Bayesian inverse problems with non-conjugate priors

Kolyan Ray[*]

Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road
Cambridge CB3 0WA
United Kingdom
e-mail: k.m.ray@maths.cam.ac.uk

Abstract: We investigate the frequentist posterior contraction rate of non-parametric Bayesian procedures in linear inverse problems in both the mildly and severely ill-posed cases. A theorem is proved in a general Hilbert space setting under approximation-theoretic assumptions on the prior. The result is applied to non-conjugate priors, notably sieve and wavelet series priors, as well as in the conjugate setting. In the mildly ill-posed setting minimax optimal rates are obtained, with sieve priors being rate adaptive over Sobolev classes. In the severely ill-posed setting, oversmoothing the prior yields minimax rates. Previously established results in the conjugate setting are obtained using this method. Examples of applications include deconvolution, recovering the initial condition in the heat equation and the Radon transform.

AMS 2000 subject classifications: Primary 62G20; secondary 62G05, 62G08.
Keywords and phrases: Rate of contraction, posterior distribution, non-parametric hypothesis testing.

Contents

1. Introduction

1.1. Outline

In this paper, we consider the problem of using Bayesian methods to estimate
an unknown parameter f from an observation Y generated from the model

$$Y \equiv Y^{(n)} = Af + \frac{1}{\sqrt{n}}Z. \tag{1.1}$$

Here we assume that f is an element of a separable Hilbert space $\mathbb{H}_1$, $A :
\mathbb{H}_1 \to \mathbb{H}_2$ is a known, injective, continuous linear operator into another Hilbert
space $\mathbb{H}_2$ and Z is a Gaussian white noise. Many specific examples of regression
fall under this general framework, such as deconvolution, recovery of the initial
condition of the heat equation and the Radon transform (see Section 1.3 for
details).

In the Bayesian framework, we treat the unknown element f as a random
variable and assign to it a prior distribution Π, defined on a σ-algebra $\mathcal{B}$ of (a
subset of) the parameter space $\mathbb{H}_1$. We then condition on the observed data Y to
update this distribution to obtain the posterior distribution $\Pi(\cdot|Y)$ and thus ob-
tain a sequence of data-driven random probability distributions. The Bayesian
then draws his inference about f based entirely on the posterior distribution.
Recently, much focus has been given to the development of nonparametric pro-
cedures, where the support of Π is infinite-dimensional.

We wish to study the asymptotic behaviour of the posterior distribution under
the frequentist assumption that the data Y is generated from the model (1.1)
for some true parameter f_0. We shall measure this behaviour by considering if
and at what rate the posterior contracts to the true f_0 as $n \to \infty$ as defined
in [11]. This question has been the object of much study in recent years (see
e.g. [11, 13, 16, 29, 31] for some examples), but the situation of inverse problems
has only recently been considered and then only in the conjugate setting [1,
20, 21], where explicit posterior expressions are available. We shall use a novel
approach to study possibly non-conjugate priors; we also recover some of the
results from [20, 21].

While it is of considerable theoretical interest to understand the behaviour of
Bayesian procedures in the non-conjugate setting, there are also strong practi-
cal reasons to do so. Although non-conjugate priors are more involved from
a computational perspective, they are increasingly finding use due to their
greater modelling flexibility and interpretability [15]. Meanwhile, advances in

Markov chain sampling methods have meant that such procedures are increasingly tractable in practice (e.g. [26]). For example, in the case of sieved priors discussed below we have that, conditional on the random truncation level M, the problem reduces to the case of a finite-dimensional model with Gaussian noise. When the prior product marginals are non-Gaussian, it is therefore possible to sample from the conditional posterior distribution using a finite dimensional MCMC scheme.

Our method of proof follows the testing approach introduced in [11] and thus does not rely on explicit computation of the posterior. A key ingredient to using this approach is the construction of suitable tests for the problem

$$H_0 : f = f_0 \qquad H_A : f \in \{f : ||f - f_0||_{\mathbb{H}_1} \geq \xi_n\} \qquad (1.2)$$

with exponentially decaying type-II errors for some sequence $\xi_n \to 0$. We follow the approach of [13] of using the concentration properties of appropriate centred linear estimators to construct suitable plug-in tests. If the operator A in (1.1) is compact, it effectively "smooths" f and so makes it more difficult to distinguish between the alternatives H_0 and H_A based on the observation Y. To deal with this, we use general analogues of the Fourier techniques used in constructing linear estimators in the case of density deconvolution [24]. Due to the inverse nature of the problem, it is natural to construct such estimators using a diagonalizing basis for A. Moreover, since our approach requires good approximation properties within the support of the prior, we consider priors that are naturally characterized by (small modifications of) such a basis.

A key requirement of this testing approach is that the prior distribution assigns sufficient mass to a neighbourhood of the true parameter f_0. In this framework, this corresponds to establishing lower bounds for the probability that Af is contained in small-ball centred at Af_0 (the "small-ball problem") under the prior. The inverse nature of the problem turns out to be of assistance with this condition, since A shrinks f towards the origin. In effect, A changes the geometry of the problem by converting an $\mathbb{H}_2$-ball into a larger $\mathbb{H}_1$-ellipsoid, whose precise size increases with the level of ill-posedness. We shall rely on this notion in our proofs and expand upon the details below.

We apply our general result to prove contraction rates in a number of situations commonly arising in Bayesian inference, some adaptive and some not. For instance, in the case of sieve priors with random truncation, we show that under weak conditions in the mildly ill-posed setting, the procedure is fully rate adaptive (up to logarithmic factors) over Sobolev classes as in the direct case [2]. In the mildly ill-posed setting, similar adaptation results are obtained in the recent work of [19] using direct methods in the case of a hierarchical, conditionally Gaussian prior and an empirical Bayes approach. In the severely ill-posed case, our results suggest that one should calibrate the prior according to the operator A at hand. In this case, oversmoothing the prior by a suitable factor is sufficient to obtain a minimax rate of contraction. This is not surprising since centred linear estimators in the severely ill-posed case are often adaptive (see [24] for results on density estimation) and our tests are built around such estimators.

In this setting, unless the prior satisfies an analytic smoothness condition, the bias of the linear estimator dominates its variance [5, 24] and consequently the minimum of the prior smoothness and the unknown true smoothness determines the rate. Since we construct our tests using a bias-variance decomposition of a linear estimator, it seems reasonable that our rate will reflect this.

When considering the specific example of deconvolution, we also consider a wavelet series prior on $[0, 1]$. While it is canonical to work in the diagonalizing basis of A, in this case the Fourier basis, our results allow some flexibility in considering different yet closely related bases; in particular, this allows us to consider priors constructed using band-limited wavelets. This turns out to have useful consequences since we can use the functional characterization properties of wavelets to reflect a greater variety of prior assumptions - notably we consider Hölder smoothness assumptions in addition to Sobolev ones.

Unless otherwise stated, $\langle \cdot, \cdot \rangle_i$ and $||\cdot||_i$ denote the inner product and norm of the Hilbert space $\mathbb{H}_i$, $i = 1, 2$. For $x, y \in \mathbb{R}$ we use the notation $x \lesssim y$ to denote that $x \leq Ky$ for some universal constant K. For sequences $\{a_n\}$ and $\{b_n\}$ we write $a_n \simeq b_n$ to mean that there exist constants $C_1, C_2 > 0$ such that $C_1 a_n \leq b_n \leq C_2 a_n$ for all $n \geq 1$. We may also sometimes use the same letter to denote a constant that varies from line to line.

1.2. Linear inverse problems

The Gaussian white noise Z in (1.1) is the iso-normal or iso-Gaussian process for $\mathbb{H}_2$. Since Z is not realisable as a Gaussian random element of $\mathbb{H}_2$, we interpret the model in process form (as in [3]), that is we consider $Z = (Z_h : h \in \mathbb{H}_2)$ as a mean-zero Gaussian process with covariance $\mathbb{E}Z_h Z_{h'} = \langle h, h' \rangle_2$. In this form, (1.1) is interpreted as observing the Gaussian process $Y = (Y_h : h \in \mathbb{H}_2)$, where

$$Y_h = \langle Af, h \rangle_2 + \frac{Z_h}{\sqrt{n}}.$$

It is statistically equivalent to observe the subprocess $(Y_{h_k} : k \in \mathbb{N})$, for any orthonormal basis $\{h_k\}_{k \in \mathbb{N}}$ of $\mathbb{H}_2$. This corresponds to observing the sequence (Y_{h_k}), where Y_{h_k} are distributed as $N(\langle Af, h_k \rangle_2, n^{-1})$ independently.

As is natural in inverse problems we consider bases $\{e_k\}$ of $\mathbb{H}_1$ that diagonalize A. Denote by A^* the adjoint of the operator A. If A is a compact operator, then we can use the singular value decomposition (SVD) to obtain such a basis. Applying the spectral theorem to the compact self-adjoint operator $A^*A : \mathbb{H}_1 \to \mathbb{H}_1$, we know that A^*A has a discrete spectrum consisting of positive eigenvalues $\{\rho_k^2\}_{k \in \mathbb{N}}$ (possibly together with 0) and a corresponding orthonormal basis $\{e_k\}$ of $\mathbb{H}_1$ of eigenfunctions (see e.g. [28]). We then have a conjugate orthonormal basis $\{g_k\}$ of the range of A in $\mathbb{H}_2$ defined by the equality $Ae_k = \rho_k g_k$. Letting $f_k := \langle f, e_k \rangle_1$, the action of A on f has a simple form when considered in this basis: $Af = A(\sum_k f_k e_k) = \sum_k \rho_k f_k g_k$. Writing $Y_k := Y_{g_k}$, (1.1) is statistically equivalent to observing the sequence (Y_k) of independent observations, where Y_k has distribution $N(\rho_k f_k, n^{-1})$. The task of estimating

f thus reduces to that of estimating the sequence $\{f_k\}$ from the sequence of independent observations (Y_k).

Whilst priors based on a decomposition of f in the $\{e_k\}$ basis are frequently natural, it is often of interest to consider slightly more general types of bases. We therefore consider any basis whose elements consist of finite linear combinations of the $\{e_k\}$.

Condition 1. *Suppose that $\{\phi_k\}$ is an orthonormal basis for $\mathbb{H}_1$ such that for each k, the set $\{l : |\langle \phi_k, e_l \rangle_1| \neq 0\}$ is finite.*

This seemingly small extension actually has large implications for the possible choice of priors. For example, if the SVD is the Fourier basis (e.g. deconvolution - see Section 1.3.1 for more details), then Condition 1 corresponds to a band-limited basis. Band-limited wavelets have been used in the deconvolution setting (e.g. [17, 27]), and this allows us to use the superior characterization properties of wavelets to create priors that model Hölder smoothness conditions rather than Sobolev smoothness conditions, which we do using periodized Meyer wavelets in Section 3.3.

In any case, we shall assume the existence of such an orthonormal basis $\{e_k\}$ of eigenvectors of A^*A, though we do not necessarily assume that A is compact. The principle additional case we include is the white noise model, when A is the identity operator. If $\rho_k \to 0$, the problem is ill-posed since the noise to signal ratio of the components tends to infinity as $k \to \infty$. Recovering f from Y is then an inverse problem. The severity of this ill-posedness can be characterized by the rate of decay of $\rho_k \to 0$; the faster this rate, the more difficult the estimation problem. We shall classify the problem using the following classes that are standard in the statistical literature.

Condition (M). *We say that the problem is mildly ill-posed with regularity α if*

$$C_1(1+k^2)^{-\alpha/2} \leq |\rho_k| \leq C_2(1+k^2)^{-\alpha/2} \quad \text{as } k \to \infty$$

for some constants $C_1, C_2 > 0$ and $\alpha \geq 0$.

Condition (S). *We say that the problem is severely ill-posed with regularity β if*

$$C_1(1+k^2)^{-\alpha_0/2}e^{-c_0 k^\beta} \leq |\rho_k| \leq C_2(1+k^2)^{-\alpha_1/2}e^{-c_0 k^\beta} \quad \text{as } k \to \infty$$

for some constants $C_1, C_2, \beta > 0$ and $\alpha_0, \alpha_1 \in \mathbb{R}$.

The polynomial terms in Condition (S) are included to add flexibility, but do not characterize the problem since they are dominated by the exponential terms.

1.3. Examples

Note that if $\mathbb{H}_1 = L^2([0,1])$ and $\mathbb{H}_2 = H^1([0,1])$ then we can rewrite (1.1) in the more classical white noise form

$$dY(t) = \tilde{A}f(t)dt + n^{-1/2}dW(t),$$

where W is a standard Brownian motion on $[0,1]$ and $\tilde{A}f(t) = \frac{d}{dt}Af(t)$. In this setting, the direct case corresponds to taking A to be the identity operator, so that $\tilde{A}f(t) = f'(t)$. Our results apply to the following situations amongst others (see [7] for a general overview of inverse problems).

1.3.1. Deconvolution

A common problem in signal and image processing is periodic deconvolution (see e.g. [17]). Consider the 1-dimensional case on the torus $\mathbb{T} = [0,1)$ and, assuming that f is a 1-periodic function, define

$$Af(t) = \int_0^t f * \mu(s)ds, \qquad t \in [0,1], \tag{1.3}$$

for some known finite signed measure μ, where $f * \mu$ stands for convolution on $\mathbb{T}$ and where addition is defined modulo 1. This fits into the above framework since $||f * \mu||_{L^2} \leq ||f||_{L^2} ||\mu||_{TV}$ by the Minkowski integral inequality and where $||\cdot||_{TV}$ denotes the total variation norm for measures. For such a μ, we can therefore consider A as a map from $L^2([0,1])$ to $H^1([0,1])$. We observe Y arising from the model $dY_t = f * \mu(t)dt + n^{-1/2}dW_t$, where W is a standard Brownian motion on $[0,1]$. The SVD basis is the Fourier basis $e_k(x) = e^{2\pi ikx}$, $k \in \mathbb{Z}$, with associated eigenvalues given by the Fourier coefficients of μ, namely $\rho_k = \hat{\mu}_k = \int_0^1 e_k(x)d\mu(x)$. The problem can be either mildly (e.g. [17]) or severely ill-posed depending on the choice of measure μ. Note that the Dirac measure δ_0 is admissible under this model and corresponds to the direct observation case. This situation can be generalized to higher dimensions.

1.3.2. Heat equation

Consider the periodic boundary problem for the 1-dimensional heat equation

$$\frac{\partial}{\partial t}u(x,t) = \frac{\partial^2}{\partial x^2}u(x,t), \qquad u(x,0) = f(x), \qquad u(0,t) = u(1,t),$$

where $u : [0,1] \times [0,T] \to \mathbb{R}$ and the initial condition $f \in L^2([0,1])$ is 1-periodic. The task is to recover the initial condition f from a noisy observation of u at time T. The solution to this problem is given by

$$u(x,T) = \sqrt{2}\sum_{k=1}^{\infty} f_k e^{-\pi^2 k^2 T} \sin(k\pi x),$$

where $f_k = \langle f, e_k \rangle_{L^2}$ with $e_k(x) = \sqrt{2}\sin(k\pi x)$. Thus we can express $u(\cdot,T) = Af$ with $\rho_k = e^{-\pi^2 k^2 T}$. Recovering f from an observation $u(\cdot,T)$ corrupted by a white noise of intensity $n^{-1/2}$ thus leads to a severely ill-posed inverse problem satisfying Condition (S) with $\beta = 2$. This problem has been studied in the Bayesian context under conjugate Gaussian priors in [21].

1.3.3. Radon transform

Another example is given by the Radon transform, which is used in computerized tomography (see [18] for more details). Let $D = \{x \in \mathbb{R}^2 : ||x|| \leq 1\}$ and suppose that $f : D \to \mathbb{R}$ is some function in $L^2(D)$ (with Lebesgue measure) that we wish to estimate based on observations of the integrals of f along all lines intersecting D. Parametrize the lines by the length $s \in [0,1]$ of their perpendicular from the origin and the angle $\varphi \in [0, 2\pi)$ of the perpendicular to the x-axis. The Radon transform is defined as

$$Af(s, \varphi) = \frac{\pi}{2\sqrt{1-s^2}} \int_{-\sqrt{1-s^2}}^{\sqrt{1-s^2}} f(s\cos\varphi - t\sin\varphi, s\sin\varphi + t\cos\varphi)dt,$$

where $(s, \varphi) \in S = [0,1] \times [0, 2\pi)$. The Radon transform can be considered as a map $A : L^2(D) \to L^2(S, \mu)$, where $d\mu(s, \varphi) = 2\pi^{-1}\sqrt{1-s^2}\, ds\, d\varphi$ and consequently fits into the framework of (1.1). Considered as such, A is a bijective and bounded operator with SVD that can be computed using Zernike polynomials, leading to a mildly ill-posed problem satisfying Condition (M) with $\alpha = 1/2$ (see [18] for more details).

1.4. The posterior distribution and other preliminaries

In the non-conjugate situation, it is in general not possible to obtain a closed form expression for the posterior distribution. For $f \in \mathbb{H}_1$, let $\mathbb{P}_f$ denote the law of the model (1.1) so that Y is an iso-Gaussian process with drift Af under $\mathbb{P}_f$. Using the sequence space model, $\mathbb{P}_f$ is statistically equivalent to

$$\bigotimes_{k=1}^{\infty} N\left(\rho_k f_k, n^{-1}\right).$$

Kakutani's product martingale theorem (c.f. Theorem 2.7 of [8]) shows that for any $f \in \mathbb{H}_1$, this measure is equivalent to $\bigotimes_{k=1}^{\infty} N(0, n^{-1})$ with affinity $\exp(-\frac{n}{8}\sum_k \rho_k^2 f_k^2) > 0$. The family of distributions $(\mathbb{P}_f : f \in \mathbb{H}_1)$ is therefore dominated by the law $\mathbb{P}_0$ (denoting here the law of a pure white noise rather than the "true" law $\mathbb{P}_{f_0}$) with density

$$\frac{d\mathbb{P}_f}{d\mathbb{P}_0} = \exp\left(\sqrt{n}\sum_{k=1}^{\infty} \rho_k f_k Z_k - \frac{n}{2}\sum_{k=1}^{\infty} \rho_k^2 f_k^2\right),$$

where $Z_k = Z_{g_k}$. This is "almost" the Cameron-Martin theorem and if Z were realizable as a Gaussian element in $\mathbb{H}_2$, then this expression would reduce to $\exp(\sqrt{n}\langle Af, Z\rangle_2 - \frac{n}{2}||Af||_2^2)$. Since under $\mathbb{P}_0$, $Z_k = \sqrt{n}Y_k$, we can express the posterior distribution via Bayes' formula:

$$\Pi(B|Y) = \frac{\int_B e^{n\sum_k \rho_k f_k Y_k - \frac{n}{2}||Af||_2^2} d\Pi(f)}{\int_{\mathcal{P}} e^{n\sum_k \rho_k f_k Y_k - \frac{n}{2}||Af||_2^2} d\Pi(f)}, \qquad B \in \mathcal{B}, \qquad (1.4)$$

where $\mathcal{P}$ is the support of the prior Π. Obtaining an expression of this form for the posterior makes it possible to use the approach of Theorem 2.1 of [11], a fact that we shall use implicitly in the proof of Theorem 2.1.

We shall classify the smoothness of functions via the Sobolev scales with respect to the basis $\{e_k\}$. For $s \geq 0$ define

$$H^s(\mathbb{H}_1) := \left\{ f \in \mathbb{H}_1 : ||f||^2_{H^s(\mathbb{H}_1)} := \sum_{k=1}^{\infty} f_k^2 (1 + k^2)^s < \infty \right\},$$

where $f_k = \langle f, e_k \rangle_1$. We shall generally omit reference to the underlying space $\mathbb{H}_1$ when there is no confusion possible. For $s > 0$ we define the dual space

$$H^{-s}(\mathbb{H}_1) := (H^s(\mathbb{H}_1))^*.$$

It can be shown (Proposition 9.16 in [10]) that the operator norm on $(H^s(\mathbb{H}_1))^*$ is equivalent to the $||\cdot||_{H^{-s}(\mathbb{H}_1)}$-norm defined above (extended to negative indices), so that H^{-s} consists exactly of the linear functionals L acting on H^s for which $||L||_{H^{-s}}$ is finite. In particular, since every $f \in \mathbb{H}_1$ yields the continuous linear functional $g \mapsto \langle g, f \rangle_1$ on H^s, we can consider $\mathbb{H}_1$ as a subspace of $H^{-s}(\mathbb{H}_1)$.

Note that this concept of smoothness is intrinsically linked to the operator A through the choice of the basis $\{e_k\}$. To be precise, the space H^s should be indexed by both $\mathbb{H}_1$ and A, since it quantifies smoothness with respect to the operator A, but we omit this explicit link to simplify notation. For $\gamma > 0$, it is known [7] that the minimax rate of estimation over any fixed ball of H^γ is $n^{-\gamma/(2\alpha+2\gamma+1)}$ under Condition (M) and $(\log n)^{-\gamma/\beta}$ under Condition (S). Minimax rates are attained by a number of methods, such as generalized Tikhonov regularization amongst others [3, 7]. In general, we shall use α and β to refer to parameters quantifying the ill-posedness of the problem (1.1), γ to refer to the smoothness of the true function f_0 and δ to quantify the prior smoothness.

A key ingredient in proving contraction rates is establishing lower bounds for the small-ball probability of Af about Af_0 (see (2.6) below). As mentioned above, if A is compact then it changes the geometry of the problem by converting it into a small-ellipsoid problem in $\mathbb{H}_1$. Under Condition (M),

$$||Af||_2^2 = \left|\left| \sum_{k=1}^{\infty} \rho_k f_k e_k \right|\right|_2^2 = \sum_{k=1}^{\infty} \rho_k^2 f_k^2 \leq C_2 \sum_{k=1}^{\infty} f_k^2 (1 + k^2)^{-\alpha} = C_2 ||f||_{H^{-\alpha}}^2,$$

so that we are actually considering the small-ball probability of f under the weaker negative Sobolev norm $H^{-\alpha}$, since the dimensions of the ellipsoid correspond to the singular values of A. To establish (2.6) in the mildly ill-posed case, it is therefore sufficient to prove

$$\Pi_n(f \in \mathcal{P} : C_2 ||f - f_0||_{H^{-\alpha}} \leq \varepsilon_n) \geq e^{-Cn\varepsilon_n^2}. \tag{1.5}$$

In fact, the greater the ill-posedness of (1.1), the greater the prior mass assigned to an $\mathbb{H}_2$-neighbourhood of Af_0, and consequently the "nicer" the geometry of

the problem. As a concrete example, if $\{e_k\}$ is the Fourier basis acting on the torus $\mathbb{T} = [0,1)$, then the singular values $\{\rho_k\}$ act as Fourier multipliers and we recover the usual definition of (negative) Sobolev smoothness via Fourier series on $\mathbb{T}$. Using the same notion, Condition (S) induces an even weaker norm with exponential weighting.

2. General contraction results

To prove posterior contraction in a number of settings, we prove a general result along the lines of Theorems 2 and 3 of [13] adapted to inverse problems. We quantify the effects of the operator A through a sequence of factors $\{\delta_k\}$. Consider the set of indices

$$A_k = \{l : |\langle \phi_m, e_l \rangle_1| \neq 0 \text{ for some } 1 \leq m \leq k\} \tag{2.1}$$

and define

$$\delta_k = \inf_{i \in A_k} |\rho_i|, \tag{2.2}$$

that is we take the smallest ρ_i such that one of the first k basis elements $\phi_1, \ldots, \phi_k$ has a non-zero component in the e_i direction. By Condition 1 and since A is injective, we know that for any $k \in \mathbb{N}$, A_k is finite and consequently $\delta_k > 0$ and the $\{\delta_k\}$ form a decreasing sequence. Note that if we are working directly in the spectral basis $\{e_k\}$ with the singular values $\{\rho_k\}$ arranged in decreasing order, we simply recover $\delta_k = \rho_k$.

Theorem 2.1. *Consider the white noise model (1.1) and let $\{\phi_k\}$ be an orthonormal basis of $\mathbb{H}_1$ satisfying Condition 1. Let $\mathcal{P} \subset \mathbb{H}_1$ and let Π_n denote a sequence of priors defined on a σ-algebra of $\mathcal{P}$. Let $\varepsilon_n, \xi_n \to 0$ be sequences of positive numbers and $k_n \to \infty$ be a sequence of positive integers such that $\sqrt{n}\varepsilon_n \to \infty$ as $n \to \infty$,*

$$k_n \leq cn\varepsilon_n^2 \qquad and \qquad \frac{\varepsilon_n}{\delta_{k_n}} \leq C_1 \xi_n \tag{2.3}$$

for some $c, C_1 > 0$ and all $n \geq 1$, and where δ_k is defined by (2.2) with respect to $\{\phi_k\}$. Denote by P_m the projection operator onto the linear span of $\{\phi_k : 1 \leq k \leq m\}$ and let $\mathcal{P}_n$ be a sequence of subsets of

$$\{f \in \mathcal{P} : ||P_{k_n}(f) - f||_1 \leq C_2 \xi_n\} \tag{2.4}$$

for some $C_2 > 0$. Moreover, assume that there exists $C > 0$ such that, for sufficiently large n,

$$\Pi_n(\mathcal{P}_n^c) \leq e^{-(C+4)n\varepsilon_n^2}, \tag{2.5}$$

$$\Pi_n(f \in \mathcal{P} : ||Af - Af_0||_2 \leq \varepsilon_n) \geq e^{-Cn\varepsilon_n^2}. \tag{2.6}$$

Suppose that Y has law $\mathbb{P}_{f_0}$, where $f_0 \in \mathbb{H}_1$ is such that $||P_{k_n}(f_0) - f_0||_1 = O(\xi_n)$. Then there exists a constant $M < \infty$ such that

$$\Pi(f \in \mathcal{P} : ||f - f_0||_1 \geq M\xi_n | Y) \to 0$$

as $n \to \infty$ in $\mathbb{P}_{f_0}$-probability.

In an analogy to the frequentist approach, the quantity $\varepsilon_n/\delta_{k_n}$ in (2.3) represents the variance term of the centred linear estimator used to test (1.2), while ξ_n represents its bias. In the mildly ill-posed setting of Condition (M), the optimal outcome is to balance these terms so that (2.3) is an equality (up to constants). Taking $k_n \simeq n\varepsilon_n^2$ gives the optimal result using this method, yielding rate $\xi_n \simeq n^\alpha \varepsilon_n^{2\alpha+1}$.

In the severely ill-posed setting of Condition (S) it is known (see [5] for the case of density deconvolution) that the bias strictly dominates the variance as long as the true function is "rougher" than the operator A. By this we mean that if f_0 strictly falls within some Sobolev class, or satisfies some weaker analytic condition than Condition (S), then ξ_n will be of strictly larger order than $\varepsilon_n/\delta_{k_n}$ so that (2.3) will be a strict inequality (which must be verified in practice) and we take $k_n = o(n\varepsilon_n^2)$ as $n \to \infty$. Since our method relies on the approximation properties of the prior, the prior bias is equally important as the true bias in determining the contraction rate in this case.

3. Main results

We analyse the contraction properties of a number of priors in the inverse problem setting under the assumption that Y has law $\mathbb{P}_{f_0}$ for some unknown $f_0 \in \mathbb{H}_1$.

3.1. Sieve priors

Consider a sieve prior in the orthonormal basis $\{e_k\}$ that diagonalizes the operator A^*A. We take

$$f = \sum_{k=1}^{M} f_k e_k, \tag{3.1}$$

where M has probability mass function h on $\mathbb{N}$ with distribution function H. We take the $\{f_k\}$ to be independent (real or complex as required) random variables with density $\tau_k^{-1} q(\tau_k^{-1} \cdot)$, for some sequence $\{\tau_k\}$ to be specified below, and for q some fixed density. The prior can thus be expressed as

$$\Pi = \sum_{m=1}^{\infty} h(m) \Pi_m,$$

where $\Pi_m(x_1, \ldots, x_m) = \prod_{k=1}^{m} \frac{1}{\tau_k} q(\frac{x_k}{\tau_k})$. Priors of this form have been studied (e.g. [29, 34]) and, under suitable conditions on h and Π_m, are adaptive over Sobolev smoothness classes in the non ill-posed case [2, 16]. Upon suitable calibration of the prior with respect to A, this adaptation property extends to the ill-posed case when considered over the classes $H^\gamma(\mathbb{H}_1)$ for $\gamma > 0$. We firstly make the following assumption on q.

Condition 2. *The density* $q : \mathbb{R}(\text{or } \mathbb{C}) \to [0, \infty)$ *satisfies*

$$De^{-d|x|^w} \leq q(x)$$

for all $x \in \mathbb{R}$ *(or* $\mathbb{C}$*) and some constants* $D, d > 0$ *and* $w \geq 1$.

Condition 2 is very mild and requires only that q is is supported on the whole of $\mathbb{R}$ (or $\mathbb{C}$) and does not decay faster than any exponentiated polynomial; this includes many standard densities, such as the Gaussian, Laplace, Cauchy and Student's t-distributions. Our first result shows that if the true parameter is actually of the form (3.1), then in the mildly ill-posed case we recover a $\sqrt{n}$-rate up to a logarithmic factor.

Proposition 3.1. *Suppose that A satisfies Condition (M) with regularity α and that the true function f_0 is a finite series in the $\{e_k\}$-basis. Let $0 < h(m) \leq Be^{-bm}$ for some constants $B, b > 0$ and all $m \in \mathbb{N}$ and suppose that the density q satisfies Condition 2 for some $w \geq 1$. Then for a sufficiently large constant $C > 0$,*

$$\Pi\left(f \in \mathbb{H}_1 : ||f - f_0||_1 > C\,\frac{(\log n)^{\alpha+1/2}}{\sqrt{n}} \,\middle|\, Y \right) \to 0$$

in $\mathbb{P}_{f_0}$-probability as $n \to \infty$.

When the true regression function is not exactly of this form, we naturally expect a nonparametric rate of convergence. The next result deals with the case where we consider a general function lying in some Sobolev class H^γ, $\gamma > 0$. We introduce a parameter $\gamma_0 \leq \gamma$ that represents a known a-priori lower bound on the unknown smoothness and allows use of a more tightly concentrated prior. Note that the choice $\gamma_0 = 0$ is valid in the following theorem and so a non-trivial lower bound is not necessarily assumed.

Proposition 3.2. *Suppose that the true function f_0 is in $H^\gamma(\mathbb{H}_1)$ for some $\gamma > 0$ and that A satisfies Condition (M) with regularity α. Consider the prior Π described above with $B_1 e^{-b_1 m} \leq h(m) \leq B_2 e^{-b_2 m}$ for all $m \in \mathbb{N}$, for some constants $B_1, B_2, b_1, b_2 > 0$, and with density q satisfying Condition 2 for some $w \geq 1$. Suppose moreover that the scale parameters satisfy $B_3(1 + k^2)^{-\gamma_0/2}(\log k)^{-1/w} \leq \tau_k \leq B_4(1 + k^2)^{(\alpha+1)/2}$ for some $B_3, B_4 > 0$ and $\gamma_0 \leq \gamma$. Then for a sufficiently large constant $C > 0$,*

$$\Pi\left(f \in \mathbb{H}_1 : ||f - f_0||_1 > C\,\frac{(\log n)^\eta}{n^{\gamma/(2\alpha+2\gamma+1)}} \,\middle|\, Y \right) \to 0$$

in $\mathbb{P}_{f_0}$-probability as $n \to \infty$, where $\eta = \frac{(2\alpha+1)(\alpha+\gamma)}{2\alpha+2\gamma+1}$.

We firstly note that this prior gives a fully adaptive convergence rate over all the Sobolev classes H^γ up to a logarithmic factor, with this rate being uniform over f_0 in balls in H^γ. Expressed in classical regularization terminology, we have that the rate does not saturate as the truth becomes smoother.

It is worth commenting on the bounds needed on $\{\tau_k\}$, both of which are used to establish the small-ball condition (2.6), and which depend on the operator A and the lower bound γ_0. Note that the choices $\tau_k \equiv \tau$ for all k, corresponding to the $\{f_k\}$ being i.i.d., or decaying coefficients $\tau_k \asymp (\log k)^{-1/w}$ both satisfy the conditions of Proposition 3.2 and require no assumptions on the unknown smoothness. The requirements on $\{\tau_k\}$ are therefore no real imposition, merely adding flexibility when calibrating the prior, and the resulting procedure is

truly rate adaptive. The lower bound reflects that the prior cannot (up to a logarithmic factor) pick coefficients that decay faster than those of f_0. If a non-trivial lower bound $\gamma_0 > 0$ is a-priori known, then smoothing the prior to incorporate this information would yield a more concentrated prior, thereby reducing the size of credible sets whilst not affecting the rate. The upper bound is extremely mild and actually allows the size of the components to increase with k. It ensures that the moments of $(Af)_k$ (assuming they exist) are $O(1)$ as $k \to \infty$, so that the prior component moments cannot grow faster than the operator A can regularize them, thus allowing the use of larger variances than would be possible in the direct case ($\alpha = 0$). The conditions on h require it to be of exponential type and are needed both to control the prior mass for the bias condition (2.5) and to establish the small-ball condition (2.6). They are of the same form as in the direct case (c.f. Condition A_5 of [2]).

When working in the severely ill-posed case, we must calibrate our prior to the degree of ill-posedness (i.e. the parameter β). When the true parameter is a finite series in the $\{e_k\}$ basis, we again recover a $\sqrt{n}$-rate up to some strictly subpolynomial factor that grows more quickly than the logarithmic factor arising in the mildly ill-posed case in Proposition 3.1.

Proposition 3.3. *Suppose that A satisfies Condition (S) and that the true function f_0 is a finite series in the $\{e_k\}$-basis. Suppose that q satisfies Condition 2 for some $w \geq 1$, let $h(m) > 0$ for all $m \in \mathbb{N}$ and suppose that $1 - H(m) \lesssim e^{-bm^{\beta+1}}$ as $m \to \infty$ for some constant b. Then for a sufficiently large constant $C > 0$,*

$$\Pi\left(f \in \mathbb{H}_1 : ||f - f_0||_1 > C\,\frac{w_n}{\sqrt{n}}\,\middle|\, Y\right) \to 0$$

in $\mathbb{P}_{f_0}$-probability as $n \to \infty$, where $w_n = (\log n)^{\frac{2\alpha_0 + \beta + 1}{2(\beta+1)}} \exp(c(\log n)^{\frac{\beta}{\beta+1}})$ grows more slowly than any power of n.

Since the bias strictly dominates the variance in the severely ill-posed case, the bias resolution level k_n grows more slowly than the balancing term $n\varepsilon_n^2$ in (2.3) (which is a strict inequality). This reduces the size of the approximating sets $\mathcal{P}_n$ in Theorem 2.1, so that we need a sharper control on the tail of the distribution H of M to establish the bias condition (2.5). Since we take $k_n \simeq (\log n)^{1/\beta}$ to account the second part of (2.3), we must calibrate H according to the ill-posedness of the problem; indeed the more difficult the problem (larger β) the thinner tails we require.

From a frequentist perspective, it is entirely reasonable to calibrate the prior according to the inverse problem, since the operator A is assumed known. While from a pure Bayesian perspective this may seem unduly restrictive, the dependence of the prior on the ill-posedness factor β seems reasonable in this instance, given that the prior already makes implicit use of knowledge of the operator A through the choice of a diagonalizing basis. To the best of our knowledge, the Bayesian procedures thus far analysed from a frequentist perspective in both the mildly and severely ill-posed settings [19, 20, 21] all make strong use of knowledge of A through the choice of diagonalizing basis.

In the general case where $f_0 \in H^\gamma$, the dominating behaviour of the bias means we need a more careful control of the approximation error. We therefore assume that the density q is a standard Gaussian distribution. Note that δ in the following proposition corresponds to the Sobolev smoothness of a prior element.

Proposition 3.4. *Suppose that the true function f_0 is in $H^\gamma(\mathbb{H}_1)$ for some $\gamma > 0$ and that A satisfies Condition (S). Suppose that the prior Π satisfies $h(m) \geq B_1 e^{-bm^{\beta+1}}$ for all $m \geq 1$ and that $1 - H(m) \leq B_2 \exp(-bm^{\beta+1})$ as $m \to \infty$ for some constants $B_1, B_2, b > 0$. Suppose moreover that the density q is standard Gaussian and that the scale parameters satisfy $\tau_k = (1+k^2)^{-\delta/2-1/4}$ for some $\delta > \beta/2$. Then for a sufficiently large constant $C > 0$,*

$$\Pi\left(f \in \mathbb{H}_1 : ||f - f_0||_1 > C \left(\log n\right)^{-\frac{(\delta-\beta/2)\wedge\gamma}{\beta}} \,\Big|\, Y\right) \to 0$$

in $\mathbb{P}_{f_0}$-probability as $n \to \infty$.

Note that the two conditions on H are mutually satisfiable and that the exponential tails used in Propositions 3.1 and 3.2 satisfy this tail condition corresponding to $\beta = 0$. In the severely ill-posed case, oversmoothing the prior by a factor of $\beta/2$ yields the minimax rate of convergence. This factor increases with the ill-posedness of the problem and arises from the lower bounds used for the small-ball probability of Af. The lack of adaptation in this case results from the combination of the constraints (2.3) and (2.5), which are more stringent in the dominating bias case.

3.2. Gaussian priors

Consider now the conjugate situation where we take Π to be a Gaussian measure on $\mathbb{H}_1$. The conjugate situation provides a canonical example in that the posterior distribution can be computed explicitly in this situation, and so provides a useful reference point for the accuracy of our approach. Recall that a Gaussian distribution has support equal to the closure of its reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ (see [32] for more details); since the posterior has the same support, consistency is only achievable when Af_0 is contained in this set.

A Gaussian distribution $N(\nu, \Lambda)$ on $\mathbb{H}_1$ is characterized by a mean element $\nu \in \mathbb{H}_1$ and a covariance operator $\Lambda : \mathbb{H}_1 \to \mathbb{H}_1$, which is a positive semi-definite, self-adjoint and trace class linear operator. A random element G in $\mathbb{H}_1$ has $N(\nu, \Lambda)$ distribution if and only if the stochastic process $(\langle G, h\rangle_1 : h \in \mathbb{H}_1)$ is a Gaussian process with

$$\mathbb{E}\langle G, h\rangle_1 = \langle \nu, h\rangle_1, \qquad \text{cov}(\langle G, h\rangle_1, \langle G, h'\rangle_1) = \langle h, \Lambda h'\rangle_1.$$

We now take the prior to be a mean-zero Gaussian distribution so that $f \sim \Pi = N(0, \Lambda)$. We shall make the following assumption as in [20, 21].

Condition 3. *Suppose that the operators A^*A and Λ have the same set of eigenvectors $\{e_k\}$ with eigenvalues $\{\rho_k^2\}$ and $\{\tau_k^2\}$ respectively, with $\tau_k^2 = (1 + k^2)^{-\delta-1/2}$ and ρ_k satisfying either Condition (M) or (S) as specified.*

The parameter δ represents the smoothness of the prior in that $f \in H^s(\mathbb{H}_1)$ for all $s < \delta$ almost surely. In particular, $\mathbb{E} \lVert f \rVert_{H^s}^2 = \sum_{k=1}^\infty (1+k^2)^{s-\delta-1/2} < \infty$ if and only if $s < \delta$. The mildly ill-posed case is dealt with in [20] using the conjugacy of the prior and we recover the same rates using our testing approach combined with the results of [31] (though we do not consider the case of scaling). We firstly obtain (a subset of) the results of Theorem 4.1 of [20].

Proposition 3.5. *Suppose that A satisfies Condition (M), that $f_0 \in H^\gamma(\mathbb{H}_1)$ for some $\gamma > 0$, and assign f the Gaussian prior distribution $N(0, \Lambda)$, where Λ satisfies Condition 3. Then for a sufficiently large constant $C > 0$,*

$$\Pi \left(f \in \mathbb{H}_1 : \lVert f - f_0 \rVert_1 > C n^{-\frac{\delta \wedge \gamma}{2\alpha + 2\delta + 1}} \,\Big|\, Y \right) \to 0$$

in $\mathbb{P}_{f_0}$-probability as $n \to \infty$.

We therefore obtain the minimax rate of convergence only when the prior smoothness matches the true unknown smoothness. While this prior is not adaptive, it is reassuring that if the true smoothness is known then the optimal rate of convergence is attainable. Given that this result is obtained using the testing approach introduced in [11], it should be possible to apply the ideas of [33] in using a Gaussian random field with inverse Gamma bandwidth to construct an adaptive Gaussian prior. However, we do not pursue such an argument here since it is beyond the scope of the present article. Consider now the severely ill-posed analogue.

Proposition 3.6. *Suppose that A satisfies Condition (S), that $f_0 \in H^\gamma(\mathbb{H}_1)$ for some $\gamma > 0$, and assign f the Gaussian prior distribution $N(0, \Lambda)$, where Λ satisfies Condition 3 for some $\delta > \beta/2$. Then for a sufficiently large constant $C > 0$,*

$$\Pi \left(f \in \mathbb{H}_1 : \lVert f - f_0 \rVert_1 > C \left(\log n \right)^{-\frac{(\delta - \beta/2) \wedge \gamma}{\beta}} \,\Big|\, Y \right) \to 0$$

in $\mathbb{P}_{f_0}$-probability as $n \to \infty$.

A gap arises in our rates when the prior undersmooths (i.e. $\gamma + \beta/2 < \delta$), since in the case of the heat equation ($\beta = 2$), [21] obtain rate $(\log n)^{-\frac{\delta \wedge \gamma}{2}}$. This gap appears to arise in Lemma 5.2 from our bound for the covering number of the unit ball of the RKHS of Af, which is used to lower bound the small-ball probability of Af using the techniques of [22]. This lower bound seems difficult to improve and so this gap may be an artefact of our proof.

3.3. Uniform wavelet series

The approach used in this section can be generalized to any band-limited orthonormal basis for a general inverse problem in the sense of Condition 1. However, for ease of exposition, we restrict ourselves to the specific case of periodic deconvolution using wavelets. Therefore, consider the case of deconvolution under the standard white noise model on $[0,1]$ described in Section 1.3.1 so that

A is given by (1.3) with SVD given by the Fourier basis. Suppose that we have an a-priori belief that the true function f_0 satisfies some Hölder smoothness condition rather than a Sobolev condition. We shall expand upon the uniform wavelet series introduced in [13] by creating a hierarchical prior that uniformly distributes the wavelet coefficients on a Hölder ball of random radius.

Let (Φ, Ψ) denote the Meyer scaling and wavelet function (see [25] for more details). As usual, define the dilated and translated wavelet at resolution level j and scale position $k/2^j$ by $\Phi_{jk}(x) = 2^{j/2}\Phi(2^j x - k)$, $\Psi_{jk}(x) = 2^{j/2}\Psi(2^j x - k)$ for $j, k \in \mathbb{Z}$. The system of wavelet functions provides a multiresolution analysis of $L^2(\mathbb{R})$. By periodizing the wavelet functions

$$\phi_{jk}(x) = \sum_{m \in \mathbb{Z}} \Phi_{jk}(x + m), \qquad \psi_{jk}(x) = \sum_{m \in \mathbb{Z}} \Psi_{jk}(x + m),$$

we obtain a natural multiresolution analysis for periodic functions in $L^2([0,1])$. We thus have the following expansion for any periodic function $f \in L^2([0,1])$:

$$f = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0 k}\phi_{j_0 k} + \sum_{l=j_0}^{\infty} \sum_{k=0}^{2^l-1} \beta_{lk}\psi_{lk},$$

where the wavelet coefficients are given by $\alpha_{jk} = \langle f, \phi_{jk} \rangle_{L^2}$ and $\beta_{lk} = \langle f, \psi_{lk} \rangle_{L^2}$.

Meyer wavelets are band limited: in particular the Fourier transform $F_{\mathbb{R}}[\Psi](w) = \int_{\mathbb{R}} \Psi(x)e^{-2\pi i w x}dx$ over $\mathbb{R}$ satisfies $\mathrm{supp}(F[\Psi]) \subset \{w : |w| \in [1/3, 4/3]\}$. This implies that the periodized wavelets are themselves band-limited with $\mathrm{supp}(F_{\mathbb{T}}[\psi]) \subset \mathbb{Z} \cap \{w : |w| \in [1/3, 4/3]\}$ (c.f. Theorem 8.31 in [10]), where $F_{\mathbb{T}}[\psi](m) = \int_0^1 \psi(x)e^{-2\pi i m x}dx$ denotes the mth Fourier coefficient of ψ. In particular, each wavelet function has finite Fourier series and so the periodized Meyer wavelet basis satisfies Condition 1. As mentioned in the introduction, band-limited wavelets have been employed to great effect in the deconvolution problem by a number of authors (see for example [17, 27] for references).

In [13], it is assumed that a quantitative upper bound is known on the C^δ-norm of the unknown function. We shall relax this to the case where it is simply known that $||f_0||_{C^\delta} < \infty$. A natural way to circumvent this problem is to treat the unknown radius B of our Hölder ball as a hyperparameter and assign to it a prior distribution, thus creating a hierarchical model. Assign to B a probability distribution H, which for simplicity we restrict to the natural numbers $\mathbb{N}$, with probability mass function h. Given B, we then consider the periodic function

$$U_\delta(x) = u\phi(x) + \sum_{l=0}^{\infty} \sum_{k=0}^{2^l-1} 2^{-l(\delta+1/2)}u_{lk}\psi_{lk}(x),$$

where $u, u_k \sim U(-B, B)$ are i.i.d. Now by the wavelet characterization of the Besov spaces $B_{pq}^s([0,1])$ (see for instance Definition 1 of [13]), we have that $U_\delta \in C^\delta([0,1]) = B_{\infty\infty}^\delta([0,1])$ almost surely and in particular $||U_\delta||_{B_{\infty\infty}^\delta} \leq B$.

Denote the law of U_δ given B by $\Pi^{\delta,B}$ so that our full prior can be expressed as

$$\Pi^{\delta,H} = \sum_{r=1}^{\infty} h(r)\Pi^{\delta,r},$$

giving a sieve-type prior. We consider only the mildly ill-posed case.

Proposition 3.7. *Suppose that A is of the form* (1.3) *and satisfies Condition (M) and that f_0 is periodic and in $C^\gamma([0,1])$ for some $\gamma > 0$. Suppose that the distribution H satisfies $h(r) \geq e^{-Dr^\nu}$ for all $r \in \mathbb{N}$ and $1 - H(r) \lesssim e^{-Dr^\nu}$ as $r \to \infty$ for some constants $D > 0$ and $1/\delta < \nu \leq \infty$. Then there exists a finite constant C such that*

$$\Pi^{\delta,H}\left(f \in \mathcal{P} : ||f - f_0||_{L^2} \geq C\xi_n\,|\,Y\right) \to 0$$

in $\mathbb{P}_{f_0}$-probability as $n \to \infty$, where

$$\xi_n = \begin{cases} n^{-\frac{\delta-1/\nu}{2\alpha+2(\delta-1/\nu)+1}} & \text{if } \delta < \gamma + \frac{1}{\nu} \\ n^{-\frac{\gamma}{2\alpha+2\gamma+1}}(\log n)^\eta & \text{if } \delta = \gamma + \frac{1}{\nu} \end{cases},$$

where $\eta = \frac{(2\alpha+1)(\alpha+\gamma)}{2\alpha+2\gamma+1}$. If H satisfies the sharper tail condition $1 - H(r) \lesssim \exp(-e^{Dr^\nu})$ as $r \to \infty$ for some constants $D > 0$ and $\nu > 0$, then the rate improves to

$$\xi_n = n^{-\frac{\delta}{2\alpha+2\delta+1}}(\log n)^{\eta'}$$

for all $\delta \leq \gamma$, where $\eta' = \frac{(2\alpha+1)((\alpha+\delta)\vee(1/\nu))}{2\alpha+2\delta+1}$.

As well as the prior smoothness, the thickness of the tail of H, as measured by ν, affects the rate. When $\delta < \gamma + \frac{1}{\nu}$, we attain the optimal rate of convergence for a $(\delta - 1/\nu)$-smooth function, that is we lose $1/\nu$ degrees of smoothness. This is entirely due to the bias constraint (2.5): the bias of a typical element arising from $\Pi^{\delta,B}$ is proportional to B, and the approximation errors therefore grow on average with the thickness of the tail of H. Note that this penalty disappears (or is relegated to logarithmic terms) if we take H to have compact support ($\nu = \infty$) or a double exponential tail. We obtain the minimax rate of convergence, up to logarithmic terms, only if the prior smoothness matches the underlying smoothness of f_0 up to the correction term $\frac{1}{\nu}$. Finally, note that if we take $\nu = \infty$ and the prior oversmooths the true parameter f_0, then we do not have posterior consistency since f_0 does not lie in the support of $\Pi^{\delta,H}$.

The assumptions on H mirror those sometimes placed on the prior distribution of the scale parameter in a Dirichlet mixtures of normal distributions [12]. Our results therefore mirror those in Theorem 1 of [12] in that we lose a factor in our rates due to the hierarchical prior needing to be able to approximate the true parameter f_0. We finally note that a sharp rate is also only attained in that situation when the hyperprior on the scale parameter has compact support.

4. Proof of Theorem 2.1

A key step in the proof of Theorem 2.1 is the construction of nonparametric tests for suitably separated alternatives in $\mathbb{H}_1$. The tests are constructed based on the norm of a simple plug-in estimator of f_0, which is then split using a standard bias-variance decomposition. We require an exponential bound on the type-II error of our test and can attain this using Borell's inequality [4]. We can construct a suitable linear estimator for f_0 using band-limited (in the sense of the $\{e_k\}$-basis) elements in a similar fashion to the deconvolution density estimators based on Fourier techniques studied in [17] and [27].

Suppose that $\{\phi_k\}$ is an orthonormal basis of $\mathbb{H}_1$ satisfying Condition 1. Writing $\phi_{k,i} = \langle \phi_k, e_i \rangle_1$ and using that $\{g_k\}$ is the conjugate basis to $\{e_k\}$ for A,

$$\langle f, \phi_k \rangle_1 = \langle f, \sum_i \phi_{k,i} \rho_i^{-1} A^* g_i \rangle_1 = \langle Af, \sum_i \phi_{k,i} \rho_i^{-1} g_i \rangle_2 =: \langle Af, \tilde{\phi}_k \rangle_2,$$

where

$$\tilde{\phi}_k = \sum_i \rho_i^{-1} \phi_{k,i} g_i.$$

Recall that by Condition 1, only finitely many of the $\phi_{k,i}$ are non-zero. In particular, note that if $\phi_k = e_k$, then we simply have $\tilde{\phi}_k = \rho_k^{-1} g_k$. In this way, we derive a (not necessarily orthonormal) basis of the range of A that is conjugate to $\{\phi_k\}$. We can therefore express the coordinates of f in the $\{\phi_k\}$ basis of $\mathbb{H}_1$ in terms of the action of $\{\tilde{\phi}_k\}$ on Af. Considering this action, define

$$\tilde{y}_k := Y_{\tilde{\phi}_k} = \langle f, \phi_k \rangle_1 + \frac{1}{\sqrt{n}} \tilde{Z}_k,$$

where $\tilde{Z}_k = Z_{\tilde{\phi}_k}$ are (not necessarily independent) mean-zero Gaussian random variables with covariance $\mathbb{E}\tilde{Z}_k \tilde{Z}_l = \langle \tilde{\phi}_k, \tilde{\phi}_l \rangle_2$. Thus the sequence $\{\tilde{y}_k\}$ provides an unbiased estimator of the coefficients of the true regression function f in the basis $\{\phi_k\}$. The sequence $(\tilde{Z}_k)$ is independent if and only if $\{\tilde{\phi}_k\}$ forms an orthogonal sequence, which is the case when $\phi_k = e_k$. This suggests a natural linear estimator of f:

$$f_n = \sum_{k=1}^{k_n} \tilde{y}_k \phi_k,$$

where the resolution level k_n is to be specified. Recall that we write P_k for the orthogonal projection operator onto the linear span of $\{\phi_l : 1 \leq l \leq k\}$. The estimator f_n then decomposes immediately into its bias and variance parts

$$f_n = P_{k_n}(f) + \frac{1}{\sqrt{n}} \sum_{k=1}^{k_n} \tilde{Z}_k \phi_k.$$

We now construct an exponential inequality for the fluctuations of the random part of f_n, that is the centred term $f_n - \mathbb{E}f_n$, following the method presented

in Section 3.1 of [13]. By the Hahn-Banach theorem and the separability of $\mathbb{H}_1$, there exists a countable and dense subset B_0 of the unit ball of $\mathbb{H}_1' = \mathbb{H}_1$ such that

$$||f||_1 = \sup_{h \in B_0} |\langle h, f \rangle_1|.$$

The norm of the variance part of our estimator can thus be written

$$||f_n - \mathbb{E}f_n||_1 = \sup_{h \in B_0} \frac{1}{\sqrt{n}} \left| \sum_{k=1}^{k_n} \tilde{Z}_k \langle h, \phi_k \rangle_1 \right| =: \sup_{h \in B_0} |G(h)|,$$

where $G = (G(h) : h \in B_0)$ is a centred Gaussian process indexed by a countable set. Applying the version of Borell's inequality for the supremum of Gaussian processes ([23], page 134) gives

$$e^{-x^2/2\sigma^2} \geq \mathbb{P}\left(\sup_{h \in B_0} |G(h)| - \mathbb{E} \sup_{h \in B_0} |G(h)| \geq x \right)$$
$$= \mathbb{P}\left(||f_n - \mathbb{E}f_n||_1 - \mathbb{E}||f_n - \mathbb{E}f_n||_1 \geq x \right), \qquad (4.1)$$

where $\sigma^2 = \sup_{h \in B_0} \mathbb{E}G(h)^2$ is the weak variance of G. By Jensen's inequality, the expectation can be controlled as

$$\mathbb{E}||f_n - \mathbb{E}f_n||_1 \leq \frac{1}{\sqrt{n}} \left(\sum_{k=1}^{k_n} \mathbb{E}\tilde{Z}_k^2 \right)^{1/2} = \frac{1}{\sqrt{n}} \left(\sum_{k=1}^{k_n} ||\tilde{\phi}_k||_2^2 \right)^{1/2}.$$

Recall the definitions (2.1) and (2.2) of the sets A_k and quantities δ_k. Since the $\{\delta_k\}$ form a decreasing sequence

$$||\tilde{\phi}_k||_2^2 = \sum_{i \in A_k} \rho_i^{-2} \phi_{k,i}^2 \leq \frac{1}{\delta_k^2} \sum_{i \in A_k} \phi_{k,i}^2 \leq \frac{1}{\delta_{k_n}^2},$$

so that

$$\mathbb{E}||f_n - \mathbb{E}f_n||_1 \leq \frac{\sqrt{k_n}}{\delta_{k_n}\sqrt{n}}.$$

Considering the weak variance σ^2, we have that for $h \in B_0$,

$$n\mathbb{E}G(h)^2 = \sum_{k=1}^{k_n} \sum_{l=1}^{k_n} \langle h, \phi_k \rangle_1 \langle h, \phi_l \rangle_1 \mathbb{E}\tilde{Z}_k \tilde{Z}_l$$
$$= \sum_{k=1}^{k_n} \sum_{l=1}^{k_n} \langle h, \phi_k \rangle_1 \langle h, \phi_l \rangle_1 \langle \tilde{\phi}_k, \tilde{\phi}_l \rangle_2 = \left|\left| \sum_{k=1}^{k_n} \langle h, \phi_k \rangle_1 \tilde{\phi}_k \right|\right|_2^2.$$

While the basis $\{\tilde{\phi}_k\}$ is in general not orthogonal, it is sufficient that each finite sequence forms a Riesz sequence (whose constants vary with the number

of terms). Since the A_k's form an increasing sequence of sets and using the definition of $\tilde{\phi}_k$,

$$\left\| \sum_{k=1}^{k_n} \langle h, \phi_k \rangle_1 \tilde{\phi}_k \right\|_2^2 = \left\| \sum_{k=1}^{k_n} \langle h, \phi_k \rangle_1 \sum_{i \in A_k} \rho_i^{-1} \phi_{k,i} g_i \right\|_2^2$$

$$= \sum_{i \in A_{k_n}} \left(\sum_{k=1}^{k_n} \langle h, \phi_k \rangle_1 \rho_i^{-1} \langle \phi_k, e_i \rangle_1 \right)^2$$

$$\leq \frac{1}{\delta_{k_n}^2} \sum_{i=1}^{\infty} \left| \left\langle \sum_{k=1}^{k_n} \langle h, \phi_k \rangle_1 \phi_k, e_i \right\rangle_1 \right|^2$$

$$= \frac{1}{\delta_{k_n}^2} \left\| \sum_{k=1}^{k_n} \langle h, \phi_k \rangle \phi_k \right\|_1^2 \leq \frac{1}{\delta_{k_n}^2} \|h\|_1^2.$$

Combining these yields

$$\sigma^2 \leq \frac{1}{n\delta_{k_n}^2} \sup_{h \in B_0} \|h\|_1^2 \leq \frac{1}{n\delta_{k_n}^2}.$$

Substituting these bounds into Borell's inequality gives

$$\mathbb{P} \left(\|f_n - \mathbb{E} f_n\|_1 \geq x + \frac{\sqrt{k_n}}{\delta_{k_n} \sqrt{n}} \right) \leq \exp \left(-\frac{1}{2} n\delta_{k_n}^2 x^2 \right),$$

which, upon letting $x = \frac{\sqrt{2L}\varepsilon_n}{\delta_{k_n}}$ for some constant L, gives

$$\mathbb{P} \left(\|f_n - \mathbb{E} f_n\|_1 \geq \frac{1}{\delta_{k_n}} \left(\sqrt{2L}\varepsilon_n + \sqrt{\frac{k_n}{n}} \right) \right) \leq e^{-Ln\varepsilon_n^2}.$$

Since $k_n \leq cn\varepsilon_n^2$ for some constant $c > 0$, we have that for all $n \geq 1$,

$$\mathbb{P} \left(\|f_n - \mathbb{E} f_n\|_1 \geq M \frac{\varepsilon_n}{\delta_{k_n}} \right) \leq e^{-Ln\varepsilon_n^2} \tag{4.2}$$

for some constant $M = M(L, c)$ large enough.

Proof of Theorem 2.1. Following the proof of Theorem 2.1 in [11] almost exactly line by line, but using formula (1.4) for the posterior distribution in the inverse setting, we recover an analogous theorem for the sampling model (1.1). In particular, it is sufficient to construct tests (indicator functions) $\phi_n = \phi_n(Y; f_0)$ such that

$$\mathbb{E}_{f_0} \phi_n \to 0, \qquad \sup_{f \in \mathcal{P}_n : \|f - f_0\|_1 \geq M\xi_n} \mathbb{E}_f (1 - \phi_n) \leq e^{-(C+4)n\varepsilon_n^2}, \tag{4.3}$$

where the constant $C > 0$ matches that in (2.6) (the analogue of (2.4) in [11] for (1.1)). Recall that we are testing the hypotheses (1.2).

We can now consider the plug-in test $\phi_n(Y) = 1\{||f_n - f_0||_1 \geq M_0\xi_n\}$, where the constant M_0 is to be selected below. Recall that we have assumed that the contraction rate ξ_n satisfies $\frac{\varepsilon_n}{\delta_{k_n}} \leq c\xi_n$ for some $c > 0$ and all $n \geq 1$. The type-I error satisfies

$$\begin{aligned}
\mathbb{E}_{f_0}\phi_n &= \mathbb{P}_{f_0}(||f_n - f_0||_1 \geq M_0\xi_n) \\
&\leq \mathbb{P}_{f_0}(||f_n - \mathbb{E}_{f_0}f_n||_1 \geq M_0\xi_n - ||\mathbb{E}_{f_0}f_n - f_0||_1).
\end{aligned}$$

By hypothesis, the bias of f_0 satisfies $||P_{k_n}(f_0) - f_0||_1 \leq D\xi_n$ for some $D > 0$. Letting $L_1 > 0$ be some constant, we can take M_0 sufficiently large so that applying (4.2) gives

$$\mathbb{E}_{f_0}\phi_n \leq \mathbb{P}_{f_0}\left(||f_n - \mathbb{E}_{f_0}f_n||_1 \geq (M_0 - D)\xi_n\right) \leq e^{-L_1 n\varepsilon_n^2} \to 0$$

as $n \to \infty$.

Now consider $f \in \mathcal{P}_n$ such that $||f - f_0||_1 \geq M\xi_n$. Letting $L_2 > 0$ be some constant, we can pick M sufficiently large so that applying the triangle inequality and (4.2),

$$\begin{aligned}
\mathbb{E}_f(1 - \phi_n) &= \mathbb{P}_f(||f_n - f_0||_1 \leq M_0\xi_n) \\
&\leq \mathbb{P}_f(||f_0 - f||_1 - ||f - \mathbb{E}f_n||_1 - ||\mathbb{E}f_n - f_n||_1 \leq M_0\xi_n) \\
&\leq \mathbb{P}_f((M - C - M_0)\xi_n \leq ||\mathbb{E}f_n - f_n||_1) \leq e^{-L_2 n\varepsilon_n^2},
\end{aligned}$$

since by assumption $\sup_{f \in \mathcal{P}_n} ||f - \mathbb{E}f_n||_1 \leq C_2\xi_n$. This verifies (4.3). $\qquad\square$

5. Other proofs

Before proceeding, we recall some facts that will be used when applying Theorem 2.1 to the examples presented in Section 3. Recall that both the sieve and Gaussian priors of Sections 3.1 and 3.2 are defined directly in the spectral basis $\{e_k\}$. For simplicity, we assume below that the singular values $\{\rho_k\}$ are arranged in decreasing order so that the ill-posedness factor (2.2) takes the simple form $\delta_k = \rho_k$.

Establishing contraction results in these cases therefore reduces to verifying the conditions of Theorem 2.1: the bias conditions on the prior (2.5) and true parameter f_0, the small-ball condition (2.6) and balancing the rate (2.3). Recall also that in the mildly ill-posed case (Condition (M) with regularity α), it is optimal to balance the terms in (2.3) so that we take resolution level $k_n \simeq n\varepsilon_n^2$ yielding contraction rate $\xi_n \simeq n^\alpha \varepsilon_n^{2\alpha+1}$. In the severely ill-posed case, (2.3) is generally a strict inequality, which must be verified in practice.

5.1. Proofs of Section 3.1 (Sieve priors)

Proof of Proposition 3.1. By hypothesis, the true regression function takes the form $f_0 = \sum_{k=1}^{m_0} f_{0,k}e_k$ for some $m_0 \in \mathbb{N}$. We first verify the small-ball condition (2.6). Let f be a finite series generated from Π, conditionally on $M = m_0$. As noted in Section 1.2, since A satisfies Condition (M), it is sufficient to prove

(1.5) to establish (2.6). Therefore,

$$
\mathbb{P}\left(||f - f_0||_{H^{-\alpha}} \le \varepsilon_n\right) = \mathbb{P}\left(\sum_{k=1}^{m_0} |f_k - f_{0,k}|^2 (1+k^2)^{-\alpha} \le \varepsilon_n^2\right)
$$

$$
\ge \mathbb{P}\left(|f_k - f_{0,k}|^2 (1+k^2)^{-\alpha} \le \frac{\varepsilon_n^2}{m_0}, \text{ for } k = 1, \ldots, m_0\right)
$$

$$
= \prod_{k=1}^{m_0} \mathbb{P}\left(|f_k - f_{0,k}| \le \frac{\varepsilon_n (1+k^2)^{\alpha/2}}{\sqrt{m_0}}\right) \tag{5.1}
$$

by the independence of the f_k's. Now if X is complex-valued with density $q :$
$\mathbb{C} \to [0, \infty)$ satisfying Condition 2, then for all $z \in \mathbb{C}$ and $t > 0$,

$$
\mathbb{P}\left(|X - z| \le t\right) \ge \int_0^t \int_0^{2\pi} D e^{-d|z+re^{i\theta}|^w} dr\, d\theta
$$

$$
\ge 2\pi D \int_0^t e^{-d(|z|+r)^w} dr \ge 2\pi D t e^{-d(|z|+t)^w}. \tag{5.2}
$$

If X is real-valued, then the same estimate holds without the π term; we shall therefore stick to the real-valued case, but note that everything below holds also in the complex case with slightly different constants. Let $\alpha_{n,k} = \frac{\varepsilon_n (1+k^2)^{\alpha/2}}{\sqrt{m_0}}$ and note that for fixed k, $\alpha_{n,k} \to 0$ as $n \to \infty$ since $\varepsilon_n \to 0$. Thus there exists $E > 0$ such that $\alpha_{n,k} \le E$ for all $1 \le k \le m_0$ and $n \ge 1$. Using (5.2), we lower bound the right-hand side of (5.1) by

$$
\prod_{k=1}^{m_0} 2D \frac{\alpha_{n,k}}{\tau_k} e^{-d\tau_k^{-w}(|f_{0,k}| + \alpha_{n,k})^w}
$$

$$
\ge C_1 \exp\left(\sum_{k=1}^{m_0} \log\left(\frac{\alpha_{n,k}}{\tau_k}\right) - d \sum_{k=1}^{m_0} \tau_k^{-w} 2^{w-1} \left(|f_{0,k}|^w + \alpha_{n,k}^w\right)\right)
$$

$$
\ge C_2 \exp\left(m_0 \log \varepsilon_n + \sum_{k=1}^{m_0} \log \frac{(1+k^2)^{\alpha/2}}{\tau_k}\right)
$$

$$
\ge C_3 e^{C_4 \log \varepsilon_n},
$$

where we have used that $(a+b)^w \le 2^{w-1}(a^w + b^w)$ for $a, b \ge 0$ and $w \ge 1$. Now since m_0 is fixed and $h(m_0) > 0$ by assumption,

$$
\Pi(f \in \mathcal{P} : ||Af - Af_0||_2 \le \varepsilon_n) \ge h(m_0) C_3 e^{C_4 \log \varepsilon_n} \ge e^{C_5 \log \varepsilon_n}
$$

for some constant $C_5 > 0$. The choice $\varepsilon_n = (\frac{\log n}{n})^{1/2}$ then satisfies (2.6).

Consider now the bias constraint (2.5). Take k_n to be an integer satisfying $L_1 n \varepsilon_n^2 \le k_n \le L_2 n \varepsilon_n^2$ for some constants L_1, L_2, and let $\mathcal{P}_n = \{f \in \mathbb{H}_1 : f = \sum_{k=1}^{k_n} f_k e_k\}$. By the assumptions on h, we have $\Pi(\mathcal{P}_n^c) \le C e^{-bk_n} \le e^{-Ln\varepsilon_n^2}$, where L is a constant that can be made arbitrarily large by choosing L_1 sufficiently large. Now for all $f \in \mathcal{P}_n$, we have the trivial bias result $||f - P_{k_n}(f)||_1 =$

0, so that choosing L large enough to match the constant used to establish (2.6) above, we verify (2.5). Finally, for the true function f_0 the bias condition follows immediately since $||f_0 - P_{k_n} f_0||_1 = 0$ for $k_n \geq m_0$. Applying Theorem 2.1 with

$$\xi_n = \frac{\varepsilon_n}{\delta_{k_n}} \leq C\varepsilon_n k_n^\alpha = C' n^\alpha \varepsilon_n^{2\alpha+1} = C' \frac{(\log n)^{\alpha+1/2}}{\sqrt{n}}$$

completes the proof. $\qquad\square$

Proof of Proposition 3.2. By the triangle inequality

$$||f - f_0||_{H^{-\alpha}} \leq ||f - P_{j_n}(f_0)||_{H^{-\alpha}} + ||P_{j_n}(f_0) - f_0||_{H^{-\alpha}},$$

where j_n is to be selected below. Since $f_0 \in H^\gamma$,

$$||P_{j_n}(f_0) - f_0||_{H^{-\alpha}}^2 = \sum_{k=j_n+1}^{\infty} |f_{0,k}|^2 (1+k^2)^{-\alpha} \leq C j_n^{-2(\alpha+\gamma)} ||f_0||_{H^\gamma}^2.$$

Taking $j_n \simeq \varepsilon_n^{-\frac{1}{\alpha+\gamma}}$ gives

$$\mathbb{P}\left(||f - f_0||_{H^{-\alpha}} \leq \varepsilon_n\right) \geq \mathbb{P}\left(||P_{j_n}(f_0) - f||_{H^{-\alpha}} \leq c'\varepsilon_n\right)$$

for some $c' > 0$. Let $\alpha_{n,k} = \frac{\varepsilon_n(1+k^2)^{\alpha/2}}{\sqrt{j_n}}$, and suppose that f is a finite series in the $\{e_k\}$ basis of degree j_n. Then using (5.2) as in the proof of Proposition 3.1,

$$\mathbb{P}\left(||f - P_{j_n}(f_0)||_{H^{-\alpha}} \leq \varepsilon_n\right) \geq \prod_{k=1}^{j_n} D' \alpha_{n,k} \tau_k^{-1} e^{-d\tau_k^{-w}(|f_{0,k}|+\alpha_{n,k})^w}$$

$$\geq \exp\left(j_n \log C_1 + \sum_{k=1}^{j_n} \log\left(\frac{\alpha_{n,k}}{\tau_k}\right)\right.$$

$$\left. - C_2 \sum_{k=1}^{j_n} \tau_k^{-w}\left(|f_{0,k}|^w + \alpha_{n,k}^w\right)\right). \quad (5.3)$$

By the hypotheses on $\{\tau_k\}$, $\sum_{k=1}^{j_n} \log(\tau_k^{-1}(1+k^2)^{\alpha/2}) \geq -E_1 j_n \log j_n$ for some $E_1 > 0$. Since $f_0 \in H^\gamma$, we have $|f_{0,k}| \leq (1+k^2)^{-\gamma/2} ||f_0||_{H^\gamma} \leq C(f_0)k^{-\gamma}$ for all $k \geq 1$. Moreover, for $k \leq j_n$, note that $\alpha_{n,k} \simeq j_n^{-\alpha-\gamma-1/2}(1+k^2)^{\alpha/2} \leq E_2 j_n^{-\gamma-1/2}$. Substituting these bounds into (5.3) and using that $\tau_k \geq B_3(1+k^2)^{-\gamma_0/2}(\log k)^{-1/w}$ yields the lower bound

$$\exp\left(C_3 j_n \log \varepsilon_n - C_4 j_n \log j_n + \sum_{k=1}^{j_n} \log\left(\frac{(1+k^2)^{\alpha/2}}{\tau_k}\right)\right.$$

$$\left. - C_5 \sum_{k=1}^{j_n} \tau_k^{-w}(k^{-\gamma w} + j_n^{-(\gamma+1/2)w})\right)$$

$$\geq \exp\left(-C_6 j_n \log j_n - E_1 j_n \log j_n - C_7 \sum_{k=1}^{j_n} \log k\right) \geq \exp\left(-C_8 j_n \log j_n\right),$$

where we have also used that $\log \varepsilon_n \simeq -\log j_n$. In conclusion, using the lower bound on h, we have shown that

$$\mathbb{P}(||f - f_0||_{H^{-\alpha}} \leq \varepsilon_n) \geq h(j_n)e^{-C_9 j_n \log j_n} \geq e^{-C_{10}\varepsilon_n^{-1/(\alpha+\gamma)} \log \frac{1}{\varepsilon_n}}.$$

Condition (2.6) is then satisfied by the choice $\varepsilon_n = (\frac{\log n}{n})^{\frac{\alpha+\gamma}{2\alpha+2\gamma+1}}$.

Again take $\mathcal{P}_n = \{f = \sum_{k=1}^{k_n} f_k e_k\}$, where k_n is an integer satisfying $L_1 n\varepsilon_n^2 \leq k_n \leq L_2 n\varepsilon_n^2$. Proceeding as above, we get $||f - P_{k_n}(f)||_1 = 0$ for all $f \in \mathcal{P}_n$ and $\Pi(\mathcal{P}_n^c) \leq e^{-Ln\varepsilon_n^2}$ for a suitable constant L, thereby verifying (2.5). This yields contraction rate

$$\xi_n = \frac{\varepsilon_n}{\delta_{k_n}} \leq C\varepsilon_n(n\varepsilon_n^2)^\alpha = C(\log n)^{\frac{(2\alpha+1)(\alpha+\gamma)}{2\alpha+2\gamma+1}} n^{-\frac{\gamma}{2\alpha+2\gamma+1}}.$$

Finally, for the true regression element f_0,

$$||f_0 - P_{k_n}(f_0)||_1 \leq Ck_n^{-\gamma} ||f_0||_{H^\gamma} \simeq (n\varepsilon_n^2)^{-\gamma} = (\log n)^{-\frac{2\gamma(\alpha+\gamma)}{2\alpha+2\gamma+1}} n^{-\frac{\gamma}{2\alpha+2\gamma+1}} \leq \xi_n$$

as required. Applying Theorem 2.1 completes the proof. □

Proof of Proposition 3.3. By exactly the same reasoning as in the proof of Proposition 3.1, (2.6) is satisfied with $\varepsilon_n = \sqrt{(\log n)/n}$. Take k_n to be an integer satisfying $(L_1 \log n)^{1/(\beta+1)} \leq k_n \leq (L_2 \log n)^{1/(\beta+1)}$ for some constants L_1 and L_2. Again taking $\mathcal{P}_n = \{f = \sum_{k=1}^{k_n} f_k e_k\}$ yields $\Pi(\mathcal{P}_n^c) \lesssim e^{-bk_n^{\beta+1}} \leq e^{-Ln\varepsilon_n^2}$ for some constant L that can be made arbitrarily large by increasing L_1. This verifies (2.5) and the bias condition on f_0 follows exactly as above. Since the bias in both cases is equal to 0 for sufficiently large n, we can apply Theorem 2.1 with contraction rate

$$\xi_n = \frac{\varepsilon_n}{\delta_{k_n}} \leq C\varepsilon_n(1 + k_n^2)^{\alpha_0/2}e^{c_0 k_n^\beta} \leq C'\frac{(\log n)^{\frac{1}{2}+\frac{\alpha_0}{\beta+1}}e^{c_0(L_2 \log n)^{\beta/(\beta+1)}}}{\sqrt{n}} = \frac{w_n}{\sqrt{n}}.$$

□

Proof of Proposition 3.4. The proof is similar to that of Proposition 3.2, though we must notably keep more careful track of the constants involved due to the exponentiation resulting from the severe ill-posedness. If A satisfies Condition (S), consider the norm induced analogously to the Sobolev norm $H^{-\alpha}$ in the mildly ill-posed case:

$$||f||_A^2 := \sum_{k=1}^\infty |f_k|^2(1 + k^2)^{-\alpha_1}e^{-2c_0 k^\beta}.$$

Taking $j_n^{-(\alpha_1+\gamma)}e^{-c_0(j_n+1)^\beta} \simeq \varepsilon_n$ and using the same truncation argument as in the proof of Proposition 3.2 gives $||P_{j_n}(f_0) - f_0||_A \leq c\varepsilon_n$ for some constant $c > 0$. Thus for f a finite series of degree j_n in the $\{e_k\}$ basis (and using that q standard normal satisfies Condition 2 for $w = 2$), we can lower bound the probability $\mathbb{P}(||P_{j_n}(f_0) - f||_A \leq c\varepsilon)$ by

$$\exp\left(j_n \log C_1 + \sum_{k=1}^{j_n} \log\left(\frac{\tilde{\alpha}_{n,k}}{\tau_k}\right) - C_2 \sum_{k=1}^{j_n} \tau_k^{-2}\left(|f_{0,k}|^2 + \tilde{\alpha}_{n,k}^2\right)\right), \qquad (5.4)$$

where $\tilde\alpha_{n,k} = j_n^{-1/2}\varepsilon_n(1+k^2)^{\alpha_1/2}e^{c_0k^\beta} \leq Cj_n^{-\gamma-1/2}e^{c_0(k^\beta-(j_n+1)^\beta)} \leq Cj_n^{-\gamma-1/2}$ for $k \leq j_n$ and by the definition of j_n. Now since $\tau_k = (1+k^2)^{-\frac{\delta}{2}-\frac{1}{4}}$ and $f_0 \in H^\gamma$, we have that

$$\sum_{k=1}^{j_n}\log(\tau_k^{-1}\tilde\alpha_{n,k}) \geq j_n\log\varepsilon_n - \frac{1}{2}j_n\log j_n \geq E_1 j_n^{\beta+1},$$

$$\sum_{k=1}^{j}\tau_k^{-2}|f_{0,k}|^2 = \sum_{k=1}^{j}k^{2\delta+1-2\gamma}k^{2\gamma}|f_{0,k}|^2 \leq j^{(2\delta-2\gamma+1)\vee 0}\,||f_0||_{H^\gamma}^2,$$

$$\sum_{k=1}^{j_n}\tau_k^{-2}\tilde\alpha_{n,k}^2 \leq Cj_n^{-2(\gamma+1/2)}\sum_{k=1}^{j_n}k^{2(\delta+1/2)} \leq E_2 j_n^{1+2(\delta-\gamma)}$$

for some constants $E_1, E_2 > 0$. Substituting these into (5.4) gives the lower bound $\exp(-C_3 j_n^{1+\theta})$, where $\theta = \max(\beta, 2(\delta-\gamma))$. In conclusion, the small ball probability satisfies

$$\mathbb{P}\left(||Af - Af_0||_2 \leq \varepsilon_n\right) \geq h(j_n)e^{-C_3 j_n^{1+\theta}} \geq B_1 e^{-C_4 j_n^{1+\theta}} \geq e^{-C_5\left(\log\frac{1}{\varepsilon_n}\right)^{\frac{1+\theta}{\beta}}},$$

so that (2.6) is satisfied by the choice $\varepsilon_n = (\log n)^{\frac{1+\theta}{2\beta}}n^{-1/2}$.

Take k_n to be an integer satisfying $(a_1\log n)^{1/\beta} \leq k_n \leq (a_2\log n)^{1/\beta}$ for some constants a_1 and a_2. For this choice of k_n, (2.3) is verified for the choice $\xi_n = (\log n)^{-\frac{\delta-\theta/2}{\beta}}$:

$$\frac{\varepsilon_n}{\delta_{k_n}} \leq D\varepsilon_n(1+k_n^2)^{\alpha_0/2}e^{c_0 k_n^\beta} \leq D'(\log n)^{\frac{2\alpha_0+\beta+1}{2\beta}}e^{(c_0 a_2-1/2)\log n} = o\left(\xi_n\right)$$

as long as we take $c_0 a_2 < 1/2$. Recall that for $f \in \mathrm{supp}(\Pi_m)$ we have Karhunen-Loève expansion $f = \sum_{k=1}^{m}\tau_k\zeta_k e_k$, where $\{\zeta_k\}$ are i.i.d. standard normal random variables. Thus for any such f, we can bound the bias by $||P_{k_n}(f) - f||_1^2 \leq \sum_{k=k_n+1}^{\infty}\tau_k^2\zeta_k^2$. We verify (2.5) by applying Borell's inequality in a similar fashion to that used in the proof of Theorem 2.1. Using the same notation, write $||P_{k_n}(f) - f||_1 = \sup_{h\in B_0} G_n(h)$, where B_0 is a weak*-dense subset of $\{h \in \mathbb{H}_1 : ||h||_1 \leq 1\}$ and G_n is the Gaussian processes

$$G_n(h) = \langle h, P_{k_n}(f) - f\rangle_1 = \sum_{k=k_n+1}^{\infty}\tau_k\zeta_k\langle h, e_k\rangle_1.$$

We can control the bias and weak variance terms as follows. Using that $\sum_{k=k_n+1}^{\infty}k^{-w} \leq k_n^{1-w}/(w-1)$ for $w > 1$ and applying Jensen's inequality to the bias gives $\mathbb{E}\,||P_{k_n}(f) - f||_1 \leq \sqrt{\sum_{k=k_n+1}^{\infty}\tau_k^2} \leq k_n^{-\delta}$. For the variance, note that for any $h \in B_0$,

$$\mathbb{E}G_n(h)^2 = \sum_{k=k_n+1}^{\infty}\tau_k^2|\langle h, e_k\rangle_1|^2 \leq \tau_{k_n+1}^2\,||h||_1^2 \leq \tau_{k_n}^2 \simeq k_n^{-2\delta-1}.$$

Using these bounds, apply Borell's inequality for the supremum of a Gaussian process as in (4.1) with $x = \sqrt{2Ln\varepsilon_n^2 k_n^{-2\delta-1}}$ to obtain

$$\mathbb{P}\left(||P_{k_n}(f) - f||_1 \geq L'\left(k_n^{-\delta} + \sqrt{n\varepsilon_n^2}k_n^{-\delta-1/2}\right)\right) \leq e^{-Ln\varepsilon_n^2}, \qquad (5.5)$$

where L' is some constant that increases with L. Substituting in our choices of ε_n and k_n yields that for $n \geq N$,

$$\mathbb{P}\left(||P_{k_n}(f) - f||_1 \geq M(N, L)(\log n)^{-\frac{2\delta-\theta}{2\beta}}\right) \leq e^{-Ln\varepsilon_n^2},$$

where the constant M increases with L. Let $\mathcal{P}_n = \{f \in \mathbb{H}_1 : ||P_{k_n}(f) - f||_1 \leq M\xi_n\}$ for a sufficiently large constant M, so that $\Pi(\mathcal{P}_n^c) \leq e^{-Ln\varepsilon_n^2}$ for $\xi_n = (\log n)^{-\frac{\delta-\theta/2}{\beta}}$. This is satisfied by our above choice of ε_n and so, choosing L sufficiently large to match the constant obtained in the small-ball probability above, this verifies (2.5). Lastly, as $f_0 \in H^\gamma$, then $||P_{k_n}(f_0) - f_0||_1 \leq Ck_n^{-\gamma} = O(\xi_n)$ exactly as above. Apply Theorem 2.1 to finish. □

5.2. Proofs of Section 3.2 (Gaussian priors)

The small-ball asymptotics of a Gaussian measure in a Hilbert space have been exactly characterized by Sytaya [30] and using the techniques of large deviations in [9]. However, while exact, the asymptotic expression is rather complicated and relies on the solution of an implicit equation that does not yield an explicit rate in terms of the radius of the shrinking ball. We therefore obtain suitable lower bounds using either direct lower bound methods [14] or the link with the metric entropy of the unit ball of the RKHS [22] (both of which yield the same result).

As mentioned above, a Gaussian distribution has support equal to the closure of its RKHS $\mathbb{H}$ and so posterior consistency is only achievable when Af_0 is contained in this set. Since f is a Gaussian random variable in a Hilbert space with Karhunen-Loève expansion $f =^d \sum_k \tau_k \zeta_k e_k$, where the $\{\zeta_k\}$ are i.i.d. standard normal random variables, we can easily characterize its RKHS in terms of ellipsoids (see [32] for more details). Letting $\mathbb{H}_f$ denote the RKHS of f, we have that if $a = \sum_k a_k e_k$, then

$$a \in \mathbb{H}_f \quad \Leftrightarrow \quad ||a||_{\mathbb{H}_f}^2 := \sum_{k=1}^\infty \frac{a_k^2}{\tau_k^2} < \infty.$$

The RKHS norm therefore consists of a weighted ℓ_2-norm, weighting the eigenvectors of Λ with the inverse of its eigenvalues. Recall that the concentration function of a Gaussian random variable W in a Banach space $(\mathbb{B}, ||\cdot||)$ with RKHS $\mathbb{H}$ is defined as

$$\phi_{w_0}(\varepsilon) := \inf_{h \in \mathbb{H}:||h-w_0||<\varepsilon} ||h||_{\mathbb{H}}^2 - \log\mathbb{P}(||W|| < \varepsilon).$$

By Theorem 2.1 of [31], choosing ε_n to satisfy $\phi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2$ is sufficient to obtain the lower bound $\mathbb{P}(||W - w_0|| \leq 2\varepsilon_n) \geq e^{-n\varepsilon_n^2}$, and consequently establish (2.6).

We firstly establish upper bounds for the concentration function ϕ_{Af_0} of the Gaussian random variable Af. When the prior oversmooths the true parameter, the approximation error in $\phi_{Af_0}(\varepsilon)$ dominates as $\varepsilon \to 0$, whereas when it undersmooths the centred small ball probability dominates. This is quantified by the following lemma.

Lemma 5.1. *Suppose that $f \sim N(0, \Lambda)$, where Λ satisfies Condition 3, and let $f_0 \in H^\gamma(\mathbb{H}_1)$ for some $\gamma > 0$. Then Af is Gaussian random variable in the Hilbert space $\mathbb{H}_2$. If A satisfies Condition (M), then Af has RKHS equal to $H^{\alpha+\delta+1/2}(\mathbb{H}_2)$ (where $H^s(\mathbb{H}_2)$ is the Sobolev scale with respect to $\{g_k\}$) and the concentration function of Af satisfies*

$$\phi_{Af_0}(\varepsilon) \leq C \begin{cases} \varepsilon^{-\frac{2\delta-2\gamma+1}{\alpha+\gamma}} & \text{if } \gamma \leq \delta \\ \varepsilon^{-\frac{1}{\alpha+\delta}} & \text{if } \gamma \geq \delta \end{cases}$$

as $\varepsilon \to 0$ for some $C = C(\alpha, \delta, f_0)$. If A satisfies Condition (S) then Af has RKHS equal to

$$\mathbb{H}_{Af} = \left\{ b = \sum_{k=1}^\infty b_k g_k \in \mathbb{H}_2 : ||b||_{\mathbb{H}_{Af}}^2 = \sum_{k=1}^\infty b_k^2 (1+k^2)^{\alpha_0+\delta+1/2} e^{2c_0 k^\beta} < \infty \right\}$$

and the concentration function of Af satisfies

$$\phi_{Af_0}(\varepsilon) \leq C \begin{cases} \left(\log \frac{1}{\varepsilon}\right)^{\frac{2\delta-2\gamma+1}{\beta}} & \text{if } \gamma + \frac{\beta}{2} \leq \delta \\ \left(\log \frac{1}{\varepsilon}\right)^{1+1/\beta} & \text{if } \gamma + \frac{\beta}{2} \geq \delta \end{cases}$$

as $\varepsilon \to 0$ for some $C = C(\alpha_0, \beta, \delta, f_0)$.

Proof. It is obvious that Af is a Gaussian element in $\mathbb{H}_2$ with $Af \sim N(0, A\Lambda A^*)$. By Condition 3, $A\Lambda A^*$ has eigenvectors $\{g_k\}$ with corresponding eigenvalues $\{\tau_k^2 \rho_k^2\}$. Consider firstly the case where A satisfies Condition (M). Using the above remark about Gaussian measures in Hilbert spaces, we have that for any $b = \sum_{k=1}^\infty b_k g_k \in \mathbb{H}_2$,

$$||b||_{\mathbb{H}_{Af}}^2 = \sum_{k=1}^\infty \frac{b_k^2}{\tau_k^2 \rho_k^2} \simeq \sum_{k=1}^\infty b_k^2 (1+k^2)^{\alpha+\delta+1/2} = ||b||_{H^{\alpha+\delta+1/2}(\mathbb{H}_2)}^2,$$

so that $\mathbb{H}_{Af} = H^{\alpha+\delta+1/2}(\mathbb{H}_2)$.

Letting $f_0 = \sum_{k=1}^\infty f_{0,k} e_k$, define $h_j = \sum_{k=1}^j \rho_k f_{0,k} g_k$ to be the projection of Af_0 onto its first j coordinates in the conjugate basis $\{g_k\}$. Then

$$||h_j - Af_0||_2^2 = \sum_{k=j+1}^\infty \rho_k^2 |f_{0,k}|^2 \leq C \sum_{k=j+1}^\infty (1+k^2)^{-\alpha} |f_{0,k}|^2 \leq C(f_0, A) j^{-2\alpha-2\gamma},$$

since $f_0 \in H^\gamma$. Taking $j \simeq \varepsilon^{-1/(\alpha+\gamma)}$ gives $||h_j - Af_0||_2 \leq \varepsilon$ and

$$||h_j||^2_{\mathbb{H}_{Af}} \leq \sum_{k=1}^{j} \tau_k^{-2} f_{0,k}^2 \leq C'(f_0, A) j^{(2\delta-2\gamma+1)\vee 0} \simeq \varepsilon^{-\frac{2\delta-2\gamma+1}{\alpha+\gamma}\wedge 0},$$

thereby giving a bound on the first term of ϕ_{Af_0}. For the second term we use the explicit lower bound (4.5.2) from Example 4.5 in [14]:

$$\mathbb{P}(||Af||_2 < \varepsilon) = \mathbb{P}\left(\sum_{k=1}^{\infty}(1+k^2)^{-\alpha-\delta-1/2}\zeta_k^2 < \varepsilon^2\right)$$
$$\geq B\varepsilon^{\rho(3-w)}\exp\left(-w(1+\rho)^\rho\varepsilon^{-2\rho}\right),$$

where ζ_k are i.i.d. standard normals, $B > 0$ is a constant, $w = \alpha + \delta + 1/2$ and $\rho = (2w-1)^{-1} = (2\alpha+2\delta)^{-1}$. Using these values gives

$$\phi_0(\varepsilon) \leq -\log B - \rho(3-w)\log\varepsilon + w(1+\rho)^\rho\varepsilon^{-2\rho} \leq C\varepsilon^{-1/(\alpha+\delta)}$$

as $\varepsilon \to 0$ for some constant $C = C(\alpha, \delta, B)$. Comparing these two rates, we see that the approximation term dominates when $\gamma \leq \delta$ while the centred small-ball term term dominates when $\gamma \geq \delta$, thus giving the desired form for $\phi_{Af_0}(\varepsilon)$.

In the case of Condition (S), substituting in the lower bounds for the eigenvalues $\{\rho_k\}$ gives the specified $\mathbb{H}_{Af}$. If we repeat the approximation argument above, taking h_j with $j \simeq (\log\frac{1}{\varepsilon})^{1/\beta}$, then $||h_j - Af_0||_2 \leq \varepsilon$ and

$$||h_j||^2_{\mathbb{H}_{Af}} \leq \sum_{k=1}^{j} |f_{0,k}|^2(1+k^2)^{\delta+1/2} \leq Cj^{(2\delta-2\gamma+1)\vee 0} \simeq \left(\log\frac{1}{\varepsilon}\right)^{\frac{(2\delta-2\gamma+1)\vee 0}{\beta}}.$$

The centred small-ball probability can be dealt with using results on Gaussian processes that link this quantity to the metric entropy of the unit ball of the RKHS [22]. Applying Theorem 2 of [22] and using Lemma 5.2 below, we get $\phi_0(\varepsilon) \lesssim (\log\frac{1}{\varepsilon})^{1+1/\beta}$. It is also possible to derive this result using a careful rearrangement of the lower bounds proved in [14]. Balancing these terms we have that this quantity dominates when $\delta \leq \gamma + \frac{\beta}{2}$ and the approximation term dominates otherwise, hence the result. $\square$

Lemma 5.2. *Consider the RKHS $\mathbb{H}_{Af}$ of Af under Condition (S) as described in Lemma 5.1, and let K_{Af} denote the unit ball of $\mathbb{H}_{Af}$. Then the covering number $N(K_{Af}, ||\cdot||_{H_2}, \epsilon)$ of K_{Af} with the usual Hilbert space distance satisfies*

$$\log N(K_{Af}, ||\cdot||_{H_2}, \epsilon) \lesssim \left(\log\frac{1}{\epsilon}\right)^{1+1/\beta}.$$

Using this result and Theorem 2 of [22], we obtain the bound $-\log\mathbb{P}(||Af||_2 < \varepsilon) \leq C(\log\frac{1}{\varepsilon})^{1+1/\beta}$. This matches the bounds obtained in [6] when considering the general setting of heat kernels ($\beta = 2$) on manifolds.

Proof. Writing $b = \sum_{k=1}^{\infty} b_k g_k$, we know that for any $b \in K_{Af}$ we have $|b_k| \leq C(1+k^2)^{-\alpha_0 - \delta - 1/2} e^{-c_0 k^\beta} \leq C e^{-c_0 k^\beta}$, so that K_{Af} is contained in the infinite rectangle

$$\prod_{k=1}^{\infty} \left[-C e^{-c_0 k^\beta}, C e^{-c_0 k^\beta} \right].$$

Taking $J = D(\log \frac{1}{\epsilon})^{1/\beta}$ for a suitable constant D, we that for $k \geq J$, the width of the above intervals is smaller that $\epsilon/2$. Thus any point in the infinite rectangle is within $\epsilon/2$ of the finite dimensional cube $X = \prod_{k=1}^{J} [-C e^{-c_0 k^\beta}, C e^{-c_0 k^\beta}]$ and so it suffices to construct an $\epsilon/2$ cover for this latter set. By considering a J-dimensional cube, we see that it is enough to cover this set by a considering a regular lattice with distance $\epsilon/(2\sqrt{J})$ between adjacent vertices. Therefore

$$N\left(X, ||\cdot||_{eucl}, \frac{\epsilon}{2}\right) \leq \prod_{k=1}^{J} \frac{2C e^{-c_0 k^\beta}}{\epsilon/(2\sqrt{J})} = \left(\frac{C'\sqrt{J}}{\epsilon}\right)^J e^{-c_0 \sum_{k=1}^{J} k^\beta}.$$

Now by a simple integral comparison test, $\sum_{k=1}^{J} k^\beta \geq J^{\beta+1}/(\beta+1)$, so that the logarithm of the right-hand side is bounded above by

$$C'' J \left(\log J + \log \frac{1}{\epsilon}\right) - c_0 \frac{J^{\beta+1}}{\beta+1} \leq C''' \left(\log \frac{1}{\epsilon}\right)^{1+1/\beta}.$$

$\square$

Proof of Proposition 3.5. Let us verify the small ball Condition (2.6). Let $\mathbb{H}_{Af}$ denote the RKHS of Af and ϕ_{Af_0} denote the concentration function of Af at Af_0. Since Af is a Gaussian random element in $\mathbb{H}_2$, we have by Theorem 2.1 of [31] that if Af_0 is contained in the $\mathbb{H}_2$-closure of $\mathbb{H}_{Af}$ and ε_n satisfies $\phi_{Af_0}(\varepsilon_n) \leq n\varepsilon_n^2$, then $\mathbb{P}(||Af - Af_0||_2 < 2\varepsilon_n) \geq e^{-n\varepsilon_n^2}$. By Lemma 5.1, the choice $\varepsilon_n = n^{-\frac{\alpha+\gamma\wedge\delta}{2\alpha+2\delta+1}}$ satisfies this condition in both the cases $\gamma \geq \delta$ and $\gamma \leq \delta$, thereby verifying (2.6).

Recall that we have Karhunen-Loève expansion $f = \sum_{k=1}^{\infty} \tau_k \zeta_k e_k$, where $\{\zeta_k\}$ are i.i.d. standard normal random variables. Proceeding as in the proof of Proposition 3.4 and taking $k_n \simeq n\varepsilon_n^2$ in (5.5), we obtain that for $n \geq N$,

$$\mathbb{P}\left(||P_{k_n}(f) - f||_1 \geq M(L,N)(n\varepsilon_n^2)^{-\delta}\right) \leq e^{-Ln\varepsilon_n^2},$$

where the constant M increases with L. Let $\mathcal{P}_n = \{f \in \mathbb{H}_1 : ||P_{k_n}(f) - f||_1 \leq M\xi_n\}$ for a sufficiently large constant M, so that $\Pi(\mathcal{P}_n^c) \leq e^{-Ln\varepsilon_n^2}$ as long as $(n\varepsilon_n^2)^{-\delta} \leq C\xi_n$ for some $C > 0$. This is satisfied by our above choice of ε_n and so, choosing L sufficiently large to match the constant obtained in the small-ball probability above, this verifies (2.5). Finally, since $f_0 \in H^\gamma$, we again recover that $||P_{k_n}(f_0) - f_0||_1 \leq C k_n^{-\gamma} ||f_0||_{H^\gamma} \simeq (n\varepsilon_n^2)^{-\gamma}$, which is smaller than $\xi_n = \varepsilon_n(n\varepsilon_n^2)^\alpha$ for our choice of ε_n. Applying Theorem 2.1 completes the proof. $\square$

Proof of Proposition 3.6. Consider firstly the case where $\gamma + \frac{\beta}{2} \leq \delta$. As above, (2.6) is verified if $\phi_{Af_0}(\varepsilon_n) \leq n\varepsilon_n^2$. By Lemma 5.1, the choice $\varepsilon_n = (\log n)^{\frac{\delta-\gamma+1/2}{\beta}} n^{-1/2}$ satisfies this condition. Now let k_n be an integer satisfying $(L_1 \log n)^{1/\beta} \leq k_n \leq (L_2 \log n)^{1/\beta}$ for some constants L_1, L_2, and which therefore satisfies $k_n \leq cn\varepsilon_n^2$ for some constant c and the above choice of ε_n. The quantity in the left-hand side of (2.3) then satisfies

$$\frac{\varepsilon_n}{\delta_{k_n}} \leq C\varepsilon_n (1 + k_n^2)^{\alpha_0/2} e^{c_0 k_n^\beta} \leq C'(\log n)^\eta n^{L_2 c_0 - 1/2} = o\left((\log n)^{-\gamma/\beta}\right)$$

as $n \to \infty$ provided that $L_2 c_0 < 1/2$. To verify (2.5), substitute our choices of ε_n and k_n into (5.5) to get

$$e^{-Ln\varepsilon_n^2} \geq \mathbb{P}\left(||P_{k_n}(f) - f||_1 \geq C(\log n)^{-\frac{\delta}{\beta}} + C'(\log n)^{-\frac{\gamma}{\beta}}\right).$$

Since $\delta \geq \gamma + \frac{\beta}{2}$ the second term is asymptotically larger, so that taking L sufficiently large, we obtain the required exponential inequality (2.5) with rate $(\log n)^{-\gamma/\beta}$. Since $f_0 \in H^\gamma$, we have that exactly as above $||P_{k_n}(f_0) - f_0||_1 \leq Ck_n^{-\gamma} \leq C'(\log n)^{-\gamma/\beta}$, so that we can apply Theorem 2.1.

Consider now the case where $\gamma + \frac{\beta}{2} \geq \delta$. Arguing as above, the choice $\varepsilon_n = (\log n)^{\frac{\beta+1}{2\beta}} n^{-1/2}$ satisfies the small-ball condition (2.6) and for the bias we recover the exponential inequality

$$e^{-Ln\varepsilon_n^2} \geq \mathbb{P}\left(||P_{k_n}(f) - f||_1 \geq C(\log n)^{-\frac{(\delta-\beta/2)}{\beta}}\right).$$

By our choice of δ, the above rate is larger than the bias of f_0 and so yields the contraction rate. $\qquad\square$

5.3. Proofs of Section 3.3 (Uniform wavelet series)

Since we are working the deconvolution setting described in Section 1.3.1 we firstly note that the Sobolev scale with respect to the Fourier basis corresponds to the classical notion of Sobolev smoothness on $\mathbb{T}$, so that $H^s(\mathbb{H}_1) = H^s([0,1])$. As mentioned above, periodized Meyer wavelets are band limited and so satisfy Condition 1 which is needed for Theorem 2.1. Moreover, since $\text{supp}(F_\mathbb{T}[\psi]) \subset [-a, a]$ for some $a > 0$, we have by the standard properties of the Fourier transform that the dilated and translated wavelets satisfy $\text{supp}(F_\mathbb{T}[\psi_{jk}]) \subset [-2^j a, 2^j a]$. Recalling definition (2.2), we therefore have that under Condition (M),

$$\delta_{2^j} = \inf_{m \in \mathbb{Z}:|m| \leq 2^j a} |F_\mathbb{T}[\mu](m)| \leq C(1 + 2^{2j})^{-\alpha/2}.$$

Since the ill-posedness affects the rate ξ_n through (2.3), we see that using the periodized Meyer wavelet basis rather than the SVD (Fourier basis) only affects the constants and does not negatively affect the rate. In this section note that $||\cdot||_2$ refers to the $L^2([0,1])$-norm rather than the $\mathbb{H}_2$-norm.

Proof of Proposition 3.7. We firstly verify the small-ball condition (2.6). Consider the case where $\delta \leq \gamma$. Using the wavelet characterization of the periodic Besov space $B_{22}^s([0,1]) = H^s([0,1])$ for $s \in \mathbb{R}$ gives

$$
\begin{aligned}
||h||_{H^s}^2 &= \left\{ |\bar{\alpha}(h)| + \left(\sum_{l=0}^{\infty} \left(2^{ls} ||\beta_{l\cdot}(h)||_{\ell_2} \right)^2 \right)^{1/2} \right\}^2 \\
&\leq 4 \max \left\{ |\bar{\alpha}(h)|^2, \sum_{l=0}^{\infty} 2^{2ls} \sum_{k=0}^{2^l-1} \beta_{lk}(h)^2 \right\}.
\end{aligned}
\tag{5.6}
$$

Let $\bar{\alpha}$, β_{lk} denote the wavelet coefficients of f_0 and note that if $||f_0||_{C^\gamma} \leq B$ then $|\bar{\alpha}| \leq B$ and $|\beta_{lk}| \leq B2^{-l(\gamma+1/2)}$ for all l, k. By (5.6), we lower bound $\mathbb{P}(||f_0 - U_\delta||_{H^{-\alpha}} \leq \varepsilon_n)$ by

$$
\begin{aligned}
&\mathbb{P}\left(\max \left\{ |\bar{\alpha} - u|^2, \sum_{l=0}^{\infty} 2^{-2l\alpha} \sum_{k=0}^{2^l-1} |\beta_{lk} - 2^{-l(\delta+1/2)} u_{lk}|^2 \right\} \leq c_1 \varepsilon_n^2 \right) \\
&= \mathbb{P}\left(|\bar{\alpha} - u|^2 \leq c_1 \varepsilon_n^2 \right) \mathbb{P}\left(\sum_{l=0}^{\infty} 2^{-2l\alpha} \sum_{k=0}^{2^l-1} |\beta_{lk} - 2^{-l(\delta+1/2)} u_{lk}|^2 \leq c_1 \varepsilon_n^2 \right) \quad (5.7)
\end{aligned}
$$

using the independence of u and the u_{lk}'s. The first probability satisfies

$$
\mathbb{P}\left(|\bar{\alpha} - u| \leq \sqrt{c_1} \varepsilon_n \right) \geq \left(\frac{\sqrt{c_1} \varepsilon_n}{2B} \right) = e^{c_2 + \log(\varepsilon_n/B)} \geq e^{c_3 \log(\varepsilon_n/B)}
$$

for some constant $c_3 = c_3(\Phi, \Psi)$. Let $b_{lk} = 2^{l(\gamma+1/2)} \beta_{lk}$ and pick $J = J(n)$ as defined below. The second probability in (5.7) becomes

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{l=0}^{\infty} 2^{-l(2\alpha+2\gamma+1)} \sum_{k=0}^{2^l-1} |b_{lk} - 2^{-l(\delta-\gamma)} u_{lk}|^2 \leq c_1 \varepsilon_n^2 \right) \\
&\geq \mathbb{P}\left(\sum_{l=0}^{\infty} 2^{-2l(\alpha+\gamma)} \sup_{0 \leq k < 2^l} |b_{lk} - 2^{-l(\delta-\gamma)} u_{lk}|^2 \leq c_1 \varepsilon_n^2 \right) \\
&\geq \mathbb{P}\left(\sum_{l=0}^{J} 2^{-2l(\alpha+\gamma)} \sup_{0 \leq k < 2^l} |b_{lk} - 2^{-l(\delta-\gamma)} u_{lk}|^2 + CB^2 \sum_{l=J+1}^{\infty} 2^{-2l(\alpha+\delta)} \leq c_1 \varepsilon_n^2 \right).
\end{aligned}
$$

Pick the truncation level $J = J(n)$ so that $B^2 2^{-2J(\alpha+\delta)} \simeq \varepsilon_n^2$, that is $2^J \simeq (\varepsilon_n/B)^{-1/(\alpha+\delta)}$. Note that since $|b_{lk}| \leq B$ and $\delta \leq \gamma$, we can lower bound the individual probabilities via

$$
\mathbb{P}\left(|b_{lk} - 2^{-l(\delta-\gamma)} u_{lk}| \leq c\varepsilon_n \right) \geq \left(\frac{c\varepsilon_n}{2^{l(\gamma-\delta)+1} B} \right) > 0.
$$

Then, choosing the constants defining $J(n)$ appropriately, we have

$$\mathbb{P}\left(\sum_{l=0}^{J} 2^{-2l(\alpha+\gamma)} \sup_{0\leq k<2^l} |b_{lk} - 2^{-l(\delta-\gamma)}u_{lk}|^2 \leq c_1\varepsilon_n^2 - c(\alpha,\delta)B^2 2^{-2J(\alpha+\delta)}\right)$$

$$\geq \mathbb{P}\left(\max_{0\leq l\leq J} \sup_{0\leq k<2^l} |b_{lk} - 2^{-l(\delta-\gamma)}u_{lk}| \leq c_4\varepsilon_n\right)$$

$$= \prod_{l=0}^{J}\prod_{k=0}^{2^l-1} \mathbb{P}\left(|b_{lk} - 2^{-l(\delta-\gamma)}u_{lk}| \leq c_4\varepsilon_n\right) \geq \prod_{l=0}^{J}\prod_{k=0}^{2^l-1}\left(\frac{c_4\varepsilon_n}{2^{l(\gamma-\delta)+1}B}\right)$$

$$\geq \exp\left(c_5 \log(\varepsilon_n/B)\sum_{l=0}^{J} 2^l - c_6\sum_{l=0}^{J} l2^l\right) \geq e^{c_7(\varepsilon_n/B)^{-1/(\alpha+\delta)}\log(\varepsilon_n/B)},$$

for $n \geq N(\alpha,\delta,B,\psi)$ and we have used that $J \simeq -\log(\varepsilon_n/B)$ in the last line. Using (5.7) and that $h(B_0) > 0$ for some $B_0 \geq ||f_0||_{C^\gamma}$, we have that for $n \geq N(\alpha,\delta,B_0,\psi)$,

$$\mathbb{P}(||f_0 - U_\delta||_{H^{-\alpha}} \leq \varepsilon_n) \geq h(B_0)e^{c_3 \log(\varepsilon_n/B_0)}e^{c_7(\varepsilon_n/B_0)^{-1/(\alpha+\delta)}\log(\varepsilon_n/B_0)}$$

$$\geq e^{c_8\varepsilon_n^{-1/(\alpha+\delta)}\log \varepsilon_n}, \tag{5.8}$$

so that (2.6) is satisfied by the choice $\varepsilon_n \simeq \left(\frac{\log n}{n}\right)^{\frac{\alpha+\delta}{2\alpha+2\delta+1}}$.

Consider now the case $\gamma < \delta \leq \gamma + \frac{1}{\nu}$, where we can establish (2.6) in a similar fashion by using an approximation argument. Recall that $f_0 \in C^\gamma([0,1])$ and let h_r be the best $H^{-\alpha}$-approximation of f_0 such that $||h||_{C^\delta} \leq r$. Write $h_r = \theta_0\phi + \sum_{l=0}^{\infty}\sum_{k=0}^{2^l-1} \theta_{lk}\psi_{lk}$, where $|\theta| \leq r$ and $|\theta_{lk}| \leq r2^{-l(\delta+1/2)}$, and recall that the wavelet coefficients of f_0 satisfy $|\beta_{lk}| \leq B_0 2^{-l(\gamma+1/2)}$ for $B_0 \geq ||f_0||_{C^\gamma}$. Let l_r be the smallest integer such that $2^{l_r(\delta-\gamma)} > r/B_0$, so that in particular $\theta_{lk} = \beta_{lk}$ for all $l < l_r$. Then

$$||f_0 - h_r||_{H^{-\alpha}}^2 \leq \sum_{l=l_r}^{\infty} 2^{-2l(\alpha+\gamma)}\left(B_0 - r2^{-l(\delta-\gamma)}\right)^2 \leq CB_0^2 2^{-2l_r(\alpha+\gamma)},$$

so that $||f_0 - h_r||_{H^{-\alpha}} \leq C(f_0)r^{-\frac{\alpha+\gamma}{\delta-\gamma}}$ by the definition of l_r. Pick r_n to be the smallest integer such that $r_n^{-\frac{\alpha+\gamma}{\delta-\gamma}} \leq \frac{\varepsilon_n}{2C}$, so that by the triangle inequality, $\mathbb{P}(||f_0 - U_\delta||_{H^{-\alpha}} \leq \varepsilon_n) \geq \mathbb{P}(||h_{r_n} - U_\delta||_{H^{-\alpha}} \leq c\varepsilon_n)$ for some $1/2 \leq c < 1$. Since $||h_{r_n}||_{C^\delta} \leq r_n$, we use (5.8) to obtain

$$\mathbb{P}\left(||f_0 - U_\delta||_{H^{-\alpha}} \leq \varepsilon_n\right) \geq h(r_n)\exp\left(c_1\left(\frac{\varepsilon_n}{r_n}\right)^{-\frac{1}{\alpha+\delta}}\log\frac{\varepsilon_n}{r_n}\right).$$

Since $\delta \leq \gamma + \frac{1}{\nu}$, $h(r) \geq e^{-Dr^\nu}$ for all $r \in \mathbb{N}$, and $r_n \geq c_2\varepsilon_n^{-\frac{\delta-\gamma}{\alpha+\gamma}}$ for some $c_2 > 0$ and sufficiently large n, we obtain the lower bound $\exp(-d_2\varepsilon_n^{-\frac{1}{\alpha+\gamma}}\log\frac{1}{\varepsilon_n})$. Bounding this from below by $e^{-Cn\varepsilon_n^2}$ yields the choice $\varepsilon_n = \left(\frac{\log n}{n}\right)^{\frac{\alpha+\gamma}{2\alpha+2\gamma+1}}$.

Consider now the bias condition (2.5) and, using the notation of wavelets, take $k_n = 2^{J_n} \simeq n\varepsilon_n^2$. Let $\mathcal{B}_r = \mathrm{supp}(\Pi^{\delta,r})$ denote the $C^\delta([0,1])$-ball of radius r. Let r_n be an integer satisfying $(L_1 n\varepsilon_n^2)^{1/\nu} \leq r_n \leq (L_2 n\varepsilon_n^2)^{1/\nu}$ for some constants L_1, L_2 and take $\mathcal{P}_n = \mathcal{B}_{r_n}$. Then $\Pi(\mathcal{P}_n^c) = 1 - H(r_n) \lesssim e^{-Dr_n^\nu} \leq e^{-Ln\varepsilon_n^2}$, where L is a constant that can be made sufficiently large by increasing L_1. Now for all functions $f \in \mathcal{B}_r$, $\sup_k |\beta_{lk}(f)| \leq r 2^{-l(\delta+1/2)}$ for all $l \geq 0$. Consequently,

$$||K_{J_n}(f) - f||_2^2 = \sum_{l=J_n}^{\infty} \sum_{k=0}^{2^l-1} |\beta_{lk}(f)|^2 \leq \sum_{l=J_n}^{\infty} \sum_{k=0}^{2^l-1} r^2 2^{-l(2\delta+1)} \leq Cr^2 2^{-2\delta J_n},$$

so that for all $f \in \mathcal{P}_n$,

$$||K_{J_n}(f) - f||_2 \leq C'(n\varepsilon_n^2)^{1/\nu-\delta} \leq C''\xi_n = C'''\varepsilon_n(n\varepsilon_n^2)^\alpha,$$

which is verified with the choice $\varepsilon_n = n^{-\frac{\alpha+\delta-1/\nu}{2\alpha+2\delta-2/\nu+1}}$. Comparing this rate to the rates obtained when verifying (2.6) we obtain the minimal choices $\varepsilon_n = n^{-\frac{\alpha+\delta-1/\nu}{2\alpha+2\delta-2/\nu+1}}$ when $\delta < \gamma + \frac{1}{\nu}$ and $\varepsilon_n = (\log n/n)^{\frac{\alpha+\gamma}{2\alpha+2\gamma+1}}$ when $\delta = \gamma + \frac{1}{\nu}$. For the true function $f_0 \in C^\gamma([0,1])$, using a standard approximation bound gives $||K_{J_n}(f_0) - f_0||_2 \leq C(f_0)2^{-\gamma J_n} \simeq (n\varepsilon_n^2)^{-\gamma} = O(\xi_n)$ for all the above choices of ε_n. In both cases, apply Theorem 2.1 to obtain rate $\xi_n = \varepsilon_n(n\varepsilon_n^2)^\alpha$.

Consider now the stronger tail condition $1 - H(r) \lesssim \exp(-e^{Dr^\nu})$ as $r \to \infty$ for some $\nu > 0$. When $\delta \leq \gamma$, (2.6) is satisfied as above by the choice $\varepsilon \simeq (\frac{\log n}{n})^{\frac{\alpha+\delta}{\alpha+2\delta+1}}$. Letting r_n be an integer satisfying $(\log(L_1 n\varepsilon_n^2))^{1/\nu} \leq r_n \leq (\log(L_2 n\varepsilon_n^2))^{1/\nu}$ for some constants L_1, L_2 and taking $\mathcal{P}_n$ as above we obtain $\Pi(\mathcal{P}_n^c) \lesssim \exp(-e^{Dr_n^\nu}) \leq e^{-Ln\varepsilon_n^2}$ for some constant L that can be made arbitrarily large by increasing L_1. Using the above bias calculations, $||K_{J_n}(f) - f||_2 \leq Cr_n 2^{-\delta J_n} \leq C'r_n(n\varepsilon_n^2)^{-\delta}$ and so setting this equal to $\xi_n = n^\alpha \varepsilon_n^{2\alpha+1}$ yields that (2.5) is satisfied by the choice $\varepsilon_n = (\log n)^{\frac{1/v}{2\alpha+2\delta+1}} n^{-\frac{\alpha+\delta}{2\alpha+2\delta+1}}$. Substituting this expression into that of ξ_n gives the desired contraction rate. $\square$

Acknowledgements

References

[1] AGAPIOU, S., LARSSON, S. and STUART, A. M. (2013). Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Process. Appl.* **123** 3828–3860.

[2] ARBEL, J., GAYRAUD, G. and ROUSSEAU, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scand. J. Stat.* **40** 549–570.

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

[4] Borell, C. (1975). The Brunn-Minkowski inequality in Gauss space. *Invent. Math.* **30** 207–216. MR0399402

[5] Butucea, C. and Tsybakov, A. B. (2007). Sharp optimality in density deconvolution with dominating bias. II. *Teor. Veroyatn. Primen.* **52** 336–349. MR2742504

[6] Castillo, I., Kerkyacharian, G. and Picard, D. (2013). Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields.* To appear.

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

[8] Da Prato, G. (2006). *An introduction to infinite-dimensional analysis. Universitext.* Springer-Verlag, Berlin. Revised and extended from the 2001 original by Da Prato. MR2244975

[9] Dembo, A., Mayer-Wolf, E. and Zeitouni, O. (1995). Exact behavior of Gaussian seminorms. *Statist. Probab. Lett.* **23** 275–280. MR1340163

[10] Folland, G. B. (1999). *Real analysis*, second ed. *Pure and Applied Mathematics (New York).* John Wiley & Sons Inc., New York. Modern techniques and their applications, A Wiley-Interscience Publication. MR1681462

[11] Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[12] Ghosal, S. and van der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. MR2336864

[13] Giné, E. and Nickl, R. (2011). Rates on contraction for posterior distributions in L^r-metrics, $1 \le r \le \infty$. *Ann. Statist.* **39** 2883–2911. MR3012395

[14] Hoffmann-Jørgensen, J., Shepp, L. A. and Dudley, R. M. (1979). On the lower tail of Gaussian seminorms. *Ann. Probab.* **7** 319–342. MR525057

[15] Holmes, C. and Denison, G. (1999). Bayesian wavelet analysis with a model complexity prior. In *Bayesian Statistics 6: proceedings of the sixth Valencia international meeting* 769–776. Clarendon Press, Oxford.

[16] Huang, T.-M. (2004). Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.* **32** 1556–1593. MR2089134

[17] Johnstone, I. M., Kerkyacharian, G., Picard, D. and Raimondo, M. (2004). Wavelet deconvolution in a periodic setting. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 547–573. MR2088290

[18] Johnstone, I. M. and Silverman, B. W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18** 251–280. MR1041393

[19] Knapik, B. T., Szabó, B. T., van der Vaart, A. W. and van Zanten, J. H. (2012). Bayes procedures for adaptive inference in inverse problems for the white noise model. arXiv:1209.3628.

[20] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39** 2626–2657. MR2906881

[21] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2013). Bayesian recovery of the initial condition for the heat equation. *Comm. Statist. Theory Methods* **42** 1294–1313. MR3031282

[22] KUELBS, J. and LI, W. V. (1993). Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal.* **116** 133–157. MR1237989

[23] LEDOUX, M. (2001). *The concentration of measure phenomenon. Mathematical Surveys and Monographs* **89**. American Mathematical Society, Providence, RI. MR1849347

[24] LOUNICI, K. and NICKL, R. (2011). Global uniform risk bounds for wavelet deconvolution estimators. *Ann. Statist.* **39** 201–231. MR2797844

[25] MEYER, Y. (1992). *Wavelets and operators. Cambridge Studies in Advanced Mathematics* **37**. Cambridge University Press, Cambridge. Translated from the 1990 French original by D. H. Salinger. MR1228209

[26] NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804

[27] PENSKY, M. and VIDAKOVIC, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *Ann. Statist.* **27** 2033–2053. MR1765627

[28] RUDIN, W. (1991). *Functional analysis*, second ed. *International Series in Pure and Applied Mathematics*. McGraw-Hill Inc., New York. MR1157815

[29] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337

[30] SYTAYA, G. N. (1974). On some asymptotic representations of the Gaussian measure in a Hilbert space. *Theory of Stochastic Processes,* Ukrainian Academy of Sciences, **2** 93–104.

[31] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663

[32] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh. Inst. Math. Stat. Collect.* **3** 200–222. Inst. Math. Statist., Beachwood, OH. MR2459226

[33] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442

[34] ZHAO, L. H. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.* **28** 532–552. MR1790008