Nonparametric Bernstein – von Mises Theorems

Ismaël Castillo and Richard Nickl

CNRS - Laboratoire Probabilités et Modèles Aléatoires
Universités Paris VI & VII
175, Rue de Chevaleret,
75013 Paris, France.

ismael.castillo@upmc.fr

Statistical Laboratory
Department of Pure Mathematics and Mathematical Statistics
University of Cambridge
CB3 0WB Cambridge, UK.

r.nickl@statslab.cam.ac.uk

Abstract: We prove Bernstein – von Mises theorems for nonparametric Bayes priors in the Gaussian white noise model. We demonstrate how such results justify Bayes methods as efficient frequentist inference procedures in a variety of concrete nonparametric problems. Particularly we investigate frequentist coverage properties of Bayesian credible sets. Applications include Sobolev goodness-of fit tests, general classes of linear and nonlinear functionals, and credible bands for Fourier coefficients and autoconvolutions. The assumptions cover non-conjugate product priors defined on general orthonormal bases of L^2 satisfying weak conditions.

AMS 2000 subject classifications: Primary 62G20; secondary 62G15, 62G08.
Keywords and phrases: Bayesian inference, plug-in property, efficiency.

1. Introduction

Consider observing a random sample $X^{(n)}$ of size n, or at noise level $n^{-1/2}$, drawn from distribution P_f^n indexed by some unknown parameter $f \in \mathcal{F}$. The Bayesian paradigm views the sample as having law P_f^n conditionally on f, that is, $X^{(n)}|f \sim P_f^n$, where f is drawn from some *prior probability distribution* Π on some σ-field $\mathcal{B}$ of $\mathcal{F}$. The random variable $f|X^{(n)}$ then has a law on $\mathcal{F}$ which is known as the *posterior distribution*, denoted by $\Pi(\cdot|X^{(n)})$. Bayesian inference on f is then entirely based on this posterior distribution – it gives access to point estimates for f, credible sets and tests in a natural way.

It is of interest to analyse the behaviour of $\Pi(\cdot|X^{(n)})$ under the frequentist sampling assumption that $X^{(n)}$ is drawn from $P_{f_0}^n$ for some fixed nonrandom $f_0 \in \mathcal{F}$, in particular it seems important to understand to which extent Bayesian procedures based on the posterior lead to valid frequentist inference.

Quite remarkably if $\mathcal{F}$ is a *finite-dimensional* space then in most situations posterior-based inference is not only valid from a frequentist perspective, but in fact asymptotically optimal. Perhaps the most fundamental explanation for this phenomenon is given by the Bernstein – von Mises (BvM) theorem, first discovered by Laplace [26], developed by von Mises (see p.156f. in [40]) and put into the framework of modern parametric statistics by Le Cam [27]. It states that, under mild and universal assumptions on the prior, the posterior distribution approximately equals a normal distribution on $\mathcal{F}$, centered at an efficient estimator $\hat{f}_n$ for f_0, and with covariance $i(f_0)$ the Cramér-Rao information bound in the statistical model considered:

$$\sup_{B\in\mathcal{B}}\left|\Pi(B|X^{(n)}) - N(\hat{f}_n, i(f_0))(B)\right| \to 0 \tag{1}$$

as $n \to \infty$ in $P^n_{f_0}$-probability. Practically this means that posterior-based inference asymptotically coincides with inference based on standard efficient, $1/\sqrt{n}$-consistent frequentist estimators of f_0, and that Bayesian methods can be rigorously justified from an asymptotic frequentist point of view.

The last decade has seen remarkable activity in the development of *nonparametric* Bayes procedures, where $\mathcal{F}$ is taken to be an infinite-dimensional space, typically consisting of functions or infinite vectors: nonparametric regression, classification, density estimation, normal means and Gaussian white noise models come to mind, and a variety of nonparametric priors have been devised in the literature for such models. Posteriors in such models can be computed efficiently by algorithms such as MCMC, and they provide broadly applicable Bayesian inferential tools for nonparametric problems. It is natural to ask whether an analogue of (1) can still be proved in such situations, as it would give a general justification for the use of nonparametric Bayes procedures. Although remarkable progress has been made in the understanding of the frequentist properties of nonparametric Bayes procedures – we refer here only to some of the key papers such as [17, 18, 34, 36, 38] and references therein – a satisfactory answer to the BvM-question seems not to have been found. A first reason is perhaps that it is not immediately clear what $N(\hat{f}_n, i(f_0))$ should be replaced by in the infinite-dimensional situation – Gaussian distributions over infinite dimensional spaces $\mathcal{F}$ are much more complex objects, and their existence (in the form relevant here) depends on the topology that $\mathcal{F}$ is endowed with. Likewise whether $1/\sqrt{n}$-efficient estimators $\hat{f}_n$ for f_0 exist or not depends strongly on the notion of distance on $\mathcal{F}$ that one employs, and many of the commonly used loss functions in nonparametric statistics (such as L^p-type loss) are not admissible.

A first step to understand this phenomenon better is thus to look for loss functions on $\mathcal{F}$ for which efficient frequentist estimators $\hat{f}_n$ with certain Gaussian limit distributions exist. This leads naturally to the setting of empirical processes: For example in the situation where one observes a random sample $X_1, \dots, X_n$ from a law P on $[0, 1]$ we know that the empirical measure $P_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ is a $1/\sqrt{n}$-efficient estimator for P in the space of bounded functions $\ell^\infty(\mathcal{H})$ on any P-Donsker class $\mathcal{H}$ of functions $h : [0, 1] \to \mathbb{R}$. More concretely this means that for such $\mathcal{H}$

$$\|P_n - P\|_{\mathcal{H}} = \sup_{h\in\mathcal{H}} \left| \int hd(P_n - P) \right| = \sup_{h\in\mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^{n}(h(X_i) - Eh(X)) \right| \quad (2)$$

is $O_P(n^{-1/2})$, in fact $\sqrt{n}(P_n - P)$ converges in distribution in $\ell^\infty(\mathcal{H})$ to a tight Gaussian random variable $N(0, i(P))$ over $\ell^\infty(\mathcal{H})$, and the covariance structure $i(P)$ of the limiting variable achieves the Cramér-Rao information bound for the fully nonparametric model (see Section 3.11.1 in [39]).

A natural setting for nonparametric Bernstein-von Mises theorems is thus to embed the parameter space $\mathcal{F}$ into an $\ell^\infty(\mathcal{H})$-type space. The purpose of the present paper is to investigate this approach rigorously in the situation of the Gaussian white noise model, and with $\mathcal{H}_s$ a ball in a Sobolev space of order $s > 1/2$ – this makes the mathematical analysis tractable without any severe loss of conceptual generality. Our main results will imply that for a large and relevant class of product priors Π that satisfy mild assumptions, and which do not require conjugacy, one has

$$\sup_{A\in\mathcal{A}} \left| \Pi(A|X^{(n)}) - \mathcal{N}(\hat{f}_n, i(f_0))(A) \right| \to^{P^n_{f_0}} 0 \quad (3)$$

where $\mathcal{N}$ is a Gaussian measure on $\ell^\infty(\mathcal{H}_s)$ centered at an efficient estimator $\hat{f}_n$ of f_0, both to be defined in a precise manner, and where the classes $\mathcal{A}$ consist of measurable subsets of $\ell^\infty(\mathcal{H}_s)$ that have uniformly smooth boundaries for the measure $\mathcal{N}$. The result is proved by showing that the (suitably shifted) posterior converges weakly (in $P^n_{f_0}$-probability) to a canonical Gaussian measure $\mathcal{N}$ in $\ell^\infty(\mathcal{H}_s)$, and by exploiting the uniformity classes for weak convergence towards $\mathcal{N}$. We should note that some restrictions on the class $\mathcal{A}$ are necessary as one can show that in the infinite-dimensional situation the Bernstein-von Mises theorem cannot hold uniformly in all Borel sets of $\ell^\infty(\mathcal{H}_s)$ (see below Definition 1 for more discussion). Our assumptions apply in particular to priors that produce posteriors which achieve frequentist optimal contraction rates in stronger loss functions (such as L^2-distance) and which resemble the state of the art prior choices in the nonparametric Bayes literature.

Our abstract results clearly only gain relevance through the fact that we can demonstrate their applicability: The general result (3) implies that posterior-based credible regions give asymptotically exact frequentist confidence sets in a variety of concrete problems of nonparametric inference, chosen to highlight the scope of our techniques: Our examples, which are given in Section 2, include weighted L^2-ellipsoid credible regions for the unknown parameter f_0, linear functionals defined on L^2 such as moments of f_0, credible bands for the Fourier coefficients of f_0, a general class of nonlinear functionals defined on L^2 such as the squared L^2-norm $\|f_0\|_2^2$, and simultaneous credible bands for the auto-convolution $f_0 * f_0$.

We note that a key point in these applications is related to the notion of the 'plug-in property' of a nonparametric estimator, coined by Bickel and Ritov

[3]. A standard nonparametric estimator that is rate-optimal in a standard loss function (such as L^p-loss) is said to have the plug-in property if it simultaneously achieves the $1/\sqrt{n}$-rate for a large class of linear functionals, that is, if (2) holds with P_n replaced by the nonparametric estimator. Standard frequentist estimators such as kernel, wavelet and nonparametric maximum likelihood estimators satisfy this property, in fact one can even prove a corresponding uniform central limit theorem in $\ell^\infty(\mathcal{H})$ for such estimators, see Kiefer and Wolfowitz [25], Nickl [29], Giné and Nickl [21, 22]. Our results imply that this is also true in the Bayesian situation: The posterior contracts at the optimal rate in L^2-loss and at the same time satisfies a Bernstein-von Mises theorem in $\ell^\infty(\mathcal{H}_s)$. Likewise, formal Bayes estimators such as the posterior mean are efficient frequentist estimators of f_0 in $\ell^\infty(\mathcal{H}_s)$ while achieving the optimal nonparametric rate for f_0 in L^2.

There is some recent work on the BvM phenomenon for nonparametric procedures that needs mentioning. Leahu [28] derives interesting results on the possibility and impossibility of BvM-theorems for undersmoothing priors – his negative results will be relevant below. His positive findings are, however, strongly tied to the Gaussian conjugate situation, and do not address efficiency questions. Rivoirard and Rousseau [31] consider BvM-type results for linear functionals of certain models of probability density functions. A number of BvM-type results have been obtained for the fixed finite-dimensional posterior with dimension increasing to infinity: Ghosal [15] and Bontemps [6] consider regression with a finite number of regressors, Ghosal [16] and Clarke and Ghosal [9] consider exponential families, and the case of discrete probability distributions is treated in Boucheron and Gassiat [7]. We believe that our approach allows for a unifying framework for these results: We show that a 'functional' BvM-result holds, giving access to many linear functionals in a uniform way. We finally note related work on semiparametric BvM-results in Castillo [8] and Bickel and Kleijn [2].

The outline of this article is as follows: In the next two subsections we introduce a general notion of the nonparametric Bernstein – von Mises phenomenon. In Section 2 we show that when this phenomenon holds, posterior-based inference is valid from a frequentist point of view in a variety of concrete examples from nonparametric statistics. In Section 3 we prove that for a large class of natural priors on L^2, the BvM phenomenon indeed occurs.

1.1. The Weak Nonparametric Bernstein – von Mises Phenomenon

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

$$dX^{(n)}(t) = f(t)dt + \frac{1}{\sqrt{n}}dW(t), \quad t \in [0,1]. \tag{4}$$

Let Π be a prior Borel probability distribution on L^2 and let $\Pi(\cdot|X^{(n)})$ be the posterior distribution on L^2 given the observed trajectory $X^{(n)}$.

Except in conjugate situations the proof of a Bernstein – von Mises type result rests typically on the fact that efficient estimation at the rate $1/\sqrt{n}$ is possible. In the nonparametric situation this rules out L^p-type loss functions, but leads one to consider weaker norms of the type $\|\cdot\|_{\mathcal{H}}$ discussed in (2). For the particular choice of $\mathcal{H}_s$ equal to an order-s Sobolev-ball we can understand this better by using simple but useful Hilbert space duality arguments in the nested scale of Sobolev spaces $\{H_2^r\}_{r\in\mathbb{R}}$ on $[0,1]$: We define these in precise detail below, but note for the moment that $H_2^0 = L^2$ and that the norm-continuous imbeddings

$$H_2^r \subseteq H_2^t,\ r \geq t,$$

hold, so to weaken the norm beyond L^2 means that we should decrease r to be negative. For $s > 0$ the space H_2^{-s} can be realised in an isometric way as a closed subspace of $\ell^\infty(\mathcal{H}_s)$. Although we shall never use this isometry and stick to negative order Sobolev spaces from now on, it heuristically explains the connection to the discussion surrounding (2) above. Some more thought shows that one should increase s so far that the Gaussian experiment in (4) is well defined as a tight random element of H_2^{-s}. This happens precisely as soon as s exceeds $1/2$, in fact we show below that the random trajectory $dX^{(n)}$ defines a tight Gaussian Borel random variable $\mathbb{X}^{(n)}$ on H_2^{-s} with mean f and covariance $n^{-1}I$. Thus if we denote by $\mathbb{W}$ the centered Gaussian Borel random variable on H_2^{-s} with covariance I, then (4) can be written as

$$\mathbb{X}^{(n)} = f + \frac{1}{\sqrt{n}}\mathbb{W}, \tag{5}$$

a natural Gaussian shift experiment in H_2^{-s}. One can show moreover that $\mathbb{X}^{(n)}$ is an efficient estimator for f for the loss function of H_2^{-s}.

Now any prior and posterior probability distribution on L^2 defines a tight Borel probability measure on H_2^{-s} simply by the continuous injection $L^2 \subset H_2^{-s}$. On H_2^{-s} and for $z \in H_2^{-s}$, define the measurable transformation

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

Let $\Pi_n = \Pi(\cdot|\mathbb{X}^{(n)})$ be the posterior distribution on H_2^{-s}, and let $\Pi_n \circ \tau_{\mathbb{X}^{(n)}}^{-1}$ be its image 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_2^{-s}-neighborhoods of the efficient estimator $\mathbb{X}^{(n)}$. To compare probability distributions on H_2^{-s} we may use any metric for weak convergence of probability measures, and we choose the bounded Lipschitz metric here for convenience (it is defined in Section 4.2 below). Let $\mathcal{N}$ be the standard Gaussian probability measure on H_2^{-s} with mean zero and covariance I – its existence is ensured in Section 1.2.

Definition 1. *Consider data generated from equation* (4) *under a fixed function* f_0*, and denote by* $P_{f_0}^n$ *the distribution of* $\mathbb{X}^{(n)}$*. Let* $s > 1/2$ *and let* β *be the bounded-Lipschitz metric for weak convergence of probability measures on* H_2^{-s}*. We say that a prior* Π *satisfies the weak Bernstein – von Mises phenomenon in*

H_2^{-s} *if, as* $n \to \infty$,

$$\beta(\Pi_n \circ \tau^{-1}_{\mathbb{X}^{(n)}}, \mathcal{N}) \to^{P^n_{f_0}} 0. \tag{6}$$

We note that the fact that the result is phrased in a way in which $\mathcal{N}$ is independent of n is important. [The reason is that the metric β does not induce a uniformity structure for the topology of weak convergence (see the Remark on p.413 in [11]).]

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_0 in the Gaussian shift experiment (5) in H_2^{-s}-loss, and we shall see how this carries over to sufficiently regular real-valued functionals $\Psi(f_0)$, see Section 2.4 below.

One may ask by analogy to the finite-dimensional situation whether a *strong* Bernstein-von Mises phenomenon, where β is replaced by the total variation norm, can be proved. It follows from Theorem 2 in [28] that already in the Gaussian conjugate situation such a result is not possible unless one restricts to very specific priors (which in particular do not satisfy the plug-in property that will be relevant below).

Now with weak instead of total variation convergence we cannot infer that $\Pi_n \circ \tau^{-1}_{\mathbb{X}^{(n)}}$ and $\mathcal{N}$ are approximately the same for every Borel set in H_2^{-s}, but only for sets B that are continuity sets for the probability measure $\mathcal{N}$. As we shall see this includes, for instance, any fixed norm ball $B(0, M)$ in H_2^{-s}, so that for n large

$$\Pi(B(\mathbb{X}^{(n)}, M/\sqrt{n})|\mathbb{X}^{(n)}) \sim \mathcal{N}(B(0, M))$$

with $P^n_{f_0}$-probability close to one. For statistical applications of the Bernstein – von Mises phenomenon one typically needs some uniformity in B, and this is where total variation results would be particularly useful. Weak convergence in H_2^{-s} implies that $\Pi_n \circ \tau^{-1}_{\mathbb{X}^{(n)}}$ is close to $\mathcal{N}$ uniformly in certain classes of subsets of H_2^{-s} whose boundaries are sufficiently regular relative to the measure $\mathcal{N}$ (see Subsection 4.2), and we show below how this allows for enough uniformity to deal with a variety of concrete nonparametric statistical problems.

The Bernstein – von Mises phenomenon in (6) will often be complemented by convergence of moments, that is, convergence of the Bochner integrals (e.g., p.100 in [1])

$$\int_{H_2^{-s}} f d\Pi_n \circ \tau^{-1}_{\mathbb{X}^{(n)}}(f) \to^{P^n_{f_0}} \int_{H_2^{-s}} f d\mathcal{N}(f) = 0$$

as $n \to \infty$ in H_2^{-s}. This implies that the posterior mean $\bar{f}_n$ of Π_n satisfies

$$\|\bar{f}_n - \mathbb{X}^{(n)}\|_{-s,2} = o_P(n^{-1/2}), \tag{7}$$

so in semiparametric terminology the posterior mean is asymptotically linear in H_2^{-s} with respect to $\mathbb{X}^{(n)}$, so in particular efficient for f_0. Therefore, if (7) holds, centering Bayesian credible sets at $\bar{f}_n$ instead of $\mathbb{X}^{(n)}$ makes no asymptotic difference in H_2^{-s}.

1.2. Sobolev Spaces and White Noise

Before we proceed we need to fix ideas and formally introduce negative order Sobolev spaces. Their definition will be given in terms of orthonormal bases of L^2 which will also be relevant for the assumptions on the priors we need.

Denote by $\langle f, g\rangle = \int_0^1 f(x)\overline{g(x)}dx$ the standard inner product on L^2. We shall work with a general orthonormal basis of L^2 that satisfies the following weak regularity condition. While notationally it reflects a wavelet type basis $\{\psi_{lk} : l \geq J_0, 0 \leq k \leq 2^l - 1\}$ of CDV-type [10], with notational convention $\psi_{J_0 k} = \phi_k$, it also includes the standard trigonometric basis $\psi_{lk}(x) \equiv e_l(x) = e^{2\pi i l x}$ and bases of standard Karhunen-Loève expansions.

Definition 2. *Let $S \in \mathbb{N}$. By an S-regular basis $\{\psi_{lk} : l \in \mathcal{L}, k \in \mathcal{Z}_l\}$ of L^2 with index sets $\mathcal{L} \subset \mathbb{Z}, \mathcal{Z}_l \subset \mathbb{Z}$ and characteristic sequence a_l we shall mean any of the following:*
a) $\psi_{lk} \equiv e_l$ is S-times differentiable with all derivatives in L^2, $|\mathcal{Z}_l| = 1$, $a_l = |l| \wedge 1$, and $\{e_l : l \in \mathcal{L}\}$ forms an orthonormal basis of L^2.
b) ψ_{lk} is S-times differentiable with all derivatives in L^2, $\mathcal{L} = \mathbb{N} \cup \{0\}$, $a_l = |\mathcal{Z}_l| = 2^l$, and $\{\psi_{lk} : l \in \mathcal{L}, k \in \mathcal{Z}_l\}$ is an orthonormal basis of L^2.

Define for $0 \leq s < S$ the Sobolev spaces as

$$H_2^s := H_2^s([0,1]) := \left\{ f \in L^2([0,1]) : \|f\|_{s,2}^2 := \sum_{l \in \mathcal{L}} a_l^{2s} \sum_{k \in \mathcal{Z}_l} |\langle \psi_{lk}, f\rangle|^2 < \infty \right\},$$

which are Hilbert spaces that may depend on the basis functions used, but since for the examples mentioned above this is not the case we suppress it in the notation. Moreover

$$\left\{ \sum_{l \in \mathcal{L}'} \sum_{k \in \mathcal{Z}_l'} c_{lk}\psi_{lk} : c_{lk} \in \mathbb{C},\ \mathcal{Z}_l' \subset \mathcal{Z}_l, \mathcal{L}' \subset \mathcal{L} \text{ finite} \right\}$$

forms a dense subset of H_2^s: For fixed $\mathcal{L}'$ finite, $\mathcal{Z}_l'$ and $c_{lk} = \langle f, \psi_{lk}\rangle$ these sums are precisely all the finite-dimensional L^2-projections $\pi_V(f)$ of $f \in L^2$ onto V, where $V \equiv V_{\mathcal{L}', \mathcal{Z}_l'} = span\{\psi_{lk} : l \in \mathcal{L}', k \in \mathcal{Z}_l'\}$.

For $s > 0$ we define the dual space

$$H_2^{-s}([0,1]) := (H_2^s[0,1])^*.$$

Using standard duality arguments (as in Proposition 9.16 in [13]) one shows the following: H_2^{-s} consists precisely of those linear forms L acting on H_2^s for which the $\|L\|_{-s,2}$-norms (defined as above also for negative s) are finite, where now $\langle \psi_{lk}, L\rangle = L(\psi_{lk})$, well defined since $S > s, \psi_{lk} \in H_2^s$. In fact the so-defined norm $\|\cdot\|_{-s,2}$ is equivalent to the standard operator norm on $(H_2^s[0,1])^*$. Moreover every $f \in L^2$ gives rise to a continuous linear form on H_2^s by using the $\langle \cdot, \cdot \rangle$ duality, so we can view L^2 as a subspace of H_2^{-s}. By reflexivity of H_2^s one concludes

$$(H_2^{-s}([0,1]))^* = H_2^s([0,1])$$

up to isomorphism, that is, any linear continuous map $K : H_2^{-s} \to \mathbb{R}$ is of the form $K : L \mapsto L(g)$ for some $g \in H_2^s$, and if L itself is a functional coming from integrating against an L^2-function f_L, then $L(g) = \langle g, f_L \rangle$.

For any $f \in H_2^s \subseteq L^2$ ($s \geq 0$) and dW standard white noise we have a random linear application

$$\mathbb{W} : f \mapsto \int_0^1 f(t)dW(t) \sim N(0, \|f\|_2^2). \tag{8}$$

For any $s > 1/2$, the $\|\mathbb{W}\|_{-s,2}$-norm is finite almost surely since, by Fubini's theorem, for g_{lk} independent $N(0,1)$ variables,

$$E\|\mathbb{W}\|_{-s,2}^2 = \sum_l a_l^{-2s} \sum_k E g_{lk}^2 < \infty,$$

so $\mathbb{W} \in H_2^{-s}$ almost surely, measurable for the cylindrical σ-algebra, and by separability of H_2^{-s} also for the Borel σ-algebra (p.374 in [5]). By Ulam's theorem (Theorem 7.1.4 in [11]), $\mathbb{W}$ is thus tight in H_2^{-s}. The Gaussian variable $\mathbb{W}$ has mean zero and covariance I diagonal for the L^2-inner product,

$$E\mathbb{W}(g)\mathbb{W}(h) = \langle g, h \rangle, \ \forall \ g, h \in H_2^s.$$

We call the law $\mathcal{N}$ of $\mathbb{W}$ a standard, or canonical, Gaussian probability measure on the Hilbert space H_2^{-s} (note that it is the isonormal Gaussian measure for the inner product of L^2 but *not* for the one of H_2^{-s}). In the same way the random trajectory $dX^{(n)}$ from (4) defines a tight Gaussian Borel random variable $\mathbb{X}^{(n)}$ on H_2^{-s} with mean f_0 and covariance $n^{-1}I$.

2. Confidence Sets for Nonparametric Bayes Procedures

2.1. Weighted L^2 Credible Ellipsoids

For $s > 1/2$, denote by

$$B(g, r) = \{f \in H_2^{-s} : \|g - f\|_{-s,2} \leq r\}$$

the norm ball in H_2^{-s} of radius r centered at g. In terms of an orthonormal basis $\{\psi_{lk}\}$ of L^2 from Definition 2 this corresponds to L^2-ellipsoids of the form

$$\left\{ \{c_{lk}\} : \sum_{l,k} a_l^{-2s} |c_{lk} - \langle g, \psi_{lk} \rangle|^2 \leq r^2 \right\}$$

where coefficients in the tail are downweighted by a_l^{-2s}. A frequentist goodness of fit test of a null hypothesis $H_0 : f = f_0$ could for instance be based on the test statistic $\|f_0 - \mathbb{X}^{(n)}\|_{-s,2}$, resembling in nature a Kolmogorov-Smirnov-type procedure (as it has power against arbitrary alternatives f). In directional statistics these are sometimes called Sobolev statistics/tests, see [19, 24].

A Bayesian approach consists in using the quantiles of the posterior directly, that is to solve for $M_n \equiv M(\mathbb{X}^{(n)}, \alpha)$ such that

$$\Pi(f : \|f - \mathbb{X}^{(n)}\|_{-s,2} \leq M_n/\sqrt{n}|\mathbb{X}^{(n)}) = 1 - \alpha \tag{9}$$

where $0 < \alpha < 1$ is some fixed significance level. The resulting credible set

$$C_n = \left\{ f : \|f - \mathbb{X}^{(n)}\|_{-s,2} \leq M_n/\sqrt{n} \right\} \tag{10}$$

is a random $\|\cdot\|_{-s,2}$-ball. The weak Bernstein-von Mises phenomenon in H_2^{-s} implies that this credible ball asymptotically coincides with the exact $(1-\alpha)$-confidence set built using the efficient estimator $\mathbb{X}^{(n)}$ for f_0, in particular the width of the credible ball is of order $O_P(n^{-1/2})$, and M_n converges to a finite constant in probability.

Theorem 1. *Suppose the weak Bernstein-von Mises phenomenon in the sense of Definition 1 holds. Let C_n be the credible region from (10) with M_n chosen as in (9). Then*

$$P_{f_0}^n(f_0 \in C_n) \to 1 - \alpha$$

as $n \to \infty$. If in addition (7) holds then the same result holds true if C_n is centered at the posterior mean $\bar{f}_n$ of $\Pi(\cdot|\mathbb{X}^{(n)})$.

One may wish to intersect C_n further with the support of $\Pi(\cdot|\mathbb{X}^{(n)})$. Theorem 1 holds for such credible sets as well as long as f_0 is in the H_2^{-s}-support of $\Pi(\cdot|\mathbb{X}^{(n)})$ (which in all natural situations will be the case).

2.2. Credible bands for Fourier coefficients

Suppose one wants to recover the Fourier coefficients

$$\hat{f}(m) = \int_0^1 e^{2\pi i m t} f(t) dt, \quad m \in \mathbb{N}, |m| \leq N,$$

up to some fixed frequency $N \in \mathbb{N}$, and is interested in inference on all the $\{\hat{f}(m)\}_{|m|\leq N}$ simultaneously. It is a 'discrete' version of the problem studied in the sampling setting in [12], where several applications can be found. Let

$$\ell_N^\infty = \{f : [-N, N] \cap \mathbb{Z} \to \mathbb{C} : \|f\|_{\infty,N} := \max_{|m|\leq N} |f(m)| < \infty\}$$

be the space of bounded complex-valued functions on the integers $m, |m| \leq N$. The Fourier transform $\mathcal{F} : f \mapsto \{\hat{f}(m)\}_{|m|\leq N}$ maps L^2 into ℓ_N^∞.

Given the empirical Fourier coefficients

$$\phi_n(m) = \int_0^1 e^{2\pi i m t} dX^{(n)}(t), \ m \in \mathbb{Z},$$

whose restriction to the integers $m, |m| \leq N$, is an element of ℓ_N^∞ almost surely, one solves for M_n in

$$\Pi_n \circ \mathcal{F}^{-1}\left(g : \max_{|m|\leq N} |g(m) - \phi_n(m)| \leq M_n/\sqrt{n}\right) = 1 - \alpha \tag{11}$$

where $\Pi_n = \Pi(\cdot|\mathbb{X}^{(n)})$ is the posterior, and constructs a credible band

$$C_n = \left\{g : \max_{|m|\leq N} |g(m) - \phi_n(m)| \leq M_n/\sqrt{n}\right\} \tag{12}$$

of functions $g : [-N, N] \cap \mathbb{Z} \to \mathbb{C}$. Again, the proof of the following theorem establishes that M_n converges in probability to a finite constant.

Theorem 2. *Suppose the weak Bernstein-von Mises phenomenon in the sense of Definition 1 holds. Let C_n be the credible band from (12) with M_n chosen as in (11). Then*

$$P_{f_0}^n\left(\{\widehat{f}(m)\}_{|m|\leq N} \in C_n\right) \to 1 - \alpha$$

as $n \to \infty$. If in addition (7) holds then the same result holds true if C_n is centered at $\mathcal{F}\bar{f}_n$ where $\bar{f}_n$ is the posterior mean of $\Pi(\cdot|X^{(n)})$.

2.3. Credible bands for self-convolutions

Suppose now we are interested in estimating the function

$$f_0 * f_0 = \int_0^1 f_0(\cdot - t) f_0(t) dt$$

where addition is *mod*-1 (so the convolution of f_0 with itself on the unit circle). The related problem in density estimation was studied in the papers [14, 20, 29, 30, 32, 33], where it is shown that $f_0 * f_0$ can be estimated at the $1/\sqrt{n}$-rate even when this is impossible for f_0. See particularly [14] for applications. Assume f_0 is one-periodic and contained in H_2^s for some $s > 1/2$, and that the posterior is supported in $L^2([0,1)) \equiv L^2_{per}([0,1))$ which, in this subsection, denotes the subspace of L^2 consisting of one-periodic functions. We will assume that the basis used to define H_2^s is such that $(\sum_m |\hat{f}(m)|^2(1 + |m|)^{2s})^{1/2}$ is an equivalent norm on H_2^s (which is the case for CDV- or periodised wavelets and trigonometric bases of L^2).

By standard properties of convolutions $\kappa : f \mapsto f * f$ maps $L^2([0,1))$ into $C([0,1))$, the space of bounded continuous periodic functions on $[0,1)$ equipped with the uniform norm $\|\cdot\|_\infty$. If $\Pi_n = \Pi(\cdot|\mathbb{X}^{(n)})$ with posterior mean $\bar{f}_n \in L^2([0,1))$, we can construct a confidence band for $f_0 * f_0$ by solving for M_n such that

$$\Pi_n \circ \kappa^{-1}(g : \|g - \bar{f}_n * \bar{f}_n\|_\infty \leq M_n/\sqrt{n}) = 1 - \alpha \tag{13}$$

with resulting credible band

$$C_n = \{g : \|g - \bar{f}_n * \bar{f}_n\|_\infty \leq M_n/\sqrt{n}\}. \tag{14}$$

Theorem 3. *Suppose the weak Bernstein-von Mises phenomenon in the sense of Definition 1 holds, and that $f_0 \in H_2^s$. Assume moreover (7) and that for some sequence $r_n = o(n^{-1/2})$,*

$$\|\bar{f}_n - f_0\|_2^2 = O_P(r_n), \quad \Pi_n(f : \|f - f_0\|_2^2 > r_n) = o_P(1).$$

Let C_n be the credible band from (14) with M_n chosen as in (13). Then

$$P_{f_0}^n(f_0 * f_0 \in C_n) \to 1 - \alpha$$

as $n \to \infty$.

If f_0 is s-Hölder on $[0, 1)$ for some $s > 1/2$ then the priors from Condition 1 with σ_l and $\gamma = s$ chosen as in Remark 1 are admissible in the above theorem with $r_n = n^{-2s/(2s+1)}$, cf. Corollaries 1, 2 below and the results in Subsection 3.4. Note that any choice $s > 1/2$ will do to obtain order $1/\sqrt{n}$-width credible bands.

2.4. Plug-in credible sets for functionals

If the goal of statistical inference is a possibly nonlinear real-valued functional $\Psi(f_0)$ of f_0 one can use the posterior $\Pi(\cdot \,|\, \mathbb{X}^{(n)})$ in a natural way to construct 'plug-in' credible sets.

2.4.1. Linear functionals

Let L be any linear form on L^2 given by

$$L(f) = \langle f, g_L \rangle = \int_0^1 f(t) g_L(t) dt, \quad f \in L^2,$$

where $g_L \in H_2^s, s > 1/2$, and $g_L \neq 0$. If $\Pi_n = \Pi(\cdot|\mathbb{X}^{(n)})$ is the posterior one may construct credible sets for $L(f_0)$ based on the induced law $\Pi_n^L = \Pi_n \circ L^{-1}$ in several ways: For example one solves for $M_n = M(\mathbb{X}^{(n)}, L, \alpha)$ in

$$\Pi_n^L(z : |z - L(\mathbb{X}^{(n)})| \leq M_n/\sqrt{n}) = 1 - \alpha \tag{15}$$

which gives rise to the credible set

$$C_n = \left\{ z : |z - L(\mathbb{X}^{(n)})| \leq M_n/\sqrt{n} \right\} \tag{16}$$

for $L(f_0)$. An alternative way to build the credible set is discussed below in a more general setting.

Theorem 4. *Suppose the weak Bernstein-von Mises phenomenon in the sense of Definition 1 holds. Let $L = \langle \cdot, g_L \rangle$ be a linear functional on L^2 where $0 \neq$*

$g_L \in H^s_2, s > 1/2$. Let $\beta_{\mathbb{R}}$ be the bounded-Lipschitz metric for weak convergence on $\mathbb{R}$, and define $\theta_t : x \mapsto \sqrt{n}(x-t)$ for $t, x \in \mathbb{R}$. Then

$$\beta_{\mathbb{R}}(\Pi_n^L \circ \theta^{-1}_{L(\mathbb{X}^{(n)})}, N(0, \|g_L\|_2^2)) \to^{P^n_{f_0}} 0.$$

Moreover let C_n be the credible region from (16) with M_n chosen as in (15). Then

$$P^n_{f_0}(L(f_0) \in C_n) \to 1-\alpha$$

as $n \to \infty$. If in addition (7) holds then the same result holds true if C_n is centered at $L(\bar{f}_n)$ where $\bar{f}_n$ is the posterior mean of $\Pi(\cdot|\mathbb{X}^{(n)})$.

The proofs imply that M_n converges to a finite positive constant in probability, and the credible set C_n thus has frequentist length of order $1/\sqrt{n}$. Moreover, the 'plug-in' posterior $\Pi_n \circ L^{-1}$ has the approximate shape of a normal distribution centered at the efficient estimator $L(\mathbb{X}^{(n)})$ of $L(f_0)$ with variance $\|g_L\|_2^2/n$. This implies in particular that the width of the credible set C_n is asymptotically efficient from the semiparametric perspective; in fact $\|g_L\|_2^2$ is the semiparametric Cramér-Rao bound for estimating $L(f_0)$ from observations in the Gaussian white noise model.

The fact that any integral functional $\int_0^1 f(t)g_L(t)dt,\ g_L \in H^s_2, s > 1/2$, is covered gives rise to a rich class of examples. For instance the functionals

$$\int_0^1 t^\alpha f(t)dt, \int_0^1 |t|^\alpha f(t)dt,\ \alpha \in \mathbb{N},$$

are covered, so nonparametric Bayes posteriors can be used with good confidence for inference on moment type functionals. The assumption $s > 1/2$ is intrinsic to our methods and cannot be relaxed.

2.4.2. Smooth nonlinear functionals

We next consider statistical inference for nonlinear functionals of f_0 that satisfy a good quadratic approximation in L^2 at f_0, more precisely, we assume that $\Psi : L^2 \to \mathbb{R}$ satisfies,

$$\Psi(f_0 + h) - \Psi(f_0) = D\Psi_{f_0}[h] + O(\|h\|_2^2), \tag{17}$$

uniformly in $h \in L^2$ and for some $D\Psi_{f_0} : L^2 \to \mathbb{R}$ linear and continuous that has a (nonzero) L^2-Riesz representer $\dot{\Psi}_{f_0} \in H^s_2$ for some $s > 1/2$. This setting includes several standard examples discussed in more detail at the end of this section, but also the linear functionals discussed above.

Note that now Ψ cannot necessarily be evaluated at $\mathbb{X}^{(n)}$ (think of $\Psi(f) = \|f\|_2^2$). However, since the posterior is supported in L^2 with probability one, the following Bayesian credible set can be constructed for $\Psi(f_0)$: For $\Pi_n = \Pi(\cdot|\mathbb{X}^{(n)})$ the posterior distribution, set $\Pi_n^\Psi = \Pi_n \circ \Psi^{-1}$ and solve for reals μ_n, ν_n such that

$$\Pi_n^\Psi((-\infty, \mu_n]) = \Pi_n^\Psi((\nu_n, +\infty)) = \frac{\alpha}{2}, \tag{18}$$

that is, μ_n, ν_n are the $\alpha/2$ and $1-\alpha/2$ quantiles of Π_n^Ψ. Set

$$C_n = C_n(\mathbb{X}^{(n)}, \alpha) = (\mu_n, \nu_n]. \tag{19}$$

Theorem 5. *Suppose the weak Bernstein-von Mises phenomenon in the sense of Definition 1 holds. Consider a functional Ψ such that (17) is satisfied. Assume moreover either that Ψ is linear or that for some sequence $r_n = o(n^{-1/2})$,*

$$\Pi_n(f : \|f - f_0\|_2^2 > r_n) = o_P(1).$$

Let C_n be the credible set from (19) with μ_n, ν_n chosen as in (18). Then, as $n \to \infty$,

$$P_{f_0}^n(\Psi(f_0) \in C_n) \to 1 - \alpha.$$

Similar to the previous result, the shape of the 'plug-in' posterior $\Pi_n \circ \Psi^{-1}$ is approximately Gaussian, this time centered at $\Psi(f_0) + \langle \dot\Psi_{f_0}/\sqrt{n}, \mathbb{W}\rangle$, and with variance $\|\dot\Psi_{f_0}\|_2^2/n$. More precisely, for $\beta_{\mathbb{R}}$ the bounded-Lipschitz metric for weak convergence,

$$\beta_{\mathbb{R}}\left(\Pi_n^\Psi \circ \theta^{-1}_{\Psi(f_0)+\frac{1}{\sqrt{n}}\langle \dot\Psi_{f_0}, \mathbb{W}\rangle}, N(0, \|\dot\Psi_{f_0}\|_2^2)\right) \to^{P_{f_0}^n} 0.$$

In fact, as follows from the proof of Theorem 5, the random quantiles μ_n, ν_n admit the expansion,

$$\mu_n = \Psi(f_0) + \frac{1}{\sqrt{n}}\langle \dot\Psi_{f_0}, \mathbb{W}\rangle + \frac{\Phi_*^{-1}(\frac{\alpha}{2})}{\sqrt{n}} + o_P(1/\sqrt{n})$$
$$\nu_n = \Psi(f_0) + \frac{1}{\sqrt{n}}\langle \dot\Psi_{f_0}, \mathbb{W}\rangle + \frac{\Phi_*^{-1}(1-\frac{\alpha}{2})}{\sqrt{n}} + o_P(1/\sqrt{n}),$$

with Φ_* the distribution function of a $N(0, \|\dot\Psi_{f_0}\|^2)$ variable. Again, $\|\dot\Psi_{f_0}\|_2^2$ is the semiparametric efficiency bound for estimating $\Psi(f_0)$ in the Gaussian white noise model, which shows that the asymptotic width of the credible set C_n for $\Psi(f_0)$ is optimal in the semiparametric sense.

If f_0 is γ-Hölder for some $\gamma > 1/2$ then the priors from Condition 1 with σ_l chosen as in Remark 1 are admissible in the above theorem with $r_n = n^{-2\gamma/(2\gamma+1)}$, cf. Corollaries 1, 2 and Subsection 3.4 below.

Examples include the standard quadratic functionals such as $\Psi(f) = \int f^2(t)dt$ or composite functionals of the form $\Psi(f) = \int \phi(f(x), x)dx$ and the like. Some functionals may necessitate some straightforward modifications of our proofs: For instance $\int |f(t)|^p dt$ requires differentiation on L^p instead of L^2, and for the entropy functional $\int f(t)\log f(t)dt$ one assumes $f_0 \geq \zeta > 0$ on $[0,1]$ and differentiates Ψ on L^∞. In these situations, to control remainder terms, one may use contraction results in $L^p, 2 < p \leq \infty$, instead of L^2, such as the ones in [23]. Our assumption $\gamma > 1/2$ falls short of the critical assumption $\gamma > 1/4$ necessary for $1/\sqrt{n}$-estimability of some of these functionals, a phenomenon intrinsic to general plug-in procedures.

3. Bernstein-von Mises Theorems in White Noise

We now develop general tools that allow to prove that priors satisfy the Bernstein-von Mises phenomenon in the sense of Definition 1, and show how they can be successfully applied to a wide variety of natural classes of product priors.

For $f \in L^2$ consider again observing a random trajectory in the white noise model (4). Given an orthonormal basis from Definition 2, the white noise model is equivalent to observing the action of $\mathbb{X}^{(n)}$ on the basis, i.e.,

$$\mathbb{X}_{lk}^{(n)} = \theta_{lk} + \frac{1}{\sqrt{n}}\varepsilon_{lk}, \quad k \in \mathcal{Z}_l, l \in \mathcal{L},$$

where $\theta_{lk} = \langle f, \psi_{lk} \rangle$, $\varepsilon_{lk} \sim^{i.i.d.} N(0,1)$. Let Π be a prior Borel probability distribution on L^2 which induces a prior, also denoted by Π, on infinite sequences $\{\theta_{lk}\} \in \ell^2$. Let $\Pi(\cdot|\mathbb{X}^{(n)})$ be the posterior distribution and let $\Pi(\theta_{lk} \,|\, \mathbb{X}^{(n)})$ denote the marginal posterior on the coordinate θ_{lk}.

3.1. Contraction Results in H_2^{-s}

In this subsection we 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. Further assume, for some fixed density φ on the real line,

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

Condition 1. *Suppose that there exists a finite constant $M > 0$ s.t.*

$$(\mathbf{P1}) \qquad \sup_{l \in \mathcal{L}, k \in \mathcal{Z}_l} \frac{|\theta_{0,lk}|}{\sigma_l} \le M, \quad \theta_0 = \{\theta_{0,lk}\} = \{\langle f_0, \psi_{lk}\rangle\}$$

Suppose also that φ is s.t. there exists $\tau > M$ and $0 < c_\varphi \le C_\varphi < \infty$ with

$$(\mathbf{P2}) \quad \varphi(x) \le C_\varphi \;\; \forall x \in \mathbb{R}, \; \varphi(x) \ge c_\varphi \;\; \forall x \in (-\tau, \tau), \; \int_{\mathbb{R}} x^2 \varphi(x) dx < \infty.$$

Some discussion of this condition is in order: We allow 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. For Gaussian processes Condition 1 applies simply by verifying that the L^2-basis provided by the Karhunen-Loève expansion of the process satisfies the conditions of Definition 2. This includes in particular Brownian motion: The corresponding φ is then the standard Gaussian density and $\sigma_l = 1/(\pi(l + \frac{1}{2}))$ are the square-roots of the eigenvalues of the operator.

Through condition (**P1**), this allows for signals $f_0 \equiv (\theta_{0,lk})$ whose coefficients on the basis decrease at least as fast as $1/l$. For primitives of Brownian motion similar remarks apply, with stronger but natural decay restrictions on $\langle f_0, \psi_{lk}\rangle$.

In principle making the prior rougher allows for more signals through Condition (**P1**), but this may harm the performance of the posterior in stronger loss functions than the one considered in the next theorem.

Theorem 6. *Consider data generated from equation (4) under a fixed function f_0 with coefficients $\theta_0 = \{\theta_{0,lk}\} = \{\langle f_0, \psi_{lk}\rangle\}$, and denote by $P^n_{f_0}$ the distribution of $\mathbb{X}^{(n)}$. Then if the product prior Π and f_0 satisfy Condition 1 we have for every $s > 1/2$ that, as $n \to \infty$,*

$$P^n_{f_0} \int \|f - f_0\|^2_{-s,2} d\Pi(f \mid \mathbb{X}^{(n)}) = O\left(\frac{1}{n}\right).$$

Proof. We write $\mathbb{X} = \mathbb{X}^{(n)}$ and $E = P^n_{f_0}$ throughout the proof to ease notation. We also decompose the indexing set $\mathcal{L}$ into

$$\mathcal{J}_n := \{l \in \mathcal{L},\ \sqrt{n}\sigma_l \geq S_0\}$$

and its complement, where S_0 is a fixed positive constant. The quantity we wish to bound equals, by definition of the negative Sobolev norm and Fubini's theorem

$$\sum_{l,k} a_l^{-2s} P^n_{f_0} \int (\theta_{lk} - \theta_{0,lk})^2 d\Pi(\theta_{lk} \mid \mathbb{X}).$$

Define further

$$B_{lk}(\mathbb{X}) := \int (\theta_{lk} - \theta_{0,lk})^2 d\Pi(\theta_{lk} \mid \mathbb{X})$$

whose $P^n_{f_0}$-expectation we now bound.

Using the independence structure of the prior we have $\Pi(\theta_{lk} \mid \mathbb{X}) = \Pi(\theta_{lk} \mid \mathbb{X}_{lk})$, and under $P^n_{f_0}$,

$$\begin{aligned} B_{lk}(\mathbb{X}) &= \frac{\int (\theta_{lk} - \theta_{0,lk})^2 e^{-\frac{n}{2}(\theta_{lk}-\theta_{0,lk})^2 + \sqrt{n}\varepsilon_{lk}(\theta_{lk}-\theta_{0,lk})} \varphi_{lk}(\theta_{lk}) d\theta_{lk}}{\int e^{-\frac{n}{2}(\theta_{lk}-\theta_{0,lk})^2 + \sqrt{n}\varepsilon_{lk}(\theta_{lk}-\theta_{0,lk})} \varphi_{lk}(\theta_{lk}) d\theta_{lk}} \\ &= \frac{1}{n} \frac{\int v^2 e^{-\frac{v^2}{2} + \varepsilon_{lk} v} \frac{1}{\sqrt{n}\sigma_l} \varphi\left(\frac{\theta_{0,lk} + n^{-1/2} v}{\sigma_l}\right) dv}{\int e^{-\frac{v^2}{2} + \varepsilon_{lk} v} \frac{1}{\sqrt{n}\sigma_l} \varphi\left(\frac{\theta_{0,lk} + n^{-1/2} v}{\sigma_l}\right) dv} =: \frac{1}{n} \frac{N_{lk}}{D_{lk}}(\varepsilon_{lk}). \end{aligned}$$

Consider first indices in $\mathcal{J}_n^c$. Taking a smaller domain of integration in the denominator makes the integral smaller

$$D_{lk}(\varepsilon_{lk}) \geq \int_{-\sqrt{n}\sigma_l}^{\sqrt{n}\sigma_l} e^{-\frac{v^2}{2} + \varepsilon_{lk} v} \frac{1}{\sqrt{n}\sigma_l} \varphi\left(\frac{\theta_{0,lk} + n^{-1/2} v}{\sigma_l}\right) dv.$$

To simplify the notation we suppose that $\tau > M + 1$. If this is not the case, one multiplies the bounds of the integral in the last display by a small enough constant. The argument of the function φ in the previous display stays in $[-M+1, M+1]$ under **(P1)**. Under assumption **(P2)** this implies that the value of φ in the last expression is bounded from below by c_φ. Next applying Jensen's inequality with the logarithm function

$$\begin{aligned}\log D_{lk}(\varepsilon_{lk}) &\geq \log(2c_\varphi) - \int_{-\sqrt{n}\sigma_l}^{\sqrt{n}\sigma_l} \frac{v^2}{2}\frac{dv}{2\sqrt{n}\sigma_l} + \varepsilon_{lk}\int_{-\sqrt{n}\sigma_l}^{\sqrt{n}\sigma_l} v\frac{dv}{2\sqrt{n}\sigma_l} \\ &= \log(2c_\varphi) - (\sqrt{n}\sigma_l)^2/6.\end{aligned}$$

Thus, $D_{lk}(\varepsilon_{lk}) \geq 2c_\varphi e^{-(\sqrt{n}\sigma_l)^2/6}$, which is bounded away from zero for indices in $\mathcal{J}_n^c$. Now about the numerator, let us split the integral defining N_{lk} into two parts $\{v : \ |v| \leq \sqrt{n}\sigma_l\}$ and $\{v : \ |v| > \sqrt{n}\sigma_l\}$. That is $N_{lk}(\varepsilon_{lk}) = (I) + (II)$. Taking the expectation of the first term and using Fubini's theorem,

$$\begin{aligned}E(I) &= \int_{-\sqrt{n}\sigma_l}^{\sqrt{n}\sigma_l} v^2 e^{-\frac{v^2}{2}} E[e^{\varepsilon_{lk}v}]\frac{1}{\sqrt{n}\sigma_l}\varphi\left(\frac{\theta_{0,lk}+n^{-1/2}v}{\sigma_l}\right)dv \\ &\leq 2n\sigma_l^2 C_\varphi/3.\end{aligned}$$

The expectation of the second term is bounded by first applying Fubini's theorem as before and then changing variables back

$$\begin{aligned}E(II) &= \int_{|v|>\sqrt{n}\sigma_l} v^2 e^{-\frac{v^2}{2}} E[e^{\varepsilon_{lk}v}]\frac{1}{\sqrt{n}\sigma_l}\varphi\left(\frac{\theta_{0,lk}+n^{-1/2}v}{\sigma_l}\right)dv \\ &= \int_{\frac{\theta_{0,lk}}{\sigma_l}-1}^{+\infty}\left(\sqrt{n}\sigma_l u - \sqrt{n}\sigma_l\frac{\theta_{0,lk}}{\sigma_l}\right)^2\varphi(u)du \\ &+ \int_{-\infty}^{\frac{\theta_{0,lk}}{\sigma_l}-1}\left(\sqrt{n}\sigma_l u - \sqrt{n}\sigma_l\frac{\theta_{0,lk}}{\sigma_l}\right)^2\varphi(u)du \\ &\leq 2n\sigma_l^2\left[\frac{\theta_{0,lk}^2}{\sigma_l^2} + \int_{-\infty}^{+\infty} u^2\varphi(u)du\right].\end{aligned}$$

Thus, using **(P1)** again, $E(I) + E(II)$ is bounded on $\mathcal{J}_n^c$ by a fixed constant times $n\sigma_l^2$. In particular, there exists a fixed constant independent of n, k, l such that $E(nB_{lk}(X))$ is bounded from above by a constant on $\mathcal{J}_n^c$.

Now about the indices in $\mathcal{J}_n$. For such l, k, using **(P1)-(P2)** one can find $L_0 > 0$ depending only on S_0, M, τ such that, for any v in $(-L_0, L_0)$,

$$\varphi\left(\frac{\theta_{0,lk}+n^{-1/2}v}{\sigma_l}\right) \geq c_\varphi.$$

Thus the denominator $D_{lk}(\varepsilon_{lk})$ can be bounded from below by

$$D_{lk}(\varepsilon_{lk}) \geq c_\varphi\int_{-L_0}^{L_0} e^{-\frac{v^2}{2}+\varepsilon_{lk}v}\frac{1}{\sqrt{n}\sigma_l}dv.$$

On the other hand, the numerator can be bounded above by

$$N_{lk}(\varepsilon_{lk}) \leq C_\varphi \int v^2 e^{-\frac{v^2}{2}+\varepsilon_{lk}v} \frac{1}{\sqrt{n}\sigma_l} dv,$$

Putting these two bounds together leads to

$$B_{lk}(\varepsilon_{lk}) \leq \frac{C_\varphi}{c_\varphi} \frac{\int v^2 e^{-\frac{v^2}{2}+\varepsilon_{lk}v} dv}{\int_{-L_0}^{L_0} e^{-\frac{v^2}{2}+\varepsilon_{lk}v} dv}.$$

The last quantity has a distribution independent of l, k. Let us thus show that

$$Q(L_0) = E\left[\frac{\int v^2 e^{-\frac{1}{2}(v-\varepsilon)^2} dv}{\int_{-L_0}^{L_0} e^{-\frac{1}{2}(v-\varepsilon)^2} dv}\right].$$

is finite for every $L_0 > 0$, where $\varepsilon \sim N(0,1)$. In the numerator we substitute $u = v - \varepsilon$. Using the inequality $(u+\varepsilon_{lk})^2 \leq 2v^2 + 2\varepsilon_{lk}^2$, the second moment of a standard normal variable appears, and this leads to the bound

$$Q(L_0) \leq CE\left[\frac{1+\varepsilon^2}{\int_{-L_0}^{L_0} e^{-\frac{1}{2}(v-\varepsilon)^2} dv}\right]$$

for some finite constant $C > 0$. Denote by g the density of a standard normal variable, by Φ its distribution function and $\bar{\Phi} = 1 - \Phi$. It is enough to prove that the following quantity is finite

$$q(L_0) := \int_{-\infty}^{+\infty} \frac{(1+u^2)g(u)}{\bar{\Phi}(u-L_0) - \bar{\Phi}(u+L_0)} du = 2\int_0^{+\infty} \frac{(1+u^2)g(u)}{\bar{\Phi}(u-L_0) - \bar{\Phi}(u+L_0)} du,$$

since the integrand is an even function. Using the standard inequalities

$$\frac{1}{\sqrt{2\pi}} \frac{u^2}{1+u^2} \frac{1}{u} e^{-u^2/2} \leq \bar{\Phi}(u) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{u} e^{-u^2/2}, \qquad u \geq 1,$$

it follows that for any $\delta > 0$, one can find $M_\delta > 0$ such that, for any $u \geq M_\delta$, it holds

$$(1-\delta)\frac{1}{u} e^{-u^2/2} \leq \sqrt{2\pi}\bar{\Phi}(u) \leq \frac{1}{u} e^{-u^2/2}, \qquad u \geq M_\delta.$$

Set $A_\delta = 2L_0 \vee M_\delta$. Then for $\delta < 1 - e^{-2L_0}$ we deduce

$$\begin{aligned} q(L_0) &\leq 2\int_0^{A_\delta} \frac{(1+u^2)g(u)}{\bar{\Phi}(A_\delta - L_0) - \bar{\Phi}(A_\delta + L_0)} du \\ &\quad + 2\sqrt{2\pi} \int_{A_\delta}^{+\infty} (u-L_0)(1+u^2) \frac{e^{\frac{1}{2}(u-L_0)^2} g(u)}{1-\delta-e^{-2L_0}} du \\ &\leq C(A_\delta, L_0) + \frac{2e^{-L_0^2/2}}{1-\delta-e^{-2L_0}} \int_{A_\delta}^{+\infty} u(1+u^2)e^{-L_0 u} du < +\infty. \end{aligned}$$

Conclude that $\sup_{l,k} P_{f_0}^n |B_{lk}(\mathbb{X})| = O(1/n)$. Since $\sum_{l,k} a_l^{-2s} < \infty$ the result follows. $\square$

Theorem 7. *With the notation of Theorem 6, suppose the product prior* Π *and* f_0 *satisfy Condition 1. Then*

$$P_{f_0}^n \int \|f - f_0\|_2^2 d\Pi(f \,|\, \mathbb{X}^{(n)}) = O\left(\sum_{l,k} (\sigma_l^2 \wedge n^{-1})\right).$$

Proof. With the notation used in the proof of Theorem 6, using Fubini's Theorem,

$$P_{f_0}^n \int \|f - f_0\|_2^2 d\Pi(f \,|\, \mathbb{X}^{(n)}) = \sum_{l,k} P_{f_0}^n \int (\theta_{lk} - \theta_{0,lk})^2 d\Pi(\theta_{lk} \,|\, \mathbb{X}) = \sum_{l,k} P_{f_0}^n B_{lk}(\mathbb{X}).$$

In the proof of Theorem 6, the following two bounds have been obtained, with the notation $\mathcal{J}_n := \{l \in \mathcal{L},\ \sqrt{n}\sigma_l \geq S_0\}$,

$$\sup_{l \in \mathcal{J}_n,\, k} P_{f_0}^n B_{lk}(\mathbb{X}) = O\left(n^{-1}\right)$$
$$\sup_{l \notin \mathcal{J}_n, 1 \leq k \leq 2^l} P_{f_0}^n B_{lk}(\mathbb{X}) = O\left(\sigma_l^2\right).$$

For any $l \in \mathcal{J}_n^c$, by definition of $\mathcal{J}_n$ it holds $\sigma_l^2 < S_0^2 n^{-1}$, thus $\sigma_l^2 \leq (1 \vee S_0^2)(\sigma_l^2 \wedge n^{-1})$. Similarly, if $l \in \mathcal{J}_n$ we have $n^{-1} \leq (1 \vee S_0^{-2})(\sigma_l^2 \wedge n^{-1})$. ☐

Corollary 1. *Set* $\sigma_l = |l|^{-\frac{1}{2}-\gamma}$ *or* $\sigma_l = 2^{-(\frac{1}{2}+\gamma)l}$ *depending on the chosen S-regular basis of type either a) or b). Suppose that the conditions of Theorem 7 are satisfied. Then*

$$P_{f_0}^n \int \|f - f_0\|_2^2 d\Pi(f \,|\, \mathbb{X}^{(n)}) = O\left(n^{-\frac{2\gamma}{2\gamma+1}}\right).$$

Proof. For both types of basis $\sum_l |\mathcal{Z}_l|(\sigma_l^2 \wedge n^{-1}) = O(n^{-\frac{2\gamma}{2\gamma+1}})$. ☐

Remark 1. The previous choice of σ_l entails a regularity condition of f_0 through Condition **(P1)**, namely $\sup_k |\theta_{0,lk}| \leq M\sigma_l$. If $\sigma_l = 2^{-(\frac{1}{2}+\gamma)l}$ this amounts to the standard Hölderian condition if one uses a CDV wavelet basis, or a periodised wavelet basis – any f_0 in the Besov space $B_{\infty\infty}^{\gamma}$ satisfies **(P1)** for such bases. For other bases similar remarks apply.

Corollary 2. *Denote by* $\bar{f}_n := \bar{f}_n(\mathbb{X}^{(n)}) := \int f d\Pi(f \,|\, \mathbb{X}^{(n)})$ *the posterior mean associated to the posterior distribution. Under the conditions of Theorem 7,*

$$P_{f_0}^n \|\bar{f}_n - f_0\|_2^2 = O\big(\sum_{l,k} (\sigma_l^2 \wedge n^{-1})\big).$$

Proof. The Cauchy-Schwarz inequality implies

$$\begin{aligned} P_{f_0}^n \|\bar{f}_n - f_0\|_2^2 &= P_{f_0}^n \sum_{l,k} \left[\int (\theta_{lk} - \theta_{0,lk}) d\Pi(\theta_{lk} \,|\, \mathbb{X})\right]^2 \\ &\leq P_{f_0}^n \sum_{l,k} \left[\int (\theta_{lk} - \theta_{0,lk})^2 d\Pi(\theta_{lk} \,|\, \mathbb{X})\right] \end{aligned}$$

and one can apply Theorem 7. ☐

3.2. Convergence of Finite Dimensional Distributions

Consider again the posterior distribution $\Pi_n \equiv \Pi(\cdot \mid \mathbb{X}^{(n)})$ on L^2 from the beginning of this section (not necessarily arising from a product measure). Let V be any of the finite-dimensional projection subspaces of L^2 defined in Section 1.2, equipped with the L^2-norm, and recall that π_V denotes the orthogonal projection onto V in L^2. For $z \in H_2^{-s}$ define the transformation

$$T_z \equiv T_{z,V} : f \mapsto \sqrt{n}\, \pi_V(f - z)$$

from H_2^{-s} to V, and consider the image measure $\Pi_n \circ T_z^{-1}$. The finite-dimensional space V carries a natural Lebesgue product measure on it.

Condition 2. *Suppose that $\Pi \circ \pi_V^{-1}$ has a Lebesgue-density $d\Pi_V$ in a neighborhood of $\pi_V(f_0)$ that is continuous and positive at $\pi_V(f_0)$. Suppose also that for every $\delta > 0$ there exists a fixed L^2-norm ball $C = C_\delta$ in V such that, for n large enough,*

$$P_{f_0}^n(\Pi_n \circ T_{f_0}^{-1})(C^c) < \delta.$$

This condition requires that the projected prior has a continuous density at $\pi_V(f_0)$ and that the image of the posterior distribution under the finite-dimensional projection onto V concentrates on a $1/\sqrt{n}$-neighborhood of $\pi_V(f_0)$. It thus ensures the classical finite-dimensional assumptions required for a (local) Bernstein-von Mises theorem in the space V, and the following theorem is proved in a way similar to the classical parametric proof (Chapter 10 in [35]) due to Le Cam [27].

Denote by $\|\cdot\|_{TV}$ the total variation norm on the space of finite signed measures on V, and let $N(0, I)$ be a standard Gaussian measure on V.

Theorem 8. *Consider data generated from equation (4) under a fixed function f_0 and denote by $P_{f_0}^n$ the distribution of $\mathbb{X}^{(n)}$. Assume Condition 2. Then we have, as $n \to \infty$,*

$$\|\Pi_n \circ T_{\mathbb{X}^{(n)}}^{-1} - N(0, I)\|_{TV} \to^{P_{f_0}^n} 0.$$

Proof. If $W_V = \pi_V(\mathbb{W})$, a standard Gaussian variable on V, and if $\tilde{\Pi}_{n,V} = \Pi_n \circ T_{f_0}^{-1}$, it suffices to prove that $\|\tilde{\Pi}_{n,V} - N(W_V, I)\|_{TV}$ converges to zero in $P_{f_0}^n$-probability. In the following, denote by λ the Lebesgue measure on V and by λ_C its restriction to a measurable set C.

Define $\tilde{\Pi}_{n,V}^C$, the posterior distribution $\tilde{\Pi}_{n,V}$ based on the prior restricted to a measurable set C and renormalised, that is, for B a Borel subset of V,

$$\tilde{\Pi}_{n,V}^C(B) = \frac{\int_B e^{-\|h\|^2/2 + \langle h, W_V \rangle} d\tilde{\Pi}_V^C(h)}{\int e^{-\|g\|^2/2 + \langle g, W_V \rangle} d\tilde{\Pi}_V^C(g)}$$

where $\tilde{\Pi}_V = \Pi \circ T_{f_0,V}^{-1}$ and where $\mu^C(B) = \mu(B \cap C)/\mu(C)$ for any probability measure μ. A simple computation shows

$$P_{f_0}^n \|\tilde{\Pi}_{n,V} - \tilde{\Pi}_{n,V}^C\|_{TV} \le 2P_{f_0}^n \tilde{\Pi}_{n,V}(C^c) < 2\delta,$$

using the hypothesis of the theorem, and likewise, if $N^C(W_V, I)$ is the restricted and renormalised normal distribution, $\|N(W_V, I) - N^C(W_V, I)\|_{TV} < \delta$, for every $\delta > 0$ and for $C = C_\delta$ a ball of large enough radius. It thus suffices to prove

$$P_{f_0}^n \|\tilde{\Pi}_{n,V}^C - N^C(W_V, I)\|_{TV} < \delta$$

for every $\delta > 0$ and n large enough. The total variation distance $\|\tilde{\Pi}_{n,V}^C - N^C(W_V, I)\|_{TV}$ is bounded by twice

$$\int \left(1 - \frac{dN^C(W_V, I)(h)}{1_C e^{-\|h\|^2/2+\langle h, W_V\rangle} d\tilde{\Pi}_V(h) / \int_C e^{-\|g\|^2/2+\langle g, W_V\rangle} d\tilde{\Pi}_V(g)}\right)^+ d\tilde{\Pi}_{n,V}^C(h) \leq$$

$$\int \int \left(1 - \frac{e^{-\|g\|^2/2+\langle g, W_V\rangle} d\tilde{\Pi}_V(g) dN^C(W_V, I)(h)}{e^{-\|h\|^2/2+\langle h, W_V\rangle} d\tilde{\Pi}_V(h) dN^C(W_V, I)(g)}\right)^+ dN^C(W_V, I)(g) d\tilde{\Pi}_{n,V}^C(h)$$

$$\leq c \int \int \left(1 - \frac{d\tilde{\Pi}_V(g)}{d\tilde{\Pi}_V(h)}\right)^+ d\lambda_C(g) d\tilde{\Pi}_{n,V}^C(h),$$

where we used $(1 - EY)^+ \leq E(1 - Y)^+$ in the first inequality and where the constant $c \equiv c(W_V)$ in the previous display is an upper bound for the density of $N^C(W_V, I)(g)$ with respect to λ_C. This constant is random but bounded in $P_{f_0}^n$-probability since W_V is tight.

Now note that the last display is random through W_V only. So, considering convergence to zero under $P_{f_0}^n$ amounts to considering convergence to zero under the marginal distribution $P_{f_0,V}^n$ on the subspace V. Under $P_{f_0,V}^n$, the variable W_V has law $N(0, I)$. We have to take the expectation of the display with respect to this law, that we denote by P_{W_V}. That is, dP_{W_V} has Lebesgue-density proportional to $e^{-\|w\|^2/2} d\lambda(w)$ on V.

Define, for $c(V)$ a normalising constant,

$$\begin{aligned} dP_C^n(w) &= c(V) \left(\int e^{-\|k-w\|^2/2} d\tilde{\Pi}_V^C(k)\right) d\lambda(w) \\ &= \left(\int e^{-\|k\|^2/2+\langle k, w\rangle} d\tilde{\Pi}_V^C(k)\right) dP_{W_V}(w), \end{aligned} \tag{20}$$

a probability measure with respect to which dP_{W_V} is contiguous, see Lemma 1 below, so that it suffices to show convergence to zero under dP_C^n instead of dP_{W_V}. The P_C^n-expectation of the quantity in the last but one display equals the expectation of the integrand under

$$d\tilde{\Pi}_{n,V}^C(h) dP_C^n(w) d\lambda_C(g) = e^{-\|h-w\|^2/2} dw d\tilde{\Pi}_V^C(h) d\lambda_C(g),$$

the latter identity following from Fubini's theorem and

$$\int_C e^{-\frac{\|k\|^2}{2}+\langle k, w\rangle} \frac{e^{-\frac{\|h\|^2}{2}+\langle h, w\rangle} d\tilde{\Pi}_V^C(h)}{\int e^{-\frac{\|m\|^2}{2}+\langle m, w\rangle} d\tilde{\Pi}_V^C(m)} d\tilde{\Pi}_V^C(k) e^{-\frac{\|w\|^2}{2}} dw = e^{-\frac{\|h-w\|^2}{2}} dw d\tilde{\Pi}_V^C(h)$$

which is less than or equal to, for n large enough and using that $d\Pi_V$ is continuous at and thus bounded near $\pi_V(f_0)$,

$$c' \int\int\int \left(1 - \frac{d\tilde{\Pi}_V(g)}{d\tilde{\Pi}_V(h)}\right)^+ e^{-\|h-w\|^2/2} dw d\lambda_C(g) d\lambda_C(h)$$
$$= c'' \int\int \left(1 - \frac{d\Pi_V(\pi_V(f_0) + g/\sqrt{n})}{d\Pi_V(\pi_V(f_0) + h/\sqrt{n})}\right)^+ d\lambda_C(g) d\lambda_C(h),$$

which converges to zero by dominated convergence and continuity of $d\Pi_V$ at $\pi_V(f_0)$. ☐

Lemma 1. *The probability measure P_{W_V} is contiguous with respect to the probability measure P_C^n defined by (20).*

Proof. Suppose $P_C^n(A_n) \to 0$, for a sequence of measurable sets A_n. This implies

$$\int_{A_n} \left[\inf_{k \in C} e^{-\|k\|^2/2 + \langle k, w\rangle}\right] dP_{W_V}(w) \to 0.$$

Since C is compact, the infimum of the continuous function in the display is attained for some fixed γ in C. Thus

$$\int_{A_n} e^{-\|\gamma\|^2/2 + \langle \gamma, w\rangle} e^{-\|w\|^2/2} d\lambda(w) = \int_{A_n} e^{-\|\gamma - w\|^2/2} d\lambda(w) \to 0.$$

Since $N(\gamma, I)$ and $N(0, I)$ are mutually contiguous (by, e.g., Le Cam's first lemma, see [35], Chapter 6), the result follows. ☐

Remark 2. Alternatively, in the case of product priors, one can apply Theorem 1 in [8]. By independence of the Gaussian coordinate experiments $\langle \psi_{lk}, \mathbb{X}^{(n)}\rangle \equiv \theta_{0,lk} + \frac{1}{\sqrt{n}}\varepsilon_{lk}$, when estimating one or more generally any finite number of the $\theta_{0,lk}$'s, there is no loss of information with respect to the case where all other $\theta_{0,lk}$'s would be known. Since the model is exactly LAN, condition (N) in [8] is satisfied with a zero remainder and condition (C) in [8] amounts to asking that the full posterior concentrate at some rate $\varepsilon_n \to 0$ in the L^2-norm (which for product priors is implied by Corollary 1).

3.3. A BvM-theorem in H_2^{-s}

Let $\Pi_n = \Pi(\cdot|\mathbb{X}^{(n)})$ be the posterior distribution on L^2. Under the following Condition 3, which depends on a positive real $r > 0$ to be specified in the sequel, we will prove that a weak Bernstein-von Mises phenomenon holds true in H_2^{-s}. For the product priors considered above we will then verify Condition 3 below.

Condition 3. *Suppose for every $\varepsilon > 0$ there exists a constant $0 < M \equiv M(\varepsilon) < \infty$ independent of n such that, for any $n \geq 1$, some $r > 1/2$,*

$$P_{f_0}^n \Pi \left[\left\{ f : \|f - f_0\|_{-r,2}^2 > \frac{M}{n} \right\} |\mathbb{X}^{(n)} \right] \leq \varepsilon. \tag{21}$$

Assume moreover that the conclusion of Theorem 8 holds true for every V (i.e., the finite-dimensional distributions converge).

On H_2^{-s} and for $z \in H_2^{-s}$, define the measurable map

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

Recalling the definitions from Section 1.2, consider $\Pi_n \circ \tau_{\mathbb{X}^{(n)}}^{-1}$, a Borel probability measure on H_2^{-s}. Let $\mathcal{N}$ be a standard Gaussian measure on H_2^{-s}.

Theorem 9. *Fix $s > r > 1/2$ and assume Condition 3 for such r. If β is the bounded-Lipschitz metric for weak convergence of probability measures on H_2^{-s} then, as $n \to \infty$,*

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

Proof. It is enough to show that for every $\varepsilon > 0$ there exists $N = N(\varepsilon)$ large enough such that for all $n \geq N$,

$$P_{f_0}^n \left(\beta(\Pi_n \circ \tau_{\mathbb{X}^{(n)}}^{-1}, \mathcal{N}) > 4\varepsilon \right) < 4\varepsilon,$$

Fix $\varepsilon > 0$ and let V_J be the finite-dimensional subspace of H_2^{-r} spanned by $\{\psi_{lk} : k \in \mathcal{Z}_l, l \in \mathcal{L}, |l| \leq J\}$, for some integer $J \geq 1$. Writing $\tilde{\Pi}_n$ for $\Pi_n \circ \tau_{\mathbb{X}^{(n)}}^{-1}$ we see from the triangle inequality

$$\beta(\tilde{\Pi}_n, \mathcal{N}) \leq \beta(\tilde{\Pi}_n, \tilde{\Pi}_n \circ \pi_{V_J}^{-1}) + \beta(\tilde{\Pi}_n \circ \pi_{V_J}^{-1}, \mathcal{N} \circ \pi_{V_J}^{-1}) + \beta(\mathcal{N} \circ \pi_{V_J}^{-1}, \mathcal{N}).$$

The middle term converges to zero in $P_{f_0}^n$-probability for every V_J, by convergence of the finite-dimensional distributions (Condition 3 and since the total variation distance dominates β). Next we handle the first term. Set $Q = M = M(\varepsilon^2/4)$ and consider the random subset D of H_2^{-s} defined as

$$D = \{g : \|g + \mathbb{W}\|_{-r,2}^2 \leq Q\}.$$

Under $P_{f_0}^n$ we have $\tilde{\Pi}_n(D) = \Pi_n(D_n)$, where $D_n = \{f : \|f - f_0\|_{-r,2}^2 \leq Q/n\}$ is the complement of the set appearing in (21). In particular, using Condition 3 and Markov's inequality yields $P_{f_0}^n(\tilde{\Pi}_n(D^c) > \varepsilon/4) \leq \varepsilon^2/\varepsilon = \varepsilon$.

If $Y_n \sim \tilde{\Pi}_n$ (conditional on $\mathbb{X}^{(n)}$), then $\pi_{V_J}(Y_n) \sim \tilde{\Pi}_n \circ \pi_{V_J}^{-1}$. For F any bounded function on H_2^{-s} of Lipschitz-norm less than one

$$\left| \int_{H_2^{-s}} F d\tilde{\Pi}_n - \int_{H_2^{-s}} F d(\tilde{\Pi}_n \circ \pi_{V_J}^{-1}) \right| = \left| E_{\tilde{\Pi}_n} \left[F(Y_n) - F(\pi_{V_J}(Y_n)) \right] \right|$$
$$\leq E_{\tilde{\Pi}_n} \left[\|Y_n - \pi_{V_J}(Y_n)\|_{-s,2} 1_D(Y_n) \right] + 2\tilde{\Pi}_n(D^c),$$

where $E_{\tilde{\Pi}_n}$ denotes expectation under $\tilde{\Pi}_n$ (given $\mathbb{X}^{(n)}$). With $y_{lk} = \langle Y_n, \psi_{lk}\rangle$,

$$\begin{aligned} E_{\tilde{\Pi}_n}\left[\|Y_n - \pi_{V_J}(Y_n)\|^2_{-s,2} 1_D(Y_n)\right] &= E_{\tilde{\Pi}_n}\left[\sum_{l>J} a_l^{-2s} \sum_k |y_{lk}|^2 1_D(Y_n)\right] \\ &= E_{\tilde{\Pi}_n}\left[\sum_{l>J} a_l^{2(r-s)} a_l^{-2r} \sum_k |y_{lk}|^2 1_D(Y_n)\right] \\ &\le a_J^{2(r-s)} E_{\tilde{\Pi}_n}\left[\|Y_n\|^2_{-r,2} 1_D(Y_n)\right] \le 2a_J^{2(r-s)}\left[Q + \|\mathbb{W}\|^2_{-r,2}\right]. \end{aligned}$$

From the definition of β one deduces that for large enough J,

$$\beta(\tilde{\Pi}_n, \tilde{\Pi}_n \circ \pi_{V_J}^{-1}) \le \varepsilon + 2\tilde{\Pi}_n(D^c) + \sqrt{2} a_J^{r-s} \|\mathbb{W}\|_{-r,2}.$$

Conclude that $P^n_{f_0}(\beta(\tilde{\Pi}_n, \tilde{\Pi}_n \circ \pi_{V_J}^{-1}) > 2\varepsilon) < 2\varepsilon$ for J large enough, combining the previous deviation bound for $\tilde{\Pi}_n(D^c)$ and that $\|\mathbb{W}\|_{-r,2}$ is bounded in probability, which follows from after (8) above. A similar (though simpler) argument leads to

$$P^n_{f_0}(\beta(\mathcal{N} \circ \pi_{V_J}^{-1}, \mathcal{N}) > \varepsilon) < \varepsilon,$$

using again that any random variable with law $\mathcal{N}$ has square integrable Hilbert-norm on H_2^{-r}. This concludes the proof. □

3.4. The BvM-theorem for Product Priors

Combining the contraction result Theorem 6 with Theorems 8 and 9 we now provide a rich class of nonparametric product priors for which the weak Bernstein – von Mises theorem in the sense of Definition 1 holds. Specific choices of σ_l allow in particular to obtain Bickel and Ritov's plug-in property for the posterior, relevant in Theorems 3 and 5.

Theorem 10. *Suppose the assumptions of Theorem 6 are satisfied and that φ is continuous in a neighborhood of $\{\theta_{0,lk}\}$ for every $k \in \mathcal{Z}_l, l \in \mathcal{L}$. Let $s > 1/2$. Then for β the bounded Lipschitz metric for weak convergence of probability measures on H_2^{-s} we have, as $n \to \infty$,*

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

Proof. We only need to verify Condition 3 with some $1/2 < r < s$ so that we can apply Theorem 9. From Theorem 6 with any such r in place of s, we see that

$$nP^n_{f_0} \int \|f - f_0\|^2_{-r,2} d\Pi(f|\mathbb{X}^{(n)}) = O(1), \tag{22}$$

which verifies the first part of Condition 3 for some M large enough using Markov's inequality. The second part follows from verifying Condition 2 to invoke Theorem 8: Let V be arbitrary. If V_J is defined as in the proof of Theorem 9, and if J is the smallest integer such that $V \subset V_J$, then

$$\|\pi_V(f - f_0)\|^2_2 \le \|\pi_{V_J}(f - f_0)\|^2_2 \le a_J^{2r} \|f - f_0\|^2_{-r,2}$$

so that the second part of Condition 2 follows from the estimate (22) and Markov's inequality, for C a fixed norm ball of squared diameter of order $a_J^{2r}M^2$. The first part of Condition 2 follows from the fact that $\Pi \circ T_{f_0}^{-1}$ is a product measure in V with bounded marginals φ_{lk} constant in k, and from the continuity assumption on φ. □

Theorem 11. *Suppose the assumptions of Theorem 6 are satisfied and that φ is continuous in a neighborhood of $\{\theta_{0,lk}\}$ for every $k \in \mathcal{Z}_l, l \in \mathcal{L}$. Let $s > 1/2$, let Y_n be a random variable drawn from $\Pi_n \circ \tau_{\mathbb{X}^{(n)}}^{-1}$ (conditional on $\mathbb{X}^{(n)}$), and let $\bar{f}_n$ be the (Bochner-) mean of the posterior distribution $\Pi(\cdot|\mathbb{X}^{(n)})$. Then*

$$E[Y_n \,|\, \mathbb{X}^{(n)}] = \sqrt{n}(\bar{f}_n - \mathbb{X}^{(n)}) \to^{P_{f_0}^n} 0$$

in H_2^{-s} as $n \to \infty$.

Proof. Note that

$$\begin{aligned} E[\|Y_n\|_{-s,2}^2 \,|\, \mathbb{X}^{(n)}] &= \int \|h\|_{-s,2}^2 d\Pi_n \circ \tau_{\mathbb{X}^{(n)}}^{-1}(h) \\ &\leq 2n \int \|f - f_0\|_{-s,2}^2 d\Pi(f|\mathbb{X}^{(n)}) + 2\|\mathbb{W}\|_{-s,2}^2 = O_{P_{f_0}^n}(1) \end{aligned}$$

by Theorem 6 and since $\|\mathbb{W}\|_{-s,2} < \infty$ almost surely, as after (8). Moreover $Y_n \to N$ weakly in H_2^{-s} in $P_{f_0}^n$-probability, where $N \sim \mathcal{N}$, by Theorem 10. By a standard uniform integrability argument (using that $\{Y_n : n \in \mathbb{N}\}$ has H_2^{-s}-norms with uniformly bounded second moments and converges to N weakly), and arguing as in the last paragraph of Section 4.2 below, we conclude

$$E[Y_n \,|\, \mathbb{X}^{(n)}] \to EN$$

in H_2^{-s} in $P_{f_0}^n$-probability, which implies the result since $EN = 0$. □

4. Remaining Proofs

4.1. *Proofs for Section 2*

Theorem 1. By Corollary 6.8.5 in [5] the image measure $\mathcal{N} \circ (\|\cdot\|_{-s,2})^{-1}$ of $\mathcal{N}$ under the norm mapping is absolutely continuous on $[0, \infty)$, so the mapping

$$\Phi : t \mapsto \mathcal{N}(B(0,t)) = \mathcal{N}(\overline{B(0,t)}) = \mathcal{N} \circ (\|\cdot\|_{-s,2})^{-1}([0,t])$$

is uniformly continuous and increasing on $[0, \infty)$. In fact, the mapping is strictly increasing on $[0, \infty)$: using the results on p.213-214 in [37], is suffices to show that any shell $\{f : s < \|f\|_{-s,2} < t\}, s < t$, contains an element of the RKHS L^2 of $\mathcal{N}$, which is obvious as L^2 is dense in H_2^{-s}. Thus Φ has a continuous inverse $\Phi^{-1} : [0, 1) \to [0, \infty)$. Since Φ is uniformly continuous for every $\epsilon > 0$ there exists $\delta > 0$ small enough such that $|\Phi(t+\delta) - \Phi(t)| < \epsilon$ for every $t \in [0,\infty)$. Now

$$\mathcal{N}(\partial_\delta B(0,t)) = \mathcal{N}(B(0,t+\delta)) - \mathcal{N}(B(0,t-\delta)) = |\Phi(t+\delta) - \Phi(t-\delta)| < 2\epsilon$$

for $\delta > 0$ small enough, independently of t. Using (25) below we deduce that the balls $\{B(0,t)\}_{0\le t<\infty}$ form a $\mathcal{N}$-uniformity class, and we can thus conclude from Definition 1 and the results in Section 4.2 below that

$$\sup_{0\le t<\infty} \left| \Pi(f : \|f - \mathbb{X}^{(n)}\|_{-s,2} \le t/\sqrt{n} | \mathbb{X}^{(n)}) - \mathcal{N}(B(0,t)) \right| \to 0$$

in $P^n_{f_0}$-probability, as $n \to \infty$. This combined with (9) gives

$$\mathcal{N}(B(0,M_n)) = \mathcal{N}(B(0,M_n)) - \Pi(f : \|f - \mathbb{X}^{(n)}\|_{-s,2} \le M_n/\sqrt{n} | \mathbb{X}^{(n)}) + 1 - \alpha,$$

which converges to $1-\alpha$ as $n \to \infty$ in $P^n_{f_0}$-probability, and thus, by the continuous mapping theorem,

$$M_n \to^{P^n_{f_0}} \Phi^{-1}(1-\alpha) \tag{23}$$

as $n \to \infty$. Now using this last convergence in probability,

$$\begin{aligned} P^n_{f_0}(f_0 \in C_n) &= P^n_{f_0}(f_0 \in B(\mathbb{X}^{(n)}, M_n/\sqrt{n})) \\ &= P^n_{f_0}(0 \in B(\mathbb{W}, M_n)) \\ &= P^n_{f_0}(0 \in B(\mathbb{W}, \Phi^{-1}(1-\alpha))) + o(1) \\ &= \mathcal{N}(B(0, \Phi^{-1}(1-\alpha)) + o(1) \\ &= \Phi(\Phi^{-1}(1-\alpha)) + o(1) = 1 - \alpha + o(1) \end{aligned}$$

which completes the proof of the first claim. The second claim follows from the same arguments combined with (7) which implies that

$$P^n_{f_0}(f_0 \in B(\bar{f}_n, M_n/\sqrt{n})) - P^n_{f_0}(f_0 \in B(\mathbb{X}^{(n)}, M_n/\sqrt{n})) \to 0$$

in $P^n_{f_0}$-probability, as $n \to \infty$. □

Theorem 2. The mapping $\mathcal{F} : f \mapsto \hat{f} = \langle f, e^{2\pi i m \cdot}\rangle_{|m|\le N, m\in\mathbb{Z}}$ is linear and continuous from $H_2^{-s} \to \ell^\infty_N$ in view of $\|\hat{f}\|_{\infty,N} \le C\|f\|_{-s,2}$ for $C = \max_m \|e^{2\pi i m\cdot}\|_{s,2} < \infty$. Therefore, by Definition 1 and the continuous mapping theorem we have (as in Section 4.2 below), as $n \to \infty$,

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

where β is the bounded Lipschitz metric for weak convergence in ℓ^∞_N. If we define $\hat{W}(u) = \int_0^1 e^{2\pi i u t} dW(t), u \in \mathbb{Z}$, then the $\{\hat{W}(u)\}_{|u|\le N}$ are i.i.d. $N(0,1)$ so that $\hat{W} \in \ell^\infty_N$ almost surely. Moreover $\max_{|u|\le N} |\hat{W}(u)|$, as a finite maximum of i.i.d. Gaussians, has an absolutely continuous distribution, in fact it is not difficult to prove that the mapping

$$\Phi_N : t \mapsto \Pr(\|\hat{W}\|_{\infty,N} \le t) = \Pr\left(\max_{|u|\le N} |\hat{W}(u)| \le t\right)$$

is uniformly continuous and strictly increasing on $[0,\infty)$ with continuous inverse Φ_N^{-1}. Conclude that, as in the proof of Theorem 1, the norm balls $\{h \in \ell_N^\infty : \|h\|_{\infty,N} \le t\}_{t\ge 0}$ form a $\mathcal{N}\circ\mathcal{F}^{-1}$-uniformity class of subsets of ℓ_N^∞. Since, by linearity, $(\Pi_n \circ \tau_{\mathbb{X}^{(n)}}^{-1})\circ\mathcal{F}^{-1} = (\Pi_n\circ\mathcal{F}^{-1})\circ\theta_{\phi_n}^{-1}$ where $\theta_{\phi_n} : \ell_N^\infty \to \ell_N^\infty$ is given by $g \mapsto \sqrt{n}(g-\phi_n)$, we obtain

$$\mathcal{N}\circ\mathcal{F}^{-1}\left(h : \|h\|_{\infty,N} \le M_n\right) - (\Pi_n\circ\mathcal{F}^{-1})\circ\theta_{\phi_n}^{-1}\left(h : \|h\|_{\infty,N}\le M_n\right) \to^{P_{f_0}^n} 0$$

and thus $M_n \to^{P_{f_0}^n} \Phi_N^{-1}(1-\alpha)$ as $n\to\infty$. Now

$$\begin{aligned}
P_{f_0}^n(\mathcal{F}f_0 \in C_n) &= P_{f_0}^n\left(\|\widehat{f}_0 - \phi_n\|_{\infty,N} \le M_n/\sqrt{n}\right)\\
&= P_{f_0}^n\left(\|\hat{W}\|_{\infty,N} \le M_n\right)\\
&= P_{f_0}^n\left(\|\hat{W}\|_{\infty,N} \le \Phi_N^{-1}(1-\alpha)\right) + o(1)\\
&= \mathcal{N}\circ\mathcal{F}^{-1}\left(h : \|h\|_{\infty,N} \le \Phi_N^{-1}(1-\alpha)\right) + o(1)\\
&= \Phi_N(\Phi_N^{-1}(1-\alpha)) + o(1) = 1-\alpha+o(1)
\end{aligned}$$

which completes the proof of the first part. The second follows as in the proof of Theorem 1, using (7) and the continuous mapping theorem for $\mathcal{F}(\cdot)$. ☐

Theorem 3. Since $f_0 \in L^1 \cap H_2^s$ we see by Fourier inversion on the circle, the Cauchy-Schwarz inequality and our assumption on the equivalent Sobolev norm that

$$\begin{aligned}
\|f * f_0\|_\infty &\le \sum_m |\hat{f}(m)|(1+|m|)^{-s}(1+|m|)^s|\widehat{f}_0(m)|\\
&\le \left(\sum_m |\hat{f}(m)|^2(1+|m|)^{-2s}\right)^{1/2}\left(\sum_m |\hat{f}_0(m)|^2(1+|m|)^{2s}\right)^{1/2}\\
&\le C'\|f\|_{-s,2},
\end{aligned}$$

in particular $f * f_0$, for $f \in H_2^{-s}, f_0 \in H_2^s$, defines a continuous function on $[0,1)$ (by Fourier inversion), and the mapping $\lambda : f \mapsto 2f * f_0$ is linear and continuous from H_2^{-s} to $C([0,1))$ (this argument is taken from Theorem 1 in [30]). By Definition 1 and the continuous mapping theorem we thus have

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

as $n\to\infty$, where β is the bounded Lipschitz metric for weak convergence in $C([0,1))$. Moreover from Corollary 6.8.5 in [5] we deduce as in the proof of Theorem 1 that norm balls $\{f : \|f\|_\infty \le t\}_{0\le t<\infty}$ are $\mathcal{N}\circ\lambda^{-1}$ uniformity classes for weak convergence, and that the mapping

$$\Phi_\lambda : t \mapsto \mathcal{N}\circ\lambda^{-1}(f : \|f\|_\infty \le t)$$

from $[0,\infty)$ to $[0,1)$ is continuous and increasing. In fact, it is strictly increasing, using the results on p.213-214 in [37] combined with the fact that the RKHS of

$\mathbb{W} * f_0$, equal to $L^2 * f_0$, contains functions of arbitrary supremum norm. Denote by Φ_λ^{-1} the continuous inverse of Φ_λ. Conclude, as in the previous proofs, that

$$\mathcal{N} \circ \lambda^{-1}(f : \|f\|_\infty \le M_n) - (\Pi_n \circ \lambda^{-1}) \circ \theta^{-1}_{\mathbb{X}^{(n)} * f_0}(f : \|f\|_\infty \le M_n) \to^{P^n_{f_0}} 0$$

as $n \to \infty$, where now $\theta_{\mathbb{X}^{(n)} * f_0} : g \mapsto \sqrt{n}(g - \mathbb{X}^{(n)} * f_0)$ maps $C([0,1)) \to C([0,1))$.

Thus, using the hypotheses on $\bar{f}_n$ and the posterior contraction rate, the decomposition

$$f * f - g * g = 2(f-g) * g + (f-g) * (f-g)$$

and the convolution inequality $\|h * h'\|_\infty \le \|h\|_2 \|h'\|_2$ we see,

$$\begin{aligned} 1 - \alpha &= \Pi_n \circ \kappa^{-1}(g : \|g - \bar{f}_n * \bar{f}_n\|_\infty \le M_n/\sqrt{n}) \\ &= \Pi_n(f : \|f * f - \bar{f}_n * \bar{f}_n\|_\infty \le M_n/\sqrt{n}) \\ &\le \Pi_n(f : 2\|(f - \mathbb{X}^{(n)}) * f_0\|_\infty \le M_n/\sqrt{n} + r_n) + o_P(1) \\ &\le \Pi_n(f : 2\sqrt{n}\|(f - \mathbb{X}^{(n)}) * f_0\|_\infty \le M_n + \delta_n) + o_P(1), \end{aligned}$$

with $\delta_n = r_n\sqrt{n} = o(1)$ as $n \to \infty$ by assumption. Using the weak convergence property established above,

$$1 - \alpha \le \Phi_\lambda(M_n + \delta_n) + o_P(1).$$

Similarly, one obtains the inequality in the other direction

$$1 - \alpha \ge \Phi_\lambda(M_n - \delta_n) + o_P(1).$$

From this we conclude $M_n \to^{P^n_{f_0}} \Phi_\lambda^{-1}(1-\alpha)$ as $n \to \infty$. Now as above

$$\begin{aligned} P^n_{f_0}(f_0 * f_0 \in C_n) &= P^n_{f_0}(\|f_0 * f_0 - \bar{f}_n * \bar{f}_n\|_\infty \le M_n/\sqrt{n}) \\ &= P^n_{f_0}(2\|(\bar{f}_n - f_0) * f_0\|_\infty \le M_n/\sqrt{n}) + o(1) \\ &= P^n_{f_0}(2\sqrt{n}\|(\mathbb{X}^{(n)} - f_0) * f_0\|_\infty \le \Phi_\lambda^{-1}(1-\alpha)) + o(1) \\ &= P^n_{f_0}(2\|\mathbb{W} * f_0\|_\infty \le \Phi_\lambda^{-1}(1-\alpha)) + o(1) \\ &= \Phi_\lambda(\Phi_\lambda^{-1}(1-\alpha)) + o(1) = 1 - \alpha + o(1) \end{aligned}$$

completing the proof. □

Theorem 4. The proof is similar to, in fact simpler than, the previous ones, using the continuous mapping theorem to deduce

$$\beta_{\mathbb{R}}((\Pi_n \circ \tau^{-1}_{\mathbb{X}^{(n)}}) \circ L^{-1}, \mathcal{N} \circ L^{-1}) \to^{P_{f_0}} 0$$

as $n \to \infty$, and that the intervals $\{[-t,t]\}_{t \ge 0}$ are $\mathcal{N} \circ L^{-1}$-uniformity classes for weak convergence. We leave the details to the reader. □

Theorem 5. The following notation is used in the proof

$$\theta_n^* = \Psi(f_0) + \langle \dot{\Psi}_{f_0}, \frac{\mathbb{W}}{\sqrt{n}} \rangle \qquad \text{and} \qquad \Phi_*(\cdot) = N(0, \|\dot{\Psi}_{f_0}\|_2^2)((-\infty, \cdot]).$$

By definition of the quantile μ_n it holds

$$\begin{aligned} \frac{\alpha}{2} &= \Pi_n \circ \Psi^{-1}((-\infty, \mu_n]) = \Pi_n(\Psi(f) \leq \mu_n) \\ &= \Pi_n \left(\Psi(f) - \Psi(f_0) \leq \mu_n - \Psi(f_0)\right) \\ &= \Pi_n \left(\langle \dot{\Psi}_{f_0}, f - \mathbb{X}^{(n)} \rangle \leq \mu_n - \theta_n^* - \left[\Psi(f) - \Psi(f_0) - \langle \dot{\Psi}_{f_0}, f - f_0 \rangle \right] \right). \end{aligned}$$

The assumed contraction of the posterior in a L^2-neighborhood of f_0 at rate r_n together with (17) and the fact that $\sqrt{n} r_n = o(1)$ imply the existence of $\delta_n \to 0$ such that

$$\begin{aligned} \frac{\alpha}{2} &\leq \Pi_n(\sqrt{n}\langle \dot{\Psi}_{f_0}, f - \mathbb{X}^{(n)} \rangle \leq \sqrt{n}(\mu_n - \theta_n^*) + \delta_n) + o_P(1) \\ \frac{\alpha}{2} &\geq \Pi_n(\sqrt{n}\langle \dot{\Psi}_{f_0}, f - \mathbb{X}^{(n)} \rangle \leq \sqrt{n}(\mu_n - \theta_n^*) - \delta_n) + o_P(1). \end{aligned}$$

Using the continuous mapping theorem and Definition 1,

$$\beta_{\mathbb{R}}(\Pi_n \circ \tau_{\mathbb{X}^{(n)}}^{-1} \circ (D\Psi_{f_0})^{-1}, \mathcal{N} \circ (D\Psi_{f_0})^{-1}) \to^{P_{f_0}^n} 0$$

as $n \to \infty$. Note that $\mathcal{N} \circ (D\Psi_{f_0})^{-1}$ has distribution function Φ_*. Since the sets $\{(-\infty, t],\ t \in \mathbb{R}\}$ form a uniformity class for weak convergence towards a normal distribution, we obtain

$$\begin{aligned} \frac{\alpha}{2} &\leq \Phi_*(\sqrt{n}(\mu_n - \theta_n^*) + \delta_n) + o_P(1) \\ \frac{\alpha}{2} &\geq \Phi_*(\sqrt{n}(\mu_n - \theta_n^*) - \delta_n) + o_P(1). \end{aligned}$$

From this we deduce the following expansion for μ_n^*, as $n \to \infty$,

$$\mu_n = \theta_n^* + \frac{1}{\sqrt{n}} \Phi_*^{-1}(\frac{\alpha}{2}) + o_P(1/\sqrt{n}).$$

The quantile ν_n expands similarly, with $\Phi_*^{-1}(\frac{\alpha}{2})$ replaced by $\Phi_*^{-1}(1 - \frac{\alpha}{2})$. Now by definition of θ_n^*,

$$\begin{aligned} &P_{f_0}^n(\Psi(f_0) \in (\mu_n, \nu_n]) = \\ &P_{f_0}^n \left(\langle \dot{\Psi}_{f_0}, \frac{\mathbb{W}}{\sqrt{n}} \rangle \in \left[\frac{\Phi_*^{-1}(\alpha/2)}{\sqrt{n}} + o_P\left(\frac{1}{\sqrt{n}}\right), \frac{\Phi_*^{-1}(1-\alpha/2)}{\sqrt{n}} + o_P\left(\frac{1}{\sqrt{n}}\right) \right] \right) \\ &= P_{f_0}^n \left(\langle \dot{\Psi}_{f_0}, \mathbb{W} \rangle \in [\Phi_*^{-1}(\alpha/2), \Phi_*^{-1}(1-\alpha/2)] \right) + o(1) = 1 - \alpha + o(1), \end{aligned}$$

completing the proof. □

4.2. Some Weak Convergence Facts

Let μ, ν be Borel probability measures on a separable metric space (S, d). We call a family $\mathcal{U}$ of measurable real-valued functions defined on S a $\mu-$uniformity class for weak convergence if for any sequence μ_n of Borel probability measures on S that converges weakly to μ we also have

$$\sup_{u \in \mathcal{U}} \left| \int_S u(s)(d\mu_n - d\mu)(s) \right| \to 0 \tag{24}$$

as $n \to \infty$. Necessary and sufficient conditions for classes $\mathcal{U}$ of functions or sets $\{1_A : A \in \mathcal{A}\}$ to form uniformity classes are given in Billingsley and Topsøe [4]. For any subset A of S, define $A^\delta = \{x \in S : d(x, A) < \delta\}$ and the δ-boundary of A by $\partial_\delta A = \{x \in S : d(x, A) < \delta, d(x, A^c) < \delta\}$. A family $\mathcal{A}$ of measurable subsets of S is a μ-uniformity class if and only if

$$\lim_{\delta \to 0} \sup_{A \in \mathcal{A}} \mu(\partial_\delta A) = 0, \tag{25}$$

see Theorem 2 in [4]. For classes of functions a similar characterisation is available using moduli of continuity of the involved functions, see Theorem 1 in [4]. In particular the bounded Lipschitz metric

$$\beta(\mu, \nu) = \sup_{u \in BL(1)} \left| \int_S u(s)(d\mu - d\nu)(s) \right|$$

tests against the class

$$BL(1) = \left\{ f : S \to \mathbb{R}, \ \sup_{s \in S} |f(s)| + \sup_{s \neq t, s,t \in S} \frac{|f(s) - f(t)|}{d(s,t)} \leq 1 \right\},$$

which is a uniformity class for any probability measure μ. It is well known that β metrises weak convergence of probability measures on any separable metric space (e.g., [11], Theorem 11.3.3).

We conclude with the following observation, which was used repeatedly in our proofs: Let $\mathcal{P}(S)$ denote the space of Borel probability measures on S, let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let $\mu_n : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathcal{P}(S), n \in \mathbb{N}$, be random probability measures on S, and let μ be a fixed probability measure on S. If $\beta(\mu_n, \mu) \to^{\mathbb{P}} 0$ as $n \to \infty$, and if $\mathcal{U}$ is a μ-uniformity class for μ, then the convergence in (24) holds true in $\mathbb{P}$-probability, as is easily proved by contradiction and passing to almost surely convergent subsequences. Likewise, if (T, d') is a metric space and $F : S \to T$ is a continuous mapping, then $\beta(\mu_n \circ F^{-1}, \mu \circ F^{-1}) \to^P 0$.

References

[1] A. Araujo and E. Giné. *The central limit theorem for real and Banach valued random variables.* John Wiley & Sons, New York-Chichester-Brisbane, 1980.

[2] P. J. Bickel and B. J. K. Kleijn. The semiparametric Bernstein-von Mises theorem. *Ann. Statist.*, 40:206–237, 2012.
[3] P. J. Bickel and Y. Ritov. Nonparametric estimators which can be "plugged-in". *Ann. Statist.*, 31(4):1033–1053, 2003.
[4] P. Billingsley and F. Topsøe. Uniformity in weak convergence. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 7:1–16, 1967.
[5] V. I. Bogachev. *Gaussian measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1998.
[6] D. Bontemps. Bernstein-von mises theorems for gaussian regression with increasing number of regressors. *Ann. Statist.*, 39:2557–2584, 2011.
[7] S. Boucheron and E. Gassiat. A Bernstein-von Mises theorem for discrete probability distributions. *Electron. J. Stat.*, 3:114–148, 2009.
[8] I. Castillo. A semiparametric Bernstein-von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields*, 152:53–99, 2012.
[9] B. Clarke and S. Ghosal. Reference priors for exponential families with increasing dimension. *Electron. J. Stat.*, 4:737–780, 2010.
[10] A. Cohen, I. Daubechies, and P. Vial. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, 1(1):54–81, 1993.
[11] R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.
[12] A. Feuerverger and R. A. Mureika. The empirical characteristic function and its applications. *Ann. Statist.*, 5(1):88–97, 1977.
[13] G. B. Folland. *Real analysis*. Pure and Applied Mathematics (New York). John Wiley & Sons Inc., New York, second edition, 1999. Modern techniques and their applications, A Wiley-Interscience Publication.
[14] E. W. Frees. Estimating densities of functions of observations. *J. Amer. Statist. Assoc.*, 89(426):517–525, 1994.
[15] S. Ghosal. Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, 5(2):315–331, 1999.
[16] S. Ghosal. Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.*, 74(1):49–68, 2000.
[17] S. Ghosal, J. K. Ghosh, and A.W. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.
[18] S. Ghosal and A.W. van der Vaart. Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.*, 35(1):192–223, 2007.
[19] E. Giné. Invariant tests for uniformity on compact Riemannian manifolds based on Sobolev norms. *Ann. Statist.*, 3(6):1243–1266, 1975.
[20] E. Giné and D. M. Mason. On local U-statistic processes and the estimation of densities of functions of several sample variables. *Ann. Statist.*, 35(3):1105–1145, 2007.
[21] E. Giné and R. Nickl. Uniform central limit theorems for kernel density estimators. *Probab. Theory Related Fields*, 141(3-4):333–387, 2008.
[22] E. Giné and R. Nickl. Uniform limit theorems for wavelet density estimators. *Ann. Probab.*, 37(4):1605–1646, 2009.

[23] E. Giné and R. Nickl. Rates of contraction for posterior distributions in l^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.*, 39:2883–2911, 2011.

[24] P. E. Jupp. Data-driven Sobolev tests of uniformity on compact Riemannian manifolds. *Ann. Statist.*, 36(3):1246–1260, 2008.

[25] J. Kiefer and J. Wolfowitz. Asymptotically minimax estimation of concave and convex distribution functions. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 34(1):73–85, 1976.

[26] P.S. Laplace. Mémoire sur les formules qui sont fonctions de très grands nombres et sur leurs applications aux probabilités. *Oeuvres de Laplace*, 12:301–345, 357–412, 1810.

[27] L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York, 1986.

[28] H. Leahu. On the Bernstein-von Mises phenomenon in the Gaussian white noise model. *Electron. J. Stat.*, 5:373–404, 2011.

[29] R. Nickl. Donsker-type theorems for nonparametric maximum likelihood estimators. *Probab. Theory Related Fields*, 138(3-4):411–449, 2007. correction ibid., 141, 331-332.

[30] R. Nickl. On convergence and convolutions of random signed measures. *J. Theoret. Probab.*, 22(1):38–56, 2009.

[31] V. Rivoirard and J. Rousseau. Bernstein-von Mises theorems for linear functionals of the density. *preprint*, 2011.

[32] A. Schick and W. Wefelmeyer. Root-n consistent density estimators for sums of independent random variables. *J. Nonparametr. Stat.*, 16(6):925–935, 2004.

[33] A. Schick and W. Wefelmeyer. Root-n consistent density estimators of convolutions in weighted L_1-norms. *J. Statist. Plann. Inference*, 137(6):1765–1774, 2007.

[34] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714, 2001.

[35] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[36] A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463, 2008.

[37] A. W. van der Vaart and J. H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. Collect.*, pages 200–222. Inst. Math. Statist., Beachwood, OH, 2008.

[38] A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675, 2009.

[39] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

[40] R. von Mises. *Wahrscheinlichkeitsrechnung*. Deuticke, Leipzig-Vienna, 1931.