Can we trust Bayesian uncertainty quantification from Gaussian process priors with squared exponential covariance kernel?

Hadji, M.A.; Szabo, B.

SIAM/ASA J. Uncertainty Quantification
Vol. 9, No. 1, pp. 185–230

Can We Trust Bayesian Uncertainty Quantification from Gaussian Process Priors with Squared Exponential Covariance Kernel?*

Amine Hadji† and Botond Szabó‡

Abstract. We investigate the frequentist coverage properties of credible sets resulting from Gaussian process priors with squared exponential covariance kernel. First, we show that by selecting the scaling hyperparameter using the maximum marginal likelihood estimator in the (slightly modified) squared exponential covariance kernel, the corresponding L_2-credible sets will provide overconfident, misleading uncertainty statements for a large, representative subclass of the functional parameters in the context of the Gaussian white noise model. Then we show that by either blowing up the credible sets with a logarithmic factor or modifying the maximum marginal likelihood estimator with a logarithmic term, one can get reliable uncertainty statements and adaptive size of the credible sets under some additional restriction. Finally, we demonstrate in a numerical study that the derived negative and positive results extend beyond the Gaussian white noise model to the nonparametric regression and classification models for small sample sizes as well. The performance of the squared exponential covariance kernel is also compared to the Matérn covariance kernel.

Key words. credible set, frequentist coverage, empirical Bayes, adaptation, asymptotics

AMS subject classifications. 68Q25, 68R10, 68U05

DOI. 10.1137/19M1253010

1. Introduction.

Bayesian methods are routinely used in various fields of applications. One very appealing advantage of the Bayesian framework is that it readily provides built-in uncertainty quantification. In nonparametric problems the remaining uncertainty in the Bayesian procedure is visualized via plotting credible bands, i.e., bands accumulating a prescribed fraction (typically 95%) of the posterior mass. Gaussian processes (GPs) are popular and frequently used choices for prior distributions in high- and infinite-dimensional models. Areas of possible application include machine learning [23], astronomy [11], genomics [17], linguistics [20], and epidemiology [3]. Gaussian processes are characterized by their mean and covariance kernel. Typically one considers centered Gaussian priors. For the covariance kernel, arguably one of the most frequently used choices is the squared exponential kernel $K(\cdot,\cdot)$, i.e., for centered GP G_t, $t \in T$,

$$K(s,t) = E[G_s G_t] = b\exp\{-a(t-s)^2\}, \qquad s,t \in T, \tag{1.1}$$

*Received by the editors March 28, 2019; accepted for publication (in revised form) September 21, 2020; published electronically February 9, 2021. The research was executed while the second author was working at Leiden University.
https://doi.org/10.1137/19M1253010
Funding: This research was supported by the Netherlands Science Foundation NWO, and by the European Research Council under ERC Grant Agreement 320637.

†Mathematical Institute, Leiden University, Leiden, 2333CA, Netherlands (m.a.hadji@math.leidenuniv.nl).
‡Department of Mathematics, Vrije Universiteit Amsterdam, 1081 HV Amersterdam, Netherlands (b.t.szabo@vu.nl).

for given hyperparameters $a, b > 0$; see, for instance, [23]. In this paper, we focus on the effect of the hyperparameter a, which has been investigated in the context of various models, including nonparametric regression, density estimation, and classification; see, for instance, [35, 8]. We note that the behavior of a is qualitatively comparable for any fixed b. Hence we take the hyperparameter $b = 1$ fixed for simplicity.

In our work we adopt a frequentist perspective; i.e., we assume that the data is generated from some unknown but fixed probability distribution P_{f_0} indexed by a true underlying functional parameter of interest f_0. Our goal is to recover f_0 using Bayesian methodology and to reliably quantify the corresponding uncertainty. In other words, we are interested in the limitations and guarantees of Bayesian techniques for recovering the underlying functional parameter f_0 and quantifying our confidence in the procedure. Please note that our model and goals are substantially different than kriging (or GP emulation) [1], where we observe a computationally expensive deterministic computer model f at a limited number of points, and our goal is to predict its value at other points. In our analysis we consider the asymptotic regime while we assume that sample size or signal-to-noise ratio increases indefinitely. The frequentist properties of Bayesian methods have been extensively studied in the literature in general high-dimensional and nonparametric settings; see, for instance, [12, 13]. In these papers it was shown that under relatively mild conditions on the prior and the likelihood function the posterior distribution contracts around the true parameter of interest at (in many instances) an optimal rate in various high-dimensional and nonparametric problems. These results show not only that the point estimators (e.g., posterior mean, δ-posterior mode) resulting from the posterior provide typically good recovery of the truth under the frequentist data generating process but also that the spread of the posterior is not too large. Therefore most of the posterior mass is concentrated in a ball centered around the underlying functional parameter f_0 with optimal diameter.

In the literature, due to its importance the posterior associated to GP priors were intensively investigated. With appropriately chosen scaling or regularity hyperparameters, depending on the smoothness of the underlying functional parameter of interest, they can provide a nearly optimal (minimax) contraction rate (see, for instance, [35, 37, 8, 5, 40, 14]) and reliable confidence statements [18, 9, 41, 39]. However, in practice the regularity of the underlying function is typically not known, and hence it is not feasible to tune the GP prior manually. To overcome this problem one either endows the hyperparameters with an additional layer of prior distribution, resulting in the so-called hierarchical prior distribution, or estimates them from the data (typically using the maximum marginal likelihood estimator). Both of these methods (typically) result in optimal recovery adapting to the (unknown) regularity of f_0; see, for instance, [36, 31, 27, 14].

In our work we focus on investigating the reliability of the empirical and hierarchical Bayes procedures for GP priors for uncertainty quantification. Under the assumption that our observations are generated via a true f_0, we investigate whether the credible sets contain this function. In our theoretical analysis we consider the Gaussian white noise model, which is closely related to various nonparametric models and can be thought of as the idealized, continuous observation version of the nonparametric regression model.

Although Bayesian methods are routinely used for uncertainty quantification, it is rather unclear whether they can provide reliable confidence statements. In fact, it is well known that

it is impossible to construct confidence sets (based on either frequentist or Bayesian methods) which have optimal size over a wide range of regularity classes and provide reliable uncertainty quantification simultaneously. More precisely, let us assume that we have a collection of functional classes Θ^β indexed by some regularity hyperparameter $\beta \in B$, and let us denote the minimax estimation rate corresponding to this class by $r_{n,\beta}$ (where n denotes the sample size or signal-to-noise ratio). Then, in general, it is not possible to construct an "honest" confidence set $\hat{C}_n$ which achieves simultaneously that

$$\liminf_n \inf_{\beta \in B} \inf_{f \in \Theta^\beta} P_f(f \in \hat{C}_n) \geq 1 - \alpha,$$
$$\liminf_n \inf_{\beta \in B} \inf_{f \in \Theta^\beta} P_f(\|f\| \leq \hat{C} r_{n,\beta}) \geq 1 - \alpha$$

for some given significance level $\alpha > 0$; see, for instance, [21, 7, 26, 16]. Therefore one has to introduce additional assumptions on the functional parameter f_0 to obtain confidence sets with optimal size and reliable uncertainty quantification. In the literature, additional, arguably mild, constraints were proposed to overcome this problem; see, for instance, the monograph [16] for a collection of such approaches.

The coverage properties of credible sets have been investigated only recently; see, for instance, [32, 30, 28, 2, 29, 25] and references therein for various combinations of models and GP priors. In these papers it was shown that for appropriate choices of the prior distribution both the hierarchical and empirical Bayes procedures can provide in various nonparametric models reliable uncertainty statements under some additional regularity assumption on the underlying functional parameter (e.g., self-similarity assumption, polished tail condition, excessive bias assumption, etc). In our work we focus on the GP prior with (approximately) squared exponential kernel and show that both the empirical and hierarchical Bayes procedures result in unreliable uncertainty statements for a large, representative class of functions satisfying the above-mentioned mild regularity assumptions. The derived negative theoretical results are also demonstrated via a simulation study and empirically extended to the nonparametric regression and classification models. This troubling finding might shatter our trust in this popular and frequently applied GP prior. However, we propose a simple and intuitive fix for this problem by slightly modifying the maximum likelihood estimator used in the empirical Bayes method. This modification corrects for the haphazard, overconfident uncertainty statements both theoretically and numerically for small sample sizes.

The paper is organized as follows. In section 2.1 we introduce the Gaussian white noise model and the considered Bayesian approach in detail. In section 2.2 we present first our negative findings on the coverage properties of Bayesian L_2-credible sets and then propose different modifications correcting the haphazard behavior of the posterior by either blowing up the credible sets by a logarithmic factor or (slightly) modifying the marginal maximum likelihood estimator. We show that the proposed methods indeed correct the overconfident uncertainty statements and result in reliable uncertainty quantification for polished tail and self-similar functions, respectively. We demonstrate our findings in a simulation study in section 3 and discuss the derived results and possible extensions in section 4. The proofs of the above results are deferred to the appendix. Finally, as a by-product we also derive contraction rates for the empirical and hierarchical Bayes procedures for a wide range of

priors on the rescaling hyperparameter, extending the results available in the literature. The contraction rate results and their proofs are also deferred to the appendix.

1.1. Notation. For two positive sequences a_n, b_n we use the notation $a_n \lesssim b_n$ if there exists a universal positive constant C such that $a_n \leq Cb_n$. Along the same lines $a_n \asymp b_n$ denotes that $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously. For $f \in L_2[0,1]$ we denote the standard L_2-norm as $\|f\|_2^2 = \int_0^1 f(x)^2 dx$ and let $\mathrm{diam}(S)$ denote the ℓ_2-diameter of the set $S \subset \ell_2$. Throughout the paper, c and C denote global constants whose values may change from one line to another. The dependence of the constants c, C on the model parameters are denoted by subindexes, e.g., c_β, $C_{\beta,m,M}$.

2. Main results.

2.1. Model description. We consider the Gaussian white noise model

$$Y(t) = \int_0^t f_0(s)ds + \frac{1}{\sqrt{n}} W_t, \quad t \in [0,1], \tag{2.1}$$

where $f_0 \in L_2[0,1]$ is the unknown function of interest, and W_t denotes the Brownian motion. Let P_0 and E_0 denote the corresponding probability measure and expected value, respectively. This model is closely related to the popular nonparametric regression and density estimation models [22, 6] and can be used as a platform for investigating more complex statistical models; see, for instance, [33, 16]. In the Bayesian approach we endow the unknown function of interest f_0 with a prior distribution representing our initial belief. In our work we investigate the popular GP prior with rescaled squared exponential kernel (1.1). Let us consider the sequence representation of the Gaussian white noise model. For an orthonormal basis ψ_i, $i = 1, 2, \ldots$, (e.g., the Fourier basis) let us denote the sequence decomposition of the functions $f_0(t), Y(t)$, and W_t by $Y_i = \langle Y(t), \psi_i(t) \rangle_2$, $f_{0,i} = \langle f_0, \psi_i(t) \rangle_2$, and $Z_i = \langle W_t, \psi_i(t) \rangle_2 \overset{iid}{\sim} N(0,1)$, $i = 1, 2, \ldots$, respectively. Then the equivalent sequence model can be given in the form

$$Y_i = f_{0,i} + \frac{1}{\sqrt{n}} Z_i, \quad i = 1, 2, \ldots.$$

Slightly abusing our notation, we denote by f_0 both the functional parameter in the Gaussian white noise model and the sequential parameter $f_0 = (f_{0,1}, f_{0,2}, \ldots)$ in the sequence model. It is common to assume that the true function f_0 belongs to a hyperrectangle regularity class, i.e.,

$$f_0 \in \Theta^\beta(M) = \left\{ f \in \ell_2 : \sup_{i \geq 1} f_i^2 i^{2\beta+1} \leq M \right\},$$

for some (typically unknown) $\beta, M > 0$. The class $\Theta^\beta(M)$ is closely related to Sobolev-type regularity classes $S^\beta(M) = \{f \in \ell_2 : \sum_{i \geq 1} f_i^2 i^{2\beta} \leq M\}$, and the derived results can easily be extended to them; see, for instance, [32]. We note that the minimax estimation rate for the above hyperrectangle regularity class is $n^{-\beta/(1+2\beta)}$; i.e., there exists $C_\beta > 0$ such that

$$\inf_{\hat{f}} \sup_{f \in \Theta^\beta(M)} E_0 \|f - \hat{f}\|_2 \geq C_\beta n^{-\beta/(1+2\beta)},$$

where the infimum is taken over all estimators; see, for instance, [10].

In view of Mercel's theorem we can represent the GP prior with squared exponential kernel as

$$G_t = \sum_{i=1}^{\infty} \lambda_i \xi_i \psi_i(t),$$

where λ_i, ψ_i, $i = 1, 2, \ldots$, are the eigenvalues and eigenfunctions of the squared exponential covariance kernel, and ξ_i are independent and identically distributed (i.i.d.) standard normal random variables; see, for instance, Chapter 4.3 of [24]. The corresponding coefficients λ_i can be approximated as $\lambda_i^2 \approx a^{-1} e^{-i/a}$; see, for instance, [24, 4]. In the rest of the paper, for convenience we (mainly) work with the prior

$$f|a \sim \bigotimes_{i=1}^{\infty} N(0, a^{-1} e^{-i/a}) \tag{2.2}$$

in the sequence model. Note that in view of $Y|f_0 \sim \bigotimes_{i=1}^{\infty} N(f_{0,i}, n^{-1})$ and the choice of the prior $\Pi_a(\cdot)$ in (2.2), the corresponding posterior $\Pi_a(\cdot|Y)$ takes the form

$$f\Big|a, Y \sim \bigotimes_{i=1}^{\infty} \mathcal{N}\left(\frac{nY_i}{ae^{i/a} + n}, \frac{1}{ae^{i/a} + n}\right). \tag{2.3}$$

The behavior of the posterior distribution is very sensitive to the choice of the hyperparameter a. Since the optimal choice of a depends on the (typically) unknown regularity parameter β of the underlying functional parameter of interest f_0, in practice one uses data-driven procedures for selecting a. The two most commonly applied Bayesian techniques for selecting the hyperparameter are the hierarchical Bayes and the marginal likelihood empirical Bayes methods. In the hierarchical Bayes method the hyperparameter a is endowed with a prior distribution π (also called hyperprior distribution), resulting in a two-level, hierarchical prior distribution

$$\Pi(\cdot) = \int_0^{\infty} \Pi_a(\cdot)\pi(a)da.$$

For technical reasons, we introduce the following assumptions on the hyperprior density function $\pi(\cdot)$ supported on $[1, A_n]$.

Assumption 1. Let us assume that for some $c_1 > 0$ there exist $c_2, c_6 \geq 0$ and $c_3, c_4, c_5 > 0$ such that

$$c_4^{-1} a^{-c_3} \exp(-c_2 a) \leq \pi(a) \leq c_4 a^{-c_5} \exp(-c_6 a) \tag{2.4}$$

for all $c_1 \leq a \leq A_n$.

Note that among others, the exponential, the gamma, and the inverse gamma distributions (restricted to $[1, A_n]$) satisfy Assumption 1.

In contrast to this, in the empirical Bayes approach we take the maximum marginal likelihood estimator (MMLE), i.e.,

$$\hat{a}_n := \arg \max_{a \in [1, A_n]} \ell_n(a), \tag{2.5}$$

where the marginal log-likelihood function (with respect to the measure $\bigotimes_{i=1}^{\infty} N(0,1)$) is

$$\ell_n(a) = -\frac{1}{2}\sum_{i=1}^{\infty}\left(\log\left(1+\frac{n}{ae^{i/a}}\right) - \frac{n^2 Y_i^2}{ae^{i/a}+n}\right)$$

and the parameter $A_n = o(n)$ restricts the parameter space to a compact interval, which is advantageous from both practical and analytical perspectives. Then the estimator $\hat{a}_n$ is plugged into the posterior distribution (2.3), resulting in the empirical Bayes posterior $\Pi_{\hat{a}_n}(\cdot|Y)$.

We show in section A.2 that both of these methods result in optimal recovery for the functional parameter of interest f_0. These results are of interest in their own right, but our main focus lies on the reliability of Bayesian uncertainty quantification resulting from both the hierarchical and the empirical Bayes procedures, and hence we have deferred the contraction rate results to the appendix.

2.2. Uncertainty quantification. In our work we investigate the reliability of the built-in uncertainty quantification of the above data-driven posterior distributions. For convenience let $\Pi_n(\cdot|Y)$ denote both the hierarchical and the empirical Bayes posterior distributions in the following. In Bayesian methods the remaining uncertainty of the procedure is visualized by the credible set. We consider ℓ_2-credible balls centered around the posterior mean; i.e., we analyze credible sets in the form

$$\hat{C}_{n,\alpha} = \{f \in \ell_2 : \|f - \hat{f}\|_2 \le r_\alpha\}, \tag{2.6}$$

where $\hat{f}$ is the posterior mean, and the radius r_α is chosen such that $\Pi(f \in \hat{C}_{n,\alpha}|Y) = 1-\alpha$ for some prescribed significance level $\alpha > 0$.

We are interested in the frequentist properties of ℓ_2-credible balls resulting from the data-driven credible balls. We denote by r_α the radius of the ℓ_2-ball centered around the posterior mean $\hat{f}$ and accumulating a $1-\alpha$ fraction of the posterior mass, i.e.,

$$\Pi(f : \|f - \hat{f}\|_2 \le r_\alpha|Y) = 1-\alpha.$$

In our analysis we introduce some additional flexibility by considering inflated credible balls, i.e..

$$\hat{C}_n(L_n) = \{f : \|f - \hat{f}\|_2 \le L_n r_\alpha\}, \tag{2.7}$$

for some blown up factor $L_n \ge 1$ possibly depending on n. As a first step we note that the size of the credible set for both the empirical and the hierarchical Bayes procedures adapts to the minimax rate (actually, the diameter of the set is even a logarithmic factor faster than the minimax rate for the empirical Bayes method).

Corollary 2.1. *Both the hierarchical and the empirical Bayes credible sets defined in* (2.7) *have rate adaptive size; i.e., for every $\beta_0 > 0$ and $M > 0$,*

$$\sup_{\beta\ge\beta_0}\sup_{f\in\Theta^\beta(M)} P_{\theta_0}\left(diam(\hat{C}_n(1)) \ge M_n n^{-\beta/(1+2\beta)}(\log n)^{-1/(1+2\beta)}\right) \to 0,$$

where the sequence M_n goes to infinity arbitrarily slowly in the case of the empirical Bayes method, and $M_n \gg \log n$ in the case of the hierarchical Bayes method.

Proof. The proof is given in Appendices D and G. ∎

2.2.1. Coverage of credible sets—negative results. Next we investigate how much we can trust the above derived data-driven Bayesian uncertainty quantification from a frequentist perspective. We would like to know whether the true function f_0 is included in the (blown up) credible set, i.e., if any fixed $\beta_0 > 0$,

$$\inf_{f_0 \in \cup_{\beta \geq \beta_0} \Theta^\beta(M)} P_0(f_0 \in \hat{C}_n(L_n)) \geq 1 - \alpha,$$

holds for some sufficiently large choice of L_n. Since it is impossible to construct honest confidence sets with rate adaptive size, and in view of the adaptive size of the credible sets (see Corollary 2.1), they must have poor frequentist coverage properties at least for certain functional parameters f_0. Actually the radius of the credible sets decays even faster than the minimax rate for the empirical Bayes method, which already implies impossibility of coverage. Nevertheless it is of interest to quantify the set of functions for which the Bayesian uncertainty quantification is truthworthy.

First, we note that a representative subset of the hyperrectangle $\Theta^\beta(M)$ is the set

$$\Theta_s^\beta(m, M) = \left\{ f \in \Theta^\beta(M) : \min_{i \geq 1} i^{1+2\beta} f_i^2 \geq m \right\} \tag{2.8}$$

for some parameters $0 < m \leq M$. Let us refer to this subclass of sequential parameters as self-similar signals following similar terminology in [15, 32]. It was shown in the latter paper that the minimax rate over $\Theta_s^\beta(m, M)$ is the same as that over $\Theta^\beta(M)$. The next theorem shows that both the hierarchical and the empirical Bayes procedures provide unreliable uncertainty quantification over this representative subclass of functions unless it is blown up with at least a logarithmic factor.

Theorem 2.2. *Let us take arbitrary $L_n = o(\sqrt{\log n})$. Then the empirical and hierarchical Bayes credible sets blown up by L_n have frequentist coverage tending to zero for every self-similar signal, i.e., for every $0 < m \leq M$,*

$$\sup_{f_0 \in \Theta_s^\beta(m,M)} P_0(f_0 \in \hat{C}_n(L_n)) \to 0.$$

Proof. See Appendix D. ∎

This negative result draws a dark picture, as it tells us that one cannot trust Bayesian uncertainty quantification resulting from the investigated prior even if one allows a certain amount of adjustment (i.e., by blowing up (not too fast) the set with a sequence tending to infinity). Since the prior (2.2) is very closely related to the GP with squared exponential covariance kernel, this gives the intuition that one has to be very cautious when working with squared exponential kernel, as the corresponding Bayesian uncertainty statements are (typically) unreliable. In the next subsection we will be completing the results we have by deriving some positive results on the coverage properties of the credible sets. First, we show that for analytic functions the (slightly inflated) credible sets provide reliable uncertainty quantification, and then we show that by either blowing up the credible sets by a logarithmic factor or slightly adjusting the maximum marginal likelihood estimator, one gets reliable uncertainty statements for a large subclass of functions, including the self-similar functions.

2.2.2. Coverage of credible sets—positive results. Let us consider the set of analytic-type functions defined as

$$f_0 \in A^\gamma(M) = \left\{ f \in \ell_2 : \sum_{i=1}^{\infty} f_i^2 e^{2i\gamma} \leq M \right\}$$

for some $\gamma > 0$. Note that the investigated prior (2.2) is more suitable for this class of functions due to the exponential decay of the variances. We show below that, indeed, for the class $A^\gamma(M)$ both the empirical and the hierarchical Bayes procedures provide reliable uncertainty quantification. Note, however, that the present class of functions is substantially smaller than $\Theta^\beta(M)$ for any $\beta > 0$.

Theorem 2.3. *The inflated empirical and hierarchical Bayes credible sets $\hat{C}_n(L)$ have frequentist coverage tending to one over the class $f_0 \in A^\gamma(M)$ for any $\gamma \geq 1/2$ and sufficiently large constant $L > 0$, i.e.,*

$$\inf_{f_0 \in A^\gamma(M)} P_0(f_0 \in \hat{C}_n(L)) \to 1.$$

Furthermore, the size of the credible set is (nearly) optimal; i.e., for some sufficiently large constant $C > 0$,

$$\inf_{f_0 \in A^\gamma(M)} P_0\big(diam(\hat{C}_n(1)) \leq C n^{-1/2} \log n\big) \to 1.$$

Proof. See Appendix C. ∎

Next we investigate the behavior of the credible sets by allowing a logarithmic inflating factor. Since the size of the inflated credible sets is still nearly minimax, the credible sets fail to cover all functional parameter f_0 of interest in view of the nonexistence result of adaptive confidence sets [7, 26]. Therefore we restrict the investigated class of functions to the so-called polished tail class introduced in [32, 28]. We say that a sequential parameter $f \in \ell_2(M)$ belongs to the class of polished tail signals denoted by $\Theta_{pt}(L_0, N_0, \rho)$ for some $L_0, \rho, N_0 > 0$ if

$$\sum_{i=N}^{\infty} f_i^2 \leq L_0 \sum_{i=N}^{\rho N} f_i^2 \quad \text{for all } N \geq N_0.$$

The above assumption basically requires that the knowledge of the sequential parameter f up to a certain coordinate enables us to draw conclusions about the tail of the sequence. We require the energy (sum of the squared coefficients) of the tail to be dominated by the energy of a finitely large block of coefficients. This condition also makes sense intuitively; as the signal can be observed only up to some limit in the stochastic model, the fluctuation in the later coordinates can equally likely be caused by the noise. Therefore to make reliable uncertainty statements we have to assume that the tail behavior of the signal hidden by the noise is not substantial and can be extrapolated by information available at a given signal-to-noise ratio. In [32] it was shown that the above assumption is mild from statistical, topological, and Bayesian perspectives as well.

The next theorem states that when the sequential parameter f_0 is restricted to polished tail sequences, both the empirical and the hierarchical Bayes credible balls blown up by a $\log n$ factor (i.e., $\hat{C}_n(L \log^{3/2} n)$) are an honest frequentist confidence set for large enough L.

Theorem 2.4. *For any $L_0, N_0, \rho \geq 1$ there exists a constant L such that*

$$\inf_{f_0 \in \Theta_{pt}(L_0,N_0,\rho)} P_0(f_0 \in \hat{C}_n(L \log^{3/2} n)) \to 1,$$

where $\hat{C}_n$ denotes either the empirical or the hierarchical Bayes credible sets under Assumption 1.

Proof. See Appendix B and section G.3. ∎

So one can achieve reliable uncertainty quantification on an arguably large subset of the function space by blowing up the standard credible set with a slowly varying term. This, however, is not very appealing, as a practitioner would righteously hesitate to introduce the artificial logarithmic blow up. Therefore, we propose another method, where one does not have to introduce a logarithmic blow up factor but instead adjust the maximum marginal likelihood estimator. Investigating the proof of Theorem 2.2 one can see that the MMLE $\hat{a}_n$, given in (2.5), is too small, and that the empirical Bayes procedure is basically oversmoothing. One can compensate for this by undersmoothing the procedure. We propose adjusting the MMLE by a multiplicative logarithmic factor of

$$\tilde{a}_n = \log(n)\hat{a}_n. \tag{2.9}$$

Then the corresponding empirical Bayes credible set (blown up by a sufficiently large constant $L > 0$) results in reliable uncertainty quantification for self-similar functions $\Theta_s^\beta(m, M)$.

Theorem 2.5. *For any $0 < m \leq M$ there exists a constant $L > 0$ such that*

$$\inf_{f_0 \in \Theta_s^\beta(m,M)} P_0(f_0 \in \tilde{C}_n(L)) \to 1,$$

where $\tilde{C}_n(1)$ denotes the credible set resulting from the empirical Bayes posterior with hyperparameter $\tilde{a}_n$.

Proof. The proof of the theorem is deferred to Appendix F. ∎

3. Numerical analysis. In this section we investigate the numerical properties of the GP prior with (approximately) squared exponential covariance kernel. First, we consider the Gaussian white noise model and the prior (2.2). We show that the corresponding Bayesian uncertainty quantification is misleading for various regularly behaving functions. We also demonstrate that a different choice of the covariance kernel or a modified version of the empirical Bayes procedure results in more accurate uncertainty statements. Then we consider (from practical point of view) the more relevant nonparametric regression and classification models, where we also demonstrate the suboptimal behavior of the (standard) empirical Bayes method with squared exponential covariance kernel and show that the proposed modification results in superior performance compared to it. We also consider GP priors with Matérn covariance kernels and show that although they pose good recovery and uncertainty quantification properties, their run times are substantially slower than using squared exponential kernels. We have carried out the simulations using an Intel Core i7-8700 CPU operating at 3.40 GHz with 16 GB of RAM.

3.1. Gaussian white noise model. First, we demonstrate the suboptimal performance of the GP with (approximately) squared exponential covariance kernel (2.2) compared to modified versions of the empirical Bayes procedure and to the GP prior with polynomially decaying variances in the series representation; see [19, 32]. Let us consider the function $f_1 \in L_2[0,1]$ given by their Fourier coefficients $f_{1,i} = i^{-3/2}\sin(i)$, for $i = 1, 2, \ldots$, respectively, relative to the Fourier eigenbasis $\psi_i(t) = \sqrt{2}\cos(\pi(i-1/2)t)$. Note that although the function lies outside of the self-similar function class (2.8), it has essentially the same behavior. In Figure 1 we visualize the 95% credible sets (light blue or light red), the posterior mean (blue or red), and the true function (black) by simulating 2000 draws from the empirical Bayes posterior distribution and plotting the closest 95% of them in the L_2-norm to the posterior mean. We note that all credible sets were constructed without any inflation factor, i.e., $L_n = 1$ was taken (except for the case where the choice $L_n = \log n$ was prespecified). The credible sets are drawn for signal-to-noise ratio $n = 100$, 500, 1000, and 5000, respectively. We also plot the same credible sets blown up by a $\log n$ factor, the credible sets obtained by the modified empirical Bayes procedure (where the MMLE $\hat{a}_n$ of the scaling parameter a was multiplied by $\log n$), and the empirical Bayes credible sets corresponding to the prior $f \sim \otimes_{i=1}^{\infty} N(0, i^{-1-2\alpha})$, with hyperparameter α estimated by the MMLE. One can see that the standard marginal likelihood empirical Bayes method provides credible sets that are too narrow and fail to cover the underlying true function. Also note that both modification of the empirical Bayes credible sets and use of the prior with polynomially decaying variances provide good coverage, but in contrast to the overly conservative approach of inflating the credible sets with a logarithmic factor, the modification of the MMLE results in a more informative uncertainty statement (i.e., smaller credible sets).

3.2. Nonparametric regression and classification. In this section we demonstrate in a simulation study that the results derived for the Gaussian white noise model generalize to more complicated statistical models as well. We consider specifically the popular nonparametric regression and classification models. The empirical Bayes posteriors, posterior means, and credible sets are computed in both cases using the MATLAB package GPML.

In the nonparametric regression model we assume we are observing pairs of random variables $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, where

$$Y_i = f_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad X_i \overset{iid}{\sim} U(0,1), \quad i = 1, \ldots, n,$$

and the aim is to estimate the unknown nonparametric regression function f_0. In the Bayesian approach we endow f_0 with a GP prior with squared exponential kernel and estimate the tuning parameter using the MMLE.

In this simulation study we take the Fourier coefficients of the underlying true function f_2 to be $f_{2,i} = i^{-3/2}\cos(i)$, $i = 1, 2, \ldots$. We take $\sigma^2 = 1/2$, but in the procedure it is considered to be unknown and estimated with the MMLE $\hat{\sigma}^2$. We plot the true function (black), the posterior mean (blue), and the posterior pointwise credible intervals (dashed blue) $[\hat{f}(x) - q_{0.025}\sqrt{\hat{c}(x,x)}, \hat{f}(x) + q_{0.025}\sqrt{\hat{c}(x,x)}]$, where $\hat{f}$ is the posterior mean, q_α is the αth quantile of the standard normal distribution, and $\hat{c}(.,.)$ is the posterior covariance kernel. We consider the MMLE empirical Bayes method with and without the $\log n$ inflation factor for the credible set, the modified empirical Bayes method (where the MMLE was multiplied by

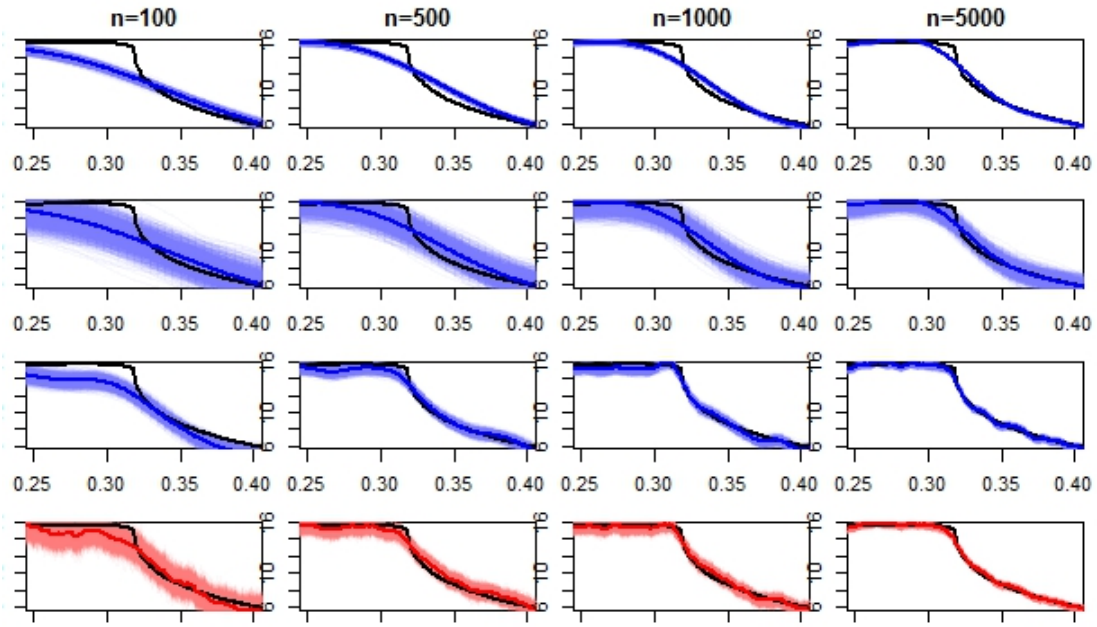

Figure 1. *Empirical Bayes credible sets for the function f_1 (drawn in black) zoomed in to the interval $x \in [0.25, 0.4]$. Top row: credible set (in light blue) and posterior mean (blue curve) corresponding to the prior with exponentially decaying variance. Second row: credible set (in light blue) blown up by a $\log n$ factor ($L_n = \log n$) and posterior mean (blue curve) corresponding to the prior with exponentially decaying variance. Third row: credible set (in light blue) and posterior mean (blue curve) corresponding to the prior with exponentially decaying variance and modified empirical Bayes procedure (the rescaling factor is multiplied by $\log n$). Bottom row: credible set (in light red) and posterior mean (red curve) corresponding to the prior with polynomially decaying variance. From left to right the signal-to-noise ratio is $n = 100, 500, 1000, 5000$.*

$\log n$), and, finally, the empirical Bayes method for Matérn covariance kernel with estimating either the regularity or the scale tuning parameter from the data. We take the sample sizes to be $n = 100, 500, 1000$, and 2000. Observe in Figure 2 that the standard MMLE empirical Bayes method provides unreliable uncertainty quantification at certain points, while the two modified squared exponential credible sets and the empirical Bayes credible sets from the Matérn kernel (with data-driven choice of the regularity hyperparameter) capture the underlying functional parameter of interest better. Also note that by multiplying the MMLE of the scaling parameter by a $\log n$ factor in the squared exponential kernel case we do not get an overly conservative credible set, unlike in the case when the radius is inflated with a logarithmic factor. Finally, we note that the computational times corresponding to the Matérn kernel are higher than for the squared exponential kernel. Estimating the regularity hyperparameter of the kernel is time-consuming as the eigenfunctions depend on it. Alternatively, one can consider a rescaled Matérn covariance kernel with fixed regularity. This method is typically faster; however, optimal recovery of the underlying function is possible only up to the smoothness level $\alpha + d/2$, where α denotes the regularity of the prior; see, for instance, [31]. Therefore, we choose α large enough ($\alpha = 10$), which then seemingly slows down the computations. The run times are collected in Table 3.

We also investigate empirically the frequentist coverage probabilities of the pointwise credible sets by repeating the experiment 100 times and reporting the frequency that the function at given points (we consider $x = (0.25, 0.3188, 0.75)$ with $0.3188 = \text{argmax}_{x\in[0,1]} f_2(x)$) is included in the credible interval; see Table 1. Moreover, Table 2 shows the average size of the pointwise credible intervals (i.e., $2q_{0.025}\sqrt{\hat{c}(x,x)}$) depending on the sample size n and the procedure used to compute the credible sets. We note that our theoretical results concern L_2-credible sets and do not extend directly to pointwise credible intervals. However, the results in Table 1 indicate that the MMLE empirical Bayes method for squared exponential kernel has suboptimal pointwise coverage compared to the other above-mentioned GP methods.

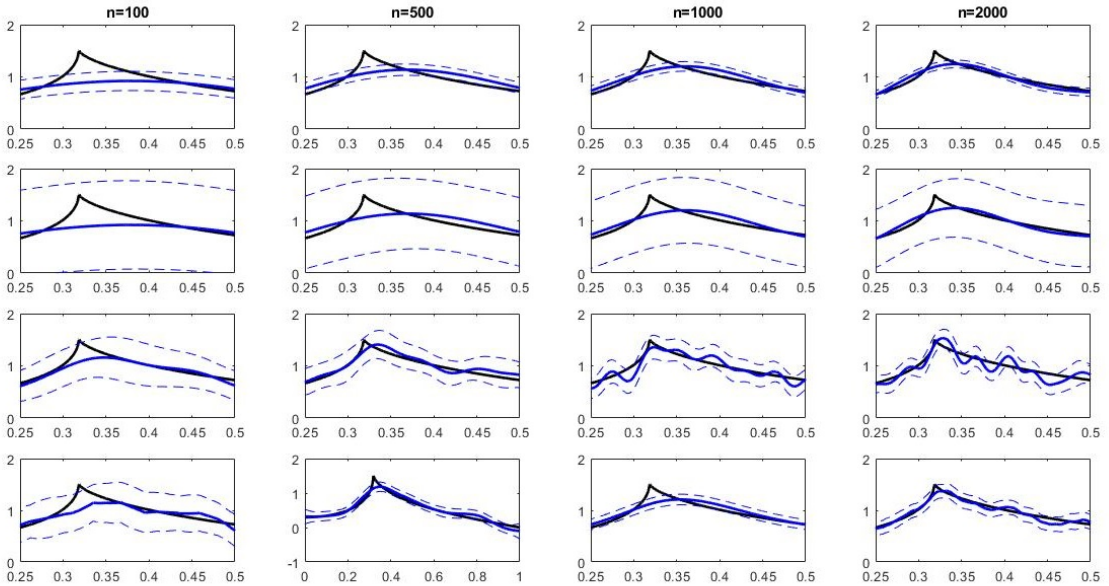

Figure 2. *Empirical Bayes credible sets for the regression function f_2 (drawn in black) zoomed in to the interval $x \in [0.25, 0.5]$. The posterior means are drawn as solid blue lines, while the 95% pointwise credible sets are dashed blue curves. We plot in the top row the MMLE empirical Bayes method, in the second row the MMLE empirical Bayes method with a $\log n$ blow up factor, in the third row the modified MMLE empirical Bayes method using a squared exponential GP prior, and in the last row the empirical Byes credible sets using a Matérn kernel with data-driven choice for the regularity hyperparameter. From left to right the sample size is $n = 100, 500, 1000, 2000$.*

Next, we consider the nonparametric classification problem. Let us assume that we observe the binary random variables $Y_1, Y_2, \ldots, Y_n \in \{0, 1\}$, with

$$P(Y_i = 1) = p(X_i), \qquad X_i \overset{iid}{\sim} U(0,1), \quad i = 1, \ldots, n,$$

for some nonparametric function $p(x) : [0,1] \mapsto [0,1]$. We write $p(x)$ in the form $p(x) = \psi(f(x))$, with $\psi(x) = e^x/(1+e^x)$, for some function $f(x) : [0,1] \mapsto \mathbb{R}$. In the Bayesian approach we endow the functional parameter $f(x)$ with a GP prior with squared exponential or Matérn covariance kernel.

Table 1

Frequency of $f_2(x)$ being inside the corresponding credible interval for the squared exponential and Matérn Gaussian process prior at given points $x \in \{0.25, 0.3188, 0.75\}$. Method 1: squared exponential kernel MMLE empirical Bayes procedure. Method 2: squared exponential kernel empirical Bayes procedure with $\log n$ blow up factor. Method 3: squared exponential kernel modified empirical Bayes procedure (MMLE multiplied by $\log n$). Method 4: Matérn kernel with smoothness MMLE empirical Bayes. Method 5: Matérn kernel with rescaling MMLE empirical Bayes and $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000$.

	$x = 0.25$			$x = 0.3188$			$x = 0.75$		
$n =$	100	500	1000	100	500	1000	100	500	1000
Method 1	0.84	0.69	0.57	0.01	0.01	0.00	0.98	0.92	0.97
Method 2	1.00	1.00	1.00	0.96	0.98	1.00	1.00	1.00	1.00
Method 3	0.98	0.98	0.97	0.35	0.55	0.50	0.99	0.96	0.98
Method 4	0.99	1.00	1.00	0.12	0.35	0.51	1.00	1.00	1.00
Method 5	0.98	1.00	1.00	0.08	0.30	0.47	0.99	1.00	1.00

Table 2

Average size of the pointwise credible intervals (i.e., $2q_{0.025}\sqrt{\hat{c}(x,x)}$) for $f_2(x)$ in the regression model. Method 1: squared exponential kernel MMLE empirical Bayes procedure. Method 2: squared exponential kernel empirical Bayes procedure with $\log n$ blow up factor. Method 3: squared exponential kernel modified empirical Bayes procedure (MMLE multiplied by $\log n$). Method 4: Matérn kernel with smoothness MMLE empirical Bayes. Method 5: Matérn kernel with rescaling MMLE empirical Bayes and $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000$.

$n =$	100	500	1000
Method 1	0.3956	0.2367	0.1814
Method 2	1.8218	1.4711	1.2533
Method 3	0.7541	0.5279	0.4262
Method 4	0.6346	0.4308	0.3446
Method 5	0.5151	0.3338	0.263

Table 3

Average run time of the EB (empirical Bayes) methods for f_2 in the regression model. Method 1: squared exponential covariance kernel. Method 4: Matérn covariance kernel and MMLE for the regularity hyperparameter. Method 5: Matérn covariance kernel and MMLE for the scaling hyperparameter with fixed regularity $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000, 5000$.

$n =$	100	500	1000	5000	10000	200000
Method 1	0.74 s	2.75 s	10.84 s	3.7 m	25.2 m	1.2 h
Method 4	1.48 s	13.93 s	43.83 s	16.7 m	3.8 h	12.5 h
Method 5	1.37 s	11.15 s	33.5 s	12.3 m	2.8 h	10.5 h

We design experiments for the nonparametric classification model that are similar to those for the nonparametric regression model above, with sample sizes $n = 100, 500, 1000$, and 2000 and the same f_2 as above. We plot the pointwise credible intervals for f_2 corresponding to the empirical Bayes procedure, with and without a $\log n$ inflation factor, and to the modified empirical Bayes procedure (where the MMLE is multiplied by a $\log n$ factor); see Figure 3. One can observe that the standard MMLE empirical Bayes procedure produces unreliable uncertainty statements, while by blowing up the credible sets with a logarithmic factor we get overly conservative uncertainty quantifications. These problems are resolved by considering the modified empirical Bayes method, which captures the shape of the underlying functional

parameter better and provides more reliable uncertainty statements. We also collect the empirical estimation of the frequentist coverage probabilities of the underlying functional parameter $f_2(x)$ at points $x = (0.25, 0.3188, 0.75)$ in Table 4 and the computation time for different methods in Table 6, underlying the conclusions drawn from the figures above. Note that, similarly to Table 1, Table 4 provides pointwise, not L_2, coverage results. Moreover, Table 5 shows the size of the average credible interval.

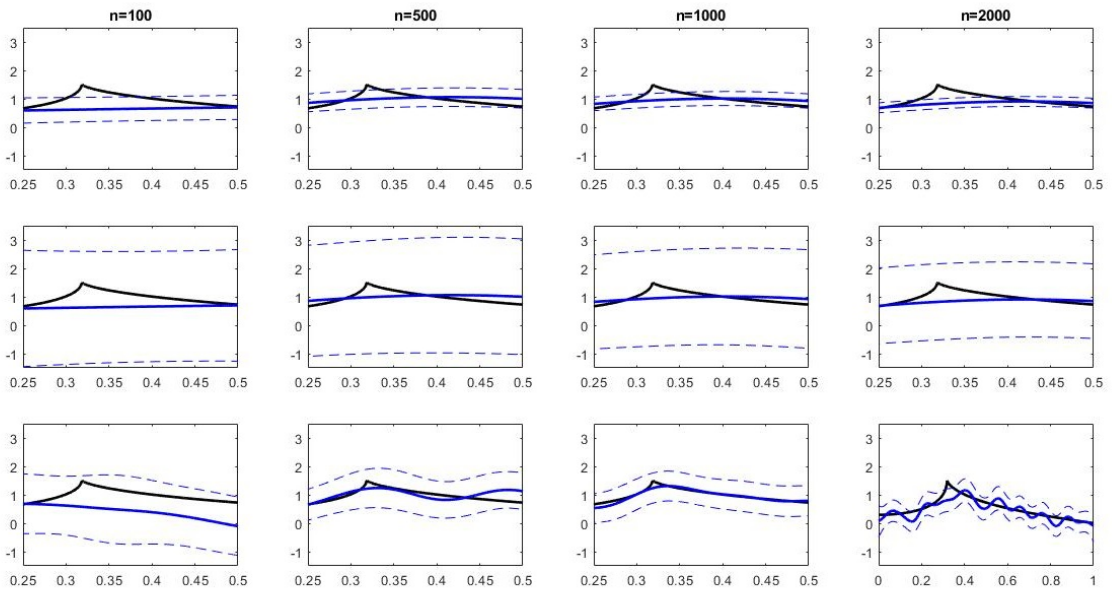

Figure 3. *Empirical Bayes credible sets using squared exponential GP priors in the classification model for the function f_2 (drawn in black). The posterior means are drawn in solid blue line, while the 95% pointwise credible intervals are dashed blue curves. We plotted in the first row the MMLE empirical Bayes method, in the second row the MMLE empirical Bayes method with a $\log n$ blow up factor, and in the the third row the modified MMLE empirical Bayes method. From left to right the sample size is $n = 100, 500, 1000, 2000$.*

Table 4

Frequency of $f_2(x)$ being inside the corresponding credible interval for squared exponential and Matérn GP prior in the logistic regression model. Method 1: squared exponential kernel MMLE empirical Bayes procedure. Method 2: squared exponential kernel empirical Bayes procedure with $\log n$ blow up factor. Method 3: squared exponential kernel modified empirical Bayes procedure (MMLE multiplied by $\log n$). Method 4: Matérn kernel with MMLE for the smoothness. Method 5: Matérn kernel with MMLE for the scaling and taking $\alpha = 10$. From left to right the sample size is $n = 100, 200, 500$.

	$x = 0.25$			$x = 0.3188$			$x = 0.75$		
$n =$	100	200	500	100	200	500	100	200	500
Method 1	0.90	0.90	0.89	0.29	0.16	0.12	0.92	0.88	0.85
Method 2	1.00	1.00	1.00	0.98	0.98	1.00	1.00	1.00	1.00
Method 3	0.91	0.94	0.95	0.42	0.36	0.45	0.94	0.94	0.95
Method 4	0.94	0.96	0.95	0.32	0.27	0.42	0.95	0.96	0.96
Method 5	0.94	0.94	0.96	0.30	0.32	0.35	0.96	0.96	0.96

Table 5

Average size of the pointwise credible intervals $2q_{0.025}\sqrt{\hat{c}(x,x)}$ *for* f_2 *in the logistic regression model. The methods and the sample sizes are the same as in Table 4.*

	$n = 100$	$n = 200$	$n = 500$
Method 1	3.2672	0.8209	0.3485
Method 2	15.0461	4.3495	2.1661
Method 3	3.6777	1.2675	0.7575
Method 4	3.5409	1.1186	0.6212
Method 5	3.4040	0.9698	0.4848

Table 6

Average run time of the EB methods for f_2 *in the logistic regression model. Method* 1: *squared exponential covariance kernel. Method* 4: *Matérn covariance kernel and MMLE for the regularity hyperparameter, Method* 5: *Matérn covariance kernel and MMLE for the scaling hyperparameter with fixed regularity* $\alpha = 10$. *From left to right the sample size is* $n = 100, 500, 1000, 5000$.

$n =$	100	500	1000	5000
Method 1	2.23 s	30.81 s	5.9 m	2.8 h
Method 4	4.77 s	3.1 m	23.9 m	11.1 h
Method 5	4.42 s	3 m	15.1 m	8.2 h

4. Discussion. We have shown that the MMLE empirical Bayes method for the GP prior with (a slightly modified version of the) squared exponential covariance kernel produces a misleading uncertainty statement in the context of the Gaussian white noise model. The derived negative results were demonstrated on a simulation study in the context of the Gaussian white noise model and extended to the nonparametric regression and classification models as well. Hence we can conclude that one has to be very cautious when applying empirical Bayes methods with squared exponential GPs for uncertainty quantification, as typically they provide misleading confidence statements, due to the oversmoothing behavior of the MMLE. We note that the bad performance of the prior (2.2) is not due to the rescaling factor a^{-1} in the variance, because similar (but easier) computations show that the prior without the a^{-1} factor behaves suboptimally as well.

One can compensate for the haphazard uncertainty statements by blowing up the credible sets with a $\log n$ factor; however, this approach is not appealing from a practical perspective, as demonstrated in our simulation study as well. Instead we propose modifying the MMLE by multiplying it with $\log n$ to compensate for the oversmoothing. This procedure is less conservative than the previous one and hence provides more accurate information about the uncertainty of the method. One can also consider different covariance kernels, with polynomially decaying eigenvalues, such as the Matérn kernel; however, these procedures can be computationally less appealing, as demonstrated in the simulation study.

Appendix A. Some properties of the MMLE.

A.1. Deterministic bounds. As a first step we provide deterministic bounds for the marginal maximum likelihood estimator $\hat{a}_n$ of the rescaling hyperparameter a. Let us introduce

the following functions for $a \in [1, \infty)$:

$$h_n(a, f_0) := \frac{1}{\log^2(n/a)} \sum_{i=1}^{\infty} \frac{n^2 i e^{i/a} f_{0,i}^2}{a(ae^{i/a} + n)^2}, \tag{A.1}$$

$$g_n(a, f_0) := \frac{1}{\log^2(n/a)} \sum_{i=2a}^{\infty} \frac{n^2 (i-a) e^{i/a} f_{0,i}^2}{a(ae^{i/a} + n)^2}. \tag{A.2}$$

These functions are derived from the expected value of the score function; see Appendix E. Then let us define the deterministic bounds $\underline{a}_n$ and $\overline{a}_n$ for $\hat{a}_n$ with the help of the functions h_n and g_n as

$$\begin{aligned} \underline{a}_n &:= \sup\{a \in [1, A_n] : g_n(a, f_0) \geq B \log n\}, \\ \overline{a}_n &:= \sup\{a \in [K_0, A_n] : h_n(a, f_0) \geq b\}, \end{aligned} \tag{A.3}$$

with some $b, B, K_0 > 0$ to be specified later and $A_n = o(n)$ given in (2.5). Then we show that these bounds sandwich $\hat{a}_n$ with high probability.

Theorem A.1. *The MMLE $\hat{a}_n$ satisfies*

$$\inf_{f_0 \in \ell_2(M)} P_0(\underline{a}_n \leq \hat{a}_n \leq \overline{a}_n) \to 1 \tag{A.4}$$

for $\underline{a}_n, \overline{a}_n$ defined in (A.3).

Proof. See Appendix E. ∎

We also derive upper bounds for $\overline{a}_n$ in the case when the true function belongs to the hyperrectangle with regularity hyperparameter β or to the analytic function class A^γ and a lower bound for $\underline{a}_n$ in the case of self-similar functions $f_0 \in \Theta^\beta(m, M)$.

Proposition A.2. *For every $\beta \geq \beta_0$ and $\gamma > 0$ there exist $C_{\beta,b,M}, C_{\gamma,b,M} > 0$ such that*

$$\begin{aligned} \sup_{f_0 \in \Theta^\beta(M)} \overline{a}_n &\leq C_{\beta,b,M} n^{1/(1+2\beta)} (\log n)^{-1-1/(1+2\beta)}, \\ \sup_{f_0 \in A^\gamma(M)} \overline{a}_n &\leq C_{\gamma,b,M}, \\ \inf_{f_0 \in \Theta^\beta(m,M)} \underline{a}_n &\geq C_{\beta,B,m} n^{1/(1+2\beta)} (\log n)^{-1-2/(1+2\beta)}. \end{aligned}$$

Proof. Let us start with the proof of the first inequality. We show that for any $b > 0$ the inequality $h_n(a, f_0) < b$ holds for $a \geq C_{\beta,b,M} n^{1/(1+2\beta)} (\log n)^{-1-1/(1+2\beta)}$. Let us introduce the notation $I_a \equiv a \log(n/a)$. Note that by using the inequalities $ae^{i/a} + n \geq n$ and $ae^{i/a} + n \geq ae^{i/a}$, for all $a \geq 1$, and the sum of geometric series, we get

$$\begin{aligned} h_n(a, f_0) &\leq \frac{M}{\log^2(n/a)} \Big(\frac{1}{a} \sum_{i=1}^{I_a} e^{i/a} i^{-2\beta} + \frac{n^2}{a^3} \sum_{i > I_a} e^{-i/a} i^{-2\beta} \Big) \\ &\leq C_\beta \frac{M}{\log^2(n/a)} \Big(I_a^{-2\beta} e^{I_a/a} + \frac{n^2}{a^2} I_a^{-2\beta} e^{-I_a/a} \Big) \\ &\leq 2 C_\beta M a^{-1-2\beta} n \big(\log(n/a) \big)^{-2-2\beta} \end{aligned}$$

for some constant $C_\beta > 0$ depending only on β. For $A_n \geq a \geq Kn^{1/(1+2\beta)}(\log n)^{-1-1/(1+2\beta)}$ the preceding display is bounded by a multiple of $2C_\beta M K^{-1-2\beta}$. Then for a sufficiently large choice of the constant $K = C_{\beta,b,M}$ (depending only on β, b, and M), we get that $h_n(a, f_0) < b$ for any $a \geq C_{\beta,b,M} n^{1/(1+2\beta)}(\log n)^{-1-1/(1+2\beta)}$.

The proof of the second inequality of the statement follows similarly, i.e., we prove that for $a \geq C_{\gamma,b,M}$ we have $h_n(a, f_0) < b$. Note that by the sum of geometric series we get for every $a \geq 1/\gamma$

$$\begin{aligned} h_n(a, f_0) &\leq \frac{M}{\log^2(n/a)} \Big(\frac{1}{a} \sum_{i=1}^{I_a} i e^{i/a} e^{-2\gamma i} + \frac{n^2}{a^3} \sum_{i > I_a} i e^{-i/a} e^{-2\gamma i} \Big) \\ &\leq \frac{M}{a \log^2(n/a)} \sum_{i=1}^{\infty} i e^{-\gamma i} \leq \frac{M}{a(1 - e^{-\gamma})^2 \log^2(n/a)}, \end{aligned}$$

which is bounded from above by arbitrarily small b for a sufficiently large choice of the constant $C_{\gamma,b,M}$.

Finally, we deal with the lower bound for $\underline{a}_n$. Note that by using the inequalities $ae^{i/a} + n \geq n$ and $ae^{i/a} + n \geq ae^{i/a}$, for all $a \geq 1$, and the sum of geometric series, we get

$$\begin{aligned} g_n(a, f_0) &\geq \frac{m}{4 \log^2(n/a)} \frac{n^2}{a^3} \sum_{i > I_a} (i - a) e^{-i/a} i^{-2\beta} \\ &\geq c_\beta \frac{m}{\log^2(n/a)} \frac{n^2}{a^2} I_a^{-2\beta} e^{-I_a/a} \\ &= c_\beta m a^{-1-2\beta} n \big(\log(n/a) \big)^{-2-2\beta} \end{aligned}$$

for some $c_\beta > 0$ depending only on β. For $1 \leq a \leq Kn^{1/(1+2\beta)}(\log n)^{-1-2/(1+2\beta)}$ the preceding display is bounded by a multiple of $mc_\beta K^{-1-2\beta} \log n$. Then for a sufficiently small choice of the constant $C_{\beta,B,m}$, we get $g_n(a, f_0) \geq B \log n$ for any $a \leq C_{\beta,B,m} n^{1/(1+2\beta)}(\log n)^{-1-2/(1+2\beta)}$. ∎

In the next lemma we show that under the polished tail condition the deterministic bounds $\underline{a}_n, \overline{a}_n$ are close to each other.

Lemma A.3. *For every $L_0, \rho, N_0 \geq 1$ we have*

$$\sup_{f_0 \in \Theta_{pt}(L_0, N_0, \rho)} \frac{\overline{a}_n \log(n/\overline{a}_n)}{\underline{a}_n \log(n/\underline{a}_n)} \leq K \log^2 n,$$

with $K = 8.1e^4 \rho^2 L_0 B/b$ for n large enough.

Proof. First, note that since $\underline{a}_n \leq \overline{a}_n$, there is nothing to prove in the trivial case $\underline{a}_n = A_n$ or $\overline{a}_n = K_0$. Hence $h_n(\overline{a}_n, f_0) \leq b$ and $g_n(a, f_0) < B \log n$, for all $a > \underline{a}_n$, hold. Further assume that $\underline{a}_n \leq \rho^{-2} \overline{a}_n$; otherwise, the statement is trivial.

Let us divide the interval $[\rho^j, \rho^{j+1})$ into subintervals $[\rho^{j+k/\lceil \log n \rceil}, \rho^{j+(k+1)/\lceil \log n \rceil})$, $k = 0, 1, \ldots, \lceil \log n \rceil - 1$ and introduce the notation

$$k_j = \operatorname{argmax}_{k=0,\ldots,\lceil \log n \rceil - 1} \vartheta_{j,k}, \quad \text{where } \vartheta_{j,k} = \sum_{i=\rho^{j+k/\lceil \log n \rceil}}^{\rho^{j+(k+1)/\lceil \log n \rceil}} f_{0,i}^2,$$

with the notational convenience $\sum_{i=a}^{b} c_i = \sum_{i=\lceil a \rceil}^{\lfloor b \rfloor} c_i$, applied later on as well.

Then by the polished tail condition,

$$\sum_{i=\rho^j}^{\infty} f_{0,i}^2 \le L_0 \sum_{i=\rho^j}^{\rho^{j+1}} f_{0,i}^2 \le L_0 \log(n) \vartheta_{j,k_j}$$

for $j \ge \log_\rho N_0$. Note that for every $a > 0$ there exists an $\tilde{a} \in (a, \rho^2 a)$ such that

$$I_{\tilde{a}} \equiv \tilde{a} \log(n/\tilde{a}) \in \big[\rho^{j+k_j/\lceil \log n \rceil}, \rho^{j+(k_j+1)/\lceil \log n \rceil}\big) \tag{A.5}$$

for some $j \in \mathbb{N}$, and let us denote this j by $J_{\tilde{a}}$. Then

$$\sum_{i=e^{-1/\log n} I_{\tilde{a}}}^{e^{1/\log n} I_{\tilde{a}}} f_{0,i}^2 \ge \vartheta_{J_{\tilde{a}}, k_{J_{\tilde{a}}}}.$$

Let us take any $a_1 \le \rho^{-2} a_2$ and denote by $\tilde{a}_1 \in (a_1, \rho^2 a_1)$ the value satisfying (A.5). Then in view of $\exp\{e^{1/\log n} \log(n/a)\} \le \exp\{(1 + 2/\log n)\log(n/a)\} \le e^2 n/a$ for $n \ge e$ combined with the previous inequalities, we get that

$$\begin{aligned}
\frac{h_n(a_2, f_0)}{h_n(\tilde{a}_1, f_0)} &\le \frac{\tilde{a}_1 \log^2(\frac{n}{\tilde{a}_1})}{a_2 \log^2(\frac{n}{a_2})} 4e^2 \frac{\sum_{i=1}^{I_{\tilde{a}_1}} i e^{i/a_2} f_{0,i}^2 + \sum_{i=I_{\tilde{a}_1}}^{I_{a_2}} i e^{i/a_2} f_{0,i}^2 + \frac{n^2}{a_2^2} \sum_{i=I_{a_2}}^{\infty} i e^{-i/a_2} f_{0,i}^2}{\sum_{i=1}^{e^{1/\log n} I_{\tilde{a}_1}} i e^{i/\tilde{a}_1} f_{0,i}^2} \\
&\le \frac{\tilde{a}_1 \log^2(\frac{n}{\tilde{a}_1})}{a_2 \log^2(\frac{n}{a_2})} 4e^2 \Big(1 + \frac{\sum_{i=I_{\tilde{a}_1}}^{I_{a_2}} i e^{i/a_2} f_{0,i}^2 + n \log(\frac{n}{a_2}) \sum_{i=I_{a_2}}^{\infty} f_{0,i}^2}{\sum_{i=e^{-1/\log n} I_{\tilde{a}_1}}^{e^{1/\log n} I_{\tilde{a}_1}} i e^{i/\tilde{a}_1} f_{0,i}^2}\Big).
\end{aligned}$$

Since $i e^{i/\tilde{a}_1} > e^{-2} n \log(n/\tilde{a}_1)$ for $i \ge e^{-1/\log n} I_{\tilde{a}_1}$, and $i e^{i/a_2} \le n \log(n/a_2)$ for $i \le I_{a_2}$, we can see that

$$\frac{\sum_{i=I_{\tilde{a}_1}}^{I_{a_2}} i e^{i/a_2} f_{0,i}^2 + n \log(\frac{n}{a_2}) \sum_{i=I_{a_2}}^{\infty} f_{0,i}^2}{\sum_{i=e^{-1/\log n} I_{\tilde{a}_1}}^{e^{1/\log n} I_{\tilde{a}_1}} i e^{i/\tilde{a}_1} f_{0,i}^2} \le e^2 \frac{\log(\frac{n}{a_2})}{\log(\frac{n}{\tilde{a}_1})} \frac{\sum_{i=I_{\tilde{a}_1}}^{\infty} f_{0,i}^2}{\sum_{i=e^{-1/\log n} I_{\tilde{a}_1}}^{e^{1/\log n} I_{\tilde{a}_1}} f_{0,i}^2}.$$

Moreover, since

$$\sum_{i=I_{\tilde{a}_1}}^{\infty} f_{0,i}^2 \le L_0 \log(n) \vartheta_{J_{\tilde{a}_1}, k_{J_{\tilde{a}_1}}} \le L_0 \log(n) \sum_{i=e^{-1/\log n} I_{\tilde{a}_1}}^{e^{1/\log n} I_{\tilde{a}_1}} f_{0,i}^2,$$

when we combine this with the preceding computations we get that

$$\frac{h_n(a_2, f_0)}{h_n(\tilde{a}_1, f_0)} \le 4e \frac{\tilde{a}_1 \log^2(\frac{n}{\tilde{a}_1})}{a_2 \log^2(\frac{n}{a_2})} \Big(1 + L_0 e^2 \log(n) \frac{\log(\frac{n}{a_2})}{\log(\frac{n}{\tilde{a}_1})}\Big). \tag{A.6}$$

Furthermore, let us note that for any $\underline{a}_n < a \leq A_n$,

$$h_n(a, f_0) \leq 2g_n(a, f_0) + \frac{2e^2}{\log^2(\frac{n}{a})} \sum_{i=1}^{2a} f_{0,i}^2 \leq 2B \log(n) + o(1).$$

Then by taking $a_1 = \underline{a}_n$, $\tilde{a}_1 \in (\underline{a}_n, \rho^2 \underline{a}_n)$, and $a_2 = \overline{a}_n$ in (A.6) we get that

$$\begin{aligned} \frac{b}{2B \log(n) + o(1)} \leq \frac{h_n(\overline{a}_n, f_0)}{h_n(\tilde{a}_1, f_0)} &\leq 4e^4(1 + o(1))L_0 \log(n) \frac{\tilde{a}_1 \log(\frac{n}{\tilde{a}_1})}{\overline{a}_n \log(\frac{n}{\overline{a}_n})} \\ &\leq 4e^4 \rho^2 (1 + o(1))L_0 \log(n) \frac{\underline{a}_n \log(\frac{n}{\underline{a}_n})}{\overline{a}_n \log(\frac{n}{\overline{a}_n})}. \end{aligned}$$

After rearranging the preceding inequality we arrive to our statement. ∎

A.2. Contraction rates. In this section we provide the contraction rate results for both the empirical and hierarchical Bayes procedures. First, we show that the empirical Bayes method achieves the (up to a logarithmic factor) optimal minimax contraction rate around the truth for unknown regularity hyperparameter $\beta > 0$.

Theorem A.4. *The maximum marginal likelihood empirical Bayes posterior corresponding to the prior (2.2) achieves the minimax adaptive contraction rate (up to a logarithmic factor), i.e., for given $M, \beta_0 > 0$ we have*

$$\sup_{\beta \geq \beta_0} \sup_{f \in \Theta^\beta(M)} E_0[\Pi_{\hat{a}_n}(\|f - f_0\|_2 \geq M_n(n/\log^2 n)^{-\beta/(1+2\beta)} | Y)] \to 0 \tag{A.7}$$

for any sequence M_n tending to infinity.

Proof. See section A.3. ∎

Using our findings on the empirical Bayes method we can extend the results on the hierarchical Bayes method derived in the literature [36, 4] (where, typically, an inverse gamma hyperprior was considered) by allowing other, more general choices of the hyperprior distribution as well.

Theorem A.5. *Let us assume that the hyperprior π satisfies Assumption 1. Then the corresponding hierarchical Bayes posterior achieves the minimax contraction rate (up to a logarithmic factor), i.e., for given $\beta_0, M > 0$ we have*

$$\sup_{\beta \geq \beta_0} \sup_{f \in \Theta^\beta(M)} E_0[\Pi(\|f - f_0\|_2 \geq M_n(n/\log^2 n)^{-\beta/(1+2\beta)} | Y)] \to 0 \tag{A.8}$$

for some arbitrary sequence M_n tending to infinity.

Proof. See section G.1. ∎

A.3. Proof of Theorem A.4. Let us introduce the shorthand notation

$$\varepsilon_n := n^{-\beta/(1+2\beta)} (\log n)^{2\beta/(1+2\beta)}.$$

In view of Markov's inequality and Theorem A.1, for every $\beta > 0$,

$$\sup_{f_0\in\Theta^\beta(M)} E_0[\Pi_{\hat{a}_n}(\|f-f_0\|_2 \geq M_n\varepsilon_n|Y)] \leq \frac{1}{M_n^2\varepsilon_n^2} \sup_{f_0\in\Theta^\beta(M)} E_0\Big[\sup_{a\in[\underline{a}_n,\overline{a}_n]} R_n(a)\Big] + o(1), \tag{A.9}$$

where

$$R_n(a) = \int \|f-f_0\|_2^2 \Pi_a(df|Y)$$

is the posterior risk. We show below that both

$$\sup_{f_0\in\Theta^\beta(M)} \sup_{a\in[\underline{a}_n,\overline{a}_n]} E_0[R_n(a)] = O(\varepsilon_n^2) \qquad \text{and} \tag{A.10}$$

$$\sup_{f_0\in\Theta^\beta(M)} E_0[\sup_{a\in[\underline{a}_n,\overline{a}_n]} |R_n(a) - E_0(R_n(a))|] = o(\varepsilon_n^2) \tag{A.11}$$

hold, which results in the right-hand side of (A.9) vanishing as $n\to\infty$, which concludes the proof of Theorem A.4.

A.3.1. Bound for the expected posterior risk (A.10). First, note that by elementary computations,

$$R_n(a) = \sum_{i=1}^\infty (\hat{f}_{a,i} - f_{0,i})^2 + \sum_{i=1}^\infty \frac{1}{ae^{i/a}+n},$$

where $\hat{f}_{a,i} = n(ae^{i/a}+n)^{-1}Y_i$ is the ith coefficient of the posterior mean. Therefore the expectation of $R_n(a)$ is given by

$$E_0R_n(a) = \sum_{i=1}^\infty \frac{a^2e^{2i/a}}{(ae^{i/a}+n)^2} f_{0,i}^2 + \sum_{i=1}^\infty \frac{n}{(ae^{i/a}+n)^2} + \sum_{i=1}^\infty \frac{1}{ae^{i/a}+n}. \tag{A.12}$$

Note that the second and third terms do not contain f_0 and that the second term is bounded by the third. By Lemma H.2 (with $r=0$ and $l=1$) and Proposition A.2 the latter is further bounded for $a\leq\overline{a}_n$ by a multiple of

$$\frac{a}{n}\log\big(\frac{n}{a}\big) \leq \frac{\overline{a}_n}{n}\log\big(\frac{n}{\overline{a}_n}\big) \leq C_{\beta,b,M} n^{-2\beta/(1+2\beta)}(\log n)^{-1/(1+2\beta)},$$

since the function $a\mapsto a\log(n/a)$ is monotone increasing for $a\leq n/e$. It remains to deal with the first term on the right-hand side of (A.12), which we divide into three parts, and we show that each part has the stated order. First, note that for $f_0\in\Theta^\beta(M)$,

$$\sum_{i=(n/\log^2 n)^{1/(1+2\beta)}}^\infty \frac{a^2e^{2i/a}}{(ae^{i/a}+n)^2} f_{0,i}^2 \leq \sum_{i=(n/\log^2 n)^{1/(1+2\beta)}}^\infty Mi^{-1-2\beta} \leq \frac{M}{2\beta}(n/\log^2 n)^{-2\beta/(1+2\beta)}.$$

Next, note that for $a\leq\overline{a}_n$, in view of Proposition A.2,

$$\begin{aligned}\sum_{i=1}^{2a} \frac{a^2e^{2i/a}}{(ae^{i/a}+n)^2} f_{0,i}^2 &\leq \sum_{i=1}^{2a} \frac{a^2e^{2i/a}}{n^2} f_{0,i}^2 \leq \frac{a^2e^4}{n^2}\sum_{i=1}^{2a} f_{0,i}^2 \\ &\leq e^4\frac{\overline{a}_n^2}{n^2} \leq e^4 M C_{\beta,b,M}^2 n^{-\frac{4\beta}{1+2\beta}}(\log n)^{-2-\frac{2}{1+2\beta}}.\end{aligned}$$

Furthermore, notice that the maximum of the function $i \mapsto e^{i/a}/(i-a)$ over $[2a, I_a]$ is attained at $i = I_a$, because the function is increasing for $i > 2a$ and $n > 0$. In addition, for $a > \underline{a}_n$ we have $g_n(a, f_0) < B \log n$, and hence for any $\underline{a}_n < a \le \bar{a}_n$,

$$\begin{aligned}\sum_{i=2a}^{I_a} \frac{a^2 e^{2i/a}}{(ae^{i/a}+n)^2} f_{0,i}^2 &\le \frac{a}{n} \frac{\log^2(n/a)}{(\log(n/a)-1)} \sum_{i=2a}^{I_a} \frac{n^2 e^{i/a}(i-a)}{a \log^2(n/a)(ae^{i/a}+n)^2} f_{0,i}^2 \\ &\le 2\bar{a}_n n^{-1} \log(n/\bar{a}_n) g_n(a, f_0) \\ &\le 2\bar{a}_n n^{-1} \log^2 n \le 2C_{\beta,b,M} n^{-2\beta/(1+2\beta)} (\log n)^{2\beta/(1+2\beta)},\end{aligned}$$

where the last inequality follows from Proposition A.2.

It remains to deal with the terms between $I_{\underline{a}_n} = \underline{a}_n \log(n/\underline{a}_n)$ and $(n/\log^2 n)^{1/(1+2\beta)}$. Let $J = J(n)$ be the smallest integer such that

$$\big(1 + \frac{1}{\log n}\big)^J \underline{a}_n \log\big(\frac{n}{\underline{a}_n}\big) \ge (n/\log^2 n)^{1/(1+2\beta)},$$

and let

$$n_j := \big(1 + \frac{1}{\log n}\big)^j I_{\underline{a}_n}.$$

Note that the sequence n_j is increasing. For notational convenience, we also introduce b_j such that $b_j e^{n_j/b_j} = n$ and $b_j < n_j$. Now we have for any $a \ge 1$,

$$\begin{aligned}\sum_{i=I_{\underline{a}_n}}^{(n/\log^2 n)^{1/(1+2\beta)}} \frac{a^2 e^{2i/a}}{(ae^{i/a}+n)^2} f_{0,i}^2 &\le \sum_{j=0}^{J-1} \sum_{i=n_j}^{n_{j+1}} f_{0,i}^2 \\ &\le 4e^2 \sum_{j=0}^{J-1} \sum_{i=n_j}^{n_{j+1}} \frac{n b_j e^{i/b_j}}{(b_j e^{i/b_j}+n)^2} f_{0,i}^2. \qquad \text{(A.13)}\end{aligned}$$

By elementary computations we get that $b_j \asymp n_j/\log n_j$; therefore (A.13) is further bounded by constant times

$$\frac{1}{n} \sum_{j=0}^{J-1} \frac{1}{\log n_j} \sum_{i=n_j}^{n_{j+1}} \frac{n^2 (i-b_j) e^{i/b_j}}{(b_j e^{i/b_j}+n)^2} f_{0,i}^2 \le \frac{1}{n} \sum_{j=0}^{J-1} \frac{b_j \log^2 n}{\log n_j} g_n(b_j, f_0).$$

Since $b_j \ge \underline{a}_n$ we have $g_n(b_j, f_0) \le B \log n$ for all $j \ge 0$. Then by the sum of geometric series we get that

$$\begin{aligned}\frac{1}{n} \sum_{j=0}^{J-1} \frac{n_j}{\log^2 n_j} \log^3 n &\le 2(1+2\beta)^2 \frac{\log n}{n} \frac{I_{\underline{a}_n}(1+1/\log n)^J}{1/\log n} \\ &\le 2(1+2\beta)^2 n^{-2\beta/(1+2\beta)} (\log n)^{2-2/(1+2\beta)},\end{aligned}$$

which concludes the proof of assertion (A.10).

A.3.2. Bound for the centered posterior risk (A.11). Note that

$$R_n(a) - E_0 R_n(a) = \mathbb{V}(a)/n - 2\mathbb{W}(a)/\sqrt{n}, \quad \text{where}$$
$$\mathbb{V}(a) = n^2 \sum_{i=1}^{\infty} \frac{1}{(ae^{i/a}+n)^2}(Z_i^2 - 1) \quad \text{and} \quad \mathbb{W}(a) = n \sum_{i=1}^{\infty} \frac{ae^{i/a} f_{0,i}}{(ae^{i/a}+n)^2} Z_i.$$

Therefore it is sufficient to show that there exists a constant $K = K_{\beta,M,b,B} > 0$ such that

$$E_0\left(\sup_{a \in [\underline{a}_n, \overline{a}_n]} |\mathbb{V}(a)|/n \right) \le K n^{-2\beta/(1+2\beta)} (\log n)^{-1/(1+2\beta)},$$
$$E_0\left(\sup_{a \in [\underline{a}_n, \overline{a}_n]} |\mathbb{W}(a)/\sqrt{n}| \right) \le K n^{-2\beta/(1+2\beta)}.$$

We deal with the two processes above separately.

For the process $\mathbb{V}$, Corollary 2.2.5 in [38] implies that

$$E_0\left[\sup_{a \in [\underline{a}_n, \overline{a}_n]} |\mathbb{V}(a)| \right] \lesssim \sup_{a \in [\underline{a}_n, \overline{a}_n]} \sqrt{V_0(\mathbb{V}(a))} + \int_0^{diam_n} \sqrt{N(\varepsilon, [\underline{a}_n, \overline{a}_n], d_n)} d\varepsilon,$$

where $d_n^2(a_1, a_2) = V_0(\mathbb{V}(a_1) - \mathbb{V}(a_2))$, $diam_n$ is the d_n-diameter of $[\underline{a}_n, \overline{a}_n]$, $V_0(\mathbb{V}(a))$ is the variance of $\mathbb{V}(a)$ with respect to the distribution P_0, and $N(\varepsilon, B, d_n)$ is the covering number of the set B with ε-radius balls relative to the d_n semimetric. The variance of $\mathbb{V}(a)$ is equal to

$$V_0(\mathbb{V}(a)) = 2n^4 \sum_{i=1}^{\infty} \frac{1}{(ae^{i/a}+n)^4}$$

since $V(Z_i^2) = 2$. Using Lemma H.2 (with $r = 0$ and $l = 4$) we can conclude that the variance of $\mathbb{V}(a)$ is bounded from above by a multiple of $a \log(n/a)$, and hence $diam_n \lesssim \sqrt{\overline{a}_n \log n}$. In view of Lemma A.6, the distance $d_n(a_1, a_2)$ is bounded from above by a multiple of $|a_1 - a_2| \log^{3/2} n$, and hence the interval $[\underline{a}_n, \overline{a}_n]$ can be covered with a constant times $\overline{a}_n \varepsilon^{-1} \log^{3/2} n$ amount of ε-balls relative to the d_n semimetric. In view of the above computation and Proposition A.2,

$$E_0\left[n^{-1} \sup_{a \in [\underline{a}_n, \overline{a}_n]} |\mathbb{V}(a)| \right] \lesssim (\overline{a}_n/n) \log n \le C_{\beta,b,M} n^{-2\beta/(1+2\beta)} (\log n)^{-1/(1+2\beta)}.$$

The process $\mathbb{W}$ can be dealt with similarly to $\mathbb{V}$. The main difference is the bounding of the variance of $\mathbb{W}$, which we describe in detail. First, note that

$$V_0\left(\frac{\mathbb{W}(a)}{\sqrt{n}} \right) = n \sum_{i=1}^{\infty} \frac{a^2 e^{2i/a}}{(ae^{i/a}+n)^4} f_{0,i}^2.$$

Let us split the sum at I_a, and by applying the inequality $ae^{i/a} + n \ge n$, we get

$$n \sum_{i=1}^{I_a} \frac{a^2 e^{2i/a}}{(ae^{i/a}+n)^4} f_{0,i}^2 \le \frac{1}{n^3} \sum_{i=1}^{I_a} a^2 e^{2i/a} f_{0,i}^2 \le \frac{\|f_0\|_2^2}{n}.$$

Then by noting that the function $i \mapsto e^{i/a}/((i-a)(ae^{i/a}+n)^2)$ is decreasing on $[I_a, \infty)$, recalling that $g_n(a, f_0) \le B \log n$, for all $a \ge \underline{a}_n$, and in view of Proposition A.2, for $a \le \overline{a}_N$,

$$\begin{aligned} n \sum_{i=I_a}^{\infty} \frac{a^2 e^{2i/a}}{(ae^{i/a}+n)^4} f_{0,i}^2 &\le \frac{a \log^2(n/a)}{4n^2(\log(n/a)-1)} \sum_{i=I_a}^{\infty} \frac{n^2(i-a)e^{i/a}}{a \log^2(n/a)(ae^{i/a}+n)^2} f_{0,i}^2 \\ &\le an^{-2} \log(n/a) g_n(a, f_0) \le B\overline{a}_n n^{-2} \log^2 n \\ &\le 2BC_{\beta,b,M} n^{-(1+4\beta)/(1+2\beta)} (\log n)^{2\beta/(1+2\beta)}, \end{aligned}$$

and hence $\mathrm{diam}_n = O(n^{-\frac{1/2+2\beta}{1+2\beta}} (\log n)^{\frac{\beta}{1+2\beta}})$. Then in view of Lemma A.6 the covering number of the interval $[\underline{a}_n, \overline{a}_n]$ is bounded by $C_M \varepsilon^{-1} (\overline{a}_n/\sqrt{n}) \log n$ with respect to the semimetric $d_n(a_1, a_2) = V_0\big(\mathbb{W}(a_1)/\sqrt{n} - \mathbb{W}(a_2)/\sqrt{n}\big)$, and the rest of the proof follows as above.

A.3.3. Bounds for the semimetrics associated to $\mathbb{V}$ and $\mathbb{W}$.

Lemma A.6. *For any $1 \le a_1 \le a_2$ and $f_0 \in \ell_2(M)$ we have*

$$\begin{aligned} V_0\Big(\mathbb{V}(a_1) - \mathbb{V}(a_2)\Big) &\lesssim (a_1 - a_2)^2 \log^3 n, \\ V_0\Big(\mathbb{W}(a_1) - \mathbb{W}(a_2)\Big) &\lesssim (a_1 - a_2)^2 \log^2 n, \end{aligned}$$

with constants only depending on M.

Proof. The left-hand side of the first inequality is equal to

$$n^4 \sum_{i=1}^{\infty} (\phi_i(a_1) - \phi_i(a_2))^2 V(Z_i^2),$$

where $\phi_i(a) = (ae^{i/a}+n)^{-2}$. The square of the derivative of ϕ_i is given by $\phi_i'(a)^2 = 4\phi_i(a)^3 e^{2i/a}(i-a)^2/a^2$, and hence in view of Lemma H.3 the preceding display is bounded above by a multiple of

$$\begin{aligned} (a_1 - a_2)^2 n^4 \sup_{a \in [a_1, a_2]} \sum_{i=1}^{\infty} \frac{e^{2i/a}(i-a)^2}{a^2(ae^{i/a}+n)^6} &\le (a_1 - a_2)^2 n^4 \sup_{a \in [a_1, a_2]} \sum_{i=1}^{\infty} \frac{e^{2i/a}(i^2+a^2)}{a^2(ae^{i/a}+n)^6} \\ &\lesssim (a_1 - a_2)^2 \sup_{a \in [a_1, a_2]} \frac{\log(n/a)}{a} \Big(1 + \log^2\big(\frac{n}{a}\big)\Big) \end{aligned}$$

with the help of Lemma H.1 (first with $m = 2$ and then with $m = 0$) and Lemma H.2 (with $r = 1$ and $l = 4$).

We next consider the process $\mathbb{W}(a)$. The left-hand side of the second inequality in the statement of the lemma is equal to

$$n^2 \sum_{i=1}^{\infty} (\phi_i(a_1) - \phi_i(a_2))^2 f_{0,i}^2 V_0(Z_i),$$

with $\phi_i(a) = ae^{i/a}/(ae^{i/a}+n)^2$. Note that $|\phi_i'(a)| \le (i+a)a^{-2}\phi_i(a)$, and hence in view of Lemma H.1 (first with $m=2$ and then with $m=0$) and Lemma H.3 the preceding display is bounded by

$$4(a_1-a_2)^2n^2 \sup_{a\in[a_1,a_2]} \frac{1}{a^2}\sum_{i=1}^{\infty} \frac{e^{2i/a}(i^2/a^2+1)}{(ae^{i/a}+n)^4} f_{0,i}^2$$
$$\le 4(a_1-a_2)^2 \sup_{a\in[a_1,a_2]} \frac{1}{a^2}\big(\log^2(n/a)+1\big)\|f_0\|_2^2,$$

which concludes the proof of the lemma. ∎

Appendix B. Proof of the empirical Bayes part of Theorem 2.4. First, note that we get the empirical Bayes credible set by plugging the estimator $\hat{a}_n$ into the credible ball $\hat{C}_{a,\alpha}$, defined as

$$\hat{C}_{a,\alpha} = \{f \in L_2 : \|f - \hat{f}_a\|_2 \le Lr_\alpha\},$$

which satisfies that

$$\Pi_a(\hat{C}_{a,\alpha}|Y) = 1-\alpha,$$

where $\hat{f}_a$ is the posterior mean for fixed hyperparameter $a>0$.

The proof of the statement is then based on the deterministic bounds for the MMLE $\hat{a}_n$ derived in Theorem A.1, and their distance is investigated in Lemma A.3.

Note that $f_0 \in \hat{C}_n(L\log^{3/2} n)$ if and only if $\|f_0 - \hat{f}_{\hat{a}_n}\|_2 \le L(\log n)^{3/2} r_\alpha$. Therefore by the triangle inequality it is sufficient to verify that

$$\|W(\hat{a}_n)\|_2 \le L(\log n)^{3/2} r_\alpha(\hat{a}_n) - \|B(\hat{a}_n, f_0)\|_2 \tag{B.1}$$

holds with high probability, where $W(a) = \hat{f}_a - E_0\hat{f}_a$ and $B(a,f_0) = E_0\hat{f}_a - f_0$ are the centered posterior mean and the bias of the posterior mean for fixed hyperparameter $a>0$, respectively. Note that the ith coefficients of these vectors take the forms

$$W_i(a) = \frac{n(Y_i - f_{0,i})}{ae^{i/a}+n} \quad \text{and} \quad B_i(a,f_0) = \frac{ae^{i/a}f_{0,i}}{ae^{i/a}+n}.$$

We prove below that there exist constants $C_1, C_2 > 0$ depending on ρ, L_0, B, and b such that for large enough n,

$$\inf_{\underline{a}_n \le a \le \overline{a}_n} r_\alpha^2(a) \ge \frac{\underline{a}_n}{3n}\log\big(\frac{n}{\underline{a}_n}\big), \tag{B.2}$$

$$\inf_{f_0\in\Theta_{pt}(L_0,N_0,\rho)} P_0\Big(\sup_{\underline{a}_n\le a\le\overline{a}_n} \|W(a)\|_2^2 \le C_1\frac{\underline{a}_n}{n}\log\big(\frac{n}{\underline{a}_n}\big)\log^2 n\Big) \to 1, \tag{B.3}$$

$$\sup_{f_0\in\Theta_{pt}(L_0,N_0,\rho)} \sup_{\underline{a}_n\le a\le\overline{a}_n} \|B(a,f_0)\|_2^2 \le C_2\frac{\underline{a}_n}{n}\log\big(\frac{n}{\underline{a}_n}\big)\log^3 n. \tag{B.4}$$

Hence in view of Theorem A.1 assertion (B.1) holds with probability tending to one for a large enough choice of L under the polished tail assumption.

Proof of (B.2). The radius $r_\alpha(a)$, given in (2.6), is defined as $P(U_n(a) < r_\alpha^2(a)) = 1-\alpha$ with $U_n(a) := \sum_{i=1}^{\infty} \frac{1}{ae^{i/a}+n} Z_i^2$, where the Z_i's are i.i.d. $N(0,1)$. We show below that

$$\liminf_{n\to\infty} \inf_{a\in[\underline{a}_n,\overline{a}_n]} E\Big[\frac{nU_n(a)}{a\log(n/a)}\Big] > \frac{1}{2}, \tag{B.5}$$

$$E\left[\sup_{a\in[\underline{a}_n,\overline{a}_n]} \frac{n|U_n(a)-E[U_n(a)]|}{a\log(n/a)}\right] \to 0. \tag{B.6}$$

Then by Markov's inequality with probability tending to one we have

$$\inf_{a\in[\underline{a}_n,\overline{a}_n]} \frac{nU_n(a)}{a\log(n/a)} > 1/3,$$

and hence (B.2) follows from the definition of $r_\alpha(a)$.

Assertion (B.5) follows as

$$E[U_n(a)] \geq \sum_{i=1}^{I_a} \frac{1}{ae^{i/a}+n} \geq \frac{I_a}{2n} \geq \frac{a}{2n}\log\big(\frac{n}{a}\big).$$

To verify (B.6), it suffices by Corollary 2.2.5 in [38] (applied with $\psi(x) = x^2$) to show that there exist $K_1, K_2 > 0$ such that for any $a \in [\underline{a}_n, \overline{a}_n]$,

$$V\Big(\frac{nU_n(a)}{a\log(n/a)}\Big) \leq K_1 \frac{1}{a\log(n/a)}, \tag{B.7}$$

$$\int_0^{diam_n} \sqrt{N(\varepsilon,[\underline{a}_n,\overline{a}_n],d_n)}d\varepsilon \leq \sqrt{A_n/n} = o(1), \tag{B.8}$$

where d_n is the semimetric defined by $d_n^2(a_1,a_2) := V(\frac{nU_n(a_1)}{a_1\log(n/a_1)} - \frac{nU_n(a_2)}{a_2\log(n/a_2)})$, $diam_n$ is the diameter of the interval $[\underline{a}_n, \overline{a}_n]$ relative to d_n, and $N(\varepsilon, S, d_n)$ is the minimal number of d_n-balls of radius ε needed to cover the set S.

First, note that in view of Lemma H.2 (with $r = 0$ and $l = 2$) we have

$$V\Big(\frac{nU_n(a)}{a\log(n/a)}\Big) = \frac{2n^2}{a^2\log^2(n/a)} \sum_{i=1}^{\infty} \frac{1}{(ae^{i/a}+n)^2} \lesssim \frac{1}{a\log(n/a)}.$$

As a consequence one can see that $diam_n \lesssim (\underline{a}_n \log(n/\underline{a}_n))^{-1/2}$. By Lemma B.1, $d_n(a_1,a_2) \lesssim a_1^{-3/2}\log^{1/2}(n/a_1)n^{-1}|a_1-a_2|$, and hence

$$N(\varepsilon,[\underline{a}_n,\overline{a}_n],d_n) \lesssim \varepsilon^{-1}\log^{1/2}(n/\underline{a}_n)\underline{a}_n^{-3/2}\overline{a}_n/n.$$

Therefore one can conclude that

$$\int_0^{diam_n} \sqrt{N(\varepsilon,[\underline{a}_n,\overline{a}_n],d_n)}d\varepsilon = \frac{\overline{a}_n^{1/2}\log^{\frac{1}{4}}(n/\underline{a}_n)}{\underline{a}_n^{3/4}n^{1/2}} \int_0^{C(\underline{a}_n\log(\frac{n}{\underline{a}_n}))^{-\frac{1}{2}}} \varepsilon^{-\frac{1}{2}}d\varepsilon \lesssim \sqrt{A_n/n}.$$

Proof of (B.3). The variable $\|W(a)\|_2^2$ is distributed as $\sum_{i=1}^{\infty} \frac{n}{(ae^{i/a}+n)^2} Z_i^2$, with $Z_i \overset{iid}{\sim} N(0,1)$. Observe that

$$E_0[\|W(a)\|_2^2] = \sum_{i=1}^{\infty} \frac{n}{(ae^{i/a}+n)^2} \quad \text{and} \quad V_0(\|W(a)\|_2^2) = 2\sum_{i=1}^{\infty} \frac{n^2}{(ae^{i/a}+n)^4}.$$

Further note that by applying Lemma H.2 (with $r=0$ and $l=2$) we get

$$\frac{a}{n}\log\big(\frac{n}{a}\big) \le \frac{4I_a n}{(ae^{I_a/a}+n)^2} \le \sum_{i=1}^{I_a} \frac{4n}{(ae^{i/a}+n)^2} \le \sum_{i=1}^{\infty} \frac{4n}{(ae^{i/a}+n)^2} \le C\frac{a}{n}\log\big(\frac{n}{a}\big)$$

for some universal constant $C>0$, while by applying the same lemma (with $r=0$ and $l=4$), we see that the variance is bounded above by a multiple of $an^{-2}\log(n/a)$. Then reasoning similar to the previous proof results in

$$\inf_{f_0\in\ell_2(M)} \Big(\sup_{\underline{a}_n \le a \le \overline{a}_n} \|W(a)\|_2^2 \le C_2(\overline{a}_n/n)\log(n/\overline{a}_n) \Big) \overset{P_0}{\to} 1. \tag{B.9}$$

Then in view of Lemma A.3, the right-hand side of the inequality in the preceding probability statement is further bounded from above by constant times $(\underline{a}_n/n)\log(n/\underline{a}_n)\log^2 n$.

Proof of (B.4). First, note that

$$\|B(a,f_0)\|_2^2 \le \sum_{i=1}^{I_a} n^{-2}a^2e^{2i/a}f_{0,i}^2 + \sum_{i=I_a}^{\infty} f_{0,i}^2.$$

To bound the first term on the right-hand side, we use the inequalities $a/n \le \log(n/a)$ for $a \le A_n$ and $\sum_{i=1}^{\infty} f_{0,i}^2 < \infty$, and we further note that the function $i \mapsto e^{i/a}/(i-a)$ is monotone increasing on the interval $[2a, I_a]$ and hence takes its maximum at I_a. Therefore in view of Lemma A.3 the first part of the bias for functions satisfying the polished tail condition is bounded by

$$\begin{aligned}
\sup_{\underline{a}_n \le a \le \overline{a}_n} \sum_{i=1}^{I_a} \frac{a^2e^{2i/a}f_{0,i}^2}{n^2} &\le \sup_{\underline{a}_n \le a \le \overline{a}_n} \sum_{i=1}^{2a} \frac{a^2e^{2i/a}f_{0,i}^2}{n^2} + \sup_{\underline{a}_n \le a \le \overline{a}_n} \frac{a}{n}\frac{\log^2(\frac{n}{a})}{(\log(\frac{n}{a})-1)} g_n(a,f_0) \\
&\le \frac{e^4\overline{a}_n^2}{n^2}\sum_{i=1}^{2a} f_{0,i}^2 + (B+o(1))\frac{\overline{a}_n}{n}\log(n)\log(\frac{n}{\overline{a}_n}) \\
&\le (B+o(1))\frac{\overline{a}_n}{n}\log(n)\log(\frac{n}{\overline{a}_n}) \\
&\le K_{\rho,L_0,B,b}\frac{\underline{a}_n}{n}\log\big(\frac{n}{\underline{a}_n}\big)\log^3 n
\end{aligned}$$

for some constant $K_{\rho,L_0,B,b}$ depending on ρ, L_0, B, and b. Furthermore, in view of the polished tail assumption we have

$$\sum_{i=I_{\underline{a}_n}}^{\infty} f_{0,i}^2 \le L_0 \sum_{i=I_{\underline{a}_n}}^{\rho I_{\underline{a}_n}} f_{0,i}^2 \le \log(\frac{n}{\underline{a}_n}) \sum_{i=I_{\tilde{a}_n}}^{I_{\tilde{a}_n}+\rho\tilde{a}_n} f_{0,i}^2$$

for some $\tilde{a}_n \in [\underline{a}_n, \rho\underline{a}_n]$. Therefore, by using Lemma A.3, we have

$$\sum_{i=I_{\underline{a}_n}}^{\infty} f_{0,i}^2 \lesssim \log(\frac{n}{\underline{a}_n}) \sum_{i=I_{\tilde{a}_n}}^{I_{\tilde{a}_n}+\rho\tilde{a}_n} \frac{n^2(i-\tilde{a}_n)e^{i/\tilde{a}_n}}{\tilde{a}_n \log^2(\frac{n}{\tilde{a}_n})(\tilde{a}_n e^{i/\tilde{a}_n}+n)^2} f_{0,i}^2 \frac{\tilde{a}_n}{n} \log(\frac{n}{\tilde{a}_n}),$$
$$\leq \log(\frac{n}{\underline{a}_n}) g_n(\tilde{a}_n, f_0) \frac{\tilde{a}_n}{n} \log(\frac{n}{\tilde{a}_n}) \leq K_{\rho,L_0,B,b} \log^2(\frac{n}{\underline{a}_n}) \log(n) \frac{\underline{a}_n}{n}$$

for some large enough constant $K_{\rho,L_0,B,b} > 0$. Combining the two bounds, we see that (B.4) holds.

Lemma B.1. *There exists a $K > 0$ such that for any $1 < a_1 < a_2$.*

$$V\Big(\frac{U_n(a_1)}{a_1 \log(n/a_1)} - \frac{U_n(a_2)}{a_2 \log(n/a_2)}\Big) \leq K(a_1 - a_2)^2 \frac{\log(n/a_1)}{a_1^3 n^2}. \tag{B.10}$$

Proof. First, note that

$$V\Big(\frac{U_n(a_1)}{a_1 \log(n/a_1)} - \frac{U_n(a_2)}{a_2 \log(n/a_2)}\Big) = 2\sum_{i=1}^{\infty}(\phi_i(a_1) - \phi_i(a_2))^2 \tag{B.11}$$

with $\phi_i(a) := \frac{1}{a\log(n/a)(ae^{i/a}+n)}$. The derivative of $\phi_i(a)$ is given as $\phi_i'(a) = \phi_i(a)(\frac{2(i-a)e^{i/a}}{a(ae^{i/a}+n)} + \frac{1}{a\log(n/a)} - \frac{1}{a})$, so we can see that $|\phi_i'(a)| \lesssim (\frac{(i+a)e^{i/a}}{a(ae^{i/a}+n)} \vee \frac{1}{a})\phi_i(a)$. Thus in view of Lemma H.3 the right-hand side of (B.11) is bounded by a multiple of

$$(a_1 - a_2)^2 \sup_{a\in[a_1,a_2]} \sum_{i=1}^{\infty} \Big(\frac{(i^2+a^2)e^{2i/a}}{a^2(ae^{i/a}+n)^2} \vee \frac{1}{a^2}\Big)\phi_i(a)^2.$$

Then in view of Lemma H.1 (first with $m = 2$ and then with $m = 0$) and Lemma H.2 (first with $r = 1$ and $l = 2$ and then with $r = 0$ and $l = 2$), the preceding display is further bounded by the right-hand side of (B.10), finishing the proof of the statement. ∎

Appendix C. Proof of Theorem 2.3.

We use the notation introduced in Appendix B. First, recall that $f_0 \in \hat{C}_n(L)$ if and only if $\|f_0 - \hat{f}\|_2 \leq Lr_\alpha(\hat{a}_n)$. We show below that

$$\inf_{f_0\in A^\gamma(M)} P_0\Big(\sup_{\underline{a}_n\leq a\leq\overline{a}_n} \|W(a)\|_2^2 \leq C_1 \frac{\underline{a}_n}{n} \log\big(\frac{n}{\underline{a}_n}\big)\Big) \to 1, \tag{C.1}$$

$$\sup_{f_0\in A^\gamma(M)} \sup_{\underline{a}_n\leq a\leq\underline{a}_n} \|B(a, f_0)\|_2^2 \leq C_2 \frac{\underline{a}_n}{n} \log\big(\frac{n}{\underline{a}_n}\big) \tag{C.2}$$

for some constants $C_1, C_2 > 0$ depending only on M, which, together with (B.2) and Theorem A.1, results in the statement.

The proof of assertion (C.1) follows by combining (B.9) and the second inequality of Proposition A.2. Next, note that similarly to the proof of (B.4), we get that

$$\|B(a, f_0)\|_2^2 \leq \sum_{i=1}^{I_a} \frac{a^2 e^{2i/a} f_{0,i}^2}{n^2} + \sum_{i=I_a}^{\infty} f_{0,i}^2 \lesssim \frac{\overline{a}_n}{n} \log\big(\frac{n}{\overline{a}_n}\big) + \sum_{i=I_{\underline{a}_n}}^{\infty} f_{0,i}^2.$$

Furthermore,

$$\sum_{i=I_{\underline{a}_n}}^{\infty} f_{0,i}^2 = \sum_{i=I_{\underline{a}_n}}^{\infty} e^{-2i\gamma} e^{2i\gamma} f_{0,i}^2 \leq M e^{-2I_{\underline{a}_n}\gamma} = M\big(\frac{\underline{a}_n}{n}\big)^{2\underline{a}_n\gamma} \leq M\frac{\underline{a}_n}{n}\log\big(\frac{n}{\underline{a}_n}\big)$$

for $\gamma \geq 1/2$, finishing the proof of (C.2) and concluding the proof of the theorem.

Appendix D. Proof of Theorem 2.2 and the empirical Bayes part of Corollary 2.1. In this proof we again use the notation introduced in Appendix B.

First, note that $f_0 \in \hat{C}_n(L_n)$ implies that $\|B(\hat{a}_n, f_0)\|_2 \leq L_n r_\alpha(\hat{a}_n) + \|W(\hat{a}_n)\|_2$, which, combined with Theorem A.1, provides the upper bound

$$P_0(f_0 \in \hat{C}_n(L_n)) \leq P_0\Big(\inf_{a\leq \overline{a}_n} \|B(a, f_0)\|_2 \leq L_n \sup_{a\leq \overline{a}_n} r_\alpha(a) + \sup_{a\leq \overline{a}_n} \|W(a)\|_2\Big) + o(1). \tag{D.1}$$

The proof of assertion (B.2) also shows that there exists constants $C_1 > 0$ such that

$$\sup_{a\leq \overline{a}_n} r_\alpha^2(a) \leq C_1 \frac{\overline{a}_n}{n}\log(\frac{n}{\overline{a}_n}). \tag{D.2}$$

Then in view of assertion (B.9) and Proposition A.2, both the squared radius $r_\alpha(a)^2$ and the variance term $\|W(a)\|_2^2$ are bounded by$C_{\beta,b,M} n^{-2\beta/(1+2\beta)}(\log n)^{-1/(1+2\beta)}$ for some $C_{\beta,b,M} > 0$.

Since for $f_0 \in \Theta_s^\beta(m, M)$ we have $\|B(a, f_0)\|_2^2 = \sum_{i=1}^\infty \frac{a^2 e^{2i/a} f_{0,i}^2}{(ae^{i/a}+n)^2}$, the bias is bounded from below by

$$\|B(a, f_0)\|_2^2 \geq m\sum_{i=I_a}^{\infty} i^{-1-2\beta} > \frac{m}{2\beta} I_a^{-2\beta} \geq \frac{m}{2\beta} a^{-2\beta}\log^{-2\beta}(n/a).$$

As the function $a \mapsto a^{-2\beta}\log^{-2\beta}\big(n/a\big)$ is monotone decreasing for $a \leq A_n$, we see that $\inf_{a\leq \overline{a}_n} \|B(a, f_0)\|_2^2 \geq m/(2\beta)\overline{a}_n^{-2\beta}\log^{-2\beta}(n/\overline{a}_n)$. Hence in view of Proposition A.2, the bias is bounded from below by $c_{m,\beta,b,B,M} n^{-2\beta/(1+2\beta)}\log^{2\beta/(1+2\beta)} n$ for some $c_{m,\beta,b,B,M} > 0$. Thus, the above inequalities imply that for arbitrary $f_0 \in \Theta_s^\beta(m, M)$ the right-hand side of (D.1) is further bounded by

$$\sup_{f_0\in\ell_2(M)} P_0\big(n^{-\beta/(1+2\beta)}(\log n)^{\beta/(1+2\beta)} \leq L_n C n^{-\beta/(1+2\beta)}(\log n)^{-(1/2)/(1+2\beta)}\big) + o(1) = o(1)$$

for arbitrary $L_n = o(\sqrt{\log n})$ and C depending on m, β, b, B, and M, which concludes the proof of the theorem.

Appendix E. Proof of Theorem A.1. First, note that the derivative of the marginal likelihood function $\ell_n(a)$ is

$$\mathbb{M}_n(a) = \frac{1}{2}\Big(\sum_{i=1}^{\infty} \frac{n^2 Y_i^2 e^{i/a}(i-a)}{a(ae^{i/a}+n)^2} - \sum_{i=1}^{\infty} \frac{n(i-a)}{a^2(ae^{i/a}+n)}\Big), \tag{E.1}$$

with expected value

$$E_0[\mathbb{M}_n(a)] = \frac{1}{2}\Big(\sum_{i=1}^{\infty} \frac{n^2(i-a)e^{i/a}f_{0,i}^2}{a(ae^{i/a}+n)^2} - \sum_{i=1}^{\infty} \frac{n^2(i-a)}{a^2(ae^{i/a}+n)^2}\Big). \tag{E.2}$$

In the following subsections we show, with the help of the score function $\mathbb{M}_n(a)$, that the marginal likelihood function $\ell_n(a)$ with probability tending to one has its global maximum outside the set $[1, \underline{a}_n) \cup (\overline{a}_n, A_n]$.

E.1. $\mathbb{M}_n(a)$ on $[1, \underline{a}_n)$. In this subsection we derive that the process $\mathbb{M}_n(a)$ is bounded from below by $-C_B \log^2(n/a)$ on $[1, \underline{a}_n]$, for some $C_B > 0$, and is bigger than $e^{-5/2}B\log^3(n/\underline{a}_n)$ on the interval

$$\mathcal{I}_n \equiv \Big[\frac{\log(n/\underline{a}_n)}{1+\log(n/\underline{a}_n)}\underline{a}_n, \underline{a}_n\Big] \tag{E.3}$$

with probability going to one, where B is the parameter in the definition of $\underline{a}_n$. Hence with probability tending to one for every $a \in [1, \underline{a}_n]/\mathcal{I}_n$,

$$\begin{aligned}
\ell_n(\underline{a}_n) - \ell_n(a) &\geq \int_a^{\frac{\log(n/\underline{a}_n)}{1+\log(n/\underline{a}_n)}\underline{a}_n} \mathbb{M}_n(\tilde{a})d\tilde{a} + \int_{\mathcal{I}_n} \mathbb{M}_n(\tilde{a})d\tilde{a} \\
&\geq -(\underline{a}_n - a)C\log^2(n/\underline{a}_n) + \frac{\tilde{c}_0 B\underline{a}_n \log^3(n/\underline{a}_n)}{\log(n/\underline{a}_n)} \\
&\geq (B\tilde{c}_0/2)\underline{a}_n \log^2(n/\underline{a}_n)
\end{aligned}$$

for $B > 2\tilde{c}_0^{-1}C$. Therefore the global maximum of $\ell_n(a)$ lies outside the interval $[1, \underline{a}_n)$ with probability tending to one. It remains to show the stated lower bounds for $\mathbb{M}_n(a)$.

By leaving out the nonnegative stochastic part, we get the lower bound

$$\mathbb{M}_n(a) \geq \frac{1}{2}\Big(\sum_{i=1}^{a} \frac{n^2(i-a)e^{i/a}Y_i^2}{a(ae^{i/a}+n)^2} - \sum_{i=1}^{\infty} \frac{n(i-a)}{a^2(ae^{i/a}+n)}\Big). \tag{E.4}$$

In view of Lemma E.1 the deterministic part in (E.4) is bounded from below by a negative constant times $\log^2(n/a)$. The stochastic part is bounded from below by $-C\sum_{i=1}^a Y_i^2$, and since $E_0[\sum_{i=1}^a Y_i^2] = \sum_{i=1}^a f_{0,i}^2 + an^{-1}$ and $V_0(\sum_{i=1}^a Y_i^2) = 2n^{-1}\sum_{i=1}^a f_{0,i}^2 + an^{-2} \to 0$ for all $a \leq A_n$ it follows from Chebyshev's inequality that the sum $\sum_{i=1}^a Y_i^2$ is bounded with probability going to one for all $f_0 \in \ell_2(M)$.

Next we deal with the lower bound on the interval $a \in \mathcal{I}_n$. First, note that $Y_i^2 \geq f_{0,i}^2 + 2f_{0,i}Z_i/\sqrt{n}$, which implies

$$\mathbb{M}_n(a) \geq \frac{1}{2}\Big(\sum_{i=1}^{a} \frac{n^2(i-a)e^{i/a}Y_i^2}{a(ae^{i/a}+n)^2} + \log^2(\frac{n}{a})g_n(a, f_0) + \mathbb{H}_n(a) - \sum_{i=1}^{\infty} \frac{n(i-a)}{a^2(ae^{i/a}+n)}\Big),$$

with the centered Gaussian process

$$\mathbb{H}_n(a) = \sum_{i=2a}^{\infty} \frac{n^{3/2}(i-a)e^{i/a}f_{0,i}Z_i}{a(ae^{i/a}+n)^2}. \tag{E.5}$$

Note that

$$V_0\Big(\frac{\mathbb{H}_n(a)}{\log^2(n/a)}\Big) = \frac{1}{\log^4(n/a)}\sum_{i=2a}^{\infty}\frac{n^3(i-a)^2e^{2i/a}f_{0,i}^2}{a^2(ae^{i/a}+n)^4}V_0(Z_i)$$
$$\leq \frac{ng_n(a,f_0)}{a\log^2(n/a)}\max_{i\geq 2a}\frac{(i-a)e^{i/a}}{(ae^{i/a}+n)^2} \geq \frac{g_n(a,f_0)}{a\log(n/a)},$$

and hence the diameter of the interval $\mathcal{I}_n$ with respect to the metric

$$d_n^2(a_1,a_2) = V_0\Big(\frac{\mathbb{H}_n(a_1)}{\log^2(n/a_1)} - \frac{\mathbb{H}_n(a_2)}{\log^2(n/a_2)}\Big)$$

is bounded by a multiple of $\sup_{a\in\mathcal{I}_n} g_n(a,f_0)^{1/2}(a\log(n/a))^{-1/2}$.

Next, we give an upper bound for the covering number of the interval $\mathcal{I}_n$. Let us take ε-balls centered at $a\in\mathcal{I}_n$, with $2a\in\mathbb{N}$. To cover the remaining part of the interval $\mathcal{I}_n$ it is sufficient to cover all intervals of the form $(a,a+1/2)$, $2a\in\mathbb{N}\cap 2\mathcal{I}_n$. Note that on these intervals, for every $a_1,a_2\in(a,a+1/2)$ we have $\lfloor 2a_1\rfloor - \lfloor 2a_2\rfloor = 0$. Hence in view of Lemma E.2 we have $d_n(a_1,a_2)\lesssim |a_1-a_2|\sup_{a\in\mathcal{I}_n}\sqrt{\log(n/a)g_n(a,f_0)/a^3}$. Thus the covering number of the interval $(a,a+1/2)$ relative to d_n is bounded from above by a multiple of $\varepsilon^{-1}\sup_{a\in\mathcal{I}_n}\sqrt{\log(n/a)g_n(a,f_0)/a^3}$, which implies that the covering number of the whole interval $\mathcal{I}_n$ is bounded from above by constant times $\varepsilon^{-1}\sup_{a\in\mathcal{I}_n}\sqrt{\log^{-1}(n/a)g_n(a,f_0)/a} + \underline{a}_n/\log(n/\underline{a}_n)$.

We show below that for any $c_0>2$,

$$e^{-2c_0}B\log n + o(1) \leq g_n(a,f_0) \leq e^{c_0}B\log n + o(1) \quad \text{for } a\in\mathcal{I}_n \tag{E.6}$$

hold. Therefore the covering number of $\mathcal{I}_n$ is bounded from above by a multiple of $\underline{a}_n + \varepsilon^{-1}\sqrt{\log^{-1}(n/\underline{a}_n)\log(n)/\underline{a}_n}$.

By Corollary 2.2.5 in [38] (applied with $\psi(x)=e^{x^2}-1$) it follows that

$$E_0\Big[\sup_{a\in\mathcal{I}_n}\Big|\frac{\mathbb{H}_n(a)}{\log^2(n/a)} - \frac{\mathbb{H}_n(\underline{a}_n)}{\log^2(n/\underline{a}_n)}\Big|\Big]$$
$$\lesssim \int_0^{C\log^{1/2}(n)I_{\underline{a}_n}^{-1/2}}\sqrt{\log\big(\underline{a}_n+\varepsilon^{-1}\sqrt{\log(n)\log^{-1}(n/\underline{a}_n)/\underline{a}_n}\big)}d\varepsilon$$
$$\lesssim \int_0^{C\log^{1/2}(n)I_{\underline{a}_n}^{-1/2}}\sqrt{\log\underline{a}_n}d\varepsilon + \int_0^1\log(1/\varepsilon)d\varepsilon = O(1).$$

Therefore the process $\mathbb{M}_n(a)$ can be bounded from below on $a\in\mathcal{I}_n$ by

$$\mathbb{M}_n(a) \geq 2^{-1}\inf_{a\in\mathcal{I}_n}\Big\{\log^2(n/a)\Big(Be^{-5}\log n - C\Big)$$
$$+\sum_{i=1}^{a}\frac{n^2(i-a)e^{i/a}Y_i^2}{a(ae^{i/a}+n)^2} - \sum_{i=1}^{\infty}\frac{n(i-a)}{a^2(ae^{i/a}+n)}\Big\}$$

with probability going to one. In view of (E.6), and since the third and fourth terms on the right-hand side of the preceding display are bounded from below by a fixed negative constant, we get that with probability tending to one, $\mathbb{M}_n(a) \geq e^{-5/2} B \log^3(n/\underline{a}_n)$.

It remains to verify assertion (E.6). First, note that

$$\frac{n^2}{\log^2(n/\underline{a}_n)} \sum_{i=c_0 I_{\underline{a}_n}}^{\infty} \frac{(i-\underline{a}_n)e^{i/\underline{a}_n}}{\underline{a}_n(\underline{a}_n e^{i/\underline{a}_n}+n)^2} f_{0,i}^2 \leq \frac{n^2}{\underline{a}_n^3 \log^2(n/\underline{a}_n)} \sum_{i=c_0 I_{\underline{a}_n}}^{\infty} i e^{-i/\underline{a}_n} f_{0,i}^2$$
$$\lesssim \frac{A_n^{c_0-2}}{n^{c_0-2}\log(n/\underline{a}_n)} \|f_0\|_2^2 = o(1).$$

Furthermore, in view of the inequality $c_0 I_{\underline{a}_n}(a^{-1} - \underline{a}_n^{-1}) \leq c_0$, for $a \in \mathcal{I}_n$, we have that

$$g_n(a, f_0) \geq \frac{n^2}{\log^2(n/a)} \sum_{i=2\underline{a}_n}^{c_0 I_{\underline{a}_n}} \frac{(i-a)e^{i/a}}{a(ae^{i/a}+n)^2} f_{0,i}^2$$
$$\geq \frac{n^2}{e^{2c_0}\log^2(n/\underline{a}_n)} \sum_{i=2\underline{a}_n}^{c_0 I_{\underline{a}_n}} \frac{(i-\underline{a}_n)e^{i/\underline{a}_n}}{\underline{a}_n(\underline{a}_n e^{i/\underline{a}_n}+n)^2} f_{0,i}^2.$$

By combining the preceding two displays, we get that

$$g_n(a, f_0) \geq e^{-2c_0} g_n(\underline{a}_n, f_0) - \frac{e^{-2c_0} n^2}{\log^2(n/\underline{a}_n)} \sum_{i=c_0 I_{\underline{a}_n}}^{\infty} \frac{(i-\underline{a}_n)e^{i/\underline{a}_n}}{\underline{a}_n(\underline{a}_n e^{i/\underline{a}_n}+n)^2} f_{0,i}^2$$
$$\geq e^{-2c_0} B \log n + o(1),$$

which finishes the proof of the first inequality in (E.6). The proof of the second inequality follows accordingly.

Lemma E.1. *There exists a constant $K > 0$ such that for any $a \in [1, n)$,*

$$\sum_{i=1}^{\infty} \frac{n(i-a)}{a^2(ae^{i/a}+n)} \leq K \log^2(n/a).$$

Proof. Note that

$$\sum_{i=1}^{\infty} \frac{n(i-a)}{a^2(ae^{i/a}+n)} \leq \sum_{i=1}^{\infty} \frac{ni}{a^2(ae^{i/a}+n)} \leq \sum_{i=1}^{I_a} \frac{i}{a^2} + \sum_{i=I_a}^{\infty} \frac{nie^{-i/a}}{a^3}$$
$$\lesssim \log^2(n/a) + \frac{\log(n/a)}{a} \lesssim \log^2(n/a). \qquad \blacksquare$$

Lemma E.2. *There exists a constant $K > 0$ such that for any $0 < a_1 < a_2$, $\lfloor 2a_2 \rfloor - \lfloor 2a_1 \rfloor = 0$,*

$$V_0\Big(\frac{\mathbb{H}_n(a_1)}{\log(n/a_1)^2} - \frac{\mathbb{H}_n(a_2)}{\log(n/a_2)^2}\Big) \leq K(a_1 - a_2)^2 \sup_{a \in [a_1, a_2]} \frac{\log(n/a) g_n(a, f_0)}{a^3}.$$

Proof. Recall that the left-hand side of the display in the lemma was denoted by $d_n^2(a_1,a_2)$, and note that

$$d_n^2(a_1,a_2)=\sum_{i=2a_2}^{\infty}\big(\phi_i(a_1)-\phi_i(a_2)\big)^2 n^3 f_{0,i}^2 \tag{E.7}$$

with $\phi_i(a):=\frac{(i-a)e^{i/a}}{\log(n/a)^2 a(ae^{i/a}+n)^2}$. Then by elementary, but cumbersome, computations we get that $|\phi_i'(a)|\lesssim ia^{-2}\phi_i(a)$. Thus, in view of Lemma H.3 the right-hand side of (E.7) is bounded by

$$\begin{aligned}&n^3(a_1-a_2)^2\sup_{a\in[a_1,a_2]}\sum_{i=2a}^{\infty}\frac{i^2}{a^4}\phi_i(a)^2 f_{0,i}^2\\&\qquad\lesssim (a_1-a_2)^2\sup_{a\in[a_1,a_2]} g_n(a)\sup_{i\in\mathbb{N}}\frac{ni^3e^{i/a}}{a^5\log^2(n/a)(ae^{i/a}+n)^2}.\end{aligned}$$

Then the statement of the lemma follows by applying Lemma H.1 (with $m=3$). ∎

E.2. $\mathbb{M}_n(a)$ on $[\overline{a}_n, A_n]$. We prove that for a sufficiently large choice of $K_0>0$ in the definition of $\overline{a}_n$,

$$\limsup_n \sup_{f_0\in\ell_2(M)}\sup_{a\in[\overline{a}_n,A_n]} E_0\Big[\frac{\mathbb{M}_n(a)}{\log^2(n/a)}\Big]<-2^{-5}, \tag{E.8}$$

$$\limsup_n \sup_{f_0\in\ell_2(M)} E_0\left[\sup_{a\in[\overline{a}_n,A_n]}\frac{|\mathbb{M}_n(a)-E_0[\mathbb{M}_n(a)]|}{\log^2(n/a)}\right]\le 2^{-6}. \tag{E.9}$$

These imply that with probability tending to one, $\mathbb{M}_n(a)<-2^{-6}\log^2(n/a)$ for every $a\in[\bar{a}_n,A_n]$, and hence the marginal likelihood function $\ell_n(a)$ is monotone decreasing and does not attain its global (or local) maximum on the interval $[\overline{a}_n,A_n]$, i.e.,

$$\inf_{f_0\in\ell_2(M)} P_0(\hat{a}_n\le\overline{a}_n)\to 1. \tag{E.10}$$

Proof of assertion (E.8). In view of $h_n(a,f_0)\le b$ for all $a\in[\overline{a}_n,A_n]$ (assuming that $\overline{a}_n>K_0$), we get that

$$\begin{aligned}E_0\Big[\frac{\mathbb{M}_n(a)}{\log^2(n/a)}\Big]&=\frac{1}{2}\Big(h_n(a,f_0)-\frac{1}{\log^2(n/a)}\sum_{i=1}^{\infty}\frac{n^2(i-a)}{a^2(ae^{i/a}+n)^2}\Big)\\&\le\frac{1}{2}\Big(b-\frac{1}{\log^2(n/a)}\sum_{i=1}^{\infty}\frac{n^2(i-a)}{a^2(ae^{i/a}+n)^2}\Big).\end{aligned}$$

In view of Lemma H.2 (with $r=0$ and $l=2$), we have $\sum_{i=1}^{\infty}\frac{n^2}{a(ae^{i/a}+n)^2}\lesssim\log(n/a)$. Furthermore,

$$\sum_{i=1}^{\infty}\frac{in^2}{a^2(ae^{i/a}+n)^2}\ge\sum_{i=1}^{I_a}\frac{i}{4a^2}=\frac{I_a(I_a+1)}{8a^2}\ge 2^{-3}\log^2\big(\frac{n}{a}\big),$$

which implies that

$$E_0\big[\mathbb{M}_n(a)/\log^2(n/a)\big] \le (b - 2^{-3} + o(1))/2,$$

which concludes the proof of assertion (E.8) for a small enough choice of b ($b < 2^{-4}$ is small enough).

Proof of assertion (E.9). In view of Corollary 2.2.5 in [38] (applied with $\psi(x) = x^2$) it is sufficient to show that there exist universal constants $K_1, K_2 > 0$ such that for any $a \in [\overline{a}_n, A_n]$,

$$V_0\big(\mathbb{M}_n(a)/\log^2(n/a)\big) \le K_1/\log(n/a), \tag{E.11}$$

$$\int_0^{diam_n} \sqrt{N(\varepsilon, [\overline{a}_n, A_n], d_n)}\, d\varepsilon \le K_2/K_0^{1/4}, \tag{E.12}$$

where d_n is the semimetric defined by $d_n^2(a_1, a_2) := V_0(\frac{\mathbb{M}_n(a_1)}{\log^2(n/a_1)} - \frac{\mathbb{M}_n(a_2)}{\log^2(n/a_2)})$, $diam_n$ is the diameter of $[\overline{a}_n, A_n]$ relative to d_n. and $N(\varepsilon, S, d_n)$ is the minimal number of d_n-balls of radius ε needed to cover the set S, since by a sufficiently large choice of K_0 ($K_0 \ge (2^6 K_2)^4$ is sufficiently large) assertion (E.9) holds.

Note that Lemma E.3 immediately implies assertion (E.11) and

$$diam_n \lesssim \sup_{a \in [\overline{a}_n, A_n]} (a \log(n/a))^{-1/2} \lesssim \log^{-1/2} n.$$

Then let us introduce the cover

$$[\overline{a}_n, A_n] \subset \bigcup_{k=0}^{K_n - 1} [2^k \overline{a}_n, 2^{k+1} \overline{a}_n]$$

with $K_n = \lceil \log(A_n/\overline{a}_n) \rceil$. In view of Lemma E.4, on the interval $a_1, a_2 \in [2^k \overline{a}_n, 2^{k+1} \overline{a}_n]$,

$$d_n(a_1, a_2) \lesssim \big(2^k \overline{a}_n\big)^{-3/2} \log^{1/2}(n) |a_1 - a_2|,$$

and hence

$$N(\varepsilon, [\overline{a}_n, A_n], d_n) \lesssim \sum_{k=0}^{K_n - 1} \frac{\log^{1/2}(n)}{\varepsilon \big(2^k \overline{a}_n\big)^{1/2}} \lesssim \frac{\log^{1/2}(n)}{\varepsilon \overline{a}_n^{1/2}}.$$

This results in

$$\int_0^{diam_n} \sqrt{N(\varepsilon, [\overline{a}_n, A_n], d_n)}\, d\varepsilon \le K_2/\overline{a}_n^{1/4} \le K_2/K_0^{1/4}.$$

Lemma E.3. *For all $a \in [\overline{a}_n, A_n]$, we have $V_0\big(\mathbb{M}_n(a)/\log^2(n/a)\big) \lesssim (a \log(n/a))^{-1}$.*

Proof. We know that the Y_i's are independent and that $V_0(Y_i^2) = 2/n^2 + 4f_{0,i}^2/n$, so the variance is equal to

$$V_0\Big(\frac{\mathbb{M}_n(a)}{\log^2(n/a)}\Big) = \frac{1}{4}\sum_{i=1}^{\infty}\frac{n^4 V_0(Y_i^2)e^{2i/a}(i-a)^2}{a^2\log^4(n/a)(ae^{i/a}+n)^4}$$

$$= \frac{1}{2}\sum_{i=1}^{\infty}\frac{n^2 e^{2i/a}(i-a)^2}{a^2\log^4(n/a)(ae^{i/a}+n)^4} + \sum_{i=1}^{\infty}\frac{n^3 e^{2i/a}(i-a)^2 f_{0,i}^2}{a^2\log^4(n/a)(ae^{i/a}+n)^4}. \tag{E.13}$$

In view of $(i-a)^2 \le a^2 + i^2$, for any $a, i > 0$, and by applying Lemma H.1 (with $m = 2$) and Lemma H.2 (first with $r = 2$ and $l = 4$ and then with $r = 1$ and $l = 2$) the first sum in (E.13) is bounded from above by a multiple of

$$\sum_{i=1}^{\infty}\frac{n^2 e^{2i/a}}{\log^4(n/a)(ae^{i/a}+n)^4} + \sum_{i=1}^{\infty}\frac{ne^{i/a}}{a\log^2(n/a)(ae^{i/a}+n)^2}$$

$$\lesssim \frac{1}{a\log^3(n/a)} + \frac{1}{a\log(n/a)} \lesssim \frac{1}{a\log(n/a)}.$$

Similarly, following from Lemma H.1 (with $m = 1$ and $m = -1$) and $h_n(a) \le b$ for $a \ge \overline{a}_n$, the second sum in (E.13) is bounded by a multiple of

$$\Big(\max_{i\in\mathbb{N}}\frac{ane^{i/a}}{i\log^2(n/a)(ae^{i/a}+n)^2} + \max_{i\in\mathbb{N}}\frac{ine^{i/a}}{a\log^2(n/a)(ae^{i/a}+n)^2}\Big)h_n(a,f_0)$$

$$\lesssim \Big(\frac{1}{a\log^3(n/a)} + \frac{1}{a\log(n/a)}\Big) \lesssim \frac{1}{a\log(n/a)},$$

which concludes the proof of the lemma. ∎

Lemma E.4. *For all $1 \le a_1 < a_2 < A_n$, we have*

$$d_n^2(a_1,a_2) \le C_0(a_1-a_2)^2 \sup_{a\in[a_1,a_2]}\frac{\log(n/a)}{a^3}(1+h_n(a,f_0))$$

for some universal constant $C_0 > 0$.

Proof. Note that

$$d_n^2(a_1,a_2) = n^4\sum_{i=1}^{\infty}(\phi_i(a_1)-\phi_i(a_2))^2 V_0(Y_i^2),$$

with $\phi_i(a) = \frac{e^{i/a}(i-a)}{2a\log^2(n/a)(ae^{i/a}+n)^2}$. By elementary computations one can see that $|\phi_i(a)'|^2 \lesssim (i^2a^{-4}+a^{-2})\phi_i^2(a)$, and hence in view of Lemma H.3,

$$d_n^2(a_1,a_2) \lesssim (a_1-a_2)^2 n^4 \sup_{a\in[a_1,a_2]}\sum_{i=1}^{\infty}\frac{e^{2i/a}(i^4+a^4)}{a^6\log^4(n/a)(ae^{i/a}+n)^4}V_0(Y_i^2).$$

Since $V_0(Y_i^2) = 2/n^2 + 4f_{0,i}^2/n$, the preceding sum is bounded by

$$\sum_{i=1}^{\infty} \frac{2e^{2i/a}(i^4+a^4)}{a^6 n^2 \log^4(n/a)(ae^{i/a}+n)^4} + \sum_{i=1}^{\infty} \frac{4e^{2i/a}(i^4+a^4)}{a^6 n \log^4(n/a)(ae^{i/a}+n)^4} f_{0,i}^2. \tag{E.14}$$

Then in view of Lemma H.1 (applied with $m = 4$ and $m = 0$) and Lemma H.2 (applied with $r = 1$ and $l = 2$) the first term of (E.14) is bounded from above by a multiple of

$$\sum_{i=1}^{\infty} \frac{e^{i/a}}{a^3 n^3 (ae^{i/a}+n)^2} \lesssim \frac{\log(n/a)}{a^3 n^4}.$$

Similarly, in view of Lemma H.1 (with $m = 3$ and $m = -1$) the second term of (E.14) is bounded by

$$\begin{aligned}\max_{i\in\mathbb{N}} &\frac{\big((i/a)^3 + (i/a)^{-1}\big)e^{i/a}}{a^2 n^3 \log^2(n/a)(ae^{i/a}+n)^2} h_n(a, f_0) \\ &\lesssim \Big(\frac{\log(n/a)}{a^3 n^4} + \frac{1}{n^5 a^2}\Big) h_n(a, f_0) \lesssim \frac{\log(n/a)}{a^3 n^4} h_n(a, f_0),\end{aligned}$$

which concludes the proof of the lemma. ∎

Appendix F. Proof of Theorem 2.5. Similarly to the previous appendices, we use the notation introduced in Appendix B. We show below that there exists a constant $c > 0$ depending only on m, M, and β_0 such that

$$\inf_{\beta\ge\beta_0} \inf_{f_0\in\Theta_s^\beta(m,M)} P_0(\hat{a}_n \ge c(n/\log n)^{1/(1+2\beta)}/\log n) \to 1, \tag{F.1}$$

which, combined with Proposition A.2 and Theorem A.1, results in

$$\inf_{\beta\ge\beta_0} \inf_{f_0\in\Theta_s^\beta(m,M)} P_0(c(n/\log n)^{1/(1+2\beta)} \le \tilde{a}_n \le C(n/\log n)^{1/(1+2\beta)}) \to 1$$

for some positive constants c, C depending on b, B, m, M, and β. Let us introduce the notation

$$\tilde{\mathcal{I}}_n = [c(n/\log n)^{\frac{1}{1+2\beta}}, C(n/\log n)^{\frac{1}{1+2\beta}}].$$

As before, note that $f_0 \in \hat{C}_n(L)$ is equivalent to $\|f_0 - \hat{f}\|_2 \le L r_\alpha(\tilde{a}_n)$, and hence by proving that

$$\begin{aligned}\inf_{a\in\tilde{\mathcal{I}}_n} r_\alpha^2(a) &\ge C_1(n/\log n)^{-2\beta/(1+2\beta)}, \\ \inf_{\beta\ge\beta_0} \inf_{f_0\in\Theta_s^\beta(m,M)} P_0\Big(\inf_{a\in\tilde{\mathcal{I}}_n} \|W(a)\|_2^2 &\le C_2(n/\log n)^{-2\beta/(1+2\beta)}\Big) \to 1, \\ \sup_{\beta\ge\beta_0} \sup_{f_0\in\Theta_s^\beta(m,M)} \sup_{a\in\tilde{\mathcal{I}}_n} \|B(a, f_0)\|_2^2 &\le C_3(n/\log n)^{-2\beta/(1+2\beta)}\end{aligned}$$

hold for some constants $C_1, C_2, C_3 > 0$, the statement of the theorem follows immediately. The proofs of the first two inequalities follow from (B.2) and (B.9) (with $\underline{a}_n$ and $\overline{a}_n$ replaced

by a multiple of $(n/\log n)^{1/(1+2\beta)}$), respectively. To prove the last inequality we note that for $f_0 \in \Theta_s^\beta(m, M)$, $a \in \tilde{\mathcal{I}}_n$, and $\beta \geq \beta_0$ we have that

$$\begin{aligned}\|B(a, f_0)\|_2^2 &\lesssim \sum_{i=1}^{I_a/2} a^2 e^{2i/a} n^{-2} i^{-1-2\beta} + \sum_{i=I_a/2}^{\infty} i^{-1-2\beta} \lesssim a/n + I_a^{-2\beta} \\ &= o\big((n/\log n)^{-2\beta/(1+2\beta)}\big).\end{aligned}$$

It remains to prove assertion (F.1). Let us introduce the slightly modified version of $\underline{a}_n$ as

$$\underline{a}'_n := \sup\{a \in [1, A_n] : g_n(a, f_0) \geq B\}$$

for some sufficiently large constant $B > 0$ to be specified later. Then we show below that

$$P_0(\hat{a}_n \geq \underline{a}'_n) \to 1 \quad \text{and} \quad \underline{a}'_n \geq c(n/\log n)^{1/(1+2\beta)}/\log n \tag{F.2}$$

for some sufficiently small constant $c > 0$.

For the second statement note that

$$g_n(a, f_0) \geq \frac{m}{\log^2(n/a)} n^2 \sum_{i=I_a}^{\infty} e^{-i/a} i^{-2\beta} \gtrsim mna^{-1-2\beta} \log^{-2-2\beta}(n/a), \tag{F.3}$$

and hence for any fixed $B > 0$ there exists a small enough $c > 0$ such that the right-hand side of the preceding display with $a = c(n/\log n)^{1/(1+2\beta)}/\log n$ is bigger than B. It remains to deal with the first part of (F.2). We show below that with probability tending to one, $\inf_{a \in [\underline{a}'_n/2, \underline{a}'_n]} \mathbb{M}_n(a) \geq cB \log^2(n/a)$, for some small enough constant $c > 0$, not depending on B. Then with probability tending to one for any $a \in [1, \underline{a}'_n/2]$ we have

$$\begin{aligned}\ell_n(\underline{a}_n) - \ell_n(a) &\geq \int_a^{\underline{a}'_n/2} \mathbb{M}_n(\tilde{a}) d\tilde{a} + \int_{[\underline{a}'_n/2, \underline{a}'_n]} \mathbb{M}_n(\tilde{a}) d\tilde{a} \\ &\geq -(\underline{a}'_n/2 - a) C \log^2(n/\underline{a}_n) + cB(\underline{a}'_n/2) \log^2(n/\underline{a}'_n) \\ &\geq (c/4) B \underline{a}'_n \log^2(n/\underline{a}'_n)\end{aligned}$$

for a large enough choice of $B > 0$, and hence the global maximum of $\ell_n(a)$ lies outside the interval $[1, \underline{a}'_n]$.

It remains to verify the lower bound for $M_n(a)$. First, note that for $a \leq A_n = o(n)$,

$$\begin{aligned}g_n(a, f_0) &\leq \frac{M}{\log^2(n/a)} \Big(\frac{1}{a} \sum_{i=2a}^{I_a} e^{i/a} i^{-2\beta} + \frac{n^2}{a^3} \sum_{i=I_a}^{\infty} e^{-i/a} i^{-2\beta} \Big) \\ &\leq c_{M,\beta} n a^{-1-2\beta} (\log n)^{-2-2\beta},\end{aligned}$$

and hence $\underline{a}'_n \leq c'_{M,\beta} B^{-1/(1+2\beta)} (n/\log n)^{1/(1+2\beta)}/\log n$. Therefore in view of (F.3), for every $a \geq \underline{a}'_n/2$ we have $g_n(a, f_0) \geq c_{M,\beta,m} B$ for some positive constant $c_{M,\beta,m} > 0$ not depending on B. Similarly, we can show that $g_n(a, f_0) \leq c'_{M,\beta,m} B$, for every $a \geq \underline{a}'_n/2$, for some $c'_{M,\beta,m} > 0$

not depending on B. Then following the same line of reasoning as in section E.1, with the only main difference being that instead of the interval given in (E.3) we are working with the interval $[\underline{a}'_n/2, \underline{a}'_n]$, we get that with probability going to one,

$$\begin{aligned}\inf_{a\in[\underline{a}'_n/2,\underline{a}'_n]} \mathbb{M}_n(a) &\geq 2^{-1} \inf_{a\in[\underline{a}'_n/2,\underline{a}'_n]} \Big\{ \log^2(n/a)\Big(c_{M,\beta,m}B - \sqrt{c'_{M,\beta,m}B}\Big) \\ &\quad + \sum_{i=1}^{a} \frac{n^2(i-a)e^{i/a}Y_i^2}{a(ae^{i/a}+n)^2} - \sum_{i=1}^{\infty} \frac{n(i-a)}{a^2(ae^{i/a}+n)} \Big\} \\ &\gtrsim B\log^2(n/\underline{a}'_n)\end{aligned}$$

for a large enough choice of $M > 0$, which finishes the proof of the theorem.

Appendix G. Proofs for the hierarchical Bayes procedure. In this section we prove the results on the hierarchical Bayes procedure (i.e., Theorems A.5 and 2.4 and Corollary 2.1) based on the results derived for the empirical Bayes procedure. First, we state that under the conditions of Theorem A.5 the hyperposterior distribution on the hyperparameter a concentrates most of its mass on the interval $\mathcal{I}_n = [\underline{a}_n \log(n)/(1+\log n), C\overline{a}_n]$ for some large enough constant $C > 0$.

Lemma G.1. *If $a \sim \pi(.)$ such that π verifies Assumption 1, then for sufficiently large $C > 0$ we have for every $\beta_0 > 0$ that*

$$\inf_{\beta>\beta_0} \inf_{f_0\in\Theta^\beta(M)} E_0\Pi\Big(\underline{a}_n \log(n)/(1+\log n) \leq a \leq C\overline{a}_n | Y\Big) = 1 + o(1/n).$$

G.1. Proof of Theorem A.5. Take $\varepsilon_n = (n/\log^2 n)^{-\beta/(1+2\beta)}$. Then following from Lemma G.1, we have

$$\begin{aligned}\sup_{f_0\in\Theta^\beta(M)} E_0\Pi\big(f : \|f - f_0\|_2 > M_n\varepsilon_n | Y\big) &\leq \sup_{f_0\in\Theta^\beta(M)} \Big(E_0\Pi(a \notin \mathcal{I}_n | Y) \\ &\quad + E_0 \sup_{a\in\mathcal{I}_n} \Pi_a\big(f : \|f - f_0\|_2 > M_n\varepsilon_n | Y\big)\Big) = o(1),\end{aligned}$$

where the last equation follows by arguments to those given in (A.9) and the displays below it (the only difference is that the supremum is taken over the interval $\mathcal{I}_n$ instead of $[\underline{a}_n, \overline{a}_n]$, but it only changes the constant factors, which do not play an essential role). This concludes the proof of the theorem.

G.2. Proof of Theorem 2.2—hierarchical Bayes part. In this proof we again use the notation introduced in Appendix B.

Let $a' := n^{1/(1+2\beta)}(\log n)^{-1-1/(1+2\beta)} \asymp \overline{a}_n \asymp \underline{a}_n$ with probability going to one thanks to Proposition A.2. One can see that in the hierarchical case,

$$P_0(f_0 \in \hat{C}_n(L_n)) \leq P_0\Big(\|B(a', f_0)\|_2 \leq L_n r_\alpha + \|W(a')\|_2 + \|\hat{f} - \hat{f}_{a'}\|_2\Big) + o(1), \tag{G.1}$$

which is a slightly modified version of (D.1) thanks to the triangle inequality. In order to prove that the right-hand side tends to zero, it is sufficient to show that there exist constants

$\tilde{C}_1, \tilde{C}_2, \tilde{C}_3 > 0$ such that

$$r_\alpha^2 \leq \tilde{C}_1 n^{-2\beta/(1+2\beta)} \log(n)^{-1/(1+2\beta)}, \tag{G.2}$$

$$P_0(\|W(a')\|_2^2 \leq \tilde{C}_2 n^{-2\beta/(1+2\beta)} \log(n)^{-1/(1+2\beta)}) \to 1, \tag{G.3}$$

$$\|B(a', f_0)\|_2^2 \geq \tilde{C}_3 n^{-2\beta/(1+2\beta)} \log(n)^{2\beta/(1+2\beta)}, \tag{G.4}$$

$$P_0(\|\hat{f} - \hat{f}_{a'}\|_2 \leq \tilde{C}_4 n^{-2\beta/(1+2\beta)} \log(n)^{-1/(1+2\beta)}) \to 1. \tag{G.5}$$

The bounds on the variance and the bias are obtained in a manner similar to that in Appendix D and section G.3. Next, we deal with assertion (G.5).

By Jensen's inequality, Fubini's theorem, and the triangle inequality, one can obtain that

$$\begin{aligned} \|\hat{f} - \hat{f}_{a'}\|_2 &= \| \int (\hat{f}_a - \hat{f}_{a'}) \Pi(da|Y) \|_2^2 \\ &\leq \sum_{i=1}^{\infty} \int (\hat{f}_{a,i} - \hat{f}_{a',i})^2 \Pi(da|Y) \\ &\leq \sup_{a_1 \in \mathcal{I}_n} \sum_{i=1}^{\infty} (\hat{f}_{a_1,i} - \hat{f}_{a',i})^2 \Pi(a_1 \in \mathcal{I}_n|Y) + \sup_{a_1 \notin \mathcal{I}_n} \sum_{i=1}^{\infty} (\hat{f}_{a_1,i} - \hat{f}_{a',i})^2 \Pi(a_1 \notin \mathcal{I}_n|Y). \end{aligned} \tag{G.6}$$

Starting with the first term, we use the trivial bound 1 for $\Pi(.|Y)$. We have with P_0-probability tending to one that

$$\begin{aligned} \sup_{a_1 \in \mathcal{I}_n} \sum_{i=1}^{\infty} (\hat{f}_{a_1,i} - \hat{f}_{a',i})^2 \leq & \sup_{a_1 \in \mathcal{I}_n} \sum_{i=1}^{\infty} (E_0 \hat{f}_{a_1,i} - E_0 \hat{f}_{a',i})^2 \\ & + \sup_{a_1 \in \mathcal{I}_n} \sum_{i=1}^{\infty} (\hat{f}_{a_1,i} - E_0 \hat{f}_{a1,i})^2 + \sum_{i=1}^{\infty} (\hat{f}_{a',i} - E_0 \hat{f}_{a',i})^2. \end{aligned} \tag{G.7}$$

The two last terms on the right-hand side are bounded by $n^{-2\beta/(1+2\beta)} \log(n)^{-1/(1+2\beta)}$ from (B.3). The first term can be written as $\sup_{a_1 \in \mathcal{I}_n} \sum_{i=1}^{\infty} (g_i(a_1) - g_i(a'))^2$ for $g_i(a) = n f_{0,i}^2/(a e^{i/a} + n)$. The derivative of $g_i(a)$ is $-n f_{0,i}^2 (a - i) e^{i/a}/(a(a e^{i/a} + n)^2)$. Without loss of generality, when $a_1 < a'$, writing the difference as the integral of $g_i'(a)$, applying the Cauchy–Schwarz inequality to its squares, and then interchanging the sum and the integral, we get that

$$\begin{aligned} \sum_{i=1}^{\infty} (E_0 \hat{f}_{a_1,i} - E_0 \hat{f}_{a',i})^2 &= \sum_{i=1}^{\infty} \Big(\int_{a_1}^{a'} g_i'(a) da \Big)^2 \leq \sum_{i=1}^{\infty} (a' - a_1) \int_{a_1}^{a'} g_i'(a)^2 da \\ &= (a' - a_1) \int_{a_1}^{a'} \sum_{i=1}^{\infty} g_i'(a)^2 da \leq (a' - a_1)^2 \sup_{a \in \mathcal{I}_n} \sum_{i=1}^{\infty} g_i'(a)^2 da \\ &\leq (a' - a_1)^2 \sup_{a \in \mathcal{I}_n} \sum_{i=1}^{\infty} \frac{n^2 f_{0,i}^4 (i - a)^2 e^{2i/a}}{a^2 (a e^{i/a} + n)^4}. \end{aligned}$$

For fixed a, the sum in the preceding display is bounded from above by constant times

$$\sum_{i=1}^{\infty} \frac{n^2 i^{-2-4\beta}(i-a)^2 e^{2i/a}}{a^2(ae^{i/a}+n)^4} \leq \frac{1}{a^2 n^2}\sum_{i=1}^{I_a}(i^2+a^2)i^{-2-4\beta}e^{2i/a} + \frac{n^2}{a^6}\sum_{i>I_a}(i^2+a^2)i^{-2-4\beta}e^{-2i/a}$$
$$\lesssim a^{-3-4\beta}\log\big(\frac{n}{a}\big)^{1-4\beta}.$$

Therefore, one can see that

$$\sup_{a_1\leq \bar{a}_n}\sum_{i=1}^{\infty}(E_0\hat{f}_{a_1,i}-E_0\hat{f}_{a',i})^2 \lesssim \sup_{a_1\in\mathcal{I}_n}(a'-a_1)^2\underline{a}_n^{-3-4\beta}\log(n)^{1-4\beta}$$
$$\lesssim n^{-1-2\beta/(1+2\beta)}\log(n)^{7+1/(1+2\beta)} = o(1/n),$$

with probability tending to one using Proposition A.2.

It is left to deal with the second term on the right-hand side of (G.6). Following from (G.7), we get with P_0-probability tending to one that

$$\sup_{a_1\notin\mathcal{I}_n}\sum_{i=1}^{\infty}(\hat{f}_{a_1,i}-\hat{f}_{a',i})^2 \leq 2\sup_{a_1\notin\mathcal{I}_n}\sum_{i=1}^{\infty}(E_0\hat{f}_{a_1,i})^2 + \sup_{a_1\notin\mathcal{I}_n}\sum_{i=1}^{\infty}(\hat{f}_{a_1,i}-E_0\hat{f}_{a_1,i})^2$$
$$+2\sum_{i=1}^{\infty}(E_0\hat{f}_{a',i})^2+\sum_{i=1}^{\infty}(\hat{f}_{a',i}-E_0\hat{f}_{a',i})^2, \tag{G.8}$$

where all terms on the right-hand side are $O(1)$. Since

$$E_0\Pi(a\notin\mathcal{I}_n|Y) = o(1/n),$$

applying Markov's inequality leads to the second term on the right-hand side of (G.6) being of lower order than n^{-1}.

It remains to deal with assertion (G.2). We show below that

$$r_\alpha \leq \tilde{r} := \sup_{a\in\mathcal{I}_n}\Big(\|\hat{f}-\hat{f}_a\|_2 + r_{\alpha/2}(a)\Big). \tag{G.9}$$

Then in view of the inequality

$$\sup_{a\in\mathcal{I}_n}\|\hat{f}-\hat{f}_a\|_2 \leq \|\hat{f}-\hat{f}_{a'}\|_2 + \sup_{a\in\mathcal{I}_n}\|\hat{f}_a-\hat{f}_{a'}\|_2$$

and assertions (G.5), (G.7), and (D.2), we get that with probability tending to one, $r_\alpha^2 \leq \tilde{C}_1 n^{-2\beta/(1+2\beta)}\log(n)^{-1/(1+2\beta)}$, and since r_α is deterministic, the inequality holds almost surely.

Finally, we verify assertion (G.9). Note that

$$\Pi(f:\|\hat{f}-f\|_2\leq\tilde{r}|Y) \geq \int_{\mathcal{I}_n}\Pi_a(f:\|f-\hat{f}\|_2\leq\tilde{r}|Y)\pi(a|Y)da$$
$$\geq \int_{\mathcal{I}_n}\Pi_a(f:\|f-\hat{f}_a\|_2\leq r_{\alpha/2}(a)|Y)\pi(a|Y)da$$
$$\geq \int_{\mathcal{I}_n}(1-\alpha/2)\pi(a|Y)da < 1-\alpha$$

for large enough n, which concludes the proof of our theorem for the hierarchical Bayes method.

G.3. Proof of Theorem 2.4—hierarchical Bayes part. Let us introduce the notation $W = \hat{f} - E_0\hat{f}$ and $B(f_0) = E_0\hat{f} - f_0$ for the centered hierarchical posterior mean and the bias of the posterior mean, respectively. Then $P_0(f_0 \in \hat{C}(L\log n))$ if and only if

$$\|W\|_2 \leq L\log(n) r_\alpha - \|B(f_0)\|_2 \tag{G.10}$$

holds. Using assertions (B.2), (B.3), and (B.4) we show below that there exist constants $\tilde{C}_1, \tilde{C}_2, \tilde{C}_3 > 0$, such that

$$r_\alpha^2 \geq \tilde{C}_1(\underline{a}_n/n)\log(n/\underline{a}_n), \tag{G.11}$$

$$\inf_{f_0 \in \Theta_{pt}(L_0,N_0,\rho)} P_0\big(\|W\|_2^2 \leq \tilde{C}_2(\underline{a}_n/n)\log(n/\underline{a}_n)\log^2 n\big) \to 1, \tag{G.12}$$

$$\|B(f_0)\|_2^2 \leq \tilde{C}_3(\underline{a}_n/n)\log^2\big(n/\underline{a}_n\big)\log n, \tag{G.13}$$

which results in (G.10) for a sufficiently large choice of $L > 0$.

Proof of (G.11). Let us take any $\alpha' > \alpha$ and note that in view of (B.2) we have

$$\inf_{a\in\mathcal{I}_n} r_{\alpha'}(a)^2 \geq C_1(\underline{a}_n/n)\log(n/\underline{a}_n).$$

Next, in view of Lemma G.1 and Anderson's lemma, we get for arbitrary $r \leq \inf_{a\in\mathcal{I}_n} r_{\alpha'}(a)$ that

$$\begin{aligned}\Pi(f:\ \|f-\hat{f}\|_2 \leq r|Y) &= \int_{\mathcal{I}_n} \Pi_a(f:\ \|f-\hat{f}\|_2 \leq r|Y)\pi(a|Y)da + o(1)\\ &\leq \int_{\mathcal{I}_n} \Pi_a(f:\ \|f-\hat{f}_a\|_2 \leq r_{\alpha'}(a)|Y)\pi(a|Y)da + o(1)\\ &\leq 1-\alpha' + o(1),\end{aligned}$$

and hence $r_\alpha^2 \geq \inf_{a\in\mathcal{I}_n} r_{\alpha'}(a)^2 \geq C_1(\underline{a}_n/n)\log(n/\underline{a}_n)$.

Proof of (G.12). Note that by the triangle inequality, Fubini's theorem, assertion (B.3), and Lemma G.1 we get that under the polished tail condition with P_0-probability tending to one,

$$\begin{aligned}\|W\|_2 &= \Big\|\int(\hat{f}_a - E_0\hat{f}_a)\pi(a|Y)da\Big\|_2\\ &\leq \sup_{a\in\mathcal{I}_n}\|W(a)\|_2\pi(\mathcal{I}_n|Y) + \sup_{1\leq a\leq A_n}\|W(a)\|_2\pi(\mathcal{I}_n^c|Y)\\ &\leq (C_2\underline{a}_n/n)^{1/2}\log(n/\underline{a}_n)^{1/2}\log n + o(1/n),\end{aligned}$$

where $\pi(\mathcal{I}_n|Y)$ denotes (by slightly abusing our notation) the posterior probability that the hyperparameter a lies in the interval $\mathcal{I}_n$ and in the last inequality we used in view of the proof of assertion (B.3) that $\sup_{1\leq a\leq A_n}\|W(a)\|_2 = O(1)$.

Proof of (G.13). Similarly to the proof of (G.12) we get that

$$\begin{aligned}\|B(f_0)\|_2^2 &\lesssim \sup_{a\in\mathcal{I}_n}\|B(a,f_0)\|_2^2 + o(\sup_{a\in[1,A_n]}\|B(a,f_0)\|_2^2/n)\\ &\leq C_3(\underline{a}_n/n)\log^2\big(n/\underline{a}_n\big)\log n + o(1/n),\end{aligned}$$

where the last inequality follows from $\|B(a,f_0)\|_2^2 \leq \|f_0\|_2^2 = O(1)$, which finishes the proof of the theorem.

G.4. Proof of Corollary 2.1. Let $\varepsilon_n = (n/\log^2 n)^{-\beta/(1+2\beta)}$, and first note that in view of assertions (G.12) and (G.13), combined with the triangle inequality and Proposition A.2, we have with P_0-probability tending to one that

$$\|f_0 - \hat{f}\|_2 \le \|W\|_2 + \|B(f_0)\|_2 \lesssim \sqrt{\underline{a}_n/n}\log(n/\underline{a}_n) \lesssim \varepsilon_n.$$

Then in view of Theorem A.5 and by again applying the triangle inequality, we get with probability tending to one that

$$\Pi(f : \|f - \hat{f}\|_2 \le M_n\varepsilon_n|Y) \ge \Pi(f : \|f - f_0\|_2 \le M_n\varepsilon_n - \|f_0 - \hat{f}\|_2|Y) = 1 - o(1),$$

which concludes the proof of the corollary.

G.5. Proof of Lemma G.1. In Appendix E it was shown that $\mathbb{M}_n(a) = \frac{\partial \ell_n(a)}{\partial a}$ satisfies, for positive constants K_1, K_2 and K_3,

$$\frac{\mathbb{M}_n(a)}{\log^2(n/a)} \begin{cases} \le -K_1 & \text{for } a \ge \overline{a}_n, \\ \ge K_2\log(n/\underline{a}_n) & \text{for } a \in [\underline{a}_n^*, \underline{a}_n], \\ \ge -K_3 & \text{for } a \le \underline{a}_n^*, \end{cases}$$

where $\underline{a}_n^* = \underline{a}_n \log n/(1+\log n)$. Furthermore, the constant K_2 can be chosen arbitrarily large by choosing B large enough, while the constant K_3 is fixed.

For $a \ge C\overline{a}_n$ with $C \ge 3$, we have

$$\ell_n(a) - \ell_n(2\overline{a}_n) \le -K_1\log^2(n/\overline{a}_n)(a - 2\overline{a}_n) \le -K_4\log^2(n/\overline{a}_n)\overline{a}_n$$

with $K_4 = K_1(C-2)$. Consequently $e^{\ell_n(a)} \le e^{\ell_n(2\overline{a}_n) - K_4\log^2(n/\overline{a}_n)\overline{a}_n}$ for $a \ge C\overline{a}_n$. Since also $e^{\ell_n(a)} \ge e^{\ell_n(2\overline{a}_n)}$ for $a \in [\overline{a}_n, 2\overline{a}_n]$, we find

$$\Pi(a \ge C\overline{a}_n|Y) \le \frac{\int_{C\overline{a}_n}^{\infty} e^{\ell_n(a)}\pi(a)da}{\int_{\overline{a}_n}^{2\overline{a}_n} e^{\ell_n(a)}\pi(a)da} \le \frac{\Pi\big([C\overline{a}_n, \infty)\big)e^{-K_4\log^2(n/\overline{a}_n)\overline{a}_n}}{\Pi([\overline{a}_n, 2\overline{a}_n])}. \tag{G.14}$$

Note that by Assumption 1,

$$\Pi\big([\overline{a}_n, 2\overline{a}_n]\big) \gtrsim \overline{a}_n^{1-c_3}e^{-c_2\overline{a}_n} \gg e^{-K_4\log^2(n/\overline{a}_n)\overline{a}_n},$$

and hence the right-hand side of (G.14) tends to zero.

The analysis of the left tail follows similarly. Note that for $a < \underline{a}_n^*/2$ we have $\ell_n(\underline{a}_n^*) - \ell_n(a) \ge -K_3(\underline{a}_n^* - a)\log^2(n/\underline{a}_n)$, hence $e^{\ell_n(a)} \le e^{\ell_n(\underline{a}_n^*) + K_3\underline{a}_n\log^2(n/\overline{a}_n)}$, and analogously for $(\underline{a}_n + \underline{a}_n^*)/2 < a < \underline{a}_n$ we have $\ell_n(a) - \ell_n(\underline{a}_n^*) \ge K_2(a - \underline{a}_n^*)\log^3(n/\underline{a}_n)$, which implies $e^{\ell_n(a)} \ge e^{\ell_n(\underline{a}_n^*) + K_2(\underline{a}_n/4)\log^2(n/\underline{a}_n)}$. Therefore,

$$\Pi(a \le \underline{a}_n^*|Y) \le \frac{\int_1^{\underline{a}_n^*} e^{\ell_n(a)}\pi(a)da}{\int_{(\underline{a}_n + \underline{a}_n^*)/2}^{\underline{a}_n} e^{\ell_n(a)}\pi(a)da} \le \frac{\Pi([1, \underline{a}_n])e^{K_3\underline{a}_n\log^2(n/\underline{a}_n)}}{\Pi([(\underline{a}_n + \underline{a}_n^*)/2, \underline{a}_n])e^{K_2(\underline{a}_n/4)\log^2(n/\underline{a}_n)}}. \tag{G.15}$$

Since

$$\Pi([(\underline{a}_n + \underline{a}_n^*)/2, \underline{a}_n])^{-1} \lesssim \log(n)\underline{a}_n^{c_5 - 1}e^{c_6\underline{a}_n} \ll e^{K_2(\underline{a}_n/8)\log^2(n/\underline{a}_n)}$$

for a large enough choice of K_2, the right-hand side of (G.15) tends to zero, which finishes the proof of the lemma.

Appendix H. Technical lemmas.

Lemma H.1. *Let $i, m \in \mathbb{N}$ and $a \geq 1$; then for any $n/a \geq e^m$,*

$$\frac{ne^{i/a}i^m}{a^m(ae^{i/a}+n)^2} \leq \frac{1}{a}\log^m\big(\frac{n}{a}\big) \vee e\frac{a^{-m}}{n}.$$

Proof. Assume first that $i \leq I_a \equiv a\log(n/a)$. Note that the function $f(x) = e^{x/a}(x/a)^m$ is monotone decreasing on $(-\infty, -ma]$ and monotone increasing on $[-ma, \infty]$. Then by the inequality $ae^{i/a} + n \geq n$,

$$\frac{ne^{i/a}i^m}{a^m(ae^{i/a}+n)^2} \leq \frac{e^{i/a}(i/a)^m}{n} \leq \frac{1}{a}\log^m\big(\frac{n}{a}\big) \vee e\frac{a^{-m}}{n}.$$

Next assume that $i > I_a$. Note that the derivative of the function $f(x) = e^{-x/a}x^m$ is $f'(x) = e^{-x/a}x^{m-1}(m - x/a)$, and hence the function $f(i)$ is monotone decreasing for $i \geq am$. Thus for $n/a \geq e^m$, $f(i)$ takes its maximum at $i = I_a$, which implies that

$$\frac{ne^{i/a}i^m}{a^m(ae^{i/a}+n)^2} \leq \frac{ne^{-i/a}i^m}{a^{m+2}} \leq \frac{1}{a}\log^m\big(\frac{n}{a}\big). \qquad \blacksquare$$

Lemma H.2. *Let $l > r \geq 0$; then for $n/a \geq e^{l-r}$,*

$$\sum_{i=1}^{\infty}\frac{e^{ir/a}}{(ae^{i/a}+n)^l} \lesssim \frac{n^{r-l}}{a^{r-1}}\log\big(\frac{n}{a}\big).$$

Proof. First, note that following from the inequality $ae^{i/a}+n \geq n$ and the sum of geometric series we get

$$\sum_{i=1}^{I_a}\frac{e^{ir/a}}{(ae^{i/a}+n)^l} \leq n^{-1}\sum_{i=1}^{I_a}e^{ir/a} \lesssim \frac{n^{r-l}}{a^{r-1}}\log\big(\frac{n}{a}\big),$$

where $I_a \equiv a\log(n/a)$. Then similarly, using the inequality $ae^{i/a} + n \geq ae^{i/a}$ and the sum of geometric series,

$$\sum_{I_a}^{\infty}\frac{e^{ir/a}}{(ae^{i/a}+n)^l} \leq a^{-l}\sum_{I_a}^{\infty}e^{(r-l)i/a} \leq \frac{n^{r-l}}{a^r}\frac{1}{e^{(l-r)/a}-1} \lesssim \frac{n^{r-l}}{a^{r-1}}\log\big(\frac{n}{a}\big)$$

because $e^{(l-r)/a} - 1 \geq \frac{l-r}{a}$ and $\log\big(\frac{n}{a}\big) \geq l - r$ for $\frac{n}{a} \geq e^{l-r}$. $\blacksquare$

Lemma H.3 (Lemma C.11 of [34]). *For any stochastic process $(V_a : a > 0)$ with continuously differentiable sample paths $a \mapsto V_a$, with derivative written as $\dot{V}_a$,*

$$E(V_{a_2} - V_{a_1})^2 \leq (a_2 - a_1)^2 \sup_{a\in[a_1,a_2]} E\dot{V}_a^2.$$

Appendix I. Extra simulation study. The purpose of this appendix is to reinforce the evidence shown in section 3. To this end, we show via plots and numerical properties the suboptimal performance of the GP with (approximately) squared exponential covariance kernel compared to other methods in the nonparametric regression model specifically. In this simulation study we take the Fourier coefficients of the underlying true function f_3 to be $f_{3,i} = i^{-3/2}\cos(i)$, $i = 1, 2, \ldots$. We take $\sigma^2 = 1/2$, but in the procedure it is considered to be unknown and estimated with the MMLE $\hat{\sigma}^2$. We take the sample sizes to be $n = 500, 1000, 5000$, and 10000. Observe in Figure 4 and Table 7 that the standard MMLE empirical Bayes method provides unreliable uncertainty quantification at certain points, especially compared to the other three methods. Furthermore, Table 8 shows that the size of the credible sets at those points contract for all methods. One can also observe through different running times in Table 9 that while the Matérn covariance kernels might provide robust credible sets, they substantially slow down the computations for large n.

We also investigate empirically the frequentist coverage probabilities of the pointwise credible sets by repeating the experiment 100 times and reporting the frequency that the function at given points (we consider $x = (0.25, 0.6474, 0.75)$ with $0.6474 = \text{argmax}_{x\in[0,1]} f_3(x)$) is included in the credible interval; see Table 1. Moreover, Table 2 shows the average size of the pointwise credible intervals (i.e., $2q_{0.025}\sqrt{\hat{c}(x,x)}$) depending on the sample size n and the procedure used to compute the credible sets. One can observe behavior similar to what we have described above.

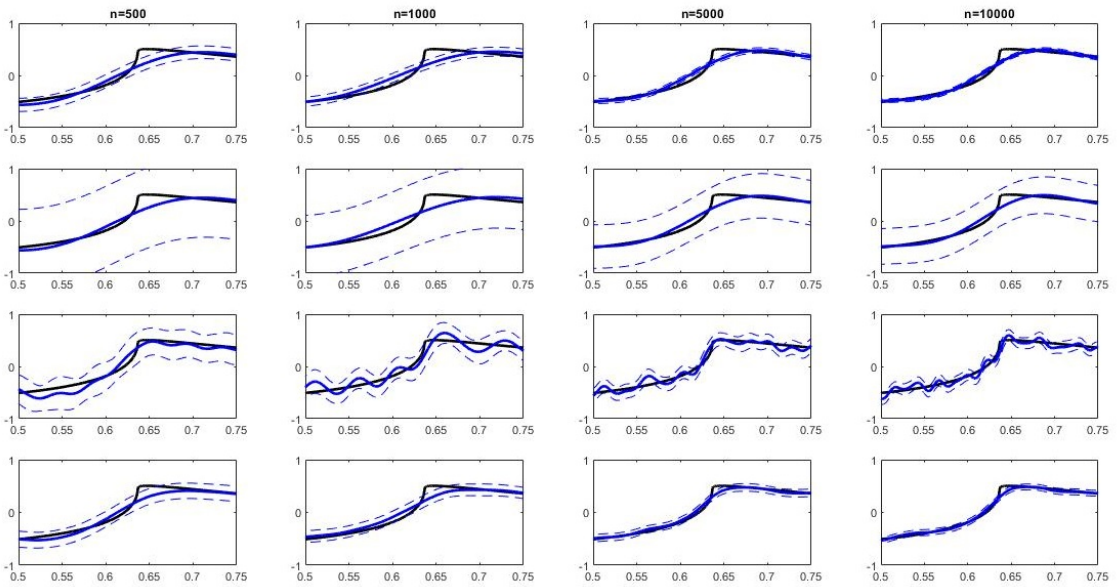

Figure 4. *Empirical Bayes credible sets for the regression function f_2 (drawn in black), zoomed in to the interval $x \in [0.5, 0.75]$. The posterior means are drawn as solid blue lines, while the 95% pointwise credible sets are dashed blue curves. We plot in the first row the MMLE empirical Bayes method, in the second row the MMLE empirical Bayes method with a $\log n$ blow up factor, in the third row the modified MMLE empirical Bayes method using squared exponential Gaussian process prior, and in the fourth row the empirical Bayes credible sets using a Matérn kernel with data-driven choice for the regularity hyperparameter. From left to right the sample size is $n = 500, 1000, 5000, 10000$.*

Table 7

Frequency of $f_3(x)$ being inside the corresponding credible interval for the squared exponential and Matérn GP prior at given points $x \in \{0.25, 0.3188, 0.75\}$. Method 1: squared exponential kernel MMLE empirical Bayes procedure. Method 2: squared exponential kernel empirical Bayes procedure with $\log n$ blow up factor. Method 3: squared exponential kernel modified empirical Bayes procedure (MMLE multiplied by $\log n$). Method 4: Matérn kernel with smoothness MMLE empirical Bayes. Method 5: Matérn kernel with rescaling MMLE empirical Bayes and $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000$.

	$x = 0.25$			$x = 0.6474$			$x = 0.75$		
$n =$	100	500	1000	100	500	1000	100	500	1000
Method 1	0.00	0.00	0.00	0.86	0.82	0.71	0.00	0.01	0.00
Method 2	0.94	1.00	1.00	0.64	1.00	1.00	1.00	1.00	1.00
Method 3	0.11	0.12	0.17	0.95	0.95	0.96	0.62	0.47	0.75
Method 4	0.00	0.13	0.14	0.86	0.86	0.90	0.24	0.27	0.35
Method 5	0.00	0.08	0.10	0.83	0.84	0.87	0.19	0.19	0.34

Table 8

Average size of the pointwise credible intervals (i.e., $2q_{0.025}\sqrt{\hat{c}(x,x)}$) for $f_3(x)$ in the regression model. Method 1: squared exponential kernel MMLE empirical Bayes procedure, Method 2: squared exponential kernel empirical Bayes procedure with $\log n$ blow up factor. Method 3: squared exponential kernel modified empirical Bayes procedure (MMLE multiplied by $\log n$). Method 4. Matérn kernel with smoothness MMLE empirical Bayes. Method 5: Matérn kernel with rescaling MMLE empirical Bayes and $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000$.

$n =$	100	500	1000
Method 1	0.4184	0.2335	0.1804
Method 2	1.9270	1.4516	1.2459
Method 3	0.7949	0.5271	0.4267
Method 4	0.6694	0.4292	0.3446
Method 5	0.5439	0.2987	0.2625

Table 9

Average run time of the EB methods for f_3 in the regression model. Method 1: squared exponential covariance kernel. Method 4: Matérn covariance kernel and MMLE for the regularity hyperparameter. Method 5: Matérn covariance kernel and MMLE for the scaling hyperparameter with fixed regularity $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000, 5000, 10000$.

$n =$	100	500	1000	5000	10000
Method 1	0.82 s	2.70 s	10.66 s	3.6 m	19.1 m
Method 4	1.61 s	14.52 s	45.77 s	19 m	4.2 h
Method 5	1.37 s	11.45 s	34.29 s	14.1 m	2.4 h

We also consider a multidimensional version of the previous regression with $d = 10$. The Fourier coefficients of the underlying true function f_4 become $f_{4,i} = \prod_{k=1}^{10}(i_k^{-3/2}\cos(i_k))$, $i_k = 1, 2, \ldots,$ for all $k = 1, 2, \ldots, 10$, relative to the Fourier eigenbasis $\psi_i(t) = 32\prod_{k=1}^{10}\cos(\pi(i_k - 1/2)t)$. We have collected the frequentist coverage probabilities of the pointwise credible sets at given points (we consider $x = (\{0.25\}^{10}, \{0.3188\}^{10}, \{0.75\}^{10})$) in Table 10 and note that we can draw conclusions similar to those in the $d = 1$ dimensional case. Surprisingly, the computation times for the posterior in higher dimension is of a similar order to that in their one-dimensional counterpart, and hence they are omitted.

Table 10

Frequency of $f_2(x)$ being inside the corresponding credible interval for the squared exponential and Matérn GP prior in the multivariate ($d = 10$) regression model. Method 1: *squared exponential kernel MMLE empirical Bayes procedure. Method* 2: *squared exponential kernel empirical Bayes procedure with* $\log n$ *blow up factor. Method* 3: *squared exponential kernel modified empirical Bayes procedure (MMLE multiplied by* $\log n$*). Method* 4: *Matérn kernel with smoothness MMLE empirical Bayes, Method* 5: *Matérn kernel with rescaling MMLE empirical Bayes and* $\alpha = 10$*. From left to right the sample size is* $n = 100, 500, 1000$.

	$x = \{0.25\}^{10}$			$x = \{0.3188\}^{10}$			$x = \{0.75\}^{10}$		
$n =$	100	500	1000	100	500	1000	100	500	1000
Method 1	1.00	0.97	0.96	0.85	0.76	0.70	1.00	0.98	0.95
Method 2	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00
Method 3	1.00	1.00	1.00	0.86	0.87	0.89	1.00	1.00	1.00
Method 4	1.00	1.00	1.00	0.96	0.95	0.97	1.00	1.00	1.00
Method 5	1.00	1.00	1.00	0.95	0.95	0.94	1.00	1.00	1.00

REFERENCES

[1] L. S. Bastos and A. O'Hagan, *Diagnostics for Gaussian process emulators*, Technometrics, 51 (2009), pp. 425–438.

[2] E. Belitser, *On coverage and local radial rates of credible sets*, Ann. Statist., 45 (2017), pp. 1124–1151, https://doi.org/10.1214/16-AOS1477.

[3] S. Bhatt, D. Weiss, E. Cameron, D. Bisanzio, B. Mappin, U. Dalrymple, K. Battle et al., *The effect of malaria control on Plasmodium falciparum in Africa between* 2000 *and* 2015, Nature, 526 (2015), pp. 207–211.

[4] A. Bhattacharya and D. Pati, *Adaptive Bayesian inference in the Gaussian sequence model using exponential-variance priors*, Statist. Probab. Lett., 103 (2015), pp. 100–104.

[5] A. Bhattacharya, D. Pati, and D. Dunson, *Anisotropic function estimation using multi-bandwidth Gaussian processes*, Ann. Statist., 42 (2014), pp. 352–381.

[6] L. D. Brown and M. G. Low, *Asymptotic equivalence of nonparametric regression and white noise*, Ann. Statist., 24 (1996), pp. 2384–2398, https://doi.org/10.1214/aos/1032181159.

[7] T. Cai and M. Low, *An adaptation theory for nonparametric confidence intervals*, Ann. Statist., 32 (2004), pp. 1805–1840.

[8] I. Castillo, G. Kerkyacharian, and D. Picard, *Thomas Bayes' walk on manifolds*, Probab. Theory Relat. Fields, 158 (2014), pp. 665–710, https://doi.org/10.1007/s00440-013-0493-0.

[9] I. Castillo and R. Nickl, *Nonparametric Bernstein–von Mises theorems in Gaussian white noise*, Ann. Statist., 41 (2013), pp. 1999–2028, https://doi.org/10.1214/13-AOS1133.

[10] D. L. Donoho, *Statistical estimation and optimal recovery*, Ann. Statist., 22 (1994), pp. 238–270, https://doi.org/10.1214/aos/1176325367.

[11] J. Faraway, A. Mahabal, and J. Berger, *Robust Bayesian displays for standard inferences concerning a normal mean*, Comput. Statist. Data Anal., 33 (2000), pp. 381–399.

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

[13] S. Ghosal and A. van der Vaart, *Convergence rates of posterior distributions for noniid observations*, Ann. Statist., 35 (2007), pp. 192–223.

[14] S. Ghosal and A. van der Vaart, *Fundamentals of Nonparametric Bayesian Inference*, Cambridge Ser. Statist. Probab. Math. 44, Cambridge University Press, 2017, https://doi.org/10.1017/9781139029834.

[15] E. Giné and R. Nickl, *Confidence bands in density estimation*, Ann. Statist., 38 (2010), pp. 1122–1170, https://doi.org/10.1214/09-AOS738.

[16] E. Giné and R. Nickl, *Mathematical Foundations of Infinite-Dimensional Statistical Models*, Cambridge Ser. Statist. Probab. Math. 40, Cambridge University Press, 2016.

[17] A. Kalaitzis and N. Lawrence, *A simple approach to ranking differentially expressed gene expression*

time courses through Gaussian process regression, BMC Bioinform., 12 (2011), 180.
[18] B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten, *Bayesian inverse problems with Gaussian priors*, Ann. Statist., 39 (2011), pp. 2626–2657, https://doi.org/10.1214/11-AOS920.
[19] B. T. Knapik, B. T. Szabó, A. W. van der Vaart, and J. H. van Zanten, *Bayes procedures for adaptive inference in inverse problems for the white noise model*, Probab. Theory Relat. Fields, 164 (2016), pp. 771–813, https://doi.org/10.1007/s00440-015-0619-7.
[20] T. Koriyama and T. Kobayashi, *Prosody generation using frame-based Gaussian process regression and classification for statistical parametric speech synthesis*, in Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4929–4933.
[21] M. Low, *On nonparametric confidence intervals*, Ann. Statist., 25 (1997), pp. 2547–2554.
[22] M. Nussbaum, *Asymptotic equivalence of density estimation and Gaussian white noise*, Ann. Statist., 24 (1996), pp. 2399–2430, https://doi.org/10.1214/aos/1032181160.
[23] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.
[24] C. E. Rasmussen, *Gaussian processes in machine learning*, in Advanced Lectures on Machine Learning, Springer, 2004, pp. 63–71.
[25] K. Ray, *Adaptive Bernstein–von Mises theorems in Gaussian white noise*, Ann. Statist., 45 (2017), pp. 2511–2536, https://doi.org/10.1214/16-AOS1533.
[26] J. Robins and A. van der Vaart, *Adaptive nonparametric confidence sets*, Ann. Statist., 34 (2006), pp. 229–253, https://doi.org/10.1214/009053605000000877.
[27] J. Rousseau and B. Szabó, *Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator*, Ann. Statist., 45 (2017), pp. 833–865.
[28] J. Rousseau and B. Szabó, *Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors*, Ann. Statist., 48 (2020), pp. 2155–2179, https://doi.org/10.1214/19-AOS1881.
[29] S. Sniekers and A. van der Vaart, *Adaptive Bayesian credible sets in regression with a Gaussian process prior*, Electron. J. Stat., 9 (2015), pp. 2475–2527, https://doi.org/10.1214/15-EJS1078.
[30] B. T. Szabó, A. W. Vaart, and J. H. van Zanten, *Honest Bayesian confidence sets for the L^2-norm*, J. Statist. Plann. Inf., 166 (2015), pp. 36–51, https://doi.org/10.1016/j.jspi.2014.06.005.
[31] B. T. Szabo, A. W. van der Vaart, and J. H. van Zanten, *Empirical Bayes scaling of Gaussian priors in the white noise model*, Electron. J. Stat., 7 (2013), pp. 991–1018, https://doi.org/10.1214/13-EJS798.
[32] B. T. Szabo, A. W. van der Vaart, and J. H. van Zanten, *Frequentist coverage of adaptive nonparametric Bayesian credible sets*, Ann. Statist., 43 (2015), pp. 1391–1428.
[33] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, revised and extended from the 2004 French original, translated by Vladimir Zaiats, Springer Ser. Statist., Springer, 2009.
[34] S. van der Pas, B. Szabó, and A. van der Vaart, *Adaptive posterior contraction rates for the horseshoe*, Electron. J. Statist., 11 (2017), pp. 3196–3225, https://doi.org/10.1214/17-EJS1316.
[35] A. van der Vaart and J. H. van Zanten, *Bayesian inference with rescaled Gaussian process priors*, Electron. J. Statist., 1 (2007), pp. 433–448.
[36] A. van der Vaart and J. H. van Zanten, *Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth*, Ann. Statist., 37 (2009), pp. 2655–2675.
[37] A. W. van der Vaart and J. H. van Zanten, *Rates of contraction of posterior distributions based on Gaussian process priors*, Ann. Statist., 36 (2008), pp. 1435–1463.
[38] A. W. Van Der Vaart and J. A. Wellner, *Weak convergence*, in Weak Convergence and Empirical Processes, Springer, 1996, pp. 16–28.
[39] Y. Yang, A. Bhattacharya, and D. Pati, *Frequentist Coverage and sup-norm Convergence Rate in Gaussian Process Regression*, preprint, https://arxiv.org/abs/1708.04753, 2017.
[40] Y. Yang and D. B. Dunson, *Bayesian manifold regression*, Ann. Statist., 44 (2016), pp. 876–905, https://doi.org/10.1214/15-AOS1390.
[41] W. W. Yoo and S. Ghosal, *Supremum norm posterior contraction and credible sets for nonparametric multivariate regression*, Ann. Statist., 44 (2016), pp. 1069–1102, https://doi.org/10.1214/15-AOS1398.