Submitted to the Annals of Statistics

DEEP HORSESHOE GAUSSIAN PROCESSES

BY ISMAËL CASTILLO[1,a], AND THIBAULT RANDRIANARISOA[2,b]

[1]Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, [a]ismael.castillo@upmc.fr

[2]Department of Statistical Sciences, University of Toronto, [b]t.randrianarisoa@utoronto.ca

Deep Gaussian processes have recently been proposed as natural objects to fit, similarly to deep neural networks, possibly complex features present in modern data samples, such as compositional structures. Adopting a Bayesian nonparametric approach, it is natural to use deep Gaussian processes as prior distributions, and use the corresponding posterior distributions for statistical inference. We introduce the deep Horseshoe Gaussian process Deep–HGP, a new simple prior based on deep Gaussian processes with a squared-exponential kernel, that in particular enables data-driven choices of the key lengthscale parameters. For nonparametric regression with random design, we show that the associated posterior distribution recovers the unknown true regression curve optimally in terms of quadratic loss, up to a logarithmic factor, in an adaptive way. The convergence rates are *simultaneously* adaptive to both the smoothness of the regression function and to its structure in terms of compositions. The dependence of the rates in terms of dimension are explicit, allowing in particular for input spaces of dimension increasing with the number of observations.

1. Introduction.

Gaussian processes (henceforth GPs) are among the most used machine learning methods, with applications ranging from inference in regression models to classification, see e.g. [33] for an overview. Due to their flexibility, in recent years GPs have been used as tools for geometric inference and deep learning. Before turning to deep Gaussian processes, and since our results are also relevant for standard GPs, we start with a brief overview of recent results for Gaussian processes.

A particularly natural field of application where there now exists at least partial theory to explain and validate practical successes of GPs is that of *Bayesian nonparametrics*: the posterior distribution corresponding to taking a GP as prior distribution on functions can be used for function estimation as well as for the practically essential task of *uncertainty quantification*. In a regression setting, when putting a GP prior distribution on the unknown regression function, the corresponding posterior distribution can often be efficiently implemented [33] and comes with theoretical convergence guarantees: the works [46, 48, 8] indeed show that the posterior contraction rate in terms of relevant loss functions (e.g. L^2–loss for regression) is completely determined (both upper and lower bounds) by the behaviour of its concentration function. Shortly thereafter, van der Vaart and van Zanten also showed that statistical *adaptation to smoothness* was possible with GPs with optimal minimax contraction rates by simply drawing at random its scaling parameter [49] in fixed design regression; see [32, 47] for extensions to random design regression and [42] to inverse problems. Results on uncertainty quantification include [40], [50] in nonparametric models and [9, 13] in semiparametric settings.

Let us mention a few applications of posterior distributions arising from GPs that illustrate their flexibility and are related to the setting considered below.

MSC2020 subject classifications: Primary 62G20.

Keywords and phrases: Bayesian nonparametrics, Deep Gaussian processes, multibandwidth Gaussian process priors, fractional posteriors, high-dimensional regression.

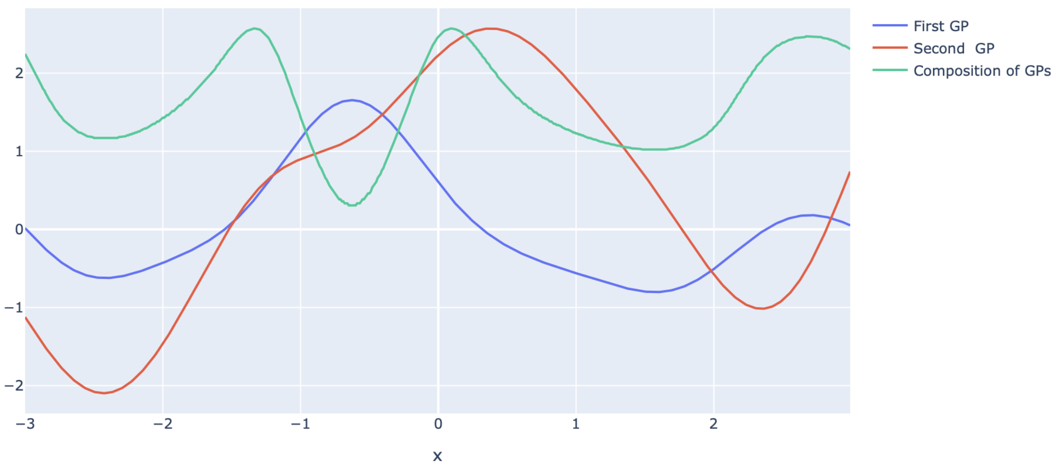
GPs flexibility: geometric settings. In modern statistical models, it is frequent that data naturally sit on a geometric object such as a compact manifold (one can think of a sphere, a swissroll etc.). It is tempting to use GPs in this setting as well, although some care is needed in their construction. For instance, the celebrated GP with squared–exponential kernel (thereafter SqExp) has no immediate analog in a manifold setting, as replacing the euclidian metric in the exponential defining SqExp with the geodesic distance does not form a covariance kernel. This can be remediated by using a kernel coming from heat equation solutions on the manifold [11], and this kernel can be shown to be a natural geometric analog of SqExp. Alternatively, one may put a prior directly on the ambient space equipped with the standard euclidean metric: the authors in [51] obtain a posterior rate that under some (smoothness) conditions adapts to the unknown dimension of the manifold with a rescaled SqExp exponential GP, when the loss function is the quadratic loss but restricted to sit on the manifold; this is further refined in the recent work [41].

GPs flexibility: adaptation to anisotropy and variable selection. By drawing independent lengthscale parameters along different dimensions, [5] shows that posteriors arising from SqExp GPs contract at near-optimal minimax anisotropic rates. A related problem is that of variable selection in (possibly high-dimensional) regression. The unknown regression function may indeed depend only on a few coordinates (although these are not known in advance). By considering variable selection type priors and then drawing lengthscale parameters of SqExp GPs, [52] and [24] provide theory for this setting and respectively investigate optimal rates and variable selection properties for the corresponding posterior distributions.

Recent years have seen a number of remarkable applications of *deep learning* methods, where ‘deep’ typically refers to a certain (often compositional) structure in terms of a number of layers. For instance, deep neural networks are now routinely trained for image or speech processing, giving excellent empirical performance. Theoretical understanding in terms of convergence of statistical procedures is recent and includes [25, 39] for results on empirical risk minimisers for classes of deep neural networks with ReLU activation function in regression settings. A Bayesian counterpart of the results in [39] with theoretical guarantees is considered in [34], where spike-and-slab priors are placed on network coefficients. Sampling directly from the corresponding posterior can be costly due to the combinatorial nature of the search of nonzero network coefficients; the works [4, 15] consider theory and implementation for mean-field variational Bayes versions thereof; the work [10] considers rescaled heavy-tailed priors on weights. Among similarities between GPs and neural networks, it has been shown in [30, 23, 17] that both deep and shallow Bayesian neural networks with random parameters (appropriately rescaled to avoid degeneracy) and with all layers of width growing to infinity asymptotically behave like GPs, with covariance kernel depending on the network structure. The Bayesian approach we describe in the next paragraphs avoids the use of large networks using activation functions by modelling layers directly through independent Gaussian processes.

Deep Gaussian processes [16] (henceforth DeepGPs) correspond to iterated compositions of Gaussian processes and broadly speaking can be seen as a possible Bayesian analogue of deep neural networks. Figure 1 depicts the sample path of a simple DeepGP obtained from two independent GPs with squared-exponential kernels. The random paths resulting from DeepGPs have greater modelling flexibility compared to single Gaussian processes, enabling for instance to capture different spatial behaviours; [22] shows that single GPs cannot reach optimal rates for compositional structures, see also [1]. While the infinite-width limits of deep Bayesian neural networks behave like GPs, forcing instead some layers of the network to be of fixed width while letting others grow leads to a deepGP (see Section 7 of [19]). One then indeed obtains in the limit the composition of the limiting GPs in-between the fixed layers. There is a lot of recent activity for providing efficient sampling methods for deepGPs [18, 35, 36]. Yet, theory is just starting to emerge.

FIGURE 1. *Composition of two Gaussian processes with* SqExp *covariance kernel* $K(s,t) = e^{-(s-t)^2}$.



The recent seminal work by Finocchio and Schmidt–Hieber [19] on deepGPs shows that using a model selection prior to select active variables in the successive Gaussian processes, and conditioning individual GP sample paths to verify certain smoothness constraints, the induced posterior distributions contract nearly optimally and adaptively in quadratic loss for compositional structures in regression, for a variety of kernel choices. Focusing on compositions of constrained GPs (with bounded sample paths and derivatives), the paper [3] uses an adapted concentration function for deepGPs and derives near-optimal contraction rates in density estimation and classification. In [29], using a recursive representation, the authors derive convergence rates for the posterior mean of deepGPs in a regression setting both in a noiseless case and with noisy data. In [1], the authors investigate the use of deepGPs for a class of nonlinear inverse problems.

This work follows the footsteps of [19] and aims at answering the following two questions. The first concerns the possibility to obtain theory and optimality results for a deepGP construction as simple as possible that comes closer to current implementations of deepGPs in practice. The second concerns the possibility to allow for a high-dimensional ambient space as well as a smaller intrinsic dimension.

1. *Can deepGP priors avoid an explicit model selection step?*
   While the deepGP prior construction in [19] is completely natural and 'canonical' from the theoretical perspective, both the conditioning step (to match smoothness constraints) and the model selection prior (for which the posterior on submodels is often expensive to compute) make posterior sampling more challenging in view of practical implementation. One main objective here is to try to simplify the construction of the prior as much as possible while keeping optimality properties, and thereby come closer to the practically used deepGPs, for which lengthscale parameters are often kept free and then adjusted in an empirical Bayes [16] or hierarchical Bayes [38] fashion. In view of this last observation, we propose a prior with a 'soft' model selection based on a prior on lengthscales instead of the previous 'hard' model selection prior.
2. *How do deepGPs scale with respect to 'dimension'?*
   Below we shall allow in some results the input space dimension d to grow with n. Even though any method must then face a 'curse of dimensionality', if the effective 'intrinsic'

dimension of the problem remains fixed or very slowly grows with n, it is conceivable that rates of convergence can still be obtained. While recent work on deep methods has shown that convergence rates only depending on intrinsic dimension(s) can be derived [39], most results are quite generous in the dependence on dimension of the constant factor in the rate. In particular, we are not aware of works allowing for input and 'intrinsic' dimensions to possibly grow with n ([31] considers an example of Bayesian deep ReLU network with growing d but *fixed* intrinsic dimension). We demonstrate below that our construction can adapt to the intrinsic dimension even for a high-dimensional ambient space (with d sublinear in n). This requires a careful tracking of the dependence on dimension, in particular revisiting earlier results in the GP literature to make the dependence on d precise.

The main contributions of the paper are as follows:

1. we introduce a new idea of *freezing–of–paths* for multi-bandwidth Gaussian processes. The benefit of random lengthscales of a stationary kernel for adaptation to *smoothness* has been established for a while [49, 47, 5, 32]. Such an adaptation to the regularity of the underlying truth is made possible by letting the lengthscales grow polynomially with n in a suitable way with sufficient probability under the prior. In the present paper, we show that letting lengthscales appropriately *vanish* (instead of diverge) enables adaptation to *structure* (instead of smoothness) or in other words to adapt to sparsity in the covariates dependence, by 'freezing' irrelevant dimensions through the corresponding posterior distributions. Intuitively, sample paths become almost constant in the directions with vanishing lengthscales, performing effectively a form of 'soft' model selection.
2. we show that the previous two effects of lengthscale parameters, namely adaptation to smoothness and to structure (by using respectively diverging and vanishing lengthscales), can be obtained using a *single* prior for lengthscales: the *horseshoe* distribution [7], that both puts a lot of mass near zero and at the tails, is shown to lead to optimal contraction rates with near-optimal scaling in terms of dimension. Our results also include exponential prior distributions on lengthscales as in earlier contributions on Gaussian processes (e.g. [49]), although dependence on dimension may not be optimal in 'large d' regimes.
3. we study a high-dimensional setting where the input space has growing dimension combined with a compositional structure and functions in the composition having few active coordinates; in particular we allow the input dimension to grow polynomially with n and the number of actually relevant variables in the input layer to grow slowly with n. A main technical contribution of the paper consists in deriving dimension-dependent analogues of the inequalities that are at the heart of GP regression theory with a squared-exponential kernel. Namely, we give precise dependence on ambient and intrinsic dimensions of the metric entropy of the unit ball of the RKHS of the covariance kernel, of the small ball probability of the GP and on quantities measuring approximation properties of this RKHS.

We note that the results are not only relevant for deep learning applications, but also already for shallow (standard) Gaussian processes, for which the freezing-of-paths effect described above is shown to lead to effective 'variable selection' in that the achieved convergence rate only depends on the number of truly present variables. Also, from the technical perspective, in order to leverage the smaller intrinsic dimensionality of the problem, a key new ingredient in the proofs consists in replacing the prior by a 'low-dimensional' oracle GP defined on the relevant coordinates. Finally in this paper for technical convenience we mostly focus on *tempered* posterior distributions, for which the likelihood in Bayes formula is raised to a fixed power ρ smaller than 1; we do however also obtain results for the standard posterior ($\rho = 1$) when the nonparametric prior is coupled with an appropriate prior on the noise variance.

The paper is organised as follows: Section 2 introduces the statistical model and deep Gaussian process priors. We recall the main elements of the frequentist theoretical analysis of GP regression in Section 2.2. Our main results are split into two parts. Section 3 considers a setting without compositions, and Theorem 2 therein shows that under mild conditions multibandwidth GPs effectively achieve a form of variable selection through a freezing-of-paths effect. Section 4 considers adaptation to compositional structures: Theorems 3 and 4 therein demonstrate that deep horseshoe GPs lead to near-minimax optimal contraction rates both in fixed dimensions and in the high-dimensional case, while Section 4.2 explains how results for tempered posteriors can be transferred to standard posteriors. A discussion follows in Section 5. Proofs are provided in Section 6. A key result underlying the proofs and bounding the GPs' concentration function is presented in Section 7. Auxiliary lemmas and their proofs can be found in the appendix [12].

Notation. For two real numbers a, b, we let $a \wedge b = \min(a,b)$, $a \vee b = \max(a,b)$. We denote by ϕ the density of a standard normal random variable. The ε-covering number $N(\varepsilon, S, D)$ of a semimetric space S equipped with a semimetric D is the minimal number of balls of radius ε needed to cover S. For a vector $A = (A_1, \ldots, A_d) \in \mathbb{R}_+^d$, denote

$$\bar{A} := \max_{i=1,\ldots,d} A_i, \qquad |A| = \sum_i A_i,$$

and $A_I = (A_i)_{i \in I}$ for $I \subset [\![1, n]\!]$, the set of all integers between 1 and n. Also, for any vector $\boldsymbol{x} \in \mathbb{R}^d$, we note $|\boldsymbol{x}|_\infty := \max_i |x_i|$ its supnorm. For f integrable on $\mathbb{R}^d$, let $\hat{f}(t) = \int_{\mathbb{R}^d} e^{-i\langle t,s\rangle} f(s) ds$ denote its Fourier transform, with $\langle \cdot, \cdot \rangle$ the euclidean inner product. In the following, $C, C_1, c_1, C_2, c_2, \ldots$ denote absolute constants whose values may change from line to line.

2. The deep horseshoe GP prior.

Consider a nonparametric regression model with random design, where one observes $(X, Y) := (X_i, Y_i)_{1 \leqslant i \leqslant n}$, with $X_1, \ldots, X_n$ independent identically distributed design points sampled from a probability measure μ on I^d for I an interval of $\mathbb{R}$ chosen for simplicity to be $[-1, 1]$ in the sequel and

$$Y_i = f_0(X_i) + \varepsilon_i, \tag{1}$$

for $f_0 : I^d \to \mathbb{R}$ an unknown regression function and ε_i independent $\mathcal{N}(0, \sigma_0^2)$ errors, with σ_0 assumed known for simplicity. We consider estimation of f_0 with respect to the integrated quadratic loss

$$\|f_0 - f\|_{L^2(\mu)}^2 = \int (f_0 - f)^2 d\mu.$$

For a given regression function f, let P_f denote the distribution of one observation (X_i, Y_i) under model (1), which has density

$$p_f(x, y) = \left(2\pi\sigma_0^2\right)^{-1/2} e^{-(y - f(x))^2/(2\sigma_0^2)}$$

with respect to $\mu \otimes \lambda$, for λ the Lebesgue measure on $\mathbb{R}$.

For a real $\beta > 0$, $\lfloor \beta \rfloor$ the largest integer strictly smaller than β and r an integer, let $\mathcal{C}^\beta[-1,1]^r$ denote the classical Hölder space equipped with the norm $\|\cdot\|_{\beta,\infty}$. It consists of functions $f : [-1,1]^r \to \mathbb{R}$ whose norm defined as

$$\|f\|_{\beta,\infty} = \max\left(\max_{|\boldsymbol{\alpha}| \leqslant \lfloor \beta \rfloor} \|\partial^{\boldsymbol{\alpha}} f\|_\infty, \max_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}| = \lfloor \beta \rfloor} \sup_{\boldsymbol{x}, \boldsymbol{y} \in [-1,1]^r,\, \boldsymbol{x} \neq \boldsymbol{y}} \frac{|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{y})|}{|\boldsymbol{x} - \boldsymbol{y}|_\infty^{\beta - \lfloor \beta \rfloor}} \right) \tag{2}$$

is finite, with the multi-index notation $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ and $\partial^{\boldsymbol{\alpha}} = \partial^{\alpha_1} \ldots \partial^{\alpha_d}$. We note that functions with finite Hölder norm are bounded, for any $\beta > 0$.

2.1. *Structural assumptions for multivariate regression.* In order to assess the performance of machine learning methods, a popular benchmark is the regression setting (1) equipped with some 'structural' assumptions. In the unconstrained case where only a smoothness condition is assumed on f_0, rates for β–Hölder smooth functions are typically of the form $n^{-\beta/(2\beta+d)}$, and so are prone to the curse of dimensionality (the rate becomes extremely slow for large d). A common approach is to assume that the multivariate regression function f_0 admits a certain *unknown* 'structure', of 'effective dimension d^*' possibly much smaller than d. For instance, in the simplest setting considered below, f_0 may only depend on a small but unknown number of coordinates. The goal is then to find algorithms that are able to achieve optimal risk bounds that adapt to the unknown underlying structure, and that therefore scale with d^* instead of d.

A first basic setting: effective variable selection. Let us first consider the simple setting where $f_0 : [-1,1]^d = I^d \to \mathbb{R}$ only depends on d^* variables, that is

$$f_0(x_1,\ldots,x_d) = g(x_{i_1},\ldots,x_{i_{d*}}), \tag{3}$$

for some $g \in \mathcal{C}^\beta(I^{d^*})$ and $\beta > 0$. The subset of indices $i_1,\ldots,i_{d*}$ is unknown to the statistician and the target convergence rate in quadratic loss is $n^{-\beta/(2\beta+d^*)}$. We call this setting *effective variable selection*, where by this we mean that one aims at achieving the same performance as if the indices of the truly present variables were known. We note that we do not consider here the problem of *actual* variable selection, where the goal would be to recover this set of indices, and which would require some further conditions; we refer to [24] for more details on this task.

Define, for $K,\beta > 0$ and $d^* \leqslant d$ two integers, and recalling that $I = [-1,1]$,

$$\mathcal{F}_{VS}(K,\beta,d,d^*) = \Big\{ f_0 : I^d \to \mathbb{R} \text{ such that (3) holds for some } g \in \mathcal{C}^\beta(I^{d^*}), \\ \|g\|_\infty \leqslant 1 \text{ and } \|g\|_{\beta,\infty} \leqslant K \Big\}. \tag{4}$$

As in recent works in deep learning and similar to [19], we assume that an upper-bound M_0 is known for the true function f_0 and without loss of generality we assume $M_0 = 1$.

Compositional structure. Following [39], suppose that f can be written as a composition

$$f = h_q \circ \cdots \circ h_0, \tag{5}$$

with $h_i : I^{d_i} \to I^{d_{i+1}}$, for (d_i) a sequence of integers such that $d_0 = d$ and $d_{q+1} = 1$. Since h_i takes values in $\mathbb{R}^{d_{i+1}}$, one may write $h_i = (h_{ij})$, where h_{ij} for $j = 1,\ldots,d_{i+1}$ are its univariate coordinate functions.

The compositional class $\mathcal{F}_{deep}(\lambda,\beta,K)$. Let us further assume that h_{ij}'s as above only depend on a subset $\mathcal{S}_{ij}$ of at most $t_i \leqslant d_i$ variables, and h_{ij} restricted to the variables $\mathcal{S}_{ij}$ belongs to $\mathcal{C}^{\beta_i}(I^{t_i})$. For $\lambda = (q,d_1,\ldots,d_q,t_0,\ldots,t_q)$, $\beta = (\beta_0,\ldots,\beta_q)$ and $K > 0$, let

$$\mathcal{F}_{deep}(\lambda,\beta,K) = \Big\{ h_q \circ \cdots \circ h_0 : I^d \to I,\ h_i : I^{d_i} \to I^{d_{i+1}},\ h_{ij} \in \mathcal{F}_{VS}(K,\beta_i,d_i,t_i) \Big\}. \tag{6}$$

These compositional classes encompass several well-studied structural models in the literature. For instance, in an additive model with a fixed number $D > 0$ of covariates, the regression function can be expressed as

$$f_0(x_1,\ldots,x_D) = \sum_{i=1}^{D} g_i(x_i) = h_1 \circ h_0(x), \tag{7}$$

where $h_0(x) = (g_1(x_1), \dots, g_d(x_D))$ and $h_1(y) = \sum_{i=1}^D y_i$. The original function f_0 is then the composition of two functions: one where each component is univariate and depends on a single variable (h_0) so that $d_1 = t_0 = 1$, and another that depends on all variables but is infinitely smooth (h_1), so $t_1 = D$. Therefore, further assuming $g_i \in \mathcal{C}^\beta(I)$, $\|g_i\|_\infty \leqslant 1$ and $\|g_i\|_{\beta,\infty} \leqslant K$ and , we have $f_0 \in \mathcal{F}_{\text{deep}}((1,1,1,D),(\beta,\beta'),\max(D,K))$ for arbitrarily large β' (taking $M_0 = D$ in this case). We refer the reader to Section 4 of [39] for more examples highlighting the link between compositional classes and usual structural constraints in nonparametric regression.

Minimax optimal rate. The minimax rate of estimation in quadratic loss over this class

$$(r_n^*)^2 = \inf_T \sup_{f \in \mathcal{F}_{deep}(\lambda,\beta,K)} E_f \|T - f\|_2^2,$$

for T ranging over the set of estimators of f, is, up to logarithmic factors,

$$r_n^* \asymp \max_{i=0,\dots,q} n^{-\frac{\beta_i \alpha_i}{2\beta_i\alpha_i + t_i}}, \qquad \text{where } \alpha_i := \prod_{l=i+1}^q (\beta_l \wedge 1), \tag{8}$$

under the mild condition $t_i \leqslant \min(d_0, \dots, d_{i-1})$, see [39]. For example, in the above additive model, for $\beta' \geqslant 1 \vee D\beta$, the rate becomes, up to a multiplicative constant depending polynomially on D,

$$n^{-\frac{\beta}{2\beta+1}} \vee n^{-\frac{\beta'}{2\beta'+D}} = n^{-\frac{\beta}{2\beta+1}}.$$

These fast rates are attainable because for both functions in the composition (7), fast rates can be achieved without suffering from the curse of dimensionality (D will still feature as a multiplicative constant in the rate, but importantly not in the exponent of $1/n$).

2.2. *Key ingredients.* *Posterior distributions and frequentist analysis.* Given a prior distribution Π on regression functions, the posterior mass $\Pi[B \mid X, Y]$ of a measurable set B is given by Bayes' formula: this is the next display for $\rho = 1$. More generally, one may set, for any $\rho \in (0, 1]$,

$$\Pi_\rho[B \mid X, Y] = \frac{\int_B \prod_{1 \leqslant i \leqslant n} p_f(X_i, Y_i)^\rho d\Pi(f)}{\int \prod_{1 \leqslant i \leqslant n} p_f(X_i, Y_i)^\rho d\Pi(f)}.$$

When $\rho = 1$ this is the usual conditional probability that f belongs to B given the data. If $0 < \rho < 1$, this quantity is called *ρ–posterior* (or tempered posterior). Its use is very much widespread in machine learning, in particular in PAC–Bayesian settings [14, 54]. We use the tempered posterior in our main results, and also provide results for the standard posterior $\rho = 1$ and an augmented prior in Section 4.2. Links with the case $\rho = 1$ are also further discussed in Section 5.

Gaussian process (ρ–) posteriors: theory. For any centered Gaussian process W on the Banach space of continuous functions on I^d equipped with the $\|\cdot\|_\infty$ norm, the probability measure of any ball $\{f : \|f - g\|_\infty < \varepsilon\}$ is lower bounded by a quantity depending on the mass of the centered ball of radius ε and on how well g can be approximated by elements of the RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ of the covariance kernel of the process. More precisely, according to Proposition 11.19 of [21], for any $\varepsilon > 0$,

$$P[\|W - g\|_\infty < \varepsilon] \geqslant e^{-\varphi_g(\varepsilon/2)},$$

$$\varphi_g(\varepsilon) = \inf_{h \in \mathbb{H}:\ \|h-g\|_\infty \leqslant \varepsilon} \frac{1}{2}\|h\|_{\mathbb{H}}^2 - \log P[\|W\|_\infty < \varepsilon]. \tag{9}$$

The function φ_g in (9) is called the *concentration function* of the Gaussian process W and plays a key role for contraction rates of GPs [48, 8]. It is the sum of two terms: the first term with the infimum is the approximation term whereas the second term is called the small ball term (the probability within the log is the small ball probability of the process W).

In nonparametric regression with fixed design, [49] proved that posterior contraction rates that are adaptive to the unknown smoothness of the regression function are achievable for stationary Gaussian process priors, with a dilatation parameter of the sample paths distributed as a Gamma variable. As a particular case, consider the *squared exponential* process SqExp defined as the zero-mean Gaussian process with covariance kernel $K(s,t) = \exp(-\|t-s\|^2)$ (and $\|\cdot\|$ the euclidean norm) on $[-1,1]^d$. Next, for $k, \theta > 0$, one sets

$$A^d \sim \text{Gamma}\,(k,\theta)$$

$$f \mid A \sim \left\{W_{At} : \; t \in [-1,1]^d\right\}.$$

This construction induces a prior on the Banach space of continuous functions for which the posterior concentrates in the empirical L_2-norm at rate $\varepsilon_n \asymp n^{-\beta/(2\beta+d)}$ (up to a log factor) whenever f_0 has β-Hölder regularity, $\beta > 0$.

Although this rate coincides with the minimax estimation rate over a ball in $\mathcal{C}^\beta[-1,1]^d$, it becomes very slow for large d. When the regression function f_0 depends on a small number of variables d^* only, a special case of the add-GP prior from [52] gives optimal posterior contraction rates $n^{-\beta/(2\beta+d^*)}$ without the need to estimate d^*. This is achieved by the introduction of an additional layer in the prior, drawing via Bernoulli random variables in which direction the Gaussian sample paths have to be dilated (the sample paths being constant in the other directions). From a practical point-of-view, this 'hard' selection of variables adds a combinatorial complexity to posterior sampling. Similarly, if the design points are located on a d^*-dimensional Riemannian manifold, $d^* < d$, of the ambient space $[-1,1]^d$, we expect the faster rate $n^{-\beta/(2\beta+d^*)}$ to be attainable. The work [51] achieves it with a dilated Gaussian process as well, the dilatation factor A being distributed as $A^{d^*} \sim \text{Gamma}\,(k,\theta)$. One may note that this approach requires an estimate of d^* to be applied and that posterior contraction rates are obtained for local distances *on the manifold* (such as the empirical L_2-norm) only, but not on the ambient space.

Technically, in the works above, adaptation to smoothness is achieved using that the distribution on A^d in the last display selects large values for the dilatation parameters with large enough probability, together with a study of the dependence of the concentration function (9) on the lengthscale parameter. In particular, [49] (isotropic case) and [5] (anisotropic case) develop a theory of approximation of (Hölder–)smooth functions in $\mathbb{R}^d$ by elements of RKHS from such Gaussian processes, in the case $A \to \infty$. In our results below on 'effective' variable selection, the regime $A \to 0$ corresponding to small lengthscale will be particularly relevant, see Section 3 for more on this. We note also that our results on posterior contraction will be expressed in the natural global L^2 loss (in contrast to a loss e.g. restricted only on active directions).

2.3. *Deep Horseshoe Gaussian Process prior.* We introduce a Gaussian process prior with independent inverse–lengthscales distributed following a half-horseshoe distribution. This distribution possesses two interesting properties for our goals. Its density has a pole at 0, which allows to 'freeze' irrelevant dimensions, drawing small inverse lengthscales with high probability. It also has heavy tails, so that it performs an adequate scaling on the ambient dimensions with sufficiently large probability.

The single-layer case. In the following, we use the real map

$$\Psi : \; x \mapsto (x \wedge 1) \; \vee \; (-1).$$

For π a density function on $\mathbb{R}_+$, consider the following prior Π on regression functions f

$$\begin{aligned} A_j &\overset{\text{i.i.d.}}{\sim} \pi \\ f \mid (A_1, \ldots, A_d) &\sim \Psi(W^A), \end{aligned} \tag{10}$$

where

$$W^A = \left\{ W_{(A_1 s_1, \ldots, A_d s_d)} : \; s = (s_1, \ldots, s_d) \in [-1,1]^d \right\},$$

for W the squared exponential process SqExp. For a given value of A, we call W^A a multi-bandwidth Gaussian process.

Although the theory below can be applied to arbitrary scaling distributions π, we consider two main examples in the sequel: an exponential and a horseshoe distribution. Let us define π_τ, $\tau > 0$, as the (half-) *horseshoe* density (introduced in [7]) i.e. the density of a random variable X_τ distributed as

$$\begin{aligned} \xi &\sim C^+(0,1) \\ X_\tau \mid \xi \; &\sim \mathcal{N}^+(0, \tau^2 \xi^2), \end{aligned}$$

with $C^+(0,1)$ a standard half-Cauchy distribution and $\mathcal{N}^+(\mu, \sigma^2)$ the half-normal distribution of $|X|$, $X \sim \mathcal{N}(\mu, \sigma^2)$. We refer to [44] for posterior contraction results in the case of estimation of sparse vectors, with priors based on π_τ and discussions on the influence of τ.

When $\pi = \pi_\tau$ the horseshoe density on $\mathbb{R}_+$ in (10), we call the above hierarchical prior the *Horseshoe Gaussian Process* prior and denote it $\mathsf{HGP}(\tau)$.

The multi-layer case. In order to perform inference in more complex models, we introduce a deepGP-type prior, mixing ideas from [19] and the just introduced prior (10).

We first place discrete priors Π_q on the number of layers q and $\Pi_d[d_1, \ldots, d_q | q]$ on the successive ambient dimensions in the composition (5). We assume that $\Pi_q[q] > 0$ and $\Pi_d[d_1, \ldots, d_q | q] > 0$ for any integers $q \geqslant 0$ and $d_i \geqslant 1$.

Given $q, d_1, \ldots, d_q$, we define a random regression function $f = W_q \circ \cdots \circ W_0$ where, for $i = 0, \ldots, q$, the map $W_i \colon I^{d_i} \to \mathbb{R}^{d_{i+1}}$ is a multivariate Gaussian process indexed by I^{d_i} and to which Ψ is applied element-wise. We assume that for $j = 1, \ldots, d_{i+1}$, the coordinates $(W_i)_j$ are independently (accross i and j) and identically (across j) distributed as a prior of the form (10). Constraining the sample paths between -1 and 1 ensures that the composition is well-defined.

The *Deep Horseshoe Gaussian Process* $\mathsf{Deep\text{–}HGP}$ is a special case of this construction where the prior on lengthscales is a horseshoe prior: it is defined as the hierarchical prior

$$\begin{aligned} q &\sim \Pi_q \\ d_1, \ldots, d_q \mid q &\sim \Pi_d[\cdot | q] \\ g_{ij} \mid (q, d_1, \ldots, d_q) &\overset{\text{ind.}}{\sim} \mathsf{HGP}(\tau_i) \\ f \mid (q, d_1, \ldots, d_q, g_{ij}) &= g_q \circ \cdots \circ g_0 \end{aligned} \tag{11}$$

for $\tau_i > 0$, $i = 0, \ldots, q$. In the above, $g_i = (g_{ij})_j$ with $g_{ij} : I^{d_i} \to I$, $1 \leqslant j \leqslant d_{i+1}$. At the j–th level of the composition, there are d_{j+1} coordinate functions (g_{ij}) distributed as the HGP process, each of these functions depending on d_j parameters (and setting $d_{q+1} = 1$ the output dimension). Note that on each layer j, all d_j variables are present simultaneously as input to each Gaussian process component, although the scalings (distributed as horseshoe random variables) calibrate the 'strength' (or 'importance') of these variables.

In the construction (11), the depth q and dimensions d_i are given prior distributions, which is perhaps the most natural Bayesian way to model these unknown quantities. However, simulating from the posterior distributions of q and d may be expensive, so it is common practice in this setting to take these parameters to be fixed 'large enough' values, say $q = q_{max}$ and $d_i = d_{max}$ for all $i \leqslant q$. This leads to the following simpler prior Deep–HGP(q_{max}, d_{max})

$$g_{ij} \overset{\text{ind.}}{\sim} \mathsf{HGP}(\tau_i)$$
$$f \mid (g_{ij}) = g_q \circ \cdots \circ g_0 \tag{12}$$

for $\tau_i > 0$, $i = 0, \ldots, q_{max}$ and now $g_i = (g_{ij})_j$ with $g_{ij} : I^{d_{max}} \to I$, for $1 \leqslant j \leqslant d_{max}$ and $i > 0$. Our main result on deep Gaussian processes in fixed dimensions, Theorem 3 below, shows that both constructions of Deep–HGP, with either random or fixed q, d, lead to near-optimal and adaptive rates for compositions, with the only restriction for Deep–HGP(q_{max}, d_{max}) being that the 'true' dimensions are indeed smaller than q_{max}, d_{max}.

In the following, we use q, d_i to denote both the parameters of the class $\mathcal{F}_{deep}(\lambda, \beta, K)$ in (6) and the hyperparameters of the prior (11), as the context will make it clear what we are referring to.

3. Main results I: shallow case and "freezing of paths".

To gain intuition on our proposed procedure, we now build up progressively the results, starting from a standard smoothness condition in regression (no composition) where the regression function depends on an effective number of coordinates possibly smaller than the input dimension d. Section 3.1 presents a simple oracle result while Section 3.2 considers more precise results allowing for adaptation and growing dimension. The 'deep' case of compositional structures is considered in Section 4. Recall that σ_0^2 is the variance of the noise in (1) and set

$$\xi := 2\sigma_0^2 / \sqrt{1 + 4\sigma_0^2}. \tag{13}$$

For simplicity, in Sections 3 and 4 we mostly focus on ρ–posteriors for given $\rho < 1$ (e.g. $\rho = 1/2$) and assume that σ_0^2 is known. We refer to Section 4.2 below for results on the standard posterior and possibly unknown noise variance.

3.1. *"Freezing of paths" for effective variable selection: a new property of scalings of Gaussian processes.*

The first result assumes that the true regression function depends on a number d^* of coordinates only, and that for now the indices of the active variables are *known*.

THEOREM 1 (freezing of paths). *Let $d \geqslant 2$ be a fixed integer and for $K \geqslant 1, \beta > 0$, set $\mathcal{F}(K) := \mathcal{F}_{VS}(K, \beta, d, d^*)$. Fix $\rho \in (0,1)$, let $f_0 \in \mathcal{F}(K)$ and suppose*

$$f_0(x_1, \ldots, x_d) = g_0(x_{i_1}, \ldots, x_{i_{d^*}}),$$

with $S_0 := \{i_1, \ldots, i_{d^*}\} \subset \{1, \ldots, n\}$ and some $1 \leqslant d^* \leqslant d$. Let Π be a multibandwidth prior

$$f(x) \sim \Psi(W_{(a_1 x_1, \ldots, a_d x_d)})$$

with W a d–dimensional SqExp Gaussian process with deterministic scaling parameters

$$a_i = \begin{cases} n^{\frac{1}{2\beta + d^*}}, & \text{if } i \in S_0 \\ 1/\sqrt{n}, & \text{if } i \notin S_0 \end{cases}. \tag{14}$$

Then, there exists $M > 0$ such that the ρ–posterior $\Pi_\rho[\cdot \mid X, Y]$ verifies

$$\sup_{f_0 \in \mathcal{F}(K)} E_{f_0} \Pi_\rho \left[f : \|f - f_0\|_{L^2(\mu)} \geqslant M \log^{1+d^*}(n)\, n^{-\frac{\beta}{2\beta + d^*}} \mid X, Y \right] \to 0.$$

Theorem 1 shows that by taking GP lengthscale parameters that are very small, of order $1/\sqrt{n}$, for the coordinates j such that f_0 in fact does not depend on x_j, and taking lengthscales equal to the 'standard nonparametric cut-off' $n^{1/(2\beta+d^*)}$ (for estimating β–smooth functions in dimension d^*) on the other coordinates, leads to an optimal minimax contraction rate $n^{-\beta/(2\beta+d^*)}$ for the integrated squared loss up to a logarithmic factor for the ρ–posterior distribution (the power in the log factor is improved in the next result). Inspection of the proof shows that for $i \notin S_0$, one may take a_i to be any value smaller than $C/\sqrt{n}$ for some fixed $C > 0$.

The intuition behind the result is that taking a small lengthscale for coordinate i 'freezes' the GP path along this coordinate, making it almost constant in that coordinate, which corresponds to the limiting case $a_i = 0$. Note that Theorem 1 is an 'oracle' result in that it assumes both β and d^* (and even the indices i_j) to be *known*. Adaptive versions are considered below. While the result is somewhat expected if one sets $a_i = 0$ for $i \notin S_0$ (this would correspond to a 'hard' variable selection), the fact that the rate remains optimal for small but non-zero values of a_i suggests that there may be room for a 'soft' variable selection procedure that would allow for small a_i in a data-driven way: this is the purpose of our next Theorem.

3.2. *Single layer setting: horseshoe GP.* The next statement is our main result on effective variable selection for (non-deep) Gaussian processes – recall that by this we mean achieving the same rates as if active coordinates were known (not recovering the truly active ones) –. It is a non-asymptotic result that allows for dimensions varying with n. It is stated for an arbitrary prior π on lengthscales. We then particularise it over the next paragraphs by stating simpler asymptotic versions and giving examples of lengthscale priors that satisfy the conditions. We consider both fixed dimensions and the case where both d^*, d vary with n.

THEOREM 2 (Single layer, generic result). *Let $1 \leqslant d^* \leqslant d$ be two integers and for $K \geqslant 1, \beta > 0$, set $\mathcal{F}(K) := \mathcal{F}_{VS}(K,\beta,d,d^*)$. Let ξ be as in (13) and fix $0 < \rho < 1$. Let Π be a multibandwidth prior (10) with density π on scaling parameters that satisfies*

$$(15)\qquad \left(\int_0^{\frac{\xi}{8d\sqrt{\rho n}}} \pi(a)da\right)^{d-d^*} \left(\int_{a^*}^{2a^*} \pi(a)da\right)^{d^*} \geqslant 2\exp(-n\rho\varepsilon_n^2/2),$$

where $a^* = a_n^*$ and ε_n verify $a^* \geqslant 1$, that $1/\sqrt{n} \leqslant \varepsilon_n \leqslant 1$, and

$$(16)\qquad 64\xi^{-2} \geqslant \varepsilon_n^2 \geqslant \{B_1 {a^*}^{-2\beta}\} \vee \{B_2 {a^*}^{d^*} \log^{1+d^*}(n)/n\},$$

where $B_1 = c_1K^2c_2^{d^*}$ and $B_2 = K^2(c_3 d^{*c_4})^{d^*}$, with $c_1,\dots,c_4 \geqslant 1$ constants depending only on β, ξ and ρ. Then, there exists $M = M(\rho,\xi) > 0$ such that, for $n \geqslant 3$,

$$\sup_{f_0\in\mathcal{F}(K)} E_{f_0}\Pi_\rho[f : \|f-f_0\|_{L^2(\mu)} \geqslant M\varepsilon_n \,|\, X, Y] \leqslant \frac{1}{n\varepsilon_n^2} + e^{-\rho n\varepsilon_n^2}.$$

Theorem 2 gives a contraction of the ρ–posterior distribution at rate ε_n around the true f_0, provided $n\varepsilon_n^2$ is suitably large. A more explicit expression of ε_n is given in the next Corollary.

COROLLARY 1 (Optimal a^* and posterior rate). *Optimising a^* in (16), leads to setting*

$$(17)\qquad (a^*)^{2\beta+d^*} = (B_1 n)/(B_2 \log^{1+d^*}(n)) \vee 1.$$

Condition (16) then becomes, for a^* as in (17),

$$(18)\qquad 64\xi^{-2} \geqslant \varepsilon_n^2 \geqslant \left[B_3(\log n)^{\frac{2\beta(1+d^*)}{2\beta+d^*}} n^{-\frac{2\beta}{2\beta+d^*}}\right] \vee \left[B_2\log^{1+d^*}(n)/n\right],$$

where $B_3 = K^2 c_5^{d^*}(d^*)^{c_6 d^*/(2\beta+d^*)}$, recalling $B_2 = K^2(c_3 d^{*c_4})^{d^*}$, and $c_3, \dots, c_6 \geqslant 1$ are constants only depending on β, ξ, ρ. If additionally (15) holds (conditions for this are given below), then the ρ–posterior rate can be taken as the right-hand side of (18).

In the asymptotic regime $n \to \infty$, for ε_n taken equal to the right hand-side of (18), it follows that $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$, so that provided (15) holds, the posterior mass in the last display of Theorem 2 goes to 0, and the posterior contracts to f_0 at rate ε_n asymptotically. For reasonable (i.e. fixed or growing slowly with n) values of d^*, the first term in (18) dominates and, again under (15), the resulting rate ε_n goes to 0 with n. Next we investigate a few examples of priors for which (15) holds with a resulting ε_n given by (18).

The proof of Theorem 2 is based on considering an oracle process defined on the d^* relevant dimensions. The rates can then carry over to the full prior thanks to Condition (15), which ensures that the lengthscale prior π tunes down irrelevant dimensions, so that the difference between the processes is small with high probability. These deviations are controlled via new dimension–dependent estimates of characteristics of the squared-exponential Gaussian process (via its RKHS), see Theorem 5 and Lemmas A.2–A.4 [12], coupled with concentration of measures tools. This theorem is a main theoretical novelty of the paper: a lengthscale prior which puts large mass on small values allows to 'freeze' irrelevant directions, so that the overall prior behaves like a smaller-dimensional one.

The case $A \to 0$ corresponding to a *freezing-of-paths* effect has not been studied so far to the best of our knowledge; although it is conceivable that a study of the RKHS of the SqExp process in the small A regime (similar to the case $A \to \infty$ discussed in Section 2.2) would also lead to a proof of Theorem 2, the case $A \to 0$ proves to be quite challenging: for instance, one may note that constant functions do not belong to the RKHS of the squared-exponential process; one should then find the best possible rate of approximation of constants by elements of the RKHS. This is why, in the present work, we have followed a different route by directly comparing to an oracle process, as explained in the previous paragraph.

Fixed dimensions. Let us now examine the case where the dimensions d, d^* are fixed, independent of n. We derive conditions on two natural priors: an exponential prior, as used e.g. in [48] for adaptation to smoothness, and a horseshoe prior.

EXAMPLE 1 (Exponential prior with fixed scaling λ). Consider $\pi(a) = \lambda e^{-\lambda a} 1\!\!1_{a>0}$ an exponential prior of parameter $\lambda > 0$. A simple calculation, see Lemma G.16 [12], shows that (15) is verified if

$$n\varepsilon_n^2 \geqslant (2/\rho)\left[d \log\left(\frac{16d\sqrt{\rho n}}{\xi\lambda}\right) + 2\lambda d^* a^* + \log 2\right], \tag{19}$$

as long as $\lambda \in [1/a^*, 8d\sqrt{\rho n}/\xi]$. As a particular case, for fixed d, d^*, K, λ, and a^* as in (17), this condition is automatically satisfied for large enough n if (18) holds.

EXAMPLE 2 (Horseshoe prior with fixed parameter τ). Consider $\pi = \pi_\tau$ a horseshoe prior of parameter $\tau > 0$. Then (15) is verified if, setting $e_0 = 2/(2\pi)^{3/2}$,

$$n\varepsilon_n^2 \geqslant (2/\rho)\left[d \log\left(\frac{8d\sqrt{\rho n}}{\xi e_0 \tau}\right) + d^* \log(10a^*/\tau) + \log 2\right], \tag{20}$$

as long as $\tau \in [\xi/(8d\sqrt{\rho n}), a^*]$, see Lemma G.17 [12]. In particular, for fixed d, d^*, K, τ, and a^* as in (17), this condition is automatically satisfied for large enough n if (18) holds.

To obtain (20), one simply uses that the horseshoe density is bounded from below by a constant on the integration interval; this is sensible for a fixed τ but can be significantly improved for small τ, as seen in Corollary 3. Corollary 2 is a direct consequence of above examples and Corollary 1.

COROLLARY 2 (Fixed dimensions). *In the setting of Theorem 2, suppose the input dimension d is fixed (independent of n). Let π be either the exponential prior or the horseshoe prior with fixed (independent of n) respective parameters λ and τ. Then for large enough $M > 0$ (depending on β, d^* only), as $n \to \infty$,*

$$\sup_{f_0 \in \mathcal{F}(K)} E_{f_0} \Pi_\rho[f : \|f - f_0\|_{L^2(\mu)} \geqslant M\varepsilon_n \,|\, X, Y] \to 0,$$

where ε_n is given by $\varepsilon_n = (\log n)^{\frac{2\beta(1+d^*)}{2\beta+d^*}} n^{-\frac{2\beta}{2\beta+d^*}}$. In particular, the posterior distribution achieves the minimax convergence rate up to a logarithmic factor.

An important consequence of Corollary 2 is that it is possible to derive a (near-)optimal rate adapting to the unknown number d^* and coordinates of the active variables with continuous priors, that is, even without setting the scaling exactly to zero on certain coordinates (i.e. without performing a 'hard model selection'). Even more surprisingly at first, such 'soft model selection' is possible (at least with tempered posteriors) using a prior not putting a particularly large amount of mass near 0, such as an exponential prior. In particular, simple priors on scalings such as exponentials or gamma distributions considered in [48] have prior mass permitting for 'enough' variable selection in order for their (tempered)–posterior distribution to contract at near optimal rate, without using oracle knowledge of which coordinates are active or not.

At this point it may seem as if effective variable selection can be achieved at no cost with just simple random scalings on coordinates. This is not (completely) so, the reason being that the dependence on the input dimension d in the convergence rate that arises from e.g. putting exponential priors on scalings is far from optimal. This can be seen from (19), or similarly (20) for π_τ with fixed τ, as follows. Recall that (19)–(20) are non-asymptotic conditions, so one may let d, d^* depend on n. Suppose for instance $d = n^\delta$ for some $\delta > 0$ and d^* is fixed, say $d^* = 1$ to fix ideas. Then (19) cannot hold with ε_n the lower bound in (18): indeed, the latter is up to a log-factor of order $n^{1/(2\beta+1)}$, so is a $o(n^\delta)$ as soon as $\delta > 1/(2\beta + 1)$, which shows that in this setting the rate is suboptimal for large enough β's.

The previous comments naturally make one wonder if effective variable selection is still feasible with a better dependence on dimensions with continuous scaling. The next section investigates this, in a setting where d can go to ∞ with n.

High-dimensional variable selection. Let us now study the problem of inference for a small number of truly active covariates if d is possibly allowed to depend on n. In the high-dimensional sparsity adaptation problem as in the first setting of Section 2.1, the work [52] derives up to constants the minimax rate of estimation for the squared $\|\cdot\|_{L^2(\mu)}$ loss which is up to a constant factor *depending on* K, β and d^*,

$$(\epsilon_n^*)^2 \asymp n^{-\frac{2\beta}{2\beta+d^*}} + \frac{d^*}{n}\log(d/d^*).$$

The first term corresponds to the rate of estimation of a low-dimensional function $g \in \mathcal{C}^\beta[-1,1]^{d^*}$ and the second is the rate for the variable selection problem. Under Condition (21) below, the first term dominates. Note however that, as the dependence of the constants in d^* is not explicit in the last display, this result allows for $d \to \infty$ and a fixed d^* but not both d^*, d going to infinity.

Let us consider the sparse high-dimensional setting where d can go to infinity; we also allow the effective dimension d^* to slowly grow with n: more precisely for some $\delta < 1/2$ and $C_1, C_2 > 0$ suppose

$$1 \leqslant d^* \leqslant (\log n)^{1/2-\delta}, \qquad 1 \leqslant d^* \leqslant d \leqslant C_1 n^{C_2}. \tag{21}$$

One may hope to obtain a convergence rate that depends on the effective dimension d^* only, not on d. In Appendix E [12], Corollary E.1, we derive a lower bound result that shows that under mild conditions on the design distribution, and if the radius of the considered Hölder ball is not too large, the minimax rate for the integrated quadratic risk is bounded below by $C_2 D_2^{d^*} n^{-2\beta/(2\beta+d^*)}$, for constants C_2, D_2 independent of d^*. We show below that this rate can be achieved by a well-chosen horseshoe GP in the regime (21), up to a slowly-varying term $D_3^{d^*}(\log n)^c$. To do so, one first determines a horseshoe scaling parameter τ for which condition (15) holds, and then we state the sparse high-dimensional result as Corollary 3 (more details on optimality can be found below in Appendix E [12], Corollary E.1).

EXAMPLE 3 (Horseshoe prior with vanishing parameter τ). Consider $\pi = \pi_\tau$ a horseshoe prior of parameter $\tau > 0$. Then (15) is verified for large enough n if

$$10 a^* e^{-n\rho\varepsilon_n^2/2d^*} \leqslant \tau \leqslant \frac{\xi}{d^2}\frac{1}{\sqrt{\rho n}}. \tag{22}$$

For a^*, ε_n as in (17)–(18), the last display holds for large enough n and fixed K (or more generally $K \leqslant C^{d^*}$ for some $C > 1$) if one sets for some $\delta > 0$

$$\tau = \tau^* := (n^{1+\delta} d^4)^{-1/2}.$$

For a proof of both claims, see Lemma G.18 [12].

COROLLARY 3 (High-dimensional horseshoe GP). *In the setting of Theorem 2, suppose d^*, d verify Condition (21). Let Π be the multibandwidth prior (10) with horseshoe scaling density π_{τ^*} and $\tau^* = (n^{1+\delta}d^4)^{-1/2}$, $\delta > 0$. Then, for $K \geqslant 1$, there exists $M = M(\xi, \rho) > 0$ such that*

$$\sup_{f_0 \in \mathcal{F}(K)} E_{f_0} \Pi_\rho[f : \|f - f_0\|_{L^2(\mu)} \geqslant M\varepsilon_n \,|\, X, Y] \to 0,$$

as $n \to \infty$ where, for some constant C that depends on β, ξ, ρ only,

$$\varepsilon_n^2 = K^2 C^{d^*} (\log n)^{\frac{2\beta(1+d^*)}{2\beta+d^*}} n^{-\frac{2\beta}{2\beta+d^*}}.$$

In particular, for $K^2 \leqslant C^{d^*}$, the rate ε_n^2 is of order $n^{-2\beta/(2\beta+d^*)}$, up to a smaller order term at most of order $C^{2d^*}(\log n)^{\frac{2\beta(1+d^*)}{2\beta+d^*}}$.

The rate ε_n^2 obtained in Corollary 3 is optimal in the minimax sense up to the smaller order term $K^2 C^{d^*}$ (up to a log factor). As mentioned above, as long as K does not grow faster than $C_4^{d^*}$ this is a slower order term compared to the main term $n^{-2\beta/(2\beta+d^*)}$ (in regime (21)) in the minimax lower bound from Corollary E.1 [12]: in this case the rate is minimax optimal up to a slower order term $C_5^{d^*}$. We also note that a growth in C^{d^*} for the radius K of the Hölder ball is typical for functions in Hölder spaces of dimension d^*, see Appendix D [12], where this is checked for functions of product form.

The idea behind Corollary 3 is that for small values of the parameter τ, the horseshoe distribution becomes very 'sparse' in the sense that most nonzero values are very close to 0: this is reminiscent of the high-dimensional statistics literature for sparse models, see e.g. [45, 43], where near-optimal posterior rates for horseshoe posteriors are derived in sparse settings. We now turn to a deep learning setting, where the prior process is allowed to have several Gaussian compositional layers.

4. Main results II: deep simultaneous adaptation to structure and smoothness.

4.1. *Multilayer setting: Deep Horseshoe GP.* We now consider the problem of adaptation to an unknown compositional structure, first in the fixed dimensional case. The following result shows that, assuming the regression function can be expressed as a composition (5), such adaptation can be achieved with a prior mimicking this structure and organizing Gaussian processes in layers. In particular, in the Deep–HGP prior, the distribution on the scalings of the individual Gaussian processes allows for adaptation to the regularity as we have seen above, but also adaptation to a sparse network of compositions.

THEOREM 3. *Let $\lambda = (q, d_1, \dots, d_q, t_0, \dots, t_q)$, $\beta = (\beta_0, \dots, \beta_q)$, $d \geqslant 1$, $K \geqslant 1$ and suppose $f_0 \in \mathcal{F}_{deep}(\lambda, \beta, K)$. Let Π be the Deep–HGP prior (11) with fixed parameters $\tau_i > 0$. Then, for any $0 < \rho < 1$, $\Pi_\rho[\cdot \mid X, Y]$ contracts to f_0 at the rate*

$$\varepsilon_n = \max_{i=0,\dots,q} (\log n)^{\frac{\alpha_i \beta_i (1+t_i)}{2\alpha_i\beta_i + t_i}} n^{-\frac{\alpha_i\beta_i}{2\alpha_i\beta_i+t_i}}$$

in $\|\cdot\|_{L^2(\mu)}$ distance, where $\alpha_i = \prod_{l=i+1}^{q} (\beta_l \wedge 1)$: for any $M_n \to \infty$

$$E_{f_0} \Pi_\rho [f : \|f - f_0\|_{L^2(\mu)} \geqslant M_n \varepsilon_n \mid X, Y] \to 0.$$

The same conclusion holds for the prior Deep–HGP(q_{max}, d_{max}) as in (12), provided the parameters of λ verify $q \leqslant q_{max}$ and $d_i \leqslant d_{max}$ for all $i \leqslant d_{max}$.

Theorem 3 shows that the fractional posterior attains the minimax rate of convergence of contraction (8) over the class $\mathcal{F}_{\text{deep}}(\lambda, \beta, K)$ up to the logarithmic factor $\log^\gamma n$, $\gamma = \max_{i=0,\dots,q} 2\alpha_i\beta_i(1 + t_i)/(2\alpha_i\beta_i + t_i)$. For simplicity, we only considered the situation where inverse-bandwidths are distributed as horseshoe random variables. As above in the fixed d setting with GPs, one can derive similar results for exponential priors on scalings. However, given the benefits of the horseshoe prior in high-dimensional settings (see also below), we focus on this choice.

Theorem 3 can be compared to Theorem 2 of [19], providing rates for a deepGP construction over a compositional functional class. The present result on the Deep–HGP prior now shows that adaptation to the structure can be achieved with ρ–posteriors without imposing a hyperprior on all the parameters describing the structure. Even if a prior can still be put on the depth and width of the composition, as in the first part of the statement, it is enough to chose these deterministically under the mild condition above. As in our results of Section 3.2 on (single-layer) GPs, instead of imposing a 'hard' selection of relevant variables on each layer, a continuous distribution on the lengthscales, with sufficient mass on small values, is enough for *simultaneous* adaptation to smoothness and sparse compositional structure. The proof of Theorem 3 can be found in Section 6.3 for random q, d (see Appendix H [12] for the version for deterministic q, d).

REMARK 1 (Benign overfitting). The prior with deterministic q, d in the last part of the statement of Theorem 3 closely matches current practice in deep GPs implementation: recent works show that a depth of just a few layers already enables important gains compared to traditional (single-layer, or 'shallow') GPs, see e.g. [53] ($q = 2$), [55] ($q = 2$ and $q = 3$), [38] ($q = 2$). Interestingly, our result shows that 'overestimating' q and d (i.e. choosing q_{max}, d_{max} 'too large' in the above statement) does not prevent one to still obtain an *adaptive* rate (i.e. knowledge of 'true' q, d is not needed). In this sense we see that a form of *benign overfitting* is at play, with an overfitted architecture specified by (q_{max}, d_{max}) still leading to the optimal minimax rate with adaptation both to smoothness and structure.

In view of Corollary 3, one naturally wonders if the Deep–HGP prior is able to perform adaptation to the structure and the regularity in a high-dimensional framework as well. More precisely, suppose $f_0 : [-1,1]^d \to \mathbb{R}$ belongs to $\mathcal{F}_{\text{deep}}(\lambda, \beta, K)$ with $d = d_0$ and t_0 possibly depending on n. As in (21), we suppose t_0, d_0 verify, for some $\delta < 1/2$ and $C_1, C_2 > 0$,

$$1 \leqslant t_0 \leqslant (\log n)^{1/2-\delta}, \qquad 1 \leqslant t_0 \leqslant d_0 = d \leqslant C_1 n^{C_2}. \tag{23}$$

The next result shows that letting the horseshoe scale parameter τ_0 of the first layer in the composition vanish in the Deep–HGP prior, while keeping the other scale parameters $\tau_i, i = 1, \dots q$ across the other layers, constant, is enough to still obtain a near-minimax rate of contraction for the fractional posterior. Choosing τ_0 appropriately small (although independent of the true unknown t_0) allows one to obtain sparsity on the first layer, mitigating the effect of the growing input dimension d.

THEOREM 4. *Under the same assumptions as in Theorem 3, now suppose t_0 and d_0 satisfy Condition (23), and that $d_1, \dots, d_q$ are fixed, non n–dependent integers. Let Π be the Deep–HGP prior with $\tau_0 = d^{-2}n^{-1}$ and $\tau_i > 0, i = 1, \dots, q$ be fixed. For any $0 < \rho < 1$, there exists c_1, c_2 depending on β_0 and ρ such that the ρ–posterior contracts to f_0 at the rate*

$$\varepsilon_n^2 = \left[c_1 K^2 c_2^{t_0} (\log n)^{\frac{2\beta_0\alpha_0(1+t_0)}{2\beta_0\alpha_0+t_0}} n^{-\frac{2\beta_0\alpha_0}{2\beta_0\alpha_0+t_0}} \right] \vee \max_{i=1,\dots,q} (\log n)^{\frac{2\alpha_i\beta_i(1+t_i)}{2\alpha_i\beta_i+t_i}} n^{-\frac{2\alpha_i\beta_i}{2\alpha_i\beta_i+t_i}}$$

in $\|\cdot\|_{L^2(\mu)}$ distance, that is, for any $M_n \to \infty$,

$$E_{f_0}\Pi_\rho[f : \|f - f_0\|_{L^2(\mu)} \geqslant M_n\varepsilon_n \,|\, X, Y] \to 0.$$

To the best of our knowledge, Theorem 4 is the first result on deep methods in high-dimensional regression where both the input dimension $d_0 = d_0(n)$ and first effective dimension $t_0 = t_0(n)$ are allowed to grow with n. It combines both the ability of the horseshoe prior to select relevant dimensions in the input space and its ability to perform model selection in presence of a compositional parameter. It is particularly interesting given that these methods are most often applied to high-dimensional problems.

Compared to the condition on τ in Corollary 3, the preceding result requires a smaller τ_0. In the present setting, given the flexibility of the sample paths from deep Gaussian processes, it is necessary to 'stabilize' each GP to avoid 'wild' behavior. From a technical point of view, it translates into more restrictive prior mass conditions for these GPs and the need for more efficient variable selection. This is achieved with a horseshoe prior that is more peaked near 0, given the choice of the smaller parameter.

REMARK 2. In Theorem 4, we let the dimensions indexed by $i = 0$ on the first layer of the composition to possibly grow with n. Extending this result to a situation where $d_i, i > 0$ can also grow with n is not straightforward. Indeed, inspection of the proof of Theorems 3 and 4 shows that the rate involves a multiplicative factor $\sum_{i=1}^{q+1} d_i$ whose dependence on the inner dimensions of the composition is linear and thus prevents polynomially growing d_i's. As this factor does not involve d_0, it still allows for t_0, d_0 as in (23).

4.2. *Results for standard posteriors.* We now derive results for the posterior distribution (that is $\rho = 1$ in the fractional posterior). To do so one keeps the same prior on the regression function f as in the previous results, the only difference being that we now also put a prior on the noise variance σ^2. The following Proposition 1 is inspired by an idea of The Thien Mai [28], who in high-dimensional regression under sparsity, notices the link between standard posterior and fractional posterior for an updated prior in that sparse setting. Here, in the different context of nonparametric regression, we use a different n-dependent prior on σ^2.

The model now allows for possibly unknown noise variance: the observations (Y_i, X_i) are an independent sample

$$Y_i = f(X_i) + \sigma\eta_i, \tag{24}$$

with $X_i \sim \mu$ and $\eta_i \sim \mathcal{N}(0,1)$, where we treat both f and the variance parameter σ^2 as unknown. We denote by P_{f,σ^2} the distribution of (X_1, Y_1) from model (24) given f, σ^2.

Let Π be a prior on (f, σ^2) defined as a product $\Pi_f \otimes \pi_{\sigma^2}$. We take $\pi_{\sigma^2} = \pi_{\sigma^2,n}$ to be, for some fixed $b \in (0,1)$,

$$\pi_{\sigma^2} = \text{Gamma}\Big(\Big\{\frac{1-b}{2}\Big\}n + 1, b\Big), \tag{25}$$

where Gamma(a_1, a_2) denotes a Gamma distribution with shape parameter $a_1 > 0$ and rate parameter $a_2 > 0$, of density proportional to $x \to x^{a_1-1}e^{-a_2 x}$ on $(0, \infty)$. This prior induces a posterior distribution $\Pi[\cdot \mid (X,Y)]$ jointly on f and σ^2, and the next result examines the behaviour of its marginal on f, that is $\Pi[f \in \cdot\,, \sigma^2 \in \mathbb{R}^+ \mid (X,Y)]$.

PROPOSITION 1. *Let Π be a prior on (f, σ^2) of product form $\Pi_f \otimes \pi_{\sigma^2}$, with π_{σ^2} given by (25). Suppose, for some rate $\varepsilon_n = o(1)$ with $n\varepsilon_n^2 \to \infty$ and a constant $D > 0$, that*

$$\Pi_f[\|f - f_0\|_\infty \leqslant \varepsilon_n] \geqslant e^{-Dn\varepsilon_n^2}. \tag{26}$$

Then, for a constant $M > 0$ large enough, as $n \to \infty$, for D_b the b–Rényi divergence, the marginal posterior distribution on f given observations from (24) verifies

$$E_{f_0,\sigma_0^2}\Pi\left[D_b(p_{f,2\sigma_0^2}, p_{f_0,2\sigma_0^2}) \leqslant M\varepsilon_n \mid (X,Y)\right] \to 1.$$

If $\Pi[\|f\|_\infty \leqslant K \mid (X,Y)] = 1 + o(1)$ for some fixed constant $K > 0$, we also have

$$E_{f_0,\sigma_0^2}\Pi\left[\|f - f_0\|^2 \leqslant M\varepsilon_n \mid (X,Y)\right] \to 1.$$

This result shows that, modulo defining an appropriate prior on σ^2, results for fractional posteriors (that effectively only require the prior mass condition as in the statement), also hold for the (marginal in f of the) classical posterior; see Corollary 4 below. Proposition 1 is also of independent interest, as it holds for any choice of prior on Π_f in random design regression; its proof can be found in the Appendix [12], Section I. We also note that sampling from the marginal posterior on f from the above augmented prior does not add any fundamental difficulty, see Section 5.

COROLLARY 4. *Provided the respective priors Π_f are replaced with augmented priors $\Pi_f \otimes \pi_{\sigma^2}$, with π_{σ^2} the prior (25) on σ^2, the results stated in Theorems 1–4 still hold for the corresponding standard marginal posterior in f (instead of the fractional posterior corresponding to Π_f in these statements). Specifically, the nonparametric convergence rates derived in Theorems 1–4 remain valid under these modifications.*

To prove this result, note that Theorems 1–4 are established by proving the corresponding inequality (26), as outlined in the preliminaries of Section 6 (the slightly different constants can be accommodated by checking the condition (27) therein for $\varepsilon_n' = R\varepsilon_n$ and a large enough constant $R > 0$). Given that the prior distributions have samples almost surely bounded by 1, Corollary 4 follows from Proposition 1.

5. Discussion and open questions. In this work, we provide theoretical guarantees for the convergence of posterior distributions using deep Gaussian process priors. One key insight is in the role played by lengthscale parameters: not only do these enable adaptation to smoothness, but they can also at the same time perform an effective variable selection (adaptation to sparsity) by 'freezing' the Gaussian process paths in suitable directions, a point relevant also for standard (non-deep) Gaussian processes; this has not been recognised so far in the literature to the best of our knowledge. The fact that adaptation to smoothness and structure can be performed *simultaneously* is particularly appealing computationally in that there is no need to include a model selection part in the building of the prior (if that was the case, posterior sampling would require to have access to the posterior distribution over models, which is often costly to implement). The obtained deepGP prior is then simple enough so that it corresponds to recently proposed algorithms, see below for more on this.

We also derive new results on deep methods for high-dimensional input spaces, and on the way obtain explicit dimension-dependent constants for the characteristics of the involved GPs.

The use of fractional posteriors. Many of our results consider fractional posterior distributions, where the parameter ρ can be taken to be any constant in $(0,1)$. The main reason for this is of technical nature: in order for one to use the general theory of convergence of Bayesian posteriors in [20], one needs to build sieve sets, capturing most prior mass, whose entropy or 'complexity' is well controlled. However, especially in complex settings such as deep learning models, sieves can be difficult to construct, in particular since the probability of the complement of the sieve is required to have a form of exponentially fast decrease, with at the same time the requirement to control the entropy of the sieve set. This difficulty leads [19] to condition sample paths of Gaussian processes to verify certain smoothness constraints. This can be avoided using ρ–posteriors, since convergence of these can be guaranteed under prior mass conditions only [6, 26, 27, 54], so we do not need to condition on boundedness of derivatives in our prior construction. This is an advantage also computationally, as adding more conditioning constraints may typically slow down MCMC samplers.

Beyond fractional posteriors, we have also obtained results for the standard posterior ($\rho = 1$) and avoid the just-mentioned technical difficulties: the idea then has been to use an augmented prior that also models the noise variance. For the original prior (without a prior on σ^2) one can conjecture that the standard posterior also achieves optimal rates; although it seems delicate to prove this using the current tools available for proving posterior convergence, given that construction of sieves (while keeping the prior simple) looks particularly difficult, it is conceivable that, at least in say regression settings, one may be able to state an adapted version of the generic theorem of [20]. Although beyond the scope of the present contribution, one may note that promising results are obtained in [2], where in regression with heavy-tailed priors both standard and fractional posteriors are shown to converge at the same rate up to constants, for a prior for which the construction of sieves seems also presently out of reach, which suggests that posterior convergence for $\rho = 1$ under prior mass conditions only is not exceptional.

On the other hand, we argue that, at least for the set of applications of Bayesian (possibly tempered) posteriors considered here, one does not loose much with fractional posteriors, except slightly in the constants in the convergence rate: here as ρ is fixed (and can be chosen e.g. to be $\rho = 1/2$) we did not keep the dependence in ρ in the constants, but it is shown in [27] that nonparametric rates arising from ρ–fractional posteriors are typically the same as for the usual posterior but with effective sample size $n' = n\rho$; for $\rho = 1/2$ the constant in terms of ρ in the rate is not particularly large then, so is not a main concern. Also, regarding sampling algorithms in practice, most sampling methods such as MCMC are of similar difficulty with the fractional or the original likelihood, so this is not a main concern computationally (we discuss below sampling from deepGP fractional posteriors from Example 1). One loses, though, the interpretation of the posterior as a conditional distribution and possibly efficiency for $\sqrt{n}$–estimable parameters that comes with the Bernstein–von Mises theorem, which will not hold as such for ρ–posteriors but again typically with a variance inflated by $1/\rho$, see [27] for more discussion. However, again, this is not a main concern here, as we are interested in estimation rates up to constants. Another interesting topic not covered in the present work is uncertainty quantification. It should be possible to prove that, modulo certain (unavoidable) structural conditions on the regression function such as self-similarity, certain credible sets from the deepGP (possibly, fractional) posterior are also confidence sets. In case one uses a fractional posterior, this may slightly increase the radius compared to the classical posterior, although, we expect, not in a significant way if ρ is kept far away from 0; we refer to [2] for an empirical study of the influence of ρ on nonparametric credible sets.

Implementation. Although our main focus here is on theoretical guarantees, we note that sampling from the deep Gaussian process fractional posteriors with exponential priors on GP lengthscales (Example 1) is readily available using the R package `deepgp` [37]. The later provides MCMC samples for standard posteriors ($\rho = 1$) using Vecchia approximation; one can similarly obtain MCMC samples from fractional posteriors with any $\rho < 1$ using the following remark. Note that using a fractional likelihood with a given $\rho \in (0, 1)$ to form the fractional posterior in Gaussian regression with independent errors $\mathcal{N}(0, \sigma^2)$ is equivalent to using the standard likelihood in the case the errors are *misspecified as* independent $\mathcal{N}(0, \sigma^2/\rho)$. Since posterior sampling is conditional on the observed values, and the `deepgp` package allows for specifying a given noise level, it suffices to specify it to the misspecified value σ^2/ρ (while data will truly be generated with errors of variance σ^2). We also note that for the implementation of the augmented prior $\Pi_f \otimes \pi_{\sigma^2}$ discussed in Section 4.2, it is enough to sample from the marginal posterior in f, and `deepgp` allows for a Gamma prior on σ^2. We refer to [38] for illustrations and details on the sampling schemes (we note that it should also be possible to modify the code, which presently allows for Gamma priors, to include horseshoe priors on lengthscales as in Examples 2–3).

Considering fixed deterministic composition depth q and width d (in the spirit of Remark 1) and given lengthscale parameters, the prior considered in the present work (but without Ψ) coincides with that considered in the paper [16] introducing deepGPs, where the kernel is termed ARD (Automatic Relevance Determination) and the lengthscales are called weights. In [16], the weights/lengthscales are then calibrated using a variational approach. Many different posterior approximating schemes for deepGPs have been proposed over the last few years, using in particular variational approximations; we refer to [38] for an overview and discussion. Obtaining theoretical guarantees for these different approaches, in particular for the Deep HGP posteriors introduced here, is an interesting avenue for future work.

6. Proof of the main results.

We denote by ν the spectral measure of the squared-exponential SqExp process. Let us recall that this process is stationary with covariance $K(s,t) = \exp(-\|s-t\|^2) = k(s-t)$ and that by Bochner's theorem $k(t) = \int e^{-t\langle u,t\rangle}\nu(du)$; in particular it follows that ν has Lebesgue density $u \to \exp(-\|u\|^2/4)/(2^d\pi^{d/2})$.

Preliminaries: reducing the problem to a prior mass condition. Given a rate (ε_n), Theorem F.1 and Proposition F.1 in Appendix F [12] ensure that if Π satisfies, for $f_0 \in C[-1,1]^d$ and $\rho \in (0,1)$,

$$\Pi\left[\|f - f_0\|_\infty \leqslant \frac{2\sigma_0^2}{\sqrt{1+4\sigma_0^2}}\varepsilon_n\right] \geqslant e^{-n\rho\varepsilon_n^2}, \tag{27}$$

then the fractional posterior is such that, for D_ρ the ρ–Rényi divergence as in (22),

$$E_0\Pi_\rho\left(\eta:\ \frac{1}{n}D_\rho(p^n_\eta, p^n_{\eta_0}) \geqslant 4\frac{\rho\varepsilon_n^2}{1-\rho}\,|\,X\right) \leqslant e^{-n\rho\varepsilon_n^2} + (n\varepsilon_n^2)^{-1}.$$

Let us now focus on the considered regression model. Assuming that the regression functions we consider are bounded, say $\|f\|_\infty, \|g\|_\infty \leqslant 1$, using $1 - x \leqslant -\log x$ and $1 - e^{-x} \geqslant xe^{-x}$ both for $x \geqslant 0$ and the additivity of the Rényi divergence for densities of independent observations, it follows that

$$\frac{1}{n}D_\rho(p_f^{\otimes n}, p_g^{\otimes n}) = -\frac{1}{1-\rho}\log\int e^{-\frac{\rho-\rho^2}{2\sigma_0^2}(f-g)^2}d\mu \geqslant \frac{1}{1-\rho}\left[1 - \int e^{-\frac{\rho-\rho^2}{2\sigma_0^2}(f-g)^2}d\mu\right]$$

$$\geqslant \frac{\rho}{2\sigma_0^2}e^{-2\frac{\rho-\rho^2}{\sigma_0^2}}\|f-g\|^2_{L_2(\mu)}\,. \tag{28}$$

We note that under the regularity assumptions on f_0 and with the use of the ‘link’ function Ψ, the boundedness assumption is satisfied for both f_0 and a draw f from the posterior in our different theorems. Consequently, in the following, the proofs consist in proving (27) for f_0 as in the different statements and for the different priors considered. This will imply

$$E_0\Pi_\rho\left(f:\ \|f-g\|^2_{L_2(\mu)} \geqslant \frac{8\sigma_0^2}{1-\rho}e^{2\frac{\rho-\rho^2}{\sigma_0^2}}\varepsilon_n^2\,|\,X\right) \leqslant e^{-n\rho\varepsilon_n^2} + (n\varepsilon_n^2)^{-1},$$

which suffices to conclude.

6.1. *Proof of Theorem 1.*

The proof of the theorem is a (simpler) variation on the proof of Theorem 2 to follow: therein, it suffices to work with the fixed values of scaling parameters specified in (14). For easy reference the precise argument is given below the end of the proof of Theorem 2.

6.2. *Proof of Theorem 2.*

Given $A \in \mathbb{R}^d_+$, let us denote by $\varphi^A_{f_0}$ the concentration function of the Gaussian process W^A as in (10) and its RKHS by $\mathbb{H}^A$. To derive the result, we prove below that (27) is satisfied, since any $f_0 \in \mathcal{F}(K)$ satisfies $\|f_0\|_\infty \leqslant 1$ and $\left\|\Psi(W^A)\right\|_\infty \leqslant 1$ by definition. Also, since $\left\|f_0 - \Psi(W^A)\right\|_\infty \leqslant \left\|f_0 - W^A\right\|_\infty$, we bound this last quantity instead as it then implies (27) for $f = \Psi(W^A)$.

Since we assume $f_0 \in \mathcal{F}_{VS}(K, \beta, d, d^*)$, we have $f_0(x_1, \dots, x_d) = g_0(x_{i_1}, \dots, x_{i_{d^*}})$ for some $g_0 \in \mathcal{F}(K, \beta, d^*)$. Let us set, for $S_0 = \{1, \dots, d\}\backslash\{i_1, \dots, i_{d^*}\}$ and $\xi = 2\sigma_0^2/\sqrt{1+4\sigma_0^2}$,

$$I_i = \begin{cases} [0, \xi/(8d\sqrt{\rho n})], & \text{if } i \in S_0,\\ [a^*, 2a^*], & \text{otherwise,}\end{cases}$$

for $a^* \in [1, n]$ as in the statement of the Theorem. For a given vector A, let us introduce $\tilde{A} = (\tilde{A}_1, \dots, \tilde{A}_d)$ with $\tilde{A}_i = A_i\mathbb{1}_{i\in\{i_1,\dots,i_{d^*}\}}$ for $i = 1, \dots, d$. For any $\varepsilon > 0$,

$$\Pi\left[f:\ \|f-f_0\|_\infty < \xi\varepsilon\right] = P\left[\left\|W^A - f_0\right\|_\infty < \xi\varepsilon\right]$$

$$\geqslant \int_{I_1}\cdots\int_{I_d} P\left[\left\|W^A - f_0\right\|_\infty < \xi\varepsilon\,|\,A\right]\prod_{i=1}^d \pi(A_i)dA_i.$$

One may now split the contribution of A_i's into subsets of indices as follows

$$\begin{aligned} P\left[\left\|W^A - f_0\right\|_\infty < \xi\varepsilon\,|\,A\right] &\geqslant P\left[\left\|W^{\tilde{A}} - f_0\right\|_\infty < \xi\varepsilon/2\,,\ \left\|W^{\tilde{A}} - W^A\right\|_\infty < \xi\varepsilon/2\,|\,A\right]\\ &\geqslant P\left[\left\|W^{\tilde{A}} - f_0\right\|_\infty < \xi\varepsilon/2\,|\,A\right] - P\left[\left\|W^{\tilde{A}} - W^A\right\|_\infty > \xi\varepsilon/2\,|\,A\right].\end{aligned}$$

In what follows we bound from below and above respectively the quantities, for $\eta = \xi\varepsilon/2$,

$$P_1(\eta) = \int_{I_1} \cdots \int_{I_d} P\left[\left\|W^{\tilde{A}} - f_0\right\|_\infty < \eta \,|\, A\right] \prod_{i=1}^{d} \pi(A_i) dA_i, \tag{29}$$

$$P_2(\eta) = \int_{I_1} \cdots \int_{I_d} P\left[\left\|W^{\tilde{A}} - W^{A}\right\|_\infty > \eta \,|\, A\right] \prod_{i=1}^{d} \pi(A_i) dA_i. \tag{30}$$

In the bounds P_1, P_2 to follow, we use that the involved scaling parameters A_i belong to the respective intervals I_i defined above.

Starting with (29), denote $A^* = \left(A_{i_1}, \ldots, A_{i_{d*}}\right) \in \mathbb{R}_+^{d*}$. Conditionally on the A_i's, the Gaussian process $W^{\tilde{A}}$, interpreted as a process on variables indexed by $i_1, \ldots, i_{d*}$ only, has the same distribution as the Gaussian process W^{A^*} on $[-1,1]^{d*}$ (they are both centered with same covariance kernel; in slight abuse of notation we denote also W for the process in the d^*–dimensional space). Since W^{A^*} is independent of A_i for $i \in S_0$, for $\eta > 0$,

$$P\left[\left\|W^{\tilde{A}} - f_0\right\|_\infty < \eta \,|\, A\right] = P\left[\left\|W^{A^*} - g_0\right\|_\infty < \eta \,|\, A^*\right].$$

The term P_1 can then be bounded from below by

$$\begin{aligned} P_1(\eta) &= \prod_{i \in S_0} \pi(I_i) \cdot \int_{I_{i_1}} \cdots \int_{I_{i_{d*}}} P\left[\left\|W^{A^*} - g_0\right\|_\infty < \eta \,|\, A^*\right] \prod_{j=1}^{d*} \pi(A_{i_j}) dA_{i_j} \\ &\geqslant \prod_{i \in S_0} \pi(I_i) \cdot \int_{I_{i_1}} \cdots \int_{I_{i_{d*}}} e^{-\varphi_{g_0}^{A^*}(\eta/2)} \prod_{j=1}^{d*} \pi(A_{i_j}) dA_{i_j}, \end{aligned}$$

where we use Lemma A.1 to bound from below the probability in the display.

We now use Theorem 5 applied to g_0, a function with input dimension d^*. We set $\varepsilon = \varepsilon_n$ to be chosen so that $\eta = \xi\varepsilon_n/2$. Suppose,

$$8 \geqslant \quad \xi\varepsilon_n \quad \geqslant 4\mathcal{C}_1 K (a^*)^{-\beta}, \tag{31}$$

$$\rho n \varepsilon_n^2/2 \geqslant \mathcal{C}_2 K^2 (a^*)^{d^*} + (C {d^*}^c a^*)^{d^*} \log^{1+d^*}(8a^*/\varepsilon_n), \tag{32}$$

where $\mathcal{C}_i = \mathcal{C}_i(\beta, d^*), i = 1, 2$ are constants of the form $c_i C_i^{d^*}$ as in the statement of Theorem 5. Up to making $\mathcal{C}_1, \mathcal{C}_2$ larger, one can always assume $\mathcal{C}_1, \mathcal{C}_2 \geqslant 1$. Below we will also use that if $\varepsilon_n \leqslant 1, a^* \geqslant 1$ verifying the last display exist, then $a^* \leqslant n$. Indeed, the last term in the second inequality is bounded from below by $(C {d^*}^c a^*)^{d^*} \log^{1+d^*}(8a^*)$. As $8a^* \geqslant e$ one must have $(C {d^*}^c a^*)^{d^*} \leqslant \rho n \varepsilon_n^2/2$ where $C \geqslant 1$, so that $a^* \leqslant (n\varepsilon_n^2)^{1/d^*} \leqslant n$.

By Theorem 5 we have $\varphi_{g_0}^{A^*}(\varepsilon_n) \leqslant \rho n \varepsilon_n^2/2$, uniformly for $a^* \leqslant A_{i_j} \leqslant 2a^*$. One deduces

$$P_1(\xi\varepsilon_n/2) \geqslant \prod_{i \in S_0} \pi(I_i) \cdot e^{-\rho n \varepsilon_n^2/2} \cdot \prod_{i \notin S_0} \pi(I_i).$$

Let us now deal with the term P_2 in (30). First one notes that, given A,

$$X := W^{\tilde{A}} - W^{A}$$

is a centered Gaussian process. In order to bound it, one first computes

$$\begin{aligned} \sigma^2 &= \sup_{t \in [-1,1]^d} \mathbb{E}[X^2(t) \,|\, A] = \sup_{t \in [-1,1]^d} 2\left(1 - e^{-\sum_{i \in S_0} A_i^2 t_i^2}\right) \\ &= 2\left(1 - e^{-\sum_{i \in S_0} A_i^2}\right) \leqslant 2\,(d - d^*) \max_{i \in S_0} A_i^2, \end{aligned}$$

using $e^{-u} \geqslant 1-u$ for any u and $|S_0| = d - d^*$. Setting $M \coloneqq e^{-\sum_{i \in S_0} A_i^2 s_i^2}$, $N \coloneqq e^{-\sum_{i \in S_0} A_i^2 t_i^2}$, $P \coloneqq e^{-\sum_{i \notin S_0} A_i^2 (s_i - t_i)^2}$, $Q \coloneqq e^{-\sum_{i \in S_0} A_i^2 (s_i - t_i)^2}$,

$$\begin{aligned} D\left(s,t\right)^2 &\coloneqq \mathbb{E}[|X(s) - X(t)|^2 \mid A] \\ &= \mathbb{E}\left[\left(W^{\tilde{A}}(t) - W^{A}(t) - W^{\tilde{A}}(s) + W^{A}(s)\right)^2 \mid A\right] \\ &= 4 + 2PM - 2PQ - 2M - 2N + 2PN - 2P \\ &= 2\left(1-Q\right) + 2\left(1-P\right)\left(1 + Q - M - N\right). \end{aligned}$$

For $s, t \in [-1,1]^d$, using again $e^{-u} \geqslant 1-u$,

$$1 - Q \leqslant |s-t|_\infty^2 \ (d-d^*) \ \max_{i \in S_0} A_i^2, \quad 1 - P \leqslant |s-t|_\infty^2 \ d^* \max_{i \notin S_0} A_i^2.$$

For any $x, y, z \geqslant 0$, we have $1 + e^{-z} - e^{-x} - e^{-y} \leqslant 2 - e^{-x} - e^{-y} \leqslant x + y$, using the inequalities $e^{-z} \leqslant 1$ and $e^{-x} \geqslant 1 - x$. Deduce

$$1 + Q - M - N \leqslant \sum_{i \in S_0} (s_i^2 + t_i^2) A_i^2 \leqslant 2\left(d - d^*\right) \max_{i \in S_0} A_i^2.$$

Combining the previous bounds one obtains, for any $s, t \in [-1,1]^d$,

$$D(s,t)^2 \leqslant 2\,|s-t|_\infty^2 \left(d-d^*\right) \max_{i \in S_0} A_i^2 \left[1 + 2d^* \max_{i \notin S_0} A_i^2\right].$$

On the other hand, using $D(s,t)^2 \leqslant 2\mathbb{E}[X(s)^2 + X(t)^2 \mid A] \leqslant 4\sigma^2$ for any s, t, we have

$$\sup_{s,t \in [0,1]^d} D(s,t) \leqslant 2\sigma.$$

According to Lemma C.9 [12], since $X(0) = 0$ almost surely, we have

$$\mathbb{E}\left[\|X\|_\infty \mid A\right] \leqslant 4\sqrt{2} \int_0^\sigma \sqrt{\log\ 2N\left(\epsilon, [-1,1]^d, D\right)} d\epsilon.$$

Writing $Z^2 = 2\left(d - d^*\right) \max_{i \in S_0} A_i^2 \left[1 + 2d^* \max_{i \notin S_0} A_i^2\right]$, this quantity is upper bounded by

$$4\sqrt{2} \int_0^\sigma \sqrt{\log\ 2N\left(\epsilon/Z, [-1,1]^d, |\cdot|_\infty\right)} d\epsilon \leqslant 4\sqrt{2d} \int_0^\sigma \sqrt{\log\ \frac{4Z}{\epsilon}} d\epsilon,$$

using $2N(\eta, [-1,1]^d, |\cdot|_\infty) \leqslant (4/\eta)^d$ for $\eta \leqslant 1$, which is the case here since $\sigma/Z \leqslant 1$ (see also below). By a change of variable $v = \sqrt{\log(4Z/\epsilon)}$, the upper bound is, with $C' = 32\sqrt{2}$,

$$4\sqrt{2d} \int_{\sqrt{\log\ \frac{4Z}{\sigma}}}^\infty 8Zv^2 e^{-v^2} dv = C'\sqrt{d}Z \int_{\sqrt{\log\ \frac{4Z}{\sigma}}}^\infty v^2 e^{-v^2} dv.$$

Integrating by parts, $\int_a^\infty v^2 e^{-v^2} dv = ae^{-a^2}/2 + \int_a^\infty e^{-v^2} dv/2$. For $a \geqslant 1$ we have $\int_a^\infty e^{-v^2} dv \leqslant \int_a^\infty v^2 e^{-v^2} dv$ so that $\int_a^\infty v^2 e^{-v^2} dv \leqslant ae^{-a^2}$. Let us apply this to the previous bound, noting that $\log 4Z/\sigma \geqslant \log 4 \geqslant 1$, using the upper bound on σ obtained above and the definition of Z. One obtains, using that $\sigma \to \sigma\sqrt{\log(4Z/\sigma)}$ is increasing on $[0, 4Z/\sqrt{e}]$, and that $\sigma^2 \leqslant \bar{\sigma}^2 := 2(d-d^*) \max_{i \in S_0} A_i^2 \leqslant Z^2 \leqslant (4Z/e^{1/2})^2$,

$$\mathbb{E}\left[\|X\|_\infty \mid A\right] \leqslant 8\sqrt{2d}\sigma\sqrt{\log 4Z/\sigma} \leqslant 16d \max_{i \in S_0} A_i \sqrt{\frac{1}{2} \log\left(16[1 + 2d^* \max_{i \notin S_0} A_i^2]\right)}.$$

Assuming $\max_{i\notin S_0} A_i^2 \geqslant 1$, we obtain for some universal $c_1 > 0$,

$$\mathbb{E}\left[\|X\|_\infty \mid A\right] \leqslant c_1 d \cdot \max_{i\in S_0} A_i \cdot \sqrt{\log\left(1 + 2d^* \max_{i\notin S_0} A_i^2\right)}.$$

One can now use Gaussian concentration to control the deviations of $\|X\|_\infty$ from its expectation. Suppose

$$\varepsilon_n \geqslant 4\xi^{-1} c_1 d \cdot \max_{i\in S_0} A_i \cdot \sqrt{\log\left(1 + 2d^* \max_{i\notin S_0} A_i^2\right)}. \tag{33}$$

Combining Lemma C.8 [12] and the above bound on $\mathbb{E}\left[\|X\|_\infty \mid A\right]$ gives

$$P\left[\|X\|_\infty > \xi\varepsilon_n/2 \,|\, A\right] \leqslant P\left[\|X\|_\infty - \mathbb{E}\,\|X\|_\infty > \xi\varepsilon_n/4 \,|\, A\right] \leqslant e^{-\xi^2\varepsilon_n^2/32\sigma^2}. \tag{34}$$

Recall that, for $A_i \in I_i$,

$$\max_{i\in S_0} A_i \leqslant \xi/(8d\sqrt{\rho n}),$$

which gives $\sigma^2 \leqslant \xi^2/(32\rho n)$, so that the last but one display is bounded from above by $e^{-\rho n\varepsilon_n^2}$, uniformly for A_i in the corresponding interval I_i. One deduces

$$P_2(\xi\varepsilon_n/2) \leqslant e^{-\rho n\varepsilon_n^2} \prod_{i=1}^d \pi(I_i).$$

Combining this with the obtained lower-bound on $P_1(\xi\varepsilon_n/2)$, one gets, using $e^{-\rho n\varepsilon_n^2/2} \geqslant 2e^{-\rho n\varepsilon_n^2}$ if $n\varepsilon_n^2 \geqslant 1/4\rho$, that $P_1(\xi\varepsilon_n/2) - P_2(\xi\varepsilon_n/2) \geqslant e^{-\rho n\varepsilon_n^2/2}\prod \pi(I_i)/2$, so that

$$\Pi\left[f:\ \|f - f_0\|_\infty < \xi\varepsilon_n\right] \geqslant e^{-\rho n\varepsilon_n^2/2} \prod_{i=1}^d \pi(I_i)/2 \geqslant e^{-\rho n\varepsilon_n^2}, \tag{35}$$

where we used (15).

Let us now optimise in terms of ε_n verifying the conditions (31)–(32)–(33). Since

$$\max_{i\in S_0} A_i \leqslant \xi/(8d\sqrt{\rho n}), \qquad \max_{i\notin S_0} A_i \leqslant n$$

for $A_i \in I_i$, we have that (33) holds if, for some $c_2 > 0$ depending on ρ,

$$\varepsilon_n^2 \geqslant c_2 \frac{\log(1 + 2d^* n^2)}{n}. \tag{36}$$

Now turning to (31) – (32), recalling $\mathcal{C}_i = \mathcal{C}_i(\beta, d^*) \geqslant 1$ and $K \geqslant 1$, it suffices to have, using $\varepsilon_n \geqslant 1/\sqrt{n}$ and $a^* \leqslant n$ as noted earlier,

$$\varepsilon_n^2 \geqslant \{B_1 a^{*-2\beta}\} \vee \{B_2 a^{*d^*} \log^{1+d^*}(n)/n\}, \tag{37}$$

where $B_1 = C(\xi^{-1}\mathcal{C}_1 K)^2$ and $B_2 = C\rho^{-1}K^2\mathcal{C}_2(c_1 d^{*c_2})^{d^*}$, with c_1, c_2, C universal constants. We note that (37) implies (36) for C large enough, using $a^* \geqslant 1$ and $n \geqslant 3$ (which implies $\log(d^*) \lesssim \log^{1+d^*}(n)$). This concludes the proof of Theorem 2, provided $\varepsilon_n \leqslant 8\xi^{-1}$.

Proof of Theorem 1. One follows the proof of Theorem 2, noting that the fixed values of scaling parameters specified in (14) belong to the intervals I_i from the proof of Theorem 2, and where now there is no conditioning on A_i (those have given values now).

Let us set $a^* = n^{1/(2\beta+s)}$, $\varepsilon_n^2 = M\log^{1+s}(n)n^{-2\beta/(2\beta+s)}$ and $d^* = s$. Then the conditions (31), (32), (33) and (37) arising on ε_n in the proof of Theorem 2 are satisfied for M large enough depending on s, K, β, ξ and ρ. This gives

$$\Pi\left[f:\ \|f - f_0\|_\infty < \xi\varepsilon_n\right] \geqslant e^{-\rho n\varepsilon_n^2}.$$

One concludes similarly as for the proof of Theorem 2, using the discussion following (27).

6.3. *Proof of Theorem 3*. The proof of Theorem 2 needs to be suitably generalized and modified: as we shall see below, the considered L^∞ balls for the various layers of the composition need to have carefully chosen radii. To obtain the results, one needs to verify the prior mass condition (27) for ε_n as in the statement of the theorem. For any $f_0 \in \mathcal{F}_{\text{deep}}(\lambda, \beta, K)$ where $\lambda = (q, d_1, \dots, d_q, t_0, \dots, t_q)$, we now have

$$\begin{aligned}\Pi\left[f : \|f - f_0\|_\infty < \xi\varepsilon_n\right] \geqslant & \Pi_q[\{q\}]\Pi[\{d_1, \dots, d_q\}|\, q]\\ & \qquad \Pi\left[\|g_q \circ \dots \circ g_0 - h_q \circ \dots \circ h_0\|_\infty < \xi\varepsilon_n \mid q, d_1, \dots, d_q\right].\end{aligned}$$

Let us now fix $0 \leqslant i \leqslant q$ and $1 \leqslant j \leqslant d_i$, such that we focus on the marginal distribution of g_{ij}. Lemma C.7 [12] indeed ensures that for any $\varepsilon_n(i) > 0$, denoting $\mathbf{d} = (d_1, \dots, d_q)$,

$$\begin{aligned}&\Pi\left[\|g_q \circ \dots \circ g_0 - h_q \circ \dots \circ h_0\|_\infty < \xi q \prod_{i=0}^{q}[2^{|\beta_i - 1|} t_i K \vee 1] \max_{i=0,\dots,q} \varepsilon_n(i)^{\alpha_i} \mid q, \mathbf{d}\right]\\ &\qquad \geqslant \prod_{i=0}^{q}\prod_{j=1}^{d_{i+1}} \Pi\left[\left\|W^{A_{ij}} - h_{ij}\right\|_\infty \leqslant \xi^{1/\alpha_i}\varepsilon_n(i) \mid q, \mathbf{d}\right], \qquad (38)\end{aligned}$$

where $W^{A_{ij}}$ is as in (10) (with random bandwidths A_{ij}) and $\alpha_i = \prod_{l=i+1}^{q}(\beta_l \wedge 1)$. Since the h_{ij} are bounded by 1 in supnorm, the factors in the above product are lower bounded by

$$\Pi\left[\|g_{ij} - h_{ij}\|_\infty \leqslant \xi^{1/\alpha_i}\varepsilon_n(i) \mid q, d_1, \dots, d_q\right].$$

If we can find $\varepsilon_n(i)$ such that the above quantity is lower bounded by $e^{-\rho n \varepsilon_n(i)^{2\alpha_i}}$, then we can verify that $\varepsilon_n = C(K, \lambda)\max_{i=0,\dots,q}\varepsilon_n(i)^{\alpha_i}$ such that $n\varepsilon_n^2 \to \infty$, for

$$C(K, \lambda) = \left[q\prod_{i=0}^{q}[2^{|\beta_i - 1|}t_i K \vee 1]\right] \vee \sum_{i=0}^{q} d_{i+1},$$

is a posterior contraction rate thanks to (27). Having $n\varepsilon_n^2 \to \infty$ ensures that the remaining mass in Theorem F.1 in Appendix F [12] is vanishing, so that ε_n is indeed a contraction rate. Indeed, up to the constant factor

$$L = \Pi_q[\{q\}]\Pi[\{d_1, \dots, d_q\}|\, q]$$

independent of n, we can derive (27) from the lower bound on the right-hand side of (38)

$$Le^{-\rho n \varepsilon_n^2}.$$

Indeed, as long as $n\varepsilon_n^2 \to \infty$, we could replace ε_n with $C\varepsilon_n$, $C \geqslant 1$, for C such that $Le^{-\rho n \varepsilon_n^2} \geqslant e^{-\rho n C \varepsilon_n^2}$. Since the left side of (27) increases when replacing ε_n with $C\varepsilon_n$, (27) would be satisfied with $C\varepsilon_n$. This is enough as we seek to express ε_n up to a large enough constant.

From here, we can continue as in the proof of Theorem 2. Since we assume $h_{ij} \in \mathcal{F}_{VS}(K, \beta_i, d_i, t_i)$, we have $h_{ij}(x_1, \dots, x_d) = f_{ij}(x_{k_1}, \dots, x_{k_{t_i}})$ for some $f_{ij} \in \mathcal{F}(K, \beta_i, t_i)$. Let us set, for $S_0 = \{1, \dots, d_i\}\backslash\{k_1, \dots, k_{t_i}\}$, $\xi = 2\sigma_0^2/\sqrt{1 + 4\sigma_0^2}$, and

$$v_{i,n} := \frac{\xi\varepsilon_n(i)^{1-\alpha_i}}{8\sqrt{\rho n d_i}},$$

the intervals

$$I_k = \begin{cases}[0, v_{i,n}], & \text{if } k \in S_0,\\ [a^*, 2a^*], & \text{otherwise,}\end{cases}$$

for $a^* \in [1, n/2]$. Let's also consider an arbitrary vector A_{ij} such that $(A_{ij})_k \in I_k$ for $1 \leqslant k \leqslant d$ (in the following, we note $\mathcal{A}_k = (A_{ij})_k$ for simplicity). If we can show $\prod_{k=1}^{d_i} \pi(I_k) \geqslant 2e^{-\rho n \varepsilon_n(i)^{2\alpha_i}/2}$, and, $\mathcal{C}_{1,i}, \mathcal{C}_{2,i}$ constants of the form $c_{j,i} C_{j,i}^{t_i}$, $j = 1, 2$, as in the statement of Theorem 5,

$$4 \geqslant \xi^{1/\alpha_i} \varepsilon_n(i)/2 \geqslant 4\mathcal{C}_{1,i} K (a^*)^{-\beta_i}, \tag{39}$$

$$\rho n \varepsilon_n(i)^{2\alpha_i}/2 \geqslant \mathcal{C}_{2,i} K^2 (a^*)^{t_i} + (C t_i^c a^*)^{t_i} \log^{1+t_i}(8a^*/\xi^{1/\alpha_i}\varepsilon_n(i)), \tag{40}$$

(counterparts of (31) and (32)) and, for some $c_1 > 0$,

$$\varepsilon_n(i) \geqslant 4\xi^{-1/\alpha_i} c_1 d_i \cdot \max_{k \in S_0} \mathcal{A}_k \cdot \sqrt{\log\left(1 + 2t_i \max_{k \notin S_0} \mathcal{A}_k^2\right)}, \tag{41}$$

$$\max_{k \in S_0} \mathcal{A}_k \leqslant \frac{\xi \varepsilon_n(i)^{1-\alpha_i}}{8\sqrt{\rho n d_i}} \tag{42}$$

(the first is a counterpart of (33), the second ensures that an upper bound $e^{-\xi^2 \varepsilon_n(i)^2/32\sigma_i^2}$ obtained as in (34) is further upper bounded by $\exp(-\rho n \varepsilon_n(i)^{2\alpha_i})$, as $\sigma_i^2 \leqslant 2d_i \max_k \mathcal{A}_k^2$). Under these conditions, we can conclude

$$\Pi\left[\left\|W^{A_{ij}} - h_{ij}\right\|_\infty \leqslant \xi^{1/\alpha_i}\varepsilon_n(i) \mid q, d_1, \ldots, d_q\right] \geqslant e^{-\rho n \varepsilon_n(i)^{2\alpha_i}}$$

in the same way we obtained (35).

From (42), which is satisfied by definition of $v_{i,n}$, and $\max_{k \notin S_0} A_k \leqslant n$, (41) is satisfied whenever

$$\varepsilon_n(i)^{4\alpha_i - 2} \geqslant c_2 \frac{\log(1 + 2t_i n^2)}{n} \tag{43}$$

for some $c_2 > 0$. Now turning to (39) – (40), recalling $\mathcal{C}_{j,i} = \mathcal{C}_{j,i}(\beta_i, t_i) \geqslant 1$ and $K_+ = K \vee 1$, it suffices to have, using $\varepsilon_n(i) \geqslant 1/\sqrt{n}$ and $a^* \leqslant n$ as noted earlier,

$$\varepsilon_n(i)^{2\alpha_i} \geqslant \{B_1 a^{*-2\alpha_i\beta_i}\} \vee \{B_2 a^{*t_i} \log^{1+t_i}(n)/n\}, \tag{44}$$

where $B_1^{1/\alpha_i} = C(\mathcal{C}_{1,i} K_+ t_i^2)^2$ and $B_2 = C\rho^{-1} K_+{}^2 \mathcal{C}_{2,i}(c_1 t_i^{c_2})^{t_i}$, with c_1, c_2, C universal constants. If $\varepsilon_n(i) \leqslant 1$, we note that (44) implies (43) for C large enough, using $\alpha_i \leqslant 1$, $a^* \geqslant 1$ and $n \geqslant 3$ (which implies $\log(t_i) \lesssim \log^{1+t_i}(n)$).

Optimising a^* in (37), leads to setting

$$(a^*)^{2\alpha_i\beta_i + t_i} = (B_1 n)/(B_2 \log^{1+t_i}(n)) \vee 1. \tag{45}$$

Condition (44) then becomes, for a^* as in (45),

$$\varepsilon_n(i)^{2\alpha_i} \geqslant \left[B_3 (\log n)^{\frac{2\alpha_i\beta_i(1+t_i)}{2\alpha_i\beta_i + t_i}} n^{-\frac{2\alpha_i\beta_i}{2\alpha_i\beta_i + t_i}}\right] \vee \left[B_2 \log^{1+t_i}(n)/n\right], \tag{46}$$

where $B_3 = K^2 c_5^{t_i} t_i^{t_i \frac{c_6}{2\alpha_i\beta_i + t_i}}$, recalling $B_2 = K^2(c_3 t_i^{c_4})^{t_i}$, and $c_3, \ldots, c_6 \geqslant 1$ are constants only depending on β_i, ρ. For $\varepsilon_n(i)$ equal to the lower bound in (46), we indeed have $\varepsilon_n(i) \leqslant 1$ and, for n large enough, $\varepsilon_n(i)^2 = B_3^{1/\alpha_i} (\log n)^{\frac{2\beta_i(1+t_i)}{2\alpha_i\beta_i + t_i}} n^{-\frac{2\beta_i}{2\alpha_i\beta_i + t_i}}$. Condition (42) is then satisfied by definition.

It now remains to prove that $\prod_{k=1}^{d_i} \pi(I_k) \geqslant e^{-\rho n \varepsilon_n(i)^{2\alpha_i}/2}$ given the definition of $v_{i,n}$ and condition (45). Using the fact that $1 \leqslant a^* \leqslant n/2$, a straightforward modification of the proof of Lemma G.17 [12] gives that it is satisfied for a parameter $\tau_i > 0$ satisfying

$$n\varepsilon_n(i)^{2\alpha_i} \geqslant (2/\rho)\left[t_i \log(10a^*/\tau_i) - d_i \log(v_{i,n} e_0 \tau_i) + \log 2\right], \tag{47}$$

whenever $v_{i,n} < \tau_i < a^*$. This last condition is satisfied for any fixed $\tau_i > 0$ as $v_{i,n} \to 0$ and $a^* \to \infty$. Also, equation (47) is satisfied for large enough n as the left-hand side has a polynomial growth and the right-hand side has a logarithmic growth in n.

This concludes the proof of Theorem 3 in the case of the prior $\mathsf{Deep–HGP}$. The proof for the prior $\mathsf{Deep–HGP}(q_{max}, d_{max})$ is very similar and can be found in Appendix H [12].

6.4. *Proof of Theorem 4.* We proceed as in the proof of Theorem 3 but with the new horseshoe prior with shrinking parameter τ_0 on the lengthscales of the first layer, with special attention to that layer of GPs, $i = 0$, as d_0, t_0 may now go to infinity. As in Corollary 3, for $i = 0$, we now have

$$\varepsilon_n(0) \geqslant \left[B_3(\log n)^{\frac{2\alpha_0\beta_0(1+t_0)}{2\alpha_0\beta_0+t_0}} n^{-\frac{2\alpha_0\beta_0}{2\alpha_0\beta_0+t_0}}\right] \vee \left[B_2 \log^{1+t_0}(n)/n\right], \tag{48}$$

which as $n \to \infty$, under (23), is equal to $\varepsilon_n(0) = C^{t_0} n^{-\frac{2\alpha_0\beta_0}{2\alpha_0\beta_0+t_0}} (\log n)^{\frac{2\alpha_0\beta_0(1+t_0)}{2\alpha_0\beta_0+t_0}}$, for C depending on K, ρ and β_i, $i \geqslant 0$. Also the condition on τ_0 becomes

$$10a^* e^{-n\rho\varepsilon_n(0)^{2\alpha_0}/4t_0} \leqslant \tau_0 \leqslant C_0 \frac{\varepsilon_n(0)^{1-\alpha_0}}{\sqrt{nd_0^4}} \tag{49}$$

for a^* as in (45) and C_0 depending ξ and ρ, via a slight modification of Lemma G.18 [12]. As it is satisfied under the assumption of the theorem, this concludes the proof.

7. Dimension-dependent bounds for multibandwidth SqExp Gaussian processes.

In order to prove posterior contraction rates for deep GPs, a key step is to derive an upper bound for the concentration function (9). The next Theorem enables us in particular to revisit Lemmas 4.2 and 4.3 from [5], with explicit multiplicative constants depending on the ambient dimension d in the result. This is a novel contribution to the literature on squared-exponential GPs, to the best of our knowledge. Also, these results allow us to deploy the HGP and $\mathsf{Deep–HGP}$ priors in the high-dimensional setting. For simplicity we do not consider here the anisotropic case in which the function f_0 can have varying smoothness across coordinates, although this could be done as well, following the approach of [5]. We focus on the variable selection aspect of the problem, assuming the same regularity on the active directions of f_0.

THEOREM 5. *Let W^A be a SqExp Gaussian process in dimension $d \geqslant 1$ with deterministic vector of scalings $A = (A_1, \dots, A_d)$ with $a \leqslant A_i \leqslant 2a$ for $i = 1, \dots, d$ and some $a \geqslant \sqrt{\log(2)/d}/2$. Let $\varphi^A_{f_0}(\varepsilon)$ be the concentration function of W^A. Suppose $f_0 \in \mathcal{F}(\beta, K, d)$ for some $\beta, K > 0$. There exist constants $\mathcal{C}_1(\beta, d)$ and $\mathcal{C}_2(\beta, d)$ depending only on β, d and a universal $c, C > 0$ such that, if*

$$\mathcal{C}_1(\beta, d) K^2 a^{-\beta} \leqslant \varepsilon \leqslant 4,$$

then

$$\varphi^A_{f_0}(\varepsilon) \leqslant \mathcal{C}_2(\beta, d) K^2 a^d + (Ca)^d d^{cd} \log^{1+d}(2a/\varepsilon).$$

Moreover, for $i = 1, 2$ one can take $\mathcal{C}_i(\beta, d) = c_i(\beta) C_i(\beta)^d$ for some constants $c_i(\beta), C_i(\beta)$ that depend only on β.

The proof of this result can be found in Appendix A [12].

Acknowledgments. The authors would like to thank François Bachoc, Agnès Lagnoux, Johannes Schmidt-Hieber, Aad van der Vaart as well as two referees for insightful comments. TR did most of his work on the project while a PhD student at Sorbonne Université and a postdoctoral fellow in the Department of Decision Sciences and BIDSA at Bocconi University.

Funding. This work is co-funded by ANR GAP project (ANR-21-CE40-0007) and the European Union (ERC, BigBayesUQ, project number: 101041064).

REFERENCES

[1] ABRAHAM, K. and DEO, N. (2023). Deep Gaussian Process Priors for Bayesian Inference in Nonlinear Inverse Problems. Arxiv preprint 2312.14294.

[2] AGAPIOU, S. and CASTILLO, I. (2024). Heavy-tailed Bayesian nonparametric adaptation. *Ann. Statist.* **52** 1433–1459. https://doi.org/10.1214/24-aos2397 MR4804815

[3] BACHOC, F. and LAGNOUX, A. (2025). Posterior contraction rates for constrained deep Gaussian processes in density estimation and classification. *Communications in Statistics - Theory and Methods* **54** 774–811. https://doi.org/10.1080/03610926.2024.2321185

[4] BAI, J., SONG, Q. and CHENG, G. (2020). Efficient Variational Inference for Sparse Deep Learning with Theoretical Guarantee. In *Advances in Neural Information Processing Systems* **33** 466–476.

[5] BHATTACHARYA, A., PATI, D. and DUNSON, D. (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *The Annals of Statistics* **42** 352–381.

[6] BHATTACHARYA, A., PATI, D. and YANG, Y. (2019). Bayesian fractional posteriors. *The Annals of Statistics* **47** 39 – 66. https://doi.org/10.1214/18-AOS1712

[7] CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465-480. https://doi.org/10.1093/biomet/asq017

[8] CASTILLO, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* **2** 1281–1299. https://doi.org/10.1214/08-EJS273 MR2471287

[9] CASTILLO, I. (2012). A semiparametric Bernstein-von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields* **152** 53–99. MR2875753

[10] CASTILLO, I. and EGELS, P. (2024). Posterior and variational inference for deep neural networks with heavy-tailed weights. arXiv preprint 2406.03369.

[11] CASTILLO, I., KERKYACHARIAN, G. and PICARD, D. (2014). Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields* **158** 665–710. https://doi.org/10.1007/s00440-013-0493-0 MR3176362

[12] CASTILLO, I. and RANDRIANARISOA, T. (2024). Supplementary material to 'Deep Horseshoe Gaussian Processes'.

[13] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** 2353–2383. https://doi.org/10.1214/15-AOS1336 MR3405597

[14] CATONI, O. (2004). *Statistical learning theory and stochastic optimization*. *Lecture Notes in Mathematics* **1851**. Springer-Verlag, Berlin Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001. https://doi.org/10.1007/b99352 MR2163920

[15] CHÉRIEF-ABDELLATIF, B. (2020). Convergence Rates of Variational Inference in Sparse Deep Learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*. *Proceedings of Machine Learning Research* **119** 1831–1842.

[16] DAMIANOU, A. and LAWRENCE, N. D. (2013). Deep Gaussian Processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. *Proceedings of Machine Learning Research* **31** 207–215.

[17] DE G. MATTHEWS, A. G., HRON, J., ROWLAND, M., TURNER, R. E. and GHAHRAMANI, Z. (2018). Gaussian Process Behaviour in Wide Deep Neural Networks. In *International Conference on Learning Representations*.

[18] DUTORDOIR, V., SALIMBENI, H., HAMBRO, E., MCLEOD, J., LEIBFRIED, F., ARTEMEV, A., VAN DER WILK, M., HENSMAN, J., DEISENROTH, M. P. and JOHN, S. (2021). GPflux: A Library for Deep Gaussian Processes. *arXiv e-prints* arXiv:2104.05674. https://doi.org/10.48550/arXiv.2104.05674

[19] FINOCCHIO, G. and SCHMIDT-HIEBER, J. (2023). Posterior contraction for deep Gaussian process priors. *Journal of Machine Learning Research* **24** 1–49.

[20] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. https://doi.org/10.1214/aos/1016218228 MR1790007

[21] Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781139029834 MR3587782

[22] Giordano, M., Ray, K. and Schmidt-Hieber, J. (2022). On the inability of Gaussian process regression to optimally learn compositional functions. In *Advances in Neural Information Processing Systems* **35** 22341–22353.

[23] Hazan, T. and Jaakkola, T. (2015). Steps toward deep kernel methods from infinite neural networks. *arXiv preprint arXiv:1508.05133*.

[24] Jiang, S. and Tokdar, S. T. (2021). Variable selection consistency of Gaussian process regression. *Ann. Statist.* **49** 2491–2505. https://doi.org/10.1214/20-aos2043 MR4338372

[25] Kohler, M. and Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *Ann. Statist.* **49** 2231–2249. https://doi.org/10.1214/20-aos2034 MR4319248

[26] Kruijer, W. and van der Vaart, A. (2013). Analyzing posteriors by the information inequality. In *From probability to statistics and back: high-dimensional models and processes*. *Inst. Math. Stat. (IMS) Collect.* **9** 227–240. Inst. Math. Statist., Beachwood, OH. https://doi.org/10.1214/12-IMSCOLL916 MR3202636

[27] L'Huillier, A., Travis, L., Castillo, I. and Ray, K. (2023). Semiparametric Inference Using Fractional Posteriors. *Journal of Machine Learning Research* **24** 1–61.

[28] Mai, T. T. (2024). Adaptive posterior concentration rates for sparse high-dimensional linear regression with random design and unknown error variance. arXiv preprint 2405.19016.

[29] Moriarty-Osborne, C. and Teckentrup, A. L. (2024). Convergence rates of non-stationary and deep Gaussian process regression. arXiv preprint arXiv:2312.07320.

[30] Neal, R. M. (2012). *Bayesian learning for neural networks* **118**. Springer Science & Business Media.

[31] Ohn, I. and Lin, L. (2024). Adaptive variational Bayes: optimality, computation and applications. *Ann. Statist.* **52** 335–363. https://doi.org/10.1214/23-aos2349 MR4718418

[32] Pati, D., Bhattacharya, A. and Cheng, G. (2015). Optimal Bayesian estimation in random covariate design with a rescaled Gaussian process prior. *J. Mach. Learn. Res.* **16** 2837–2851. MR3450525

[33] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

[34] Rocková, V. and Polson, N. (2018). Posterior Concentration for Sparse Deep Learning. In *Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018* 938–949.

[35] Salimbeni, H. and Deisenroth, M. (2017). Doubly Stochastic Variational Inference for Deep Gaussian Processes. In *Advances in Neural Information Processing Systems* **30**.

[36] Salimbeni, H., Dutordoir, V., Hensman, J. and Deisenroth, M. (2019). Deep Gaussian Processes with Importance-Weighted Variational Inference. In *Proceedings of the 36th International Conference on Machine Learning*. *Proceedings of Machine Learning Research* **97** 5589–5598.

[37] Sauer, A. (2022). deepgp: Deep Gaussian Processes using MCMC R package version 1.1.1.

[38] Sauer, A., Cooper, A. and Gramacy, R. B. (2023). Vecchia-Approximated Deep Gaussian Processes for Computer Experiments. *Journal of Computational and Graphical Statistics* **32** 824-837. https://doi.org/10.1080/10618600.2022.2129662

[39] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.* **48** 1875–1897. https://doi.org/10.1214/19-AOS1875 MR4134774

[40] Szabó, B., van der Vaart, A. W. and van Zanten, H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** 1391–1428. (with discussion).

[41] Tang, T., Wu, N., Cheng, X. and Dunson, D. (2024). Adaptive Bayesian Regression on Data with Low Intrinsic Dimensionality. arXiv preprint arXiv:2407.09286.

[42] Teckentrup, A. L. (2020). Convergence of Gaussian process regression with estimated hyper-parameters and applications in Bayesian inverse problems. *SIAM/ASA J. Uncertain. Quantif.* **8** 1310–1337. https://doi.org/10.1137/19M1284816 MR4164077

[43] van der Pas, S., Szabó, B. and van der Vaart, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* **12** 1221–1274. With a rejoinder by the authors. https://doi.org/10.1214/17-BA1065 MR3724985

[44] van der Pas, S. L., Kleijn, B. J. K. and van der Vaart, A. W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics* **8** 2585 – 2618. https://doi.org/10.1214/14-EJS962

[45] van der Pas, S. L., Kleijn, B. J. K. and van der Vaart, A. W. (2014). The horseshoe estimator: posterior concentration around nearly black vectors. *Electron. J. Stat.* **8** 2585–2618. https://doi.org/10.1214/14-EJS962 MR3285877

[46] van der Vaart, A. and van Zanten, H. (2007). Bayesian inference with rescaled Gaussian process priors. *Electron. J. Stat.* **1** 433–448. https://doi.org/10.1214/07-EJS098 MR2357712

[47] van der Vaart, A. and van Zanten, H. (2011). Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res.* **12** 2095–2119. MR2819028

[48] van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. https://doi.org/10.1214/009053607000000613 MR2418663

[49] van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. https://doi.org/10.1214/08-AOS678 MR2541442

[50] Yang, Y., Bhattacharya, A. and Pati, D. (2017). Frequentist coverage and sup-norm convergence rate in Gaussian process regression. *arXiv e-prints* arXiv:1708.04753. https://doi.org/10.48550/arXiv.1708.04753

[51] Yang, Y. and Dunson, D. B. (2016). Bayesian manifold regression. *The Annals of Statistics* **44** 876 – 905. https://doi.org/10.1214/15-AOS1390

[52] Yang, Y. and Tokdar, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43** 652–674. https://doi.org/10.1214/14-AOS1289 MR3319139

[53] Yu, W., Wade, S., Bondell, H. D. and Azizi, L. (2023). Nonstationary Gaussian process discriminant analysis with variable selection for high-dimensional functional data. *J. Comput. Graph. Statist.* **32** 588–600. https://doi.org/10.1080/10618600.2022.2098136 MR4592932

[54] Zhang, T. (2006). From ϵ-entropy to KL-entropy: analysis of minimum information complexity density estimation. *Ann. Statist.* **34** 2180–2210. https://doi.org/10.1214/009053606000000704 MR2291497

[55] Zhao, Z., Emzir, M. and Särkkä, S. (2021). Deep state-space Gaussian processes. *Stat. Comput.* **31** Paper No. 75, 26. https://doi.org/10.1007/s11222-021-10050-6 MR4319943