Posterior and Variational Inference for Deep Neural Networks with Heavy-Tailed Weights

Ismaël Castillo[*]   ISMAEL.CASTILLO@SORBONNE-UNIVERSITE.FR
LPSM, Sorbonne Université
4, place Jussieu
75005, Paris, France

Paul Egels[*]   PAUL.EGELS@SORBONNE-UNIVERSITE.FR
LPSM, Sorbonne Université
4, place Jussieu
75005, Paris, France

Editor: Pierre Alquier

Abstract

We consider deep neural networks in a Bayesian framework with a prior distribution sampling the network weights at random. Following a recent idea of Agapiou and Castillo (2024), who show that heavy-tailed prior distributions achieve automatic adaptation to smoothness, we introduce a simple Bayesian deep learning prior based on heavy-tailed weights and ReLU activation. We show that the corresponding posterior distribution achieves near-optimal minimax contraction rates, *simultaneously* adaptive to both intrinsic dimension and smoothness of the underlying function, in a variety of contexts including nonparametric regression, geometric data and Besov spaces. While most works so far need a form of model selection built-in within the prior distribution, a key aspect of our approach is that it does not require to sample hyperparameters to learn the architecture of the network. We also provide variational Bayes counterparts of the results, that show that mean-field variational approximations still benefit from near-optimal theoretical support.

Keywords: Bayesian deep neural networks, fractional posterior distributions, heavy-tailed priors, Variational Bayes, overparametrized regime.

1. Introduction

The last decade has seen a remarkable expansion of the use of deep neural networks (DNNs) through a broad range of applications such as imaging, natural language processing, inverse problems to name a few. In parallel to this spectacular empirical success, theory to provide understanding of the performance of these methods is emerging. Among the key elements playing a role in the mathematical and statistical analysis of DNNs are their approximation properties (e.g. Yarotsky (2017); Schmidt-Hieber (2020); Kohler and Langer (2021)), the choice of parameters in particular the network's *architecture*, and the convergence of the sampling or optimisation schemes (e.g. gradient descent, variational criteria etc.) involved in the training.

---

[*]Equal contribution.

Using a Bayesian approach for inference with deep neural networks is particularly appealing, among others for the natural way to quantify uncertainty through credible sets of the posterior distribution. A first contribution on theoretical understanding in this direction is the work by Polson and Ročková (2018), who show, building up in particular on the approximation theory derived in Schmidt-Hieber (2020), that assigning spike–and–slab prior distributions to the weights of a DNN with ReLU activation leads to a near-minimax posterior convergence rate in nonparametric regression. While from the mathematical perspective spike–and–slab priors can be seen as a form of theoretical ideal, they can suffer from a particularly high sampling complexity; indeed, sampling from the associated posterior distributions often faces combinatorial difficulties, as one needs to explore a high number of possible candidate models. While a possible answer, as we discuss in more details below, is to deploy a Variational Bayes (VB) approximation, it is particularly desirable (and even if VB is chosen as an approximation) to develop simpler prior distributions, that still retain most mathematical properties but are easier to implement in practice.

A number of recent contributions have considered prior distributions beyond spike–and–slab (SAS) priors for the DNN's weights. From the theoretical perspective let us cite the work by Lee and Lee (2022), who broadens the results for SAS of Polson and Ročková (2018) and considers a class of shrinkage priors (whose characteristics depend on smoothness parameters of the function to be estimated); the work by Kong et al. (2023) considers masked Bayesian neural networks, where the 'mask' determines the position of zero weights, and which allows for more efficient reparametrisation in view of computations compared to SAS posteriors. The preprint Kong and Kim (2024) uses non-sparse Gaussian priors on weights, building on the approximation theory of Kohler and Langer (2021). From the algorithmical perspective, the work by Ghosh et al. (2019) considers using horseshoe priors on weights. More discussion on the previously mentioned priors and their relation to the present work can be found in Section 4 below.

Let us mention also two lines of work in Bayesian deep learning that are somewhat different from our approach but with natural connections, in particular in terms of the tails of the distributions arising in the networks. The first considers DNNs in the large width limits, following seminal ideas of Neal (1996): using Gaussian priors on weights, Sell and Singh (2023) use the approximation by a Gaussian process obtained obtained in large width to propose posterior sampling schemes; the work by Lee et al. (2023) instead considers the large width limit with heavier-tailed distributions on weights, obtaining a mixture of Gaussian processes in the limit. The second line of work has no activation functions involved, but models directly random functions accross layers, which leads to deep Gaussian processes as introduced in Damianou and Lawrence (2013) and recently investigated in terms of convergence by Finocchio and Schmidt-Hieber (2023) and Castillo and Randrianarisoa (2024).

Variational Bayes is a particularly popular approach in machine learning and statistics, where the idea is to approximate a possibly complex posterior distribution by an element of a simpler class of distributions, effectively requiring to solve an optimisation problem, see e.g. Blei et al. (2017). This is a method of choice for approximating posteriors in deep neural networks, with a statistical-computational trade-off through the choice of the variational class, which is often taken to be mean-field, thereby requiring a form of independence along a certain parametrisation. Following general ideas from Alquier and Ridgway (2020) (see

also Yang et al. (2020); Zhang and Gao (2020)), the work by Chérief-Abdellatif (2020) proves that taking a mean-field variational class combined to the spike-and-slab prior of Polson and Ročková (2018) leads to a variational posterior that converges at the same near-optimal rates towards the true regression function. Simulations using stochastic gradient optimisation for this variational posterior, and some extensions, are considered in Bai et al. (2020). Using simpler sieve-type priors on DNNs but requiring a form of model selection, the work Ohn and Lin (2024) considers adaptation within a family of variational posteriors.

For more discussion and further references on Bayesian deep learning we refer to Papamarkou et al. (2024), Alquier (2024), Castillo (2024), Chapter 4, and to the review paper Arbel et al. (2023).

Let us now turn to a summary of the contributions of the paper. We prove that putting a suitably rescaled heavy-tailed distribution on deep neural network weights leads to an optimal rate of convergence (up to logarithmic factors) of the corresponding posterior distribution in nonparametric regression with random design. The best rate is automatically attained without assuming any knowledge of regularity parameters of the unknown regression function, achieving in particular *simultaneous adaptation* to *both* smoothness and intrinsic dimensions when the regression function has a compositional structure with effective intrinsic dimensions (typically) of smaller order than the input dimension. We derive similar results for scalable (mean-field) Variational Bayes approximations of the posterior distribution; each coordinate of the VB approximation has only two parameters to be fitted, one for location and one for scale.

To help putting the results into perspective, we note that many fundamental statistical results so far on DNNs are often stated with an 'oracle' choice of the network architecture: for wide and moderately deep DNNs for instance, if the network width is well chosen in terms of both smoothness and dimensionality parameters, Schmidt-Hieber (2020) and Kohler and Langer (2021) show that an empirical risk minimiser on neural networks of the *prescribed* (oracle) architecture achieves a near-optimal minimax contraction rate for compositional classes. Similarly, Suzuki (2018) shows that deep learning methods are 'adaptive' to the intrinsic dimension for anisotropic smoothness classes, with adaptation meaning here that the achieved rate depends only on the intrinsic dimension (as opposed to the input dimension); however the results are achieved with a network architecture that still depends on the smoothness and intrinsic dimension parameters, which are typically unknown in practice. Building up on the seminal approximation results of the just cited papers (as well as Nakada and Imaizumi (2020) for data on geometric objects), we show that well-chosen heavy tailed weight distributions enable one to obtain statistical adaptation, simultaneously in terms of smoothness and dimension parameters, using Bayesian fractional posterior distributions or their mean-field variational counterparts. Beyond classes of compositions in nonparametric random design regression, to illustrate the flexibility of our method we consider two other applications: geometrical data through the use of the Minkowski dimension as in Nakada and Imaizumi (2020) and anisotropic classes as in Suzuki and Nitanda (2021). For both we show that simultaneous statistical adaptation is achieved for fractional posteriors and mean-field VB.

Outline of the paper. In Section 2, we introduce the statistical framework, deep neural networks, and the proposed method; the heavy-tailed prior distribution on network weights

is defined, as well as the fractional posterior distribution and its variational approximation. In Section 3 we present our main results, declined along three settings: compositional classes in Sections 3.1–3.2, geometric data in Section 3.3 and anisotropic classes in Section 3.4; in each setting results are derived for the fractional posterior distribution, and a mean-field VB approximation. in Section 3.5, we discuss the impact of the network architecture, particularly its depth. Then, we also explore extensions of our results in multiple directions: in Section 3.6, we consider the choice of activation function; in Section 3.7, we address the question of knowledge about the noise variance; and in Section 3.8, we examine the case of the standard (non-fractional) posterior. Section 4 contains a discussion on possible choices of prior for Bayesian deep neural networks, discussion on sampling and algorithms and future work and open questions. Section 5 contains the proof of the main results. The Appendix includes a number of short technical lemmas, a lemma on bounds of weights of neural networks and proofs for all possible extensions of our main results.

Notation. For any real number β, let $\lfloor \beta \rfloor$ be the largest integer strictly smaller than β. We write the set of positive real numbers as $\mathbb{R}_{>0}$. For any two sequences we write $a_n \lesssim b_n$ for inequality up to a constant and $a_n \asymp b_n$ whenever $a_n \gtrsim b_n$ is also satisfied. For two real numbers a, b, we write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For $T \geq 1$ an integer, $[T]$ denotes the set of integers $\{1, \ldots, T\}$. For $1 \leq p \leq \infty$, we denote $|\cdot|_p$ (resp. $\|\cdot\|_p$) the ℓ^p (resp. L^p) norms and $\|\cdot\|_{L^p(\mu)}$ when we want to specify the integrating measure, say μ. For $d \geq 1$ an integer, $F \geq 0$ and $D \subset \mathbb{R}^d$, let $\mathcal{C}_d^\beta(D, F)$ be the ball of d-variate β-Hölder functions

$$\mathcal{C}_d^\beta(D, F) := \left\{ f : D \subset \mathbb{R}^d \to \mathbb{R} : \sum_{|\mathbf{k}| < \beta} ||\partial^{\mathbf{k}} f||_\infty + \sum_{|\mathbf{k}| = \lfloor \beta \rfloor} \sup_{\substack{x,y \in D \\ x \neq y}} \frac{|\partial^{\mathbf{k}} f(x) - \partial^{\mathbf{k}} f(y)|}{|x - y|_\infty^{\beta - \lfloor \beta \rfloor}} \leq F \right\}.$$

Elements of this ball have all their partial derivatives up to order $\lfloor \beta \rfloor$ bounded, and their partial derivatives of order $\lfloor \beta \rfloor$ are $(\beta - \lfloor \beta \rfloor)$-Hölder-continuous. We denote by $\mathrm{KL}(P, Q)$ and $D_\alpha(P, Q)$ respectively the Kullback-Leibler and Rényi divergence of order α between probability measures P and Q (see Appendix A for their definitions).

2. Bayesian deep neural networks

Let us now introduce the notation and statistical framework used to express our results.

2.1 Statistical framework

We consider the nonparametric random design Gaussian regression model with possibly large (but fixed) dimension $d \geq 1$: one observes n independent and identically distributed (iid) pairs of random variables $(X_i, Y_i) \in [0, 1]^d \times \mathbb{R}$, $1 \leq i \leq n$, with, for $\tau_0 > 0$,

$$Y_i = f_0(X_i) + \tau_0 \xi_i. \tag{1}$$

The design points (X_i) are iid draws from a distribution P_X on $[0, 1]^d$ and, independently, the noise variables $(\xi_i)_i$ are iid with $\mathcal{N}(0, 1)$ distribution. From now on and unless stated otherwise we assume for simplicity that τ_0 is known and take $\tau_0 = 1$. In Section 3.7 we

extend our results to the case where the noise level τ_0 is unknown. The function f_0 belongs to a parameter set $\mathcal{F}$ of functions to be specified below. Writing P_{f_0} for the distribution of (X_1, Y_1), the joint distribution of the observations $(X, Y) := \{(X_i, Y_i)_i\}$ is $P_{f_0}^{\otimes n}$. The inferential goal here is to recover the 'true' function f_0 from the observed data $(X_i, Y_i)_i$, and our contraction rate results will be given in terms of the integrated quadratic loss

$$||f_0 - f||^2_{L^2(P_X)} = \int (f_0 - f)^2 \, dP_X.$$

The minimax rate of estimation of a β–smooth function f_0, in the Hölder sense say, in squared-integrated loss in this model is known to be of order $n^{-2\beta/(2\beta+d)}$, which suffers from the curse of dimensionality, as this rate becomes very slow for large d unless f_0 is very smooth. In recent years regularity classes based on compositions of functions have emerged as a benchmark for the performance of deep learning methods: indeed, these classes enable to model smaller effective dimensions compared to the input dimension d, with optimal rates, to be described in more details below, that only depend on the intrinsic dimensions at the successive steps of the composition.

Compositional structure. Let $q \geq 1$ be an integer, $K > 0$, $\mathbf{d} = (d_0, \ldots, d_{q+1})$, $\mathbf{t} = (t_0, \ldots t_q)$ vectors of ambient and effective dimensions such that $t_i \leq d_i$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_q)$ a vector of smoothness parameters. Following Schmidt-Hieber (2020), define a class of compositions as

$$\mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K) := \Big\{ f = g_q \circ \cdots \circ g_0 \, : \, g_i = (g_{ij})_{j \in [d_{i+1}]} : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}},$$

$$g_{ij} \in \mathcal{C}_{t_i}^{\beta_i}([a_i, b_i]^{t_i}, K), \quad |a_i|, |b_i| \leq K \Big\},$$

where we denote with a slight abuse of notation $\mathcal{C}_{t_i}^{\beta_i}$ the Hölder ball of functions $g_{ij} : [a_i, b_i]^{d_i} \to \mathbb{R}$ that can be written

$$g_{ij}(z_1, \ldots, z_{d_i}) = h_{ij}(z_{q_1}, \ldots, z_{q_{t_i}}),$$

for $\{q_1, \ldots, q_{t_i}\}$ a subset of t_i indices among $\{1, \ldots, d_i\}$ and $h_{ij} : [a_i, b_i]^{t_i} \to \mathbb{R}$ a Hölder function. We also set $d_0 = d$ and $d_{q+1} = 1$ (for an alternative definition, see also Finocchio and Schmidt-Hieber (2023)). Let us emphasize that even if the g_{ij}'s take as input a vector of size d_i, the value of the output depends on at most $t_i \leq d_i$ variables, which allows one to model low-dimensional structures. We further note that a function f_0 from $\mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$ does not have a unique decomposition as a composition of g_i's. This is not a concern here, as we are solely interested in estimating f_0 and not g_i's themselves. Many models, such as additive models and sparse tensor decompostions, fall within this framework, see e.g. Schmidt-Hieber (2020), Section 4, for more discussion.

The minimax rate of estimation of f_0 over the above compositional class, in terms of the quadratic loss as defined above turns out to depend only on the effective dimensions t_i and 'effective smoothness' parameters, for $0 \leq i \leq q$,

$$\beta_i^* := \beta_i \prod_{k=i+1}^{q} (\beta_k \wedge 1). \tag{2}$$

An intuition behind this is that Hölder regularities larger than 1 cannot in general 'propagate' through inner layers of composition to make the overall Hölder smoothness larger than individual regularities of the components, but regularities smaller than one can decrease the overall smoothness.

Schmidt-Hieber (2020), Theorem 3, shows that if $t_i \leq \min(d_0, \ldots, d_{q+1})$ for all i, then

$$\phi_n^* := \max_{0 \leq i \leq q} \left\{ n^{-\frac{\beta_i^*}{2\beta_i^* + t_i}} \right\}. \tag{3}$$

is a lower bound for the minimax rate of estimation in the regression model for compositional functions and squared-integrated loss.

2.2 Deep ReLU neural networks

The rectified linear unit (ReLU) activation function is defined for any real x as

$$\rho(x) := \max(0, x).$$

When x is a multi-dimensional vector, we simply apply ρ coordinate–wise. Throughout most of this paper we focus on DNNs with ReLU activation as it is one of the most commonly used and studied activation functions in deep learning. We provide a simple way to extend our results to other smooth enough activation functions in Section 3.6. Let $L \geq 1$ be an integer and $\mathbf{r} = (r_0, \ldots, r_{L+1}) \in \mathbb{N}^{L+2}$, we say that a function $f : \mathbb{R}^{r_0} \to \mathbb{R}^{r_{L+1}}$ is the realisation of a ReLU neural network with architecture (or structure) $(L, \mathbf{r})$ if there is a family of matrix-vectors $\{(W_l, v_l)\}_{l \in [L+1]}$, with $(W_l, v_l) \in \mathbb{R}^{r_l \times r_{l-1}} \times \mathbb{R}^{r_l}$ such that

$$f = A_{L+1} \circ \rho \circ \cdots \circ \rho \circ A_1, \tag{4}$$

where A_l is the affine map $y \mapsto W_l y + v_l$. Here L and $\mathbf{r}$ are respectively called the *length* (or *depth*) and the *width vector* of the network architecture. Here we take the convention that the composition in (4) ends with an affine transformation with a shift (as in Kohler and Langer (2021)) and not simply a linear one. The network weights are the entries of the matrices (W_l) and shifts vectors (v_l) of the successive layers. The total number of weight parameters of a network with architecture $(L, \mathbf{r})$ is

$$T = T(L, \mathbf{r}) := \sum_{l=0}^{L} r_l r_{l+1} + \sum_{l=1}^{L+1} r_l. \tag{5}$$

Let $\theta := (\theta_k)_{1 \leq k \leq T} \in \mathbb{R}^T$ be a vector containing all the network weights written in a given order (the choice of the order does not matter, but is fixed once and forall). Given a family of matrices-vectors $\{(W_l, v_l)\}_{l \in [L+1]}$ specifying a neural network, to simplify the notation it is sometimes easier to consider the shift vector v_l to be the 0-th column of W_l so that the set of parameters is $(\theta_k : k \in [T]) \equiv (W_l^{(ij)} : l \in [L+1], (i,j) \in \{1, \ldots, r_l\} \times \{0, \ldots, r_{l-1}\})$. This is made rigorous from the fact that

$$\begin{pmatrix} v^{(1)} & W^{(11)} & \cdots & W^{(1r_{l-1})} \\ \vdots & \vdots & & \vdots \\ v^{(r_l)} & W^{(r_l 1)} & \cdots & W^{(r_l r_{l-1})} \end{pmatrix} \begin{pmatrix} 1 \\ x^{(1)} \\ \vdots \\ x^{(r_{l-1})} \end{pmatrix} = Wx + v.$$

The set of networks $\mathcal{F}(L, \mathbf{r})$ of architecture $(L, \mathbf{r})$. Let $\mathcal{F}(L, \mathbf{r})$ be the set of all realisations of ReLU neural networks with structure $(L, \mathbf{r})$, for $L \geq 1$ a given depth and $\mathbf{r}$ a vector of widths. Also define $\mathcal{F}(L, \mathbf{r}, s)$ as the set of realisations of neural networks that have at most s active parameters, that is $\mathcal{F}(L, \mathbf{r}, s) := \{f \in \mathcal{F}(L, \mathbf{r}), \#\{k \in [T]; \theta_k \neq 0\} \leq s\}$, with $\#S$ the cardinality of a finite set S.

2.3 Heavy-tailed priors on deep neural networks

Frequentist analysis of tempered posteriors. The reconstruction of f_0 from the data $(X_i, Y_i)_i$ given by the regression model (1) will be conducted through a (generalised–) Bayesian analysis. From a prior distribution Π on a (measurable) parameter space $\mathcal{F}$ containing f_0, given $\alpha \in (0, 1)$, one can construct a data-dependent probability measure on $\mathcal{F}$ called tempered posterior distribution, given by, for any measurable $A \subset \mathcal{F}$,

$$\Pi_\alpha[A \,|\, X, Y] := \frac{\int_A e^{-\frac{\alpha}{2} \sum_{i=1}^n (Y_i - f(X_i))^2} d\Pi(f)}{\int e^{-\frac{\alpha}{2} \sum_{i=1}^n (Y_i - f(X_i))^2} d\Pi(f)}. \tag{6}$$

Setting α equal to 1 in the last display leads to the classical posterior distribution, the conditional distribution of f given X, Y in model (1), when f is equipped with a prior distribution Π. Here we (mostly) work with a tempered (or 'fractional') posterior instead, which corresponds to rising the Gaussian likelihood to a power α, here taken to be smaller than 1, which effectively downweights the influence of the data; a result for the classical posterior ($\alpha = 1$) and an augmented prior is discussed in Section 3.8. Tempered posteriors are particular cases of so-called Gibbs posteriors, where the (log–)likelihood is itself replaced by a user-chosen empirical quantity; both objects are particularly popular in both machine learning and statistical applications, see e.g. L'Huillier et al. (2023) (or the more general references Ghosal and van der Vaart (2017); Alquier (2024)) for more discussion. The reason we consider fractional posteriors with $\alpha < 1$ is mainly of technical nature and related to our choice of prior below; we refer to the discussion in Section 4 for more on this.

To analyse the convergence of the fractional posterior distribution, we take a frequentist approach, which means that we analyse the behaviour of $\Pi_\alpha[\cdot \,|\, X, Y]$ in (6) under the assumption that the data has effectively been generated according to model (1) with f_0 a fixed 'true' regression function. The goal is then to characterise the rate at which the fractional posterior distribution concentrates its mass around the true function f_0 (if at all). It turns out that a sufficient condition for this is, for contraction in terms of a Rényi–type divergence, that the prior distribution puts enough mass on a certain neighborhood of f_0 (see e.g. Bhattacharya et al. (2019), L'Huillier et al. (2023) and references therein); the version we use in the present paper can be found as Lemma 17 in Appendix A below. For more context and references on general theory of frequentist analysis of (possibly generalised–) posterior distributions, we refer to the book by Ghosal and van der Vaart (2017) and the monograph Castillo (2024).

Known upper-bound on f_0. As is common in the analysis of deep learning algorithms, we assume for simplicity that an upper-bound M_0 on the true regression function f_0 is known: $\|f_0\|_\infty \leq M_0$. Our results continue to hold in slightly weaker form if this is not the case.

Heavy-tailed priors on network coefficients. For a given depth L and width vector $\mathbf{r}$, we define a prior distribution on $\mathcal{F}(L, \mathbf{r})$ as follows. Importantly, in our method, the choice of $L, \mathbf{r}$ are *deterministic* and given in (11), (12) (or (13)) below.

Let T be the total number of parameters in the network, and recall the notation $[T] = \{1, \dots, T\}$ and that we order network parameters in a fixed arbitrary order. The heavy-tailed prior samples the network coefficients as

$$\forall k \in [T], \qquad \theta_k = \sigma_k \, \zeta_k, \tag{7}$$

- where the scaling factors σ_k for $k \in [T]$ are positive real numbers to be chosen below;

- the random variables ζ_k's for $k \in [T]$ are independent identically distributed with a given heavy-tailed density h on $\mathbb{R}$ that satisfies the following conditions (H1) to (H3): there exist constants $c_1, c_2 > 0$ and $\kappa \geq 0$ such that

(H1) h is symmetric, positive, bounded, and decreasing on $[0, +\infty)$,

(H2) for all $x \geq 0$,
$$\log\left(1/h(x)\right) \leq c_1(1 + \log^{1+\kappa}(1+x)),$$

(H3) for all $x \geq 1$,
$$\overline{H}(x) := \int_x^{+\infty} h(u)du \leq \frac{c_2}{x}.$$

Let us now give typical examples for (σ_k) and h as well as some intuition behind this choice of prior. The conditions on h accommodate many standard choices of densities with polynomial tails, such as Cauchy and most Student densities: all these verify (H2) with $\kappa = 0$ for a sufficiently large constant $c_1 > 0$; a typical possible choice of σ_k allowed in all theorems below is $\exp(-\log^{2(1+\delta)} n)$ for $\delta > 0$, independent of k (i.e. one takes this same value for all network coefficients), with a decrease in n just slightly faster than a polynomial (which would correspond to $\exp(-a \log n)$).

This type of priors has been recently introduced by Agapiou and Castillo (2024) for use in a number of classical nonparametric models such as regression, density estimation and classification, where it is shown to achieve optimal convergence rates (up to log factors) for tempered posteriors with automatic adaptation to smoothness; efficient simulations using infinite-dimensional MCMC schemes are also considered therein.

There are two key ideas behind this choice of prior. First, the heavy tailed distribution enables one to model large values if needed (here 'large' will mean at most of the order of a polynomial in n); second, the scaling factors are taken sufficiently small to prevent overfitting. Unlike for spike-and-slab priors (e.g. Polson and Ročková (2018), Kong et al. (2023)) the prior sets no coefficients exactly to zero. This is often an advantage computationally, as modelling exact zero coefficients in the prior typically requires performing a form of discrete model selection in the posterior sampling, which can have a large computational cost even for moderate sample sizes. Note also that the architecture of the network is non-random: we refer to Section 3 for two possible specific choices of $L, \mathbf{r}$. This is in contrast to model-selection type priors where a prior is put on $\mathbf{r}, L$ or both. Again, performing model selection for instance on the width $\mathbf{r}$ is a form of hyper-parameter selection problem which often leads

to additional computational cost, as one needs to obtain the posterior on the hyperparameter; see also the discussion section in Agapiou and Castillo (2024), where similar comments are made with respect to Gaussian process priors in regression, for which hyperparameters need to be calibrated (e.g. the key lengthscale parameter if one uses a squared-exponential kernel).

The overall prior Π *on DNNs.* We are now in position to define the prior on functions f we consider:

1. choose the architecture $(L, \mathbf{r})$ according to one of the two deterministic choices (11)–(12) or (11)–(13) as specified at the beginning of Section 3 below;

2. given $(L, \mathbf{r})$, sample network weights θ_k according the heavy-tailed prior as defined above; next form the neural network realisation f as in (4).

A final remark is that the choice of architecture parameters we make below places us in the *overparametrized regime*: namely, our prior always fits more weights than the optimal number predicted by the classical bias-variance trade-off, see the Discussion below for more on this.

2.4 Heavy-tailed mean-field tempered variational approximations

Let us now turn to a variational approach, where the idea is to approximate the tempered posterior on a subset $\mathcal{S}$ of (simpler) probability distributions on the parameter space $\mathcal{F}$. The variational tempered approximation (on $\mathcal{S}$) $\hat{Q}_\alpha = \hat{Q}_\alpha(X, Y)$ is defined as the best approximation of the tempered posterior $\Pi_\alpha[\,\cdot\,|X, Y]$ by elements of $\mathcal{S}$ in the KL–sense

$$\hat{Q}_\alpha := \underset{Q \in \mathcal{S}}{\operatorname{argmin}} \, \operatorname{KL}(Q, \Pi_\alpha[\,\cdot\,|X, Y]), \tag{8}$$

where the $\operatorname{KL}(P, Q)$ is Kullback-Leibler divergence between P and Q (see Appendix A). Below we consider a *mean-field* approximation, which consists in taking a class $\mathcal{S}$ of distributions of product form, assuming $\mathcal{F}$ has a natural parametrisation in coordinates $(\theta_1, \ldots, \theta_T)$, which is the case here for DNNs by taking the collection of network weights. The individual elements of the product are taken below to belong to a scale-location family of heavy tailed distributions.

For h a (heavy-tailed) density function on $\mathbb{R}$ and $\mu \in \mathbb{R}, \varsigma > 0$, we let $h_{\mu,\varsigma}$ be the shifted and re-scaled density $h_{\mu,\varsigma} : x \mapsto h((x - \mu)/\varsigma)/\varsigma$ and set

$$\mathcal{H} = \mathcal{H}(h) := \{Q \, : \, dQ(x) = h_{\mu,\varsigma}(x)dx, \ \mu \in \mathbb{R}, \varsigma > 0\} \, .$$

That is, Q belongs to $\mathcal{H}(h)$ if it has a density with respect to the Lebesgue measure $h_{\mu,\varsigma}$ on $\mathbb{R}$ for some μ, ς. Although the class $\mathcal{H}(h)$ can be defined for an arbitrary function h, note that we use the same notation 'h' as for the prior density from the previous section, since in our variational results below (see Theorem 5), we use a shifted-and-rescaled function that matches the prior density. For h a density and $\lambda > 0$, we denote its λ–th moment by

$$m_\lambda(h) := \int |x|^\lambda h(x)\, dx. \tag{9}$$

Definition 1 (Heavy-tailed mean-field class) *Let h be a density function satisfying (H1), (H2), (H3) with finite second order moment $m_2(h) < \infty$. For $T \geq 1$ an integer, we define the heavy-tailed mean-field class (generated from h) as*

$$\mathcal{S}_{HT}(h) := \left\{ Q \,:\, dQ(\theta) = \prod_{k=1}^{T} dQ_k(\theta_k), \text{ with } Q_k \in \mathcal{H}(h) \right\}. \tag{10}$$

Mean-field variational approximations such as (10) are particularly popular (Blei et al., 2017) in that, by removing dependencies along a decomposition of the parameter space $\mathcal{F}$ as above, they often allow for scalable computation of the minimizer defined by the optimisation problem (8). The mean-field class (10) corresponds to our heavy-tailed prior on DNNs defined by (7). Variational approximation for spike-and-slab priors on neural networks have been studied previously by Chérief-Abdellatif (2020) and Bai et al. (2020). Adaptive variational concentration results for hierarchical uniform priors have been recently obtained by Ohn and Lin (2024). Concentration tools for variational approximations are recalled in Appendix A following ideas of Alquier and Ridgway (2020), Yang et al. (2020), Zhang and Gao (2020) (see also Castillo (2024) for an overview). For more context and discussion we also refer to the review paper on variational Bayes by Blei et al. (2017) and the review on Bayesian neural networks Arbel et al. (2023).

3. Main results

We now consider the convergence of tempered posteriors for heavy tailed-priors on the network weights. To illustrate the flexibility of our approach, we explore *simultaneous adaptation* to intrinsic dimension and smoothness in three different settings: nonparametric regression with compositional structures, geometric data through the control of Minkowski's dimension and anisotropic Besov spaces. For the last two, our method is the first to achieve simultaneous adaptation to dimension and smoothness to the best of our knowledge.

A given overparametrized architecture. In learning with deep neural networks, a key aspect is the choice of architecture $(L, \mathbf{r})$. In our approach, we choose a logarithmic depth: fix $\delta > 0$ and let

$$L := \lceil \log^{1+\delta} n \rceil. \tag{11}$$

Also, we fix beforehand a (relatively) large common width for network layers by setting

$$\mathbf{r}(\sqrt{n}) := (d, \lceil \sqrt{n} \rceil, \ldots, \lceil \sqrt{n} \rceil, 1), \tag{12}$$

with $\mathbf{r}(\sqrt{n})$ of size $L + 2$. Below we will also consider a slightly larger choice of width

$$\mathbf{r}(n) := (d, n, \ldots, n, 1), \tag{13}$$

with $\mathbf{r}(n)$ again a line vector of size $L + 2$. More generally, it is easily seen from the proofs of the results below that any choice of widths r_l of polynomial order $r_l \asymp n^{a_l}$, for *arbitrary* given $a_l \geq 1/2$ is compatible with the obtained rates, naturally placing the prior in the *overparametrized* regime (and already so for the smallest possible choice $a_l = 1/2$). Since one main choice of prior on weights below places the same distribution on each weight, this

means that the prior places no special 'penalty' term for picking more coefficients.

Choice of scaling factors σ_k. We consider two main choices of scaling parameters σ_k for the heavy-tailed prior on coefficients (7). The main choice discussed in the paper, and referred to as *constant* $\sigma_k s$ (although the constant depends on n, but not on k) is, for $\delta > 0$ as in (11) and any $k = 1, \ldots, T$,

$$\sigma_k = e^{-(\log n)^{2(1+\delta)}}, \tag{14}$$

or equivalently $\log(1/\sigma_k) = \log^{2(1+\delta)} n$, which is independent of k.

We next consider a more general choice that includes (14) as a special case, and referred to as *directed* $\sigma_k s$. It involves a little more notation, as the network weights may now depend on their relative positions in the network (recall their ordering is fixed though arbitrary) so may be skipped at first read. Recall from the notation as in Section 2.2 that we interpret shift vectors v_l as the 0-th columns of the weight matrices W_l. Suppose we work with an architecture $(L, \mathbf{r})$, so that the network weights (θ_k) are in one–to–one correspondance with coefficients $W_l^{(ij)}$ for $l \in [L+1]$, $i \in [r_l]$ and $j \in \{0, 1, \ldots, r_{l-1}\}$. We set, whenever θ_k corresponds to $W_l^{(ij)}$,

$$\sigma_k = \sigma_l^{(ij)}, \tag{15}$$

where for any admissible l, i, j as above, and $\delta > 0$ as in (11),

$$\log^{2(1+\delta)}(i \vee j) \leq \log(1/\sigma_l^{(ij)}) \leq \log^{2(1+\delta)} n. \tag{16}$$

The choice (14) corresponds to taking the second inequality in (16) to be an equality. If one takes instead equality in the first inequality in (16), one obtains a (double-index) version of the weights considered in Agapiou and Castillo (2024) (who set $\sigma_k = \exp(-\log^2 k)$ with one index k). The later choice enables one, with higher prior probability, to have larger values on a number of weights (those with small k), which may be useful in practice to avoid too many very small weights. We refer to Section 4 for more discussion on the choice of weights.

In the sequel, for $\kappa \geq 0$ the parameter in condition (H2) on the prior density ($\kappa = 0$ for polynomial tails) and $\delta > 0$ as in (11), we denote

$$\gamma = 2(1+\delta)(1+\kappa) + 1. \tag{17}$$

3.1 Adaptive contraction rate in regression

For a class of compositions $\mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$ and γ as in (17), let us introduce the rate

$$\phi_n := \max_{0 \leq i \leq q} \left(\frac{\log^\gamma n}{n} \right)^{\frac{\beta_i^*}{2\beta_i^* + t_i}}. \tag{18}$$

This coincides up to the logarithmic factor with the optimal rate ϕ_n^* in the minimax sense as in (3) for estimating a f_0 in the regression model if it is a composition as in the class $\mathcal{G}$.

Theorem 2 *Consider data from the nonparametric random design regression model* (1), *with $\tau_0 = 1$, $f_0 \in \mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$ and arbitrary unknown parameters. Let Π be a heavy-tailed DNN prior as described in Section 2.3 with*

- *architecture $(L, \mathbf{r})$ with L as in (11) and $\mathbf{r} = \mathbf{r}(\sqrt{n})$ as in (12);*

- *network weights $\theta_k = \sigma_k \zeta_k$ independent as in (7), with*

  - *scaling factors σ_k verifying either (14), or more generally (16);*
  - *random ζ_k with heavy-tailed density h satisfying (H1)–(H3).*

Then for any $\alpha \in (0,1)$, for $M > 0$ large enough, D_α the α–Rényi divergence, as $n \to \infty$,

$$E_{f_0} \Pi_\alpha \left[\{ f \, : \, D_\alpha(f, f_0) \geq M\phi_n^2 \} \mid X, Y \right] \to 0.$$

Let us now formulate a corollary in terms of the squared integrated loss. For $B > 0$, let $C_B(\cdot)$ be the 'clipping' function $C_B(x) = (-B) \vee (x \wedge B)$. Define a clipped tempered posterior as $\Pi_\alpha^B[\cdot|X,Y] = \Pi_\alpha[\cdot|X,Y] \circ C_B^{-1}$. Equivalently, if $f \sim \Pi_\alpha[\cdot|X,Y]$, then

$$g = C_B(f) \sim \Pi_\alpha^B[\cdot|X,Y].$$

The next Corollary shows that the rate above in terms of Rényi divergence carries over to the $L^2(P_X)$ loss as long as one works with a clipped posterior, which only requires to assume that an upper bound on $\|f_0\|_\infty$ is known, which is often assumed in theory for deep learning methods (Corollary 3 itself is a direct application of Lemma 21 below). In the sequel, for simplicity we always state results in terms of D_α, but similarly results in the $L^2(P_X)$ loss can be derived as in the next Corollary.

Corollary 3 *Under the setting and conditions of Theorem 2, and $\Pi_\alpha^B[\cdot|X,Y]$ as defined in the last display, for any $B > \|f_0\|_\infty$, any $\alpha \in (0,1)$, as $n \to \infty$,*

$$E_{f_0} \Pi_\alpha^B \left[\{ f \, : \, \|f - f_0\|_{L^2(P_X)} \geq M\phi_n \} \mid X, Y \right] \to 0.$$

Theorem 2 shows that, when performing a re-scaling of every coefficient of wide structured deep neural networks, by a well-chosen factor σ_k, the heavy-tailed prior leads to nearly minimax adaptive concentration of the tempered posterior regarding both the smoothness and the hidden compositional structure of the true function f_0. Indeed, neither the smoothness parameters (β_i) nor the intrinsic dimensions (t_i) were used to define the prior Π. Theorem 7 below shows that a similar prior also leads to adaptation to low-dimensional structure on the input data (X_i).

Theory of deep neural network approximation (see for example Kohler and Langer (2021) or Lu et al. (2021)) suggests that, given a true regression function f_0 in a model displaying low-dimensional structure (for example compositional structure on f_0) there is a 'minimal' approximating network structure $(L^*(\boldsymbol{\beta}, \mathbf{t}), \mathbf{r}^*(\boldsymbol{\beta}, \mathbf{t}))$ in terms of depth and width. To achieve adaptation one could recover this structure via model selection with a hierarchical prior (see for example Polson and Ročková (2018), Kong et al. (2023), Ohn and Lin (2024)). Here instead we show in Theorem 2 (resp. Theorem 7 below) that with our heavy-tailed prior no hierarchical step is required to achieve adaptation and it is sufficient to take a fixed ('overparametrized') network architecture, independent of $\boldsymbol{\beta}$ and $\mathbf{t}$ (resp. independent of β and t^* in the result below) and draw independent coefficients. We refer to Section 3.5 for more discussion on the choice of the architecture.

We allow for some flexibility in the choice of the decay coefficients σ_k. One can for example chose the same decay $\sigma_k = \exp(-\log^{2(1+\delta)} n)$ for every coefficient so that the θ_k's are independent and identically distributed. We refer to Section 4 for a brief discussion on other possible choices.

Remark 4 *a) The rate ϕ_n includes a logarithmic factor which we did not try to optimise; one can make the power γ slightly smaller by taking a slightly different depth, albeit with a slightly different choice of σ_k's. For simplicity we do not pursue such refinements here. b) Recall that we have assumed that an upper bound M_0 on f_0 is known: this is only used through the fact that we take a 'clipping' function to state Corollary 3, where B has to be larger than M_0. If this is not assumed, then the conclusion of Theorem 2 still holds in α–Rényi divergence.*

3.2 Contraction of mean-field variational approximation

Let us now consider taking a variational approach, and instead of sampling from the tempered posterior distribution $\Pi_\alpha[\cdot \,|\, X, Y]$, sample from the variational approximation $\hat{Q}_\alpha$ defined in (8), where the variational class is mean-field with heavy-tailed components.

Theorem 5 *Consider data from the nonparametric random design regression model (1), with $\tau_0 = 1$, $f_0 \in \mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$ and arbitrary unknown parameters. Let Π be a heavy-tailed DNN prior as in Theorem 2, with heavy-tailed density h such that $m_{2\vee(1+\kappa)}(h) \leq c_3$ for some $c_3 > 0$ and κ as in (H2). Suppose the scaling factors σ_k verify (14).*

Let $\mathcal{S} = \mathcal{S}_{HT}(h)$ be the mean-field class generated from h as in Definition 1 and let $\hat{Q}_\alpha$ be the associated variational tempered posterior as in (8).

Then for any $\alpha \in (0, 1)$, for ϕ_n as in (18), for M large enough, as $n \to \infty$,

$$E_{f_0} \int D_\alpha(f, f_0) d\hat{Q}_\alpha(f) \leq M\phi_n^2.$$

In particular, for any diverging sequence $M_n \to \infty$

$$E_{f_0} \hat{Q}_\alpha \left[\left\{ f \,:\, D_\alpha(f, f_0) \geq M_n \phi_n^2 \right\} \right] \to 0.$$

Theorem 5 establishes contraction of the heavy-tailed mean-field variational approximation with the same adaptive rate as in Theorem 2. The definition (1) of the variational class is independent of any unknown quantities linked to f_0 such as t and $\boldsymbol{\beta}$, has no hierarchical structure and only requires two parameters (m, ς), avoiding the choice of a sparsity level hyperparameter required in model selection priors such as spike and slab priors (as in e.g. Chérief-Abdellatif (2020); Bai et al. (2020)). The moment restriction on h allows for a result in expectation at the level of the variational posterior and guarantees consistency of the variational posterior mean (see also Remark 19 below), restricting however the choice of the prior to fixed decay as in (14). Up to somewhat more technical proofs, we believe that one could also weaken the conditions to derive a result in probability as above but without moment condition on h, as well as to allow for scalings verifying the more general decay (16). For simplicity here we do not investigate such further refinements.

Similarly as for Theorem 2, one may deduce from Theorem 5 a convergence rate result for the variational tempered posterior $\hat{Q}_\alpha$ in terms of the $L^2(P_X)$ loss by post-processing $\hat{Q}_\alpha$ through a simple 'clipping' operation.

3.3 Geometric data

Rather than making assumptions about the structure of the unknown regression function f_0 (e.g. in terms of compositions as in Sections 3.1 and 3.2), we can instead permit the observed data itself to exhibit a low-dimensional structure.

For instance, one can assume that the distribution P_X of the design points (X_i) is supported on a smooth manifold of lower dimension than d. This notion is frequently employed to describe intrinsic dimensionality and various results are available to approximate functions on such manifolds using deep neural networks, see for example Chen et al. (2019), Schmidt-Hieber (2019) and Fang et al. (2024). Here we follow the more general setting of Nakada and Imaizumi (2020), which includes design points sitting on a manifold as a special case while also allowing for rougher sets (such as fractals, although for the discussion below we focus on a manifold example for simplicity) via the following notion of dimension.

Definition 6 (Minkowski Dimension) *Let $E \subset [0,1]^d$, the (ε, ∞)-covering number of E, written $\mathcal{N}(E, \varepsilon)$ is the minimal number of ℓ_∞-balls of radius $\varepsilon > 0$ necessary to cover E. The upper Minkowski dimension of E is defined as*

$$\dim_M E := \inf\{t \geq 0 \,:\, \limsup_{\varepsilon \downarrow 0} \mathcal{N}(E, \varepsilon)\varepsilon^t = 0\}.$$

The Minkowski dimension gives insight on how the covering number of E changes compared to the decrease of the radius of the covering balls.

Assumption 1 *Suppose that $t^* := \dim_M \operatorname{Supp} P_X < d$.*

Assumption 1 is satisfied in the case that $\operatorname{Supp} P_X$ is a smooth compact manifold of dimension d^* strictly less than d, as then $t^* \leq d^* < d$, as established in Lemma 9 in Nakada and Imaizumi (2020).

We show below that, under Assumption 1, our construction of heavy-tailed priors can be adapted to lead to hierarchical-free adaptation to the Minkowski dimension of the support of the design points (X_i). Let us set, for κ the constant appearing in (H2) and $t > 0$,

$$\varepsilon_n(t) := \left(n/\log^{2(1+\kappa)+1} n\right)^{-\beta/(2\beta+t)}. \tag{19}$$

Theorem 7 *Consider data from the nonparametric random design regression model* (1), *with $\tau_0 = 1$, $f_0 \in \mathcal{C}_d^\beta([0,1]^d)$. Suppose Assumption 1 holds and let Π be a heavy-tailed DNN prior as described in Section 2.3 with*

- *architecture $(L, \mathbf{r})$ with $L = \lceil \log n \rceil$ and $\mathbf{r} = \mathbf{r}(n)$ as in* (13);

- *network weights $\theta_k = \sigma_k \zeta_k$ independently as in* (7), *with*

  - *scaling factors σ_k verifying* (14);

  - *random ζ_k with heavy-tailed density h satisfying (H1)–(H3).*

14

Let $\varepsilon_n(t)$ as in (19), for a given arbitrary $t > t^*$. Then for any $\alpha \in (0,1)$, for M large enough, as $n \to \infty$,

$$E_{f_0}\Pi_\alpha\left[\{f \,:\, D_\alpha(f, f_0) \geq M\varepsilon_n(t)\} \mid X, Y\right] \to 0.$$

Assume further that $m_{2\vee(1+\kappa)}(h) < \infty$ and let $\hat{Q}_\alpha$ be the heavy-tailed tempered variational approximation defined by (8) on $\mathcal{S}_{HT}(h)$. Then the contraction of $\hat{Q}_\alpha$ as stated in Theorem 5 holds here with ϕ_n replaced by ε_n.

As a special case, if Supp P_X is a smooth compact manifold of dimension d^* strictly less than d, Theorem 7 gives a contraction rate $\varepsilon_n(t)$ in particular for any $t > d^*$, that is a rate $\varepsilon_n(d^* + \eta)$ for any fixed arbitrarily small $\eta > 0$, which is near-optimal (the optimal minimax rate being $\varepsilon_n(d^*)$). The proof of Theorem 7 in given in Appendix C.3.

In Theorem 7, we use the choice $\mathbf{r} = \mathbf{r}(n)$ as in (13), that is, we increase the network width quadratically compared to (12). Indeed, in order to establish the above posterior concentration, we rely on an approximation result of functions in such a setting by ReLU-deep neural networks by Nakada and Imaizumi (2020). Such a result, recalled as Lemma 25 below, uses sparse approximations (i.e. many network weights are equal to zero), in a similar spirit as the corresponding result in Schmidt-Hieber (2020), which has the effect to increase (quadratically) the width of the network compared to Theorem 2 and Theorem 5, for which we used the recent fully-connected approximation results of Kohler and Langer (2021) (as opposed to Schmidt-Hieber (2020), Theorem 5). It should also be possible to use fully-connected approximation results in the present setting, for example in the spirit of Theorem 6.3 of Jiao et al. (2023) (which, similar to Kohler and Langer (2021), has at most polynomially increasing weights).

Also, here we have considered only the case of constant scaling factors (14). The same remarks as in the discussion below Theorem 5 apply: the statement can presumably be adapted to directed decay of the network weights up to using a slightly more technical proof.

A number of other geometrical settings of interest could also be considered within the framework of the present work. Among others, one could consider the case of design points not exactly supported *on* the manifold but rather in a neighborhood of it. While such extensions are very interesting, they fall beyond the scope of the present contribution. We refer to the Discussion in Section 4 for some comments on how to possibly extend our results to design points lying in a tubular neighborhood of an unknown manifold.

3.4 Anisotropic Besov classes

Instead of assuming the true function to be in a Hölder space as in Theorem 2, we now investigate the case that the smoothness of the true function is both depending on the location (non-homogeneous smoothness) and the direction (anisotropic). For simplicity of presentation we focus on the setting without additional compositional structure, but one could also derive results in case of compositions as in Section 3.1; the case of a single function is already illustrative of the adaptation properties at stake. In the following we recall a definition of anisotropic Besov spaces through the moduli of smoothness. For properties of these spaces and equivalent characterizations (for example through multiresolution analysis) we refer to Giné and Nickl (2015), Section 4.3. and to the monograph Triebel (1983).

For a function $f : \mathbb{R}^d \to \mathbb{R}$, the m-th difference of f in the direction $v \in \mathbb{R}^d$ is

$$\Delta_v^m(f)(x) := \Delta_v^{m-1}(f)(x+v) - \Delta_v^{m-1}(f)(x), \qquad \Delta_v^0(f)(x) := f(x).$$

Definition 8 *For a function $f \in L^p([0,1]^d)$, where $p \in [1, \infty)$, the m-th modulus of smoothness of f is defined by*

$$\omega_{m,p}(f, u) := \sup_{v \in \mathbb{R}^d \,:\, |v_i| \le u_i} \|\Delta_v^m(f)\|_p, \qquad for \ u = (u_1, \ldots, u_d), \ \ u_i > 0.$$

Definition 9 (Anisotropic Besov space) *Let $p \in [1, \infty)$, $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}_{>0}^d$ and $m := \max_i \lfloor \beta_i \rfloor + 1$. Let the seminorm $|\cdot|_{B_{pp}^\beta}$ be defined as*

$$|f|_{B_{pp}^\beta}^p := \sum_{k=0}^{\infty} \left[2^k w_{m,p} \left(f, (2^{-k/\beta_1}, \ldots, 2^{-k/\beta_d}) \right) \right]^p.$$

For $\|f\|_{B_{pp}^\beta} = \|f\|_p + |f|_{B_{pp}^\beta}$ and $M > 0$, the anisotropic Besov ball of radius M on $[0,1]^d$ is

$$B_{pp}^\beta(M) := \left\{ f \in L^p([0,1]^d) \ : \ \|f\|_{B_{p,p}^\beta} \le M \right\}.$$

Optimal minimax rates on such anisotropic Besov balls can be expressed in terms of the harmonic mean smoothness, defined as

$$\tilde{\beta}^{-1} := \sum_{k=1}^{d} \beta_k^{-1}. \tag{20}$$

Let us set, for γ the constant given by (17) and $\tilde{\beta}$ defined by (20),

$$\epsilon_n = (n / \log^\gamma n)^{-\tilde{\beta}/(2\tilde{\beta}+1)}. \tag{21}$$

Theorem 10 *Consider data from the nonparametric random design regression model (1), with $\tau_0 = 1$, $f_0 \in B_{pp}^\beta(1)$ and $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}_{>0}^d$. Assume $p < 2$ and $\tilde{\beta} > 1/p$, and let Π be a heavy-tailed DNN prior as described in Section 2.3 with*

- *architecture $(L, \mathbf{r})$ with L as in (11) and $\mathbf{r} = \mathbf{r}(n)$ as in (13);*

- *network weights $\theta_k = \sigma_k \zeta_k$ independently as in (7), with*

  - *scaling factors σ_k verifying (14);*
  - *random ζ_k with heavy-tailed density h satisfying (H1)–(H3).*

Then for any $\alpha \in (0,1)$, for $M > 0$ large enough it holds, as $n \to \infty$, for ϵ_n as in (21)

$$E_{f_0} \Pi_\alpha \left[\{ f \ : \ D_\alpha(f, f_0) \ge M \epsilon_n \} \mid X, Y \right] \to 0.$$

Assume further that $m_{2 \vee (1+\kappa)}(h) < \infty$ and let $\hat{Q}_\alpha$ be the heavy-tailed tempered variational approximation defined by (8) on $\mathcal{S}_{HT}(h)$. Then the contraction of $\hat{Q}_\alpha$ as stated in Theorem 5 holds here with ϕ_n replaced by ϵ_n.

Whenever $p < 2$ and $\tilde{\beta} > 1/p$, functions in B_{pp}^{β} are continuous and in-homogeneously smooth in terms of the L^2–norm (see for instance Suzuki and Nitanda (2021) and Donoho and Johnstone (1998)). An interesting feature of Theorem 10 is the adaptation to the anisotropic smoothness it provides. Anisotropy itself may help mitigate the curse of dimensionality. Indeed, when $\beta = (\beta_0, \ldots, \beta_0)$, we have $\tilde{\beta} = \beta_0/d$, and one obtains the usual rate for a homogeneously smooth function in $\mathbb{R}^d$. However, when f_0 is not very smooth only in a small number of directions, the estimation rate $n^{-\tilde{\beta}/(2\tilde{\beta}+1)}$ depends only on this specific small number of coordinates rather than on the whole ambient dimension d.

The proof of Theorem 10 can be found in appendix C.4. Similar remarks as for Theorem 7 can be made. In particular, . Also, the proof uses the sparse approximation of Suzuki and Nitanda (2021) recalled as Lemma 26 below, which again explains why it is natural to use the width $\mathbf{r} = \mathbf{r}(n)$ as in (13).

3.5 Influence of the network architecture

In the next lines, we discuss possible choices of the architecture. We consider in particular our proposed choices (that is, (12) or (13)), but also contrast these with different ones, examining the impact on rates and optimality. For simplicity we discuss the simplest setting of estimating a β–Hölder function (without the compositional structure), first in dimension 1 (see the end of the Section for the case of dimension d), that is, on $[0, 1]$, for which the minimax rate in squared integrated norm is of order $\varepsilon_n^2 = n^{-2\beta/(2\beta+1)}$.

Deterministic Architecture. In this work, we have chosen a prior with a fixed (deterministic) width and depth. This choice aligns with common practice for deep neural networks where practitioners typically select a fixed architecture prior to learning network weights (e.g. by gradient descent or a variational optimisation algorithm). Among Bayesian approaches so far for deep neural networks, those that have claimed adaptation with respect to the smoothness parameter $\beta > 0$ have involved sampling either the nonzero coefficients, or the architecture itself, from a hyper-prior: this is the case in Polson and Ročková (2018) where the prior is a spike-and-slab distribution that sets to zero a number of coefficients at random and very recently Kong and Kim (2024), where the width in the architecture is random. A key insight from our results is that the deterministic (and overparametrized) choice of architecture, where all networks coefficients are non-zero under the prior, is not detrimental to optimality and statistical adaptation properties.

Moderate depth. In the present work, we have restricted the depth of the network to grow logarithmically with the sample size n. This restriction is primarily technical, since, as we explain in more details now, our arguments (as well as those used so far in the theoretical study of Bayesian DNNs, to the best of our knowledge) are presently difficult to extend to the 'very deep' case. A crucial challenge in posterior contraction analysis (even in the non-tempered case, $\alpha = 1$) is verifying the so-called 'mass condition' (Lemma 17), which ensures the prior places enough probability mass near the true function. In the present setting, this is achieved using the sufficient condition in Lemma 24. This result quantifies how much two network realisations are far appart if we control individual distances between weights. Using this 'error propagation' Lemma requires controlling the distance between

the 'true' coefficients (of a neural network approximation of f_0) and those sampled from the prior, uniformly over the network (in particular independently of the depth). Notably, the required precision for this control increases polynomially with width (for fixed depth) and exponentially with depth (for fixed width). This makes analyzing 'very deep' priors significantly more challenging, as it would require having a very precise prior mass control for coefficients very deep in the network. We believe that such a result is beyond the scope of the present paper as it would involve arguments specifically tied to the study of very deep neural networks, but an interesting direction for future work (in particular the understanding of depth-dependence one should put on the decay coefficients when randomly initializing a neural network is of significant interest and has been studied for instance, in the case of very-deep Res-Nets in a slightly different context by Marion et al. (2022)).

Choice of the width. Having fixed the network depth L to a slowly growing logarithmic term as in (11), the only degree of freedom left on the architecture is now the width of the network (the activation function is in principle another degree of freedom; activations beyond ReLU are discussed in Section 3.6 below). The fully-connected approximation result from Kohler and Langer (2021), recalled in Proposition 33 below, implies that for a logarithmic depth L as above, there is an oracle width $r^* := \mathbf{r}(N^*)$ such that networks with architecture (L, r^*) approximate β–Hölder functions with the desired target accuracy ε_n, and the corresponding width N^* is proportional to $n^{1/\{2(2\beta+1)\}}$ (recall we have assumed here that $d = 1$). We say that a neural network architecture is underparametrized (resp. overparamterized) if its width parameter is smaller (resp. greater) than r^*.

overparametrized vs underparametrized architecture. Consider a prior construction where the network width is deterministic and set to n^a for some $a > 0$, leading to an architecture $(L, \mathbf{r}(n^a))$. There are two cases

1. $a = 1/2$ (our choice) or $a > 1/2$. As shown in Theorem 2, when $a = 1/2$, the heavy-tailed prior with architecture $(L, \mathbf{r}(n^a))$ achieves the minimax adaptive posterior contraction rate. It is not hard to check that the result of Theorem 2 continues to go through larger values of a, that is $a \geq 1/2$. For $a \geq 1/2$, we are always in the *overparametrized* regime since $n^a \gg N^* = n^{1/\{2(2\beta+1)\}}$ for any possible smoothness $\beta > 0$.

2. $a < 1/2$. Then there exists a regularity index $\beta_{min} > 0$ for which $a < 1/(2(2\beta_{min}+1))$ by definition and in that case the architecture is *undersmoothing*. In this case, it is possible to show that approximation by a ReLU-DNN with architecture $(L, \mathbf{r}(n^a))$ cannot be fast enough to reach the desired optimal estimation rate ε_n for f_0. Indeed, applying Lemma 1 from Schmidt-Hieber (2020) (a lower bound in the L^2-norm is also available, as noted in the remark therein), one obtains the following lower bound on the approximation rate

$$\sup_{f_0 \in \mathcal{C}_d^{\beta_{min}}([0,1])} \inf_{f \in \mathcal{F}(L, \mathbf{r}(n^a))} \|f - f_0\|_\infty \gtrsim \eta_n, \tag{22}$$

where $\eta_n \approx n^{-2a\beta_{min}}$, up to logarithmic factors. Since $a < 1/(2(2\beta_{min}+1))$, this rate is strictly slower than the minimax rate $n^{-\beta_{min}/(2\beta_{min}+1)}$ for that regularity. The proof of Theorem 2 would then give a suboptimal contraction rate.

Our choice: an overparametrized architecture yet leading to optimal rates. In conclusion, the smallest deterministic width for the architecture (given L has been chosen as above), for which one still achieves statistical adaptation to smoothness, is the overparametrized choice $n^{1/2}$ corresponding to $a = 1/2$ (our choice in Theorem 2). Finally, in the case of estimation on $[0,1]^d$, the oracle width within the setting of Kohler and Langer (2021) is $n^{d/\{2(2\beta+d)\}}$, so once again the overparametrized width $n^{1/2}$ is the smallest one that still verifies $n^{1/2} \geq n^{d/\{2(2\beta+d)\}}$ for any value of $\beta > 0, d \geq 1$.

3.6 Extension to other activation functions

All results so far have been stated in the case where the activation ρ in (4) is the ReLU function $x \mapsto \max(0,x)$. We now see how to extend our main first result, Theorem 2, to other, sigmoid-type, smooth, activation functions. Here we state this extension in a slightly informal way, and refer the reader to the Appendix Section C.6 for a formal statement that adapts Theorem 2 to other activation functions (Theorem 32 therein). One can easily follow the steps developed in Section C.6 to also adapt Theorems 7 and 10. Within the proof of Theorem 2, only two properties are required from the activation function ρ:

- a result similar to Lemma 24 should hold; the latter gives a control of how errors on individual weights 'propagate' into an error at the level of the function realised by the neural network;

- there should exist a wide neural network having ρ as activation function, such that the neural network approximates the function f_0 with the correct rate and such that the weights of the network are bounded 'sub-exponentially' in that the magnitude of the weights is at most $\exp(c_1 \log^{c_2} n))$ for some positive constants c_1, c_2.

Note that instead of requiring a polynomial control of the weights in the second point above, one allows for a slightly larger growth in n, so that the condition is significantly milder, a fact we crucially exploit in our proof for ρ-activations.

Let us now comment on how to check these two properties. Regarding the first point, it can be checked that if the activation ρ is Lipschitz-continuous, Lemma 24 still holds up to a constant (see Lemma 29). To ensure the second required property on ρ, one could prove an approximation result for DNNs with activation ρ as long as such result provides an appropriate control on the magnitude of the weights of the approximating network (this is the approach we take in Appendix D to prove Theorem 2). We propose here another approach that, under some smoothness conditions on ρ, enables to start from a ReLU approximation with polynomially bounded weights to obtain the required approximation with activation ρ.

Definition 11 *A function $\rho : \mathbb{R} \to \mathbb{R}$ is said to be admissible if*

- *ρ is non-decreasing, bounded and Lipschitz-continuous,*

- *ρ is three times continuously differentiable, with bounded derivatives and there exists a point $x_\rho \in \mathbb{R}$ such that $\rho'(x_\rho) \neq 0$ and $\rho''(x_\rho) \neq 0$,*

19

- *there exists two real numbers $a < b$, and a constant $c_\rho > 0$, such that,*

$$\sup_{x \in \mathbb{R}} |x\rho(x) - \max(ax, bx)| \leq c_\rho.$$

The smoothness conditions in Definition 11 ensure that ρ is a sigmoid-type function. In particular the third condition states that ρ should converge asymptotically at least as fast as $1/|x|$ towards the horizontal line $\{y = a\}$ (resp. $\{y = b\}$) when $x \to -\infty$ (resp. $y \to +\infty$). These conditions are satisfied by most sigmoid-type activations commonly used in deep-learning, for instance the logistic activation $x \mapsto (1 + e^{-x})^{-1}$ satisfies the conditions with $a = 0$ and $b = 1$. Both the hyperbolic tangent (tanh) and the error (erf) activation functions satisfy these conditions with $a = -1$ and $b = 1$.

Claim. If ρ is admissible according to Definition 11, every previous Theorem still holds with ρ in place of the ReLU activation.

Let us now give a brief intuition of the proof, referring to Appendix C.6 and Theorem 32 therein for a formal statement and proof. Noticing that for $a < b$ and any $x \in \mathbb{R}$, we have

$$\text{ReLU}(x) = \frac{1}{b - a}(\max(ax, bx) - ax),$$

If ρ is admissible results from Kohler et al. (2022) show that the ReLU function can be approximated by a ρ-DNN with constant width and depth and polynomially bounded weights. From this it follows (Lemma 31) that any ReLU-DNN approximation results with logarithmic depth transposes to a ρ-DNN result with same architecture and sub-exponential weights.

Let us emphasize that, in the current proof, going from a ReLU result to an admissible ρ result requires one to be able to work with weights going to infinity with n at a possibly slightly faster than polynomial rate (but slower than exponential). This shows that our heavy-tailed priors are also suited for a wide range of sigmoidal activation functions, which might be of interest in practice. Indeed, as we have mentioned, and as opposed to priors requiring hyper-priors on the architecture, our priors with fixed architecture are suitable for both MCMC and variational procedures; even though ReLU activations are popular in methods involving backpropagation because of the gradient being essentially a Boolean (therefore very fast to compute), in MCMC algorithms involving more complex dynamics (for example HMC Neal (2011)) it might be of interest to work with a smoother activation function so that the target log-density is smooth, as suggested in Dinh et al. (2024).

3.7 Extension to unknown noise level τ_0

So far in model (1), we have taken the noise variance τ_0^2 to be equal to 1; however in practice this quantity may be unknown. If τ_0 is unknown, let us write P_{f_0,τ_0} for the distribution of (X_1, Y_1) under model (1), the parameters being now $(f, \tau^2) \in \mathcal{F} \times \mathbb{R}_{>0}$. We put a prior Π on (f, τ^2) of the form $\Pi = \Pi_f \otimes \pi_{\tau^2}$ where Π_f is a heavy-tailed DNN prior as before and π_{τ^2} is a distribution on $\mathbb{R}_{>0}$ with a density bounded away from 0 on a neighborhood of τ_0:

$$d\pi_{\tau^2}(x) = q(x)dx, \qquad \inf_{x \in [\tau_0^2 - \epsilon, \tau_0^2 + \epsilon]} q(x) \geq r_0, \tag{23}$$

for some $\epsilon > 0$ and $r_0 > 0$, which is satisfied by standard choices of priors on variance parameters, such as exponential or inverse-gamma distributions with fixed parameters.

Contraction of the tempered posterior as stated in Theorems (2, 7 and 10) still holds true when τ_0 is unknown as long as one chooses a prior on τ^2 satisfying Condition (23). We only state below the generalisation of Theorem 2, the same holds true for the other Theorems.

Theorem 12 *Consider data from the nonparametric random design regression model* (1), *with unknown* $\tau_0 > 0$ *and* $f_0 \in \mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$ *for arbitrary unknown parameters. Let* $\Pi = \Pi_f \otimes \pi_{\tau^2}$ *be a prior on* (f, τ^2) *where* Π_f *is the heavy-tailed DNN prior as in Theorem 2 and* π_{τ^2} *is a prior on* $\mathbb{R}_{>0}$ *satisfying condition* (23). *Recall that* ϕ_n *is the rate as in* (18). *Then for any* $\alpha \in (0, 1)$, *for* $M > 0$ *large enough, as* $n \to \infty$,

$$E_{f_0}\Pi_\alpha\left[\left\{(f, \tau^2) : D_\alpha(P_{f,\tau^2}, P_{f_0,\tau_0^2}) \leq M\phi_n^2\right\} \mid X, Y\right] \to 1.$$

Furthermore, assuming $\|f_0\|_\infty \leq B$, *the clipped posterior* $\Pi_\alpha^B[\cdot|X, Y]$ *satisfies, as* $n \to \infty$,

$$E_{f_0}\Pi_\alpha^B\left[\left\{(f, \tau^2) : \|f - f_0\|_{L^2(P_X)} \leq M\phi_n, |\tau^2 - \tau_0^2| \leq M\phi_n\right\} \mid X, Y\right] \to 1.$$

The proof of Theorem 12 can be found in Appendix C.5. We note that the second part of the statement entails posterior contraction around the unknown variance parameter τ_0^2.

3.8 Extension to standard posteriors ($\alpha = 1$)

For technical reasons we have so far focused on fractional posterior distributions with $\alpha < 1$. This enables one to focus on prior mass conditions, which is particularly handy when dealing with heavy-tailed priors. Although one can conjecture (this has been shown to be true for heavy-tailed series prior in the technically easier Gaussian white noise regression model in Agapiou and Castillo (2024)) that all previous results continue to go through under exactly the same prior and conditions, this seems technically non-trivial and presently hard to reach with currently available techniques, see the discussion in Section 4 for more on this.

However, one considers an augmented prior that also models the variance of the noise, which one would do anyways under unknown noise level as in the previous Section 3.7, then one can use a recent idea from Castillo and Randrianarisoa (2024) (who took inspiration from an idea of Mai (2024), see also the references therein, in a different high-dimensional context) who show that, in the present random design regression model, augmenting a given prior on f with a *well-chosen* prior on τ^2 enables one to transfer an already obtained contraction rate for a fractional posterior on f with $\alpha < 1$ (such as ones obtained above) to a contraction rate for the standard posterior ($\alpha = 1$) on f.

As in the previous Section 3.7, we write P_{f_0,τ_0} for the distribution of (X_1, Y_1) under model (1), the parameters being $(f, \tau^2) \in \mathcal{F} \times \mathbb{R}_{>0}$. We put a prior Π on (f, τ^2) of the form $\Pi = \Pi_f \otimes \pi_{\tau^2}$ where Π_f is a heavy-tailed DNN prior as before and $\pi_{\tau^2} = \pi_{\tau^2,n}$ the following distribution on $\mathbb{R}_{>0}$, for some fixed $b \in (0, 1)$,

$$\pi_{\tau^2} = \text{Gamma}\left(\left\{\frac{1-b}{2}\right\}n + 1, b\right), \tag{24}$$

where $\text{Gamma}(a_1, a_2)$ 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 $\Pi[\cdot \,|\, (X, Y)]$ jointly on f and τ^2. Similar to the result on the clipped posterior in Corollary 3, let $\Pi^B[\cdot \,|\, (X, Y)]$ denote the clipped posterior (where the clipping function operates only on the f part, not on τ^2).

The next result shows that, if the conditions of Theorem 2 are satisfied for the prior Π_f, then the classical marginal posterior on f, that is $\Pi[f \in \cdot\,,\ \tau^2 \in \mathbb{R}_{>0} \,|\, (X, Y)]$ also contracts at the same rate as the one from Theorem 2.

Theorem 13 *Let Π be a prior on (f, τ^2) of product form $\Pi_f \otimes \pi_{\tau^2}$, with π_{σ^2} given by (24). Suppose the conditions of Theorem 2 are satisfied. Then for $\Pi^B[\cdot|X, Y]$ the clipped posterior as defined above, for any $B > \|f_0\|_\infty$,*

$$E_{f_0} \Pi^B \left[\{ f \,:\, \|f - f_0\|_{L^2(P_X)} \geq M\phi_n \} \,|\, X, Y \right] \to 0.$$

That is, the conclusion of Corollary 3 holds for the clipped marginal in f of the classical posterior $\Pi[\cdot \,|\, (X, Y)]$ $(\alpha = 1)$.

Let us note that this result also automatically handles the more practical setting where, as in the last subsection, the noise variance τ_0^2 is unknown.

4. Discussion

A main key take-away from this work is that putting suitable heavy-tailed distribution on weights of deep neural networks enable automatic and simultaneous adaptation to intrinsic dimension and smoothness. This prior distribution allows for a *soft selection* of relevant weights in the network, that get higher values under the posterior, while less important weights can get very small values. The approach does not use a 'hard' variable selection approach, which would either set some weights to zero, and hence require the need to sample from the posterior distribution on the support points, or attempt to sample the network architecture at random, which would require to sample from the posterior on $(L, \mathbf{r})$, thus requiring in both cases to sample from 'support' hyperparameters, which can be costly computationally.

Another key idea is that the approach allows for a wide variety of choice of *over-parametrized* architectures. We have seen, say in the setting of standard nonparametric regression to fix ideas, that any choice of a common width for the layers that is polynomial in n with a power at least $1/2$, that it $r_i \asymp n^a$ with $a > 1/2$ is compatible with optimal rates with our method. We note that we have not investigated here the case of very deep networks with polynomial depth, but the idea could in principle apply there too. We also allow for a fairly wide variety of scalings σ_k of the individual weights, showing that the optimality of the rates is robust to different choices of scalings.

4.1 Discussion: priors for Bayesian deep neural networks

Priors for Bayesian deep neural networks. For networks of large width (possibly increasing with the sample size, but finite) and moderate depth, a number of natural prior distributions

have been proposed in recent works. We provide now a quick overview of recent results and refer to Fortuin (2022) for a recent specific review on choices of priors for BNNs.

Let us first discuss some differences of our approach with another natural family of priors that directly induces sparsity. One approach directly applies sparsity to the weights by drawing exact zeros through spike-and-slab (SAS) priors, which were among the first to receive frequentist theoretical guarantees for full posteriors by Polson and Ročková (2018) and for their variational approximations by Chérief-Abdellatif (2020). A tractable variational Bayes (VB) algorithm was later proposed by Bai et al. (2020) to handle these priors, allowing for a continuous relaxation of the spike-and-slab and updating the variational parameters through stochastic gradient descent. Another type of sparse prior enforces sparsity at the node level, for instance via masking functions, as explored by Kong et al. (2023), who also provide an MCMC algorithm for this setting. One difference is that here we use the recent 'dense' approximation results of Kohler and Langer (2021) that enable us to be more parsimonious (if this is desired) in terms of the total number of parameters that should be sampled in the posterior. Contrary to our 'dense' approach, these sparse priors require some additional hyper-priors on the sparsity level, increasing computational costs. Updating the proofs from Polson and Ročková (2018), one could also deploy SAS priors with denser architecture but SAS priors still have more (hyper-)parameters overall, with the probability of being nonzero to be tuned as well as the slab parameter (the same remark applies for their variational counterparts). We also note that currently all these priors achieve adaptation through a form of hyper-prior on the architecture, which complicates the use of adaptive MCMC algorithms and may make them intractable as the width of the model grows. This challenge reinforces the need for more scalable (though only approximative) VB algorithms, where (in contrast to MCMC) adaptivity can be managed through sampling over different architectures, a process made significantly more efficient and parallelizable by recent techniques from Ohn and Lin (2024). Contrary to SAS, our heavy-tailed priors enable both sampling from the full posterior with usual MCMC methods (see the paragraph on simulations below) as well as the use of mean-field variational approximation.

Two other recent natural constructions are the priors considered in Lee and Lee (2022) and the ones in Kong and Kim (2024). The work Kong and Kim (2024) adapts, as we do, the approximation theory of Kohler and Langer (2021) to allow for dense (non-sparse) approximations, thus avoiding the need for forcing some weights to be zero – as a side note, let us mention that more or less simultaneously the need for a control on the amplitude of the weights for use in Bayesian arguments has been recognised by several authors that have adapted Theorem 2 in Kohler and Langer (2021) to include quantitative bounds on network coefficients; see for example Theorem K.1 in Ohn and Lin (2024), and Theorem 1 in Kong and Kim (2024) and Appendix D for our own construction –. However, an important difference of our work with the approach of Kong and Kim (2024) is that the latter requires to sample from the width of the network to achieve adaptation; as noted above, such hyper-parameter sampling can be costly for simulations, as it requires access to a good approximation of the posterior on the width (similar to the calibration of the cutoff of sieve priors in nonparametrics, see Agapiou and Castillo (2024) for more discussion on this); this hyper-parameter sampling is not required here, as we use a deterministic 'overparametrized' architecture.

23

The work Lee and Lee (2022) allows for heavy-tailed priors, such as mixture of Gaussians, although crucially the conditions of Theorem 3 therein are non-adaptive (i.e. the priors' parameters depend on the unknown smoothness of the regression function). A main insight of the present work, following the idea recently introduced in Agapiou and Castillo (2024) in nonparametric settings, is that by taking an ('overparametrized') width, heavy tails on weights make the (tempered) posterior automatically adapt to unknown structural parameters, let it be smoothness, intrinsic dimension(s) or compositional structures. The simulation aspect of the work Lee and Lee (2022) is also relevant here, since therein the authors develop and run MCMC algorithms that enable to sample from the corresponding neural network posteriors.

Finally, let us mention the work by Ghosh et al. (2019), who provide algorithms to sample from variational approximations of posteriors corresponding to horseshoe priors on weight parameters. Although the authors provide no theoretical back-up for their empirical results, this is related in spirit to our approach; in fact, although there are some notable differences in the prior's choice (e.g. the horseshoe has a pole at zero and has a tuning sparsity parameter) we think that our approach and proofs could presumably be adapted in order to provide theoretical understanding and validation of their method; this idea is also supported by current work in progress Agapiou et al. (2025), where we derive theory for horseshoe priors for nonparametrics.

4.2 Discussion: from Gaussian to Heavy-tailed weights

As we have seen, two main families of priors for Bayesian DNNs are 1) sparsity-inducing priors and 2) priors on fully-connected architectures. In this section, we focus on the latter: perhaps the most natural choice of prior on the weights seems to be an isotropic Gaussian.

We discuss two ways in which heavy-tailed priors naturally emerge in this context. First, in a classical frequentist setting, where Gaussian weights are used as initializations for finding an empirical risk minimizer via gradient descent, the weights tend to become increasingly heavy-tailed during training. Second, in a Bayesian framework, the output distribution of a neural network with isotropic Gaussian weights becomes progressively more heavy-tailed as the network depth increases.

Regarding the first point: starting from an isotropic Gaussian initialization of weights, empirical evidence indicates that when using MCMC, in order to improve the posterior performance, the likelihood has to be raised to a certain power $\alpha > 1$, this is the so-called "cold-posterior effect" (Wenzel et al. (2020)) suggesting that the Gaussian prior is misspecified (in the sense that it does not correctly represent the data). When training a feedforward neural network in a frequentist manner, such as with stochastic gradient descent, it has been observed (for instance by Fortuin et al. (2022)) that the weights become progressively more heavy-tailed during training. This suggests that an heavy-tailed prior would be less misspecified and indeed Fortuin et al. (2022) show that the influence of the "cold-posterior effect" tends to vanish for heavy-tailed priors such as Laplace and t-Student. These priors perform well in reconstruction tasks for any power α, whereas the Gaussian prior appears to require more tuning for this parameter. Note that we show in all our theorems that the heavy-tailed posterior contracts at minimax rate for any $\alpha < 1$; the

aforementioned experimental studies on "cold posteriors" suggest that for our heavy-tailed prior the choice of α has little effect on the predictive capabilities of the estimate, contrary to its Gaussian counterpart.

Now, turning to the second point: interestingly, heavy-tailed priors can heuristically be connected to the simple isotropic Gaussian prior. Indeed, it has been showed by Vladimirova et al. (2019) that while starting from an isotropic Gaussian prior, the marginal distribution of the post and pre-activations (the input and output of the neurons) become more and more heavy-tailed as the depth increases, which indicates that heavier-tailed distributions are a natural occurrence in neural-network priors. We refer to the review paper of Arbel et al. (2023) for more discussion and references on this.

In summary, this discussion reveals interesting links between isotropic Gaussian priors on weights and heavy-tailed ones; a better understanding, in particular from the theoretical point of view, is left for future work.

4.3 Discussion: algorithms.

Algorithms. Although an in-depth simulation study is beyond the scope of the present paper, we note that algorithms are readily available to sample from the heavy-tailed tempered posteriors introduced here. As mentioned above, since there are no structure hyper-parameters to sample from (such as the network architecture, or the position of the non-zero coefficients in the case of SAS priors), one can use an MCMC algorithm such as the one from Lee and Lee (2022). Another possibility is the use a variational Bayes algorithm for heavy tailed priors, such as the one considered in Ghosh et al. (2019) for the horseshoe prior. This is left for future work. We note that having both feasible MCMC and variational algorithms makes it an interesting setting. Indeed, it is known that mean-field VB, not being as complex as the posterior, may distorts certain aspects of the latter, such as posteriors for finite-dimensional functionals. On the other hand, unlike VB, the full posterior on network will retain correlations, which are believed to be particularly important for deep neural networks. It will be interesting to compare the behaviour of both algorithms through an extended simulation study; this will be considered elsewhere.

In principle possible, hyperparameter posterior sampling (such as posterior sampling from complexity-based priors with hyper-priors on the complexity) is often believed to be delicate in terms of mixing of MCMC (a classical reference discussing this is, for example, Brooks et al. (2003)), and one may heuristically think that such hyper-sampling, on top of possibly begin time and/or ressource consuming, may add some 'noise' to the MCMC output. One may argue that any prior distribution, including a prior on the complexity for instance, has a 'constant' to be tuned, but we believe that it is fair to say that any prior with 'one level of hierarchy less' is (at least philosophically) computationally simpler.

Although MCMC sampling is made easier with a fixed architecture, it is expected that for very large networks, using current methods, MCMC sampling will become computationally intractable, the posterior being a multi-modal distribution of very high dimensionality. Recent techniques to accelerate sampling using MCMC for BNNs have been obtained by Hron et al. (2022) and Pezzetti et al. (2025). As the size of the architecture grows, a practician might consider using VB to obtain a computable approximation of the posterior, perhaps at the cost of the quality of uncertainty quantification, another choice would be for example to

use a Laplace approximation, see e.g. Arbel et al. (2023) for more references on this aspect. Although our results provide theoretical support for a Heavy-tailed mean-field VB approximation it is the case that most practical algorithms are developed with Gaussian priors in mind and a study of efficient VB algorithms for more general distributions is an interesting research path. Note that in the Gaussian case, up until now adaptation is theoretically supported only when an hyper-prior is used on the architecture (e.g. as in the work of Kong and Kim (2024)). In order to obtain a scalable variational result for such a prior one would have to adapt parallelization techniques, using for example the approach suggested by Ohn and Lin (2024), where a variational class with uniform prior has been studied. This remains to be done to the best of our knowledge.

Another relevant aspect of the approach in practice is the possibility to make different choices for scaling sequences (σ_k). For simplicity we have mostly focused on the constant choice as in (14). Other choices may be interesting, in particular for small sample sizes, for instance taking scalings matching the lower bound in (16). Indeed, since one expects a number of coefficients to be of significant amplitude, it may help to have at least a few scalings of the order of a constant, as is the case for the latter choice. This may accelerate, in small or moderate samples, the 'soft selection' of the weights of highest amplitude in the network. Again we leave investigating this for future work.

4.4 Discussion: fractional posteriors ($\alpha < 1$) and classical posteriors ($\alpha = 1$)

The main (technical) reason for which we have primarily focused on fractional posteriors is that proving convergence for these *only* requires a prior-mass control. In particular, one avoids the necessity to build sieve sets for which the complexity (e.g. entropy) needs to be controlled, as in the usual conditions using a generic theorem such as the one of Ghosal et al. (2000). A first reason why building a sieve set is difficult here is that we use a overparametrized prior, that is, all weights in the neural network are modeled, so one cannot take as sieve the set of neural networks built on the optimal 'oracle' architecture (one that gives the optimal convergence rate). A second reason is connected to heavy-tails: the usual contraction Theorem in Ghosal et al. (2000) requires exponentially fast decrease of the prior mass of the complement of the sieve sets. But working with heavy-tailed priors on the weights makes verifying this difficult, unless the sieve sets are very large.

To illustrate this discussion, it is helpful to consider the example of priors on functions defined as random series on a given orthonormal basis. In the case of normally distributed coefficients, the work van der Vaart and van Zanten (2008) shows optimal posterior contraction following the generic approach of Ghosal et al. (2000); this has been extended to p-exponential priors in Agapiou et al. (2021) ($p \in [1,2]$, the case $p = 1$ corresponding to Laplace priors on coefficients, the case $p = 2$ recovering the Gaussian case), their proofs strongly relying on exponential-type decrease of complements of sieve sets. The case of heavy-tailed coefficients has been recently investigated in Agapiou and Castillo (2024); therein, it is proved that the (classical) posterior distribution for a heavy-tailed series prior indeed converges at a near-optimal adaptive minimax rate for a very specific model, namely the Gaussian white noise: although one cannot apply the generic contraction theorem, for that model one is able to prove convergence by a direct analysis of the posterior. For more

complex models (such as random design regression or classification), proving this is an open problem. We refer to the discussion in the Appendix of Agapiou and Castillo (2024) for more technical details on how one could attempt to prove a posterior rate for the classical posterior in general settings.

In Section 3.8, we provide a result for the classical posterior ($\alpha = 1$) for the augmented prior that also models the noise variance. By exploiting a link between classical and fractional posterior in this case, one is able to circumvent the use of sieve sets and control of the complexity of these.

Coming back to fractional posteriors ($\alpha < 1$), although our results do not enable to discriminate between different choices of α, this choice is an important practical question. More generally, the question of how to choose α for fractional posteriors goes much beyond the scope of the present contribution, but we give now a few pointers to recent work on this point. One motivation for the use of α–posteriors for statistical inference is their greater robustness to model misspecification compared to the usual Bayesian posterior, as discussed for instance in works by Grünwald (2012) and by Bissiri et al. (2016). The latter work also discusses possible ways to choose α. Other works in this direction include Holmes and Walker (2017); Lyddon et al. (2019); Syring and Martin (2019). Although beyond the scope of the paper, extending the results to possibly misspecified models with additional insight on the choice of α is an interesting research direction.

4.5 Discussion: possible extensions with geometric data

In Section 3.3, we work under Assumption 1, which assumes that the covariates (X_i) are supported on a compact set with bounded Minkowski dimension. A natural extension of this model would involve adding noise to these covariates. For example, one could assume that observations are sampled from points on a compact manifold with some additive noise locally perpendicular to the manifold. Alternatively, one could assume the covariates are supported within a tubular region around a compact manifold.

Another interesting but more complex model would combine both anisotropy (as in Section 3.4) and manifold assumptions by, for instance, allowing the true function to exhibit different levels of smoothness along and orthogonal to the manifold. Although this is beyond the scope of the present paper, such a model has been studied within a nonparametric Bayesian framework (with simpler priors, not neural network ones) in density estimation by Berenfeld et al. (2024).

We now discuss what happens in the case of a tubular region. Assume that there exists a compact manifold $M \subset [0,1]^d$ with dimension $d^* < d$ and a "noise level" $\delta > 0$ such that the distribution of the covariates P_X is supported on the δ-tubular region around M,

$$M_\delta := \{x \in [0,1]^d \, : \, d(x, M) \leq \delta\}.$$

The first step for applying our analysis (as in Theorem 7 for instance) would be to establish an approximation result for f_0 on M_δ using wide DNNs. Since this assumption is strictly weaker than Assumption 1, one would expect a slower approximation rate that depends on both d^* and δ. In fact, when δ is too large, minimax approximation results may not be feasible due to the increased difficulty of the problem. For a Hölder-smooth

$f_0 \in \mathcal{C}^\beta(F)$, one such approximation result is given in Jiao et al. (2023). By following the proof of their Theorem 6.1 and applying our usual enlargement of the network (for instance as in Proposition 14), one can show that there exists a DNN realization $\tilde{f}$, with depth $\log^2 n$ and width $\sqrt{n}$, such that as $n \to \infty$,

$$\|\tilde{f} - f_0\|_{L^2(P_X)} \lesssim n^{-\frac{\beta}{2\beta + d_\eta}},$$

where $d_\eta = O(d^* \log d)$, provided that the tubular region is not too large in the sense that $\delta \leq C n^{-\frac{\beta}{2\beta + d_\eta}} (\log n)^{-4\beta/d_\eta}$, with a constant C explicitly derived by Jiao et al. (2023). Notably, this rate includes a factor of $d^* \log d$ rather than d^* as in the exact manifold case, resulting in a slightly slower rate.

A second step would involve verifying that the weights of this approximating neural network are bounded polynomially in terms of the sample size. This could be achieved by following the proofs in Jiao et al. (2023) to track these bounds, in a similar way as we did for the results by Kohler and Langer in Kohler and Langer (2021) in Appendix D.

4.6 Further work and open questions

Let us mention a few further natural questions arising from the present results beyond the just-mentioned aspects on simulations. First, although we are able to derive a result for the classical posterior with an augmented prior, the question is left open as to whether posterior contraction still holds *without* putting a prior on the noise variance. In the setting of random series priors, the simulation study in Agapiou and Castillo (2024) conducted in regression (as well as density and classification models) for both fractional and classical posteriors suggests that there is no visible phase transition when α increases from e.g. $1/2$ up until 1 (included); we conjecture that all the theoretical results presented here (with no prior on τ^2) still go through for standard posteriors ($\alpha = 1$).

Also, as mentioned above, a promising direction is to derive results for horseshoe priors, for instance in the spirit of Ghosh et al. (2019), as well as to compare with the present heavy-tailed priors. We are currently investigating this in the simpler setting of white noise regression with series priors in Agapiou et al. (2025).

Finally, it would be interesting to understand more the effect of overparametrization with our approach: as we have noted, we can choose a polynomial width with large power for the network; also, for the specific choice of constant σ_k the prior is exchangeable and does not particularly penalise a high number of large coefficients; one may then think that this setting may be particularly appropriate to test for the presence of a *double descent* phenomenon, which features a decrease of the risk for very overfitted models. It is conceivable that for very large widths one sees this appear in the present context: this deserves further investigation.

5. Proofs of the main results

In this section we give the proofs of the main results for the compositional models.

5.1 Fully connected ReLU DNN approximation

We recall here the approximation result we will use to study posterior contraction towards compositional functions as in Theorems 2 and 5.

Proofs are done in appendix C. For Proposition 14 we follow the same approach as Schmidt-Hieber (2020, Theorem 1), simply using the fully-connected architecture from Kohler and Langer (2021) for the elementary building blocks, as it is better suited to our prior. Note that Kohler and Langer (2021) does not give a explicit bound on the magnitude of the coefficients in the approximating neural network $\tilde{f}$. Such a bound is fundamental in the study of the properties of our prior and is provided in appendix D.

Proposition 14 *Let $f \in \mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$, L as in (11) and ϕ_n as in (18). There exists a ReLU network*

$$\tilde{f} \in \mathcal{F}(L, \mathbf{r}(\sqrt{n}), s)$$

with

$$s \lesssim n\phi_n^2 \log^{1-\gamma} n$$

such that for sufficiently large n,

$$\|f - \tilde{f}\|_\infty \lesssim \phi_n.$$

Additionally, all coefficients of $\tilde{f}$ satisfy $|\tilde{\theta}_k| \leq n^{c_\beta}$, where $c_\beta \geq 1$ is a constant only depending on $(\beta_i)_i$.

Remark 15 *We can be more precise in Proposition 14 and construct the network $\tilde{f}$ in such a way that the position of the active coefficients is known. From the proof given in Appendix C.2 it follows that there is an integer*

$$r^* \asymp \max_{0 \leq i \leq q} \left(\frac{n}{\log^\gamma n} \right)^{\frac{1}{2} \frac{t_i}{2\beta_i^* + t_i}},$$

such that if $(\tilde{W}_l, \tilde{v}_l)$ are the coefficients of $\tilde{f}$, we have for all $l \in [L+1]$, $\tilde{W}_l^{(ij)} = 0$ whenever $i \vee j > r^*$ and $\tilde{v}_k^{(i)} = 0$ whenever $i > r^*$.

5.2 Proof of Theorem 2

Proof In view of Lemma 17, it suffices to show that there exists $C > 0$ such that for sufficiently large n,

$$\Pi\left(\|f - f_0\|_\infty \leq \phi_n \right) \geq e^{-Cn\phi_n^2}.$$

Applying Proposition 14, we obtain $\tilde{f}_0 \in \mathcal{F}(L, \mathbf{r}(\sqrt{n}), s)$ a ReLU neural network that approximates f_0 for large enough n with $\|f_0 - \tilde{f}_0\|_\infty \leq \phi_n/2$ and $s \lesssim n\phi_n^2 \log^{1-\gamma} n$. Using the triangle inequality,

$$\Pi\left(\|f - f_0\|_\infty \leq \phi_n \right) \geq \Pi\left(\|f - \tilde{f}_0\|_\infty \leq \phi_n/2 \right).$$

It suffice to control the difference in the supremum norm between two networks of the same structure. For this purpose we use Lemma 24. Let θ_k (resp. $\tilde{\theta}_k$) be the coefficients of f (resp. $\tilde{f}_0$). We know from Proposition 14 that $\sup_k |\tilde{\theta}_k| \leq n^{c_\beta}$ where $c_\beta \geq 1$. Recall that $V = \prod_{l=0}^{L}(r_l + 1)$ thus using lemma 24 and independence,

$$\Pi\left(\|f - \tilde{f}_0\|_\infty \leq \phi_n/2 \right) \geq \prod_{k=1}^{T} \Pi\left(|\tilde{\theta}_k - \theta_k| \leq \frac{\phi_n}{2n^{c_\beta L}V(L+1)} , \, |\theta_k| \leq n^{c_\beta} \right). \tag{25}$$

From remark 15 there is an integer

$$r^* \asymp \max_{0 \leq i \leq q} \left(\frac{n}{\log^{\gamma} n} \right)^{\frac{1}{2} \frac{t_i}{2\beta_i^* + t_i}} \tag{26}$$

such that, for all $l \in [L+1]$, $\tilde{W}_l^{(ij)} = 0$ whenever $i \vee j > r^*$. Let

$$z_n := \phi_n / (2n^{c_\beta L} V(L+1)), \tag{27}$$

note that $z_n \to 0$ when $n \to \infty$, so that $0 < z_n < 1$ when n is large enough.

We can then split the right hand side in (25) and first take care of the possibly large coefficients (but still bounded by n^{c_β}) using the heavy tails properties of the prior. To do so we use Lemma 27 with z_n, on every coordinate such that $i \vee j \leq r^*$. We get a large enough constant $C_0 > 0$ such that when n is sufficiently large,

$$\prod_{l=1}^{L+1} \prod_{i \vee j \leq r^*} \Pi \left(|W_l^{(ij)} - \tilde{W}_l^{(ij)}| \leq z_n \,, \, |W_l^{(ij)}| \leq n^{c_\beta} \right)$$

$$\geq \prod_{l=1}^{L+1} \prod_{i \vee j \leq r^*} \underbrace{z_n \exp \left(-C_0 \log^{1+\kappa} \left((1+n^{c_\beta})(\sigma_l^{(ij)})^{-1} \right) \right)}_{:=A}.$$

Recall that $V = \prod_{l=0}^{L} (r_l + 1) = (d+1)(\lceil \sqrt{n} \rceil + 1)^L$ and $L = \lceil \log^{1+\delta} n \rceil$, thus there is C' a large enough constant such that, when n is sufficiently large,

$$2n^{c_\beta L}(L+1)V \leq e^{C' \log^{2+\delta} n}. \tag{28}$$

Using $\log \left((\sigma_l^{(ij)})^{-1} \right) \leq \log^{2(1+\delta)} n$ from (16), this leads to the bound, for C_3 a large enough constant

$$A \geq \phi_n e^{-C' \log^{2+\delta} n - C_0 \log^{1+\kappa} \left((1+n^{c_\beta})(\sigma_l^{(ij)})^{-1} \right)} \geq \phi_n e^{-C_3 \log^{2(1+\delta)(1+\kappa)} n}. \tag{29}$$

Now since $\phi_n \asymp (n/\log^{\gamma} n)^{-\nu}$ for some $\nu > 0$, there exists a constant $C_4 > 0$ such that

$$\prod_{l=1}^{L+1} \prod_{i \vee j \leq r^*} A \geq e^{-C_4 (r^*)^2 L \log^{\gamma-1} n}.$$

Finally since $(r^*)^2 L \leq s \lesssim n\phi_n^2 \log^{1-\gamma} n$, we get for the indices such that $i \vee j \leq r^*$, a large enough constant $C > 0$ such that

$$\prod_{l=1}^{L+1} \prod_{i \vee j \leq r^*} \Pi \left(|W_l^{(ij)} - \tilde{W}_l^{(ij)}| \leq z_n \,, \, |W_l^{(ij)}| \leq n^{c_\beta} \right) \geq e^{-Cn\phi_n^2}.$$

Now we take care of the small coefficients using the decay (σ_k). For the indices such that $i \vee j > r^*$, recall that we have $\tilde{W}_l^{(ij)} = 0$ for all l, leaving us to bound

$$\prod_{l=1}^{L+1} \prod_{i \vee j > r^*} \Pi \left(|W_l^{(ij)}| \leq z_n \right) = \prod_{l=1}^{L+1} \prod_{i \vee j > r^*} \left[1 - 2\overline{H} \left(z_n / \sigma_l^{(ij)} \right) \right],$$

here we have used the symmetry of h and the fact that $z_n < n^{c_\beta}$ for n sufficiently large. Using $\log^{2(1+\delta)}(i \vee j) \leq \log(1/\sigma_l^{(ij)})$ from (16) and the expression of r^* given by (26), we have a large enough constant C_5 such that whenever $i \vee j > r^*$, we have

$$1/\sigma_l^{(ij)} \geq e^{\log^{2(1+\delta)}(i \vee j)} \geq e^{\log^{2(1+\delta)} r^*} \geq e^{C_5 \log^{2(1+\delta)} n}.$$

Recall that by assumption (H3), $\overline{H}(x) \leq c_2/x$ for all $x \geq 1$. Noting that $\log(1 - 2x) \geq -4x$ whenever $x < 1/4$ and using the bound (28) on V, one gets that for sufficiently large n satisfying

$$z_n e^{C_5 \log^{2(1+\delta)} n} \geq 1 \vee 4c_2,$$

there is a constant C_6 such that

$$1 - 2\overline{H}\left(z_n/\sigma_l^{(ij)}\right) \geq 1 - 2c_2 z_n^{-1} e^{-C_5 \log^{2(1+\delta)} n} \geq \exp(-C_6 z_n^{-1} e^{-C_5 \log^{2(1+\delta)} n}).$$

Recall T is the total number of parameters of the network, this leads to the lower bound

$$\prod_{l=1}^{L+1} \prod_{i \vee j > r^*} \Pi\left(|W_l^{(ij)}| \leq z_n\right) \geq \exp(-C_6 T z_n^{-1} e^{-C_5 \log^{2(1+\delta)} n}).$$

Whenever n is large enough and for our choice of architecture, we have $T \lesssim n \log^{1+\delta} n$, using again the bound (28) on V one gets for n sufficiently large,

$$T n^{c_\beta L} V(L+1) \lesssim n(\log^{(1+\delta)} n) e^{C' \log^{2+\delta} n} \lesssim n \phi_n^3 e^{C_5 \log^{2(1+\delta)} n}.$$

Finally we have a large enough constant $C > 0$ such that

$$\prod_{l=1}^{L+1} \prod_{i \vee j > r^*} \Pi\left(|W_l^{(ij)}| \leq z_n\right) \geq e^{-Cn\phi_n^2}.$$

∎

5.3 Proof of Theorem 5

Proof We give the proof for $\kappa = 0$, giving the fastest rate, the proof in the general case being similar. In this setting $\gamma = 2(1 + \delta) + 1$. Let $\tilde{f}_0 \in \mathcal{F}(L, \mathbf{r}(\sqrt{n}), s)$ be the network approximating f_0 from Proposition 14. Using $\|f - f_0\|_{L^2(P_X)}^2 \leq 2\|f - \tilde{f}_0\|_{L^2(P_X)}^2 + 2\|\tilde{f}_0 - f_0\|_{L^2(P_X)}^2$, Lemma 18 gives

$$E_{f_0}\left(\int D_\alpha(f, f_0) \, d\hat{Q}_\alpha(f)\right) \leq \frac{1}{1-\alpha}\|\tilde{f}_0 - f_0\|_{L^2(P_X)}^2$$
$$+ \inf_{Q \in \mathcal{S}}\left\{\frac{\alpha}{1-\alpha} \int \|f - \tilde{f}_0\|_{L^2(P_X)}^2 \, dQ(f) + \frac{\mathrm{KL}(Q, \Pi)}{n(1-\alpha)}\right\}.$$

31

Since $\|\tilde{f}_0 - f_0\|^2_{L^2(P_X)} \le \phi_n^2$ when n is large enough, it is enough to show that $\mathcal{S}_{HT}(h)$ contains a distribution Q^* satisfying the extended prior mass condition (see Remark 19),

$$\int \|f - \tilde{f}_0\|^2_{L^2(P_X)} \, dQ^*(f) \le \phi_n^2 \qquad \text{and} \qquad \mathrm{KL}(Q^*, \Pi) \le n\phi_n^2. \tag{30}$$

For h the heavy-tailed density taken as prior on coefficients, let us define $Q^* := \bigotimes_{k=1}^T Q_k^* \in \mathcal{S}_{HT}(h)$ by setting for all $k \in [T]$,

$$\frac{dQ_k^*}{d\theta_k}(\theta_k) := \frac{1}{\sigma_k} h\left(\frac{\theta_k - \tilde{\theta}_k}{\sigma_k}\right) = h_{\tilde{\theta}_k, \sigma_k}(\theta_k),$$

where $\sigma_k = e^{-\log^{2(1+\delta)} n}$ and $\{\tilde{\theta}_k\}$ are the coefficients of $\tilde{f}_0$, such that $Q_k^* \in \mathcal{H}(h)$.

Using Lemma 24 and setting $w_n := n^{2c_\beta L} V^2 (L+1)^2$, we get

$$\int \|f - \tilde{f}_0\|^2_{L^2(P_X)} \, dQ^*(f) \le \int \|f - \tilde{f}_0\|^2_\infty \, dQ^*(f) \le w_n \int \max_k |\theta_k - \tilde{\theta}_k|^2 \, dQ^*(f).$$

Recall the definition (9) of the moments of h, from the definition of Q^* it follows,

$$\int \max_k |\theta_k - \tilde{\theta}_k|^2 \, dQ^*(f) \le \sum_{k=1}^T \int |\theta_k - \tilde{\theta}_k|^2 dQ_k^*(\theta_k) = m_2(h) \sum_{k=1}^T \sigma_k^2 = m_2(h) T e^{-2\log^{2(1+\delta)} n}.$$

Using inequality (28), we get $w_n T \le e^{C \log^{2+\delta} n}$ for $C > 0$ sufficiently large. Since $m_2(h) \le c_3^{2/(2\vee(1+\kappa))}$ by assumption, we have $m_2(h) w_n T \sigma_k^2 \le \phi_n^2$ if n is large enough.

We just showed that Q^* satisfies the first condition in (30), let us check the second one. We can write $\Pi = \bigotimes_{k=1}^T \Pi_k$ where,

$$\frac{d\Pi_k}{d\theta_k}(\theta_k) = \frac{1}{\sigma_k} h\left(\frac{\theta_k}{\sigma_k}\right) = h_{0,\sigma_k}(\theta_k).$$

By additivity using independence we have $\mathrm{KL}(Q^*, \Pi) = \sum_{k=1}^T \mathrm{KL}(Q_k^*, \Pi_k)$. Let us write $S_0 = \{k \in [T], \tilde{\theta}_k \ne 0\}$ the support of the collection of coefficients of $\tilde{f}_0$, since $Q_k^* = \Pi_k$ whenever $\tilde{\theta}_k = 0$, we get

$$\mathrm{KL}(Q^*, \Pi) = \sum_{k \in S_0} \mathrm{KL}(Q_k^*, \Pi_k).$$

Let $k \in S_0$. Using the fact that h is bounded,

$$\mathrm{KL}(Q_k^*, \Pi_k) = \int \log\left(\frac{h\left(\frac{\theta - \tilde{\theta}_k}{\sigma_k}\right)}{h\left(\frac{\theta}{\sigma_k}\right)}\right) \frac{1}{\sigma_k} h\left(\frac{\theta - \tilde{\theta}_k}{\sigma_k}\right) d\theta$$

$$\le \log \|h\|_\infty + \underbrace{\int \log\left(\frac{1}{h\left(\frac{\theta}{\sigma_k}\right)}\right) \frac{1}{\sigma_k} h\left(\frac{\theta - \tilde{\theta}_k}{\sigma_k}\right) d\theta}_{I_k}.$$

Using the heavy-tailed assumption (H2) with $\kappa = 0$ on h, for all $\theta > 0$, we get

$$\log\left(\frac{1}{h(\theta\sigma_k^{-1})}\right) \leq c_1(1 + \log(1 + \theta\sigma_k^{-1})).$$

Splitting the integral I_k in two and using the symmetry of h leads to

$$I_k \leq \int_0^{+\infty} \log\left(\frac{1}{h(\theta\sigma_k^{-1})}\right)\frac{1}{\sigma_k}h\left(\frac{\theta - \tilde{\theta}_k}{\sigma_k}\right)d\theta + \int_0^{+\infty} \log\left(\frac{1}{h(\theta\sigma_k^{-1})}\right)\frac{1}{\sigma_k}h\left(\frac{\theta + \tilde{\theta}_k}{\sigma_k}\right)d\theta$$

$$\leq \underbrace{\int_0^{+\infty} c_1\left[1 + \log(1 + \theta\sigma_k^{-1})\right]\frac{1}{\sigma_k}h\left(\frac{\theta - \tilde{\theta}_k}{\sigma_k}\right)d\theta}_{I_k^-} + \underbrace{\int_0^{+\infty} c_1\left[1 + \log(1 + \theta\sigma_k^{-1})\right]\frac{1}{\sigma_k}h\left(\frac{\theta + \tilde{\theta}_k}{\sigma_k}\right)d\theta}_{I_k^+}.$$

One can suppose $\tilde{\theta}_k > 0$, given the symmetry in I_k^+ and I_k^-. Considering first I_k^-, we have, by changing variables,

$$I_k^- \leq c_1 + c_1 \int_{-\tilde{\theta}_k/\sigma_k}^{+\infty} \log\left(1 + \frac{\tilde{\theta}_k}{\sigma_k} + u\right)h\left(u\right)du.$$

The integral in the last display is bounded as follows

$$\int_{-\tilde{\theta}_k/\sigma_k}^{+\infty} \log\left(1 + \frac{\tilde{\theta}_k}{\sigma_k} + u\right)h\left(u\right)du \leq \log\left(1 + 2\frac{\tilde{\theta}_k}{\sigma_k}\right) + \int_{\tilde{\theta}_k/\sigma_k}^{+\infty} \log\left(1 + 2u\right)h\left(u\right)du$$

$$\leq \log\left(1 + 2\frac{\tilde{\theta}_k}{\sigma_k}\right) + 2m_1(h).$$

For I_k^+, we have, changing variables

$$\int_{\tilde{\theta}_k/\sigma_k}^{+\infty} \log\left(1 - \frac{\tilde{\theta}_k}{\sigma_k} + u\right)h\left(u\right)du \leq \int_{\tilde{\theta}_k/\sigma_k}^{+\infty} \log\left(1 + u\right)h\left(u\right)du \leq m_1(h).$$

In the case of $\kappa > 0$ one gets $m_{1+\kappa}(h)$ instead of the first order moment, this quantity is also bounded by assumption. Finally, we have $I_k \lesssim 1 + \log\left(1 + 2\frac{|\tilde{\theta}_k|}{\sigma_k}\right)$ for n large enough. Using the fact that $|\tilde{\theta}_k| \leq n^{c_\beta}$ we obtain

$$\mathrm{KL}(Q_k^*, \Pi_k) \lesssim 1 + \log\left(1 + 2\frac{|\tilde{\theta}_k|}{\sigma_k}\right) \lesssim \log^{2(1+\delta)} n = \log^{\gamma-1} n.$$

Recalling that $|S_0| = s \lesssim n\phi_n^2 \log^{1-\gamma} n$, we finally get,

$$\mathrm{KL}(Q^*, \Pi) \lesssim s\log^{\gamma-1} n \lesssim n\phi_n^2,$$

which concludes the proof. ∎

Acknowledgments

Appendix A. Tempered posterior contraction

In this appendix we recall some results regarding the concentration of posterior distributions.

Definition 16 *Let $\alpha \in (0,1)$ and P, Q be two probability measures. The Kullback-Liebler (KL) divergence between P and Q is defined by*

$$\mathrm{KL}(P, Q) := \int \log\left(\frac{dP}{dQ}\right) dP \quad \text{if } P \ll Q, \text{ and } +\infty \text{ otherwise.}$$

Let μ be any measure satisfying $P \ll \mu$ and $Q \ll \mu$. The α-Rényi divergence between P and Q is defined by

$$D_\alpha(P, Q) := \frac{1}{\alpha - 1} \log\left(\int \left(\frac{dP}{d\mu}\right)^\alpha \left(\frac{dQ}{d\mu}\right)^{1-\alpha} d\mu\right).$$

For an overview on properties of such divergences we refer to van Erven and Harremoes (2014). In particular, if P_{f_0, τ_0^2} is the probability measure such that $(X_i, Y_i) \overset{i.i.d.}{\sim} P_{f_0, \tau_0^2}$ in the regression setting (1), using additivity of the Rényi divergence for independent observations, for any $f \in \mathcal{F}$ and $\tau^2 > 0$, we have

$$D_\alpha(P_{f,\tau^2}, P_{f_0,\tau_0^2}) = \frac{1}{n} D_\alpha(P_{f,\tau^2}^{\otimes n}, P_{f_0,\tau_0^2}^{\otimes n}). \tag{31}$$

When τ_0 is supposed to be known (e.g. $\tau_0 = 1$) we simply denote

$$D_\alpha(f, f_0) := D_\alpha(P_{f,\tau_0^2}, P_{f_0,\tau_0^2}).$$

A.1 Contraction Lemmas : case of known variance $\tau_0 = 1$.

In order to simplify the reading of this section we first state the results we use to prove our Theorems in the case where the noise variance $\tau_0 = 1$ is known. A more general version used in the unknown variance case can be found below.

Lemma 17 *Let $f_0 \in \mathcal{F}$ and assume $\tau_0 = 1$ is known. Let Π be a probability measure on $\mathcal{F}$. For any $\alpha \in (0,1)$, any positive sequence $(\varepsilon_n)_n$ such that $\varepsilon_n \to 0$, $n\varepsilon_n^2 \to \infty$ and*

$$\Pi\left[\{f \,:\, \|f - f_0\|_\infty \le \varepsilon_n\}\right] \ge e^{-n\varepsilon_n^2},$$

there is a constant $C > 0$ such that as n tends to infinity,

$$E_{f_0} \Pi_\alpha \left[\left\{ f \,:\, D_\alpha(f, f_0) \geq C \frac{\alpha \varepsilon_n^2}{1 - \alpha} \right\} \mid X, Y \right] \to 0,$$

where E_{f_0} denotes the expectation under P_{f_0}.

This is Theorem 4.1 in L'Huillier et al. (2023) written in our regression setting, noting that the KL–neighborhood therein simplifies to an L^∞–neighborhood via Lemma 20. The version of this result that accounts for possible unknown noise variance τ_0 is Lemma 23 below.

Lemma 18 *Let $f_0 \in \mathcal{F}$ and assume $\tau_0 = 1$ is known. Let Π a probability measures on $\mathcal{F}$ and $\mathcal{S}$ a set of probability measure on $\mathcal{F}$. For $\alpha \in (0, 1)$ let $\hat{Q}_\alpha$ be the tempered variational approximation on $\mathcal{S}$ defined by (8). Then*

$$E_{f_0} \left(\int D_\alpha(f, f_0) \, d\hat{Q}_\alpha(f) \right) \leq \inf_{Q \in \mathcal{S}} \left\{ \frac{\alpha}{2(1 - \alpha)} \int \| f - f_0 \|^2_{L^2(P_X)} \, dQ(f) + \frac{\mathrm{KL}(Q, \Pi)}{n(1 - \alpha)} \right\}.$$

This is Theorem 2.6 from Alquier and Ridgway (2020) written in our regression framework, using (A.2) with $\tau_0 = \tau = 1$, to express the Kullback divergence therein as an $L^2(P_X)$-squared norm.

Remark 19 *If there is $\varepsilon_n > 0$ such that*

$$E_{f_0} \left(\int D_\alpha(f, f_0) \, d\hat{Q}_\alpha(f) \right) \leq \varepsilon_n^2,$$

then by Markov's inequality, for any $M_n \to \infty$, as $n \to \infty$,

$$E_{f_0} \hat{Q}_\alpha \left[\{ f \,:\, D_\alpha(f, f_0) \geq M_n \varepsilon_n^2 \} \right] \to 0.$$

Moreover, if we consider the clipped posterior $\hat{Q}_\alpha^B$ as in Corollary 3 we obtain convergence of the variational posterior mean in $L^2(P_X)$-distance, as then Lemma 21 gives

$$E_{f_0} \left\| \int f \, d\hat{Q}_\alpha^B(f) - f_0 \right\|_{L^2(P_X)} \leq E_{f_0} \int \| f - f_0 \|_{L^2(P_X)} \, d\hat{Q}_\alpha^B(f) \lesssim \varepsilon_n.$$

A.2 Extended Lemmas : unknown variance case

We provide now a sufficient condition for tempered posterior contraction, recalled in Lemma 23 below, it involves putting enough prior mass on the Kullback-neighborhood

$$B_n((f_0, \tau_0^2), \varepsilon_n) := \{ (f, \tau^2) \,:\, \mathrm{KL}(P_{f_0, \tau_0^2}, P_{f, \tau^2}) \leq \varepsilon_n^2, \, V_2(P_{f_0, \tau_0^2}, P_{f, \tau^2}) \leq \varepsilon_n^2 \}, \tag{32}$$

where

$$V_2(P_{f_0, \tau_0^2}, P_{f, \tau^2}) := \int \left(\log \left(\frac{dP_{f_0, \tau_0^2}}{dP_{f, \tau^2}} \right) - \mathrm{KL}(P_{f_0, \tau_0^2}, P_{f, \tau^2}) \right)^2 \, dP_{f_0, \tau_0^2}.$$

In the regression setting, simple calculations allow us to identify the KL-neighborhood (32) as an L^∞-type ball.

Lemma 20 *One can find a constant $C > 0$ sufficiently large, such that it holds, for any $\varepsilon_n \to 0$ as $n \to \infty$,*

$$\{(f, \tau^2) : ||f - f_0||_\infty \le \varepsilon_n, \, |\tau^2 - \tau_0^2| \le \varepsilon_n^2\} \subset B_n((f_0, \tau_0^2), C\varepsilon_n).$$

In particular when $\tau_0 = 1$ is known,

$$\{f : ||f - f_0||_\infty \le \varepsilon_n\} \subset B_n(f_0, C\varepsilon_n).$$

Proof Simple calculations (see for instance Lemma 19 in Castillo and Randrianarisoa (2024)) lead to

$$\mathrm{KL}(P_{f_0,\tau_0}, P_{f,\tau}) = \frac{1}{2\tau^2} ||f - f_0||_{L^2(P_X)}^2 + \frac{1}{2} \left(\log(\frac{\tau^2}{\tau_0^2}) + \frac{\tau_0^2}{\tau^2} - 1 \right),$$

$$V_2(P_{f_0,\tau_0}, P_{f,\tau}) = \frac{\tau_0^2}{\tau^4} ||f - f_0||_{L^2(P_X)}^2 + \frac{(\tau_0^2 - \tau^2)^2}{2\tau^4} + \frac{1}{4\tau^4} \left(||f_0 - f||_{L^4(P_X)}^4 - ||f_0 - f||_{L^2(P_X)}^4 \right).$$

Using for any $p \ge 1$, $||f_0 - f||_{L^p(P_X)} \le ||f - f_0||_\infty \le \varepsilon_n$ and $|\tau^2 - \tau_0^2| \le \varepsilon_n^2$ as well as $\varepsilon_n \to 0$ when $n \to \infty$, we get

$$\mathrm{KL}(P_{f_0,\tau_0^2}, P_{f,\tau^2}) \le \frac{\varepsilon_n^2}{\tau_0^2 - \varepsilon_n^2} + \frac{1}{2} \log(1 + \frac{\varepsilon_n^2}{\tau_0^2}) \le C\varepsilon_n^2$$

$$V_2(P_{f_0,\tau_0^2}, P_{f,\tau^2}) \le \frac{\tau_0^2 \varepsilon_n^2}{(\tau_0^2 - \varepsilon_n^2)^2} + \frac{3\varepsilon_n^4}{4(\tau_0^2 - \varepsilon_n^2)^2} \le C\varepsilon_n^2.$$

$\blacksquare$

The contraction Theorems for tempered posteriors in Section 3 are obtained applying Lemmas 17 and 18 above (or Lemma 23 in the case of unknown variance), therefore the rates are formulated in terms of Rényi divergences. In the random design Gaussian regression model (1), one can relate such rates to usual $L^2(P_X)$ ones, provided the true function f_0 has a known upper bound M_0.

Lemma 21 *Let $P_{f,\tau^2}, P_{f_0,\tau_0^2}$ be probability measures of the regression model (1) and let $\alpha \in (0, 1)$. Assume $||f||_\infty, ||f_0||_\infty \le M_0$ for some $M_0 > 0$. One can find a constant $C = C(\alpha, \tau_0^2, M_0)$ such that for any $\varepsilon_n \to 0$, as $n \to \infty$, it holds that*

$$\{(f, \tau^2) : D_\alpha(P_{f,\tau^2}, P_{f_0,\tau_0^2}) \le \varepsilon_n^2\} \subset \{(f, \tau^2) : ||f - f_0||_{L^2(P_X)} \le C\varepsilon_n, \, |\tau^2 - \tau_0^2| \le C\varepsilon_n\}.$$

In particular, if $\tau_0^2 = 1$ is known, we have

$$D_\alpha(f, f_0) \ge \frac{\alpha}{2} e^{-2M_0^2 \alpha(1-\alpha)} ||f - f_0||_{L^2(P_X)}^2.$$

Proof Let us express $D_\alpha(P_{f,\tau^2}, P_{f_0,\tau_0^2})$, more explicitly. By using the standard formula expressing the D_α-divergence between two univariate Gaussians we get

$$D_\alpha(\mathcal{N}(f(x), \tau^2), \mathcal{N}(f_0(x), \tau_0^2)) = \frac{\alpha}{2\tau_\alpha^2} (f(x) - f_0(x))^2 + \frac{1}{1-\alpha} \log \frac{\tau_\alpha}{\tau_0^{1-\alpha} \tau^\alpha},$$

where $\tau_\alpha^2 := (1-\alpha)\tau^2 + \alpha\tau_0^2$, see e.g. van Erven and Harremoes (2014) eq. (10), one has

$$
\begin{aligned}
D_\alpha(P_{f,\tau^2}, P_{f_0,\tau_0^2}) &= \frac{1}{\alpha - 1} \log \int e^{(\alpha-1)D_\alpha(\mathcal{N}(f(x),\tau^2),\mathcal{N}(f_0(x),\tau_0^2))} \, dP_X(x) \\
&= \frac{1}{1-\alpha} \log \left(\frac{\tau_\alpha}{\tau_0^{1-\alpha}\tau^\alpha} \right) + \frac{1}{\alpha - 1} \log \int e^{-\frac{\alpha(1-\alpha)}{2\tau_\alpha^2}(f(x)-f_0(x))^2} \, dP_X(x) \\
&= \qquad\qquad (I) \qquad\qquad + \qquad\quad (II).
\end{aligned}
$$

We have both $(I) \geq 0$ (e.g. by concavity of the logarithm) and $(II) \geq 0$. From the assumption $D_\alpha(P_{f,\tau^2}, P_{f_0,\tau_0^2}) \leq \varepsilon_n^2$ it follows that $(I) \leq \varepsilon_n^2$ and $(II) \leq \varepsilon_n^2$. From Lemma 22 below, there is a constant $C > 0$, such that $|\tau^2 - \tau_0^2| \leq C\varepsilon_n$. Note that $(II) = D_\alpha(P_{f,\tau_\alpha^2}, P_{f_0,\tau_\alpha^2})$. Since $|\tau^2 - \tau_0^2| \leq C\varepsilon_n$, using $\varepsilon_n \to 0$, we get $\tau_\alpha^2 \leq \tau_0^2 + C(1-\alpha)\varepsilon_n \leq 2\tau_0^2$ for n large enough. Therefore $(II) \geq D_\alpha(P_{f,2\tau_0^2}, P_{f_0,2\tau_0^2})$, where we used that the α-Divergence is a non-increasing function of the variance parameter. Now, using first $1 - x \leq -\log x$ and next $1 - e^{-x} \geq xe^{-x}$, leads to

$$
\begin{aligned}
(II) &\geq \frac{1}{1-\alpha} \left[1 - \int \exp\left(\frac{\alpha(\alpha-1)}{4\tau_0^2}(f-f_0)^2 \right) dP_X \right] \\
&\geq \frac{\alpha}{4\tau_0^2} \int e^{\alpha(\alpha-1)(f-f_0)^2/(4\tau_0^2)} (f-f_0)^2 \, dP_X \\
&\geq \frac{\alpha}{4\tau_0^2} e^{-2M_0^2\alpha(1-\alpha)} \|f - f_0\|_{L^2(P_X)}^2.
\end{aligned}
$$

Therefore, $\|f - f_0\|_{L^2(P_X)} \leq C\varepsilon_n$ for a suitable constant $C > 0$. $\blacksquare$

Lemma 22 *Let $\psi = \psi_\alpha$ be a function defined on $\mathbb{R}_{>0}$, for $0 < \alpha < 1$ and $\tau_0 > 0$, by*

$$
\psi(\tau^2) := \log\left(\frac{(1-\alpha)\tau^2 + \alpha\tau_0^2}{(\tau^2)^{1-\alpha}(\tau_0^2)^\alpha} \right).
$$

There is a constant $c = c(\alpha, \tau_0^2) > 0$, such that for any $\varepsilon_n \to 0$, as $n \to \infty$,

$$
\{\tau^2 : \psi(\tau^2) \leq c\varepsilon_n^2\} \subset \{|\tau^2 - \tau_0^2| \leq \varepsilon_n\}.
$$

Proof Denoting $\tau_\alpha^2 := (1-\alpha)\tau^2 + \alpha\tau_0^2$, one gets

$$
\psi'(\tau^2) = \alpha(1-\alpha)(1 - \tau_0^2/\tau^2)(\tau_\alpha^2)^{-1},
$$

so that $\psi'(\tau_0^2) = 0$, ψ is decreasing from $+\infty$ to $\psi(\tau_0^2) = 0$ on $(0, \tau_0^2]$ and increasing from 0 to $+\infty$ on $[\tau_0^2, \infty)$. Also,

$$
\psi''(\tau^2) = \frac{1-\alpha}{(\tau^2)^2} - \frac{(1-\alpha)^2}{(\tau_\alpha^2)^2},
$$

so that $\psi''(\tau_0^2) =: 2m > 0$. By continuity of ψ'', there exists an interval $I := [\tau_0^2 \pm \eta]$, for some $\eta = \eta(\tau_0^2, \alpha)$, such that $\psi'' \geq m$ on I. Set

$$
c := c(\tau_0^2, \alpha) = \min(m/2, \psi(\tau_0^2 - \eta), \psi(\tau_0^2 + \eta)).
$$

Let τ^2 verify $\psi(\tau^2) \leq c\varepsilon_n^2$. For n large enough we have $\psi(\tau^2) \leq c$ and by monotonicity of ψ on each side of τ_0^2 one must have $|\tau^2 - \tau_0^2| \leq \eta$, i.e. $\tau^2 \in I$. Recall $\psi(\tau_0^2) = \psi'(\tau_0^2) = 0$, computing the Taylor expansion of ψ at τ_0^2 gives, for any $\tau^2 \in I$, and some $\zeta \in I$,

$$\psi(\tau^2) = \psi''(\zeta)(\tau^2 - \tau_0^2)^2/2 \geq (m/2)(\tau^2 - \tau_0^2)^2,$$

where we used that $\psi'' \geq m$ on I. Therefore as soon as $\psi(\tau^2) \leq c\varepsilon_n$ and n is large enough, we have $|\tau^2 - \tau_0^2| \leq \sqrt{2c/m}\varepsilon_n$, which gives the result using $2c/m \leq 1$ by definition of c. ∎

Lemma 23 *Let $f_0 \in \mathcal{F}$ and $\tau_0 > 0$ and Π be a probability measure on $\mathcal{F} \times \mathbb{R}_{>0}$. For any $\alpha \in (0,1)$, any positive sequence $(\varepsilon_n)_n$ such that $\varepsilon_n \to 0$, $n\varepsilon_n^2 \to \infty$ and*

$$\Pi\left[\{(f,\tau^2) : \|f - f_0\|_\infty \leq \varepsilon_n, |\tau^2 - \tau_0^2| \leq \varepsilon_n^2\}\right] \geq e^{-n\varepsilon_n^2},$$

there is a constant $M > 0$ such that, as $n \to \infty$,

$$E_{f_0}\Pi_\alpha\left[\left\{(f,\tau^2) : D_\alpha(P_{f,\tau^2}, P_{f_0,\tau_0^2}) \geq \frac{M\alpha}{1-\alpha}\varepsilon_n^2\right\} \mid X, Y\right] \to 0,$$

where E_{f_0} denotes the expectation under P_{f_0,τ_0^2}.

Appendix B. Additional properties of ReLU DNNs

We list here some well-known properties of ReLU neural networks, which are very useful for combining different networks:

- *Width enlargement*: $\mathcal{F}(L, \mathbf{r}, s) \subseteq \mathcal{F}(L, \mathbf{r}', s')$ whenever $\mathbf{r} \leq \mathbf{r}'$ component-wise and $s \leq s'$.

- *Composition*: Given $f \in \mathcal{F}(L, \mathbf{r})$ and $g \in \mathcal{F}(L', \mathbf{r}')$ with $r_{L+1} = r'_0$, we can define the network whose realization $g \circ f$ is in the class $\mathcal{F}(L + L', (\mathbf{r}, r'_1, \ldots, r'_{L'+1}))$.

- *Depth synchronization*: To synchronize the number of hidden layers between two networks, we can add layers with the identity matrix as many times as desired,

$$\mathcal{F}(L, \mathbf{r}, s) \subset \mathcal{F}(L + q, (\underbrace{r_0, \ldots, r_0}_{q \text{ times}}, \mathbf{r}), s + q\,r_0).$$

- *Parallelization*: Given $f \in \mathcal{F}(L, \mathbf{r})$ and $g \in \mathcal{F}(L, \mathbf{r}')$ with $r_0 = r'_0$, one can simultaneously realizes f and g within a joint network whose realization (f, g) is in the class $\mathcal{F}(L, (r_0, r_1 + r'_1, \ldots, r_{L+1} + r'_{L+1}))$.

Following is a very useful Lemma giving a bound on the distance between two ReLU neural networks realizations in terms of the distance between their coefficients. These types of inequalities are well established in the literature. The version presented below appears in the proof of Lemma 3 in Suzuki (2018).

Lemma 24 *Let f and $f^* \in \mathcal{F}(L, \mathbf{r})$ with total number of parameters T and coefficients (θ_k) and (θ_k^*). Suppose that for all $k \in [T]$, $|\theta_k| \leq b$, $|\theta_k^*| \leq b$, and $|\theta_k - \theta_k^*| \leq \delta$. Define*

$$V := \prod_{l=0}^{L} (r_l + 1), \tag{33}$$

then

$$||f - f^*||_{L^\infty([0,1]^{r_0})} \leq \delta V (b \vee 1)^L (L + 1).$$

For reader's convenience we recall below the approximation results that will be used to prove Theorems 7 and 10, these are respectively borrowed from Nakada and Imaizumi (2020) and Suzuki and Nitanda (2021) where the constants are made explicit.

Lemma 25 (Theorem 5 in Nakada and Imaizumi (2020)) *Let $f \in \mathcal{C}_d^\beta([0,1]^d, K)$ and assume that $\dim_M \operatorname{Supp} P_X < t$ holds with $t < d$. Let $\varepsilon > 0$. There is a ReLU neural network $\tilde{f}$ with constant depth $L = L(\beta, d, t)$ and sparsity $\tilde{s} \lesssim \varepsilon^{-t/\beta}$ such that, when ε is small enough,*

$$||f - \tilde{f}||_{L^\infty(P_X)} \leq \varepsilon,$$

moreover there is a constant $c = c(\beta, d, t)$ such that the weights $\{\tilde{\theta}_k, k \in [T]\}$ of $\tilde{f}$ all satisfy $|\tilde{\theta}_k| \leq \varepsilon^{-c}$.

Lemma 26 (Proposition 2 in Suzuki and Nitanda (2021)) *Let $p \in (1,2)$, $\beta \in \mathbb{R}_{>0}^d$ such that $\tilde{\beta} > 1/p$ and $f \in B_{pp}^\beta(1)$. Let N be a large enough positive integer, there exists $\tilde{f} \in \mathcal{F}(L, (d, r, \ldots, r, 1), s)$ with $L \lesssim \log N$, $r \lesssim N$ and $s \lesssim NL$ such that*

$$||f - \tilde{f}||_\infty \lesssim N^{-\tilde{\beta}},$$

additionally all the weights in $\tilde{f}$ are bounded by a universal constant.

Appendix C. Additional proofs

C.1 A technical Lemma

Lemma 27 *Let $0 < \sigma, z \leq 1$, $0 < |\tilde{\theta}| \leq B$ for some $B > 1$ and $\theta = \sigma \cdot \zeta$, where ζ is a random variable on $\mathbb{R}$ with heavy-tailed density h satisfying properties (H1), (H2) and (H3). Denoting by P_θ the induced distribution on θ, there is a large enough constant $C > 0$ such that,*

$$P_\theta(|\theta - \tilde{\theta}| \leq z, |\theta| \leq B) \geq z \exp(-C \log^{1+\kappa}((B + z)\sigma^{-1})).$$

Proof Write $B_- := (-B) \vee (\tilde{\theta} - z)$ and $B^+ := B \wedge (\tilde{\theta} + z)$ so that

$$P_\theta(|\theta - \tilde{\theta}| \leq z, |\theta| \leq B) = \int_{B_-}^{B_+} \frac{1}{\sigma} h\left(\frac{x}{\sigma}\right) dx.$$

Using the fact that $|\tilde{\theta}| < B$ one gets, when $\tilde{\theta} > 0$,

$$B^+ = B \wedge (\tilde{\theta} + z) \geq \tilde{\theta}, \qquad B_- = \tilde{\theta} - z$$

and when $\tilde{\theta} < 0$,

$$B^+ = \tilde{\theta} + z, \qquad B_- = (-B) \vee (\tilde{\theta} - z) \leq \tilde{\theta}.$$

Using the symmetry of h, one can reduce the problem to the case $\tilde{\theta} > 0$. Using the positivity of h and the fact that $\sigma \leq 1$, one gets

$$P_\theta(|\theta - \tilde{\theta}| \leq z\,,\, |\theta| \leq B) \geq \int_{\tilde{\theta}-z}^{\tilde{\theta}} h\left(\frac{x}{\sigma}\right)\, dx.$$

Using the monotonicity of h, we have

$$\int_{\tilde{\theta}-z}^{\tilde{\theta}} h\left(\frac{x}{\sigma}\right)\, dx \geq z \times h\left(\frac{|\tilde{\theta}| \vee |\tilde{\theta} - z|}{\sigma}\right) \geq zh((B+z)\sigma^{-1}).$$

Now simply use the fact that $(B + z)\sigma^{-1} > 1$ and the heavy tail property (H2) of h to conclude. ∎

C.2 Proof of Proposition 14

Proof First express $f = g_q \circ \cdots \circ g_0 : [0,1]^d \to \mathbb{R}$ as the composition of functions defined on hypercubes $[0,1]^{t_i}$. To do this, for all $1 \leq i \leq q-1$, define

$$h_0 := \frac{g_0}{2K} + \frac{1}{2}, \quad h_i := \frac{g_i(2K \cdot -K)}{2K} + \frac{1}{2}, \quad h_q := g_q(2K \cdot -K). \tag{34}$$

Such that $f_0 = h_q \circ \cdots \circ h_0$ and for all j,

$$h_{0j} \in \mathcal{C}_{t_0}^{\beta_0}([0,1]^{t_0}, 1), \quad h_{ij} \in \mathcal{C}_{t_i}^{\beta_i}([0,1]^{t_i}, (2K)^{\beta_i}), \quad h_{qj} \in \mathcal{C}_{t_q}^{\beta_q}([0,1]^{t_q}, K(2K)^{\beta_q}).$$

We can now approximate each of the h_{ij} by a network realisation using Lemma 33, before that, let us provide a propagation result that links the approximation quality of h_{ij} to that of the total composition.

Lemma 28 (Lemma 3 from Schmidt-Hieber (2020)) *Let $h_i = (h_{ij})_j$ be functions as in (34) with $K \geq 1$. There is a constant $C = C(K, \boldsymbol{\beta})$ such that for any functions $\tilde{h}_i := (\tilde{h}_{ij})_j$ with $\tilde{h}_{ij} : [0,1]^{t_i} \to [0,1]$, we have,*

$$\|h_q \circ \cdots \circ h_0 - \tilde{h}_q \circ \cdots \circ \tilde{h}_q\|_{L^\infty([0,1]^d)} \leq C \sum_{i=0}^{q} \||h_i - \tilde{h}_i|\|_{L^\infty([0,1]^{d_i})}^{\prod_{k=i+1}^{q}(\beta_k \wedge 1)}.$$

We can now apply Lemma 33 to each of the h_{ij} separately, using at $i \in \{0, \ldots, q\}$ fixed, $M = M_i := \left\lceil \left(\frac{n}{\log^\gamma n}\right)^{\frac{1}{2(2\beta_i^* + t_i)}} \right\rceil$. We get for all (i,j), $\tilde{h}_{ij} \in \mathcal{F}(L_i', (t_i, r_i', \ldots r_i', 1))$ such that

$$\|h_{ij} - \tilde{h}_{ij}\|_\infty \lesssim M_i^{-2\beta_i}, \tag{35}$$

where $L_i' := C_\ell \log n$ and $r_i' := C_w M_i^{t_i}$ with C_ℓ and C_w two constants given by lemma 33 only depending on $(\boldsymbol{\beta}, \mathbf{t})$.

To apply lemma 28 we need to make sure each $\tilde{h}_{ij}$ with $i < q$ takes values in $[0,1]$. This can be done through the two layer network $(1-(1-x)_+)_+$. Let $L_i = L'_i + 2$ and keep calling the new networks $\tilde{h}_{ij}$. Since the h_{ij} takes values in $[0,1]$ inequality (35) still holds with the two extra layers.

Now computing the networks $\tilde{h}_{ij}$ in parallel lends $\tilde{h}_i = (\tilde{h}_{ij})_{j \in [d_{i+1}]}$ in the class

$$\mathcal{F}(L_i, (d_i, r_i, \ldots, r_i, d_{i+1})),$$

where $r_i := d_{i+1} r'_i$. Now from inequality (35) we immediately get,

$$\||h_i - \tilde{h}_i|_\infty\|_{L^\infty([0,1]^{d_i})}^{\prod_{k=i+1}^{q}(\beta_k \wedge 1)} \lesssim M_i^{-2\beta_i^*}. \tag{36}$$

Finally we compute the composite network $\tilde{f} := \tilde{h}_q \circ \cdots \circ \tilde{h}_1$ in the class

$$\mathcal{F}(L_0 + \cdots + L_q, (d, r, \ldots, r, 1)), \tag{37}$$

where $r := \max_{0 \le i \le q} r_i$. Using lemma 28 and (36), for n large enough, we get,

$$\|f - \tilde{f}\|_\infty \lesssim \sum_{i=0}^{q} M_i^{-2\beta_i^*} \lesssim \max_i M_i^{-2\beta_i^*} \le \phi_n.$$

Now we have for sufficiently large n, $L_0 + \cdots + L_q = (q+1)C_\ell \log n \le \log^{1+\delta} n$ and $r \le \lceil \sqrt{n} \rceil$. Thus, for n sufficiently large the space (37) can be embedded into

$$\mathcal{F}(\lceil \log^{1+\delta} n \rceil, (d, \lceil \sqrt{n} \rceil, \ldots, \lceil \sqrt{n} \rceil, 1), s),$$

where

$$s \lesssim r^2 \log n \lesssim \max_{0 \le i \le q} \left(\frac{n}{\log^\gamma n} \right)^{\frac{t_i}{2\beta_i^* + t_i}} \times \log n \lesssim n\phi_n^2 \log^{1-\gamma} n.$$

Now for the bound on the coefficients, in view of proposition 34, for sufficiently large n every coefficient θ of the network satisfies

$$|\theta| \le \max_{0 \le i \le q} \lceil M_i^{2(\beta_i+1)} \rceil^2 \le n^{c_\beta},$$

where $c_\beta \ge 1$ is a constant only depending on $(\beta_i)_i$.

This completes the proof. ∎

C.3 Proofs for data with low Minkowski dimension support

Proof [Proof of Theorem 7]The proof is very similar to that of Theorem 2. Since we will use a sparse approximation result the position of the zeros in the network is not precisely known.

In view of Lemma 17 it suffices to show that there exists $C > 0$ such that for sufficiently large n,

$$\Pi\left(\|f - f_0\|_\infty \le \varepsilon_n\right) \ge e^{-Cn\varepsilon_n^2}.$$

We apply Lemma 25 for $t > t^*$ and $2\varepsilon = \varepsilon_n := \left(n/\log^{2(1+\kappa)+1} n\right)^{-\beta/(2\beta+t)}$. When n is large enough there is a network $\tilde{f}_0$ with constant length, number of non-zero parameters $\tilde{s} \lesssim \varepsilon_n^{-t/\beta}$ and magnitude of coefficients $\max_k |\tilde{\theta}_k| \leq \varepsilon^{-c} \leq n^c$, such that $||f_0 - \tilde{f}_0||_{L^2(P_X)} \leq \varepsilon_n/2$. Since $\tilde{s} \leq n$ when n is large enough, we can embed $\tilde{f}_0$ in the space $\mathcal{F}(\log n, (d, n, \ldots, n, 1), s)$ where $s \lesssim \tilde{s} \log n$. Write (θ_k) (resp. $(\tilde{\theta}_k)$) for the coefficients of f (resp. $\tilde{f}_0$). Recall that there is a constant $C_V > 0$, such that

$$V := \prod_{l=1}^{L}(r_l + 1) \leq e^{C_V \log^2 n}. \tag{38}$$

Let $S_0 = \{k \in [T] : \tilde{\theta}_k \neq 0\}$, recall $|S_0| \leq s \lesssim \varepsilon_n^{-t/\beta} \log n$ and let

$$z_n := \varepsilon_n/(2n^{cL}V(\log n + 1)).$$

Using triangle inequality and applying Lemma 24 we need to bound from below

$$\Pi\left(||f - f_0||_{L^2(P_X)} \leq \varepsilon_n\right) \geq \prod_{k=1}^{T}\Pi\left(|\theta_k - \tilde{\theta}_k| \leq z_n, |\theta_k| \leq n^c\right).$$

We can follow the proof of Theorem 2, splitting the product in the last display on whether or not k is in S_0. For $k \in S_0$ we apply Lemma 27 and get a large enough constant $C_7 > 0$ such that

$$\prod_{k \in S_0}\Pi\left(|\theta_k - \tilde{\theta}_k| \leq z_n, |\theta_k| \leq n^c\right) \geq z_n^s e^{-C_7 s \log^{2(1+\kappa)} n}.$$

Using the bound (38) we have a large enough constant C_V' such that

$$2n^{cL}V(\log n + 1) \leq e^{C_V' \log^2 n}.$$

Finally, use $s \log^{2(1+\kappa)} n \lesssim \varepsilon_n^{-t/\beta} \log^{2(1+\kappa)+1} n \leq n\varepsilon_n^2$ to get $C > 0$ large enough such that

$$\prod_{k \in S_0}\Pi\left(|\theta_k - \tilde{\theta}_k| \leq z_n, |\theta_k| \leq n^c\right) \geq e^{-Cn\varepsilon_n^2}.$$

For the zero coefficients, $k \notin S_0$ we can simply follow the proof of Theorem 2 and get

$$\prod_{k \notin S_0}\Pi\left(|\theta_k| \leq z_n\right) \geq e^{-Cn\varepsilon_n^2},$$

which, according to Lemma 17 gives us concentration of the tempered posterior in terms of the α-Rényi divergence D_α.

For the concentration of the variational posterior, using the existence of $\tilde{f}_0 \in \mathcal{F}(L, \mathbf{r}(n), s)$ such that for n large enough $||f_0 - \tilde{f}_0||_{L^2(P_X)} \leq \varepsilon_n/2$ and $s \log^{2(1+\kappa)} n \lesssim n\varepsilon_n^2$, one can easily follow the steps of the proof of Theorem 5 to get to the result. ∎

C.4 Proofs for anisotropic Besov spaces

Proof [Proof of Theorem 10] We apply Lemma 26 with $N = \lceil n/\log^\gamma n \rceil^{1/(2\tilde\beta+1)}$. We get $\tilde f_0 \in \mathcal{F}(L, (d, \Lambda, \ldots, \Lambda, 1), s)$ such that when n is large enough, $L \lesssim \log n$, $\Lambda \lesssim N$ and

$$s \lesssim NL \lesssim \epsilon_n^{-1/\tilde\beta} \log n \lesssim n\epsilon_n^2 \log^{1-\gamma} n.$$

The network realization $\tilde f_0$ satisfies $\|f_0 - \tilde f_0\|_\infty \lesssim \epsilon_n/2$ and the coefficients of $\tilde f_0$ are uniformly bounded by a constant. When n is large enough we can embed $\tilde f_0$ into the space $\mathcal{F}(\log^{1+\delta} n, \mathbf{r}(n), s)$ and assume that $\max_k |\tilde\theta_k| \le n$. From here one can follow the proofs of Theorems 5 and 7 to get the desired results. ∎

C.5 Proofs for unknown noise level τ_0 and the case $\alpha = 1$

Proof [Proof of Theorem 12] In view of Lemma 23 (a generalized version of Lemma 17) it is enough to show

$$\Pi\left[\{(f, \tau^2) : \|f - f_0\|_\infty \le \phi_n, |\tau^2 - \tau_0^2| \le \phi_n^2\}\right] \ge e^{-Cn\phi_n^2},$$

for some sufficiently large constant $C > 0$. Since $\Pi = \Pi_f \otimes \pi_{\tau^2}$, it suffices to verify

$$\Pi_f(\{f : \|f - f_0\|_\infty \le \phi_n) \ge e^{-(C-1)n\phi_n^2},$$
$$\pi_{\tau^2}(\{\tau^2 : |\tau^2 - \tau_0^2| \le \phi_n^2\}) \ge e^{-n\phi_n^2}.$$

Since Π_f is the heavy-tailed DNN prior used in Theorem 2, the first inequality of the last display as been proved in Section 5.2. The second inequality immediately follows combining Condition (23) and the fact that ϕ_n is polynomial in n^{-1}. A direct application of Lemma 23 leads to contraction of $\Pi_\alpha[\cdot|X, Y]$ around (f_0, τ_0^2) at rate ϕ_n in α-Rényi divergence. For the clipped posterior $\Pi_\alpha^B[\cdot|X, Y]$, one uses Lemma 21 to relate the α-Rényi neighborhood to the $L^2(P_X)$-type one. ∎

Proof [Proof of Theorem 13] We apply Proposition 1 in Castillo and Randrianarisoa (2024): the latter shows that, provided that one chooses the prior π_{τ^2} as in (24), that the prior on f verifies the prior mass condition, for some $D > 0$,

$$\Pi_f[\|f - f_0\|_\infty \le \varepsilon_n] \ge e^{-Dn\varepsilon_n^2}, \tag{39}$$

and that the posterior on f is clipped by $B \ge \|f_0\|_\infty$, then

$$E_{f_0,\tau_0}\Pi[\|f - f_0\|_{L^2(P_X)}^2 \le M\varepsilon_n^2 \,|\, (X, Y)] \to 1.$$

Since (39) is verified within the proof of Theorem 2 for $\varepsilon_n = c\phi_n$ for some constant $c > 0$, and since we work under the conditions of the latter Theorem, the proof is complete. ∎

C.6 Proofs for general activation functions

We want to prove that if ρ is admissible according to Definition 11, all of our results still hold for ρ-DNNs. As stated in Section 3.6 it is sufficient to show that a result similar to Lemma 24 still holds for ρ-DNNs and that ReLU : $x \mapsto \max(0,x)$ can be approximated by a ρ-DNN. We provide in Lemma 29 a version of Lemma 24 for Lispchitz-continuous activations; it follows by a similar proof as for a ReLU activation, tracking the Lipschitz constant throughout the proof, see e.g. the proof of Proposition 1 in Ohn and Kim (2019), Appendix B therein.

Lemma 29 *Let $\rho : \mathbb{R} \to \mathbb{R}$ be a C_ρ-Lipschitz-continuous activation function and f, f^* two ρ–DNNs with width vector $\mathbf{r} = (d, N, \ldots, N, 1)$, depth L and coefficients (θ_k) and (θ_k^*). Suppose that for all k, $|\theta_k| \le b$, $|\theta_k^*| \le b$, and $|\theta_k - \theta_k^*| \le \delta$.*

$$||f - f^*||_\infty \le \delta L (C_\rho (b \vee 1)(N+1))^L.$$

For the approximation result we use the following Lemma.

Lemma 30 *Assume ρ is admissible according to definition 11, for any $D \ge 1$ and for any $K \ge 1$ both taken large enough, there exists a ρ-DNN, $\tilde{f}_{\text{ReLU}}^{\rho,D}$, with constant width and depth, such that*

$$\sup_{x \in [-D,D]} |\tilde{f}_{\text{ReLU}}^{\rho,D}(x) - \text{ReLU}(x)| \lesssim \frac{1}{K}.$$

Additionally all the weights of $\tilde{f}_{ReLU}^{\rho,D}$ satisfy $|\tilde{\theta}| \lesssim K^2 D^6$.

Proof From Definition 11 of an admissible activation function, there exists $a < b \in \mathbb{R}$, such that for any $x \in \mathbb{R}$ and any $K \ge 1$,

$$|x\rho(Kx) - \max(ax, bx)| \lesssim 1/K.$$

Notice that for all $x \in \mathbb{R}$, using $a < b$, we have

$$\text{ReLU}(x) = \max(0,x) = \frac{1}{b-a}(\max(ax, bx) - ax). \tag{40}$$

From Lemma 1 and Lemma 2 in Kohler et al. (2022), for any $D \ge 1$ and any $K \ge 1$ both large enough, there exists two ρ-networks f_{id}^D and $f_\times^D$, with constant width and depth, such that the absolute values of their weights is bounded respectively by $B(f_{id}^D) \lesssim KD^2$ and $B(f_\times^D) \lesssim K^2 D^6$ and such that,

$$\sup_{x \in [-D,D]} |f_{id}^D(x) - x| \lesssim 1/K \qquad \text{and} \qquad \sup_{x,y \in [-D,D]} |f_\times^D(x,y) - xy| \lesssim 1/K.$$

For $K \ge 1$ large enough and any $x \in [-D, D]$ we have $|f_{id}^D(x)| \le D + |f_{id}^D(x) - x| \le 2D$. We set

$$\tilde{f}_{ReLU}^{\rho,D}(x) := \frac{1}{b-a} \left[f_\times^{2D}(f_{id}^D(x), \rho(Kx)) - a f_{id}^D(x) \right],$$

which matches the expression (40) of the ReLU activation and gives the desired ρ-network. ∎

Using Lemma 30 we can transpose the ReLU-DNN approximations to ρ-DNN results.

Lemma 31 *Assume ρ is Lispchitz-continuous and locally quadratic. Let $f_0 : [0,1]^d \to \mathbb{R}$ be a function such that, there exists a large enough $N \geq 1$ and a ReLU-DNN, $\tilde{f} \in \mathcal{F}(L, (d, N, \ldots, N, 1))$ with length $L \asymp \log(N)$, satisfying*

$$||f_0 - \tilde{f}||_\infty \lesssim N^{-\eta},$$

where $\eta > 0$, and the weights of $\tilde{f}$ are bounded by N^c for some $c \geq 1$. Then there exists a ρ-DNN, $\tilde{f}^\rho$ with width $\tilde{N} \asymp N$ and depth $\tilde{L} \asymp L$, such that

$$||f_0 - \tilde{f}^\rho||_\infty \lesssim N^{-\eta},$$

with weights bounded by $e^{\tilde{c}\log^2 N}$, where $\tilde{c} \geq 1$ is a large enough constant depending only on c and η.

Proof By assumption we have a sequence of affine transformations $(A_l)_{1 \leq l \leq L+1}$ such that the ReLU-network realization $\tilde{f} \in \mathcal{F}(L, (d, N, \ldots, N, 1))$ is written

$$\tilde{f} := A_{L+1} \circ \text{ReLU} \circ A_L \circ \cdots \circ \text{ReLU} \circ A_1 \tag{41}$$

and is a good approximation of f_0. Here ReLU refers to the real valued function $x \mapsto x \vee 0$, we recall that in (41), the one-dimensional function ReLU is applied coordinate-wise to the output of the affine transformations A_l. Note that for any $x \in [0,1]^d$, we have, $|A_1 x|_\infty \leq (d+1)N^c \leq N^{c+1}$, for N large enough. Let $K \geq 1$, to be chosen (large enough) below and let $\tilde{f}_1^\rho$ be the ρ-DNN of Lemma 30 with precision K and $D = D_1 := N^{c+1}$, there is a constant $C > 0$ such that, for all $x \in [0,1]^d$, as N gets large enough,

$$|\tilde{f}_1^\rho \circ A_1(x) - \text{ReLU} \circ A_1(x)|_\infty \leq \frac{C}{K}.$$

By Lemma 30, weights of $\tilde{f}_1^\rho$ are bounded by $K^2 D^6$. Using $|\text{ReLU}(x)| \leq |x|$, we have

$$
\begin{aligned}
|A_2 \circ \tilde{f}_1^\rho \circ A_1(x)|_\infty &\leq (N+1)N^c|\tilde{f}_1^\rho \circ A_1(x)|_\infty \\
&\leq (N+1)N^c(|\text{ReLU} \circ A_1(x)|_\infty + C/K) \\
&\leq (N+1)N^c(D_1 + C/K) \\
&\leq 2N^{2(c+1)}(1 + 1/K) =: D_2.
\end{aligned}
$$

By recursion, we can set, for all $1 \leq l \leq L+1$, (with the convention $B_1^\rho = A_1$),

$$D_l := N^{l(c+1)}2^{l-1}(1+1/K)^{l-1} \quad \text{and} \quad B_l^\rho := A_l \circ \tilde{f}_{l-1}^\rho \circ A_{l-1} \circ \cdots \circ \tilde{f}_1^\rho \circ A_1, \tag{42}$$

such that the ρ-networks $\tilde{f}_l^\rho$ are generated by Lemma 30, their weights are bounded by $K^2 D_l^6$, and they satisfy

$$|\tilde{f}_l^\rho \circ B_l^\rho(x) - \text{ReLU} \circ B_l^\rho(x)|_\infty \leq \frac{C}{K} \quad \text{and} \quad |B_l(x)|_\infty \leq D_l. \tag{43}$$

Now define

$$\tilde{f}^\rho := B_{L+1} = A_{L+1} \circ \tilde{f}_L^\rho \circ A_L \circ \cdots \circ \tilde{f}_1^\rho \circ A_1. \tag{44}$$

Equation (44) is similar to the definition of $\tilde{f}$, given in equation (41), once one replaces the ReLU activations with their ρ-approximations. Recall that each ρ-network realization $\tilde{f}_l^\rho :$ $\mathbb{R} \to \mathbb{R}$ has constant width and depth and is applied coordinate wise to the N-dimensional outputs $B_{l-1}^\rho(x)$. Therefore the ρ-network $\tilde{f}_l$ has width $\tilde{N} \asymp N$ and depth $\tilde{L} \asymp L$ and weights bounded by $K^2 D_l^6$. To conclude it is enough to show that, taking K appropriately large, we get $||\tilde{f}^\rho - \tilde{f}||_\infty \lesssim N^{-\eta}$. Define, for any $1 \le l \le L$, (with convention $E_L = A_{L+1}$),

$$E_l := A_{L+1} \circ \mathrm{ReLU} \circ A_L \circ \cdots \circ \mathrm{ReLU} \circ A_{l+1}.$$

Since ReLU is 1-Lipschitz (with respect to the sup-norm) and A_l is N^{c+1}-Lipschitz, one easily checks that E_l is Λ_l-Lipschitz with $\Lambda_l \le N^{(c+1)(L+1-l)}$. Recall (42) the definition of D_l and B_l^ρ, writing $\tilde{f}^\rho - \tilde{f}$ as a telescopic sum and using (43), one gets

$$|\tilde{f}^\rho(x) - \tilde{f}(x)|_\infty = \left| \sum_{l=1}^L E_l \circ \tilde{f}_l^\rho \circ B_l^\rho(x) - E_l \circ \mathrm{ReLU} \circ B_l^\rho(x) \right|_\infty$$

$$\lesssim K^{-1} \sum_{l=1}^L \Lambda_l \le K^{-1} \sum_{l=1}^L N^{l(c+1)} \le \frac{N^{(L+1)(c+1)}}{K}.$$

Now, since $L \asymp \log N$, to get $||\tilde{f}^\rho - \tilde{f}||_\infty \lesssim N^{-\eta}$, it is sufficient to take K large enough, so that $N^{(L+1)(c+1)+\eta} \lesssim K$. We take $K := e^{\tilde{c}_1/2 \log^2 N}$ for $\tilde{c}_1 \ge 1$ large enough, depending only on c and η. The weights of $\tilde{f}^\rho$ are all bounded by $D_{L+1}^6 K^2 = N^{6(c+1)(L+1)} 2^L (1+1/K)^L e^{\tilde{c}_1 \log^2 N} \le e^{\tilde{c} \log^2 N}$, where $\tilde{c} \ge 1$ is a large enough constant. ∎

Theorem 32 *Consider ρ to be an admissible sigmoid activation according to definition 11. Suppose the conditions of Theorem 2 are satisfied, Π being a ρ-DNN prior on f, with depth $L = \lceil \log^{1+\delta} n \rceil$, width $\mathbf{r} = \mathbf{r}(\sqrt{n})$ and decay (σ_k) on the weights given by*

$$\log^{3(1+\delta)}(i \vee j) \le \log(1/\sigma_l^{(ij)}) \le \log^{3(1+\delta)} n.$$

Then for $\Pi_\alpha^B[\cdot|X,Y]$ the clipped posterior as defined above, for any $B > \|f_0\|_\infty$,

$$E_{f_0} \Pi_\alpha^B \left[\{f : \|f - f_0\|_{L^2(P_X)} \ge M\phi_n' \} \mid X, Y \right] \to 0,$$

where ϕ_n' is the rate given in (18) with $\gamma' = 3(1+\delta)(1+\kappa) + 1$.

Proof We check that the steps of the proof of Theorem 2 still hold. One start from the ReLU-approximation given by Lemma 33, this provides the exact setting to apply the transfer Lemma 31 with $\eta = 2\beta/d$. This shows that any Hölder function in $\mathcal{C}^\beta([0,1^d])$ can be approximated with precision $N^{-2\beta/d}$ with a ρ-DNN of depth $L \asymp \log N$, width $r \asymp N$ and weights bounded by $e^{\tilde{c} \log^2 N}$, for N large enough (here N plays the role of M^d in Lemma 33). We obtain an equivalent of Prop 14 for ρ-DNNs following the proof given in Appendix C.2, for $f_0 \in \mathcal{G}(q, \mathbf{d}, \mathbf{t}, \boldsymbol{\beta}, K)$ and ϕ_n' as in (18) with $\gamma' = 3(1+\delta)(1+\kappa)+1$, there is a ρ-DNN $\tilde{f}_0$ with depth $L = \lceil \log^{1+\delta} n \rceil$, width $\lceil \sqrt{n} \rceil$ and sparsity $s \lesssim n(\phi_n')^2 \log^{1-\gamma} n$ such that $||f_0 - \tilde{f}_0||_\infty \lesssim \phi_n'$, Additionally, all coefficients of $\tilde{f}_0$ satisfy $|\tilde{\theta}_k| \le e^{c \log^2 n}$, where $c \ge 1$

is a large enough constant. Note that the differences from Prop 14 is the magnitude of the weight increasing sub-exponentially as $\exp(c \log^2 n)$ (instead of polynomially as n^c) this induces some extra log-factors in the rate ϕ'_n and the decay σ_k. To conclude we check that the prior mass still holds under these new conditions. Notice that in the proof of Theorem 2 the magnitude bound is required first in equation (28) which then becomes

$$2e^{cL \log^2 n}(L+1)V \leq e^{C' \log^{3+\delta} n}$$

and equation (29) that involves the control of

$$C' \log^{3+\delta} n - C_0 \log^{1+\kappa}\left((1 + e^{c \log^2 n})(\sigma_l^{(ij)})^{-1}\right).$$

Taking (σ_k) such that $\log((\sigma_l^{(ij)})^{-1}) \leq \log^{3(1+\delta)} n$ shows that equation (29) becomes

$$A \geq \phi_n e^{-C_3 \log^{3(1+\delta)(1+\kappa)} n}.$$

The condition $\log^{3(1+\delta)}(i \vee j) \leq \log(1/\sigma_l^{(ij)})$ ensures the rest of the proof follows the one of Theorem 2. ∎

Appendix D. Bound on the coefficients

In this section we give the main steps of the proof of Kohler and Langer (2021) leading to Lemma 33. Here the goal is to track the bounds on the coefficients of the constructed network in order to obtain a quantitative upper-bound in terms of the number of parameters. For the individual construction of each subnetwork and their approximating capacity we refer to the proof of Kohler and Langer (2021).

The goal is to approximate $f \in \mathcal{C}_d^\beta([0,1]^d, F)$ with precision $M^{-2\beta}$ on hypercubes of size M^{-2} using local Taylor polynomials of f. The network used to find on which of these sub-cubes we are, will lead to the coefficients of highest magnitude through the important quantity

$$B_M := \lceil M^{2(\beta+1)} \rceil.$$

Note that, in the following, if g is the realization of a neural network with coefficients θ_k, we denote the scale of its coefficients as

$$B(g) := \max_k |\theta_k|.$$

Recall the Lemma

Lemma 33 (Theorem 2 in Kohler and Langer (2021)) *Let $f \in \mathcal{C}_d^\beta([0,1]^d, F)$ be a function of regularity $\beta > 0$, and let $M \geq 2$ be an integer such that the inequality*

$$M^{2\beta} \geq C(1 \vee F)^{4(\beta+1)}$$

holds for a sufficiently large constant $C \geq 1$.

Let $L, r \in \mathbb{N}$ satisfy

$$L \geq 5 + \lceil \log_4(M^{2\beta}) \rceil \left(\lceil \log_2(d \vee \lfloor \beta \rfloor + 1) \rceil + 1 \right)$$

and

$$r \geq 64 \binom{d + \lfloor \beta \rfloor}{d} 2^d d^2 (\lfloor \beta \rfloor + 1) M^d.$$

There exists a ReLU network $\tilde{f}_{wide} \in \mathcal{F}(L, (d, r, \ldots, r, 1))$ such that

$$||f - \tilde{f}_{wide}||_\infty \lesssim (1 \vee F)^{4(\beta+1)} M^{-2\beta}.$$

Proposition 34 Let $B_M := \lceil M^{2(\beta+1)} \rceil$, every coefficient θ of the network $\tilde{f}_{wide}$ of the previous Lemma satisfies

$$|\theta| \leq \max\{2(F \vee 1)e^{2d}, B_M^2\}.$$

Here are the step leading to Proposition 34.

D.1 Recursive definition of Taylor polynomials

Lemma 35 Let $f \in \mathcal{C}_d^\beta([0,1]^d, F)$ and $x_0 \in \mathbb{R}^d$. Define

$$T_{f,\beta,x_0}(x) := \sum_{j \in \mathbb{N}^d; |j| \leq \lfloor \beta \rfloor} (\partial^j f)(x_0) \frac{(x - x_0)^j}{j!},$$

then, for all $x \in \mathbb{R}^d$,

$$|f(x) - T_{f,\beta,x_0}(x)| \lesssim ||x - x_0||_\infty^\beta.$$

We conduct the following study on $[-1,1]^d$ to match Kohler and Langer (2021) results, the same can be done on the unit cube $[0,1]^d$. Set $\mathcal{P}_2 := \{C_{k,2}\}_{k \in [M^{2d}]}$ a partition of $[-1,1]^d$ into M^{2d} cubes and for all $x \in \mathbb{R}^d$, $(C_{\mathcal{P}_2}(x))_g$ the leftmost corner of the cube in $\mathcal{P}_2$ that contains x, from the previous lemma,

$$\forall x \in [-1,1]^d, \quad |f(x) - T_{f,\beta,(C_{\mathcal{P}_2}(x))_g}(x)| \lesssim M^{-2\beta}.$$

The crux of the proof is to approximate these polynomials by ReLU neural networks, to do so a recursive definition is given. Note that to compute $T_{f,\beta,(C_{\mathcal{P}_2}(x))_g}$ we need to compute about M^{2d} derivatives of f. Although it is possible to do so in a single layer of width M^{2d} a significant improvement here from sparse neural classes comes from the fact that between 2 layers of size M^d in a fully-connected network there are about M^{2d} connections. Therefore the derivatives are set as the weights of the networks, decreasing quadratically the width of the structure. To do this Kohler and Langer (2021) introduce another partition of broader scale, namely $\mathcal{P}_1 := \{C_{k,1}\}_{k \in [M^d]}$ a partition of $[-1,1]^d$ into M^d cubes.

Now for any $i \in [M^d]$, let $\{\tilde{C}_{j,i}\}_{j \in [M^d]}$ be the cubes of $\mathcal{P}_2$ that are in $C_{i,1}$. Thus $\mathcal{P}_2 = \{\tilde{C}_{i,j}\}_{1 \leq i,j \leq M^d}$ and for all $x \in \mathbb{R}^d$,

$$T_{f,\beta,(C_{\mathcal{P}_2}(x))_g}(x) = \sum_{1 \leq i,j \leq M^d} T_{f,\beta,(\tilde{C}_{i,j})_g}(x) \, \mathbf{1}_{\tilde{C}_{j,i}}(x) \tag{45}$$

In order to compute this as a neural network, the expression (45) of the local Taylor polynomial of f can be written recursively as follows:

1. first find in which broader cube of $\mathcal{P}_1$ the point x is located and compute derivatives of f on each smaller cube of size $1/M^2$ inside,

$$\phi_{1,1} := x,$$

$$\phi_{2,1} := \sum_{i=1}^{M^d} (C_{i,1})_g \, \mathbf{1}_{C_{i,1}}(x),$$

$$\phi_{3,1}^{(l,j)} := \sum_{i=1}^{M^d} (\partial^l f)((\tilde{C}_{j,i})_g) \, \mathbf{1}_{C_{i,1}}(x),$$

for all $j \in [M^d]$ and $|l| \leq \lfloor \beta \rfloor$.

2. then find the small cube of $\mathcal{P}_2$ containing x and gather the corresponding derivatives from the first step; to do this first note that for all $1 \leq k, i \leq M^d$,

$$v_k := (\tilde{C}_{k,i})_g - (C_{i,1})_g \tag{46}$$

takes values in $\{0, 2/M^2, \ldots, 2(M-1)/M^2\}$ and there is $j \in [M^d]$ such that,

$$C_{\mathcal{P}_2}(x) = \mathcal{A}^{(j)} := \{x \in \mathbb{R}^d, \ \phi_{2,1}^{(k)} + v_j^{(k)} \leq x^{(k)} < \phi_{2,1}^{(k)} + v_j^{(k)} + 2/M^2, \quad \forall k \in \{1, \ldots, d\}\}.$$

Then set,

$$\phi_{1,2} := x,$$

$$\phi_{2,2} := \sum_{j=1}^{M^d} (\phi_{2,1} + v_j) \, \mathbf{1}_{\mathcal{A}^{(j)}}(\phi_{1,1}),$$

$$\phi_{3,2}^{(l)} := \sum_{j=1}^{M^d} \phi_{3,2}^{(l,j)} \, \mathbf{1}_{\mathcal{A}^{(j)}}(\phi_{1,1}).$$

3. finally compute the local Taylor polynomial as

$$\phi_{1,3} := \sum_{j \in \mathbb{N}^d; |j| \leq \lfloor \beta \rfloor} \frac{\phi_{3,2}^{(j)}}{j!} (\phi_{1,2} - \phi_{2,2})^j.$$

Lemma 36 $\phi_{1,3} = T_{f,\beta,(C_{\mathcal{P}_2}(x))_g}(x)$

In the next step we construct an approximation of $\phi_{1,3}$ by ReLU neural networks. This approximation will be correct as soon as x is not too close to one of the boundaries of a $\mathcal{P}_2$ cube. More precisely, if C is a cube and δ is a strictly positive real number, let C_δ^0 be the set of points in C at a distance of at least δ from the boundary of C.

Lemma 37 *There is $\tilde{f} \in \mathcal{F}(L, (d, N, \ldots, N, 1))$ such that $\|\tilde{f}\|_\infty \leq 2(F \vee 1)e^{2d}$ and*

$$\forall x \in \bigcup_{j=1}^{M^{2d}} (C_{j,2})^0_{1/M^{2\beta+2}}, \quad |f(x) - \tilde{f}(x)| \lesssim M^{-2\beta}.$$

Furthermore, $L \asymp \log M$, $N \lesssim M^d$ and $B(\tilde{f}) \leq F \vee B_M^2$ with $B_M := \lceil M^{2\beta+2} \rceil$.

The boundary condition arises from the fact that through ReLU networks, only continuous piecewise affine functions can be realized. Therefore, it is necessary to approximate the indicator functions in the definition of $\phi_{1,3}$. Moreover, we will see that the accuracy of the approximation of these indicators controls the magnitude of the largest coefficients of the network.

D.2 Auxiliary networks for the construction of $\phi_{1,3}$

To approximate polynomial functions through ReLU networks the first step is to approximate the multiplication between two numbers, this construction comes from Schmidt-Hieber (2020)

Lemma 38 *For any positive integer R there is $\hat{f}_{mult} \in \mathcal{F}(R, 18)$ with $B(\hat{f}_{mult}) \leq 4$, such that,*

$$\forall x, y \in [-1, 1], \quad |\hat{f}_{mult}(x, y) - xy| \leq 4^{-R}.$$

Starting from Lemma 38, it is quite easy to provide an approximation result for polynomials in several variables. Let $N \in \mathbb{N}$, denote the set of d-variables polynomials of total degree at most N by,

$$\mathcal{P}_N := \text{Vect}\left(\prod_{k=1}^d (x^{(k)})^{r_k}, \ r_1 + \cdots + r_d \leq N \right).$$

Since $\dim \mathcal{P}_N = \binom{d+N}{d}$, write $m_1, \ldots, m_{\binom{d+N}{d}}$ all the monomials in $\mathcal{P}_N$. Let $r_1, \ldots, r_{\binom{d+N}{d}}$ be real numbers, for all $x \in [-1, 1]^d$ and $y_1, \ldots, y_{\binom{d+N}{d}} \in [-1, 1]$, set

$$\mathfrak{p}(x, y_1, \ldots, y_{\binom{d+N}{d}}) := \sum_{k=1}^{\binom{d+N}{d}} r_k \, y_k \, m_k(x).$$

Lemma 39 *Let $R \geq \log_4(2 \times 4^{2(N+1)})$ be an integer and $r_1, \ldots, r_{\binom{d+N}{d}}$ real numbers, set $\overline{r(\mathfrak{p})} := \max |r_k|$. There is a fully connected neural network of length $R\lceil \log_2(N+1) \rceil$ and width $18(N+1)\binom{d+N}{d}$ whose realisation $\hat{f}_\mathfrak{p}$ is such that $B(\hat{f}_\mathfrak{p}) \leq 4 \vee \overline{r(\mathfrak{p})}$, and,*

$$\left| \hat{f}_\mathfrak{p}(x, y_1, \ldots, y_{\binom{d+N}{d}}) - \mathfrak{p}(x, y_1, \ldots, y_{\binom{d+N}{d}}) \right| \lesssim \overline{r(\mathfrak{p})} 4^{-R},$$

for all $x \in [-1, 1]^d, y_1, \ldots, y_{\binom{d+N}{d}} \in [-1, 1]$.

As seen in the previous Section one also needs to approximate indicator functions, this is where the coefficients of highest magnitude will appear.

Lemma 40 *Let $R \in \mathbb{N}$ and $a, b \in \mathbb{R}^d$ such that for all $b^{(i)} - a^{(i)} \geq 2/R$ for all $i \in \{1, \ldots, d\}$. Let*

$$K_{1/R} := \left\{ x \in \mathbb{R}^d, \, x^{(i)} \notin [a^{(i)}, a^{(i)} + 1/R[\cup]b^{(i)} - 1/R, b^{(i)}], \text{ for all } i \in [d] \right\}.$$

1. *There is $\hat{f}_{ind[a,b[} \in \mathcal{F}(2, 2d)$ with $B(\hat{f}_{ind[a,b[}) \leq \max(R, 1/R, |a|_\infty, |b|_\infty)$ such that*

$$\forall x \in K_{1/R}, \quad \hat{f}_{ind[a,b[}(x) = \mathbf{1}_{[a,b[}(x),$$

   and

$$\forall x \in \mathbb{R}^d, \quad |\hat{f}_{ind[a,b[}(x) - \mathbf{1}_{[a,b[}(x)| \leq 1.$$

2. *Let $|s| \leq R$, there is $\hat{f}_{test}(., a, b, s) \in \mathcal{F}(2, 2(2d+2))$ with $B(\hat{f}_{test}) \leq \max(R^2, 1/R, |a|_\infty, |b|_\infty)$ such that*

$$\forall x \in K_{1/R}, \quad \hat{f}_{test}(x, a, b, s) = s\mathbf{1}_{[a,b[}(x),$$

   and

$$\forall x \in \mathbb{R}^d, \quad |\hat{f}_{test}(x, a, b, s) - s\mathbf{1}_{[a,b[}(x)| \leq |s|.$$

Also it is clear that one can realise the identity function of $\mathbb{R}^d$ with a ReLU network of arbitrary length, write f_{id}^ℓ such network of length $\ell \geq 2$. Choose $l_1, \ldots, l_{\binom{d+\lfloor\beta\rfloor}{d}}$ such that

$$\left\{ l_1, \ldots, l_{\binom{d+\lfloor\beta\rfloor}{d}} \right\} = \{(s_1, \ldots, s_d) \in \mathbb{N}^d, \, s_1 + \cdots + s_d \leq \lfloor\beta\rfloor\}.$$

Here are the networks deployed to approximate $\phi_{1,3}$, where for the $\hat{f}_{ind}$ and $\hat{f}_{test}$ networks we use Lemma 40 with $R = B_M := \lceil M^{2\beta+2} \rceil$. Let $\mathbf{J} := (1, \ldots, 1)^T \in \mathbb{R}^d$,

$$\hat{\phi}_{1,1} := f_{id}^2(x),$$

$$\hat{\phi}_{2,1} := \sum_{i=1}^{M^d} (C_{i,1})_g \, \hat{f}_{ind\, C_{i,1}}(x),$$

$$\hat{\phi}_{3,1}^{(l,j)} := \sum_{i=1}^{M^d} (\partial^l f)((\tilde{C}_{j,i})_g) \hat{f}_{ind\, C_{i,1}}(x),$$

$$\hat{\phi}_{1,2} := f_{id}^2(\hat{\phi}_{1,1}),$$

$$\hat{\phi}_{2,2}^{(k)} := \sum_{i=1}^{M^d} \hat{f}_{test}(\hat{\phi}_{1,1}, \hat{\phi}_{2,1} + v_i, \hat{\phi}_{2,1} + v_i + 2/M^2 \cdot \mathbf{J}, \hat{\phi}_{2,1}^{(k)} + v_i^{(k)}),$$

$$\hat{\phi}_{3,2}^{(l)} := \sum_{i=1}^{M^d} \hat{f}_{test}(\hat{\phi}_{1,1}, \hat{\phi}_{2,1} + v_i, \hat{\phi}_{2,1} + v_i + 2/M^2 \cdot \mathbf{J}, \hat{\phi}_{3,1}^{(l,i)}).$$

Now, for any $u \in \{1, \ldots, \binom{d+\lfloor \beta \rfloor}{d}\}$ set $y_u := \hat{\phi}_{3,2}^{(l_u)}$ and $r_u := 1/(l_u)!$, the approximating network of $\phi_{3,1}$ is given by Lemma 39 with

$$R = B_{M,\mathfrak{p}} := \lceil \log_4(2 \cdot 4^{2(\beta+1)} \cdot (2 \vee F)^{2(\beta+1)}) \rceil,$$

through

$$\hat{\phi}_{1,3} = \hat{f}_{\mathfrak{p}}(\hat{\phi}_{1,2} - \hat{\phi}_{2,2}, y_1, \ldots y_{\binom{d+\lfloor \beta \rfloor}{d}}).$$

The network $\hat{\phi}_{1,3}$ satisfies conditions of Lemma 37, we only provide the bound on the coefficients and refer to Kohler and Langer (2021) for a proof on the approximating properties. It is clear that $B(\hat{\phi}_{1,1}) = B(\hat{\phi}_{1,2}) = 1$. Using Lemma 40 one gets

$$B(\hat{\phi}_{2,1}) \leq \max \left\{ \max_i |(C_{i,1})_g|_\infty, B_M^2, 1/B_M \right\} \leq \max(2/M, B_M^2) \leq B_M^2,$$

and for all $l \in \{1, \ldots, \binom{d+\lfloor \beta \rfloor}{d}\}$ and $j \in [M^d]$,

$$B(\hat{\phi}_{3,1}^{(l,j)}) \leq \max \left\{ \max_i |(\partial^l f)((\tilde{C}_{j,i})_g)|, B_M^2, 1/B_M \right\} \leq F \vee B_M^2.$$

From the proof of Kohler and Langer (2021) it appears that the networks previously defined $(\hat{\phi}_{2,1}^{(k)}, \hat{\phi}_{3,1}^{(l,j)})$ are all in the set of bounded functions $\{g : ||g||_\infty \leq 1 \vee F\}$, thus using Lemma 40 one gets, for all $k \in \{1, \ldots, d\}$ and $l \in \{1, \ldots, \binom{d+\lfloor \beta \rfloor}{d}\}$,

$$B(\hat{\phi}_{2,2}^{(k)}) \leq F \vee B_M^2 \quad \text{and} \quad B(\hat{\phi}_{3,2}^{(l)}) \leq F \vee B_M^2.$$

One can check that $|\hat{\phi}_{1,2} - \hat{\phi}_{2,2}| \leq 2$ and $||\hat{\phi}_{3,2}||_\infty \leq F$. Now since $\overline{r(\mathfrak{p})} = \max |1/(l_u)!| \leq 1$ and $B_M \geq 4$ from $M \geq 2$, Lemma 39 yields,

$$B(\hat{\phi}_{1,3}) \leq \max\{F, B_M^2, B_{M,\mathfrak{p}}\} \leq F \vee B_M^2.$$

D.3 Approximation of $w_{\mathcal{P}_2}(x) \cdot f(x)$

The next step in the proof deals with handling the case where x is such that the approximation provided by Lemma 37 is no longer valid. To address this, define a weight function $w_{\mathcal{P}_2}$ for all $x \in \mathbb{R}^d$ by

$$w_{\mathcal{P}_2}(x) := \prod_{k=1}^d \left(1 - M^2 \left| (C_{\mathcal{P}_2}(x))_g^{(k)} + \frac{1}{M^2} - x^{(k)} \right| \right)_+. \tag{47}$$

This function reaches its maximum at the center of the cube of $\mathcal{P}_2$ containing x and is zero outside. Moreover, it takes very small values near the boundaries of the cube. More precisely, we have,

$$\text{for all } x \in \bigcup_{j=1}^{M^{2d}} C_{j,2} \setminus (C_{j,2})_{1/M^{2\beta+2}}^0, \quad w_{\mathcal{P}_2}(x) \leq M^{-2\beta}. \tag{48}$$

In this step we want to construct a network that will realise a good approximation of $w_{\mathcal{P}_2}(x) \cdot f(x)$ on the whole cube. In the final step we will aggregate these approximations using a partition of unity argument to get back to f.

Lemma 41 *There is $\tilde{f} \in \mathcal{F}(L,(d,N,\ldots,N,1))$ with $L \lesssim \log M$, $N \lesssim M^d$ and $B(\tilde{f}) \leq B_M^2 \vee F$, such that*

$$\forall x \in [-1,1]^d, \quad |\tilde{f}(x) - w_{\mathcal{P}_2}(x) \cdot f(x)| \lesssim M^{-2\beta}.$$

First approximate $w_{\mathcal{P}_2}$ with wide neural networks.

Lemma 42 *There is a network $\hat{f}_{w_{\mathcal{P}_2}} \in \mathcal{F}(L,(d,N,\ldots,N,1))$ such that $\|\hat{f}_{w_{\mathcal{P}_2}}\|_\infty \leq 2$ and*

$$\forall x \in \bigcup_{i=1}^{M^{2d}} (C_{i,2})^0_{1/M^{2\beta+2}}, \quad |\hat{f}_{w_{\mathcal{P}_2}}(x) - w_{\mathcal{P}_2}(x)| \leq M^{-2\beta},$$

with $L \lesssim \log M$, $N \lesssim M^d$ and $B(\hat{f}_{w_{\mathcal{P}_2}}) \leq B_M^2 \vee F$.

From (47) it is clear that such network exists using $\hat{\phi}_{2,2}$ to get $(C_{\mathcal{P}_2}(x))_g$ and $\hat{\phi}_{1,2}$ to get x. Those yield good approximations since x is "far" from the boundary of the cube and from previous calculations,

$$B(\hat{f}_{w_{\mathcal{P}_2}}) \leq \max\{4, B(\hat{\phi}_{1,2}), B(\hat{\phi}_{2,2}), M^2, 1/M^2\} \leq B_M^2 \vee F.$$

Lemmas 37 and 42 already show that there is a network that will approximate correctly $w_{\mathcal{P}_2}(x) \cdot f(x)$ whenever x is at least $M^{-2(\beta+1)}$ away from the boundaries of $C_{\mathcal{P}_2}(x)$. The idea now is to change a bit this network so that it is zero whenever x is too close to the boundaries. The resulting network will be a good approximation of $w_{\mathcal{P}_2}(x) \cdot f(x)$ on the whole cube because of (48). The following lemma defines a network that checks the relative position of x with respect to the boundaries of $C_{\mathcal{P}_2}(x)$.

Lemma 43 *There is $\hat{f}_{check\mathcal{P}_2} \in \mathcal{F}(5,(d,N,\ldots,N,1))$ such that $\|\hat{f}_{check\mathcal{P}_2}\|_\infty \leq 1$ and*

$$\forall x \notin \bigcup_{i=1}^{M^{2d}} (C_{i,2})^0_{1/M^{2\beta+2}} \setminus (C_{i,2})^0_{2/M^{2\beta+2}}, \quad \hat{f}_{check\mathcal{P}_2}(x) = \mathbf{1}\Big(\bigcup_{j=1}^{M^{2d}} C_{j,2} \setminus (C_{j,2})^0_{1/M^{2\beta+2}}\Big),$$

with $N \asymp M^d$ and $B(\hat{f}_{check\mathcal{P}_2}) \leq B_M^2$.

First find which cube of $\mathcal{P}_1$ contains x using Lemma 40 with $R = B_M$,

$$\hat{f}_1 := 1 - \sum_{i=1}^{M^d} \hat{f}_{ind(C_{i,1})^0_{1/M^{2\beta+2}}},$$

previous calculations have shown that $B(\hat{f}_1) \leq B_M^2$. Then find which smaller cube of $\mathcal{P}_2$ contains x using lemma 40 again with $R = B_M$. Recall $\mathbf{J} = (1,\ldots,1)^T \in \mathbb{R}^d$,

$$\hat{f}_2(x) := 1 - \sum_{j=1}^{M^d} \hat{f}_{test}(f_{id}^2(x), \hat{\phi}_{2,1} + v_j + M^{-2(\beta+1)} \cdot \mathbf{J}, \hat{\phi}_{2,1} + v_j + 2/M^2 \cdot \mathbf{J} + M^{-2(\beta+1)} \cdot \mathbf{J}, 1).$$

From previous calculations we know $B(\hat{f}_2) \leq \max\{B_M^2, B(\hat{\phi}_{2,1})\} \leq B_M^2$. Recall $\rho(x) = \max\{0, x\}$ and define

$$\hat{f}_{checkP_2}(x) := 1 - \rho(1 - \hat{f}_2(x) - f_{id}^2(\hat{f}_1(x))).$$

This network satisfies conditions of Lemma 43.

To construct a network satisfying the conditions of Lemma 41 let $\hat{f}_{\mathcal{P}_2}$ be the network defined by Lemma 37 and $B_{true} := 2(1 \vee F)e^{2d}$ be an upper bound on its supremum norm. Set

$$\hat{f}_{\mathcal{P}_2,true} := \rho(\hat{f}_{\mathcal{P}_2} - B_{true} \cdot \hat{f}_{checkP_2}) - \rho(-\hat{f}_{\mathcal{P}_2} - B_{true} \cdot \hat{f}_{checkP_2}),$$

this network is zero whenever x is less than $M^{-2(\beta+1)}$ close from the boundaries of $C_{\mathcal{P}_2}$ and otherwise it will be a good approximation of f. Thus $\tilde{f} = \hat{f}_{mult}(\hat{f}_{\mathcal{P}_2,true}, \hat{f}_{w\mathcal{P}_2})$ is always a good approximation of $w_{\mathcal{P}_2}(x) \cdot f(x)$ and satisfies conditions of Lemma 41. Indeed, we have

$$B(\hat{f}_{\mathcal{P}_2,true}) \leq \max\{B_{true}, B(\hat{f}_{\mathcal{P}_2}), B(\hat{f}_{checkP_2})\} \leq \max\{2(F \vee 1)e^{2d}, B_M^2\},$$

and

$$B(\tilde{f}) \leq \max\{B(\hat{f}_{\mathcal{P}_2,true}), B(\hat{f}_{w\mathcal{P}_2})\} \leq \max\{2(F \vee 1)e^{2d}, B_M^2\}.$$

D.4 Partition of unity through $w_{\mathcal{P}_2}$

For the final step, construct a partition of unity using the $w_{\mathcal{P}_2}$ weight functions. To do so shift the partitions $\mathcal{P}_1$ and $\mathcal{P}_2$ by shifting at least one of the coordinates by $1/M^2$. This leads to two sets of 2^d partitions $\{P_{1,v}, v \in \{1, \ldots, 2^d\}\}$ and $\{P_{2,v}, v \in \{1, \ldots, 2^d\}\}$. Enlarging the original hypercube if necessary the resulting family of weight functions $\{w_{\mathcal{P}_2,v}\}$ is a partition of unity on $[-1,1]^d$. If we set $\hat{f}_v$ to be the network given by Lemma 41 for the partition $\mathcal{P}_{2,v}$ then $f_{wide} := \sum_{v=1}^{2^d} \hat{f}_v$ yields a good approximation of $f = \sum_v w_{\mathcal{P}_2,v} f$ and satisfies all the conditions of Lemma 33, in particular

$$B(f_{wide}) \leq \max\{2(F \vee 1)e^{2d}, B_M^2\}.$$

References

S. Agapiou and I. Castillo. Heavy-tailed Bayesian nonparametric adaptation. *Ann. Statist.*, 52(4):1433–1459, 2024. ISSN 0090-5364,2168-8966. doi: 10.1214/24-aos2397. URL https://doi.org/10.1214/24-aos2397.

S. Agapiou, M. Dashti, and T. Helin. Rates of contraction of posterior distributions based on p-exponential priors. *Bernoulli*, 27(3):1616–1642, 2021. ISSN 1350-7265. doi: 10.3150/20-bej1285. URL https://doi.org/10.3150/20-bej1285.

S. Agapiou, I. Castillo, and P. Egels. Heavy-tailed and horseshoe priors for regression and sparse besov rates. *arXiv preprint arXiv:2505.15543*, 2025.

P. Alquier. User-friendly introduction to PAC-bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303, 2024.

P. Alquier and J. Ridgway. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475 – 1497, 2020.

J. Arbel, K. Pitas, M. Vladimirova, and V. Fortuin. A primer on Bayesian neural networks: Review and debates, 2023. arXiv eprint 2309.16314.

J. Bai, Q. Song, and G. Cheng. Efficient variational inference for sparse deep learning with theoretical guarantee. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

C. Berenfeld, P. Rosa, and J. Rousseau. Estimating a density near an unknown manifold: a Bayesian nonparametric approach. *Ann. Statist.*, 52(5):2081–2111, 2024. ISSN 0090-5364,2168-8966. doi: 10.1214/24-aos2423. URL https://doi.org/10.1214/24-aos2423.

A. Bhattacharya, D. Pati, and Y. Yang. Bayesian fractional posteriors. *The Annals of Statistics*, 47(1):39 – 66, 2019.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(5):1103–1130, 2016.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, Apr. 2017.

S. P. Brooks, P. Giudici, and G. O. Roberts. Efficient construction of reversible jump markov chain monte carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):3–39, 2003.

I. Castillo. *Bayesian nonparametric statistics, St Flour lecture notes*, volume LI. Springer, 2024.

I. Castillo and T. Randrianarisoa. Deep Horseshoe Gaussian Processes. 2024. arXiv eprint 2403.01737.

M. Chen, H. Jiang, W. Liao, and T. Zhao. Efficient approximation of deep ReLU networks for functions on low dimensional manifolds. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

B.-E. Chérief-Abdellatif. Convergence rates of variational inference in sparse deep learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1831–1842, 13–18 Jul 2020.

A. Damianou and N. D. Lawrence. Deep Gaussian processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 207–215, 29 Apr–01 May 2013.

V. C. Dinh, L. S. T. Ho, and C. V. Nguyen. Hamiltonian monte carlo on reLU neural networks is inefficient. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879 – 921, 1998.

Y. Fang, I. Ohn, V. Gupta, and L. Lin. Intrinsic and extrinsic deep learning on manifolds. *Electronic Journal of Statistics*, 18(1):1160 – 1184, 2024.

G. Finocchio and J. Schmidt-Hieber. Posterior contraction for deep Gaussian process priors. *Journal of Machine Learning Research*, 24(66):1–49, 2023.

V. Fortuin. Priors in bayesian deep learning: A review. *International Statistical Review*, 90 (3):563–591, 2022.

V. Fortuin, A. Garriga-Alonso, S. W. Ober, F. Wenzel, G. Rätsch, R. E. Turner, M. van der Wilk, and L. Aitchison. Bayesian neural network priors revisited. *The Tenth International Conference on Learning Representations, ICLR*, 2022.

S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500 – 531, 2000.

S. Ghosh, J. Yao, and F. Doshi-Velez. Model selection in Bayesian neural networks via horseshoe priors. *Journal of Machine Learning Research*, 20(182):1–46, 2019.

E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge university press, 2015.

P. Grünwald. The Safe Bayesian: Learning the learning rate via the mixability gap. *Lecture Notes in Comput. Sci.*, 7568:169–183, 2012.

C. C. Holmes and S. G. Walker. Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503, 2017.

J. Hron, R. Novak, J. Pennington, and J. Sohl-Dickstein. Wide bayesian neural networks have a simple weight posterior: theory and accelerated sampling. In *International conference on machine learning*, pages 8926–8945. PMLR, 2022.

Y. Jiao, G. Shen, Y. Lin, and J. Huang. Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691–716, 2023.

M. Kohler and S. Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231 – 2249, 2021.

M. Kohler, A. Krzyżak, and S. Langer. Estimation of a function of low local dimensionality by deep neural networks. *IEEE transactions on information theory*, 68(6):4032–4042, 2022.

I. Kong and Y. Kim. Posterior concentrations of fully-connected Bayesian neural networks with general priors on the weights, 2024. arxiv eprint 2403.14225.

I. Kong, D. Yang, J. Lee, I. Ohn, G. Baek, and Y. Kim. Masked Bayesian neural networks: Theoretical guarantee and its posterior inference. In *International Conference on Machine Learning*, pages 17462–17491. PMLR, 2023.

H. Lee, F. Ayed, P. Jung, J. Lee, H. Yang, and F. Caron. Deep neural networks with dependent weights: Gaussian Process mixture limit, heavy tails, sparsity and compressibility, 2023. Arxiv eprint 2205.08187.

K. Lee and J. Lee. Asymptotic properties for Bayesian neural network in Besov space. In *Advances in Neural Information Processing Systems*, volume 35, pages 5641–5653, 2022.

A. L'Huillier, L. Travis, I. Castillo, and K. Ray. Semiparametric inference using fractional posteriors. *The Journal of Machine Learning Research*, 24(1):18619–18679, 2023.

J. Lu, Z. Shen, H. Yang, and S. Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

S. P. Lyddon, C. C. Holmes, and S. G. Walker. General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478, 2019.

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

P. Marion, A. Fermanian, G. Biau, and J.-P. Vert. Scaling resnets in the large-depth regime. 2022. arXiv eprint arXiv:2206.06929.

R. Nakada and M. Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174): 1–38, 2020.

R. M. Neal. *Priors for Infinite Networks*, pages 29–53. Springer New York, New York, NY, 1996.

R. M. Neal. *MCMC using Hamiltonian dynamics*, pages 113–162. 2011.

I. Ohn and Y. Kim. Smooth function approximation by deep neural networks with general activation functions. *Entropy*, 21(7), 2019.

I. Ohn and L. Lin. Adaptive variational Bayes: Optimality, computation and applications. *The Annals of Statistics*, 52(1):335–363, 2024.

T. Papamarkou et al. Position paper: Bayesian deep learning in the age of large-scale AI. 2024. arxiv eprint 2402.00809.

L. Pezzetti, S. Favaro, and S. Peluchetti. Function-space mcmc for bayesian wide neural networks. 2025. arXiv eprint arXiv:2408.14325.

N. G. Polson and V. Ročková. Posterior concentration for sparse deep learning. *Advances in Neural Information Processing Systems*, 31, 2018.

J. Schmidt-Hieber. Deep ReLU network approximation of functions on a manifold, 2019. arXiv eprint 1908.00695.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020.

T. Sell and S. S. Singh. Trace-class Gaussian priors for Bayesian learning of neural networks with MCMC. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):46–66, 01 2023.

T. Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *7th International Conference on Learning Representations, ICLR 2019*, 2018.

T. Suzuki and A. Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. *Advances in Neural Information Processing Systems*, 34:3609–3621, 2021.

N. Syring and R. Martin. Calibrating general posterior credible regions. *Biometrika*, 106 (2):479–486, 2019.

H. Triebel. *Theory of Function Spaces*. Birkhäuser Basel, 1983.

A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463, 2008. ISSN 0090-5364. doi: 10.1214/009053607000000613. URL https://doi.org/10.1214/009053607000000613.

T. van Erven and P. Harremoes. Rényi divergence and Kullback–Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

M. Vladimirova, J. Verbeek, P. Mesejo, and J. Arbel. Understanding priors in bayesian neural networks at the unit level. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, 2019.

F. Wenzel, K. Roth, B. S. Veeling, J. Świątkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowozin. How good is the bayes posterior in deep neural networks really? *International Conference on Machine Learning*, 2020.

Y. Yang, D. Pati, and A. Bhattacharya. α-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886 – 905, 2020.

D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.

F. Zhang and C. Gao. Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180–2207, 2020.