

# Supplementary Material

## A Auxiliary Lemmas

This section contains a collection of results that are needed in the proofs, most of which are classical theorems in statistical learning.

Lemma A.1 uses a high-probability upper bound on the local Rademacher complexity Bartlett et al. [2005] to control the maximal deviation between empirical means and true means of a bounded function, typically a loss function.

**Lemma A.1.** *[(Theorem 2.1 in Bartlett et al. [2005])] Let  $\mathcal{F}$  be a class of functions that map  $\mathcal{X}$  into  $[a, b]$ . Assume that there is some  $r \geq 0$  such that for every  $f \in \mathcal{F}$ ,  $\text{Var}[f(X_i)] \leq r$ . Then, for every  $x > 0$ , with probability at least  $1 - 2e^{-x}$  over the data  $S = [x_1, x_2, \dots, x_n]^T$*

$$\sup_{f \in \mathcal{F}} (Pf - P_n f) \leq \inf_{\alpha \in (0,1)} \left( 2 \frac{1+\alpha}{1-\alpha} \mathbb{E}_\sigma \mathcal{R}_n \mathcal{F} + \sqrt{\frac{2rx}{n}} + (b-a) \left( \frac{1}{3} + \frac{1}{\alpha} + \frac{1+\alpha}{2\alpha(1-\alpha)} \right) \frac{x}{n} \right). \quad (8)$$

We denote that  $Pf = E_{X \sim P}[f(X)]$  and  $P_n f = \frac{1}{n} \sum_{i=1}^n f(x_i)$ . For a class  $\mathcal{F}$ , set  $\mathcal{R}_n \mathcal{F} = \sup_{f \in \mathcal{F}} \mathcal{R}_n f$ . Besides, we represent empirical Rademacher complexity as  $\mathbb{E}_\sigma \mathcal{R}_n \mathcal{F} = \mathbb{E}_\sigma \left[ \frac{1}{n} \sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i f(x_i) \right]$  where  $\sigma_i$  are Rademacher random variables.

**Corollary A.1.** *Let  $\mathcal{F}$  be a class of functions that map  $\mathcal{X}$  into  $[-M, M]$ . Assume that there is some  $r \geq 0$  such that for every  $f \in \mathcal{F}$ ,  $\text{Var}[f(X)] \leq r$ . Then for every  $\epsilon > 0$ , with probability at least  $1 - \epsilon$  over the data  $S = [x_1, x_2, \dots, x_n]^T$*

$$\sup_{f \in \mathcal{F}} (Pf - P_n f) \leq 6\mathbb{E}_\sigma \mathcal{R}_n \mathcal{F} + \sqrt{\frac{2r \log(\frac{2}{\epsilon})}{n}} + \frac{32M \log(\frac{2}{\epsilon})}{3n} \quad (9)$$

*Proof.* It suffices to take  $\alpha = \frac{1}{2}$ ,  $\epsilon = 2e^{-x}$  and  $b - a = 2M$  in Eq. (8) □

Lemma A.2 is a slight variant of the standard Dudley entropy integral bound on the empirical Rademacher complexity.

**Lemma A.2** (Lemma A.5 in Bartlett et al. [2017]). *Let  $\mathcal{F}$  be a real-valued function class taking values in  $[0, M]$  and assume that  $\mathbf{0} \in \mathcal{F}$ . Then we have*

$$\mathcal{R}(\mathcal{F}(\mathbf{X})) \leq \inf_{a>0} \left( \frac{4a}{\sqrt{n}} + \frac{12}{n} \int_a^{M\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{F}(\mathbf{X}), \epsilon, \|\cdot\|_2)} d\epsilon \right). \quad (10)$$

Lemma A.3 is attributed to Maurey, and applied to bounding the covering number generally.

**Lemma A.3.** *In a Hilbert space  $(\mathcal{H}, \|\cdot\|)$ , let  $U \in \mathcal{H}$  be given with the representation  $U = \sum_{l=1}^N a_l W_l$ , where  $W_l \in \mathcal{H}$  and  $a_l \geq 0$ . Then, for any positive integer  $k$ , there exists a choice of nonnegative integers  $(k_1, \dots, k_N)$  such that  $\sum_{i=1}^N k_i = k$  and*

$$\left\| U - \frac{\|\mathbf{a}\|_1}{k} \sum_{l=1}^N k_l V_l \right\|^2 \leq \frac{\|\mathbf{a}\|_1}{k} \sum_{l=1}^N a_l \|V_l\|^2 \leq \frac{\|\mathbf{a}\|_1^2}{k} \max_i \|V_i\|^2, \quad (11)$$

where  $\mathbf{a} = (a_1, \dots, a_N)$  and  $\|\mathbf{a}\|_1 = \sum_{l=1}^N a_l$ .

*Proof.* Denote  $\beta = \|\mathbf{a}\|_1$ . Let  $W_1, \dots, W_k$  be  $k$  i.i.d. r.v. such that  $P(W_1 = \beta V_i) = \frac{a_i}{\beta}$ . Assume that  $W = \sum_{i=1}^k \frac{W_i}{k}$ , then  $EW = EW_1 = \sum_i \beta V_i \cdot \frac{a_i}{\beta} = U$ .

529 On the other hand,

$$\begin{aligned}
E[\|W - U\|^2] &= \frac{1}{k^2} E[\|\sum_i (U - W_i)\|^2] \\
&= \frac{1}{k^2} \left[ E\left[\sum_i \|U - W_i\|^2\right] + E\left[\sum_{i \neq j} \langle U - W_i, U - W_j \rangle\right] \right] \\
&= \frac{1}{k^2} E\left[\sum_i \|U - W_i\|^2\right] = \frac{1}{k} E[\|U - W_i\|^2] = \frac{1}{k} (E[\|W_1\|^2] - \|U\|^2) \quad (12) \\
&\leq \frac{1}{k} E[\|W_1\|^2] = \frac{1}{k} \sum_i \frac{\alpha_i}{\beta} \beta^2 \|V_i\|^2 = \frac{\beta}{k} \sum_i \alpha_i \|V_i\|^2 \\
&\leq \frac{\beta^2}{k} \max_i \|V_i\|^2
\end{aligned}$$

530 Therefore there exists realization  $(j_1, \dots, j_k) \in \{1, \dots, d\}^k$  such that  $\hat{W}_k = \beta V_{j_i}$ ,  $\hat{W} = \frac{\sum \hat{W}_i}{k}$  and  
531  $\|U - \hat{W}\| \leq E[\|W - U\|^2]$ . Finally we conclude our result by taking  $k_i = \sum_i \mathbf{1}_{[j_i=i]}$ .  $\square$

## 532 B Comparison with Different Regularization Techniques

533 In this section we discuss the theoretical implications of different regularization techniques, such  
534 as standard  $L_2$  penalties for ResNet50 He et al. [2016],  $L_1$  penalties for KANs Liu et al. [2025],  
535 dropout for BERT Devlin et al. [2019] and  $L_{1.5}$  penalties in our work.

536  **$L_1$ -regularization** For MLPs,  $L_1$  regularization of linear weights is used to favor sparsity. In  
537 KANs, linear weights are replaced by learnable activation functions, so [Liu et al., 2025] define the  
538  $L_1$  norm of an activation function  $\phi$  to be its average magnitude over its  $N_p$  input, i.e.

$$|\phi|_1 \equiv \frac{1}{N_p} \sum_{s=1}^{N_p} \left| \phi(x^{(s)}) \right|.$$

539 Then for a KAN layer  $\Phi$  with  $n_{\text{in}}$  inputs and  $n_{\text{out}}$  outputs, [Liu et al., 2025] define the  $L_1$  norm of  $\Phi$   
540 to be the sum of  $L_1$  norms of all activation functions, i.e.,

$$|\Phi|_1 \equiv \sum_{i=1}^{n_{\text{in}}} \sum_{j=1}^{n_{\text{out}}} |\phi_{i,j}|_1.$$

541 The total training objective  $\mathcal{L}_{\text{total}}$  is the prediction loss  $\mathcal{L}_{\text{pred}}$  plus  $L_1$  regularization of all KAN  
542 layers:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \lambda \sum_{l=0}^{L-1} |\Phi_l|_1,$$

543 where  $\lambda$  controls overall regularization magnitude. The problem is in the sparsification which is  
544 claimed to be critical to KAN's interpretability. For efficiency, Efficient-KAN<sup>2</sup> instead replaces the  
545  $L_1$  regularization on samples with the  $L_1$  regularization on the weights, which is more common in  
546 neural networks and is compatible with the reformulation:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \lambda \sum_{l=1}^L \|\mathbf{W}_l\|_1.$$

547  **$L_2$ -regularization**  $L_2$  regularization promotes smooth optimization landscapes and improves  
548 generalization by penalizing large weights. The training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \lambda \sum_{l=0}^{L-1} \|\mathbf{W}_l\|_2.$$

---

<sup>2</sup><https://github.com/Blealtan/efficient-kan>

Theoretically,  $L_2$  regularization interacts with the implicit regularization induced by residual connections, further constraining the optimization path to smoother, more stable solutions. This synergy ensures weight shrinkage toward zero while preserving gradient propagation through skip connections Zaeemzadeh et al. [2020], Neyshabur [2017].

**Dropout** Dropout Srivastava et al. [2014] is a popular and effective heuristic for preventing large neural networks from overfitting. Indeed, dropout improves the stability bounds generically Hardt et al. [2016]. From the point of view of stochastic gradient descent, dropout is equivalent to setting a fraction of the gradient weights to zero. In our paper,  $L_{1.5}$ -regularization also improve the bounds.

**$L_{1.5}$ -regularization** The design of  $L_{1.5}$ -loss originates from generalization bounds (Theorem 4.5), which regularize  $\|\mathbf{W}_i^T\|_{2,2,1}$  in  $R_{\mathcal{W}_1^L}$ , leading to an improved bias-variance trade-off (Corollary E.1).

## C Supplementary Definition

### C.1 Norms of Vectors, Matrices and Three-Dimensional Tensors

In this section, we define the norms of vectors (e.g.,  $\mathbf{x}_i$ ), matrices (e.g., the input matrix  $X$ ) and three-dimensional tensors (e.g., the weight tensor  $\mathbf{W}_i$  of KANs). Let  $1 \leq p, r \leq \infty$ . Given a vector  $\alpha = (a_1, a_2, \dots, a_m)^T \in \mathbb{R}^m$ , the  $L_p$  norm of  $\alpha$  is  $\|\alpha\|_p = (\sum_i^m a_i^p)^{\frac{1}{p}}$ . Given a matrix  $A = (\alpha_1, \alpha_2, \dots, \alpha_n) \in \mathbb{R}^{m \times n}$  where the  $i$ -th column of  $A$  is  $\alpha_i \in \mathbb{R}^m$ , we use  $\|A\|_{p,r} = \|(\|\alpha_1\|_p, \|\alpha_2\|_p, \dots, \|\alpha_n\|_p)\|_r$  to represent the  $(p, r)$  norm of matrix  $A$ , denoted as  $\|A\|_{p,r}$ . Especially, the standard  $L_1$  induced norm and  $L_\infty$  induced norm of matrix  $A$  can be represented respectively by  $\|A\|_{1,\infty}$  and  $\|A\|_{\infty,1}$ . Besides, we use  $\|A\|_\sigma$  to represent the standard spectral norm (also called  $L_2$  induced norm) of  $A$ . To define norms for three-dimensional tensors similarly to the matrix  $(p, r)$  norms, we can generalize the structure systematically. Let  $\mathcal{A} \in \mathbb{R}^{m \times n \times k}$  be a three-dimensional tensor. Its elements are denoted by  $\mathcal{A}_{ijk}$ . The tensor  $\mathcal{A}$  can be thought of as a stack of matrices, where the  $k$ -th matrix slice is  $\mathcal{A}_{:,k}$  (i.e., fixing the third index of the tensor). Let  $1 \leq p, r, q \leq \infty$ , we define the  $(p, r, q)$  norm of  $\mathcal{A}$  as:  $\|\mathcal{A}\|_{p,r,q} = \left( \sum_{k=1}^p \|\mathcal{A}_{:,k}\|_{p,r}^q \right)^{1/q}$ , where  $\|\mathcal{A}_{:,k}\|_{p,r} = \left\| \left( \|\mathcal{A}_{:,1k}\|_p, \|\mathcal{A}_{:,2k}\|_p, \dots, \|\mathcal{A}_{:,nk}\|_p \right) \right\|_r$  for each slice  $\mathcal{A}_{:,k} \in \mathbb{R}^{m \times n}$ .

### C.2 Sub-Gaussian Distribution

Our analysis focuses on sub-Gaussian distribution  $\mathbb{P}$ , so we introduce sub-Gaussian random variables and sub-Gaussian random vectors for preparation. Given a random variable  $x$ , if there exists  $K > 0$  such that the tail of  $x$  satisfies  $\mathbb{P}\{|x| \geq t\} \leq 2 \exp(-t^2/K^2)$ ,  $\forall t \geq 0$ , then we call  $x$  a sub-Gaussian random variable, where the quantity  $K^2$  is named the sub-Gaussian variance proxy. The sub-Gaussian norm of  $x$ , denoted  $\|x\|_{\psi_2}$ , is defined as  $\|x\|_{\psi_2} := \inf \{t > 0 : \mathbb{E}[\exp(x^2/t^2)] \leq 2\}$ . It can be deduced by Markov's inequality that there exists  $c > 0$  such that  $\mathbb{P}\{|x| \geq t\} \leq 2 \exp(-ct^2/\|x\|_{\psi_2}^2)$ , indicating that the tail decays slower as the sub-Gaussian norm becomes larger. A random vector  $\mathbf{x}$  in  $\mathbb{R}^n$  is called *sub-Gaussian* if the one-dimensional marginals  $\langle X, x \rangle$  are sub-Gaussian random variables for all  $x \in \mathbb{R}^n$ . The sub-Gaussian norm of  $X$  is defined as  $\|\mathbf{x}\|_{\psi_2} := \sup_{\mathbf{y} \in S^{n-1}} \|\langle \mathbf{x}, \mathbf{y} \rangle\|_{\psi_2}$ . In particular, a random vector with independent bounded coordinates is a sub-Gaussian random vector.

### C.3 Lipschitz Norm

In this section, we discuss Lipschitz properties of various function family.

**Definition C.1** (Lipschitz functions). Let  $(X, d_X)$  and  $(Y, d_Y)$  be metric spaces. A function  $f : X \rightarrow Y$  is called *Lipschitz* if there exists  $L \in \mathbb{R}$ , such that

$$d_Y(f(u), f(v)) \leq L \cdot d_X(u, v), \forall u, v \in X.$$

The infimum of all  $L$  in this definition is called the Lipschitz norm of  $f$  and is denoted  $\|f\|_{\text{Lip}}$ .

591 In other words, Lipschitz functions may not blow up distances between points too much. Lipschitz  
 592 functions with  $\|f\|_{\text{Lip}} \leq 1$  are usually called contractions. Specifically, the layer normalization  
 593  $\tilde{\mathbf{x}} = \tanh(\mathbf{x})$  is a contraction with  $\tilde{x}_i = \frac{e^{x_i} - e^{-x_i}}{e^{x_i} + e^{-x_i}}$ .

594 **Lemma C.1.** *If  $\sigma : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is  $\rho$ -Lipschitz along every coordinate, then it is  $\rho$ -Lipschitz according*  
 595 *to  $\|\cdot\|_p$  for any  $p \geq 1$ .*

596 *Proof.* For any  $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ ,

$$\|\sigma(\mathbf{x}) - \sigma(\mathbf{x}')\|_p = \left( \sum_i |\sigma(\mathbf{x})_i - \sigma(\mathbf{x}')_i|^p \right)^{1/p} \leq \left( \sum_i \rho^p |\mathbf{x}_i - \mathbf{x}'_i|^p \right)^{1/p} = \rho \|\mathbf{x} - \mathbf{x}'\|_p \quad (13)$$

597  $\square$

598 **Lemma C.2.** *Let  $\Sigma = \{\sigma_i : [-1, 1] \rightarrow \mathbb{R} \mid i \in [G]\}$  be a function family. If  $\sigma_i$  is  $\rho_i$ -Lipschitz, then*  
 599  *$M_\Sigma : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times G}$  is  $\left(\sum_{i=1}^G \rho_i^p\right)^{\frac{1}{p}}$ -Lipschitz according to  $\|\cdot\|$  for any  $p \geq 1$ , where the  $(i, j)$ -th*  
 600 *element of the matrix  $M_\Sigma(\mathbf{x})$  is given by  $f_j(x_i)$ .*

601 *Proof.* For any  $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ ,

$$\begin{aligned} \|M_\Sigma(\mathbf{x}) - M_\Sigma(\mathbf{x}')\|_p &= \left( \sum_i \|\sigma_i(\mathbf{x}) - \sigma_i(\mathbf{x}')\|_p^p \right)^{\frac{1}{p}} \leq \left( \sum_i \rho_i^p \|\mathbf{x} - \mathbf{x}'\|_p^p \right)^{\frac{1}{p}} \\ &= \left( \sum_i \rho_i^p \right)^{\frac{1}{p}} \|\mathbf{x} - \mathbf{x}'\|_p = \|\rho\|_p \|\mathbf{x} - \mathbf{x}'\|_p, \end{aligned}$$

602 where  $\rho := (\rho_1, \dots, \rho_G)^T$ .  $\square$

603 Therefore, we define  $\|\Sigma\|_{\text{Lip}} := \|M_\Sigma\|_{\text{Lip}}$  for brevity. We lead to Definition 4.2 when we take  $p = 2$ .

604 Now we prove Lemma E.1.

605 **Lemma E.1.** *If  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  and  $l : \mathbb{R} \rightarrow \mathbb{R}$  are both Lipschitz continuous functions with Lipschitz*  
 606 *norms  $\|\sigma\|_{\text{Lip}}$  and  $\|l\|_{\text{Lip}}$ , respectively, then the composition  $l \circ \sigma(x)$  is Lipschitz continuous with*  
 607 *Lipschitz norm  $\|l \circ \sigma\|_{\text{Lip}} \leq \|l\|_{\text{Lip}} \|\sigma\|_{\text{Lip}}$ .*

608 *Proof.* By the definition of Lipschitz continuity and Lipschitz norm, for any  $x, x' \in \mathbb{R}$ ,

$$\begin{aligned} |l(\sigma(x)) - l(\sigma(x'))| &\leq \|l\|_{\text{Lip}} |\sigma(x) - \sigma(x')| \\ &\leq \|l\|_{\text{Lip}} \|\sigma\|_{\text{Lip}} |x - x'| \end{aligned}$$

609 Thus, the composition  $l \circ \sigma$  satisfies the Lipschitz norm  $\|l \circ \sigma\|_{\text{Lip}} \leq \|l\|_{\text{Lip}} \|\sigma\|_{\text{Lip}}$ .  $\square$

610 Below, we discuss the Lipschitz property of various basis function families.

611 • When  $\Sigma$  is a set of univariate B-spline basis functions with degree  $p$  and knots  $\{\xi_i\}_k$ , we  
 612 have  $\rho_i \leq \frac{2p}{\Delta}$  where  $\Delta = \max_i(\xi_{k+p} - \xi_k)$ . By default of Liu et al. [2025], we set  $p = 3$   
 613 with grid size 5. To ensure the continuity and smoothness of the interpolation, it is usually  
 614 necessary to extend  $p$  points at each end of the original data points. We get the knots  $\xi_k$  as

$$[-2.2, -1.8, -1.4, -1.0, -0.6, -0.2, 0.2, 0.6, 1.0, 1.4, 1.8, 2.2]$$

615 with  $\Delta = 1.2$  and  $\rho_k \leq \frac{2p}{\Delta} = 5$ . See Figure 2 for the visualization of B-spline basis  
 616 functions. Finally, the Lipschitz norm of  $\Sigma$  can be bounded as  $\|\Sigma\|_{\text{Lip}} \leq 10\sqrt{2}$ .

617 • When  $\Sigma = \{\cos(kx), \sin(kx)\}_k$  is a set of Fourier basis functions Xu et al. [2024a], we  
 618 have  $\rho_k \leq k$  based on Lemma C.3. To ensure the consistency of the number of parameters,  
 619 we set  $k \in [4]$  and deduce that  $\|\Sigma\|_{\text{Lip}} \leq 2\sqrt{15}$ .

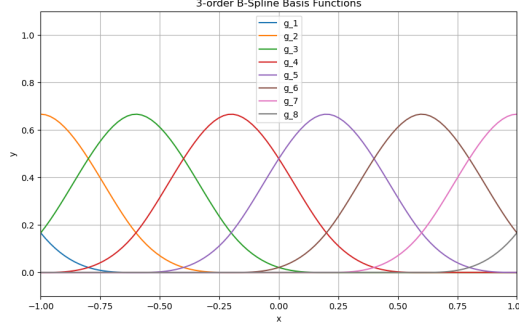


Figure 2: Visualization of the 3-order B-spline basis functions

**Lemma C.3.**  $\cos(kx)$  and  $\sin(kx)$  are  $k$ -Lipschitz continuous.

*Proof.* We only prove the case of the cosine function; the sine function is similar. For any  $x, x' \in \mathbb{R}$ ,

$$\begin{aligned}
 |\cos(kx) - \cos(ky)| &= \left| -2 \sin\left(\frac{kx + ky}{2}\right) \sin\left(\frac{kx - ky}{2}\right) \right| \\
 &= 2 \left| \sin\left(\frac{kx + ky}{2}\right) \right| \left| \sin\left(\frac{kx - ky}{2}\right) \right| \\
 &\leq 2 \left| \sin\left(\frac{kx - ky}{2}\right) \right| \\
 &\leq 2 \left| \frac{kx - ky}{2} \right| \\
 &= k|x - y|
 \end{aligned} \tag{14}$$

where the last inequality is due to the fact that  $\left| \sin\left(\frac{kx - ky}{2}\right) \right| \leq \left| \frac{kx - ky}{2} \right|$ .  $\square$

- When  $\Sigma = \{T_k\}_k$  is a set of Chebyshev polynomialSS et al. [2024], which is defined as  $T_k(x) = \cos(k \arccos(x))$  (please see Figure 3 for the visualization) and can also be expressed using the explicit polynomial form:

$$\begin{aligned}
 T_0(x) &= 1 \\
 T_1(x) &= x \\
 T_2(x) &= 2x^2 - 1 \\
 T_3(x) &= 4x^3 - 3x \\
 T_n(x) &= 2xT_{n-1}(x) - T_{n-2}(x) \text{ for } n \geq 2,
 \end{aligned}$$

we can calculate the Lipschitz norm directly due to its polynomial nature. These polynomials are a sequence of orthogonal functions, so only a smaller number of basis functions are needed. We set  $|\Sigma| = 4$  like SS et al. [2024] and in this case,  $\|\Sigma\|_{\text{Lip}} \leq \sqrt{14}$ .

#### C.4 Einsum

To define the product between tensors, we first introduce the definition of matrix inner product.

The inner product of two matrices  $A, B \in \mathbb{R}^{m \times n}$  is defined as the sum of the products of their corresponding entries. Mathematically, it is given by:

$$\langle A, B \rangle = \sum_{i=1}^m \sum_{j=1}^n A_{ij} B_{ij}.$$

Alternatively, this can be expressed using the trace of the product of  $A$  and  $B^T$

$$\langle A, B \rangle = \text{tr}(A^T B).$$

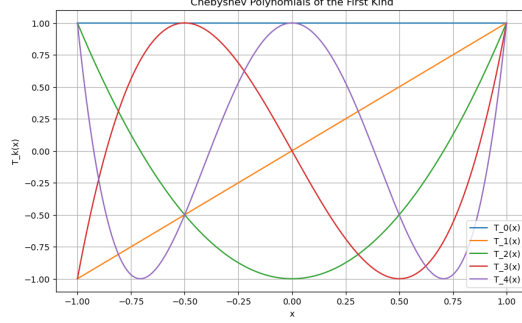


Figure 3: Visualization of the Chebyshev polynomials of the first kind

Specifically, the inner product induces the Frobenius norm:

$$\|A\|_F = \sqrt{\langle A, A \rangle},$$

with the Cauchy-Schwarz inequality stated as

$$|\langle A, B \rangle| \leq \|A\|_F \cdot \|B\|_F. \quad (15)$$

Below we define the Einstein summation operation (“ $\circ$ ” between  $W \in \mathbb{R}^{m,n,l}$  and  $\Sigma \in \mathbb{R}^{n,l}$  as

$$\text{einsum}(W, \Sigma)_i = (W \circ \Sigma)_i := \sum_{j=1}^n \sum_{k=1}^l W_{i,j,k} \Sigma_{j,k}, i = 1, \dots, m.$$

We can also represent  $W \circ X \in \mathbb{R}^m$  as  $(W \circ \Sigma)_i = \langle W_i, X \rangle$  where  $W_i = W_{i::} \in \mathbb{R}^{n,l}$ .

## D Supplementary Experiments

### D.1 Function Approximation

**Datasets** Feynman Udrescu et al. [2020] is a symbolic regression and function approximation dataset of physical equations collected from Feynman’s textbook. We implement a function approximation task using the Feynman dataset, as suggested by [Liu et al., 2025] to investigate whether LipKANs can learn better activation functions compared to KANs.

To demonstrate that LipKANs preserves the powerful expressive capacity of KANs, we first compare the function approximation capabilities of LipKANs with baselines. We utilize the function generation script provided by PyKAN Liu et al. [2025] and assess model performance on the Feynman dataset. We exclude some equations since the function generation script returns NaN in input and inf in label (such as “Jackson 11.38 (Doppler)”).

In Table 2, we present the RMSE of each model on several examples, where the second row of each cell represents the Lip version of the corresponding model. It is clear in Table 2 that the Lip version of KANs not only does not diminish the expressive power but even outperforms in many functions. Among all the baselines, the original KAN achieves the best fitting performance, while Fourier-KAN exhibits the worst fitting performance. Rational-KAN and RBF-KAN perform closely in RMSE. MLP is not comparable to the KAN-based architecture in the experiment, reflecting the strong function approximation ability of KAN-based methods.

## E Proof of Main Results

### E.1 Proof of Lemma 4.2

As outlined in the text, constructing a whole-network cover through induction on layers requires minimal assumptions about the norms imposed on the weight matrices. In this subsection, we delve into a more general analysis of this approach. The structure of the networks is the same as before;

Table 2: RMSE Comparison between KANs (the first row of each cell) and LipKANs (the second row of each cell) on Feynman Dataset

Feynman Eq.	Original Formula	MLP	KAN	Fourier-KAN	Rational-KAN	RBF-KAN	Cheby-KAN
I.6.20a	$\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$	0.04859	0.00047 <b>0.00039</b>	0.02359 0.02884	0.00219 0.00385	0.00082 0.00482	0.04859 0.04761
I.9.18	$\frac{G \cdot m_1 \cdot m_2}{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$	0.42783	0.41967 0.41965	0.54147 0.49766	0.42203 0.42004	0.41933 0.41955	0.41933 0.41954
I.24.6	$\frac{1}{4}m(\omega^2 + \omega_0^2)x^2$	0.04384	0.00877 <b>0.00871</b>	0.10813 0.07405	0.02019 0.03187	0.01101 0.03415	0.04750 0.04821
I.13.4	$\frac{1}{2}m(v^2 + u^2 + w^2)$	0.23311	0.01264 <b>0.01187</b>	0.27906 0.17017	0.01203 0.07661	0.05021 0.09352	0.26101 0.23197
II.35.18	$\frac{n_0}{\exp\left(\frac{\mu_B}{k_B T}\right) + \exp\left(-\frac{\mu_B}{k_B T}\right)}$	0.13784	0.06437 0.06220	0.15890 0.14590	0.04542 0.08346	<b>0.03442</b> 0.07837	0.17830 0.17846
III.4.33	$\frac{\hbar\omega}{\exp\left(\frac{\hbar\omega}{k_B T}\right) - 1}$	0.27705	<b>0.18712</b> 0.18774	0.42472 0.29129	0.21685 0.20695	0.19196 0.21895	0.33203 0.35953

namely, given tensors  $\mathcal{W} = (\mathbf{W}_1, \dots, \mathbf{W}_L)$ , define the mapping  $F_{\mathcal{W}}$  as  $F_{\mathcal{W}}(\mathbf{x}) = \text{KAN}_{\mathcal{W}}(\mathbf{x})$  in Eq. (1). More generally for  $i \leq L$  define  $\mathcal{W}_1^i := (\mathbf{W}_1, \dots, \mathbf{W}_i)$  and

$$F_{\mathcal{W}_1^i}(\mathbf{x}) = (\mathbf{W}_i \circ \Sigma \circ \dots \circ \mathbf{W}_1 \circ \Sigma)\mathbf{x}$$

with the convention  $F_{\emptyset}(\mathbf{x}) = \mathbf{x}$ . For brevity, we write  $F_{\mathcal{W}_1^i}(\mathbf{x})$  as  $F_{\mathcal{W}_1^i}$  when data  $\mathbf{x}$  is fixed.

**Lemma 4.2** Let  $(\epsilon_1, \dots, \epsilon_L)$  be given. Under Assumption 1 and 2, each  $\sigma_k$  is Lipschitz continuous with  $\rho_k := \|\sigma_k\|_{\text{Lip}}$ . Let  $\rho := (\rho_1, \dots, \rho_G)$ . Suppose the tensors  $\mathcal{W}_1^L = (\mathbf{W}_1, \dots, \mathbf{W}_L)$  lie within  $\mathcal{B}_1 \times \dots \times \mathcal{B}_L$  where  $\mathcal{B}_i$  are arbitrary classes with the property that each  $\mathbf{W}_i \in \mathcal{B}_i$  has  $\|\mathbf{W}_i\|_{\sigma} \leq c_i$ . Then, letting  $\tau_1 = \epsilon_1$  and  $\tau_l = \sum_{j=1}^l (\prod_{j=i+1}^l c_j \|\rho\|_2) \epsilon_i$ , the neural net images  $\mathcal{H}_X := \{F_{\mathcal{W}_1^L}(X) : \mathcal{W}_1^L \in \mathcal{B}_1 \times \dots \times \mathcal{B}_L\}$  have a covering number bound:

$$\mathcal{N}(\mathcal{H}_X, \tau_L, \|\cdot\|_2) \leq \prod_{i=1}^L \sup_{\substack{(\mathbf{W}_1, \dots, \mathbf{W}_{i-1}) \\ \forall j < i, \mathbf{W}_j \in \mathcal{B}_j}} \mathcal{N}\left(\left\{\mathbf{W}_i \circ \Sigma\left(F_{\mathcal{W}_1^{i-1}}(X)\right) : \mathbf{W}_i \in \mathcal{B}_i\right\}, \epsilon_i, \|\cdot\|_2\right).$$

*Proof.* Inductively construct covers  $\mathcal{F}_1, \dots, \mathcal{F}_L$  of  $\mathcal{W}_2, \dots, \mathcal{W}_{L+1}$  as follows,

- Choose an proper  $\epsilon_1$ -cover  $\mathcal{F}_{\mathcal{W}_1^1}$  of  $\{F_{\mathcal{W}_1^1}, \mathbf{W}_1 \in \mathcal{B}_1\}$ , thus

$$|\mathcal{F}_1| = \mathcal{N}(\{F_{\mathcal{W}_1^1}, \mathbf{W}_1 \in \mathcal{B}_1\}, \epsilon_1, \|\cdot\|_2) =: N_1$$

- For every element  $F \in \mathcal{F}_1$ , construct an  $\epsilon_{i+1}$ -cover  $\mathcal{G}_{i+1}(F)$  of

$$\{\mathbf{W}_{i+1} \circ \Sigma(F) : \mathbf{W}_{i+1} \in \mathcal{B}_{i+1}\}.$$

The covers are proper, so  $F = F_{\mathcal{W}_1^{i-1}}$  for some  $(\mathbf{W}_1, \dots, \mathbf{W}_i) \in \mathcal{B}_1 \times \dots \times \mathcal{B}_i$ . It follows that

$$|\mathcal{G}_{i+1}(F)| \leq \sup_{\substack{(\mathbf{W}_1, \dots, \mathbf{W}_i) \\ \forall j \leq i, \mathbf{W}_j \in \mathcal{B}_j}} \mathcal{N}\left(\left\{\mathbf{W}_{i+1} F_{\mathcal{W}_1^i}(Z) : \mathbf{W}_{i+1} \in \mathcal{B}_{i+1}\right\}, \epsilon_{i+1}, \|\cdot\|_{i+2}\right) =: N_{i+1}$$

Construct the cover

$$\mathcal{F}_{i+1} := \cup_{F \in \mathcal{F}_i} \mathcal{G}_{i+1}(F), \quad (16)$$

with cardinality

$$|\mathcal{F}_{i+1}| \leq |\mathcal{F}_i| \cdot N_{i+1} \leq \prod_{l=1}^{i+1} N_l.$$

Below, we show that  $\mathcal{F}_{i+1}$  is an  $\tau_{i+1}$ -cover of  $\{F_{\mathcal{W}_1^{i+1}} : (\mathbf{W}_1 \times \dots \times \mathbf{W}_{i+1}) \in \mathcal{B}_1 \times \dots \times \mathcal{B}_{i+1}\}$ .

For any  $F_{\mathcal{W}_1^{i+1}} = (\mathbf{W}_{i+1} \circ \Sigma \circ \dots \circ \mathbf{W}_1 \circ \Sigma)\mathbf{x}$ , by construction, we can find  $\hat{F}_1 \in \mathcal{F}_1$  such that

$$\|\hat{F}_1 - F_{\mathcal{W}_1^1}\| \leq \epsilon_1 = \tau_1.$$

Now suppose we can find  $\hat{F}_i \in \mathcal{F}_i$  such that

$$\|\hat{F}_i - F_{\mathcal{W}_1^i}\| \leq \tau_i$$

By the construction of  $\mathcal{F}_{i+1}$  in Eq. (16), we can find  $\hat{F}_{i+1} \in \mathcal{F}_{i+1}$  such that

$$\|\hat{F}_{i+1} - \mathbf{W}_{i+1} \circ \Sigma(\hat{F}_i)\|_2 \leq \epsilon_{i+1}$$

Then we use the triangle inequality of norm  $\|\cdot\|_2$ , we have

$$\begin{aligned} \|\hat{F}_{i+1} - F_{\mathcal{W}_1^{i+1}}\|_2 &= \|\hat{F}_{i+1} - \mathbf{W}_{i+1} \circ \Sigma(\hat{F}_i) + \mathbf{W}_{i+1} \circ \Sigma(\hat{F}_i) - F_{\mathcal{W}_1^{i+1}}\|_2 \\ &\leq \|\hat{F}_{i+1} - \mathbf{W}_{i+1} \circ \Sigma(\hat{F}_i)\|_2 + \|\mathbf{W}_{i+1} \circ \Sigma(\hat{F}_i) - F_{\mathcal{W}_1^{i+1}}\|_2 \\ &\leq \epsilon_{i+1} + \|\mathbf{W}_{i+1}\|_\sigma \|\Sigma(\hat{F}_i) - F_{\mathcal{W}_1^i}\|_2 \\ &\leq \epsilon_{i+1} + c_{i+1} \|\rho\|_2 \tau_i = \tau_{i+1} \end{aligned}$$

where the second  $\leq$  is due to Cauchy-Schwarz inequality in Eq. (15).  $\square$

## E.2 Proof of Lemma 4.3

The proof relies upon the Lemma A.3, which is stated in terms of sparsifying convex hulls, and in its use here is inspired by covering number bounds for linear predictors.

**Lemma 4.3** Let conjugate exponents  $(p, q)$  and  $(u, v)$  be given with  $p \leq 2$  as well as positive reals  $(a, b, \epsilon)$  and positive integer  $m$ . Let data  $X$  be given with  $\|\Sigma(X)\|_p \leq b$ . Then

$$\ln \mathcal{N} \left( \left\{ \mathbf{W} \circ \Sigma(X) : \mathbf{W} \in \mathbb{R}^{d' \times d \times G}, \|\mathbf{W}\|_{q,q,v} \leq a \right\}, \epsilon, \|\cdot\|_2 \right) \leq \left\lceil \frac{a^2 b^2 d'^{\frac{2}{u}}}{\epsilon^2} \right\rceil \ln(2dd'G),$$

where  $\lceil x \rceil$  returns the smallest integer greater than or equal to  $x$ .

*Proof.* Let data  $X \in \mathbb{R}^{n \times d}$  be given, with weight tensor  $\mathbf{W} \in \mathbb{R}^{d \times G \times d'}$ . For brevity, we write tensor

$$\hat{\Sigma} := \Sigma(X) \in \mathbb{R}^{n \times d \times G} \text{ with } \hat{\Sigma}_{ijk} = [g_k(\mathbf{x}_i)]_j \text{ and } \mathbf{W} \circ \Sigma(\mathbf{x}) \in \mathbb{R}^{n \times d'}. \text{ Set } k := \left\lceil \frac{a^2 b^2 d'^{\frac{2}{u}}}{\epsilon^2} \right\rceil.$$

We obtain tensor  $\bar{\Sigma} \in \mathbb{R}^{n \times d \times G}$  by rescaling the first dimension of tensor  $\hat{\Sigma}$  to have unit  $p$ -norm:

$$\bar{\Sigma}_{:jk} := \frac{\hat{\Sigma}_{:jk}}{\|\hat{\Sigma}_{:jk}\|_p}.$$

Define  $\alpha \in \mathbb{R}^{d \times G \times d'}$  to be a rescaling tensor such that for each  $i \in [d']$ ,

$$\alpha_{::i} = \begin{pmatrix} \|\Sigma_{:1,1}\|_p & \|\Sigma_{:1,2}\|_p & \cdots & \|\Sigma_{:1,G}\|_p \\ \|\Sigma_{:2,1}\|_p & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \|\Sigma_{:d,1}\|_p & \cdots & \cdots & \|\Sigma_{:d,G}\|_p \end{pmatrix}$$

which satisfies  $\mathbf{W} \circ \tilde{\Sigma} = (\alpha \odot \mathbf{W}) \circ \bar{\Sigma}$  where  $\odot$  denotes element-wise product. Note that

$$\begin{aligned} \|\alpha\|_{p,p,u} &= \left( \sum_{i=1}^{d'} \|\alpha_{::i}\|_{p,p}^u \right)^{\frac{1}{u}} \\ &= (d')^{\frac{1}{u}} \|\alpha_{::i}\|_{p,p} \\ &= (d')^{\frac{1}{u}} \left( \sum_j^d \sum_k^G \|\Sigma_{:jk}\|_p^p \right)^{\frac{1}{p}} \\ &= (d')^{\frac{1}{u}} \left( \sum_i^n \sum_j^d \sum_k^G \|\Sigma_{ijk}\|_p^p \right)^{\frac{1}{p}} \\ &= (d')^{\frac{1}{u}} \|\tilde{\Sigma}\|_p \end{aligned}$$

696 Define  $\bar{\mathbf{W}} = \alpha \odot \mathbf{W}$ , whereby using conjugacy of  $\|\cdot\|_{p,r,u}$  and  $\|\cdot\|_{q,s,v}$  gives

$$\|\bar{\mathbf{W}}\|_1 \leq \langle \alpha, |\mathbf{W}| \rangle \leq \|\alpha\|_{p,p,u} \|\mathbf{W}\|_{q,q,v} \leq (d')^{\frac{1}{u}} \|\Sigma\|_p a =: \bar{a}$$

697 Now we have

$$\begin{aligned} \mathbf{W} \circ \tilde{\Sigma} &= \bar{\mathbf{W}} \circ \bar{\Sigma} = \left( \sum_i^d \sum_j^G \sum_k^{d'} \bar{\mathbf{W}}_{ijk} E_{ijk} \right) \circ \bar{\Sigma} \\ &= \|\bar{\mathbf{W}}\|_1 \left( \sum_i^d \sum_j^G \sum_k^{d'} \frac{\bar{\mathbf{W}}_{ijk}}{\|\bar{\mathbf{W}}\|_1} E_{ijk} \circ \bar{\Sigma} \right) \end{aligned} \quad (17)$$

698 where  $E_{ijk} \in \mathbb{R}^{d \times G \times d'}$  is a tensor where the element at position  $(i, j, k)$  is 1, and all other elements  
699 are 0.

700 To derive the desired cover, we define

$$\{V_1, \dots, V_N\} = \{g E_{ijk} \circ \bar{\Sigma} \mid g \in \{-1, +1\}, i \in [d], j \in [G], k \in [d']\},$$

701 and construct a cover

$$\mathcal{C} := \left\{ \frac{\bar{a}}{k} \sum_{i=1}^N k_i V_i : k_i \geq 0, \sum_{i=1}^N k_i = k \right\} = \left\{ \frac{\bar{a}}{k} \sum_{j=1}^k V_{i_j} : (i_1, \dots, i_k) \in [N]^k \right\},$$

702 where  $k_i$ 's are integers and  $N := 2dd'G$ . By construction, we have  $|\mathcal{C}| \leq N^k$ , and

$$\max_i \|V_i\|_2 \leq \max_{j,k} \frac{\|\Sigma_{:jk}\|_p}{\|\Sigma_{:jk}\|_2} \leq 1,$$

703 where the last inequality is due to  $p \leq 2$ .

704 To use Lemma A.3, We can view  $V_i$ 's as elements in the Hilbert space  $\mathbb{R}^{n \times d'}$  equipped with the  
705 inner product  $\langle A, \tilde{A} \rangle = \text{trace} \left( A^\top \tilde{A} \right)$  defined in Appendix C.4 and norm  $\|A\|_2$ . Following Eq. (17),  
706  $\mathbf{W} \circ \tilde{\Sigma}$  lies in the convex hull of  $\{V_1, \dots, V_N\}$ , i.e.

$$\mathbf{W} \circ \tilde{\Sigma} = \bar{\mathbf{W}} \circ \bar{\Sigma} \in \bar{a} \cdot \text{conv} \{V_1, \dots, V_N\}$$

707 Combining the preceding constructions with Lemma A.3, there exist nonnegative integers  
708  $(k_1, \dots, k_N)$  with  $\sum_i k_i = k$  such that

$$\|\mathbf{W} \circ \tilde{\Sigma} - \frac{\bar{a}}{k} \sum_{i=1}^N k_i V_i\|_2^2 = \|\bar{\mathbf{W}} \circ \bar{\Sigma} - \frac{\bar{a}}{k} \sum_{i=1}^N k_i V_i\|_2^2 \leq \frac{\bar{a}^2}{k} \max_i \|V_i\|_2 \leq \frac{a^2 d'^{\frac{2}{u}} \|\Sigma\|_p^2}{k} \leq \epsilon^2.$$

709 The desired cover element is thus  $\frac{\bar{a}}{k} \sum_i k_i V_i \in \mathcal{C}$  and the result follows.  $\square$

### 710 E.3 Proof of Theorem 4.4

711 The whole-network covering bound now follows by the general norm covering number in Lemma 4.2,  
712 and the matrix covering lemma in Lemma 4.3.

713 **Theorem 4.4** Under Assumption 2, we can bound

$$\log \mathcal{N}(\mathcal{H}_X, \epsilon, \|\cdot\|_2) \leq \frac{R_{\mathcal{W}_1^L}^2 \log(2\tilde{d}^2 G)}{\epsilon^2},$$

714 where  $\tilde{d} := \max_i d_i$ .

715 *Proof.* Given tensors  $\mathcal{W}_1^{i-1} := (\mathbf{W}_1, \dots, \mathbf{W}_{i-1})$ , we define

$$\mathbf{X}^{l-1} := F_{\mathcal{W}_1^{i-1}}(\mathbf{X}) = (\mathbf{W}_{i-1} \circ \Sigma \circ \dots \circ \mathbf{W}_1 \circ \Sigma) \mathbf{X}$$

716 Set  $p = q = 2, u = \infty, v = 1$  in Lemma 4.3, define  $\mathcal{B}_i = \{\mathbf{W} \in \mathbb{R}^{d_i \times d_{i-1} \times G} : \|\mathbf{W}\|_\sigma \leq$   
 717  $c_i, \|\mathbf{W}\|_{2,2,1} \leq a_i\}$ , then we can decompose the whole covering number based on Lemma 4.2.

$$\begin{aligned}
 & \log \mathcal{N}(\mathcal{H}_{\mathbf{x}}, \tau_L, \|\cdot\|_2) \\
 & \leq \sum_{i=1}^L \sup_{\substack{(\mathbf{W}_1, \dots, \mathbf{W}_{i-1}) \\ \forall j < i, \mathbf{W}_j \in \mathcal{B}_j}} \log \mathcal{N}\left(\left\{\mathbf{W}_i \circ \left(\Sigma F_{\mathcal{W}_1^{i-1}}(\mathbf{x})\right) : \mathbf{W}_i \in \mathcal{B}_i\right\}, \epsilon_i, \|\cdot\|_2\right) \\
 & = \sum_{i=1}^L \sup_{\substack{(\mathbf{W}_1, \dots, \mathbf{W}_{i-1}) \\ \forall j < i, \mathbf{W}_j \in \mathcal{B}_j}} \log \mathcal{N}\left(\left\{\mathbf{W}_i \circ \left(\Sigma F_{\mathcal{W}_1^{i-1}}(\mathbf{x})\right) - \mathbf{W}_i \circ \Sigma(\mathbf{0}) : \mathbf{W}_i \in \mathcal{B}_i\right\}, \epsilon_i, \|\cdot\|_2\right) \\
 & = \sum_{i=1}^L \sup_{\substack{(\mathbf{W}_1, \dots, \mathbf{W}_{i-1}) \\ \forall j < i, \mathbf{W}_j \in \mathcal{B}_j}} \log \mathcal{N}\left(\left\{\mathbf{W}_i \circ (\Sigma(\mathbf{x}^{l-1}) - \Sigma(\mathbf{0})) : \mathbf{W}_i \in \mathcal{B}_i\right\}, \epsilon_i, \|\cdot\|_2\right) \\
 & \leq \sum_{i=1}^L \frac{a_i^2 \|\Sigma(\mathbf{x}^{l-1}) - \Sigma(\mathbf{0})\|_2^2}{\epsilon_i^2} \log(2d_i d_{i-1} G)
 \end{aligned} \tag{18}$$

718 Below, we use the Lipschitz property of  $\Sigma$  to bound the term  $\|\Sigma(\mathbf{x}^{l-1}) - \Sigma(\mathbf{0})\|_2$  as follows:

$$\begin{aligned}
 \|\Sigma(\mathbf{x}^{l-1}) - \Sigma(\mathbf{0})\|_2 & \leq \|\Sigma\|_{\text{Lip}} \|\mathbf{x}^{l-1}\|_2 \\
 & \leq \|\Sigma\|_{\text{Lip}} \|\mathbf{W}_{l-1} \circ \Sigma(\mathbf{x}^{l-2})\|_2 \\
 & \leq \|\Sigma\|_{\text{Lip}} \|\mathbf{W}_{l-1}\|_\sigma \|\Sigma(\mathbf{x}^{l-2})\|_2 \\
 & \leq \|\Sigma\|_{\text{Lip}} \|\mathbf{W}_{l-1}\|_\sigma \|\Sigma(\mathbf{x}^{l-2}) - \Sigma(\mathbf{0}) + \Sigma(\mathbf{0})\|_2 \\
 & \leq \|\Sigma\|_{\text{Lip}} \|\mathbf{W}_{l-1}\|_\sigma (\|\Sigma(\mathbf{x}^{l-2}) - \Sigma(\mathbf{0})\|_2 + \|\Sigma(\mathbf{0})\|_2) \\
 & \leq \|\Sigma\|_{\text{Lip}} \|\mathbf{W}_{l-1}\|_\sigma (\|\Sigma\|_{\text{Lip}} \|\mathbf{x}^{l-2}\|_2 + \|\Sigma\|_\infty) \\
 & \leq \left( \|\Sigma\|_{\text{Lip}}^l \|\mathbf{X}\|_2 \prod_{i=1}^{l-1} \|\mathbf{W}_i\|_\sigma \right) + \left( \sum_{j=1}^{l-1} \|\Sigma\|_{\text{Lip}}^j \|\Sigma\|_\infty \left( \prod_{i=l-j}^{l-1} \|\mathbf{W}_i\|_\sigma \right) \right) \\
 & \leq \left( C^l D \prod_{i=1}^{l-1} c_i \right) + \left( \sum_{j=1}^{l-1} C^j E \left( \prod_{i=l-j}^{l-1} c_i \right) \right)
 \end{aligned}$$

719 where  $\|\Sigma\|_\infty := \sqrt{\sum_k^G \|g_k\|_\infty^2}$  represents the  $\infty$  norm of function family  $\Sigma$ . Now we suppose  
 720  $\|\mathbf{W}_i\|_\sigma \leq \sigma_i$  and  $\|\Sigma\|_\infty \leq E$ , then we have

$$\|\Sigma(\mathbf{x}^{l-1}) - \Sigma(\mathbf{0})\|_2 \leq \left( C^l D \prod_{i=1}^{l-1} c_i \right) + \left( \sum_{j=1}^{l-1} C^j E \left( \prod_{i=l-j}^{l-1} c_i \right) \right) \tag{19}$$

721 Plugging Eq. (19) into Eq. (18), we obtain that

$$\begin{aligned}
 \log \mathcal{N}(\mathcal{H}_{\mathbf{x}}, \tau_L, \|\cdot\|_2) & \leq \sum_{i=1}^L \frac{a_i^2 \|\Sigma(\mathbf{x}^{l-1}) - \Sigma(\mathbf{0})\|_2^2}{\epsilon_i^2} \log(2d_i d_{i-1} G) \\
 & \leq \sum_{i=1}^L \frac{a_i^2 \left( C^l D \prod_{j=1}^{i-1} c_j + \sum_{j=1}^{i-1} C^j E \left( \prod_{k=i-j}^{i-1} c_k \right) \right)^2}{\epsilon_i^2} \log(2d^2 G)
 \end{aligned} \tag{20}$$

722 Define

$$\alpha_i = a_i^{\frac{2}{3}} \left( \prod_{j=i+1}^L C c_j \right)^{\frac{2}{3}} \left( C^l D \prod_{j=1}^{i-1} c_j + \sum_{j=1}^{i-1} C^j E \left( \prod_{k=i-j}^{i-1} c_k \right) \right)^{\frac{2}{3}}$$

723 and  $\tilde{\alpha} = \sum_{i=1}^L \alpha_i$ . For any  $\epsilon > 0$ , set

$$\epsilon_i = \frac{\alpha_i \epsilon}{\tilde{\alpha} \prod_{j=i+1}^L C c_j}$$

724 Then we have  $s_L = \sum_{i=1}^L \left( \prod_{j=i+1}^L C c_j \right) \epsilon_i = \epsilon$  and hence

$$\log \mathcal{N}(\mathcal{H}_{\mathbf{x}}, \epsilon, \|\cdot\|_2) \leq \frac{\tilde{\alpha}^3 \log(2\tilde{d}^2 G)}{\epsilon^2}$$

725

□

#### 726 E.4 Proof of Theorem 4.5

727 We prove our main result by integrating Theorem 4.4 with standard properties of Rademacher  
728 complexity in Section A.

729 **Theorem 4.5** (Generalization Bounds for KANs) Under Assumption 2 and 3, let fixed Lipschitz  
730 activations  $\Sigma = \{\sigma_k \mid k \in [G]\}$  and weight tensors  $\mathcal{W}_1^L = \{\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_L\}$  be given. Then for  
731  $(\mathbf{x}, y), (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$  drawn iid from distribution  $\mathbb{P}$ , we have with probability greater than  
732  $1 - \epsilon$ ,

$$\begin{aligned} R(F_{\mathcal{W}_1^L}) - \hat{R}(F_{\mathcal{W}_1^L}) &\leq \frac{144\sqrt{\zeta}\{\log(nM/(3\sqrt{\zeta})) \vee 1\}}{n} + \sqrt{\frac{4M^2 \log(2/\epsilon)}{n}} + \frac{32M \log(2/\epsilon)}{3n} \\ &\leq \mathcal{O}\left(\frac{\|X\|_2 R_{\mathcal{W}_1^L} \max_i |\rho(y_i)|}{n} \log(\tilde{d}^2 G) + \sqrt{\frac{1/\epsilon}{n}}\right), \end{aligned}$$

733 where  $\zeta = R_{\mathcal{W}_1^L}^2 \|X\|_2^2 \log(2\tilde{d}^2 G) \max_i \rho^2(y_i)$ .

734 *Proof.* Define the loss class  $\mathcal{L}(S) := \{(\mathcal{L}(f(\mathbf{x}_1), y_1), \dots, \mathcal{L}(f(\mathbf{x}_n), y_n)) \mid f \in \mathcal{H}\}$ . By theorem 4.4,  
735 we have

$$\log \mathcal{N}(\mathcal{L}(S), \epsilon, \|\cdot\|_2) \leq \frac{\tilde{\alpha}^3 \log(2\tilde{d}^2 G) \max_i \rho^2(y_i)}{\epsilon^2} = \frac{\zeta}{\epsilon^2}$$

736 Lemma A.2 implies that

$$\begin{aligned} \mathcal{R}(\mathcal{L}(S)) &\leq \inf_{a>0} \left( \frac{4a}{\sqrt{n}} + \frac{12}{n} \int_a^{\sqrt{n}M} \sqrt{\frac{\zeta}{\epsilon^2}} d\epsilon \right) \\ &= \inf_{a>0} \left( \frac{4a}{\sqrt{n}} + \frac{12\sqrt{\zeta}}{n} \log(M\sqrt{n}/a) \right) \\ &\leq \frac{12\sqrt{\zeta}}{n} + \frac{12\sqrt{\zeta} \log(nM/(3\sqrt{\zeta}))}{n} \\ &\leq \frac{24\sqrt{\zeta}\{\log(nM/(3\sqrt{\zeta})) \vee 1\}}{n}. \end{aligned}$$

737 Using Lemma A.1 with  $r = 2M^2$ , we have with probability greater than  $1 - \epsilon$ ,

$$\begin{aligned} \mathbb{E}[\mathcal{L}(f(x), y)] &\leq \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f(\mathbf{x}_i), y_i) + \frac{144\sqrt{\zeta}\{\log(nM/(3\sqrt{\zeta})) \vee 1\}}{n} \\ &\quad + \sqrt{\frac{4M^2 \log(2/\epsilon)}{n}} + \frac{32M \log(2/\epsilon)}{3n} \\ &= \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f(\mathbf{x}_i), y_i) + \mathcal{O}\left(\frac{\tilde{\alpha}^{\frac{3}{2}} \sqrt{\ln(\tilde{d}^2 G)}}{n}\right) \end{aligned}$$

738 for any  $f \in \mathcal{H}$ , which completes our proof. □

#### 739 E.5 Other Theoretical Results

740 Based on Assumption 1, which states that the distribution  $\mathbb{P}$  satisfies the sub-Gaussian property, we  
741 can bound  $\|X\|_2$  in Eq. (6) with high probability, leading to more applicable bounds in Corollary E.1.

742 **Corollary E.1** (Sub-Gaussian Generalization Bounds for KANs). *Under Assumption 1, 2 and 3, let*  
 743 *fixed Lipschitz activations  $\Sigma = \{\sigma_k : k \in [G]\}$  and weight tensor  $\mathcal{W}_1^L = \{\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_L\}$  be*  
 744 *given. Then for  $(\mathbf{x}, y), (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$  drawn iid from sub-Gaussian distribution  $\mathbb{P}$ , we have*  
 745 *with probability greater than  $1 - \epsilon - \delta$ ,*

$$R(F_{\mathcal{W}_1^L}) - \hat{R}(F_{\mathcal{W}_1^L}) \leq \mathcal{O} \left( \frac{K R_{\mathcal{W}_1^L} \max_i |\rho(y_i)|}{\sqrt{n}} \log(\tilde{d}^2 G) \sqrt{\tilde{d} \log \left( \frac{1}{\delta} \right)} + \sqrt{\frac{1/\epsilon}{n}} \right), \quad (21)$$

746 where  $K$  was defined in Assumption 1.

747 *Proof.* We conclude the result immediately by plugging Eq. (2) into Eq. (6).  $\square$

748 We introduce Lemma E.1 that ensures the Lipschitz property of the composition of Lipschitz functions.

749 **Lemma E.1.** *If  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  and  $l : \mathbb{R} \rightarrow \mathbb{R}$  are both Lipschitz continuous functions with Lipschitz*  
 750 *norms  $\|\sigma\|_{\text{Lip}}$  and  $\|l\|_{\text{Lip}}$ , respectively, then the composition  $l \circ \sigma(x)$  is Lipschitz continuous with*  
 751 *Lipschitz norm  $\|l \circ \sigma\|_{\text{Lip}} \leq \|l\|_{\text{Lip}} \|\sigma\|_{\text{Lip}}$ .*

752 The proof of Lemma E.1 is provided in Appendix C.3. Lemma E.1 demonstrates that composing with  
 753 a univariate function having a small Lipschitz norm reduces the Lipschitz norm of the activations  
 754 family, thereby reducing the complexity of the functions represented by LipKANs.

755 **Corollary 5.1** (Generalization Bounds for LipKANs) Under Assumption 1, 2, and 3, let fixed  
 756 Lipschitz activations  $\Sigma = \{\sigma_k : k \in [G]\}$ , weight tensor  $\mathcal{W}_1^L = \{\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_L\}$ , and  
 757 Lipschitz function  $l : \mathbb{R} \rightarrow \mathbb{R}$  with  $\|l\|_{\text{Lip}} \leq 1$ . Then for  $(\mathbf{x}, y), (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$  drawn i.i.d.  
 758 from sub-Gaussian distribution  $\mathbb{P}$ , we have with probability greater than  $1 - \epsilon - \delta$ ,

$$R(F_{\mathcal{W}_1^L, l}) - \hat{R}(F_{\mathcal{W}_1^L, l}) \leq \mathcal{O} \left( \frac{K R_{\mathcal{W}_1^L, l} \max_i |\rho(y_i)|}{\sqrt{n}} \log(\tilde{d}^2 G) \sqrt{\tilde{d} \log \left( \frac{1}{\delta} \right)} + \sqrt{\frac{1/\epsilon}{n}} \right),$$

759 where the right-hand side of the inequality is also bounded by the upper bound in Eq. (21).

760 *Proof.* By Lemma E.1,

$$\begin{aligned} R_{\mathcal{W}_1^L, l} &= \left( \|\ell \circ \Sigma\|_{\text{Lip}}^L \prod_{i=1}^L \|\mathbf{W}_i\|_{\sigma} \right) \left( \sum_{i=1}^L \|\mathbf{W}_i^T\|_{2,2,1}^{2/3} \right)^{\frac{3}{2}} \\ &\leq \left( \|\Sigma\|_{\text{Lip}}^L \prod_{i=1}^L \|\mathbf{W}_i\|_{\sigma} \right) \left( \sum_{i=1}^L \|\mathbf{W}_i^T\|_{2,2,1}^{2/3} \right)^{\frac{3}{2}}, \end{aligned}$$

761 where the right hand side of the inequality represents the complexity of KANs, thereby proving the  
 762 conclusion.  $\square$