BAYESIAN INVERSE PROBLEMS WITH GAUSSIAN PRIORS

BY B. T. KNAPIK[1], A. W. VAN DER VAART AND J. H. VAN ZANTEN

*VU University Amsterdam, VU University Amsterdam and
Eindhoven University of Technology*

The posterior distribution in a nonparametric inverse problem is shown to contract to the true parameter at a rate that depends on the smoothness of the parameter, and the smoothness and scale of the prior. Correct combinations of these characteristics lead to the minimax rate. The frequentist coverage of credible sets is shown to depend on the combination of prior and true parameter, with smoother priors leading to zero coverage and rougher priors to conservative coverage. In the latter case credible sets are of the correct order of magnitude. The results are numerically illustrated by the problem of recovering a function from observation of a noisy version of its primitive.

1. Introduction. In this paper we study a Bayesian approach to estimating a parameter μ from an observation Y following the model

$$(1.1) \qquad Y = K\mu + \frac{1}{\sqrt{n}}Z.$$

The unknown parameter μ is an element of a separable Hilbert space H_1, and is mapped into another Hilbert space H_2 by a known, injective, continuous linear operator $K : H_1 \to H_2$. The image $K\mu$ is perturbed by unobserved, scaled Gaussian white noise Z. There are many special examples of this infinite-dimensional regression model, which can also be viewed as an idealized version of other statistical models, including density estimation. The inverse problem of estimating μ has been studied by both statisticians and numerical mathematicians (see, e.g., [3, 6, 8, 24, 26, 33] for reviews), but rarely from a theoretical Bayesian perspective; exceptions are [7] and [11].

The Bayesian approach to (1.1) consists of putting a prior on the parameter μ, and computing the posterior distribution. We study Gaussian priors, which are conjugate to the model, so that the posterior distribution is also Gaussian and easy to derive. Our interest is in studying the properties of this posterior distribution, under the "frequentist" assumption that the data Y has been generated according to the model (1.1) with a given "true" parameter μ_0. We investigate whether and at what rate the posterior distributions contract to μ_0 as $n \to \infty$ (as in [15]), but have

as main interest the performance of credible sets for measuring the uncertainty about the parameter.

A Bayesian *credible set* is defined as a central region in the posterior distribution of specified posterior probability, for instance, 95%. As a consequence of the Bernstein–von Mises theorem credible sets for smooth *finite-dimensional* parametric models are asymptotically equivalent to confidence regions based on the maximum likelihood estimator (see, e.g., [31], Chapter 10), under mild conditions on the prior. Thus, "Bayesian uncertainty" is equivalent to "frequentist uncertainty" in these cases, at least for large n. However, there is no corresponding Bernstein–von Mises theorem in nonparametric Bayesian inference, as noted in [12]. The performance of Bayesian credible sets in these situations has received little attention, although in practice such sets are typically provided as indicators of uncertainty, for instance, based on the spread of the output of a (converged) MCMC run. The paper [7] did tackle this issue and came to the alarming conclusion that Bayesian credible sets have frequentist coverage zero. If this were true, many data analysts would (justifiably) distrust the spread in the posterior distribution as a measure of uncertainty. For other results see [4, 13, 14] and [18].

The model considered in [7] is equivalent to our model (1.1), and a good starting point for studying these issues. More precisely, the conclusion of [7] is that *for almost every parameter* μ_0 *from the prior the coverage of a credible set* (*of any level*) *is* 0. In the present paper we show that this is only part of the story, and, taken by itself, the conclusion is misleading. The coverage depends on the true parameter μ_0 and the prior together, and it can be understood in terms of a bias-variance trade-off, much as the coverage of frequentist nonparametric procedures. A nonparametric procedure that oversmoothes the truth (too big a bandwidth in a frequentist procedure, or a prior that puts too much weight on "smooth" parameters) will be biased, and a confidence or credible region based on such a procedure will be both too concentrated and wrongly located, giving zero coverage. On the other hand, undersmoothing does work (to a certain extent), also in the Bayesian setup, as we show below. In this light we reinterpret the conclusion of [7] to be valid only in the oversmoothed case (notwithstanding a conjecture to the contrary in the Introduction of this paper; see page 905, answer to objection 4). In the undersmoothed case credible regions are conservative in general, with coverage tending to 1. The good news is that typically they are of the correct order of magnitude, so that they do give a reasonable idea of the uncertainty in the estimate.

Of course, whether a prior under- or oversmoothes depends on the regularity of the true parameter. In practice, we may not want to consider this known, and adapt the prior smoothness to the data. In this paper we do consider the effect of changing the "length scale" of a prior, but do not study data-dependent length scales. The effect of setting the latter by, for example, an empirical or full Bayes method will require further study.

Credible sets are by definition "central regions" in the posterior distribution. Because the posterior distribution is a random probability measure on the Hilbert

space H_1, a "central ball" is a natural shape of such a set, but it has the disadvantage that it is difficult to visualize. If the Hilbert space is a function space, then *credible bands* are more natural. These correspond to simultaneous credible intervals for the function at a point, and can be obtained from the (marginal) posterior distributions of a set of linear functionals. Besides the full posterior distribution, we therefore study its marginals for linear functionals. The same issue of the dependence of coverage on under- and oversmoothing arises, except that "very smooth" linear functionals cancel the inverse nature of the problem, and do allow a Bernstein–von Mises theorem for a large set of priors. Unfortunately point evaluations are usually not smooth in this sense.

Thus, we study two aspects of inverse problems—recovering the full parameter μ (Section 4) and recovering linear functionals (Section 5). We obtain the rate of contraction of the posterior distribution in both settings, in its dependence on parameters of the prior. Furthermore, and most importantly, we study the "frequentist" coverage of credible regions for μ in both settings, for the same set of priors. In the next section we give a more precise statement of the problem, and in Section 3 we describe the priors that we consider and derive the corresponding posterior distributions. In Section 6 we illustrate the results by simulations and pictures in the particular example that K is the Volterra operator. Technical proofs are placed in Sections 7 and 8 at the end of the paper.

Throughout the paper $\langle \cdot, \cdot \rangle_1$ and $\| \cdot \|_1$, and $\langle \cdot, \cdot \rangle_2$ and $\| \cdot \|_2$ denote the inner products and norms of the Hilbert spaces H_1 and H_2. The adjoint of an operator A between two Hilbert spaces is denoted by A^T. The Sobolev space S^β with its norm $\| \cdot \|_\beta$ is defined in (2.2). For two sequences (a_n) and (b_n) of numbers $a_n \asymp b_n$ means that $|a_n/b_n|$ is bounded away from zero and infinity as $n \to \infty$, and $a_n \lesssim b_n$ means that a_n/b_n is bounded.

2. Detailed description of the problem. The noise process Z in (1.1) is the standard normal or *iso-Gaussian process* for the Hilbert space H_2. Because this is not realizable as a random element in H_2, the model (1.1) is interpreted in process form (as in [3]). The iso-Gaussian process is the zero-mean Gaussian process $Z = (Z_h : h \in H_2)$ with covariance function $\mathrm{E} Z_h Z_{h'} = \langle h, h' \rangle_2$, and the measurement equation (1.1) is interpreted in that we observe a Gaussian process $Y = (Y_h : h \in H_2)$ with mean and covariance functions

$$(2.1) \qquad \mathrm{E} Y_h = \langle K\mu, h \rangle_2, \qquad \mathrm{cov}(Y_h, Y_{h'}) = \frac{1}{n} \langle h, h' \rangle_2.$$

Sufficiency considerations show that it is statistically equivalent to observe the subprocess $(Y_{h_i} : i \in \mathbb{N})$, for any orthonormal basis $h_1, h_2, \ldots$ of H_2.

If the operator K is compact, then the *spectral decomposition* of the self-adjoint operator $K^T K : H_1 \to H_1$ provides a convenient basis. In the compact case the operator $K^T K$ possesses countably many positive eigenvalues κ_i^2 and there is a corresponding orthonormal basis (e_i) of H_1 of eigenfunctions (hence, $K^T K e_i = \kappa_i^2 e_i$

for $i \in \mathbb{N}$; see, e.g., [23]). The sequence (f_i) defined by $Ke_i = \kappa_i f_i$ forms an orthonormal "conjugate" basis of the range of K in H_2. An element $\mu \in H_1$ can be identified with its sequence (μ_i) of coordinates relative to the eigenbasis (e_i), and its image $K\mu = \sum_i \mu_i Ke_i = \sum_i \mu_i \kappa_i f_i$ can be identified with its coordinates $(\mu_i \kappa_i)$ relative to the conjugate basis (f_i). If we write Y_i for Y_{f_i}, then (2.1) shows that $Y_1, Y_2, \ldots$ are independent Gaussian variables with means $EY_i = \mu_i \kappa_i$ and variance $1/n$. Therefore, a concrete equivalent description of the statistical problem is to *recover the sequence (μ_i) from independent observations $Y_1, Y_2, \ldots$ with $N(\mu_i \kappa_i, 1/n)$-distributions*.

In the following we do not require K to be compact, but we do assume the existence of an orthonormal basis of eigenfunctions of $K^T K$. The main additional example we then cover is the *white noise model*, in which K is the identity operator. The description of the problem remains the same.

If $\kappa_i \to 0$, this problem is ill-posed, and the recovery of μ from Y an *inverse problem*. The ill-posedness can be quantified by the speed of decay $\kappa_i \downarrow 0$. Although the tools are more widely applicable, we limit ourselves to the *mildly ill-posed* problem (in the terminology of [6]) and assume that the decay is polynomial: for some $p \geq 0$,

$$\kappa_i \asymp i^{-p}.$$

Estimation of μ is harder if the decay is faster (i.e., p is larger).

The difficulty of estimation may be measured by the minimax risks over the scale of *Sobolev spaces* relative to the orthonormal basis (e_i) of eigenfunctions of $K^T K$. For $\beta > 0$ define

$$(2.2) \qquad \|\mu\|_\beta = \sqrt{\sum_{i=1}^{\infty} \mu_i^2 i^{2\beta}} \qquad \text{if } \mu = \sum_{i=1}^{\infty} \mu_i e_i.$$

Then the Sobolev space of order β is $S^\beta = \{\mu \in H_1 : \|\mu\|_\beta < \infty\}$. The minimax rate of estimation over the unit ball of this space relative to the loss $\|t - \mu\|_1$ of an estimate t for μ is $n^{-\beta/(1+2\beta+2p)}$. This rate is attained by various "regularization" methods, such as generalized *Tikhonov* and *Moore–Penrose* regularization [1, 3, 6, 16, 19]. The Bayesian approach is closely connected to these methods: in Section 3 the posterior mean is shown to be a regularized estimator.

Besides recovery of the full parameter μ, we consider estimating linear functionals $L\mu$. The minimax rate for such functionals over Sobolev balls depends on L as well as on the parameter of the Sobolev space. Decay of the coefficients of L in the eigenbasis may alleviate the level of ill-posedness, with rapid decay even bringing the functional in the domain of "regular" $n^{-1/2}$-rate estimation.

3. Prior and posterior distributions. We assume a mean-zero Gaussian prior for the parameter μ. In the next three paragraphs we recall some essential facts on Gaussian distributions on Hilbert spaces.

A *Gaussian distribution* $N(v, \Lambda)$ on the Borel sets of the Hilbert space H_1 is characterized by a *mean* v, which can be any element of H_1, and a *covariance operator* $\Lambda : H_1 \rightarrow H_1$, which is a nonnegative-definite, self-adjoint, linear operator *of trace class*: a compact operator with eigenvalues (λ_i) that are summable $\sum_{i=1}^{\infty} \lambda_i < \infty$ (see, e.g., [25], pages 18–20). A random element G in H_1 is $N(v, \Lambda)$-distributed if and only if the stochastic process $(\langle G, h \rangle_1 : h \in H_1)$ is a Gaussian process with mean and covariance functions

$$(3.1) \qquad \mathrm{E}\langle G, h \rangle_1 = \langle v, h \rangle_1, \qquad \mathrm{cov}(\langle G, h \rangle_1, \langle G, h' \rangle_1) = \langle h, \Lambda h' \rangle_1.$$

The coefficients $G_i = \langle G, \varphi_i \rangle_1$ of G relative to an orthonormal eigenbasis (φ_i) of Λ are independent, univariate Gaussians with means the coordinates (v_i) of the mean vector v and variances the eigenvalues λ_i.

The iso-Gaussian process Z in (1.1) may be thought of as a $N(0, I)$-distributed Gaussian element, for I the identity operator (on H_2), but as I is not of trace class, this distribution is not realizable as a proper random element in H_2. Similarly, the data Y in (1.1) can be described as having a $N(K\mu, n^{-1}I)$-distribution.

For a stochastic process $W = (W_h : h \in H_2)$ and a continuous, linear operator $A : H_2 \rightarrow H_1$, we define the transformation AW as the stochastic process with coordinates $(AW)_h = W_{A^T h}$, for $h \in H_1$. If the process W arises as $W_h = \langle W, h \rangle_2$ from a random element W in the Hilbert space H_2, then this definition is consistent with identifying the random element AW in H_1 with the process $(\langle AW, h \rangle_1 : h \in H_1)$, as in (3.1) with $G = AW$. Furthermore, if A is a *Hilbert–Schmidt* operator (i.e., AA^T is of trace class), and $W = Z$ is the iso-Gaussian process, then the process AW can be realized as a random variable in H_1 with a $N(0, AA^T)$-distribution.

In the Bayesian setup the prior, which we take $N(0, \Lambda)$, is the marginal distribution of μ, and the noise Z in (1.1) is considered independent of μ. The joint distribution of (Y, μ) is then also Gaussian, and so is the conditional distribution of μ given Y, the posterior distribution of μ. In general, one must be a bit careful with manipulating possibly "improper" Gaussian distributions (see [20]), but in our situation the posterior is a proper Gaussian conditional distribution on H_1.

PROPOSITION 3.1 (Full posterior). *If μ is $N(0, \Lambda)$-distributed and Y given μ is $N(K\mu, n^{-1}I)$-distributed, then the conditional distribution of μ given Y is Gaussian $N(AY, S_n)$ on H_1, where*

$$(3.2) \qquad S_n = \Lambda - A(n^{-1}I + K\Lambda K^T)A^T,$$

and $A : H_2 \rightarrow H_2$ is the continuous linear operator

$$(3.3) \quad A = \Lambda^{1/2}\left(\frac{1}{n}I + \Lambda^{1/2}K^T K \Lambda^{1/2}\right)^{-1}\Lambda^{1/2}K^T = \Lambda K^T\left(\frac{1}{n}I + K\Lambda K^T\right)^{-1}.$$

The posterior distribution is proper (i.e., S_n has finite trace) and equivalent (in the sense of absolute continuity) to the prior.

PROOF. Identity (3.3) is a special case of the identity $(I + BB^T)^{-1}B = B(I + B^T B)^{-1}$, which is valid for any compact, linear operator $B: H_1 \to H_2$. That S_n is of trace class is a consequence of the fact that it is bounded above by Λ (i.e., $\Lambda - S_n$ is nonnegative definite), which is of trace class by assumption.

The operator $\Lambda^{1/2} K^T K \Lambda^{1/2}: H_1 \to H_1$ has trace bounded by $\|K^T K\| \operatorname{tr}(\Lambda)$ and hence is of trace class. It follows that the variable $\Lambda^{1/2} K^T Z$ can be defined as a random element in the Hilbert space H_1, and so can AY, for A given by the first expression in (3.3). The joint distribution of (Y, μ) is Gaussian with zero mean and covariance operator

$$\begin{pmatrix} n^{-1}I + K\Lambda K^T & K\Lambda \\ \Lambda K^T & \Lambda \end{pmatrix}.$$

Using this with the second form of A in (3.3), we can check that the cross covariance operator of the variables $\mu - AY$ and Y (the latter viewed as a Gaussian stochastic process in $\mathbb{R}^{H_2}$) vanishes and, hence, these variables are independent. Thus, the two terms in the decomposition $\mu = (\mu - AY) + AY$ are conditionally independent and degenerate given Y, respectively. The distribution of $\mu - AY$ is zero-mean Gaussian with covariance operator $\operatorname{Cov}(\mu - AY) = \operatorname{Cov}(\mu) - \operatorname{Cov}(AY)$, by the independence of $\mu - AY$ and AY. This gives the form of the posterior distribution.

The final assertion may be proved by explicitly comparing the Gaussian prior and posterior. Easier is to note that it suffices to show that the model consisting of all $N(K\mu, n^{-1}I)$-distributions is dominated. In that case the posterior can be obtained using Bayes' rule, which reveals the normalized likelihood as a density relative to the (in fact, *any*) prior. To prove domination, we may consider equivalently the distributions $\bigotimes_{i=1}^{\infty} N(\kappa_i \mu_i, n^{-1})$ on $\mathbb{R}^{\infty}$ of the sufficient statistic (Y_i) defined as the coordinates of Y relative to the conjugate spectral basis. These distributions, for $(\mu_i) \in \ell_2$, are equivalent to the distribution $\bigotimes_{i=1}^{\infty} N(0, n^{-1})$, as can be seen with the help of Kakutani's theorem, the affinity being $\exp(-\sum_i \kappa_i^2 \mu_i^2 / 8) > 0$. (This argument actually proves the well-known fact that the Gaussian shift experiment obtained by translating the standard normal distribution on $\mathbb{R}^{\infty}$ over its RKHS ℓ_2 is dominated.) □

In the remainder of the paper we study the asymptotic behavior of the posterior distribution, under the assumption that $Y = K\mu_0 + n^{-1/2}Z$ for a fixed $\mu_0 \in H_1$. The posterior is characterized by its *center* AY, the posterior mean, and its *spread*, the posterior covariance operator S_n. The first depends on the data, but the second is deterministic. From a frequentist-Bayes perspective both are important: one would like the posterior mean to give a good estimate for μ_0, and the spread to give a good indication of the uncertainty in this estimate.

The posterior mean is a regularization, of the Tikhonov type, of the naive estimator $K^{-1}Y$. It can also be characterized as a penalized least squares estimator

(see [21, 27]): it minimizes the functional

$$\mu \mapsto \|Y - K\mu\|_2^2 + \frac{1}{n}\|\Lambda^{-1/2}\mu\|_1^2.$$

The penalty $\|\Lambda^{-1/2}\mu\|_1$ is interpreted as ∞ if μ is not in the range of $\Lambda^{1/2}$. Because this range is precisely the *reproducing kernel Hilbert space* (RKHS) of the prior (cf. [32]), with $\|\Lambda^{-1/2}\mu\|_1$ as the RKHS-norm of μ, the posterior mean also fits into the general regularization framework using RKHS-norms (see [22]). In any case the posterior mean is a well-studied point estimator in the literature on inverse problems. In this paper we add a Bayesian interpretation to it, and are (more) concerned with the full posterior distribution.

Next consider the posterior distribution of a linear functional $L\mu$ of the parameter. We are not only interested in continuous, linear functionals $L\mu = \langle \mu, l \rangle_1$, for some given $l \in H_1$, but also in certain discontinuous functionals, such as point evaluation in a Hilbert space of functions. The latter entail some technicalities. We consider *measurable linear functionals relative to the prior* $N(0, \Lambda)$, defined in [25], pages 27–29, as Borel measurable maps $L : H_1 \to \mathbb{R}$ that are linear on a measurable linear subspace $\underline{H}_1 \subset H_1$ such that $N(0, \Lambda)(\underline{H}_1) = 1$. This definition is exactly right to make the marginal posterior Gaussian.

PROPOSITION 3.2 (Marginal posterior). *If μ is $N(0, \Lambda)$-distributed and Y given μ is $N(K\mu, n^{-1}I)$-distributed, then the conditional distribution of $L\mu$ given Y for a $N(0, \Lambda)$-measurable linear functional $L : H_1 \to \mathbb{R}$ is a Gaussian distribution $N(LAY, s_n^2)$ on $\mathbb{R}$, where*

$$(3.4) \qquad s_n^2 = (L\Lambda^{1/2})(L\Lambda^{1/2})^T - LA(n^{-1}I + K\Lambda K^T)(LA)^T,$$

and $A : H_2 \to H_2$ is the continuous linear operator defined in (3.3).

PROOF.   As in the proof of Proposition 3.1, the first term in the decomposition $L\mu = L(\mu - AY) + LAY$ is independent of Y. Therefore, the posterior distribution is the marginal distribution of $L(\mu - AY)$ shifted by LAY. It suffices to show that this marginal distribution is $N(0, s_n^2)$.

By Theorem 1 on page 28 in [25], there exists a sequence of continuous linear maps $L_m : H_1 \to \mathbb{R}$ such that $L_m h \to Lh$ for all h in a set with probability one under the prior $\Pi = N(0, \Lambda)$. This implies that $L_m \Lambda^{1/2} h \to L\Lambda^{1/2} h$ for *every* $h \in H_1$. Indeed, if $V = \{h \in H_1 : L_m h \to Lh\}$ and $g \notin V$, then $V_1 := V + g$ and V are disjoint measurable, affine subspaces of H_1, where $\Pi(V) = 1$. The range of $\Lambda^{1/2}$ is the RKHS of Π and, hence, if g is in this range, then $\Pi(V_1) > 0$, as Π shifted over an element from its RKHS is equivalent to Π. But then V and V_1 are not disjoint.

Therefore, from the first definition of A in (3.3) we see that $L_m A \to LA$, and, hence, $L_m(\mu - AY) \to L(\mu - AY)$, almost surely. As L_m is continuous,

the variable $L_m(\mu - AY)$ is normally distributed with mean zero and variance $L_m S_m L_m^T = (L_m \Lambda^{1/2})(L_m \Lambda^{1/2})^T - L_m A(n^{-1}I + K \Lambda K^T)(L_m A)^T$, for S_n given by (3.2). The desired result follows upon taking the limit as $m \to \infty$. $\quad\square$

As shown in the preceding proof, $N(0, \Lambda)$-measurable linear functionals L automatically have the further property that $L\Lambda^{1/2} : H_1 \to \mathbb{R}$ is a continuous linear map. This shows that LA and the adjoint operators $(L\Lambda^{1/2})^T$ and $(LA)^T$ are well defined, so that the formula for s_n^2 makes sense. If L is a *continuous* linear operator, one can also write these adjoints in terms of the adjoint L^T of L, and express s_n^2 in the covariance operator S_n of Proposition 3.1 as $s_n^2 = L S_n L^T$. This is exactly as expected.

In the remainder of the paper we study the full posterior distribution $N(AY, S_n)$, and its marginals $N(LAY, s_n^2)$. We are particularly interested in the influence of the prior on the performance of the posterior distribution for various true parameters μ_0. We study this in the following setting.

ASSUMPTION 3.1.   The operators $K^T K$ and Λ have the same eigenfunctions (e_i), with eigenvalues (κ_i^2) and (λ_i), satisfying

$$(3.5) \qquad \lambda_i = \tau_n^2 i^{-1-2\alpha}, \qquad C^{-1} i^{-p} \le \kappa_i \le C i^{-p}$$

for some $\alpha > 0$, $p \ge 0$, $C \ge 1$ and $\tau_n > 0$ such that $n\tau_n^2 \to \infty$. Furthermore, the true parameter μ_0 belongs to S^β for some $\beta > 0$: that is, its coordinates $(\mu_{0,i})$ relative to (e_i) satisfy $\sum_{i=1}^\infty \mu_{0,i}^2 i^{2\beta} < \infty$.

The setting of Assumption 3.1 is a Bayesian extension of the *mildly ill-posed inverse problem* (cf. [6]). We refer to the parameter β as the "regularity" of the true parameter μ_0. In the special case that H_1 is a function space and (e_i) its Fourier basis, this parameter gives smoothness of μ_0 in the classical Sobolev sense. Because the coefficients (μ_i) of the prior parameter μ are normally $N(0, \lambda_i)$-distributed, under Assumption 3.1 we have $E \sum_i i^{2\alpha'} \mu_i^2 = \tau_n^2 \sum_i i^{2\alpha'} \lambda_i < \infty$ if and only if $\alpha' < \alpha$. Thus, α is "almost" the smoothness of the parameters generated by the prior. This smoothness is modified by the *scaling factor* τ_n. Although this leaves the relative sizes of the coefficients μ_i, and hence the qualitative smoothness of the prior, invariant, we shall see that scaling can completely alter the performance of the Bayesian procedure. Rates $\tau_n \downarrow 0$ increase, and rates $\tau_n \uparrow \infty$ decrease the regularity.

4. Recovering the full parameter.   We denote by $\Pi_n(\cdot|Y)$ the posterior distribution $N(AY, S_n)$, derived in Proposition 3.1. Our first theorem shows that it contracts as $n \to \infty$ to the true parameter at a rate ε_n that depends on all four parameters α, β, τ_n, p of the (Bayesian) inverse problem.

THEOREM 4.1 (Contraction).   *If μ_0, (λ_i), (κ_i) and (τ_n) are as in Assumption* 3.1, *then* $E_{\mu_0}\Pi_n(\mu : \|\mu - \mu_0\|_1 \geq M_n\varepsilon_n|Y) \to 0$, *for every* $M_n \to \infty$, *where*

$$(4.1) \qquad \varepsilon_n = (n\tau_n^2)^{-\beta/(1+2\alpha+2p)\wedge 1} + \tau_n(n\tau_n^2)^{-\alpha/(1+2\alpha+2p)}.$$

The rate is uniform over μ_0 in balls in S^β. In particular:

  (i) *If $\tau_n \equiv 1$, then $\varepsilon_n = n^{-(\alpha\wedge\beta)/(1+2\alpha+2p)}$.*
  (ii) *If $\beta \leq 1 + 2\alpha + 2p$ and $\tau_n \asymp n^{(\alpha-\beta)/(1+2\beta+2p)}$, then $\varepsilon_n = n^{-\beta/(1+2\beta+2p)}$.*
  (iii) *If $\beta > 1 + 2\alpha + 2p$, then $\varepsilon_n \gg n^{-\beta/(1+2\beta+2p)}$, for every scaling τ_n.*

The minimax rate of convergence over a Sobolev ball S^β is of the order $n^{-\beta/(1+2\beta+2p)}$ (see [6]). By (i) of the theorem the posterior contraction rate is the same if the regularity of the prior is chosen to match the regularity of the truth ($\alpha = \beta$) and the scale τ_n is fixed. Alternatively, the optimal rate is also attained by appropriately scaling ($\tau_n \asymp n^{(\alpha-\beta)/(1+2\beta+2p)}$, determined by balancing the two terms in ε_n) a prior that is regular enough ($\beta \leq 1 + 2\alpha + 2p$). In all other cases (no scaling and $\alpha \neq \beta$, or any scaling combined with a rough prior $\beta > 1 + 2\alpha + 2p$), the contraction rate is slower than the minimax rate.

That "correct" specification of the prior gives the optimal rate is comforting to the true Bayesian. Perhaps the main message of the theorem is that even if the prior mismatches the truth, it may be scalable to give the optimal rate. Here, similar as found by [29] in a different setting, a smooth prior can be scaled to make it "rougher" to any degree, but a rough prior can be "smoothed" relatively little (namely, from α to any $\beta \leq 1 + 2\alpha + 2p$). It will be of interest to investigate a full or empirical Bayesian approach to set the scaling parameter.

Bayesian inference takes the spread in the posterior distribution as an expression of uncertainty. This practice is not validated by (fast) contraction of the posterior. Instead we consider the frequentist coverage of credible sets. As the posterior distribution is Gaussian, it is natural to center a credible region at the posterior mean. Different shapes of such a set could be considered. The natural counterpart of the preceding theorem is to consider balls. In the next section we also consider bands. (Alternatively, one might consider ellipsoids, depending on geometry of the support of the posterior.)

Because the posterior spread S_n is deterministic, the radius is the only degree of freedom when we choose a ball, and we fix it by the desired "credibility level" $1 - \gamma \in (0, 1)$. A *credible ball* centered at the posterior mean AY takes the form, where $B(r)$ denotes a ball of radius r around 0,

$$(4.2) \qquad AY + B(r_{n,\gamma}) := \{\mu \in H_1 : \|\mu - AY\|_1 < r_{n,\gamma}\},$$

where the radius $r_{n,\gamma}$ is determined so that

$$(4.3) \qquad \Pi_n(AY + B(r_{n,\gamma})|Y) = 1 - \gamma.$$

Because the posterior spread S_n is not dependent on the data, neither is the radius $r_{n,\gamma}$. The frequentist *coverage* or confidence of the set (4.2) is

$$(4.4) \qquad \qquad \mathrm{P}_{\mu_0}\big(\mu_0 \in AY + B(r_{n,\gamma})\big),$$

where under the probability measure P_{μ_0} the variable Y follows (1.1) with $\mu = \mu_0$. We shall consider the coverage as $n \to \infty$ for fixed μ_0, uniformly in Sobolev balls, and also along sequences μ_0^n that change with n.

The following theorem shows that the relation of the coverage to the credibility level $1 - \gamma$ is mediated by all parameters of the problem. For further insight, the credible region is also compared to the "correct" frequentist confidence ball $AY + B(\tilde{r}_{n,\gamma})$, which has radius $\tilde{r}_{n,\gamma}$ chosen so that the probability in (4.4) with $r_{n,\gamma}$ replaced by $\tilde{r}_{n,\gamma}$ is equal to $1 - \gamma$.

THEOREM 4.2 (Credibility). *Let μ_0, (λ_i), (κ_i), and τ_n be as in Assumption* 3.1, *and set $\tilde{\beta} = \beta \wedge (1 + 2\alpha + 2p)$. The asymptotic coverage of the credible region* (4.2) *is:*

(i) 1, *uniformly in μ_0 with $\|\mu_0\|_\beta \leq 1$, if $\tau_n \gg n^{(\alpha - \tilde{\beta})/(1 + 2\tilde{\beta} + 2p)}$; in this case* $\tilde{r}_{n,\gamma} \asymp r_{n,\gamma}$.

(ii) 1, *for every fixed $\mu_0 \in S^\beta$, if $\beta < 1 + 2\alpha + 2p$ and $\tau_n \asymp n^{(\alpha - \tilde{\beta})/(1 + 2\tilde{\beta} + 2p)}$; c, along some μ_0^n with $\sup_n \|\mu_0^n\|_\beta < \infty$, if $\tau_n \asymp n^{(\alpha - \tilde{\beta})/(1 + 2\tilde{\beta} + 2p)}$ (any $c \in [0, 1)$).*

(iii) 0, *along some μ_0^n with $\sup_n \|\mu_0^n\|_\beta < \infty$, if $\tau_n \ll n^{(\alpha - \tilde{\beta})/(1 + 2\tilde{\beta} + 2p)}$.*

If $\tau_n \equiv 1$, then the cases (i), (ii) *and* (iii) *arise if $\alpha < \beta$, $\alpha = \beta$ and $\alpha > \beta$, respectively. In case* (iii) *the sequence μ_0^n can then be chosen fixed.*

The theorem is easiest to interpret in the situation without scaling ($\tau_n \equiv 1$). Then oversmoothing the prior [case (iii): $\alpha > \beta$] has disastrous consequences for the coverage of the credible sets, whereas undersmoothing [case (i): $\alpha < \beta$] leads to conservative confidence sets. Choosing a prior of correct regularity [case (ii): $\alpha = \beta$] gives mixed results.

Inspection of the proofs shows that the lack of coverage in case of oversmoothing arises from a bias in the positioning of the posterior mean combined with a posterior spread that is smaller even than in the optimal case. In other words, the posterior is off mark, but believes it is very right. The message is that (too) smooth priors should be avoided; they lead to overconfident posteriors, which reflect the prior information rather than the data, even if the amount of information in the data increases indefinitely.

Under- and correct smoothing give very conservative confidence regions (coverage equal to 1). However, (i) and (ii) also show that the credible ball has the same order of magnitude as a correct confidence ball ($1 \geq \tilde{r}_{n,\gamma}/r_{n,\gamma} \gg 0$), so that the spread in the posterior does give the correct *order* of uncertainty. This at first sight surprising phenomenon is caused by the fact that the posterior distribution

concentrates near the boundary of a ball around its mean, and is not spread over the inside of the ball. The coverage is 1, because this sphere is larger than the corresponding sphere of the frequentist distribution of AY, even though the two radii are of the same order.

By Theorem 4.1 the optimal contraction rate is obtained (only) by a prior of the correct smoothness. Combining the two theorems leads to the conclusion that priors that slightly undersmooth the truth might be preferable. They attain a nearly optimal rate of contraction and the spread of their posterior gives a reasonable sense of uncertainty.

Scaling of the prior modifies these conclusions. The optimal scaling $\tau_n \asymp n^{(\alpha-\beta)/(1+2\alpha+2p)}$ found in Theorem 4.1, possible if $\beta < 1 + 2\alpha + 2p$, is covered in case (ii). This rescaling leads to a balancing of square bias, variance and spread, and to credible regions of the correct order of magnitude, although the precise (uniform) coverage can be any number in $[0, 1]$. Alternatively, bigger rescaling rates are covered in case (i) and lead to coverage 1. The optimal or slightly bigger rescaling rate seems the most sensible. It would be interesting to extend these results to data-dependent scaling.

5. Recovering linear functionals of the parameter.

We denote by $\Pi_n(\mu : L\mu \in \cdot |Y)$ the posterior distribution of the linear functional L, as described in Proposition 3.2. A continuous, linear functional $L : H_1 \to \mathbb{R}$ can be identified with an inner product $L\mu = \langle \mu, l \rangle_1$, for some $l \in H_1$, and hence with a sequence (l_i) in ℓ_2 giving its coordinates in the eigenbasis (e_i).

As shown in the proof of Proposition 3.2, for L in the larger class of $N(0, \Lambda)$-measurable linear functionals, the functional $L\Lambda^{1/2}$ is a continuous linear map on H_1 and hence can be identified with an element of H_1. For such a functional L we define a sequence (l_i) by $l_i = (L\Lambda^{1/2})_i/\sqrt{\lambda_i}$, for $((L\Lambda^{1/2})_i)$ the coordinates of $L\Lambda^{1/2}$ in the eigenbasis. The assumption that L is a $N(0, \Lambda)$-measurable linear functional implies that $\sum_i l_i^2 \lambda_i < \infty$, but (l_i) need not be contained in ℓ_2; if $(l_i) \in \ell_2$, then L is continuous and the definition of (l_i) agrees with the definition in the preceding paragraph.

We measure the smoothness of the functional L by the size of the coefficients l_i, as $i \to \infty$. First we assume that the sequence is in S^q, for some q.

THEOREM 5.1 (Contraction). *If μ_0, (λ_i), (κ_i) and (τ_n) are as in Assumption 3.1 and the representer (l_i) of the $N(0, \Lambda)$-measurable linear functional L is contained in S^q for $q \geq -\beta$, then $\mathrm{E}_{\mu_0}\Pi_n(\mu : |L\mu - L\mu_0| \geq M_n \varepsilon_n |Y) \to 0$, for every sequence $M_n \to \infty$, where*

$$\varepsilon_n = (n\tau_n^2)^{-(\beta+q)/(1+2\alpha+2p)\wedge 1} + \tau_n(n\tau_n^2)^{-(1/2+\alpha+q)/(1+2\alpha+2p)\wedge(1/2)}.$$

The rate is uniform over μ_0 in balls in S^β. In particular:

(i) *If $\tau_n \equiv 1$, then $\varepsilon_n = n^{-(\beta\wedge(1/2+\alpha)+q)/(1+2\alpha+2p)} \vee n^{-1/2}$.*

(ii) *If $q \leq p$ and $\beta + q \leq 1 + 2\alpha + 2p$ and $\tau_n \asymp n^{(1/2+\alpha-\beta)/(2\beta+2p)}$, then* $\varepsilon_n = n^{-(\beta+q)/(2\beta+2p)}$.

(iii) *If $q \leq p$ and $\beta + q > 1 + 2\alpha + 2p$, then $\varepsilon_n \gg n^{-(\beta+q)/(2\beta+2p)}$ for every scaling τ_n.*

(iv) *If $q \geq p$ and $\tau_n \gtrsim n^{(1/2+\alpha-\tilde{\beta}+p-q)/(2\tilde{\beta}+2q)}$, where $\tilde{\beta} = \beta \wedge (1 + 2\alpha + 2p - q)$, then $\varepsilon_n = n^{-1/2}$.*

If $q \geq p$, then the smoothness of the functional L cancels the ill-posedness of the operator K, and estimating $L\mu$ becomes a "regular" problem with an $n^{-1/2}$ rate of convergence. Without scaling the prior ($\tau_n \equiv 1$), the posterior contracts at this rate [see (i) or (iv)] if the prior is not too smooth ($\alpha \leq \beta - 1/2 + q - p$). With scaling, the rate is also attained, with any prior, provided the scaling parameter τ_n does not tend to zero too fast [see (iv)]. Inspection of the proof shows that too smooth priors or too small scale creates a bias that slows the rate.

If $q < p$, where we take q the "biggest" value such that $(l_i) \in S^q$, estimating $L\mu$ is still an inverse problem. The minimax rate over a ball in the Sobolev space S^β is known to be bounded above by $n^{-(\beta+q)/(2\beta+2p)}$ (see [8, 9, 16]).

This rate is attained without scaling [see (i): $\tau_n \equiv 1$] if and only if the prior smoothness α is equal to the true smoothness β minus $1/2$ ($\alpha + 1/2 = \beta$). An intuitive explanation for this apparent mismatch of prior and truth is that regularity of the parameter in the Sobolev scale ($\mu_0 \in S^\beta$) is not the appropriate type of regularity for estimating a linear functional $L\mu$. For instance, the difficulty of estimating a function at a point is determined by the regularity in a neighborhood of the point, whereas the Sobolev scale measures global regularity over the domain. The fact that a Sobolev space of order β embeds continuously in a Hölder space of regularity $\beta - 1/2$ might give a quantitative explanation of the "loss" in smoothness by $1/2$ in the special case that the eigenbasis is the Fourier basis. In our Bayesian context we draw the conclusion that the prior must be adapted to the inference problem if we want to obtain the optimal frequentist rate: for estimating the global parameter, a good prior must match the truth ($\alpha = \beta$), but for estimating a linear functional a good prior must consider a Sobolev truth of order β as having regularity $\alpha = \beta - 1/2$.

If the prior smoothness α is not $\beta - 1/2$, then the minimax rate may still be attained by scaling the prior. As in the global problem, this is possible only if the prior is not too rough [$\beta + q \leq 1 + 2\alpha + 2p$, cases (ii) and (iii)]. The optimal scaling when this is possible [case (ii)] is the same as the optimal scaling for the global problem [Theorem 4.1(ii)] *after* decreasing β by $1/2$. So the "loss in regularity" persists in the scaling rate. Heuristically this seems to imply that a simple data-based procedure to set the scaling, such as empirical or hierarchical Bayes, cannot attain simultaneous optimality in both the global and local senses.

In the application of the preceding theorem, the functional L, and hence the sequence (l_i), will be given. Naturally, we apply the theorem with q equal

to the largest value such that $(l_i) \in S^q$. Unfortunately, this lacks precision for the sequences (l_i) that decrease exactly at some polynomial order: a sequence $l_i \asymp i^{-q-1/2}$ is in $S^{q'}$ for every $q' < q$, but not in S^q. In the following theorem we consider these sequences, and the slightly more general ones such that $|l_i| \leq i^{-q-1/2}\mathcal{S}(i)$, for some slowly varying sequence $\mathcal{S}(i)$. Recall that $\mathcal{S}:[0,\infty) \to \mathbb{R}$ is *slowly varying* if $\mathcal{S}(tx)/\mathcal{S}(t) \to 1$ as $t \to \infty$, for every $x > 0$. [For these sequences $(l_i) \in S^{q'}$ for every $q' < q$, $(l_i) \notin S^{q'}$ for $q' > q$, and $(l_i) \in S^q$ if and only if $\sum_i \mathcal{S}^2(i)/i < \infty$.]

THEOREM 5.2 (Contraction). *If μ_0, (λ_i), (κ_i) and (τ_n) are as in Assumption* 3.1 *and the representer (l_i) of the $N(0, \Lambda)$-measurable linear functional L satisfies $|l_i| \leq i^{-q-1/2}\mathcal{S}(i)$ for a slowly varying function $\mathcal{S}$ and $q > -\beta$, then the result of Theorem* 5.1 *is valid with*

$$(5.1) \quad \varepsilon_n = (n\tau_n^2)^{-(\beta+q)/(1+2\alpha+2p)\wedge 1}\gamma_n + \tau_n(n\tau_n^2)^{-(1/2+\alpha+q)/(1+2\alpha+2p)\wedge(1/2)}\delta_n,$$

where, for $\rho_n = (n\tau_n^2)^{1/(1+2\alpha+2p)}$,

$$\gamma_n^2 = \begin{cases} \mathcal{S}^2(\rho_n), & \text{if } \beta + q < 1 + 2\alpha + 2p, \\ \displaystyle\sum_{i \leq \rho_n} \frac{\mathcal{S}^2(i)}{i}, & \text{if } \beta + q = 1 + 2\alpha + 2p, \\ 1, & \text{if } \beta + q > 1 + 2\alpha + 2p, \end{cases}$$

$$\delta_n^2 = \begin{cases} \mathcal{S}^2(\rho_n), & \text{if } q < p, \\ \displaystyle\sum_{i \leq \rho_n} \frac{\mathcal{S}^2(i)}{i}, & \text{if } q = p, \\ 1, & \text{if } q > p. \end{cases}$$

This has the same consequences as in Theorem 5.1, *up to the addition of slowly varying terms.*

Because the posterior distribution for the linear functional $L\mu$ is the one-dimensional normal distribution $N(LAY, s_n^2)$, the natural *credible interval* for $L\mu$ has endpoints $LAY \pm z_{\gamma/2}s_n$, for z_γ the (lower) standard normal γ-quantile. The *coverage* of this interval is

$$P_{\mu_0}(LAY + z_{\gamma/2}s_n \leq L\mu_0 \leq LAY - z_{\gamma/2}s_n),$$

where Y follows (1.1) with $\mu = \mu_0$. To obtain precise results concerning coverage, we assume that (l_i) behaves polynomially up to a slowly varying term, first in the situation $q < p$ that estimating $L\mu$ is an inverse problem. Let $\tilde{\tau}_n$ be the (optimal) scaling τ_n that equates the two terms in the right-hand side of (5.1). This satisfies $\tilde{\tau}_n \asymp n^{(1/2+\alpha-\tilde{\beta})/(2\tilde{\beta}+2p)}\eta_n$, for a slowly varying factor η_n, where $\tilde{\beta} = \beta \wedge (1 + 2\alpha + 2p - q)$.

THEOREM 5.3 (Credibility). *Let μ_0, (λ_i), (κ_i) and (τ_n) be as in Assumption* 3.1, *and let $|l_i| = i^{-q-1/2}\mathcal{S}(i)$ for $q < p$ and a slowly varying function $\mathcal{S}$. Then the asymptotic coverage of the interval $LAY \pm z_{\gamma/2}s_n$ is:*

(i) *in $(1 - \gamma, 1)$, uniformly in μ_0 such that $\|\mu_0\|_\beta \leq 1$ if $\tau_n \gg \tilde{\tau}_n$.*

(ii) *in $(1 - \gamma, 1)$, for every $\mu_0 \in S^\beta$, if $\tau_n \asymp \tilde{\tau}_n$ and $\beta + q < 1 + 2\alpha + 2p$; in $(0, c)$, along some μ_0^n with $\sup_n \|\mu_0^n\|_\beta < \infty$ if $\tau_n \asymp \tilde{\tau}_n$ [any $c \in (0, 1)$].*

(iii) *0 along some μ_0^n with $\sup_n \|\mu_0^n\|_\beta < \infty$ if $\tau_n \ll \tilde{\tau}_n$.*

In case (iii) *the sequence μ_0^n can be taken a fixed element μ_0 in S^β if $\tau_n \lesssim n^{-\delta}\tilde{\tau}_n$ for some $\delta > 0$.*

Furthermore, if $\tau_n \equiv 1$, then the coverage takes the form as in (i), (ii) *and* (iii) *if $\alpha < \beta - 1/2$, $\alpha = \beta - 1/2$, and $\alpha > \beta - 1/2$, respectively, where in case* (iii) *the sequence μ_0^n can be taken a fixed element.*

Similarly, as in the nonparametric problem, oversmoothing leads to coverage 0, while undersmoothing gives conservative intervals. Without scaling the cut-off for under- or oversmoothing is at $\alpha = \beta - 1/2$; with scaling the cut-off for the scaling rate is at the optimal rate $\tilde{\tau}_n$.

The conservativeness in the case of undersmoothing is less extreme for functionals than for the full parameter, as the coverage is strictly between the credibility level $1 - \gamma$ and 1. The general message is the same: oversmoothing is disastrous for the interpretation of credible band, whereas undersmoothing gives bands that at least have the correct order of magnitude, in the sense that their width is of the same order as the variance of the posterior mean (see the proof). Too much undersmoothing is also undesirable, as it leads to very wide confidence bands, and may cause that $\sum_i l_i^2\lambda_i$ is no longer finite (see measurability property).

The results (i) and (ii) are the same for every $q < p$, even if $\tau_n \equiv 1$. Closer inspection would reveal that for a given μ_0 the exact coverage depends on q [and $\mathcal{S}(i)$] in a complicated way.

If $q \geq p$, then the smoothness of the functional L compensates the lack of smoothness of K^{-1}, and estimating $L\mu$ is not a true inverse problem. This drastically changes the performance of credible intervals. Although oversmoothing again destroys their coverage, credible intervals are exact confidence sets if the prior is not too smooth. We formulate this in terms of a Bernstein–von Mises theorem.

The Bernstein–von Mises theorem for parametric models asserts that the posterior distribution approaches a normal distribution centered at an efficient estimator of the parameter and with variance equal to its asymptotic variance. It is the ultimate link between Bayesian and frequentist procedures. There is no version of this theorem for infinite-dimensional parameters [12], but the theorem may hold for "smooth" finite-dimensional projections, such as the linear functional $L\mu$ (see [5]).

In the present situation the posterior distribution of $L\mu$ is already normal by the normality of the model and the prior: it is a $N(LAY, s_n^2)$-distribution by Proposition 3.2. To speak of a Bernstein–von Mises theorem, we also require the following:

(i) That the (root of the) spread s_n of the posterior distribution is asymptotically equivalent to the standard deviation t_n of the centering variable LAY.

(ii) That the sequence $(LAY - L\mu_0)/t_n$ tends in distribution to a standard normal distribution.

(iii) That the centering LAY is an asymptotically efficient estimator of $L\mu$.

We shall show that (i) happens if and only if the functional L cancels the ill-posedness of the operator K, that is, if $q \geq p$ in Theorem 5.2. Interestingly, the rate of convergence t_n must be $n^{-1/2}$ up to a slowly varying factor in this case, but it could be strictly slower than $n^{-1/2}$ by a slowly varying factor increasing to infinity.

Because LAY is normally distributed by the normality of the model, assertion (ii) is equivalent to saying that its bias tends to zero faster than t_n. This happens provided the prior does not oversmooth the truth too much. For very smooth functionals ($q > p$) there is some extra "space" in the cut-off for the smoothness, which (if the prior is not scaled: $\tau_n \equiv 1$) is at $\alpha = \beta - 1/2 + q - p$, rather than at $\alpha = \beta - 1/2$ as for the (global) inverse estimating problem. Thus, the prior may be considerably smoother than the truth if the functional is very smooth.

Let $\|\cdot\|$ denote the total variation norm between measures. Say that $l \in \mathcal{R}^q$ if $|l_i| = i^{-q-1/2}\mathcal{S}(i)$ for a slowly varying function $\mathcal{S}$. Write

$$B_n = \sup_{\|\mu\|_\beta \lesssim 1} |LAK\mu - L\mu|$$

for the maximal bias of LAY over a ball in the Sobolev space S^β. Finally, let $\tilde{\tau}_n$ be the (optimal) scaling τ_n in that it equates the two terms in the right-hand side of (5.1).

THEOREM 5.4 (Bernstein–von Mises). *Let μ_0, (λ_i), and (κ_i) be as in Assumption 3.1, and let l be the representer of the $N(0, \Lambda)$-measurable linear functional L:*

(i) *If $l \in S^p$, then $s_n/t_n \to 1$; in this case $nt_n^2 \to \sum_i l_i^2/\kappa_i^2$. If $l \in \mathcal{R}^q$, then $s_n/t_n \to 1$ if and only if $q \geq p$; in this case $n \mapsto nt_n^2$ is slowly varying.*

(ii) *If $l \in S^q$ for $q \geq p$, then $B_n = o(t_n)$ if either $\tau_n \gg n^{(\alpha+1/2-\beta)/(2\beta+2q)}$ or $(\tau_n \equiv 1$ and $\alpha < \beta - 1/2 + q - p)$. If $l \in \mathcal{R}^q$ for $q \geq p$, then $B_n = o(t_n)$ if $(\tau_n \gg \tilde{\tau}_n)$ or $(\tau_n \equiv 1$ and $\alpha < \beta - 1/2 + q - p)$ or $(q = p, \tau_n \equiv 1$ and $\alpha = \beta - 1/2 + q - p)$ or $[q > p, \tau_n \equiv 1$ and $\alpha = \beta - 1/2 + q - p$ and $\mathcal{S}(i) \to 0$ as $i \to \infty]$.*

(iii) *If $l \in S^p$ or $l \in \mathcal{R}^p$ and $B_n = o(t_n)$, then $\mathrm{E}_{\mu_0}\|\Pi_n(L\mu \in \cdot|Y) - N(LAY, t_n^2)\| \to 0$ and $(LAY - L\mu_0)/t_n$ converges under μ_0 in distribution to a standard*

normal distribution, uniformly in $\|\mu_0\|_\beta \lesssim 1$. If $l \in S^p$, then this is also true with LAY and t_n^2 replaced by $\sum_i Y_i l_i / \kappa_i$ and its variance $n^{-1} \sum_i l_i^2 / \kappa_i^2$.

In both cases (iii), *the asymptotic coverage of the credible interval $LAY \pm z_{\gamma/2} s_n$ is $1 - \gamma$, uniformly in $\|\mu_0\|_\beta \lesssim 1$. Finally, if the conditions under* (ii) *fail, then there exists μ_0^n with $\sup_n \|\mu_0^n\|_\beta < \infty$ along which the coverage tends to an arbitrarily low value.*

The observation Y in (1.1) can be viewed as a reduction by sufficiency of a random sample of size n from the distribution $N(K\mu, I)$. Therefore, the model fits in the framework of i.i.d. observations, and "asymptotic efficiency" can be defined in the sense of semiparametric models discussed in, for example, [2, 30] and [31]. Because the model is shift-equivariant, it suffices to consider local efficiency at $\mu_0 = 0$. The one-dimensional submodels $N(K(th), I)$ on the sample space $\mathbb{R}^{H_2}$, for $t \in \mathbb{R}$ and a fixed "direction" $h \in H_1$, have likelihood ratios

$$\log \frac{dN(tKh, I)}{dN(0, I)}(Y) = t Y_{Kh} - \frac{1}{2} t^2 \|Kh\|_2^2.$$

Thus, their *score function* at $t = 0$ is the (Kh)th coordinate of a single observation $Y = (Y_h : h \in H_2)$, the *score operator* is the map $\tilde{K} : H_1 \to L_2(N(0, I))$ given by $\tilde{K}h(Y) = Y_{Kh}$, and the *tangent space* is the range of $\tilde{K}$. [We denote the score operator by the same symbol K as in (1.1); if the observation Y *were* realizable in H_2, and not just in the bigger sample space $\mathbb{R}^{H_2}$, then Y_{Kh} would correspond to $\langle Y, Kh \rangle_2$ and, hence, the score would be exactly Kh for the operator in (1.1) after identifying H_2 and its dual space.] The adjoint of the score operator restricted to the closure of the tangent space is the operator $\tilde{K}^T : \overline{\tilde{K}H_1} \subset L_2(N(0, I)) \to H_1$ that satisfies $\tilde{K}^T(Y_g) = K^T g$, where K^T on the right is the adjoint of $K : H_1 \to H_2$. The functional $L\mu = \langle l, \mu \rangle_1$ has derivative l. Therefore, by [28] asymptotically regular sequences of estimators exist, and the local asymptotic minimax bound for estimating $L\mu$ is finite, if and only if l is contained in the range of K^T. Furthermore, the variance bound is $\|m\|_2^2$ for $m \in H_2$ such that $K^T m = l$.

In our situation the range of K^T is S^p, and if $l \in S^p$, then by Theorem 5.4(iii) the variance of the posterior is asymptotically equivalent to the variance bound and its centering can be taken equal to the estimator $n^{-1} \sum_i Y_i l_i / \kappa_i$, which attains this variance bound. Thus, the theorem gives a semiparametric Bernstein–von Mises theorem, satisfying every of (i), (ii), (iii) in this case. If only $l \in \mathcal{R}^p$ and not $l \in S^p$, the theorem still gives a Bernstein–von Mises type theorem, but the rate of convergence is slower than $n^{-1/2}$, and the standard efficiency theory does not apply.

6. Example—Volterra operator. The classical *Volterra operator* $K : L^2[0, 1] \to L^2[0, 1]$ and its adjoint K^T are given by

$$K\mu(x) = \int_0^x \mu(s) \, ds, \qquad K^T \mu(x) = \int_x^1 \mu(s) \, ds.$$

The resulting problem (1.1) can also be written in "signal in white noise" form as follows: observe the process $(Y_t : t \in [0, 1])$ given by $Y_t = \int_0^t \int_0^s \mu(u) \, du \, ds + n^{-1/2} W_t$, for a Brownian motion W.

The eigenvalues, eigenfunctions of $K^T K$ and conjugate basis are given by (see [17]), for $i = 1, 2, \ldots,$

$$\kappa_i^2 = \frac{1}{(i - 1/2)^2 \pi^2}, \qquad e_i(x) = \sqrt{2} \cos((i - 1/2)\pi x),$$

$$f_i(x) = \sqrt{2} \sin((i - 1/2)\pi x).$$

The (f_i) are the eigenfunctions of KK^T, relative to the same eigenvalues, and $Ke_i = \kappa_i f_i$ and $K^T f_i = \kappa_i e_i$, for every $i \in \mathbb{N}$.

To illustrate our results with simulated data, we start by choosing a true function μ_0, which we expand as $\mu_0 = \sum_i \mu_{0,i} e_i$ on the basis (e_i). The data are the function

$$Y = K\mu_0 + \frac{1}{\sqrt{n}} Z = \sum_i \mu_{0,i} \kappa_i f_i + \frac{1}{\sqrt{n}} Z.$$

It can be generated relative to the conjugate basis (f_i) as a sequence of independent Gaussian random variables $Y_1, Y_2, \ldots$ with $Y_i \sim N(\mu_{0,i} \kappa_i, n^{-1/2})$. The posterior distribution of μ is Gaussian with mean AY and covariance operator S_n, given in Proposition 3.1. Under Assumption 3.1 it can be represented in terms of the coordinates (μ_i) of μ relative to the basis (e_i) as (conditionally) independent Gaussian variables $\mu_1, \mu_2, \ldots$ with

$$\mu_i | Y \sim N\left(\frac{n \lambda_i \kappa_i Y_i}{1 + n \lambda_i \kappa_i^2}, \frac{\lambda_i}{1 + n \lambda_i \kappa_i^2}\right).$$

The (marginal) posterior distribution for the function μ at a point x is obtained by expanding $\mu(x) = \sum_i \mu_i e_i(x)$, and applying the framework of linear functionals $L\mu = \sum_i l_i \mu_i$ with $l_i = e_i(x)$. This shows that

$$\mu(x) | Y \sim N\left(\sum_i \frac{n \lambda_i \kappa_i Y_i e_i(x)}{1 + n \lambda_i \kappa_i^2}, \sum_i \frac{\lambda_i e_i(x)^2}{1 + n \lambda_i \kappa_i^2}\right).$$

We obtained (marginal) posterior credible bands by computing for every x a central 95% interval in the normal distribution on the right-hand side.

Figure 1 illustrates these bands for $n = 1{,}000$. In every one of the 10 panels in the figure the black curve represents the function μ_0, defined by the coefficients $i^{-3/2} \sin(i)$ relative to e_i ($\beta = 1$). The 10 panels represent 10 independent realizations of the data, yielding 10 different realizations of the posterior mean (the red curves) and the posterior credible bands (the green curves). In the left five panels the prior is given by $\lambda_i = i^{-2\alpha-1}$ with $\alpha = 1$, whereas in the right panels the prior corresponds to $\alpha = 5$. Each of the 10 panels also shows 20 realizations from the posterior distribution.
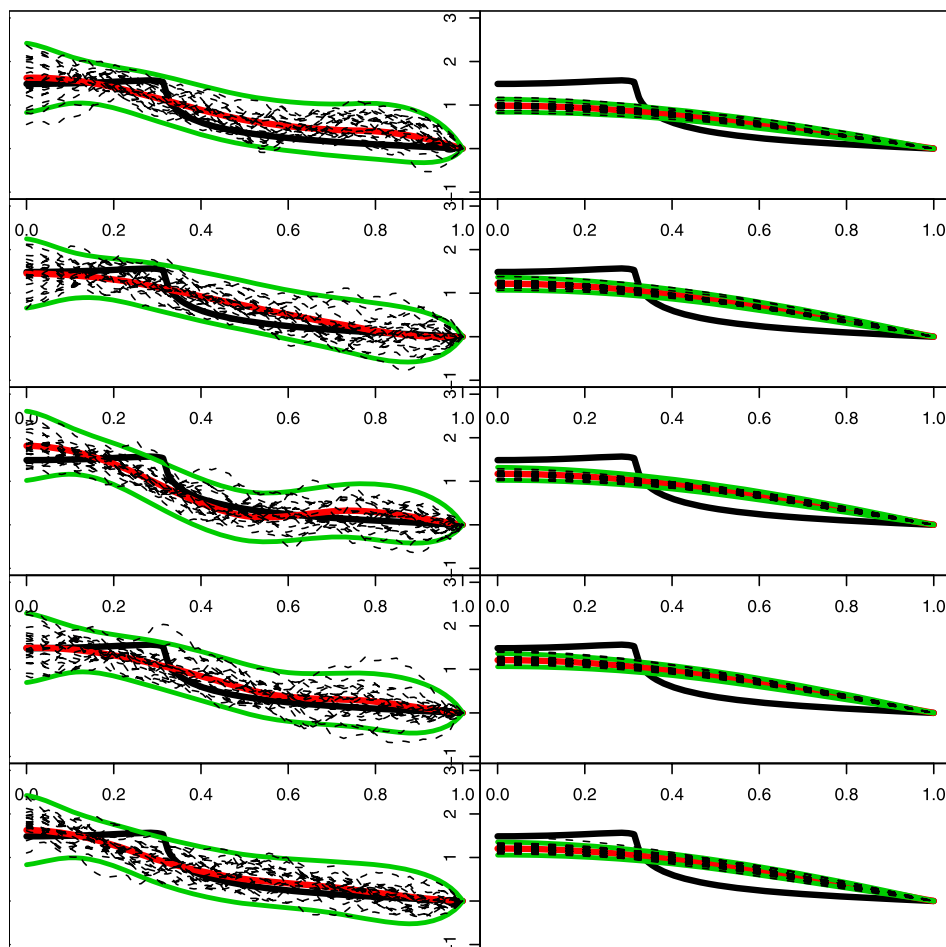
FIG. 1. *Realizations of the posterior mean* (*red*) *and* (*marginal*) *posterior credible bands* (*green*), *and* 20 *draws from the posterior* (*dashed curves*). *In all ten panels* $n = 1{,}000$ *and* $\beta = 1$. *Left 5 panels*: $\alpha = 1$; *right 5 panels*: $\alpha = 5$. *True curve* (*black*) *given by coefficients* $\mu_{0,i} = i^{-3/2} \sin(i)$.

Clearly, the posterior mean is not estimating the true curve very well, even for $n = 1{,}000$. This is mostly caused by the intrinsic difficulty of the inverse problem: better estimation requires bigger sample size. A comparison of the left and right panels shows that the rough prior ($\alpha = 1$) is aware of the difficulty: it produces credible bands that in (almost) all cases contain the true curve. On the other hand, the smooth prior ($\alpha = 5$) is overconfident; the spread of the posterior distribution poorly reflects the imprecision of estimation.

Specifying a prior that is too smooth relative to the true curve yields a posterior distribution which gives both a bad reconstruction and a misguided sense of uncer-
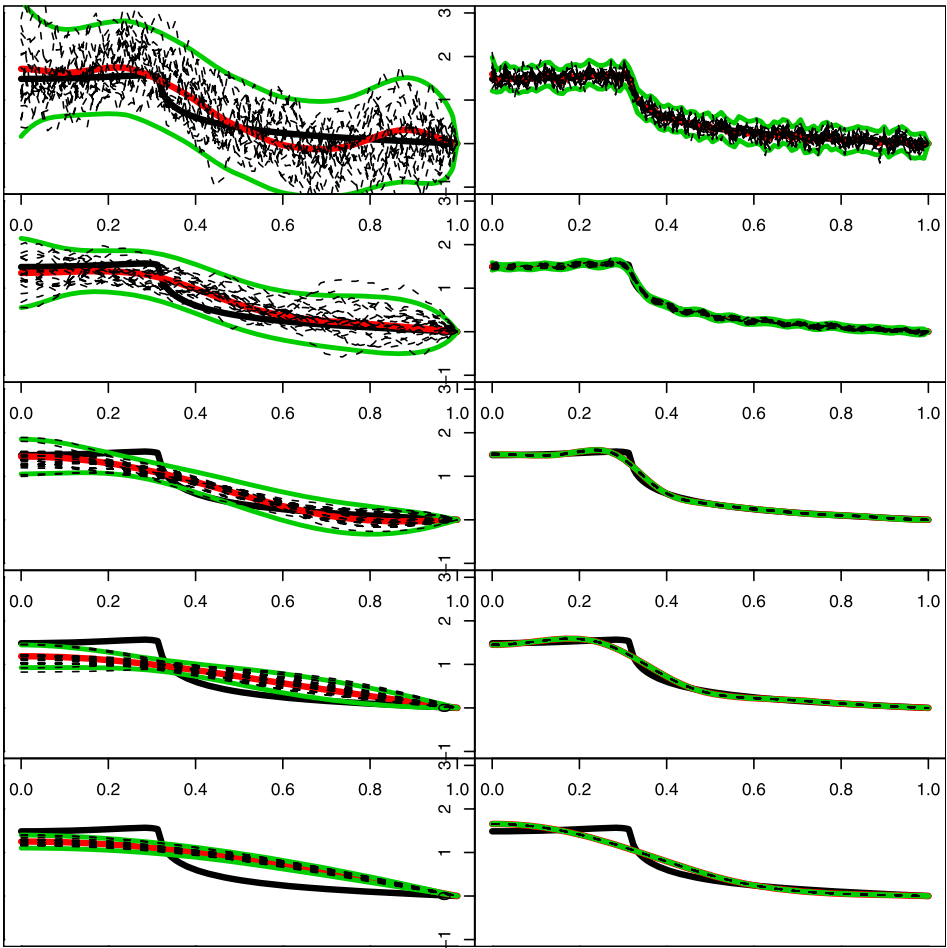
FIG. 2. *Realizations of the posterior mean* (red) *and* (marginal) *posterior credible bands* (green), *and* 20 *draws from the posterior* (dashed curves). *In all ten panels* $\beta = 1$. *Left* 5 *panels*: $n = 1,000$ *and* $\alpha = 0.5, 1, 2, 3, 5$ (*top to bottom*); *right* 5 *panels*: $n = 10^8$ *and* $\alpha = 0.5, 1, 2, 3, 5$ (*top to bottom*). *True curve* (black) *given by coefficients* $\mu_{0,i} = i^{-3/2} \sin(i)$.

tainty. Our theoretical results show that the inaccurate quantification of estimation error remains even as $n \to \infty$.

The reconstruction, by the posterior mean or any other posterior quantiles, will eventually converge to the true curve. However, specification of a too smooth prior will slow down this convergence significantly. This is illustrated in Figure 2. Every one of its 10 panels is similarly constructed as before, but now with $n = 1,000$ and $n = 10^8$ for the five panels on the left-hand and right-hand side, respectively, and with $\alpha = 1/2, 1, 2, 3, 5$ for the five panels from top to bottom. At first sight $\alpha = 1$ seems better (see the left column in Figure 2), but leads to zero coverage because

of the mismatch close to the bump (see the right column), while $\alpha = 1/2$ captures the bump. For $n = 10^8$ the posterior for this optimal prior has collapsed onto the true curve, whereas the smooth posterior for $\alpha = 5$ still has major difficulty in recovering the bump in the true curve (even though it "thinks" it has captured the correct curve, the bands having collapsed to a single curve in the figure).

7. Proofs.

7.1. *Proof of Theorem* 4.1. The second moment of a Gaussian distribution on H_1 is equal to the square norm of its mean plus the trace of its covariance operator. Because the posterior is Gaussian $N(AY, S_n)$, it follows that

$$\int \|\mu - \mu_0\|_1^2 \, d\Pi_n(\mu|Y) = \|AY - \mu_0\|_1^2 + \operatorname{tr}(S_n).$$

By Markov's inequality, the left-hand side is an upper bound to $M_n^2 \varepsilon_n^2 \Pi_n(\mu : \|\mu - \mu_0\|_1 \geq M_n \varepsilon_n | Y)$. Therefore, it suffices to show that the expectation under μ_0 of the right-hand side of the display is bounded by a multiple of ε_n^2. The expectation of the first term is the mean square error of the posterior mean AY, and can be written as the sum $\|AK\mu_0 - \mu_0\|_1^2 + n^{-1}\operatorname{tr}(AA^T)$ of its square bias and "variance." The second term $\operatorname{tr}(S_n)$ is deterministic. Under Assumption 3.1 the three quantities can be expressed in the coefficients relative to the eigenbasis (e_i) as

$$(7.1) \quad \|AK\mu_0 - \mu_0\|_1^2 = \sum_i \frac{\mu_{0,i}^2}{(1 + n\lambda_i\kappa_i^2)^2} \asymp \sum_i \frac{\mu_{0,i}^2}{(1 + n\tau_n^2 i^{-1-2\alpha-2p})^2},$$

$$(7.2) \quad \frac{1}{n}\operatorname{tr}(AA^T) = \sum_i \frac{n\lambda_i^2\kappa_i^2}{(1 + n\lambda_i\kappa_i^2)^2} \asymp \sum_i \frac{n\tau_n^4 i^{-2-4\alpha-2p}}{(1 + n\tau_n^2 i^{-1-2\alpha-2p})^2},$$

$$(7.3) \quad \operatorname{tr}(S_n) = \sum_i \frac{\lambda_i}{1 + n\lambda_i\kappa_i^2} \asymp \sum_i \frac{\tau_n^2 i^{-1-2\alpha}}{1 + n\tau_n^2 i^{-1-2\alpha-2p}}.$$

By Lemma 8.1 (applied with $q = \beta$, $t = 0$, $u = 1 + 2\alpha + 2p$, $v = 2$ and $N = n\tau_n^2$), the first can be bounded by $\|\mu_0\|_\beta^2 (n\tau_n^2)^{-(2\beta)/(1+2\alpha+2p)\wedge 2}$, which accounts for the first term in the definition of ε_n. By Lemma 8.2 [applied with $\mathcal{S}(i) = 1$, $q = -1/2$, $t = 2 + 4\alpha + 2p$, $u = 1 + 2\alpha + 2p$, $v = 2$, and $N = n\tau_n^2$], and again Lemma 8.2 [applied with $\mathcal{S}(i) = 1$, $q = -1/2$, $t = 1 + 2\alpha$, $u = 1 + 2\alpha + 2p$, $v = 1$ and $N = n\tau_n^2$], both the second and third expressions are of the order the square of the second term in the definition of ε_n.

The consequences (i) and (ii) follow by verification after substitution of τ_n as given. To prove consequence (iii), we note that the two terms in the definition of ε_n are decreasing and increasing in τ_n, respectively. Therefore, the maximum of these two terms is minimized with respect to τ_n by equating the two terms. This minimum (assumed at $\tau_n = n^{-(1+\alpha+2p)/(3+4\alpha+6p)}$) is much bigger than $n^{-\beta/(1+2\beta+2p)}$ if $\beta > 1 + 2\alpha + 2p$.

7.2. *Proof of Theorem* 5.1.   By Proposition 3.2 the posterior distribution is $N(LAY, s_n^2)$, and, hence, similarly as in the proof of Theorem 4.1, it suffices to show that

$$\mathrm{E}_{\mu_0}|LAY - L\mu_0|^2 + s_n^2 = |LAK\mu_0 - L\mu_0|^2 + \frac{1}{n}\|LA\|_1^2 + s_n^2$$

is bounded above by a multiple of ε_n^2. Under Assumption 3.1 the expressions on the right can be written

$$(7.4) \qquad LAK\mu_0 - L\mu_0 = -\sum_i \frac{l_i\mu_{0,i}}{1 + n\lambda_i\kappa_i^2} \lesssim \sum_i \frac{|l_i\mu_{0,i}|}{1 + n\tau_n^2 i^{-1-2\alpha-2p}},$$

$$t_n^2 := \frac{1}{n}\|LA\|_1^2 = \sum_i \frac{l_i^2 n\lambda_i^2\kappa_i^2}{(1 + n\lambda_i\kappa_i^2)^2}$$

$$(7.5)$$

$$\asymp n\tau_n^4 \sum_i \frac{l_i^2 i^{-2-4\alpha-2p}}{(1 + n\tau_n^2 i^{-1-2\alpha-2p})^2},$$

$$(7.6) \qquad s_n^2 = \sum_i \frac{l_i^2\lambda_i}{1 + n\lambda_i\kappa_i^2} \asymp \tau_n^2 \sum_i \frac{l_i^2 i^{-1-2\alpha}}{1 + n\tau_n^2 i^{-1-2\alpha-2p}}.$$

By the Cauchy–Schwarz inequality the square of the bias (7.4) satisfies

$$(7.7) \qquad |LAK\mu_0 - L\mu_0|^2 \lesssim \|\mu_0\|_\beta^2 \sum_i \frac{l_i^2 i^{-2\beta}}{(1 + n\tau_n^2 i^{-1-2\alpha-2p})^2}.$$

By Lemma 8.1 (applied with $q = q, t = 2\beta, u = 1 + 2\alpha + 2p, v = 2$ and $N = n\tau_n^2$) the right-hand side of this display can be further bounded by $\|\mu_0\|_\beta^2\|l\|_q^2$ times the square of the first term in the sum of two terms that defines ε_n. By Lemma 8.1 (applied with $q = q, t = 2 + 4\alpha + 2p, u = 1 + 2\alpha + 2p, v = 2$ and $N = n\tau_n^2$) and again Lemma 8.1 (applied with $q = q, t = 1 + 2\alpha, u = 1 + 2\alpha + 2p, v = 1$ and $N = n\tau_n^2$), the right-hand sides of (7.5) and (7.6) are bounded above by $\|l\|_q^2$ times the square of the second term in the definition of ε_n.

Consequences (i)–(iv) follow by substitution, and, in the case of (iii), optimization over τ_n.

7.3. *Proof of Theorem* 5.2.   This follows the same lines as the proof of Theorem 5.1, except that we use Lemma 8.2 (with $q = q, t = 2\beta, u = 1 + 2\alpha + 2p, v = 2$ and $N = n\tau_n^2$) and Lemma 8.2 (with $q = q, t = 2 + 4\alpha + 2p, u = 1 + 2\alpha + 2p, v = 2$ and $N = n\tau_n^2$) and again Lemma 8.2 (with $q = q, t = 1 + 2\alpha, u = 1 + 2\alpha + 2p, v = 1$ and $N = n\tau_n^2$) to bound the three terms (7.5)–(7.7).

7.4. *Proof of Theorem* 4.2. Because the posterior distribution is $N(AY, S_n)$, by Proposition 3.1, the radius $r_{n,\gamma}$ in (4.3) satisfies $P(U_n < r_{n,\gamma}^2) = 1 - \gamma$, for U_n a random variable distributed as the square norm of an $N(0, S_n)$-variable. Under (1.1) the variable AY is $N(AK\mu_0, n^{-1}AA^T)$-distributed, and, thus, the coverage (4.4) can be written as

$$(7.8) \qquad P(\|W_n + AK\mu_0 - \mu_0\|_1 \leq r_{n,\gamma})$$

for W_n possessing a $N(0, n^{-1}AA^T)$-distribution. For ease of notation let $V_n = \|W_n\|_1^2$.

The variables U_n and V_n can be represented as $U_n = \sum_i s_{i,n} Z_i^2$ and $V_n = \sum_i t_{i,n} Z_i^2$, for $Z_1, Z_2, \ldots$ independent standard normal variables, and $s_{i,n}$ and $t_{i,n}$ the eigenvalues of S_n and $n^{-1}AA^T$, respectively, which satisfy

$$s_{i,n} = \frac{\lambda_i}{1 + n\lambda_i\kappa_i^2} \asymp \frac{\tau_n^2 i^{-2\alpha-1}}{1 + n\tau_n^2 i^{-2\alpha-2p-1}},$$

$$t_{i,n} = \frac{n\lambda_i^2\kappa_i^2}{(1 + n\lambda_i\kappa_i^2)^2} \asymp \frac{n\tau_n^4 i^{-4\alpha-2p-2}}{(1 + n\tau_n^2 i^{-2\alpha-2p-1})^2},$$

$$s_{i,n} - t_{i,n} = \frac{\lambda_i}{(1 + n\lambda_i\kappa_i^2)^2} \asymp \frac{\tau_n^2 i^{-2\alpha-1}}{(1 + n\tau_n^2 i^{-2\alpha-2p-1})^2}.$$

Therefore, by Lemma 8.2 (applied with $\mathcal{S} \equiv 1$ and $q = -1/2$; always the first case),

$$EU_n = \sum_i s_{i,n} \asymp \tau_n^2 (n\tau_n^2)^{-2\alpha/(1+2\alpha+2p)},$$

$$EV_n = \sum_i t_{i,n} \asymp \tau_n^2 (n\tau_n^2)^{-2\alpha/(1+2\alpha+2p)},$$

$$E(U_n - V_n) = \sum_i (s_{i,n} - t_{i,n}) \asymp \tau_n^2 (n\tau_n^2)^{-2\alpha/(1+2\alpha+2p)},$$

$$\mathrm{var}\, U_n = 2\sum_i s_{i,n}^2 \asymp \tau_n^4 (n\tau_n^2)^{-(1+4\alpha)/(1+2\alpha+2p)},$$

$$\mathrm{var}\, V_n = 2\sum_i t_{i,n}^2 \asymp \tau_n^4 (n\tau_n^2)^{-(1+4\alpha)/(1+2\alpha+2p)}.$$

We conclude that the standard deviations of U_n and V_n are negligible relative to their means, and also relative to the difference $E(U_n - V_n)$ of their means. Because $U_n \geq V_n$, we conclude that the distributions of U_n and V_n are asymptotically completely separated: $P(V_n \leq v_n \leq U_n) \to 1$ for some v_n [e.g., $v_n = E(U_n + V_n)/2$]. The numbers $r_{n,\gamma}^2$ are $1 - \gamma$-quantiles of U_n, and, hence, $P(V_n \leq r_{n,\gamma}^2(1 + o(1))) \to 1$. Furthermore, it follows that

$$r_{n,\gamma}^2 \asymp \tau_n^2 (n\tau_n^2)^{-2\alpha/(1+2\alpha+2p)} \asymp EU_n \asymp EV_n.$$

The square norm of the bias $AK\mu_0 - \mu_0$ is given in (7.1), where it was noted that

$$B_n := \sup_{\|\mu_0\|_\beta \lesssim 1} \|AK\mu_0 - \mu_0\|_1 \asymp (n\tau_n^2)^{-\beta/(1+2\alpha+2p)\wedge 1}.$$

The bias B_n is decreasing in τ_n, whereas EU_n and $\operatorname{var} U_n$ are increasing. The scaling rate $\tilde{\tau}_n \asymp n^{(\alpha-\tilde{\beta})/(1+2\tilde{\beta}+2p)}$ balances the square bias B_n^2 with the variance EV_n of the posterior mean, and hence with $r_{n,\gamma}^2$.

Case (i). In this case $B_n \ll r_{n,\gamma}$. Hence, $P(\|W_n + AK\mu_0 - \mu_0\|_1 \le r_{n,\gamma}) \ge P(\|W_n\|_1 \le r_{n,\gamma} - B_n) = P(V_n \le r_{n,\gamma}^2(1 + o(1))) \to 1$, uniformly in the set of μ_0 in the supremum defining B_n.

Case (iii). In this case $B_n \gg r_{n,\gamma}$. Hence, $P(\|W_n + AK\mu_0^n - \mu_0^n\|_1 \le r_{n,\gamma}) \le P(\|W_n\|_1 \ge B_n - r_{n,\gamma}) \to 0$ for any sequence μ_0^n (nearly) attaining the supremum in the definition of B_n. If $\tau_n \equiv 1$, then B_n and $r_{n,\gamma}$ are both powers of $1/n$ and, hence, $B_n \gg r_{n,\gamma}$ implies that $B_n \gtrsim r_{n,\gamma}n^\delta$, for some $\delta > 0$. The preceding argument then applies for a fixed μ_0 of the form $\mu_{0,i} \asymp i^{-1/2-\beta-\varepsilon}$, for small $\varepsilon > 0$, that gives a bias that is much closer than n^δ to B_n.

Case (ii). In this case $B_n \asymp r_{n,\gamma}$. If $\beta < 1 + 2\alpha + 2p$, then by the second assertion (first case) of Lemma 8.1 the bias $\|AK\mu_0 - \mu_0\|_1$ at a fixed μ_0 is of strictly smaller order than the supremum B_n. The argument of (i) shows that the asymptotic coverage then tends to 1.

Finally, we prove the existence of a sequence μ_0^n along which the coverage is a given $c \in [0, 1)$. The coverage (7.8) with μ_0 replaced by μ_0^n tends to c if, for $b_n = AK\mu_0^n - \mu_0^n$ and z_c a standard normal quantile,

$$(7.9) \qquad \frac{\|W_n + b_n\|_1^2 - E\|W_n + b_n\|_1^2}{\operatorname{sd}\|W_n + b_n\|_1^2} \rightsquigarrow N(0, 1),$$

$$(7.10) \qquad \frac{r_{n,\gamma}^2 - E\|W_n + b_n\|_1^2}{\operatorname{sd}\|W_n + b_n\|_1^2} \to z_c.$$

Because W_n is mean-zero Gaussian, we have $E\|W_n + b_n\|_1^2 = E\|W_n\|_1^2 + \|b_n\|_1^2$ and $\operatorname{var}\|W_n + b_n\|_1^2 = \operatorname{var}\|W_n\|_1^2 + 4\operatorname{var}\langle W_n, b_n\rangle_1$. Here $\|W_n\|_1^2 = V_n$ and the distribution of $\langle W_n, b_n\rangle_1$ is zero-mean Gaussian with variance $\langle b_n, n^{-1}AA^T b_n\rangle_1$. With $t_{i,n}$ the eigenvalues of $n^{-1}AA^T$, display (7.10) can be translated in the coefficients $(b_{n,i})$ of b_n relative to the eigenbasis, as

$$(7.11) \qquad \frac{r_{n,\gamma}^2 - EV_n - \sum_i b_{n,i}^2}{\sqrt{\operatorname{var} V_n + 4\sum_i t_{i,n}b_{n,i}^2}} \to z_c.$$

We choose $(b_{n,i})$ differently in the cases that $\beta \le 1 + 2\alpha + 2p$ and $\beta \ge 1 + 2\alpha + 2p$, respectively. In both cases the sequence has exactly one nonzero coordinate. We denote this coordinate by b_{n,i_n}, and set, for numbers d_n to be determined,

$$b_{n,i_n}^2 = r_{n,\gamma}^2 - EV_n - d_n \operatorname{sd} V_n.$$

Because $r_{n,\gamma}^2$, EV_n and $r_{n,\gamma}^2 - EV_n$ are of the same order of magnitude, and sd V_n is of strictly smaller order, for bounded or slowly diverging d_n the right-hand side of the preceding display is equivalent to $(r_{n,\gamma}^2 - EV_n)(1 + o(1))$. Consequently, the left-hand side of (7.11) is equivalent to

$$\frac{d_n \text{ sd } V_n}{\sqrt{\text{var } V_n + 4t_{i_n,n}(r_{n,\gamma}^2 - EV_n)(1 + o(1))}}.$$

The remainder of the argument is different in the two cases.

Case $\beta \leq 1 + 2\alpha + 2p$. We choose $i_n \asymp (n\tau_n^2)^{1/(1+2\alpha+2p)}$. It can be verified that $t_{i_n,n}(r_{n,\gamma}^2 - EV_n)/\text{var } V_n \asymp 1$. Therefore, for $c \in [0, 1]$, there exists a bounded or slowly diverging sequence d_n such that the preceding display tends to z_c.

The bias b_n results from a parameter μ_0^n such that $b_{n,i} = (1 + n\lambda_i\kappa_i^2)^{-1}(\mu_0^n)_i$, for every i. Thus, μ_0^n also has exactly one nonzero coordinate, and this is proportional to the corresponding coordinate of b_n, by the definition of i_n. It follows that

$$i_n^{2\beta}(\mu_0^n)_{i_n}^2 \asymp i_n^{2\beta}b_{n,i_n}^2 \lesssim i_n^{2\beta}(r_{n,\gamma}^2 - EV_n) \asymp 1$$

by the definition of τ_n. It follows that $\|\mu_0^n\|_\beta \lesssim 1$.

Case $\beta \geq 1 + 2\alpha + 2p$. We choose $i_n = 1$. In this case $\tau_n \to 0$ and it can be verified that $t_{i_n,n}(r_{n,\gamma}^2 - EV_n)/\text{var } V_n \to 0$. Also,

$$(\mu_0^n)_1^2 \asymp (1 + n\tau_n^2)^2 b_{n,1}^2 \lesssim (1 + n\tau_n^2)^2 EV_n.$$

This is $O(1)$, because τ_n is chosen so that EV_n is of the same order as the square bias B_n^2, which is $(n\tau_n^2)^{-2}$ in this case.

It remains to prove the asymptotic normality (7.9). We can write

$$\|W_n + b_n\|_1^2 - E\|W_n + b_n\|_1^2 = \sum_i t_{i,n}(Z_i^2 - 1) + 2b_{n,i_n}\sqrt{t_{i_n,n}}Z_{i_n}.$$

The second term is normal by construction. The first term has variance $2\sum_i t_{i,n}^2$. With some effort it can be seen that

$$\sup_i \frac{t_{i,n}^2}{\sum_i t_{i,n}^2} \to 0.$$

Therefore, by a slight adaptation of the Lindeberg–Feller theorem (to infinite sums), we have that $\sum_i t_{i,n}(Z_i^2 - 1)$ divided by its standard deviation tends in distribution to the standard normal distribution. Furthermore, the preceding display shows that this conclusion does not change if the i_nth term is left out from the infinite sum. Thus, the two terms converge jointly to asymptotically independent standard normal variables, if scaled separately by their standard deviations. Then their scaled sum is also asymptotically standard normally distributed.

7.5. *Proof of Theorem* 5.3.   Under (1.1) the variable LAY is $N(LAK\mu_0, t_n^2)$-distributed, for t_n^2 given in (7.5). It follows that the coverage can be written, with W a standard normal variable,

$$(7.12) \qquad \qquad P(|Wt_n + LAK\mu_0 - L\mu_0| \le -s_n z_{\gamma/2}).$$

The bias $LAK\mu_0 - L\mu_0$ and posterior spread s_n^2 are expressed as a series in (7.4) and (7.6).

In the proof of Theorem 5.2 s_n and t_n were seen to have the same order of magnitude, given by the second term in ε_n given in (5.1), with a slowly varying term δ_n as given in the theorem,

$$(7.13) \qquad \qquad s_n \asymp t_n \asymp \tau_n (n\tau_n^2)^{-(1/2+\alpha+q)/(1+2\alpha+2p)} \delta_n.$$

Furthermore, $t_n \le s_n$ for every n, as every term in the infinite series (7.5) is $n\lambda_i \kappa_i^2 / (1 + n\lambda_i \kappa_i^2) \le 1$ times the corresponding term in (7.6).

Because W is centered, the coverage (7.12) is largest if the bias $LAK\mu_0 - L\mu_0$ is zero. It is then at least $1 - \gamma$, because $t_n \le s_n$; remains strictly smaller than 1, because $t_n \asymp s_n$; and tends to exactly $1 - \gamma$ iff $s_n/t_n \to 1$. By Theorem 5.4(i) the latter is impossible if $q < p$. The analysis for nonzero μ_0 depends strongly on the size of the bias relative to t_n.

The supremum of the bias satisfies, for γ_n the slowly varying term given in Theorem 5.2,

$$(7.14) \quad B_n := \sup_{\|\mu_0\|_\beta \lesssim 1} |LAK\mu_0 - L\mu_0| \asymp (n\tau_n^2)^{-((\beta+q)/(1+2\alpha+2p))\wedge 1} \gamma_n.$$

That the left-hand side of (7.14) is smaller than the right-hand side was already shown in the proof of Theorem 5.2, with the help of Lemma 8.2. That this upper bound is sharp follows by considering the sequence μ_0^n defined by, with $\tilde{B}_n$ the right-hand side of the preceding display,

$$\mu_{0,i}^n = \frac{1}{\tilde{B}_n} \frac{i^{-2\beta} l_i}{1 + n\tau_n^2 i^{-1-2\alpha-2p}}.$$

[This is the sequence that gives equality in the application of the Cauchy–Schwarz inequality to derive (7.7).] Using Lemma 8.2, it can be seen that $\|\mu_0^n\|_\beta \lesssim 1$ and that the bias at μ_0^n is of the order $\tilde{B}_n$.

By Lemma 8.3, the bias at a *fixed* $\mu_0 \in S^\beta$ is of strictly smaller order than the supremum B_n if $\beta + q < 1 + 2\alpha + 2p$.

The maximal bias B_n is a decreasing function of the scaling parameter τ_n, while the standard deviation t_n and root-spread s_n increase with τ_n. The scaling rate $\tilde{\tau}_n$ in the statement of the theorem balances B_n with $s_n \asymp t_n$.

Case (i). If $\tau_n \gg \tilde{\tau}_n$, then $B_n \ll t_n$. Hence, the bias $LAK\mu_0 - L\mu_0$ in (7.12) is negligible relative to $t_n \asymp s_n$, uniformly in $\|\mu_0\|_\beta \lesssim 1$, and the coverage is asymptotic to $P(|Wt_n| \le -s_n z_{\gamma/2})$, which is asymptotically strictly between $1 - \gamma$ and 1.

Case (iii). If $\tau_n \ll \tilde{\tau}_n$, then $B_n \gg t_n$. If $b_n = LAK\mu_0^n - L\mu_0^n$ is the bias at a sequence μ_0^n that (nearly) attains the supremum in the definition of B_n, then the coverage at μ_0^n satisfies $P(|Wt_n + b_n| \leq -s_n z_{\gamma/2}) \leq P(|Wt_n| \geq b_n - s_n|z_{\gamma/2}|) \to 0$, as $b_n \asymp B_n \gg s_n$. By the same argument, the coverage also tends to zero for a fixed μ_0 in S^β with bias $b_n = LAK\mu_0 - L\mu_0 \gg t_n$. For this we choose $\mu_{0,i} = i^{-\beta}i^q l_i \tilde{S}(i)$ for a slowly varying function such that $\sum_i \mathcal{S}^2(i)\tilde{S}^2(i)/i < \infty$. The latter condition ensures that $\|\mu_0\|_\beta < \infty$. By another application of Lemma 8.2, the bias at μ_0 is of the order [cf. (7.4)]

$$\sum_i \frac{l_i \mu_{0,i}}{1 + n\tau_n^2 i^{-1-2\alpha-2p}} = \sum_i \frac{(l_i \tilde{S}^{1/2}(i))^2 i^{-\beta+q}}{1 + n\tau_n^2 i^{-1-2\alpha-2p}} \asymp (n\tau_n^2)^{-(\beta+q)/(1+2\alpha+2p)\wedge 1}\tilde{\gamma}_n,$$

where, for $\rho_n = (n\tau_n^2)^{1/(1+2\alpha+2p)}$,

$$\tilde{\gamma}_n^2 = \begin{cases} \mathcal{S}^2(\rho_n)\tilde{S}(\rho_n), & \text{if } \beta + q < 1 + 2\alpha + 2p, \\ \sum_{i \leq \rho_n} \frac{\mathcal{S}^2(i)\tilde{S}(i)}{i}, & \text{if } \beta + q = 1 + 2\alpha + 2p, \\ 1, & \text{if } \beta + q > 1 + 2\alpha + 2p. \end{cases}$$

Therefore, the bias at μ_0 has the same form as the maximal bias B_n; the difference is in the slowly varying factor $\tilde{\gamma}_n$. If $\tau_n \leq \tilde{\tau}_n n^{-\delta}$, then $B_n \gtrsim t_n n^{\delta'}$ for some $\delta' > 0$ and, hence, $b_n \asymp B_n \tilde{\gamma}_n/\gamma_n \gg t_n$.

Case (ii). If $\tau_n \asymp \tilde{\tau}_n$, then $B_n \asymp t_n$. If $b_n = LAK\mu_0^n - L\mu_0^n$ is again the bias at a sequence μ_0^n that nearly assumes the supremum in the definition of B_n, we have that $P(|Wt_n + db_n| \leq -s_n z_{\gamma/2}) \leq P(|Wt_n| \geq db_n - s_n|z_{\gamma/2}|)$ attains an arbitrarily small value if d is chosen sufficiently large. This is the coverage at the sequence $d\mu_0^n$, which is bounded in S^β. On the other hand, the bias at a fixed $\mu_0 \in S^\beta$ is of strictly smaller order than the supremum B_n, and, hence, the coverage at a fixed μ_0 is as in case (i).

If the scaling rate is fixed to $\tau_n \equiv 1$, then it can be checked from (7.13) and (7.14) that $B_n \ll t_n$, $B_n \asymp t_n$ and $B_n \gg t_n$ in the three cases $\alpha < \beta - 1/2$, $\alpha = \beta - 1/2$ and $\alpha > \beta - 1/2$, respectively. In the first and third cases the maximal bias and the spread differ by more than a polynomial term n^δ; in the second case it must be noted that the slowly varying terms γ_n and δ_n are equal [to $\mathcal{S}(\rho_n)$]. It follows that the preceding analysis (i), (ii), (iii) extends to this situation.

7.6. *Proof of Theorem* 5.4. (i). The two quantities s_n and t_n are given as series in (7.6) and (7.5). Every term in the series (7.5) is $n\lambda_i\kappa_i^2/(1 + n\lambda_i\kappa_i^2) \leq 1$ times the corresponding term in the series (7.6). Therefore, $s_n/t_n \to 1$ if and only if the series are determined by the terms for which these numbers are "close to" 1, that is, $n\lambda_i\kappa_i^2$ is large. More precisely, we show below that $s_n/t_n \to 1$ if and only if, for every $c > 0$,

$$(7.15) \qquad \sum_{n\lambda_i\kappa_i^2 \leq c} \frac{l_i^2\lambda_i}{1 + n\lambda_i\kappa_i^2} = o\left(\sum_i \frac{l_i^2\lambda_i}{1 + n\lambda_i\kappa_i^2}\right).$$

If $l \in S^p$, then the series on the left is as in Lemma 8.1 with $q = p$, $u = 1 + 2\alpha + 2p$, $v = 1$, $N = n\tau_n^2$ and $t = 1 + 2\alpha$. Hence, $(t + 2q)/u \geq v$, and the display follows from the final assertion of the lemma. If $l_i = i^{-q-1/2}\mathcal{S}(i)$ for a slowly varying function $\mathcal{S}$, then the series is as in Lemma 8.2, with the same parameters, and by the last statement of the lemma the display is true if and only if $(t + 2q)/u \geq v$, that is, $q \geq p$.

To prove that (7.15) holds iff $s_n/t_n \to 1$, write $s_n^2 = A_n + B_n$, for A_n and B_n the sums over the terms in (7.6) with $n\lambda_i\kappa_i^2 > c$ and $n\lambda_i\kappa_i^2 \leq c$, respectively, and, similarly, $t_n^2 = C_n + D_n$. Then

$$\frac{D_n}{B_n} \leq \frac{c}{1 + c} \leq \frac{C_n}{A_n} \leq 1.$$

It follows that

$$\frac{t_n^2}{s_n^2} = \frac{C_n + D_n}{A_n + B_n} = \frac{C_n/A_n + (D_n/B_n)(B_n/A_n)}{1 + B_n/A_n} \leq \frac{1 + c/(1 + c)(B_n/A_n)}{1 + B_n/A_n}.$$

Because $x \mapsto (1 + rx)/(1 + x)$ is strictly decreasing from 1 at $x = 0$ to $r < 1$ at $x = \infty$ (if $0 < r < 1$), the right-hand side of the equation is asymptotically 1 if and only if $B_n/A_n \to 0$, and otherwise its liminf is strictly smaller. Thus, $t_n/s_n \to 1$ implies that $B_n/A_n \to 0$. Second,

$$\frac{t_n^2}{s_n^2} \geq \frac{C_n}{A_n + B_n} = \frac{C_n/A_n}{1 + B_n/A_n} \geq \frac{c/(1 + c)}{1 + B_n/A_n}.$$

It follows that $\liminf t_n^2/s_n^2 \geq c/(1 + c)$ if $B_n/A_n \to 0$. This being true for every $c > 0$ implies that $t_n/s_n \to 1$.

(i) Second assertion. If $l \in S^p$, then we apply Lemma 8.1 with $q = p$, $t = 1 + 2\alpha$, $u = 1 + 2\alpha + 2p$, $v = 1$ and $N = n\tau_n^2$ to see that $s_n^2 \asymp \tau_n^2(n\tau_n^2)^{-v} = n^{-1}$. Furthermore, the second assertion of the lemma with $(uv - t)/2 = p$ shows that $ns_n^2 \to \|l\|_p^2 = \sum_i l_i^2/\kappa_i^2$ in the case that $\kappa_i = i^{-p}$. The proof can be extended to cover the slightly more general sequence (κ_i) in Assumption 3.1.

If $l \in \mathcal{R}^q$, then we apply Lemma 8.2 with $q = p$, $t = 1 + 2\alpha$, $u = 1 + 2\alpha + 2p$, $v = 1$ and $N = n\tau_n^2$ to see that $s_n^2 \asymp n^{-1}\sum_{i \leq N^{1/u}} \mathcal{S}^2(i)/i$.

(ii) If $l \in S^q$, then the bias is bounded above in (7.7), and in the proof of Theorem 5.1 its supremum B_n over $\|\mu_0\|_\beta \lesssim 1$ is seen to be bounded by $(n\tau_n^2)^{-(\beta+q)/(1+2\alpha+2p)\wedge 1}$, the first term in the definition of ε_n in the statement of this theorem. This upper bound is $o(n^{-1/2})$ iff the stated conditions hold. [Here we use that $\mathcal{S}^2(N) \ll \sum_{i \leq N} \mathcal{S}^2(i)/i$ as $N \to \infty$, as noted in the proof of Lemma 8.2.]

The supremum of the bias B_n in the case that $l \in \mathcal{R}^q$ is given in (7.14). It was already seen to be $o(t_n)$ if $\tau \gg \tilde{\tau}_n$ in the proof of case (i) of Theorem 5.3. If $\tau_n = 1$, we have that $B_n \asymp n^{-(\beta+q)/(1+2\alpha+2p)\wedge 1}\gamma_n$, for γ_n the slowly varying factor given in the statement of Theorem 5.2. Furthermore, we have $s_n \asymp t_n \asymp n^{-1/2}\delta_n$, for

δ_n the slowly varying factor in the same statement. Under the present conditions, $\delta_n \asymp 1$ if $q > p$ and $\delta_n^2 \asymp \sum_{i \leq \rho_n} \mathcal{S}^2(i)/i$ if $q = p$. We can now verify that $B_n = o(t_n)$ if and only if the conditions as stated hold.

(iii) The total variation distance between two Gaussian distributions with the same expectation and standard deviations s_n and t_n tends to zero if and only if $s_n/t_n \to 1$. Similarly, the total distance between two Gaussians with the same standard deviation s_n and means μ_n and ν_n tends to zero if and only if $\mu_n - \nu_n = o(s_n)$. Therefore, it suffices to show that $(LAY - \sum_i Y_i l_i/\kappa_i)/s_n \to 0$ if $l \in S^p$. Because the bias was already seen to be $o(t_n)$ and $s_n \asymp n^{-1/2}$ if $l \in S^p$, it suffices to show that $LAZ - \sum_i Z_i l_i/\kappa_i \to 0$. Under Assumption 3.1 this difference is equal to

$$\sum_i \frac{\kappa_i \lambda_i l_i Z_i}{n^{-1} + \kappa_i^2 \lambda_i} - \sum_i Z_i \frac{l_i}{\kappa_i} = \sum_i \frac{Z_i l_i}{\kappa_i}\left(\frac{1}{1 + n\kappa_i^2 \lambda_i}\right).$$

If $\sum_i l_i^2/\kappa_i^2 < \infty$, then the variance of this expression is seen to tend to zero by dominated convergence.

The final assertion of the theorem follows along the lines of the proof of Theorem 5.3. $\quad\square$

8. Technical lemmas.

LEMMA 8.1. *For any $q \geq 0$, $t \geq -2q$, $u > 0$ and $v \geq 0$, as $N \to \infty$,*

$$\sup_{\|\xi\|_q \leq 1} \sum_i \frac{\xi_i^2 i^{-t}}{(1 + Ni^{-u})^v} \asymp N^{-((t+2q)/u)\wedge v}.$$

Moreover, for every fixed $\xi \in S^q$, as $N \to \infty$,

$$N^{((t+2q)/u)\wedge v} \sum_i \frac{\xi_i^2 i^{-t}}{(1 + Ni^{-u})^v} \to \begin{cases} 0, & \text{if } (t+2q)/u < v, \\ \|\xi\|_{(uv-t)/2}^2, & \text{if } (t+2q)/u \geq v. \end{cases}$$

The last assertion remains true if the sum is limited to the terms $i \leq cN^{1/u}$, for any $c > 0$.

PROOF. In the range $i \leq N^{1/u}$ we have $Ni^{-u} \leq 1 + Ni^{-u} \leq 2Ni^{-u}$, while $1 \leq 1 + Ni^{-u} \leq 2$ in the range $i > N^{1/u}$. Thus, deleting either the first or second term, we obtain

$$\sum_{i \leq N^{1/u}} \frac{\xi_i^2 i^{-t}}{(1 + Ni^{-u})^v} \asymp \sum_{i \leq N^{1/u}} \xi_i^2 i^{2q} \frac{i^{uv-t-2q}}{N^v} \leq \|\xi\|_q^2 N^{-((t+2q)/u)\wedge v},$$

$$\sum_{i > N^{1/u}} \frac{\xi_i^2 i^{-t}}{(1 + Ni^{-u})^v} \asymp \sum_{i > N^{1/u}} \xi_i^2 i^{2q} i^{-t-2q} \leq N^{-(t+2q)/u} \sum_{i > N^{1/u}} \xi_i^2 i^{2q}.$$

The inequality in the first line follows by bounding i in $i^{uv-t-2q}$ by $N^{1/u}$ if $uv - t - 2q > 0$, and by 1 otherwise. This proves the upper bound for the supremum.

The lower bound follows by considering the two sequences (ξ_i) given by $\xi_i = i^{-q}$ for $i \sim N^{1/u}$ and $\xi_i = 0$ otherwise (showing that the supremum is bigger than $N^{-(t+2q)/u}$), and given by $\xi_1 = 1$ and $\xi_i = 0$ otherwise (showing that the supremum is bigger than N^{-v}).

The second line of the preceding display shows that the sum over the terms $i > N^{1/u}$ is $o(N^{-(t+2q)/u})$. Furthermore, the first line can be multiplied by $N^{(t+2q)/u}$ to obtain

$$N^{(t+2q)/u} \sum_{i \leq N^{1/u}} \frac{\xi_i^2 i^{-t}}{(1 + Ni^{-u})^v} \asymp \sum_{i \leq N^{1/u}} \xi_i^2 i^{2q} \left(\frac{i}{N^{1/u}} \right)^{uv-t-2q}.$$

If $(t + 2q)/u < v$, then $uv - t - 2q > 0$ and this tends to zero by dominated convergence. Also,

$$N^v \sum_i \frac{\xi_i^2 i^{-t}}{(1 + Ni^{-u})^v} = \sum_i \xi_i^2 i^{uv-t} \left(\frac{Ni^{-u}}{1 + Ni^{-u}} \right)^v.$$

If $(t + 2q)/u \geq v$, then $q \geq (uv - t)/2$ and, hence, $\xi \in S^{(uv-t)/2}$, and the right-hand side tends to $\sum_i \xi_i^2 i^{uv-t}$ by dominated convergence.

The final assertion needs to be proved only in the case that $(t + 2q)/u \geq v$, as in the other case the whole sum tends to 0. The sum over the terms $i > N^{1/u}$ was seen to be always $o(N^{-(t+2q)/u})$, which is $o(N^{-v})$ if $(t + 2q)/u \geq v$. The final assertion for $c = 1$ follows, because the sum over the terms $i \leq N^{1/u}$ was seen to have the exact order N^{-v} (if $\xi \neq 0$). For general c the proof is analogous, or follows by scaling N. □

LEMMA 8.2. *For any $t, v \geq 0$, $u > 0$, and (ξ_i) such that $|\xi_i| = i^{-q-1/2}S(i)$ for $q > -t/2$ and a slowly varying function $S : (0, \infty) \to (0, \infty)$, as $N \to \infty$,*

$$\sum_i \frac{\xi_i^2 i^{-t}}{(1 + Ni^{-u})^v} \asymp \begin{cases} N^{-(t+2q)/u} S^2(N^{1/u}), & \text{if } (t + 2q)/u < v, \\ N^{-v} \sum_{i \leq N^{1/u}} S^2(i)/i, & \text{if } (t + 2q)/u = v, \\ N^{-v}, & \text{if } (t + 2q)/u > v. \end{cases}$$

Moreover, for every $c > 0$, the sum on the left is asymptotically equivalent to the same sum restricted to the terms $i \leq cN^{1/u}$ if and only if $(t + 2q)/u \geq v$.

PROOF. As in the proof of the preceding lemma, we split the infinite series in the sum over the terms $i \leq N^{1/u}$ and $i > N^{1/u}$. For the first part of the series

$$\sum_{i \leq N^{1/u}} \frac{\xi_i^2 i^{-t}}{(1 + Ni^{-u})^v} \asymp \sum_{i \leq N^{1/u}} S(i)^2 \frac{i^{uv-t-2q-1}}{N^v}.$$

If $uv - t - 2q > 0$ [i.e., $(t + 2q)/u < v$], the right-hand side is of the order $N^{-(t+2q)/u}\mathcal{S}^2(N^{1/u})$, by Theorem 1(b) on page 281 in [10], while if $uv - t - 2q < 0$, it is of the order N^{-v} by Lemma on page 280 in [10]. Finally, if $uv - t - 2q = 0$, then the right-hand side is identical to $N^{-v}\sum_{i \le N^{1/u}}\mathcal{S}^2(i)/i$.

The other part of the infinite series satisfies, by Theorem 1(a) on page 281 in [10],

$$\sum_{i > N^{1/u}} \frac{\xi_i^2 i^{-t}}{(1 + Ni^{-u})^v} \asymp \sum_{i > N^{1/u}} \mathcal{S}(i)^2 i^{-t-2q-1} \asymp N^{-(t+2q)/u}\mathcal{S}^2(N^{1/u}).$$

This is never bigger than the contribution of the first part of the sum, and of equal order if $(t + 2q)/u < v$. If $(t + 2q)/u > v$, then the leading polynomial term is strictly smaller than N^{-v}. If $(t + 2q)/u = v$, then the leading term is equal to N^{-v}, but the slowly varying part satisfies $\mathcal{S}^2(N^{1/u}) \ll \sum_{i \le N^{1/u}}\mathcal{S}^2(i)/i$, by Theorem 1(b) on page 281 in [10]. Therefore, in both cases the preceding display is negligible relative to the first part of the sum. This proves the final assertion of the lemma for $c = 1$. The proof for general $c > 0$ is analogous. $\square$

By the Cauchy–Schwarz inequality, for any $\mu \in S^{t/2}$,

$$\left| \sum_i \frac{\xi_i \mu_i}{1 + Ni^{-u}} \right|^2 \le \|\mu\|_{t/2}^2 \sum_i \frac{\xi_i^2 i^{-t}}{(1 + Ni^{-u})^2}.$$

The preceding lemma gives the exact order of the right-hand side. The application of the Cauchy–Schwarz inequality is sharp, in that there is equality for some $\mu \in S^{t/2}$. However, this μ depends on N. For fixed $\mu \in S^{t/2}$ the left-hand side is strictly smaller than the right-hand side.

LEMMA 8.3. *For any $t, u \ge 0$, $\mu \in S^{t/2}$ and (ξ_i) such that $|\xi_i| = i^{-q-1/2}\mathcal{S}(i)$ for $0 < t + 2q < 2u$ and a slowly varying function $\mathcal{S} : (0, \infty) \to (0, \infty)$, as $N \to \infty$,*

$$\sum_i \frac{|\xi_i \mu_i|}{1 + Ni^{-u}} \ll N^{-(t+2q)/(2u)}\mathcal{S}(N^{1/u}).$$

PROOF. We split the series in two parts, and bound the denominator $1 + Ni^{-u}$ by Ni^{-u} or 1. By the Cauchy–Schwarz inequality, for any $r > 0$,

$$\left| \sum_{i \le N^{1/u}} \frac{|\xi_i \mu_i|}{Ni^{-u}} \right|^2 \le \frac{1}{N^2} \sum_{i \le N^{1/u}} \frac{\mathcal{S}^2(i)i^r}{i} \sum_{i \le N^{1/u}} \mu_i^2 i^{2u-2q-r}$$

$$\asymp \frac{1}{N^2}\mathcal{S}^2(N^{1/u})N^{r/u}$$

$$\times \sum_{i \le N^{1/u}} \mu_i^2 i^t \left(\frac{i}{N^{1/u}} \right)^{2u-2q-r-t} N^{(2u-2q-r-t)/u},$$

$$\left| \sum_{i > N^{1/u}} \frac{|\xi_i \mu_i|}{1} \right|^2 \le \sum_{i > N^{1/u}} \frac{\mathcal{S}^2(i)}{i} i^{-2q} \sum_{i > N^{1/u}} \mu_i^2 \asymp \mathcal{S}^2(N^{1/u}) N^{-2q/u} \sum_{i > N^{1/u}} \mu_i^2.$$

The terms in the remaining series in the right-hand side of the first inequality are bounded by $\mu_i^2 i^t$ and tend to zero pointwise as $N \to \infty$ if $2u - 2q - r - t > 0$. If $t + 2q < 2u$, then there exists $r > 0$ such that the latter is true, and for this r the sum tends to zero by the dominated convergence theorem. The other terms collect to $N^{-(t+2q)/(u)} \mathcal{S}^2(N^{1/u})$. The sum in the right-hand side of the second inequality is bounded by $\sum_{i > N^{1/u}} \mu_i^2 i^t N^{-t/u} = o(N^{-t/u})$. $\quad\square$

REFERENCES

[1] BERTERO, M. (1986). Regularization methods for linear inverse problems. In *Inverse Problems* (*Montecatini Terme*, 1986). *Lecture Notes in Math.* **1225** 52–112. Springer, Berlin. MR0885511

[2] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

[3] BISSANTZ, N., HOHAGE, T., MUNK, A. and RUYMGAART, F. (2007). Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.* **45** 2610–2636 (electronic). MR2361904

[4] BONTEMPS, D. (2011). Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *Ann. Statist.* **39** 2557–2584.

[5] CASTILLO, I. (2011). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields*. To appear. DOI:10.1007/s00440-010-0316-5.

[6] CAVALIER, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems* **24** 034004, 19. MR2421941

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

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

[9] DONOHO, D. L. and LOW, M. G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.* **20** 944–970. MR1165601

[10] FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications. Vol. II*, 2nd ed. Wiley, New York. MR0270403

[11] FLORENS, J. and SIMONI, A. (2011). Regularizing priors for linear inverse problems. Preprint.

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

[13] GHOSAL, S. (1999). Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli* **5** 315–331. MR1681701

[14] GHOSAL, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.* **74** 49–68. MR1790613

[15] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[16] GOLDENSHLUGER, A. and PEREVERZEV, S. V. (2003). On adaptive inverse estimation of linear functionals in Hilbert scales. *Bernoulli* **9** 783–807. MR2047686

[17] HALMOS, P. R. (1967). *A Hilbert Space Problem Book*. Van Nostrand, Princeton. MR0208368

[18] LEAHU, H. (2011). On the Bernstein–von Mises phenomenon in the Gaussian white noise model. *Electron. J. Stat.* **5** 373–404.

[19] MAIR, B. A. and RUYMGAART, F. H. (1996). Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.* **56** 1424–1444. MR1409127

[20] MANDELBAUM, A. (1984). Linear estimators and measurable linear transformations on a Hilbert space. *Z. Wahrsch. Verw. Gebiete* **65** 385–397. MR0731228

[21] MATHÉ, P. and PEREVERZEV, S. V. (2006). Regularization of some linear ill-posed problems with discretized random noisy data. *Math. Comp.* **75** 1913–1929 (electronic). MR2240642

[22] NASHED, M. Z. and WAHBA, G. (1974). Generalized inverses in reproducing kernel spaces: An approach to regularization of linear operator equations. *SIAM J. Math. Anal.* **5** 974–987. MR0358405

[23] RUDIN, W. (1991). *Functional Analysis*, 2nd ed. McGraw-Hill, New York. MR1157815

[24] RUYMGAART, F. H. (1993). A unified approach to inversion problems in statistics. *Math. Methods Statist.* **2** 130–146. MR1257980

[25] SKOROHOD, A. V. (1974). *Integration in Hilbert Space. Ergebnisse der Mathematik und Ihrer Grenzgebiete* **79**. Springer, New York. Translated from the Russian by Kenneth Wickwire. MR0466482

[26] STUART, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numer.* **19** 451–559. MR2652785

[27] TIKHONOV, A. N. and ARSENIN, V. Y. (1977). *Solutions of Ill-posed Problems*. V. H. Winston and Sons, Washington, DC. Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics. MR0455365

[28] VAN DER VAART, A. (1991). On differentiable functionals. *Ann. Statist.* **19** 178–204. MR1091845

[29] VAN DER VAART, A. and VAN ZANTEN, H. (2007). Bayesian inference with rescaled Gaussian process priors. *Electron. J. Stat.* **1** 433–448 (electronic). MR2357712

[30] VAN DER VAART, A. W. (1988). *Statistical Estimation in Large Parameter Spaces. CWI Tract* **44**. Stichting Mathematisch Centrum Centrum voor Wiskunde en Informatica, Amsterdam. MR0927725

[31] VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

[32] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics*: *Contributions in Honor of Jayanta K. Ghosh. Inst. Math. Stat. Collect.* **3** 200–222. IMS, Beachwood, OH. MR2459226

[33] VAN ROOIJ, A. C. M. and RUYMGAART, F. H. (2001). Abstract inverse estimation with application to deconvolution on locally compact abelian groups. *Ann. Inst. Statist. Math.* **53** 781–798. MR1880812

B. T. KNAPIK
A. W. VAN DER VAART
DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCES
VU UNIVERSITY AMSTERDAM
DE BOELELAAN 1081A
1081 HV AMSTERDAM
THE NETHERLANDS
E-MAIL: b.t.knapik@vu.nl
          aad@cs.vu.nl

J. H. VAN ZANTEN
DEPARTMENT OF MATHEMATICS
EINDHOVEN UNIVERSITY OF TECHNOLOGY
EINDHOVEN
THE NETHERLANDS
E-MAIL: j.h.v.zanten@tue.nl