Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2020, Vol. 56, No. 3, 2081–2107
https://doi.org/10.1214/19-AIHP1029

www.imstat.org/aihp

Bayesian linear inverse problems in regularity scales

Shota Gugushvili[a], Aad van der Vaart[b] and Dong Yan[b]

[a]*Biometris, Wageningen University & Research, The Netherlands. E-mail: gugushvili@gmail.com*
[b]*Mathematical Institute, Leiden University, The Netherlands. E-mail: avdvaart@math.leidenuniv.nl; d.yan@math.leidenuniv.nl*

Received 22 March 2018; revised 2 September 2019; accepted 7 October 2019

Abstract. We obtain rates of contraction of posterior distributions in inverse problems defined by scales of smoothness classes. We derive abstract results for general priors, with contraction rates determined by 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. The proofs are based on general testing and approximation arguments, without explicit calculations on the posterior distribution. We are thus not restricted to priors based on the singular value decomposition of the operator. We illustrate the results with examples of inverse problems resulting from differential equations.

Résumé. Nous obtenons le taux de contraction des distributions a posteriori dans les problèmes inverses définis par des classes d'échelles de régularité. Nous obtenons des résultats abstraits pour des lois a posteriori générales déterminées par des approximations de type Galerkin. Le taux dépend du niveau de concentration de la loi a priori au voisinage des vrais paramètres et de la probabilité a priori de l'ensemble des paramètres avec approximation Galerkin inférieure. Nous appliquons le résultat abstrait a trois types de lois a priori : au cas des séries aléatoires non conjuguées, montrant ainsi que ces mesures a priori donnent une récupération presque optimale sous des hypothèses assez générales ; au cas des mesures gaussiennes ; et au cas des mélanges de gaussiennes, où il est également démontré que ces derniers sont presque optimaux et adaptatifs. Les preuves sont basées sur des tests statistiques et arguments d'approximation, sans calculs explicites sur la loi a posteriori. Nous ne sommes donc pas limités aux lois a priori basées sur la décomposition en valeurs singulières de l'opérateur. Nous illustrons les résultats par des exemples de problèmes inverses résultant d'équations différentielles.

MSC: Primary 62G20; secondary 35R30

Keywords: Adaptive estimation; Gaussian prior; Hilbert scale; Linear inverse problem; Nonparametric Bayesian estimation; Posterior contraction rate; Random series prior; Regularity scale; White noise

1. Introduction

In a statistical inverse problem one observes a noisy version of a transformed signal Af and wishes to recover the unknown parameter f. In this paper we consider linear inverse problems of the type

$$Y^{(n)} = Af + \frac{1}{\sqrt{n}}\xi, \tag{1.1}$$

where $A : H \to G$ is a known bounded linear operator between separable Hilbert spaces H and G, and ξ is a stochastic 'noise' process, which is multiplied by the scalar 'noise level' $n^{-1/2}$. The problem is to infer f from the observation $Y^{(n)}$. To this purpose we assume that the *forward operator* A is injective, but we shall be interested in the case that the inverse A^{-1}, defined on the range of A is not continuous (or equivalently the range of A is not closed in G). The problem of recovering f from $Y^{(n)}$ is then *ill-posed*, and *regularization* methods are necessary in order to 'invert' the operator A. These consist of constructing an approximation to A^{-1}, with natural properties such as boundedness and whose domain includes the data $Y^{(n)}$, and applying this to $Y^{(n)}$. By the discontinuity of the inverse A^{-1}, the noise present in the observation is necessarily multiplied, and regularization is focused on balancing the error in the approximation to

A^{-1} to the size of the magnified noise, in order to obtain a solution that is as close as possible to the true signal f. In this article we study this through the convergence rates of the regularized solutions to a true parameter f, as $n \to \infty$, i.e. as the noise level tends to zero. In particular, we consider contraction rates of posterior distributions resulting from a Bayesian approach to the problem.

There is a rich literature on inverse problems. The case that the noise ξ is a bounded *deterministic* perturbation, has been particularly well studied, and various general procedures and methods to estimate the convergence rates of regularized solutions have been proposed. See the monographs [14,29]. The case of stochastic noise is less studied, but is receiving increasing attention. In this paper we shall be mostly interested in the case that ξ is white noise indexed by the Hilbert space G, i.e. the *isonormal process*, which is characterized by the requirement that $\langle \xi, w\rangle_G$ is a zero-mean Gaussian variable with variance $\|w\|_G^2$, for every $w \in G$, where $\langle \cdot, \cdot \rangle_G$ and $\| \cdot \|_G$ are the inner product and norm in in G. Actually the isonormal process cannot be realized as a Borel-measurable map into G, and hence we need to interpret (1.1) in a generalized sense. In our measurement model the observation $Y^{(n)}$ will be a stochastic process $(Y^{(n)}(w) : w \in G)$ such that

$$Y^{(n)}(w) = \langle Af, w\rangle_G + \frac{1}{\sqrt{n}}\xi(w), \quad w \in G, \tag{1.2}$$

where $\xi = (\xi(w) : w \in G)$ is the iso-normal process, i.e. a zero-mean Gaussian process with covariance function $\mathbb{E}(\xi(w_1)\xi(w_2)) = \langle w_1, w_2\rangle_G$. The processes $Y^{(n)}$ and ξ are viewed as measurable maps in the *sample space* $\mathbb{R}^G$, with its product σ-field. Statistical sufficiency considerations show that the observation can also be reduced to the vector $(Y^{(n)}(w_1), Y^{(n)}(w_2), \ldots)$, which takes values in the sample space $\mathbb{R}^\infty$, for any orthonormal basis $(w_i)_{i\in\mathbb{N}}$ of G. Since in that case the variables $\xi(w_1), \xi(w_2), \ldots$ are stochastically independent standard normal variables, the coordinates $Y^{(n)}(w_i)$ of this vector are independent random variables with normal distributions with means $\langle Af, w_i\rangle_G$ and variance $1/n$. This is known as the *Gaussian sequence model* in statistics, albeit presently the 'drift function' Af involves the operator A. See [4,27] and references therein.

An alternative method to give a rigorous interpretation to white noise ξ, is to embed G into a bigger space in which ξ can be realized as a Borel measurable map, or to think of ξ as a cylindrical process. See e.g., [51]. For G a set of functions on an interval, one can also realize ξ as a stochastic integral relative to Brownian motion, which takes its values in the 'abstract Wiener space' attached to G. We shall not follow these constructions, as they imply the stochastic process version (1.2), which is easier to grasp and will be the basis for our proofs.

It is also possible to consider the model (1.1) with a noise variable ξ that takes its values inside the Hilbert space G. In this paper we briefly note some results on this 'coloured noise' model, but our main focus is model (1.2).

The study of statistical (nonparametric) linear inverse problems was initiated by Wahba in 1970s in [60]. The 1990s paper [12] used wavelet shrinkage methods, while around 2000, the authors of [9] investigated (1.1) in the linear partial differential equations setting, while a systematic study of Gaussian sequence models was presented in [8]. A review of work until 2008 is given in [7]. The connection of regularization methods to the Bayesian approach was recognized early on. However, the study of the recovery properties of posterior distributions was started only in [31,32]. A review of the Bayesian approach to inverse problems, with many examples, is given in [52].

In the present paper we follow the Bayesian approach. This consists of putting a probability measure on f, the *prior*, that quantifies one's prior beliefs on f, and next, after collecting the data, updating the prior to the *posterior* measure, through Bayes' formula. As always, this is the conditional distribution of f given $Y^{(n)}$ in the model, where f follows the prior measure Π, a Borel probability distribution on H, and given f the variable $Y^{(n)}$ has the conditional distribution on $\mathbb{R}^G$ determined by (1.2). For a given $f \in H$ the latter conditional distribution is dominated by its distribution under $f = 0$. The Radon–Nikodym densities $y \mapsto p_f^{(n)}(y)$ of the conditional distributions can be chosen jointly measurable in (y, f), and by Bayes' formula the posterior distribution of f is the Borel measure on H given by

$$\Pi_n\big(f \in B|Y^{(n)}\big) = \frac{\int_B p_f^{(n)}(Y^{(n)})\, d\Pi(f)}{\int p_f^{(n)}(Y^{(n)})\, d\Pi(f)}. \tag{1.3}$$

The form of the densities $p_f^{(n)}$ is given by the (abstract) Cameron–Martin formula, but will not be needed in the following (see Lemma 9.1). 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 'center' of the posterior distribution, but also provides a way to conduct uncertainty quantification, through the spread in the posterior distribution.

One hopes that as the noise level tends to zero, i.e. $n \to \infty$, the posterior measures (1.3) will contract to a Dirac measure at f_0 if in reality $Y^{(n)}$ is generated through the model (1.2) with $f = f_0$. We shall be interested in the *rate* of

contraction. Following [16,18,19] 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 $n \to \infty$,

$$\Pi_n\big(f : \|f - f_0\|_H > M\varepsilon_n | Y^{(n)}\big) \overset{\mathbb{P}_{f_0}^{(n)}}{\to} 0. \tag{1.4}$$

We shall use the general approach to establishing such rates of contraction, based on a prior mass condition and testing condition, explained in [18]. This was adapted to the inverse setup in [30], who in a high level result show how to obtain an inverse rate from a rate in the forward problem and a continuity modulus of the restriction of the operator to suitable sets on which the posterior concentrates.

Much of the existing work on statistical inverse problems is based on the singular value decomposition (SVD) of the operator A; see, e.g., [7]. When A is compact, the operator A^*A, where A^* is the adjoint of A, can be diagonalized with respect to an orthonormal *eigenbasis*, with eigenvalues tending to zero. The observation $Y^{(n)}$ can then be reduced to noisy observations on the Fourier coefficients of Af in the eigenbasis, which are multiples of the Fourier coefficients of f, and the problem is to recover the latter. In the frequentist setup thresholding or other regularization methods can be applied to reduce the weight of estimates on coefficients corresponding to smaller eigenvalues, in which the noise will overpower the signal. In the Bayesian setup one may design a prior by letting the Fourier coefficients be (independent) random variables, with smaller variances for smaller eigenvalues. These singular value methods have several disadvantages, as pointed out in [11,12]. First, the eigenbasis functions might not be easy to compute. Second, and more importantly, these functions are directly linked to the operator A, and need not be related to the function space (smoothness class) that is thought to contain the true signal f. Consequently, the parameter of interest f may not have a simple, parsimonious representation in the eigenbasis expansion, see [12]. Furthermore, it is logical to consider the series expansion of the signal f in other bases than the eigenbasis, for instance, in the situation that one can only measure noisy coefficients of the signal f in a given basis expansion, due to a particular experimental setup. See [20,40] for further discussion.

One purpose of the present paper is to work with priors that directly relate to common bases (e.g., splines or wavelets bases) and function spaces, rather than to the operator through its singular value decomposition. We succeed in this aim under the assumption that the operator A respects a given scale of function spaces. In Section 2 we first set up such a scale in an abstract manner, and then introduce a smoothing assumption on the operator A in terms of this scale. Next in Section 4–Section 7 we consider priors defined in terms of the scale, rather than the operator. Thus operator and prior are assumed related, but only indirectly, through the scale.

A canonical example are Sobolev spaces, with the operator A being an integral operator. This Sobolev space setup with wavelet basis was investigated in [11,12]. In deterministic inverse problems, a more general setup, considering A that acts along nested Hilbert spaces, *Hilbert scales*, was initiated by Natterer in [42] and further developed in, amongst others, [26,39,40]. In the Bayesian context Hilbert scales were used in [15], under the assumption that the noise ξ is a proper Gaussian element in G, and in [1], but under rather intricate assumptions.

A second purpose of the present paper is to allow priors that are not necessarily Gaussian. In the linear inverse problem Gaussian priors are easy, as they lead to Gaussian posterior distributions, which can be studied by direct means. Most of the results on Bayesian inverse problems fall in this framework [1,15,31,32], exceptions being [49] and [30].

Thus in this paper we investigate a Bayesian approach to linear inverse problems that is not based on the SVD and does cover non-conjugate, non-Gaussian priors.

The white noise model represents a limiting case (in an appropriate sense) of the inverse regression model

$$Y_i = Af(x_i) + z_i, \quad i = 1, \ldots, n,$$

where z_i are independent standard normal random variables. Insights gained in inverse problems in the white noise model shed light on the behaviour of statistical procedures in the inverse regression model, which is the one encountered in actual practice, as the signal f can be typically observed only on a discrete grid of points. It is next at times possible to extend theoretical results obtained in the white noise setting to those in the inverse regression setting.

The paper is organized as follows. In Section 2 we introduce in greater detail our setup along with the assumptions that will be used in this article. We also present some examples for illustration. Next we present a general contraction theorem in Section 3, and apply this to two main special cases, series priors and Gaussian priors in Section 4 and Section 6. The section on Gaussian priors is preceded by a discussion in Section 5 of Hilbert scales generated by unbounded operators, which next serve as inverse covariance operators. Since the simple Gaussian prior is not fully adaptive, we introduce Gaussian mixture priors to obtain adaptation in Section 7. In Section 8 we discuss several extensions of the present work. Section 9 contains the proofs, and an appendix presents background to some of the tools we need in the proofs.

Notation 1.1. 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 formalize the structure of the inverse problem that will be worked out in this article.

Smoothness scales
The function f in (1.1) is an element of a Hilbert space H. We embed this space as the space $H = H_0$ in a 'scale of smoothness classes', defined 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:

(i) For $s < t$ the space H_t is a dense subspace of H_s and $\|f\|_s \lesssim \|f\|_t$, for $f \in H_t$.
(ii) For $s \geq 0$ and $f \in H_0$ viewed as element of $H_{-s} \supset H_0$,

$$\|f\|_{-s} = \sup_{\|g\|_s \leq 1} \langle f, g \rangle_0, \quad f \in H_0. \tag{2.1}$$

The notion of scales of smoothness classes is standard in the literature on inverse problems. In the preceding definition we have stripped it to the bare essentials needed in our general result on posterior contraction. Concrete examples, as well as more involved structures such as Hilbert scales, are introduced below.

Remark 2.2. We may also start with Hilbert spaces H_s for $s \geq 0$ only satisfying (i) and next define H_{-s} for $s \geq 0$ to be the dual space H_s^*. We next embed H_{-s} for $s \geq 0$ in H_0 through identifying H_0 and its dual $H_0^* \subset H_s^*$ (the restriction of a continuous linear map from H_0 to $\mathbb{R}$ to domain H_s is contained in H_s^*), and the *norm duality* (2.1) will be automatic.

It is important that we only 'flip' H_0 in this contruction. Every Hilbert space H_s can be identified with its dual H_s^* in the usual way, but this involves the inner product in H_s, and is different from the identification of H_s^* with the 'bigger space' H_{-s} for $s \neq 0$.

More generally (2.1) is implied if, for $s > 0$, the space H_{-s} can be identified with the dual space H_s^* of H_s and the embedding $\iota : H_0 \to H_{-s}$ is the adjoint of the embedding $\iota : H_s \to H_0$, after the usual identification of H_0 and its dual space H_0^*. (The three nested spaces $H_{-s} \supset H_0 \supset H_s$ then form a 'Gelfand triple'.) Indeed, by definition the image $\iota^* f$ of $f \in H_0 = H_0^*$ under the adjoint $\iota^* : H_0^* \to H_s^*$ is the map $g \mapsto (\iota^* f)(g) = \langle \iota g, f \rangle_0 = \langle g, f \rangle_0$ from $H_s \to \mathbb{R}$. The norm of this map as an element of H_s^* is $\sup_{\|g\|_s \leq 1} (\iota^* f)(g)$. The norm duality follows if $\iota^* f$ is identified with the element $f \in H_0 \subset H_{-s}$.

We assume that the smoothness scale allows good finite-dimensional approximations, as in the following condition.

Assumption 2.3 (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

$$\inf_{g \in V_j} \|f - g\|_0 \lesssim \delta(j, s) \|f\|_s, \tag{2.2}$$

$$\|g\|_s \lesssim \frac{1}{\delta(j, s)} \|g\|_0, \quad \forall g \in V_j. \tag{2.3}$$

This assumption is also common in the literature on inverse problems. The two inequalities (2.2) and (2.3) are known as inequalities of Jackson and Bernstein type, respectively, see, e.g., [5]. The approximation property (2.2) shows that 'smooth elements' $f \in H_s$ are well approximated in $\|\cdot\|_0$ 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. Naturally one expects the numbers $\delta(j, s)$ that control the approximation to be decreasing in both j and s. In our examples we shall mostly have polynomial dependence $\delta(j, s) = j^{-s/d}$, in the case that H_0 consists of functions on a d-dimensional domain. The stability property (2.3) quantifies the smoothness norm of the projections in terms of the approximation numbers. Both conditions are assumed up to a maximal order of smoothness $S > 0$, and it follows from (2.3) that V_j must be contained in the space H_S.

The approximation property (2.2) can also be stated in terms of the 'approximation numbers' of the canonical embedding $\iota : H_s \to H_0$. The jth *approximation number* of a general bounded linear operator $T : G \to H$ between normed spaces is defined as

$$a_j(T : G \to H) = \inf_{U : \operatorname{Rank} U < j} \sup_{f : \|f\|_G \leq 1} \big\| (T - U) f \big\|_H, \tag{2.4}$$

where the infimum is taken over all linear operators $U : G \to H$ of rank less than j. It is immediate from the definitions that the numbers $\delta(j, s)$ in (2.2) can be taken equal to the approximation numbers $a_j(\iota : H_s \to H_0)$. The set of approximation numbers $a_j(\iota : H_{s+t} \to H_t)$ of the canonical embedding describes many characteristics of the smoothness scale $(H_s)_{s\in\mathbb{R}}$. We give a brief discussion in Appendix B.

Example 2.4 (Sobolev classes). The most important examples of smoothness classes satisfying Definition 2.1 are fractional Sobolev spaces on a bounded domain $\mathcal{D} \subset \mathbb{R}^d$. For a natural number $s \in \mathbb{N}$ the Sobolev space of order s can be defined by

$$H_s(\mathcal{D}) = W^{s,2}(\mathcal{D}) := \left\{ f \in \mathcal{D}'(\mathcal{D}) : \|f\|_s := \sum_{|\alpha|\le s} \left\| D^\alpha f \right\|_{L^2(\mathcal{D})} < \infty \right\}.$$

Here $\mathcal{D}'(\mathcal{D})$ is the space of generalized functions on $\mathcal{D}$ (distributions), i.e. the topological dual space of the space $C_c^\infty(\mathcal{D})$ of infinitely differentiable functions with compact support in $\mathcal{D}$; the sum ranges over the multi-indices $\alpha = (\alpha_1, \ldots, \alpha_d) \in (\{0\} \cup \mathbb{N})^d$ with $|\alpha| := \sum_{i=1}^s \alpha_i \le s$; and D^α is the differential operator

$$D^\alpha := \frac{\partial^{\alpha_1} \partial^{\alpha_2} \cdots \partial^{\alpha_d}}{\partial x_1^{\alpha_1} x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}.$$

The definition can be extended to $s \in \mathbb{R}\backslash\mathbb{N}$ in several ways. All constructions are equivalent to the Besov space $B^s_{2,2}(\mathcal{D})$, see [55,56].

It is well known that the approximation numbers of the scale of Sobolev spaces satisfy Assumption 2.3 with $\delta(j, t) = j^{-t/d}$, see [25].

Example 2.5 (Sequence spaces). Suppose $(\phi_i)_{i\in\mathbb{N}}$ is a given orthonormal sequence in a given Hilbert space H, and $1 \le b_i \uparrow \infty$ is a given sequence of numbers. For $s \ge 0$, 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

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

Then $H_0 = H$ is embedded in H_s, for every $s > 0$, and the norms $\|f\|_s$ are increasing in s. Every space H_s is a Hilbert space; in fact H_s is isometric to H_0 under the map $(f_i) \to (f_i b_i^s)$, where we have identified the series with their coefficients for simplicity of notation.

For $s < 0$, we equip the elements $f = \sum_{i\in\mathbb{N}} f_i \phi_i$ 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 metric completion of H under this norm. The space H_s is isometric to the set of all sequences $(f_i)_{i\in\mathbb{N}}$ with $\sum_{i\in\mathbb{N}} f_i^2 b_i^{2s} < \infty$ equipped with the norm given on the right hand side of the preceding display, but the series $\sum_{i\in\mathbb{N}} f_i \phi_i$ may not possess a concrete meaning, for instance as a function if H is a function space.

By Parseval's identity the inner product on $H = H_0$ is given by $\langle f, g\rangle_0 = \sum_{i\in\mathbb{N}} f_i g_i$, and the norm duality (2.1) follows with the help of the Cauchy–Schwarz inequality.

The natural approximation spaces for use in Assumption 2.3 are $V_j = \mathsf{Span}(\phi_i : i < j)$. Inequalities (2.2)–(2.3) are satisfied with the approximation numbers taken equal to $\delta(j, t) = b_j^{-t}$.

The forward operator A in the model (1.1) is a bounded linear operator $A : H \to G$ between the separable Hilbert spaces H and G, and is assumed to be smoothing. The following assumption makes this precise. This assumption is satisfied in many examples and is common in the literature (for instance [11,21,42]).

In Definition 2.1 the space H is embedded as $H = H_0$ in the smoothness scale $(H_s)_{s\in\mathbb{R}}$ and hence has norm $\|\cdot\|_0$.

Assumption 2.6 (Smoothing property of A). For some $\gamma > 0$ the operator $A : H_{-\gamma} \to G$ is injective and bounded and, for every $f \in H_0$,

$$\|Af\| \simeq \|f\|_{-\gamma}. \tag{2.5}$$

Example 2.7 (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

$$\lambda_i \simeq i^{-2\gamma}. \tag{2.6}$$

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 2.5, then (2.5) is satisfied.

Indeed, we can write A in polar decomposition as $Af = U(A^*A)^{1/2}f$, for a partial isometry $U : \text{Range}(A) \to G$, and then have $Af = U\sum_i f_i\sqrt{\lambda_i}\phi_i$, so that $\|Af\| = \|\sum_i f_i i^{-\gamma}\phi_i\|_0 \simeq \|f\|_{-\gamma}$.

Thus constructions using the singular value decomposition of A can always be accommodated in the more general setup described in the preceding.

Example 2.8 (Poisson equation). The operator $A : L^2(0,1) \to L^2(0,1)$ defined by the differential equation $(Af)'' = f$ with Dirichlet boundary conditions $Af(0) = Af(1) = 0$ is smoothing with $\gamma = 2$ in the Sobolev scale given in Example 2.4. This is shown in Sections 10.4 and 11.2 in [24].

Example 2.9 (Symm's equation [29]). Consider the Laplace equation $\Delta u = 0$ in a bounded set $\Omega \subset \mathbb{R}^2$ with boundary condition $u = g$ on the boundary $\partial\Omega$. The singular layer potential, a boundary integral

$$u(x) = -\frac{1}{\pi}\int_{\partial\Omega} h(y)\ln|x-y|\,ds(y), \quad x \in \Omega,$$

solves the boundary value problem if and only if the density h, belonging to the space $C(\partial\Omega)$ of continuous functions on $\partial\Omega$, solves *Symm's equation*

$$-\frac{1}{\pi}\int_{\partial\Omega} h(y)\ln|x-y|\,ds(y) = g(x), \quad x \in \partial\Omega. \tag{2.7}$$

Assume the boundary $\partial\Omega$ has a parametrization of the form $\{\rho(s), s \in [0,2\pi]\}$, for some 2π-periodic analytic function $\rho : [0,2\pi] \to \mathbb{R}^2$ such that $|\dot\rho(s)| > 0$ for all s. Then Symm's equation takes the following form,

$$Af(z) := -\frac{1}{\pi}\int_0^{2\pi} \log\bigl|\rho(z) - \rho(s)\bigr| f(s)\,ds = g\bigl(\rho(z)\bigr), \quad z \in [0,2\pi],$$

where $f(s) = h(\rho(s))|\dot\rho(s)|$. It is shown in Theorem 3.18 of [29] that the operator A satisfies (2.5) with $\gamma = 1$ and $(H_s)_{s\in\mathbb{R}}$ the periodic Sobolev spaces on $[0,2\pi]$.

Example 2.10 (Radon transform). Inverting the Radon transform was recently studied in the Bayesian framework by [41], who studied the posterior distribution of smooth functionals for general Gaussian priors, but not the inversion of the whole function. The SVD of the transform is known (see [61,62]) and can be used to put the problem in our framework, in the spirit of Example 2.7. This would give a Bayesian parallel to the rate results in [28]. We do not know if other standard smoothness scales could be used within our framework as well.

Remark 2.11. For all our purposes the smoothing condition (2.5) can be relaxed to (A.6)–(A.7). This relaxation covers the situation where there exists an operator A_0 that satisfies (2.5) and is a 'version' of A in that the two operators possess a common inverse, such as when A and A_0 are defined to solve a differential equation with different boundary conditions. Lemma A.3 shows that the relaxed version of the smoothing condition is then satisfied by the map $f \mapsto [Af]$ of f in the class of Af in the quotient space $G/R(A - A_0)$.

3. General result

In this section we present a general theorem on posterior contraction. We form the posterior distribution $\Pi_n(\cdot|Y^{(n)})$ as in (1.3), given a prior Π on the space $H = H_0$ and an observation $Y^{(n)}$, whose conditional distribution given f is determined by the model (1.2). We study this random distribution under the assumption that $Y^{(n)}$ follows the model (1.2) 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 result is based on an extension of the testing approach of [19] to the inverse problem (1.2). It resembles the approach in [44,45,49,54] or [30], except that the inverse problem is handled with the help of the Galerkin method,

which is a well known strategy in numerical analysis to solve the operator equation $y = Af$ for f, in particular for differential and integral operators. The Galerkin method has several variants, which are useful depending on the properties of the operator involved. Here we use the least squares method, which is of general application; for other variants and background, see e.g., [29]. In Appendix A we give a self-contained derivation of the necessary inequalities, exactly in our framework. We note that the Galerkin method only appears as a tool to state and derive a posterior contraction rate. In our context it does not enter into the solution of the inverse problem, which is achieved through the Bayesian method.

Let $W_j = AV_j \subset G$ be the image under A of a finite-dimensional approximation space V_j linked to the smoothness scale $(H_s)_{s\in\mathbb{R}}$ as in Assumption 2.3, and let $Q_j : G \to W_j$ be the orthogonal projection onto W_j. If $A : H \to G$ is injective, then A is a bijection between the finite-dimensional vector spaces V_j and W_j, and hence for every $f \in H$ there exists $f^{(j)} \in V_j$ such that $Af^{(j)} = Q_j Af$. The element $f^{(j)}$ is called the *Galerkin solution* to Af in V_j. By the projection theorem in Hilbert spaces it is characterized by the property that $f^{(j)} \in V_j$ together with the orthogonality relations

$$\langle Af^{(j)}, w\rangle = \langle Af, w\rangle, \quad w \in W_j. \tag{3.1}$$

The idea of the Galerkin inversion is to project the (complex) object Af onto the finite-dimensional space W_j, and next find the inverse image $f^{(j)}$ of the projection, in the finite-dimensional space V_j, as in the diagram: Clearly the Galerkin solution to an element $f \in V_j$ is f itself, but in general $f^{(j)}$ is an approximation to f, which will be better for increasing j, but increasingly complex. The following theorem uses a dimension $j = j_n$ that balances approximation to complexity, where the complexity is implicitly determined by a testing criterion.

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 $\gamma > 0$, and let $f^{(j)}$ denote the Galerkin solution to Af relative to linear subspaces V_j associated to $(H_s)_{s\in\mathbb{R}}$ as in Assumption 2.3. Let $f_0 \in H_\beta$ for some $\beta \in (0, S)$, 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*

$$j_n \leq cn\varepsilon_n^2, \tag{3.2}$$

$$\eta_n \geq \frac{\varepsilon_n}{\delta(j_n, \gamma)}, \tag{3.3}$$

$$\eta_n \geq \delta(j_n, \beta). \tag{3.4}$$

Consider prior probability distributions Π on H_0 satisfying

$$\Pi\big(f : \|Af - Af_0\| < \varepsilon_n\big) \geq e^{-n\varepsilon_n^2}, \tag{3.5}$$

$$\Pi\big(f : \big\|f^{(j_n)} - f\big\|_0 > \eta_n\big) \leq e^{-4n\varepsilon_n^2}. \tag{3.6}$$

Then the posterior distribution in the model (1.2) contracts at the rate η_n at f_0, i.e. for a sufficiently large constant M we have $\Pi_n(f : \|f - f_0\|_0 > M\eta_n | Y^{(n)}) \to 0$, in probability under the law of $Y^{(n)}$ given by (1.2) with $f = f_0$.

Proof. The Kullback–Leibler divergence and variation between the distributions of $Y^{(n)}$ under two functions f and f_0 are given by $n\|Af - Af_0\|^2/2$ and twice this quantity, respectively. (At a referee's request, a proof is provided in Lemma 9.1.) Therefore the neighbourhoods $B_{n,2}(f_0, \varepsilon)$ in (8.19) of [19] contain the ball $\{f \in H_0 : \|Af - Af_0\| \leq \varepsilon\}$. By assumption (3.5) this has prior mass at least $e^{-n\varepsilon_n^2}$.

Because the quotient of the left sides of (3.5) and (3.6) is $o(e^{-2n\varepsilon_n^2})$, the posterior probability of the set $\{f : \|f^{(j_n)} - f\|_0 > \eta_n\}$ tends to zero, by Theorem 8.20 in [19].

By a variation of Theorem 8.22 in [19] it is now sufficient to show the existence of tests τ_n such that, for some $M > 0$,

$$P_{f_0}^{(n)}\tau_n \to 0, \qquad \sup_{\substack{f:\|f-f_0\|_0 > M\eta_n,\\ \|f^{(j_n)}-f\|_0 \leq \eta_n}} P_f^{(n)}(1 - \tau_n) \leq e^{-4n\varepsilon_n^2}.$$

Indeed, in the case that the prior mass condition (8.20) in Theorem 8.22 of [19] can be strengthened to (8.22), as is the case in our setup in view of (3.5), it suffices to verify (8.24) only for a single value of j. Furthermore, we can apply Theorem 8.22 with the metrics $d_n(f, g) = \|f - g\|_0\varepsilon_n/\eta_n$ in order to reduce the restriction $d_n(\theta, \theta_{n,0}) > M\varepsilon_n$ to $\|f - f_0\|_0 > M\eta_n$.

Fix any orthonormal basis $(\bar\psi_i)_{i<j}$ of $W_j = AV_j$ and define

$$\bar Y_j = \sum_{i<j} Y^{(n)}_{\bar\psi_i}\bar\psi_i = \sum_{i<j}\langle Af,\bar\psi_i\rangle\bar\psi_i + \frac{1}{\sqrt n}\sum_{i<j}\xi_{\bar\psi_i}\bar\psi_i$$
$$= Q_jAf + \frac{1}{\sqrt n}\bar\xi_j,$$

where $\bar\xi_j := \sum_{i<j}\xi_{\bar\psi_i}\bar\psi_i$. The latter is "standard normal in the finite-dimensional space W_j": because $(\xi_{\bar\psi_i})_{i<j}$ are i.i.d. standard normal variables, the variable $\langle\bar\xi_j, w\rangle = \sum_{i<j}\xi_{\bar\psi_i}\langle\bar\psi_i, w\rangle$ is $N(0, \|Q_jw\|^2)$-distributed, for every $w\in G$.

Let the operator $R_j : G\mapsto V_j$ be defined as $R_j = A^{-1}Q_j$, where A^{-1} is the inverse of A, which is well defined on the range $W_j = AV_j$ of Q_j. Then by definition R_jAf is equal to the Galerkin solution $f^{(j)}$ to Af. By the preceding display $R_j\bar Y_j$ is a well-defined Gaussian random element in V_j, satisfying

$$R_j\bar Y_j = f^{(j)} + \frac{1}{\sqrt n}R_j\bar\xi_j. \tag{3.7}$$

The variable $R_j\bar\xi_j$ is a Gaussian random element in V_j with strong and weak second moments

$$\mathrm E\|R_j\bar\xi_j\|_0^2 \le \|R_j\|^2\mathrm E\|\bar\xi_j\|^2 = \|R_j\|^2\mathrm E\sum_{i<j}\xi^2_{\bar\psi_i} = \|R_j\|^2(j-1) \lesssim \frac{j}{\delta(j,\gamma)^2},$$

$$\sup_{\|f\|_0\le1}\mathrm E\langle R_j\bar\xi_j, f\rangle_0^2 = \sup_{\|f\|_0\le1}\mathrm E\big\langle\bar\xi_j, R_j^*f\big\rangle^2 = \sup_{\|f\|_0\le1}\big\|Q_jR_j^*f\big\|^2 \le \big\|R_j^*\big\|^2 \lesssim \frac{1}{\delta(j,\gamma)^2}.$$

In both cases the inequality on $\|R_j\| = \|R_j^*\|$ at the far right side follows from (A.3).

The first inequality implies that the first moment $\mathrm E\|R_j\bar\xi_j\|_0$ of the variable $\|R_j\bar\xi_j\|_0$ is bounded above by $\sqrt j/\delta(j,\gamma)$. By Borell's inequality (e.g. Lemma 3.1 in [37] and subsequent discussion), applied to the Gaussian random variable $R_j\bar\xi_j$ in H_0, we see that there exist positive constants a and b such that, for every $t>0$,

$$\Pr\left(\|R_j\bar\xi_j\|_0 > t + a\frac{\sqrt j}{\delta(j,\gamma)}\right) \le e^{-bt^2\delta(j,\gamma)^2}.$$

For $t = 2\sqrt n\eta_n/\sqrt b$ and η_n, ε_n and j_n satisfying (3.2), (3.3) and (3.4) this yields, for some $a_1>0$,

$$\Pr\big(\|R_{j_n}\bar\xi_{j_n}\|_0 > a_1\sqrt n\eta_n\big) \le e^{-4n\varepsilon_n^2}. \tag{3.8}$$

We apply this to bound the error probabilities of the tests

$$\tau_n = 1\big\{\|R_{j_n}\bar Y_{j_n} - f_0\|_0 \ge M_0\eta_n\big\}, \tag{3.9}$$

where M_0 is a given constant, to be determined.

Under f_0, the decomposition (3.7) is valid with $f=f_0$, and hence $R_j\bar Y_j - f_0 = n^{-1/2}R_j\bar\xi_j + f_0^{(j)} - f_0$. By the triangle inequality it follows that $\tau_n = 1$ implies that $n^{-1/2}\|R_{j_n}\bar\xi_{j_n}\|_0 \ge M_0\eta_n - \|f_0^{(j)} - f_0\|_0$. By (A.5) the assumption that $f_0\in H_\beta$ implies that $\|f_0^{(j)} - f_0\|_0 \le M_1\delta(j,\beta)$, for some M_1, which at $j=j_n$ is further bounded by $M_1\eta_n$, by assumption (3.4). Hence the probability of an error of the first kind satisfies

$$P^{(n)}_{f_0}\tau_n \le \Pr\left(\frac{1}{\sqrt n}\|R_{j_n}\bar\xi_{j_n}\|_0 \ge (M_0-M_1)\eta_n\right).$$

For $M_0 - M_1 > a_1$, the right side is bounded by $e^{-4n\varepsilon_n^2}$, by (3.8).

Under f the decomposition (3.7) gives that $R_j\bar Y_j - f_0 = n^{-1/2}R_j\bar\xi_j + f^{(j)} - f_0$. By the triangle inequality $\tau_n = 0$ implies that $n^{-1/2}\|R_{j_n}\bar\xi_{j_n}\|_0 \ge \|f^{(j_n)} - f_0\|_0 - M_0\eta_n$. For f such that $\|f-f_0\|_0 > M\eta_n$ and $\|f - f^{(j_n)}\|_0 \le \eta_n$, we have $\|f^{(j_n)} - f_0\| \ge (M-1)\eta_n$. Hence the probability of an error of the second kind satisfies

$$P^{(n)}_f(1-\tau_n) \le \Pr\left(\frac{1}{\sqrt n}\|R_{j_n}\bar\xi_{j_n}\|_0 \ge (M-1-M_0)\eta_n\right).$$

For $M-1-M_0 > a_1$, this is bounded by $e^{-4n\varepsilon_n^2}$, by (3.8).

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 (3.5) is the usual *prior mass condition* for the 'direct problem' of estimating Af (see [16]). It determines the rate of contraction ε_n of the posterior distribution of Af to Af_0. 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 (3.6), and the inverse problem, through the various approximation rates.

Remark 3.2. It would be possible to obtain the theorem as a corollary of Theorem 2.1 in [30]. We would take the sets $\mathcal{S}_n^c$ in the latter high-level result equal to the sets $\{f : \|f^{(j_n)} - f\|_0 > \eta_n\}$ appearing in (3.6). To verify the conditions of [30] for this choice, most of the preceding proof would be needed. Since the next theorem appears not to be a consequence of this approach, and its proof uses the preceding proof, we have given a direct proof instead.

The theorem applies to a true function f_0 that is 'smooth' of order β (i.e., $f_0 \in H_\beta$). For a prior that is constructed to give an optimal contraction rate for multiple values of β simultaneously, the theorem may not give the best result. The following theorem refines Theorem 3.1 by considering a mixture prior of the form

$$\Pi = \int \Pi_\tau \, dQ(\tau), \tag{3.10}$$

where Π_τ is a prior on H, for every given 'hyperparameter' τ running through some measurable space, and Q is a prior on this hyperparameter. The idea is to *adapt* the prior to multiple smoothness levels through the hyperparameter τ.

Theorem 3.3. *Consider the setup and assumptions of Theorem 3.1 with a prior of the form (3.10). Assume that (3.2), (3.3), (3.4) and (3.5) hold, but replace (3.6) by the pair of conditions, for numbers $\eta_{n,\tau}$ and $C > 0$ and every τ,*

$$\Pi_\tau\big(f : \big\|f^{(j_n)} - f\big\|_0 > \eta_{n,\tau}\big) \le e^{-4n\varepsilon_n^2}, \tag{3.11}$$

$$\Pi_\tau\big(f : \|f - f_0\|_0 < 2\eta_{n,\tau}\big) \le e^{-4n\varepsilon_n^2}, \quad \forall \tau \text{ with } \eta_{n,\tau} \ge C\eta_n. \tag{3.12}$$

Then the posterior distribution in the model (1.2) contracts at the rate η_n at f_0, i.e. for a sufficiently large constant M we have $\Pi_n(f : \|f - f_0\|_0 > M\eta_n | Y^{(n)}) \to 0$, in probability under the law of $Y^{(n)}$ given by (1.2) with $f = f_0$.

Proof. We take the parameter of the model as the pair (f, τ), which receives the joint prior given by $f|\tau \sim \Pi_\tau$ and $\tau \sim Q$. With abuse of notation, we denote this prior also by Π. The likelihood still depends on f only, but the joint prior gives rise to a posterior distribution on the pair (f, τ), which we also denote by $\Pi_n(\cdot|Y^{(n)})$, by a similar abuse of notation.

By (3.10) and (3.11)–(3.12),

$$\Pi\big((f,\tau) : \big\|f^{(j_n)} - f\big\|_0 > \eta_{n,\tau}\big) \le e^{-4n\varepsilon_n^2},$$

$$\Pi\big((f,\tau) : \eta_{n,\tau} \ge C\eta_n, \|f - f_0\|_0 < 2\eta_{n,\tau}\big) \le e^{-4n\varepsilon_n^2}.$$

In view of (3.5) and Theorem 8.20 in [19], the posterior probabilities of the two sets in the left sides tend to zero. As in the proof of Theorem 3.1, we can apply a variation of Theorem 8.22 in [19] to see that it is now sufficient to show the existence of tests τ_n such that, for some $M \ge 2C$,

$$P_{f_0}^{(n)}\tau_n \to 0, \qquad \sup_{\substack{(f,\tau):\|f-f_0\|_0 > M\eta_n \vee 2\eta_{n,\tau},\\ \|f^{(j_n)} - f\|_0 \le \eta_{n,\tau}}} P_f^{(n)}(1 - \tau_n) \le e^{-4n\varepsilon_n^2}.$$

(Note that $M\eta_n \vee 2\eta_{n,\tau} = M\eta_n$ if $\eta_{n,\tau} < C\eta_n$ and $M \ge 2C$.) We use the tests defined in (3.9), as in the proof of Theorem 3.1. The latter proof shows that the tests are consistent. We adapt the bound on the power, as follows.

By the triangle inequality $\tau_n = 0$ implies that, for (f, τ) with $\|f - f_0\|_0 > M\eta_n \vee 2\eta_{n,\tau}$ and $\|f^{(j_n)} - f\|_0 \le \eta_{n,\tau}$,

$$\begin{aligned} n^{-1/2}\|R_{j_n}\bar{\xi}_{j_n}\|_0 &\ge \big\|f^{(j_n)} - f_0\big\|_0 - M_0\eta_n \ge \|f - f_0\|_0 - \big\|f^{(j_n)} - f\big\|_0 - M_0\eta_n \\ &\ge M\eta_n \vee 2\eta_{n,\tau} - \eta_{n,\tau} - M_0\eta_n \ge (M/2 - M_0)\eta_n. \end{aligned}$$

Hence by (3.8) the probability of an error of the second kind is bounded by $e^{-4n\varepsilon_n^2}$, for M sufficiently large that $M/2 - M_0 > a_1$. □

In a typical application of the preceding theorem the priors Π_τ for τ such that $\eta_{n,\tau} \geq C\eta_n$ will be the priors on 'rough' functions, with 'intrinsic' contraction rate $\eta_{n,\tau}$ slower than η_n. These 'bad' priors do not destroy the overall contraction rate, because they put little mass near the true function f_0, by condition (3.12). It is necessary to address these priors explicitly in the conditions, because they will typically fail the approximation condition (3.6), which must be relaxed to (3.11). A further generalization might be to allow the truncation levels j_n to depend on τ, but this will not be needed for our examples.

Inspection of the proof shows that the posterior probability of the sets $\{\tau : \eta_{n,\tau} \gtrsim C\eta_n\}$ tends to zero. This means that the posterior correctly disposes of the models that are 'too rough', for the given true function f_0. In general there is no similar protection against models that are too smooth, but this does not affect the contraction rate.

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 2.3. Consider a prior defined as the law of the random series

$$f = \sum_{i=1}^{M} f_i \phi_i, \tag{4.1}$$

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

Condition 4.1 (Random series prior).

(i) The probability density function p_M of M satisfies, for some positive constants b_1, b_2,

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

(ii) The variable f_i has density $p(\cdot/\kappa_i)/\kappa_i$, for a given probability density p on $\mathbb{R}$ and a constant $\kappa_i > 0$ such that, for some $C > 0$ and $w > 0$, $\alpha, \beta_0 > 0$,

$$p(x) \gtrsim e^{-C|x|^w}, \tag{4.2}$$

$$i^{-\beta_0/d}(\log i)^{-1/w} \lesssim \kappa_i \lesssim i^{\alpha}. \tag{4.3}$$

Priors of this type were studied in [2,49], and applied to inverse problems in the SVD framework in [49] (see Section 3.1 of the latter paper for discussion). For Gaussian variables f_j and degenerate M the series (4.1) is a Gaussian process, and has been more widely studied, but we focus here on the non-Gaussian case. Since the basis $(\phi_i)_{i\in\mathbb{N}}$ used in the prior is linked to the smoothness class $(H_s)_{s\in\mathbb{R}}$, rather than to the operator A, the prior is not restricted to the SVD framework. Of course, in the theorem below we do require the operator to be smoothing in the same smoothness scale, thus maintaining a link between prior and operator.

The assumption on the density p_M is mild and is satisfied, for instance, by the Poisson distribution. The assumption on the density p is mild as well, and is satisfied by many distributions with full support in $\mathbb{R}$, including the Gaussian and Laplace distributions. The parameter β_0 in (4.3) must be a lower bound on the smoothness of the true parameter f_0. Apart from this, condition (4.3) is also very mild, and allows the scale parameters κ_i to tend both to zero or to infinity.

The preceding random series prior is not conjugate to the inverse problem (1.1). In general the resulting posterior distribution will not have a closed form expression, but must be computed using simulation, such as Markov chain Monte Carlo, or approximated using an optimisation method, such as variational approximation. However, the contraction rate of the posterior distribution can be established without the help of an explicit expression for the posterior distribution, as shown in the following theorem.

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 2.3 with $\delta(j,s) = j^{-s/d}$ relative to smoothness classes $(H_s)_{s\in\mathbb{R}}$ as in Definition 2.1. Assume that*

$\|Af\| \simeq \|f\|_{-\gamma}$ for some $\gamma > 0$, and let $f_0 \in H_\beta$ for some $\beta \in (0, S)$. Then, for the random series prior defined in (4.1) and satisfying Definition 4.1 with $\beta_0 \le \beta$, and sufficiently large $\underline{M} > 0$, for $\tau = (\beta+\gamma)(1+2\gamma/d)/(2\beta+2\gamma+d)$,

$$\Pi_n\big(f : \|f - f_0\|_0 > \underline{M} n^{-\beta/(2\beta+2\gamma+d)} (\log n)^\tau \,|\, Y^{(n)}\big) \overset{\mathbb{P}_{f_0}^{(n)}}{\to} 0.$$

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., [11]). 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 [17] and Chapter 10 in [19] for general discussion of adaptation in the Bayesian sense.)

The proof of the theorem is deferred to Section 9; it will be based on Theorem 3.1.

Example 4.3 (Wavelet basis). Let p be a standard normal density, p_M a standard Poisson probability mass function, and set the scaling parameters κ_i equal to 1 (no scaling).

Consider an S-regular orthonormal wavelet basis $\{\phi_{j,k}\}$ for the space of square-integrable functions on the d-dimensional torus $(0, 2\pi]^d$. We can renumber the index (j,k) into $\mathbb{N}$ by ordering the basis functions by their multiresolution levels, $2^{jd} + k$, and next construct the random series prior (4.1).

An S-regular orthonormal wavelet basis is known to correspond to the scale of Sobolev spaces up to smoothness level S. Therefore, by Theorem 4.2, the contraction rate of the posterior distribution is $n^{-\beta/(2\beta+2\gamma+d)}$ times a logarithmic factor whenever the operator is smoothing relative to the Sobolev scale and the true function f_0 belongs to the Sobolev space of order β, for $\beta_0 \le \beta < S$. Thus the posterior distributions are adaptive up to a logarithmic factor to the scale of Sobolev spaces of orders between β_0 and S.

For increasing $\beta \ge S$ the rate given by the theorem still improves. However, the 'regularity' β defined by the scale $(H_s)_{s\in\mathbb{R}}$ may then not coincide with the Sobolev scale.

5. Hilbert scales

A *Hilbert scale* is a special type of smoothness scale $(H_s)_{s\in\mathbb{R}}$, as in Definition 2.1, generated by an unbounded operator. Such a scale is particularly useful in connection to differential operators and Gaussian priors, as considered in the next sections. For reference we include a short summary on Hilbert scales, and some examples. Extended discussions of Hilbert scales in the context of regularization theory can be found e.g. in Chapter 8 of [14], and a general treatment of the subject in [34].

A Hilbert scale is generated by an unbounded operator $L : D(L) \subset H_0 \to H_0$, with domain $D(L)$ such that

(a) $D(L)$ is dense in H_0 (i.e. 'L is densely defined'),
(b) $D(L) = D(L^*)$,
(c) $\langle Lx, y\rangle = \langle x, Ly\rangle$ for all $x, y \in D(L)$ (i.e. 'L is symmetric'),
(d) $\langle Lx, x\rangle \ge \kappa\|x\|^2$, for all $x \in D(L)$, and some $\kappa > 0$.

The set $D(L^*)$ in (b) is the domain of the adjoint L^* of L, which is *defined* as the set of all $y \in H$ such that the map $x \mapsto \langle Lx, y\rangle$ from $D(L)$ to $\mathbb{R}$ is continuous. Thus $D(L^*)$ depends on the domain $D(L)$, which is considered part of the definition of L and is restricted by (a) only. Together, requirements (b) and (c) are equivalent to the requirement that L be *self-adjoint*. The latter is important for the existence of a spectral decomposition, used below.

The domain of the k-th power of the operator L is defined, by induction for $k = 2, 3, \ldots$, as $D(L^k) = \{f \in D(L^{k-1}) : Lf \in D(L)\}$ (with $L^1 = L$). All powers L^k, for $k \in \mathbb{N}$, are defined on

$$H_\infty := \bigcap_{k\in\mathbb{N}} D\big(L^k\big). \tag{5.1}$$

It can be shown that H_∞ is dense in H_0 (Lemma 8.17 in [14]). Next, using spectral theory, fractional powers L^s can be defined as well on the domain H_∞, for every $s \in \mathbb{R}$, through integration with respect to the spectral family (E_λ) of L, i.e.

$$L^s := \int_{\mathbb{R}} \lambda^s \, dE_\lambda = \int_\kappa^\infty \lambda^s \, dE_\lambda.$$

This allows to define an inner product on H_∞ by, for $h, g \in H_\infty$ and $s \in \mathbb{R}$,

$$\langle h, g\rangle_s := \langle L^s h, L^s g\rangle. \tag{5.2}$$

Definition 5.1 (Hilbert scales). The Hilbert space H_s is the completion of H_∞ with respect to the norm induced by the inner product $\langle\cdot,\cdot\rangle_s$ defined in (5.2). The family $(H_s)_{s\in\mathbb{R}}$ is called the *Hilbert scale generated by* L.

The following proposition, adapted from Proposition 8.19 in [14], lists basic properties of Hilbert scales.

Proposition 5.2. *Let L be a densely defined unbounded operator satisfying (a)–(d). Then the Hilbert scale $(H_s)_{s\in\mathbb{R}}$ is a smoothness scale in the sense of Definition 2.1, with*

(i) $\|f\|_s \le \kappa^{s-t}\|f\|_t$, *for* $f \in H_t$, *and* $s < t$.
(ii) $\|f\|_s \le \|f\|_r^{\lambda}\|f\|_t^{1-\lambda}$, *for* $\lambda = (t-s)/(t-r)$, *and* $r < s < t$.

Furthermore, for any $s, t \in \mathbb{R}$ the operator L^{t-s} has a unique extension from H_∞ to a bounded, self-adjoint operator $L^{t-s} : H_t \to H_s$, satisfying

(iii) $\|L^{t-s}f\|_s \simeq \|f\|_t$, *for* $f \in H_t$.
(iv) $L^{t-s} = L^t L^{-s}$.
(v) $(L^s)^{-1} = L^{-s}$.

Somewhat abusing notation, we have denoted the extension of L^{t-s} in the proposition using the same symbol L^{t-s}. Taking $s = 0$ or $t = 0$, we see that $L^s : H_s \to H_0$ and $L^s : H_0 \to H_{-s}$ are norm isomorphisms, for every $s \in \mathbb{R}$. In particular, the unbounded densely defined operator $L : D(L) \subset H_0 \to H_0$ that generates the scale can be extended to a bounded operator $L : H_1 \to H_0$, by strengthening the norm on its domain, and also to a bounded operator $L : H_0 \to H_{-1}$, by extending its range space and weakening the norm of its range space. Moreover, the inverse map is a norm isomorphism $L^{-1} : H_0 \to H_1$, and hence is certainly bounded as an operator $L^{-1} : H_0 \to H_0$.

The eigenvalues of L^{-1} are closely connected to the approximation numbers in Assumption 2.3.

Proposition 5.3. *If $L^{-1} : H_0 \to H_0$ is compact with eigenvalues $\lambda_j \downarrow 0$, then Assumption 2.3 is satisfied in the Hilbert scale $(H_s)_{s\in\mathbb{R}}$ generated by L, with $\delta(j,t) \simeq \lambda_j^t$ and $S = \infty$. In fact, there exist linear spaces V_j of dimension $j-1$ such that, for $s \ge 0$ and $t \in \mathbb{R}$,*

$$\inf_{g\in V_j} \|f - g\|_t \lesssim \delta(j,s)\|f\|_{s+t}, \tag{5.3}$$

$$\|g\|_{s+t} \lesssim \frac{1}{\delta(j,s)}\|g\|_t, \quad \forall g \in V_j. \tag{5.4}$$

Proof. Because $L^{-1} : H_0 \to H_0$ is compact, there exists an orthonormal basis $(\phi_i)_{i\in\mathbb{N}}$ of eigenfunctions in H_0. It may be checked that $f = \sum_{i\in\mathbb{N}} f_i\phi_i$ has $L^s f = \sum_{i\in\mathbb{N}} f_i\lambda_i^{-s}\phi_i$, and square norm $\|f\|_s^2 = \sum_{i\in\mathbb{N}} f_i^2\lambda_i^{-2s}$, provided the latter series converges. Take V_j equal to the linear span of the first $j-1$ eigenfunctions. Then $f - P_j f = \sum_{i\ge j} f_i\phi_i$ and hence $\|f - P_j f\|_t^2 = \sum_{i\ge j} f_i^2\lambda_i^{-2t} \le \lambda_j^{2s}\sum_{i\ge j} f_i^2\lambda_i^{-2t-2s} \le \lambda_j^{2s}\|f\|_{s+t}^2$, for $s, t \ge 0$, and for $f \in V_j$ we have $\|f\|_{s+t}^2 = \sum_{i<j} f_i^2\lambda_i^{-2s-2t} \le \lambda_j^{-2s}\sum_{i\le j} f_i^2\lambda_i^{-2t} = \lambda_j^{-2s}\|f\|_t^2$. □

The sequence spaces of Example 2.5 are one class of examples of Hilbert scales, generated by the operator $L : (f_i) \mapsto (f_i b_i)$. More intricate Hilbert scales arise from (elliptic) differential operators. These are useful in that they can incorporate boundary conditions, which are then automatically inherited by a Gaussian prior attached to such a scale. The following one-dimensional example is simplistic, but illustrative.

Example 5.4 (Sobolev scales). Consider the one-dimensional negative Laplacian

$$-\Delta = -\frac{d^2}{dx^2}$$

as an operator on the space $C_c^\infty(0,1)$ of infinitely often differentiable functions with compact support in $(0,1)$, viewed as subset of $L^2(0,1)$, with range space $L^2(0,1)$. On this domain this operator is not self-adjoint, but it has a self-adjoint

extension (with differentiation interpreted in the sense of distributions) to the space of all functions $f \in W^{2,2}(0,1)$ satisfying the *Dirichlet boundary condition*

$$f(0) = 0 = f(1). \tag{5.5}$$

(See Theorem 4.23 in [23].) The eigenfunctions of the Laplacian under the Dirichlet boundary condition are the functions $x \mapsto \sin(j\pi x)$, for $j \in \mathbb{N}$, with eigenvalues of the order $b_j \asymp j^{-1}$. The corresponding Hilbert scale can also be described as the sequence space generated by this orthogonal basis.

Because the Laplacian is a second derivative it is natural to half the scale parameter, or equivalently use the root negative Laplacian $L := \sqrt{-\Delta}$ as the generator of the scale (where the root is defined through the spectral decomposition).

The boundary conditions play an important role in defining the scale. Technically they are needed to create a domain on which the operator is self-adjoint. An alternative choice to the Dirichlet is the *Cauchy boundary condition*

$$f'(0) = 0 = f(1).$$

This leads to the sequence scale generated by the eigenfunctions $x \mapsto \cos((j-1/2)\pi x)$, for $j \in \mathbb{N}$, and is different from the Dirichlet scale. Again the eigenvalues of L^{-1} are of the order j^{-1}.

Incidentally, it is shown in [43] that the full Sobolev scale ($s \in \mathbb{R}$) of Example 2.4 is not a Hilbert scale for any generating operator L. Also in that sense the boundary conditions are essential.

Example 5.5 (Abel operator). For a given kernel function $K : (0,1) \times (0,1) \to \mathbb{R}$ and $\alpha \in (0,1]$, consider the operator $A : L^2(0,1) \to L^2(0,1)$ given by

$$Af(x) = \frac{1}{\Gamma(\alpha)} \int_0^x (x-s)^{\alpha-1} K(x,s) f(s)\, ds.$$

For $K = 1$ this gives the classical *Abel operator*. Under mild smoothness conditions on K, it is shown in [22], Theorem 1, that A is smoothing (i.e. (2.5) holds) of order $\gamma = 1$ for the Sobolev scale generated by the root negative Laplacian under the Cauchy boundary condition, described in Example 5.4.

While in the preceding examples the boundary consists of just two points, for multi-dimensional domains the boundary is continuous, and the restrictions of functions in a smoothness class to the boundary form an infinite-dimensional function space. By choosing an appropriate generating operator, we can construct a Hilbert scale of functions that automatically satisfy a desired boundary condition.

Consider a second order elliptic differential operator $L : D(L) \subset L^2(\mathcal{D}) \to L^2(\mathcal{D})$ on a bounded domain $\mathcal{D} \subset \mathbb{R}^d$ with a smooth boundary. To generate a Hilbert scale the operator must be self-adjoint, which involves both the form of the operator and its domain $D(L)$, where different domains will lead to different Hilbert scales. Self-adjointness requires both the structural property $\int_{\mathcal{D}} (Lf) g\, d\lambda = \int_{\mathcal{D}} f(Lg)\, d\lambda$ and equality of the domains of L and its adjoint L^*, where the domain of L^* is by definition the set of g such that the left side of the preceding equality is a continuous function of $f \in D(L) \subset L^2(\mathcal{D})$. The latter implies restrictions on the domain, which are typically revealed through partial integrations.

One may start from L as an operator on the space of C^∞-functions with support within $\mathcal{D}$. The closure of this operator (defined by the closure of its graph $\{(f, Lf) : f \in C_c^\infty(\mathcal{D})\}$ in $L^2(\mathcal{D}) \times L^2(\mathcal{D})$), is known as the *minimal realization* associated with L, while the *maximal realization* has domain of definition $\{f \in L^2(\mathcal{D}) : \exists u \in L^2(\mathcal{D})$ such that $Lf = u$ weakly$\}$. (See Definitions 4.1–4.2 in [23]; a self-adjoint operator is always closed, which explains 'minimal'.) Neither of these operators need to be self-adjoint, but there always exist self-adjoint operators with a domain between these two extremes.

For example, the minimal domain of the d-dimensional Laplacian operator $L = -\Delta$ is given by $\{f \in W^{2,2}(\mathcal{D}) : f|_{\partial\mathcal{D}} = \nabla f|_{\partial\mathcal{D}} = 0\}$ (see Theorem 10.19 in [50]) and the maximal domain contains the full Sobolev space $W^{2,2}(\mathcal{D})$ given in Example 2.4 (see Exercise 10.11 in [50]). Two possible domains on which L is self-adjoint are (see Theorems 10.19 and 10.20 in [50]):

$$\big\{f \in W^{2,2}(\mathcal{D}) : f|_{\partial\mathcal{D}} = 0\big\},$$
$$\big\{f \in W^{2,2}(\mathcal{D}) : \nabla f|_{\partial\mathcal{D}} = 0\big\}.$$

These correspond to the *Dirichlet* and *Neumann* boundary conditions, respectively. More sophisticated boundary conditions are possible as well, see [23,38].

In the Bayesian setup we model a function through a prior. When a true function is known to satisfy certain boundary conditions, as in many problems involving differential forward operators, we can incorporate these in the prior by choosing an appropriate generating operator. For an operator A defined in terms of the Laplacian and the same boundary conditions the smoothing condition (2.5) will be satisfied. The following is a another example of a pair of L and A.

Example 5.6 (Volterra). Consider the operator $A : L^2((0,1)^2) \to L^2((0,1)^2)$ on functions $f : (0,1)^2 \to \mathbb{R}$ on the unit square satisfying the differential equation

$$D_{x,y} Af = f, \quad D_{x,y} = \frac{\partial^2}{\partial x\, \partial y}.$$

We can render the solution of the equation unique by imposing boundary conditions. Two solutions are given by

$$Af(x,y) = \int_0^x \int_0^y f(s,t)\, ds\, dt,$$

$$A_0 f(x,y) = Af(x,y) - \int_0^1 Af(x,t)\, dt - \int_0^1 Af(s,y)\, ds + \int_0^1 \int_0^1 Af(s,t)\, ds\, dt.$$

The first satisfies the boundary conditions $Af(x,0) = Af(0,y) = 0$, while the second is obtained from the first by subtracting its projection on the set of all functions of the form $(x,y) \mapsto g_1(x) + g_2(y)$, which forms the kernel of the differential operator. Other boundary conditions will still give different versions of the operator.

We claim that A_0 is smoothing of order $\gamma = 1$ for the Hilbert scale generated by the root L of $D_{x,y}^2$ with Dirichlet boundary condition, while A is smoothing relative to the scale of L combined with Cauchy boundary condition.

The scale under the Dirichlet boundary condition is generated by the orthogonal system of eigenfunctions $e_{k,l} : (x,y) \mapsto \sin(k\pi x)\sin(l\pi y)$, for $(k,l) \in \mathbb{N}^2$, the tensor product of the basis of the one-dimensional Dirichlet–Laplacian as in Example 5.4, with corresponding eigenvalues are $k^2 l^2 \pi^4$. By explicit calculation

$$Ae_{k,l}(x,y) = \frac{1}{kl\pi^2}\big[\cos(k\pi x)\cos(l\pi y) - \cos(k\pi x) - \cos(l\pi y) + 1\big],$$

$$A_0 e_{k,l}(x,y) = \frac{1}{kl\pi^2}\cos(k\pi x)\cos(l\pi y).$$

The functions $(x,y) \mapsto \cos(k\pi x)\cos(l\pi y)$, for $(k,l) \in (\mathbb{N} \cup \{0\})^2$ form an orthogonal basis of $L^2((0,1)^2)$. We conclude that for $f = \sum_{k,l} f_{k,l} e_{k,l}$,

$$\|Af\|^2 \simeq \sum_{k,l} \frac{f_{k,l}^2}{k^2 l^2} + \sum_k \Big(\sum_l \frac{f_{k,l}}{kl}\Big)^2 + \sum_l \Big(\sum_k \frac{f_{k,l}}{kl}\Big)^2 + \Big(\sum_{k,l} \frac{f_{k,l}}{kl}\Big)^2,$$

$$\|A_0 f\|^2 \simeq \sum_{k,l} \frac{f_{k,l}^2}{k^2 l^2} \simeq \|f\|_{-1}^2,$$

where $\|\cdot\|_{-1}$ refers to the scale of L with Dirichlet boundary condition. The first equation shows that the operator A is not smoothing in this scale, but in general satisfies $\|Af\| \gtrsim \|f\|_{-1}$.

On the other hand, the Cauchy boundary condition generates the system of eigenfunctions $(x,y) \mapsto \cos((k - 1/2)\pi x)\cos((l - 1/2)\pi y)$, for $(k,l) \in \mathbb{N}^2$. These can be seen to be also the eigenfunctions of A^*A, and hence the smoothing property of A fits the‘ SVD framework, as in Example 2.7.

The two versions A and A_0 possess the same inverse operator, namely the differential operator $D_{x,y}$ used for their definitions. This suggests that from the point of view of reconstructing f in the inverse problem it should not matter whether one is provided with a noisy version of either Af or $A_0 f$ as input data, seemingly contradicting the fact that the operators are smoothing in different scales. This paradox may be resolved by considering A or A_0 as maps into the quotient space $L^2((0,1)^2)/N(D_{x,y})$, where N denotes the kernel of the operator. The map $f \mapsto [Af] = [A_0 f]$ into the class of Af in this quotient space is injective and can be shown to be appropriately smoothing (see (A.6)–(A.7)), and consequently both scales can be used with both operators (cf. Remark 2.11).

6. Gaussian priors

If the function f in (1.1) is equipped with a Gaussian prior, then the corresponding posterior distribution will be Gaussian as well. Furthermore, the posterior mean will then be equal to the solution found by the method of Tikhonov-type regularization (see e.g. [15,31,52]). Although this allows to study the posterior mean and the full posterior distribution by direct methods, in this section we derive the rate of posterior contraction from the general result Theorem 3.1. An advantage of this approach is that the proof can be extended to mixtures of Gaussian priors, which is important to obtain optimal recovery rates for true functions of different smoothness levels. See Section 7.

Centred Gaussian distributions on a separable Hilbert space correspond bijectively to covariance operators. By definition a random variable F with values in H_0 is Gaussian if $\langle F, g\rangle_0$ is normally distributed, for every $g \in H_0$, and it has zero mean if these variables have zero means. The variances of these variables can then be written as

$$\mathrm{E}\langle F, g\rangle_0^2 = \langle Cg, g\rangle_0,$$

for a linear operator $C : H_0 \to H_0$, called the *covariance operator*. A covariance operator C is necessarily self-adjoint, nonnegative, and of *trace class*, i.e., $\sum_{i\in\mathbb{N}}\langle C\phi_i, \phi_i\rangle < \infty$, for some (and then every) orthonormal basis $(\phi_i)_{i\in\mathbb{N}}$ of H_0; and every operator with these properties generates a Gaussian distribution.

In the setting of a Hilbert scale $(H_s)_{s\in\mathbb{R}}$ generated by the operator L it is natural to choose a Gaussian prior with covariance operator of the form $L^{-2\alpha}$, for some $\alpha > 0$. If L^{-1} has eigenvalues λ_j, then this operator is of trace class if $\sum_{j\in\mathbb{N}}\lambda_j^{-2\alpha} < \infty$. Thus α must be chosen big enough for the Gaussian prior to exist as a 'proper' prior on H_0. For instance, if $\lambda_j \simeq j^{-1/d}$, then every choice $\alpha > d/2$ yields a proper prior.

This leads to the following theorem on posterior contraction rates for Gaussian priors, the proof of which is given in Section 9.

Theorem 6.1 (Gaussian prior). *Consider a Hilbert scale $(H_s)_{s\in\mathbb{R}}$ generated by an operator L as in the preceding such that $L^{-1} : H_0 \to H_0$ is compact with eigenvalues λ_j satisfying $\lambda_j \simeq j^{-1/d}$. Suppose the operator $A : H_0 \to G$ satisfies $\|Af\| \simeq \|f\|_{-\gamma}$, assume that $f_0 \in H_\beta$, for some $\beta > 0$, and let the prior be zero-mean Gaussian with covariance operator $L^{-2\alpha}$, for some $\alpha > d/2$. Then the posterior distribution satisfies, for sufficiently large $M > 0$,*

$$\Pi_n\big(f : \|f - f_0\|_0 > Mn^{-((\alpha-d/2)\wedge\beta)/(2\alpha+2\gamma)} | Y^{(n)}\big) \overset{\mathbb{P}_{f_0}^{(n)}}{\to} 0.$$

If F is distributed according to the prior in the preceding theorem, then $L^s F$ is also zero-mean Gaussian distributed, with covariance operator $L^{2s-2\alpha}$, which has eigenvalues $j^{-(2\alpha-2s)/d}$. For $s < \alpha - d/2$, this operator is of trace class and hence $L^s F$ is a proper random variable in H_0. In other words, the distribution of F gives probability 1 to $L^{-s}H_0 = H_s$, for every $s < \alpha - d/2$. The prior in the preceding theorem can therefore be interpreted as being 'almost' of regularity $\alpha - d/2$. The rate $n^{-((\alpha-d/2)\wedge\beta)/(2\alpha+2\gamma)}$ is therefore comparable to the rate obtained in Theorem 3.5 in [49] and Theorem 4.1 in [31] (with the scaling parameter fixed to 1), except that the parameter α in the latter references is denoted presently by $\alpha - d/2$.

An improvement of the present theorem is that the covariance operator of the Gaussian prior is not directly linked to the operator A, but only weakly so by (2.5). For example, we may construct a prior by a random series (see Theorem I.23 in Appendix I.6, [19]), in any basis corresponding to the smoothness scale. We illustrate this below by using the wavelet basis for an inverse problem given by a differential operator, after first noting that the singular value setup is covered as well.

Example 6.2 (SVD). The scale of smoothness classes constructed in Example 2.5 and Example 2.7 is the Hilbert scale attached to the operator L given by $Lf = \sum_{i\in\mathbb{N}} b_i f_i \phi_i$ defined on the domain of functions $f = \sum_{i\in\mathbb{N}} f_i\phi_i$, with $\sum_{i\in\mathbb{N}} b_i^2 f_i^2 < \infty$. Under assumption (2.6) this operator can also be expressed as $L = (A^*A)^{-1/(2\gamma)}$, and depends on the operator A through its eigenfunctions. A Gaussian prior with covariance operator $L^{-2\alpha}$ corresponds to modelling the coefficients f_i relative to the basis ϕ_i as independent zero-mean normal variables F_i with variances $b_i^{-2\alpha}$. This follows, because in that case $\mathrm{E}\langle F, g\rangle_0^2 = \sum_{i\in\mathbb{N}} b_i^{-2\alpha} g_i^2 = \langle L^{-2\alpha} g, g\rangle_0^2$, for every $g \in H_0$.

Thus in this case the prior coincides with the ones in the literature studied under the SVD framework, e.g. [31,32]. In the present more general setting L need not be directly linked to A, except that the operator must possess the smoothing property Definition 2.6.

Example 6.3 (Sobolev scales, wavelet prior). Let $\{\phi_{j,k}\}_{(j,k)\in\Lambda}$, be an S-regular orthonormal wavelet basis in $L^2(\mathbb{T})$, on $\mathbb{T} := (0, 2\pi]$. Let $f_{j,k} = \int_{\mathbb{T}} f(x)\phi_{j,k}(x)\,dx$ be the wavelet coefficients of a function f. By Parseval's identity, the map

$U : f \mapsto \{f_{j,k}\}$ is a unitary operator $U : L^2(\mathbb{T}) \to \ell^2(\Lambda)$. The multiplication operator $m : \{f_{j,k}\} \mapsto \{2^j f_{j,k}\}$ on $\ell^2(\Lambda)$ has s-th power given by $m^s : \{f_{j,k}\} \mapsto \{2^{js} f_{j,k}\}$. Then $L := U^* m U$ has s-th power $L^s := U^* m^s U$ and generates a Hilbert scale $(H_s)_{s\in\mathbb{R}}$. For $f \in H_s$, we have

$$\|f\|^2_{H_s(\mathbb{T})} = \sum_{j=0}^{\infty} 2^{2js} \sum_{k=0}^{2^j-1} f_{j,k}^2.$$

This norm can be shown to be equivalent to the standard Sobolev norm, for $0 \le s < S$.

The Gaussian prior with covariance operator $L^{-2\alpha}$ can be represented by a random series of the form

$$F = \sum_{(j,k)\in\Lambda} F_{j,k}\phi_{j,k},$$

where $F_{j,k} \sim \mathcal{N}(0, 2^{-2j\alpha})$ are independent random variables. This prior corresponds to the Hilbert scale, but does not refer to an operator A. For instance, the eigenbasis of the operator in Example 2.9 is the Fourier basis (see [29]), and not the wavelet basis. Thus we have constructed a Gaussian prior that is not related to the eigenbasis, but attains the same contraction rate.

It may be noted that the scale $(H_s)_{s\in\mathbb{R}}$ is well defined for every $s \in \mathbb{R}$, and with the preceding prior Theorem 6.1 is applicable to the full scale, and gives a contraction rate relative to the scale, which is optimal when $\beta = \alpha - d/2$. However, the scale agrees with the Sobolev scale only for $\beta < S$, and hence the optimality is in the Sobolev sense only if $\beta < S$. This restriction is typical when working with an approximation scheme such as wavelets or splines. One can of course choose a suitably large value of S, or may mix over multiple wavelet bases, as in the next section.

As mentioned in Section 1, there are many works on Bayesian inverse problems with Gaussian priors. The setup of the preceding theorem is similar to [1,15], arguably closer to [1]. While we mainly treat the white noise case, our results can be extended to cover the noise structure in [1], and hence also cover the model in [15]. On the other hand, we differ from [1] in the following sense. First, unlike Assumption 3.1 in [1], our characterization of the smoothing property of the operator A, i.e. Definition 2.6, is simple, and in principle, our setup can also be extended to severely ill-posed problems, see Section 8. Second, our proof strategy is different, as we do not use Gaussian conjugacy, which is the main tool in [1]. This also allows us to obtain posterior contraction rates for non-conjugate priors in Section 4, and for Gaussian mixtures in Section 7.

7. Gaussian mixtures

The posterior contraction rate resulting from a zero-mean Gaussian prior with covariance operator $L^{-2\alpha}$, as considered in Section 6, is equal to the minimax rate $n^{-\beta/(2\beta+2\gamma+d)}$ (see [11]) 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. [33,53,57]). In this section we consider mixtures of the mean-zero Gaussian priors with covariance operators $\tau^2 L^{-2\alpha}$ over the 'hyperparameter' 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 in Section 6, 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.

Condition 7.1. The distribution Q of τ has support $[0, \infty)$ and satisfies

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

Theorem 7.2 (Gaussian mixture prior). *Consider a Hilbert scale $(H_s)_{s\in\mathbb{R}}$ generated by an operator L as in the preceding such that $L^{-1} : H_0 \to H_0$ is compact with eigenvalues λ_j satisfying $\lambda_j \simeq j^{-1/d}$. Suppose the operator $A : H_0 \to G$ satisfies $\|Af\| \simeq \|f\|_{-\gamma}$, assume that $f_0 \in H_\beta$, for some $\beta \in (0, \alpha]$, and let the prior be a mixture of the zero-mean Gaussian distributions with covariance operators $\tau^2 L^{-2\alpha}$ over the parameters τ equipped with*

a prior satisfying Definition 7.1, for some $\alpha > d/2$. Then the posterior distribution satisfies, for sufficiently large $M > 0$,

$$\Pi_n\big(f : \|f - f_0\|_0 > Mn^{-\beta/(2\beta+2\gamma+d)} | Y^{(n)}\big) \overset{\mathbb{P}_{f_0}^{(n)}}{\to} 0.$$

The proof is given in Section 9.

8. Discussion and comments

In this section we comment on the present setup and discuss directions in which the results in this article can be extended.

Coloured noise

We have examined the case that the noise ξ in model (1.1) is white noise. Statistical estimation in the case that the noise is a proper centred Gaussian random element in G, as studied in [15], is easier in terms of minimax rates (if in both cases the noise is scaled to the same unit), as this would imply that the noise is less variable. By inspection of our proofs one sees that the concentration inequalities that drive the testing criterion remain valid if the covariance operator of the noise is bounded above by the identity, as is assumed in [1,3]. As a consequence, the proof of Theorem 3.1 goes through and the theorem remains valid, as do the corollaries in the later sections. However, for truly coloured noise the result may be suboptimal, as one may expect a faster posterior contraction rate, which will incorporate the decrease of the noise variance in certain directions. The methods of the present paper can be adapted to this case as long as the covariance operator fits the scale of smoothness classes, as in [15]. A sharp result in full generality may be difficult to attain, as it will be the outcome of the interaction of the directions of decrease in the noise, the true parameter and the prior.

Approximation numbers of embeddings

In the corollaries to the main result we have assumed that the approximation numbers $\delta(j, s)$ of the canonical embedding $\iota : H_s \to H_0$ are of polynomial order $j^{-s/d}$. This order matches the approximation numbers of Sobolev spaces on d-dimensional, bounded domains, and seems common. Other decay rates do arise, e.g., an exponential rate in severely ill-posed problems (as in the heat equation considered in [32]), or a logarithmic rate (as in [6]). The general Theorem 3.1 remains valid, but its corollaries must be adapted. For Gaussian priors in logarithmic or exponential scales, this is relatively straightforward using the general theory of approximation numbers, which relates these to singular values and metric entropy. See the discussion in Appendix B.

9. Proofs

Lemma 9.1. *For $\theta = (\theta_1, \theta_2, \ldots)$ let P_θ be the distribution of the random element $(X_1 + \theta_1, X_2 + \theta_2, \ldots)$ in $\mathbb{R}^\infty$ for $X_1, X_2, \ldots$ i.i.d. mean-zero normal variables with variance σ^2. If $\theta \in \ell^2$, then P_θ is absolutely continuous relative to P_0 with log likelihood*

$$\log \frac{dP_\theta}{dP_0}(X_1, X_2, \ldots) = \frac{1}{\sigma^2}\sum_{i=1}^\infty \theta_i X_i - \frac{1}{2\sigma^2}\sum_{i=1}^\infty \theta_i^2,$$

where the first series converges almost surely and in second mean. The expectation and variance of minus this variable are $\sum_{i=1}^\infty \theta_i^2/(2\sigma^2)$ and twice this quantity, respectively.

Proof. That the series converges in L^2 is clear from the fact that $\theta \in \ell^2$; the almost sure convergence next follows from the Itô–Nisio theorem. The expectation and variance of the right side are easy to compute as limits.

Write Λ_∞ for the right side of the display, and Λ_n for the expression obtained by replacing the infinite sums by the sums from 1 to n. Thus $\Lambda_n \to \Lambda_\infty$ almost surely. Since $\mathrm{E}_0 e^{2\Lambda_n} = e^{\sum_{i=1}^n \theta_i^2/\sigma^2}$ is uniformly bounded in n, it follows that e^{Λ_n} is uniformly integrable and hence converges in mean to e^{Λ_∞}. In particular, the mean of the latter variable is 1, the mean of the former variables.

It follows that the Borel measure on $\mathbb{R}^\infty$ defined by $B \mapsto \mathrm{E}_0 1_B(X) e^{L_\infty}$ is a probability measure. For every Borel set B it is the limit of $\mathrm{E}_0 1_B(X) e^{L_n}$, which is $P_\theta(B)$ if B depends only on the first n coordinates, as e^{L_n} is the density of the distribution of $(X_1 + \theta_1, \ldots, X_n + \theta_n)$ with respect to its distribution at $\theta = 0$. Since the Borel σ-field on $\mathbb{R}^\infty$ is generated by the algebra of all cylinder sets, it follows that P_θ and the measure $B \mapsto \mathrm{E}_0 1_B(X) e^{L_\infty}$ agree. □

9.1. *Proof of Theorem 4.2*

The theorem is a corollary to Theorem 3.1 and uses arguments as in the proof of Proposition 3.2 in [49].

First we determine ε_n to satisfy the prior mass condition (3.5) of the direct problem. Let P_j be the projection onto the linear span of the first $j-1$ basis elements ϕ_i. By the assumption on A and the triangle inequality, for any $i_n \in \mathbb{N}$,

$$\begin{aligned}\|Af - Af_0\| &\lesssim \|f - f_0\|_{-\gamma} \lesssim \|f - P_{i_n} f_0\|_{-\gamma} + \|P_{i_n} f_0 - f_0\|_{-\gamma} \\ &\lesssim \|f - P_{i_n} f_0\|_{-\gamma} + \delta(i_n, \gamma)\delta(i_n, \beta)\|f_0\|_\beta, \end{aligned} \tag{9.1}$$

by (A.1), if $0 \le \beta, \gamma < S$. Here $\delta(i_n, \gamma)\delta(i_n, \beta) = i_n^{-(\gamma+\beta)/d} \simeq \varepsilon_n$ if $i_n \simeq \varepsilon_n^{-d/(\gamma+\beta)}$.

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 $\|\phi_j\|_{-\gamma} \le \delta(j, \gamma)\|\phi_j\|_0 \lesssim j^{-\gamma/d}$, for every j, by (A.1). Consequently, for $f = \sum_{i=1}^{i_n-1} f_i\phi_i \in V_{i_n}$ and $f_0 = \sum_i f_{0,i}\phi_i$, by the triangle inequality,

$$\|f - P_{i_n} f_0\|_{-\gamma} \lesssim \sum_{i=1}^{i_n-1} |f_i - f_{0,i}| i^{-\gamma/d}.$$

It follows that there exists a constant $a > 0$ such that

$$\begin{aligned}\Pi\big(f : \|f - P_{i_n} f_0\|_{-\gamma} < a\varepsilon\big) &\ge \Pi\left(\big((f_i), M\big) : \sum_{i=1}^{i_n-1} |f_i - f_{0,i}| i^{-\gamma/d} < \varepsilon, M = i_n - 1\right) \\ &\ge \prod_{i=1}^{i_n} \Pi\left(f_i : |f_i - f_{0,i}| < \frac{\varepsilon i^{\gamma/d}}{i_n}\right) \Pi(M = i_n - 1) \\ &\ge \prod_{i=1}^{i_n} \int_0^{\varepsilon i^{\gamma/d}/(\kappa_i i_n)} p\left(x + \frac{f_{0,i}}{\kappa_i}\right) dx e^{-b_1 i_n},\end{aligned}$$

in view of Definition 4.1. By (4.2) of the latter assumption, the integral $\int_0^r p(x+\mu)\,dx$ is bounded below by a constant times $re^{-C(r+|\mu|)^w}$. It follows that for ε such that $\varepsilon i^{\gamma/d}/(\kappa_i i_n) \le 1$, for $i \le i_n$, the preceding display is lower bounded by a multiple of

$$\varepsilon^{i_n}\left[\prod_{i=1}^{i_n} \frac{i^{\gamma/d}}{\kappa_i i_n}\right] \exp\left[-C\sum_{i=1}^{i_n}\left(1 + \frac{|f_{0,i}|}{\kappa_i}\right)^w\right] e^{-b_1 i_n}.$$

By (4.3), we have $i^{\gamma/d}/\kappa_i \gtrsim (1/i)^{\gamma/d-\alpha}$, which is bounded below by 1 if $\gamma/d - \alpha \ge 0$ and by $(1/i_n)^{\alpha-\gamma/d}$ otherwise, and hence always by $(1/i_n)^\alpha$. This shows that the first term in square brackets is bounded below by $(a_2/i_n^{\alpha+1})^{i_n}$, for some $a_2 > 0$. Since $f_0 \in H_\beta$, by assumption, the norm duality (2.1) gives that $|f_{0,i}| = |\langle f_0, \phi_i\rangle_0| \le \|f_0\|_\beta \|\phi_i\|_{-\beta} \lesssim i^{-\beta/d}$. Together with (4.3) this gives that $|f_{0,i}|/\kappa_i \lesssim i^{(\beta_0-\beta)/d}(\log i)^{1/w} \le (\log i)^{1/w}$, whence minus the exponent in the second term in square brackets is bounded by a multiple of $i_n(1 + (\log i_n)^{1/w})^w$. We conclude that there exists a constant $a_3 > 0$ such that

$$\Pi\big(f : \|f - P_{i_n} f_0\|_{-\gamma} < a\varepsilon\big) \ge \varepsilon^{i_n} e^{-a_3 i_n \log i_n} e^{-b_1 i_n},$$

for every $\varepsilon > 0$ such that $\varepsilon i^{\gamma/d}/(\kappa_i i_n) \le 1$, for every $i \le i_n$. Since $i^{\gamma/d}/\kappa_i \lesssim i^{(\gamma+\beta_0)/d}(\log i)^{1/w}$, again by (4.3), a sufficient condition for the latter is that $\varepsilon i_n^{(\gamma+\beta_0)/d}(\log i_n)/i_n \le 1$.

Combining this with (9.1), we see that (3.5) is satisfied for ε_n such that there exists i_n with

$$i_n^{-(\gamma+\beta)/d} \lesssim \varepsilon_n, \qquad i_n \log i_n \lesssim n\varepsilon_n^2, \qquad \varepsilon_n i_n^{(\gamma+\beta_0)/d}(\log i_n) \le i_n.$$

This leads to the rates

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

(The third requirement is easily satisfied and remains inactive.) We can choose a sufficiently large proportionality constant in $\simeq$ when defining ε_n, so that (3.5) is satisfied for ε_n, since the left and right sides of (3.5) are increasing and decreasing in ε_n, respectively.

Since the Galerkin projection $f^{(j)}$ is equal to f itself if $f \in V_j$, we have that $\|f^{(j_n)} - f\|_0 = 0$ for the random series $f = \sum_{i=1}^{M} f_i \phi_i$ if $M < j_n$. By (ii) of Definition 4.1 it follows that, for some $b_2' > 0$ and every $\eta_n > 0$,

$$\Pi\big(f : \big\|f^{(j_n)} - f\big\|_0 > \eta_n\big) \leq \Pi(M \geq j_n) \leq e^{-b_2' j_n}.$$

Hence (3.6) is satisfied for $j_n = n\varepsilon_n^2/(4b_2')$. Thus we choose

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

with a sufficiently large constant in $\simeq$. Then (3.2) is satisfied and it remains to solve η_n from (3.3) and (3.4). This leads to the inequalities

$$\eta_n \geq \varepsilon_n j_n^{\gamma/d} \simeq n^{-\beta/(2\beta+2\gamma+d)} (\log n)^{(1+2\gamma/d)(\beta+\gamma)/(2\beta+2\gamma+d)},$$
$$\eta_n \geq j_n^{-\beta/d} \simeq n^{-\beta/(2\beta+2\gamma+d)} (\log n)^{-\beta(2\beta+2\gamma)/((2\beta+2\gamma+d)d)}.$$

The rate is the maximum of the rates at the right hand sides, which coincides with the first rate. This concludes the proof.

9.2. *Proof of Theorem 6.1*

The theorem is a corollary to Theorem 3.1. The main tasks are to determine ε_n satisfying the prior mass condition (3.5) of the direct problem, and next to identify η_n from the prior mass condition (3.6) and the other conditions.

The first task is achieved in the following lemma.

Lemma 9.2. *Under the assumptions of Theorem 6.1, for $f_0 \in H_\beta$, as $\varepsilon \downarrow 0$,*

$$-\log \Pi\big(f : \|Af - Af_0\| < \varepsilon\big) \lesssim \begin{cases} \varepsilon^{-d/(\alpha+\gamma-d/2)} & \text{if } d/2 < \alpha \leq \beta + d/2, \\ \varepsilon^{-(2\alpha-2\beta)/(\beta+\gamma)} & \text{if } \alpha > \beta + d/2. \end{cases} \tag{9.2}$$

Proof. Since by assumption $\|Af - Af_0\| \simeq \|f - f_0\|_{-\gamma}$, the probability in the left side is the decentered small ball probability $\Pi(f : \|f - f_0\|_{-\gamma} < a\varepsilon)$ of the Gaussian random variable F distributed according to the prior and viewed as map into $H_{-\gamma} \supset H_0$, for some $a > 0$. Because F has covariance operator $L^{-2\alpha}$ as a map in H_0, its reproducing kernel Hilbert space (or Cameron–Martin space) $\mathbb{H}$ (which does not depend on its range space) is equal to the range of $L^{-\alpha}$ under the norm $\|L^{-\alpha}h\|_{\mathbb{H}} = \|h\|_0$ (see e.g., Example I.14 of [19]). Since $L^{-\alpha} : H_0 \to H_\alpha$ is a norm isometry, by (iii) of Proposition 5.2, this is the Hilbert space H_α with its natural norm $\|\cdot\|_\alpha$. The left side of (9.2) is therefore up to constants equivalent to

$$\inf_{h \in H_\alpha : \|h - f_0\|_{-\gamma} < \varepsilon} \|h\|_\alpha^2 - \log \Pi\big(\|f\|_{-\gamma} < \varepsilon\big). \tag{9.3}$$

See [35,36,58], or Section 11.2, in particular, Proposition 11.19 in [19].

By (A.1) $\|P_j f_0 - f_0\|_{-\gamma} \lesssim \delta(j,\gamma)\delta(j,\beta)\|f_0\|_\beta$, which is bounded above by ε for $j \simeq \varepsilon^{-d/(\beta+\gamma)}$. Thus for this value of j the first term in (9.3) is bounded above by

$$\|P_j f_0\|_\alpha \lesssim \begin{cases} \|P_j f_0\|_\beta & \text{if } \alpha \leq \beta, \\ 1/\delta(j, \alpha - \beta)\|P_j f_0\|_\beta & \text{if } \alpha > \beta \end{cases}$$

by (5.4). Here $\|P_j f_0\|_\beta \leq \|P_j f_0 - f_0\|_\beta + \|f_0\|_\beta \leq (\delta(j,0) + 1)\|f_0\|_\beta$, by (5.3). It follows that the contribution of the decentering in (9.3) is of order 1 if $\alpha \leq \beta$ and is bounded above by a term of order $\varepsilon^{-2(\alpha-\beta)/(\beta+\gamma)}$ if $\alpha > \beta$.

By Lemma B.1, the metric entropy $\log N(\varepsilon, \{f \in H_\alpha : \|f\|_\alpha \leq 1\}, \|\cdot\|_{-\gamma})$ is of the order $\varepsilon^{-d/(\alpha+\gamma)}$. Hence, by [35] (see Lemma 6.2 in [59]),

$$-\log \Pi\big(\|f\|_{-\gamma} < \varepsilon\big) \simeq \varepsilon^{-d/(\alpha+\gamma-d/2)}.$$

Finally, the assertion of the lemma follows from discussion by cases. □

It follows that (3.5) is satisfied for

$$\varepsilon_n \geq n^{-(\beta\wedge(\alpha-d/2)+\gamma)/(2\alpha+2\gamma)}. \tag{9.4}$$

The next step of the proof is to bound the prior probability in (3.6).

Lemma 9.3. *Under the assumptions of Theorem 6.1, there exist $a, b > 0$, such that for every $j \in \mathbb{N}$ and $t > 0$,*

$$\Pi\big(f : \|f^{(j)} - f\|_0 > t + aj^{1/2-\alpha/d}\big) \leq e^{-bt^2 j^{2\alpha/d}}.$$

Proof. We have $f^{(j)} - f = (R_j A - I)f$, for $R_j = A^{-1}Q_j$. Therefore, the probability on the left concerns the random variable $(R_j A - I)F$, if F is a variable distributed according to the prior Π. Since F is zero-mean normal with covariance operator $L^{-2\alpha}$, this variable is zero-mean Gaussian with covariance operator $(R_j A - I)L^{-2\alpha}(R_j A - I)^*$. We shall compute the weak and strong second moments of the variable $(R_j A - I)F$, and next apply Borell's inequality for the norm of a Gaussian variable to obtain the exponential bound.

Because $\langle (R_j A - I)F, g\rangle_0 = \langle F, (R_j A - I)^* g\rangle_0$ is zero-mean Gaussian with variance $\|L^{-\alpha}(R_j A - I)^* g\|_0^2 = \|(R_j A - I)^* g\|_{-\alpha}^2$, the weak second moment of $(R_j A - I)F$ is given by

$$\sup_{\|g\|_0\leq 1} \mathrm{E}\big\langle (R_j A - I)F, g\big\rangle_0^2 = \sup_{\|g\|_0\leq 1} \big\|(R_j A - I)^* g\big\|_{-\alpha}^2.$$

By the norm duality (2.1), the right side is equal to

$$\sup_{\|g\|_0\leq 1} \sup_{\|f\|_\alpha\leq 1} \big\langle f, (R_j A - I)^* g\big\rangle_0^2 \leq \sup_{\|f\|_\alpha\leq 1} \big\|(R_j A - I)f\big\|_0^2 \lesssim \delta(j,\alpha)^2$$

in view of (A.5).

The strong second moment of the Gaussian variable $(R_j A - I)F$ is equal to the trace of its covariance operator. As $\mathrm{Trace}(S^*S) = \sum_i \|S\phi_i\|^2 = \sum_i \sum_j \langle S\phi_i, \phi_j\rangle^2 = \sum_i \|S^*\phi_i\|^2$, for any orthonormal basis (ϕ_i) and operator S, we have

$$\mathrm{E}\big\|(R_j A - I)F\big\|_0^2 = \sum_{i\in\mathbb{N}} \big\|(R_j A - I)L^{-\alpha}\phi_i\big\|_0^2.$$

For the orthonormal basis of eigenfunctions of L^{-1} and V_j the span of the first $j-1$ of these eigenfunctions, as in Proposition 5.3, $L^{-\alpha}V_j \subset V_j$, and hence $(R_j A - I)L^{-\alpha}\phi_i$ vanishes for $i < j$. For $i \geq j$ the latter element is the difference $g^{(j)} - g$ of the Galerkin solution $g^{(j)}$ to $g = L^{-\alpha}\phi_i$. Therefore, by (A.5) the preceding display is bounded above by a multiple of

$$\sum_{i\geq j} \delta(i,\alpha)^2 \big\|L^{-\alpha}\phi_i\big\|_\alpha^2 = \sum_{i\geq j} \delta(i,\alpha)^2 \|\phi_i\|_0^2 \lesssim j^{1-2\alpha/d},$$

where we used the estimate $\sum_{i>j} i^{-b} \leq j^{1-b}/(b-1)$, for $b > 1$.

Since the first moment of $\|(R_j A - I)F\|_0$ is bounded by the root of its second moment, the lemma follows by Borell's inequality (see e.g. Lemma 3.1 and subsequent discussion in [37]). □

For $t^2 = 4n\varepsilon_n^2/(bj_n^{2\alpha/d})$ and $j = j_n$ the bound in the preceding lemma becomes $e^{-4n\varepsilon_n^2}$. Hence (3.6) 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 (3.5), given in (9.4). Next we solve for η_n under the constraints (3.3) and (3.4). 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

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

$$\eta_n \geq n^{-(\beta\wedge(\alpha-d/2)+\gamma)/(2\alpha+2\gamma)} j_n^{\gamma/d},$$

$$\eta_n \geq j_n^{-\beta/d}.$$

Depending on the relation between α and $\beta + d/2$, two situations need to be discussed separately.

(i) $\alpha \le \beta + d/2$. 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 $\eta_n \ge n^{-(\alpha-d/2)/(2\alpha+2\gamma)}$, while the third becomes $\eta_n \ge n^{-\beta/(2\alpha+2\gamma)}$ and becomes inactive.
(ii) $\alpha > \beta + d/2$. We choose $j_n \simeq n^{d/(2\alpha+2\gamma)} \le n\varepsilon_n^2$, and then see that all three requirements reduce to $\eta_n \ge n^{-\beta/(2\alpha+2\gamma)}$.

Finally, we apply Theorem 3.1 to complete the proof.

9.3. *Proof of Theorem 7.2*

Let Π_τ denote the zero-mean Gaussian distribution on H with covariance operator $\tau^2 L^{-2\alpha}$ (where $\alpha > d/2$).

Lemma 9.4. *Under the assumptions of Theorem 7.2, for $f_0 \in H_\beta$ and $\beta \le \alpha$, as $\varepsilon \downarrow 0$,*

$$-\log \Pi_\tau\big(f : \|Af - Af_0\| < \varepsilon\big) \lesssim \frac{1}{\tau^2}\left(\frac{1}{\varepsilon}\right)^{(2\alpha-2\beta)/(\beta+\gamma)} + \left(\frac{\tau}{\varepsilon}\right)^{d/(\alpha+\gamma-d/2)}.$$

Lemma 9.5. *Under the assumptions of Theorem 7.2, for $f_0 \in H_\beta$ and $\beta \le \alpha$, as $\varepsilon \downarrow 0$,*

$$-\log \Pi_\tau\big(f : \|f\|_0 < \varepsilon\big) \gtrsim \left(\frac{\tau}{\varepsilon}\right)^{d/(\alpha-d/2)}.$$

Lemma 9.6. *Under the assumptions of Theorem 7.2, there exist $a, b > 0$ such that, for every $j \in \mathbb{N}$ and $x, \tau > 0$,*

$$\Pi_\tau\big(f : \big\|f^{(j)} - f\big\|_0 > \tau x + \tau a j^{1/2-\alpha/d}\big) \le e^{-bx^2 j^{2\alpha/d}}.$$

Proofs. The proof of the first lemma follows the same lines as the proof of Lemma 9.2, except that now the Cameron–Martin space of the measure Π_τ on $H_{-\gamma}$ is H_α equipped with the norm $\|\cdot\|_{\mathbb{H}} = \frac{1}{\tau}\|\cdot\|_\alpha$ rather than its natural norm. The second lemma follows similarly, but considers the centered probability only. The third lemma is immediate from Lemma 9.3 as Π_τ is the law of τF, for F the Gaussian variable with the law Π as in the latter lemma, and the map $f \mapsto f^{(j)} - f$ is linear. □

As preparation for the proof of Theorem 7.2, we first show that the minimax rate can be obtained by a Gaussian prior with the deterministic scaling, dependent on β, given by

$$\tau_n = n^{(\alpha-d/2-\beta)/(2\beta+2\gamma+d)}. \tag{9.5}$$

Theorem 9.7. *Assume the conditions on the Hilbert scale, the forward operator A and the true parameter f_0 in Theorem 6.1 hold. Suppose that the priors Π are zero-mean Gaussian with covariance operators $\tau_n^2 L^{-2\alpha}$ with τ_n as given in (9.5) and $\alpha > d/2$. Then for $\beta \le \alpha$, the posterior distribution satisfies, for sufficiently large $M > 0$,*

$$\Pi_n\big(f : \|f - f_0\|_0 > Mn^{-\beta/(2\beta+2\gamma+d)} | Y^{(n)}\big) \overset{\mathbb{P}_{f_0}^{(n)}}{\to} 0.$$

Proof. The theorem is a corollary to Theorem 3.1. The proof follows the same lines as the proof of Theorem 6.1. By Lemma 9.4, inequality (3.5) is satisfied for

$$\varepsilon_n \gtrsim n^{-(\beta+\gamma)/(2\beta+2\gamma+d)}.$$

By Lemma 9.6, inequality (3.6) is satisfied for

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

We choose $j_n \simeq n\varepsilon_n^2$, and the minimal solution $\varepsilon_n = n^{-(\beta+\gamma)/(2\beta+2\gamma+d)}$ to the second last display. It is then straightforward to verify that (3.3), (3.4) and (3.6) are satisfied for $\eta_n \simeq n^{-\beta/(2\beta+2\gamma+d)}$. □

Theorem 7.2 is a corollary of Theorem 3.3, with the choices

$$\eta_n \simeq n^{-\beta/(2\beta+2\gamma+d)}, \qquad \varepsilon_n \simeq n^{-(\beta+\gamma)/(2\beta+2\gamma+d)},$$
$$j_n \simeq n\varepsilon_n^2 = n^{d/(2\beta+2\gamma+d)}.$$

Conditions (3.2), (3.3), and (3.4) are satisfied for these choices. It remains to verify (3.5), and (3.11)–(3.12).

For ease of notation, for the moment, define η_n and ε_n as in the preceding display, with exact equality (i.e., with the constant set equal 1). Let τ_n be the 'optimal' scaling rate defined in (9.5).

Verification of (3.5). For $\tau \asymp \tau_n$ and $\varepsilon \asymp \varepsilon_n$ as given and $\beta \le \alpha$, both terms in the right side of Lemma 9.4 are of the order $n\varepsilon_n^2$. The lemma yields, for $\tau_n \le \tau \le 2\tau_n$ and some constant $a_1 > 0$,

$$-\log \Pi_\tau\big(f : \|Af - Af_0\| < \varepsilon_n\big) \le a_1 n\varepsilon_n^2.$$

This shows that

$$\begin{aligned}\Pi\big(f : \|Af - Af_0\| < \varepsilon_n\big) &= \int_0^\infty \Pi_\tau\big(f : \|Af - Af_0\| < \varepsilon_n\big)\, dQ(\tau)\\ &\ge e^{-a_1 n\varepsilon_n^2} Q(\tau_n, 2\tau_n).\end{aligned}$$

If $\alpha - d/2 < \beta$, then $\tau_n \to 0$, and Definition 7.1 on Q gives that

$$-\log Q(\tau_n, 2\tau_n) \lesssim \tau_n^{-2} = n^{(2\beta-2\alpha+d)/(2\beta+2\gamma+d)} \le n^{d/(2\beta+2\gamma+d)} = n\varepsilon_n^2,$$

if $\beta \le \alpha$. If $0 < \beta < \alpha - d/2$, then $\tau_n \to \infty$, and Definition 7.1 on Q gives that

$$\begin{aligned}-\log Q(\tau_n, 2\tau_n) &\lesssim \tau_n^{d/(\alpha-d/2)} = n^{(d(\alpha-d/2-\beta)/(\alpha-d/2)(2\beta+2\gamma+d))}\\ &\le n^{d/(2\beta+2\gamma+d)} = n\varepsilon_n^2.\end{aligned}$$

Finally if $\alpha - d/2 = \beta$, then $\tau_n = 1$ and $Q(\tau_n, 2\tau_n) \gtrsim 1$. Thus in all three cases $Q(\tau_n, 2\tau_n)$ is bounded below by a power of $e^{-n\varepsilon_n^2}$. Combining this with the preceding, we see that $\Pi(f : \|Af - Af_0\| \le \varepsilon_n) \ge e^{-a_2 n\varepsilon_n^2}$, for some positive constant a_2, which we can take bigger than 1. Then (3.5) is satisfied for ε_n equal to $\sqrt{a_2}$ times the current ε_n.

Verification of (3.12). Lemma 9.5 gives that

$$\Pi_\tau\big(f : \|f - f_0\|_0 < 2\eta_{n,\tau}\big) \le \Pi_\tau\big(f : \|f\|_0 < 2\eta_{n,\tau}\big) \le e^{-a_3(\tau/\eta_{n,\tau})^{d/(\alpha-d/2)}},$$

for some constant a_3. This is bounded above by $e^{-4a_2 n\varepsilon_n^2}$ if

$$\eta_{n,\tau} = 2a_4 \tau n^{(d/2-\alpha)/(2\beta+2\gamma+d)} = 2a_4 \tau \eta_n/\tau_n,$$

for a sufficiently small constant $a_4 > 0$.

Verification of (3.11). Choosing $x = a_4\eta_n/\tau_n = \eta_{n,\tau}/(2\tau)$ in Lemma 9.6, we see that the left side of (3.11) is bounded above by $e^{-4a_2 n\varepsilon_n^2}$ if j_n satisfies

$$a j_n^{1/2-\alpha/d} \le a_4\eta_n/\tau_n \quad \text{and} \quad b a_4^2 (\eta_n/\tau_n)^2 j_n^{2\alpha/d} \ge 4a_2 n\varepsilon_n^2.$$

Both inequalities become equalities for j_n of the order $j_n \asymp n^{d/(2\beta+2\gamma+d)}$, as indicated at the beginning of the proof. Since $1/2 - \alpha/d < 0$ and $2\alpha/d > 0$, the left side of the first inequality is decreasing in j_n and the left side of second inequality is increasing. Thus both inequalities are satisfied for $j_n = a_5 n^{d/(2\beta+2\gamma+d)}$ and a sufficienty large constant a_5.

Finally we choose ε_n and j_n in Theorem 3.3 equal to $\sqrt{a_2}$ and a_5 times the orders indicated at the beginning of the proof. Then (3.2) is satisfied, and (3.3) and (3.4) are satisfied if η_n is chosen of the indicated order times a sufficiently large constant.

Appendix A: Galerkin projection

In this section we collect some (well known) results on the Galerkin method. Consider a scale of smoothness classes $(H_s)_{s\in\mathbb{R}}$ as in Definition 2.1.

Lemma A.1. *If V_j is a finite-dimensional space as in Assumption 2.3 such that (2.2) and (2.3) hold, then, for $P_j : H_0 \to V_j$ the orthogonal projection onto V_j, and $0 \le s, t < S$,*

$$\|f - P_j f\|_{-t} \lesssim \delta(j,t)\delta(j,s)\|f\|_s, \quad f \in H_0, \tag{A.1}$$

$$\|g\|_s \lesssim \frac{1}{\delta(j,s)\delta(j,t)}\|g\|_{-t}, \quad g \in V_j. \tag{A.2}$$

Proof. By the dual norm relation in (ii) of Definition 2.1, and the orthogonality of $f - P_j f$ to V_j,

$$\|f - P_j f\|_{-t} = \sup_{\|g\|_t \le 1} \langle f - P_j f, g\rangle_0 = \sup_{\|g\|_t \le 1} \langle f - P_j f, g - P_j g\rangle_0$$
$$\le \|f - P_j f\|_0 \sup_{\|g\|_t \le 1} \|g - P_j g\|_0,$$

by the Cauchy–Schwarz inequality. Here $\|f - P_j f\|_0 \lesssim \delta(j,s)\|f\|_s$ and $\|g - P_j g\|_0 \lesssim \delta(j,t)\|g\|_t$, both by (2.2). Inequality (A.1) follows.

For the second inequality we have, for $g \in V_j$,

$$\|g\|_0 = \sup_{f \in V_j : \|f\|_0 \le 1} \langle g, f\rangle_0 \lesssim \sup_{f \in V_j : \|f\|_0 \le 1} \|g\|_{-t}\|f\|_t,$$

again by the dual norm relation. Here we can bound $\|f\|_t$ by $\|f\|_0/\delta(j,t)$, with the help of (2.3). We obtain (A.2) by first bounding $\|g\|_s$ with the help of (2.3) and next using the preceding display. □

Let $A : H \to G$ be an injective bounded operator between separable Hilbert spaces, and let V_j be a finite-dimensional subspace of H. The Galerkin solution $f^{(j)} \in V_j$ to the image Af of an element f is defined (also see Section 3) as the element in V_j such that $Af^{(j)}$ is equal to the orthogonal projection of Af onto the image space $W_j = AV_j$. Thus, if $Q_j : G \to W_j$ denotes the orthogonal projection onto W_j, then the Galerkin solution can be written as

$$f^{(j)} = R_j A f \quad \text{for } R_j = A^{-1} Q_j,$$

where the inverse A^{-1} is well defined on the linear subspace W_j.

If the operators $R_j A$ are uniformly bounded with respect to j, then the convergence rate $\|f^{(j)} - f\|_0$ of the Galerkin solution to f is known to be of the same order as the distance $\|P_j f - f\|_0$ of f to its projection on V_j. (See Section 3.2 and Theorem 3.7 in [29], or the proof below.) In particular, if $f \in H_s$ and V_j satisfies (2.2), then the convergence rate is given by $\delta(j,s)$.

In order to control the stochastic noise term ξ in the observation scheme (1.1), it is necessary also to control the norms of the operators R_j. The following lemma summarizes the properties of the Galerkin projection needed in the proof of our main result.

Lemma A.2. *If V_j is a finite-dimensional space as in Assumption 2.3 such that (2.2) and (2.3) 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*

$$\|R_j\| \lesssim_A \frac{1}{\delta(j,\gamma)}, \tag{A.3}$$

$$\|R_j A\| \lesssim_A 1. \tag{A.4}$$

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

$$\big\|f^{(j)} - f\big\|_0 \lesssim_A \delta(j,s)\|f\|_s. \tag{A.5}$$

Proof. For $g \in G$ we have $R_j g \in V_j$ and hence by (A.2),

$$\|R_j g\|_0 \lesssim \frac{1}{\delta(j,\gamma)}\|R_j g\|_{-\gamma} \simeq \frac{1}{\delta(j,\gamma)}\|AR_j g\| = \frac{1}{\delta(j,\gamma)}\|Q_j g\|,$$

since $AR_j = Q_j$. Because $\|Q_j g\| \le \|g\|$, we conclude that $\|R_j\| \lesssim 1/\delta(j,\gamma)$.

By definition $f^{(j)} = R_j A f$, and $R_j A$ acts as the identity on V_j. Therefore $f^{(j)} - P_j f = R_j A(f - P_j f)$, and hence

$$\big\|f^{(j)} - P_j f\big\|_0 \le \|R_j\|\big\|A(f - P_j f)\big\| \simeq \|R_j\|\|f - P_j f\|_{-\gamma} \le \|R_j\|\delta(j,\gamma)\|f\|_0,$$

by (A.1). By the preceding paragraph $\|R_j\|\delta(j,\gamma) \lesssim 1$, so that the right side is bounded above by a multiple of $\|f\|_0$. By the triangle inequality

$$\|R_j A f\|_0 = \big\|f^{(j)}\big\|_0 \le \big\|f^{(j)} - P_j f\big\|_0 + \|P_j f - f\|_0 + \|f\|_0 \lesssim \|f\|_0,$$

in view of the preceding display and the fact that $\|P_j f - f\|_0 \le \|f\|_0$. This shows that $\|R_j A\| \lesssim 1$.

Finally, since $f^{(j)} - f = (R_j A - I)(f - P_j f)$, we have that

$$\|f^{(j)} - f\|_0 = \|(R_j A - I)(f - P_j f)\|_0 \leq \big(\|R_j A\| + 1\big)\|f - P_j f\|_0.$$

Inequality (A.5) follows by the boundedness of $\|R_j A\|$ and (2.2). □

As is clear from the proof, the smoothing assumption $\|Af\| \simeq \|f\|_{-\gamma}$ can be relaxed to the pair of inequalities

$$\|Af\| \lesssim \|f\|_{-\gamma}, \quad f \perp V_j, \tag{A.6}$$

$$\|Af\| \gtrsim \|f\|_{-\gamma}, \quad f \in R(R_j). \tag{A.7}$$

This helps to cover cases in which the smoothing condition is satisfied for a modification of the operator A, but not A itself, for example a modification taking different boundary conditions of a differential operator into account.

We introduce a *modified Galerkin solution* to Af to cover such a case. Let $A_0, A : H \to G$ be injective bounded operators between separable Hilbert spaces that possess a common inverse in the sense of existence of a linear map $B : D(B) \subset G \to H$ with domain $D(B)$ containing the linear span of the ranges of A_0 and A such that that $BA_0 = I = BA$. For simplicity of notation, write $B = A^- = A_0^-$. Intuitively, for the inverse problem, taking $A_0 f$ or Af as input data should be equivalent. However, it may be that A_0 is smoothing in a given scale $(H_s)_{s\in\mathbb{R}}$, whereas A is not. In that case we reconstruct as follows. Assume that $\Phi = A - A_0$ has closed range, and let $P_\Phi : G \to G$ be the orthogonal projection onto this range. Now let $Q_j : G \to G$ be the orthogonal projection onto the finite-dimensional space $(I - P_\Phi)AV_j$, and set

$$f^{(j)} = R_j Af, \quad \text{for } R_j = A^- Q_j (I - P_\Phi). \tag{A.8}$$

Thus after removing the "irrelevant part" of Af that does not influence the inversion, we project onto the finite-dimensional space $(I - P_\Phi)AV_j$ of similarly cleaned functions Af with $f \in V_j$, and finally invert.

Lemma A.3. *If V_j is a finite-dimensional space as in Assumption 2.3 such that (2.2) and (2.3) hold, and $A_0, A : H_0 \to G$ are bounded linear operators with common inverse satisfying $\|A_0 f\| \simeq \|f\|_{-\gamma}$ for every $f \in H_0$, then the operators $R_j : G \to H_0$ and $R_j A : H_0 \to H_0$ and $f^{(j)} = R_j Af$ as in (A.8) satisfy (A.3), (A.4) and (A.5).*

Proof. The operator $[A] : H \to G/\Phi(H)$ mapping $f \in H$ into the class of Af in the quotient space $G/\Phi(H)$ is one-to-one, since $[Af] = 0$ implies $Af \in R(\Phi)$ and hence $f = BAf = 0$, since $B\Phi = 0$. Identifying $[g] \in \tilde{G} := G/\Phi(H)$ with the function $(I - P_\Phi)g$ with norm $\|[g]\|_{\tilde{G}} = \|(I - P_\Phi)g\|_G$, we see that $R_j Af$ as in (A.8) is actually the Galerkin solution to $[A]f$. It suffices to show that $[A] : H \to \tilde{G}$ is smoothing in the sense of (A.6). Now $\|[Af]\|_{\tilde{G}} = \|(I - P_\Phi)A_0 f\|_G \leq \|A_0 f\|_G \simeq \|f\|_{-\gamma}$, for every $f \in H$. Furthermore, for every f such that $A_0 f \perp R(\Phi)$, the inequality is an equality. This is true for $f = R_j g$, since $A_0 R_j g = Q_j (I - P_\Phi)g \in (I - P_\Phi)AV_j$. □

Appendix B: Approximation numbers and metric entropy

The jth *approximation number* of a bounded linear operator $T : G \to H$ between normed spaces is defined as

$$a_j(T : G \to H) = \inf_{U:\operatorname{Rank} U < j} \|T - U\|_{G\to H}, \tag{B.1}$$

where the infimum is taken over all linear operators $U : G \to H$ of rank (i.e., dimension of the range space) strictly less than j, and the norm on the right is the operator norm $\|T - U\|_{G\to H} = \sup_{f:\|f\|_G \leq 1} \|(T - U)f\|_H$. The approximation numbers measure the possibility of approximating an operator by simpler operators of finite-dimensional rank. There is a rich literature on approximation numbers. The main purpose of the present section is to note their relationship to singular values and to metric entropy. Metric entropy plays an important role in the characterization of contraction rates of Bayesian posterior distributions.

If $G \subset H$, we can take T equal to the embedding $\iota : G \to H$, and then by linearity we see that there exists an operator U of rank smaller than j such that

$$\|f - Uf\|_H \lesssim a_j(\iota : G \to H)\|f\|_G, \quad \forall f \in G.$$

If H is a Hilbert space, then the minimizing finite-rank operator U is of course the orthogonal projection P_j on V_j. However, the approximation numbers also 'search' an optimal projection space. If we take $G = H_s$ and $H = H_0$, then the range space V_j of U satisfies the approximation property (2.2), with the numbers $\delta(j, s)$ taken equal to the approximation numbers $a_j(\iota : H_s \to H_0)$.

The approximation number is an example of an *s-number*, as introduced in [46]. In general s-numbers are defined as maps $T \mapsto (s_j(T))_{j\in\mathbb{N}}$, attaching to every operator T a sequence of nonnegative numbers $s_j(T)$, satisfying certain axiomatic properties. In general, approximation numbers attached to operators $T : H \to H$ are the 'largest' possible s-numbers, but on Hilbert spaces there is only one s-number: all s-numbers are the same (see 2.11.9 in [47]). Because the singular values are also s-numbers, the latter unicity yields the important relation that the approximation numbers of operators on Hilbert spaces are equal to their singular values. Recall here that the singular values of a compact operator $T : G \to H$ are the roots of the eigenvalues of the self-adjoint operator $T^*T : G \to G$.

The finite-rank approximations U that (nearly) achieve the infimum in the definition of the approximation numbers for different j are not a-priori ordered. However, in many cases there exists a basis $(\phi_i)_{i\in\mathbb{N}}$ such that the projections on the linear span of the first $j-1$ basis elements achieve the infimum. For Sobolev spaces e.g. spline bases, the Fourier basis, or wavelet bases are all 'optimal' in this sense (see [10,48]).

Approximation numbers are strongly connected to metric entropy. In the literature the connection is usually made through the notion of 'entropy numbers', which are defined as follows. The j-th *entropy number* $e_j(T)$ of an operator $T : G \to H$ is defined as the infimum of the numbers $\varepsilon > 0$ so that the image $T(U_G) \subset H$ of the unit ball U_G in G can be covered by 2^{j-1} balls of radius ε in H; or more formally, with U_H the unit ball in H,

$$e_j(T) = \inf\left\{\varepsilon > 0 : T(U_G) \subset \bigcup_{i=1}^{2^{j-1}} (h_i + \varepsilon U_H), \text{for some } h_1, \ldots, h_{2^{j-1}} \in H\right\}.$$

The function $j \mapsto e_j(T)$ is roughly the inverse function of the metric entropy of $T(U_G)$ relative to the metric induced by $\|\cdot\|_H$. Recall that the *metric entropy* of a metric space (U, d) is the logarithm of the covering number $N(\varepsilon, U, d)$, which is the minimal number of d-balls of radius $\varepsilon > 0$ needed to cover the space U. Presently we consider the metric entropy $H(\varepsilon, T) = \log N(\varepsilon, T(U_G), \|\cdot\|_H)$ of $T(U_G)$ under the metric of H. Roughly we have that

$$N\big(\varepsilon, T(U_G), \|\cdot\|_H\big) \simeq 2^{j-1}, \quad \text{if } e_j(T) \simeq \varepsilon.$$

If we use the logarithm at base 2, then the map $\varepsilon \mapsto H(\varepsilon, T)$ is approximately inverse to the map $j \mapsto e_j(T)$.

Now it is proved in [13] that for any operator $T : G \to H$ between Hilbert spaces with infinite-dimensional ranges:

$$e_{j+1}(T) \le 2a_{J+1}(T) \le 2\sqrt{2}e_{J+2}(T),$$

for any natural numbers j, J satisfying:

$$j \log 2 \ge 2\sum_{i=1}^{J} \log \frac{3a_i(T)}{a_{J+1}(T)}.$$

As shown in [13] this relationship between entropy numbers and approximation numbers may be solved to derive the entropy number from the approximation numbers in many cases.

The following lemma gives one example, important to the present paper.

Lemma B.1 (Metric entropy). *For a smoothness scale $(H_s)_{s\in\mathbb{R}}$ satisfying (2.2) with $\delta(j, s) = j^{-s/d}$, and $s > 0$ and $t \ge 0$,*

$$\log N\big(\varepsilon, \big\{f \in H_s : \|f\|_s \le 1\big\}, \|\cdot\|_{-t}\big) \sim \varepsilon^{-d/(s+t)}. \tag{B.2}$$

Proof. By (A.1) the approximation number $a_j(\iota : H_s \to H_{-t})$ is of the order $\delta(j, s)\delta(j, t) = j^{-(s+t)/d}$. It is shown in [13] that the entropy numbers $e_j(\iota : H_s \to H_{-t})$ are of the order $j^{-(s+t)/d}$. By the preceding reasoning this can be inverted to obtain the order of the metric entropy of the image of the unit ball in H_{-t}. □

In a similar way it is possible to invert approximation numbers that are not of the polynomial form $j^{-s/d}$. There are many examples of this type, for instance, involving additional logarithmic terms, or exponentially decreasing rates.

Acknowledgements

We thank two anonymous referees for the comments and suggestions that have led to an improved presentation (and some corrections), in particular relating to Section 5. The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

References

[1] S. Agapiou, S. Larsson and A. M. Stuart. Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Process. Appl.* **123** (10) (2013) 3828–3860. MR3084161 https://doi.org/10.1016/j.spa.2013.05.001

[2] J. Arbel, G. Gayraud and J. Rousseau. Bayesian optimal adaptive estimation using a sieve prior. *Scand. J. Stat.* **40** (3) (2013) 549–570. MR3091697 https://doi.org/10.1002/sjos.12002

[3] N. Bissantz, T. Hohage, A. Munk and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.* **45** (6) (2007) 2610–2636. MR2361904 https://doi.org/10.1137/060651884

[4] L. D. Brown and M. G. Low. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** (6) (1996) 2384–2398. MR1425958 https://doi.org/10.1214/aos/1032181159

[5] C. Canuto, M. Hussaini, A. Quarteroni and T. Zang. *Spectral Methods: Fundamentals in Single Domains. Scientific Computation*. Springer, Berlin Heidelberg, 2010. MR2340254

[6] I. Castillo and R. Nickl. Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** (4) (2013) 1999–2028. MR3127856 https://doi.org/10.1214/13-AOS1133

[7] L. Cavalier. Nonparametric statistical inverse problems. *Inverse Probl.* **24** (3) (2008) 034004. MR2421941 https://doi.org/10.1088/0266-5611/24/3/034004

[8] L. Cavalier and A. Tsybakov. Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields* **123** (3) (2002) 323–354. MR1918537 https://doi.org/10.1007/s004400100169

[9] P.-L. Chow, I. A. Ibragimov and R. Z. Khasminskii. Statistical approach to some ill-posed problems for linear partial differential equations. *Probab. Theory Related Fields* **113** (3) (1999) 421–441. MR1679030 https://doi.org/10.1007/s004400050212

[10] A. Cohen. *Numerical Analysis of Wavelet Methods*. *Studies in Mathematics and Its Applications*. Elsevier, Amsterdam, 2003. MR1990555

[11] A. Cohen, M. Hoffmann and M. Reiß. Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM J. Numer. Anal.* **42** (4) (2004) 1479–1501. MR2114287 https://doi.org/10.1137/S0036142902411793

[12] D. L. Donoho. Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Appl. Comput. Harmon. Anal.* **2** (2) (1995) 101–126. MR1325535 https://doi.org/10.1006/acha.1995.1008

[13] R. Edmunds. Inequalities between entropy and approximation numbers of compact maps. *Z. Anal. Anwend.* **7** (3) (1988) 223–227. MR0951120 https://doi.org/10.4171/ZAA/299

[14] H. Engl, M. Hanke and A. Neubauer. *Regularization of Inverse Problems*. *Mathematics and Its Applications*. Springer, Netherlands, 2000. MR1408680

[15] J.-P. Florens and A. Simoni. Regularizing priors for linear inverse problems. *Econometric Theory* **32** (01) (2016) 71–121. MR3442503 https://doi.org/10.1017/S0266466614000796

[16] S. Ghosal, J. K. Ghosh and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.* **28** (2) (2000) 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

[17] S. Ghosal, J. Lember and A. van der Vaart. Nonparametric Bayesian model selection and averaging. *Electron. J. Stat.* **2** (2008) 63–89. MR2386086 https://doi.org/10.1214/07-EJS090

[18] S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for non-iid observations. *Ann. Statist.* **35** (1) (2007) 192–223. MR2332274 https://doi.org/10.1214/009053606000001172

[19] S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2017. MR3587782 https://doi.org/10.1017/9781139029834

[20] A. Goldenshluger and S. V. Pereverzev. Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations. *Probab. Theory Related Fields* **118** (2) (2000) 169–186. MR1790080 https://doi.org/10.1007/s440-000-8013-3

[21] A. Goldenshluger and S. V. Pereverzev. On adaptive inverse estimation of linear functionals in Hilbert scales. *Bernoulli* **9** (5) (2003) 783–807. MR2047686 https://doi.org/10.3150/bj/1066418878

[22] R. Gorenflo and M. Yamamoto. Operator-theoretic treatment of linear Abel integral equations of first kind. *Jpn. J. Ind. Appl. Math.* **16** (1) (1999) 137–161. MR1676342 https://doi.org/10.1007/BF03167528

[23] G. Grubb. *Distributions and Operators*. *Graduate Texts in Mathematics*. Springer, New York, 2010. MR2453959

[24] M. Haase. *Functional Analysis: An Elementary Introduction*. *Graduate Studies in Mathematics*. Am. Math. Soc., Providence, 2014. MR3237610

[25] D. Haroske and H. Triebel. *Distributions, Sobolev Spaces, Elliptic Equations*. *EMS Monographs in Mathematics*. European Mathematical Society, 2008. MR2375667

[26] M. Hegland. Variable Hilbert scales and their interpolation inequalities with applications to Tikhonov regularization. *Appl. Anal.* **59** (1–4) (1995) 207–223. MR1378036 https://doi.org/10.1080/00036819508840400

[27] I. Ibragimov and R. Has'minskii. *Statistical Estimation: Asymptotic Theory. Stochastic Modelling and Applied Probability*. Springer, New York, 2013.

[28] I. M. Johnstone and B. W. Silverman. Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18** (1) (1990) 251–280. MR1041393 https://doi.org/10.1214/aos/1176347500

[29] A. Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems*. *Applied Mathematical Sciences*. Springer, Berlin, 2011. MR3025302 https://doi.org/10.1007/978-1-4419-8474-6

[30] B. Knapik and J.-B. Salomond. A general approach to posterior contraction in nonparametric inverse problems. *Bernoulli* **24** (3) (2018) 2091–2121. MR3757524 https://doi.org/10.3150/16-BEJ921

[31] B. Knapik, A. van der Vaart and J. van Zanten. Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39** (5) (2011) 2626–2657. MR2906881 https://doi.org/10.1214/11-AOS920

[32] B. Knapik, A. van der Vaart and J. van Zanten. Bayesian recovery of the initial condition for the heat equation. *Comm. Statist. Theory Methods* **42** (7) (2013) 1294–1313. MR3031282 https://doi.org/10.1080/03610926.2012.681417

[33] B. T. Knapik, B. T. Szabó, A. W. van der Vaart and J. H. van Zanten. Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Related Fields* **164** (3–4) (2016) 771–813. MR3477780 https://doi.org/10.1007/s00440-015-0619-7

[34] S. Krein and Y. Petunin. Scales of Banach spaces. *Russian Math. Surveys* **21** (2) (1966) 85. MR0193499

[35] J. Kuelbs and W. Li. Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal.* **116** (1) (1993) 133–157. MR1237989 https://doi.org/10.1006/jfan.1993.1107

[36] J. Kuelbs, W. Li and W. Linde. The Gaussian measure of shifted balls. *Probab. Theory Related Fields* **98** (2) (1994) 143–162. MR1258983 https://doi.org/10.1007/BF01192511

[37] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*, **23**. Springer-Verlag, Berlin, 1991. MR1102015 https://doi.org/10.1007/978-3-642-20212-4

[38] J. L. Lions and E. Magenes. *Non-Homogeneous Boundary Value Problems and Applications*, 1st edition. *Die Grundlehren der Mathematischen Wissenschaften* **1**. Springer-Verlag, Berlin Heidelberg, 1972. MR0350177

[39] B. A. Mair and F. H. Ruymgaart. Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.* **56** (5) (1996) 1424–1444. MR1409127 https://doi.org/10.1137/S0036139994264476

[40] P. Mathé and S. V. Pereverzev. Optimal discretization of inverse problems in Hilbert scales. Regularization and self-regularization of projection methods. *SIAM J. Numer. Anal.* **38** (6) (2001) 1999–2021. MR1856240 https://doi.org/10.1137/S003614299936175X

[41] F. Monard, R. Nickl and G. P. Paternain. Efficient nonparametric Bayesian inference for X-ray transforms. *Ann. Statist.* **47** (2) (2019) 1113–1147. MR3909962 https://doi.org/10.1214/18-AOS1708

[42] F. Natterer. Error bounds for Tikhonov regularization in Hilbert scales. *Appl. Anal.* **18** (1–2) (1984) 29–37. MR0762862 https://doi.org/10.1080/00036818408839508

[43] A. Neubauer. When do Sobolev spaces form a Hilbert scale? *Proc. Amer. Math. Soc.* **103** (2) (1988) 557–562. MR0943084 https://doi.org/10.2307/2047179

[44] R. Nickl. Bernstein-von Mises theorems for statistical inverse problems I: Schrödinger equation, 2018. Available at arXiv:1707.01764v3 [math.ST]. https://doi.org/10.1214/19-ejs1609

[45] R. Nickl and J. Söhl. Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *Ann. Statist.* **45** (4) (2017) 1664–1693. MR3670192 https://doi.org/10.1214/16-AOS1504

[46] A. Pietsch. s-numbers of operators in Banach spaces. *Studia Math.* **51** (3) (1974) 201–223. MR0361883 https://doi.org/10.4064/sm-51-3-201-223

[47] A. Pietsch. *Eigenvalues and S-Numbers. Mathematik und Ihre Anwendungen in Physik und Technik*. Akademische Verlagsgesellschaft Geest & Portig, 1987. MR0917067

[48] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*. *Springer Series in Computational Mathematics*. Springer, Berlin Heidelberg, 2009. MR1299729

[49] K. Ray. Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat.* **7** (2013) 2516–2549. MR3117105 https://doi.org/10.1214/13-EJS851

[50] K. Schmüdgen. *Unbounded Self-Adjoint Operators on Hilbert Space*. *Graduate Texts in Mathematics*. Springer, Netherlands, 2012. MR2953553 https://doi.org/10.1007/978-94-007-4753-1

[51] A. V. Skorohod. *Integration in Hilbert Space*, 1st edition. *Ergebnisse der Mathematik und Ihrer Grenzgebiete* **79**. Springer-Verlag, Berlin Heidelberg, 1974. MR0466482

[52] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numer.* **19** (2010) 451–559. MR2652785 https://doi.org/10.1017/S0962492910000061

[53] B. Szabó, A. van der Vaart and H. van Zanten. Empirical Bayes scaling of Gaussian priors in the white noise model. *Electron. J. Stat.* **7** (2013) 991–1018. MR3044507 https://doi.org/10.1214/13-EJS798

[54] M. Trabs. Bayesian inverse problems with unknown operators. *Inverse Probl.* **34** (8) (2018) 085001. MR3817293 https://doi.org/10.1088/1361-6420/aac3aa

[55] H. Triebel. *Function Spaces and Wavelets on Domains*. *EMS Tracts in Mathematics*. European Mathematical Society, 2008. MR2455724 https://doi.org/10.4171/019

[56] H. Triebel. *Theory of Function Spaces. Modern Birkhäuser Classics*. Springer, Basel, 2010. MR3024598

[57] A. van der Vaart. Bayesian regularization. In *Proceedings of the International Congress of Mathematicians. Volume IV* 2370–2385. Hindustan Book Agency, New Delhi, 2010. MR2827976

[58] A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** (3) (2008) 1435–1463. MR2418663 https://doi.org/10.1214/009053607000000613

[59] A. W. van der Vaart and J. H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. In *Collections* 200–222, **3**. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2008. MR2459226 https://doi.org/10.1214/074921708000000156

[60] G. Wahba. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.* **14** (4) (1977) 651–667. MR0471299 https://doi.org/10.1137/0714044

[61] Y. Xu. Reconstruction from Radon projections and orthogonal expansion on a ball. *J. Phys. A* **40** (26) (2007) 7239–7253. MR2344454 https://doi.org/10.1088/1751-8113/40/26/010

[62] Y. Xu, O. Tischenko and C. Hoeschen. Approximation and reconstruction from attenuated Radon projections. *SIAM J. Numer. Anal.* **45** (1) (2007) 108–132. MR2285847 https://doi.org/10.1137/05064388X