Elementary Superexpressive Activations

Dmitry Yarotsky

Skoltech
Skolkovo Institute of Science and Technology

ICML 2021

Fixed-size neural network can approximate functions with **arbitrary accuracy**

Maiorov-Pinkus (1999): There exists an analytic, sigmoidal and strictly increasing activation function σ such that **any** $f \in C([0,1]^d)$ can be approximated with **arbitrary accuracy** by a two-hidden-layer network of a **fixed size**:

$$\widehat{f}(\mathbf{x}) = \sum_{i=1}^{6d+3} a_i \sigma\Big(\sum_{j=1}^{3d} c_{ij} \sigma(\mathbf{w}^{ij} \cdot \mathbf{x} + \theta_{ij}) + \gamma_i \Big)$$

Note: activation function σ is **very complex, non-elementary**

Our results

1. Simple examples of **elementary** activations with a similar "superexpressiveness" property
2. Proof that most commonly used activations are not "superexpressive"

A related previous work

Shen, Yang & Zhang (2020): A three-layer network using the floor $\lfloor\cdot\rfloor$, the exponential 2^x and the step function $\mathbf{1}_{x\geq 0}$ as activations can approximate Lipschitz functions with an exponentially small error $O(e^{-cW})$, where W is the number of weights.

"Superexpressiveness": a formal definition

Let $\mathcal{A}$ be a family of univariate activation functions

Different neurons can be equipped with different activations from $\mathcal{A}$

Call $\mathcal{A}$ **superexpressive** if:

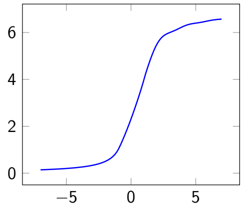
- for any d, there exists a fixed d-input network architecture with a fixed number of neurons using activations from $\mathcal{A}$ so that any $f \in C([0,1]^d)$ can be approximated with any accuracy in the uniform norm $\|\cdot\|_\infty$ by such a network

Elementary superexpressive activations

Main Theorem 1: Each of the following families of activation functions is superexpressive:

$$\mathcal{A}_1 = \{\sigma_1, \lfloor\cdot\rfloor\},$$
$$\mathcal{A}_2 = \{\sin, \arcsin\},$$
$$\mathcal{A}_3 = \{\sigma_3\},$$

σ_3



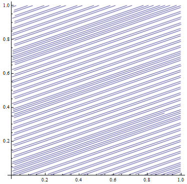
where σ_1 is any function that is real analytic and non-polynomial in some interval $(\alpha, \beta) \subset \mathbb{R}$, and

$$\sigma_3(x) = \begin{cases} -\frac{1}{x}, & x < -1, \\ \frac{1}{\pi}\left(x \arcsin x + \sqrt{1-x^2}\right) + \frac{3}{2}x, & x \in [-1, 1], \\ 7 - \frac{3}{x} + \frac{\sin x}{\pi x^2}, & x > 1. \end{cases}$$

The function σ_3 is $C^1(\mathbb{R})$, bounded, and strictly monotone increasing.

Key proof ideas

- Have a periodic piecewise linear function in the network
- If not directly available, generate such a function by superpositions, multiplications and differentiations
- Divide the domain $[0,1]^d$ into M sub-domains and map them to M points $1,\ldots,M$, with a large M
- Use the density of an irrational winding on the M-dimensional torus to fit the values at the points $1,\ldots,M$



(https://en.wikipedia.org/wiki/Linear_flow_on_the_torus)

Absence of superexpressiveness

Main Theorem 2. Let $\mathcal{A}$ be a family of finitely many **piecewise Pfaffian** activation functions. Then $\mathcal{A}$ is not superexpressive.

Pfaffian functions[1]

- A **Pfaffian chain**: a sequence $f_1, \ldots, f_l$ of real analytic functions on a common connected domain $U \subset \mathbb{R}^d$ such that

$$\frac{\partial f_i}{\partial x_j}(\mathbf{x}) = P_{ij}(\mathbf{x}, f_1(\mathbf{x}), \ldots, f_i(\mathbf{x})), 1 \leq i \leq l,\ 1 \leq j \leq d$$

  for some polynomials P_{ij}
- A **Pfaffian function** in the chain $f_1, \ldots, f_l$: a function on U that can be expressed as a polynomial P in the variables $(\mathbf{x}, f_1(\mathbf{x}), \ldots, f_l(\mathbf{x}))$
- **Complexity** of the Pfaffian function f: the triplet (l, α, β) consisting of the length l of the chain, the maximum degree α of the polynomials P_{ij}, and the degree β of the polynomial P

All **elementary** functions are Pfaffian when considered on suitable domains

[1] Khovanskii, Fewnomials (1991)

Examples and properties of Pfaffian functions

The following functions are Pfaffian:

1. polynomials on $U = \mathbb{R}^d$,
2. e^x on $\mathbb{R}$,
3. $\ln x$ on $\mathbb{R}_+$,
4. $\arcsin x$ on $(-1, 1)$,
5. $\sin x$ is Pfaffian on any bounded interval (A, B), with complexity depending on $B - A$.

But $\sin x$ is **not** Pfaffian on whole $\mathbb{R}$!

The solution bound for Pfaffian functions

Theorem (Khovanskii). Let $f_1, \ldots, f_d$ be Pfaffian d-variable functions with a common Pfaffian chain on a connected domain U. Then the number of nondegenerate solutions of the system $f_1(\mathbf{x}) = \ldots = f_d(\mathbf{x}) = 0$ is bounded by a finite number only depending on the complexities of the functions $f_1, \ldots, f_d$.

Common practical activations are not superexpressive

Call an activation σ **piecewise Pfaffian** if its domain can be divided into **finitely many** intervals on which σ is Pfaffian.

Most practical activations are piecewise Pfaffian, e.g.:

- $\sigma(x) = \tanh x$
- $\sigma(x) = (1 + e^{-x})^{-1}$ (standard sigmoid)
- $\sigma(x) = \max(0, x)$ (ReLU)
- $\sigma(x) = \max(ax, x)$ (leaky ReLU)
- $\sigma(x) = e^{-x^2}$ (Gaussian)
- $\sigma(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$ (step function)
- $\sigma(x) = \ln(1 + e^x)$ (softplus)
- $\sigma(x) = \begin{cases} a(e^x - 1), & x < 0 \\ x, & x \geq 0 \end{cases}$ (ELU)