پانزدهمین سمینار احتمال
و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان

15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan

Kolmogorov-Arnold Networks (KAN): Revolutionizing Prediction in Machine Learning

Hossein Haghbin*

Department of Statistics,Faculty of Intelligent Systems Engineering and Data Sciences, Persian Gulf University, Iran

Abstract

Kolmogorov–Arnold Networks (KANs) are a novel class of neural networks inspired by the Kolmogorov–Arnold representation theorem. By replacing traditional weight matrices with learnable univariate function modules, KANs provide a fundamentally different approach to modeling multivariate functions. This paper introduces the core concepts behind KANs, compares their structure and performance to traditional multilayer perceptrons (MLPs), and demonstrates their advantages in both prediction tasks.

Keywords: Kolmogorov-Arnold Networks, Nonlinear Regression, Function Approximation.
Mathematics Subject Classification (2020): 93E35, 60J20, 68W20.

1 Introduction

In recent years, deep learning has achieved remarkable success across a wide range of applications, including computer vision, natural language processing, and scientific computing. Classical neural architectures—such as multilayer perceptrons (MLPs), convolutional neural networks (CNNs), and transformers—typically rely on compositions of affine transformations and pointwise nonlinear activation functions. While effective, these architectures often act as black boxes and require large parameter counts and extensive data to perform well.

A foundational theoretical result underlying modern neural networks is the *Universal Approximation Theorem*, which states that a feedforward neural network with a single hidden layer containing a finite number of neurons can approximate any continuous function on a compact domain, given a suitable activation function Cybenko (1989); Hornik (1991). While this theorem establishes the expressive power of MLPs in theory, in practice, the required number of neurons may be very large, and the learned representations often lack interpretability.

On the other hand, the *Kolmogorov–Arnold Representation Theorem* provides a constructive way to represent any multivariate continuous function as a finite sum of univariate functions composed with addition Kolmogorov (1957). This formulation has inspired alternative architectures to MLPs, such as *Kolmogorov–Arnold Networks (KANs)*, recently introduced by

*Corresponding author, haghbin@pgu.ac.ir

Liu et al. Liu et al. (2024). Unlike MLPs, which use scalar weights and fixed activation functions, KANs leverage adaptive univariate functions—typically modeled as splines—to build each layer.

KANs bring multiple advantages: they align with a solid mathematical foundation, enhance interpretability, and often require fewer parameters for competitive performance, especially in low-data regimes. Related studies have also explored spline-based models and adaptive activations, but KANs offer a unified and principled approach that integrates these ideas into the core of the neural architecture.

This paper aims to:

- Provide a formal introduction to Kolmogorov–Arnold Networks (KANs),
- Contrast their structure and training mechanisms with traditional MLPs,
- Demonstrate their empirical performance in regression and classification tasks, and
- Discuss the interpretability and theoretical implications of such architectures.

2 Foundations and Motivation

One of the foundational results supporting the use of neural networks for function approximation is the Universal Approximation Theorem (UAT). This theorem formalizes the expressive power of feedforward neural networks, particularly multilayer perceptrons (MLPs).

Theorem 2.1 (Universal Approximation Theorem Hornik (1991))**.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a nonconstant, bounded, and continuous activation function. Then, for any continuous function $f : [0,1]^n \to \mathbb{R}$ and any $\epsilon > 0$, there exists a feedforward neural network with a single hidden layer of the form*

$$F(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \, \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i),$$

such that

$$\sup_{\mathbf{x} \in [0,1]^n} |F(\mathbf{x}) - f(\mathbf{x})| < \epsilon.$$

This theorem established that MLPs, even with just one hidden layer, are capable of approximating any continuous function on compact subsets of $\mathbb{R}^n$, provided that the width of the hidden layer is sufficiently large. However, in practice, deeper architectures with multiple hidden layers are often used to improve learning efficiency and generalization. Despite their universal expressiveness, MLPs typically rely on a composition of affine transformations and nonlinear activations, which may lead to difficulties in capturing localized or highly structured features in the data.

Although MLPs are theoretically universal approximators, their practical application faces challenges related to interpretability, training stability, and sample efficiency. Their reliance on global basis functions (like ReLU or sigmoid) can lead to inefficient parameter usage and poor local generalization. These limitations motivate the exploration of alternative architectures grounded in more structured mathematical representations of functions.

3 Kolmogorov–Arnold Networks (KANs)

The Kolmogorov–Arnold Representation Theorem provides a constructive guarantee that any multivariate continuous function can be decomposed into a finite composition of continuous univariate functions and additions. Formally, for any continuous function $f : [0,1]^n \to \mathbb{R}$, there exist continuous univariate functions ϕ_q and continuous functions ψ_p, such that:

$$f(x_1, \ldots, x_n) = \sum_{q=1}^{2n+1} \phi_q \left(\sum_{p=1}^{n} \psi_p^{(q)}(x_p) \right).$$

This result suggests that complex multivariate mappings do not inherently require multivariate nonlinearities, but instead can be constructed through composition and summation of univariate transformations. While the original proof is non-constructive and impractical for numerical purposes, it forms the theoretical backbone of Kolmogorov–Arnold Networks (KANs).

KANs operationalize this representation in a learnable framework by replacing traditional weight matrices and fixed activation functions used in Multilayer Perceptrons (MLPs) with learnable univariate functions. Each unit in a KAN layer performs a sum of such univariate functions applied to the individual coordinates of the input vector. Specifically, the j-th output of layer ℓ is computed as:

$$\Phi_\ell(\mathbf{x})_j = \sum_{i=1}^{d_{\ell-1}} \phi_{\ell,j,i}(x_i),$$

where each $\phi_{\ell,j,i} : \mathbb{R} \to \mathbb{R}$ is a learnable function, commonly parameterized using B-splines or Fourier series. The entire network is constructed as a composition of such layers:

$$\mathrm{KAN}(\mathbf{x}) = (\Phi_L \circ \Phi_{L-1} \circ \cdots \circ \Phi_1)(\mathbf{x}),$$

where L is the depth of the network.

To contrast this with traditional MLPs, consider a deep L-layer MLP defined by:

$$\mathrm{MLP}(\mathbf{x}) = (\mathbf{W}_L \circ \sigma_{L-1} \circ \mathbf{W}_{L-1} \circ \cdots \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x}),$$

where each $\mathbf{W}_\ell$ is a learnable weight matrix and σ_ℓ is a fixed elementwise activation function (e.g., ReLU or GELU).

In MLPs, expressive power arises primarily through the combination of linear transformations and fixed nonlinearities. In contrast, KANs achieve expressivity through flexible, data-driven univariate functions and additive combinations. This not only enables improved approximation fidelity but also offers enhanced interpretability by isolating nonlinear transformations at a per-dimension level.

4 Simulation Study: Nonlinear Regression

To evaluate the functional approximation capabilities of Kolmogorov–Arnold Networks (KANs) in comparison with conventional Multi-Layer Perceptrons (MLPs), we conducted a comprehensive simulation study on a synthetic nonlinear regression problem. The ground truth function is defined as a nonlinear interaction between two input variables:

$$y = \sin(3x_1) \cdot \cos(2x_2) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.05^2),$$

where $x_1, x_2 \sim \mathrm{Uniform}(-1, 1)$. A total of 5,000 samples were generated, with 80% allocated for training and 20% for testing. All input features were standardized to zero mean and unit variance prior to model fitting.

Two model families KANs and MLPs were considered. Each was evaluated in both shallow and deep configurations. Shallow architectures consisted of a single hidden layer, whereas deep models incorporated two or more hidden layers. The width of each hidden layer was systematically varied to examine the trade-offs between model capacity and performance.

Model performance was assessed on the test set usingMean Squared Error (MSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2) and the total number of trainable parameters (# Par.) metrics. The numerical results of the simulation are presented in Table 1. Across both shallow and deep configurations, KANs consistently outperform MLPs in terms of MSE and R^2, often with comparable or fewer trainable parameters. Notably, the shallow KAN model with structure $[2, 3, 1]$ achieves a remarkably low MSE of 0.0082 and an R^2 of

Table 1: Comparison of KAN and MLP on Nonlinear Regression Task

Type	**Model**	**Structure**	**# Par.**	**MSE**	**MAE**	R^2
Shallow	KAN	[2,2,1]	84	0.0238	0.1194	0.9727
		[2,3,1]	126	**0.0082**	**0.0715**	**0.9906**
		[2,4,1]	168	0.0107	0.0764	0.9878
	MLP	[2,16,1]	65	0.2212	0.3770	0.7460
		[2,32,1]	129	0.0484	0.1642	0.9444
		[2,64,1]	257	0.0405	0.1448	0.9535
Deep	KAN	[2,3,2,1]	196	0.0220	0.1133	0.9748
		[2,4,2,1]	193	0.0230	0.1100	0.9735
		[2,3,4,1]	252	0.1409	0.2864	0.8382
	MLP	[2,16,8,1]	335	0.0113	0.0822	0.9870
		[2,16,16,1]	308	0.0177	0.1001	0.9797
		[2,32,16,1]	369	0.0133	0.0812	0.9847

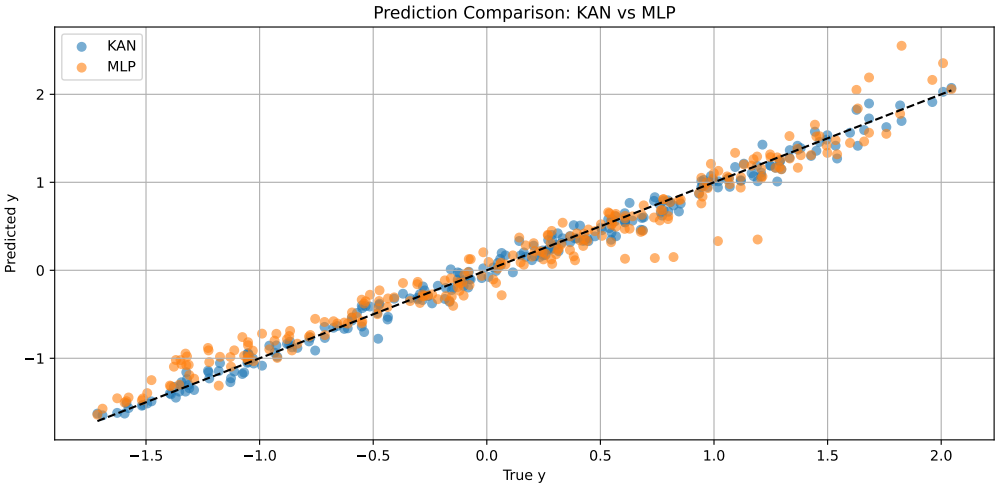


Figure 1: Prediction Comparison: KAN[2,3,1] vs MLP[2,64,1]

0.9906, despite having fewer parameters than its MLP counterpart $[2, 64, 1]$, which attains an MSE of 0.0405 and R^2 of 0.9535.

To complement the quantitative analysis, Figure 1 shows scatter plots comparing the predicted and actual values for the best-performing shallow models: KAN $[2, 3, 1]$ and MLP $[2, 64, 1]$. The predictions from the KAN model align more tightly with the diagonal reference line, demonstrating better adherence to the true output values. This result corroborates the superior regression accuracy of KAN, even with a significantly smaller model size.

In addition, Figure 2 presents surface plots of the true function alongside the outputs produced by the same two models across a uniform grid of input values. The KAN surface accurately reconstructs the complex nonlinear patterns of the target function, maintaining smooth transitions and capturing fine-grained structure. In contrast, the MLP surface exhibits oversmoothing and visible approximation errors, particularly in boundary regions. This visual evidence reinforces the claim that KANs can achieve high-fidelity approximations with fewer parameters and better generalization.

The findings of this study underscore the expressive efficiency of Kolmogorov–Arnold Networks for smooth nonlinear regression. KANs consistently outperform MLPs in shallow regimes, offering superior accuracy with lower model complexity. While MLPs begin to

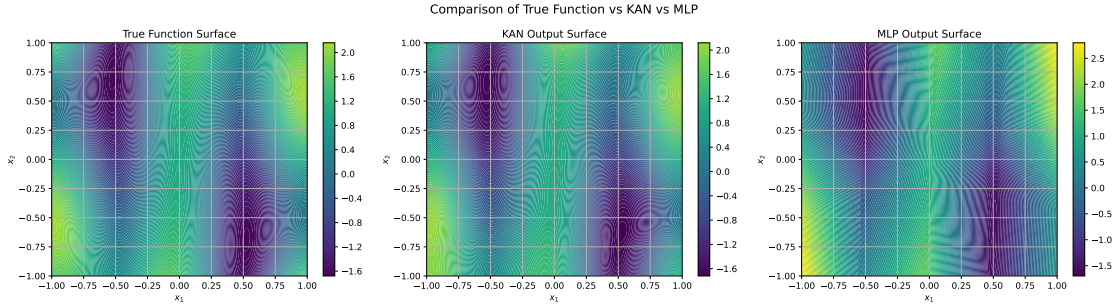


Figure 2: Surface Comparison of True Function, KAN[2,3,1], and MLP[2,64,1] Outputs

close the performance gap in deeper and wider configurations, they require significantly more parameters to do so. These results highlight the potential of KANs as a compact and interpretable alternative to traditional feedforward architectures, particularly in applications where parameter efficiency and functional interpretability are critical.

References

Cybenko, G. (1989), Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, **2**, 303–314.

Hornik, K. (1991), Approximation capabilities of multilayer feedforward networks. *Neural Networks*, **4**(2), 251–257.

Kolmogorov, A. N. (1957), On the representation of continuous functions of many variables by superpositions of continuous functions of one variable and addition. *Doklady Akademii Nauk SSSR*, **114**, 953–956.

Liu, X., et al. (2024), Kolmogorov–Arnold Networks. *arXiv preprint*, arXiv:2402.12782. https://arxiv.org/abs/2402.12782