ are as above. To define the notions of information and efficiency in our model, we assume for simplicity that the considered model is LAN with linear paths, which falls in the framework considered in [82], so we can borrow from that paper the definitions and their implications for efficiency. In fact, assumption (3.4) is essentially weaker than the *uniform* type of control on $R_n(\theta, f)$ required below. All examples considered in the sequel admit such a LAN expansion, at least for a well-chosen parametrisation of the model.

Semiparametric structure. Here we define the notions of least favorable direction and efficient Fisher information following [82]. Let $\overline{\mathcal{F}}$ be the closure in $\mathcal{V}_{\eta_0}$ of the linear span of all elements of the type $(0, f - f_0)$, where f belongs to $\mathcal{F}$. Let us define the element $(0, \gamma(\cdot)) \in \overline{\mathcal{F}}$ as the orthogonal projection of the vector $(1, 0)$ onto the closed subspace $\overline{\mathcal{F}}$. The element γ is called *least favorable direction*. For any $(s, g) \in \mathcal{V}_{\eta_0}$, one has the following decomposition

$$\|s, g\|_L^2 \quad = \quad (\|1, 0\|_L^2 - \|0, \gamma\|_L^2)s^2 + \|0, g + s\gamma\|_L^2. \tag{3.5}$$

The coefficient of s^2 is called *efficient Fisher information* and is denoted by $\tilde{I}_{\eta_0} = \|1, 0\|_L^2 - \|0, \gamma\|_L^2$. If γ is zero, we say there is *no loss of information* and denote the information simply by I_{η_0}. Note also that since $\|\cdot\|_L$ is a norm, $I_{\eta_0} = \|1, 0\|_L^2$ is always nonzero. In that case it can be checked that this information I_{η_0} equals the information in the model where f would be known (that is, the standard Fisher information). If $\tilde{I}_{\eta_0}$ itself is nonzero, let us also denote

$$\Delta_{n,\eta_0} = \tilde{I}_{\eta_0}^{-1} W_n(1, -\gamma).$$

An estimator $\hat{\theta}_n$ of θ_0 is said asymptotically linear and efficient if $\sqrt{n}(\hat{\theta}_n - \theta_0) = \Delta_{n,\eta_0} + o_{P_{\eta_0}^{(n)}}(1)$.

When an approximation such as (3.4) holds, the model asymptotically looks like a Gaussian shift experiment with inner-product $\langle \cdot, \cdot \rangle_L$. How much *information* is available for estimating a given parameter is completely encoded in the inner-product. Observe that from (3.5), one deduces $\|s, g\|_L^2 \geq (\|1, 0\|_L^2 - \|0, \gamma\|_L^2)s^2 = \tilde{I}_{\eta_0} s^2$ with equality when $g = -s\gamma$. The quantity $\tilde{I}_{\eta_0}$ represents the 'smallest curvature' of paths approaching η_0.

The case without loss of information. We first state a result applying when there is no loss of information, that is when $\gamma = 0$. If $\gamma = 0$, it holds $\|h, a\|_L^2 = \|h, 0\|_L^2 + \|0, a\|_L^2$ and $I_{\eta_0} = \tilde{I}_{\eta_0} = \|1, 0\|_L^2$ as well as $\Delta_{n,\eta_0} = W_n(1, 0)/\|1, 0\|_L^2$.

Concentration **(C)**. Let $\varepsilon_n \to 0$ be a sequence such that $n\varepsilon_n^2 \to \infty$. The statistical model and the prior Π satisfy condition **(C)** with rate ε_n if there exists a sequence of measurable sets $\mathcal{F}_n$ in $\mathcal{F}$ such that, as $n \to \infty$, in $P_{\eta_0}^{(n)}$-probability,

$$\Pi\left(\{\eta \in \Theta \times \mathcal{F}_n, \quad \|\eta - \eta_0\|_L \leq \varepsilon_n\} \mid X^{(n)}\right) \quad \to \quad 1,$$

$$\Pi^{\theta=\theta_0}\left(\{f \in \mathcal{F}_n, \quad \|0, f - f_0\|_L \leq \varepsilon_n/\sqrt{2}\} \mid X^{(n)}\right) \quad \to \quad 1.$$

Local shape **(N).** Let R_n be defined by (3.3) and let ε_n and $\mathcal{F}_n$ be as in **(C)**. Let us denote $V_n = \{(\theta, f) \in \Theta \times \mathcal{F}_n, \quad \|\theta - \theta_0, f - f_0\|_L \leq \varepsilon_n\}$. The model satisfies **(N)** with rate ε_n over the sieve $\mathcal{F}_n$ if

$$\sup_{(\theta,f)\in V_n} \frac{|R_n(\theta, f) - R_n(\theta_0, f)|}{1 + n(\theta - \theta_0)^2} \quad = \quad o_{P_{\eta_0}^{(n)}}(1).$$

Theorem 3.1 *Let us assume that the prior Π on (θ, f) satisfies* **(P)** *and that the model and prior verify conditions* **(C)**, **(N)**. *Suppose there is no information loss. Then it holds, as $n \to \infty$,*

$$\sup_B \left|\Pi\left(B \times \mathcal{F} \mid X^{(n)}\right) - N\left(\theta_0 + \frac{1}{\sqrt{n}}\Delta_{n,\eta_0}, \frac{1}{n} I_{\eta_0}^{-1}\right)(B)\right| \to 0,$$

in $P_{\eta_0}^{(n)}$-probability, where the supremum is taken over all measurable sets B in Θ. In words, the total variation distance between the marginal in θ of the posterior distribution and a Gaussian distribution centered at $\theta_0 + \frac{1}{\sqrt{n}}\Delta_{n,\eta_0}$, of variance $\frac{1}{n}I_{\eta_0}^{-1}$, converges to zero, in $P_{\eta_0}^{(n)}$-probability.

Condition **(C)** means that the posterior concentrates at ε_n-rate around the true η_0 in terms of $\|\cdot\|_L$. This is an Hilbert-norm, often corresponding to a weighted L^2-space: often, one may apply the general rate Theorem 0.1. Sometimes this does not suffice and one may have to devise a specific argument, or apply a different technique, such as the sup-norm rate approach from Chapter 2. Condition **(N)** controls how much the likelihood ratio differs locally from the one of a Gaussian experiment and is studied below. Note that assumptions **(P)** -the parametric part of the prior must charge θ_0-, **(N)** -which is about the shape of the model- and **(C)** -which enables us to localize in a neighborhood of the true η_0- compare in their spirit to the assumptions for the *parametric* Bernstein-von Mises theorem as stated in Le Cam and Yang [72], §7.3, Prop.1. One can also note that Theorem 3.1 actually yields a result in the particular case where f is known. In this case, the Theorem implies that if posterior concentration occurs at rate $\varepsilon_n = M_n n^{-1/2}$ for some $M_n \to \infty$ (for instance $M_n = \log n$ say) and if the uniform LAN property **(N)** in θ holds in that neighborhood of size $M_n n^{-1/2}$ then the parametric BVM theorem holds.

The case with information loss. Here we shall restrict our investigations to Gaussian priors for π_f. More precisely suppose π_f is the distribution associated to a centered Gaussian process taking its values almost surely in a separable Banach space $\mathbb{B}$. Let $\mathbb{H}$ be the Reproducing Kernel Hilbert Space of the Gaussian process. We shall assume that the space $\mathbb{H}$ is 'large enough' so that the least favorable direction γ above can be approximated by elements of $\mathbb{H}$. Suppose that there exists $\rho_n \to 0$ and a sequence γ_n of elements in $\mathbb{H}$ such that for all n, $\gamma_n - \gamma$ belongs to $\mathcal{G}_{\eta_0}$ and

$$\|\gamma_n\|_{\mathbb{H}}^2 \leq 2n\rho_n^2 \quad \text{and} \quad \|0, \gamma_n - \gamma\|_L \leq \rho_n. \tag{3.6}$$

Concentration **($\mathcal{C}$)**. The model verifies condition **($\mathcal{C}$)** with rate ε_n if there exists a sequence of measurable sets $\mathcal{F}_n$ in $\mathcal{F}$ such that, if $\mathcal{F}_n(\theta) = \mathcal{F}_n + (\theta - \theta_0)\gamma_n$,

$$\Pi\left(\{\eta \in \Theta \times \mathcal{F}_n, \quad \|\eta - \eta_0\|_L \leq \varepsilon_n\} \mid X^{(n)}\right) \quad \to \quad 1,$$

$$\inf_{|\theta-\theta_0|\tilde{I}_{\eta_0}^{1/2} \leq \varepsilon_n} \Pi^{\theta=\theta_0}\left(\{f \in \mathcal{F}_n(\theta), \quad \|0, f - f_0\|_L \leq \varepsilon_n/2\} \mid X^{(n)}\right) \quad \to \quad 1,$$

as $n \to \infty$, in $P_{\eta_0}^{(n)}$-probability. Suppose the neighborhoods in (3.1)-(3.2) verify, for some $c, d > 0$

$$\Pi(B_{KL,n}(\eta_0, d\varepsilon_n)) \geq e^{-cn\varepsilon_n^2} \quad \text{and} \quad \pi_f(B_{KL,n}^{\theta=\theta_0}(f_0, d\varepsilon_n)) \geq e^{-cn\varepsilon_n^2}.$$

Local Shape **($\mathcal{N}$)**. Let $V_n = \{(\theta, f) \in \Theta \times \mathcal{F}_n, \quad \|\theta - \theta_0, f - f_0\|_L \leq 2\varepsilon_n\}$. Assume that for any (θ, f) in V_n, the function $f - (\theta - \theta_0)\gamma_n$ belongs to $\mathcal{F}$ and that

$$\sup_{(\theta,f)\in V_n} \frac{|R_n(\theta, f) - R_n(\theta_0, f - (\theta - \theta_0)\gamma_n)|}{1 + n(\theta - \theta_0)^2} \quad = \quad o_{P_{\eta_0}^{(n)}}(1).$$

Our last assumption is related to how well the least favorable direction γ is approximated by elements of $\mathbb{H}$. As $n \to \infty$ suppose, with ε_n as in **($\mathcal{C}$)** and ρ_n, γ_n as in (3.6),

$$\textbf{($\mathcal{E}$)} \qquad \sqrt{n}\varepsilon_n\rho_n = o(1) \quad \text{and} \quad W_n(0, \gamma - \gamma_n) = o_{P_{\eta_0}^{(n)}}(1).$$

Theorem 3.2 *Let us assume that the prior $\Pi = \pi_\theta \otimes \pi_f$ on (θ, f) satisfies* **(P)**, *that π_f is a Gaussian prior, that $\tilde{I}_{\eta_0} > 0$ and that the least favorable direction γ can be approximated according to (3.6). Suppose that conditions* **($\mathcal{C}$)**, **($\mathcal{N}$)** *and* **($\mathcal{E}$)** *are satisfied. Then it holds*

$$\sup_B \left|\Pi\left(B \times \mathcal{F} \mid X^{(n)}\right) - N\left(\theta_0 + \frac{1}{\sqrt{n}}\Delta_{n,\eta_0}, \frac{1}{n}\tilde{I}_{\eta_0}^{-1}\right)(B)\right| \to 0,$$

as $n \to \infty$, in $P_{\eta_0}^{(n)}$-probability, where the supremum is taken over all measurable sets B in Θ.

The assumptions are similar in nature to the ones of Theorem 3.1, with additional requirements about the least favorable direction γ and Gaussianity of π_f. Note that if γ happens to belong to the RKHS $\mathbb{H}$ of π_f, then (3.6) and $(\boldsymbol{\mathcal{E}})$ are trivially satisfied.

Discussion of the assumptions. Assumptions (**C**)–$(\boldsymbol{\mathcal{C}})$ ask for the posterior to contract around η_0 at a preliminary rate ε_n. This is quite reasonable, especially since this rate does not have to be the 'best possible' rate. Of course the faster ε_n, the easier to check assumptions (**N**)–$(\boldsymbol{\mathcal{N}})$ become. These last assumptions control how far the model is from Gaussianity asymptotically. Without assuming (local, asymptotic) Gaussianity somewhere, one cannot of course hope to converge towards a normal distribution as in the above Theorems. The control has to be uniform in both θ and f, which may sometimes be a relatively strong requirement, although the assumptions provide the flexibility of checking it only on a *sieve* $\mathcal{F}_n$.

Assumption $(\boldsymbol{\mathcal{E}})$ is really what makes the difference between both cases (with or without information loss). One can make the following comments on it

- It is typically not too difficult to check, especially in models where an explicit expression of γ is available. One may note that even a qualitative knowledge of γ can possibly be enough: one only needs to know how well γ can be approximated by elements of $\mathbb{H}$.
- However, this assumption is often the most restrictive of the three. What typically matters are the respective 'regularities' of f_0 and γ, which in turn determine the rates ε_n and ρ_n. Consider the simple case where say $\varepsilon_n = \rho_n$. The condition becomes $\sqrt{n}\varepsilon_n^2 = o(1)$, which means the rate ε_n should not be too slow. In particular, taking a Gaussian prior π_f that *oversmooths* too much can easily destroy the condition. Indeed, we have seen in Chapter 1 that the convergence rate for Gaussian processes drops quickly to very slow rates in the oversmoothing case.
- One cannot avoid such a condition in general. In the following we consider an example where the BvM theorem does not to hold for some priors in most of the zone determined by the condition. Also, the condition will also arise for other priors than Gaussian.

If one compares the conditions here with sets of sufficient conditions for semiparametric frequentist point estimators to be efficient, see e.g. [98]-Chapter 25.8, one notes the presence of a so-called 'no-bias' condition, e.g. (25.52) in [98], which in terms of rates may lead to conditions such as $\sqrt{n}\varepsilon_n^2 \to 0$ as above. By analogy we talk for $(\boldsymbol{\mathcal{E}})$ of a *no-bias condition*. This terminology will be further justified below.

An attractive aspect of $(\boldsymbol{\mathcal{E}})$ is that, while it is certainly not always sharp, it is completely explicit in terms of prior and model. It shows well which aspects are at stake for solving the semiparametric problem: there is an interaction between 'fairly good estimation of the nonparametric part f_0' and 'estimation of the least favorable direction' (in the i.i.d. case, the latter is closely related to the so-called efficient score function). In other words, good estimation of the nuisance part is not the only requirement. In fact, one can show that adaptive priors (with respect to estimation of f) may not verify the BvM theorem, precisely because they perform poorly on the least-favorable direction (or efficient score) part.

Example: Translation model. Consider the translation model (9). To ensure identifiability, one assumes that θ belongs to $\Theta = [-\tau_0, \tau_0] \subset]-1/4, 1/4[$, a set of diameter smaller than $1/2$.

Let $\mathcal{F}$ be the linear space of all *symmetric* square-integrable functions $f : [-1/2, 1/2] \to \mathbb{R}$ and, for simplicity in the definitions of the classes of functions below, such that $\int_0^1 f(u)du = 0$. We extend any $f \in \mathcal{F}$ by 1-periodicity and denote its real Fourier coefficients by $f_k = \sqrt{2}\int_0^1 f(u)\cos(2\pi k u)du$, $k \geq 1$. Note that we can still denote by f_0 the true function f. Let us denote $\varepsilon_k(\cdot) = \cos(2\pi k \cdot)$ for $k \geq 0$. Also, let $\|\cdot\|_2$ be the L^2-norm over $[-1/2, 1/2]$.

TRANSLATION MODEL: SMOOTHNESS CONDITIONS. A function $f = (f_k)_{k\geq 1}$ is said to fulfill conditions **(R)** if there exist reals $\rho > 0$, $L > 0$ and $\beta > 1$ such that $|f_1| \geq \rho$ and $\sum_{k\geq 1} k^{2\beta} f_k^2 \leq L^2$.

TRANSLATION MODEL: LAN-NORM. It can be checked that if f_0 satisfies **(R)**, the model is LAN with linear paths and with LAN-norm the Hilbert norm given by, for a real h and a square integrable 1-periodic $a(\cdot)$,

$$\|h, a\|_L^2 = \Big(\int_{-1/2}^{1/2} f_0'(u)^2 du \Big) h^2 + \int_{-1/2}^{1/2} a(u)^2 du.$$

We see that the norm 'splits' in two independent parts, one for the parametric component of interest represented by h, and the other for the nuisance part. From this it can be deduced that there is no information loss in this model.

TRANSLATION MODEL: PRIOR AND STATEMENT. For the parametric part π_θ of the prior, let us choose the uniform measure on $[-1/4, 1/4]$. The nonparametric part π_f is a family of Gaussian priors parameterised by a real parameter α similar to (1.8). Let $\{\nu_k\}_{k\geq 1}$ be a sequence of independent standard normal random variables and for any $k > 0$ and $\alpha > 1$, let $\sigma_k = k^{-1/2-\alpha}$. The prior π_f^α is the distribution generated by

$$f(\cdot) = \sum_{k=1}^{\infty} \sigma_k \nu_k \varepsilon_k(\cdot). \tag{3.7}$$

One may also define a prior truncated at $k(n)$, a strictly increasing sequence of integers. In that case the entropy bounds involved to get posterior convergence are easier to obtain. This also explains why for this prior the domain where the BVM-theorem holds is slightly larger in the following theorem. In both cases α can be seen as the 'regularity' of the prior, as opposed to the (unknown) regularity β of f_0.

Theorem 3.3 *Suppose that f_0 satisfies* **(R)** *with regularity $\beta > 1$. Let the prior π_θ satisfy* **(P)** *and let π_f be defined by (3.7) for some $\alpha > 1$. Then conditions* **(C)** *and* **(N)** *of Theorem 3.1 are satisfied for pairs (β, α) such that the corresponding point in Figure 3.1 lies in the shaded area. In particular, the BvM theorem holds in this region. For the prior $\pi_{f,k(n)}^\alpha$ with $k(n) = \lfloor n^{1/(2\alpha+1)} \rfloor$, the same holds in the region delimited by the 'triangle'-curve.*

The region for which $\beta > 1$ and $\alpha > 1$ delimited by the 'square'-curve in Figure 3.1 can be regarded as the 'best possible' region, since it describes true functions and priors which have at least one derivatives in a weak (L^2-) sense. This condition on β is necessary to have a finite Fisher information, which here equals $\|f_0'\|_2^{-1}$. Thus with this respect the results of Theorem 3.3 are quite sharp, in that only a small strip in the region where α or β are very close to 1 is not covered by Theorem 3.1. More precisely, the region where BVM holds is defined by $\alpha > 1 + \sqrt{3}/2$ (resp. $\alpha > 3/2$ for the truncated prior), $\beta > 3/2$ and, finally, $\alpha < (3\beta - 2)/(4 - 2\beta)$, which corresponds to the non-linear curve in Figure 3.1. These mild conditions arise when checking **(N)**.

For instance for $\beta = 2$, any prior of the type (1.8) with $\alpha > 1 + \sqrt{3}/2$ will do. This means also that in model (9) for $\beta \geq 2$, *no condition on the nonparametric concentration rate ε_n of the posterior is needed* to get the semiparametric BVM theorem. It can be easily seen that if $\beta = 2$ and α increases, ε_n becomes slower and slower (in fact if $\beta = 2$ and $\alpha \geq 2$, then ε_n can be as slow as $n^{-2/(2\alpha+1)}$, as we have seen in Chapter 1).

One could consider extending the results of Theorem 3.3 to other families of priors, for instance non-Gaussian series, such as truncated series of weighted Laplace laws on the Fourier coefficients. For infinite Laplace series, similar results may hold as well, but this is possibly much harder, as the heavier tails make the construction of sieves more delicate.

EXAMPLE: COX'S PROPORTIONAL HAZARDS MODEL. The observations are a random sample from the distribution of the variable (T, δ, Z), where $T = X \wedge Y$, $\delta = \mathbb{1}_{X \leq Y}$, for some real-valued

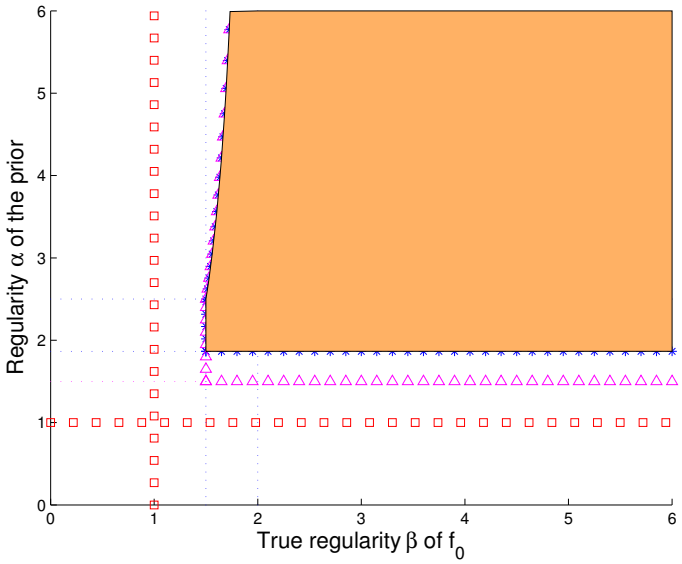

Figure 3.1: Translation model. Possible choices for π_f

random variables X, Y and Z. We assume that the variable Z, called covariate, is bounded by M and admits a continuous density φ with respect to Lebesgue's measure on $[-M, M]$. Suppose that given Z, the variables X and Y are independent and that there exists a real $\tau > 0$ such that $P_\eta(X > \tau) > 0$ and $P_\eta(Y \geq \tau) = P_\eta(Y = \tau) > 0$. The conditional hazard function α of X knowing Z is defined by $\alpha(x)dx = P_\eta(X \in [x, x + dx] \mid X \geq x,\ Z)$. The Cox model assumes that $\alpha(x) = e^{\theta Z}\lambda(x)$, where λ is an unknown hazard function and θ a real parameter. For notational simplicity, we have assumed that Z and θ are one-dimensional.

Let us assume that λ_0 is continuous and that there exists a $\rho > 0$ such that, for all x in $[0, \tau]$, one has $\lambda_0(x) \geq \rho > 0$. We denote $\Lambda(x) = \int_0^x \lambda(u)du$. We also assume that Y given $Z = z$ admits a continuous density g_z with respect to Lebesgue's measure on $[0, \tau)$ with distribution function G_z and that there exists $\rho' > 0$ such that $g_z(t) \geq \rho'$ for almost all z, t. Finally we assume that the possible values of θ lie in some compact interval $[-\theta_M, \theta_M]$.

Cox' model: LAN norm. In this semiparametric framework the unknown parameter is the pair (θ, λ). Equivalently under the above assumptions one may consider instead $\eta = (\theta, r)$, where we denote $r := \log \lambda$. Indeed, some calculations show that the Cox model fulfills the LAN property with linear paths when parameterised by $\eta = (\theta, r)$, with LAN Hilbert norm equal to, for a real h and a square integrable function a on $[0, 1]$,

$$\|h, a\|_L^2 = \int_0^\tau \{h^2 M_2(u) + 2ha(u)M_1(u) + a(u)^2 M_0(u)\}d\Lambda_0,$$

where $M_i(u) = E_{\eta_0}(\mathbf{1}_{u \leq T} Z^i e^{\theta_0 Z})$, for any $u \in [0, \tau]$ and any integer i. One can check that there is a *loss of information* in this model and that the least favorable direction $\gamma(\cdot)$ has a simple expression in terms of the functions M_i above. Namely, $\gamma = M_1/M_0$. Explicitly, for any $u \in [0, \tau]$,

$$\gamma(u) = \frac{M_1}{M_0}(u) = \frac{\int_0^\tau (1 - G_z(u-))ze^{\theta_0 z - \Lambda_0(u)e^{\theta_0 z}}\varphi(z)dz}{\int_0^\tau (1 - G_z(u-))e^{\theta_0 z - \Lambda_0(u)e^{\theta_0 z}}\varphi(z)dz}. \tag{3.8}$$

Cox' model: Prior. We construct Π as $\pi_\theta \otimes \pi_f$ with π_θ having a positive continuous density with respect to Lebesgue measure on $[-\theta_M, \theta_M]$. As prior π_f on $r = \log \lambda$, we considered the family of *Riemann-Liouville type processes* introduced in (12) and parameterised by the 'regularity' parameter $\alpha > 0$.

Theorem 3.4 *Suppose that* $\log \lambda_0$ *belongs to* $\mathcal{C}^\beta[0,\tau]$ *with* $\beta > 3/2$. *Suppose the least favorable direction* γ *in* (3.8) *has Hölder regularity at least* $2\beta/3$. *Let the prior* π_θ *be defined as described above and* π_f *be a Riemann-Liouville type process with parameter* $\alpha > 3/2$. *Then the conditions* **(P)**, $(\mathcal{C})$, $(\mathcal{N})$ *and* $(\mathcal{E})$ *of Theorem 3.2 are satisfied for pairs* (β, α) *such that the corresponding point in Figure 3.2 lies in the shaded area. In particular, the BvM theorem holds when* $\alpha > 3/2$ *and* $\alpha < 4\beta/3 - 1/2$.

The main difference with Theorem 3.3 is the triangular shape of the obtained region. The origin of this shape is due to a "no-bias"-type condition. As the proof reveals, conditions $(\mathcal{N})$ and $(\mathcal{E})$ both carry conditions imposing a constraint on the rate, and the strongest of the two leads to ask for $n^{3/4}\varepsilon_n^2 \to 0$.

Also, note that the regularity condition on γ is not difficult to check, due to the existence of an explicit form (3.8) for γ. For instance if Y and Z are independent, in which case $G_z(u)$ does not depend on z, then γ has the same regularity as Λ_0 so its Hölder regularity is at least $\beta + 1 > 2\beta/3$ and the condition is verified.

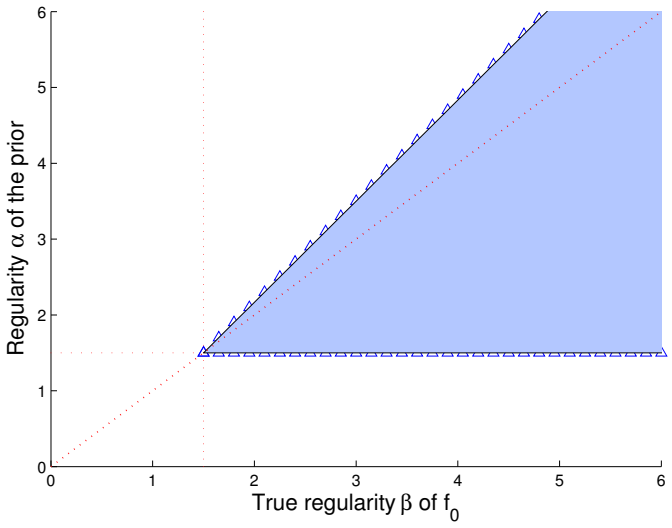

Figure 3.2: Cox' model. Possible choices for π_f

3.2 Semiparametric BvM, extensions

Natural questions arise from the results in the previous section

1. In the case with information loss, what about non-Gaussian priors ?
2. No-bias type conditions such as $(\mathcal{E})$ seem to be important. Can one avoid them ?
3. The case of implicit semiparametric functionals $f \to \psi(f)$ is not covered by these results.

Our aim in this short Section will simply be to give some insight into these questions via some specific results. About Questions 1. and 3., below we state a result in the density estimation model from [P13] which gives a partial answer for 'smooth' functionals. Next in the case of Gaussian priors we give an example where the BvM phenomenon does not hold in the region of parameters where $(\mathcal{E})$ is not satisfied.

We note that Theorem 3.5 below is a fairly special case of our main result in [P13]. One of our goals in [P13] is to shed some light on complex functionals where a study 'at first order' is not

enough. A typical example is, in the white noise model, estimation of $\psi(f) = \int f^2$, for Sobolev regularities of f below $\beta = 1/2$. For $\beta < 1/2$, second order terms become important and for $\beta < 1/4$ the estimation rate drops below the standard $\sqrt{n}$-rate. We shall not discuss further issues related to second order considerations here and refer to [P13] for a precise statement and the application to the mentioned quadratic functional.

Density estimation example. Consider the density model (3). We suppose that the true density f_0 is *bounded away* from 0 and ∞ on $[0,1]$. We consider $A_n = \{f,\ \|f - f_0\|_1 \le \varepsilon_n\}$ where ε_n is a positive sequence decreasing to 0. Let us define

$$L^2(f_0) = \{\varphi : [0,1] \to \mathbb{R}, \quad \int_0^1 \varphi(x)^2 f_0(x)dx < \infty\}.$$

For any φ in $L^2(f_0)$, we write $F_0(\varphi)$ as shorthand for $\int_0^1 \varphi(x) f_0(x)dx$.

Set, for any positive density f on $[0,1]$,

$$\eta = \log f, \quad \eta_0 = \log f_0, \qquad h = \sqrt{n}(\eta - \eta_0).$$

One can write the following LAN-type expansion

$$\ell_n(\eta) - \ell_n(\eta_0) = \sqrt{n}F_0(h) + \frac{1}{\sqrt{n}}\sum_{i=1}^n [h(X_i) - F_0(h)] = -\frac{1}{2}\|h\|_L^2 + W_n(h) + R_n(\eta, \eta_0),$$

with the notation, for any g in $L^2(f_0)$,

$$\|g\|_L^2 = \int_0^1 (g - F_0(g))^2 f_0, \quad W_n(g) = \mathbb{G}_n g = \frac{1}{\sqrt{n}}\sum_{i=1}^n [g(X_i) - F_0(g)], \tag{3.9}$$

and $R_n(\eta, \eta_0) = \sqrt{n}P_{f_0}h + \|h\|_L^2/2$. Note that $\|\cdot\|_L$ is an Hilbertian norm induced by the inner-product $\langle g_1, g_2\rangle_L = \int g_1 g_2 f_0$ defined on the space $\mathcal{H}_T := \{g \in L^2(P_{f_0}),\ \int g f_0 = 0\} \subset \mathcal{H} = L^2(f_0)$.

We consider functionals $\psi(f)$ of the density f, which are differentiable relative to the tangent set $\mathcal{H}_T$ with efficient influence function $\tilde{\psi}_{f_0}$, see [98], Chap. 25. Let us set

$$\psi(f) - \psi(f_0) = \langle \frac{f - f_0}{f_0}, \tilde{\psi}_{f_0}\rangle_L + \tilde{r}(f, f_0). \tag{3.10}$$

Theorem 3.5 *Let f_0 be bounded away from 0 and ∞ on $[0,1]$ and let Π be a prior on f. Let ψ be a differentiable functional relative to the tangent set $\mathcal{H}_T$, with efficient influence function $\tilde{\psi}_{f_0}$ bounded on $[0,1]$. Let $\tilde{r}$ be defined by (3.10). Suppose that for some $\varepsilon_n \to 0$ it holds*

$$\Pi[f :\ \|f - f_0\|_1 \le \varepsilon_n \,|\, X^{(n)}] \to 1, \tag{3.11}$$

in $P_{f_0}^{(n)}$-probability and that, for $A_n = \{f,\ \|f - f_0\|_1 \le \varepsilon_n\}$,

$$\sup_{f \in A_n} \tilde{r}(f, f_0) = o(1/\sqrt{n}).$$

Set $\eta_t = \eta - \frac{t}{\sqrt{n}}\tilde{\psi}_{f_0} - \log \int_0^1 e^{\eta - \frac{t}{\sqrt{n}}\tilde{\psi}_{f_0}}$ and assume that, in $P_{f_0}^{(n)}$-probability,

$$\frac{\int_{A_n} e^{\ell_n(\eta_t) - \ell_n(\eta_0)} d\Pi(\eta)}{\int e^{\ell_n(\eta) - \ell_n(\eta_0)} d\Pi(\eta)} \to 1. \tag{3.12}$$

Then, for $\hat{\psi}_n$ any linear efficient estimator of $\psi(f)$, the BvM theorem holds for the functional ψ. That is, the posterior distribution of $\sqrt{n}(\psi(f) - \hat{\psi}_n)$ is asymptotically Gaussian with mean 0 and variance $\|\tilde{\psi}_{f_0}\|_L^2$, in $P_{f_0}^{(n)}$-probability.

The semiparametric efficiency bound for estimating $\psi(f)$ at f_0 is $\|\tilde{\psi}_{f_0}\|_L^2$, so Theorem 3.5 yields the BvM Theorem for the functional $\psi(f)$, with efficient limiting distribution.

Theorem 3.5 has three main conditions: first, a concentration of the posterior around the true f_0, similar in spirit to (**C**) above. Second, the remainder term $\tilde{r}$ of the functional expansion is assumed small, which ensures that a study 'at first order' is sufficient. Third, condition (3.12) expresses that some perturbation of the parameter along the efficient influence function $\tilde{\psi}_{f_0}$ can be ignored as $n \to \infty$ in a ratio of integrals. Making this condition more explicit would be very desirable but seems difficult without saying more on the prior Π. A further informal interpretation of this conditions is that priors should behave well under a 'change of variable' $\eta \to \eta - t\tilde{\psi}_{f_0}/\sqrt{n}$.

In [P13], as far as first order results are concerned, we investigate two classes of priors, which were not considered before from the BvM perspective: random histograms and Gaussian process priors. In doing so, we develop change of variables formulas for histograms which we also use in [P12]. The message is similar to that of the previous Section: for the semiparametric BvM to hold one needs the prior to approximate sufficiently well the influence function of the functional (in the previous section the 'least favorable direction') and possibly also the true function. In [P13], we also consider the possibility of BvM results for a class of adaptive priors, namely random histograms with a random cut-off-level. One notable conclusion is that if the efficient influence function and the true density have very different regularities, then the use of an adaptive prior for f may actually prevent the semiparametric BvM to hold. A simple example with Haar-histogram priors is given in [P13], Section 4.4.

THE ROLE OF THE BIAS. We now turn to an example from [P8] illustrating what can happen in the region of parameters where $(\boldsymbol{\mathcal{E}})$ does not hold. Consider the following semiparametric 'alignment of curves' problem. Let θ belong to $\Theta \subset [-\tau, \tau]$, with $0 < \tau < 1/2$. Let f belong to $\mathcal{F} = L^2[0,1]$. For simplicity of treatment, we assume that f is 1-periodic. The observations consist of the paths

$$dY(t) = f(t)dt + \frac{1}{\sqrt{n}} dW_1(t)$$
$$dZ(t) = f(t-\theta)dt + \frac{1}{\sqrt{n}} dW_2(t),$$

where W_1, W_2 are independent standard Brownian motions and $t \in [0,1]$. Both the real θ and the function f are unknown, making the model semiparametric in $\eta = (\theta, f)$. Let X denote the coupled observation of (Y, Z). We omit the index n to simplify the notation.

Consider a prior $\Pi = \pi_\theta \otimes \pi_f$ on the pair $\eta = (\theta, f)$. As prior π_θ on θ, we simply take the uniform distribution on the interval $[-\tau, \tau]$. For π_f we consider two examples of prior distributions. Let $\{\nu_k\}_{k \geq 1}$ be a sequence of independent $N(0,1)$ random variables. We define, for any real $\alpha > 1$ and u in $\mathbb{R}$, the priors

$$\pi_f^\alpha \sim \sqrt{2} \sum_{k=1}^{\infty} \left[(2k)^{-\frac{1}{2}-\alpha} \nu_{2k} \cos(2\pi k u) + (2k)^{-\frac{1}{2}-\alpha} \nu_{2k+1} \sin(2\pi k u) \right]$$
$$\pi_f^{\alpha,*} \sim \sqrt{2} \sum_{k=1}^{\infty} \left[(2k)^{-\frac{1}{2}-\alpha} \nu_{2k} \cos(2\pi k u) + (2k+1)^{-\frac{1}{2}-\alpha} \nu_{2k+1} \sin(2\pi k u) \right].$$

Thus, the prior π_f^α draws random functions with Gaussian Fourier coefficients of variance equal on even and odd harmonics to the same constant times $k^{-1-2\alpha}$. The prior $\pi_f^{\alpha,*}$ is the same, except that the variance of the kth Fourier coefficient is simply $k^{-1-2\alpha}$.

Provided the true function f_0 has at least one derivative in the L^2-sense, one can set

$$\gamma := -f_0'/2 \qquad \text{and} \qquad \tilde{I} := \frac{1}{2} \int_0^1 f_0'^2(u) du.$$

One can check that these are respectively the least favorable direction and the efficient information in the model as defined in the previous section. So, there is a loss of information in this model if

f_0' is nonzero, which should be assumed anyways for identifiability reasons. So, assuming $\tilde{I} > 0$, we also denote $\Delta := -\tilde{I}^{-1} \int_0^1 [\gamma(u)dW_1(u) - \gamma(u-\theta_0)dW_2(u)]$. A special example illustrating our results is the function $f_0^{[\beta]}$, defined for $\beta > 3/2$ by

$$f_0^{[\beta]}(u) = \sqrt{2} \sum_{k=1}^{\infty} \left[(2k)^{-\frac{1}{2}-\beta} \cos(2\pi k u) + (2k+1)^{-\frac{1}{2}-\beta} \sin(2\pi k u) \right]. \tag{3.13}$$

The following Proposition is a consequence of the main result in [P8].

Proposition 3.1 *Let $\eta_0 = (\theta_0, f_0)$ with $\theta_0 \in \Theta$ and f_0 the function $f_0^{[\beta]}$ defined in (3.13). Let $\Pi^\alpha = \pi_\theta \otimes \pi_f^\alpha$ and $\Pi^{\alpha,*} = \pi_\theta \otimes \pi_f^{\alpha,*}$. Take $\alpha = 4$ and $\beta = 2$. As $n \to \infty$, it holds*

$$\|\Pi^\alpha(\cdot \times \mathcal{F} \,|\, X) - N(\theta_0 + \frac{\Delta}{\sqrt{n}}, \frac{\tilde{I}^{-1}}{n})(\cdot)\| \to^{P_{\eta_0}^{(n)}} 0.$$

In particular, the semiparametric Bernstein-von Mises theorem holds for Π^α. On the other hand, for any $\delta > 4/9$ and any $M > 0$, as $n \to \infty$,

$$\Pi^{\alpha,*}(|\theta - \theta_0| \leq M n^{-\delta} \,|\, X) \to^{P_{\eta_0}^{(n)}} 0.$$

In particular, the marginal of the Bayesian posterior for $\Pi^{\alpha,*}$ is not $\sqrt{n}$-consistent.

One can check that condition $(\boldsymbol{\mathcal{E}})$ of Theorem 3.2 asks for $\alpha < 2\beta - 3/2$ and indeed $(\alpha, \beta) = (4, 2)$ does not meet this condition. In fact, some explicit computations can be carried out in the present model, and the following can be shown. If one takes a prior for π_f with general Fourier coefficients $\{\sigma_{2k}, \sigma_{2k+1}\}$, under some mild conditions on those as well as on f_0, one can check that,

$$\|\Pi^\sigma(\cdot \times \mathcal{F} \,|\, X) - N(\theta_0 + \frac{\Delta + \zeta_n^\sigma}{\sqrt{n}}, \frac{\tilde{I}^{-1}}{n})\| \to^{P_{\eta_0}^{(n)}} 0,$$

where ζ_n^σ is given by, for $\{f_{0,k}\}$ the Fourier coefficients of f_0,

$$\zeta_n^\sigma = \frac{\pi}{\sqrt{n}} \sum_{k=1}^{\infty} k f_{0,2k} f_{0,2k+1} (\sigma_{2k+1}^{-2} - \sigma_{2k}^{-2}) \left\{ \frac{(2n)^2}{(2n + \sigma_{2k}^{-2})(2n + \sigma_{2k+1}^{-2})} \right\}. \tag{3.14}$$

Note that if $\sigma_{2k} = \sigma_{2k+1}$ then ζ_n^σ is zero, which explains the first part of Proposition 3.1. On the other hand, if σ_{2k} and σ_{2k+1} are different, ζ_n^σ may not tend to 0 as $n \to \infty$. More precisely, this happens for $\pi_f^{\alpha,*}$ above as soon as $\alpha \geq 2\beta - 1/2$. This is the case for $(\alpha, \beta) = (4, 2)$, hence the second part of Proposition 3.1.

As a conclusion, the semiparametric BvM does not hold for some Gaussian priors in most of the complement of the region delimited by condition $(\boldsymbol{\mathcal{E}})$. The departure from BvM takes the form of a *bias* term, here ζ_n^σ given by (3.14), as announced.

3.3 Nonparametric BvM in Gaussian white noise

In view of the previous semiparametric BvM results, it is natural to ask whether a *nonparametric* BvM theorem can be formulated. If so is it still possible to deduce applications to confidence sets ?

Introduction. A natural place to start is the Gaussian white noise model (1). Even for such a simple model, Cox (1993) [30] and Freedman (1999) [41] have shown the impossibility of a nonparametric BvM result in a strict L^2-setting. Leahu (2011) [73] derived interesting results on the possibility and impossibility of BvM-theorems in model (1): Leahu shows that a BvM-result in white noise, in the total variation sense and for Gaussian conjugate priors, can only hold for

priors inducing a heavy undersmoothing (such priors do not in fact induce L^2 random functions). Though these results are nice mathematically, as a consequence of the roughness the induced credible sets are typically too large for most nonparametric applications.

Also, a number of BvM-type results have been obtained for finite-dimensional posteriors with dimension increasing to infinity: Ghosal [45] and Bontemps [19] consider regression with a finite number of regressors, Ghosal [46] and Clarke and Ghosal [28] consider exponential families, and the case of discrete probability distributions is treated in Boucheron and Gassiat [21]. Another recent further result in this direction is Panov and Spokoiny [85], which also allows for possible model misspecification. These results are formulated in terms of the total variation distance.

A NEW APPROACH. The idea we propose in [P10] consists of two parts 1) enlarge the space in which results are formulated and 2) change the notion of convergence in the result. In view of the negative results mentioned above, that rule out the existence of BvM in a pure L^2 setting, part 1) is quite natural: one defines a space common to all components of the white noise model. But this space enlargement has another effect: since the space is larger, the norm becomes weaker. Hence tightness at a 'fast', parametric, $1/\sqrt{n}$ rate becomes possible and enables part 2), that is weak convergence of probability measures in the enlarged space for broad classes of prior distributions.

AN ENLARGED SPACE. To stay in a Hilbert space setting while enlarging L^2, a natural class comes to mind, that of *negative*-order Sobolev spaces $\{H_2^r\}_{r<0}$ on $[0,1]$, defined similarly as usual Sobolev spaces, but with negative orders. To obtain sharp results we need 'logarithmic' Sobolev spaces, for a real s and $\delta > 0$,

$$H_2^{s,\delta} := \left\{ f : \quad \|f\|_{s,2,\delta}^2 := \sum_{l\geq 0} \frac{(2^l)^{2s}}{l^{2\delta}} \sum_{k=0}^{2^l-1} \langle \psi_{lk}, f\rangle^2 < \infty \right\}, \tag{3.15}$$

where $\{\psi_{lk}\}$ is a wavelet basis on $[0,1]$ (a Fourier-type basis is also possible here, but wavelets will be crucially needed in the next Section, so for easy reference we write already in terms of wavelet notation). The space should be large enough so that the Gaussian experiment in (1) can be realised as a tight random element in that space. The critical value for this to be the case turns out to be $s = -1/2$. So, define the collection of Hilbert spaces

$$H := H_2^{-1/2,\delta}, \quad \|\cdot\|_H := \|\cdot\|_{-1/2,2,\delta}, \quad \delta > 1/2. \tag{3.16}$$

If we denote by $\mathbb{W}$ the centered Gaussian Borel random variable on H with covariance I, then the Gaussian white noise model (1) can be written as

$$\mathbb{X}^{(n)} = f + \frac{1}{\sqrt{n}}\mathbb{W}, \tag{3.17}$$

a natural Gaussian shift experiment in the Hilbert space H. For any $\delta > 1/2$, a simple calculation reveals that the $\|\mathbb{W}\|_{-1/2,2,\delta}$-norm converges almost surely and $\mathbb{W}$ is thus tight in $H_2^{-1/2,\delta}$.

We denote by $\mathcal{N}$ the law of $\mathbb{W}$ a standard, or canonical, Gaussian probability measure on the Hilbert space H. Now we are ready to define the notion of convergence we consider: it is simply weak convergence, but in the space H. It is convenient to consider a distance which metrises it. We choose the so-called bounded-Lipschitz metric, see e.g. Dudley (2002) [37], Theorem 11.3.3.

BOUNDED-LIPSCHITZ METRIC. The space of bounded Lipschitz functions on the space H is

$$BL_H(1) = \left\{ f : H \to \mathbb{R}, \quad \sup_{s\in H} |f(s)| + \sup_{s\neq t, s,t\in H} |f(s) - f(t)|/d(s,t) \leq 1 \right\}.$$

The bounded Lipschitz metric on probability measures on H is $\beta := \beta_H$ defined as

$$\beta_H(\mu,\nu) = \sup_{u\in BL_H(1)} \left| \int_H u(s)(d\mu - d\nu)(s) \right|. \tag{3.18}$$

The metric β metrises weak convergence of probability measures on H.

NOTION OF WEAK BVM IN THE SPACE H. Let Π be a prior on L^2. It naturally induces a prior on H by the injection $L^2 \to H$. Let

$$\Pi_n = \Pi(\cdot \,|\, X^{(n)}) = \Pi(\cdot \,|\, \mathbb{X}^{(n)})$$

denote the corresponding posterior distribution on H given the data from the white noise model (1), or equivalently, from (3.17). On H and for $z \in H$, define the transformation

$$\tau_z : f \mapsto \sqrt{n}(f - z). \tag{3.19}$$

Let $\Pi_n \circ \tau_{\mathbb{X}^{(n)}}^{-1}$ be the image of the posterior law under $\tau_{\mathbb{X}^{(n)}}$. The shape of $\Pi_n \circ \tau_{\mathbb{X}^{(n)}}^{-1}$ reveals how the posterior concentrates on $1/\sqrt{n}$-H-neighborhoods of the efficient estimator $\mathbb{X}^{(n)}$.

Definition 3.1 *Consider the white noise model (1) viewed in H as (3.17). Under a fixed function f_0, denote by $P_{f_0}^{(n)}$ the distribution of $\mathbb{X}^{(n)}$. Let β be the bounded Lipschitz metric for weak convergence of probability measures on H. We say that a prior Π satisfies the weak Bernstein - von Mises phenomenon in H if, as $n \to \infty$,*

$$\beta(\Pi_n \circ \tau_{\mathbb{X}^{(n)}}^{-1}, \mathcal{N}) \to^{P_{f_0}^{(n)}} 0.$$

Thus when the weak Bernstein-von Mises phenomenon holds the posterior necessarily has the approximate shape of an infinite-dimensional Gaussian distribution. Moreover, we require this Gaussian distribution to equal $\mathcal{N}$ – the canonical choice in view of efficiency considerations. The covariance of $\mathcal{N}$ is the Cramér-Rao bound for estimating f in the Gaussian shift experiment (3.17) in H-loss, and this can be seen to carry over to sufficiently regular real-valued functionals.

At this point one may wonder whether the notion of weak convergence in H is strong enough to include interesting applications.

UNIFORMITY CLASSES FOR WEAK CONVERGENCE. Since we have a statement in terms of weak convergence (as opposed e.g. to total variation) we cannot infer that $\Pi_n \circ \tau_{\mathbb{X}^{(n)}}^{-1}$ and $\mathcal{N}$ are approximately the same for every Borel set in H, but only for sets B that are continuity sets for the probability measure $\mathcal{N}$. For statistical applications of the BvM phenomenon one typically needs some uniformity in B. Weak convergence in H implies that $\Pi_n \circ \tau_{\mathbb{X}^{(n)}}^{-1}$ is close to $\mathcal{N}$ uniformly in classes of subsets of H whose boundaries are sufficiently regular relative to the measure $\mathcal{N}$.

We call a family $\mathcal{U}$ of measurable real-valued functions defined on H *a $\mathcal{N}-$uniformity class for weak convergence* if for any sequence μ_n of Borel probability measures on H that converges weakly to $\mathcal{N}$ we also have

$$\sup_{u \in \mathcal{U}} \left| \int_H u(s)(d\mu_n - d\mathcal{N})(s) \right| \to 0 \tag{3.20}$$

as $n \to \infty$. For any subset A of H, define the δ-boundary of A by $\partial_\delta A = \{x \in H : d(x, A) < \delta, d(x, A^c) < \delta\}$. By Theorem 2 of Billingsley and Topsøe [12], a family $\mathcal{A}$ of measurable subsets of H is a $\mathcal{N}$-uniformity class if and only if

$$\lim_{\delta \to 0} \sup_{A \in \mathcal{A}} \mathcal{N}(\partial_\delta A) = 0. \tag{3.21}$$

This typically allows for enough uniformity to deal with a variety of concrete nonparametric statistical problems. A key property for this is that, by general properties of Gaussian measures on separable Banach spaces, the collection of all centered balls (or rather, in a L^2-perspective, ellipsoids) for the $\|\cdot\|_H$-norm verify (3.21) and thus form a $\mathcal{N}$-uniformity class.

APPLICATION: WEIGHTED L^2-CREDIBLE ELLIPSOIDS. Recall that H stands for the space $H(\delta)$ from (3.16) for some arbitrary choice of $\delta > 1/2$. Denote by $B(g, r) = \{f \in H : \|f - g\|_H \leq r\}$

the norm ball in H of radius r centered at g. In terms of the wavelet basis $\{\psi_{lk}\}$ of L^2, this corresponds to L^2-ellipsoids of radius r

$$\Big\{\{c_{lk}\}: \quad \sum_{l,k} l^{-2\delta}2^{-l}|c_{lk} - \langle g, \psi_{lk}\rangle|^2 \leq r^2\Big\}.$$

To find the appropriate radius, one may use the quantiles of the posterior distribution. Given $\alpha > 0$, one solves for $R_n \equiv R(\mathbb{X}^{(n)}, \alpha)$ such that

$$\Pi(f : \|f - T_n\|_H \leq R_n/\sqrt{n} \,|\, \mathbb{X}^{(n)}) = 1 - \alpha, \tag{3.22}$$

where $T_n = \mathbb{X}^{(n)}$. By its mere definition, a $\|\cdot\|_H$-ball centred at T_n of radius R_n constitutes a level $(1-\alpha)$-*credible* set for the posterior distribution. The weak Bernstein-von Mises phenomenon in H implies that this credible ball asymptotically coincides with the exact $(1-\alpha)$-confidence set built using the efficient estimator $\mathbb{X}^{(n)}$ for f, by the following result.

Theorem 3.6 *Suppose the weak Bernstein-von Mises phenomenon in the sense of Definition 3.1 holds. Given $0 < \alpha < 1$ consider the credible set*

$$C_n = \Big\{f : \ \|f - \mathbb{X}^{(n)}\|_H \leq R_n/\sqrt{n}\Big\} \tag{3.23}$$

where $R_n \equiv R(\mathbb{X}^{(n)}, \alpha)$ is such that $\Pi(C_n|\mathbb{X}^{(n)}) = 1 - \alpha$. Then, as $n \to \infty$,

$$P^n_{f_0}(f_0 \in C_n) \to 1 - \alpha \quad \textit{and} \quad R_n = O_P(1).$$

Also, the posterior mean $\bar{f}_n$ may replace $\mathbb{X}^{(n)}$ in (3.23) as long as $\|\bar{f}_n - \mathbb{X}^{(n)}\|_H = o_P(n^{-1/2})$.

One may ask: what is 'size' of the confidence set C_n in (3.23) ? Indeed, C_n has a 'small' radius $1/\sqrt{n}$ but in terms a *weak* norm, namely the H-norm. So far it would seem that we cannot re-translate this result back in L^2-terms. We can do so by using a helpful 'interpolation' idea.

Interpolation idea. Let $B_{\alpha,\|\cdot\|_\infty}(g, R)$ denote a ball in the Hölder space of functions of order $\alpha > 0$ on $[0,1]$, and $B_{\|\cdot\|_2}(g, R)$ a ball for the standard L^2-norm on $[0,1]$.

Let $c_1, c_2 > 0$ be given and let $g \in L^2$. Then there exists $c_3 > 0$ such that for $n \geq 1$,

$$B_{\|\cdot\|_H}\left(g, \frac{c_1}{\sqrt{n}}\right) \ \cap \ B_{\|\cdot\|_{\alpha,\infty}}(0, c_2) \ \subset \ B_{\|\cdot\|_2}\left(g, \frac{c_3 l_n}{n^{\frac{\alpha}{2\alpha+1}}}\right), \tag{3.24}$$

where $l_n = \log^{2\delta} n$. So, provided it is possible to further intersect the previous credible set with a α-Hölder ball of fixed radius, the resulting set is automatically included in a L^2-ball of radius precisely the standard minimax nonparametric rate for α-regular functions, up to a logarithmic term (one may note that the log-term comes from the logarithmic correction to the space H).

Example of nonparametric confident credible set. For the sake of simplicity, consider first a uniform wavelet prior Π on L^2 arising from the law of the random wavelet series, for $\alpha > 0$,

$$U_{\alpha,M} = \sum_l \sum_k 2^{-l(\alpha+1/2)} u_{lk} \psi_{lk}(\cdot), \tag{3.25}$$

where the u_{lk} are i.i.d. uniform on $[-M, M]$ for some $M > 0$ and indexes l, k vary as usual. Such priors model functions that lie in a fixed Hölder ball of $\|\cdot\|_{\alpha,\infty}$-radius M, with posteriors $\Pi(\cdot \,|\, \mathbb{X}^{(n)})$ contracting about f_0 at the L^2-minimax rate within logarithmic factors if $\|f_0\|_{\alpha,\infty} \leq M$, see [50]. Of course in practice using such a prior means that an upper-bound on the α-Hölder norm is known, which may not always be the case. We note below that the method can be adapted if this is not the case.

In this situation it is natural to intersect the credible set C_n with a Hölder ball

$$\mathcal{C}_n = \big\{f : \|f\|_{\alpha,\infty} \leq M, \quad \|f - \bar{f}_n\|_H \leq R_n/\sqrt{n}\big\}, \tag{3.26}$$

where R_n is as in (3.22) with $T_n = \bar{f}_n$. By definition of the prior Π induced by $U_{\alpha,M}$ above, we have $\|f\|_{\alpha,\infty} \leq M$, Π-almost surely, so also Π_n-almost surely. In particular $\Pi(\mathcal{C}_n|\mathbb{X}^{(n)}) = 1-\alpha$, so $\mathcal{C}_n$ is a credible set of level $1-\alpha$. Theorem 3.6 implies the following result.

Corollary 3.1 *Consider observations from (3.17) under a function $f_0 \in C^\alpha$ with $\|f_0\|_{\alpha,\infty} < M$. Let Π be the law of $U_{\alpha,M}$, let $\Pi(\cdot|\mathbb{X}^{(n)})$ the associated posterior and let $\mathcal{C}_n$ be as in (3.26). Then*

$$P^n_{f_0}(f_0 \in \mathcal{C}_n) \to 1-\alpha$$

as $n \to \infty$ and the L^2-diameter $|\mathcal{C}_n|_2$ of $\mathcal{C}_n$ satisfies, for some $\kappa > 0$,

$$|\mathcal{C}_n|_2 = O_P(n^{-\alpha/(2\alpha+1)}(\log n)^\kappa).$$

More generally, to avoid the use of a fixed bound M in (3.25), one may use more general priors that have infinite support on each coordinate. Then instead of intersecting with an α-Hölder ball of given radius M, it is enough to intersect it with a ball whose radius is a type of 'posterior quantile' for a α-order norm. The results then parallel those for the uniform priors, but without the boundedness constraint. A precise statement is given in [P10], Corollary 2.

Conclusion so far on credible sets. The meaning of the results we have presented so far is as follows: provided one can show the weak nonparametric BvM for standard nonparametric priors modelling α–smooth functions - this will be the object of the next paragraphs - it is possible to deduce *confident credible sets* which have diameter equal to the nonparametric minimax rate of convergence for such problems, up to a slowly varying factor (this extra log-term can in fact be replaced by an *arbitrary* factor $M_n \to \infty$, up to a slightly different definition of the space H). It is important to note that for the argument to go through, α–smooth priors should be allowed, as opposed to priors inducing a too severe undersmoothing.

We also note that such confidence sets are for *fixed regularity* α (i.e. one should know α, or a lower bound on it, to at least 'undersmooth'). Construction of *adaptive* confident credible sets is an interesting further problem, but is qualitatively somewhat different: in particular, rates will typically change, unless something more is assumed on the considered functions.

BvM for product priors in Gaussian white noise. Let us consider priors of the form $\Pi = \otimes_{lk}\pi_{lk}$ defined on the coordinates of the orthonormal basis $\{\psi_{lk}\}$, where π_{lk} are probability distributions with Lebesgue density φ_{lk} on the real line, with the following assumptions. For some fixed density φ on the real line and admissible indexes k, l,

$$\varphi_{lk}(\cdot) = \frac{1}{\sigma_l}\varphi\left(\frac{\cdot}{\sigma_l}\right) \quad \forall k, \qquad \text{with } \sigma_l > 0.$$

Condition 3.1 *Suppose that for a finite constant $M > 0$,*

$$\sup_{l,k} \frac{|\langle f_0, \psi_{lk}\rangle|}{\sigma_l} \leq M.$$

Suppose also that for some $\tau > M$ and $0 < c_\varphi \leq C_\varphi < \infty$

$$\varphi(x) \leq C_\varphi \quad \forall x \in \mathbb{R}, \quad \varphi(x) \geq c_\varphi \quad \forall x \in (-\tau, \tau), \quad \text{and} \quad \int_{\mathbb{R}} x^2\varphi(x)dx < \infty.$$

This allows for a rich variety of base priors φ, such as Gaussian, sub-Gaussian, Laplace, most Student laws, or more generally any law with positive continuous density and finite second moment, but also uniform priors with large enough support. The full prior on f considered here is thus a sum of independent terms over the basis $\{\psi_{lk}\}$, including many, especially non-Gaussian, processes. One may also consider Gaussian processes such as Brownian motion, even if their Karhunen-Loève expansion is not a (localised) wavelet basis, as long as it is smooth enough, see [P10] for details.

First, one shows the following intermediate result

Theorem 3.7 *Consider data from the white noise model (3.17) under a fixed function f_0 with coefficients $\{\langle f_0, \psi_{lk}\rangle\}$ over the basis $\{\psi_{lk}\}$. Then if the product prior Π and f_0 satisfy Condition 3.1, we have, as $n \to \infty$,*

$$E_{f_0}^{(n)} \int \|f - f_0\|_H^2 d\Pi(f \mid \mathbb{X}^{(n)}) = O\left(\frac{1}{n}\right).$$

The conclusion of this Theorem enables a tightness argument at rate $1/\sqrt{n}$ in the proof of the weak BvM theorem stated next. Interestingly, even if the result of Theorem 3.7 is in terms of a 'weak' norm (the rate is 'fast' though !), it does not seem possible to derive this convergence rate using a testing approach as in the general rate Theorem 0.1. Instead, we use a type of *multiscale* approach, which we later formalised for more complex norms as in Section 2.4.

Theorem 3.8 *Suppose the assumptions of Theorem 3.7 are satisfied and that φ is continuous near $\{\langle f_0, \psi_{lk}\rangle\}$ for every indexes l, k. Then for β the bounded Lipschitz metric for weak convergence of probability measures on H, as $n \to \infty$ we have $\beta(\Pi_n \circ \tau_{\mathbb{X}^{(n)}}^{-1}, \mathcal{N}) \to^{P_{f_0}^{(n)}} 0$.*

We note that the results of Theorem 3.8 as well as those presented above on credible sets, which are its consequences, can be seen to be uniform ('honest') in all f_0 that satisfy Condition 3.1 with fixed constant M.

Further applications and uniform semiparametrics. Another important set of applications of the weak nonparametric BvM theorem is related to continuous functionals. Indeed, by the continuous mapping theorem, it immediately follows that the weak convergence result in Definition 3.1 implies weak convergence of the image measures through any *continuous* mapping $\psi : H \to \mathcal{Y}$, for some given space $\mathcal{Y}$. Applications include semiparametric BvM results for linear and smooth nonlinear functionals, credible bands for selfconvolutions, etc., see [P10], Section 2. In this perspective, one may see the weak nonparametric BvM as a semiparametric BvM 'uniform in many functionals'.

This also leads to a natural question: is the choice of space H canonical ? What if the goal is to obtain credible sets in different norms than $\|\cdot\|_2$, such as $\|\cdot\|_\infty$? We consider this next.

3.4 Nonparametric BvM and Donsker's theorem

In the previous section we have considered the white noise model and confidence-sets results linked to the $\|\cdot\|_2$-norm. In addition to the question of obtaining results in terms of different norms, it is natural to consider the nonparametric BvM question for other statistical models as well. Here we shall focus on density estimation, following [P14]. We note that the construction below can also be followed in the white noise model as an alternative to the construction in the previous section.

We define Hölder-type spaces C^s of continuous functions on $[0,1]$:

$$C^s([0,1]) = \left\{ f \in C([0,1]) : \|f\|_{s,\infty} := \sup_{l,k} 2^{l(s+1/2)} |\langle \psi_{lk}, f\rangle| < \infty \right\}. \tag{3.27}$$

Density model, limiting distribution. Consider the density model (3) where we observe $X_1, \ldots, X_n$ i.i.d. from law P with density f on $[0,1]$.

In analogy with the white noise case, the first step is to identify the limiting distribution for the BvM result. In white noise, one identifies the limit via the equation (3.17), $\mathbb{X}^{(n)} = f + n^{-1/2}\mathbb{W}$, where $\mathbb{X}^{(n)}$ can be seen as an estimator of f. In the density model, let us take an intermediate step via projections onto the wavelet basis $\{\psi_{lk}\}$, which we assume to be a localised basis such as the Haar or CDV bases used in Section 2.4.

A natural estimate of $\langle f, \psi_{lk}\rangle$ is given by $P_n\psi_{lk} \equiv \langle P_n, \psi_{lk}\rangle = \frac{1}{n}\sum_{i=1}^n \psi_{lk}(X_i)$.

THE P-WHITE BRIDGE PROCESS. By the central limit theorem, for k, l fixed and as $n \to \infty$, the random variable $\sqrt{n}(P_n - P)(\psi_{lk})$ converges in distribution to

$$\mathbb{G}_P(\psi_{lk}) \sim N(0, \mathrm{Var}_P(\psi_{lk}(X_1))). \tag{3.28}$$

In analogy to the white noise process $\mathbb{W}$, the process $\mathbb{G}_P$ arising from (3.28) can be defined as the Gaussian process indexed by the Hilbert space

$$L^2(P) \equiv \left\{ f : [0,1] \to \mathbb{R} : \int_0^1 f^2 dP < \infty \right\}$$

with covariance function $\mathbb{E}\left[\mathbb{G}_P(g)\mathbb{G}_P(h)\right] = \int_0^1 (g - Pg)(h - Ph)dP$. We call $\mathbb{G}_P$ the *P-white bridge process*. Now we turn to the definition of spaces similar to the large space H of the previous section.

MULTISCALE SPACES $\mathcal{M}(w)$ AND $\mathcal{M}_0(w)$. For monotone increasing weighting sequences $w = (w_l : l \geq J_0 - 1), w_l \geq 1$, we define multi-scale sequence spaces

$$\mathcal{M} \equiv \mathcal{M}(w) \equiv \left\{ x = \{x_{lk}\} : \quad \|x\|_{\mathcal{M}(w)} \equiv \sup_l \frac{\max_k |x_{lk}|}{w_l} < \infty \right\}. \tag{3.29}$$

The space $\mathcal{M}(w)$ is a non-separable Banach space (it is isomorphic to ℓ_∞). However, the weighted sequences in $\mathcal{M}(w)$ that vanish at infinity form a separable closed subspace for the same norm, which leads us to define

$$\mathcal{M}_0 = \mathcal{M}_0(w) = \left\{ x \in \mathcal{M}(w) : \quad \lim_{l \to \infty} \max_k \frac{|x_{lk}|}{w_l} = 0 \right\}. \tag{3.30}$$

Furthermore, we call a sequence (w_l) *admissible* if $w_l/\sqrt{l} \uparrow \infty$ as $l \to \infty$.

P-WHITE BRIDGE AS TIGHT MEASURE ON $\mathcal{M}_0(w)$. The idea behind the definition of the enlarged space $\mathcal{M}_0(w)$ is, as for H, to find a 'smallest' (separable) large space the limit $\mathbb{G}_P$ belongs to. The next proposition also applies to white noise $\mathbb{W}$.

Proposition 3.2 *Let $\mathbb{G}_P$ be a P-white bridge. For $\omega = (\omega_l) = \sqrt{l}$ we have $E\|\mathbb{G}_P\|_{\mathcal{M}(\omega)} < \infty$. If $w = (w_l)$ is admissible then $\mathbb{G}_P$ defines a tight Gaussian Borel probability measure in $\mathcal{M}_0(w)$.*

TRUNCATED EMPIRICAL MEASURE, CONVERGENCE. Any P with bounded density f has coefficients $\langle f, \psi_{lk} \rangle \in \ell_2 \subset \mathcal{M}_0(w)$. We would like to formulate a statement such as

$$\sqrt{n}(P_n - P) \to^d \mathbb{G}_P \text{ in } \mathcal{M}_0,$$

as $n \to \infty$, paralleling (3.17) in the Gaussian white noise setting. The fluctuations of $\sqrt{n}(P_n - P)(\psi_{lk})/\sqrt{l}$ along k are stochastically bounded for l such that $2^l \leq n$, but are unbounded for high frequencies. Thus the empirical process $\sqrt{n}(P_n - P)$ will not define an element of $\mathcal{M}_0$ for every admissible sequence w. In our nonparametric setting we can restrict to frequencies at levels $l, 2^l \leq n$, and introduce an appropriate 'projection' $P_n(j)$ of the empirical measure P_n onto V_j via

$$\langle P_n(j), \psi_{lk} \rangle = \begin{cases} \langle P_n, \psi_{lk} \rangle & \text{if } \ l \leq j \\ 0 & \text{if } \ l > j, \end{cases} \tag{3.31}$$

which defines a tight random variable in $\mathcal{M}_0$. The following theorem shows that $P_n(j)$ estimates P efficiently in $\mathcal{M}_0$ if j is chosen appropriately. Note that the natural choice $j = L_n$ such that

$$2^{L_n} \sim N^{1/(2\gamma+1)},$$

where $N = n$ (if $\gamma > 0$) or $N = n/\log n$ (if $\gamma \geq 0$), is possible.

Theorem 3.9 *Let $w = (w_l)$ be admissible. Suppose P has density f in $C^\gamma([0,1])$ for some $\gamma \geq 0$. Let j_n be such that*

$$\sqrt{n}2^{-j_n(\gamma+1/2)}w_{j_n}^{-1} = o(1), \quad \frac{2^{j_n}j_n}{n} = O(1).$$

Then we have, as $n \to \infty$,

$$\sqrt{n}(P_n(j_n) - P) \to^d \mathbb{G}_P \text{ in } \mathcal{M}_0(w).$$

WEAK NONPARAMETRIC BVM IN $M_0(w)$. As in the previous section, we metrise weak convergence of laws in $\mathcal{M}_0(w)$ via $\beta_{\mathcal{M}_0(w)}$ defined in the same way as (3.18), where now β_S denotes weak convergence on a space S, and view the prior Π on the functional parameter $f \in L^2$ as a prior on sequence space ℓ_2 under the wavelet isometry $L^2 \cong \ell_2$.

Definition 3.2 *Let w be admissible, let Π be a prior and $\Pi(\cdot \mid X^{(n)})$ the corresponding posterior distribution on $\ell_2 \subset \mathcal{M}_0 = \mathcal{M}_0(w)$, obtained from observations $X^{(n)}$ in the density model. Let $\tilde{\Pi}_n$ be the image measure of $\Pi(\cdot \mid X^{(n)})$ under the mapping*

$$\tau : f \mapsto \sqrt{n}(f - T_n)$$

where $T_n = T_n(X^{(n)})$ is an estimator of f in $\mathcal{M}_0$. Then we say that Π satisfies the weak Bernstein von Mises phenomenon in $\mathcal{M}_0$ with centering T_n if, for $X^{(n)} \sim P_{f_0}^n$ and fixed f_0, as $n \to \infty$,

$$\beta_{\mathcal{M}_0}(\tilde{\Pi}_n, \mathcal{N}) \to^{P_{f_0}} 0,$$

where $\mathcal{N}$ is the law in $\mathcal{M}_0$ of $\mathbb{G}_{P_0}$, $f_0 \in L^\infty$.

WEAK NONPARAMETRIC BVM, DENSITY MODEL. We define multi-scale priors Π on some space $\mathcal{F}$ of probability density functions f giving rise to absolutely continuous probability measures. Suppose the true density f_0 is bounded away from 0 and ∞.

We now introduce possible values for a cut-off parameter L_n. For $\alpha > 0$, let $j_n = j_n(\alpha)$ and $l_n = l_n(\alpha)$ be the *largest* integers such that

$$2^{j_n} \leq n^{\frac{1}{2\alpha+1}}, \qquad 2^{l_n} \leq \left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}}. \tag{3.32}$$

and set, in slight abuse of notation, either

$$L_n = j_n \quad (\forall n \geq 1) \quad \text{or} \quad L_n = l_n \quad (\forall n \geq 1). \tag{3.33}$$

We consider two classes of priors, the examples **(S), (H)** of respectively log-density and random histograms priors from Section 2.4. The priors **(S), (H)** are 'multiscale' priors where high frequencies are ignored – corresponding to truncated series priors considered frequently in the nonparametric Bayes literature. The resulting posterior distributions $\Pi(\cdot|X^{(n)})$ attain minimax optimal contraction rates up to logarithmic terms in Hellinger and L^2-distance ([100], [88], [P13]) and L^∞-distance ([P12]) as we stated in Section 2.4. Clearly other priors are of interest as well, for instance priors without or with random high-frequency cut-off, or Dirichlet mixtures of normals etc. While our current proofs do not cover such situations, one can note that our proof strategy via simultaneous control of many linear functionals is applicable in such situations as well. Generalising the scope of our techniques is an interesting direction of future research.

The projection $P_n(j)$ as in (3.31), with the choice $j = L_n$ from (3.33), defines a tight random variable in $\mathcal{M}_0$. For $z \in \mathcal{M}_0$, the map $\tau_z : f \mapsto \sqrt{n}(f - z)$ maps $\mathcal{M}_0 \to \mathcal{M}_0$, and we can define the shifted posterior $\Pi(\cdot \mid X^{(n)}) \circ \tau_{P_n(L_n)}^{-1}$. The following theorem shows that the above priors satisfy a weak BvM theorem in $\mathcal{M}_0$ in the sense of Definition 3.2, with efficient centring $P_n(L_n)$ (cf. Theorem 3.9). Denote the law $\mathcal{L}(\mathbb{G}_{P_0})$ of $\mathbb{G}_{P_0}$ from Proposition 3.2 by $\mathcal{N}$.

Theorem 3.10 *Let $\mathcal{M}_0 = \mathcal{M}_0(w)$ for any admissible $w = (w_l)$. Let $X^{(n)} = (X_1, \dots, X_n)$ i.i.d. from law P_0 with density $f_0 \in \mathcal{F}_0$. Let Π be a prior on the set of probability densities $\mathcal{F}$ that is*

1. *either of type* **(S)**, *in which case one assumes* $\log f_0 \in C^\alpha$ *for some* $\alpha > 1$,

2. *or of type* **(H)**, *and one assumes* $f_0 \in C^\alpha$ *for some* $1/2 < \alpha \leq 1$.

Suppose the prior parameters satisfy (3.33), (2.38) *and* (2.40). *Let* $\Pi(\cdot \,|\, X^{(n)})$ *be the induced posterior distribution on* $\mathcal{M}_0$. *Then, as* $n \to \infty$,

$$\beta_{\mathcal{M}_0}(\Pi(\cdot \,|\, X^{(n)}) \circ \tau^{-1}_{P_n(L_n)}, \mathcal{N}) \to^{P_0^{\mathbb{N}}} 0. \tag{3.34}$$

Application to credible bands. As the multiscale spaces $\mathcal{M}_0(w)$ are defined via maxima of collections of wavelet coefficients, they are particularly well-suited to the study of the supremum norm. Indeed, by analogy to the study of the credible sets in the previous section, which were shown to have (nearly) optimal diameter in L^2, it is possible up to minor adaptations to carry out the same method to obtain credible sets which are confidence *bands* having an optimal diameter (up to an arbitrary undersmoothing factor $M_n \to \infty$) in $L^\infty[0,1]$. We refer to [P14], Section 4.2 for explicit statement. Instead here we shall present in some detail a different application.

Bayesian Donsker's theorem. Whenever a prior on f satisfies the weak Bernstein-von Mises phenomenon in the sense of Definition 3.2, we can deduce from the continuous mapping theorem many BvMs via continuous functionals from $\mathcal{M}_0(w)$ to arbitrary spaces.

One may do so for integral functionals $L_g(f) = \int_0^1 g(x)f(x)dx$ *simultaneously* for many g's satisfying bounds on the decay of their wavelet coefficients. More precisely a bound $\sum_k |\langle g, \psi_{lk}\rangle| \leq c_l$ for all l combined with a weak BvM for (w_l) such that $\sum c_l w_l < \infty$ is sufficient. Let us illustrate this in a key example $g_t = 1_{[0,t]}, t \in [0,1]$, where we can derive results paralleling the classical Donsker theorem for distribution functions and its BvM version for the Dirichlet process proved by Albert Lo (1983) [78]. For simplicity we restrict to situations where the posterior $f \,|\, X^{(n)}$ is supported in L^2, and where the centering T_n in Definition 3.2 is contained in L^2. In that case the primitives

$$F(t) = \int_0^t f(x)dx, \quad \mathbb{T}_n(t) = \int_0^t T_n(x)dx$$

define random variables in the separable space $C([0,1])$ of continuous functions on $[0,1]$, and we can formulate a BvM-result in that space. Different centerings, such as the empirical distribution function, are discussed below.

Theorem 3.11 *Let* Π *be a prior supported in* $L^2([0,1])$ *and suppose the weak Bernstein - von Mises phenomenon in the sense of Definition 3.2 holds true in* $\mathcal{M}_0(w)$ *for some sequence* (w_l) *such that* $\sum_l w_l 2^{-l/2} < \infty$, *and with centering* $T_n \in L^2$. *Define the posterior cumulative distribution function*

$$F(t) = \int_0^t f(x)dx, \ t \in [0,1]. \tag{3.35}$$

Let G *be a* P_0*-Brownian bridge* $(G(t) \equiv G_{P_0}(t) : t \in [0,1])$, $dP_0(x) = f_0(x)dx$, $f_0 \in L^\infty$,. *If* $X^{(n)} \sim P^n_{f_0}$ *for some fixed* f_0 *then as* $n \to \infty$,

$$\beta_{C([0,1])}(\mathcal{L}(\sqrt{n}(F - \mathbb{T}_n) \,|\, X^{(n)}), \mathcal{L}(G)) \to^{P^n_{f_0}} 0, \tag{3.36}$$

$$\beta_{\mathbb{R}}(\mathcal{L}(\sqrt{n}\|F - \mathbb{T}_n\|_\infty \,|\, X^{(n)}), \mathcal{L}(\|G\|_\infty)) \to^{P^n_{f_0}} 0. \tag{3.37}$$

Let us now apply this result to the case of priors **(S)** or **(H)**, for which the weak BvM has been obtained above with centering the truncated empirical measure $P_n(L_n)$. Theorem 3.11 leads to a result with centering the primitive of $P_n(L_n)$. One can check that this can be replaced by

$$F_n(t) = \frac{1}{n}\sum_{i=1}^n 1_{[0,t]}(X_i), \quad t \in [0,1],$$

the empirical distribution function based on a sample $X_1, \ldots, X_n$.

Corollary 3.2 *Let* Π *be a prior of type* **(S)** *or* **(H)** *and suppose the conditions of Theorem 3.10 are satisfied. Then, as* $n \to \infty$,

$$\beta_{L^\infty([0,1])}(\mathcal{L}(\sqrt{n}(F - F_n) \mid X^{(n)}), \mathcal{L}(G_{P_0})) \to^{P^n_{f_0}} 0,$$

$$\beta_{\mathbb{R}}(\mathcal{L}(\sqrt{n}\|F - F_n\|_\infty \mid X^{(n)}), \mathcal{L}(\|G_{P_0}\|_\infty)) \to^{P^n_{f_0}} 0.$$

This result is the Bayesian analogue of Donsker's theorem for the empirical distribution function F_n. To our knowledge, most results of this kind in a Bayesian context have been obtained under some form of, at least partial, conjugacy of the model and prior. Note that here the results are obtained from general principles. Other examples of priors are currently under investigation.

3.5 Perspectives

As mentioned early in the Introduction, the measure of spread naturally provided by the posterior distribution is potentially a very interesting tool for building confidence regions.

In semiparametric problems, if the Bernstein-von Mises theorem holds, the posterior quantiles immediately give natural and asymptotically efficient confidence regions. As we have seen, obtaining semiparametric BvM results is a complex question in general: in particular, the choice of the prior on the nonparametric part is crucial. It would be interesting to develop semiparametric BvMs for further classes of priors, and in particular provide further functional 'change of variable' guarantees that ensure that no additional bias appears in the limiting posterior marginal.

In nonparametric problems, constructing confidence sets is an interesting and challenging question. In the previous sections we have provided some tools and a possible way to construct fixed regularity confident credible sets. Building adaptive confidence sets is a natural further question, though this becomes often a somewhat qualitatively different problem: in particular, rates may change, as is well known from the frequentist analysis of the problem, see Low (1997) [80]. Recent contributions on the question, where some further references can be found, include [51] following a non-Bayesian approach and [96], where a Bayesian approach is considered.

Publications

All papers are available on my webpage http://www.proba.jussieu.fr/~castillo

[P1] I. Castillo. "Penalized profile likelihood methods and second order properties in semiparametric models". PhD thesis. Université Paris-Sud Orsay, 2006.

[P2] I. Castillo, C. Lévy-Leduc, and C. Matias. Exact adaptive estimation of the shape of a periodic function with unknown period corrupted by white noise. *Mathematical Methods of Statistics* 15, pp. 146–175, 2006.

[P3] I. Castillo. Semi-parametric second-order efficient estimation of the period of a signal. *Bernoulli* 13, pp. 910–932, 2007.

[P4] I. Castillo. Lower bounds for posterior rates with Gaussian process priors. *Electronic Journal of Statistics* 2, pp. 1281–1299, 2008.

[P5] I. Castillo and J.-M. Loubes. Estimation of the distribution of random shifts deformation. *Mathematical Methods of Statistics* 18, pp. 21–42, 2009.

[P6] I. Castillo and E. Cator. Semiparametric shift estimation based on the cumulated periodogram for non-regular functions. *Electronic Journal of Statistics* 5, pp. 102–126, 2011.

[P7] I. Castillo. A semiparametric Bernstein-von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields* 152, pp. 53–99, 2012.

[P8] I. Castillo. Semiparametric Bernstein–von Mises theorem and bias, illustrated with Gaussian process priors. *Sankhya A* 74, pp. 194–221, 2012.

[P9] I. Castillo and A. van der Vaart. Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *The Annals of Statistics* 40, pp. 2069–2101, 2012.

[P10] I. Castillo and R. Nickl. Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *The Annals of Statistics* 41, pp. 1999–2028, 2013.

[P11] I. Castillo, G. Kerkyacharian, and D. Picard. Thomas Bayes' walk on manifolds. *Probability Theory and Related Fields* 158, pp. 665–710, 2014.

[P12] I. Castillo. On Bayesian Supremum norm contraction rates. *The Annals of Statistics.* 42, pp. 2058–2091, 2014.

[P13] I. Castillo and J. Rousseau. A Bernstein–von Mises theorem for smooth functionals in semiparametric models. Submitted. arXiv:1305.4482, 2013.

[P14] I. Castillo and R. Nickl. On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *The Annals of Statistics.* 42, pp. 1941–1969, 2014.

[P15] I. Castillo, J. Schmidt-Hieber, and A. van der Vaart. Bayesian linear regression with sparse priors. Submitted. arXiv:1403.0735, 2014.

Bibliography

[1] F. Abramovich et al. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* 34, pp. 584–653, 2006.

[2] P. Alquier. *Contributions to statistical learning in sparse models.* Habilitation à diriger des recherches, Université Paris 7. 2013.

[3] P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Stat.* 5, pp. 127–145, 2011.

[4] E. Arias-Castro and K. Lounici. Estimation and variable selection with exponential weights. *Electron. J. Stat.* 8, pp. 328–354, 2014.

[5] A. Barron. *The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions.* Tech. rep. 7. Dept. Statistics, Univ. Illinois, Champaign, 1988.

[6] A. Barron, M. J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* 27, pp. 536–561, 1999.

[7] T. Bayes. An essay towards solving a problem in the doctrine of chances. *R. Soc. Lond. Philos. Trans.* 53, pp. 370–418, 1763.

[8] E. Belitser and S. Ghosal. Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.* 31 Dedicated to the memory of Herbert E. Robbins, pp. 536–559, 2003.

[9] A. Bhattacharya, D. Pati, and D. Dunson. Anisotropic function estimation using multi-bandwidth Gaussian processes. *Ann. Statist.* 42, pp. 352–381, 2014.

[10] P. J. Bickel and B. J. K. Kleijn. The semiparametric Bernstein-von Mises theorem. *Ann. Statist.* 40, pp. 206–237, 2012.

[11] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* 37, pp. 1705–1732, 2009.

[12] P. Billingsley and F. Topsøe. Uniformity in weak convergence. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 7, pp. 1–16, 1967.

[13] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* 65, pp. 181–237, 1983.

[14] L. Birgé. "Robust testing for independent nonidentically distributed variables and Markov chains". In: *Specifying statistical models (Louvain-la-Neuve, 1981).* Vol. 16. Lecture Notes in Statist. Springer, New York, 1983. Pp. 134–162.

[15] L. Birgé. Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.* 3, pp. 259–282, 1984.

[16] L. Birgé. "Robust tests for model selection". In: *From probability to statistics and back: high-dimensional models and processes.* Vol. 9. Inst. Math. Stat. (IMS) Collect. Inst. Math. Statist., Beachwood, OH, 2013. Pp. 47–64.

[17] L. Birgé. About the non-asymptotic behaviour of Bayes estimators. Arxiv preprint arXiv:1402.3695, 2014.

[18] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* 3, pp. 203–268, 2001.

[19] D. Bontemps. Bernstein-von Mises theorems for Gaussian regression with increasing number of regressors. *Ann. Statist.* 39, pp. 2557–2584, 2011.

[20] L. Bottolo and S. Richardson. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Anal.* 5, pp. 583–618, 2010.

[21] S. Boucheron and E. Gassiat. A Bernstein-von Mises theorem for discrete probability distributions. *Electron. J. Stat.* 3, pp. 114–148, 2009.

[22] L. Breiman, L. Le Cam, and L. Schwartz. Consistent estimates and zero-one sets. *Ann. Math. Statist.* 35, pp. 157–161, 1964.

[23] P. Bühlmann and S. van de Geer. *Statistics for High-dimensional Data*. Berlin: Springer, 2011.

[24] P. Bühlmann et al. Correlated variables in regression: Clustering and sparse estimation. *J. Statist. Plann. Inference* 143, pp. 1835–1858, 2013.

[25] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* 35, pp. 2313–2351, 2007.

[26] C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika* 97, pp. 465–480, 2010.

[27] O. Catoni. *Statistical learning theory and stochastic optimization.* Vol. 1851. Lecture Notes in Mathematics. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. Springer-Verlag, Berlin, 2004. Pp. viii+272.

[28] B. Clarke and S. Ghosal. Reference priors for exponential families with increasing dimension. *Electron. J. Stat.* 4, pp. 737–780, 2010.

[29] A. Cohen, I. Daubechies, and P. Vial. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* 1, pp. 54–81, 1993.

[30] D. D. Cox. An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* 21, pp. 903–923, 1993.

[31] A. S. Dalalyan and A. B. Tsybakov. "Aggregation by exponential weighting and sharp oracle inequalities". In: *Learning theory*. Vol. 4539. Lecture Notes in Comput. Sci. Berlin: Springer, 2007. Pp. 97–111.

[32] P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *Ann. Statist.* 14 With a discussion and a rejoinder by the authors, pp. 1–67, 1986.

[33] D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. on Information Theory* 52, pp. 6–18, 2006.

[34] D. L. Donoho et al. Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* 54 With discussion and a reply by the authors, pp. 41–81, 1992.

[35] D. L. Donoho et al. Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B* 57 With discussion and a reply by the authors, pp. 301–369, 1995.

[36] J. L. Doob. "Application of the theory of martingales". In: *Le Calcul des Probabilités et ses Applications*. Colloques Internationaux du CNRS, no. 13. CNRS, Paris, 1949. Pp. 23–27.

[37] R. M. Dudley. *Real analysis and probability*. Cambridge, UK, 2002. Pp. x+555.

[38] S. Efromovich. On sharp adaptive estimation of multivariate curves. *Math. Methods Statist.* 9, pp. 117–139, 2000.

[39] J. Fabius. Asymptotic behavior of Bayes' estimates. *Ann. Math. Statist.* 35, pp. 846–856, 1964.

[40] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.* 1, pp. 209–230, 1973.

[41] D. Freedman. On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* 27, pp. 1119–1140, 1999.

[42] D. A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* 34, pp. 1386–1403, 1963.

[43] E. I. George. The variable selection problem. *J. Amer. Statist. Assoc.* 95, pp. 1304–1308, 2000.

[44] E. I. George and D. P. Foster. Calibration and empirical Bayes variable selection. *Biometrika* 87, pp. 731–747, 2000.

[45] S. Ghosal. Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli* 5, pp. 315–331, 1999.

[46] S. Ghosal. Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Mult. Anal.* 74, pp. 49–68, 2000.

[47] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. "Consistency issues in Bayesian nonparametrics". In: *Asymptotics, nonparametrics, and time series*. Vol. 158. Statist. Textbooks Monogr. Dekker, New York, 1999. Pp. 639–667.

[48] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.* 28, pp. 500–531, 2000.

[49] S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* 35, pp. 192–223, 2007.

[50] E. Giné and R. Nickl. Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.* 39, pp. 2883–2911, 2011.

[51] E. Giné and R. Nickl. Confidence bands in density estimation. *Ann. Statist.* 38, pp. 1122–1170, 2010.

[52] G. K. Golubev. Reconstruction of sparse vectors in white Gaussian noise. *Problemy Peredachi Informatsii* 38, pp. 75–91, 2002.

[53] A. Grigor'yan. *Heat kernel and analysis on manifolds*. Vol. 47. AMS/IP Studies in Advanced Mathematics. Providence, RI: American Mathematical Society, 2009. Pp. xviii+482.

[54] W. Härdle et al. *Wavelets, approximation, and statistical applications*. Vol. 129. Lecture Notes in Statistics. New York: Springer-Verlag, 1998. Pp. xviii+265.

[55] M. Hoffmann, J. Rousseau, and J. Schmidt-Hieber. On adaptive posterior concentration rates. Arxiv preprint arXiv:1305.5270, 2013.

[56] H. Ishwaran and J. S. Rao. Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Statist.* 33, pp. 730–773, 2005.

[57] W. Jiang and C.-H. Zhang. General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* 37, pp. 1647–1684, 2009.

[58] I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* 32, pp. 1594–1649, 2004.

[59] R. de Jonge and J. H. van Zanten. Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Ann. Statist.* 38, pp. 3300–3320, 2010.

[60] G. Kerkyacharian and D. Picard. Minimax or maxisets? *Bernoulli* 8, pp. 219–253, 2002.

[61] G. Kerkyacharian et al. "Learning out of leaders". In: *Multiscale, nonlinear and adaptive approximation*. Springer, Berlin, 2009. Pp. 295–324.

[62] Y. Kim. The Bernstein-von Mises theorem for the proportional hazard model. *Ann. Statist.* 34, pp. 1678–1700, 2006.

[63] Y. Kim and J. Lee. A Bernstein-von Mises theorem in the nonparametric right-censoring model. *Ann. Statist.* 32, pp. 1492–1512, 2004.

[64] J. F. C. Kingman. Completely random measures. *Pacific J. Math.* 21, pp. 59–78, 1967.

[65] J. Kuelbs and W. V. Li. Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal.* 116, pp. 133–157, 1993.

[66] P. Laplace. Mémoire sur les formules qui sont fonctions de très grands nombres et sur leurs applications aux probabilités. *Oeuvres de Laplace* 12, pp. 301–345, 1810.

[67] M. Lavine. Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* 20, pp. 1222–1235, 1992.

[68] L. Le Cam. On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Publ. Statist.* 1, pp. 277–329, 1953.

[69] L. Le Cam. *Théorie asymptotique de la décision statistique*. Séminaire de Mathématiques Supérieures, No. 33 (Été, 1968). Les Presses de l'Université de Montréal, Montreal, Que., 1969. P. 140.

[70] L. Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.* 1, pp. 38–53, 1973.

[71] L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. New York: Springer-Verlag, 1986. Pp. xxvi+742.

[72] L. Le Cam and G. L. Yang. *Asymptotics in statistics*. Springer Series in Statistics. Some basic concepts. New York: Springer-Verlag, 1990. Pp. viii+180.

[73] H. Leahu. On the Bernstein-von Mises phenomenon in the Gaussian white noise model. *Electron. J. Stat.* 5, pp. 373–404, 2011.

[74] M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Vol. 23. Isoperimetry and processes. Berlin: Springer-Verlag, 1991.

[75] P. J. Lenk. The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.* 83, pp. 509–516, 1988.

[76] T. Leonard. Density estimation, stochastic processes and prior information. *J. Roy. Statist. Soc. Ser. B* 40 With discussion, pp. 113–146, 1978.

[77] W. V. Li and Q.-M. Shao. "Gaussian processes: inequalities, small ball probabilities and applications". In: *Stochastic processes: theory and methods*. Vol. 19. Handbook of Statist. Amsterdam: North-Holland, 2001. Pp. 533–597.

[78] A. Lo. Weak convergence for Dirichlet processes. *Sankhyā* 45, pp. 105–111, 1983.

[79] K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics* 2, pp. 90–102, 2008.

[80] M. G. Low. On nonparametric confidence intervals. *Ann. Statist.* 25, pp. 2547–2554, 1997.

[81] R. D. Mauldin, W. D. Sudderth, and S. C. Williams. Pólya trees and random distributions. *Ann. Statist.* 20, pp. 1203–1221, 1992.

[82] B. McNeney and J. A. Wellner. Application of convolution theorems in semiparametric models with non-i.i.d. data. *J. Statist. Plann. Inference* 91, pp. 441–480, 2000.

[83] R. von Mises. *Wahrscheinlichkeitsrechnung*. Vienna: Deuticke, 1931.

[84] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* 83, pp. 1023–1036, 1988.

[85] M. Panov and V. Spokoiny. Critical dimension in semiparametric Bernstein – von Mises Theorem. *ArXiv e-prints* arXiv1310.7796.

[86] S. Petrone and L. Wasserman. Consistency of Bernstein polynomial posteriors. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 64, pp. 79–100, 2002.

[87] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* 39, pp. 731–771, 2011.

[88] V. Rivoirard and J. Rousseau. Bernstein-von Mises theorems for linear functionals of the density. *Ann. Statist.* 40, pp. 1489–1523, 2012.

[89] V. Rockova and E. I. George. EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association* To appear, 2014.

[90] J. Rousseau. Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.* 38, pp. 146–180, 2010.

[91] A. Schreck et al. A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection. *ArXiv e-prints* arXiv 1312.5658, 2013.

[92] L. Schwartz. On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 4, pp. 10–26, 1965.

[93] J. G. Scott and J. O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* 38, pp. 2587–2619, 2010.

[94] X. Shen. Asymptotic normality of semiparametric and nonparametric posterior distributions. *J. Amer. Statist. Assoc.* 97, pp. 222–235, 2002.

[95] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.* 29, pp. 687–714, 2001.

[96] B. Szabó, A. W. van der Vaart, and H. van Zanten. Frequentist coverage of adaptive nonparametric Bayesian credible sets. Arxiv preprint arXiv:1310.4489, 2013.

[97] R. Tibshirani. Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* 58, pp. 267–288, 1996.

[98] A. W. van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998. Pp. xvi+443.

[99] A. W. van der Vaart and H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* 37, pp. 2655–2675, 2009.

[100] A. W. van der Vaart and H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* 36, pp. 1435–1463, 2008.

[101] A. W. van der Vaart and H. van Zanten. Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res.* 12, pp. 2095–2119, 2011.

[102] A. W. van der Vaart and H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. *IMS Collections* 3, pp. 200–222, 2008.

[103] S. Walker. New approaches to Bayesian consistency. *Ann. Statist.* 32, pp. 2028–2043, 2004.

[104] S. Walker, A. Lijoi, and I. Prünster. On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.* 35, pp. 738–746, 2007.

[105] M. Yuan and Y. Lin. Efficient empirical Bayes variable selection and estimation in linear models. *J. Amer. Statist. Assoc.* 100, pp. 1215–1225, 2005.

[106] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* 38, pp. 894–942, 2010.

[107] C.-H. Zhang and J. Huang. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* 36, pp. 1567–1594, 2008.

[108] L. H. Zhao. Bayesian aspects of some nonparametric problems. *Ann. Statist.* 28, pp. 532–552, 2000.