

Hierarchical Besov-Laplace priors for spatially inhomogeneous binary classification

Patric Dolmeta¹ and Matteo Giordano^{1*}

¹ESOMAS Department, University of Turin, Corso Unione Sovietica 218/bis, Turin, 10134, Italy.

*Corresponding author(s). E-mail(s): matteo.giordano@unito.it;
Contributing authors: patric.dolmeta@unito.it;

Abstract

We study nonparametric Bayesian binary classification, in the case where the unknown probability response function is possibly spatially inhomogeneous, for example, being generally flat across the domain but presenting localized sharp variations. We consider a hierarchical procedure based on the popular Besov-Laplace priors from inverse problems and imaging, with a carefully tuned hyper-prior on the regularity parameter. We show that the resulting posterior distribution concentrates towards the ground truth at optimal rate, automatically adapting to the unknown regularity. To implement posterior inference in practice, we devise an efficient Markov chain Monte Carlo (MCMC) algorithm based on recent ad-hoc dimension-robust methods for Besov-Laplace priors. We then test the considered approach in extensive numerical simulations, where we obtain a solid corroboration of the theoretical results.

Keywords: Adaptation; Besov spaces; frequentist analysis of Bayesian procedures; minimax-optimal; posterior contraction rates

1 Introduction

Consider the binary classification problem: To predict a 0-1 response Y from the value of a possibly multi-dimensional covariate X . This task is canonically approached by modeling Y , conditionally given X , as a Bernoulli random variable with success probability $p(X)$. Given labeled data $D^{(n)} := \{(Y_i, X_i)\}_{i=1}^n$, the goal is then to obtain an estimate \hat{p} of the ‘probability response function’ $x \mapsto p(x) = \Pr(Y = 1|X = x)$. Using the latter, new unlabeled inputs X_{n+1}, X_{n+2}, \dots , can be classified based on

the plug-in estimates $\hat{p}(X_{n+1}), \hat{p}(X_{n+2}), \dots$; for example, to belong to class 1 if the corresponding probabilities are greater than a certain threshold.

In this article, we consider the nonparametric Bayesian approach to the binary classification problem. This entails modeling p with a suitable prior probability measure Π on a function space, and then combining Π with the data, via Bayes' theorem, to form the posterior distribution $\Pi(\cdot|D^{(n)})$, that is the conditional law of $p|D^{(n)}$. Following the Bayesian paradigm, $\Pi(\cdot|D^{(n)})$ represents the updated belief about p after observing $D^{(n)}$, furnishing point estimates as well as uncertainty quantification. See Section 2 for details, and (Ghosal and van der Vaart 2017, Chapter 1) for a general overview of the methodology. The arguably most widely adopted prior distributions for classification surfaces are the ones based on Gaussian processes, for which there exists an extensive literature providing methodological and computational strategies, e.g. Lenk (1988); Choudhuri et al. (2007); Nickisch and Rasmussen (2008), as well as theoretical performance guarantees, e.g. Ghosal and Roy (2006); van der Vaart and van Zanten (2008, 2009). Further, see (Rasmussen and Williams 2006, Chapter 3). Other commonly used approaches include procedures with Dirichlet process priors, Dirichlet process mixture models, and mixture of experts; see Gelfand and Kuo (1991); Jara et al. (2007); Wang et al. (2010), where many additional references can be found.

In many applications, it is natural to expect that the target probability response function exhibits both general trends as well as localized features, whose correct detection is central to the efficacy of the employed classification procedure. For instance, there may be unknown ‘critical values’ of the covariates that induce sharp variations in the probability of success. To model this scenario, we employ the popular Besov-Laplace priors from the inverse problems and imaging literature, Lassas et al. (2009). These are defined via wavelet series expansions with independent random coefficients following rescaled Laplace distributions, furnishing an infinite-dimensional version of the celebrated total-variation prior of Rudin et al. (1992), while also maintaining a favorable log-concave structure that enhances computation, Bui-Thanh and Ghattas (2015), and analytical study, Agapiou et al. (2021). Besov-Laplace priors are known to give rise to sparse and edge-preserving reconstructions at the level of the maximum-a-posteriori (MAP) estimator, that perform well in the recovery of ‘spatially inhomogeneous’ objects, namely ones that are generally flat and smooth across the domain but exhibit sudden increases or decreases (or even jumps) in some localized areas, like images. See e.g. Leporini and Pesquet (2001); Bioucas-Dias (2006); Vänskä et al. (2009); Sakhaee and Entezari (2015); Kekkonen et al. (2023) and references therein. In contrast, Gaussian priors are suited to model functions with milder variability, and have been shown to be unable to optimally reconstruct more structured signals, Agapiou et al. (2021); Giordano et al. (2022); Agapiou and Wang (2024). See Section 3 for an illustration with synthetic data.

The study of the large sample properties of posterior distributions based on Besov-Laplace priors has been recently initiated in the seminal paper by Agapiou et al. (2021), following the landmark developments in the theory of the frequentist analysis of nonparametric Bayesian procedures over the last two decades, Ghosal et al. (2000); Shen and Wasserman (2001); Ghosal and van der Vaart (2007); van der Vaart and van Zanten (2008); Giné and Nickl (2011). Among their results, they showed that, in the

white noise model, properly tuned Besov-Laplace priors achieve minimax-optimal posterior contraction rates towards ground truths p_0 in the Besov scale B_1^α , $\alpha \geq 0$. These function spaces are defined via wavelet series expansions with ℓ^1 -type penalties on the wavelet coefficients, measuring local variations in an L^1 -sense and allowing for spatial inhomogeneity. See Section 1.1 for definitions, and (Donoho and Johnstone 1998, Section 1) for a detailed description of the connection to the space of bounded variation functions. These results were later extended to various statistical models, including drift estimation for diffusion processes, Giordano and Ray (2022), density estimation, Giordano (2023) and nonlinear inverse problems, Agapiou and Wang (2024). We further refer the reader to the recent investigation by Dolera et al. (2024), as well as to earlier related results by Castillo and Nickl (2014); Ray (2013); Arbel et al. (2013).

In the context of nonparametric binary classification, the recent work by Giordano (2025) showed that Besov-Laplace priors can yield optimal reconstruction of spatially inhomogeneous probability response functions. However, a limitation of their result is the lack of ‘adaptation’, that is, the often unrealistic requirement of knowing the regularity of the ground truth in order to correctly tune the procedure to achieve the optimal rate. See (Ghosal and van der Vaart 2017, Chapter 10) for a general overview of the problem of adaptation in Bayesian nonparametrics. To our knowledge, for Besov-Laplace priors, this issue has so far been investigated only by Giordano (2023) in density estimation, using the hierarchical Bayesian approach, and by Agapiou and Savva (2024) in the white noise model for both hierarchical and empirical methods.

Here, we build on the latter studies, and consider a hierarchical prior for classification surfaces obtained by randomizing the regularity hyper-parameter within a base rescaled Besov-Laplace prior (combined with a suitable link function). We show that the resulting posterior distribution achieves optimal posterior contraction rates towards any $p_0 \in B_1^\alpha$, simultaneously for all α greater than a minimal threshold, without requiring knowledge of α and thereby adapting to the smoothness of p_0 , cf. Theorem 1. The proof is based on the general approach to posterior contraction rates in total variation distance, Ghosal et al. (2000), which we pursue by carefully constructing the hyper-prior for the smoothness. See Appendix A.

A secondary contribution of this article is an investigation of the implementation aspects of Besov-Laplace priors for binary classification. Since in the setting at hand the posterior distribution is not available in closed form, we devise an efficient sampling scheme employing recent ad-hoc dimension robust Markov chain Monte Carlo (MCMC) techniques, Chen et al. (2018). We then test the considered methodology in several simulation studies, providing a solid corroboration of the theory, cf. Section 3. In the experiments, we consider both spatially homogeneous and inhomogeneous ground truths, in one- and bi-dimensional scenarios.

The remainder of the paper is organized as follows. Section 1.1 summarizes basic definitions and the main notation used throughout. The hierarchical Besov-Laplace prior for probability response functions is constructed in Section 2.1. The main asymptotic result is provided in Section 2.2. The employed MCMC scheme is outlined in Section 2.3. In Section 3, we present the simulation studies. A summary of results and outlook on related research questions is included in Section 4. The proofs are developed in the Appendix A.

1.1 Main notation

In the following, we take the d -dimensional unit cube $[0, 1]^d$, $d \in \mathbb{N}$, as the primary working domain. For $p \in [1, \infty]$, let $L^p([0, 1]^d)$ be the usual Lebesgue spaces of p -integrable functions defined on $[0, 1]^d$, and let $\|\cdot\|_p$ be their norms. Write $\langle \cdot, \cdot \rangle_2$ for the inner product of $L^2([0, 1]^d)$.

Let $(\psi_l, l \in \mathbb{N})$ be an orthonormal wavelet basis of $L^2([0, 1]^d)$, ordered with a single index, comprising S -regular, $S \in \mathbb{N}$, compactly supported and boundary corrected Daubechies wavelets; see (Lassas et al. 2009, Appendix A) for definitions and details. For $\alpha \in [0, S]$ and $p \in [1, \infty)$, define the (wavelet-based) Besov spaces

$$B_p^\alpha([0, 1]^d) := \left\{ w = \sum_{l=1}^{\infty} w_l \psi_l : \|w\|_{B_p^\alpha}^p := \sum_{l=1}^{\infty} l^{p(\alpha/d+1/2)-1} |w_l|^p < \infty \right\},$$

cf. (Lassas et al. 2009, Appendix A). For $p = \infty$, the spaces $B_\infty^\alpha([0, 1]^d)$, $\alpha \geq 0$, are defined as above by replacing the ℓ^p -type norm with the corresponding ℓ^∞ -type one. The (fixed) regularity S of the underlying wavelet basis can be taken arbitrarily large; thus, we tacitly imply that the condition $\alpha < S$ be satisfied throughout. The traditional Hilbert-Sobolev spaces $H^\alpha([0, 1]^d)$ and Hölder spaces $C^\alpha([0, 1]^d)$ are known to be contained within the above family. Specifically, for all $\alpha \geq 0$, $B_2^\alpha([0, 1]^d) = H^\alpha([0, 1]^d)$, e.g. (Giné and Nickl 2016, p. 370), and $C^\alpha([0, 1]^d) \subseteq B_\infty^\alpha([0, 1]^d)$, with equality holding when $\alpha \notin \mathbb{N}$, e.g. (Giné and Nickl 2016, p. 370). On the other hand, for $p = 1$, the B_1^α -Besov scale is known to suitably model spatially inhomogeneous functions with localized ‘spiky’ or ‘blocky’ features. For example, the space of bounded variation functions $BV([0, 1]^d)$, which is of particular interest in many applications, is closely related to $B_1^1([0, 1]^d)$; see (Donoho and Johnstone 1998, Section 1).

When no confusion can arise, we at times omit the underlying domain in the notation, writing, for example, B_p^α for $B_p^\alpha([0, 1]^d)$. We use the symbols \lesssim , \gtrsim , and \simeq for one- and two-sided inequalities holding up to multiplicative constants that are independent of all the involved quantities. For a set \mathcal{W} and a metric δ on \mathcal{W} , the covering number $N(\varepsilon; \mathcal{W}, \delta)$, $\varepsilon > 0$, is defined as the minimum number of balls of δ -radius equal to ε needed to contain \mathcal{W} in their union.

2 Hierarchical Besov-Laplace priors for binary classification

Let $D^{(n)} = \{(Y_i, X_i)\}_{i=1}^n$ be binary-labeled classification data generated according to the statistical model

$$\begin{aligned} Y_i | X_i &\stackrel{\text{iid}}{\sim} \text{Bernoulli}(p(X_i)), \\ X_i &\stackrel{\text{iid}}{\sim} \mu_X. \end{aligned} \tag{1}$$

Here, μ_X is a probability distribution on $[0, 1]^d$ and $p : [0, 1]^d \rightarrow [0, 1]$ is a probability response function, also referred to as the ‘classification surface’. We consider

the problem of nonparametrically estimating p from observations $D^{(n)}$. Throughout, we assume that μ_X be known and impose the minimal requirement that it be absolutely continuous with respect to the Lebesgue measure, with a bounded and bounded away from zero probability density function (p.d.f.), that we also denote by μ_X in slight abuse of notation. If μ_X were unknown, standard density estimation techniques, e.g. [Tsybakov \(2009\)](#); [Giné and Nickl \(2016\)](#); [Ghosal and van der Vaart \(2017\)](#) could be used to make inference on it based on the marginal sample X_1, \dots, X_n .

We denote by $Q_p^{(n)}$ the joint (product) law of $D^{(n)}$, and by $E_p^{(n)}$ the expectation with respect to it. The likelihood is given by

$$L^{(n)}(p) = \prod_{i=1}^n p(X_i)^{Y_i} (1 - p(X_i))^{1-Y_i} \mu_X(X_i). \quad (2)$$

2.1 The prior model

We adopt the nonparametric Bayesian approach, cf. [Ghosal and van der Vaart \(2017\)](#), modeling p with a prior distribution Π supported on the collection \mathcal{P} of measurable functions defined on $[0, 1]^d$ and with values in $[0, 1]$. In particular, we address the case where p may be spatially inhomogeneous, possibly presenting localized features such as spikes or sharp increases or decreases. Since low-integrability Besov spaces are known to provide an effective mathematical model for functions of this type, we make the assumption that p belongs to the B_1^α -Besov scale; see Section 1.1 for definitions and details. We then employ Besov-Laplace prior, which constitutes a particular case within the general class of ‘Besov space priors’ introduced by [Lassas et al. \(2009\)](#). These are popular in inverse problems and imaging due to their ‘edge-preserving and sparsity promoting’ properties, e.g. [Leporini and Pesquet \(2001\)](#); [Vänskä et al. \(2009\)](#); [Kekkonen et al. \(2023\)](#), and have been recently shown to yield optimal recovery, in various statistical models, of spatially inhomogeneous functions in Besov spaces; see [Agapiou et al. \(2021\)](#) as well as [Giordano and Ray \(2022\)](#); [Giordano \(2023\)](#); [Agapiou and Wang \(2024\)](#).

Building on the latter references, we define priors for classification surfaces starting from rescaled Besov-Laplace random functions

$$W_n(x) = \frac{1}{n^{d/(2\alpha+d)}} \sum_{l=1}^{\infty} l^{-(\frac{\alpha}{d}-\frac{1}{2})} w_l \psi_l(x), \quad x \in [0, 1]^d, \quad \alpha > d, \quad w_l \stackrel{\text{iid}}{\sim} \text{Laplace}, \quad (3)$$

with $(\psi_l, l \in \mathbb{N})$ a single-index orthonormal wavelet basis of $L^2([0, 1]^d)$ generating the collection of Besov spaces, cf. Section 1.1, and where the Laplace (or double exponential) distribution has p.d.f. proportional to $e^{-|t|/2}$, $t \in \mathbb{R}$. In [Agapiou et al. \(2021\)](#), the law of W_n in (3) is called a ‘rescaled $(\alpha - d)$ -regular Laplace prior’ in view of the fact that its realizations belong almost surely to $B_p^\beta([0, 1]^d)$ for all $\beta < \alpha - d$ and all $p \in [1, \infty]$, cf. ([Agapiou et al. 2021](#), Lemma 5.2). In the terminology of [Lassas et al. \(2009\)](#), (3) defines a (rescaled) ‘ B_1^α -prior’.

In the aforementioned frequentist analysis literature, the smoothness hyperparameter α was shown to be key in driving the speed of concentration of posterior

distributions associated to rescaled Besov-Laplace priors, requiring a precise matching to the regularity of the ground truth in order to achieve minimax-optimal posterior contraction rates. Since assuming knowledge of the true smoothness is typically unrealistic, here we seek a fully data-driven procedure able to automatically ‘adapt’ to it, achieving optimal performances for a wide range of regularities. To do so, we employ the hierarchical Bayesian approach, e.g. (Ghosal and van der Vaart 2017, Chapter 10), randomizing α in (3) via a hyper-prior distribution. Specifically, we model $\alpha \sim \Sigma_n$, where Σ_n is an n -dependent absolutely continuous distribution supported on the increasing interval $(d, \log n]$, with p.d.f.

$$\sigma_n(\alpha) = \frac{e^{-n^{d/(2\alpha+d)}}}{\zeta_n}, \quad \alpha \in (d, \log n], \quad (4)$$

whose normalization constant satisfies $\zeta_n \simeq \log n$. This construction is motivated by previous findings in the literature showing that such hyper-prior distributions enjoy certain universal adaptation properties. See Lember and van der Vaart (2007) for results in density estimation, and van Waaij and van Zanten (2016) for drift estimation for diffusion processes. An analogous choice also underpins the adaptive posterior contraction rates for Besov-Laplace priors in density estimation derived by Giordano (2023).

For W_n as in (3), and $\alpha \sim \Sigma_n$ with hyper-prior p.d.f. as in (4), we conclude the construction of the prior distribution Π for probability response functions, whose range is equal to $[0, 1]$, via a transformation through a smooth and strictly increasing link function $H : \mathbb{R} \rightarrow [0, 1]$. For concreteness, we take the the logistic (or ‘sigmoid’) link $H(t) = e^t/(e^t + 1)$, $t \in \mathbb{R}$, and let Π_n be the law of the random function

$$p_{W_n}(x) := H[W_n(x)] = \frac{e^{W_n(x)}}{e^{W_n(x)} + 1}, \quad x \in [0, 1]^d. \quad (5)$$

Other link functions (such as the probit link $H = \Phi$, with Φ the standard normal cumulative distribution function) could be used as well under suitable regularity conditions. We refer to Π_n as a hierarchical rescaled Besov-Laplace prior for classification surfaces.

Given labeled binary classification data $D^{(n)}$ arising as in (1), the posterior distribution $\Pi(\cdot|D^{(n)})$ of $p|D^{(n)}$ is then given, according to Bayes’ formula, e.g. (Ghosal and van der Vaart 2017, p.7), by

$$\Pi_n(A|D^{(n)}) = \frac{\int_A L^{(n)}(p)d\Pi(p)}{\int_{\mathcal{P}} L^{(n)}(p)d\Pi(p)}, \quad A \subseteq \mathcal{P} \text{ measurable}, \quad (6)$$

with $L^{(n)}$ the likelihood from (2). Following the Bayesian paradigm, $\Pi_n(\cdot|D^{(n)})$ represents the updated belief about p after observing the data. It furnishes point estimates and uncertainty quantification. See Section 3 for a concrete illustration with synthetic data.

Remark 1 (Rescaling) Similar rescaling terms to the sequence $n^{-d/(2\alpha+d)}$ introduced in (3) underpin essentially all existing frequentist analyses of Besov-Laplace priors, e.g. Agapiou et al. (2021); Giordano (2023); Agapiou and Wang (2024). By uniformly shrinking the prior draws, this enforces additional regularization that yields tight complexity bounds for the ‘sieves’ associated to the prior distribution, which play a crucial role in the pursuit of the testing approach to posterior contraction rates (e.g. Ghosal and van der Vaart (2017)). See the discussion after Theorem 1 in Giordano (2023) for further insights. A notable exception is the recent investigation by Dolera et al. (2024) in the white noise model, which is based on a different proof strategy. However, it remains unclear whether the latter could be extended to the problem at hand.

2.2 Adaptive posterior contraction rates

In this section, we characterize the asymptotic behavior of the posterior distribution as $n \rightarrow \infty$ under the frequentist assumption that the data $D^{(n)} \sim Q_{p_0}^{(n)}$ have been generated according to the observation model (1) by some possibly spatially inhomogeneous true probability response function $p_0 \in B_1^{\alpha_0}([0, 1]^d)$, for some $\alpha_0 > d$. The following result quantifies, via the notion of ‘posterior contraction rates’, cf. (Ghosal and van der Vaart 2017, Chapter 8), the speed at which $\Pi(\cdot | D^{(n)})$ concentrates around p_0 in L^1 -distance.

Theorem 1 Let Π_n be a hierarchical rescaled Besov-Laplace prior for probability response functions constructed as in Section 2.1. Let $D^{(n)} = \{(Y_i, X_i)\}_{i=1}^n \sim Q_{p_0}^{(n)}$ be a random sample of labeled binary classification data arising from model (1) for some fixed $p = p_0 \in B_1^{\alpha_0}([0, 1]^d)$ for some $\alpha_0 > d$, satisfying $\inf_{x \in [0, 1]^d} p_0(x) > 0$. Then, for $M > 0$ large enough, as $n \rightarrow \infty$,

$$E_{p_0}^{(n)} \left[\Pi_n \left(p : \|p - p_0\|_1 > M n^{-\frac{\alpha_0}{2\alpha_0+d}} \middle| D^{(n)} \right) \right] \rightarrow 0.$$

Theorem 1 entails that, with $Q_{p_0}^{(n)}$ -probability tending to one, $\Pi_n(\cdot | D^{(n)})$ puts all of its probability mass in small neighborhoods of p_0 with L^1 -radius shrinking at rate $n^{-\alpha_0/(2\alpha_0+d)}$. Consequently, draws from the posterior distribution provide increasingly better reconstruction of the ground truth.

The rate $n^{-\alpha_0/(2\alpha_0+d)}$ is known to be optimal, in the minimax sense, for the L^1 -distance over the Besov space $B_1^{\alpha_0}([0, 1]^d)$, e.g. Donoho and Johnstone (1998). In Theorem 1, this is achieved via a nonparametric Bayesian procedure that does not require knowledge of the true regularity, but rather adapts to α_0 in the wide range (d, ∞) . We then conclude that hierarchical rescaled Besov-Laplace priors achieve adaptive posterior contraction rates in binary classification. This is in line with the existing adaptation results for hierarchical Besov-Laplace priors in density estimation, Giordano (2023), and in the white noise model, Agapiou and Savva (2024).

For spatially homogeneous true probability response functions belonging to traditional Hölder spaces, van der Vaart and van Zanten (2009) proved adaptation for hierarchical Gaussian priors with randomized length-scale. Our result provide a parallel to this for ground truths in the B_1^α -Besov scale and hierarchical rescaled Besov-Laplace priors. On the other hand, Gaussian priors have been shown to be unable to

optimally reconstruct spatially inhomogeneous functions, Agapiou and Wang (2024), and thus cannot be expected to achieve optimal posterior contraction rates, even non-adaptive ones, in the setting of Theorem 1, regardless of any specific tuning or randomization. We provide a numerical illustration of this phenomenon in Section 3 via synthetic data.

Remark 2 (Boundedness away from zero) In Theorem 1, the assumption that the ground truth be positive throughout the domain guarantees that p_0 be in the range of the composition with respect to the link function H . Since, by construction, the prior Π_n is supported over such functional compositions, this restriction appears to be necessary for posterior consistency in the considered setting. Similar assumptions are also imposed for the posterior contraction rates for Gaussian priors in binary classification derived by van der Vaart and van Zanten (2008). We refer the reader to (Ghosal and van der Vaart 2017, Chapter 9.5.6) for results for priors based on the Dirichlet process in the case where the true probability response function is not necessarily bounded away from zero. Investigating such scenario in the presence of spatial inhomogeneity is an interesting open question.

2.3 Posterior sampling

For the observation model (1), the posterior distribution resulting from the considered hierarchical rescaled Besov-Laplace prior, is not available in closed form. We then employ an MCMC method to draw approximate samples from $\Pi(\cdot|D^{(n)})$ and concretely implement posterior inference.

Specifically, within a Gibbs-type scheme to handle the hierarchical construction, we resort to the ‘whitened pre-conditioned Crank-Nicolson’ (wpCN) algorithm of Chen et al. (2018), which is a Metropolis-Hastings-type technique applicable to prior distributions that can be expressed (in our case, conditionally) as transformation of a Gaussian white noise. For the hierarchical rescaled Besov-Laplace priors from Section 2.2, we observe that, for fixed α , the random function W_n in (3) is equal in distribution to

$$T_\alpha^{(n)}(\xi)(x) := \sum_{l=1}^{\infty} T_{\alpha,l}^{(n)}(w_l) \psi_l(x), \quad x \in [0, 1]^d \quad w_l \stackrel{\text{iid}}{\sim} N(0, 1), \quad (7)$$

where

$$\xi(x) := \sum_{l=1}^{\infty} w_l \psi_l(x), \quad x \in [0, 1]^d, \quad w_l \stackrel{\text{iid}}{\sim} N(0, 1), \quad (8)$$

is a Gaussian white noise process indexed by $[0, 1]^d$, and where the ‘whitening transformation’ T is given by

$$T_{\alpha,l}^{(n)}(w) := n^{-\frac{d}{2\alpha+d}} l^{-(\frac{\alpha}{d}-\frac{1}{2})} \operatorname{sgn}(w) [-\log(2 - 2\Phi(|w|))], \quad w \in \mathbb{R}, \quad l \in \mathbb{N}.$$

Starting from some initialization for α (e.g. a fixed ‘cold start’ $\alpha = d+1$), and given an initial white noise sample ξ_0 (obtained from (8) by drawing independent standard normal random coefficients), with $w_0 := T(\xi_0)$ the corresponding initialization for $w = H^{-1} \circ p$, each step of the wpCN-within-Gibbs algorithm alternates samples from:

1. The full conditional distribution of the smoothness parameter α , via a simple random walk Metropolis-Hastings algorithm, namely:

- propose $\alpha_* := \max\{\alpha_{s-1} + \delta_1 Z, d\}$, where $\delta_1 > 0$ is a fixed step-size and Z is an independent standard Gaussian random variable.
- Set

$$\alpha_s := \begin{cases} \alpha_*, & \text{with probability } \min \left\{ 1, \frac{L^{(n)}(H \circ T_{\alpha_*}^{(n)}(\xi_{s-1}))}{L^{(n)}(H \circ T_{\alpha_{s-1}}^{(n)}(\xi_{s-1}))} \times \frac{\sigma_n(\alpha_*)}{\sigma_n(\alpha_{s-1})} \right\}, \\ \alpha_{s-1}, & \text{otherwise,} \end{cases}$$

with $L^{(n)}$ the likelihood from (2) and σ_n the hyper-prior p.d.f. from (4).

2. The full conditional distribution of the infinite-dimensional parameter ω via the wpCN algorithm, namely:

- Construct the whitened proposal $\xi^* := \sqrt{1 - 2\delta_2} \xi_{s-1} + \sqrt{2\delta_2} \chi$, where $\delta_2 \in (0, 1/2)$ is a fixed step-size and χ is an independent Gaussian white noise.
- Set

$$\xi_s := \begin{cases} \xi^*, & \text{with probability } \min \left\{ 1, \frac{L^{(n)}(H \circ T_{\alpha_s}^{(n)}(\xi^*))}{L^{(n)}(H \circ \omega_{s-1})} \right\}, \\ \xi_{s-1}, & \text{otherwise.} \end{cases}$$

- Set $\omega_s = T_{\alpha_s}^{(n)}(\xi_s)$.

In practice, we implement the first step within the above wpCN routine by truncating the series in (7) and (8) at some pre-specified level $L \in \mathbb{N}$, sufficiently high as to guarantee that the resulting numerical approximation error is negligible compared to the statistical one (e.g. taking L proportional to n). The second operation necessitates the evaluation of the proposal likelihood, which is straightforward for the observation model (1), cf. (2).

The resulting Markov chain $(\alpha_s, \omega_s)_{s=0}^\infty$ has limiting distribution equal to the joint posterior distribution of (α, ω) , [Chen et al. \(2018\)](#). Moreover, the underlying pCN-type structure is known to give rise to dimension-robust acceptance probabilities, [Cotter et al. \(2013\)](#). This implies desirable mixing properties and efficient convergence towards equilibrium, even under high discretization dimensions (i.e., truncation levels).

3 Simulation studies

We assess the empirical performance of the considered hierarchical rescaled Besov-Laplace prior for binary classification via numerical experiments with synthetic data. For one- and bi-dimensional domains $[0, 1]^d$, $d = 1, 2$, we fix spatially homogeneous and inhomogeneous true probability response functions p_0 , generate independent and identically distributed (i.i.d.) inputs $X_i \stackrel{\text{iid}}{\sim} \text{Uniform}([0, 1]^d)$, conditionally on which we sample the labels Y_1, \dots, Y_n according to model (1). We then perform posterior inference via the wpCN-within-Gibbs MCMC algorithm for posterior sampling outlined in Section 2.3.

For comparison, we also consider a hierarchical Gaussian prior defined via the law of a conditionally centered stationary Gaussian process $G := (G_x, x \in [0, 1]^d)$ with square-exponential covariance kernel, whose length-scale is randomized through an inverse-Gamma hyper-prior distribution,

$$E[G_x G_y | A] = e^{-A|x-y|^2}, \quad x, y \in [0, 1]^d, \quad A^d \sim \Gamma(1, 1).$$

Combined with suitable link functions H (including e.g. the logistic one), this was shown by [van der Vaart and van Zanten \(2009\)](#) to achieve adaptive posterior contraction rates towards true probability response functions with traditional Hölder regularity. However, it is generally expected to be unable to optimally recover spatially inhomogeneous ground truths, in view of the known sub-optimality of Gaussian priors in this case, cf. the discussion after Theorem 1.

Posterior inference with the above hierarchical Gaussian prior is implemented via a Metropolis-within-Gibbs MCMC sampling algorithm similar to the one from Section 2.3, where the wpCN routine for Besov-Laplace priors used therein is replaced by the standard pCN method, [Cotter et al. \(2013\)](#). All the numerical experiments were carried out in R on an Intel(R) Core(TM) i7-10875H 2.30GHz processor with 32 GB of RAM. Each MCMC run was iterated for 25,000 steps, discarding the first 10,000 as burn-in. Maximum per-experiment computation times were of around 20 minutes.

3.1 Univariate experiments

We start considering a univariate spatially homogeneous scenario, setting the ground truth to

$$p_0(x) = \frac{1}{1 + e^{-(9x-5)}}, \quad x \in [0, 1], \quad (9)$$

namely a scaled and shifted sigmoid function, restricted to the unit interval $[0, 1]$. See Figure 1. With this setup, we sampled labeled binary classification data $D^{(n)}$ from model (1), with $\mu_X = \text{Uniform}([0, 1])$ and increasing sample sizes $n = 50, 200, 1000, 5000$. The observations are represented by the rugs in Figure 2.

We then performed posterior inference based on the hierarchical rescaled Besov-Laplace priors from Section 2.1 and the hierarchical Gaussian prior described above (for which the logistic link function was also chosen). Figure 1 shows the obtained (MCMC approximations to the) ‘posterior means’ $\bar{p}_n := H \circ E^{\Pi_n}[w|D^{(n)}]$ for $n = 200, 1000, 5000$. Uncertainty is quantified and visualized by the associated point-wise 95%-credible intervals. As expected from Theorem 1 and the existing theory for hierarchical Gaussian priors with randomized length-scale, [van der Vaart and van Zanten \(2009\)](#), both procedures achieve a clear convergence towards the spatially homogeneous ground truth, with practically indistinguishable reconstructions at the largest sample size. This visual comparison is corroborated by Table 1, where L^1 -estimation errors, averaged over 50 replications of each experiment, are reported, jointly with the corresponding standard deviations.

For the hierarchical rescaled Besov-Laplace prior, we used Daubechies-8 maximally symmetric (i.e. ‘Symmelet-8’) functions, with symmetric boundary reflection, implemented in the R package `wavethresh`, truncating the series after the first 1,024 terms.

All runs of the MCMC algorithms were initialized at cold, uninformative starts. The step sizes for the wpCN and pCN routines were chosen within the range [0.001, 0.05], depending on the sample size, to achieve a stabilization of the acceptance probabilities at around 30% after burn-in.

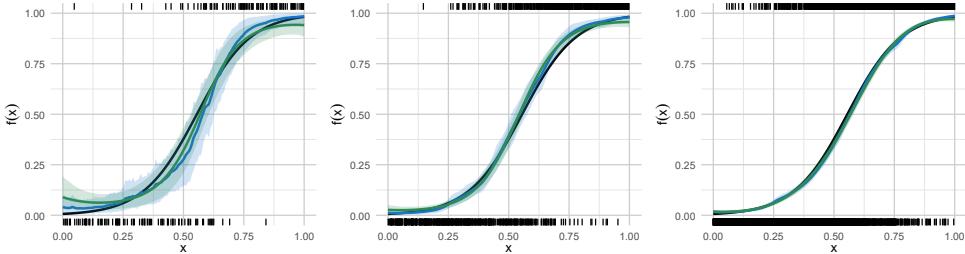


Fig. 1: Left to right: Posterior means for Gaussian (solid green) and Besov-Laplace (solid blue) priors, pointwise 95%-credible intervals (shaded regions) for $n = 200, 1000, 5000$, respectively. The ground truth p_0 from (9) is shown in solid black. Rugs at the bottom represent the covariate values labeled 0, while rugs at the top represent covariates labeled 1.

	n	50	200	1000	5000
Gaussian	$\ \bar{p}_n - p_0\ _1$	0.16 (0.02)	0.05 (0.01)	0.02 (0.005)	0.01 (0.003)
	$\ \bar{p}_n - p_0\ _1 / \ p_0\ _1$	0.27 (0.03)	0.09 (0.03)	0.04 (0.008)	0.02 (0.005)
Laplace	$\ \bar{p}_n - p_0\ _1$	0.21 (0.05)	0.05 (0.02)	0.02 (0.007)	0.01 (0.003)
	$\ \bar{p}_n - p_0\ _1 / \ p_0\ _1$	0.37 (0.09)	0.09 (0.03)	0.04 (0.011)	0.02 (0.005)

Table 1: Average L^1 -estimation errors for the posterior mean (and their standard deviations) over 50 repeated experiments with the ground p_0 from (9).

We next consider the step-like spatially inhomogeneous probability response function

$$p_0(x) = \begin{cases} 0.9 & x \in [0, 0.4) \\ 0.2 & x \in [0.4, 0.7) \\ 0.5 & x \in [0.7, 1], \end{cases} \quad (10)$$

cf. Figure 2 below, for which the obtained results are summarized in Figure 2 below. Unlike the preceding spatially homogeneous scenario, here a marked difference between the performance of the two procedures emerges. In particular, the hierarchical Gaussian prior appears to be unable to correctly detect the sharp jumps of the ground truth at inputs $x = 0.4, 0.7$, significantly over-smoothing the edges of the blocks even as the sample size increases. This is in line with the known sub-optimality of Gaussian priors, [Agapiou and Wang \(2024\)](#), and generally of linear procedures, [Donoho and Johnstone \(1998\)](#), in the presence of spatial inhomogeneity. On the contrary, the

hierarchical rescaled Besov Laplace priors achieves progressively more faithful reconstructions. Table 2 reports the obtained average L^1 -estimation errors for the posterior mean, where the latter procedure is shown to outperform the former across all values of n , with a more significant improvement in the recovery at the largest sample size.

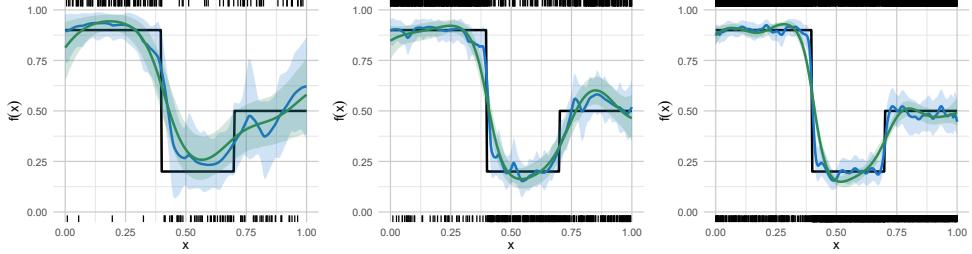


Fig. 2: Left to right: Posterior means for Gaussian (solid green) and Laplace (solid blue) priors, pointwise 95%-credible intervals (shaded regions) for $n = 200, 1000, 5000$, respectively. The ground truth p_0 from (10) is shown in solid black.

	n	50	200	1000	5000
Gaussian	$\ \bar{p}_n - p_0\ _{L^1}$	0.26 (0.03)	0.17 (0.03)	0.09 (0.01)	0.08 (0.006)
	$\ \bar{p}_n - p_0\ _{L^1} / \ p_0\ _{L^1}$	0.36 (0.05)	0.19 (0.04)	0.14 (0.01)	0.12 (0.009)
Laplace	$\ \bar{p}_n - p_0\ _{L^1}$	0.25 (0.03)	0.15 (0.03)	0.09 (0.01)	0.06 (0.003)
	$\ \bar{p}_n - p_0\ _{L^1} / \ p_0\ _{L^1}$	0.35 (0.04)	0.17 (0.04)	0.14 (0.01)	0.09 (0.005)

Table 2: Average L^1 -estimation errors for the posterior mean (and their standard deviations) over 50 repeated experiments with the ground truth p_0 from (10).

3.2 Bivariate experiments

Over the unit square $[0, 1]^2$, we also consider two scenarios, respectively with:

1. A spatially homogeneous true probability response function given by

$$f(x_1, x_2) = \frac{1}{4} f_{SN}(x_1, x_2; (0.4, 0.6), 0.05I_2, (3, -2)), \quad (x_1, x_2) \in [0, 1]^2, \quad (11)$$

where f_{SN} denotes the (bivariate) skew-normal p.d.f. and I_2 is the identity matrix in $\mathbb{R}^{2,2}$. See the last panel of Figure 3;

2. A spatially inhomogeneous ground truth with a square block component

$$p_0(x_1, x_2) = 0.4 \prod_{h=1}^2 (1 + \text{sgn}(x_h - b_h))(1 - \text{sgn}(x_h - c_h)), \quad (x_1, x_2) \in [0, 1]^2, \quad (12)$$

where the block extremes are $b = (0.1, 0.1)$ and $c = (0.5, 0.5)$, cf. Figure 4 (last panel).

For both, the obtained posterior mean estimates \bar{p}_n based on the hierarchical rescaled Besov-Laplace prior and the hierarchical Gaussian prior are shown in Figures 3 and 4, respectively, for increasing sample sizes $n = 200, 1000, 5000$. The associated L^1 -estimation errors are reported in Tables 3 and 4, respectively.

The numerical results broadly reinforce the findings from the univariate experiments from Section 3.1. For the homogeneous ground truth, both procedures deliver excellent recoveries, achieving very similar estimation errors, that steadily decrease as n grows. On the other hand, the hierarchical Gaussian prior appears to be unable to correctly reconstruct the edges of the block component of the true probability response function (12), which are visibly oversmoothed. These are instead precisely reconstructed by the hierarchical rescaled Besov-Laplace prior. For the latter, the obtained estimation errors are lower across all values of n , and particularly for the largest sample size.

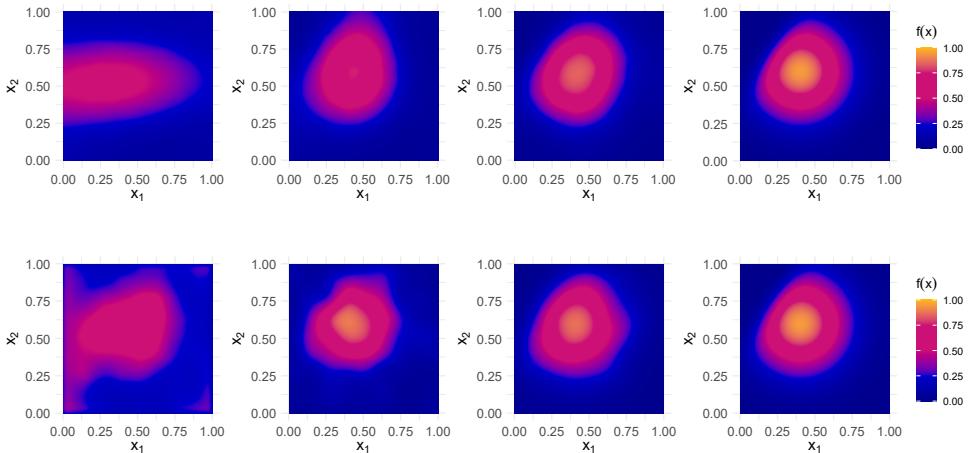


Fig. 3: Top to bottom, left to right: posterior means for Gaussian (top) and Besov-Laplace (bottom) priors for increasing sample sizes $n = 200, 1000, 5000$, in case of the spatially homogeneous ground truth (11), shown in the rightmost panel of both rows.

	n	50	200	1000	5000
Gaussian	$\ \bar{p}_n - p_0\ _{L^1}$	0.26 (0.01)	0.18 (0.03)	0.07 (0.006)	0.04 (0.003)
	$\ \bar{p}_n - p_0\ _{L^1} / \ p_0\ _{L^1}$	0.75 (0.02)	0.50 (0.09)	0.19 (0.02)	0.12 (0.01)
Laplace	$\ \bar{p}_n - p_0\ _{L^1}$	0.28 (0.03)	0.17 (0.01)	0.07 (0.007)	0.05 (0.003)
	$\ \bar{p}_n - p_0\ _{L^1} / \ p_0\ _{L^1}$	0.80 (0.06)	0.48 (0.02)	0.21 (0.02)	0.13 (0.01)

Table 3: Average L^1 -estimation errors for the posterior mean (and their standard deviations) over 50 repeated experiments with the ground p_0 from (11).

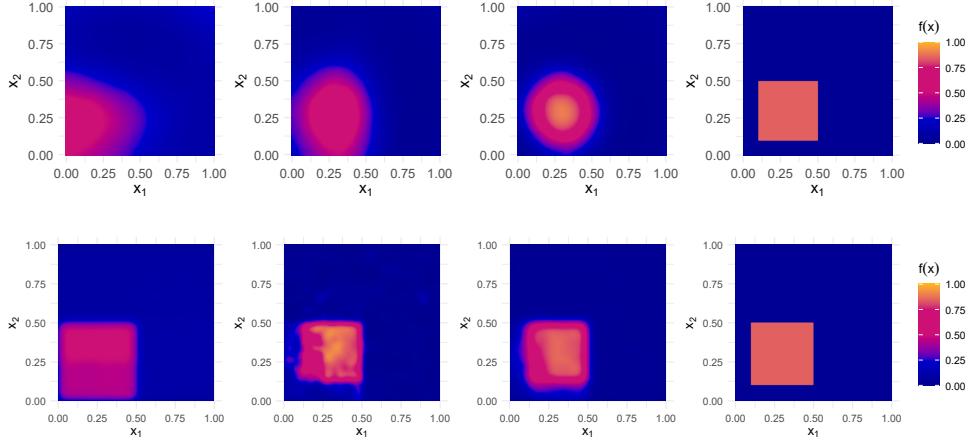


Fig. 4: Top to bottom, left to right: posterior means for Gaussian (top) and Besov-Laplace (bottom) priors for increasing sample sizes $n = 200, 1000, 5000$, in case of the spatially homogeneous ground truth (12), shown in the rightmost panel of both rows.

	n	50	200	1000	5000
Gaussian	$\ \bar{p}_n - p_0\ _{L^1}$	0.28 (0.05)	0.24 (0.02)	0.17 (0.01)	0.14 (0.003)
	$\ \bar{p}_n - p_0\ _{L^1} / \ p_0\ _{L^1}$	0.84 (0.12)	0.72 (0.05)	0.51 (0.03)	0.42 (0.009)
Laplace	$\ \bar{p}_n - p_0\ _{L^1}$	0.28 (0.05)	0.21 (0.01)	0.14 (0.02)	0.08 (0.003)
	$\ \bar{p}_n - p_0\ _{L^1} / \ p_0\ _{L^1}$	0.84 (0.11)	0.62 (0.02)	0.50 (0.05)	0.25 (0.008)

Table 4: Average L^1 -estimation errors for the posterior mean (and their standard deviations) over 50 repeated experiments with the ground p_0 from (12).

4 Discussion

In this work, we have studied a nonparametric Bayesian approach to the recovery of spatially inhomogeneous binary classification surfaces based on hierarchical rescaled

Besov-Laplace priors. Our main theoretical result (Theorem 1) shows that, for ground truths p_0 belonging to the B_1^α -scale, the posterior contracts around p_0 at the optimal rate in L^1 -distance, automatically adapting to the smoothness. For implementation, we have outlined an MCMC posterior sampling algorithm in Section 2.3, which we have applied in the simulation studies of Section 3, demonstrating the practical feasibility and effective performance of the considered procedure.

The employed prior is constructed from base rescaled Besov-Laplace random elements, cf. (3), and involves the randomization of the smoothness hyper-parameter by a carefully tuned hyper-prior, cf. (4). Recently, Agapiou and Savva (2024) achieved adaptive posterior contraction rates towards spatially inhomogeneous ground truths in the white noise by independently randomizing, via somewhat more natural hyper-priors, both the regularity and the scaling terms. It remains unclear at the present stage whether their investigation could be extended to the binary classification setting.

Lastly, let us also mention the important issue of providing a frequentist validation for the associated uncertainty quantification, since it is known that, in infinite-dimensional statistical models, credible sets can have asymptotically zero coverage even if they arise from consistent posterior distributions; see Cox (1993). An established approach to derive such guarantees is via so-called ‘nonparametric Bernstein von-Mises theorems’, Castillo and Nickl (2013). A result of this kind for Besov-Laplace prior has been recently established, in the context of drift estimation for diffusion processes, by Giordano and Ray (2025). Pursuing these results for the considered hierarchical priors for probability response functions is an interesting open question for future research.

Acknowledgements. The authors gratefully acknowledge the “de Castro” Statistics Initiative for supporting this research. M.G. is also grateful for the partial financial support by MUR, PRIN project 2022CLTYP4.

Code availability. The code to reproduce all the presented numerical experiments is available at <https://github.com/PatricDolmeta/Besov-Laplace-Classification>.

Appendix A Proofs

A.1 Proof of Theorem 1

We apply the general program for posterior contraction rates in total variation distance set forth in Ghosal et al. (2000); see also (Ghosal and van der Vaart 2017, Theorem 8.9). Recall from Section 2 that for a data pair $(X, Y) \sim Q_p^{(1)}$, $p \in \mathcal{P}$, arising as in model (1), the joint probability density function is given by

$$q_p(y, x) = p(x)^y (1 - p(x))^{1-y} \mu_X(x), \quad y \in \{0, 1\}, \quad x \in [0, 1]^d.$$

For two probability response functions $p_1, p_2 \in \mathcal{P}$, the total variation distance between the laws $Q_{p_1}^{(1)}, Q_{p_2}^{(1)}$ is then given by

$$TV(Q_{p_1}^{(1)}, Q_{p_2}^{(1)}) = \frac{1}{2} \|q_{p_1} - q_{p_2}\|_{L^1(\{0,1\} \times [0,1]^d)}$$

$$\begin{aligned}
&= \frac{1}{2} \left[\int_{[0,1]^d} |p_1(x) - p_2(x)| \mu_X(x) dx \right. \\
&\quad \left. + \int_{[0,1]^d} |1 - p_1(x) - 1 + p_2(x)| \mu_X(x) dx \right] = \|p_1 - p_2\|_{L^1([0,1]^d, \mu_X)}
\end{aligned}$$

cf. eq. (B.1) in [Ghosal and van der Vaart \(2017\)](#). Since μ_X is bounded and bounded away from zero by assumption, we conclude that the latter is equivalent to the standard L^1 -distance $\|p_1 - p_2\|_1$. Further, let

$$K(p_1, p_2) := E_{p_1}^{(1)} \left[\log \frac{p_1(Y, X)}{p_2(Y, X)} \right]; \quad V(p_1, p_2) := E_{p_1}^{(1)} \left| \log \frac{p_1(Y, X)}{p_2(Y, X)} - K(p_1, p_2) \right|^2,$$

be the Kullback-Leibler divergence and variation, respectively. Theorem 8.9 in [Ghosal and van der Vaart \(2017\)](#) then yields that, if for some positive sequence $\varepsilon_n \rightarrow 0$ such that $n\varepsilon_n^2 \rightarrow \infty$, the hierarchical rescaled Besov-Laplace prior Π_n for classification surfaces from Section 2.1 is shown to satisfy

$$\Pi_n(p : K(p_0, p) \leq \varepsilon_n^2, V(p_0, p) \leq \varepsilon_n^2) \geq e^{-Cn\varepsilon_n^2}, \quad (\text{A1})$$

for some constant $C > 0$, and all $n \in \mathbb{N}$ large enough, as well as

$$\Pi_n(\mathcal{P}_n^c) \leq e^{-(C+4)n\varepsilon_n^2}, \quad (\text{A2})$$

for measurable sets $\mathcal{P}_n \subseteq \mathcal{P}$ satisfying

$$\log N(\varepsilon_n; \mathcal{P}_n, \|\cdot\|_1) \lesssim n\varepsilon_n^2, \quad (\text{A3})$$

then $\Pi_n(\cdot | D^{(n)})$ contracts towards p_0 at rate ε_n in total variation (i.e., standard L^1 -distance).

We verify conditions (A1) - (A3) with $\varepsilon_n = Mn^{-\alpha_0/(2\alpha_0+d)}$ with $M > 0$ a large enough constant. Let Π_{W_n} be the law of the hierarchical rescaled Besov-Laplace random element W_n from (3) with smoothness hyper-prior $\alpha \sim \Sigma_n$ as in (4). Note that, since the logistic link H is strictly increasing and smooth, it possesses a strictly increasing and smooth inverse, $H^{-1} : [0, 1] \rightarrow \mathbb{R}$. Thus, in view of the positivity assumption $\inf_{x \in [0,1]^d} p_0(x) > 0$, we have $p_0 = H \circ w_0$ for $w_0 := H^{-1} \circ p_0$. Further, recalling that $p_0 \in B_1^{\alpha_0}$, we can conclude by Theorem 10 in [Bourdaud and Sickel \(2011\)](#), that $w_0 \in B_1^{\alpha_0}$ as well. Then, by Lemma 2.8 in [Ghosal and van der Vaart \(2017\)](#), for all measurable and bounded $w : [0, 1]^d \rightarrow \mathbb{R}$,

$$\max \{K(p_0, H \circ w), V(p_0, H \circ w)\} \lesssim \|w_0 - w\|_{L^2([0,1]^d, \mu_X)}^2 \simeq \|w_0 - w\|_2^2.$$

The prior probability in (A1) is then bounded below by, for some $c_1 > 0$, by

$$\Pi_{W_n} \left(w : \|w - w_0\|_2 \leq c_1 M n^{-\frac{\alpha_0}{2\alpha_0+d}} \right)$$

which, upon choosing $M > 0$ large enough, is greater than $e^{-c_2 n^{d/(2\alpha_0+d)}} = e^{-c_2 n \varepsilon_n^2}$ for some $c_2 > 0$ by Lemma 2 below.

Turning to the construction of the sieves \mathcal{P}_n from (A2) and (A3), by Lemma 3 below, there exists sufficiently large constants $b_1, b_2 > 0$ such the sets

$$\mathcal{W}_n := \left\{ w = w^{(1)} + w^{(2)} : \|w^{(1)}\|_1 \leq b_1 n^{-\frac{\alpha_*}{2\alpha_*+d}}, \|w^{(2)}\|_{B_1^{\alpha_*+d}} \leq b_1 n^{\frac{d}{2\alpha_*+d}} \right\},$$

with $\alpha_* = \alpha_0/(1 + b_2/\log n)$, satisfy

$$\Pi_{W_n}(\mathcal{W}_n^c) \leq e^{-(c_2+4)n^{d/(2\alpha_0+d)}} = e^{-(c_2+4)n\varepsilon_n^2}.$$

Set $\mathcal{P}_n := \{H \circ w, w \in \mathcal{W}_n\}$. Then, by construction, $\Pi_n(\mathcal{P}_n^c) \leq \Pi_{W_n}(\mathcal{W}_n^c) \leq e^{-(c_2+4)n\varepsilon_n^2}$, showing that (A2) is indeed verified. Lastly, by Lemma 3.2 in van der Vaart and van Zanten (2008), for all $w_1, w_2 \in \mathcal{W}_n$,

$$\|H \circ w_1 - H \circ w_2\|_1 \lesssim \|w_1 - w_2\|_1$$

and therefore

$$\log N(\varepsilon_n; \mathcal{P}_n, \|\cdot\|_1) \lesssim \log N(\varepsilon_n; \mathcal{W}_n, \|\cdot\|_1).$$

Since, $n^{-\alpha_*/(2\alpha_*+d)} \lesssim \varepsilon_n$, cf. (A7) below, and since, by construction, \mathcal{W}_n is a $b_1 n^{-\alpha_*/(2\alpha_*+d)}$ -enlargement in L^1 -distance of the set $\{w : \|w\|_{B_1^{\alpha_*+d}} \leq b_1 n^{d/(2\alpha_*+d)}\}$, the metric entropy of interest is upper bounded by a multiple of

$$\begin{aligned} \log N(\varepsilon_n; \left\{ w : \|w\|_{B_1^{\alpha_*+d}} \leq b_1 n^{\frac{d}{2\alpha_*+d}} \right\}, \|\cdot\|_1) &\lesssim \left(\frac{b_1 n^{\frac{d}{2\alpha_*+d}}}{\varepsilon_n} \right)^{\frac{d}{\alpha_*+d}} \\ &\lesssim n^{\frac{d}{2\alpha_*+d}} \lesssim n\varepsilon_n^2, \end{aligned}$$

having used the metric entropy inequality in Theorem 4.3.36 in Giné and Nickl (2016) and again (A7). This concludes the verification of (A3) and the proof of Theorem 1 in view of the equivalence between the total variation distance and the standard L^1 -distance. \square

A.2 Auxiliary results

The next two auxiliary lemmas provide key quantitative properties of the information geometry of the hierarchical rescaled Besov-Laplace priors. They adapt to the present setting previous findings from Giordano (2023), which in turn were based on the investigations of Lember and van der Vaart (2007) and van Waaij and van Zanten (2016), where similar hyper-prior for the smoothness were considered.

Lemma 2 Let Π_{W_n} be the law of the hierarchical rescaled Besov-Laplace random element W_n from (3) with smoothness hyper-prior $\alpha \sim \Sigma_n$ as in (4). Let $w_0 \in B_1^{\alpha_0}([0, 1]^d)$, for some $\alpha_0 > d$. Then, for sufficiently large $B_1, B_2 > 0$,

$$\Pi_{W_n} \left(w : \|w - w_0\|_2 \leq B_1 n^{-\frac{\alpha_0}{2\alpha_0+d}} \right) \geq e^{-B_2 n^{d/(2\alpha_0+d)}}.$$

Proof For $\alpha > d$, let $\varepsilon_{\alpha,n} := n^{-\alpha/(2\alpha+d)}$ and let $\Pi_{W_{\alpha,n}}$ be the law of

$$W_{\alpha,n} := \frac{W_\alpha}{n\varepsilon_{\alpha,n}^2}, \quad W_\alpha := \sum_{l=1}^{\infty} l^{-(\frac{\alpha}{d}-\frac{1}{2})} W_l \psi_l. \quad (\text{A4})$$

Then, by construction,

$$\begin{aligned} \Pi_{W_n} \left(w : \|w - w_0\|_2 \leq B_1 n^{-\frac{\alpha_0}{2\alpha_0+d}} \right) \\ = \int_d^{\log n} \Pi_{W_{\alpha,n}} (w : \|w - w_0\|_2 \leq B_1 \varepsilon_{\alpha_0,n}) \sigma_n(\alpha) d\alpha \\ \geq \int_{\alpha_0}^{\alpha_0+1/\log n} \Pi_{W_{\alpha,n}} (w : \|w - w_0\|_2 \leq B_1 \varepsilon_{\alpha_0,n}) \sigma_n(\alpha) d\alpha. \end{aligned}$$

For a truncation level $L_n \in \mathbb{N}$ to be chosen below, the wavelet projection of $P_{L_n} w_0$ of $w_0 \in B_1^{\alpha_0}([0, 1]^d)$ satisfies

$$\|w_0 - P_{L_n} w_0\|_2 = \sum_{l>L_n} l^{-(\frac{\alpha_0}{d}-\frac{1}{2})} l^{(\frac{\alpha_0}{d}-\frac{1}{2})} |w_{0,l}| \leq L_n^{-(\frac{\alpha_0}{d}-\frac{1}{2})} \|w_0\|_{B_1^{\alpha_0}}.$$

Thus, taking $L_n \simeq n^{\alpha_0/[(2\alpha_0+d)(\alpha_0/d-1/2)]}$ (note that this is greater than the usual order $n^{1/(2\alpha_0+d)}$), we have

$$\|w_0 - P_{L_n} w_0\|_2 \lesssim n^{-\frac{\alpha_0}{2\alpha_0+d}} = \varepsilon_{\alpha_0,n}.$$

Furthermore, for all $\alpha \in [\alpha_0, \alpha_0 + 1/\log n]$, it also holds

$$\begin{aligned} \|P_{L_n} w_0\|_{B_1^\alpha} &= \sum_{l \leq L_n} l^{(\frac{\alpha}{d}-\frac{\alpha_0}{d})} l^{(\frac{\alpha_0}{d}-\frac{1}{2})} |\langle w_0, \psi_l \rangle_2| \\ &\leq L_n^{\frac{(\alpha-\alpha_0)}{d}} \|w_0\|_{B_1^{\alpha_0}} \lesssim n^{\frac{\alpha_0(\alpha_0+\log^{-1} n-\alpha_0)}{(2\alpha_0+d)(\alpha_0-d/2)}} = e^{\frac{\alpha_0}{(2\alpha_0+d)(\alpha_0-d/2)\log n}} \log n \lesssim 1. \end{aligned}$$

Hence, by the triangle inequality, for $B_1 > 0$ large enough and some $c_1 > 0$,

$$\Pi_{W_{\alpha,n}} (w : \|w - w_0\|_2 \leq B_1 \varepsilon_{\alpha_0,n}) \geq \Pi_{W_{\alpha,n}} (w : \|w - P_{L_n} w_0\|_2 \leq c_1 \varepsilon_{\alpha_0,n}).$$

On the unit cube, we can further lower bound this quantity by,

$$\Pi_{W_{\alpha,n}} (w : \|w - P_{L_n} w_0\|_\infty \leq c_1 \varepsilon_{\alpha_0,n}),$$

which, by the decentering-inequality (32) in [Giordano \(2023\)](#) for the rescaled Besov-Laplace random element $W_{\alpha,n}$, is greater than

$$\begin{aligned} e^{-\|P_{L_n} w_0\|_{B_1^\alpha} n \varepsilon_{\alpha,n}^2} \Pi_{W_{\alpha,n}} (w : \|w\|_\infty \leq c_1 \varepsilon_{\alpha_0,n}) \\ \geq e^{-c_2 n \varepsilon_{\alpha_0,n}^2} \Pi_{W_\alpha} \left(w : \|w\|_\infty \leq c_1 n \varepsilon_{\alpha_0,n} \varepsilon_{\alpha,n}^2 \right). \end{aligned}$$

By the centered small ball inequality (34) in [Giordano \(2023\)](#) (noting that W_α coincides with W there with the choice $t = \alpha - d > 0$),

$$\Pi_{W_\alpha} \left(w : \|w\|_\infty \leq c_1 n \varepsilon_{\alpha_0,n} \varepsilon_{\alpha,n}^2 \right) \geq e^{-(c_3(\alpha-d)+c_4)(c_1 n \varepsilon_{\alpha_0,n} \varepsilon_{\alpha,n}^2)^{-d/(\alpha-d)}} \geq e^{-c_5 n \varepsilon_{\alpha_0,n}^2},$$

for $c_3, c_4, c_5 > 0$. Here, we used the fact that

$$\begin{aligned} \left(n \varepsilon_{\alpha_0,n} \varepsilon_{\alpha,n}^2 \right)^{-\frac{d}{\alpha-d}} &= \left(n n^{-\frac{\alpha_0}{2\alpha_0+d}} n^{\frac{2\alpha}{2\alpha+d}} \right)^{-\frac{d}{\alpha-d}} \\ &= \left(n^{\frac{4\alpha\alpha_0+2d\alpha_0+2d\alpha+d^2-2\alpha\alpha_0-d\alpha_0-4\alpha\alpha_0-2d\alpha}{(2\alpha+d)(2\alpha_0+d)}} \right)^{-\frac{d}{\alpha-d}} \end{aligned}$$

$$= \left(n^{\frac{\alpha_0 d + d^2 - 2\alpha_0}{(2\alpha+d)(2\alpha_0+d)}} \right)^{-\frac{d}{\alpha-d}} = \left(n^{\frac{d}{2\alpha_0+d}} \right)^{\frac{2\alpha\alpha_0 - d\alpha_0 - d^2}{(\alpha-d)(2\alpha+d)}} \leq n\varepsilon_{\alpha_0,n}^2$$

which holds provided the exponent is smaller than 1, since $n\varepsilon_{\alpha_0,n}^2 = n^{d/2\alpha_0+d}$. To see this, note that

$$(\alpha - d)(2\alpha + d) - (2\alpha\alpha_0 - d\alpha_0 - d^2) \geq 0$$

following from $\alpha \geq \alpha_0 \geq d$. Combining the previous bounds, we find that for all $\alpha \in [\alpha_0, \alpha_0 + 1/\log n]$, for sufficiently large $B_1 > 0$,

$$\Pi_{W_{\alpha,n}}(w : \|w - w_0\|_\infty \leq B_1\varepsilon_{\alpha_0,n}) \geq e^{-c_6 n\varepsilon_{\alpha_0,n}^2}.$$

As a result,

$$\begin{aligned} & \int_{\alpha_0}^{\alpha_0+1/\log n} \Pi_{W_{\alpha,n}}(w : \|w - w_0\|_2 \leq B_1\varepsilon_{\alpha_0,n}) \sigma_n(\alpha) d\alpha \\ & \geq \int_{\alpha_0}^{\alpha_0+\frac{1}{\log n}} e^{-c_6 n\varepsilon_{\alpha_0,n}^2} \sigma_n(\alpha) d\alpha \\ & = e^{-c_6 n\varepsilon_{\alpha_0,n}^2} \int_{\alpha_0}^{\alpha_0+\frac{1}{\log n}} \frac{e^{-n\varepsilon_{\alpha,n}^2}}{\zeta_n} d\alpha \\ & \geq \frac{1}{\log n} \times \frac{e^{-n\varepsilon_{\alpha_0,n}^2}}{\zeta_n} \gtrsim (\log n)^{-2} e^{-n\varepsilon_{\alpha_0,n}^2} \geq e^{-c_7 n\varepsilon_{\alpha_0,n}^2}, \end{aligned}$$

for some $c_7 > 0$, where we used that $\sigma_n(\alpha)$ is increasing in α , that the length of the integration interval is $1/\log n$ and that the normalization constant of σ_n satisfies $\zeta_n \simeq \log n$. The claim then follows taking $B_2 = c_6 + c_7 > 0$. \square

Lemma 3 Let Π_{W_n} be the law of the hierarchical rescaled Besov-Laplace random element W_n from (3) with smoothness hyper-prior $\alpha \sim \Sigma_n$ as in (4). For fixed $\alpha_0 > d$, and $A_1, A_2 > 0$, let $\alpha_* = \alpha_0/(1 + A_2/\log n)$ and define the set

$$\mathcal{W}_n = \left\{ w = w^{(1)} + w^{(2)}, \|w^{(1)}\|_1 \leq A_1 n^{-\frac{\alpha_*}{2\alpha_*+d}}, \|w^{(2)}\|_{B_1^{\alpha_*+d}} \leq A_2 n^{\frac{d}{2\alpha_*+d}} \right\}. \quad (\text{A5})$$

Then, for all $K > 0$, there exist sufficiently large A_1, A_2 , such that for $n \in \mathbb{N}$ large enough,

$$\Pi_{W_n}(\mathcal{W}_n^c) \leq e^{-Kn^{d/(2\alpha_0+d)}}.$$

Proof For all $A_1, A_2 > 0$, the probability of interest is equal to

$$\int_d^{\alpha_*} \Pi_{W_{\alpha,n}}(\mathcal{W}_n^c) \sigma_n(\alpha) d\alpha + \int_{\alpha_*}^{\log n} \Pi_{W_{\alpha,n}}(\mathcal{W}_n^c) \sigma_n(\alpha) d\alpha. \quad (\text{A6})$$

We start analyzing the first integral, which is smaller than

$$\int_d^{\alpha_*} \sigma_n(\alpha) d\alpha \leq \alpha_* \frac{e^{-n\varepsilon_{\alpha_*,n}^2}}{\zeta_n} \leq e^{-c_1 n\varepsilon_{\alpha_*,n}^2},$$

having upper bounded the size of the integration interval by α_* and exploited the fact that the hyper-prior p.d.f. σ_n is increasing in α and has normalizing constant $\zeta_n \simeq \log n$. By comparing $n\varepsilon_{\alpha_*,n}^2$ and $n\varepsilon_{\alpha_0,n}^2$, we obtain,

$$\frac{n\varepsilon_{\alpha_*,n}^2}{n\varepsilon_{\alpha_0,n}^2} = n^{\frac{dA_2+d\log n}{dA_2+(2\alpha_0+d)\log n} - \frac{d}{2\alpha_0+d}}$$

$$= n^{\frac{2dA_2\alpha_0}{(2\alpha_0+d)^2 \log n + dA_2(2\alpha_0+d)}} = e^{\frac{2dA_2\alpha_0 \log n}{(2\alpha_0+d)^2 \log n + dA_2(2\alpha_0+d)}},$$

which shows that, for $c_2 := e^{\frac{2dA_2\alpha_0}{(2\alpha_0+d)^2}} > 1$,

$$\begin{aligned} n\varepsilon_{\alpha_*,n}^2 &= n\varepsilon_{\alpha_0,n}^2 e^{\frac{2dA_2\alpha_0 \log n}{(2\alpha_0+d)^2 \log n + dA_2(2\alpha_0+d)}} \\ &= n\varepsilon_{\alpha_0,n}^2 e^{\frac{dA_2\alpha_0}{(2\alpha_0+d)^2}} e^{\frac{dA_2\alpha_0(\log n(2\alpha_0+d) - dA_2)}{(2\alpha_0+d)^2(\log n(2\alpha_0+d) + dA_2)}} \geq \sqrt{c_2} n\varepsilon_{\alpha_0,n}^2, \end{aligned}$$

as well as

$$n\varepsilon_{\alpha_0,n}^2 e^{\frac{2dA_2\alpha_0}{(2\alpha_0+d)^2}} e^{-\frac{2d^2 A_2^2 \alpha_0}{(2\alpha_0+d)^2(\log n(2\alpha_0+d) + dA_2)}} \leq c_2 n\varepsilon_{\alpha_0,n}^2$$

holding for all $n \in \mathbb{N}$ large enough. For any $K > 0$, we can then take A_2 large enough so that

$$\int_d^{\alpha_*} \Pi_{W_{\alpha,n}}(\mathcal{W}_n^c) \sigma_n(\alpha) d\alpha \leq e^{-c_1 n \varepsilon_{\alpha_*,n}^2} \leq e^{-c_1 \sqrt{c_2} n \varepsilon_{\alpha_0,n}^2} \leq e^{-(K+1) n \varepsilon_{\alpha_0,n}^2}.$$

Concerning the second integral in (A6), we write

$$\begin{aligned} \Pi_{W_{\alpha,n}}(\mathcal{W}_n) \\ = \Pi_{W_\alpha} \left(w = w^{(1)} + w^{(2)} : \|w^{(1)}\|_1 \leq A_1 n \varepsilon_{\alpha_*,n} \varepsilon_{\alpha,n}^2, \|w^{(2)}\|_{B_1^{\alpha_*+d}} \leq A_1 n^2 \varepsilon_{\alpha_*,n}^2 \varepsilon_{\alpha,n}^2 \right). \end{aligned}$$

Letting

$$\begin{aligned} \bar{\mathcal{W}}_n := \left\{ \bar{w} = \bar{w}^{(1)} + \bar{w}^{(2)} + \bar{w}^{(3)} : \|\bar{w}^{(1)}\|_1 \leq n \varepsilon_{\alpha_*,n} \varepsilon_{\alpha,n}^2, \|\bar{w}^{(2)}\|_{H^{\alpha-d/2}} \leq \sqrt{\bar{A}_1 n \varepsilon_{\alpha_*,n}^2}, \right. \\ \left. \|\bar{w}^{(3)}\|_{B_1^\alpha} \leq \bar{A}_1 n \varepsilon_{\alpha_*,n}^2 \right\}, \end{aligned}$$

the two-level concentration inequality (33) in Giordano (2023) implies, using again (A7), for $c_3, c_4 > 0$,

$$\begin{aligned} \Pi_{W_\alpha}(\bar{\mathcal{W}}_n) &\geq 1 - \frac{1}{\Pi_{W_\alpha}(w : \|w\|_1 \leq n \varepsilon_{\alpha_*,n} \varepsilon_{\alpha,n}^2)} e^{-c_3 \bar{A}_1 n \varepsilon_{\alpha_*,n}^2} \\ &\geq 1 - \frac{1}{\Pi_{W_\alpha}(w : \|w\|_1 \leq n \varepsilon_{\alpha_*,n} \varepsilon_{\alpha,n}^2)} e^{-c_4 \bar{A}_1 n \varepsilon_{\alpha_0,n}^2}. \end{aligned}$$

As $\|w\|_1 \leq \|w\|_\infty$, noting that $\alpha \geq \alpha_* = \alpha_0 \log n / (M + \log n) > d$ for all n large enough since $\alpha_0 > d$, by the centred small ball inequality (34) in Giordano (2023), we have for all $\alpha_* < \alpha \leq \log n$,

$$\begin{aligned} \Pi_{W_\alpha} \left(w : \|w\|_1 \leq n \varepsilon_{\alpha_*,n} \varepsilon_{\alpha,n}^2 \right) &\geq e^{-(c_5(\alpha-d) + c_6)(n \varepsilon_{\alpha_*,n} \varepsilon_{\alpha,n}^2)^{-d/(\alpha-d)}} \\ &\geq e^{-c_7 \log n (n \varepsilon_{\alpha_*,n} \varepsilon_{\alpha,n}^2)^{-d/(\alpha-d)}}. \end{aligned}$$

Using again (A7) and the fact that

$$\begin{aligned} \log n \left(n \varepsilon_{\alpha_*,n} \varepsilon_{\alpha,n}^2 \right)^{-\frac{d}{\alpha-d}} &= \log n \left(n^{\frac{-2\alpha\alpha_* - d\alpha_* + 2d\alpha_* + d^2}{(2\alpha_*+d)(2\alpha+d)}} \right)^{-\frac{d}{\alpha-d}} \\ &= \log n \left(n \varepsilon_{\alpha_*,n}^2 \right)^{\frac{2\alpha\alpha_* - d\alpha_* - d^2}{(2\alpha+d)(\alpha-d)}} \leq n \varepsilon_{\alpha_*,n}^2, \end{aligned}$$

as the last exponent is strictly smaller than one, by virtue of $\alpha \geq \alpha_*$, we obtain

$$\Pi_{W_\alpha} \left(w : \|w\|_1 \leq n \varepsilon_{\alpha_*,n} \varepsilon_{\alpha,n}^2 \right) \geq e^{-c_7 n \varepsilon_{\alpha_*,n}^2} \geq e^{-c_8 n \varepsilon_{\alpha_0,n}^2}.$$

For sufficiently large $\bar{A}_1 > 0$, it then follows that for all $\alpha \in [\alpha_*, \log n]$

$$\Pi_{W_\alpha}(\bar{\mathcal{W}}_n) \geq 1 - e^{-(c_4 \bar{A}_1 - c_8)n \varepsilon_{\alpha_0,n}^2} \geq 1 - e^{-(K+1)n \varepsilon_{\alpha_0,n}^2}. \quad (\text{A7})$$

Now let $P_{L_n} \bar{w}^{(2)}$ be the wavelet projection of the term $\bar{w}^{(2)}$ in the definition of $\bar{\mathcal{W}}_n$, at resolution $L_n \in \mathbb{N}$ such that $L_n \simeq n^{\frac{d}{2\alpha+d}}$. Then,

$$\begin{aligned}\|\bar{w}^{(2)} - P_{L_n} \bar{w}^{(2)}\|_1 &\leq L_n^{-\left(\frac{\alpha}{d}-\frac{1}{2}\right)} \|\bar{w}^{(2)}\|_{H^{\alpha-d/2}} \\ &\lesssim n^{-\frac{\alpha-d/2}{2\alpha+d}} \sqrt{n} \varepsilon_{\alpha_*, n} = n^{\frac{d}{2\alpha+d}} \varepsilon_{\alpha_*, n} = n \varepsilon_{\alpha, n}^2 \varepsilon_{\alpha_*, n}.\end{aligned}$$

Also,

$$\begin{aligned}\|P_{L_n} \bar{w}^{(2)}\|_{B_1^\alpha} &\lesssim \sqrt{L_n} \|\bar{w}^{(2)}\|_{H^{\alpha-d/2}} \\ &\lesssim n^{\frac{d/2}{2\alpha+d}} n^{\frac{d/2}{2\alpha_*+d}} = n^{\frac{d\alpha_*+d^2/2+d\alpha+d^2/2}{(2\alpha+d)(2\alpha_*+d)}} \leq n^{\frac{d(d+\alpha_*)}{(2\alpha+d)(2\alpha_*+d)}} \leq n \varepsilon_{\alpha_*, n}^2\end{aligned}$$

since $d + \alpha + \alpha_* < 2\alpha + d$ whenever $\alpha \geq \alpha_*$. Thus, setting $\tilde{w}^{(1)} := \bar{w}^{(1)} + (\bar{w}^{(2)} - P_{L_n} \bar{w}^{(2)})$ and $\tilde{w}^{(2)} := \bar{w}^{(3)} + P_{L_n} \bar{w}^{(2)}$ yields that for all $\alpha \in [\alpha_*, \log n]$ and all n and A_1 large enough,

$$\bar{\mathcal{W}}_n \subseteq \widetilde{\mathcal{W}}_n := \{\tilde{w} = \tilde{w}^{(1)} + \tilde{w}^{(2)} : \|\tilde{w}^{(1)}\|_1 \leq \tilde{A}_1 n \varepsilon_{\alpha, n}^2 \varepsilon_{\alpha_*, n}, \|\tilde{w}^{(2)}\|_{B_1^\alpha} \leq \tilde{A}_1 n \varepsilon_{\alpha_*, n}^2\}.$$

In view of (A7),

$$\Pi_{W_\alpha}(\widetilde{\mathcal{W}}_n) \geq 1 - e^{-(K+1)n\varepsilon_{\alpha_0, n}^2}. \quad (\text{A8})$$

We conclude showing that, choosing sufficiently large $A_1 > 0$,

$$\widetilde{\mathcal{W}}_n \subseteq \{w = w^{(1)} + w^{(2)} : \|w^{(1)}\|_1 \leq A_1 n \varepsilon_{\alpha, n}^2 \varepsilon_{\alpha_*, n}, \|w^{(2)}\|_{B_1^{\alpha_*+d}} \leq A_1 n^2 \varepsilon_{\alpha, n}^2 \varepsilon_{\alpha_*, n}^2\} \quad (\text{A9})$$

for all $\alpha \in [\alpha_*, \log n]$ and all $n \in \mathbb{N}$ large enough. To this aim, we start with the high-regularity case, $\alpha \in [\alpha_* + d, \log n]$. Then

$$\|\tilde{w}^{(2)}\|_{B_1^{\alpha_*+d}} \leq \|\tilde{w}^{(2)}\|_{B_1^\alpha} \leq \tilde{R} n \varepsilon_{\alpha_*, n}^2 \leq \tilde{R} n^2 \varepsilon_{\alpha, n}^2 \varepsilon_{\alpha_*, n}^2$$

since $n \varepsilon_{\alpha, n}^2 \rightarrow \infty$. The inclusion (A9) thus follows with $w^{(1)} = \tilde{w}^{(1)}$, $w^{(2)} = \tilde{w}^{(2)}$, and $R = \tilde{R}$. For the remaining range $\alpha \in [\alpha_*, \alpha_* + d)$, we consider the wavelet projection $P_{L_n} \tilde{w}^{(2)}$ of $\tilde{w}^{(2)}$ at resolution level $L_n \simeq n^{\frac{d^2}{(2\alpha+d)(\alpha_*+d-\alpha)}}$. Then,

$$\|P_{L_n} \tilde{w}^{(2)}\|_{B_1^{\alpha_*+d}} \leq L_n^{\frac{(\alpha_*+d-\alpha)}{d}} \|\tilde{w}^{(2)}\|_{B_1^\alpha} \lesssim n^{\frac{d}{(2\alpha+d)}} n \varepsilon_{\alpha_*, n}^2 = n^2 \varepsilon_{\alpha, n}^2 \varepsilon_{\alpha_*, n}^2,$$

and, using the continuous embedding of B_1^0 into L^1 (e.g., eq. (21), p.169 in [Schmeisser and Triebel \(1987\)](#)),

$$\begin{aligned}\|\tilde{w}^{(2)} - P_{L_n} \tilde{w}^{(2)}\|_1 &\lesssim \|\tilde{w}^{(2)} - P_{L_n} \tilde{w}^{(2)}\|_{B_1^0} \\ &\leq L_n^{-\frac{\alpha}{d}} \|\tilde{w}^{(2)}\|_{B_1^\alpha} \\ &\lesssim n^{-\frac{d\alpha}{(2\alpha+d)(\alpha_*+d-\alpha)}} n^{\frac{d}{2\alpha_*+d}} = n^{\frac{d^3+\alpha_*d^2-2\alpha^2d}{(2\alpha+d)(2\alpha_*+d)(\alpha_*+d-\alpha)}}.\end{aligned}$$

The inclusion (A9) follows upon showing that the right hand side is smaller than

$$n \varepsilon_{\alpha, n}^2 \varepsilon_{\alpha_*, n} = n^{\frac{d}{2\alpha+d}} n^{-\frac{\alpha_*}{2\alpha_*+d}} = n^{\frac{d^2+\alpha_*d-2\alpha\alpha_*}{(2\alpha_*+d)(2\alpha+d)}} = n^{\frac{d^3+\alpha_*d^2-\alpha^2d-\alpha\alpha_*d+2\alpha^2\alpha_*-2\alpha\alpha_*^2}{(2\alpha+d)(2\alpha_*+d)(\alpha_*+d-\alpha)}}.$$

Indeed, the difference between the numerators of the exponents equals

$$\begin{aligned}\Delta(\alpha) &= d^3 + \alpha_* d^2 - 2\alpha^2 d - d^3 - \alpha_* d^2 + \alpha^2 d + \alpha \alpha_* d - 2\alpha^2 \alpha_* + 2\alpha \alpha_*^2 \\ &= -(2\alpha_* + d)\alpha^2 + (2\alpha_*^2 + d\alpha_*)\alpha,\end{aligned}$$

which, as a function of α , is a downward-pointing parabola with maximum attained at

$$\alpha_v := \frac{2\alpha_*^2 + d\alpha_*}{4\alpha_* + d} < \alpha_*.$$

Since $\Delta(\alpha)$ is decreasing for $\alpha > \alpha_*$, for all $\alpha \in [\alpha_*, \alpha_* + d]$,

$$\begin{aligned}\Delta(\alpha) &\leq \Delta(\alpha_*) \\ &= -(2\alpha_* + d)\alpha_*^2 + (2\alpha_*^2 + d\alpha_*)\alpha_* = 0.\end{aligned}$$

This shows as required that $\|\tilde{w}^{(2)} - P_{L_n} \tilde{w}^{(2)}\|_1 \lesssim n\varepsilon_{\alpha,n}^2 \varepsilon_{\alpha_*,n}$, so that taking $w^{(1)} := \tilde{w}^{(1)} + (\tilde{w}^{(2)} - P_{L_n} \tilde{w}^{(2)})$ and $w^{(2)} := P_{L_n} \tilde{w}^{(2)}$, the desired inclusion (A9) follows for large enough $A_1 > 0$. By (A8), we then conclude

$$\Pi_{W_{\alpha,n}}(\mathcal{W}_n) \geq \Pi_{W_\alpha}(\widetilde{\mathcal{W}}_n) \geq 1 - e^{-(K+1)n\varepsilon_{\alpha_0,n}^2}.$$

Finally, combined with (A6), this yield

$$\begin{aligned}\Pi_{W_n}(\mathcal{W}_n^c) &\leq e^{-(K+1)n\varepsilon_{\alpha_0,n}^2} + \int_{\alpha_*}^{\log n} e^{-(K+1)n\varepsilon_{\alpha_0,n}^2} \sigma_n(\alpha) d\alpha \\ &\leq 2e^{-(K+1)n\varepsilon_{\alpha_0,n}^2} \leq e^{-Kn\varepsilon_{\alpha_0,n}^2}.\end{aligned}$$

□

References

- Agapiou S, Dashti M, Helin T. Rates of contraction of posterior distributions based on p -exponential priors. Bernoulli. 2021;27(3):1616 – 1642. <https://doi.org/10.3150/20-BEJ1285>, <https://doi.org/10.3150/20-BEJ1285>.
- Agapiou S, Savva A. Adaptive inference over Besov spaces in the white noise model using p -exponential priors. Bernoulli. 2024;30(3):2275–2300. <https://doi.org/https://doi.org/10.3150/23-BEJ1673>.
- Agapiou S, Wang S. Laplace priors and spatial inhomogeneity in Bayesian inverse problems. Bernoulli. 2024;30(2):878–910. <https://doi.org/https://doi.org/10.3150/22-BEJ1563>.
- Arbel J, Gayraud G, Rousseau J. Bayesian optimal adaptive estimation using a sieve prior. Scand J Stat. 2013;40(3):549–570. <https://doi.org/10.1002/sjos.12002>, <https://doi.org/10.1002/sjos.12002>.
- Bioucas-Dias JM. Bayesian wavelet-based image deconvolution: a GEM algorithm exploiting a class of heavy-tailed priors. IEEE Trans Image Process. 2006;15(4):937–951. <https://ezproxy-prd.bodleian.ox.ac.uk:2102/10.1109/TIP.2005.863972>, <https://doi.org/10.1109/TIP.2005.863972>.
- Bourdaud G, Sickel W. Composition operators on function spaces with fractional order of smoothness. In: Harmonic analysis and nonlinear partial differential equations RIMS Kôkyûroku Bessatsu, B26. Kyoto: Res. Inst. Math. Sci. (RIMS); 2011. p. 93–132.
- Bui-Thanh T, Ghattas O. A scalable algorithm for MAP estimators in Bayesian inverse problems with Besov priors. Inverse Probl Imaging. 2015;9(1):27–53. <https://doi-org.ezp.lib.cam.ac.uk/10.3934/ipi.2015.9.27>, <https://doi.org/10.3934/ipi.2015.9.27>.

Castillo I, Nickl R. Nonparametric Bernstein–von Mises Theorems in Gaussian white noise. *Ann Statist.* 2013;41(4):1999–2028. <https://doi.org/https://doi.org/10.1214/13-AOS1133>.

Castillo I, Nickl R. On the Bernstein–von Mises Phenomenon for nonparametric Bayes procedures. *Ann Statist.* 2014;42(5):1941–1969. <https://doi.org/https://doi.org/10.1214/aos/1017938917>.

Chen V, Dunlop M, Papaspiliopoulos O, Stuart A. Robust MCMC Sampling with Non-Gaussian and Hierarchical Priors in High Dimensions. arXiv preprint arXiv:180303344. 2018 03;<https://doi.org/https://arxiv.org/abs/1803.03344>.

Choudhuri N, Ghosal S, Roy A. Nonparametric binary regression using a Gaussian process prior. *Stat Methodol.* 2007;4(2):227–243. <https://doi.org/https://doi.org/10.1016/j.stamet.2006.07.003>.

Cotter SL, Roberts GO, Stuart AM, White D. MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster. *Stat Sci.* 2013;28(3):424–446. <https://doi.org/https://doi.org/10.1214/13-STS421>.

Cox DD. An analysis of Bayesian inference for nonparametric regression. *Ann Statist.* 1993;21(2):903–923. <https://doi-org.ezp.lib.cam.ac.uk/10.1214/aos/1176349157>, <https://doi.org/10.1214/aos/1176349157>.

Dolera E, Favaro S, Giordano M. On strong posterior contraction rates for Besov–Laplace priors in the white noise model. arXiv preprint arXiv:241106981. 2024;<https://doi.org/https://doi.org/10.48550/arXiv.2411.06981>.

Donoho DL, Johnstone IM. Minimax estimation via wavelet shrinkage. *Ann Statist.* 1998;26(3):879–921. <https://doi.org/10.1214/aos/1024691081>, <https://doi.org/10.1214/aos/1024691081>.

Gelfand AE, Kuo L. Nonparametric Bayesian bioassay including ordered polytymous response. *Biometrika.* 1991;78(3):657–666. <https://doi.org/https://doi.org/10.2307/2337035>.

Ghosal S, Ghosh JK, van der Vaart AW. Convergence rates of posterior distributions. *Ann Statist.* 2000;28(2):500–531. <https://doi.org/https://doi.org/10.1214/aos/1016218228>.

Ghosal S, Roy A. Posterior Consistency of Gaussian Process Prior for Nonparametric Binary Regression. *Ann Statist.* 2006;p. 2413–2429. <https://doi.org/https://doi.org/10.1214/009053606000000795>.

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

- Ghosal S, van der Vaart AW. Fundamentals of Nonparametric Bayesian Inference. New York: Cambridge University Press; 2017.
- Giné E, Nickl R. Rates of contraction for posterior distributions in L^r -metrics, $1 \leq r \leq \infty$. Ann Statist. 2011;39(6):2883–2911. <https://doi.org/https://doi.org/10.1214/11-AOS924>.
- Giné E, Nickl R. Mathematical foundations of infinite-dimensional statistical models. New York: Cambridge University Press; 2016.
- Giordano M. Besov-Laplace priors in density estimation: optimal posterior contraction rates and adaptation. Electron J Stat. 2023;17(2):2210 – 2249. <https://doi.org/10.1214/23-EJS2161>, <https://doi.org/10.1214/23-EJS2161>.
- Giordano M. Bayesian Inference with Besov-Laplace Priors for Spatially Inhomogeneous Binary Classification Surfaces. Studies in classification, data analysis, and knowledge organization. 2025;p. 202–213. https://doi.org/https://doi.org/10.1007/978-3-032-03042-9_18.
- Giordano M, Ray K. Nonparametric Bayesian inference for reversible multidimensional diffusions. Ann Statist. 2022;50(5):2872–2898. <https://doi.org/10.1214/22-aos2213>, <https://doi.org/10.1214/22-aos2213>.
- Giordano M, Ray K. Semiparametric Bernstein-von Mises theorems for reversible diffusions. arXiv preprint arXiv:250516275. 2025;<https://doi.org/https://doi.org/10.48550/arXiv.2505.16275>.
- Giordano M, Ray K, Schmidt-Hieber J. On the inability of Gaussian process regression to optimally learn compositional functions. A Adv Neural Inf Process. 2022;35:22341–22353. <https://doi.org/https://doi.org/10.48550/arXiv.2205.07764>.
- Jara A, Garcia-Zattera MJ, Lesaffre E. A Dirichlet process mixture model for the analysis of correlated binary responses. Comput Stat Data Anal. 2007;51(11):5402–5415. <https://doi.org/https://doi.org/10.1016/j.csda.2006.09.010>.
- Kekkonen H, Lassas M, Saksman E, Siltanen S. Random tree Besov priors—towards fractal imaging. Inverse Probl Imaging. 2023;17(2):507–531. <https://doi.org/https://doi.org/10.3934/ipi.2022059>.
- Lassas M, Saksman E, Siltanen S. Discretization-invariant Bayesian inversion and Besov space priors. Inverse Probl Imaging. 2009;3(1):87–122. <http://dx.doi.org/10.3934/ipi.2009.3.87>, <https://doi.org/10.3934/ipi.2009.3.87>.
- Lember J, van der Vaart A. On universal Bayesian adaptation. Statistics & Decisions International Mathematical Journal for Stochastic Methods and Models. 2007;25(2):127–152. <https://doi.org/https://doi.org/10.1524/stnd.2007.25.2.127>.

Lenk PJ. The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J Am Stat Assoc.* 1988;83(402):509–516. <https://doi.org/https://doi.org/10.2307/2288870>.

Leporini D, Pesquet JC. Bayesian wavelet denoising: Besov priors and non-Gaussian noises. *Signal Process.* 2001;81(1):55–67. <https://www.sciencedirect.com/science/article/pii/S0165168400001900>, [https://doi.org/https://doi.org/10.1016/S0165-1684\(00\)00190-0](https://doi.org/https://doi.org/10.1016/S0165-1684(00)00190-0).

Nickisch H, Rasmussen CE. Approximations for binary Gaussian process classification. *J Mach Learn Res.* 2008;9(10):2035–2078. <https://doi.org/http://jmlr.org/papers/v9/nickisch08a.html>.

Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Adaptive Computation and Machine Learning, Cambridge, MA: MIT Press; 2006.

Ray K. Bayesian inverse problems with non-conjugate priors. *Electron J Stat.* 2013;7:2516–2549. <https://doi.org/http://dx.doi.org/10.1214/13-EJS851>.

Rudin LI, Osher S, Fatemi E. Nonlinear total variation based noise removal algorithms. In: Experimental mathematics: computational issues in nonlinear science, vol. 60 Los Alamos (NM); 1992. p. 259–268.

Sakhaee E, Entezari A. Spline-based sparse tomographic reconstruction with Besov priors. In: Ourselin S, Styner MA, editors. Medical Imaging 2015: Image Processing, vol. 9413 Orlando, Florida, United States: SPIE; 2015. p. 101 – 108.

Schmeisser HJ, Triebel H. Topics in Fourier analysis and function spaces. A Wiley-Interscience Publication, Chichester: John Wiley & Sons, Ltd.; 1987.

Shen X, Wasserman L. Rates of convergence of posterior distributions. *Ann Statist.* 2001;p. 687–714. <https://doi.org/https://doi.org/10.1214/aos/1009210686>.

Tsybakov AB. Introduction to nonparametric estimation. Springer Series in Statistics, New York: Springer; 2009.

van der Vaart AW, van Zanten JH. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann Statist.* 2008;36(3):1435–1463. <https://doi.org/https://doi.org/10.1214/009053607000000613>.

van der Vaart AW, van Zanten JH. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann Statist.* 2009;37(5B):2655–2675. <https://doi.org/10.1214/08-AOS678>, <https://doi.org/10.1214/08-AOS678>.

Vänskä S, Lassas M, Siltanen S. Statistical X-ray tomography using empirical Besov priors. *Int J Tomogr Stat.* 2009;11(S09):3–32. <https://doi.org/http://www.ceser.in/ceserp/index.php/ijts/article/view/67>.

van Waaij J, van Zanten H. Gaussian process methods for one-dimensional diffusions: Optimal rates and adaptation. *Electron J Stat.* 2016;10(1):628–645. <https://doi.org/https://doi.org/10.1214/16-EJS1117>.

Wang C, Liao X, Carin L, Dunson DB, Blei D. Classification with Incomplete Data Using Dirichlet Process Priors. *J Mach Learn Res.* 2010;11(12). [https://doi.org/http://jmlr.org/papers/v11/wang10a.html](http://jmlr.org/papers/v11/wang10a.html).