# Posterior Contraction Rates for Sparse Kolmogorov-Arnold Networks in Anisotropic Besov Spaces

**Jeunghun Oh**[1] **Jaeyong Lee**[1] **Lizhen Lin**[2] **Kyeongwon Lee**[3*]
[1]Department of Statistics, Seoul National University
[2]Department of Mathematics, University of Maryland
[3]Department of Statistics, Inha University
nomad1994@snu.ac.kr leejyc@gmail.com
lizhen01@umd.edu kwlee@inha.ac.kr

## Abstract

We study posterior contraction rates for sparse Bayesian Kolmogorov-Arnold networks (KANs) over anisotropic Besov spaces, providing a statistical foundation for KANs from a Bayesian point of view. We show that sparse Bayesian KANs equipped with spike-and-slab-type sparsity priors attain a near-minimax posterior contraction rate. In particular, the contraction rate depends on the intrinsic anisotropic smoothness of the underlying function. Moreover, by placing a hyperprior on a single model-size parameter, the resulting posterior adapts to unknown anisotropic smoothness and still achieves the corresponding near-minimax rate. A distinctive feature of our results, compared with those for standard multilayer perceptrons, is that the KAN depth can be kept fixed: owing to the expressive power of learnable spline edge functions, the required approximation complexity is controlled through the network width, the range and size of the spline grid, and parameter sparsity. Our analysis develops theoretical tools tailored to sparse spline-edge architectures, including approximation and complexity bounds for Bayesian KANs. We then extend to compositional Besov spaces and show that the contraction rates depend on *layerwise smoothness and effective dimension* of the underlying compositional structure, thereby effectively avoiding the curse of dimensionality. Together, these results advance the theoretical understanding of Bayesian neural networks and establish rigorous Bayesian guarantees for KANs.

## 1 Introduction

Kolmogorov-Arnold Networks are a recently introduced alternative to multilayer perceptrons (MLPs) (Liu et al., 2025, 2024). This design is motivated by the Kolmogorov-Arnold representation theorem, which states that any continuous multivariate function can be expressed through sums and compositions of suitable univariate functions. By representing the edges of the network as linear combinations of univariate B-splines, KANs offer greater modeling flexibility than conventional fully connected neural networks that use fixed activation functions. KANs are also more interpretable than MLPs, since each edge function can be visualized and interpreted directly. Motivated by these properties, KANs have recently been applied in a wide range of domains, including time series analysis (Vaca-Rubio et al., 2024), graph learning (Kiamari et al., 2024; Bresson et al., 2025), and differential equations (Koenig et al., 2024).

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In high-dimensional nonparametric regression, statistical performance is strongly affected by both the ambient dimension and the regularity structure of the underlying regression function. The deterioration of estimation rates with dimension, the so-called curse of dimensionality, is a familiar manifestation of this dependence. In addition, when the smoothness of the target function varies across coordinates or the function exhibits local irregularities, isotropic smoothness classes such as Hölder or Sobolev spaces may fail to capture the effective complexity of the problem. Anisotropic Besov spaces offer a flexible alternative by encoding direction-dependent regularity through a smoothness vector (Kerkyacharian et al., 2001; Suzuki and Nitanda, 2021); the corresponding contraction rates in this paper are governed by the intrinsic smoothness which will be presented later.

Recent works have established near-minimax guarantees for Bayesian deep learning procedures over anisotropic function spaces. For fractional posteriors, Egels and Castillo (2025) proved adaptive near-minimax contraction rates for deep ReLU networks with heavy-tailed priors over several function classes, including anisotropic Besov spaces. For standard posteriors, Lee et al. (2025) showed that sparse Bayesian neural networks equipped with sparsity-inducing priors attain near-optimal contraction rates over anisotropic Besov and composite anisotropic Besov spaces, together with adaptive results through hierarchical priors. For related results on (near-)minimax guarantees for multilayer perceptrons (MLPs), see Suzuki (2018); Schmidt-Hieber (2020); Suzuki and Nitanda (2021) for frequentist approaches and Polson and Ročková (2018); Lee and Lee (2022); Kong and Kim (2025) for Bayesian approaches.

As mentioned, the accumulated body of works has been developed for ReLU-based MLPs, and it remains unclear whether analogous results can be established for spline-edge architectures such as KANs. Recently, Kratsios et al. (2026) established optimal approximation rates for residual KANs in Besov norms, together with learning guarantees based on complexity bounds, providing theoretical support for the expressive power of KANs. However, to the best of our knowledge, the rigorous theoretical support for Bayesian KANs with respect to the posterior contraction rates has not previously been established. We aim to close such gaps in recent literature by proposing sparse Bayesian KANs and providing the rigorous theory.

One of the key features of our work is that fixed-depth KANs attain the approximation and near-optimal posterior contraction rates. However, in the aforementioned studies on ReLU-based MLPs, the depth of the networks in order to obtain theoretical guarantees should grow with the sample size, typically, at an order of $\log n$. Therefore, our conclusion suggests one of promising features of KANs for expressive efficiency (Wang et al., 2025).

To summarize, our main contributions are threefold. First, we develop fundamental results for sparse Bayesian KANs. We prove that sparse KANs approximate functions over anisotropic Besov spaces. Then, we propose sparse Bayesian KANs with sparsity-inducing priors including spike-and-slab and Bernoulli spike-and-slab and prove that KANs attain near-minimax posterior contraction rates, with the rate depending on the intrinsic smoothness; our construction keeps the KAN depth fixed and controls approximation complexity through sparse spline-edge representations rather than depth growth or increasingly fine grid spacing. Second, we establish adaptation to unknown anisotropic smoothness by placing a hyperprior on a single model-size parameter that jointly controls the width, grid size, and grid range of the KAN architecture. Third, we extend the approximation theorem to compositional anisotropic Besov spaces, obtaining rates governed by layerwise smoothness and effective dimension.

## 2 Model assumptions and setup

**Notation.** Let $\mathbb{N}$ and $\mathbb{R}$ denote the sets of natural numbers and real numbers, respectively. Moreover, let $\mathbb{R}_+$ denote the set of nonnegative real numbers and $\mathbb{R}_{++}$ the set of positive real numbers. For $a \in \mathbb{R}$, let $\lfloor a \rfloor$ and $\lceil a \rceil$ be the floor and ceiling functions. For $n \in \mathbb{N}$, let $[n]$ denote the set $\{1, 2, \ldots, n\}$. For each $a, b \in \mathbb{R}$, write $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. For a real-valued vector $\mathbf{v}$, denote its $\ell_p$-norm by $\|\mathbf{v}\|_p$, $1 \leq p \leq \infty$, and let $\|\mathbf{v}\|_0$ be the number of nonzero coordinates. For a measurable function $f : [0,1]^d \to \mathbb{R}$ and a measure $\mu$ on $[0,1]^d$ with $d \in \mathbb{N}$, define $\|f\|_{L^p(\mu)} := (\int_{[0,1]^d} |f(x)|^p d\mu(x))^{1/p}$ for $0 < p < \infty$. When $p = \infty$, define $\|f\|_\infty := \sup_{x \in [0,1]^d} |f(x)|$. Let $\delta_0$ denote the point mass at 0. For two sequences $a_n, b_n \geq 0$, if $a_n \leq C b_n$ holds for some universal constant $C > 0$, then we write $a_n \lesssim b_n$; equivalently, we write $b_n \gtrsim a_n$. If $a_n \lesssim b_n \lesssim a_n$, then we write $a_n \asymp b_n$. For convenience, define the uniformly bounded

function class $\mathcal{UB} := \{f : [0,1]^d \to \mathbb{R} : \|f\|_\infty \leq 1\}$. For a normed space $\mathcal{F}$, let $U(\mathcal{F})$ denote the unit ball of $\mathcal{F}$. Let $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)$ denote the $\epsilon$-covering number of the set $\mathcal{F}$ with respect to the norm $\|\cdot\|$.

**Regression model.** We observe data $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ with $(X_i, Y_i) \in [0,1]^d \times \mathbb{R}$, from the following random design nonparametric regression model:

$$Y_i = f_0(X_i) + \varepsilon_i, \qquad Y_i \,|\, X_i \overset{ind}{\sim} N(f_0(X_i), \sigma_0^2), \qquad X_i \overset{iid}{\sim} P_X$$

where $f_0 : [0,1]^d \to \mathbb{R}$ is the true regression function, and $\sigma_0^2 > 0$ is the variance of the observation noise. The distribution $P_X$ denotes the distribution of the design points $X_i$.

### 2.1 Kolmogorov–Arnold Networks

#### 2.1.1 B-splines

Following the original formulation of Liu et al. (2025), we assume that each edge in the network is given by a sum of a B-spline expansion and a SiLU activation. B-spline bases admit several constructions depending on the knot, or grid, specification such as allowing repeated or nonuniform knots (Schumaker, 2007). In this work, we restrict attention to *fixed-knot KANs*, whose edges are built from B-splines on a fixed, equally spaced knot grid with no repeated knots.

Specifically, let $m \in \mathbb{N}$ be the spline degree and let $G \in \mathbb{N}$ be the number of grid intervals. For an extended interval $[a, b]$, define a knot vector $\boldsymbol{\xi} = (\xi_{-m}, \cdots, \xi_{G+m}) \in [a,b]^{G+2m+1}$ whose entries are strictly increasing and equally spaced, i.e. $\xi_k < \xi_{k+1}$ and $\Delta = \xi_{k+1} - \xi_k$, $\forall k$. We denote by $\{B_{k,m}(\cdot; \boldsymbol{\xi})\}_{k=1}^{G+m}$ the collection of degree $m$ B-spline basis functions generated by $\boldsymbol{\xi}$ (De Boor, 1978). We take $[\xi_0, \xi_G]$ as the estimation domain, while the extended interval $[a, b]$ is used to mitigate boundary effects.

#### 2.1.2 KAN specification

We proceed with introducing the KAN architecture by defining each edge of KANs. Let $L$ denote the depth of the network, $\mathbf{D} = (D_0, \ldots, D_L)$ its layer widths, and $\mathbf{G} = (G_0, \ldots, G_{L-1})$ the per-layer numbers of grid intervals, with $D_0 = d$ and $D_L = 1$. For each $l = 0, \ldots, L-1$, the edge function $\phi_{i,j}^{(l)} : \mathbb{R} \to \mathbb{R}$ connecting node $j \in [D_l]$ in layer $l$ to node $i \in [D_{l+1}]$ in layer $l+1$ is defined by

$$\phi_{i,j}^{(l)}(x) = \sum_{k=1}^{G_l+m} \theta_{i,j,k}^{(l)} B_{k,m}(x; \boldsymbol{\xi}^{(l)}) + \theta_{i,j,0}^{(l)} \,\mathrm{SiLU}(x), \quad x \in \mathbb{R} \tag{1}$$

where $B_{k,m}(x; \boldsymbol{\xi}^{(l)})$ denotes the B-spline basis generated by the layer $l$ knots $\boldsymbol{\xi}^{(l)} \in [a_l, b_l]^{G_l+2m+1}$, and $\mathrm{SiLU}(x) = x/(1+e^{-x})$. We set $[a_l, b_l] = [-H, H]$, $l = 1, \ldots, L-1$ for $H > 0$ and if $l = 0$, any fixed points $a_0 < \xi_0^{(0)} < \xi_{G_0}^{(0)} < b_0$ such that $[\xi_0^{(0)}, \xi_{G_0}^{(0)}]^d \supseteq [0,1]^d$. Each $\theta_{i,j,k}^{(l)} \in \mathbb{R}$ is the corresponding edge weight, and we collect the weights of edge $(i,j)$ at layer $l$ into the vector $\boldsymbol{\theta}_{i,j}^{(l)} = (\theta_{i,j,0}^{(l)}, \ldots, \theta_{i,j,G_l+m}^{(l)}) \in \mathbb{R}^{G_l+m+1}$. The transformation $\Phi_l : \mathbb{R}^{D_l} \to \mathbb{R}^{D_{l+1}}$ at layer $l$ is then given by

$$\Phi_l(\mathbf{x}) = \Big( \sum_{j=1}^{D_l} \phi_{1,j}^{(l)}(x_j), \ldots, \sum_{j=1}^{D_l} \phi_{D_{l+1},j}^{(l)}(x_j) \Big)^\top \tag{2}$$

with weight tensor $\mathbf{W}^{(l)} = (\boldsymbol{\theta}_{i,j}^{(l)})_{i,j} \in \mathbb{R}^{D_{l+1} \times D_l \times (G_l+m+1)}$. The Kolmogorov-Arnold network (KAN) $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}$ is then defined as $f_{\boldsymbol{\theta}} = \Phi_{L-1} \circ \cdots \circ \Phi_0$ where $\boldsymbol{\theta} = (\mathbf{W}^{(l)})_l = (\theta_{i,j,k}^{(l)})_{i,j,k,l}$ collects all the parameters; we also use the vectorized representation $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)$ when convenient. The total number of parameters is $T(L, \mathbf{D}, \mathbf{G}) = \sum_{l=0}^{L-1} D_{l+1} D_l (G_l + m + 1)$. For a fixed structure $(L, \mathbf{D}, \mathbf{G}, H)$, where $L \in \mathbb{N}$, $\mathbf{D} \in \mathbb{N}^{L+1}$, $\mathbf{G} \in \mathbb{N}^L$, $H > 0$, and parameters $\boldsymbol{\theta} = (\theta_{i,j,k}^{(l)})_{i,j,k,l}$, we define the class of KANs of order $m$ :

$$KAN(L, \mathbf{D}, \mathbf{G}, H; m) := \left\{ f_{\boldsymbol{\theta}} = \Phi_{L-1} \circ \cdots \circ \Phi_0 \;\middle|\; \begin{array}{l} \boldsymbol{\theta} = (\mathbf{W}^{(l)})_l \text{ such that} \\ \Phi_l \text{ is defined by Eq. (1)-Eq. (2),} \\ \mathbf{W}^{(l)} \in \mathbb{R}^{D_{l+1} \times D_l \times (G_l+m+1)}, \ \forall l \end{array} \right\}. \tag{3}$$

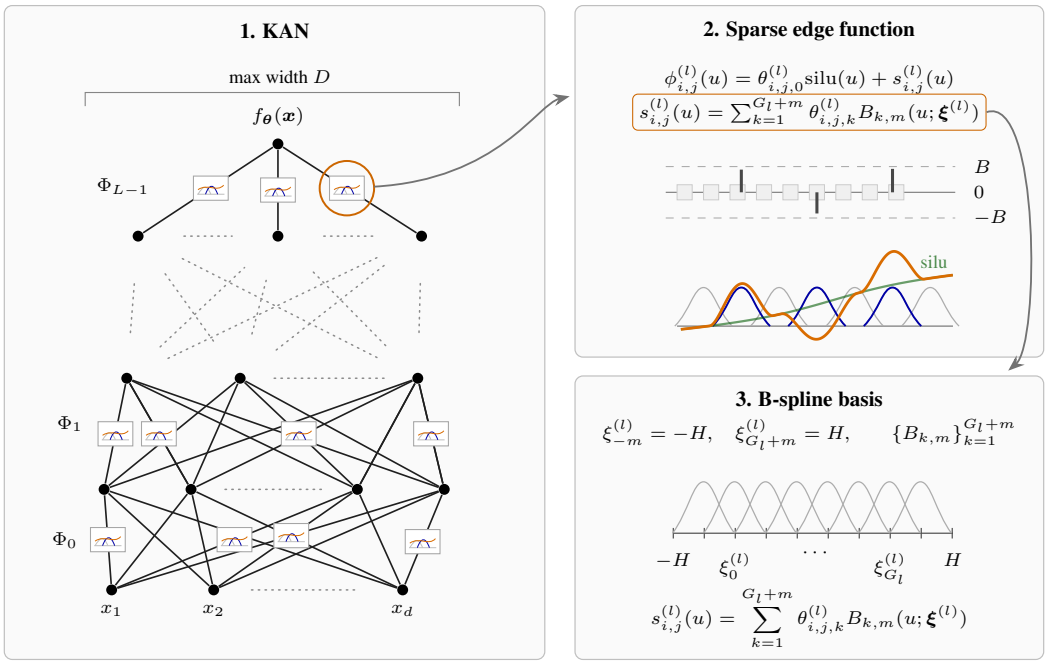



Figure 1: Sparse fixed-knot B-spline KAN construction. The left panel shows a KAN with layer widths bounded by $D$; edges carry learnable univariate functions. The upper-right panel zooms in on one edge function with coefficient-level sparsity, the uniform coefficient bound $B$, and a SiLU component. The lower-right panel zooms further into the B-spline basis used in the spline part of the edge function.

For the posterior contraction analysis, we further specialize to KANs with equal hidden width and a uniform fixed-knot grid: for $D, G \in \mathbb{N}$ and $H > 0$,

$$\mathbf{D} = (d, D, \ldots, D, 1), \quad \mathbf{G} = (G_0, G, \ldots, G), \quad [a_l, b_l] = [-H, H] \text{ for } l = 1, \ldots, L-1$$

where $G_0$ is fixed. For notational simplicity, we write KANs with equal-width and uniform fixed-knot structure with parameter $\boldsymbol{\theta}$ as $f_{\boldsymbol{\theta}} \in KAN(L, D, G, H)$, dropping the boldface to indicate $D$ and $G$ are now scalars. Under this setting, the total number of parameters is given by $T = T(L, D, G) = dD(G_0 + m + 1) + ((L-2)D^2 + D)(G + m + 1)$.

### 2.2 Anisotropic Besov spaces

In this section, we introduce anisotropic Besov spaces, the function class to which we assume the true function belongs. The anisotropic modulus of smoothness is defined similarly to that of the isotropic Besov space case, but by taking each direction into account (Nikol'skii, 1975). For a function $f : \mathbb{R}^d \to \mathbb{R}$, $r$th-order finite difference of $f$ in direction $\mathbf{h} \in \mathbb{R}^d$ is defined by $\Delta_h^r(f)(\mathbf{x}) = \sum_{j=0}^r \binom{r}{j}(-1)^{r-j} f(\mathbf{x} + j\mathbf{h})$ if $\mathbf{x} + r\mathbf{h} \in [0,1]^d$ or $\Delta_h^r(f)(\mathbf{x}) = 0$ otherwise. Given a directional radius $\mathbf{t} = (t_1, \ldots, t_d) \in \mathbb{R}_{++}^d$, the anisotropic modulus of smoothness of a function $f \in L^p$ is defined by $w_{r,p}(f, t) = \sup_{\mathbf{h} \in \mathbb{R}^d : |h_i| \le t_i} \|\Delta_h^r(f)\|_{L^p}$.

**Definition 2.1** (Anisotropic Besov Space: $\mathcal{B}_{p,q}^{\mathbf{s}}([0,1]^d)$)**.** *For $p, q \in (0, \infty]$ and $\mathbf{s} = (s_1, \ldots, s_d) \in \mathbb{R}_{++}^d$, let $r := \max_i \lfloor s_i \rfloor + 1$. For a function $f \in L^p$, the anisotropic Besov seminorm $|\cdot|_{\mathcal{B}_{p,q}^{\mathbf{s}}}$ is*

$$|f|_{\mathcal{B}_{p,q}^{\mathbf{s}}} := \begin{cases} \left(\sum_{k=0}^{\infty} [2^k w_{r,p}(f, (2^{-k/s_1}, \ldots, 2^{-k/s_d}))]^q\right)^{1/q}, & q < \infty \\ \sup_{k \ge 0} [2^k w_{r,p}(f, (2^{-k/s_1}, \ldots, 2^{-k/s_d}))], & q = \infty. \end{cases}$$

*The norm on the space $\mathcal{B}_{p,q}^{\mathbf{s}}([0,1]^d)$ is defined by $\|f\|_{\mathcal{B}_{p,q}^{\mathbf{s}}} := \|f\|_{L^p} + |f|_{\mathcal{B}_{p,q}^{\mathbf{s}}}$, and $\mathcal{B}_{p,q}^{\mathbf{s}}([0,1]^d) := \{f \in L^p : \|f\|_{\mathcal{B}_{p,q}^{\mathbf{s}}} < \infty\}$.*

The anisotropic smoothness vector $\mathbf{s}$ can be summarized by a single *intrinsic smoothness* index $\tilde{\mathbf{s}} := (\sum_{j=1}^d s_j^{-1})^{-1}$, which captures the joint regularity of $f$ across all coordinates and governs the

minimax convergence rate. Let $\underline{\mathbf{s}} := \min_j s_j$ be the smallest smoothness among all directions, and $\overline{\mathbf{s}} := \max_j s_j$ the largest smoothness. Intuitively, a small $s_j$ indicates that $f$ is rough along the $j$th coordinate, while a large $s_j$ indicates higher regularity or smoothness in that direction. The isotropic case $s_1 = \cdots = s_d = s$ recovers the standard isotropic Besov space, in which case $\tilde{\mathbf{s}} = s/d$.

**Remark 2.2** (Intrinsic smoothness and intrinsic dimension)**.** *The intrinsic smoothness $\tilde{\mathbf{s}}$ can be equivalently expressed through the anisotropy-induced dimension $d_{\mathrm{int}}(\mathbf{s}) := \underline{\mathbf{s}}/\tilde{\mathbf{s}} = \sum_{j=1}^{d} \underline{\mathbf{s}}/s_j \le d$. Indeed, the classical anisotropic minimax rate (Kerkyacharian et al., 2001; Hoffman and Lepski, 2002) satisfies $n^{-\tilde{\mathbf{s}}/(2\tilde{\mathbf{s}}+1)} = n^{-\underline{\mathbf{s}}/(2\underline{\mathbf{s}}+d_{\mathrm{int}}(\mathbf{s}))}$. Thus $d_{\mathrm{int}}(\mathbf{s})$ is the dimension index induced by anisotropy: a roughest coordinate contributes one full dimension, while a smoother coordinate contributes only the fraction $\underline{\mathbf{s}}/s_j$. In the isotropic case $d_{\mathrm{int}}(\mathbf{s}) = d$, and the layerwise quantity $t^{*(j)} = \underline{\mathbf{s}}^{(j)}/\tilde{s}^{(j)}$ below is the same index for each compositional component.*

## 3 Main results

In this section, we present posterior contraction rate results for sparse Bayesian KANs over anisotropic Besov spaces. For the main results, we impose the following assumptions.

**Assumptions.**

- **(A1)** For some $0 < p, q \le \infty$ and $\mathbf{s} \in \mathbb{R}^d_{++}$, the true function satisfies $f_0 \in U(\mathcal{B}^{\mathbf{s}}_{p,q}) \cap \mathcal{UB}$ and $(1/p - 1/2)_+ < \tilde{\mathbf{s}}$. In the likelihood,
$$Y_i | X_i \sim N(\mathrm{clip} \circ f(X_i), \sigma^2),$$
where $\mathrm{clip}(x) := \min\{1, \max\{-1, x\}\}$.
- **(A2)** The design distribution $P_X$ is supported on $[0,1]^d$ and admits a bounded density function $p_X$, i.e. there exists a constant $C_X > 0$ such that $\|p_X\|_{L^\infty} \le C_X$.
- **(A3)** There exist constants $0 < \underline{\sigma} < \overline{\sigma} < \infty$ such that $\underline{\sigma} < \sigma_0 < \overline{\sigma}$.
- **(A4)** The spline degree of KANs satisfies $m \ge 2$ and $0 < \overline{\mathbf{s}} < \min\{m, m - 1 + 1/p\}$.

Assumption (A1) is a standard condition adopted in theoretical studies over Besov spaces. It means that the true function $f_0$ belongs to a bounded Besov class, and the condition $\tilde{\mathbf{s}} > (1/p - 1/2)_+$ is the minimal condition for $L^2$-integrability and for the embedding $\mathcal{B}^{\tilde{\mathbf{s}}}_{p,q} \hookrightarrow L^2$. Meanwhile, (A2)–(A3) are frequently used in the analysis of posterior contraction for Bayesian models. Assumption (A2) is natural once the inputs $X$ have been normalized to $[0,1]^d$ in the fixed-dimensional regime $d = O(1)$. Assumption (A3) requires only mild prior knowledge of the range in which $\sigma_0$ lies.

**Remark 3.1.** *The uniform boundedness condition in* (A1) *is imposed with bound* 1 *only for simplicity. More generally, one may take any sufficiently large $F > 0$, define $\mathcal{UB}_F := \{f : [0,1]^d \to \mathbb{R} : \|f\|_\infty \le F\}$ and $\mathrm{clip}_F(x) := \min\{F, \max\{-F, x\}\}$. We set $F = 1$ without loss of generality. The convention applies to* (D1)*, to be introduced later.*

**Proof strategy.** All results below are proved through a common pipeline. Let $A_n := \{(f, \sigma) : \|f - f_0\|_{L^2(P_X)} \le \epsilon_n/2,\ |\sigma - \sigma_0| \le \epsilon_n/2\}$. By Lemma C.1, the posterior contracts to $(f_0, \sigma_0)$ at rate $\epsilon_n$, i.e., for any $M_n \to \infty$ it asymptotically assigns no mass outside $M_n\epsilon_n$-neighborhoods of $(f_0, \sigma_0)$, provided that there exists a set of functions $\mathcal{F}_n$ (a sieve) such that, for a sufficiently large $C > 0$,

$$\text{(i) } \log \mathcal{N}\big(\epsilon_n, \mathcal{F}_n, \|\cdot\|_{L^2(P_X)}\big) \lesssim n\epsilon_n^2, \quad \text{(ii) } -\log \Pi_n(A_n) \lesssim n\epsilon_n^2, \quad \text{(iii) } \Pi_n(\mathcal{F}_n^c) = o\big(e^{-Cn\epsilon_n^2}\big).$$

We construct $\mathcal{F}_n$ to consist of KANs of fixed depth whose parameters are sparse and bounded. For a model-size parameter $N$, Lemma B.4 constructs such a KAN with $O(N)$ nonzero parameters that exactly reproduces a B-spline approximation of $f_0$ with $L^2$-error $O(N^{-\tilde{\mathbf{s}}})$. For (ii), Lemma B.4 and the Lipschitz bound in Corollary B.14 ensure that every parameter in a small $\ell_\infty$-ball around that of the approximating KAN yields $f_{\boldsymbol{\theta}}$ within $\epsilon_n/2$ of $f_0$; the prior mass of this ball then bounds $-\log \Pi_n(A_n)$ by a multiple of $N \log n$. For (i), the same Lipschitz bound yields the covering bound of order $N \log n$ in Lemma B.15, so (i) holds whenever $N \log n \lesssim n\epsilon_n^2$. Condition (iii) follows from tail bounds of the prior on the parameters. Requiring both $N^{-\tilde{\mathbf{s}}} \lesssim \epsilon_n$ and $N \log n \lesssim n\epsilon_n^2$ leads to $N \asymp n^{1/(2\tilde{\mathbf{s}}+1)}$ and $\epsilon_n = n^{-\tilde{\mathbf{s}}/(2\tilde{\mathbf{s}}+1)} (\log n)^{1/2}$.

### 3.1 Spike-and-slab prior

The result in this section can be summarized as follows. For any function $f_0$ satisfying (A1), there exists an $m$-degree sparse KAN approximator $f^\star = f_{\boldsymbol{\theta}^\star} \in KAN(L_0, D_n, G_n, H_n; m)$ with parameter $\boldsymbol{\theta}^\star$ satisfying $\|\boldsymbol{\theta}^\star\|_\infty \le B_n^\star$ and $\|\boldsymbol{\theta}^\star\|_0 \le S_n$, attaining approximation error $\epsilon_n \asymp n^{-\tilde{\mathrm{s}}/(2\tilde{\mathrm{s}}+1)}$. Under a spike-and-slab prior, the posterior distribution contracts toward $f_0$ at a rate that matches this approximation error up to a logarithmic factor. The optimal architecture of the given sparse KANs is as follows:

$$L_0 = 3 + 2\lceil \log_2 d \rceil, \quad D_n = 2dN_n, \quad G_n = 4H_n - 2m, \quad H_n = \lceil 2B_n^\star \rceil \tag{4}$$

with sparsity $S_n := \lceil S_0 N_n \rceil$ for some $S_0 > 0$, where $N_n := \lceil n^{1/(2\tilde{\mathrm{s}}+1)} \rceil$, $B_n^\star = B_0 N_n^\beta$ for some $B_0 > 0$, and $\beta = (1+\kappa)\max\{(1/p - \tilde{\mathrm{s}})_+, \tilde{\mathrm{s}}/\underline{\mathrm{s}}\}$ with $\kappa = \frac{2\omega}{\tilde{\mathrm{s}}-\omega}$ if $p < 2$ and $\kappa = 0$ if $p \ge 2$ and $\omega = (1/p - 1/2)_+$.

Equation (4) highlights two structural features of the construction. First, the network depth is held fixed at the finite value $L_0$, while approximation power is acquired through the width $D_n$, the grid size $G_n$, and the grid range $H_n$. The induced grid spacing $2H_n/(G_n + 2m) = 1/2$ remains bounded away from zero, so the construction does not rely on increasingly fine grids. Second, the entire architecture parameters $(L_0, D_n, G_n, H_n, S_n)$ are governed by a single model-size parameter $N_n$, yielding a particularly parsimonious specification. The weight bound $B_n^\star$ of the oracle network $f^\star$ also scales with $N_n$; since $B_n^\star$ enters the prior conditions stated below, the slab distribution must be adjusted accordingly.

Let $T_n := T(L_0, D_n, G_n)$ be the total number of model parameters under the architecture Eq. (4). The spike-and-slab (SAS) prior places a joint distribution on the weight vector $\boldsymbol{\theta} = (\theta_t)_{t=1}^{T_n} \in \mathbb{R}^{T_n}$, the sparsity index $\boldsymbol{\gamma} = (\gamma_t)_{t=1}^{T_n} \in \{0,1\}^{T_n}$ and the noise variance $\sigma^2$:

$$\begin{aligned} \pi(\boldsymbol{\theta} \mid \boldsymbol{\gamma}) &= \prod_{t=1}^{T_n} \Big[ \gamma_t \widetilde{\pi}_{SL}(\theta_t) + (1-\gamma_t)\delta_0(\theta_t) \Big], \\ \pi(\boldsymbol{\gamma}) &= \binom{T_n}{S_n}^{-1} \mathbf{1}\{\boldsymbol{\gamma} \in \{0,1\}^{T_n},\ \|\boldsymbol{\gamma}\|_0 = S_n\}, \\ \pi(d\sigma^2) &= \pi_{\sigma^2}(\sigma^2)\, d\sigma^2 \end{aligned} \tag{5}$$

Here, $\widetilde{\pi}_{SL}$ is a continuous slab density that governs the nonzero coordinates of $\theta$ and $\pi_{\sigma^2}$ is the prior density function of the noise variance. The prior on $\boldsymbol{\gamma}$ is uniform over the set of $T_n$ dimensional binary vectors with exactly $S_n$ active coordinates.

For the spike-and-slab prior (5) to support our contraction results, we require

- **(B1)** There exists a constant $c_1 > 0$ such that $\log \inf_{|u| \le B_n^\star + 1} \widetilde{\pi}_{SL}(u) \ge -c_1 \log n$ for all sufficiently large $n$.
- **(B2)** There exist constants $c_2, c_3, \alpha > 0$ and a sequence $\tau_n \lesssim B_n^\star$ such that for all $t > 0$, $\int_{|u|>t} \widetilde{\pi}_{SL}(u)\, du \le c_2 \exp\left[ -c_3 \left( \frac{t}{\tau_n} \right)^\alpha \right]$.
- **(B3)** $\pi_{\sigma^2}$ is supported on $[\underline{\sigma}^2, \overline{\sigma}^2]$, and there exists a constant $c_\sigma$ such that $\inf_{\sigma \in [\underline{\sigma}, \overline{\sigma}]} \pi_{\sigma^2}(\sigma^2) \ge c_\sigma > 0$.

Conditions (B1)–(B3) reflect the prior properties needed for the general posterior contraction framework of Ghosal et al. (2000); Ghosal and van der Vaart (2007).

**Theorem 3.2** (Spike-and-slab prior)**.** *Suppose Assumptions* (A1)–(A4) *hold, and that the architecture and prior of the sparse Bayesian KAN are specified as in* (4) *and* (5)*, with the prior satisfying* (B1)–(B3)*. Then the posterior distribution contracts toward the true value* $(f_0, \sigma_0^2)$ *at the rate*

$$\epsilon_n = n^{-\frac{\tilde{\mathrm{s}}}{2\tilde{\mathrm{s}}+1}} (\log n)^{1/2}. \tag{6}$$

*In other words, for any sequence* $M_n \to \infty$*,*

$$\Pi\left( (\boldsymbol{\theta}, \sigma) : \|\mathrm{clip} \circ f_{\boldsymbol{\theta}} - f_0\|_{L^2(P_X)} + \left|\sigma^2 - \sigma_0^2\right| > M_n \epsilon_n \;\middle|\; \mathcal{D}_n \right) \to 0 \tag{7}$$

*in* $P_{f_0,\sigma_0}^{(n)}$*-probability as* $n \to \infty$*.*

*Proof.* See Appendix C.2. □

**Remark 3.3.** *Theorem 3.2 is a near-optimal result, it differs from known minimax convergence rate $n^{-\tilde{\mathbf{s}}/(2\tilde{\mathbf{s}}+1)}$ over anisotropic Besov function spaces only by the logarithmic factor $(\log n)^{1/2}$ (Kerkyacharian et al., 2001; Hoffman and Lepski, 2002).*

The following are standard families of slab priors that satisfy (B1)–(B2). In each case the slab scale $\tau_n$ is tuned in accordance with the oracle weight bound $B_n^\star \asymp n^{\beta/(2\tilde{\mathbf{s}}+1)}$.

**Example 3.4** (Uniform slab distribution)**.** *Slab distribution $\tilde{\pi}_{SL} = U(-\tau_n, \tau_n)$ where $\tau_n = C_\tau\, n^{\beta/(2\tilde{\mathbf{s}}+1)} \geq B_n^\star + 1$ for some $C_\tau > 0$.*

**Example 3.5** (Sub-Weibull slab distribution)**.** $\tilde{\pi}_{SL} = \mathrm{subW}(1/\alpha)$, $\alpha > 0$ *(Vladimirova et al., 2020). When the density is given by $\tilde{\pi}_{SL}(u) \propto \frac{1}{\tau_n} e^{-\left|\frac{u}{\tau_n}\right|^\alpha}$, $\tau_n = C_\tau\, n^{\frac{\beta}{2\tilde{\mathbf{s}}+1}}$ for some $C_\tau > 0$, the prior distribution satisfies both* (B1)–(B2)*. The case $\alpha = 2$ corresponds to the Gaussian distribution, and the case $\alpha = 1$ corresponds to the Laplace distribution.*

### 3.2 Bernoulli spike-and-slab prior

We now consider a Bernoulli variant of the spike-and-slab prior, in which the inclusion indicators $\gamma_t$ are i.i.d. Bernoulli draws rather than uniform on a fixed-cardinality support. Let $\rho_n \in (0, 1)$ denote the coordinatewise inclusion probability.

The Bernoulli spike-and-slab prior is then defined by

$$\begin{aligned} \pi(\boldsymbol{\theta} \mid \boldsymbol{\gamma}) &= \prod_{t=1}^{T_n} \Big[\gamma_t \widetilde{\pi}_{SL}(\theta_t) + (1-\gamma_t)\delta_0(\theta_t)\Big], \\ \pi(\boldsymbol{\gamma}) &= \prod_{t=1}^{T_n} \rho_n^{\gamma_t}(1-\rho_n)^{1-\gamma_t}, \\ \pi(d\sigma^2) &= \pi_{\sigma^2}(\sigma^2)\, d\sigma^2 \end{aligned} \tag{8}$$

Specific admissible choices of $\rho_n$ are discussed in Eq. (9) and Remark 3.7 below.

**Theorem 3.6** (Bernoulli spike-and-slab)**.** *Suppose Assumptions* (A1)–(A4) *hold, that the architecture and prior of the sparse Bayesian KAN are specified as in* (4) *and* (8) *with the slab and variance prior satisfying* (B1)–(B3)*, and that the inclusion probability $\rho_n$ satisfies*

$$T_n^{-1} \lesssim \rho_n, \quad \log\left(\frac{S_n}{T_n \rho_n}\right) \gtrsim \log n \tag{9}$$

*Then the posterior achieves the same rate in* (6) *and the result* (7) *holds.*

*Proof.* See Appendix C.3. □

**Remark 3.7.** *Condition Eq.* (9) *is satisfied, for instance, by inclusion probabilities $\rho_n$ such that, for some $0 < c_\rho < 1/(2\tilde{\mathbf{s}}+1)$, $N_n^{-(2+\beta)} \lesssim \rho_n \lesssim N_n^{-(1+\beta)} n^{-c_\rho}$. On the other hand, $\rho_n \asymp T_n^{-1}$ is a particularly simple choice. In this case, $T_n \rho_n \asymp 1$ and $\log \frac{S_n}{T_n \rho_n} \asymp \log S_n \asymp \log N_n \asymp \log n$, so Eq.* (9) *is satisfied.*

### 3.3 Adaptive spike-and-slab prior

A key question in adaptive nonparametric estimation is whether a single prior can automatically select a model suited to the unknown smoothness while still attaining the near-minimax contraction rate on each smoothness class. The architecture (4) depends on $\mathbf{s}$ and, through $\beta$, on $p$ as well. So the corresponding sparse Bayesian KAN requires oracle knowledge of the regularity of $f_0$. We now relax this by placing a hyperprior on a single model-size parameter $N$ and controlling the remaining architectural quantities through conservative upper bounds that cover a prespecified smoothness envelope.

Let $\mathfrak{S} \subset \mathbb{R}_{++}^d$ denote the smoothness envelope over which adaptation is desired. Define $\tilde{s}_{\min} := \inf_{\mathbf{s}\in\mathfrak{S}} (\sum_{j=1}^d s_j^{-1})^{-1} > \omega$ where $\omega = (1/p - 1/2)_+$, and set $\beta_{ad} := 1 + \kappa_{ad}$, where $\kappa_{ad}$ is fixed

arbitrary as $\kappa_{ad} > 2\omega/(\tilde{s}_{\min} - \omega)$ if $p < 2$ and $\kappa_{ad} > 0$ if $p \geq 2$. Since $\tilde{\mathbf{s}} \leq \underline{\mathbf{s}}$, $(1/p - \tilde{\mathbf{s}})_+ < 1/2$ and $\kappa \leq 2\omega/(\tilde{s}_{\min} - \omega)$, the exponent $\beta$ satisfies $\beta \leq 1 + \kappa < \beta_{ad}$ for every $\mathbf{s} \in \mathfrak{S}$.

Now for each $N \in \mathbb{N}$, which is a model size parameter, we define the adaptive sparse KANs architecture by

$$L_0 = 3 + 2\lceil \log_2 d \rceil, \quad D(N) = 2dN, \quad G(N) = 4H(N) - 2m, \quad H(N) = \lceil 2B^\star(N) \rceil.$$

where $B^\star(N) := B_{ad} N^{\beta_{ad}}$ with $B_{ad} > 0$ chosen large enough. Also, we denote the sparsity by $S(N) := \lceil S_0 N \rceil$ for $S_0 > 0$. And let the total number of parameters be $T(N) := T(L_0, D(N), G(N))$.

Building on the previous settings, the adaptive spike-and-slab prior is defined as:

$$\begin{aligned} \pi_N(N) &\propto \exp(-\lambda_N N \log N) \\ \pi(\boldsymbol{\theta} \mid \boldsymbol{\gamma}, N) &= \prod_{t=1}^{T(N)} \Big[ \gamma_t \widetilde{\pi}_{SL}(\theta_t \mid N) + (1 - \gamma_t)\delta_0(\theta_t) \Big] \\ \pi(\boldsymbol{\gamma} \mid N) &= \binom{T(N)}{S(N)}^{-1} \mathbf{1}\{\boldsymbol{\gamma} \in \{0,1\}^{T(N)},\ \|\boldsymbol{\gamma}\|_0 = S(N)\} \\ \pi(d\sigma^2) &= \pi_{\sigma^2}(\sigma^2)\, d\sigma^2, \end{aligned} \tag{10}$$

where $\lambda_N > 0$ is fixed. Assume that the following conditions hold. These are conditional versions of (B1)-(B2).

- **(C1)** There exists a constant $c_1 > 0$ such that for all $N$, $\log \inf_{|u| \leq B^\star(N)+1} \widetilde{\pi}_{SL}(u|N) \geq -c_1 \log N$.
- **(C2)** There exist constants $c_2, c_3 > 0$ and a sequence $\tau_N \lesssim N^{\beta_{\mathrm{ad}}}$ such that for all $t > 0$, $\int_{|u|>t} \widetilde{\pi}_{SL}(u \mid N)\, du \leq c_2 \exp\Big[-c_3 \left(\frac{t}{\tau_N}\right)^{\alpha}\Big]$.
- **(C3)** $\pi_{\sigma^2}$ satisfies (B3).

**Theorem 3.8** (Adaptive spike-and-slab prior)**.** *Suppose that Assumptions* (A1)–(A4) *hold for some unknown $s \in \mathfrak{S}$, and that the prior in Eq.* (10) *satisfies* (C1)–(C3)*. Then the induced posterior contracts at the same rate $\epsilon_n$ in* (6)*; in particular,* (7) *holds for $\epsilon_n$.*

*Proof.* See Appendix C.4. □

### 3.4 Extension to compositional anisotropic Besov spaces

We now extend the main results to compositional anisotropic Besov spaces, following the function-space definition of Suzuki and Nitanda (2021); Lee et al. (2025). In this setting, the regression function admits a layered compositional structure in which each layer depends only on a low-dimensional subspace of its inputs (Schmidt-Hieber, 2020), so that the effective approximation complexity is determined by the *layerwise smoothness and intrinsic dimension* rather than by the ambient dimension $d$. Structures of this kind arise naturally in regression. The additive model $f(\mathbf{x}) = \sum_{i=1}^{d} g_i(x_i)$ with $g_i \in \mathcal{B}_{p,q}^{s_i}([0,1])$ and $s_i > 1/p$, and the low-dimensional projection model $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x} + \mathbf{b})$ with $g \in \mathcal{B}_{p,q}^{\mathbf{s}}([0,1]^t)$, $\tilde{s} > 1/p$ and $\mathbf{A} \in \mathbb{R}^{t \times d}$, are special cases corresponding to $t^{(j)} = 1$ and $J = 2$, respectively (Lee et al., 2025).

**Definition 3.9** (Compositional anisotropic Besov space)**.** *Let $J \in \mathbb{N}$ be the compositional depth and let $\mathbf{d} = (d^{(0)}, d^{(1)}, \ldots, d^{(J)}) \in \mathbb{N}^{J+1}$, $d^{(0)} = d$, $d^{(J)} = 1$, be a sequence of dimensions. Let $\mathbf{t} = (t^{(1)}, \ldots, t^{(J)})$ be a sequence of effective dimensions with $1 \leq t^{(j)} \leq d^{(j-1)}$, and let $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(J)}$ be layerwise smoothness vectors with $\mathbf{s}^{(j)} \in \mathbb{R}_{++}^{t^{(j)}}$. The compositional anisotropic Besov space $\mathcal{B}_{p,q}^{\mathbf{d},\mathbf{t},\mathbf{s}}([0,1]^d)$ is the class of functions of the form $f = f_J \circ f_{J-1} \circ \cdots \circ f_1 : [0,1]^d \to \mathbb{R}$ such that, for $j = 1, \ldots, J-1$, $f_j = (f_{j,1}, \ldots, f_{j,d^{(j)}}) : [0,1]^{d^{(j-1)}} \to [0,1]^{d^{(j)}}$ and $f_J : [0,1]^{d^{(J-1)}} \to \mathbb{R}$. Moreover, for every $j \in [J]$ and $k \in [d^{(j)}]$, there exist a subset $I_{j,k} \subset [d^{(j-1)}]$ with $|I_{j,k}| = t^{(j)}$ and a function $\tilde{f}_{j,k} \in \mathcal{B}_{p,q}^{\mathbf{s}^{(j)}}([0,1]^{t^{(j)}})$ such that $f_{j,k}(\mathbf{x}) = \tilde{f}_{j,k}(\mathbf{x}_{I_{j,k}})$, where $\mathbf{x}_I$ denotes the subvector of $\mathbf{x}$ indexed by $I$.*

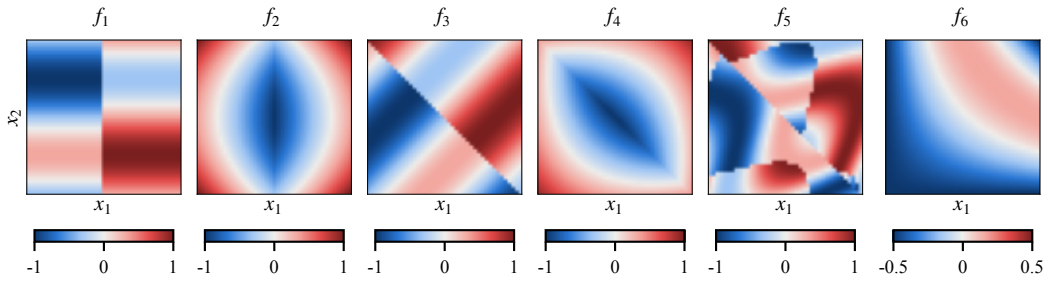



Figure 2: Target functions of the simulation study in Appendix D, normalized to $[-1, 1]$: $f_1$ and $f_2$ from Lee et al. (2025), their $45^\circ$ rotations $f_3$ and $f_4$, the composition $f_5$, and Friedman's function $f_6$ on $[0, 1]^{10}$, shown on $(x_1, x_2)$ with $x_3 = \cdots = x_{10} = 1/2$.

Figure 2 shows examples of the function classes considered in this paper. The functions $f_1$ (a jump in $x_1$ plus a periodic term in $x_2$) and $f_2$ (a kink in $x_1$ plus a quadratic term in $x_2$) are anisotropic, with smoothness differing across coordinates. The functions $f_3$ and $f_4$ are $f_1$ and $f_2$ composed with a rotation, so that the direction of low regularity is no longer aligned with a coordinate axis, and $f_5 = f_3\big((f_3 + 1)/2, (f_4 + 1)/2\big)$ composes $f_3$ and $f_4$ into a two-layer function with the structure of Definition 3.9 above. Friedman's function $f_6$ (Friedman, 1991) depends on only five of its ten inputs through components of at most two variables, so its effective dimension is much smaller than the ambient dimension $d = 10$.

For each layer $j \in [J]$, define the layerwise intrinsic smoothness and minimum smoothness by $\tilde{s}^{(j)} := (\sum_{i=1}^{t^{(j)}} (s_i^{(j)})^{-1})^{-1}$, $\underline{\mathbf{s}}^{(j)} := \min_{1 \le i \le t^{(j)}} s_i^{(j)}$, $\bar{\mathbf{s}}^{(j)} := \max_{1 \le i \le t^{(j)}} s_i^{(j)}$. The rate of posterior contraction is governed by an intrinsic smoothness that accounts for the propagation of approximation errors across layers. For each $j \in [J]$, define $t^{*(j)} := \frac{\underline{\mathbf{s}}^{(j)}}{\tilde{s}^{(j)}}$, $\alpha_j := \underline{\mathbf{s}}^{(j)} - t^{*(j)}/p$, $\tilde{s}^{*(j)} := \tilde{s}^{(j)} \prod_{k=j+1}^{J} \{(\underline{\mathbf{s}}^{(k)} - \frac{t^{*(k)}}{p}) \wedge 1\}$, and let $j^* := \arg\min_{j \in [J]} \tilde{s}^{*(j)}$, $t^* := t^{*(j^*)}$, $\tilde{s}^* := \tilde{s}^{*(j^*)}$ denote the intrinsic layer, intrinsic dimension, and intrinsic smoothness, respectively.

**Assumptions.** We impose the following assumption on the compositional structure.

- **(D1)** The true function satisfies $f_0 \in \mathcal{B}_{p,q}^{\mathbf{d},\mathbf{t},\mathbf{s}}([0,1]^d) \cap \mathcal{UB}$. Furthermore, $\tilde{s}^{(1)} > (1/p - 1/2)_+$, $\tilde{s}^{(j)} > 1/p$ for $j = 2, \ldots, J$ and assume $Y_i | X_i \sim N(\mathrm{clip} \circ f(X_i), \sigma^2)$.
- **(D2)** For each $j \in [J]$, $0 < \bar{\mathbf{s}}^{(j)} < \min\{m, m - 1 + 1/p\}$, and if $q > 1$, $\alpha_j \neq 1$ for $j = 2, \ldots, J$.

For $j \ge 2$, the condition $\tilde{s}^{(j)} > 1/p$ in (D1) guarantees, via the embedding $\mathcal{B}_{p,q}^{\mathbf{s}^{(j)}} \hookrightarrow \mathcal{B}_{\infty,q}^{\gamma_j \mathbf{s}^{(j)}}$ with $\gamma_j := 1 - 1/(p\tilde{s}^{(j)}) > 0$ (Suzuki and Nitanda, 2021, Proposition 1), that the components of the $j$-th layer are Hölder continuous with exponent $\alpha_j \wedge 1$ (Lemma B.8). An approximation error of order $N^{-\tilde{s}^{(j)}}$ incurred at an inner layer is raised to the power of the Hölder exponent of each outer layer it passes through, so that it becomes $N^{-\tilde{s}^{*(j)}}$ at the output; this is precisely what the definition of $\tilde{s}^{*(j)}$ captures.

Under (D1), (D2), (A2), and (A3), there exist constants $L_{\mathrm{cp}} \in \mathbb{N}$ and $D_{\mathrm{cp}}, S_{\mathrm{cp}}, B_{\mathrm{cp}}, \beta_{\mathrm{cp}} > 0$ such that the compositional architecture sequence is given by

$$\begin{aligned} N_n^* &:= \lceil n^{1/(2\tilde{s}^*+1)} \rceil, \quad D_n^{\mathrm{cp}} := D_{\mathrm{cp}} N_n^*, \quad G_n^{\mathrm{cp}} := 4H_n^{\mathrm{cp}} - 2m, \\ B_n^{\mathrm{cp},*} &:= B_{\mathrm{cp}} (N_n^*)^{\beta_{\mathrm{cp}}}, \quad H_n^{\mathrm{cp}} := \lceil 2B_n^{\mathrm{cp},*} \rceil, \quad S_n^{\mathrm{cp}} := \lceil S_{\mathrm{cp}} N_n^* \rceil. \end{aligned} \tag{11}$$

We use the spike-and-slab prior Eq. (5) with the compositional architecture sequence in Eq. (11), where the slab density satisfies (B1)–(B2) with $B_n^\star = B_n^{\mathrm{cp},*}$ and $\tau_n \lesssim (N_n^*)^{\beta_{\mathrm{cp}}}$, and the variance prior satisfies (B3).

**Theorem 3.10** (Compositional anisotropic Besov spaces)**.** *Suppose* (D1)*,* (D2)*,* (A2)*, and* (A3)*, and that the architecture of the sparse Bayesian KAN is specified as in* (11) *and its prior in* (5) *satisfies*

(B1)–(B3). *Then, for* $\epsilon_n^* := n^{-\tilde{s}^*/(2\tilde{s}^*+1)}(\log n)^{1/2}$ *and any sequence* $M_n \to \infty$,

$$\Pi\big((\boldsymbol{\theta}, \sigma) : \|\operatorname{clip} \circ f_{\boldsymbol{\theta}} - f_0\|_{L^2(P_X)} + |\sigma^2 - \sigma_0^2| > M_n \epsilon_n^* \mid \mathcal{D}_n\big) \to 0$$

*in* $P_{f_0,\sigma_0}^{(n)}$*-probability as* $n \to \infty$.

*Proof.* See Appendix C.5. □

**Remark 3.11.** *The rate* $\epsilon_n^* = n^{-\tilde{s}^*/(2\tilde{s}^*+1)}(\log n)^{1/2}$ *depends only on the intrinsic smoothness* $\tilde{s}^*$*, not on the ambient dimension* $d$*. When* $t^{(j)} \ll d^{(j-1)}$ *for each layer, the curse of dimensionality is substantially mitigated.*

**Remark 3.12** (Numerical illustration)**.** *For the target functions in Figure 2, Appendix D reports the estimation results of sparse Bayesian KANs together with a Bayesian ReLU network, BART (Chipman et al., 2010), a Gaussian process, and a random forest, for sample sizes* $n$ *from* $250$ *to* $4{,}000$*.*

## 4 Conclusion and discussion

In this paper, we established posterior contraction rates for sparse Bayesian KANs over anisotropic Besov spaces. We showed that sparse Bayesian KANs equipped with spike-and-slab priors achieve near-minimax posterior contraction rates, with the rate governed by the intrinsic anisotropic smoothness of the underlying function. Moreover, by introducing a hyperprior on a single model-size parameter, we demonstrated that the posterior can adapt to unknown smoothness while still attaining the corresponding near-optimal rate. We further extended our analysis to compositional Besov spaces, showing that sparse Bayesian KANs attain near-minimax rates determined by the layerwise smoothness and effective dimensions of the underlying compositional structure.

A key insight of our analysis is that, in contrast to standard ReLU-based architectures, the depth of KANs can be kept fixed, while approximation complexity is controlled through spline-grid range and size, width, and sparsity. This highlights KANs' mechanism for achieving statistical efficiency: rather than relying on increasingly deep architectures, they leverage flexible spline-based edge functions and sparse representations to capture anisotropic regularity. Our results show that this structural advantage translates into optimal statistical efficiency in a Bayesian framework.

There are several future directions to explore. First, while our analysis focuses on fixed-knot spline constructions, it would be of interest to study adaptive or data-driven knot placement, which could further enhance empirical performance. Second, it would be of interest to develop a theory for generative models in which the generator or the score function is modeled by the KAN architecture. Third, although our theory captures statistical optimality, understanding the computational aspects of posterior inference for sparse KANs, including scalable variational or MCMC methods, remains an important problem.

## Acknowledgments and Disclosure of Funding

Jaeyong Lee was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2023R1A2C1003050 and RS-2024-00406127). Lizhen Lin would like to acknowledge the generous support of NSF grant DMS 2503119. Kyeongwon Lee was supported by the National Research Foundation of Korea (NRF) funded by the Korea government (MSIT) (RS-2026-25486448) and INHA UNIVERSITY Research Grant (2026).

## References


Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljacic, T. Hou, and M. Tegmark. KAN: Kolmogorov-Arnold networks. In *International Conference on Learning Representations*, volume 2025, pages 70367–70413, 2025.

Z. Liu, P. Ma, Y. Wang, W. Matusik, and M. Tegmark. KAN 2.0: Kolmogorov-Arnold networks meet science. *arXiv preprint arXiv:2408.10205*, 2024.

C. J. Vaca-Rubio, L. Blanco, R. Pereira, and M. Caus. Kolmogorov-Arnold networks (kans) for time series analysis. In *2024 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6. IEEE, 2024.

M. Kiamari, M. Kiamari, and B. Krishnamachari. GKAN: Graph Kolmogorov-Arnold networks. *arXiv preprint arXiv:2406.06470*, 2024.

R. Bresson, G. Nikolentzos, G. Panagopoulos, M. Chatzianastasis, J. Pang, and M. Vazirgiannis. KAGNNs: Kolmogorov-Arnold networks meet graph learning. *Transactions on Machine Learning Research*, 2025:1–29, 2025.

B. C. Koenig, S. Kim, and S. Deng. KAN-ODEs: Kolmogorov–Arnold network ordinary differential equations for learning dynamical systems and hidden physics. *Computer Methods in Applied Mechanics and Engineering*, 432:117397, 2024.

G. Kerkyacharian, O. Lepski, and D. Picard. Nonlinear estimation in anisotropic multi-index denoising. *Probability theory and related fields*, 121(2):137–170, 2001.

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

P. Egels and I. Castillo. Posterior and variational inference for deep neural networks with heavy-tailed weights. *Journal of Machine Learning Research*, 26(122):1–58, 2025.

K. Lee, L. Lin, J. Park, and S. Jeong. Posterior contraction for sparse neural networks in Besov spaces with intrinsic dimensionality. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

T. Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2018.

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

N. G. Polson and V. Ročková. Posterior concentration for sparse deep learning. *Advances in neural information processing systems*, 31, 2018.

K. Lee and J. Lee. Asymptotic properties for bayesian neural network in besov space. *Advances in Neural Information Processing Systems*, 35:5641–5653, 2022.

I. Kong and Y. Kim. Posterior concentrations of fully-connected bayesian neural networks with general priors on the weights. *Journal of Machine Learning Research*, 26(94):1–60, 2025.

A. Kratsios, B. J. Kim, and T. Furuya. Approximation rates in Besov norms and sample-complexity of Kolmogorov-Arnold networks with residual connections. *Neural Networks*, page 108797, 2026.

Y. Wang, J. Siegel, Z. Liu, and T. Hou. On the expressiveness and spectral bias of kans. In *International Conference on Learning Representations*, volume 2025, pages 27492–27511, 2025.

L. Schumaker. *Spline Functions: Basic Theory*. Cambridge university press, 2007.

C. De Boor. *A practical guide to splines*, volume 27. Springer, New York, 1978.

S. M. Nikol'skii. *Approximation of functions of several variables and imbedding theorems*. Springer Science & Business Media, 1975.

M. Hoffman and O. Lepski. Random rates in anisotropic regression (with a discussion and a rejoinder by the authors). *The Annals of Statistics*, 30(2):325–396, 2002.

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

S. Ghosal and A. W. van der Vaart. Convergence rates of posterior distributions for non-iid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

M. Vladimirova, S. Girard, H. Nguyen, and J. Arbel. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.

J. H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, March 1991. doi: 10.1214/aos/1176347963. URL `https://doi.org/10.1214/aos/1176347963`.

H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, March 2010. doi: 10.1214/09-AOAS285. URL `https://doi.org/10.1214/09-AOAS285`.

C. K. Chui. *An Introduction to Wavelets*. Academic Press, San Diego, 1992.

R. A. DeVore and G. G. Lorentz. *Constructive approximation*, volume 303. Springer Science & Business Media, 1993.

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

F. Xie and Y. Xu. Adaptive Bayesian nonparametric regression using a kernel mixture of polynomials with application to partial linear models. *Bayesian Analysis*, 15(1):159 – 186, 2020.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

G. Petrillo. Very fast Bayesian additive regression trees on GPU. *arXiv preprint arXiv:2410.23244*, 2024.

M. D. Hoffman and A. Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.

# A Preliminaries

We introduce the notation and definitions needed to understand the proofs in this appendix. Let $(L, \mathbf{D}, \mathbf{G}, H)$ be the structure of KANs as defined in Eq. (3) where $L \in \mathbb{N}$, $\mathbf{D} \in \mathbb{N}^{L+1}$, $\mathbf{G} \in \mathbb{N}^L$, $H > 0$, and with parameters $\boldsymbol{\theta} = (\theta_{i,j,k}^{(l)})_{i,j,k,l}$.

$$KAN_c(L, \mathbf{D}, \mathbf{G}, H; m) := \left\{ f_{\boldsymbol{\theta}} \in KAN(L, \mathbf{D}, \mathbf{G}, H; m) \mid \theta_{i,j,0}^{(l)} = 0 \text{ for all } l, i, j \right\}$$

The class $KAN_c$ consists of KANs whose activation functions are constructed only from the spline components, without the SiLU activation. In this work, we analyze posterior contraction of the model under common width and grid size. That is, when the width and grid are given by $\mathbf{D} = (d, D, D, \ldots, 1)$ and $\mathbf{G} = (G_0, G, \ldots, G)$ where $G_0$ and the corresponding knot vector $\boldsymbol{\xi}^{(0)} \in [a_0, b_0]^{G_0+2m+1}$ s.t. $[\xi_0^{(0)}, \xi_{G_0}^{(0)}] \supseteq [0,1]^d$, $\Delta_0 = \xi_{k+1}^{(0)} - \xi_k^{(0)}$ are fixed, respectively, we write, without boldface notation,

$$KAN(L, D, G, H; m) := KAN(L, \mathbf{D}, \mathbf{G}, H; m),$$
$$KAN_c(L, D, G, H; m) := KAN_c(L, \mathbf{D}, \mathbf{G}, H; m)$$

for simplicity.

Next, we describe the properties of B-splines. Let $\boldsymbol{\xi} = (\xi_{-m}, \cdots, \xi_{G+m}) \in [a, b]^{G+2m+1}$, with $a < b$, be a knot vector. We assume that its coordinates are strictly increasing without repetition, i.e., $\xi_k < \xi_{k+1}$ for all $k$, and equally spaced with $\Delta = \xi_{k+1} - \xi_k$. These B-splines satisfy the nonnegativity and local support properties

$$B_{k,m}(x) \geq 0 \text{ for all } x, \quad B_{k,m}(x) = 0 \text{ if } x \notin [\xi_{k-m-1}, \xi_k], \quad (k = 1, \ldots, G+m).$$

They also satisfy the partition of unity property

$$\begin{aligned} \sum_{k=1}^{G+m} B_{k,m}(x) &= 1, \quad x \in [\xi_0, \xi_G], \\ 0 \leq \sum_{k=1}^{G+m} B_{k,m}(x) &< 1, \quad \text{otherwise.} \end{aligned} \tag{12}$$

A B-spline curve is a function $s_m : \mathbb{R} \to \mathbb{R}$ given by a linear combination of the basis functions $\{B_k\}_{k=1}^{G+m}$. We denote the coefficients of the basis functions by $\mathbf{w} = (w_1, \ldots, w_{G+m}) \in \mathbb{R}^{G+m}$, that is,

$$s_m(x) := \sum_{k=1}^{G+m} w_k B_{k,m}(x).$$

Since the shape of each curve depends on $(\mathbf{w}, \boldsymbol{\xi})$, we write $s_m(x; \mathbf{w}, \boldsymbol{\xi})$ when we need to emphasize this dependence. We introduce one more property of the B-spline curve $s_m$: its derivative $s_m'$ is again a B-spline curve of degree $m-1$. Specifically, following De Boor (1978),

$$\begin{aligned} s_m'(x) &= \sum_{k=0}^{G+m} w_k' B_{k,m-1}(x), \\ w_k' &:= \frac{m(w_{k+1} - w_k)}{\xi_k - \xi_{k-m}} = \frac{w_{k+1} - w_k}{\Delta}, \quad k = 0, \ldots, G+m, \end{aligned} \tag{13}$$

where $w_0 = w_{G+m+1} := 0$, and $B_{k,m-1}(\cdot\,; \boldsymbol{\xi})$, $k = 0, \ldots, G+m$, are the B-splines of degree $m-1$ generated by the same knot vector, with $B_{k,m-1}$ supported on $[\xi_{k-m}, \xi_k]$. The boundary terms $k = 0$ and $k = G+m$ vanish on $[\xi_0, \xi_G]$, but they are needed for the formula to hold on all of $\mathbb{R}$. Furthermore, writing $\mathbf{w}' = (w_0', \ldots, w_{G+m-1}', w_{G+m}')$ and applying Eq. (12) to the B-splines of degree $m-1$, we obtain

$$|s_m'(x)| \leq \|\mathbf{w}'\|_\infty \leq \frac{2\|\mathbf{w}\|_\infty}{\Delta}, \quad \forall x. \tag{14}$$

Finally, when we emphasize the spline part of the edge function $\phi_{i,j}^{(l)}$ in the layer $\Phi^{(l)}$, we write

$$\phi_{i,j}^{(l)}(x) = s_m(x; \boldsymbol{\theta}_{i,j,-0}^{(l)}) + \theta_{i,j,0}^{(l)} \text{ SiLU}(x),$$

where $\boldsymbol{\theta}_{i,j}^{(l)} = (\theta_{i,j,0}^{(l)}, \boldsymbol{\theta}_{i,j,-0}^{(l)}) \in \mathbb{R}^{G_l+m+1}$.

## B Auxiliary results

### B.1 Exact B-spline representations

For an equally spaced knot vector $\boldsymbol{\xi} = (\xi_{-m}, \dots, \xi_{G+m})$ with spacing $\Delta$, $B_{k,m}(\cdot\,; \boldsymbol{\xi})$ is supported on $[\xi_{k-m-1}, \xi_k]$, and we write $s_m(\cdot\,; \mathbf{w}, \boldsymbol{\xi}) = \sum_{k=1}^{G+m} w_k B_{k,m}(\cdot\,; \boldsymbol{\xi})$.

**Lemma B.1** (Local polynomial reproduction)**.** *Let $\boldsymbol{\xi}$ be an equally spaced knot vector with spacing $\Delta > 0$, $I = [u_-, u_+] \subseteq [\xi_0, \xi_G]$, $P(u) = \sum_{r=0}^{m} a_r u^r$, and $m \geq 1$. Then there exists $\mathbf{w} \in \mathbb{R}^{G+m}$ such that*

*(i)* $s_m(u; \mathbf{w}, \boldsymbol{\xi}) = P(u)$ *for* $u \in I$*;*

*(ii)* $w_k = 0$ *whenever* $(\xi_{k-m-1}, \xi_k) \cap I = \emptyset$*; in particular,* $\|\mathbf{w}\|_0 \leq \lceil (u_+ - u_-)/\Delta \rceil + m + 1$*;*

*(iii)* $\|\mathbf{w}\|_\infty \leq \sum_{r=0}^{m} |a_r| R^r$*, where* $R := \max\{|u_-|, |u_+|\} + (m+1)\Delta$.

*Proof.* Consider the Marsden identity (De Boor, 1978)

$$(u - y)^m = \sum_{k=1}^{G+m} \prod_{i=k-m}^{k-1} (\xi_i - y)\, B_{k,m}(u; \boldsymbol{\xi}), \qquad u \in [\xi_0, \xi_G],\ y \in \mathbb{R}.$$

Regarding both sides as polynomials in $y$ and comparing the coefficients of $(-y)^{m-r}$, we obtain

$$u^r = \sum_{k=1}^{G+m} \lambda_{k,r} B_{k,m}(u; \boldsymbol{\xi}), \qquad \lambda_{k,r} := \binom{m}{r}^{-1} e_r(\xi_{k-m}, \dots, \xi_{k-1}), \qquad u \in [\xi_0, \xi_G], \tag{15}$$

where $e_r$ denotes the $r$th elementary symmetric polynomial. We set $w_k := \sum_r a_r \lambda_{k,r}$ if $(\xi_{k-m-1}, \xi_k) \cap I \neq \emptyset$, and $w_k := 0$ otherwise.

- (i): The omitted $B_{k,m}$ vanish on $I$.
- (ii): $(\xi_{k-m-1}, \xi_k) \cap I \neq \emptyset \iff \xi_k \in (u_-,\ u_+ + (m+1)\Delta)$, and this open interval contains at most $\lceil (u_+ - u_-)/\Delta \rceil + m + 1$ knots.
- (iii): If $w_k \neq 0$, then $[\xi_{k-m-1}, \xi_k]$ meets $I$, so $|\xi_i| \leq R$ for $k - m \leq i \leq k - 1$, and hence $|\lambda_{k,r}| \leq R^r$. □

Since the Marsden identity holds only on the basic interval $[\xi_0, \xi_G]$, the first layer requires $[\xi_0^{(0)}, \xi_{G_0}^{(0)}] \supseteq [0, 1]$. The localization in (ii) is the key point. The global coefficients grow as large as $|\lambda_{k,r}| \asymp \max_i |\xi_i|^r \asymp H^r$, but an edge whose input lies in $[0, 1]$ uses only the $O(m)$ basis functions whose supports meet $[0, 1]$, so its coefficients do not depend on $H$.

Let $\psi_m := \mathbf{1}_{[0,1)}^{*(m+1)}$ be the cardinal B-spline of degree $m$. It has knots $0, 1, \dots, m+1$ and support $[0, m+1]$, and coincides with the cardinal B-spline of order $m$ (i.e., of degree $m$) in Suzuki and Nitanda (2021).

**Lemma B.2** (Cardinal B-splines on a half-integer grid)**.** *Let $H \in \mathbb{N}$ with $H \geq m + 1$, let $G := 4H - 2m$, and let $\boldsymbol{\xi}$ be the equally spaced knot vector on $[-H, H]$. Then*

*(i)* $\Delta = 1/2$*,* $\xi_k = -H + (k+m)/2$*, and* $[0, 1] \subseteq [\xi_0, \xi_G] = [-H + m/2,\ H - m/2]$*;*

*(ii) there exists* $\mathbf{w}$ *with* $\|\mathbf{w}\|_0 = m + 2$ *and* $\|\mathbf{w}\|_\infty \leq 2$ *such that* $s_m(\cdot\,; \mathbf{w}, \boldsymbol{\xi}) = \psi_m$ *on* $\mathbb{R}$*;*

*(iii)* $0 \leq \psi_m \leq 1$.

*Proof.* (i) We have $\Delta = 2H/(G + 2m) = 1/2$, and since $H \geq m + 1$, it follows that $\xi_0 = -H + m/2 \leq 0$ and $\xi_G = H - m/2 \geq 1$.

(ii) Since the knots are equally spaced, $B_{k,m}(u;\boldsymbol{\xi}) = \psi_m(2u - 2\xi_{k-m-1})$, and hence $\psi_m(2\,\cdot - r) = B_{k_r,m}(\cdot\,;\boldsymbol{\xi})$ with $k_r := 2H + r + 1 \in [G + m]$ for $r = 0, \dots, m+1$. By the two-scale relation (Chui, 1992)

$$\psi_m(u) = 2^{-m} \sum_{r=0}^{m+1} \binom{m+1}{r} \psi_m(2u - r), \tag{16}$$

we may take $w_{k_r} = 2^{-m}\binom{m+1}{r} \le 2$ and $w_k = 0$ otherwise.

(iii) Since $\psi_m$ is a convolution of nonnegative functions, $\psi_m \ge 0$, and by the partition of unity, $\psi_m \le \sum_\ell \psi_m(\cdot - \ell) = 1$. □

**Lemma B.3** (Exact multiplication by fixed-knot $KAN_c$)**.** *Let $m \ge 2$ and $d \ge 2$, let $(H, G, \boldsymbol{\xi})$ be as in Lemma B.2, and let $L_\times := 2\lceil \log_2 d \rceil$. There exist $KAN_c$ layers $\Phi_0^\times, \dots, \Phi_{L_\times - 1}^\times$ with knot vector $\boldsymbol{\xi}$ and zero SiLU weights, input dimension $d$, output dimension $1$, and intermediate widths $n_\ell \le \lfloor 3d/2 \rfloor$, such that $\Pi_d := \Phi_{L_\times-1}^\times \circ \cdots \circ \Phi_0^\times$ satisfies the following:*

- *(i) $\Pi_d(\mathbf{u}) = u_1 \cdots u_d$ for $\mathbf{u} \in [0,1]^d$;*
- *(ii) if $\mathbf{u} \in [0,1]^d$, then all intermediate outputs lie in $[0,1]^{n_\ell}$;*
- *(iii) the number of nonzero weights is at most $\big(7(d-1) + 2\lceil \log_2 d \rceil\big)(m+3)$;*
- *(iv) every weight is bounded in absolute value by $(m+3)^2/2$.*

*Proof.* Applying Lemma B.1 with $I = [0,1]$ and $\Delta = 1/2$, so that $R = (m+3)/2$, we obtain B-spline curves $\phi_{\mathrm{id}}$, $\phi_{\mathrm{half}}$, and $\phi_{\mathrm{sq}}$ such that, for $t \in [0,1]$,

$$\varphi_{\mathrm{id}}(t) = t, \quad \varphi_{\mathrm{half}}(t) = t/2, \quad \varphi_{\mathrm{sq}}(t) = 2t^2. \tag{17}$$

Note that $\varphi_{\mathrm{sq}}$ requires degree $m \ge 2$. Each curve has at most $m+3$ nonzero coefficients, whose magnitudes are bounded by $R$, $R/2$, and $2R^2$, respectively.

This allows us to implement the binary product $\Pi_2$ through addition and composition of B-splines. That is, for $(u_1, u_2) \in [0,1]^2$,

$$\begin{aligned}
\phi^{(1)}(u_1,u_2) &:= \big(\varphi_{\mathrm{half}}(u_1) + \varphi_{\mathrm{half}}(u_2),\ \varphi_{\mathrm{half}}(u_1),\ \varphi_{\mathrm{half}}(u_2)\big) = \Big(\tfrac{u_1+u_2}{2}, \tfrac{u_1}{2}, \tfrac{u_2}{2}\Big) \in [0,1]^3,\\
\phi^{(2)}(v_1,v_2,v_3) &:= \varphi_{\mathrm{sq}}(v_1) - \varphi_{\mathrm{sq}}(v_2) - \varphi_{\mathrm{sq}}(v_3),\\
\Pi_2 &:= \phi^{(2)} \circ \phi^{(1)}, \qquad \Pi_2(u_1,u_2) = 2\big(\tfrac{u_1+u_2}{2}\big)^2 - 2\big(\tfrac{u_1}{2}\big)^2 - 2\big(\tfrac{u_2}{2}\big)^2 = u_1u_2.
\end{aligned} \tag{18}$$

Since each output coordinate is a sum of univariate edges, both maps are of the form Eq. (2). That is, $\Pi_2$ is implemented by $\phi^{(2)} \circ \phi^{(1)}$, which has depth 2, width 3, and 7 nonzero edges. Since the arguments are halved, all intermediate values lie in $[0,1]$.

We now construct a $KAN_c$ implementing $\Pi_d$ by repeating this composition along a binary tree. The proof proceeds by induction. Let $\ell = 0, \dots, L_\times/2 - 1$, and let the initial width be $n_0 := d$. First, for $\ell = 0$, the layers $\Phi_1^\times \circ \Phi_0^\times$ can be constructed as follows:

$$\begin{aligned}
\Phi_0^\times(u_1,\dots,u_d) &:= \big(\phi^{(1)}(u_1,u_2),\ \phi^{(1)}(u_3,u_4),\ \dots,\ \phi^{(1)}(u_{d-1},u_d)\big),\\
\Phi_1^\times \circ \Phi_0^\times(u_1,\dots,u_d) &:= \big(\phi^{(2)}\circ\phi^{(1)}(u_1,u_2),\ \phi^{(2)}\circ\phi^{(1)}(u_3,u_4),\ \dots,\ \phi^{(2)}\circ\phi^{(1)}(u_{d-1},u_d)\big)\\
&= \big(u_1u_2,\ u_3u_4,\ \dots,\ u_{d-1}u_d\big).
\end{aligned}$$

If $d$ is odd, we additionally append $\varphi_{\mathrm{id}}$ as the last component of each layer. The width of the output layer of $\Phi_1^\times \circ \Phi_0^\times$ is $n_1 = \lceil d/2 \rceil$, and the width of the intermediate layer $\Phi_0^\times$ is $3\lfloor d/2 \rfloor + (d \bmod 2)$. More generally, for $\ell > 0$, we define

$$\begin{aligned}
\Phi_{2\ell}^\times(v_1,\dots,v_{n_\ell}) &:= \big(\phi^{(1)}(v_1,v_2),\dots,\phi^{(1)}(v_{n_\ell-1},v_{n_\ell})\big),\\
\Phi_{2\ell+1}^\times\big(t_1,\dots,t_{3\lfloor n_\ell/2\rfloor}\big) &:= \big(\phi^{(2)}(t_1,t_2,t_3),\dots,\phi^{(2)}(t_{3\lfloor n_\ell/2\rfloor-2},t_{3\lfloor n_\ell/2\rfloor-1},t_{3\lfloor n_\ell/2\rfloor})\big).
\end{aligned}$$

If $n_\ell$ is odd, we again append $\phi_{\mathrm{id}}$ as the last component of each layer. The width of the output layer is $n_{\ell+1} = \lceil n_\ell/2 \rceil$, and the width of the intermediate layer is $3\lfloor n_\ell/2 \rfloor + (n_\ell \bmod 2)$. Repeating this

up to $\ell = L_\times/2 - 1$, the final output is $\Phi^\times_{L_\times - 1} \circ \cdots \circ \Phi^\times_0(u_1, \ldots, u_d) = u_1 \cdots u_d$. This proves (i) and (ii).

Next, we check the depth and width of this network. Since pairing the variables two at a time to form binary products is implemented by a $KAN_c$ module of depth 2, the total depth of the network is $L_\times = 2\lceil \log_2 d \rceil$. The width of each intermediate layer is $3\lfloor n_\ell/2 \rfloor + (n_\ell \bmod 2) \le \lfloor 3d/2 \rfloor$.

The network $\Phi^\times_{L_\times - 1} \circ \cdots \circ \Phi^\times_0$ contains exactly $d - 1$ copies of the binary product module $\Pi_2$, each of which has 7 nonzero edges. When the output width $n_\ell$ is odd, at most one additional component $\phi_{\mathrm{id}}$ is added per layer. Moreover, each edge has at most $m + 3$ nonzero coefficients. Altogether, the number of nonzero weights of the network is at most $(7(d-1) + 2\lceil \log_2 d \rceil)(m+3)$, which proves (iii).

Finally, for (iv), substituting the conditions of the present lemma into Lemma B.1(iii) gives $R = (m+3)/2$, so by the coefficient bounds following Eq. (17), the largest weight bound is $2R^2 = (m+3)^2/2$. This completes the proof. □

Lemma B.3 corresponds to the product network of Schmidt-Hieber (2020, Lemma A.3). Since ReLU networks only approximate products, their depth grows with the required accuracy, whereas B-spline edges with $m \ge 2$ reproduce quadratic maps exactly, so the depth depends only on $d$. The condition $m \ge 2$ in (A4) is used at this step.

### B.2 Approximation theorem

**Lemma B.4** (Approximation by fixed-knot $KAN_c$)**.** *Suppose that* (A1) *and* (A4) *hold. Then there exist constants $S_0, B_0 > 0$, independent of $N$, such that for every $N \in \mathbb{N}$ there exists*

$$f_N^\star \in KAN_c\big(L_0, D(N), G(N), H(N); m\big), \qquad L_0 = 3 + 2\lceil \log_2 d \rceil,$$

*satisfying*

$$\|f_0 - f_N^\star\|_{L^2} \lesssim N^{-\tilde{\mathbf{s}}}. \tag{19}$$

*Moreover, there exists a parameter vector $\boldsymbol{\theta}_N^\star$ realizing $f_N^\star$ such that*

$$\|\boldsymbol{\theta}_N^\star\|_\infty \le B^\star(N), \qquad \|\boldsymbol{\theta}_N^\star\|_0 \le S_0 N.$$

*Here,*

$$D(N) := 2dN, \quad G(N) := 4H(N) - 2m, \quad H(N) := \lceil 2B^\star(N) \rceil, \quad B^\star(N) := B_0 N^\beta,$$

*and*

$$\begin{aligned} \beta &:= (1+\kappa)\max\{(1/p - \tilde{\mathbf{s}})_+, \tilde{\mathbf{s}}/\underline{\mathbf{s}}\}, \\ \kappa &:= \begin{cases} \frac{2\omega}{\tilde{\mathbf{s}} - \omega}, & p < 2, \\ 0, & p \ge 2, \end{cases} \end{aligned} \tag{20}$$

*with $\omega := (1/p - 1/2)_+$.*

**Proof sketch and layer-wise structure.** We first summarize the construction in Lemma B.4. By Lemma 2 of Suzuki and Nitanda (2021), $f_0$ is approximated in $L^2$ at the rate $N^{-\tilde{\mathbf{s}}}$ by a tensor-product B-spline expansion $f_N^{\mathrm{sp}} = \sum_{(k,j)\in E_N} \alpha_{k,j} M_{k,j}^{d,m}$ with $|E_N| \le N$. Here $M_{k,j}^{d,m}(\mathbf{x}) = \prod_{i=1}^d \psi_m(z_{k,j,i}(x_i))$, $z_{k,j,i}(x_i) = 2^{\lfloor k\underline{\mathbf{s}}/s_i \rfloor} x_i - j_i$, and $\psi_m$ is the cardinal B-spline of degree $m$. The proof shows that $f_N^{\mathrm{sp}}$ is realized *exactly* by a fixed-knot $KAN_c$ through the following pipeline:

$$\begin{aligned} \mathbf{x} &\xmapsto{\Phi_0} \big(z_{k,j,i}(x_i)\big)_{k,j,i} \xmapsto{\Phi_1} \big(\psi_m(z_{k,j,i}(x_i))\big)_{k,j,i} \\ &\xmapsto{\Phi_2, \ldots, \Phi_{L_0-2}} \big(M_{k,j}^{d,m}(\mathbf{x})\big)_{(k,j)\in E_N} \xmapsto{\Phi_{L_0-1}} f_N^{\mathrm{sp}}(\mathbf{x}). \end{aligned}$$

(i) By the Marsden identity (De Boor, 1978), B-splines of degree $m$ exactly reproduce polynomials of degree at most $m$, so each affine map $z_{k,j,i}$ is implemented by a single edge in the first layer.

(ii) The choice $G(N) = 4H(N) - 2m$ yields the hidden-layer knot spacing $2H(N)/(G(N) + 2m) = 1/2$, and the two-scale relation $\psi_m(u) = 2^{-m} \sum_{r=0}^{m+1} \binom{m+1}{r} \psi_m(2u - r)$ (Chui, 1992) expresses $\psi_m$ in terms of $m + 2$ basis functions with coefficients independent of $N$.

(iii) Since $uv = 2(\frac{u+v}{2})^2 - 2(\frac{u}{2})^2 - 2(\frac{v}{2})^2$ involves only quadratic univariate maps, two KAN layers compute the product exactly when $m \geq 2$. A binary tree of such modules computes $M_{k,j}^{d,m}$ using $2\lceil \log_2 d \rceil$ layers (Lemma B.3).

(iv) The output layer forms the linear combination with coefficients $\alpha_{k,j}$. Since each of the $|E_N| \leq N$ blocks uses $O(1)$ nodes and nonzero weights, the width and sparsity grow linearly in $N$, while the depth does not depend on $N$; see table 1.

Table 1: Layer-wise structure of a single block $(k, j) \in E_N$ of the oracle network $f_N^\star$ in Lemma B.4. The $|E_N| \leq N$ blocks are arranged in parallel, and $H = H(N)$.

| Layer | Operation | Domain $\to$ Range | Nonzero weights | Weight bound |
|---|---|---|---|---|
| $\Phi_0$ | Affine map $z_{k,j,i}$ | $[0,1] \to [-H, H]$ | $\leq d(G_0 + m)$ | $\lesssim N^{(1+\kappa)\tilde{\mathbf{s}}/\underline{\mathbf{s}}}$ |
| $\Phi_1$ | Cardinal B-spline $\psi_m$ | $[-H, H] \to [0,1]$ | $d(m+2)$ | $\leq 2$ |
| $\Phi_2, \ldots, \Phi_{L_0-2}$ | Product tree | $[0,1]^d \to [0,1]$ | $O(dm)$ | $O(m^2)$ |
| $\Phi_{L_0-1}$ | $t \mapsto \alpha_{k,j} t$ | $[0,1] \to \mathbb{R}$ | $\leq m + 3$ | $\lesssim N^{(1+\kappa)(1/p-\tilde{\mathbf{s}})_+}$ |
| Total: depth $L_0 = 3 + 2\lceil \log_2 d \rceil$, width $2dN$, knot spacing $1/2$ | | | $\leq S_0 N$ | $\leq B_0 N^\beta$ |

*Proof. Step 0: Sparse spline approximation.* We apply Lemma 2 of Suzuki and Nitanda (2021) with $r = 2$; its conditions are satisfied under (A1) and (A4). The number of terms in Eq. (11) of that lemma is at most $C_0 2^{\|K\|_{\mathbf{s}}}$ for a constant $C_0$ depending only on $(d, m, \kappa)$, where $\|K\|_{\mathbf{s}} := \sum_i \lfloor K\underline{\mathbf{s}}/s_i \rfloor \geq K$. We therefore set $N_0 := \lceil C_0 \rceil$ and, for each $N \geq N_0$, choose the largest $K$ such that $C_0 2^{\|K\|_{\mathbf{s}}} \leq N$. Then $2^K \leq N$, and with $K^* = \lceil (1+\kappa)K \rceil$, there exist an index set $E_N \subset \{(k,j) : 0 \leq k \leq K^*,\ j \in \mathbb{Z}^d\}$ and coefficients $(\alpha_{k,j})$ such that the function

$$
\begin{aligned}
f_N^{\mathrm{sp}} &:= \sum_{(k,j)\in E_N} \alpha_{k,j} M_{k,j}^{d,m}, \quad |E_N| \leq N, \quad \max_{(k,j)\in E_N} |\alpha_{k,j}| \lesssim N^{(1+\kappa)(1/p-\tilde{\mathbf{s}})_+}, \\
M_{k,j}^{d,m}(\mathbf{x}) &:= \prod_{i=1}^{d} \psi_m\big(z_{k,j,i}(x_i)\big), \quad z_{k,j,i}(x_i) := 2^{a_{k,i}} x_i - j_i, \quad a_{k,i} := \lfloor k\underline{\mathbf{s}}/s_i \rfloor,
\end{aligned}
\tag{21}
$$

with $j_i \in \{-m, \ldots, 2^{a_{k,i}}\}$, satisfies

$$
\|f_0 - f_N^{\mathrm{sp}}\|_{L^2} \lesssim N^{-\tilde{\mathbf{s}}}. \tag{22}
$$

We now construct, layer by layer, $f_N^\star = \Phi_{L_0-1}^\star \circ \cdots \circ \Phi_0^\star \in KAN_c$ that exactly realizes $f_N^{\mathrm{sp}}$ on $[0,1]^d$. We denote the parameters of this network by $\boldsymbol{\theta}_N^\star$. In each layer, all edges not explicitly specified are set to 0.

For $N < N_0$, we set $f_N^\star := 0$ (i.e., $\boldsymbol{\theta}_N^\star = \mathbf{0}$), which gives $\|f_0 - f_N^\star\|_{L^2} \lesssim 1 \leq N_0^{\tilde{\mathbf{s}}} N^{-\tilde{\mathbf{s}}}$. Hence, in what follows, we prove the claim for $N \geq N_0$.

We choose $B_0 \geq (m+1)/2$ so that $H(N) \geq m + 1$ ($B_0$ will be enlarged finitely many more times below). Since the hidden-layer knot vector is equally spaced on $[-H(N), H(N)]$, Lemmas B.2 and B.3 apply with $H = H(N)$. Since $\|K\|_{\mathbf{s}} \geq K\underline{\mathbf{s}}/\tilde{\mathbf{s}} - d$ and $2^{\|K\|_{\mathbf{s}}} \leq N$, we have $2^K \leq (2^d N)^{\tilde{\mathbf{s}}/\underline{\mathbf{s}}}$. Therefore, from $K^* \leq (1+\kappa)K + 1$ and $\tilde{\mathbf{s}} \leq \underline{\mathbf{s}}$, we obtain, with $c_1 := 2^{1+d(1+\kappa)}$,

$$
\max_{(k,j)\in E_N,\ i\in[d]} 2^{a_{k,i}} \leq 2^{K^*} \leq c_1 N^{(1+\kappa)\tilde{\mathbf{s}}/\underline{\mathbf{s}}}. \tag{23}
$$

*Step 1: Affine layer $\Phi_0^\star$.* Choose an injection $\nu_0 : E_N \times [d] \to [D(N)]$, which is possible since $d|E_N| \leq D(N)$. Let $\boldsymbol{\xi}^{(0)}$ be the first-layer knot vector with spacing $\Delta_0 > 0$ satisfying $[\xi_0^{(0)}, \xi_{G_0}^{(0)}] \supseteq [0,1]$. Applying Lemma B.1 to $\boldsymbol{\xi}^{(0)}$ with $I = [0,1]$ and $P(u) = 2^{a_{k,i}} u - j_i$, we obtain

$$
(\Phi_0^\star(\mathbf{x}))_{\nu_0(k,j,i)} = \phi^{\star(0)}_{\nu_0(k,j,i),i}(x_i) = z_{k,j,i}(x_i), \quad x_i \in [0,1].
$$

We denote the coefficients of each edge by $\big(\theta^{\star(0)}_{\nu_0(k,j,i),i,l}\big)_l$. Since $|j_i| \le m + 2^{a_{k,i}} \le (m+1)2^{a_{k,i}}$ and $z_{k,j,i}([0,1]) = [-j_i,\ 2^{a_{k,i}} - j_i] \subseteq [-2^{a_{k,i}},\ 2^{a_{k,i}} + m]$, it follows from Eq. (23) that, with $R_0 := 1 + (m+1)\Delta_0$,

$$\begin{aligned} \max_l \Big|\theta^{\star(0)}_{\nu_0(k,j,i),i,l}\Big| &\le R_0 2^{a_{k,i}} + |j_i| \le \big((R_0+1)c_1 + m\big)N^{(1+\kappa)\tilde{\mathbf{s}}/\underline{\mathbf{s}}}, \\ \max_{(k,j)\in E_N,\ i\in[d]}\ \sup_{x_i\in[0,1]} |z_{k,j,i}(x_i)| &\le (c_1+m)N^{(1+\kappa)\tilde{\mathbf{s}}/\underline{\mathbf{s}}}. \end{aligned} \tag{24}$$

Each edge has at most $G_0 + m$ coefficients.

*Step 2: Cardinal B-spline layer $\Phi^\star_1$.* Using Lemma B.2, we choose the edges $\phi^{\star(1)}_{\nu_0(k,j,i),\nu_0(k,j,i)}$ of the layer $\Phi^\star_1$ such that

$$\phi^{\star(1)}_{\nu_0(k,j,i),\nu_0(k,j,i)} = \psi_m \quad \text{on } \mathbb{R}.$$

The coefficients of each edge satisfy $\|\boldsymbol{\theta}^{\star(1)}_{\nu_0(k,j,i)}\|_0 \le m+2$ and $\|\boldsymbol{\theta}^{\star(1)}_{\nu_0(k,j,i)}\|_\infty \le 2$, and the corresponding knot vector $\boldsymbol{\xi}^{(1)}$ is equally spaced on $[-H(N), H(N)]$ with spacing $1/2$. Therefore,

$$\big(\Phi^\star_1 \circ \Phi^\star_0(\mathbf{x})\big)_{\nu_0(k,j,i)} = \psi_m(z_{k,j,i}(x_i)) \in [0,1].$$

*Step 3: Product layers $\Phi^\star_2, \dots, \Phi^\star_{L_\times+1}$.* If $d = 1$, then $L_\times = 0$ and this step is void. Let $d \ge 2$, let $\Pi_d = \Phi^\times_{L_\times - 1} \circ \cdots \circ \Phi^\times_0$ be the product network of Lemma B.3, and let $n^\times_\ell$ denote the output width of $\Phi^\times_{\ell-1}$. That is, $n^\times_0 = d$, $n^\times_{L_\times} = 1$, and $n^\times_\ell \le \lfloor 3d/2 \rfloor$ otherwise. For each $\ell = 0, \dots, L_\times$, we choose pairwise disjoint node sets $J_\ell(k,j) \subseteq [D(N)]$ with $|J_\ell(k,j)| = n^\times_\ell$ for $(k,j) \in E_N$, where $J_0(k,j) := \{\nu_0(k,j,1), \dots, \nu_0(k,j,d)\}$, and identify the elements of each $J_\ell(k,j)$ with $[n^\times_\ell]$ in increasing order. Each layer uses at most $\lfloor 3d/2 \rfloor |E_N| \le D(N)$ nodes.

For $\ell = 0, \dots, L_\times - 1$, we define the parallel combination $\bigoplus_{(k,j)\in E_N} \Phi^\times_\ell : \mathbb{R}^{D(N)} \to \mathbb{R}^{D(N)}$ of $\Phi^\times_\ell$ as the layer obtained by copying the corresponding edges of $\Phi^\times_\ell$ only to the edges with $p \in J_\ell(k,j)$ and $q \in J_{\ell+1}(k,j)$, and setting all remaining edges, including those between different blocks, to 0. That is,

$$\Big(\bigoplus_{(k,j)\in E_N} \Phi^\times_\ell(\mathbf{v})\Big)_{J_{\ell+1}(k,j)} = \Phi^\times_\ell\big(\mathbf{v}_{J_\ell(k,j)}\big), \qquad (k,j) \in E_N,$$

which is a $KAN_c$ layer of the form Eq. (2). We now set

$$\Phi^\star_{\ell+2} := \bigoplus_{(k,j)\in E_N} \Phi^\times_\ell, \qquad \ell = 0, \dots, L_\times - 1.$$

Let $\mathbf{u}^{(k,j)}(\mathbf{x}) := \big(\psi_m(z_{k,j,i}(x_i))\big)_{i=1}^d \in [0,1]^d$. Since the blocks are computed independently of one another, induction on $\ell$ gives

$$\big(\Phi^\star_{\ell+2} \circ \cdots \circ \Phi^\star_0(\mathbf{x})\big)_{J_{\ell+1}(k,j)} = \Phi^\times_\ell \circ \cdots \circ \Phi^\times_0\big(\mathbf{u}^{(k,j)}(\mathbf{x})\big).$$

In particular, for $\ell = L_\times - 1$, $J_{L_\times}(k,j)$ is a single node, and by Lemma B.3(i), its value is $\Pi_d(\mathbf{u}^{(k,j)}(\mathbf{x})) = M^{d,m}_{k,j}(\mathbf{x})$. By Lemma B.3(ii), all intermediate values lie in $[0,1]$, and both the number and the magnitude of the nonzero weights per block depend only on $(d,m)$.

*Step 4: Output layer $\Phi^\star_{L_\times+2}$.* For each block $(k,j)$, we define the edge from its output node to the network output by Lemma B.1 with $I = [0,1]$ and $P(t) = \alpha_{k,j}t$; this edge has at most $m+3$ coefficients, each bounded in absolute value by $|\alpha_{k,j}|(m+3)/2$. Then $f^\star_N(\mathbf{x}) = f^{\mathrm{sp}}_N(\mathbf{x})$ for $\mathbf{x} \in [0,1]^d$, and Eq. (22) yields Eq. (19).

*Step 5: Architecture and parameter bounds.* From Steps 1–4 and Lemma B.3(iii)(iv), we obtain

$$\begin{aligned} &\text{Depth} : L_0 = 1 + 1 + L_\times + 1 = 3 + 2\lceil \log_2 d \rceil, \\ &\text{Width} : \max\{d, \lfloor 3d/2 \rfloor\}|E_N| \le 2dN = D(N), \\ &\text{Sparsity} : \|\boldsymbol{\theta}^\star_N\|_0 \le S_0|E_N| \le S_0 N, \\ &\qquad S_0 := d(G_0+m) + d(m+2) + \big(7(d-1) + 2\lceil \log_2 d \rceil\big)(m+3) + (m+3), \\ &\text{Weights} : \|\boldsymbol{\theta}^\star_N\|_\infty \le \max\Big\{\big((R_0+1)c_1+m\big)N^{(1+\kappa)\tilde{\mathbf{s}}/\underline{\mathbf{s}}},\ 2,\ \tfrac{(m+3)^2}{2},\ \tfrac{m+3}{2}\max_{(k,j)\in E_N}|\alpha_{k,j}|\Big\} \\ &\qquad \le B_0 N^\beta. \end{aligned} \tag{25}$$

The last inequality holds for sufficiently large $B_0$, since $\beta \geq (1+\kappa)\tilde{\mathbf{s}}/\underline{\mathbf{s}}$ and $\beta \geq (1+\kappa)(1/p - \tilde{\mathbf{s}})_+$ by Eq. (20), together with Eq. (21). As for the ranges of the layers, the second bound in Eq. (24) does not exceed the first, which is in turn bounded by $B^\star(N)$ by Eq. (25). Hence,

$$\max_{k,j,i} \sup_{x_i \in [0,1]} |z_{k,j,i}(x_i)| \leq B^\star(N) \leq H(N),$$

that is, the inputs to $\Phi_1^\star$ lie in $[-H(N), H(N)]$, and the inputs to all subsequent layers lie in $[0,1] \subseteq [-H(N), H(N)]$. Moreover, since $H(N) \leq 2B_0 N^\beta + 1$, we have $H(N) \asymp N^\beta$ and $G(N) = 4H(N) - 2m \geq 2m + 4$. Therefore, $f_N^\star \in KAN_c(L_0, 2dN, 4H(N) - 2m, H(N); m)$, and all the claims hold. □

**Remark B.5** (Enlarging the architecture)**.** *The proof above uses the architecture only through the following three properties: (a) $H(N) \in \mathbb{N}$ with $H(N) \geq \max\{m+1,\ \max_{k,j,i} \sup_{x_i} |z_{k,j,i}(x_i)|\}$; (b) $G(N) = 4H(N) - 2m$, so that the hidden-layer knot spacing is $1/2$; and (c) all weights are bounded by $B^\star(N)$. The representations in Lemmas B.1 to B.3 use only half-integer knots in $[-(m+1)/2,\ m+1]$. Hence, for any $B_0' \geq B_0$ and $\beta' \geq \beta$, with $H'(N) := \lceil 2B_0' N^{\beta'} \rceil$, the same network is realized in $KAN_c(L_0, D(N), 4H'(N) - 2m, H'(N); m)$, up to a reindexing of the hidden-layer basis functions. In particular, $\kappa$ in Eq.* (20) *may be replaced by any $\kappa' \geq \kappa$, and for $p \geq 2$, $\kappa = 0$ is the smallest admissible value.*

### B.3 Compositional approximation

**Lemma B.6** (Layer-wise approximation by fixed-knot $KAN_c$)**.** *For $t \in \mathbb{N}$, $\mathbf{s} \in \mathbb{R}_{++}^t$, and $r \in \{2, \infty\}$, let*

$$\tilde{\mathbf{s}} := \Big(\sum_{i=1}^t s_i^{-1}\Big)^{-1}, \qquad \underline{\mathbf{s}} := \min_i s_i, \qquad \overline{\mathbf{s}} := \max_i s_i, \qquad \omega_r := (1/p - 1/r)_+.$$

*Assume that $\omega_r < \tilde{\mathbf{s}}$ and $0 < \overline{\mathbf{s}} < \min\{m,\ m - 1 + 1/p\}$, and set*

$$\kappa_r := \frac{2\omega_r}{\tilde{\mathbf{s}} - \omega_r}, \qquad \beta_r := (1+\kappa_r)\max\big\{(1/p - \tilde{\mathbf{s}})_+,\ \tilde{\mathbf{s}}/\underline{\mathbf{s}}\big\},$$
$$B_r^*(N) := B_r N^{\beta_r}, \quad H_r^*(N) := \lceil 2B_r^*(N) \rceil$$

*with $\kappa_r = 0$ if $\omega_r = 0$. Then there exist constants $L_r \in \mathbb{N}$ and $D_r, S_r, B_r > 0$, depending only on $t, m, p, \mathbf{s}, r$ and an upper bound on $\|g_0\|_{\mathcal{B}^{\mathbf{s}}_{p,q}([0,1]^t)} \vee \|g_0\|_\infty$ such that for every $N \in \mathbb{N}$ and every bounded function $g_0$ with $g_0|_{[0,1]^t} \in \mathcal{B}^{\mathbf{s}}_{p,q}([0,1]^t)$, there exists $g_N$, realized by a parameter vector $\boldsymbol{\theta}_N$, satisfying*

$$g_N \in KAN_c\big(L_r,\ D_r N,\ 4H_r^*(N) - 2m,\ H_r^*(N);\ m\big),$$
$$\|\boldsymbol{\theta}_N\|_\infty \leq B_r^*(N), \qquad \|\boldsymbol{\theta}_N\|_0 \leq S_r N, \qquad \|g_0 - g_N\|_{L^r([0,1]^t)} \lesssim N^{-\tilde{\mathbf{s}}}.$$

*Proof.* For $r = 2$, the claim follows by applying Lemma B.4 to the $t$-variate function $g_0$. For $r = \infty$, Lemma 2 of Suzuki and Nitanda (2021) applies since $\omega_\infty = 1/p < \tilde{\mathbf{s}}$. With $N_0$ and $K$ chosen as in Step 0 of the proof of Lemma B.4, now with $\|K\|_{\mathbf{s}} = \sum_{i=1}^t \lfloor K\underline{\mathbf{s}}/s_i \rfloor$, it yields for $N \geq N_0$ an expansion $g_N^{\mathrm{sp}}$ of the form Eq. (21) with $|E_N| \leq N$, $k \leq K^* = \lceil (1+\kappa_\infty)K \rceil$, $\max|\alpha_{k,j}| \lesssim N^{(1+\kappa_\infty)(1/p - \tilde{\mathbf{s}})_+}$, and $\|g_0 - g_N^{\mathrm{sp}}\|_{L^\infty([0,1]^t)} \lesssim N^{-\tilde{\mathbf{s}}}$; in particular, Eq. (23) holds with $(d, \kappa)$ replaced by $(t, \kappa_\infty)$. Steps 1–5 of that proof use only these properties, not the norm in which the error is measured, so the same construction realizes $g_N^{\mathrm{sp}}$ with depth $3 + 2\lceil \log_2 t \rceil$, width $2tN$, sparsity $S_r N$, and weights bounded by $B_r N^{\beta_r}$. For $N < N_0$, set $g_N = 0$ as in Step 0. □

**Lemma B.7** (Approximate clipping)**.** *Let $m \geq 2$, $H \in \mathbb{N}$ with $H \geq 2m + 3$, $G := 4H - 2m$, and let $\boldsymbol{\xi}$ be the knot vector of Lemma B.2. For $a \geq 1$, let $\eta_a := 2/(ea)$. There exist KAN layers $\Phi_0^{\mathrm{rc}} : \mathbb{R} \to \mathbb{R}^3$ and $\Phi_1^{\mathrm{rc}} : \mathbb{R}^3 \to \mathbb{R}$ with knot vector $\boldsymbol{\xi}$ such that $R_a := \Phi_1^{\mathrm{rc}} \circ \Phi_0^{\mathrm{rc}}$ satisfies the following:*

*(i)* $R_a(\mathbb{R}) \subseteq [0,1]$;

*(ii)* $|R_a(u) - v| \leq |u - v| + 2\eta_a$ *for all $u \in \mathbb{R}$ and $v \in [0,1]$;*

*(iii) $R_a$ has at most $18m + 40$ nonzero weights, each bounded in absolute value by $(m + 3)a/2$.*

*Proof.* Write $\mathrm{clip}_{[\alpha,\beta]}(w) := \min\{\beta, \max\{\alpha, w\}\}$ and $F_a(x) := a^{-1}\,\mathrm{SiLU}(ax)$. Since $|\mathrm{SiLU}(x) - x_+| = |x|/(1 + e^{|x|}) \le e^{-1}$,

$$\sup_{x\in\mathbb{R}} |F_a(x) - x_+| \le (ea)^{-1} = \eta_a/2. \tag{26}$$

We approximate $\mathrm{clip}_{[0,1]}(w) = w_+ - (w - 1)_+$ by replacing $x_+$ with $F_a$. Since the spline edges vanish outside $[\xi_{-m}, \xi_{G+m}]$ and the SiLU edges cannot produce a constant, the first layer maps $u$ into a bounded interval by a spline $C$ and produces the constant 1 by a second spline $S$.

Let $\varrho := (m + 1)/2$, $I := [-1 - 2\varrho,\ 2 + 2\varrho]$, $\mathcal{K} := \{k : (\xi_{k-m-1}, \xi_k) \cap I \ne \emptyset\}$, and $\xi_k^* := (\xi_{k-m} + \cdots + \xi_{k-1})/m$, which equals $\lambda_{k,1}$ in Eq. (15). Define

$$C(u) := \sum_{k\in\mathcal{K}} \mathrm{clip}_{[-\varrho,1+\varrho]}(\xi_k^*)\, B_{k,m}(u;\boldsymbol{\xi}), \qquad S(u) := \sum_{k\in\mathcal{K}} B_{k,m}(u;\boldsymbol{\xi}), \qquad u \in \mathbb{R}.$$

Since $H \ge 2m + 3$, we have $I \subseteq [\xi_0, \xi_G]$. We claim that

(a) $0 \le S \le 1$ on $\mathbb{R}$, and $S = 1$ on $I$;

(b) $C(\mathbb{R}) \subseteq [-\varrho, 1 + \varrho]$;

(c) $|C(u) - v| \le |u - v|$ for $u \in I$ and $v \in [0, 1]$.

Claims (a) and (b) follow from $B_{k,m} \ge 0$ and the partition of unity on $[\xi_0, \xi_G]$: every $B_{k,m}$ that does not vanish on $I$ has $k \in \mathcal{K}$, and $C(u)$ is a nonnegative combination of values in $[-\varrho, 1 + \varrho] \ni 0$ with total weight $S(u) \le 1$. For (c), we first show that $C(u) = u$ on $[0, 1]$. If $B_{k,m}(u) \ne 0$, then $|\xi_k^* - u| \le (m + 1)\Delta = \varrho$, so the clipping is inactive, and $C(u) = \sum_k \xi_k^* B_{k,m}(u) = u$ by Eq. (15). Next, $C$ is 1-Lipschitz on $I$: by Eq. (13), for $u \in I$, $C'(u)$ is a nonnegative combination, with total weight at most 1, of the difference quotients $\{\mathrm{clip}_{[-\varrho,1+\varrho]}(\xi_{k+1}^*) - \mathrm{clip}_{[-\varrho,1+\varrho]}(\xi_k^*)\}/\Delta \in [-1, 1]$ with $k, k + 1 \in \mathcal{K}$ (the boundary terms $k = \min\mathcal{K} - 1$ and $k = \max\mathcal{K}$ vanish on $I$, since $B_{k,m-1}$ is supported on $[\xi_{k-m}, \xi_k]$ while $\xi_{\min\mathcal{K}-1} \le \inf I$ and $\xi_{\max\mathcal{K}-m} \ge \sup I$ by the definition of $\mathcal{K}$). Hence, for $u \in I$ and $v \in [0, 1] \subseteq I$, $|C(u) - v| = |C(u) - C(v)| \le |u - v|$. Let $\varphi_{\mathrm{c}}$ be the spline edge of Lemma B.1 that equals the constant $\eta_a/(1 + 2\eta_a)$ on $[-\varrho, 1 + \varrho]$, and set

$$\begin{aligned} \Phi_0^{\mathrm{rc}}(u) &:= \big(C(u),\ aC(u),\ a\{C(u) - S(u)\}\big), \\ \Phi_1^{\mathrm{rc}}(z, y_1, y_2) &:= \varphi_{\mathrm{c}}(z) + \frac{\mathrm{SiLU}(y_1) - \mathrm{SiLU}(y_2)}{a(1 + 2\eta_a)}. \end{aligned} \tag{27}$$

Each coordinate of $\Phi_0^{\mathrm{rc}}$ is a single spline edge, and $\Phi_1^{\mathrm{rc}}$ consists of one spline edge and two SiLU weights. By (b), $\varphi_{\mathrm{c}}(C(u)) = \eta_a/(1 + 2\eta_a)$, so with $w := C(u)$ and $s := S(u)$,

$$R_a(u) = \frac{F_a(w) - F_a(w - s) + \eta_a}{1 + 2\eta_a}.$$

Fix $u \in \mathbb{R}$ and let $c := w_+ - (w - s)_+ = \mathrm{clip}_{[0,s]}(w) \in [0, 1]$. By Eq. (26), $F_a(w) - F_a(w - s) = c + \delta$ with $|\delta| \le \eta_a$, and therefore

$$R_a(u) = \frac{c + \delta + \eta_a}{1 + 2\eta_a} \in [0, 1], \qquad |R_a(u) - c| = \frac{|\delta + \eta_a(1 - 2c)|}{1 + 2\eta_a} \le 2\eta_a,$$

which proves (i). For (ii), let $v \in [0, 1]$. If $u \in I$, then $s = 1$ by (a), so $c = \mathrm{clip}_{[0,1]}(C(u))$ and $|c - v| \le |C(u) - v| \le |u - v|$ by (c). If $u \notin I$, then $|u - v| \ge 1 + 2\varrho \ge 1 \ge |c - v|$. In both cases, $|R_a(u) - v| \le |c - v| + 2\eta_a \le |u - v| + 2\eta_a$. By Lemma B.1(ii), $|\mathcal{K}| \le \lceil |I|/\Delta \rceil + m + 1 = 5m + 11$ and $\|\varphi_{\mathrm{c}}\|_0 \le 3m + 5$, so the number of nonzero weights is at most $3(5m + 11) + (3m + 5) + 2 = 18m + 40$. The weights of $\Phi_0^{\mathrm{rc}}$ are bounded by $a(1 + \varrho) = (m + 3)a/2$, and those of $\Phi_1^{\mathrm{rc}}$ by 1. □

**Lemma B.8** (Hölder continuity of the outer layers)**.** *Assume* (D1) *and* (D2)*, and for $k = 2, \ldots, J$, let $\gamma_k := 1 - 1/(p\tilde{s}^{(k)})$ and $\alpha_k := \gamma_k \underline{\mathbf{s}}^{(k)} = \underline{\mathbf{s}}^{(k)} - t^{*(k)}/p$. Then, for all $k \ge 2$ and $i \in [d^{(k)}]$, $\tilde{f}_{k,i}$*

*has a continuous representative, and there exist constants $A_k, \Lambda_j > 0$ such that*

$$\left|\tilde{f}_{k,i}(\mathbf{y}) - \tilde{f}_{k,i}(\mathbf{y}')\right| \le A_k \|\mathbf{y} - \mathbf{y}'\|_\infty^{\alpha_k \wedge 1},$$

$$|F_{0,j}(\mathbf{u}) - F_{0,j}(\mathbf{v})| \le \Lambda_j \|\mathbf{u} - \mathbf{v}\|_\infty^{\Gamma_j}, \qquad \Gamma_j := \prod_{k=j+1}^{J} (\alpha_k \wedge 1),$$

*for all $\mathbf{y}, \mathbf{y}' \in [0,1]^{t^{(k)}}$, $\mathbf{u}, \mathbf{v} \in [0,1]^{d^{(j)}}$, and $j \in [J]$. Here, $F_{0,j} := f_{0,J} \circ \cdots \circ f_{0,j+1}$, $F_{0,J} := \mathrm{id}$, and $\Gamma_J := 1$.*

*Proof.* Fix $k \ge 2$ and $i$, and write $f := \tilde{f}_{k,i}$, $t := t^{(k)}$, and $s'_\ell := \gamma_k s_\ell^{(k)}$, so that $\min_\ell s'_\ell = \alpha_k$. Since $\tilde{s}^{(k)} > 1/p$, we have $\gamma_k \in (0,1]$, and by Suzuki and Nitanda (2021, Proposition 1, item 2) with $(p_1, p_2) = (p, \infty)$, $f \in \mathcal{B}^{\mathbf{s}'}_{\infty,q}([0,1]^t)$. Moreover, by Suzuki and Nitanda (2021, Proposition 1, item 4), $\mathcal{B}^{\mathbf{s}^{(k)}}_{p,q}([0,1]^t) \hookrightarrow C^0([0,1]^t)$, so $f$ has a continuous representative; we identify $f$ with it throughout the proof.

Let $\omega_{r,\ell}(f,\tau) := \sup_{|h|\le\tau} \left\|\Delta^r_{h\mathbf{e}_\ell}(f)\right\|_{L^\infty}$ with $r = \max_\ell \lfloor s'_\ell \rfloor + 1$. Restricting the supremum in the anisotropic modulus to $\mathbf{h} = h\mathbf{e}_\ell$ gives $\omega_{r,\ell}(f, 2^{-n/s'_\ell}) \le 2^{-n} |f|_{\mathcal{B}^{\mathbf{s}'}_{\infty,q}}$, and by monotonicity,

$$\omega_{r,\ell}(f,\tau) \le 2|f|_{\mathcal{B}^{\mathbf{s}'}_{\infty,q}} \tau^{s'_\ell}, \qquad \tau \in (0,1]. \tag{28}$$

Fix $\ell$ and the remaining coordinates $\mathbf{y}_{-\ell}$, and let $g(u) := f(\mathbf{y}_{-\ell}, u)$ for $u \in [0,1]$, where $u$ occupies the $\ell$th coordinate. Denote by $\omega_r(g,\cdot)$ the univariate $r$th modulus of smoothness of $g$ on $[0,1]$. Since $f$ is continuous, $\omega_r(g,\tau) \le \omega_{r,\ell}(f,\tau)$ and $\|g\|_{L^\infty} \le \|f\|_{L^\infty}$ for every $\mathbf{y}_{-\ell}$, while $\omega_{1,\ell}(f,\tau) \le \sup_{\mathbf{y}_{-\ell}} \omega_1(g,\tau)$. Applying the Marchaud inequality (DeVore and Lorentz, 1993, Ch. 2) to $g$, whose constant does not depend on $\mathbf{y}_{-\ell}$, taking the supremum over $\mathbf{y}_{-\ell}$, and substituting Eq. (28), we obtain

$$\omega_{1,\ell}(f,\tau) \lesssim \tau\Big(\int_\tau^1 \frac{\omega_{r,\ell}(f,u)}{u^2}\, du + \|f\|_{L^\infty}\Big) \lesssim \begin{cases} \tau^{s'_\ell \wedge 1}, & s'_\ell \ne 1, \\ \tau \log(e/\tau), & s'_\ell = 1, \end{cases} \qquad \tau \in (0,1]. \tag{29}$$

- $\alpha_k < 1$: Since $s'_\ell \ge \alpha_k$ and $\tau\log(e/\tau) \lesssim \tau^{\alpha_k}$, we have $\omega_{1,\ell}(f,\tau) \lesssim \tau^{\alpha_k}$ for all $\ell$.
- $\alpha_k > 1$: Since $s'_\ell > 1$ for all $\ell$, we have $\omega_{1,\ell}(f,\tau) \lesssim \tau$.
- $\alpha_k = 1$: By (D2), $q \le 1$. For $\ell$ with $s'_\ell = 1$, the definition of the quasi-norm and $q \le 1$ yield

  $$\int_\tau^1 \frac{\omega_{r,\ell}(f,u)}{u^2}\, du \le \sum_{n\ge 0} 2^n \omega_{r,\ell}(f, 2^{-n}) \le \Big(\sum_{n\ge 0} \{2^n \omega_{r,\ell}(f,2^{-n})\}^q\Big)^{1/q} \le |f|_{\mathcal{B}^{\mathbf{s}'}_{\infty,q}},$$

  and hence the first inequality in Eq. (29) gives $\omega_{1,\ell}(f,\tau) \lesssim \tau$.

In all cases, $\omega_{1,\ell}(f,\tau) \lesssim \tau^{\alpha_k \wedge 1}$.

Changing one coordinate at a time gives $|f(\mathbf{y}) - f(\mathbf{y}')| \le \sum_\ell \omega_{1,\ell}(f, |y_\ell - y'_\ell|) \le A_k \|\mathbf{y} - \mathbf{y}'\|_\infty^{\alpha_k \wedge 1}$. Since $f_{0,k,i}(\mathbf{u}) = \tilde{f}_{k,i}(\mathbf{u}_{I_{k,i}})$, we have $\|f_{0,k}(\mathbf{u}) - f_{0,k}(\mathbf{v})\|_\infty \le A_k \|\mathbf{u} - \mathbf{v}\|_\infty^{\alpha_k \wedge 1}$. Backward induction starting from $j = J$ (with $\Gamma_J = 1$) then yields

$$\begin{aligned} |F_{0,j-1}(\mathbf{u}) - F_{0,j-1}(\mathbf{v})| &= |F_{0,j}(f_{0,j}(\mathbf{u})) - F_{0,j}(f_{0,j}(\mathbf{v}))| \le \Lambda_j \|f_{0,j}(\mathbf{u}) - f_{0,j}(\mathbf{v})\|_\infty^{\Gamma_j} \\ &\le \Lambda_j A_j^{\Gamma_j} \|\mathbf{u} - \mathbf{v}\|_\infty^{(\alpha_j \wedge 1)\Gamma_j} = \Lambda_j A_j^{\Gamma_j} \|\mathbf{u} - \mathbf{v}\|_\infty^{\Gamma_{j-1}}, \end{aligned}$$

where the induction hypothesis applies because $f_{0,j}$ maps $[0,1]^{d^{(j-1)}}$ into $[0,1]^{d^{(j)}}$. □

We elaborate on the roles of (D1) and (D2). Since Hölder exponents multiply under composition, an error of order $N^{-\tilde{s}^{(j)}}$ in the $j$th layer becomes $N^{-\tilde{s}^{(j)} \prod_{k>j} (\alpha_k \wedge 1)} = N^{-\tilde{s}^{*(j)}}$ at the output. If $\alpha_k > 1$, Lipschitz continuity suffices, which explains the truncation $\wedge 1$. For the innermost layer,

only the $L^2$-approximation condition $\tilde{s}^{(1)} > (1/p - 1/2)_+$ is required. The last condition in (D2) excludes only the boundary case $\alpha_j = 1$ when $q > 1$. In that case, elements of $\mathcal{B}^{\gamma_j \mathbf{s}^{(j)}}_{\infty,q}$ need not be Lipschitz (e.g., $u \mapsto u \log u$ for $q = \infty$), so the factor $\alpha_j \wedge 1 = 1$ would incur an additional logarithmic loss through the term $\tau \log(e/\tau)$.

**Lemma B.9** (Compositional approximation by fixed-knot KAN)**.** *Assume* (D1) *and* (D2)*. There exist constants $L_{\mathrm{cp}} \in \mathbb{N}$ and $D_{\mathrm{cp}}, S_{\mathrm{cp}}, B_{\mathrm{cp}}, \beta_{\mathrm{cp}} > 0$ such that for every $N \in \mathbb{N}$, there exists $f_N^{\mathrm{cp},*}$, realized by a parameter vector $\boldsymbol{\theta}_N^{\mathrm{cp}}$, satisfying*

$$f_N^{\mathrm{cp},*} \in KAN\big(L_{\mathrm{cp}},\, D_{\mathrm{cp}}N,\, 4H^*_{\mathrm{cp}}(N) - 2m,\, H^*_{\mathrm{cp}}(N);\, m\big),$$
$$\|\boldsymbol{\theta}_N^{\mathrm{cp}}\|_\infty \le B^*_{\mathrm{cp}}(N), \qquad \|\boldsymbol{\theta}_N^{\mathrm{cp}}\|_0 \le S_{\mathrm{cp}}N,$$

*and*

$$\|f_0 - f_N^{\mathrm{cp},*}\|_{L^2(P_X)} \lesssim N^{-\tilde{s}^*}. \tag{30}$$

*Here, $B^*_{\mathrm{cp}}(N) := B_{\mathrm{cp}}N^{\beta_{\mathrm{cp}}}$, $H^*_{\mathrm{cp}}(N) := \lceil 2B^*_{\mathrm{cp}}(N)\rceil$, and $L_{\mathrm{cp}} = \sum_{j=1}^{J}(3+2\lceil\log_2 t^{(j)}\rceil)+2(J-1)$.*

*Proof.* If $J = 1$, the claim reduces to Lemma B.4; hence we assume $J \ge 2$.

Let $r^{(1)} := 2$ and $r^{(j)} := \infty$ for $j \ge 2$. By (D1) and (D2), the hypotheses of Lemma B.6 hold for $(t^{(j)}, \mathbf{s}^{(j)}, r^{(j)})$, namely $\omega_{r^{(j)}} < \tilde{s}^{(j)}$ and $\bar{\mathbf{s}}^{(j)} < \min\{m,\, m - 1 + 1/p\}$. Let $L_j, D_j, S_j, B_j, \beta_j$, and $H^*_j(N)$ denote its constants, so that $L_j = 3 + 2\lceil\log_2 t^{(j)}\rceil$. Since each $\tilde{f}_{j,k}$ is bounded (it takes values in $[0,1]$ for $j < J$, and for $j = J$ it is continuous by Lemma B.8 or equals $f_0 \in \mathcal{UB}$) and there are finitely many $(j,k)$, the constants may be taken uniform in $k$. Thus, for all $j \in [J]$ and $k \in [d^{(j)}]$, there exists

$$h_{j,k,N} \in KAN_c\big(L_j,\, D_jN,\, 4H^*_j(N) - 2m,\, H^*_j(N);\, m\big),$$

realized by $\boldsymbol{\theta}_{j,k,N}$ with $\|\boldsymbol{\theta}_{j,k,N}\|_\infty \le B_jN^{\beta_j}$ and $\|\boldsymbol{\theta}_{j,k,N}\|_0 \le S_jN$, such that

$$\|\tilde{f}_{j,k} - h_{j,k,N}\|_{L^{r^{(j)}}([0,1]^{t^{(j)}})} \lesssim N^{-\tilde{s}^{(j)}}. \tag{31}$$

For $j \ge 2$, both $\tilde{f}_{j,k}$ (Lemma B.8) and $h_{j,k,N}$ are continuous, so Eq. (31) holds pointwise on $[0,1]^{t^{(j)}}$.

For $j \ge 2$, the first layer of $h_{j,k,N}$ becomes a hidden layer of the composite network: its knot vector is $\boldsymbol{\xi}$ of Lemma B.2 instead of $\boldsymbol{\xi}^{(0)}$, and its input lies in $[0,1] \subseteq [\xi_0, \xi_G]$ as shown below. Step 1 in the proof of Lemma B.4 applies to this layer with $\Delta_0 = 1/2$, so that $R_0 = (m+3)/2$; let $B'_j \ge B_j$ denote the resulting weight bound, and set $B'_1 := B_1$. To place all components on a common grid, set

$$\beta_{\mathrm{cp}} := \max\Big\{\max_{j\in[J]}\beta_j,\ \max_{j\in[J-1]}\tilde{s}^{(j)}\Big\}, \qquad B_{\mathrm{cp}} := \max\Big\{m+2,\ \max_{j\in[J]}B'_j\Big\},$$

and let $H := H^*_{\mathrm{cp}}(N)$ and $G := 4H - 2m$. Then $H \ge 2(m+2) > 2m+3$, so Lemma B.7 applies on the grid $(G,H)$, and since $B_{\mathrm{cp}} \ge B'_j$ and $\beta_{\mathrm{cp}} \ge \beta_j$, Remark B.5 realizes each $h_{j,k,N}$ on the grid $(G,H)$ with the same weights. We extend $h_{j,k,N}$ to $d^{(j-1)}$ input variables by setting the edges from the coordinates outside $I_{j,k}$ to zero, and place the $d^{(j)}$ components, all of depth $L_j$, in parallel to obtain $h_{j,N} := (h_{j,1,N}, \dots, h_{j,d^{(j)},N})$.

For $j \in [J-1]$, let $a_{j,N} := N^{\tilde{s}^{(j)}}$, $g_{j,k,N} := R_{a_{j,N}} \circ h_{j,k,N}$, and $g_{j,N} := (g_{j,1,N}, \dots, g_{j,d^{(j)},N})$; let $g_{J,N} := h_{J,N}$. By Lemma B.7(i), $g_{j,N}(\mathbb{R}^{d^{(j-1)}}) \subseteq [0,1]^{d^{(j)}}$ for $j \in [J-1]$. Since $f_{0,j,k}$ takes values in $[0,1]$, Lemma B.7(ii) gives, for $j \in [J-1]$ and $\mathbf{y} \in [0,1]^{d^{(j-1)}}$,

$$|g_{j,k,N}(\mathbf{y}) - f_{0,j,k}(\mathbf{y})| \le \Big|h_{j,k,N}(\mathbf{y}_{I_{j,k}}) - \tilde{f}_{j,k}(\mathbf{y}_{I_{j,k}})\Big| + 2\eta_{a_{j,N}},\ 2\eta_{a_{j,N}} = \frac{4}{e}N^{-\tilde{s}^{(j)}}. \tag{32}$$

For $2 \le j \le J-1$, Eq. (32) and the pointwise form of Eq. (31) give the first bound below; for $j = J$, it is Eq. (31) itself:

$$\sup_{\mathbf{y}\in[0,1]^{d^{(j-1)}}} \|g_{j,N}(\mathbf{y}) - f_{0,j}(\mathbf{y})\|_\infty \lesssim N^{-\tilde{s}^{(j)}}, \qquad j = 2, \dots, J. \tag{33}$$

For $j = 1$, the marginal density of $X_{I_{1,k}}$ is bounded by $C_X$ by (A2), so Eq. (32) and Eq. (31) give

$$\|g_{1,k,N} - f_{0,1,k}\|_{L^2(P_X)} \le C_X^{1/2} \Big\| h_{1,k,N} - \tilde{f}_{1,k} \Big\|_{L^2([0,1]^{t^{(1)}})} + 2\eta_{a_{1,N}} \lesssim N^{-\tilde{s}^{(1)}}. \tag{34}$$

Let $f_N^{\mathrm{cp},*} := g_{J,N} \circ \cdots \circ g_{1,N}$. The $j$th block $h_{j,N}$ has depth $L_j$, and each of the $J-1$ range-correction modules adds two layers, so the depth of $f_N^{\mathrm{cp},*}$ is $\sum_{j=1}^{J} L_j + 2(J-1) = L_{\mathrm{cp}}$. The $d^{(j)}$ parallel components of $h_{j,N}$ have width at most $D_j N$ each, and the range-correction module following it has width $3d^{(j)}$, so padding every hidden layer by zero edges to the width $D_{\mathrm{cp}} N$ with $D_{\mathrm{cp}} := \max_j d^{(j)}(D_j + 3)$ suffices. The number of nonzero parameters is at most $\sum_{j=1}^{J} d^{(j)}(S_j + 18m + 40)\, N =: S_{\mathrm{cp}} N$. Every weight of $h_{j,N}$ is bounded in absolute value by $B'_j N^{\beta_j} \le B_{\mathrm{cp}} N^{\beta_{\mathrm{cp}}}$, and every weight of a range-correction module by $(m+3)a_{j,N}/2 \le (m+2)N^{\beta_{\mathrm{cp}}} \le B_{\mathrm{cp}} N^{\beta_{\mathrm{cp}}}$, since $a_{j,N} = N^{\tilde{s}^{(j)}} \le N^{\beta_{\mathrm{cp}}}$. Hence $f_N^{\mathrm{cp},*}$ belongs to the class stated in the lemma.

Let $G_{0,N} := \mathrm{id}$ and $G_{j,N} := g_{j,N} \circ \cdots \circ g_{1,N}$. As shown above, $G_{j,N}([0,1]^d) \subseteq [0,1]^{d^{(j)}}$ for $j \in [J-1]$. Since $F_{0,j}$ is evaluated only on $[0,1]^{d^{(j)}}$, the telescoping identity

$$f_0 - f_N^{\mathrm{cp},*} = \sum_{j=1}^{J} \Big[ F_{0,j} \circ f_{0,j} \circ G_{j-1,N} - F_{0,j} \circ g_{j,N} \circ G_{j-1,N} \Big]$$

is well defined on $[0,1]^d$. Writing $\mathbf{y} := G_{j-1,N}(\mathbf{x})$, Lemma B.8, Eq. (33), and Eq. (34) give

$$|F_{0,j}(f_{0,j}(\mathbf{y})) - F_{0,j}(g_{j,N}(\mathbf{y}))| \le \Lambda_j \|f_{0,j}(\mathbf{y}) - g_{j,N}(\mathbf{y})\|_\infty^{\Gamma_j} \lesssim N^{-\tilde{s}^{(j)}\Gamma_j} = N^{-\tilde{s}^{*(j)}}, \qquad j \ge 2,$$

$$\begin{aligned} \|F_{0,1} \circ f_{0,1} - F_{0,1} \circ g_{1,N}\|_{L^2(P_X)}^2 &\le \Lambda_1^2 \Big( \int \|f_{0,1} - g_{1,N}\|_\infty^2 \, dP_X \Big)^{\Gamma_1} \\ &\le \Lambda_1^2 \Big( \sum_k \|f_{0,1,k} - g_{1,k,N}\|_{L^2(P_X)}^2 \Big)^{\Gamma_1} \lesssim N^{-2\tilde{s}^{*(1)}}, \end{aligned}$$

where, for $j = 1$, we applied Jensen's inequality to the concave map $u \mapsto u^{\Gamma_1}$ ($\Gamma_1 \le 1$). Summing over $j$ and using $\tilde{s}^* = \min_j \tilde{s}^{*(j)}$ yields Eq. (30). □

### B.4 Model complexity

#### B.4.1 Layerwise bounds for KANs

**Lemma B.10.** *Let $L, m \in \mathbb{N}$ ($m \ge 2$), $\mathbf{D} \in \mathbb{N}^{L+1}$, $\mathbf{G} \in \mathbb{N}^L$, and $H > 0$. For any $f_{\boldsymbol{\theta}} \in KAN(L, \mathbf{D}, \mathbf{G}, H; m)$, write $f_{\boldsymbol{\theta}} = \Phi_{L-1} \circ \cdots \circ \Phi_0$. Let its parameters be denoted by $\boldsymbol{\theta} = (\mathbf{W}^{(l)}) = (\theta_{i,j,k}^{(l)})$. Also, let the minimum spacing of the knots in layer $l$ of $f_{\boldsymbol{\theta}}$ be*

$$\Delta^{(l)} = \min_k |\xi_{k-m-1}^{(l)} - \xi_{k-m}^{(l)}|, \quad l = 1, \ldots, L-1$$

*Then the following holds. For each $l = 1, \ldots, L-1$, there exists a constant $C_l > 0$ such that*

$$|(\Phi_{L-1} \circ \cdots \circ \Phi_l)(\mathbf{x}) - (\Phi_{L-1} \circ \cdots \circ \Phi_l)(\mathbf{x}^*)| \le C_l \cdot \|\mathbf{x} - \mathbf{x}^*\|_\infty, \quad \textit{for all } \mathbf{x}, \mathbf{x}^* \in \mathbb{R}^{D_l}.$$

*Here,*

$$C_l := \prod_{k=l}^{L-1} \left( \frac{2m}{\Delta^{(k)}} + m \right) \|\mathbf{W}^{(k)}\|_\infty D_k$$

*Proof.* We prove the claim by induction on the number of composed functions, $h = L - l$. First, suppose that $h = 1$. Fix $\mathbf{x}, \mathbf{x}^* \in \mathbb{R}^{D_{L-1}}$. By the derivative properties Eq. (13) and Eq. (14) of the B-spline curve $s_m$, together with the mean value theorem,

$$\begin{aligned} &|\phi_{1,j}^{(L-1)}(x_j) - \phi_{1,j}^{(L-1)}(x_j^*)| \\ &\le |s_m(x_j; \boldsymbol{\theta}_{1,j,-0}^{(L-1)}) - s_m(x_j^*; \boldsymbol{\theta}_{1,j,-0}^{(L-1)})| + |\theta_{1,j,0}^{(L-1)}| \, |\operatorname{SiLU}(x_j) - \operatorname{SiLU}(x_j^*)| \\ &\le \left( \frac{2\|\boldsymbol{\theta}_{1,j,-0}^{(L-1)}\|_\infty}{\Delta^{(L-1)}} + c_S |\theta_{1,j,0}^{(L-1)}| \right) \cdot |x_j - x_j^*| \\ &\le \big( 2m(\Delta^{(L-1)})^{-1} + m \big) \cdot \|\mathbf{W}^{(L-1)}\|_\infty \cdot |x_j - x_j^*| \end{aligned} \tag{35}$$

for $j = 1, \ldots, D_{L-1}$. The second inequality follows from Eq. (14) and the mean value theorem applied to SiLU, where $c_{\mathrm{S}} := \sup_z \left|\mathrm{SiLU}'(z)\right| \in (1, 1.1)$ (note that SiLU is not 1-Lipschitz). The last inequality uses $2 \leq 2m$ and $c_{\mathrm{S}} < 2 \leq m$. Summing over $j = 1, \ldots, D_{L-1}$ yields

$$|\Phi_{L-1}(\mathbf{x}) - \Phi_{L-1}(\mathbf{x}^*)| \leq \big(2m(\Delta^{(L-1)})^{-1} + m\big)\|\mathbf{W}^{(L-1)}\|_\infty D_{L-1} \cdot \|\mathbf{x} - \mathbf{x}^*\|_\infty \tag{36}$$

Now suppose that the conclusion holds for some $h \geq 1$. We show that the inequality also holds for $h+1$. Let $\mathbf{x}, \mathbf{x}^* \in \mathbb{R}^{D_{L-h-1}}$. By the induction hypothesis, we have

$$\begin{aligned}
&|\Phi_{L-1} \circ \cdots \circ \Phi_{L-h-1}(\mathbf{x}) - \Phi_{L-1} \circ \cdots \circ \Phi_{L-h-1}(\mathbf{x}^*)| \\
&\leq \left( \prod_{k=L-h}^{L-1} \big(2m(\Delta^{(k)})^{-1} + m\big) \cdot \|\mathbf{W}^{(k)}\|_\infty \cdot D_k \right) \cdot \|\Phi_{L-h-1}(\mathbf{x}) - \Phi_{L-h-1}(\mathbf{x}^*)\|_\infty.
\end{aligned} \tag{37}$$

Moreover, for $i = 1, \ldots, D_{L-h}$, by the same argument as in Eq. (35), we have

$$|\phi_{i,j}^{(L-h-1)}(x_j) - \phi_{i,j}^{(L-h-1)}(x_j^*)| \leq \big(2m(\Delta^{(L-h-1)})^{-1} + m\big)\|\mathbf{W}^{(L-h-1)}\|_\infty \cdot |x_j - x_j^*|$$

for $j = 1, \ldots, D_{L-h-1}$. Applying this for all $i$, as in Eq. (36), gives

$$\|\Phi_{L-h-1}(\mathbf{x}) - \Phi_{L-h-1}(\mathbf{x}^*)\|_\infty \leq \big(2m(\Delta^{(L-h-1)})^{-1} + m\big) \cdot \|\mathbf{W}^{(L-h-1)}\|_\infty \cdot D_{L-h-1} \cdot \|\mathbf{x} - \mathbf{x}^*\|_\infty.$$

Substituting this into Eq. (37) completes the proof. □

**Lemma B.11.** *Let $L, m \in \mathbb{N}$, $\mathbf{D} \in \mathbb{N}^{L+1}$, $\mathbf{G} \in \mathbb{N}^L$, and $H > 0$. Let $f_{\boldsymbol{\theta}} \in KAN(L, \mathbf{D}, \mathbf{G}, H; m)$, $f_{\boldsymbol{\theta}} = \Phi_{L-1} \circ \cdots \circ \Phi_0$, and write $\boldsymbol{\theta} = (\mathbf{W}^{(l)})_l = (\theta_{i,j,k}^{(l)})_{i,j,k,l}$. Assume that $|a_0| \vee |b_0| \geq 1$. Then the following holds: For each $l = 0, \ldots, L-1$,*

$$\|\Phi_l \circ \cdots \circ \Phi_0(\mathbf{x})\|_\infty \leq 2 \cdot (|a_0| \vee |b_0|) \cdot \prod_{k=0}^{l} D_k \cdot \Big(\sum_{h=0}^{l} \prod_{k=l-h}^{l} \|\mathbf{W}^{(k)}\|_\infty\Big) \quad \textit{for all } \mathbf{x} \in [a_0, b_0]^{D_0}.$$

*Proof.* Fix $\mathbf{x} = (x_1, \ldots, x_{D_0}) \in [a_0, b_0]^{D_0}$. When $l = 0$, for $i = 1, \ldots, D_1$,

$$\begin{aligned}
|\phi_{i,j}^{(0)}(x_j)| &\leq |s_m(x_j; \boldsymbol{\theta}_{i,j,-0}^{(0)}, \boldsymbol{\xi}^{(0)})| + |\theta_{i,j,0}^{(0)} \cdot \mathrm{SiLU}(x_j)| \\
&\leq \|\mathbf{W}^{(0)}\|_\infty + (|a_0| \vee |b_0|) \cdot \|\mathbf{W}^{(0)}\|_\infty \\
&\leq 2 \cdot (|a_0| \vee |b_0|) \cdot \|\mathbf{W}^{(0)}\|_\infty \quad (j = 1, \ldots, D_0)
\end{aligned}$$

The second inequality follows from the partition of unity property Eq. (12) of B-splines and the bound $|\mathrm{SiLU}(z)| = |z/(1+e^{-z})| \leq |z| \leq |a_0| \vee |b_0|$ for all $a_0 \leq z \leq b_0$. Hence,

$$\|\Phi_0(\mathbf{x})\|_\infty \leq 2 \cdot (|a_0| \vee |b_0|) \cdot D_0 \cdot \|\mathbf{W}^{(0)}\|_\infty \quad (i = 1, \ldots, D_1)$$

and the conclusion follows.

Next, suppose that the conclusion holds for some $l \geq 0$. We show that it also holds for $l+1$. Denote the output of the previous layer $\Phi_l : \mathbb{R}^{D_l} \to \mathbb{R}^{D_{l+1}}$ by $z_j := [\Phi_l \circ \cdots \circ \Phi_0(\mathbf{x})]_j$, $j = 1, \ldots, D_{l+1}$. Then, by an argument similar to the case $l = 0$ and by the induction hypothesis,

$$\begin{aligned}
|\phi_{i,j}^{(l+1)}(z_j)| &\leq \|\mathbf{W}^{(l+1)}\|_\infty + \|\mathbf{W}^{(l+1)}\|_\infty \cdot \sup_{\mathbf{x}} |z_j| \\
&\leq 2 \cdot (|a_0| \vee |b_0|) \cdot \prod_{k=0}^{l} D_k \cdot \left[ \|\mathbf{W}^{(l+1)}\|_\infty + \|\mathbf{W}^{(l+1)}\|_\infty \cdot \Big(\sum_{h=0}^{l} \prod_{k=l-h}^{l} \|\mathbf{W}^{(k)}\|_\infty\Big) \right], \\
&\leq 2 \cdot (|a_0| \vee |b_0|) \cdot \prod_{k=0}^{l} D_k \cdot \Big(\sum_{h=0}^{l+1} \prod_{k=l+1-h}^{l+1} \|\mathbf{W}^{(k)}\|_\infty\Big)
\end{aligned}$$

for $j = 1, \ldots, D_{l+1}$. In the second inequality, we used $2 \cdot (|a_0| \vee |b_0|) \cdot \prod_{k=0}^{l} D_k \geq 1$. Finally, since $|[\Phi_{l+1} \circ \cdots \circ \Phi_0(\mathbf{x})]_i| \leq D_{l+1} \cdot \max_j |\phi_{i,j}^{(l+1)}(z_j)|$, substituting the above bound and rearranging completes the proof. □

**Lemma B.12.** *Let $L, m \in \mathbb{N}$, $\mathbf{D} \in \mathbb{N}^{L+1}$, $\mathbf{G} \in \mathbb{N}^L$, and $H > 0$. Let $f$, $f^* \in KAN(L, \mathbf{D}, \mathbf{G}, H; m)$ be denoted by $f = \Phi_{L-1} \circ \cdots \circ \Phi_0$ and $f^* = \Phi^*_{L-1} \circ \cdots \circ \Phi^*_0$, and denote each edge by $(\phi^{(l)}_{i,j})$ and $(\phi^{*(l)}_{i,j})$. Their respective parameters are denoted by $\boldsymbol{\theta} := (\mathbf{W}^{(l)}) = (\theta^{(l)}_{i,j,k})$ and $\boldsymbol{\theta}^* := (\mathbf{W}^{*(l)}) = (\theta^{*(l)}_{i,j,k})$. Further assume that $|a_0| \vee |b_0| \geq 1$. Then the following holds:*

*Given $\epsilon > 0$, if $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty < \epsilon$, then the following holds. For each $l = 1, \ldots, L-1$, there exists a constant $C^*_l > 0$ such that*

$$\|\Phi_l \circ (\Phi^*_{l-1} \circ \ldots \Phi^*_0)(\mathbf{x}) - \Phi^*_l \circ (\Phi^*_{l-1} \circ \ldots \Phi^*_0)(\mathbf{x})\|_\infty \leq \epsilon \cdot C^*_l \quad \textit{for all } \mathbf{x} \in [a_0, b_0]^{D_0}$$

*Here,*

$$C^*_l := 2 \cdot (|a_0| \vee |b_0|) \cdot \prod_{k=0}^{l} D_k \cdot \left(1 + \sum_{h=0}^{l-1} \prod_{k=l-1-h}^{l-1} \|\mathbf{W}^{*(k)}\|_\infty\right)$$

*Moreover, if we define the case $l = 0$, then*

$$\|\Phi_0(\mathbf{x}) - \Phi^*_0(\mathbf{x})\|_\infty \leq \epsilon \cdot C^*_0 \quad \textit{for all } \mathbf{x} \in [a_0, b_0]^{D_0}$$

*where the constant is*

$$C^*_0 := 2 \cdot (|a_0| \vee |b_0|) \cdot D_0$$

*Proof.* Let $\epsilon > 0$ be given, and suppose that $\|\boldsymbol{\theta} - \boldsymbol{\theta^*}\|_\infty < \epsilon$. Also fix $\mathbf{x} = (x_1, \ldots, x_{D_0}) \in [a_0, b_0]^{D_0}$. First, we handle the case $l = 0$ separately. By the property Eq. (12) and the bound $|\operatorname{SiLU}(x)| \leq |x| \leq |a_0| \vee |b_0|, \forall x \in [a_0, b_0]$, for $i = 1, \ldots, D_1$,

$$\begin{aligned} |\phi^{(0)}_{i,j}(x_j) - \phi^{*(0)}_{i,j}(x_j)| &\leq \|\mathbf{W}^{(0)} - \mathbf{W}^{*(0)}\|_\infty (1 + |x_j|) \\ &\leq \epsilon \cdot 2 \cdot (|a_0| \vee |b_0|) \end{aligned} \tag{38}$$

for every $j = 1, \ldots, D_0$. Therefore, substituting this bound into $\|\Phi_0(\mathbf{x}) - \Phi^*_0(\mathbf{x})\|_\infty \leq D_0 \max_{i,j} |\phi^{(0)}_{i,j}(x_j) - \phi^{*(0)}_{i,j}(x_j)|$ completes the proof for this case.

Now let $l = 1, \ldots, L-1$ be arbitrary, and we prove the stated conclusion. Denote the output of the previous layer $\Phi^*_{l-1} : \mathbb{R}^{D_{l-1}} \to \mathbb{R}^{D_l}$ by $\mathbf{z}^* = (z^*_1, \ldots, z^*_{D_l}) \in \mathbb{R}^{D_l}$ s.t. $z^*_j := [\Phi^*_{l-1} \circ \cdots \circ \Phi^*_0(\mathbf{x})]_j$, $j = 1, \ldots, D_l$. Then, by an argument similar to Eq. (38), for $i = 1, \ldots, D_{l+1}$ and $j = 1, \ldots, D_l$,

$$\begin{aligned} |(\phi^{(l)}_{i,j} - \phi^{*(l)}_{i,j})(z^*_j)| &\leq \|\mathbf{W}^{(l)} - \mathbf{W}^{*(l)}\|_\infty \cdot (1 + \sup_{\mathbf{x}} |z^*_j|) \\ &\leq \epsilon \cdot (1 + \sup_{\mathbf{x}} |z^*_j|) \end{aligned}$$

Applying Lemma B.11, we have

$$|z^*_j| \leq 2(|a_0| \vee |b_0|) \cdot \prod_{k=0}^{l-1} D_k \cdot \left(\sum_{h=0}^{l-1} \prod_{k=l-1-h}^{l-1} \|\mathbf{W}^{*(k)}\|_\infty\right)$$

Combining this with $1 \leq 2(|a_0| \vee |b_0|) \cdot \prod_{k=0}^{l-1} D_k$ gives

$$\|\Phi_l(\mathbf{z}^*) - \Phi^*_l(\mathbf{z}^*)\|_\infty \leq \epsilon \cdot 2(|a_0| \vee |b_0|) \cdot \prod_{k=0}^{l} D_k \cdot \left(1 + \sum_{h=0}^{l-1} \prod_{k=l-1-h}^{l-1} \|\mathbf{W}^{*(k)}\|_\infty\right)$$

which completes the proof. □

### B.4.2 Parameter-to-function Lipschitz inequality

**Lemma B.13.** *Let $m \in \mathbb{N}$ with $m \geq 2$ be given, and let $L \in \mathbb{N}$, $\mathbf{D} \in \mathbb{N}^{L+1}$, $\mathbf{G} \in \mathbb{N}^L$, $H > 0$, and $B > 0$. For given functions $f, f^* \in KAN(L, \mathbf{D}, \mathbf{G}, H; m)$, write $f = \Phi_{L-1} \circ \cdots \circ \Phi_0$ and $f^* = \Phi^*_{L-1} \circ \cdots \circ \Phi^*_0$. Suppose also that their respective parameters $\boldsymbol{\theta} = (\mathbf{W}^{(l)})_{l=0}^{L-1}$ and $\boldsymbol{\theta}^* = (\mathbf{W}^{*(l)})_{l=0}^{L-1}$ satisfy*

$$\|\boldsymbol{\theta}\|_\infty \leq B, \ \|\boldsymbol{\theta}^*\|_\infty \leq B.$$

*Then the following holds:*

*For any $\epsilon > 0$, if $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty < \epsilon$, then there exists a constant $K(L, \mathbf{D}, \mathbf{G}, H, B; a_0, b_0, m)$ such that*

$$\|f - f^*\|_{L^\infty} \leq \epsilon K(L, \mathbf{D}, \mathbf{G}, H, B; a_0, b_0, m)$$

*Here,*

$$\begin{aligned} &K(L, \mathbf{D}, \mathbf{G}, H, B\,;\, a_0, b_0, m) \\ &:= 2(|a_0| \vee |b_0|) L m^{L-1} \left[\prod_{l=1}^{L-1} \left(\frac{G_l + 2m}{H} + 1\right)\right] \cdot \left(\prod_{l=0}^{L-1} D_l\right) \cdot \left(\sum_{l=0}^{L-1} B^l\right) \end{aligned}$$

*Proof.* For the given $\epsilon > 0$, assume that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty < \epsilon$. For the network representation $f = \Phi_{L-1} \circ \cdots \circ \Phi_0$, define the following partial composition notation. For each $l = 1, \ldots, L-1$, let

$$\begin{aligned} T_l^+(f)(\cdot) &:= \Phi_{l-1} \circ \cdots \circ \Phi_0(\cdot), \\ T_l^-(f)(\cdot) &:= \Phi_{L-1} \circ \cdots \circ \Phi_l(\cdot). \end{aligned}$$

Also define $T_0^+(f)(\mathbf{x}) = \mathbf{x}$ when $l = 0$, and define $T_L^-(f)(\mathbf{x}) = \mathbf{x}$ when $l = L$. Now fix $\mathbf{x} \in [a_0, b_0]^{D_0}$. Then, by a telescoping argument that replaces one layer at a time and by the triangle inequality,

$$|f(\mathbf{x}) - f^*(\mathbf{x})| \leq \sum_{l=0}^{L-1} \left|T_{l+1}^-(f) \circ \Phi_l \circ T_l^+(f^*)(\mathbf{x}) - T_{l+1}^-(f) \circ \Phi_l^* \circ T_l^+(f^*)(\mathbf{x})\right| \tag{39}$$

Applying Lemma B.10 to the right-hand side of Eq. (39), we obtain

$$\begin{aligned} |f(\mathbf{x}) - f^*(\mathbf{x})| \leq \sum_{l=0}^{L-2} C_{l+1} &\left\|\Phi_l \circ T_l^+(f^*)(\mathbf{x}) - \Phi_l^* \circ T_l^+(f^*)(\mathbf{x})\right\|_\infty \\ &+ \left|\Phi_{L-1} \circ T_{L-1}^+(f^*)(\mathbf{x}) - \Phi_{L-1}^* \circ T_{L-1}^+(f^*)(\mathbf{x})\right|, \end{aligned} \tag{40}$$

where, for $l' = 1, \ldots, L-1$,

$$C_{l'} := \prod_{k=l'}^{L-1} \left(\frac{2m}{\Delta^{(k)}} + m\right) \|\mathbf{W}^{(k)}\|_\infty D_k.$$

Next, applying Lemma B.12 to Eq. (40), we obtain the following upper bound:

$$|f(\mathbf{x}) - f^*(\mathbf{x})| \leq \left(\sum_{l=1}^{L-1} C_l C_{l-1}^* + C_{L-1}^*\right) \epsilon. \tag{41}$$

Here,

$$C_0^* := 2(|a_0| \vee |b_0|) D_0$$

and, for $l' = 1, \ldots, L-1$,

$$C_{l'}^* := 2(|a_0| \vee |b_0|) \prod_{k=0}^{l'} D_k \left(1 + \sum_{k'=0}^{l'-1} \prod_{k=l'-1-k'}^{l'-1} \|\mathbf{W}^{*(k)}\|_\infty\right).$$

Since

$$\Delta^{(k)} = \frac{b_k - a_k}{G_k + 2m} = \frac{2H}{G_k + 2m}, \qquad \|\mathbf{W}^{(k)}\|_\infty \leq B, \qquad \|\mathbf{W}^{*(k)}\|_\infty \leq B$$

we have $2m/\Delta^{(k)} + m = m\{(G_k + 2m)/H + 1\}$, and hence, for $l = 1, \ldots, L-1$,

$$\begin{aligned} C_l &\leq \prod_{k=l}^{L-1} \left\{m\left(\frac{G_k + 2m}{H} + 1\right)\right\} B D_k \\ &\leq m^{L-l} B^{L-l} \left[\prod_{k=l}^{L-1} \left(\frac{G_k + 2m}{H} + 1\right)\right] \left(\prod_{k=l}^{L-1} D_k\right). \end{aligned}$$

Moreover,

$$C_{l-1}^* \le 2(|a_0| \vee |b_0|) \left(\prod_{k=0}^{l-1} D_k\right) \left(\sum_{r=0}^{l-1} B^r\right).$$

Therefore,

$$\begin{aligned} C_l C_{l-1}^* &\le 2(|a_0| \vee |b_0|) m^{L-l} \left[\prod_{k=l}^{L-1} \left(\frac{G_k + 2m}{H} + 1\right)\right] \left(\prod_{k=0}^{L-1} D_k\right) B^{L-l} \left(\sum_{r=0}^{l-1} B^r\right) \\ &\le 2(|a_0| \vee |b_0|) m^{L-1} \left[\prod_{k=1}^{L-1} \left(\frac{G_k + 2m}{H} + 1\right)\right] \left(\prod_{k=0}^{L-1} D_k\right) \left(\sum_{r=0}^{L-1} B^r\right), \end{aligned}$$

where the last inequality uses $m \ge 1$ and

$$B^{L-l} \sum_{r=0}^{l-1} B^r = \sum_{r=L-l}^{L-1} B^r \le \sum_{r=0}^{L-1} B^r$$

The term corresponding to the last layer can be bounded in the same way as

$$C_{L-1}^* \le 2(|a_0| \vee |b_0|) \left(\prod_{k=0}^{L-1} D_k\right) \left(\sum_{r=0}^{L-1} B^r\right)$$

and hence it is also bounded by the same common constant above. Therefore, the right-hand side of Eq. (41) is bounded by

$$\epsilon \cdot 2(|a_0| \vee |b_0|) L m^{L-1} \left[\prod_{l=1}^{L-1} \left(\frac{G_l + 2m}{H} + 1\right)\right] \left(\prod_{l=0}^{L-1} D_l\right) \left(\sum_{l=0}^{L-1} B^l\right)$$

Since $\mathbf{x} \in [a_0, b_0]^{D_0}$ was arbitrary,

$$\|f - f^*\|_{L^\infty} \le \epsilon K(L, \mathbf{D}, \mathbf{G}, H, B\,;\, a_0, b_0, m)$$

This completes the proof. □

**Corollary B.14.** *Let $L, D, G, m \in \mathbb{N}$, $H > 0$ with $m \ge 2$, and $B \ge 1$. For $f_{\boldsymbol{\theta}}$, $f_{\boldsymbol{\theta}^*} \in KAN(L, D, G, H; m)$, assume that*

$$\|\boldsymbol{\theta}\|_\infty \vee \|\boldsymbol{\theta}^*\|_\infty \le B$$

*For any $\epsilon > 0$, if $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty < \epsilon$, then there exists $K(L, D, G, H, B; a_0, b_0, d, m)$ such that*

$$\|f_{\boldsymbol{\theta}} - f_{\boldsymbol{\theta}^*}\|_{L^\infty} \le \epsilon K(L, D, G, H, B; a_0, b_0, d, m)$$

*Here,*

$$K(L, D, G, H, B; a_0, b_0, d, m) = 2(|a_0| \vee |b_0|) L^2 m^{L-1} d \left(\frac{G + 2m}{H} + 1\right)^{L-1} D^{L-1} B^L.$$

*Proof.* The result follows directly by substituting $\mathbf{D} = (d, D, \ldots, D, 1)$ and $\mathbf{G} = (G_0, G \ldots, G)$ into $K(L, \mathbf{D}, \mathbf{G}, H, B; a_0, b_0, d, m)$ in Lemma B.13. □

### B.4.3 Covering number bound for sparse KANs

**Lemma B.15** (Covering number bound)**.** *Let $L, D, G, m \in \mathbb{N}$ with $m \ge 2$, $H > 0$, $B \ge 1$, and $1 \le S \le T(L, D, G)$, where $T(L, D, G) = dD(G_0 + m + 1) + ((L-2)D^2 + D)(G + m + 1)$. Then, for any $\epsilon > 0$, the following holds:*

$$\begin{aligned} &\log \mathcal{N}\Big(\epsilon, \Big\{f_{\boldsymbol{\theta}} \in KAN(L, D, G, H; m) : \|\boldsymbol{\theta}\|_\infty \le B,\ \|\boldsymbol{\theta}\|_0 \le S\Big\}, \|\cdot\|_{L^\infty}\Big) \\ &\lesssim\ S \log \frac{eT(L, D, G)}{S} + S \log\left(1 + \frac{2BK(L, D, G, H, B; a_0, b_0, d, m)}{\epsilon}\right) \end{aligned} \tag{42}$$

*where*

$$K(L, D, G, H, B; a_0, b_0, d, m) = 2(|a_0| \vee |b_0|) L^2 m^{L-1} d \left(\frac{G + 2m}{H} + 1\right)^{L-1} D^{L-1} B^L$$

*Proof.* Write $T := T(L, D, G)$, and define the parameter set as

$$\Theta(B, S) := \{\boldsymbol{\theta} \in \mathbb{R}^T : \|\boldsymbol{\theta}\|_\infty \le B,\ \|\boldsymbol{\theta}\|_0 \le S\}.$$

Write $\delta := \epsilon / K(L, D, G, H, B; a_0, b_0, d, m)$. By Corollary B.14, every $\delta$-cover of $\Theta(B, S)$ in $\|\cdot\|_\infty$ induces an $\epsilon$-cover of the KAN class in Eq. (42). Therefore, the left-hand side of Eq. (42) is bounded above by

$$\log \mathcal{N}\left(\delta, \Theta(B, S), \|\cdot\|_\infty\right).$$

We now decompose $\Theta(B, S)$ according to the support set $\mathrm{supp}(\boldsymbol{\theta}) = \{i : \theta_i \neq 0\}$. For $I \subset \{1, \dots, T\}$, define

$$\Theta_I(B) := \{\boldsymbol{\theta} \in \mathbb{R}^T : \mathrm{supp}(\boldsymbol{\theta}) \subset I,\ \|\boldsymbol{\theta}\|_\infty \le B\}.$$

Then

$$\Theta(B, S) = \bigcup_{s=0}^{S} \bigcup_{\substack{I \subset \{1,\dots,T\} \\ |I| = s}} \Theta_I(B).$$

For a fixed $I$ s.t. $|I| = s$, covering the set $\Theta_I(B)$ under the $\ell_\infty$ metric is equivalent to covering $[-B, B]^s$ under the $\ell_\infty$ metric. Hence,

$$\mathcal{N}\big(\delta, \Theta_I(B), \|\cdot\|_\infty\big) \le \left\lceil \frac{2B}{\delta} \right\rceil^s.$$

Moreover, since there are $\binom{T}{s}$ possible support sets of size $s = |I|$, we have

$$\begin{aligned}
\mathcal{N}\left(\delta, \Theta(B, S), \|\cdot\|_\infty\right) &\le \sum_{s=1}^{S} \binom{T}{s} \left\lceil \frac{2B}{\delta} \right\rceil^s \\
&\le \left(1 + \frac{2B}{\delta}\right)^S \sum_{s=0}^{S} \binom{T}{s} \\
&\le 1 + \left(1 + \frac{2B}{\delta}\right)^S \left(\frac{eT}{S}\right)^S.
\end{aligned}$$

In the last inequality, we used $\sum_{s=1}^{S} \binom{T}{s} \le \left(\frac{eT}{S}\right)^S$. Finally, substituting back $\delta = \epsilon / K(L, D, G, H, B; a_0, b_0, d, m)$ and taking logarithms completes the proof. □

# C Proofs on posterior concentration

In this section, we present the proofs of the main theorems corresponding to Section 3. Appendix C.1 is a standard lemma used in Appendices C.2 to C.5. It is a reformulation of Theorem 4 of Ghosal and van der Vaart (2007) for independent observations in the framework of a nonparametric regression model with Gaussian errors.

## C.1 Lemmas for posterior concentration

**Lemma C.1** (Sufficient conditions for posterior concentration)**.** *Let $\mathcal{F}_n \subset \mathcal{UB}$ be a sieve, and define*

$$\mathcal{M}_n := \mathcal{F}_n \times [\underline{\sigma}, \overline{\sigma}].$$

*Also let $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$. Define*

$$A_n := \left\{(f, \sigma) : \|f - f_0\|_{L^2(P_X)} \le \epsilon_n/2,\ |\sigma - \sigma_0| \le \epsilon_n/2\right\}.$$

*Suppose that, for a sufficiently large constant $C > 0$, the following three conditions hold:*

$$\log \mathcal{N}\big(\epsilon_n, \mathcal{F}_n, \|\cdot\|_{L^2(P_X)}\big) \lesssim n\epsilon_n^2, \tag{43}$$

$$-\log \Pi_n(A_n) \lesssim n\epsilon_n^2, \tag{44}$$

$$\Pi_n(\mathcal{M}_n^c) = o\big(e^{-Cn\epsilon_n^2}\big). \tag{45}$$

*Then*

$$\Pi_n\Big((f, \sigma) : \|f - f_0\|_{L^2(P_X)} + \left|\sigma^2 - \sigma_0^2\right| > M_n\epsilon_n \ \Big|\ \mathcal{D}_n\Big) \to 0$$

*holds in $P_{f_0, \sigma_0}^{(n)}$-probability for any sequence $M_n \to \infty$.*

*Proof.* Write the density of a single observation $(X, Y)$ as

$$p_{f,\sigma}(x,y) := p_X(x)\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y-f(x))^2}{2\sigma^2}\right].$$

On the bounded variance interval $[\underline{\sigma}, \overline{\sigma}]$, the Hellinger distance, the Kullback–Leibler divergence, and the second moment of the log-likelihood ratio for the Gaussian regression model are all controlled by

$$\rho\big((f,\sigma),(g,\tau)\big) := \|f-g\|_{L^2(P_X)} + |\sigma-\tau|.$$

More precisely, by the calculations in Lee et al. (2025), there exist constants $c_1, c_2, c_3 > 0$ such that, for all $(f,\sigma),(g,\tau) \in \mathcal{M}_n$,

$$h^2(p_{f,\sigma}, p_{g,\tau}) \le c_1\Big(\|f-g\|^2_{L^2(P_X)} + |\sigma-\tau|^2\Big), \tag{46}$$

$$K(p_{f_0,\sigma_0}, p_{f,\sigma}) \le c_2\Big(\|f-f_0\|^2_{L^2(P_X)} + |\sigma-\sigma_0|^2\Big), \tag{47}$$

$$V(p_{f_0,\sigma_0}, p_{f,\sigma}) \le c_3\Big(\|f-f_0\|^2_{L^2(P_X)} + |\sigma-\sigma_0|^2\Big) \tag{48}$$

hold. Here, $K$ denotes the Kullback–Leibler divergence, and $V$ denotes the second moment of the log-likelihood ratio.

By Eq. (46), combining an $L^2(P_X)$-cover of $\mathcal{F}_n$ and an absolute-value cover of the interval $[\underline{\sigma}, \overline{\sigma}]$ yields a Hellinger cover of $\mathcal{M}_n$. That is, for some constant $c > 0$,

$$\log \mathcal{N}\big(c\eta, \mathcal{M}_n, h\big) \le \log \mathcal{N}\big(\eta, \mathcal{F}_n, \|\cdot\|_{L^2(P_X)}\big) + \log \mathcal{N}\big(\eta, [\underline{\sigma}, \overline{\sigma}], |\cdot|\big).$$

The last term satisfies

$$\log \mathcal{N}\big(\eta, [\underline{\sigma}, \overline{\sigma}], |\cdot|\big) \le \log\left(1 + \frac{\overline{\sigma}-\underline{\sigma}}{\eta}\right) \lesssim \log(\eta^{-1}),$$

and since $n\epsilon_n^2 \to \infty$, this is of smaller order than Eq. (43) when $\eta = \epsilon_n$. Hence,

$$\log \mathcal{N}\big(c\epsilon_n, \mathcal{M}_n, h\big) \lesssim n\epsilon_n^2.$$

Secondly, by Eq. (47) and Eq. (48), on $A_n$ we have

$$K(p_{f_0,\sigma_0}, p_{f,\sigma}) \vee V(p_{f_0,\sigma_0}, p_{f,\sigma}) \lesssim \epsilon_n^2.$$

Therefore, $A_n$ is contained in a Kullback-Leibler neighborhood. This proves that Eq. (44) is a sufficient condition.

Condition Eq. (45) is precisely the standard prior probability condition outside the sieve.

Now applying the theorem in Ghosal and van der Vaart (2017) to $\mathcal{M}_n$, the standard posterior contraction theorem directly implies from Eq. (43)–Eq. (45) that

$$\Pi_n\big(\rho\big((f,\sigma),(f_0,\sigma_0)\big) > M_n\epsilon_n \;\big|\; \mathcal{D}_n\big) \to 0$$

in $P^{(n)}_{f_0,\sigma_0}$-probability. Finally, since $\sigma, \sigma_0 \in [\underline{\sigma}, \overline{\sigma}]$,

$$\big|\sigma^2 - \sigma_0^2\big| = |\sigma-\sigma_0||\sigma+\sigma_0| \le 2\overline{\sigma}|\sigma-\sigma_0|.$$

Applying Lemma B.1 of Xie and Xu (2020), the desired conclusion

$$\Pi_n\Big((f,\sigma) : \|f-f_0\|_{L^2(P_X)} + \big|\sigma^2-\sigma_0^2\big| > M_n\epsilon_n \;\Big|\; \mathcal{D}_n\Big) \to 0$$

follows. □

### C.2 Proof of Theorem 3.2

*Proof.* We define the sieve for applying Lemma C.1 as follows.

$$\begin{aligned}
\mathcal{F}_n^{\mathrm{spi}} &:= \big\{ f_{\boldsymbol{\theta}} \in KAN(L_0, D_n, G_n, H_n; m) : \|\boldsymbol{\theta}\|_\infty \le \overline{B}_n,\ \|\boldsymbol{\theta}\|_0 \le S_n \big\} \\
\widetilde{\mathcal{F}}_n^{\mathrm{spi}} &:= \big\{ \mathrm{clip} \circ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}} \in \mathcal{F}_n^{\mathrm{spi}} \big\}
\end{aligned} \tag{49}$$

and set $\widetilde{\mathcal{M}}_n^{\mathrm{spi}} = \widetilde{\mathcal{F}}_n^{\mathrm{spi}} \times [\underline{\sigma}, \overline{\sigma}]$. Here, we take the truncation sequence $\overline{B}_n$, which is used only for the sieve, as

$$\overline{B}_n := \overline{B}_0 N_n^{\beta+1/\alpha} (\log n)^{1/\alpha} \tag{50}$$

where $\overline{B}_0 > 0$ is a sufficiently large constant chosen so that the calculation of the sieve-complement probability closes. We now show that the three conditions Eq. (43)–Eq. (45) in Lemma C.1 are satisfied for the rate $\epsilon_n$ in Eq. (6). Note that Lemma C.1 is applied with $\mathcal{F}_n = \widetilde{\mathcal{F}}_n^{\mathrm{spi}}$ and $\mathcal{M}_n = \widetilde{\mathcal{M}}_n^{\mathrm{spi}}$.

First, we show (43). By the density condition in (A2),

$$\|g\|_{L^2(P_X)} \le \|p_X\|_{L^\infty}^{1/2} \|g\|_{L^2} \le C_X^{1/2} \|g\|_{L^\infty}$$

and since clip is 1-Lipschitz, an $L^\infty$-cover of $\mathcal{F}_n^{\mathrm{spi}}$ induces an $L^\infty$-cover, hence an $L^2(P_X)$-cover, of $\widetilde{\mathcal{F}}_n^{\mathrm{spi}}$. Therefore,

$$\log \mathcal{N}\big(\epsilon_n, \widetilde{\mathcal{F}}_n^{\mathrm{spi}}, \|\cdot\|_{L^2(P_X)}\big) \le \log \mathcal{N}\big(C_X^{-1/2}\epsilon_n, \mathcal{F}_n^{\mathrm{spi}}, \|\cdot\|_{L^\infty}\big).$$

Applying Lemma B.15 to the sieve (49), we obtain

$$\log \mathcal{N}\big(C_X^{-1/2}\epsilon_n, \mathcal{F}_n^{\mathrm{spi}}, \|\cdot\|_{L^\infty}\big) \lesssim S_n \log \frac{eT_n}{S_n} + S_n \log\left(1 + \frac{\overline{B}_n K_n}{C_X^{-1/2}\epsilon_n}\right), \tag{51}$$

where $T_n := T(L_0, D_n, G_n)$ and $K_n := K(L_0, D_n, G_n, H_n, \overline{B}_n; m)$. Substituting the values of the relevant sequences gives

$$\begin{aligned} T_n &\lesssim N_n^{2+\beta} \\ K_n &= 2(|a_0| \vee |b_0|) L_0^2 m^{L_0-1} d \left(\frac{G_n + 2m}{H_n} + 1\right)^{L_0-1} D_n^{L_0-1} \overline{B}_n^{L_0} \\ &\lesssim N_n^{L_0(\beta+1)-1+L_0/\alpha} (\log n)^{L_0/\alpha} \end{aligned} \tag{52}$$

Moreover, using $S_n = \lceil S_0 N_n \rceil$ and $N_n = \lceil n^{1/(2\tilde{\mathbf{s}}+1)} \rceil$ in Eq. (51), we see that $\log \frac{eT_n}{S_n} \lesssim \log n$ and $\log\left(1 + \frac{\overline{B}_n K_n}{\epsilon_n}\right) \lesssim \log n$. Therefore,

$$S_n \log \frac{eT_n}{S_n} + S_n \log\left(1 + \frac{\overline{B}_n K_n}{\epsilon_n}\right) \lesssim N_n \log n \asymp n\epsilon_n^2 \tag{53}$$

holds.

Secondly, we prove (44). Applying Lemma B.4 with $N = N_n$, there exists $f_n^\star \in KAN_c(L_0, D_n, G_n, H_n; m)$ with some parameter $\boldsymbol{\theta}_n^\star$ satisfying $\|\boldsymbol{\theta}_n^\star\|_\infty \le B_n^\star$ and $\|\boldsymbol{\theta}_n^\star\|_0 \le S_n$ such that

$$\|f_0 - f_n^\star\|_{L^2} \lesssim N_n^{-\tilde{\mathbf{s}}} \tag{54}$$

holds. Here, $B_n^\star := B_0 N_n^\beta$ for some $B_0 > 0$. We then define the oracle model space by

$$\mathcal{F}_n^\star := \{f_{\boldsymbol{\theta}} \in KAN(L_0, D_n, G_n, H_n; m) : \|\boldsymbol{\theta}\|_\infty \le B_n^\star, \|\boldsymbol{\theta}\|_0 \le S_n\}. \tag{55}$$

Since

$$\frac{\overline{B}_n}{B_n^\star} = \frac{\overline{B}_0}{B_0} N_n^{1/\alpha} (\log n)^{1/\alpha} \to \infty, \tag{56}$$

we have $B_n^\star \le \overline{B}_n$ for all sufficiently large $n$. Thus, $f_n^\star \in \mathcal{F}_n^\star \subset \mathcal{F}_n^{\mathrm{spi}}$ for sufficiently large $n$. In addition, $N_n^{-\tilde{\mathbf{s}}} \le n^{-\tilde{\mathbf{s}}/(2\tilde{\mathbf{s}}+1)} = \epsilon_n (\log n)^{-1/2}$ for sufficiently large $n$. Hence, for sufficiently large $n$,

$$\|f_0 - f_n^\star\|_{L^2} \le \frac{\epsilon_n}{4C_X^{1/2}} \tag{57}$$

holds, and by applying (A2), we obtain

$$\|f_0 - f_n^\star\|_{L^2(P_X)} \le \|p_X\|_{L^\infty}^{1/2} \|f_0 - f_n^\star\|_{L^2} \le \epsilon_n/4. \tag{58}$$

Since $f_0 \in \mathcal{UB}$, we have $\mathrm{clip} \circ f_0 = f_0$. Moreover, the clipping map clip is 1-Lipschitz. Therefore, for any $f \in \mathcal{F}_n^\star$,

$$\begin{aligned} \|\,\mathrm{clip}\circ f - f_0\|_{L^2(P_X)} &= \|\,\mathrm{clip}\circ f - \mathrm{clip}\circ f_0\|_{L^2(P_X)} \\ &\le \|\,\mathrm{clip}\circ f - \mathrm{clip}\circ f_n^\star\|_{L^2(P_X)} + \|\,\mathrm{clip}\circ f_n^\star - \mathrm{clip}\circ f_0\|_{L^2(P_X)} \\ &\le \|f - f_n^\star\|_{L^2(P_X)} + \|f_n^\star - f_0\|_{L^2(P_X)} \\ &\le C_X^{1/2}\|f - f_n^\star\|_{L^2} + \epsilon_n/4. \end{aligned} \tag{59}$$

Combining the preceding bounds, for the prior mass $\Pi_n(A_n)$ in Eq. (44), we have

$$\begin{aligned} \Pi_n(A_n) &\ge \Pi_n\left(f \in \mathcal{F}_n^{\mathrm{spi}} : \|f - f_n^\star\|_{L^2} \le \frac{\epsilon_n}{4C_X^{1/2}}\right)\Pi\left(\sigma : |\sigma - \sigma_0| \le \frac{\epsilon_n}{2}\right) \\ &\ge \Pi_n\left(f \in \mathcal{F}_n^\star : \|f - f_n^\star\|_{L^\infty} \le \frac{\epsilon_n}{4C_X^{1/2}}\right)\Pi\left(\sigma : |\sigma - \sigma_0| \le \frac{\epsilon_n}{2}\right). \end{aligned} \tag{60}$$

In the second inequality, we used the fact that the $L^\infty$ norm dominates the $L^2$ norm. Now, for the parameter vector $\boldsymbol{\theta}_n^\star$ realizing $f_n^\star$, let

$$I_n^\star := \{t : \theta_{n,t}^\star \neq 0\}, \qquad S_n^\star := |I_n^\star|. \tag{61}$$

Then $S_n^\star \le S_n$. Since the spike-and-slab prior selects exactly $S_n$ nonzero coordinates, choose any set $J_n^\star \subset \{1, \dots, T_n\}$ such that $I_n^\star \subset J_n^\star$ and $|J_n^\star| = S_n$ and denote a vector $\mathbf{1}_{J_n^\star} \in \{0,1\}^{T_n}$ as

$$(\mathbf{1}_{J_n^\star})_t = \begin{cases} 1, & t \in J_n^\star \\ 0, & t \notin J_n^\star. \end{cases} \tag{62}$$

That is, we enlarge the parameter support of the actual approximator to an $S_n$-sized support. We now apply Corollary B.14 to the first term in the last inequality of (60). For $f_{\boldsymbol{\theta}}, f_n^\star \in \mathcal{F}_n^\star$, if $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n^\star\|_\infty \le \delta_n$, then

$$\|f_{\boldsymbol{\theta}} - f_n^\star\|_{L^\infty} \le K_n^\star\|\boldsymbol{\theta} - \boldsymbol{\theta}_n^\star\|_\infty \le \frac{\epsilon_n}{4C_X^{1/2}}. \tag{63}$$

where $K_n^\star := K(L_0, D_n, G_n, H_n, B_n^\star; a_0, b_0, d, m)$. Recall that $K_n^\star \asymp D_n^{L_0-1}(B_n^\star)^{L_0} \asymp N_n^{L_0-1}(B_n^\star)^{L_0} = N_n^{(1+\beta)L_0-1}$. Define

$$\delta_n := \frac{\epsilon_n}{4C_X^{1/2}K_n^\star}. \tag{64}$$

Then the lower bound in Eq. (60) is

$$\Pi_n\Big(\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_n^\star\|_\infty \le \delta_n \mid \boldsymbol{\gamma} = \mathbf{1}_{J_n^\star}\Big)\pi_n(\boldsymbol{\gamma} = \mathbf{1}_{J_n^\star})\Pi(\sigma : |\sigma - \sigma_0| \le \epsilon_n/2). \tag{65}$$

First, since $\pi_n(\boldsymbol{\gamma} = \mathbf{1}_{J_n^\star}) = \binom{T_n}{S_n}^{-1}$,

$$-\log \pi_n(\boldsymbol{\gamma} = \mathbf{1}_{J_n^\star}) \le S_n \log \frac{eT_n}{S_n} \lesssim N_n \log n$$

holds. Next, since $\|\boldsymbol{\theta}_n^\star\|_\infty \le B_n^\star$ and $\delta_n \le 1$ for all sufficiently large $n$, combining this with (B1) yields

$$\begin{aligned} \Pi_n\Big(\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_n^\star\|_\infty \le \delta_n \mid \boldsymbol{\gamma} = \mathbf{1}_{J_n^\star}\Big) &\ge \prod_{t \in J_n^\star}\int_{\theta_{n,t}^\star - \delta_n}^{\theta_{n,t}^\star + \delta_n} \widetilde{\pi}_{SL}(u)\,du \\ &\ge \left(2\delta_n \inf_{|u| \le B_n^\star + 1}\widetilde{\pi}_{SL}(u)\right)^{S_n} \\ &\ge (2\delta_n n^{-c_1})^{S_n}. \end{aligned}$$

Therefore, it remains to bound the right-hand side of

$$-\log \Pi_n\Big(\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_n^\star\|_\infty \le \delta_n \mid \boldsymbol{\gamma} = \mathbf{1}_{J_n^\star}\Big) \le S_n \log \delta_n^{-1} + c_1 S_n \log n \tag{66}$$

by $n\epsilon_n^2 = n^{1/(2\tilde{\mathbf{s}}+1)} \log n$. The second term is immediate, so we treat the first term.

$$\delta_n^{-1} \asymp K_n^\star \epsilon_n^{-1} = N_n^{(1+\beta)L_0 - 1 + \tilde{\mathbf{s}}} (\log n)^{-1/2}$$

Thus, $\log \delta_n^{-1} \lesssim \log n$, which completes this part. Finally, for the variance term in Eq. (65), by (B3) and $\sigma_0 \ge \underline{\sigma} > 0$, for all sufficiently large $n$,

$$\Pi_n(|\sigma - \sigma_0| \le \epsilon_n/2) \ge c_\sigma \Big[ \big(\sigma_0 + \epsilon_n/2\big)^2 - \big(\sigma_0 - \epsilon_n/2\big)^2 \Big] = 2c_\sigma \sigma_0 \epsilon_n, \tag{67}$$

and hence

$$-\log \Pi_n(|\sigma - \sigma_0| \le \epsilon_n/2) \lesssim \log(\epsilon_n^{-1}) \lesssim \log n. \tag{68}$$

This verifies (44).

Finally, we prove (45). By (B3), the prior distribution on the variance is supported on $[\underline{\sigma}^2, \overline{\sigma}^2]$. Since the sufficient lemma is applied with $\mathcal{F}_n = \widetilde{\mathcal{F}}_n^{\mathrm{spi}}$, where

$$\widetilde{\mathcal{F}}_n^{\mathrm{spi}} := \left\{ \mathrm{clip} \circ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}} \in \mathcal{F}_n^{\mathrm{spi}} \right\},$$

we have

$$\Pi_n\big((\widetilde{\mathcal{M}}_n^{\mathrm{spi}})^c\big) = \Pi_n\big((\widetilde{\mathcal{F}}_n^{\mathrm{spi}})^c\big) \le \Pi_n\big((\mathcal{F}_n^{\mathrm{spi}})^c\big). \tag{69}$$

Moreover, since the sparsity index in (5) always has exactly $S_n$ active coordinates,

$$\Pi_n\big(\|\boldsymbol{\theta}\|_0 > S_n\big) = 0.$$

Therefore,

$$\Pi_n\big((\mathcal{F}_n^{\mathrm{spi}})^c\big) = \Pi_n\big(\|\boldsymbol{\theta}\|_\infty > \overline{B}_n\big).$$

Conditionally on $\boldsymbol{\gamma}$, the number of nonzero coordinates is $S_n$, and all zero coordinates are equal to $0$. Thus, by the union bound and (B2),

$$\begin{aligned}
\Pi_n\big(\|\boldsymbol{\theta}\|_\infty > \overline{B}_n\big) &= \sum_{\substack{\boldsymbol{\gamma} \in \{0,1\}^{T_n} \\ \|\boldsymbol{\gamma}\|_0 = S_n}} \pi_n(\boldsymbol{\gamma}) \Pi_n\big(\|\boldsymbol{\theta}\|_\infty > \overline{B}_n \mid \boldsymbol{\gamma}\big) \\
&\le \sum_{\substack{\boldsymbol{\gamma} \in \{0,1\}^{T_n} \\ \|\boldsymbol{\gamma}\|_0 = S_n}} \pi_n(\boldsymbol{\gamma}) \sum_{t:\gamma_t = 1} \Pi_n\big(|\theta_t| > \overline{B}_n \mid \gamma_t = 1\big) \\
&\le S_n \int_{|u| > \overline{B}_n} \widetilde{\pi}_{SL}(u)\, du \\
&\le c_2 S_n \exp\left[ -c_3 \left( \frac{\overline{B}_n}{\tau_n} \right)^\alpha \right].
\end{aligned} \tag{70}$$

By (B2), we have $\tau_n \lesssim B_n^\star$, and by the definition of the truncation sequence,

$$B_n^\star = B_0 N_n^\beta, \qquad \overline{B}_n = \overline{B}_0 N_n^{\beta + 1/\alpha} (\log n)^{1/\alpha}.$$

Therefore, there exists a constant $C_\tau > 0$ such that

$$\left( \frac{\overline{B}_n}{\tau_n} \right)^\alpha \ge C_\tau \overline{B}_0^\alpha N_n \log n.$$

Substituting this into Eq. (69) and Eq. (70), we obtain

$$\Pi_n\big((\widetilde{\mathcal{M}}_n^{\mathrm{spi}})^c\big) \le c_2 S_n \exp\Big[ -c_3 C_\tau \overline{B}_0^\alpha N_n \log n \Big].$$

On the other hand, since $S_n = \lceil S_0 N_n \rceil$, we have $\log S_n \lesssim \log n$. Hence,

$$\log \Pi_n\big((\widetilde{\mathcal{M}}_n^{\mathrm{spi}})^c\big) \le \log c_2 + \log S_n - c_3 C_\tau \overline{B}_0^\alpha N_n \log n.$$

Moreover,

$$n\epsilon_n^2 = n^{1/(2\tilde{\mathbf{s}}+1)} \log n \asymp N_n \log n.$$

Therefore, by choosing $\overline{B}_0 > 0$ sufficiently large, for the sufficiently large constant $C > 0$ appearing in Lemma C.1, we have

$$\Pi_n\big((\widetilde{\mathcal{M}}_n^{\mathrm{spi}})^c\big) = o\left( e^{-Cn\epsilon_n^2} \right).$$

This proves Eq. (45).

□

### C.3 Proof of Theorem 3.6

*Proof.* We define a sieve as in the proof of Appendix C.2. The proof follows almost the same structure. Let the truncation sequence $\overline{B}_n$ be defined as in Eq. (50), and let the sieve be as in Eq. (49), except that the sparsity level $S_n$ is adjusted.

$$\begin{aligned}
\mathcal{F}_n^{\mathrm{ber}} &:= \left\{ f_{\boldsymbol{\theta}} \in KAN(L_0, D_n, G_n, H_n; m) : \|\boldsymbol{\theta}\|_\infty \le \overline{B}_n,\ \|\boldsymbol{\theta}\|_0 \le \overline{S}_n \right\} \\
\widetilde{\mathcal{F}}_n^{\mathrm{ber}} &:= \left\{ \mathrm{clip} \circ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}} \in \mathcal{F}_n^{\mathrm{ber}} \right\} \\
\widetilde{\mathcal{M}}_n^{\mathrm{ber}} &:= \widetilde{\mathcal{F}}_n^{\mathrm{ber}} \times [\underline{\sigma}, \overline{\sigma}]
\end{aligned}$$

where $\overline{S}_n := \overline{S}_0 S_n$ for some sufficiently large $\overline{S}_0 > 0$.

We first show that (43) holds. As in the proof of Appendix C.2, by (A2) and Lemma B.15, we have

$$\log \mathcal{N}\big(C_X^{-1/2}\epsilon_n, \mathcal{F}_n^{\mathrm{ber}}, \|\cdot\|_{L^\infty}\big) \lesssim \overline{S}_n \log \frac{eT_n}{\overline{S}_n} + \overline{S}_n \log\left(1 + \frac{\overline{B}_n K_n}{C_X^{-1/2}\epsilon_n}\right). \tag{71}$$

The remaining argument is identical to the steps in Eq. (52)–Eq. (53), and the right-hand side can be controlled by $n\epsilon_n^2$.

Next, we prove (44). Following the same steps as in Eq. (54)–Eq. (64) in the proof of the prior mass condition in Appendix C.2, we start from Eq. (65).

$$\begin{aligned}
&\Pi_n(A_n) \\
&\ge \Pi_n\Big(\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_n^\star\|_\infty \le \delta_n \mid \boldsymbol{\gamma} = \mathbf{1}_{J_n^\star}\Big)\pi_n(\boldsymbol{\gamma} = \mathbf{1}_{J_n^\star})\Pi(\sigma : |\sigma - \sigma_0| \le \epsilon_n/2).
\end{aligned} \tag{72}$$

The first probability on the right-hand side is handled in the same way as in Eq. (66). Next, for the second probability,

$$\pi_n(\boldsymbol{\gamma} = \mathbf{1}_{J_n^\star}) = \rho_n^{S_n}(1-\rho_n)^{T_n - S_n}$$

and its negative logarithm is

$$S_n \log \frac{1}{\rho_n} + (T_n - S_n)\log\frac{1}{1-\rho_n}.$$

By the first condition in Eq. (9), there exists $c_{\rho,1} > 0$ such that $\rho_n \ge c_{\rho,1} T_n^{-1}$ for all sufficiently large $n$. Therefore,

$$S_n \log \frac{1}{\rho_n} \le S_n \log \frac{T_n}{c_{\rho,1}} \lesssim S_n \log n \lesssim n\epsilon_n^2.$$

Next, the second condition in Eq. (9) implies that $T_n\rho_n \le S_n n^{-c_{\rho,2}}$ for some $c_{\rho,2} > 0$. Since $\rho_n \to 0$, for all sufficiently large $n$,

$$(T_n - S_n)\log\frac{1}{1-\rho_n} \le 2T_n\rho_n \lesssim S_n. \tag{73}$$

Thus, the second probability is also controlled. The third probability, corresponding to the variance parameter, is handled in the same way as in Eq. (67)–Eq. (68).

Now we show that (45) holds. As in Eq. (69), it suffices to show that

$$\Pi_n\big((\mathcal{F}_n^{\mathrm{ber}})^c\big) = o(e^{-Cn\epsilon_n^2}).$$

Under the Bernoulli spike-and-slab prior in Eq. (8), there are two ways to leave the sieve: either the number of nonzero coordinates exceeds $\overline{S}_n$, or some nonzero coordinate exceeds $\overline{B}_n$. Thus,

$$\Pi_n\big((\mathcal{F}_n^{\mathrm{ber}})^c\big) \le \Pi_n(\|\boldsymbol{\gamma}\|_0 > \overline{S}_n) + \Pi_n\Big(\|\boldsymbol{\theta}\|_\infty > \overline{B}_n\Big). \tag{74}$$

Write $\|\boldsymbol{\gamma}\|_0 \sim \mathrm{Bin}(T_n, \rho_n)$, and denote its mean by $\mu_n = T_n\rho_n$. From Eq. (73), we have $\mu_n \le S_n/e$ for all sufficiently large $n$. Therefore, by the Chernoff bound for the binomial distribution,

$$\Pi_n(\|\boldsymbol{\gamma}\|_0 > \overline{S}_n) \le \left(\frac{e\mu_n}{\overline{S}_n}\right)^{\overline{S}_n}.$$

Taking negative logarithms and using the second condition in Eq. (9), we obtain

$$-\log \Pi_n(\|\boldsymbol{\gamma}\|_0 > \overline{S}_n) \geq \overline{S}_n \left(\log \frac{\overline{S}_n}{\mu_n} - 1\right) \gtrsim \overline{S}_n \log n.$$

Since $\overline{S}_n = \overline{S}_0 S_n$ and $S_n \log n \asymp n\epsilon_n^2$, choosing $\overline{S}_0 > 0$ sufficiently large makes the preceding negative logarithm sufficiently larger than $Cn\epsilon_n^2$. Therefore,

$$\Pi_n(\|\boldsymbol{\gamma}\|_0 > \overline{S}_n) = o(e^{-Cn\epsilon_n^2})$$

holds.

Next, the second probability is handled as follows:

$$\Pi_n\Big(\|\boldsymbol{\theta}\|_\infty > \overline{B}_n\Big) = \Pi_n\Big(\max_{1\leq t\leq T_n} \gamma_t|\theta_t| > \overline{B}_n\Big) \leq \sum_{t=1}^{T_n} \Pi_n\Big(\gamma_t = 1,\ |\theta_t| > \overline{B}_n\Big).$$

The probability in the last sum is bounded, exactly as in Eq. (70), by

$$\begin{aligned}\sum_{t=1}^{T_n} \Pi_n\Big(\gamma_t = 1,\ |\theta_t| > \overline{B}_n\Big) &\leq T_n \rho_n \int_{|u|>\overline{B}_n} \tilde{\pi}_{SL}(u)du \\ &\leq \mu_n c_2 \exp\left[-c_3 \left(\frac{\overline{B}_n}{\tau_n}\right)^\alpha\right].\end{aligned}$$

Similarly, taking negative logarithms gives

$$\left(\frac{\overline{B}_n}{\tau_n}\right)^\alpha \gtrsim N_n \log n \asymp n\epsilon_n^2.$$

Therefore, by choosing $\overline{B}_0 > 0$ sufficiently large, this term is of order $o(e^{-Cn\epsilon_n^2})$. □

### C.4 Proof of Theorem 3.8

*Proof.* Fix the unknown smoothness vector $\mathbf{s} = (s_1, \ldots, s_d) \in \mathfrak{S}$, and write $\tilde{\mathbf{s}}$, $\underline{\mathbf{s}}$, and $\omega = (1/p - 1/2)_+$ as defined in Section 3. For brevity, throughout this proof let

$$N_n := \left\lceil n^{1/(2\tilde{\mathbf{s}}+1)} \right\rceil, \qquad \epsilon_n := \epsilon_n(\mathbf{s}) = n^{-\tilde{\mathbf{s}}/(2\tilde{\mathbf{s}}+1)} (\log n)^{1/2},$$

so that $N_n \log n \asymp n\epsilon_n^2$.

Next, we define the cutoff

$$\overline{N}_n := \left\lceil \overline{N}_0 N_n \right\rceil,$$

where the constant $\overline{N}_0 > 0$ will be chosen sufficiently large below. For each $1 \leq N \leq \overline{N}_n$, let

$$\mathcal{F}_n^{\mathrm{ad}}(N) := \left\{ f_{\boldsymbol{\theta}} \in KAN(L_0, D(N), G(N), H(N); m) : \|\boldsymbol{\theta}\|_\infty \leq \overline{B}_n(N),\ \|\boldsymbol{\theta}\|_0 \leq S(N) \right\},$$

where the truncation sequence is

$$\overline{B}_n(N) := \overline{B}_0\, N^{\beta_{\mathrm{ad}}} (N_n \log n)^{1/\alpha},$$

and $S(N)$ is the sparsity level used in the adaptive prior. The constant $\overline{B}_0 > 0$ will be chosen sufficiently large when controlling the sieve-complement probability.

Define the clipped sieve for each $N$ by

$$\widetilde{\mathcal{F}}_n^{\mathrm{ad}}(N) := \left\{ \mathrm{clip} \circ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}} \in \mathcal{F}_n^{\mathrm{ad}}(N) \right\}.$$

Finally, define the full adaptive sieves by

$$\mathcal{F}_n^{\mathrm{ad}} := \bigcup_{N=1}^{\overline{N}_n} \mathcal{F}_n^{\mathrm{ad}}(N), \qquad \widetilde{\mathcal{F}}_n^{\mathrm{ad}} := \bigcup_{N=1}^{\overline{N}_n} \widetilde{\mathcal{F}}_n^{\mathrm{ad}}(N),$$

and set

$$\widetilde{\mathcal{M}}_n^{\mathrm{ad}} := \widetilde{\mathcal{F}}_n^{\mathrm{ad}} \times [\underline{\sigma}, \overline{\sigma}].$$

We verify conditions Eq. (43)–Eq. (45) of Lemma C.1 for the rate $\epsilon_n$. Note that Lemma C.1 is applied with $\mathcal{F}_n = \widetilde{\mathcal{F}}_n^{\mathrm{ad}}$ and $\mathcal{M}_n = \widetilde{\mathcal{M}}_n^{\mathrm{ad}}$.

We first verify the entropy condition (43). By (A2), $\|g\|_{L^2(P_X)} \le C_X^{1/2}\|g\|_{L^\infty}$. Moreover, since clip is 1-Lipschitz, an $L^\infty$-cover of $\mathcal{F}_n^{\mathrm{ad}}(N)$ induces an $L^\infty$-cover, hence an $L^2(P_X)$-cover, of $\widetilde{\mathcal{F}}_n^{\mathrm{ad}}(N)$. Therefore,

$$\log \mathcal{N}\Big(\epsilon_n,\, \widetilde{\mathcal{F}}_n^{\mathrm{ad}},\, \|\cdot\|_{L^2(P_X)}\Big) \lesssim \log \sum_{N=1}^{\overline{N}_n} \mathcal{N}\Big(C_X^{-1/2}\epsilon_n,\, \mathcal{F}_n^{\mathrm{ad}}(N),\, \|\cdot\|_{L^\infty}\Big). \tag{75}$$

Applying Lemma B.15 to each $\mathcal{F}_n^{\mathrm{ad}}(N)$ gives

$$\begin{aligned}
&\log \mathcal{N}\Big(C_X^{-1/2}\epsilon_n,\, \mathcal{F}_n^{\mathrm{ad}}(N),\, \|\cdot\|_{L^\infty}\Big) \\
&\quad\lesssim S(N)\log\frac{eT(N)}{S(N)} \\
&\qquad + S(N)\log\left(1 + \frac{\overline{B}_n(N)\,K(L_0, D(N), G(N), H(N), \overline{B}_n(N); a_0, b_0, d, m)}{C_X^{-1/2}\epsilon_n}\right).
\end{aligned} \tag{76}$$

Since $D(N) \asymp N$, $G(N)/H(N) \asymp 1$, $H(N) \asymp N^{\beta_{\mathrm{ad}}}$, and $L_0$ is fixed, we have $T(N) \lesssim N^{2+\beta_{\mathrm{ad}}}$. Moreover, uniformly over $1 \le N \le \overline{N}_n$,

$$\begin{aligned}
\log T(N) &\lesssim \log n, \\
\log \overline{B}_n(N) &\lesssim \log n, \\
\log K\big(L_0, D(N), G(N), H(N), \overline{B}_n(N); a_0, b_0, d, m\big) &\lesssim \log n.
\end{aligned}$$

Hence the right-hand side of Eq. (76) is bounded by $C \cdot S(N)\log n \lesssim N \log n$. Substituting into Eq. (75),

$$\log \mathcal{N}\Big(\epsilon_n,\, \widetilde{\mathcal{F}}_n^{\mathrm{ad}},\, \|\cdot\|_{L^2(P_X)}\Big) \lesssim \log \overline{N}_n + \overline{N}_n \log n \lesssim N_n \log n \asymp n\epsilon_n^2,$$

where the last equivalence follows from the definitions of $N_n$ and $\epsilon_n$. This verifies Eq. (43).

Next, we establish the prior-mass condition (44). Applying Lemma B.4 with $N = N_n$ and then Remark B.5, which applies here because $\beta < \beta_{ad}$ implies $B_0 N_n^\beta \le B^\star(N_n)$ for all sufficiently large $n$, there exists

$$f_n^\star \in KAN_c\big(L_0,\, D(N_n),\, G(N_n),\, H(N_n);\, m\big)$$

with realizing parameter vector $\boldsymbol{\theta}_n^\star$ satisfying

$$\|\boldsymbol{\theta}_n^\star\|_0 \le S(N_n), \qquad \|\boldsymbol{\theta}_n^\star\|_\infty \le B^\star(N_n),$$

and

$$\|f_0 - f_n^\star\|_{L^2} \lesssim N_n^{-\tilde{\mathbf{s}}} \lesssim \epsilon_n (\log n)^{-1/2}.$$

Therefore, for all sufficiently large $n$, by (A2),

$$\|f_0 - f_n^\star\|_{L^2(P_X)} \le \epsilon_n/4.$$

Moreover, since $B^\star(N_n) \le \overline{B}_n(N_n)$ for all sufficiently large $n$, we have $f_n^\star \in \mathcal{F}_n^{\mathrm{ad}}(N_n)$.

Let $I_n^\star := \{t : \theta_{n,t}^\star \neq 0\}$. Choose $J_n^\star \subset \{1, \dots, T(N_n)\}$ such that $I_n^\star \subset J_n^\star$ and $|J_n^\star| = S(N_n)$, and set $\mathbf{1}_{J_n^\star} \in \{0,1\}^{T(N_n)}$ as in Eq. (62). Define

$$K_n^\star := K\big(L_0, D(N_n), G(N_n), H(N_n), B^\star(N_n); a_0, b_0, d, m\big), \qquad \delta_n := \frac{\epsilon_n}{4 C_X^{1/2} K_n^\star}.$$

Since $\log K_n^\star \lesssim \log n$, we have $\delta_n \le 1$ and $\log \delta_n^{-1} \lesssim \log n$ for all sufficiently large $n$. By Corollary B.14, if $\|\boldsymbol{\theta}-\boldsymbol{\theta}_n^\star\|_\infty \le \delta_n$, then $\|f_{\boldsymbol{\theta}}-f_n^\star\|_{L^\infty} \le \epsilon_n/(4C_X^{1/2})$, and hence $\|f_{\boldsymbol{\theta}}-f_n^\star\|_{L^2(P_X)} \le \epsilon_n/4$. Since $f_0 \in \mathcal{UB}$, we have $\mathrm{clip}\circ f_0 = f_0$. Moreover, clip is 1-Lipschitz. Therefore,

$$\begin{aligned}\|\,\mathrm{clip}\circ f_{\boldsymbol{\theta}} - f_0\|_{L^2(P_X)} &= \|\,\mathrm{clip}\circ f_{\boldsymbol{\theta}} - \mathrm{clip}\circ f_0\|_{L^2(P_X)} \\ &\le \|\,\mathrm{clip}\circ f_{\boldsymbol{\theta}} - \mathrm{clip}\circ f_n^\star\|_{L^2(P_X)} + \|\,\mathrm{clip}\circ f_n^\star - \mathrm{clip}\circ f_0\|_{L^2(P_X)} \\ &\le \|f_{\boldsymbol{\theta}} - f_n^\star\|_{L^2(P_X)} + \|f_n^\star - f_0\|_{L^2(P_X)} \le \epsilon_n/2.\end{aligned}$$

Let $A_n$ denote the KL neighborhood in Lemma C.1, where Lemma C.1 is applied with $\mathcal{F}_n = \widetilde{\mathcal{F}}_n^{\mathrm{ad}}$. Then

$$\begin{aligned}\Pi_n(A_n) \ge \pi_N(N_n)\cdot \Pi\big(\boldsymbol{\gamma} = \mathbf{1}_{J_n^\star} \mid N = N_n\big) \\ \times \Pi\big(\|\boldsymbol{\theta}-\boldsymbol{\theta}_n^\star\|_\infty \le \delta_n \mid \boldsymbol{\gamma} = \mathbf{1}_{J_n^\star},\, N = N_n\big) \\ \times \Pi(|\sigma-\sigma_0| \le \epsilon_n/2)\,.\end{aligned} \tag{77}$$

We bound each factor in turn. Since $\pi_N(N) \propto \exp(-\lambda_N N\log N)$,

$$-\log \pi_N(N_n) \lesssim N_n \log N_n \lesssim N_n \log n. \tag{78}$$

The combinatorial prior assigns $\Pi(\boldsymbol{\gamma} = \mathbf{1}_{J_n^\star} \mid N = N_n) = \binom{T(N_n)}{S(N_n)}^{-1}$, so

$$-\log \Pi\big(\boldsymbol{\gamma} = \mathbf{1}_{J_n^\star} \mid N = N_n\big) \le S(N_n)\log\frac{eT(N_n)}{S(N_n)} \lesssim N_n \log n. \tag{79}$$

By (C1), using $\|\boldsymbol{\theta}_n^\star\|_\infty \le B^\star(N_n)$ and $\delta_n \le 1$,

$$\begin{aligned}\Pi\big(\|\boldsymbol{\theta}-\boldsymbol{\theta}_n^\star\|_\infty \le \delta_n \mid \boldsymbol{\gamma} = \mathbf{1}_{J_n^\star},\, N = N_n\big) \\ \ge \left[2\delta_n \inf_{|u|\le B^\star(N_n)+1} \widetilde{\pi}_{SL}(u \mid N_n)\right]^{S(N_n)}.\end{aligned}$$

Therefore,

$$-\log \Pi\big(\boldsymbol{\theta} : \|\boldsymbol{\theta}-\boldsymbol{\theta}_n^\star\|_\infty \le \delta_n \mid \boldsymbol{\gamma} = \mathbf{1}_{J_n^\star},\, N = N_n\big) \lesssim S(N_n)\{\log\delta_n^{-1} + \log n\} \lesssim N_n \log n. \tag{80}$$

By (C3), the same calculation as in Eq. (67)–Eq. (68) gives

$$-\log \Pi(|\sigma-\sigma_0| \le \epsilon_n/2) \lesssim \log(\epsilon_n^{-1}) \lesssim \log n. \tag{81}$$

Combining Eq. (78)–Eq. (81) in Eq. (77),

$$-\log \Pi_n(A_n) \lesssim N_n \log n \asymp n\epsilon_n^2,$$

which verifies Eq. (44).

Finally, we verify the sieve-complement condition (45). By (C3), the variance prior is supported on $[\underline{\sigma}^2, \overline{\sigma}^2]$. Since Lemma C.1 is applied with $\mathcal{F}_n = \widetilde{\mathcal{F}}_n^{\mathrm{ad}}$ and $\mathcal{M}_n = \widetilde{\mathcal{M}}_n^{\mathrm{ad}}$, we have

$$\begin{aligned}\Pi_n^{\mathrm{ad}}\big((\widetilde{\mathcal{M}}_n^{\mathrm{ad}})^c\big) = \Pi_n^{\mathrm{ad}}\big((\widetilde{\mathcal{F}}_n^{\mathrm{ad}})^c\big) \\ \le \pi_N(N > \overline{N}_n) + \sum_{N=1}^{\overline{N}_n} \pi_N(N)\,\Pi\big(\|\boldsymbol{\theta}\|_\infty > \overline{B}_n(N) \mid N\big)\,.\end{aligned}$$

Indeed, for $N \le \overline{N}_n$, the event $\|\boldsymbol{\theta}\|_\infty \le \overline{B}_n(N)$ implies $f_{\boldsymbol{\theta}} \in \mathcal{F}_n^{\mathrm{ad}}(N)$, and hence $\mathrm{clip}\circ f_{\boldsymbol{\theta}} \in \widetilde{\mathcal{F}}_n^{\mathrm{ad}}(N) \subset \widetilde{\mathcal{F}}_n^{\mathrm{ad}}$. For all sufficiently large $n$,

$$\pi_N(N > \overline{N}_n) \lesssim \sum_{N>\overline{N}_n} \exp(-\lambda_N N\log N) \lesssim \exp\left\{-\frac{\lambda_N}{2}\overline{N}_n \log \overline{N}_n\right\}.$$

Since $\overline{N}_n \log \overline{N}_n \asymp \overline{N}_0 N_n \log n$, choosing $\overline{N}_0$ sufficiently large gives

$$\pi_N(N > \overline{N}_n) = o\Big(e^{-Cn\epsilon_n^2}\Big) \tag{82}$$

for the constant $C > 0$ in Lemma C.1.

Since conditionally on $N$ the spike-and-slab prior selects exactly $S(N)$ nonzero coordinates, by the union bound and (C2),

$$\Pi\big(\|\boldsymbol{\theta}\|_\infty > \overline{B}_n(N) \mid N\big) \le S(N) \int_{|u|>\overline{B}_n(N)} \widetilde{\pi}_{SL}(u \mid N)\, du \le c_2 S(N) \exp\left[-c_3 \left(\frac{\overline{B}_n(N)}{\tau_N}\right)^{\alpha}\right].$$

Since $\tau_N \lesssim N^{\beta_{\mathrm{ad}}}$, the definition of $\overline{B}_n(N)$ implies, uniformly over $1 \le N \le \overline{N}_n$,

$$\left(\frac{\overline{B}_n(N)}{\tau_N}\right)^{\alpha} \gtrsim \overline{B}_0^{\alpha} N_n \log n.$$

Therefore,

$$\sum_{N=1}^{\overline{N}_n} \pi_N(N)\, \Pi\big(\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_\infty > \overline{B}_n(N) \mid N\big) \le c_2 S(\overline{N}_n) \exp\Big[-c_4 \overline{B}_0^{\alpha} N_n \log n\Big].$$

Since $S(\overline{N}_n) \lesssim N_n$ and $N_n \log n \asymp n\epsilon_n^2$, choosing $\overline{B}_0 > 0$ sufficiently large gives

$$\sum_{N=1}^{\overline{N}_n} \pi_N(N)\, \Pi\big(\|\boldsymbol{\theta}\|_\infty > \overline{B}_n(N) \mid N\big) = o\Big(e^{-Cn\epsilon_n^2}\Big). \tag{83}$$

Combining Eq. (82) and Eq. (83) verifies Eq. (45). Applying Lemma C.1 completes the proof. □

### C.5 Proof of Theorem 3.10

*Proof.* Let $N_n^* := \lceil n^{1/(2\tilde{s}^*+1)} \rceil$ and $\epsilon_n^* := n^{-\tilde{s}^*/(2\tilde{s}^*+1)} (\log n)^{1/2}$, so that $N_n^* \log n \asymp n(\epsilon_n^*)^2$. Introduce the auxiliary truncation sequence

$$\overline{B}_n^{\mathrm{cp}} := \overline{B}_{\mathrm{cp}} (N_n^*)^{\beta_{\mathrm{cp}}+1/\alpha} (\log n)^{1/\alpha},$$

where $\overline{B}_{\mathrm{cp}} > 0$ is chosen sufficiently large so that the sieve-complement calculation closes. Define the raw sieve

$$\mathcal{F}_n^{\mathrm{cp}} := \Big\{ f_{\boldsymbol{\theta}} \in KAN(L_{\mathrm{cp}}, D_n^{\mathrm{cp}}, G_n^{\mathrm{cp}}, H_n^{\mathrm{cp}}; m) : \|\boldsymbol{\theta}\|_\infty \le \overline{B}_n^{\mathrm{cp}},\ \|\boldsymbol{\theta}\|_0 \le S_n^{\mathrm{cp}} \Big\}$$

and the corresponding clipped sieve

$$\widetilde{\mathcal{F}}_n^{\mathrm{cp}} := \{ \mathrm{clip} \circ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}} \in \mathcal{F}_n^{\mathrm{cp}} \}.$$

Set

$$\widetilde{\mathcal{M}}_n^{\mathrm{cp}} := \widetilde{\mathcal{F}}_n^{\mathrm{cp}} \times [\underline{\sigma}, \overline{\sigma}].$$

We verify conditions Eq. (43)–Eq. (45) of Lemma C.1 for the rate $\epsilon_n^*$. Note that Lemma C.1 is applied with $\mathcal{F}_n = \widetilde{\mathcal{F}}_n^{\mathrm{cp}}$ and $\mathcal{M}_n = \widetilde{\mathcal{M}}_n^{\mathrm{cp}}$.

Now we prove that (43) holds. Let

$$T_n^{\mathrm{cp}} := T(L_{\mathrm{cp}}, D_n^{\mathrm{cp}}, G_n^{\mathrm{cp}})$$

and

$$K_n^{\mathrm{cp}} := K(L_{\mathrm{cp}}, D_n^{\mathrm{cp}}, G_n^{\mathrm{cp}}, H_n^{\mathrm{cp}}, \overline{B}_n^{\mathrm{cp}}; a_0, b_0, d, m).$$

By Assumption (A2), $\|g\|_{L^2(P_X)} \le C_X^{1/2} \|g\|_{L^\infty}$. Moreover, since clip is 1-Lipschitz, an $L^\infty$-cover of $\mathcal{F}_n^{\mathrm{cp}}$ induces an $L^\infty$-cover, hence an $L^2(P_X)$-cover, of $\widetilde{\mathcal{F}}_n^{\mathrm{cp}}$. Therefore, by Lemma B.15,

$$\begin{aligned}
\log \mathcal{N}\big(\epsilon_n^*, \widetilde{\mathcal{F}}_n^{\mathrm{cp}}, \|\cdot\|_{L^2(P_X)}\big) &\le \log \mathcal{N}\big(C_X^{-1/2}\epsilon_n^*, \mathcal{F}_n^{\mathrm{cp}}, \|\cdot\|_{L^\infty}\big) \\
&\lesssim S_n^{\mathrm{cp}} \log \frac{e T_n^{\mathrm{cp}}}{S_n^{\mathrm{cp}}} + S_n^{\mathrm{cp}} \log\left(1 + \frac{\overline{B}_n^{\mathrm{cp}} K_n^{\mathrm{cp}}}{C_X^{-1/2}\epsilon_n^*}\right).
\end{aligned}$$

Since $L_{\mathrm{cp}}$ is a fixed constant, we have $T_n^{\mathrm{cp}} \lesssim (N_n^*)^{2+\beta_{\mathrm{cp}}}$ and $\log K_n^{\mathrm{cp}} \lesssim \log n$, so the right-hand side is bounded by

$$S_n^{\mathrm{cp}} \log n \asymp N_n^* \log n \asymp n(\epsilon_n^*)^2,$$

verifying Eq. (43).

Next, we show (44). Applying Lemma B.9 with $N = N_n^*$, there exists $f_n^{\mathrm{cp},*} \in KAN(L_{\mathrm{cp}}, D_n^{\mathrm{cp}}, G_n^{\mathrm{cp}}, H_n^{\mathrm{cp}}; m)$ with realizing parameter vector $\boldsymbol{\theta}_n^{\mathrm{cp},*}$ satisfying $\|\boldsymbol{\theta}_n^{\mathrm{cp},*}\|_\infty \le B_n^{\mathrm{cp},*}$ and $\|\boldsymbol{\theta}_n^{\mathrm{cp},*}\|_0 \le S_n^{\mathrm{cp}}$, and such that

$$\|f_0 - f_n^{\mathrm{cp},*}\|_{L^2(P_X)} \lesssim (N_n^*)^{-\tilde{s}^*} \lesssim \epsilon_n^* (\log n)^{-1/2} \le \frac{\epsilon_n^*}{4}$$

for all sufficiently large $n$. Since $B_n^{\mathrm{cp},*} \le \overline{B}_n^{\mathrm{cp}}$ for all sufficiently large $n$, we have $f_n^{\mathrm{cp},*} \in \mathcal{F}_n^{\mathrm{cp}}$.

Let $I_n^{\mathrm{cp},*} := \{t : \theta_{n,t}^{\mathrm{cp},*} \neq 0\}$ and choose $J_n^{\mathrm{cp},*} \subset \{1, \dots, T_n^{\mathrm{cp}}\}$ with $I_n^{\mathrm{cp},*} \subset J_n^{\mathrm{cp},*}$ and $|J_n^{\mathrm{cp},*}| = S_n^{\mathrm{cp}}$. Define

$$K_n^{\mathrm{cp},*} := K\big(L_{\mathrm{cp}}, D_n^{\mathrm{cp}}, G_n^{\mathrm{cp}}, H_n^{\mathrm{cp}}, B_n^{\mathrm{cp},*}; a_0, b_0, d, m\big), \qquad \delta_n^{\mathrm{cp}} := \frac{\epsilon_n^*}{4 C_X^{1/2} K_n^{\mathrm{cp},*}}.$$

Since $\log K_n^{\mathrm{cp},*} \lesssim \log n$, we have $\delta_n^{\mathrm{cp}} \le 1$ and $\log(\delta_n^{\mathrm{cp}})^{-1} \lesssim \log n$ for all sufficiently large $n$. By Corollary B.14, $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n^{\mathrm{cp},*}\|_\infty \le \delta_n^{\mathrm{cp}}$ implies

$$\|f_{\boldsymbol{\theta}} - f_n^{\mathrm{cp},*}\|_{L^\infty} \le \frac{\epsilon_n^*}{4 C_X^{1/2}},$$

and hence

$$\|f_{\boldsymbol{\theta}} - f_n^{\mathrm{cp},*}\|_{L^2(P_X)} \le \frac{\epsilon_n^*}{4}.$$

Since $f_0 \in \mathcal{UB}$, we have $\mathrm{clip} \circ f_0 = f_0$. Moreover, clip is 1-Lipschitz. Therefore,

$$\begin{aligned}
\|\mathrm{clip} \circ f_{\boldsymbol{\theta}} - f_0\|_{L^2(P_X)} &= \|\mathrm{clip} \circ f_{\boldsymbol{\theta}} - \mathrm{clip} \circ f_0\|_{L^2(P_X)} \\
&\le \|\mathrm{clip} \circ f_{\boldsymbol{\theta}} - \mathrm{clip} \circ f_n^{\mathrm{cp},*}\|_{L^2(P_X)} + \|\mathrm{clip} \circ f_n^{\mathrm{cp},*} - \mathrm{clip} \circ f_0\|_{L^2(P_X)} \\
&\le \|f_{\boldsymbol{\theta}} - f_n^{\mathrm{cp},*}\|_{L^2(P_X)} + \|f_n^{\mathrm{cp},*} - f_0\|_{L^2(P_X)} \\
&\le \frac{\epsilon_n^*}{2}.
\end{aligned}$$

Let $A_n$ denote the KL neighborhood in Lemma C.1, where Lemma C.1 is applied with $\mathcal{F}_n = \widetilde{\mathcal{F}}_n^{\mathrm{cp}}$. Using the same prior mass decomposition as in the proof of Theorem 3.2 (cf. Eq. (60)–Eq. (68)),

$$\Pi_n(A_n) \ge \Pi_n\big(\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_n^{\mathrm{cp},*}\|_\infty \le \delta_n^{\mathrm{cp}} \,\big|\, \boldsymbol{\gamma} = \mathbf{1}_{J_n^{\mathrm{cp},*}}\big)\, \pi_n\big(\boldsymbol{\gamma} = \mathbf{1}_{J_n^{\mathrm{cp},*}}\big)\, \Pi_n(\sigma : |\sigma - \sigma_0| \le \epsilon_n^*/2)\,.$$

Since $\pi_n(\boldsymbol{\gamma} = \mathbf{1}_{J_n^{\mathrm{cp},*}}) = \binom{T_n^{\mathrm{cp}}}{S_n^{\mathrm{cp}}}^{-1}$, we have

$$-\log \pi_n(\boldsymbol{\gamma} = \mathbf{1}_{J_n^{\mathrm{cp},*}}) \le S_n^{\mathrm{cp}} \log \frac{e T_n^{\mathrm{cp}}}{S_n^{\mathrm{cp}}} \lesssim N_n^* \log n.$$

Moreover, by (B1), using $\|\boldsymbol{\theta}_n^{\mathrm{cp},*}\|_\infty \le B_n^{\mathrm{cp},*}$ and $\delta_n^{\mathrm{cp}} \le 1$,

$$\begin{aligned}
&\Pi_n\big(\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_n^{\mathrm{cp},*}\|_\infty \le \delta_n^{\mathrm{cp}} \,\big|\, \boldsymbol{\gamma} = \mathbf{1}_{J_n^{\mathrm{cp},*}}\big) \\
&\quad \ge \left(2\delta_n^{\mathrm{cp}} \inf_{|u| \le B_n^{\mathrm{cp},*}+1} \widetilde{\pi}_{SL}(u)\right)^{S_n^{\mathrm{cp}}},
\end{aligned}$$

and hence

$$-\log \Pi_n\big(\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_n^{\mathrm{cp},*}\|_\infty \le \delta_n^{\mathrm{cp}} \,\big|\, \boldsymbol{\gamma} = \mathbf{1}_{J_n^{\mathrm{cp},*}}\big) \lesssim S_n^{\mathrm{cp}} \{\log(\delta_n^{\mathrm{cp}})^{-1} + \log n\} \lesssim N_n^* \log n.$$

Finally, by (B3), the same calculation as in Eq. (67)–Eq. (68) gives

$$-\log \Pi_n(\sigma : |\sigma - \sigma_0| \le \epsilon_n^*/2) \lesssim \log((\epsilon_n^*)^{-1}) \lesssim \log n.$$

Combining these bounds yields

$$-\log \Pi_n(A_n) \lesssim N_n^* \log n \asymp n(\epsilon_n^*)^2,$$

which verifies Eq. (44).

Finally, we show that (45) holds. By (B3), the variance prior is supported on $[\underline{\sigma}^2, \overline{\sigma}^2]$. Since Lemma C.1 is applied with $\mathcal{F}_n = \widetilde{\mathcal{F}}_n^{\mathrm{cp}}$ and $\mathcal{M}_n = \widetilde{\mathcal{M}}_n^{\mathrm{cp}}$, we have

$$\Pi_n\big((\widetilde{\mathcal{M}}_n^{\mathrm{cp}})^c\big) = \Pi_n\big((\widetilde{\mathcal{F}}_n^{\mathrm{cp}})^c\big) \le \Pi_n\big((\mathcal{F}_n^{\mathrm{cp}})^c\big).$$

Since the spike-and-slab prior Eq. (5) fixes exactly $S_n^{\mathrm{cp}}$ active coordinates, the only way to leave $\mathcal{F}_n^{\mathrm{cp}}$ is for some active slab coefficient to exceed $\overline{B}_n^{\mathrm{cp}}$. By the union bound and (B2),

$$\Pi_n\big((\mathcal{F}_n^{\mathrm{cp}})^c\big) \le S_n^{\mathrm{cp}} \cdot c_2 \exp\left[-c_3 \left(\frac{\overline{B}_n^{\mathrm{cp}}}{\tau_n}\right)^{\alpha}\right].$$

Using $\tau_n \lesssim (N_n^*)^{\beta_{\mathrm{cp}}}$ and the definition of $\overline{B}_n^{\mathrm{cp}}$,

$$\left(\frac{\overline{B}_n^{\mathrm{cp}}}{\tau_n}\right)^{\alpha} \gtrsim N_n^* \log n \asymp n(\epsilon_n^*)^2.$$

Choosing $\overline{B}_{\mathrm{cp}}$ sufficiently large gives

$$\Pi_n\big((\widetilde{\mathcal{M}}_n^{\mathrm{cp}})^c\big) = o(e^{-Cn(\epsilon_n^*)^2}),$$

verifying Eq. (45). Applying Lemma C.1 completes the proof. □

# D Numerical study

This section reports a numerical study of sparse Bayesian KANs (BKANs) on the target functions in Figure 2.

## D.1 Setup

**Data.** For each target $f_0$, we generate $X_i \sim U([0,1]^d)$ and $Y_i = f_0(X_i) + \varepsilon_i$ with $\varepsilon_i \sim N(0, 0.1^2)$, where $d = 2$ for $f_1, \ldots, f_5$ and $d = 10$ for $f_6$. Every target is affinely normalized so that its range lies in $[-1, 1]$:

$$f_1(\mathbf{x}) = \tfrac{2}{3}\big\{\mathbf{1}(x_1 \geq 1/2) + \sin(2\pi x_2) - 1/2\big\}, \quad f_2(\mathbf{x}) = \tfrac{8}{3}\big\{|x_1 - 1/2| + (x_2 - 1/2)^2\big\} - 1,$$
$$f_3(\mathbf{x}) = f_1(R\mathbf{x}), \quad f_4(\mathbf{x}) = f_2(R\mathbf{x}), \quad f_5(\mathbf{x}) = f_3\big((f_3(\mathbf{x}) + 1)/2,\ (f_4(\mathbf{x}) + 1)/2\big),$$

where $R$ is the rotation by $45^\circ$ about $(1/2, 1/2)$, so that the jump of $f_3$ lies on the line $x_1 + x_2 = 1$, and $f_6(\mathbf{x}) = \{10\sin(\pi x_1 x_2) + 20(x_3 - 1/2)^2 + 10x_4 + 5x_5 - 15\}/15$ is the first function of Friedman (1991), for which the noise standard deviation is $1/15$ (that is, 1 on the original scale). For each target and replication, we draw a training pool of $4{,}000$ observations and use its first $n \in \{250, 500, 1{,}000, 2{,}000, 4{,}000\}$ observations as the training set, together with a validation set of size 100 and two test sets of size $1{,}000$: one with the noiseless values $f_0(\tilde{X}_i)$, used for the estimation error, and one with noisy responses $\tilde{Y}_i = f_0(\tilde{X}_i) + \tilde{\varepsilon}_i$, used for the predictive intervals. Every configuration is replicated five times.

**Models.** The BKAN uses a compact instance of the architecture (4) with $d = 2$: depth $L_0 = 5$, width $D_n = 4$ with $N_n = 1$, cubic splines ($m = 3$), and hidden-layer knots on $[-H_n, H_n]$ with $H_n = 2$ and $G_n = 4H_n - 2m = 2$, so that the knot spacing is $1/2$; the first layer uses three knot intervals on $[0, 1]$. This gives 368 parameters, counting the spline coefficients and the SiLU weight of every edge. For $f_6$ we keep the same depth and width, that is, the layer widths $(10, 4, 4, 4, 4, 1)$ with 592 parameters. As a continuous relaxation of the spike-and-slab prior (5), each coefficient receives the Gaussian mixture prior $(1 - \pi)N(0, 0.05^2) + \pi N(0, 1)$ with $\pi = 0.1$, and the noise precision $\sigma^{-2}$ receives a $\mathrm{Gamma}(1, 0.02)$ prior. For comparison, we consider (i) a Bayesian ReLU network (BNN) with four hidden layers of width 10 (width 12 for $f_6$), whose number of parameters matches that of the BKAN, with the same Gaussian mixture prior on every weight and bias and the same noise prior; (ii) BART (Chipman et al., 2010) with 50 trees and four chains of $1{,}000$ burn-in and $1{,}000$ posterior draws; (iii) a Gaussian process (GP) regression with an isotropic RBF kernel of fixed length scale $0.2$ for every target, including $f_6$ with $d = 10$, with the noise precision integrated out under the same Gamma prior; and (iv) a random forest (RF) (Breiman, 2001) with 500 trees.

**Inference and evaluation.** The BKAN and the BNN are fitted by mean-field Gaussian variational inference (VI). We report the restart with the smallest RMSE of the posterior predictive mean on the validation set; the validation set is not used by BART, the GP, or the RF. We evaluate the posterior mean $\hat{f}$ by the test RMSE $\{\sum_{i=1}^{1000}(\hat{f}(\tilde{X}_i) - f_0(\tilde{X}_i))^2/1000\}^{1/2}$ on the noiseless test set. For every Bayesian method we also report the pointwise $90\%$ predictive interval for a new response, given by the $5\%$ and $95\%$ quantiles of the posterior predictive draws of $Y$ (for the GP, of its predictive mixture), by its coverage of the noisy test responses $\tilde{Y}_i$ and its average width; since the noise standard deviation is $0.1$, the width of this interval cannot fall below about $0.33$.

**Computing environment.** All experiments were run on a Mac mini with an Apple M4 chip (10-core CPU), without GPU acceleration. The BKAN, the BNN, NUTS, and BART are implemented in `JAX` with `NumPyro` and `bartz` (Petrillo, 2024), in single precision; each fit runs on a single core, and ten fits run in parallel. The four NUTS chains of one fit also run one after another on a single core.

**Choice of the inference algorithm.** To assess the quality of VI, we compare it with the No-U-Turn sampler (NUTS) (Hoffman and Gelman, 2014) for the same BKAN on $f_1$ and $f_2$ at $n = 500$, running four chains of $1{,}500$ warm-up and $1{,}000$ posterior draws, each started at the initial value of the corresponding VI restart. As shown in Table 2 and Figure 3, NUTS attains a smaller RMSE and a coverage closer to the nominal level, while VI yields slightly wider and conservative intervals. Figure 4 shows that the two posteriors are nevertheless close in function space: both posterior means

recover the jump of $f_1$ and the kink of $f_2$, with both slightly smoothing the true function at $x_1 = 1/2$, while the VI intervals are somewhat wider than the NUTS intervals across the domain, in line with the higher coverage of VI in Table 2. However, NUTS takes 55–60 times longer than VI, about seven hours per fit. We use VI for all reported results.

Table 2: NUTS against VI for the same BKAN ($N_n = 1$, $n = 500$), averaged over five replications (standard deviation of the RMSE in parentheses). Coverage and width refer to the pointwise 90% predictive interval for a new response on the noisy test set; time is the median wall-clock time per fit.

| Target | Inference | RMSE | Coverage | Width | Minutes |
|---|---|---|---|---|---|
| $f_1$ | NUTS | 0.0339 (0.0020) | 0.902 | 0.346 | 414.7 |
| | VI | 0.0491 (0.0088) | 0.946 | 0.418 | 7.5 |
| $f_2$ | NUTS | 0.0237 (0.0026) | 0.896 | 0.332 | 456.4 |
| | VI | 0.0327 (0.0024) | 0.943 | 0.400 | 7.5 |

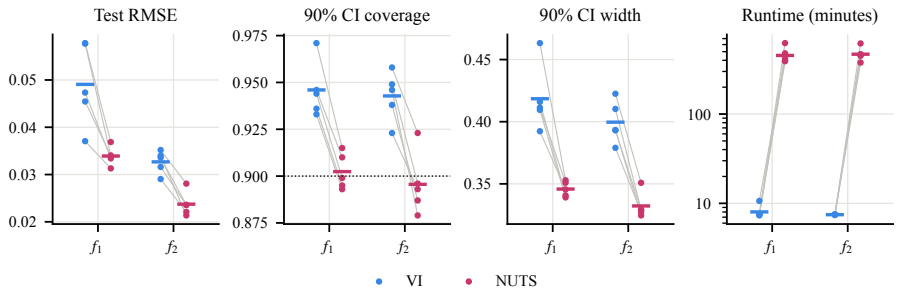
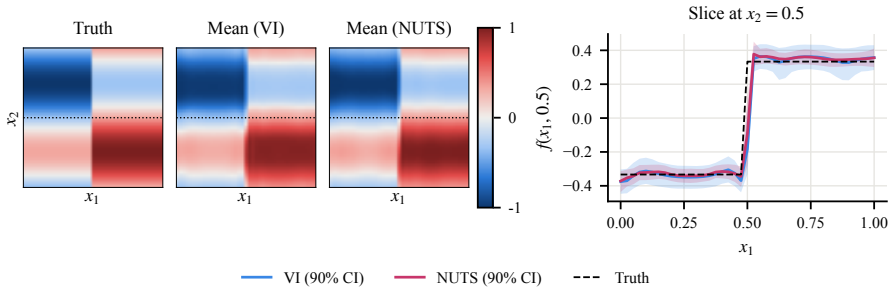



Figure 3: NUTS against VI for the same BKAN ($n = 500$). Dots are replications (paired by grey lines) and bars means (geometric mean for runtime). Coverage and width refer to the 90% predictive interval on the noisy test set, and the dotted line marks its nominal coverage.

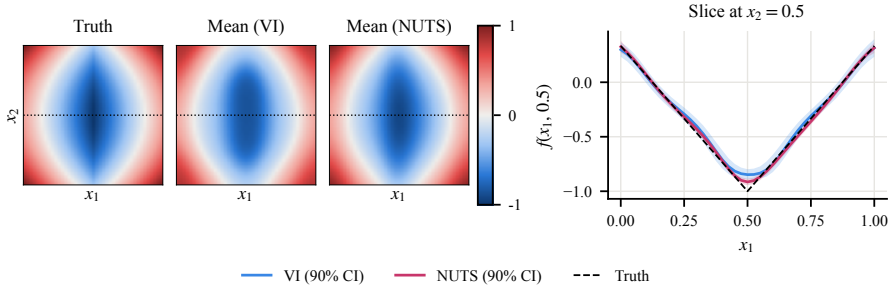
### D.2 Results

**Estimation error.** Figure 5 shows the test RMSE against the sample size. The RMSE of the BKAN decreases steadily in $n$ for every target, and the BKAN attains the smallest or near-smallest error in most settings, in particular for $f_2$, $f_4$, $f_5$, and $f_6$ at large $n$. For the composite target $f_5$, the BKAN has the smallest error for $n \geq 500$, in line with Theorem 3.10, and for $f_6$, the BKAN and the BNN adapt to the five inactive inputs, whereas the GP with a fixed isotropic kernel fails to do so. The exceptions are $f_1$, where BART is comparable to or slightly better than the BKAN, and $f_3$, where the BNN has a smaller error, especially at small $n$. Both are explained by the shape of the targets: the jump of $f_1$ is aligned with a coordinate axis, which a single tree split represents exactly, and the jump of $f_3$ lies on the hyperplane $x_1 + x_2 = 1$, which a single ReLU unit represents exactly. Even for $f_3$, the RMSE of the BKAN decreases faster than that of the other methods as $n$ grows.

**Posterior mean surfaces.** Figure 6 compares the posterior mean surfaces with the true functions at $n = 250$ and $n = 4{,}000$. At $n = 250$, the BKAN recovers $f_1$, $f_2$, $f_4$, and $f_6$ well but smooths out the diagonal jump of $f_3$, which the BNN locates more sharply. The tree-based methods, BART and RF, reproduce the axis-aligned jump of $f_1$ but show staircase artifacts on the rotated targets $f_3$, $f_4$, and $f_5$, and the GP oversmooths the discontinuities and nearly flattens $f_6$. At $n = 4{,}000$, where the design is dense enough, the BKAN recovers the diagonal jump of $f_3$ and the layered pattern of $f_5$.




(a) $f_1$




(b) $f_2$

Figure 4: VI against NUTS for the same BKAN ($N_n = 1$, $n = 500$, first replication) on (a) $f_1$ and (b) $f_2$: the true function and the posterior means (left), and their slices along the dotted line $x_2 = 1/2$ with pointwise 90% credible intervals (right).

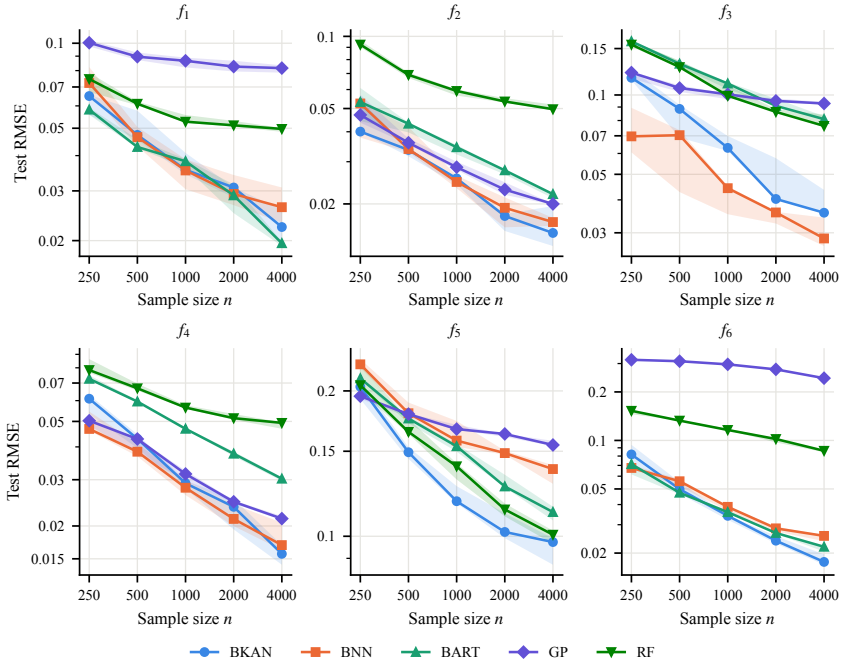
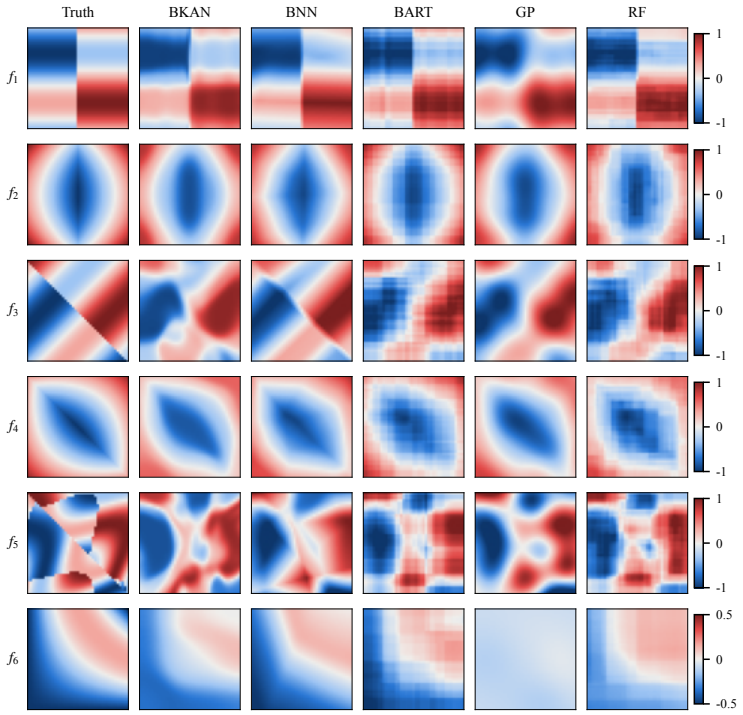
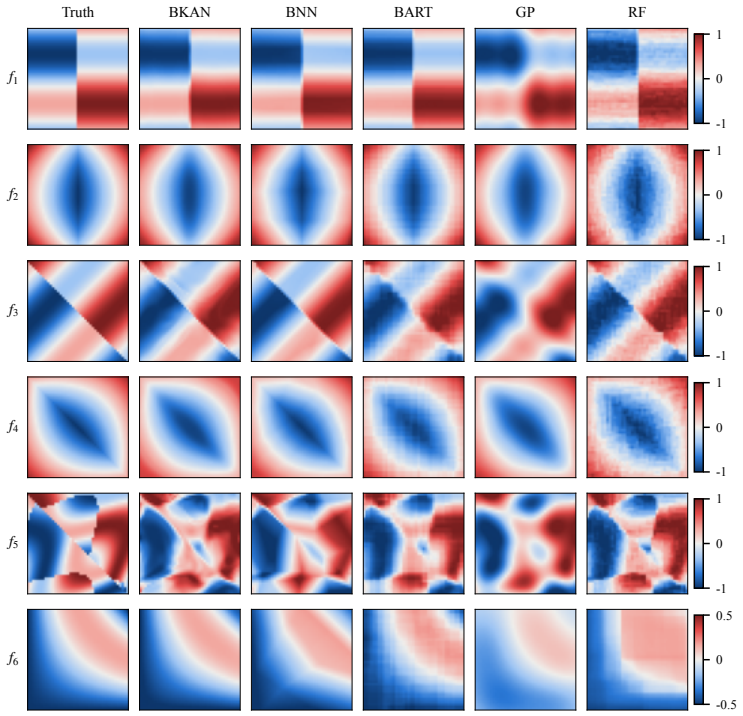



Figure 5: Test RMSE against the sample size $n$ (log–log). Lines are medians over five replications and bands the interquartile range.




(a) $n = 250$




(b) $n = 4{,}000$

Figure 6: Posterior mean surfaces (first replication) next to the true functions (first column of each panel) at (a) $n = 250$ and (b) $n = 4{,}000$. For $f_6$, the surfaces are shown on $(x_1, x_2)$ with $x_3 = \cdots = x_{10} = 1/2$.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's main contribution as a theoretical analysis of posterior contraction rates for sparse Bayesian Kolmogorov–Arnold networks over anisotropic Besov spaces, and accurately reflect the paper's theoretical scope and assumptions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses its limitations in Section 4, Conclusion and Discussion. Specifically, we point out that the theoretical analysis is based on fixed knot placement, and that extending the framework to adaptive or data-driven knot selection is an important direction for future work. We also acknowledge that the scalability of Bayesian KANs remains an important open problem.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper provides the full set of assumptions required for the theoretical results, organized as Assumptions A–D. All theorems, lemmas, and corollaries are stated clearly, and their proofs are either provided or explicitly referenced. The complete proofs and supporting results are included in the supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix D and supplementary material specify the target functions, the data-generating process, the BKAN architecture and priors, and the settings of every method. The data are simulated from fixed seeds, so the code in the supplementary material regenerates them.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplementary material contains the code that generates the data, fits every method, and produces the tables and figures, together with a README. No external data are used; every dataset is simulated from a fixed seed by the code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix D and supplementary material report the data splits, sample sizes, architectures, priors, and hyperparameters of every method, the optimizer and its settings, and how the reported variational restart is chosen. The hyperparameters were fixed in advance and were not tuned on the test data.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Every configuration is replicated five times with independent data and inference seeds. Table 2 reports the mean and the standard deviation (not the standard error) of the RMSE over the replications, Figure 5 shows the median and the interquartile range over the replications, and Figure 3 shows every replication as a point. The variability captured is that of the simulated data and of the inference seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All runs used one Mac mini with an Apple M4 chip (Appendix D). The reported experiments took about 150 CPU-hours in total.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper focuses on mathematical aspects of Bayesian neural networks and does not raise direct societal implications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any data or models. It does not pose risks related to model misuse, dual use, or data privacy.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The experiments use various python packages, which are named in Appendix D. No datasets or pretrained models are used.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The supplementary code is documented in its README, which describes the design of the study, the environment, and the commands that regenerate the data, run every method, and produce the tables and figures.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper exclusively discusses the theory of deep learning and does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This question is not applicable because the paper does not involve crowdsourcing or research with human subjects. The paper focuses exclusively on the theory of deep learning.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The theoretical results and methodological developments in the paper are derived independently.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.