Semiparametric Bernstein-von Mises theorem and bias, illustrated with Gaussian process priors

Ismaël Castillo [◇] , [*]

[◇] CNRS, Laboratoire Probabilités et Modèles Aléatoires (LPMA).
175, rue du Chevaleret, 75013 Paris, France.
E-mail: ismael.castillo@upmc.fr

Abstract

A semiparametric model is considered where the functional of interest is a shift parameter between two curves. A surprising example is provided where two at first sight indistinguishable Gaussian priors lead to quite different behaviors of the posterior distribution of the functional of interest. This phenomenon also illustrates that a condition introduced in [4] of approximation of the least favorable direction by the Gaussian prior is almost necessary for the Bernstein-von Mises theorem to hold.

Key words and phrases: Bayesian non-parametrics, Bernstein-von Mises theorem, semiparametric models, Gaussian process priors
AMS Classification: 62G05, 62G20.

1 Introduction

Recent years have seen the development of a theory for the behavior of Bayesian procedures -estimation, testing etc.- in high dimensional contexts, typically nonparametric. Understanding the behavior of posterior distributions when one puts a prior on an unknown function or high dimensional unknown parameter is the main goal. Some conditions are needed to guarantee appropriate behavior of the nonparametric posterior. The counter-example of Diaconis and Freedman [7] for estimation of the center of symmetry (which can in fact be seen as a semiparametric problem) illustrates this fact, since even innocent-looking priors can lead to inconsistency in the posterior. The pioneering works [11], [20], [12] give sufficient conditions that ensure posterior convergence at a nonparametric rate for some distances of interest.

In semiparametric problems, the quality of a Bayesian procedure can be measured using the marginal of the posterior with respect to the parameter of interest θ. Consistency is

[*]Work partly supported by ANR Grant 'Banhdits' ANR-2010-BLAN-0113-03.

a first requirement, we refer to [26] for a review of existing results. A very desirable further property of the convergence is given by the so-called Bernstein-von Mises theorem (abbreviated BvM in the sequel). Its conclusion is the convergence in total variation of the marginal posterior towards a Gaussian distribution at rate constant times $1/\sqrt{n}$ and centered around a frequentist efficient estimator of θ. From it one directly deduces that Bayesian confidence intervals asymptotically coincide with optimal frequentist ones. The extension of the BvM theorem to semiparametric frameworks has recently sparked off works by several authors, starting with [19], which provided a first set of general conditions, some of which were however a bit implicit in nature or hard to check in practice. The work [16] considered the proportional hazards model and obtained the semiparametric BvM for Lévy-type priors, exploiting their partial conjugacy in this model. In [4], a set of simple and easy-to-interpret sufficient conditions were given for the semiparametric BvM to hold in a quite general semiparametric framework, focusing mostly on Gaussian process priors. In [18], generic condition are given for estimation of linear functionals of the density in the case of sieve-type priors. In [1], a set of sufficient conditions for the semiparametric BvM to hold is given. We also mention the parallel direction of research considering the validity of BvM results in nonparametric settings [6], [8], [15] or growing dimension settings, see for instance [10], [2].

Back to the semiparametric framework, in [4], a novel condition consisted in measuring the quality of approximation of the least favorable direction of the model by the RKHS of the Gaussian prior. The present work shows that a condition of this type is needed in general for the BvM theorem to hold, thus complementing [4]. We focus mostly on a model of alignment of curves, which makes appear an interesting phenomenon of independent interest that we describe below. We also exhibit another semiparametric model in white noise, where the aforementioned condition is strictly necessary.

Let θ a real belonging to an open interval $\Theta \subset [-\tau, \tau]$, with $0 < \tau < 1/2$. Let f be an element of $L^2[0,1]$. For simplicity of treatment, we assume that f is 1-periodic. One observes in continuous time 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 $t \in [0,1]$, $n \geq 1$ is an integer measuring the amount of "information" present in the model and W_1, W_2 are independent standard Brownian motions. Both the real θ and the function f are unknown, making the model semiparametric in $\eta = (\theta, f)$. For an overview on semiparametric models, we refer to [22], Chapter 25. We are interested in estimation of θ from the Bayesian perspective. We denote by X the coupled observation of (Y, Z).

This type of model is of particular interest in signal processing, where it naturally arises in problems of alignment of noisy curves or images, see for instance [9]. For the sake of simplicity, the Gaussian white noise version of the model is considered here, but analogous results hold in discretized versions of it as well.

We follow a Bayesian approach and assign a priori probability distributions π_θ and π_f to the unknown θ and f. Those yield a prior $\Pi = \pi_\theta \otimes \pi_f$ on the pair $\eta = (\theta, f)$. This prior is updated with the data X leading to the posterior distribution $\Pi(\cdot|X)$. We assume that there exists a "true value" $\eta_0 = (\theta_0, f_0)$ of the parameter and are interested in the convergence of the posterior under the corresponding law $P_{\eta_0}^{(n)}$ of the observations X, as $n \to +\infty$. We often drop the index n, and denote $\mathbf{E}_{\eta_0}$ the expectation under this law.

As prior π_θ on θ, we simply take the uniform distribution on the interval $[-\tau, \tau]$. For f we consider two prior distributions on the space of 1-periodic square integrable functions. Let $\{\nu_k\}_{k\geq 1}$ be a sequence of independent $\mathcal{N}(0,1)$ random variables. We define, for any real $\alpha > 1$, the distributions (the variable u belongs to $\mathbb{R}$)

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

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

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

$$\gamma_{\eta_0} = -f_0'/2 \qquad \text{and} \qquad \tilde{I}_{\eta_0} = \frac{1}{2}\int_0^1 f_0'^2(u)du.$$

These quantities will be further interpreted in Section 2. In the sequel, we often drop the index η_0 and simply write $\tilde{I}$ and γ. We also denote

$$\Delta = -\tilde{I}^{-1}\int_0^1 [\gamma(u)dW_1(u) - \gamma(u-\theta_0)dW_2(u)] \overset{not.}{=} \tilde{I}^{-1}\mathcal{W}(1,-\gamma).$$

A special example illustrating our results is the function $f_0^{[\beta]}$, defined for $\beta > 3/2$ by

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

The condition $\beta > 3/2$ ensures that $f_0^{[\beta]}$ is a continuously differentiable ($\mathcal{C}^1$) function. One could also take same coefficients in front of cosine and sine, that is $(2k)^{-\frac{1}{2}-\beta}$. The conclusion of the Proposition below would be the same.

For any prior Π on $\Theta \times \mathcal{F}$, let $\Pi(\cdot \times \mathcal{F} \,|\, X)$ denote the marginal distribution on θ of the posterior distribution with respect to Π given the observation of the data $X = (Y, Z)$. Let $\|\cdot - \cdot\|$ denote the total variation distance between positive measures on $\mathbb{R}$ equipped with the Lebesgue σ-field.

The following statement is a consequence of the main result of the paper, see Theorem 1 in Section 2. The details on what is precisely meant by the semiparametric Bernstein-von Mises theorem are postponed to Section 2.

Proposition 1. *Let θ_0 belong to Θ and f_0 be the function $f_0^{[\beta]}$ defined in (1). Let us set $\Pi^\alpha = \pi_\theta \otimes \pi_f^\alpha$ and $\Pi^{\alpha,*} = \pi_\theta \otimes \pi_f^{\alpha,*}$. Take $\alpha = 4$ and $\beta = 2$. As $n \to +\infty$, it holds*

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

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 $n \to +\infty$,

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

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

One of the surprising aspects of the above result is that the two considered priors Π^α and $\Pi^{\alpha,*}$ share the same properties insofar as estimation of f_0 is concerned. More formally, if one considers the posterior distributions corresponding to Π^α and $\Pi^{\alpha,*}$ for nonparametric estimation of f_0 in our model, they have precisely the same rate of convergence towards f_0. Combining nonparametric techniques from [23], [3], it can be checked that, for $f_0 = f_0^{[\beta]}$, $\beta > 1$ and with $\| \cdot \|_2$ the L^2-norm on the interval $[0,1]$, there exist some positive constants a, b such that, denoting $\varepsilon_n = n^{-\alpha \wedge \beta/(2\alpha+1)}$,

$$\mathbf{E}_{\eta_0} \bar{\Pi}(a\varepsilon_n^{\alpha,\beta} \leq \|f - f_0\|_2 \leq b\varepsilon_n^{\alpha,\beta} \,|\, X) \to 1,$$

as $n \to +\infty$ and the preceding display holds for both $\bar{\Pi} = \Pi^\alpha$ and $\bar{\Pi} = \Pi^{\alpha,*}$. Note however that both posterior distributions have radically different behaviors insofar as estimation of the functional θ_0 of the full law P_{η_0} is concerned.

One could notice that π_f^α is stationary, while $\pi_f^{\alpha,*}$ is not. However, we will see below that many non-stationary priors will work here too (in the sense that they satisfy the Bernstein-von Mises theorem) while $\pi_f^{\alpha,*}$ does not.

The outline of the paper is as follows. In Section 2, we introduce the main notation and assumptions. We also explain how this paper is related to [4]. We then state our main result, and provide some discussion. Section 3 is devoted to the proof of the main result. Finally, in the appendix Section 4, it is checked that some particular priors verify the general assumptions.

2 Posterior concentration in the curve alignment model

Let $\{\varepsilon_p\}_{p \geq 1}$ denote the Fourier basis that we number as follows

$$\varepsilon_1(\cdot) = 1, \quad \varepsilon_{2k}(\cdot) = \sqrt{2}\cos(2\pi k \cdot), \quad \varepsilon_{2k+1}(\cdot) = \sqrt{2}\sin(2\pi k \cdot), \quad k \geq 1.$$

Let Θ be an open sub-interval of $[-\tau, \tau]$, where $0 < \tau < 1/2$. We assume that "the true" parameter θ_0 belongs to Θ.

Let $\mathcal{F}$ denote the linear space of all square-integrable functions on $[0,1]$, extended by periodicity to $\mathbb{R}$. Any element f in $\mathcal{F}$ has Fourier coefficients $f_k = \int_0^1 f(u)\varepsilon_k(u)du$, any $k \geq 1$. We further impose $f_1 = \int_0^1 f = 0$ for any f in $\mathcal{F}$.

The last requirement is for notational simplicity. The results of this paper immediately extend to the case where f_1 is not necessarily zero by adding a first Fourier term to the priors too. Next we state the assumed regularity conditions on "the true" f_0.

Condition (**R**). Assume that f_0 belongs to $\mathcal{F}$, is continuously differentiable and, for some $\beta > 1$,

$$f_{0,1} = 0, \qquad |f_{0,2}| > 0, \qquad \sum_{k\geq 1} k^{2\beta} \{f_{0,2k}^2 + f_{0,2k+1}^2\} < +\infty$$

The condition $|f_{0,2}| > 0$ ensures identifiability in imposing 1 as the smallest period of f. The last condition ensures some (Sobolev) smoothness. The case of more irregular nuisance functions, for instance if $\beta < 1$, is also interesting, but one leaves the set of "smooth" models, see [5] for some frequentist results.

As an example, the function $f_0^{[\beta]}$ defined by (1) fulfills conditions (**R**) for $\beta > 3/2$ (as noted above, asking $\beta > 3/2$ for this specific function guarantees that $f_0^{[\beta]}$ is $\mathcal{C}^1$).

Also, in this article we shall focus on the case of 1-dimensional θ, which could be extended to the multi-dimensional case of $\theta \in \mathbb{R}^d$, $d > 1$ without much effort.

2.1 Semiparametric structure

Likelihood. For any pair (θ, f) in $\Theta \times \mathcal{F}$, the probability $\mathbf{P}_{\theta,f}$ of observing the pair of paths (Y, Z) given (θ, f) is related to the probability $\mathbf{P}_0$ of observing (Y, Z) given $f = 0$ through Girsanov's formula.

The likelihood of $X = (Y, Z)$ in our model is given by the Radon-Nikodym derivative

$$\begin{aligned}\frac{d\mathbf{P}_{\theta,f}}{d\mathbf{P}_0}(X) =& \exp\left(n\int_0^1 f(t)dY(t) - \frac{n}{2}\int_0^1 f(t)^2dt\right)\\ &\times \exp\left(n\int_0^1 f(t-\theta)dZ(t) - \frac{n}{2}\int_0^1 f(t-\theta)^2dt\right)\\ =& \exp\left(n\int_0^1 \{f(t)dY(t) + f(t-\theta)dZ(t)\} - n\int_0^1 f(t)^2dt\right),\end{aligned}$$

noticing that 1-periodicity of f implies that the two quadratic terms are the same. We denote by $\ell_n(\theta, f)$ the log-likelihood.

LAN expansion. The log-likelihood-difference $\Lambda_n(\theta, f) = \ell_n(\theta, f) - \ell_n(\theta_0, f_0)$ under the true (θ_0, f_0) is obtained replacing Y, Z by their expressions and equals

$$\begin{aligned}\Lambda_n(\theta, f) =& -\frac{n}{2}\int_0^1 (f - f_0)^2(t)dt - \frac{n}{2}\int_0^1 \{f(t-\theta) - f_0(t-\theta_0)\}^2(t)dt\\ &+ \sqrt{n}\int_0^1 (f - f_0)(t)dW_1(t) + \sqrt{n}\int_0^1 \{f(t-\theta) - f(t-\theta_0)\}dW_2(t).\end{aligned}$$

For any true parameter (θ_0, f_0), we define an inner product $\langle \cdot, \cdot \rangle_L$ on $\mathbb{R} \times \mathcal{F}$ by, for any $(h_1, a_1), (h_2, a_2)$ in $\mathbb{R} \times \mathcal{F}$,

$$\langle\, (h_1, a_1)\,,\, (h_2, a_2)\,\rangle_L = \langle a_1, a_2 \rangle_2 + \langle a_1 - h_1 f_0', a_2 - h_2 f_0' \rangle_2,$$

with $\langle \cdot, \cdot \rangle_2$ the usual inner product on $L^2[0,1]$. The norm associated to $\langle \cdot, \cdot \rangle_L$ will be denoted $\| \cdot \|_L$ and for brevity $\|h, a\|_L$ stands for $\|(h,a)\|_L$. This norm defines an Hilbert space structure on $\mathbb{R} \times \mathcal{F}$.

For any $(h, a) \in \mathbb{R} \times \mathcal{F}$, we also denote

$$\mathcal{W}(\,(h,a)\,) = \int_0^1 a(u) dW_1(u) + \int_0^1 (a - h f_0')(u - \theta_0) dW_2(u).$$

Notice that for any $d \geq 1$ and any fixed $v_1, \ldots, v_d$ each in $\mathbb{R} \times \mathcal{F}$, the variable $W(v_1, \ldots, v_d)$ is centered multivariate Gaussian, of covariance structure $(\langle\, v_i\,,\, v_j\,\rangle_L)_{1 \leq i,j \leq d}$.

In the sequel we use as shorthand notation, for any (θ, f) in $\Theta \times \mathcal{F}$,

$$h_\theta = \sqrt{n}(\theta - \theta_0), \quad a_f = \sqrt{n}(f - f_0),$$

and also, for the following normalized remainder term of the Taylor expansion of f_0,

$$D_n(t, h_\theta) = \sqrt{n}\{f_0(t - \theta) - f_0(t - \theta_0)\} + h_\theta f_0'(t - \theta_0).$$

The previous definitions enable to rewrite the log-likelihood difference $\Lambda_n(\theta, f)$ as

$$\Lambda_n(\theta, f) = -\frac{n}{2} \| \,\theta - \theta_0, f - f_0\, \|_L^2 + \sqrt{n} \mathcal{W}(\,\theta - \theta_0, f - f_0\,) + R_n(\theta, f), \tag{2}$$

where the remainder term $R_n(\theta, f)$ is the sum of the four following terms $R_{n,i}$, $1 \leq i \leq 4$,

$$\begin{aligned}
R_{n,1}(\theta, f) &= \int_0^1 (a_f(t - \theta) - a_f(t - \theta_0)) dW_2(t) \\
R_{n,2}(\theta, f) &= \int_0^1 D_n(t, h_\theta) dW_2(t) \\
R_{n,3}(\theta, f) &= -\int_0^1 (a_f - h_\theta f_0')(t - \theta_0)\, [a_f(t - \theta) - a_f(t - \theta_0) + D_n(t, h_\theta)]\, dt \\
R_{n,4}(\theta, f) &= -\frac{1}{2} \int_0^1 [a_f(t - \theta) - a_f(t - \theta_0) + D_n(t, h_\theta)]^2\, dt.
\end{aligned}$$

While (2) is an identity valid for any θ, f, one can for a moment consider it on shrinking neighborhoods of size $1/\sqrt{n}$ of the true (θ_0, f_0). For any fixed (t, a) in $\Theta \times \mathcal{F}$,

$$\Lambda_n(\theta_0 + \frac{t}{\sqrt{n}}, f_0 + \frac{a}{\sqrt{n}}) = -\frac{1}{2} \| \,t, a\, \|_L^2 + \mathcal{W}(\,t, a\,) + R_n(\theta_0 + \frac{t}{\sqrt{n}}, f_0 + \frac{a}{\sqrt{n}}),$$

where the remainder term, as can be checked from the previous expressions, tends to zero in probability as $n \to +\infty$ (use the continuity at θ_0 of $\theta \to \int_0^1 (a(t - \theta) - a(t - \theta_0))^2 dt$

for any square-integrable function a and the fact that for any f_0 satisfying **(R)**, we have $\int_0^1 D_n(t, \theta_0 + t/\sqrt{n})^2 dt = o(1)$ as $n \to +\infty$, similarly to [4], Lemma 5).

This expansion is called local asymptotic normality (LAN) property of the considered sequence of statistical experiments, see for instance [17] or [25], Section 3.11. The expansion here has also the property that the tangent space coincides with $\mathbb{R} \times \mathcal{F}$ and that the approximating paths $(\theta_0 + t/\sqrt{n}, f_0 + a/\sqrt{n})$ are linear.

The fact that the expansion holds means that the model ressembles asymptotically a (shift) Gaussian experiment, as studied in the (more general) theory of limiting experiments due to Le Cam. The inner-product $\langle \cdot, \cdot \rangle_L$ should be understood as a generalization of the Fisher information metric arising in smooth parametric models. One can study concepts like optimality of estimation of θ - efficiency in this context- directly from the Hilbert space structure generated by $\langle \cdot, \cdot \rangle_L$.

Efficient information and least favorable direction. It follows from the general results for semiparametric models in this context, see [17] or [25], Section 3.11, that the optimal amount of information when estimating θ for unknown f is given by the projection of the vector $(1, 0)$ into the orthogonal for $\langle \cdot, \cdot \rangle_L$ of the subspace $\{0\} \times \mathcal{F}$. If we denote by $(0, \gamma)$ the orthogonal projection of $(1, 0)$ onto $\{0\} \times \mathcal{F}$, we have, for any (t, a) in $\mathbb{R} \times \mathcal{F}$,

$$\| t, a \|_L^2 = \| 1, -\gamma \|_L^2 t^2 + \| 0, a - t\gamma \|_L^2.$$

Notice that by positivity of the norm this quantity is always larger than $\| 1, -\gamma \|_L^2 t^2$ and that this lower bound is achieved when $a = t\gamma$. In the model considered here, a simple calculation shows that $\gamma = -f_0'/2$, which is called least favorable direction. The quantity

$$\tilde{\mathcal{I}} = \| 1, -\gamma \|_L^2 = \frac{1}{2} \int_0^1 f_0'^2(u) du > 0$$

is called efficient information. Efficient semiparametric estimators of θ have variance $\tilde{\mathcal{I}}^{-1}/\sqrt{n}$ asymptotically and such an estimator $\tilde{\theta}_n$ is called linear efficient if it has the expansion, as $n \to +\infty$,

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) = \tilde{\mathcal{I}}^{-1} \mathcal{W}(1, -\gamma) + o_{P_{\eta_0}^n}(1),$$

which can be rewritten as $\tilde{\theta}_n = \theta_0 + \Delta/\sqrt{n} + o_P(1/\sqrt{n})$, with $\Delta = \tilde{\mathcal{I}}^{-1} \mathcal{W}(1, -\gamma)$. It is important to note that in the model under consideration, there is a loss of information, in the sense that the information in the semiparametric context is smaller than in the parametric context where f is known, where one can check that it equals $\|1, 0\|_L^2 = \|f_0'\|_2^2$.

2.2 Prior

The prior π_θ on θ is chosen to be the uniform distribution on the interval $[-\tau, \tau]$. The prior on the nuisance f is defined as the distribution of the Gaussian process

$$\pi_f^\sigma \sim \sqrt{2} \sum_{k=1}^{+\infty} \left[\sigma_{2k} \nu_{2k} \cos(2\pi k u) + \sigma_{2k+1} \nu_{2k+1} \sin(2\pi k u) \right],$$

for some square integrable positive sequence $\{\sigma_k\}_{k\geq 1}$. This induces a prior on the pair (θ, f) denoted $\Pi^\sigma = \pi_\theta \otimes \pi_f^\sigma$. Among those priors, we also denote

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

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

We denote $\Pi^\alpha = \pi_\theta \otimes \pi_f^\alpha$ and $\Pi^{\alpha,*} = \pi_\theta \otimes \pi_f^{\alpha,*}$. We could also consider sieve priors, for which the sum is truncated at some cut-off $k(n)$. For instance, all results obtained in the sequel for the priors $\pi_f^\alpha, \pi_f^{\alpha,*}$ are also valid when the sum is truncated at $k(n) = \lfloor n^{1/(2\alpha+1)} \rfloor$. These cutted distributions make the prior depend on n, which makes the notation slightly more involved, so for simplicity we focus on the case of the infinite prior.

We shall assume that the sequence $\{\sigma_k\}$ is decreasing. Let $\{\gamma_k\}$ denote the Fourier coefficients of $\gamma = -f_0'/2$. We assume that σ fulfills the following technical requirement, as $n \to +\infty$,

$$\sum_{k\geq 1}(1 \wedge n^{-1}\sigma_k^{-2})\gamma_k^2 = o(1). \tag{3}$$

Equation (3) is automatically fulfilled for instance for $f_{0,k} = f_{0,k}^{[\beta]} = k^{-\frac{1}{2}-\beta}$ and $\sigma_k \sim k^{-1/2-\alpha}$, for any $\beta > 1$ and $\alpha > 0$. Since σ is decreasing and square integrable, and thus tends to 0, there exists a largest integer K_n such that $n\sigma_{K_n}^2 \geq 1$. We define the sequence

$$\gamma_{[n]} = \sum_{k\leq K_n}\gamma_k\varepsilon_k(\cdot).$$

Associated to the Gaussian prior π_f^σ is the Reproducing Kernel Hilbert Space (RKHS) $\mathbb{H}^\sigma$, see [24] for basic properties. It can be checked that for the prior at stake

$$\mathbb{H}^\sigma = \{h = (h_k)_{k\geq 1}, \quad \sum_{k\geq 1}\sigma_k^{-2}h_k^2 < +\infty\}.$$

When no confusion is possible, we drop the index σ in the notation for $\mathbb{H}^\sigma$. This space is equipped with the Hilbert-space norm, for $h \in \mathbb{H}$,

$$\|h\|_{\mathbb{H}}^2 = \sum_{k\geq 1}\sigma_k^{-2}h_k^2.$$

One can notice that (3) ensures some approximation of the least favorable direction γ by elements of the RKHS $\mathbb{H}$ of the prior. More precisely, (3) implies the existence of a sequence $\rho_n \to 0$ such that

$$\inf_{h\in\mathbb{H},\ \|h-\gamma\|_2<\rho_n}\|h\|_{\mathbb{H}}^2 \leq n\rho_n^2.$$

Indeed, the sequence $\gamma_{[n]}$ defined above belongs to $\mathbb{H}$ and, under Condition (3), is such that $\|\gamma_{[n]} - \gamma\|_2^2 + n^{-1}\|\gamma_{[n]}\|_{\mathbb{H}}^2$ tends to 0.

To a Gaussian process $(Z(t))_{t\in[0,1]}$ defined on the probability space Ω with covariance function $K(\cdot,\cdot)$ and associated RKHS $\mathbb{H}$, one can always associate a map U from the linear span of the functions $t \to K(\cdot,t)$ into $L^2(\Omega)$ by

$$U : \sum_{i=1}^{p} a_i K(\cdot,t_i) \to \sum_{i=1}^{p} a_i Z(t_i,\omega).$$

It is easily seen that U is an isometry, which can be extended to an isometry from $\mathbb{H}$ into $L^2(\Omega)$, see for instance [24][Section 3]. In our framework, explicit calculations easily reveal that for the prior π_f^σ, denoting $\{\varepsilon_k\}$ the Fourier basis, U is the map described by

$$\begin{array}{rccc} U : & \mathbb{H} & \to & L^2(\Omega) \\ & \sum_{k=1}^{+\infty} \kappa_k \varepsilon_k(\cdot) & \to & \sum_{k=1}^{+\infty} \sigma_k^{-1} \kappa_k \nu_k. \end{array}$$

There is a slightly more compact way to think about $U\hbar$ for $\hbar \in \mathbb{H}$. Let f denote a function drawn according to the prior, that is $f(\cdot) = \sum_{k\geq 1} \sigma_k \nu_k \varepsilon_k(\cdot)$. Let us write $\hbar(\cdot) = \sum_{k\geq 1} \kappa_k \varepsilon_k(\cdot)$. Then notice that by the preceding identity, $U\hbar = \sum_{k=1}^{+\infty} \sigma_k^{-2}(\sigma_k\nu_k)\kappa_k$, which can be interpreted as an inner product "$\langle f, \hbar\rangle_{\mathbb{H}}$" in $\mathbb{H}$, although f does not belong to $\mathbb{H}$.

2.3 Bernstein-von Mises Theorem, sufficient conditions

We now have a prior $\Pi = \Pi^\sigma = \pi_\theta \otimes \pi_f^\sigma$ on the pair (θ, f), which combined with the data X leads to the posterior distribution using Bayes formula. We are interested in the marginal distribution in θ, which is given by, for any measurable $B \subset \Theta$,

$$\Pi^\sigma(B \,|\, X) = \frac{\int_B \int_{\mathcal{F}} e^{\ell_n(\theta,f)} d\pi_f^\sigma(f) d\pi_\theta(\theta)}{\int_\Theta \int_{\mathcal{F}} e^{\ell_n(\theta,f)} d\pi_f^\sigma(f) d\pi_\theta(\theta)}.$$

We also introduce a notation for the posterior distribution in the model where θ is known to be θ_0, which will appear as a technical tool in our results. For a measurable set C in $\mathcal{F}$,

$$\Pi^{\theta=\theta_0}(C \,|\, X) = \frac{\int_C e^{\ell_n(\theta_0,f)} d\pi_f^\sigma(f)}{\int_{\mathcal{F}} e^{\ell_n(\theta_0,f)} d\pi_f^\sigma(f)}.$$

We also denote $E_\Pi^{\theta=\theta_0}(\cdot \,|\, X)$ the expectation with respect to this measure.

In the present semiparametric context, a very desirable result for the marginal posterior $\Pi^\sigma(\cdot \times \mathcal{F} \,|\, X)$ is the so-called Bernstein-von Mises phenomenon, which asserts that the total variation distance between the marginal posterior and the normal $\mathcal{N}(\tilde{\theta}_n, 1/(n\tilde{\mathcal{I}}))$ converges (in $P_{\eta_0}^{(n)}$-probability) to 0 as $n \to +\infty$, where $\tilde{\theta}_n$ is an efficient estimator of θ in the semiparametric sense and $\tilde{\mathcal{I}}$ the efficient information. Alternatively, one can center the target normal distribution at the point $\theta_0 + \Delta/\sqrt{n}$ or at any linear and efficient estimator of θ_0 (this might be the case for a maximum likelihood-type estimator $\hat{\theta}^{MLE}$, but the

previous centering provides more flexibility, since one does not have to prove efficiency of $\hat{\theta}^{MLE}$, see [21] for a discussion).

We now briefly describe the ideas behind Theorem 2 in [4] and specialize the conditions to the framework of the curve alignment model under consideration. The result in itself is not used in this paper, so the proof presented in Section 3 is independent of it, though related. First, there should exist a sequence γ_n of elements in $\mathbb{H}$ such that, for some sequence $\rho_n \to 0$,

$$\|0, \gamma - \gamma_n\|_L \leq \rho_n \qquad \text{and} \qquad \|\gamma_n\|_{\mathbb{H}}^2 \leq n\rho_n^2. \tag{4}$$

Since in our model the LAN-norm of $(0, a)$ is a multiple of the L^2-norm of a, we have already checked above that under (3) this condition is satisfied for the choice $\gamma_n = \gamma_{[n]}$.

Concentration ($\mathbf{C_1}$). This condition requires the existence of a rate $\varepsilon_n \to 0$ and of a sequence of measurable sets $\mathcal{F}_n$ in $\mathcal{F}$ such that, if $\mathcal{F}_n(\theta)$ denotes $\mathcal{F}_n + (\theta - \theta_0)\gamma_n$,

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

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

as $n \to +\infty$, in $P_{\eta_0}^{(n)}$-probability.

This condition asks for the posterior distribution to concentrate at some rate ε_n towards the true $\eta_0 = (\theta_0, f_0)$, in terms of the LAN-norm. This assumption has a non-parametric flavour and can typically be checked using Bayesian nonparametric techniques, as developed for instance in [11] (see [23] for specific techniques for Gaussian processes). Often, results are first obtained in a distance for which some tests are known to exist. Some extra work might then be needed to obtain results in terms of $\|\cdot\|_L$. Notice however that the condition just ask for the existence of a rate, which need not be very fast (of course a too slow rate makes the next condition harder to check). One might also directly look for a rate in the distance of interest, see for instance [13] for some results related to the sup-norm.

Local Shape ($\mathbf{N_1}$). Setting $V_n = \{(\theta, f) \in \Theta \times \mathcal{F}_n, \quad \|\theta - \theta_0, f - f_0\|_L \leq 2\varepsilon_n\}$, this condition requires that for any (θ, f) in V_n,

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

This condition means that the expansion of the log-likelihood around the true parameter should not be too far from the one which would be obtained in a Gaussian shift experiment. This condition enables the posterior distribution to have a Gaussian shape in the limit $n \to +\infty$. The less "complex" (or "large") the sieves $\mathcal{F}_n$ are, the easiest this condition to verify. For Gaussian priors, Borell's inequality is often a useful tool to construct sieves verifying this conditions, see Section 4.

We refer to [4] for further comments on how to check conditions $(\mathbf{C_1})$ and $(\mathbf{N_1})$ in general. Here, as verified in Section 4, both conditions hold for instance for the priors Π^α and $\Pi^{\alpha,*}$, when choosing ε_n proportional to $n^{-\alpha\wedge\beta/(2\alpha+1)}$, and $\alpha > 1+\sqrt{3}/2$, $\beta \geq 2$.

Conditions $(\mathbf{C_1})$, $(\mathbf{N_1})$ will be part of the assumptions of Theorem 1 below. We note that $(\mathbf{C_1})$ is slightly weaker than (C') in [4] (the condition on the Kullback-Leibler neighborhoods will not be needed in the proof of Theorem 1).
The last assumption made in [4] is related to how well the least favorable direction γ can be approximated by elements of the RKHS $\mathbb{H}$ of the Gaussian prior. Given an approximation γ_n of γ such that (4) is fulfilled with rate ρ_n, condition **(E)** assumes that, as $n \to +\infty$,

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

The better the approximation of γ, the faster the rate ρ_n and the weaker this condition becomes, so one typically chooses γ_n such that this approximation is the best possible.

We note that the condition $\mathcal{W}(0, \gamma - \gamma_n) = o_P(1)$ is automatically verified in our context when taking $\gamma_n = \gamma_{[n]}$, by definition of $\mathcal{W}$ and because $\|0, \gamma - \gamma_n\|_2$ tends to 0.

2.4 Main result

Let us denote, for any sequence σ as specified in Section 2.2,

$$\zeta_n^\sigma = \frac{\pi}{\sqrt{n}} \sum_{k=1}^{+\infty} k f_{0,2k} f_{0,2k+1} (\sigma_{2k+1}^{-2} - \sigma_{2k}^{-2}) \left\{ \frac{(2n)^2}{(2n + \sigma_{2k}^{-2})(2n + \sigma_{2k+1}^{-2})} \right\}. \tag{5}$$

Theorem 1. *Let $\eta_0 = (\theta_0, f_0)$ belong to $\Theta \times \mathcal{F}$, where f_0 satisfies $(\mathbf{R})$. Suppose that σ is a decreasing square integrable sequence and satisfies (3). Suppose that the corresponding prior Π^σ satisfies conditions $(\mathbf{C_1})$ and $(\mathbf{N_1})$. Then, if ζ_n^σ is defined by (5), as $n \to +\infty$,*

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

Theorem 1 gives the exact expression up to the order $o_P(1/\sqrt{n})$ of the centering term of the semiparametric Bernstein-von Mises theorem in the model of alignment of curves, for a fairly large family of priors Π^σ, provided two basic requirements of concentration $(\mathbf{C_1})$ and shape $(\mathbf{N_1})$ are met. Notice that the semiparametric BvM theorem holds if and only if $\zeta_n^\sigma = o(1)$ as $n \to +\infty$. We illustrate the behavior of ζ_n^σ below for priors having a polynomial decrease. The main consequences of this result are

- Structural nonparametric assumptions such as non-parametric $(\mathbf{C_1})$ posterior rate ε_n of estimation of $\eta = (\theta, f)$ and approximately Gaussian shape $(\mathbf{N_1})$ of the model are not enough to characterize the semiparametric problem when the least favorable direction γ is nonzero (Theorem 1 in [4] shows that these conditions are sufficient if $\gamma = 0$, for general, possibly non-Gaussian, priors). Indeed, a bias term ζ_n^σ potentially appears depending on the choice of σ, and ζ_n^σ becomes dominant for some regularities,

typically (but not always) when a smooth prior (σ_k^{-1} large) is combined with a less regular f ($f_{0,k}$ goes slowly to 0). This confirms the need, as already suggested in [4], of avoiding oversmoothing in the prior choice. The safe way indicated in [4] being "choosing a prior with [regularity] α as small as allowed by the theory". This illustrates the necessity of a condition of the type **(E)**. If this condition is not met, then the bias terms can become dominant, as will be seen in the examples below.

- One step further, this result can be interpreted in a more refined way at the light of how the least favorable direction is approximated through the prior. The fact that $\gamma = -f_0'/2$ (here) plays a role can be guessed when recalling that $\gamma_{2k} = -k\pi f_{0,2k+1}$ and $\gamma_{2k+1} = k\pi f_{0,2k}$. Thus ζ_n^σ can be read as a twisted inner-product between f_0 and γ, where the prior comes in through σ. The fact that the expression cancels if $\sigma_{2k} = \sigma_{2k+1}$ is related to the fact that f_0 and γ are orthogonal in $L^2[0,1]$ (recall that f_0 is 1-periodic). This makes it possible for two close priors to behave differently. In general, when γ and f_0 are not orthogonal, the contribution of the bias when oversmoothing (α large) is analogous for priors of the same regularity, see the example of the amplitude estimation model below.

- To synthesize, the present case corresponds to a critical situation where validity of BvM depends on much more refined properties of the prior than its only regularity. Though the orthogonality of f_0 and γ makes it possible for BvM to be verified even when oversmoothing (for priors where σ_{2k} is equal or very close to σ_{2k+1}), in most cases a second order term appears in the form of ζ_n^σ which introduces an extra bias.

Let us now illustrate the behavior of ζ_n^σ in function of σ and f_0. We consider the simple case of $f_0 = f_0^{[\beta]}$ such that $f_{0,k}^{[\beta]} = k^{-1/2-\beta}$. We consider the priors $\Pi^\sigma = \Pi^\alpha$ and $\Pi^\sigma = \Pi^{\alpha,*}$. First it should be noted that both priors satisfy $(\mathbf{C_1})$, $(\mathbf{N_1})$ for instance under the condition that $\alpha > 1 + \sqrt{3}/2 = 1.87$ and $\beta \geq 2$ (the condition is even satisfied in a slightly larger zone, see Section 4 or [4], Fig. 1). The proof of this is very similar to the proof of the corresponding result for the translation model studied in [4]. For completeness the details are included in Section 4. Once this has been noted, we immediately get, for $\Pi^\sigma = \Pi^\alpha$, that $\zeta_n^\sigma = 0$. On the other hand, if $\Pi^\sigma = \Pi^{\alpha,*}$, then ζ_n^σ is of the order of $\kappa_n^{2\alpha-2\beta+1} n^{-1/2}$, where $\kappa_n = \lfloor n^{1/(2\alpha+1)} \rfloor$, if $2\alpha - 2\beta + 1 > 0$ (for smaller α's it is of the order of $1/\sqrt{n}$, with an extra $\log n$ in case of equality). Thus, according to the conclusion of Theorem 1, there is an extra bias appearing in the centering if this term is not a $o(1)$ as $n \to +\infty$, that is if $\alpha \geq 2\beta - 1/2$. This includes Proposition 1, where we took $\alpha = 4$ and $\beta = 2$.

Remark. Another surprising aspect of the above is that, in the particular case where one would assume beforehand that the true f_0 has Fourier coefficients $k^{-1/2-\lambda}$ for some (unknown) $\lambda > 0$, the prior $\pi_f^{*,\alpha}$ with coefficients $k^{-1/2-\alpha}$ would in principle look more natural than π_f^α. Theorem 1 shows however that, if $\alpha = 4$, the corresponding prior is not suitable for estimating θ_0 when $\lambda = 2$, in that the posterior marginal does not concentrate at rate $1/\sqrt{n}$ around θ_0 (this phenomenon appears for α's at least slightly larger than β, if $\alpha = \beta$ then both priors lead to the BvM theorem).

2.5 Discussion

An examination of the proof of Theorem 1 (or of the main result in [4]) shows that the following quantity is responsible for the "extra-bias" phenomenon when it exists. Let γ_n be an approximating sequence in $\mathbb{H}$ such that (4) is satisfied with the fastest possible rate ρ_n. Let us set

$$\xi_n(f) = -\sqrt{n}\langle (0, f - f_0), (0, \gamma - \gamma_n)\rangle_L + \frac{U\gamma_n}{\sqrt{n}}. \tag{6}$$

For this term to be negligible, we need, with the notation $h = \sqrt{n}(\theta - \theta_0)$,

$$\log \Pi^{\theta=\theta_0}(e^{h\xi_n(f)}|X) = o_P(1 + h^2).$$

As examplified by priors Π^α and $\Pi^{*,\alpha}$, failure of this condition induces an extra bias in the Bernstein-von Mises theorem. So, what is required is that *joint estimation of the true f_0 and the least favorable direction γ be good enough.* Also, as we have noted already, *concentration properties of the posterior around the true f_0 are not enough* for the BvM theorem to hold, since from this viewpoint Π^α and $\Pi^{\alpha,*}$ cannot be distinguished. Something more is needed, which involves approximation of the least favorable direction γ.

What is generally the order of the terms in (6) ?

Let us briefly explain how (6) was bounded in [4], resulting in the condition $\sqrt{n}\varepsilon_n\rho_n \to 0$. For the first term in ξ_n, by construction $\|0, f - f_0\|_L$ concentrates at rate ε_n. Then, by assumption on γ_n, the term $\|0, \gamma - \gamma_n\|_L$ is of the order ρ_n. Cauchy-Schwarz inequality leads to an upper- bound $\sqrt{n}\varepsilon_n\rho_n$ for this term. For the second term, under the prior $U\gamma_n$ is $\mathcal{N}(0, \|\gamma_n\|_L^2)$ distributed. Hence the event

$$\mathcal{H} = \{\omega, \quad |U\gamma_n(\omega)| > M\sqrt{n}\varepsilon_n\|\gamma_n\|_{\mathbb{H}}\}$$

has prior probability at most $\exp(-Mn\varepsilon_n^2/2)$. From this fact one can check that the posterior probability of the event $\mathcal{H}$ tends to zero, using Lemma 1 in [12]. This means one can restrict to the event $\mathcal{H}$, which leads to a bound proportional to $\varepsilon_n\|\gamma_n\|_{\mathbb{H}}$ for the term $U\gamma_n/\sqrt{n}$. By assumption on γ_n, the previous bound is at most of order $\sqrt{n}\varepsilon_n\rho_n$. This shows that (6) is bounded by $(\sqrt{n}\varepsilon_n\rho_n)h$. Hence the condition $\sqrt{n}\varepsilon_n\rho_n \to 0$.

Of course, when considering specific models, the two terms considered above might present some simplifications which do not make the condition sharp. This is what happens in our model. Indeed, the fact that *f_0 and γ are orthogonal in $L^2(0,1)$* implies some simplifications in (6). In that one can actually interpret ζ_n^σ as a "second order term".

However, even here the condition $\sqrt{n}\varepsilon_n\rho_n \to 0$ is already fairly precise, as one can see by looking at the example of $f_0^{[\beta]}$ and of prior Π^*. As seen above, ζ_n^σ can for some priors be as large as $\sqrt{n}n^{-2\beta/(2\alpha+1)}$, which must tend to zero if one wants the BvM theorem to hold, while the condition that $\sqrt{n}\varepsilon_n\rho_n \to 0$ amounts to asking for the slightly stronger condition $\sqrt{n}n^{(1-2\beta)/(2\alpha+1)} \to 0$. We notice that in this case, it is still possible for well chosen priors to verify BvM even in the zone where the previous conditions do not hold (for instance, here, by choosing $\sigma_{2k} = \sigma_{2k+1}$). However, it appears that in general, oversmoothing will typically result in extra bias, and the next example shows what seems to be the prototypical

situation, where the BvM theorem fails to hold if $\sqrt{n}\varepsilon_n\rho_n$ does not tend to 0, at least for some true parameter $\eta_0 = (\theta_0, f_0)$.

Necessity of the condition $\sqrt{n}\varepsilon_n\rho_n \to 0$ in general

Here we show that there are models (in fact presumably most models) where failure of $\sqrt{n}\varepsilon_n\rho_n \to 0$ implies that the BvM theorem does not hold. As what preceeds suggests, let us consider a model where $(0, f_0)$ and $(0, \gamma)$ are not orthogonal for $\langle \cdot, \cdot \rangle_L$. For $t \in [0, 1]$, let the observation process be given by the path X defined by

$$dX(t) = \theta f(t)dt + \frac{1}{\sqrt{n}}dW(t),$$

for some square integrable function f and a positive real θ. It can be checked that the least favorable direction is given by $\gamma(\cdot) = f_0(\cdot)/\theta_0$. In particular, the inner product $\langle (0, f_0), (0, \gamma) \rangle_L$ is strictly positive. For the function $f_0 = f_0^{[\beta]}$, and prior Π^α, where say $\alpha \geq \beta$ (for $\alpha < \beta$ the considered condition is empty if $\beta > 1/2$), combining results in [23] and [3] one can check that $\varepsilon_n = n^{-\beta/(1+2\alpha)}$ is the precise rate of convergence of the posterior when $f_0 = f_0^{[\beta]}$, up to a multiplicative constant. Also, ρ_n is of the same order as ε_n when $\alpha \geq \beta$, since γ is equal to f_0 up to a constant. Computations similar to the ones in Lemma 1 below show that the extra bias term induced by (6) when $\alpha \geq \beta$ is of the order of $h(\sqrt{n}n^{-2\beta/(1+2\alpha)})$ which can thus be identified to $h(\sqrt{n}\varepsilon_n\rho_n)$ up to a multiplicative constant, so the above claim is established.

Overspecification, a remark

In this paragraph we would like to comment on the special subcase of the main model of this paper corresponding to the case of symmetric (i.e. even) functions f_0, for which $f_{0,2k+1} = 0$. If one knows beforehand that the curve is symmetric, then the model can be seen to reduce (from the semiparametric perspective) to the problem of center of symmetry estimation in Gaussian white noise, see [14], for which the semiparametric BvM theorem has been obtained in [3]. In particular, there is no information loss in this model, and the efficient information equals the Fisher information in the parametric case, that is $\|f_0'\|^2$.

Now a natural question is: does one recover the results in that specific (sub-)model using a prior such as Π^α ? The answer is no, since Theorem 1 applied with the prior Π^α ($\alpha = 2$, say) implies the concentration of the posterior with variance $\tilde{I}^{-1} = 2\|f_0'\|^{-2}$, so the (optimal) Bernstein-von Mises theorem does not hold for this prior. This could seem unexpected: indeed, under conditions similar to $(\mathbf{C_1}), (\mathbf{N_1})$, the BvM theorem is established in [3] for a prior putting mass only on symmetric functions (that is $\sigma_{2k+1} = 0$). Adding the antisymmetric part of the prior by having nonzero σ_{2k+1}'s - the prior is somehow overspecified - yields suboptimality. From the prior mass point of view, both priors however do charge in about the same way neighborhoods of the true (θ_0, f_0), but, again, have quite different behaviors in the limit (the difference is less important here, with a loss only in terms of the constant in the rate).

In fact, this phenomenon is very common, already in simple parametric models. If the parameter takes the form $\theta = (\theta_1, \theta_2)$ and the Fisher information matrix $(I_{ij})_{i\leq 2, j\leq 2}$ is invertible but non-diagonal, then there is a loss of information as far as estimation of θ_1 is

concerned. If θ_2 is known, say $\theta_0 = (\theta_{01}, 0)$, but the prior charges both θ_1 and θ_2, then the posterior in θ achieves the (suboptimal) information $I_{11} - (I_{12}^2/I_{22})$. The present remark constitutes an infinite-dimensional analogue of this fact.

3 Proof of Theorem 1

The proof follows to some extent the steps in [4], up to the fact that here bias terms, which in some cases do not vanish, need to be dealt with. In particular, we introduce an approximating sequence γ_n of γ in the RKHS $\mathbb{H}$ of the prior. We point out the following variant of the proof: one could, once the likelihood localized in an appropriate neighborhood with $(\mathbf{C_1}), (\mathbf{N_1})$, write down an analogue of Lemma 1 without introducing γ_n and changing variables (the computations are still relatively similar, though), exploiting the partial conjugacy of the prior in the nonparametric component. The advantage of the proof below is that it is more generic. Up to Lemma 1, which is model-specific, it can be used for different models, even in situations where the model is not partially conjugate, provided one can evaluate the posterior probability appearing in Lemma 1.

Proof of Theorem 1. First let us gather a few properties of the approximating sequence $\gamma_{[n]}$ defined as

$$\gamma_n = \gamma_{[n]} = \sum_{k=1}^{K_n} \gamma_k \varepsilon_k,$$

where K_n was defined as the largest integer such that $n\sigma_k^2 \geq 1$. Due to (3), as $n \to +\infty$,

$$\|0, \gamma - \gamma_n\|_L = o(1), \qquad n^{-1}\|\gamma_n\|_{\mathbb{H}}^2 = o(1).$$

As a direct consequence of (3), we also have that

$$\sum_{k \geq 1} (\sigma_k^2 n \wedge \sigma_k^{-2} n^{-1})^2 \gamma_k^2 = o(1).$$

The proof starts similarly as in [4]. One difference is that we will not take the indicator that $U\gamma_n$ is smaller than some quantity. Indeed, here the main point is that $U\gamma_n$ is one of the terms responsible for the "extra bias phenomenon", when it exists. To be able to see this, it is thus important to let $U\gamma_n$ vary freely.

Now, condition $(\mathbf{C_1})$ enables one to restrict the study of the posterior $\Pi(\cdot \,|\, X)$ to a neighborhood of the true $\eta_0 = (\theta_0, f_0)$ of some size ε_n in terms of $\|\cdot\|_L$. Similarly to [4], setting $V_n = \{(\theta, f) \in \Theta \times \mathcal{F}_n,\ \|\theta - \theta_0, f - f_0\|_L \leq \varepsilon_n\}$, one first shows that, due to $(\mathbf{C_1})$, it is enough to focus on the posterior restricted to V_n that is $\Pi^{V_n}(\cdot \,|\, X)$. Then, we directly have the bounds for any measurable $B \subset \Theta$,

$$\frac{P_2(B)}{Q_2} \leq \Pi^{V_n}(B \,|\, X^{(n)}) \leq \frac{P_1(B)}{Q_1}, \tag{7}$$

where the explicit expression of $P_1(B)$ is given by

$$P_1(B) = \int_B \mathbf{1}_{\tilde{I}(\theta-\theta_0)^2 \le \varepsilon_n^2} \left[\int_{\mathcal{F}_n} \mathbf{1}_{\|0, f-f_0+(\theta-\theta_0)\gamma_n\|_L \le 2\varepsilon_n} e^{\ell_n(\eta)-\ell_n(\eta_0)} d\pi_f(f) \right] d\pi_\theta(\theta),$$

and similarly for $Q_1, P_2(B)$ and Q_2, with slightly different constants in front of ε_n's.

Now let us expand $\ell_n(\eta) - \ell_n(\eta_0)$ using the LAN-type expansion (2)

$$\ell_n(\theta, f) - \ell_n(\theta_0, f_0) = \Delta\ell_n^{(1)}(\theta) + \Delta\ell_n^{(2)}(\theta, f),$$

where $\Delta\ell_n^{(1)}(\theta)$ is the parametric part and $\Delta\ell_n^{(2)}(\theta, f)$ contains the terms depending on f,

$$\begin{aligned}
\Delta\ell_n^{(1)}(\theta) &= -n\tilde{I}(\theta-\theta_0)^2/2 + \sqrt{n}(\theta-\theta_0)\mathcal{W}(1,-\gamma). \\
\Delta\ell_n^{(2)}(\theta, f) &= -n\|0, f + (\theta-\theta_0)\gamma - f_0\|_L^2/2 + \sqrt{n}\mathcal{W}(0, f+(\theta-\theta_0)\gamma - f_0) + R_n(\theta, f).
\end{aligned}$$

The first term factorizes from the integral with respect to f, while the second term can be decomposed making appear the sequence γ_n

$$\begin{aligned}
\Delta\ell_n^{(2)}(\theta, f) = &-n\|0, f+(\theta-\theta_0)\gamma_n - f_0\|_L^2/2 + \sqrt{n}\mathcal{W}(0, f+(\theta-\theta_0)\gamma_n - f_0) && \text{(i)} \\
&+ R_n(\theta_0, f+(\theta-\theta_0)\gamma_n) + R_n(\theta, f) - R_n(\theta_0, f+(\theta-\theta_0)\gamma_n) && \text{(ii)} \\
&- \sqrt{n}h\langle\ (0, f-f_0+(\theta-\theta_0)\gamma_n)\ ,\ (0, \gamma-\gamma_n)\ \rangle_L && \text{(iii)} \\
&- h^2\|0, \gamma-\gamma_n\|_L^2/2 + h\mathcal{W}(0, \gamma-\gamma_n). && \text{(iv)}
\end{aligned}$$

Note that (iv) is always a $o_P(1+h^2)$ by assumption. Also, we have that (ii) reduces to $R_n(\theta_0, f+(\theta-\theta_0)\gamma_n)$ since the two other terms combine into a $o_P(1+h^2)$ due to $(\mathbf{N_1})$.

From the preceding arguments $\Delta\ell_n^{(2)}(\theta, f)$ is, up to a $o_P(1+h^2)$, equal to a term depending only on $f+(\theta-\theta_0)\gamma_n$. Let us denote it $\Delta\ell_n^{(3)}(f+(\theta-\theta_0)\gamma_n)$. Then

$$\begin{aligned}
\Delta\ell_n^{(3)}(f+(\theta-\theta_0)\gamma_n) = &-n\|0, f+(\theta-\theta_0)\gamma_n - f_0\|_L^2/2 \\
&+ \sqrt{n}\mathcal{W}(0, f+(\theta-\theta_0)\gamma_n - f_0) + R_n(\theta_0, f+(\theta-\theta_0)\gamma_n) \\
&- \sqrt{n}h\langle\ (0, f-f_0+(\theta-\theta_0)\gamma_n)\ ,\ (0, \gamma-\gamma_n)\ \rangle_L.
\end{aligned}$$

This leads to the following upper-bound on $P_1(B)$,

$$P_1(B) \le \int_{B;\ \tilde{I}(\theta-\theta_0)^2 \le \varepsilon_n^2} e^{\Delta\ell_n^{(1)}(\theta)} \underbrace{\left[\int_{\mathcal{F}_n,\ \|0, f+(\theta-\theta_0)\gamma_n - f_0\|_L \le 2\varepsilon_n} e^{\Delta\ell_n^{(3)}(f+(\theta-\theta_0)\gamma_n - f_0)} d\pi_f(f) \right]}_{\mathcal{I}_n(\theta)} d\pi_\theta(\theta).$$

As in [4], since γ_n belongs to $\mathbb{H}$, one can now change variables in the Gaussian measure into brackets by setting $g = f + (\theta-\theta_0)\gamma_n$. Doing so, the bracket $\mathcal{I}_n = \mathcal{I}_n(\theta)$ above becomes

$$\mathcal{I}_n = \int_{g \in \mathcal{F}_n(\theta),\ \|0, g-f_0\|_L \le 2\varepsilon_n} e^{\Delta\ell_n^{(3)}(g-f_0)} \left\{ e^{(\theta-\theta_0)U\gamma_n - (\theta-\theta_0)^2\|\gamma_n\|_{\mathbb{H}}^2/2} \right\} d\pi_f(g).$$

Let us notice that $(\theta-\theta_0)^2\|\gamma_n\|^2_{\mathbb{H}}$ is a $o(h^2)$ with our choice of γ_n. Next, as in [4], let us recognize in some terms defining $\Delta\ell_n^{(3)}$ (actually all except the inner product) a likelihood in the model where $\theta=\theta_0$. Indeed, the difference of log-likelihoods in this model, say $\Delta\ell_n^{\theta=\theta_0}$ equals

$$\begin{aligned}\Delta\ell_n^{\theta=\theta_0} =& \ell_n^{\theta=\theta_0}(g)-\ell_n^{\theta=\theta_0}(f_0)\\ =& -n\|0,g-f_0\|_L^2/2+\sqrt{n}\mathcal{W}(0,g-f_0)+R_n(\theta_0,g).\end{aligned}$$

Therefore, up to a normalizing factor which is independent of θ and f, let us recognize in $\mathcal{I}_n$ a posterior expectation, in the model where $\theta=\theta_0$, with respect to the prior π_f and with observations X. This expectation was introduced in Section 2 with the notation $E_\Pi^{\theta=\theta_0}(\cdot\,|\,X)$. If $\propto$ stands for proportionality up to a constant,

$$\mathcal{I}_n \propto E_\Pi^{\theta=\theta_0}(\mathbf{1}_{g\in\mathcal{F}_n(\theta),\ \|0,g-f_0\|_L\le 2\varepsilon_n} e^{(\theta-\theta_0)U\gamma_n-\sqrt{n}h\langle(0,g-f_0),(0,\gamma-\gamma_n)\rangle_L}\,|\,X).$$

Due to Lemma 2, the indicator can be deleted, up to a negligible term. We can now apply Lemma 1 to obtain, with $h=\sqrt{n}(\theta-\theta_0)$,

$$\mathcal{I}_n(\theta)\propto e^{h\zeta_n^\sigma}e^{o_P(1+h^2)}.$$

This leads to the following bound on $P_1(B)$.

$$P_1(B)\le\int_{B;\ \tilde{I}(\theta-\theta_0)^2\le\varepsilon_n^2} e^{-h^2\tilde{I}/2+h\mathcal{W}(1,-\gamma)+h\zeta_n^\sigma+o_P(1+h^2)}d\pi_\theta(\theta).$$

Now coming back to (7), very similar (upper or lower) bounds can be obtained in exactly the same way as above for each of the terms $P_2(B), Q_1, Q_2$. Provided the Lemmas are proved, this leads to the result, by simple manipulations on the (by now) parametric-type likelihood. □

Lemma 1. *Suppose that* (3) *holds. Then, denoting* $h=\sqrt{n}(\theta-\theta_0)$, *as* $n\to+\infty$,

$$\begin{aligned}&E_\Pi^{\theta=\theta_0}(e^{(\theta-\theta_0)U\gamma_n-\sqrt{n}h\langle(0,f-f_0),(0,\gamma-\gamma_n)\rangle_L}\,|\,X)\\ &=E_\Pi^{\theta=\theta_0}(e^{(\theta-\theta_0)U\gamma_n}\,|\,X)\cdot E_\Pi^{\theta=\theta_0}(e^{-\sqrt{n}h\langle(0,f-f_0),(0,\gamma-\gamma_n)\rangle_L}\,|\,X)\\ &=e^{h\zeta_n^\sigma+o_P(1+h^2)}.\end{aligned}$$

Proof of Lemma 1. We consider the case where K_n is odd, which slightly simplifies the presentation, the case K_n even being analogous, as explained below. We first focus on the term $e^{(\theta-\theta_0)U\gamma_n}$. The explicit expression of Uh for any h in $\mathbb{H}$ is given in Section 2. Let f_k be a shorthand notation for $\langle f,\varepsilon_k\rangle_{L^2}$. Notice that under the prior it is distributed as a Gaussian variable of variance σ_k^2, since $f_k=\sigma_k\nu_k$.

Note that if $\theta=\theta_0$, the model consists in "observing f twice". More precisely, observing the paths (Y,Z) is equivalent to observing the collection of pairs $\int\varepsilon_k(\cdot)dY(\cdot), \int\varepsilon_k(\cdot)dZ(\cdot)$,

any $k \geq 1$, but also to observing the collection $\int \varepsilon_k(\cdot)dY(\cdot), \int \varepsilon_k(\cdot - \theta_0)dZ(\cdot)$; $k \geq 1$. Denoting by y_k, z_k this last pair, we have

$$y_k = f_k + \frac{1}{\sqrt{n}}\zeta_k^{(1)} \qquad \text{and} \qquad z_k = f_k + \frac{1}{\sqrt{n}}\zeta_k^{(2)},$$

where $\zeta_k^{(1)}, \zeta_k^{(2)}$ are independent standard normal. From this it is fairly direct to see that the posterior distribution of f given X is the product of the posteriors $f_k \,|\, y_k, z_k$. The later are normal by conjugacy of the normal prior under $E_\Pi^{\theta=\theta_0}$ and simple calculations lead to

$$f_k \,|\, X \underset{\text{under } E_\Pi^{\theta=\theta_0}}{\overset{\mathcal{L}}{=}} N\left(\frac{n}{2n+\sigma_k^{-2}}(y_k + z_k), \frac{1}{2n+\sigma_k^{-2}}\right).$$

The posterior expectation of $\exp((\theta-\theta_0)U\gamma_n)$ is now computed using the expression of $U\gamma_n$ together with the simple identity, for any V of law $N(a, \sigma^2)$, $\log \mathbf{E}(e^{\mu V}) = \mu a + \mu^2\sigma^2/2$. For compactness in the following formulas, let us set $\psi_k(\theta) = (\theta - \theta_0)\sigma_k^{-2}\gamma_{n,k}$. We have, using the preceding identities, that the posterior at stake equals

$$\begin{aligned}
\prod_{k\leq K_n} E_\Pi^{\theta=\theta_0}(e^{\psi_k(\theta)f_k} \,|\, X) &= \prod_{k\leq K_n} \exp\left[\frac{n\psi_k(\theta)}{2n+\sigma_k^{-2}}(y_k + z_k) + \frac{\psi_k(\theta)^2}{2(2n+\sigma_k^{-2})}\right] \\
&= \prod_{k\leq K_n} \exp\left[\psi_k(\theta) f_{0,k}\frac{2n}{2n+\sigma_k^{-2}}\right] \qquad \text{(a)} \\
&\quad \times \exp\left[\psi_k(\theta)\frac{2\sqrt{n}}{2n+\sigma_k^{-2}}(\zeta_k^{(1)} + \zeta_k^{(2)})\right] \qquad \text{(b)} \\
&\quad \times \exp\left[\frac{\psi_k(\theta)^2}{2(2n+\sigma_k^{-2})}\right] \qquad \text{(c)}
\end{aligned}$$

The main term turns out to be (a). We explicit it next, splitting the sum in even and odd k's. Since K_n is odd, one can write $K_n = 2M_n + 1$, for some integer $M_n \geq 1$.

$$\begin{aligned}
\text{(a)} &= \exp\left[\pi(\theta-\theta_0)\sum_{p\leq M_n}\left[-\sigma_{2p}^{-2} p f_{0,2p+1} f_{0,2p}\left(\frac{2n}{2n+\sigma_{2p}^{-2}}\right) + \sigma_{2p+1}^{-2} p f_{0,2p} f_{0,2p+1}\left(\frac{2n}{2n+\sigma_{2p+1}^{-2}}\right)\right]\right] \\
&= \exp\left[\pi(\theta-\theta_0)\sum_{p\leq M_n} p f_{0,2p+1} f_{0,2p}(\sigma_{2p+1}^{-2} - \sigma_{2p}^{-2})\left\{\frac{(2n)^2}{(2n+\sigma_{2p}^{-2})(2n+\sigma_{2p+1}^{-2})}\right\}\right]
\end{aligned}$$

Turning to the study of (b) and (c),

$$\begin{aligned}
\mathbf{E}_{\eta_0}(\text{b}) &= \exp\left[\sum_{k\leq K_n}\frac{\psi_k(\theta)^2}{2n}\left(\frac{n}{2n+\sigma_k^{-2}}\right)^2\right] \\
(\text{c}) &= \exp\left[\sum_{k\leq K_n}\frac{\psi_k(\theta)^2}{2n}\frac{n}{2n+\sigma_k^{-2}}\right].
\end{aligned}$$

Condition (3) and the definition of K_n directly imply $\mathbf{E}_{\eta_0}(\mathrm{b}) + (\mathrm{c}) = \exp(o(1+h^2))$.

Now focusing on the second term, note that the expression $\langle (0,f_0),(0,\gamma-\gamma_n)\rangle_L$ equals $2\int_0^1 f_0(\gamma-\gamma_n)$. But, since f_0 is a $\mathcal{C}^1$, 1-periodic function by assumption,

$$\int_0^1 f_0(u)\gamma(u)du = -\int_0^1 f_0(u)f_0'(u)/2 = 0.$$

Since K_n is odd, the same argument shows that $\int_0^1 f_0\gamma_n$ is zero. Thus we focus on $\langle (0,f),(0,\gamma-\gamma_n)\rangle_L = 2\int f(\gamma-\gamma_n)$. Let us set $w_k(\theta) = -2(\theta-\theta_0)n\gamma_k$, any $k > K_n$. Similar to the case $k \le K_n$,

$$\begin{aligned}
\prod_{k>K_n} E_\Pi^{\theta=\theta_0}(e^{w_k(\theta)f_k} \mid X) &= \prod_{k>K_n} \exp\left[\frac{nw_k(\theta)}{2n+\sigma_k^{-2}}(y_k+z_k) + \frac{w_k(\theta)^2}{2(2n+\sigma_k^{-2})}\right] \\
&= \prod_{k>K_n} \exp\left[w_k(\theta)f_{0,k}\frac{2n}{2n+\sigma_k^{-2}}\right] && \text{(a')} \\
&\quad \times \exp\left[w_k(\theta)\frac{2\sqrt{n}}{2n+\sigma_k^{-2}}(\zeta_k^{(1)}+\zeta_k^{(2)})\right] && \text{(b')} \\
&\quad \times \exp\left[\frac{w_k(\theta)^2}{2(2n+\sigma_k^{-2})}\right] && \text{(c')}
\end{aligned}$$

Similarly as for (a)-(b), using (3) one verifies that $\mathbf{E}_{\eta_0}$(b') and (c') are $\exp(o(1+h^2))$. To write down (a'), we split the sum along even and odd $k's$,

$$\begin{aligned}
(\mathrm{a}') &= \exp\left[2\pi(\theta-\theta_0)n\sum_{p>M_n}\left[pf_{0,2p+1}f_{0,2p}\left(\frac{2n}{2n+\sigma_{2p}^{-2}}\right) - pf_{0,2p}f_{0,2p+1}\left(\frac{2n}{2n+\sigma_{2p+1}^{-2}}\right)\right]\right] \\
&= \exp\left[\pi(\theta-\theta_0)(2n)^2\sum_{p>M_n} pf_{0,2p+1}f_{0,2p}\left\{\frac{\sigma_{2p+1}^{-2}-\sigma_{2p}^{-2}}{(2n+\sigma_{2p}^{-2})(2n+\sigma_{2p+1}^{-2})}\right\}\right].
\end{aligned}$$

Regrouping (a) and (a') concludes the proof in the case of odd K_n. When K_n is even, the argument is similar, except at the split point $k = K_n$. In that case one uses the fact that an extra contribution comes from $-2\int_0^1 f_0\gamma_n = 2M_n f_{0,K_n}f_{0,K_n+1}$ to regroup the terms for $k = K_n$ and $k = K_n+1$, leading to the same expression of the leading terms. □

Lemma 2. *Suppose that* (3) *holds. Then, as* $n \to +\infty$,

$$\begin{aligned}
E_\Pi^{\theta=\theta_0}(\mathbf{1}_{g\in\mathcal{F}_n(\theta),\ \|0,g-f_0\|_L\le 2\varepsilon_n} e^{\left\{(\theta-\theta_0)U\gamma_n - \sqrt{n}h\langle (0,g-f_0),(0,\gamma-\gamma_n)\rangle_L\right\}} \mid X) \\
= (1+o_P(1))e^{h\zeta_n^\sigma + o_P(1+h^2)}.
\end{aligned}$$

Proof. Let $\lambda_n(\theta,g)$ be the term into the braces. Applying Cauchy-Schwarz inequality,

$$\begin{aligned}
&E_\Pi^{\theta=\theta_0}(\left[1-\mathbf{1}_{g\in\mathcal{F}_n(\theta),\ \|0,g-f_0\|_L\le 2\varepsilon_n}\right] e^{\lambda_n(\theta,g)} \mid X) \\
&\quad \le E_\Pi^{\theta=\theta_0}(\left[1-\mathbf{1}_{g\in\mathcal{F}_n(\theta),\ \|0,g-f_0\|_L\le 2\varepsilon_n}\right] \mid X)^{1/2} E_\Pi^{\theta=\theta_0}(e^{2\lambda_n(\theta,g)} \mid X)^{1/2} \\
&\quad \le o_P(1)E_\Pi^{\theta=\theta_0}(e^{2\lambda_n(\theta,g)} \mid X)^{1/2}.
\end{aligned}$$

The bound by $o_P(1)$ results from the second part of assumption $(\mathbf{N_1})$. The last square root can be computed exactly in the same way as in Lemma 1. In fact, we can notice that the leading term in the computations in the proof of Lemma 1 is linear in $\theta - \theta_0$ (for both cases $k > K_n$ and $k \leq K_n$), so we have that

$$E_\Pi^{\theta=\theta_0}(e^{2\lambda_n(\theta,g)} \mid X)^{1/2} = E_\Pi^{\theta=\theta_0}(e^{\lambda_n(\theta,g)} \mid X) e^{o_P(1+h^2)}(1 + o_P(1)).$$

This leads to the assertion of the Lemma. □

4 Appendix, checking conditions $(\mathbf{C_1})$ - $(\mathbf{N_1})$

Let us check that the priors $\Pi^\alpha, \Pi^{\alpha,*}$ verify conditions $(\mathbf{C_1})$, $(\mathbf{N_1})$ for some rate ε_n, in some domain of values of the regularity parameters (α, β). The arguments are very similar to the ones used in [4] for the translation parameter estimation, [4], eq. (9). In fact, here we obtain the same set of parameters (α, β) for which $(\mathbf{C_1})$, $(\mathbf{N_1})$ are satified, see [4], Fig. 1.

First, we check the concentration condition $(\mathbf{C_1})$ following the approach in [12]. The first step is to show a concentration in terms of a distance for which tests with exponential decrease exist. Given the true parameter (θ_0, f_0) and another parameter (θ_1, f_1), let us set

$$\phi_n = \mathbf{1}\{2\int_0^1 (f_1 - f_0)(t - \theta_0)dY(t) + 2\int_0^1 \{f_1(t - \theta_1) - f_0(t - \theta_0)\}dZ(t) > \|f_1\|^2 - \|f_0\|^2\}.$$

Simple calculations analogous to Lemma 5 in [12] show that this test enables to test the true (θ_0, f_0) versus a ball with appropriate exponential decrease of the error probabilities, see [4], eq. (4) or [12], eq. (2.2). The corresponding testing distance d_T is given by

$$d_T((\theta_1, f_1), (\theta_2, f_2))^2 = \|f_1 - f_2\|^2 + \|f_1(\cdot - \theta_1) - f_2(\cdot - \theta_2)\|^2,$$

One then relates d_T^2 to the squared-distance $\|f_1 - f_2\|_2^2 + (\theta_1 - \theta_2)^2$. This is easily done by adapting Lemma 4 in [4] to the case of not-necessarily symmetric f. Once those distances are related, the verifications of the entropy and prior mass conditions are done exactly as in [4], Section 4.1.1, thus leading to $(\mathbf{C_1})$. One also verifies that the rate ε_n can be taken proportional to $n^{-\alpha\wedge\beta/(2\alpha+1)}$.

Now we check $(\mathbf{N_1})$. The term $R_n(\theta_0, f + (\theta - \theta_0)\gamma)$ is zero so one focuses on $R_n(\theta, f)$. We first introduce a sieve $\mathcal{F}_n$ on which it is possible to restrict the supremum in condition $(\mathbf{N_1})$. Let us introduce the Hilbert space of functions

$$\mathbb{B}^p = \{f = \sum_{k\geq 1} f_k \varepsilon_k(\cdot), \quad \sum_{k\geq 1} k^{2p} f_k^2 < +\infty\}, \qquad p \geq 1,$$

equipped with the norm $\|f\|_{2,p}^2 = \sum_{k\geq 1} k^{2p} f_k^2$. The idea is to use Borell's inequality in the form of [24], Theorem 5.1. This result exactly tells us that overwhelming probability, the Gaussian prior (either Π^α or $\Pi^{\alpha,*}$) draws functions g which can be written

$$g = \varepsilon_n v_0 + \sqrt{n}\varepsilon_n w_0, \quad \text{with } v_0 \in \mathbb{B}_1^0,\ w \in \mathbb{H}_1^\alpha, \tag{8}$$

but also, for $1 \leq p < \alpha$, and some rate $\alpha_n \to 0$ to be specified,

$$g = \alpha_n v + \sqrt{n}\alpha_n w, \quad \text{with } v \in \mathbb{B}_1^p,\ w \in \mathbb{H}_1^\alpha, \tag{9}$$

where $\mathbb{H}_1^\alpha$ denotes the unit ball of the RKHS of the prior (we use the same notation $\mathbb{H}_1^\alpha$ for Π^α and $\Pi^{\alpha,*}$ though the corresponding spaces differ slightly) and $\mathbb{B}_1^p$ the unit ball of the space $\mathbb{B}^p$. As in [4], one can then define a sieve $\mathcal{F}_n$ as the intersection of the set of functions defined by (8) and (9). Under some conditions on α_n, Borell's inequality implies that the complement $\mathcal{F} \setminus \mathcal{F}_n$ has probability less than $\exp(-n\varepsilon_n^2)$, see [4], Lemma 13. Thus it is possible to restrict the study of the posterior (and of $(\mathbf{N_1})$) to $\mathcal{F}_n$.

We first deal with the deterministic terms $R_{n,3}, R_{n,4}$. To control $R_{n,4}$, it is enough to bound from above separately $\int (a_f(t-\theta) - a_f(t-\theta_0))^2 dt$ and $\int D_n(t,h)^2 dt$. This last term can be bounded as in [4], Lemma 5 (adapting slightly the proof to accomodate to not necessarily symmetric functions f), leading to a bound in $o(1+h^2)$. The first term is bounded using the decomposition (9) in the form $f = \alpha_n v + w_n$, with $\|w_n\|_{\mathbb{H}_1^\alpha}^2 \leq n\alpha_n^2$,

$$\begin{aligned}
&\int_0^1 (a_f(t-\theta) - a_f(t-\theta_0))^2 dt \\
&\lesssim\ n\alpha_n^2 \int_0^1 (v(t-\theta) - v(t-\theta_0))^2 dt + n\int_0^1 ((w_n - f_0)(t-\theta) - (w_n - f_0)(t-\theta_0))^2 dt.
\end{aligned}$$

The bounds on the respective variances have been derived in [4], see the bounds to (22)-(23). The first term is a $O(\alpha_n^2 h^2)$ and the second is a $O((1+h^2)\alpha_n^2 n^{2/(1+2\alpha)})$. Thus both are $o(1+h^2)$ provided that $\alpha_n = o(n^{-1/(1+2\alpha)})$.

To bound $R_{n,3}$, we develop the product and bound again each term separately. One resulting term is $\int (a_f - hf_0')(t-\theta_0) D_n(t,h) dt$ and, similar to Lemma 6 in [4], is a $o(1+h^2)$ as soon as $\varepsilon_n = o(n^{-1+\beta/2})$. Another term is $h\int f_0'(t-\theta_0)(a_f(t-\theta) - a_f(t-\theta_0))dt$. Using Cauchy-Schwarz inequality, we can re-use the bound of the previous display. The last term to bound is $\int a_f(t-\theta_0)(a_f(t-\theta) - a_f(t-\theta_0))dt$. First we notice that, for any w in $L^2[0,1]$ 1-periodic of Fourier coefficients w_k, expanding the function on the Fourier basis,

$$\int_0^1 w(t-\theta_0)(w(t-\theta_0) - w(t-\theta))dt = \sum_{k\geq 1} \sin^2(\pi k(\theta - \theta_0))(w_{2k}^2 + w_{2k+1}^2).$$

Applying this to the function $a_f = \sqrt{n}(f - f_0)$ and using the inequality $\sin(x) \leq x$ enables us to bound the quantity at stake by a constant times $h^2 \sum_{k\geq 1} k^2 (f_{0,k} - f_k)^2$. We split this sum along indexes $k \leq k(n)$ and $k > k(n)$, with $k(n) = \lfloor n^{1/(1+2\alpha)} \rfloor$. The sum up to $k(n)$ leads to the bound $h^2 k(n)^2 \|f - f_0\|^2 \leq h^2 k(n)^2 \varepsilon_n^2$. Due to the expressions of $k(n)$ and ε_n, this is a $o(h^2)$ when $\alpha \wedge \beta \geq 1$. The sum for $k > k(n)$ is bounded noticing that $\sum_{k>k(n)} k^2 f_{0,k}^2 = o(1)$ since $\beta > 1$ and using the decomposition (9) as follows

$$\sum_{k>k(n)} k^2 f_k^2 \leq \alpha_n^2 \sum_{k>k(n)} k^2 v_k^2 + n\alpha_n^2 \sum_{k>k(n)} k^2 w_k^2.$$

Since $v \in \mathbb{B}_1^p$ with $p > 1$ the first term is a $o(\alpha_n^2)$. Since $w \in \mathbb{H}_1^\alpha$, we have that

$$\sum_{k>k(n)} k^2 w_k^2 \leq k(n)^{1-2\alpha} \sum_{k>k(n)} k^{1+2\alpha} w_k^2 = o(k(n)^{1-2\alpha}).$$

By definition of $k(n)$, we conclude that the term at stake is a $o(h^2)$ if $n^{2/(1+2\alpha)}\alpha_n^2 = o(1)$. The stochastic terms $R_{n,1}$ and $R_{n,2}$ are exactly the same (up to the symmetry assumption on f, which does not change the proofs) as in [4], see eq. (16)-(18), so we can borrow the proofs.

The imposed conditions on ε_n, α_n found above are the same as in [4], section 4.1.3, where it is checked that those are satisfied as soon as $\alpha > 1 + \sqrt{3}/2$, $\beta > 3/2$ and, if $\beta < 2 \wedge \alpha$, also $\alpha < (3\beta - 2)/(4 - 2\beta)$. This is the zone depicted in [4], Fig. 1. It includes in particular the rectangle $\alpha > 1 + \sqrt{3}/2$, $\beta \geq 2$, where $(\mathbf{N_1})$ is therefore satisfied. □

Acknowledgments. The author acknowledges the hospitality of the Statistics departements of the Vrije Universiteit Amsterdam and TU Eindhoven/Eurandom for a two weeks stay during the preparation of this work. The author would also like to thank Dominique Picard for an insightful comment.

References

[1] P. Bickel and B. Kleijn. The semiparametric Bernstein-von Mises theorem. Preprint.

[2] S. Boucheron and E. Gassiat. A Bernstein-von Mises theorem for discrete probability distributions. Electron. J. Stat., 3:114–148, 2009.

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

[4] I. Castillo. A semiparametric Bernstein-von Mises theorem for Gaussian process priors. Probability Theory and Related Fields, 2011. To appear.

[5] I. Castillo and E. Cator. Semiparametric shift estimation based on the cumulated periodogram for non-regular functions. Electron. J. Statist., 5:102–126, 2011.

[6] D. D. Cox. An analysis of Bayesian inference for nonparametric regression. Ann. Statist., 21(2):903–923, 1993.

[7] P. Diaconis and D. Freedman. On the consistency of Bayes estimates. Ann. Statist., 14(1):1–67, 1986. With a discussion and a rejoinder by the authors.

[8] D. Freedman. On the Bernstein-von Mises theorem with infinite-dimensional parameters. Ann. Statist., 27(4):1119–1140, 1999.

[9] F. Gamboa, J.-M. Loubes, and E. Maza. Semi-parametric estimation of shifts. Electron. J. Stat., 1:616–640, 2007.

[10] 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.

[11] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. Ann. Statist., 28(2):500–531, 2000.

[12] S. Ghosal and A. W. van der Vaart. Convergence rates of posterior distributions for noniid observations. Ann. Statist., 35(1), 2007.

[13] E. Giné and R. Nickl. On the uniform consistency of nonparametric Bayes estimates. Preprint.

[14] I. A. Ibragimov and R. Z. Has′minskiĭ. Statistical estimation, volume 16 of Applications of Mathematics. Springer-Verlag, New York, 1981.

[15] I. Johnstone. High dimensional Bernstein-von Mises: simple examples. In Festschrift for Lawrence D. Brown, volume 6 of Inst. Math. Stat. Collect., pages 87–98. 2010.

[16] Y. Kim. The Bernstein-von Mises theorem for the proportional hazard model. Ann. Statist., 34(4):1678–1700, 2006.

[17] B. McNeney and J. A. Wellner. Application of convolution theorems in semiparametric models with non-i.i.d. data. J. Statist. Plann. Inference, 91(2):441–480, 2000.

[18] J. Rousseau and V. Rivoirard. Bernstein-von Mises theorem for linear functionals of the density. Preprint.

[19] X. Shen. Asymptotic normality of semiparametric and nonparametric posterior distributions. J. Amer. Statist. Assoc., 97(457):222–235, 2002.

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

[21] A. van der Vaart. The statistical work of Lucien Le Cam. Ann. Statist., 30(3):631–682, 2002. Dedicated to the memory of Lucien Le Cam.

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

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

[24] A. W. van der Vaart and H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. IMS Collections, 3:200–222, 2008.

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

[26] Y. Wu and S. Ghosal. Posterior consistency for some semi-parametric problems. Sankhyā, 70(2, Ser. A):267–313, 2008.