Kolmogorov-Arnold Networks for High-Dimensional Estimation: a Method of Sieves Approach

Sami Abdurahman

October 17, 2025

Abstract

This paper introduces a novel sieve extremum estimator based on Kolmogorov-Arnold Networks (KANs), designed for nonparametric estimation in high-dimensional settings involving time series data. By integrating KANs with a method of sieves estimation approach and leveraging sparsity, our method achieves asymptotic rates of convergence in the Euclidean norm that are independent of the covariate dimension, thereby mitigating the curse of dimensionality. Specifically, our approach yields an explicit convergence rate of $o_P(n^{-1/4})$. This framework accommodates diverse applications such as high-dimensional conditional density estimation and nuisance function estimation within Double/Debiased Machine Learning, areas where traditional deep neural networks often struggle due to their "black box" nature, reliance on i.i.d. assumptions, and slower convergence rates. Our proposed KAN sieve extremum estimator overcomes these limitations by learning activation functions using B-splines, and its theoretical framework rigorously permits stationary mixing processes.

Author Affiliation: Toronto Metropolitan University, Department of Economics

Email: sami.abdurahman@torontomu.ca

Computations were performed on the Niagara supercomputer at the SciNet HPC Consortium. SciNet is funded by Innovation, Science and Economic Development Canada; the Digital Research Alliance of Canada; the Ontario Research Fund: Research Excellence; and the University of Toronto.

Full code is available at: https://github.com/samiabd8/SieveKAN

1 Introduction

In modern statistical and economic applications, flexible nonparametric function estimation methods with wide-applicability are increasingly important for modeling complex relationships arising from high-dimensional datasets. Examples of these methods include regression models based on kernels and penalized splines. However, a fundamental challenge facing the efficacy of these methods in high-dimensional settings is the curse of dimensionality (CoD). This phenomenon dictates that the sample size required for a given estimation accuracy grows exponentially with the dimensionality of the covariate space. As a direct consequence of this, derived rates of convergence will be prohibitively slow, rendering the use of a wide range of traditional nonparametric estimation techniques impractical in these settings. Although approaches such as generalized additive models (GAMs) and spline estimation methods may seek to mitigate the CoD through imposing structural assumptions (e.g., additivity), these assumptions[1] will generally fail to capture the compositional and interactive nature inherent in many complex relationships arising from real-world phenomena.

One of these flexible methods for nonparametric estimation that has demonstrated strong empirical performance in a variety of machine learning (ML) tasks (e.g., classification, natural language processing) in recent years is deep neural networks (DNNs), largely due to their powerful approximation capabilities for complex functions. Despite their demonstrated empirical excellence (see Goodfellow et al., 2016 for a comprehensive overview), the application of DNNs for rigorous statistical inference and econometric analysis presents many significant challenges. The most critical of which arises from the "black box" nature of these network architectures, which hinders interpretability in a manner that models such as GAMs do not. Furthermore, much of the recent work in the literature that attempts to bridge DNNs with statistical theory for inference[2] rely on independent and identically distributed (i.i.d.) assumptions. While such assumptions simplify the analysis on the approximation capabilities

[1] Commonly found in the literatures on nonparametric estimation and high-dimensional statistics.

[2] Typically falling under the category of "Deep ReLU networks", which we will discuss at length throughout this paper

of these structures, they render DNNs unusable for time series and panel data structures common in economic applications. Lastly, obtaining precise asymptotic convergence rates for the application of these structures under high-dimensionality is a highly nontrivial task due to the CoD. As we will show in this paper, commonly derived convergence rates in the literature require extremely restrictive conditions (explicitly stated or otherwise) that break down in such settings. Such conditions tie the smoothness of the function to be estimated to the dimension of the covariates and require degrees of differentiability that are generally unrealistic for even a modest number of regressors.

Introduced by Liu et al. (2024), Kolmogorov-Arnold Networks (KANs) represent a promising alternative to the general DNN architecture (multilayer perceptrons, or MLPs) that addresses some of these limitations. By replacing fixed activations in MLP structures with learned spline functions, KANs do not suffer from the "black box" nature associated with DNNs. As the name suggests, the universal approximation capabilities of KANs arises from the Kolmogorov-Arnold Representation theorem (KART). This theorem in Kolmogorov (1956) posits that complex multivariate functions can be represented as compositions of univariate functions. Formally, this theorem states:

Kolmogorov-Arnold Representation Theorem: for a smooth $f : [0,1]^d \to \mathbb{R}$,

$$f(\boldsymbol{x}) = f(x_1, \ldots, x_d) = \sum_{q=1}^{2d+1} \Phi_q \left(\sum_{p=1}^{d} \phi_{q,p}(x_p) \right) \tag{1}$$

This paper develops a novel and theoretically grounded sieve extremum estimator based on deep KANs for nonparametric estimation in high-dimensional settings with time series data. By integrating KANs with the sieve estimation framework of Chen and Shen (1998) and leveraging a sparse version of the Kolmogorov-Arnold representation theorem, we can achieve explicit asymptotic convergence rates in the $L_2(\cdot)$ norm that do not depend on the dimension of the covariates (d). This directly mitigates the CoD in a manner similar to sparse additive models. Our sieve estimation approach relies on a finite-dimensional space

that grows with sample size n in order to control network complexity growth in addition to sparsity, while simultaneously ensuring sufficient network approximation power. Unlike MLP architectures where the approximation power is tied to the network width (the number of neurons in the hidden layers) leading to intractable complexity, the expressiveness of KANs hinges on the flexibility of the splines in our learnable activation functions.

Our proposed KAN sieve extremum estimator achieves a dimension-independent convergence rate of $o_P(n^{-1/4})$, a desirable target rate for nonparametric function estimation as we will discuss. Under this sieve KAN framework, we are able to accommodate a wide range of practical applications involving time series and panel data. Such applications include the high-dimensional conditional density estimation example that we consider in simulations. We also demonstrate how our estimator can be used for nuisance function estimation in the Double/Debiased Machine Learning (DML) framework of Chernozhukov et al. (2018). DML has emerged as an indispensable tool in high-dimensional statistics for constructing valid inference for finite-dimensional parameters. We discuss how our proposed estimator can be extended to semiparametric structural models without DML (which also hinges on an i.i.d assumption), a task left to future research as it is beyond the scope of our proposed framework; due to our reliance on convergence in the $L_2(\mu)$ norm.

1.1 Related Literature

Many of the recent contributions to the literature on DNNs revolved around networks using the Rectified Linear Units (ReLU) activation function[3] due to their convenient and relatively simple properties. Yarotsky (2017, 2018) develops the approximation theory for this class of DNNs, while Farrell et al. (2021) and Schmidt-Hieber (2020) propose rigorous frameworks for their application in semiparametric inference and nonparametric regression, respectively. The former derives nonasymptotic high probability bounds for ReLU DNNs, while the latter

[3]Defined as $\sigma(x) = \max\{0, x\}$.

considers a sparsely connected[4] architecture and obtains an oracle inequality. In an earlier study addressing the CoD and potential avenues for circumventing it in the context of ReLU DNNs[5], Mhaskar and Poggio (2016) define *relative dimensions*. This concept provides quantitative measurements of sparsity tied to parameters of the DNN and leverages compositional function assumptions to provide a plausible explanation for the improved performance of DNNs relative to "shallow" (single hidden-layer) networks. We contrast this with our proposed sieve KAN using a conditional density estimation example in the results section of this paper, highlighting its limited applicability for general DNNs.

Linking the improved performance of artificial neural networks (aNNs) when additional layers are added in MLPs to the KART, Schmidt-Hieber (2021) presents an interpretation of this representation theorem using deep ReLU networks under simple modifications. This constituted a breakthrough in the application of the KART in aNNs, once thought to be irrelevant[6] in this context. Although several papers in the time between Kolmogorov (1956) and Schmidt-Hieber (2021) have attempted to show the usefulness of the KART in aNNs (e.g., Sprecher, 1996, 1997), these were limited to network structures with two hidden-layers and fixed activation functions.

Applying this crucial insight, Liu et al. (2024) propose an alternative to MLPs in the form of KANs, which derive their universal approximation capabilities from the KART rather than the universal approximation theorem[7]. While both of these DNN architectures are fully-connected, KANs replace the fixed activation functions in MLPs with *learnable* activation functions in the form of B-splines. The authors demonstrate how by defining the layers of KANs in terms of matrices of univariate functions ($\mathbf{\Phi}_l$), the Kolmogorov-Arnold representation can be generalized to arbitrary widths and depths. As a result, KANs are effectively models with MLPs on the outside and splines on the inside, the latter contributing to sig-

[4]With respect to the network weights, where a compositional structure assumption is imposed on the target function.

[5]With a focus on deep convolutional networks, see LeCun et al. (2015) for an overview.

[6]e.g., Girosi and Poggio (1989), titled: "Representation Properties of Networks: Kolmogorov's Theorem Is Irrelevant".

[7]Introduced in Hornik et al. (1989), a foundational paper in the literature on DNNs and machine learning.

nificantly improved accuracy as shown by Liu et al. (2024) in a wide range of applications. Deriving generalization bounds for KANs, Zhang and Zhou (2024) show how it is possible to guarantee that the bound scales with the ℓ_1 norm of the coefficients when the activation function is composed of B-spline basis functions. In both of these KAN papers, ℓ_1-regularization plays an essential role. While these papers lay the groundwork for KANs with learned activation functions in the form of B-splines, our proposed framework achieves explicit rates of convergence and an oracle inequality as a result of our sieve KAN M-estimator.

In a major contribution to the literature on sieve extremum estimation (Grenander, 1981), Chen and Shen (1998) obtain convergence rates for these nonparametric and semiparametric estimators involving time series data; in addition to establishing the root-N asymptotic normality of "plug-in" sieve extremum estimates with respect to smooth functionals. The method of sieves entails estimating a function belonging to a possibly infinite-dimensional parameter space Θ using a sequence of approximating parameter spaces (growing in complexity as $n \to \infty$) that are dense in Θ. This M-estimator framework derives theoretical results using a quasi-maximum-likelihood estimation (QMLE) method which does not suffer from the issues associated with infinite-dimensional maximum likelihood[8]. Crucially, the extension of sieve estimation to time series data in Chen and Shen (1998) allows for the application of this method to a wide range of economic applications that would otherwise not be possible under the i.i.d assumptions made in the existing literature (e.g., White and Wooldridge, 1991). Among the applications involving time series models considered to demonstrate this sieve extremum estimator, rates of convergence for nonlinear ARX models using aNNs are obtained. Chen and White (1999) obtain improved approximation rates for shallow sieve aNN estimators building directly upon this framework, see Abdurahman (2025) for a full exposition. Shen et al. (2023) investigate the asymptotic properties of shallow sieve network estimators and discuss how their results can potentially be extended to DNNs with Lipschitz continuous activation functions. To the best of our knowledge, there does not exist a frame-

[8]Namely, slow convergence rates and inconsistency.

work integrating the method of sieves with modern DNN architectures in order to establish the consistency of such networks in nonparametric regression.

The remainder of this paper is organized as follows. Section 2 defines the target function space and outlines our proposed KAN sieve extremum estimator, including the key parameters of the sieve KAN architecture and its assumptions. Section 3 presents our main theoretical results on approximation error and convergence rates. We present results of Monte Carlo simulations for our CDE and DML (double/debiased mean regression) examples in Section 4. Section 5 concludes with a summary and a discussion of directions for future research.

2 Sieve Kolmogorov-Arnold Networks

We begin by defining the relevant metric spaces and their properties, in addition to the notation used throughout this paper. Given a sequence of covariates $\{X_t\}_{t=1}^n$ with dimension d, let μ be a probability measure on $\mathbb{R}^d$ (e.g., the distribution of $X_t \in \mathbb{R}^d$). We define $L_2(\mu)$ as the space of all measurable functions $h : \mathbb{R}^d \to \mathbb{R}$ such that $\int_{\mathbb{R}^d} |h(x)|^2 d\mu(x) < \infty$, equipped with the following norm:

$$\|h\|_{L_2(\mu)} = \left(\int_{\mathbb{R}^d} |h(x)|^2 d\mu(x) \right)^{1/2}$$

To simplify the notation, let $\|\cdot\| \equiv \|\cdot\|_{L_2(\mu)}$ unless stated otherwise.

In order to characterize the space of the target function, we define the multi-index $\lambda = (\lambda_1, \ldots, \lambda_d)^\intercal$, which represents a vector of non-negative integers with order $|\lambda| = \sum_{i=1}^d \lambda_i$. Let D^λ represent the weak partial derivative operator corresponding to λ, where:

$$D^\lambda h(x) = \frac{\partial^{|\lambda|} h(x)}{\partial x_1^{\lambda_1} \ldots \partial x_d^{\lambda_d}}$$

The target function space is defined as a Sobolev space $\mathcal{F}_d^{m+1} \equiv W_2^{m+1}(\mu)$. In addition to

consisting of all measurable functions $h : \mathbb{R}^d \to \mathbb{R}$ such that $h \in L_2(\mu)$, this space contains its weak derivatives $D^\lambda h \in L_2(\mu)$ for all multi-indices λ with $|\lambda| \leq m+1$. The norm for a Sobolev space $W_2^{m+1}(\mu)$ is given by:

$$\|h\|_{W_2^{m+1}(\mu)} = \left[\sum_{|\lambda|=0}^{m+1} \left(\|D^\lambda h\|_{L_2(\mu)} \right)^2 \right]^{1/2}$$

As we demonstrate in our model assumptions, this definition allows us to quantify the smoothness of the true function we seek to estimate. As a result, we are able to link the properties of the proposed sieve KAN architecture that features (learned) spline activation functions to this function space. Critically, this target space allows us to work towards convergence rates where the dependence on the smoothness parameter m is absorbed into constants; unlike certain "minimax" convergence rates[9] commonly found in the literature on deep ReLU networks. For example, a common minimax estimation rate in nonparametric estimation takes on the form of $n^{-m/(2m+d)}$, which is prohibitively slow in high-dimensional settings. By avoiding such a dependency, our approach is a key step towards mitigating the CoD in DNN architectures.

Although our analysis relies on the smoothness provided by the Sobolev space, we are only able to obtain convergence rates in the $L_2(\mu)$ norm rather than in the stronger Sobolev norm (the latter implying convergence in the former). We will demonstrate why this is the case when applying spline approximation theory in our key results, as rates of convergence in the Sobolev norm would not be independent of m under the sieve KAN architecture. This is one of the main ways our approach differs from Chen and White (1999), where results for consistent nonparametric sieve network estimators (in the form of single hidden-layer aNNs) are obtained in the Sobolev norm. Nonetheless, this weaker form of convergence is sufficient for the applications we consider, as well as for important extensions[10] left to future

[9] Where the smoothness degree of the target function m is tied to d; see Schmidt-Heiber (2020) for a full discussion in relation to DNNs.

[10] Namely, obtaining asymptotic normal linear plug-in estimators in a semiparametric model framework, where the nonparametric component is estimated using our proposed KAN sieve extremum estimator.

work. Furthermore, given how quickly convergence rates involving non-static conditions on m breakdown in high-dimensional settings (see the Appendix for an illustrative example), this tradeoff is justified by how it enables the proper theoretical grounding for analysis in these settings.

We assume the target function $f_0 \in \mathcal{F}_d^{m+1}$ has a sparse compositional structure (Assumption **K.3**), where the number of relevant functions in the Kolmogorov-Arnold representation is given by $s_0 = O(\log n)$ or takes on the value of a fixed constant. In the context of the possibly infinite-dimensional parameter space $\mathcal{F}_d^{m+1}$, this assumption is comparable to those made in sparse additive model frameworks[11], as it entails approximating a multivariate function using a collection of univariate functions[12]. As we demonstrate throughout this paper, a sieve KAN can be viewed as a generalization of these models to sparse compositional structures (thus allowing for complex interactions), embedded within a DNN architecture. In terms of multiplicative interactions, Liu et al. (2024) demonstrate how even a simple two-layer KAN structure is capable of approximating functions such as $y = x_1 x_2$. In this example, it can be observed that the KAN computes $x_1 x_2$ by leveraging $2x_1 x_2 = (x_1 + x_2)^2 - (x_1^2 + x_2^2)$, where the spline functions learn the individual parts of this approximation. The main consequence of this sparse compositional assumption imposed on the target function $f_0 \in \mathcal{F}_d^{m+1}$ is that it allows for the dimension of the covariates, d, to grow with the sample size at even a polynomial rate: $d = O(n^{\zeta_d})$ for some $\zeta_d > 0$. As we show later on, the assumption regarding s_0 can be relaxed to allow for a polynomial sparsity growth rate with respect to n (i.e., $s_0 = O(n^{\zeta_s})$, $\zeta_s > 0$), subject to a simple condition.

In either case, we are able to derive explicit asymptotic convergence rates in the L_2 norm that do not depend on d at all. The ability of our proposed framework to permit dimensions that scale as the sample increases without causing the metric entropy to grow uncontrollably[13] makes it a very promising direction for the application of sparse deep net-

[11] Such as Raskutti et al. (2009), Meier et al. (2009) and Friedman (1991).

[12] Also referred to one dimensional, or "1D" functions

[13] By leveraging spline approximation theory in addition to the method of sieves, as we discuss in the next section.

works in high-dimensional settings; namely due to how assumptions commonly made in high-dimensional models such as $s_0 = O(1)$ (which implies that the target function can be approximated by a fixed number of univariate functions) can be overly restrictive in many practical applications. As a result, we mitigate the CoD by exploiting sparsity in a manner similar to sparse additive models and bound the network entropy by using a sieve estimation approach. We note that a similar claim is also made in Liu et al. (2024), one that is debatable for general KAN classes in the absence of sparsity and appropriate mechanisms for controlling complexity[14]. Without these theoretical underpinnings, the network complexity *implicitly scales with* d and renders general KAN classes vulnerable to exponential growth of said complexity, especially in high-dimensional settings. Going forward, we use the symbol $\asymp$ to denote asymptotic equivalence, which is important for our network parameter definitions.

2.1 KAN Sieve Extremum Estimator

We integrate our proposed KAN architecture in the sieve extremum estimator framework of Chen and Shen (1998), which allows for time series data (Assumption **K.1**) and avoids the reliance on the otherwise restrictive i.i.d assumption (e.g., as in Schmidt-Heiber, (2020) and Farrell et al., (2021), among many other works on deep ReLU networks). This is of significance as many of the potential high-dimensional applications a researcher might be interested in using deep networks for generally involve panel data. We aim to extend this framework to semiparametric structural models, although due to certain limitations involving the norm of our convergence rates discussed later in the paper, we leave this task to future research. Instead, we consider semiparametric applications in the double/debiased machine learning framework introduced by Chernozhukov et al. (2018). Our proposed KAN sieve extremum estimator[15], $\theta_n \in \Theta_n$, achieves the required rate of convergence (precisely $o(n^{-1/4})$ in the $L_2(\mu)$ norm) for nuisance function estimation under Neyman-orthogonal moments/scores in DML.

[14] i.e., the sparse compositional form imposed on the target function and our sieve estimation approach.
[15] To avoid confusion, "sieve KAN" refers to the architecture with network weights ω_n.

For a stationary process $Z_t \equiv \{Y_t, X_t\}_{t=1}^n$ and given a marginal probability measure P_0 where $\theta_0 \equiv \Gamma(P_0)$ (a value indexing the data-generating process), we define the following:

$$\hat{\theta}_n \equiv \arg\max_{\theta \in \Theta_n} \mathcal{L}_n(\theta) \tag{2}$$

$$\mathcal{L}_n(\theta) \equiv \frac{1}{n} \sum_{t=1}^{n} l(Z_t, \theta) - \lambda_n R(\theta) \tag{3}$$

$$\lambda_n \asymp O\left(\sqrt{\frac{\log p_n}{n}}\right) \tag{4}$$

Here, $l(\cdot)$ is a quasi-log-likelihood function belonging to a class of general loss functions, permitting a variety of nonparametric regression problems. For example, $l(Z_t, \theta) = -\frac{1}{2}(Y_t - \theta(X_t))^2$ in the case of mean regression and $l(Z_t, \theta) = (Y_t - \theta(X_t))(\tau - \mathbf{1}_{Y_t \leq \theta(\cdot)}(\cdot))$ when applied in quantile regression models (see Koenker, 2005). The ℓ_1-regularization penalty term λ_n is asymptotically equivalent to a term that accounts for the (log) total number of network parameters relative to the sample size, chosen based on an oracle inequality (Assumption **S**). Lastly, $R(\theta)$ is a regularizer term used to represent the constrained coefficients of our learned B-spline activation functions, which we expand upon in the following section.

2.2 Sieve KAN Architecture

The network architecture of our sieve KAN estimator has three key parameters: the number of layers (L_n), the network width (W_n; corresponding to the number of nodes W_n^l in a given layer $l \in [1, \ldots, L_n]$) and the size of the B-spline grid (G_n) for the learnable activation functions which makes the network interpretable. Unlike how MLPs place fixed activation functions on nodes, the learnable activation functions belong on *edges* (or, "in-between" layers) so that the KAN nodes simply perform summation. This compositional structure

can be represented as follows, in contrast with the structure of a deep MLP:

$$\text{KAN}(\boldsymbol{x}) = \big(\boldsymbol{\Phi}_{L_n-1} \circ \boldsymbol{\Phi}_{L_n-2} \circ \ ... \ \circ \boldsymbol{\Phi}_1 \circ \boldsymbol{\Phi}_0\big)\boldsymbol{x} \tag{5}$$

$$\text{MLP}(\boldsymbol{x}) = \big(\boldsymbol{W}_{L-1} \circ \sigma \circ \boldsymbol{W}_{L-2} \circ \sigma \circ \ ... \ \circ \boldsymbol{W}_1 \circ \sigma \circ \boldsymbol{W}_0\big)\boldsymbol{x} \tag{6}$$

where σ denotes a fixed activation (e.g., ReLU) and $\boldsymbol{\Phi}_l$ represents a collection of learnable univariate spline functions at layer l. In other words, the typical weight matrices on the edges of MLP structures are replaced with univariate functions in the form of B-splines. To reflect this, we define $r_n \coloneqq L_n W_n^2$ as our main sieve parameter (note: $r_n \equiv W_n$ in the case of single hidden layer aNNs from CW1999) representing the number of total *potential* edges in our KAN architecture.

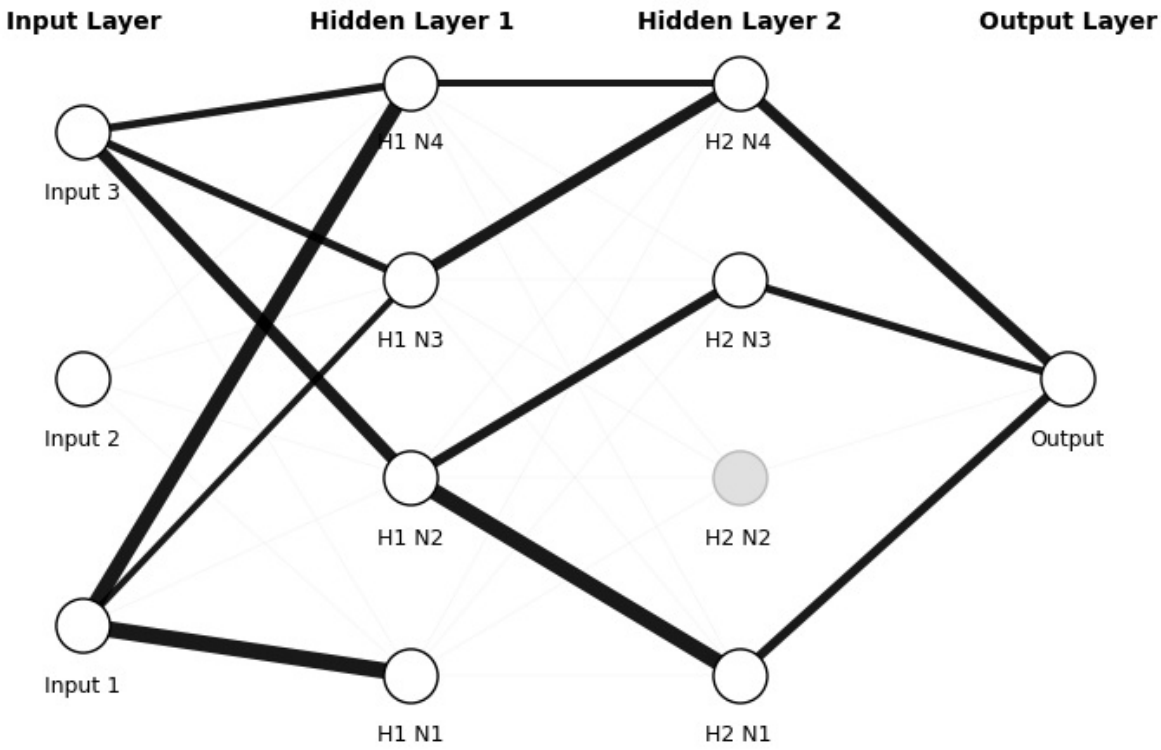

Figure 1: A Sparse KAN With Active Paths Highlighted

The difference between r_n and the number of *active* paths (denoted by s_n and determined by s_0 in addition to the ℓ_1 constraint Δ_n: another key parameter in our sieve space definition)

can be seen in Figure 1, where the solid lines represent learned activation functions with nonzero B-spline coefficients under ℓ_1-regularization. In this example, the second input variable (x_2) is irrelevant to the true target function and there is an inactive node present in the second hidden layer, lacking any activation functions to sum as a result of the ℓ_1-regularization. Using r_n, we define our total parameters in the KAN as $p_n \coloneqq O(r_n G_n)$ as in Liu et al. (2024) where $p = O(LW^2G)$. For each $l \in [1, \ldots, L_n]$, a layer within the sieve KAN can be represented as follows:

$$\mathbf{\Phi}_l = \sum_{i=1}^{s_n} \phi_{i,j} \left(\sum_{j \in \mathcal{J}_i} \phi_{l,i,j}(h_{l-1,j}) \right) \tag{7}$$

where the network output is given by $f(\cdot) = \mathbf{\Phi}_{L_n}$. The index of these learned activation functions allows us to determine the location of active paths inside the network, which is of significance due to the many potential paths in sieve KANs used for high-dimensional applications[16].

The fact that the sieve KAN learns activation functions instead of weight matrices is what allows us to fix W_n with respect to the sample size when $s_0 = O(1)$ in the first place, attempting to do so with MLP structures[17] is incompatible with our sieve estimation approach. Under a hypothetical framework incorporating MLPs with sieve estimation, ensuring that the MLP simultaneously possesses a bounded entropy and a degree of approximation power sufficient for high-dimensional estimation is an extremely challenging task. This is due to how the approximation power of an MLP is tied to its width and depth, which are typically required to grow with the sample size in order to achieve convergence rates comparable to the aforementioned minimax rates in nonparametric regression. As a result of this inherent dependency, it is difficult to control the metric entropy of such a sieve without inordinately restricting the approximation capabilities of the MLP.

[16] e.g., in our first simulation DGP, where $r_n = 2000$ despite $n = 500$.

[17] See Farrell et al. (2021) for a discussion on this class of "fixed width" networks, which are especially likely to struggle in high-dimensions.

To the contrary, due to how the expressiveness KANs are instead linked to the flexibility of the splines (i.e., is capable of being increased by simply expanding the grid size), this architecture guarantees the controlled growth of the sieve space. In these structures, the width determines the number of interactions between input variables and by extension, the diversity of interactions within a given layer. This makes KANs well-suited for sieve estimation in high-dimensions, in addition to how they do not suffer from the "black box" nature of MLPs. Likewise, stacking additional layers in a KAN permits for richer representations of f_0. As Schmidt-Hieber (2020) states: "...our theory suggests that for nonparametric regression, scaling the network depth with the sample size is natural." KAN architectures significantly contribute to this framework that harmonizes DNNs with a method of sieves estimation approach for nonparametric regression.

Let $\Theta \equiv A \times \mathcal{F}_d^{m+1}$ be the parameter space where $A \subset \mathbb{R}^a$: a compact set representing the finite-dimensional space of our linear parameters of interest and Θ is equipped with the pseudometric $\|\cdot\|$. The sieve space is given by $\Theta_n \equiv A \times \mathcal{K}_n$, where for a vector of network "weights" $\boldsymbol{\omega}_n \equiv (p_n, \{\boldsymbol{c}_j\}_{j=1}^{r_n})$, we define the sieve KAN class:

$$\mathcal{K}_n = \Bigg\{ f(\boldsymbol{x}; \boldsymbol{\omega}_n) \mid f(\boldsymbol{x}) = \Phi_{L_n}(\Phi_{L_n-1}(\dots \Phi_1(\boldsymbol{x})))), \\ \boldsymbol{\omega}_n \equiv (p_n, \{c_{j,k}\}_{j=1,k=1}^{r_n,G_n}),\ \sum_{j=1}^{r_n} \|\boldsymbol{c}_j\| \leq \Delta_n \Bigg\} \tag{8}$$

Here, $\boldsymbol{\omega}_n$ consists of the network architecture parameters p_n and a collection of all B-spline coefficients across the r_n potential edges within the KAN. In practice, activation functions with non-zero coefficients that do not belong to the active paths s_n, i.e., do not contribute to the network output, are pruned. See Figure 1 for an example of such a function, corresponding to the first input edge.

Following Liu et al. (2024), we define our learnable activation function to include a basis function $b(x)$ in order to provide increased stability as part of our network initialization. During training, the initial activation function will take on a weighted sum of the basis

function and the B-spline function as follows:

$$\phi(x) = w_b b(x) + w_s \text{spline}(x) \tag{9}$$

$$b(x) = \text{silu}(x) = x/(1+e^{-x}), \quad \text{spline}(x) = \sum_{k=1}^{G_n} c_k B_k(x) \tag{10}$$

where the weights (w_b, w_s) are updated during training in addition to the learned B-spline coefficients. Note that the basis function here is simply a combination of the ReLU and sigmoid activation functions (resulting in the "sigmoid linear unit", or "SiLU" activation). See Abdurahman (2025) for a discussion on the useful properties of this "hybrid" activation function with respect to gradient stabilization in the context of (single hidden layer) sieve networks. The B-spline basis functions $B_{j,k}(x)$ are defined recursively[18]: for order k and a (non-decreasing) sequence of knot point t_j, the B-spline $B_{j,k}(x)$ is given by:

$$B_{j,k}(x) = \frac{x - t_j}{t_{j+k} - t_p} B_{j,k-1}(x) + \frac{t_{j+k+1} - x}{t_{j+k+1} - t_{j+1}} B_{j+1,k-1}(x)$$

with the base case for order $k = 1$ defined as:

$$B_{j,1}(x) = \begin{cases} 1 & \text{if } t_j \leq x < t_{j+1} \\ 0 & \text{otherwise} \end{cases}$$

In this framework, $\text{silu}(x)$ is equivalent to a fixed bias term that vanishes and the approximation power of the KAN is derived entirely from the learned B-splines. This is due to how our initialization strategy entails starting with $w_b = 1$ and $w_s = 0$, before gradually converging to the case where $w_b = 0$ and $w_s = 1$. As a result of this, we only consider $\phi(x) = \sum_{k=1}^{G_n} c_k B_k(x)$ in the proofs of Theorem 1-3, see the Appendix for the full exposition on this initialization method. In order to simplify the analysis in our key theorems/results,

[18]This definition is commonly referred to as the Cox-de Boor recursion formula: see Cox (1972) and de Boor (1978).

we fix the polynomial degree of our splines as order $k = 4$, with G_n intervals accordingly ($G_n + 1$ knot points in total: $t_0 < t_1 < \ ... \ < t_{G_n}$). We note that in practice, KANs typically use cubic splines for a wide range of general applications, so this formalization does not incur additional costs or tradeoffs beyond in certain edge cases outside of the scope of nonparametric estimation (e.g., symbolic regression).

2.3 Key Parameters and Model Assumptions

For cases where A is a nonempty set (e.g., in semiparametric mean regression, where $\alpha_0 \in A$ is the coefficient for the linear), we have: $\theta_0 \equiv (\alpha_0, f_0) \in \Theta$ and $\theta_n \equiv (\alpha_n, f(\cdot; \boldsymbol{\omega}_n)) = (\alpha_n, \pi_n f_0) \in \Theta_n$, where π_n is a projection operator representing the best possible approximation of f_0 in the sieve space. Under the pseudometric $\|\cdot\|$ on Θ, we define Θ_n as dense in Θ if for any $\theta \in \Theta$, there exists a $\pi_n\theta \in \Theta_n$ such that $\|\theta - \pi_n\theta\| \to 0$ as $n \to \infty$. Furthermore, we define $\pi_n\theta_0 \equiv \inf_{\theta\in\Theta_n}\|\theta - \theta_0\|$. We let $E[\cdot]$ denote the expectation under measure P_0 henceforth.

We define our KAN parameters $p_n \equiv (L_n, W_n, G_n)$ as follows:

$$L_n \asymp O(log\ n), \quad W_n \gtrsim s_0 \asymp O(\log n), \quad G_n \asymp n^{\gamma} \tag{11}$$

$$\Delta_n \asymp O(\log n) \vee O(n^{\zeta_\delta}), \quad \frac{1}{16} < \gamma < 1 \tag{12}$$

where γ is a parameter required to guarantee sufficient approximation power in order to deliver our $o_P(n^{-1/4})$ convergence rate. With respect to selecting a suitable width such that $W_n \geq s_0$, we recommend the use of cross-validation. In theory, overparameterization in any of the parameters p_n does not harm the model performance due to the adaptability of our sieve space (see the oracle inequality). We verify this in the simulation analysis across many DGPs, defining $W_n = C_w \log n$ for some constant $C_W > 0$ and evaluating a wide range of potential values. In the case of $s_0 = O(1)$, we simply let $W_n = C_W$ and focus our attention on the case above in proofs involving the metric entropy of this KAN architecture.

The requirement that the width of the sieve KAN grows at a rate equivalent to that of s_0 arises from our sparse compositional structure on $\theta_0 \in \mathcal{F}_d^{m+1}$ (Assumption **K.3(a)** below). Simply put, this ensures that we have enough potential paths to approximate the active edge functions in θ_0, even as $n \to \infty$. For simplicity, we set $L_n = \lfloor \log n \rfloor$ in practice.

Sieve KAN assumptions:

K.1(a): Data $\{Z_t\} = \{Y_t', X_t'\}'$ is i.i.d or m-dependent ($m \in \mathbb{N}$), or

K.1(b): β-mixing stationary with $\beta(j) \leq \beta_0 j^{-\xi}, \xi > 2$ for some $\beta_0 > 0,\ \xi > 2$, or

K.1(c): ϕ-mixing (uniform mixing) stationary with $\psi(j) \leq \psi_0 j^{-\xi}$ for some $\psi_0 > 0,\ \xi > 2$.

K.2(a): The parameter space Θ is equipped with the pseudometric $\|\cdot\|$, Θ_n is dense in Θ and $\mathcal{F}_d^{m+1} \equiv W_2^{m+1}(\mu)$ is a Sobolev space.

K.2(b): The quasi-log-likelihood function $l(Z_t, \theta)$ satisfies:

$$\inf_{\theta \in \Theta_n, \|\theta - \theta_0\| \geq \epsilon} \text{Var}[l(Z_t, \theta_0) - l(Z_t, \theta)] \geq C_1 \epsilon^{2\rho}, \quad \text{for all small } \epsilon > 0$$

where $C_1 > 0$ and $\rho = 1$ for quadratic losses.

K.2(c): (Entropy) There exists a function $U(\cdot)$ and some $\kappa > 0$ such that for any $\delta > 0$:

$$\sup_{\theta \in \Theta_n, \|\theta_0 - \theta\| \leq \delta} |l(\theta) - l(\theta_0)| \leq \delta^{\kappa} U(Z_t)$$

Our initial set of assumptions are standard in the literature on sieve extremum estimators. Assumption **K.1** permits data dependence structures beyond the baseline i.i.d case used to derive the entropy bounds of DNNs. Since m-dependence is generally an unrealistic assumption for economic and financial time series[19], the β-mixing and uniform mixing assumptions are necessary for controlling serial dependence in the context of empirical processes. Assumption **K.2(b)** guarantees that θ_0 is identifiable and that $l(\cdot)$ is well-behaved

[19]To see why this is the case, consider a Moving Average process of order m (denoted $MA(m)$) with a sequence of i.i.d shocks ϵ_t defined by $X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_n \epsilon_{t-m}$. For a finite m, this implies that the autocorrelation of this process is zero for all lags greater than m, thus ruling out the long-term persistence common in macroeconomic shocks, to give an example.

in a neighborhood surrounding it. The size of these neighborhoods in Θ_n is dictated by the term δ^κ in Assumption **K.2(c)**, thus preventing the sieve from becoming overly fine.

Sieve KAN assumptions (continued):

K.3(a): (Sparse Compositional Structure) The target function $f_0 \in \mathcal{F}_{s_0}^{m+1},\ f_0 : [0,1]^{s_0} \to \mathbb{R}$ takes on the following compositional form:

$$f_0(\mathbf{x}) = \sum_{i=1}^{s_0} \phi_i \left(\sum_{j \in \mathcal{J}_i} \phi_{ij}(x_j) \right)$$

with $|\mathcal{J}_i| \ll d$, and $d = O(n^{\zeta_d})$ for some $\zeta_d > 0$. Furthermore, $\|\phi_i\|_\infty, \|\phi_{i,j}\|_\infty \leq B$.

K.3(b): (Sparsity Growth) Let $W_n \geq s_0$. For the $s_n \geq s_0$ corresponding to the best approximation $\pi_n \theta_0$:

i. $s_0 = O(1)$, or

ii. $s_0 \asymp O(\log n)$

K.4(a): Each edge function $\phi_i, \phi_{i,j}$ in $f_0 \in \mathcal{F}_{s_0}^{m+1}$ is continuously differentiable (C^4) with $m \geq 3$ and with uniformly bounded derivatives.

K.4(b): The edge functions $\phi_i, \phi_{i,j}$ and their derivatives (up to order 2) are Lipschitz continuous.

Under these assumptions, the KAN sieve extremum estimator delivers the convergence rate: $\|\hat{\theta}_n - \theta_0\| = o_P(n^{-1/4})$. We note that while this estimator does not suffer from the COD as demonstrated in the proofs, d is naturally still relevant to our network parameters through s_0 and the subset $\mathcal{J}_i$. This is despite how estimation methods relying on splines are notorious for failing to produce precise estimates (Chen, 2007) in high-dimensional settings due to the COD. This will be demonstrated in future work extending the sieve KAN framework to structural semiparametric models, using the penalized sieve minimum distance estimator of Chen and Pouzo (2009) as a comparison. Although there are existing spline estimation

approaches that attempt to exploit sparsity in high-dimensional environments[20] similar to our proposed KAN sieve estimator, these methods are likely to perform poorly on the truly compositional structures (i.e., where imposing additivity is insufficient) found in economic and financial applications.

In practice, we do not apply entropy regularization, a method that penalizes the average magnitude of the activation functions commonly applied in KANs (see the approach outlined in Liu et al., 2024, Appendix C therein). While this may assist in reducing the number of redundant functions in general KAN architectures (in settings where $n \gg d$), this form of regularization effectively distributes the task of learning the structure of f_0 across the many potential paths. Instead, our bound Δ_n promotes the opposite, forcing the network to use each edge in the s_n active paths as efficiently as possible *without* sacrificing approximation power. As we demonstrate in the simulations, this remains true even as the data-generating process strays from strict $d \gg n$ cases.

The following assumption arises from the high-dimensional M-estimation framework of Negahban et al. (2012). Under the three general conditions (including a Restricted Strong Convexity (RSC) condition), their **Theorem 1** implies the oracle inequality in **Assumption S**. We formalize this connection in **Lemma 1** in the Appendix, where we then extend their result[21] to the case where s_0 grows at a polynomial rate in n. Arising from this proof is the condition $\zeta_s < 0.5$

Assumption S: (Sparse Recovery Condition) Given the quasi-log-likelihood function $l(\cdot)$ and regularization parameter λ_n, the KAN sieve estimator $\hat{\theta}_n$ that maximizes:

$$\frac{1}{n}\sum_{t=1}^{n} l(Z_t, \theta) - \lambda_n \left(\sum_{j=1}^{r_n} \sum_{k=1}^{G_n} |c_{j,k}| \right) \tag{13}$$

[20] e.g., Friedman (1991), Stone (1985)

[21] With an original proof accounting for mixing processes under **K.1**, see Yu (1994).

satisfies the following oracle inequality:

$$\mathbb{E}\left[\hat{\mathcal{L}}_n(\hat{\theta}_n) - \inf_{\theta \in \mathcal{K}_n} \hat{\mathcal{L}}_n(\theta)\right] \leq C_0 \left(\frac{s_0 \log p_n}{n} + \|\pi_n \theta_0 \ - \theta_0\|\right) \tag{14}$$

3 Theoretical Results

To simplify notation, we assume $\theta_0 \equiv f_0 \in \Theta$ and $\theta_n \equiv f(\cdot; \boldsymbol{\omega}_n) \in \Theta_n$ (i.e., A is an empty set; see our earlier parameter space definitions). A potential extension of the result from Chen and Shen (1998) concerning linear plug-in estimators is discussed in the following section. Proofs of each of the following theorems can be found in the Appendix, in addition to a demonstration of why these results do not extend to rates of convergence in the Sobolev norm.

3.1 Convergence Rates in the L_2 Norm

Theorem 1 (Sieve KAN Approximation Error): Under Assumptions **K.1-K.4**, there exists a sieve approximation $\pi_n \theta_0 \in \mathcal{K}_n$ such that under $G_n \asymp n^{\gamma}$ and given B-splines activation functions of fixed interval length (with uniform knots):

$$\|\pi_n \theta_0 - \theta_0\| \leq C s_0 G_n^{-4} = o(n^{-1/4})$$

The proof of **Theorem 1** relies on the sparse compositional structure imposed on the target function and the smooth univariate edge functions (Assumptions **K.3** and **K.4**). Using results from spline approximation theory (DeVore and Lorentz, 1993), we show that each edge function is approximated with error $O(G_n^{-k})$. The resulting approximation function of the sieve KAN is independent of d, a key result for mitigating the CoD. We note that due to the results from approximation theory applied in the proof of this theorem, we cannot apply the adaptive grid approach of Liu et al. (2024). However, due to our optimization

approach that transitions away from hybrid activations with a fixed component to B-splines alone ($w_s = 1$), this is not problematic. As discussed earlier, the constraint Δ_n on the spline coefficients leads to a similar result in practice.

Theorem 2 (Sieve KAN Entropy Bound): Under Assumptions **K.1-K.3**, and given sieve space $\mathcal{K}_n$:

$$\log \mathcal{N}(\epsilon, \mathcal{K}_n, \|\cdot\|_{L_2(\mu)}) \leq C_1 s_n \log\left(\frac{r_n}{s_n}\right) + C_2 s_n G_n \log\left(\frac{\Delta_n}{\epsilon}\right)$$

In this proof, the metric entropy bound is broken down into two components. The first, $s_n \log(r_n/s_n)$, is a combinatorial term arising from sparse edge selection (Raskutti et al., 2009) which is central to our sparse compositional form assumption. The B-spline complexity term $s_n G_n \log(\Delta_n/\epsilon)$ arises via the ℓ_1-constraint imposed on the spline coefficients (van der Vaart and Wellner, 1996). This highlights the role of Δ_n in ensuring sparse paths in the sieve KAN architecture and the dependence on effective parameters ($s_n G_n$) only, crucial for high-dimensional estimation.

Our approach differs from Zhang and Zhou (2024), where generalization bounds for KANs with activation functions represented by linear combinations of basis functions[22] are established for supervised learning tasks. Given covariates $X_i \in \mathbb{R}^d$ coefficient matrices B_i, Lipschitz constants C_j and (for layers $l \in \{1, \ldots, L\}$) an output space $\mathcal{H}_L(X) = \{\Psi_L \circ \cdots \circ \Psi_1(X) : \Psi_l \in \mathcal{F}_1\}$, they obtain the covering number bound:

$$\log \mathcal{N}\left(\mathcal{H}_L(X), \epsilon, \|\cdot\|_2\right) \leq \frac{\tilde{\alpha}^3 \log\left(2\tilde{d}\tilde{p}\right)}{\epsilon^2}$$

where $\tilde{\alpha}$[23] is a function of coefficients and Lipschitz constants, $\tilde{d} = \max_i d_i$ and $\tilde{p} = \max_i p_i$

[22] In addition to activation functions lying in a low-rank Reproducing Kernel Hilbert Space (RKHS). An extension of the sieve KAN framework involving activation functions in a low-rank RKHS is a promising direction for future research, in the context of both deriving tighter metric entropy bounds and obtaining rates of convergence in the Sobolev norm. This would entail shifting dependence on s_0 to rank r_i, potentially allowing for dimension-free rates in non-sparse settings; albeit ones with low intrinsic dimensions.

[23] Defined as $\tilde{\alpha} = \sum_{i=1}^{L} \alpha_i$ where $\alpha_i = B_i^{2/3} c_i^{2/3} (\Pi_{j=i+1} \ldots$

(the total number of univariate functions). In both Zhang and Zhou (2024) and Liu et al. (2024), the high-dimensional cases evaluated in their simulation analysis involve a sample size of $n = 10,000$ and $d = 100$ (versus $d = 10$ in the low-dimensional cases). From this covering number, it is easy to see how the metric entropy of general KANs grows uncontrollably through $\tilde{\alpha} \approx O(d^2)$ under large widths ($W \asymp d$) and unconstrained B-spline coefficients in truly high-dimensional settings (e.g., $d \gg n$) due to the CoD.

Theorem 3 (Sieve KAN Estimation Error): Under Assumptions **K.1-K.3**, given sieve space $\mathcal{K}_n$ and letting $\frac{1}{16} < \gamma < 1$:

$$\|\hat{\theta}_n - \pi_n \theta_0\| = o_P(n^{-1/4})$$

The proof of **Theorem 3** involves splitting the entropy integral at $\epsilon_0 = n^{-1}$ at showing that each term is $o_P(n^{-1/4})$. Obtaining convergence rates in the Sobolev norm would lead to a condition that $m > d$ as we show in the Appendix, which would re-introduce the CoD. Instead, the estimation error achieves a convergence rate of $o_P(n^{-1/4})$ which holds uniformly over $d = O(n^{\zeta_d})$. Combining this result with the result from **Theorem 1** earlier, we have:

Corollary 1 (Sieve KAN Convergence Rate): Under Assumptions **K.1-K.4**, the KAN sieve extremum estimator achieves $\|\hat{\theta}_n - \theta_0\| = o_P(n^{-1/4})$.

To see this, note that by the triangle inequality:

$$\|\hat{\theta}_n - \theta_0\| \leq \|\hat{\theta}_n - \pi_n \theta_0\| + \|\pi_n \theta_0 - \theta_0\| \tag{15}$$

$$\|\hat{\theta}_n - \theta_0\| = o_P(n^{-1/4}) + o(n^{-1/4}) = o_P(n^{-1/4}) \tag{16}$$

Remark: We note that by letting $\omega_n \to \infty$ and $\Delta_n \to \infty$ as $n \to \infty$, we obtain a sequence of increasingly flexible network architectures; making this analogous to the result from White (1990) for KANs under sparsity, a foundational result originally in terms of universal approximator MLPs. In establishing the consistency of sieve aNN estimators, White

(1990) demonstrates how the approximation power of the network increases in tandem with the "experience" of the aNN[24] while controlling network complexity sufficiently well. In the context of sparse compositional target functions, our framework represents a class of *connectionist sieve networks* with multiple hidden layers that guarantees an (explicitly) asymptotic convergence rate, the first of its kind.

Theorem 4 (Asymptotic Normality):

3.2 Double/Debiased Machine Learning Example

Chernozhukov et al. (2018) address how the naive application of machine learning algorithms in semiparametric models with low-dimensional causal parameters can lead to biased estimates. This arises from the regularization methods commonly employed in these models for nuisance function estimation, which perform well by reducing variance (overfitting) in a manner consistent with the standard variance-bias tradeoff. As a result of this overfitting and regularization bias, estimates of the causal parameter obtained by using the fitted nuisance function as a plug-in estimator will suffer from a heavy degree of bias. Furthermore, they demonstrate how the Donsker conditions typically imposed in semiparametric models to control complexity of the parameter break down in high-dimensional settings. The DML framework resolves these issues through two key ingredients: Neyman orthogonality and sample splitting.

We consider the partially linear regression model (Robinson, 1988) in our simulations. This model is defined in terms of the following equations:

$$Y = D\alpha_0 + g_0(X) + U, \qquad E[U|X, D] = 0 \tag{17}$$

$$D = m_0(X) + V, \qquad E[V|X] = 0 \tag{18}$$

where the sample splitting procedure (in the case of two folds) entails estimating m_0 using

[24]Referring to the sample size of the dataset the network is trained on

half of the dataset and then estimating g_0 with the remaining half. Neyman orthogonality is a condition that ensures moment conditions defining α_0 are insensitive to perturbations in the nuisance functions η_0.

Definition 1 (Neyman Orthogonality): The score $\psi(\cdot;\cdot)$ satisfies Neyman orthogonality if:

$$\partial_\eta E[\psi(Z;\alpha_0,\eta_0)][\eta-\eta_0]=0 \quad \text{for all } \eta\in \mathrm{T}_n$$

The main advantage of our procedure estimating nuisance functions using sieve KANs over widely used machine learning methods (e.g., Lasso, random forests) is how it is capable of capturing complex compositional structures in the data-generating process (DGP). In their empirical example, Chernozhukov et al. (2018) discuss how "Deep Learning methods" were experimented with but how the results were omitted due to computational and stability issues. We encounter no such issues across all of our Monte Carlo simulations and compare the results against those of a Lasso model using the same ℓ_1 penalty, λ_n.

Corollary 2 (DML Estimation): Under Assumptions **K.1(a)-K.4** and Assumption **S**, the DML estimator $\hat{\alpha}$ where $(\hat{g}_n,\hat{m}_n)\in\mathcal{K}_n$ satisfies:

$$\sigma^{-1}\sqrt{n}(\hat{\alpha}-\alpha_0)\rightsquigarrow N(0,1), \quad \sigma^2=(E[V^2])^{-1}E[V^2U^2](E[V^2])^{-1}$$

which is a result that holds when σ^2 is replaced with $\hat{\sigma}^2$ (Theorem 4.1, Chernozhukov et al. (2018)). A formal verification of Assumption **3.2** from Chernozhukov et al. (2018), which imposes the rate requirement $\epsilon_n=o_P(n^{-1/4})$ for the upper bound on the rate of convergence of $\hat{\eta}$ under the $L_2(\mu)$ norm, is left to the Appendix.

4 Results

4.1 Monte Carlo Simulations

In our simulations, we compare the performance of sieve KANs against Lasso in estimating $\alpha_0 = 1.5$. Across each DGP, we generate $d = 500$ covariates in total with $S = 20$ relevant covariates and set use five folds in the sample spliting procedure. After implementing this split for each fold and an additional split to obtain the validation set, we are left with a training sample size of $n_{\text{train}} = 360$ observations for the first two DGPs and $n_{\text{train}} = 1440$ for the DGP #3. The validation set serves an important purpose in making sure our fitted models generalize well to unseen data and we report these loss metrics. We fix the sieve KAN parameter determining the grid size as $\gamma = 0.5$ for all simulations.

In each DGP, the true partially linear regression model is generated using the following nonlinear sparse functions and computed orthogonal residuals:

$$m_0(\boldsymbol{X}) = \sum_{j=1}^{S} \frac{0.5}{j+1} \sin(X_j) + \epsilon \tag{19}$$

$$g_0(\boldsymbol{X}) = \sum_{j=1}^{S} \frac{0.3}{j+1} X_j^2 + \nu \tag{20}$$

$$\epsilon \sim N(0, 0.25), \quad \nu \sim N(0, 0.25) \tag{21}$$

where $\hat{V} = D - \hat{m}(\boldsymbol{X})$, and $\hat{U} = Y - \hat{g}(\boldsymbol{X})$ are used to obtain $\hat{\alpha} = \frac{\sum \hat{V}_i (Y_i - \hat{g})}{\sum \hat{V}_i D_i}$.

In DGP #1, we find a bias of 0.0055 (i.e., $\hat{\alpha} = 1.5055$), significantly lower than that of the Lasso model with a similar variance. Holding all else constant and significantly increasing the width of the sieve KAN, we obtain identical results with the exception of the ratio of spline coefficients driven near zero in our learned activation functions. This is in line with expectations as the true signal in DGP #2 does not change and demonstrates how overparameterization of W_n does not impact the approximation power of the sieve KAN. The bias of this model is further reduced in DGP #3, which results in a significantly lower

Table 1: Simulation Results: Sieve KAN vs. Lasso in DML ($\alpha_0 = 1.5$)

Simulation	Method	n	d	W_n	Bias	RMSE	Sparsity	Validation Loss	95% CI Coverage
DGP #1	KAN	500	500	20	0.0055	0.0736	0.761	0.9427	[1.288, 1.723]
	Lasso	500	500	20	-0.1338	0.1537	0.989	1.2628	[1.120, 1.613]
DGP #2	KAN	500	500	50	0.0055	0.0744	0.881	0.9536	[1.288, 1.723]
	Lasso	500	500	50	-0.1367	0.1587	0.992	1.2633	[1.122, 1.605]
DGP #3	KAN	2000	500	50	0.0008	0.0352	0.986	0.9590	[1.392, 1.609]
	Lasso	2000	500	50	-0.1331	0.1387	0.989	1.2079	[1.235, 1.498]

Notes: Results based on 1000 simulations per DGP. "Sparsity" refers to the fraction of zero parameters (i.e., spline coefficients driven close to zero in the sieve KAN case) relative to the total number.

standard error of $\hat{\alpha}$ when increasing the sample size to $n = 2000$ ($n_{\text{train}} > d$). The relative increase in the "Sparsity" ratio arises from the grid size doubling (G_n under $\gamma = 0.5$) and the additional layers included under this larger sample size (holding $S = 20$ constant). We confirm that the sieve KAN generalizes well to unseen data, as the validation loss is consistently close to but less than 1.0000. The bias arising from underfitting in the Lasso model under λ_n persists across DGPs and the 95% confidence interval fails to cover α_0 in DGP #3. This results in many coefficients of relevant covariates being driven to zero, particularly in the case of $\hat{m}$ due to the nonlinearity in the true model.

4.2 Conditional Density Estimation Example

We now consider a conditional density estimation (CDE) example to motivate a discussion on the shortcomings of deep MLPs in high-dimensional settings, as well as how these are addressed by sieve KANs. Suppose we seek to employ a deep MLP in a sparse estimation problem involving covariates with high-dimensions and naively attempt to implement ℓ_1 regularization (or the more common "elastic net" method combining it with ℓ_2 regularization). This applies a penalty $\sum_{i,j} |w_{ij}|$ on all weights across every layer, in theory reducing the contributions of weights significantly influenced by irrelevant features. Achieving this in practice with single hidden-layer networks is more feasible relative to DNNs[25], due to

[25]See Shen et al. (2023).

the complex interconnected nature of the weights in the latter architecture. Consider the following function:

$$f_0(x_1, x_2, \ldots x_{10}) = -3.0 + 2\sin(x_1) + 0.1(x_2^2) + \epsilon, \quad \epsilon \sim N(0, 0.25)$$

where the true density of the relevant variables (x_1 and x_2) takes on a bullseye shape, $d = 10$ and $n = 10,000$.

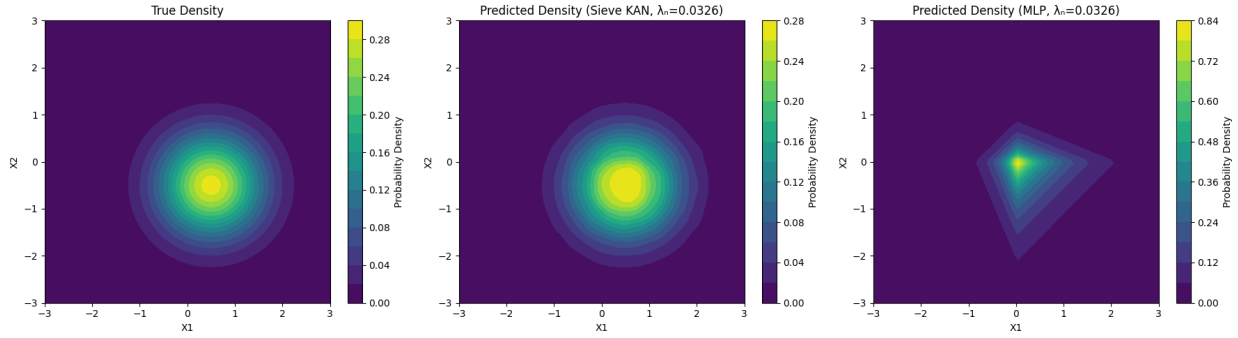

Figure 2: CDE Example Predicted Densities

As seen in Figure 2, a deep MLP (in this case, a ReLU DNN) applying ℓ_1 regularization will struggle to learn the shape of the true density. However in the absence of this regularization, the MLP is more prone to overfitting in noisy DGPs[26], highlighting a key tension in the potential application of deep MLPs in high-dimensions. As for our sieve KAN, the predicted density is much closer to the true density in terms of shape and scale. We note that due to how $n \gg d$, this prediction can likely be improved by increasing λ_n and/or G_n through γ (initially set to $\gamma = 0.3$ to obtain a grid size similar to those in DGP #1 and #2 in the earlier simulations).

[26]See Schmidt-Hieber (2020)

References

[1] Belloni, A., & Chernozhukov, V. (2011). L1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics*, *39*(1), 82–130.

[2] Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of Econometrics*, *6B*, 5549-5632. Elsevier.

[3] Chen, X., Linton, O., & Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, *71*(5), 1591-1608.

[4] Chen, X., & Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, *66*(2), 289-314.

[5] Chen, X., & White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, *45*(2), 682-691.

[6] DeVore, R. A. (1998). Nonlinear approximation. *Acta Numerica*, *7*, 51-150.

[7] Farrell, M. H., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, *89*(1), 181-213.

[8] Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1-67.

[9] Grenander, U. (1981). *Abstract Inference*. Wiley, New York.

[10] Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359-366.

[11] Hornik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, *3*(5), 551-560.

[12] Koenker, R. (2005). *Quantile regression* (Vol. 38). Cambridge university press.

[13] Kolmogorov, A. N. (1956). On the representation of continuous functions of several variables as superpositions of continuous functions of a smaller number of variables. *Dokl. Akad. Nauk*, *108*(2), 179-182.

[14] Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., ... & Tegmark, M. (2024). KAN: Kolmogorov-Arnold networks. *arXiv preprint arXiv:2404.19756*.

[15] Meier, L., Van de Geer, S., & Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics*, *37*(6B), 3730-3751.

[16] Mhaskar, H. N., & Poggio, T. (2016). Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, *14*(06), 829-848.

[17] Raskutti, G., Wainwright, M. J., & Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, *57*(10), 6976-6994.

[18] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, *48*(4), 1851-1875.

[19] Schmidt-Hieber, J. (2021). The Kolmogorov–Arnold representation theorem revisited. *Neural Networks*, *137*, 119-126.

[20] Shen, X., Jiang, C., Sakhanenko, L., & Lu, Q. (2023). Asymptotic properties of neural network sieve estimators. *Journal of nonparametric statistics*, *35*(4), 839-868.

[21] Sprecher, D. A. (1996). A numerical implementation of Kolmogorov's superpositions. *Neural networks*, *9*(5), 765-772.

[22] Sprecher, D. A. (1997). A numerical implementation of Kolmogorov's superpositions II. *Neural networks*, *10*(3), 447-457.

[23] Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science* (Vol. 47). Cambridge University Press.

[24] White, H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, *3*(5), 535-549.

[25] White, H. (1992). Nonparametric estimation of conditional quantiles using neural networks. In C. Page & R. LePage (Eds.), *Computing Science and Statistics* (pp. 190-199). Springer.

[26] Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, *94*, 103-114.

[27] Yarotsky, D. (2018). Optimal approximation of continuous functions by very deep ReLU networks. In *Conference on Learning Theory* (COLT 2018) (pp. 639-649). PMLR.

[28] Zhang, X., & Zhou, H. (2024). Generalization bounds and model complexity for kolmogorov-arnold networks. *arXiv preprint arXiv:2410.08026*.

5 Appendix

5.1 Proofs of Main Results

Proof of Theorem 1: Under **K.3(a)**, θ_0 admits a sparse compositional structure with s_0 active edges given by ϕ_q and $\phi_{q,p}$ which are sufficiently smooth (**K.3(b)**) to be approximated by B-splines of order $k = 4$. Let $\delta_T = \max_{0 \leq j \leq N_j} t_{j+1} - t_j$ where the number of intervals is given by $N_j = G_n - k + 1$. By Theorem 7.3 of DeVore and Lorentz (1993) we obtain an approximation error of $O(G_n^{-4})$, since $\delta_T = O(G_n^{-1})$.

The constants arising from errors terms propagated across each layer[27] (e.g., Lipschitz constants) in addition to smoothness parameter m are absorbed by the constant $C > 0$. To see this, consider the following bound of the $L_2(\mu)$ error of the approximation of $\phi \in C^k$ (DeVore and Lorentz, 1993):

$$\|\phi - \pi_n \phi\| \leq C(\|\phi^k\|_\infty) G_n^{-k}$$

Since the overall approximation error accumulates across every active edge, it is bounded by $Cs_0 G_n^{-4}$. In the case of logarithmic sparsity growth where $s_0 = \log n$ and $W_n = C_W \log n$, we have

$$\|\pi_n \theta_0 - \theta_0\| = O(\log n \cdot n^{-4\gamma})$$

For this approximation error term to be $o(n^{-1/4})$, we require $-4\gamma < -\frac{1}{4} \Rightarrow \gamma > \frac{1}{16}$ (the first of two key conditions on γ) guaranteeing the following limit

$$\lim_{n \to \infty} \frac{\log n \cdot n^{-4\gamma}}{n^{-1/4}} = 0$$

noting that the polynomial term dominates here and the case where $s_0 = O(1)$ follows trivially. In the case where $s_0 = O(n^{\zeta_s})$, this condition becomes $\gamma > \frac{\zeta_s}{4} + \frac{1}{16}$.

[27] See Schmidt-Heiber (2021)

Proof of Theorem 2: In this proof we account for two components, the term arising from selection of (active) edges and the basis function approximation. For the former, we have the term $s_n \log\left(\frac{r_n}{s_n}\right)$ which is a generalization bound of the combinatorial complexity of sparse function classes (Theorem 2 of Raskutti et al., 2009); highlighting the s_n active edges from the r_n potential edges in total. As for the latter, we note that across each active edge, the activation function is approximated by G_n basis functions with ℓ_1 constrained coefficients. By Lemma 2.6.11 of van der Vaart and Wellner (1996), the complexity of these coefficients within an ℓ_1 ball is bounded by $s_n G_n \log\left(\frac{\Delta_n}{\epsilon}\right)$. Given constants $C_1 > 0$ and $C_2 > 0$ and combining both components, the metric entropy is bounded.

Proof of Theorem 3: We begin by letting $\delta_n = n^{-1/4-\eta}$ for some arbitrarily small $\eta > 0$ and verify the following entropy integral condition:

$$\int_0^{\delta_n} \sqrt{\frac{\log \mathcal{N}(\epsilon, \mathcal{K}_n, \|\cdot\|_{L_2(\mu)})}{n}} d\epsilon \leq C_3 \delta_n^2$$

From **Theorem 2**, we have: $\log \mathcal{N}(\epsilon, \mathcal{K}_n, \|\cdot\|_{L_2(\mu)}) \leq C_1 s_n \log\left(\frac{r_n}{s_n}\right) + C_2 s_n G_n \log\left(\frac{\Delta_n}{\epsilon}\right)$, consider the case where $s_n \asymp s_0 = O(\log n)$ for which we defined $W_n = C_W \log n$ accordingly earlier. Then we have $r_n = O((\log n)^3)$ by definition, leading to the expression:

$$\log \mathcal{N}(\epsilon, \mathcal{K}_n, \|\cdot\|_{L_2(\mu)}) \leq C(\log n) \log\left(\frac{O((\log n)^3)}{O(\log n)}\right) + C'(\log n) n^\gamma \log(\Delta_n/\epsilon) \tag{22}$$

$$= O(\log n \log(\log n)) + O(\log n \cdot n^\gamma \log(\Delta_n/\epsilon)) \tag{23}$$

where in the case of $s_0 = O(1)$, the RHS term is simply equal to $O(\log(\log n)) + O(n^\gamma \log(\Delta_n/\epsilon))$. In both cases, the dominant term of $\log \mathcal{N}(\epsilon, \mathcal{K}_n, \|\cdot\|_{L_2(\mu)})$ is $O(s_0 G_n \log(\Delta_n/\epsilon))$. We decompose this by splitting the integral at $\epsilon_0 = n^{-1}$ and show that both terms are $o_P(n^{-1/4})$:

$$\int_0^{\delta_n} \sqrt{\frac{\log \mathcal{N}(\epsilon, \mathcal{K}_n, \|\cdot\|_{L_2(\mu)})}{n}} d\epsilon = \underbrace{\int_0^{\epsilon_0} \cdots d\epsilon}_{(I)} + \underbrace{\int_{\epsilon_0}^{\delta_n} \cdots d\epsilon}_{(II)} \tag{24}$$

Term (I):

$$(I) \leq \epsilon_0 \sqrt{\frac{\log \mathcal{N}(\epsilon_0, \mathcal{K}_n, \|\cdot\|_{L_2(\mu)})}{n}} \tag{25}$$

For $s_0 = O(\log n)$, we have $\log \mathcal{N}(\epsilon_0, \mathcal{K}_n, \|\cdot\|_{L_2(\mu)}) = O(\log n \cdot n^\gamma \log n) = O(n^\gamma (\log n)^2)$. Therefore:

$$(I) = O\left(n^{-1}\sqrt{n^\gamma (\log n)^2}\right) = O\left(n^{\frac{\gamma}{2}-1}\log n\right)$$

and in order for this term to be $o_P(n^{-1/4})$, we require $\frac{\gamma}{2} - 1 < -1/4 \implies \frac{\gamma}{2} < 3/4 \implies \gamma < 3/2$, which is satisfied by $\gamma < 1$. The same condition arises when $s_0 = O(1)$ since we end up with $(I) = O\left(n^{\frac{\gamma}{2}-1}(\log n)^{1/2}\right)$ for this case.

Term (II):

Using the aforementioned dominant term of $\log \mathcal{N}(\epsilon, \mathcal{K}_n, \|\cdot\|_{L_2(\mu)})$, we have:

$$\begin{aligned}
(II) &\leq \sqrt{\frac{Cs_0}{n}} \int_{\epsilon_0}^{\delta_n} \sqrt{G_n \log(\Delta_n/\epsilon)} d\epsilon && (26)\\
&= O\left(\sqrt{\frac{s_0}{n}}\sqrt{G_n}\delta_n \sqrt{\log(\Delta_n/\delta_n)}\right) && (27)\\
&= O\left(n^{\frac{\zeta_s}{2}-\frac{1}{2}}(n^\gamma)^{1/2} n^{-1/4-\eta}(\log n)^{1/2}\right) \quad (\text{where } \zeta_s \to 0 \text{ for } s_0 = O(\log n)) && (28)\\
&= O\left(n^{\frac{\zeta_s+\gamma}{2}-\frac{3}{4}-\eta}(\log n)^{1/2}\right) && (29)
\end{aligned}$$

For term (II) to be $o_P(n^{-1/4})$, we require $\frac{\zeta_s+\gamma}{2} - \frac{3}{4} - \eta < -\frac{1}{4} \implies \frac{\zeta_s+\gamma}{2} < \frac{1}{2} + \eta \implies \zeta_s + \gamma < 1 + 2\eta$. For the cases considered here we have $\zeta_s = 0$ for $s_0 = O(1)$ and $\zeta_s \to 0$ for $s_0 = O(\log n)$ (by L'Hopital's rule, $\forall \zeta_s > 0$). Thus for a sufficiently small $\eta > 0$, the key condition for guaranteeing our $o_P(n^{-1/4})$ rate is that $\gamma < 1$ (or $\zeta_s + \gamma < 1$ under polynomial sparsity growth). By Theorem 1 of Chen and Shen (1998), we conclude that $\|\hat{\theta}_n - \pi_n \theta_0\| = o_P(n^{-1/4})$.