Bayesian Linear Inverse Problems in Regularity Scales with Discrete Observations

Yan, Dong; Gugushvili, Shota; van der Vaart, Aad

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Bayesian Linear Inverse Problems in Regularity Scales with Discrete Observations

Dong Yan and Aad van der Vaart
TU Delft, Delft, Netherlands
Shota Gugushvili
Wageningen University & Research, Wageningen, The Netherlands

Abstract

We obtain rates of contraction of posterior distributions in inverse problems with discrete observations. In a general setting of smoothness scales we derive abstract results for general priors, with contraction rates determined by discrete Galerkin approximation. The rate depends on the amount of prior concentration near the true function and the prior mass of functions with inferior Galerkin approximation. We apply the general result to non-conjugate series priors, showing that these priors give near optimal and adaptive recovery in some generality, Gaussian priors, and mixtures of Gaussian priors, where the latter are also shown to be near optimal and adaptive.

AMS (2000) subject classification. 62G20; 35R30.
Keywords and phrases. Adaptive estimation, Gaussian prior, Hilbert scale, Linear inverse problem, Nonparametric Bayesian estimation, Posterior contraction rate, Random series prior, Regularity scale, Regression, Fixed design, Interpolation, Galerkin.

1   Introduction

In an inverse problem one observes a noisy version of a transformed signal Af and wishes to recover the unknown signal f. The *forward operator A* is assumed to be injective, but its inverse is discontinuous, so that a naive inversion would amplify the noise. In this paper we study a Bayesian method for regularising the inverse in a setting with discrete observations. We assume that the transformed signal is a function $Af : \mathcal{D} \to \mathbb{R}$ on a bounded domain $\mathcal{D} \subset \mathbb{R}^d$, and for given *design points* $\{x_1, \ldots, x_n\} \subset \mathcal{D}$, we observe $Y^n = (Y_1, \ldots, Y_n)$ defined by

$$Y_i = Af(x_i) + Z_i, \qquad i = 1, \cdots, n, \tag{1}$$

with Z_i i.i.d. standard normal random variables. This *regression model* (1) may be compared to its continuous version, the *white noise model*, in which the observation is given by

$$Y^{(n)} = Af + \frac{1}{\sqrt{n}}\xi, \qquad (2)$$

where ξ is Gaussian (white) noise. There is a rich literature on the continuous problem, even though the regression problem is more realistic in practice.

The review papers Cavalier (2008), Stuart (2010) provide an overview to the field of inverse problems up to the time around 2010. The former paper focuses on frequentist studies, and the latter one concentrates on Bayesian methods, as does the present paper. The book Nickl (2023) gives a recent overview, including nonlinear problems. Bayesian methods for the white noise problem (2) were studied in among others Agapiou et al. (2013), Gugushvili et al. (2020), Knapik and Salomond (2018), Knapik et al. (2011, 2013), but the regression model is less explored, except in the case that A is identity (e.g. Sniekers and van der Vaart (2020, 2015)). In Gugushvili et al. (2018) Gaussian conjugacy was used to study a problem with one-dimensional $\mathcal{D}$.

The Bayesian approach consists of putting a *prior* probability measure on f, and, after collecting the data, updating this to the *posterior* probability measure. If $y \mapsto p_f^{(n)}(y)$ denotes the (Gaussian) density of the data $Y^n = (Y_1, \ldots, Y_n)$, then this is the Borel measure on H, given, by Bayes's formula, as

$$\Pi_n(f \in B \mid Y^n) = \frac{\int_B p_f^{(n)}(Y^n)\, d\Pi_n(f)}{\int p_f^{(n)}(Y^n)\, d\Pi_n(f)}. \qquad (3)$$

In the Bayesian paradigm the posterior distribution encompasses all the necessary information for inference on f. An attractive feature of the Bayesian approach is that it not only offers an estimation procedure, through a measure of 'centre' of the posterior distribution, but also provides a way to conduct uncertainty quantification, through the spread of the posterior distribution.

If one thinks of the Bayesian approach as a method and not merely as a philosophy, then one wants the posterior measure (3) to recover the function f_0 if in reality $Y_1, \ldots, Y_n$ are generated according to the model (1) with $f = f_0$ and $n \to \infty$. We shall be interested in the *rate of posterior contraction*. Following Ghosal et al. (2000), Ghosal and van der Vaart (2007, 2017), we say that a sequence $\varepsilon_n \downarrow 0$ is a rate of posterior contraction to f_0

if, for a fixed sufficiently large constant M, and $n \to \infty$,

$$\Pi_n\big(f : \|f - f_0\| > M\varepsilon_n \mid Y^n\big) \overset{\mathbb{P}_{f_0}^{(n)}}{\to} 0. \tag{4}$$

We are interested in the fastest rate for which this is satisfied, and compare this to the rate of estimation of non-Bayesian methods.

The paper is organised as follows. Section 2 introduces our setup along with the assumptions, and introduces an interpolation device. Section 3 presents two general theorems, which are applied to two special cases, series priors and (mixtures of) Gaussian priors, in Sections 4 and 5. Section 6 contains some of the proofs.

The paper extends results for the white noise model obtained in Gugushvili et al. (2020) to the case of discrete observations. While we repeat necessary definitions, we refer to the latter paper for further examples and discussion.

1.1 Notation The symbols $\lesssim, \gtrsim, \simeq$ mean $\leq, \geq, =$ up to a positive multiple independent of n, (or another asymptotic parameter). The constant may be stated explicitly in subscripts, and e.g. $\lesssim_f$ means that it depends on f.

2 Setup

In this section we formulate the assumptions on the operator A and present the interpolation device to handle the discrete observational scheme. The first part of the section borrows from Gugushvili et al. (2020).

2.1 Smoothing The function f in (1) is an element of a Hilbert space $(H, \langle \cdot, \cdot \rangle)$. We embed this space as the space $H = H_0$ in a 'scale of smoothness classes', as is common in the literature on inverse problems, as follows.

Definition 2.1 (Smoothness scale). For every $s \in \mathbb{R}$ the space H_s is an infinite-dimensional, separable Hilbert space, with inner product $\langle \cdot, \cdot \rangle_s$ and induced norm $\| \cdot \|_s$. The spaces $(H_s)_{s \in \mathbb{R}}$ satisfy the following conditions:

1. For $s < t$ the space H_t is a dense subspace of H_s and $\|f\|_s \lesssim \|f\|_t$, for $f \in H_t$.

2. For $s \geq 0$ and $f \in H_0$ viewed as element of $H_{-s} \supset H_0$,

$$\|f\|_{-s} = \sup_{\|g\|_s \leq 1} \langle f, g \rangle, \qquad f \in H_0. \tag{5}$$

Assumption 2.2 *(Approximation). For every $j \in \mathbb{N}$ and $s \in (0, S)$, for some $S > 0$, there exists a $(j-1)$-dimensional linear subspace $V_j \subset H_0$ and a number $\delta(j, s)$ such that $\delta(j, s) \to 0$ as $j \to \infty$, and such that*

$$\inf_{g \in V_j} \|f - g\| \lesssim \delta(j, s) \|f\|_s, \tag{6}$$

$$\|g\|_s \lesssim \frac{1}{\delta(j, s)} \|g\|, \qquad \forall g \in V_j. \tag{7}$$

The two inequalities (6) and (7) are known as inequalities of Jackson and Bernstein type, respectively, see, e.g., Canuto et al. (2010). The approximation property (6) shows that 'smooth elements' $f \in H_s$ are well approximated in $\|\cdot\|$ by their projection onto a finite-dimensional space V_j, with approximation error tending to zero as the dimension of V_j tends to infinity. We shall focus on the case of polynomial dependence $\delta(j, s) \simeq j^{-s/d}$, where d is the dimension of the domain of the functions in H_0.

The Sobolev spaces $H_s(\mathcal{D}) = W^{s,2}(\mathcal{D})$ provide perhaps the most important example of a smoothness scale (see e.g. Haroske and Triebel (2008), Triebel (2008, 2010)), which can also be restricted by boundary conditions, as common in the theory of differential equations. The following example gives a smoothness scale described through basis functions. Because one can always choose an appropriate basis, for our purposes not much is lost by considering only this concrete example.

Example 2.3 *(Sequence spaces). Suppose $(\phi_i)_{i \in \mathbb{N}}$ is a given orthonormal sequence in the Hilbert H, and $1 \leq b_i \uparrow \infty$ is a given sequence of numbers. For $s \geq 0$, define H_s as the set of all elements $f = \sum_{i \in \mathbb{N}} f_i \phi_i \in H$ with $\sum_{i \in \mathbb{N}} b_i^{2s} f_i^2 < \infty$, equipped with the norm*

$$\|f\|_s = \left(\sum_{i \in \mathbb{N}} b_i^{2s} f_i^2 \right)^{1/2}.$$

For $s < 0$, equip the elements $f = \sum_{i=1}^{\infty} f_i \phi_i$ of H, where $(f_i) \in \ell_2$, with the norm as in the display, which is now automatically finite, and next define H_s as the completion of H under this norm.

It can be checked that these spaces form a smoothness scale as in Definition 2.1 and that Assumption 2.2 is satisfied with the spaces $V_j = \mathsf{Span}(\phi_i : i < j)$, and the approximation numbers $\delta(j, s) = b_j^{-s}$.

The forward operator A in the model (1) is assumed to be a bounded linear operator $A : H \to G$ between the separable Hilbert spaces H and G, which is 'smoothing'. The following assumption, which is standard in the

literature (see Cohen et al. (2004), Goldenshluger and Pereverzev (2003), Mathé and Pereverzev (2001), Natterer (1984)) makes this precise.

Assumption 2.4 *(Smoothing property of A). For some $\gamma > 0$, the operator $A : H_{-\gamma} \to G$ is injective and bounded and, for every $f \in H_0$,*

$$\|Af\| \simeq \|f\|_{-\gamma}. \tag{8}$$

While the preceding assumption is sufficient for our main results, a stronger smoothing property is often satisfied and may be helpful when analysing particular priors. We now assume that both $H = H_0$ and $G = G_0$ are embedded are embedded in smoothness scales $(H_s)_{s\in\mathbb{R}}$ and $(G_s)_{s\in\mathbb{R}}$ as in Definition 2.1. Somewhat abusing notation, we write the norms of both scales as $\|\cdot\|_s$.

Assumption 2.5 *(Strengthened smoothing property of A). For some $\gamma > 0$, the operator $A : H_{s-\gamma} \to G_s$ is injective and bounded for every $s \geq 0$, and, for every $f \in H_0 \cap H_{s-\gamma}$,*

$$\|Af\|_s \simeq \|f\|_{s-\gamma}. \tag{9}$$

Example 2.6 *(SVD). If the operator $A : H \to G$ is compact, then the positive self-adjoint operator $A^*A : H \to H$ possesses a countable orthonormal basis of eigenfunctions ϕ_i, which can be arranged so that the corresponding sequence of eigenvalues λ_i decreases to zero. If A is injective, then all eigenvalues, whose roots are known as the* singular values *of A, are strictly positive. Suppose that there exists $\gamma > 0$ such that they are of order*

$$\lambda_i \simeq i^{-2\gamma}.$$

If we construct the smoothness classes $(H_s)_{s\in\mathbb{R}}$ from the basis $(\phi_i)_{i\in\mathbb{N}}$ and the numbers $b_i = i$ as in Example 2.3, then (8) is satisfied, since $\|Af\|^2 = \langle A^T A f, f \rangle = \sum_i \lambda_i^2 f_i^2 \simeq \|f\|_{-\gamma}^2$.

If we also construct smoothness classes $(G_s)_{s\in\mathbb{R}}$ from the orthonormal sequence $U\phi_i$ in G, where $U : \mathrm{Range}(A) \to G$ is the partial isometry in the polar decomposition $Af = U(A^*A)^{1/2}f$ of A, then (9) will also be satisfied, for every $s \geq 0$.

Concrete examples of operators A satisfying Assumptions 2.4 or 2.5 can be found in Gugushvili et al. (2020).

    2.2 Interpolation The sampling scheme (1) collects discrete data, but the operator A acts on a continuous function, which we wish to estimate on its full domain. In this section we describe an interpolation technique

that maps discrete signals to the continuous domain. A similar technique has been used in the context of proving asymptotic equivalence of the white noise model and nonparametric regression; see Reiß (2008) and the references therein. The assumptions are also inspired by the theory from the field of numerical analysis, see Canuto et al. (2010).

The range space G of the operator $A : H \to G$ is a collection of functions $g : \mathcal{D} \to \mathbb{R}$, equipped with a pre-inner product $\langle \cdot, \cdot \rangle$. The design points give rise to a "discrete" semi-inner product and semi-norm on G given by

$$\langle g, h \rangle_n := \frac{1}{n} \sum_{i=1}^{n} g(x_i) h(x_i) \qquad \|g\|_n := \sqrt{\frac{1}{n} \sum_{i=1}^{n} g^2(x_i)}.$$

(The notation $\| \cdot \|_n$ clashes with the notation $\| \cdot \|_s$ for the norms of the smoothness scales, but this should not lead to confusion as n will never appear as smoothness level.)

For every $n \in \mathbb{N}$, we fix an n-dimensional subspace $\widetilde{W}_n \subset G$, and, for $g \in G$, denote by $\mathcal{I}_n g$ the element of $\widetilde{W}_n$ that interpolates g at the design points, i.e.

$$\mathcal{I}_n g(x_i) = g(x_i), \qquad i = 1, \ldots, n.$$

This defines a map $\mathcal{I}_n : G \to \widetilde{W}_n \subset G$. Existence and uniqueness of this map is addressed in Lemma 2.9 below. (In order that the point evaluations $g(x_i)$ be well defined, the domain G may be restricted to continuous functions.)

Assumption 2.7 (*Interpolation*). *For every $n \in \mathbb{N}$ and $s \in (s_d, S_d)$, for some $S_d > s_d \geq 0$, there exist an n-dimensional subspace $\widetilde{W}_n \subset G$ and a number $\delta_d(n, s)$ such that*

$$\|w\|_n \simeq \|w\|, \qquad \forall w \in \widetilde{W}_n, \tag{10}$$

$$\|\mathcal{I}_n g - g\| \lesssim \delta_d(n, s) \|g\|_s. \tag{11}$$

In this assumption $\| \cdot \|_s$ are the norms of a smoothness scale $(G_s)_{s \in \mathbb{R}}$ as in Definition 2.1, in which the space G is embedded as $G = G_0$ The approximation numbers $\delta_d(n, s)$ will often be the same as the approximation rates $\delta(n, s)$ in Assumption 2.2. However, the approximation (11) is typically not true for every smoothness level $s > 0$, but only for s in a range (s_d, S_d). For instance, for Sobolev scales the lower bound s_d is typically equal to $d/2$ for d the dimension of the domain $\mathcal{D}$ of the functions in G, and the upper bound S_d is the regularity of the basis elements used to define the scale.

Fix an arbitrary orthonormal basis $e_{1,n}, \ldots, e_{n,n}$ of $\widetilde{W}_n$ relative to the discrete inner product $\langle \cdot, \cdot \rangle_n$, and given the discrete data $(Y_1, \ldots, Y_n)$ as in (1), define

$$Y^{(n)} = \sum_{i=1}^{n} \frac{1}{n} \sum_{j=1}^{n} Y_j e_{i,n}(x_j)\, e_{i,n}. \tag{12}$$

This embeds the discrete data as a 'continuous signal' into the space $\widetilde{W}_n \subset G$. If the observations satisfy (1), then the continuous observation $Y^{(n)}$ can be decomposed as

$$Y^{(n)} = \sum_{i=1}^{n} \langle Af, e_{i,n} \rangle_n e_{i,n} + \frac{1}{n} \sum_{j=1}^{n} Z_j \sum_{i=1}^{n} e_{i,n}(x_j)\, e_{i,n} = \mathcal{I}_n Af + \frac{1}{\sqrt{n}} \xi^{(n)},$$

where $\xi^{(n)}$ is a Gaussian random variable with values in the space in $\widetilde{W}_n$. Loosely speaking, as $n \to \infty$ the operators $\mathcal{I}_n$ should tend to the identity operator, and the mean of the signal $Y^{(n)}$ should become more representative of the full signal Af. As shown in the following lemma, the noise $\xi^{(n)}$ tends to white noise in the Hilbert space G.

Lemma 2.8 *The variable $\xi^{(n)}$ defined in the preceding display is a Gaussian random element in $\widetilde{W}_n \subset G$ with mean zero. Under (10) its covariance operator is up to multiplicative constants that do not depend on n bounded below and above by the orthogonal projection $\tilde{Q}_n : G \to \widetilde{W}_n$ relative to the continuous inner product $\langle \cdot, \cdot \rangle$.*

Proof For $g \in G$, we can write $\langle \xi^{(n)}, g \rangle = n^{-1/2} \sum_{j=1}^{n} Z_j \sum_{i=1}^{n} e_{i,n}(x_j) \langle e_{i,n}, g \rangle$. The expectation of this variable vanishes, while the variance is given by

$$\mathrm{var}\langle \xi^{(n)}, g \rangle = \frac{1}{n} \sum_{j=1}^{n} \left(\sum_{i=1}^{n} e_{i,n}(x_j) \langle e_{i,n}, g \rangle \right)^2 = \left\| \sum_{i=1}^{n} \langle e_{i,n}, g \rangle e_{i,n} \right\|_n^2 = \sum_{i=1}^{n} \langle e_{i,n}, g \rangle^2.$$

by the orthogonality of the basis.

The orthogonal projection $\tilde{Q}_n g$ of g onto $\widetilde{W}_n$ relative to the continuous inner product is equal to $\tilde{Q}_n g = \sum_{i=1}^{n} \alpha_i e_{i,n}$, for $\alpha = (\alpha_1, \ldots, \alpha_n)^T = \Sigma_n^{-1} \beta$, where $\beta = (\langle e_{i,n}, g \rangle)_{i=1..n}$ and Σ_n is the Gram matrix $(\langle e_{i,n}, e_{j,n} \rangle)$. Hence $\|\tilde{Q}_n g\|^2 = \alpha^T \Sigma_n \alpha = \beta^T \Sigma_n^{-1} \beta$. Because Σ_n is bounded above and below by a multiple of the identity, by Lemma 2.9 below, it follows that $\|\tilde{Q}_n g\|^2 \simeq \beta^T \beta = \sum_{i=1}^{n} \langle e_{i,n}, g \rangle^2$.

Lemma 2.9 *Suppose that $\widetilde{W}_n \subset G$ is a finite-dimensional subspace so that* (10) *holds. Then for every $g \in G$ there exists a function $\mathcal{I}_n g \in \widetilde{W}_n$ that interpolates g at the points $x_1, \ldots, x_n$ and*

- $\mathcal{I}_n g$ *is the orthogonal projection of g onto $\widetilde{W}_n \subset G$ relative to $\langle \cdot, \cdot \rangle_n$.*
- $\|g\|_n \simeq \|\mathcal{I}_n g\|$, *for every $g \in G$.*

Furthermore, for any orthonormal basis $e_{1,n}, \ldots, e_{n,n}$ of $\widetilde{W}_n$ relative to $\langle \cdot, \cdot \rangle_n$, the Gram matrix $(\langle e_{i,n}, e_{j,n} \rangle)_{i,j=1..n}$ is bounded above and below by the identity matrix times a constant that does not depend on n.

Proof Under (10), the linear map $R : \widetilde{W}_n \subset G \to \mathbb{R}^n$ given by $(Rw)_i = w(x_i)$, for $i = 1, \ldots, n$, has kernel 0 and hence is a bijection. Thus there is a solution $w = \mathcal{I}_n g \in \widetilde{W}_n$ to $Rw = \big(g(x_i)\big)_{i=1..n}$, for every $g \in G$, and this is unique.

The function $\mathcal{I}_n g$ is contained in $\widetilde{W}_n$ by definition, and satisfies $\langle g - \mathcal{I}_n g, h \rangle_n = n^{-1} \sum_{i=1}^n g(x_i) h(x_i) - n^{-1} \sum_{i=1}^n \mathcal{I}_n g(x_i) h(x_i) = 0$, for every $h \in \widetilde{W}_n$. Thus $\mathcal{I}_n g$ is the orthogonal projection of g onto $\widetilde{W}_n$ relative to $\langle \cdot, \cdot \rangle_n$. Because the functions g and $\mathcal{I}_n g$ coincide at the design points, we have $\|g\|_n = \|\mathcal{I}_n g\|_n$, which is equivalent to $\|\mathcal{I}_n g\|$, by (10).

The final statement is the equivalence $\|\sum_{i=1}^n \alpha_i e_{i,n}\|^2 \simeq \alpha^T \alpha$, for $\alpha \in \mathbb{R}^n$. Now $\|\sum_{i=1}^n \alpha_i e_{i,n}\| \simeq \|\sum_{i=1}^n \alpha_i e_{i,n}\|_n$, by (10), and this is equal to the Euclidean norm of α if $e_{1,n}, \ldots, e_{n,n}$ is orthonormal relative to $\langle \cdot, \cdot \rangle_n$.

The following two examples exhibit suitable discretisation spaces, both with equidistant design points.

Example 2.10 (*Trigonometric Polynomials*). *This example is adapted from Section 2.3 in Reiß (2008). Let $\mathcal{D} = \mathbb{T}^d = (0,1]^d$, for $d \in \mathbb{N}$, and consider the set of $n = m^d$ design points $\{k/m\}_{k \in \{1, \cdots, m\}^d}$, for a given odd natural number m. In this case, the Fourier system*

$$e_k(x) = e^{i2\pi \langle k, x \rangle_{\mathbb{R}^d}}, \qquad k = (k_1, \cdots, k_d) \in \mathbb{Z}^d,$$

is not only orthonormal in the continuous space $L^2(\mathbb{T}^d)$, but ('modulo m') also with respect to the discrete inner product $\langle \cdot, \cdot \rangle_n$, i.e.

$$\langle e_j, e_k \rangle_n = \begin{cases} 1, & \text{if } j_l \equiv k_l \mod m, \quad \forall l \in \{1, \cdots, d\}, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The scale of isotropic Sobolev spaces $H_s(\mathbb{T}^d)$ is defined in terms of the Fourier coefficients $f_k = \int_{\mathbb{T}^d} f(x) e_k(x) \, dx$ of functions $f \in L^2(\mathbb{T}^d)$, as (for $|k|$ any

norm on $\mathbb{R}^d$)

$$H_s(\mathbb{T}^d) := \left\{ f \in L^2(\mathbb{T}^d) : \ \|f\|_{H_s}^2 := \sum_{k \in \mathbb{Z}^d} (1 + |k|)^{2s} |f_k|^2 < \infty \right\}.$$

For smoothness levels $s \in \mathbb{N}$, this norm is equivalent to the canonical Sobolev norm $\sum_{|l|_1 \le s} \|D^l f\|_{L^2(\mathbb{T}^d)}$.

The spaces V_j obtained as the linear span of the basis elements, ordered suitably, satisfy Assumption 2.2. Due to (13), the space $\widetilde{W}_n = \mathsf{Span}\{e_k : |k|_\infty \le (m-1)/2\}$ satisfies (10) with $C_1 = C_2 = 1$.

As noted in Reiß (2008), the following estimates hold for $f \in H_s(\mathbb{T}^D)$,

$$\begin{aligned}
\|f - Q_n f\|_{L^2} &\lesssim n^{-s/d} \|f\|_{H_s}, \\
\|Q_n f - \mathcal{I}_n f\| &\lesssim_d n^{-s/d} \|f\|_{H_s}. \qquad \text{if } s > d/2.
\end{aligned}$$

Here Q_n is the orthogonal projection on $\widetilde{W}_n$. Consequently (11) is fulfilled for $s > s_d = d/2$.

Example 2.11 (Wavelets). *This example is adapted from Section 3.3 in Reiß (2008). Let $\mathcal{D} = \mathbb{T}^d = (0,1]^d$, for $d \in \mathbb{N}$, and consider the design points $\{k 2^{-j}\}_{k \in \{1, \cdots, 2^j\}^d}$, where $n = 2^{jd}$ for some $j \in \mathbb{N}$. We consider a multi-resolution analysis $\{V_j\}_{j \ge 0}$ on $L^2(\mathbb{T}^d)$ obtained by periodisation and tensor products. Let $\widetilde{\phi}$ be a standard orthonormal scaling function of an S-regular multi-resolution analysis for $L^2(\mathbb{R})$, with compact support in $[S-1, S]$. In particular, the polynomial exactness condition is satisfied: $\sum_{k \in \mathbb{Z}} k^q \widetilde{\phi}(x - k) - x^q$ is a polynomial of maximal degree $q - 1$ for $q \in [0, S-1]$. As shown in Reiß (2008), the functions*

$$e_{j,k}(x_1, \ldots, x_d) = \sum_{m \in \mathbb{Z}^d} 2^{jd/2} \prod_{i=1}^d \widetilde{\phi}(2^j x_i - k_i + 2^j m_i),$$

are well defined and form an orthonormal basis in $L^2(\mathbb{T}^d)$. Furthermore, for $\widetilde{W}_n := V_j = \mathsf{Span}\{e_{j,k} \mid k \in \{1, \cdots, 2^j\}^d\}$ with $n = 2^{jd} \ge 2S - 1$, conditions (10) is satisfied with constants C_1, C_2 that depend only on $\widetilde{\phi}$. Moreover, for the functions $e_{j,k}$ belong to the Besov space $B_{2,2}^s(\mathbb{T})$, for $s < S$, and, for every f in this Besov space and $d/2 < s < S$,

$$\|f - \mathcal{I}_n f\|_{L^2} \lesssim n^{-s/d} \|f\|_{B_{2,2}^s}.$$

Thus (11) *is satisfied, with the smoothness scale* $(H_s)_{s \in \mathbb{R}}$ *taken equal to the canonical Sobolev spaces (i.e. Besov spaces $B_{2,2}^s$) on* $\mathbb{T}^d$.

Other examples of suitable discretisation spaces are provided by orthogonal polynomials, for instance the systems of Legendre, Chebyshev, or Jacobi polynomials, etc., for suitably chosen design points. Canonical Sobolev spaces on $\mathbb{T} = (0,1]$ are Hilbert scales, and Assumption 2.2 is satisfied by standard Sturm-Liouville theory (see 5.2 in Canuto et al. (2010)): the polynomials form infinitely differentiable orthogonal bases in $L^2(\mathbb{T})$. Assumption 2.7 is satisfied with Gaussian quadrature points as design points (Section 5.3 in Canuto et al. (2010)). These results can be extended to multivariate domains by using tensor products.

3   General Contraction Rates

In this section we present two general theorems on posterior contraction, which we apply to different priors in later sections. The theorems address the situation of discrete observations. After making the appropriate modifications, their proofs borrow from the proofs for the continuous case in Gugushvili et al. (2020).

We form the posterior distribution $\Pi_n(\cdot \mid Y^n)$ as in (3), given a prior Π on the space $H = H_0$ and an observation $Y^n = (Y_1, \ldots, Y_n)$, whose conditional distribution given f is determined by the model (1). We study this random distribution under the assumption that Y^n follows the model (1) for a given 'true' function $f = f_0$, which we assume to be an element of H_β in a given smoothness scale $(H_s)_{s \in \mathbb{R}}$, as in Definition 2.1.

The theorem is stated in terms of the Galerkin solution to the continuous inverse problem, which is defined as follows. (See e.g., Kirsch (2011) for a general introduction to the Galerkin method and Appendix A for the two results needed in our proof.) Let $W_j = AV_j \subset G$ be the image under the operator A of a finite-dimensional approximation space V_j linked to the smoothness scale $(H_s)_{s \in \mathbb{R}}$ as in Assumption 2.2, and let $Q_j : G \to W_j$ be the orthogonal projection onto W_j. The *Galerkin solution* to Af is the inverse image $f^{(j)} \in V_j$ of $Q_j Af$. The idea is to simplify Af by projection before inverting the map A. (A solution $f^{(j)} \in V_j$ to the equation $Af^{(j)} = Q_j Af$ exists, under our assumption that $A : H \to G$ is injective, which implies that A is a bijection between the finite-dimensional vector spaces V_j and W_j.)

In our current setting we have no access to the continuous function Af, but only to its values at the design points. We overcome this by using the Galerkin solution $f^{(j,n)} = A^{-1} Q_j \mathcal{A}_n f$ to the interpolation $\mathcal{A}_n f$ of the dis-

crete signal, where $\mathcal{A}_n = \mathcal{I}_n A$, for $\mathcal{I}_n$ the interpolation operator defined in Section 2.2. This *discrete Galerkin solution* is illustrated in the following diagram

$$H \ni f \quad \xrightarrow{\mathcal{A}_n = \mathcal{I}_n A} \quad \mathcal{A}_n f \in \widetilde{W}_n \subset G$$

$$Q_j \Big\downarrow$$

$$H \supset V_j \ni f^{(j,n)} \xleftarrow{\quad A^{-1} \quad} Q_j \mathcal{A}_n f \in W_j \subset G$$

In this scheme the space $\widetilde{W}_n$, used to construct the continuous interpolation, may or may not be equal to $W_n = AV_n$. Setting it equal to W_n simplifies the scheme, but then the interpolation properties in Assumption 2.7 must be verified for AV_n.

Theorem 3.1 *For smoothness classes $(H_s)_{s\in\mathbb{R}}$ as in Definition 2.1, assume that $\|Af\| \simeq \|f\|_{-\gamma}$, for some $\gamma > 0$. Let $f^{(j,n)}$ denote the discrete Galerkin solution to $\mathcal{A}_n f = \mathcal{I}_n Af$ relative to linear subspaces V_j associated to $(H_s)_{s\in\mathbb{R}}$ as in Assumption 2.2 and interpolation spaces $\widetilde{W}_n$ satisfying (10)-(11). Let $f_0 \in H_\beta$ and $Af_0 \in G_{\beta+\gamma}$ for some $\beta \in (s_d - \gamma, S_d - \gamma)$, and for $\eta_n \geq \varepsilon_n \downarrow 0$ such that $n\varepsilon_n^2 \to \infty$, and $j_n \in \mathbb{N}$ such that $j_n \to \infty$, and some $c > 0$, assume*

$$j_n \leq cn\varepsilon_n^2, \tag{14}$$

$$\eta_n \geq \frac{\varepsilon_n}{\delta(j_n, \gamma)}, \tag{15}$$

$$\eta_n \geq \delta(j_n, \beta) \vee \frac{\delta_d(n, \beta+\gamma)}{\delta(j_n, \gamma)}. \tag{16}$$

Consider prior probability distributions Π on H_0 satisfying

$$\Pi\big(f \in H : \|Af - Af_0\|_n < \varepsilon_n\big) \geq e^{-n\varepsilon_n^2}, \tag{17}$$

$$\Pi\big(f \in H : \|f^{(j_n, n)} - f\| > \eta_n\big) \lesssim e^{-4n\varepsilon_n^2}. \tag{18}$$

Then the posterior distribution in the model (1) contracts at the rate η_n at f_0, i.e. for a sufficiently large constant M we have $\Pi_n\big(f : \|f - f_0\| > M\eta_n \mid Y_1, \ldots, Y_n\big) \to 0$, in probability if $Y_1, \ldots, Y_n$ follow (1) with $f = f_0$.

Proof Let $p_f^{(n)}$ be the (multivariate-normal) density of Y^n. The Kullback-Leibler divergence $P_{f_0}^{(n)} \log(p_{f_0}^{(n)}/p_f^{(n)})$ and variation $P_{f_0}^{(n)}\big(\log(p_{f_0}^{(n)}/p_f^{(n)})\big)^2$ can be computed to be $n\|Af - Af_0\|_n^2/2$ and twice this quantity, respectively.

Therefore, by (17) the set where these quantities are bounded above by $n\varepsilon_n^2$ has prior probability at least $e^{-n\varepsilon_n^2}$.

It follows by Lemma 8.2 of Ghosal and van der Vaart (2017) that the events $B_n := \{\int p_f^{(n)}/p_{f_0}^{(n)}(y^n)\,d\Pi(f) \geq e^{-2n\varepsilon_n^2}\}$ satisfy $P_{f_0}^{(n)}(B_n) \to 1$.

Furthermore, by (18) and Theorem 8.20 in the same reference, the posterior probability $\Pi_n\big(\|f^{(j_n,n)} - f\| > \eta_n \mid Y^n\big)$ tends to zero in $P_{f_0}^{(n)}$ probability.

Let $\mathcal{F}_n = \{f : \|f - f_0\| > M\eta_n, \|f^{(j_n,n)} - f\| \leq \eta_n\}$. We show below that there exist tests τ_n such that $P_{f_0}^{(n)}\tau_n \to 0$ and $P_f^{(n)}(1 - \tau_n) \leq e^{-4n\varepsilon_n^2}$, for every $f \in \mathcal{F}_n$. Then to prove the assertion of the theorem, we bound $\Pi_n\big(f : \|f - f_0\| > M\eta_n \mid Y_1, \ldots, Y_n\big)$ by

$$\tau_n + 1_{B_n^c} + \Pi_n\big(\|f^{(j_n,n)} - f\| > \eta_n \mid Y^n\big) + \Pi_n(\mathcal{F}_n \mid Y^n)1_{B_n}(1 - \tau_n).$$

The expectations under $P_{f_0}^{(n)}$ of the first three terms tend to zero by the preceding. In the fourth term we bound the denominator in Bayes's formula (3) divided by the likelihood at f_0 by $e^{-2n\varepsilon_n^2}$, and next bound its expectation by

$$e^{2n\varepsilon_n^2}P_{f_0}^{(n)}\int_{\mathcal{F}_n}\frac{p_f^{(n)}}{p_{f_0}^{(n)}}\,d\Pi(f)(1 - \tau_n) \leq e^{2n\varepsilon_n^2}\sup_{f\in\mathcal{F}_n}P_f^{(n)}(1 - \tau_n) \leq e^{-2n\varepsilon_n^2} \to 0.$$

This proof of the theorem can now be completed by constructing the tests τ_n.

Define the operator $R_j : G \mapsto V_j$ by $R_j = A^{-1}Q_j$, where $Q_j : G \to W_j$ is the orthogonal projection onto $W_j = AV_j$ and A^{-1} is the inverse of A (restricted to W_j). Consider the tests defined by, for a given constant M_0 determined later and $Y^{(n)}$ defined in (12),

$$\tau_n = 1\big\{\|R_{j_n}Y^{(n)} - f_0\| \geq M_0\eta_n\big\}. \tag{19}$$

Since by definition $f^{(j,n)} = R_j\mathcal{A}_n f$, we have

$$R_jY^{(n)} = R_j\mathcal{A}_n f + \frac{1}{\sqrt{n}}R_j\xi^{(n)} = f^{(j,n)} + \frac{1}{\sqrt{n}}R_j\xi^{(n)}. \tag{20}$$

The variable $R_j \xi^{(n)} = R_j Q_j \xi^{(n)}$ is a centred Gaussian random element in V_j with strong and weak second moments

$$\mathrm{E}\big\|R_j \xi^{(n)}\big\|^2 \leq \|R_j\|^2 \mathrm{E}\|Q_j \xi^{(n)}\|^2 \lesssim \|R_j\|^2 j \lesssim \frac{j}{\delta(j,\gamma)^2},$$

$$\sup_{\|f\|\leq 1} \mathrm{E}\langle R_j \xi^{(n)}, f\rangle^2 = \sup_{\|f\|\leq 1} \mathrm{E}\langle \xi^{(n)}, R_j^* f\rangle_G^2 \lesssim \sup_{\|f\|\leq 1} \|R_j^* f\|_G^2 \leq \|R_j^*\|^2 \lesssim \frac{1}{\delta(j,\gamma)^2}.$$

In both cases the inequality on $\|R_j\| = \|R_j^*\|$ at the far right side follows from (A3), and we also use that, by Lemma 2.8, the covariance operator of $\xi^{(n)}$ is bounded above by a multiple of the projection onto $\widetilde{W}_n$, and hence the identity, so that the covariance operator of $Q_j \xi^{(n)}$ is bounded above by a multiple of Q_j.

The first inequality shows that the first moment $\mathrm{E}\|R_j \xi^{(n)}\|$ of the variable $\|R_j \xi^{(n)}\|$ is bounded above by $\sqrt{j}/\delta(j,\gamma)$. By Borell's inequality (e.g. Lemma 3.1 in Ledoux and Talagrand (1991) and subsequent discussion), applied to the Gaussian random variable $R_j \xi^{(n)}$ in H_0, we see that, there exist positive constants a and b such that, for every $t > 0$,

$$\Pr\Big(\|R_j \xi^{(n)}\| > t + a\frac{\sqrt{j}}{\delta(j,\gamma)}\Big) \leq e^{-bt^2 \delta(j,\gamma)^2}.$$

For $t = 2\sqrt{n}\eta_n/\sqrt{b}$ and η_n, ε_n and j_n satisfying (14), (15) and (16) this yields, for some $a_1 > 0$,

$$\Pr\Big(\|R_{j_n} \xi^{(n)}\| > a_1 \sqrt{n}\eta_n\Big) \leq e^{-4n\varepsilon_n^2}. \tag{21}$$

We apply this to bound the two error probabilities of the tests τ_n.

Under f_0 the decomposition (20) is valid with $f = f_0$, and hence $R_j Y^{(n)} - f_0 = n^{-1/2} R_j \xi^{(n)} + f_0^{(j,n)} - f_0$. By the triangle inequality it follows that $\tau_n = 1$ implies that $n^{-1/2}\|R_{j_n} \xi^{(n)}\| \geq M_0 \eta_n - \|f_0^{(j,n)} - f_0\|$. By the triangle inequality followed by (A3) and (11), and (A5),

$$\|f_0^{(j,n)} - f_0\| \leq \|R_j\|\,\|\mathcal{I}_n A f_0 - A f_0\| + \|R_j A f_0 - f_0\|$$
$$\lesssim \frac{\delta_d(n, \beta + \gamma)}{\delta(j_n, \gamma)}\|A f_0\|_{\beta+\gamma} + \delta(j_n, \beta)\|f_0\|_\beta \leq M_1 \eta_n,$$

by assumption (16). Hence the probability of an error of the first kind satisfies

$$P_{f_0}^{(n)} \tau_n \leq \Pr\Big(\frac{1}{\sqrt{n}} \| R_{j_n} \xi^{(n)} \| \geq (M_0 - M_1) \eta_n \Big),$$

For $M_0 - M_1 > a_1$, the right side is bounded by $e^{-4n\varepsilon_n^2}$, by (21).

Under f the decomposition (20) gives that $R_j Y^{(n)} - f_0 = n^{-1/2} R_j \xi^{(n)} + f^{(j,n)} - f_0$. By the triangle inequality $\tau_n = 0$ implies that $n^{-1/2} \| R_{j_n} \xi^{(n)} \| \geq \| f^{(j_n,n)} - f_0 \| - M_0 \eta_n$. For f such that $\| f - f_0 \| > M \eta_n$ and $\| f - f^{(j_n,n)} \| \leq \eta_n$, we have $\| f^{(j_n,n)} - f_0 \| \geq (M-1)\eta_n$. Hence the probability of an error of the second kind satisfies

$$P_f^{(n)} (1 - \tau_n) \leq \Pr\Big(\frac{1}{\sqrt{n}} \| R_{j_n} \xi^{(n)} \| \geq (M - 1 - M_0) \eta_n \Big),$$

For $M - 1 - M_0 > a_1$, this is bounded by $e^{-4n\varepsilon_n^2}$, by (21).

We can first choose M_0 large enough so that $M_0 - M_1 > a_1$, and next M large enough so that $M - 1 - M_0 > a_1$, to finish the proof.

Inequality (17) is the usual *prior mass condition* for the 'direct problem' of estimating Af at the design points (see Ghosal et al. (2000)). It determines the rate of contraction ε_n of the posterior distribution of Af to Af_0 relative to the discrete seminorm $\| \cdot \|_n$. The rate of contraction η_n of the posterior distribution of f is slower due to the necessity of (implicitly) inverting the operator A. The theorem shows that the rate η_n depends on the combination of the prior, through (18), and the inverse problem, through the various approximation rates.

The theorem has a similar form as the theorem obtained in Gugushvili et al. (2020) in the case of observation of a continuous signal in white noise. Presently the Galerkin projection $f^{(j,n)}$ in (18) incorporates the errors of both inversion and discretisation. The factor $\delta_d(n, \beta + \gamma)/\delta(j_n, \gamma)$ in (16), which arises from having discrete observations only, will typically be negligible relative to $\delta(j_n, \beta)$.

The following refinement of the preceding theorem is appropriate for certain priors that are constructed to give an optimal contraction rate for multiple values of the smoothness β of the true parameter simultaneously. It considers a mixture prior of the form

$$\Pi = \int \Pi_\tau \, dQ(\tau), \tag{22}$$

where Π_τ is a prior on H, for every given 'hyper-parameter' τ running through some measurable space, and Q is a prior on this hyper-parameter. The idea is to adapt the prior to multiple smoothness levels through the hyper-parameter τ. The proof is very similar to the proof of Theorem 3.3 in Gugushvili et al. (2020) and is omitted.

Theorem 3.2 *Consider the setup and assumptions of Theorem 3.1 with a prior of the form* (22). *Assume that* (14), (15), (16) *and* (17) *hold, but replace* (18) *by the pair of conditions, for numbers $\eta_{n,\tau}$ and $C > 0$ and every τ,*

$$\Pi_\tau\big(f : \|f^{(j_n,n)} - f\| > \eta_{n,\tau}\big) \le e^{-4n\varepsilon_n^2}, \tag{23}$$

$$\Pi_\tau\big(f : \|f - f_0\| < 2\eta_{n,\tau}\big) \le e^{-4n\varepsilon_n^2}, \quad \forall \tau \text{ with } \eta_{n,\tau} \ge C\eta_n. \tag{24}$$

Then the posterior distribution in the model (1) *contracts at the rate η_n at f_0, i.e. for a sufficiently large constant M we have $\Pi_n\big(f : \|f - f_0\| > M\eta_n \mid Y_1, \ldots, Y_n\big) \to 0$, in probability if $Y_1, \ldots, Y_n$ follow* (1) *with $f = f_0$.*

4 Random Series Priors

Suppose that $\{\phi_i\}_{i\in\mathbb{N}}$ is an orthonormal basis of $H = H_0$ that gives optimal approximation relative to the scale of smoothness classes $(H_s)_{s\in\mathbb{R}}$ in the sense that the linear spaces $V_j = \mathsf{Span}\{\phi_i\}_{i<j}$ satisfy Assumption 2.2. Consider a prior defined as the law of the random series

$$f = \sum_{i=1}^M f_i\phi_i, \tag{25}$$

where M is a random variable in $\mathbb{N}$ independent from the independent random variables $f_1, f_2, \ldots$ in $\mathbb{R}$.

Assumption 4.1 *(Random series prior).*

1. *The probability density function p_M of M satisfies, for some positive constants b_1, b_2,*

$$e^{-b_1 k} \lesssim p_M(k) \lesssim e^{-b_2 k}, \qquad \forall k \in \mathbb{N}.$$

2. *The variable f_i has density $p(\cdot/\kappa_i)/\kappa_i$, for a given Lebesgue probability density p on $\mathbb{R}$ and a positive constant κ_i such that, for some $C > 0$*

and $0 < v < w < \infty$, $\beta_0 > 0$ and $\alpha > 0$,

$$e^{-C|x|^w} \lesssim p(x) \lesssim e^{-C|x|^v}, \tag{26}$$

$$i^{-\beta_0/d} \lesssim \kappa_i \lesssim i^\alpha. \tag{27}$$

Priors of this type were studied in Arbel et al. (2013), and applied to inverse problems in the SVD framework in Ray (2013). The assumptions on p_M, p and the scale parameters κ_i are mild and satisfied by many examples. For Gaussian variables f_j and degenerate M, the series (25) is a Gaussian process.

The basis functions are connected to the operator A, because they must generate the smoothness scale such that Assumption 2.5 holds, but this allows considerable freedom. For instance, a basis on wavelets may be appropriate for differential operators that are smoothing in the Sobolev scale.

Theorem 4.2 (*Random Series Prior*). *Let $(\phi_i)_{i\in\mathbb{N}}$ be an orthonormal basis of H_0 such that the spaces $V_j = \mathsf{Span}\{\phi_i\}_{i<j}$ satisfy Assumption 2.2 with $\delta(j, s) = j^{-s/d}$ relative to smoothness classes $(H_s)_{s\in\mathbb{R}}$ as in Definition 2.1. Assume that Assumption 2.7 holds with $\delta_d(n, s) = n^{-s/d}$ and sufficiently large S_d. Suppose Assumption 2.5 holds for the operator A, and $f_0 \in H_\beta$ for some $\beta \in (0, S]$ such that $\beta + \gamma > s_d \vee d/2$. Then, for the random series prior defined in (25) and satisfying Assumption 4.1 with $\beta_0 \le \beta$, and sufficiently large $M > 0$, for $\tau = (\beta + \gamma)(1 + 2\gamma/d)/(2\beta + 2\gamma + d)$,*

$$\Pi_n\Big(f : \|f - f_0\| > Mn^{-\beta/(2\beta+2\gamma+d)}(\log n)^\tau \mid Y^{(n)}\Big) \overset{\mathbb{P}^{(n)}_{f_0}}{\to} 0.$$

The rate $n^{-\beta/(2\beta+2\gamma+d)}$ is known to be the minimax rate of estimation of a β-regular function on a d-dimensional domain, in an inverse problem with inverse parameter γ (see, e.g., Cohen et al. (2004)). The assumption that $\delta(j, s) = j^{-s/d}$ places the setup of the theorem in this setting, and hence the rate of contraction obtained in the preceding theorem is the minimax rate up to a logarithmic factor. The rate is adaptive to the regularity of β of the true parameter, which is not used in the construction of the prior, apart from the assumption that $\beta \ge \beta_0$. (See Ghosal et al. (2008) and Chapter 10 in Ghosal and van der Vaart (2017) for general discussion of adaptation in the Bayesian sense.)

The proof of the theorem is deferred to Section 6; it will be based on Theorem 3.1.

5 Gaussian Priors and Gaussian Mixtures

If the function f in (1) is equipped with a Gaussian prior, then the corresponding posterior distribution is Gaussian as well. The posterior mean is then equal to the solution found by the method of Tikhonov-type regularization (see e.g. Florens and Simoni (2016), Knapik et al. (2011), Stuart (2010)). Because a Gaussian prior possesses a fixed smoothness level, it is usually scaled with a hyper-parameter, leading to a mixture of Gaussian priors. In this section we derive a rate of contraction both for Gaussian priors of fixed smoothness and for scaled Gaussian priors.

By definition a random variable F with values in a Hilbert space H_0 is Gaussian if $\langle F, g \rangle$ is normally distributed, for every $g \in H_0$. It is *centred* if these variables have zero means, and it has *covariance operator* $C : H_0 \to H_0$ if $E\langle F, g \rangle^2 = \langle Cg, g \rangle$, for every $g \in H_0$.

A Gaussian prior can also, and without loss of generality, be described through a basis expansion. Given an orthonormal basis $(\phi_j)_{j \in \mathbb{N}}$ of H_0 and a sequence of independent variables $F_j \sim N(0, \sigma_j^2)$, for given $\sigma_j^2 > 0$ with $\sum_j \sigma_j^2 < \infty$, the variable $F = \sum_{j=1}^{\infty} F_j \phi_j$ is Gaussian with covariance operator $C_\sigma g = \sum_j \sigma_j^2 g_j \phi_j$ if $g = \sum_j g_j \phi_j$.

It is customary to relate such a prior to the generator of a *Hilbert scale* of spaces, as follows. We restrict to the case that σ_j is polynomial in $1/j$. We define spaces H_s as in Example 2.3, with $b_j = j^{1/d}$. The operator $Lg = \sum_j b_j g_j \phi_j$ is an isometry $L : H_s \to H_{s-1}$, for every s, and is said to generate the Hilbert scale $(H_s)_{s \in \mathbb{R}}$. For $\alpha > d/2$, the operator $L^{-2\alpha}$ has finite trace $\sum_j j^{-2\alpha/d}$ and defines a Gaussian distribution on H_0. Because $L^s F$ has covariance operator $L^{2s-2\alpha}$, which has finite trace if $s < \alpha - d/2$, this variable is a proper Gaussian variable in H_0, for every $s < \alpha - d/2$. Hence F is a proper Gaussian variable in $H_s = L^{-s}H_0$ for the same values of s. We may describe this by saying that the prior with covariance operator $L^{-2\alpha}$ is '(nearly) of smoothness $\alpha - d/2$'.

The following theorem gives a posterior contraction rate for this prior. The proof is deferred to Section 6.

Theorem 5.1 (*Gaussian Prior*). *Suppose Assumption 2.5 holds for the operator A relative to the Hilbert scale $(H_s)_{s \in \mathbb{R}}$ generated by the operator L^{-1}, as indicated. Assume that Assumption 2.7 holds with $\delta_d(n, s) = n^{-s/d}$ and $s_d = d/2$. Assume $f_0 \in H_\beta$, for some $\beta > 0$ such that $\beta + \gamma > d/2$, and let the prior be zero-mean Gaussian with covariance operator $L^{-2\alpha}$, for some $\alpha > d/2 \vee \beta$ such that $\alpha + \gamma > d$. Then the posterior distribution satisfies,*

for sufficiently large $M > 0$,

$$\Pi_n\Big(f : \|f - f_0\| > M n^{-((\alpha - d/2)\wedge\beta)/(2\alpha + 2\gamma)} \mid Y^{(n)}\Big) \overset{\mathbb{P}^{(n)}_{f_0}}{\to} 0.$$

The Gaussian prior in the preceding theorem has a fixed smoothness $\alpha - d/2$, which results in a contraction rate that depends on both the prior smoothness and the smoothness β of the true function. This rate is equal to the minimax rate $n^{-\beta/(2\beta + 2\gamma + d)}$ (see Cohen et al. (2004)) only when $\alpha - d/2 = \beta$, i.e., when the prior smoothness $\alpha - d/2$ matches the true smoothness β. By mixing over Gaussian priors of varying smoothness, the minimax rate can often be obtained simultaneously for a range of values β (cf. Knapik et al. (2016), van der Vaart (2010), Szabó et al. (2013)).

In the following theorem we consider mixtures of the mean-zero Gaussian priors with covariance operators $\tau^2 L^{-2\alpha}$ over the 'hyper-parameter' τ. Thus the prior Π is the distribution of τF, where F is a zero-mean Gaussian variable in H_0 with covariance operator $L^{-2\alpha}$ as before, and τ is an independent scale parameter. The variable $1/\tau^a$ may be taken to possess a Gamma distribution for some given $0 < a \le 2$, or, more generally, should satisfy the following mild condition.

Assumption 5.2 *The distribution Q of τ has support $[0, \infty)$ and satisfies*

$$\begin{cases} -\log Q\big((t, 2t)\big) \lesssim t^{-2}, & \text{as } t \downarrow 0, \\ -\log Q\big((t, 2t)\big) \lesssim t^{d/(\alpha - d/2)}, & \text{as } t \to \infty. \end{cases}$$

Theorem 5.3 *(Gaussian mixture prior). Suppose Assumption 2.5 holds for the operator A relative to the Hilbert scale $(H_s)_{s\in\mathbb{R}}$ generated by the operator L^{-1}, as indicated. Assume that Assumption 2.7 holds with $\delta_d(n, s) = n^{-s/d}$ and $s_d = d/2$. Assume that $f_0 \in H_\beta$, for some $\beta > 0$ such that $\beta + \gamma > d/2$, and let the prior be a mixture of the zero-mean Gaussian $\tau^2 L^{-2\alpha}$ over the parameters τ equipped with a prior satisfying Assumption 5.2, for some $\alpha > d/2$ such that $\alpha + \gamma > d$. Then the posterior distribution satisfies, for sufficiently large $M > 0$,*

$$\Pi_n\Big(f : \|f - f_0\| > M n^{-\beta/(2\beta + 2\gamma + d)} \mid Y^{(n)}\Big) \overset{\mathbb{P}^{(n)}_{f_0}}{\to} 0.$$

The proof of the theorem is deferred to Section 6.

6 Proofs

6.1 Proof of Theorem 4.2 The theorem is a corollary to Theorem 3.1. We shall verify the conditions with

$$\varepsilon_n \simeq (\log n/n)^{(\beta+\gamma)/(2\beta+2\gamma+d)}, \quad j_n \simeq n^{d/(2\beta+2\gamma+d)}(\log n)^{(2\beta+2\gamma)/(2\beta+2\gamma+d)}.$$

Let P_j be the orthogonal projection of H on the linear span of the first $j-1$ basis elements ϕ_j, and define an additional sequence of integers by

$$i_n \simeq (n/\log n)^{d/(2\beta+2\gamma+d)}.$$

By the orthogonality of the basis (ϕ_i), the function ϕ_j is orthogonal to the space V_j spanned by $(\phi_i)_{i<j}$. Hence $P_j\phi_j = 0$, so that $\|\phi_j\|_{-s} \leq \delta(j,s)\|\phi_j\| \lesssim j^{-s/d}$, for every j and $s \geq 0$, by (A1). The same estimate is also true for negative smoothness levels s such that $0 < -s < S$, directly by assumption (7). Therefore, by the triangle inequality, we have

$$\|f\|_s \lesssim \sum_j |f_j| j^{s/d}, \qquad \text{if } f = \sum_j f_j \phi_j.$$

Furthermore, since $f_0 \in H_\beta$ by assumption, the norm duality (5) gives that $|f_{0,i}| = |\langle f_0, \phi_i \rangle| \leq \|f_0\|_\beta \|\phi_i\|_{-\beta} \lesssim i^{-\beta/d}$.

First we verify the prior condition (17) of the direct problem. By Lemma 2.9, $\|Af - Af_0\|_n \simeq \|\mathcal{I}_n(Af - Af_0)\|$. By several applications of the triangle inequality, since $\beta + \gamma \in (s_d, S_d)$, $\|Af\|_{\beta+\gamma} \simeq \|f\|_\beta$ and $\|Af - Af_0\| \simeq \|f - f_0\|_{-\gamma}$,

$$\begin{aligned}
\|\mathcal{I}_n(Af- Af_0)\| &\leq \|\mathcal{I}_n Af - Af\| + \|\mathcal{I}_n Af_0 - Af_0\| + \|Af - Af_0\| \\
&\lesssim \delta_d(n,\beta+\gamma)\big(\|f\|_\beta + \|f_0\|_\beta\big) + \|f - P_{i_n}f_0\|_{-\gamma} + \|f_0 - P_{i_n}f_0\|_{-\gamma} \\
&\lesssim \delta_d(n,\beta+\gamma)\|f - P_{i_n}f_0\|_\beta + \|f - P_{i_n}f_0\|_{-\gamma} \\
&\quad + \delta_d(n,\beta+\gamma)\big(\|P_{i_n}f_0\|_\beta + \|f_0\|_\beta\big) + \delta(i_n,\gamma)\delta(i_n,\beta)\|f_0\|_\beta,
\end{aligned}$$

by (A1). The last term is of the order $\delta(i_n,\gamma)\delta(i_n,\beta) = i_n^{-(\gamma+\beta)/d} \simeq \varepsilon_n$, while the second last term is bounded above by $\delta_d(n,\beta+\gamma)\big(\|f_0\|/\delta(i_n,\beta) + \|f_0\|_\beta\big) \ll \varepsilon_n$, if $(\beta+\gamma)^2 > \beta d/2$. For $f = \sum_{i=1}^{i_n-1} f_i\phi_i \in V_{i_n}$, the sum of the first two terms on the right is bounded above by

$$\delta_d(n,\beta+\gamma)\sum_{i=1}^{i_n-1} |f_i - f_{0,i}| i^{\beta/d} + \sum_{i=1}^{i_n-1} |f_i - f_{0,i}| i^{-\gamma/d} \lesssim \sum_{i=1}^{i_n-1} |f_i - f_{0,i}| i^{-\gamma/d}.$$

By arguments as in Ray ([2013](#)), it is shown in Section 8.1 of Gugushvili et al. ([2020](#)) that the right side of this equation is bounded above by ε_n with prior probability at least $e^{-n\varepsilon_n^2/2}$. Since also $\Pi(M = i_n - 1) \geq e^{-b_1 i_n} \geq e^{-n\varepsilon_n^2/2}$, it follows that ([17](#)) is satisfied.

Next we verify ([18](#)). Since $\Pi(M \geq j_n) \leq e^{-b_2' j_n} \leq e^{-4n\varepsilon_n^2}$, by Assumption [4.1](#), we may intersect the event in ([18](#)) with the event $M < j_n$. For $M < j$ the random series $f = \sum_{i=1}^{M} f_i \phi_i$ is contained in V_j and the Galerkin approximation $R_j A f$ of f is exact. Since $f^{(j,n)} = R_j \mathcal{I}_n A f$, the triangle inequality followed by ([A3](#)) and ([11](#)) give, for $s + \gamma \in (s_d, S_d)$ and $f \in V_j$, in view of ([7](#)),

$$
\begin{aligned}
\|f^{(j,n)} - f\| &\leq \|R_j \mathcal{I}_n A f - R_j A f\| + \|R_j A f - f\| \\
&\leq \frac{\delta_d(n, s + \gamma)}{\delta(j, \gamma)} \|f\|_s + 0 \leq \frac{\delta_d(n, s + \gamma)}{\delta(j, \gamma)\delta(j, s)} \|f\|.
\end{aligned}
$$

We conclude that it suffices to prove that

$$
\Pi\left(\sum_{i=1}^{j_n-1} f_i^2 > \eta_n^2 \left(\frac{n}{j_n}\right)^{(2\gamma+2s)/d}\right) \leq e^{-4n\varepsilon_n^2}. \tag{28}
$$

By Assumption [4.1](#) we have that

$$
\mathrm{E} \sum_{i=1}^{j_n-1} f_i^2 \simeq \sum_{i=1}^{j_n-1} \kappa_i^2 \lesssim j_n^{2\alpha+1}.
$$

$$
\mathrm{E}\left|\sum_{i=1}^{j_n-1} (f_i^2 - \mathrm{E}f_i^2)\right| \lesssim \sqrt{\sum_{i=1}^{j_n-1} \kappa_i^4} \lesssim j_n^{2\alpha+1/2}.
$$

Furthermore, by the tail bound on the density p we further have that the $\psi_{v/2}$ Orlicz norm of f_i^2 is bounded above by κ_i^2. Therefore, (see e.g. Proposition A.1.6 in van der Vaart and Wellner ([1996](#))), for q conjugate to $v/2$,

$$
\left\|\sum_{i=1}^{j_n-1} (f_i^2 - \mathrm{E}f_i^2)\right\|_{\psi_{v/2}} \lesssim \begin{cases} j_n^{2\alpha+1/2} + (\log j_n)^{2/v} \max_{i<j_n} \kappa_i^2, & \text{if } v \leq 2, \\ j_n^{2\alpha+1/2} + (\sum_{i<j_n} \kappa_i^{2q})^{1/q}, & \text{if } 2 < v \leq 4, \end{cases}
$$

In both cases the first term $j_n^{2\alpha+1/2}$ dominates. Provided

$$
\eta_n^2 (n/j_n)^{(2\gamma+2s)/d} = n^{(4(s+\gamma)(\beta+\gamma)/d-2\beta)/(2\beta+2\gamma+d)} (\log n)^a \gtrsim j_n^{2\alpha+1},
$$

we can first centre $\sum_{i=1}^{j_n-1} f_i^2$ at mean zero, and next bound the tail of the centred variable with the help of the Orlicz norm. This will give a bound on the left side of (28) of the type

$$\frac{1}{\psi_{v/2}}\Big(\frac{n^{(4(s+\gamma)(\beta+\gamma)/d-2\beta)/(2\beta+2\gamma+d)}(\log n)^a}{j_n^{2\alpha+1/2}}\Big).$$

For sufficiently large s the argument is an arbitrarily large power of n times a logarithmic factor and hence this is easily smaller than $e^{-4n\varepsilon_n^2}$. Thus condition (18) holds provided s, which was required to belong to $(s_d-\gamma, S_d-\gamma)$, can be chosen sufficiently large. This can be worked out to lead to the requirement that $2(\beta+\gamma)(S_d+\gamma) > d^2(v^{-1} \vee (\alpha+1/2)) + \beta d$.

  6.2 Proof of Theorem 5.1 The theorem is a corollary to Theorem 3.1 with

$$\varepsilon_n \simeq n^{-(\beta\wedge(\alpha-d/2)+\gamma)/(2\alpha+2\gamma)}. \tag{29}$$

We verify that ε_n satisfies the prior mass condition (17) of the direct problem, and next identify η_n from the other prior mass condition (18).

In both cases we reduce the problem to the case of continuous observations by estimating the interpolation error, using the following lemma.

Lemma 6.1 *Under the assumptions of Theorem 5.1, there exists a constant $c > 0$ so that for $s < \alpha - d/2$ with $s + \gamma \in (s_d, S_d)$ and any $t > 0$,*

$$\Pi\big(\|\mathcal{I}_n Af - Af\| > t\big) \lesssim e^{-ct^2 n^{(2s+2\gamma)/d}}.$$

Proof Since the prior of f is a centred Gaussian distribution with covariance operator $L^{-2\alpha}$, we have that $f \in H_s$, almost surely, for $s < \alpha - d/2$, and consequently, $Af \in G_{s+\gamma}$ almost surely. Assumptions (11) and (9) give

$$\|\mathcal{I}_n Af - Af\| \lesssim \delta_d(n, s+\gamma)\|Af\|_{s+\gamma} \simeq \delta_d(n, s+\gamma)\|f\|_s = \delta_d(n, s+\gamma)\|L^s f\|.$$

The variable $L^s F$, for $F \sim \Pi$, possesses a centred Gaussian distribution in H_0 with covariance operator $L^{-2(\alpha-s)}$, which is finite trace since $\alpha-s > d/2$. Therefore, Borell's inequality gives $\Pi\big(\|L^s F\| > t\big) \leq e^{-ct^2}$, for every $t > 0$ and a suitable constant $c > 0$. Together with the preceding display, this implies the lemma.

For the proof of (17), we first note that $\|Af - Af_0\|_n \simeq \|\mathcal{I}_n(Af - Af_0)\|$, by Lemma 2.9. Next, by the triangle inequality,

$$\|\mathcal{I}_n(Af - Af_0)\| \leq \|\mathcal{I}_n Af - Af\| + \|\mathcal{I}_n Af_0 - Af_0\| + \|Af - Af_0\|. \tag{30}$$

Since $Af_0 \in G_{\beta+\gamma}$ and $\beta + \gamma > d/2 = s_d$, the middle term satisfies $\|\mathcal{I}_n Af_0 - Af_0\| \le \delta_d(n, \beta+\gamma)\|Af_0\|_{\beta+\gamma} \simeq n^{-(\beta+\gamma)/d} \ll n^{-(\beta+\gamma)/(2\alpha+2\gamma)} \le \varepsilon_n$, if $\alpha + \gamma > d/2$, and hence is negligible in the context of (17).

Because $\alpha + \gamma > d$ by assumption, we have $d/2 - \gamma < \alpha - d/2$. For s in the interval $(d/2 - \gamma, < \alpha - d/2)$, we have $s + \gamma > d/2 = s_d$ and hence Lemma 6.1 gives that $\Pi\big(\|\mathcal{I}_n Af - Af\| > d\varepsilon_n\big)$ is bounded above by $e^{-c_1 \varepsilon_n^2 n^{(2s+2\gamma)/d}} \ll e^{-n\varepsilon_n^2}$, since $s + \gamma > d/2$.

Using the inequality $\Pr(X + Y < \varepsilon) \ge \Pr(X < \varepsilon/2) - \Pr(Y > \varepsilon/2)$, for any random variables X, Y, and (30), we see the validity of (17) if $\Pi\big(\|Af - Af_0\| < \varepsilon_n\big) \ge e^{-n\varepsilon_n^2}$. This prior estimate is the same as the prior mass condition for the white noise model, which was studied in Gugushvili et al. (2020). The prior estimate follows from the following lemma, which is Lemma 9.2 from the latter paper.

Lemma 6.2 *Under the assumptions of Theorem 5.1, for $f_0 \in H_\beta$, as $\varepsilon \downarrow 0$,*

$$-\log \Pi\big(f : \|Af - Af_0\| < \varepsilon\big) \lesssim \begin{cases} \varepsilon^{-d/(\alpha+\gamma-d/2)}, & \text{if } d/2 < \alpha \le \beta + d/2, \\ \varepsilon^{-(2\alpha-2\beta)/(\beta+\gamma)}, & \text{if } \alpha > \beta + d/2. \end{cases}$$

The next step of the proof is to bound the prior probability in (18), which involves the discrete Galerkin approximation $f^{(j,n)} = R_j \mathcal{I}_n Af$, where $R_j = A^{-1} Q_j$. By the triangle inequality,

$$\begin{aligned} \|f^{(j,n)} - f\| &\le \|R_j Af - f\| + \|R_j \mathcal{I}_n Af - R_j Af\| \\ &\le \|f^{(j)} - f\| + \frac{1}{\delta(j,\gamma)}\|\mathcal{I}_n Af - Af\|, \end{aligned} \tag{31}$$

by (A3) and the definition of $f^{(j)} = R_j Af$. The contribution of the second term on the right to (18) can be bounded by $\Pi\big(\|\mathcal{I}_n Af - Af\| > \delta(j_n, \gamma)\eta_n\big) \le \Pi\big(\|\mathcal{I}_n Af - Af\| > \varepsilon_n\big)$, because $\delta(j_n, \gamma)\eta_n \ge \varepsilon_n$, by (15). The last probability was seen to be $o(e^{-n\varepsilon_n^2})$ in the preceding. It follows that in the verification of (18) the discrete Galerkin solution $f^{(j_n,n)}$ may be replaced by the continuous Galerkin solution $f^{(j_n)}$. The prior estimates for this continuous case are given by the following lemma, which is Lemma 9.3 in Gugushvili et al. (2020). (In its proof $\|(R_j A - I)L^{-\alpha}\phi\|_0$ should be bounded by $\|L^{-\alpha}\phi_0\|_0 \le \delta(i, \alpha)$.)

Lemma 6.3 *Under the assumptions of Theorem 5.1, there exist $a, b > 0$, such that for every $j \in \mathbb{N}$ and $t > 0$,*

$$\Pi\big(f : \|f^{(j)} - f\| > t + aj^{1/2-\alpha/d}\big) \le e^{-bt^2 j^{2\alpha/d}}.$$

For $t^2 = n\varepsilon_n^2/(4bj_n^{2\alpha/d})$ the bound in the preceding lemma becomes $e^{-4n\varepsilon_n^2}$. Hence (18) is satisfied for

$$\eta_n \gtrsim \sqrt{n}\varepsilon_n/j_n^{\alpha/d} + j_n^{1/2-\alpha/d}.$$

Here we choose ε_n the minimal solution that satisfies the direct prior mass condition (17), given in (29). Next we solve for η_n under the constraints, (15) and (16). The first of these constraints, $j_n \leq n\varepsilon_n^2$, shows that the first term on the right side of the preceding display always dominates the second term. Therefore, we obtain the requirements $j_n \leq n\varepsilon_n^2$ and

$$\begin{aligned}
\eta_n &\geq \sqrt{n}\,n^{-(\beta\wedge(\alpha-d/2)+\gamma)/(2\alpha+2\gamma)}, \\
\eta_n &\geq n^{-(\beta\wedge(\alpha-d/2)+\gamma)/(2\alpha+2\gamma)}j_n^{\gamma/d}, \\
\eta_n &\geq j_n^{-\beta/d}.
\end{aligned}$$

Depending on the relation between α and $\beta + d/2$, two situations arise:

1. $\alpha < \beta + d/2$. We choose $j_n \simeq n^{d/(2\alpha+2\gamma)} = n\varepsilon_n^2$ and then see that the first two requirements in the preceding display both reduce to $\eta_n \geq n^{-(\alpha-d/2)/(2\alpha+2\gamma)}$, while the third becomes $\eta_n \geq n^{-\beta/(2\alpha+2\gamma)}$ and becomes inactive.

2. $\alpha > \beta + d/2$. We choose $j_n \simeq n^{d/(2\alpha+2\gamma)} \leq n\varepsilon_n^2$, and then see that all three requirements reduce to $\eta_n \geq n^{-\beta/(2\alpha+2\gamma)}$.

6.3 Proof of Theorem 5.3 Theorem 5.3 is a corollary of Theorem 3.2, with the choices

$$\begin{aligned}
\eta_n &\simeq n^{-\beta/(2\beta+2\gamma+d)}, & \varepsilon_n &\simeq n^{-(\beta+\gamma)/(2\beta+2\gamma+d)}, \\
j_n &\simeq n\varepsilon_n^2 = n^{d/(2\beta+2\gamma+d)}, & \eta_{n,\tau} &\simeq \tau\,n^{(d/2-\alpha)/(2\beta+2\gamma+d)}.
\end{aligned}$$

Conditions (14), (15), and (16) are satisfied for these choices. Condition (24) is exactly the same as in the continuous observation scheme and is verified in Gugushvili et al. (2020). It remains to verify (17) and (23).

For (17), we use that $\|Af - Af_0\|_n \simeq \|\mathcal{I}_n(Af - Af_0)\|$ and next (30). The middle term on the right is $o(\varepsilon_n)$. Therefore, it suffices to show, for some τ_n,

$$\int_{\tau_n}^{2\tau_n} \Pi_\tau\big(\|\mathcal{I}_n Af - Af\| + \|Af - Af_0\| < \varepsilon_n\big)\,dQ(\tau) \geq e^{-n\varepsilon_n^2}.$$

In Gugushvili et al. (2020) this inequality is shown to be true without the term $\|\mathcal{I}_n A f - A f\|$, for $\tau_n = n^{-(\alpha - d/2 - \beta)/(2\beta + 2\gamma + d)}$, and where we can replace ε_n by $\varepsilon_n/2$. Thus it suffices to show that

$$\sup_{\tau_n < \tau < 2\tau_n} \Pi_\tau\big(\|\mathcal{I}_n A f - A f\| > \varepsilon_n/2\big) = o(e^{-n\varepsilon_n^2}).$$

Because Π_τ is a scaling of $\Pi = \Pi_1$, the left side is bounded above by $\Pi\big(\|\mathcal{I}_n A f - A f\| > \varepsilon_n/(2\tau_n)\big)$. By Lemma 6.1 this is $o(e^{-n\varepsilon_n^2})$ if $(\varepsilon_n/\tau_n)^2 n^{(2s+2\gamma)/d} \gg n\varepsilon_n^2$, where $s < \alpha - d/2$ can be chosen arbitrarily close to $\alpha - d/2$. (As noted in the preceding section, it can be chosen so that $s + \gamma > d/2$, since $\alpha + \gamma > d$, by assumption.) This leads to the inequality $\alpha + \gamma > (d/2)(2\beta + \gamma + d)/(2\beta + 2\gamma)$, which is implied by the assumptions $\alpha + \gamma > d$ and $\beta + \gamma > d/2$.

For the proof of (23), we employ (31). As the continuous time equivalent that replaces $f^{(j,n)}$ by $f^{(j)}$, is shown to be true in Gugushvili et al. (2020), it suffices to show that, for every τ,

$$\Pi_\tau\Big(\frac{1}{\delta(j_n, \gamma)}\|\mathcal{I}_n A f - A f\| > \eta_{n,\tau}\Big) \le e^{-4n\varepsilon_n^2}.$$

By Lemma 6.1 this true if $(\eta_{n,\tau}/\tau)^2 j_n^{-2\gamma/d} n^{(2s+2\gamma)/d} \gg n\varepsilon_n^2$. Since $(\eta_{n,\tau}/\tau) j_n^{-\gamma/d} \simeq \varepsilon_n/\tau_n$, this reduces to the same inequality as in the preceding paragraph.

Declarations

Competing interests The authors declare to have no conflicts of interest.

References

Agapiou, S., Larsson, S., Stuart, A.M.: Posterior contraction rates for the bayesian approach to linear ill-posed inverse problems. Stochastic Processes and their Applications **123**(10), 3828–3860 (2013)

Arbel, J., Gayraud, G., Rousseau, J.: Bayesian optimal adaptive estimation using a sieve prior. Scandinavian Journal of Statistics **40**(3), 549–570 (2013). https://doi.org/10.1002/sjos.12002

Canuto, C., Hussaini, M.Y., Quarteroni, A., Zang, T.A.: Spectral Methods: Fundamentals in Single Domains. Scientific Computation. Springer (2010)

Cavalier, L.: Nonparametric statistical inverse problems. Inverse Problems **24**(3), 034004 (2008)

Cohen, A., Hoffmann, M., Reiß, M.: Adaptive wavelet galerkin methods for linear inverse problems. SIAM Journal on Numerical Analysis **42**(4), 1479–1501 (2004) http://dx.doi.org/10.1137/S0036142902411793. https://doi.org/10.1137/S0036142902411793

Florens, J.-P., Simoni, A.: Regularizing priors for linear inverse problems. Econometric Theory **32**(01), 71–121 (2016)

Ghosal, S., Ghosh, J.K., van der Vaart, A.W.: Convergence rates of posterior distributions. The Annals of Statistics **28**(2), 500–531 (2000). https://doi.org/10.1214/aos/1016218228

Ghosal, S., Lember, J., van der Vaart, A.: Nonparametric Bayesian model selection and averaging. Electron. J. Stat. **2**, 63–89 (2008). https://doi.org/10.1214/07-EJS090

Ghosal, S., van der Vaart, A.: Convergence rates of posterior distributions for non-iid observations. Ann. Statist. **35**(1), 192–223 (2007). https://doi.org/10.1214/009053606000001172

Ghosal, S., van der Vaart, A.: Fundamentals of Nonparametric Bayesian Inference. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press (2017)

Goldenshluger, A., Pereverzev, S.V.: On adaptive inverse estimation of linear functionals in hilbert scales. Bernoulli **9**(5), 783–807 (2003). https://doi.org/10.3150/bj/1066418878

Gugushvili, S., van der Vaart, A.W., Yan, D.: Bayesian inverse problems with partial observations. Trans. A. Razmadze Math. Inst. **172**(3, part A), 388–403 (2018). https://doi.org/10.1016/j.trmi.2018.09.002

Gugushvili, S., van der Vaart, A., Yan, D.: Bayesian linear inverse problems in regularity scales. Ann. Inst. Henri Poincaré Probab. Stat. **56**(3), 2081–2107 (2020). https://doi.org/10.1214/19-AIHP1029

Haroske, D., Triebel, H.: Distributions, Sobolev Spaces, Elliptic Equations. EMS Monographs in mathematics. European Mathematical Society (2008)

Kirsch, A.: An Introduction to the Mathematical Theory of Inverse Problems. Applied Mathematical Sciences. Springer (2011)

Knapik, B., Salomond, J.-B.: A general approach to posterior contraction in nonparametric inverse problems. Bernoulli **24**(3), 2091–2121 (2018). https://doi.org/10.3150/16-BEJ921

Knapik, B.T., Szabó, B.T., van der Vaart, A.W., van Zanten, J.H.: Bayes procedures for adaptive inference in inverse problems for the white noise model. Probab. Theory Related Fields **164**(3-4), 771–813 (2016). https://doi.org/10.1007/s00440-015-0619-7

Knapik, B., van der Vaart, A., van Zanten, J.: Bayesian inverse problems with gaussian priors. The Annals of Statistics **39**(5), 2626–2657 (2011)

Knapik, B., van der Vaart, A., van Zanten, J.: Bayesian recovery of the initial condition for the heat equation. Communications in Statistics - Theory and Methods (2013) **42**(7), 1294–1313 (2013)

Ledoux, M., Talagrand, M.: Probability in Banach Spaces vol. 23, p. 480. Springer, Berlin (1991)

Mathé, P., Pereverzev, S.V.: Optimal discretization of inverse problems in hilbert scales. regularization and self-regularization of projection methods. SIAM Journal on Numerical Analysis **38**(6), 1999–2021 (2001). https://doi.org/10.1137/S003614299936175X

Natterer, F.: Error bounds for tikhonov regularization in hilbert scales. Applicable Analysis **18**(1-2), 29–37 (1984). https://doi.org/10.1080/00036818408839508

Nickl, R.: Bayesian Non-linear Inverse Problems. Zürich Lectures in Advanced Mathematics. ETH, Zürich (2023)

Ray, K.: Bayesian inverse problems with non-conjugate priors. Electron. J. Statist. **7**, 2516–2549 (2013). https://doi.org/10.1214/13-EJS851

Reiß, M.: Asymptotic equivalence for nonparametric regression with multivariate and random design. Ann. Statist. **36**(4), 1957–1982 (2008). https://doi.org/10.1214/07-AOS525

Sniekers, S., van der Vaart, A.: Adaptive Bayesian credible bands in regression with a Gaussian process prior. Sankhya A **82**(2), 386–425 (2020). https://doi.org/10.1007/s13171-019-00185-0

Sniekers, S., van der Vaart, A.: Credible sets in the fixed design model with Brownian motion prior. J. Statist. Plann. Inference **166**, 78–86 (2015). https://doi.org/10.1016/j.jspi.2014.07.008

Stuart, A.M.: Inverse problems: A bayesian perspective. Acta Numerica **19**, 451–559 (2010). https://doi.org/10.1017/S0962492910000061

Szabó, B., van der Vaart, A., van Zanten, H.: Empirical bayes scaling of gaussian priors in the white noise model. Electron. J. Statist. **7**, 991–1018 (2013)

Triebel, H.: Function Spaces and Wavelets on Domains. EMS tracts in mathematics. European Mathematical Society (2008)

Triebel, H.: Theory of Function Spaces. Modern Birkhäuser Classics. Springer (2010)

van der Vaart, A.W., Wellner, J.A.: Weak Convergence and Empirical Processes. Springer Series in Statistics, p. 508. Springer (1996). https://doi.org/10.1007/978-1-4757-2545-2. With applications to statistics. http://dx.doi.org/10.1007/978-1-4757-2545-2

van der Vaart, A.: Bayesian regularization. In: Proceedings of the International Congress of Mathematicians. Volume IV, pp. 2370–2385. Hindustan Book Agency, New Delhi (2010)

Appendix A Galerkin Projection

In this section we state for easy reference two results on the Galerkin method. Proofs may be found in the literature, or exactly in the present form in Gugushvili et al. (2020).

Lemma 6.4 *If V_j is a finite-dimensional space as in Assumption 2.2 such that (6) and (7) hold, then, for $P_j : H_0 \to V_j$ the orthogonal projection onto V_j, and $0 \leq s, t < S$,*

$$\|f - P_j f\|_{-t} \lesssim \delta(j,t)\delta(j,s)\|f\|_s, \qquad f \in H_0, \tag{A1}$$

$$\|g\|_s \lesssim \frac{1}{\delta(j,s)\delta(j,t)}\|g\|_{-t}, \qquad g \in V_j. \tag{A2}$$

Lemma 6.5 *If V_j is a finite-dimensional space as in Assumption 2.2 such that (6) and (7) hold, and $A : H_0 \to G$ is a bounded linear operator satisfying $\|Af\| \simeq \|f\|_{-\gamma}$ for every $f \in H_0$, then the norms of the operators $R_j : G \to H_0$ and $R_j A : H_0 \to H_0$ satisfy*

$$\|R_j\| \lesssim_A \frac{1}{\delta(j,\gamma)}, \tag{A3}$$

$$\|R_j A\| \lesssim_A 1. \tag{A4}$$

Furthermore, for $f \in H_s$ the Galerkin solution $f^{(j)} \in V_j$ to Af satisfies

$$\|f^{(j)} - f\| \lesssim_A \delta(j,s) \|f\|_s. \tag{A5}$$

Dong Yan and Aad van der Vaart
DIAM, TU Delft, Mekelweg 4,
Delft 2628 CD, Netherlands
E-mail: a.w.vandervaart@tudelft.nl

Shota Gugushvili
Biometris, Plant Sciences Group,
Wageningen University & Research,
P.O. Box 16, Wageningen 6700 AA,
The Netherlands
E-mail: shota.gugushvili@wur.nl