Electronic Journal of Statistics
Vol. 7 (2013) 991–1018
ISSN: 1935-7524
DOI: 10.1214/13-EJS798

Empirical Bayes scaling of Gaussian priors in the white noise model

B. T. Szabó,

Department of Mathematics
Eindhoven University of Technology
e-mail: b.szabo@tue.nl

A. W. van der Vaart

Mathematical Institute
Leiden University
e-mail: avdvaart@math.leidenuniv.nl

and

J. H. van Zanten

Korteweg-de Vries institute for Mathematics
University of Amsterdam
e-mail: j.h.vanzanten@uva.nl

Abstract: The performance of nonparametric estimators is heavily dependent on a bandwidth parameter. In nonparametric Bayesian methods this parameter can be specified as a hyperparameter of the nonparametric prior. The value of this hyperparameter may be made dependent on the data. The empirical Bayes method is to set its value by maximizing the marginal likelihood of the data in the Bayesian framework. In this paper we analyze a particular version of this method, common in practice, that the hyperparameter scales the prior variance. We characterize the behavior of the random hyperparameter, and show that a nonparametric Bayes method using it gives optimal recovery over a scale of regularity classes. This scale is limited, however, by the regularity of the unscaled prior. While a prior can be scaled up to make it appropriate for arbitrarily rough truths, scaling cannot increase the nominal smoothness by much. Surprisingy the standard empirical Bayes method is even more limited in this respect than an oracle, deterministic scaling method. The same can be said for the hierarchical Bayes method.

AMS 2000 subject classifications: Primary 62G05, 62G15; secondary 62G20.
Keywords and phrases: Adaptation, hyper-rectangle, Gaussian white noise, normal means model, bandwidth, rate of contraction.

Received December 2012.

1. Introduction

Recent years have seen increasing use of Bayesian methods in high-dimensional or nonparametric statistical problems. It is known from both theory (e.g. [13,

*Research supported by Netherlands Organization for Scientific Research NWO.

15, 5]) and practice that the (asymptotic) performance of such methods is sensitive to the fine properties of the prior that is employed. This dependence can be alleviated by adapting the prior to the data through one or more tuning parameters, so-called *hyperparameters*. In the case of function estimation such parameters can for instance describe the degree of regularity of a prior, a length scale, or a bandwidth.

Two tuning methods are widely used. The first is to endow the hyperparameters with a prior distribution, and leads to fully Bayesian procedures, referred to as *hierarchical Bayes*. The frequentist behavior of such methods has been studied in e.g. [2, 14, 20, 23, 27], where it was found that, if the priors are well chosen, they can yield adaptive, rate-optimal recovery for a range of nonparametric statistical problems. A second possible approach is to estimate the hyperparameters from the data, e.g. by using a likelihood-based method. This approach is not fully Bayesian, and commonly called *empirical Bayes*, but is often computationally convenient and therefore commonly used in practice.

The theoretical performance of empirical Bayes methods in nonparametric problems has been studied in only a limited number of special cases, see for instance [1, 17]. Because a general understanding of such methods appears difficult at this time, in this paper we focus on the important case that the hyperparameter is a scale parameter of a Gaussian prior. This situation was first considered in work on spline smoothing (see [29]), where the posterior mean for a (multiply integrated, scaled and released) Brownian motion as a prior for an unknown function is a penalized least squares estimator, and choosing the scale parameter of the prior is equivalent to choosing the smoothing parameter (that multiplies the penalty).

We consider the scaled Gaussian priors in the particular case of the Gaussian white noise model, which allows tractable formulas. In view of the close relation between this model and many other nonparametric models, it is expected that our findings generalize. However, since we deliberately consider a particular method, this does not follow from general results on equivalence of experiments and thus will require further investigation.

The term *empirical Bayes* is used in various ways (see [22, 10, 30, 16] for the original and alternative uses). In our situation it means determining a suitable value of a (scaling) parameter of a prior from the data, which could still refer to different methods. Specifically, we study the maximum likelihood estimator (MLE) for the scale parameter based on the *marginal Bayesian likelihood* (see (2.5) below). This is a natural method, which attempts to take the best of both worlds. The method is also of interest by its close relation to the "full" (hierarchical Bayes) method. These two methods differ only in that empirical Bayes takes the MLE for the (univariate) marginal Bayesian likelihood, whereas hierarchical Bayes equips the (univariate) parameter of this marginal likelihood with a prior. Within our framework these methods perform equivalently, as we show in Section 2.3.

We investigate the behavior of the empirical Bayes method in a frequentist set-up: the method is (empirical) Bayesian, but it is evaluated under the assumption that the data are generated under a given "true" parameter. In this

situation minimax optimal rates can be used as a benchmark for performance. However, it is not our primary aim to construct minimax estimators, or even to exhibit priors that lead to minimax posterior means. Rather the particular (scaled) priors and specific likelihood-based empirical Bayes method for choosing the scaling parameter are the starting points. We aim at establishing their performance, as they are natural and widely applied choices. For the aim of minimax estimation there are various other methods (see e.g. [4, 21, 9]).

The results of this paper are a step towards a more general understanding of empirical Bayes methods. They concern the behaviour of the empirical Bayes scaling parameter and contraction of resulting plug-in posterior distribution. We study contraction of the full posterior distribution rather than a summary measure, such as a posterior mean. The full posterior is important for the use of the Bayesian method for uncertainty quantification, for instance through *credible sets*: sets of prescribed posterior probability. We hope to report on this involved issue in a future paper. Understanding the behaviour of the empirical Bayes scaling parameter will also be essential in this investigation.

In an earlier paper [19] we considered the performance of posterior distributions based on the same priors, but with *deterministic scaling*. It turned out that for a given base prior and a given true regularity level there is an *optimal scaling rate*. It is natural to compare the empirical Bayes method, which gives a data-dependent rate, to the performance with this optimal rate, which would be available to an *oracle*. Here we found the following somewhat surprising result. While it is known that the oracle procedure fails to be minimax if the regularity of the true parameter is higher than a level dependent on the unscaled prior (see [25] and the next section), it turns out that the empirical Bayes method fails to follow the oracle if the regularity of the true parameter exceeds an even lower bound. This finding may motivate the investigation of different empirical Bayes schemes. On the positive side our results show that empirical Bayes works adequately if the base prior does not (or only little) undersmooth the true parameter.

In the next section we give a precise description of the problem, and state our main findings. In Section 3 we illustrate the results with some simulations and pictures. Sections 4 and 5 contain the proofs.

We write $a \lesssim b$ for $a \leq Cb$ for a constant C that is universal or fixed in the context, and $a_n \asymp b_n$ if $a_n/b_n \to 1$.

2. Main results

2.1. Setup

To be able to derive concrete results we consider a relatively tractable nonparametric model: the Gaussian sequence model, or, equivalently the signal-in-white-noise model, and sometimes called the normal means model. This model often serves as a platform to investigate the behavior of statistical procedures, see for instance [7, 24, 2, 5, 12, 6] for studies on various aspects of non- and over-smoothing procedures in this setting.

We assume we observe a sequence $X = (X_1, X_2, \ldots)$ satisfying

$$X_i = \theta_{0,i} + \frac{1}{\sqrt{n}} Z_i, \qquad i = 1, 2, \ldots, \tag{2.1}$$

for $\theta_0 = (\theta_{0,1}, \theta_{0,2}, \ldots)$ an unknown element of $\ell^2 = \{\theta \in \mathbb{R}^\infty : \|\theta\|^2 = \sum_k \theta_k^2 < \infty\}$ and $Z_1, Z_2, \ldots$ independent, standard normal random variables. We denote the "true" distribution of X by P_0 and the corresponding expectations by E_0. All results refer to this distribution, although in the next paragraphs we adopt a Bayesian point of view in which the parameter is random to motivate the posterior distribution and the empirical Bayes likelihood.

This model is equivalent to the *signal-in-white-noise model*, in which we observe the process $(Y_t : 0 \le t \le 1)$ given by

$$Y_t = \int_0^t f_0(s)\, ds + \frac{1}{\sqrt{n}} W_t, \qquad t \in [0, 1],$$

with $f_0 \in L^2[0,1]$ an unknown function and W a standard Brownian motion. Indeed, if e_i is an orthonormal basis of $L^2[0,1]$, then the variables $X_i = \int_0^1 e_i(s)\, dY_s$ satisfy (2.1), with $\theta_{0,i} = \langle f_0, e_i \rangle$ the Fourier coefficients of f_0 relative to the basis e_i. In Section 3 we illustrate our findings by simulated data in this setting.

The variance of the errors in (2.1) is taken equal to the known value $1/n$. It is clear from the signal-in-white-noise representation that a possible parameter σ^2, changing the variance in σ^2/n, would be 'estimable' without error from the data, e.g. by n times the quadratic variation $[Y]_1$ of the signal $(Y_t : 0 \le t \le 1)$. Thus it is no loss of generality not to introduce such an additional parameter in the model; taking it equal to unity simplifies the notation. A different situation would arise, were the signal observed only on a discrete time set. We guess that similar phenomena will occur in this different model, but to verify this will require significant additional technical work. Including an additional variance parameter would be natural in this work.

We assume that the parameter θ_0 belongs to a hyper rectangle in ℓ^2, i.e.

$$\theta_{0,k}^2 \le C^2 k^{-1-2\beta}, \tag{2.2}$$

for (unknown) constants and $C, \beta > 0$. In the case that the $\theta_{0,k}$'s are the Fourier coefficients of some unknown function, this roughly means assuming that the function has "regularity" of the order β. It is known that the minimax rate of estimation relative to the ℓ^2-norm over hyper rectangles of this form is of the order $n^{-\beta/(1+2\beta)}$ (see [8]).

In the Bayesian set-up the model (2.1) is viewed as giving the conditional distribution of X given the parameter θ_0, inference on the unknown parameter θ_0 begins by postulating a prior distribution for θ_0. We consider the family of priors

$$\Pi_\tau = \bigotimes_{k=1}^{\infty} N(0, \tau^2 k^{-1-2\alpha}) \tag{2.3}$$

on $\mathbb{R}^\infty$, where $\alpha > 0$ is a fixed parameter and $\tau > 0$ is a scaling parameter that will be set by an empirical Bayes approach. In other words, under the prior Π_τ the coordinates $\theta_{0,k}$ of θ_0 are independent, centered Gaussian variables with variances $\tau^2 k^{-1-2\alpha}$. The parameter α determines the speed at which the variances tend to zero. It can be interpreted as the baseline "regularity" of the unscaled prior. Indeed, for fixed $\tau > 0$ and any $s < \alpha$, the prior Π_τ gives full mass to the Sobolev space $H^s = \{\theta \in \ell^2 : \sum_k \theta_k^2 k^{2s} < \infty\}$.

In this paper we stick to this prior. The fact that the prior does give mass zero to the Sobolev space of order α motivated [31] (also see [3]) to consider various modifications, such as block dependent priors. Another alternative would be to mix priors of the form (2.3) over the value of α. Estimating α by empirical or hierarchical Bayes (with $\tau = 1$ fixed) is considered in [18].

Under the (conditional) model (2.1) and the prior (2.3) the coordinates $(\theta_{0,k}, X_k)$ of the vector (θ_0, X) are independent, and hence the conditional distribution of θ_0 given X factorizes over the coordinates as well. Thus the computation of the posterior distribution reduces to countably many posterior computations in conjugate normal models. It is straightforward to verify that the posterior distribution $\Pi_\tau(\cdot \,|\, X)$ is given by

$$\Pi_\tau(\cdot \,|\, X) = \bigotimes_{k=1}^{\infty} N\Big(\frac{n\tau^2}{n\tau^2 + k^{1+2\alpha}} X_k, \frac{\tau^2}{n\tau^2 + k^{1+2\alpha}}\Big). \tag{2.4}$$

In the empirical Bayes approach we subsequently replace the hyperparameter τ by a data-driven choice $\hat{\tau}_n$. In the Bayesian setting described by the conditional distributions $\theta \,|\, \tau \sim \Pi_\tau$ and $X \,|\, (\theta, \tau) \sim \otimes_k N(\theta_k, 1/n)$, it holds that

$$X \,|\, \tau \sim \bigotimes_{k=1}^{\infty} N(0, \tau^2 k^{-1-2\alpha} + 1/n).$$

The corresponding log-likelihood for τ (relative to an infinite product of $N(0, 1/n)$-distributions) is given by

$$\ell_n(\tau) = -\frac{1}{2} \sum_{k=1}^{\infty} \Big(\log\Big(1 + \frac{\tau^2 n}{k^{1+2\alpha}}\Big) - \frac{\tau^2 n^2}{k^{1+2\alpha} + \tau^2 n} X_k^2 \Big). \tag{2.5}$$

We shall prove that with P_0-probability going to one, ℓ_n attains a global maximum on $(0, \infty)$, and denote the point where this is attained by $\hat{\tau}_n$. (If the point of global maximum is not unique, any global maximum can be chosen.) Outside the event on which ℓ_n has a global maximum, $\hat{\tau}_n$ can be set to an arbitrary value.

The *empirical Bayes posterior* is now defined as the random measure $\Pi_{\hat{\tau}_n}(\cdot|X)$ obtained by substituting $\hat{\tau}_n$ for τ in the posterior distribution (2.4), i.e.

$$\Pi_{\hat{\tau}_n}(B|X) = \Pi_\tau(B|X)\Big|_{\tau=\hat{\tau}_n}$$

for measurable subsets $B \subset \ell^2$. The results presented in the next subsection concern the rate at which the empirical Bayes posterior contracts to the true parameter θ_0. Furthermore, we characterize the behavior of $\hat{\tau}_n$ itself.

If the true parameter satisfies next to (2.2) also the reverse inequality (with a constant $c \leq C$), then it turns out that $\hat{\tau}_n$ has a precise behavior, and the performance of the posterior can be established by uniformity arguments. The more difficult case is to consider $\hat{\tau}_n$ and $\Pi_{\hat{\tau}_n}(\cdot|X)$ under general θ_0 in the rectangle described by (2.2).

2.2. Main results

If the prior is not rescaled, i.e. we use the prior Π_τ for some fixed value of τ, then the posterior (2.4) contracts to θ_0 at the optimal rate $n^{-\beta/(1+2\beta)}$ if and only if $\alpha = \beta$ (cf. [26, 5, 19, 11]). That is, the Bayesian procedure performs optimally if and only if the "regularities" of the prior and the unknown parameter match.

This relationship changes if the parameter $\tau = \tau_n$ is chosen to tend to zero or infinity with n. Two situations arise: if the prior does not under-smooth the unknown parameter too much, then the optimal rate can still be attained, whereas in the other case the posterior gives suboptimal recovery no matter the scaling ([19, 25]). More precisely,

(i) If θ_0 satisfies (2.2) for $\beta \leq 1+2\alpha$, then for the choice $\tau = \tau_n = n^{(\alpha-\beta)/(1+2\beta)}$, and every $M_n \to \infty$,

$$\Pi_{\tau_n}(\theta : \|\theta - \theta_0\|_2 > M_n n^{-\frac{\beta}{1+2\beta}} \mid X) \overset{\mathrm{P}_0}{\to} 0.$$

(ii) If $\beta > 1 + 2\alpha$, then this posterior probability tends to 1 for some θ_0 satisfying (2.2).

The optimal rescaling rate $\tau_n = n^{(\alpha-\beta)/(1+2\beta)}$ in case (i) depends on the unknown parameter β that measures the smoothness of the true parameter. We therefore call it the *oracle rescaling rate*. Our aim is to compare the performance of the empirical Bayes procedure to that of the oracle procedure.

Remarkably, the performance of the empirical Bayes procedure cuts the range $\beta \leq 1+2\alpha$, where optimal deterministic scaling is possible, into two subregimes. If $\beta < 1/2+\alpha$, then the empirical Bayes posterior matches the oracle procedure and contracts at the optimal rate $n^{-\beta/(2\beta+1)}$ to θ_0. On the other hand, if $1/2+\alpha \leq \beta < 1+2\alpha$, then the empirical Bayes procedure performs strictly worse than the oracle. The message is that smooth priors perform well from the perspective of contraction rates; if empirical Bayes scaling is used, then a good prior should under-smooth the truth by at most $1/2$ level of regularity.

Besides the empirical Bayes posterior, we study the empirical Bayes rescaling rate $\hat{\tau}_n$ itself. In our first theorem we give upper and lower bounds for its magnitude. For given nonzero θ_0 consider the functions $h_n : (0,\infty) \to (0,\infty)$ defined by

$$h_n(\tau) = \sum_{k=1}^{\infty} \frac{(\tau^2 n)^{\frac{2\alpha}{1+2\alpha}} k^{1+2\alpha} \theta_{0,k}^2}{(k^{1+2\alpha} + \tau^2 n)^2}. \tag{2.6}$$

For fixed n the function h_n is positive on $(0,\infty)$ and tends to zero as $\tau \to \infty$, by dominated convergence, for any nonzero $\theta_0 \in \ell_2$. Therefore, for positive

constants $l < L$ we can define

$$\overline{\tau}_n = \sup\Big\{\tau > 0 : h_n(\tau) \geq l/n\Big\}, \tag{2.7}$$

$$\underline{\tau}_n = \sup\Big\{\tau > 0 : h_n(\tau) \geq L/n\Big\}. \tag{2.8}$$

In the next theorem we show that $\hat{\tau}_n$ belongs with probability tending to one to the interval $[\underline{\tau}_n, \overline{\tau}_n]$, provided l is chosen sufficiently small and L sufficiently big.

The function h_n and the bounds $\underline{\tau}_n \leq \overline{\tau}_n$ depend on the unknown true parameter θ_0. For typical θ_0 the upper and lower bounds have the same order of magnitude. In particular, this is true for θ_0 satisfying the exact asymptotic behavior $\theta_{0,k}^2 \asymp C^2 k^{-1-2\beta}$, in which case, for some constants d depending on α, β, C and l (see Section 4.4),

$$\overline{\tau}_n \asymp \begin{cases} d\, n^{\frac{\alpha-\beta}{1+2\beta}}, & \text{if } \beta < \alpha + 1/2, \\ d\, n^{\frac{-1}{4+4\alpha}} (\log n)^{\alpha/(4+4\alpha)}, & \text{if } \beta = \alpha + 1/2, \\ d\, n^{\frac{-1}{4+4\alpha}}, & \text{if } \beta > \alpha + 1/2. \end{cases} \tag{2.9}$$

The cut-off at $\beta = \alpha + 1/2$ is clearly visible in this bound.

The exact asymptotic behavior $\theta_{0,k}^2 \asymp C^2 k^{-1-2\beta}$ may be considered a worst case for θ_0 belonging to the hyper rectangle (2.2). For general θ_0 that are not "in the boundary" of any rectangle (for any β), the behaviour of $\hat{\tau}_n$, may be complicated, but we shall see in (the proof of) Theorem 2.2 that the lower and upper bounds $\underline{\tau}_n$ and $\overline{\tau}_n$ are sufficiently sharp to analyze the behaviour of the empirical Bayes posterior distribution of θ.

Theorem 2.1. *Suppose (2.2) holds. If $\theta_0 \neq 0$, then the constants l and L in (2.7) and (2.8) can be chosen such that $\mathrm{P}_0(\underline{\tau}_n < \hat{\tau}_n < \overline{\tau}_n) \to 1$. If in addition $\theta_{0,k}^2 \geq c^2 k^{-1-2\beta}$ for some $c > 0$, then $\underline{\tau}_n$ and $\overline{\tau}_n$ are of the same order. Moreover, if $\theta_{0,k}^2 \asymp C^2 k^{-1-2\beta}$, then $\hat{\tau}_n/\overline{\tau}_n$ tends in probability to a constant, and $\overline{\tau}_n$ satisfies (2.9). Finally, if $\theta_0 = 0$, then $\hat{\tau}_n = O_P(1/\sqrt{n})$.*

The worst case upper bound $\overline{\tau}_n$ in (2.9) has the same order as the optimal rescaling rate $n^{(\alpha-\beta)/(1+2\beta)}$ if $\beta < 1/2 + \alpha$, but not if $\beta \geq 1/2 + \alpha$. The theorem shows that the empirical Bayes procedure selects the common order whenever the lower and upper bounds have the same order, in particular when $\theta_{0,k}^2 \asymp C^2 k^{-1-2\beta}$. Hence in the latter case the empirical Bayes procedure selects the proper oracle scaling rate if $\beta < \alpha + 1/2$, but not in the other case, i.e. only if the baseline "regularity" α of the prior is sufficiently large compared to the regularity β of the truth θ_0. This suggests that the empirical Bayes posterior will match the oracle only in the case $\beta < \alpha + 1/2$, and performs sub-optimally if $\beta \geq \alpha + 1/2$. The following theorem, which is the main result of this paper, states that this is true under the general assumption (2.2).

Theorem 2.2. *If θ_0 satisfies (2.2), then*

$$\Pi_{\hat{\tau}_n}(\theta : \|\theta - \theta_0\|_2 \leq M_n \varepsilon_{n,\alpha,\beta} \,|\, X) \xrightarrow{\mathrm{P}_0} 1,$$

for every sequence $M_n \to \infty$, *where*

$$\varepsilon_{n,\alpha,\beta} = \begin{cases} n^{-\beta/(1+2\beta)}, & \text{if } \beta < 1/2+\alpha, \\ n^{-\beta/(1+2\beta)}(\log n)^{(1/2)/(1+2\beta)}, & \text{if } \beta = 1/2+\alpha, \\ n^{-(1/2+\alpha)/(2+2\alpha)}, & \text{if } \beta > 1/2+\alpha. \end{cases}$$

Furthermore, if in addition $\theta_{0,k}^2 \geq c^2 k^{-1-2\beta}$ *for some* $c > 0$, *then, for all sufficiently small* $m > 0$,

$$\Pi_{\hat{\tau}_n}(\theta : \|\theta - \theta_0\|_2 < m\varepsilon_{n,\alpha,\beta} | X) \xrightarrow{P_0} 0.$$

Finally, if $\beta > 1/2+\alpha$, *this is true for any* $\theta_0 \neq 0$ *that satisfies (2.2).*

The first assertion of the theorem shows that the empirical Bayes procedure attains the optimal rate if $\beta < 1/2+\alpha$, but a slower rate in the other cases. The rate $n^{-(1/2+\alpha)/(2+2\alpha)}$ in the case that $\beta > 1/2+\alpha$ is the optimal rate for the value $\beta = 1/2+\alpha$ at the cut point. If (2.2) holds for some $\beta > 1/2+\alpha$, then it also holds for $\beta = 1/2+\alpha$. Therefore an interpretation is that the empirical Bayes procedure with a prior of regularity α is incapable to exploit regularity (2.2) in the true function θ_0 beyond level $1/2+\alpha$. The second and third assertions of the theorem show that the rates are sharp. The third, final assertion shows in a very strong sense that the deterioration of the rate in the third case is caused completely by the prior.

The good news is that the empirical Bayes procedure repairs any amount of prior over-smoothing, at least as far as contraction rates are concerned.

2.3. Hierarchical Bayes

Instead of substituting a random value for $\hat{\tau}_n$ into the posterior distribution for θ, the hierarchical Bayes approach models τ with a prior distribution λ, and next performs a full Bayes analysis with the mixture prior $\int_0^\infty \Pi_\tau \, d\lambda(\tau)$ on θ. Here Π_τ is the prior on θ with scale τ, as given in (2.3). Besides a posterior distribution on θ, this also yields a posterior distribution for τ, which can be written in the form

$$\Pi(\tau \in B \,|\, X) = \frac{\int_B e^{\ell_n(\tau)} \, d\lambda(\tau)}{\int e^{\ell_n(\tau)} \, d\lambda(\tau)},$$

for ℓ_n the marginal log likelihood of X given τ, given in (2.5). By definition the empirical Bayes value $\hat{\tau}_n$ is the point of maximum of the integrand in the integrals on the right. Thus the two methods are closely related. The link is made formal in the following theorem, which implies that the hierarchical Bayes method copies both the good *and* the bad behavour (suboptimality if $\beta \geq \alpha + 1/2$) of the empirical Bayes method.

We restrict to the inverse Gamma distribution as a prior for τ^2. Inspection of the proof shows that the theorem goes through for many other priors λ. Define $\underline{\tau}_n$ and $\overline{\tau}_n$ as before by (2.7) and (2.8), where the constant L in (2.8) is chosen sufficiently large.

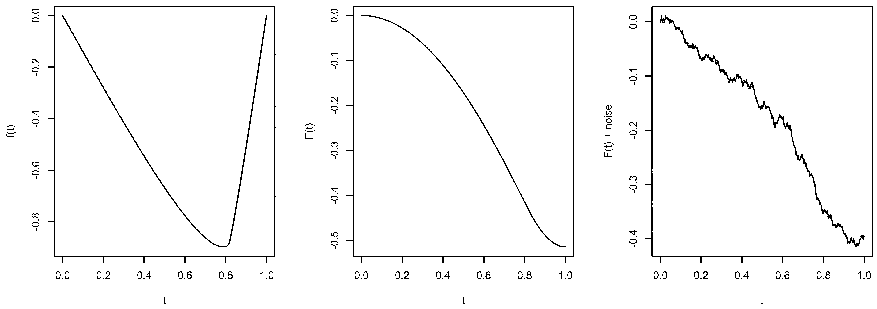

FIG 1. *The true function, its primitive function, and the noisy observation of the primitive function.*

Theorem 2.3. *If* $1/\tau^2 \sim \Gamma(a,b)$ *for some constants* $a,b>0$ *and* $\theta_0 \neq 0$ *satisfies (2.2), then, for sufficiently large* M *(and* L *in (2.8)),*

$$\Pi\big(\underline{\tau}_n/2 \leq \tau \leq M\overline{\tau}_n \,|\, X\big) \stackrel{\mathrm{P}_0}{\to} 1.$$

As a consequence the posterior distribution of θ *relative to the prior* $\int_0^\infty \Pi_\tau \, d\lambda(\tau)$ *has the same properties as* $\Pi_{\hat{\tau}_n}(\cdot \,|\, X)$ *given in Theorem 2.2.*

3. Some Simulation Results

To illustrate the main results we simulated data from the signal-in-white-noise model

$$dY_t = f_0(t)\,dt + \frac{1}{\sqrt{n}} dW_t, \qquad t \in [0,1],$$

for $n = 200$ and the true function f_0 given by

$$f_0(t) = \sum_k \theta_{0,k} \sqrt{2} \sin(k\pi t).$$

The Fourier coefficients of this function are given by $\theta_{0,k} = k^{-2.25}\sin(10k)$, corresponding to a true regularity level as in (2.2) given by $\beta = 1.75$. Figure 1 shows the function f_0, its primitive, and the noisy observation Y.

We put the Gaussian prior (2.3) on (the Fourier coefficients of) f_0, with prior regularity level $\alpha = 1.75$, and determined an appropriate scaling parameter $\hat{\tau}_n$ by the empirical Bayes method. The left panel of Figure 2 shows the true signal f_0 (black) and the posterior mean (red). The right panel shows the empirical log-likelihood for τ.

The empirical Bayes reconstruction is satisfying. To illustrate that the scale parameter τ of the prior really matters, we also computed the posterior means with scaling parameter 20 times larger and 20 times smaller than the empirical

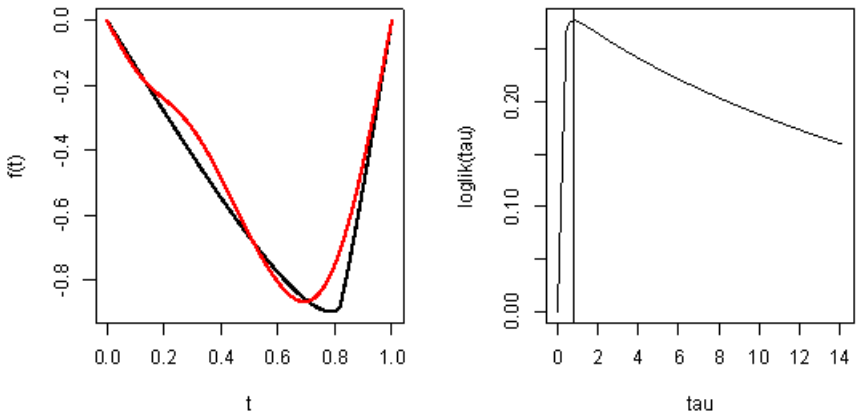

FIG 2. *Left panel: the true signal f_0 (black) and the posterior mean (red). Right panel: the empirical log-likelihood for τ, with indicated point of maximum $\hat{\tau}_n$. The prior smoothness $\alpha = 1.75$ is equal to the true regularity of the signal.*

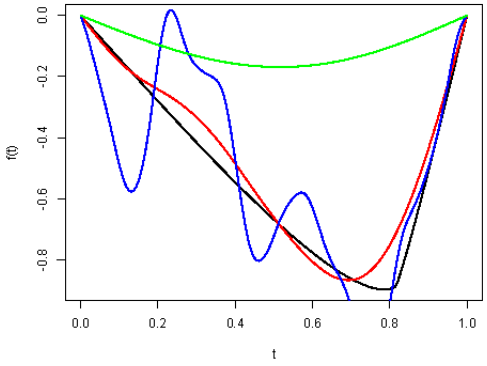

FIG 3. *The true function (black) and empirical Bayes constructions with scaling 20 times the maximum likelihood estimator (blue), the maximum likelihood estimator divided by 20 (green), and the maximum likelihood estimator (black). The prior smoothness $\alpha = 1.75$ is equal to the true regularity of the signal.*

Bayes value. This leads to under-smoothing (blue), and over-smoothing (green), respectively, as shown in Figure 3.

In an attempt to visualize the cut-off at $\beta = \alpha + 1/2$ we repeated the procedure for various prior regularities α near β, every time choosing the scaling by the empirical Bayes method. The results are shown in Figure 4. The theory claims that big values α (i.e. over-smoothing) work fine, as they can and will be corrected by the choice of the scale parameter $\hat{\tau}_n$, but values α below $\beta - 1/2$ cannot be corrected, and lead to suboptimal reconstruction. This is illustrated

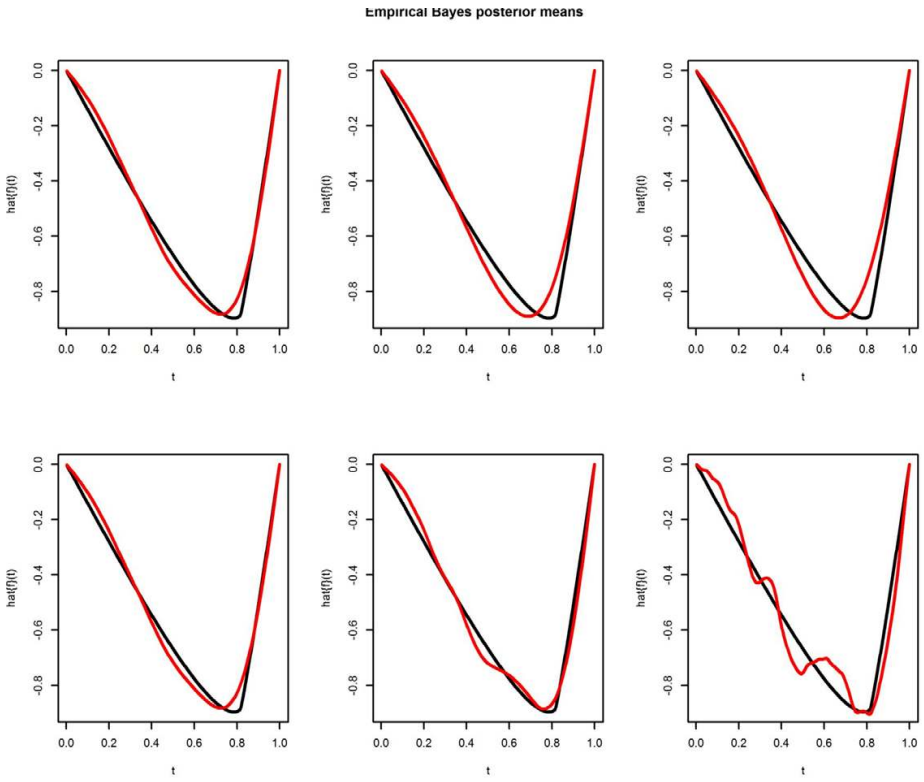

FIG 4. *Top panel from left to right: true signal and empirical Bayes posterior means for the priors with regularities* $\alpha = \beta, \beta + 1/2, \beta + 1$. *Bottom panel: true signal and posterior means for the priors with regularities* $\alpha = \beta, \beta - 1/2, \beta - 1$.

in Figure 4, in which the prior smoothness increases in steps of $1/2$ from β to $\beta + 1$ in the top panels, and decreases from β to $\beta - 1$ in the top panels. The last reconstruction, for $\alpha = \beta - 1$, is clearly not satisfactory.

The theory says that the empirical and hierarchical Bayes methods do not differ much. We illustrate this in Figure 5, which is the hierarchical Bayes version of Figure 2. Instead of the likelihood for τ, the picture shows the posterior distribution of this parameter in the right panel. We used the inverse Gamma distribution for the square scaling parameter τ^2, which is conjugate to the Gaussian location family. The posterior distribution was computed by a Gibbs sampler. Finally Figure 6 is the hierarchical Bayes counterpart of Figure 4. The estimates are computed based on the same (simulated) datasets, and show the same pattern: an undersmoothed base prior (right panels in the bottom row) cannot be corrected by Bayesian (posterior) averaging over a scale parameter, whereas overmoothed base priors (top row) can.

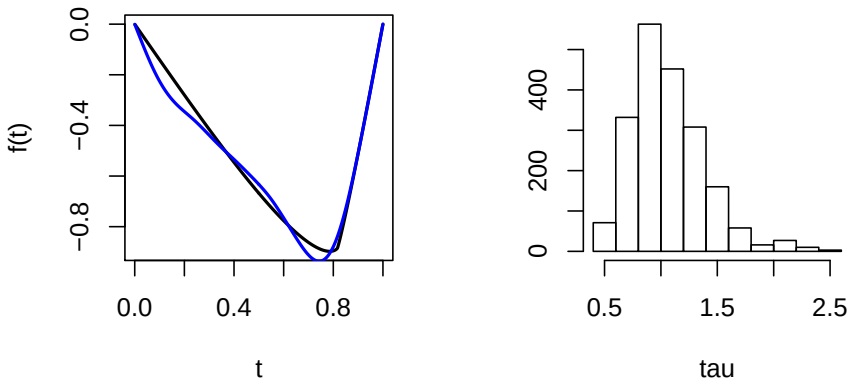

FIG 5. *Left panel: the true signal f_0 black and the hierarchical Bayes posterior mean (blue). Right panel: MCMC sample of size 2000 from the posterior distribution of τ after a burn-in period of size 10000. The prior smoothness $\alpha = 1.75$ is equal to the true regularity of the signal.*

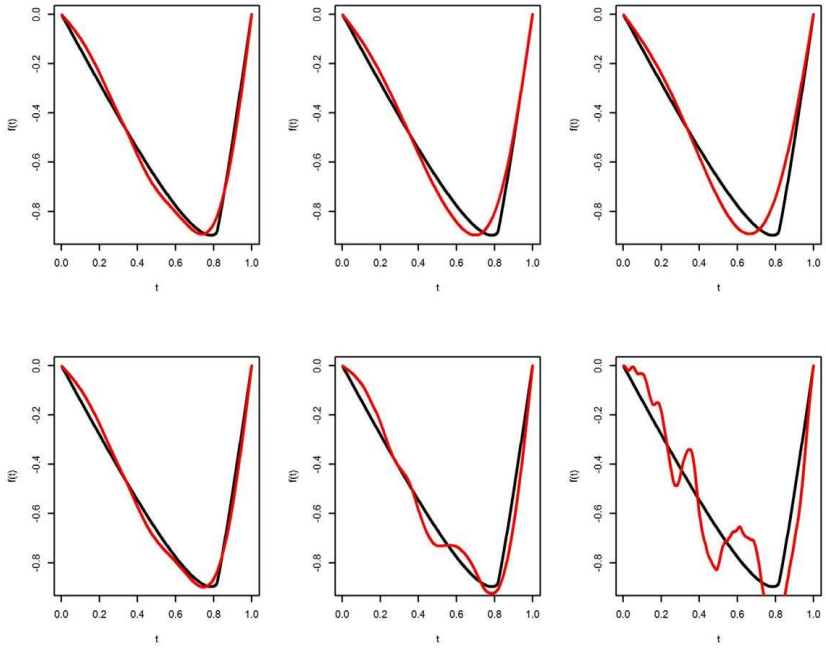

FIG 6. *Top panel from left to right: true signal and hierarchical Bayes posterior means for the priors with regularities $\alpha = \beta, \beta + 1/2, \beta + 1$. Bottom panel: true signal and posterior means for the priors with regularities $\alpha = \beta, \beta - 1/2, \beta - 1$.*

4. Proof of Theorem 2.1

Every term of the series (2.5) that defines ℓ_n is a smooth function of τ. With the help of the dominated convergence theorem, it is straightforward to see that the function ℓ_n is (P_0-a.s.) continuously differentiable on $(0,\infty)$, with derivative given by the series of term-wise derivatives. It will be convenient first to substitute $\nu^{1+2\alpha} = \tau^2 n$, and then differentiate with respect to ν. The resulting derivative map $\mathbb{M}_n$ is given by

$$\mathbb{M}_n(\nu) = \frac{1+2\alpha}{2}\Big(\sum_{k=1}^{\infty} \frac{n\nu^{2\alpha}k^{1+2\alpha}}{(k^{1+2\alpha}+\nu^{1+2\alpha})^2} X_k^2 - \sum_{k=1}^{\infty} \frac{\nu^{2\alpha}}{k^{1+2\alpha}+\nu^{1+2\alpha}}\Big). \tag{4.1}$$

In the new parametrization the upper and lower bounds become

$$\underline{\nu}_n = \sup\big\{\nu \geq 0 : nh(\nu) \geq L\big\},$$
$$\overline{\nu}_n = \sup\big\{\nu \geq 0 : nh(\nu) \geq l\big\},$$

for the function $h : (0,\infty) \to (0,\infty)$ given by

$$h(\nu) = \sum_{k=1}^{\infty} \frac{\nu^{2\alpha}k^{1+2\alpha}\theta_{0,k}^2}{(k^{1+2\alpha}+\nu^{1+2\alpha})^2}. \tag{4.2}$$

In the following subsections we prove that if the constants $l, L > 0$ are sufficiently small and large, respectively, then with probability tending to 1,

(i) the function $\mathbb{M}_n$ is strictly negative and bounded away from 0 on $(\overline{\nu}_n, \infty)$,
(ii) larger than any given constant on $(\underline{\nu}_n/2, \underline{\nu}_n)$,
(iii) bounded below by a fixed constant on $(0, \underline{\nu}_n/2)$.

Property (i) shows that the primitive function of $\mathbb{M}_n$ (and hence the log marginal Bayesian likelihood ℓ_n) is decreasing on $(\overline{\nu}_n, \infty)$, whence an absolute maximum is taken to the left of $\overline{\nu}_n$. The pair of properties (ii) and (iii) imply that the primitive function of $\mathbb{M}_n$ increases more on $(\underline{\nu}_n/2, \underline{\nu}_n)$ than it possibly decreases on $(0, \underline{\nu}_n/2)$. Thus an absolute maximum of ℓ_n is taken to the right of $\underline{\nu}_n$. We conclude that the absolute maximum of ℓ_n is taken in the interval $[\underline{\nu}_n, \overline{\nu}_n]$, which is the first assertion of Theorem 2.1.

Because the constant in (iii) may be negative, it does not follow that $\mathbb{M}_n$ is nonnegative throughout $(0, \underline{\nu}_n/2)$. Thus our proof does not exclude additional *local* maxima on this interval. In fact such local maxima may exist for irregular θ_0, as we illustrate in Section 4.6.

We consider the special cases that $\theta_{0,k}^2 \geq c^2 k^{-1-2\beta}$ for a constant $c > 0$, $\theta_{0,k}^2 \asymp C^2 k^{-1-2\beta}$, or $\theta_0 = 0$, separately in Sections 4.4 and 4.5. In particular, in Section 4.4 we derive concrete bounds on $\underline{\nu}_n$ and $\overline{\nu}_n$ in these cases. In particular it is seen that $\overline{\nu}_n \lesssim n^{1/(1+2\beta)}$ under (2.2).

4.1. Asymptotic behavior of $\mathbb{M}_n$ on $(\overline{\nu}_n, \infty)$

In this section we prove that if l in the definition of $\overline{\nu}_n$ is small enough, then

$$\limsup_{n\to\infty} \sup_{\nu\geq\overline{\nu}_n} \mathrm{E}_0\mathbb{M}_n(\nu) < 0, \tag{4.3}$$

$$\sup_{\nu\geq\overline{\nu}_n} |\mathbb{M}_n(\nu) - \mathrm{E}_0\mathbb{M}_n(\nu)| \xrightarrow{\mathrm{P}_0} 0. \tag{4.4}$$

This shows that $\mathbb{M}_n$ is negative throughout $[\overline{\nu}_n, \infty)$ with probability tending to one, so that the empirical likelihood is strictly decreasing on this interval.

For the proof of (4.3) we note that, since $\mathrm{E}_0 X_k^2 = \theta_{0,k}^2 + 1/n$,

$$\frac{2}{1+2\alpha}\mathrm{E}_0\mathbb{M}_n(\nu) = nh(\nu) - \sum_{k=1}^{\infty} \frac{\nu^{-1}}{\bigl((k/\nu)^{1+2\alpha}+1\bigr)^2},$$

for h defined in (4.2). By considering Riemann sums (cf. Lemma A.1 in the appendix) we see that for $\nu\to\infty$ the second term on the right converges to the positive constant $c_\alpha := \int_0^\infty (x^{1+2\alpha}+1)^{-2}\,dx$. By the definition of $\overline{\nu}_n$ we have $nh(\nu)\leq l$ for $\nu\geq\overline{\nu}_n$. It follows that (4.3) is satisfied for $l < c_\alpha$.

For the proof of (4.4) it suffices, by Corollary 2.2.5 in [28] applied with $\psi(x) = x^2$, to show that $\mathrm{Var}_0\,\mathbb{M}_n(\overline{\nu}_n)\to 0$ and

$$\int_0^{\mathrm{diam}_n} \sqrt{N(\varepsilon, (\overline{\nu}_n,\infty), d_n)}\,d\varepsilon \to 0,$$

where d_n is the semi-metric defined by $d_n^2(\nu_1,\nu_2) = \mathrm{Var}_0(\mathbb{M}_n(\nu_1) - \mathbb{M}_n(\nu_2))$, diam_n is the diameter of $(\overline{\nu}_n,\infty)$ relative to d_n, and $N(\varepsilon, B, d)$ is the minimal number of d-balls of radius ε needed to cover the set B.

The random variables X_k^2 are independent and $\mathrm{Var}_0\,X_k^2 = 2/n^2 + 4\theta_{0,k}^2/n$. Hence, by (2.2),

$$\mathrm{Var}_0\,\mathbb{M}_n(\nu) \lesssim \sum_{k=1}^{\infty} \frac{\nu^{4\alpha}k^{2+4\alpha}(1+n\theta_{0,k}^2)}{(k^{1+2\alpha}+\nu^{1+2\alpha})^4} \lesssim \bigl(1+nh(\nu)\bigr)\frac{1}{\nu}. \tag{4.5}$$

(The first part can be handled by splitting the sum in the parts $k\leq\nu$ and $k>\nu$ and bound $k^{1+2\alpha}+\nu^{1+2\alpha}$ below by $\nu^{1+2\alpha}$ and $k^{1+2\alpha}$, respectively; for the second part we use the inequality $xy/(x+y)^2\leq 1$, valid for $xy>0$, and the definition of h.) For $\nu\geq\overline{\nu}_n$ we have that $nh(\nu)$ is bounded by l, and hence the right side is bounded by a multiple of $1/\nu$.

It follows that $\mathrm{Var}_0\,\mathbb{M}_n(\overline{\nu}_n)\to 0$ as required. Furthermore, combination with the triangle inequality shows that the d_n-diameter of the set $[\overline{\nu}_n,\infty)$ is bounded by a multiple of $1/\sqrt{\overline{\nu}_n}$.

Next we consider the covering number $N(\varepsilon, [\overline{\nu}_n,\infty), d_n)$. Because the d_n-diameter of the set $[\nu,\infty)$ is bounded above by a multiple of $1/\sqrt{\nu}$, for a large

enough constant A the interval $[A/\varepsilon^2, \infty)$ is included in a single d_n-ball of radius ε. For the remaining interval we have

$$[\overline{\nu}_n, A/\varepsilon^2] \subset \bigcup_{k=0}^{K} [A/(2^{k+1}\varepsilon^2), A/(2^k\varepsilon^2)]$$

for $K \lesssim 1 + (\log(A/(\varepsilon^2\overline{\nu}_n)))_+$. By Lemma 4.1 (below) on each of the relevant intervals $[A/(2^{k+1}\varepsilon^2), A/(2^k\varepsilon^2)]$ appearing on the right:

$$d_n(\nu_1, \nu_2) \lesssim 2^k\varepsilon^2|\nu_1 - \nu_2|.$$

It follows that $N(\varepsilon, [A/(2^{k+1}\varepsilon^2), A/(2^k\varepsilon^2)], d_n) \lesssim 1/\varepsilon$. Putting things together we obtain

$$N(\varepsilon, [\overline{\nu}_n, \infty), d_n) \lesssim \frac{1}{\varepsilon}\Big(1 + \Big(\log\frac{1}{\varepsilon^2\overline{\nu}_n}\Big)_+\Big)$$

and hence

$$\int_0^{\operatorname{diam}_n} \sqrt{N(\varepsilon, (\overline{\nu}_n, \infty), d_n)}\, d\varepsilon \lesssim (\overline{\nu}_n)^{-1/4} \to 0.$$

This concludes the proof of (4.4).

Lemma 4.1. *For any* $0 < \nu_1 < \nu_2 < \infty$,

$$\operatorname{Var}_0(\mathbb{M}_n(\nu_1) - \mathbb{M}_n(\nu_2)) \lesssim \frac{1}{\nu_1^3}\Big(1 + \frac{\nu_2}{\nu_1}\Big)^{4\alpha-2}\big(1 + nh(\nu_1)\big)\,|\nu_1 - \nu_2|^2.$$

Proof. The random variables X_k^2 are independent and $\operatorname{Var}_0 X_k^2 = 2/n^2 + 4\theta_{0,k}^2/n$. Hence, by (2.2), the left hand side is bounded by a constant times

$$\sum_{k=1}^{\infty}\Big(\frac{\nu_1^{2\alpha}}{(k^{1+2\alpha} + \nu_1^{1+2\alpha})^2} - \frac{\nu_2^{2\alpha}}{(k^{1+2\alpha} + \nu_2^{1+2\alpha})^2}\Big)^2 k^{2+4\alpha}(1 + n\theta_{0,k}^2). \tag{4.6}$$

The function $f_k : (0,\infty) \to (0,\infty)$ defined by $f_k(\nu) = \nu^{2\alpha}/(k^{1+2\alpha} + \nu^{1+2\alpha})^2$ has derivative satisfying $|f_k'(\nu)| \lesssim \nu^{2\alpha-1}/(k^{1+2\alpha} + \nu^{1+2\alpha})^2$, which is bounded above by $(\nu_1 \vee \nu_2)^{2\alpha-1}/(k^{1+2\alpha} + \nu_1^{1+2\alpha})^2$ on the interval $[\nu_1, \nu_2]$. Therefore, by the mean value theorem,

$$d_n^2(\nu_1, \nu_2) \lesssim |\nu_1 - \nu_2|^2(\nu_1 \vee \nu_2)^{4\alpha-2}\sum_k \frac{k^{2+4\alpha}(1 + n\theta_{0,k}^2)}{(k^{1+2\alpha} + \nu_1^{1+2\alpha})^4}.$$

We can bound this in the same way as (4.5). □

4.2. Asymptotic behavior of $\mathbb{M}_n$ *on* $(0, n^{1/(3+6\alpha)}]$

In this section we show that if $\theta_0 \neq 0$, then there exists a constant $K > 0$ such that with P_0-probability tending to 1, it holds that $\mathbb{M}_n(\nu) \geq K\nu^{2\alpha}n^{1/3}$ on $(0, n^{1/(3+6\alpha)}]$.

For $\nu^{1+2\alpha} \le n^{1/3}$ we have

$$\frac{2}{(1+2\alpha)\nu^{2\alpha}}\mathbb{M}_n(\nu) \ge \sum_{k=1}^{\infty} \frac{nk^{1+2\alpha}}{(k^{1+2\alpha}+n^{1/3})^2} X_k^2 - \sum_{k=1}^{\infty} \frac{1}{k^{1+2\alpha}}.$$

Since $\mathrm{E}_0 X_k^2 = 1/n + \theta_{0,k}^2$ and $2xy \ge x + y$ for $x, y \ge 1$, the expected value of the right-hand side is bounded below by

$$\sum_{k=1}^{\infty} \frac{k^{1+2\alpha}}{(k^{1+2\alpha}+n^{1/3})^2} + \frac{1}{4} n^{1/3} \sum_{k=1}^{\infty} \frac{\theta_{0,k}^2}{k^{1+2\alpha}} - \sum_{k=1}^{\infty} \frac{1}{k^{1+2\alpha}},$$

which, for n large enough, is bounded below by a constant times $n^{1/3}$ if $\theta_0 \neq 0$. Since $\mathrm{Var}_0\, X_k^2 = 2/n^2 + \theta_{0,k}^2 \lesssim 1/n^2 + 1/(nk^{1+2\beta})$, the variance is bounded by a constant times

$$\sum_{k=1}^{\infty} \frac{k^{2+4\alpha}}{(k^{1+2\alpha}+n^{1/3})^4} + n \sum_{k=1}^{\infty} \frac{k^{1+4\alpha-2\beta}}{(k^{1+2\alpha}+n^{1/3})^4},$$

which is (easily) bounded by $n^{1/3}$ for n large enough. The proof of the statement is now completed by an application of Chebychev's inequality.

4.3. Asymptotic behavior of $\mathbb{M}_n$ on $(n^{1/(3+6\alpha)}, \underline{\nu}_n)$

In this section we show that if the constant L in the definition of $\underline{\nu}_n$ is chosen large enough, then $\mathbb{M}_n$ is bounded uniformly below by a fixed (negative) constant on $(n^{1/(3+6\alpha)}, \underline{\nu}_n)$ and by an arbitrarily large constant on $(\underline{\nu}_n/2, \underline{\nu}_n)$, with probability tending to 1.

Since $X_k^2 \ge \theta_{0,k}^2 + 2\theta_{0,k} Z_k/\sqrt{n}$, we have

$$\frac{2}{1+2\alpha}\mathbb{M}_n(\nu) \ge nh(\nu) + 2\sqrt{n}\mathbb{H}(\nu) - \sum_{k=1}^{\infty} \frac{\nu^{2\alpha}}{k^{1+2\alpha}+\nu^{1+2\alpha}},$$

for h given in (4.2) and

$$\mathbb{H}(\nu) = \sum_{k=1}^{\infty} \frac{\nu^{2\alpha} k^{1+2\alpha} \theta_{0,k}}{(k^{1+2\alpha}+\nu^{1+2\alpha})^2} Z_k. \tag{4.7}$$

The last term on the right tends to $-\int_0^{\infty} (x^{1+2\alpha}+1)^{-1}\, dx$, as $\nu \to \infty$. It suffices to prove that the sum $nh(\nu) + 2\sqrt{n}\mathbb{H}(\nu)$ of the remaining terms has the desired properties.

We have that

$$\mathrm{Var}_0\, \mathbb{H}(\nu) = \sum_{k=1}^{\infty} \frac{\nu^{4\alpha} k^{2+4\alpha} \theta_{0,k}^2}{(k^{1+2\alpha}+\nu^{1+2\alpha})^4} \lesssim \frac{1}{\nu} h(\nu). \tag{4.8}$$

We shall show that the sequence of random variables

$$G_n := \frac{1}{n^{1/(6+12\alpha)}} \sup_{n^{1/(3+6\alpha)} \le \nu \le \underline{\nu}_n} \frac{|\mathbb{H}(\nu)|}{\sqrt{h(\nu)/\nu}}$$

tends in probability to zero. Then for every $\nu \ge n^{1/(3+6\alpha)}$,

$$nh(\nu) + 2\sqrt{n}\mathbb{H}(\nu) \ge nh(\nu) - 2\sqrt{nh(\nu)}G_n \ge \begin{cases} -G_n^2, & \nu > 0, \\ nh(\nu)/2, & nh(\nu) \ge 16G_n^2, \end{cases}$$

because $f(x) = x - 2\sqrt{x}g$ possesses minimal value $-g^2$ on $(0,\infty)$ and is bounded below by $x/2$ for $x \ge 16g^2$. It follows that the left side is bounded below on $(n^{1/(3+6\alpha)}, \underline{\nu}_n)$ by a negative constant that tends to zero, and is "big" whenever $nh(\nu)$ is big.

The definition of $\underline{\nu}_n$ implies that $nh(\underline{\nu}_n) \ge L$. Furthermore, for any ν in $[\underline{\nu}_n/2, \underline{\nu}_n]$,

$$h(\nu) = \sum_{k=1}^{\infty} \frac{\nu^{2\alpha} k^{1+2\alpha} \theta_{0,k}^2}{(k^{1+2\alpha} + \nu^{1+2\alpha})^2} \ge \sum_{k=1}^{\infty} \frac{(\underline{\nu}_n/2)^{2\alpha} k^{1+2\alpha} \theta_{0,k}^2}{(k^{1+2\alpha} + \underline{\nu}_n^{1+2\alpha})^2} = 2^{-2\alpha} h(\underline{\nu}_n).$$

It follows that $nh(\nu) \ge 2^{-2\alpha}L$ for any ν in $[\underline{\nu}_n/2, \underline{\nu}_n]$, whence $nh(\nu) \ge 16G_n^2$ with probability tending to 1, and hence $nh(\nu) + 2\sqrt{n}\mathbb{H}(\nu) \ge nh(\nu)/2 \ge 2^{-2\alpha}L/2$. This can be made arbitrarily large by choice of L.

Finally we prove that $G_n \to 0$ in probability. The process $\mathbb{H}(\nu)/\sqrt{h(\nu)/\nu}$ is Gaussian. Lemma 4.2 (below) shows that on the interval $[\nu, 2\nu]$ its intrinsic metric is bounded above by a multiple of $|\cdot|/\nu$. It follows that the covering number of this interval relative to the Gaussian metric is bounded above by a multiple of $1/\varepsilon$. Since $\underline{\nu}_n$ is bounded by a power of n, the interval $(1, \underline{\nu}_n]$ can be covered with $O(\log n)$ intervals of this type, and hence has covering number bounded above by a multiple of $\log n/\varepsilon$. By Corollary 2.2.5 in [28], applied with $\psi(x) = e^{x^2} - 1$, it follows that

$$\mathrm{E}_0 \sup_{1 < \nu < \underline{\nu}_n} \left| \frac{\mathbb{H}(\nu)}{\sqrt{h(\nu)/\nu}} - \frac{\mathbb{H}(\underline{\nu}_n)}{\sqrt{h(\underline{\nu}_n)/\underline{\nu}_n}} \right| \lesssim \sqrt{\log\log n}.$$

Together with the fact that $\mathrm{Var}_0\, \mathbb{H}(\nu) \lesssim h(\nu)/\nu$, this shows that G_n is of the order $O_P(n^{-1/(6+12\alpha)}\sqrt{\log\log n})$.

Lemma 4.2. *For any $0 < \nu_1 < \nu_2 < \infty$,*

$$\mathrm{Var}_0\Big(\frac{\mathbb{H}(\nu_1)}{\sqrt{h(\nu_1)/\nu_1}} - \frac{\mathbb{H}(\nu_2)}{\sqrt{h(\nu_2)/\nu_2}}\Big) \lesssim \frac{1}{\nu_1^2}\Big(\frac{\nu_2}{\nu_1}\Big)^{2+6\alpha} |\nu_1 - \nu_2|^2.$$

Proof. The left side of the lemma is equal to

$$\sum_{k=1}^{\infty} \Big(\frac{\nu_1^{1/2+2\alpha}}{\sqrt{h(\nu_1)}(k^{1+2\alpha} + \nu_1^{1+2\alpha})^2} - \frac{\nu_2^{1/2+2\alpha}}{\sqrt{h(\nu_2)}(k^{1+2\alpha} + \nu_2^{1+2\alpha})^2} \Big)^2 k^{2+4\alpha} \theta_{0,k}^2.$$

The function g_k defined by $g_k(\nu) = \nu^{1/2+2\alpha}h(\nu)^{-1/2}(k^{1+2\alpha} + \nu^{1+2\alpha})^{-2}$ has derivative satisfying $|g_k'(\nu)| \lesssim \nu^{2\alpha-1/2}h(\nu)^{-1/2}(k^{1+2\alpha}+\nu^{1+2\alpha})^{-2}\big[1+\nu|h'/h(\nu)|\big]$. Since it can be checked that $|h'(\nu)| \le h(\nu)/\nu$, the factor $1 + \nu|h'/h(\nu)|$ is uniformly bounded. Therefore, by the mean value theorem the left side of the lemma is bounded by

$$|\nu_1 - \nu_2|^2 \sum_k \sup_{\nu_1\le\nu\le\nu_2} \frac{\nu^{4\alpha-1}}{h(\nu)(k^{1+2\alpha} + \nu^{1+2\alpha})^4}\theta_{0,k}^2 k^{2+4\alpha}.$$

The sum is bounded by

$$\frac{\sum_k \sup_{\nu_1\le\nu\le\nu_2} \frac{\nu^{4\alpha-1}\theta_{0,k}^2 k^{2+4\alpha}}{(k^{1+2\alpha}+\nu^{1+2\alpha})^4}}{\inf_{\nu_1\le\nu\le\nu_2} h(\nu)} \le \frac{\sum_k \frac{(\nu_1\vee\nu_2)^{2\alpha-2}\theta_{0,k}^2 k^{1+2\alpha}}{(k^{1+2\alpha}+\nu_1^{1+2\alpha})^2}}{\sum_k \frac{\nu_1^{2\alpha}k^{1+2\alpha}\theta_{0,k}^2}{(k^{1+2\alpha}+\nu_2^{1+2\alpha})^2}}.$$

Because $(C + x_2)/(C + x_1) \le x_2/x_1$ if $x_2 \ge x_1$ and $C > 0$, we can replace the denominator $(k^{1+2\alpha}+\nu_1^{1+2\alpha})^2$ in the series in the numerator by the denominator of the series in the denominator at the cost of a factor $(\nu_2/\nu_1)^{2+4\alpha}$, after which the two series cancel. □

4.4. Asymptotic behavior for special choices of θ_0

For $\theta_{0,k}^2 = C^2k^{-1-2\beta}$ the function h given by (4.2) satisfies, as $\nu \to \infty$,

$$h(\nu) = C^2\sum_{k=1}^{\infty} \frac{n\nu^{2\alpha}k^{2(\alpha-\beta)}}{\big(k^{1+2\alpha} + \nu^{1+2\alpha}\big)^2} \asymp C^2 \begin{cases} \nu^{-1-2\beta}c_{\alpha,\beta}, & \beta < 1/2 + \alpha, \\ \nu^{-1-2\beta}\log\nu, & \beta = 1/2 + \alpha, \\ \nu^{-2-2\alpha}c_{\alpha,\beta}, & \beta > 1/2 + \alpha, \end{cases} \tag{4.9}$$

(cf. Lemma A.1) for the constants $c_{\alpha,\beta}$ defined by

$$c_{\alpha,\beta} = \begin{cases} \int_0^\infty \frac{x^{2(\alpha-\beta)}}{(x^{1+2\alpha}+1)^2}\,dx, & \beta < 1/2 + \alpha, \\ \sum_{k=1}^\infty k^{2\alpha-2\beta}, & \beta > 1/2 + \alpha. \end{cases}$$

In this case the definition of $\overline{\nu}_n$ readily gives that

$$\overline{\nu}_n \asymp \begin{cases} (C^2c_{\alpha,\beta}n/l)^{\frac{1}{1+2\beta}}, & \text{if } \beta < 1/2 + \alpha, \\ (C^2n\log n/l)^{\frac{1}{1+2\beta}}, & \text{if } \beta = 1/2 + \alpha, \\ (C^2c_{\alpha,\beta}n/l)^{\frac{1}{2+2\alpha}}, & \text{if } \beta > 1/2 + \alpha. \end{cases} \tag{4.10}$$

Furthermore, by its definition $\underline{\nu}_n$ satisfies the same equation with L instead of l.

If θ_0 satisfies the one-sided inequality $\theta_{0,k}^2 \le C^2k^{-1-2\beta}$ (or $\theta_{0,k}^2 \ge c^2k^{-1-2\beta}$), then the function h can be upper bounded as previously (or lower bounded with c instead of C, respectively). By its definition the upper bound $\overline{\nu}_n$ can then be upper bounded by the right side of (4.10) (or $\underline{\nu}_n$ can be lower bounded by this expression with c replacing C, respectively). Thus given both the upper and lower bound on θ_0, the two quantities $\overline{\nu}_n$ and $\underline{\nu}_n$ have the same order.

Finally assume again that $\theta_{0,k}^2 \asymp C^2 k^{-1-2\beta}$. Relation (4.4) is then valid also with $\underline{\nu}_n$ replacing $\overline{\nu}_n$: $\sup_{\nu \geq \underline{\nu}_n} |\mathbb{M}_n(\nu) - \mathrm{E}_0 \mathbb{M}_n(\nu)| \to 0$ in probability, in all three cases. Since $\hat{\nu}_n$ is a zero of $\mathbb{M}_n$ and is contained in $[\underline{\nu}_n, \overline{\nu}_n]$, it follows that $\mathrm{E}_0 \mathbb{M}_n(\nu)_{|\nu=\hat{\nu}_n} \to 0$ in probability. Again employing (4.9), we conclude that $C^2 n \hat{\nu}_n^{-1-2\beta} c_{\alpha,\beta} - c_\alpha$ or $C^2 n \hat{\nu}_n^{-1-2\beta} \log \hat{\nu}_n - c_\alpha$ or $C^2 n \hat{\nu}_n^{-2-2\alpha} c_{\alpha,\beta} - c_\alpha$ tends to zero in probability in the three cases, respectively, for the constants $c_{\alpha,\beta}$ defined previously and

$$c_\alpha = \int_0^\infty \frac{1}{(x^{1+2\alpha}+1)^2}\, dx.$$

This readily gives that $\hat{\tau}_n / \overline{\tau}_n$ tends to a constant in probability.

4.5. The special case $\theta_0 = 0$

If $\theta_0 = 0$, then the expected value $\mathrm{E}_0 \mathbb{M}_n(\nu)$, given at the beginning of Section 4.1, tends to a negative constant as $\nu \to \infty$, and is 0 only at $\nu = 0$. Thus it is negative and bounded away from zero on every interval $[\nu, \infty)$ for $\nu > 0$.

Furthermore, in this case the function h vanishes and hence the computations in Section 4.1 show that $\mathrm{Var}_0 \mathbb{M}_n(\nu) \lesssim 1/\nu$ for every $\nu > 0$, and that the upper bound on $\mathrm{Var}_0(\mathbb{M}_n(\nu_1) - \mathbb{M}_n(\nu_2))$ given by Lemma 4.1 is valid without the factor $(1 + nh(\nu_1))$ in its right side. Similar arguments as in Section 4.1 then show that $\mathbb{M}_n$ tends to its expectation uniformly on every sequence of intervals $[\nu_n, \infty)$ with $\nu_n \to \infty$.

Combination of these findings shows that $\mathrm{P}_0(\hat{\nu}_n \leq \nu_n) \to 1$ for every $\nu_n \to \infty$. This is equivalent to $n\hat{\tau}_n^2$ being bounded in probability.

4.6. Example: multiple local maxima

We construct a fixed parameter θ_0 and a subsequence $n_j \to \infty$ such that, with probability tending to 1, the random map $\mathbb{M}_{n_j}$ is strictly negative somewhere in the interval $[0, \underline{\nu}_{n_j}]$. We fix $0 < \beta < \alpha + 1/2$ and for (large) positive constants A, B and C to be determined later and $j \in \mathbb{N}$,we set $\nu_j = A^j$ and $n_j = B\nu_j^{1+2\beta}$ and define θ_0 by

$$\theta_{0,k} = \begin{cases} k^{-1/2-\beta} & \text{if } \nu_j \leq k \leq 2\nu_j \text{ for some } j \in \mathbb{N}, \\ 0 & \text{otherwise.} \end{cases}$$

We shall show that by choosing the constants A, B and C sufficiently large, we can ensure that $n_j h(\nu_j/C)$ becomes arbitrarily small (positive) and $n_j h(\nu_j) > L$ for j large enough. The latter implies that $\nu_j/C < \nu_j < \underline{\nu}_{n_j}$, and the former that $\mathrm{E}_0 \mathbb{M}_{n_j}(\nu_j/C)$ is smaller than a negative constant. Using (4.5) we then also get that $\mathrm{Var}_0 \mathbb{M}_{n_j}(\nu_j/C) \lesssim 1/\nu_j \to 0$, and the claim follows by Chebychev's inequality.

To upper bound $n_j h(\nu_j/C)$ we split the sum in the definition of h into three parts. The sum over the indices $k < 2\nu_{j-1}$ is bounded by

$$n_j\Big(\frac{\nu_j}{C}\Big)^{-2-2\alpha}\sum_{k=1}^{2\nu_{j-1}} k^{2\alpha-2\beta} \lesssim n_j\Big(\frac{\nu_j}{C}\Big)^{-2-2\alpha}\nu_{j-1}^{1+2\alpha-2\beta} = A^{-1-2\alpha+2\beta}BC^{2+2\alpha}.$$

The second sum is over the indices $\nu_j < k < 2\nu_j$ and is bounded by

$$n_j\Big(\frac{\nu_j}{C}\Big)^{2\alpha}\sum_{k=\nu_j}^{2\nu_j} k^{-2-2\alpha-2\beta} \lesssim n_j\Big(\frac{\nu_j}{C}\Big)^{2\alpha}\nu_j^{-1-2\alpha-2\beta} = BC^{-2\alpha}.$$

Finally, since $\nu_j < \nu_{j+1}$, we have the same bound for the sum over $k > \nu_j + 1$:

$$n_j\Big(\frac{\nu_j}{C}\Big)^{2\alpha}\sum_{k=\nu_{j+1}}^{\infty} k^{-2-2\alpha-2\beta} \lesssim n_j\Big(\frac{\nu_j}{C}\Big)^{2\alpha}\nu_{j+1}^{-1-2\alpha-2\beta} = BC^{-2\alpha}.$$

We conclude that $n_j h(\nu_j/C) \lesssim A^{-1-2\alpha+2\beta}BC^{2+2\alpha} + BC^{-2\alpha}$. For the lower bound we note that

$$\begin{aligned} n_j h(\nu_j) &\geq n_j \sum_{k=\nu_j}^{2\nu_j} \frac{\nu_j^{2\alpha}k^{2(\alpha-\beta)}}{(k^{1+2\alpha}+\nu_j^{1+2\alpha})^2} \\ &\geq \frac{1}{4}n_j\nu_j^{2\alpha}\sum_{k=\nu_j}^{2\nu_j} k^{-2-2\alpha-2\beta} \gtrsim n_j\nu_j^{2\alpha}\nu_j^{-1-2\alpha-2\beta} = B. \end{aligned}$$

To complete the construction, observe that by choosing B large enough we can ensure that $n_j h(\nu_j) > L$. By next choosing C large enough and then A large enough we can make $n_j h(\nu_j/C)$ arbitrarily small.

5. Proof of Theorem 2.2

It is convenient to continue to work with the parametrization $\nu^{1+2\alpha} = \tau^2 n$. Slightly abusing notation we denote by Π_ν the same prior as Π_τ for $\nu^{1+2\alpha} = \tau^2 n$ and similarly for the posterior, so

$$\Pi_\nu(\cdot \mid X) = \bigotimes_{k=1}^{\infty} N\Big(\frac{\nu^{1+2\alpha}}{k^{1+2\alpha}+\nu^{1+2\alpha}}X_k, \frac{\nu^{1+2\alpha}/n}{k^{1+2\alpha}+\nu^{1+2\alpha}}\Big).$$

In this notation the empirical Bayes posterior is

$$\Pi_{\hat\nu_n}(\cdot \mid X) = \Pi_\nu(\cdot \mid X)\Big|_{\nu=\hat\nu_n},$$

where $\hat\nu_n$ is the (or rather a) zero of the random function $\mathbb{M}_n$ on $(0,\infty)$ defined by (4.1).

Because $\|\theta-\theta_0\|^2=\sum(\theta_k-\theta_{k,0})^2$, we have, with $\hat\theta_{\nu,k}=\nu^{1+2\alpha}(k^{1+2\alpha}+\nu^{1+2\alpha})^{-1}X_k$ the posterior mean,

$$\int\|\theta-\theta_0\|^2\,\Pi_\nu(d\theta\,|\,X)=\sum_k(\hat\theta_{\nu,k}-\theta_{0,k})^2+\frac{1}{n}\sum_{k=1}^\infty\frac{\nu^{1+2\alpha}}{k^{1+2\alpha}+\nu^{1+2\alpha}}. \tag{5.1}$$

By Markov's inequality the left side divided by $(M_n\varepsilon_n)^2$ is an upper bound on $\Pi_\nu(\theta:\|\theta-\theta_0\|\geq M_n\varepsilon_n|X)$, for any $M_n\varepsilon_n>0$. We like to show that the latter probability evaluated at $\nu=\hat\nu_n$ tends to zero for the appropriate rate $\varepsilon_n=\varepsilon_{n,\alpha,\beta}$ and any $M_n\to\infty$. By Theorem 2.1 with probability going to 1, the empirical Bayes rescaling rate $\hat\nu_n$ belongs to the interval $[\underline\nu_n,\overline\nu_n]$. Therefore, to prove Theorem 2.2 it suffices to show that the expectation of the supremum of this expression over $\nu\in[\underline\nu_n,\overline\nu_n]$ is of the appropriate order ε_n^2. We shall first show that the supremum of the expectations has the right order, and next that the expectation of the supremum has the same order.

5.1. Posterior risk for scaling in $[\underline\nu_n,\overline\nu_n]$

The second term of (5.1) is deterministic. The expectation of the first term can be split in square bias and variance terms. We find that the expectation of (5.1) is given by

$$\sum_{k=1}^\infty\frac{k^{2+4\alpha}\theta_{0,k}^2}{(k^{1+2\alpha}+\nu^{1+2\alpha})^2}+\frac{1}{n}\sum_{k=1}^\infty\frac{\nu^{2+4\alpha}}{(k^{1+2\alpha}+\nu^{1+2\alpha})^2}+\frac{1}{n}\sum_{k=1}^\infty\frac{\nu^{1+2\alpha}}{k^{1+2\alpha}+\nu^{1+2\alpha}}.$$

In this section we prove that the supremum of this expression over $\nu\in[\underline\nu_n,\overline\nu_n]$ is bounded by a constant times $n^{-2\beta/(1+2\beta)}+\overline\nu_n/n$. In Section 4.4 it was seen that under (2.2) the upper bound $\overline\nu_n$ is bounded above by the right side of (4.10), which shows that $\overline\nu_n/n\asymp\varepsilon_{n,\alpha,\beta}^2$, the (square) order claimed in Theorem 2.2. The first term $n^{-2\beta/(1+2\beta)}$ is smaller than this order, in all three cases.

The series in the second and third terms are bounded by a multiple of ν (and asymptotic to ν times a constant as $\nu\to\infty$), and hence the suprema of these terms over $\nu\in[\underline\nu_n,\overline\nu_n]$ are bounded by a multiple of $\overline\nu_n/n$.

The first series is decreasing in ν and hence it suffices to consider it at $\nu=\underline\nu_n$. Its terms are bounded above by $\theta_{0,k}^2$. Therefore, in view of (2.2), we have for $N\sim n^{1/(1+2\beta)}$,

$$\sum_{k>N}\frac{k^{2+4\alpha}\theta_{0,k}^2}{(k^{1+2\alpha}+\underline\nu_n^{1+2\alpha})^2}\leq\sum_{k>N}\theta_{0,k}^2\lesssim C^2n^{-2\beta/(1+2\beta)}.$$

By the definition of $\underline\nu_n$ (and continuity of the series) we have, for $\nu\geq\underline\nu_n$,

$$\nu h(\nu)\equiv\sum_{k=1}^\infty\frac{\nu^{1+2\alpha}k^{1+2\alpha}\theta_{0,k}^2}{(k^{1+2\alpha}+\nu^{1+2\alpha})^2}\leq\frac{L}{n}\nu. \tag{5.2}$$

(The function h is as in (4.2).) As a first consequence we have

$$\sum_{k\le\underline{\nu}_n}\frac{k^{2+4\alpha}\theta_{0,k}^2}{(k^{1+2\alpha}+\underline{\nu}_n^{1+2\alpha})^2}\le\sum_{k\le\underline{\nu}_n}\frac{\underline{\nu}_n^{1+2\alpha}k^{1+2\alpha}\theta_{0,k}^2}{(k^{1+2\alpha}+\underline{\nu}_n^{1+2\alpha})^2}\le\frac{L}{n}\underline{\nu}_n\le\frac{L}{n}\overline{\nu}_n,$$

It remains to consider the terms between $\underline{\nu}_n$ and N. For $\nu\le k\le 2\nu$ and any $\nu>0$ we have that $\nu^{1+2\alpha}k^{1+2\alpha}/(k^{1+2\alpha}+\nu^{1+2\alpha})^2\ge 1/(2^{1+2\alpha}+3)$. Therefore, as a second consequence of (5.2),

$$\frac{1}{2^{1+2\alpha}+3}\sum_{\nu<k\le 2\nu}\theta_{0,k}^2\le\sum_k\frac{\nu^{1+2\alpha}k^{1+2\alpha}\theta_{0,k}^2}{(k^{1+2\alpha}+\nu^{1+2\alpha})^2}\le\frac{L}{n}\nu,$$

for $\nu\ge\underline{\nu}_n$. For L large enough that $\underline{\nu}_n 2^L\ge N$ we have

$$\sum_{\underline{\nu}_n<k\le N}\frac{k^{2+4\alpha}\theta_{0,k}^2}{(k^{1+2\alpha}+\underline{\nu}_n^{1+2\alpha})^2}\le\sum_{l=1}^{L}\sum_{\underline{\nu}_n 2^{l-1}<k\le\underline{\nu}_n 2^l}\theta_{0,k}^2\lesssim\sum_{l=1}^{L}\frac{L}{n}\underline{\nu}_n 2^{l-1}.$$

For $\underline{\nu}_n 2^L\sim N$ this is bounded above by a multiple of $LN/n\lesssim n^{-2\beta/(1+2\beta)}$.

5.2. *Uniform result for the posterior risk*

In this section we bound the quantity

$$\mathrm{E}_0\sup_{\nu\in[\underline{\nu}_n,\overline{\nu}_n]}\Big|\sum_k(\hat\theta_{\nu,k}-\theta_{0,k})^2-\mathrm{E}_0\sum_k(\hat\theta_{\nu,k}-\theta_{0,k})^2\Big|.$$

Using the explicit expressions for the $\hat\theta_{\nu,k}$ we see that the random variable in the supremum is the absolute value of $\mathbb{V}(\nu)/n-2\mathbb{W}(\nu)/\sqrt{n}$, where

$$\mathbb{V}(\nu)=\sum_k\frac{\nu^{2+4\alpha}}{(k^{1+2\alpha}+\nu^{1+2\alpha})^2}(Z_k^2-1),\qquad\mathbb{W}(\nu)=\sum\frac{\nu^{1+2\alpha}k^{1+2\alpha}\theta_{0,k}}{(k^{1+2\alpha}+\nu^{1+2\alpha})^2}Z_k.$$

We deal with the two processes separately.

By comparison with Riemann sums (cf. Lemma A.1) we see that $\mathrm{Var}_0\,\mathbb{V}(\nu)\asymp\nu\int_0^\infty(x^{1+2\alpha}+1)^{-4}\,dx$ as $\nu\to\infty$. By Lemma 5.1 below the standard deviation metric of $\mathbb{V}$ is bounded above by a multiple of $|\cdot|/\sqrt{\nu}$ on the interval $[\nu,2\nu]$. Therefore the covering number of this interval relative to the standard deviation metric is bounded above by a multiple of $\sqrt{\nu}/\varepsilon$. Covering the interval $[\underline{\nu}_n,\overline{\nu}_n]$ with the intervals $(2^{-m-1}\overline{\nu}_n,2^{-m}\overline{\nu}_n]$, for $m=0,1,2,\ldots$, we see that its covering number is bounded above by $\sum_m\sqrt{2^{-m}\overline{\nu}_n}/\varepsilon\lesssim\sqrt{\overline{\nu}_n}/\varepsilon$. By Corollary 2.2.5 in [28] applied with $\psi(x)=x^2$, it follows that

$$\mathrm{E}_0\sup_{\underline{\nu}_n\le\nu\le\overline{\nu}_n}|\mathbb{V}(\nu)|\lesssim\sqrt{\overline{\nu}_n}+\int_0^{\sqrt{\overline{\nu}_n}}\sqrt{\sqrt{\overline{\nu}_n}/\varepsilon}\,d\varepsilon\lesssim\sqrt{\overline{\nu}_n}.$$

Divided by n this yields $\sqrt{\overline{\nu}_n}/n \le \overline{\nu}_n/n$.

It remains to deal with the process $\mathbb{W}$. Because $\mathbb{W}(\nu) = \nu\mathbb{H}(\nu)$ for $\mathbb{H}$ given in (4.7), we have by (4.8) that $\operatorname{Var}_0 \mathbb{W}(\nu) = \nu h(\nu)$, for h given in (4.2). By (5.2) we have that $\operatorname{Var}_0 \mathbb{W}(\nu) \le \nu L/n$ for $\nu \ge \underline{\nu}_n$. Furthermore, by Lemma 5.1 (below) the standard deviation metric of $\mathbb{W}$ is bounded above by a multiple of $|\cdot|\, h(\nu)/\nu \lesssim |\cdot|/\sqrt{\nu n}$ on an interval $[\nu, 2\nu]$ with $\nu \ge \underline{\nu}_n$. By the same reasoning as in the preceding paragraph this shows that the covering number of the interval $[\underline{\nu}_n, \overline{\nu}_n]$ relative to the standard deviation metric is bounded above by $\sqrt{\overline{\nu}_n/n}/\varepsilon$. By Corollary 2.2.5 in [28] applied with $\psi(x) = e^{x^2} - 1$, it follows that

$$\mathrm{E}_0 \sup_{\underline{\nu}_n \le \nu \le \overline{\nu}_n} |\mathbb{W}(\mu)| \lesssim \sqrt{\frac{\overline{\nu}_n}{n}} + \int_0^{\sqrt{\overline{\nu}_n/n}} \sqrt{\log \sqrt{\frac{\overline{\nu}_n}{n}} \frac{1}{\varepsilon}}\, d\varepsilon \lesssim \sqrt{\log n} \sqrt{\frac{\overline{\nu}_n}{n}}.$$

Divided by $\sqrt{n}$ this yields the order $\sqrt{\overline{\nu}_n \log n}/n \le \overline{\nu}_n/n$.

Lemma 5.1. *For any $0 < \nu_1 < \nu_2 < \infty$,*

$$\operatorname{Var}_0(\mathbb{V}(\nu_1) - \mathbb{V}(\nu_2)) \le \frac{1}{\nu_1}\Big(\frac{\nu_2}{\nu_1}\Big)^{2+8\alpha} |\nu_1 - \nu_2|^2,$$
$$\operatorname{Var}_0(\mathbb{W}(\nu_1) - \mathbb{W}(\nu_2)) \le \frac{1}{\nu_1}\Big(\frac{\nu_2}{\nu_1}\Big)^{4\alpha} h(\nu_1) |\nu_1 - \nu_2|^2,$$

where h is given by (4.2).

Proof. The left side of the first inequality of the lemma takes the form $\sum_k (h_k(\nu_1) - h_k(\nu_2))^2$, for the function h_k equal to $h_k(\nu) = \sqrt{2}\nu^{2+4\alpha}/(k^{1+2\alpha} + \nu^{1+2\alpha})^2$. The derivative of this function satisfies $|h_k'(\nu)| \lesssim \nu^{1+4\alpha}/(k^{1+2\alpha} + \nu^{1+2\alpha})^2$. The assertion follows by similar, but simpler, arguments as in the proof of Lemma 4.1.

The left side of the second inequality of the lemma can be written in the form $\sum_k (h_k(\nu_1) - h_k(\nu_2))^2 k^{2+4\alpha} \theta_{0,k}^2$, this time for the function h_k given by $h_k(\nu) = \nu^{1+2\alpha}/(k^{1+2\alpha} + \nu^{1+2\alpha})^2$. The derivative of this function satisfies $|h_k'(\nu)| \lesssim \nu^{2\alpha}/(k^{1+2\alpha} + \nu^{1+2\alpha})^2$. By the mean value theorem the left side of the lemma is bounded by a multiple of

$$|\nu_1 - \nu_2|^2 \sum_k \sup_{\nu_1 < \nu < \nu_2} \frac{\nu^{4\alpha} k^{2+4\alpha} \theta_{0,k}^2}{(k^{1+2\alpha} + \nu^{1+2\alpha})^4}$$
$$\le |\nu_1 - \nu_2|^2 \Big(\frac{\nu_2}{\nu_1}\Big)^{4\alpha} \sum_k \frac{\nu_1^{4\alpha} k^{2+4\alpha} \theta_{0,k}^2}{(k^{1+2\alpha} + \nu_1^{1+2\alpha})^4}.$$

The series in the right side is bounded by $\nu_1^{-1} h(\nu_1)$. □

5.3. Proof that the rates in Theorem 2.2 are sharp

If $\theta_1, \theta_2, \ldots$ are independent Gaussian variables with means μ_k and variances σ_k^2, then Chebychev's inequality shows that $\|\theta - \mu\|_2^2 = \sum_k (\theta_k - \mu_k)^2$ satisfies,

for any $c > 0$,

$$\Pr\Bigl(\|\theta-\mu\|_2^2 \le \sum_k \sigma_k^2 - c\sqrt{2\sum_k \sigma_k^4}\Bigr) \le \frac{1}{c^2}.$$

Because a Gaussian distribution gives most probability to a ball of given radius if this is centered at its mean (by Anderson's lemma), this inequality remains true if μ is replaced by a different element of ℓ_2.

Under the posterior distribution $\Pi_\nu(\cdot|X)$ the coordinates $\theta_1, \theta_2, \ldots$ are independent Gaussian variables with variances $\sigma_k^2 = n^{-1}\nu^{1+2\alpha}/(k^{1+2\alpha}+\nu^{1+2\alpha})$, and, for positive constants d_α, c_α, as $\nu \to \infty$,

$$\sum_k \sigma_k^2 = \frac{1}{n}\sum_k \frac{\nu^{1+2\alpha}}{(k^{1+2\alpha}+\nu^{1+2\alpha})} \asymp d_\alpha \frac{\nu}{n},$$
$$\sum_k \sigma_k^4 = \frac{1}{n^2}\sum_k \frac{\nu^{2+4\alpha}}{(k^{1+2\alpha}+\nu^{1+2\alpha})^2} \asymp c_\alpha \frac{\nu}{n^2}.$$

It follows that, for $d < d_\alpha$, uniformly in $\nu \ge \underline{\nu}_n$, for any $c > 0$,

$$\Pi_\nu\Bigl(\theta : \|\theta-\theta_0\|_2^2 \le \frac{d\nu}{n} - c\frac{\sqrt{c_\alpha \nu}}{n}\Bigr) \le \frac{1}{c^2}.$$

For $\nu \ge \underline{\nu}_n$ and $c = (d/\sqrt{c_\alpha})\sqrt{\underline{\nu}_n}/2$ we have $d\nu - c\sqrt{c_\alpha \nu} \ge d\underline{\nu}_n/2$, and hence

$$\sup_{\nu \ge \underline{\nu}_n} \Pi_\nu\bigl(\theta : \|\theta-\theta_0\|_2^2 \le d\underline{\nu}_n/(2n)\bigr) \lesssim \frac{1}{\underline{\nu}_n}. \tag{5.3}$$

Because $\mathrm{P}_0(\hat{\nu}_n \ge \underline{\nu}_n) \to 1$, the empirical Bayes posterior probability of the event in the left side is with probability tending to one bounded by the supremum, and hence tends to zero.

Under the assumption that $\theta_{0,k}^2 \ge c^2 k^{-1-2\beta}$ the sequence $\underline{\nu}_n/n$ was seen to be of the order $\overline{\nu}_n$ in Section 4.4. Because $\overline{\nu}_n/n \asymp \varepsilon_{n,\alpha,\beta}^2$, the second assertion of Theorem 2.2 follows by (5.3).

If $\theta_0 \ne 0$, then there exists $k \in \mathbb{N}$ such that

$$h(\nu) \ge \frac{\nu^{2\alpha}k^{1+2\alpha}}{(k^{1+2\alpha}+\nu^{1+2\alpha})^2} \asymp \frac{k^{1+2\alpha}}{\nu^{2+2\alpha}}, \tag{5.4}$$

as $\nu \to \infty$. In view of its definition it follows that $\underline{\nu}_n \gtrsim n^{1/(2+2\alpha)}$. Again using (5.3) we obtain a lower bound on the order of the square rate equal to $\underline{\nu}_n/n \asymp n^{(1+2\alpha)/(2+2\alpha)}$. For $\beta > 1/2+\alpha$ this rate is equal to $\varepsilon_{n,\alpha,\beta}^2$. (In the other case it is a valid lower bound, but strictly smaller than $\varepsilon_{n,\alpha,\beta}^2$ and hence of less interest.)

6. Proof of Theorem 2.3

As noted following (5.4), if $\theta_0 \ne 0$, then $\overline{\nu}_n \ge \underline{\nu}_n$ tends to infinity as $n \to \infty$ at least at the rate $n^{1/(2+2\alpha)}$. (The corresponding sequences $\overline{\tau}_n \ge \underline{\tau}_n$ can tend

both to zero or infinity, depending on α and θ_0.) Let $\tau_n(\nu)$ be the solution to $\nu^{1+2\alpha} = n\tau^2$, and let $L_n(\tau) = \exp \ell_n(\tau)$ be the marginal likelihood.

In Section 4 (see (i)–(iii) in its introduction) it was seen that $\mathbb{M}_n(\nu) = (d/d\nu)\ell_n\big(\tau_n(\nu)\big)$ satisfies, for positive constants c_1, c_2, c_3,

$$\mathbb{M}_n(\nu) \begin{cases} \le -c_1, & \text{for } \nu \ge \overline{\nu}_n, \\ \ge c_2, & \text{for } \nu \in [\underline{\nu}_n/2, \underline{\nu}_n], \\ \ge -c_3, & \text{for } \nu \in [0, \underline{\nu}_n/2]. \end{cases}$$

Furthermore, the constant c_2 can be chosen arbitrarily large by choosing L in (2.8) large enough, while the constant c_3 is fixed.

For $\tau_n(\nu) \ge M\overline{\tau}_n$ and $\tau_n(\nu_1) = 2\overline{\tau}_n$, we have $\nu \ge M^{2/(1+2\alpha)}\overline{\nu}_n$ and $\nu_1 = 2^{2/(1+2\alpha)}\overline{\nu}_n$. Choose $M \ge 2$. Since both are greater than $\overline{\nu}_n$, it follows that

$$\ell_n\big(\tau_n(\nu)\big) - \ell_n\big(\tau_n(\nu_1)\big) \le -c_1(\nu - \nu_1) \le -c_4\overline{\nu}_n,$$

for $c_4 = c_1(M^{2/(1+2\alpha)} - 2^{2/(1+2\alpha)})$. Consequently $L_n(\tau) \le L_n(2\overline{\tau}_n)e^{-c_4\overline{\nu}_n}$ for $\tau \ge M\overline{\tau}_n$. Since also $L_n(\tau) \ge L_n(2\overline{\tau}_n)$, for $\tau \in [\overline{\tau}_n, 2\overline{\tau}_n]$, we find

$$\Pi(\tau \ge M\overline{\tau}_n \,|\, X) \le \frac{\int_{M\overline{\tau}_n}^{\infty} L_n(\tau)\, d\lambda(\tau)}{\int_{\overline{\tau}_n}^{2\overline{\tau}_n} L_n(\tau)\, d\lambda(\tau)} \le \frac{\lambda(M\overline{\tau}_n, \infty)e^{-c_4\overline{\nu}_n}}{\lambda(\overline{\tau}_n, 2\overline{\tau}_n)}.$$

Here c_4 can be made arbitrarily large by choice of a large M, and $\overline{\nu}_n\overline{\tau}_n^2 = \overline{\nu}_n^{2+2\alpha}/n$ is bounded away from 0. If $\Gamma := 1/\tau^2$ possesses a Gamma distribution with shape a and rate b, then

$$\lambda(\overline{\tau}_n, 2\overline{\tau}_n) = \Pr\Big(\frac{1}{4\overline{\tau}_n^2} \le \Gamma \le \frac{1}{\overline{\tau}_n^2}\Big) \asymp \begin{cases} \Big(\frac{1}{\overline{\tau}_n^2}\Big)^a, & \text{if } \overline{\tau}_n \to \infty, \\ \Big(\frac{1}{\overline{\tau}_n^2}\Big)^{a-1} e^{-b/(4\overline{\tau}_n^2)}, & \text{if } \overline{\tau}_n \to 0, \\ 1, & \text{if } 0 \ll \overline{\tau}_n \ll \infty. \end{cases}$$

In all cases this is much bigger than $e^{-c_4\overline{\nu}_n}$ if M and hence c_4 is chosen big enough. Arguing, if necessary, along subsequences, we conclude that the right side of the second last display tends to zero.

The analysis of the left tail is similar, but slightly more complicated, because the different lower bounds on $\mathbb{M}_n$ on the two subintervals of $[0, \underline{\tau}_n]$. The difference $\ell_n\big(\tau_n(\nu_1)\big) - \ell_n\big(\tau_n(\nu)\big)$ is bounded below by $c_2(\nu_1 - \nu)$ if $\underline{\nu}_n/2 \le \nu \le \nu_1 \le \underline{\nu}_n$, and bounded below by $c_2(\nu_1 - \underline{\nu}_n/2) - (\underline{\nu}_n/2 - \nu)c_3$ if $\nu \le \underline{\nu}_n/2 \le \nu_1 \le \underline{\nu}_n$. If $c_2 > 2c_3$ the difference can be seen to be bounded below by $c_5\underline{\nu}_n$ if both $3\underline{\nu}_n/4 \le \nu_1 \le \underline{\nu}_n$ and $\nu_1 - \nu \ge c_6\underline{\nu}_n$, for $c_5 = (c_2/4 - c_3/2) \wedge c_2c_6$. Let $\tilde{\nu}_n = (1/4)^{1/(1+2\alpha)}\underline{\nu}_n$, so that $\tau_n(\tilde{\nu}_n) = \underline{\tau}_n/2$, and $\nu_1 = \big((3/4)^{1/(1+2\alpha)} \vee (3/4)\big)\underline{\nu}_n$. Then for $\nu \le \tilde{\nu}_n$ we have $\nu_1 - \nu \ge c_6\underline{\nu}_n$, for $c_6 = (3/4)^{1/(1+2\alpha)} - (1/4)^{1/(1+2\alpha)}$ and $\nu_1 \ge 3\underline{\nu}_n/4$ and hence $\ell_n\big(\tau_n(\nu_1)\big) - \ell_n\big(\tau_n(\nu)\big) \ge c_5\underline{\nu}_n$. Consequently $L_n(\tau) \le L_n\big(\tau_n(\nu_1)\big)e^{-c_5\underline{\nu}_n}$ for $\tau \le \underline{\tau}_n/2$. Since $\underline{\tau}_n/2 \le \tau_n(\nu_1) \le \underline{\tau}_n$, we also have that $L_n(\tau) \ge L_n\big(\tau_n(\nu_1)\big)$ on $\big[\tau_n(\nu_1), \underline{\tau}_n\big]$. It follows that

$$\Pi(\tau \le \underline{\tau}_n/2 \,|\, X) \le \frac{\int_0^{\underline{\tau}_n/2} L_n(\tau)\, d\lambda(\tau)}{\int_{\tau_n(\nu_1)}^{\underline{\tau}_n} L_n(\tau)\, d\lambda(\tau)} \le \frac{\lambda(0, \underline{\tau}_n/2)e^{-c_5\underline{\nu}_n}}{\lambda(\tau_n(\nu_1), \underline{\tau}_n)}.$$

Here $\tau_n(\nu_1) = \tau_n(c_7\underline{\nu}_n) = c_8\underline{\tau}_n$, for constants $c_7, c_8 \in (0,1)$. The right side tends to zero by the same arguments as before, provided that c_5 is sufficiently big. This can be achieved by choosing L and hence c_2 sufficiently big.

The final assertion of the theorem follows by the arguments in Section 5. These show that the posterior distributions $\Pi_\tau(\cdot \mid X)$ have the desired properties uniformly for $\tau \in [\underline{\tau}_n, \overline{\tau}_n]$. The latter interval can be stretched to $[\underline{\tau}_n/2, M\overline{\tau}_n]$ by merely notational changes to the arguments as given.

Appendix

Lemma A.1. *For $r, s, t \geq 0$ with $st > 1$ consider $f(x) = x^{-r}(x^s+1)^{-t}$ and set $f_\nu = \sum_{k=1}^\infty \nu^{-1} f(k/\nu)$. Then as $\nu \to \infty$,*

(i) if $r < 1$, then $f_\nu \asymp \int_0^\infty f(x)\,dx$.
(ii) if $r = 1$, then $f_\nu \asymp \log \nu$.
(iii) if $r > 1$, then $f_\nu \asymp \nu^{r-1} \sum_{k=1}^\infty k^{-r}$.

Proof. Assertion (iii) is an immediate consequence of the fact that $\sum_{k=1}^\infty k^{-r} \times ((k/\nu)^s + 1)^{-t} \to \sum_{k=1}^\infty k^{-r}$, by the dominated convergence theorem. For assertions (i) and (ii) we note that, since f is decreasing,

$$\int_{1/\nu}^\infty f(x)\,dx \leq f_\nu \leq \frac{1}{\nu} f(1/\nu) + \int_{1/\nu}^\infty f(x)\,dx.$$

The integral converges at ∞ by the assumption that $st > 1$. In case (i) it also converges at 0, while $\nu^{-1} f(1/\nu) \asymp \nu^{r-1} \to 0$ as $\nu \to \infty$. In case (ii), we have for every $\varepsilon > 0$, since $\int_a^b x^{-1}\,dx = \log(b/a)$,

$$\frac{\log(\varepsilon\nu)}{(\varepsilon^s+1)^t} \leq \int_{1/\nu}^\varepsilon f(x)\,dx \leq \log(\varepsilon\nu).$$

This shows that $\int_{1/\nu}^\infty f(x)\,dx \asymp \log \nu$. Finally $\nu^{-1} f(1/\nu) \to 1$ in this case. □

References

[1] Belitser, E., and Enikeeva, F. Empirical Bayesian test of the smoothness. *Math. Methods Statist. 17*, 1 (2008), 1–18. MR2400361

[2] Belitser, E., and Ghosal, S. Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist. 31*, 2 (2003), 536–559. Dedicated to the memory of Herbert E. Robbins. MR1983541

[3] Brown, L. D., and Zhao, L. H. Estimators for Gaussian models having a block-wise structure. *Statist. Sinica 19*, 3 (2009), 885–903. MR2536135

[4] Cai, T. T., Low, M. G., and Zhao, L. H. Sharp adaptive estimation by a blockwise method. *J. Nonparametr. Stat. 21*, 7 (2009), 839–850. MR2572586

[5] Castillo, I. Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat. 2* (2008), 1281–1299. MR2471287

[6] Cox, D. D. An analysis of Bayesian inference for nonparametric regression. *Ann. Statist. 21*, 2 (1993), 903–923. MR1232525

[7] Donoho, D. L. Statistical estimation and optimal recovery. *Ann. Statist. 22*, 1 (1994), 238–270. MR1272082

[8] Donoho, D. L., Liu, R. C., and MacGibbon, B. Minimax risk over hyperrectangles, and implications. *Ann. Statist. 18*, 3 (1990), 1416–1437. MR1062717

[9] Efroĭmovich, S. Y., and Pinsker, M. S. Estimation of square-integrable probability density of a random variable. *Problemy Peredachi Informatsii 18*, 3 (1982), 19–38. MR0711898

[10] Efron, B., and Morris, C. Limiting the risk of Bayes and empirical Bayes estimators. II. The empirical Bayes case. *J. Amer. Statist. Assoc. 67* (1972), 130–139. MR0323015

[11] Florens, J., and Simoni, A. Regularizing priors for linear inverse problems. *preprint*.

[12] Freedman, D. On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist. 27*, 4 (1999), 1119–1140. MR1740119

[13] Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. Convergence rates of posterior distributions. *Ann. Statist. 28*, 2 (2000), 500–531. MR1790007

[14] Ghosal, S., Lember, J., and van der Vaart, A. Nonparametric Bayesian model selection and averaging. *Electron. J. Stat. 2* (2008), 63–89. MR2386086

[15] Ghosh, J. K., and Ramamoorthi, R. V. *Bayesian nonparametrics*. Springer Series in Statistics. Springer-Verlag, New York, 2003. MR1992245

[16] Jiang, W., and Zhang, C.-H. General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist. 37*, 4 (2009), 1647–1684. MR2533467

[17] Johnstone, I. M., and Silverman, B. W. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist. 32*, 4 (2004), 1594–1649. MR2089135

[18] Knapik, B. T., Szabó, B., van der Vaart, A. W., and van Zanten, J. H. Bayes procedures for adaptive inference in nonparametric inverse problems. *Preprint, arXiv:1209.3628 [math.ST]* (2012).

[19] Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. Bayesian inverse problems with Gaussian priors. *Ann. Statist. 39*, 5 (2011), 2626–2657. MR2906881

[20] Lember, J., and van der Vaart, A. On universal Bayesian adaptation. *Statist. Decisions 25*, 2 (2007), 127–152. MR2388859

[21] Pinsker, M. S. Optimal filtration of square-integrable signals in Gaussian noise.

[22] Robbins, H. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I* (Berkeley and Los Angeles, 1956), University of California Press, pp. 157–163. MR0084919

[23] Scricciolo, C. Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. *Ann. Statist. 34*, 6 (2006), 2897–2920. MR2329472

[24] Tsybakov, A. B. *Introduction à l'estimation non-paramétrique*, vol. 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2004. MR2013911

[25] van der Vaart, A., and van Zanten, H. Bayesian inference with rescaled Gaussian process priors. *Electron. J. Stat. 1* (2007), 433–448 (electronic). MR2357712

[26] van der Vaart, A. W., and van Zanten, J. H. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist. 36*, 3 (2008), 1435–1463. MR2418663

[27] van der Vaart, A. W., and van Zanten, J. H. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist. 37*, 5B (2009), 2655–2675. MR2541442

[28] van der Vaart, A. W., and Wellner, J. A. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics. MR1385671

[29] Wahba, G. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B 40*, 3 (1978), 364–372. MR0522220

[30] Zhang, C.-H. General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist. 33*, 1 (2005), 54–100. MR2157796

[31] Zhao, L. H. Bayesian aspects of some nonparametric problems. *Ann. Statist. 28*, 2 (2000), 532–552. MR1790008