

Bayesian optimal adaptive estimation using a sieve prior

Julyan Arbel, Ghislaine Gayraud, Judith Rousseau

▶ **To cite this version:**

Julyan Arbel, Ghislaine Gayraud, Judith Rousseau. Bayesian optimal adaptive estimation using a sieve prior. Scandinavian Journal of Statistics, 2013, 10.1002/sjos.12002 . hal-01203280

HAL Id: hal-01203280
https://hal.science/hal-01203280

Submitted on 24 Sep 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:1204.2392v2 [math.ST] 18 Feb 2013

Bayesian optimal adaptive estimation using a sieve prior

Julyan Arbel[1,2], *Ghislaine Gayraud*[1,3] *and Judith Rousseau*[1,4]

[1] *Laboratoire de Statistique, CREST, France*

[2] *INSEE, France*

[3] *LMAC, Université de Technologie de Compiègne, France*

[4] *Université Paris Dauphine, France*

Abstract

We derive rates of contraction of posterior distributions on nonparametric models resulting from sieve priors. The aim of the paper is to provide general conditions to get posterior rates when the parameter space has a general structure, and rate adaptation when the parameter space is, *e.g.*, a Sobolev class. The conditions employed, although standard in the literature, are combined in a different way. The results are applied to density, regression, nonlinear autoregression and Gaussian white noise models. In the latter we have also considered a loss function which is different from the usual l^2 norm, namely the pointwise loss. In this case it is possible to prove that the adaptive Bayesian approach for the l^2 loss is strongly suboptimal and we provide a lower bound on the rate.

Keywords adaptation, minimax criteria, nonparametric models, rate of contraction, sieve prior, white noise model.

1 Introduction

The asymptotic behaviour of posterior distributions in nonparametric models has received growing consideration in the literature over the last ten years.

Many different models have been considered, ranging from the problem of density estimation in i.i.d. models (Barron et al., 1999; Ghosal et al., 2000), to sophisticated dependent models (Rousseau et al., 2012). For these models, different families of priors have also been considered, where the most common are Dirichlet process mixtures (or related priors), Gaussian processes (van der Vaart and van Zanten, 2008), or series expansions on a basis (such as wavelets, see Abramovich et al., 1998).

In this paper we focus on a family of priors called *sieve priors*, introduced as *compound priors* and discussed by Zhao (1993, 2000), and further studied by Shen and Wasserman (2001). It is defined for models $(\mathcal{X}^{(n)}, A^{(n)}, P_{\boldsymbol{\theta}}^{(n)} : \boldsymbol{\theta} \in \Theta)$, $n \in \mathbb{N}\backslash\{0\}$, where $\Theta \subseteq \mathbb{R}^{\mathbb{N}}$, the set of sequences. Let A be a σ-field associated to Θ. The observations are denoted X^n, where the asymptotics are driven by n. The probability measures $P_{\boldsymbol{\theta}}^{(n)}$ are dominated by some reference measure μ, with density $p_{\boldsymbol{\theta}}^{(n)}$. Remark that such an infinite-dimensional parameter $\boldsymbol{\theta}$ can often characterize a functional parameter, or a curve, $\boldsymbol{f} = \boldsymbol{f}_{\boldsymbol{\theta}}$. For instance, in regression, density or spectral density models, $\boldsymbol{f}$ represents a regression function, a log density or a log spectral density respectively, and $\boldsymbol{\theta}$ represents its coordinates in an appropriate basis $\boldsymbol{\psi} = (\psi_j)_{j\geq 1}$ (*e.g.*, a Fourier, a wavelet, a log spline, or an orthonormal basis in general). In this paper we study frequentist properties of the posterior distributions as n tends to infinity, assuming that data X^n are generated by a measure $P_{\boldsymbol{\theta}_0}^{(n)}$, $\boldsymbol{\theta}_0 \in \Theta$. We study in particular rates of contraction of the posterior distribution and rates of convergence of the risk.

A sieve prior Π is expressed as

$$\boldsymbol{\theta} \sim \Pi(\,\cdot\,) = \sum_{k=1}^{\infty} \pi(k)\Pi_k(\,\cdot\,), \tag{1}$$

where $\sum_k \pi(k) = 1$, $\pi(k) \geq 0$, and the Π_k's are prior distributions on so-called sieve spaces $\Theta_k = \mathbb{R}^k$. Set $\boldsymbol{\theta}_k = (\theta_1, \ldots, \theta_k)$ the finite-dimensional vector of the first k entries of $\boldsymbol{\theta}$. Essentially, the whole prior Π is seen as a hierarchical prior, see Figure 1. The hierarchical parameter k, called threshold parameter, has prior π. Conditionally on k, the prior on $\boldsymbol{\theta}$ is Π_k which is supposed to have mass only on Θ_k (this amounts to say that the priors on the remaining entries θ_j, $j > k$, are point masses at 0). We assume that Π_k is an independent prior on the coordinates θ_j, $j = 1, \ldots, k$, of $\boldsymbol{\theta}_k$ with a unique probability density g once rescaled by positive $\boldsymbol{\tau} = (\tau_j)_{j\geq 1}$. Using the same notation Π_k for probability

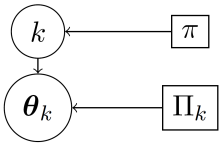


Fig. 1: Graphical representation of the hierarchical structure of the *sieve prior* given by Equation (1)

and density with Lebesgue measure or $\mathbb{R}^k$, we have

$$\forall \boldsymbol{\theta}_k \in \Theta_k, \quad \Pi_k\left(\boldsymbol{\theta}_k\right) = \prod_{j=1}^{k} \frac{1}{\tau_j} g\left(\frac{\theta_j}{\tau_j}\right). \tag{2}$$

Note that the quantities Π, Π_k, π, $\boldsymbol{\tau}$ and g could depend on n. Although not purely Bayesian, data dependent priors are quite common in the literature. For instance, Ghosal and van der Vaart (2007) use a similar prior with a deterministic cutoff $k = \lfloor n^{1/(2\alpha+1)} \rfloor$ in application 7.6.

We will also consider the case where the prior is truncated to an l^1 ball of radius $r_1 > 0$ (see the nonlinear AR(1) model application in Section 2.3.3)

$$\forall \boldsymbol{\theta}_k \in \Theta_k, \quad \Pi_k\left(\boldsymbol{\theta}_k\right) \propto \prod_{j=1}^{k} \frac{1}{\tau_j} g\left(\frac{\theta_j}{\tau_j}\right) \mathbb{I}(\sum_{j=1}^{k} |\theta_j| \leq r_1). \tag{3}$$

The posterior distribution $\Pi(\cdot|X^n)$ is defined by, for all measurable sets B of Θ,

$$\Pi(B|X^n) = \frac{\int_B p_{\boldsymbol{\theta}}^{(n)}(X^n) d\Pi(\boldsymbol{\theta})}{\int_\Theta p_{\boldsymbol{\theta}}^{(n)}(X^n) d\Pi(\boldsymbol{\theta})}. \tag{4}$$

Given the sieve prior Π, we study the rate of contraction of the posterior distribution in $P_{\boldsymbol{\theta}_0}^{(n)}-$probability with respect to a semimetric d_n on Θ. This rate is defined as the best possible (*i.e.* the smallest) sequence $(\epsilon_n)_{n\geq 1}$ such that

$$\Pi\left(\boldsymbol{\theta} :\, d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq M\epsilon_n^2 | X^n\right) \underset{n\to\infty}{\longrightarrow} 0,$$

in $P_{\boldsymbol{\theta}_0}^{(n)}$ probability, for some $\boldsymbol{\theta}_0 \in \Theta$ and a positive constant M, which can be chosen as large as needed. We also derive convergence rates for the posterior loss $\Pi(d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|X^n)$ in $P_{\boldsymbol{\theta}_0}^{(n)}-$probability.

The posterior concentration rate is optimal when it coincides with the minimax rates of convergence, when $\boldsymbol{\theta}_0$ belongs to a given functional class, associated to the same semimetric d_n. Typically these minimax rates of convergence are defined for functional classes indexed by a smoothness parameter Sobolev, Hölder, or more generally Besov spaces.

The objective of this paper is to find mild generic assumptions on the sieve prior Π of the form (1), on models $P_{\boldsymbol{\theta}}^{(n)}$ and on d_n, such that the procedure adapts to the optimal rate in the minimax sense, both for the posterior distribution and for the risk. Results in Bayesian nonparametrics literature about contraction rates are usually of two kinds. Firstly, general assumptions on priors and models allow to derive rates, see for example Shen and Wasserman (2001); Ghosal et al. (2000); Ghosal and van der Vaart (2007). Secondly, other papers focus on a particular prior and obtain contraction rates in a particular model, see for instance Belitser and Ghosal (2003) in the white noise model, De Jonge and van Zanten (2010) in regression, and Scricciolo (2006) in density. The novelty of this paper is that our results hold for a family of priors (sieve priors) without a specific underlying model, and can be applied to different models.

An additional interesting property that is sought at the same time as convergence rates is adaptation. This means that, once specified a loss function (a semimetric d_n on Θ), and a collection of classes of different smoothnesses for the parameter, one constructs a procedure which is independent of the smoothness, but which is rate optimal (under the given loss d_n), within each class. Indeed, the optimal rate naturally depends on the smoothness of the parameter, and standard straightforward estimation techniques usually use it as an input. This is all the more an important issue that relatively few instances in the Bayesian literature are available in this area. That property is often obtained when the unknown parameter is assumed to belong to a discrete set, see for example Belitser and Ghosal (2003). There exist some results in the context of density estimation by Huang (2004), Scricciolo (2006), Ghosal et al. (2008), van der Vaart and van Zanten (2009), Rivoirard and Rousseau (2012a), Rousseau (2010) and Kruijer et al. (2010), in regression by De Jonge and van Zanten (2010), and in spectral density estimation by Rousseau and Kruijer (2011). What enables adaptation in our results is the thresholding induced by the prior on k: the posterior distribution of parameter k concentrates around values that are the typical efficient size of models of the true smoothness.

As seen from our assumptions in Section 2.1 and from the general results (The-

orem 1 and Corollary 1), adaptation is relatively straightforward under sieve priors defined by (1) when the semimetric is a global loss function which acts like the Kullback-Leibler divergence, the l^2 norm on $\boldsymbol{\theta}$ in the regression problem, or the Hellinger distance in the density problem. If the loss function (or the semimetric) d_n acts differently, then the posterior distribution (or the risk) can be quite different (suboptimal). This is illustrated in Section 3.2 for the white noise model (16) when the loss is a local loss function as in the case of the estimation of $\boldsymbol{f}(t)$, for a given t, where $d_n(\boldsymbol{f}, \boldsymbol{f}_0) = (\boldsymbol{f}(t) - \boldsymbol{f}_0(t))^2$. This phenomenon has been encountered also by Rousseau and Kruijer (2011). It is not merely a Bayesian issue: Cai et al. (2007) show that an optimal estimator under global loss cannot be locally optimal at each point $\boldsymbol{f}(t)$ in the white noise model. The penalty between global and local rates is at least a $\log n$ term. Abramovich et al. (2004) and Abramovich et al. (2007a) obtain similar results with Bayesian wavelet estimators in the same model.

The paper is organized as follows. Section 2 first provides a general result on rates of contraction for the posterior distribution in the setting of sieve priors. We also derive a result in terms of posterior loss, and show that the rates are adaptive optimal for Sobolev smoothness classes. The section ends up with applications to the density, the regression function and the nonlinear autoregression function estimation. In Section 3, we study more precisely the case of the white noise model, which is a benchmark model. We study in detail the difference between global or pointwise losses in this model, and provide a lower bound for the latter loss, showing that sieve priors lead to suboptimal contraction rates. Proofs are deferred to the Appendix.

Notations

We use the following notations. Vectors are written in bold letters, for example $\boldsymbol{\theta}$ or $\boldsymbol{\theta}_0$, while light-face is used for their entries, like θ_j or θ_{0j}. We denote by $\boldsymbol{\theta}_{0k}$ the projection of $\boldsymbol{\theta}_0$ on its first k coordinates, and by $p_{0k}^{(n)}$ and $p_0^{(n)}$ the densities of the observations in the corresponding models. We denote by d_n a semimetric, by $\|\cdot\|_2$ the l^2 norm (on vectors) in Θ or the L^2 norm (on curves $\boldsymbol{f}$), and by $\|\cdot\|_{2,k}$ the l^2 norm restricted to the first k coordinates of a parameter. Expectations $\mathbb{E}_0^{(n)}$ and $\mathbb{E}_{\boldsymbol{\theta}}^{(n)}$ are defined with respect to $P_{\boldsymbol{\theta}_0}^{(n)}$ and $P_{\boldsymbol{\theta}}^{(n)}$ respectively. The same notation $\Pi(\cdot|X^n)$ is used for posterior probability or posterior expectation. The expected posterior risk and the frequentist risk

relative to d_n are defined and denoted by $\mathcal{R}_n^{d_n}(\boldsymbol{\theta}_0) = \mathbb{E}_0^{(n)}\Pi(d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|X^n)$ and $R_n^{d_n}(\boldsymbol{\theta}_0) = \mathbb{E}_0^{(n)}(d_n^2(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0))$ respectively (for an estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$), where the mention of $\boldsymbol{\theta}_0$ might be omitted (*cf.* Robert, 2007, Section 2.3). We denote by φ the standard Gaussian probability density.

Let K denote the Kullback-Leibler divergence $K(f,g) = \int f \log(f/g) d\mu$, and $V_{m,0}$ denote the m^{th} centered moment $V_{m,0}(f,g) = \int f |\log(f/g) - K(f,g)|^m$ $d\mu$, with $m \geq 2$.

Define two additional divergences $\widetilde{K}$ and $\widetilde{V}_{m,0}$, which are expectations with respect to $p_0^{(n)}$, $\widetilde{K}(f,g) = \int p_0^{(n)} |\log(f/g)| d\mu$ and $\widetilde{V}_{m,0}(f,g) = \int p_0^{(n)} |\log(f/g) - K(f,g)|^m d\mu$.

We denote by C a generic constant whose value is of no importance and we use $\lesssim$ for inequalities up to a multiple constant.

2 General case

In this section we give a general theorem which provides an upper bound on posterior contraction rates ϵ_n. Throughout the section, we assume that the sequence of positive numbers $(\epsilon_n)_{n\geq 1}$, or $(\epsilon_n(\beta))_{n\geq 1}$ when we point to a specific value of smoothness β, is such that $\epsilon_n \underset{n\to\infty}{\longrightarrow} 0$ and $n\epsilon_n^2/\log n \underset{n\to\infty}{\longrightarrow} \infty$.

We introduce the following numbers

$$j_n = \lfloor j_0 n\epsilon_n^2/\log(n) \rfloor,\ k_n = \lfloor M_0 j_n \log(n)/L(n) \rfloor, \tag{5}$$

for $j_0 > 0, M_0 > 1$, where L is a slow varying function such that $L \leq \log$, hence $j_n \leq k_n$. We use k_n to define the following approximation subsets of Θ

$$\Theta_{k_n}(Q) = \left\{ \boldsymbol{\theta} \in \Theta_{k_n} : \|\boldsymbol{\theta}\|_{2,k_n} \leq n^Q \right\},$$

for $Q > 0$. Note that the prior actually charges a union of spaces of dimension k, $k \geq 1$, so that $\Theta_{k_n}(Q)$ can be seen as a union of spaces of dimension $k \leq k_n$. Lemma 2 provides an upper bound on the prior mass of $\Theta_{k_n}(Q)$.

It has been shown (Ghosal et al., 2000; Ghosal and van der Vaart, 2007; Shen and Wasserman, 2001) that an efficient way to derive rates of contraction of posterior distributions is to bound from above the numerator of (4) using tests (and k_n for the increasing sequence $\Theta_{k_n}(Q)$), and to bound from below its denominator

using an approximation of $p_0^{(n)}$ based on a value $\boldsymbol{\theta} \in \Theta_{j_n}$ close to $\boldsymbol{\theta}$. The latter is done in Lemma 3 where we use j_n to define the finite component approximation $\boldsymbol{\theta}_{0j_n}$ of $\boldsymbol{\theta}_0$, and we show that the prior mass of the following Kullback-Leibler neighbourhoods of $\boldsymbol{\theta}_0$, $\mathcal{B}_n(m)$, $n \in \mathbb{N}^*$, are lower bounded by an exponential term:

$$\mathcal{B}_n(m) = \left\{ \boldsymbol{\theta} : K\left(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) \leq 2n\epsilon_n^2, V_{m,0}\left(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) \leq 2^{m+1}(n\epsilon_n^2)^{m/2} \right\}.$$

Define two neighbourhoods of $\boldsymbol{\theta}_0$ in the sieve space Θ_{j_n}, $\widetilde{\mathcal{B}}_n(m)$, similar to $\mathcal{B}_n(m)$ but using $\widetilde{K}$ and $\widetilde{V}_{m,0}$, and $\mathcal{A}_n(H_1)$, an l^2 ball of radius n^{-H_1}, $H_1 > 0$:

$$\widetilde{\mathcal{B}}_n(m) = \left\{ \boldsymbol{\theta} \in \Theta_{j_n} : \widetilde{K}\left(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) \leq n\epsilon_n^2, \widetilde{V}_{m,0}\left(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) \leq \left(n\epsilon_n^2\right)^{m/2} \right\},$$
$$\mathcal{A}_n(H_1) = \left\{ \boldsymbol{\theta} \in \Theta_{j_n} : \left\|\boldsymbol{\theta}_{0j_n} - \boldsymbol{\theta}\right\|_{2,j_n} \leq n^{-H_1} \right\}.$$

2.1 Assumptions

The following technical assumptions are involved in the subsequent analysis, and are discussed at the end of this section. Recall that the true parameter is $\boldsymbol{\theta}_0$, under which the observations have density $p_0^{(n)}$.

A_1 Condition on $p_0^{(n)}$ and ϵ_n. For n large enough and for some $m > 0$,

$$K\left(p_0^{(n)}, p_{0j_n}^{(n)}\right) \leq n\epsilon_n^2 \quad \text{and} \quad V_{m,0}\left(p_0^{(n)}, p_{0j_n}^{(n)}\right) \leq \left(n\epsilon_n^2\right)^{m/2}.$$

A_2 Comparison between norms. The following inclusion holds in Θ_{j_n}

$$\exists H_1 > 0, \text{ s.t. } \mathcal{A}_n(H_1) \subset \widetilde{\mathcal{B}}_n(m).$$

A_3 Comparison between d_n and l^2. There exist three non negative constants D_0, D_1, D_2 such that, for any two $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_{k_n}(Q)$,

$$d_n(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq D_0 k_n^{D_1} \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_{2,k_n}^{D_2}.$$

A_4 Test Condition. There exist two positive constants c_1 and $\zeta < 1$ such

that, for every $\boldsymbol{\theta}_1 \in \Theta_{k_n}(Q)$, there exists a test $\phi_n(\boldsymbol{\theta}_1) \in [0,1]$ which satisfies

$$\mathbb{E}_0^{(n)}(\phi_n(\boldsymbol{\theta}_1)) \leq e^{-c_1 n d_n^2(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1)} \quad \text{and}$$
$$\sup_{d_n(\boldsymbol{\theta},\boldsymbol{\theta}_1)<\zeta d_n(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1)} \mathbb{E}_{\boldsymbol{\theta}}^{(n)}\left(1-\phi_n(\boldsymbol{\theta}_1)\right) \leq e^{-c_1 n d_n^2(\boldsymbol{\theta}_0,\boldsymbol{\theta}_1)}.$$

A_5 On the prior Π. There exist positive constants $a, b, G_1, G_2, G_3, G_4, H_2, \alpha$ and τ_0 such that π satisfy

$$\forall k = 1, 2, \ldots, \quad e^{-akL(k)} \leq \pi(k) \leq e^{-bkL(k)}, \tag{6}$$

where the function L is a slow varying function introduced in Equation (5); g satisfy

$$\forall \theta \in \mathbb{R}, \quad G_1 e^{-G_2|\theta|^\alpha} \leq g(\theta) \leq G_3 e^{-G_4|\theta|^\alpha}. \tag{7}$$

The scales $\boldsymbol{\tau}$ defined in Equation (2) satisfy the following conditions

$$\max_{j\geq 1} \tau_j \leq \tau_0, \tag{8}$$
$$\min_{j\leq k_n} \tau_j \geq n^{-H_2}, \tag{9}$$
$$\sum_{j=1}^{j_n} |\theta_{0j}|^\alpha / \tau_j^\alpha \leq C j_n \log n. \tag{10}$$

Remark 1.

- Conditions $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$ are local in that they need to be checked at the true parameter $\boldsymbol{\theta}_0$ only. They are useful to prove that the prior puts sufficient mass around Kullback-Leibler neighbourhoods of the true probability. Condition $\boldsymbol{A_1}$ is a limiting factor to the rate: it characterizes ϵ_n through the capacity of approximation of $p_0^{(n)}$ by $p_{0j_n}^{(n)}$: the smoother $p_0^{(n)}$, the closer $p_0^{(n)}$ and $p_{0j_n}^{(n)}$, and the faster ϵ_n. In many models, they are ensured because $K(p_0^{(n)}, p_{\boldsymbol{\theta}_{j_n}}^{(n)})$ and $V_{m,0}(p_0^{(n)}, p_{\boldsymbol{\theta}_{j_n}}^{(n)})$ can be written locally (meaning around $\boldsymbol{\theta}_0$) in terms of the l^2 norm $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{j_n}\|_2$ directly. Smoothness assumptions are then typically required to control $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{j_n}\|_2$.

  It is the case for instance for Sobolev and Besov smoothnesses (*cf.* Equation (12)). The control is expressed with a power of j_n, whose comparison to ϵ_n^2 provides in turn a tight way to tune the rate (*cf.* the proof of

Proposition 1).

Note that the constant H_1 in Condition $\boldsymbol{A_2}$ can be chosen as large as needed: if $\boldsymbol{A_2}$ holds for a specified positive constant H_0, then it does for any $H_1 > H_0$. This makes the condition quite loose. A more stringent version of $\boldsymbol{A_2}$, if simpler, is the following.

$\boldsymbol{A_2'}$ **Comparison between norms**. For any $\boldsymbol{\theta} \in \Theta_{j_n}$

$$\begin{aligned} \widetilde{K}\left(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) &\leq Cn \left\|\boldsymbol{\theta}_{0j_n} - \boldsymbol{\theta}\right\|_{2,j_n}^2 \text{ and} \\ \widetilde{V}_{m,0}\left(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) &\leq Cn^{m/2} \left\|\boldsymbol{\theta}_{0j_n} - \boldsymbol{\theta}\right\|_{2,j_n}^m . \end{aligned}$$

This is satisfied in the Gaussian white noise model (see Section 3).

- Condition $\boldsymbol{A_3}$ is generally mild. The reverse is more stringent since d_n may be bounded, as is the case with the Hellinger distance. $\boldsymbol{A_3}$ is satisfied in many common situations, see for example the applications later on. Technically, this condition allows to switch from a covering number (or entropy) in terms of the l^2 norm to a covering number in terms of the semimetric d_n.

- Condition $\boldsymbol{A_4}$ is common in the Bayesian nonparametric literature. A review of different models and their corresponding tests is given in Ghosal and van der Vaart (2007) for example. The tests strongly depend on the semimetric d_n.

- Condition $\boldsymbol{A_5}$ concerns the prior. Equations (6) and (7) state that the tails of π and g have to be at least exponential or of exponential type. For instance, if π is the geometric distribution, $L = 1$, and if it is the Poisson distribution, $L(k) = \log(k)$ (both are slow varying functions). Laplace and Gaussian distributions are covered by g, with $\alpha = 1$ and $\alpha = 2$ respectively. These equations aim at controlling the prior mass of $\Theta_{k_n}^c(Q)$, the complement of $\Theta_{k_n}(Q)$ in Θ (see Lemma 2). The case where the scale $\boldsymbol{\tau}$ depends on n is considered in Babenko and Belitser (2009, 2010) in the white noise model. Here the constraints on $\boldsymbol{\tau}$ are rather mild since they are allowed to go to zero polynomially as a function of n, and must be upper bounded. Rivoirard and Rousseau (2012a) study a family of scales $\boldsymbol{\tau} = (\tau_j)_{j\geq 1}$ that are decreasing polynomially with j. Here the prior is more general and encompasses both frameworks. Equations (6) - (10) are needed in Lemmas 2 and 3 for bounding respectively $\Pi(\mathcal{B}_n(m))$

from below and $\Pi(\Theta_{k_n}^c(Q))$ from above. A smoothness assumption on $\boldsymbol{\theta}_0$ is usually required for Equation (10).

2.2 Results

2.2.1 Concentration and posterior loss

The following theorem provides an upper bound for the rate of contraction of the posterior distribution.

Theorem 1. *If Conditions* $\boldsymbol{A_1}$ *-* $\boldsymbol{A_5}$ *hold, then for* M *large enough and for* L *introduced in Equation (5),*

$$\mathbb{E}_0^{(n)}\Pi\left(\boldsymbol{\theta}:\ d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0) \geq M\frac{\log n}{L(n)}\epsilon_n^2|X^n\right) = \mathcal{O}\left((n\epsilon_n^2)^{-m/2}\right) \underset{n\to\infty}{\longrightarrow} 0.$$

Proof. See the Appendix. □

The convergence of the posterior distribution at the rate ϵ_n implies that the expected posterior risk converges (at least) at the same rate ϵ_n, when d_n is bounded.

Corollary 1. *Under the assumptions of Theorem 1, with a value of* m *in Conditions* $\boldsymbol{A_1}$ *and* $\boldsymbol{A_2}$ *such that* $(n\epsilon_n^2)^{-m/2} = \mathcal{O}(\epsilon_n^2)$*, and if* d_n *is bounded on* Θ*, then the expected posterior risk given* $\boldsymbol{\theta}_0$ *and* Π *converges at least at the same rate* ϵ_n

$$\mathcal{R}_n^{d_n} = \mathbb{E}_0^{(n)}\Pi(d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0)|X^n) = \mathcal{O}\left(\frac{\log n}{L(n)}\epsilon_n^2\right).$$

Proof. Denote D the bound of d_n, *i.e.* for all $\boldsymbol{\theta},\ \boldsymbol{\theta}' \in \Theta$, $d_n(\boldsymbol{\theta},\boldsymbol{\theta}') \leq D$. We have

$$\begin{aligned}\mathcal{R}_n^{d_n} &\leq M\frac{\log n}{L(n)}\epsilon_n^2 + \mathbb{E}_0^{(n)}\Pi\left(\mathbb{I}\left(d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0) \geq M\frac{\log n}{L(n)}\epsilon_n^2\right) d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0)|X^n\right)\\ &\leq M\frac{\log n}{L(n)}\epsilon_n^2 + D\mathbb{E}_0^{(n)}\Pi\left(\boldsymbol{\theta}:\ d_n^2(\boldsymbol{\theta},\boldsymbol{\theta}_0) \geq M\frac{\log n}{L(n)}\epsilon_n^2|X^n\right)\end{aligned}$$

so $\mathcal{R}_n^{d_n} = \mathcal{O}(\log n/L(n)\epsilon_n^2)$ by Theorem 1 and the assumption on m. □

Remark 2. The condition on m in Corollary 1 requires $n\epsilon_n^2$ to grow as a power of n. When $\boldsymbol{\theta}_0$ has Sobolev smoothness β, this is the case since ϵ_n^2 is typically of

order $(n/\log n)^{-2\beta/(2\beta+1)}$. The condition on m boils down to $m \geq 4\beta$. When $\boldsymbol{\theta}_0$ is smoother, *e.g.* in a Sobolev space with exponential weights, the rate is typically of order $\log n/\sqrt{n}$. Then a common way to proceed is to resort to an exponential inequality for controlling the denominator of the posterior distribution of Equation (4) (see *e.g.* Rivoirard and Rousseau, 2012b).

Remark 3. We can note that this result is meaningful from a non Bayesian point of view as well. Indeed, let $\widehat{\boldsymbol{\theta}}$ be the posterior mean estimate of $\boldsymbol{\theta}$ with respect to Π. Then, if $\boldsymbol{\theta} \to d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ is convex, we have by Jensen's inequality

$$d_n^2(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) \leq \Pi\left(d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|X^n\right),$$

so the frequentist risk converges at the same rate ϵ_n

$$R_n^{d_n} = \mathbb{E}_0^{(n)}(d_n^2(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)) \leq \mathbb{E}_0^{(n)}\Pi\left(d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0)|X^n\right) = \mathcal{R}_n^{d_n} = \mathcal{O}\left(\frac{\log n}{L(n)}\epsilon_n^2\right).$$

Note that we have no result for general pointwise estimates $\widehat{\boldsymbol{\theta}}$, for instance for the MAP. This latter was studied in Abramovich et al. (2007b, 2010).

2.2.2 Adaptation

When considering a given class of smoothness for the parameter $\boldsymbol{\theta}_0$, the minimax criterion implies an optimal rate of convergence. Posterior (resp. risk) adaptation means that the posterior distribution (resp. the risk) concentrates at the optimal rate for a class of possible smoothness values.

We consider here Sobolev classes $\Theta_\beta(L_0)$ for univariate problems defined by

$$\Theta_\beta(L_0) = \left\{\boldsymbol{\theta} : \sum_{j=1}^{\infty} \theta_j^2 j^{2\beta} < L_0\right\},\ \beta > 1/2,\ L_0 > 0 \tag{11}$$

with smoothness parameter β and radius L_0. If $\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)$, then one has the following bound

$$\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{0j_n}\|_2^2 = \sum_{j=j_n+1}^{\infty} \theta_{0j}^2 j^{2\beta} j^{-2\beta} \leq L_0 j_n^{-2\beta}. \tag{12}$$

Donoho and Johnstone (1998) give the global (*i.e.* under the l^2 loss) minimax rate $n^{-\beta/(2\beta+1)}$ attached to the Sobolev class of smoothness β. We show that

under an additional condition between K, $V_{m,0}$ and l^2, the upper bound ϵ_n on the rate of contraction can be chosen equal to the optimal rate, up to a $\log n$ term.

Proposition 1. *Let L_0 denote a positive fixed radius, and $\beta_2 \geq \beta_1 > 1/2$. If for n large enough, there exists a positive constant C_0 such that*

$$\sup_{\beta_1 \leq \beta \leq \beta_2} \sup_{\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)} K\left(p_0^{(n)}, p_{0j_n}^{(n)}\right) \leq C_0 n \left\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{0j_n}\right\|_2^2, \text{ and}$$
$$\sup_{\beta_1 \leq \beta \leq \beta_2} \sup_{\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)} V_{m,0}\left(p_0^{(n)}, p_{0j_n}^{(n)}\right) \leq C_0^m n^{m/2} \left\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{0j_n}\right\|_2^m, \qquad (13)$$

and if Conditions $\boldsymbol{A_2}$ - $\boldsymbol{A_5}$ hold with constants independent of $\boldsymbol{\theta}_0$ in the set $\cup_{\beta_1 \leq \beta \leq \beta_2} \Theta_\beta(L_0)$, then for M sufficiently large,

$$\sup_{\beta_1 \leq \beta \leq \beta_2} \sup_{\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)} \mathbb{E}_0^{(n)} \Pi\left(\boldsymbol{\theta} : \ d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq M \frac{\log n}{L(n)} \epsilon_n^2(\beta) | X^n\right) \underset{n\to\infty}{\longrightarrow} 0,$$

with

$$\epsilon_n(\beta) = \epsilon_0 \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}},$$

and ϵ_0 depending on L_0, C_0 and the constants involved in the assumptions, but not depending on β.

Remark 4. In the standard case where d_n is the l^2 norm, ϵ_n is the *optimal rate of contraction*, up to a $\log n$ term (which is quite common in Bayesian nonparametric computations).

Proof. Let $\beta \in [\beta_1, \beta_2]$ and $\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)$. Then $\boldsymbol{\theta}_0$ satisfies Equation (12), and Condition (13) implies that

$$K\left(p_0^{(n)}, p_{0j_n}^{(n)}\right) \leq C_0 L_0 n j_n^{-2\beta}, \ V_{m,0}\left(p_0^{(n)}, p_{0j_n}^{(n)}\right) \leq C_0 L_0^m n^{m/2} j_n^{-m\beta}.$$

For given $\boldsymbol{\theta}_0$ and β, the result of Theorem 1 holds if Condition $\boldsymbol{A_1}$ is satisfied. This is the case if we choose $\epsilon_n(\beta, \boldsymbol{\theta}_0) \geq C_0 L_0 j_n^{-\beta}$, provided that the bounds in Conditions $\boldsymbol{A_2}$ - $\boldsymbol{A_5}$ and in Equation (13) are uniform. Combined with $j_n = \lfloor j_0 n \epsilon_n^2 / \log n \rfloor$, it gives as a tight choice $\epsilon_n(\beta, \boldsymbol{\theta}_0) = \epsilon_0(\beta, \boldsymbol{\theta}_0)(\log n / n)^{\beta/(2\beta+1)}$ with $\epsilon_0(\beta, \boldsymbol{\theta}_0) \leq (L_0 C_0 j_0^{-\beta})^{1/(2\beta+1)}$. So there exists a bound $\epsilon_0 > 0$ such that $\sup_{\beta_1 \leq \beta \leq \beta_2} \sup_{\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)} \epsilon_0(\beta, \boldsymbol{\theta}_0) = \epsilon_0 < \infty$, which concludes the proof. □

2.3 Examples

In this section, we apply our results of contraction of Sections 2.2.1 and 2.2.2 to a series of models. The Gaussian white noise example is studied in detail in Section 3. We suppose in each model that $\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)$, where $\Theta_\beta(L_0)$ is defined in Equation (11).

Throughout, we consider the following prior Π on Θ (or on a curve space $\mathcal{F}$ through the coefficients of the functions in a basis). Let the prior distribution π on k be Poisson with parameter λ, and given k, the prior distribution on θ_j/τ_j, $j = 1, \ldots, k$ be standard Gaussian,

$$\begin{aligned} k &\sim \text{Poisson}(\lambda), \\ \frac{\theta_j}{\tau_j} \,|\, k &\sim \mathcal{N}(0,1),\ j = 1, \ldots, k, \text{ independently.} \end{aligned} \tag{14}$$

It satisfies Equation (6) with function $L(k) = \log(k)$ and Equation (7) with $\alpha = 2$. Choose then $\tau_j^2 = \tau_0 j^{-2q}$, $\tau_0 > 0$, with $q > 1/2$. It is decreasing and bounded from above by τ_0 so Equation (8) is satisfied. Additionally,

$$\min_{j \le k_n} \tau_j = \tau_{k_n} = k_n^{-2q} \ge n^{-H_2}$$

for H_2 large enough, so Equation (9) is checked. Since $\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)$,

$$\tau_0^2 \sum_{j=1}^{j_n} \theta_{0j}^2/\tau_j^2 = \sum_{j=1}^{j_n} \theta_{0j}^2 j^{2q} = \sum_{j=1}^{j_n} \theta_{0j}^2 j^{2\beta} j^{2q-2\beta} \le j_n \sum_{j=1}^{j_n} \theta_{0j}^2 j^{2\beta} \le j_n L_0,$$

as soon as $2q - 2\beta \le 1$. Hence by choosing $1/2 < q \le 1$, Equation (10) is verified for all $\beta > 1/2$. The prior Π thus satisfies Condition $\boldsymbol{A_5}$.

Since Condition $\boldsymbol{A_5}$ is satisfied, we will show in the three examples that Conditions $\boldsymbol{A_2}$ - $\boldsymbol{A_4}$ and Condition (13) hold, thus Proposition 1 applies: the posterior distribution attains the optimal rate of contraction, up to a $\log n$ term, that is $\epsilon_n = \epsilon_0 (\log n/n)^{\beta/(2\beta+1)}$, for a distance d_n which is specific to each model. This rate is adaptive in a range of smoothness $[\beta_1, \beta_2]$.

2.3.1 Density

Let us consider the density model in which the density $\boldsymbol{p}$ is unknown, and we observe i.i.d. data

$$X_i \sim \boldsymbol{p}, \quad i = 1, 2, \ldots, n,$$

where $\boldsymbol{p}$ belongs to $\mathcal{F}$,

$$\mathcal{F} = \left\{ \boldsymbol{p} \text{ density on } [0,1] : \boldsymbol{p}(0) = \boldsymbol{p}(1) \text{ and } \log \boldsymbol{p} \in L^2(0,1) \right\}.$$

Equality $\boldsymbol{p}(0) = \boldsymbol{p}(1)$ is mainly used for ease of computation. We define the parameter $\boldsymbol{\theta}$ of such a function $\boldsymbol{p}$, and write $\boldsymbol{p} = \boldsymbol{p_\theta}$, as the coefficients of $\log \boldsymbol{p_\theta}$ in the Fourier basis $\boldsymbol{\psi} = (\psi_j)_{j \geq 1}$, *i.e.* it can be represented as

$$\log \boldsymbol{p_\theta}(x) = \sum_{j=1}^{\infty} \theta_j \psi_j(x) - c(\boldsymbol{\theta}),$$

where $c(\boldsymbol{\theta})$ is a normalizing constant. We assign a prior to $\boldsymbol{p_\theta}$ by assigning the sieve prior Π of Equation (14) to $\boldsymbol{\theta}$.

A natural choice of metric d_n in this model is the Hellinger distance $d_n(\boldsymbol{\theta}, \boldsymbol{\theta'}) = h(\boldsymbol{p_\theta}, \boldsymbol{p_{\theta'}}) = \left(\int \left(\sqrt{\boldsymbol{p_\theta}} - \sqrt{\boldsymbol{p_{\theta'}}} \right)^2 d\mu \right)^{1/2}$. Lemma 2 in Ghosal and van der Vaart (2007) shows the existence of tests satisfying $\boldsymbol{A_4}$ with the Hellinger distance.

Rivoirard and Rousseau (2012b) study this model in detail (Section 4.2.2) in order to derive a Bernstein-von Mises theorem for the density model. They prove that Conditions $\boldsymbol{A_2}$, $\boldsymbol{A_3}$ and (13) are valid in this model (see the proof of Condition (C) for $\boldsymbol{A_2}$ and (13), and the proof of Condition (B) for $\boldsymbol{A_3}$). With $D_1 = D_2 = 1$, Condition $\boldsymbol{A_3}$ is written $h(\boldsymbol{p_\theta}, \boldsymbol{p_{\theta'}}) \leq D_0 k_n \left\| \boldsymbol{\theta} - \boldsymbol{\theta'} \right\|_{2,k_n}$.

2.3.2 Regression

Consider now the following nonparametric regression model

$$X_i = \boldsymbol{f}(t_i) + \sigma \xi_i, \quad i = 1, \ldots, n,$$

with the regular fixed design $t_i = i/n$ in $[0,1]$, i.i.d. centered Gaussian errors ξ_i with variance σ^2. The unknown σ case is studied in an unpublished paper by Rousseau and Sun. They endow σ with an Inverse Gamma (conjugate) prior. They show that this one dimensional parameter adds an $n \log(\sigma/\sigma_0)$ term in the

Kullback-Leibler divergence but does not alter the rates by considering three different cases for σ, either $\sigma < \sigma_0/2$, $\sigma > 3\sigma_0/2$, or $\sigma \in [\sigma_0/2, 3\sigma_0/2]$.

We consider now in more detail the σ known case. Denote $\boldsymbol{\theta}$ the coefficients of a regression function $\boldsymbol{f}$ in the Fourier basis $\boldsymbol{\psi} = (\psi_j)_{j\geq 1}$. So for all $t \in [0,1]$, $\boldsymbol{f}$ can be represented as $\boldsymbol{f}(t) = \sum_{j=1}^{\infty} \theta_j \psi_j(t)$. We assign a prior to $\boldsymbol{f}$ by assigning the sieve prior Π of Equation (14) to $\boldsymbol{\theta}$.

Let $\mathbb{P}_t^n = n^{-1}\sum_{i=1}^n \delta_{t_i}$ be the empirical measure of the covariates t_i's, and define the square of the empirical norm by $\|\boldsymbol{f}\|_{\mathbb{P}_t^n}^2 = n^{-1}\sum_{i=1}^n \boldsymbol{f}^2(t_i)$. We use $d_n = \|\cdot\|_{\mathbb{P}_t^n}$.

Let $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{f}$ the corresponding regression. Basic algebra (see for example Lemma 1.7 in Tsybakov, 2009) provides, for any two j and k,

$$\frac{1}{n}\sum_{i=1}^n \psi_j(t_i)\psi_k(t_i) = \delta_{jk},$$

where δ_{jk} stands for Kronecker delta. Hence

$$\|\boldsymbol{f}\|_{\mathbb{P}_t^n}^2 = \frac{1}{n}\sum_{i=1}^n \sum_{j,k} \theta_j\theta_k\psi_j(t_i)\psi_k(t_i) = \|\boldsymbol{\theta}\|_2^2 = \|\boldsymbol{f}\|_2^2, \tag{15}$$

where the last equality is Parseval's. It ensures Condition $\boldsymbol{A_3}$ with $D_0 = D_2 = 1$ and $D_1 = 0$.

The densities $\mathcal{N}(\boldsymbol{f}(t_i), \sigma^2)$ of X_i's are denoted $p_{\boldsymbol{f},i}$, $i = 1, \ldots, n$, and their product $p_{\boldsymbol{f}}^{(n)}$. The quantity $\boldsymbol{f}_{0j_n}$ denotes the truncated version of $\boldsymbol{f}_0$ to its first j_n terms in the Fourier basis.

We have $2K(p_{\boldsymbol{f}_0,i}, p_{\boldsymbol{f},i}) = V_{2,0}(p_{\boldsymbol{f}_0,i}, p_{\boldsymbol{f},i}) = \sigma^{-2}(\boldsymbol{f}_0(t_i) - \boldsymbol{f}(t_i))^2$ and $V_{m,0}(p_{\boldsymbol{f}_0,i}, p_{\boldsymbol{f},i}) = \sigma_m\sigma^{m-2}|\boldsymbol{f}_0(t_i) - \boldsymbol{f}(t_i)|^2$ for $m \geq 2$, where σ_m is the (non centered) $m^{\text{th}}-$moment of a standard Gaussian variable. So using Equation (15) we get

$$2K(p_{\boldsymbol{f}_0}^{(n)}, p_{\boldsymbol{f}}^{(n)}) = V_{2,0}(p_{\boldsymbol{f}_0}^{(n)}, p_{\boldsymbol{f}}^{(n)}) = n\sigma^{-2}\|\boldsymbol{f}_0 - \boldsymbol{f}\|_{\mathbb{P}_t^n}^2 = n\sigma^{-2}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|_2^2$$

which proves Condition (13).

Additionally, both $2\widetilde{K}(p_{\boldsymbol{f}_{0j_n}}^{(n)}, p_{\boldsymbol{f}}^{(n)})$ and $\widetilde{V}_{2,0}(p_{\boldsymbol{f}_{0j_n}}^{(n)}, p_{\boldsymbol{f}}^{(n)})$ are upper bounded by $n\sigma^{-2}(2\|\boldsymbol{f}_{0j_n} - \boldsymbol{f}\|_{\mathbb{P}_t^n}^2 + \|\boldsymbol{f}_0 - \boldsymbol{f}_{0j_n}\|_{\mathbb{P}_t^n}^2)$. Let $\boldsymbol{\theta} \in \mathcal{A}_n(H_1)$, for a certain $H_1 > 0$. Then, using (15) again, the bound is less than

$$n\sigma^{-2}(n^{-H_1} + L_0 j_n^{-2\beta}) \leq Cn\epsilon_n^2$$

for $H_1 > 2\beta/(2\beta+1)$, which ensures Condition $\boldsymbol{A_2}$.

Ghosal and van der Vaart (2007) state in Section 7.7 that tests satisfying $\boldsymbol{A_4}$ exist with $d_n = \|\cdot\|_{\mathbb{P}_t^n}$.

2.3.3 Nonlinear AR(1) model

As a nonindependent illustration, we consider the following Markov chain: the nonlinear autoregression model whose observations $X^n = (X_1, \ldots, X_n)$ come from a stationary time series $X_t, t \in \mathbb{Z}$, such that

$$X_i = \boldsymbol{f}(X_{i-1}) + \xi_i, \quad i = 1, 2, \ldots, n,$$

where the function $\boldsymbol{f}$ is unknown and the residuals ξ_i are standard Gaussian and independent of $(X_1, \ldots, X_{i-1})$. We suppose that X_0 is drawn in the stationary distribution.

Suppose that regression functions $\boldsymbol{f}$ are in $L_2(\mathbb{R})$, and uniformly bounded by a constant M_1 (a bound growing with n could also be considered here). We use Hermite functions $\boldsymbol{\psi} = (\psi_j)_{j\geq 1}$ as an orthonormal basis of $\mathbb{R}$, such that for all $x \in \mathbb{R}$, $\boldsymbol{f}(x) = \boldsymbol{f_\theta}(x) = \sum_{j=1}^{\infty} \theta_j \psi_j(x)$. This basis is uniformly bounded (by Cramér's inequality). Consider the sieve prior Π in its truncated version (3) for $\boldsymbol{\theta}$, with radius r_1 a (possibly large) constant independent of k and n.

We show that Conditions $\boldsymbol{A_1}$-$\boldsymbol{A_4}$ are satisfied, along the lines of Ghosal and van der Vaart (2007) Sections 4 and 7.4. Denote $p_{\boldsymbol{\theta}}(y|x) = \varphi(y - \boldsymbol{f_\theta}(x))$ the transition density of the chain, where $\varphi(\cdot)$ is the standard normal density distribution, and where reference measures relative to x and y are denoted respectively by ν and μ. Define $r(y) = \frac{1}{2}(\varphi(y - M_1) + \varphi(y + M_1))$, and set $d\nu = r d\mu$. Then Ghosal and van der Vaart (2007) show that the chain $(X_i)_{1\leq i\leq n}$ has a unique stationary distribution and prove the existence of tests satisfying $\boldsymbol{A_4}$ relative to the Hellinger semidistance d whose square is given by

$$d^2(\boldsymbol{\theta}, \boldsymbol{\theta'}) = \int\int \left(\sqrt{p_{\boldsymbol{\theta}}(y|x)} - \sqrt{p_{\boldsymbol{\theta'}}(y|x)}\right)^2 d\mu(y) d\nu(x).$$

They show that d is bounded by $\|\cdot\|_2$ (which proves Condition $\boldsymbol{A_3}$) and that

$$K(p_0, p_{\boldsymbol{\theta}}) = V_{2,0}(p_0, p_{\boldsymbol{\theta}}) \lesssim \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\|_2^2.$$

Thus Equation (13) holds. Condition $\boldsymbol{A_2}$ follows from inequalities $\widetilde{K}(p_{0j_n}, p_{\boldsymbol{\theta}}) \lesssim \sum_{j=1}^{j_n} |\theta_{0j} - \theta_j|$ and $\widetilde{V}_{2,0}(p_{0j_n}, p_{\boldsymbol{\theta}}) \lesssim \|\boldsymbol{\theta}_{0j_n} - \boldsymbol{\theta}\|^2_{2,j_n}$ for $\boldsymbol{\theta} \in \Theta_{j_n}$.

3 Application to the white noise model

Consider the Gaussian white noise model

$$dX^n(t) = \boldsymbol{f}_0(t)dt + \frac{1}{\sqrt{n}}dW(t), \quad 0 \leq t \leq 1, \tag{16}$$

in which we observe processes $X^n(t)$, where $\boldsymbol{f}_0$ is the unknown function of interest belonging to $L^2(0,1)$, $W(t)$ is a standard Brownian motion, and n is the sample size. We assume that $\boldsymbol{f}_0$ lies in a Sobolev ball, $\Theta_\beta(L_0)$, see (11). Brown and Low (1996) show that this model is asymptotically equivalent to the nonparametric regression (assuming $\beta > 1/2$). It can be written as the equivalent infinite normal mean model using the decomposition in a Fourier basis $\boldsymbol{\psi} = (\psi_j)_{j\geq 1}$ of $L^2(0,1)$,

$$X_j^n = \theta_{0j} + \frac{1}{\sqrt{n}}\xi_j,\ j = 1, 2, \ldots \tag{17}$$

where $X_j^n = \int_0^1 \psi_j(t)\, dX^n(t)$ are the observations, $\theta_{0j} = \int_0^1 \psi_j(t)\, \boldsymbol{f}_0(t)dt$ the Fourier coefficients of $\boldsymbol{f}_0$, and $\xi_j = \int_0^1 \psi_j(t)dW(t)$ are independent standard Gaussian random variables. The function $\boldsymbol{f}_0$ and the parameter $\boldsymbol{\theta}_0$ are linked through the relation in $L^2(0,1)$, $\boldsymbol{f}_0 = \sum_{j=1}^\infty \theta_{0j}\psi_j$.

In addition to results in concentration, we are interested in comparing the risk of an estimate $\hat{\boldsymbol{f}}_n$ corresponding to basis coefficients $\hat{\boldsymbol{\theta}}_n$, under two different losses: the global L^2 loss (if expressed on curves $\boldsymbol{f}$, or l^2 loss if expressed on $\boldsymbol{\theta}$),

$$R_n^{L^2}(\boldsymbol{\theta}_0) = \mathbb{E}_0^{(n)} \left\| \hat{\boldsymbol{f}}_n - \boldsymbol{f}_0 \right\|_2^2 = \mathbb{E}_0^{(n)} \sum_{j=1}^\infty \left(\hat{\theta}_{nj} - \theta_{0j} \right)^2,$$

and the local loss at point $t \in [0,1]$,

$$R_n^{\mathrm{loc}}(\boldsymbol{\theta}_0, t) = \mathbb{E}_0^{(n)} \left(\hat{\boldsymbol{f}}_n(t) - \boldsymbol{f}_0(t) \right)^2 = \mathbb{E}_0^{(n)} \left(\sum_{j=1}^\infty a_j \left(\hat{\theta}_{nj} - \theta_{0j} \right) \right)^2,$$

with $a_j = \psi_j(t)$. Note that the difference between global and local risks expres-

sions in basis coefficients comes from the parenthesis position with respect to the square: respectively the sum of squares and the square of a sum.

We show that sieve priors allow to construct adaptive estimate in global risk. However, the same estimate does not perform as well under the pointwise loss, which illustrates the result of Cai et al. (2007). We provide a lower bound for the pointwise rate.

3.1 Adaptation under global loss

Consider the global l^2 loss on $\boldsymbol{\theta}_0$. The likelihood ratio is given by

$$\frac{p_0^{(n)}}{p_{\boldsymbol{\theta}}^{(n)}}(X^n) = \exp\left(n\langle \boldsymbol{\theta}_0 - \boldsymbol{\theta}, X^n\rangle - \frac{n}{2}\left\|\boldsymbol{\theta}_0\right\|_2^2 + \frac{n}{2}\left\|\boldsymbol{\theta}\right\|_2^2\right),$$

where $\langle .,. \rangle$ denotes the l^2 scalar product. We choose here the l^2 distance as $d_n(\boldsymbol{\theta}, \boldsymbol{\theta}') = \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2$. Let us check that Conditions $\boldsymbol{A_2}$ - $\boldsymbol{A_4}$ and Condition (13) hold.

The choice of d_n ensures Condition $\boldsymbol{A_3}$ with $D_0 = D_2 = 1$ and $D_1 = 0$. The test statistic of $\boldsymbol{\theta}_0$ against $\boldsymbol{\theta}_1$ associated with the likelihood ratio is $\phi_n(\boldsymbol{\theta}_1) = \mathbb{I}(2\langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_0, X^n\rangle > \|\boldsymbol{\theta}_1\|_2^2 - \|\boldsymbol{\theta}_0\|_2^2)$. With Lemma 5 of Ghosal and van der Vaart (2007) we have that $\mathbb{E}_0^{(n)}(\phi_n(\boldsymbol{\theta}_1)) \leq e^{-n\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|_2^2/4}$ and $\mathbb{E}_{\boldsymbol{\theta}}^{(n)}(1 - \phi_n(\boldsymbol{\theta}_1)) \leq e^{-n\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|_2^2/4}$ for $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_1\|_2 \leq \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|_2/4$. It provides a test as in Condition $\boldsymbol{A_4}$ with $c_1 = \zeta = 1/4$.

Moreover, following Lemma 6 of Ghosal and van der Vaart (2007) we have

$$K\left(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) = n\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\|_2^2/2 \text{ and } V_{2,0}\left(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) = n\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\|_2^2.$$

For $m \geq 2$, we have

$$\begin{aligned} V_{m,0}\left(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) &= \int p_0^{(n)} \left|\log\left(p_0^{(n)}/p_{\boldsymbol{\theta}}^{(n)}\right) - K\left(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right)\right|^m d\mu \\ &= n^m \int p_0^{(n)} \left|\langle \boldsymbol{\theta}_0 - \boldsymbol{\theta}, X^n - \boldsymbol{\theta}_0\rangle\right|^m d\mu \\ &\leq n^m \left\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\right\|_2^m \int p_0^{(n)} \left\|X^n - \boldsymbol{\theta}_0\right\|_2^m d\mu. \end{aligned}$$

The centered $m^{\text{th}}-$moment of the Gaussian variable X^n is proportional to $n^{-m/2}$, so $V_{m,0}\left(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) \lesssim n^{m/2}\left\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\right\|_2^m$, and Condition (13) is satisfied.

The same calculation shows that Condition $\boldsymbol{A_2'}$ is satisfied: for all $\boldsymbol{\theta} \in \Theta_{j_n}$, $\widetilde{K}\left(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) = \frac{n}{2}\left\|\boldsymbol{\theta}_{0j_n} - \boldsymbol{\theta}\right\|_{2,j_n}^2$ and $\widetilde{V}_{m,0}\left(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}\right) \lesssim n^{m/2}\left\|\boldsymbol{\theta}_{0j_n} - \boldsymbol{\theta}\right\|_{2,j_n}^m$.

Conditions $\boldsymbol{A_2}$ - $\boldsymbol{A_4}$ and Condition (13) hold, if moreover $\boldsymbol{A_4}$ is satisfied, then by Proposition 1, the procedure is adaptive, which is expressed in the following proposition.

Proposition 2. *Under the prior* Π *defined in Equations (14), the global* l^2 *rate of posterior contraction is optimal adaptive for the Gaussian white noise model, i.e. for* M *large enough and* $\beta_2 \geq \beta_1 > 1/2$

$$\sup_{\beta_1 \leq \beta \leq \beta_2} \sup_{\boldsymbol{\theta}_0 \in \Theta_\beta(L_0)} \mathbb{E}_0^{(n)} \Pi\left(\boldsymbol{\theta} : \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\|_2^2 \geq M \frac{\log n}{L(n)} \epsilon_n^2(\beta) | X^n\right) \underset{n\to\infty}{\longrightarrow} 0,$$

with $\epsilon_n(\beta) = \epsilon_0 \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}$.

The distance here is not bounded, so Corollary 1 does not hold. For deriving a risk rate, we need a more subtle result than Theorem 1 that we can obtain when considering sets $\mathcal{S}_{n,j}(M) = \left\{\boldsymbol{\theta} : M\frac{\log n}{L(n)}(j+1)\epsilon_n^2 \geq \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\|_2^2 \geq M\frac{\log n}{L(n)} j \epsilon_n^2\right\}$, $j = 1, 2, \ldots$ instead of $\mathcal{S}_n(M) = \left\{\boldsymbol{\theta} : \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\|_2^2 \geq M\frac{\log n}{L(n)}\epsilon_n^2\right\}$. Then the bound of the expected posterior mass of $\mathcal{S}_{n,j}(M)$ becomes

$$\mathbb{E}_0^{(n)} \Pi(\mathcal{S}_{n,j}(M) | X^n) \leq c_7 \left(nj\epsilon_n^2\right)^{-m/2} \tag{18}$$

for a fixed constant c_7. Hence we obtain the following rate of convergence in risk.

Proposition 3. *Under Condition (13) with* $m \geq 5$*, the expected posterior risk given* $\boldsymbol{\theta}_0$ *and* Π *converges at least at the same rate* ϵ_n

$$\mathcal{R}_n^{L^2}(\boldsymbol{\theta}_0) = \mathbb{E}_0^{(n)} \Pi\left[\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\|_2^2 | X^n\right] = \mathcal{O}\left(\epsilon_n^2\right),$$

for any $\boldsymbol{\theta}_0$*. So the procedure is risk adaptive as well (up to a* $\log(n)$ *term).*

Proof. We have

$$\begin{aligned}\mathcal{R}_n^{L^2}(\boldsymbol{\theta}_0) &\leq \mathbb{E}_0^{(n)}\Pi\left[\left(\mathbb{I}(\boldsymbol{\theta}\notin\mathcal{S}_n(M))+\sum_{j\geq 1}\mathbb{I}(\boldsymbol{\theta}\in\mathcal{S}_{n,j}(M))\right)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2\,|X^n\right]\\ &\leq M\frac{\log n}{L(n)}\epsilon_n^2\left(1+\sum_{j=1}^{\infty}(j+1)\mathbb{E}_0^{(n)}\Pi(\mathcal{S}_{n,j}(M)|X^n)\right).\end{aligned}$$

Due to (18), the last sum in j converges as soon as $m \geq 5$. This is possible in the white noise setting because the conditions are satisfied whatever m. So $\mathcal{R}_n^{L^2}(\boldsymbol{\theta}_0) = \mathcal{O}\left(\epsilon_n^2\right)$. □

We have shown that conditional to the existence of a sieve prior for the white noise model satisfying $\boldsymbol{A_5}$ (*cf.* Section 2.3), the procedure has minimax rates (up to a $\log(n)$ term) both in contraction and in risk. We now study the asymptotic behaviour of the posterior under the local loss function.

3.2 Lower bound under pointwise loss

The previous section derives rates of convergence under the global loss. Here, under the pointwise loss, we show that the risk deteriorates as a power n factor compared to the benchmark minimax pointwise risk $n^{-(2\beta-1)/2\beta}$ (note the difference with the global minimax rate $n^{-2\beta/(2\beta+1)}$, both given for risks on squares). We use the sieve prior defined as a conditional Gaussian prior in Equation (14). Denote by $\hat{\boldsymbol{\theta}}_n$ the Bayes estimate of $\boldsymbol{\theta}$ (the posterior mean). Then the following proposition gives a lower bound on the risk (pointwise square error) under a pointwise loss:

Proposition 4. *If the point t is such that $a_j = \psi_j(t) = 1$ for all j ($t = 0$), then for all $\beta \geq q$, for all $L_0 > 0$, a lower bound on the risk rate under pointwise loss is given by*

$$\sup_{\boldsymbol{\theta}_0\in\Theta_\beta(L_0)} R_n^{\mathrm{loc}}(\boldsymbol{\theta_0}, t) \gtrsim n^{-\frac{2\beta-1}{2\beta+1}}/\log^2 n.$$

Proof. See the Appendix. □

Cai et al. (2007) show that a global optimal estimator cannot be pointwise optimal. The sieve prior leads to an (almost up to a $\log n$ term) optimal global

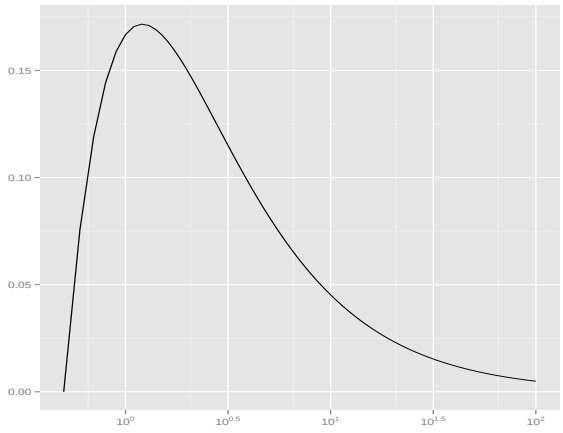


Fig. 2: Variation of the exponent of the penalty in a log scale for β between $1/2$ and 100; it is maximum for $\beta = (1+\sqrt{2})/2$

risk and Proposition 4 shows that the pointwise risk associated to the posterior mean $\hat{\boldsymbol{\theta}}_n$ is suboptimal with a power of n penalty, whose exponent is

$$\frac{2\beta-1}{2\beta} - \frac{2\beta-1}{2\beta+1} = \frac{2\beta-1}{2\beta(2\beta+1)}.$$

The maximal penalty is for $\beta = (1+\sqrt{2})/2$, and it vanishes as β tends to $1/2$ and $+\infty$ (see the Figure 2). Abramovich et al. (2007a) also derive such a power n penalty on the maximum local risk of a globally optimal Bayesian estimate, as well as on the reverse case (maximum global risk of a locally optimal Bayesian estimate).

Remark 5. This result is not anecdotal and illustrates the fact that the Bayesian approach is well suited for loss functions that are related to the Kullback-Leibler divergence (*i.e.* often the l^2 loss). The pointwise loss does not satisfy this since it corresponds to an unsmooth linear functional of $\boldsymbol{\theta}$. This possible suboptimality of the posterior distribution of some unsmooth functional of the parameter has already been noticed in various other cases, see for instance Rivoirard and Rousseau (2012b) or Rousseau and Kruijer (2011). The question of the existence of a fully Bayesian adaptive procedure to estimate $\boldsymbol{f}_0(t) = \sum_{j=1}^{\infty} a_j \theta_{0j}$ remains an open question.

Acknowledgements

We would like to thank the referees for their valuable comments which have helped to improve the manuscript.

References

Abramovich, F., Amato, U., and Angelini, C. (2004). On optimality of bayesian wavelet estimators. *Scand. J. Stat.*, 31(2):217–234.

Abramovich, F., Angelini, C., and De Canditiis, D. (2007a). Pointwise optimality of bayesian wavelet estimators. *Ann. Inst. Statist. Math.*, 59(3):425–434.

Abramovich, F., Grinshtein, V., and Pensky, M. (2007b). On optimality of bayesian testimation in the normal means problem. *Ann. Statist.*, 35(5):2261–2286.

Abramovich, F., Grinshtein, V., Petsa, A., and Sapatinas, T. (2010). On bayesian testimation and its application to wavelet thresholding. *Biometrika*, 97(1):181–198.

Abramovich, F., Sapatinas, T., and Silverman, B. (1998). Wavelet thresholding via a bayesian approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(4):725–749.

Babenko, A. and Belitser, E. (2009). On the posterior pointwise convergence rate of a Gaussian signal under a conjugate prior. *Statist. Probab. Lett.*, 79(5):670–675.

Babenko, A. and Belitser, E. (2010). Oracle convergence rate of posterior under projection prior and Bayesian model selection. *Math. Methods Statist.*, 19(3):219–245.

Barron, A., Schervish, M., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 17(2):536–561.

Belitser, E. and Ghosal, S. (2003). Adaptive bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.*, 31(2):536–559.

Brown, L. and Low, M. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398.

Cai, T., Low, M., and Zhao, L. (2007). Trade-offs between global and local risks in nonparametric function estimation. *Bernoulli*, 13(1):1–19.

De Jonge, R. and van Zanten, J. (2010). Adaptive nonparametric bayesian inference using location-scale mixture priors. *Ann. Statist.*, 38(6):3300–3320.

Donoho, D. and Johnstone, I. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921.

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

Ghosal, S., Lember, J., and van der Vaart, A. W. (2008). Nonparametric bayesian model selection and averaging. *Electron. J. Stat.*, 2:63–89.

Ghosal, S. and van der Vaart, A. W. (2007). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):697–723.

Huang, T. (2004). Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.*, 32(4):1556–1593.

Kruijer, W., Rousseau, J., and van der Vaart, A. W. (2010). Adaptive bayesian density estimation with location-scale mixtures. *Electron. J. Stat.*, 4:1225–1257.

Rivoirard, V. and Rousseau, J. (2012a). Bernstein-von Mises theorem for linear functionals of the density. *Ann. Statist.*, 40(3):1489–1523.

Rivoirard, V. and Rousseau, J. (2012b). Posterior concentration rates for infinite-dimensional exponential families. *Bayesian Anal.*, 7(2):311–334.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Verlag.

Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.*, 38(1):146–180.

Rousseau, J., Chopin, N., and Liseo, B. (2012). Bayesian nonparametric estimation of the spectral density of a long or intermediate memory gaussian process. *Ann. Statist.*, 40(2):964–995.

Rousseau, J. and Kruijer, W. (2011). Adaptive bayesian estimation of a spectral density. *Preprint*.

Scricciolo, C. (2006). Convergence rates for bayesian density estimation of infinite-dimensional exponential families. *Ann. Statist.*, 34(6):2897–2920.

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

Tsybakov, A. (2009). *Introduction to nonparametric estimation*. Springer Verlag.

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

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

Zhao, L. (1993). *Frequentist and Bayesian aspects of some nonparametric estimation problems*. PhD thesis, Ph. D. thesis. Cornell University.

Zhao, L. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.*, 28(2):532–552.

Corresponding author: Julyan Arbel, E37, Laboratoire de Statistique, CREST, 92245 Malakoff, France.
E-mail: *julyan.arbel@m4x.org*

A Appendix

A.1 Three technical lemmas

Set $\mathcal{S}_n(M) = \{\boldsymbol{\theta} : d_n^2(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \geq M \frac{\log n}{L(n)} \epsilon_n^2\}$ and recall that $\Theta_{k_n}(Q) = \{\boldsymbol{\theta} \in \Theta_{k_n} : \|\boldsymbol{\theta}\|_{2,k_n} \leq n^Q\}$, $Q > 0$. We begin with three technical lemmas.

Lemma 1. *If Conditions* $\boldsymbol{A_3}$ *and* $\boldsymbol{A_4}$ *hold, then there exists a test* $\boldsymbol{\phi}_n$ *such that for* M *large enough, there exists a constant* c_2 *such that*

$$\mathbb{E}_0^{(n)}(\boldsymbol{\phi}_n) \leq e^{-c_2 M \frac{\log n}{L(n)} n\epsilon_n^2} \quad \textit{and} \quad \mathbb{E}_{\boldsymbol{\theta}}^{(n)}(1 - \boldsymbol{\phi}_n) \leq e^{-c_2 M \frac{\log n}{L(n)} n\epsilon_n^2},$$

for all $\boldsymbol{\theta} \in \mathcal{S}_n(M) \cap \Theta_{k_n}(Q)$.

Proof. Set $r_n = \left(\sqrt{M\frac{\log n}{L(n)}}\frac{\zeta\epsilon_n}{D_0 k_n^{D_1}}\right)^{1/D_2}$. The set $\mathcal{S}_n(M)\cap\Theta_{k_n}(Q)$ is compact relative to the l^2 norm. Let a covering of this set by l^2 balls of radius r_n and centre $\boldsymbol{\theta}^{(i)}$. Its number of elements is $\eta_n \lesssim (Cn^Q/r_n)^{k_n} \lesssim \exp(Ck_n\log n) \lesssim \exp(C\frac{\log n}{L(n)}n\epsilon_n^2)$ due to relation (5).

For each centre $\boldsymbol{\theta}^{(i)} \in \mathcal{S}_n(M)\cap\Theta_{k_n}(Q)$, there exists a test $\phi_n(\boldsymbol{\theta}^{(i)})$ satisfying Condition $\boldsymbol{A_4}$. We define the test $\boldsymbol{\phi}_n = \max_i \phi_n(\boldsymbol{\theta}^{(i)})$ which satisfies

$$\mathbb{E}_0^{(n)}(\boldsymbol{\phi}_n) \le \eta_n e^{-c_1 M\frac{\log n}{L(n)}n\epsilon_n^2} \le e^{C\frac{\log n}{L(n)}n\epsilon_n^2 - c_1 M\frac{\log n}{L(n)}n\epsilon_n^2} \le e^{-c_2 M\frac{\log n}{L(n)}n\epsilon_n^2},$$

for M large enough and a constant c_2.

Here, Condition $\boldsymbol{A_3}$ allows to switch from the coverage in term of the l^2 distance to a covering expressed in term of d_n: each $\boldsymbol{\theta}\in\mathcal{S}_n(M)\cap\Theta_{k_n}(Q)$ which lies in a l^2 ball of centre $\boldsymbol{\theta}^{(i)}$ and of radius r_n in the covering of size η_n also lies in a d_n ball of adequate radius

$$d_n(\boldsymbol{\theta},\boldsymbol{\theta}^{(i)}) \le D_0 k_n^{D_1}\|\boldsymbol{\theta}-\boldsymbol{\theta}^{(i)}\|_2^{D_2} \le D_0 k_n^{D_1} r_n^{D_2} = \zeta\epsilon_n\sqrt{M\frac{\log n}{L(n)}}.$$

Then there exists a constant c_2 (the minimum with the previous one)

$$\sup_{\boldsymbol{\theta}\in\mathcal{S}_n(M)\cap\Theta_{k_n}(Q)} \mathbb{E}_{\boldsymbol{\theta}}^{(n)}(1-\boldsymbol{\phi}_n) \le e^{-c_2 M\frac{\log n}{L(n)}n\epsilon^2},$$

hence the result follows. □

Lemma 2. *Under Condition $\boldsymbol{A_5}$, for any constant $c_6>0$, there exist positive constants Q, C and M_0 such that*

$$\Pi(\Theta_{k_n}^c(Q)) \le Ce^{-c_6 n\epsilon_n^2}, \tag{19}$$

where M_0 is introduced in the definition (5) of k_n, and $\Theta_{k_n}^c(Q)$, the complementary of $\Theta_{k_n}(Q)$, is taken in Θ.

Proof. $\Theta_{k_n}^c(Q)$ is written by $\Theta_{k_n}^c(Q) = \{\boldsymbol{\theta}\in\Theta : \|\boldsymbol{\theta}\|_{2,k_n} > n^Q \text{ or } \exists j > k_n \text{ s.t. } \theta_j\neq 0\}$, so its prior mass is less than $\pi(k>k_n) + \sum_{k\le k_n}\pi_k\Pi_k(\boldsymbol{\theta}\in\Theta_k : \|\boldsymbol{\theta}\|_{2,k} > n^Q)$, where the last sum is less than $\Pi_{k_n}(\boldsymbol{\theta}\in\Theta_{k_n} : \|\boldsymbol{\theta}\|_{2,k_n} > n^Q)$ because its terms are increasing.

The prior mass of sieves that exceed k_n is controlled by Equation (6). We have

$$\pi\left(k \geq k_n\right) \leq \sum_{j \geq k_n} e^{-bjL(j)} \leq \sum_{j \geq k_n} e^{-bjL(k_n)} \leq Ce^{-bk_nL(k_n)}.$$

Since L is a slow varying function, we have $k_nL(k_n) \gtrsim j_n \log(n) \gtrsim n\epsilon_n^2$. Hence $\pi\left(k \geq k_n\right) \leq Ce^{-c_6 n\epsilon_n^2}$ for a constant c_6 as large as needed since it is determined by constant M_0 in Equation (5).

Then by the second part of Condition (7), $\Pi_{k_n}\left(\boldsymbol{\theta} \in \Theta_{k_n} : \|\boldsymbol{\theta}\|_{2,k_n} > n^Q\right)$ is less than

$$\begin{aligned} & \int_{\|\boldsymbol{\theta}\|_{2,k_n} > n^Q} \prod_{j=1}^{k_n} g(\theta_j/\tau_j)/\tau_j d\theta_j, \\ \leq \quad & (G_3 n^{H_2})^{k_n} \int_{\|\boldsymbol{\theta}\|_{2,k_n} > n^Q} \exp(-G_4 \sum_{j=1}^{k_n} |\theta_j|^\alpha/\tau_j^\alpha)\, d\theta_i, \end{aligned} \tag{20}$$

by using the lower bound on the τ_j's of Equation (9).

If $\alpha \geq 2$, then applying Hölder inequality, one obtains

$$n^{2Q} \leq \|\boldsymbol{\theta}\|_{2,k_n}^2 \leq \|\boldsymbol{\theta}\|_{\alpha,k_n}^2 k_n^{1-2/\alpha},$$

which leads to

$$\|\boldsymbol{\theta}\|_{\alpha,k_n}^\alpha \geq k_n^{1-\alpha/2} n^{Q\alpha}.$$

If $\alpha < 2$, then a classical result states that the l^α norm $\|\,.\,\|_\alpha$ is larger than the l^2 norm $\|\,.\,\|_2$, *i.e.*

$$\|\boldsymbol{\theta}\|_{\alpha,k_n}^\alpha \geq \|\boldsymbol{\theta}\|_{2,k_n}^\alpha \geq n^{Q\alpha}.$$

Eventually the upper bound τ_0 on the τ_j's of Equation (8) provides

$$\sum_{j=1}^{k_n} |\theta_j|^\alpha/\tau_j^\alpha \geq \tau_0^{-\alpha} n^{Q\alpha} \min(k_n^{1-\alpha/2}, 1).$$

The integral in the right-hand side of (20) is bounded by

$$\exp(-\frac{G_4}{2}\|\boldsymbol{\theta}\|_{2,k_n}^\alpha/\tau_0^\alpha) \int_{\Theta_{k_n}} \exp(-\frac{G_4}{2} \sum_{j=1}^{k_n} |\theta_j|^\alpha/\tau_j^\alpha)\, d\theta_i.$$

The last integral is bounded by C^{k_n}, so

$$\Pi_{k_n}\left(\boldsymbol{\theta} \in \Theta_{k_n} : \|\boldsymbol{\theta}\|_{2,k_n} > n^Q\right) \leq C^{k_n \log n} \exp(-\frac{G_4}{2}\tau_0^{-\alpha} n^{Q\alpha} \min(k_n^{1-\alpha/2}, 1)).$$

The right-hand side of the last inequality can be made smaller than $Ce^{-c_6 n\epsilon_n^2}$ for any constant C and c_6 provided that Q is chosen large enough. This entails result (19).

In the truncated case (3), we note that if $\sum_{j=1}^{k_n} |\theta_j| \leq r_1$, then $\sum_{j=1}^{k_n} \theta_j^2 \leq r_1^2$, so that for n large enough, $\Pi(\Theta_{k_n}^c(Q)) = \pi(k \geq k_n)$, and the rest of the proof is similar. □

Lemma 3. *Under Conditions* $\boldsymbol{A_1}$, $\boldsymbol{A_2}$ *and* $\boldsymbol{A_5}$, *there exists* $c_4 > 0$ *such that*

$$\Pi(\mathcal{B}_n(m)) \geq e^{-c_4 n\epsilon_n^2}.$$

Proof. Let $\boldsymbol{\theta} \in \mathcal{A}_n(H_1)$. For n large enough, Conditions $\boldsymbol{A_1}$ and $\boldsymbol{A_2}$ imply that

$$K(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}) \leq K(p_0^{(n)}, p_{0j_n}^{(n)}) + \widetilde{K}(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)}) \leq 2n\epsilon_n^2,$$

and

$$\begin{aligned} V_{m,0}(p_0^{(n)}, p_{\boldsymbol{\theta}}^{(n)}) &= \int p_0^{(n)} \Big|\log(p_0^{(n)}/p_{0j_n}^{(n)}) - K(p_0^{(n)}, p_{0j_n}^{(n)}) + \\ &\quad \log(p_{0j_n}^{(n)}/p_{\boldsymbol{\theta}}^{(n)}) - \int p_0^{(n)} \log(p_{0j_n}^{(n)}/p_{\boldsymbol{\theta}}^{(n)}) d\mu\Big|^m d\mu \\ &\leq 2^m (V_{m,0}(p_0^{(n)}, p_{0j_n}^{(n)}) + \widetilde{V}_{m,0}(p_{0j_n}^{(n)}, p_{\boldsymbol{\theta}}^{(n)})) \leq 2^{m+1} \left(n\epsilon_n^2\right)^{\frac{m}{2}}, \end{aligned}$$

which yields $\mathcal{A}_n(H_1) \subset \mathcal{B}_n(m)$ so that a lower bound for $\Pi(\mathcal{B}_n(m))$ is given by $\Pi(\mathcal{A}_n(H_1))$. Note that for $H_0 > H_1$, then

$$\mathcal{A}_n(H_0) \subset \mathcal{A}_n(H_1) \subset \mathcal{B}_n(m). \tag{21}$$

We have

$$\Pi(\mathcal{A}_n(H_1)) = \sum_{k=1}^{\infty} \pi(k)\Pi_k(\mathcal{A}_n(H_1)) \geq \pi(j_n)\Pi_{j_n}(\mathcal{A}_n(H_1)).$$

By the first part of Condition (6) we have

$$\pi(j_n) \geq e^{-j_n L(j_n)} \geq e^{-\frac{c_4}{2} n\epsilon_n^2}, \tag{22}$$

for c_4 large enough. Now by the first part of Condition (7) and by Condition (8)

$$\begin{aligned}\Pi_{j_n}(\mathcal{A}_n(H_1)) &= \int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_{0j_n}\|_{2,j_n}\leq n^{-H_1}} \prod_{j=1}^{j_n} g(\theta_j/\tau_j)/\tau_j d\theta_j \qquad (23)\\ &\geq (G_1/\tau_0)^{j_n} \int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_{0j_n}\|_{2,j_n}\leq n^{-H_1}} \exp(-G_2 \sum_{j=1}^{j_n} |\theta_j|^\alpha/\tau_j^\alpha) d\theta_j.\end{aligned}$$

We can bound above $\tau_j^{-\alpha}$ by $n^{\alpha H_2}$ by Equation (9) as $j \leq j_n \leq k_n$. We write $|\theta_j|^\alpha \leq 2^\alpha \left(|\theta_{0j}|^\alpha + |\theta_j - \theta_{0j}|^\alpha\right)$. First, Equation (10) gives

$$\sum_{j=1}^{j_n} |\theta_{0j}|^\alpha / \tau_j^\alpha \leq C j_n \log n.$$

Then, if $\alpha \geq 2$

$$\sum_{j=1}^{j_n} |\theta_j - \theta_{0j}|^\alpha \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_{0j_n}\|_{2,j_n}^\alpha \leq n^{-\alpha H_1},$$

and if $\alpha < 2$ then Hölder's inequality provides

$$\sum_{j=1}^{j_n} |\theta_j - \theta_{0j}|^\alpha \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_{0j_n}\|_{2,j_n}^\alpha \, j_n^{1-\alpha/2} \leq n^{-\alpha H_1} j_n^{1-\alpha/2}.$$

In both cases we have

$$\sum_{j=1}^{j_n} |\theta_j|^\alpha / \tau_j^\alpha \leq 2^\alpha (C j_n \log n + n^{\alpha(H_2-H_1)} j_n^{1-\alpha/2}),$$

so choosing $H_2 \leq H_1$ ensures to bound the latter by $j_n \log n$. Last, the integral of the ball in dimension j_n, centered around $\boldsymbol{\theta}_{0j_n}$, and of radius n^{-H_1}, is at least equal to $e^{-C j_n \log n}$, for some given positive constant C.

Noting that $j_n = \lfloor j_0 n\epsilon_n^2 / \log(n) \rfloor$ and choosing H_1 large enough, which is possible by Equation (21), ensures the existence of $c_4 > 0$ such that $\Pi_{j_n}(\mathcal{A}_n(H_1)) \geq e^{-\frac{c_4}{2} n\epsilon_n^2}$. Combining this with (22) allows to conclude.

In the truncated case (3), we can first choose r_1 larger than $2\sum_{j=1}^{j_n} |\theta_{0j}|$. If $\boldsymbol{\theta} \in \mathcal{A}_n(H_1)$, then $\sum_{j=1}^{j_n} |\theta_j| \leq \sum_{j=1}^{j_n} (|\theta_j - \theta_{0j}| + |\theta_{0j}|) \leq \sqrt{j_n} n^{-H_1} + r_1/2 \leq r_1$ for n and H_1 large enough. So the expression of integral (23) is still valid. □

A.2 Theorem 1

Proof. (of **Theorem 1**)

Express the quantity of interest $\Pi\left(\mathcal{S}_n(M)|X^n\right)$ in terms of N_n, $\widetilde{N_n}$ and D_n defined as follows

$$\frac{\int_{\mathcal{S}_n(M)\cap\Theta_{k_n}(Q)} p_{\boldsymbol{\theta}}^{(n)}/p_{\boldsymbol{\theta}_0}^{(n)} d\Pi(\boldsymbol{\theta}) + \int_{\mathcal{S}_n(M)\cap\Theta_{k_n}^c(Q)} p_{\boldsymbol{\theta}}^{(n)}/p_{\boldsymbol{\theta}_0}^{(n)} d\Pi(\boldsymbol{\theta})}{\int_\Theta p_{\boldsymbol{\theta}}^{(n)}/p_{\boldsymbol{\theta}_0}^{(n)} d\Pi(\boldsymbol{\theta})} := \frac{N_n + \widetilde{N_n}}{D_n}.$$

Denote $\rho_n(c_3) = \exp(-(c_3+1)n\epsilon_n^2)\Pi(\mathcal{B}_n(m))$ for $c_3 > 0$. Introduce $\boldsymbol{\phi}_n$ the test statistic of Lemma 1, and take the expectation of the posterior mass of $\mathcal{S}_n(M)$ as follows

$$\begin{aligned}
&\mathbb{E}_0^{(n)}\left(\frac{N_n+\widetilde{N_n}}{D_n}\left(\boldsymbol{\phi}_n + 1 - \boldsymbol{\phi}_n\right)\left(\mathbb{I}(D_n \le \rho_n(c_3)) + \mathbb{I}(D_n > \rho_n(c_3))\right)\right) \\
&\le \mathbb{E}_0^{(n)}\left(\boldsymbol{\phi}_n\right) + \mathbb{E}_0^{(n)}\left(\frac{N_n+\widetilde{N_n}}{D_n}\left(1-\boldsymbol{\phi}_n\right)\left(\mathbb{I}(D_n \le \rho_n(c_3)) + \mathbb{I}(D_n > \rho_n(c_3))\right)\right) \\
&\le \mathbb{E}_0^{(n)}\left(\boldsymbol{\phi}_n\right) + \mathbb{P}_0^{(n)}\left(D_n \le \rho_n(c_3)\right) + \frac{\mathbb{E}_0^{(n)}\left(N_n\left(1-\boldsymbol{\phi}_n\right)\right) + \mathbb{E}_0^{(n)}(\widetilde{N_n})}{\rho_n(c_3)}. \qquad (24)
\end{aligned}$$

Lemma 10 in Ghosal and van der Vaart (2007) gives $\mathbb{P}_0^{(n)}\left(D_n \le \rho_n(c_3)\right) \lesssim \left(n\epsilon_n^2\right)^{-m/2}$ for every $c_3 > 0$.

Fubini's theorem entails that $\mathbb{E}_0^{(n)}(N_n(1-\boldsymbol{\phi}_n)) \le \sup_{\mathcal{S}_n(M)\cap\Theta_{k_n}(Q)} \mathbb{E}_{\boldsymbol{\theta}}^{(n)}(1-\boldsymbol{\phi}_n)$. Along with $\mathbb{E}_0^{(n)}(\boldsymbol{\phi}_n)$, it is upper bounded in Lemma 1 by $e^{-c_2 M \frac{\log n}{L(n)} n\epsilon_n^2}$.

Lemma 2 implies that $\mathbb{E}_0^{(n)}(\widetilde{N_n}) \le \Pi(\Theta_{k_n}^c(Q)) \le e^{-c_6 n\epsilon_n^2}$ and Lemma 3 yields $\Pi_n(\mathcal{B}_n(m)) \ge e^{-c_4 n\epsilon_n^2}$. Constants c_3 and c_4 are fixed, so we can choose M, M_0 and Q large enough for c_6 to be sufficiently large (see proof of Lemma 2), such that $\min(M\frac{\log n}{L(n)}c_2, c_6) > c_3 + c_4 + 1$. It implies that the third term in Equation (24) is bounded above by $e^{-c_5 n\epsilon_n^2}$ for some positive c_5. Finally,

$$\mathbb{E}_0^{(n)}\Pi\left(\mathcal{S}_n(M)|X^n\right) = \mathcal{O}\left(\left(n\epsilon_n^2\right)^{-m/2}\right) \underset{n\to\infty}{\longrightarrow} 0,$$

since $n\epsilon_n^2 \underset{n\to\infty}{\longrightarrow} \infty$. □

A.3 Proposition 4

The proof of the lower bound in the local risk case uses the next lemma, whose proof follows from Cauchy-Schwarz' inequality.

Lemma 4. *If* $\mathbb{E}(B_n^2) = o(\mathbb{E}(A_n^2))$, *then* $\mathbb{E}((A_n + B_n)^2) = \mathbb{E}(A_n^2)(1 + o(1))$.

Proof. (of **Proposition 4**)

The coordinates of $\hat{\boldsymbol{\theta}}_n$ are $\hat{\theta}_{nj} = \Pi(\theta_j|X^n) = \sum_{k=1}^{\infty} \pi(k|X^n)\widetilde{\theta}_{nj}(k)$, with $\widetilde{\theta}_{nj}(k) = \tau_j^2/(\tau_j^2 + 1/n)X_j^n$ if $k \geq j$, and $\widetilde{\theta}_{nj}(k) = 0$ otherwise (see Zhao, 2000).

Denote $u_j(X^n) = \sum_{k \geq j} \pi(k|X^n) = \pi(k \geq j|X^n)$, so that $\hat{\theta}_{nj} = u_j(X^n)\tau_j^2/(\tau_j^2 + 1/n)X_j^n$. Denote $K_n = n^{1/(2\beta+1)}$ and $J_n = n^{1/2\beta}$. Most of the posterior mass on k is concentrated before K_n, in the sense that there exists a constant c such that

$$\mathbb{E}_0^{(n)}\left(u_{K_n}(X^n)\right) \lesssim \exp\left(-cK_n\right). \tag{25}$$

This follows from the exponential inequality

$$P_{\boldsymbol{\theta}_0}^{(n)}[u_{K_n}(X^n) > \exp(-cK_n)] \lesssim \exp(-cK_n),$$

which is obtained by classic arguments in line with Theorem 1: writing the posterior quantity $u_{K_n}(X^n)$ as a ratio N_n/D_n, and then using Fubini's theorem, Chebyshev's inequality and an upper bound on $\pi(k > K_n)$.

Due to Relation (17), we split in three the sum in the risk

$$R_n^{\mathrm{loc}}(\boldsymbol{\theta}_0, t) = \mathbb{E}_0^{(n)}\left(\sum_{i=1}^{\infty} a_i[(1 - u_i(X^n)\tfrac{\tau_i^2}{\tau_i^2+1/n})\theta_{0i} - u_i(X^n)\tfrac{\tau_i^2}{\tau_i^2+1/n}\tfrac{\xi_i}{\sqrt{n}}]\right)^2$$

by centring the stochastic term X_i^n and writing $1 - u_i(X^n)\frac{\tau_i^2}{\tau_i^2+1/n} = \frac{1}{n}\frac{1}{\tau_i^2+1/n} + \frac{\tau_i^2}{\tau_i^2+1/n}(1 - u_i(X^n))$. The idea of the proof is to show that there is a leading term in the sum, and to apply Lemma 4.

Let $R_1 = \left(\sum_{i=1}^{\infty} a_i \frac{1}{n\tau_i^2+1}\theta_{0i}\right)^2$, $R_2 = \mathbb{E}_0^{(n)}\left(\sum_{i=1}^{\infty} a_i \frac{\tau_i^2}{\tau_i^2+1/n}(1 - u_i(X^n))\theta_{0i}\right)^2$ and $R_3 = \mathbb{E}_0^{(n)}\left(\sum_{i=1}^{\infty} a_i \frac{\tau_i^2}{\tau_i^2+1/n}u_i(X^n)\frac{\xi_i}{\sqrt{n}}\right)^2$. By using Cauchy-Schwarz' inequal-

ity

$$R_1 = \left(\sum_{i=1}^{\infty} a_i \tfrac{1}{n\tau_i^2+1}\theta_{0i}\right)^2 = \left(\sum_{i=1}^{\infty} a_i \tfrac{i^{-\beta}}{n\tau_i^2+1}\theta_{0i} i^{\beta}\right)^2 \lesssim L_0 \sum_{i=1}^{\infty} \tfrac{i^{-2\beta}}{(ni^{-2q}+1)^2},$$

because the a_i's are bounded. If $2\beta - 4q > 1$, then we can write

$$R_1 \lesssim \frac{1}{n^2}\sum_{i=1}^{\infty} i^{-2\beta+4q} \lesssim \frac{1}{n^2},$$

and if $2\beta - 4q \leq 1$, then comparing to an integral provides

$$\begin{aligned} R_1 &\lesssim \int_1^{\infty} \tfrac{x^{-2\beta}}{(nx^{-2q}+1)^2}dx \\ &\lesssim \left(n^{1/2q}\right)^{1-2\beta} \int_{n^{-1/2q}}^{\infty} \tfrac{y^{-2\beta}}{(y^{-2q}+1)^2}dy \\ &\lesssim n^{-\frac{2\beta-1}{2q}} \lesssim n^{-\frac{2\beta-1}{2\beta}}, \end{aligned}$$

where the last inequality holds because q is chosen such that $q \leq \beta$. Then $R_1 = \mathcal{O}(n^{-(2\beta-1)/2\beta})$.

For $k = 2, 3$, denote $R_k(b_n, c_n)$ the partial sum of R_k from $j = b_n$ to c_n. Then $R_2(1, J_n)$ is the larger term in the decomposition, and is treated at the end of the section. The upper part $R_2(J_n, \infty)$ is easily bounded by

$$R_2(J_n, \infty) \lesssim \left(\sum_{i=J_n}^{\infty} |\theta_{0i}|\, i^{\beta} i^{-\beta}\right)^2 \lesssim J_n^{-2\beta+1} = \mathcal{O}\left(n^{-\frac{2\beta-1}{2\beta}}\right).$$

We split $R_3(1, J_n)$ in two parts $R_{3,1}(1, J_n)$ and $R_{3,2}(1, J_n)$ by writing $u_i(X^n) = u_{J_n}(X^n) + \pi(i \leq k < J_n | X^n)$ for all $i \leq J_n$:

$$\begin{aligned} nR_3(1, J_n) &\lesssim \mathbb{E}_0^{(n)}\left(\sum_{j=1}^{J_n} \pi(j|X^n) \sum_{i=1}^{j} a_i \tfrac{\tau_i^2}{\tau_i^2+1/n}\xi_i\right)^2 \\ &\quad + \mathbb{E}_0^{(n)}\left(u_{J_n}(X^n) \sum_{i=1}^{J_n} a_i \tfrac{\tau_i^2}{\tau_i^2+1/n}\xi_i\right)^2 \\ &:= R_{3,1}(1, J_n) + R_{3,2}(1, J_n). \end{aligned}$$

Let $\Gamma_{jn}(X^n) = \sum_{i=1}^{j} a_i \frac{\tau_i^2}{\tau_i^2 + 1/n} \xi_i$. We have $\sum_{j=1}^{J_n} \pi(j|X^n) \leq 1$ so we can apply Jensen's inequality,

$$\begin{aligned} R_{3,1}(1, J_n) &\leq \mathbb{E}_0^{(n)} \left(\sum_{j=1}^{J_n} \pi(j|X^n) \Gamma_{jn}(X^n)^2 \right) \\ &\leq \mathbb{E}_0^{(n)} \max_{j \leq J_n} \left\{ \Gamma_{jn}(X^n)^2 \right\}. \end{aligned}$$

Noting that $(\Gamma_{jn}(X^n))_{1 \leq j \leq J_n}$ is a martingale, we get using Doob's inequality

$$R_{3,1}(1, J_n) \leq \mathbb{E}_0^{(n)} \Gamma_{J_n n}(X^n)^2 = \sum_{i=1}^{J_n} \left(a_i \frac{\tau_i^2}{\tau_i^2 + 1/n} \right)^2 \lesssim J_n.$$

The second term $R_{3,2}(1, J_n)$ can be upper bounded in the same way as for $R_3(J_n, \infty)$ in Equation (26) below by noting that

$$R_{3,2}(1, J_n) \lesssim \mathbb{E}_0^{(n)} \left[u_{J_n}(X^n)^2 \left(\sum_{i=K_n}^{\infty} \frac{\tau_i^2}{\tau_i^2 + 1/n} |\xi_i| \right)^2 \right].$$

For the upper part $R_3(J_n, \infty)$, we use the bound (25) on $\mathbb{E}_0^{(n)} \left(u_{K_n}(X^n) \right)$,

$$\begin{aligned} n R_3(J_n, \infty) &\lesssim \mathbb{E}_0^{(n)} \left(\sum_{i=K_n}^{\infty} \frac{\tau_i^2}{\tau_i^2 + 1/n} u_i(X^n) |\xi_i| \right)^2 \\ &\lesssim \mathbb{E}_0^{(n)} \left[u_{K_n}(X^n)^2 \left(\sum_{i=K_n}^{\infty} \frac{\tau_i^2}{\tau_i^2 + 1/n} |\xi_i| \right)^2 \right] \\ &\lesssim \left[\mathbb{E}_0^{(n)} u_{K_n}(X^n)^4 \right]^{1/2} \left[\mathbb{E}_0^{(n)} \left(\sum_{i=K_n}^{\infty} \frac{\tau_i^2}{\tau_i^2 + 1/n} |\xi_i| \right)^4 \right]^{1/2} \\ &\lesssim \left[\mathbb{E}_0^{(n)} u_{K_n}(X^n) \right]^{1/2} \left[\left(\sum_{i=K_n}^{\infty} \frac{\tau_i^2}{\tau_i^2 + 1/n} \right)^4 \right]^{1/2} \\ &\lesssim e^{-c_2 K_n / 2} n^{1/q}, \end{aligned} \tag{26}$$

where we bound the different moments of $|\xi_i|$ by a unique constant and then use $\sum_{i=K_n}^{\infty} \tau_i^2 / (\tau_i^2 + 1/n) = \mathcal{O}(n^{1/2q})$. Then $R_3 = \mathcal{O}(n^{-(2\beta-1)/2\beta})$.

To sum up, $R_2(1, J_n)$ is the only remaining term. We build an example where it is of greater order than $n^{-(2\beta-1)/2\beta}$. Let $\boldsymbol{\theta}_0$ be defined by its coordinates $\theta_{0i} =$

$i^{-\beta-1/2}\left(\log(i+1)\right)^{-1}$ such that the series $\sum_i \theta_{0i}^2 i^{2\beta}$ converge, so $\boldsymbol{\theta}_0$ belongs to the Sobolev ball of smoothness β. It is assumed that $a_i = \psi_i(t) = 1$, so all terms in the sum $R_2(1, J_n)$ are *positive*, hence

$$R_2(1, J_n) \geq \frac{1}{4}\mathbb{E}_0^{(n)}\left(\sum_{i=K_n}^{J_n}(1 - u_i(X^n))\theta_{0i}\right)^2,$$

noting that for $i \leq J_n$, we have $n\tau_i^2 \geq n^{1-q/\beta} \geq 1$ because $q \leq \beta$ and $n \geq 1$, so $\tau_i^2/(\tau_i^2 + 1/n) \geq 1/2$. Moreover, $u_i(X^n)$ decreases with i, so

$$R_2(1, J_n) \geq \frac{1}{4}\mathbb{E}_0^{(n)}\left((1 - u_{K_n}(X^n))^2\right)\left(\sum_{i=K_n}^{J_n}\theta_{0i}\right)^2,$$

where $\mathbb{E}_0^{(n)}\left((1 - u_{K_n}(X^n))^2\right)$ is lower bounded by a positive constant for n large enough. Comparing the series $\sum_{i=K_n}^{J_n}\theta_{0i}$ to an integral shows that it is bounded from below by $K_n^{-\beta+1/2}/\log n$. We obtain by using Lemma 4 that $R_n^{\mathrm{loc}}(\boldsymbol{\theta}_0, t) = R_2(1, J_n)(1 + o(1)) \gtrsim n^{-\frac{2\beta-1}{2\beta+1}}/\log^2 n$, which ends the proof. □