Adaptive Priors based on Splines with Random Knots

Eduard Belitser and Paulo Serra
Department of Mathematics, Eindhoven University of Technology

March 15, 2013

Abstract

Splines are useful building blocks when constructing priors on nonparametric models indexed by functions. Recently it has been established in the literature that hierarchical priors based on splines with a random number of equally spaced knots and random coefficients in the B-spline basis corresponding to those knots lead, under certain conditions, to adaptive posterior contraction rates, over certain smoothness functional classes. In this paper we extend these results for when the location of the knots is also endowed with a prior. This has already been a common practice in MCMC applications, where the resulting posterior is expected to be more "spatially adaptive", but a theoretical basis in terms of adaptive contraction rates was missing. Under some mild assumptions, we establish a result that provides sufficient conditions for adaptive contraction rates in a range of models.

Keywords:Adaptive estimation, bayesian non-parametric, optimal contraction rate, spline, random knots.

1 Introduction

The Bayesian approach in statistics has become quite popular in recent years as an alternative to classical *frequentist* methods. The main appeal of the Bayesian methodology is its conceptual simplicity: given a model for the observed data $X \sim P_f$, $f \in \mathcal{F}$, some space of functions, put a prior on the unknown parameter f and draw inferences based on the resulting posterior $\Pi(f|X)$. Knowledge about the model under study can also be be incorporated into the inference procedure via the prior. However, some seemingly "correct" priors can lead to unreasonable posteriors, especially in nonparametric models. It is therefore desirable to place ourselves in a setting where it is possible to assess the quality of the resulting posterior from some objective point of view.

This gave rise to the development of the notion of contraction rate (cf. Ghosal et al. (2000)), a Bayesian analog of a convergence rate: data is assumed to come from a fixed probability measure $P_0 = P_{f_0}$ for a "true" $f_0 \in \mathcal{F}$; the contraction rate is then the smallest radius such that the posterior mass in a Hellinger ball of probability measures around P_0 converges to 1 in P_0-probability as some information index such as a sample size goes to infinity.

Some general results about posterior contraction rates establish sufficient conditions on prior distributions such that the resulting posteriors attain a certain contraction rate. In this spirit, when studying specific priors, some authors now choose to present their results in the form of say *meta-theorems* which claim that sufficient conditions (such as the ones in Ghosal et al. (2000)) required to attain a certain range of contraction rates hold for their choice of prior; cf. de Jonge and van Zanten (2012), Shen and Ghosal (2012), van der Vaart and van Zanten (2008) and further references therein. We adopt this practice here as well.

In the case where f_0 is a function from some functional space of smoothness α, the posterior contraction rate is typically compared to the convergence rate of the minimax risk (called optimal rate) over that space in the estimation problem. For example, if we observe a sample of size n and want to estimate a univariate α-smooth function (e.g., density or regression function), the typical optimal rate is of order $n^{-\alpha/(2\alpha+1)}$, possibly up to a logarithmic factor depending on the risk function. If the smoothness parameter α is unknown, and one wants to build estimators which attain the optimal rate corresponding to α but do not depend explicitly on α, one speaks of an adaptation problem. In a Bayesian context, the adaptation problem consists in finding a prior which leads to the optimal posterior contraction rate (usually up to a logarithmic factor) for any α-smooth function of interest and does not depend on the smoothness parameter α. Such priors are called rate adaptive. There is a growing number of papers, where this problem has been studied in different settings; cf. de Jonge and van Zanten (2012), Shen and Ghosal (2012), van der Vaart and van Zanten (2008), van der Vaart and van Zanten (2009) and Belitser and Ghosal (2003) among others.

Splines, in particular, can be used when constructing adaptive priors. A spline (cf. de Boor (1978)) is a piecewise polynomial function designed to have a certain level of smoothness which is referred to as its order. Splines are easy to store, differentiate, integrate and evaluate on a computer, and are extensively used in practice for constructing good, parsimonious approximations of smooth functions. The points at which the different polynomial pieces of a spline connect are called knots. If an order (read: maximal polynomial degree) and a set of knots is fixed, then the space of all splines with that order and those knots forms a linear space which admits a basis of so called B-splines. Any spline of a fixed order is consequently characterized by a set of knots and its coordinates in the B-splines basis corresponding to those knots. Randomly generating a number of knots and, given those, generating random coordinates in the corresponding B-spline basis with equally spaced knots results in a random spline whose law can be used as a prior. If, given the number of knots, the coordinates in the corresponding B-spline basis are chosen to be independent and normally distributed, then the resulting spline has a conditionally Gaussian law and was studied by de Jonge and van Zanten (2012) by using Reproducing Kernel Hilbert Space techniques. Shen and Ghosal (2012) propose a more general, random series prior: the coefficients in the series are not necessarily independent or Gaussian and a basis other than the B-spline basis can also be used.

The case where the locations of the knots are also random is not covered by the results of either de Jonge and van Zanten (2012) or Shen and Ghosal (2012). However when practitioners put a prior on the number of knots they almost invariably also put a prior on the

locations of the knots (e.g., Denison et al. (1998), Di Matteo et al. (2001), Sharef et al. (2010)) – a Poisson process is a popular choice. Their motivation for allowing arbitrarily located knots seems to be twofold. Firstly, this is attractive from the implementation point of view: designing reversible jump MCMC samplers is much simpler if any collection of knots is allowed since new knots can be inserted at arbitrary positions causing only localized changes in the spline. Secondly, the resulting posterior based on the prior with random locations of the knots is expected to be more "spatially adaptive": the function of interest may not have a fixed level of smoothness throughout its support, it may consist of rough and smooth pieces. To sustain an adequate level of accuracy over the whole support, more knots are needed in rough pieces and less in smooth ones. Therefore, to make it at least possible for the resulting posterior to pick up eventual spatial features of the function, the prior has to be flexible enough to model random locations of the knots.

In this paper, we extend the results of de Jonge and van Zanten (2012), and those of Shen and Ghosal (2012) in respect to the prior with random knots: we add one more level to the hierarchical spline prior by putting a prior on the location of the knots of the spline as well, making, in fact, the basis functions also random. Under some mild assumptions on the proposed hierarchical spline prior, we establish our main result for the proposed prior, providing sufficient conditions for adaptive, optimal contraction rates of the resulting posterior in a range of models (among others: density estimation, nonparametric regression, binary regression, Poisson regression, and classification). In doing so, we provide a theoretical basis for the common practice of using randomly located knots in spline based priors.

2 Notation and preliminaries on splines

First we introduce some notation. For $d \in \mathbb{N}$ and $1 \leq p < \infty$ denote by $\|\boldsymbol{x}\|_p = \big(\sum_{i=1}^d |x_i|^p\big)^{1/p}$ the l_p-norm of $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and by $\|\boldsymbol{x}\|_\infty = \max_{i=1,\ldots,d} |x_i|$. For $1 \leq p < \infty$ let the L_p-norm of a function f on [0,1] be $\|f\|_p = \big(\int_0^1 |f(x)|^p \, dx\big)^{1/p}$ and $\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|$.

We use $\lesssim$ (respectively $\gtrsim$) to denote smaller (respectively greater) or equal up to a constant, the symbols $a \vee b$ and $a \wedge b$ stand for $\max\{a, b\}$ and $\min\{a, b\}$ respectively. The covering number $N(\epsilon, S, d)$ of a subset S of a metric space with balls of size ϵ is the smallest number of balls (with respect to distance d) of radius ϵ needed to cover S.

Now we provide some preliminaries on splines, which can be found, for example, in Schumaker (2007). A function is called a spline is of order $q \in \mathbb{N}$, with respect to a certain partition of its support, if it is $q-2$ times continuously differentiable and when restricted to each interval in this partition, coincides with a polynomial of degree at most $q-1$. Consider $q \in \mathbb{N}$, $q \geq 2$, which will be fixed throughout the remainder of this text. For any $j \in \mathbb{N}$, such that $j \geq q$ let $\mathcal{K}_j = \{(k_1, \ldots, k_{j-q}) \in (0,1)^{j-q} : 0 < k_1 < \cdots < k_{j-q} < 1\}$. We will refer to a vector $\boldsymbol{k} = \boldsymbol{k}_j \in \mathcal{K}_j$ as a set of inner knots; the index j in $\boldsymbol{k}_j$ will sometimes be used to emphasize the dependence on j. A vector $\boldsymbol{k} \in \mathcal{K}_j$ will be said to induce the partition $\big\{[k_0, k_1), [k_1, k_2), \ldots, [k_{j-q}, k_{j-q+1}]\big\}$, with $k_0 = 0$ and $k_{j-q+1} = 1$. For any $\boldsymbol{k} \in \mathcal{K}_j$ we will call $M(\boldsymbol{k}) = \max_{i=1}^{j-q+1} |k_i - k_{i-1}|$ the mesh size of the partition

induced by $\boldsymbol{k}$ and $m(\boldsymbol{k}) = \min_{i=1}^{j-q+1} |k_i - k_{i-1}|$ the sparseness of the partition induced by $\boldsymbol{k}$. For a $\boldsymbol{k} \in \mathcal{K}_j$, denote by $\mathcal{S}^{\boldsymbol{k}} = \mathcal{S}_q^{\boldsymbol{k}}$ the linear space of splines of order q on $[0,1]$ with simple knots $\boldsymbol{k}$ (see the definition of knot multiplicity in Schumaker (2007)). This space has dimension j and admits a basis of so called B-splines $\{B_1^{\boldsymbol{k}}, \ldots, B_j^{\boldsymbol{k}}\}$. The construction of $\{B_1^{\boldsymbol{k}}, \ldots, B_j^{\boldsymbol{k}}\}$ involves the knots $k_{-q+1}, \ldots, k_{-1}, k_0, k_1, \ldots, k_{j-q}, k_{j-q+1}, k_{j-q+2}, \ldots, k_j$, with arbitrary extra knots $k_{-q+1} \leq \cdots \leq k_{-1} \leq k_0 = 0$ and $1 = k_{j-q+1} \leq k_{j-q+2} \leq \cdots \leq k_j$. Usually one takes $k_{-q+1} = \cdots = k_{-1} = k_0 = 0$ and $1 = k_{j-q+1} = \cdots = k_j$, and we adopt this choice here as well. These basis functions are nonnegative: $B_i^{\boldsymbol{k}}(x) \geq 0$, for all $x \in [0,1]$. Besides, they have local support and form a partition of unity:

$$B_i^{\boldsymbol{k}}(x) = 0 \ \text{ for } x \notin [k_{-q+i}, k_i], \quad \sum_{i=1}^{j} B_i^{\boldsymbol{k}}(x) = 1 \ \text{ for all } x \in [0,1]. \tag{1}$$

To refer explicitly to the coordinates $\boldsymbol{a} = (a_1, \ldots, a_j) \in \mathbb{R}^j$ of a spline on a specific B-spline basis with inner knots $\boldsymbol{k}$, we write $s_{\boldsymbol{a},\boldsymbol{k}}(x) = \sum_{i=1}^{j} a_i B_i^{\boldsymbol{k}}(x)$, $x \in [0,1]$. Since $\sum_{i=1}^{j} B_i^{\boldsymbol{k}}(x) = 1$, it is easy to see that for any $s_{\boldsymbol{a},\boldsymbol{k}}, s_{\boldsymbol{b},\boldsymbol{k}} \in \mathcal{S}_q^{\boldsymbol{k}}$

$$\|s_{\boldsymbol{a},\boldsymbol{k}} - s_{\boldsymbol{b},\boldsymbol{k}}\|_2 \leq \|s_{\boldsymbol{a},\boldsymbol{k}} - s_{\boldsymbol{b},\boldsymbol{k}}\|_\infty \leq \|\boldsymbol{a} - \boldsymbol{b}\|_\infty \leq \|\boldsymbol{a} - \boldsymbol{b}\|_2. \tag{2}$$

Splines have good approximation properties for sufficiently smooth functions provided they are defined on a partition with appropriately small mesh size. We say that a function f on $[0,1]$ belongs to a generic smoothness class $\mathcal{F}_\alpha$, $\alpha > 0$, if for any set of inner knots $\boldsymbol{k}$ there exists a spline $s_{\boldsymbol{a},\boldsymbol{k}} \in \mathcal{S}_q^{\boldsymbol{k}}$ such that for some bounded C_f

$$\|f - s_{\boldsymbol{a},\boldsymbol{k}}\|_\infty \leq C_f M^\alpha(\boldsymbol{k}). \tag{3}$$

We will also be assuming that $\mathcal{F}_\alpha$ is contained in a Lipschitz class: $\mathcal{F}_\alpha \subseteq \mathcal{L}(\kappa_\alpha, L_\alpha) = \{f : |f(x_1) - f(x_2)| \leq L_\alpha |x_1 - x_2|^{\kappa_\alpha}, x_1, x_2 \in [0,1]\}$ for some $\kappa_\alpha, L_\alpha > 0$.

A leading example of a smoothness class $\mathcal{F}_\alpha$ is the Hölder space $\mathcal{H}_\alpha = \mathcal{H}_\alpha(L, [0,1])$, $0 < \alpha \leq q$, which is the collection of all functions f that have bounded derivatives up to order $\alpha_0 = \lfloor \alpha \rfloor = \max\{z \in \mathbb{Z} : z < \alpha\}$ and such that the α_0-th derivative satisfies the Hölder condition $|f^{(\alpha_0)}(x) - f^{(\alpha_0)}(y)| \leq L|x - y|^{\alpha - \alpha_0}$, for $L > 0$ and $x, y \in [0,1]$. In this case, a well-known spline approximation result (cf. de Boor (1978)) claims that (3) holds with $C_f = C_q \|f^{(\alpha)}\|_\infty$ for some constant C_q depending only on q. Other examples of smoothness classes for which the approximation property (3) hold, include α-times continuously differentiable functions, Sobolev and Besov spaces; cf. Theorems 6.21, 6.25 and 6.31 in Schumaker (2007).

3 Main Result

We begin by describing a hierarchical prior on $\mathcal{S} = \mathcal{S}_q = \cup_{j=q}^{\infty} \cup_{\boldsymbol{k} \in \mathcal{K}_j} \mathcal{S}_q^{\boldsymbol{k}}$: first draw a number $J \in \mathbb{N}$, $J \geq q$; then, given J, generate independently $(J - q)$ inner knots $\boldsymbol{K}_J \in \mathcal{K}_j$ and also independently, J B-spline coefficients $\boldsymbol{\theta} \in \mathbb{R}^J$. Our prior on $\mathcal{S}$ will be the law of

the random spline $s_{\boldsymbol{\theta},\boldsymbol{K}_J}$. We impose the following conditions on this prior. For $c_1, c_2 > 0$, $0 \le t_1, t_2 \le 1$ and all sufficiently large j,

$$\mathbb{P}(J > j) \lesssim \exp\big(-c_1 j \log^{t_1} j\big), \tag{4}$$

$$\mathbb{P}(J = j) \gtrsim \exp\big(-c_2 j \log^{t_2} j\big). \tag{5}$$

For some $\tau \ge 1$, $c_3 > 0$, $0 \le t_3 \le 1$, and all $j \ge q$,

$$\mathbb{P}\big(m(\boldsymbol{K}_j) < \delta(j) | J = j\big) = 0, \tag{6}$$

$$\mathbb{P}\big(M(\boldsymbol{K}_j) \le \tau/j | J = j\big) \gtrsim \exp\big(-c_3 j \log^{t_3} j\big), \tag{7}$$

where $\delta(i)$ is a positive, strictly decreasing function on $\mathbb{N}$. Without loss of generality assume that $\delta(i) \le 1$, $i \in \mathbb{N}$. For each $j \ge q$, the conditional distribution of $\boldsymbol{\theta} \in \mathbb{R}^j$ satisfies the following condition: for any $M > 0$ there exists $c_0 = c_0(M)$ such that

$$\mathbb{P}\big(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_\infty \le \epsilon | J = j\big) \gtrsim \exp\big(-c_0 j \log(1/\epsilon)\big) \tag{8}$$

for all $\epsilon > 0$ and all $\boldsymbol{\theta}_0 \in \mathbb{R}^j$ such that $\|\boldsymbol{\theta}_0\|_\infty \le M$.

For examples of particular choices on the components of our hierarchical prior which verify these conditions we refer the reader to Section 5.

Denote $\mathcal{C}^j(M) = [-M, M]^j$. The following theorem is our main result.

Theorem 1. *Let $\|f_0\|_\infty < M$ and $f_0 \in \mathcal{F}_\alpha$ so that (3) holds with C_{f_0}. Let $\epsilon_n, \bar{\epsilon}_n$ be two positive sequences such that $\epsilon_n \ge \bar{\epsilon}_n$, $\epsilon_n \to 0$ as $n \to \infty$ and $n\bar{\epsilon}_n^2 > 1$. Assume that there exist sequences $J_n, \bar{J}_n > q$, $M_n > 0$ and a constant $c_M \ge c_1$ satisfying:*

$$J_n \log\Big[\frac{J_n(M_n \vee 1)}{\epsilon_n \delta(J_n)}\Big] \lesssim n\epsilon_n^2, \tag{9}$$

$$\frac{n\bar{\epsilon}_n^2}{\log^{t_1} J_n} \le J_n, \quad P\big(\boldsymbol{\theta} \notin \mathcal{C}^j(M_n) | J = j\big) \lesssim \exp(-c_M n\bar{\epsilon}_n^2),\ q \le j \le J_n, \tag{10}$$

$$\Big[\frac{\bar{\epsilon}_n}{\tau^\alpha C_{f_0}}\Big]^{-1/\alpha} \le \bar{J}_n, \quad \log^{t_2 \vee t_3} \bar{J}_n \lesssim \log\frac{1}{\bar{\epsilon}_n}. \tag{11}$$

Let $\mathcal{S}_n = \cup_{j=q}^{J_n} \cup_{\boldsymbol{k} \in \mathcal{K}_j^{\delta(j)}} \big\{s_{\boldsymbol{\theta},\boldsymbol{k}} \in \mathcal{S}_q^{\boldsymbol{k}} : \|\boldsymbol{\theta}\|_\infty \le M_n\big\}$, where $\mathcal{K}_j^\delta = \{\boldsymbol{k} \in \mathcal{K}_j : m(\boldsymbol{k}) > \delta\}$. Then it holds that

$$\log N(\epsilon_n, \mathcal{S}_n, \|\cdot\|_2) \lesssim n\epsilon_n^2, \tag{12}$$

$$P\big(s_{\boldsymbol{\theta},\boldsymbol{K}_J} \notin \mathcal{S}_n\big) \lesssim \exp\big(-c_1 n\bar{\epsilon}_n^2\big), \tag{13}$$

$$P\big(\|s_{\boldsymbol{\theta},\boldsymbol{K}_J} - f_0\|_\infty \le 2\bar{\epsilon}_n\big) \gtrsim \exp\big\{-(c_0(M) + c_2 + c_3)\bar{J}_n \log(1/\bar{\epsilon}_n)\big\}. \tag{14}$$

Remark 1. Consider constants $c_4, c_5 > 0$ and a function $\delta(\cdot)$ as above. If condition (6) is replaced by

$$\sum_{j=q}^{J_n} \mathbb{P}\big(J = j\big)\mathbb{P}\big(m(\boldsymbol{K}_j) < \delta(j) | J = j\big) \le c_5 \exp(-c_4 n), \tag{6'}$$

then the conclusions of Theorem 1 remain valid so long as J_n is a sequence satisfying (9) and (10) (cf. Section 5 and Remark 4 for a comparison of (6) and (6').)

Proof. First we establish (12). Let $L_n(j) = 4M_n j(q+1)(\delta(j))^{-(q+1)}$ and $j > q$. Let $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{m_1}\}$ be an $\epsilon_n/2$-net of the set $\{\boldsymbol{\theta} \in \mathbb{R}^j : \|\boldsymbol{\theta}\|_\infty \le M_n\}$ and let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{m_2}\}$ be an $\epsilon_n/(2L_n(j))$-net of $\{\boldsymbol{x} \in \mathbb{R}^{j-q} : \boldsymbol{x} \in (0,1)^{j-q}\}$, both with respect to the $\|\cdot\|_\infty$-norm. Then, by using (2) and Lemma 2 (Lemma 2 is applicable since $\epsilon_n/(2L_n(j)) \le \delta(j)$ for sufficiently large n), $\{s_{\boldsymbol{\theta}_k, \boldsymbol{x}_l}, k = 1, \ldots, m_1,\ l = 1, \ldots m_2\}$ forms an ϵ_n-net of $\cup_{\boldsymbol{k} \in \mathcal{K}_j^{\delta(j)}} \big\{ s_{\boldsymbol{\theta}, \boldsymbol{k}} \in \mathcal{S}_q^{\boldsymbol{k}} : \|\boldsymbol{\theta}\|_\infty \le M_n \big\}$ with respect to the $\|\cdot\|_\infty$-norm. By using this fact, we obtain

$$
\begin{aligned}
&N\big(\epsilon_n, \mathcal{S}_n, \|\cdot\|_2\big) \le N\big(\epsilon_n, \mathcal{S}_n, \|\cdot\|_\infty\big) \\
&\le \sum_{j=q}^{J_n} N\Big(\epsilon_n, \cup_{\boldsymbol{k} \in \mathcal{K}_j^{\delta(j)}} \big\{ s_{\boldsymbol{\theta}, \boldsymbol{k}} \in \mathcal{S}_q^{\boldsymbol{k}} : \|\boldsymbol{\theta}\|_\infty \le M_n \big\}, \|\cdot\|_\infty \Big) \\
&\le \sum_{j=q}^{J_n} \Big[N\Big(\frac{\epsilon_n}{2}, \big\{\boldsymbol{\theta} \in \mathbb{R}^j : \|\boldsymbol{\theta}\|_\infty \le M_n\big\}, \|\cdot\|_\infty\Big) N\Big(\frac{\epsilon_n}{2L_n(j)}, (0,1)^{j-q}, \|\cdot\|\infty\Big)\Big] \\
&\le J_n \Big[\frac{2(M_n \vee 1)}{\epsilon_n}\Big]^{J_n} \Big[\frac{2L_n(J_n)}{\epsilon_n}\Big]^{J_n - q} \le J_n \Big(\frac{16(q+1)(M_n \vee 1)^2 J_n}{\epsilon_n^2 (\delta(j))^{q+1}}\Big)^{J_n}.
\end{aligned}
$$

The last relation and (9) imply (12):

$$
\log N\big(\epsilon_n, \mathcal{S}_n, \|\cdot\|_2\big) \lesssim J_n \log\Big[\frac{J_n(M_n \vee 1)}{\epsilon_n \delta(J_n)}\Big] \lesssim n\epsilon_n^2.
$$

Now we check (13). From the definition of $\mathcal{S}_n$, the relations (4), (6) and (10), it follows that

$$
\begin{aligned}
\mathbb{P}\big(s_{\boldsymbol{\theta}, \boldsymbol{K}_J} \notin \mathcal{S}_n\big) &\le \mathbb{P}\big(J > J_n\big) + \sum_{j=q}^{J_n} \mathbb{P}\big(J = j\big) \mathbb{P}\big(m(\boldsymbol{K}_j) < \delta(j) | J = j\big) \\
&\quad + \sum_{j=q}^{J_n} \mathbb{P}\big(J = j\big) \mathbb{P}\big(\boldsymbol{\theta} \notin \mathcal{C}^j(M_n) | J = j\big) \\
&\lesssim \exp\big\{ - c_1 J_n \log^{t_1} J_n \big\} + 0 + \exp\big\{ - c_M n \bar{\epsilon}_n^2 \big\} \\
&\lesssim \exp\big\{ - c_1 n \bar{\epsilon}_n^2 \big\}.
\end{aligned}
$$

It remains to prove (14). First note that, by using (3) and (11), for all $j \ge \bar{J}_n$ and for all sets of knots $\boldsymbol{k}_j \in \mathcal{K}_j$ such that $M(\boldsymbol{k}_j) \le \tau/j$, there exists a spline $s_{\boldsymbol{\theta}_0, \boldsymbol{k}_j} \in \mathcal{S}_q^{\boldsymbol{k}_j}$ (of course, $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0(\boldsymbol{k}_j) = \boldsymbol{\theta}_0(\boldsymbol{k}_j, f_0)$) such that

$$
\|f_0 - s_{\boldsymbol{\theta}_0, \boldsymbol{k}_j}\|_\infty \le C_{f_0} M^\alpha(\boldsymbol{k}_j) \le C_{f_0} \tau^\alpha \bar{J}_n^{-\alpha} \le \bar{\epsilon}_n. \tag{15}
$$

Since $\|f_0\|_\infty < M$ and $\bar{J}_n$ must grow with n in view of (11), it follows from Lemma 3 and (15) $\|\boldsymbol{\theta}_0(\boldsymbol{k}_j)\|_\infty \le M$ for all $\boldsymbol{k}_j \in \mathcal{K}_j$ such that $M(\boldsymbol{k}_j) \le \tau/\bar{J}_n$ for $j \ge \bar{J}_n$.

Introduce the events: $E_1^j = \{M(\boldsymbol{K}_j) \le \tau/j\}$, $E_2^j = \{\|f_0 - s_{\boldsymbol{\theta}_0(\boldsymbol{K}_j), \boldsymbol{K}_j}\|_\infty \le \bar{\epsilon}_n\}$, $E_3^j = \{\|\boldsymbol{\theta}_0(\boldsymbol{K}_j) - \boldsymbol{\theta}\|_\infty \le \bar{\epsilon}_n\}$, $E_4^j = \{\|f_0 - s_{\boldsymbol{\theta}, \boldsymbol{K}_j}\|_\infty \le 2\bar{\epsilon}_n\}$ and $E_5^j = \{\|\boldsymbol{\theta}_0(\boldsymbol{K}_j)\|_\infty \le M\}$.

Using the argument from the previous paragraph, the triangle inequality, (2) and (15), we obtain that

$$E_1^{\bar{J}_n} \subseteq E_2^{\bar{J}_n}, \quad E_1^{\bar{J}_n} \subseteq E_5^{\bar{J}_n}, \quad E_2^j \cap E_3^j \subseteq E_4^j, \quad j \geq q. \tag{16}$$

Combining (5), (7), (8), (11) and (16), we prove (14):

$$\begin{aligned}
&\mathbb{P}\big(\|s_{\boldsymbol{\theta},\boldsymbol{K}_J} - f_0\|_\infty \leq 2\bar{\epsilon}_n\big) = \mathbb{P}(E_4^J) \geq \mathbb{P}(J = \bar{J}_n)\mathbb{P}\big(E_4^{\bar{J}_n}|J = \bar{J}_n\big) \\
&\quad \geq \mathbb{P}(J = \bar{J}_n)\mathbb{P}\big(E_2^{\bar{J}_n} \cap E_3^{\bar{J}_n}|J = \bar{J}_n\big) \\
&\quad \geq \mathbb{P}(J = \bar{J}_n)\mathbb{P}\big(E_1^{\bar{J}_n} \cap E_3^{\bar{J}_n} \cap E_5^{\bar{J}_n}|J = \bar{J}_n\big) \\
&\quad = \mathbb{P}(J = \bar{J}_n)\mathbb{E}\big[P\big(E_1^{\bar{J}_n} \cap E_3^{\bar{J}_n} \cap E_5^{\bar{J}_n}|J = \bar{J}_n, \boldsymbol{K}_{\bar{J}_n}\big)\big] \\
&\quad = \mathbb{P}(J = \bar{J}_n)\mathbb{E}\big[\mathbb{I}\{\boldsymbol{K}_{\bar{J}_n} \in E_1^{\bar{J}_n} \cap E_5^{\bar{J}_n}\}\mathbb{P}\big(E_3^{\bar{J}_n}|J = \bar{J}_n, \boldsymbol{K}_{\bar{J}_n}\big)\big] \\
&\quad \geq \mathbb{P}(J = \bar{J}_n)\mathbb{P}\big(E_1^{\bar{J}_n}|J = \bar{J}_n\big) \inf_{\|\boldsymbol{\theta}_0\|_\infty \leq M} \mathbb{P}\big(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_\infty \leq \bar{\epsilon}_n|J = \bar{J}_n\big) \\
&\quad \gtrsim \exp\big(-(c_2 + c_3)\bar{J}_n \log^{t_2 \vee t_3} \bar{J}_n\big) \exp\big(-c_0(M)\bar{J}_n \log(1/\bar{\epsilon}_n)\big) \\
&\quad \gtrsim \exp\big(-(c_0(M) + c_2 + c_3)\bar{J}_n \log(1/\bar{\epsilon}_n)\big).
\end{aligned}$$

□

Remark 2. If the range of the underlying curve f_0 is contained in some known interval $[a, b] \subset \mathbb{R}$, then, according to Lemma 3 and the proof of property (14), the prior on $\boldsymbol{\theta} \in \mathbb{R}^j$ can be chosen to be supported on, say, $[a-1, b+1]^j$ so that (8) has to hold only for $\boldsymbol{\theta}_0 \in [a-1, b+1]^j$. Condition (10) will trivially be satisfied for $M_n > (1-a) \wedge (b+1)$.

Remark 3. If (20) is assumed instead of (7), the proof of (14) can then be simplified a lot, as in this case one can condition on the event $\{\boldsymbol{K}_{\bar{J}_n} = \bar{\boldsymbol{k}}_{\bar{J}_n}\}$ so that $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0(\bar{\boldsymbol{k}}_{\bar{J}_n})$ becomes fixed and $\mathbb{P}(E_1^J|J = \bar{J}_n, \boldsymbol{K}_{\bar{J}_n} = \bar{\boldsymbol{k}}_{\bar{J}_n}) = 1$.

Remark 4. Condition (6) is used in the proof of Theorem 1 exclusively to enforce $\sum_{j=q}^J \mathbb{P}\big(J = j\big)\mathbb{P}\big(m(\boldsymbol{K}_j) < \delta(j)|J = j\big)$ to be zero. Inspection of the proof shows, however, that it would suffice to require this sum to be upper-bounded by a multiple of $\exp\big\{-c_1 n\bar{\epsilon}_n^2\big\}$. Although this would be a weaker requirement, typically the sequence $\bar{\epsilon}_n$ will depend on the unknown smoothness α. Note however that since $\epsilon_n \geq \bar{\epsilon}_n$ and ϵ_n will obviously be taken to converge to 0, then for large enough n, $c_1 n\bar{\epsilon}_n^2 < n$. This allows the term $\sum_{j=q}^J \mathbb{P}\big(J = j\big)\mathbb{P}\big(m(\boldsymbol{K}_j) < \delta(j)|J = j\big)$ to be absorbed into the remaining terms of the bound on $\mathbb{P}\big(s_{\boldsymbol{\theta},\boldsymbol{K}_J} \notin \mathcal{S}_n\big)$ in the proof. Consequently, as claimed, Theorem 1 also holds if (6') is assumed instead of (6).

4 Implications of the main result

We clarify now the relevance of our result. Consider a family of models $\mathcal{P} = \big\{P_f : f \in \mathcal{F}_\mathcal{A}\big\}$, $\mathcal{F}_\mathcal{A} = \cup_{\alpha \in \mathcal{A}} \mathcal{F}_\alpha$, with densities p_f with respect to some common dominating measure. Assume that we observe a sample $\boldsymbol{X}^{(n)} = (X_1, \ldots, X_n) \sim p_{f_0}^{(n)}$, $X_i \overset{ind}{\sim} p_{f_0}$, $f_0 \in \mathcal{F}_\alpha$ for some unknown smoothness $\alpha \in \mathcal{A}$. The Bayesian approach consists of putting a prior

measure Π on $\mathcal{F} \subseteq \mathcal{F}_{\mathcal{A}}$ which, together with the likelihood $p_f^{(n)}$, leads to the posterior distribution $\Pi(\cdot|\boldsymbol{X}^{(n)})$ via Bayes' formula:

$$\Pi\big(A|\boldsymbol{X}^{(n)}\big) = \frac{\int_A p_f^{(n)}(\boldsymbol{X}^{(n)})\, d\Pi(f)}{\int_{\mathcal{F}} p_f^{(n)}(\boldsymbol{X}^{(n)})\, d\Pi(f)}$$

for a measurable $A \subseteq \mathcal{F}$. The asymptotic behavior of the posterior distribution can be studied from the point of view of the probability measure $P_0 = P_{f_0}$; see Ghosal et al. (2000).

For two densities p_f and p_g with $f, g \in \mathcal{F}_{\mathcal{A}}$, define the (squared) Hellinger metric $h^2(p_f, p_g) = 2\big(1 - \mathbb{E}_g \sqrt{p_f(X)/p_g(X)}\big)$, Kullback-Leibler divergence $K(p_f, p_g) = -\mathbb{E}_g \log\big(p_f(X)/p_g(X)\big)$ and the Csiszár f- divergence $V(p_f, p_g) = \mathbb{E}_g \log^2\big(p_f(X)/p_g(X)\big)$. Define also the ball $B(\epsilon_n, f_0) = \big\{f \in \mathcal{F} : K(f, f_0) \leq \epsilon^2, V(f, f_0) \leq \epsilon^2\big\}$.

The following theorem is the main result of Ghosal et al. (2000) (for a version involving two sequences ϵ_n and $\bar{\epsilon}_n$ cf. also Ghosal and van der Vaart (2001)) which makes a statement about the asymptotic behavior of a posterior measure.

Theorem 2 (Theorem 2.1 of Ghosal et al. (2000))**.** *Suppose that for two positive sequences $\epsilon_n \geq \bar{\epsilon}_n$ such that $n\bar{\epsilon}_n^2 > 1$ and $\epsilon_n \to 0$ as $n \to \infty$, sets $\mathcal{F}_n \subseteq \mathcal{F}$ and constants $c_1, c_2, c_3, c_4 > 0$, the following conditions hold:*

$$\log N\big(\epsilon_n, \mathcal{F}_n, h\big) \leq c_1 n\epsilon_n^2, \tag{17}$$

$$\Pi(\mathcal{F} \backslash \mathcal{F}_n) \leq c_2 e^{-(c_3+4)n\bar{\epsilon}_n^2}, \tag{18}$$

$$\Pi(B(\bar{\epsilon}_n, f_0)) \geq c_4 e^{-c_3 n\bar{\epsilon}_n^2}. \tag{19}$$

Then, for large enough $M > 0$, $\Pi\big(f \in \mathcal{F} : h(p_f, p_{f_0}) \geq M\epsilon_n|\boldsymbol{X}^{(n)}\big) \to 0$ as $n \to \infty$ in P_{f_0}-probability.

The conditions of this theorem require the existence of a *sieve* $\mathcal{F}_n$ with small entropy (17) which contains most of the prior mass (18) and which enough prior mass around the parameter f_0 which indexes the "true" underlying measure of the data. Assume now that the models in $\mathcal{P}$ are such that for d^2 being h^2, K or V, $d^2(p_f, p_{f_0}) \lesssim \|f - f_0\|_2^2$. If in addition one can prove that in the considered model $h(p_f, p_{f_0}) \gtrsim \|f - f_0\|_2$, then Theorem 2 delivers a contraction rate ϵ_n with respect to the L_2-distance as well. Some examples of models for which the above relations between norms can be established are, among others, density estimation, non-parametric regression, binary regression, Poisson regression and classification; cf. Ghosal et al. (2000), de Jonge and van Zanten (2012), Shen and Ghosal (2012). In this case one can apply our meta-theorem (Theorem 1) to obtain an adaptive contraction rate which essentially verifies (17)–(19) for our spline-based prior. We summarize this in the following theorem.

Theorem 3. *Let Π be the spline prior described in Section 3. Consider a family of models $\mathcal{P} = \big\{P_f : f \in \mathcal{F}_{\mathcal{A}}\big\}$, $\mathcal{F}_{\mathcal{A}} = \cup_{\alpha \in \mathcal{A}} \mathcal{F}_\alpha$, with densities p_f with respect to some common dominating measure. Assume also that the models in $\mathcal{P}$ are such that for d^2*

being h^2, K or V, $d^2(p_f, p_{f_0}) \lesssim \|f - f_0\|_2^2$. Take an i.i.d. sample $\boldsymbol{X}^{(n)} = (X_1, \ldots, X_n)$, $X_i \sim p_{f_0}$, $f_0 \in \mathcal{F}_\alpha$, $\|f_0\|_\infty < M$, for some unknown smoothness $\alpha \in \mathcal{A}$. Consider a prior Π which verifies (4) through (8) for certain constants c_1, c_2, c_3, t_1, t_2 and t_3. Assume also that either $\alpha \leq 1$ or $t_2 \wedge t_3 = 1$.

Then, for large enough $C > 0$, $\Pi\big(f \in \mathcal{F} : h(p_f, p_{f_0}) \geq C\epsilon_n | \boldsymbol{X}^{(n)}\big) \to 0$ as $n \to \infty$ in P_0-probability for $\epsilon_n = C_3 n^{-\alpha/(2\alpha+1)} (\log n)^{\alpha/(2\alpha+1)+(1-(t_1 \wedge t_3))/2}$. If $h(p_f, p_{f_0}) \gtrsim \|f - f_0\|_2$ then in the previous statement the Hellinger distance may be replaced by the L_2 distance and the statement remains valid.

Proof. We have that for some constant $k > 0$ and $\mathcal{F} = \mathcal{S}$, $\mathcal{F}_n = \mathcal{S}_n$,

$$\begin{aligned} N\big(\epsilon_n, \mathcal{F}_n, h\big) &\leq N(\epsilon_n/k, \mathcal{F}_n, \|\cdot\|_2), \\ \Pi(\mathcal{F} \backslash \mathcal{F}_n) &= P\big(s_{\boldsymbol{\theta}, \boldsymbol{K}_J} \notin \mathcal{F}_n\big), \\ \Pi(B(\bar{\epsilon}_n, f_0)) &\geq P\big(\|s_{\boldsymbol{\theta}, \boldsymbol{K}_J} - f_0\|_\infty \leq \bar{\epsilon}_n/k\big). \end{aligned}$$

The first inequality follows from the fact that by assumption $h(p_f, p_g) \leq k\|f - g\|_2$ and so an ϵ/k cover of $\mathcal{F}_n$ according to $\|\cdot\|_2$ induces an ϵ cover of $\mathcal{F}_n$ according to h. Then, since for d^2 being K or V, $d^2(p_f, p_{f_0}) \leq k\|f - f_0\|_2^2$, we have $B(\bar{\epsilon}_n, f_0) \supset \big\{f \in \mathcal{F} : \|f - f_0\|_2 \leq \epsilon/k\big\}$ and the last inequality follows.

By assumption $f_0 \in \mathcal{F}_\alpha$ satisfies the conditions of Theorem 1; assume (3) holds for some C_{f_0}. Consider then a prior that satisfies (4)–(8). Let us present a choice of quantities M_n, $\delta(j)$, J_n, $\bar{J}_n$, ϵ_n and $\bar{\epsilon}_n$ which meet conditions (9)–(11). First of all, sequence M_n can be taken as a polynomial in n (for instance, for normal or exponential conditional priors for $\boldsymbol{\theta} \in \mathbb{R}^j$ in (10)) and $1/\delta(j)$ as a polynomial in j. Next, note that there is no $\bar{J}_n$ that satisfies (11) unless $\alpha \leq 1$ or $t_2 \wedge t_3 = 1$. If either $\alpha > 1$ or $t_2 \wedge t_3 < 1$, then the best possible choices are $\bar{J}_n = \tau C_{f_0}^{1/\alpha} (\bar{\epsilon}_n)^{-1/\alpha}$, $\bar{\epsilon}_n = C_1 (\log n/n)^{\alpha/(2\alpha+1)}$ for sufficiently large C_1, $J_n = C_2 n^{1/(2\alpha+1)} (\log n)^{2\alpha/(2\alpha+1)-t_1}$ for sufficiently large C_2, and finally,

$$\epsilon_n = C_3 n^{-\alpha/(2\alpha+1)} (\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$$

for sufficiently large C_3. Since these quantities satisfy (9)–(11), Theorem 1 implies conditions (12)–(14) for the quantities defined above. Finally, applying Theorem 2, we conclude that the contraction rate of the resulting posterior is at most ϵ_n, which appears to be optimal (up to a logarithmic factor) in a minimax sense over the Hölder class $\mathcal{H}_\alpha$ (also over α-smooth Sobolev class).

□

Remark 5. A priori, it may be unknown whether $\alpha > 1$ or not, or it may be simply known that $\alpha \leq 1$. We can however always ensure the condition $t_2 \wedge t_3 < 1$ by an appropriate choice of prior. For example, we take a geometric prior on J so that $t_2 = 0$ and a prior on $\boldsymbol{K}_j$ such that (20) (which implies (7)) holds with, say, $t_3 = 0$.

Remark 6. The common practice, in applications, of endowing the location of the knots with a Poisson point process prior results in a prior that does not verify assumption (6). Assumption (6'), however, permits this so long as a large enough point mass is placed at an equally spaced knot vector. This very simple modification assures that our Theorem 3 may be applied to show that these priors result in a rate adaptive posteriors.

5 Examples of Priors

We give now examples of particular choices for the several components of our hierarchical prior which verify conditions (4) through (8) and (6').

As for the prior on the number of basis functions, assumptions (4) and (5) hold for the geometric, Poisson and negative binomial distributions; cf. Shen and Ghosal (2012). Assumption (8), on the other hand, will trivially hold if we assume, for example, the coordinates of $\boldsymbol{\theta} \in \mathbb{R}^j$ to be (conditionally on $J = j$) independent and identically distributed according to a density uniformly bounded away from zero on the interval $[-M, M]$.

There is an ample choice of priors on $\boldsymbol{K}_J$, given $J = j$, which satisfy condition (6). First note that this condition enforces the prior on the location of the knots, for each $J = j$, to be such that, with probability 1, adjacent knots are at least $\delta(j)$ apart. The function $1/\delta(j)$ can be taken as a polynomial in j of high degree which makes the requirement less restrictive. If a certain sequence ϵ_n verifies the conditions of Theorem 1 then an increase in the exponent of $1/\delta(j)$ can be accommodated by making ϵ_n larger by a multiplicative factor (cf. condition (9).)

A simple choice for the prior on $\boldsymbol{K}_J$, given $J = j$, is to pick $(j - q)$ knots uniformly at random, without replacement, on a uniform $\delta(j)$-sparse grid. This construction is possible if δ is chosen in such a way that $\lfloor 1/\delta(j) \rfloor > j - q$ for all j. Another example is to take, for each j, the $(j - q)$ inner knots in $\boldsymbol{K}_j$ to be generated sequentially in the following way: add a knot K_1 uniformly at random on the interval $[\delta(j), 1 - \delta(j)]$, then a knot K_2 uniformly at random on the interval $[\delta(j), 1 - \delta(j)] \backslash (K_1 - \delta(j), K_1 + \delta(j))$ and so on. Finally, take the ordered $\boldsymbol{K}_j = (K_{(1)}, \ldots, K_{(j-q)})$. This construction is always possible if $1/\delta(j)$ grows faster than $2(j - q + 1)$. (If J is Poisson distributed, these points are simply distributed like a homogeneous Poisson process, conditioned to have all points at least $\delta(J)$ apart.) Note that for this construction, the probability $\mathbb{P}\big(m(\boldsymbol{K}_j) > \delta(j) | J = j\big)$ is at least $(1 - 2(j - q)\delta(j))^{j-q}$ which is very close to one if j is large and $1/\delta(j)$ is a large power of j, say. Clearly, condition (6) is satisfied for these two constructions since all prior mass is concentrated on partitions with sparseness larger than $\delta(j)$.

It is also easy to see that condition (7) is verified for the knot vectors obtained from one of these two constructions. In fact, condition (7) is trivially fulfilled if, for some $0 \leq t_3 < 1$,

$$\mathbb{P}(\boldsymbol{K}_j = \bar{\boldsymbol{k}}_j) \gtrsim \exp\big(- c_3 j \log^{t_3} j \big), \tag{20}$$

where $\bar{\boldsymbol{k}}_j \in \mathcal{K}_j$ is the set of $(j - q)$ equally spaced inner knots. This suggests a mechanism to assure that any prior which verifies (6) can be slightly modified to also verify (7): given $J = j$, generate a Bernoulli random variable X with success probability, say, $\exp(-c_3 j \log^{t_3} j)$; if $X = 1$, then take $\boldsymbol{K}_j = \bar{\boldsymbol{k}}_j$, otherwise pick the knots in $\boldsymbol{K}_j$ according to any procedure which verifies (6), for instance one of two procedures described above. The resulting prior will trivially satisfy both (6) and (7).

Condition (6) necessarily excludes some partitions from the support of the prior (and then also from the support of the posterior.) As mentioned before very few partitions will be excluded so long as $1/\delta(j)$ is a large enough power of j. It is nonetheless of interest to design a weaker alternative for condition (6). Condition (6′) plays this role, in that

it allows priors on $\boldsymbol{K}$ which have *any* partition of $[0,1]$ into non-empty intervals in its support.

Assuming condition (6') instead of (6) consequently allows us to put positive mass on any vector of simple knots in a straightforward way: generate a Bernoulli random variable with success probability $1 - c_5 \exp(-c_4 n)$; if $X = 1$ take $\boldsymbol{K}_j = \bar{\boldsymbol{k}}_j$, equally spaced; if $X = 0$ then take an arbitrary $\boldsymbol{K}_j$ (for example independent, uniformly distributed points on $[0,1]$.) So long as we take $1/\delta(j) = j$ and $\tau \geq q$ then conditions (6') and (7) are verified. This procedure, although simpler, does place little prior mass on knot vectors with inhomogeneous distributions.

An alternative, less degenerate prior, which verifies (6') and (7) can be obtained in the following way: given $J = j$, first, generate a Bernoulli random variable X_1 with success probability $c_5 \exp(-c_4 n)$; if $X_1 = 1$ distribute the $(j - q)$ knots arbitrarily; if $X_1 = 0$ then generate another Bernoulli random variable X_2 with success probability, $\exp(-j)$; if $X_2 = 1$ then take $(j - q)$ equally spaced knots $\bar{\boldsymbol{k}}_j$; If $X_2 = 0$, then place the knots such that (6) is verified. This procedure should allow good control on the prior on the knots while not excluding any knot vectors.

Note that the priors described above which verify (4) through (8) do not depend on the sample size n, as prescribed by the Bayesian paradigm. Condition (6') is a weaker requirement then condition (6) but it will, introduce a dependence on the sample size n in the prior.

6 Technical results

In this section we collect some technical results. Lemmas 1 and 2 are needed to bound the entropy number of the sieves $\mathcal{S}_n$ in Theorem 1. Lemma 3 claims in essence that if some bounds on the range of the function f_0 are known, then this knowledge can be incorporated into the prior on the coefficients $\boldsymbol{\theta}$.

Theorem 4.26 of Schumaker (2007) claims that if all the inner knots of a B-spline are simple, then the B-spline is continuous, uniformly over its support, with respect to its knots. In Lemma 2 we establish a slightly stronger result (a Lipschitz-type property): if we take two splines with the same coefficients in their respective B-spline basis, then the L_∞ distance between the splines can be bounded by a multiple of the l_∞ distance between the two sets of knots, as long as the sets of knots are sufficiently sparse. First, we present a preliminary lemma. Denote the $(r+1)$-th order divided difference of a function h over the points $t_1, \ldots, t_{r+1}$ as $[t_1, \ldots, t_{r+1}]h = ([t_2, \ldots, t_{r+1}]h - [t_1, \ldots, t_r]h)/(t_{r+1} - t_1)$, with $[t_i]h = h(t_i)$. If $t_1 = \cdots = t_{r+1}$ then $[t_1, \ldots, t_{r+1}]h = h^{(r)}(t_1)/r!$ for a function h with enough derivatives at t_1.

Lemma 1. *Let $i \in \{1, \ldots, r\}$, $r \geq 2$, $(k_1, \ldots, k_{r+1}) \in (0,1)^{r+1}$. Assume $k_{v+1} - k_v > \delta > 0$ for $v = 0, \ldots, i-1, i+1, \ldots, r$ and $k_{i+1} - k_i = 0$. For fixed $x \in [0,1]$ take the function $h(y) = (x-y)_+^{q-1}$ with $y \in [0,1]$ and $q \geq 2$. Then the divided difference $\left|[k_1, \ldots, k_{r+1}]h\right| \leq 4/\delta^r$ for $x \neq k_i$.*

Proof. Notice that $|h'(y)| = (q-1)(x-y)_+^{q-2} \leq (q-1) \leq 1/\delta$ for $x \neq y$, as $q \geq 2$ and thus $\delta < k_2 - k_1 < 1 \leq \frac{1}{q-1}$. Next, if $v = i-1$, $\left|[k_{v+1}, k_{v+2}]h\right| = |h'(k_{v+1})| \leq 1/\delta$;

if $v \neq i-1$, $\big|[k_{v+1}, k_{v+2}]h\big| = |h(k_{v+2}) - h(k_{v+1})|/|k_{v+2} - k_{v+1}| \leq 2/\delta$. We conclude $\big|[k_{v+1}, k_{v+2}]h\big| \leq 2/\delta$ as long as $x \neq k_i$.

For $j = 2, \ldots, r$, define $\gamma_j = \min_{v=1,\ldots,r+1-j} |k_{v+j} - k_v| \geq (j-1)\delta$. Now we make use of Theorem 2.56 from Schumaker (2007) and the previous bound:

$$\big|[k_1, \ldots, k_{r+1}]h\big| \leq \sum_{v=0}^{r-1} \binom{r-1}{v} \frac{\big|[k_{v+1}, k_{v+2}]h\big|}{\gamma_2 \ldots \gamma_r} \leq \frac{2^r}{(r-1)!\delta^r} \leq \frac{4}{\delta^r}$$

holds for all $x \neq k_i$. This completes the proof of the Lemma.

□

Lemma 2. *Let $\boldsymbol{\theta} \in \mathbb{R}^j$ satisfies $\|\boldsymbol{\theta}\|_\infty \leq M$ and let $\boldsymbol{k}, \boldsymbol{k}' \in \mathcal{K}_j^\delta = \{\boldsymbol{k} \in \mathcal{K}_j : m(\boldsymbol{k}) > \delta\}$ be such that $\|\boldsymbol{k} - \boldsymbol{k}'\|_\infty \leq \delta$. Then $\|s_{\boldsymbol{\theta},\boldsymbol{k}} - s_{\boldsymbol{\theta},\boldsymbol{k}'}\|_\infty \leq L\|\boldsymbol{k} - \boldsymbol{k}'\|_\infty$, for $L = 4j(q+1)M\delta^{-(q+1)}$.*

Proof. Define $\boldsymbol{k}^l = (k_1^l, \ldots, k_{j-q}^l) = (k_1', \ldots, k_l', k_{l+1}, \ldots, k_{j-q})$ for $l = 0, \ldots, j-q$, such that $\boldsymbol{k}^0 = \boldsymbol{k}$ and $\boldsymbol{k}^{j-q} = \boldsymbol{k}'$. We get

$$\begin{aligned}
\big\|s_{\boldsymbol{\theta},\boldsymbol{k}} - s_{\boldsymbol{\theta},\boldsymbol{k}'}\big\|_\infty &= \Big\| \sum_{i=1}^{j} \theta_i B_i^{\boldsymbol{k}^0} - \sum_{i=1}^{j} \theta_i B_i^{\boldsymbol{k}^{j-q}} \Big\|_\infty \leq M \Big\| \sum_{i=1}^{j} (B_i^{\boldsymbol{k}^0} - B_i^{\boldsymbol{k}^{j-q}}) \Big\|_\infty \\
&\leq jM \max_{1\leq i\leq j} \big\|B_i^{\boldsymbol{k}^0} - B_i^{\boldsymbol{k}^{j-q}}\big\|_\infty \leq jM \max_{1\leq i\leq j} \sum_{l=0}^{j-q-1} \big\|B_i^{\boldsymbol{k}^l} - B_i^{\boldsymbol{k}^{l+1}}\big\|_\infty \\
&\leq (q+1)jM \max_{1\leq i\leq j} \max_{0\leq l\leq j-q-1} \big\|B_i^{\boldsymbol{k}^l} - B_i^{\boldsymbol{k}^{l+1}}\big\|_\infty,
\end{aligned}$$

The last inequality follows from (1) and the fact that the inner knots of $B_i^{\boldsymbol{k}^l}$ and $B_i^{\boldsymbol{k}^{l+1}}$ differ only at the $(l+1)$-th entry.

Theorem 4.27 of Schumaker (2007) gives explicit expressions for the derivative of a B-spline with respect to one of its knots. These expressions are in terms of the divided differences which satisfy the conditions of Lemma 1, so that combining this with Lemma 1 for $r = q+1$ (the maximal number of knots in the support of a B-spline) yields that this derivative is bounded in absolute value by $4\delta^{-(q+1)}$, except at $x = k_{l+1}^l$, where it is not defined. Then, as $\|\boldsymbol{k}^l - \boldsymbol{k}^{l+1}\|_\infty \leq \|\boldsymbol{k} - \boldsymbol{k}'\|_\infty$, we obtain that, for $x \neq k_{l+1}^l$, $l = 0, \ldots, j-q-1$,

$$\big|B_i^{\boldsymbol{k}^l}(x) - B_i^{\boldsymbol{k}^{l+1}}(x)\big| \leq |k_{l+1}^{l+1} - k_{l+1}^l| \sup_{k_{l+1}^l \in (0,1)} \Big|\frac{\partial B_i^{\boldsymbol{k}^l}(x)}{\partial k_{l+1}^l}\Big| \leq \frac{4\|\boldsymbol{k} - \boldsymbol{k}'\|_\infty}{\delta^{q+1}}.$$

Since splines are continuous for all $q > 1$, so is $s_{\boldsymbol{\theta},\boldsymbol{k}} - s_{\boldsymbol{\theta},\boldsymbol{k}'}$ and we conclude that the same bound must also hold for $x = k_{l+1}^l$. Combining the above two relations concludes the proof.

□

The properties of B-splines allow to relate the range of the coefficients of the approximating spline to the range of the approximated function. The following lemma generalizes Lemma 1 of Shen and Ghosal (2012) for non-equally spaced knots.

Lemma 3. *Let $f \in \mathcal{F}^\alpha$ (so that (3) holds), $a < b$, $\varepsilon > 0$. Assume that $f(x) \in [a+\varepsilon, b-\varepsilon]$ for all $x \in [0,1]$. Then there exits a positive constant $\delta = \delta(\mathcal{F}^\alpha, \varepsilon)$ such that for any $\boldsymbol{k} \in \mathcal{K}_j$, $j \geq q$, such that $M(\boldsymbol{k}) \leq \delta$, the coefficients $\boldsymbol{a}$ of the approximating spline $s_{\boldsymbol{a},\boldsymbol{k}}$ in (3) can be taken to be contained in (a,b).*

Proof. Fix q, j and inner knots $\boldsymbol{k}$, assume $I = [a,b]$, $a < b$ and $a+\varepsilon < f < b-\varepsilon$, for some $\varepsilon > 0$.

We use results from section 4.6 of Schumaker (2007) on dual basis of B-splines. If $B_1^{\boldsymbol{k}}, \ldots, B_j^{\boldsymbol{k}}$ is the B-spline basis associated with the inner knots $\boldsymbol{k}$, then there exists a dual basis $\lambda_1, \ldots, \lambda_j$ of linear functionals such that, for each $i, r = 1, \ldots, j$, $\lambda_r B_i^{\boldsymbol{k}} = 1$ if $i = r$ and is 0 otherwise. As a consequence, we obtain that $\lambda_i s_{\boldsymbol{a},\boldsymbol{k}} = a_i$, and since $\sum_{i=1}^j B_i^{\boldsymbol{k}}(x) = 1$, it follows that $\lambda_i c = c$ for any constant c and all $i = 1, \ldots, j$. This dual basis is not necessarily unique and, according to Theorem 4.41 from Schumaker (2007), can be taken such that $|\lambda_i f| \leq C_1 \sup_{x \in I_i} |f(x)|$ where I_i represents the support of $B_i^{\boldsymbol{k}}$ and constant C_1 depends only on q. Each I_i consists of at most q adjacent intervals in the partition induced by $\boldsymbol{k}$ and thus the length of I_i is bounded by $qM(\boldsymbol{k})$.

Let $s_{\boldsymbol{a},\boldsymbol{k}}$ be such that (3) is fulfilled for f. Then for any constant c

$$\begin{aligned} |a_i - c| &= \left|\lambda_i s_{\boldsymbol{a},\boldsymbol{k}} - \lambda_i f + \lambda_i f - c\right| \leq \left|\lambda_i (s_{\boldsymbol{a},\boldsymbol{k}} - f)\right| + \left|\lambda_i (f - c)\right| \\ &\leq C_1 C_f M^\alpha(\boldsymbol{k}) + C_1 \sup_{x \in I_i} |f(x) - c|. \end{aligned}$$

Take $c = \inf_{x \in I_i} f(x)$ and recall that $f \in \mathcal{F}_\alpha \subseteq \mathcal{L}(\kappa_\alpha, L_\alpha)$. Using the Lipschitz property, we derive that $\sup_{x \in I_i} |f(x) - c| = \sup_{x \in I_i} f(x) - \inf_{x \in I_i} f(x) \leq L_\alpha (q\, M(\boldsymbol{k}))^{\kappa_\alpha}$ and therefore

$$|a_i - \inf_{x \in I_i} f(x)| \leq C_1 C_f M^\alpha(\boldsymbol{k}) + C_1 L_\alpha (q\, M(\boldsymbol{k}))^{\kappa_\alpha} \leq C_2 M^{\alpha \wedge \kappa_\alpha}(\boldsymbol{k}).$$

In the same way, if we take $c = \sup_{x \in I_i} f(x)$, we derive that $\sup_{x \in I_i} |f(x) - c| \leq L_\alpha (q\, M(\boldsymbol{k}))^{\kappa_\alpha}$ and thus $\left|a_i - \sup_{x \in I_i} f(x)\right| \leq C_2 M^{\alpha \wedge \kappa_\alpha}(\boldsymbol{k})$.

Now for $\delta = (\varepsilon / (2C_2))^{1/(\alpha \wedge \kappa_\alpha)}$ conclude that if $M(\boldsymbol{k}) \leq \delta$, then $a_i \geq \inf_{x \in I_i} f(x) - C_2 M^{\alpha \wedge \kappa_\alpha}(\boldsymbol{k}) \geq \inf_{x \in I_i} f(x) - \varepsilon/2 > a$. For the same choice of δ we have $a_i \leq \sup_{x \in I_i} f(x) + C_2 M^{\alpha \wedge \kappa_\alpha}(\boldsymbol{k}) \leq \sup_{x \in I_i} f(x) + \varepsilon/2 < b$.

□

References

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

de Boor, C. (1978). *A practical guide to splines*. Springer-Verlag, New York.

de Jonge, R. and van Zanten, H. (2012). "Adaptive estimation of multivariate functions using conditionally Gaussian tensr-product spline priors." *pre-print*.

Denison, D., Mallick, B., and Smith, A. (1998). "Bayesian MARS." *Statistics and Computing*, 8: 337–346.

Di Matteo, I., Genovese, C., and Kaas, R. (2001). "Bayesian curve-fitting with free-knot splines." *Biometrika*, 88(4): 1055–1071.

Ghosal, S., Ghosh, J., and van der Vaart, A. (2000). "Convergence rates of posterior distributions." *Ann. Statist.*, 28(2): 500–531.

Ghosal, S. and van der Vaart, A. (2001). "Entropies and Rates of Convergence for Maximum Likelihood and Bayes Estimation for Mixtures of Normal Densities." *Ann. Statist.*, 29(5): 1233–1263.

Schumaker, L. (2007). *Spline functions: basic theory*. John Wiley & Sons, New York.

Sharef, E., Strawderman, R., Ruppert, D., Cowen, M., and Halasyamani, L. (2010). "Bayesian adaptive B-spline estimation in proportional frailty models." *Electron. J. Stat.*, 4: 606–642.

Shen, W. and Ghosal, S. (2012). "MCMC-free adaptive Bayesian procedures using random series prior." *pre-print*.

van der Vaart, A. and van Zanten, H. (2008). "Rates of contraction of posterior distributions based on Gaussian process priors." *Ann. Statist.*, 36(3): 1435–1463.

van der Vaart, A. W. and van Zanten, J. H. (2009). "Adaptive Bayesian Estimation Using A Gaussian Random Field With Inverse Gamma Bandwidth." *Ann. Statist.*, 37(5B): 2655–2675.