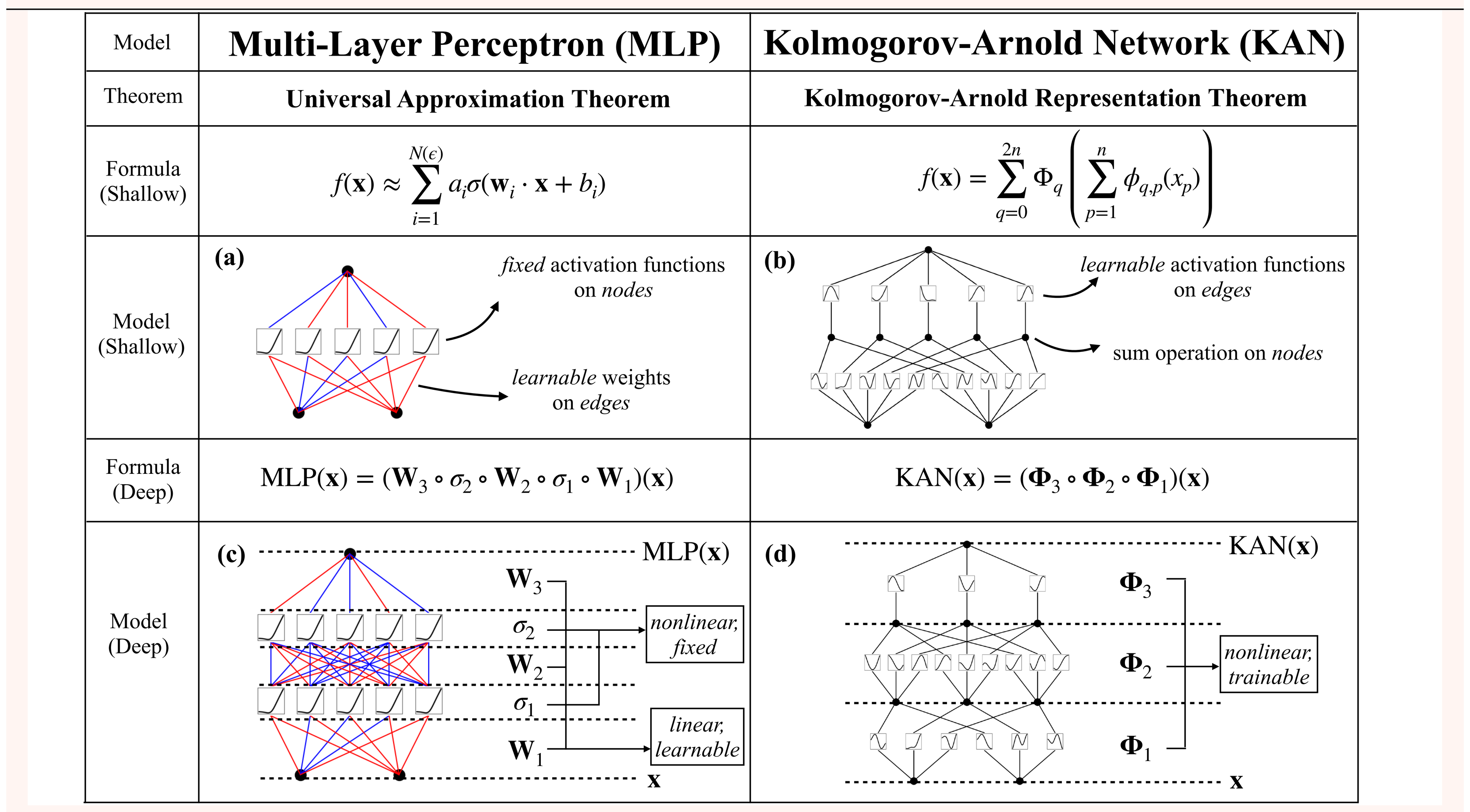
On the expressiveness and spectral bias of KANs

Yixuan Wang[1] Jonathan W. Siegel[2] Ziming Liu[3] Thomas Y. Hou[1]
[1] Caltech [2]MIT [3]Texas A&M

ICLR April 24-28, 2025

Kolmogorov–Arnold Networks (KANs) [2]



For a continuous $f : [0,1]^n \to \mathbb{R}$

$$f(\mathbf{x}) = f(x_1, \cdots, x_n) = \sum_{q=0}^{2n} \Phi_q\Big(\sum_{p=1}^{n} \phi_{q,p}(x_p)\Big). \tag{1}$$

where $\phi_{q,p} : [0,1] \to \mathbb{R}$ and $\Phi_q : \mathbb{R} \to \mathbb{R}$ are continuous.

- Summing and composition of univariate functions. Potentially address the curse of dimensionality (COD).
- Φ_q and $\phi_{q,p}$ not necessarily smooth. We may need more than two layers.

We parametrize the learnable activation functions by B-splines.

Approximation Theory

Suppose that a function $f(\mathbf{x})$ admits a smooth representation

$$f = (\mathbf{\Phi}_{L-1} \circ \mathbf{\Phi}_{L-2} \circ \cdots \circ \mathbf{\Phi}_1 \circ \mathbf{\Phi}_0)\mathbf{x}\,, \tag{2}$$

where $\Phi_{l,i,j}$ are smooth with derivatives uniformly bounded up to $k+1$-th order. Then using k-th order B-splines with $G+1$ grid points as activation functions, there exist $\Phi^G_{l,i,j}$ such that for any $0 \le m \le k$, we have the bound

$$\|f - (\mathbf{\Phi}^G_{L-1} \circ \mathbf{\Phi}^G_{L-2} \circ \cdots \circ \mathbf{\Phi}^G_1 \circ \mathbf{\Phi}^G_0)\mathbf{x}\|_{C^m} \le CG^{-k-1+m}. \tag{3}$$

In particular for L^2 or RMSE, we have the scaling law $k+1$. Informally, such functions are dense in the class of continuous functions, by [1].

Leveraging the 1D structure to get better scaling laws

Expressiveness and Spectral bias: KANs and MLPs

- ReLU-k MLP with width W, depth L can be represented by Spline-k-KAN with width W, depth 2L, grid size 2.
- k-KAN with width W, depth L, grid size G can be represented by ReLU-k MLP with width $(G+2k+1)W^2$, depth 2L.
- $O(G^2W^4L)$ parameter count for MLP, $O(GW^2L)$ for KAN in this formulation.
- Sharp if we restrict the depth of MLP.
- MLPs have difficulty learning high frequencies, i.e. have spectral bias.
- Splines, as building blocks of KANs, have no spectral bias.

Function Fitting

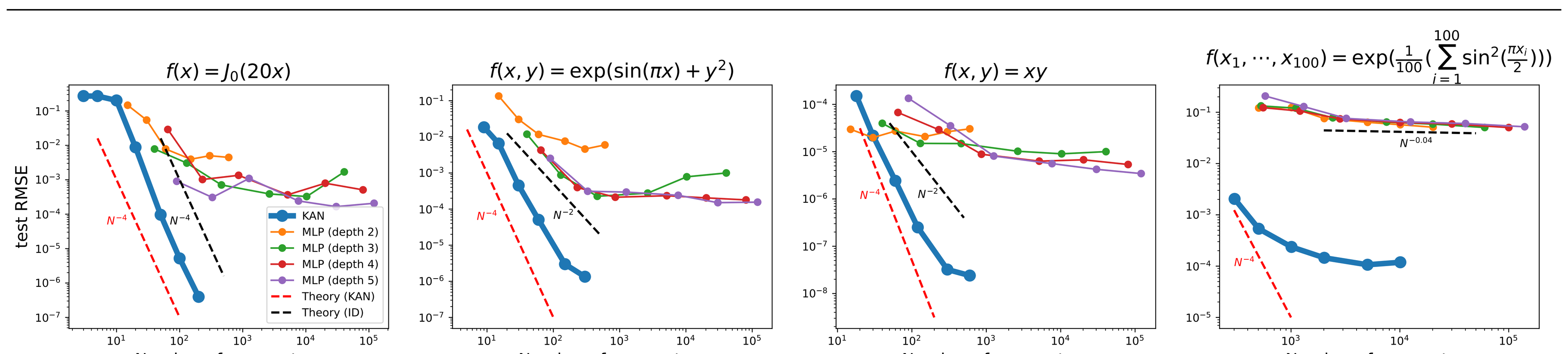


Figure 1. KANs almost saturate the fastest scaling law by theory ($\alpha = 4$), while MLPs scale slowly and plateau.

Function Fitting: 1D Waves of Different Frequencies

$$f(x) = \sum A_i \sin(2\pi k_i z + \varphi_i)\,, \quad k = (5, 10, \cdots, 45, 50).$$

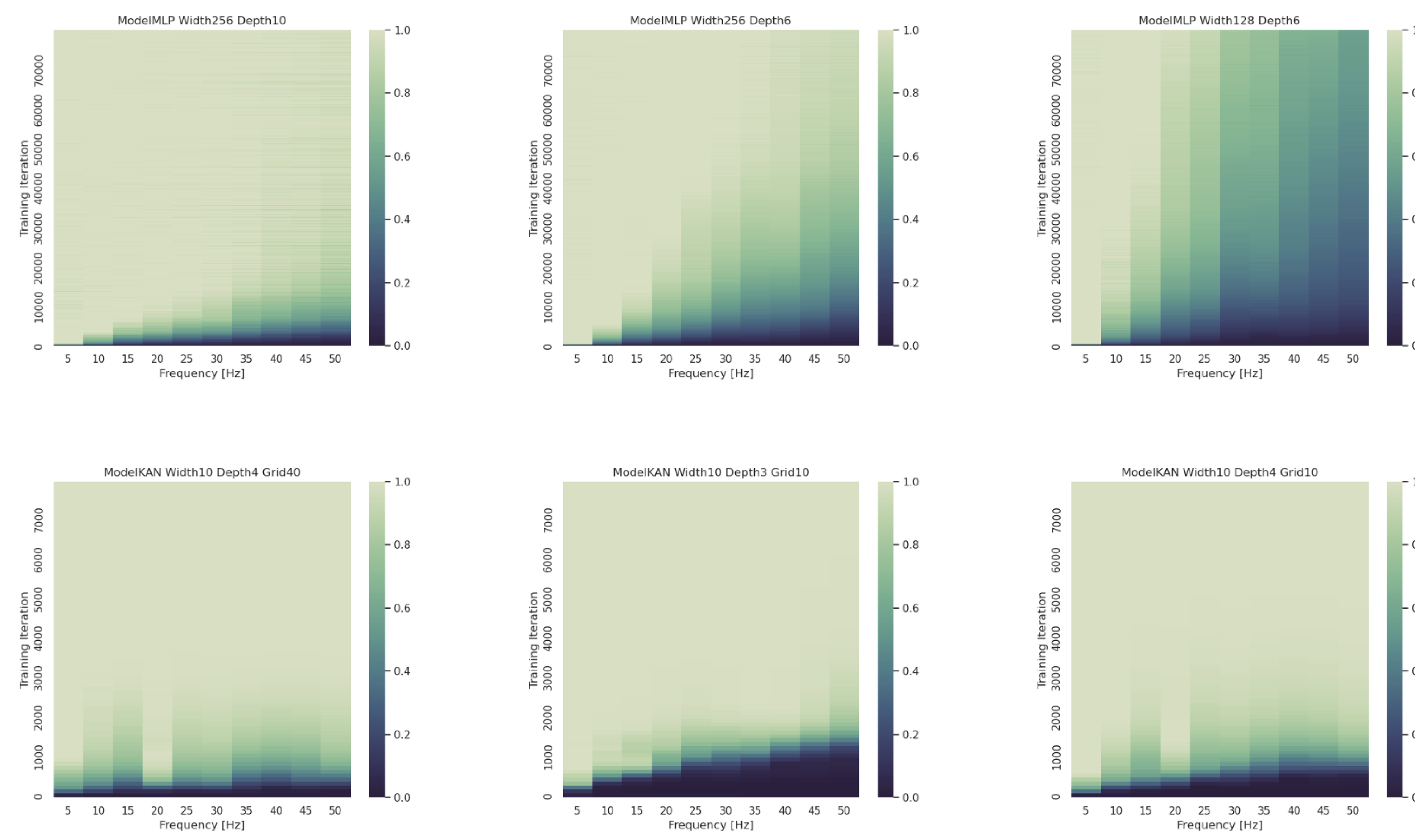


Figure 2. KANs learn high frequencies much faster, even with 10x fewer epoches.

PDE Solving: 2D Poisson of Different Frequencies

MLP of width 256, depth 6; KAN of width 10, depth 2, grid size 20.

$$-\Delta u = f, \;\; u = \sin(\pi x)\sin(\pi y) + \frac{1}{k}\sin(k\pi x)\sin(k\pi y).$$

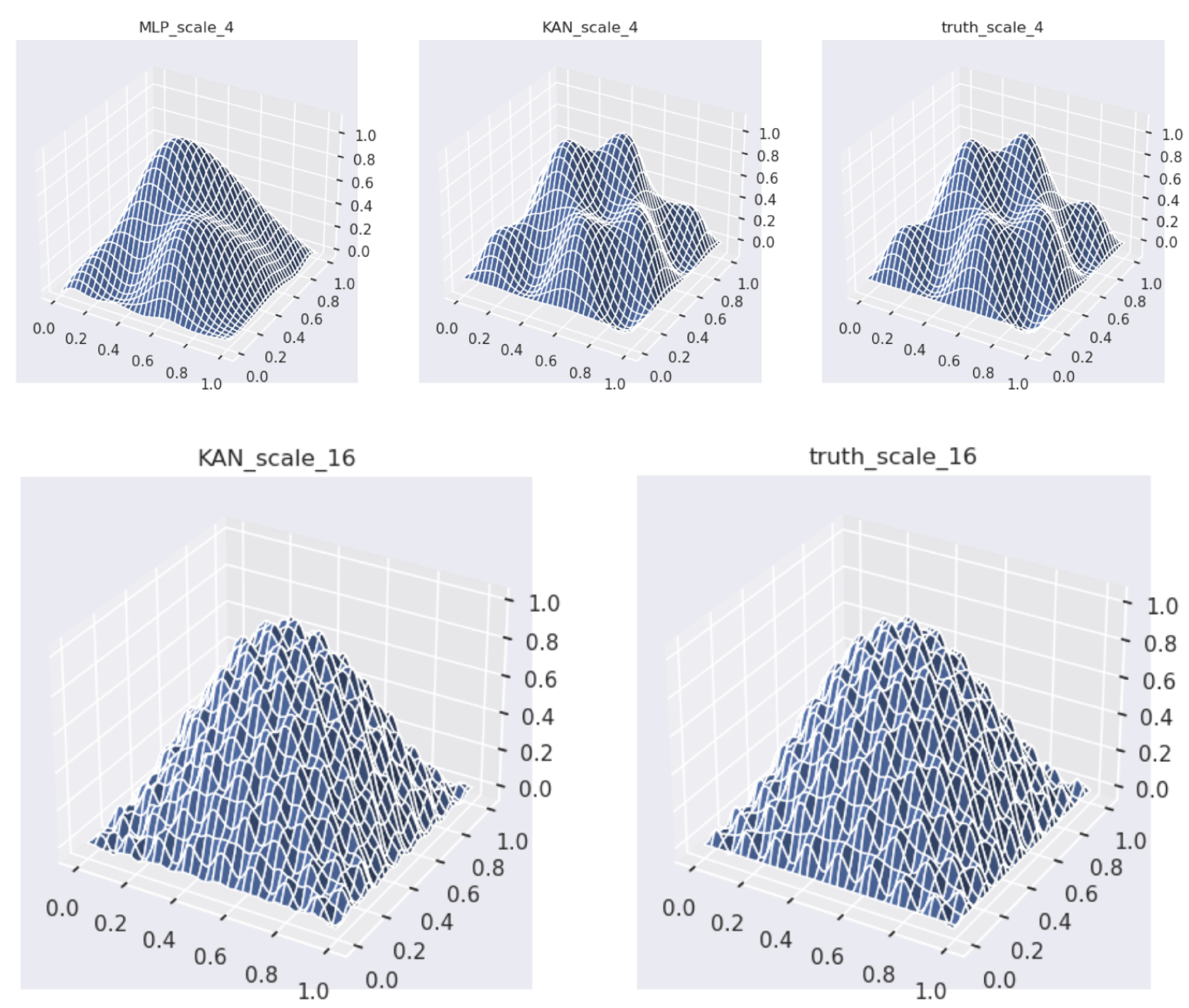


Figure 3. MLPs struggle to learn high frequency information without frequency encoding.

References

[1] Ming-Jun Lai and Zhaiming Shen.
The kolmogorov superposition theorem can break the curse of dimensionality when approximating high dimensional functions.
arXiv preprint arXiv:2112.09963, 2021.

[2] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark.
Kan: Kolmogorov-arnold networks.
arXiv preprint arXiv:2404.19756, ICLR 2025 oral, 2024.

[3] Yixuan Wang, Jonathan W Siegel, Ziming Liu, and Thomas Y Hou.
On the expressiveness and spectral bias of kans.
arXiv preprint arXiv:2410.01803, ICLR 2025, 2024.

https://roywangyx.github.io/
roywang@caltech.edu