Journal of Machine Learning Research 24 (2023) 1-49 Submitted 5/21; Revised 8/22; Published 2/23

Posterior Contraction for Deep Gaussian Process Priors

Gianluca Finocchio GIANLUCA.FINOCCHIO@UNIVIE.AC.AT
Faculty of Business, Economics and Statistics
University of Vienna
1090 Vienna, Austria

Johannes Schmidt-Hieber A.J.SCHMIDT-HIEBER@UTWENTE.NL
Faculty of Electrical Engineering, Mathematics and Computer Science
University of Twente
7522 NB Enschede, The Netherlands

Editor: Mingyuan Zhou

Abstract

We study posterior contraction rates for a class of deep Gaussian process priors in the nonparametric regression setting under a general composition assumption on the regression function. It is shown that the contraction rates can achieve the minimax convergence rate (up to $\log n$ factors), while being adaptive to the underlying structure and smoothness of the target function. The proposed framework extends the Bayesian nonparametric theory for Gaussian process priors.

Keywords: Bayesian nonparametric regression, contraction rates, deep Gaussian processes, uncertainty quantification, neural networks

1. Introduction

In the multivariate nonparametric regression model with random design distribution μ supported on $[-1,1]^d$, we observe n i.i.d. pairs $(\mathbf{X}_i, Y_i) \in [-1,1]^d \times \mathbb{R}$, $i = 1, \ldots, n$, with $\mathbf{X}_i \sim \mu$,

$$Y_i = f^*(\mathbf{X}_i) + \varepsilon_i, \quad i = 1, \ldots, n \tag{1}$$

and ε_i independent standard normal random variables that are independent of the design vectors $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$. We aim to recover the true regression function $f^* : [-1,1]^d \to \mathbb{R}$ from the sample. Here it is assumed that the true regression function itself is a composition of a number of unknown simpler functions. This comprises several important cases including generalized additive models. As recently proved by Giordano et al. (2022), Gaussian process priors are suboptimal for learning such regression functions. Meanwhile, Schmidt-Hieber (2020) proved that sparsely connected deep neural networks are able to pick up the underlying composition structure and achieve nearly-minimax $L^2(\mu)$-estimation rates over a compositional function class.

Deep Gaussian process priors (DGPs), as conceived by Neal (1996); Damianou and Lawrence (2013); Cutajar et al. (2017), can be viewed as a Bayesian analogue of deep networks. While deep nets are built on a hierarchy of individual network layers, DGPs are based on iterations of Gaussian processes. Compared to neural networks, DGPs have moreover the advantage that the posterior can be used for uncertainty quantification. This makes them potentially attractive for AI applications with strict safety requirements, such as automated driving and health.

In Bayesian nonparametric regression without compositional constraints, Gaussian process priors are a natural choice and a comprehensive literature is available, see for instance the book (Chapter 11) by Ghosal and van der Vaart (2017). The seminal contribution of van der Vaart and van Zanten (2008a) fully characterizes the contraction rate of the posterior distribution for a Gaussian process prior by the small-ball probabilities of the prior and a property of the associated reproducing kernel Hilbert space (RKHS). From this, one can deduce that the best possible rates are achieved when the smoothness of the prior matches that of the unknown regression function. Adaptation to the smoothness can be achieved via rescaled smooth Gaussian processes and an inverse gamma hyperprior on the rescaling parameter as shown by van der Vaart and van Zanten (2009).

In this work we extend the theory of Gaussian process priors to derive posterior contraction rates for DGPs to learn compositional functions. Inspired by model selection priors, we propose a hierarchical prior construction. First, we assign prior weights to possible composition structures. Then, given a composition structure, the final DGP prior puts suitable Gaussian processes on all functions in this model.

The main contribution of this work is to provide an explicit construction for a large class of deep Gaussian process priors and to characterize the corresponding posterior contraction rates. By controlling the composition of Gaussian processes, we obtain tight bounds for the entropy and the decentered small-ball probabilities of DGPs. We also provide examples of DGP priors inducing nearly-minimax posterior contraction rates. In particular, if there is some low-dimensional structure in the composition, the posterior will not suffer from the curse of dimensionality.

Stabilization enhancing methods such as dropout and batch normalization are crucial for the performance of deep learning. In particular, batch normalization guarantees that the signal sent through a trained network cannot explode. We argue that for deep Gaussian processes similar effects play a role. In Figure 1 below we visualize the effect of composing independent copies of a Gaussian process, the resulting trajectories are rougher and more versatile than those generated by the original process alone. This may, however, lead to wild behavior of the sample paths. As we aim for a fully Bayesian approach, the only possibility is to induce stability through the selection of the prior. We enforce stability by conditioning each individual Gaussian process to lie in a set of 'stable' paths. To achieve nearly-optimal contraction rates, these sets have to be carefully selected and depend on the optimal contraction rate itself.

The mathematical analysis requires to answer various questions involving the composition of GPs. To our knowledge, the closest results in the literature are bounds on the centred small-ball probabilities of iterated processes. They have been obtained for self-similar processes by Aurzada and Lifshits (2009) and for time-changed self-similar processes by Kobayashi (2016). A good reference on the literature of iterated processes is given by Arendarczyk (2017). In a different line of research, Funaki (1979) studies iterated Brownian motions (IBMs) as solutions of high-order parabolic stochastic differential equations (SDEs), whereas Burdzy (1993) focuses on their path properties. The composition of general processes in relation with high-order parabolic and hyperbolic SDEs has been studied by

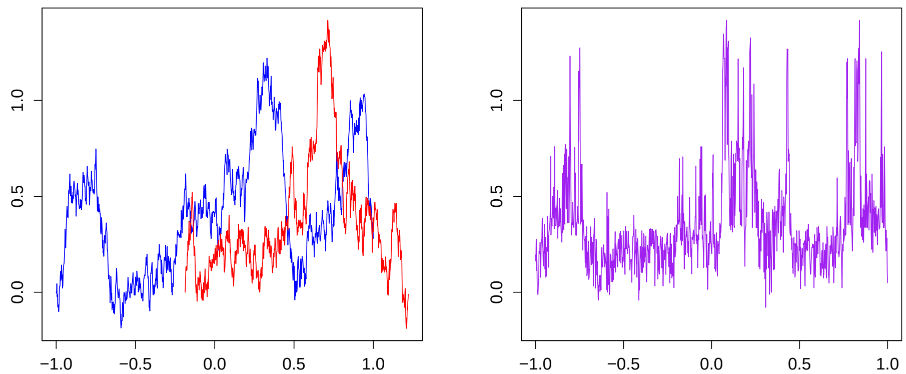

Figure 1: Composition of Gaussian processes results in rougher and more versatile sample paths. On the left: trajectories of two independent copies of a standard Brownian motion. On the right: the composition (red $\circ$ blue) of the trajectories.

Hochberg and Orsingher (1996). More recently, the infinite iteration of Brownian motions has been studied by Curien and Konstantopoulos (2014); Casse and Marckert (2016).

The article is structured as follows. In Section 2 we formalize the model and give an explicit parametrization of the underlying graph and the smoothness index. Section 3 provides a detailed construction of the deep Gaussian process prior. In Section 4 we state the main posterior contraction results. In Section 5 we present a construction achieving optimal contraction rates and provide explicit examples in Section 6. Section 7 compares Bayes with DGPs and deep learning. It also contains a subsection on computational challenges. All proofs are deferred to the appendix.

Notation: Vectors are denoted by bold letters, e.g. $\mathbf{x} := (x_1, \dots, x_d)^\top$. For $S \subseteq \{1, \dots, d\}$, we write $\mathbf{x}_S = (x_i)_{i \in S}$ and $|S|$ for the cardinality of S. As usual, we define $|\mathbf{x}|_p := (\sum_{i=1}^d |\mathbf{x}_i|^p)^{1/p}$, $|\mathbf{x}|_\infty := \max_i |\mathbf{x}_i|$, $|\mathbf{x}|_0 := \sum_{i=1}^d \mathbf{1}(\mathbf{x}_i \neq 0)$, and write $\|f\|_{L^p(D)}$ for the L^p norm of f on D. If there is no ambiguity concerning the domain D, we also write $\|\cdot\|_p$. For two sequences $(a_n)_n$ and $(b_n)_n$ we write $a_n \lesssim b_n$ if there exists a constant C such that $a_n \leq C b_n$ for all n. Moreover, $a_n \asymp b_n$ means that $(a_n)_n \lesssim (b_n)_n$ and $(b_n)_n \lesssim (a_n)_n$. For positive sequences $(a_n)_n$ and $(b_n)_n$ we write $a_n \ll b_n$ if a_n/b_n tends to zero when n tends to infinity.

2. Composition Structure on the Regression Function

In this section, we introduce the compositional class for the regression function f^* in the nonparametric model (1). As we are interested in prediction, the aim is to learn the true regression function, but not its underlying compositional structure.

Consider the class of functions f which can be written as the composition of $q+1$ functions, that is, $f = g_q \circ g_{q-1} \circ \ldots \circ g_1 \circ g_0$, for functions $g_i : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}}$ with $d_0 = d$ and $d_{q+1} = 1$. If f takes values in the interval $[-1, 1]$, rescaling $h_i = g_i(\|g_{i-1}\|_\infty \cdot)/\|g_i\|_\infty$ with $\|g_{-1}\|_\infty := 1$ leads to the alternative representation

$$f = h_q \circ h_{q-1} \circ \ldots \circ h_1 \circ h_0 \tag{2}$$

for functions $h_i : [-1,1]^{d_i} \to [-1,1]^{d_{i+1}}$. We also write $h_i = (h_{ij})^\top_{j=1,\ldots,d_{i+1}}$, with $h_{ij} : [-1,1]^{d_i} \to [-1,1]$. The representation can be modified if f takes values outside $[-1,1]$, but to avoid unnecessary technical complications, we do not consider this case here. Although the function h_i in the representation of f is defined on $[-1,1]^{d_i}$, we allow each component function h_{ij} to possibly only depend on a subset of t_i 'active' variables $\mathcal{S}_{ij} \subseteq \{1,\ldots,d_i\}$ for some $t_i \le d_i$. Writing for a subset of indices S, $(\cdot)_S : \mathbf{x} \mapsto \mathbf{x}_S = (x_i)_{i\in S}$, define

$$\overline{h}_{ij} : [-1,1]^{t_i} \to [-1,1], \quad \mathbf{x}_{\mathcal{S}_{ij}} \mapsto h_{ij}(\mathbf{x}_{\mathcal{S}_{ij}}, \mathbf{x}_{\mathcal{S}_{ij}^c}). \tag{3}$$

As h_{ij} does not depend on $\mathbf{x}_{\mathcal{S}_{ij}^c}$, this function is well-defined.

To define suitable function classes and priors on composition functions, it is natural to first associate to each composition structure a directed graph. The nodes in the graph are arranged in $q+2$ layers with $q+1$ the number of components in (2). The number of nodes in each layer is given by the integer vector $\mathbf{d} := (d, d_1, \ldots, d_q, 1) \in \mathbb{N}^{q+2}$ storing the dimensions of the components h_i appearing in (2). In the graph, we draw an edge between the j-th node in the $i+1$-st layer and the k-th node in the i-th layer if and only if $k \in \mathcal{S}_{ij}$. The number t_i is the in-degree of the nodes in layer $i+1$.

For any i, the subsets corresponding to different nodes $j = 1,\ldots,d_{i+1}$, are combined into $\mathcal{S}_i := (\mathcal{S}_{i1}, \ldots, \mathcal{S}_{id_{i+1}})$ and $\mathcal{S} := (\mathcal{S}_0, \ldots, \mathcal{S}_q)$. Setting $\mathbf{t} := (t_0, \ldots, t_q)$, we summarize the previous quantities into the hyper-parameter

$$\lambda := (q, \mathbf{d}, \mathbf{t}, \mathcal{S}), \tag{4}$$

which we refer to as the graph of the function f in (2). The set of all possible graphs is denoted by Λ.

As an example consider the function $f(x_1,\ldots,x_5) = h_1(h_{01}(x_1,x_3,x_4), h_{02}(x_1,x_4,x_5), h_{03}(x_2,x_4,x_5))$ with corresponding graph representation displayed in Figure 2. This function is the composition of two layers and takes a five-dimensional input, while each of its components never involves more than three variables at once (the in-degree of all nodes is 3). In this case, we have $q = 1$, $d_0 = 5, d_1 = 3, d_2 = 1$ and $t_0 = t_1 = 3$. The active sets are $\mathcal{S}_{01} = \{1,3,4\}, \mathcal{S}_{02} = \{1,4,5\}, \mathcal{S}_{03} = \{2,4,5\}$, and $\mathcal{S}_{11} = \{1,2,3\}$. Another example shown in Figure 2 are generalized additive models (GAMs), where the functions are of the form $f(x_1,\ldots,x_d) = h_{21}(\sum_{i=1}^d h_{0i}(x_i))$. The first layer of the graph computes the function $x_i \mapsto h_{0i}(x_i)$, the second layer sums these functions, and the last layer finally computes the GAM $h_{21}(\sum_{i=1}^d h_{0i}(x_i))$. In this case we have $q = 2$, $\mathbf{d} = (d,d,1,1)$, $\mathbf{t} = (1,d,1)$, $\mathcal{S}_{0i} = \{i\}$, for $i = 1,\ldots,d$, $\mathcal{S}_{11} = \{1,\ldots,d\}$, and $\mathcal{S}_{21} = \{1\}$.

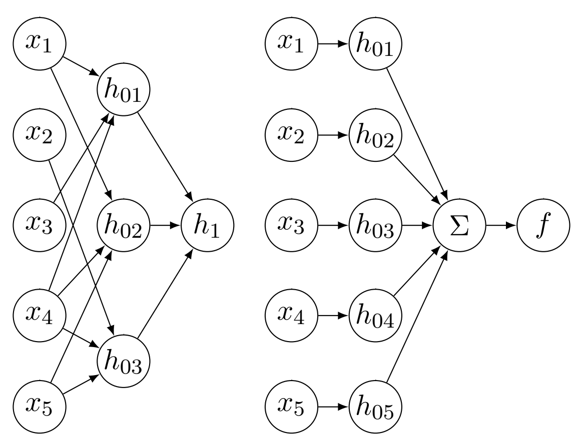

Figure 2: Graph representation of the example function (left) and the generalized additive model (right).

We impose smoothness conditions on all functions in the composition. A function has Hölder smoothness index $\beta > 0$ if all partial derivatives up to order $\lfloor \beta \rfloor$ exist and are bounded, and the partial derivatives of order $\lfloor \beta \rfloor$ are $(\beta - \lfloor \beta \rfloor)$-Hölder. Here, the ball of β-smooth Hölder functions of radius K is defined as

$$\mathcal{C}_r^\beta(K) = \Big\{ f : [-1,1]^r \to [-1,1] : \\ 2r \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|<\lfloor\beta\rfloor} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + 2^{\beta-\lfloor\beta\rfloor} \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor\beta\rfloor} \sup_{\substack{\mathbf{x},\mathbf{y}\in[-1,1]^r \\ \mathbf{x}\neq\mathbf{y}}} \frac{|\partial^{\boldsymbol{\alpha}} f(\mathbf{x}) - \partial^{\boldsymbol{\alpha}} f(\mathbf{y})|}{|\mathbf{x}-\mathbf{y}|_\infty^{\beta-\lfloor\beta\rfloor}} \leq K \Big\}, \tag{5}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r) \in \mathbb{N}^r$ is a multi-index, $|\boldsymbol{\alpha}| := |\boldsymbol{\alpha}|_1$ and $\partial^{\boldsymbol{\alpha}} = \partial^{\alpha_1} \cdots \partial^{\alpha_r}$. The factors $2r$ and $2^{\beta-\lfloor\beta\rfloor}$ guarantee the embedding $\mathcal{C}_r^\beta(K) \subseteq \mathcal{C}_r^{\beta'}(K)$ whenever $\beta' \leq \beta$, see Lemma 15.

For $\beta = 1$, we recover the Lipschitz functions $|f(\mathbf{x}) - f(\mathbf{y})| \leq \frac{K}{2}|\mathbf{x}-\mathbf{y}|_\infty$. If for a positive integer β all the partial derivatives $\partial^{\boldsymbol{\alpha}} f$ with $|\boldsymbol{\alpha}| \leq \beta$ exist and are continuous, then it follows from the definition of the Hölder functions and first order Taylor expansion that $f \in \mathcal{C}_r^\beta(K)$ for all sufficiently large K. Moreover, if $f(x_1, \ldots, x_r) = g_1(x_1) \cdot \ldots \cdot g_r(x_r)$ with $g_1, \ldots, g_r \in \mathcal{C}_1^\beta(K)$, then there exists a finite K' such that $f \in \mathcal{C}_r^\beta(K')$, see Lemma 14 for a detailed statement and a proof.

We impose now Hölder smoothness on the functions $\overline{h}_{ij}$ in (3), assuming that $\overline{h}_{ij} \in \mathcal{C}_{t_i}^{\beta_i}(K)$, with $\beta_i \in [\beta_-, \beta_+]$ for some known and fixed lower and upper bounds β_-, β_+ satisfying $0 < \beta_- \leq \beta_+ < +\infty$. The smoothness indices of the $q+1$ components are collected into the vector

$$\boldsymbol{\beta} := (\beta_0, \ldots, \beta_q) \in [\beta_-, \beta_+]^{q+1} =: I(\lambda). \tag{6}$$

Combined with the graph parameter (4), the compositional functions (2) are completely described by

$$\eta := (\lambda, \boldsymbol{\beta}) = (q, \mathbf{d}, \mathbf{t}, \mathcal{S}, \boldsymbol{\beta}). \tag{7}$$

We refer to η as the composition structure of the regression function f in (2). The set of all possible choices of $\eta = (\lambda, \boldsymbol{\beta})$ with $\lambda \in \Lambda$ and $\boldsymbol{\beta} \in I(\lambda)$ is denoted by Ω.

Throughout the following, we assume that the true regression function f^* belongs to the function space $\mathcal{F}(\eta^*, K)$, for some unknown $\eta^* \in \Omega$, where

$$\mathcal{F}\big(\eta, K\big) := \Big\{ f = h_q \circ h_{q-1} \circ \ldots \circ h_1 \circ h_0 : \; h_i = (h_{ij})_j : [-1,1]^{d_i} \to [-1,1]^{d_{i+1}}, \\ \overline{h}_{ij} \in \mathcal{C}_{t_i}^{\beta_i}\big(K\big) \Big\}, \tag{8}$$

for some known $K > 0$. Given a true regression function f^*, there might be several compositional function decompositions η for which $f^* \in \mathcal{F}(\eta, K)$. Our main result on posterior contraction rates states that, if $f^* \in \mathcal{F}(\eta, K)$ for a given η, then the posterior contracts with a rate $\varepsilon_n(\eta)$. This means that if $f^* \in \mathcal{F}(\eta_j, K)$ for different choices $\eta_1, \ldots, \eta_m$, the smallest rate $\wedge_{j=1}^m \varepsilon_n(\eta_j)$ can be achieved.

3. Deep Gaussian Process Priors

In this section, we construct deep Gaussian process priors as prior on composition functions. The hierarchical prior construction assigns first specific prior weights to all composition structures and, for any fixed structure, Gaussian process priors on all the individual functions that occur in the representation. To achieve fast contraction rates, the prior weights on the composition structures need to be selected carefully. Due to the complexity, the construction is split into several steps.

Step 0. Choice of Gaussian processes. For a centered Gaussian process $X = (X_t)_{t\in T}$, the covariance operator viewed as a function on $T \times T$, that is, $(s,t) \mapsto k(s,t) = \mathbb{E}[X_s X_t]$ is a positive semidefinite function. We refer to the reproducing kernel Hilbert space generated by k as the RKHS corresponding to the Gaussian process X, following the formalization by van der Vaart and van Zanten (2008b).

For any dimension $r = 1, 2, \ldots,$ and any $\beta > 0$, pick a centered Gaussian process $\widetilde{G}^{(\beta,r)} = (\widetilde{G}^{(\beta,r)}(\mathbf{u}))_{\mathbf{u}\in[-1,1]^r}$ on the Banach space of continuous functions from $[-1,1]^r$ to $\mathbb{R}$ equipped with the supremum norm. We will use this process later as a prior for β-Hölder smooth functions defined on $[-1,1]^r$ in the compositional structure. Write $\|\cdot\|_{\mathbb{H}^{(\beta,r)}}$ for the RKHS-norm of the reproducing kernel Hilbert space $\mathbb{H}^{(\beta,r)}$ corresponding to $\widetilde{G}^{(\beta,r)}$. For positive Hölder radius K, we call

$$\varphi^{(\beta,r,K)}(u) := \sup_{f\in\mathcal{C}_r^\beta(K)} \inf_{g:\|g-f\|_\infty\le u} \|g\|^2_{\mathbb{H}^{(\beta,r)}} - \log \mathbb{P}\big(\big\|\widetilde{G}^{(\beta,r)}\big\|_\infty \le u\big), \tag{9}$$

the concentration function over $\mathcal{C}_r^\beta(K)$. This is the global version of the local concentration function appearing in the posterior contraction theory for Gaussian process priors byvan der Vaart and van Zanten (2008a). For any $0 < \alpha \le 1$, let $\varepsilon_n(\alpha,\beta,r)$ be such that

$$\varphi^{(\beta,r,K)}\big(\varepsilon_n(\alpha,\beta,r)^{1/\alpha}\big) \le n\varepsilon_n(\alpha,\beta,r)^2. \tag{10}$$

Step 1. Deep Gaussian processes. We now define a corresponding DGP $G^{(\eta)}$ on a given composition structure $\eta = (q, \mathbf{d}, \mathbf{t}, \mathcal{S}, \boldsymbol{\beta})$. Let $\mathbb{B}_\infty(R) := \{f : \sup_{\mathbf{x}\in[-1,1]^r} |f(\mathbf{x})| \le R\}$ be the supremum unitary ball with radius R. For simplicity, we suppress the dependence on r. Recall that the Hölder ball radius K in (8) is assumed to be known. With $\alpha_i := \prod_{\ell=i+1}^q (\beta_\ell \wedge 1)$, the subset of paths

$$\mathcal{D}_i(\eta, K) := \mathbb{B}_\infty(1) \cap \Big(\mathcal{C}_{t_i}^{\beta_i}(K) + \mathbb{B}_\infty\big(2\varepsilon_n(\alpha_i,\beta_i,t_i)^{1/\alpha_i}\big)\Big), \tag{11}$$

contains all functions that belong to the supremum unitary ball $\mathbb{B}_\infty(1)$ and are at most $2\varepsilon_n(\alpha_i,\beta_i,t_i)^{1/\alpha_i}$-away in supremum norm from the Hölder-ball $\mathcal{C}_{t_i}^{\beta_i}(K)$. With $\widetilde{G}^{(\beta,r)}$ the centred Gaussian process in Step 0, write $\overline{G}_i^{(\beta_i,t_i)}$ for the process $\widetilde{G}^{(\beta_i,t_i)}$ conditioned on the event $\{\widetilde{G}^{(\beta_i,t_i)} \in \mathcal{D}_i(\eta,K)\}$. Recall that for an index set S, the function $(\cdot)_S$ maps a vector to the components in S. For each $i = 0,\ldots,q$, $j = 1,\ldots,d_{i+1}$, define the component functions $G_{ij}^{(\eta)}$ to be independent copies of the processes $\overline{G}_i^{(\beta_i,t_i)} \circ (\cdot)_{\mathcal{S}_{ij}} : [-1,1]^{d_i} \to [-1,1]$. Finally, set $G_i^{(\eta)} := (G_{ij}^{(\eta)})_{j=1}^{d_{i+1}}$ and define the deep Gaussian process $G^{(\eta)} := G_q^{(\eta)} \circ \ldots \circ G_0^{(\eta)} : [-1,1]^d \to [-1,1]$. We denote by $\Pi(\cdot|\eta)$ the distribution of $G^{(\eta)}$.

Step 2. Structure prior. We now construct a hyperprior on the underlying composition structure. For $I(\lambda)$ as in (6) and for any function $a(\eta) = a(\lambda, \boldsymbol{\beta})$, it is convenient to define

$$\int a(\eta)\, d\eta := \sum_{\lambda \in \Lambda} \int_{I(\lambda)} a(\lambda, \boldsymbol{\beta})\, d\boldsymbol{\beta}.$$

Let γ be a probability density on the possible composition structures, that is, $\int \gamma(\eta)\, d\eta = 1$. We can construct such a measure γ by first choosing a distribution on the number of compositions q. Given q one can then select distributions on the ambient dimensions $\mathbf{d}$, the effective dimensions $\mathbf{t}$, the active sets $\mathcal{S}$ and finally the smoothness indices $\boldsymbol{\beta} \in [\beta_-, \beta_+]^{q+1}$ via the conditional density formula $\gamma(\eta) = \gamma(\lambda)\gamma(\boldsymbol{\beta}|\lambda) = \gamma(q)\gamma(\mathbf{d}|q)\gamma(\mathbf{t}|\mathbf{d}, q)\gamma(\mathcal{S}|\mathbf{t}, \mathbf{d}, q)\gamma(\boldsymbol{\beta}|\lambda)$. For a sequence $\varepsilon_n(\eta)$ satisfying

$$\varepsilon_n(\eta) \geq \max_{i=0,\ldots,q} \varepsilon_n(\alpha_i, \beta_i, t_i), \quad \text{with} \quad \alpha_i := \prod_{\ell=i+1}^{q} \big(\beta_\ell \wedge 1\big), \tag{12}$$

and $|\mathbf{d}|_1 = 1 + \sum_{i=0}^q d_i$, consider the hyperprior

$$\pi(\eta) := \frac{e^{-\Psi_n(\eta)}\gamma(\eta)}{\int e^{-\Psi_n(\eta)}\gamma(\eta)\, d\eta}, \quad \text{with} \quad \Psi_n(\eta) := n\varepsilon_n(\eta)^2 + e^{e^{|\mathbf{d}|_1}}. \tag{13}$$

The denominator is positive and finite, since $0 < e^{-\Psi_n(\eta)} \leq 1$ and $\int \gamma(\eta)\, d\eta = 1$.

Step 3. DGP prior. We consider deep Gaussian process priors of the form

$$\Pi(df) := \int_\Omega \Pi(df|\eta)\pi(\eta)\, d\eta, \tag{14}$$

where Ω is the set of all valid composition structures, $\Pi(\cdot|\eta)$ is the distribution of the DGP $G^{(\eta)}$ and $\pi(\eta)$ is the structure prior on η.

Some remarks on the DGP prior. Step 0 allows for a large class of Gaussian processes, since the only requirement is that the concentration function inequality (10) admits solutions for any $0 < \alpha \leq 1$. Lemma 5 shows that it is often enough to check (10) for $\alpha = 1$ only.

Composing Gaussian processes leads to wild sample paths with non-negligible probability. We overcome the issue in Step 1, by conditioning the Gaussian processes onto the sets $\mathcal{D}_i(\eta, K)$. This is well-defined since $\mathcal{D}_i(\eta, K)$ contains the sup-norm ball $\mathbb{B}_\infty\big(2\varepsilon_n(\alpha_i, \beta_i, t_i)^{1/\alpha_i}\big)$ and Gaussian processes with continuous sample paths give positive mass to $\mathbb{B}_\infty(R)$ for any positive radius $R > 0$. By choosing suitable Gaussian processes, the conditioning is not very restrictive. Brownian motion, for instance, is known to have sample paths that are almost surely β-Hölder continuous for any $\beta < 1/2$. The Hölder norm is, however, random. The conditioning step requires then to constraint the prior to Brownian motion sample paths with Hölder norm bounded by K.

Step 2 of the DGP prior construction introduces the rate $\varepsilon_n(\eta)$. This sequence will later be shown to be the posterior contraction rate if the true regression function f^* is an element of the compositional space $\mathcal{F}(\eta, K)$. To achieve fast posterior contraction rates, it is desirable to pick a small $\varepsilon_n(\eta)$ in (12). On the contrary, Assumption 2 below imposes additional restrictions on the rate.

The prior $\pi(\eta)$ on the composition structure η should be viewed as a model selection prior, as discussed (in Chapter 10) by Ghosal and van der Vaart (2017). As always, some care is needed to avoid posterior contraction on models that are too large and consequently lead to overfitting and suboptimal posterior contraction rates. This is achieved by the carefully chosen exponent Ψ_n in (13), which depends on the sample size and penalizes large composition structures.

4. Main Results

Denote by $\Pi\big(\cdot|\mathbf{X},\mathbf{Y}\big)$ the posterior distribution corresponding to a DGP prior Π constructed as above and $(\mathbf{X},\mathbf{Y}) = (\mathbf{X}_i, Y_i)_i$ a sample from the nonparametric regression model (1). For normalizing factor $Z_n := \int p_f/p_{f^*}(\mathbf{X},\mathbf{Y})\,\Pi(df)$ and any Borel measurable $\mathcal{A}$ in the Banach space of continuous functions on $[-1,1]^d$,

$$\Pi\big(\mathcal{A}|\mathbf{X},\mathbf{Y}\big) = Z_n^{-1}\int_{\mathcal{A}} \frac{p_f}{p_{f^*}}(\mathbf{X},\mathbf{Y})\,\Pi(df), \quad \Pi(df) := \int_{\Omega} \Pi(df|\eta)\pi(\eta)\,d\eta, \tag{15}$$

where $(p_f/p_{f^*})(\mathbf{X},\mathbf{Y})$ denotes the likelihood ratio. With a slight abuse of notation, for any subset of composition structures $\mathcal{M}\subseteq\Omega$, we set

$$\Pi\big(\eta\in\mathcal{M}|\mathbf{X},\mathbf{Y}\big) := Z_n^{-1}\int \frac{p_f}{p_{f^*}}(\mathbf{X},\mathbf{Y})\int_{\mathcal{M}} \Pi(df|\eta)\pi(\eta)\,d\eta, \tag{16}$$

which is the contribution of the composition structures $\mathcal{M}\subseteq\Omega$ to the posterior mass.

Before we can state the results, we first need to impose some conditions. The first condition is on the distribution γ appearing in the graph prior in Step 2 of the DGP prior construction. While uniformity of $\gamma(\cdot|\lambda)$ is the most natural choice and simplifies the mathematical analysis considerably, the next condition also states that all graphs have to be charged with non-negative mass and requires $\int\sqrt{\gamma(\eta)}\,d\eta$ to be finite. The latter restricts the amount of prior mass that can be assigned to complex composition structures.

Assumption 1 *We assume that, for any graph λ, the measure $\gamma(\cdot|\lambda)$ is the uniform distribution on the hypercube of possible smoothness indices $I(\lambda) = [\beta_-,\beta_+]^{q+1}$. Furthermore, we assume that the distribution γ is independent of n, that it assigns positive mass $\gamma(\eta) > 0$ to all composition structures η, and that it satisfies $\int\sqrt{\gamma(\eta)}\,d\eta < +\infty$.*

The prior on γ can be constructed hierarchically by using the decomposition $\gamma(\eta) = \gamma(q)\gamma(\mathbf{d}|q)\gamma(\mathbf{t}|\mathbf{d},q)\gamma(\mathcal{S}|\mathbf{t},\mathbf{d},q)(\beta_+-\beta_-)^{-q-1}$. It is natural to draw q and all components of $\mathbf{d}|q$ independently from distributions on the positive integers and generate $\mathbf{t}|(\mathbf{d},q)$ and $\mathcal{S}|(\mathbf{t},\mathbf{d},q)$ independently from uniform distributions on the respective sample spaces. The distribution γ assigns positive mass to all composition structures. The next result shows that the previous assumption is satisfied under suitable moment conditions.

Lemma 1 *If the prior γ is generated hierarchically as above, where q is drawn from a distribution with finite moment $\mathbb{E}_q[A^q]$ for all $A>0$, and all components $d_1|q,\ldots,d_q|q$ are i.i.d. with $\mathbb{E}_{d_1|q}[d_1^3 2^{d_1^2}] < \infty$, then, γ satisfies Assumption 1.*

The second assumption deals with the rates in Step 2 of the DGP prior construction. We impose a lower bound on $\varepsilon_n(\alpha, \beta, r)$ and also require that, uniformly over all models $\eta = (\lambda, \boldsymbol{\beta}), \eta' = (\lambda, \boldsymbol{\beta}')$ sharing the same graph λ and having similar smoothness indices $\boldsymbol{\beta}, \boldsymbol{\beta}'$, the corresponding rates $\varepsilon_n(\eta), \varepsilon_n(\eta')$ only differ by multiplicative factors.

Assumption 2 *We assume the following on the rates appearing in the construction of the prior.*

(i) For any positive integer r, any $\beta > 0$, let $Q_1(\beta, r, K)$ be the constant from Lemma 19. Then, the sequences $\varepsilon_n(\alpha, \beta, r)$ solving the concentration function inequality (10) *are chosen in such a way that*

$$\varepsilon_n(\alpha, \beta, r) \geq Q_1(\beta, r, K)^{\frac{\beta}{2\beta+r}} n^{-\frac{\beta\alpha}{2\beta\alpha+r}}. \tag{17}$$

(ii) There exists a constant $Q \geq 1$ such that the following holds. For any $n > 1$, any graph $\lambda = (q, \mathbf{d}, \mathbf{t}, \mathcal{S})$ with $|\mathbf{d}|_1 = 1 + \sum_{i=0}^q d_i \leq \log(2 \log n)$, and any $\boldsymbol{\beta}' = (\beta_0', \ldots, \beta_q'), \boldsymbol{\beta} = (\beta_0, \ldots, \beta_q) \in I(\lambda)$ satisfying $\beta_i' \leq \beta_i \leq \beta_i' + 1/\log^2 n$ for all $i = 0, \ldots, q$, the rates relative to the composition structures $\eta = (\lambda, \boldsymbol{\beta})$ and $\eta' = (\lambda, \boldsymbol{\beta}')$ satisfy

$$\varepsilon_n(\eta) \leq \varepsilon_n(\eta') \leq Q\varepsilon_n(\eta). \tag{18}$$

The previous assumptions are checked for a number of examples in Section 6. The rate $\varepsilon_n(\eta)$ associated to a composition structure η can be viewed as measure of the complexity of this structure, where larger rates $\varepsilon_n(\eta)$ correspond to more complex models. Our first result states that the posterior concentrates on small models in the sense that all posterior mass is asymptotically allocated on a set

$$\mathcal{M}_n(C) := \big\{\eta : \varepsilon_n(\eta) \leq C\varepsilon_n(\eta^*)\big\} \cap \big\{\eta : |\mathbf{d}|_1 \leq \log(2 \log n)\big\} \tag{19}$$

with sufficiently large constant C. This shows that the posterior not only concentrates on models with fast rates $\varepsilon_n(\eta)$ but also on graph structures with number of nodes in each layer bounded by $\log(2 \log n)$. The proof is given in Appendix A.2.

Theorem 2 (Model selection) *Let $\Pi\big(\cdot|\mathbf{X}, \mathbf{Y}\big)$ be the posterior distribution corresponding to a DGP prior Π constructed as in Section 3 and satisfying Assumptions 1 and 2. Let $\eta^* = (\lambda^*, \boldsymbol{\beta}^*)$ for some $\boldsymbol{\beta}^* \in (\beta_-, \beta_+)^{q^*+1}$ and suppose $\varepsilon_n(\eta^*) \leq 1/(4Q)$. Then, for a positive constant $C = C(\eta^*)$,*

$$\sup_{f^* \in \mathcal{F}(\eta^*, K)} \mathbb{E}_{f^*}\big[\Pi\big(\eta \notin \mathcal{M}_n(C) \big| \mathbf{X}, \mathbf{Y}\big)\big] \xrightarrow{n \to \infty} 0,$$

where $\mathbb{E}_{f^*}$ denotes the expectation with respect to $\mathbb{P}_{f^*}$, the true distribution of the sample $(\mathbf{X}, \mathbf{Y})$.

Denote by μ the distribution of the covariate vector $\mathbf{X}_1$ and write $L^2(\mu)$ for the weighted L^2-space with respect to the measure μ. The next result shows that the posterior distribution achieves contraction rate $\varepsilon_n(\eta^*)$ up to a $\log n$ factor. The proof is based on the testing approach developed by Ghosal et al. (2000) and given in Appendix A.2.

Theorem 3 (Posterior contraction) *Let $\Pi\big(\cdot\,|\mathbf{X},\mathbf{Y}\big)$ be the posterior distribution corresponding to a DGP prior Π constructed as in Section 3 and satisfying Assumptions 1 and 2. Let $\eta^* = (\lambda^*, \boldsymbol{\beta}^*)$ for some $\boldsymbol{\beta}^* \in (\beta_-, \beta_+)^{q^*+1}$ and suppose $\varepsilon_n(\eta^*) \le 1/(4Q)$. Then, for a positive constant $L = L(\eta^*)$,*

$$\sup_{f^* \in \mathcal{F}(\eta^*, K)} \mathbb{E}_{f^*}\left[\Pi\left(\|f - f^*\|_{L^2(\mu)} \ge L(\log n)^{1+\log K}\varepsilon_n(\eta^*)\big|\mathbf{X},\mathbf{Y}\right)\right] \xrightarrow{n\to\infty} 0,$$

where $\mathbb{E}_{f^*}$ denotes the expectation with respect to $\mathbb{P}_{f^*}$, the true distribution of the sample $(\mathbf{X},\mathbf{Y})$.

Remark 4 *Our proving strategy allows for the following modification to the construction of the DGP prior. The concentration functions $\varphi^{(\beta,r,K)}$ in (9) are defined globally over the Hölder-ball $\mathcal{C}_r^\beta(K)$. The concentration function inequality in (10) essentially requires that the closure $\overline{\mathbb{H}}^{(\beta,r)}$ of the RKHS of the underlying Gaussian process $\widetilde{G}^{(\beta,r)}$ contains the whole Hölder ball. There are classical examples for which this is too restrictive, and one might want to weaken the construction by considering a subset $\mathcal{H}_r^\beta(K) \subseteq \mathcal{C}_r^\beta(K)$. This can be done by replacing $\mathcal{C}_{t_i}^{\beta_i}(K)$ with the corresponding subset $\mathcal{H}_{t_i}^{\beta_i}(K)$ in the definition of the conditioning sets $\mathcal{D}_i(\eta, K)$ in (11) and the function class $\mathcal{F}(\eta, K)$ in (8). As a consequence, this also reduces the class of functions for which the posterior contraction rates derived in Theorem 3 hold.*

We do not impose an a-priori known upper bound on the complexity of the underlying composition structure (2). While we think that this is natural in practice, it causes extra technical complications resulting for instance in the appearance of the bound $|\mathbf{d}|_1 \le \log(2\log n)$ in Assumption 2 and in the definition of the model class $\mathcal{M}_n(C)$ in (19). If we additionally assume that the true composition structure satisfies $|\mathbf{d}|_1 \le D$ for a known upper bound D, then the factor $e^{e^{|\mathbf{d}|_1}}$ in (13) can be avoided. Moreover, the $(\log n)^{1+\log K}$-factor occurring in the posterior contraction rate is somehow an artifact of the proof, and could be replaced by K^D, see the proof of Lemma 20 for more details. A trade-off regarding the choice of K appears. To allow for larger classes of functions and a weaker constraint induced by the conditioning on (11), we want to select a large K. On the contrary, large K results in slower posterior contraction guarantees.

The main result requires that all smoothness indices lie in the interval $[\beta_-, \beta_+]$ with $0 < \beta_- < \beta_+ < \infty$. As commonly observed in nonparametric Bayes, extension to $(0, \infty)$ is highly non-trivial as many constants depend in an intricate way on β_-, β_+ and quickly explode in the limits $\beta_- \to 0$ and $\beta_+ \to \infty$. A prototypical example for the latter is the constant e^{β_+} in Lemma 7 below. One could also wonder what would happen in the misspecified case where the true $\boldsymbol{\beta}^*$ lies outside of the compact set $[\beta_-, \beta_+]^{q^*+1}$. For nonparametric regression using GPs, it is known that extending the range of $\boldsymbol{\beta}^*$ while keeping the prior on $\boldsymbol{\beta}$ as before, the rates become suboptimal by a polynomial factor in the sample size n due to the mismatch between the smoothness of the prior and that of the true regression function, Castillo (2008) showed this for the Riemann-Liouville prior.

For mathematical convenience, all functions in the function class (8) have range $[-1, 1]$. In general, any function f mapping to $[-a, a]$ can be represented as in (2) where all intermediate components h_{ij} in the compositional class have range $[-1, 1]$ and only the very last h_q has range $[-a, a]$. In the construction of the DGP prior, we then need to restrict for $i = q$ in (11) to the set $\mathbb{B}_\infty(a)$ instead. The mathematical analysis can be extended. In particular, in the Hellinger-$L^2(\mu)$ equivalence in (28), the constant factor becomes $e^{-a^2/2}$.

By making appropriate changes in the proofs, one can also include in the nonparametric model (1) a known standard deviation $\sigma^* > 0$ instead of taking standard normal noise. As the analysis is already quite technical, we have chosen to avoid additional parameters that are not of primary focus.

We view the proposed Bayesian analysis rather as a proof of concept than something that is straightforward implementable or computationally efficient. The main obstacles towards a scalable Bayesian method are the combinatorial nature of the set of graphs as well as conditioning the sample paths to neighborhoods of Hölder functions, see also Section 7 for a more in-depth discussion.

5. Nearly Optimal Contraction Rates

Theorem 2.5 by Ghosal et al. (2000) ensures the existence of a frequentist estimator converging to the true parameter with the posterior contraction rate. This implies that the fastest possible posterior contraction rate is the minimax estimation rate. For the prediction loss $\|f - g\|_{L^2(\mu)}$, the minimax estimation rate over the class $\mathcal{F}(\eta, K)$ is, up to some logarithmic factors,

$$\mathfrak{r}_n(\eta) = \max_{i=0,\ldots,q} n^{-\frac{\beta_i\alpha_i}{2\beta_i\alpha_i + t_i}}, \quad \text{with} \quad \alpha_i := \prod_{\ell=i+1}^{q} \big(\beta_\ell \wedge 1\big), \tag{20}$$

as shown by Schmidt-Hieber (2020). This rate is attained by suitable estimators based on sparsely connected deep neural networks. It is also shown in the recent work by Giordano et al. (2022) that there are compositional structures for which GP priors lead to a posterior contraction rate that is suboptimal by a polynomial factor compared with $\mathfrak{r}_n(\eta)$.

Below we derive sufficient conditions that are simpler than Assumption 2, apply to standard examples of Gaussian processes in the Bayes literature and imply an optimal posterior contraction rate $\mathfrak{r}_n(\eta)$ up to $\log n$ factors.

The first result shows that the solution to the concentration function inequality for arbitrary $0 < \alpha \leq 1$ can be deduced from the solution for $\alpha = 1$. The proof is in Appendix A.3.

Lemma 5 *Let $\varepsilon_n(1, \beta, r)$ be a solution to the concentration function inequality* (10) *for $\alpha = 1$. Then, any sequence $\varepsilon_n(\alpha, \beta, r) \geq \varepsilon_{m_n}(1, \beta, r)^\alpha$ where m_n is chosen such that $m_n \varepsilon_{m_n}(1, \beta, r)^{2-2\alpha} \leq n$, solves the concentration function inequality for arbitrary $\alpha \in (0, 1]$.*

The next result identifies sequences $\varepsilon_n(\alpha, \beta, r)$ satisfying inequality (10) provided the solution $\varepsilon_n(1, \beta, r)$ for $\alpha = 1$ is, up to $\log n$ factors, $n^{-\beta/(2\beta+r)}$. The proof is given in Appendix A.3.

Lemma 6 *Let $n \geq 3$.*

(i) If the sequence $\varepsilon_n(1, \beta, r) = C_1 (\log n)^{C_2} n^{-\beta/(2\beta+r)}$ solves the concentration function inequality (10) *for $\alpha = 1$, with constants $C_1 \geq 1$ and $C_2 \geq 0$, then, any sequence*

$$\varepsilon_n(\alpha, \beta, r) \geq C_1^2 (2\beta + 1)^{2C_2} (\log n)^{C_2(2\beta+2)} \ n^{-\frac{\beta\alpha}{2\beta\alpha + r}}$$

solves the concentration function inequality for arbitrary $\alpha \in (0, 1]$.

(ii) If there are constant $C_1' \geq 1$ *and* $C_2' \geq 0$ *such that the concentration function satisfies*

$$\varphi^{(\beta,r,K)}(\delta) \leq C_1'(\log \delta^{-1})^{C_2'} \delta^{-\frac{r}{\beta}}, \quad \textit{for all } 0 < \delta \leq 1, \tag{21}$$

then, any sequence

$$\varepsilon_n(\alpha, \beta, r) \geq C_1'(\beta \log n)^{C_2'}\, n^{-\frac{\beta\alpha}{2\beta\alpha+r}}$$

solves the concentration function inequality (10) *for arbitrary* $\alpha \in (0, 1]$.

To verify Assumption 2, we need to pick suitable sequences $\varepsilon_n(\eta) \geq \max_{i=0,\ldots,q} \varepsilon_n(\alpha_i, \beta_i, t_i)$. In the previous lemma, $\varepsilon_n(\alpha, \beta, r) = C_1(\beta, r)(\log n)^{C_2(\beta,r)} n^{-\beta\alpha/(2\beta\alpha+r)}$ for some constants $C_1(\beta, r) \geq 1$ and $C_2(\beta, r) \geq 0$. A suitable choice in this case is

$$\varepsilon_n(\eta) = \widetilde{C}_1(\eta)(\log n)^{\widetilde{C}_2(\eta)} \mathfrak{r}_n(\eta), \tag{22}$$

provided the constants $\widetilde{C}_j(\eta) := \max_{i=0,\ldots,q} \sup_{\beta\in[\beta_-,\beta_+]} C_j(\beta, t_i)$, $j \in \{1, 2\}$ are finite. This can be checked by verifying that $\beta \mapsto C_j(\beta, r)$, $j \in \{1, 2\}$ are bounded functions on $[\beta_-, \beta_+]$.

Lemma 7 *The rates* $\varepsilon_n(\eta)$ *in* (22) *satisfy condition (ii) in Assumption 2 with* $Q = e^{\beta_+}$.

Let Π be a DGP prior constructed with the Gaussian processes and rates given in this section. Then, the corresponding posterior satisfies Theorem 3 and contracts with rate $\mathfrak{r}_n(\eta^*)$ up to the multiplicative factor $L(\eta^*)\widetilde{C}_1(\lambda^*, K)(\log n)^{1+\log K+\widetilde{C}_2(\lambda^*,K)}$. An examination of the proof gives $L(\eta^*) = M \cdot 10Ce^{\beta_+}$ for $M > 0$ a sufficiently large universal constant.

6. Examples of DGP Priors

In this section we verify the abstract conditions for standard choices of Gaussian processes $\{\widetilde{G}^{(\beta,r)} : \beta \in [\beta_-, \beta_+]\}$ and show that they achieve nearly-optimal posterior contraction rates.

6.1 Lévy's Fractional Brownian Motion

Assume that the upper bound β_+ on the possible range of smoothness indices is bounded by one. A zero-mean Gaussian process X^β is called a Lévy fractional Brownian motion of order $\beta \in (0, 1)$ if

$$X^\beta(0) = 0, \quad \mathbb{E}\left[|X^\beta(\mathbf{u}) - X^\beta(\mathbf{u}')|^2\right] = |\mathbf{u} - \mathbf{u}'|_2^{2\beta}, \quad \forall \mathbf{u}, \mathbf{u}' \in [-1, 1]^r.$$

The covariance function of the process is $\mathbb{E}[X^\beta(\mathbf{u})X^\beta(\mathbf{u}')] = \frac{1}{2}(|\mathbf{u}|_2^{2\beta} + |\mathbf{u}'|_2^{2\beta} - |\mathbf{u} - \mathbf{u}'|_2^{2\beta})$. In Chapter 3 of their book, Cohen and Istas (2013) provide the following representation for a r-dimensional β-fractional Brownian motion. Denote by $\widehat{f}(\boldsymbol{\xi}) := (2\pi)^{-r/2} \int_{\mathbb{R}^r} e^{i\mathbf{u}^\top \boldsymbol{\xi}} f(\mathbf{u}) d\mathbf{u}$ the Fourier transform of the function f. For $W = (W(\mathbf{u}))_{\mathbf{u}\in[-1,1]^r}$ a multidimensional Brownian motion, and C_β a positive constant depending only on β, r,

$$X^\beta(\mathbf{u}) = \int_{\mathbb{R}^r} \frac{e^{-i\mathbf{u}^\top \boldsymbol{\xi}} - 1}{C_\beta^{1/2} |\boldsymbol{\xi}|_2^{\beta+r/2}} \widehat{W}(d\boldsymbol{\xi}),$$

in distribution, where $\widehat{W}(d\boldsymbol{\xi})$ is the Fourier transform of the Brownian random measure $W(d\mathbf{u})$, see Section 2.1.6 of the book by Cohen and Istas (2013) for definitions and properties. The same reference defines, for all $\varphi \in L^2([-1,1]^r)$, the integral operator

$$(I^\beta \varphi)(\mathbf{u}) := \int_{\mathbb{R}^r} \overline{\widehat{\varphi}(\boldsymbol{\xi})} \frac{e^{-i\mathbf{u}^\top \boldsymbol{\xi}} - 1}{C_\beta^{1/2} |\boldsymbol{\xi}|_2^{\beta + r/2}} \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}.$$

The RKHS $\mathbb{H}^\beta$ of X^β is given in Section 3.3 of the same book as

$$\mathbb{H}^\beta = \Big\{ I^\beta \varphi : \varphi \in L^2([-1,1]^r) \Big\}, \quad \langle I^\beta \varphi, I^\beta \varphi' \rangle_{\mathbb{H}^\beta} = \langle \varphi, \varphi' \rangle_{L^2([-1,1]^r)}.$$

The process X^β is always zero at $\mathbf{u} = 0$. To release it at zero, let $Z \sim \mathcal{N}(0,1)$ be independent of X^β and consider the process $\mathbf{u} \mapsto Z + X^\beta(\mathbf{u})$. The RKHS of the constant process $\mathbf{u} \mapsto Z$ is the set $\mathbb{H}^Z$ of all constant functions and, according to Lemma I.18 by Ghosal and van der Vaart (2017), the RKHS of $Z + X^\beta$ is the direct sum $\mathbb{H}^Z \oplus \mathbb{H}^\beta$.

The next result is proved in Appendix A.4 and can be viewed as the multidimensional extension of the RKHS bounds in Theorem 4 by Castillo (2008). Whereas the original proof relies on kernel smoothing and Taylor approximations, we use a spectral approach. Write

$$\mathcal{W}_r^\beta(K) := \Big\{ h : [-1,1]^r \to [-1,1] : \int_{\mathbb{R}^r} |\widehat{h}(\boldsymbol{\xi})|^2 (1 + |\boldsymbol{\xi}|_2)^{2\beta} \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} \le K \Big\}, \tag{23}$$

for the β-Sobolev ball of radius K.

Lemma 8 *Let $\beta \in [\beta_-, 1]$ and $Z + X^\beta = (Z + X^\beta(\mathbf{u}))_{\mathbf{u} \in [-1,1]^r}$ the fractional Brownian motion of order β released at zero. Fix $h \in \mathcal{C}_r^\beta(K) \cap \mathcal{W}_r^\beta(K)$. Set $\phi_\sigma = \sigma^{-r} \phi(\cdot/\sigma)$ with ϕ a suitable regular kernel and $\sigma < 1$. Then, $\|h * \phi_\sigma - h\|_\infty \le K R_\beta \sigma^\beta$ and $\|h * \phi_\sigma\|^2_{\mathbb{H}^Z \oplus \mathbb{H}^\beta} \le K^2 L_\beta^2 \sigma^{-r}$ for some constants R_β, L_β that depend only on β, r.*

The next lemma shows that for Lévy's fractional Brownian motion released at zero nearly-optimal posterior contraction rates can be obtained. For that we need to restrict the definition of the global concentration function to the smaller class $\mathcal{C}_r^\beta(K) \cap \mathcal{W}_r^\beta(K)$. The proof of the lemma is in Appendix A.4.

Lemma 9 *Let $\beta_+ \le 1$ and work on the reduced function spaces $\mathcal{H}_r^\beta(K) = \mathcal{C}_r^\beta(K) \cap \mathcal{W}_r^\beta(K)$ as outlined in Remark 4. For $\{\widetilde{G}^{(\beta,r)} : \beta \in [\beta_-, \beta_+]\}$ the family of Levy's fractional Brownian motions $Z + X^\beta$ released at zero, there exist sequences $\varepsilon_n(\eta) = C_1(\eta)(\log n)^{C_2(\eta)} \mathfrak{r}_n(\eta)$ such that Assumption 2 holds.*

6.2 Truncated Wavelet Series

Let $\{\psi_{j,k} : j \in \mathbb{N}_+,\ k = 1, \dots 2^{jr}\}$ be an orthonormal wavelet basis of $L^2([-1,1]^r)$. For any $\varphi \in L^2([-1,1]^r)$, we denote by $\varphi = \sum_{j=1}^\infty \sum_{k=1}^{2^{jr}} \lambda_{j,k}(\varphi) \psi_{j,k}$ its wavelet expansion. The quantities $\lambda_{j,k}(\varphi)$ are the corresponding real coefficients. For any $\beta > 0$, we denote by $\mathcal{B}_{\infty,\infty,\beta}$ the Besov space of functions φ with finite

$$\|\varphi\|_{\infty,\infty,\beta} := \sup_{j \in \mathbb{N}} 2^{j(\beta + \frac{r}{2})} \max_{k=1,\dots 2^{jr}} |\lambda_{j,k}(\varphi)|.$$

We assume that the wavelet basis is s-regular with $s > \beta_+$. For i.i.d. random variables $Z_{j,k} \sim \mathcal{N}(0,1)$, consider the Gaussian process induced by the truncated series expansion

$$X^\beta(\mathbf{u}) := \sum_{j=1}^{J_\beta} \sum_{k=1}^{2^{jr}} \frac{2^{-j(\beta+\frac{r}{2})}}{\sqrt{jr}} Z_{j,k}\psi_{j,k}(\mathbf{u}),$$

where the maximal resolution J_β is chosen as the integer closest to the solution J of the equation $2^J = n^{1/(2\beta+r)}$, see Section 4.5 by van der Vaart and van Zanten (2008a). The RKHS of the process X^β is given in the proof of Theorem 4.5 by van der Vaart and van Zanten (2008a) as the set $\mathbb{H}^\beta$ of functions $\varphi = \sum_{j=1}^{J_\beta} \sum_{k=1}^{2^{jr}} \lambda_{j,k}(\varphi)\psi_{j,k}$ with coefficients $\lambda_{j,k}(\varphi)$ satisfying $\|\varphi\|_{\mathbb{H}^\beta}^2 := \sum_{j=1}^{J_\beta} \sum_{k=1}^{2^{jr}} jr2^{2j(\beta+r/2)}\lambda_{j,k}(\varphi)^2 < \infty$.

For this family of Gaussian processes, it is rather straightforward to verify that conditioning on a neighbourhood of β-smooth functions as in Step 1 of the deep Gaussian process prior construction is not a restrictive constraint. The next result shows that, with high probability, the process X^β belong to the $\mathcal{B}_{\infty,\infty,\beta}$-ball of radius $(1+K')\sqrt{2\log 2}$, with $K' > \sqrt{3}$. In view of Section 4.3.6 by Giné and Nickl (2016), the Besov space $\mathcal{B}_{\infty,\infty,\beta}$ contains the Hölder space $\mathcal{C}_r^\beta$ for any $\beta > 0$, and they coincide whenever β is not an integer. Moreover, $\mathcal{B}_{\infty,\infty,\beta'} \subset \mathcal{C}_r^\beta$ for all $\beta' > \beta$.

Lemma 10 *Let X^β be the truncated wavelet process. Then, for any $K' > \sqrt{3}$,*

$$\mathbb{P}\left(\|X^\beta\|_{\infty,\infty,\beta} \le (1+K')\sqrt{2\log 2}\right) \ge 1 - \frac{4}{2^{rK'^2} - 4}.$$

The proof is postponed to Appendix A.4. The probability in the latter display converges quickly to one. As an example consider $K' = 2$. Since $r \ge 1$, the bound implies that more than 2/3 of the simulated sample paths $\mathbf{u} \mapsto X^\beta(\mathbf{u})$ lie in the Besov ball $\mathcal{B}_{\infty,\infty,\beta}(3\sqrt{2\log 2})$.

The next lemma shows that for the truncated series expansion, nearly-optimal posterior contraction rates can be achieved. The proof of the lemma is deferred to Appendix A.4.

Lemma 11 *For $\{\widetilde{G}^{(\beta,r)} : \beta \in [\beta_-, \beta_+]\}$ the family of truncated Gaussian processes X^β there exist sequences $\varepsilon_n(\eta) = C_1(\eta)(\log n)^{C_2(\eta)}\mathfrak{r}_n(\eta)$ such that Assumption 2 holds.*

6.3 Stationary Process

A zero-mean Gaussian process $X^\nu = (X^\nu(\mathbf{u}))_{\mathbf{u}\in[-1,1]^r}$ is called stationary if its covariance function can be represented by a spectral density measure ν on $\mathbb{R}^r$ as

$$\mathbb{E}[X^\nu(\mathbf{u})X^\nu(\mathbf{u}')] = \int_{\mathbb{R}^r} e^{-i(\mathbf{u}-\mathbf{u}')^\top \boldsymbol{\xi}}\nu(\boldsymbol{\xi})d\boldsymbol{\xi},$$

see Example 11.8 by Ghosal and van der Vaart (2017). We consider stationary Gaussian processes with radially decreasing spectral measures that have exponential moments, that is, $\int e^{c|\boldsymbol{\xi}|_2}\nu(\boldsymbol{\xi})d\boldsymbol{\xi} < +\infty$ for some $c > 0$. Such processes have smooth sample paths thanks to Proposition I.4 by Ghosal and van der Vaart (2017). An example is the square-exponential process with spectral measure $\nu(\boldsymbol{\xi}) = 2^{-r}\pi^{-r/2}e^{-|\boldsymbol{\xi}|_2^2/4}$. For any $\varphi \in L^2(\nu)$, set $(H^\nu\varphi)(\mathbf{u}) :=$

$\int_{\mathbb{R}^r} e^{i\boldsymbol{\xi}^\top \mathbf{u}} \varphi(\boldsymbol{\xi})\nu(\boldsymbol{\xi})d\boldsymbol{\xi}$. The RKHS of X^ν is given in Lemma 11.35 by Ghosal and van der Vaart (2017) as $\mathbb{H}^\nu = \{H^\nu\varphi : \varphi \in L^2(\nu)\}$ with inner product $\langle H^\nu\varphi, H^\nu\varphi'\rangle_{\mathbb{H}^\nu} = \langle \varphi, \varphi'\rangle_{L^2(\nu)}$.

For every $\beta \in [\beta_-, \beta_+]$, take $\widetilde{G}^{(\beta,r)}$ to be the rescaled process $X^\nu(a\cdot) = (X^\nu(a\mathbf{u}))_{\mathbf{u}\in[-1,1]^r}$ with scaling

$$a = a(\beta, r) = n^{\frac{1}{2\beta+r}} (\log n)^{-\frac{1+r}{2\beta+r}}. \tag{24}$$

The process $\widetilde{G}^{(\beta,r)}$ thus depends on the sample size n. We prove the next result in Appendix A.4.

Lemma 12 *For $\{\widetilde{G}^{(\beta,r)} : \beta \in [\beta_-, \beta_+]\}$ the family of rescaled stationary processes $X^\nu(a\cdot)$, there exist sequences $\varepsilon_n(\eta) = C_1(\eta)(\log n)^{C_2(\eta)}\mathfrak{r}_n(\eta)$ such that Assumption 2 holds.*

7. DGP Priors, Wide Neural Networks and Regularization

In this section, we explore similarities and differences between deep learning and the Bayesian analysis based on deep Gaussian process priors. Both methods are based on the likelihood. It is moreover known that standard random initialization schemes in deep learning converge to Gaussian processes in the wide limit. Since the initialization is crucial for the success of deep learning, this suggests that the initialization could act in a similar way as a Gaussian prior in the Bayesian world. Next to a proper initialization scheme, stability enhancing regularization techniques such as batch normalization are widely studied in deep learning and a comparison might help us to identify conditions that constraint the potentially wild behavior of deep Gaussian process priors. Below we investigate these aspects in more detail.

It has been argued in the literature that Bayesian neural networks and regression with Gaussian process priors are intimately connected. In Bayesian neural networks, we generate a function valued prior distribution by using a neural network and drawing the network weights randomly. Recall that a neural network with a single hidden layer is called shallow, and a neural network with a large number of units in all hidden layers is called wide. If the network weights in a shallow and wide neural network are drawn i.i.d., and the scaling of the variances is such that the prior does not become degenerate, then, it has been argued by Neal (1996) that the prior will converge in the wide limit to a Gaussian process prior and expressions for the covariance structure of the limiting process are known. One might be tempted to believe that for a deep neural network one should obtain a deep Gaussian process as a limit distribution. If the width of all hidden layers tends simultaneously to infinity, Matthews et al. (2018) proves that this is false and that one still obtains a Gaussian limit. The covariance of the limiting process is, however, more complicated and can be given via a recursion formula, where each step in the recursion describes the change of the covariance by a hidden layer. Matthews et al. (2018) shows moreover in a simulation study that Bayesian neural networks and Gaussian process priors with appropriate choice of the covariance structure behave indeed similarly.

It is conceivable that if one keeps the width of some hidden layers fixed and let the width of all other hidden layers tend to infinity, the Bayesian neural network prior will converge, if all variances are properly scaled, to a deep Gaussian process. By stacking for instance two shallow networks as indicated in Figure 3 and making the first and last hidden layer wide, the limit is the composition of two Gaussian process and thus, a deep Gaussian process. In

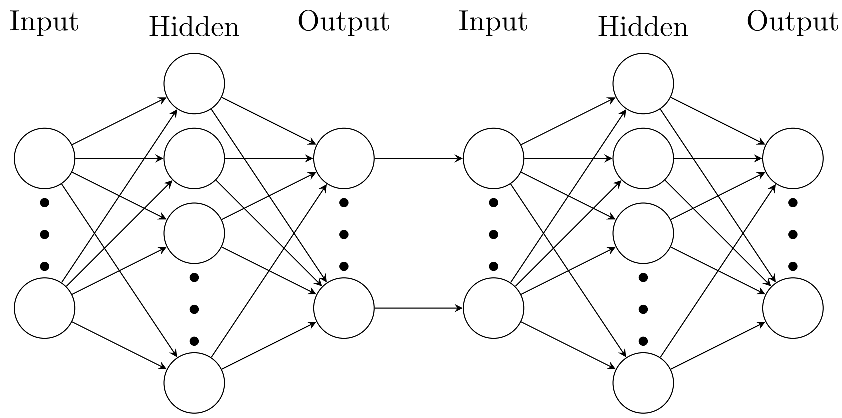

Figure 3: Schematic stacking of two shallow neural networks.

the hierarchical deep Gaussian prior construction in Section 3, we pick in a first step a prior on composition structures. For Bayesian neural networks this is comparable with selecting first a hyperprior on neural network architectures.

Even more recently, Peluchetti and Favaro (2020) as well as Hayou et al. (2019) studied the behaviour of neural networks with random weights when both depth and width tend to infinity.

While the discussion so far indicates that Bayesian neural networks and Bayes with deep Gaussian process priors are similar methods, we now show that deep learning with randomly initialized network weights behaves similarly as a Bayes estimator with respect to a deep Gaussian process prior. The random network initialization means that the deep learning algorithm is initialized approximately by a deep Gaussian process. As mentioned earlier, initialization is crucial for the success of deep learning and indeed acts as a prior. Denote by $-\ell$ the negative log-likelihood/cross entropy. Whereas in deep learning we fit a function by iteratively decreasing the cross-entropy using gradient descent method, the posterior is proportional to $\exp(\ell)\times$prior and concentrates on elements in the support of the prior with small cross entropy. Bayesian sampling methods such as stochastic gradient Langevin dynamics are closely related to noisy stochastic gradient descent, as shown by Welling and Teh (2011), and Gibbs sampling of the posterior has a similar flavor as coordinate-wise descent methods for the cross-entropy.

The only theoretical result that we are aware of examining the relationship between deep learning and Bayesian neural networks is by Polson and Ročková (2018). It proves that for a neural network prior with network weights drawn i.i.d. from a suitable spike-and-slab prior, the full posterior behaves similarly as the global minimum of the cross-entropy (that is, the empirical risk minimizer) based on sparsely connected deep neural networks.

As a last point we now compare stabilization techniques for deep Gaussian process priors and deep learning. In Step 1 of the deep Gaussian process prior construction, we have conditioned the individual Gaussian processes to map to $[-1, 1]$ and to generate sample paths in small neighborhoods of a suitable Hölder ball. This induces regularity in the prior and avoids the wild behaviour of the composed sample paths due to bad realizations of individual components. We argue that this form of regularization has a similar flavor as batch normalization, cf. Section 8.7 by Goodfellow et al. (2016). The purpose of batch normalization is to avoid vanishing or exploding gradients due to the composition of several functions in deep neural networks. The main idea underlying batch normalization is to normalize the outputs from a fixed hidden layer in the neural network before they become the input of the next hidden layer. The normalization step is different from the conditioning proposed for the compositions of Gaussian processes. In fact, for batch normalization, mean and variance of the outputs are estimated based on a subsample and an affine transformation is applied such that the outputs are approximately centered and have variance one. One of the key differences is that this normalization invokes the distribution of the underlying design, while the conditioning proposed for deep Gaussian processes is independent of the distribution of the covariates. One suggestion that we can draw from this comparison is that instead of conditioning the processes to have sample paths in $[-1,1]$, it might also be interesting to apply the normalization $f \mapsto f(t)/\sup_{t\in[-1,1]^r}|f(t)|$ between any two compositions. This also ensures that the output maps to $[-1,1]$ and is closer to batch normalization. A data-dependent normalization of the prior cannot be incorporated in the fully Bayesian framework considered here and would result in an empirical Bayes method.

7.1 Computational Challenges of DGP Priors

For complex tasks, sampling from the posterior is a notoriously hard computational problem. Several algorithmic breakthroughs resulted in computationally scalable procedures even in the particularly problematic case of high-dimensional model selection. While this is not the main focus of this work, we briefly discuss in this section aspects regarding the implementation of DGP priors and the underlying computational challenges.

Model selection prior. A major computational challenge is the large number of potential composition structures. A naive sampling method needs to visit all these composition structures. Below we describe several methods to reduce the number of candidate structures.

Firstly, we argue that many models can in principle be omitted as they do not lead to faster contraction rates. As an example consider functions of the form $f = h_1 \circ h_0$ for univariate functions h_i having smoothness $\beta_i \leq 1$, that is, $h_i \in \mathcal{C}_1^{\beta_i}(K)$ for $i = 0, 1$. From the definition of Hölder spaces, we can immediately infer that f has Hölder smoothness $\beta_0\beta_1$. This means that we can alternatively represent f as a degenerate composition model $f = h_0$ with $q = 0$ and h_0 a $\beta_0\beta_1$-smooth function. The optimal contraction rate $\mathfrak{r}_n(\eta)$ defined in (20) is in both cases $n^{-\beta_0\beta_1/(2\beta_0\beta_1+1)}$. Writing the function as a composition model, does therefore not lead to a faster contraction rate. The prior does consequently also not need to put any mass on this composition.

For a general composition structure $f = h_q \circ h_{q-1} \circ \ldots \circ h_0$, let i^* be an index dominating the rate in the sense that $\mathfrak{r}_n(\eta) = \max_{i=0,\ldots,q} n^{-\beta_i\alpha_i/(2\beta_i\alpha_i+t_i)} = n^{-\beta_{i^*}\alpha_{i^*}/(2\beta_{i^*}\alpha_{i^*}+t_{i^*})}$. In many cases, it can be checked that the rates for the two composition structures

$$f = \underbrace{h_q \circ \ldots \circ h_{i^*}}_{=:H_1} \circ \underbrace{h_{i^*-1} \circ \ldots \circ h_0}_{=:H_0} = H_1 \circ H_0$$

are the same. In all these cases, it is therefore sufficient to consider the simpler composition structure $H_1 \circ H_0$ only.

The next lemma provides conditions such that the number of compositions q can be reduced by one.

Lemma 13 *Suppose that f is a function with composition structure $\eta = (q, \mathbf{d}, \mathbf{t}, \mathcal{S}, \boldsymbol{\beta})$ and assume that $\beta_+ = K = 1$. If there exists an index $j \in \{1, \ldots, q\}$ with $t_j = t_{j-1} = 1$, then, f can also be written as a function with composition structure $\eta' := (q-1, \mathbf{d}_{-j}, \mathbf{t}_{-j}, \mathcal{S}_{-j}, \boldsymbol{\beta}')$, where $\mathbf{d}_{-j}, \mathbf{t}_{-j}, \mathcal{S}_{-j}$ denote $\mathbf{d}, \mathbf{t}, \mathcal{S}$ with entries $d_j, t_j, \mathcal{S}_j$ removed, respectively, and $\boldsymbol{\beta}' := (\beta_0, \ldots, \beta_{j-2}, \beta_{j-1}\beta_j, \beta_{j+1}, \ldots, \beta_q)$. Moreover the induced posterior contraction rates agree, that is, $\mathfrak{r}_n(\eta) = \mathfrak{r}_n(\eta')$.*

The discussion above calls for some notion of equivalence classes on composition structures. Two composition structures are said to be equivalent if one of them can be reduced to the other and both lead to the same rate $\mathfrak{r}_n(\eta)$. It is enough to assign prior mass to only one composition structure in each of the equivalence classes. The equivalence classes can be computed a priori.

Even for large sample sizes, it makes little difference whether the contraction rate is $n^{-1/10}$ or $n^{-1/11}$, say. While compositional structure can lead to much faster contraction rates, the biggest gains occur for structures with small q and small t_i. This indicates that if we are willing to loose a bit in the contraction rates, the number of candidate composition structures can be significantly further reduced.

An alternative is to combine posterior sampling with a greedy model selection method. By first sampling from small composition structures, we can successively increase the complexity of the composition structures. In each iteration, we keep the composition structures with the most pronounced posterior concentration.

Additional information about the underlying composition structure can be incorporated as well. In nonparametric statistics, the generalized additive models and the single index model are two examples of widely studied structural assumptions imposed on the target function. Both constraints can also be rewritten as compositional constraint of the form $f = h_q \circ \ldots \circ h_0$ and correspond therefore to specific compositional structures, more details are given by Schmidt-Hieber (2020). This means that if we suspect that those structural assumptions represent the true regression function, we can put higher (or all) prior weight on the corresponding composition structures.

Conditioning of Gaussian Processes. Step 1 of the deep Gaussian process prior construction requires to condition Gaussian processes to the unitary ball $\mathbb{B}_\infty(1)$ and some enlargement of a Hölder ball. Following Section 4 by Swiler et al. (2020), conditioning the sample paths to the unitary ball is a bound constraint where the range of the process is restricted to some compact interval. One proposed method is to add a warping function to the output of the Gaussian process in such a way to restrict the range to the desired interval, introducing nonlinear layers when dealing with compositions. Another solution involves discrete constraints using truncated Gaussian distributions. This requires an expensive rejection sampling procedure that has been recently made more efficient by Ray et al. (2020); Lopez Lopera (2019). An alternative solution involves constrained maximum likelihood optimization. This approach has been proposed by Pensoneault et al. (2020) to enforce non-negativity but can be easily extended to other inequality constraints. Another method, proposed by Maatouk and Bay (2017) is based on finite dimensional approximations of Gaussian processes.

Restricting the sample paths of Gaussian processes to neighborhoods of a Hölder ball is considerably more challenging from a computational point of view. Lemma 10 shows that for truncated wavelet processes this constraint is not very restrictive.

Variational Bayes. The variational Bayes approach introduced by Titsias (2009) shows how to approximate the posterior distribution induced by a Gaussian process prior. Nieman et al. (2021) gives contraction rates for such an approximation and shows that minimax rates are still achievable even when the KL divergence between the true posterior and the variational posterior does not vanish in the limit $n \to \infty$. This approach has been extended by Damianou and Lawrence (2013). Conditionally on a fixed composition structure η, let $\Pi(\cdot|\mathbf{X}, \mathbf{Y}, \eta)$ denote the posterior distribution. This is the same as putting all mass of the model selection prior π on η. Therefore, sampling from $\Pi(\cdot|\mathbf{X}, \mathbf{Y}, \eta)$ does not involve any model selection. Damianou and Lawrence (2013) proposes a procedure to sample from an approximation of the distribution $\Pi(\cdot|\mathbf{X}, \mathbf{Y}, \eta)$ arising from a DGP prior $\Pi(\cdot|\eta)$ obtained by composing square-exponential processes (radial basis function kernels). The same procedure has been reformulated by Cutajar et al. (2017) in terms of random features expansions.

To sample approximately from the full posterior $\Pi(\cdot|\mathbf{X}, \mathbf{Y}) = \int \Pi(\cdot|\mathbf{X}, \mathbf{Y}, \eta)\pi(\eta)\, d\eta$ for our hierarchical prior, one can first sample a compositional structure $\eta \sim \pi(\eta)$ and afterwards use the method from Damianou and Lawrence (2013) to approximate the distribution $\Pi(\cdot|\mathbf{X}, \mathbf{Y}, \eta)$. Let $Q(\cdot|\mathbf{X}, \mathbf{Y}, \eta)$ denote the distribution of the approximation and $Q(\cdot|\mathbf{X}, \mathbf{Y})$ be the induced approximation of the posterior $\Pi(\cdot|\mathbf{X}, \mathbf{Y})$. If for the variational Bayes step, we have some approximation guarantee of the form $\sup_\eta \|\Pi(\cdot|\mathbf{X}, \mathbf{Y}, \eta) - Q(\cdot|\mathbf{X}, \mathbf{Y}, \eta)\|_{TV} < \gamma$, then also $\|\Pi(\cdot|\mathbf{X}, \mathbf{Y}) - Q(\cdot|\mathbf{X}, \mathbf{Y})\|_{TV} < \gamma$.

Acknowledgments

We would like to thank Joris Bierkens and Ismaël Castillo for helpful discussions. We are extremely grateful to the Associate Editor and three anonymous referees for their constructive comments and suggestions that resulted in a major improvement of the work. The research is part of the project *Nonparametric Bayes for high-dimensional models: contraction, credible sets, computations* (with project number 613.001.604) financed by the Dutch Research Council (NWO).

Appendix A. Proofs

We provide here the proofs for the results presented in the main sections, together with auxiliary material.

A.1 Proofs for Section 2

Lemma 14 *If $f(x_1,\ldots,x_r)=g_1(x_1)\cdot\ldots\cdot g_r(x_r)$ with $g_1,\ldots,g_r\in\mathcal{C}_1^{\beta}(K)$, then there exists a finite K' such that $f\in\mathcal{C}_r^{\beta}(K')$.*

Proof [Proof of Lemma 14] Without loss of generality, assume $K\geq 1$. For any $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_r)\in\mathbb{N}^r$ and $\mathbf{x}=(x_1,\ldots,x_r)$, we have $\partial^{\boldsymbol{\alpha}}f(\mathbf{x})=\partial^{\alpha_1}g_1(x_1)\cdots\partial^{\alpha_r}g_r(x_r)$. By assumption, $\sum_{\alpha_j\leq\lfloor\beta\rfloor}\|\partial^{\alpha_j}g_j\|_\infty\leq K$ for all $j=1,\ldots,r$, thus

$$\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|\leq\lfloor\beta\rfloor}\|\partial^{\boldsymbol{\alpha}}f\|_\infty\leq\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|\leq\lfloor\beta\rfloor}\left(\prod_{j=1}^r\|\partial^{\alpha_j}g_j\|_\infty\right)\leq\prod_{j=1}^r\left(\sum_{\alpha_j\leq\lfloor\beta\rfloor}\|\partial^{\alpha_j}g_j\|_\infty\right)\leq K^r. \qquad (25)$$

For real numbers $a_1,\ldots,a_r,b_1,\ldots,b_r$, we can expand the difference of the products by a telescoping sum, finding that $a_1\cdot\ldots\cdot a_r-b_1\cdot\ldots\cdot b_r=\sum_{j=1}^r(\prod_{s=1}^{j-1}a_s)(a_j-b_j)\prod_{t=j+1}^r b_t$. Therefore, $|a_1\cdot\ldots\cdot a_r-b_1\cdot\ldots\cdot b_r|\leq(\prod_{s=1}^r|a_s|\vee|b_s|\vee 1)\sum_{j=1}^r|a_j-b_j|$. Applying this to $\partial^{\boldsymbol{\alpha}}f(\mathbf{x})=\partial^{\alpha_1}g_1(x_1)\cdots\partial^{\alpha_r}g_r(x_r)$ and $\partial^{\boldsymbol{\alpha}}f(\mathbf{y})=\partial^{\alpha_1}g_1(y_1)\cdots\partial^{\alpha_r}g_r(y_r)$ gives

$$\sup_{\substack{\mathbf{x},\mathbf{y}\in[-1,1]^r\\ \mathbf{x}\neq\mathbf{y}}}\frac{|\partial^{\boldsymbol{\alpha}}f(\mathbf{x})-\partial^{\boldsymbol{\alpha}}f(\mathbf{y})|}{|\mathbf{x}-\mathbf{y}|_\infty^{\beta-\lfloor\beta\rfloor}}\leq\left(\prod_{j=1}^r(\|\partial^{\alpha_j}g_j\|_\infty\vee 1)\right)\sum_{j=1}^r\sup_{\substack{x_j,y_j\in[-1,1]\\ x_j\neq y_j}}\frac{|\partial^{\alpha_j}g_j(x_j)-\partial^{\alpha_j}g_j(x_j)|}{|x_j-y_j|^{\beta-\lfloor\beta\rfloor}}.$$

For any $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_r)$ such that $|\boldsymbol{\alpha}|=\lfloor\beta\rfloor$, the function $\partial^{\alpha_j}g_j$ belongs to $\mathcal{C}_r^{\beta-\alpha_j}(K)$ by assumption. Therefore, $\sup_{x\neq y\in[-1,1]}|\partial^{\alpha_j}g_j(x)-\partial^{\alpha_j}g_j(y)|/|x-y|^{\beta-\lfloor\beta\rfloor}\leq K$. With the bound in (25) and $K\geq 1$, we obtain

$$\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor\beta\rfloor}\sup_{\substack{\mathbf{x},\mathbf{y}\in[-1,1]^r\\ \mathbf{x}\neq\mathbf{y}}}\frac{|\partial^{\boldsymbol{\alpha}}f(\mathbf{x})-\partial^{\boldsymbol{\alpha}}f(\mathbf{y})|}{|\mathbf{x}-\mathbf{y}|_\infty^{\beta-\lfloor\beta\rfloor}}\leq rK\cdot\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor\beta\rfloor}\left(\prod_{j=1}^r(\|\partial^{\alpha_j}g_j\|_\infty\vee 1)\right)\leq rK^{r+1}.$$

The proof is complete by taking $K'\geq 2rK^r+2^{\beta-\lfloor\beta\rfloor}rK^{r+1}$.

■

Proof [Proof of Lemma 1] Given $q,\mathbf{d}$, the components of $\mathbf{t}=(t_0,t_1,\ldots,t_q)$ are drawn independently from $\mathrm{Unif}\{1,\ldots,d_i\}$. Given $q,\mathbf{d},\mathbf{t}$, $\mathcal{S}=(\mathcal{S}_0,\mathcal{S}_1,\ldots,\mathcal{S}_q)$ with $\mathcal{S}_i=(\mathcal{S}_{i1},\ldots,\mathcal{S}_{id_{i+1}})$ are drawn independently from $\mathrm{Unif}\{1,\ldots,\binom{d_i}{t_i}\}$. By construction,

$$\begin{aligned}&\int\sqrt{\gamma(\eta)}\,d\eta\\ &=\sum_{\lambda}\sqrt{(\beta_+-\beta_-)^{q+1}}\sqrt{\gamma(\lambda)}\\ &=\sum_q\sqrt{(\beta_+-\beta_-)^{q+1}}\sqrt{\gamma(q)}\sum_{d_1,\ldots,d_q}\prod_{j=1}^q\sqrt{\gamma(d_j|q)}\sum_{t_0,\ldots,t_q}\sqrt{\prod_{i=0}^q\frac{1}{d_i}}\sum_{\mathcal{S}_0,\ldots,\mathcal{S}_q}\sqrt{\prod_{i=0}^q\prod_{j=1}^{d_{i+1}}\frac{1}{\binom{d_i}{t_i}}}.\end{aligned}$$

Because of the inequality $\binom{n}{k} \leq \sum_{q=0}^{n} \binom{n}{q} = 2^n$, we have

$$\sum_{\mathcal{S}_0,\ldots,\mathcal{S}_q} \sqrt{\prod_{i=0}^{q} \prod_{j=1}^{d_{i+1}} \frac{1}{\binom{d_i}{t_i}}} = \prod_{i=0}^{q} \prod_{j=1}^{d_{i+1}} \binom{d_i}{t_i} \sqrt{\prod_{i=0}^{q} \prod_{j=1}^{d_{i+1}} \frac{1}{\binom{d_i}{t_i}}} = \sqrt{\prod_{i=0}^{q} \binom{d_i}{t_i}^{d_{i+1}}} \leq \sqrt{\prod_{i=0}^{q} 2^{d_i d_{i+1}}}.$$

The sum $\sum_{t_0,\ldots,t_q}$ is over $\prod_{i=0}^{q} d_i$ many terms. Using that $2^{d_i d_{i+1}} \leq 2^{d_i^2/2} 2^{d_{i+1}^2/2}$, the previous displays combined give

$$\int \sqrt{\gamma(\eta)}\, d\eta \leq \sqrt{d 2^{d^2}} \sum_q \sqrt{(\beta_+ - \beta_-)^{q+1}} \sqrt{\gamma(q)} \sum_{d_1,\ldots,d_q} \prod_{j=1}^{q} \sqrt{\gamma(d_j|q) d_j 2^{d_j^2}}. \tag{26}$$

Cauchy-Schwarz inequality gives

$$\sum_{d_j} \sqrt{\gamma(d_j|q) d_j 2^{d_j^2}} \leq \sqrt{\sum_{d_j} \gamma(d_j|q) d_j^3 2^{d_j^2}} \sqrt{\sum_{d_j} \frac{1}{d_j^2}} = \sqrt{\frac{\pi^6}{6} \mathbb{E}_{d_1|q}[d_1^3 2^{d_1^2}]} =: \kappa$$

Due to $\sum_{d_1,\ldots,d_q} \prod_{j=1}^{q} = \prod_{j=1}^{q} \sum_{d_j}$, (26) combined with Cauchy-Schwarz inequality yields

$$\begin{aligned}
\int \sqrt{\gamma(\eta)}\, d\eta &\leq \sqrt{d 2^{d^2}} \sum_q \sqrt{(\beta_+ - \beta_-)^{q+1}} \sqrt{\gamma(q)} \kappa^q \\
&\leq \sqrt{d 2^{d^2}} \sqrt{\sum_q (\beta_+ - \beta_-)^{2q+2} \gamma(q) e^q \kappa^{2q}} \sqrt{\sum_q e^{-q}} \\
&= \sqrt{d 2^{d^2}} \sqrt{\mathbb{E}_q\big[(\beta_+ - \beta_-)^{2q+2} e^q \kappa^{2q}\big]} \frac{1}{\sqrt{e-1}} \\
&< \infty,
\end{aligned}$$

using for the last step that by assumption $\mathbb{E}_q[A^q] < \infty$, for all $A > 0$. The proof is complete. ■

A.2 Proofs for Section 4

Information geometry in the nonparametric regression model. The following results are fairly standard in the nonparametric Bayes literature. As we are aiming for a self-contained presentation of the material, these facts are reproduced here. Let P_f be the law of *one* observation $(\mathbf{X}_i, Y_i)$. The Kullback-Leibler divergence in the nonparametric regression model is

$$\mathrm{KL}\left(P_f, P_g\right) = \frac{1}{2} \int (f(\mathbf{x}) - g(\mathbf{x}))^2 d\mu(\mathbf{x}) \leq \frac{1}{2} \|f - g\|_{L^\infty([-1,1]^d)}^2$$

with μ the distribution of the covariates $\mathbf{X}_1$. Using that $\mathbb{E}_f[\log dP_f/dP_g] = \mathrm{KL}(P_f, P_g)$, $\mathrm{Var}(Z) \leq \mathbb{E}[Z^2]$, $\mathbb{E}_f[Y|\mathbf{X}] = f(\mathbf{X})$ and $\mathbb{E}_f[(Y - f(\mathbf{X}))^2|\mathbf{X}] = 1$, we also have that

$$\begin{aligned}
V_2(P_f, P_g) &:= \mathbb{E}_f\Big[\Big|\log\frac{dP_f}{dP_g} - \mathrm{KL}(P_f, P_g)\Big|^2\Big] \\
&\leq \mathbb{E}_f\Big[\Big|\log\frac{dP_f}{dP_g}\Big|^2\Big] \\
&= \mathbb{E}_f\Big[\Big(Y\big(f(\mathbf{X}) - g(\mathbf{X})\big) - \frac{1}{2}f(\mathbf{X})^2 + \frac{1}{2}g(\mathbf{X})^2\Big)^2\Big] \\
&= \mathbb{E}_f\Big[\Big(Y\big(f(\mathbf{X}) - g(\mathbf{X})\big) - \frac{1}{2}\big(f(\mathbf{X}) - g(\mathbf{X})\big)\big(f(\mathbf{X}) + g(\mathbf{X})\big)\Big)^2\Big] \\
&= \mathbb{E}_f\Big[\big(f(\mathbf{X}) - g(\mathbf{X})\big)^2\Big(Y - \frac{1}{2}\big(f(\mathbf{X}) + g(\mathbf{X})\big)\Big)^2\Big] \\
&= \int \big(f(\mathbf{x}) - g(\mathbf{x})\big)^2 + \frac{1}{4}\big(f(\mathbf{x}) - g(\mathbf{x})\big)^4 \, d\mu(\mathbf{x}).
\end{aligned}$$

In particular, for $\varepsilon \leq 1$, $\|f - g\|_\infty \leq \varepsilon/2$ implies that $V_2(P_f, P_g) \leq \varepsilon^2$ and therefore

$$B_2(P_f, \varepsilon) = \Big\{g : \mathrm{KL}(P_f, P_g) < \varepsilon^2, V_2(P_f, P_g) < \varepsilon^2\Big\} \supseteq \Big\{g : \|f - g\|_\infty \leq \frac{\varepsilon}{2}\Big\}. \tag{27}$$

We derive posterior contraction rates for the Hellinger distance. This can then be related to the $\|\cdot\|_{L^2(\mu)}$-norm as explained below. Using the moment generating function of a standard normal distribution, the Hellinger distance for one observation $(\mathbf{X}, Y)$ becomes

$$\begin{aligned}
d_H(P_f, P_g) = 1 - \int \sqrt{dP_f dP_g} &= 1 - \int \sqrt{dP_f/dP_g}\, dP_g \\
&= 1 - \mathbb{E}_g\Big[e^{\frac{1}{4}(Y - g(\mathbf{X}))^2 - \frac{1}{4}(Y - f(\mathbf{X}))^2}\Big] \\
&= 1 - \mathbb{E}\Big[\mathbb{E}_g\Big[e^{\frac{1}{2}(Y - g(\mathbf{X}))(f(\mathbf{X}) - g(\mathbf{X}))}\Big|\mathbf{X}\Big]e^{-\frac{1}{4}(f(\mathbf{X}) - g(\mathbf{X}))^2}\Big] \\
&= 1 - \mathbb{E}\Big[e^{-\frac{1}{8}(f(\mathbf{X}) - g(\mathbf{X}))^2}\Big] \\
&= 1 - \int e^{-\frac{1}{8}(f(\mathbf{x}) - g(\mathbf{x}))^2}\, d\mu(\mathbf{x}).
\end{aligned}$$

Since $1 - e^{-x} \leq x$ and μ is a probability measure, we have that $d_H(P_f, P_g) \leq \frac{1}{8}\int (f(\mathbf{x}) - g(\mathbf{x}))^2 \, d\mu(\mathbf{x})$. Due to $1 - e^{-x} \geq e^{-x}x$, we also find

$$d_H(P_f, P_g) \geq \frac{e^{-Q^2/2}}{8}\int \big(f(\mathbf{x}) - g(\mathbf{x})\big)^2 \, d\mu(\mathbf{x}), \quad \text{for all } f, g, \text{ with } \|f\|_\infty, \|g\|_\infty \leq Q. \tag{28}$$

By Proposition D.8 by Ghosal and van der Vaart (2017), for any f, g, there exists a test such that $\mathbb{E}_f\phi \leq \exp(-\frac{n}{8}d_H(P_f, P_g)^2)$ and $\sup_{h: d_H(P_h, P_g) < d_H(P_f, P_g)/2}\mathbb{E}_h[1 - \phi] \leq \exp(-\frac{n}{8}d_H(P_f, P_g)^2)$. This means that for the Hellinger distance, the test condition in (8.2) by Ghosal and van der Vaart (2017) holds for $\xi = 1/2$.

Function spaces. The next result shows that the Hölder-balls defined in this paper are nested.

Lemma 15 *If $0 < \beta' \leq \beta$, then, for any positive integer r and any $K > 0$, we have $\mathcal{C}_r^{\beta}(K) \subseteq \mathcal{C}_r^{\beta'}(K)$.*

Proof [Proof of Lemma 15] If $\lfloor \beta' \rfloor = \lfloor \beta \rfloor$ the embedding follows from the definition of the Hölder-ball and the fact that $\sup_{\mathbf{x},\mathbf{y}\in[-1,1]^r} |\mathbf{x}-\mathbf{y}|_\infty^{\beta-\beta'} = 2^{\beta-\beta'}$. If $\lfloor \beta' \rfloor < \lfloor \beta \rfloor$, it remains to prove $\mathcal{C}^{\beta}(K) \subseteq \mathcal{C}^{\lfloor \beta' \rfloor + 1}(K)$. This follows from first order Taylor expansion,

$$2 \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor \beta' \rfloor} \sup_{\substack{\mathbf{x},\mathbf{y}\in[-1,1]^r \\ \mathbf{x}\neq\mathbf{y}}} \frac{|\partial^{\boldsymbol{\alpha}} f(\mathbf{x}) - \partial^{\boldsymbol{\alpha}} f(\mathbf{y})|}{|\mathbf{x}-\mathbf{y}|_\infty} \leq 2 \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor \beta' \rfloor} \big\| \big|\nabla(\partial^{\boldsymbol{\alpha}} f)\big|_1 \big\|_\infty$$
$$\leq 2r \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor \beta' \rfloor+1} \big\|\partial^{\boldsymbol{\alpha}} f\big\|_\infty,$$

and the definition of the Hölder-ball in (5). ■

The following is a slight variation of Lemma 3 by Schmidt-Hieber (2020).

Lemma 16 *Let $h_{ij} : [-1,1]^{t_i} \to [-1,1]$ be as in (2). Assume that, for some $K \geq 1$ and $\eta_i \geq 0$, $|h_{ij}(\mathbf{x}) - h_{ij}(\mathbf{y})|_\infty \leq \eta_i + K|\mathbf{x}-\mathbf{y}|_\infty^{\beta_i \wedge 1}$ for all $\mathbf{x},\mathbf{y} \in [-1,1]^{t_i}$. Then, for any functions $\widetilde{h}_i = (\widetilde{h}_{ij})_j^\top$ with $\widetilde{h}_{ij} : [-1,1]^{t_i} \to [-1,1]$,*

$$\big\| h_q \circ \ldots \circ h_0 - \widetilde{h}_q \circ \ldots \circ \widetilde{h}_0 \big\|_{L^\infty[-1,1]^d} \leq K^q \sum_{i=0}^{q} \eta_i^{\alpha_i} + \big\| |h_i - \widetilde{h}_i|_\infty \big\|_\infty^{\alpha_i}.$$

with $\alpha_i = \prod_{\ell=i+1}^{q} \beta_\ell \wedge 1$.

Proof [Proof of Lemma 16.] We prove the assertion by induction over q. For $q = 0$, the result is trivially true. Assume now that the statement is true for a positive integer k. To show that the assertion also holds for $k+1$, define $H_k = h_k \circ \ldots \circ h_0$ and $\widetilde{H}_k = \widetilde{h}_k \circ \ldots \circ \widetilde{h}_0$. By triangle inequality,

$$\begin{aligned}
&\big|h_{k+1} \circ H_k(\mathbf{x}) - \widetilde{h}_{k+1} \circ \widetilde{H}_k(\mathbf{x})\big|_\infty \\
&\leq |h_{k+1} \circ H_k(\mathbf{x}) - h_{k+1} \circ \widetilde{H}_k(\mathbf{x})\big|_\infty + |h_{k+1} \circ \widetilde{H}_k(\mathbf{x}) - \widetilde{h}_{k+1} \circ \widetilde{H}_k(\mathbf{x})\big|_\infty \\
&\leq \eta_{k+1} + K\big|H_k(\mathbf{x}) - \widetilde{H}_k(\mathbf{x})\big|_\infty^{\beta_{k+1}\wedge 1} + \||h_{k+1} - \widetilde{h}_{k+1}|_\infty\|_\infty.
\end{aligned}$$

Together with the induction hypothesis and the inequality $(y+z)^\alpha \leq y^\alpha + z^\alpha$ which holds for all $y, z \geq 0$ and all $\alpha \in [0,1]$, the induction step follows. ■

The next result is a corollary of Theorem 8.9 by Ghosal and van der Vaart (2017).

Lemma 17 *Denote the data by $\mathcal{D}_n$ and the (generic) posterior by $\Pi(\cdot|\mathcal{D}_n)$. Let $(A_n)_n$ be a sequence of events and $B_2(P_{f^*}, \varepsilon)$ as in (27). Assume that*

$$e^{2na_n^2} \frac{\Pi(A_n)}{\Pi(B_2(P_{f^*}, a_n))} \xrightarrow{n\to\infty} 0, \tag{29}$$

for some positive sequence $(a_n)_n$. *Then,*

$$\mathbb{E}_{f^*}\big[\Pi(A_n|\mathcal{D}_n)\big] \xrightarrow{n\to\infty} 0,$$

where $\mathbb{E}_{f^*}$ *is the expectation with respect to* P_{f^*}.

We can now prove Theorem 2.

Proof [Proof of Theorem 2] By definition (16), the quantity $\Pi\big(\eta \notin \mathcal{M}_n(C)|\mathbf{X},\mathbf{Y}\big)$ denotes the posterior mass of the functions whose models are in the complement of $\mathcal{M}_n(C)$. In view of Lemma 17, it is sufficient to show condition (29) for $A_n = \{\eta \notin \mathcal{M}_n(C)\} =: \mathcal{M}_n^c(C)$ and a_n proportional to $\varepsilon_n(\eta^*)$. We now prove that

$$e^{2na_n^2}\frac{\int_{\mathcal{M}_n^c(C)}\pi(\eta)\,d\eta}{\Pi(B_2(P_{f^*},a_n))} \to 0, \tag{30}$$

for $a_n = 4K^{q^*}(q^*+1)Q\varepsilon_n(\eta^*)$ and Π the deep Gaussian process prior. The next result deals with the lower bound on the denominator. For any hypercube I, we introduce the notation $\mathrm{diam}(I) := \sup_{\boldsymbol{\beta},\boldsymbol{\beta}'\in I}|\boldsymbol{\beta}-\boldsymbol{\beta}'|_\infty$.

Lemma 18 *Let* Π *be a DGP prior satisfying the assumptions of Theorem 2. With* Q *the universal constant from Assumption 2,* $R^* := 4K^{q^*}(q^*+1)Q$ *and sufficiently large sample size* n, *we have*

$$\Pi\Big(B_2\big(P_{f^*},2R^*\varepsilon_n(\eta^*)\big)\Big) \ge e^{-|\mathbf{d}^*|_1 Q^2 n\varepsilon_n(\eta^*)^2}\pi(\lambda^*,I_n^*),$$

where $I_n^* := \{\boldsymbol{\beta}=(\beta_0,\ldots,\beta_{q^*}) : \beta_i \in [\beta_i^*-b_n,\beta_i^*], \forall i\}$ *and* $b_n := 1/\log^2 n$.

Proof [Proof of Lemma 18] By construction (27), the set $B_2(P_{f^*},2R^*\varepsilon_n(\eta^*))$ is a superset of $\{g : \|f^*-g\|_\infty \le R^*\varepsilon_n(\eta^*)\}$, so that

$$\Pi\Big(B_2\big(P_{f^*},2R^*\varepsilon_n(\eta^*)\big)\Big) \ge \Pi\Big(f^* + \mathbb{B}_\infty\left(R^*\varepsilon_n(\eta^*)\right)\Big).$$

We then localize the probability in the latter display in the neighborhood I_n^* around the true $\boldsymbol{\beta}^* = (\beta_0^*,\ldots,\beta_q^*)$. Since $\boldsymbol{\beta}^* \in I(\lambda^*) = (\beta_-,\beta_+)^{q^*+1}$, we can always choose n large enough such that $I_n^* \subseteq I(\lambda^*)$. With $f^* = h_{q^*}^*\circ\ldots\circ h_0^*$ and $R^* = 4K^{q^*}(q^*+1)Q$,

$$\begin{aligned}&\Pi\Big(\big\{g : \|f^*-g\|_\infty \le R^*\varepsilon_n(\eta^*)\big\}\Big)\\ &\quad\ge \int_{I_n^*}\mathbb{P}\Big(\big\|h_{q^*}^*\circ\ldots\circ h_0^* - G_{q^*}^{(\lambda^*,\boldsymbol{\beta})}\circ\ldots\circ G_0^{(\lambda^*,\boldsymbol{\beta})}\big\|_\infty \le R^*\varepsilon_n(\eta^*)\Big)\pi(\lambda^*,\boldsymbol{\beta})d\boldsymbol{\beta}.\end{aligned} \tag{31}$$

Fix any $\boldsymbol{\beta} \in I_n^*$. Both $G_{ij}^{(\lambda^*,\boldsymbol{\beta})} = \overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})}\circ(\cdot)_{\mathcal{S}_{ij}^*}$ and $h_{ij}^* = \overline{h}_{ij}^*\circ(\cdot)_{\mathcal{S}_{ij}^*}$ map $[-1,1]^{d_i^*}$ into $[-1,1]$. They also depend on the same subset of variables $\mathcal{S}_{ij}^*$. By construction, the process $\overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})}$ is an independent copy of the conditioned Gaussian process $\widetilde{G}^{(\beta_i,t_i^*)}|\{\widetilde{G}^{(\beta_i,t_i^*)} \in \mathcal{D}_i(\lambda^*,\boldsymbol{\beta},K)\}$. The function $\overline{h}_{ij}^*$ belongs by definition to the space $\mathcal{C}_{t_i^*}^{\beta_i^*}(K)$ and satisfies $|\overline{h}_{ij}^*(\mathbf{x}) - \overline{h}_{ij}^*(\mathbf{y})| \le$

$K|\mathbf{x}-\mathbf{y}|_\infty^{\beta_i^*\wedge 1}$ for all $\mathbf{x},\mathbf{y}\in[-1,1]^{t_i^*}$ and all $i=0,\dots,q^*$; $j=1,\dots,d_{i+1}^*$. By Lemma 16 with $\eta_i=0$, we thus find

$$\big\|f^*-G^{(\lambda^*,\boldsymbol{\beta})}\big\|_\infty\le K^{q^*}\sum_{i=0}^{q^*}\max_{j=1,\dots,d_{i+1}^*}\|\overline{h}_{ij}^*-\overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})}\|_\infty^{\alpha_i^*},$$

where $\alpha_i^*=\prod_{\ell=i+1}^{q^*}\beta_\ell^*\wedge 1$. Since $\alpha_i^*\ge\alpha_i=\prod_{\ell=i+1}^{q^*}(\beta_\ell\wedge 1)$ for $\boldsymbol{\beta}\in I_n^*$, if $\|\overline{h}_{ij}^*-\overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})}\|_\infty$ is smaller than one, the latter display is bounded above by

$$\big\|f^*-G^{(\lambda^*,\boldsymbol{\beta})}\big\|_\infty\le K^{q^*}\sum_{i=0}^{q^*}\max_{j=1,\dots,d_{i+1}^*}\|\overline{h}_{ij}^*-\overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})}\|_\infty^{\alpha_i}.$$

Set $\delta_{in}:=\varepsilon_n(\alpha_i,\beta_i,t_i^*)^{1/\alpha_i}$. By Assumption 2 and the definition of $\varepsilon_n(\eta)$ in (12), we have $\delta_{in}\le(Q\varepsilon_n(\eta^*))^{1/\alpha_i}$ and so $4\delta_{in}<1$ since we are also assuming $\varepsilon_n(\eta^*)<1/(4Q)$. If $\|\overline{h}_{ij}^*-\overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})}\|_\infty\le 4\delta_{in}$ for all $i=0,\dots,q^*$ and $j=1,\dots,d_{i+1}^*$ then $\|f^*-G^{(\lambda^*,\boldsymbol{\beta})}\|_\infty\le R^*\varepsilon_n(\eta^*)$. Consequently,

$$\begin{aligned}&\mathbb{P}\Big(\big\|h_{q^*}^*\circ\ldots\circ h_0^*-G_{q^*}^{(\lambda^*,\boldsymbol{\beta})}\circ\ldots\circ G_0^{(\lambda^*,\boldsymbol{\beta})}\big\|_\infty\le R^*\varepsilon_n(\eta^*)\Big)\\&\quad\ge\prod_{i=0}^{q^*}\prod_{j=1}^{d_{i+1}^*}\mathbb{P}\Big(\big\|\overline{h}_{ij}^*-\overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})}\big\|_\infty\le 4\delta_{in}\Big).\end{aligned}\tag{32}$$

We now lower bound the probabilities on the right hand side. If $\overline{h}_{ij}^*\in\mathcal{C}_{t_i^*}^{\beta_i^*}(K)$, then, $(1-2\delta_{in})\overline{h}_{ij}^*\in\mathcal{C}_{t_i^*}^{\beta_i^*}((1-2\delta_{in})K)$ and by the embedding property in Lemma 15, we obtain $(1-2\delta_{in})\overline{h}_{ij}^*\in\mathcal{C}_{t_i^*}^{\beta_i}(K)$.

When $\|(1-2\delta_{in})\overline{h}_{ij}^*-\widetilde{G}^{(\beta_i,t_i^*)}\|_\infty\le 2\delta_{in}$, the Gaussian process $\widetilde{G}^{(\beta_i,t_i^*)}$ is at most $2\delta_{in}$-away from $(1-2\delta_{in})\overline{h}_{ij}^*\in\mathcal{C}_{t_i^*}^{\beta_i}(K)$. Consequently, $\{\|(1-2\delta_{in})\overline{h}_{ij}^*-\widetilde{G}^{(\beta_i,t_i^*)}\|_\infty\le 2\delta_{in}\}\subseteq\{\widetilde{G}^{(\beta_i,t_i^*)}\in\mathcal{D}_i(\lambda^*,\boldsymbol{\beta},K)\}$, where the bound for the unitary ball follows from the triangle inequality. Moreover $\|(1-2\delta_{in})\overline{h}_{ij}^*-\widetilde{G}^{(\beta_i,t_i^*)}\|_\infty\le 2\delta_{in}$ implies $\|\overline{h}_{ij}^*-\widetilde{G}^{(\beta_i,t_i^*)}\|_\infty\le 4\delta_{in}$. The concentration function property in Lemma I.28 by Ghosal and van der Vaart (2017) gives

$$\mathbb{P}\Big(\big\|(1-\delta_{in})\overline{h}_{ij}^*-\widetilde{G}^{(\beta_i,t_i^*)}\big\|_\infty\le 2\delta_{in}\Big)\ge\exp\Big(-\varphi^{(\beta_i,t_i^*,K)}\big(\delta_{in}\big)\Big).$$

Together with the concentration function inequality in (10), we find

$$\begin{aligned}\mathbb{P}\Big(\big\|\overline{h}_{ij}^*-\overline{G}_{ij}^{(\lambda^*,\boldsymbol{\beta})}\big\|_\infty\le 4\delta_{in}\Big)&\ge\frac{\mathbb{P}\big(\|(1-\delta_{in})\overline{h}_{ij}^*-\widetilde{G}^{(\beta_i,t_i^*)}\|_\infty\le 2\delta_{in}\big)}{\mathbb{P}\big(\widetilde{G}^{(\beta_i,t_i^*)}\in\mathcal{D}_i(\lambda^*,\boldsymbol{\beta},K)\big)}\\&\ge\mathbb{P}\Big(\big\|(1-\delta_{in})\overline{h}_{ij}^*-\widetilde{G}^{(\beta_i,t_i^*)}\big\|_\infty\le 2\delta_{in}\Big)\\&\ge\exp\Big(-\varphi^{(\beta_i,t_i^*,K)}\big(\delta_{in}\big)\Big)\\&\ge\exp\Big(-n\varepsilon_n(\alpha_i,\beta_i,t_i^*)^2\Big)\\&\ge\exp\Big(-Q^2n\varepsilon_n(\eta^*)^2\Big),\end{aligned}$$

where Q is the universal constant from Assumption 2. With $|\mathbf{d}^*|_1 = 1 + \sum_{j=0}^{q^*} d_j^*$, (31), (32) and the previous display we recover the claim. ■

The latter result shows that

$$\Pi\Big(B_2\big(P_{f^*}, 4K^{q^*}(q^*+1)Q\varepsilon_n(\eta^*)\big)\Big) \geq e^{-|\mathbf{d}^*|_1 Q^2 n\varepsilon_n(\eta^*)^2} \int_{I_n^*} \pi(\lambda^*, \boldsymbol{\beta})\, d\boldsymbol{\beta}, \tag{33}$$

with $I_n^* = \{\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{q^*}) : \beta_i \in [\beta_i^* - 1/\log^2 n, \beta_i^*], \forall i\}$. Recall that, by construction (13), $\pi(\eta) \propto e^{-\Psi_n(\eta)}\gamma(\eta)$ with $\Psi_n(\eta) = n\varepsilon_n(\eta)^2 + e^{e^{|\mathbf{d}|_1}}$. For any $\boldsymbol{\beta} \in I_n^*$, we have $\Psi_n(\lambda^*, \boldsymbol{\beta}) \leq \Psi_n(\lambda^*, \boldsymbol{\beta}^*)$, and Assumption 1 gives $\gamma(\lambda^*, \boldsymbol{\beta}) = \gamma(\lambda^*)\gamma(\boldsymbol{\beta}|\lambda^*)$ with $\gamma(\lambda^*) > 0$ independent of n and $\gamma(\cdot|\lambda^*)$ the uniform distribution over $I(\lambda^*) = [\beta_-, \beta_+]^{q^*+1}$. Thus, $\gamma(I_n^*|\lambda^*) = |I_n^*|/|I(\lambda^*)|$ and $|I_n^*| = (1/\log^2 n)^{q^*+1}$, so that by Assumption 2

$$\begin{aligned}\frac{\pi(\eta)}{\int_{I_n^*} \pi(\lambda^*, \boldsymbol{\beta})\, d\boldsymbol{\beta}} &\leq \frac{e^{Q^2 n\varepsilon_n(\eta^*)^2 + e^{e^{|\mathbf{d}^*|_1}} - \Psi_n(\eta)}\gamma(\eta)}{\gamma(\lambda^*)\gamma(I_n^*|\lambda^*)} \\ &= \frac{|I(\lambda^*)|}{\gamma(\lambda^*)} e^{Q^2 n\varepsilon_n(\eta^*)^2 + e^{e^{|\mathbf{d}^*|_1}} - \Psi_n(\eta)} e^{2(q^*+1)\log\log n}\gamma(\eta).\end{aligned} \tag{34}$$

The quantities $\gamma(\lambda^*)$, $|I(\lambda^*)| = (\beta_+ - \beta_-)^{q^*+1}$ and $\exp(e^{e^{|\mathbf{d}^*|_1}})$ are constants independent of n. We can finally verify condition (30) by showing that, with $a_* = 4K^{q^*}(q^*+1)Q$,

$$e^{2a_*^2 n\varepsilon_n(\eta^*)^2} e^{Q^2 n\varepsilon_n(\eta^*)^2 + 2(q^*+1)\log\log n} \sum_{\lambda} \int_{\boldsymbol{\beta}:(\lambda,\boldsymbol{\beta})\notin\mathcal{M}_n(C)} e^{-\Psi_n(\eta)}\gamma(\eta)\, d\boldsymbol{\beta} \to 0.$$

By the lower bound in Assumption 2 (i), we have $n\varepsilon_n(\eta^*)^2 \geq n\mathfrak{r}_n(\eta^*)^2 \gg 2(q^*+1)\log\log n$, since the quantity $n\mathfrak{r}_n(\eta^*)^2$ is a positive power of n by definition (20). The complement of the set $\mathcal{M}_n(C)$ is the union of $\{\eta : \varepsilon_n(\eta) > C\varepsilon_n(\eta^*)\}$ and $\{\eta : |\mathbf{d}|_1 > \log(2\log n)\}$. Over these sets, by construction (13), we have either $\Psi_n(\eta) > C^2 n\varepsilon_n(\eta^*)^2$ or $\Psi_n(\eta) > n^2$. Therefore, the term $e^{-\Psi_n(\eta)}$ decays faster than either $e^{-C^2 n\varepsilon_n(\eta^*)^2}$ or e^{-n^2}. In the first case, the latter display converges to zero for sufficiently large $C > 0$. In the second case, the latter display converges to zero since $\varepsilon_n(\eta^*) < 1$ and $n\varepsilon_n(\eta^*)^2 < n \ll n^2$. This completes the proof. ■

We need some preliminary notation and results before proving Theorem 3.

Entropy bounds. Previous bounds for the metric entropy of Hölder-balls, e.g. Proposition C.5 by Ghosal and van der Vaart (2017), are of the form $\log \mathcal{N}\big(\delta, \mathcal{C}_r^\beta(K), \|\cdot\|_\infty\big) \leq Q_1(\beta, r, K)\delta^{-r/\beta}$ for some constant $Q_1(\beta, r, K)$ that is hard to control. An exception is Theorem 8 by Bolley (2010) that, however, only holds for $\beta \leq 1$. We derive an explicit bound on the constant $Q_1(\beta, r, K)$ for all $\beta > 0$. The proof is given in Appendix A.2.1.

Lemma 19 *For any positive integer r, any $\beta > 0$ and $0 < \delta < 1$, we have*

$$\begin{aligned}\mathcal{N}\left(\delta, C_r^\beta(K), \|\cdot\|_\infty\right) &\leq \left(\frac{4eK^2 r^\beta}{\delta} + 1\right)^{(\beta+1)^r} \left(2^{\beta+2}eKr^\beta + 1\right)^{4^r(\beta+1)^r r^r (2eK)^{\frac{r}{\beta}}\delta^{-\frac{r}{\beta}}} \\ &\leq e^{Q_1(\beta,r,K)\delta^{-\frac{r}{\beta}}}\end{aligned}$$

with $Q_1(\beta, r, K) := (1 + eK)4^{r+1}(\beta + 3)^{r+1}r^{r+1}(8eK^2)^{r/\beta}$. *For any* $0 < \alpha \leq 1$ *and any sequence* $\delta_n \geq Q_1(\beta, r, K)^{\beta/(2\beta+r)}n^{-\beta\alpha/(2\beta\alpha+r)}$, *we also have*

$$\log \mathcal{N}\left(\delta_n^{1/\alpha}, C_r^\beta(K), \|\cdot\|_\infty\right) \leq n\delta_n^2. \tag{35}$$

Support of DGP prior and local complexity. For any graph $\lambda = (q, \mathbf{d}, \mathbf{t}, \mathcal{S})$ and any $\boldsymbol{\beta} \in [\beta_-, \beta_+]^{q+1}$, denote by $\Theta_n(\lambda, \boldsymbol{\beta}, K)$ the space of functions $f : [-1, 1]^d \to [-1, 1]$ for which there exists a decomposition $f = h_q \circ \ldots \circ h_0$ such that $h_{ij} : [-1, 1]^{d_i} \to [-1, 1]$ and $\overline{h}_{ij} \in \mathcal{D}_i(\lambda, \boldsymbol{\beta}, K)$, for all $i = 0, \ldots, q$; $j = 1, \ldots, d_{i+1}$ and $\mathcal{D}_i(\lambda, \boldsymbol{\beta}, K)$ as defined in (11). Differently speaking

$$\Theta_n(\lambda, \boldsymbol{\beta}, K) := \Theta_{q,n}(\lambda, \boldsymbol{\beta}, K) \circ \cdots \circ \Theta_{0,n}(\lambda, \boldsymbol{\beta}, K) \tag{36}$$

with

$$\Theta_{i,n}(\lambda, \boldsymbol{\beta}, K) := \Big\{h_i : [-1, 1]^{d_i} \to [-1, 1]^{d_{i+1}} : \overline{h}_{ij} \in \mathcal{D}_i(\lambda, \boldsymbol{\beta}, K), j = 1, \ldots, d_{i+1}\Big\}.$$

By construction, the support of the deep Gaussian process $G^{(\eta)} \sim \Pi(\cdot|\eta)$ is contained in $\Theta_n(\lambda, \boldsymbol{\beta}, K)$. For a subset $B \subseteq [\beta_-, \beta_+]^{q+1}$ we also set $\Theta_n(\lambda, B, K) := \cup_{\boldsymbol{\beta} \in B}\Theta_n(\lambda, \boldsymbol{\beta}, K)$. The next lemma provides a bound for the covering number of $\Theta_n(\lambda, B, K)$. Recall that $\operatorname{diam}(B) = \sup_{\boldsymbol{\beta}, \boldsymbol{\beta}' \in B} |\boldsymbol{\beta} - \boldsymbol{\beta}'|_\infty$. We postpone the proof to Appendix A.2.1.

Lemma 20 *Suppose that Assumption 2 holds and let* λ *be a graph such that* $|\mathbf{d}|_1 \leq \log(2 \log n)$. *Let* $B \subseteq [\beta_-, \beta_+]^{q+1}$ *with* $\operatorname{diam}(B) \leq 1/\log^2 n$. *Then, with* $R_n := 5Q(2 \log n)^{1+\log K}$,

$$\sup_{\boldsymbol{\beta} \in B} \frac{\log \mathcal{N}\left(R_n\varepsilon_n(\lambda, \boldsymbol{\beta}), \Theta_n(\lambda, B, K), \|\cdot\|_\infty\right)}{n\varepsilon_n(\lambda, \boldsymbol{\beta})^2} \leq \frac{R_n^2}{25}.$$

Proof [Proof of Theorem 3] For any $\rho > 0$, introduce the complement Hellinger ball $\mathcal{H}^c(f^*, \rho) := \{f : d_H(P_f, P_{f^*}) > \rho\}$. The convergence with respect to the Hellinger distance d_H implies convergence in $L^2(\mu)$ thanks to (28). As a consequence, we show that

$$\sup_{f^* \in \mathcal{F}(\eta^*, K)} \mathbb{E}_{f^*}\Big[\Pi\Big(\mathcal{H}^c\big(f^*, L_n\varepsilon_n(\eta^*)\big)\Big|(\mathbf{X}, \mathbf{Y})\Big)\Big] \to 0,$$

with $L_n := MR_n$, $R_n := 10QC(2 \log n)^{1+\log K}$ and $M > 0$ a sufficiently large universal constant to be determined. With $\mathcal{M}_n(C) = \{\eta : \varepsilon_n(\eta) \leq C\varepsilon_n(\eta^*)\} \cap \{\eta : |\mathbf{d}|_1 \leq \log(2 \log n)\}$ and the notation in (16), we denote by $\Pi(\cdot \cap \mathcal{M}_n(C)|\mathbf{X}, \mathbf{Y})$ the contribution of the composition structures $\mathcal{M}_n(C)$ to the posterior mass.

We denote by $\mathcal{L}_n(C)$ the set of graphs that are realized by some good composition structure, that is,

$$\mathcal{L}_n(C) := \big\{\lambda \in \Lambda : \exists \boldsymbol{\beta} \in I(\lambda),\ \eta = (\lambda, \boldsymbol{\beta}) \in \mathcal{M}_n(C)\big\}. \tag{37}$$

By construction, all graphs in $\mathcal{L}_n(C)$ have bounded dimension $|\mathbf{d}|_1 \leq \log(2 \log n)$ and so part (ii) in Assumption 2 can be applied. For any $\lambda \in \mathcal{L}_n(C)$, partition $I(\lambda) = [\beta_-, \beta_+]^{q+1}$ into hypercubes of diameter $1/\log^2 n$ and let $B_1(\lambda), \ldots, B_{N(\lambda)}(\lambda)$ be the $N(\lambda)$ blocks that contain at least one $\boldsymbol{\beta} \in I(\lambda)$ that is realized by some composition structure in $\mathcal{M}_n(C)$. The blocks may contain also values of $\boldsymbol{\beta}$ for which $(\lambda, \boldsymbol{\beta}) \notin \mathcal{M}_n(C)$. The set $\mathcal{M}_n(C)$ is contained in the enlargement

$$\widetilde{\mathcal{M}}_n(C) := \bigcup_{\lambda \in \mathcal{L}_n(C)} \bigcup_{k=1}^{N(\lambda)} \big(\lambda, B_k(\lambda)\big).$$

Thanks to Theorem 2 and the enlarged set of structures $\widetilde{\mathcal{M}}_n(C)$, it is enough to show, for sufficiently large constants M, C,

$$\sup_{f^* \in \mathcal{F}(\eta^*, K)} \mathbb{E}_{f^*}\Big[\Pi\Big(\mathcal{H}^c\big(f^*, MR_n\varepsilon_n(\eta^*)\big) \cap \widetilde{\mathcal{M}}_n(C)\Big|(\mathbf{X}, \mathbf{Y})\Big)\Big] \to 0. \tag{38}$$

Fix any $f^* \in \mathcal{F}(\eta^*, K)$. Since there is no ambiguity, we shorten the notation to $\mathcal{H}_n^c = \mathcal{H}^c(f^*, MR_n\varepsilon_n(\eta^*))$ and rewrite

$$\mathbb{E}_{f^*}\Big[\Pi\Big(\mathcal{H}_n^c \cap \widetilde{\mathcal{M}}_n(C)\Big|(\mathbf{X}, \mathbf{Y})\Big)\Big] = \mathbb{E}_{f^*}\left[\frac{\int_{\mathcal{H}_n^c} \Pi(df \cap \widetilde{\mathcal{M}}_n(C)|\mathbf{X}, \mathbf{Y})}{\int \Pi(df|\mathbf{X}, \mathbf{Y})}\right].$$

We follow the steps of the proof of Theorem 8.14 by Ghosal and van der Vaart (2017). In their notation we use $\varepsilon_n = R_n\varepsilon_n(\eta^*)$ and $\xi = 1/2$ for contraction with respect to Hellinger loss. Set

$$A_n^* = \Big\{\int \Pi(df|\mathbf{X}, \mathbf{Y}) \geq \Pi\Big(B_2\big(f^*, R_n\varepsilon_n(\eta^*)\big)\Big)e^{-2R_n^2 n\varepsilon_n(\eta^*)^2}\Big\}.$$

Then $\mathbb{P}_{f^*}(A_n^*)$ tends to 1, thanks to Lemma 8.10 by Ghosal and van der Vaart (2017) applied with $D = 1$. Since $1 = \mathbf{1}(A_n^*) + \mathbf{1}(A_n^{*,c})$, we have

$$\mathbb{E}_{f^*}\Big[\Pi\Big(\mathcal{H}_n^c \cap \widetilde{\mathcal{M}}_n(C)\Big|(\mathbf{X}, \mathbf{Y})\Big)\Big] \leq \mathbb{P}_{f^*}(A_n^{*,c}) + \mathbb{E}_{f^*}\left[\mathbf{1}(A_n^*)\frac{\int_{\mathcal{H}_n^c} \Pi(df \cap \widetilde{\mathcal{M}}_n(C)|\mathbf{X}, \mathbf{Y})}{\int \Pi(df|\mathbf{X}, \mathbf{Y})}\right]$$

and $\mathbb{P}_{f^*}(A_n^{*,c}) \to 0$ when $n \to +\infty$. It remains to show that the second terms on the right side tends to zero.

Let $\phi_{n,k}(\lambda)$ be arbitrary statistical tests to be chosen later. Test are to be understood as $\phi_{n,k}(\lambda) = \phi_{n,k}(\lambda)(\mathbf{X}, \mathbf{Y})$ measurable functions of the sample $(\mathbf{X}, \mathbf{Y})$, taking values in $[0, 1]$. Then, $1 = \phi_{n,k}(\lambda) + (1 - \phi_{n,k}(\lambda))$. Using the definition of $\Pi(df \cap \widetilde{\mathcal{M}}_n(C)|\mathbf{X}, \mathbf{Y})$ and Fubini's theorem, we find

$$\mathbb{E}_{f^*}\Big[\Pi\Big(\mathcal{H}_n^c \cap \widetilde{\mathcal{M}}_n(C)\Big|(\mathbf{X}, \mathbf{Y})\Big)\Big] \leq \mathbb{P}_{f^*}(A_n^{*,c}) + \mathbb{E}_{f^*}\left[T_1 + T_2\right], \tag{39}$$

where

$$T_1 := \mathbf{1}(A_n^*)\frac{\sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \phi_{n,k}(\lambda) \int_{B_k(\lambda)} \pi(\lambda, \boldsymbol{\beta}) \Big(\int_{\mathcal{H}_n^c} \frac{p_f}{p_{f^*}}(\mathbf{X}, \mathbf{Y})\Pi(df|\lambda, \boldsymbol{\beta})\Big)\, d\boldsymbol{\beta}}{\int \Pi(df|\mathbf{X}, \mathbf{Y})},$$

$$T_2 := \mathbf{1}(A_n^*)\frac{\sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \int_{B_k(\lambda)} \pi(\lambda, \boldsymbol{\beta}) \Big(\int_{\mathcal{H}_n^c} (1 - \phi_{n,k}(\lambda))\frac{p_f}{p_{f^*}}(\mathbf{X}, \mathbf{Y})\Pi(df|\lambda, \boldsymbol{\beta})\Big)\, d\boldsymbol{\beta}}{\int \Pi(df|\mathbf{X}, \mathbf{Y})}.$$

We bound T_1 by using $\mathbf{1}(A_n^*) \leq 1$, together with

$$\frac{\int_{B_k(\lambda)} \pi(\lambda, \boldsymbol{\beta}) \left(\int_{\mathcal{H}_n^c} \frac{p_f}{p_{f^*}}(\mathbf{X}, \mathbf{Y}) \Pi(df|\lambda, \boldsymbol{\beta}) \right) d\boldsymbol{\beta}}{\int \Pi(df|\mathbf{X}, \mathbf{Y})} \leq 1,$$

so that

$$\mathbb{E}_{f^*}[T_1] \leq \sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \mathbb{E}_{f^*}[\phi_{n,k}(\lambda)]. \tag{40}$$

We bound T_2 using the definition of A_n^*, and obtain

$$T_2 \leq \frac{\sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \int_{B_k(\lambda)} \pi(\lambda, \boldsymbol{\beta}) \left(\int_{\mathcal{H}_n^c} (1 - \phi_{n,k}(\lambda)) \frac{p_f}{p_{f^*}}(\mathbf{X}, \mathbf{Y}) \Pi(df|\lambda, \boldsymbol{\beta}) \right) d\boldsymbol{\beta}}{\Pi\left(B_2\big(f^*, R_n \varepsilon_n(\eta^*)\big)\right) e^{-2R_n^2 n \varepsilon_n(\eta^*)^2}}.$$

For large n, $R_n = 10QC(2\log n)^{1+\log K} \geq 4QK^{q^*}(q^*+1)$. By Lemma 18, we find

$$\Pi\left(B_2\big(f^*, R_n \varepsilon_n(\eta^*)\big)\right) \geq e^{-R_n^2 n \varepsilon_n(\eta^*)^2} \pi(\lambda^*, I_n^*), \tag{41}$$

with $I_n^* := \{\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{q^*}) : \beta_i \in [\beta_i^* - b_n, \beta_i^*], \forall i\}$ and $b_n := 1/\log^2 n$. By the construction of the prior π in (13), the denominator term $e^{-\Psi_n(\eta)}$ is bounded above by 1 and $\int_\Omega \gamma(\eta)\, d\eta = 1$, thus

$$\pi(\lambda^*, I_n^*) = \frac{\int_{I_n^*} e^{-\Psi_n(\lambda^*, \boldsymbol{\beta})} \gamma(\lambda^*, \boldsymbol{\beta}) d\boldsymbol{\beta}}{\int_\Omega e^{-\Psi_n(\eta)} \gamma(\eta)\, d\eta} \geq \int_{I_n^*} e^{-\Psi_n(\lambda^*, \boldsymbol{\beta})} \gamma(\lambda^*, \boldsymbol{\beta}) d\boldsymbol{\beta}.$$

Furthermore, by condition (ii) in Assumption 2, we have $\varepsilon_n(\lambda^*, \boldsymbol{\beta}) \leq Q\varepsilon_n(\lambda^*, \boldsymbol{\beta}^*)$ for any $\boldsymbol{\beta} \in I_n^*$, which gives

$$\pi(\lambda^*, I_n^*) \geq \exp(-e^{e^{|\mathbf{d}^*|_1}})\, e^{-Q^2 n \varepsilon_n(\eta^*)^2} \gamma(\lambda^*, I_n^*),$$

with proportionality constant $\exp(-e^{e^{|\mathbf{d}^*|_1}}) > 0$ independent of n. Again by construction, the measure $\gamma(\lambda^*, I_n^*)$ can be split into the product $\gamma(\lambda^*)\gamma(I_n^*|\lambda^*)$ where $\gamma(\cdot|\lambda^*)$ is the uniform measure on $I(\lambda^*) = [\beta_-, \beta_+]^{q^*+1}$ and the quantity $\gamma(\lambda^*) > 0$ is a constant independent of n. Thus, with $c(\eta^*) := \exp(-e^{e^{|\mathbf{d}^*|_1}})\gamma(\lambda^*)/|I(\lambda^*)|$, we obtain $\pi(\lambda^*, I_n^*) \geq c(\eta^*)|I_n^*|e^{-Q^2 n \varepsilon_n(\eta^*)^2}$. In the discussion after (34) we have shown that $n\varepsilon_n(\eta^*)^2$ is a positive power of n and so, for a large enough n depending only on η^*, one has $|I_n^*| = (1/\log^2 n)^{q^*+1} > e^{-n\varepsilon_n(\eta^*)^2}$. This results in

$$\pi(\lambda^*, I_n^*) \geq c(\eta^*) e^{-(Q^2+1) n \varepsilon_n(\eta^*)^2} \geq c(\eta^*) e^{-R_n^2 n \varepsilon_n(\eta^*)^2}. \tag{42}$$

By putting together the small-ball probability bound (41) and the hyperprior bound (42), we recover

$$T_2 \leq c(\eta^*)^{-1} \frac{\sum_{\lambda \in \mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \int_{B_k(\lambda)} \pi(\lambda, \boldsymbol{\beta}) \left(\int_{\mathcal{H}_n^c} (1 - \phi_{n,k}(\lambda)) \frac{p_f}{p_{f^*}}(\mathbf{X}, \mathbf{Y}) \Pi(df|\lambda, \boldsymbol{\beta}) \right) d\boldsymbol{\beta}}{e^{-4R_n^2 n \varepsilon_n(\eta^*)^2}}, \tag{43}$$

with proportionality constant $c(\eta^*)^{-1}$ independent of n and depending only on $\eta^*,\ \beta_-,\ \beta_+$. We now bound the numerator of T_2 using Fubini's theorem and the inequality $\mathbb{E}_{f^*}[(1-\phi_{n,k}(\lambda))(p_f/p_{f_*})(\mathbf{X},\mathbf{Y})] \le \mathbb{E}_f[1-\phi_{n,k}(\lambda)]$,

$$\begin{aligned}
&\mathbb{E}_{f^*}\left[\sum_{\lambda\in\mathcal{L}_n(C)}\sum_{k=1}^{N(\lambda)}\int_{B_k(\lambda)}\pi(\lambda,\boldsymbol{\beta})\left(\int_{\mathcal{H}_n^c}(1-\phi_{n,k}(\lambda))\frac{p_f}{p_{f^*}}(\mathbf{X},\mathbf{Y})\Pi(df|\lambda,\boldsymbol{\beta})\right)d\boldsymbol{\beta}\right]\\
&\quad=\sum_{\lambda\in\mathcal{L}_n(C)}\sum_{k=1}^{N(\lambda)}\int_{B_k(\lambda)}\pi(\lambda,\boldsymbol{\beta})\left(\int_{\mathcal{H}_n^c}\mathbb{E}_{f^*}\left[(1-\phi_{n,k}(\lambda))\frac{p_f}{p_{f^*}}(\mathbf{X},\mathbf{Y})\right]\Pi(df|\lambda,\boldsymbol{\beta})\right)d\boldsymbol{\beta}\\
&\quad\le\sum_{\lambda\in\mathcal{L}_n(C)}\sum_{k=1}^{N(\lambda)}\int_{B_k(\lambda)}\pi(\lambda,\boldsymbol{\beta})\int_{\mathcal{H}_n^c}\mathbb{E}_f\left[(1-\phi_{n,k}(\lambda))\right]\Pi(df|\lambda,\boldsymbol{\beta})d\boldsymbol{\beta}. \qquad (44)
\end{aligned}$$

With the supports $\Theta_n(\lambda,\boldsymbol{\beta},K)$ in (36) and any $k=1,\ldots,N(\lambda)$, consider $\Theta_n(\lambda,B_k(\lambda),K)=\cup_{\boldsymbol{\beta}\in B_k(\lambda)}\Theta_n(\lambda,\boldsymbol{\beta},K)$. Now, for any fixed λ,k choose tests $\phi_{n,k}(\lambda)$ according to Theorem D.5 by Ghosal and van der Vaart (2017), so that for $f\in\Theta_n(\lambda,B_k(\lambda),K)\cap\mathcal{H}^c(f^*,MR_n\varepsilon_n(\eta^*))$ we have, for some universal constant $\widetilde{K}>0$,

$$\begin{aligned}
\mathbb{E}_{f^*}[\phi_{n,k}(\lambda)]&\le c_k(\lambda)\mathcal{N}\left(\frac{R_n\varepsilon_n(\eta^*)}{2},\Theta_n(\lambda,B_k(\lambda),K),\|\cdot\|_\infty\right)\frac{e^{-\widetilde{K}M^2R_n^2n\varepsilon_n(\eta^*)^2}}{1-e^{-\widetilde{K}M^2R_n^2n\varepsilon_n(\eta^*)^2}},\\
\mathbb{E}_f[1-\phi_{n,k}(\lambda)]&\le c_k(\lambda)^{-1}e^{-\widetilde{K}M^2R_n^2n\varepsilon_n(\eta^*)^2},
\end{aligned}$$

with choice of coefficients

$$c_k(\lambda)^2:=\frac{\pi(\lambda,B_k(\lambda))}{\mathcal{N}\left(\frac{R_n\varepsilon_n(\eta^*)}{2},\Theta_n(\lambda,B_k(\lambda),K),\|\cdot\|_\infty\right)}.$$

Let us denote by $\rho_k(\lambda)$ the local complexities

$$\rho_k(\lambda):=\sqrt{\pi(\lambda,B_k(\lambda))}\cdot\sqrt{\mathcal{N}\left(\frac{R_n\varepsilon_n(\eta^*)}{2},\Theta_n(\lambda,B_k(\lambda),K),\|\cdot\|_\infty\right)}.$$

Combining this with the bound on T_1 in (40) and the bounds on T_2 in (43)-(44), gives

$$\begin{aligned}
\mathbb{E}_{f^*}\left[T_1\right]&\le\frac{e^{-\widetilde{K}M^2R_n^2n\varepsilon_n(\eta^*)^2}}{1-e^{-\widetilde{K}M^2R_n^2n\varepsilon_n(\eta^*)^2}}\sum_{\lambda\in\mathcal{L}_n(C)}\sum_{k=1}^{N(\lambda)}\rho_k(\lambda),\\
\mathbb{E}_{f^*}\left[T_2\right]&\le c(\eta^*)^{-1}e^{(4-\widetilde{K}M^2)R_n^2n\varepsilon_n(\eta^*)^2}\sum_{\lambda\in\mathcal{L}_n(C)}\sum_{k=1}^{N(\lambda)}\rho_k(\lambda).
\end{aligned}$$

It remains to show that both expectations in the latter display tend to zero when $n\to+\infty$.

Since $M>0$ can be chosen arbitrarily large, we choose it in such a way that $\widetilde{K}M^2>5$ and the proof is complete if we can show that

$$\sum_{\lambda\in\mathcal{L}_n(C)}\sum_{k=1}^{N(\lambda)}\rho_k(\lambda)\lesssim e^{R_n^2n\varepsilon_n(\eta^*)^2}, \qquad (45)$$

for some proportionality constant independent of n. Fix $\lambda \in \mathcal{L}_n(C)$ and $k = 1, \ldots, N(\lambda)$. By construction, there exists $\boldsymbol{\beta} \in B_k(\lambda)$ such that $(\lambda, \boldsymbol{\beta}) \in \mathcal{M}_n(C)$. Since $R_n = 10QC(2\log n)^{1+\log K}$, using $C\varepsilon_n(\eta^*) \geq \varepsilon_n(\eta)$ together with Lemma 20, we find

$$\log \mathcal{N}\left(\frac{R_n\varepsilon_n(\eta^*)}{2}, \Theta_n(\lambda, B, K), \|\cdot\|_\infty\right) \leq \frac{R_n^2}{25} n\varepsilon_n(\eta^*)^2.$$

This results in

$$\begin{aligned}\sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \rho_k(\lambda) &\leq \exp\left(\frac{1}{50}R_n^2 n\varepsilon_n(\eta^*)^2\right) \sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \sqrt{\pi(\lambda, B_k(\lambda))} \\ &= \exp\left(\frac{1}{50}R_n^2 n\varepsilon_n(\eta^*)^2\right) z_n^{-\frac{1}{2}} \sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \sqrt{\gamma(\lambda, B_k(\lambda))},\end{aligned}$$

where $z_n = \sum_\lambda \int_{I(\lambda)} e^{-\Psi_n(\lambda,\boldsymbol{\beta})}\gamma(\lambda, \boldsymbol{\beta})d\boldsymbol{\beta}$ is the normalization term in (13). By the localization argument in (42), we know that $z_n \geq \pi(\lambda^*, I_n^*) \gtrsim e^{-R_n^2 n\varepsilon_n(\eta^*)^2}$, with proportionality constant independent of n. Thus,

$$\sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \rho_k(\lambda) \lesssim \exp\left(\frac{26}{50}R_n^2 n\varepsilon_n(\eta^*)^2\right) \sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \sqrt{\gamma(\lambda, B_k(\lambda))}.$$

Since $R_n = 10QC(2\log n)^{1+\log K} \gg 1$, it is sufficient for (45) that

$$\sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \sqrt{\gamma(\lambda, B_k(\lambda))} \lesssim e^{\frac{1}{2}n\varepsilon_n(\eta^*)^2}, \tag{46}$$

for some proportionality constant that can be chosen independent of n. To see this, observe that by Assumption 1, we have $\gamma(\lambda, \boldsymbol{\beta}) = \gamma(\lambda)\gamma(\boldsymbol{\beta}|\lambda)$ with $\gamma(\cdot|\lambda)$ the uniform distribution over $I(\lambda)$. Thus

$$\sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \sqrt{\gamma(\lambda, B_k(\lambda))} = \sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \sqrt{|B_k(\lambda)|}\sqrt{\gamma(\lambda, \boldsymbol{\beta}_k(\lambda))},$$

where $\boldsymbol{\beta}_k(\lambda)$ is the center point of the hypercube $B_k(\lambda)$. Since $|B_k(\lambda)| = (1/\log^2 n)^{q+1}$ and $(q+1) \leq |\mathbf{d}|_1 \leq \log(2\log n)$, we have $\log(|B_k(\lambda)|^{-1}) = 2(q+1)(\log n) \leq 4\log^2 n \ll n\varepsilon_n(\eta^*)^2$. Therefore, for a sufficiently large n depending only on η^*, we find $|B_k(\lambda)|^{-1} \leq e^{n\varepsilon_n(\eta^*)^2}$. The above discussion yields the following bound on the latter display,

$$\begin{aligned}\sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \sqrt{\gamma(\lambda, B_k(\lambda))} &= \sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} \frac{1}{\sqrt{|B_k(\lambda)|}}|B_k(\lambda)|\sqrt{\gamma(\lambda, \boldsymbol{\beta}_k(\lambda))} \\ &\leq e^{\frac{1}{2}n\varepsilon_n(\eta^*)^2} \sum_{\lambda\in\mathcal{L}_n(C)} \sum_{k=1}^{N(\lambda)} |B_k(\lambda)|\sqrt{\gamma(\lambda, \boldsymbol{\beta}_k(\lambda))}.\end{aligned}$$

This is enough to obtain (46) since, by Assumption 1,

$$\sum_{\lambda\in\mathcal{L}_n(C)}\sum_{k=1}^{N(\lambda)}|B_k(\lambda)|\sqrt{\gamma(\lambda,\boldsymbol{\beta}_k(\lambda))}\leq\sum_{\lambda\in\Lambda}\int_{I(\lambda)}\sqrt{\gamma(\lambda,\boldsymbol{\beta})}d\boldsymbol{\beta}=\int_{\Omega}\sqrt{\gamma(\eta)}\,d\eta$$

is a finite constant independent of n.

Since all bounds are independent of the particular choice of f^* and only depend on the function class $\mathcal{F}(\eta^*,K)$, this concludes the proof of the uniform statement (38). ■

A.2.1 Proofs of auxiliary results

Proof [Proof of Lemma 19] We follow the proof of Theorem 2.7.1 by van der Vaart and Wellner (1996) and provide explicit expressions for all constants. We start by covering the interval $[-1,1]^r$ with a grid of width $\tau=(\delta/c(\beta))^{1/\beta}$, where $c(\beta):=er^\beta+2K$. The grid consists of M points $\mathbf{x}_1,\ldots,\mathbf{x}_M$ with

$$M\leq\frac{vol([-2,2]^r)}{\tau^r}=4^r c(\beta)^{\frac{r}{\beta}}\delta^{-\frac{r}{\beta}}. \tag{47}$$

For any $h\in\mathcal{C}_r^\beta(K)$ and any $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_r)\in\mathbb{N}^r$ with $|\boldsymbol{\alpha}|_1=\alpha_1+\ldots+\alpha_r\leq\lfloor\beta\rfloor$, set

$$A^{\boldsymbol{\alpha}}h:=\left(\left\lfloor\frac{\partial^{\boldsymbol{\alpha}}h(\mathbf{x}_1)}{\tau^{\beta-|\boldsymbol{\alpha}|_1}}\right\rfloor,\ldots,\left\lfloor\frac{\partial^{\boldsymbol{\alpha}}h(\mathbf{x}_M)}{\tau^{\beta-|\boldsymbol{\alpha}|_1}}\right\rfloor\right). \tag{48}$$

The vector $\tau^{\beta-|\boldsymbol{\alpha}|_1}A^{\boldsymbol{\alpha}}h$ consists of the values $\partial^{\boldsymbol{\alpha}}h(\mathbf{x}_i)$ discretized on a grid of mesh-width $\tau^{\beta-|\boldsymbol{\alpha}|_1}$. Since $\partial^{\boldsymbol{\alpha}}h\in\mathcal{C}_r^{\beta-|\boldsymbol{\alpha}|_1}(K)$ and $\tau<1$ by construction, the entries of the vector in the latter display are integers bounded in absolute value by

$$\left\lfloor\frac{|\partial^{\boldsymbol{\alpha}}h(\mathbf{x}_i)|}{\tau^{\beta-|\boldsymbol{\alpha}|_1}}\right\rfloor\leq\left\lfloor\frac{K}{\tau^{\beta-|\boldsymbol{\alpha}|_1}}\right\rfloor\leq\left\lfloor\frac{K}{\tau^{\beta}}\right\rfloor. \tag{49}$$

Let $h,\widetilde{h}\in\mathcal{C}_r^\beta(K)$ be two functions such that $A^{\boldsymbol{\alpha}}h=A^{\boldsymbol{\alpha}}\widetilde{h}$ for all $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}|_1\leq\lfloor\beta\rfloor$. We now show that $\|h-\widetilde{h}\|_\infty\leq\delta$. For any $\mathbf{x}\in[-1,1]^r$, let $\mathbf{x}_i$ be the closest grid vertex, so that $|\mathbf{x}-\mathbf{x}_i|_\infty\leq\tau$. Taylor expansion around $\mathbf{x}_i$ gives

$$\begin{aligned}(h-\widetilde{h})(\mathbf{x})&=\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1\leq\lfloor\beta\rfloor}\partial^{\boldsymbol{\alpha}}(h-\widetilde{h})(\mathbf{x}_i)\frac{(\mathbf{x}-\mathbf{x}_i)^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!}+R,\\ R&=\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1=\lfloor\beta\rfloor}\left[\partial^{\boldsymbol{\alpha}}(h-\widetilde{h})(\mathbf{x}_\xi)-\partial^{\boldsymbol{\alpha}}(h-\widetilde{h})(\mathbf{x}_i)\right]\frac{(\mathbf{x}-\mathbf{x}_i)^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!},\end{aligned} \tag{50}$$

with $\mathbf{x}_{\xi_i} = \mathbf{x}_i + \xi_i(\mathbf{x} - \mathbf{x}_i)$ and a suitable $\xi_i \in [0,1]$. With $|\partial^{\boldsymbol{\alpha}}(h - \widetilde{h})|_{\beta-|\boldsymbol{\alpha}|_1}$ the Hölder seminorm of the function $\partial^{\boldsymbol{\alpha}}(h - \widetilde{h}) \in \mathcal{C}_r^{\beta-|\boldsymbol{\alpha}|_1}(2K)$, we bound the remainder R by

$$\begin{aligned}
|R| &\leq \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1=\lfloor\beta\rfloor} \left|\partial^{\boldsymbol{\alpha}}(h-\widetilde{h})(\mathbf{x}_\xi) - \partial^{\boldsymbol{\alpha}}(h-\widetilde{h})(\mathbf{x}_i)\right| \frac{\tau^{|\boldsymbol{\alpha}|_1}}{\boldsymbol{\alpha}!} \\
&\leq \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1=\lfloor\beta\rfloor} \left|\partial^{\boldsymbol{\alpha}}(h-\widetilde{h})\right|_{\beta-|\boldsymbol{\alpha}|_1} \tau^{\beta-|\boldsymbol{\alpha}|_1} \frac{\tau^{|\boldsymbol{\alpha}|_1}}{\boldsymbol{\alpha}!} \\
&\leq 2K\tau^\beta.
\end{aligned}$$

Plugging this into the bound (50) gives

$$\begin{aligned}
|(h-\widetilde{h})(\mathbf{x})| &\leq \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1\leq\lfloor\beta\rfloor} \left|\partial^{\boldsymbol{\alpha}}(h-\widetilde{h})(\mathbf{x}_i)\right| \frac{\tau^{|\boldsymbol{\alpha}|_1}}{\boldsymbol{\alpha}!} + 2K\tau^\beta \\
&= \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1\leq\lfloor\beta\rfloor} \tau^{\beta-|\boldsymbol{\alpha}|_1} \left|\frac{\partial^{\boldsymbol{\alpha}}(h-\widetilde{h})(\mathbf{x}_i)}{\tau^{\beta-|\boldsymbol{\alpha}|_1}}\right| \frac{\tau^{|\boldsymbol{\alpha}|_1}}{\boldsymbol{\alpha}!} + 2K\tau^\beta.
\end{aligned}$$

In view of definition (48), we denote $A^{\boldsymbol{\alpha}}(h-\widetilde{h})(\mathbf{x}_i) = \lfloor \partial^{\boldsymbol{\alpha}}(h-\widetilde{h})(\mathbf{x}_i)/\tau^{\beta-|\boldsymbol{\alpha}|_1} \rfloor$ and $B^{\boldsymbol{\alpha}}(h-\widetilde{h})(\mathbf{x}_i) = \partial^{\boldsymbol{\alpha}}(h-\widetilde{h})(\mathbf{x}_i)/\tau^{\beta-|\boldsymbol{\alpha}|_1} - A^{\boldsymbol{\alpha}}(h-\widetilde{h})(\mathbf{x}_i)$. Thus $A^{\boldsymbol{\alpha}}(h-\widetilde{h})(\mathbf{x}_i) = 0$ by assumption on $h, \widetilde{h}$ and $|B^{\boldsymbol{\alpha}}(h-\widetilde{h})(\mathbf{x}_i)| < 1$. We now prove

$$\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1\leq\lfloor\beta\rfloor} \frac{1}{\boldsymbol{\alpha}!} \leq er^\beta. \tag{51}$$

In fact, for any positive integer k, consider the multinomial distribution induced by a fair r-sided die over k independent rolls. The corresponding p.m.f. is $\boldsymbol{\alpha} \mapsto r^{-k}k!/\boldsymbol{\alpha}!$ and is supported on $\{\boldsymbol{\alpha} : |\boldsymbol{\alpha}|_1 = k\}$. Since the p.m.f. sums to one, we have $\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1=k} 1/\boldsymbol{\alpha}! = r^k/k!$. By summing over $k = 0, \ldots, \lfloor\beta\rfloor$, one finds $\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1\leq\lfloor\beta\rfloor} 1/\boldsymbol{\alpha}! = \sum_{k=0}^{\lfloor\beta\rfloor}\sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1=k} 1/\boldsymbol{\alpha}! \leq er^\beta$, proving (51). Combining the discussion above together with the previous bounds, yields

$$|(h-\widetilde{h})(\mathbf{x})| \leq \tau^\beta \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1\leq\lfloor\beta\rfloor} \frac{1}{\boldsymbol{\alpha}!} + 2K\tau^\beta \leq \tau^\beta\left(er^\beta + 2K\right) = \delta, \tag{52}$$

proving that, if two functions $h, \widetilde{h} \in \mathcal{C}_r^\beta(K)$ have $A^{\boldsymbol{\alpha}}h = A^{\boldsymbol{\alpha}}\widetilde{h}$ for all $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}|_1 \leq \lfloor\beta\rfloor$, then $\|h-\widetilde{h}\|_\infty \leq \delta$.

The quantity $\mathcal{N}(\delta, \mathcal{C}_r^\beta(K), \|\cdot\|)$ is bounded above by cardinality $\#\mathcal{A}$ of the set of matrices

$$\mathcal{A} = \left\{ Ah = (A^{\boldsymbol{\alpha}}h)^\top_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1\leq\lfloor\beta\rfloor} : h \in \mathcal{C}_r^\beta(K) \right\}.$$

The rows of the matrix Ah consist of the row vectors $A^{\boldsymbol{\alpha}}h$. Since we consider $\boldsymbol{\alpha} : |\boldsymbol{\alpha}|_1 \leq \lfloor\beta\rfloor$, the matrix Ah can have at most $(\lfloor\beta\rfloor+1)^r$ rows. Any matrix Ah has moreover M columns.

To complete the counting argument, we first explain the underlying idea. If two neighboring grid points $\mathbf{x}_i, \mathbf{x}_j$ are selected such that $|\mathbf{x}_i - \mathbf{x}_j|_\infty < 2\tau$, say, then, $\partial^{\boldsymbol{\alpha}}h(\mathbf{x}_i) \approx \partial^{\boldsymbol{\alpha}}h(\mathbf{x}_j)$, whenever $|\boldsymbol{\alpha}|_1 < \beta$. Since the ℓ-th column of Ah contains the discretized entries

$(\partial^{\boldsymbol{\alpha}} h(\mathbf{x}_\ell))_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1\leq\lfloor\beta\rfloor}$, the number of possible realizations of the i-th and j-th column vector can be bounded by the possible realizations of the i-th column vector times a factor that describes the number of possible deviations of the values in the j-th column vector.

We now show that this factor is bounded by $2^{\beta+1}c(\beta)$. To see this, observe that Taylor expansion gives

$$\partial^{\boldsymbol{\alpha}} h(\mathbf{x}_j) = \sum_{\mathbf{k}:|\boldsymbol{\alpha}|_1+|\mathbf{k}|_1<\lfloor\beta\rfloor} \partial^{\boldsymbol{\alpha}+\mathbf{k}} h(\mathbf{x}_i)\frac{(\mathbf{x}_j-\mathbf{x}_i)^{\mathbf{k}}}{\mathbf{k}!} + \sum_{\mathbf{k}:|\boldsymbol{\alpha}|_1+|\mathbf{k}|_1=\lfloor\beta\rfloor} \partial^{\boldsymbol{\alpha}+\mathbf{k}} h(\mathbf{x}_\xi)\frac{(\mathbf{x}_j-\mathbf{x}_i)^{\mathbf{k}}}{\mathbf{k}!},$$

for some $\mathbf{x}_\xi$ on the line with endpoints $\mathbf{x}_i, \mathbf{x}_j$. By replacing $\widetilde{h}$ by 0, h by $\partial^{\boldsymbol{\alpha}} h$, β by $\beta-|\boldsymbol{\alpha}|_1$ and τ by 2τ, we can argue as for (52) to find

$$\begin{aligned}\left|\partial^{\boldsymbol{\alpha}} h(\mathbf{x}_j) - \sum_{\mathbf{k}:|\boldsymbol{\alpha}|_1+|\mathbf{k}|_1\leq\lfloor\beta\rfloor} \tau^{\beta-|\boldsymbol{\alpha}|_1-|\mathbf{k}|_1} A^{\boldsymbol{\alpha}+\mathbf{k}} h(\mathbf{x}_i)\frac{(\mathbf{x}_j-\mathbf{x}_i)^{\mathbf{k}}}{\mathbf{k}!}\right| &\leq 2^{\beta-|\boldsymbol{\alpha}|_1} c(\beta-|\boldsymbol{\alpha}|_1)\tau^{\beta-|\boldsymbol{\alpha}|_1} \\ &\leq 2^{\beta} c(\beta)\tau^{\beta-|\boldsymbol{\alpha}|_1}.\end{aligned}$$

This shows that, if the i-th column of Ah is fixed, the values $\partial^{\boldsymbol{\alpha}} h(\mathbf{x}_j)$ range over an interval of length at most $2\cdot 2^{\beta}c(\beta)\tau^{\beta-|\boldsymbol{\alpha}|_1}$. The entry $\lfloor \partial^{\boldsymbol{\alpha}} h(\mathbf{x}_j)/\tau^{\beta-|\boldsymbol{\alpha}|_1}\rfloor$ can attain therefore at most $2^{\beta+1}c(\beta)\tau^{\beta-|\boldsymbol{\alpha}|_1}/\tau^{\beta-|\boldsymbol{\alpha}|_1}+1 = 2^{\beta+1}c(\beta)+1$ different values. As there are at most $(\beta+1)^r$ many rows, for fixed i-th column of Ah, the j-th column of Ah can attain at most $(2^{\beta+1}c(\beta)+1)^{(\beta+1)^r}$ different values.

Without loss of generality, assume that the points $\mathbf{x}_1,\ldots,\mathbf{x}_M$ are ordered in such a way that for each $j>1$, there exists $i<j$, such that $|\mathbf{x}_i-\mathbf{x}_j|_\infty < 2\tau$. This determines then also the ordering of the columns of the matrix Ah. In view of Equation (49), the first column of Ah can attain at most $(2K\tau^{-\beta}+1)^{(\beta+1)^r}$ different values. For each of the $M-1$ remaining columns, we can use the argument above and find

$$\mathcal{N}(\delta, \mathcal{C}_r^\beta(K), \|\cdot\|_\infty) \leq \#\mathcal{A} \leq (2K\tau^{-\beta}+1)^{(\beta+1)^r}\cdot(2^{\beta+1}c(\beta)+1)^{(M-1)(\beta+1)^r}.$$

Since $x+y\leq xy$ for all $x,y\geq 2$, using $c(\beta)=er^\beta+2K\leq 2eKr^\beta$, the bound on M in (47) and the definition of $\tau=(\delta/c(\beta))^{1/\beta}$, the first assertion of the lemma follows.

For the bound on the constant $Q_1(\beta,r,K)$, we take the logarithm. With $\log(x+1)\leq\log(2x)$ for all $x>1$, we get

$$\begin{aligned}&\log\left(\left(\frac{4eK^2r^\beta}{\delta}+1\right)^{(\beta+1)^r}\left(2^{\beta+2}eKr^\beta+1\right)^{4^r(\beta+1)^r r^r(2eK)^{\frac{r}{\beta}}\delta^{-\frac{r}{\beta}}}\right)\\ &\leq (\beta+1)^r\log\left(\frac{8eK^2r^\beta}{\delta}\right)+4^r(\beta+1)^r r^r(2eK)^{\frac{r}{\beta}}\delta^{-\frac{r}{\beta}}\log\left(2^{\beta+3}eKr^\beta\right)\\ &=: A_1+A_2.\end{aligned}$$

Observe that $\log(x)<x^a/a$ for all $a,x>0$, then

$$\log\left(\frac{8eK^2r^\beta}{\delta}\right)\leq\frac{\beta}{r}(8eK^2r^\beta)^{\frac{r}{\beta}}\delta^{-\frac{r}{\beta}} = \beta r^{r-1}(8eK^2)^{\frac{r}{\beta}}\delta^{-\frac{r}{\beta}},$$

which yields $A_1 \leq (\beta+1)^{r+1} r^{r-1} (8eK^2)^{r/\beta} \delta^{-r/\beta}$. Furthermore, using that $\log x < x$ for all $x > 0$,

$$\log\left(2^{\beta+3} eKr^{\beta}\right) \leq 2(\beta+3) + \beta r + eK \leq (r+2)(\beta+3) + eK.$$

Since $r > 1$, the latter display is smaller than $4(\beta+3)r + eK \leq 4eK(\beta+3)r$ and so $A_2 \leq 4eK(\beta+3)r \cdot 4^r (\beta+1)^r r^r (2eK)^{r/\beta} \delta^{-r/\beta}$. Combining the bounds for A_1 and A_2, we find

$$\begin{aligned}\frac{A_1 + A_2}{\delta^{-\frac{r}{\beta}}} &\leq (\beta+1)^{r+1} r^{r-1} (8eK^2)^{\frac{r}{\beta}} + eK(\beta+3) 4^{r+1} r^{r+1} (\beta+1)^r (2eK)^{\frac{r}{\beta}} \\ &\leq (1+eK) 4^{r+1} (\beta+3)^{r+1} r^{r+1} (8eK^2)^{\frac{r}{\beta}},\end{aligned}$$

and the right hand side is $Q_1(\beta, r, K)$ by definition.

We now prove the entropy bound in (35). Let $Q_1 = Q_1(\beta, r, K)$, $C_1 = Q_1^{\beta\alpha/(2\beta\alpha+r)}$ and $\mathfrak{r}_n = n^{-\beta\alpha/(2\beta\alpha+r)}$. By construction, $\mathfrak{r}_n^{-r/\beta\alpha} = n\mathfrak{r}_n^2$ and $Q_1 C_1^{-r/\beta\alpha} = C_1^2$. For any sequence $\delta_n \geq C_1 \mathfrak{r}_n$, the first part of the proof gives

$$\begin{aligned}\log \mathcal{N}\left(\delta_n^{\frac{1}{\alpha}}, \mathcal{C}_r^{\beta}(K), \|\cdot\|_\infty\right) &\leq \log \mathcal{N}\left((C_1\mathfrak{r}_n)^{\frac{1}{\alpha}}, \mathcal{C}_{t_i}^{\beta_i}(K), \|\cdot\|_\infty\right) \\ &\leq Q_1 C_1^{-\frac{r}{\beta\alpha}} \mathfrak{r}_n^{-\frac{r}{\beta\alpha}} \\ &= C_1^2 n \mathfrak{r}_n^2 \\ &\leq n\delta_n^2.\end{aligned}$$

The proof is complete. ■

Proof [Proof of Lemma 20] Fix any $\boldsymbol{\beta} \in [\beta_-, \beta_+]^{q+1}$. In a first step we show that, with $R := 5K^q(q+1)$ and $\delta_{in}(\lambda, \boldsymbol{\beta}) = \varepsilon_n(\alpha_i, \beta_i, t_i)^{1/\alpha_i}$, we have

$$\mathcal{N}\left(R\varepsilon_n(\lambda, \boldsymbol{\beta}), \Theta_n(\lambda, \boldsymbol{\beta}, K), \|\cdot\|_\infty\right) \leq \prod_{i=0}^{q} \mathcal{N}\left(3\delta_{in}(\lambda, \boldsymbol{\beta}), \Theta_{i,n}(\lambda, \boldsymbol{\beta}, K), \|\cdot\|_\infty\right). \tag{53}$$

For any $i = 0, \ldots, q$, let $g_{i,1}, \ldots, g_{i,N_i}$ be the centers of a $3\delta_{in}(\lambda, \boldsymbol{\beta})$-covering of $\Theta_{i,n}(\lambda, \boldsymbol{\beta}, K)$. Then, any function $g_q \circ \cdots \circ g_0 \in \Theta_n(\lambda, \boldsymbol{\beta}, K)$ belongs to a ball around a composition of centers $g_{q,k_q} \circ \cdots \circ g_{0,k_0}$ for some $\mathbf{k} = (k_0, \ldots, k_q)$ and such that $\|g_i - g_{i,k_i}\|_\infty \leq 3\delta_{in}(\lambda, \boldsymbol{\beta})$. By definition of $\Theta_{i,n}(\lambda, \boldsymbol{\beta}, K)$, the components $(g_{ij,k_i})_j$ of g_{i,k_i} satisfy $|g_{ij,k_i}(\mathbf{x}) - g_{ij,k_i}(\mathbf{y})| \leq 2\delta_{in}(\lambda, \boldsymbol{\beta}) + K|\mathbf{x}-\mathbf{y}|_\infty^{\beta_i \wedge 1}$, for all $\mathbf{x}, \mathbf{y} \in [-1,1]^{d_i}$. Using Lemma 16, the definition of $\delta_{in}(\lambda, \boldsymbol{\beta})$ and the fact that $\alpha_i \leq 1$, gives

$$\begin{aligned}\left\|g_q \circ \cdots \circ g_0 - g_{q,k_q} \circ \cdots \circ g_{0,k_0}\right\|_\infty &\leq K^q \sum_{i=0}^{q} (2\delta_{in}(\lambda, \boldsymbol{\beta}))^{\alpha_i} + (3\delta_{in}(\lambda, \boldsymbol{\beta}))^{\alpha_i} \\ &\leq 5K^q(q+1)\varepsilon_n(\lambda, \boldsymbol{\beta}) \\ &= R\varepsilon_n(\lambda, \boldsymbol{\beta}).\end{aligned}$$

Since there are $N_0 \times \cdots \times N_q$ centers, this concludes the first part of the proof.

We now cover the set $\Theta_n(\lambda, B, K) = \cup_{\boldsymbol{\beta} \in B} \Theta_n(\lambda, \boldsymbol{\beta}, K)$. Denote by $\underline{\boldsymbol{\beta}}$ the smallest element in B, that is, for any $\boldsymbol{\beta} \in B$ we have $\underline{\beta}_i \leq \beta_i$, for all $i = 0, \ldots, q$. Since B is contained in the closed hypercube $[\beta_-, \beta_+]^{q+1}$, we have $\underline{\boldsymbol{\beta}} \in [\beta_-, \beta_+]^{q+1}$. We now show that, for any $i = 0, \ldots, q$ and $\boldsymbol{\beta} \in B$, the set $\Theta_{i,n}(\lambda, \boldsymbol{\beta}, K)$ is contained in the set $\Theta_{i,n}(\lambda, \underline{\boldsymbol{\beta}}, K)$. In fact, by definition, any function $h_i \in \Theta_{i,n}(\lambda, \boldsymbol{\beta}, K)$ satisfies $\overline{h}_{ij} \in \mathcal{D}_i(\lambda, \boldsymbol{\beta}, K)$ and, as a consequence of the embedding in Lemma 15 and the fact that $\delta_{in}(\lambda, \boldsymbol{\beta}) \leq \delta_{in}(\lambda, \underline{\boldsymbol{\beta}})$ by the rate comparison condition (ii) in Assumption 2, one has $\mathcal{C}_{t_i}^{\beta_i}(K) + \mathbb{B}_\infty(2\delta_{in}(\lambda, \boldsymbol{\beta})) \subseteq \mathcal{C}_{t_i}^{\underline{\beta}_i}(K) + \mathbb{B}_\infty(2\delta_{in}(\lambda, \underline{\boldsymbol{\beta}}))$. Thus, $\mathcal{D}_i(\lambda, \boldsymbol{\beta}, K) \subseteq \mathcal{D}_i(\lambda, \underline{\boldsymbol{\beta}}, K)$ and $\Theta_{i,n}(\lambda, \boldsymbol{\beta}, K) \subseteq \Theta_{i,n}(\lambda, \underline{\boldsymbol{\beta}}, K)$. Together with (53), we obtain

$$\begin{aligned}
\mathcal{N}\left(R\varepsilon_n(\lambda, \underline{\boldsymbol{\beta}}), \Theta_n(\lambda, B, K), \|\cdot\|_\infty\right) &\leq \mathcal{N}\left(R\varepsilon_n(\lambda, \underline{\boldsymbol{\beta}}), \Theta_n(\lambda, \underline{\boldsymbol{\beta}}, K), \|\cdot\|_\infty\right) \\
&\leq \prod_{i=0}^{q} \mathcal{N}\left(3\delta_{in}(\lambda, \underline{\boldsymbol{\beta}}), \Theta_{i,n}(\lambda, \underline{\boldsymbol{\beta}}, K), \|\cdot\|_\infty\right).
\end{aligned}$$

We now use the definition of $\Theta_{i,n}(\lambda, \underline{\boldsymbol{\beta}}, K)$ and upper bound the metric entropy by removing the constraint $\mathbb{B}_\infty(1)$ in the definition of $\mathcal{D}_i(\lambda, \underline{\boldsymbol{\beta}}, K)$. This gives

$$\mathcal{N}\left(3\delta_{in}(\lambda, \underline{\boldsymbol{\beta}}), \Theta_{i,n}(\lambda, \underline{\boldsymbol{\beta}}, K), \|\cdot\|_\infty\right) \leq \prod_{j=i}^{d_{i+1}} \mathcal{N}\left(3\delta_{in}(\lambda, \underline{\boldsymbol{\beta}}), \mathcal{C}_{t_i}^{\underline{\beta}_i}(K) + \mathbb{B}_\infty\left(2\delta_{in}(\lambda, \underline{\boldsymbol{\beta}})\right), \|\cdot\|_\infty\right).$$

Any function in $\mathcal{C}_{t_i}^{\underline{\beta}_i}(K) + \mathbb{B}_\infty(2\delta_{in}\left(\lambda, \underline{\boldsymbol{\beta}}\right))$ is at most, in sup-norm distance, $2\delta_{in}(\lambda, \underline{\boldsymbol{\beta}})$-away from some function in $\mathcal{C}_{t_i}^{\underline{\beta}_i}(K)$. Therefore, by applying Lemma 19 with $r = t_i$, $\beta = \underline{\beta}_i$, $\alpha = \underline{\alpha}_i$, and $\delta_n = \delta_{in}(\eta)$,

$$\begin{aligned}
\mathcal{N}\left(3\delta_{in}(\lambda, \underline{\boldsymbol{\beta}}), \mathcal{C}_{t_i}^{\underline{\beta}_i}(K) + \mathbb{B}_\infty\left(2\delta_{in}(\lambda, \underline{\boldsymbol{\beta}})\right), \|\cdot\|_\infty\right) &\leq \mathcal{N}\left(\delta_{in}(\lambda, \underline{\boldsymbol{\beta}}), \mathcal{C}_{t_i}^{\underline{\beta}_i}(K), \|\cdot\|_\infty\right) \\
&\leq e^{n\varepsilon_n(\lambda, \underline{\boldsymbol{\beta}})^2}.
\end{aligned}$$

Assumption 2 ensures that $\varepsilon_n(\lambda, \underline{\boldsymbol{\beta}}) \leq Q\varepsilon_n(\lambda, \boldsymbol{\beta})$, thus combining the last inequalities gives

$$\begin{aligned}
\log \mathcal{N}\left(RQ\varepsilon_n(\lambda, \boldsymbol{\beta}), \Theta_n(\lambda, B, K), \|\cdot\|_\infty\right) &\leq \log \mathcal{N}\left(R\varepsilon_n(\lambda, \underline{\boldsymbol{\beta}}), \Theta_n(\lambda, B, K), \|\cdot\|_\infty\right) \\
&\leq \sum_{i=0}^{q} \sum_{j=1}^{d_{i+1}} n\varepsilon_n(\lambda, \underline{\boldsymbol{\beta}})^2 \\
&= |\mathbf{d}|_1 n\varepsilon_n(\lambda, \underline{\boldsymbol{\beta}})^2 \\
&\leq Q^2 |\mathbf{d}|_1 n\varepsilon_n(\lambda, \boldsymbol{\beta})^2.
\end{aligned}$$

Since $R = 5K^q(q+1)$, we use that $(q+1) \leq |\mathbf{d}|_1 \leq \log(2\log n)$, together with $\log(2\log n) \leq 2\log n$. Thus, $K^q(q+1) \leq (2\log n)^{1+\log K}$ and, with $R_n = 5Q(2\log n)^{1+\log K}$,

$$\log \mathcal{N}\left(R_n\varepsilon_n(\lambda, \boldsymbol{\beta}), \Theta_n(\lambda, B, K), \|\cdot\|_\infty\right) \leq \frac{R_n^2}{25} n\varepsilon_n(\lambda, \boldsymbol{\beta})^2,$$

which concludes the proof. ∎

A.3 Proofs for Section 5

Proof [Proof of Lemma 5] Fix β, r and let ε_T be such that $\varphi^{(\beta,r,K)}(\varepsilon_T) \leq T\varepsilon_T^2$ for all $T \geq 1$. For this choice of (β, r), we show that the concentration function inequality (10) holds for any $0 < \alpha \leq 1$ with $\varepsilon_n(\alpha, \beta, r) := \varepsilon_{m_n}^{\alpha}$, where the sequence m_n is chosen such that $m_n \varepsilon_{m_n}^{2-2\alpha} \leq n$. To see this, observe that

$$\varphi^{(\beta,r,K)}\big(\varepsilon_n(\alpha,\beta,r)^{1/\alpha}\big) = \varphi^{(\beta,r,K)}\big(\varepsilon_{m_n}\big) \leq m_n\varepsilon_{m_n}^2 \leq n\varepsilon_{m_n}^{2\alpha} = n\varepsilon_n(\alpha,\beta,r)^2.$$

By Lemma 3 by Castillo (2008), the function $u \mapsto \varphi^{(\beta,r,K)}(u)$ is strictly decreasing on $u \in (0, +\infty)$, thus any $\overline{\varepsilon}_n(\alpha,\beta,r) \geq \varepsilon_n(\alpha,\beta,r)$ satisfies

$$\varphi^{(\beta,r,K)}\big(\overline{\varepsilon}_n(\alpha,\beta,r)^{1/\alpha}\big) \leq \varphi^{(\beta,r,K)}\big(\varepsilon_n(\alpha,\beta,r)^{1/\alpha}\big) \leq n\varepsilon_n(\alpha,\beta,r)^2 \leq n\overline{\varepsilon}_n(\alpha,\beta,r)^2,$$

which concludes the proof. ∎

Proof [Proof of Lemma 6] *(i):* By Lemma 5, the sequence $\varepsilon_n(\alpha,\beta,r)$ can be obtained from $\varepsilon_n(1,\beta,r)$ via $\varepsilon_n(\alpha,\beta,r) = \varepsilon_{m_n}(1,\beta,r)^{\alpha}$, for any sequence m_n satisfying $m_n\varepsilon_{m_n}(1,\beta,r)^{2-2\alpha} \leq n$. We verify this for the sequence $m_n = C_3(\log n)^{-C_4} n^{(2\beta+r)/(2\beta\alpha+r)}$ with

$$C_3 := \big[C_1(2\beta+1)^{C_2}\big]^{-\frac{(2-2\alpha)(2\beta+r)}{2\beta\alpha+r}} \quad \text{and} \quad C_4 := \frac{(2-2\alpha)(2\beta+r)}{2\beta\alpha+r}C_2.$$

Since $C_1 \geq 1$ and $n \geq 3$, we must have $C_3 \leq 1$, $(\log n)^{-C_4} \leq 1$ and thus also $\log(m_n) \leq (2\beta+1)\log(n)$. Consequently,

$$\begin{aligned}
m_n\varepsilon_{m_n}(1,\beta,r)^{2-2\alpha} &= m_n\left[C_1(\log m_n)^{C_2} m_n^{-\frac{\beta}{2\beta+r}}\right]^{2-2\alpha} \\
&= C_1^{2-2\alpha}(\log m_n)^{C_2(2-2\alpha)} m_n^{\frac{2\beta+r-2\beta+2\beta\alpha}{2\beta+r}} \\
&\leq C_1^{2-2\alpha}\big((2\beta+1)\log(n)\big)^{C_2(2-2\alpha)} m_n^{\frac{2\beta\alpha+r}{2\beta+r}} \\
&\leq C_1^{2-2\alpha}(2\beta+1)^{C_2(2-2\alpha)} C_3^{\frac{2\beta\alpha+r}{2\beta+r}} (\log n)^{C_2(2-2\alpha)-C_4\frac{2\beta\alpha+r}{2\beta+r}} n \\
&= n.
\end{aligned}$$

As shown in the proof of Lemma 5, any sequence $\overline{\varepsilon}_n(\alpha,\beta,r) \geq \varepsilon_n(\alpha,\beta,r)$ is still a solution to the concentration function inequality (10). We now derive a simple upper bound for $\varepsilon_n(\alpha,\beta,r)$. Using that $\alpha \leq 1$ and $\log(m_n) \leq (2\beta+1)\log(n)$, we find

$$\begin{aligned}
\varepsilon_n(\alpha,\beta,r) &= \varepsilon_{m_n}(1,\beta,r)^{\alpha} \\
&\leq C_1^{\alpha}\log(m_n)^{\alpha C_2} m_n^{-\frac{\beta\alpha}{2\beta+r}} \\
&\leq C_1(2\beta+1)^{C_2} C_3^{-\frac{\beta\alpha}{2\beta+r}} (\log n)^{C_2+C_4\frac{\beta\alpha}{2\beta\alpha+r}} n^{-\frac{\beta\alpha}{2\beta\alpha+r}}.
\end{aligned}$$

Using the definition of C_3 together with $0 < \alpha \leq 1$, $(2-2\alpha) \leq 2$ and $2\beta\alpha/(2\beta\alpha+r) \leq 1$ yields

$$C_3^{-\frac{\beta\alpha}{2\beta+r}} = \big[C_1(2\beta+1)^{C_2}\big]^{\frac{(2-2\alpha)\beta\alpha}{2\beta\alpha+r}} \leq C_1(2\beta+1)^{C_2}.$$

Similarly, we get

$$C_4\frac{\beta\alpha}{2\beta\alpha+r}\le C_2\frac{(2-2\alpha)(2\beta+r)}{2\beta\alpha+r}\cdot\frac{1}{2}\le\frac{2\beta+r}{2\beta\alpha+r}\le C_2(2\beta+1).$$

The two previous displays recover the first assertion.

(ii): By assumption, for any $\delta\in(0,1)$, we have $\varphi^{(\beta,r,K)}(\delta)\le C_1'(\log\delta^{-1})^{C_2'}\delta^{-\frac{r}{\beta}}$. We now choose $\delta=\varepsilon_n(\alpha,\beta,r)^{1/\alpha}$ and $\varepsilon_n(\alpha,\beta,r)=C_1'(\beta\log n)^{C_2'}n^{-\beta\alpha/(2\beta\alpha+r)}$. Since $C_1'\ge 1$, $C_2'\ge 0$, and $\log n\ge 1$,

$$\log\varepsilon_n(\alpha,\beta,r)^{-\frac{1}{\alpha}}\le-\frac{1}{\alpha}\log n^{-\frac{\beta\alpha}{2\beta\alpha+r}}\le\frac{\beta}{2\beta\alpha+r}\log n\le\beta\log n.$$

Similarly,

$$\varepsilon_n(\alpha,\beta,r)^{-\frac{r}{\beta\alpha}}\le\left(n^{-\frac{\beta\alpha}{2\beta\alpha+r}}\right)^{-\frac{r}{\beta\alpha}}=n\cdot n^{-\frac{2\beta\alpha}{2\beta\alpha+r}}\le\frac{n\varepsilon_n(\alpha,\beta,r)^2}{(C_1')^2(\beta\log n)^{2C_2'}},$$

and therefore,

$$\varphi^{(\beta,r,K)}\left(\varepsilon_n(\alpha,\beta,r)^{\frac{1}{\alpha}}\right)\le C_1'\left(\log\varepsilon_n(\alpha,\beta,r)^{-\frac{1}{\alpha}}\right)^{C_2'}\varepsilon_n(\alpha,\beta,r)^{-\frac{r}{\beta\alpha}}\le n\varepsilon_n(\alpha,\beta,r)^2.$$

■

Proof [Proof of Lemma 7.] It is sufficient to show that for $n>1$, any composition graph $\lambda=(q,\mathbf{d},\mathbf{t},\mathcal{S})$, and any $\boldsymbol{\beta}'=(\beta_0',\ldots,\beta_q'),\boldsymbol{\beta}=(\beta_0,\ldots,\beta_q)\in I(\lambda)$ satisfying $\beta_i'\le\beta_i\le\beta_i'+1/\log^2 n$ for all $i=0,\ldots,q$, the rates relative to the composition structures $\eta=(\lambda,\boldsymbol{\beta})$ and $\eta'=(\lambda,\boldsymbol{\beta}')$ satisfy $\varepsilon_n(\eta)\le\varepsilon_n(\eta')\le e^{\beta_+}\varepsilon_n(\eta)$.

Since $\varepsilon_n(\eta)=\widetilde{C}_1(\eta)(\log n)^{\widetilde{C}_2(\eta)}\mathfrak{r}_n(\eta)$ with $\widetilde{C}_j(\eta):=\max_{i=0,\ldots,q}\sup_{\beta\in[\beta_-,\beta_+]}C_j(\beta,t_i)$, $j\in\{1,2\}$, we have that $\widetilde{C}_j(\eta)=\widetilde{C}_j(\eta')$ and it is thus sufficient to prove $\mathfrak{r}_n(\eta)\le\mathfrak{r}_n(\eta')\le e^{\beta_+}\mathfrak{r}_n(\eta)$.

Using that $\mathfrak{r}_n(\eta)=\max_{i=0,\ldots,q}n^{-\beta_i\alpha_i/(2\beta_i\alpha_i+t_i)}$ and the fact that the function $x\mapsto x/(2x+t_i)$ is strictly increasing for $x>0$ (its derivative is $x\mapsto t_i/(2x+t_i)^2$), the first inequality $\mathfrak{r}_n(\eta)\le\mathfrak{r}_n(\eta')$ follows. For the second inequality, rewriting the expressions and simplifying the exponents gives

$$\frac{\mathfrak{r}_n(\eta')}{\mathfrak{r}_n(\eta)}\le\max_{i=0,\ldots,q}\min_{j=0,\ldots,q}n^{-\frac{\beta_i'\alpha_i'}{2\beta_i'\alpha_i'+t_i}+\frac{\beta_j\alpha_j}{2\beta_j\alpha_j+t_j}}\le\max_{i=0,\ldots,q}n^{-\frac{\beta_i'\alpha_i'}{2\beta_i'\alpha_i'+t_i}+\frac{\beta_i\alpha_i}{2\beta_i\alpha_i+t_i}}\le\max_{i=0,\ldots,q}n^{\beta_i\alpha_i-\beta_i'\alpha_i'}.$$

We conclude the proof by showing that $|\beta_i\alpha_i-\beta_i'\alpha_i'|\le\beta_+/\log n$. For $u,u',v,v'\ge 0$, we have that $|uv-u'v'|\le u|v-v'|+v'|u-u'|$. In particular, if $u,v'\le 1$, then also $|uv-u'v'|\le|v-v'|+|u-u'|$. By iterating this argument, we find that

$$|\alpha_i-\alpha_i'|\le\sum_{\ell=i+1}^{q}\big|(1\wedge\beta_\ell)-(1\wedge\beta_\ell')\big|\le\sum_{\ell=i+1}^{q}\big|\beta_\ell-\beta_\ell'\big|\le\frac{q-i}{\log^2 n}$$

and thus, $\beta_i\alpha_i-\beta_i'\alpha_i'\le\beta_+|\alpha_i-\alpha_i'|+\alpha_i'|\beta_i-\beta_i'|\le\beta_+(q+1)/\log^2 n$. Since we are restricting ourselves to graphs λ such that $|\mathbf{d}|_1=1+\sum_{i=0}^q d_i\le\log(2\log n)$, we must have $q+1\le\log(2\log n)$. Since $\log x\le x/2$ for all $x>0$, we find $(q+1)\le\log n$. This gives $(q+1)/\log^2 n\le 1/\log n$ and thus Assumption 2 (ii) holds with $Q=e^{\beta_+}$. ■

A.4 Proofs for Section 6

Proof [Proof of Lemma 8.] For the first part of the proof, take a kernel ϕ with $R_\beta := \int_{\mathbb{R}^r} |\mathbf{v}|_\infty^\beta \phi(\mathbf{v}) d\mathbf{v} < +\infty$. Using that $h \in \mathcal{C}_r^\beta(K)$, $\beta \leq 1$ and the change of variable $\mathbf{v}' = \mathbf{v}/\sigma$, we immediately get

$$
\begin{aligned}
\big|(h * \phi_\sigma)(\mathbf{u}) - h(\mathbf{u})\big| &\leq \int_{\mathbb{R}^r} \phi_\sigma(\mathbf{v}) |h(\mathbf{u}-\mathbf{v}) - h(\mathbf{u})|\, d\mathbf{v} \\
&\leq K \int_{\mathbb{R}^r} |\mathbf{v}|_\infty^\beta \phi_\sigma(\mathbf{v})\, d\mathbf{v} \\
&= K \int_{\mathbb{R}^r} |\mathbf{v}|_\infty^\beta \sigma^{-r} \phi(\mathbf{v}/\sigma)\, d\mathbf{v} \\
&= K \int_{\mathbb{R}^r} |\sigma \mathbf{v}'|_\infty^\beta \phi(\mathbf{v}')\, d\mathbf{v}' \\
&\leq K R_\beta \sigma^\beta.
\end{aligned}
$$

This shows $\|h * \phi_\sigma - h\|_\infty \leq K R_\beta \sigma^\beta$ and concludes the first part of the proof.

We now deal with the RKHS norm. Notice that

$$
\begin{aligned}
(h * \phi_\sigma)(\mathbf{u}) &= \int_{\mathbb{R}^r} \widehat{h}(\boldsymbol{\xi}) \widehat{\phi_\sigma}(\boldsymbol{\xi}) e^{-i\mathbf{u}^\top \boldsymbol{\xi}} \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} \\
&= C_\beta^{1/2} \int_{\mathbb{R}^r} \widehat{h}(\boldsymbol{\xi}) \widehat{\phi_\sigma}(\boldsymbol{\xi}) |\boldsymbol{\xi}|_2^{\beta + r/2} \frac{e^{-i\mathbf{u}^\top \boldsymbol{\xi}} - 1}{C_\beta^{1/2} |\boldsymbol{\xi}|_2^{\beta + r/2}} \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} + (h * \phi_\sigma)(0).
\end{aligned}
$$

The RKHS of $Z + X^\beta$ is the direct sum of the space of constant functions $\mathbb{H}^Z$ and $\mathbb{H}^\beta$. If the term $(h * \phi_\sigma)(0)$ is finite, it is a constant and thus belongs to $\mathbb{H}^Z$. Then, the function $h * \phi_\sigma$ is a candidate element of $\mathbb{H}^Z \oplus \mathbb{H}^\beta$ since $h * \phi_\sigma - (h * \phi_\sigma)(0)$ has been represented as a potential element of the RKHS of X^β. We now bound their norm using the isometry property of the norm $\|\cdot\|_{\mathbb{H}^\beta}$, so that

$$
\|h * \phi_\sigma\|_{\mathbb{H}^Z \oplus \mathbb{H}^\beta}^2 \leq 2|(h * \phi_\sigma)(0)|^2 + 2C_\beta \int_{\mathbb{R}^r} |\widehat{h}(\boldsymbol{\xi})|^2 |\widehat{\phi_\sigma}(\boldsymbol{\xi})|^2 |\boldsymbol{\xi}|_2^{2\beta + r} \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}}.
$$

By the change of variable $\boldsymbol{\xi}' = \sigma \boldsymbol{\xi}$, the fact that $\boldsymbol{\xi} \mapsto (1 + |\boldsymbol{\xi}|_2)^\beta \widehat{h}(\boldsymbol{\xi})$ has L^2-norm bounded by K, the property $\widehat{\phi_\sigma}(\boldsymbol{\xi}) = \widehat{\phi}(\sigma \boldsymbol{\xi})$, and choosing ϕ such that $M^2 := \sup_{\boldsymbol{\xi} \in \mathbb{R}^r} |\widehat{\phi}(\boldsymbol{\xi})|^2 |\boldsymbol{\xi}|_2^r < +\infty$, we can bound

$$
\begin{aligned}
&\int |\widehat{h}(\boldsymbol{\xi})|^2 |\widehat{\phi_\sigma}(\boldsymbol{\xi})|^2 |\boldsymbol{\xi}|_2^{2\beta+r} \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} \\
&= \sigma^{-2\beta-2r} \int_{\mathbb{R}^r} |\widehat{h}(\boldsymbol{\xi}/\sigma)|^2 |\widehat{\phi}(\boldsymbol{\xi})|^2 |\boldsymbol{\xi}|_2^{2\beta+r} \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} \\
&= \sigma^{-2\beta-2r} \int_{\mathbb{R}^r} \frac{(1+|\boldsymbol{\xi}/\sigma|_2)^{2\beta}}{(1+|\boldsymbol{\xi}/\sigma|_2)^{2\beta}} |\widehat{h}(\boldsymbol{\xi}/\sigma)|^2 |\widehat{\phi}(\boldsymbol{\xi})|^2 |\boldsymbol{\xi}|_2^{2\beta+r} \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} \\
&\leq \sigma^{-2\beta-2r} \sup_{\boldsymbol{\xi}\in\mathbb{R}^r} \frac{|\widehat{\phi}(\boldsymbol{\xi})|^2 |\boldsymbol{\xi}|_2^{2\beta+r}}{(1+|\boldsymbol{\xi}/\sigma|_2)^{2\beta}} \int_{\mathbb{R}^r} (1+|\boldsymbol{\xi}/\sigma|_2)^{2\beta} |\widehat{h}(\boldsymbol{\xi}/\sigma)|^2 \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} \\
&\leq \sigma^{-2\beta-2r} \sup_{\boldsymbol{\xi}\in\mathbb{R}^r} \frac{|\widehat{\phi}(\boldsymbol{\xi})|^2 |\boldsymbol{\xi}|_2^{2\beta+r}}{|\boldsymbol{\xi}/\sigma|_2^{2\beta}} \int_{\mathbb{R}^r} \sigma^r (1+|\boldsymbol{\xi}|_2)^{2\beta} |\widehat{h}(\boldsymbol{\xi})|^2 \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} \\
&\leq K^2 M^2 \sigma^{-r}.
\end{aligned}
$$

Similarly, by choosing ϕ such that $N^2 := (2\pi)^{-r/2} \int_{\mathbb{R}^r} |\widehat{\phi}(\boldsymbol{\xi})|^2 \, d\boldsymbol{\xi} < +\infty$, we obtain

$$
\begin{aligned}
|(h * \phi_\sigma)(0)|^2 &\leq \left(\int_{\mathbb{R}^r} |\widehat{h}(\boldsymbol{\xi})| |\widehat{\phi_\sigma}(\boldsymbol{\xi})| \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} \right)^2 \\
&= \left(\int_{\mathbb{R}^r} |\widehat{h}(\boldsymbol{\xi})| (1+|\boldsymbol{\xi}|_2)^\beta \frac{|\widehat{\phi}(\sigma\boldsymbol{\xi})|}{(1+|\boldsymbol{\xi}|_2)^\beta} \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} \right)^2 \\
&\leq \left(\int_{\mathbb{R}^r} |\widehat{h}(\boldsymbol{\xi})|^2 (1+|\boldsymbol{\xi}|_2)^{2\beta} \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} \right) \left(\int_{\mathbb{R}^r} \frac{|\widehat{\phi}(\sigma\boldsymbol{\xi})|^2}{(1+|\boldsymbol{\xi}|_2)^{2\beta}} \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} \right) \\
&\leq K^2 \sigma^{-r} \int_{\mathbb{R}^r} \frac{|\widehat{\phi}(\boldsymbol{\xi})|^2}{(1+|\boldsymbol{\xi}/\sigma|_2)^{2\beta}} \frac{d\boldsymbol{\xi}}{(2\pi)^{r/2}} \\
&\leq K^2 N^2 \sigma^{-r}.
\end{aligned}
$$

The proof is complete by taking $L_\beta^2 := 2(C_\beta + 1)(M^2 \vee N^2)$. Since this will be useful for the proof of Lemma 9, the explicit form of the constant C_β is given in (3.67) by Cohen and Istas (2013) as

$$
C_\beta = \frac{\pi^{1/2}\Gamma(\beta+1/2)}{2^{r/2}\beta\Gamma(2\beta)\sin(\pi\beta)\Gamma(\beta+r/2)},
$$

and depends only on $\beta,\ r$. ■

Proof [Proof of Lemma 9] We show that:
(a) Lemma 6 (ii) holds for some $C_1'(\beta, r) \geq 1$ and $C_2'(\beta, r) = 0$.
(b) Any sequence $\varepsilon_n(\alpha, \beta, r) \geq C_1'(\beta, r) n^{-\beta\alpha/(2\beta\alpha+r)}$ solves (10).
(c) Assumption 2 (i) holds for $\varepsilon_n(\alpha, \beta, r) = C_1(\beta, r) n^{-\beta\alpha/(2\beta\alpha+r)}$ with $C_1(\beta, r) := C_1'(\beta, r) \vee Q_1(\beta, r, K)^{\beta/(2\beta+r)}$.
(d) $\sup_{\beta\in[\beta_-,\beta_+]} C_1(\beta, r) < +\infty$.
(e) Assumption 2 (ii) holds for $\varepsilon_n(\eta)$ of the form (22).

By Lemma 6, (a) $\implies$ (b) $\implies$ (c) and by Lemma 7, (d) $\implies$ (e). Thus it remains to prove (a) and (d).

Proof of (a): Fix $\beta \in [\beta_-, \beta_+]$. We have to show that, for all $\delta \in (0,1)$, $\varphi^{(\beta,r,K)}(\delta) \leq C_1'(\beta,r)\delta^{-r/\beta}$, for some constant $C_1'(\beta,r) \geq 1$ depending only on $\beta,\ r,\ K$. We denote the small-ball probability term by $\varphi_0^{(\beta,r)}(\delta) := -\log \mathbb{P}(\|Z + X^\beta\|_\infty \leq \delta)$ and show that

$$(A): \quad \sup_{\delta\in(0,1)} \delta^{r/\beta}\varphi_0^{(\beta,r)}(\delta) < +\infty, \quad (B): \quad \sup_{\delta\in(0,1)} \delta^{r/\beta}\big(\varphi^{(\beta,r,K)}(\delta) - \varphi_0^{(\beta,r)}(\delta)\big) < +\infty.$$

To prove (A), observe that the process $Z + X^\beta$ is the sum of two independent processes, thus its small-ball probability can be bounded by $\log \mathbb{P}(\|Z + X^\beta\|_\infty < \delta) \geq \log \mathbb{P}(\|Z\|_\infty < \delta/2) + \log \mathbb{P}(\|X^\beta\|_\infty < \delta/2)$. It is then sufficient to study the small-ball probabilities of Z and X^β, separately. We now show the following condition, which implies (A),

$$\sup_{\delta\in(0,1)} -\delta^{r/\beta}\log \mathbb{P}(\|X^\beta\|_\infty < \delta) < +\infty, \quad \sup_{\delta\in(0,1)} -\delta^{r/\beta}\log \mathbb{P}(|Z| < \delta) < +\infty. \tag{54}$$

Sharp bounds are known, see Theorem 5.1 by Li and Shao (2001), for the small-ball probability of the fractional Brownian motion X^β. In particular, for $0 < \delta < 1$, we have $-\log \mathbb{P}(\|X^\beta\|_\infty \leq \delta) \leq c_X(\beta,r)\delta^{-r/\beta}$ for a finite constant $c_X(\beta,r)$ depending only on $\beta,\ r$. Since Z is a standard normal, we have $\mathbb{P}(|Z| \leq \delta) = (2\pi)^{-1/2}\int_{-\delta}^{\delta} e^{-x^2/2}dx \geq 2\delta e^{-\delta^2/2}/\sqrt{2\pi}$. With the universal constant $c := 2/\sqrt{2\pi}$, this gives $-\log \mathbb{P}(|Z| \leq \delta) \leq \log(c^{-1}\delta^{-1}) + \delta^2/2$. Therefore, using $\log(x) \leq x^a/a$ for all $x > 1, a > 0$, we get

$$\sup_{\delta\in(0,1)} -\delta^{r/\beta}\log \mathbb{P}(|Z| < \delta) \leq \sup_{\delta\in(0,1)} \delta^{r/\beta}\Big(\frac{\beta}{r}\left(\frac{1}{c\delta}\right)^{r/\beta} + \frac{\delta^2}{2}\Big) = \frac{\beta}{c^{r/\beta}r} + \frac{1}{2} =: c_Z(\beta,r).$$

This concludes the proof of (A).

To prove (B), we apply Lemma 8. In particular, with finite constants $R(\beta,r), L(\beta,r)$ depending only on $\beta,\ r$, take $\sigma = (KR(\beta,r))^{-1/\beta}\delta^{1/\beta}$. Then, any function $h \in \mathcal{C}_r^\beta(K) \cap \mathcal{W}_r^\beta(K)$ can be well approximated by the convolution $h*\phi_\sigma$ in such a way that $\|h - h*\phi_\sigma\|_\infty \leq \delta$ and $\|h*\phi_\sigma\|^2_{\mathbb{H}^Z\oplus\mathbb{H}^\beta} \leq K^2L(\beta,r)^2\delta^{-r/\beta}$. This proves (B) because it gives

$$\sup_{\delta\in(0,1)} \frac{\varphi^{(\beta,r,K)}(\delta) - \varphi_0^{(\beta,r)}(\delta)}{\delta^{-r/\beta}} \leq \sup_{\delta\in(0,1)} \frac{K^2L(\beta,r)^2\delta^{-r/\beta}}{\delta^{-r/\beta}} = K^2L(\beta,r)^2.$$

We have thus concluded the proof of (a), that is, for any $\beta \in [\beta_-, \beta_+]$, condition (ii) in Lemma 6 holds with finite constants

$$C_1'(\beta,r) := c_X(\beta,r) + c_Z(\beta,r) + K^2L(\beta,r)^2, \quad C_2'(\beta,r) = 0.$$

Proof of (d): With the definition of $C_1(\beta,r)$, we want to show that

$$\sup_{\beta\in[\beta_-,\beta_+]} C_1'(\beta,r) \vee Q_1(\beta,r,K)^{\beta/(2\beta+r)} < +\infty.$$

The constant $Q_1(\beta,r,K)$ is given explicitly in Lemma 19 and depends continuously on $\beta > 0$. Thus, $\sup_{\beta\in[\beta_-,\beta_+]} Q_1(\beta,r,K) =: \widetilde{Q}_1 < +\infty$. Since the function $\beta \mapsto \beta/(2\beta+r)$ is increasing for $\beta > 0$, we also have $Q_1(\beta,r,K)^{\beta/(2\beta+r)} \leq \widetilde{Q}_1^{\beta_+/(2\beta_++r)}$.

In the previous part of the proof we have found $C_1'(\beta, r) = c_Z(\beta, r) + c_X(\beta, r) + K^2 L(\beta, r)^2$, thus it remains to prove

$$\sup_{\beta \in [\beta_-, \beta_+]} c_Z(\beta, r) + c_X(\beta, r) + K^2 L(\beta, r)^2 < +\infty. \tag{55}$$

By examining the proof of Lemma 8, we know that $\sup_{\beta \in [\beta_-, \beta_+]} K^2 L(\beta, r)^2 < +\infty$. The explicit form of $c_Z(\beta)$ is given in (54) and so, with $c = 2/\sqrt{2\pi}$,

$$\sup_{\beta \in [\beta_-, \beta_+]} c_Z(\beta, r) \leq \frac{\beta_+}{r c^{\beta_+}} + \frac{1}{2} < +\infty.$$

We now show that the properties of $c_X(\beta, r)$ can be deduced from Theorem 5.2 by Li and Shao (2001). We observe that $\mathbb{E}[|X^\beta(\mathbf{u}) - X^\beta(\mathbf{u}')|^2] = |\mathbf{u} - \mathbf{u}'|_2^{2\beta}$. Furthermore, the function $\beta \mapsto \mathbb{E}[X^\beta(\mathbf{u}) X^\beta(\mathbf{u}')]$ is continuous for all fixed $\mathbf{u}, \mathbf{u}' \in [-1, 1]^r$. In the notation of Theorem 5.2 by Li and Shao (2001), we can take $\sigma_\beta(\delta) := \delta^\beta$ and check that, with $c_1 := 1/2^{\beta_+}$ and $c_2 := 1$, $c_1 \sigma_\beta(2\delta \wedge 1) \leq \sigma_\beta(\delta) \leq c_2 \sigma_\beta(2\delta \wedge 1)$ for all $\delta \in (0, 1)$. The constants c_1, c_2 are chosen to be independent of β in the compact interval $[\beta_-, \beta_+]$. From this, one obtains a constant $c_X(r)$ that only depends on c_1, c_2 and such that $-\log \mathbb{P}(\|X^\beta\|_\infty < \delta) \leq c_X(r) \delta^{-r/\beta}$ for all $\delta > 0$. This shows that the quantity $c_X(\beta, r)$ in (55) can be replaced by $c_X(r)$ and thus is bounded, concluding the proof of (d). ■

Proof [Proof of Lemma 10] Using the definition of X^β yields

$$\|X^\beta\|_{\infty, \infty, \beta} = \max_{j=1, \ldots, J_\beta} \frac{1}{\sqrt{jr}} \max_{k=1, \ldots, 2^{jr}} |Z_{j,k}|.$$

It is known that $\mathbb{E}[\max_{k=1,\ldots,2^{jr}} Z_{j,k}] \leq \sqrt{2\log(2^{jr})}$, a reference is Lemma 2.3 by Massart (2007). For $K' > 1$, using symmetry of $Z_{j,k}$ and the Borell-TIS inequality, e.g. Theorem 2.1.1 by Adler and Taylor (2007),

$$\begin{aligned}\mathbb{P}\left(\max_{k=1,\ldots,2^{jr}} |Z_{j,k}| \geq (1 + K')\sqrt{2\log(2^{jr})} \right) &\leq 2\mathbb{P}\left(\max_{k=1,\ldots,2^{jr}} Z_{j,k} \geq (1 + K')\sqrt{2\log(2^{jr})} \right) \\ &\leq 4\exp\left(-K'^2 \log(2^{jr}) \right).\end{aligned}$$

Combining this with the union bound and the formula for the geometric sum, we obtain for any $K' > 2/\sqrt{r}$,

$$\mathbb{P}\left(\exists j = 1, \ldots, J_\beta, \max_{k=1,\ldots,2^{jr}} |Z_{j,k}| \geq (1 + K')\sqrt{2\log(2^{jr})} \right) \leq \sum_{j=1}^{J_\beta} 2^{j(2 - rK'^2)} \leq \frac{1}{1 - 2^{2 - rK'^2}} - 1.$$

Therefore, with $K' > \sqrt{3}$, on an event with probability at least $1 - 4/(2^{rK'^2} - 4)$, we find

$$\|X^\beta\|_{\infty, \infty, \beta} \leq \max_{j=1,\ldots,J_\beta} \frac{(1 + K')\sqrt{2\log(2^{jr})}}{\sqrt{jr}} = (1 + K')\sqrt{2\log 2}.$$

■

Proof [Proof of Lemma 11] In view of Section 4.3.6 by Giné and Nickl (2016), the Besov space $\mathcal{B}_{\infty,\infty,\beta}$ contains the Hölder space $\mathcal{C}_r^\beta$ for any $\beta > 0$, and they coincide whenever $\beta \notin \mathbb{N}$. Thus there exists K' such that $\mathcal{C}_r^\beta(K) \subseteq \mathcal{B}_{\infty,\infty,\beta}(K')$. We show that:

(a) The assumption of Lemma 6 (i) holds for some $C_1'(\beta,r) \geq 1$ and $C_2'(\beta,r) = 3/2$.

(b) Any sequence $\varepsilon_n(\alpha,\beta,r) \geq C_1'(\beta,r)^2(2\beta+1)^3(\log n)^{3(\beta+1)}n^{-\beta\alpha/(2\beta\alpha+r)}$ solves the concentration function inequality (10).

(c) Assumption 2 (i) holds by taking $\varepsilon_n(\alpha,\beta,r) = C_1(\beta,r)(\log n)^{C_2(\beta,r)}n^{-\beta\alpha/(2\beta\alpha+r)}$ with $C_1(\beta,r) := C_1'(\beta,r)^2(2\beta+1)^3 \vee Q_1(\beta,r,K)^{\beta/(2\beta+r)}$ and $C_2(\beta,r) := 3(\beta+1)$.

(d) $\sup_{\beta\in[\beta_-,\beta_+]} C_1(\beta,r) < +\infty$ and $\sup_{\beta\in[\beta_-,\beta_+]} C_2(\beta,r) < +\infty$.

(e) Assumption 2 (ii) holds for an $\varepsilon_n(\eta)$ of the form (22).

By Lemma 6, (a) $\implies$ (b) $\implies$ (c) and by Lemma 7, (d) $\implies$ (e). Thus it remains to prove (a) and (d).

Proof of (a): We denote the small-ball probability term by $\varphi_0^{(\beta,r)}(\delta) := -\log \mathbb{P}(\|X^\beta\|_\infty \leq \delta)$ and the RKHS term by $\varphi^{(\beta,r,K)}(\delta) - \varphi_0^{(\beta,r)}(\delta)$. We start with the RKHS term. The proof of Theorem 4.5 by van der Vaart and van Zanten (2008a) shows that any function $h \in \mathcal{B}_{\infty,\infty,\beta}(K')$ can be well approximated by its projection h^{J_β} at truncation level J_β. In fact, one has $\|h - h^{J_\beta}\|_\infty \leq K'2^{-J_\beta\beta}/(2^\beta - 1)$ and, with coefficients $\omega_j = 2^{-j(\beta+r/2)}/\sqrt{jr}$,

$$\|h^{J_\beta}\|_{\mathbb{H}^\beta}^2 = \sum_{j=1}^{J_\beta}\sum_{k=1}^{2^{jr}} \lambda_{j,k}(h)^2\omega_j^{-2} \leq K'^2 rJ_\beta \sum_{j=1}^{J_\beta} 2^{jr} \leq K'^2 rJ_\beta^2 2^{J_\beta r}.$$

Recall that J_β is defined as the closest integer to the solution J of $2^J = n^{1/(2\beta+r)}$. By definition, we always have $J_\beta \leq 1 + \log_2 n/(2\beta+r)$ and so $2^{J_\beta} \leq 2n^{1/(2\beta+r)}$ and $2^{-J_\beta\beta} \geq 2^{-\beta}n^{-\beta/(2\beta+r)}$. With all the above, the choice

$$\delta_n := K'\frac{(2^\beta+1)^2}{(2^\beta-1)}\sqrt{r2^r}J_\beta^{3/2}2^{-J_\beta\beta},$$

implies $\|h - h^{J_\beta}\|_\infty < \delta_n$ and

$$\begin{aligned}
\varphi^{(\beta,r,K)}(\delta_n) - \varphi_0^{(\beta,r)}(\delta_n) &\leq K'^2 rJ_\beta^2 2^{J_\beta r} \\
&\leq K'^2 r2^r J_\beta^2 n^{\frac{r}{2\beta+r}} \\
&\leq K'^2\frac{(2^\beta+1)^2}{(2^\beta-1)^2} r2^r J_\beta^2 n^{\frac{r}{2\beta+r}} \\
&\leq nK'^2\frac{(2^\beta+1)^4}{(2^\beta-1)^2} r2^r J_\beta^2 2^{-2J_\beta\beta} \\
&\leq n\delta_n^2.
\end{aligned}$$

We now study the small-ball probability. The proof of Theorem 4.5 by van der Vaart and van Zanten (2008a) shows that, for any sequence $\delta_n \in (0,1)$,

$$\varphi_0^{(\beta,r)}(\delta_n) \leq -\sum_{j=1}^{J_\beta} 2^{jr}\log\left(2\Phi\left(\frac{\delta_n 2^{j\beta}}{\widetilde{K}(\beta)+j^2r^2}\right) - 1\right),$$

where $\widetilde{K}(\beta)$ is chosen in such a way that the function $x \mapsto x^{\beta/r}/(\widetilde{K}(\beta)+\log_2^2(x))$ is increasing for $x \geq 1$. Taking the derivative and imposing it to be positive for $x > 1$ yields $\beta\widetilde{K}(\beta)/r + \log_2^2(x) - 2\log(x)/\log_2^2(2) > 0$, which is solved for any $\widetilde{K}(\beta) \geq 4r/\beta$. van der Vaart and van Zanten (2008a) also shows that the function $f(y) = -\log(2\Phi(y) - 1)$ is decreasing and can be bounded above by $f(y) \leq 1 + |\log y|$ on any interval $y \in [0, c]$. Thus we find

$$\varphi_0^{(\beta,r)}(\delta_n) \leq \sum_{j=1}^{J_\beta} 2^{jr} \left(1 + \left|\log\left(\frac{\delta_n 2^{j\beta}}{\widetilde{K}(\beta) + j^2 r^2}\right)\right|\right).$$

We now show that for sufficiently large n, we have $\delta_n \leq (\widetilde{K}(\beta) + J_\beta^2 r^2) 2^{-J_\beta \beta}$. For this, one has to verify the inequality

$$K' \frac{(2^\beta + 1)^2}{(2^\beta - 1)} \sqrt{r 2^r} J_\beta^{3/2} \leq \widetilde{K}(\beta) + J_\beta^2 r^2,$$

which holds for sufficiently large n since $J_\beta^{3/2} \ll J_\beta^2$. Therefore, using that the function $j \mapsto 1 + \log((\widetilde{K}(\beta) + j^2 r^2)/(\delta_n 2^{j\beta}))$ is decreasing, together with $1 + \log(x) \leq 2\log(x)$ for all $x \geq e$, and $\log(a/b) = \log(a) + \log(1/b)$,

$$\begin{aligned}
\varphi_0^{(\beta,r)}(\delta_n) &\leq \sum_{j=1}^{J_\beta} 2^{jr} \left(1 + \log\left(\frac{\widetilde{K}(\beta) + j^2 r^2}{\delta_n 2^{j\beta}}\right)\right) \\
&\leq J_\beta 2^{J_\beta r} \left(1 + \log\left(\frac{\widetilde{K}(\beta) + r^2}{\delta_n 2^\beta}\right)\right) \\
&\leq 2 J_\beta 2^{J_\beta r} \left[\log\left(\frac{\widetilde{K}(\beta) + r^2}{2^\beta}\right) + \log\left(\frac{1}{\delta_n}\right)\right].
\end{aligned}$$

Then, with the definition of δ_n,

$$\begin{aligned}
\varphi_0^{(\beta,r)}(\delta_n) &\leq 2\left[\log\left(\frac{\widetilde{K}(\beta) + r^2}{2^\beta}\right) + \log\left(\frac{1}{\delta_n}\right)\right] J_\beta 2^{J_\beta r} \\
&\leq 2\left[\log\left(\frac{\widetilde{K}(\beta) + r^2}{2^\beta}\right) + \log\left(\frac{2^{J_\beta \beta}}{K' \frac{(2^\beta+1)^2}{(2^\beta-1)} \sqrt{r 2^r J_\beta^3}}\right)\right] J_\beta 2^{J_\beta r} \\
&\leq 2\left[\log\left(\frac{\widetilde{K}(\beta) + r^2}{2^\beta}\right) + \log\left(2^{J_\beta \beta}\right)\right] J_\beta 2^{J_\beta r} \\
&\leq 2\left[\log\left(\frac{\widetilde{K}(\beta) + r^2}{2^\beta}\right) + \beta\log(2)\right] J_\beta^2 2^{J_\beta r} \\
&\leq 2\left[\log\left(\frac{\widetilde{K}(\beta) + r^2}{2^\beta}\right) + \beta\log(2)\right] 2^r J_\beta^2 n^{\frac{r}{2\beta+r}}.
\end{aligned}$$

The last term in the latter display only depends on n through the quantity $J_\beta^2 n^{\frac{r}{2\beta+r}}$. Since $J_\beta^2 \ll J_\beta^3$, for large enough n the latter display is smaller than $K'^2 \frac{(2^\beta+1)^2}{(2^\beta-1)^2} 2^r J_\beta^3 n^{\frac{r}{2\beta+r}} \leq n\delta_n^2$.

This concludes the proof of (a), since we have shown that the sequence

$$\varepsilon_n(1,\beta,r) := K' \frac{(2^\beta+1)^2}{(2^\beta-1)}\sqrt{r2^r}J_\beta^{3/2}2^{-J_\beta\beta},$$

solves the concentration function inequality (10) for $\alpha = 1$.

Proof of (d): The constant $Q_1(\beta, r, K)$ is given explicitly in Lemma 19 and depends continuously on $\beta > 0$, so $\sup_{\beta\in[\beta_-,\beta_+]} Q_1(\beta,r,K) =: \widetilde{Q}_1 < +\infty$. Since the function $\beta \mapsto \beta/(2\beta+r)$ is increasing for $\beta > 0$, we also have $Q_1(\beta,r,K)^{\beta/(2\beta+r)} \leq \widetilde{Q}_1^{\beta_+/(2\beta_++r)}$.

All the quantities involved in the construction of the rate $\varepsilon_n(1,\beta,r)$ are explicit. It is immediate to see that they are all bounded on the compact interval $\beta \in [\beta_-,\beta_+]$ since $\beta_- > 0$.

This concludes the proof of (d). ■

Proof [Proof of Lemma 12] We show that:
(a) Lemma 6 (i) holds for some $C_1'(\beta,r) \geq 1$ and $C_2'(\beta,r) = (1+r)\beta/(2\beta+r)$.
(b) Any sequence $\varepsilon_n(\alpha,\beta,r) \geq C_1'(\beta,r)^2(2\beta+1)^{2C_2'(\beta,r)}(\log n)^{(2\beta+2)C_2'(\beta,r)}n^{-\beta\alpha/(2\beta\alpha+r)}$ solves the concentration function inequality (10).
(c) Assumption 2 (i) holds for any $\varepsilon_n(\alpha,\beta,r) = C_1(\beta,r)(\log n)^{C_2(\beta,r)}n^{-\beta\alpha/(2\beta\alpha+r)}$ with $C_1(\beta,r) := C_1'(\beta,r)^2(2\beta+1)^{2C_2'(\beta,r)} \vee Q_1(\beta,r,K)^{\beta/(2\beta+r)}$ and $C_2(\beta,r) := (2\beta+2)(1+r)\beta/(2\beta+r)$.
(d) $\sup_{\beta\in[\beta_-,\beta_+]} C_1(\beta,r) < +\infty$ and $\sup_{\beta\in[\beta_-,\beta_+]} C_2(\beta,r) < +\infty$.
(e) Assumption 2 (ii) holds for an $\varepsilon_n(\eta)$ of the form (22).

By Lemma 6, (a) $\implies$ (b) $\implies$ (c) and by Lemma 7, (d) $\implies$ (e). Thus it remains to prove (a) and (d).

Proof of (a): Let $\varphi_a^{(\beta,r,K)}$ the concentration function of the rescaled process $X^\nu(a\cdot)$, then Lemma 11.55 and Lemma 11.56 by Ghosal and van der Vaart (2017) show that, for all $0 < \delta < 1$,

$$\varphi_a^{(\beta,r,K)}(\delta) \leq \Big(C(r)\big(\log(a\delta^{-1})\big)^{1+r} + D(r)\Big)a^r,$$

where $C(r)$ and $D(r)$ are constants that only depend on r and the spectral measure ν of X^ν. It is sufficient to solve the concentration function inequality (10) for $\alpha = 1$. The solution is given in Section 11.5.2 by Ghosal and van der Vaart (2017) as $\varepsilon_n(1,\beta,r) = C_1'(\beta,r)(\log n)^{(1+r)\beta/(2\beta+r)}n^{-\beta/(2\beta+r)}$, for some constant $C_1'(\beta,r) \geq 1$ depending on $\beta,\ r,\ K$.

Proof of (d): The constant $Q_1(\beta, r, K)$ is given explicitly in Lemma 19 and depends continuously on $\beta > 0$, so $\sup_{\beta\in[\beta_-,\beta_+]} Q_1(\beta,r,K) =: \widetilde{Q}_1 < +\infty$. Since the function $\beta \mapsto \beta/(2\beta+r)$ is increasing for $\beta > 0$, we also have $Q_1(\beta,r,K)^{\beta/(2\beta+r)} \leq \widetilde{Q}_1^{\beta_+/(2\beta_++r)}$.

The dependence on β in the concentration function bound only appears in the scaling $a = a(\beta,r)$, since the constants $C(r)$ and $D(r)$ are independent of β. The right side of the latter display depends continuously on the scaling $a = a(\beta,r)$, which in turn is continuous in $\beta \in [\beta_-,\beta_+]$ by construction (24). This gives

$$\sup_{\beta\in\beta\in[\beta_-,\beta_+]} C_1(\beta,r) < +\infty, \qquad \sup_{\beta\in\beta\in[\beta_-,\beta_+]} C_2(\beta,r) \leq (2\beta_++1)(1+r)\frac{\beta_+}{2\beta_++r}.$$

■

A.5 Proofs for Section 7

Proof [Proof of Lemma 13] For two functions $g_k \in \mathcal{C}_1^{\beta_k}(1)$, $k = 1, 2$, and $\beta_1, \beta_2 \leq 1$, we have that $|g_2(g_1(x)) - g_2(g_1(y))| \leq |g_1(x) - g_1(y)|^{\beta_2} \leq |x - y|^{\beta_1 \beta_2}$. Hence, $g_2 \circ g_1 \in \mathcal{C}_1^{\beta_1 \beta_2}(1)$. We now write

$$f = h_q \circ \cdots \circ h_0 = h_q \circ \cdots \circ h_{j+1} \circ \widetilde{h}_j \circ h_{j-2} \circ \cdots \circ h_0,$$

with $\widetilde{h}_j := h_j \circ h_{j-1}$. The right hand side can be written as composition structure $\eta' := (q-1, \mathbf{d}_{-j}, \mathbf{t}_{-j}, \mathcal{S}_{-j}, \boldsymbol{\beta}')$, with $\mathbf{d}_{-j}, \mathbf{t}_{-j}, \mathcal{S}_{-j}, \boldsymbol{\beta}'$ as defined in the statement of the lemma. Due to $\beta_+ \leq 1$, we have $\mathfrak{r}_n(\eta) = \max_{i=0,\ldots,q} n^{-\frac{\gamma_i}{2\gamma_i + t_i}}$, with $\gamma_i = \prod_{\ell=i}^q \beta_\ell$ and it follows that $\mathfrak{r}_n(\eta) = \mathfrak{r}_n(\eta')$. ■

References

Robert J. Adler and Jonathan E. Taylor. *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer, New York, 2007.

Marek Arendarczyk. On the asymptotics of supremum distribution for some iterated processes. *Extremes*, 20(2):451–474, 2017.

Frank Aurzada and Mikhail Lifshits. On the small deviation problem for some iterated processes. *Electron. J. Probab.*, 14:no. 68, 1992–2010, 2009.

François Bolley. Quantitative concentration inequalities on sample path space for mean field interaction. *ESAIM Probab. Stat.*, 14:192–209, 2010.

Krzysztof Burdzy. Some path properties of iterated Brownian motion. In *Seminar on Stochastic Processes, 1992 (Seattle, WA, 1992)*, volume 33 of *Progr. Probab.*, pages 67–87. Birkhäuser Boston, Boston, MA, 1993.

Jérôme Casse and Jean-François Marckert. Processes iterated ad libitum. *Stochastic Process. Appl.*, 126(11):3353–3376, 2016.

Ismaël Castillo. Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.*, 2:1281–1299, 2008.

Serge Cohen and Jacques Istas. *Fractional Fields and Applications*, volume 73 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Heidelberg, 2013. With a foreword by Stéphane Jaffard.

Nicolas Curien and Takis Konstantopoulos. Iterating Brownian motions, ad libitum. *J. Theoret. Probab.*, 27(2):433–448, 2014.

Kurt Cutajar, Edwin V Bonilla, Pietro Michiardi, and Maurizio Filippone. Random feature expansions for deep gaussian processes. In *International Conference on Machine Learning*, pages 884–893. PMLR, 2017.

Andreas Damianou and Neil Lawrence. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

Tadahisa Funaki. Probabilistic construction of the solution of some higher order parabolic differential equation. *Proc. Japan Acad. Ser. A Math. Sci.*, 55(5):176–179, 1979.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2017.

Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York, 2016.

Matteo Giordano, Kolyan Ray, and Johannes Schmidt-Hieber. On the inability of Gaussian process regression to optimally learn compositional functions. *arXiv e-prints*, art. arXiv:2205.07764, May 2022.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2016.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In *International Conference on Machine Learning*, pages 2672–2680. PMLR, 2019.

Kenneth J. Hochberg and Enzo Orsingher. Composition of stochastic processes governed by higher-order parabolic and hyperbolic equations. *J. Theoret. Probab.*, 9(2):511–532, 1996.

Kei Kobayashi. Small ball probabilities for a class of time-changed self-similar processes. *Statist. Probab. Lett.*, 110:155–161, 2016.

W. V. Li and Q.-M. Shao. Gaussian processes: inequalities, small ball probabilities and applications. In *Stochastic Processes: Theory and Methods*, volume 19 of *Handbook of Statist.*, pages 533–597. North-Holland, Amsterdam, 2001.

Andres Felipe Lopez Lopera. *Gaussian Process Modelling Under Inequality Constraints*. Theses, Université de Lyon, September 2019. URL `https://tel.archives-ouvertes.fr/tel-02863891`.

Hassan Maatouk and Xavier Bay. Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, 49(5):557–582, 2017.

Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Alexander Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv e-prints*, art. arXiv:1804.11271, April 2018.

Radford M Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer, New York, NY, 1996.

Dennis Nieman, Botond Szabo, and Harry van Zanten. Contraction rates for sparse variational approximations in Gaussian process regression. *arXiv e-prints*, art. arXiv:2109.10755, September 2021.

Stefano Peluchetti and Stefano Favaro. Infinitely deep neural networks as diffusion processes. In *International Conference on Artificial Intelligence and Statistics*, pages 1126–1136. PMLR, 2020.

Andrew Pensoneault, Xiu Yang, and Xueyu Zhu. Nonnegativity-enforced Gaussian process regression. *Theoretical and Applied Mechanics Letters*, 10(3):182–187, 2020. ISSN 2095-0349.

Nicholas G Polson and Veronika Ročková. Posterior concentration for sparse deep learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 930–941. Curran Associates, Inc., 2018.

Pallavi Ray, Debdeep Pati, and Anirban Bhattacharya. Efficient Bayesian shape-restricted function estimation with constrained Gaussian process priors. *Statistics and Computing*, 30(4):839–853, 2020. ISSN 0960-3174.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.*, 48(4):1875–1897, 2020.

Laura P. Swiler, Mamikon Gulian, Ari L. Frankel, Cosmin Safta, and John D. Jakeman. A survey of constrained gaussian process regression: Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing*, 1(2):119–156, 2020. ISSN 2689-3967.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL https://proceedings.mlr.press/v5/titsias09a.html.

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

A. W. van der Vaart and J. H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. (IMS) Collect.*, pages 200–222. Inst. Math. Statist., Beachwood, OH, 2008b.

Aad W. van der Vaart and Harry van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675, 2009.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.