Kolmogorov-Arnold Networks

jiahong.long@

Cruise · UC San Diego

10 June 2024

Welcome to the inaugural Sim-Eval literature review!

I'm trying to keep this informal – teaching is the best way of learning, and we all benefit from sharing knowledge.

My background is in math, so expect a bit more of the underlying theory here!

On the agenda:

- MLPs and their limitations
- What KANs do better
- What KANs do worse

Some bookkeeping notes:

For simplicity, assume that the functions we care about are *scalar-valued* and *of multiple variables*, i.e.

$$f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$$

Parts of this are intentionally abstract, so as to not get into the weeds of technical examples. Expect a fair bit of hand-waving.

The original paper is on `ArXiv:2404.19756v1` at `https://arxiv.org/pdf/2404.19756v1`.

Primer: a gentle (re)introduction to the multi-layer perceptron

Canonically, a "shallow" multi-layer perceptron is *two layers*. (This will be important for the literature!) A multi-layer perceptron in the shallow case is represented as

$$\hat{f}\colon\ \mathbb{R}^n \longrightarrow \mathbb{R}$$

$$\hat{f}(x) = \sum_{i}^{N(\varepsilon)} a_i \sigma \left[\mathbf{W}_i \mathbf{x} + b_i\right]$$

Note the width $N(\varepsilon)$ is a function of the precision ε.

$$\hat{f}(x) = \sum_{i}^{N(\varepsilon)} a_i \sigma \left[\mathbf{W}_i \mathbf{x} + b_i \right]$$

where:

b_i	is the bias *(affine!)*
$\mathbf{x}$	is the i th input vector
$\mathbf{W}_i$	is the i th weight matrix *(learned!)*
σ	is the activation function *(e.g. ReLU, sigmoid,* tanh*...)*
a_i	are elements of the outermost weight matrix
$N(\varepsilon)$	is the number of neurons

In the shallow case, a_i can be denoted by a row vector.

$$\hat{f}(x) = \sum_{i}^{N(\varepsilon)} a_i \sigma \left[\mathbf{W}_i \mathbf{x} + b_i \right]$$

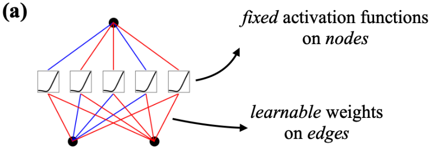

Figure: *A shallow 2-layer multi-layer perceptron.* $d = 2$, $n = 5$.

Interesting side note: more generally, for deep ($d > 2$) networks, a multi-layer perceptron is really just a composition of (affine!) linear transformations separated by non-linear activations.

$$MLP(\mathbf{x}) = [\mathbf{W}_d \circ \sigma_d \circ \mathbf{W}_{d-1} \circ \sigma_{d-1} \circ \ldots \mathbf{W}_1 \circ \sigma_1] (\mathbf{x})$$

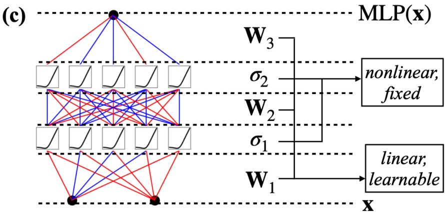

Figure: *A deep multi-layer perceptron.*

$$\hat{f}(x) = \sum_{i}^{N(\varepsilon)} a_i \sigma \left[\mathbf{W}_i \mathbf{x} + b_i\right]$$

Question: What can a multi-layer perceptron represent?
Answer Anything!*

Question: What can a multi-layer perceptron represent?
Answer Anything!*

Let $D \subset \mathbb{R}^n$ be compact[1]. $f \colon D \longrightarrow \mathbb{R}$ be an *arbitrary nonlinear function*, and let $\hat{f} \colon D \longrightarrow \mathbb{R}$ denote a shallow (*n.b.* 2-layer) multi-layer perceptron, denoted by

$$\hat{f}(x) = \sum_{i}^{N(\varepsilon)} a_i \sigma \left[\mathbf{W}_i \mathbf{x} + b_i \right]$$

where $N(\varepsilon)$ is the *number of neurons*. In the shallow case this is $=$ the width.

[1] Compact denotes some notion of "closed and bounded" – the n-dimensional equivalent of a closed interval $[a, b] \subset \mathbb{R}$.

Let $D \subset \mathbb{R}^n$ be compact. $f\colon D \longrightarrow \mathbb{R}$ be an *arbitrary nonlinear function*, and let $\hat{f}\colon D \longrightarrow \mathbb{R}$ denote a shallow (*n.b.* 2-layer) multi-layer perceptron, denoted by

$$\hat{f}(x) = \sum_{i}^{N(\varepsilon)} a_i \sigma \left[\mathbf{W}_i \mathbf{x} + b_i \right]$$

Theorem

Universal approximation (2-layer network). *For arbitrary $\varepsilon \in \mathbb{R} > 0$, there exists $N(\varepsilon)$ such that*

$$|f(x) - \hat{f}(x)| \leq \varepsilon$$

$$\hat{f}(x) = \sum_{i}^{N(\varepsilon)} a_i \sigma \left[\mathbf{W}_i \mathbf{x} + b_i\right]$$

Theorem

Universal approximation (2-layer network). *For arbitrary $\varepsilon \in \mathbb{R} > 0$, there exists $N(\varepsilon)$ such that*

$$|f(x) - \hat{f(x)}| \leq \varepsilon$$

So we can model basically anything! But...

Theorem

Universal approximation (2-layer network). *For arbitrary* $\varepsilon \in \mathbb{R} > 0$, *there exists* $N(\varepsilon)$ *such that*

$$|f(x) - \hat{f(x)}| \leq \varepsilon$$

The million-dollar question:

What is $N(\varepsilon)$?

What is $N(\varepsilon)$?

We don't know, in general! The universal approximation theorem guarantees no bounds on N. For deep networks, we do know that it's possibly poorly behaved ($N \propto \exp(d)$, the layer depth of the network).

Why does this make sense? Because we are fitting a "mostly linear" model to an "arbitrary non-linear" function. We need "a lot of linear pieces" to get good at modeling funky nonlinear functions.

Figure: *It takes a lot of line segments to approximate this quadratic, and as soon as we leave $[-3, 3]$, the error in our approximation blows up!*

Enter the notion of *neural scaling laws.*

Neural scaling laws formalize the notion of "mostly linear things approximate nonlinearities inefficiently" – we can generally say that the *training*[2] loss ℓ decreases according to

$$\ell \propto N^{-\alpha}$$

where α is the *scaling exponent.*

[2] i.e. overfits