A general approach to posterior contraction in nonparametric inverse problems

Bartek Knapik, Jean-Bernard Salomond

Submitted to Bernoulli
arXiv: arXiv:1407.0335

A general approach to posterior contraction in nonparametric inverse problems

BARTEK KNAPIK[1,*] and JEAN-BERNARD SALOMOND[2,**]

[1]*Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands. E-mail:* [*]b.t.knapik@vu.nl

[2]*Université Paris-Est, Laboratoire d'Analyse et de Mathématiques Appliquées (UMR 8050), UPEM, UPEC, CNRS, F-94010, Créteil, France. E-mail:* [**]jean-bernard.salomond@u-pec.fr

In this paper we propose a general method to derive an upper bound for the contraction rate of the posterior distribution for nonparametric inverse problems. We present a general theorem that allows us to derive contraction rates for the parameter of interest from contraction rates of the related direct problem of estimating transformed parameter of interest. An interesting aspect of this approach is that it allows us to derive contraction rates for priors that are not related to the singular value decomposition of the operator. We apply our result to several examples of linear inverse problems, both in the white noise sequence model and the nonparametric regression model, using priors based on the singular value decomposition of the operator, location-mixture priors and splines prior, and recover minimax adaptive contraction rates.

Keywords: Bayesian nonparametrics, nonparametric inverse problems, posterior distribution, rate of contraction, modulus of continuity.

1. Introduction

Statistical approaches to inverse problems have been initiated in the 1960's and since then many estimation methods have been developed. Inverse problems arise naturally when one observes the object of interest only indirectly. Mathematically speaking, this phenomenon is easily modeled by the introduction of an operator K modifying the object of interest f, such that the observation at hand comes from the model

$$Y^n \sim P^n_{Kf}, \tag{1.1}$$

where f is assumed to belong to a parameter space $\mathcal{F}$, and $n \to \infty$ reflects the increasing amount of information in the observation. In many applications the operator K is assumed to be injective. However, in the most interesting cases its inverse is not continuous, thus the parameter of interest f cannot be reconstructed by a simple inversion of the operator. Such problems are said to be *ill-posed*. Several methods dealing with the discontinuity of the inverse operator have been proposed in the literature. The most famous one is to conduct the inference while imposing some regularity constraints on the parameter of interest f. These so-called regularization methods have been widely studied in the literature both from a theoretical and applied perspective, see [12, 18] for reviews.

A Bayesian approach to inverse problems is therefore particularly interesting, as it is well known that putting a prior distribution on the functional parameter yields a natural regularization.

This property of the Bayesian approach is particularly interesting for model choice, but it has proved also useful in many estimation procedures, as shown in [35] in the case of overfitted mixtures models, in [9] in the case of nonparametric models where regularization is necessary, in [37] in the semiparametric problem of estimating a monotone density at the boundaries of its support, or in [28] in the white noise setting.

In this paper we study the behaviour of the posterior distribution when the amount of information goes to infinity (e.g. when the number of data points n goes to infinity or when the level of the noise goes to 0) under the frequentist assumption that the data Y^n are generated from model (1.1) for some true unknown parameter f_0. Asymptotic properties of the posterior distribution in nonparametric models have been studied for many years. Some first results about consistency of Bayes procedures date back to Schwartz [38]. Her ideas were further refined and extended in an unpublished work of Barron [5], and can be also found in other works, e.g., [4], [22]. The next natural step is to consider the rate at which the neighborhoods of the truth can shrink, yet still capture most of the posterior mass. In other words, the interest lies in finding an upper bound for the rate at which the posterior concentrates around f_0. This is also the main focus of this paper. The aforementioned consistency results served as a starting point for two seminal papers on rates of convergence of posterior distributions by Ghosal et al. [23] and Shen and Wasserman [42]. Understanding of the whole posterior distribution is necessary for uncertainty quantification, see a recent paper by Szabó et al. [43] for an overview, but is also directly related to asymptotic properties of Bayes point estimators, see, e.g., Theorem 2.5 in [23]. In Bayesian nonparametrics it is important to understand the impact of the prior distribution on the posterior. In particular, some aspects of the prior may be inherited by the posterior when the amount of information grows to infinity and may thus be highly influential for the quality and speed of recovery.

Asymptotic properties of the Bayesian approach to nonparametric linear inverse problems have recently received a growing interest. Knapik et al. [28], Agapiou et al. [1], and Florens and Simoni [21] were the first to study posterior contraction rates under conjugate prior in the so-called mildly ill-posed setting (in the terminology of [12]). These were followed by two papers by Knapik et al. [29] and Agapiou et al. [2], studying Bayesian recovery of the initial condition for heat equation and related extremely and severely ill-posed inverse problems. One type of priors studied in [29] leads to a rate-adaptive Bayesian procedure. The paper by Ray [33] was the first study of the posterior contraction rates in the non-conjugate sequence setting. Considering non-conjugate prior is particularly interesting as it allows some additional flexibility of the model. However, the approach presented in [33] is only valid for priors that are closely linked to the *singular value decomposition* (SVD) of the operator. Moreover, in [33] several rate-adaptive priors were considered, both in the mildly and severely ill-posed setting. It should be noted, however, that some of the bounds on contraction rates in the severely ill-posed setting obtained in that paper are not optimal and do not agree with the bounds found in [29] or [2], probably due to proof techniques. Similar adaptive results, in the conjugate mildly ill-posed setting, using empirical and hierarchical Bayes approach were obtained in [27].

There is a rich literature on the problem of deriving posterior contraction rate in the direct problem setting, i.e. estimating Kf in (1.1). Since the seminal papers of Ghosal et al. [23] and Shen and Wasserman [42], general conditions on the prior distribution for which the posterior contracts at a certain rate have been derived in various cases. In particular, Ghosal and van der Vaart in [24] give a number of conditions for non independent and identically distributed

data. However, such results cannot be applied directly to ill-posed inverse problems, and to the authors' best knowledge, no analogous results exist in the inverse problem literature. In this paper we propose a unified general approach to posterior contraction in nonparametric inverse problems, and illustrate it for specific linear inverse problems.

To understand why the existing general posterior contraction results are not suited for nonparametric inverse problems consider an abstract setting in which the parameter space $\mathcal{F}$ is an arbitrary metrizable topological vector space and let K be a continuous injective mapping $K : \mathcal{F} \ni f \mapsto Kf \in K\mathcal{F}$. Let d and d_K denote some metrics or semi-metrics on $\mathcal{F}$ and $K\mathcal{F}$, respectively. Any prior Π on f imposes a prior on Kf through the continuous mapping K. Recall that the true parameter of interest f_0 belongs to $\mathcal{F}$. General posterior contraction results (e.g., in [23] or [24]) rely on several natural metrics related to the model (1.1) and therefore control the distance between Kf_0 and Kf in the d_K metric. On the other hand, our interest lies in the recovery of f_0, and therefore the control of the distance between f_0 and f in the d metric is desirable. Since the operator K does not have a continuous inverse and the problem is ill-posed, even if $d_K(Kf, Kf_0)$ is small, the distance $d(f, f_0)$ between f and the true f_0 can be arbitrarily large. In other words, there is no equivalence between the metrics d and d_K and therefore the existing theory of posterior contraction does not allow obtaining bounds on posterior contraction rates for the recovery of f_0.

Even if the problem is ill-posed, there exist subsets $\mathcal{S}_n$ of $\mathcal{F}$ such that the inverse of the operator K restricted to $K\mathcal{S}_n$ is continuous. We can thus easily derive posterior contraction rate for $f \in \mathcal{S}_n$ from posterior contraction rate for Kf by inverting the operator K. For suitably chosen priors, the sets $\mathcal{S}_n$ will capture most of the posterior mass, and we can thus extend the contraction result to the whole parameter space $\mathcal{F}$. The sets $\mathcal{S}_n$, thought of as sieves approximating the parameter space $\mathcal{F}$, have already been considered in [23] allowing some additional flexibility and are often incorporated in results on posterior contraction for various models. However, their principal role was not to enable the change of metrics, but rather alleviate the usual entropy condition. In our approach we first assume the existence of a contraction result for the so-called direct problem (that is the recovery of Kf) that can be derived using general posterior contraction literature. Next, we choose a sequence of subsets $\mathcal{S}_n$ in such a way that the inversion of the operator K on $K\mathcal{S}_n$, so also the change of metrics, can be controlled and at the same this sets are big enough (in terms of the posterior mass). The latter condition can be verified by imposing additional sufficient conditions on the prior (on f). We are then able to show that the posterior distribution for the parameter of interest f contracts at a given rate.

The rest of the paper is organized as follows: we present the main result in Section 2 and discuss how it relates to other results using the concept of sieves to control contraction in other metrics. We then apply our result in various settings. We first consider the white noise sequence model in Section 3, where we present a general construction of the sets $\mathcal{S}_n$, and recover many of the existing results with much less effort. We also observe an interesting interplay between optimality of Bayesian procedures for estimating f and Kf. In Section 4 we apply our method in the nonparametric inverse regression setting, considering new families of priors that need not be related to the SVD, and leading to optimal Bayesian procedures. Proofs of Sections 2–4 are placed in Section 5. We conclude the paper with a discussion in Section 6.

For two sequences (a_n) and (b_n) of numbers, $a_n \asymp b_n$ means that $|a_n/b_n|$ is bounded away from zero and infinity as $n \to \infty$, $a_n \lesssim b_n$ means that a_n/b_n is bounded, $a_n \sim b_n$ means that

$a_n/b_n \to 1$ as $n \to \infty$, and $a_n \ll b_n$ means that $a_n/b_n \to 0$ as $n \to \infty$. For two real numbers a and b, we denote by $a \vee b$ their maximum, and by $a \wedge b$ their minimum. For a sequence of random variables $X^n = (X_1, \dots, X_n) \sim P_f^n$ and any measurable function ψ with respect to P_f^n, we denote by $\mathrm{E}_f \psi$ the expectation of $\psi(X^n)$ with respect to P_f^n and when $f = f_0$ we will write E_0 instead of E_{f_0}.

2. General theorem

Assume that the observations Y^n come from model (1.1) and that P_{Kf}^n admit densities p_{Kf}^n relative to a σ-finite measure μ^n. To avoid complicated notations, we drop the superscript n in the rest of the paper. Let $\mathcal{F}$ and $K\mathcal{F}$ be metric spaces, and let d and d_K denote metrics on both spaces, respectively.

In this section we present the main result of this paper which gives an upper bound on the posterior contraction rate under some general conditions on the prior. We call the estimation of Kf given the observations Y the *direct problem*, and the estimation f given Y the *inverse problem*. The main idea is to control the change of metrics d_K and d. If the posterior distribution concentrates around Kf_0 for the metric d_K at a certain rate in the direct problem, applying the change of metrics will give us an upper bound on the posterior contraction rate for the metric d in the inverse problem. However, since the operator K does not posses a continuous inverse, the change of metrics cannot be controlled over the whole space $K\mathcal{F}$. A way to circumvent this issue is to only focus on a sequence of sets of high posterior mass for which the change of metric is feasible. More precisely, for a set $\mathcal{S} \subset \mathcal{F}$, $f_0 \in \mathcal{F}$ and a fixed $\delta > 0$ we call the quantity

$$\omega(\mathcal{S}, f_0, d, d_K, \delta) := \sup\big\{ d(f, f_0) : f \in \mathcal{S}, d_K(Kf, Kf_0) \leq \delta \big\} \tag{2.1}$$

the *modulus of continuity*. We note that in this definition we do not assume $f_0 \in \mathcal{S}$. This is thus a local version of the modulus of continuity considered in [17] or [26]. On the one hand, the sets $\mathcal{S}_n$ need to be big enough to capture most of the posterior mass. On the other hand, one has to be able to control the distance between the elements of $\mathcal{S}_n$ and f_0, given the distance between Kf and Kf_0 is small. Since the operator K is unbounded, this suggests that the sets $\mathcal{S}_n$ cannot be too big.

Theorem 2.1. *Let $\epsilon_n \to 0$ and let Π the prior distribution on f be such that*

$$\mathrm{E}_0 \Pi\big(\mathcal{S}_n^c \,|\, Y^n\big) \to 0, \tag{2.2}$$

for some sequence of sets $(\mathcal{S}_n)$, $\mathcal{S}_n \subset \mathcal{F}$, and for any positive sequence M_n

$$\mathrm{E}_0 \Pi\big(f : d_K(Kf, Kf_0) \geq M_n \epsilon_n \,|\, Y^n\big) \to 0. \tag{2.3}$$

Then

$$\mathrm{E}_0 \Pi\big(f : d(f, f_0) \geq \omega(\mathcal{S}_n, f_0, d, d_K, M_n \epsilon_n) \,|\, Y^n\big) \to 0.$$

The proof is elementary and can be found in Section 5.1.

The idea behind Theorem 2.1 is simple and was used to change metrics also in direct problems. For instance Castillo and van der Vaart [11] considered the multivariate normal mean model in the situation that the mean vector is sparse. They use the fact that the posterior concentrates along certain subspaces on which it is easy to control an ℓ_q-like metric with the standard Euclidean metric for $q < 2$. Hoffmann et al. [26] also use concentration of the posterior on specific sets to control the L_∞ metric with the L_2 metric in the white noise model.

Castillo et al. [10] extended the ideas of [11] to the sparse linear regression model, in which the recovery of the parameter of the model is an inverse problem. Similar reasoning was also used in [44] to study posterior contraction in the special case of Gaussian elliptic inverse problem, and in [16] to investigate asymptotic properties of empirical Bayes procedures for density deconvolution. However, these papers consider specific inverse problems only, whereas Theorem 2.1 allows deriving contraction rates for a wide variety of inverse problem models for which the prior is not necessarily related to the spectral decomposition of the operator K, e.g., when the operator does not admit singular value decomposition, as in Section 4.1.

The interpretation of the theorem is the following: given a properly chosen sequence of sets $\mathcal{S}_n$, the rate of posterior contraction $M_n\epsilon_n$ in the direct problem restricted to the given sequence can be translated to the rate of posterior contraction in the inverse setting. Note that the sequence M_n is often chosen to grow to infinity as slowly as needed (see, e.g., in [23] or [24]), making ϵ_n the effective rate of posterior contraction. Also, in both contraction results of Section 4, the sequence M_n need not be diverging and is chosen to be constant. Since the operator K is injective and continuous, any prior Π on f induces a prior on Kf, and the general posterior contraction results can be applied to obtain the rate of contraction in the direct problem of estimating Kf.

Next, the choice of $\mathcal{S}_n$ is crucial as it is the principal component in the control of the change of metric. In particular, the contraction rate $M_n\epsilon_n$ for the direct problem may not be optimal, and still lead to an optimal contraction rate $\omega(\mathcal{S}_n, f_0, d, d_K, M_n\epsilon_n)$ for the inverse problem with a well-suited choice of $\mathcal{S}_n$. As shown in Section 3.3, it is possible in some cases to obtain optimal recovery of f without having optimal recovery of Kf. In this example, we can choose S_n small enough so that the change of metrics can be control very precisely. This widens the possible choice of priors leading to optimal contraction rates and shows that the change of metric is the crucial part here. However, in most cases, the priors considered in this paper lead to optimal recovery for both f and for Kf.

To control the posterior mass of the sets $\mathcal{S}_n$ we can usually alter the proofs of contraction results for the direct problems. Here we present a standard argument leading to (2.2). Define the usual Kullback–Leibler neighborhoods by

$$B_n(Kf_0, \epsilon) = \Big\{ f \in \mathcal{F} : -\int p_{Kf_0} \log \frac{p_{Kf}}{p_{Kf_0}}\, d\mu \le n\epsilon^2, \\ \int p_{Kf_0} \Big(\log \frac{p_{Kf}}{p_{Kf_0}} \Big)^2 d\mu \le n\epsilon^2, \Big\}, \tag{2.4}$$

The following lemma adapted from [24] gives general conditions on the prior such that (2.2) is satisfied.

Lemma 2.1 (Lemma 1 in [24]). *Let $\epsilon_n \to 0$ and let $(\mathcal{S}_n)$ be a sequence of sets $\mathcal{S}_n \subset \mathcal{F}$. If Π is*

the prior distribution on f satisfying

$$\frac{\Pi(\mathcal{S}_n^c)}{\Pi(B_n(Kf_0, \epsilon_n))} \lesssim \exp(-2n\epsilon_n^2),$$

then

$$\mathrm{E}_0\Pi\big(\mathcal{S}_n^c \,|\, Y^n\big) \to 0.$$

For clarity of presentation the results in this section are stated for a fixed f_0, but we note that they are easily extended to uniform results over certain sets, i.e., balls of fixed radius and regularity, or union of balls of fixed radii over compact range of regularity parameter (see results of Section 3).

3. Sequence white noise model

Our first examples are based on the well-studied infinite-dimensional normal mean model. In the Bayesian context the problem of direct estimation of infinitely many means has been studied, among others, in [7, 24, 42, 45].

We consider the white noise setting, where we observe an infinite sequence $Y^n = (Y_1, Y_2, \ldots)$ satisfying

$$Y_i = \kappa_i f_i + \frac{1}{\sqrt{n}} Z_i, \tag{3.1}$$

where $Z_1, Z_2, \ldots$ are independent standard normal random variables, $f = (f_1, f_2, \ldots) \in \ell_2$ is the infinite-dimensional parameter of interest and (κ_i) is a known sequence that may converge to 0 as $i \to \infty$. If this is the case (so when the operator K does not possess a continuous inverse) the modulus of continuity defined in (2.1) is infinite when $\mathcal{S} = \mathcal{F}$.

Even though this model is rather abstract, it is mathematically tractable and it enables rigorous results and proofs. Moreover, it can be seen as an idealized version of other statistical models through equivalence results see, e.g., [8, 31, 32]. Both white noise examples of inverse problems presented in this section have already been studied in the Bayesian literature. We present them here for several reasons. First, the direct version of the normal mean model attracted a lot of attention in the Bayesian literature, e.g. providing contraction results for estimation of Kf in the mildly ill-posed setting. Therefore, we choose this example to illustrate how Theorem 2.1 works in practice. In particular, it allows us to make it clear how one could construct a sequence of sets $\mathcal{S}_n$. In the severely ill-posed case we study truncated (or sieve) priors leading to optimal recovery of the parameter of interest. Our results improve the findings of [29] and [2]. In addition, we can show that optimal contraction for f does not necessarily require optimal recovery of Kf.

3.1. Computation of a modulus

In this section we first present an example of the sequence of sets $\mathcal{S}_n$, and later present how the modulus of continuity for this sequence can be computed in a standard inverse problem

setting. We now suppose that $\mathcal{F}$ and $K\mathcal{F}$ are separable Hilbert spaces, denoted $(\mathbb{H}_1, \|\cdot\|_{\mathbb{H}_1})$ and $(\mathbb{H}_2, \|\cdot\|_{\mathbb{H}_2})$ respectively. We note that the sets $\mathcal{S}_n$ resemble the sets $\mathcal{P}_n$ considered in [33].

As already noted, the operator K restricted to certain subsets of the domain $\mathbb{H}_1$ might have a finite modulus of continuity defined in (2.1). Clearly, one wants to construct a sequence of sets $\mathcal{S}_n$ that in a certain sense approaches the full domain $\mathbb{H}_1$. This is understood in terms of the remaining prior mass condition in Theorem 2.1. Moreover, since we do not require f_0 to be in $\mathcal{S}_n$, we need to be able to control the distance between f_0 and $\mathcal{S}_n$.

A natural guess is to consider finite-dimensional projections of $\mathbb{H}_1$. In this section we go beyond this concept. To get some intuition, consider the Fourier basis of $\mathbb{H}_1$. The ill-posedness can be then viewed as too big an amplification of the high frequencies through the inverse of the operator K. Therefore, one wants to control the higher frequencies in the signal, and thus in the parameter f.

Since $\mathbb{H}_1$ is a separable Hilbert space, there exist an orthonormal basis (e_i) and each element $f \in \mathbb{H}_1$ can be viewed as an element of ℓ_2 and

$$\|f\|_{\mathbb{H}_1} = \sum_{i=1}^{\infty} f_i^2.$$

For given sequences of positive numbers $k_n \to \infty$ and $\rho_n \to 0$, and a constant $c \geq 0$ we define

$$\mathcal{S}_n := \Big\{ f \in \ell_2 : \sum_{i>k_n} f_i^2 \leq c\rho_n^2 \Big\}. \tag{3.2}$$

If the operator K is compact, then the spectral decomposition of the self-adjoint operator $K^TK : \mathbb{H}_1 \to \mathbb{H}_1$ provides a convenient orthonormal basis. In the compact case the operator K^TK possesses countably many positive eigenvalues κ_i^2 and there is a corresponding orthonormal basis (e_i) of $\mathbb{H}_1$ of eigenfunctions, and the sequence $(\tilde{e}_i)$ defined by $Ke_i = \kappa_i\tilde{e}_i$ forms an orthonormal conjugate basis of the range of K in $\mathbb{H}_2$. Therefore, both f and Kf can be associated with sequences in ℓ_2. Since the problem is ill-posed when $\kappa_i \to 0$, we can assume without loss of generality that the sequence κ_i is decreasing.

Let k_n, ρ_n, and c in the definition of $\mathcal{S}_n$ be fixed. Then for any $g \in \mathcal{S}_n$

$$\begin{aligned}
\|g\|_{\mathbb{H}_1}^2 &= \sum_{i=1}^{\infty} g_i^2 = \sum_{i\leq k_n} g_i^2 + \sum_{i>k_n} g_i^2 \\
&\leq \sum_{i\leq k_n} g_i^2 + c\rho_n^2 = \sum_{i\leq k_n} \kappa_i^{-2}\kappa_i^2 g_i^2 + c\rho_n^2 \\
&\leq \kappa_{k_n}^{-2} \sum_{i\leq k_n} \kappa_i^2 g_i^2 + c\rho_n^2 \leq \kappa_{k_n}^{-2}\|Kg\|_{\mathbb{H}_2}^2 + c\rho_n^2.
\end{aligned}$$

Let f_n be the projection of f_0 on the first k_n coordinates, i.e., $f_{n,i} = f_{0,i}$ for $i \leq k_n$ and 0 otherwise. Moreover, we assume that f_0 belongs to some smoothness class described by a decreasing sequence (s_i):

$$\|f_0\|_s^2 = \sum_{i=1}^{\infty} s_i^{-2} f_{0,i}^2 < \infty.$$

For instance, the usual Sobolev space of regularity β is defined in that way with $s_i = i^{-\beta}$. Therefore, we have

$$\|f_n - f_0\|_{\mathbb{H}_1} \leq s_{k_n}\|f_0\|_s, \qquad \|Kf_n - Kf_0\|_{\mathbb{H}_2} \leq s_{k_n}\kappa_{k_n}\|f_0\|_s.$$

Using the triangle inequality twice and keeping in mind that $f - f_n \in \mathcal{S}_n$ we obtain

$$\begin{aligned}\|f - f_0\|_{\mathbb{H}_1} &\leq \|f - f_n\|_{\mathbb{H}_1} + \|f_n - f_0\|_{\mathbb{H}_1} \\ &\leq \kappa_{k_n}^{-1}\|Kf - Kf_n\|_{\mathbb{H}_2} + \sqrt{c}\rho_n + s_{k_n}\|f_0\|_s \\ &\leq \kappa_{k_n}^{-1}\big(\|Kf - Kf_0\|_{\mathbb{H}_2} + \kappa_{k_n}s_{k_n}\|f_0\|_s\big) + \sqrt{c}\rho_n + s_{k_n}\|f_0\|_s \\ &= \kappa_{k_n}^{-1}\|Kf - Kf_0\|_{\mathbb{H}_2} + \sqrt{c}\rho_n + 2\|f_0\|_s s_{k_n}. \end{aligned} \tag{3.3}$$

We then find an upper bound for the modulus of continuity with this specific choice of $\mathcal{S}_n$ is

$$\omega(\mathcal{S}_n, f_0, \|\cdot\|_{\mathbb{H}_1}, \|\cdot\|_{\mathbb{H}_2}, \delta) \lesssim \kappa_{k_n}^{-1}\delta + \rho_n + s_{k_n}. \tag{3.4}$$

3.2. Mildly ill-posed problems

In this section we consider the model (3.1), where $C^{-1}i^{-p} \leq \kappa_i \leq Ci^{-p}$ for some $p \geq 0$ and $C \geq 1$. Since the κ_i's decay polynomially, the operator is *mildly* ill-posed. Such problems are well studied in the frequentist literature, and we refer the reader to [12] for a comprehensive overview. There are also several papers on properties of Bayes procedures for such problems. The first studies of posterior contraction in mildly ill-posed operators were obtained in [28] and [1]. Later, adaptive priors leading to the optimal minimax rate of contraction (up to slowly varying factors) were studied in [33] and [27]. Similar problem, with a different noise structure, has been studied in [21]. The main purpose of this section is to show how Theorem 2.1 can be applied to such problems and how existing results on contraction rates for Kf in the sequence setting can be used to obtain posterior contraction rates for f without explicit computations as in aforementioned papers.

We put a product prior on f of the form

$$\Pi = \bigotimes_{i=1}^{\infty} N(0, \lambda_i),$$

where $\lambda_i = i^{-1-2\alpha}$, for some $\alpha > 0$. Furthermore, the true parameter f_0 is assumed to belong to S^β for some $\beta > 0$:

$$S^\beta = \Big\{f \in \ell_2 : \|f\|_\beta^2 := \sum f_i^2 i^{2\beta} < \infty\Big\}. \tag{3.5}$$

Therefore, $\|Kf_0\|_{\beta+p}^2$ is finite, the prior on f induces the prior on Kf such that $(Kf)_i \sim N(0, \lambda_i\kappa_i^2)$, and one can deduce from the results of [45] and [7] that

$$\sup_{\|Kf_0\|_{\beta+p} \leq R} \mathrm{E}_0\Pi\big(f : \|Kf - Kf_0\| \geq M_n n^{-\frac{(\alpha\wedge\beta)+p}{1+2\alpha+2p}} \,\big|\, Y^n\big) \to 0.$$

In order to apply Theorem 2.1 we need to construct the sequence of sets $\mathcal{S}_n$ and verify condition (2.2). We use the construction as in (3.2), and we verify the remaining posterior mass condition along the lines of Lemma 2.1.

Theorem 3.1. *Suppose the true f_0 belongs to S^β for $\beta > 0$. Then for every $R > 0$ and $M_n \to \infty$*

$$\sup_{\|f_0\|_\beta \le R} \mathrm{E}_0 \Pi\big(f : \|f - f_0\| \ge M_n n^{-\frac{(\alpha\wedge\beta)}{1+2\alpha+2p}} \,\big|\, Y^n\big) \to 0.$$

The proof of this theorem is postponed to Section 5.2.1.

The upper bound on the posterior contraction rate obtained in this theorem agrees with the ones already obtained in the existing literature (see, for instance, [27, 28, 33]). We note that the prior used above requires the knowledge of the true regularity parameter β in order to achieve minimax optimal rate of recovery. Moreover, we note that the prior with $\alpha = \beta$ leads to optimal recovery of both f and Kf.

The prior used in this section is rather simple and is not hierarchical, i.e., is not aimed at adaptive recovery. We have already pointed out that [33] and [27] studied adaptive Bayesian approach to mildly ill-posed inverse problems and obtained optimal rates (up to logarithmic factors). We would also like to point out that recent studies of adaptive approaches to the sequence white noise model [e.g. 3, 27] already consider its inverse version (i.e., allowing $\kappa_i \neq 1$). In a recent work Belitser [6] even obtained adaptive posterior contraction rate in a setting equivalent to the one considered here that could be used both for the estimation of Kf and the estimation of f. Therefore, even though one could consider the existing approaches studied in the literature to achieve adaptation (by first showing optimal contraction for Kf and then applying Theorem 2.1 to prove contraction for f), this will not be treated here for the sake of simplicity (in the latter cases also to avoid rather artificial application of Theorem 2.1).

3.3. Severely and extremely ill-posed problems

We again consider the sequence white noise setting, where we observe an infinite sequence $Y^n = (Y_1, Y_2, \ldots)$ as in (3.1) where $\kappa_i \asymp \exp(-\gamma i^p)$ for some $p \ge 1$ and $\gamma > 0$. We first consider estimation of Kf_0 that will be later used to obtain the rate of contraction of the posterior around f_0. We put a product prior on f of the form

$$\Pi = \bigotimes_{i=1}^{k_n} N(0, \lambda_i),$$

where $\lambda_i = i^{-\alpha}\exp(-\xi i^p)$, for $\alpha \ge 0$, $\xi > 0$, and some $k_n \to \infty$. We choose k_n solving $1 = n\lambda_i \exp(-2\gamma i^p) = n i^{-\alpha}\exp(-(\xi + 2\gamma)i^p)$. Using the Lambert function W one can show that

$$k_n = \Big(\frac{\alpha}{p(\xi+2\gamma)} W\Big(n^{\frac{p}{\alpha}} \frac{p(\xi+2\gamma)}{\alpha}\Big)\Big)^{1/p} = \Big(\frac{\log n}{\xi+2\gamma} + O(\log\log n)\Big)^{1/p}, \qquad (3.6)$$

see also Lemma A.4. in [29]. Note that in this case we have $\exp(k_n^p) = (nk_n^{-\alpha})^{1/(\xi+2\gamma)}$, so we can avoid exponentiating k_n. Therefore, we do not have to specify the constant in front of the $\log\log n$ term in the definition of k_n, and we may assume that k_n is of the order $(\log n)^{1/p}$.

Note that the hyperparameters of the prior do not depend on f_0, but only on K, which is known. For $\mathcal{S}_n$ as in (3.2) with k_n as above and $c = 0$, the prior is supported on $\mathcal{S}_n$ and the first condition of Theorem 2.1 is trivially satisfied. Regardless of the choice of ξ and α (as long as $\alpha \geq 0$ and $\xi > 0$) the following theorem shows that the posterior contracts at the optimal minimax rate $(\log n)^{-\beta/p}$ for the inverse problem of estimating f_0 (cf. [29] or [2] and references therein), so the prior is rate-adaptive.

In this section we consider deterministically truncated Gaussian priors. Similar priors in the extremely ill-posed setting are considered in [33], but in this paper the truncation level is endowed with a hyper-prior and the bound on the posterior contraction is suboptimal. Other papers on Bayesian approach to severely and extremely ill-posed inverse problems do not consider truncated priors. In [29] the optimal rate is achieved for the priors with exponentially decaying or polynomially decaying variances (in the latter case the speed of decay leading to optimal rate is closely related to the regularity of the truth). Similar results for the priors with polynomially decaying variances are presented in [33] and [2]. However, in the former case the rate for undersmoothing priors is worse than the rate obtained in the other papers.

Theorem 3.2. *Suppose the true f_0 belongs to S^β for $\beta > 0$. Then for every $R > 0$ and $M_n \to \infty$*

$$\sup_{\|f_0\|_\beta \leq R} \mathrm{E}_0 \Pi\big(f : \|f - f_0\| \geq M_n (\log n)^{-\frac{\beta}{p}} \;\big|\; Y^n\big) \to 0.$$

The proof of this Theorem is postponed to Section 5.2.2. The prior considered in this theorem might seem unnatural, since λ_i's do not coincide with the type of regularity of the truth and the prior puts mass only on analytic functions of growing complexity. However, similar approaches are quite common in the Bayesian literature, for instance when finite mixtures models are considered. Moreover, this prior has also some computational advantages, since the corresponding posterior can be handled numerically.

Inspection of the proof shows that the deterministic truncation is suboptimal for the estimation of Kf_0, since the resulting upper bound is polynomially slower than the minimax rate $n^{-1/2}(\log n)^{1/2p}$. It sheds light on an interesting, although counterintuitive property of the Bayesian approach to inverse problems: one may not need optimal contraction for the estimation of Kf_0 to get optimal contraction for the estimation of f_0. This phenomenon should be interpreted in the following way: since the operator K regularizes the parameter f_0, one could compensate the suboptimal contraction of the posterior for the direct problem, by a sharper control of the deviation between f and f_0 in (3.3) when f is in $\mathcal{S}_n$. Indeed, when ξ increases (which slows down the upper bound on the posterior contraction for Kf_0), the truncation level k_n decreases. As a result, the sets $\mathcal{S}_n$ become smaller, so the sharper control of $d(f, f_0)$ is indeed possible. In the specific setting of sequence white noise model it might seem artificial. However, this observation could prove useful in more complex settings, especially because it widens the class of possible prior distributions giving optimal contraction rates.

***Remark* 3.1.** If an upper bound $\widetilde{\beta}$ on the regularity of the true f_0 is known, one can also take $\xi = 0$ and $\alpha \geq 1 + 2\widetilde{\beta}$ and the assertion of Theorem 3.2 stays valid. In this case the upper bound on the posterior contraction rate for Kf_0 is logarithmically slower than the minimax rate.

4. Regression

We now consider the inverse regression model with Gaussian residuals

$$Y_i = Kf(x_i) + \sigma\epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \quad i = 1, \ldots, n, \tag{4.1}$$

where the covariates x_i are fixed in a covariate space $\mathcal{X}$. In the sequel, we either choose $\mathcal{X} = [0,1]$ or $\mathcal{X} = \mathbb{R}$. In the following we consider the noise level $\sigma > 0$ to be known although one could also think of putting a prior on it and estimate it in the direct model. Nonparametric regression models have been studied in the literature for direct problems, and frequentist properties of the posterior distribution are well known for a wide variety of priors. In [24], Ghosal and van der Vaart give general conditions on the prior such that the posterior contracts at a given rate. Nonparametric inverse regression models are also used in practice, for instance in econometrics where one considers instrumental variable as in [20]. However, to the authors' best knowledge, contraction rates for these models have only been considered in [44].

In this setting, a common choice for the metrics d and d_K are the usual l_2 norms

$$d(f,g)^2 = n^{-1}\sum_{i=1}^{n}(f(x_i) - g(x_i))^2 = \|f - g\|_n^2, \quad d_K(f,g) = d(Kf, Kg).$$

For $a \in \mathbb{R}^k$, $k \in \mathbb{N}^*$, and $f \in L_2$, we denote the standard Euclidean and L_2 norms by

$$\|a\|_k = \Big(\sum_{i=1}^{k} a_i^2\Big)^{1/2}, \qquad \|f\| = \Big(\int f^2\Big)^{1/2},$$

respectively.

We now consider two examples of inverse regression problems, namely *numerical differentiation* and *deconvolution on* $\mathbb{R}$. For these sampling models, we study the frequentist properties of the posterior distribution for standard prior that have not been considered for inverse regression problems so far.

4.1. Numerical differentiation using spline prior

In this section, we consider the inverse regression problem (4.1) with the operator K between $L_1[0,1]$ and the space of functions differentiable almost everywhere on the interval $[0,1]$ (see also Chapter 7 of [36]) defined by

$$Kf(x) = \int_0^x f(t)dt, \qquad \text{for } x \in [0,1]. \tag{4.2}$$

We note that the operator K is not defined between two Hilbert spaces, hence goes beyond the concept of singular value decomposition. This model is particularly useful for numerical differentiation, for instance, and has been well studied in the literature. In particular, in [12] a related problem of estimating a derivative of a square integrable function is presented and it is shown that the SVD basis is the Fourier basis. Moreover, the operator is mildly ill-posed of degree 1 (cf. Section 3). We consider a prior on f that is well-suited if the true regression function f_0 belongs to the Hölder space $\mathcal{H}(\beta, L)$ for some $\beta > 0$, that is f_0 is $\beta_0 = \lfloor \beta \rfloor$ times differentiable and

$$\|f_0\|_\beta = \sup_{x \neq y} \frac{|f^{(\beta_0)}(x) - f^{(\beta_0)}(y)|}{|x-y|^{\beta-\beta_0}} \leq L.$$

Since Kf_0 is $(\beta_0 + 1)$ times differentiable, it also holds that $f_0 \in \mathcal{H}(\beta, L)$ implies then $Kf_0 \in \mathcal{H}(\beta + 1, L)$.

We construct a prior on f by considering its decomposition in a B-spline basis. A definition of the B-spline basis can be found in [13]. For a fixed positive integer $q > 1$ called the degree of the basis, and a given partition of $[0, 1]$ in m subintervals of the form $((i-1)/m, i/m]$, the space of splines is a collection of function $f(0, 1] \to \mathbb{R}$ that are $q - 2$ times differentiable and if restricted to one of the sets $((i-1)/m, i/m]$, are polynomial of degree at most q. An interesting feature of the space of splines is that it forms a J-dimensional linear space with the so called B-spline basis denoted $(B_{1,q}, \dots, B_{J,q})$, for $J = m + q - 1$. Priors based on the decomposition of the function f in the B-spline basis of order q have been considered in the regression setting in, e.g., [24] and [40], and are commonly used in practice. Here we construct a different version of the prior that will prove to be useful to derive contraction rate for the direct problem and the inverse problem.

Let the prior distribution on f be defined as follows:

$$\Pi := \begin{cases} J \sim \Pi_J \\ a_1, \dots a_J \overset{iid}{\sim} \Pi_{a,J} \\ f(x) = J \sum_{j=1}^{J-1} (a_{j+1} - a_j) B_{j,q-1}(x). \end{cases} \tag{4.3}$$

Given the definition of $B_{j,q}$ in [13], standard computations give

$$B'_{j,q}(x) = J\big(B_{j,q-1}(x) - B_{j+1,q-1}(x)\big)$$

which in turn gives

$$Kf(x) = \sum_{j=1}^{J} a_j B_{j,q}(x). \tag{4.4}$$

This explains why we choose a prior as in (4.3) since it leads to the usual spline prior on Kf. Note that the condition that $Kf(0) = 0$ can be imposed by a specific choice of nodes for the B-spline basis (see [13]). To compute the modulus of continuity for this model, we need to impose some conditions on the design. Let Σ_n^q be a matrix defined by its elements

$$(\Sigma_n^q)_{i,j} = \frac{1}{n} \sum_{l=1}^{n} B_{i,q}(x_l) B_{j,q}(x_l), \quad i, j = 1, \dots, J.$$

Similarly to [24], we ask that the design points satisfy the following conditions:

D1 for all $\mathbf{v}_1 \in \mathbb{R}^J$

$$J^{-1}\|\mathbf{v}_1\|_J^2 \asymp \mathbf{v}_1'\Sigma_n^q\mathbf{v}_1$$

D2 for all $\mathbf{v}_2 \in \mathbb{R}^{J-1}$

$$(J-1)^{-1}\|\mathbf{v}_2\|_{J-1}^2 \asymp \mathbf{v}_2'\Sigma_n^{(q-1)}\mathbf{v}_2.$$

Condition **D1** is natural when considering B-splines priors in a regression setting, and both conditions are satisfied for a wide variety of designs. Consider for instance the uniform design $x_i = i/n$ for $i = 1, \dots, n$. Then given Lemma 4.2 in [23], we get that for $\mathbf{v}_1 \in \mathbb{R}^J$, $\mathbf{v}_2 \in \mathbb{R}^{J-1}$

$$\|\mathbf{v}_1\|_J^2 J^{-1} \lesssim \Big\|\sum_{j=1}^{J}\mathbf{v}_{1,j}B_{j,q}\Big\|^2 \lesssim \|\mathbf{v}_1\|_J^2 J^{-1},$$

$$\|\mathbf{v}_2\|_{J-1}^2 (J-1)^{-1} \lesssim \Big\|\sum_{j=1}^{J-1}\mathbf{v}_{2,j}B_{j,q-1}\Big\|^2 \lesssim \|\mathbf{v}_2\|_{J-1}^2 (J-1)^{-1},$$

and the constants depend only on q. Furthermore, we have that

$$\Big\|\sum_{j=1}^{J}\mathbf{v}_{1,j}B_{j,q}\Big\|^2 = \mathbf{v}_1'\Sigma_n^q\mathbf{v}_1 + O\big(n^{-1}\big),$$

where the remainder depends only on q. Similarly,

$$\Big\|\sum_{j=1}^{J-1}\mathbf{v}_{2,j}B_{j,q-1}\Big\|^2 = \mathbf{v}_2'\Sigma_n^{q-1}\mathbf{v}_2 + O\big(n^{-1}\big).$$

Thus **D1** and **D2** are satisfied for the uniform design for all $J = o(n)$.

We now go on and derive conditions on the prior such that the posterior contracts at the minimax adaptive rate (up to a $\log n$ factor). The prior we consider is not conjugate, and does not depend on the singular value decomposition of the operator K for obvious reasons.

Theorem 4.1. *Let $Y^n = (Y_1, \dots, Y_n)$ be a sample from (4.1) with $\mathcal{X} = [0,1]$ and Π be a prior for f as in (4.3). Suppose that Π_J is such that for some constants $c_d, c_u > 0$ and $t \geq 0$,*

$$\exp(-c_d j(\log j)^t) \leq \Pi_J(j \leq J \leq 2j), \quad \Pi_J(J > j) \lesssim \exp(-c_u j(\log j)^t), \tag{4.5}$$

for all $J > 1$, and suppose that $\Pi_{a,J}$ is such that for all $a_0 \in \mathbb{R}^J$, $\|a_0\|_\infty \leq H$, there exists a constant c_2 depending only on H such that

$$\Pi_{a,J}(\|a - a_0\|_J \leq \epsilon) \geq \exp(-c_2 J\log(1/\epsilon)) \tag{4.6}$$

*Let $\Theta(\beta, L, H) = \{f \in \mathcal{H}(\beta, L), \|f\|_\infty \leq H\}$. If the design $(x_1, \dots, x_n)$ satisfies conditions **D1** and **D2**, then for all L and for all $\beta \leq q$ there exits a constant $C > 0$ that depends only on q, L, H and Π such that if $f_0 \in \mathcal{H}(\beta, L)$, then*

$$\sup_{\beta \leq q-1}\sup_{f_0 \in \Theta(\beta,L,H)} \mathrm{E}_0\Pi\big(\|f - f_0\|_n \geq Cn^{-\frac{\beta}{2\beta+3}}(\log n)^{3r} \;\big|\; Y^n\big) \to 0, \tag{4.7}$$

with $r = (1 \vee t)(\beta+1)/(2\beta+3)$.

Condition (4.5) is for instance satisfied by the Poisson or geometric distribution. A similar condition is considered in [40]. Condition (4.6) is satisfied for usual choices of priors, such as the product of J independent copies of a distribution that admits a continuous density. Similar results hold for functions that are not uniformly bounded, with additional conditions on the tails of $\Pi_{a,J}$. This will only require additional computations similar to those in [40], and will thus not be treated here.

This theorem gives theoretical validation for a family of priors that are widely used in practice for regression problems and are easy to implement. A key feature here is that we can control the transformation of a spline basis function by the operator K through (4.4), which in turn allows us to control the change of norms. This point is highly interesting as it gives guidelines for the construction of priors for inverse problems. Namely, it suggests that a prior whose geometry does not change too much through the application of the operator K could lead to optimal contraction for the inverse problem.

4.2. Deconvolution using mixture priors

In this section, we consider the model (4.1), where K is the convolution operator in $\mathbb{R}$. This model is widely used in practice, especially when considering auxiliary variables in a regression setting or for image deblurring. For a convolution kernel $\lambda \in L_2(\mathbb{R})$ symmetric around 0, and for all $f \in L_2(\mathbb{R})$, we define K as

$$Kf(x) = \lambda \star f(x) = \int_{\mathbb{R}} f(u)\lambda(x-u)du, \qquad \text{for } x \in \mathbb{R}. \tag{4.8}$$

To the authors' best knowledge, theoretical properties of Bayesian nonparametric approach to this nonparametric regression model have not been studied in the literature. In this setting we consider a mixture type prior on f, and derive an upper bound for the posterior contraction rate. Mixture priors are common in the Bayesian literature: [25], [24] and [41] consider mixtures of Gaussian kernels, [30] consider location scale mixture and [34] studies mixtures of betas. Nonetheless, since they do not fit well into the usual setting based on the SVD of the operator, mixture priors have not be considered in the literature for ill-posed inverse problems. In our case, they proved particularly well suited for the deconvolution problem.

Let $Y^n = (Y_1, \dots, Y_n)$ be sampled from model (4.1) for a true regression function $f_0 \in L_2(\mathbb{R})$ with $\mathcal{X} = \mathbb{R}$, and assume that for $c_x > 0$, for all $i = 1, \dots, n$, $x_i \in [-c_x \log n, c_x \log n]$. It is equivalent to imposing tail conditions on the design distribution in the random design setting. We choose a prior that is well suited for f_0 in the Sobolev ball $W^\beta(L)$, for some $\beta > 0$. To avoid technicalities, we will also assume that f_0 has finite support, that we may choose to be $[0, 1]$ without loss of generality. Similar results should hold for function with support on $\mathbb{R}$ with additional assumptions on the tails of f_0 but are not treated here.

For a collection of kernels Ψ_v that depend on the parameter v, a positive integer J and a sequence of nodes $(z_1, \dots, z_J)$ we consider the following decomposition of the regression function f from the model (4.1)

$$f(\cdot) = \sum_{j=1}^{J} w_j \Psi_v(\cdot - z_j),$$

where $(w_1, \dots, w_J) \in \mathbb{R}^J$ is a sequence of weights. We choose Ψ_j proportional to a Gaussian kernel of variance v^2 and the uniform sequence of nodes $z_j = j/J$ for j such that $j/J \in [-2c_x \log n, 2c_x \log n]$

$$\Psi_{j,v}(x) = \Psi_v(x - z_j) = \frac{1}{\sqrt{2\pi v^2}} e^{-\frac{(x-j/J)^2}{2v^2}},$$

The choice of a Gaussian kernel is fairly natural in the nonparametric literature. In our specific case it will prove to be particularly well suited. The main advantage of Gaussian kernels in this case is that we can easily compute the Fourier transform of f and thus use a similar approach as in Section 3.1 to control the modulus of continuity. We consider the following prior distribution on f

$$\Pi := \begin{cases} J \sim \Pi_J \\ v \sim \Pi_v \\ w_1, \dots, w_J | J \sim \bigotimes_{j=1}^{J} N(0,1) \end{cases} \tag{4.9}$$

We use a specific Gaussian prior for the weights $(w_1, \dots, w_J)$ in order to use the results on Reproducing Kernel Hilbert Spaces following [14] to derive contraction rate for the direct problem. However, we believe that the following result should hold for more general classes of priors, but the computations would be more involved.

Following [19], we define the degree of ill-posedness of the problem through the Fourier transform of the convolution kernel. For $p > 0$, we say that the problem is mildly ill-posed of degree p if there exist some constants $c, C > 0$ such that for $\hat{\lambda}$, the Fourier transform of λ,

$$\hat{\lambda}(t) = \int \lambda(u) e^{itu} du,$$

we have for $|t|$ sufficiently large

$$c|t|^{-p} \leq |\hat{\lambda}(t)| \leq C|t|^{-p}, \quad p \in \mathbb{N}^*, \tag{4.10}$$

For all $f_0 \in W^\beta(L)$, we have that $Kf_0 \in W^{\beta+p}(L')$ for $L' = LC$. Under these conditions, the following Theorem gives an upper bound on the posterior contraction rate.

Theorem 4.2. *Let $Y^n = (Y_1, \dots, Y_n)$ be sampled from (4.1) with $\mathcal{X} = \mathbb{R}$ and assume that the design satisfies $(x_1, \dots, x_n) \in [-c_x \log n, c_x \log n]^n$. Let f_0 be such that for $\beta \in \mathbb{N}^*$ and $M > 0$, $f_0 \in W^\beta(L)$ with support on $[0,1]$ and $\|f_0\|_\infty \leq M$. Consider K as in (4.8) with λ satisfying (4.10). Let Π be a prior distribution as in (4.9) with*

$$\Pi_J(J = j) \asymp j^{-s},$$

$$v^{-q} \exp\Big(-\frac{c_d}{v} \log(1/v)^u\Big) \lesssim \Pi_v(v) \lesssim v^{-q} \exp\Big(-\frac{c_u}{v} \log(1/v)^u\Big),$$

for some positive constants s, c_u, c_d, q, and u. Then there exist constants C and r depending only on Π, L, K and M such that

$$\mathrm{E}_0 \Pi\big(\|f - f_0\| \geq C n^{-\frac{\beta}{1+2\beta+2p}} (\log n)^r \;\big|\; Y^n\big) \to 0.$$

Note that the prior does not depend on the regularity β of the true f_0 and the posterior contracts at the minimax rate. Our approach is thus adaptive. Moreover, the prior does not depend on the degree of ill-posedness either. It is thus well suited for a wide variety of convolution kernels. In particular, this can be useful when the operator is only partially known, as in this case when the regularity of the kernel may not be accessible. However, this is beyond the scope of this article.

We prove Theorem 4.2 by applying Theorem 2.1 together with Lemma 2.1. A first difficulty is to define the sets $\mathcal{S}_n$ on which we can control the modulus of continuity. A second problem is to derive the posterior contraction rate for the direct problem, given that in our setting Kf is supported on the real line: [14] derived the posterior contraction rate only for Hölder smooth functions with bounded support. However, their results directly extend to the case of convolution of Sobolev functions with bounded support given the results of [39]. The complete proof of this Theorem is postponed to Section 5.3.2.

5. Proofs

5.1. Proof of the main theorem

Proof of Theorem 2.1. By the definition of the modulus of continuity

$$\begin{aligned}
&\mathrm{E}_0\Pi\big(f : d(f,f_0) \geq \omega(\mathcal{S}_n, f_0, d, d_K, M_n\epsilon_n) \,|\, Y^n\big) \\
&\qquad \leq \mathrm{E}_0\Pi\big(f \in \mathcal{S}_n : d(f,f_0) \geq \omega(\mathcal{S}_n, f_0, d, d_K, M_n\epsilon_n) \,|\, Y^n\big) + \mathrm{E}_0\Pi(\mathcal{S}_n^c \,|\, Y^n) \\
&\qquad \leq \mathrm{E}_0\Pi\big(f \in \mathcal{S}_n : d_K(Kf,Kf_0) \geq M_n\epsilon_n \,|\, Y^n\big) + \mathrm{E}_0\Pi(\mathcal{S}_n^c \,|\, Y^n).
\end{aligned}$$

Together with (2.2) and (2.3) it completes the proof. □

5.2. Proofs of Section 3

5.2.1. Mildly ill-posed problems

Proof of Theorem 3.1. We first note that if $\|f\|_\beta \leq R$, then $\|Kf\|_{\beta+p} \leq CR$. Next we verify the condition of Lemma 2.1. Let

$$k_n = n^{\frac{1}{1+2\alpha+2p}}, \quad \rho_n = n^{-\frac{(\alpha\wedge\beta)}{1+2\alpha+2p}}, \quad \epsilon_n = n^{-\frac{(\alpha\wedge\beta)+p}{1+2\alpha+2p}}.$$

Note that

$$n\epsilon_n^2 = n\cdot n^{-\frac{2(\alpha\wedge\beta)+2p}{1+2\alpha+2p}} = n^{\frac{1+2\alpha-2(\alpha\wedge\beta)}{1+2\alpha+2p}} = \epsilon_n^{-\frac{1+2\alpha-2(\alpha\wedge\beta)}{(\alpha\wedge\beta)+p}},$$

hence $\Pi(B_n(Kf_0,\epsilon_n)) \gtrsim \exp(-C_2 n\epsilon_n^2)$ uniformly over a Sobolev ball of radius R (see Lemma 5.1 at the end of this subsection).

Note also that

$$\rho_n^2 k_n^{1+2\alpha} = n^{-\frac{2(\alpha\wedge\beta)}{1+2\alpha+2p}} \cdot n^{\frac{1+2\alpha}{1+2\alpha+2p}} = n^{\frac{1+2\alpha-2(\alpha\wedge\beta)}{1+2\alpha+2p}} = n\epsilon_n^2,$$

and given $c \geq 2(1+2\alpha)/\alpha$ we have $\Pi(\mathcal{S}_n^c) \leq \exp(-(c/8)n\epsilon_n^2)$ by Lemma 5.2.

Hence

$$\frac{\Pi(\mathcal{S}_n^c)}{\Pi(B_n(Kf_0,\epsilon_n))} \lesssim \exp\Big(-\Big(\frac{c}{8}-C_2\Big)n\epsilon_n^2\Big),$$

uniformly over a ball of radius R. The condition of Lemma 2.1 is verified upon choosing $c = 8(2+C_2)\vee 2(1+2\alpha)/\alpha$.

Finally, we note that (cf. (3.4))

$$\begin{aligned}\omega(\mathcal{S}_n, f_0, \|\cdot\|,\|\cdot\|, M_n\epsilon_n) &\\ &\lesssim M_n n^{\frac{p}{1+2\alpha+2p}}\cdot n^{-\frac{(\alpha\wedge\beta)+p}{1+2\alpha+2p}} + n^{-\frac{(\alpha\wedge\beta)}{1+2\alpha+2p}} + n^{-\frac{\beta}{1+2\alpha+2p}}\\ &\lesssim M_n n^{-\frac{(\alpha\wedge\beta)}{1+2\alpha+2p}},\end{aligned}$$

which ends the proof. □

Lemma 5.1. *Suppose $f_0 \in S^\beta$. Then for every $R>0$ there exist positive constants C_1, C_2 such that for all $\epsilon \in (0,1)$,*

$$\inf_{\|f_0\|_\beta \le R} \Pi(B_n(Kf_0,\epsilon)) \ge C_1 \exp\Big(-C_2\epsilon^{-\frac{1+2\alpha-2(\alpha\wedge\beta)}{(\alpha\wedge\beta)+p}}\Big).$$

Proof. This proof is adapted from [7]. Recall that in the white noise model the ℓ_2 balls and Kullback–Leibler neighborhoods are equivalent. By independence, for any N,

$$\begin{aligned}&\Pi\Big(\sum_{i=1}^\infty (\kappa_i f_i - \kappa_i f_{0,i})^2 \le \epsilon^2\Big)\\ &\quad\ge \Pi\Big(\sum_{i=1}^N (\kappa_i f_i - \kappa_i f_{0,i})^2 \le \epsilon^2/2\Big)\Pi\Big(\sum_{i=N+1}^\infty (\kappa_i f_i - \kappa_i f_{0,i})^2 \le \epsilon^2/2\Big).\end{aligned} \tag{5.1}$$

Also

$$\sum_{i=N+1}^\infty (\kappa_i f_i - \kappa_i f_{0,i})^2 \le 2\sum_{i=N+1}^\infty \kappa_i^2 f_i^2 + 2\sum_{i=N+1}^\infty \kappa_i^2 f_{0,i}^2. \tag{5.2}$$

The second sum in the display above is less than or equal to

$$2N^{-2\beta-2p}\sum_{i=N+1}^\infty i^{2\beta} f_{0,i}^2 \le 2N^{-2\beta-2p}\|f_0\|_\beta^2 < \frac{\epsilon^2}{4},$$

whenever $N > N_1 = (8\|f_0\|_\beta^2)^{1/(2\beta+2p)}\epsilon^{-1/(\beta+p)}$.

By Chebyshev's inequality, the first sum on the right-hand side of (5.2) is less than $\epsilon^2/4$ with probability at least

$$1 - \frac{8}{\epsilon^2}\sum_{i=N+1}^\infty \mathrm{E}_\Pi(\kappa_i^2 f_i^2) = 1 - \frac{8}{\epsilon^2}\sum_{i=N+1}^\infty i^{-1-2\alpha-2p} \ge 1 - \frac{4}{(\alpha+p)N^{2(\alpha+p)}\epsilon^2} > 1/2$$

if $N > N_2 = (8/(\alpha+p))^{1/(2\alpha+2p)}\epsilon^{-1/(\alpha+p)}$.

To bound the first term in (5.1) we apply Lemma 6.2 in [7] with $\xi_i = \kappa_i f_{0,i}$ and $\delta^2 = \epsilon^2/2$. Note that

$$\begin{aligned}\sum_{i=1}^{N} i^{1+2\alpha+2p}\xi_i^2 &= \sum_{i=1}^{N} i^{1+2\alpha+2p}\cdot i^{-2p} f_{0,i}^2 \\ &= \sum_{i=1}^{N} i^{1+2\alpha-2\beta} f_{0,i}^2 i^{2\beta} \leq N^{(1+2\alpha-2\beta)\vee 0}\|f_0\|_\beta^2.\end{aligned}$$

Therefore,

$$\begin{aligned}&\Pi\Big(\sum_{i=1}^{N}(\kappa_i f_i - \kappa_i f_{0,i})^2 \leq \epsilon^2/2\Big) \\ &\qquad \geq \exp\Big(-\Big(1+2\alpha+2p+\frac{\log 2}{2}\Big)N\Big)\exp\Big(-N^{(1+2\alpha-2\beta)\vee 0}\|f_0\|_\beta^2\Big) \\ &\qquad\quad \times \Pr\Big(\sum_{i=1}^{N} V_i^2 \leq 2\delta^2 N^{1+2\alpha+2p}\Big).\end{aligned}$$

The last term, by the central limit theorem, is at least $1/4$ if $2\delta^2 N^{1+2\alpha+2p} > N$ and N is large, that is, $N > N_3 = \epsilon^{-1/(\alpha+p)}$ and $N > N_4$, where N_4 does not depend on f_0. Choosing $N = \max\{N_1, N_2, N_3, N_4\}$ we obtain

$$\begin{aligned}&\Pi(f : \|Kf - Kf_0\| \leq \epsilon) \\ &\qquad \geq \frac{1}{8}\exp\Big(-\Big(1+2\alpha+2p+\frac{\log 2}{2}\Big)N\Big)\exp\Big(-N^{(1+2\alpha-2\beta)\vee 0}\|f_0\|_\beta^2\Big).\end{aligned}$$

Consider $\alpha \geq \beta$. Then $\exp(-N) \geq \exp(-N^{(1+2\alpha-2\beta)})$ so

$$\Pi(f : \|Kf - Kf_0\| \leq \epsilon) \geq \frac{1}{8}\exp\Big(-C_3 N^{(1+2\alpha-2\beta)}\Big),$$

for some constant C_3 that depends only on α, β, p and $\|f_0\|_\beta^2$. Moreover, since $\epsilon < 1$ and $\alpha \geq \beta$, N is dominated by $\epsilon^{-1/(\beta+p)}$ and we can write

$$\Pi(f : \|Kf - Kf_0\| \leq \epsilon) \geq \frac{1}{8}\exp\Big(-C_4\epsilon^{-\frac{1+2\alpha-2\beta}{\beta+p}}\Big),$$

where C_4 depends on f_0 again through $\|f_0\|_\beta^2$ only.

Now consider $\alpha < \beta$. Similar arguments lead to

$$\Pi(f : \|Kf - Kf_0\| \leq \epsilon) \geq \frac{1}{8}\exp\Big(-C_5\epsilon^{-\frac{1}{\alpha+p}}\Big),$$

for some constant C_5 that depends only on α, β, p and $\|f_0\|_\beta^2$. □

Lemma 5.2. *Let ρ_n be an arbitrary sequence tending to 0, c be an arbitrary constant, and let the sequence $k_n \to \infty$ satisfy $k_n^{2\alpha} \geq 2(1+2\alpha)/(\alpha c \rho_n^2)$. Then*

$$\Pi(\mathcal{S}_n^c) \leq \exp\Big(-\frac{c}{8}\rho_n^2 k_n^{1+2\alpha}\Big).$$

Proof. For $W_1, W_2, \ldots$ independent standard normal random variables

$$\Pi(\mathcal{S}_n^c) = \Pr\Big(\sum_{i>k_n} \lambda_i W_i^2 > c\rho_n^2\Big).$$

For some $t > 0$

$$\begin{aligned}
&\Pr\Big(\sum_{i>k_n} \lambda_i W_i^2 > c\rho_n^2\Big) \\
&\quad = \Pr\Big(\exp\Big(t\sum_{i>k_n} \lambda_i W_i^2\Big) > \exp(tc\rho_n^2)\Big) \leq \exp(-tc\rho_n^2)\mathrm{E}\exp\Big(t\sum_{i>k_n} \lambda_i W_i^2\Big) \\
&\quad = \exp(-tc\rho_n^2)\prod_{i>k_n} \mathrm{E}\exp(t\lambda_i W_i^2) = \exp(-tc\rho_n^2)\prod_{i>k_n}(1-2t\lambda_i)^{-1/2}.
\end{aligned}$$

We first applied Markov's inequality, and later used properties of the moment generating function. Here we additionally assume that $2t\lambda_i < 1$ for $i > k_n$.

We take the logarithm of the right-hand side of the previous display. Since $\log(1-y) \geq -y/(1-y)$, we have

$$\begin{aligned}
&-tc\rho_n^2 + \sum_{i>k_n} \log(1-2t\lambda_i)^{-1/2} \\
&\qquad = -tc\rho_n^2 - \frac{1}{2}\sum_{i>k_n} \log(1-2t\lambda_i) \leq -tc\rho_n^2 + \frac{1}{2}\sum_{i>k_n} \frac{2t\lambda_i}{1-2t\lambda_i}.
\end{aligned}$$

We continue with the latter term, noticing that $1 - 2t\lambda_i > 1 - 2tk_n^{-1-2\alpha}$ for $i > k_n$

$$\frac{1}{2}\sum_{i>k_n} \frac{2t\lambda_i}{1-2t\lambda_i} \leq \frac{t}{1-2tk_n^{-1-2\alpha}}\sum_{i>k_n} i^{-1-2\alpha}.$$

Since $x^{-1-2\alpha}$ is decreasing, we have that

$$\sum_{i>k_n} i^{-1-2\alpha} \leq \int_{k_n}^{\infty} x^{-1-2\alpha}\,dx + k_n^{-1-2\alpha} = \frac{k_n^{-2\alpha}}{2\alpha} + k_n^{-1-2\alpha} \leq k_n^{-2\alpha}\frac{1+2\alpha}{2\alpha},$$

noting that $k_n > 1$ for n large enough. Finally

$$-tc\rho_n^2 + \sum_{i>k_n} \log(1-2t\lambda_i)^{-1/2} \leq -tc\rho_n^2 + \frac{1+2\alpha}{2\alpha}\frac{t}{1-2tk_n^{-1-2\alpha}}k_n^{-2\alpha}.$$

Thus for $t = k_n^{1+2\alpha}/4$

$$\Pi(\mathcal{S}_n^c) \leq \exp\Big(-\frac{c}{4}\rho_n^2 k_n^{1+2\alpha} + \frac{1+2\alpha}{4\alpha}k_n\Big) \leq \exp\Big(-\frac{c}{8}\rho_n^2 k_n^{1+2\alpha}\Big),$$

since $k_n^{2\alpha} \geq 2(1+2\alpha)/(\alpha c\rho_n^2)$. □

5.2.2. Severely and extremely ill-posed problems

Proof of Theorem 3.2. Assume for brevity that we have the exact equality $\kappa_i = \exp(-\gamma i^p)$. Dealing with the general case is straightforward, but makes the proofs somewhat lengthier.

Since $Y_i|f_i \sim N(\kappa_i f_i, n^{-1})$ and $f_i \sim N(0, \lambda_i)$ for $i \leq k_n$, the posterior distribution (for Kf) can be written as $(Kf)_i|Y^n \sim N(\sqrt{nt_{i,n}}Y_i, v_{i,n})$ for $i \leq k_n$, where

$$v_{i,n} = \frac{\lambda_i \kappa_i^2}{1 + n\lambda_i \kappa_i^2}, \qquad t_{i,n} = \frac{n\lambda_i^2 \kappa_i^4}{(1 + n\lambda_i \kappa_i^2)^2}.$$

Since the posterior is Gaussian, we have

$$\int \|Kf - Kf_0\|^2 \, d\Pi(Kf|Y^n) = \|\widehat{Kf} - Kf_0\|^2 + \sum_{i \leq k_n} v_{i,n}, \tag{5.3}$$

where $\widehat{Kf}$ denotes the posterior mean and can be rewritten as:

$$\begin{aligned}\widehat{Kf} &= \Big(\frac{n\lambda_i \kappa_i^2}{1 + n\lambda_i \kappa_i^2} Y_i\Big)_{i=1}^{k_n} = \Big(\frac{n\lambda_i \kappa_i^3 f_{0,i}}{1 + n\lambda_i \kappa_i^2} + \frac{\sqrt{n}\lambda_i \kappa_i^2 Z_i}{1 + n\lambda_i \kappa_i^2}\Big)_{i=1}^{k_n} \\ &=: \mathrm{E}\widehat{Kf} + (\sqrt{t_{i,n}} Z_i)_{i=1}^{k_n}.\end{aligned}$$

By Markov's inequality the left side of (5.3) is an upper bound to $M_n^2 \varepsilon_n^2$ times the desired posterior probability. Therefore, in order to show that $\Pi(f : \|Kf - Kf_0\| \geq M_n \varepsilon_n | Y^n)$ goes to zero in probability, it suffices to show that the expectation (under the true f_0) of the right hand side of (5.3) is bounded by a multiple of ε_n^2. The last term is deterministic. As for the first term we have

$$\mathrm{E}\|\widehat{Kf} - Kf_0\|^2 = \|\mathrm{E}\widehat{Kf} - Kf_0\|^2 + \sum_{i \leq k_n} t_{i,n}.$$

We also observe

$$\|\mathrm{E}\widehat{Kf} - Kf_0\|^2 = \sum_{i \leq k_n} \frac{\kappa_i^2 f_{0,i}^2}{(1 + n\lambda_i \kappa_i^2)^2} + \sum_{i > k_n} \kappa_i^2 f_0^2.$$

Note that $t_{i,n} \leq n^{-1}$ and $s_{i,n} \leq n^{-1}$, hence

$$\sum_{i \leq k_n} v_{i,n} \lesssim n^{-1} k_n \asymp n^{-1} (\log n)^{\frac{1}{p}}, \quad \sum_{i \leq k_n} t_{i,n} \lesssim n^{-1} k_n \asymp n^{-1} (\log n)^{\frac{1}{p}}.$$

By Lemma 5.3

$$\sum_{i \leq k_n} \frac{\kappa_i^2 f_{0,i}^2}{(1 + n\lambda_i \kappa_i^2)^2} + \sum_{i > k_n} \kappa_i^2 f_{0,i}^2 \lesssim \|f_0\|_\beta^2 n^{-\frac{2\gamma}{\xi + 2\gamma}} (\log n)^{-\frac{2\beta}{p} + \frac{2\gamma\alpha}{p(\xi + 2\gamma)}}.$$

Therefore, the posterior contraction rate for the direct problem is given by

$$(\log n)^{-\frac{\beta}{p} + \frac{\gamma\alpha}{p(\xi + 2\gamma)}} n^{-\frac{\gamma}{\xi + 2\gamma}},$$

and is uniform over Sobolev balls of fixed radius. [This bound is also valid if $\xi = 0$ and $\alpha \geq 1 + 2\beta$.]

By (3.4) an upper bound for the modulus of continuity is given by

$$\begin{aligned}\omega(\mathcal{S}_n, f_0, \|\cdot\|, \|\cdot\|, M_n\epsilon_n) &\lesssim M_n \exp(\gamma k_n^p)\epsilon_n + k_n^{-\beta} \\ &\lesssim M_n n^{\frac{\gamma}{\xi+2\gamma}} (\log n)^{-\frac{\gamma\alpha}{p(\xi+2\gamma)}} \epsilon_n + (\log n)^{-\frac{\beta}{p}} \\ &\lesssim M_n (\log n)^{-\frac{\beta}{p}},\end{aligned}$$

which ends the proof. □

Lemma 5.3. *It holds that*

$$\sum_{i\leq k_n} \frac{\kappa_i^2 f_{0,i}^2}{(1+n\lambda_i\kappa_i^2)^2} + \sum_{i>k_n} \kappa_i^2 f_{0,i}^2 \lesssim \|f_0\|_\beta^2 n^{-\frac{2\gamma}{\xi+2\gamma}} (\log n)^{-\frac{2\beta}{p}+\frac{2\gamma\alpha}{p(\xi+2\gamma)}}.$$

Proof. As for the first sum we have

$$\begin{aligned}\sum_{i\leq k_n} \frac{\kappa_i^2 f_{0,i}^2}{(1+n\lambda_i\kappa_i^2)^2} &\leq n^{-2} \sum_{i\leq k_n} \lambda_i^{-2}\kappa_i^{-2} i^{-2\beta} i^{2\beta} f_{0,i}^2 \\ &= n^{-2} \sum_{i\leq k_n} i^{2(\alpha-\beta)} \exp(2(\xi+\gamma)i^p) i^{2\beta} f_{0,i}^2,\end{aligned}$$

and for k_n large enough all terms $i^{2(\alpha-\beta)}\exp(2(\xi+\gamma)i^p)$ are dominated by $k_n^{2(\alpha-\beta)}\exp(2(\xi+\gamma)k_n^p)$, so

$$\sum_{i\leq k_n} \frac{\kappa_i^2 f_{0,i}^2}{(1+n\lambda_i\kappa_i^2)^2} \leq n^{-2} k_n^{2(\alpha-\beta)} \exp(2(\xi+\gamma)k_n^p)\|f_0\|_\beta^2. \tag{5.4}$$

As for the second sum we note that

$$\sum_{i>k_n} \kappa_i^2 f_{0,i}^2 = \sum_{i>k_n} \exp(-2\gamma i^p) i^{-2\beta} i^{2\beta} f_{0,i}^2,$$

and since $\exp(-2\gamma i^p) i^{-2\beta}$ is monotone decreasing

$$\sum_{i>k_n} \kappa_i^2 f_{0,i}^2 \leq \exp(-2\gamma k_n^p) k_n^{-2\beta} \|f_0\|_\beta^2. \tag{5.5}$$

Recall that $\exp(k_n^p) = (nk_n^{-\alpha})^{1/(\xi+2\gamma)}$ and therefore we can rewrite the bounds in (5.4) and (5.5) as

$$n^{-2} k_n^{2(\alpha-\beta)} \left(nk_n^{-\alpha}\right)^{\frac{2(\xi+\gamma)}{\xi+2\gamma}} = n^{-\frac{2\gamma}{\xi+2\gamma}} k_n^{-2\beta+\frac{2\gamma\alpha}{\xi+2\gamma}},$$

and

$$k_n^{-2\beta} \left(nk_n^{-\alpha}\right)^{-\frac{2\gamma}{\xi+2\gamma}} = n^{-\frac{2\gamma}{\xi+2\gamma}} k_n^{-2\beta+\frac{2\gamma\alpha}{\xi+2\gamma}}.$$

Finally, since k_n in this case can be taken of the order $(\log n)^{1/p}$, we obtain the desired upper bound. □

5.3. Proofs of Section 4

5.3.1. Numerical differentiation using spline prior

We first compute an upper bound for the modulus of continuity. For $a \in \mathbb{R}^J$ we define $\Delta(a) \in \mathbb{R}^{J-1}$ such that $\Delta(a)_i = a_{i+1} - a_i$, for $i = 1, \dots, (J-1)$. Given conditions **D1** and **D2** we get,

$$\begin{aligned}\|f\|_n^2 = J^2 \Delta(a)' \Sigma_n^{q-1} \Delta(a) &\lesssim J^2 \frac{1}{J-1} \|\Delta(a)\|_{J-1}^2 \\ &\lesssim J^2 \frac{1}{J-1} \|a\|_J^2 \lesssim J^2 \|Kf\|_n^2.\end{aligned}$$

To apply Theorem 2.1, we first need to derive a contraction rate for Kf. Note that in this case we simply have a standard non parametric regression model with a spline prior. This model has been extensively studied in the literature (see, e.g., [24] or [15]) and we can easily adapt their results to derive minimax adaptive contraction rates.

Lemma 5.4. *Let Π be as in Theorem 4.1. Let Y_n be sampled form model 4.1 with $f = f_0$ and assume that $f_0 \in \Theta(\beta, L, H)$ with $\beta \leq q - 1$. Then there exists a constant C depending only on H, L, Π, and q such that*

$$\mathrm{E}_0 \Pi\big(\|Kf - Kf_0\|_n \geq C n^{-\frac{\beta+1}{2\beta+3}} (\log n)^r \;\big|\; Y_n\big) \to 0$$

with $r = (1 \vee t)\beta/(2\beta+1)$.

Similar results have been proved in [40], however the authors do not give a direct proof of their result. Here this lemma gives us directly the posterior contraction rate for the direct problem. The proof of this lemma is postponed to the end of this subsection.

Proof of Theorem 4.1. We now derive the posterior contraction rate of the posterior distribution for the inverse problem. We first get an upper bound for the modulus of continuity, for $f \in \mathcal{S}_n$. Using standard approximation results on splines (e.g. [13]), we have that for all J there exists $a^0 \in \mathbb{R}^J$ such that

$$\Big\| f_0 - \sum_{j=1}^{J-1} (a_{j+1}^0 - a_j^0)(B_{j,q-1}) \Big\|_\infty \leq (J-1)^{-\beta} \|f_0\|_\infty,$$

and

$$\Big\| Kf_0 - \sum_{j=1}^{J} a_j^0 B_{j,q} \Big\|_\infty \leq J^{-\beta-1} \|Kf_0\|_\infty.$$

We thus deduce that for $J \geq 2$,

$$\begin{aligned}\|f - f_0\|_n &\leq \|f - f_{a^0}\|_n + \|f_{a^0} - f_0\|_n \\ &\leq C J^{-1} \|Kf - Kf_n\|_n + \|f_{a^0} - f_0\|_n \\ &\leq C J^{-1} \|Kf - Kf_0\|_n + \|Kf_{a^0} - Kf_0\|_n + \|f_{a^0} - f_0\|_n.\end{aligned}$$

We can thus deduce an upper bound for the modulus of continuity

$$\omega(S_n, f_0, \|\cdot\|_n, \|\cdot\|_n, \delta) \leq J_n \delta.$$

Applying Theorem 2.1 gives

$$\mathrm{E}_0 \Pi\big(\|f - f_0\|_n \geq C n^{-\frac{\beta}{2\beta+3}} (\log n)^r \;\big|\; Y^n\big) \to 0,$$

for a constant $C > 0$ depending only on $\|f_0\|_\infty$, r, and Π. □

Proof of Lemma 5.4. We prove the lemma using Theorem 4 in [24]. Let $\beta \leq q$ and f_0 be in $\mathcal{H}(\beta, L)$ and set $\epsilon_n = C n^{-(\beta+1)/(2\beta+3)} (\log n)^r$ with $r = (1 \vee t)\beta/(2\beta+1)$. Set $J_n := J_0 n \epsilon_n^2 \log(n)^{-t}$ for a fixed constant $J_0 > 0$ and consider the sets $\mathcal{S}_n$ defined by

$$\mathcal{S}_n := \big\{ J \leq J_n, a \in \mathbb{R}^J \big\}$$

We first control the local entropy function $N(\epsilon, \{J, a \in \mathcal{S}_n : \|Kf - Kf_0\|_n \leq \epsilon_n\}, \|\cdot\|_n)$. By using the same reasoning as in the proof of Theorem 12 in [24], for all $J \in S_n$ we get

$$\log(N(\epsilon, \{J, a \in \mathcal{S}_n : \|Kf - Kf_0\|_n \leq \epsilon_n\}, \|\cdot\|_n)) \leq n\epsilon_n^2.$$

The prior mass of the set $\mathcal{S}_n$ is easily controlled using condition (4.5):

$$\Pi(\mathcal{S}_n^c) = \Pi_J(J > J_n) \leq \exp(-c_u J_n (\log J_n)^t).$$

We now need to control the prior mass of Kullback–Leibler neighborhoods of Kf_0. Note that this condition will also be useful to apply Lemma 2.1 and thus derive the posterior contraction rate for the direct problem. Let $B_n(Kf_0, \epsilon)$ be defined as in (2.4).

Using the results of Section 7.3 in [24], setting $\tilde{J}_n = J_n (\log n)^{-r/\beta}$ we deduce that for some constant c depending only on σ

$$B_n(Kf_0, \epsilon_n) \supset \{\tilde{J}_n \leq J \leq 2\tilde{J}_n, \|Kf - Kf_0\|_n^2 \leq c\epsilon_n^2\}.$$

Standard approximation results on splines give that for all J there exists a sequence $a_0 = (a_{0,1}, \ldots, a_{0,J})$ such that

$$\Big\| Kf_0 - \sum_{j=1}^{J} a_{0,j} B_{j,q} \Big\|_n \leq J^{-\beta-1} \|Kf_0\|_\beta \leq J^{-\beta-1} L.$$

Given condition **D1** on the design, we thus have that for a constant $c' > 0$ depending only on σ and L

$$B_n(Kf_0, \epsilon_n) \supset \big\{ \tilde{J}_n \leq J \leq 2\tilde{J}_n, \|a - a_0\|_{\tilde{J}_n} \leq c' \tilde{J}_n^{1/2} \epsilon_n \big\}.$$

Therefore, we obtain a lower bound on the prior mass of a Kullback–Leibler neighbourhood of Kf_0:

$$\begin{aligned} \Pi(B_n(Kf_0, \epsilon_n)) &\geq \Pi\big(\tilde{J}_n \leq J \leq 2\tilde{J}_n, \|a - a_0\|_n \leq c' \tilde{J}_n^{1/2} \epsilon_n\big) \\ &\geq \exp\big(-\tilde{J}_n (c_d (\log \tilde{J}_n)^t + c_2 \log(\tilde{J}_n^{-1/2} \epsilon_n^{-1}))\big). \end{aligned}$$

We thus have for $C_2 > 0$,

$$\frac{\Pi(\mathcal{S}_n^c)}{\Pi(B_n(Kf_0, \epsilon_n))} \leq \exp(-C_2 J_n (\log J_n)^t), \tag{5.6}$$

which together with Theorem 4 in [24] ends the proof. □

5.3.2. Deconvolution using mixture priors

Proof of Theorem 4.2. We first specify the sets $\mathcal{S}_n$ for which we can control the modulus of continuity. Denoting $\hat{f}$ the Fourier transform of f, for any sequence a_n going to infinity and $I_n = [-a_n, a_n]$ we define for $a > 0$

$$\mathcal{S}_n = \Big\{ f : \int_{I_n} |\hat{f}(t)|^2 dt \geq a \int_{I_n^c} |\hat{f}(t)|^2 dt \Big\}. \tag{5.7}$$

We control the modulus of continuity $\omega(\mathcal{S}_n, f_0, \|\cdot\|, \|\cdot\|, \delta)$ in a similar way as in Section 3.1. First consider $f \in \mathcal{S}_n$, and denote $\hat{f}_n(\cdot) = \hat{f}(\cdot)\mathbb{I}_{I_n}(\cdot)$. We then have

$$\|f\|^2 = \|\hat{f}\|^2 \leq (1+a)\|\hat{f}_n\|^2 \lesssim a_n^{2p} \int_{I_n} |\hat{f}|^2 |\hat{\lambda}|^2 \lesssim a_n^{2p} \|Kf\|^2.$$

Note that for $f_0 \in W^\beta(L)$ we have for $f_{0,n}(x) = \int \hat{f}_{0,n}(t) e^{-itx} dt$

$$\|f_0 - f_{0,n}\| \leq 2a_n^{-\beta} L, \qquad \|Kf_0 - Kf_{0,n}\| \leq 2a_n^{-(\beta+p)} L',$$

which gives

$$\omega(\mathcal{S}_n, f_0, \|\cdot\|, \|\cdot\|, \delta) \lesssim a_n^p \delta + a_n^{-\beta}. \tag{5.8}$$

We now control the prior mass of $\mathcal{S}_n^c$ in order to apply Lemma 2.1. Denote by $l_n = \lfloor a_n/(2\Pi J) \rfloor$, $L_n = \lceil a_n/(2\Pi J) \rceil$. We have

$$\begin{aligned}
\int_{I_n} |\hat{f}(t)|^2 dt &\geq 2\pi J \int_{-L_n}^{l_n} e^{-4\pi^2 t^2 v^2} \Big| \sum_{j=1}^J w_j e^{2\pi j t} \Big| dt \\
&= 2\pi J \sum_{l=-L_n}^{l_n} \int_l^{l+1} e^{-4\pi^2 t^2 v^2} \Big| \sum_{j=1}^J w_j e^{2\pi j t} \Big| dt \\
&= 2\pi J \int_0^1 \Big| \sum_{j=1}^J w_j e^{2\pi j t} \Big| \sum_{l=-L_n}^{l_n} e^{-4\pi^2 (t+l)^2 v^2} dt \\
&\geq 2\pi J \sum_{l=-L_n}^{l_n} e^{-4\pi^2 (1+|l|)^2 v^2} \int_0^1 \Big| \sum_{j=1}^J w_j e^{2\pi j t} \Big| dt,
\end{aligned}$$

and similarly we get

$$\int_{I_n^c} |\hat{f}(t)|^2 dt \leq 2\pi J \int_0^1 \Big|\sum_{j=1}^J w_j e^{2\pi j t}\Big| \Big(\sum_{l=-\infty}^{-L_n} e^{-4\pi^2(t+l)^2 v^2} + \sum_{l=l_n}^{\infty} e^{-4\pi^2(t+l)^2 v^2} \Big) dt$$
$$\leq 2\pi J \Big(\sum_{l=-\infty}^{-L_n} e^{-4\pi^2 l^2 v^2} + \sum_{l=l_n}^{\infty} e^{-4\pi^2 l^2 v^2} \Big) \int_0^1 \Big|\sum_{j=1}^J w_j e^{2\pi j t}\Big| dt.$$

We thus deduce that for absolute constants $C' > 0$

$$\Pi(\mathcal{S}_n^c) \leq \Pi(v \leq J/a_n) \lesssim e^{-C' a_n \log a_n}.$$

We end the proof by combining this result (choosing $a_n = n\epsilon_n^2$) with Lemma 2.1, Lemma 5.5, and Theorem 2.1. □

Lemma 5.5. *Let Y^n be sample from* (4.1) *with K defined by* (4.8)*. Let Π be as in Theorem 4.2. For all $\beta \in \mathbb{N}^*$ if $f_0 \in W^\beta(L)$ with support on $[0,1]$ and $||f_0||_\infty \leq M$, we have for $C > 0$ large enough if $\epsilon_n = n^{-(\beta+p)/(1+2\beta+2p)}(\log n)^r$, where r is some constant,*

$$\mathrm{E}_0 \Pi(\|Kf - Kf_0\| \geq C\epsilon_n | Y^n) \to 0,$$

and

$$\Pi(\|Kf - Kf_0\| \leq \epsilon_n) \geq e^{-n\epsilon_n^2}.$$

Proof. This proof is based on the results of [14] and [39]. We adapt the results of [14] to our setting in order to control the posterior mass of the Kullback–Leibler neighbourhoods of Kf_0 and the posterior contraction rate for the direct problem. Following their notation we have that $K\Psi_v \in \mathcal{P}_\infty$, and thus the small ball probability $\Pi(\|f\|_\infty \leq \epsilon)$ can be controlled by their Lemma 3.3. We then extend their Lemma 3.5 to our setting. Note that with Lemma 9 of [39], Lemma 3.4 of [14] holds for the same $T_{\alpha,v}$ with $\alpha = \beta + p$. Choosing h to be as in the proof of Lemma 3.5 of [14] and denoting $\omega_0 = f_0 \star \lambda$, we have

$$h(x) = \sum_{j/J \in [-2c_x \log n, 2c_x \log n]} T_{\alpha,v}(\omega_0) \frac{1}{Jv} \Psi\Big(\frac{x - j/J}{v}\Big),$$

and thus deduce

$$\|h\|_{H^{J,v}}^2 \leq 2c_x \|T_{\alpha,v}(\omega_0)\|^2 \log n.$$

Using their decomposition (3.8), we control $|h(x) - \Psi_v \star T_{\alpha,v}(\omega_0)(x)|$ along the same lines as in their computations on page 3312. We have

$$|h(x) - \Psi_v \star T_{\alpha,v}(\omega_0)(x)|$$
$$\leq \Big| h(x) - \int_{-2c_x \log n}^{2c_x \log n} T_{\alpha,v}(\omega_0)(y) \Psi_v(x-y) dy \Big|$$
$$+ \Big| \int_{-\infty}^{-2c_x \log n} T_{\alpha,v}(\omega_0)(y) \Psi_v(x-y) dy \Big| + \Big| \int_{2c_x \log n}^{\infty} T_{\alpha,v}(\omega_0)(y) \Psi_v(x-y) dy \Big|$$

The first term on the right hand side of the above display can be controlled as in the proof of Lemma 3.5 of [14]. For the last two terms, we have

$$\Big|\int_{-\infty}^{-2c_x \log n} T_{\alpha,v}(\omega_0)(y)\Psi_v(x-y)dy\Big| + \Big|\int_{2c_x \log(n)}^{\infty} T_{\alpha,v}(\omega_0)(y)\Psi_v(x-y)dy\Big| \\ \lesssim \|T_{\alpha,v}(\omega_0)\|_\infty e^{-\frac{c_x^2(\log n)^2}{2v^2}} v^{-1}.$$

Following the same proof of Theorem 2.2 of [14], we get

$$\mathrm{E}_0\Pi(\|Kf - Kf_0\| \geq Cn^{-\frac{\beta+p}{1+2\beta+2p}}(\log n)^r | Y^n) \to 0,$$

and similarly to their equation (2.5) we get, with $\epsilon_n = n^{-(\beta+p)/(1+2\beta+2p)}(\log n)^r$, where r is some constant,

$$\Pi(\|Kf - Kf_0\| \leq \epsilon_n) \geq e^{-n\epsilon_n^2}.$$

□

6. Discussion

In this paper we propose a new approach to the problem of deriving posterior contraction rates for linear ill-posed inverse problems. More precisely, we put a prior on the parameter of interest f that naturally imposes the prior on Kf, leading to a certain rate of contraction in the direct problem. Next, we consider a sequence of sets on which the operator K possesses a continuous inverse. Then, we impose additional conditions on the prior (or the posterior itself) under which the posterior contracts at a certain rate in the inverse problem setting.

This is a great advantage of the Bayesian approach in this setting as when the posterior distribution is known to contract at a given rate in the direct problem, one only has to consider subset of high prior mass for which the norm of the inverse of the operator may be handled. Our result seems to show that the main difficulty when considering linear inverse problems is to control the change of metrics form d_K to d, which is dealt here by considering the modulus of continuity as introduced in [17] and [26]. It is also to be noted that contrariwise to existing methods, we do not require a Hilbertian structure for the parameter space, see for instance the example treated in Section 4.1. This could be particularly useful when considering nonlinear operators, and is of potential interest when considering the case of partially known operators.

We recovered (a subset of) the existing results from [28], [29], [1], [2], and [33]. Our approach should be viewed as a generalization of the ideas presented in the last paper and the existing sieve method used in the literature on posterior contraction. Furthermore, we were able to go beyond the sequence setting as well as derive posterior contraction rates for prior distributions that were not covered by the existing theory. We also treated an operator that does not admit singular value decomposition. In this sense, the approach proposed in this paper is more general, and we believe more natural, than the existing ones.

Acknowledgements

The authors would like to thank the editor, the associate editor, and the referees for their comments which helped to improve this paper. The authors are also grateful to Judith Rousseau and Eduard Belitser for helpful discussions and comments. This work is partially funded by the STAR cluster, ANR Bandhits, VICI Safe Statistics and Labex ECODEC. This work is part of the second author's PhD.

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(10):3828–3860.

[2] Agapiou, S., Stuart, A. M., and Zhang, Y.-X. (2014). Bayesian posterior contraction rates for linear severely ill-posed inverse problems. *J. Inverse Ill-Posed Probl.*, 22(3):297–321.

[3] Arbel, J., Gayraud, G., and Rousseau, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scand. J. Stat.*, 40(3):549–570.

[4] Barron, A., Schervish, M. J., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27(2):536–561.

[5] Barron, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical report, University of Illinois, Dept. of Statistics.

[6] Belitser, E. (in press). On coverage and local radial rates of credible sets. *Ann. Statist.*

[7] Belitser, E. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.*, 31(2):536–559.

[8] Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398.

[9] Castillo, I. (2014). On Bayesian supremum norm contraction rates. *Ann. Statist.*, 42(5):2058–2091.

[10] Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.*, 43(5):1986–2018.

[11] Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *Ann. Statist.*, 40(4):2069–2101.

[12] Cavalier, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems*, 24(3):034004, 19.

[13] De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag New York.

[14] de Jonge, R. and van Zanten, J. H. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Ann. Statist.*, 38(6):3300–3320.

[15] de Jonge, R. and van Zanten, J. H. (2012). Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electron. J. Stat.*, 6:1984–2001.

[16] Donnet, S., Rivoirard, V., Rousseau, J., and Scricciolo, C. (2014). Posterior concentration rates for empirical Bayes procedures, with applications to Dirichlet Process mixtures. *arXiv preprint arXiv:1406.4406*.

[17] Donoho, D. L. and Liu, R. C. (1991). Geometrizing rates of convergence. II. *Ann. Statist.*, 19(2):633–667.
[18] Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of inverse problems*, volume 375. Springer.
[19] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3):1257–1272.
[20] Florens, J.-P. and Simoni, A. (2012a). Nonparametric estimation of an instrumental regression: a quasi-Bayesian approach based on regularized posterior. *J. Econometrics*, 170(2):458–475.
[21] Florens, J.-P. and Simoni, A. (2012b). Regularized posteriors in linear ill-posed inverse problems. *Scand. J. Stat.*, 39(2):214–235.
[22] Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.*, 27(1):143–158.
[23] Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.
[24] Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.*, 35(1):192–223.
[25] Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263.
[26] Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.*, 43(5):2259–2295.
[27] Knapik, B. T., Szabó, B. T., van der Vaart, A. W., and van Zanten, J. H. (2016). Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Related Fields*, 164(3):771–813.
[28] Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.*, 39(5):2626–2657.
[29] 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.
[30] Kruijer, W., Rousseau, J., and van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.*, 4:1225–1257.
[31] Meister, A. (2011). Asymptotic equivalence of functional linear regression and a white noise inverse problem. *Ann. Statist.*, 39(3):1471–1495.
[32] Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.*, 24(6):2399–2430.
[33] Ray, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat.*, 7:2516–2549.
[34] Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.*, 38(1):146–180.
[35] Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(5):689–710.
[36] Rudin, W. (1987). *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition.
[37] Salomond, J.-B. (2014). Concentration rate and consistency of the posterior distribution for

selected priors under monotonicity constraints. *Electron. J. Statist.*, 8(1):1380–1404.
[38] Schwartz, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete*, 4:10–26.
[39] Scricciolo, C. (2014). Adaptive Bayesian density estimation in L^p-metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Anal.*, 9(2):475–520.
[40] Shen, W. and Ghosal, S. (2015). Adaptive bayesian procedures using random series priors. *Scandinavian Journal of Statistics*, 42(4):1194–1213. 10.1111/sjos.12159.
[41] Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640.
[42] Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714.
[43] Szabó, B., van der Vaart, A. W., and van Zanten, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.*, 43(4):1391–1428.
[44] Vollmer, S. J. (2013). Posterior consistency for Bayesian inverse problems through stability and regression results. *Inverse Problems*, 29(12):125011.
[45] Zhao, L. H. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.*, 28(2):532–552.