Asymptotic theory for Bayesian nonparametric procedures in inverse problems

Kolyan Michael Ray

St Catharine's College

University of Cambridge

A dissertation submitted for the degree of

Doctor of Philosophy

November 2014

Abstract

The main goal of this thesis is to investigate the frequentist asymptotic properties of nonparametric Bayesian procedures in inverse problems and the Gaussian white noise model. In the first part, we study the frequentist posterior contraction rate of nonparametric Bayesian procedures in linear inverse problems in both the mildly and severely ill-posed cases. This rate provides a quantitative measure of the quality of statistical estimation of the procedure. 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.

In the second part of this thesis, we investigate Bernstein–von Mises type results for adaptive nonparametric Bayesian procedures in both the Gaussian white noise model and the mildly ill-posed inverse setting. The Bernstein–von Mises theorem details the asymptotic behaviour of the posterior distribution and provides a frequentist justification for the Bayesian approach to uncertainty quantification. We establish weak Bernstein–von Mises theorems in both a Hilbert space and multiscale setting, which have applications in L^2 and L^∞ respectively. This provides a theoretical justification for plug-in procedures, for example the use of certain credible sets for sufficiently smooth linear functionals. We use this general approach to construct optimal frequentist confidence sets using a Bayesian approach. We also provide simulations to numerically illustrate our approach and obtain a visual representation of the different geometries involved.

Acknowledgements

I would like to extend my deepest gratitude to my supervisor, Richard Nickl, for guiding me throughout my PhD and suggesting the topics that form the basis of this thesis. He introduced me to a fascinating and active branch of mathematical statistics, which I hope to continue exploring in the years to come.

I would like to thank my friends with whom I have shared this PhD experience: Marc, Julio, Bati, Kostas (my gym-mate), Damon and Meline, Maria, Sara and Ed. I have had a great four years, both within the CMS and outside, and this is largely due to all of you

Finally, I would like to give special thanks to my parents and brother, Nikhil, who have encouraged me in my scholastic pursuits ever since I was a child. This thesis would not have been possible without your many years of love and support and I shall always be grateful to have such a wonderful family.

I also gratefully acknowledge the financial support of the Engineering and Physical Sciences Research Council (EPSRC) which has made this thesis possible.

Statement of Originality

I hereby declare that my dissertation entitled "Asymptotic theory for Bayesian nonparametric procedures in inverse problems" is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other University. I further state that no part of my dissertation has already been or is concurrently submitted for any such degree of diploma or other qualification.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where explicitly stated in the text. Below is a more detailed account.

Chapter 1 contains introductory and background material. It is review work that has been done personally.

Chapter 2 consists of original work on establishing contraction rates in linear inverse problems using an abstract testing approach. It is the result of my own work and is essentially the content of [71].

Chapter 3 consists of original work on establishing weak Bernstein–von Mises theorems for nonparametric Bayesian procedures with applications to uncertainty quantification. It is the result of my own work and is the content of [72].

Contents

Chapter 1

Introduction

The principle goal of statistics is to make inference based on observed data, that is to translate observations into conclusions. Collecting data is an uncertain process, with possible errors arising from measurement imperfections or natural variations within a population under study. A statistical procedure must therefore take into account these observational imperfections by factoring in the notion of randomness. A key question is therefore how best to account for this randomness, before making conclusions from the data.

To perform inference, it is necessary to reformulate a statistical problem in an appropriate language. Mathematics provides just such a language, being both convenient and powerful. A *model* is a mathematical approximation of some phenomenon of interest. Once a model has been established, a rigorous study can be applied to the problem based on the rules of mathematics. A model can rarely fully capture all the fine detail of a problem, but it can be invaluable for capturing key features and making predictions. We do not concern ourselves with the question of selecting an appropriate model here, focusing purely on the stage from which a model has been selected.

More formally, a model $\mathcal{P}$ is a collection of candidate probability distributions for the underlying distribution. This can range from the space of all probability distributions to much smaller classes, based on a-priori knowledge of the specific statistical problem at hand. In the frequentist paradigm, an observation Y, taking values in some measurable space $\mathcal{Y}$, is assumed to be generated from some fixed true probability distribution $\mathbb{P}_0$ belonging to a model $\mathcal{P}$. The statistician then seeks to make inference about (some feature of) $\mathbb{P}_0$ based on Y.

Rather than dealing purely in abstract models, it can be fruitful to introduce an indexing of the model that carries a more interpretable meaning. We *parametrize* a model by considering a parameter space $\mathcal{F}$ together with some map taking an element $f \in \mathcal{F}$ to an element of $\mathcal{P}$, which be denote $\mathbb{P}_f$. It is natural to consider maps $f \mapsto \mathbb{P}_f$ that are

bijections, which we henceforth assume. We can therefore rewrite

$$\mathcal{P} = \{\mathbb{P}_f : f \in \mathcal{F}\}.$$

In this case, the notion of true distribution therefore corresponds to some true parameter $f_0 \in \mathcal{F}$. Perhaps the simplest example of this is the normal distribution on $\mathbb{R}$, which can be characterized by its mean μ and its variance σ^2, so that $f = (\mu, \sigma^2)$ and $\mathbb{P}_f = N(\mu, \sigma^2)$. It is possible to parametrize a model by itself, that is $\mathcal{F} = \mathcal{P}$, which is often the case when $\mathcal{P}$ is infinite-dimensional. Indeed, situations where $\mathcal{F}$ is infinite-dimensional are the principle focus of this thesis. It is of significant interest to consider the case of *model misspecification* when the truth f_0 does not lie in the targeted parameter space. This is an interesting question, but one that we do not address here, where we always assume that the model is *well-specified*, that is $\mathbb{P}_0 \in \mathcal{P}$ (for some results of a similar flavour to those of this thesis in the misspecified case see [50, 51]).

The complexity and features of the model can therefore be efficiently characterized by the parameter space $\mathcal{F}$. We call a model *parametric* when the parameter space $\mathcal{F}$ is a subset of a finite-dimensional space, for example $\mathbb{R}^k$ or $\mathbb{Z}^k$ for some $k \geq 1$. In slightly misleading terminology, we call a model *nonparametric* if the parameter space $\mathcal{F}$ is infinite-dimensional, for example the space of all probability distributions that admit a density function. Parametric models are usually simpler to use and compute, and work well in many instances. However, they impose a high level of rigidity that can yield poor results if the true distribution does not fit the strong constraints imposed by the parametric model. On the other hand, nonparametric models allow a much richer class of target densities and can flexibly model a wider variety of phenomena. Whilst more difficult to compute, recent computational advances mean that nonparametric models are increasingly finding use in practice. This extra generality means that the mathematics underpinning such models is usually more involved than in the finite-dimensional case, and we seek to shed light on some approaches here.

Within this thesis we focus on the Bayesian approach to statistical inference. This is a flexible modelling framework that possesses a conceptual simplicity that renders it easily adaptable to a wide variety of problems. The Bayesian paradigm can be extended to a philosophical perspective, namely *subjective Bayes*, and this is a subject of much discussion. However, our goals are more pragmatic and we therefore concern ourselves with its mathematical theory rather than its philosophical underpinnings. In particular, we take a dual approach here in that we seek to study the Bayesian approach as a statistical estimation procedure from a frequentist perspective.

We shall investigate a number of frequentist properties of Bayesian nonparametric procedures in order to assess their quality for statistical estimation when applied to data. This is a central question for a number of reasons. Given the myriad of statistical procedures

available, it is important for a user to be able to compare procedures to select the most appropriate one for the problem at hand. Moreover, once a procedure has been selected, it is important to have some measure of the quality of the answers it provides.

Our focus in this thesis will be on *asymptotic*, or large sample, performance of estimators. Consider data $Y = Y^{(n)}$, where the index n denotes the sample size or quality of the observation. We are interested in the case where $n \to \infty$, that is we can take n "sufficiently large" for our analysis. Such analyses are useful since they can be used to approximate statistical procedures as well as study their quality. Moreover, asymptotic study often provides a first step in the theoretical study of statistical estimators, often providing insights into their finite-sample performance. We therefore seek a quantitative understanding of the asymptotic properties of Bayesian nonparametric procedures.

1.1 Bayesian statistics from a frequentist perspective

In this section, we introduce the Bayesian approach to statistics and its study from a frequentist point of view, in particular with regards to asymptotic results. We assume that the reader is relatively well versed in the frequentist theory of statistics, providing only a very brief review of certain key notions used in this thesis. We take a rather abstract approach to introducing Bayesian statistics and then later specialize to the setting of the white-noise model and linear inverse problems for more detailed results.

1.1.1 Frequentist measures of estimation quality

We very quickly recall some frequentist notions of quality of statistical estimation. We define the *minimax risk* or *rate* over a class $\mathcal{F}$ for some loss function $L : \mathcal{F} \times \mathcal{F} \to [0, \infty]$ by

$$r_{n,\mathcal{F}} = \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f L(\hat{f}_n(Y^{(n)}, f)), \tag{1.1.1}$$

where $\mathbb{E}_f$ denotes the expectation with respect to $\mathbb{P}_f$ and the infimum is taken over all estimators of f, that is measurable functions $\hat{f}_n : \mathcal{Y} \to \mathcal{F}$. An estimator $\hat{f}_n$ is called a *minimax estimator* if its risk attains the minimax bound $r_{n,\mathcal{F}}$. In a slight abuse of notation, an estimator is often called minimax if it attains the rate (in terms of n) implied by the minimax risk even though it does achieve the optimal constant in (1.1.1). In particular in the Bayesian nonparametrics literature, estimators are often deemed minimax if they attain the correct polynomial rate whilst ignoring logarithmic factors. We shall use the notion of minimax rate to measure the convergence rate of Bayesian methods to the true parameter f_0 under the law $\mathbb{P}_{f_0}$ it generates. For more details see Lehmann and Casella [63].

A more sophisticated measure is the notion of coverage probability. In particular we seek to quantify the error in our estimation procedure by constructing a (data-driven)

set $C_n \subset \mathcal{F}$ such that the true f_0 lies in C_n with a prescribed probability. This provides a range of estimates whilst making a quantitative statement about the reliability of the set C_n via the coverage probability. For $\gamma \in (0,1)$, a set $C_n = C_n(Y^{(n)})$ is said to be a *confidence set* with confidence level $1-\gamma$ if

$$\inf_{f \in \mathcal{F}} \mathbb{P}_f(f \in C_n) \geq 1 - \gamma.$$

The above statement is a finite-sample statement and can therefore be difficult to obtain in practice. A weaker notion is that of an *(honest) asymptotic confidence set*, where we require the statement only to hold as the data size or quality tends to infinity:

$$\liminf_{n \to \infty} \inf_{f \in \mathcal{F}} \mathbb{P}_f(f \in C_n) \geq 1 - \gamma. \tag{1.1.2}$$

The above definitions are overly conservative in that they require only a lower bound on the coverage probability. It is of great theoretical interest to construct confidence sets that are *exact*, that is there is an equality in (1.1.2). We note that the above definitions require uniformity in the parameter space $\mathcal{F}$. It is a much simpler exercise to construct sets that satisfy (1.1.2) pointwise (i.e. without the infimum over $\mathcal{F}$). However in this case, the sample size for which sufficiently high probability is achieved depends strongly on the unknown parameter f_0. Thus from a practical point of view, such pointwise asymptotic statements are of little value.

1.1.2 Prior and posterior distributions

For a statistical model $\mathcal{P}$, let $\mathcal{F}$ denote its parameter space, which we assume to be a Polish space with associated σ-algebra Σ. Suppose that we observe data $Y \equiv Y^{(n)}$ taking values in a Polish space $\mathcal{Y}$ with associated σ-algebra $\mathcal{T}$. In contrast to the frequentist approach, the Bayesian considers the parameter f to be a random variable taking value in $\mathcal{F}$. More formally, we should denote this measurable map by $\tilde{f}$ to differentiate it from an element of $\mathcal{F}$. However, in a slight abuse of notation, we simply denote this map by f since it is usually clear from the context whether this refers to a fixed value or a random variable. The Bayesian therefore considers (Y, f) as a joint random variable taking values in the space $\mathcal{Y} \times \mathcal{F}$ and supposes that there exists a probability measure

$$\overline{\Pi} : \Sigma \times \mathcal{T} \to [0,1], \tag{1.1.3}$$

which is not necessarily a product measure. As a result, there may be a (possibly complex) dependence structure between Y and f. This forms the underlying probabilistic model for the Bayesian.

By conditioning on the parameter $\tilde{f} = f$, for some $f \in \mathcal{F}$, we therefore obtain the law of the observed data under this model. In particular, by consider the conditional

distributions $\mathbb{P}_f = \overline{\Pi}(\cdot \mid \tilde{f} = f)$, we define a model $\mathcal{P} = \{\mathbb{P}_f : f \in \mathcal{F}\}$. We define the *prior* distribution to be the marginal distribution

$$\Pi(\cdot) = \overline{\Pi}(\mathcal{Y} \times \cdot) : \mathcal{T} \to [0, 1].$$

In this framework, specifying the prior and the model $\mathcal{P} = \{\mathbb{P}_f : f \in \mathcal{F}\}$ completely characterizes the Bayesian probability measure $\overline{\Pi}$. Unlike the frequentist, assuming the existence of a joint distribution $\overline{\Pi}$ allows one to condition on the data to obtain the *posterior* distribution

$$\Pi_{f|Y} : \Sigma \times \mathcal{T} \to [0, 1].$$

Assume now that the model $\mathcal{P} = \{\mathbb{P}_f : f \in \mathcal{F}\}$ is dominated by a σ-finite measure μ. The posterior distribution can then be expressed in terms of the densities $p_f = \frac{d\mathbb{P}_f}{d\mu} : \mathcal{Y} \to [0, \infty)$ via *Bayes rule*,

$$\Pi(B \mid Y) = \frac{\int_B p_f(Y) d\Pi(f)}{\int_{\mathcal{P}} p_f(Y) d\Pi(f)}, \qquad B \in \mathcal{T}. \tag{1.1.4}$$

From a mathematical perspective, Bayes rule provides a simple and powerful method to evaluate the posterior distribution. Computing it in practice is the subject of significant study and is discussed in Section 1.1.6.

Up until this point we have made no assumptions on the joint density $\overline{\Pi}$, which we now seek to place in a frequentist context. In particular, given a (parametrized) frequentist model $\mathcal{P} = \{\mathbb{P}_f : f \in \mathcal{F}\}$, we can define a Bayesian joint probability distribution $\overline{\Pi}$ on $\mathcal{Y} \times \mathcal{F}$ by specifying a prior distribution on the parameter space $\mathcal{F}$ and using the model to provide the marginal distributions. In this context, the prior distribution is then introduced by the statistician to extend the frequentist model to one of the form (1.1.3) rather than as a marginal distribution of some underlying joint distribution. In this way it is possible to separate the prior distribution from the data generating model. Using Bayes rule (1.1.4), it is then possible to condition on the data to define the posterior distribution and thus obtain a sequence of data-driven random probability measures which can be used as generalized frequentist estimators.

This additional conditioning is key to the Bayesian approach and within this framework, the posterior provides the natural way to incorporate data. The conceptual simplicity of this approach is that any desired inferential information, such as point estimators, measures of risk or predictions, can be extracted from a single object, the posterior distribution. From a theoretical (rather than computational) point of view, the inference procedure is fully automated once a prior has been specified.

The principle conceptual difficulty for the Bayesian lies in specifying the prior. This can reflect prior knowledge or beliefs, which may or may not be justified, or may be designed to be efficient for the problem at hand. The prior represents to what extent we believe in the different possibilities represented by the parameter space $\mathcal{F}$, with larger

probabilities corresponding to higher a-priori plausibility. Prior selection is the equivalent of the frequentist problem of selecting tuning parameters, and in much the same way the quality of statistical performance can rely heavily on this choice. Recently, much focus has been given to the development of nonparametric procedures, where the support of Π is infinite-dimensional. As discussed below, nonparametric procedures are especially sensitive to prior selection due to the absence of a Bernstein–von Mises theorem.

1.1.3 Posterior consistency and contraction

A basic quality that is desirable for a statistical procedure is *consistency*. In frequentist statistics, an estimator $\hat{f}_n$ of f is called consistent if it converges to f_0 (in probability) under the true distribution $\mathbb{P}_{f_0}$ of the data. The Bayesian analogue is to consider consistency of the posterior distribution.

Definition 1 (Posterior consistency)**.** *We say the posterior distribution* $\Pi_n(\cdot \mid Y^{(n)})$ *is* consistent *at* f_0 *if for every neighbourhood* W *of* f_0

$$\Pi_n(W^c \mid Y^{(n)}) \to^{\mathbb{P}_{f_0}} 0 \quad \text{as } n \to \infty.$$

When considering a metric space $(\mathcal{F}, d)$, one can restrict to $W_\delta = \{f \in \mathcal{F} : d(f, f_0) \leq \delta\}$ for all $\delta > 0$ to obtain a more familiar expression for consistency. However, the above definition also allows one to consider the notion of *weak consistency*, where one takes W to be a weak neighbourhood of f_0.

Posterior consistency says that the posterior distribution will eventually concentrate around the true parameter f_0. In other words, consistency says that the data will eventually overwhelm the (possibly incorrect) prior beliefs of a statistician, as represented by the prior distribution. However, it says nothing about the quality of estimation of the procedure, only that the posterior will correctly identify the truth eventually. In parametric models, posterior consistency follows under mild conditions due to the Bernstein–von Mises theorem.

However in infinite-dimensions, posterior consistency can fail even for seemingly simple priors. There exist priors whose influence is so strong that even an infinite amount of data can not override an incorrect positioning of the prior. A classical example is given in Freedman [34], who constructs a prior that puts positive mass on every neighbourhood of the true distribution and yet whose posterior converges to the wrong distribution. Not only is the Bayesian wrong, but he is eventually certain of his incorrect answer. This issue is discussed at length in Diaconis and Freedman [30], who further show that this behaviour is not isolated or limited to pathological functions. In a topological sense, almost every pair (f_0, Π) of truth and prior is inconsistent.

While the situation is indeed more complicated in infinite-dimensions, there exist positive general theorems for establishing posterior consistency. Doob's consistency theorem

[31] states a given prior is consistent at every point f apart from possibly some null set of the prior. While this provides a Bayesian justification, it gives no intuition as to the null set involved and so we can not be sure if a given f_0 falls into this set. In particular, in infinite-dimensional cases this null set can actually be extremely large [30].

Theorem 1.1.1 (Doob's consistency theorem)**.** *Suppose that both the parameter space $\mathcal{F}$ and the sample space $\mathcal{Y}$ are Polish spaces endowed with their Borel σ-algebras and that the map $\mathcal{F} \to \mathcal{P}$ given by $f \mapsto \mathbb{P}_f$ is bijective. Then the sequence of posterior distributions is consistent Π-almost surely.*

Schwartz [75] provides a more satisfactory answer from a frequentist point of view, permitting one to establish consistency at any given f_0 in the model.

Theorem 1.1.2 (Schwartz's consistency theorem)**.** *Let $\mathcal{P}$ be a model with a metric d, dominated by some σ-finite measure μ and assume that $\mathbb{P}_0 \in \mathcal{P}$. Let Π be a prior on $\mathcal{P}$ and assume that the following hold:*

(i) For every $\varepsilon > 0$,

$$\Pi\left(\mathbb{P} \in \mathcal{P} : -\mathbb{P}_0 \log \frac{p}{p_0} \leq \varepsilon\right) > 0,$$

(ii) for every $\varepsilon > 0$, there exist a sequence of tests ϕ_n such that

$$\mathbb{P}_0 \phi_n \to 0, \qquad \sup_{\mathbb{P}: d(\mathbb{P}, \mathbb{P}_0) > \varepsilon} \mathbb{P}(1 - \phi_n) \to 0.$$

Then for any $\epsilon > 0$, as $n \to \infty$,

$$\Pi(d(\mathbb{P}, \mathbb{P}_0) \geq \epsilon \mid Y^{(n)}) \to 0 \qquad \mathbb{P}_0 - a.s.$$

A related notion is contraction, which quantifies the rate at which the posterior distribution converges to the truth and thus provides a quantitative measure of posterior accuracy. Rather than considering a fixed ball of radius $\varepsilon > 0$, we now let the radius depend on n and seek the smallest ε_n such that a consistency-type statement holds. This yields the following definition.

Definition 2 (Posterior rate of contraction)**.** *We say that the posterior distribution $\Pi_n(\cdot \mid Y^{(n)})$ contracts around the point f_0 with rate $\varepsilon_n \downarrow 0$ if*

$$\Pi_n(f : d(f, f_0) \geq M_n \varepsilon_n \mid Y^{(n)}) \to 0$$

in $\mathbb{P}_{f_0}$-probability, for every sequence $M_n \to \infty$.

We note that posterior contraction trivially implies posterior consistency. The posterior contraction rate provides an upper bound for the speed in the sense that any sequence ε_n'

that tends to zero more slowly is also a rate of contraction. Usually, the aim is to establish a fastest rate of contraction, which depends on the statistical model $\mathcal{P}$ and metric d, and in particular whether it matches the minimax rate of estimation.

The posterior distribution yields a point estimator that converges to the true parameter f_0 at a rate equal to the rate of contraction. The minimax rate of estimation consequently provides a fundamental lower bound for the posterior contraction rate.

Proposition 1.1.3. *Suppose that the posterior distribution contracts to the true parameter f_0 at rate ε_n with respect to the metric d on $\mathcal{F}$. Define $\hat{f}_n$ to be the centre of a smallest d-ball that contains posterior mass at least $1/2$. Then for any $\eta > 0$, there exists $M > 0$ such that*

$$\mathbb{P}_{f_0}\left(d(\hat{f}_n, f_0) > M\varepsilon_n\right) \leq \eta.$$

In finite-dimensions, contraction at the optimal $1/\sqrt{n}$ rate follows from the Bernstein–von Mises theorem. In infinite (and finite) dimensions, Ghosal et al. [36] provide a contraction analogue to Schwartz's result that deals with any given f_0 in the model. This theorem requires that the prior puts sufficient mass on shrinking Kullback-Leibler neighbourhoods around the true parameter. This general result relies on the possibility of testing for the truth against an alternative which consists of the complement of a shrinking ball around it. In particular, the authors require an exponentially decreasing type-II error. The existence of such tests is a highly non-trivial matter and depends strongly on the metric under consideration. Define the *Hellinger distrance* between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ that are absolutely continuous with respect to a third measure μ by

$$h(\mathbb{P}, \mathbb{Q})^2 = \frac{1}{2}\int\left(\sqrt{\frac{d\mathbb{P}}{d\mu}} - \sqrt{\frac{d\mathbb{Q}}{d\mu}}\right)^2 d\mu.$$

In the Hellinger distance, it can be shown [60] that tests with the required error bounds exist as long as the sets being tested are convex. The required tests can then be constructed by piecing together such smaller tests and using a union bound, as long as the number of smaller tests is not too large. This size can be measured via the metric entropy of the parameter space under consideration, which is essentially the complement of a ball about the truth. The *metric entropy* $N(\mathcal{F}, d, \epsilon)$ is the smallest number of ϵ-balls in the metric d required to cover the set $\mathcal{F}$. Much of the general theory of contraction therefore involves verifying the (non-trivial) conditions of Theorem 1.1.4 and has been restricted to the Hellinger distance. We note that if the densities in the model are uniformly bounded, then contraction in the Hellinger distance immediately implies contraction in L^2. We state the following result in the case of i.i.d. density estimation.

Theorem 1.1.4 (Ghosal et al.)**.** *Suppose that $Y_1, ..., Y_n$ are i.i.d. observations arising from a distribution $\mathbb{P}_0$ admitting a Lebesgue density p_0. Let Π be a prior on some set $\mathcal{P}$ of*

probability distributions admitting Lebesgue densities (with distribution $\mathbb{P}$ having density p). Denote by $\mathbb{P}_0^n$ the n-fold product measure for $\mathbb{P}_0$. Suppose that for some sequence $\varepsilon_n \to 0$ with $n\varepsilon_n^2 \to \infty$, there exists some constant $C > 0$ and sets $\mathcal{P}_n \subset \mathcal{P}$ such that

(i) $\log N(\mathcal{P}_n, h, \varepsilon_n) \leq n\varepsilon_n^2$,

(ii) $\Pi(\mathcal{P} \backslash \mathcal{P}_n) \leq \exp(-(C+4)n\varepsilon_n^2)$,

(iii) $\Pi(\mathbb{P} \in \mathcal{P} : -\mathbb{P}_0(\log \frac{p}{p_0}) \leq \varepsilon_n^2, \mathbb{P}_0(\log \frac{p}{p_0})^2 \leq \varepsilon_n^2) \geq \exp(-Cn\varepsilon_n^2)$.

Then for a sufficiently large constant $M > 0$, we have

$$\Pi(\mathbb{P} \in \mathcal{P} : h(\mathbb{P}, \mathbb{P}_0) \geq M\varepsilon_n \mid Y_1, ..., Y_n) \to 0$$

in $\mathbb{P}_0^n$-probability.

It is worth noting that this general testing approach often results in a logarithmic gap with the sharpest contraction rate. This result has been extended to the non-i.i.d. setting in Ghosal and van der Vaart [37], where different tests are constructed in several settings. In the white noise model, likelihood ratios are employed, while alternative tests satisfying the required conditions are also constructed based on the results of Birgé [9, 10] in the case of Markov processes and stationary Gaussian time series. The case of Gaussian process priors is treated in detail in van der Vaart and van Zanten [82] in several settings, such as density estimation, classification and regression. Theorem 2.1 of [82] shows that conditions analogous to those in Theorem 1.1.4 are implied by a single condition on the concentration function of the Gaussian process (see (2.5.6) for the exact definition). The latter results will be important in Chapter 2.

Extending this approach to other distances necessitates alternative constructions of the required tests. In the L^p spaces, Giné and Nickl [41] use the concentration properties of linear centered kernel-type density estimators, derived using empirical process techniques. This replaces the metric entropy condition in Theorem 1.1.4 with an approximation theoretic condition that the sets $\mathcal{P}_n$ are contained in

$$\{f \in \mathcal{F} : ||K_{J_n}(f) - f||_p \leq C(K)\tilde{\varepsilon}_n\},$$

where K_J is a suitable projection operator and $\tilde{\varepsilon}_n$ is the rate of contraction (no longer equal to the ε_n in Theorem 1.1.4). In particular, such linear type estimators can be analyzed in a number of settings and in Chapter 2 we study them in the case of linear inverse problems to obtain an equivalent contraction theorem.

Nonparametric priors typically involve the use of tuning or hyper parameters, whose choice influences the accuracy of the posterior. Procedures that automatically select these parameters in a data-driven manner and achieve optimal performance over multiple parameter classes are termed *adaptive*. Since many qualitative properties of the parameter of

interest are usually unknown, such as its smoothness or regularity, it is of both theoretical and practical importance to study such procedures.

1.1.4 Credibility

The Bayesian analogue of a frequentist confidence set is a credible set. We say that $C_n = C_n(Y^{(n)})$ is a *credible set* for f with *credibility* $1-\gamma$ if

$$\Pi(f \in C_n(Y^{(n)}) \mid Y^{(n)}) \geq 1-\gamma. \tag{1.1.5}$$

From the point of view of posterior-based inference, this is the natural notion of uncertainty quantification. There are of course a wide variety of possible choices of the set C_n that satisfy (1.1.5). It therefore makes sense to select a set that has minimal size in some sense, which can be defined via some geometric notion, in particular by considering a minimal ball in some metric. This topic is investigated in detail in Chapter 3.

The use of credible sets rather than confidence sets can be considered an advantage of the Bayesian approach since Bayesian credible sets can be computed by simulation, whereas in many situations it can be difficult to construct frequentist confidence sets. In particular, the Bayesian generates a number of posterior draws and then keeps a prescribed fraction, discarding the remainder according to some rule. This rule often has a geometric interpretation, such as minimizing some metric, but this is not strictly necessary. From an applied perspective, the practitioner ultimately seeks a practical and effective rule for sorting through posterior draws and such geometric interpretations can be viewed as somewhat artificial in applications. A key question is whether such a method has a theoretical justification and if credible sets are also frequentist confidence sets.

A frequentist theoretical justification for posterior based inference using any (Borel) credible set in finite dimensions is provided by the Bernstein–von Mises theorem (discussed in more detail in Section 1.1.5). In infinite dimensions, such a result does not hold in full generality and the situation is far more subtle. In nonparametric situations there has been as of yet relatively little study into the frequentist coverage of Bayesian credible sets. Early negative results came to alarming conclusions: Cox [27] considered the case of fixed design regression with a Gaussian prior and showed that for almost every parameter from the prior, the coverage of the ℓ_2-credible ball is 0. However there have recently been positive results in the case of Gaussian white noise [53, 61, 79], circumventing the need for a BvM by explicitly studying the coverage properties of certain specific credible sets. This will one of the subjects of study of Chapter 3

1.1.5 Bernstein–von Mises theorems

For parametric models, the asymptotic behaviour of the posterior distribution can be described in detail via the Bernstein–von Mises theorem. This remarkable result establishes

conditions on the prior under which the posterior is approximately a normal distribution centered at an efficient estimator of the true parameter, such as the maximum likelihood estimator, with covariance equal to the Cramér-Rao bound. Surprisingly, this deep result holds under very mild conditions, requiring only that the prior charges a neighbourhood of the true parameter. We explain the principle ideas in the classical setting of density estimation with i.i.d observations so that $Y^{(n)} = (Y_1, ..., Y_n)$ and $\mathbb{P}^n$ denotes the n-fold product measure.

The first results concerning the normal limiting behaviour of a posterior distribution date back to Laplace [58], with later results from Bernstein [5] and von Mises [86]. The result was formalized by Doob [31] and then put into the framework of modern statistics by Le Cam [60]. We follow here the approach presented in van der Vaart [81].

A model $\mathcal{P}$ is *locally asymptotically normal* (LAN) if for every sequence (h_n) in $\mathbb{R}^m$ with $h_n \to h$,

$$\log \frac{d\mathbb{P}^n_{f_0+h_n/\sqrt{n}}}{d\mathbb{P}^n_{f_0}} = h_n^T \Delta_{n,f_0} - \frac{1}{2} h_n^T I_{f_0} h_n + o_{\mathbb{P}^n_{f_0}}(1),$$

where the derivative is the Radon-Nikodym derivative, I_{f_0} is the Fisher information matrix at f_0 and

$$\Delta_{n,f_0} = \frac{1}{\sqrt{n}} \dot{\ell}_{n,f_0}(Y^{(n)})$$

with ℓ_{n,f_0} denoting the log-likelihood function of the model. Since $\mathbb{P}_{f_0} \dot{\ell}_{n,f_0} = 0$ and $-\mathbb{P}_{f_0} \ddot{\ell}_{n,f_0} = \mathbb{P}_{f_0} \dot{\ell}^2_{n,f_0} = n I_{f_0}$, we have by the central limit theorem that Δ_{n,f_0} is asymptotically normal with mean zero and variance I_{f_0}. The LAN condition means that a local expansion of the log-likelihood is of the same form as in the standard Gaussian shift experiment. As a result, this gives the shape of the limiting distribution (and hence of the posterior distribution). We consider a version adapted from Theorem 10.1 of [81].

Theorem 1.1.5 (parametric Bernstein–von Mises)**.** *Suppose that the model $\mathcal{P} = (\mathbb{P}_f : f \in \mathcal{F})$ is locally asymptotically normal at f_0, where $\mathcal{F} \subset \mathbb{R}^m$ is open, $m \geq 1$. Suppose moreover that the Fisher information matrix I_{f_0} is non-singular and that for every $\varepsilon > 0$, there exists a sequence of tests ϕ_n such that*

$$\mathbb{P}^n_{f_0} \phi_n \to 0, \qquad \sup_{||f - f_0|| \geq \varepsilon} \mathbb{P}^n_f (1 - \phi_n) \to 0. \tag{1.1.6}$$

Furthermore, let the prior measure be absolutely continuous in a neighbourhood of f_0 with a continuous positive density at f_0. Then the corresponding posterior distributions satisfy

$$\sup_{B \in \mathcal{T}} \left| \Pi_n(\sqrt{n}(f - f_0) \in B \mid Y^{(n)}) - N\left(\Delta_{n,f_0}, I_{f_0}^{-1}\right)(B) \right| \to^{\mathbb{P}_{f_0}} 0 \quad \text{as } n \to \infty, \tag{1.1.7}$$

where the supremum is over all measurable subsets of $\mathcal{F}$.

Note that if for $g \in \mathbb{R}^m$ we define the mapping $\tau_g : \mathbb{R}^m \to \mathbb{R}^m$ by

$$\tau_g : f \mapsto \sqrt{n}(f - g),$$

then we can then rewrite assertion (1.1.7) using the total variation norm as

$$\left|\left|\Pi(\cdot \mid Y^{(n)}) \circ \tau_{f_0}^{-1} - N(\Delta_{n,f_0}, I_{f_0}^{-1})\right|\right|_{TV} \to^{\mathbb{P}_{f_0}} 0. \tag{1.1.8}$$

Since the Gaussian measure depends only on the model $\mathcal{P}$ and not the prior, this signifies that the effect of the prior distribution is asymptotically negligible. The testing condition (1.1.6) is similar to that required in Theorem 1.1.2 for contraction and in fact exists under the conditions of Theorem 1.1.5, as shown by Lemma 10.3 of [81].

Lemma 1.1.6. *Under the conditions of Theorem 1.1.5, there exists for every $M_n \to \infty$ a sequence of tests ϕ_n and a constant $c > 0$ such that, for every sufficiently large n and every $||f - f_0|| \geq M_n/\sqrt{n}$,*

$$\mathbb{P}_{f_0}^n \phi_n \to 0, \qquad \mathbb{P}_f^n (1 - \phi_n) \leq e^{-cn(||f-f_0||^2 \wedge 1)}.$$

The BvM theorem therefore holds in great generality and justifies the use of posterior-based inference as an efficient frequentist procedure. From a practical perspective, the importance of this result lies in the uniformity achieved over all Borel sets in (1.1.7). This establishes the asymptotic equivalence of credible sets and confidence sets and so justifies the Bayesian approach to uncertainty quantification.

In infinite-dimensions, the situation is more subtle and far less clear. In particular, the full Bernstein–von Mises theorem does not generalize to the nonparametric setting. Several counterexamples of such a result have been studied involving Gaussian priors, notably by Cox [27] and Freedman [33] - see also the related contributions [47, 61]. In particular, Freedman considers the Gaussian white noise model in ℓ_2 with a conjugate Gaussian prior. He shows that the frequentist variance of the ℓ_2 square norm functional $T_n = ||f - \hat{f}||_2^2$, where $\hat{f}$ denotes the posterior mean, is asymptotically smaller than its Bayesian variance. As a consequence, ℓ_2-credible balls $\{f : ||f - \hat{f}||_2^2 \leq R^2\}$ will not have the correct frequentist coverage probabilities.

However, there has been recent progress in investigating Bernstein-von Mises phenomena in the full nonparametric setting. Castillo and Nickl [21, 22] consider the formulation (1.1.8) of the BvM but weaken the notion of convergence, replacing convergence in total variation by weak convergence metrized using the bounded Lipschitz metric β_S (see Chapter 3 for full definitions). In particular, rather than the classical L^p spaces, they consider weaker topologies which admit $1/\sqrt{n}$-consistent estimators and where Gaussian limits are possible.

We say a family of measurable real-valued functions $\mathcal{U}$ defined on a separable metric

space (S, d) is a *μ-uniformity class* for weak convergence if for any sequence μ_n of Borel probability measures on S converging weakly to μ, we have, as $n \to \infty$,

$$\sup_{u \in \mathcal{U}} \left| \int_S u(s)(d\mu_n - d\mu)(s) \right| \to 0.$$

In particular, taking $\mathcal{U} = \{1_A : A \in \mathcal{A}\}$ allows one to establish uniform statements over a class $\mathcal{A}$ of sets. For $A \subset S$ define the δ-enlargement $A^\delta = \{x \in S : d(x, A) < \delta\}$ and δ-boundary $\partial_\delta A = \{x \in S : d(x, A) < \delta, d(x, A^c) < \delta\}$. Restricting to the case of indicator functions, Billingsley and Topsøe [8] show that a family of subsets $\mathcal{A}$ is a μ-uniformly class if and only if

$$\lim_{\delta \to 0} \sup_{A \in \mathcal{A}} \mu(\partial_\delta A) = 0 \tag{1.1.9}$$

(they also establish a similar characterization for functions using moduli of continuity). While using a weak convergence approach loses the uniformity over all Borel sets in (1.1.7), Castillo and Nickl [21, 22] nonetheless establish uniformity statements over certain classes of sets whose geometry is amenable to a Gaussian limit in the sense of (1.1.9). Such results are relevant in Chapter 3.

1.1.6 Computational aspects

While posterior inference is conceptually simple, computing (attributes of) the posterior distribution is central to its practical applicability. In many cases, conjugate models are sufficient and so the posterior can be computed explicitly. However, non-conjugate models are often used in practice and the issue of their computation can be dealt with using Markov chain Monte Carlo (MCMC) techniques.

MCMC methods sample from a probability distribution by constructing a Markov chain with equilibrium distribution equal to the target distribution. Using such a chain, one can sample a draw from an approximation to the posterior without having to invoke an explicit form for it. Repeated sampling can then be used to approximate any desired feature of the posterior, such as the posterior mean or a credible set. Commonly used methods include the Metropolis-Hastings algorithm, which allows one to draw from a probability distribution if one can compute its density function up to its normalizing constant. This is an attractive feature, since calculating the normalizing constant can be difficult in practice. Gibbs sampling exploits the fact that for multivariate distributions it is often easier to sample from the conditional distributions than the joint distribution (which may not be known). The Gibbs sampling algorithm generates an instance for each coordinate of the variable from the target distribution, conditional on the current values of the other variables. This yields a Markov chain whose equilibrium distribution is the desired target. Other variations and methods include Metropolis-within-Gibbs, slice sampling, approximate Bayesian computation and expectation propagation.

1.2 Inverse problems

Nonparametric inverse problem arise in many fields, such as medical imaging (X-ray tomography), astronomy (blurred images of the Hubble Space Telescope), geophysics (reflection seismology), genomics (gene expressions) and mathematical finance (volatility calibration) to name but a few. They arise when the parameter of interest is not directly observable, but rather some transformation of it, possibly with the addition of noise. In many instances this transformation is not (continuously) invertible and we say the problem is *ill-posed*. Such ill-posedness prevents a naive inversion of the observation and usually requires some form of regularization to make sensible inference. In the Bayesian framework, such regularization can be provided by the choice of the prior distribution, which introduces the necessary extra information.

In this thesis we consider a particular class of inverse problems that covers a diverse number of examples, namely linear inverse problems under Gaussian noise. This is related to the infinite-dimensional normal mean model and is an idealized version of many models, such as density estimation or fixed design regression.

1.2.1 General formulation of linear inverse problems

In this thesis, we consider the problem of estimating an unknown parameter f from an observation Y generated from the model

$$Y \equiv Y^{(n)} = Af + \frac{1}{\sqrt{n}}Z. \tag{1.2.1}$$

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 1.2.2 for details).

The Gaussian white noise Z in (1.2.1) 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 [11]), 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, (1.2.1) 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.

In inverse problems it is natural to consider bases $\{e_k\}$ of $\mathbb{H}_1$ that diagonalize A, thus making the problem more tractable by decoupling (1.2.1) into a sequence of independent signal plus noise problems. 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. [74]). 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: $Af = A\left(\sum_k f_k e_k\right) = \sum_k \rho_k f_k g_k$. Writing $Y_k := Y_{g_k}$, (1.2.1) 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).

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. We say that the problem is *mildly ill-posed* with regularity p if

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

for some constants $C_1, C_2 > 0$ and $p \ge 0$. We say that the problem is *severely ill-posed* with regularity γ if

$$C_1(1+k^2)^{-p_0/2}e^{-c_0 k^\gamma} \le |\rho_k| \le C_2(1+k^2)^{-p_1/2}e^{-c_0 k^\gamma} \quad \text{as } k \to \infty$$

for some constants $C_1, C_2, \gamma > 0$ and $p_0, p_1 \in \mathbb{R}$. The polynomial terms p_0 and p_1 are included to add flexibility, but do not characterize the problem since they are dominated by the exponential terms.

1.2.2 Examples

Note that if $\mathbb{H}_1 = \mathbb{H}_2 = L^2([0,1])$ then we can rewrite (1.2.1) in the more classical white noise form

$$dY(t) = (Af)(t)dt + n^{-1/2}dW(t),$$

where W is a standard Brownian motion on $[0,1]$. In particular, taking A to be the identity operator, we recover the classical white noise model in $L^2([0,1])$, which is asymptotically equivalent to fixed design nonparametric regression with Gaussian errors [15] in the sense of Le Cam [60]. Our results apply to the following situations amongst others (see [26] for a general overview of inverse problems).

Deconvolution

A common problem in signal and image processing is periodic deconvolution (see e.g. [48]). Consider the 1-dimensional case on the torus $\mathbb{T} = [0,1)$ and, assuming that f is a 1-periodic function, define

$$Af(t) = \int_0^t f * \mu(s)ds, \qquad t \in [0,1], \tag{1.2.2}$$

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. [48]) 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.

Heat equation

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

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

where $u : [0,1] \times [0,T] \to \mathbb{R}$ and the initial condition $f \in L^2([0,1])$ satisfies $f(0) = f(1) = 0$. 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 with $\gamma = 2$. This problem has been studied in the Bayesian context under conjugate Gaussian priors in Knapik et al. [54] and Agapiou et al. [3].

Radon transform

Another example is given by the Radon transform, which is used in computerized tomography (see [49] 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 (1.2.1). 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 with $p = 1/2$ (see [49] for more details).

Chapter 2

Rates of contraction for non-conjugate priors

In this chapter, we study posterior contraction in linear inverse problems using an abstract testing approach based on approximation-theoretic assumptions on the prior. This chapter is structured as follows: Section 2.1 details the problem and mathematical preliminaries, Section 2.2 contains the general contraction theorem, Section 2.3 contains applications of this result to concrete priors, Sections 2.4 and 2.5 contain proofs of Sections 2.2 and 2.3 respectively and Section 2.6 describes possible extensions of using this approach.

2.1 Introduction

2.1.1 Outline

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

$$Y \equiv Y^{(n)} = Af + \frac{1}{\sqrt{n}}Z. \tag{2.1.1}$$

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 1.2.2 for details).

We wish to study the asymptotic behaviour of the posterior distribution under the frequentist assumption that the data Y is generated from the model (2.1.1) 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 Definition 2. This question has been the object of much study in recent years (see e.g. [36, 41, 46, 76, 82] for some

examples), but the situation of inverse problems has only recently been considered and then only in the conjugate setting [2, 53, 54], 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 [53, 54].

While it is of considerable theoretical interest to understand the behaviour of Bayesian procedures in the non-conjugate setting, there are also strong practical 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 [45]. In many domains, interpretability is a very desirable quality in a model since practitioners usually prefer transparent models to black-box ones. Interpretable models provide meaningful information, particularly when only a few key factors are used. For example, such a concern motivated the use of nonparametric priors in gene expression modelling [55]. Meanwhile, advances in Markov chain sampling methods have meant that such procedures are increasingly tractable in practice (e.g. [66] or Section 1.1.6). 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 [36] 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

$$H_0 : f = f_0 \qquad H_A : f \in \{f : ||f - f_0||_{\mathbb{H}_1} \geq \xi_n\} \tag{2.1.2}$$

with exponentially decaying type-II errors for some sequence $\xi_n \to 0$. We follow the approach of [41] of using the concentration properties of appropriate centred linear estimators to construct suitable plug-in tests. If the operator A in (2.1.1) 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 [64]. 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 [4]. In the mildly ill-posed setting, similar adaptation results are obtained in the recent work of [52] 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 [64] 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 [17, 64] 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 conveniently model Hölder smoothness using uniform random variables. An alternative approach is to consider correctly scaled random Gaussian series [77].

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.

2.1.2 Linear inverse problems

The Gaussian white noise Z in (2.1.1) 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 [11]), 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, (2.1.1) 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.

In this chapter, we henceforth assume the existence of an orthonormal basis $\{e_k\}$ of $\mathbb{H}_1$ consisting of eigenvectors of A^*A, such as the SVD when A is compact. Define the conjugate orthonormal basis $\{g_k\}$ of the range of A in $\mathbb{H}_2$ by the equality $Ae_k = \rho_k g_k$ for some constants $\{\rho_k\}$. For further details see Section 1.2.1. Letting $f_k := \langle f, e_k \rangle_1$, the action of A on f has a simple form when considered in this basis: $Af = A\left(\sum_k f_k e_k\right) = \sum_k \rho_k f_k g_k$. Writing $Y_k := Y_{g_k}$, (2.1.1) 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 1.2.2 for more details), then Condition 1 corresponds to a band-limited basis. Band-limited wavelets have been used in the deconvolution setting (e.g. [48, 69]), 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 2.3.3.

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 ill-posed 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 p if*

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

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

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

$$C_1(1+k^2)^{-p_0/2}e^{-c_0 k^\gamma} \leq |\rho_k| \leq C_2(1+k^2)^{-p_1/2}e^{-c_0 k^\gamma} \qquad \text{as } k \to \infty$$

for some constants $C_1, C_2, \gamma > 0$ and $p_0, p_1 \in \mathbb{R}$.

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.

2.1.3 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 (2.1.1) 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 [28]) shows that for any $f \in \mathbb{H}_1$, this measure is equivalent to $\bigotimes_{k=1}^{\infty} N(0, n^{-1})$ with affinity $\exp\left(-\frac{n}{8}\sum_k \rho_k^2 f_k^2\right) > 0$. 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

$$\frac{d\mathbb{P}_f}{d\mathbb{P}_0} = \exp\left(\sqrt{n}\sum_{k=1}^{\infty}\rho_k f_k Z_k - \frac{n}{2}\sum_{k=1}^{\infty}\rho_k^2 f_k^2\right), \tag{2.1.3}$$

where $Z_k = Z_{g_k}$. If Z were realizable as a Gaussian element in $\mathbb{H}_2$, then this expression would reduce to $\exp\left(\sqrt{n}\langle Af, Z\rangle_2 - \frac{n}{2}||Af||_2^2\right)$. As it is, (2.1.3) makes sense whenever the drift component Af lies in the Cameron-Martin space of Z, that is $||Af||_2 < \infty$. Since under $\mathbb{P}_0$, $Z_k = \sqrt{n}Y_k$, we can express the posterior distribution via Bayes' formula:

$$\Pi(B|Y) = \frac{\int_B e^{n\sum_k \rho_k f_k Y_k - \frac{n}{2}||Af||_2^2} d\Pi(f)}{\int_{\mathcal{P}} e^{n\sum_k \rho_k f_k Y_k - \frac{n}{2}||Af||_2^2} d\Pi(f)}, \qquad B \in \mathcal{B}, \tag{2.1.4}$$

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 [36], a fact that we shall use implicitly in the proof of Theorem 2.2.1.

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 [32]) 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 $\beta > 0$, it is known [26] that the minimax rate of estimation over any fixed ball of H^β is $n^{-\beta/(2p+2\beta+1)}$ under Condition (M) and $(\log n)^{-\beta/\gamma}$ under Condition (S). Minimax rates are attained by a number of methods, such as generalized Tikhonov regularization amongst others [11, 26]. In general, we shall use p and γ to refer to parameters quantifying the ill-posedness of the problem (2.1.1), β to refer to the smoothness of the true function f_0 and α 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 (2.2.5) 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),

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

so that we are actually considering the small-ball probability of f under the weaker negative Sobolev norm H^{-p}, since the dimensions of the ellipsoid correspond to the singular values of A. To establish (2.2.5) in the mildly ill-posed case, it is therefore sufficient to prove

$$\Pi_n(f \in \mathcal{P} : C_2 \, ||f - f_0||_{H^{-p}} \leq \varepsilon_n) \geq e^{-Cn\varepsilon_n^2}. \tag{2.1.5}$$

In fact, the greater the ill-posedness of (2.1.1), 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.

2.2 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 [41] adapted to inverse problems. We quantify the effects of the operator A through a sequence of factors $\{\delta_k\}$. Consider the set of indices

$$A_k = \{l : |\langle \phi_m, e_l \rangle_1| \neq 0 \text{ for some } 1 \leq m \leq k\} \tag{2.2.1}$$

and define

$$\delta_k = \inf_{i \in A_k} |\rho_i|, \tag{2.2.2}$$

that is we take the smallest ρ_i such that one of the first k basis elements $\phi_1, ..., \phi_k$ 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 2.2.1. *Consider the white noise model* (2.1.1) *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* $n \to \infty$*,*

$$k_n \leq cn\varepsilon_n^2 \qquad \text{and} \qquad \frac{\varepsilon_n}{\delta_{k_n}} \leq C_1 \xi_n \tag{2.2.3}$$

for some $c, C_1 > 0$ *and all* $n \geq 1$*, and where* δ_k *is defined by* (2.2.2) *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*

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

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

$$\Pi_n(\mathcal{P}_n^c) \leq e^{-(C+4)n\varepsilon_n^2}, \tag{2.2.4}$$

$$\Pi_n(f \in \mathcal{P} : ||Af - Af_0||_2 \leq \varepsilon_n) \geq e^{-Cn\varepsilon_n^2}. \tag{2.2.5}$$

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

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

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 (2.2.3) represents the variance term of the centred linear estimator used to test (2.1.2), 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 (2.2.3) is an equality (up to constants). Taking $k_n \simeq n\varepsilon_n^2$ gives the optimal result using this method, yielding rate $\xi_n \simeq n^p \varepsilon_n^{2p+1}$.

In the severely ill-posed setting of Condition (S) it is known (see [17] 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 (2.2.3) 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.

2.3 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 $f_0 \in \mathbb{H}_1$.

2.3.1 Sieve priors

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

$$f = \sum_{k=1}^{M} f_k e_k, \tag{2.3.1}$$

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 $\Pi_m(x_1, ..., x_m) = \prod_{k=1}^m \frac{1}{\tau_k} q\left(\frac{x_k}{\tau_k}\right)$. Priors of this form have been studied (e.g. [76, 87]) and, under suitable conditions on h and Π_m, are adaptive over Sobolev smoothness classes in the non ill-posed case [4, 46]. Upon suitable calibration of the prior with respect to A, this adaptation property extends to the ill-posed case when considered over the classes $H^\beta(\mathbb{H}_1)$ for $\beta > 0$. We firstly make the following assumption on q.

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

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

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

Condition 2 is very mild and requires only that q 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 (2.3.1), then in the mildly ill-posed case we recover a $\sqrt{n}$-rate up to a logarithmic factor.

Proposition 2.3.1. *Suppose that A satisfies Condition (M) with regularity p 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 $C > 0$,*

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

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^β, $\beta > 0$. We introduce a parameter $\beta_0 \leq \beta$ 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 $\beta_0 = 0$ is valid in the following theorem and so a non-trivial lower bound is not necessarily assumed.

Proposition 2.3.2. *Suppose that the true function f_0 is in $H^\beta(\mathbb{H}_1)$ for some $\beta > 0$ and that A satisfies Condition (M) with regularity p. 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 $B_3(1 + k^2)^{-\beta_0/2}(\log k)^{-1/w} \leq \tau_k \leq B_4(1 + k^2)^{(p+1)/2}$ for some $B_3, B_4 > 0$ and $\beta_0 \leq \beta$. Then for a sufficiently large constant $C > 0$,*

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

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

We firstly note that this prior gives a fully adaptive convergence rate over all the Sobolev classes H^β up to a logarithmic factor, with this rate being uniform over f_0 in balls in H^β. 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 (2.2.5), and which depend on the operator A and the lower bound β_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 2.3.2 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 $\beta_0 > 0$ 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 ($p = 0$). The conditions on h require it to be of exponential type and are needed both to control the prior mass for the bias condition (2.2.4) and to establish the small-ball condition (2.2.5). They are of the same form as in the direct case (c.f. Condition A_5 of [4]).

When working in the severely ill-posed case, we must calibrate our prior to the degree of ill-posedness (i.e. the parameter γ). 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 2.3.1.

Proposition 2.3.3. *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* $1 - H(m) \leq B\exp(-bm^{\gamma+1})$ *as* $m \to \infty$ *for some constants* $B, b > 0$*. 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* $w_n = (\log n)^{\frac{2p_0+\gamma+1}{2(\gamma+1)}} \exp\left(c(\log n)^{\frac{\gamma}{\gamma+1}}\right)$ *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 (2.2.3) (which is a strict inequality). This reduces the size of the approximating sets $\mathcal{P}_n$ in Theorem 2.2.1, so that we need a sharper control on the tail of the distribution H of M to establish the bias condition (2.2.4). Since we take $k_n \simeq (\log n)^{1/\gamma}$ to account the second part of (2.2.3), we must calibrate H according to the ill-posedness of the problem; indeed the more difficult the problem (larger γ) 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 γ 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 [52, 53, 54] all make strong use of knowledge of A through the choice of diagonalizing basis.

In the general case where $f_0 \in H^\beta$, 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 α in the following proposition corresponds to the Sobolev smoothness of a prior element.

Proposition 2.3.4. *Suppose that the true function f_0 is in $H^\beta(\mathbb{H}_1)$ for some $\beta > 0$ and that A satisfies Condition (S). Suppose that the prior Π satisfies $h(m) \geq B_1 e^{-bm^{\gamma+1}}$ for all $m \geq 1$ and that $1 - H(m) \leq B_2 \exp(-bm^{\gamma+1})$ 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 $\tau_k = (1+k^2)^{-\alpha/2-1/4}$ for some $\alpha > \gamma/2$. Then for a sufficiently large constant $C > 0$,*

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

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 2.3.1 and 2.3.2 satisfy this tail condition corresponding to $\gamma = 0$. The upper bound for $1 - H$ is again needed to control the approximating sets $\mathcal{P}_n$ as in Proposition 2.3.3, while the lower bound on h is need to ensure the prior puts sufficient mass at the truth to verify (2.2.5). In the severely ill-posed case, oversmoothing the prior by a factor of $\gamma/2$ 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 (2.2.3) and (2.2.4), which are more stringent in the dominating bias case.

2.3.2 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 [83] 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

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

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 [53, 54].

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 $\tau_k^2 = (1+k^2)^{-\alpha-1/2}$ and ρ_k satisfying either Condition (M) or (S) as specified.*

The parameter α represents the smoothness of the prior in that $f \in H^s(\mathbb{H}_1)$ for all $s < \alpha$ almost surely. In particular, $\mathbb{E}\,||f||_{H^s}^2 = \sum_{k=1}^{\infty}(1+k^2)^{s-\alpha-1/2} < \infty$ if and only if $s < \alpha$. The mildly ill-posed case is dealt with in [53] using the conjugacy of the prior and we recover the same rates using our testing approach combined with the results of [82]. We firstly obtain the results of Theorem 4.1 of [53] in the case where the prior has no additional scaling (which could be treated similarly).

Proposition 2.3.5. *Suppose that A satisfies Condition (M), that $f_0 \in H^\beta(\mathbb{H}_1)$ for some $\beta > 0$, and assign f the Gaussian prior distribution $N(0, \Lambda)$, where Λ satisfies Condition 3. Then for a sufficiently large constant $C > 0$,*

$$\Pi\left(f \in \mathbb{H}_1 : ||f - f_0||_1 > Cn^{-\frac{\alpha\wedge\beta}{2p+2\alpha+1}} \,\middle|\, Y\right) \to 0$$

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 [36], it should be possible to apply the ideas of [84] 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 thesis. Consider now the severely ill-posed analogue.

Proposition 2.3.6. *Suppose that A satisfies Condition (S), that $f_0 \in H^\beta(\mathbb{H}_1)$ for some $\beta > 0$, and assign f the Gaussian prior distribution $N(0, \Lambda)$, where Λ satisfies Condition 3 for some $\alpha > \gamma/2$. Then for a sufficiently large constant $C > 0$,*

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

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

A gap arises in our rates when the prior undersmooths (i.e. $\beta + \gamma/2 < \alpha$), since in the case of the heat equation ($\gamma = 2$), [54] obtain rate $(\log n)^{-\frac{\alpha\wedge\beta}{2}}$. This gap appears to arise in Lemma 2.5.2 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 [57]. It may be possible to obtain a sharper result through a more careful study of the small ball probability, but at present this lower bound seems difficult to improve and so this gap may be an artefact of our proof.

2.3.3 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 1.2.2 so that A is given by (1.2.2) 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 [41] 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 [65] 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 $f \in L^2([0,1])$:

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

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

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 $\text{supp}(F[\Psi]) \subset \{w : |w| \in [1/3, 4/3]\}$. This implies that the periodized wavelets are themselves band-limited with $\text{supp}(F_{\mathbb{T}}[\psi]) \subset \mathbb{Z} \cap \{w : |w| \in [1/3, 4/3]\}$ (c.f. Theorem 8.31 in [32]), 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 [48, 69] for references).

A convenient notion of smoothness is given by the Besov scale of function spaces.

Definition 3. *Let ϕ, ψ denote the periodized Meyer wavelets described above, and $\bar{\alpha}_{j_0 k}(f) = \int_0^1 \phi_{j_0 k} f$ and $\bar{\beta}_{lk}(f) = \int_0^1 \psi_{lk} f$ denote the wavelet coefficients of $f \in L^p([0,1])$. The Besov space $B^s_{pq}([0,1])$ is defined as the set of functions $\{f \in L^p([0,1]) : ||f||_{s,p,q} < \infty\}$ where*

$$||f||_{s,p,q} = ||\bar{\alpha}_{j_0(\cdot)}||_p + \left(\sum_{l=j_0}^{\infty}\left(2^{l(s+1/2-1/p)}||\bar{\beta}_{l(\cdot)}(f)||_p\right)^q\right)^{1/q}$$

with the obvious modification in the case $q = \infty$.

We note some standard embeddings and identifications. For $C^s([0,1])$ the Hölder(-Zygmund when $s \in \mathbb{N}$) spaces, we have $B^s_{\infty\infty}([0,1]) = C^s([0,1])$, while $B^s_{22}([0,1]) = H^s_2([0,1])$ where $H^s_2([0,1])$ are the standard L^2-Sobolev spaces.

In [41], it is assumed that a quantitative upper bound is known on the C^α-norm of the unknown function. We shall relax this to the case where it is simply known that $||f_0||_{C^\alpha} < \infty$. 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

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

where $u, u_k \sim U(-B, B)$ are i.i.d.. We then have that $U_\alpha \in C^\alpha([0,1]) = B^\alpha_{\infty\infty}([0,1])$ almost surely and in particular $||U_\alpha||_{B^\alpha_{\infty\infty}} \leq B$. Denote the law of U_α given B by $\Pi^{\alpha,B}$ so

that our full prior can be expressed as

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

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

Proposition 2.3.7. *Suppose that A is of the form (1.2.2) and satisfies Condition (M) and that f_0 is periodic and in $C^\beta([0,1])$ for some $\beta > 0$. Suppose that the distribution H satisfies $h(r) \geq e^{-Dr^\nu}$ for all $r \in \mathbb{N}$ and $1 - H(r) \lesssim e^{-Dr^v}$ as $r \to \infty$ for some constants $D > 0$ and $1/\alpha < \nu \leq \infty$. Then there exists a finite constant C such that*

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

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

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

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

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

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

As well as the prior smoothness, the thickness of the tail of H, as measured by ν, affects the rate. When $\alpha < \beta + \frac{1}{\nu}$, we attain the optimal rate of convergence for a $(\alpha - 1/\nu)$-smooth function, that is we lose $1/\nu$ degrees of smoothness. This is entirely due to the bias constraint (2.2.4): the bias of a typical element arising from $\Pi^{\alpha,B}$ is proportional to B, and the approximation errors therefore grow on average with the thickness of the tail of H. 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 note that the above framework includes the case where we take B deterministic, corresponding to $H = \delta_B$ with $\nu = \infty$. 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 $\Pi^{\alpha,H}$.

The assumptions on H mirror those sometimes placed on the prior distribution of the scale parameter in a Dirichlet mixtures of normal distributions [38]. Our results therefore mirror those in Theorem 1 of [38] 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.

2.4 Proof of Theorem 2.2.1

A key step in the proof of Theorem 2.2.1 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 [13]. 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 [48] and [69].

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 [41]. 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 ([62], page 134) gives

$$\begin{aligned} e^{-x^2/2\sigma^2} &\geq \mathbb{P}\left(\sup_{h \in B_0} |G(h)| - \mathbb{E} \sup_{h \in B_0} |G(h)| \geq x \right) \\ &= \mathbb{P}\left(||f_n - \mathbb{E}f_n||_1 - \mathbb{E}\,||f_n - \mathbb{E}f_n||_1 \geq x \right), \end{aligned} \tag{2.4.1}$$

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 (2.2.1) and (2.2.2) 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 $\tilde{\phi}_k$,

$$\left|\left|\sum_{k=1}^{k_n}\langle h,\phi_k\rangle_1\tilde{\phi}_k\right|\right|_2^2 = \left|\left|\sum_{k=1}^{k_n}\langle h,\phi_k\rangle_1\sum_{i\in A_k}\rho_i^{-1}\phi_{k,i}g_i\right|\right|_2^2$$
$$= \sum_{i\in A_{k_n}}\left(\sum_{k=1}^{k_n}\langle h,\phi_k\rangle_1\rho_i^{-1}\langle\phi_k,e_i\rangle_1\right)^2$$
$$\leq \frac{1}{\delta_{k_n}^2}\sum_{i=1}^{\infty}\left|\left\langle\sum_{k=1}^{k_n}\langle h,\phi_k\rangle_1\phi_k,e_i\right\rangle_1\right|^2$$
$$= \frac{1}{\delta_{k_n}^2}\left|\left|\sum_{k=1}^{k_n}\langle h,\phi_k\rangle\phi_k\right|\right|_1^2 \leq \frac{1}{\delta_{k_n}^2}\left|\left|h\right|\right|_1^2.$$

Combining these yields

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

Substituting these bounds into Borell's inequality gives

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

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

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

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

$$\mathbb{P}\left(\left|\left|f_n-\mathbb{E}f_n\right|\right|_1 \geq M\frac{\varepsilon_n}{\delta_{k_n}}\right) \leq e^{-Ln\varepsilon_n^2} \tag{2.4.2}$$

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

Proof of Theorem 2.2.1. By Theorem 2.5.3, it is sufficient to construct tests (indicator functions) $\phi_n = \phi_n(Y; f_0)$ such that

$$\mathbb{E}_{f_0}\phi_n \to 0, \qquad \sup_{f\in\mathcal{P}_n:||f-f_0||_1\geq M\xi_n} \mathbb{E}_f(1-\phi_n) \leq e^{-(C+4)n\varepsilon_n^2}, \tag{2.4.3}$$

where the constant $C > 0$ matches that in (2.2.5). Recall that we are testing the hypotheses (2.1.2).

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 (2.4.2) 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 (2.4.2),

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

since by assumption $\sup_{f\in\mathcal{P}_n} ||f - \mathbb{E}f_n||_1 \leq C_2\xi_n$. This verifies (2.4.3). □

2.5 Other proofs

Before proceeding, we recall some facts that will be used when applying Theorem 2.2.1 to the examples presented in Section 2.3. Recall that both the sieve and Gaussian priors of Sections 2.3.1 and 2.3.2 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 (2.2.2) takes the simple form $\delta_k = \rho_k$.

Establishing contraction results in these cases therefore reduces to verifying the conditions of Theorem 2.2.1: the bias conditions on the prior (2.2.4) and true parameter f_0, the small-ball condition (2.2.5) and balancing the rate (2.2.3). Recall also that in the mildly ill-posed case (Condition (M) with regularity p), it is optimal to balance the terms

in (2.2.3) so that we take resolution level $k_n \simeq n\varepsilon_n^2$ yielding contraction rate $\xi_n \simeq n^p \varepsilon_n^{2p+1}$. In the severely ill-posed case, (2.2.3) is generally a strict inequality, which must be verified in practice.

2.5.1 Proofs of Section 2.3.1 (Sieve priors)

Proof of Proposition 2.3.1. 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 (2.2.5). Let f be a finite series generated from Π, conditionally on $M = m_0$. As noted in Section 2.1.2, since A satisfies Condition (M), it is sufficient to prove (2.1.5) to establish (2.2.5). Therefore,

$$\begin{aligned}
\mathbb{P}\left(||f - f_0||_{H^{-p}} \leq \varepsilon_n\right) &= \mathbb{P}\left(\sum_{k=1}^{m_0} |f_k - f_{0,k}|^2 (1+k^2)^{-p} \leq \varepsilon_n^2\right) \\
&\geq \mathbb{P}\left(|f_k - f_{0,k}|^2 (1+k^2)^{-p} \leq \frac{\varepsilon_n^2}{m_0}, \text{ for } k = 1, ..., m_0\right) \qquad (2.5.1) \\
&= \prod_{k=1}^{m_0} \mathbb{P}\left(|f_k - f_{0,k}| \leq \frac{\varepsilon_n (1+k^2)^{p/2}}{\sqrt{m_0}}\right)
\end{aligned}$$

by the independence of the f_k's.

Note that if q satisfies Condition 2 with $0 \leq w < 1$, then it also satisfies the same condition with $w = 1$ and possible different constants $D', d' > 0$ instead of $D, d > 0$. In what follows, we therefore take $w \geq 1$. 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 $t > 0$,

$$\begin{aligned}
\mathbb{P}\left(|X - z| \leq t\right) &\geq \int_0^t \int_0^{2\pi} D e^{-d|z + re^{i\theta}|^w} dr\, d\theta \\
&\geq 2\pi D \int_0^t e^{-d(|z|+r)^w} dr \geq 2\pi D t e^{-d(|z|+t)^w}.
\end{aligned} \qquad (2.5.2)$$

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 $\alpha_{n,k} = \frac{\varepsilon_n (1+k^2)^{p/2}}{\sqrt{m_0}}$ 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 $\alpha_{n,k} \leq E$ for all $1 \leq k \leq m_0$ and $n \geq 1$. Using (2.5.2),

we lower bound the right-hand side of (2.5.1) by

$$\begin{aligned}
&\prod_{k=1}^{m_0} 2D\frac{\alpha_{n,k}}{\tau_k}e^{-d\tau_k^{-w}(|f_{0,k}|+\alpha_{n,k})^w} \\
&\geq C_1 \exp\left(\sum_{k=1}^{m_0}\log\left(\frac{\alpha_{n,k}}{\tau_k}\right) - d\sum_{k=1}^{m_0}\tau_k^{-w}2^{w-1}\left(|f_{0,k}|^w + \alpha_{n,k}^w\right)\right) \\
&\geq C_2 \exp\left(m_0 \log\varepsilon_n + \sum_{k=1}^{m_0}\log\frac{(1+k^2)^{p/2}}{\tau_k}\right) \\
&\geq C_3 e^{C_4\log\varepsilon_n},
\end{aligned}$$

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

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

for some constant $C_5 > 0$. The choice $\varepsilon_n = \left(\frac{\log n}{n}\right)^{1/2}$ then satisfies (2.2.5).

Consider now the bias constraint (2.2.4). Take k_n to be an integer satisfying $L_1 n\varepsilon_n^2 \leq k_n \leq L_2 n\varepsilon_n^2$ 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 $\Pi(\mathcal{P}_n^c) \leq Ce^{-bk_n} \leq e^{-Ln\varepsilon_n^2}$, 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 (2.2.5) above, we verify (2.2.4). 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 2.2.1 with

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

completes the proof. □

Proof of Proposition 2.3.2. By the triangle inequality

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

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

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

Taking $j_n \simeq \varepsilon_n^{-\frac{1}{p+\beta}}$ gives

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

for some $c' > 0$. Let $\alpha_{n,k} = \frac{\varepsilon_n(1+k^2)^{p/2}}{\sqrt{j_n}}$, and suppose that f is a finite series in the $\{e_k\}$ basis of degree j_n. Then using (2.5.2) as in the proof of Proposition 2.3.1,

$$\begin{aligned}\mathbb{P}\left(||f - P_{j_n}(f_0)||_{H^{-p}} \leq \varepsilon_n\right) &\geq \prod_{k=1}^{j_n} D'\alpha_{n,k}\tau_k^{-1}e^{-d\tau_k^{-w}(|f_{0,k}|+\alpha_{n,k})^w} \\ &\geq \exp\left(j_n \log C_1 + \sum_{k=1}^{j_n}\log\left(\frac{\alpha_{n,k}}{\tau_k}\right)\right. \\ &\qquad \left. -C_2\sum_{k=1}^{j_n}\tau_k^{-w}\left(|f_{0,k}|^w + \alpha_{n,k}^w\right)\right). \end{aligned} \tag{2.5.3}$$

By the hypotheses on $\{\tau_k\}$,

$$\sum_{k=1}^{j_n}\log\left(\tau_k^{-1}(1+k^2)^{p/2}\right) \geq -E_1 j_n \log j_n,$$

for some $E_1 > 0$. Since $f_0 \in H^\beta$, we have $|f_{0,k}| \leq (1+k^2)^{-\beta/2}\,||f_0||_{H^\beta} \leq C(f_0)k^{-\beta}$ for all $k \geq 1$. Moreover, for $k \leq j_n$, note that

$$\alpha_{n,k} \simeq j_n^{-p-\beta-1/2}(1+k^2)^{p/2} \leq E_2 j_n^{-\beta-1/2},$$

for some $E_2 > 0$. Substituting these bounds into (2.5.3) and using that $\tau_k \geq B_3(1+k^2)^{-\beta_0/2}(\log k)^{-1/w}$ yields the lower bound

$$\begin{aligned}&\exp\left(C_3 j_n \log\varepsilon_n - C_4 j_n \log j_n + \sum_{k=1}^{j_n}\log\left(\frac{(1+k^2)^{p/2}}{\tau_k}\right)\right. \\ &\qquad \left. -C_5\sum_{k=1}^{j_n}\tau_k^{-w}(k^{-\beta w} + j_n^{-(\beta+1/2)w})\right) \\ &\geq \exp\left(-C_6 j_n\log j_n - E_1 j_n \log j_n - C_7\sum_{k=1}^{j_n}\log k\right) \geq \exp\left(-C_8 j_n \log j_n\right),\end{aligned}$$

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

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

Condition (2.2.5) is then satisfied by the choice $\varepsilon_n = \left(\frac{\log n}{n}\right)^{\frac{p+\beta}{2p+2\beta+1}}$.

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 (2.2.4). This yields contraction rate

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

Finally, for the true regression element f_0,

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

as required. Applying Theorem 2.2.1 completes the proof. □

Proof of Proposition 2.3.3. By exactly the same reasoning as in the proof of Proposition 2.3.1, (2.2.5) is satisfied with $\varepsilon_n = \sqrt{(\log n)/n}$. Take k_n to be an integer satisfying $(L_1 \log n)^{1/(\gamma+1)} \leq k_n \leq (L_2 \log n)^{1/(\gamma+1)}$ for some constants L_1 and L_2. Again taking $\mathcal{P}_n = \left\{f = \sum_{k=1}^{k_n} f_k e_k\right\}$ yields $\Pi(\mathcal{P}_n^c) \lesssim e^{-bk_n^{\gamma+1}} \leq e^{-Ln\varepsilon_n^2}$ for some constant L that can be made arbitrarily large by increasing L_1. This verifies (2.2.4) 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 2.2.1 with contraction rate

$$\xi_n = \frac{\varepsilon_n}{\delta_{k_n}} \leq C\varepsilon_n(1 + k_n^2)^{p_0/2} e^{c_0 k_n^\gamma} \leq C' \frac{(\log n)^{\frac{1}{2} + \frac{p_0}{\gamma+1}} e^{c_0(L_2 \log n)^{\gamma/(\gamma+1)}}}{\sqrt{n}} = \frac{w_n}{\sqrt{n}}.$$

□

Proof of Proposition 2.3.4. The proof is similar to that of Proposition 2.3.2, 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^{-p} in the mildly ill-posed case:

$$||f||_A^2 := \sum_{k=1}^{\infty} |f_k|^2 (1 + k^2)^{-p_1} e^{-2c_0 k^\gamma}.$$

Taking $j_n^{-(p_1+\beta)} e^{-c_0(j_n+1)^\gamma} \simeq \varepsilon_n$ and using the same truncation argument as in the proof of Proposition 2.3.2 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 $\mathbb{P}\left(||P_{j_n}(f_0) - f||_A \leq c\varepsilon\right)$ by

$$\exp\left(j_n \log C_1 + \sum_{k=1}^{j_n} \log\left(\frac{\tilde{\alpha}_{n,k}}{\tau_k}\right) - C_2 \sum_{k=1}^{j_n} \tau_k^{-2}\left(|f_{0,k}|^2 + \tilde{\alpha}_{n,k}^2\right)\right), \tag{2.5.4}$$

where $\tilde{\alpha}_{n,k} = j_n^{-1/2}\varepsilon_n(1+k^2)^{p_1/2}e^{c_0 k^\gamma} \leq Cj_n^{-\beta-1/2}e^{c_0(k^\gamma-(j_n+1)^\gamma)} \leq Cj_n^{-\beta-1/2}$ for $k \leq j_n$ and by the definition of j_n. Now since $\tau_k = (1+k^2)^{-\frac{\alpha}{2}-\frac{1}{4}}$ and $f_0 \in H^\beta$, we have that

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

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

$$\sum_{k=1}^{j_n} \tau_k^{-2}\tilde{\alpha}_{n,k}^2 \leq Cj_n^{-2(\beta+1/2)} \sum_{k=1}^{j_n} k^{2(\alpha+1/2)} \leq E_2 j_n^{1+2(\alpha-\beta)}$$

for some constants $E_1, E_2 > 0$. Substituting these into (2.5.4) gives the lower bound $\exp\left(-C_3 j_n^{1+\theta}\right)$, where $\theta = \max(\gamma, 2(\alpha-\beta))$. In conclusion, the small ball probability satisfies

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

so that (2.2.5) is satisfied by the choice $\varepsilon_n = (\log n)^{\frac{1+\theta}{2\gamma}} n^{-1/2}$.

Take k_n to be an integer satisfying $(a_1 \log n)^{1/\gamma} \leq k_n \leq (a_2 \log n)^{1/\gamma}$ for some constants a_1 and a_2. For this choice of k_n, (2.2.3) is verified for the choice $\xi_n = (\log n)^{-\frac{\alpha-\theta/2}{\gamma}}$:

$$\frac{\varepsilon_n}{\delta_{k_n}} \leq D\varepsilon_n(1+k_n^2)^{p_0/2}e^{c_0 k_n^\gamma} \leq D'(\log n)^{\frac{2p_0+\gamma+1}{2\gamma}} e^{(c_0 a_2 - 1/2)\log n} = o\left(\xi_n\right)$$

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 (2.2.4) by applying Borell's inequality in a similar fashion to that used in the proof of Theorem 2.2.1. 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 $\mathbb{E}\,||P_{k_n}(f) - f||_1 \leq \sqrt{\sum_{k=k_n+1}^\infty \tau_k^2} \leq k_n^{-\alpha}$. For the variance, note that for any $h \in B_0$,

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

Using these bounds, apply Borell's inequality for the supremum of a Gaussian process as in (2.4.1) with $x = \sqrt{2Ln\varepsilon_n^2 k_n^{-2\alpha-1}}$ to obtain

$$\mathbb{P}\left(||P_{k_n}(f) - f||_1 \geq L'\left(k_n^{-\alpha} + \sqrt{n\varepsilon_n^2}k_n^{-\alpha-1/2}\right)\right) \leq e^{-Ln\varepsilon_n^2}, \tag{2.5.5}$$

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

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

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 $\xi_n = (\log n)^{-\frac{\alpha-\theta/2}{\gamma}}$. 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 (2.2.4). Lastly, as $f_0 \in H^\beta$, then $||P_{k_n}(f_0) - f_0||_1 \leq Ck_n^{-\beta} = O(\xi_n)$ exactly as above. Apply Theorem 2.2.1 to finish. □

2.5.2 Proofs of Section 2.3.2 (Gaussian priors)

The small-ball asymptotics of a Gaussian measure in a Hilbert space have been exactly characterized by Sytaya [78] and using the techniques of large deviations in [29]. 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 [44] or the link with the metric entropy of the unit ball of the RKHS [57] (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 [83] 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

$$\phi_{w_0}(\varepsilon) := \inf_{h\in\mathbb{H}:||h-w_0||<\varepsilon} ||h||_{\mathbb{H}}^2 - \log \mathbb{P}(||W|| < \varepsilon). \tag{2.5.6}$$

By Theorem 2.1 of [82], 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 (2.2.5).

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

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

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

$$\mathbb{H}_{Af} = \left\{ b = \sum_{k=1}^{\infty} b_k g_k \in \mathbb{H}_2 : ||b||_{\mathbb{H}_{Af}}^2 = \sum_{k=1}^{\infty} b_k^2 (1+k^2)^{p_0+\alpha+1/2} e^{2c_0 k^\gamma} < \infty \right\}$$

and the concentration function of Af satisfies

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

as $\varepsilon \to 0$ for some $C = C(p_0, \gamma, \alpha, f_0)$.

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 $b = \sum_{k=1}^{\infty} b_k g_k \in \mathbb{H}_2$,

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

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

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

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

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

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

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 [44]:

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

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

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

as $\varepsilon \to 0$ for some constant $C = C(p, \alpha, B)$. Comparing these two rates, we see that the approximation term dominates when $\beta \leq \alpha$ while the centred small-ball term term dominates when $\beta \geq \alpha$, 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 $j \simeq (\log\frac{1}{\varepsilon})^{1/\gamma}$, then $||h_j - Af_0||_2 \leq \varepsilon$ and

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

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 [57]. Applying Theorem 2 of [57] and using Lemma 2.5.2 below, we get $\phi_0(\varepsilon) \lesssim \left(\log\frac{1}{\varepsilon}\right)^{1+1/\gamma}$. It is also possible to derive this result using a careful rearrangement of the lower bounds proved in [44]. Balancing these terms we have that this quantity dominates when $\alpha \leq \beta + \frac{\gamma}{2}$ and the approximation term dominates otherwise, hence the result. □

Lemma 2.5.2. *Consider the RKHS $\mathbb{H}_{Af}$ of Af under Condition (S) as described in Lemma 2.5.1, and let K_{Af} denote the unit ball of $\mathbb{H}_{Af}$. Then the covering number $N(K_{Af}, ||\cdot||_{\mathbb{H}_2}, \epsilon)$ of K_{Af} with the usual Hilbert space distance satisfies*

$$\log N(K_{Af}, ||\cdot||_{\mathbb{H}_2}, \epsilon) \lesssim \left(\log\frac{1}{\epsilon}\right)^{1+1/\gamma}.$$

Using this result and Theorem 2 of [57], we obtain the bound $-\log\mathbb{P}(||Af||_2 < \varepsilon) \leq$

$C\left(\log\frac{1}{\varepsilon}\right)^{1+1/\gamma}$. This matches the bounds obtained in [20] when considering the general setting of heat kernels ($\gamma = 2$) on manifolds.

Proof. Writing $b = \sum_{k=1}^{\infty} b_k g_k$, we know that for any $b \in K_{Af}$ we have $|b_k| \leq C(1 + k^2)^{-p_0-\alpha-1/2} e^{-c_0 k^\gamma} \leq C e^{-c_0 k^\gamma}$, so that K_{Af} is contained in the infinite rectangle

$$\prod_{k=1}^{\infty} \left[-Ce^{-c_0 k^\gamma}, Ce^{-c_0 k^\gamma}\right].$$

Taking $J = D(\log\frac{1}{\epsilon})^{1/\gamma}$ 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 $X = \prod_{k=1}^{J} \left[-Ce^{-c_0 k^\gamma}, Ce^{-c_0 k^\gamma}\right]$ 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

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

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

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

□

Proof of Proposition 2.3.5. Let us verify the small ball Condition (2.2.5). 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 [82] 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 $\mathbb{P}\left(||Af - Af_0||_2 < 2\varepsilon_n\right) \geq e^{-n\varepsilon_n^2}$. By Lemma 2.5.1, the choice $\varepsilon_n = n^{-\frac{p+\beta\wedge\alpha}{2p+2\alpha+1}}$ satisfies this condition in both the cases $\beta \geq \alpha$ and $\beta \leq \alpha$, thereby verifying (2.2.5).

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 2.3.4 and taking $k_n \simeq n\varepsilon_n^2$ in (2.5.5), we obtain that for $n \geq N$,

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

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 $(n\varepsilon_n^2)^{-\alpha} \leq C\xi_n$ 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 (2.2.4). Finally, since $f_0 \in H^\beta$, we again recover that $||P_{k_n}(f_0) - f_0||_1 \leq Ck_n^{-\beta} ||f_0||_{H^\beta} \simeq (n\varepsilon_n^2)^{-\beta}$, which is smaller than $\xi_n = \varepsilon_n(n\varepsilon_n^2)^p$ for our choice of ε_n. Applying Theorem 2.2.1 completes the proof. □

Proof of Proposition 2.3.6. Consider firstly the case where $\beta + \frac{\gamma}{2} \leq \alpha$. As above, (2.2.5) is verified if $\phi_{Af_0}(\varepsilon_n) \leq n\varepsilon_n^2$. By Lemma 2.5.1, the choice $\varepsilon_n = (\log n)^{\frac{\alpha-\beta+1/2}{\gamma}} n^{-1/2}$ satisfies this condition. Now let k_n be an integer satisfying $(L_1 \log n)^{1/\gamma} \leq k_n \leq (L_2 \log n)^{1/\gamma}$ 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 (2.2.3) then satisfies

$$\frac{\varepsilon_n}{\delta_{k_n}} \leq C\varepsilon_n(1 + k_n^2)^{p_0/2} e^{c_0 k_n^\gamma} \leq C'(\log n)^\eta n^{L_2 c_0 - 1/2} = o\left((\log n)^{-\beta/\gamma}\right)$$

as $n \to \infty$ provided that $L_2 c_0 < 1/2$. To verify (2.2.4), substitute our choices of ε_n and k_n into (2.5.5) to get

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

Since $\alpha \geq \beta + \frac{\gamma}{2}$ the second term is asymptotically larger, so that taking L sufficiently large, we obtain the required exponential inequality (2.2.4) with rate $(\log n)^{-\beta/\gamma}$. Since $f_0 \in H^\beta$, we have that exactly as above $||P_{k_n}(f_0) - f_0||_1 \leq Ck_n^{-\beta} \leq C'(\log n)^{-\beta/\gamma}$, so that we can apply Theorem 2.2.1.

Consider now the case where $\beta + \frac{\gamma}{2} \geq \alpha$. Arguing as above, the choice $\varepsilon_n = (\log n)^{\frac{\gamma+1}{2\gamma}} n^{-1/2}$ satisfies the small-ball condition (2.2.5) and for the bias we recover the exponential inequality

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

By our choice of α, the above rate is larger than the bias of f_0 and so yields the contraction rate. □

2.5.3 Proofs of Section 2.3.3 (Uniform wavelet series)

Since we are working the deconvolution setting described in Section 1.2.2 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 2.2.1. 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 (2.2.2), we therefore have that

under Condition (M),

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

Since the ill-posedness affects the rate ξ_n through (2.2.3), 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 2.3.7. We firstly verify the small-ball condition (2.2.5). Consider the case where $\alpha \leq \beta$. Using the wavelet characterization of the periodic Besov space $B^s_{22}([0,1]) = H^s([0,1])$ for $s \in \mathbb{R}$ gives

$$\begin{aligned} ||h||^2_{H^s} &= \left\{ |\bar{\alpha}(h)| + \left(\sum_{l=0}^{\infty} \left(2^{ls} \left|\left| \bar{\beta}_{l\cdot}(h) \right|\right|_{\ell_2} \right)^2 \right)^{1/2} \right\}^2 \\ &\leq 4 \max \left\{ |\bar{\alpha}(h)|^2, \sum_{l=0}^{\infty} 2^{2ls} \sum_{k=0}^{2^l-1} \bar{\beta}_{lk}(h)^2 \right\}. \end{aligned} \tag{2.5.7}$$

Let $\bar{\alpha}$, $\bar{\beta}_{lk}$ denote the wavelet coefficients of f_0 and note that if $||f_0||_{C^\beta} \leq B$ then $|\bar{\alpha}| \leq B$ and $|\bar{\beta}_{lk}| \leq B2^{-l(\beta+1/2)}$ for all l, k. By (2.5.7), we lower bound $\mathbb{P}(||f_0 - U_\alpha||_{H^{-p}} \leq \varepsilon_n)$ by

$$\begin{aligned} &\mathbb{P}\left(\max\left\{ |\bar{\alpha}-u|^2, \sum_{l=0}^{\infty} 2^{-2lp} \sum_{k=0}^{2^l-1} |\bar{\beta}_{lk} - 2^{-l(\alpha+1/2)} u_{lk}|^2 \right\} \leq c_1\varepsilon_n^2 \right) \\ &= \mathbb{P}\left(|\bar{\alpha}-u|^2 \leq c_1\varepsilon_n^2 \right) \mathbb{P}\left(\sum_{l=0}^{\infty} 2^{-2lp} \sum_{k=0}^{2^l-1} |\bar{\beta}_{lk} - 2^{-l(\alpha+1/2)} u_{lk}|^2 \leq c_1\varepsilon_n^2 \right) \end{aligned} \tag{2.5.8}$$

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 $b_{lk} = 2^{l(\beta+1/2)}\bar{\beta}_{lk}$ and pick $J = J(n)$ as defined

below. The second probability in (2.5.8) becomes

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

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

$$\mathbb{P}\left(|b_{lk} - 2^{-l(\alpha-\beta)} u_{lk}| \le c\varepsilon_n\right) \ge \left(\frac{c\varepsilon_n}{2^{l(\beta-\alpha)+1} B}\right) > 0.$$

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

$$\begin{aligned}
&\mathbb{P}\left(\sum_{l=0}^{J} 2^{-2l(p+\beta)} \sup_{0\le k<2^l} |b_{lk} - 2^{-l(\alpha-\beta)} u_{lk}|^2 \le c_1 \varepsilon_n^2 - c(p,\alpha) B^2 2^{-2J(p+\alpha)}\right) \\
&\ge \mathbb{P}\left(\max_{0\le l\le J} \sup_{0\le k<2^l} |b_{lk} - 2^{-l(\alpha-\beta)} u_{lk}| \le c_4 \varepsilon_n\right) \\
&= \prod_{l=0}^{J} \prod_{k=0}^{2^l-1} \mathbb{P}\left(|b_{lk} - 2^{-l(\alpha-\beta)} u_{lk}| \le c_4 \varepsilon_n\right) \ge \prod_{l=0}^{J} \prod_{k=0}^{2^l-1} \left(\frac{c_4 \varepsilon_n}{2^{l(\beta-\alpha)+1} B}\right) \\
&\ge \exp\left(c_5 \log(\varepsilon_n/B) \sum_{l=0}^{J} 2^l - c_6 \sum_{l=0}^{J} l 2^l\right) \ge e^{c_7 (\varepsilon_n/B)^{-1/(p+\alpha)} \log(\varepsilon_n/B)},
\end{aligned}$$

for $n \ge N(p,\alpha,B,\psi)$ and we have used that $J \simeq -\log(\varepsilon_n/B)$ in the last line. Using (2.5.8) and that $h(B_0) > 0$ for some $B_0 \ge ||f_0||_{C^\beta}$, we have that for $n \ge N(p,\alpha,B_0,\psi)$,

$$\begin{aligned}
\mathbb{P}(||f_0 - U_\alpha||_{H^{-p}} \le \varepsilon_n) &\ge h(B_0) e^{c_3 \log(\varepsilon_n/B_0)} e^{c_7 (\varepsilon_n/B_0)^{-1/(p+\alpha)} \log(\varepsilon_n/B_0)} \\
&\ge e^{c_8 \varepsilon_n^{-1/(p+\alpha)} \log \varepsilon_n},
\end{aligned} \tag{2.5.9}$$

so that (2.2.5) is satisfied by the choice $\varepsilon_n \simeq \left(\frac{\log n}{n}\right)^{\frac{p+\alpha}{2p+2\alpha+1}}$.

Consider now the case $\beta < \alpha \le \beta + \frac{1}{\nu}$, where we can establish (2.2.5) in a similar fashion by using an approximation argument. Recall that $f_0 \in C^\beta([0,1])$ and let h_r be the best H^{-p}-approximation of f_0 such that $||h||_{C^\alpha} \le r$. Write $h_r = \theta_0 \phi + \sum_{l=0}^{\infty} \sum_{k=0}^{2^l-1} \theta_{lk} \psi_{lk}$, where $|\theta| \le r$ and $|\theta_{lk}| \le r 2^{-l(\alpha+1/2)}$, and recall that the wavelet coefficients of f_0 satisfy $|\bar{\beta}_{lk}| \le B_0 2^{-l(\beta+1/2)}$ for $B_0 \ge ||f_0||_{C^\beta}$. Let l_r be the smallest integer such that $2^{l_r(\alpha-\beta)} >$

r/B_0, so that in particular $\theta_{lk} = \bar{\beta}_{lk}$ for all $l < l_r$. Then

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

so that $||f_0 - h_r||_{H^{-p}} \leq C(f_0) r^{-\frac{p+\beta}{\alpha-\beta}}$ by the definition of l_r. Pick r_n to be the smallest integer such that $r_n^{-\frac{p+\beta}{\alpha-\beta}} \leq \frac{\varepsilon_n}{2C}$, so that by the triangle inequality, $\mathbb{P}\left(||f_0 - U_\alpha||_{H^{-p}} \leq \varepsilon_n\right) \geq \mathbb{P}\left(||h_{r_n} - U_\alpha||_{H^{-p}} \leq c\varepsilon_n\right)$ for some $1/2 \leq c < 1$. Since $||h_{r_n}||_{C^\alpha} \leq r_n$, we use (2.5.9) to obtain

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

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

Consider now the bias condition (2.2.4) and, using the notation of wavelets, take $k_n = 2^{J_n} \simeq n\varepsilon_n^2$. Let $\mathcal{B}_r = \text{supp}(\Pi^{\alpha,r})$ denote the $C^\alpha([0,1])$-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 $f \in \mathcal{B}_r$, $\sup_k |\bar{\beta}_{lk}(f)| \leq r2^{-l(\alpha+1/2)}$ for all $l \geq 0$. Consequently,

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

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

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

which is verified with the choice $\varepsilon_n = n^{-\frac{p+\alpha-1/\nu}{2p+2\alpha-2/\nu+1}}$. Comparing this rate to the rates obtained when verifying (2.2.5) we obtain the minimal choices $\varepsilon_n = n^{-\frac{p+\alpha-1/\nu}{2p+2\alpha-2/\nu+1}}$ when $\alpha < \beta + \frac{1}{\nu}$ and $\varepsilon_n = (\log n/n)^{\frac{p+\beta}{2p+2\beta+1}}$ when $\alpha = \beta + \frac{1}{\nu}$. For the true function $f_0 \in C^\beta([0,1])$, using a standard approximation bound gives $||K_{J_n}(f_0) - f_0||_2 \leq C(f_0)2^{-\beta J_n} \simeq (n\varepsilon_n^2)^{-\beta} = O(\xi_n)$ for all the above choices of ε_n. In both cases, apply Theorem 2.2.1 to obtain rate $\xi_n = \varepsilon_n(n\varepsilon_n^2)^p$.

Consider now the stronger tail condition $1 - H(r) \lesssim \exp\left(-e^{Dr^\nu}\right)$ as $r \to \infty$ for some $\nu > 0$. When $\alpha \leq \beta$, (2.2.5) is satisfied as above by the choice $\varepsilon \simeq \left(\frac{\log n}{n}\right)^{\frac{p+\alpha}{p+2\alpha+1}}$. 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 $\Pi(\mathcal{P}_n^c) \lesssim \exp\left(-e^{Dr_n^\nu}\right) \leq e^{-Ln\varepsilon_n^2}$ for some constant

L that can be made arbitrarily large by increasing L_1. Using the above bias calculations, $||K_{J_n}(f) - f||_2 \leq Cr_n 2^{-\alpha J_n} \leq C' r_n (n\varepsilon_n^2)^{-\alpha}$ and so setting this equal to $\xi_n = n^p \varepsilon_n^{2p+1}$ yields that (2.2.4) is satisfied by the choice $\varepsilon_n = (\log n)^{\frac{1/v}{2p+2\alpha+1}} n^{-\frac{p+\alpha}{2p+2\alpha+1}}$. Substituting this expression into that of ξ_n gives the desired contraction rate. □

2.5.4 Abstract contraction results

Following the proof of Theorem 2.1 in [36] with the formula (2.1.4) for the posterior distribution in the inverse setting, we recover an analogous theorem for the sampling model (2.1.1). We include the details for completeness.

Theorem 2.5.3. *Let $\varepsilon_n \to 0$ be a sequence such that $\sqrt{n}\varepsilon_n \to \infty$ as $n \to \infty$. Suppose that the sequence of priors (Π_n) satisfies for some $C > 0$*

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

Suppose moreover that there exists a sequence $\mathcal{P}_n \subset \mathcal{P}$ such that $\Pi_n(\mathcal{P}_n^c) \leq e^{-(C+4)n\varepsilon_n^2}$ and for which there exist tests $\phi_n = \phi_n(Y^{(n)})$ such that

$$\mathbb{E}_{f_0}\phi \to 0, \qquad \sup_{f \in \mathcal{P}_n : ||f - f_0||_{\mathbb{H}_1} \geq M\xi_n} \mathbb{E}_f(1 - \phi_n) \leq Le^{-(C+4)n\varepsilon_n^2}. \tag{2.5.10}$$

Then the posterior distribution $\Pi_n(\cdot|Y)$ contracts about f_0 at rate ξ_n in $||\cdot||_{\mathbb{H}_1}$, where M is a fixed constant.

In this section, we denote by $\mathbb{P}_f$ the law of the model (2.1.1) when the operator A equals the identity (as opposed to $\mathbb{P}_f$ in the rest of Chapter 2, which implicitly considers general A). Let $w^f = \frac{d\mathbb{P}_f}{d\mathbb{P}_0}$ denote the density of $\mathbb{P}_h$ with respect to the law $\mathbb{P}_0$ of the pure white noise process. We can write the log-likelihood ratio as

$$\log \frac{w^h}{w^{h_0}}(Y) = \sqrt{n} \sum_k (h_k - h_{0,k}) Z_k - \frac{n}{2} \left(||h||_{\mathbb{H}_2}^2 - ||h_0||_{\mathbb{H}_2}^2 \right).$$

Under $\mathbb{P}_{h_0}$, the above expression can be rewritten as

$$\begin{aligned} \log \frac{w^h}{w^{h_0}}(Y) &= \sqrt{n} \sum_k (h_k - h_{0,k})(\sqrt{n} h_{0,k} + \tilde{Z}_k) - \frac{n}{2} \left(||h||_{\mathbb{H}_2}^2 - ||h_0||_{\mathbb{H}_2}^2 \right) \\ &= \sqrt{n} \sum_k (h_k - h_{0,k}) \tilde{Z}_k - \frac{n}{2} ||h - h_0||_{\mathbb{H}_2}^2, \end{aligned} \tag{2.5.11}$$

where the $(\tilde{Z}_k)$ are i.i.d. standard normal random variables under $\mathbb{P}_{h_0}$. In particular, taking an expectation under $\mathbb{P}_{h_0}$ of this likelihood ratio yields

$$\mathbb{E}_{h_0} \log \frac{w^h}{w^{h_0}}(Y) = -\frac{n}{2} ||h - h_0||_{\mathbb{H}_2}^2. \tag{2.5.12}$$

Proof of Theorem 2.5.3. By assumption on the tests

$$\mathbb{E}_{f_0}\left[\Pi_n(f \in \mathcal{P} : ||f - f_0||_{\mathbb{H}_1} \geq M\xi_n|Y)\phi_n\right] \leq \mathbb{E}_{f_0}\phi_n \to 0.$$

Thus we need only consider the remaining event $(1 - \phi_n)$, that is

$$\Pi_n(f \in \mathcal{P} : ||f - f_0||_{\mathbb{H}_1} \geq M\xi_n|Y)(1 - \phi_n) = \frac{\int_{\{||f-f_0||_{\mathbb{H}_1} \geq M\xi_n\}} \frac{w^{Af}}{w^{Af_0}}(Y)d\Pi_n(f)(1 - \phi_n)}{\int_{\mathcal{P}} \frac{w^{Af}}{w^{Af_0}}(Y)d\Pi_n(f)}.$$

By Lemma 2.5.4 we have that for all $c > 0$ and probability measures ν with support in

$$B_n = \{f \in \mathcal{P} : ||Af - Af_0||^2_{\mathbb{H}_2} \leq 2\varepsilon_n^2\}$$

one has

$$\mathbb{P}^n_{f_0}\left(\int_B \frac{w^{Af}}{w^{Af_0}}(Y)d\nu(f) \leq e^{-(c+1)n\epsilon^2}\right) \leq \frac{2}{c^2 n\epsilon^2}.$$

Let $c = 1$ and $\nu = \frac{\Pi_n|_{B_n}}{\Pi_n(B_n)}$ and consider the events

$$A_n = \left\{\int_{B_n} \frac{w^{Af}}{w^{Af_0}}(Y)d\Pi_n(f) \geq \Pi_n(B_n)e^{-2n\varepsilon_n^2} \geq e^{-(C+2)n\varepsilon_n^2}\right\}.$$

By Lemma 2.5.4, $\mathbb{P}_{f_0}A_n^c \leq \frac{2}{c^2 n\varepsilon_n^2}$ and so $\mathbb{P}_{f_0}(A_n) \to 1$ as $n \to \infty$. Thus

$$\begin{aligned}
&\mathbb{P}_{f_0}\left(\frac{\int_{\{||f-f_0||_{\mathbb{H}_1} \geq M\xi_n\}} \frac{w^{Af}}{w^{Af_0}}(Y)d\Pi_n(f)(1 - \phi_n)}{\int_{\mathcal{P}} \frac{w^{Af}}{w^{Af_0}}(Y)d\Pi_n(f)} > \epsilon\right) \\
&\leq \mathbb{P}_{f_0}(A_n^c) + \mathbb{P}_{f_0}\left((1 - \phi_n)e^{(C+2)n\varepsilon_n^2} \int_{\{||f-f_0||_{\mathbb{H}_1} \geq M\xi_n\}} \frac{w^{Af}}{w^{Af_0}}(Y)d\Pi_n(f) \geq \epsilon\right).
\end{aligned}$$

Since w^{Af} and w^{Af_0} are densities (Radon-Nikodym derivatives),

$$\mathbb{E}_{f_0}\left[\frac{w^{Af}}{w^{Af_0}}(Y)\right] = 1 \quad \text{and} \quad \mathbb{E}_{f_0}\left[\frac{w^{Af}}{w^{Af_0}}(Y)(1 - \phi_n)\right] = \mathbb{E}_f\left[(1 - \phi_n)\right].$$

Consequently, we have

$$\begin{aligned}
&\mathbb{E}_{f_0}\left[(1 - \phi_n)\int_{\{||f-f_0||_{\mathbb{H}_1} \geq M\xi_n\}} \frac{w^{Af}}{w^{Af_0}}(Y)d\Pi_n(f)\right] \\
&\leq \int_{\mathcal{P}_n^c} \mathbb{E}_{f_0}\left[\frac{w^{Af}}{w^{Af_0}}(Y)\right]d\Pi_n(f) + \sup_{f\in\mathcal{P}_n:||f-f_0||_{\mathbb{H}_1} \geq M\xi_n} \mathbb{E}_{f_0}\left[(1 - \phi_n)\frac{w^{Af}}{w^{Af_0}}(Y)\right] \\
&= \Pi_n(\mathcal{P}_n^c) + \sup_{f\in\mathcal{P}_n:||f-f_0||_{\mathbb{H}_1} \geq M\xi_n} \mathbb{E}_f(1 - \phi_n).
\end{aligned}$$

Combined with Markov's inequality and (2.5.10), we have the result. □

Lemma 2.5.4. *Let $\epsilon > 0$ and ν be a probability measure with support on the set*

$$B = \{f \in \mathcal{P} : ||h - h_0||_{\mathbb{H}_2} \leq 2\epsilon^2\}.$$

Then for every $c > 0$

$$\mathbb{P}_{h_0}\left(\int_B \frac{w^h}{w^{h_0}}(Y)d\nu(h) \leq e^{-(c+1)n\epsilon^2}\right) \leq \frac{2}{c^2 n\epsilon^2}.$$

Proof. By Jensen's inequality we have

$$\log\left(\int \frac{w^h}{w^{h_0}}(Y)d\nu(h)\right) \geq \int \log\left(\frac{w^h}{w^{h_0}}(Y)\right)d\nu(h).$$

We can then bound the probability in question by

$$\begin{aligned}
\mathbb{P}_{h_0}\left(\int_B \frac{w^h}{w^{h_0}}(Y)d\nu(h) \leq e^{-(c+1)n\epsilon^2}\right) &\leq \mathbb{P}_{h_0}\left(\int \log\left(\frac{w^h}{w^{h_0}}(Y)\right)d\nu(h) \leq -(c+1)n\epsilon^2\right) \\
&= \mathbb{P}_{h_0}\left(\int \log\frac{w^h}{w^{h_0}}(Y)d\nu(h) - \mathbb{E}_{g_0}\int \log\frac{w^h}{w^{h_0}}(Y)d\nu(h)\right. \\
&\qquad \left. \leq -n(c+1)\epsilon^2 - \mathbb{E}_{h_0}\int \log\frac{w^h}{w^{h_0}}(Y)d\nu(h)\right).
\end{aligned} \tag{2.5.13}$$

Using Fubini's theorem, the definition of B and the expression (2.5.12) for $\mathbb{E}_{h_0}\frac{w^h}{w^{h_0}}(Y)$ we have

$$-\mathbb{E}_{h_0}\int \log\frac{w^h}{w^{h_0}}(Y)d\nu(h) = \int -\mathbb{E}_{h_0}\log\frac{w^h}{w^{h_0}}(Y)d\nu(h) = \int \frac{n}{2}\,||h - h_0||^2_{\mathbb{H}_2}\,d\nu(h) \leq n\epsilon^2. \tag{2.5.14}$$

Moreover, taking the variance with respect to Y and using (2.5.11)

$$\begin{aligned}
\mathrm{var}_{h_0}\left(\int \log\frac{w^h}{w^{h_0}}(Y)d\nu(h)\right) &= \mathrm{var}_{h_0}\left(\int \sqrt{n}\sum_k (h_k - h_{0,k})\tilde{Z}_k d\nu(h)\right) \\
&\leq n\mathbb{E}_{h_0}\left(\int \sum_k (h_k - h_{0,k})\tilde{Z}_k d\nu(h)\right)^2 \\
&\leq n\int \mathbb{E}_{h_0}\left(\sum_k (h_k - h_{0,k})\tilde{Z}_k\right)^2 d\nu(h) \\
&= n\int ||h - h_0||^2_{\mathbb{H}_2}\,d\nu(h) \leq 2n\epsilon^2.
\end{aligned}$$

Applying the bound (2.5.14) to (2.5.13) and then applying Chebychev's inequality with

the previous display yields

$$\mathbb{P}_{h_0}\left(\int \log \frac{w^h}{w^{h_0}}(Y)d\nu(h) - \mathbb{E}_{h_0}\int \log \frac{w^h}{w^{h_0}}(Y)d\nu(h) \leq -nc\epsilon^2\right)$$
$$\leq \frac{1}{c^2n^2\epsilon^4}\mathrm{var}_{h_0}\left(\int \log \frac{w^h}{w^{h_0}}(Y)d\nu(h)\right) \leq \frac{2}{c^2n\epsilon^2}.$$

□

2.6 Possible extensions

The testing approach based on the concentration properties of estimators that is introduced in this chapter could possibly be extended in a number of directions. A natural question is to what degree Condition 1 can be relaxed with the goal of analzing priors not based around the SVD of A. The following extension we discuss was suggested by Madhuresh Roy. The required exponential inequalities for the tests can be obtained for any basis $\{\phi_k\}$ of $\mathbb{H}_1$ that lies in the "Cameron-Martin" space of the eigenpair $\{\rho_i, e_i\}$, that is for which

$$||\phi_k||_\rho^2 := \sum_{i=1}^{\infty} \rho_i^{-2}|\langle \phi_k, e_i\rangle|^2 < \infty \qquad k = 1, 2, ... \tag{2.6.1}$$

In particular, note that Condition 1 implies (2.6.1). Under Condition 1 the sum in (2.6.1) is finite, which allows one to conveniently separate the ill-posedness in the rate by extracting the factor $1/\delta_k = 1/(\inf_{i\in A_k}|\rho_i|)$ (see (2.2.1) and (2.2.2)) from the sum. Under (2.6.1), the exponential inequality (2.4.2) is replaced by

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

where $\sigma_r^2 = \sum_{k=1}^r ||\phi_k||_\rho^2$, yielding that the rate ξ_n must satisfy $\varepsilon_n\sigma_{k_n} \leq C\xi_n$, where as usual k_n is the degree of the linear estimator. A contraction rate can be computed by studying the size of the quantities σ_{k_n} instead of δ_{k_n}.

Such an approach can be applied to other statistical settings, such as nonparametric regression and deconvolution density estimation, where exponential inequalities are already known for linear estimators. However, in spite of there being conceptually little difference, this latter setting involves significant technical hurdles due to the additional structural constraints involved in modelling densities. In particular, the more complicated prior models require significant work to establish the required small-ball estimates.

There is also the possibility of extending this approach to non-linear inverse problems by constructing analogous tests based on frequentist estimators (or otherwise). In particular, it may be possible to use the more developed frequentist theory for non-linear

inverse problems to extend this study to the Bayesian framework. This will be the object of future study.

Chapter 3

Bernstein–von Mises theorems for adaptive Bayesian nonparametric procedures

In this chapter, we investigate Bernstein–von Mises theorems for adaptive nonparametric procedures in linear inverse problems. We use this general approach to construct optimal frequentist confidence sets based on the posterior distribution and illustrate this method via a numerical study. This chapter is structured as follows: Section 3.1 outlines the general approach, Section 3.2 introduces the required mathematical material, Section 3.3 contains the Bernstein–von Mises results, Section 3.4 contains the results on confidence sets, Section 3.5 provides a numerical study and Sections 3.6 and 3.7 contain proofs.

3.1 Introduction

A key aspect of statistical inference is uncertainty quantification and the Bayesian approach to this problem is to use the posterior distribution to generate a *credible set*, that is a region of prescribed posterior probability (often 95%). This can be considered an advantage of the Bayesian approach since Bayesian credible sets can be computed by simulation. The Bayesian generates a number of posterior draws and then keeps a prescribed fraction, discarding the remainder which are considered "extreme" in some sense. From a frequentist perspective, key questions are whether such a method has a theoretical justification and what is an effective rule for determining which draws to discard. A natural approach is to characterize such draws using a geometric notion, in particular by considering a minimal ball in some metric.

In finite dimensions, the Euclidean distance has a clear interpretation as the natural measure of size. However in infinite dimensions such a notion is less clear-cut: the L^2 metric is the natural generalization of the Euclidean norm, but lacks a clear visual inter-

pretation, while L^∞ can be easily visualized but is more difficult to treat mathematically. From the Bayesian perspective of simulating credible sets, the practitioner ultimately seeks a practical and effective rule for sorting through posterior draws and such geometric interpretations can be viewed as somewhat artificial impositions. The aim of this article is therefore to study possible geometric choices of credible sets that behave well from a frequentist asymptotic perspective.

We study the behaviour of the posterior distribution $\Pi(\cdot \mid Y^{(n)})$ when $Y^{(n)}$ is drawn from the probability distribution $\mathbb{P}_{f_0}$ for some non-random true $f_0 \in \mathcal{F}$ as the data size or quality $n \to \infty$. From such a viewpoint, the theoretical justification for posterior based inference using any (Borel) credible set in finite dimensions is provided by the Bernstein–von Mises (BvM) theorem (see [60, 81]). This deep result establishes mild conditions on the prior under which the posterior is approximately a normal distribution centered at an efficient estimator of the true parameter. It thus provides a powerful tool to study the asymptotic behaviour of Bayesian procedures and justifies the use of Bayesian simulations for uncertainty quantification.

A BvM in infinite-dimensions fails to hold in even very simple cases. Freedman [33] showed that in the basic conjugate ℓ_2 sequence space setting with both Gaussian priors and data, the BvM does not hold for ℓ_2-balls centered at the posterior mean – see also the related contributions [27, 47, 61]. The resulting message is that despite their intuitive interpretation, credible sets based on posterior draws using an ℓ_2-based selection procedure do not behave as in classical parametric models. Recently, Castillo and Nickl [21, 22] have established fully infinite-dimensional BvMs by considering weaker topologies than the classical L^p spaces. Their focus lies on considering spaces which admit $1/\sqrt{n}$-consistent estimators and where Gaussian limits are possible, unlike L^p-type loss. In particular, for the posterior to have a weak limit requires tightness of the limit distribution, taken to be the ℓ_2-Gaussian white noise process, which can only occur in strictly weaker topologies than L^p (see (3.2.4) below). Credible regions selected using these different geometries are shown to behave well, generating asymptotically exact frequentist confidence sets. In this paper, we explore this approach in practice via both theoretical results for *adaptive* priors, as well as numerical simulations.

Before going into more abstract detail, it is useful to consider an example from [22] to numerically illustrate this approach in practice. Suppose that we observe $Y_1, ..., Y_n$ i.i.d. samples from an unknown density f_0 on $[0, 1]$. We take a simple histogram prior Π,

$$f = 2^{L_n} \sum_{k=0}^{2^{L_n}-1} h_k 1_{I_k^{L_n}}, \qquad I_0^{L_n} = [0, 2^{-L_n}], \quad I_{k,L_n} = (k2^{-L_n}, (k+1)2^{-L_n}], \quad k \geq 1,$$

where the h_k are drawn from a $\mathcal{D}(1, ..., 1)$-Dirichlet distribution on the unit simplex in $\mathbb{R}^{2^L}$. Here we ignore adaptation issues and select $L = L_n$ based on the smoothness of the

true function. Consider the standard Haar wavelets

$$\psi_{-1,0} = 1_{[0,1]}, \quad \psi_{lk} = 2^{l/2}\left(1_{\left(\frac{k}{2^l}, \frac{k+1/2}{2^l}\right]} - 1_{\left(\frac{k+1/2}{2^l}, \frac{k+1}{2^l}\right]}\right),$$

where $l \in \{-1, 0, 1, ...\}$ and $k = 0, ..., 2^l - 1$. Letting $w_l = l^{1/2+\epsilon}$ for $\epsilon > 0$ small, consider the multiscale credible ball

$$C_n = \left\{ f : \max_{k,l \leq L_n} w_l^{-1} |\langle f - \hat{f}_n, \psi_{lk} \rangle| \leq R_n n^{-1/2} \right\}, \tag{3.1.1}$$

where $\hat{f}_n$ denotes the posterior mean and $R_n = R(Y_1, ..., Y_n)$ is chosen such that $\Pi(C_n \mid Y_1, ..., Y_n) = 0.95$. By Proposition 1 of [22], $\mathbb{P}_{f_0}(f_0 \in C_n) \to 0.95$ as $n \to \infty$, whereas no such result is available for the L^∞-credible ball. Due to the conjugacy of the Dirichlet distribution with multinomial sampling, the posterior distribution can be computed straightforwardly and R_n can be easily obtained by simulation.

For convenience we take f_0 to be a Laplace distribution with location parameter $1/2$ and scale parameter 5 that is truncated to $[0, 1]$, that is $f_0(x) \propto e^{-5|x-1/2|} 1_{[0,1]}(x)$ with $f_0 \in H_2^s([0, 1))$ for $s < 3/2$. In Figure 3.1, we plotted the true density (solid black) and the posterior mean (red) in the cases $n = 1000, 2000, 5000, 10000$. We generated 100,000 posterior draws and plotted the 95% closest to the posterior mean in the $\mathcal{M}(w)$ sense (grey) to simulate C_n. We also used the posterior draws to generate a 95% credible band in L^∞ by estimating Q_n satisfying $\Pi(f : ||f - \hat{f}_n||_\infty \leq Q_n \mid Y) = 0.95$ and then plotting $\hat{f}_n \pm Q_n$ (dashed black).

We see that the L^∞ diameter of C_n is strictly greater than that of the L^∞-credible band, with this difference particularly marked at the peak of the density. However, the diameter of C_n is spatially heterogeneous and has greatest width at the peak, whilst having smaller width around points where the true density is more regular. In all cases, C_n contains the true f_0, whereas the L^∞ confidence band has more difficulty capturing the peak.

The main message of this numerical example is that simulating the credible set C_n, which uses a slightly different geometry, yields a set that does not look particularly strange in practice and in fact resembles an L^∞ credible band. Both approaches are methodologically similar, the only difference being the rule for discarding posterior draws. From a theoretical point of view, the difference between the two sets is far more significant, with C_n yielding exact coverage statements at the expense of unbounded L^∞ diameter. It is however possible to improve upon the naive implementation of such sets to also obtain the optimal L^∞ diameter (see Proposition 1 of [22] and related results below). Modifying the geometry in such a way to obtain an exact coverage statement therefore comes at little additional cost from a practitioner's perspective.

Nonparametric priors typically involve the use of tuning or hyper parameters, and it

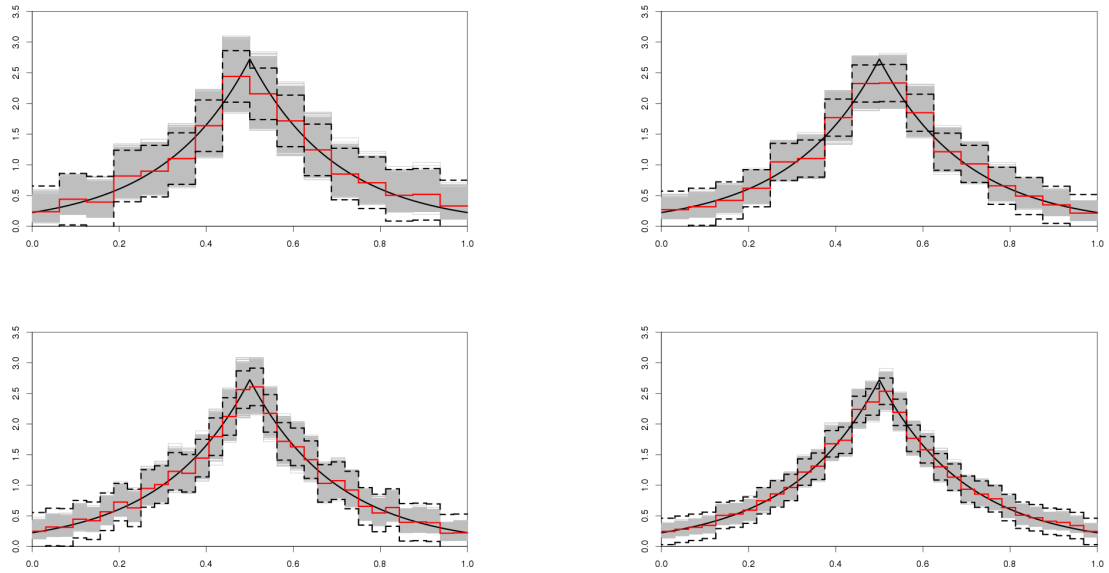

Figure 3.1: *Credible sets based on the Dirichlet prior with the true density function (solid black), the posterior mean (red), a 95% credible band in* L^∞ *(dashed black) and the set* C_n *given in* (3.1.1) *(grey). We have* $n = 1000, 2000, 5000$ *and* 10000 *respectively.*

is a key challenge to study procedures that select these parameters automatically in a data-driven manner. This approach avoids the need to make unreasonably strong prior assumptions on the qualitative properties of the unknown parameter of interest, since incorrect calibration of the prior can lead to suboptimal performance (see e.g. [53]). It therefore makes sense to use an automatic procedure, unless a practitioner is particularly confident that their prior correctly captures the fine details of the unknown parameter, such as its level of smoothness or regularity. Adaptive procedures are in fact widely used in practice, with hyper parameters commonly selected using a hyperprior or an empirical Bayes method. In the case of Gaussian white noise, a number of Bayesian procedures have been shown to be rate adaptive over common smoothness classes. Most such frequentist analyses restrict attention to obtaining contraction rates and do not study coverage properties of credible sets. The focus of this paper is therefore to investigate nonparametric BvMs for adaptive priors, with the goal of studying the coverage properties of credible sets.

In the case of Gaussian white noise, there has been recent work [53, 61] circumventing the need for a BvM by explicitly studying the coverage properties of certain specific credible sets. Of particular relevance is a nice recent paper by Szabó et al. [79], where the authors use an empirical Bayes approach combined with scaling up the radius of ℓ_2-balls to obtain adaptive confidence sets under a so-called *polished tail condition*. Their more hands on approach relies on explicit prior computations and provides an alternative to the more

general abstract point of view taken here. One of our principle goals is exact coverage statements and this seems more difficult to obtain using such an explicit approach. Since adaptive confidence sets do not exist in full generality, we also require self-similarity conditions on the true parameter to exclude certain "difficult" functions [40],[42],[16]. Such conditions ensure that functions look equally smooth at all resolution levels and allow consistent estimation of the unknown regularity, something that is not possible in general. In the absence of such a condition a function may "mislead" an adaptive procedure into thinking a function is overly smooth, leading to a serious failure in the statistical procedure. For example, in Theorem 3.1 of [79] the authors construct a Bayesian credible set that has zero coverage asymptotically. We shall consider the procedure of [79] in Section 3.3.1 and obtain exact coverage statements under the self-similarity condition introduced there.

Another key motivation in studying BvMs is to establish the plug-in property of Bickel and Ritov [7]. In high dimensions, the problem of estimating functionals of the unknown parameter is involved and the nonparametric BvM allows for the simultaneous estimation of a large class of such functionals at $1/\sqrt{n}$-rate. In this case, the results of Castillo and Nickl [21] describe the behaviour of the induced posterior for sufficiently regular (both linear and non-linear) functionals. In particular, this justifies the use of the induced posterior as an efficient procedure with correct uncertainty quantification.

We note other work dealing with BvM results in the nonparametric setting. Leahu [61] has expanded upon the problem of Freedman [33] to study the impact of prior smoothness on the existence of BvM theorems in the conjugate Gaussian sequence space model. Bickel and Kleijn [6], Castillo [18] and Castillo and Rousseau [23] provide sufficient conditions for semiparametric BvMs, while Rivoirard and Rousseau [73] consider linear functionals of probability densities. For the case of finite-dimensional posteriors with increasing dimension, see Ghosal [35] and Bontemps [12] for the case of regression or Boucheron and Gassiat [14] for discrete probability distributions. Much of the approach taken here can equally be applied to other statistical settings such as sparsity, but we restrict to the nonparametric regime for ease of exposition.

3.2 Statistical setting

3.2.1 Function spaces and the white noise model

We use the usual notation $L^p = L^p([0,1])$ for p-times Lebesgue integrable functions and denote by ℓ_p the usual sequence spaces. We consider the canonical white noise model, which is equivalent to the fixed design Gaussian regression model with known variance.

For $f \in L^2 = L^2([0,1])$, consider observing the trajectory

$$dY_t^{(n)} = (Af)(t)dt + \frac{1}{\sqrt{n}}dZ_t, \qquad t \in [0,1], \tag{3.2.1}$$

where dZ is a standard white noise and $A : L^2([0,1]) \mapsto L^2([0,1])$ is a known, injective and continuous linear operator. By considering the action of the orthonormal basis $\{e_\lambda\}_{\lambda\in\Lambda}$ on (3.2.1), it is statistically equivalent to consider the Gaussian sequence space model

$$Y_\lambda^{(n)} \equiv Y_\lambda = \rho_\lambda f_\lambda + \frac{1}{\sqrt{n}}Z_\lambda, \qquad \lambda \in \Lambda, \tag{3.2.2}$$

where the $(Z_\lambda)_{\lambda\in\Lambda}$ are i.i.d. standard normal random variables, the unknown parameter of interest $f = (f_\lambda)_{\lambda\in\Lambda}$ is assumed to be in ℓ_2 (i.e. square summable) and $\{\rho_k\}$ are known constants. We denote by $\mathbb{P}_{f_0}$ or $\mathbb{P}_0$ the law of Y arising from (3.2.2) under the true function f_0. In the following, Λ will cover two principal cases: a Fourier-type basis and a wavelet basis. In the ℓ_2-setting, (3.2.2) can be interpreted purely in sequence form with $\Lambda = \mathbb{N}$ and we do not need to associate to it a time index $t \in [0,1]$ as in (3.2.1). We consider the moderately ill-posed case where

$$C_1 k^{-p} \leq |\rho_k| \leq C_2 k^{-p}, \qquad k = 1, 2, ...,$$

for some $C_1, C_2 > 0$ and $p \geq 0$. The parameter p determines the level of ill-posedness of the problem and quantifies the observed signal to noise ratio. In the case where we do generate the model (3.2.2) by considering the action of an $L^2([0,1])$-basis $\{e_\lambda\}$, we have that $\tilde{f}(t) = \sum_{\lambda\in\Lambda} \rho_\lambda f_\lambda e_\lambda(t)$ in (3.2.1). For a more general overview of inverse problems see Section 1.2 or Cavalier [26].

In our setting, the ℓ_2-Sobolev spaces $\{H_2^s\}_{s\in\mathbb{R}}$ are insufficient to sharply characterize the law of the limiting distribution of the posterior (see discussion after Theorems 3.3.1 and 3.3.2). We therefore consider Sobolev spaces at the logarithmic level. For $s, \delta \geq 0$, define

$$H^{s,\delta} \equiv H_2^{s,\delta} := \left\{ f \in \ell_2 : ||f||_{s,2,\delta}^2 := \sum_{k=1}^{\infty} k^{2s} (\log k)^{-2\delta} |f_k|^2 < \infty \right\}.$$

From this we recover the usual definition of the Sobolev spaces $H^s \equiv H_2^s = H_2^{s,0}$ and by duality we define for $s > 0$, $H_2^{-s} := (H_2^s)^*$. Note that these spaces equal the classical periodic Sobolev spaces if we restrict to the square integrable periodic function $f \in L^2_{per}([0,1))$ and consider $f_k = \langle f, e_k \rangle_2$, where $(e_k(\cdot) = e^{2\pi i k \cdot} : k \in \mathbb{Z})$ is the classical Fourier basis. By standard Hilbert space duality arguments, we can consider ℓ_2 as a subspace of H_2^{-s} and can similarly define the logarithmic spaces for $s < 0$ and $\delta \geq 0$ using the above series definition, yielding spaces satisfying the continuous embeddings $H_2^r \subset H_2^{r,\delta} \subset H_2^s$ for all

$s < r$. In the ℓ_2-setting we shall classify smoothness via the Sobolev *hyper rectangles*

$$\mathcal{Q}(\beta, R) = \left\{ f \in \ell_2 : \sup_{k \geq 1} k^{2\beta+1} f_k^2 \leq R \right\}$$

for $\beta \geq 0$, where $\beta = 0$ corresponds to the ℓ_2-hyper rectangle.

In L^∞ we consider a multiscale approach so that $\Lambda = \{(j,k) : j \geq 0, k = 0, ..., 2^j - 1\}$. In particular, we consider an S-regular ($S \geq 0$) wavelet basis of $L^2([0,1])$:

$$\{\psi_{lk} : l \geq J_0 - 1,\ k = 0, ..., 2^l - 1\}, \qquad J_0 \in \mathbb{N}.$$

For notational simplicity, denote the scaling function ϕ by the first wavelet $\psi_{(J_0-1)0}$. We consider either periodized wavelets or boundary corrected wavelets (see e.g. [65] for more details). Moreover, in certain applications we require in addition that the wavelets satisfy a localization property

$$c(\phi) = \sup_{x \in [0,1]} \sum_k |\phi(x-k)| < \infty, \qquad c(\psi) = \sup_{x \in [0,1]} \sum_k |\psi(x-k)| < \infty. \tag{3.2.3}$$

This property is satisfied by a number of wavelets, for example Meyer wavelets. The sequence model (3.2.2) therefore corresponds to estimating the wavelet coefficients $f_{lk} = \langle f, \psi_{lk} \rangle$, for all $(l,k) \in \Lambda$, since any function $f \in L^2$ generates such a wavelet sequence. Conversely, any such sequence (f_{lk}) generates the wavelet series of a function (or distribution if the sequence is not in ℓ_2) $\sum_{(l,k)} f_{lk} \psi_{lk}$. In the multiscale setting, we shall not consider the ill-posed case, that is we assume that $\rho_\lambda = \rho_{lk} \equiv 1$ for all $\lambda = (l,k) \in \Lambda$. This is mainly due to technical considerations since our method of proof only extends in a straightforward manner to operators that are diagonalized by a wavelet basis. Since this is not the case for the majority of inverse problems of interest, we omit this generalization.

In the $L^\infty([0,1])$-setting we consider multiscale spaces: for a monotone increasing sequence $w = (w_l)_{l \geq 1}$ with $w_l \geq 1$, define

$$\mathcal{M} = \mathcal{M}(w) = \left\{ x = (x_{lk}) : ||x||_{\mathcal{M}(w)} := \sup_{l \geq 0} \frac{1}{w_l} \max_k |x_{lk}| < \infty \right\}$$

(for further references to multiscale statistics see [22]). A separable closed subspace is obtained by considering the restriction

$$\mathcal{M}_0 = \mathcal{M}_0(w) = \left\{ x \in \mathcal{M}(w) : \lim_{l \to \infty} \frac{1}{w_l} \max_k |x_{lk}| = 0 \right\},$$

that is those (weighted) sequences in $\mathcal{M}(w)$ that converge to 0. Note that $\mathcal{M}$ contains the space ℓ_2, since $||x||_{\mathcal{M}} \leq ||x||_{\ell_2}$ as $w_l \geq 1$. In this setting, we consider norm-balls in the

Besov spaces $B^{\beta}_{\infty,\infty}([0,1])$,

$$\mathcal{H}(\beta, R) = \{f = (f_{lk})_{(l,k)\in\Lambda} : |f_{lk}| \leq R2^{-l(\beta+1/2)},\ \forall (l,k) \in \Lambda\}.$$

We recall that $B^{\beta}_{\infty,\infty}([0,1]) = C^{\beta}([0,1])$, the classical Hölder (-Zygmund in the case $\beta \in \mathbb{N}$) spaces. For more details on these embeddings and identifications see [65]. Whether an ℓ_2-white noise defines a tight random element of $\mathcal{M}_0(w)$ depends on the weighting sequence (w_l).

Definition 4. *We call a sequence* $\{w_l\}_{l\geq 1}$ admissible *if* $w_l/\sqrt{l} \nearrow \infty$ *as* $l \to \infty$.

Let $Z = \{Z_\lambda = \langle Z, e_\lambda\rangle : \lambda \in \Lambda\}$, where $Z_\lambda \sim N(0,1)$ i.i.d., denote the Gaussian white noise in (3.2.2). We have from [21, 22] that for $\delta > 1/2$ and $w_l = \sqrt{l}$,

$$\mathbb{E}\,||Z||_{-1/2,2,\delta} < \infty, \qquad \mathbb{E}\,||Z||_{\mathcal{M}(w)} < \infty. \tag{3.2.4}$$

Moreover, for $\delta > 1/2$ and (w_l) an admissible sequence, Z defines a tight Gaussian Borel probability measure on $H_2^{-1/2,\delta}$ and $\mathcal{M}_0(w)$ respectively.

To establish weak convergence of the posterior distribution, we require tightness of the limit distribution by Prokhorov's theorem. Since the law of Z is tight in these spaces, we can consider (3.2.1) as a Gaussian shift model. Denoting by $\mathbb{Z}$ the centered Gaussian Borel random variable in either $H_2^{-1/2,\delta}$ or $\mathcal{M}_0(w)$ with covariance equal to the (ℓ_2-)identity, (3.2.1) can be rewritten as

$$\mathbb{Y}^{(n)} = f + \frac{1}{\sqrt{n}}\mathbb{Z}, \tag{3.2.5}$$

where the above inequality is in the $H_2^{-1/2,\delta}$- or $\mathcal{M}_0(w)$-sense. By (3.2.4) and since

$$\sqrt{n}(\mathbb{Y}^{(n)} - f) = \mathbb{Z} \quad \text{ in } H_2^{-1/2,\delta} \text{ or } \mathcal{M}_0(w),$$

it immediately follows that $\mathbb{Y}^{(n)}$ is an efficient estimator for f in either norm.

Among the two classes $\{H_2^{s,\delta}\}_{s\in\mathbb{R},\delta\geq 0}$ and $\{\mathcal{M}_0(w)\}_w$ of spaces considered, one can show that $s = -1/2$, $\delta > 1/2$ and admissibility of w determine the minimal spaces where the law of the ℓ_2-white noise Z is tight (see [21, 22] for further discussion). We therefore focus attention on these spaces since they provide the threshold for which a weak convergence approach can work. For convenience, we denote $H \equiv H(\delta) \equiv H_2^{-p-1/2,\delta}$. We further denote the law of Z in H or $\mathcal{M}_0(w)$ by $\mathcal{N}$ as appropriate.

3.2.2 Weak Bernstein–von Mises phenomena

To use the notion of weak convergence in defining a nonparametric BvM we need to metrize the weak convergence of probability distributions. For μ and ν probability measures on a

metric space (S, d), we define the bounded Lipschitz metric by

$$\beta_S(\mu, \nu) = \sup_{u:||u||_{BL}\leq 1} \left| \int_S u(s)(d\mu(s) - d\nu(s)) \right|, \tag{3.2.6}$$

$$||u||_{BL} = \sup_{s\in S} |u(s)| + \sup_{s,t\in S: s\neq t} \frac{|u(s) - u(t)|}{d(s,t)}.$$

Thus for random variables taking values in (S, d), $X_n \to^d X$ if and only if $\beta_S(\mathcal{L}(X_n), \mathcal{L}(X)) \to 0$, where $\mathcal{L}(X)$ denotes the law of X. In particular, we shall consider the choices $S = H(\delta) = H_2^{-p-1/2,\delta}$ or $S = H^{-p-s}$ for $s > 1/2$ in ℓ_2 and $S = \mathcal{M}_0(w)$ for $\{w_l\}_{l\geq 1}$ an admissible sequence in L^∞.

Due to the continuous embeddings $\ell_2 \subset H$ and $\ell_2 \subset \mathcal{M}_0(w)$, any Borel probability measure on ℓ_2 yields a tight Borel probability measure on H and $\mathcal{M}_0(w)$. Consider a prior Π on ℓ_2 and let $\Pi_n = \Pi(\cdot \mid Y^{(n)})$ denote the posterior distribution based on data (3.2.2). For $z \in S$ (where here we require in addition that S is a vector space), consider the map $\tau_z : S \to S$ given by

$$\tau_z : f \mapsto \sqrt{n}(f - z).$$

Let $\Pi_n \circ \tau^{-1}_{\mathbb{Y}^{(n)}}$ denote the image measure of the posterior distribution (considered as a measure on H or $\mathcal{M}_0(w)$) under the map $\tau_{\mathbb{Y}^{(n)}}$. Thus for any Borel set B arising from these topologies,

$$\Pi_n \circ \tau^{-1}_{\mathbb{Y}^{(n)}}(B) = \Pi(\sqrt{n}(f - \mathbb{Y}^{(n)}) \in B \mid Y),$$

so that we can more intuitively write $\Pi_n \circ \tau^{-1}_{\mathbb{Y}^{(n)}} = \mathcal{L}(\sqrt{n}(f - \mathbb{Y}^{(n)}) \mid Y^{(n)})$, where $\mathcal{L}(f \mid Y^{(n)})$ denotes the law of f under the posterior. Recalling that we denote by $\mathcal{N}$ the law of the white noise Z in (3.2.2) as an element of S, we define the notion of nonparametric BvM.

Definition 5. *Consider data generated from* (3.2.2) *under a fixed function f_0 and denote by $\mathbb{P}_{f_0}$ the distribution of $Y^{(n)}$. We say that a prior Π satisfies a weak Bernstein-von Mises phenomenon in S if, as $n \to \infty$,*

$$\beta_S(\Pi_n \circ \tau^{-1}_{\mathbb{Y}^{(n)}}, \mathcal{N}) = \beta_S(\mathcal{L}(\sqrt{n}(f - \mathbb{Y}^{(n)}) \mid Y^{(n)}), \mathcal{N}) \to^{\mathbb{P}_{f_0}} 0.$$

Here S is taken to be one of $H(\delta)$ for $\delta > 1/2$, H^{-p-s} for $s > 1/2$ or $\mathcal{M}_0(w)$ for $(w_l)_{l\geq 1}$ an admissible sequence.

The left-hand side consists of the rescaled posterior distribution centered at $\mathbb{Y}$, an efficient estimator of f_0 in S. The weak BvM says that the (scaled and centered) posterior distribution asymptotically looks like an infinite-dimensional Gaussian distribution in some 'weak' sense, quantified by the bounded Lipschitz metric (3.2.6). The lack of a BvM in total variation prevents the user from deducing that $\Pi_n \circ \tau_{\mathbb{Y}}^{-1}$ and $\mathcal{N}$ are asymptotically close, uniformly over all ℓ_2-Borel sets (see e.g. Theorem 2 of [61]). On the other hand,

weak convergence in S implies that these two probability measures are approximately equal on certain classes of sets, whose boundaries behave smoothly with respect to the measure $\mathcal{N}$ (for more discussion see Sections 1.1 and 4.1 of [21]). This allows a uniform control over certain geometric classes of subsets and allows the user to perform useful inference in certain cases discussed below. We also note that by Proposition 4 of [22], a BvM in $\mathcal{M}_0(w)$ implies one in $H(\delta)$, where the parameter δ depends on the growth of the sequence (w_l). Neither notion is strictly more general since the limiting case $w_l = \sqrt{l}$ yields the condition $\delta > 1$ by that Proposition, rather than the threshold $\delta > 1/2$.

It is interesting to note that following this approach requires the geometry of the credible sets to depend explicitly on the level of ill-posedness p of the problem (since we consider an $H_2^{-p-1/2,\delta}$-ball). In the moderately ill-posed setting, only linear functionals whose representors are in $H_2^{p+1/2}$ are estimable at a $1/\sqrt{n}$-rate (see e.g. [53]). In view of this, we see that these spaces are sharp within the classes $\{H_2^s\}_{s\in\mathbb{R}}$ since a weak BvM in a stronger topology would entail the uniform estimation of less regular functionals at a $1/\sqrt{n}$-rate.

The study of adaptive BvM results naturally leads to the topic of adaptive frequentist confidence sets. It is known that confidence sets with radius of optimal order over a class of submodels nested by regularity that also possess honest (i.e. uniform in the parameter f_0) coverage do not exist in full generality (see [42, 68] for recent references). We therefore require additional assumptions on the parameters to be estimated and so consider self-similar functions, whose regularity is similar at both small and large scales. Such conditions have been considered in Giné and Nickl [40], Hoffmann and Nickl [42] and Bull [16] and ensure that we remove those functions whose norms (measuring smoothness) are difficult to estimate and which statistically look smoother than they actually are. The smoothness of such parameters can be accurately estimated and this information can in turn be used to construct adaptive confidence sets, which we do in a Bayesian way. We firstly consider the ℓ_2-type self-similarly assumption found in Szabó et al. [79].

Definition 6. *Fix integer $N_0 \geq 2$ and parameters $\tau \geq 0$, $\rho > 1$. We say that a function $f \in \mathcal{Q}(\beta, R)$ is* self-similar *if*

$$\sum_{k=N}^{\lceil \rho N \rceil} f_k^2 \geq \varepsilon_N R N^{-2\beta} \qquad \text{for all} \quad N \geq N_0,$$

for some sequence $\{\varepsilon_N\} \in (0,1)$ with $\varepsilon_N \geq (\log N)^{-\tau}$. We denote the class of self-similar elements of $\mathcal{Q}(\beta, R)$ by $\mathcal{Q}_{SS}(\beta, R, \varepsilon)$.

The parameters ε_j are permitted to depend on the resolution level j merely to weaken the condition slightly, since logarithmic (or smaller) deviations from the polynomial lower bound have little effect. This condition can be shown to be necessary for the likelihood based procedures considered here, though it is possible to do better using a strictly fre-

quentist approach [68] - see also [67] for further discussion. In L^∞ we consider Condition 3 of Giné and Nickl [40], which can only be slightly relaxed [16].

Definition 7. *Fix positive integer j_0. We say that a function $f \in \mathcal{H}(\beta, R)$ is* self-similar *if there exists a constant $\varepsilon > 0$ such that*

$$||K_j(f) - f||_\infty \geq \varepsilon 2^{-j\beta} \qquad \text{for all} \quad j \geq j_0.$$

We denote the class of self-similar elements of $\mathcal{H}(\beta, R)$ by $\mathcal{H}_{SS}(\beta, R, \varepsilon)$.

In particular, since $f \in \mathcal{H}(\beta, R)$, we have that $||K_j(f) - f||_\infty \asymp 2^{-j\beta}$ for all $j \geq j_0$. What we really require is that there is at least one significant coefficient at the level $(n/\log n)^{1/(2\beta+1)}$ that the posterior distribution can detect. However, this level depends also on unknown constants in practice (see proof of Proposition 3.4.3) and so we require a statement for all (sufficiently large) resolution levels as in Definition 7. See Giné and Nickl [40] and also Bull [16] for further discussion about this condition. These conditions conveniently allow concise and efficient proofs of BvM phenomena, but can possibly be relaxed.

3.3 Bernstein–von Mises Results

3.3.1 Empirical and hierarchical Bayes in ℓ_2

We continue the frequentist analysis of the adaptive priors studied in [52, 79, 80] in ℓ_2. For $\alpha > 0$ define the product prior on the ℓ_2-coordinates by the product measure

$$\Pi_\alpha = \bigotimes_{k=1}^{\infty} N(0, k^{-2\alpha-1}), \tag{3.3.1}$$

so that the coordinates are independent. A draw from this distribution will almost surely (under the prior) be in all Sobolev spaces $H_2^{\alpha'}$ for $\alpha' < \alpha$. We use the notational convention of [79] in that the first coordinate of f_0 is assumed to be 0, since for $k = 1$, the prior does not depend on the smoothness parameter α. Otherwise, this results in some (minor) technical nuisance in establishing a parametric BvM for the projections of the hierarchical prior. As mentioned in [79], this can be circumvented by trivially changing the prior variances to $(k+1)^{-2\alpha-1}$ in (3.3.1), but setting $f_{0,1} = 0$ is notationally simpler.

If $f_0 \in H^\beta$ and $\alpha = \beta$, it has been shown [53] that the posterior contracts at the minimax rate of convergence, while if $\alpha \neq \beta$, then strictly suboptimal rates are achieved. Since the true smoothness β is generally unknown, two data-driven procedures have been considered in [52]. The empirical Bayes procedure consists of selecting the smoothness

parameter by using a likelihood-based approach. Namely, we consider the estimate

$$\hat{\alpha}_n = \underset{\alpha \in [0, \log n / v_n]}{\operatorname{argmax}} \ell_n(\alpha), \tag{3.3.2}$$

where $v_n \to \infty$ is any sequence such that $v_n = O(\log n)$ as $n \to \infty$ and

$$\ell_n(\alpha) = -\frac{1}{2} \sum_{k=1}^{\infty} \left(\log \left(1 + \frac{n}{k^{2\alpha+1} \rho_k^{-2}} \right) - \frac{n^2}{k^{2\alpha+1} \rho_k^{-2} + n} Y_k^2 \right)$$

is the marginal log-likelihood for α in the joint model (f, Y) in the Bayesian setting (relative to the infinite product measure $\otimes_{k=1}^{\infty} N(0,1)$). The case $v_n \asymp \log n$ corresponds to prior knowledge of an upper bound on the smoothness, whereas taking $v_n = o(\log n)$ allows the method to eventually cover the entire range of Sobolev scales. The introduction of v_n is needed to establish a parametric BvM for the finite dimensional projections of the empirical Bayes procedure (see Theorem 3.6.2). The posterior distribution is then defined via the plug-in procedure

$$\Pi_{\hat{\alpha}_n}(\cdot \mid Y) = \Pi_{\alpha}(\cdot \mid Y) \mid_{\alpha = \hat{\alpha}_n}.$$

If there exist multiple maxima to (3.3.2), then any of them can be selected.

A fully Bayesian approach is to put a hyperprior on the parameter α. This yields the hierarchical prior distribution

$$\Pi = \int_0^{\infty} \lambda(\alpha) \Pi_{\alpha} d\alpha,$$

where λ is a positive Lebesgue density on $(0, \infty)$ satisfying the following assumption (Assumption 2.4 of [52]).

Condition 4. *Assume that for every $c_1 > 0$ there exists $c_2 \geq 0, c_3 \in \mathbb{R}$, with $c_3 > 1$ if $c_2 = 0$ and $c_4 > 0$ such that*

$$c_4^{-1} \alpha^{-c_3} \exp(-c_2 \alpha) \leq \lambda(\alpha) \leq c_4 \alpha^{-c_3} \exp(-c_2 \alpha)$$

for $\alpha \geq c_1$.

The exponential, gamma and inverse gamma distributions satisfy Condition 4 for example. Knapik et al. [52] showed that both of these procedures contract to the true parameter adaptively at the (almost) minimax rate, uniformly over Sobolev balls of fixed radius, and the result follows similarly for Sobolev hyper rectangles. In general, it is impossible to estimate the smoothness β of f_0 from the data Y. However, if the true parameter is self-similar in the sense of Definition 6, β can be estimated by either $\hat{\alpha}_n$ or the posterior median of $\lambda(\cdot \mid Y)$ at rate $O_{\mathbb{P}_0}(1/\log n)$ (see Lemmas 3.6.7 and 3.6.8 below). Both procedures satisfy a weak BvM in the sense of Definition 5.

Theorem 3.3.1. *Consider the empirical Bayes procedure described above. For every $\beta, R > 0$ and $s > 1/2$, we have*

$$\sup_{f_0 \in \mathcal{Q}(\beta,R)} \beta_{H^{-p-s}}(\Pi_{\hat{\alpha}_n} \circ \tau_{\mathbb{Y}}^{-1}, \mathcal{N}) \to^{\mathbb{P}_0} 0$$

as $n \to \infty$. Moreover, for $\delta > 2$ we have the (slightly) stronger convergence

$$\sup_{f_0 \in \mathcal{Q}_{SS}(\beta,R,\varepsilon)} \beta_{H(\delta)}(\Pi_{\hat{\alpha}_n} \circ \tau_{\mathbb{Y}}^{-1}, \mathcal{N}) \to^{\mathbb{P}_0} 0$$

as $n \to \infty$.

Theorem 3.3.2. *Consider the hierarchical Bayes procedure described above, where the prior density λ satisfies Condition 4. For every $\beta, R > 0$ and $s > 1/2$, we have*

$$\sup_{f_0 \in \mathcal{Q}(\beta,R)} \beta_{H^{-p-s}}(\Pi_n \circ \tau_{\mathbb{Y}}^{-1}, \mathcal{N}) \to^{\mathbb{P}_0} 0$$

as $n \to \infty$. Moreover, for $\delta > 2$ we have the (slightly) stronger convergence

$$\sup_{f_0 \in \mathcal{Q}_{SS}(\beta,R,\varepsilon)} \beta_{H(\delta)}(\Pi_n \circ \tau_{\mathbb{Y}}^{-1}, \mathcal{N}) \to^{\mathbb{P}_0} 0$$

as $n \to \infty$.

The requirement of self-similarity for a weak BvM in $H(\delta)$ could conceivably be relaxed, but such an assumption is natural since it is anyway needed for the construction of adaptive confidence sets in Section 3.4.1. Weakening the Sobolev exponent $-1/2$ by an arbitrary amount renders this assumption unnecessary, but results in a polynomial suboptimality in the diameter of confidence sets derived using this method. It is not clear whether this is a fundamental limit or a technical artefact of the proof.

Whilst minimax optimality is clearly desirable from a theoretical frequentist perspective, it may be too stringent a goal in our context. Using a purely Bayesian point of view, we derive an analogous result to Doob's almost sure consistency result (Theorem 1.1.1). Specifically, a weak BvM holds in $H(\delta)$ for $\delta > 2$ for prior draws, almost surely under both the empirical Bayes and hierarchical priors. For this, it is sufficient to show that prior draws are self-similar almost surely.

Proposition 3.3.3. *Let $f \sim \Pi_\alpha$, where Π_α is the conditional prior distribution given in* (3.3.1)*. Then, Π_α-almost surely, f is self-similar in the sense of Definition 6 and hence satisfies a weak BvM in $H(\delta)$ for $\delta > 2$. Consequently, if $f \sim \Pi$ is drawn from the full hierarchical prior distribution, then f satisfies a weak BvM in $H(\delta)$ for $\delta > 2$, Π-almost surely.*

In particular, f satisfies Definition 6 with smoothness α and parameters $\tau = 0$, $\rho > 1$ and $\varepsilon_N = \varepsilon(\alpha, \rho, R) > 0$ sufficiently small and random N_0 sufficiently large, Π_α-almost

surely. As a simple corollary to Theorems 3.3.1 and 3.3.2, we have that the rescaled posteriors merge weakly (with respect to weak convergence on $H(\delta)$) in the sense of Diaconis and Freedman [30]. By Proposition 2.1 of [70], we immediately have that the unscaled posteriors merge weakly with respect to the ℓ_2-topology since they are both consistent. However, in the case of bounded Lipschitz functions (rather than the full case of continuous and bounded functions), we can improve this result to obtain a rate of convergence.

Corollary 3.3.4. *For every $\beta, R > 0$, $s > 1/2$ and $\delta > 2$, we have*

$$\sup_{f_0 \in \mathcal{Q}(\beta,R)} \beta_{H^{-p-s}}(\Pi_n \circ \tau_{\mathbb{Y}}^{-1}, \Pi_{\hat{\alpha}_n} \circ \tau_{\mathbb{Y}}^{-1}) \to^{\mathbb{P}_0} 0$$

$$\sup_{f_0 \in \mathcal{Q}_{SS}(\beta,R,\varepsilon)} \beta_{H(\delta)}(\Pi_n \circ \tau_{\mathbb{Y}}^{-1}, \Pi_{\hat{\alpha}_n} \circ \tau_{\mathbb{Y}}^{-1}) \to^{\mathbb{P}_0} 0$$

as $n \to \infty$. In particular, for $S = H^{-p-s}$ or $H(\delta)$ as above,

$$\sup_{u:||u||_{BL} \leq L} \left| \int_S u \, d(\Pi_n - \Pi_{\hat{\alpha}_n}) \right| = O_{\mathbb{P}_0}\left(\frac{L}{\sqrt{n}}\right).$$

This yields a rate of convergence when dealing with functionals that are sufficiently smooth, i.e. those which are continuous with respect to the S-topology. Since convergence rates are a property of the metric rather than the underlying topology and weak convergence is generally a purely topological phenomenon, these rates can not be extended to the full space of continuous and bounded functionals. In particular, since testing a measure against the unit ball of continuous and bounded functions yields the total variation norm, if we could uniformly extend to such functionals then we would obtain strong merging of the measures in the sense of total variation.

3.3.2 Slab and spike prior in L^∞

Consider the slab and spike prior, whose frequentist contraction rate has been analyzed in Castillo and van der Vaart [25], Hoffmann et al. [43] and Castillo et al. [24]. The assumptions in [43] ensure that prior draws are very sparse and only very few coefficients are fitted. We therefore modify the prior slightly so that the prior automatically fits the first few coefficients of the signal without any thresholding. This ensures that the posterior will have a rough approximation of the signal before fitting wavelet coefficients more sparsely at higher resolution levels. This makes sense from a practical point of view by preventing overly sparse models and is in fact necessary from a theoretical perspective (see Proposition 3.3.7).

Let $J_n = \lfloor \log n / \log 2 \rfloor$ be such that $n/2 < 2^{J_n} \leq n$ and define some strictly increasing sequence $j_0 = j_0(n) \to \infty$ such that $j_0(n) < J_n$. For the low resolutions $j \leq j_0$ we fit a simple product prior where we draw the f_{lk}'s independent from a bounded density g such

that

$$g(x) > 0, \qquad \forall x \in \mathbb{R}.$$

For the middle resolution levels $j_0 < j \leq J_n$, the f_{lk}'s are drawn independently from the mixture

$$\Pi_j(dx) = (1 - w_{jn})\delta_0(dx) + w_{jn}g(x)dx, \qquad n^{-K} \leq w_{j,n} \leq 2^{-j(1+\tau)},$$

for some $K > 0$ and $\tau > 1/2$. All coefficients at levels $j > J_n$ are set to 0. Since this is a product prior, it is possible to sample from the posterior distribution using an MCMC scheme on each component separately. We have a weak BvM in the multiscale space $\mathcal{M}_0(w)$, where the rate at which the admissible sequence (w_l) diverges depends on the how many coefficients we automatically fit in the prior via the sequence $j_0(n)$.

Theorem 3.3.5. *Consider the slab and spike prior defined above with lower threshold given by the strictly increasing sequence $j_0(n) \to \infty$. The posterior distribution satisfies a weak BvM in $\mathcal{M}_0(w)$ in the sense of Definition 5, that is*

$$\sup_{f_0 \in \mathcal{H}(\beta,R)} \beta_{\mathcal{M}_0(w)}(\Pi_n \circ \tau_{\mathbb{Y}}^{-1}, \mathcal{N}) \to^{\mathbb{P}_0} 0$$

as $n \to \infty$, for any admissible sequence (w_l) satisfying $w_{j_0(n)}/\sqrt{\log n} \nearrow \infty$.

Note that in the limiting case $w_l = \sqrt{l}$, we recover $j_0(n) \simeq \log n$, so that the prior automatically fits the same fixed fraction of the full coefficients for all n. Since we consider only admissible sequences, we have that compared to the full scale $2^{J_n} \simeq n$ of coefficients, the fraction of coefficients that the prior fits automatically is asymptotically vanishing. An alternative way to consider this result is in reverse: based on a desired rate in applications, we prescribe an admissible sequence $w_l = \sqrt{l}u_l$, where u_l is some divergent sequence, and then pick $j_0(n)$ appropriately. Since the rate $j_0(n)$ is obtained via an implicit relation, we include a specific case here for clarity.

Corollary 3.3.6. *Consider the slab and spike prior defined above with lower threshold*

$$j_0 = j_0(n) \simeq (\log n)^{\frac{1}{2\epsilon+1}},$$

for some $\epsilon > 0$. Then it satisfies a weak BvM in $\mathcal{M}_0(w)$ in the sense of Definition 5, that is,

$$\sup_{f_0 \in \mathcal{H}(\beta,R)} \beta_{\mathcal{M}_0(w)}(\Pi_n \circ \tau_{\mathbb{Y}}^{-1}, \mathcal{N}) \to^{\mathbb{P}_0} 0$$

as $n \to \infty$ for the admissible sequence $w_l = l^{1/2+\epsilon}u_l$, where u_l is any (arbitrarily slowly) diverging sequence.

Note that for $\epsilon > 0$, $2^{j_0(n)} = 2^{c(\log n)^{1/(2\epsilon+1)}}$ grows more slowly than any power of n.

While the requirement to fit the first few coefficients of the prior is very mild and of practical use in nonparametrics, it is naturally of interest to study the behaviour of the posterior distribution with full thresholding, that is when $j_0(n) \equiv 0$, which we denote by Π'. In general however, the full posterior contracts to the truth at a rate strictly slower than $1/\sqrt{n}$ in $\mathcal{M}(w)$, so that a $\sqrt{n}$-rescaling of the posterior can not converge weakly to a limit. This holds even for self-similar functions, which can be see from the proof of the following proposition.

Proposition 3.3.7. *Let (w_l) be any admissible sequence. Then for any $\beta > 0$ and $R > 0$, there exists $f_0 \in \mathcal{H}(\beta, R)$ such that along some subsequence (n_m),*

$$\mathbb{E}_0 \Pi'(||f - \mathbb{Y}||_{\mathcal{M}(w)} \geq M_{n_m} {n_m}^{-1/2} \mid Y^{(n_m)}) \to 1$$

for all $M_n \to \infty$ sufficiently slowly. Consequently, for such an f_0, a weak BvM in $\mathcal{M}_0(w)$ in the sense of Definition 5 can not hold.

On the level of a $\sqrt{n}$-rescaling as in Definition 5, the rescaled posterior distribution asymptotically puts vanishingly small probability mass on any given $\mathcal{M}(w)$-ball infinitely often and there is therefore no hope that it can look like the required mixture of Gaussian distribution and Dirac mass at zero. This occurs because the posterior selects non-zero coordinates by thresholding at the level $\sqrt{\log n/n}$ rather than the required $1/\sqrt{n}$ (Lemma 1 of [43]). The weighting sequence (w_l) acts to regularize the extra $\sqrt{\log n}$ factor at high frequencies, but it remains present at low frequencies. This is the reason that the weighting sequence (w_l) depends explicitly on the tresholding factor $\sqrt{\log n}$ in Theorem 3.3.5.

It seems that using such an adaptive scheme on low frequencies of the signal causes the weak BvM to fail. This prior closely resembles the frequentist practice of wavelet thresholding, where such a phenomenon has also been observed. For example, Giné and Nickl [39] require similar (though stronger) assumptions on the number of coefficients that need to be fitted automatically to obtain a central limit theorem for the distribution function of the hard thesholding wavelet estimator in density estimation (Theorem 8 of [39]).

3.4 Applications

3.4.1 Adaptive credible sets

We propose credible sets from the hierarchical or empirical Bayes procedures, which we show are adaptive frequentist confidence sets for self-similar parameters. We consider the natural Bayesian approach of using the quantiles of the posterior distribution to obtain a credible set of prescribed posterior probability. By considering sets whose geometry is amenable to the space $H(\delta)$, the weak BvM implies that such credible sets are asymptotically confidence sets.

For a given significance level $0 < \gamma < 1$, consider the credible set

$$C_n = \left\{ f : ||f - Y||_H \leq R_n/\sqrt{n} \right\}, \tag{3.4.1}$$

where $R_n = R_n(Y, \gamma)$ is chosen such that $\Pi_{\hat{\alpha}_n}(C_n|Y) = 1 - \gamma$ or $\Pi(C_n|Y) = 1 - \gamma$. Since the empirical and hierarchical Bayes procedures both satisfy a weak BvM, we have from Theorem 1 of [21] that in both cases

$$\mathbb{P}_{f_0}(f_0 \in C_n) \to 1 - \gamma \qquad \textit{and} \qquad R_n = O_p(1)$$

as $n \to \infty$, so that C_n is asymptotically an exact frequentist confidence set (of unbounded ℓ_2-diameter). We control the diameter of the set using either the estimator $\hat{\alpha}_n$ or the posterior median as a smoothness estimate, and then use the standard frequentist approach of undersmoothing. In the first case, consider

$$\tilde{C}_n = \left\{ f : ||f||_{H^{\hat{\alpha}_n - \epsilon_n}} \leq M_n, \quad ||f - Y||_H \leq R_n/\sqrt{n} \right\}, \tag{3.4.2}$$

where $M_n \to \infty$ grows more slowly than any polynomial in n, R_n is chosen as in C_n and $0 < \epsilon_n < \hat{\alpha}_n$ (chosen possibly data dependently) is such that $\epsilon_n \to 0$ and $\epsilon_n = O(1/\log n)$. Geometrically, $\tilde{C}_n$ is the intersection of two ℓ_2-ellipsoids, C_n and an $H^{\hat{\alpha}_n - \epsilon_n}$-norm ball. For a typical element f in $\tilde{C}_n$, the size of the low frequency coordinates of f are determined by C_n, while the smoothness condition in $\tilde{C}_n$ acts to regularize the elements of C_n (which are typically not in ℓ_2) by shrinking the higher frequencies.

Proposition 3.4.1. *For any $f_0 \in \mathcal{Q}_{SS}(\beta, R, \varepsilon)$, where $R \geq 1$ and $\beta \in (0, \beta_{max}]$, the confidence set $\tilde{C}_n$ given in* (3.4.2) *satisfies*

$$\mathbb{P}_{f_0}(f_0 \in \tilde{C}_n) \to 1 - \gamma \qquad \textit{and} \qquad \Pi(\tilde{C}_n|Y) = 1 - \gamma + o_{\mathbb{P}_0}(1)$$

as $n \to \infty$. If $M_n \to \infty$ grows more slowly than any power of $\log n$ (e.g. $M_n \asymp \log\log n$), then the ℓ_2-diameter of $\tilde{C}_n$ satisfies

$$|\tilde{C}_n|_2 = O_{\mathbb{P}_0}\left(n^{-\frac{\beta}{2\beta+2p+1}} (\log n)^{\frac{2\delta\beta}{2\beta+2p+1}} M_n^{\frac{2p+1}{2\beta+2p+1}} \right).$$

If $M_n \asymp (\log n)^{\zeta}$ for some $\zeta > 0$, then the ℓ_2-diameter of $\tilde{C}_n$ increases to

$$|\tilde{C}_n|_2 = O_{\mathbb{P}_0}\left(n^{-\frac{\beta}{2\beta+2p+1}} (\log n)^{\frac{2\delta\beta+\zeta(2p+1)}{2\beta+2p+1}} \right).$$

The above result is uniform over $\mathcal{Q}_{SS}(\beta, R, \varepsilon)$ as can be seen from the proof. The extra power of M_n in the diameter is similar to the penalties that commonly arise in frequentist procedures due to undersmoothing (see for example [40]) and in particular, M_n can be

taken to increase arbitrarily slowly. On the other hand, the logarithmic correction in the definition of $H(\delta)$ that is required for a weak BvM imposes a strict suboptimality on the diameter of $\tilde{C}_n$. Since the function $x \mapsto x/(x+c)$ is strictly increasing for any $c > 0$, this suboptimality is of order $(\log n)^{2\delta}$, uniformly over $\beta \geq 0$. This is the price required for using a plug-in approach in $H(\delta)$. The extra undersmoothing due to ϵ_n is necessary to obtain the posterior credibility statement, but ϵ_n can be set to 0 if the frequentist coverage statement of Proposition 3.4.1 is sufficient.

Replacing the estimate $\hat{\alpha}_n$ with the median α_n^M of the marginal posterior distribution $\lambda_n(\cdot|Y)$ yields a fully Bayesian analogue. To obtain the necessary undersmoothing over a target range $(0, \beta_{max}]$, we consider the shifted estimator $\hat{\beta}_n = \hat{\alpha}_n - (C+1)/\log n$, where $C = C(R, \beta_{max}, \varepsilon, \rho) = \max_{0<\beta\leq\beta_{max}} C(R, \beta, \varepsilon, \rho)$ is the constant appearing in Lemma 3.6.8 (which can be explicitly computed). Consider

$$\tilde{C}_n' = \left\{ f : ||f||_{H^{\hat{\beta}_n}} \leq M_n, \quad ||f - Y||_H \leq R_n/\sqrt{n} \right\}, \tag{3.4.3}$$

where $M_n \to \infty$ grows more slowly than any polynomial in n and R_n is chosen as in C_n. Taking C_n arising from the hierarchical Bayesian procedure, $\tilde{C}_n'$ is therefore a fully Bayesian object. Using the same approach as above, we have an analogue of Proposition 3.4.1.

Proposition 3.4.2. *For any $f_0 \in \mathcal{Q}_{SS}(\beta, R, \varepsilon)$, where $R \geq 1$ and $\beta \in (0, \beta_{max}]$, the confidence set $\tilde{C}_n'$ given in* (3.4.3) *satisfies*

$$\mathbb{P}_{f_0}(f_0 \in \tilde{C}_n') \to 1 - \gamma, \qquad \text{and} \qquad \Pi(\tilde{C}_n'|Y) = 1 - \gamma + o_{\mathbb{P}_0}(1)$$

as $n \to \infty$. If $M_n \to \infty$ grows more slowly than any power of $\log n$ (e.g. $M_n \asymp \log\log n$), then the ℓ_2-diameter of $\tilde{C}_n'$ satisfies

$$|\tilde{C}_n'|_2 = O_{P_0}\left(n^{-\frac{\beta}{2\beta+2p+1}} (\log n)^{\frac{2\delta\beta}{2\beta+2p+1}} M_n^{\frac{2p+1}{2\beta+2p+1}} \right).$$

If $M_n \asymp (\log n)^{\zeta}$ for some $\zeta > 0$, then the ℓ_2-diameter of $\tilde{C}_n'$ increases to

$$|\tilde{C}_n'|_2 = O_{\mathbb{P}_0}\left(n^{-\frac{\beta}{2\beta+2p+1}} (\log n)^{\frac{2\delta\beta+\zeta(2p+1)}{2\beta+2p+1}} \right).$$

3.4.2 Adaptive confidence bands

We provide a fully Bayesian construction of adaptive confidence bands using the slab and spike prior. The posterior median $\tilde{f} = (\tilde{f}_{n,lk})_{(l,k)\in\Lambda}$ (defined coordinate-wise) takes the form of a thresholding estimator (c.f. [1]), which we use to identify significant coefficients. This has the advantage of both simplicity and interpretability and also provides a natural Bayesian approach for this coefficient selection. Such an approach was used by Kueh [56]

to construct an asymptotically honest adaptive frequentist confidence set on the sphere using needlets. In that article, the coefficients are selected based on the empirical wavelet coefficients with the thresholds selected conservatively using Bernstein's inequality. In contrast, we use a Bayesian approach to automatically select the thresholding quantile constants that then yields exact coverage statements.

Let $B_{\mathcal{M}(w)}(g,R) = \{f : ||f-g||_{\mathcal{M}(w)} \leq R\}$ denote the ball of radius R and centre g in the space $\mathcal{M}(w)$. We firstly select the radius $R_n = R_n(Y,\gamma)$ such that $\Pi(B_{\mathcal{M}(w)}(Y, R_n/\sqrt{n})|Y) = 1-\gamma$, that is the $\mathcal{M}(w)$-ball centered at Y with posterior credibility $1-\gamma$. We then define the data driven width of our confidence band

$$\sigma_{n,\gamma} = \sigma_{n,\gamma}(Y) = \sup_{x\in[0,1]} \sum_{l=0}^{J_n} \frac{v_n}{\sqrt{n}} \sum_{k=0}^{2^l-1} 1_{\{\tilde{f}_{lk}\neq 0\}} |\psi_{lk}(x)|, \tag{3.4.4}$$

where (v_n) is any sequence such that $v_n \to \infty$. If ones wishes to scale different frequencies by varying amounts, one can replace (v_n) by $(v_{n,l})_{n,l\geq 0}$ (possibly random), where $v_{n,l_n} \to \infty$ for any subsequence $l_n \to \infty$. For example, a posterior choice based on the multiscale approach might be $v_{n,l} = R_n(Y,\gamma)w_l$, since $R_n = O_{\mathbb{P}_0}(1)$ [22]. Under a local self-similarity type condition as in Kueh [56], one could possibly remove the supremum in (3.4.4) to obtain a spatially adaptive procedure. However, we restrict attention to more global self-similarity conditions here for simplicity Since we consider wavelets satisfying (3.2.3), we immediately have

$$\sigma_{n,\gamma} \leq \frac{v_n}{\sqrt{n}} \sup_{x\in[0,1]} \sum_{l=0}^{J_n} \sum_{k=0}^{2^l-1} |\psi_{lk}(x)| \leq C(\psi)\frac{v_n}{\sqrt{n}} \sum_{l=0}^{J_n} 2^{l/2} \leq C' v_n < \infty \quad a.s.,$$

for all n and $\gamma \in (0,1)$. Letting π_{med} denote the projection onto the non-zero coordinates of the posterior median, we consider the set

$$D_n = \{f : ||f-Y||_{\mathcal{M}(w)} \leq R_n/\sqrt{n}, \quad ||f - \pi_{med}(Y)||_\infty \leq \sigma_{n,\gamma}(Y)\}. \tag{3.4.5}$$

This involves a two-stage procedure: we firstly calculate the required $\mathcal{M}(w)$-radius R_n and then use the posterior median to select the coefficients deemed significant.

Proposition 3.4.3. *Let $f_0 \in \mathcal{H}_{SS}(\beta, R, \varepsilon)$, where $R \geq 1$ and $\beta \in [\beta_{min}, \beta_{max}]$ for $0 < \beta_{min} \leq \beta_{max} < \infty$. Consider the slab and spike prior defined above with threshold $j_0(n) \to \infty$ and let (w_l) be any admissible sequence that satisfies $w_{j_0(n)}/\sqrt{\log n} \nearrow \infty$. Then the confidence set D_n given in* (3.4.5)*, using the choice (w_l) and $\sigma_{n,\gamma}(Y)$ defined in* (3.4.4) *for $v_n \to \infty$, satisfies*

$$\mathbb{P}_{f_0}(f_0 \in D_n) \to 1-\gamma$$

as $n \to \infty$. Moreover, the L^∞-diameter of D_n satisfies

$$|D_n|_\infty = O_{\mathbb{P}_0}\left(\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}} v_n\right).$$

Under self-similarity, D_n has radius equal to the minimax rate in L^∞ up to some factor v_n that can be taken to diverge arbitrarily slowly, again mirroring a frequentist undersmoothing penalty. The choice of the posterior median is for simplicity and can be replaced by any other suitable thresholding procedure, for example directly using the posterior mixing probabilities between the atom at zero and the continuous density component:

$$\tilde{\sigma}_{n,\gamma}(Y) = \sup_{x\in[0,1]} \sum_{l=0}^{J_n} \sum_k \frac{v_n}{\sqrt{n}} 1\{\Pi(f_{lk} = 0 \mid Y) \leq 1/2\} \left|\psi_{lk}(x)\right|.$$

An alternative to a fully Bayesian procedure would be to consider an empirical Bayes approach such as in Section 3.3.1. For example, one could use a Lepski type bandwidth choice (e.g. [40]) to estimate the truncation level $J_n(\beta)$, and hence the true smoothness β (again under Definition 7), thereby allowing an optimal truncation similar to that in (3.4.4).

3.5 Simulation examples

We now apply our approach in a numerical example. Since the key idea is to briefly illustrate this geometric approach rather than perform an extensive simulation study, we restrict to the conditionally conjugate case for simplicity. Following on from the example of the $\mathcal{M}(w)$-based credible set (3.1.1), we now consider the space $H_2^{-1/2,\delta}$.

Consider the Fourier sine basis

$$e_k(x) = \sqrt{2}\sin(k\pi x), \qquad k = 1, 2, ...,$$

and define the true function $f_{0,k} = \langle f_0, e_k\rangle_2 = k^{-3/2}\sin(k)$ so that the true smoothness is $\beta = 1$. We consider realisations of the data (3.2.2) at levels $n = 500$, 1000 and 2000 and use the empirical Bayes posterior distribution. We plotted the true f_0 (black), the posterior mean (red) and an approximation to the credible sets (grey). To simulate ℓ_2 credible balls, we sampled 2000 curves from the posterior distribution and kept the 95% closest in the ℓ_2 sense to the posterior mean and plotted them (grey). While the true ℓ_2-ball is unbounded in L^∞, this gives some visual idea of the posterior spread. We performed the same approach to obtain the full $H(\delta)$-credible set C_n given in (3.4.1) and then plotted the full adaptive confidence set $\tilde{C}_n$ given in (3.4.2) with $M_n = \log\log n$ and $\epsilon_n = 1/\log n$. We also present the approximate credibility of $\tilde{C}_n$ by considering the fraction of the simulated curves from the posterior that satisfy the extra constraint of $\tilde{C}_n$ that $||f||_{H^{\hat{\alpha}_n - \epsilon_n}} \leq M_n$.

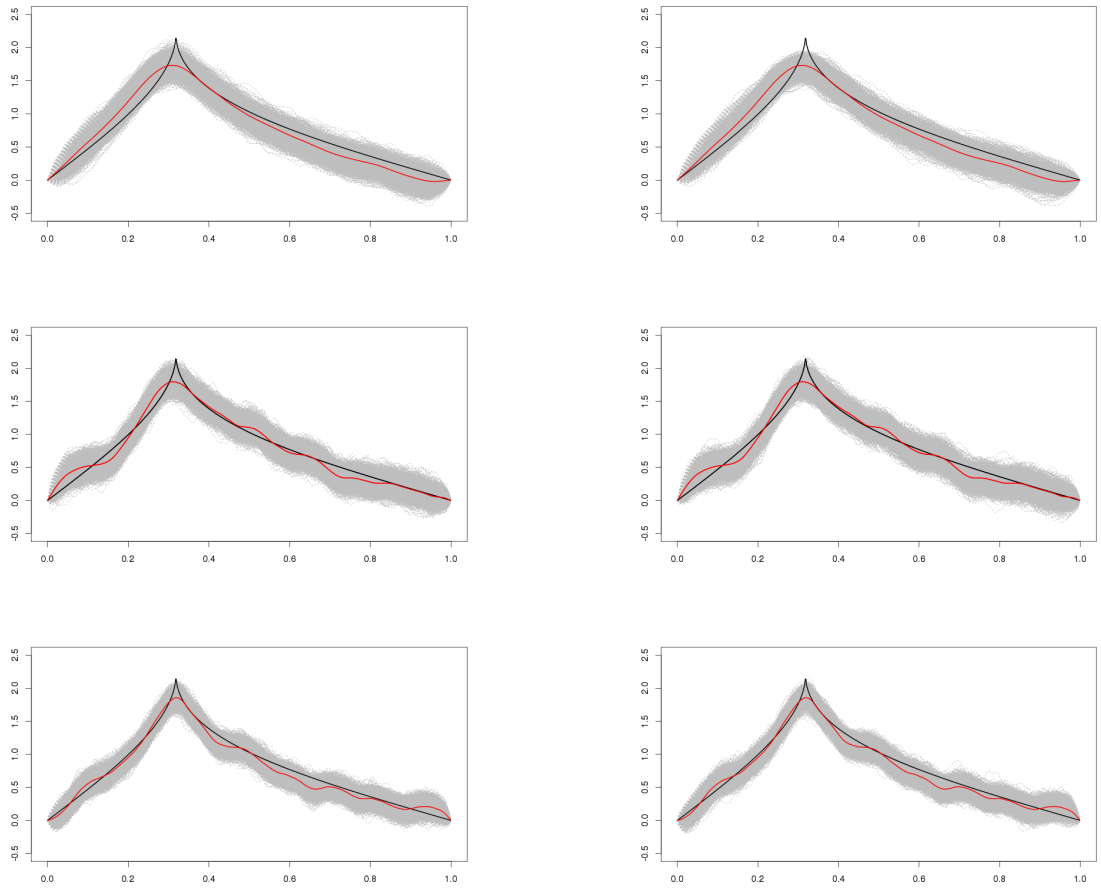

Figure 3.2: *Empirical Bayes credible sets for the Fourier sine basis with the true curve (black) and the empirical Bayes posterior mean (red). The left panels contain the ℓ_2 credible ball and the right panels contain the set $\tilde{C}_n$ given in* (3.4.2)*. From top to bottom, $n = 500, 1000, 2000$, with the right-hand side having credibility 73%, 95%, 95% respectively.*

This is given in Figure 3.2.

For a given set of 2000 posterior draws, we also computed the credibility of $\tilde{C}_n$ at a chosen significance level and the credibility of the posterior draws jointly discarded by both methods. We repeated this 10 times and the average values are presented in Figure 3.3.

The posterior distribution appears to have some difficulty visually capturing the resulting function at its peak. In fact the credible sets do "cover the true function", but do so in an ℓ_2 rather than an L^∞-sense. Indeed, any ℓ_2-type confidence ball will be unresponsive to highly localized pointwise features since they occur on a set of small Lebesgue measure (as in this case). Similar reasoning also explains the poor performance of the posterior mean at this point. The posterior mean estimates the Fourier coefficients of f_0 and hence estimates the true function via its Fourier series. The discrepancy at this point is thus

	n=1000				n=2000			
Chosen significance	95	90	85	80	95	90	85	80
Cred. of $\tilde{C}_n$	92.92	87.15	81.45	76.40	93.63	89.29	84.49	78.49
Cred. of joint rejections	0.28	1.26	2.90	4.73	0.30	1.06	2.31	4.23

Figure 3.3: *Table showing the average credibility of $\tilde{C}_n$ and the average credibility of the posterior draws that are jointly discarded by both methods (all as percentages).*

due to the poor pointwise convergence properties of Fourier series.

In Figure 3.2 we see very little difference visually between the ℓ_2 and $H(\delta)$-credible balls. However, the number of posterior draws that are jointly discarded by both methods is low (the last line in Figure 3.3) and so the two approaches do actually use different rejection criteria in practice. For example, for $n = 2000$ and significance level 95%, only 6 $(= 0.30 \times 2000/100)$ of the 2000 curves were jointly rejected by both methods on average, indicating almost completely different selection outcomes. The visual similarity in Figure 3.2 is therefore a result of the posterior draws themselves looking similar, rather than the methods performing identically.

We note that the credibility of $\tilde{C}_n$ is strictly less than the ℓ_2 credible ball due to the additional smoothness constraint in $\tilde{C}_n$, but that this difference is small by $n = 2000$. The posterior distribution already strongly regularizes the high frequencies so that the posterior draws are very regular with high probability. This can be quantitatively seen by the rapidly decaying variance term of the posterior distribution (3.6.10). This is indeed the case in the simulation, where the credibility gap is small, thereby demonstrating that most of the posterior draws already satisfy the smoothness constraint in $\tilde{C}_n$.

Finally, we repeat the same simulation using the same true function $f_{0,k} = k^{-3/2}\sin(k)$, but with basis equal to the singular value decomposition (SVD) of the Volterra operator (c.f. [53]):

$$e_k(x) = \sqrt{2}\cos((k-1/2)\pi x), \qquad k = 1, 2, ...$$

and plot this in Figure 3.4 for $n = 1000$. Unlike Figure 3.2, the resulting function has no "spike" and so both credible sets have no trouble capturing the true function.

3.6 Proofs

3.6.1 Proofs of weak BvM results in ℓ_2 (Theorems 3.3.1 and 3.3.2)

To prove a weak BvM we need to show that the posterior contracts at rate $1/\sqrt{n}$ to the truth in the relevant space and that the finite-dimensional projections of the rescaled posterior converge weakly to those of the normal law $\mathcal{N}$, which are simply standard Gaussian random variables. The latter condition is implied by a classical parametric BvM in total variation.

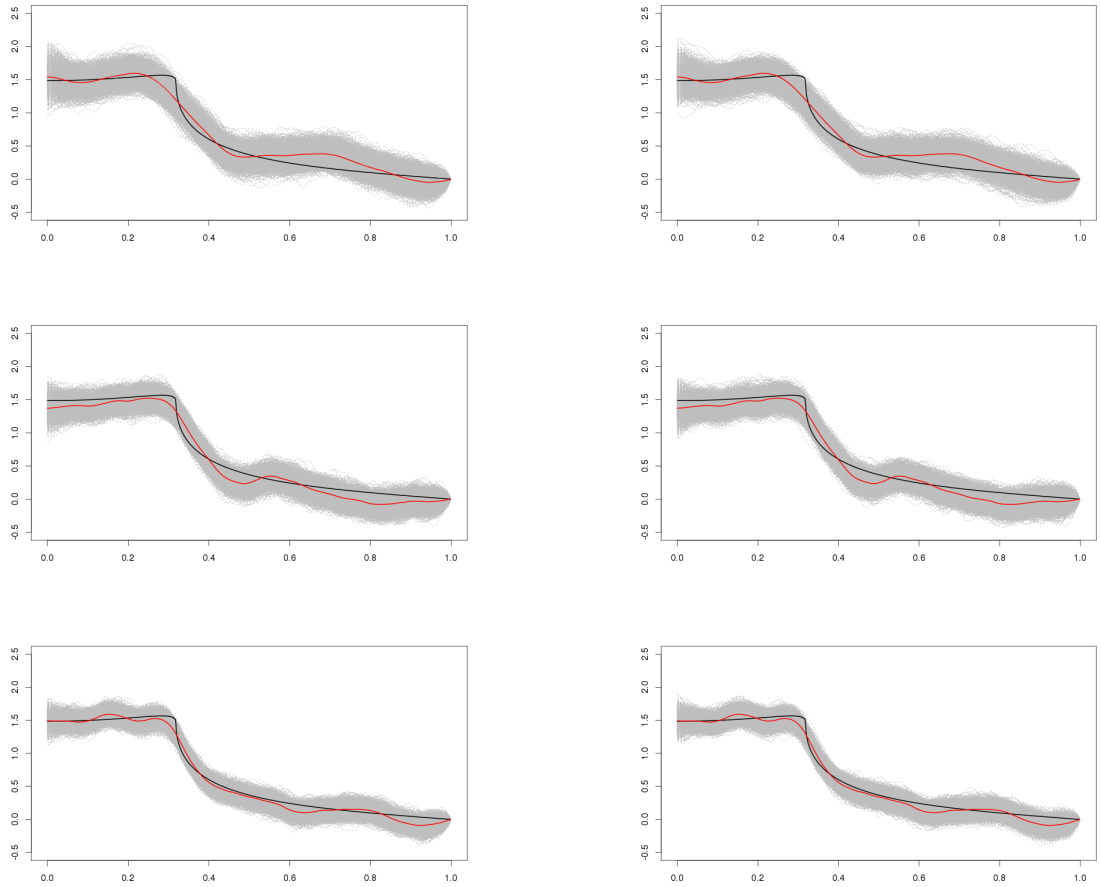

Figure 3.4: *Empirical Bayes credible sets for the Volterra SVD basis with the true curve (black) and the empirical Bayes posterior mean (red). The left panels contain the ℓ_2 credible ball and the right panels contain the set $\tilde{C}_n$ given in* (3.4.2)*. From top to bottom, $n = 500, 1000, 2000$, with the right-hand side having credibility 86%, 94%, 95% respectively.*

Theorem 3.6.1. *For every $\beta, R > 0$ and $M_n \to \infty$, we have*

$$\sup_{f_0 \in \mathcal{Q}(\beta,R)} \mathbb{E}_0 \Pi_{\hat{\alpha}_n}(f : ||f - f_0||_S \geq M_n L_n n^{-1/2} | Y) \to 0,$$

where $S = H(\delta)$ or H^{-p-s} for $s > 1/2$. If $S = H(\delta)$ then $L_n = (\log n)^{3/2}(\log\log n)^{1/2}$; if in addition $f_0 \in \mathcal{Q}(\beta, R, \varepsilon)$, then the rate improves to $L_n = 1$ for $\delta \geq 2$. If $S = H^{-p-s}$ for $s > 1/2$, then $L_n = 1$.

Proof. This contraction result is proved in the same manner as Theorem 2.3 in [52], with suitable modifications for the different norms used. The proof is presented in Section 3.7 for completeness. □

A classical parametric BvM for the projections of the empirical Bayes posterior, whose proof we delay to Section 3.6.3, can be obtained by modifying the classical arguments of Le Cam [60].

Theorem 3.6.2. *The finite dimensional projections of the empirical Bayes procedure satisfy a parametric BvM, that is for every finite dimensional subspace $V \subset \ell_2$,*

$$\sup_{f_0 \in \mathcal{Q}(\beta,R)} ||\Pi_{\hat{\alpha}_n}(\cdot|Y) \circ T_{\mathbb{Y}}^{-1} - N_V(0, I)||_{TV} \to^{\mathbb{P}_0} 0,$$

where π_V denotes the projection onto V and $T_z : f \mapsto \sqrt{n}\pi_V(f - z)$.

Proof of Theorem 3.3.1. Fix $\eta > 0$, let S denote H^{-p-s} or $H(\delta)$ as appropriate and set $\tilde{\Pi}_{\hat{\alpha}_n} = \Pi_{\hat{\alpha}_n} \circ \tau_{\mathbb{Y}}^{-1}$. By the triangle inequality, uniformly over the relevant class of functions,

$$\beta_S(\tilde{\Pi}_{\hat{\alpha}_n}, \mathcal{N}) \leq \beta_S(\tilde{\Pi}_{\hat{\alpha}_n}, \tilde{\Pi}_{\hat{\alpha}_n} \circ \pi_j^{-1}) + \beta_S(\tilde{\Pi}_{\hat{\alpha}_n} \circ \pi_j^{-1}, \mathcal{N} \circ \pi_j^{-1}) + \beta_S(\mathcal{N} \circ \pi_j^{-1}, \mathcal{N}),$$

for some $j > 0$. Using the contraction result of Theorem 3.6.1 and following the argument of Theorem 8 of [21], we deduce that the first term is smaller that $\eta/3$ for sufficiently large j (in the case of $H(\delta)$ the result holds for all $\delta > 2$ - we recall from the proof of that theorem that if the required contraction is established in $H(\delta')$, then the required tightness argument holds in $H(\delta)$ for any $\delta > \delta'$). A similar result holds for the third term. For the middle term, note that the total variation distrance dominates the bounded Lipschitz metric. For fixed j, we therefore have that for n large enough,

$$\beta_S(\tilde{\Pi}_{\hat{\alpha}_n} \circ \pi_j^{-1}, \mathcal{N} \circ \pi_j^{-1}) \leq ||\Pi_{\hat{\alpha}_n}(\cdot|Y) \circ T_{\mathbb{Y}}^{-1} - N_V(0, I)||_{TV} \leq \eta/3,$$

using Theorem 3.6.2 with $V = \text{span}\{e_k : 1 \leq k \leq j\}$. This completes the proof. □

A similar situation holds true for the fully Bayesian approach.

Theorem 3.6.3. *Suppose that the prior density λ satisfies Condition 4. Then for every $\beta, R > 0$ and $M_n \to \infty$, we have*

$$\sup_{f_0 \in \mathcal{Q}(\beta,R)} \mathbb{E}_0 \Pi \left(f : ||f - f_0||_S \geq M_n L_n n^{-1/2} | Y \right) \to 0,$$

where $S = H(\delta)$ or H^{-p-s} for $s > 1/2$. If $S = H(\delta)$ then $L_n = (\log n)^{3/2} (\log \log n)^{1/2}$; if in addition $f_0 \in \mathcal{Q}_{SS}(\beta, R, \varepsilon)$, then the rate improves to $L_n = 1$ for $\delta \geq 2$. If $S = H^{-p-s}$ for $s > 1/2$, then $L_n = 1$.

Proof. This contraction result is proved in the same manner as Theorem 2.5 in [52], with suitable modifications arising as in the proof of Theorem 3.6.1. □

Theorem 3.6.4. *Let $V \subset \ell_2$ be a finite dimensional subspace such that $\pi_V(f_0) \neq 0$, where π_V denotes the projection onto V. Then the finite dimensional projection of the hierarchical Bayesian procedure satisfies a parametric BvM, that is*

$$\sup_{\substack{f_0 \in \mathcal{Q}(\beta,R), \\ \pi_V(f_0) \neq 0}} ||\Pi_n \circ T_{\mathbb{Y}}^{-1} - N_V(0, I)||_{TV} \to^{\mathbb{P}_0} 0,$$

where $T_z : f \mapsto \sqrt{n} \pi_V(f - z)$.

Proof of Theorem 3.3.2. The proof is exactly the same as that of Theorem 3.3.1, using Theorems 3.6.3 and 3.6.4 instead of Theorems 3.6.1 and 3.6.4. □

3.6.2 Proof of weak BvM result in L^∞ (Theorem 3.3.5)

Following Theorem 2 of [43], define the sets

$$\mathcal{J}_n(\gamma) = \left\{ (j,k) \in \Lambda : |f_{0,jk}| > \gamma \sqrt{\log n / n} \right\}$$

for $\gamma > 0$. In what follows, we denote by S the support of the prior draw, that is the set of non-zero coefficients of $f = (f_{jk})_{(j,k) \in \Lambda}$ drawn from the prior. We require the following contraction result.

Theorem 3.6.5. *Consider the slab and spike prior defined in Section 3.3.2 with lower threshold given by the strictly increasing sequence $j_0(n) \to \infty$. Then for every $0 < \beta_{min} \leq \beta_{max}$, $R > 0$ and $M_n \to \infty$, we have*

$$\sup_{f_0 \in \mathcal{H}(\beta,R)} \mathbb{E}_0 \Pi(f : ||f - f_0||_{\mathcal{M}(w)} \geq M_n n^{-1/2} \mid Y) \to 0$$

uniformly over $\beta \in [\beta_{min}, \beta_{max}]$, where (w_l) is any admissible sequence satisfying $w_{j_0(n)} \geq c\sqrt{\log n}$ for some $c > 0$.

Proof of Theorem 3.6.5. Fix $\eta > 0$. Consider the event

$$A_n = \{S^c \cap \mathcal{J}_n(\overline{\gamma}) = \emptyset\} \cap \{S \cap \mathcal{J}_n^c(\underline{\gamma}) = \emptyset\} \cap \{\max_{(j,k)\in\mathcal{J}_n(\underline{\gamma})} |f_{0,jk} - f_{jk}| \leq \overline{\gamma}\sqrt{(\log n)/n}\}. \quad (3.6.1)$$

By Theorem 2 of [43], there exist constants $0 < \underline{\gamma} < \overline{\gamma} < \infty$ (independent of β and R) such that

$$\sup_{f_0 \in \cup_{\beta\in[\beta_{min},\beta_{max}]}\mathcal{H}(\beta,R)} \mathbb{E}_0\Pi(A_n^c \mid Y) \lesssim n^{-B}, \quad (3.6.2)$$

for some $B = B(\beta_{min}, \beta_{max}, R) > 0$ (this follows since the probabilities of the complements of each of the events constituting A_n satisfy the above bound individually). We then have the following decomposition for some $D = D(\eta) > 0$ large enough to be specified later,

$$\begin{aligned}
&\mathbb{E}_0\Pi\left(||f - f_0||_{\mathcal{M}} \geq M_n n^{-1/2} \mid Y\right) \\
&\quad \leq \mathbb{E}_0\Pi\left(\{||f - f_0||_{\mathcal{M}} \geq M_n n^{-1/2}\} \cap \{||\pi_{j_0}(f - f_0)||_{\mathcal{M}} \leq Dn^{-1/2}\} \cap A_n \mid Y\right) \\
&\qquad + \mathbb{E}_0\Pi\left(\{||f - f_0||_{\mathcal{M}} \geq M_n n^{-1/2}\} \cap \{||\pi_{j_0}(f - f_0)||_{\mathcal{M}} > Dn^{-1/2}\} \cap A_n \mid Y\right) \\
&\qquad + E_0\Pi(A_n^c \mid Y).
\end{aligned} \quad (3.6.3)$$

Firstly note that the first term on the right-hand side of (3.6.3) is bounded by

$$\mathbb{E}_0\left(\{||(I - \pi_{j_0})(f - f_0)||_{\mathcal{M}} \geq (M_n - D)n^{-1/2}\} \cap A_n \mid Y\right), \quad (3.6.4)$$

where I is the identity operator. Combining this with (3.6.2), we can upper bound the right hand side of (3.6.3) by

$$\begin{aligned}
&\mathbb{E}_0\Pi(\{||(I - \pi_{j_0})(f - f_0)||_{\mathcal{M}} \geq \tilde{M}_n n^{-1/2}\} \cap A_n \mid Y) \\
&\qquad + \mathbb{E}_0\Pi(||\pi_{j_0}(f - f_0)||_{\mathcal{M}} > Dn^{-1/2} \mid Y) + o(1),
\end{aligned} \quad (3.6.5)$$

where $\tilde{M}_n = M_n - D \to \infty$ as $n \to \infty$. We bound the two remaining terms in (3.6.5) separately.

For the first term in (3.6.5), we can proceed as in the proof of Theorem 2 of [43]. By the definition of the Hölder ball $\mathcal{H}(\beta, R)$, there exists $J_n(\beta)$ such that $2^{J_n(\beta)} \leq k(n/\log n)^{1/(2\beta+1)}$ for some constant $k > 0$ such that $\mathcal{J}_n(\underline{\gamma}) \subset \{(j,k) : j \leq J_n(\beta),\ k = 0, ..., 2^j - 1\}$ and

$$\sup_{f_0\in\mathcal{H}(\beta,R)} \sup_{l>J_n(\beta)} w_l^{-1} \max_k |f_{0,lk}| \leq \frac{R2^{-J_n(\beta)(\beta+1/2)}}{\sqrt{J_n(\beta)}} \leq C(\beta, R)\frac{1}{\sqrt{n}}.$$

Consider now the frequencies $j_0 \leq l \leq J_n(\beta)$. On the event A_n, we have that

$$\sup_{j_0 \leq l \leq J_n(\beta)} \frac{1}{w_l} \max_k |f_{lk} - f_{0,lk}| \leq \frac{1}{w_{j_0}} \overline{\gamma} \sqrt{\frac{\log n}{n}} \leq \frac{\overline{\gamma}}{c} \frac{1}{\sqrt{n}},$$

since $w_{j_0(n)} \geq c\sqrt{\log n}$ by hypothesis. We thus have that on the event A_n, $||(I - \pi_{j_0})(f - f_0)||_{\mathcal{M}} = O(n^{-1/2})$ for any $f_0 \in \mathcal{H}(\beta, R)$, which proves that the first term in (3.6.5) is 0 for n sufficiently large.

Consider now the second term in (3.6.5). We shall use the approach of [19] using the moment generating function to control the low frequency terms. Recall that on these coordinates we have the simple product prior $\Pi(dx_1, ..., dx_{j_0}) = \prod_{k=1}^{j_0} g(x_i)dx_i$. One can prove as in Lemma 1 of [19] that we have the subgaussian bound

$$\mathbb{E}_0 \mathbb{E}^{\Pi}(e^{t\sqrt{n}(f_{lk} - Y_{lk})} \mid Y) \leq Ce^{t^2/2}$$

for some $C > 0$. Denote by Pr the law with expectation $\mathbb{E}_0\mathbb{E}^{\Pi}$, where $\mathbb{E}^{\Pi}$ denotes the expectation under the posterior measure. By a standard application of Markov's inequality we have the subgaussian bound

$$\Pr(\sqrt{n}|f_{lk} - Y_{lk}| > v) \leq C'e^{-Cv^2}$$

for all $v > 0$ and universal constants $C, C' > 0$. We then follow the proof of Theorem 2 of [22], which we include here for completeness. For some fixed constant $M > 0$, using the above bound yields

$$\begin{aligned}
\mathbb{E}_0\mathbb{E}^{\Pi}\left(\sup_{j \leq j_0} l^{-1/2} \max_k \sqrt{n}|f_{lk} - Y_{lk}| \mid Y\right) &\leq M + \int_M^{\infty} \Pr\left(\sup_{l \leq j_0} l^{-1/2} \max_k \sqrt{n}|f_{lk} - Y_{lk}| > u\right) du \\
&\leq M + \sum_{(l,k): l \leq j_0} \int_M^{\infty} \Pr\left(\sqrt{n}|f_{lk} - Y_{lk}| > \sqrt{l}u\right) du \\
&\leq M + C' \sum_{l \leq j_0} 2^l \int_M^{\infty} e^{-Clu^2} du \\
&\leq M + C' \sum_{1 \leq j_0} 2^l e^{-ClM^2} \leq C''.
\end{aligned}$$

By Markov's inequality and then the triangle inequality, the second term in (3.6.5) is then bounded by

$$\begin{aligned}
\frac{\sqrt{n}}{D}\mathbb{E}_0\mathbb{E}^{\Pi}\left(||\pi_{j_0}(f - f_0)||_{\mathcal{M}} \mid Y\right) &\leq \frac{\sqrt{n}}{D}\mathbb{E}_0\mathbb{E}^{\Pi}\left(||\pi_{j_0}(\mathbb{Y} - f_0)||_{\mathcal{M}} \mid Y\right) + \frac{C''}{D} \\
&\leq \frac{\mathbb{E}_0||\mathbb{Z}||_{\mathcal{M}}}{D} + \frac{C''}{D}.
\end{aligned} \tag{3.6.6}$$

By Proposition 2 of [22] and the fact that (w_l) is an admissible sequence, the first term in (3.6.6) is also bounded by C/D for some $C > 0$. Taking $D = D(\eta) > 0$ sufficiently large, (3.6.6) can be then made smaller than $\eta/2$. □

Proof of Theorem 3.3.5. Fix $\eta > 0$ and denote $\tilde{\Pi}_n = \Pi_n \circ \tau_{\mathbb{Y}}^{-1}$. By the triangle inequality, uniformly over the relevant class of functions,

$$\beta_{\mathcal{M}_0}(\tilde{\Pi}_n, \mathcal{N}) \leq \beta_{\mathcal{M}_0}(\tilde{\Pi}_n, \tilde{\Pi}_n \circ \pi_j^{-1}) + \beta_{\mathcal{M}_0}(\tilde{\Pi}_n \circ \pi_j^{-1}, \mathcal{N} \circ \pi_j^{-1}) + \beta_{\mathcal{M}_0}(\mathcal{N} \circ \pi_j^{-1}, \mathcal{N}),$$

for fixed $j > 0$. Since we have a $1/\sqrt{n}$-contraction rate in $\mathcal{M}$ for the posterior from Theorem 3.6.5, we can make the first term smaller than $\eta/3$ by taking j sufficiently large, again using the arguments of Theorem 8 of [21]. We recall from the proof of that theorem that if the required contraction is established in $\mathcal{M}(\overline{w})$ for an admissible sequence $(\overline{w}_l)$, then the required tightness argument holds in $\mathcal{M}_0(w)$ for any admissible (w_l) such that $w_l/\overline{w}_l \nearrow \infty$. A similar result holds for the third term.

For the middle term, note that $j_0(n) \geq j$ for n large enough. For such n, the projected prior onto the first j coordinates is a simple product prior which satisfies the usual conditions of the parametric BvM, namely it is has a density that is positive and continuous at the true (projected) paramter, and hence converges to 0 in total variation (see Chapter 10 of [81] for more details). Since the total variation distance dominates the bounded Lipschitz metric, this completes the proof. □

3.6.3 Proofs of finite dimensional BvM results (Theorems 3.6.2 and 3.6.4)

We recall some definitions from Knapik et al. [52]. Let $h_n : (0, \infty) \to [0, \infty)$ be

$$h_n(\alpha) = \frac{1 + 2\alpha + 2p}{n^{1/(2\alpha+2p+1)} \log n} \sum_{k=1}^{\infty} \frac{n^2 k^{2\alpha+1} f_{0,k}^2 \log k}{(k^{2\alpha+1} \rho_k^{-2} + n)^2}$$

and for $0 < l < L$ define the bounds

$$\underline{\alpha}_n = \inf\{\alpha > 0 : h_n(\alpha) > l\} \wedge \sqrt{\log n},$$

$$\overline{\alpha}_n = \inf\{\alpha > 0 : h_n(\alpha) > L(\log n)^2\}.$$

We recall (a slight modification of) Theorem 2.2 from [52]:

Theorem 3.6.6 (Knapik et al.)**.** *For every $R > 0$ the constants l and L can be chosen such that*

$$\inf_{f_0 \in \mathcal{Q}(\beta, R)} \mathbb{P}_0 \left(\text{argmax}_{\alpha \geq 0} \ell_n(\alpha) \in [\underline{\alpha}_n, \overline{\alpha}_n] \right) \to 1,$$

where $\ell_n(\alpha)$ denotes the log-likelihood for α.

The proof of Theorem 3.6.2 modifies the classical approach of Le Cam [60] by showing that the projections of the conditional posteriors $\Pi_\alpha(\cdot \mid Y)$ satisfy a BvM uniformly over the typical range of the estimator $\hat{\alpha}_n$.

Proof of Theorem 3.6.2. For $f_0 \in \mathcal{Q}(\beta, R)$, fix $\epsilon > 0$, define $I_n := [\underline{\alpha}_n, \overline{\alpha}_n]$ and consider the event $A_n = \{\hat{\alpha}_n \in I_n\}$. By Theorem 3.6.6,

$$\mathbb{P}_0\left(||\Pi_{\hat{\alpha}_n}(\cdot|Y) \circ T_{\mathbb{Y}}^{-1} - N(0, I_J)||_{TV} > \epsilon\right) \leq \mathbb{P}_0\left(\sup_{\alpha \in I_n} ||\Pi_\alpha(\cdot|Y) \circ T_{\mathbb{Y}}^{-1} - N(0, I_J)||_{TV} > \epsilon\right) + o(1),$$

uniformly over $f_0 \in \mathcal{Q}(\beta, R)$ as $n \to \infty$. It is therefore sufficient to show that the first term on the right-hand side converges to 0, that is to establish the parametric BvM uniformly over the set of posteriors $\{\Pi_\alpha(\cdot \mid Y) : \alpha \in I_n\}$

For integer $J \geq 1$, let $V_J = \text{span}\{e_k : 1 \leq k \leq J\}$ be a finite-dimensional subspace of ℓ_2 and let $\pi_J : \ell_2 \to \mathbb{R}^J$ denote the projection onto $\mathbb{R}^J$. Fix the smallest J such that $V \subset V_J$ and without loss of generality it is sufficient to prove the result with V_J instead of V. We shall firstly establish this result conditional on an arbitrary compact set $K \subset \mathbb{R}^J$ and then use an approximation argument to extend this to the full space. Abbreviate the conditional posterior distributions of the scaled and centered parameter $\sqrt{n}\pi_J(f - f_0)$ by $\tilde{\Pi}_{\alpha,n} = \Pi_\alpha(\cdot \mid Y) \circ T_{f_0}^{-1}$ and let Φ_n denote the normal distribution $N(\Delta_{n,f_0}, I_J)$, where $\Delta_{n,f_0} = \sqrt{n}\pi_J(Y - f_0)$. For an arbitrary probability measure μ and a compact subset K such that $\mu(K) > 0$, define the conditional version μ^K of μ by $\mu^K(B) = \mu(B \cap K)/\mu(K)$.

Let $K \subset \mathbb{R}^J$ be a compact subset. Let $\phi_n : K \to \mathbb{R}$ denote the Lebesgue density of Φ_n, $\tilde{\pi}_{\alpha,n} : K \to \mathbb{R}$ the Lebesgue density of (rescaled and centered) prior $\Pi_\alpha \circ T_{f_0}^{-1}$ and $s_n : K \to \mathbb{R}$ the likelihood ratio $s_n(h) = p_{f_0+n^{-1/2}h}/p_{f_0}(Y) = \exp(h^T\Delta_{n,f_0} - \frac{1}{2}h^Th)$. Consider the (random) functions $\psi_{\alpha,n} : K \times K \to \mathbb{R}$ defined by

$$\psi_{\alpha,n}(g, h) = \left(1 - \frac{\phi_n(h)s_n(g)\tilde{\pi}_{\alpha,n}(g)}{\phi_n(g)s_n(h)\tilde{\pi}_{\alpha,n}(h)}\right)_+.$$

Since we are exactly in a Gaussian shift experiment, we have that for any two sequences (g_n), (h_n) in K,

$$\log \frac{\phi_n(h_n)s_n(g_n)\tilde{\pi}_{\alpha,n}(g_n)}{\phi_n(g_n)s_n(h_n)\tilde{\pi}_{\alpha,n}(h_n)} = \log \frac{\tilde{\pi}_{\alpha,n}(g_n)}{\tilde{\pi}_{\alpha,n}(h_n)}.$$

Recall that by definition, the estimator $\hat{\alpha}_n$ is constrained to the interval $[0, \log n/v_n]$, where $v_n \to \infty$ is such that $v_n = o(\log n)$. Since $\tilde{\Pi}_\alpha = \otimes_{k=1}^J N(-\sqrt{n}f_{0,k}, nk^{-2\alpha-1})$ and

$K \subset B_M(0)$ for some $M > 0$ (since K is compact), we can bound the above quotient by

$$\begin{aligned}\left|\log \frac{\tilde{\pi}_{\alpha,n}(g_n)}{\tilde{\pi}_{\alpha,n}(h_n)}\right| &= \left|-\frac{1}{2n}\sum_{k=1}^{J} k^{2\alpha+1}(g_{n,k}^2 - h_{n,k}^2) - \frac{1}{\sqrt{n}}\sum_{k=1}^{J} k^{2\alpha+1} f_{0,k}(g_{n,k} - h_{n,k})\right| \\ &\leq \frac{J^{2\alpha+1}M^2}{n} + \frac{2J^{2\alpha+1}M\,||\pi_J(f_0)||_{\mathbb{R}^J}}{\sqrt{n}} \\ &\leq C(M,R)\exp\left(\left(\frac{2\log J}{v_n} - \frac{1}{2}\right)\log n\right) \to 0\end{aligned}$$

as $n \to \infty$. In particular, the bound holds uniformly over $\alpha \in I_n$. Since the functions $\psi_{\alpha,n}$ are continuous on the compact set $K \times K$, we have that

$$\sup_{\alpha\in I_n}\sup_{g,h\in K} \psi_{\alpha,n}(g,h) \to 0 \tag{3.6.7}$$

as $n \to \infty$ uniformly over $f_0 \in \mathcal{Q}(\beta, R)$, where the convergence is deterministic (unlike in [51]) since we are exactly in a Gaussian shift experiment.

Consider now K compact containing a neighbourhood of 0 (so that $\Phi_n(K) > 0$) and let $\Xi_n = \{\tilde{\Pi}_{\alpha,n}(K) > 0 \text{ for all } \alpha \in I_n\}_*$, where the $*$ signifies the inner measurable cover set in case the event is not measurable. Moreover, for given $\eta > 0$, we have by (3.6.7) that $\sup_{\alpha\in I_n}\sup_{g,h\in K}\psi_{\alpha,n}(g,h) \leq \eta$ for all $n \geq N = N(\eta,\beta,R)$ a non-random integer. Since the total variation distance between two arbitrary probability measures P and Q can be expressed in the form $||P-Q||_{TV} = 2\int(1-p/q)_+ dQ$, we have for $n \geq N$,

$$\begin{aligned}\frac{1}{2}\mathbb{E}_0 \sup_{\alpha\in I_n} ||\tilde{\Pi}_{\alpha,n}^K - \Phi_n^K||_{TV} 1_{\Xi_n} &= \mathbb{E}_0 \sup_{\alpha\in I_n}\int\left(1 - \frac{d\Phi_n^K}{d\tilde{\Pi}_{\alpha,n}^K}\right)_+ d\tilde{\Pi}_{\alpha,n}^K 1_{\Xi_n} \\ &= \mathbb{E}_0 \sup_{\alpha\in I_n}\int_K\left(1 - \int_K \frac{s_n(g)\tilde{\pi}_{\alpha,n}(h)\phi_n(h)}{s_n(h)\tilde{\pi}_{\alpha,n}(g)\phi_n(g)} d\Phi_n^K(g)\right)_+ d\tilde{\Pi}_{\alpha,n}^K(h) 1_{\Xi_n}.\end{aligned} \tag{3.6.8}$$

Applying Jensen's inequality to the convex function $x \mapsto (1-x)_+$ for the Φ_n^K-expectation yields that (3.6.8) is bounded by

$$\mathbb{E}_0^n \sup_{\alpha\in I_n}\int_{K\times K}\psi_{\alpha,n}(g,h) d\Phi_n^K(g) d\tilde{\Pi}_{\alpha,n}^K(h) 1_{\Xi_n} \leq \eta \mathbb{E}_0^n \sup_{\alpha\in I_n}\int_{K\times K} d\Phi_n^K(g) d\tilde{\Pi}_{\alpha,n}^K(h) = \eta$$

for $n \geq N$. We thus have that for any K compact containing a neighbourhood of 0, $\mathbb{E}_0^n \sup_{\alpha\in I_n} ||\tilde{\Pi}_{\alpha,n}^K - \Phi_n^K||_{TV} 1_{\Xi_n} \to 0$.

Let $K_m = B_{M_m}(0)$ denote a sequence of balls in $\mathbb{R}^J$ centered at 0 with radii $M_m \to \infty$. Since the convergence above holds for all $m \geq 1$, we can use a diagonal argument to extract a subsequence (m_n) such that the above convergence can still be obtained going through the sequence (K_{m_n}) simultaneously. We firstly note that by the proof of Theorem 3.6.1

(see Section 3.7 for more details), we have for $s > 1/2$,

$$\frac{n}{M_n^2} \sup_{f_0 \in \mathcal{Q}(\beta,R)} \mathbb{E}_0^n \sup_{\alpha \in I_n} \int ||f - f_0||_{H^{-p-s}}^2 \, \Pi_\alpha(df|Y) \to 0 \tag{3.6.9}$$

for any $M_n \to \infty$. Consider the events $\Xi_n = \{\tilde{\Pi}_{\alpha,n}(K_{m_n}) > 0 \text{ for all } \alpha \in I_n\}_*$, which we check satisfy $\mathbb{P}_0^n(\Xi_n) \to 1$ as $n \to \infty$. Fix $0 < r < 1$ and $s > 1/2$ and define $R_n = M_{m_n} J^{-p-s} \to \infty$. Applying Markov's inequality twice yields that

$$\begin{aligned}
\mathbb{P}_0^n(\Xi_n^c) &= \mathbb{P}_0^n(\tilde{\Pi}_{\alpha,n}(K_{m_n}) = 0 \text{ for some } \alpha \in I_n) \\
&= \mathbb{P}_0^n(\Pi_\alpha(f : \sqrt{n}\,||\pi_J(f - f_0)||_2 \leq M_{m_n} \mid Y) = 0 \text{ for some } \alpha \in I_n) \\
&\leq \mathbb{P}_0^n(\Pi_\alpha(f : ||f - f_0||_{H^{-p-s}} \leq M_{m_n} J^{-p-s} n^{-1/2} \mid Y) = 0 \text{ for some } \alpha \in I_n) \\
&\leq \mathbb{P}_0^n\left(\sup_{\alpha \in I_n} \Pi_\alpha(f : ||f - f_0||_{H^{-p-s}} \geq R_n n^{-1/2} \mid Y) \geq 1 - r\right) \\
&\leq \frac{1}{1-r} \mathbb{E}_0^n \sup_{\alpha \in I_n} \Pi_\alpha(f : ||f - f_0||_{H^{-p-s}} \geq R_n n^{-1/2} \mid Y) \\
&\leq \frac{n}{(1-r)R_n^2} \mathbb{E}_0^n \sup_{\alpha \in I_n} \int ||f - f_0||_{H^{-p-s}}^2 \, \Pi_\alpha(df|Y) \to 0,
\end{aligned}$$

where the last term converges to 0 uniformly over $f_0 \in \mathcal{Q}(\beta, R)$ by (3.6.9).

As a result, we have that there exists a sequence of balls (K_{m_n}) with radii $M_{m_n} \to \infty$ such that $\mathbb{E}_0 \sup_{\alpha \in I_n} ||\tilde{\Pi}_{\alpha,n}^{K_{m_n}} - \Phi_n^{K_{m_n}}||_{TV} 1_{\Xi_n} \to 0$, where the conditional probabilities are well-defined on events Ξ_n with $\mathbb{P}_0^n(\Xi_n) \to 1$. Note that for any set K and measure Π the total variation distance satisfies $||\Pi - \Pi^K||_{TV} \leq 2\Pi(K^c)$. Since Lemma 5.2 of [51] yields that $\Phi(K_{m_n}^c) \to^{P_{f_0}^n} 0$, we have that combined with the above

$$\begin{aligned}
\mathbb{E}_0 \sup_{\alpha \in I_n} ||\tilde{\Pi}_{\alpha,n} - \Phi_n||_{TV} &\leq \mathbb{E}_0 \sup_{\alpha \in I_n} ||\tilde{\Pi}_{\alpha,n} - \tilde{\Pi}_{\alpha,n}^{K_{m_n}}||_{TV} + \mathbb{E}_0 \sup_{\alpha \in I_n} ||\tilde{\Pi}_{\alpha,n}^{K_{m_n}} - \Phi_n^{K_{m_n}}||_{TV} \\
&\quad + \mathbb{E}_0 ||\Phi_n^{K_{m_n}} - \Phi_n||_{TV} \\
&\leq 2\mathbb{E}_0 \sup_{\alpha \in I_n} \tilde{\Pi}_{\alpha,n}(K_{m_n}^c) + o(1).
\end{aligned}$$

Since $||\pi_J(f - f_0)||_{\mathbb{R}^J} \leq J^{p+s} ||f - f_0||_{H^{-p-s}}$, we have by (3.6.9) that

$$\mathbb{E}_0 \sup_{\alpha \in I_n} \tilde{\Pi}_{\alpha,n}(K_{m_n}^c) \leq \mathbb{E}_0 \sup_{\alpha \in I_n} \Pi_\alpha(f : ||f - f_0||_{H^{-p-s}} \geq M_{m_n} J^{-p-s} n^{-1/2} | Y) = o(1),$$

which completes the proof. □

Proof of Theorem 3.6.4. Recall that we assume that $f_{0,1} = 0$ (otherwise as indicated above, the following arguments still hold true under a slight modification to the prior). Let V_J be as in the proof of Theorem 3.6.2. Since the hierarchical prior is conditionally a product prior, we have that the distribution $\Pi \circ \pi_V^{-1}$ is absolutely continuous with respect

to J-dimensional Lebesgue measure with density function

$$p_J(x_1, ..., x_J) = (2\pi)^{-J/2} \int_0^\infty (J!)^{\alpha+1/2} e^{-2^{2\alpha}x_1^2 - \frac{1}{2}\sum_{k=2}^J k^{2\alpha+1}x_k^2} \lambda(\alpha) d\alpha.$$

The integral is finite on $\mathbb{R}^J \backslash \{0\}$, so that the above expression is well-defined except on a set of J-dimensional Lebesgue measure 0. The density p_J is positive on $\mathbb{R}^J$ and continuous by a dominated convergence argument.

For f_0 such that $\pi_V(f_0) \neq 0$, we verify the conditions of Theorem 2.1 of [51], which we note can be made uniform over $\mathcal{Q}(\beta, R)$ by carefully keeping track of the constants in the proof. As shown above, the prior $\Pi \circ \pi_V^{-1}$ has a Lebesgue density that is continuous and positive at $\pi_V(f_0)$. Since we are in a Gaussian white noise model, the model trivially satisfies the stochastic local asymptotic normality condition (2.1) of [51] at any point $\pi_V(f) \in \mathbb{R}^J$, with random vectors $\Delta_{n,f} = \sqrt{n}\pi_V(Y - f)$, non-singular matrix $V_f \equiv I_J$ and norming rate $\delta_n = n^{-1/2}$. For any $M_n \to \infty$,

$$\mathbb{E}_0 \Pi_n(f : ||\pi_V(f - f_0)|| \geq M_n n^{-1/2} | Y) \leq \mathbb{E}_0 \Pi_n(f : ||f - f_0||_{H^{-p-s}} \geq M_n J^{-p-s} n^{-1/2} | Y) \to 0,$$

by Theorem 3.6.3. By Theorem 2.1 of [51], we thus have that for $T_z : f \mapsto \sqrt{n}\pi_V(f - z)$,

$$\left|\left|\Pi_n \circ T_{f_0}^{-1} - N(\Delta_{n,f_0}, V_{f_0}^{-1})\right|\right|_{TV} = ||\Pi_n \circ T_Y^{-1} - N(0, I_J)||_{TV} \to 0$$

as $n \to \infty$ in $\mathbb{P}_0$-probability. □

3.6.4 Other proofs

ℓ_2 confidence sets

Self-similarity ensures that we can estimate the unknown smoothness β of the signal f_0 at the rate $O_{\mathbb{P}_0}(1/\log n)$ using the estimator $\hat{\alpha}_n$; this is necessary in order for the radius of any confidence set to adapt to the unknown smoothness. The behaviour of $\hat{\alpha}_n$ is contained in Lemma 3.11 of [79], which is summarized below for convenience.

Lemma 3.6.7 (Szabó et al.)**.** *For any $0 < \beta \leq A - 1$ and $R \geq 1$, there exist constants K_1 and K_2 such that $\mathbb{P}_0(\beta - K_1/\log n \leq \hat{\alpha}_n \leq \beta + K_2/\log n) \to 1$ uniformly over $f_0 \in \mathcal{Q}_{SS}(\beta, R, \varepsilon)$.*

As mentioned in the discussion following the lemma in [79], the constant K_2 is negative for large enough R so that the estimate $\hat{\alpha}_n$ undersmooths the true β. We recall that the posterior distribution corresponding to the prior Π_α in (3.3.1) is given by

$$\Pi_\alpha(\cdot \mid Y) = \bigotimes_{k=1}^\infty N\left(\frac{n\rho_k^{-1}}{k^{2\alpha\rho_k^{-2}+1} + n} Y_k, \frac{\rho_k^{-2}}{k^{2\alpha+1}\rho_k^{-2} + n}\right). \tag{3.6.10}$$

Proof of Proposition 3.4.1. By (3.6.10), we see that for $f \sim \Pi_\alpha(\cdot \mid Y)$, we have $E^\Pi[||f||^2_{H^{\alpha'}} \mid Y] < \infty$ if and only if $\alpha' < \alpha$, from which we deduce that $\Pi_{\hat{\alpha}_n}(||f||_{H^{\hat{\alpha}_n-\epsilon_n}} \leq M_n \mid Y) \to^{\mathbb{P}_0} 1$ since $M_n \to \infty$. Since R_n is selected such that $\Pi_{\hat{\alpha}_n}(C_n \mid Y) = 1-\gamma$, the claim about the posterior credibility of $\tilde{C}_n$ follows.

For the coverage, we have by Lemma 3.6.7 that for sufficiently large n, $\hat{\alpha}_n \leq \beta$ and consequently $||f_0||_{H^{\hat{\alpha}_n-\epsilon_n}} \leq M_n$ eventually under $\mathbb{P}_0$, uniformly over $f_0 \in \mathcal{Q}_{SS}(\beta, R)$. For n large enough, we therefore have that $\mathbb{P}_0(f_0 \in \tilde{C}_n) = \mathbb{P}_0(||f_0 - Y||_H \leq R_n/\sqrt{n})$ and the proof of coverage then follows by Theorem 1 of [21] since $H(\delta)$-balls form a uniformity class for the measure $\mathcal{N}$.

Consider firstly M_n growing more slowly than any power of $\log n$. Let $f_1, f_2 \in \tilde{C}_n$ and set $g = f_1 - f_2$. Picking $J_n \sim [nM_n^2/(\log n)^{2\delta}]^{1/(1+2\hat{\alpha}_n-2\epsilon_n+2p)}$ yields

$$\begin{aligned}
||g||_2^2 = \sum_{k=1}^{\infty} |g_k|^2 &= \sum_{k=1}^{J_n} k^{2p+1}k^{-2p-1}(\log k)^{2\delta-2\delta}|g_k|^2 + \sum_{k=J_n+1}^{\infty} k^{2(\hat{\alpha}_n-\epsilon_n)-2(\hat{\alpha}_n-\epsilon_n)}|g_k|^2 \\
&\leq J_n^{2p+1}(\log J_n)^{2\delta}\,||g||^2_{H(\delta)} + J_n^{-2(\hat{\alpha}_n-\epsilon_n)}\,||g||^2_{H^{\hat{\alpha}_n-\epsilon_n}} \\
&= O_{\mathbb{P}_0}\left(J_n^{2p+1}(\log J_n)^{2\delta}n^{-1} + J_n^{-2(\hat{\alpha}_n-\epsilon_n)}M_n^2\right) \\
&= O_{\mathbb{P}_0}\left(n^{-\frac{2(\hat{\alpha}_n-\epsilon_n)}{1+2\hat{\alpha}_n-2\epsilon_n+2p}}(\log n)^{\frac{4\delta(\hat{\alpha}_n-\epsilon_n)}{1+2\hat{\alpha}_n-2\epsilon_n+2p}}M_n^{\frac{2(2p+1)}{1+2\hat{\alpha}_n-2\epsilon_n+2p}}\right),
\end{aligned}$$

where the constants do not depend on g. Since $|\hat{\alpha}_n - \beta| = O_{\mathbb{P}_0}(1/\log n)$ by Lemma 3.6.7 and $\epsilon_n = O(1/\log n)$ by assumption, some straightforward computations yield that $||g||_2^2 = O_{\mathbb{P}_0}(n^{-2\beta/(2\beta+2p+1)}(\log n)^{4\delta\beta/(2\beta+2p+1)}M_n^{(4p+2)/(2\beta+2p+1)})$ as $n \to \infty$. If $M_n \asymp (\log n)^\zeta$, then the diameter result follows exactly as above taking $J_n \sim n^{1/(2\hat{\alpha}_n+2p+1)}(\log n)^{-2(\delta-\zeta)/(2\hat{\alpha}_n+2p+1)}$. □

We have an analogous approach in the fully Bayesian case.

Lemma 3.6.8. *The posterior median α_n^M of the marginal posterior distribution $\lambda_n(\cdot|Y)$ satisfies*

$$\inf_{f_0 \in \mathcal{Q}(\beta,R)} \mathbb{P}_0\left(\alpha_n^M \in [\underline{\alpha}_n, \overline{\alpha}_n]\right) \to 1$$

as $n \to \infty$, where $\underline{\alpha}_n$ and $\overline{\alpha}_n$ are defined in Section 3.6.3 above. Moreover, for $C = C(\beta, R, \varepsilon, \rho)$,

$$\inf_{f_0 \in \mathcal{Q}_{SS}(\beta,R,\varepsilon)} \mathbb{P}_0\left(\left|\alpha_n^M - \beta\right| \leq C/\log n\right) \to 1.$$

Proof. The proof follows directly from the proof of Theorem 2.5 of [52]. □

Proof of Proposition 3.4.2. Using Lemma 3.6.8, the proof follows in the same was as that of Proposition 3.4.1. □

L^∞ confidence bands

To prove Proposition 3.4.3 we need to understand the behaviour of the posterior median under the law $\mathbb{P}_0$, which is the content of the next lemma.

Lemma 3.6.9. *Let $\tilde{f} = \tilde{f}_n$ denote the posterior median (defined coordinate-wise) of the slab and spike prior. Then the event*

$$\begin{aligned} B_n = \{\tilde{f}_{lk} = 0 \quad \forall (l,k) \in \mathcal{J}_n^c(\underline{\gamma})\} \cap \{\tilde{f}_{lk} \neq 0 \quad \forall (l,k) \in \mathcal{J}_n(\overline{\gamma}')\} \\ \cap \{\sqrt{n}|Y_{lk} - f_{0,lk}| \leq (8l\log 2 + a\log n)^{1/2} \quad \forall l \leq J_n, \forall k = 0, ..., 2^l - 1\} \end{aligned} \tag{3.6.11}$$

satisfies $\inf_{f_0 \in \mathcal{H}(\beta,R)} \mathbb{P}_0(B_n) \to 1$ *as* $n \to \infty$, *for some constants* $0 < \underline{\gamma} < \overline{\gamma}' < \infty$ *and* $a > 0$.

Proof. We show that the $\mathbb{P}_0$-probability of each of these events individually tends to 1. For the first event

$$\begin{aligned} \{\tilde{f}_{lk} = 0 \quad \forall (l,k) \in \mathcal{J}_n^c(\underline{\gamma})\} &\supseteq \{\Pi(f_{lk} = 0 \mid Y) \geq 1/2 \quad \forall (l,k) \in \mathcal{J}_n^c(\underline{\gamma})\} \\ &\supseteq \{\Pi(f_{lk} = 0 \quad \forall (l,k) \in \mathcal{J}_n^c(\underline{\gamma})) \geq 1/2\} \\ &= \{\Pi(S \cap \mathcal{J}_n^c(\underline{\gamma}) = \emptyset) \geq 1/2\}. \end{aligned}$$

By Lemma 1 of [43] the $\mathbb{P}_0$-probability of this last event tends to 1 for some $\underline{\gamma} > 0$ as $n \to \infty$.

Consider the third event,

$$\Omega_n = \{\sqrt{n}|Y_{lk} - f_{0,lk}| \leq (8l\log 2 + a\log n)^{1/2} \quad \forall l \leq J_n, \forall k = 0, ..., 2^l - 1\},$$

where $a > 32\log 2$ is a constant. Using the standard inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ and that $l \leq J_n \leq \log n$,

$$\begin{aligned} (\sqrt{8l\log 2 + a\log n} - \sqrt{8l\log 2})^2 &= 16l\log 2 + a\log n - 2\sqrt{64l^2(\log 2)^2 + 8al\log 2\log n} \\ &\geq 16l\log 2 + a\log n - 16\log l(\log 2) - 2\sqrt{8a\log 2}\log n \\ &= a'\log n, \end{aligned}$$

where $a' > 0$ by the choice of a. Using the Borell-Sudakov-Tsirelson inequality [62] and

that $\mathbb{E}_0 \max_{1\le i\le n} |Z_i| \le \sqrt{8\log n}$ for (Z_i) i.i.d. standard normal random variables yields

$$\begin{aligned}
\mathbb{P}_0(\Omega_n^c) &\le \sum_{l=1}^{J_n} \mathbb{P}_0\left(\max_k |Z_{lk}| \ge \sqrt{2l\log 2 + \frac{1}{2}\log n}\right) \\
&\le \sum_{l=1}^{J_n} \mathbb{P}_0\left(\max_k |Z_{lk}| - \mathbb{E}_0 \max_k |Z_{lk}| \ge \sqrt{2l\log 2 + \frac{1}{2}\log n} - \sqrt{8l\log 2}\right) \\
&\le 2\sum_{l=1}^{J_n} \exp\left(-\frac{1}{2}\left(\sqrt{8l\log 2 + a\log n} - \sqrt{8l\log 2}\right)\right) \\
&\le 2J_n \exp\left(-\frac{a'}{2}\log n\right) \le 2n^{-a'/2}\log n \to 0.
\end{aligned}$$

as required. We shall lastly show that

$$\Omega_n \subset \{\tilde{f}_{lk} \ne 0 \quad \forall (l,k) \in \mathcal{J}_n(\overline{\gamma}')\}, \tag{3.6.12}$$

which then completes the proof.

Consider firstly the case $f_{0,lk} \in \mathcal{J}_n(\overline{\gamma}')$ with $f_{0,lk} > 0$. Write

$$\Pi(f_{lk} \le 0 \mid Y) = \Pi(f_{lk} = 0 \mid Y) + \Pi(f_{lk} < 0 \mid Y). \tag{3.6.13}$$

By the proof of Lemma 1 of [43], we have that on the event Ω_n and for sufficiently large $\overline{\gamma}'$, the first posterior probability in (3.6.13) is bounded above by a multiple of $n^{K+1/2-(\overline{\gamma}')^2/8}$. Again on the event Ω_n, we use (42) of [43] to bound the second term via

$$\begin{aligned}
\Pi(f_{lk} < 0 \mid Y) &= \frac{w_{jn}\int_{-\infty}^{0} e^{-\frac{n}{2}(x-Y_{lk})^2} g(x)dx}{w_{jn}\int_{-\infty}^{\infty} e^{-\frac{n}{2}(x-Y_{lk})^2} g(x)dx + (1-w_{j,n})} \\
&\le \frac{||g||_\infty \int_{-\infty}^{-\sqrt{n}Y_{lk}} e^{-\frac{1}{2}v^2} dv}{a(\pi/n)^{1/2}} = C\sqrt{n}\bar{\Phi}(\sqrt{n}Y_{lk}),
\end{aligned} \tag{3.6.14}$$

where $\bar{\Phi} = 1 - \Phi$ with Φ the distribution function of a standard normal variable. On Ω_n, we have for $l \le J_n$,

$$Y_{lk} = (Y_{lk} - f_{0,lk}) + f_{0,lk} \ge -\sqrt{\frac{2J_n\log 2 + \frac{1}{2}\log n}{n}} + \overline{\gamma}'\sqrt{\frac{\log n}{n}} \ge \delta\sqrt{\frac{\log n}{n}}$$

for some $\delta = \delta(\overline{\gamma}') > 0$ that can be made arbitrarily large by taking $\overline{\gamma}'$ large enough. Thus applying the standard tail bounds for $\bar{\Phi}$ we have that the right-hand side of (3.6.14) is bounded above by a multiple of

$$\sqrt{n}\bar{\Phi}(\delta\sqrt{\log n}) \le \frac{\sqrt{n}}{\delta\sqrt{2\pi\log n}} e^{-\frac{1}{2}\delta^2\log n} = C(\delta)\frac{n^{\frac{1}{2}-\frac{1}{2}\delta^2}}{\sqrt{\log n}}.$$

Combining the above results, we have that for sufficiently large $\overline{\gamma}'$ (and hence δ), (3.6.13) is bounded above by a constant times n^{-B} for some $B > 0$, uniformly over the positive coefficients in $\mathcal{J}_n(\overline{\gamma}')$. In particular, the posterior median satisfies $\tilde{f}_{lk} > 0$ for all $(l,k) \in \mathcal{J}_n(\overline{\gamma}')$ with $f_{lk} > 0$ and n large enough. The case $f_{0,lk} < 0$ is dealt with in exactly the same way, thereby proving (3.6.12). □

Proof of Proposition 3.4.3. By Lemma 3.6.9, it suffices to prove all the results on the event B_n defined in (3.6.11). We firstly establish the diameter of the confidence set. Taking $f_1, f_2 \in D_n$ and setting $2^{J_n(\beta)} \simeq (n/\log n)^{1/(2\beta+1)}$, we have on B_n,

$$\begin{aligned}
||f_1 - f_2||_\infty &\leq ||f_1 - \pi_{med}(Y)||_\infty + ||f_2 - \pi_{med}(Y)||_\infty \\
&\leq 2 \sup_{x\in[0,1]} \sum_{l=0}^{J_n(\beta)} \sum_{k=0}^{2^l-1} \frac{v_n}{\sqrt{n}} |\psi_{lk}(x)| \\
&\leq C(\psi) \frac{v_n}{\sqrt{n}} \sum_{l=0}^{J_n(\beta)} 2^{l/2} \leq C' \frac{2^{J_n(\beta)/2} v_n}{\sqrt{n}} = O_{\mathbb{P}_0}\left(\left(\frac{\log n}{n} \right)^{\frac{\beta}{2\beta+1}} v_n \right).
\end{aligned}$$

We now need to establish asymptotic coverage. Split $f_0 = \pi_{\mathcal{J}_n(\underline{\gamma})}(f_0) + \pi_{\mathcal{J}_n^c(\underline{\gamma})}(f_0)$. Since $\pi_{\mathcal{J}_n^c(\underline{\gamma})} \circ \pi_{med}(Y) = 0$ on B_n, we can write

$$||f_0 - \pi_{med}(Y)||_\infty \leq ||\pi_{med}(f_0 - Y)||_\infty + ||(id - \pi_{med}) \circ \pi_{\mathcal{J}_n(\underline{\gamma})}(f_0)||_\infty + ||\pi_{\mathcal{J}_n^c(\underline{\gamma})}(f_0)||_\infty. \tag{3.6.15}$$

For the third term in (3.6.15), note that since $f_0 \in \mathcal{H}(\beta, L)$,

$$\begin{aligned}
||\pi_{\mathcal{J}_n^c(\underline{\gamma})}(f_0)||_\infty &\leq \sum_{l=0}^{\infty} 2^{l/2} \max_{k:(l,k)\in\mathcal{J}_n^c(\underline{\gamma})} |\langle f_0, \psi_{lk}\rangle| \\
&\leq \sum_{l=0}^{J_n(\beta)} 2^{l/2} \underline{\gamma} \sqrt{\frac{\log n}{n}} + \sum_{l>J_n(\beta)} 2^{-l\beta} \leq C(\beta, L) \left(\frac{\log n}{n} \right)^{\frac{\beta}{2\beta+1}}.
\end{aligned} \tag{3.6.16}$$

For the second term in (3.6.15), we note that any indices remaining satisfy $(l,k) \in \mathcal{J}_n^c(\overline{\gamma}')$ and so by the same reasoning as above, this term is also $O((\log n/n)^{\beta/(2\beta+1)})$.

It is shown in the proof of Proposition 3 of [42] that self-similarity in the sense of Definition 7 implies the existence of infinitely many coefficients of significant size in the following sense. Firstly observe that for $f_0 \in \mathcal{H}_{SS}(\beta, R, \varepsilon)$ and any $j \geq j_0$,

$$||\psi||_\infty \sum_{l\geq j} 2^{l/2} \sup_k |\langle f_0, \psi_{lk}\rangle| \geq ||K_j(f) - f||_\infty \geq \varepsilon 2^{-j\beta}.$$

Let N be a fixed integer and let $j \geq j_0$. Lower bounding the maximum by the average

over the range $[j, j+N-1]$ yields

$$\begin{aligned}
\sup_{(l,k):l\geq j} |\langle f_0, \psi_{lk}\rangle| &\geq \frac{1}{N}\sum_{l=j}^{j+N-1} \sup_k |\langle f_0, \psi_{lk}\rangle| \\
&\geq \frac{2^{-(j+N)/2}}{N}\left(\sum_{l=j}^{\infty} 2^{l/2}\sup_k |\langle f_0, \psi_{lk}\rangle| - \sum_{l=j+N}^{\infty} 2^{l/2}\sup_k |\langle f_0, \psi_{lk}\rangle|\right) \\
&\geq \frac{2^{-(j+N)/2}}{N}\left(\frac{\varepsilon}{||\psi||_\infty}2^{-j\beta} - \frac{R}{1-2^{-\beta}}2^{-(j+N)\beta}\right) \\
&\geq \frac{2^{-(j+N)/2}}{2N||\psi||_\infty}\varepsilon 2^{-j\beta} \geq d(\varepsilon, R, \beta, \psi)2^{-j(\beta+1/2)}
\end{aligned}$$

for some $d(\varepsilon, R, \beta, \psi) > 0$ if N is chosen sufficiently large (but finite), depending only on $\varepsilon, R, \beta, \psi$. Let $\tilde{J}_n(\beta)$ be such that $\frac{\epsilon}{2}(n/\log n)^{1/(2\beta+1)} \leq 2^{\tilde{J}_n(\beta)} \leq \epsilon(n/\log n)^{1/(2\beta+1)}$, where $\epsilon = \epsilon(b, R, \beta, \psi) > 0$ is small enough so that $d/\epsilon^{\beta+1/2} > \overline{\gamma}'$. Using this yields

$$\sup_{(l,k):l\geq \tilde{J}_n(\beta)} |\langle f_0, \psi_{lk}\rangle| \geq \frac{d(b, R, \beta, \psi)}{\epsilon^{\beta+1/2}}\sqrt{\frac{\log n}{n}} > \overline{\gamma}'\sqrt{\frac{\log n}{n}}.$$

We therefore have that on the event B_n, there exists (l', k') with $l' \geq \tilde{J}_n(\beta)$ such that $\tilde{f}_{l'k'} \neq 0$ and a non-zero coefficient therefore appears in the definition (3.4.4) of $\sigma_{n,\gamma}$. We can thus lower bound

$$\sigma_{n,\gamma} \geq \frac{v_n}{\sqrt{n}}\sup_{x\in[0,1]} |\psi_{l'k'}(x)| \geq c(\psi)\frac{v_n 2^{\tilde{J}_n(\beta)/2}}{\sqrt{n}} = c' v_n \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}. \tag{3.6.17}$$

Now, since $v_n \to \infty$ as $n \to \infty$, we have from (3.6.16) and the remark after it that for sufficiently large n (depending on β and R), the first two terms in (3.6.15) satisfy

$$||(id - \pi_{med}) \circ \pi_{\mathcal{J}_n(\underline{\gamma})}(f_0)||_\infty + ||\pi_{\mathcal{J}_n^c(\underline{\gamma})}(f_0)||_\infty \leq C\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}} \leq \sigma_{n,\gamma}/2.$$

For the first term in (3.6.15) we recall that on B_n, the posterior median only picks up coefficients (l, k) with $l \leq J_n(\beta) \leq J_n$. Therefore on this event,

$$\begin{aligned}
||\pi_{med}(f_0 - Y)||_\infty &\leq \sup_{x\in[0,1]} \sum_{(l,k)\in med} |f_{0,lk} - Y_{lk}||\psi_{lk}(x)| \\
&\leq C(\psi)\sqrt{\frac{\log n}{n}} \sum_{(l,k):l\leq J_n(\beta)} 2^{l/2} \leq C'\left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+1}}.
\end{aligned}$$

Using the lower bound (3.6.17), we deduce that on B_n, $||\pi_{med}(f_0 - Y)||_\infty \leq \sigma_{n,\gamma}(Y)/2$ for n large enough, uniformly over $f_0 \in \mathcal{H}_{SS}(\beta, R)$. Combining all of the above yields that

$B_n \subset \{||f_0 - \pi_{med}(Y)||_\infty \leq \sigma_{n,\gamma}\}$. We therefore conclude that

$$\begin{aligned}\mathbb{P}_0(f_0 \in D_n) &= \mathbb{P}_0(\{||f_0 - Y||_{\mathcal{M}(w)} \leq R_n/\sqrt{n}\} \cap \{||f_0 - \pi_{med}(Y)||_\infty \leq \sigma_{n,\gamma}\} \cap B_n) + o(1) \\ &= \mathbb{P}_0(\{||f_0 - Y||_{\mathcal{M}(w)} \leq R_n/\sqrt{n}\} \cap B_n) + o(1) \\ &= 1 - \gamma + o(1),\end{aligned}$$

where we have used that $\mathbb{P}_0(B_n) \to 1$ and that $\mathbb{P}_0(||f_0 - Y||_{\mathcal{M}(w)} \leq R_n/\sqrt{n}) \to 1 - \gamma$ by Theorem 5 of [22]. □

Remaining proofs

Proof of Proposition 3.3.3. Fix $\rho > 1$, let $\varepsilon = \varepsilon(\alpha, \rho, R) < (1-\rho^{-2\alpha})/(2\alpha R)$ be sufficiently small so that $\varepsilon \in (0,1)$ and consider the events $A_{\alpha,N} = \{\sum_{k=N}^{\lceil \rho N \rceil} f_k^2 < \varepsilon R N^{-2\alpha}\}$. By a simple integral comparison we have that $\sum_{k=N}^{\lceil \rho N \rceil} k^{-2\alpha-1} \geq (2\alpha)^{-1} N^{-2\alpha}(1-\rho^{-2\alpha})$, so that under the conditional prior,

$$\begin{aligned}\Pi_\alpha(A_{\alpha,N}) &= \mathbb{P}\left(\sum_{k=N}^{\lceil \rho N \rceil} k^{-2\alpha-1} g_k^2 < \varepsilon R N^{-2\alpha}\right) \\ &\leq \mathbb{P}\left(\sum_{k=N}^{\lceil \rho N \rceil} k^{-2\alpha-1}(g_k^2 - 1) < \varepsilon R N^{-2\alpha} - \frac{1}{2\alpha} N^{-2\alpha}(1-\rho^{-2\alpha})\right) \\ &\leq \mathbb{P}\left(\sum_{k=N}^{\lceil \rho N \rceil} k^{-2\alpha-1}(g_k^2 - 1) < -\varepsilon' N^{-2\alpha}\right),\end{aligned}$$

where the g_k's are i.i.d. standard normal random variables and $\varepsilon' > 0$ (by the choice of ε). By (4.2) of Lemma 1 of [59] we have the exponential inequality

$$\mathbb{P}\left(\sum_{k=N}^{\lceil \rho N \rceil} k^{-2\alpha-1}(g_k^2 - 1) \leq -2\left(\sum_{k=N}^{\lceil \rho N \rceil} k^{-4\alpha-2}\right)^{1/2} \sqrt{x}\right) \leq e^{-x}.$$

For $N \geq 2$, again by an integral comparison we have that $\sum_{k=N}^{\lceil \rho N \rceil} k^{-4\alpha-2} \leq C(\alpha) N^{-4\alpha-1}$. Using this and letting $x = MN$, the exponential inequality becomes

$$\mathbb{P}\left(\sum_{k=N}^{\lceil \rho N \rceil} k^{-2\alpha-1}(g_k^2 - 1) \leq -C'(\alpha)\sqrt{M} N^{-2\alpha}\right) \leq e^{-MN}.$$

Taking M sufficiently small so that $C'(\alpha)\sqrt{M} < \varepsilon'$, we obtain that $\Pi_\alpha(A_{\alpha,N}) \leq e^{-MN}$. Since this sequence is summable in N, the result follows from the first Borel-Cantelli Lemma. □

Proof of Proposition 3.3.7. Under the law $\mathbb{P}_0$, $\sqrt{n}\mathbb{E}_0||\mathbb{Y} - f_0||_{\mathcal{M}(w)} = \mathbb{E}_0||\mathbb{Z}||_{\mathcal{M}(w)} < \infty$ by Proposition 2 of [22]. By the triangle inequality it therefore suffices to show the conclusion of Proposition 3.3.7 with $\mathbb{Y}$ replaced by f_0. Rewrite the multiscale indices $\Lambda = \{(l,k) : l \geq 0, k = 0, ..., 2^l - 1\}$ in increasing lexicographic order, so that $\Lambda = \{(l_m, k_m) : m \in \mathbb{N}\}$, where

$$\begin{aligned} l_m &= i, \qquad\qquad if \quad 2^i \leq m < 2^{i+1}, \quad i = 0, 1, 2, ..., \\ k_m &= m - 2^i, \qquad if \quad\quad 2 \leq m < 2^{i+1}, \quad i = 0, 1, 2, ... \end{aligned}$$

Consider a strictly increasing subsequence $(n_m)_{m\geq 1}$ of $\mathbb{N}$ such that $(\log n_m)/w_{l_m}^2 \to \infty$ (such a subsequence can be constructed for any admissible (w_l) since $w_l \nearrow \infty$). Define a function $f_0 \in \ell_2$ via its wavelet coefficients

$$\langle f_0, \psi_{l_m k_m} \rangle = r\sqrt{\log n_m / n_m},$$

where $r \leq \underline{\gamma}$ for $\underline{\gamma}$ the value given in the proof of Theorem 3.6.5. Since

$$2^{l_m(\beta+1/2)} |\langle f_0, \psi_{l_m k_m} \rangle| \leq r m^{\beta+1/2} \sqrt{\frac{\log n_m}{n_m}},$$

we can ensure f_0 is in any given Hölder ball $\mathcal{H}(\beta, R)$ by letting r be sufficiently small and taking the subsequence n_m to grow fast enough. Let A_n denote the event defined in (3.6.1). We have that on A_n, the posterior distribution $\Pi'(\cdot \mid Y^{(n_m)})$ assigns the (l_m, k_m) coordinate to the Dirac mass component of the distribution. Consequently, by the choice of (n_m),

$$\begin{aligned} &\mathbb{E}_0 \Pi'(\, ||f - f_0||_{\mathcal{M}} \leq M_{n_m} {n_m}^{-1/2} \mid Y^{(n_m)}) \\ &\quad = \mathbb{E}_0 \Pi'(\{||f - f_0||_{\mathcal{M}} \leq M_{n_m} {n_m}^{-1/2}\} \cap A_{n_m} \mid Y^{(n_m)}) + o(1) \\ &\quad \leq \mathbb{E}_0 \Pi'(\{|f_{l_m k_m} - f_{0, l_m k_m}| \leq M_{n_m} w_{l_m} {n_m}^{-1/2}\} \cap A_{n_m} \mid Y^{(n_m)}) + o(1) \\ &\quad = \mathbb{E}_0 \Pi'(\{r\sqrt{\log n_m / n_m} \leq M_{n_m} w_{l_m} n_m^{-1/2}\} \cap A_{n_m} \mid Y^{(n_m)}) + o(1) \\ &\quad \leq \mathbb{E}_0 \Pi'(r\sqrt{\log n_m} / w_{l_m} \leq M_{n_m} \mid Y^{(n_m)}) + o(1) \\ &\quad = o(1) \end{aligned}$$

for any sequence M_n such that $M_{n_m} = o(w_{l_m}^{-1}\sqrt{\log n_m})$ as $m \to \infty$. □

3.7 Proof of Theorem 3.6.1

This is proved in exactly the same manner as Theorem 2.3 from [52] and is included only for completeness. Throughout this section we assume that $\rho_k = k^{-p}$ to simply notation. Recall also that h_n is the function defined in Section 3.6.3. By Markov's inequality and

Theorem 3.6.6,

$$\sup_{f_0\in\mathcal{Q}(\beta,R)}\mathbb{E}_0\Pi_{\hat{\alpha}_n}(||f-f_0||_{H(\delta)}\geq M_nL_nn^{-1/2}|Y)\leq\frac{n}{M_n^2L_n^2}\sup_{f_0\in B(\beta,R)}\mathbb{E}_0\sup_{\underline{\alpha}_n<\alpha<\overline{\alpha}_n\wedge\log n}R_n(\alpha)+o(1), \tag{3.7.1}$$

where

$$R_n(\alpha)=\int||f-f_0||^2_{H(\delta)}\,\Pi_\alpha(df|Y)$$

is the posterior risk. Letting $\hat{f}_{\alpha,k}=nk^p(k^{2\alpha+2p+1}+n)^{-1}Y_k$ and using the explicit form (3.6.10) for the posterior distribution yields

$$\begin{aligned}R_n(\alpha)&=\int\sum_{k=1}^\infty k^{-2p-1}(\log k)^{-2\delta}|f_k-f_{0,k}|^2d\Pi_\alpha(df|Y)\\&=\sum_{k=1}^\infty k^{-2p-1}(\log k)^{-2\delta}\left((\hat{f}_{\alpha,k}-f_{0,k})^2+\frac{k^{2p}}{k^{2\alpha+2p+1}+n}\right).\end{aligned}$$

Using the triangle inequality, we can then split

$$\begin{aligned}\mathbb{E}_0\sup_{\underline{\alpha}_n\leq\alpha\leq\overline{\alpha}_n\wedge\log n}R_n(\alpha)\leq\mathbb{E}_0&\sup_{\underline{\alpha}_n\leq\alpha\leq\overline{\alpha}_n\wedge\log n}\left|\sum_{k=1}^\infty\frac{(\hat{f}_{\alpha,k}-f_{0,k})^2}{k^{2p+1}(\log k)^{2\delta}}-\mathbb{E}_0\sum_{k=1}^\infty\frac{(\hat{f}_{\alpha,k}-f_{0,k})^2}{k^{2p+1}(\log k)^{2\delta}}\right|\\&+\sup_{\underline{\alpha}_n\leq\alpha\leq\overline{\alpha}_n\wedge\log n}\mathbb{E}_0\sum_{k=1}^\infty\frac{(\hat{f}_{\alpha,k}-f_{0,k})^2}{k^{2p+1}(\log k)^{2\delta}}\\&+\sup_{\underline{\alpha}_n\leq\alpha\leq\overline{\alpha}_n\wedge\log n}\sum_{k=1}^\infty\frac{1}{k(\log k)^{2\delta}(k^{2\alpha+2p+1}+n)}.\end{aligned} \tag{3.7.2}$$

Bound for the expected posterior risk

Using the definition of $\hat{f}_{\alpha,k}$ and a bias-variance expansion, the second term of (3.7.2) equals

$$\sup_{\underline{\alpha}_n\leq\alpha\leq\overline{\alpha}_n\wedge\log n}\left\{\sum_{k=1}^\infty\frac{k^{4\alpha+4p+2}f_{0,k}^2}{k^{2p+1}(\log k)^{2\delta}(k^{2\alpha+2p+1}+n)^2}+n\sum_{k=1}^\infty\frac{1}{k(\log k)^{2\delta}(k^{2\alpha+2p+1}+n)^2}\right\}. \tag{3.7.3}$$

Now the second term in (3.7.3) is smaller than the third term in (3.7.2), which is itself smaller than Cn^{-1}, where $C=C(\delta)=\sum_{k=1}^\infty k^{-1}(\log k)^{-2\delta}$ is finite for $\delta>1/2$.

For the first term in (3.7.3), consider the sets

$$P_n=\{f_0\in\mathcal{Q}(\beta,R):f_{0,k}\neq0\text{ for some }k\geq2\}$$

$$Q_n=\{f_0\in\mathcal{Q}(\beta,R):f_{0,k}=0\text{ for all }k\geq2\}$$

and note that by Lemma 2.1(iv) of [52], we have $\overline{\alpha}_n < \log n$ if $f_0 \in P_n$. If $f_0 \in Q_n$, then we trivially have that the first term in (3.7.3) is bounded by $f_{0,1}^2/(1+n)^2 \leq R/n^2$. If $f_0 \in P_n$, we have $\alpha \in [\underline{\alpha}_n, \overline{\alpha}_n]$ and we split up the sum into three parts. Firstly, for $t_n = n^{1/(2\beta+2p+1)}$,

$$\sum_{k \geq t_n} \frac{k^{4\alpha+2p+1} f_{0,k}^2}{(\log k)^{2\delta}(k^{2\alpha+2p+1}+n)^2} \leq \sum_{k \geq t_n} \frac{R}{k^{2\beta+2p+2}(\log k)^{2\delta}} \leq C(\delta) R n^{-1},$$

since $\delta > 1/2$. Letting $s_n(\alpha) = n^{1/(2\alpha+2p+1)}$, we have

$$\sum_{1 \leq k \leq s_n} \frac{k^{4\alpha+2p+1} f_{0,k}^2}{(\log k)^{2\delta}(k^{2\alpha+2p+1}+n)^2} \leq \frac{R}{n^2} + \frac{1}{1+2\alpha+2p} h_n(\alpha) n^{-2+\frac{1}{2\alpha+2p+1}} \log n \max_{2 \leq k \leq s_n} \frac{k^{2\alpha+2p}}{(\log k)^{2\delta+1}}. \tag{3.7.4}$$

Now in the case of general f_0, we have $\max_{2 \leq k \leq s_n} k^{2\alpha+2p}/(\log k)^{2\delta+1} \leq \max_{2 \leq k \leq s_n} k^{2\alpha+2p} = n^{(2\alpha+2p)/(2\alpha+2p+1)}$. Since $\alpha \leq \overline{\alpha}_n$ we have that $h_n(\alpha) \leq L(\log n)^2$ and so the right-hand side of (3.7.4) is bounded by $n^{-1}(\log n)^3$.

Now suppose that $f_0 \in \mathcal{Q}_{SS}(\beta, R)$. Basic calculus shows that the function $x \mapsto x^{2\alpha+2p}(\log x)^{-2\delta-1}$ on $(0,\infty)$ has a single minimum at $e^{(2\delta+1)/(2\alpha+2p)}$ and is monotonic on either side of it. When restricted to the interval $[2, s_n]$, the function thus attains its maximum value at either 2 or s_n. By Lemma 3.6.7, $\overline{\alpha}_n \leq \beta + C_0 = C$, so that

$$\max_{2 \leq k \leq s_n} \frac{k^{2\alpha+2p}}{(\log k)^{2\delta+1}} = \max\left\{ \frac{2^{2\alpha+2p}}{(\log 2)^{2\delta+1}}, C' \frac{n^{\frac{2\alpha+2p}{2\alpha+2p+1}}}{(\log n)^{\frac{2\alpha+2p}{2\alpha+2p+1}+2\delta+1}} \right\} \leq C'' \max\left\{ 1, \frac{n^{\frac{2\alpha+2p}{2\alpha+2p+1}}}{(\log n)^{2\delta+1}} \right\}.$$

Again using that $h_n(\alpha) \leq L(\log n)^2$ yields that the right-hand side of (3.7.4) is bounded by a constant times

$$\max\left\{ n^{-2+\frac{1}{2\alpha+2p+1}} (\log n)^3, \frac{1}{n(\log n)^{2\delta-2}} \right\} = O(n^{-1})$$

for all $\alpha > \underline{\alpha}_n$ and $\delta \geq 1$. Now since $x \mapsto x/(c+x)$ is increasing for every $c > 0$, we have that the sum in (3.7.4) is maximized by the choice $\alpha = \overline{\alpha}_n$ and thus we can take $s_n = s_n(\overline{\alpha}_n)$ since the right-hand side is independent of α.

Let $J = J(n)$ be the smallest integer such that $\overline{\alpha}_n \wedge (\log n)/(1+1/\log n)^J \leq \beta$, which is bounded above by a multiple of $(\log n)(\log\log n)$ for all $\beta > 0$. Partition the summation range using the following numbers

$$b_j = 1 + 2\frac{\overline{\alpha}_n}{(1+1/\log n)^j} + 2p, \qquad j = 0, ..., J,$$

which form a decreasing sequence. We then have

$$\sum_{k=2\vee s_n}^{n^{1/(2\beta+1)}} \frac{k^{4\alpha+2p+1} f_{0,k}^2}{(\log k)^{2\delta}(k^{2\alpha+2p+1}+n)^2} \leq \sum_{j=0}^{J-1} \sum_{k=n^{1/b_j}}^{n^{1/b_{j+1}}} \frac{f_{0,k}^2}{k^{2p+1}(\log k)^{2\delta}} \leq 4 \sum_{j=0}^{J-1} \sum_{k=n^{1/b_j}}^{n^{1/b_{j+1}}} \frac{n k^{b_j-2p} f_{0,k}^2}{k(\log k)^{2\delta}(k^{b_{j+1}}+n)^2}.$$

Note that the right-hand side is independent of α and that since $(b_j - b_{j+1}) \log n = b_{j+1} - 1$, we have that $k^{b_j - b_{j+1}} \leq n^{1/\log n} = e$ for $n^{1/b_j} \leq k \leq n^{1/b_{j+1}}$. Consequently, the above is bounded by a multiple of

$$\begin{aligned}
&\frac{1}{n} \sum_{j=0}^{J-1} \sum_{k=n^{1/b_j}}^{n^{1/b_{j+1}}} \frac{n^2 k^{b_{j+1}-2p} f_{0,k}^2 (\log k) k^{b_j - b_{j+1}}}{k(\log k)^{2\delta+1}(k^{b_{j+1}}+n)^2} \\
&\quad \lesssim \frac{\log n}{n} \sum_{j=0}^{J-1} \frac{n^{1/b_{j+1}}}{b_{j+1}} h_n\left(\frac{b_{j+1}}{2} - \frac{1}{2} - p\right) \max_{n^{1/b_j} \leq k \leq n^{1/b_{j+1}}} \frac{1}{k(\log k)^{2\delta+1}} \qquad (3.7.5) \\
&\quad \lesssim \frac{L(\log n)^{2-2\delta}}{n} \sum_{j=0}^{J-1} n^{1/b_{j+1} - 1/b_j} \frac{b_j^{2\delta+1}}{b_{j+1}},
\end{aligned}$$

since by the definition of $\overline{\alpha}_n$ and b_j, we have that $h_n((b_{j+1}-1)/2) \leq L(\log n)^2$. Since $b_j > 1$,

$$n^{1/b_{j+1} - 1/b_j} = \exp\left(\frac{(b_j - b_{j+1}) \log n}{b_j b_{j+1}}\right) = \exp\left(\frac{b_{j+1} - 1}{b_j b_{j+1}}\right) \leq \exp\left(\frac{1}{b_j}\right) \leq e.$$

If $f_0 \in \mathcal{Q}(\beta, R)$, note that $b_0 \lesssim \log n$ and $b_j / b_{j+1} \leq C(1 + 1/\log n) \leq C'$. The above sum is therefore bounded by $Le(\log n)^2 n^{-1} J(n) \lesssim (\log n)^3 (\log\log n) n^{-1} = L_n^2 n^{-1}$. If in addition $f_0 \in \mathcal{Q}_{SS}(\beta, R)$, then $\overline{\alpha}_n \leq \beta + K_2/(\log n)$ for n sufficiently large by Lemma 3.6.7. We can thus bound (3.7.5) by a multiple of

$$n^{-1}(\log n)^{2-2\delta} \sum_{j=0}^{J-1} b_j^{2\delta} \leq n^{-1}(\log n)^{4-2\delta} J(n)(1 + 2\overline{\alpha}_n)^{2\delta} \leq C n^{-1}$$

for $\delta \geq 2$.

Bounds for the centered posterior risk

We now turn our attention to the first term in (3.7.2), which can be controlled using empirical process techniques. Note that after expanding out the terms, this sum is equal

to

$$\begin{aligned}\mathbb{E}_0 \sup_{\alpha\in[\underline{\alpha}_n,\overline{\alpha}_n\wedge\log n]}&\left|\sum_{k=1}^{\infty}\frac{n(Z_k^2-1)}{k(\log k)^{2\delta}(k^{2\alpha+2p+1}+n)^2}-\sum_{k=1}^{\infty}\frac{2\sqrt{n}k^{2\alpha+p}f_{0,k}Z_k}{(\log k)^{2\delta}(k^{2\alpha+2p+1}+n)^2}\right| \\ &=: \sup_{\alpha\in[\underline{\alpha}_n,\overline{\alpha}_n\wedge\log n]}\left|\frac{\mathbb{V}(\alpha)}{n}-\frac{2\mathbb{W}(\alpha)}{\sqrt{n}}\right|,\end{aligned} \tag{3.7.6}$$

where the Z_k are i.i.d. standard normals. The two terms can be controlled separately. By Corollary 2.2.5 of [85],

$$\mathbb{E}_0 \sup_{\alpha\in[\underline{\alpha}_n,\overline{\alpha}_n\wedge\log n)}|\mathbb{V}(\alpha)| \lesssim \sup_{\alpha\in[\underline{\alpha}_n\infty)}\sqrt{\mathrm{var}_0\mathbb{V}(\alpha)}+\int_0^{T_n}\sqrt{N([\underline{\alpha}_n,\overline{\alpha}_n\wedge\log n),d_n,\varepsilon)}d\varepsilon,$$

where $d_n^2(\alpha_1,\alpha_2)=\mathrm{var}_0(\mathbb{V}(\alpha_1)-\mathbb{V}(\alpha_2))$ and T_n is the d_n-diameter of $[\underline{\alpha}_n,\overline{\alpha}_n\wedge\log n)$. Now

$$\mathrm{var}_0\mathbb{V}(\alpha)=2n^4\sum_{k=1}^{\infty}\frac{1}{k^2(\log k)^{4\delta}(k^{2\alpha+2p+1}+n)^4}\le\sum_{k=1}^{\infty}\frac{2}{k^2}<\infty.$$

In particular, this implies that $T_n\le C$, for some C independent of n and $\hat{\alpha}_n$. Note that for $0<\alpha_1<\alpha_2$

$$\begin{aligned}\mathrm{var}_0(\mathbb{V}(\alpha_1)-\mathbb{V}(\alpha_2))&=\sum_{k=2}^{\infty}\frac{n^4}{k^2(\log k)^{4\delta}}\left(\frac{1}{(k^{2\alpha_1+2p+1}+n)^2}-\frac{1}{(k^{2\alpha_2+2p+1}+n)^2}\right)^2\mathrm{var}(Z_k^2)\\ &\le 2n^4\sum_{k=2}^{\infty}\frac{1}{k^2(\log k)^{4\delta}(k^{2\alpha_1+2p+1}+n)^4}\lesssim n^4\sum_{k=2}^{\infty}k^{-8\alpha_1-6}\lesssim n^4 2^{-8\alpha_1}.\end{aligned}$$

So for $\varepsilon>0$, we can cover $[K\log(n/\varepsilon),\infty)$ by a single ε-ball (by letting $\alpha_2\to\infty$), for some $K>0$. By an analogue of Lemma 6.1 in [52], we have the bound $d_n(\alpha_1,\alpha_2)\le C(\delta)|\alpha_1-\alpha_2|n^{-\frac{4p+1}{4\alpha_2+4p+2}}$ for $\delta>3/4$. Combining these facts yields that $N([\underline{\alpha}_n,\overline{\alpha}_n\wedge\log n),d_n,\varepsilon)\le C(\delta)n^{-\frac{4p+1}{4\underline{\alpha}_n+4p+2}}\varepsilon^{-1}\log(n/\varepsilon)$. We therefore have

$$\mathbb{E}_0 \sup_{\alpha\in[\underline{\alpha}_n,\overline{\alpha}_n\wedge\log n)}|\mathbb{V}(\alpha)|\le C+C(\delta)n^{-\frac{4p+1}{8\underline{\alpha}_n+8p+4}}\int_0^{C}\varepsilon^{-1/2}\sqrt{\log(n/\varepsilon)}d\varepsilon. \tag{3.7.7}$$

Using the substitution $y^2=\log(n/\varepsilon)$, integrating by parts and then applying the standard tail bound for Gaussian integrals yields

$$\begin{aligned}\int_0^{C}\varepsilon^{-1/2}\sqrt{\log(n/\varepsilon)}d\varepsilon=2\sqrt{n}\int_{\sqrt{\log(n/C)}}^{\infty}y^2e^{-\frac{1}{2}y^2}dy&=2\sqrt{C\log(n/C)}+2\sqrt{n}\int_{\sqrt{\log(n/C)}}^{\infty}e^{-\frac{1}{2}y^2}dy\\ &\le 2\sqrt{C\log(n/C)}+2\sqrt{C}(\log(n/C))^{-1/2}\\ &\le C''\sqrt{\log n}.\end{aligned}$$

Using this and that $\underline{\alpha}_n \leq \beta$, (3.7.7) is bounded above by some constant. In conclusion, the first term on the right-hand side of (3.7.6) is $O(n^{-1})$.

We perform a similar calculation on $\mathbb{W}$. Consider now $\alpha \in [\underline{\alpha}_n, \overline{\alpha}_n]$ and note that

$$\operatorname{var}_0\left(\frac{\mathbb{W}(\alpha)}{\sqrt{n}}\right) = \sum_{k=1}^{\infty} \frac{nk^{4\alpha+2p} f_{0,k}^2}{(\log k)^{4\delta}(k^{2\alpha+2p+1}+n)^4}.$$

Splitting the above sum as above and using the definition of $h_n(\alpha)$ we have

$$\begin{aligned}\sum_{k \leq n^{1/(2\alpha+2p+1)}} & \frac{nk^{4\alpha+2p} f_{0,k}^2}{(\log k)^{4\delta}(k^{2\alpha+2p+1}+n)^4} \\ & \leq \frac{f_{0,k}^2}{n^3} + h_n(\alpha) n^{-1+\frac{1}{2\alpha+2p+1}} \log n \max_{2 \leq k \leq n^{1/(2\alpha+2p+1)}} \frac{k^{2\alpha+2p-1}}{(k^{2\alpha+2p+1}+n)^2 (\log k)^{4\delta+1}}.\end{aligned}$$

If $2\alpha + 2p - 1 > 0$, then the second term above is bounded by a multiple of

$$n^{-1+\frac{1}{2\alpha+2p+1}} (\log n)^3 n^{-2+\frac{2\alpha+2p-1}{2\alpha+2p+1}} = n^{-2-\frac{1}{2\alpha+2p+1}} (\log n)^3 = O(n^{-2})$$

for all $\alpha > 0$. If $2\alpha + 2p - 1 \leq 0$, then the sum is similarly $O(n^{-2})$. For the upper part of the sum, recalling that $\alpha \in [\underline{\alpha}_n, \overline{\alpha}_n \wedge \log n]$,

$$\begin{aligned}\sum_{k > n^{1/(2\alpha+2p+1)}} \frac{nk^{4\alpha+2p} f_{0,k}^2}{(\log k)^{4\delta}(k^{2\alpha+2p+1}+n)^4} &\leq \frac{1}{n} \sum_{k > n^{1/(2\alpha+2p+1)}} \frac{n^2 k^{2\alpha+1} f_{0,k}^2 \log k}{(k^{2\alpha+2p+1}+n)^2} \frac{1}{n^{2-\frac{2\alpha+2p-1}{2\alpha+2p+1}} (\log k)^{4\delta+1}} \\ &\leq n^{-2-1/(2\alpha+2p+1)} h_n(\alpha) (\log n)^{-4\delta} (1+2\alpha+2p)^{4\delta} \\ &\leq CLn^{-2-\frac{1}{2\alpha+2p+1}} (\log n)^{3-4\delta} \leq C' n^{-2}.\end{aligned}$$

In conclusion, we have that $\sup_{\alpha \in [\underline{\alpha}_n, \overline{\alpha}_n \wedge \log n)} \sqrt{\operatorname{var}_0 n^{-1/2} \mathbb{W}(\alpha)} \leq n^{-1}$.

By following the proof of Lemma 6.1 in [52], we similarly recover that in our case the intrinsic covariance metric satisfies $\tilde{d}_n^2(\alpha_1, \alpha_2)^2 = \operatorname{var}_0(n^{-1/2}(\mathbb{W}(\alpha_1) - \mathbb{W}(\alpha_2)) \lesssim (\alpha_1 - \alpha_2)^2 n^{-2}$. Using the same reasoning as above, we have that $N([\underline{\alpha}_n, \overline{\alpha}_n \wedge \log n], \tilde{d}_n, \varepsilon) \leq Kn^{-1} \log(n/\varepsilon)/\varepsilon$ so that

$$\mathbb{E}_0 \sup_{\alpha \in [\underline{\alpha}_n, \overline{\alpha}_n \wedge \log n]} \left| \frac{\mathbb{W}(\alpha)}{\sqrt{n}} \right| \lesssim n^{-1} + n^{-1/2} \int_0^C \varepsilon^{-1/2} \sqrt{\log(n/\varepsilon)} d\varepsilon \lesssim n^{-1}.$$

We thus have that (3.7.6) is $O(n^{-1})$ and consequently so is (3.7.2), thereby completing the proof.

We note that f_0 being self-similar was only required to establish that the first term in (3.7.3) was $O(n^{-1})$. In the case H^{-p-s}, $s > 1/2$, following the steps above with this weaker norm yields the required bound.

Bibliography

[1] Abramovich, F., Sapatinas, T., and Silverman, B. W. Wavelet thresholding via a Bayesian approach. *J. R. Stat. Soc. Ser. B Stat. Methodol. 60*, 4 (1998), 725–749. (Cited on page 84.)

[2] Agapiou, S., Larsson, S., and Stuart, A. M. Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Process. Appl. 123*, 10 (2013), 3828–3860. (Cited on page 30.)

[3] Agapiou, S., Stuart, A. M., and Zhang, Y.-X. Bayesian posterior contraction rates for linear severely ill-posed inverse problems. *J. Inverse Ill-Posed Probl. 22*, 3 (2014), 297–321. (Cited on page 27.)

[4] Arbel, J., Gayraud, G., and Rousseau, J. Bayesian optimal adaptive estimation using a sieve prior. *Scand. J. Stat. 40*, 3 (2013). (Cited on pages 31, 37, and 38.)

[5] Bernstein, S. N. *Theory of probability.* Moscow, 1917. (Cited on page 21.)

[6] Bickel, P. J., and Kleijn, B. J. K. The semiparametric Bernstein-von Mises theorem. *Ann. Statist. 40*, 1 (2012), 206–237. (Cited on page 71.)

[7] Bickel, P. J., and Ritov, Y. Nonparametric estimators which can be "plugged-in". *Ann. Statist. 31*, 4 (2003), 1033–1053. (Cited on page 71.)

[8] Billingsley, P., and Topsøe, F. Uniformity in weak convergence. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 7* (1967), 1–16. (Cited on page 23.)

[9] Birgé, L. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete 65*, 2 (1983), 181–237. (Cited on page 19.)

[10] Birgé, L. Robust testing for independent nonidentically distributed variables and Markov chains. In *Specifying statistical models (Louvain-la-Neuve, 1981)*, vol. 16 of *Lecture Notes in Statist.* Springer, New York, 1983, pp. 134–162. (Cited on page 19.)

[11] Bissantz, N., Hohage, T., Munk, A., and Ruymgaart, F. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal. 45*, 6 (2007), 2610–2636. (Cited on pages 24, 31, and 34.)

[12] Bontemps, D. Bernstein-von Mises theorems for Gaussian regression with increasing number of regressors. *Ann. Statist. 39*, 5 (2011), 2557–2584. (Cited on page 71.)

[13] Borell, C. The Brunn-Minkowski inequality in Gauss space. *Invent. Math. 30*, 2 (1975), 207–216. (Cited on page 44.)

[14] Boucheron, S., and Gassiat, E. A Bernstein-von Mises theorem for discrete probability distributions. *Electron. J. Stat. 3* (2009), 114–148. (Cited on page 71.)

[15] Brown, L. D., and Low, M. G. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist. 24*, 6 (1996), 2384–2398. (Cited on page 26.)

[16] Bull, A. D. Honest adaptive confidence bands and self-similar functions. *Electron. J. Stat. 6* (2012), 1490–1516. (Cited on pages 71, 76, and 77.)

[17] Butucea, C., and Tsybakov, A. B. Sharp optimality in density deconvolution with dominating bias. II. *Teor. Veroyatn. Primen. 52*, 2 (2007), 336–349. (Cited on pages 31 and 36.)

[18] Castillo, I. A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields 152*, 1-2 (2012), 53–99. (Cited on page 71.)

[19] Castillo, I. On Bayesian supremum norm contraction rates. *Ann. Statist. 42*, 5 (2014), 2058–2091. (Cited on page 93.)

[20] Castillo, I., Kerkyacharian, G., and Picard, D. Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields 158*, 3-4 (2014), 665–710. (Cited on page 56.)

[21] Castillo, I., and Nickl, R. Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist. 41*, 4 (2013), 1999–2028. (Cited on pages 22, 23, 68, 71, 74, 76, 83, 90, 94, and 99.)

[22] Castillo, I., and Nickl, R. On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist. 42*, 5 (2014), 1941–1969. (Cited on pages 22, 23, 68, 69, 73, 74, 76, 85, 93, 94, 104, and 105.)

[23] Castillo, I., and Rousseau, J. A Bernstein–von Mises theorem for smooth functionals in semiparametric models. arXiv:1305.4482, 2013. (Cited on page 71.)

[24] Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. W. Bayesian linear regression with sparse priors. arXiv:1403.0735, 2014. (Cited on page 80.)

[25] Castillo, I., and van der Vaart, A. Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *Ann. Statist. 40*, 4 (2012), 2069–2101. (Cited on page 80.)

[26] Cavalier, L. Nonparametric statistical inverse problems. *Inverse Problems 24*, 3 (2008), 034004, 19. (Cited on pages 26, 34, and 72.)

[27] Cox, D. D. An analysis of Bayesian inference for nonparametric regression. *Ann. Statist. 21*, 2 (1993), 903–923. (Cited on pages 20, 22, and 68.)

[28] Da Prato, G. *An introduction to infinite-dimensional analysis*. Universitext. Springer-Verlag, Berlin, 2006. Revised and extended from the 2001 original by Da Prato. (Cited on page 33.)

[29] Dembo, A., Mayer-Wolf, E., and Zeitouni, O. Exact behavior of Gaussian seminorms. *Statist. Probab. Lett. 23*, 3 (1995), 275–280. (Cited on page 53.)

[30] Diaconis, P., and Freedman, D. On the consistency of Bayes estimates. *Ann. Statist. 14*, 1 (1986), 1–67. With a discussion and a rejoinder by the authors. (Cited on pages 16, 17, and 80.)

[31] Doob, J. L. Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13. Centre National de la Recherche Scientifique, Paris, 1949, pp. 23–27. (Cited on pages 17 and 21.)

[32] Folland, G. B. *Real analysis*, second ed. Pure and Applied Mathematics (New York). John Wiley & Sons Inc., New York, 1999. Modern techniques and their applications, A Wiley-Interscience Publication. (Cited on pages 34 and 42.)

[33] Freedman, D. On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist. 27*, 4 (1999), 1119–1140. (Cited on pages 22, 68, and 71.)

[34] Freedman, D. A. On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist. 34* (1963), 1386–1403. (Cited on page 16.)

[35] Ghosal, S. Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli 5*, 2 (1999), 315–331. (Cited on page 71.)

[36] Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. Convergence rates of posterior distributions. *Ann. Statist. 28*, 2 (2000), 500–531. (Cited on pages 18, 29, 30, 33, 40, and 61.)

[37] Ghosal, S., and van der Vaart, A. Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist. 35*, 1 (2007), 192–223. (Cited on page 19.)

[38] Ghosal, S., and van der Vaart, A. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist. 35*, 2 (2007), 697–723. (Cited on page 43.)

[39] Giné, E., and Nickl, R. Uniform limit theorems for wavelet density estimators. *Ann. Probab. 37*, 4 (2009), 1605–1646. (Cited on page 82.)

[40] Giné, E., and Nickl, R. Confidence bands in density estimation. *Ann. Statist. 38*, 2 (2010), 1122–1170. (Cited on pages 71, 76, 77, 83, and 86.)

[41] Giné, E., and Nickl, R. Rates on contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist. 39*, 6 (2011), 2883–2911. (Cited on pages 19, 29, 30, 35, 41, 42, and 45.)

[42] Hoffmann, M., and Nickl, R. On adaptive inference and confidence bands. *Ann. Statist. 39*, 5 (2011), 2383–2409. (Cited on pages 71, 76, and 102.)

[43] Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. On adaptive posterior concentration rates. arXiv:1305.5270, 2013. (Cited on pages 80, 82, 91, 92, 100, and 101.)

[44] Hoffmann-Jørgensen, J., Shepp, L. A., and Dudley, R. M. On the lower tail of Gaussian seminorms. *Ann. Probab. 7*, 2 (1979), 319–342. (Cited on pages 53 and 55.)

[45] Holmes, C., and Denison, G. Bayesian wavelet analysis with a model complexity prior. In *Bayesian Statistics 6: proceedings of the sixth Valencia international meeting*. Clarendon Press, Oxford, 1999, pp. 769–776. (Cited on page 30.)

[46] Huang, T.-M. Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist. 32*, 4 (2004), 1556–1593. (Cited on pages 29 and 37.)

[47] Johnstone, I. M. High dimensional Bernstein–von Mises: simple examples. In *Borrowing strength: theory powering applications—a Festschrift for Lawrence D. Brown*, vol. 6 of *Inst. Math. Stat. Collect.* Inst. Math. Statist., Beachwood, OH, 2010, pp. 87–98. (Cited on pages 22 and 68.)

[48] Johnstone, I. M., Kerkyacharian, G., Picard, D., and Raimondo, M. Wavelet deconvolution in a periodic setting. *J. R. Stat. Soc. Ser. B Stat. Methodol. 66*, 3 (2004), 547–573. (Cited on pages 26, 32, 42, and 44.)

[49] Johnstone, I. M., and Silverman, B. W. Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist. 18*, 1 (1990), 251–280. (Cited on page 27.)

[50] Kleijn, B. J. K., and van der Vaart, A. W. Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist. 34*, 2 (2006), 837–877. (Cited on page 12.)

[51] Kleijn, B. J. K., and van der Vaart, A. W. The Bernstein-Von-Mises theorem under misspecification. *Electron. J. Stat. 6* (2012), 354–381. (Cited on pages 12, 96, 97, and 98.)

[52] Knapik, B. T., Szabó, B. T., van der Vaart, A. W., and van Zanten, J. H. Bayes procedures for adaptive inference in inverse problems for the white noise model. arXiv:1209.3628, 2012. (Cited on pages 31, 39, 77, 78, 90, 91, 94, 99, 105, 107, 109, and 110.)

[53] Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. Bayesian inverse problems with Gaussian priors. *Ann. Statist. 39*, 5 (2011), 2626–2657. (Cited on pages 20, 30, 39, 40, 70, 76, 77, and 88.)

[54] Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. Bayesian recovery of the initial condition for the heat equation. *Comm. Statist. Theory Methods 42*, 7 (2013), 1294–1313. (Cited on pages 27, 30, 39, 40, and 41.)

[55] Knowles, D., and Ghahramani, Z. Nonparametric Bayesian sparse factor models with application to gene expression modeling. *Ann. Appl. Stat. 5*, 2B (2011), 1534–1552. (Cited on page 30.)

[56] Kueh, A. Locally adaptive density estimation on the unit sphere using needlets. *Constr. Approx. 36*, 3 (2012), 433–458. (Cited on pages 84 and 85.)

[57] Kuelbs, J., and Li, W. V. Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal. 116*, 1 (1993), 133–157. (Cited on pages 41, 53, and 55.)

[58] Laplace, P.-S. *Mémoire sur les formules qui sont fonctions de très grands nombres et sur leurs applications aux probabilités*, vol. 12 of *Oeuvres de Laplace*. 1810. (Cited on page 21.)

[59] Laurent, B., and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist. 28*, 5 (2000), 1302–1338. (Cited on page 104.)

[60] Le Cam, L. *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York, 1986. (Cited on pages 18, 21, 26, 68, 90, and 95.)

[61] Leahu, H. On the Bernstein-von Mises phenomenon in the Gaussian white noise model. *Electron. J. Stat. 5* (2011), 373–404. (Cited on pages 20, 22, 68, 70, 71, and 75.)

[62] Ledoux, M. *The concentration of measure phenomenon*, vol. 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001. (Cited on pages 45 and 100.)

[63] Lehmann, E. L., and Casella, G. *Theory of point estimation*, second ed. Springer Texts in Statistics. Springer-Verlag, New York, 1998. (Cited on page 13.)

[64] Lounici, K., and Nickl, R. Global uniform risk bounds for wavelet deconvolution estimators. *Ann. Statist. 39*, 1 (2011), 201–231. (Cited on pages 30 and 31.)

[65] Meyer, Y. *Wavelets and operators*, vol. 37 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1992. Translated from the 1990 French original by D. H. Salinger. (Cited on pages 41, 73, and 74.)

[66] Neal, R. M. Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist. 9*, 2 (2000), 249–265. (Cited on page 30.)

[67] Nickl, R. Discussion of: "Frequentist coverage of adaptive nonparametric Bayesian credible sets". *Ann. Statist.* (2014), To appear. (Cited on page 77.)

[68] Nickl, R., and Szabó, B. T. A sharp adaptive confidence ball for self-similar functions. arXiv:1406.3994, 2014. (Cited on pages 76 and 77.)

[69] Pensky, M., and Vidakovic, B. Adaptive wavelet estimator for nonparametric density deconvolution. *Ann. Statist. 27*, 6 (1999), 2033–2053. (Cited on pages 32, 42, and 44.)

[70] Petrone, S., Rousseau, J., and Scricciolo, C. Bayes and empirical Bayes: do they merge? *Biometrika 101*, 2 (2014), 285–302. (Cited on page 80.)

[71] Ray, K. Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat. 7* (2013), 2516–2549. (Cited on page 7.)

[72] Ray, K. Bernstein–von Mises theorems for adaptive Bayesian nonparametric procedures. arXiv:1407.3397, 2014. (Cited on page 7.)

[73] Rivoirard, V., and Rousseau, J. Bernstein-von Mises theorem for linear functionals of the density. *Ann. Statist. 40*, 3 (2012), 1489–1523. (Cited on page 71.)

[74] Rudin, W. *Functional analysis*, second ed. International Series in Pure and Applied Mathematics. McGraw-Hill Inc., New York, 1991. (Cited on page 25.)

[75] Schwartz, L. On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 4* (1965), 10–26. (Cited on page 17.)

[76] Shen, X., and Wasserman, L. Rates of convergence of posterior distributions. *Ann. Statist. 29*, 3 (2001), 687–714. (Cited on pages 29 and 37.)

[77] Stuart, A. M. The Bayesian approach to inverse problems. arXiv:1302.6989, 2013. (Cited on page 31.)

[78] Sytaya, G. N. On some asymptotic representations of the gaussian measure in a hilbert space. *Theory of Stochastic Processes,* Ukrainian Academy of Sciences*, 2* (1974), 93–104. (Cited on page 53.)

[79] Szabó, B. T., van der Vaart, A. W., and van Zanten, J. H. Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* (2014), To appear. (Cited on pages 20, 70, 71, 76, 77, and 98.)

[80] Szabó, B. T., van der Vaart, A. W., and van Zanten, J. H. Honest Bayesian confidence sets for the L^2-norm. *J. Statist. Plann. Inference* (2014), To appear. (Cited on page 77.)

[81] van der Vaart, A. W. *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. (Cited on pages 21, 22, 68, and 94.)

[82] van der Vaart, A. W., and van Zanten, J. H. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist. 36*, 3 (2008), 1435–1463. (Cited on pages 19, 29, 40, 54, and 56.)

[83] van der Vaart, A. W., and van Zanten, J. H. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, vol. 3 of *Inst. Math. Stat. Collect.* Inst. Math. Statist., Beachwood, OH, 2008, pp. 200–222. (Cited on pages 40 and 53.)

[84] van der Vaart, A. W., and van Zanten, J. H. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist. 37*, 5B (2009), 2655–2675. (Cited on page 40.)

[85] van der Vaart, A. W., and Wellner, J. A. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics. (Cited on page 109.)

[86] von Mises, R. *Wahrscheinlichkeitsrechnung*. Deuticke, Vienna, 1931. (Cited on page 21.)

[87] Zhao, L. H. Bayesian aspects of some nonparametric problems. *Ann. Statist. 28*, 2 (2000), 532–552. (Cited on page 37.)