# Scalability and uncertainty of Gaussian processes

Hadji, M.A.

CHAPTER 2

# Adaptive credible sets with a squared-exponential GP prior

*This chapter has been published as: A. Hadji, B. Szabó, "Can we trust Bayesian uncertainty quantification from Gaussian process priors with squared exponential covariance kernel?" (2021) in SIAM/ASA Journal on Uncertainty Quantification, (9), 1, 185-230*

**Abstract.** We investigate the frequentist coverage properties of credible sets resulting in from Gaussian process priors with squared exponential covariance kernel. First we show that by selecting the scaling hyper-parameter using the maximum marginal likelihood estimator in the (slightly modified) squared exponential covariance kernel the corresponding credible sets will provide overconfident, misleading uncertainty statements for a large, representative subclass of the functional parameters in 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 statement and adaptive size of the credible sets under some additional restriction. Finally we demonstrate on a numerical study that the derived negative and positive results extend beyond the Gaussian white noise model to the non-parametric 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.

## §2.1 Main results

### §2.1.1 Model description

We consider the Gaussian white noise model

$$Y(t) = \int_0^t \theta_0(s)ds + \frac{1}{\sqrt{n}}W_t, \quad t \in [0,1], \tag{2.1.1}$$

where $\theta_0 \in L_2[0,1]$ is the unknown function of interest and $W_t$ denotes the Brownian motion. Let $P_0, E_0$, and $V_0$ denote the corresponding probability measure, expected value, and variance, respectively. In the Bayesian approach we endow the unknown function of interest $\theta_0$ with a prior distribution representing our initial belief. In our

work we investigate the popular Gaussian process prior with rescaled squared exponential kernel. Let us consider the sequence representation of the Gaussian white noise model. For an orthonormal basis $\psi_i$, $i = 1, 2, ...$ (e.g. the Fourier basis) let us denote the sequence decomposition of the functions $\theta_0(t), Y(t)$, and $W_t$ by $Y_i = \langle Y(t), \psi_i(t)\rangle_2$, $\theta_{0,i} = \langle \theta_0, \psi_i(t)\rangle_2$, and $Z_i = \langle W_t, \psi_i(t)\rangle_2 \overset{iid}{\sim} N(0,1)$, $i = 1, 2, ...$, respectively. Then the equivalent sequence model can be given in the form

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

Slightly abusing our notations we denote by $\theta_0$ both the functional parameter in the Gaussian white noise model and the sequential parameter $\theta_0 = (\theta_{0,1}, \theta_{0,2}, ...)$ in the sequence model. It is common to assume that the true function $\theta_0$ belongs to a hyper-rectangle regularity class, i.e.

$$\theta_0 \in \Theta^\beta(M) = \left\{ \theta \in \ell_2 : \sup_{i \geq 1} \theta_i^2 i^{2\beta+1} \leq M \right\},$$

for some (typically unknown) $\beta, M > 0$. The class $\Theta^\beta(M)$ is closely related to Sobolev type of regularity classes $S^\beta(M) = \{\theta \in \ell_2 : \sum_{i \geq 1} \theta_i^2 i^{2\beta} \leq M\}$ and the derived results can (typically) easily be extended to them, see for instance (Szabo et al., 2015). We note that the minimax estimation rate for the above hyper-rectangle regularity class is $n^{-\beta/(2\beta+1)}$, i.e. there exists $C_\beta > 0$ such that

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

where the infimum is taken over all estimators, see for instance (Donoho, 1994).

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

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

where $\lambda_i$, $\psi_i$, $i = 1, 2, ...$ are the eigenvalues and eigenfunctions of the squared exponential covariance kernel, and $\xi_i$ are iid standard normal random variables, see for instance Chapter 4.3 of (Rasmussen, 2004). The corresponding coefficients $\lambda_i$ can be approximated as $\lambda_i^2 \approx a^{-1} e^{-i/a}$ for $\mathcal{X} = \mathbb{R}$ and with respect to the Gaussian dominated measure, see for instance (Rasmussen, 2004). In the rest of the chapter we (mainly) work with the prior

$$\theta | a \sim \bigotimes_{i=1}^{\infty} N(0, a^{-1} e^{-i/a}) \tag{2.1.2}$$

in the sequence model, for convenience. Note that in view of $Y|\theta_0 \sim \bigotimes_{i=1}^{\infty} N(\theta_{0,i}, n^{-1})$ and the choice of the prior $\Pi_a(\cdot)$ in (2.1.2), the corresponding posterior $\Pi_a(\cdot|Y)$ takes the form

$$\theta | a, Y \sim \bigotimes_{i=1}^{\infty} \mathcal{N}\left( \frac{nY_i}{ae^{i/a} + n}, \frac{1}{ae^{i/a} + n} \right). \tag{2.1.3}$$

The behavior of the posterior distribution is very sensitive on the choice of the hyper-parameter $a$. Since the optimal choice of $a$ depends on the (typically) unknown regularity parameter $\beta$ of the underlying functional parameter of interest $\theta_0$ in practice one uses data driven procedures for selecting $a$. The two most commonly applied Bayesian techniques for selecting the hyper-parameter are the hierarchical Bayes and the marginal likelihood empirical Bayes methods. In the hierarchical Bayes method the hyper-parameter $a$ is endowed with a prior distribution $\pi$ (also called hyper-prior 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 hyper-prior density function $\pi(\cdot)$ supported on $[1, A_n]$.

**Assumption 2.1.1.** *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*

$$c_4^{-1} a^{-c_3} \exp(-c_2 a) \leq \pi(a) \leq c_4 a^{-c_5} \exp(-c_6 a), \tag{2.1.4}$$

*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 2.1.1.

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

$$\hat{a}_n := \arg \max_{a \in [1, A_n]} \ell_n(a), \tag{2.1.5}$$

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 both from practical and analytical perspective. Then the estimator $\hat{a}_n$ is plugged in into the posterior distribution (2.1.3), resulting in the empirical Bayes posterior $\Pi_{\hat{a}_n}(\cdot|Y)$.

We show in Section 2.4.2 that both of these methods result in optimal recovery for the functional parameter of interest $\theta_0$. These results are of interest on their own right, but our main focus lies on the reliability of Bayesian uncertainty quantification resulting both from the hierarchical and the empirical Bayes procedures, hence we have deferred the contraction rate results to the Section 2.4.2.

## §2.1.2 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 $\ell_2$-credible balls centered around the posterior mean, i.e. we analyze credible sets in the form

$$\hat{C}_{n,\alpha} = \left\{\theta \in \ell_2 : \|\theta - \hat{\theta}\|_2 \leq r_\alpha\right\} \tag{2.1.6}$$

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

We are interested in the frequentist properties of $\ell_2$-credible balls resulting from the data driven credible balls. Then let us denote by $r_\alpha$ the radius of the $\ell_2$-ball centered around the posterior mean $\hat{\theta}$ and accumulating $1-\alpha$ fraction of the posterior mass, i.e.

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

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

$$\hat{C}_n(L_n) = \left\{\theta \,\|\theta - \hat{\theta}\|_2 \leq L_n r_\alpha\right\}, \tag{2.1.7}$$

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

**Corollary 2.1.2.** *Both the hierarchical and the empirical Bayes credible sets defined in* (2.1.7) *have rate adaptive size, i.e. for every $\beta_0 > 0$ and $M > 0$*

$$\sup_{\beta \geq \beta_0} \sup_{\theta \in \Theta^\beta(M)} P_0\left(diam(\hat{C}_n(1)) \geq M_n n^{-\beta/(2\beta+1)} (\log n)^{-1/(2\beta+1)}\right) \to 0,$$

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

*Proof.* The proof is given in Sections 2.7 and 2.10 respectively. □

#### 2.1.2.1 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 $\theta_0$ is included in the (blown up) credible set, i.e. if any fixed $\beta_0 > 0$

$$\inf_{\theta_0 \in \cup_{\beta \geq \beta_0} \Theta^\beta(M)} P_0(\theta_0 \in \hat{C}_n(L_n)) \geq 1-\alpha$$

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 2.1.2, they must have poor frequentist coverage properties at least for certain functional parameters $\theta_0$. Actually the radius of the credible sets decays even faster than the minimax rate, which already implies impossibility of coverage. Nevertheless it is of interest to quantify the set of functions for which the Bayesian uncertainty quantification is truth-worthy.

First we note that a representative subset of the hyper-rectangle $\Theta^\beta(M)$ is the set

$$\Theta_s^\beta(m, M) = \left\{ \theta \in \Theta^\beta(M) : \min_{i \geq 1} i^{1+2\beta} \theta_i^2 \geq m \right\}, \tag{2.1.8}$$

for some parameters $0 < m \leq M$. Let us refer to this subclass of sequential parameters as self-similar signals following the similar terminology of (Giné and Nickl, 2010) and (Szabo et al., 2015). It was shown in the later paper that the minimax rate over $\Theta_s^\beta(m, M)$ is the same as over $\Theta^\beta(M)$. The next theorem shows that both of the hierarchical and the empirical Bayes procedures provide unreliable uncertainty quantification over this representative sub-class of functions unless it is blown up with at least a logarithmic factor.

**Theorem 2.1.3.** *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 $0 < m \leq M$,*

$$\sup_{\theta_0 \in \Theta_s^\beta(m,M)} P_{\theta_0}(\theta_0 \in \hat{C}_n(L_n)) \to 0.$$

*Proof.* See Section 2.7. □

This negative result draws a dark picture as it tells us that one can not trust Bayesian uncertainty quantification resulting from the investigated prior, even if one allows certain amount of adjustment (i.e. by blowing up the set with a sequence tending to infinity, not too fast). Since the prior used is very closely related to the Gaussian process with squared exponential covariance kernel this gives the intuition that one has to be very cautious working with squared exponential kernel as the corresponding Bayesian uncertainty statement are (typically) unreliable. In the next subsection we will be touching the corners 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 second we show that by blowing up the credible sets by a logarithmic factor or by slightly adjusting the maximum marginal likelihood estimator, one gets reliable uncertainty statements for a large subclass of functions, including the self-similar functions.

#### 2.1.2.2 Coverage of credible sets - positive results

Let us consider the set of analytic-type functions defined as

$$\theta_0 \in A^\gamma(M) = \left\{ \theta \in \ell_2 : \sum_{i=1}^{\infty} \theta_i^2 e^{2i\gamma} \leq M \right\},$$

for some $\gamma > 0$. Note that the investigated prior (2.1.2) 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 2.1.4.** *The inflated empirical and hierarchical Bayes credible sets $\hat{C}_n(L)$ have frequentist coverage tending to one over the class $\theta_0 \in A^\gamma(M)$ for any $\gamma \geq 1/2$ and sufficiently large constant $L > 0$, i.e.*

$$\inf_{\theta_0 \in A^\gamma(M)} P_0(\theta_0 \in \hat{C}_n(L)) \to 1.$$

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

$$\inf_{\theta_0 \in A^\gamma(M)} P_0\left( diam(\hat{C}_n(1)) \leq Cn^{-1/2} \log n \right) \to 1.$$

*Proof.* See Section 2.6. □

Next we investigate the behavior of the credible sets by allowing a logarithmic inflating factor. Since the size of the inflated credible sets are still nearly minimax, the credible sets fail to cover all functional parameter $\theta_0$ of interest, in view of the non-existence result of adaptive confidence sets seen in (Cai and Low, 2004) and (Robins and van der Vaart, 2006). Therefore we restrict the investigated class of functions to the so called polished tail class, introduced in (Szabo et al., 2015) and (Rousseau and Szabo, 2020). We say that a sequential parameter $\theta \in \ell_2(M)$ 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

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

The above assumption basically requires that knowing the sequential parameter $\theta$ up to a certain coordinate enables us to draw conclusion about the tail of the sequence. We require that the energy (sum of the squared coefficients) of the tail is dominated by the energy of a finitely large block of coefficients. This condition makes also sense intuitively as in the stochastic model the signal can be observed only up to some limit, the fluctuation in the later coordinates can equally likely be caused by the noise. Therefore to make reliable uncertainty statement 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 given signal-to-noise ratio. In (Szabo et al., 2015) it was shown that the above assumption is mild from statistical, topological and Bayesian point of view as well.

The next theorem states that when the sequential parameter $\theta_0$ is restricted to polished tail sequences, then 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 honest frequentist confidence set, for large enough $L$.

**Theorem 2.1.5.** *For any $L_0, N_0, \rho \geq 1$ there exists a constant $L$ such that*

$$\inf_{\theta_0 \in \Theta_{pt}(L_0, N_0, \rho)} P_0(\theta_0 \in \hat{C}_n(L(\log n)^{3/2})) \to 1,$$

*where $\hat{C}_n$ denotes either the empirical or the hierarchical Bayes credible sets under Assumption 2.1.1.*

*Proof.* See Sections 2.5 and 2.10.3 respectively. □

Hence 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 introducing 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 2.1.3 one can see that the MMLE $\hat{a}_n$, given in (2.1.5), is too small, the empirical Bayes procedure is basically oversmoothing. One can compensate for this by undersmoothing the procedure. We propose to adjust the MMLE by a multiplicative logarithmic factor

$$\tilde{a}_n = \log(n)\hat{a}_n. \tag{2.1.9}$$

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

$$\inf_{\theta_0 \in \Theta_s^\beta(m,M)} P_0(\theta_0 \in \tilde{C}_n(L)) \to 1,$$

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

*Proof.* The proof of the theorem is deferred to Section 2.9. □

## §2.2 Numerical analysis

In this section we investigate the numerical properties of the Gaussian process prior with (approximately) squared exponential covariance kernel. First we consider the Gaussian white noise model and the prior (2.1.2). 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 the (from practical point of view) more relevant non-parametric regression and classification models, where we also demonstrate the sub-optimal 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 poses good recovery and uncertainty quantification properties, their run times are substantially slower than using squared exponential kernels.

### §2.2.1 Gaussian white noise model

First we demonstrate the sub-optimal performance of the Gaussian process with (approximately) squared exponential covariance kernel (2.1.2) compared to modified versions of the empirical Bayes procedure and to the Gaussian process prior with polynomially decaying variances in the series representation, see (Knapik et al., 2016) and (Szabo et al., 2015). Let us consider the function $\theta_1 \in L_2[0,1]$ given by their Fourier coefficients $\theta_{1,i} = i^{-3/2}\sin(i)$, for $i = 1, 2, ...$, 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 (2.1.8), it has essentially the same behavior. In Figure 2.1 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 $L_2$-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 of the case where the choice $L_n = \log n$ was pre-specified). 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 $\theta \sim \otimes_{i=1}^{\infty} N(0, i^{-1-2\alpha})$, with hyper-parameter $\alpha$ estimated by the MMLE. One can see that the standard marginal likelihood empirical Bayes method provides too narrow credible sets failing to cover the underlying true function. Also note that both modifications of the empirical Bayes credible sets and using 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 more informative uncertainty statement (i.e. smaller credible sets).

### §2.2.2 Non-parametric regression and classification

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

In the non-parametric regression model we assume to observe pairs of random variables $(X_1, Y_1), (X_2, Y_2), ...(X_n, Y_n)$, where

$$Y_i = \theta_0(X_i) + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad X_i \overset{iid}{\sim} U(0,1), \quad i = 1, ..., n,$$

and the aim is to estimate the unknown non-parametric regression function $\theta_0$. In the Bayesian approach we endow $\theta_0$ with a Gaussian process 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 $\theta_2$ to be $\theta_{2,i} = i^{-3/2}\cos(i)$, $i = 1, 2, ....$ 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 point-wise credible

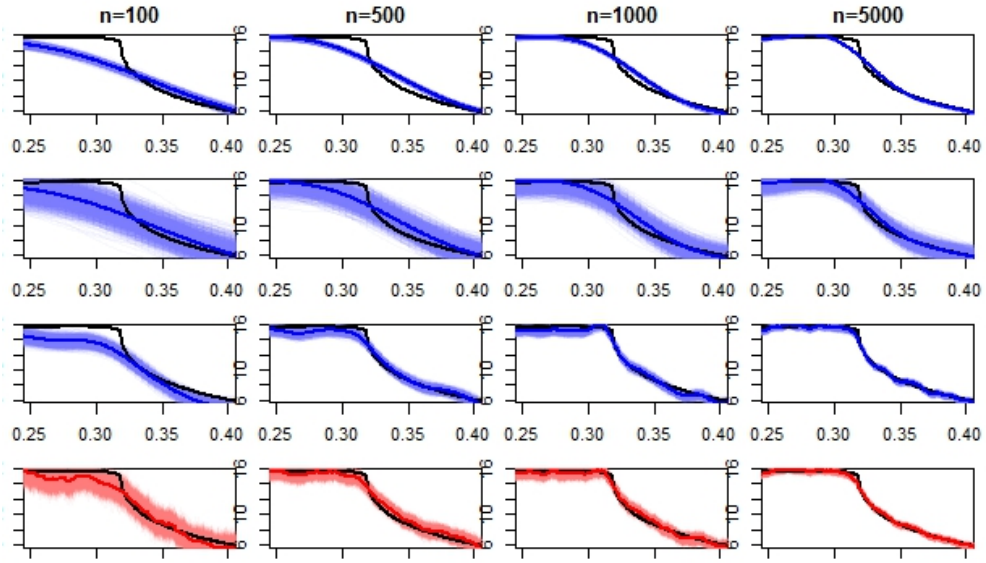


Figure 2.1: Empirical Bayes credible sets for the function $\theta_1$ (drawn in black) zoomed in to the interval $x \in [0.25, 0.4]$. First line: credible set (in light blue) and posterior mean (blue curve) corresponding to the prior with exponentially decaying variance. Second line: credible set (in 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 line: credible set (in light blue) and posterior mean (blue curve) corresponding to the prior with exponentially decaying variance and modified empirical Bayes procedure (rescaling factor multiplied by $\log n$). Last line: 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 $n = 100, 500, 1000, 5000$.

intervals (dashed blue) $[\hat{\theta}(x) - q_{0.025}\sqrt{\hat{c}(x,x)}, \hat{\theta}(x) + q_{0.025}\sqrt{\hat{c}(x,x)}]$, where $\hat{\theta}$ is the posterior mean, $q_\alpha$ the $\alpha$-th quantile of the standard normal distribution and $\hat{c}(.,.)$ 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 $\log n$), 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 size to be $n = 100, 500, 1000,$ and 2000. Observe in Figure 2.2 that the standard MMLE empirical Bayes method provides unreliable uncertainty quantification in 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 hyper-parameter) 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 computation time using the squared exponential kernel is much smaller than working with the Matérn kernel. We note that the computational times corresponding to the Matérn kernel are higher than for the squared exponential kernel. Estimating the regularity hyper-parameter 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 $\alpha$ denotes the regularity of the prior, see for instance Szabo et al. (2013). Therefore, we choose $\alpha$ large enough ($\alpha = 10$), which then seemingly slows down the computations. The different running times can be found in Table 2.3. The running time is based on the time spent computing the MMLE, the posterior mean and the point-wise posterior variance using the MatLab package "gpml" in a personal computing environment.

We also investigate empirically the frequentist coverage probabilities of the point-wise 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 $0.3188 = \text{argmax}_{x\in[0,1]}\theta_2(x)$) is included in the credible interval, see Table 2.1. Moreover, Table 2.2 shows the average size of the point-wise 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 similar behavior to what we have described above.

Note that Table 2.1 does not quite illustrate the results of Section 2.1.2 since the table shows the point-wise credible intervals whereas most of our theoretical results concern the $L_2$ credible balls. However, they still give an indication of the reliability of Bayesian uncertainty quantification. The point $x = 0.3188$, at which the maximum of $\theta_2$ is achieved, is seen as one of the clearest way to illustrate our negative results, whereas our positive results could be accepted only if the probability of $\theta_2(x)$ being inside of the corresponding credible interval goes to one for all point $x \in [0, 1]$.

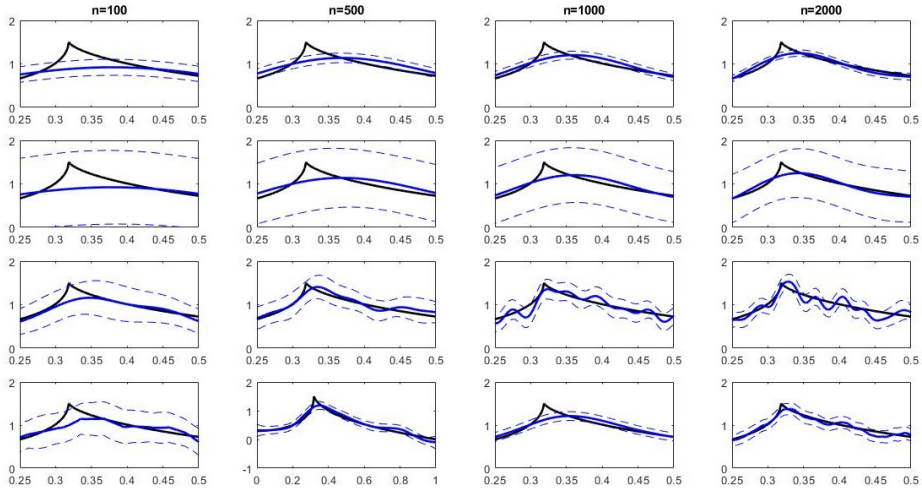


Figure 2.2: Empirical Bayes credible sets for the regression function $\theta_2$ (drawn in black), zoomed in to the interval $x \in [0.25, 0.5]$. The posterior means are drawn by solid blue line, while the 95% point-wise credible sets by dashed blue curves. In the first row we plot the MMLE empirical Bayes method, in the second row the MMLE empirical Bayes method with a $\log n$ blow up factor, the third row the modified MMLE empirical Bayes method using squared exponential Gaussian process prior, while in the fourth row we plot the empirical Byes credible sets using a Matérn kernel with data-driven choice for the regularity hyper-parameter. From left to right the sample size is $n = 100, 500, 1000, 2000$.

Next we consider the non-parametric classification problem. Let us assume that

| | $x = 0.25$ | | | $x = 0.3188$ | | | $x = 0.75$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $n =$ | 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 2.1: Frequencies that $\theta_2(x)$ is inside of 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: SE kernel MMLE empirical Bayes procedure, Method 2: SE kernel empirical Bayes procedure with $\log n$ blow up factor, Method 3: SE 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 2.2: Average size of the pointwise credible intervals (i.e. $2q_{0.025}\sqrt{\hat{c}(x,x)}$) for $\theta_2(x)$ in the regression model. Method 1: SE kernel MMLE empirical Bayes procedure, Method 2: SE kernel empirical Bayes procedure with $\log n$ blow up factor, Method 3: SE 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 | 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 |

Table 2.3: Average run time of the EB methods for $\theta_2$ in the regression model. Method 1: SE covariance kernel, Method 4: Matérn covariance kernel and MMLE for the regularity hyper-parameter, Method 5: Matérn covariance kernel and MMLE for the scaling hyper-parameter with fixed regularity $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000, 5000$

we observe the binary random variables $Y_1, Y_2, ..., Y_n \in \{0, 1\}$, with

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

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

We design similar experiments for the non-parametric classification model as for the non-parametric regression model above, with sample sizes $n = 100, 500, 1000$ and 2000 and the same $\theta_2$ as above. We plot the point-wise credible intervals for $\theta_2$ 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 2.3. 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 quantification. 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 $\theta_2(x)$ at points $x = (0.25, 0.3188, 0.75)$ in Table 2.4 and the computation time for different methods in Table 2.6, underlying the conclusions drawn from the figures above. Note that, similarly to Table 5.2, Table 2.4 does not quite illustrate our theoretical results, but is linked to it in a similar fashion as Table 5.2. Moreover, Table 2.5 shows the size of the average credible interval.

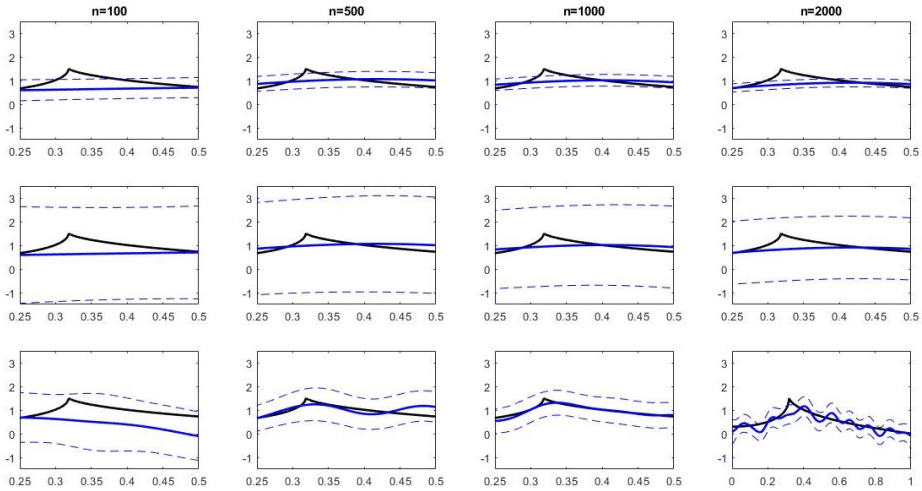


Figure 2.3: Empirical Bayes credible sets using squared exponential Gaussian process priors in the classification model for the function $\theta_2$ (drawn in black). The posterior means are drawn by solid blue line, while the 95% point-wise credible intervals by dashed blue curves. In the first row we plotted the MMLE empirical Bayes method, in the second row the MMLE empirical Bayes method with a $\log n$ blow up factor, while in the the third row the modified MMLE empirical Bayes method. From left to right the sample size is $n = 100, 500, 1000, 2000$.

## §2.3 Discussion

We have shown that the MMLE empirical Bayes method for Gaussian process prior with (a slightly modified version of the) squared exponential covariance kernel produces misleading uncertainty statement in context of the Gaussian white noise model. The derived negative results were demonstrated on a simulation study in context of the Gaussian white noise model and extended to the non-parametric 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 Gaussian

| | $x = 0.25$ | | | $x = 0.3188$ | | | $x = 0.75$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $n =$ | 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 2.4: Frequencies that $\theta_2(x)$ is inside of the corresponding credible interval for squared exponential and Matérn Gaussian process prior in the logistic regression model. Method 1: SE kernel MMLE empirical Bayes procedure, Method 2: SE kernel empirical Bayes procedure with $\log n$ blow up factor, Method 3: SE 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 $n = 100, 200, 500$.

| | 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 2.5: Average size of the pointwise credible intervals $2q_{0.025}\sqrt{\hat{c}(x,x)}$ for $\theta_2$ in the logistic regression model. The methods and the sample sizes are the same as in Table 2.4.

| $n =$ | 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 |

Table 2.6: Average run time of the EB methods for $\theta_2$ in the logistic regression model. Method 1: SE covariance kernel, Method 4: Matérn covariance kernel and MMLE for the regularity hyper-parameter, Method 5: Matérn covariance kernel and MMLE for the scaling hyper-parameter with fixed regularity $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000, 5000$

processes for uncertainty quantification as typically they provide misleading confidence statements, due to over-smoothing behavior of the MMLE. We note that the bad performance of the prior (2.1.2) 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 sub-optimally as well.

One can compensate 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 to modify the MMLE by multiplying it with $\log n$ to compensate for the over-smoothing. 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, like the Matérn kernel, however, these procedures can be computationally less appealing, as demonstrated in the simulation study.

## §2.4 Some properties of the MMLE

### §2.4.1 Deterministic bounds

As a first step we provide deterministic bounds for the marginal maximum likelihood estimator $\hat{a}_n$ of the rescaling hyper-parameter $a$. Let us introduce the following functions for $a \in [1, \infty)$:

$$h_n(a, \theta_0) := \frac{1}{\log^2\left(\frac{n}{a}\right)} \sum_{i=1}^{\infty} \frac{n^2 i e^{i/a} \theta_{0,i}^2}{a(ae^{i/a} + n)^2}, \tag{2.4.1}$$

$$g_n(a, \theta_0) := \frac{1}{\log^2\left(\frac{n}{a}\right)} \sum_{i=2a}^{\infty} \frac{n^2 (i-a) e^{i/a} \theta_{0,i}^2}{a(ae^{i/a} + n)^2}. \tag{2.4.2}$$

These functions are derived from the expected value of the score function, see Section 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

$$\begin{aligned} \underline{a}_n &:= \sup\{a \in [1, A_n] : g_n(a, \theta_0) \geq B \log n\}, \\ \overline{a}_n &:= \sup\{a \in [K_0, A_n] : h_n(a, \theta_0) \geq b\}, \end{aligned} \tag{2.4.3}$$

with some $b, B, K_0 > 0$ to be specified later and $A_n = o(n)$ given in (2.1.5). Then we show that these bounds sandwich $\hat{a}_n$ with high probability.

**Theorem 2.4.1.** *The MMLE $\hat{a}_n$ satisfies*

$$\inf_{\theta_0 \in \ell_2(M)} P_0(\underline{a}_n \leq \hat{a}_n \leq \overline{a}_n) \to 1, \tag{2.4.4}$$

*for $\underline{a}_n, \overline{a}_n$ defined in* (2.4.3).

*Proof.* See Section □

We also derive upper bounds for $\overline{a}_n$, in the case the true function belongs to the hyper-rectangle with regularity hyper-parameter $\beta$ or or to the analytic function class $A^\gamma$ and a lower bound for $\underline{a}_n$ in the case of self similar functions $\theta_0 \in \Theta^\beta(m, M)$.

**Proposition 2.4.2.** *For every $\beta \geq \beta_0$ and $\gamma > 0$ there exist $C_{\beta,b,M}, C_{\gamma,b,M} > 0$ such that*

$$\begin{aligned} \sup_{\theta_0 \in \Theta^\beta(M)} \overline{a}_n &\leq C_{\beta,b,M} n^{1/(2\beta+1)} (\log n)^{-1-1/(2\beta+1)}, \\ \sup_{\theta_0 \in A^\gamma(M)} \overline{a}_n &\leq C_{\gamma,b,M}, \\ \inf_{\theta_0 \in \Theta^\beta(m,M)} \underline{a}_n &\geq C_{\beta,B,m} n^{1/(2\beta+1)} (\log n)^{-1-2/(2\beta+1)}. \end{aligned}$$

*Proof.* Let us start with the proof of the first inequality. We show that for any $b > 0$ the inequality $h_n(a, \theta_0) < b$ holds for $a \geq C_{\beta,b,M} n^{1/(2\beta+1)} (\log n)^{-1-1/(2\beta+1)}$. 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

$$\begin{aligned} h_n(a,\theta_0) &\leq \frac{M}{\log^2\left(\frac{n}{a}\right)}\left(\frac{1}{a}\sum_{i=1}^{I_a} e^{i/a} i^{-2\beta} + \frac{n^2}{a^3}\sum_{i>I_a} e^{-i/a} i^{-2\beta}\right) \\ &\leq C_\beta \frac{M}{\log^2\left(\frac{n}{a}\right)}\left(I_a^{-2\beta} e^{I_a/a} + \frac{n^2}{a^2} I_a^{-2\beta} e^{-I_a/a}\right) \\ &\leq 2C_\beta M a^{-1-2\beta} n \left(\log\left(\frac{n}{a}\right)\right)^{-2-2\beta}, \end{aligned}$$

for some constant $C_\beta > 0$ depending only on $\beta$. For any $a > 0$ such that $A_n \geq a \geq Kn^{1/(2\beta+1)}(\log n)^{-1-1/(2\beta+1)}$ the preceding display is bounded by a multiple of $2C_\beta M K^{-1-2\beta}$. Then for sufficiently large choice of the constant $K = C_{\beta,b,M}$ (depending only on $\beta, b$ and $M$), we get that $h_n(a,\theta_0) < b$ for any $a$ larger than $C_{\beta,b,M} n^{1/(1+2\beta)}(\log n)^{-1-1/(2\beta+1)}$.

The proof of the second inequality of the statement goes similarly, i.e. we prove that for $a \geq C_{\gamma,b,M}$ we have $h_n(a,\theta_0) < b$. Note that by the sum of geometric series we get for every $a \geq 1/\gamma$

$$\begin{aligned} h_n(a,\theta_0) &\leq \frac{M}{\log^2\left(\frac{n}{a}\right)}\left(\frac{1}{a}\sum_{i=1}^{I_a} i e^{i/a} e^{-2\gamma i} + \frac{n^2}{a^3}\sum_{i>I_a} i e^{-i/a} e^{-2\gamma i}\right) \\ &\leq \frac{M}{a\log^2\left(\frac{n}{a}\right)}\sum_{i=1}^{\infty} i e^{-\gamma i} \leq \frac{M}{a(1-e^{-\gamma})^2\log^2\left(\frac{n}{a}\right)}, \end{aligned}$$

which is bounded from above by arbitrarily small $b$ for 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

$$\begin{aligned} g_n(a,\theta_0) &\geq \frac{m}{4\log^2\left(\frac{n}{a}\right)}\frac{n^2}{a^3}\sum_{i>I_a}(i-a)e^{-i/a} i^{-2\beta} \\ &\geq c_\beta \frac{m}{\log^2\left(\frac{n}{a}\right)}\frac{n^2}{a^2} I_a^{-2\beta} e^{-I_a/a} \\ &= c_\beta m a^{-1-2\beta} n\big(\log(n/a)\big)^{-2-2\beta}, \end{aligned}$$

for some $c_\beta > 0$ depending only on $\beta$. For $1 \leq a \leq Kn^{1/(2\beta+1)}(\log n)^{-1-2/(2\beta+1)}$ the preceding display is bounded by a multiple of $mc_\beta K^{-1-2\beta}\log n$. Then for sufficiently small choice of the constant $C_{\beta,B,m}$, we get that $g_n(a,\theta_0) \geq B\log n$ for any $a \leq C_{\beta,B,m} n^{1/(2\beta+1)}(\log n)^{-1-2/(2\beta+1)}$. □

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 2.4.3.** *For every $L_0, \rho, N_0 \geq 1$ we have*

$$\sup_{\theta_0 \in \Theta_{pt}(L_0, N_0, \rho)} \frac{\overline{a}_n \log\left(\frac{n}{\overline{a}_n}\right)}{\underline{a}_n \log\left(\frac{n}{\underline{a}_n}\right)} \leq K \log^2 n,$$

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

*Proof.* First of all note that since $\underline{a}_n \leq \overline{a}_n$, there is nothing to prove in the trivial cases $\underline{a}_n = A_n$ or $\overline{a}_n = K_0$. Hence $h_n(\overline{a}_n, \theta_0) \leq b$ and $g_n(a, \theta_0) < B \log n$, for all $a > \underline{a}_n$, hold. Furthermore assume that $\underline{a}_n \leq \rho^{-2}\overline{a}_n$, else the statement is trivial.

Let us divide the interval $[\rho^j, \rho^{j+1})$ into sub-intervals $[\rho^{j+\frac{k}{\lceil \log n \rceil}}, \rho^{j+\frac{k+1}{\lceil \log n \rceil}})$, $k = 0, 1, ...\lceil \log n \rceil - 1$, and introduce the notation

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

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

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

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

$$I_{\tilde{a}} \equiv \tilde{a} \log(n/\tilde{a}) \in \left[\rho^{j+\frac{k_j}{\lceil \log n \rceil}}, \rho^{j+\frac{k_j+1}{\lceil \log n \rceil}}\right) \tag{2.4.5}$$

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

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

Let us take any $a_1 \leq \rho^{-2} a_2$ and denote by $\tilde{a}_1 \in (a_1, \rho^2 a_1)$ the value satisfying (2.4.5). Then in view of $\exp\{e^{1/\log n} \log(n/a)\} \leq \exp\{(1+2/\log n)\log(n/a)\} \leq e^2 n/a$, for $n \geq e$, combined with the previous inequalities we get that the ratio $h_n(a_2, \theta_0)/h_n(\tilde{a}_1, \theta_0)$ is bounded from above by

$$\begin{aligned}
&\frac{\tilde{a}_1 \log^2\left(\frac{n}{\tilde{a}_1}\right)}{a_2 \log^2\left(\frac{n}{a_2}\right)} 4e^2 \frac{\sum_{i=1}^{I_{\tilde{a}_1}} i e^{i/a_2} \theta_{0,i}^2 + \sum_{i=I_{\tilde{a}_1}}^{I_{a_2}} i e^{i/a_2} \theta_{0,i}^2 + \frac{n^2}{a_2^2} \sum_{i=I_{a_2}}^{\infty} i e^{-i/a_2} \theta_{0,i}^2}{\sum_{i=1}^{e^{1/\log n} I_{\tilde{a}_1}} i e^{i/\tilde{a}_1} \theta_{0,i}^2} \\
&\leq \frac{\tilde{a}_1 \log^2\left(\frac{n}{\tilde{a}_1}\right)}{a_2 \log^2\left(\frac{n}{a_2}\right)} 4e^2 \left(1 + \frac{\sum_{i=I_{\tilde{a}_1}}^{I_{a_2}} i e^{i/a_2} \theta_{0,i}^2 + n \log\left(\frac{n}{a_2}\right) \sum_{i=I_{a_2}}^{\infty} \theta_{0,i}^2}{\sum_{i=e^{-1/\log n} I_{\tilde{a}_1}}^{e^{1/\log n} I_{\tilde{a}_1}} i e^{i/\tilde{a}_1} \theta_{0,i}^2}\right).
\end{aligned}$$

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

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

Moreover, since

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

combined with the preceding computations we get that

$$\frac{h_n(a_2,\theta_0)}{h_n(\tilde{a}_1,\theta_0)} \leq 4e\frac{\tilde{a}_1\log^2\left(\frac{n}{\tilde{a}_1}\right)}{a_2\log^2\left(\frac{n}{a_2}\right)}\left(1 + L_0e^2\frac{\log\left(\frac{n}{a_2}\right)}{\log\left(\frac{n}{\tilde{a}_1}\right)}\log n\right). \tag{2.4.6}$$

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

$$h_n(a,\theta_0) \leq 2g_n(a,\theta_0) + \frac{2e^2}{\log^2\left(\frac{n}{a}\right)}\sum_{i=1}^{2a}\theta_{0,i}^2 \leq 2B\log(n) + o(1).$$

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 (2.4.6) we get that

$$\begin{aligned}\frac{b}{2B\log(n)+o(1)} \leq \frac{h_n(\overline{a}_n,\theta_0)}{h_n(\tilde{a}_1,\theta_0)} &\leq 4e^4(1+o(1))L_0\log(n)\frac{\tilde{a}_1\log(\frac{n}{\tilde{a}_1})}{\overline{a}_n\log(\frac{n}{\overline{a}_n})} \\ &\leq 4e^4\rho^2(1+o(1))L_0\log(n)\frac{\underline{a}_n\log(\frac{n}{\underline{a}_n})}{\overline{a}_n\log(\frac{n}{\overline{a}_n})}.\end{aligned}$$

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

## §2.4.2 Contraction rates

In this section we provide the contraction rate results both for 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 hyper-parameter $\beta > 0$.

**Theorem 2.4.4.** *The maximum marginal likelihood empirical Bayes posterior corresponding to the prior* (2.1.2) *achieves the minimax adaptive contraction rate (up to a logarithmic factor), i.e. for given $M, \beta_0 > 0$ we have*

$$\sup_{\beta\geq\beta_0}\sup_{\theta\in\Theta^\beta(M)} E_0\left[\Pi_{\hat{a}_n}(\|\theta-\theta_0\|_2 \geq M_n\left(\frac{n}{\log^2 n}\right)^{-\beta/(2\beta+1)}|Y)\right] \to 0, \tag{2.4.7}$$

*for any sequence $M_n$ tending to infinity.*

*Proof.* See Section 2.4.3. □

Using our findings on the empirical Bayes method we can extend the results on the hierarchical Bayes method, derived in (van der Vaart and van Zanten, 2009a) and (Bhattacharya and Pati, 2015) (where typically inverse gamma hyper-prior was considered), by allowing other, more general choices of the hyper-prior distribution as well.

**Theorem 2.4.5.** *Let us assume that the hyper-prior $\pi$ satisfies Assumption 2.1.1. 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*

$$\sup_{\beta\geq\beta_0}\sup_{\theta\in\Theta^\beta(M)} E_0\left[\Pi(\|\theta-\theta_0\|_2 \geq M_n\left(\frac{n}{\log^2 n}\right)^{-\beta/(2\beta+1)}|Y)\right]\to 0, \tag{2.4.8}$$

*for some arbitrary sequence $M_n$ tending to infinity.*

*Proof.* See Section 2.10.1. □

## §2.4.3 Proof of Theorem 2.4.4

Let us introduce the shorthand notation

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

In view of Markov's inequality and Theorem 2.4.1, for every $\beta > 0$

$$\sup_{\theta_0\in\Theta^\beta(M)} E_0[\Pi_{\hat{a}_n}(\|\theta-\theta_0\|_2 \geq M_n\varepsilon_n|Y)] \leq \frac{1}{M_n^2\varepsilon_n^2}\sup_{\theta_0\in\Theta^\beta(M)} E_0[\sup_{a\in[\underline{a}_n,\overline{a}_n]} R_n(a)] + o(1), \tag{2.4.9}$$

where

$$R_n(a) = \int \|\theta-\theta_0\|_2^2\Pi_a(d\theta|Y)$$

is the posterior risk. We show below that both

$$\sup_{\theta_0\in\Theta^\beta(M)}\sup_{a\in[\underline{a}_n,\overline{a}_n]} E_0[R_n(a)] = O(\varepsilon_n^2) \qquad \text{and} \tag{2.4.10}$$

$$\sup_{\theta_0\in\Theta^\beta(M)} E_0[\sup_{a\in[\underline{a}_n,\overline{a}_n]} |R_n(a) - E_0(R_n(a))|] = o(\varepsilon_n^2) \tag{2.4.11}$$

hold, which results in that the right-hand side of (2.4.9) vanishes as $n\to\infty$, concluding the proof of Theorem 2.4.4.

### 2.4.3.1 Bound for the expected posterior risk (2.4.10)

First, note that by elementary computations

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

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

$$E_0 R_n(a) = \sum_{i=1}^{\infty} \frac{a^2 e^{2i/a}}{(ae^{i/a}+n)^2}\theta_{0,i}^2 + \sum_{i=1}^{\infty} \frac{n}{(ae^{i/a}+n)^2} + \sum_{i=1}^{\infty} \frac{1}{ae^{i/a}+n}. \tag{2.4.12}$$

Note that the second and third terms do not contain $\theta_0$, and that the second term is bounded by the third. By Lemma 2.11.2 (with $r=0$ and $l=1$) and Proposition 2.4.2 the latter is further bounded for $a \le \overline{a}_n$ by a multiple of

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

since the function $a \mapsto a\log(n/a)$ is monotone increasing for $a \le n/e$. It remained to deal with the first term on the right hand side of (2.4.12), which we divide into three parts and show that each of the parts have the stated order. First note that for $\theta_0 \in \Theta^\beta(M)$

$$\begin{aligned}\sum_{i=(n/\log^2 n)^{1/(2\beta+1)}}^{\infty} \frac{a^2 e^{2i/a}}{(ae^{i/a}+n)^2}\theta_{0,i}^2 &\le \sum_{i=(n/\log^2 n)^{1/(2\beta+1)}}^{\infty} M i^{-1-2\beta} \\ &\le \frac{M}{2\beta}\left(\frac{n}{\log^2 n}\right)^{-2\beta/(2\beta+1)}.\end{aligned}$$

Next note that for $a \le \overline{a}_n$, in view of Proposition 2.4.2,

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

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 $n > 0$. Besides, for $a > \underline{a}_n$ we have $g_n(a, f_0) < B\log n$, hence for any $\underline{a}_n < a \le \bar{a}_n$

$$\begin{aligned}\sum_{i=2a}^{I_a} \frac{a^2 e^{2i/a}}{(ae^{i/a}+n)^2}\theta_{0,i}^2 &\le \frac{a}{n}\frac{\log^2\left(\frac{n}{a}\right)}{\left(\log\left(\frac{n}{a}\right)-1\right)} \sum_{i=2a}^{I_a} \frac{n^2 e^{i/a}(i-a)}{a\log^2\left(\frac{n}{a}\right)(ae^{i/a}+n)^2}\theta_{0,i}^2 \\ &\le 2\overline{a}_n n^{-1}\log\left(\frac{n}{\overline{a}_n}\right) g_n(a,\theta_0) \\ &\le 2\overline{a}_n n^{-1}\log^2 n \le 2C_{\beta,b,M} n^{-2\beta/(2\beta+1)} (\log n)^{2\beta/(2\beta+1)},\end{aligned}$$

where the last inequality follows from Proposition 2.4.2.

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

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

and let

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

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 $a \geq 1$

$$\begin{aligned}\sum_{i=I_{\underline{a}_n}}^{(n/\log^2 n)^{1/(2\beta+1)}} \frac{a^2 e^{2i/a}}{(ae^{i/a}+n)^2}\theta_{0,i}^2 &\leq \sum_{j=0}^{J-1}\sum_{i=n_j}^{n_{j+1}} \theta_{0,i}^2 \\ &\leq 4e^2 \sum_{j=0}^{J-1}\sum_{i=n_j}^{n_{j+1}} \frac{nb_j e^{i/b_j}}{(b_j e^{i/b_j}+n)^2}\theta_{0,i}^2. \qquad (2.4.13)\end{aligned}$$

By elementary computations we get that $b_j \asymp n_j/\log n_j$, therefore (2.4.13) is further bounded by constant times

$$\frac{1}{n}\sum_{j=0}^{J-1}\frac{1}{\log n_j}\sum_{i=n_j}^{n_{j+1}}\frac{n^2(i-b_j)e^{i/b_j}}{(b_j e^{i/b_j}+n)^2}\theta_{0,i}^2 \leq \frac{1}{n}\sum_{j=0}^{J-1}\frac{b_j \log^2 n}{\log n_j} g_n(b_j,\theta_0).$$

Since $b_j \geq \underline{a}_n$ we have $g_n(b_j,\theta_0) \leq B\log n$ for all $j \geq 0$. Then by the sum of geometric series we get that

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

concluding the proof of assertion (2.4.10).

#### 2.4.3.2 Bound for the centered posterior risk (2.4.11)

Note that

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

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

$$\begin{aligned}E_0\left(\sup_{a\in[\underline{a}_n,\overline{a}_n]}\frac{|\mathbb{V}(a)|}{n}\right) &\leq Kn^{-2\beta/(2\beta+1)}(\log n)^{-1/(2\beta+1)}, \\ E_0\left(\sup_{a\in[\underline{a}_n,\overline{a}_n]}\frac{|\mathbb{W}(a)|}{\sqrt{n}}\right) &\leq Kn^{-2\beta/(1+2\beta)}.\end{aligned}$$

We deal with the two processes above, separately.

For the process $\mathbb{V}$, Corollary 2.2.5 in (Van Der Vaart and Wellner, 1996) 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 $[\underline{a}_n,\overline{a}_n]$ and $N(\varepsilon,B,d_n)$ the covering number of the set $B$ with $\varepsilon$-radius balls relative to the $d_n$ semi-metric. 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 2.11.2 (with $r=0$ and $l=4$) we can conclude that the variance of $\mathbb{V}(a)$ is bounded from above by a multiple of $a\log(n/a)$, hence $diam_n\lesssim\sqrt{\overline{a}_n\log n}$. In view of Lemma 2.4.6, the distance $d_n(a_1,a_2)$ is bounded from above by a multiple of $|a_1-a_2|\log^{3/2}n$, hence the interval $[\underline{a}_n,\overline{a}_n]$ can be covered with constant times $\overline{a}_n\varepsilon^{-1}\log^{3/2}n$ amount of $\varepsilon$-balls relative to the $d_n$ semi-metric. In view of the above computation and Proposition 2.4.2

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

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 details. First note that

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

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

$$n\sum_{i=1}^{I_a}\frac{a^2e^{2i/a}}{(ae^{i/a}+n)^4}\theta_{0,i}^2\le\frac{1}{n^3}\sum_{i=1}^{I_a}a^2e^{2i/a}\theta_{0,i}^2\le\frac{\|\theta_0\|_2^2}{n}.$$

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 $g_n(a,\theta_0)\le B\log n$, for all $a\ge\underline{a}_n$, and in view of Proposition 2.4.2, for $a\le\overline{a}_N$

$$\begin{aligned}n\sum_{i=I_a}^{\infty}\frac{a^2e^{2i/a}}{(ae^{i/a}+n)^4}\theta_{0,i}^2&\le\frac{a\log^2\left(\frac{n}{a}\right)}{4n^2\left(\log\left(\frac{n}{a}\right)-1\right)}\sum_{i=I_a}^{\infty}\frac{n^2(i-a)e^{i/a}}{a\log^2\left(\frac{n}{a}\right)(ae^{i/a}+n)^2}\theta_{0,i}^2\\&\le an^{-2}\log\left(\frac{n}{a}\right)g_n(a,\theta_0)\le B\overline{a}_nn^{-2}\log^2n\\&\le 2BC_{\beta,b,M}n^{\frac{-(4\beta+1)}{2\beta+1}}(\log n)^{2\beta/(2\beta+1)},\end{aligned}$$

hence $\mathrm{diam}_n=O(n^{-\frac{1/2+2\beta}{1+2\beta}}(\log n)^{\beta/(1+2\beta)})$. Then in view of Lemma 2.4.6 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 semi-metric $d_n(a_1,a_2)=V_0\big(\mathbb{W}(a_1)/\sqrt{n}-\mathbb{W}(a_2)/\sqrt{n}\big)$ and the rest of the proof goes as above.

#### 2.4.3.3 Bounds for the semi-metrics associated to $\mathbb{V}$ and $\mathbb{W}$

**Lemma 2.4.6.** *For any $1 \le a_1 \le a_2$ and $f_0 \in \ell_2(M)$ we have*

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

*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 $\phi_i$ is given by $\phi_i'(a)^2 = 4\phi_i(a)^3 e^{2i/a}(i-a)^2/a^2$, hence in view of Lemma 2.11.3 the preceding display is bounded above by a multiple of

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

with the help of Lemma 2.11.1 (first with $m = 2$ and then with $m = 0$), and Lemma 2.11.2 (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

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

with $\phi_i(a) = ae^{i/a}/(ae^{i/a} + n)^2$. Note that $|\phi_i'(a)| \le (i + a)a^{-2}\phi_i(a)$, hence in view of Lemma 2.11.1 (first with $m = 2$ and then with $m = 0$) and Lemma 2.11.3 the preceding display is bounded by

$$4(a_1 - a_2)^2 n^2 \sup_{a \in [a_1, a_2]} \frac{1}{a^2} \sum_{i=1}^{\infty} \frac{e^{2i/a}(\frac{i^2}{a^2} + 1)}{(ae^{i/a} + n)^4} \theta_{0,i}^2$$
$$\le 4(a_1 - a_2)^2 \sup_{a \in [a_1, a_2]} \frac{1}{a^2} \left(\log^2\left(\frac{n}{a}\right) + 1\right) \|\theta_0\|_2^2,$$

concluding the proof of the lemma. □

## §2.5 Proof of the empirical Bayes part of Theorem 2.1.5

First note that we get the empirical Bayes credible set by plugging in the estimator $\hat{a}_n$ into the credible ball $\hat{C}_{a,\alpha}$ defined as

$$\hat{C}_{a,\alpha} = \{\theta \in L_2 : \|\theta - \hat{\theta}_a\|_2 \le L r_\alpha\}$$

satisfying that

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

where $\hat{\theta}_a$ is the posterior mean for fixed hyper-parameter $a > 0$.

The proof of the statement is then based on the deterministic bounds for the MMLE $\hat{a}_n$ derived in Theorem 2.4.1 and their distance investigated in Lemma 2.4.3.

Note that $\theta_0 \in \hat{C}_n(L\log^{3/2} n)$ if and only if $\|\theta_0 - \hat{\theta}_{\hat{a}_n}\|_2 \le L(\log n)^{3/2} r_\alpha$. Therefore by triangle inequality it is sufficient to verify that

$$\|W(\hat{a}_n)\|_2 \le L(\log n)^{3/2} r_\alpha(\hat{a}_n) - \|B(\hat{a}_n,\theta_0)\|_2 \tag{2.5.1}$$

holds with high probability, where $W(a) = \hat{\theta}_a - E_0\hat{\theta}_a$ and $B(a,\theta_0) = E_0\hat{\theta}_a - \theta_0$ are the centered posterior mean and the bias of the posterior mean for fixed hyper-parameter $a > 0$, respectively. Note that the $i$th coefficient of these vectors take the form

$$W_i(a) = \frac{n(Y_i - \theta_{0,i})}{ae^{i/a}+n}, \quad \text{and} \quad B_i(a,\theta_0) = \frac{ae^{i/a}\theta_{0,i}}{ae^{i/a}+n}.$$

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

$$\inf_{\underline{a}_n \le a \le \overline{a}_n} r_\alpha^2(a) \ge \frac{\underline{a}_n}{3n}\log\left(\frac{n}{\underline{a}_n}\right), \tag{2.5.2}$$

$$\inf_{\theta_0 \in \Theta_{pt}(L_0,N_0,\rho)} P_0\left(\sup_{\underline{a}_n \le a \le \overline{a}_n} \|W(a)\|_2^2 \le C_1 \frac{\underline{a}_n}{n}\log\left(\frac{n}{\underline{a}_n}\right)\log^2 n\right) \to 1, \tag{2.5.3}$$

$$\sup_{\theta_0 \in \Theta_{pt}(L_0,N_0,\rho)} \sup_{\underline{a}_n \le a \le \underline{a}_n} \|B(a,\theta_0)\|_2^2 \le C_2 \frac{\underline{a}_n}{n}\log\left(\frac{n}{\underline{a}_n}\right)\log^3 n. \tag{2.5.4}$$

Hence in view of Theorem 2.4.1 assertion (2.5.1) holds with probability tending to one for large enough choice of $L$, under the polished tail assumption.

Proof of (2.5.2): The radius $r_\alpha(a)$, given in (2.1.6), 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 $Z_i$'s are iid $N(0,1)$. We show below that

$$\liminf_{n\to\infty} \inf_{a\in[\underline{a}_n,\overline{a}_n]} E\left[\frac{nU_n(a)}{a\log\left(\frac{n}{a}\right)}\right] > \frac{1}{2}, \tag{2.5.5}$$

$$E\left[\sup_{a\in[\underline{a}_n,\overline{a}_n]} \frac{n|U_n(a) - E[U_n(a)]|}{a\log\left(\frac{n}{a}\right)}\right] \to 0. \tag{2.5.6}$$

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

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

hence (2.5.2) follows from the definition of $r_\alpha(a)$.

Assertion (2.5.5) follows as

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

To verify (2.5.6), it suffices by Corollary 2.2.5 in (Van Der Vaart and Wellner, 1996) (applied with $\psi(x) = x^2$) to show that there exist $K_1, K_2 > 0$ such that for any $a \in [\underline{a}_n, \overline{a}_n]$

$$V\left(\frac{nU_n(a)}{a\log\left(\frac{n}{a}\right)}\right) \leq K_1 \frac{1}{a\log\left(\frac{n}{a}\right)}, \tag{2.5.7}$$

$$\int_0^{diam_n} \sqrt{N(\varepsilon, [\underline{a}_n, \overline{a}_n], d_n)}d\varepsilon \leq \sqrt{A_n/n} = o(1), \tag{2.5.8}$$

where $d_n$ is the semi-metric defined by $d_n^2(a_1, a_2) := V\left(\frac{nU_n(a_1)}{a_1\log(n/a_1)} - \frac{nU_n(a_2)}{a_2\log(n/a_2)}\right)$, $diam_n$ 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 $\varepsilon$ needed to cover the set $S$.

First note that in view of Lemma 2.11.2 (with $r = 0$ and $l = 2$) we have

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

As a consequence one can see that $diam_n \lesssim (\underline{a}_n \log(n/\underline{a}_n))^{-1/2}$. By Lemma 2.5.1, $d_n(a_1, a_2) \lesssim a_1^{-3/2}\log^{1/2}(n/a_1)n^{-1}|a_1 - a_2|$, hence

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

Therefore one can conclude that

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

Proof of (2.5.3): 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

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

Furthermore note that by applying Lemma 2.11.2 (with $r = 0$ and $l = 2$) we get

$$\frac{a}{n}\log\left(\frac{n}{a}\right) \leq \frac{4I_a n}{(ae^{I_a/a}+n)^2} \leq \sum_{i=1}^{I_a} \frac{4n}{(ae^{i/a}+n)^2} \leq \sum_{i=1}^{\infty} \frac{4n}{(ae^{i/a}+n)^2} \leq C\frac{a}{n}\log\left(\frac{n}{a}\right),$$

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

$$\inf_{\theta_0 \in \ell_2(M)} \left( \sup_{\underline{a}_n \le a \le \overline{a}_n} \|W(a)\|_2^2 \le C_2 \frac{\overline{a}_n}{n} \log\left(\frac{n}{\overline{a}_n}\right) \right) \xrightarrow{P_0} 1. \tag{2.5.9}$$

Then in view of Lemma 2.4.3, the right hand side of the inequality in the preceding probability statement is further bounded from above by constant multiplier of $(\underline{a}_n/n)\log(n/\underline{a}_n)\log^2 n$.

Proof of (2.5.4): First note that

$$\|B(a,\theta_0)\|_2^2 \le \sum_{i=1}^{I_a} n^{-2}a^2 e^{2i/a}\theta_{0,i}^2 + \sum_{i=I_a}^{\infty} \theta_{0,i}^2.$$

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 $\sum_{i=1}^{\infty}\theta_{0,i}^2 < \infty$, and furthermore note the function $i \mapsto e^{i/a}/(i-a)$ is monotone increasing on the interval $[2a, I_a]$ hence it takes its maximum at $I_a$. Therefore in view of Lemma 2.4.3 the first part of the bias for functions satisfying the polished tail condition is bounded by

$$\begin{aligned}
\sup_{\underline{a}_n \le a \le \overline{a}_n} \sum_{i=1}^{I_a} \frac{a^2 e^{2i/a}\theta_{0,i}^2}{n^2} &\le \sup_{\underline{a}_n \le a \le \overline{a}_n} \sum_{i=1}^{2a} \frac{a^2 e^{2i/a}\theta_{0,i}^2}{n^2} + \sup_{\underline{a}_n \le a \le \overline{a}_n} \frac{a}{n} \frac{\log^2\left(\frac{n}{a}\right)}{\left(\log\left(\frac{n}{a}\right)-1\right)} g_n(a,\theta_0) \\
&\le \frac{e^4 \overline{a}_n^2}{n^2} \sum_{i=1}^{2a} \theta_{0,i}^2 + (B+o(1))\frac{\overline{a}_n}{n}\log\left(\frac{n}{\overline{a}_n}\right)\log n \\
&\le (B+o(1))\frac{\overline{a}_n}{n}\log\left(\frac{n}{\overline{a}_n}\right)\log n \\
&\le K_{\rho,L_0,B,b}\frac{\underline{a}_n}{n}\log\left(\frac{n}{\underline{a}_n}\right)\log^3 n,
\end{aligned}$$

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

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

for some $\tilde{a}_n \in [\underline{a}_n, \rho\underline{a}_n]$. Therefore, by using Lemma 2.4.3,

$$\begin{aligned}
\sum_{i=I_{\underline{a}_n}}^{\infty} \theta_{0,i}^2 &\lesssim \log\left(\frac{n}{\underline{a}_n}\right) \sum_{i=I_{\tilde{a}_n}}^{I_{\tilde{a}_n}+\rho\tilde{a}_n} \frac{n^2(i-\tilde{a}_n)e^{i/\tilde{a}_n}}{\tilde{a}_n \log^2\left(\frac{n}{\tilde{a}_n}\right)(\tilde{a}_n e^{i/\tilde{a}_n}+n)^2}\theta_{0,i}^2 \frac{\tilde{a}_n}{n}\log\left(\frac{n}{\tilde{a}_n}\right), \\
&\le \log\left(\frac{n}{\underline{a}_n}\right) g_n(\tilde{a}_n,\theta_0)\frac{\tilde{a}_n}{n}\log\left(\frac{n}{\tilde{a}_n}\right) \le K_{\rho,L_0,B,b}\log^2\left(\frac{n}{\underline{a}_n}\right)\frac{\underline{a}_n}{n}\log n,
\end{aligned}$$

for some large enough constant $K_{\rho,L_0,B,b} > 0$. Combining the two bounds, we see that (2.5.4) holds.

**Lemma 2.5.1.** *There exists a $K > 0$ such that for any $1 < a_1 < a_2$*

$$V\left(\frac{U_n(a_1)}{a_1 \log\left(\frac{n}{a_1}\right)} - \frac{U_n(a_2)}{a_2 \log\left(\frac{n}{a_2}\right)}\right) \leq K(a_1 - a_2)^2 \frac{\log\left(\frac{n}{a_1}\right)}{a_1^3 n^2}. \tag{2.5.10}$$

*Proof.* First note that

$$V\left(\frac{U_n(a_1)}{a_1 \log\left(\frac{n}{a_1}\right)} - \frac{U_n(a_2)}{a_2 \log\left(\frac{n}{a_2}\right)}\right) = 2\sum_{i=1}^{\infty}(\phi_i(a_1) - \phi_i(a_2))^2 \tag{2.5.11}$$

with $\phi_i(a) := \frac{1}{a\log(n/a)(ae^{i/a}+n)}$. The derivative of $\phi_i(a)$ is given as

$$\phi_i'(a) = \phi_i(a)\Big(\frac{2(i-a)e^{i/a}}{a(ae^{i/a}+n)} + \frac{1}{a\log(n/a)} - \frac{1}{a}\Big),$$

so we can see that $|\phi_i'(a)| \lesssim \big(\frac{(i+a)e^{i/a}}{a(ae^{i/a}+n)} \vee \frac{1}{a}\big)\phi_i(a)$. Thus in view of Lemma 2.11.3 the right hand side of (2.5.11) 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 2.11.1 (first with $m = 2$ and then with $m = 0$) and Lemma 2.11.2 (first with $r = 1$ and $l = 2$ and second with $r = 0$ and $l = 2$) the preceding display is further bounded by the right hand side of (2.5.10), finishing the proof of the statement. □

## §2.6 Proof of Theorem 2.1.4

We use the notations introduced in Section 2.5.

First recall that $\theta_0 \in \hat{C}_n(L)$ if and only if $\|\theta_0 - \hat{\theta}\|_2 \leq Lr_\alpha(\hat{a}_n)$. We show below that

$$\inf_{\theta_0\in A^\gamma(M)} P_0\left(\sup_{\underline{a}_n\leq a\leq \overline{a}_n} \|W(a)\|_2^2 \leq C_1 \frac{\underline{a}_n}{n}\log\left(\frac{n}{\underline{a}_n}\right)\right) \to 1, \tag{2.6.1}$$

$$\sup_{\theta_0\in A^\gamma(M)} \sup_{\underline{a}_n\leq a\leq \underline{a}_n} \|B(a, f_0)\|_2^2 \leq C_2 \frac{\underline{a}_n}{n}\log\left(\frac{n}{\underline{a}_n}\right), \tag{2.6.2}$$

for some constants $C_1, C_2 > 0$ depending only on $M$, which together with (2.5.2) and Theorem 2.4.1 results in the statement.

The proof of assertion (2.6.1) follows by combining (2.5.9) and the second inequality of Proposition 2.4.2. Next note that similarly to the proof of (2.5.4), we get that

$$\|B(a,\theta_0)\|_2^2 \leq \sum_{i=1}^{I_a} \frac{a^2 e^{2i/a}\theta_{0,i}^2}{n^2} + \sum_{i=I_a}^{\infty} \theta_{0,i}^2 \lesssim \frac{\overline{a}_n}{n}\log\left(\frac{n}{\overline{a}_n}\right) + \sum_{i=I_{\underline{a}_n}}^{\infty} \theta_{0,i}^2.$$

Furthermore

$$\sum_{i=I_{\underline{a}_n}}^{\infty} \theta_{0,i}^2 = \sum_{i=I_{\underline{a}_n}}^{\infty} e^{-2i\gamma} e^{2i\gamma} \theta_{0,i}^2 \le M e^{-2I_{\underline{a}_n}\gamma} = M \left(\frac{\underline{a}_n}{n}\right)^{2\underline{a}_n\gamma} \le M \frac{\underline{a}_n}{n} \log\left(\frac{n}{\underline{a}_n}\right)$$

for $\gamma \ge 1/2$, finishing the proof of (2.6.2) and concluding the proof of the theorem.

## §2.7 Proof of Theorem 2.1.3 and the empirical Bayes part of Corollary 2.1.2

In the proof we use again the notations introduced in Section 2.5.

First note that $\theta_0 \in \hat{C}_n(L_n)$ implies that $\|B(\hat{a}_n, \theta_0)\|_2 \le L_n r_\alpha(\hat{a}_n) + \|W(\hat{a}_n)\|_2$, which combined with Theorem 2.4.1 provides the upper bound

$$P_0(\theta_0 \in \hat{C}_n(L_n)) \le P_0\left(\inf_{a \le \overline{a}_n} \|B(a,\theta_0)\|_2 \le L_n \sup_{a \le \overline{a}_n} r_\alpha(a) + \sup_{a \le \overline{a}_n} \|W(a)\|_2\right) + o(1). \tag{2.7.1}$$

The proof of assertion (2.5.2) also shows that there exists constants $C_1 > 0$ such that

$$\sup_{a \le \overline{a}_n} r_\alpha^2(a) \le C_1 \frac{\overline{a}_n}{n} \log\left(\frac{n}{\overline{a}_n}\right). \tag{2.7.2}$$

Then in view of assertion (2.5.9) and Proposition 2.4.2, both the squared radius $r_\alpha(a)^2$ and the variance term $\|W(a)\|_2^2$ are bounded by$C_{\beta,b,M} n^{-2\beta/(2\beta+1)} (\log n)^{-1/(2\beta+1)}$, for some $C_{\beta,b,M} > 0$.

Since for $\theta_0 \in \Theta_s^\beta(m,M)$ we have $\|B(a,\theta_0)\|_2^2 = \sum\limits_{i=1}^{\infty} \frac{a^2 e^{2i/a} \theta_{0,i}^2}{(ae^{i/a}+n)^2}$ the bias is bounded from below by

$$\|B(a,\theta_0)\|_2^2 \ge m \sum_{i=I_a}^{\infty} i^{-1-2\beta} > \frac{m}{2\beta} I_a^{-2\beta} \ge \frac{m}{2\beta} a^{-2\beta} \log^{-2\beta}\left(\frac{n}{a}\right).$$

As the function $a \mapsto a^{-2\beta} \log^{-2\beta}\big(n/a\big)$ is monotone decreasing for $a \le A_n$, we see that $\inf\limits_{a \le \overline{a}_n} \|B(a,\theta_0)\|_2^2 \ge (m/(2\beta)) \overline{a}_n^{-2\beta} \log^{-2\beta}(n/\overline{a}_n)$. Hence in view of Proposition 2.4.2 the bias is bounded from below by $c_{m,\beta,b,B,M} n^{-2\beta/(2\beta+1)} (\log n)^{2\beta/(2\beta+1)}$, for some $c_{m,\beta,b,B,M} > 0$. Thus, the above inequalities imply that for arbitrary $\theta_0 \in \Theta_s^\beta(m,M)$ the right hand side of (2.7.1) is further bounded by

$$\sup_{\theta_0 \in \ell_2(M)} P_0\left(n^{-\beta/(2\beta+1)} (\log n)^{\beta/(2\beta+1)} \le L_n C n^{-\beta/(2\beta+1)} (\log n)^{-(1/2)/(2\beta+1)}\right) + o(1),$$

which goes to 0 for arbitrary $L_n = o(\sqrt{\log n})$ and $C$ depending on $m, \beta, b, B$ and $M$, concluding the proof of the theorem.

## §2.8 Proof of Theorem 2.4.1

First note that the derivative of the marginal likelihood function $\ell_n(a)$ is

$$\mathbb{M}_n(a) = \frac{1}{2}\left(\sum_{i=1}^{\infty} \frac{n^2 Y_i^2 e^{i/a}(i-a)}{a(ae^{i/a}+n)^2} - \sum_{i=1}^{\infty} \frac{n(i-a)}{a^2(ae^{i/a}+n)}\right), \tag{2.8.1}$$

with expected value

$$E_0[\mathbb{M}_n(a)] = \frac{1}{2}\left(\sum_{i=1}^{\infty} \frac{n^2(i-a)e^{i/a}\theta_{0,i}^2}{a(ae^{i/a}+n)^2} - \sum_{i=1}^{\infty} \frac{n^2(i-a)}{a^2(ae^{i/a}+n)^2}\right). \tag{2.8.2}$$

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 of the set $[1, \underline{a}_n) \cup (\overline{a}_n, A_n]$.

### §2.8.1 $\mathbb{M}_n(a)$ 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

$$\mathcal{I}_n \equiv \left[\frac{\log\left(\frac{n}{\underline{a}_n}\right)}{1+\log\left(\frac{n}{\underline{a}_n}\right)}\underline{a}_n, \underline{a}_n\right] \tag{2.8.3}$$

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 $a \in [1, \underline{a}_n]/\mathcal{I}_n$

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

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

By leaving the non-negative stochastic part out we get the lower bound

$$\mathbb{M}_n(a) \geq \frac{1}{2}\left(\sum_{i=1}^{a} \frac{n^2(i-a)e^{i/a}Y_i^2}{a(ae^{i/a}+n)^2} - \sum_{i=1}^{\infty} \frac{n(i-a)}{a^2(ae^{i/a}+n)}\right). \tag{2.8.4}$$

In view of Lemma 2.8.1 the deterministic part in (2.8.4) 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 $E_0\sum_{i=1}^{a} Y_i^2 = \sum_{i=1}^{a}\theta_{0,i}^2 + an^{-1}$ and $V_0\big(\sum_{i=1}^{a} Y_i^2\big) = 2n^{-1}\sum_{i=1}^{a}\theta_{0,i}^2 + an^{-2} \to 0$ for all $a \le A_n$ it follows from Chebyshev's inequality that the sum $\sum_{i=1}^{a} Y_i^2$ is bounded with probability going to 1, for all $\theta_0 \in \ell_2(M)$.

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

$$\mathbb{M}_n(a) \ge \frac{1}{2}\left(\sum_{i=1}^{a}\frac{n^2(i-a)e^{i/a}Y_i^2}{a(ae^{i/a}+n)^2} + \log^2\left(\frac{n}{a}\right)g_n(a,\theta_0) + \mathbb{H}_n(a) - \sum_{i=1}^{\infty}\frac{n(i-a)}{a^2(ae^{i/a}+n)}\right),$$

with the centered Gaussian process

$$\mathbb{H}_n(a) = \sum_{i=2a}^{\infty}\frac{n^{3/2}(i-a)e^{i/a}\theta_{0,i}Z_i}{a(ae^{i/a}+n)^2}. \tag{2.8.5}$$

Note that

$$\begin{aligned} V_0\left(\frac{\mathbb{H}_n(a)}{\log^2\left(\frac{n}{a}\right)}\right) &= \frac{1}{\log^4\left(\frac{n}{a}\right)}\sum_{i=2a}^{\infty}\frac{n^3(i-a)^2e^{2i/a}\theta_{0,i}^2}{a^2(ae^{i/a}+n)^4}V_0(Z_i) \\ &\le \frac{ng_n(a,\theta_0)}{a\log^2\left(\frac{n}{a}\right)}\max_{i\ge 2a}\frac{(i-a)e^{i/a}}{(ae^{i/a}+n)^2} \ge \frac{g_n(a,\theta_0)}{a\log\left(\frac{n}{a}\right)}, \end{aligned}$$

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

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

is bounded by a multiple of $\sup_{a\in\mathcal{I}_n} g_n(a,\theta_0)^{1/2}(a\log(n/a))^{-1/2}$.

Next we give an upper bound for the covering number of the interval $\mathcal{I}_n$. Let us take $\varepsilon$-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 2.8.2 we have $d_n(a_1,a_2) \lesssim |a_1-a_2|\sup_{a\in\mathcal{I}_n}\sqrt{\log(n/a)g_n(a,\theta_0)/a^3}$. Thus the covering number of the interval $(a, a+1/2)$ relative to $d_n$ is bounded from above by a multiple of $\varepsilon^{-1}\sup_{a\in\mathcal{I}_n}\sqrt{\log(n/a)g_n(a,\theta_0)/a^3}$, which implies that the covering number of the whole interval $\mathcal{I}_n$ is bounded from above by constant times $\varepsilon^{-1}\sup_{a\in\mathcal{I}_n}\sqrt{\log^{-1}(n/a)g_n(a,\theta_0)/a + \underline{a}_n/\log(n/\underline{a}_n)}$.

We show below that for any $c_0 > 2$

$$e^{-2c_0}B\log n + o(1) \le g_n(a,\theta_0) \le e^{c_0}B\log n + o(1), \qquad \text{for } a \in \mathcal{I}_n, \tag{2.8.6}$$

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 (Van Der Vaart and Wellner, 1996) (applied with $\psi(x) = e^{x^2} - 1$) it follows that

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

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

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

with probability going to one. In view of (2.8.6) 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 remained to verify assertion (2.8.6). First note that

$$\frac{n^2}{\log^2\left(\frac{n}{\underline{a}_n}\right)}\sum_{i=c_0 I_{\underline{a}_n}}^{\infty}\frac{(i-\underline{a}_n)e^{i/\underline{a}_n}}{\underline{a}_n(\underline{a}_n e^{i/\underline{a}_n}+n)^2}\theta_{0,i}^2 \leq \frac{n^2}{\underline{a}_n^3\log^2\left(\frac{n}{\underline{a}_n}\right)}\sum_{i=c_0 I_{\underline{a}_n}}^{\infty} ie^{-i/\underline{a}_n}\theta_{0,i}^2$$
$$\lesssim \frac{A_n^{c_0-2}}{n^{c_0-2}\log\left(\frac{n}{\underline{a}_n}\right)}\|\theta_0\|_2^2 = o(1).$$

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

$$g_n(a,\theta_0) \geq \frac{n^2}{\log^2\left(\frac{n}{a}\right)}\sum_{i=2\underline{a}_n}^{c_0 I_{\underline{a}_n}}\frac{(i-a)e^{i/a}}{a(ae^{i/a}+n)^2}\theta_{0,i}^2$$
$$\geq \frac{n^2}{e^{2c_0}\log^2\left(\frac{n}{\underline{a}_n}\right)}\sum_{i=2\underline{a}_n}^{c_0 I_{\underline{a}_n}}\frac{(i-\underline{a}_n)e^{i/\underline{a}_n}}{\underline{a}_n(\underline{a}_n e^{i/\underline{a}_n}+n)^2}\theta_{0,i}^2.$$

By combining the preceding two displays we get that

$$g_n(a,\theta_0) \geq e^{-2c_0}g_n(\underline{a}_n,\theta_0) - \frac{e^{-2c_0}n^2}{\log^2\left(\frac{n}{\underline{a}_n}\right)}\sum_{i=c_0 I_{\underline{a}_n}}^{\infty}\frac{(i-\underline{a}_n)e^{i/\underline{a}_n}}{\underline{a}_n(\underline{a}_n e^{i/\underline{a}_n}+n)^2}\theta_{0,i}^2$$
$$\geq e^{-2c_0}B\log n + o(1),$$

finishing the proof of the first inequality in (2.8.6). The proof of the second inequality goes accordingly.

**Lemma 2.8.1.** *There exists a constant $K > 0$ such that for any $a \in [1, n)$*

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

*Proof.* Note that

$$\begin{aligned}\sum_{i=1}^{\infty} \frac{n(i-a)}{a^2(ae^{i/a}+n)} &\le \sum_{i=1}^{\infty} \frac{ni}{a^2(ae^{i/a}+n)} \le \sum_{i=1}^{I_a} \frac{i}{a^2} + \sum_{i=I_a}^{\infty} \frac{nie^{-i/a}}{a^3} \\ &\lesssim \log^2(n/a) + \frac{\log(n/a)}{a} \lesssim \log^2(n/a).\end{aligned}$$

□

**Lemma 2.8.2.** *There exists a constant $K > 0$ such that for any positive $a_1$ and $a_2$ such that $a_1 < a_2$, $\lfloor 2a_2 \rfloor - \lfloor 2a_1 \rfloor = 0$*

$$V_0\Big(\frac{\mathbb{H}_n(a_1)}{\log(n/a_1)^2} - \frac{\mathbb{H}_n(a_2)}{\log(n/a_2)^2}\Big) \le K(a_1-a_2)^2 \sup_{a\in[a_1,a_2]} \frac{\log(n/a) g_n(a,\theta_0)}{a^3}.$$

*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

$$d_n^2(a_1,a_2) = \sum_{i=2a_2}^{\infty} \big(\phi_i(a_1) - \phi_i(a_2)\big)^2 n^3 \theta_{0,i}^2 \tag{2.8.7}$$

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 2.11.3, the right hand side of (2.8.7) is bounded by

$$\begin{aligned}&n^3(a_1-a_2)^2 \sup_{a\in[a_1,a_2]} \sum_{i=2a}^{\infty} \frac{i^2}{a^4}\phi_i(a)^2\theta_{0,i}^2 \\ &\qquad\lesssim (a_1-a_2)^2 \sup_{a\in[a_1,a_2]} g_n(a,\theta_0) \sup_{i\in\mathbb{N}} \frac{ni^3 e^{i/a}}{a^5 \log^2(n/a)(ae^{i/a}+n)^2}.\end{aligned}$$

Then the statement of the lemma follows by applying Lemma 2.11.1 (with $m = 3$). □

## §2.8.2 $\mathbb{M}_n(a)$ on $[\overline{a}_n, A_n]$

We prove that for sufficiently large choice of $K_0 > 0$ in the definition of $\overline{a}_n$

$$\limsup_n \sup_{\theta_0\in\ell_2(M)} \sup_{a\in[\overline{a}_n,A_n]} E_0\Big[\frac{\mathbb{M}_n(a)}{\log^2(n/a)}\Big] < -2^{-5}, \tag{2.8.8}$$

$$\limsup_n \sup_{\theta_0\in\ell_2(M)} E_0\left[\sup_{a\in[\overline{a}_n,A_n]} \frac{|\mathbb{M}_n(a) - E_0[\mathbb{M}_n(a)]|}{\log^2(n/a)}\right] \le 2^{-6}. \tag{2.8.9}$$

These imply that with probability tending to one $\mathbb{M}_n(a) < -2^{-6}\log^2(n/a)$, for every $a \in [\bar{a}_n, A_n]$, 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.

$$\inf_{\theta_0 \in \ell_2(M)} P_0(\hat{a}_n \leq \overline{a}_n) \to 1. \tag{2.8.10}$$

Proof of assertion (2.8.8): In view of $h_n(a, \theta_0) \leq b$ for all $a \in [\overline{a}_n, A_n]$ (assuming that $\overline{a}_n > K_0$), we get that

$$\begin{aligned} E_0\Big[\frac{\mathbb{M}_n(a)}{\log^2(n/a)}\Big] &= \frac{1}{2}\Big(h_n(a,\theta_0) - \frac{1}{\log^2(n/a)}\sum_{i=1}^{\infty}\frac{n^2(i-a)}{a^2(ae^{i/a}+n)^2}\Big) \\ &\leq \frac{1}{2}\Big(b - \frac{1}{\log^2(n/a)}\sum_{i=1}^{\infty}\frac{n^2(i-a)}{a^2(ae^{i/a}+n)^2}\Big). \end{aligned}$$

In view of Lemma 2.11.2 (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,

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

which implies that

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

concluding the proof of assertion (2.8.8), for small enough choice of $b$ ($b < 2^{-4}$ is small enough).

Proof of assertion (2.8.9): In view of Corollary 2.2.5 in (Van Der Vaart and Wellner, 1996) (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 $a \in [\overline{a}_n, A_n]$

$$V_0\big(\mathbb{M}_n(a)/\log^2(n/a)\big) \leq K_1/\log(n/a), \tag{2.8.11}$$

$$\int_0^{diam_n}\sqrt{N(\varepsilon, [\overline{a}_n, A_n], d_n)}d\varepsilon \leq K_2/K_0^{1/4}, \tag{2.8.12}$$

where $d_n$ is the semi-metric defined by $d_n^2(a_1,a_2) := V_0\Big(\frac{\mathbb{M}_n(a_1)}{\log^2(n/a_1)} - \frac{\mathbb{M}_n(a_2)}{\log^2(n/a_2)}\Big)$, $diam_n$ 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 $\varepsilon$ needed to cover the set $S$, since by sufficiently large choice of $K_0$ ($K_0 \geq (2^6K_2)^4$ is sufficiently large) assertion (2.8.9) holds.

Note that Lemma 2.8.3 immediately implies assertion (2.8.11) and

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

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 2.8.4, when $a_1$ and $a_2$ are on the interval $[2^k\overline{a}_n, 2^{k+1}\overline{a}_n]$

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

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 2.8.3.** *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 $V_0(Y_i^2) = 2/n^2 + 4\theta_{0,i}^2/n$, so the variance is equal to

$$\begin{aligned} V_0\Big(\frac{\mathbb{M}_n(a)}{\log^2(n/a)}\Big) &= \frac{1}{4}\sum_{i=1}^{\infty} \frac{n^4V_0(Y_i^2)e^{2i/a}(i-a)^2}{a^2\log^4(n/a)(ae^{i/a}+n)^4} \\ &= \frac{1}{2}\sum_{i=1}^{\infty} \frac{n^2e^{2i/a}(i-a)^2}{a^2\log^4(n/a)(ae^{i/a}+n)^4} + \sum_{i=1}^{\infty} \frac{n^3e^{2i/a}(i-a)^2\theta_{0,i}^2}{a^2\log^4(n/a)(ae^{i/a}+n)^4}. \end{aligned} \tag{2.8.13}$$

In view of $(i-a)^2 \le a^2+i^2$, for any $a,i>0$, and by applying Lemma 2.11.1 (with $m=2$) and Lemma 2.11.2 (first with $r=2$ and $l=4$ and then with $r=1$ and $l=2$) the first sum in (2.8.13) is bounded from above by a multiple of

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

Similarly, following from Lemma 2.11.1 (with $m=1$ and $m=-1$) and $h_n(a,\theta_0) \le b$ for $a \ge \overline{a}_n$, the second sum in (2.8.13) is bounded by a multiple of

$$\begin{aligned} &\Big(\max_{i\in\mathbb{N}} \frac{ane^{i/a}}{i\log^2(n/a)(ae^{i/a}+n)^2} + \max_{i\in\mathbb{N}} \frac{ine^{i/a}}{a\log^2(n/a)(ae^{i/a}+n)^2}\Big)h_n(a,\theta_0) \\ &\qquad \lesssim \Big(\frac{1}{a\log^3(n/a)} + \frac{1}{a\log(n/a)}\Big) \lesssim \frac{1}{a\log(n/a)}, \end{aligned}$$

concluding the proof of the lemma. □

**Lemma 2.8.4.** *For all $1 \le a_1 < a_2 < A_n$, we have*

$$d_n^2(a_1,a_2) \le C_0(a_1-a_2)^2 \sup_{a\in[a_1,a_2]} \frac{\log(n/a)}{a^3}(1+h_n(a,\theta_0)),$$

*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 $|\phi_i(a)'|^2 \lesssim (i^2a^{-4} + a^{-2})\phi_i^2(a)$, hence in view of Lemma 2.11.3,

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

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

$$\sum_{i=1}^{\infty} \frac{2e^{2i/a}(i^4+a^4)}{a^6n^2\log^4(n/a)(ae^{i/a}+n)^4} + \sum_{i=1}^{\infty} \frac{4e^{2i/a}(i^4+a^4)}{a^6n\log^4(n/a)(ae^{i/a}+n)^4}\theta_{0,i}^2. \tag{2.8.14}$$

Then in view of Lemma 2.11.1 (applied with $m=4$ and $m=0$) and Lemma 2.11.2 (applied with $r=1$ and $l=2$) the first term of (2.8.14) is bounded from above by a multiple of

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

Similarly in view of Lemma 2.11.1 (with $m=3$ and $m=-1$) the second term of (2.8.14) is bounded by

$$\begin{aligned}\max_{i\in\mathbb{N}} &\frac{\big((i/a)^3 + (i/a)^{-1}\big)e^{i/a}}{a^2n^3\log^2(n/a)(ae^{i/a}+n)^2} h_n(a,\theta_0) \\ &\lesssim \Big(\frac{\log(n/a)}{a^3n^4} + \frac{1}{n^5a^2}\Big)h_n(a,\theta_0) \lesssim \frac{\log(n/a)}{a^3n^4} h_n(a,\theta_0),\end{aligned}$$

concluding the proof of the lemma. □

## §2.9 Proof of Theorem 2.1.6

Similarly to the previous sections we use the notations introduced in Section 2.5. We show below that there exists a constant $c>0$ depending only on $m, M$ and $\beta_0$ such that

$$\inf_{\beta\geq\beta_0} \inf_{\theta_0\in\Theta_s^\beta(m,M)} P_0(\hat{a}_n \geq c(n/\log n)^{1/(1+2\beta)}/\log n) \to 1, \tag{2.9.1}$$

which combined with Proposition 2.4.2 and Theorem 2.4.1 results in

$$\inf_{\beta\geq\beta_0} \inf_{\theta_0\in\Theta_s^\beta(m,M)} P_0(c(n/\log n)^{1/(1+2\beta)} \leq \tilde{a}_n \leq C(n/\log n)^{1/(1+2\beta)}) \to 1,$$

for some positive constants $c, C$ depending on $b, B, m, M$ and $\beta$. Let us introduce then 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 $\theta_0 \in \hat{C}_n(L)$ is equivalent to $\|\theta_0 - \hat{\theta}\|_2 \le Lr_\alpha(\tilde{a}_n)$, hence by proving that

$$\inf_{a\in\tilde{\mathcal{I}}_n} r_\alpha^2(a) \ge C_1(n/\log n)^{-2\beta/(1+2\beta)},$$

$$\inf_{\beta\ge\beta_0}\inf_{\theta_0\in\Theta_s^\beta(m,M)} P_0\big(\inf_{a\in\tilde{\mathcal{I}}_n}\|W(a)\|_2^2 \le C_2(n/\log n)^{-2\beta/(1+2\beta)}\big) \to 1,$$

$$\sup_{\beta\ge\beta_0}\sup_{\theta_0\in\Theta_s^\beta(m,M)}\sup_{a\in\tilde{\mathcal{I}}_n}\|B(a,\theta_0)\|_2^2 \le C_3(n/\log n)^{-2\beta/(1+2\beta)},$$

hold for some constants $C_1, C_2, C_3 > 0$, the statement of the theorem follows immediately. The proof of the first two inequalities follow from (2.5.2) and (2.5.9) (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 $\theta_0 \in \Theta_s^\beta(m, M)$, $a \in \tilde{\mathcal{I}}_n$, and $\beta \ge \beta_0$ we have that

$$\begin{aligned}\|B(a,\theta_0)\|_2^2 &\lesssim \sum_{i=1}^{I_a/2} a^2 e^{2i/a} n^{-2} i^{-1-2\beta} + \sum_{i=I_a/2}^{\infty} i^{-1-2\beta} \lesssim a/n + I_a^{-2\beta}\\ &= o\big((n/\log n)^{-2\beta/(1+2\beta)}\big).\end{aligned}$$

It remained to prove assertion (2.9.1). Let us introduce the slightly modified version of $\underline{a}_n$ as

$$\underline{a}'_n := \sup\{a \in [1, A_n] : g_n(a, \theta_0) \ge B\},$$

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

$$P_0(\hat{a}_n \ge \underline{a}'_n) \to 1, \quad \text{and} \quad \underline{a}'_n \ge c(n/\log n)^{1/(1+2\beta)}/\log n, \tag{2.9.2}$$

for some sufficiently small constant $c > 0$.

For the second statement note that

$$g_n(a,\theta_0) \ge \frac{m}{\log^2(n/a)} n^2 \sum_{i=I_a}^{\infty} e^{-i/a} i^{-2\beta} \gtrsim mna^{-1-2\beta}\log^{-2-2\beta}(n/a), \tag{2.9.3}$$

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 remained to deal with the first part of (2.9.2). We show below that with probability tending to one $\inf_{a\in[\underline{a}'_n/2,\underline{a}'_n]}\mathbb{M}_n(a) \ge cB\log^2(n/a)$, 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

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

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

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

$$g_n(a,\theta_0) \le \frac{M}{\log^2(n/a)}\Big(\frac{1}{a}\sum_{i=2a}^{I_a} e^{i/a} i^{-2\beta} + \frac{n^2}{a^3}\sum_{i=I_a}^{\infty} e^{-i/a} i^{-2\beta}\Big)$$
$$\le c_{M,\beta} n a^{-1-2\beta} (\log n)^{-2-2\beta},$$

hence $\underline{a}'_n \le c'_{M,\beta} B^{-1/(1+2\beta)} (n/\log n)^{1/(1+2\beta)}/\log n$. Therefore in view of (2.9.3) for every $a \ge \underline{a}'_n/2$ we have $g_n(a,\theta_0) \ge c_{M,\beta,m} B$, for some positive constant $c_{M,\beta,m} > 0$ not depending on $B$. Similarly we can show that $g_n(a,\theta_0) \le c'_{M,\beta,m} B$, for every $a \ge \underline{a}'_n/2$, for some $c'_{M,\beta,m} > 0$ not depending on $B$. Then following the same line of reasoning as in Section 2.8.1, with the only main difference that instead of the interval given in (2.8.3) we are working with the interval $[\underline{a}'_n/2, \underline{a}'_n]$ we get that with probability going to one

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

for large enough choice of $M > 0$, finishing the proof of the theorem.

## §2.10 Proofs for the Hierarchical Bayes procedure

In this section we prove the results on the hierarchical Bayes procedure (i.e. Theorems 2.4.5 and 2.1.5 and Corollary 2.1.2) based on the results derived for the empirical Bayes procedure. First we state that under the conditions of Theorem 2.4.5 the hyper-posterior distribution on the hyper-parameter $a$ 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 2.10.1.** *If $a \sim \pi(.)$ such that $\pi$ verifies Assumption 2.1.1 then for sufficiently large $C > 0$ we have for every $\beta_0 > 0$ that*

$$\inf_{\beta>\beta_0} \inf_{\theta_0\in\Theta^\beta(M)} E_0 \Pi\Big(\underline{a}_n \log(n)/(1+\log n) \le a \le C\overline{a}_n | Y\Big) = 1 + o(1/n).$$

## §2.10.1 Proof of Theorem 2.4.5

Take $\varepsilon_n = (n/\log^2 n)^{-\beta/(1+2\beta)}$. Then following from Lemma 2.10.1, we have

$$\sup_{\theta_0\in\Theta^\beta(M)} E_0 \Pi\big(\theta : \|\theta-\theta_0\|_2 > M_n\varepsilon_n | Y\big) \le \sup_{\theta_0\in\Theta^\beta(M)} \Big(E_0\Pi(a \notin \mathcal{I}_n | Y)$$
$$+ E_0 \sup_{a\in\mathcal{I}_n} \Pi_a\big(\theta : \|\theta-\theta_0\|_2 > M_n\varepsilon_n | Y\big)\Big) = o(1),$$

where the last equation follows by similar arguments as given in (2.4.9) 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.

## §2.10.2 Proof of Theorem 2.1.3 - Hierarchical Bayes part

In the proof we use again the notations introduced in Section 2.5.

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 2.4.2. One can see that in the hierarchical case,

$$P_0(\theta_0 \in \hat{C}_n(L_n)) \leq P_0\Big(\|B(a', \theta_0)\|_2 \leq L_n r_\alpha + \|W(a')\|_2 + \|\hat{\theta} - \hat{\theta}_{a'}\|_2\Big) + o(1), \tag{2.10.1}$$

which is a slightly modified version of (2.7.1) 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

$$r_\alpha^2 \leq \tilde{C}_1 n^{-2\beta/(1+2\beta)} \log(n)^{-1/(1+2\beta)}, \tag{2.10.2}$$

$$P_0(\|W(a')\|_2^2 \leq \tilde{C}_2 n^{-2\beta/(1+2\beta)} \log(n)^{-1/(1+2\beta)}) \to 1, \tag{2.10.3}$$

$$\|B(a', \theta_0)\|_2^2 \geq \tilde{C}_3 n^{-2\beta/(1+2\beta)} \log(n)^{2\beta/(1+2\beta)}, \tag{2.10.4}$$

$$P_0(\|\hat{\theta} - \hat{\theta}_{a'}\|_2 \leq \tilde{C}_4 n^{-2\beta/(1+2\beta)} \log(n)^{-1/(1+2\beta)}) \to 1. \tag{2.10.5}$$

The bounds on the variance and the bias are obtained in a similar manner as in Section 2.7 and 2.10.3. Next we deal with assertion (2.10.5).

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

$$\begin{aligned}
\|\hat{\theta} - \hat{\theta}_{a'}\|_2 &= \| \int (\hat{\theta}_a - \hat{\theta}_{a'}) \Pi(da|Y) \|_2^2 \qquad (2.10.6)\\
&\leq \sum_{i=1}^{\infty} \int (\hat{\theta}_{a,i} - \hat{\theta}_{a',i})^2 \Pi(da|Y) \\
&\leq \sup_{a_1 \in \mathcal{I}_n} \sum_{i=1}^{\infty} (\hat{\theta}_{a_1,i} - \hat{\theta}_{a',i})^2 \Pi(a_1 \in \mathcal{I}_n|Y) + \sup_{a_1 \notin \mathcal{I}_n} \sum_{i=1}^{\infty} (\hat{\theta}_{a_1,i} - \hat{\theta}_{a',i})^2 \Pi(a_1 \notin \mathcal{I}_n|Y).
\end{aligned}$$

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

$$\begin{aligned}
\sup_{a_1 \in \mathcal{I}_n} \sum_{i=1}^{\infty} (\hat{\theta}_{a_1,i} - \hat{\theta}_{a',i})^2 \leq & \sup_{a_1 \in \mathcal{I}_n} \sum_{i=1}^{\infty} (E_0\hat{\theta}_{a_1,i} - E_0\hat{\theta}_{a',i})^2 \qquad (2.10.7)\\
&+ \sup_{a_1 \in \mathcal{I}_n} \sum_{i=1}^{\infty} (\hat{\theta}_{a_1,i} - E_0\hat{\theta}_{a_1,i})^2 + \sum_{i=1}^{\infty} (\hat{\theta}_{a',i} - E_0\hat{\theta}_{a',i})^2
\end{aligned}$$

The two last term on the right hand-side are bounded by a constant multiplier of $n^{-2\beta/(1+2\beta)}\log(n)^{-1/(1+2\beta)}$ from (2.5.3). 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 $g_i(a)=n\theta_{0,i}^2/(ae^{i/a}+n)$. The derivative of $g_i(a)$ is $-n\theta_{0,i}^2(a-i)e^{i/a}/(a(ae^{i/a}+n)^2)$. Without loss of generality, when $a_1<a'$ writing the difference as the integral of $g_i'(a)$, applying Cauchy-Schwartz inequality to its squares and then interchanging the sum and the integral, we get that

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

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

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

Therefore, one can see that

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

with probability tending to one using Proposition 2.4.2

It is left to deal with the second term on the right hand-side of (2.10.6). Following from (2.10.7), we get with $P_0$-probability tending to 1 that

$$\begin{aligned}\sup_{a_1\notin\mathcal{I}_n}\sum_{i=1}^{\infty}(\hat{\theta}_{a_1,i}-\hat{\theta}_{a',i})^2\leq 2\sup_{a_1\notin\mathcal{I}_n}\sum_{i=1}^{\infty}(E_0\hat{\theta}_{a_1,i})^2+\sup_{a_1\notin\mathcal{I}_n}\sum_{i=1}^{\infty}(\hat{\theta}_{a_1,i}-E_0\hat{\theta}_{a_1,i})^2\\+2\sum_{i=1}^{\infty}(E_0\hat{\theta}_{a',i})^2+\sum_{i=1}^{\infty}(\hat{\theta}_{a',i}-E_0\hat{\theta}_{a',i})^2,\end{aligned}\tag{2.10.8}$$

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 (2.10.6) being of lower order than $n^{-1}$.

It remained to deal with assertion (2.10.2). We show below that

$$r_\alpha \le \tilde{r} := \sup_{a\in\mathcal{I}_n}\Big(\|\hat{\theta}-\hat{\theta}_a\|_2 + r_{\alpha/2}(a)\Big). \tag{2.10.9}$$

Then in view of the inequality

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

and assertions (2.10.5), (2.10.7), and (2.7.2) we get that with probability tending to one $r_\alpha^2 \le \tilde{C}_1 n^{-2\beta/(1+2\beta)}\log(n)^{-1/(1+2\beta)}$ and since $r_\alpha$ is deterministic the inequality holds almost surely.

Finally we verify assertion (2.10.9). Note that

$$\begin{aligned}\Pi(\theta:\ \|\hat{\theta}-\theta\|_2\le\tilde{r}|Y) &\ge \int_{\mathcal{I}_n}\Pi_a(\theta:\ \|\theta-\hat{\theta}\|_2\le\tilde{r}|Y)\pi(a|Y)da\\ &\ge \int_{\mathcal{I}_n}\Pi_a(\theta:\ \|\theta-\hat{\theta}_a\|_2\le r_{\alpha/2}(a)|Y)\pi(a|Y)da\\ &\ge \int_{\mathcal{I}_n}(1-\alpha/2)\pi(a|Y)da < 1-\alpha,\end{aligned}$$

for large enough $n$, concluding the proof of our theorem for the Hierarchical Bayes method.

## §2.10.3 Proof of Theorem 2.1.5 - Hierarchical Bayes part

Let us introduce the notations $W=\hat{\theta}-E_0\hat{\theta}$ and $B(\theta_0)=E_0\hat{\theta}-\theta_0$, for the centered hierarchical posterior mean and the bias of the posterior mean, respectively. Then $P_0(\theta_0\in\hat{C}(L\log n))$ if and only if

$$\|W\|_2 \le L\log(n)r_\alpha - \|B(\theta_0)\|_2 \tag{2.10.10}$$

holds. Using assertions (2.5.2), (2.5.3), and (2.5.4) we show below that, there exist constants $\tilde{C}_1,\tilde{C}_2,\tilde{C}_3>0$, such that

$$r_\alpha^2 \ge \tilde{C}_1(\underline{a}_n/n)\log(n/\underline{a}_n), \tag{2.10.11}$$

$$\inf_{\theta_0\in\Theta_{pt}(L_0,N_0,\rho)} P_0\big(\|W\|_2^2 \le \tilde{C}_2(\underline{a}_n/n)\log(n/\underline{a}_n)\log^2 n\big)\to 1, \tag{2.10.12}$$

$$\|B(\theta_0)\|_2^2 \le \tilde{C}_3(\underline{a}_n/n)\log^2\big(n/\underline{a}_n\big)\log n, \tag{2.10.13}$$

resulting in (2.10.10) for sufficiently large choice of $L>0$.

Proof of (2.10.11): Let us take any $\alpha'>\alpha$ and note that in view of (2.5.2) we have

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

Next, in view of Lemma 2.10.1 and Anderson's lemma, we get for arbitrary $r \leq \inf_{a\in\mathcal{I}_n} r_{\alpha'}(a)$ that

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

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 (2.10.12): Note that by triangle inequality, Fubini's theorem, assertion (2.5.3), and Lemma 2.10.1 we get that under the polished tail condition with $P_0$-probability tending to one

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

where $\pi(\mathcal{I}_n|Y)$ denotes (by slightly abusing our notation) the posterior probability that the hyper-parameter $a$ lies in the interval $\mathcal{I}_n$ and in the last inequality we used in view of the proof of assertion (2.5.3) that $\sup_{1\leq a\leq A_n}\|W(a)\|_2=O(1)$.

Proof of (2.10.13): Similarly to the proof of (2.10.12) we get that

$$\begin{aligned}\|B(\theta_0)\|_2^2 &\lesssim \sup_{a\in\mathcal{I}_n}\|B(a,\theta_0)\|_2^2+o(\sup_{a\in[1,A_n]}\|B(a,\theta_0)\|_2^2/n)\\ &\leq C_3(\underline{a}_n/n)\log^2\big(n/\underline{a}_n\big)\log n+o(1/n),\end{aligned}$$

where the last inequality follows from $\|B(a,\theta_0)\|_2^2\leq\|\theta_0\|_2^2=O(1)$, finishing the proof of the theorem.

## §2.10.4 Proof of Corollary 2.1.2

Let $\varepsilon_n=(n/\log^2 n)^{-\beta/(1+2\beta)}$ and first note that in view of assertions (2.10.12) and (2.10.13) combined with triangle inequality and Proposition 2.4.2 we have with $P_0$-probability tending to one that

$$\|\theta_0-\hat\theta\|_2\leq\|W\|_2+\|B(\theta_0)\|_2\lesssim\sqrt{\underline{a}_n/n}\log(n/\underline{a}_n)\lesssim\varepsilon_n.$$

Then in view of Theorem 2.4.5 and by applying again the triangle inequality we get with probability tending to one that

$$\Pi(\theta:\ \|\theta-\hat\theta\|_2\leq M_n\varepsilon_n|Y)\geq\Pi(\theta:\ \|\theta-\theta_0\|_2\leq M_n\varepsilon_n-\|\theta_0-\hat\theta\|_2|Y)=1-o(1),$$

concluding the proof of the corollary.

## §2.10.5 Proof of Lemma 2.10.1

In Section 2.8 it was shown that $\mathbb{M}_n(a) = \frac{\partial \ell_n(a)}{\partial a}$ satisfies, for positive constants $K_1$, $K_2$ and $K_3$,

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

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

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

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

$$\Pi(a \geq C\overline{a}_n | Y) \leq \frac{\int_{C\overline{a}_n}^{\infty} e^{\ell_n(a)}\pi(a)da}{\int_{\overline{a}_n}^{2\overline{a}_n} e^{\ell_n(a)}\pi(a)da} \leq \frac{\Pi\big([C\overline{a}_n, \infty)\big)e^{-K_4 \log^2(n/\overline{a}_n)\overline{a}_n}}{\Pi([\overline{a}_n, 2\overline{a}_n])}. \tag{2.10.14}$$

Note that by Assumption 2.1.1

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

hence the right hand side of (2.10.14) tends to zero.

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

$$\Pi(a \leq \underline{a}_n^* | Y) \leq \frac{\int_1^{\underline{a}_n^*} e^{\ell_n(a)}\pi(a)da}{\int_{(\underline{a}_n + \underline{a}_n^*)/2}^{\underline{a}_n} e^{\ell_n(a)}\pi(a)da} \leq \frac{\Pi([1, \underline{a}_n])e^{K_3\underline{a}_n \log^2(n/\underline{a}_n)}}{\Pi([(\underline{a}_n + \underline{a}_n^*)/2, \underline{a}_n])e^{K_2(\underline{a}_n/4)\log^2(n/\underline{a}_n)}}. \tag{2.10.15}$$

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 large enough choice of $K_2$, the right hand side of (2.10.15) tends to zero, finishing the proof of the lemma.

## §2.11 Technical Lemmas

**Lemma 2.11.1.** *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 $f'(x) = e^{-x/a}x^{m-1}(m - x/a)$, 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

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

□

**Lemma 2.11.2.** *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 $\frac{n}{a} \geq e^{l-r}$. □

**Lemma 2.11.3** (Lemma C.11 of (van der Pas et al., 2017))**.** *For any stochastic process $(V_a : a > 0)$ with continuously differentiable sample paths $a \mapsto V_a$, with derivative written as $\dot{V}_a$,*

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

## §2.12 Extra simulation study

The purpose of this section is to reinforce the evidence shown in Section 2.2. To this end, we will show graphically and numerically the sub-optimal performance of the Gaussian process with (approximately) squared exponential covariance kernel compared to other methods in the non-parametric regression model specifically. In this

simulation study we take the Fourier coefficients of the underlying true function $\theta_3$ to be $\theta_{3,i} = i^{-3/2}\cos(i)$, $i = 1, 2, ....$ 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 size to be $n = 500, 1000, 5000$, and 10000. Observe in Figure 2.4 that the standard MMLE empirical Bayes method provides unreliable uncertainty quantification in certain points, especially compared to the three other methods. One can also observe through different running times in Table 2.9, 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 point-wise 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 $0.6474 = \text{argmax}_{x\in[0,1]}\theta_3(x)$) is included in the credible interval, see Table 2.7. Moreover, Table 2.8 shows the average size of the point-wise 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 similar behavior to what we have described above.

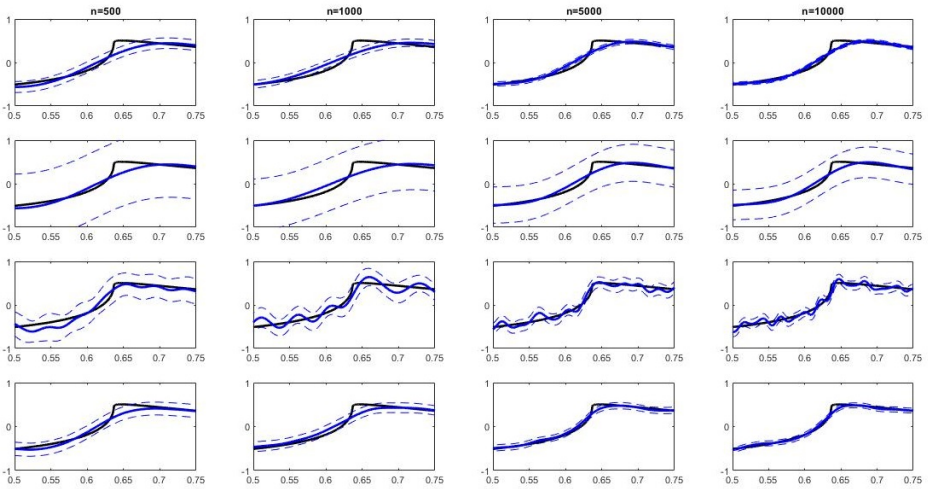


Figure 2.4: Empirical Bayes credible sets for the regression function $\theta_3$ (drawn in black), zoomed in to the interval $x \in [0.5, 0.75]$. The posterior means are drawn by solid blue line, while the 95% point-wise credible sets by dashed blue curves. In the first row we plot the MMLE empirical Bayes method, in the second row the MMLE empirical Bayes method with a $\log n$ blow up factor, the third row the modified MMLE empirical Bayes method using squared exponential Gaussian process prior, while in the fourth row we plot the empirical Byes credible sets using a Matérn kernel with data-driven choice for the regularity hyper-parameter. From left to right the sample size is $n = 500, 1000, 5000, 10000$.

We also consider a multi-variable version of the previous regression with $d = 10$ variables. The Fourier coefficients of the underlying true function $\theta_4$ become $\theta_{4,i} = \prod_{k=1}^{10}(i_k^{-3/2}\cos(i_k))$, $i_k = 1, 2, ...$ for all $k = 1, 2, ..., 10$, 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 point-wise credible sets at given points (we consider $x = (\{0.25\}^{10}, \{0.3188\}^{10}, \{0.75\}^{10})$) in Table 2.10 and note that similar conclusions

| | $x = 0.25$ | | | $x = 0.6474$ | | | $x = 0.75$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $n =$ | 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 2.7: Frequencies that $\theta_3(x)$ is inside of 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: SE kernel MMLE empirical Bayes procedure, Method 2: SE kernel empirical Bayes procedure with $\log n$ blow up factor, Method 3: SE 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 2.8: Average size of the pointwise credible intervals (i.e. $2q_{0.025}\sqrt{\hat{c}(x,x)}$) for $\theta_3(x)$ in the regression model. Method 1: SE kernel MMLE empirical Bayes procedure, Method 2: SE kernel empirical Bayes procedure with $\log n$ blow up factor, Method 3: SE 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 | 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 |

Table 2.9: Average run time of the EB methods for $\theta_3$ in the regression model. Method 1: SE covariance kernel, Method 4: Matérn covariance kernel and MMLE for the regularity hyper-parameter, Method 5: Matérn covariance kernel and MMLE for the scaling hyper-parameter with fixed regularity $\alpha = 10$. From left to right the sample size is $n = 100, 500, 1000, 5000, 10000$

can be drawn as in the $d = 1$ dimensional case. Surprisingly, the computation times for the posterior in higher dimension is of similar order as their one-dimensional counterpart, hence they are omitted.

| | $x = \{0.25\}^{10}$ | | | $x = \{0.3188\}^{10}$ | | | $x = \{0.75\}^{10}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $n =$ | 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 |

Table 2.10: Frequencies that $\theta_4(x)$ is inside of the corresponding credible interval for the squared exponential and Matérn Gaussian process prior in the multivariate ($d = 10$) regression model. Method 1: SE kernel MMLE empirical Bayes procedure, Method 2: SE kernel empirical Bayes procedure with $\log n$ blow up factor, Method 3: SE 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 $n = 100, 500, 1000$.