Credible sets in nonparametric regression

Sniekers, Suzanne

Author: Sniekers, Suzanne
Title: Credible sets in nonparametric regression
Issue Date: 2015-12-01

Chapter 2

Adaptive global credible sets

2.1 Introduction and main result

We consider the fixed design regression model, where we observe a vector $\vec{Y}_n := (Y_{1,n}, \dots, Y_{n,n})^T$ with coordinates

$$Y_{i,n} = f(x_{i,n}) + \varepsilon_{i,n}, \qquad i \in \{1, \dots, n\}. \tag{2.1}$$

Here the parameter f is an unknown function $f : \mathcal{X} \to \mathbb{R}$ on some set $\mathcal{X}$, the design points $(x_{i,n})$ are a known sequence of points in $\mathcal{X}$, and the (unobservable) errors $\varepsilon_{i,n}$ are independent standard normal random variables. We are interested in the performance of a nonparametric Bayesian approach that uses a scaled Gaussian process $\sqrt{c}W$ as a prior on f. We investigate its efficiency to reconstruct the true regression function, and its ability to quantify the remaining uncertainty in the statistical analysis through the full posterior distribution. Our main interest is in the dependence of the posterior distribution on the scaling factor $\sqrt{c}$ in the Gaussian process, which can be viewed as a bandwidth parameter that can adapt the prior and posterior distributions to the unknown regularity of the regression function. We consider empirical and hierarchical Bayes methods to determine this scaling factor, and study the properties of the resulting plug-in or full posterior distributions.

We denote the prior process for $f = \big(f(x) : x \in \mathcal{X}\big)$ by $W^c = (W^c_x : x \in \mathcal{X})$, where c is the scaling factor, and it is assumed that the process W^c is equal in distribution to the process $\sqrt{c}\,W^1$. The index set $\mathcal{X}$ may possess a special structure, but the general results allow it to be arbitrary. These results cover both one-dimensional and multidimensional domains $\mathcal{X}$.

As a particular example we consider the case that $\mathcal{X} = [0, 1]$ and W^1 is a standard Brownian motion. In this case W^c is a mean-zero Gaussian process with covariance function $\mathrm{E}W_s^cW_t^c = c\,(s\wedge t)$, and can also be obtained by taking a standard Brownian motion on the transformed time scale ct. More generally, for every *self-similar* process W^1 of order α the process $(\sqrt{c}\,W_t^1 : t \geq 0)$ is equal in distribution to $(W^1_{tc^{1/(2\alpha)}} : t \geq 0)$ and hence our present sense of scaling is equivalent to changing the *length scale* of the standard process. This applies in particular to multifold integrals (indefinite integrals) of Brownian motion, as considered in [Kimeldorf and Wahba, 1970] in connection to spline smoothing.

For a given scale c the Bayesian model is then described by

$$\begin{aligned} &f\,|\,c \sim W^c, \\ &\vec{Y}_n\,|\,f, c \sim \mathcal{N}_n(\vec{f}_n, I), \qquad\qquad \vec{f}_n = \big(f(x_{1,n}),\ldots,f(x_{n,n})\big)^T. \end{aligned} \tag{2.2}$$

The *posterior distribution* given c is by definition the conditional distribution of f given $(\vec{Y}_n, c)$ in this setup. As $\vec{Y}_n$ depends on f only through $\vec{f}_n$, the conditional distribution of f given $(\vec{Y}_n, \vec{f}_n, c)$ does not depend on the data $\vec{Y}_n$ and is the same as the conditional distribution of f given $(\vec{f}_n, c)$, which is determined by the prior only. Thus we focus on the posterior distribution of $\vec{f}_n$, which by standard Gaussian calculus can be seen to satisfy

$$\begin{aligned} &\vec{f}_n\,|\,\vec{Y}_n, c \sim \mathcal{N}_n\big(\hat{f}_{n,c}, I - \Sigma_{n,c}^{-1}\big), \\ &\hat{f}_{n,c} = (I - \Sigma_{n,c}^{-1})\vec{Y}_n, \quad \Sigma_{n,c} = I + cU_n, \end{aligned} \tag{2.3}$$

for U_n the covariance matrix of the unit scale process W^1 restricted to the design points $x_{i,n}$. For instance, for scaled Brownian motion $(U_n)_{i,j} = x_{i,n} \wedge x_{j,n}$.

If $\vec{Y}_n$ follows the model (2.1) with a continuous function f, then for fixed c the posterior mean $\hat{f}_{n,c}$ tends to $\vec{f}_n$ and the posterior covariance matrix $I - \Sigma_{n,c}^{-1}$ tends to zero as $n \to \infty$ (see [Cox, 1993, van der Vaart and van Zanten, 2008]). This remains true if $c = c_n$ is made dependent on n and allowed to tend to zero or infinity at polynomial rates. Thus the posterior distribution given $c = c_n$ contracts to the Dirac measure at f for reasonable c_n. The rate of contraction depends on c_n and the regularity of the function f jointly. A smaller value of c corresponds to less variability in the prior process, and yields a posterior distribution with a less variable mean function and a smaller covariance. This is advantageous if the true regression function f is fairly regular, but will lead to a suboptimal contraction rate and a too optimistic quantification of remaining uncertainty in the opposite case (see [van der Vaart and van Zanten, 2007] and Chapter 1). It is therefore important to adapt c to the data. We discuss three methods, which turn out to have similar behaviour, both in terms of contraction rate and uncertainty quantification, although the sets of functions for which they work differ.

In the *hierarchical Bayes* setup the parameter c is equipped with a prior, and an ordinary Bayesian analysis is carried out with the resulting mixture of normals prior for f. We shall consider the situation that c follows an inverse Gamma distribution.

In the *empirical Bayes* setup an estimator $\hat{c}_n$ of the length scale is plugged into the posterior distribution for given c. We consider two methods of estimation: a likelihood-based and a risk-based method.

The *likelihood-based empirical Bayes* method defines $\hat{c}_n$ as the maximum likelihood estimator of c within the *marginal Bayesian model* $\vec{Y}_n \,|\, c \sim \mathcal{N}(0, \Sigma_{n,c})$, which follows from (2.2). In this marginal model c is the only parameter, and its maximum likelihood estimator is

$$\hat{c}_n = \operatorname{argmin}_{c\in I_n} \Big[\log\det \Sigma_{n,c} + \vec{Y}_n^T \Sigma_{n,c}^{-1} \vec{Y}_n\Big]. \tag{2.4}$$

The restriction of c to an interval I_n away from the extremes 0 and ∞ is convenient. Throughout the chapter we shall use

$$I_n = [\log n/n, n^{m-1}],$$

where m is chosen large enough so that the minimax scaling rates for all smoothness levels are included. (If (2.12) holds, then it is chosen equal to the m in this equation.) The likelihood-based empirical Bayes procedure ought to be close to the hierarchical Bayes procedure, as the posterior density for c is proportional to the marginal density of $\vec{Y}_n$ given c times the prior density by Bayes's rule, and hence ought to concentrate around $\hat{c}_n$ in (2.4). Thus the posterior distribution with a likelihood-based empirical Bayes plug-in for the scale parameter is sometimes viewed a computationally cheaper version of a true Bayesian analysis.

The *risk-based empirical Bayes* method uses an alternative estimator for c that tries to minimize the risk of the posterior mean $\hat{f}_{n,c}$, which is given by

$$\mathrm{E}_f \big\|\hat{f}_{n,c} - \vec{f}_n\big\|^2 = \| - \Sigma_{n,c}^{-1}\vec{f}_n\|^2 + \operatorname{tr}\big((I - \Sigma_{n,c}^{-1})^2\big). \tag{2.5}$$

The first term on the right depends on the unknown function f, and hence cannot be used in a criterion to estimate c. An obvious estimate for this term is $\| - \Sigma_{n,c}^{-1}\vec{Y}_n\|^2$, but it is biased, as

$$\mathrm{E}_f \| - \Sigma_{n,c}^{-1}\vec{Y}_n\|^2 = \|\Sigma_{n,c}^{-1}\vec{f}_n\|^2 + \mathrm{E}_f\|\Sigma_{n,c}^{-1}\vec{\varepsilon}_n\|^2 = \|\Sigma_{n,c}^{-1}\vec{f}_n\|^2 + \operatorname{tr}(\Sigma_{n,c}^{-2}).$$

This motivates the estimator for c given by

$$\hat{c}_n = \operatorname{argmin}_{c\in I_n} \Big[\operatorname{tr}\big((I - \Sigma_{n,c}^{-1})^2\big) - \operatorname{tr}(\Sigma_{n,c}^{-2}) + \vec{Y}_n^T \Sigma_{n,c}^{-2} \vec{Y}_n\Big]. \tag{2.6}$$

In the special case that W^c is an $(m-1)$-fold integral of Brownian motion, this estimator was introduced in the context of regression by spline-smoothing. The posterior mean in our setup is then equal to a penalized least squares estimator for the penalty $\lambda \int f^{(m)}(x)^2\,dx$, with smoothing parameter λ equal to $1/(cn)$. See [Wahba, 1983, Cox, 1993].

In Bayesian inference the posterior distribution is used both to reconstruct the regression function f, typically by the posterior mean, and to quantify the uncertainty in this construction, using the spread of the posterior distribution. In this chapter we are interested in the accuracy of these procedures within the so-called frequentist setup, which assumes that the data $\vec{Y}_n$ are generated according to model (2.1) for a given "true function" f. The accuracy of the posterior mean as a point estimator of f can be measured by its risk function or the contraction rate of the full posterior distribution (see [Ghosal et al., 2000]), as usual. The accuracy of the uncertainty quantification can be studied through the coverage and size of *credible sets*, which are data-dependent sets of prescribed posterior probability. In connection to the empirical Bayes methods we shall first study credible sets of the form

$$\hat{C}_{n,\eta,M} = \big\{f : \|\vec{f}_n - \hat{f}_{n,\hat{c}_n}\| < M r_n(\hat{c}_n, \eta)\big\}, \tag{2.7}$$

with $\|\cdot\|$ the Euclidean norm. Here $r_n(c,\eta)$ is for given $\eta \in (0,1)$ determined such that the ball of radius $r_n(c,\eta)$ centred at the origin receives probability η under the posterior law of $\vec{f}_n - \hat{f}_{n,c}$ given a fixed c, which by (3.3) is the normal law $\mathcal{N}_n(0, I - \Sigma_{n,c}^{-1})$. In the hierarchical Bayes setup we augment the Bayesian model (2.2) with a prior on c. We then take $\eta_1, \eta_2 \in (0,1)$ and select a pair of (nontrivial) quantiles $\hat{c}_{1,n}(\eta_1) < \hat{c}_{2,n}(\eta_1)$ in the posterior distribution of c, i.e. such that the posterior $c\,|\,\vec{Y}_n$ assigns mass η_1 to the interval $\big[\hat{c}_{1,n}(\eta_1), \hat{c}_{2,n}(\eta_1)\big]$. We then consider as credible sets for f:

$$\hat{C}_{n,\eta,M} = \bigcup_{\hat{c}_{1,n}(\eta_1)<c<\hat{c}_{2,n}(\eta_1)} \big\{f : \|\vec{f}_n - \hat{f}_{n,c}\| < M r_n(c, \eta_2)\big\}. \tag{2.8}$$

This two-step construction can exploit that the credible sets for fixed c have a simple description through the radii $r_n(c,\eta)$. An alternative would be a ball around the hierarchical posterior mean $\int \hat{f}_{n,c}\,\Pi_n(dc\,|\,\vec{Y}_n)$.

The uncertainty quantification, by either (2.7) or (2.8), is deemed accurate if the sets $\hat{C}_{n,\eta,M}$ cover the true parameter f with high probability, if the data are generated according to the model (2.1). In particular, the credible sets are *honest confidence sets* at level η for a given class of functions $\mathcal{F}$ if

$$\inf_{f\in\mathcal{F}} P_f\big(f \in \hat{C}_{n,\eta,M}\big) \geq \eta.$$

The number $r_n(c,\eta)$ is the natural radius of the credible set for fixed c at level η in the Bayesian framework. The additional constant M in the definitions (2.7)–(2.8) of the credible sets is required because the Bayesian and frequentist notions of coverage are not the same, and c is estimated.

It is well known that the size of an honest confidence set for a given model $\mathcal{F}$ is determined by "worst case" members of $\mathcal{F}$ [Low, 1997, Juditsky and Lambert-Lacroix, 2003, Cai and Low, 2004, 2006, Robins and van der Vaart, 2006, Genovese and Wasserman, 2008, Hoffmann and Nickl, 2011]. For instance, if $\mathcal{F}$ contains a Hölder ball of regularity α, then the (random) diameter of the confidence set cannot be of smaller order than $\sqrt{n}\,n^{-\alpha/(2\alpha+1)}$, even if the true function is much smoother. In other words, the size of honest confidence sets cannot *adapt* to the unknown smoothness of the true regression function. On the other hand, the posterior contraction rate of the hierarchical Bayes method is known to adapt to unknown regularity, in that the rate is faster if the true function is smoother. We show below that the empirical Bayes methods adapt in a similar manner. Since the corresponding credible sets will have diameter of order the contraction rate, it follows that these sets cannot be honest over a "full" set of functions, such as a Hölder ball. Following [Giné and Nickl, 2010, Bull, 2012, Szabó et al., 2015] we lower our expectation and investigate honesty over a reduced parameter space, with certain "inconvenient" true parameters cut out, as follows.

The distribution of the data depends on the function f only through the vector $\vec{f_n}$. A convenient way to describe this vector is through its coordinates relative to the eigenbasis of the covariance matrix U_n. Write $f_{1,n},\dots,f_{n,n}$ for the coordinates of $\vec{f_n}$ relative to this basis, i.e.

$$f_{j,n} := \vec{f_n}^T e_{j,n}, \qquad j \in \{1,\dots,n\},$$

for $e_{1,n},\dots,e_{n,n}$ the orthonormal eigenbasis of U_n. Let $\lambda_{1,n},\dots,\lambda_{n,n}$ be the corresponding eigenvalues.

Definition 2.1 (Discrete polished tail)**.** We say that the function f, or the corresponding array $(f_{j,n})$, satisfies the *polished tail condition* if there exist constants L and ρ such that for all $c > 0$ and sufficiently large n it holds that

$$L \sum_{j:\rho\leq c\lambda_{j,n}\leq 1} f_{j,n}^2 \geq \sum_{j:c\lambda_{j,n}\leq 1} f_{j,n}^2. \tag{2.9}$$

The condition may be paraphrased as requiring that the "energy" of the signal f in the "large frequencies" $\{j : \rho \leq c\lambda_{j,n} \leq 1\}$ is at least a fraction L^{-1} of the "energy" in the "frequencies" $\{j : c\lambda_{j,n} \leq 1\}$. Perhaps a better name would be "self-similar", but this name is already taken in the literature for a more special property. The following example shows that the condition is similar to the polished tail condition introduced in [Szabó et al., 2015] when the eigenvalues decrease polynomially in j.

Example 2.2 (Polynomial eigenvalues)**.** If $\lambda_{j,n} \asymp K_n/j^k$, for some constants K_n and $k > 0$, then the discrete polished tail condition is equivalent to the existence of constants L and ρ such that, for all sufficiently large m (and hence sufficiently large n),

$$\sum_{j=m}^{n} f_{j,n}^2 \leq L \sum_{j=m}^{\rho m \wedge n} f_{j,n}^2. \tag{2.10}$$

Indeed, the condition $c\lambda_{j,n} \leq 1$ is equivalent to $j \geq (cK_n)^{1/k} =: J$, whence the right side of (2.9) is bounded above by $\sum_{j \geq J} f_{j,n}^2$, which is bounded above by $L \sum_{J \leq j \leq J\rho} f_{j,n}^2$ by (2.10). This is the left side of (2.9), with ρ^{-k} instead of ρ. In [Szabó et al., 2015] a condition similar to (2.10) is introduced in a continuous time setup. We comment on the relationship of these conditions in Section 2.4.

The main result of this chapter is that all three types of credible sets are honest confidence sets over polished tail parameters, of diameter that adapts to the smoothness of f. We measure smoothness through the square norms, for $\alpha > 0$,

$$\begin{aligned} \|f\|_{n,\alpha}^2 &= \frac{1}{n} \sum_{j=1}^{n} j^{2\alpha} f_{j,n}^2, \\ \|f\|_{n,\alpha,\infty}^2 &= \frac{1}{n} \sup_{1 \leq j \leq n} j^{1+2\alpha} f_{j,n}^2. \end{aligned} \tag{2.11}$$

These norms are in terms of the restriction of f to the grid $(x_{j,n})$. We comment on their relationship to norms on the full function f in Section 2.4. (In general the coefficients $f_{j,n}$ cannot be directly related to an infinite sequence of Fourier coefficients of f, but for many functions the numbers $f_{j,n}/\sqrt{n}$, which include the scaling factor $\sqrt{n}$, is close to the j^{th} Fourier coefficient.)

In the following theorem we assume that there exist constants $0 < \underline{\delta} \leq \overline{\delta} < \infty$ and $m \geq 1$ such that the eigenvalues $\lambda_{1,n}, \ldots, \lambda_{n,n}$ of U_n satisfy

$$\underline{\delta}\, \frac{n}{j^m} \leq \lambda_{j,n} \leq \overline{\delta}\, \frac{n}{j^m}. \tag{2.12}$$

Since $\vec{W}_n$ is distributed as $\sum_{j=1}^n \sqrt{\lambda_{j,n}} Z_{j,n} e_{j,n}$ for i.i.d. standard normal random variables $Z_{j,n}$, we have $\mathrm{E}\|W\|_{n,\alpha}^2 = n^{-1} \sum_{j=1}^n j^{2\alpha} \lambda_{j,n}$. For the eigenvalues (2.12) this is uniformly bounded if and only if $\alpha < (m-1)/2$. Thus these eigenvalues correspond to modelling the regression function a-priori as "almost $(m-1)/2$-smooth".

Let $\mathcal{F}_{n,L}$ be the set of all functions that satisfy the discrete polished tail condition (2.10) for given L and satisfy $\sum_{j=1}^n f_{j,n}^2 \leq dn$ for some sufficiently small constant d (that may depend on $\underline{\delta}$ and m).

Theorem 2.3. *Assume that (2.12) holds. For sufficiently large M and any $\eta > 0$ the credible sets (2.7), with $\hat{c}_n$ given by (2.4) or (2.6), and the credible sets (2.8) satisfy*

$$\inf_{f \in \mathcal{F}_{n,L}} P_f(f \in \hat{C}_{n,\eta,M}) \to 1.$$

Furthermore, for any $\alpha \in (0, m/2)$, the diameter of the credible sets $\hat{C}_{n,\eta,M}$ relative to the scaled Euclidean norm $\|\cdot\|_{n,0}$ is of the order $O_{P_f}\big(n^{-\alpha/(1+2\alpha)}\big)$, uniformly in f with $\|f\|_{n,\alpha} \lesssim 1$ or $\|f\|_{n,\alpha,\infty} \lesssim 1$. For the risk-based empirical Bayes method this is even true for $\alpha \in (0, m)$.

The theorem is a summary of the main results of the chapter as valid for all three methods. More specific results for the individual methods, with relaxations of the polished tail condition tailored to the specific method, as well as results that do not assume the eigenvalue condition (2.12), are described below. For example, these results cover functions f on a two-dimensional domain with eigenvalues of the forms (2.19) or (2.20), as introduced below.

The second and third assertions of the theorem show that the diameter of the credible sets adapts to the regularity of the true regression function. The restrictions to regularity levels $\alpha < m/2$ or $\alpha < m$ in the likelihood-based and risk-based methods stem from the prior, through the rate of decrease (2.12) of its eigenvalues, and the method used. The range $(0, m)$ is bigger than could be expected from the existing literature on Gaussian process priors. For instance, $(m/2-1)$-fold integrated Brownian motion satisfies (2.12) and has sample paths of regularity $m/2 - 1/2$. It has been documented to be an appropriate prior for functions of exactly regularity $m/2 - 1/2$, and to become appropriate for functions of regularities $\alpha \in (0, m/2]$ after appropriate (deterministic) scaling (see [van der Vaart and van Zanten, 2007, Knapik et al., 2011] and Chapter 1). The latter property is retained under random scaling by likelihood-based empirical Bayes and hierarchical Bayes methods considered in the present context (although for $\alpha = m/2$ an extra logarithmic factor may come in; see Example 2.23; the definitions of regularity in the various papers are also not directly comparable). Surprisingly the risk-based method performs better than the likelihood-based methods, in that it enlarges the good range to $\alpha \in (0, m)$. This is caused by the closer connection of the risk-based empirical Bayes method to the diameter of the credible set, yielding a more appropriate scaling factor $\hat{c}_n$ for minimizing this diameter.

The diameter of the credible sets is linked to the posterior contraction rate. The rates $O_{P_f}(n^{-\alpha/(1+2\alpha)})$ are attained irrespective of f satisfying the polished tail condition, the latter condition being important only for the coverage.

The credible sets (2.7) and (2.8) are obtained by considering balls in the space of function values of f at the design points. An alternative are (sets based on) pointwise intervals of the form

$$\hat{C}_{n,\eta,M}(x) = \{f : |f(x) - \hat{f}_{n,\hat{c}_n}(x)| < M r_n(\hat{c}_n, \eta, x)\} \tag{2.13}$$

or

$$\hat{C}_{n,\eta,M}(x) = \bigcup_{\hat{c}_{1,n}(\eta_1) < c < \hat{c}_{2,n}(\eta_1)} \{f : |f(x) - \hat{f}_{n,c}(x)| < M r_n(c, \eta_2, x)\}, \tag{2.14}$$

where $\hat{f}_{n,c}(x)$ denotes the mean of the marginal posterior distribution of $f(x)$ given c and $r_n(c, \eta, x)$ is determined so that

$$P\big(|f(x) - \hat{f}_{n,c}(x)| < r_n(c, \eta, x) \mid \vec{Y}_n, c\big) = \eta.$$

Since this marginal posterior distribution of $f(x)$ given c is normal with mean $\hat{f}_{n,c}(x)$, these intervals are easily determined. In particular, for a design point $x = x_{i,n}$ the radius $r_n(c, \eta, x)$ is equal to $z_\eta(1 - (\Sigma_{n,c}^{-1})_{i,i})^{1/2}$, for z_η the $(1+\eta)/2$-quantile of the standard normal distribution. When used simultaneously for multiple values of x, these intervals form a *credible band.*

The study of the coverage of such pointwise intervals and bands requires different techniques from those in this chapter, and appears to be tractable only for concretely specified prior processes. However, the methods developed here are suitable when measuring coverage in an averaged fashion that focuses on the fraction of the design points at which the intervals (2.13) or (2.14) cover the true function. A similar point of view was taken by [Wahba, 1983, Cai et al., 2014]. The following corollary gives such a result for a subset of design points $x_{i,n}$ that are spread evenly relative to the prior process. More precisely, let

$$s_n^2(c, x_{i,n}) := \inf_{a \in \mathbb{R}^n} \Big[c\,\mathrm{E}\big(W_{x_{i,n}}^1 - a^T \vec{W}_n^1\big)^2 + \|a\|^2\Big]$$

denote the posterior variance at the design point $x_{i,n}$ and set

$$J_n := \Big\{i : s_n^2(c, x_{i,n}) \geq \frac{C}{n} \sum_{j=1}^n s_n^2(c, x_{j,n})\Big\} \tag{2.15}$$

for some constant C that is independent of n. Then the corollary holds when considering the design points in this set.

In Corollary 1.11 of Chapter 1, we have seen that Brownian motion satisfies this condition for the set of all design points that satisfy $x_{i,n} \geq C/\sqrt{\log n}$.

The following corollary shows that the uncertainty quantification through the intervals $\hat{C}_{n,\eta,M}(x_{i,n})$ is correct at the design points in the set J_n as long this set is large enough, except possibly a fraction.

Corollary 2.4. *Assume that (2.12) holds and that the set J_n given in (2.15) satisfies $|J_n| \sim n$. Fix $\gamma \in (0,1)$, $\eta > 0$ and let $\hat{c}_n$ be given by (2.4) or (2.6). Then for sufficiently large M the credible sets defined in either (2.13) or (2.14) satisfy*

$$\inf_{f \in \mathcal{F}_{n,L}} P_f\Big(\frac{1}{n}\sum_{i \in J_n} 1\big\{f \in \hat{C}_{n,\eta,M}(x_{i,n})\big\} \geq \gamma\Big) \to 1.$$

Furthermore, if for some constant $C' > 0$ it also holds that $s_n^2(c, x_{i,n}) \leq \frac{C'}{n}\sum_{j=1}^n s_n^2(c, x_{j,n})$ for $i \in J_n$, then for any $\alpha \in (0, m/2)$ the length of the intervals $\hat{C}_{n,\eta,M}(x_{i,n})$ is of the order $O_{P_f}\big(n^{-\alpha/(1+2\alpha)}\big)$ uniformly in $i \in J_n$, uniformly in f with $\|f\|_{n,\alpha} \lesssim 1$ or $\|f\|_{n,\alpha,\infty} \lesssim 1$. For the risk-based empirical Bayes method this is even true for $\alpha \in (0, m)$.

The proof of this corollary can be found in Section 2.6.

The multiplicative constant n in (2.12) is motivated by comparison with the continuous time setup. If the covariance function $K(s,t) = \mathrm{E}W_s^1 W_t^1$ of the continuous time process W^1 has eigenfunctions e_j satisfying

$$\int K(s,t)e_j(t)\,dt = \lambda_j e_j(s),$$

then for equidistant design points one may expect that

$$\sum_{i=1}^n K(x, x_{i,n})e_j(x_{i,n}) \approx n\lambda_j e_j(x).$$

This suggests both that $\lambda_{j,n} \approx n\lambda_j$ and that the "discrete" eigenvectors $e_{j,n}$ should be close to the eigenfunctions restricted to the design points. This is a suggestion only, which already makes little sense when counting the numbers of eigenvalues involved: n versus ∞. Nevertheless, for the Brownian motion prior the correspondence is exact.

Example 2.5 (Brownian motion)**.** The Brownian motion prior permits explicit formulas for eigenbasis and eigenvalues, provided the design points are taken equal to $x_{i,n} = i/(n+1/2)$ for $i \in \{1,\dots,n\}$, a slight shift from the usual uniform grid. The formulas are interesting as they allow to make a connection to the Fourier basis (see Section 2.4).

The eigenvectors of the covariance matrix U_n of standard Brownian motion scaled to unit length are given by

$$\begin{aligned} e_{j,n} &= \frac{1}{\sqrt{n+1/2}}\big(e_j(x_{1,n}),\dots,e_j(x_{n,n})\big)^T, \\ e_j(x) &= \sqrt{2}\sin\big[\big(j-\tfrac{1}{2}\big)\pi x\big] \end{aligned} \tag{2.16}$$

for $j \in \{1,\dots,n\}$. Here the functions e_j are an orthonormal basis of $\{f \in L_2[0,1] : f(0) = 0\}$, and happen to be eigenfunctions of the covariance kernel of continuous Brownian motion. A similar correspondence is valid for Brownian bridge, but we are not aware of other examples where the continuous and discrete setups match up so closely.

The eigenvalues of U_n are given by

$$\lambda_{j,n} = \frac{1}{(4n+2)\sin^2\big((j-1/2)\pi/(2n+1)\big)}.$$

As the argument of the sine is in $[0,\pi/2]$, for which $2x/\pi \le \sin x \le x$, there exist numbers $(\underline{\delta},\overline{\delta})$ such that

$$\frac{\underline{\delta} n}{j^2} \le \frac{1}{(4n+2)\pi^2}\Big(\frac{2n+1}{j-\frac{1}{2}}\Big)^2 \le \lambda_{j,n} \le \frac{1}{16n+8}\Big(\frac{2n+1}{j-\frac{1}{2}}\Big)^2 \le \frac{\overline{\delta} n}{j^2}, \tag{2.17}$$

where this inequality holds for all n and $j \ge 1$ if we take $(\underline{\delta},\overline{\delta}) = (\pi^{-2},3)$ and for $j > 2$ and n sufficiently large if we let $\overline{\delta} = 4/10$.

Standard Brownian motion has sample paths of regularity 1/2, and has been documented to become an appropriate prior for functions of regularities $\alpha \in (0,1)$ after appropriate scaling (see Chapter 1 and [van der Vaart and van Zanten, 2007, Knapik et al., 2011]). We show in this chapter that the good range is enlarged to $\alpha \in (0,2)$ provided that the scaling by the risk-based empirical Bayes method is used.

Example 2.6 (Discrete priors)**.** Although it often helps intuition to model a function f a-priori by a Gaussian process on a "continuous" space that encompasses the design points, nothing in the preceding setup requires this. In fact, we may turn the construction around, by starting with an arbitrary orthonormal basis $e_{1,n},\dots,e_{n,n}$ and eigenvalues $\lambda_{1,n},\dots,\lambda_{n,n}$, and next define the prior covariance matrix U_n to be the matrix that has this as its eigenbasis and eigenvalues, that is, its spectral decomposition is

$$U_n = \sum_{i=1}^{n} \lambda_{i,n} e_{i,n} e_{i,n}^T. \tag{2.18}$$

Given arbitrary points $x_{1,n}, \ldots, x_{n,n}$ the vector $\vec{f_n}$ is then a-priori modelled by its coefficients $f_{i,n}$ relative to $e_{1,n}, \ldots, e_{n,n}$, which are independent $\mathcal{N}(0, c\lambda_{i,n})$-variables.

One particular example is to retain the eigenvectors of Brownian motion, but to change the corresponding eigenvalues to (2.12) for a general m. The interpretation of the norms $\|\cdot\|_{n,\alpha}$ and $\|\cdot\|_{n,\alpha,\infty}$ would be the same as for Brownian motion (as discussed in Section 2.4), but the good rates relative to these norms would now be attained for α up to m (or $m/2$) rather than 2 (or 1). Our theoretical results show only advantages to taking a larger value of m, but one might guess that a deeper analysis could change this picture.

Example 2.7 (Discrete Laplacian)**.** The *discrete Laplacian* is a useful tool to construct "smooth priors" on a discrete set of design points. For a univariate grid it is closely connected to the Brownian motion prior of Example 2.5. For a countable set $\mathcal{X}$ equipped with a neighbourhood relation $\sim$ the Laplacian is the operator acting on functions $f : \mathcal{X} \to \mathbb{R}$, defined by

$$L(f)(x) = \sum_{y: y \sim x} \bigl[f(y) - f(x)\bigr].$$

Small values of $|Lf|$ indicate that f changes little across its neighbourhoods, whence L can be used to model smoothness relative to the given neighbourhood structure.

Identification of a function $f : \mathcal{X} \to \mathbb{R}$ with the infinite vector $\bigl(f(x) : x \in \mathcal{X}\bigr)$ gives an identification of L with an infinite matrix (with $(x, y)^{\text{th}}$ element equal to 1 if $y \neq x$ and $y \sim x$; equal to $-\#\{y \sim x\}$ if $y = x$; and equal to 0 otherwise). The restriction of this matrix to the rows $x \in \{x_{1,n}, \ldots, x_{n,n}\}$ will have nonzero elements in columns $y \notin \{x_{1,n}, \ldots, x_{n,n}\}$ with $y \sim x_{i,n}$ for some i, and hence a restriction of Lf to the design points may not correspond to simply taking the appropriate $(n \times n)$-submatrix of L. This is typically solved by imposing boundary conditions, much as when considering a continuous partial differential operator.

In the example of $\mathcal{X} = \mathbb{Z}$ with the design points $x_{1,n}, \ldots, x_{n,n}$ identified with the points $1, \ldots, n$ and the neighbourhood system: $i \sim j$ if and only if $|i - j| = 1$, the discrete Laplacian is

$$L(f)(i) = \sum_{j: |j-i|=1} \bigl[f(j) - f(i)\bigr] = f(i+1) + f(i-1) - 2f(i).$$

The restriction of $L(f)$ to the design points $1, \ldots, n$ also involves the points 0 and $n+1$, and there are various ways of imposing boundary conditions. The natural choice $f(0) = f(n+1) = 0$ is known as the *Dirichlet boundary*, while the other natural choice given by $f(0) = f(1)$ and $f(n+1) = f(n)$ is the *Neumann boundary*. The eigenvectors and eigenvalues corresponding to these boundary conditions are known explicitly, and so they are for the mixed Dirichlet-Neumann conditions: $f(0) = 0$ and $f(n+1) = f(n)$. In fact, in the latter case the eigenvectors are exactly equal to $e_{j,n}$ as given in (2.16) and the eigenvalues are $-1/((n+1/2)\lambda_{j,n})$ for $\lambda_{j,n}$ as given in (2.17). This close connection to Brownian motion is not obvious, but also not entirely surprising as minus the inverse Laplacian (the twofold primitive) is the covariance operator of Brownian motion (restricted to the orthogonal complement of the constant functions) and standard Brownian motion is tied at zero. The connection invites to interpret the eigenvectors (2.16) as modelling smoothness in a discrete sense, an interpretation that also makes sense if the design points $x_{i,n}$ are linearly ordered and roughly equally spaced, but not exactly equal to $i/(n+1/2)$ as in Example 2.5. For the special grid of the latter example the norm in (2.11) corresponds exactly to the size measured by the Laplacian, in that

$$\frac{1}{n}\,\|(n^2L)^{\alpha/2}\vec{f}_n\|^2 = n^{2\alpha-1}\sum_{i=1}^{n}\frac{f_{i,n}^2}{\big((n+1/2)\lambda_{i,n}\big)^{\alpha}} \asymp \|f\|_{n,\alpha}^2.$$

(The norm on the left side is the Euclidean norm of $\mathbb{R}^n$ and the leading factor $1/n$ stabilises the sum involved in this norm; the factor n^2 preceding L corresponds to $1/h^2$, for $h \sim 1/n$ the mesh width of the grid.) Although the eigenvalues (2.17) come naturally with the discrete Laplacian, when defining the prior they might be replaced by eigenvalues (2.12) for a general m. This would correspond to describing a-priori smoothness by a power of the Laplacian. Indeed, as noted following (2.12), for these eigenvalues we have $\mathrm{E}\|W\|_{n,\alpha}^2 < \infty$ for $\alpha < (m-1)/2$. In view of the preceding display, this is equivalent to finiteness of $\frac{1}{n}\,\mathrm{E}\|(n^2L)^{\alpha/2}\vec{W}_n\|^2$. So the prior with covariance matrix (2.18), for eigenvalues (2.12) and eigenvectors (2.16), corresponds to modelling f by a Gaussian process W with finite discrete Laplacian $(n^2L)^{\alpha/2}W$ for $\alpha < (m-1)/2$.

Example 2.8 (Integrated Brownian motion)**.** Once integrated Brownian motion $W_t^1 = \int_0^t B_s\,ds$, for B standard Brownian motion, possesses covariance function $\mathrm{cov}(W_s^1, W_t^1) = s^2(3t-s)/6$ for $s \le t$. The eigenfunctions are given by

$$\begin{aligned} e_j(t) \propto &(\sin\theta_j + \sinh\theta_j)\big(\cos(t\theta_j) - \cosh(t\theta_j)\big)\\ &- (\cos\theta_j + \cosh\theta_j)\big(\sin(t\theta_j) - \sinh(t\theta_j)\big), \end{aligned}$$

where the θ_j are the positive roots of the equation $\cos(\theta)\cosh(\theta) = -1$, for $j \in \{1, 2, \ldots\}$. See [Freedman, 1999], Theorem 7. The corresponding eigenvalues are $\lambda_j = \theta_j^{-4}$ and are of the order $((2j-1)\pi/2)^{-4}$.

Thus this example appears to satisfy (2.12) with $m = 4$. However, exact expressions for the discrete eigenvectors and eigenvalues appear not known.

Example 2.9 (Functions of two arguments)**.** Functions $f : [0,1]^2 \to \mathbb{R}$ on the unit square may be modelled a-priori by the product $W^1_{s,t} = B_{1,s}B_{2,t}$ of two independent standard Brownian motions B_1 and B_2. The covariance function $\mathrm{E}W^1_{s,t}W^1_{s',t'}$ is the product $K(s,s')K(t,t')$ of the covariance functions $K(s,s') = s \wedge s'$ of the Brownian motions. While this process is not Gaussian, we can replace it by a Gaussian process with the same covariance structure. For a rectangular grid consisting of points $(x_{i,n}, x_{j,n})$ constructed from a given univariate grid $0 \le x_{1,n} < \cdots < x_{n,n} \le 1$, the covariance matrix of the n^2-dimensional vector $(W_{x_{i,n},x_{j,n}})$, for $(i,j) \in \{1, \ldots, n\}^2$, with its coordinates ordered appropriately, is the Kronecker product of two copies of the covariance matrix of the n-dimensional vector $(B_{x_{i,n}})$. The eigenvectors are the tensor products $e_{i,n} \otimes e_{j,n}$ of the univariate eigenvectors $e_{i,n}$, with corresponding eigenvalues the products $\lambda_{i,j,n} = \lambda_{i,n}\lambda_{j,n}$ of the univariate eigenvalues $\lambda_{i,n}$.

Even though in this case the eigenfunctions and eigenvalues are more naturally viewed as a two-dimensional array than a sequence, they may of course be ordered in a sequence. Then this example fits the general setup, except that n has been changed into n^2.

In particular, for the grid in Example 2.5 the eigenvectors are the discretisations of the tensor products of the sine-basis given in (2.16) and the eigenvalues satisfy

$$\lambda_{i,j,n} \asymp \frac{n^2}{i^m j^m}, \qquad (i,j) \in \{1, \ldots, n\}^2 \tag{2.19}$$

for $m = 2$. Theorem 2.3, which assumes (2.12), does not apply to this example. However, the assumptions of the general results below are satisfied, also for a general value of $m \ge 1$, and hence the message of the theorem goes through. The set of polished tail functions can be defined in the same manner by (2.10), after ordering the array of coefficients $f_{i,j,n}$ in a sequence by order of decreasing eigenvalues $\lambda_{i,j,n}$ (that is, increasing values of ij).

The square smoothness norm $\|\cdot\|_{n,\alpha}$ as in (2.11) now becomes $n^{-2}\sum_{i=1}^n \sum_{j=1}^n (ij)^{2\alpha} f_{i,j,n}^2$. While the eigenbasis is essentially the natural two-dimensional Fourier basis, the restriction imposed by this norm is a bit unusual, in its focus on the cross product ij. As the smoothness norm describes the prior process, this may be unsatisfactory. More natural "Sobolev norms" $n^{-2}\sum_{i=1}^n\sum_{j=1}^n(i^2+j^2)^\alpha f_{i,j,n}^2$ correspond to the eigenvalues

$$\lambda_{i,j,n}\asymp\frac{n^2}{(i^2+j^2)^m},\qquad (i,j)\in\{1,\ldots,n\}^2. \tag{2.20}$$

The Gaussian process W^1 corresponding to these eigenvalues has $\mathrm{E}\|W^1\|_{n^2,\alpha}^2<\infty$ for every $\alpha<m-1$, and hence may be considered "Sobolev smooth almost of order $m-1$".

For these eigenvalues the discrete polished tail condition (2.9) can be written in the form

$$\sum_{\substack{i=1\\ i^2+j^2\geq m}}^n\sum_{j=1}^n f_{i,j,n}^2\leq L\sum_{\substack{i=1\\ m\leq i^2+j^2\leq\rho m}}^n\sum_{j=1}^n f_{i,j,n}^2,$$

for sufficiently large m. The theorems below show that the credible sets corresponding to this prior cover functions that satisfy this condition.

Organisation of the chapter

We give an outline of the argument as presented in Section 2.2. Firstly, the various parts of the criterion functions used to define the estimators $\hat{c}_n$ in (2.4) and (2.6) are studied. The behaviour of these estimators is then quantified by Theorem 2.12. This result is applied in Theorem 2.17 to obtain our main result on the coverage of the credible sets as defined in (2.7). We follow up with results about contraction rates of oracle type and over various concrete models (Section 2.2). The argument in the hierarchical case in Section 2.3 has the same structure: first we quantify the behaviour of the posterior $c\mid\vec{Y}_n$ in Theorem 2.25 and this is then applied in Theorem 2.27 to obtain coverage. Again this is followed by a discussion of the contraction rates in Section 2.3.

The rest of the paper is structured as follows. Section 2.4 concerns the interpretation of the polished tail condition, which is related to a similar condition on the Fourier coefficients of f. It is shown to be satisfied with probability one under the prior. This section also discusses various alternative smoothness assumptions on the function f. Section 2.5 is a closing discussion, which addresses conditions, interpretations, and generalisations of our results. Finally Sections 2.6 and 2.7 gather technical proofs and technical lemmas.

Notation

The notation $a_n\asymp b_n$ means that a_n/b_n is bounded away from 0 and infinity, as $n\to\infty$, and $a_n\sim b_n$ means that a_n/b_n tends to 1. If a_n and b_n are functions, then we say that $a_n\asymp b_n$ or $a_n\sim b_n$ uniformly over a domain if the constants away from 0 and infinity can be chosen the same for every value in the domain, or the convergence to 1 is uniform. The notation $a \lesssim b$ means $a \leq Cb$ for a universal constant C.

For a function $g : \mathcal{X} \to \mathbb{R}$, the vector $\big(g(x_{1,n}), \ldots, g(x_{n,n})\big)$ is denoted by $\vec{g}_n$. The same notational device is used for a vector $\vec{\varepsilon}_n$ composed of variables $\varepsilon_{1,n}, \ldots, \varepsilon_{n,n}$. Unless stated otherwise the set I_n is the interval $I_n = [\log n/n, n^{m-1}]$.

2.2 Empirical Bayes

By substituting the model equation $\vec{Y}_n = \vec{f}_n + \vec{\varepsilon}_n$, we can decompose the quadratic forms in the empirical Bayes criteria (2.4) and (2.6) as

$$\vec{Y}_n^T \Sigma_{n,c}^{-k} \vec{Y}_n = \vec{f}_n^T \Sigma_{n,c}^{-k} \vec{f}_n + \vec{\varepsilon}_n^T \Sigma_{n,c}^{-k} \vec{\varepsilon}_n + 2\vec{f}_n^T \Sigma_{n,c}^{-k} \vec{\varepsilon}_n, \qquad k \in \{1,2\}. \tag{2.21}$$

We next express both $\vec{f}_n$ and $\vec{\varepsilon}_n$ relative to the orthonormal eigenbasis $e_{1,n}, \ldots, e_{n,n}$ of U_n. The coefficients of $\vec{f}_n$ are by their definition the numbers $f_{j,n}$, while the coefficients of $\vec{\varepsilon}_n$ are i.i.d. standard normal variables $Z_{j,n}$. The matrix $\Sigma_{n,c} = I + cU_n$ and its inverses $\Sigma_{n,c}^{-1}$ and $\Sigma_{n,c}^{-2}$ have the same eigenbasis as U_n, with eigenvalues $(1 + c\lambda_{j,n})$, $(1 + c\lambda_{j,n})^{-1}$ and $(1 + c\lambda_{j,n})^{-2}$, respectively, for $\lambda_{j,n}$ the eigenvalues of U_n. It follows that the two types of empirical Bayes estimators $\hat{c}_n$ minimize criteria L_n^L and L_n^R of the form

$$\begin{aligned} L_n(c,f) &:= D_{1,n}(c,f) + D_{2,n}(c) + R_{1,n}(c,f) + R_{2,n}(c) \\ &= D_n(c,f) + R_n(c,f). \end{aligned} \tag{2.22}$$

For the risk-based empirical Bayes estimator (2.6) the functions and processes $D_{1,n}, D_{2,n}, R_{1,n}$ and $R_{2,n}$ on the right side are defined by

$$\begin{aligned} D_{1,n}^R(c,f) &= \vec{f}_n^T \Sigma_{n,c}^{-2} \vec{f}_n = \sum_{j=1}^n \frac{f_{j,n}^2}{(1 + c\lambda_{j,n})^2}, \\ D_{2,n}^R(c) &= \mathrm{tr}\big((I - \Sigma_{n,c}^{-1})^2\big) = \sum_{j=1}^n \frac{(c\lambda_{j,n})^2}{(1 + c\lambda_{j,n})^2}, \\ R_{1,n}^R(c,f) &= 2\vec{f}_n^T \Sigma_{n,c}^{-2} \vec{\varepsilon}_n = 2\sum_{j=1}^n \frac{Z_{j,n} f_{j,n}}{(1 + c\lambda_{j,n})^2}, \\ R_{2,n}^R(c) &= \vec{\varepsilon}_n^T \Sigma_{n,c}^{-2} \vec{\varepsilon}_n - \mathrm{tr}(\Sigma_{n,c}^{-2}) - \sum_{j=1}^n (Z_{j,n}^2 - 1) \\ &= \sum_{j=1}^n (Z_{j,n}^2 - 1)\Big[\frac{1}{(1 + c\lambda_{j,n})^2} - 1\Big], \end{aligned} \tag{2.23}$$

whereas for the likelihood-based empirical Bayes estimator (2.4) these functions and processes are given by

$$\begin{aligned}
D^L_{1,n}(c,f) &= \vec{f}_n^T \Sigma_{n,c}^{-1} \vec{f}_n = \sum_{j=1}^n \frac{f_{j,n}^2}{1+c\lambda_{j,n}},\\
D^L_{2,n}(c) &= \log\det\Sigma_{n,c} - \operatorname{tr}\bigl(I - \Sigma_{n,c}^{-1}\bigr) = \sum_{j=1}^n \Bigl[\log(1+c\lambda_{j,n}) - \frac{c\lambda_{j,n}}{1+c\lambda_{j,n}}\Bigr],\\
R^L_{1,n}(c,f) &= 2\vec{f}_n^T \Sigma_{n,c}^{-1} \vec{\varepsilon}_n = 2\sum_{j=1}^n \frac{Z_{j,n} f_{j,n}}{1+c\lambda_{j,n}}, \qquad (2.24)\\
R^L_{2,n}(c) &= \vec{\varepsilon}_n^T \Sigma_{n,c}^{-1} \vec{\varepsilon}_n - \operatorname{tr}(\Sigma_{n,c}^{-1}) - \sum_{j=1}^n (Z_{j,n}^2 - 1)\\
&= -\sum_{j=1}^n \frac{(Z_{j,n}^2 - 1)c\lambda_{j,n}}{1+c\lambda_{j,n}}.
\end{aligned}$$

In general discussions we shall leave off the superscripts R and L, for "Risk" and "Likelihood", and denote both the risk- and likelihood-based functions by $D_{1,n}, D_{2,n}, R_{1,n}, R_{2,n}$. In both cases we have shifted the criteria by the factor $\sum_{j=1}^n (Z_{j,n}^2 - 1)$, which does not depend on c, in order that the remainder term $R_{2,n}$ be smaller.

The functions $D_{1,n}$ and $D_{2,n}$ are deterministic, whereas $R_{1,n}$ and $R_{2,n}$ are random processes. The processes $D_{1,n}$ and $R_{1,n}$ depend on f, whereas the other processes are free of the parameter. Even though the functions and processes differ in the risk- and likelihood-based cases, for instance by the power of $1 + c\lambda_{j,n}$ in the denominators, the two estimators $\hat{c}_n$ can be analysed by similar methods. In Lemma 2.14 it will be seen that under (2.12) the two functions $D_{2,n}$, even though quite different in form, are asymptotically equivalent. The following proposition shows that in both cases the stochastic process R_n is negligible relative to the deterministic process D_n.

Proposition 2.10. *If (2.12), (2.19) or (2.20) holds, then for $R_{1,n}$ and $R_{2,n}$ as given in (2.23) or (2.24) and the corresponding $D_n = D_{1,n} + D_{2,n}$ in the same display it holds that*

$$\sup_{c\in I_n} \frac{|R_{1,n}(c,f)| + |R_{2,n}(c)|}{D_n(c,f)} \overset{P_f}{\to} 0. \qquad (2.25)$$

The proof of the proposition can be found in Section 2.6. In case of the eigenvalues (2.19) or (2.20), it should be understood that n is replaced by n^2 in the assertion (and the single sums in (2.23) or (2.24) by double sums).

We may view the stochastic process $R_n = R_{1,n} + R_{2,n}$ in (2.23) or (2.24) as an "estimation error" when estimating an "ideal" criterion $D_n = D_{1,n} + D_{2,n}$. The preceding proposition essentially says that this error can be ignored. As a consequence the minimizer $\hat{c}_n$ of $L_n = D_n + R_n$ will behave similarly to the (deterministic) minimizer of D_n. The latter functions consists of a part $D_{1,n}(\cdot, f)$ that is decreasing in c, from $D_{1,n}(0, f) = \sum_{j=1}^n f_{j,n}^2$ to $D_{1,n}(\infty, f) = 0$, and a part $D_{2,n}$ that is free of f and is strictly increasing in c, from $D_{2,n}(0) = 0$ to $D_{2,n}(\infty) \geq n$. Minimizing the sum D_n of these functions can be viewed as an attempt to balance these two terms.

In the case of the risk-based empirical Bayes method $D_{1,n}(c, f)$ is exactly the square bias of the posterior mean at the true regression function f, given a fixed scale c, and $D_{2,n}(c)$ is its variance, which is independent of f (see (2.5)). The square bias is decreasing in the scale c, while the variance is increasing, and hence the empirical Bayes estimator $\hat{c}_n$ tries to balance the square bias and variance by minimizing an estimate of their sum. The likelihood-based empirical Bayes estimator is not as strongly tied to the risk, but we shall see that it performs in a similar manner. Here the essence will be that its bias term $D_{1,n}$ is bigger than the bias term of the risk-based method, while its variance term has the same order of magnitude.

For minimizing the risk the empirical Bayes methods always do the right thing. However, the coverage of the credible sets depends not on the sum of square bias and variance, but on their relationship, or rather the relationship between square bias and the *posterior variance*

$$s_n^2(c) = \mathrm{E}\big(\|\vec{f}_n - \hat{f}_{n,c}\|^2 \,|\, \vec{Y}_n, c\big) = \mathrm{tr}(I - \Sigma_{n,c}^{-1}) = \sum_{j=1}^n \frac{c\lambda_{j,n}}{1 + c\lambda_{j,n}}. \tag{2.26}$$

If for a particular f the square bias exceeds the posterior variance, then the empirical Bayes method will put a too narrow credible set too far from the truth, which it will not cover in that case. The posterior variance, although not equal to the variance terms $D_{2,n}$, has the same order of magnitude as these quantities (see Lemma 2.14). Thus a lack of coverage is caused by too small a value of $\hat{c}_n$, giving too small a prior variance and posterior variance, i.e. by "oversmoothing" the truth.

Notwithstanding the nice properties of the functions $D_{1,n}$ and $D_{2,n}$ for a given n, such oversmoothing may occur for f for which the "bias" function $c \mapsto D_{1,n}(c, f)$ changes haphazardly with n. (We describe this here in an asymptotic framework, with $n \to \infty$, but a problem will arise for every given n, albeit possibly for different f.) The point is that at different sample sizes, different aspects of f determine the behaviour of the empirical Bayes estimators $\hat{c}_n$. The assumption that f satisfies the polished tail condition prevents such haphazard behaviour for both empirical Bayes methods. When considering a given method, good behaviour can also be more precisely characterised through the corresponding function $D_{1,n}$, as follows.

Definition 2.11 (Good bias condition)**.** We say that the function f, or the corresponding array $(f_{j,n})$, satisfies the *good bias condition* relative to $D_{1,n}$ if there exists a constant $a > 0$ such that, for $c \in I_n$,

$$D_{1,n}(Kc, f) \leq K^{-a} D_{1,n}(c, f), \qquad \text{for all } K > 1. \tag{2.27}$$

As a pendant to this condition we call $D_{2,n}$ *good variance functions* if there exist constants $b, B, B' > 0$, independent of n, such that for $c \in I_n$ we have

$$Bk^b D_{2,n}(c) \leq D_{2,n}(kc) \leq B' k^b D_{2,n}(c) \qquad \text{for all} \quad k < 1. \tag{2.28}$$

Since the functions $D_{2,n}$ do not depend on f, the good variance condition merely refers to the prior process. Priors satisfying (2.12) give $D_{2,n}(c) \asymp (cn)^{1/m}$ (see Lemma 2.14) and hence yield good variance functions with $b = 1/m$.

The essence of these "good conditions" is captured in the purely analytical Lemma 2.42 in Section 2.7, which is the basis of the proof of the second assertion of the following theorem.

Theorem 2.12. *Suppose the remainder terms $R_{1,n}$ and $R_{2,n}$ satisfy (2.25). Then for any f and $\varepsilon > 0$ the empirical Bayes estimators $\hat{c}_n$ given in (2.4) and (2.6), with the corresponding function $D_n = D_{1,n} + D_{2,n}$ as given in (2.23) and (2.24), satisfy*

$$P_f\Big(D_n(\hat{c}_n, f) \leq (1+\varepsilon) \inf_{c \in I_n} D_n(c, f)\Big) \to 1.$$

Furthermore, if f satisfies the good bias condition with constant a, $D_{2,n}$ are good variance functions with constants b, B, B' and $\sum_{j=1}^n f_{j,n}^2 \leq \sup_{c \in I_n} D_{2,n}(c)$, then also

$$P_f\Big(D_{1,n}(\hat{c}_n, f) \leq B^{-1}(2+2\varepsilon)^{1+b/a} D_{2,n}(\hat{c}_n)\Big) \to 1.$$

Proof. Let $c_n \in I_n$ be a minimizer of D_n and set $\Lambda_n = \{c \in I_n : D_n(c, f) \leq (1+\varepsilon) D_n(c_n, f)\}$. For the first assertion it suffices to show that $P_f(\hat{c}_n \in \Lambda_n) \to 1$. By the definition of $\hat{c}_n$, this is the case if $\inf_{c \notin \Lambda_n} L_n(c, f)$ is with probability tending to one strictly bigger than $L_n(c_n, f)$. Since $L_n = D_n + R_n$, relation

(2.25) gives

$$\begin{aligned}\inf_{c\notin\Lambda_n} L_n(c,f) &= \inf_{c\notin\Lambda_n}\Big[D_n(c,f)\Big(1+\frac{R_n(c,f)}{D_n(c,f)}\Big)\Big]\\ &\geq \inf_{c\notin\Lambda_n} D_n(c,f)\Big(1-\sup_{c\notin\Lambda_n}\Big|\frac{R_n(c,f)}{D_n(c,f)}\Big|\Big)\\ &\geq \Big[\inf_{c\notin\Lambda_n} D_n(c,f)\Big]\big(1-o_P(1)\big).\end{aligned}$$

By the definition of Λ_n the infimum on the right side is at least equal to $(1+\varepsilon)D_n(c_n,f)$. Moreover, again by Proposition 2.10 we have that $L_n(c_n,f)\leq D_n(c_n,f)\big(1+o_P(1)\big)$. The desired result follows, as $D_n(c_n,f)$ is strictly positive.

For the proof of the second assertion we define $\tilde{c}_n$ as the unique point of intersection of the graphs of the functions $D_{1,n}$ and $D_{2,n}$, i.e. the unique solution of the equation $D_{1,n}(c,f)=D_{2,n}(c)$. If $\tilde{c}_n\in I_n$, then by the first assertion $D_n(\hat{c}_n,f)\leq(1+\varepsilon)D_n(\tilde{c}_n,f)$, whence the assertion follows from Lemma 2.42(i). If $\tilde{c}_n$ falls to the left of I_n, then $D_{1,n}(c,f)\leq D_{2,n}(c)$ throughout I_n by the monotonicity of the two functions and the assertion is trivially true. The assumption that $D_{1,n}(0,f)=\sum_{j=1}^n f_{j,n}^2$ is below the maximum value of $D_{2,n}$ prevents that $\tilde{c}_n$ falls to the right of I_n. □

The good-bias condition on f is dependent on the prior and the method through the function $D_{1,n}$, which can be $D_{1,n}^L$ or $D_{1,n}^R$. For both methods the condition is implied by the discrete polished tail condition.

Lemma 2.13. *Any f that satisfies the discrete polished tail condition also satisfies the good bias condition, for both the risk-based and likelihood-based bias functions $D_{1,n}(\cdot,f)$.*

Proof. If f satisfies the discrete polished tail condition, then

$$\begin{aligned}\sum_{j:c\lambda_{j,n}\leq 1}\frac{f_{j,n}^2}{(1+c\lambda_{j,n})^2} &\leq \sum_{j:c\lambda_{j,n}\leq 1} f_{j,n}^2 \leq L\sum_{j:\rho\leq c\lambda_{j,n}\leq 1} f_{j,n}^2\\ &\leq 4L\sum_{j:\rho\leq c\lambda_{j,n}\leq 1}\frac{f_{j,n}^2}{(1+c\lambda_{j,n})^2},\end{aligned}$$

since $1+c\lambda_{j,n}\leq 2$ for j in the range of the sum. The left side is part of the sum that defines the function $D_{1,n}^R$. Splitting this sum in the parts with $c\lambda_{j,n}\leq 1$

and with $c\lambda_{j,n} > 1$ and noting that $\rho \leq 1$, we see that

$$\begin{aligned} D^R_{1,n}(c,f) &\leq (1+4L) \sum_{j:\rho\leq c\lambda_{j,n}} \frac{f^2_{j,n}}{(1+c\lambda_{j,n})^2} \\ &\leq \frac{(1+4L)(1+\rho)}{\rho} \sum_{j:\rho\leq c\lambda_{j,n}} \frac{f^2_{j,n} c\lambda_{j,n}}{(1+c\lambda_{j,n})^3}, \end{aligned}$$

since $c\lambda_{j,n}/(1+c\lambda_{j,n}) \geq \rho/(1+\rho)$ for j in the range of the sum. The sum on the right side becomes even bigger if we let the sum range from 1 to n and is then equal to $-\frac{1}{2}c\,(D^R_{1,n})'(c,f)$. It follows that there exists $a > 0$ such that

$$\frac{(D^R_{1,n})'(c,f)}{D^R_{1,n}(c,f)} \leq -\frac{a}{c}.$$

Integrating this from c to Kc we find that $\log D^R_{1,n}(Kc,f) - \log D^R_{1,n}(c,f)$ is bounded above by $-a\log K$, and the good bias condition (2.27) follows.

The proof for the likelihood-based function $D^L_{1,n}$ differs only in that the power of the factors $(1+c\lambda_{j,n})^2$ in the denominator must be decreased from 2 to 1. □

The following lemma gives the behaviour of the three variance functions if the eigenvalues satisfy (2.12), (2.19) or (2.20). The lemma implies that these three functions are good variance functions in the sense of (2.28).

Lemma 2.14. *The functions $D^R_{2,n}$ given in (2.23), $D^L_{2,n}$ given in (2.24) and s_n given in (2.26) are strictly increasing on $[0,\infty)$. Furthermore, if (2.12) holds, then*

$$D^R_{2,n}(c) \asymp D^L_{2,n}(c) \asymp s^2_n(c) \asymp (cn)^{1/m},$$

uniformly in c in I_n as $n \to \infty$. The same is true (with n^2 instead of n) under (2.20). Moreover, if (2.19) holds, then

$$D^R_{2,n^2}(c) \asymp D^L_{2,n^2}(c) \asymp s^2_{n^2}(c) \asymp \begin{cases} (cn^2)^{1/m}\big(1+\log(cn^2)\big) & \textit{if } cn^2 \leq n^m \\ (cn^2)^{1/m}\Big(1+\log\Big(\frac{n^{2m}}{cn^2}\Big)\Big) & \textit{if } cn^2 \geq n^m \end{cases}$$

uniformly in c in I_{n^2}.

Proof. The monotonicity of $D^R_{2,n}$ and s_n is clear. Under (2.12) the function $D^R_{2,n}$ satisfies

$$(cn\underline{\delta})^2 \sum_{j=1}^n \frac{1}{(j^m+cn\underline{\delta})^2} \leq D^R_{2,n}(c) \leq (cn\overline{\delta})^2 \sum_{j=1}^n \frac{1}{(j^m+cn\overline{\delta})^2},$$

where in the second inequality we use that $x \mapsto x/(1+x)$ is increasing. By Lemma 2.43 in the appendix the sums are of the order $(\overline{\delta}cn)^{-2+1/m}$ for $c \in I_n$. The function s_n can be treated analogously.

The derivative of $D^L_{2,n}$ is given by

$$(D^L_{2,n})'(c) = \sum_{j=1}^{n} \left(\frac{\lambda_{j,n}}{1 + c\lambda_{j,n}} - \frac{\lambda_{j,n}}{(1 + c\lambda_{j,n})^2} \right) = \sum_{j=1}^{n} \frac{c\lambda^2_{j,n}}{(1 + c\lambda_{j,n})^2}.$$

The monotonicity of $D^L_{2,n}$ is a consequence of the positivity of this function. The value of $D^L_{2,n}$ at c is the integral of this derivative over the interval $[0, c]$. If (2.12) holds, then

$$\underline{\delta}^2 \int_0^c \sum_{j=1}^{n} \frac{sn^2}{(j^m + \overline{\delta}sn)^2}\, ds \leq D^L_{2,n}(c) \leq \overline{\delta}^2 \int_0^c \sum_{j=1}^{n} \frac{sn^2}{(j^m + \underline{\delta}sn)^2}\, ds.$$

By Lemma 2.43 the integrands are asymptotic to a multiple of $(sn^2)(\delta sn)^{-2+1/m} = n^{1/m}s^{-1+1/m}$ uniformly in $s \in [l_n/n, n^{m-1}]$, for any $l_n \to \infty$ and $\delta = \underline{\delta}$ and $\delta = \overline{\delta}$ respectively. The integral of the latter function over $[0, c]$ is equal to a multiple of $(cn)^{1/m}$, while its integral over $[0, l_n/n]$ is of the order $l_n^{1/m}$. The integral of $(D^L_{2,n})'$ over $[0, l_n/n]$ is bounded above by a multiple of $\int_0^{l_n/n} sn^2 \sum_{j=1}^n j^{-2m}\, ds \asymp l_n^2$. Hence both remainders are of lower order than $(cn)^{1/m}$ for $c \in I_n$ if l_n is chosen equal to, for instance, $\log\log n$.

The proof under (2.20) is the same, except that we use Lemma 2.45 instead of Lemma 2.43. The final assertion also follows along the same lines, but now employing Lemma 2.44. The details are deferred to Section 2.6. □

Coverage of the empirical Bayes credible sets

The function f is contained in the empirical Bayes credible sets (2.7) if $\|\vec{f}_n - \hat{f}_{n,\hat{c}_n}\| \leq Mr_n(\hat{c}_n, \eta)$. In view of (3.3) and (2.1), the square of the left side can be decomposed for any c as

$$\begin{aligned}\|\hat{f}_{n,c} - \vec{f}_n\|^2 &= \vec{f}_n^T \Sigma_{n,c}^{-2} \vec{f}_n - 2\vec{f}_n^T \Sigma_{n,c}^{-1}(I - \Sigma_{n,c}^{-1})\vec{\varepsilon}_n + \vec{\varepsilon}_n^T (I - \Sigma_{n,c}^{-1})^2 \vec{\varepsilon}_n \\ &= D^R_{1,n}(c, f) + D^R_{2,n}(c) + R_{3,n}(c, f) + R_{4,n}(c), \qquad (2.29)\end{aligned}$$

where the first two processes on the right are defined in (2.23) and (2.24) and

$$R_{3,n}(c, f) = -2\vec{f}_n^T \Sigma_{n,c}^{-1}(I - \Sigma_{n,c}^{-1})\vec{\varepsilon}_n = -2\sum_{j=1}^{n} \frac{(c\lambda_{j,n})Z_{j,n}f_{j,n}}{(1 + c\lambda_{j,n})^2}, \qquad (2.30)$$

$$R_{4,n}(c) = \vec{\varepsilon}_n^T (I - \Sigma_{n,c}^{-1})^2 \vec{\varepsilon}_n - \mathrm{tr}\big((I - \Sigma_{n,c}^{-1})^2\big) = \sum_{j=1}^{n} \frac{(c\lambda_{j,n})^2(Z^2_{j,n} - 1)}{(1 + c\lambda_{j,n})^2}.$$

The following proposition shows that the remainder $R_{3,n} + R_{4,n}$ is negligible relative to the deterministic process D_n, for both the likelihood-based and risk-based functions.

Proposition 2.15. *If (2.12), (2.19) or (2.20) holds, then for $R_{3,n}$ and $R_{4,n}$ given in (2.30) and $D_n = D_{1,n} + D_{2,n}$ given in (2.23) or (2.24) we have*

$$\sup_{c \in I_n} \frac{|R_{3,n}(c,f)| + |R_{4,n}(c)|}{D_n(c,f)} \overset{P_f}{\to} 0. \tag{2.31}$$

The proof of the proposition can be found in Section 2.6.

The radius $r_n(c,\eta)$ of the Bayesian credible set is the η-quantile of the posterior distribution of $\|\vec{f_n} - \hat{f}_{n,c}\|$ given c. As the distribution of $\vec{f_n} - \hat{f}_{n,c}$ does not depend on Y, the radius $r_n(c,\eta)$ is deterministic. Since the posterior distribution of $\vec{f_n} - \hat{f}_{n,c}$ is multivariate normal with mean zero and covariance matrix $I - \Sigma_{n,c}^{-1}$ (see (3.3)), the square norm is equal in distribution to the variable

$$N_n(c) = \sum_{j=1}^{n} \frac{c\lambda_{j,n} Z_{j,n}^2}{1 + c\lambda_{j,n}}, \tag{2.32}$$

where the $Z_{j,n}$ are independent standard normal random variables. The mean of this variable is by its definition the posterior variance $s_n^2(c)$, given in (2.26). The following proposition shows that the variables N_n degenerate to their mean as $n \to \infty$.

Proposition 2.16. *If (2.12), (2.19) or (2.20) holds, then*

$$\sup_{c \in I_n} \left| \frac{N_n(c)}{s_n^2(c)} - 1 \right| \overset{P}{\to} 0. \tag{2.33}$$

The proof of the proposition can be found in Section 2.6.

We are ready for our main result on coverage. The result applies to discrete polished tail functions and under every of the three eigenvalues conditions, but we give a more general statement, which takes the output of the preceding propositions as its conditions.

Theorem 2.17 (Coverage). *Suppose the following conditions hold:*

1. *the remainders $R_{1,n}$ and $R_{2,n}$ behave as in (2.25) and $R_{3,n}$ and $R_{4,n}$ behave as in (2.31),*
2. *(2.33) is satisfied,*

3. *$D^R_{2,n}(c) \asymp D^L_{2,n}(c) \asymp s_n^2(c)$ uniformly in $c \in I_n$,*

4. *the function f satisfies the good bias condition and $\sum_{j=1}^n f_{j,n}^2 \leq \sup_{c\in I_n} D_{2,n}(c)$.*

Then $P_f(f \in \hat{C}_{n,\eta,M}) \to 1$, for both the risk-based and likelihood-based credible sets (2.7) and sufficiently large M. In particular, this is true if (2.12), (2.19) or (2.20) and condition 4 above hold.

Proof. Since $N_n(c)/s_n^2(c) \to 1$ in probability uniformly in $c \in I_n$, the quantities $r_n^2(c,\eta)/s_n^2(c)$, which are the η-quantiles of the variables $N_n(c)/s_n^2(c)$, tend to 1 as well, uniformly in c. In order to see this, suppose that $\sup_{c\in I_n} |r_n^2(c,\eta)/s_n^2(c) - 1| \not\to 0$. Then there exist a subsequence $r^2_{n_k}/s^2_{n_k}$ and points $c_k \in I_n$ such that $|r^2_{n_k}(c_k,\eta)/s^2_{n_k}(c_k) - 1| > \epsilon$. We may assume that we either have $r^2_{n_k}(c_k,\eta)/s^2_{n_k}(c_k) > 1+\epsilon$ for all k or $r^2_{n_k}(c_k,\eta)/s^2_{n_k}(c_k) < 1-\epsilon$ for all k. In the latter case, we see that along this subsequence we have

$$P\left(\frac{N_{n_k}(c_k)}{s^2_{n_k}(c_k)} < \frac{r^2_{n_k}(c_k,\eta)}{s^2_{n_k}(c_k)}\right) \leq P\left(\sup_{c\in I_{n_k}} \frac{N_{n_k}(c)}{s^2_{n_k}(c)} < 1-\epsilon\right) \to 0$$

by (2.33). The case that $r_{n_k}(c_k) > 1+\epsilon$ can be treated similarly, where now this probability tends to one. In either case, this contradicts the definition of $r_n(c,\eta)$.

It follows that f is contained in the set $\hat{C}_{n,\eta,M}$ if $\|\hat{f}_{n,\hat{c}_n} - \vec{f}_n\|^2/s_n^2(\hat{c}_n) \leq M^2\big(1+o_P(1)\big)$. By the decomposition (2.29) this is equivalent to

$$\frac{D^R_{1,n}(\hat{c}_n,f) + D^R_{2,n}(\hat{c}_n) + R_{3,n}(\hat{c}_n,f) + R_{4,n}(\hat{c}_n)}{s_n^2(\hat{c}_n)} \leq M^2\big(1+o_P(1)\big).$$

By assumption $s_n^2(\hat{c}_n)$ has the same asymptotic behaviour as both $D^R_{2,n}(\hat{c}_n)$ and $D^L_{2,n}(\hat{c}_n)$, up to a multiplicative constant. If f satisfies the good bias condition for the risk-based procedure, then $D^R_{2,n}(\hat{c}_n) \gtrsim D^R_{1,n}(\hat{c}_n,f)$ with probability tending to one by Theorem 2.12, whence $D^R_n(\hat{c}_n,f) \asymp D^R_{2,n}(\hat{c}_n) \asymp s_n^2(\hat{c}_n)$. It then follows that the first two terms in the display are bounded above, while the remainder terms tend to zero by (2.31).

By definition we always have $D^R_{1,n}(c,f) \leq D^L_{1,n}(c,f)$. If f satisfies the good bias condition for the likelihood-based procedure, then $D^L_{1,n}(\hat{c}_n,f) \lesssim D^L_{2,n}(\hat{c}_n)$ with probability tending to one by Theorem 2.12, while $D^L_{2,n}(\hat{c}_n) \asymp D^R_{2,n}(\hat{c}_n)$ by assumption. It follows that again $D^R_{1,n}(\hat{c}_n,f) \lesssim D^R_{2,n}(\hat{c}_n)$, and the proof is analogous to the risk-based case, where for the last two terms we use the fact that $D^L_n(\hat{c}_n,f) \asymp D^L_{2,n}(\hat{c}_n) \asymp s_n^2(\hat{c}_n)$.

The final assertion of the theorem follows by Propositions 2.10, 2.15 and 2.16 and Lemma 2.14, which show that all assumptions hold under (2.12), (2.19) or (2.20) and the conditions on f. □

Contraction rates of the empirical Bayes posteriors

We first consider the risk-based setting. If the remainder processes in (2.29) are negligible relative to $D_n^R = D_{1,n}^R + D_{2,n}^R$ uniformly in $c \in I_n$, which is true under our three eigenvalue conditions by Proposition 2.15, then

$$\|\hat{f}_{n,\hat{c}_n} - \vec{f}_n\|^2 = O_P\big(D_n^R(\hat{c}_n, f)\big). \tag{2.34}$$

For the estimator $\hat{c}_n$ the right side is by the first assertion of Theorem 2.12 of the order (in probability)

$$\inf_{c\in I_n} D_n^R(c, f)$$

with probability tending to one. Since $D_n^R(c, f)$ is exactly the risk of the estimator $\hat{f}_{n,c}$ for a given c, these two assertions combined can be viewed as an *oracle type* inequality for the risk-based empirical Bayes plug-in posterior mean $\hat{f}_{n,\hat{c}_n}$: the empirical Bayes estimator manages to choose the best value of c for each possible f. The family of estimators $\hat{f}_{n,c}$, where $c \in I_n$, turns out be rich enough to give an optimal estimation rate for the usual regularity classes. Thus the estimator $\hat{f}_{n,\hat{c}_n}$ adapts to unknown regularity in the usual sense. We formalize this in the next theorem, together with the observation that the posterior variance also adapts correctly. From this we deduce that the full posterior distribution contracts adaptively.

Write $\Pi_c\big(\cdot \,|\, \vec{Y}_n\big)$ for the posterior distribution of $\vec{f}_n$ given c and let $\Pi_{\hat{c}_n}\big(\cdot \,|\, \vec{Y}_n\big)$ be the same object, but with c replaced by $\hat{c}_n$.

Theorem 2.18 (Contraction, risk-based EB)**.** *Suppose the following conditions hold:*

1. *the remainders $R_{1,n}$ and $R_{2,n}$ behave as in (2.25) and $R_{3,n}$ and $R_{4,n}$ behave as in (2.31),*
2. *$D_{2,n}^R(c) \asymp s_n^2(c)$ uniformly in $c \in I_n$.*

Then for $\hat{c}_n$ given by (2.6) and any sequence $M_n \to \infty$,

$$\Pi_{\hat{c}_n}\Big(w : \|\vec{w}_n - \vec{f}_n\|^2 \geq M_n \inf_{c\in I_n} E_f\|\hat{f}_{n,c} - \vec{f}_n\|^2 \,|\, \vec{Y}_n\Big) \overset{P_f}{\to} 0.$$

In particular, this is true if (2.12), (2.19) or (2.20) holds.

Proof. Let W denote a variable that given $\vec{Y}_n$ and c is distributed according to the posterior distribution of f. Then we have by Markov's inequality

$$\begin{aligned}M^2\,\Pi_c\bigl(w:\|\vec{w}_n-\vec{f}_n\|^2\ge M^2\,|\,\vec{Y}_n\bigr)&\le \mathrm{E}\bigl[\|\vec{W}_n-\vec{f}_n\|^2\,|\,\vec{Y}_n,c\bigr]\\&\le\|\hat{f}_{n,c}-\vec{f}_n\|^2+\mathrm{E}\bigl[\|\vec{W}_n-\hat{f}_{n,c}\|^2\,|\,\vec{Y}_n,c\bigr]\end{aligned}$$

for any M and c. The second term on the far right is the posterior variance $s_n^2(c)$, which by assumption is bounded by a multiple of $D_{2,n}^R(c)\le D_n^R(c,f)$ uniformly in $c\in I_n$. The first term on the far right evaluated at $c=\hat{c}_n$ is bounded above by $D_n^R(\hat{c}_n,f)$ with probability tending to one, in view of (2.29) and (2.25) and (2.31). It follows that with probability tending to one

$$\Pi_{\hat{c}_n}\bigl(w:\|\vec{w}_n-\vec{f}_n\|^2\ge M^2\,|\,\vec{Y}_n\bigr)\lesssim\frac{1}{M^2}D_n^R(\hat{c}_n,f)\lesssim\frac{1}{M^2}\inf_{c\in I_n}D_n^R(c,f)$$

by the first assertion of Theorem 2.12. Since $D_n^R(c,f)=\mathrm{E}_f\|\hat{f}_{n,c}-\vec{f}_n\|^2$, the proof is complete. □

Thus the risk-based empirical Bayes method attains a rate of contraction equal to the best estimator in the class of estimators $\hat{f}_{n,c}$, for $c\in I_n$. In standard models this class contains a rate-minimax estimator.

Example 2.19 (Sobolev norm)**.** Denote by S_n^α the set all functions f for which the discrete Sobolev norm $\|f\|_{n,\alpha}$, defined in (2.11), is bounded by 1. For eigenvalues satisfying (2.12) and $f\in S_n^\alpha$ for $\alpha\le m$ we have

$$\begin{aligned}D_{1,n}^R(c,f)&\lesssim\sum_{j=1}^n\frac{j^{2m}f_{j,n}^2}{(j^m+cn)^2}\lesssim\frac{1}{(cn)^2}\sum_{j=1}^{(cn)^{1/m}}j^{2m}f_{j,n}^2+\sum_{j=(cn)^{1/m}+1}^n f_{j,n}^2\\&\lesssim\frac{(cn)^{(2m-2\alpha)/m}}{(cn)^2}\sum_{j=1}^{(cn)^{1/m}}j^{2\alpha}f_{j,n}^2+\frac{1}{(cn)^{2\alpha/m}}\sum_{j=(cn)^{1/m}+1}^n j^{2\alpha}f_{j,n}^2\\&\le n(cn)^{-2\alpha/m}.\end{aligned}$$

In combination with Lemma 2.14 we find that

$$\frac{1}{n}D_n^R(c,f)\lesssim(cn)^{-2\alpha/m}+n^{-1}(cn)^{1/m}.$$

The argument $c=n^{m/(1+2\alpha)-1}$ equates the two terms and gives a value of the order $n^{-2\alpha/(1+2\alpha)}$. By Theorem 2.18 this is the square contraction rate of the plug-in posterior distribution with the risk-based empirical Bayes estimator (2.6) relative to the scaled Euclidean norm $\|\cdot\|_{n,0}$.

For $\alpha>m$ the order of the square bias $D_{1,n}^R(c,f)$ does not improve beyond the rate $n(cn)^{-2}$ found for $\alpha=m$ and hence nor does the contraction rate.

Example 2.20 (Hyperrectangles)**.** Denote by Θ_n^α the set all functions f for which the discrete Sobolev norm $\|f\|_{n,\alpha,\infty}$, defined in (2.11), is bounded by 1. For eigenvalues satisfying (2.12) and $f \in \Theta_n^\alpha$ we have

$$D_{1,n}^R(c,f) \leq \sum_{j=1}^n \frac{nj^{-2\alpha-1}}{(1+c\lambda_{j,n})^2} \lesssim n \sum_{j=1}^n \frac{j^{2m-2\alpha-1}}{(j^m+cn)^2} \lesssim \begin{cases} \frac{n}{(cn)^{2\alpha/m}} & \text{if } \alpha < m \\ \frac{n\log(cn)}{(cn)^2} & \text{if } \alpha = m \\ \frac{n}{(cn)^2} & \text{if } \alpha > m. \end{cases}$$

The first case follows directly by Lemma 2.43, the second by writing

$$n \sum_{j=1}^n \frac{j^{2m-2\alpha-1}}{(j^m+cn)^2} = n \sum_{j=1}^{(cn)^{1/m}} \frac{j^{2m-2\alpha-1}}{(j^m+cn)^2} + n \sum_{j=(cn)^{1/m}+1}^{n} \frac{j^{2m-2\alpha-1}}{(j^m+cn)^2}$$

and applying a variant of the lemma to the second sum. The third case follows immediately by using $j^m + cn > cn$. For $\alpha < m$ and $\alpha > m$ this is the same result as in Example 2.19, leading to the same conclusions on the contraction rate. For $\alpha = m$ the additional logarithmic factor leads to the square contraction rate $n^{-2\alpha/(2\alpha+1)}(\log n)^{1/(2\alpha+1)}$.

The likelihood-based empirical Bayes method also satisfies an oracle type inequality, but relative to a loss function that is not as closely linked to the L_2-risk of the posterior mean. Because its "bias term" $D_{1,n}^L$ is bigger (the inequality $D_{1,n}^L \geq D_{1,n}^R$ is immediate from definitions (2.23) and (2.24)), while its "variance term" $D_{2,n}^L$ has the same order of magnitude, in its attempt to balance bias and variance the likelihood-based empirical Bayes method may choose a bigger estimator $\hat{c}_n$ than the risk-based method. This may have an adverse effect on the contraction rate of the plug-in posterior distribution.

Theorem 2.21 (Contraction, likelihood-based EB)**.** *Suppose the following conditions hold:*

1. *the remainders $R_{1,n}$ and $R_{2,n}$ behave as in (2.25) and $R_{3,n}$ and $R_{4,n}$ behave as in (2.31),*
2. *$D_{2,n}^L(c) \asymp s_n^2(c)$ uniformly in $c \in I_n$.*

Then for $\hat{c}_n$ given by (2.4) and any sequence $M_n \to \infty$ we have

$$\Pi_{\hat{c_n}}\Big(w : \|\vec{w}_n - \vec{f}_n\|^2 \geq M_n \inf_{c\in I_n} D_n^L(c,f) \,|\, \vec{Y}_n\Big) \overset{P_f}{\to} 0.$$

In particular, this is true if (2.12), (2.19) or (2.20) holds.

Proof. Since $D_n^L \gtrsim D_n^R$ we obtain as in the proof of Theorem 2.18 that

$$\|\hat{f}_{n,\hat{c}_n} - \vec{f}_n\|^2 = O_P\big(D_n^L(\hat{c}_n, f)\big).$$

Next we can use the first assertion of Theorem 2.12 to replace the right hand side by the infimum of $D_n^L(c,f)$ over c. The posterior variance is of the same order as $D_{2,n}^L$ and hence the proof can be concluded as the proof of Theorem 2.18. □

Even though the loss function of the likelihood-based empirical Bayes estimator does not relate correctly to the risk in general, the method does give optimal contraction rates on the models in the preceding examples, albeit for a smaller range of regularity levels.

Example 2.22 (Sobolev norm). For $f \in S_n^\alpha$ for $\alpha \le m/2$ and eigenvalues satisfying (2.12) we have

$$\begin{aligned} D_{1,n}^L(c,f) &\lesssim \sum_{j=1}^n \frac{j^m f_{j,n}^2}{j^m + cn} \lesssim \frac{1}{cn} \sum_{j=1}^{(cn)^{1/m}} j^m f_{j,n}^2 + \sum_{j=(cn)^{1/m}+1}^{n} f_{j,n}^2 \\ &\lesssim \frac{(cn)^{(m-2\alpha)/m}}{cn} \sum_{j=1}^{(cn)^{1/m}} j^{2\alpha} f_{j,n}^2 + \frac{1}{(cn)^{2\alpha/m}} \sum_{j=(cn)^{1/m}+1}^{n} j^{2\alpha} f_{j,n}^2 \\ &\le n(cn)^{-2\alpha/m}. \end{aligned}$$

The upper bound has the same form as for the risk-based empirical Bayes method. Since $D_{2,n}^L \asymp D_{2,n}^R$, we obtain the same contraction rate results. The difference is that the rate does not improve for $\alpha \ge m/2$.

Example 2.23 (Hyperrectangles). For eigenvalues satisfying (2.12) and $f \in \Theta_n^\alpha$ we have

$$D_{1,n}^L(c,f) \le \sum_{j=1}^n \frac{nj^{-2\alpha-1}}{1+c\lambda_{j,n}} \lesssim n \sum_{j=1}^n \frac{j^{m-2\alpha-1}}{j^m + cn} \lesssim \begin{cases} \frac{n}{(cn)^{2\alpha/m}} & \text{if } \alpha < m/2 \\ c^{-1}\log(cn) & \text{if } \alpha = m/2 \\ c^{-1} & \text{if } \alpha > m/2. \end{cases}$$

This leads to the contraction rate $n^{-\alpha/(2\alpha+1)}$ relative to the scaled Euclidean norm $\|\cdot\|_{n,0}$ if $\alpha < m/2$ and the square contraction rate $n^{-2\alpha/(2\alpha+1)}(\log n)^{1/(2\alpha+1)}$ if $\alpha = m/2$.

Diameter of the empirical Bayes credible sets

The empirical Bayes credible sets inherit their diameter from the contraction rate.

Corollary 2.24. *Under the conditions of Theorems 2.18 and 2.21 the square of the diameter $Mr_n(\hat{c}_n, \eta)$ of the credible sets (2.7) is of the order $\inf_{c\in I_n} D_n^R(c,f)$ and $\inf_{c\in I_n} D_n^L(c,f)$ for the risk-based and likelihood-based empirical Bayes procedures respectively with probability tending to one.*

Proof. By Theorems 2.18 and 2.21 the empirical Bayes posterior distributions concentrate all their mass on a ball of radius of the same order as the given rate. Since the posterior distribution is Gaussian, the balls B_n of the same radius centred at the posterior mean must also have mass tending to one. By definition the credible sets are balls of posterior mass $\eta \in (0,1)$ around the posterior mean, and hence are contained in the B_n.

Alternatively, the square radius $r_n^2(\hat{c}_n, \eta)$ was seen to be of the same order as the posterior variance $s_n^2(\hat{c}_n)$, which was in turn seen to have the given order. □

2.3 Hierarchical Bayes

The hierarchical Bayes method is closely related to the likelihood-based empirical Bayes method, since the posterior density of c is proportional to the product of the the prior density π for c and the marginal likelihood that defines the latter method. More precisely, in the model (2.2) augmented with $c \sim \pi$ it holds that

$$\pi_n(c \,|\, \vec{Y}_n) \propto p(\vec{Y}_n \,|\, c)\, \pi(c) \propto \det \Sigma_{n,c}^{-1/2}\, e^{-\frac{1}{2}\vec{Y}_n^T \Sigma_{n,c}^{-1} \vec{Y}_n}\, \pi(c).$$

The likelihood-based empirical Bayes estimator (2.4) would be the posterior mode if the prior density were improper. We shall analyse the hierarchical Bayes method by exploiting this link.

We start with showing that the posterior distribution of c concentrates on the interval where the deterministic part of the likelihood-based criterion $D_n^L = D_{1,n}^L + D_{2,n}^L$ is small. This criterion is derived from minus the log marginal likelihood. On closer inspection it becomes evident that the prior density π, which we will choose inverse gamma, also plays a role and adds a term $1/c$ to this criterion. We truncate the inverse gamma prior to the interval I_n, so that c has a prior density so that, for some fixed $\kappa, \lambda > 0$,

$$\pi(c) \propto c^{-1-\kappa}\, e^{-\lambda/c}, \qquad c \in I_n.$$

Theorem 2.25. *Suppose the following conditions hold:*

1. *the remainders $R^L_{1,n}$ and $R^L_{2,n}$ satisfy (2.25),*
2. *the function $D^L_{2,n}$ is a good variance function with $D^L_{2,n}(c) \geq \log(nc)$,*
3. *there is a minimizer $c_n(f)$ of $c \mapsto D^L_n(c,f) + 2\lambda/c$ over $c \in (0,\infty)$ that satisfies $c_n(f) \in I_n$ and $2c_n(f) \in I_n$.*

Then for sufficiently large M

$$\Pi_n\Big(c : D^L_n(c,f) + \frac{1}{c} \leq M \inf_{c>0}\Big[D^L_n(c,f) + \frac{1}{c}\Big] \,|\, \vec{Y}_n\Big) \overset{P_f}{\to} 1.$$

Furthermore, if f satisfies the good bias condition relative to $D^L_{1,n}$, then

$$\Pi_n\Big(c : D^L_{1,n}(c,f) + \frac{1}{c} \lesssim D^L_{2,n}(c) \,|\, \vec{Y}_n\Big) \overset{P_f}{\to} 1.$$

Moreover, there exist constants $0 < k < K < \infty$ such that

$$\Pi_n\Big(c : c \in \big[kc_n(f), Kc_n(f)\big] \,|\, \vec{Y}_n\Big) \overset{P_f}{\to} 1.$$

In particular, these assertions are true if (2.12), (2.19) or (2.20) holds, for every f satisfying condition 3.

Proof. For every measurable set $J \subseteq I_n$ we have

$$\begin{aligned}\Pi_n\Big(c : c \in J \,|\, \vec{Y}_n\Big) &= \frac{\int_J e^{-\frac{1}{2}L^L_n(c,f)}\,\pi(c)\,dc}{\int_{I_n} e^{-\frac{1}{2}L^L_n(c,f)}\,\pi(c)\,dc}\\ &= \frac{\int_J e^{-\frac{1}{2}[D^L_n(c,f)+R^L_n(c,f)]}\,\pi(c)\,dc}{\int_{I_n} e^{-\frac{1}{2}[D^L_n(c,f)+R^L_n(c,f)]}\,\pi(c)\,dc}\end{aligned}$$

by the decomposition (2.22). Define $\ell_n(c,f) = D^L_n(c,f) + 2\lambda/c$, so that $c_n := c_n(f)$ is a minimizer of ℓ_n. In view of (2.25) we have for any $\delta > 0$

$$\ell_n(c,f)(1-\delta) \leq D^L_n(c,f) + R^L_n(c,f) + \frac{2\lambda}{c} \leq \ell_n(c,f)(1+\delta),$$

with probability tending to one. Consequently, we see that

$$\Pi_n\Big(c : c \in J \,|\, \vec{Y}_n\Big) \leq \frac{\int_J e^{-\frac{1}{2}\ell_n(c,f)(1-\delta)}\,c^{-\kappa-1}\,dc}{\int_{I_n} e^{-\frac{1}{2}\ell_n(c,f)(1+\delta)}\,c^{-\kappa-1}\,dc}.$$

with probability tending to one. Since $D^L_{2,n}$ is a good variance function, we have that $D^L_{2,n}(2c_n) \leq (B')^{-1}2^b D_{2,n}(c_n)$. Because $D^L_{1,n}$ is decreasing and $D^L_{2,n}$

is increasing, we then also have that $\ell_n(c,f) \leq (B')^{-1}2^b\ell_n(c_n,f)$ for every $c \in [c_n, 2c_n]$. Combining this with the fact that $\ell_n(c,f) \geq 2\lambda/c$, it follows that

$$\begin{aligned}\Pi_n\Big(c : \ell_n(c,f) \geq M\ell_n(c_n,f) \,|\, \vec{Y}_n\Big)&\\ &\leq \frac{\int e^{-\frac{1}{4}\ell_n(c,f)(1-\delta)}\, c^{-\kappa-1}\, dc\, e^{-\frac{1}{4}(1-\delta)M\ell_n(c_n,f)}}{e^{-\frac{1}{2B'}2^b(1+\delta)\ell_n(c_n,f)}\int_{c_n}^{2c_n} c^{-\kappa-1}\, dc}\\ &\lesssim c_n^{\kappa}e^{-\kappa\ell_n(c_n,f)}\int_0^\infty e^{-\frac{1}{2}(1-\delta)\lambda/c}\, c^{-\kappa-1}\, dc\end{aligned}$$

for $M(1-\delta) \geq (4\kappa + 2(B')^{-1}2^b)(1+\delta)$. If $c_n \to 0$, then this clearly tends to zero. If c_n is bounded away from zero, the above also tends to zero, by the assumption that $\ell_n(c,f) \geq \log(cn)$. This concludes the proof of the first assertion of the theorem.

If f satisfies the good bias condition, then, for $K > 1$,

$$D^L_{1,n}(Kc,f) + \frac{2\lambda}{Kc} \leq K^{-a}D^L_{1,n}(c,f) + \frac{2\lambda}{Kc} \leq K^{-(a\wedge 1)}\Big[D^L_{1,n}(c,f) + \frac{2\lambda}{c}\Big].$$

In other words, the function $c \mapsto D^L_{1,n}(c,f) + 2\lambda/c$ also satisfies a good bias condition.

Let $\Lambda_n = \big\{c : \ell_n(c,f) \leq M\ell_n(\tilde{c}_n,f)\big\}$, for $\tilde{c}_n$ the solution to the equation $D^L_{1,n}(c,f) + 2\lambda/c = D^L_{2,n}(c)$. Since $\ell_n(c_n,f) \leq \ell_n(\tilde{c}_n,f)$, we have that $\Pi_n(c : c \in \Lambda_n \,|\, \vec{Y}_n\big) \to 1$ by the first part of the proof. Since ℓ_n is the sum of the decreasing function $D^L_{1,n}(c,f) + 2\lambda/c$ and the increasing function $D^L_{2,n}$, which are both "good" functions, it follows that $D^L_{1,n}(c,f) + 2\lambda/c \lesssim D^L_{2,n}(c)$ for every $c \in \Lambda_n$ by Lemma 2.42(i). Furthermore, Lemma 2.42(ii) gives the existence of constants $0 < k_1 < K_1 < \infty$ with $\Lambda_n \subset [k_1\tilde{c}_n, K_1\tilde{c}_n]$. Since $c_n \in \Lambda_n$, it follows that also $\Lambda_n \subset [k_1/K_1c_n, K_1/k_1c_n]$. This proves the second and third assertions of the theorem. □

The theorem shows that under the posterior distribution the scaling c will concentrate on the set of small values of the criterion $c \mapsto D^L_n(c,f) + 1/c$. This differs by the term $1/c$ from the criterion minimized by likelihood-based empirical Bayes estimator $\hat{c}_n$ defined by (2.4), whose behaviour is given in Theorem 2.12. The additional term is due to the prior distribution. The usual prior distribution, which we consider here, has very thin tails near 0, and the extra term $1/c$ essentially prevents the posterior distribution to concentrate very close to zero.

Very small values of the scaling parameter c are advantageous for very smooth functions f. For such functions the bias term $D^L_{1,n}(c,f)$ will be very small and the balance between square bias $D^L_{1,n}(c, f)$ and variance $D^L_{2,n}(c)$ will be assumed for small c. The additional term can be viewed as adding an artificial bias term of the order $1/c$, thus shifting the bias-variance trade-off to bigger values of c.

In most cases this is not harmful. In particular, the shift will not be apparent in contraction rates over the usual smoothness models (see Example 2.29). The following example shows that this is different for very smooth f.

Example 2.26. The smoothest imaginable function f is the zero function. For $f = 0$, the bias function $D^L_{1,n}(c, f)$ in (2.24) vanishes. If the eigenvalues satisfy (2.12), then the variance $D^L_{2,n}(c)$ is of the order $(cn)^{1/m}$ by Lemma 2.14 and the criterion becomes

$$c \mapsto D^L_n(c, f) + \frac{1}{c} \asymp (cn)^{1/m} + \frac{1}{c}.$$

The right side is minimized by $c_n \asymp (1/n)^{1/(m+1)}$. Theorem 2.25 shows that the posterior distribution for the scale parameter c will concentrate on the set of c that minimize the criterion up to a multiplicative factor. This set is contained in an interval with boundaries of the order $(1/n)^{1/(m+1)}$.

The fact that this interval shrinks to zero is good, as the variance is smaller for smaller c, while the bias is negligible. However, it is a bit disappointing that the shrinkage is not faster than of order $(1/n)^{1/(m+1)}$. In comparison, the empirical Bayes estimator $\hat{c}_n$ will shrink at the order $\log n/n$, the minimal possible value permitted in our minimization scheme by Theorem 2.12.

Coverage of the hierarchical Bayes credible set

The hierarchical Bayesian credible sets cover true parameters under the same conditions as the empirical Bayes sets.

Theorem 2.27 (Coverage, HB)**.** *Suppose the following conditions hold:*

1. *the remainders $R^L_{1,n}$ and $R^L_{2,n}$ behave as in (2.25) and $R_{3,n}$ and $R_{4,n}$ behave as in (2.31),*
2. *(2.33) is satisfied,*
3. *$D^L_{2,n}$ is a good variance function with $D^L_{2,n}(c) \geq \log(nc)$,*
4. *there is a minimizer $c_n(f)$ of $c \mapsto D^L_n(c, f) + 2\lambda/c$ over $c \in (0, \infty)$ that satisfies $c_n(f) \in I_n$ and $2c_n(f) \in I_n$,*
5. *$D^R_{2,n}(c) \asymp D^L_{2,n}(c) \asymp s^2_n(c)$ uniformly in $c \in I_n$,*

6. *the function f satisfies the good bias condition.*

Then the hierarchical Bayes credible sets (2.8) satisfy $P_f(f \in \hat{C}_{n,\eta,M}) \to 1$ for sufficiently large M. In particular, this is true if (2.12), (2.19) or (2.20) holds and conditions 4 and 6 hold.

Proof. The function f is contained in $\hat{C}_{n,\eta,M}$ as soon as there exists some $c \in [\hat{c}_{1,n}(\eta_1), \hat{c}_{2,n}(\eta_1)]$ for which it holds that $\|\vec{f}_n - \hat{f}_{n,c}\| \leq M r_n(c, \eta_2)$. Since $N_n(c)/s_n^2(c) \to 1$ in probability uniformly in $c \in I_n$ by (2.33), the quantities $r_n^2(c,\eta_2)/s_n^2(c)$, which are the η_2-quantiles of the variables $N_n(c)/s_n^2(c)$, tend to 1 as well uniformly in c. In view of the decomposition (2.29) it follows that the function f is contained in $\hat{C}_{n,\eta,M}$ as soon as there exists some $c \in [\hat{c}_{1,n}(\eta_1), \hat{c}_{2,n}(\eta_1)]$ with

$$\frac{D_{1,n}^R(c,f) + D_{2,n}^R(c) + R_{3,n}(c,f) + R_{4,n}(c)}{s_n^2(c)} \leq M^2\big(1 + o_P(1)\big).$$

By assumption $s_n^2(c)$ is equivalent to both $D_{2,n}^R(c)$ and $D_{2,n}^L(c)$, up to a multiplicative constant. In particular, the second term on the left is bounded above.

By the second assertion of Theorem 2.25 the posterior probability of the set $\Lambda_n := \{c : D_{1,n}^L(c,f) \lesssim D_{2,n}^L(c)\}$ tends to one in probability. Since $\hat{c}_{1,n}(\eta_1)$ and $\hat{c}_{2,n}(\eta_1)$ are nontrivial quantiles of the posterior distribution of c, the interval $[\hat{c}_{1,n}(\eta_1), \hat{c}_{2,n}(\eta_1)]$ must intersect Λ_n with probability tending to 1. For $c = \bar{c}_n$ in this intersection it holds that $D_n^L(c,f) \asymp D_{2,n}^L(c)$ and hence $s_n^2(c)$ in the preceding display can be replaced by $D_n^L(c,f)$, up to a multiplicative constant. This shows that the remainder terms tend to zero, in view of (2.31). The first term $D_{1,n}^R(c,f)/s_n^2(c)$ is bounded by a multiple of $D_{1,n}^R(c,f)/D_n^L(c,f) \leq D_{1,n}^R(c,f)/D_{1,n}^L(c,f) \leq 1$, by definitions (2.23) and (2.24). This proves the first assertion of the theorem.

The final assertion of the theorem follows by Propositions 2.10, 2.15 and 2.16 and Lemma 2.14, which show that all remaining assumptions hold under (2.12), (2.19) or (2.20). □

Contraction rate of the hierarchical Bayes posterior

As in Section 2.2 write $\Pi_c\big(\cdot \,|\, \vec{Y}_n\big)$ for the posterior distribution of $\vec{f}_n$ given c. Then the hierarchical posterior distribution can be decomposed as

$$\Pi_n\big(w : \vec{w}_n \in B \,|\, \vec{Y}_n\big) = \int \Pi_c\big(w : \vec{w}_n \in B \,|\, \vec{Y}_n\big)\, \pi_n(c \,|\, \vec{Y}_n)\, dc$$

for $B \subseteq \mathbb{R}^n$ measurable. Here $\pi_n(c \,|\, \vec{Y}_n)$ is the posterior density of c, analysed in Theorem 2.25.

This hierarchical posterior distribution contracts to the true parameter according to an oracle inequality, with the likelihood-based criterion augmented by the extra term $1/c$.

Theorem 2.28 (Contraction rate, HB)**.** *If conditions 1, 3, 4, and 5 of Theorem 2.27 hold, then, for any sequence $M_n \to \infty$,*

$$\Pi_n\Big(w : \|\vec{w}_n - \vec{f}_n\|^2 \geq M_n \inf_{c \in I_n} \Big[D_n^L(c, f) + \frac{1}{c}\Big] \,|\, \vec{Y}_n\Big) \overset{P_f}{\to} 0.$$

Proof. Let $c_n \in I_n$ be a minimizer of $c \mapsto D_n^L(c, f) + 1/c$ and for given M_1 define a set

$$C_n = \Big\{c \in I_n : D_n^L(c, f) + 1/c \leq M_1 \big[D_n^L(c_n, f) + 1/c_n\big]\Big\}. \tag{2.35}$$

By Theorem 2.25 the posterior probability that $c \in C_n$ tends to 1 in probability, for sufficiently large M_1. Therefore, for any $M > 0$ we apply the above decomposition of the posterior to find

$$\begin{aligned}
&\Pi_n\big(w : \|\vec{w}_n - \vec{f}_n\| \geq M \,|\, \vec{Y}_n\big) \\
&\qquad \leq \sup_{c \in C_n} \Pi_c\big(w : \|\vec{w}_n - \vec{f}_n\| \geq M \,|\, \vec{Y}_n\big) + \Pi_n(c : c \notin C_n \,|\, \vec{Y}_n) \\
&\qquad \leq \frac{1}{M^2} \sup_{c \in C_n} \big[\|\hat{f}_{n,c} - \vec{f}_n\|^2 + s_n^2(c)\big] + o_P(1)
\end{aligned}$$

by Markov's inequality. In view of (2.29), this is further bounded above by

$$\frac{1}{M^2} \sup_{c \in C_n} \Big[D_{1,n}^R(c, f) + D_{2,n}^R(c) + R_{3,n}(c, f) + R_{4,n}(c) + s_n^2(c)\Big] + o_P(1).$$

Here $D_{1,n}^R \leq D_{1,n}^L$, and $D_{2,n}^R$ is of the same order as $D_{2,n}^L$ and s_n^2. It follows that the first two terms are bounded by a multiple of $\sup_{c \in C_n} D_n^L(c, f) \leq M_1\big[D_n^L(c_n) + 1/c_n\big]$. The remainder terms are of the order $D_n^L(c, f)$ uniformly in $c \in I_n$ with probability tending to one by (2.31) and hence are similarly bounded. □

Example 2.29 (Sobolev)**.** It was seen in Example 2.22 that for eigenvalues satisfying (2.12) and $f \in S_n^\alpha$ for $\alpha \leq m/2$ we have

$$D_{1,n}^L(c, f) + D_{2,n}^L(c) \lesssim n(cn)^{-2\alpha/m} + (cn)^{1/m}.$$

The upper bound on the right side has minimum value $n^{1/(2\alpha+1)}$ at $c_n \asymp n^{m/(1+2\alpha)-1}$. In this point the term $1/c_n$ is smaller than $n^{1/(2\alpha+1)}$ (for $\alpha \leq m/2$). It follows from Theorem 2.28 that on the model S_n^α the hierarchical Bayes posterior distribution contracts at the same rate as the likelihood-based empirical Bayes method.

Example 2.30 (Hyperrectangle)**.** It was seen in Example 2.23 that, for eigenvalues satisfying (2.12) and $f \in \Theta_n^\alpha$,

$$D_{1,n}^L(c,f) + D_{2,n}^L(c) \lesssim \begin{cases} n(cn)^{-2\alpha/m} + (cn)^{1/m} & \text{if } \alpha < m/2, \\ c^{-1}\log(cn) + (cn)^{1/m} & \text{if } \alpha = m/2, \\ c^{-1} + (cn)^{1/m} & \text{if } \alpha > m/2, \end{cases}$$

It follows again that the hierarchical Bayes posterior distribution contracts at the same rate as the likelihood-based empirical Bayes method.

Example 2.31 (Zero function)**.** The square bias $D_{1,n}^L$ of the function $f = 0$ is equal to zero. For eigenvalues satisfying (2.12) the minimum of $c \mapsto D_n^L(c,f) + 1/c$ is assumed at $c_n \asymp (1/n)^{1/(m+1)}$, resulting in a rate of contraction for the scaled Euclidean norm $\|\cdot\|_{n,0}$ of the order $n^{-(m/2)/(m+1)}$.

In contrast the empirical Bayes estimators attain a rate of contraction of the order $n^{-1/2}$ up to a logarithmic factor.

The same difference between the hierarchical and empirical Bayes methods exists for (sequences of) functions f with a square bias $D_{1,n}^R(c,f)$ that tends to zero at an exponential rate.

Diameter of the hierarchical Bayes credible set

The diameter of the credible sets is again of the same order as the contraction rate.

Theorem 2.32. *Under the conditions of Theorem 2.28 the diameter of the credible sets (2.8) is of the order* $\inf_{c\in I_n}\big[D_n^L(c,f)+1/c\big]$ *with probability tending to one.*

Proof. In view of Proposition 2.16, for fixed c the radius of the credible set $\{w : \|\vec{w}_n - \hat{f}_{n,c}\| < Mr_n(c,\eta_2)\}$ is of the order the posterior standard deviation $s_n(c)$ given by (2.26). Thus the triangle inequality gives that the diameter of $\hat{C}_{n,\eta,M}$ is bounded above by a multiple of

$$\sup_{\hat{c}_{1,n}(\eta_1)<c<\hat{c}_{2,n}(\eta_1)} \big[s_n(c) + \|\vec{f}_n - \hat{f}_{n,c}\|\big].$$

The supremum of the function in this display over the set C_n defined in (2.35) is shown to be of the desired order in the proof of Theorem 2.28. The theorem would follow if the interval $[\hat{c}_{1,n}(\eta_1), \hat{c}_{2,n}(\eta_1)]$ belongs to C_n with probability tending to one.

By Theorem 2.25 the posterior distribution of c concentrates all its mass on the sets C_n. Since $\hat{c}_{1,n}(\eta_1)$ and $\hat{c}_{2,n}(\eta_1)$ are nontrivial quantiles of this distribution, we can conclude that they must belong to the convex hull of C_n with probability tending to one. If this convex hull is $[c_m, c_M]$, then for any c in this convex hull

$$\begin{aligned} D_n^L(c,f) + \frac{1}{c} &= D_{1,n}^L(c,f) + \frac{1}{c} + D_{2,n}^L(c) \leq D_{1,n}^L(c_m,f) + \frac{1}{c_m} + D_{2,n}^L(c_M) \\ &\leq 2M_1\Big[D_n^L(c_n,f) + \frac{1}{c_n}\Big]. \end{aligned}$$

Thus the convex hull of C_n is contained in a set of the same form as C_n, but with the constant M_1 replaced by $2M_1$. The proof of Theorem 2.28 still shows that the supremum over this bigger set is of the desired order. □

2.4 On the polished tail condition

The parameter in the regression model (2.1) is a fixed function f, but most of the results of this chapter are driven by the representation of the restriction $\vec{f_n}$ of f to the design points in terms of the eigenvectors $e_{j,n}$ of the covariance matrix U_n of the (unscaled) prior restricted to the design points. It is clearly of interest to relate the "continuous" object f to its discrete counterparts, but this is more involved than it may seem.

In this section we investigate the relationship between the continuous and discrete setups for the special case of the Brownian motion prior.

Aliasing

For the design points $x_{i,n} = i/n_+$, where $n_+ = n + 1/2$, the eigenvectors of the covariance matrix U_n of discretized Brownian motion are given in (2.16) for $j \in \{1, \dots, n\}$. The formula shows that they are $1/\sqrt{n+}$ times the restrictions of the eigenfunctions e_j to the design points. Using this correspondence we may also define vectors $e_{j,n} \in \mathbb{R}^n$ for $j > n$, again by (2.16), by discretizing the higher frequency eigenfunctions of Brownian motion. Since the vectors $e_{1,n}, \dots, e_{n,n}$ are an orthonormal basis of $\mathbb{R}^n$, these further vectors are redundant. It turns out that their linear dependency on the vectors $e_{i,n}$ for $i \leq n$ takes a very special form:

(i) The vectors $e_{i,n}$ are $(2n+1)$-periodic in i: $e_{i+2n+1,n} = e_{i,n}$ for all i.

(ii) The vectors in the middle of a $(2n+1)$ period vanish: $e_{n+1,n} = 0$.

(iii) The vectors within a $(2n+1)$ period are anti-symmetric about the midpoint: $e_{2n+2-i,n} = -e_{i,n}$ for all i.

In particular, every $e_{j,n}$ with $j > n$ is either zero or "loads" on exactly one $e_{i,n}$ with $i \in \{1, \dots, n\}$ with coefficient 1 or -1. This leads to a simple connection between the infinite expansion of a function $f = \sum_{j=1}^{\infty} f_j e_j$ in the eigenfunctions e_j of continuous Brownian motion and the finite expansion $\vec{f_n} = \sum_{i=1}^{n} f_{i,n} e_{i,n}$ of the discretized function $\vec{f_n}$ in the eigenvectors $e_{j,n}$ of discretized Brownian motion, as follows. Assuming that the series $f(x) = \sum_{j=1}^{\infty} f_j e_j(x)$ converges pointwise, we can use (2.16), which says that $(\vec{e_j})_n = \sqrt{n_+} e_{j,n}$, and (i)-(iii) to see that the coefficients in $\vec{f_n}$ are given by

$$f_{i,n} = \sum_{j=0}^{\infty} f_j (e_{j,n})^T e_{i,n} = \sqrt{n_+} \sum_{l=0}^{\infty} (f_{(2n+1)l+i} - f_{(2n+1)l+2n+2-i}). \tag{2.36}$$

The terms of this last series correspond to the consecutive periods of lengths $(2n+1)$. Exactly two of the inner products per period are nonzero and they yield coefficients 1 and -1 respectively. The formula is an example of the *aliasing* effect in signal analysis: the energy of the function f at frequencies j higher than the Nyquist frequency n, whose fluctuations fall between the grid points, is represented at the lower frequencies.

The scaling $\sqrt{n+}$ results from the normalisation of the vectors $e_{i,n}$ in $\mathbb{R}^n$. However, even apart from the normalisation the correspondence between the discrete and continuous coefficients is imperfect. By writing (2.36) in the form

$$\frac{f_{i,n}}{\sqrt{n_+}} = f_i - f_{2n+2-i} + \sum_{l=1}^{\infty} (f_{(2n+1)l+i} - f_{(2n+1)l+2n+2-i}),$$

we see that $f_{i,n}/\sqrt{n_+}$ is in general not equal to f_i. The "harmonic frequencies" at periods $2n+1$ add to a frequency at $i \in \{1, 2, \dots, n\}$, and the frequencies mirrored around the midpoints of the blocks subtract from it.

It is clear from the preceding display that a given discrete sequence $(f_{i,n})$ can be obtained from the infinite sequence $(f_{1,n}, f_{2,n}, \dots, f_{n,n}, 0, 0, \dots)/\sqrt{n_+}$ of L^2 coefficients, but also from many other infinite sequences (f_j). Because the data model (2.1) depends on f only through the discrete sequence $(f_{i,n})$, there is clearly no hope to recover which of these infinite sequences would be the "true" sequence. Furthermore, for a given fixed infinite sequence the values of the array $(f_{i,n})$ will change with n, and for some reasonable infinite sequences the series defining the discrete coefficients may not even converge. (We obtained the preceding display under the assumption that the series $\sum_j f_j e_j(x)$ converges pointwise.) The following lemma shows that the infinite series is essentially a Fourier series, and hence this less than perfect correspondence is disappointing.

Lemma 2.33. *For a given $f : [0,1] \to \mathbb{R}$ in $L_2[0,1]$, the expansion $f = \sum_j f_j e_j$ is derived from the Fourier series of the function $x \mapsto e^{i\pi x/2} f(x)$ on $[0,2]$, where f is extended to $[0,2]$ by symmetry about 1. In particular, if $f \in C^\alpha[0,1]$ for some $\alpha > 0$ and $f(0) = 0$, then*

$$f(x) = \sum_{j=1}^{\infty} f_j e_j(x), \qquad \textit{uniformly in } x.$$

Proof. The function $x \mapsto e^{i\pi x/2} f(x)$, with f extended as indicated, is periodic (i.e. it has the same value at 0 and 2) and contained in $L_2[0,2]$. Its Fourier series can be written in the form

$$e^{i\pi x/2} f(x) = \sum_{j\in\mathbb{Z}} c_j e^{i\pi j x} \tag{2.37}$$

for some $c_j \in \mathbb{C}$ and hence

$$f(x) = \sum_{j\in\mathbb{Z}} c_j e^{i\pi(j-\frac{1}{2})x}.$$

Since f is real, the complex part of the right side vanishes, while the real part can be written in the form

$$f(x) = \sum_{j\in\mathbb{Z}} \bigl[a_j \cos\bigl(\pi x(j-1/2)\bigr) - b_j \sin\bigl((j-1/2)\pi x\bigr)\bigr],$$

for $a_j, b_j \in \mathbb{R}$. Since f is symmetric about 1, the antisymmetric cosine part vanishes, while the terms with $j \le 0$ of the sine part can be united with terms with $j \ge 1$. This gives an expansion in terms of the eigenfunctions e_j. By the orthogonality of these functions the resulting expansion is unique.

If $f \in C^\alpha[0,1]$, then the extended function $x \mapsto e^{i\pi x/2} f(x)$ is contained in $C^\alpha[0,2]$ and hence we have uniform convergence in (2.37). The uniform convergence is retained under multiplying left and right with $e^{-i\pi x/2}$. □

As a consequence of the lemma, the speed at which the f_j tend to zero as $j \to \infty$ can be interpreted in the sense of Sobolev smoothness. However, this is not easily comparable to the smoothness of the corresponding array $(f_{i,n})$. In fact, if f is contained in a Sobolev space of order α for $\alpha \le 1/2$, that is $\sum_j j^{2\alpha} f_j^2 < \infty$, then the aliased coefficients may not even be well defined.

Polished tail sequences

In [Szabó et al., 2015] a function f, or rather its infinite series of coefficients (f_j) relative to a given eigenbasis, is defined to be *polished tail* if for some $L, \rho > 0$ and all sufficiently large m,

$$\sum_{j=m}^{\infty} f_j^2 \leq L \sum_{j=m}^{\rho m} f_j^2. \tag{2.38}$$

This reduces to the "discrete polished tail" condition (2.10) if applied to the infinite sequences $(f_{1,n}, f_{2,n}, \ldots, f_{n,n}, 0, 0, \ldots)/\sqrt{n_+}$. For general sequences (f_j) the relationship is less perfect, but for typical examples the two concepts agree.

Example 2.34 (Self-similar sequences)**.** In [Szabó et al., 2015] an infinite sequence (f_j) is defined to be *self-similar* of order $\alpha > 0$ if for some positive constants M, ρ, L and every m

$$\sup_{j\geq 1} j^{1/2+\alpha}|f_j| \leq M \qquad \text{and} \qquad \sum_{j=m}^{\rho m} f_j^2 \geq M^2 L m^{-2\alpha}.$$

Particular examples are the sequences with the exact order $|f_j| \asymp j^{-1/2-\alpha}$. Self-similar sequences are easily seen to be polished tail for every $\alpha > 0$ and arbitrary $\rho > 1$. For $\alpha \leq 1/2$ the corresponding function is not necessarily well defined at every point and the series (2.36) defining the aliased coefficients may diverge. However, for $\alpha > 1/2$ the induced array $(f_{i,n})$ is well defined and also discrete polished tail in the sense of (2.10).

To see this, first note that for $\ell \geq 1$ and taking M equal to 1 for simplicity we have

$$|f_{(2n+1)\ell+i}| \vee |f_{(2n+1)\ell+2n+2-i}| \lesssim \frac{1}{n^{1/2+\alpha}\ell^{1/2+\alpha}}.$$

This shows that the series (2.36) that defines the aliased coefficients converges. Furthermore, we see that the rescaled coefficients $\tilde{f}_{i,n} = f_{i,n}/\sqrt{n_+}$ satisfy $|\tilde{f}_{i,n} - f_i| \lesssim n^{-1/2-\alpha}$, so that $|\tilde{f}_{i,n}| \lesssim i^{-1/2-\alpha} + n^{-1/2-\alpha}$ and the left side of (2.10) satisfies

$$\sum_{i=m}^{n} \tilde{f}_{i,n}^2 \lesssim \frac{1}{m^{2\alpha}} + \frac{1}{n^{2\alpha}} \lesssim \frac{1}{m^{2\alpha}}.$$

We wish to show that the right side of (2.10) is lower bounded by the expression on the right, where we may assume that m satisfies $\rho m \leq n$, because otherwise there is nothing to prove. First we note that

$$|\tilde{f}_{i,n}^2 - f_i^2| = |\tilde{f}_{i,n} - f_i|\,|\tilde{f}_{i,n} + f_i| \lesssim \frac{1}{n^{1/2+\alpha}}\Big(|f_i| + \frac{1}{n^{1/2+\alpha}}\Big).$$

It follows that for some universal constant C

$$\sum_{i=m}^{\rho m\wedge n}\tilde{f}_{i,n}^2\geq\sum_{i=m}^{\rho m}f_i^2-\frac{C(\rho-1)m}{n^{1+2\alpha}}-C\sum_{i=m}^{\rho m}\frac{|f_i|}{n^{1/2+\alpha}}\gtrsim\frac{1}{m^{2\alpha}}\Bigl(L-\frac{2C(\rho-1)}{\rho^{1/2+\alpha}}\Bigr).$$

For sufficiently large L the constant in the last display is positive.

Example 2.35. The sequence $f_j=j^{-1/2-\alpha}$ is easily seen to be polished tail for every $\alpha>0$, as is also noted in Example 2.34. We shall show that the corresponding array $(f_{i,n})$ is also discrete polished tail in the sense of (2.10), for any $\alpha>0$, thus extending Example 2.34 to the range $\alpha\in(0,1/2]$. This refinement is possible by the exact form of the f_j, which allows us to exploit cancellation of positive and negative terms in (2.36).

To prove the claim we first apply the mean value theorem to find that, for every $\ell\geq 1$,

$$|f_{(2n+1)\ell+i}-f_{(2n+1)\ell+2n+2-i}|\lesssim\frac{1}{n^{1/2+\alpha}\ell^{3/2+\alpha}}.$$

This shows that the series in (2.36) defining the discrete coefficients converges. Moreover,

$$|\tilde{f}_{i,n}|\lesssim\frac{2}{i^{1/2+\alpha}}+\sum_{\ell=1}^{\infty}|f_{(2n+1)\ell+i}-f_{(2n+1)\ell+2n+2-i}|\lesssim\frac{1}{i^{1/2+\alpha}}+\frac{1}{n^{1/2+\alpha}}.$$

Consequently, the left side of (2.10) satisfies

$$\sum_{i=m}^{n}\tilde{f}_{i,n}^2\lesssim\frac{1}{m^{2\alpha}}+\frac{1}{n^{2\alpha}}\lesssim\frac{1}{m^{2\alpha}}.$$

Furthermore, since all terms in (2.36) are positive, we also have

$$\tilde{f}_{i,n}\geq\frac{1}{i^{1/2+\alpha}}-\frac{1}{(2n+2-i)^{1/2+\alpha}}\gtrsim\frac{1}{i^{1/2+\alpha}},$$

for $i\leq cn$ and any fixed $c<1$. To bound the right side of (2.10) we may assume that m satisfies $\rho m\leq n$, because otherwise there is nothing to prove. Then choosing $c<1$ and $\rho>1$ such that $c\rho>1$, we have

$$\sum_{i=m}^{\rho m\wedge n}\tilde{f}_{i,n}^2\geq\sum_{i=m}^{c\rho m}\tilde{f}_{i,n}^2\gtrsim\sum_{i=m}^{c\rho m}\frac{1}{i^{1+2\alpha}}\geq\int_m^{c\rho m}\frac{1}{t^{1+2\alpha}}\,\mathrm{d}t\gtrsim\frac{1}{m^{2\alpha}}.$$

The right side is seen to be bigger than a multiple of the left side of (2.10). This proves the claim.

Prior polished tail sequences

According to the Bayesian model the true function f is a realisation of the prior process W^c. In this section we show that almost every such realisation gives rise to a discrete polished tail array. Consequently, for a Bayesian who believes in her prior, the polished tail condition is reasonable. For a non-Bayesian the following proposition is also of interest, as it shows that polished tail functions are abundant.

The proof of the statement will be based on the Karhunen-Loève expansion. For standard Brownian motion $W^1 = (W_t^1 : t \in [0,1])$ this is given by

$$W_t^1 = \sum_{j=1}^{\infty} \frac{Z_j}{(j-1/2)\pi} e_j(t).$$

Here $Z_1, Z_2, \ldots$ are independent standard normal random variables. We see that the prior W^c is given by $\sum_j f_j e_j$, for the infinite sequence $f_j = \sqrt{c} Z_j/((j-1/2)\pi)$. We shall show that the induced array $f_{j,n}$ defined by (2.36) is discrete polished tail, almost surely.

In fact a more general result holds for any Gaussian series with polynomially decaying singular values relative to the eigenbasis of Brownian motion.

Proposition 2.36. *For given $\alpha > 0$ and $\delta \in \mathbb{R}$ set*

$$W_t = \sum_{j=1}^{\infty} \frac{Z_j}{(j+\delta)^{1/2+\alpha}} e_j(t), \qquad t \in [0,1],$$

where $Z_1, Z_2, \ldots$ are independent standard normal random variables. Then almost every realisation of W is both polished tail in the sense of (2.38) and discrete polished tail in the sense of (2.10).

Proof. The first claim is proved in Proposition 3.5 of [Szabó et al., 2015]. To prove that W is discrete polished tail, we consider the coefficients given in (2.36), except the factor $\sqrt{n_+}$:

$$\tilde{W}_{i,n} = \sum_{l=0}^{\infty} \left(\frac{Z_{(2n+1)l+i}}{(\delta + (2n+1)l + i)^{1/2+\alpha}} - \frac{Z_{(2n+1)l+2n+2-i}}{(\delta + (2n+1)l + 2n + 2 - i)^{1/2+\alpha}} \right).$$

In view of Lévy's continuity theorem this array consists for each n of independent zero-mean normal random variables $W_{1,n}, W_{2,n}, \ldots, W_{n,n}$ with variances

$$\operatorname{var}(\tilde{W}_{i,n}) \asymp \sum_{l=0}^{\infty} \left(\frac{1}{((2n+1)l+i)^{2\alpha+1}} + \frac{1}{((2n+1)l+2n+2-i)^{2\alpha+1}} \right).$$

Now let $L, \rho > 0$ and consider the event

$$E_m = \left\{ \sum_{i=m}^{n} \tilde{W}_{i,n}^2 > L \sum_{i=m}^{\rho m} \tilde{W}_{i,n}^2 \right\}.$$

Setting

$$X = L \sum_{i=m}^{\rho m} \tilde{W}_{i,n}^2 - \sum_{i=m}^{n} \tilde{W}_{i,n}^2 = (L-1) \sum_{i=m}^{\rho m} \tilde{W}_{i,n}^2 - \sum_{i=\rho m+1}^{n} \tilde{W}_{i,n}^2,$$

we see that E_m has probability $P(E_m) = P(X < 0)$. We then have by Markov's inequality that for $\eta > 0$

$$P(E_m) = P\left(X < 0\right) \le P\left(|X - \mathrm{E}X| \ge \mathrm{E}X\right) \le \frac{\mathrm{E}|X - \mathrm{E}X|^\eta}{(\mathrm{E}X)^\eta}.$$

We proceed to bound the expectation of X. Clearly the variance of $\tilde{W}_{i,n}$ is bigger than a constant times $i^{-1-2\alpha}$. Since $i \le n$, it is also smaller than

$$\frac{1}{i^{2\alpha+1}} + \frac{3}{(2n+1)^{2\alpha+1}} + 2\int_1^\infty \frac{1}{((2n+1)x + i)^{2\alpha+1}}\,\mathrm{d}x \le \frac{1}{i^{2\alpha+1}} + \frac{L_1}{n^{2\alpha+1}}$$

for some $L_1 > 0$. It follows that

$$\begin{aligned}\mathrm{E}X &\ge (L-1)\sum_{i=m}^{\rho m} \frac{1}{i^{2\alpha+1}} - \sum_{i=\rho m+1}^{n} \frac{1}{i^{2\alpha+1}} - L_1 \sum_{i=\rho m+1}^{n} \frac{1}{n^{2\alpha+1}} \\ &\ge \frac{1}{2\alpha}\frac{1}{m^{2\alpha}}\big[(L-1)(1-\rho^{-2\alpha}) - (1+L_1)\rho^{-2\alpha}\big].\end{aligned}$$

We choose L and ρ large enough so that this is positive. Applying the Marcinkiewicz-Zygmund inequality and next Hölder's inequality with conjugate parameters $(\eta/2, \eta/(\eta-2))$, we see that $\mathrm{E}|X - \mathrm{E}X|^\eta$ is for $\eta > 2$ bounded by a constant times

$$\begin{aligned}&\mathrm{E}\bigg(\sum_{i=m}^{\rho m} (L-1)^2 \big(\tilde{W}_{i,n}^2 - \mathrm{E}\tilde{W}_{i,n}^2\big)^2 + \sum_{i=\rho m+1}^{n} \big(\tilde{W}_{i,n}^2 - \mathrm{E}\tilde{W}_{i,n}^2\big)^2 \bigg)^{\eta/2} \\ &\lesssim \mathrm{E}\left(\left(\sum_{i=m}^{n} |\tilde{W}_{i,n}^2 - \mathrm{E}\tilde{W}_{i,n}^2|^\eta i^{\eta/2} \right)^{2/\eta} \left(\sum_{i=m}^{n} i^{-\eta/(\eta-2)} \right)^{1-2/\eta} \right)^{\eta/2} \\ &= \sum_{i=m}^{n} \mathrm{E}|\tilde{W}_{i,n}^2 - \mathrm{E}\tilde{W}_{i,n}^2|^\eta i^{\eta/2} \left(\sum_{i=m}^{n} i^{-\eta/(\eta-2)} \right)^{\eta/2-1}.\end{aligned}$$

Since $\mathrm{E}|\tilde{W}_{i,n}^2 - \mathrm{E}\tilde{W}_{i,n}^2|^\eta \asymp \mathrm{var}(\tilde{W}_{i,n})^\eta \lesssim i^{-(1+2\alpha)\eta}$, we conclude that

$$\begin{aligned}\mathrm{E}|X - \mathrm{E}X|^\eta &\lesssim \sum_{i=m}^{n} i^{(1/2-(1+2\alpha))\eta} \left(\sum_{i=m}^{n} i^{-\eta/(\eta-2)}\right)^{\eta/2-1} \\ &\lesssim m^{1-(1/2+2\alpha)\eta+\eta/2-1-\eta/2} = m^{-(1/2+2\alpha)\eta},\end{aligned}$$

hence the $P(E_m)$ are bounded by a multiple of $m^{-\eta/2}$ and thus summable over m for $\eta > 2$. It follows by the Borel-Cantelli lemma that the event E_m occurs at most finitely many times with probability one. □

2.5 Discussion

The model (2.1) can also be formulated directly in terms of the coordinates $(f_{i,n})$ of $\vec{f_n}$ relative to the eigenbasis $e_{j,n}$ of the prior covariance matrix U_n. For O_n the orthogonal matrix with rows the eigenvectors $e_{j,n}$ of U_n, the definition of $f_{j,n}$ gives

$$O_n\vec{Y_n} = O_n\vec{f_n} + O_n\vec{\varepsilon_n} = \begin{pmatrix} f_{1,n} \\ f_{2,n} \\ \vdots \\ f_{n,n} \end{pmatrix} + O_n\vec{\varepsilon_n}.$$

By the orthonormality of O_n the error vector $O_n\vec{\varepsilon_n}$ is equal in distribution to $\vec{\varepsilon_n}$, whence $\tilde{Y}_n = O_n\vec{Y_n}$ can be considered a vector of observations in a normal mean model with mean vector $(f_{i,n})$. Under the prior W^c on f, given c the vector $(f_{1,n}, \dots, f_{n,n})^T = O_n^{-1}\vec{f_n}$ possesses a mean zero normal distribution with covariance matrix $cO_n^{-1}U_nO_n = \mathrm{diag}(c\lambda_{i,n})$. Prior and data model both factorise over the coordinates, and it can be seen that under the posterior distribution given c the variables $f_{1,n}, \dots, f_{n,n}$ are again independent with

$$f_{i,n} \,|\, \vec{Y_n}, c \sim \mathcal{N}\left(\frac{c\lambda_{i,n}}{1+c\lambda_{i,n}}\tilde{Y}_{i,n}, \frac{c\lambda_{i,n}}{1+c\lambda_{i,n}}\right).$$

This gives a representation of the posterior distribution different from, but equivalent to (3.3).

In this form the model resembles the infinite Gaussian sequence model (or white noise model). A difference is that presently the sequence is of length n instead of infinite, and the parameter vector $(f_{1,n}, \dots, f_{n,.n})$ changes with n, even it refers to a single true function f. The discussion in Section 2.4 shows that this difference is not trivial.

Likelihood-based empirical Bayes and hierarchical Bayes estimation of the scale parameter c in the infinite sequence model were studied in [Szabó et al., 2013].

Besides considering the finite sequence model, here we also treat the risk-based empirical Bayes method and allow more general priors. A main difference is that we have focused on the coverage of the credible sets. Such coverage is also studied in [Szabó et al., 2015], but only for the likelihood-based empirical Bayes method in the infinite-sequence model with $\mathcal{N}(0, i^{-1-2\alpha})$-priors and α taken equal to the smoothing parameter. The same model is studied in [Ray, 2015], where exact credible sets are obtained when considering Sobolev spaces with negative exponent. The focus in this chapter on balls in the space of the finite vectors $\vec{f}_n$ of function values allows us to make the connection to the correctness of a fraction of the credible intervals, as in Corollary 2.4. The present treatment also differs in its technical details and proofs, in that our results are directly formulated in terms of the criterion that is optimized, whereas [Szabó et al., 2015, 2013] make the derivative of the criterion intercede. The present approach gives better insight and allows to state the contribution of the (discrete) polished tail condition more precisely, with the possibility of generalisation to the good bias condition (2.27), which is dependent both on the method and the prior.

Throughout, we limit the estimator to the interval I_n. This is reasonable, since the optimal rate of rescaling for functions in a class of smoothness α satisfies $cn \asymp n^{\delta}$, where $\delta = m/(1+2\alpha) \in (0, m]$ (if $\alpha \in (0, m)$ or $\alpha \in (0, m/2)$ in the risk-based and likelihood-based methods).

We consider the hierarchical Bayes only with the usual inverse Gamma prior on the scaling parameter. From the proof it is not difficult to see that the result extends to more general priors. For instance if $c^{-r} \sim \Gamma(\kappa, \lambda)$, for some $r > 0$, then the theorem is again true, but with the term $1/c$ replaced by $(1/c)^r$. A choice $r \leq 1$ does not change much, but the choice $r > 1$ has an adverse effect on the rate of contraction for Sobolev classes: optimality is obtained only for $\alpha \leq (1/r + m - 1)/2$.

The assumption that the errors in the regression model are normally distributed is crucial to define the posterior distribution and credible sets. However, the derivation of the properties of these objects uses only that the errors have mean zero and finite fourth moments. Thus the standard normal model may be misspecified. This is true in particular regarding the assumption of unit variance, although it would be preferable to extend our results to allow for a prior on this variance.

The study of credible bands, rather than credible balls or credible intervals in a fractional sense, would require control of the bias of the posterior mean in a uniform sense. This involves properties of the eigenvectors of the priors and goes beyond the "ℓ_2-theory" considered in this chapter. The bias in the example of Brownian motion is considered in detail in Chapter 1. We will employ this in the study of credible bands in the next chapter.

2.6 Technical proofs

In this section we give the proofs of Corollary 2.4 and Propositions 2.10, 2.15 and 2.16.

Proof of Corollary 2.4

In the Bayesian model (2.2) we have $\vec{Y}_n = \vec{W}_n^c + \vec{\varepsilon}_n$ for independent vectors $\vec{W}_n^c$ and $\vec{\varepsilon}_n$. The marginal posterior distribution of $f(x)$ given c and $\vec{Y}_n$ is the conditional law of W_x^c given c and $\vec{Y}_n$. By the assumed Gaussianity, this is a normal law with mean the conditional expectation $\hat{f}_{n,c}(x) = \mathrm{E}(W_x^c \,|\, \vec{Y}_n, c)$ and variance equal to

$$\begin{aligned} s_n^2(c,x) &= \mathrm{var}\left[W_x^c \,|\, c, \vec{Y}_n\right] = \mathrm{var}\big[W_x^c - \mathrm{E}(W_x^c \,|\, \vec{Y}_n, c) \,|\, c\big] \\ &= \inf_a \mathrm{E}\big[(W_x^c - a^T\vec{Y}_n)^2 \,|\, c\big]. \end{aligned}$$

When evaluated at a design point $x = x_{i,n}$, this is equal to the i^{th} diagonal element of the posterior covariance matrix $I - \Sigma_{n,c}^{-1}$. Hence the sum of the posterior variances over the design points is the trace of this matrix. It follows that for all $i \in J_n$ we have

$$s_n^2(c, x_{i,n}) \gtrsim \frac{1}{n}\,\mathrm{tr}(I - \Sigma_{n,c}^{-1}) = \frac{s_n^2(c)}{n},$$

where $s_n^2(c)$ is given in (2.26). It follows that for $i \in J_n$ the radius $Mr_n(c, x_{i,n})$ of the empirical Bayes interval $\hat{C}_{n,\eta,M}(x_{i,n})$ is bounded from below (up to a universal multiple) of $Mz_\eta s_n(c)/\sqrt{n}$.

The function f fails to belong to the empirical Bayes interval $\hat{C}_{n,\eta,M}(x)$ if and only if $|f(x) - \hat{f}_{n,\hat{c}_n}(x)| \geq Mr_n(\hat{c}_n, \eta, x)$. Therefore, by Markov's inequality

$$\begin{aligned} \frac{1}{n}\sum_{i\in J_n} 1\{f \notin \hat{C}_{n,\eta,M}(x_{i,n})\} &\leq \frac{1}{n}\sum_{i\in J_n} \frac{|f(x_{i,n}) - \hat{f}_{n,\hat{c}_n}(x_{i,n})|^2}{M^2 r_n^2(\hat{c}_n, \eta, x_{i,n})} \\ &\lesssim \frac{\|\vec{f}_n - \hat{f}_{n,\hat{c}_n}\|^2}{M^2 z_\eta^2 s_n^2(\hat{c}_n)}. \end{aligned}$$

As noted in the first paragraph of the proof of Theorem 2.17, $s_n^2(\hat{c}_n)$ is asymptotic to the square radius $r_n^2(\hat{c}_n, \eta')$ of the credible balls of the form (2.7), for any $\eta' \in (0,1)$. Therefore, if the left-hand is bigger than $1-\gamma$, then $f \notin \hat{C}_{n,M',\eta}$

for M' a multiple of Mz_η. By Theorem 2.3 this is the case with probability tending to zero if M' is sufficiently large, which it is if M is large. The result then follows, since

$$\frac{1}{n}\sum_{i\in J_n} 1\{f\in \hat C_{n,\eta,M}(x_{i,n})\} + \frac{1}{n}\sum_{i\in J_n} 1\{f\notin \hat C_{n,\eta,M}(x_{i,n})\} = \frac{|J_n|}{n}\to 1.$$

If the function f fails to belong to the hierarchical interval $\hat C_{n,\eta,M}(x)$, then $|f(x)-\hat f_{n,\bar c_n}(x)| \geq Mr_n(\bar c_n,\eta_2,x)$, for $\bar c_n$ as defined in the proof of Theorem 2.27. The rest of the proof is similar to the proof of the empirical Bayes intervals.

The assertions concerning the radii are immediate from the corresponding assertions of Theorem 2.3 and the equivalences $s_n(c,x_{i,n}) \asymp s_n(c)/\sqrt{n} \asymp r_n(c,\eta)/\sqrt{n}$ uniformly for $i\in J_n$ under the extra assumption on the posterior variances.

Proof of final assertion of Lemma 2.14

That D^R_{2,n^2} and s_{n^2} behave as claimed is immediate from Lemma 2.44; we only need consider the behaviour of D^L_{2,n^2}. The derivative of this function is given by $c\mapsto c^{-1}D^R_{2,n^2}(c)$ and hence is asymptotic to $c^{-1}(cn^2)^{1/m}k_n(c)$ uniformly on the interval $[l_n/n^2, n^{2m-2}]$, for any $l_n\to\infty$. Here $k_n(c)=1+\log(cn^2)$ for $cn^2\leq n^m$ and $k_n(c)=1+\log(n^{2m}/(cn^2))$ for $cn^2\geq n^m$. Now, as $cn^2\geq l_n\to\infty$, we have for $cn^2\leq n^m$

$$\int_0^c s^{-1}(sn^2)^{1/m}k_n(s)\,ds = \int_0^{cn^2} u^{1/m-1}(1+\log u)\,du \asymp (cn^2)^{1/m}\log(cn^2),$$

since $\int_0^t u^{1/m-1}\log u\,du = mt^{1/m}\log t - m^2t^{1/m}$. Furthermore, we have for $cn^2\in[n^m,n^{2m}]$

$$\begin{aligned}\int_0^c s^{-1}(sn^2)^{1/m}k_n(s)\,ds &\asymp n\log n + \int_{n^m}^{cn^2} u^{1/m-1}\bigl(1+\log n^{2m}-\log u\bigr)\,du\\ &= n\log n + m\bigl(1+\log(n^{2m}/u)\bigr)u^{1/m}\big|_{n^m}^{cn^2} + m\int_{n^m}^{cn^2} u^{1/m-1}\,du\\ &\asymp (cn^2)^{1/m}\bigl(1+\log(n^{2m}/cn^2)\bigr).\end{aligned}$$

Combining the two displays we see that in both cases the left side is asymptotic to $(cn^2)^{1/m}k_n(c)$. This order does not change if we limit the integrals to the interval $[l_n/n^2,c]$, for $l_n\to\infty$ slowly. It follows that $D^L_{2,n^2}(c)$ has this order, provided the integral $\int_0^{l_n/n^2}(D_{2,n^2}^L)'(s)\,ds$ is of lower order. Since $(D_{2,n^2}^L)'(s) \lesssim \sum_{i=1}^n\sum_{j=1}^n (ij)^{-2m}sn^4$, the latter integral is bounded by a multiple of l_n^2, which is of lower order again if $l_n \to \infty$ sufficiently slowly.

Proof of Proposition 2.10

The proof is based on two lemmas.

Lemma 2.37. *For the functions in both (2.23) and (2.24) and any c and $s < t$ in $(0,\infty)$ we have*

$$\begin{aligned}\operatorname{var}\big[R_{1,n}(c,f)\big] &\lesssim D_{1,n}(c,f),\\ \operatorname{var}\big[R_{2,n}(c)\big] &\lesssim D_{2,n}(c),\\ \operatorname{var}\big[R_{1,n}(s,f)-R_{1,n}(t,f)\big] &\lesssim \frac{(t-s)^2D_{1,n}(s,f)}{s^2},\\ \operatorname{var}\big[R_{2,n}(s)-R_{2,n}(t)\big] &\lesssim \frac{(t-s)^2D_{2,n}(s)}{s^2}.\end{aligned}$$

Proof. For the risk-based remainder $R_{1,n}^R$ given in (2.23) we have

$$\operatorname{var}\big[R_{1,n}^R(c,f)\big] = 4\sum_{j=1}^n \frac{f_{j,n}^2}{(1+c\lambda_{j,n})^4} \le 4D_{1,n}^R(c,f).$$

The bound on the variance of the likelihood-based remainder $R_{1,n}^L$ in (2.24) is very similar. For $R_{2,n}^R$ in (2.23) we have

$$\operatorname{var}\big[R_{2,n}^R(c)\big] = 2\sum_{j=1}^n \frac{(2c\lambda_{j,n}+c^2\lambda_{j,n}^2)^2}{(1+c\lambda_{j,n})^4} \le 8\sum_{j=1}^n \frac{(c\lambda_{j,n})^2}{(1+c\lambda_{j,n})^2} = 8D_{2,n}^R(c).$$

For the likelihood-based remainder in (2.24) we have

$$\operatorname{var}\big[R_{2,n}^L(c)\big] = 2\sum_{j=1}^n \frac{(c\lambda_{j,n})^2}{(1+c\lambda_{j,n})^2} = 2D_{2,n}^R(c) \le 4D_{2,n}^L(c),$$

in view of the inequality $\log(1+x) - x/(1+x) \ge x^2/(1+x)^2/2$ for $x > 0$.

The third and fourth assertions of the lemma follow by applying Lemma 2.47. For the risk-based remainder given in (2.23), we use the lemma with the choices:

- for $R_{1,n}^R$: $(\alpha,\beta) = (0,2)$, $a_j = 2f_{j,n}$, $U_j = Z_{j,n}$ and $(\delta,\gamma) = (0,2)$, where the sum in (2.42) becomes $4D_{1,n}^R$,

- for $R^R_{2,n}$: $(\alpha, \beta) = (0, 2)$, $a_j = 1$, $U_j = Z^2_{j,n} - 1$ and $(\delta, \gamma) = (2, 2)$, where the sum in (2.42) becomes $D^R_{2,n}$.

For the likelihood-based remainder, given in (2.24), we use the lemma with the choices:

- for $R^L_{1,n}$: $(\alpha, \beta) = (0, 1)$, $a_j = 2f_{j,n}$, $U_j = Z_{j,n}$ and $(\delta, \gamma) = (0, 1)$, where the sum in (2.42) becomes $4D^L_{1,n}$,
- for $R^L_{2,n}$: $(\alpha, \beta) = (1, 0)$, $a_j = -1$, $U_j = Z^2_{j,n} - 1$ and $(\delta, \gamma) = (2, 2)$, where the sum in (2.42) will become $D^R_{2,n}$, which is bounded by a multiple of $D^L_{2,n}$.

This concludes the proof. □

Lemma 2.38. *For the functions in both (2.23) and (2.24) and any $s < t$ in I_n we have*

$$\big|D_{1,n}(s, f) - D_{1,n}(t, f)\big| \lesssim \frac{|t - s| D_{1,n}(s, f)}{s},$$
$$\big|D_{2,n}(s) - D_{2,n}(t)\big| \lesssim \frac{|t - s| s^2_n(s)}{s}.$$

Proof. By Lemma 2.46 with $(\alpha, \beta) = (0, 2)$ and $D^R_{1,n}$ as in (2.23) we have

$$|D^R_{1,n}(s, f) - D^R_{1,n}(t, f)| \leq \frac{|s - t|}{s} \sum_{j=1}^{n} \frac{f^2_{j,n}}{(1 + s\lambda_{j,n})^2} = \frac{|s - t|}{s} D^R_{1,n}(s, f).$$

The function $D^L_{1,n}$ in (2.24) can be treated similarly, with the choice $(\alpha, \beta) = (0, 1)$.

Applying Lemma 2.46 with $(\alpha, \beta) = (2, 0)$ to $D^R_{2,n}(c)$, we find

$$|D^R_{2,n}(s) - D^R_{2,n}(t)| \leq \frac{|s - t|}{s} \sum_{j=1}^{n} \frac{s\lambda_{j,n}}{(1 + s\lambda_{j,n})^2} \leq \sum_{j=1}^{n} \frac{s\lambda_{j,n}}{1 + s\lambda_{j,n}}.$$

The right side is $s^2_n(s)$, by definition (2.26). Applying the mean value theorem to $D^L_{2,n}$ in (2.24) we find for some $s \leq \xi \leq t$,

$$|D^L_{2,n}(s) - D^L_{2,n}(t)| \leq |s - t| \sum_{j=1}^{n} \frac{\xi\lambda^2_{j,n}}{(1 + \xi\lambda_{j,n})^2} \leq |s - t| \sum_{j=1}^{n} \frac{\lambda_{j,n}}{1 + \xi\lambda_{j,n}}$$
$$\leq \frac{|s - t|}{s} \sum_{j=1}^{n} \frac{s\lambda_{j,n}}{1 + s\lambda_{j,n}}.$$

This concludes the proof. □

Proof of Proposition 2.10. Applying Lemmas 2.37 and 2.38, we see that for any $s < t$ in I_n we have

$$\begin{aligned}
&\operatorname{var}\left(\frac{R_{1,n}(s,f)}{D_n(s,f)} - \frac{R_{1,n}(t,f)}{D_n(t,f)}\right)/2\\
&\le \operatorname{var}\left(\frac{R_{1,n}(s,f) - R_{1,n}(t,f)}{D_n(s,f)}\right) + \operatorname{var}\big[R_{1,n}(t,f)\big]\left(\frac{D_n(s,f) - D_n(t,f)}{D_n(s,f)D_n(t,f)}\right)^2\\
&\lesssim \frac{(t-s)^2}{s^2 D_n(s,f)} + \frac{(t-s)^2}{s^2 D_n(t,f)}\frac{D_{1,n}^2(s,f) + s_n^4(s)}{D_n^2(s,f)}\\
&\lesssim \frac{(t-s)^2}{s^{2+1/m} n^{1/m}},
\end{aligned}$$

since $D_n(s,f) \ge D_{2,n}(s) \gtrsim (sn)^{1/m} \asymp s_n^2(s)$ by Lemma 2.14. Similarly, applying Lemma 2.37 we see that

$$\operatorname{var}\left(\frac{R_{1,n}(s,f)}{D_n(s,f)}\right) \lesssim \frac{1}{D_n(s,f)} \lesssim \frac{1}{(sn)^{1/m}}$$

by Lemma 2.14. The result for $R_{1,n}$ follows from the preceding two displays, by application of Lemma 2.48. The assertion for $R_{2,n}$ is proved analogously, from the other parts of Lemmas 2.37 and 2.38. □

Proof of Proposition 2.15

In addition to Lemma 2.38 we need the following lemma.

Lemma 2.39. *For any c and any $s < t$ in $(0,\infty)$ we have*

$$\begin{aligned}
\operatorname{var}\big[R_{3,n}(c,f)\big] &\le 4D_{1,n}^R(c,f),\\
\operatorname{var}\big[R_{4,n}(c)\big] &\le 2D_{2,n}^R(c),\\
\operatorname{var}\big[R_{3,n}(s,f) - R_{4,n}(t,f)\big] &\lesssim \frac{(t-s)^2 D_{1,n}^R(s,f)}{s^2},\\
\operatorname{var}\big[R_{4,n}(s) - R_{4,n}(t)\big] &\lesssim \frac{(t-s)^2 D_{2,n}^R(s)}{s^2}.
\end{aligned}$$

Proof. For the first two inequalities we compute

$$\operatorname{var}\left[R_{3,n}(c,f)\right] = 4\sum_{j=1}^{n} \frac{(c\lambda_{j,n})^2 f_{j,n}^2}{(1+c\lambda_{j,n})^4} \le 4D_{1,n}^R(c,f),$$

$$\operatorname{var}\left[R_{4,n}(c)\right] = 2\sum_{j=1}^{n} \frac{(c\lambda_{j,n})^4}{(1+c\lambda_{j,n})^4} \le 2D_{2,n}^R(c).$$

The third and fourth inequalities follow by application of Lemma 2.47 with the following choices:

- for $R_{3,n}$: $(\alpha,\beta) = (1,1)$, $a_j = -2f_{j,n}$, $U_j = Z_{j,n}$ and $(\delta,\gamma) = (0,2)$, where the sum in (2.42) becomes $4D_{1,n}^R$.
- for $R_{4,n}$: $(\alpha,\beta) = (2,0)$, $a_j = 1$, $U_j = Z_{j,n}^2 - 1$ and $(\delta,\gamma) = (2,2)$, where the sum in (2.42) becomes $D_{2,n}^R$.

This concludes the proof. □

Proof of Proposition 2.15. Using Lemmas 2.39 and 2.38, we have for $s < t$ in I_n

$$\begin{aligned}
&\operatorname{var}\left(\frac{R_{3,n}(s,f)}{D_n^R(s,f)} - \frac{R_{3,n}(t,f)}{D_n^R(t,f)}\right)/2 \\
&\le \operatorname{var}\left(\frac{R_{3,n}(s,f)-R_{3,n}(t,f)}{D_n^R(s,f)}\right) + \operatorname{var}\left[R_{3,n}(t,f)\right]\left(\frac{D_n^R(s,f)-D_n^R(t,f)}{D_n^R(s,f)D_n^R(t,f)}\right)^2 \\
&\lesssim \frac{(t-s)^2}{s^2 D_n^R(s,f)} + \frac{(t-s)^2}{s^2 D_n^R(t,f)}\frac{D_{1,n}^R(s,f)^2 + s_n^4(s)}{D_n^R(s,f)^2} \\
&\lesssim \frac{(t-s)^2}{s^{2+1/m}n^{1/m}},
\end{aligned}$$

since $D_{1,n}^R \le D_n^R$ and $D_n^R(s,f) \ge D_{2,n}^R(s) \gtrsim (sn)^{1/m} \asymp s_n^2(s)$ by Lemma 2.14. Similarly, we have by Lemma 2.39

$$\operatorname{var}\left(\frac{R_{3,n}(s,f)}{D_n^R(s,f)}\right) \le \frac{1}{D_n^R(s,f)} \lesssim \frac{1}{(sn)^{1/m}},$$

by Lemma 2.14. The proposition with $D_n = D_n^R$ follows by an application of Lemma 2.48.

Since $D_n^L \ge D_n^R/2$, this immediately implies the proposition for the likelihood-based norming. The assertion for $R_{4,n}$ is proved analogously, from the other parts of Lemmas 2.39 and 2.38. □

Proof of Proposition 2.16

Lemma 2.40. *For $s \leq t$ we have*

$$\left|s_n^2(t) - s_n^2(s)\right| \lesssim \frac{|t-s| s_n^2(s)}{s}.$$

Proof. This is immediate from the definition of s_n^2 in (2.26) and Lemma 2.46 with $(\alpha, \beta) = (1, 0)$. □

Proof of Proposition 2.16. It is immediate from the definition of N_n that

$$\mathrm{E}\left[\frac{N_n(c)}{s_n^2(c)} - 1\right] = 0, \qquad \operatorname{var}\left[N_n(c)\right] \lesssim s_n^2(c).$$

Applying Lemma 2.47 with $(\alpha, \beta) = (1, 0)$, $a_j = 1$, $(\gamma, \delta) = (1, 1)$ and $g = s_n^2$, we find that for $s \leq t$

$$\operatorname{var}\left[N_n(s) - N_n(t)\right] \lesssim \frac{(t-s)^2 s_n^2(s)}{s^2}.$$

It follows by Lemma 2.14 that

$$\begin{aligned}
&\operatorname{var}\left(\frac{N_n(s)}{s_n^2(s)} - \frac{N_n(t)}{s_n^2(t)}\right) \\
&\quad \leq 2 \operatorname{var}\left(\frac{N_n(s) - N_n(t)}{s_n^2(s)}\right) + 2 \operatorname{var}\big[N_n(t)\big] \left(\frac{s_n^2(s) - s_n^2(t)}{s_n^2(s) s_n^2(t)}\right)^2 \\
&\quad \lesssim \frac{(t-s)^2}{s^2 s_n^2(s)} + \frac{(t-s)^2}{s^2 s_n^2(t)} \\
&\quad \lesssim \frac{(t-s)^2}{s^{2+1/m} n^{1/m}}.
\end{aligned}$$

The proposition follows by an application of Lemma 2.48. □

For Brownian motion, we can gain more insight in the behaviour of (part of) the function D_2^L.

Lemma 2.41. *For the Brownian motion prior and $c \in [\log n/n, n]$,*

$$\log \det \Sigma_{n,c} \sim \sqrt{cn}.$$

Proof. We want to find the determinant of the $n \times n$ matrix

$$\Sigma_{n,c} = c \begin{pmatrix} \frac{1}{c} + \frac{1}{n_+} & \frac{1}{n_+} & \frac{1}{n_+} & \cdots & \frac{1}{n_+} \\ \frac{1}{n_+} & \frac{1}{c} + \frac{2}{n_+} & \frac{2}{n_+} & \cdots & \frac{2}{n_+} \\ \frac{1}{n_+} & \frac{2}{n_+} & \ddots & & \vdots \\ \vdots & \vdots & & \frac{1}{c} + \frac{n-1}{n_+} & \frac{n-1}{n_+} \\ \frac{1}{n_+} & \frac{2}{n_+} & \cdots & \frac{n-1}{n_+} & \frac{1}{c} + \frac{n}{n_+} \end{pmatrix}$$

$$\sim \begin{pmatrix} 2 + \frac{c}{n_+} & -1 & 0 & \cdots & 0 \\ -1 & 2 + \frac{c}{n_+} & -1 & \cdots & 0 \\ 0 & -1 & \ddots & & \vdots \\ \vdots & \vdots & & 2 + \frac{c}{n_+} & -1 \\ 0 & 0 & \cdots & -1 & 1 + \frac{c}{n_+} \end{pmatrix}.$$

If we denote this determinant by d_n, we see that

$$d_n = \left(2 + \frac{c}{n_+}\right) d_{n-1} - d_{n-2},$$

with $d_1 = 1 + \frac{c}{n_+}$ and $d_2 = \left(2 + \frac{c}{n_+}\right)\left(1 + \frac{c}{n_+}\right) - 1$. Note that this is the same recurrence relation as (1.3) in Chapter 1. The solution is given by $d_n = A\lambda_+^n + B\lambda_-^n$, where

$$A = \frac{c^2 + cn_+(3 - \lambda_-) + n_+^2(1 - \lambda_-)}{(\lambda_+ - \lambda_-)\lambda_+ n_+^2}, \quad \lambda_\pm = 1 + \frac{c}{2n_+} \pm \frac{\sqrt{c}}{2\sqrt{n_+}}\sqrt{4 + \frac{c}{n_+}}.$$

Note that $\lambda_+\lambda_- = 1$. Since $\theta = \frac{c}{n_+} \to 0$ uniformly in $c \in I_n$, we have $\lambda_\pm \to 1$ and

$$A = \frac{(1 - \lambda_-)}{(\lambda_+ - \lambda_-)} + o(1) = \frac{\frac{1}{2}\left(\sqrt{\theta(4+\theta)} - \theta\right)}{\sqrt{\theta(4+\theta)}} + o(1) \to \frac{1}{2}.$$

It is easy to see that $B = \lambda_- A \sim A$. Furthermore, we have

$$\begin{aligned} \log(\lambda_+^n) &= n\left[\frac{\theta}{2} + \sqrt{\theta}\frac{\sqrt{4+\theta}}{2} - \frac{\theta}{2}\left(\frac{\sqrt{4+\theta}}{2}\right)^2 + O(\theta^{3/2})\right] \\ &= n\sqrt{\theta} + O(n\theta^{3/2}). \end{aligned}$$

Finally, we have

$$\log d_n - \log(A\lambda_+^n) = \log\left(1 + \frac{B}{A}\lambda_-^{2n}\right) \to 0.$$

The result follows. □

2.7 Technical results

Lemma 2.42. *Let $D_1 : I_n \to (0, \infty)$ be a decreasing function and $D_2 : I_n \to (0, \infty)$ an increasing function. Suppose that there exist $a, b, B, B' > 0$ such that*

$$D_1(Kc) \leq K^{-a} D_1(c), \qquad \textit{for any} \quad K > 1, \tag{2.39}$$

$$B' k^b D_2(c) \geq D_2(kc), \geq B k^b D_2(c) \qquad \textit{for any} \quad k < 1. \tag{2.40}$$

Let $\tilde{c}$ satisfy $D_1(\tilde{c}) = D_2(\tilde{c})$, and for a given constant $E \geq 1$, define $\Lambda = \big\{c : (D_1 + D_2)(c) \leq E\,(D_1 + D_2)(\tilde{c})\big\}$. Then

(i) $D_1(c) \leq B^{-1}(2E)^{1+b/a} D_2(c)$*, for every $c \in \Lambda$.*

(ii) $\Lambda \subset \big[(2E)^{-1/a}\tilde{c}, (2EB')^{1/b}\tilde{c}\big]$.

Proof. (i). If $c \geq \tilde{c}$, then $D_1(c) \leq D_2(c)$, since D_1 and D_2 are equal at $\tilde{c}$ and decreasing and increasing respectively. The inequality in (i) is then satisfied, since $B^{-1}(2E)^{1+b/a} \geq 1$. If $c < \tilde{c}$, then by (2.39) with $K = \tilde{c}/c$ we have

$$(\tilde{c}/c)^a D_1(\tilde{c}) \leq D_1(c).$$

If $c \in \Lambda$, then also

$$D_1(c) \leq (D_1 + D_2)(c) \leq E(D_1 + D_2)(\tilde{c}) = 2ED_1(\tilde{c})$$

by the definition of $\tilde{c}$. Concatenating these inequalities, we conclude that $(\tilde{c}/c)^a \leq 2E$, or $c \geq b_1\tilde{c}$ for $b_1 = (2E)^{-1/a} < 1$. Then, by monotonicity and (2.40),

$$D_2(c) \geq D_2(b_1\tilde{c}) \geq B b_1^b D_2(\tilde{c}).$$

This is equal to $B b_1^b D_1(\tilde{c}) \geq B b_1^b/(2E) D_1(c)$ by the second last display. This concludes the proof of (i).

(ii). The lower bound on Λ in (ii) is equivalent to the inequality $c \geq b_1\tilde{c}$, which was already obtained in the preceding proof of (i). For the upper bound we first note that for every $c \in \Lambda$ we have $D_2(c) \leq D_1(c) + D_2(c) \leq E(D_1 + D_2)(\tilde{c}) = 2ED_2(\tilde{c})$, by the definition of $\tilde{c}$. If $c > \tilde{c}$, then (2.40) gives that the right hand side is bounded above by $2EB'(\tilde{c}/c)^b D_2(c)$. Concatenation of the inequalities gives that $1 \leq 2EB'(\tilde{c}/c)^b$. □

The following lemma is applied throughout to handle the sums that occur in both the deterministic and stochastic terms of L.

Lemma 2.43. *Let $\gamma > -1$, $m \geq 1$ and $\nu \in \mathbb{R}$ such that $\gamma - m\nu < -1$. Then*

$$\sum_{j=1}^{n} \frac{j^\gamma}{(j^m + cn)^\nu} = C_{\gamma,\nu,m}(cn)^{\gamma/m-\nu+1/m}\big(1 + o(1)\big) \tag{2.41}$$

uniformly for $c \in [l_n/n, n^{m-1}/l_n]$ as $n \to \infty$, for any $l_n \to \infty$. The constant is given by

$$C_{\gamma,\nu,m} = \int_0^\infty \frac{u^\gamma}{(u^m + 1)^\nu}\, du.$$

Furthermore, the left side of (2.41) has the same order as the right side uniformly in $c \in [l_n/n, n^{m-1}]$, for any $l_n \to \infty$, possibly with a smaller constant.

Proof. If $\gamma \leq 0$, then the function $t \mapsto g(t) = t^\gamma/(t^m + cn)^\nu$ is decreasing on $[0, \infty)$, while if $\gamma > 0$ the function is unimodal with a maximum at $k(cn)^{1/m}$ for the constant $k = (\gamma/(m\nu - \gamma))^{1/m}$. In the first case we have

$$\int_1^n \frac{t^\gamma}{(t^m + cn)^\nu}\, \mathrm{d}t \leq \sum_{j=1}^{n} \frac{j^\gamma}{(j^m + cn)^\nu} \leq \int_0^n \frac{t^\gamma}{(t^m + cn)^\nu}\, \mathrm{d}t,$$

while in the second case

$$\begin{aligned}\int_1^n \frac{t^\gamma}{(t^m + cn)^\nu}\, \mathrm{d}t - g(k(cn)^{1/m}) &\leq \sum_{j=1}^{n} \frac{j^\gamma}{(j^m + cn)^\nu}\\ &\leq \int_0^n \frac{t^\gamma}{(t^m + cn)^\nu}\, \mathrm{d}t + g(k(cn)^{1/m}).\end{aligned}$$

By the change of coordinates $t^m = (cn)u^m$ we have

$$\int_a^n \frac{t^\gamma}{(t^m + cn)^\nu}\, \mathrm{d}t = (cn)^{\gamma/m-\nu+1/m} \int_{a/(cn)^{1/m}}^{n/(cn)^{1/m}} \frac{u^\gamma}{(u^m + 1)^\nu}\, \mathrm{d}u.$$

If $cn \to \infty$ with $(cn)^{1/m} \ll n$, then for both $a = 0$ and $a = 1$ the integral on the right approaches $C_{\gamma,\nu,m}$, which is finite under the conditions of the lemma. The maximum value in the second display satisfies $g(k(cn)^{1/m}) \lesssim (cn)^{(\gamma/m-\nu)}$ and hence is of lower order than the right side of the preceding display if $cn \to \infty$. This proves the first assertion of the lemma. For c as in the second assertion we still have that $cn \to \infty$, so that the lower limit of the integral tends to zero, but the upper limit $n/(cn)^{1/m}$ may remain bounded, although it is bigger than 1 by assumption. □

Lemma 2.44. *For $\gamma > -1$, $m \geq 1$ and $\nu \in \mathbb{R}$ such that $\gamma - m\nu < -1$ we have*

$$\sum_{i=1}^n \sum_{j=1}^n \frac{(ij)^\gamma}{((ij)^m + cn^2)^\nu} \asymp (cn^2)^{\gamma/m - \nu + 1/m} \cdot \begin{cases} \big(1 + \log(cn^2)\big) & \text{if } cn^2 \leq n^m \\ \big(1 + \log\big(\frac{n^{2m}}{cn^2}\big)\big) & \text{if } cn^2 \geq n^m \end{cases}$$

uniformly for $c \in [l_n/n^2, n^{2m-2}]$ as $n \to \infty$, for any $l_n \to \infty$.

Proof. Since $cn^2 \leq (ij)^m + cn^2 \leq 2cn^2$ if $(ij)^m \leq cn^2$ and $(ij)^m \leq (ij)^m + cn^2 \leq 2(ij)^m$ otherwise, the double sum is up to a constant 2^ν bounded above and below by

$$\sum_{\substack{i=1\\(ij)^m \leq cn^2}}^n \sum_{j=1}^n \frac{(ij)^\gamma}{(cn^2)^\nu} + \sum_{\substack{i=1\\(ij)^m > cn^2}}^n \sum_{j=1}^n (ij)^{\gamma - m\nu}.$$

Since $cn^2 \geq l_n \to \infty$, the first sum is never empty; the second is empty if $cn^2 = n^{2m}$ takes it maximally allowed value. To proceed we consider the cases that $N := (cn^2)^{1/m}$ is smaller or bigger than n separately. If $N \leq n$, then the second sum splits in two parts and the preceding display is equivalent to

$$\begin{aligned} &\sum_{i=1}^N \sum_{j=1}^{N/i} \frac{(ij)^\gamma}{N^{m\nu}} + \sum_{i=1}^N \sum_{j=N/i+1}^n (ij)^{\gamma - m\nu} + \sum_{i=N+1}^n \sum_{j=1}^n (ij)^{\gamma - m\nu} \\ &\asymp \sum_{i=1}^N \frac{i^\gamma (N/i)^{\gamma+1}}{N^{m\nu}} + \sum_{i=1}^N i^{\gamma - m\nu} (N/i)^{\gamma - m\nu + 1} + \sum_{i=N+1}^n i^{\gamma - m\nu} \\ &\asymp (\log N) N^{\gamma + 1 - m\nu} + (\log N) N^{\gamma - m\nu + 1} + N^{\gamma - m\nu + 1}. \end{aligned}$$

If $N > n$, then the first sum splits into two parts and we obtain the equivalent expression

$$\begin{aligned} &\sum_{i=1}^{N/n} \sum_{j=1}^n \frac{(ij)^\gamma}{N^{m\nu}} + \sum_{i=N/n+1}^n \sum_{j=1}^{N/i} \frac{(ij)^\gamma}{N^{m\nu}} + \sum_{i=N/n+1}^n \sum_{j=N/i+1}^n (ij)^{\gamma - m\nu} \\ &\asymp \sum_{i=1}^{N/n} \frac{i^\gamma n^{\gamma+1}}{N^{m\nu}} + \sum_{i=N/n+1}^n \frac{i^\gamma (N/i)^{\gamma+1}}{N^{m\nu}} + \sum_{i=N/n+1}^n i^{\gamma - m\nu} (N/i)^{\gamma - m\nu + 1} \\ &\asymp N^{\gamma - m\nu + 1} + (\log(n^2/N)) N^{\gamma - m\nu + 1} + (\log(n^2/N)) N^{\gamma + 1 - m\nu}. \end{aligned}$$

These bounds can be written in the form given by the lemma. □

Lemma 2.45. *For $m \geq 1$ and $\nu \in \mathbb{R}$ such that $-m\nu < -1$, we have*

$$\sum_{i=1}^n \sum_{j=1}^n \frac{1}{\big((i^2 + j^2)^m + cn^2\big)^\nu} \asymp (cn^2)^{-\nu + 1/m}$$

uniformly for $c \in [l_n/n^2, n^{2m-2}]$ as $n \to \infty$, for any $l_n \to \infty$.

Proof. Since the function $(s,t) \mapsto 1/\big((s^2+t^2)^m + cn^2\big)^\nu$ is decreasing in s and t, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\big((i^2+j^2)^m + cn^2\big)^\nu} \leq \int_0^n \int_0^n \frac{1}{\big((s^2+t^2)^m + cn^2\big)^\nu}\,\mathrm{d}s\,\mathrm{d}t$$

and

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\big((i^2+j^2)^m + cn^2\big)^\nu} \geq \int_1^n \int_1^n \frac{1}{\big((s^2+t^2)^m + cn^2\big)^\nu}\,\mathrm{d}s\,\mathrm{d}t.$$

Rewriting the double integrals in polar coordinates, we see that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\big((i^2+j^2)^m + cn^2\big)^\nu} \leq \frac{\pi}{2}\int_0^{\sqrt{2}n} \frac{r}{\big(r^{2m} + cn^2\big)^\nu}\,\mathrm{d}r$$

and

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\big((i^2+j^2)^m + cn^2\big)^\nu} \geq \frac{\pi}{2}\int_{\sqrt{2}}^{n} \frac{r}{\big(r^{2m} + cn^2\big)^\nu}\,\mathrm{d}r.$$

By the change of coordinates $r = \big(cn^2\big)^{\frac{1}{2m}}u$ we then have

$$\int_a^{bn} \frac{r}{\big(r^{2m} + cn^2\big)^\nu}\,\mathrm{d}r = \big(cn^2\big)^{-\nu+1/m} \int_{a/(cn^2)^{1/(2m)}}^{bn/(cn^2)^{1/(2m)}} \frac{u}{\big(u^{2m}+1\big)^\nu}\,\mathrm{d}u.$$

Since $cn^2 \to \infty$ the lower limit of this integral tends to zero. Combining this with the fact that the upper limit is bounded from below by b, the result follows. □

The following three lemmas are used to establish uniform bounds on the stochastic remainder terms.

Lemma 2.46. *Consider a function $g : (0,\infty) \to \mathbb{R}$ of the form*

$$g(c) = \frac{(c\lambda_{j,n})^\alpha}{(1+c\lambda_{j,n})^{\alpha+\beta}},$$

where $\alpha, \beta \geq 0$ are integers. Then, for $0 < s < t < \infty$,

$$|g(s) - g(t)| \leq \frac{|s-t|}{s} \frac{s\lambda_{j,n}}{(1+s\lambda_{j,n})^{2\vee(1+\beta)}}.$$

In particular, if $\beta \geq 2$, then $|g(s)-g(t)| \leq \frac{|s-t|}{s}\frac{1}{(1+s\lambda_{j,n})^2}$.

Proof. We apply the mean value theorem to the function $h(x) = \frac{x^\alpha}{(1+x)^{\alpha+\beta}}$. Note that for $x \geq 0$ we have

$$|h'(x)| = \left|\frac{x^{\alpha-1}\big(-\beta x+\alpha\big)}{(1+x)^{1+\alpha+\beta}}\right| \lesssim \frac{x^\alpha}{(1+x)^{1+\alpha+\beta}}1_{\beta\neq 0} + \frac{x^{\alpha-1}}{(1+x)^{1+\alpha+\beta}}$$
$$\leq \frac{1}{(1+x)^{1+\beta}}1_{\beta\neq 0} + \frac{1}{(1+x)^{2+\beta}} \lesssim \frac{1}{(1+x)^{2\vee(1+\beta)}}.$$

Hence

$$|g(s)-g(t)| \lesssim |s-t|\frac{\lambda_{j,n}}{(1+s\lambda_{j,n})^{2\vee(1+\beta)}} = \frac{|s-t|}{s}\frac{s\lambda_{j,n}}{(1+s\lambda_{j,n})^{2\vee(1+\beta)}}. \qquad \square$$

Lemma 2.47. *Consider the stochastic process $(U(c) : c > 0)$ given by*

$$U(c) = \sum_{j=1}^n \frac{a_j(c\lambda_{j,n})^\alpha}{(1+c\lambda_{j,n})^{\alpha+\beta}}U_j,$$

for some constants a_j, i.i.d. mean-zero random variables U_j with variance one and integers $\alpha, \beta \geq 0$. Suppose that for some $\gamma, \delta \in \{0,1,2\}$ and some non-negative function g we have

$$\sum_{j=1}^n \frac{a_j^2(s\lambda_{j,n})^\delta}{(1+s\lambda_{j,n})^\gamma} \lesssim g(s). \tag{2.42}$$

Then for $0 < s < t < \infty$ we have

$$\mathrm{var}\big(U(s)-U(t)\big) \lesssim \frac{(s-t)^2 g(s)}{s^2}.$$

Proof. We consider

$$\mathrm{var}\big[U(s)-U(t)\big] = \sum_{j=1}^n a_j^2\left[\frac{(s\lambda_{j,n})^\alpha}{(1+s\lambda_{j,n})^{\alpha+\beta}} - \frac{(t\lambda_{j,n})^\alpha}{(1+t\lambda_{j,n})^{\alpha+\beta}}\right]^2.$$

Applying the previous lemma, we see that

$$\left|\frac{(s\lambda_{j,n})^\alpha}{(1+s\lambda_{j,n})^{\alpha+\beta}} - \frac{(t\lambda_{j,n})^\alpha}{(1+t\lambda_{j,n})^{\alpha+\beta}}\right| \lesssim \frac{|s-t|}{s}\frac{s\lambda_{j,n}}{(1+s\lambda_{j,n})^2}.$$

We conclude

$$\mathrm{var}\big[U(s)-U(t)\big] \lesssim \frac{(s-t)^2}{s^2}\sum_{j=1}^n \frac{a_j^2(s\lambda_{j,n})^2}{(1+s\lambda_{j,n})^4}$$
$$\leq \frac{(s-t)^2}{s^2}\sum_{j=1}^n a_j^2\frac{(s\lambda_{j,n})^\delta}{(1+s\lambda_{j,n})^\gamma},$$

which holds for any $\gamma, \delta \in \{0,1,2\}$. The result follows. □

Lemma 2.48. *Let $l_n \to \infty$ be a given sequence of numbers. If $U_n = (U_n(s) : s \in I_n)$ are continuous stochastic processes such that for all $s < t$ in a closed interval $I_n \subset [l_n/n, \infty)$ and some $a > 0$ we have*

$$E\big[U_n(s)\big]^2 \lesssim \frac{1}{n^a s^a}, \qquad E\big[U_n(s) - U_n(t)\big]^2 \lesssim \frac{(t-s)^2}{n^a s^{2+a}},$$

then $\sup_{s\in I_n} |U_n(s)|$ tends to zero in probability.

Proof. Write $I_n = [a_n, b_n]$. For a given interval $[s_0, t_0] \subset I_n$ we have $\mathrm{E}\big[U_n(s) - U_n(t)\big]^2 \lesssim d_0^2(s,t)$, for d_0 the metric

$$d_0(s,t) = K_0|t-s|, \qquad K_0 = n^{-a/2} s_0^{-1-a/2}.$$

The d_0-diameter of $[s_0, t_0]$ is $K_0|t_0 - s_0|$ and the covering number $N(u, [s_0,t_0], d_0)$ is bounded above by $\big(K_0|t_0 - s_0|/u\big) \vee 1$. Therefore by Corollary 2.2.5 in [van der Vaart and Wellner, 1996], with $\psi(x) = x^2$, we have

$$\mathrm{E} \sup_{s,t\in[s_0,t_0]} \big[U_n(s) - U_n(t)\big]^2 \lesssim K_0^2 |t_0 - s_0|^2 = \frac{|t_0/s_0 - 1|^2}{(ns_0)^a}.$$

Fix M so that $2^{M-1} < 1/a_n \le 2^M$ and N so that $2^{N-1} < b_n \le 2^N$. Define $s_{-M} = a_n$, $s_N = b_n$ and $s_i = 2^i$ for $i \in \{-M+1, \dots, N-1\}$. Then $s_{-M} < s_{-M+1} < \cdots < s_N$ partitions I_n. Since $s_{i+1}/s_i - 1 \le 1$ for every i (in fact, equal to 1 except for the boundary values), we then have

$$\begin{aligned}
\mathrm{E} \sup_{s\in I_n} U_n(s)^2 &\le 2\mathrm{E} \max_{i\in\{-M,\dots,N-1\}} \Big[\sup_{s\in[s_i,s_{i+1}]} |U_n(s) - U_n(s_i)|^2 + U_n(s_i)^2\Big] \\
&\lesssim \sum_{i=-M}^{N-1} \Big[\frac{1^2}{(ns_i)^a} + \frac{1}{(ns_i)^a}\Big] \\
&\lesssim \frac{1}{n^a} \sum_{i=-M}^{N-1} 2^{-ia} = \frac{1}{n^a} 2^{Ma} \frac{1 - 2^{-a(M+N)}}{1 - 2^{-a}} \\
&\le \frac{1}{n^a}\Big(\frac{2}{a_n}\Big)^a \frac{1}{1 - 2^{-a}} \le \frac{1}{l_n^a} \frac{2^a}{1 - 2^{-a}},
\end{aligned}$$

by definition of M. This tends to zero, since $l_n \to \infty$. □