Ideal Bayesian Spatial Adaptation

Veronika Ročková and Judith Rousseau

January 30, 2023

Abstract

Many real-life applications involve estimation of curves that exhibit complicated shapes including jumps or varying-frequency oscillations. Practical methods have been devised that can adapt to a locally varying complexity of an unknown function (e.g. variable-knot splines, sparse wavelet reconstructions, kernel methods or trees/forests). However, the overwhelming majority of existing asymptotic minimaxity theory is predicated on homogeneous smoothness assumptions. Focusing on locally Hölderian functions, we provide new *locally adaptive* posterior concentration rate results under the supremum loss for widely used Bayesian machine learning techniques in white noise and non-parametric regression. In particular, we show that popular spike-and-slab priors and Bayesian CART are uniformly locally adaptive. In addition, we propose a new class of repulsive partitioning priors which relate to variable knot splines and which are exact-rate adaptive. For uncertainty quantification, we construct locally adaptive confidence bands whose width depends on the local smoothness and which achieve uniform asymptotic coverage under local self-similarity. To illustrate that spatial adaptation is not at all automatic, we provide lower-bound results showing that popular hierarchical Gaussian process priors fall short of spatial adaptation.

Keywords: *Bayesian CART, Partitioning Priors, Supremum Norm, Spike-and-Slab, Spatial Adaptation*

1 Spatial Adaptation

The key to practically successful curve estimation is the ability to adapt to subtle qualitative structures of the analyzed curve. Very often, interesting aspects of the estimated curve are related to spatial inhomogeneities, e.g. discontinuities or oscillations with varying frequency and/or amplitude. There is a wealth of techniques for function estimation (e.g.

kernel methods, local polynomial fitting, nearest neighbor techniques or splines) which exert various degrees of global and local adaptation. For functions with a *locally* varying complexity, however, global procedures can be woefully inaccurate, leading to overfitting smooth domains and underfitting wiggly domains.

While such weaknesses have long been recognized, the applied statistics community has has not embraced locally adaptive techniques as widely. To remediate this absence, this paper provides a methodological and theoretical inferential framework valuable for practitioners. Numerous methodological developments have spawned that are capable of local adaptation, e.g., local polynomial regression [34] or kernel estimation [10, 48, 52] with a local bandwidth selection. In the context of spline smoothing, [50] suggested adaptively selecting subsets of basis functions which pertains to selective wavelet reconstructions [22, 31] and variable knot spline techniques [25, 35, 36, 68]. Notably, [51, 71] proposed (total-variation) penalized least square estimates which correspond to regression splines with data-adaptive knot points. An alternative approach is to allow the smoothing parameter to vary locally (see [57] for piecewise constant smoothing parameters). For example, [63] suggest spline fitting with a roughness penalty whose logarithm is itself a linear spline with knot values chosen by cross-validation. Variants of such spatially adaptive penalty parameters have been widely used in practice [3, 24, 45, 46]. Besides splines and wavelets, tree based methods (CART [9, 20, 27], random forests [8] or BART [21]) are particularly appealing for recovering spatially inhomogeneous functions by adapting the placement of splits to the function itself via recursive partitioning [29]. Deep learning methods are also expected to perform well for structured curves [40, 69].

From a methodological perspective, spatially adaptive curve estimation has been tackled rather broadly. From a theoretical perspective, however, there are still gaps in understanding whether these techniques are indeed optimal and (uniformly) spatially adaptive. The majority of existing asymptotic minimaxity theory (for density or regression function estimation) is predicated on homogeneous smoothness assumptions. For example, existing

results for random forests [73, 74], deep learning [58, 64], Bayesian forests [42, 62] and other non-parametric methods such as Gaussian processes [55, 72] have been concerned with convergence rates for spatially homogeneous Hölderian functions under the L^2 global estimation loss. Here, we extend the scope of such theoretical results in two important directions. First, we focus on *both* global and local (supremum) loss providing results for uniform local adaptation. Second, we provide a frequentist framework for uncertainty quantification in the form of locally adaptive bands. Our goal is to investigate the extent to which widely used Bayesian priors (spike-and-slab priors [22, 41, 59, 75], Gaussian process priors [2, 6, 44, 66, 70] and Bayesian CART priors [18, 20, 27]) can optimally adapt to local smoothness. Before listing our contributions, we review existing theoretical results for spatially inhomogeneous functions.

The first natural question is how well an estimator performs *globally*. For the stereotypical Besov classes $B^{\alpha}_{p,q}$, one way to assess the global quality of an estimator is in terms of a L^r loss that is sharper than the norm of the Besov functional class (i.e. $p < r < \infty$), see e.g. [32] and [48]. For $p < 2$, linear estimators are known to be incapable of achieving the optimal rate [32]. For a discussion on minimax rates in Besov spaces, we refer to [30] and [26]. Unlike linear estimators, wavelet thresholding offers a powerful technology for spatially adaptive curve estimation [30]. In particular, [31] describe a selective wavelet reconstruction method called RiskShrink based on shrinkage of wavelet coefficients and show that this procedure mimics an oracle 'as well as it is possible to do so'. RiskShrink is an automatic model selection method which picks a subset of wavelet vectors and fits a model consisting only of wavelets in that subset. In this work, we investigate Bayesian variants of such strategies. Positive findings for global estimation in Besov spaces have also been reported for deep learning [40, 69], penalized least squares [51], locally variable kernels [52]. Notably, [48] propose a kernel estimator with a variable data-driven bandwidth that achieves the minimax rate of estimation over a wide scale of Besov classes and hence shares rate optimality properties with wavelet estimators.

Another, and perhaps more transparent, approach is to assess the quality of an estimator *locally*. For density estimation, [37] study adaptivity to heterogeneous smoothness, simultaneously for every point in a fixed interval, in a supremum-norm loss. The authors consider a certain notion of pointwise Hölder continuity and study dyadic histogram estimators with a variable bin size and with a Lepski-type adaptation. We adopt a similar estimation setup here, but approach it entirely from a Bayesian perspective.

Practical deployments of the Lepski-based adaptation require tuning parameters (especially of the threshold used for comparing two estimates from different scales) for which theoretical justifications may not always be available [37, 43]. Bayesian procedures, on the other hand, are known to adapt automatically to the unknown aspects of the estimation problem, even yielding rate-exact adaptation [41]. This work studies whether (rate-exact) uniform adaptation is attainable for popular Bayesian learning procedures in terms of local (supremum-norm) concentration rates. We are not aware of any other Bayesian investigation of this type in the literature. Our contributions can be grouped into four types of results. First, we show that spike-and-slab priors achieve uniform exact-rate optimal adaptation in a supremum-norm sense under the white noise model. We relax the prior assumptions of [41], allowing for considerably less sparse priors. Next, building on [18] we show that Bayesian CART is *also* uniformly locally adaptive but sacrifices a logarithmic factor. These results are obtained in the white-noise model as well as non-parametric regression with suitably regular (not necessarily equi-distant) fixed design points. Second, we show how to construct locally adaptive credible bands (with asymptotic coverage 1) whose width depends on local smoothness. This construction builds on [18, 59] who proposed non-locally adaptive bands for (empirical) wavelet coefficients with a possibly exact asymptotic coverage after a multiscale intersection. Third, we provide negative results for Gaussian process priors showing that they are incapable of local adaptation. Fourth, in the context of non-parametric regression, we propose a new class of 'repulsive' partitioning priors which penalize irregular partitions and which are locally rate-exact. These priors can

be viewed as a simplified (zero-degree) version of data-adaptive knot splines. Our results thus provide a stepping stone towards studying Bayesian variable-knot spline techniques. We investigate the numerical performance of our adaptive confidence bands in an extensive simulation study which demonstrates how misleading widely used non-locally adaptive procedures (local polynomials or regressograms [67]) may be. The numerical results for our adaptive confidence bands, on the other hand, agree well with our theory and provide a remedy. We illustrate our framework also on a real dataset of service calls at a bank call center [14].

The manuscript is organized as follows. Section 2 describes the estimation setup and reviews some facts about spatially inhomogeneous functions. Section 3 shows results for spike-and-slab and Bayesian CART priors in the white noise model. Section 4 then shows our results for non-parametric regression. Section 6 wraps up with a discussion. The proofs are in the Supplemental Materials.

2 Statistical Setting

For our theoretical development, we will consider *both* the non-parametric regression model as well as its idealized white noise counterpart. The regression model assumes n noisy samples $Y = (Y_1, \ldots, Y_n)'$ of a function $f_0 : [0,1] \to \mathbb{R}$, where

$$Y_i = f_0(x_i) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \tag{2.1}$$

with $x_i \in [0,1]$ and where $\sigma^2 > 0$ is a known scalar. The white noise model is an elegant continuous version of (2.1) defined via a stochastic differential equation

$$dY(t) = f_0(t)dt + \frac{1}{\sqrt{n}} dW(t), \quad t \in [0,1], \quad n \in \mathbb{N}, \tag{2.2}$$

where $Y(t)$ is an observation process, $W(t)$ is the standard Wiener process on $[0,1]$ and $f_0 \in L^2[0,1]$ is an unknown bounded function on $[0,1]$ to be estimated. The model (2.2) is observationally equivalent to a Gaussian sequence model after projecting the observation

process onto a wavelet basis $\{\psi_{lk} : l \geq 0, 0 \leq k \leq 2^l - 1\}$ of $L^2[0,1]$. This sequence model writes as

$$Y_{lk} = \beta^0_{lk} + \frac{\varepsilon_{lk}}{\sqrt{n}}, \quad \varepsilon_{lk} \overset{iid}{\sim} \mathcal{N}(0,1), \quad l \geq 0, \quad k = 0, \ldots, 2^l - 1, \tag{2.3}$$

where the wavelet coefficients $\beta^0_{lk} = \langle f_0, \psi_{lk} \rangle = \int_0^1 f_0(t)\psi_{lk}(t)dt$ of f_0 are indexed by a scale parameter l and a location parameter k.

Throughout this paper, we will be using the standard Haar wavelet basis $\psi_{-10}(x) = \mathbb{I}_{[0,1]}(x)$ and $\psi_{lk}(x) = 2^{l/2}\psi(2^l x - k)$ obtained with orthonormal dilation-translations of $\psi = \mathbb{I}_{(0,1/2]} - \mathbb{I}_{(1/2,1]}$. We denote with I_{lk} the *dyadic* intervals which correspond to the domain of the balanced Haar wavelets ψ_{lk}, i.e. $\mathrm{I}_{00} = (0,1], \quad \mathrm{I}_{lk} = (k2^{-l}, (k+1)2^{-l}]$ for $l \geq 0$ and $0 \leq k < 2^l$. As elaborated on in Remark 1, the concepts and constructions developed in this work apply to S-regular wavelet basis ψ_{lk} on $[0,1]$, e.g. the boundary-corrected wavelet basis of [23], suitable for estimating functions smoother than Lipschitz. We will illustrate the main ideas using the Haar basis for simplicity.

Brown and Low [12] showed asymptotic equivalence between (2.1) (with an equispaced fixed design) and (2.2) under a uniform smoothness assumption which is satisfied by, e.g., α-Hölderian smooth functions with $\alpha > 0.5$. From a sequence of optimal procedures in one problem, they also prescribed a construction of an asymptotically equivalent sequence in the other. This recipe is particularly convenient for linear estimators. For Bayesian methods, however, it is generally *not known* whether the knowledge of a (wavelet shrinkage/non-linear) minimax procedure in one problem automatically implies the optimality in the other. This is why we study Bayesian procedures in *both* models (see Section 3 for white noise and Section 4 for regression). We obtain local rate-optimality in regression under the assumption $\alpha > 1/2$ in Section 4.2 and, finally, in Section 4.3 we show that a new class of adaptive-split priors (related to variable-knot spline techniques) yields exact-rate optimality *without assuming* $\alpha > 1/2$.

In both models (2.1) and (2.2), the goal is to estimate a possibly *spatially inhomogeneous* function f_0 (see Section 2.1 below). We assess the quality of an estimator using both the

global L^r loss as well as the locally re-weighted supremum-norm loss (similarly as in [37]). In particular, for the (near) minimax rate $r_n(x)$ of adaptive estimation of f_0 at the point x, we will show that, with probability P_{f_0} tending to one, the random variable

$$\sup_{x\in[0,1]} \frac{1}{r_n(x)}|f(x) - f_0(x)|, \quad \text{with} \quad f \sim \Pi(f \mid Y) \tag{2.4}$$

denoting the posterior distribution, is stochastically bounded, thereby implying a uniform local adaptation. Such spatial adaptation is not automatic for many standard estimators. We illustrate this phenomenon on an example below. See Example G.1 in the Supplement for a smoother Doppler curve example adopted from [30].

Example 1. *(Brownian Motion) Our running example throughout the manuscript assumes that f_0 has been generated from a Brownian motion on $[0, 1/2)$ (whose almost all trajectories are locally α-Hölder continuous with $\alpha < 1/2$) and a constant function on $[1/2, 1]$. The plots of the kernel regression, smooth wavelets [14], local polynomials with global smoothing [19] and Bayesian CART estimates assuming $n = 2^{10}$ are in Figure 1. Bayesian CART wastes no splits on the flat domain (compared to regular partitioning methods [19]), showcasing its spatial adaptivity. We will investigate this example theoretically in Section 4.1 where we show that hierarchical Gaussian processes adapt to the worse regularity (determined by the Brownian motion). Beyond adaptability, Bayesian methods can also quantify uncertainty via the posterior (as seen from a companion plot in Figure 2 in Section 3.1.2). The width of the optimal band should be wider when the function is less smooth. In Section 3.1.2, we propose one such construction and show its frequentist validity.*

2.1 Spatially Inhomogeneous Functions

Below, we review several known facts about function classes with inhomogeneous smoothness. The Besov class $B^{\alpha}_{p,q}$ (which contains Hölder and Sobolev classes by setting $p = q = \infty$ and $p = q = 2$, respectively) permits spatial inhomogeneity when $p < 2$. For example, the Bump algebra (consisting of infinite mixtures of Gaussian bumps) coincides with $B^1_{1,1}$ [32]

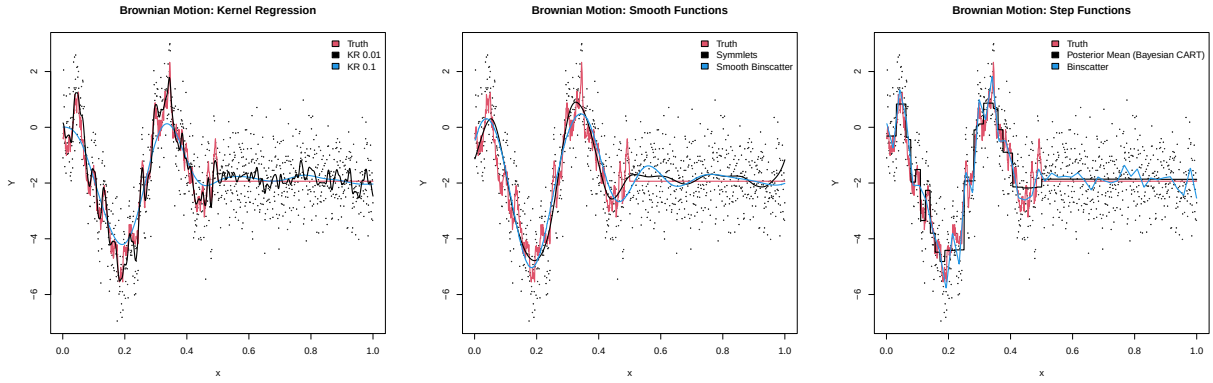

Figure 1: The Brownian motion Example 1. The left panel displays kernel regression estimates (`ksmooth` in R) with a bandwidth 0.1 (capturing well the flat part) and 0.01 (capturing well the wigglier part). The middle panel displays point estimates obtained with (a) symmlet 8 basis ([14] with $\alpha = 0.05$) and (b) the smooth binscatter [19] with $s = p = 2$. The right panel displays point estimators: (a) posterior mean for Bayesian CART with adaptive partitioning and (b) binscatter [19] histogram with $s = p = 0$ and non-adaptive (regular) partitioning.

and constitutes an interesting caricature of smoothness inhomogeneity which would *not* be allowed within the Hölder class. Another example is the total variation (TV) class (contained inside $B^1_{1,\infty}$ and containing $B^1_{1,1}$) which includes functions that may have jumps localized in one part of the domain and be very flat elsewhere [32]. For a discussion on global minimax rates in Besov spaces we refer to [26, 33].

Function spaces where the smoothness can vary from point to point have quite a rich history. Besov spaces with variable smoothness were defined by [47] and later developed by many others (see [65] and references therein). We focus on Hölderian functions which are more intuitive for a supremum-norm analysis. Indeed, the perhaps more widely accepted notion of pointwise regularity has been formalized for Hölderian functions where the exponent itself is a function[1] taking its values in $[0, \infty)$ [1]. For example, 'typical' functions in the Besov space exhibit a multifractal behavior where the Hölder exponent is a continuous function [37].

[1]Andersson [1] showed that a non-negative function is an exponent of a pointwise Hölder function if and only if it can be written as a limit inferior of a sequence of continuous functions.

Following [37], we define a set of bounded functions that are locally t-Hölder at $x \in [0,1]$

$$\mathcal{C}(t,x,M,\eta) = \left\{ f : [0,1] \to \mathbb{R}; \max\left(\|f\|_\infty, \sup_{0<|m|\leq\eta} \frac{|f(x+m)-f(x)|}{|m|^t} \right) \leq M \right\}, \quad (2.5)$$

for $t \leq 1$. We denote with $\mathcal{C}(t,M,\eta)$ the set of functions that are locally Hölder for each $x \in [0,1]$, i.e. $\mathcal{C}(t,M,\eta) = \{f : f(x) \in \mathcal{C}(t,x,M,\eta)\ \forall x \in [0,1]\}$. Throughout this work, we will make the following assumption on f_0.

Assumption 1. *Assume $f_0 \in \mathcal{C}(t,M,\eta)$ where $M(\cdot)$ and $\eta(\cdot)$ are bounded and uniformly bounded away from zero and where the smoothness function $t(\cdot)$ satisfies $0 < t_1 \leq \inf_{x\in[0,1]} t(x)$.*

It is well known that regularity, both local and global, of a function is reflected in the speed at which its wavelet coefficients decay. The following lemma formally characterizes the magnitude of multiscale coefficients in terms of the *local* Hölder smoothness.

Lemma 1. *Denote with $\beta^0_{lk} = \langle f_0, \psi_{lk}\rangle$ the multiscale coefficient of a function f_0 that satisfies Assumption 1. Let $x \in [0,1]$ and for all $l > 0$ define $k_l(x) \in \{0,1,\ldots,2^l-1\}$ such that $x \in \mathrm{I}_{l\,k_l(x)}$. For $l > 0$ and $k \in \{0,1,\ldots,2^l-1\}$ let $\eta_{lk} = \min_{x\in \mathrm{I}_{lk}} \eta(x)$ and $M_{lk} = \max_{x\in \mathrm{I}_{lk}} M(x)$. When $l \geq \log_2[1/(2\eta_{l\,k_l(x)})]$, we have $|\beta^0_{l\,k_l(x)}| \leq 2\, M_{l\,k_l(x)} 2^{-l[t(x)+1/2]}$.*

Proof. For $k = k_l(x)$, we have $|\beta^0_{lk}| = 2^{l/2}\left|\int_{k/2^l}^{(k+1/2)/2^l} [f_0(y) - f_0(y+2^{-l-1})]dy\right|$. Then

$$|\beta^0_{lk}| \leq 2^{l/2} \int_{\frac{k}{2^l}}^{\frac{(k+1/2)}{2^l}} \left[|f_0(y)-f_0(x)| + |f_0(x)-f_0(y+2^{-l-1})|\right] dy \leq 2\, M_{lk} 2^{-l[t(x)+1/2]}. \square$$

Remark 1. *While in this paper we study the case $t(\cdot) \leq 1$, our results (Theorems 1, 2, 3 and 5) can be generalized to higher-order Hölder functions for which Lemma 1 applies with S-regular wavelet basis. More precisely for the results of Sections 3 and 4.2 to be valid what we mainly need is a property in the form $|\beta^0_{l\,k_l(x)}| \leq 2\, M_{l\,k_l(x)} 2^{-l[t(x)+1/2]}$, with $t(x) > 0$ and possibly larger than 1. A natural generalisation of definition (2.5) to $t(\cdot) > 0$ is: for all x there exists $\eta(x) > 0$ such that for all $|y-x| \leq \eta(x)$ we have $\left|f(y) - \sum_{\ell=0}^{r(x)} f^{(\ell)}(x)\frac{(y-x)^\ell}{\ell!}\right| \leq M|y-x|^{t(x)}$, $r(x) = \lceil t(x)\rceil - 1$. Then if $(\psi_{lk} : l \geq 1, k = 0,1,\ldots,2^l-1)$ are C^r wavelets*

(with bounded support and $r > \max_x t(x)$, $|f_{lk}| \leq 2^{-l(t_{lk}+1/2)} M \|\psi\|_{C^r}$, since $\int u^j \psi(u) du = 0$ for $j \leq r$. Under Lipschitz assumption on $x \to t(x)$, Lemma 1 remains valid when $t(\cdot)$ take values larger than 1. There exists another definition in the literature of Hölder (or more generally Besov) function with spatially varying smoothness, such as [65] or [47]. It is not clear how these definitions relate to the above, although [65]'s approach seems related.

3 Spatial Adaptation in White Noise

Donoho et al. [30] characterized pointwise (as well as global) properties of selective wavelet reconstructions showing their near-optimality for estimating Hölderian functions at a given point $x \in [0,1]$. Here, we establish *uniform* (supremum-norm) local adaptation for *all* $x \in [0,1]$ focusing on (2.4) under the white noise model (2.3) and various priors $\Pi(f)$. Adaptive supremum-norm concentration rate results (in white noise and regression) are still few and far between with pathbreaking progress made by multiple authors including [41, 75]. To date, results exist *only* for homogeneous Hölderian functions under the spike-and-slab prior [41, 75] and, more recently, the Bayesian CART prior [18]. Both of these priors leverage certain sparsity structure on the wavelet coefficients $\{\beta_{lk}\}$. We will show that *both* of these priors achieve uniform spatial adaptation.

3.1 Bayesian CART

CART methods [20, 27, 28] and other successful software developments including MARS [35] capture local aspects of the function being estimated by recursively subdividing the predictor space. Donoho et al. [30] pointed out that *'the spatial adaptivity camp is, to date, a-theoretical and largely motivated by heuristic plausibility of the methods'*. While it has been more than 20 years since this seminal paper, there is a shortage of theoretical justifications focusing on spatial adaptation with practically used machine learning methods. Here, we resurrect this question by focusing on Bayesian CART.

Bayesian CART corresponds to a wavelet prior that prescribes a particular sparsity

structure in the wavelet reconstruction according to a *binary tree* $\mathcal{T}$ (see [18] for a more thorough exposition). A tree $\mathcal{T}$ is defined as a collection of hierarchically organized nodes (l,k) where $(l,k) \in \mathcal{T} \Rightarrow (j, \lfloor k/2^{l-j} \rfloor) \in \mathcal{T}$ for $j = 0, \ldots, l-1$. It will be useful to distinguish between two types of nodes: *internal* ones $\mathcal{T}_{int} = \{(l,k) \in \mathcal{T} : \{(l+1, 2k), (l+1, 2k+1)\} \in \mathcal{T}\} \cup (-1, 0)$ and *external* ones $\mathcal{T}_{ext} = \mathcal{T} \backslash \mathcal{T}_{int}$ which are at the bottom of the tree. We then denote with $\boldsymbol{\beta}_{\mathcal{T}} = (\beta_{lk} : (l,k) \in \mathcal{T}_{int})$ the 'active' wavelet coefficients. Similarly as with the selective wavelet reconstruction (*RiskShrink* of [31]), Bayesian CART weeds out wavelet coefficients that are outside the tree, i.e. $\beta_{lk} = 0$ when $(l,k) \notin \mathcal{T}_{int}$. Namely, for $L_{max} \equiv \lfloor \log_2 n \rfloor$ we assume the tree-shaped wavelet shrinkage prior [18]

$$\mathcal{T} \quad \sim \quad \Pi_{\mathcal{T}} \tag{3.1}$$

$$\{\beta_{lk}\}_{l \leq L_{max}, k} \,|\, \mathcal{T} \;\sim\; \pi(\boldsymbol{\beta}_{\mathcal{T}}) \otimes \bigotimes_{(l,k) \notin \mathcal{T}_{int}} \delta_0(\beta_{lk}), \tag{3.2}$$

where $\pi(\boldsymbol{\beta}_{\mathcal{T}}) = \prod_{(l,k) \in \mathcal{T}_{int}} \phi(\beta_{lk}; 0, 1)$ is an independent product of standard Gaussians[2] and where $\Pi_{\mathcal{T}}$ is the Bayesian CART prior [20]. This prior is essentially a heterogeneous Galton-Watson process with a node split probability $p_{lk} = P[(l,k) \in \mathcal{T}_{int}] = p_l = (1/\Gamma)^l$ for some $\Gamma > 2$ (see Section 2.1 in [18] and [61]).

3.1.1 Uniformly Adaptive Rate

The following theorem establishes uniform spatial adaptation of Bayesian CART in the supremum-norm sense. In other words, the posterior is shown to contract at a locally minimax rate, up to a log factor, uniformly for all $x \in [0,1]$. While very intuitive, such a result has not yet been formalized in the Bayesian literature.

Theorem 1. *Assume the Bayesian CART prior* (3.1) *and* (3.2) *with a split probability* p_{lk} *for some sufficiently large* $\Gamma > 0$. *Under the model* (2.3) *and with* t, M *and* η *satisfying Assumption 1, we have*

$$\sup_{f_0 \in \mathcal{C}(t,M,\eta)} E_{f_0} \Pi \left[f : \sup_{x \in [0,1]} \zeta_n(x) |f(x) - f_0(x)| > M_n \,\middle|\, Y \right] \to 0 \tag{3.3}$$

[2] [18] also consider correlated wavelet coefficients.

for $\zeta_n(x) = \left(\frac{n}{\log n}\right)^{\frac{t(x)}{2t(x)+1}}$ *and for any* $M_n \to \infty$ *that is faster than* $\sqrt{\log n}$.

The proof is provided in Section B.1 (Supplement). The first step in the proof of Theorem 1 is showing that trees, a-posteriori, grow deeper in domains where f_0 is less smooth. This property is summarized in Lemma B.1 in the Supplemental Materials. Supremum-norm convergence rate results are valuable for constructing confidence bands. For example, Theorem 1 implies the non-parametric Bernstein-von Mises phenomenon in the multiscale space which can be used to construct credible bands with *exact* asymptotic coverage (see, e.g., Theorem 4.1 in [18]). This set, however, is not guaranteed to have the optimal size (i.e. its diameter shrinking at the minimax rate). Here, we will focus on constructing valid *adaptive* confidence bands. With spatially varying functions (such as the local Hölder functions from Section 2.1), one would expect the width of the confidence band to vary with the smoothness $t(\cdot)$ and be wider where t is smaller. Keeping the diameter constant throughout may yield bands that are more conservative in certain areas of the sample space.

3.1.2 Locally Adaptive Bands

A reasonable requirement for band construction is that their diameter shrinks at the minimax rate of estimation, up to possibly a slow multiplication factor. When the degree of smoothness is known, multiscale[3] credible balls can be constructed (see (5) in [17]) and intersected with qualitative restrictions on f_0 to obtain 'optimal' frequentist confidence sets (which shrink at the optimal rate). We construct optimal confidence sets when the smoothness $t(\cdot)$ is *unknown* and *varying* over $[0, 1]$. Confidence bands that are simultaneously adaptive and honest, of course, do not exist in full generality [49]. Gine and Nickl [39] point out, however, that such confidence sets exist for certain generic subsets of Hölderian functions, the so-called self-similar functions [13, 38, 54, 56, 59], whose complement was shown to be negligible [13]. Under self-similarity, [18, 59] constructed adaptive credible bands for homogeneous Hölderian functions under the spike-and-slab prior and the

[3] They resemble the L^∞ balls [59].

Bayesian CART, respectively.

Here, we extend the notion of self-similarity to *inhomogeneous* Hölder classes for which it is possible to construct a *locally adaptive* confidence set $\mathcal{C}_n$ in the sense that

$$\sup_{f,g\in\mathcal{C}_n}\left[\sup_{x\in[0,1]}\frac{\zeta_n(x)}{v_n}|f(x)-g(x)|\right]=\mathcal{O}_{P_{f_0}}(1) \tag{3.4}$$

for some suitable sequence $v_n\to\infty$ and where $\zeta_n(x)=(n/\log n)^{t(x)/(2t(x)+1)}$. Note that the diameter of $\mathcal{C}_n$ depends on x and equals the minimax rate of estimation (inflated by v_n) *at every point* $x\in[0,1]$. Below, we formally introduce the notion of locally self-similar functions.

Definition 1. *(Local Self-Similarity) We say that* $f\in\mathcal{C}(t,M,\eta)$ *is* locally self-similar *at* $x\in[0,1]$ *if, for some* $c_1>0$ *and an integer* j_0*, we have* $|K_j(f)(x)-f(x)|\geq 2^{-j\,t(x)}c_1$ *for all* $j\geq j_0$*, where* $K_j(f)=\sum_{l\leq j-1}\sum_k\langle\psi_{lk},f\rangle\,\psi_{lk}$ *is the wavelet projection at level* j*. The class of all self-similar functions at* x *will be denoted by* $\mathcal{C}_{SS}(t(x),x,M(x),\eta(x))$*. Moreover, we denote with* $\mathcal{C}_{SS}(t,M,\eta)$ *a set of functions that are self-similar for all* $x\in[0,1]$*.*

For spatially heterogeneous Hölderian functions, we construct locally adaptive confidence bands whose width is varying and reflects smoothness at each given x While related to previous constructions (see e.g. [18] for the homogeneous case), its simplicity and ease of computability make our band particularly appealing in practice (see Figure 2). In addition, we are not aware of any other related frequentist band for the case of heterogeneous smoothness. We center our confidence bands around a pivot estimator, the median tree estimator [18].

Definition 2. *(The Median Tree) Given a posterior distribution* $\Pi_{\mathcal{T}}[\cdot\,|\,Y]$ *over tree-shaped coefficient subsets, we define the* median tree $\mathcal{T}_Y^*$ *as the following set of nodes*

$$\mathcal{T}_Y^*=\{(l,k),\ l\leq L_{max},\ \ \Pi[(l,k)\in\mathcal{T}_{int}\,|\,Y]\geq 1/2\}\,. \tag{3.5}$$

We define the resulting median tree estimator as $\widehat{f}_T(x)=\sum_{(l,k)\in\mathcal{T}_Y^*}Y_{lk}\psi_{lk}(x)$ which is shown to attain the near-minimax rate of estimation at each point (see the proof of Theorem

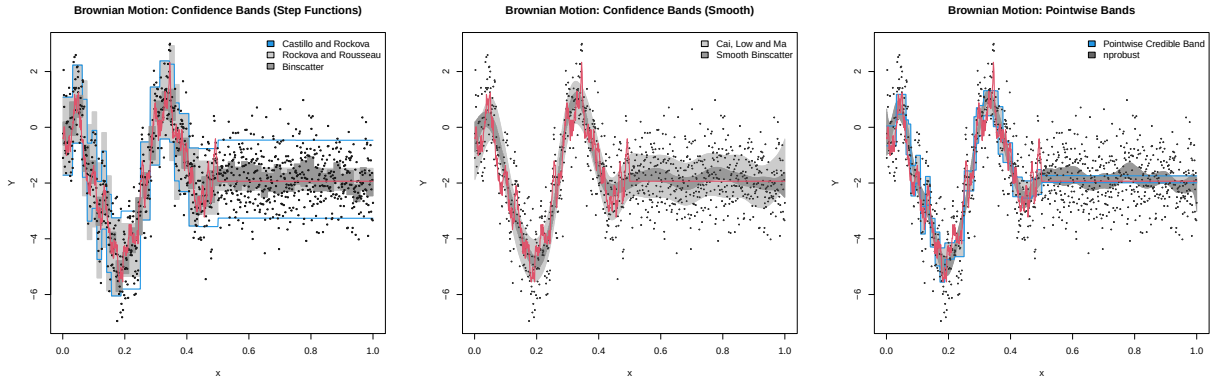

Figure 2: The Brownian motion example (Example 1 in red color) with dyadic Bayesian CART. The left panel displays confidence bands (a) the non-adaptive band of [18], (b) our adaptive band and the binscatter [19] bands with $s = p = 0$. For (a) and (b) we choose v_n so that the band contains 95% posterior probability. The middle panel displays smooth bands obtained with (a) symmlet 8 basis ([14] with $\alpha = 0.05$) and (b) the smooth binscatter [19] with $s = p = 2$. The right panel displays point-wise bands: (a) pasted 95% posterior credible intervals and the bands in [15] with $\alpha = 0.05$ implemented in software `nprobust`.

2). Next, we define the *local radius* (which varies with x) as

$$\sigma_n(x) = v_n \sqrt{\frac{\log n}{n}} \sum_{l=0}^{L_{max}} \mathbb{I}\{(l, k_l(x)) \in \mathcal{T}_Y^*\} |\psi_{lk_l(x)}(x)| \tag{3.6}$$

for some $v_n \to \infty$ to be chosen. Finally, we construct the confidence band according to the following prescription

$$\mathcal{C}_n = \left\{ f : \sup_{x \in [0,1]} \left[\frac{1}{\sigma_n(x)} |f(x) - \widehat{f}_T(x)| \right] \leq 1 \right\}. \tag{3.7}$$

Theorem 2. *Let Π be the prior as in the statement of Theorem 1. Then for $\mathcal{C}_n$ defined in (3.7) with $v_n = \mathcal{O}(\log n)$, uniformly over t, M and η that satisfy the Assumption 1*

$$\inf_{f_0 \in \mathcal{C}_{SS}(t,M,\eta)} P_{f_0}(f_0 \in \mathcal{C}_n) \to 1 \quad \text{as } n \to \infty.$$

Uniformly over $f_0 \in \mathcal{C}_{SS}(t, M, \eta)$, the diameter verifies (3.4), as $n \to \infty$, and the credibility of the band satisfies $\Pi[\mathcal{C}_n \,|\, Y] = 1 + o_{P_{f_0}}(1)$.

The proof is provided in Section B.2 (Supplement).

According to Theorem 2, the band (3.7) has asymptotic coverage 1. It is possible to intersect (3.7) with a multi-scale ball (as in [18, 59]) to obtain asymptotic coverage $1-\gamma$ for some small $\gamma > 0$ as a consequence of the non-parametric Bernstein-von Mises (BvM) theorem. The multiscale band (see Corollary 1 and 2 in the Supplement to [18]) compares suitably normalized sequences of (empirical) wavelet coefficients and its shape resembles an L_∞ band. We could implemented the intersection to "stabilize" (3.7) , i.e. to avoid overly wide bands due to large choices of v_n. However, the multiscale band, defined in (77) in the Supplement of [18], requires an ("admissible") monotone increasing weighting sequence w_l in the multiscale norm which has to be determined by the user. Instead of the multiscale intersection [18], here we choose v_n adaptively so that (3.7) captures $(1-\gamma)\%$ posterior (draws). This adaptive choice of v_n leverages posterior information and can be implemented using a grid search. In order to illustrate the practical virtue of Theorem 2, we revisit the Brownian motion example (Example 1) from Figure 2. We implement a dyadic version of the Bayesian CART algorithm [20] which splits only at dyadic rationals. We plot the adaptive band (3.7) together with a non-adaptive band obtained by taking the maximal diameter $\sigma(x)$ over the domain $[0,1]$ (as in Theorem 4 of [18]). For both cases we choose v_n adaptively such that the resulting band contains 95% posterior probability. Comparing our construction with [18] (Figure 2 on the left), we can see benefits of our locally adaptive construction, where the width is larger in the first half of the domain where the function meanders according to the Brownian motion (expected since the smoothness is smaller than 1/2). Interestingly, a regular partitioning method `binscatter` [19] does not achieve satisfactory coverage which is in line with theory in [18] showing the inability of regular histograms to achieve ℓ_∞ adaptation. In addition, smoother techniques based on symmmlets [14] or local-polynomials with global smoothing [19] also show poor coverage (Figure 2 in the middle). Finally, Figure 2 on the right shows that, expectedly, point-wise bands (95%-credible bands and [15]) do not yield satisfactory coverage. We revisit this example in a simulation study in Section 5.1.

3.2 Spike-and-Slab Priors

Spike-and-slab priors are arguably one of the most ubiquitous priors in statistics (see references [22, 41, 59] for wavelet shrinkage contexts). Compared with the Bayesian CART prior from Section 3.1, spike-and-slab priors allocate positive prior mass to *any* subset $\mathcal{T}$ of $\{(l,k) : l \leq L_{max}\}$, not just tree-shaped subsets. We define the spike-and-slab prior through the following hierarchical model.

Assumption 2. *(Spike-and-Slab Prior)*

- *Prior on* $\mathcal{T} \subseteq \{(l,k) : 0 \leq k < 2^l, l \leq L_{max}\}$*: There exist constants* $c_T, C_T > 0$ *such that*
$$c_T\, \omega_l \leq \frac{\Pi(\mathcal{T} \cup \{(l,k)\})}{\Pi(\mathcal{T})} \leq C_T\, \omega_l \quad \forall \mathcal{T} \text{ such that } (l,k) \notin \mathcal{T} \tag{3.8}$$
for some positive sequence ω_l *such that, for some* $B_\omega > 0$ *and* $\delta > 0$*,*
$$n^{-B_\omega} \leq \omega_l \leq n^{(1-\delta)/2} 2^{-l} \quad \text{for} \quad l \leq L_{max}. \tag{3.9}$$
- *There exist probability densities* $\pi_{lk}(\cdot)$ *on* $\mathbb{R}$ *such that, conditionally on* $\mathcal{T}$*,*
$$\beta_{lk} \overset{ind}{\sim} \pi_{lk} \quad \forall (l,k) \in \mathcal{T} \quad \text{and} \quad \beta_{lk} = 0, \quad \forall (l,k) \notin \mathcal{T}$$
and there exist $R, c_R, C_R > 0$ *such that*
$$c_R \leq \inf_{|\beta| \leq R} \pi_{lk}(\beta) \leq \sup_{\beta \in \mathbb{R}} \pi_{lk}(\beta) \leq C_R. \tag{3.10}$$

While seemingly similar to the prior considered in [41], our Assumption 2 is much weaker. Indeed, our prior construction subsumes the spike-and-slab prior of [41] by imposing weaker constraints on the decay of inclusion probabilities. Note that ω_l's in (3.9) are allowed to be much larger than in [41] which assume $n^{-B} \leq \omega_l \leq 2^{-j(1+\tau)}$ for some $\tau > 1/2$. This perhaps subtle difference is of great practical importance and indicates that optimal sup-norm adaptation occurs in far less sparse situations than originally perceived. Another important difference is that we *do not require* the binary indicators $\mathbb{I}(\beta_{lk} \neq 0)$ for $0 \leq k < 2^l$ and $l \leq L_{max}$ to be *iid* Bernoulli random variables. This extension allows us to consider, for example, Ising prior constructions [7] which allow the inclusion indicators to be related through a Markovian model.

Theorem 3. *Consider the model (2.3) with a prior* Π *on* $\{\beta_{lk}\}_{lk}$ *following the Assumption 2. Let* t, M *and* η *satisfy the Assumption 1. Then*

$$\sup_{f_0 \in \mathcal{C}(t,M,\eta)} E_{f_0} \Pi \left[f : \sup_{x \in [0,1]} \zeta_n(x) |f(x) - f_0(x)| > \widetilde{M} \,\middle|\, Y \right] \to 0,$$

for all sufficiently large $\widetilde{M} > 0$ *where* $\zeta_n(x) = (n/\log n)^{t(x)/(2t(x)+1)}$.

The proof is provided in Section B.3 (Supplement). Theorem 3 shows that, unlike Bayesian CART, the spike-and-slab priors achieve the *exact* rate uniformly over the entire domain $[0, 1]$ without any additional logarithmic penalty ([18] showed that the log-factor in Bayesian CART is non-negotiable). In Lemma B.5 (an analog of Lemma 1 in [41]) we show that the posterior concentrates on a subset of large enough coefficients. This fact can be used to show that the median probability model (MPM) [4, 5, 59] consisting of all coefficients with at least 50%-posterior probability of being active is an (exact) rate-optimal estimator. Following the strategy of Proposition 4.5 in [59] one can then show that Theorem 2 remains true for the spike-and-slab prior when replacing the median tree estimator with MPM and with v_n that can grow slower at a rate at least $\sqrt{\log n}$ [59]. We discuss benefits and drawbacks relative to the Bayesian CART prior in our simulation study in Section 5.1.

4 Spatial Adaptation in Non-parametric Regression

Throughout this section, we assume the canonical non-parametric regression setup (2.1) with $\sigma^2 = 1$. While nonparametric regression with a *regular* design and the white noise model are asymptotically equivalent (e.g. under the usual smoothness assumption $t(x) > 1/2$ [12]), optimality of a procedure in one setup *does not* automatically imply optimality in the other. In Section 4.2, we show rate-optimality of Bayesian CART in non-parametric regression when $t(x) > 1/2$ *without* assuming regular designs. Later in Section 4.3, we relax the restriction $t(x) > 1/2$ and propose new 'repulsive' partitioning priors (related to adaptive-knot splines) and show that they are exact-rate adaptive. First, we describe some not so optimistic findings for hierarchical Gaussian priors.

4.1 Gaussian Processes

In Figure 1 we have seen that spatial adaptation is *not attainable* by methods which are not sufficiently localized. In this section, we formally show that several practically used hierarchical Gaussian process priors *do not* lead to spatially adaptive concentration rates. While this phenomenon is not entirely surprising, it is nevertheless worthwhile to document it formally. In particular, we provide lower bound results showing sub-optimality of Gaussian processes in terms of a *global* estimation loss. To this end, we consider the following heterogeneous-smoothness assumption which aligns with Figure 1 where the function f_0 has smoothness $\alpha_1 < 1/2$ on $[0, 1/2]$ and $\alpha_2 = 1$ on $(1/2, 1]$.

Assumption 3. *Assume that the Haar wavelet decomposition of a function* f_0

$$f_0(x) = \psi_{-10}(x)\beta^0_{-10} + \sum_{l=0}^{\infty}\sum_{k\in I_l} \beta^0_{lk}\psi_{lk}(x) \tag{4.1}$$

with $I_l = \{0, 1, \dots, 2^l - 1\}$ *satisfies that for all* l*, there exists* $0 \le N_l \le 2^l - 1$ *with* $N_l 2^{-l} \asymp 1$ *and such that for some* $M_1 > 0$*,*

$$\max_{k\in I_{l1}} |\beta^0_{lk}| \le M_1 2^{-l(\alpha_1+1/2)} \quad \textit{and} \quad \max_{k\in I_{l2}} |\beta^0_{lk}| \le M_1 2^{-l(\alpha_2+1/2)} \quad \textit{with} \quad \alpha_1 < \alpha_2, \tag{4.2}$$

where $I_{l1} = \{0, 1, \dots, N_l\}$ *and* $I_{l2} = \{N_l + 1, \dots, 2^l - 1\}$.

Methods that are globally, but not locally, adaptive are expected to adapt to the worse-case scenario and attain the slower rate determined by the smaller smoothness α_1. We will formalize this intuition by assessing the quality of the reconstruction with an L^2 loss over the *entire* domain as well as the *smoother* domain determined by α_2, i.e. we define

$$\|f - f_0\|_2^2 \equiv \int_0^1 |f(x) - f_0(x)|^2 dx \quad \text{and} \quad \|f - f_0\|_{1/2,1}^2 \equiv \int_{1/2}^1 |f(x) - f_0(x)|^2 dx.$$

We now consider three hierarchical Gaussian processes on

$$f(x) = \psi_{-10}(x)\beta_{-10} + \sum_{l=0}^{\infty}\sum_{k\in I_l} \beta_{lk}\psi_{lk}(x) \tag{4.3}$$

induced through a prior on the sequence $\{\beta_{lk}\}_{lk}$. These priors were studied in [60]. This section assumes a regular design $x_i = i/n$ for $n = 2^{L_{max}+1}$ with some $L_{max} > 0$.

- (T1) (*Sieve Prior*) Let $L-1 \sim \Pi(L)$ where $\Pi(L)$ behaves either like a Poisson or a geometric distribution, truncated to $2^L \leq n$. Then conditionally on L

$$\beta_{lk} \overset{iid}{\sim} \mathcal{N}(0, \tau^2) \quad \text{for} \quad l \leq L \quad \text{and} \quad \beta_{lk} = 0 \quad \text{for} \quad l > L, \quad \tau > 0.$$

- (T2) (*Scale Parameter*) Given $\alpha > 0$ assume

$$\beta_{lk} = \mathbb{I}_{l \leq L_{max}} 2^{-l(\alpha+1/2)} Z_{lk}, \quad Z_{lk} \overset{iid}{\sim} \mathcal{N}(0, \tau^2), \quad \tau \sim \pi_\tau,$$

  where π_τ is n Inverse Gamma distribution, or more generally follows the assumptions of Lemma 3.5 of [60].

- (T3) (*Rate Parameter*) Given $\tau > 0$ assume

$$\beta_{lk} = \mathbb{I}_{l \leq L_{max}} 2^{-l(\alpha+1/2)} Z_{lk}, \quad Z_{lk} \overset{iid}{\sim} \mathcal{N}(0, \tau^2), \quad \alpha \sim \pi_\alpha$$

  where π_α is a Gamma distribution or more generally satisfies the assumptions of Lemma 3.6 of [60].

These priors have been studied in a multitude of works, see e.g. [2, 66] for the setup (T1) [6, 44, 70] for the Gaussian process priors (T2) and (T3). More recently, this framework has been studied in [60] in the case of Fourier-series priors where both lower bounds and an upper bound have been obtained in the case of non-linear regression. We adapt their proof to the wavelet basis case with functions satisfying (4.2). In this case, we note that for any L such that $2^L \leq n$ and for $f_{\beta,L}$ denoting the Haar wavelet expansions (4.3) truncated at $L \geq 1$ we have $\|f_{\beta,L} - f_{\beta_0,L}\|_n^2 = \frac{1}{n}\sum_i [f_{\beta,L}(x_i) - f_{\beta_0,L}(x_i)]^2 = \|\boldsymbol{\beta}_L - \boldsymbol{\beta}_{0,L}\|_2^2 = \|f_{\beta,L} - f_{\beta_0,L}\|_2^2$. where $\boldsymbol{\beta}_L = (\beta_{lk} : l \leq L)'$ (resp. $\boldsymbol{\beta}_{0,L}$) is the truncated version of $\boldsymbol{\beta}$ (resp $\boldsymbol{\beta}_0$).

Theorem 4. *Let f_0 satisfy (4.2) and consider either of the priors (T1)-(T3). Then for $Y = (Y_1, \ldots, Y_n)'$ arising from (2.1) with $x_i = i/n$ we have*

$$\Pi\left[\|f - f_0\|_2 \leq M_n \epsilon_n(\alpha_1) \,|\, Y\right] = 1 + o_{P_{f_0}}(1) \quad \text{as } n \to \infty, \text{ where} \tag{4.4}$$

$$\epsilon_n(\alpha_1) = \begin{cases} (n/\log n)^{-\alpha_1/(2\alpha_1+1)} & \text{under (T1), (T2) if } \alpha_1 < 2\alpha+1, \text{ and (T3)} \\ (n/\log n)^{-(2\alpha+1)/(4\alpha+4)} & \text{under (T2) if } \alpha_1 \geq 2\alpha+1. \end{cases}$$

Moreover, for all functions satisfying (4.2) *and, for some* $c > 0$,

$$\min_{k \in I_{l1}} |\beta_{lk}^0| \geq c\, 2^{-l(\alpha_1 + 1/2)} \tag{4.5}$$

we have $\Pi\left[\|f - f_0\|_{1/2,1} \leq n^{-\delta}\epsilon_n(\alpha_1) \,|\, Y\right] = o_{P_{f_0}}(1) \quad$ *as* $n \to \infty$ *for all* $\delta > 0$.

The proof is provided in Section C (Supplement). The first statement (4.4) shows that the posterior under the Gaussian sequence priors adapts to the *worse* smoothness α_1. Moreover, the second statement implies that, under a suitable identifiability condition, the posterior is *incapable* of achieving a faster rate on the smoother domain (determined by $\alpha_2 > \alpha_1$), rendering adaptation to α_2 impossible. Note that similarly to [60], the same conclusions holds if one deploys an empirical Bayesian procedure based on the marginal maximum likelihood estimator on L for T1 (resp. τ for T2 and α for T3).

4.2 Bayesian CART

This section reports positive findings in the context Bayesian CART. In particular, we show a regression analogue of Theorem 1 assuming $t(x) > 1/2$. [18] also study Bayesian CART in regression but with a *regular* design where the prior is assigned to *empirical* wavelet coefficients. This re-parametrization closely resembles the white noise model, enabling a more direct transfer of the results. Here, we follow an alternative route. A perhaps more transparent approach is to assign a prior *directly* to the actual (not empirical) wavelet coefficients (similarly as in [75]). This strategy aligns more closely with what is done in practice. We pursue this direction here and, in addition, consider designs that are not necessarily regular.

With a vector of observations $Y = (Y_1, \ldots, Y_n)'$ and $F_0 = (f_0(x_1), \ldots, f_0(x_n))'$, we can re-write (2.1) in a matrix notation: let $p = 2^{L_{max}} = \lfloor C^* \sqrt{n/\log n} \rfloor$ for some $C^* > 0$,

$$Y = X\boldsymbol{\beta}_0 + \boldsymbol{\nu}, \quad \text{where} \quad \boldsymbol{\nu} = F_0 - X\boldsymbol{\beta}_0 + \boldsymbol{\epsilon} \quad \text{with} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I_n), \tag{4.6}$$

where $\boldsymbol{\beta}_0 = (\beta_1^0, \ldots, \beta_p^0)' \in \mathbb{R}^p$ is a sparse vector of multiscale coefficients $\langle f_0, \psi_{lk} \rangle$ ordered according to $2^l + k$ and where $X = (x_{ij})_{i \leq n, j \leq p}$ with $x_{ij} = \psi_{lk}(x_i)$ when $j = 2^l + k$.

Because we assume $t(x) > 1/2$, we do not need resolutions larger than L_{max} to be able to approximate f_0 well. We will be denoting with $\mathbb{T}$ all tree-shaped subsets of nodes (l,k) such that $l \leq L_{max}$. For a tree $\mathcal{T} \in \mathbb{T}$ and a vector $\boldsymbol{\beta} \in \mathbb{R}^p$, we denote with $\boldsymbol{\beta}_{\mathcal{T}}$ the subset of coefficients inside the tree and with $\boldsymbol{\beta}_{\setminus\mathcal{T}}$ the complement. Similarly, we split the design matrix X into active covariates $X_{\mathcal{T}}$ (that correspond to $(l,k) \in \mathcal{T}_{int}$) and the complementary inactive ones $X_{\setminus\mathcal{T}}$. It will be advantageous[4] to use the unit-information g-prior for $\boldsymbol{\beta}_{\mathcal{T}}$

$$\boldsymbol{\beta}_{\mathcal{T}} \sim \mathcal{N}(0, g_n(X'_{\mathcal{T}}X_{\mathcal{T}})^{-1}) \quad \text{with } g_n = n \tag{4.7}$$

which yields the following marginal likelihood under each tree $\mathcal{T}$

$$N_Y(\mathcal{T}) = \frac{\exp\left\{-\frac{1}{2}Y'[I - X_{\mathcal{T}}\Sigma_{\mathcal{T}}X'_{\mathcal{T}}]Y\right\}}{(2\pi)^{n/2}(1+g_n)^{|\mathcal{T}|/2}},\ \Sigma_{\mathcal{T}} = c_n(X'_{\mathcal{T}}X_{\mathcal{T}})^{-1},\ c_n = g_n/(g_n+1).$$

Throughout this section, we will denote with n^L_{lk} (resp. n^R_{lk}) the number of observations that fall inside the domain of the left (resp. right) wavelet piece ψ_{lk}, i.e. $n_{lk} = \sum_{i=1}^n \mathbb{I}(x_i \in I_{lk}) = n^R_{lk} + n^L_{lk}$ and we define $\bar{n}_{lk} = \max\{n^R_{lk}, n^L_{lk}\}$ and $\underline{n}_{lk} = \min\{n^R_{lk}, n^L_{lk}\}$. The regular design $x_i = i/n$ satisfies $\bar{n}_{lk} = \underline{n}_{lk} = n/2^{l+1}$, when n is a power of 2. Here, we allow for a design $\mathcal{X} = \{x_i \in [0,1] : 1 \leq i \leq n\}$ that is *not necessarily* regular. Instead, we make the following design balance assumption.

Assumption 4. *(Balanced Design) Let $\widetilde{L}_{max}$ be such that $2^{\widetilde{L}_{max}} = \lfloor C_x n/\log n \rfloor$ for some $C_x > 0$. We say that the design $\mathcal{X}$ is υ-regular for some $\upsilon > 0$ if for any (l,k) s.t. $0 \leq l \leq \widetilde{L}_{max}$*

$$\frac{c\,n}{2^l} \leq \underline{n}_{lk} \leq \bar{n}_{lk} \leq \frac{(C+l)n}{2^l} \quad \textit{for some } c, C > 0 \tag{4.8}$$

and, for some $C_d > 0$, $0 \leq \bar{n}_{lk} - \underline{n}_{lk} \leq C_d \frac{\sqrt{n}\log^{\upsilon} n}{2^{l/2}}$.

Note that the threshold L_{max} used to construct the design matrix X is smaller than the threshold $\widetilde{L}_{max}$ in Assumption 4 and all partitioning cells induced by $\mathcal{T} \in \mathbb{T}$ are

[4] We can take advantage of certain properties of projection matrices. Other priors can be considered as well.

guaranteed to have at least one observation. The Assumption 4 is not overly restrictive. Indeed, we show in Lemma F.5 (Supplemental Materials) that the second condition is satisfied[5] with probability approaching one when $\mathcal{X}$ arises from a uniform distribution on $[0,1]$. Irregular observations ultimately induce *correlated* Haar wavelet designs X where the correlation pattern has a particular hierarchical structure described in Lemma F.1 (Supplement). As with other related consistency results in regression (see e.g. [53]), we cannot allow for too much correlation in the design X. Fortunately, *balanced* designs that satisfy Assumption 4 are not too collinear and yield diagonally dominant covariance matrices with well-behaved eigenvalues (see Lemma F.4 in the Supplement). We are now ready to state a non-parametric regression version of Theorem 1.

Theorem 5. *Assume the regression model (4.6) under Assumption 4 for some $0 \leq \upsilon < 1/2$ and $c, C, C_d > 0$ or $\upsilon = 1/2$ with $c > 2C_dC^*$. Assume the Bayesian CART tree prior $\Pi_{\mathcal{T}}$ with a split probability $p_l = (\Gamma)^{-l^{2[\upsilon+(\upsilon\vee 1)]}}$ for some sufficiently large $\Gamma > 0$ and the g-prior (4.7). Assume that t, M and η satisfy Assumption 1 with $t_1 > 1/2$, then*

$$\sup_{f_0 \in \mathcal{C}(t,M,\eta)} E_{f_0} \Pi \left[f : \sup_{x \in [0,1]} \zeta_n(x) |f(x) - f_0(x)| > M_n \,\middle|\, Y \right] \to 0$$

for $\zeta_n(x) = \left(\frac{n}{\log n}\right)^{\frac{t(x)}{2t(x)+1}}$ and for any $M_n \to \infty$ that is faster than $\log^{\upsilon+\upsilon\vee 1+1/2} n$.

The proof is provided in Section A.1 (Supplement). Note that the prior split probability decays more rapidly to accommodate the irregular design assumption. An analogous statement can be obtained for the spike-and-slab prior using a similar approach as in the proof of Theorem 3. In addition, the confidence set construction in (3.7) remains valid also under the non-parametric regression setting. Indeed, rate-optimality of the posterior implies rate-optimality of the median-tree estimator and the regression variant of Theorem 2 thus holds under the assumption $t_1 > 1/2$.

[5] for $\upsilon \geq 1/2$ due to simultaneous control of $2^{\widetilde{L}_{max}}$ coefficients

4.3 Repulsive Partition Prior

In the previous section, we studied a prior on wavelet coefficients that corresponds to recursive partitioning. In this section, we propose a different partitioning prior on piecewise constant functions which relates to variable knot spline techniques [51, 71]. We assume that a partition $S = (I_1^S, \cdots, I_J^S)$ of $[0,1]$ is not necessarily induced by a tree but arrives from a 'determinantal-type' prior

$$S \sim \pi_S \propto \prod_{j=1}^{J} |I_j^S|^B \mathbb{I}(S \in \mathbb{S}) \quad \text{for some } B > 0, \tag{4.9}$$

where $\mathbb{S}$ contains all partitions made out of blocks with endpoints belonging to a fixed grid $\mathcal{I}_n = (z_\ell : \ell \leq N_n)$ such that $z_0 = 0$ and $z_{N_n} = 1$ and

$$0 \leq z_\ell < z_{\ell+1}, \ \frac{C_1 \log n}{n} \leq z_{\ell+1} - z_\ell \leq \frac{C_2 \log n}{n}, \ \ell \leq N_n \quad \text{for some} \quad C_1, C_2 > 0. \tag{4.10}$$

Note that the size of an interval in S can be measured either in terms of its length or in terms of its number of units, i.e. number of elements in the grid $\mathcal{I}_n$ belonging to it. We refer to (4.9) as a repulsive partitioning prior because it prevents the splits from occurring too close to one another. The prior (4.9) thus rewards partitions that are more regular. The set $\mathcal{I}_n$ contains candidate knots for possible split, e.g. a subset of observed design points. While in variable knot spline techniques (such as MARS [35]) knot points are added, removed and allocated recursively using cross-vaildiation, here we let the posterior distribution choose the knots in a data-adaptive way. Given the partition $S \in \mathbb{S}$ we reconstruct the regression surface with

$$f_\beta^S(x) = \sum_{j=1}^{J} \beta_j \mathbb{I}_{I_j^S}(x), \quad \text{where} \quad (\beta_j : j \leq J) \overset{ind}{\sim} g_j. \tag{4.11}$$

Regarding the prior density g_j, we will assume that there exist $0 < c_0 \leq c_1$ and $B_0 > 0$ such that

$$c_0 \leq g_j(\beta) \quad \forall |\beta| \leq B_0, \quad \text{and} \quad \|g_j\|_\infty \leq c_1 \quad \forall j \leq J. \tag{4.12}$$

While in Section 4.2 we obtained the near-minimax rate under the assumption $t_1 > 1/2$, here we show rate-exactness *without* necessarily assuming $t_1 > 1/2$. We are interested in

bounding $\Pi(A^c_{\varepsilon_n}(\widetilde{M}) \mid D_n)$ where $D_n = \{(Y_i, x_i)\}_{i=1}^n$ and

$$A_{\varepsilon_n}(\widetilde{M}) = \left\{ \sup_{x\in(0,1)} \frac{|f^S_\beta(x) - f_0(x)|}{\varepsilon_n(x)} \leq \widetilde{M} \right\} \quad \text{and} \quad \varepsilon_n(x) = \left(\frac{n}{\log n}\right)^{-\frac{t(x)}{2t(x)+1}}. \tag{4.13}$$

For a given point $x \in [0,1]$ and a partition $S = \{I^S_j\}_{j=1}^J$ we denote with $I^S_x \in S$ the interval containing x and with $R(I^S_x)$ (resp. $L(I^S_x)$) its right (resp. left) neighbor. We then define $\mathcal{I}(x)$ as the set of intervals which contain x and the two neighboring intervals, i.e.

$$\mathcal{I}(x) = \bigcup_{S\in\mathbb{S}} \{I^S_x, R(I^S_x), L(I^S_x)\}, \tag{4.14}$$

and we define, for a given $x \in [0,1]$ and some $u_1 > 0$,

$$\Omega_{n,x}(u_1) = \left\{ |n_I - n \times p_I| \leq u_1 \sqrt{\log n \times n \times p_I} \quad \forall I \in \mathcal{I}(x) \right\} \tag{4.15}$$

where $n_I = \sum_{i=1}^n \mathbb{I}(x_i \in I)$ and p_I is a function of I which satisfies $p_0|I| \leq p_I \leq p_1|I|$ for some $0 < p_0 \leq p_1$. Our results will be conditional on a large probability event which can be loosely regarded as a design assumption. Namely, we require that the cells containing the knot points z_l (and the neighboring cells) are large enough in terms of the number of observations falling inside, i.e. we consider an intersection of events in (4.15) $\Omega_n(u_1) = \bigcap_{l=1}^{N_n} \Omega_{n,z_l}(u_1)$. Our result below will hold on this event. If the design is regular, then $\Omega_n(u_1)$ holds for any $u_1 > 0$ and $p_0 = p_1 = 1$. If the design is random (with a density bounded away from zero) then (4.15) holds with large probability if u_1 is large enough, as shown in Lemma F.6 (Supplemental Materials). Unlike in Section 4.2 where we assumed $\inf_{x\in[0,1]} t(x) > 1/2$, now we assume that $0 < t(\cdot) \leq 1$ and that $t(\cdot)$ is piecewise Hölder.

Assumption 5. *(Piecewise Hölder) Assume that there exists a fixed partition of* $[0,1]$ *into* k *intervals say* $[a_j, a_{j+1})$ *(resp.* $(a_j, a_{j+1}]$*) with* $a_0 = 0$ *and* $a_{k+1} = 1$ *such that* $t(\cdot)$ *is* α_j*-Hölder on* (a_j, a_{j+1})*, i.e. for* $L_0 > 0$ $|t(x) - t(y)| \leq L_0|x - y|^{\alpha_j}$ *for* $x, y \in (a_j, a_{j+1})$*, and such that* $t(\cdot)$ *is right (resp. left) Hölder at* a_j*.*

Theorem 6. *Consider the prior defined by* (4.10)-(4.12) *and* (4.9) *and with* $B > 9$ *and* $C_1 > 4u_1^2/p_0$ *and* $\|f_0\|_\infty < B_0$*. Under the Assumption 1 with* $t(\cdot)$ *piecewise Hölder according to the Assumption 5, there exists* $\widetilde{M} > 0$ *such that* $E_{f_0}\left[\mathbb{I}_{\Omega_n(u_1)}\Pi(A^c_{\varepsilon_n}(\widetilde{M}) \mid D_n)\right] = o(1)$*.*

The proof is provided in Section A.2 (Supplement). Theorem 6 shows that *rate-exactness* can be achieved in regression *uniformly* over $[0,1]$ for local Hölderian functions whose exponents are piece-wise Hölder. The prior construction (4.11), (4.12) and (4.9) can be regarded as a version of variable-knot zeroth order splines. Note that the partition in Assumption 5 needs not be known for our procedure to be valid.

While we have presented our result in the case of univariate densities, extension to the multivariate case are possible but perhaps a bit more tedious. More interestingly, the proving technique in Section A.2 may be extended to free-knot splines, which have typically been devised to adapt spatially but for which no proofs exist. Finally, although the repulsive prior used in Theorem 6 on the partition is not proved to be necessary, we believe that some form repulsion is necessary.

5 Performance Evaluation

We demonstrate the benefits of our locally adaptive confidence bands (relative to widely used methods in practice that are *not* spatially adaptive) in a simulation study as well as on a real data example.

5.1 Simulation Study

We considered 4 test functions exhibiting various degrees of spatial inhomogeneity following [29] (details are shown in Section G of the Supplement). One of the test functions was discussed previously in Example 1. We summarize results from 100 repetitions from the model (2.1) with $\sigma = 1$ and with $x_i = 1/n$ and $n = 2^{10}$. See Section G (Supplement) for implementation details of Metropolis-Hastings samplers for Bayesian CART and Spike-and-Slab priors.

We construct our confidence band according to (3.7) using an adaptive choice of v_n in (3.6) so that [6] the band contains $(1-\alpha)\%$ of posterior draws. We choose $\alpha \in \{0.05, 0\}$ and denote these two bands with $\mathcal{C}_n^1$ (with $\alpha = 0.05$) and $\mathcal{C}_n^2$ (with $\alpha = 0$) in our Tables. Next,

[6] We find such a v_n by grid search over $v_n = \{0.5 + k \times 0.005 : 1 \leq k \leq 100\}$.

	Bayesian CART					Spike-and-Slab					CCF	BIN1	BIN2	CLM
	$\mathcal{C}_n^1$	$\mathcal{C}_n^2$	$\widetilde{\mathcal{C}}_n$	$\mathcal{P}_n$	L_∞	$\mathcal{C}_n^1$	$\mathcal{C}_n^2$	$\widetilde{\mathcal{C}}_n$	$\mathcal{P}_n$	L_∞				
Doppler Curve														
%	7.31	1.71	1.7	42.91	2.16	2.08	0.04	0.9	37.33	1.25	16.24	16.99	6.2	12.45
Avg W	2.12	3.76	3.41	0.65	2.81	2.4	5.82	3.06	0.72	2.68	0.5	0.88	0.86	1.66
Loss_∞	2.44	2.44	2.44	2.29	2.29	1.93	1.93	1.93	1.88	1.88	3.09	2.45	0.8	3.53
Loss_2	0.24	0.24	0.24	0.21	0.21	0.2	0.2	0.2	0.18	0.18	0.52	0.16	0.05	0.47
Min W	1.12	1.96	3.41	0.32	2.81	1.25	3.03	3.06	0.34	2.68	0.39	0.56	0.56	1.66
Max W	5.35	9.56	3.41	2.32	2.81	5.63	13.73	3.06	2.08	2.68	1.45	1.54	2.2	1.66
Brownian Motion														
%	7.45	1.53	4.58	29.01	6.12	2.32	0	2.05	26.86	2.71	37.02	24.98	20.51	8.58
Avg W	1.71	3.15	2.46	0.52	2.08	2.54	5.83	2.89	0.58	2.54	0.53	0.84	0.85	1.65
Loss_∞	1.96	1.96	1.96	1.96	1.96	1.93	1.93	1.93	1.88	1.88	2.22	1.83	1.47	1.76
Loss_2	0.23	0.23	0.23	0.21	0.21	0.21	0.21	0.21	0.19	0.19	0.24	0.18	0.13	0.19
Min W	0.92	1.69	2.46	0.24	2.08	1.39	3.19	2.89	0.25	2.54	0.4	0.57	0.56	1.65
Max W	3.77	6.92	2.46	1.63	2.08	5.5	12.63	2.89	1.89	2.54	1.52	1.22	2.14	1.65
Bumps														
%	6.33	2.56	2.89	33.65	4.01	2.02	0.34	1.46	25.22	2.01	74.19	29.36	29.71	32.92
Avg W	4.33	6.26	5.38	0.7	4.37	3.83	6.39	4.53	0.87	3.85	0.41	0.95	0.93	1.61
Loss_∞	4.06	4.06	4.06	3.92	3.92	3.26	3.26	3.26	3.2	3.2	4.09	3.25	2.98	3.44
Loss_2	0.58	0.58	0.58	0.55	0.55	0.37	0.37	0.37	0.33	0.33	1.9	0.61	0.44	0.97
Min W	2.69	3.89	5.38	0.34	4.37	2.18	3.63	4.53	0.34	3.85	0.33	0.56	0.57	1.61
Max W	10.01	14.55	5.38	3.47	4.37	8.93	14.87	4.53	3.01	3.85	1.2	1.74	2.27	1.61
Blocks														
%	5.38	2.98	1.77	19.27	2.65	1.44	0.44	1.19	18.59	1.68	62.69	26.81	39.24	33.65
Avg W	4.61	6.22	6.6	0.72	5.22	3.66	5.89	5.84	0.77	4.72	0.47	0.93	0.92	1.7
Loss_∞	4.11	4.11	4.11	4.08	4.08	4.34	4.34	4.34	4.16	4.16	3.88	4.45	3.1	3.23
Loss_2	0.61	0.61	0.61	0.56	0.56	0.39	0.39	0.39	0.35	0.35	1.66	0.91	0.66	1.29
Min W	2.88	3.85	6.6	0.34	5.22	2.21	3.53	5.84	0.34	4.72	0.37	0.54	0.55	1.7
Max W	13.54	18.37	6.6	4.1	5.22	10.74	17.31	5.84	3.81	4.72	1.24	1.96	2.08	1.7

Table 1: The numbers are averages over 100 repetitions. % stands for the percentage of non-covered points $f_0(x)$ for $x \in \mathcal{X} = \{x_i : 1 \le i \le n\}$; Avg W, Min W and Max W stand for average (over $\mathcal{X}$), minimal and maximal width; $\text{Loss}_\infty = \max_{x\in\mathcal{X}} |\hat{f}(x) - f_0(x)|$ and $\text{Loss}_2 = \frac{1}{n}\sum_i (\hat{f}(x_i) - f_0(x_0))^2$ are the losses of a point estimator (posterior median for $C_n^1, \mathcal{C}_n^2, \widetilde{\mathcal{C}}_n$, posterior mean for $\mathcal{P}_n, L_\infty$, the band midpoint for CLM, BIN1 and BIN2 and a point estimator of CCF after adaptively choosing the number of bins before de-biasing).

we implement the locally *non-adaptive* band [18], denoted by $\widetilde{\mathcal{C}}_n$, which uses $\sup_{x\in[0,1]} \sigma_n(x)$ as the global diameter in (3.6). Again, we choose v_n adaptively so that $\widetilde{\mathcal{C}}_n$ contains 95% of posterior draws. We compare $\widetilde{\mathcal{C}}_n$ with a frequentist counterpart[7] [14] where the global level of truncation is estimated by performing tests on individual wavelet coefficients. We denote this method by CLM in our tables, using $\alpha = 0.05$. Next, we compare our bands to 95% credible L_∞ bands centered at the posterior mean estimator $\widehat{f}$ (i.e. $L_\infty \equiv \{f : \sup_{x\in[0,1]} |f(x) - \widehat{f}(x)| \le R_\alpha\}$, where R_α is the 95% sample quantile of $\max_{x\in\mathcal{X}} |f_i(x) - \widehat{f}(x)|$ where f_i for $1 \le i \le M$ are the posterior draws of f and $\mathcal{X} = \{x_i : 1 \le i \le n\}$. This construction is locally non-adaptive and, although similar to the multiscale credible band in [18], its coverage properties are not theoretically understood. We also included a point-wise 95% credible band (denoted by $\mathcal{P}_n$ in our Tables) and a (pointwise) band from a recent R package `nprobust` [15] (using default settings) which implements robust

[7] We used authors' Matlab code with a Symmlet 8 basis with default tuning ($\beta_0 = 3$ and $M_0 = 100$).

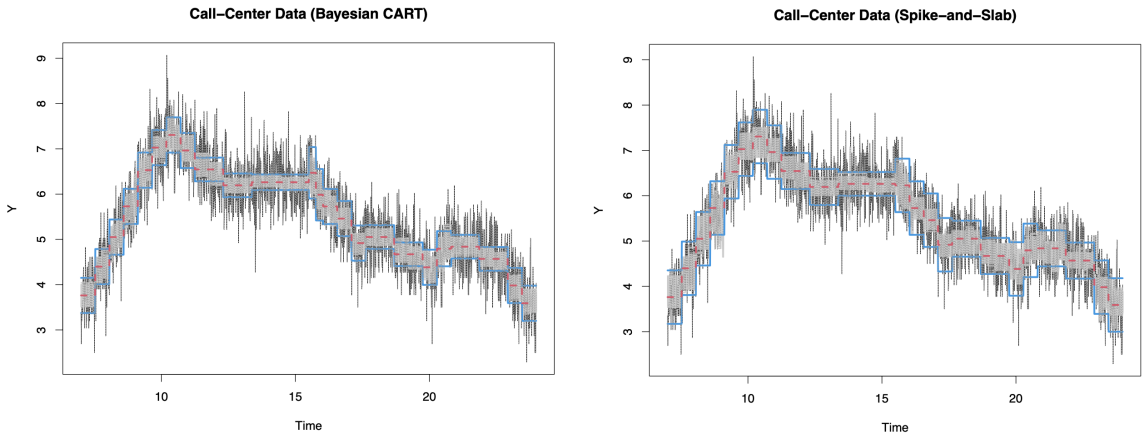

Figure 3: Confidence bands for $f_0(t)$. Left: Bayesian CART prior with $\Gamma = 1.001$. Right: Spike-and-Slab prior with $\Gamma = 2$. Blue lines are the confidence bands with an adaptively chosen v_n ($\alpha = 0.05$) and the red dashed line is the posterior median. True data marked with black (dark gray dotted) lines. The gray area are superimposed posterior samples after burnin (one line for each sample).

bias-corrected bands using local polynomial regression. This method is denoted by CCF in our Tables. Lastly, we compare our adaptive partitioning approach with the binscatter [19] (regressogram) popular among econometricians [67]. The R package `binsreg` provides confidence bands based on bias correction and adaptive selection of the number of bins. We used both piece-wise step functions (option $p = s = 0$) with a non-adaptive placement of splits (denoted as BIN1 in our Tables) as well as the recommended default option (with $p = s = 2$) based on smoother local polynomials with a global smoothing penalty across the bins (denoted as BIN2 in our Tables). While smoother approaches (CCF and BIN2) work well on the Doppler curve (Example G.1), the Bumps and Blocks design dramatically reveal the benefits of our locally adaptive (step function) approach. This simulation study shows distinctive benefits of a Bayesian approach to adaptive confidence band construction. Plots of the confidence bands for the Brownian motion example is in Figure 2 and for all 4 test functions in Section G (Figure 2 and 3).

5.2 Call Center Data

The data set, studied in [11] and [14], consists of arrival times of regular service calls to the call center of an Israeli bank from August to October in 1999. Following [11], the number of calls are assumed to arrive according to an inhomogeneous Poisson process with a mean function $\mu(t)$. We want to non-parametrically estimate $\mu(t)$ and to provide a confidence band. Similarly as in [14] we divide daily operating times (7-AM till midnight) into $n = 2048$ equispaced intervals and compute responses Y_i for i^{th} time interval as $Y_i = \sqrt{N_i + 1/4}$ where $N_i \sim Pois[\mu(t_i)]$ is the number of phone calls arriving in the i^{th} interval. These transformed data Y_i approximately follow the model (2.1) with $f_0(t) = \sqrt{\mu(t)}$ and with a fixed variance $\sigma^2 = 1/4$ [11]. We applied the Bayesian CART and Spike-and-Slab priors with $\Gamma = 1.001$ (MH acceptance rate 12%) and $\Gamma = 2$ (MH acceptance rate[8] 4%), respectively. Instead of fixing the variance at the theoretically justified value $\sigma^2 = 1/4$, we estimated it using the $(1/2, 1/2)$-inverse Gamma prior. The posterior mean and credible interval for σ^2 was $0.27(0.261, 0.296)$ for Bayesian CART and $0.28(0.263, 0.31)$, suggesting that Bayesian CART leaves less unexplained variance and yields an estimate that is closer to the true theoretical value 0.25. We also saw that the Markov chain under the Spike-and-Slab prior took longer to escape from the initialization at $(l, k) \in \{(0, 0), (1, 0), (0, 1)\}$. This example shows the benefits of tree-shaped regularization which prevents from the inclusion of spurious high-resolution signals and thereby yields smoother reconstructions and tighter bands. After burning 1 000 of the 5 000 MCMC iterations, we construct the band $\mathcal{C}_n$ in (3.7) with an adaptively chosen v_n so as the band consists of 95% posterior probability ($v_n = 0.52$ for Bayesian CART and $v_n = 1.85$ for Spike-and-Slab). The results are summarized in Figure 3 where we plot the transformed data Y_i and confidence bands for $f_0(t)$. In order to obtain bands for $\mu(t)$ one could transform the results by taking the square [11].

[8] In order to achieve the acceptance rate 12% for the Spike-and-Slab prior, we would need to decrease Γ to, say, 1.5. This results in inclusion of spurious deeper coefficients and thereby wider confidence bands.

6 Discussion

This work studies spatial adaptivity aspects of popular Bayesian machine learning procedures including Bayesian CART, Gaussian processes, spike-and-slab wavelet reconstructions and variable-knot splines. We have focused on Hölderian classes where the smoothness is varying over the function domain. We have shown uniform (near)-minimax local adaptation in the supremum-norm sense in white noise as well as non-parametric regression for Bayesian CART and spike-and-slab priors. We have also provided a valid frequentist framework for uncertainty quantification with confidence set with asymptotic coverage 1 and whose width is optimal and varies with local smoothness. We proposed a new class of repulsive partitioning priors which relate to variable-knot spline techniques and showed that they are locally rate-exact. Although we have only treated regression-type models, our results can be extended to, for example, density or Poisson intensity estimation estimation using a formulation similar to [16]. Extensions to higher dimensions $d > 1$ are straightforward using, for example, tensor products of Haar basis functions. The spike-and-slab approach extends naturally to tensor products while the Bayesian CART approach lends itself to d-ary trees (as opposed to binary trees), where each internal node has 2^d children (see Section 7.4 in the Supplement of [18]). Alternatively, multivariate Bayesian CART can be more transparently translated using anisotropic Haar wavelet basis functions which more closely resemble recursive partitioning [28]. One would need to make sure that the partition is sufficiently regular in the sense that the binary trees split roughly equally along each direction during the anisotropic dictionary construction. A similar requirement would be needed for the repulsive prior from Section 4.3.

References

[1] Andersson, P. (1997). Characterization of pointwise Hölder regularity. *Applied and Computational Harmonic Analysis 4*, 429–443.

[2] Arbel, J., G. Gayraud, and J. Rousseau (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian Journal of Statistics 40*, 549–570.

[3] Baladandayuthapani, V., B. Mallick, and R. Carroll (2005). Spatially adaptive Bayesian regression splines. *Journal of Computational and Graphical Statistics 14*, 378–394.

[4] Barbieri, M. and J. Berger (2004). Optimal predictive model selection. *The Annals of Statistics 32*, 870–897.

[5] Barbieri, M., J. Berger, E. George, and V. Ročková (2020). The median probability model and correlated variables. *Bayesian Analysis (to appear)*, 1–30.

[6] Belitzer, E. and S. Ghosal (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Annals of Statistics 31*, 536–559.

[7] Besag, J. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society (Series B) 34*, 75–83.

[8] Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

[9] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and Regression Trees*. Wadsworth and Brooks.

[10] Breiman, L., W. Meisel, and E. Purcell (1977). Variable kernel estimates of multivariate densities. *Technometrics 19*, 135–144.

[11] Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao (2005). tatistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association 100*, 36–50.

[12] Brown, L. and M. Low (1996). Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics 3*, 2384–2398.

[13] Bull, A. (2012). Honest adaptive confidence bands and self-similar functions. *Electronic Journal of Statistics 6*, 1490–1516.

[14] Cai, T., M. Low, and Z. Ma (2014). Adaptive confidence bands for non-parametric regression functions. *Journal of the American Statistical Association 109*, 1054–1070.

[15] Calonico, S., M. D. Cattaneo, and M. H. Farrell (2019). nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference. *Journal of Statistical Software 91*(8), 1–33.

[16] Castillo, I. and R. Mismer (2021). Spike and slab Polya tree posterior densities: Adaptive inference. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 57*(3), 1521 – 1548.

[17] Castillo, I. and R. Nickl (2014). On the Bernstein–von Mises theorem for nonparametric Bayes proceduress. *The Annals of Statistics 42*, 1941–1969.

[18] Castillo, I. and V. Ročková (2020). Uncertainty quantification for Bayesian CART. *The Annals of Statistics (to appear)*.

[19] Cattaneo, M., R. Crump, M. Farrell, and Y. Feng (2019). On binscatter. *ArXiv*.

[20] Chipman, H., E. George, and R. McCulloch (1997). Bayesian CART model search. *Journal of the American Statistical Association 93*, 935–960.

[21] Chipman, H., E. George, and R. McCulloch (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics 4*, 266–298.

[22] Chipman, H., E. D. Kolaczyk, and R. McCulloch (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association 92*, 1413–1421.

[23] Cohen, A., I. Daubechies, and P. Vial (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal. 1*(1), 54–81.

[24] Crainiceanu, C., D. Ruppert, and R. Carroll (2007). Spatially adaptive bayesian P-splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics 16*, 265–288.

[25] de Boor, C. (1973). Good approximation by splines with variable knots. *Spline Functions and Approximation Theory 25*, 57–72.

[26] Delyon, B. and A. Judistky (1996). On minimax wavelet estimators. *Applied and Computational Harmonic Analysis 3*, 215–228.

[27] Denison, D., B. Mallick, and A. Smith (1998). A Bayesian CART algorithm. *Biometrika 85*, 363–377.

[28] Donoho, D. (1997). CART and best-ortho-basis: a connection. *Annals of Statistics 25*, 1870–1911.

[29] Donoho, D. and I. Johnstone (1995a). Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81*, 425–455.

[30] Donoho, D. and I. Johnstone (1995b). Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society. Series B 57*, 301–369.

[31] Donoho, D. and I. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81*, 425–455.

[32] Donoho, D. and I. M. Johnstone (1998). Minimax estimation via wavelet shrinkage. *Annals of Statistics 26*, 879–921.

[33] Donoho, D. L. and I. M. Johnstone (1995c). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association 90*, 1200–1224.

[34] Fan, J. and I. Gijbels (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society. Series B 57*, 371–394.

[35] Friedman, J. (1991). Multivariate adaptive regression splines. *The Annals of Statistic 19*, 1–61.

[36] Friedman, J. and B. Silverman (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics 31*, 3–39.

[37] Gach, F., R. Nickl, and V. Spokoiny (2013). Spatially adaptive density estimation by localised Haar projections. *Annales de l'Institut Henri Poincar'e, Probabilit'es et Statistiques 49*, 900–914.

[38] Gine, E. and R. Nickl (2010). Confidence bands in density estimation. *The Annals of Statistics 38*, 1122–1170.

[39] Gine, E. and R. Nickl (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.

[40] Hayakawa, S. and T. Suzuki (2019). On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *arXiv:1905.09195*, 1–40.

[41] Hoffmann, M., J. Rousseau, and J. Schmidt-Hieber (2015). On adaptive posterior concentration rates. *The Annals of Statistics 43*, 2259–2295.

[42] Jeong, S. and V. Ročková (2020). The art of BART: On flexibility of Bayesian forests. *Manuscript*, 1–44.

[43] Katkovnik, V. and V. Spokoiny (2008). Spatially adaptive estimation via fitted local likelihood techniques. *IEEE Transactions on Signal Processing 56*, 873–880.

[44] Knapik, B., B. Szabo, and A. van der Vaart (2011). Bayesian inference problems with gaussian priors. *Annals of Statistics 39*, 2626–2657.

[45] Krivobokova, T., C. Crainiceanu, and G. Kauermann (2008). Fast adaptive penalized splines. *Journal of Compuational and Graphical Statistics 17*, 1–20.

[46] Lang, S. and A. Bretzger (2004). Bayesian P-splines. *Journal of Compuational and Graphical Statistics 13*, 183–212.

[47] Leopold, H. G. (1989). On Besov spaces of variable order of differentiation. *Journal of Analysis and its Applications 8*, 69–82.

[48] Lepski, O. V., M. E. and V. G. Spokoiny (1997). Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection. *The Annals of Statistics 25*, 929–947.

[49] Low, M. (1997). On nonparametric confidence intervals. *The Annals of Statistics 25*, 2547–2554.

[50] Luo, Z. and G. Wahba (1997). Hybrid adaptive splines. *Journal of the American Statistical Association 92*, 107–116.

[51] Mammen, E. and S. van der Geer (1997). Locally adaptive regression splines. *The Annals of Statistics 25*, 387—413.

[52] Muller, H. and U. Stadtmuller (1987). Variable bandwidth kernel estimators of regression curves. *The Annals of Statistics 15*, 182–201.

[53] Narisetty, N. and X. He (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics 42*, 789–817.

[54] Nickl, R. and B. Szabo (2016). A sharp adaptive confidence ball for self-similar functions. *Stochastic Processes and their Applications 126*, 3913–3934.

[55] Pati, D., A. Bhattacharya, and G. Cheng (2015). Optimal Bayesian estimation in random covariate design with a rescaled Gaussian process prior. *Journal of Machine Learning Research 16*, 2837–2851.

[56] Picard, D. and K. Tribouley (2000). Adaptive confidence interval for pointwise curve estimation. *The Annals of Statistics 28*, 298–335.

[57] Pintore, A., P. Speckman, and C. Holmes (2006). Spatially adaptive smoothing splines. *Biometrika 93*, 113–125.

[58] Polson, N. and Ročková (2018). Posterior concentration for sparse deep learning. *Advances in Neural Information Processing Systems 31*, 697–704.

[59] Ray, K. (2017). Adaptive Bernstein–vin Mises theorems in Gaussian white noise. *The Annals of Statistics 45*, 2511–2536.

[60] Rousseau, J. and B. Szabo (2017). Asymptotic behavior of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *The Annals of Statistics 45*, 833–865.

[61] Ročková, V. and E. Saha (2019). On theory for BART. *JMLR: Workshop and Conference Proceedings: 22nd International Conference on Artificial Intelligence and Statistics 89*, 2839–2848.

[62] Ročková, V. and S. van der Pas (2020). Posterior concentration for Bayesian regression trees and forests. *The Annals of Statistics 48*, 2108–2131.

[63] Ruppert, D. and R. Carroll (2000). Spatially-adaptive penalties for spline fitting. *Australian & New Zealand Journal of Statistics 42*, 205–223.

[64] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics 48*, 1875–1897.

[65] Schneider, J. (2007). Function spaces of varying smoothness. *Mathematische Nachrichten 280*, 1801–1826.

[66] Shen, X. and S. Ghosal (2015). Adaptive Bayesian procedures using random series priors. *Scandinavian Journal of Statistics 42*, 1194–1213.

[67] Starr, E. and B. Goldfarb (2020). Binned scatterplots: A simple tool to make research easier and better. *Strategic Management Journal 41*, 2261–2274.

[68] Stone, C. (1982). Optimal rates of convergence for nonparametric regression. *Annals of Statistics 10*, 1040–1053.

[69] Suzuki, T. (2018). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. *arXiv:1810.08033*, 1–40.

[70] Szabo, B., A. van der Vaart, and J. van Zanten (2013). Empirical Bayes scaling of Gaussian priors in the white noise model. *Electronic Journal of Statistics 7*, 991–1018.

[71] Tibshirani, R. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics 42*, 285–323.

[72] van der Vaart, A. and J. van Zanten (2008). Rates of contraction of posterior distributoins based on Gaussian process priors. *The Annals of Statistics 36*, 1435–1463.

[73] Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*, 1228–1242.

[74] Wager, S. and W. Guenther (2015). Adaptive concentration of regression trees with application to random forests. *Manuscript*.

[75] Yoo, W., V. Rivoirard, and J. Rousseau (2018). Adaptive supremum norm posterior contraction: Wavelet spike-and-slab and anisotropic Besov spaces. *arXiv:1708.01909*, 1–50.