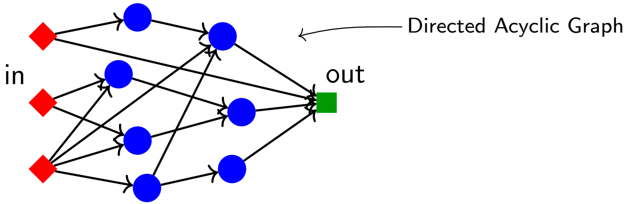
Expressiveness of neural networks

Dmitry Yarotsky

Feedforward neural networks



Implements a map $y = \widetilde{f}(\mathbf{x}, \mathbf{W}) \equiv \widetilde{f}_{\mathbf{W}}(\mathbf{x})$

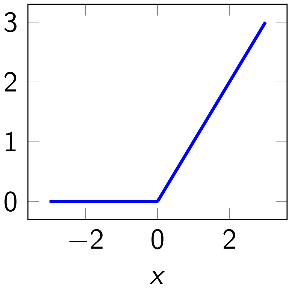
- $\mathbf{x} = (x_1, \ldots, x_\nu) \in \mathbb{R}^\nu$: input vector
- $\mathbf{W}$: the collection of all **network weights** (all tunable parameters)
- y: the (scalar) output
- A **neuron** in a hidden layer: $z_1, \ldots, z_d \mapsto \sigma\big(\sum_{m=1}^d w_m z_m + h\big)$
- Weights in a neuron: $\{w_m\}_{m=1}^d, h$ (depend on the neuron)
- σ: a (nonlinear) **activation function**
- The output neuron: $z_1, \ldots, z_d \mapsto \sum_{m=1}^d w_m z_m + h$ (no activation)

Some common activation functions

Exercise: Why does σ need to be nonlinear?

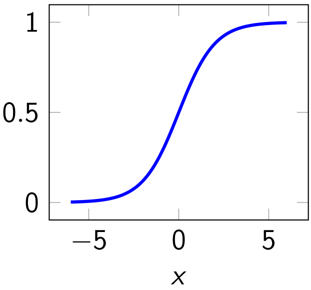
ReLU (Rectified Linear Unit)

$$\sigma(x) \equiv (x)_+ = \max(0, x)$$



Standard sigmoid

$$\sigma(x) = 1/(1 + e^{-x})$$



Piecewise linear activation functions

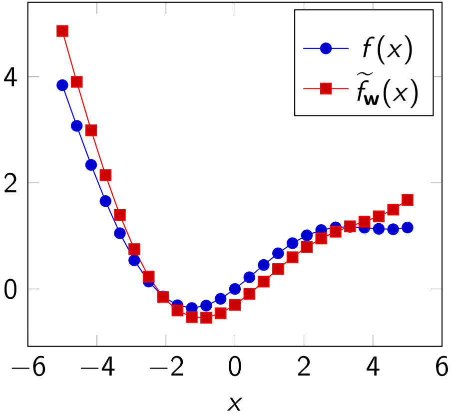
Exercise (equivalence of piecewise linear activation functions)

- Suppose that $f : \mathbb{R}^\nu \to \mathbb{R}$ is implemented by a NN with some continuous piecewise-linear activation function. Then f can also be implemented by a ReLU NN (possibly of a different architecture).
- Suppose that $f : [0,1]^\nu \to \mathbb{R}$ is implemented by a ReLU NN. Then, for any given continuous piecewise-linear activation function σ_1, f can be implemented by a σ_1-NN.

Network fitting (approximation): general idea

We try to adjust the weight vector $\mathbf{W}$ so that our network becomes close to a "ground truth" map f:

$$\widetilde{f}_{\mathbf{w}} \approx f$$



The concept of expressiveness

General idea: When the weights and possibly the architecture are varied, how significantly varies $\widetilde{f}$? How rich is the set of $\widetilde{f}$'s?

Refinements:

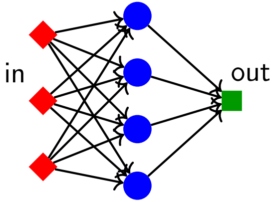
- (Regression) How efficiently can we approximate the given map $f : [0,1]^\nu \to \mathbb{R}$ by NN's?
- (Classification) How big is the set of Boolean maps $\widetilde{f} : X \to \{-1,+1\}$ implementable by NN's?
- (for ReLU networks) How many linear pieces can the function $\widetilde{f}(\mathbf{x})$ have?
- (Topology) How many connected components can $\widetilde{f}^{-1}(y)$ have?
- ...

Expressiveness = F(Network complexity)

A good survey: A. Pinkus, Approximation theory of the MLP model in neural networks, 1999

A one-hidden-layer network:

$$\widetilde{f}(\mathbf{x}) = \sum_{n=1}^{N} c_n \sigma\Big(\sum_{k=1}^{\nu} w_{nk} x_k + h_n\Big) + h$$



$\nu = 3, N = 4$

Uniform approximation on compact sets

Recall:

- A subset K of a topological space is *compact* $\stackrel{\text{def}}{\Longleftrightarrow}$ any open cover of K has a finite subcover
- A subset $K \subset \mathbb{R}^\nu$ is compact $\Longleftrightarrow$ K is bounded and closed

Uniform (or L^∞) norm on $C(K)$: $\|f\|_\infty = \sup_{\mathbf{x}\in K} |f(\mathbf{x})|$

Uniform approximation on K: given $f : K \to \mathbb{R}$, for any $\epsilon > 0$ find $\widetilde{f} : K \to \mathbb{R}$ such that $\|\widetilde{f} - f\|_\infty < \epsilon$

Uniform approximation on compact sets in $\mathbb{R}^\nu$: given $f : \mathbb{R}^\nu \to \mathbb{R}$, for any compact $K \subset \mathbb{R}^\nu$ and $\epsilon > 0$ find $\widetilde{f}$ such that $\|(f - \widetilde{f})|_K\|_\infty < \epsilon$

Exercise: Why might it be reasonable to consider uniform approximation on compact sets in $\mathbb{R}^\nu$ rather than on whole $\mathbb{R}^\nu$? Will the following theorem remain valid in this case?

The universal approximation theorem

Many versions; a nice one:

Theorem (Leshno et al.'93)

Suppose that the activation function σ is continuous. Then, the following are equivalent:

1. *Any continuous $f : \mathbb{R}^\nu \to \mathbb{R}$ can be uniformly approximated on compact sets by one-hidden-layer σ-NN's*
2. *σ is not a polynomial.*

Exercise: 1) $\implies$ 2)

The nontrivial part: 2) $\implies$ 1)

Proof of UAT: reduction to 1D case

A *ridge function* f: $f(\mathbf{x}) = g(\mathbf{x} \cdot \mathbf{q})$ for some $\mathbf{q} \in \mathbb{R}^\nu$ and $g : \mathbb{R} \to \mathbb{R}$

Lemma

Any continuous $f : \mathbb{R}^\nu \to \mathbb{R}$ can be approximated by finite linear combinations of continuous ridge functions.

By the Lemma, proving UAT is reduced to the case $\nu = 1$ (it remains to approximate $g(\cdot)$ by expressions $\sum_{n=1}^N c_n \sigma(w_n \cdot + h_n)$)

Proof of the Lemma

Approximation by trigonometric polynomials:

- Given a compact $K \subset \mathbb{R}^\nu$, approximate $f|_K$ by a smooth function f_1 supported on some $[-a, a]^\nu$
- Expand f_1 in a (multi-dimensional) Fourier series, $f_1(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbb{Z}^\nu} c_{\mathbf{k}} e^{\pi i \mathbf{k}\cdot\mathbf{x}/a}$
- By smoothness of f_1, $|c_{\mathbf{k}}| = O(|\mathbf{k}|^{-\alpha})$ for any α
- Hence, f_1 can be approximated on K in $\|\cdot\|_\infty$ by finite trigonometric polynomials
- Each trigonometric monomial is a ridge function

Exercise: Give an alternative proof using Stone-Weierstrass theorem or polynomial approximation

Theorem (Weierstrass)

For any continuous $f : [a, b] \to \mathbb{R}$ *and any* $\epsilon > 0$ *there exists a polynomial* f_1 *such that* $\max_{x\in[a,b]} |f(x) - f_1(x)| < \epsilon$.

- A *subalgebra* $A \subset C(X, \mathbb{R})$: a subspace closed under multiplication
- Subset A *separates points of* X: for any $\mathbf{x}_1, \mathbf{x}_2 \in X$ there exists $f \in A$ such that $f(\mathbf{x}_1) \neq f(\mathbf{x}_2)$

Theorem (Stone-Weierstrass)

Let X *be a compact Hausdorff space (e.g., a compact metric space). Let* A *be a subalgebra in* $C(X, \mathbb{R})$ *separating points of* X *and containing* $f \equiv 1$. *Then* A *is dense in* $C(X, \mathbb{R})$.

Application: denseness of trigonometric polynomials in $C([-a, a]^\nu, \mathbb{R})$

UAT: proof in the 1D case

Exercise: Give a direct proof in the special case of ReLU σ, by approximating f by a linear spline $\widetilde{f}$ and writing

$$\widetilde{f}(x) = \sum_{n=1}^{N} c_n (x - h_n)_+$$

In general:

1. First prove for $\sigma \in C^\infty(\mathbb{R})$
2. Then extend to general nonpolynomial $\sigma \in C(\mathbb{R})$

Proof for $\sigma \in C^\infty(\mathbb{R})$

Exercise: It suffices to show that for any n expressions $\sum_{k=0}^{N} c_k \sigma(w_k x + h_k)$ can approximate the monomial x^n

Proof of UAT. For given n, since σ is not a polynomial, we can find $x_0 \in \mathbb{R}$ such that $\frac{d^n \sigma}{dx^n}(x_0) \neq 0$. Then, the monomial x^n can be approximated by expressions $\sum_{k=0}^{n} c_k \sigma(w_k x + h_k)$:

$$\begin{aligned}
\sigma(x_0 + wx) &= \sigma(x_0) + o(1) && (w \to 0) \\
\tfrac{1}{w}(\sigma(x_0 + wx) - \sigma(x_0)) &= \tfrac{d\sigma}{dx}(x_0)x + o(1) && (w \to 0) \\
\tfrac{1}{w^2}(\sigma(x_0 + 2wx) - 2\sigma(x_0 + wx) + \sigma(x_0)) &= \tfrac{d^2\sigma}{dx^2}(x_0)x^2 + o(1) && (w \to 0) \\
&\cdots \\
\frac{1}{w^n}\sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}\sigma(x_0 + kwx) &= \tfrac{d^n\sigma}{dx^n}(x_0)x^n + o(1) && (w \to 0)
\end{aligned}$$

Proof for $\sigma \in C^\infty(\mathbb{R})$

Exercise: Prove above claim using Taylor expansion and the identity

$$\sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}k^s = \begin{cases} 0, & s = 0, \dots, n-1 \\ n!, & s = n \end{cases}$$

(verify the identity by differentiating the function $(t-1)^n$).

Remark: this approximation uses small weights w_k and large c_k

Proof for general nonpolynomial $\sigma \in C(\mathbb{R})$

Suppose that some x^m cannot be approximated by $\sum_k c_k\sigma(w_k \cdot +h_k)$

Smoothen σ by convolving with a smooth kernel:

$$\sigma_\phi = \sigma * \phi, \quad \phi \in C_0^\infty(\mathbb{R})$$

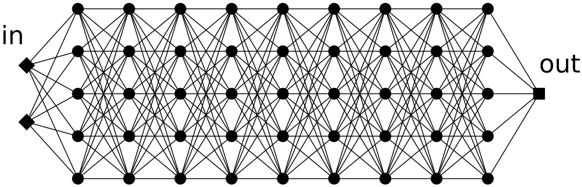
σ_ϕ can be approximated by finite linear combinations $\sum_k c_k\sigma(w_k \cdot +h_k)$.
Hence x^m cannot be approximated by $\sum_k c_k\sigma_\phi(w_k \cdot +h_k)$.

But then, from the argument for smooth σ_ϕ, we get $\frac{d^m\sigma_\phi}{dx^m} \equiv 0$, i.e. σ_ϕ is a polynomial of degree $< m$.

Since ϕ was arbitrary, σ must also be a polynomial of degree $< m$. $\square$

Absence of UAT: deep narrow networks

Fully-connected networks of "width" H and arbitrary depth.
Example ($\nu = 2$ inputs, width $H = 5$):



Theorem (Hanin & Sellke, arXiv:1710.11278)

For given ν, width-H ReLU networks approximate any $f \in C(\mathbb{R}^\nu)$ if and only if $H > \nu$.

Proof that $H > \nu$ is necessary

Claim: $f(\mathbf{x}) = \sum_{s=1}^{\nu} x_s^2$ cannot be approximated by width-ν ReLU networks.

A *level set*: $\widetilde{f}^{-1}(a)$ for some $a \in \mathbb{R}$

Lemma

Let $S \subset \mathbb{R}^\nu$ be the set of input points on which all ReLU evaluations throughout the evaluation of $\widetilde{f}$ are (strictly) positive. Then S is open and convex, $\widetilde{f}$ is affine on S, and every bounded connected component of a level set of $\widetilde{f}$ is contained in S.

Exercise: Prove the convexity, openness and affinity statements (easy)

Exercise: Derive the Theorem from the Lemma (easy)

Proof of Lemma: bounded connected components of level sets are contained in S

Suppose $\mathbf{x} \in \widetilde{f}^{-1}(a)$ is not in S. We will show that $\mathbf{x}$ belongs to an unbounded connected component of $\widetilde{f}^{-1}(a)$

Since $\mathbf{x} \notin S$, when computing $\widetilde{f}(\mathbf{x})$, at some layer k one of the ReLU's is applied to a non-positive value. Assume k is the earliest such layer.

Let $\widetilde{f}_j$ denote the action of first j hidden layers:

$$\widetilde{f}_j(\mathbf{x}) = \mathbf{ReLU} \circ A_j \circ \cdots \mathbf{ReLU} \circ A_1(\mathbf{x}) : \mathbb{R}^\nu \to \mathbb{R}^\nu,$$

where **ReLU** is component-wise ReLU

Let s be the vanishing component of $\widetilde{f}_k(\mathbf{x})$, i.e. $(\widetilde{f}_k(\mathbf{x}))_s = 0$
Then $\mathbf{ReLU}^{-1}(\widetilde{f}_k(\mathbf{x}))$ contains an infinite ray $R \ni A_k \circ \widetilde{f}_{k-1}(\mathbf{x})$:

$$R = \left\{ \mathbf{y} : y_m \begin{cases} \leq 0, & m = s \\ = (A_k \circ \widetilde{f}_{k-1}(\mathbf{x}))_m, & m \neq s \end{cases} \right\}$$

Proof of Lemma: continued

Now observe that:

- For any $\mathbf{u} \in \mathbb{R}^\nu$ and linear $A : \mathbb{R}^\nu \to \mathbb{R}^\nu$, if $A\mathbf{u}$ belongs to some unbounded connected Q, then the connected component of $A^{-1}Q$ containing $\mathbf{u}$ is unbounded[1] (Consider separately the cases of degenerate and non-degenerate A)
- The same holds if a linear transformation A is replaced by **ReLU**

Starting from $\mathbf{u} = \widetilde{f}_{k-1}(\mathbf{x})$ and $Q = R$ and applying these observations for $A_k, \mathbf{ReLU}, A_{k-1}, \mathbf{ReLU}, \ldots, A_1$, we see that $\mathbf{x}$ is contained in an unbounded connected component of
$A_1^{-1} \circ \ldots \circ \mathbf{ReLU}^{-1} \circ A_k^{-1} \circ \mathbf{ReLU}^{-1}(\widetilde{f}_k(\mathbf{x}))$, which is in turn a subset of $\widetilde{f}^{-1}(a)$. $\square$

[1] **Remark:** this wouldn't be true with $H > \nu$.

Open (?) problems

- Give a necessary and sufficient condition for a function $f \in C(\mathbb{R}^\nu)$ to be approximable by width-ν ReLU networks (no bounded connected components in level sets?)
- What are the minimal networks widths for other activation functions?

A sufficient condition for rather general activations: Kidger & Lyons, arXiv:1905.08539 (2019)

Exercise: Consider the family of ReLU networks that have width $H > \nu$ in every layer except for, say, layer 10, in which they have only $\nu - 1$ neurons. Show that this family does not have the universal approximation property.

Way forward: approximation rates

The universal approximation property: only a **qualitative** assurance of approximability

Any **quantitative** estimates of how efficiently can various functions be approximated by neural networks?

That requires us:

- to make more detailed assumptions about the class of fitted function, e.g. their smoothness (**Sobolev spaces**);
- to quantify the model complexity (**Parametric approximations**)

L^p norms/spaces

L^p norms of functions $f : \Omega \to \mathbb{R}, \Omega \subset \mathbb{R}^\nu$:

$$\|f\|_p = \begin{cases} \Big(\int_\Omega |f(\mathbf{x})|^p d\mathbf{x} \Big)^{1/p}, & 1 \leq p < \infty \\ \operatorname{ess\,sup}_{\mathbf{x} \in \Omega} |f(\mathbf{x})|, & p = \infty \end{cases}$$

Exercise: Assuming $\int_\Omega d\mathbf{x} < \infty$, show that $\|f\|_\infty = \lim_{p \to +\infty} \|f\|_p$

L^p spaces:

$$L^p(\Omega) = \{f : \Omega \to \mathbb{R} : \|f\|_p < \infty\}/\{\|f\| = 0\}$$

Exercise*: for $p \in [1, \infty]$, $L^p(\Omega)$ is a Banach space (normed + metrically complete)

Sobolev spaces: general idea

Banach spaces $\mathcal{W}^{d,p}(\Omega) = \{f : \Omega \to \mathbb{R} | \, \|f\|_{d,p} < \infty\}$

- $\Omega \subset \mathbb{R}^\nu$
- d: number of derivatives
- $p \in [1, \infty]$ (as in L^p)

$$\|f\|_{d,p} = \sum_{\mathbf{k}:|\mathbf{k}|\le d} \|D^{\mathbf{k}} f\|_p \qquad |\mathbf{k}| = \sum_{s=1}^{\nu} k_s$$

A rigorous definition ensuring completeness?

Sobolev spaces: rigorous definitions

Approach 1: first take functions $f \in C^\infty(\Omega)$ with finite norm $\|f\|_{d,p}$, then define $\mathcal{W}^{d,p}(\Omega)$ as their $\|\cdot\|_{d,p}$-completion

Approach 2: define $\mathcal{W}^{d,p}(\Omega)$ as the space of all f's having weak derivatives up to degree d in L^p

(A weak derivative $(\frac{\partial f}{\partial x_s})_{\mathrm{w}}$: $\int_\Omega (\frac{\partial f}{\partial x_s})_{\mathrm{w}}(\mathbf{x})\phi(\mathbf{x})d\mathbf{x} = -\int_\Omega f(\mathbf{x})\frac{\partial \phi}{\partial x_s}(\mathbf{x})d\mathbf{x}$ for any $\phi \in C_0^\infty(\Omega)$)

The two approaches are equivalent for $p < \infty$ (Meyers-Serrin theorem), but not for $p = \infty$ (Def.2 gives a larger space)

Exercise: Let $f(x) = |x|$. Then $f \in \mathcal{W}^{1,\infty}([-1,1])$ in the sense of Def.2, but not Def.1.

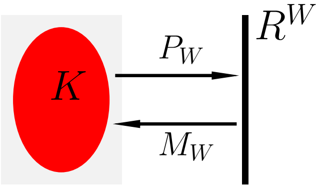
Sobolev spaces: further properties

Exercise: Describe the values d, p, ν for which $f \in \mathcal{W}^{d,p}(\mathbb{R}^\nu)$ may have a singularity $\sim |\mathbf{x}|^\alpha$ with $\alpha < 0$.

Exercise: (With Def.2) For $d \geq 1$, $\mathcal{W}^{d,\infty}$ consists of functions that are globally Lipschitz along with their derivatives up to degree $d-1$.

(**Def:** f is *Lipschitz* (with Lipschitz constant L) if $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y}$)

Parametric approximation



Suppose approximation has the form $\widetilde{f}_W = M_W(P_W(f))$, where

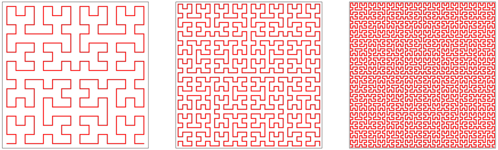
- $f \in K$, where K is the set of functions that we want to approximate
- $K \subset \mathcal{F}$, where $\mathcal{F}$ is a normed functional space (e.g. $\mathcal{F} = C([0,1])$)
- W : number of parameters
- $P_W : K \to \mathbb{R}^W$: **parameter assignment** map ($f \mapsto \mathbf{W}$)
- $M_W : \mathbb{R}^W \mapsto \mathcal{F}$: **reconstruction** map ($\mathbf{W} \mapsto \widetilde{f}$)

E.g., in the case of ANNs, P_W corresponds to fitting the weights using some fixed architecture

Isn't this framework too general?

The class of general parametric approximations is too wide!

Exercise: Let K be compact. Then, for $W = 1$, there is a smooth maps M_W such that $K \subset \overline{M_W(\mathbb{R})}$.



(https://en.wikipedia.org/wiki/Space-filling_curve)

Linear M_W: a good class of approximations, but the linearity constraint is too restrictive

A reasonable framework admitting nonlinear M_W, but avoiding unnatural examples?

Continuous parametric approximations

Key requirement: parameter assignment P_W is **continuous**
(Remark: no assumption on the reconstruction map M_W!)

Exercise: Why does this requirement exclude "Peano curve" constructions?

Optimal approximation (a.k.a. *continuous nonlinear W-width*):

$$h_W = \inf_{P_W \text{ cont.}, M_W} \sup_{f \in K} \|f - M_W(P_W(f))\|$$

Key result: Let K be a ball in $\mathcal{W}^{d,p}([0,1]^\nu)$ and $\mathcal{F} = L^p([0,1]^\nu)$. Then

$$\boxed{h_W \asymp W^{-d/\nu}}$$

Remark: $\frac{\nu}{d}$ can be interpreted as a "complexity" of the ball K

Remark: Typically, in "classical" linear approximation methods (e.g., Fourier series expansion or wavelets), the approximation rate agrees with the optimal continuous rate: $\|\widetilde{f}_W - f\|_\infty = \widetilde{O}(W^{-d/\nu})$

The lower bound

Theorem (DeVore, Howard, Micchelli 1989)

Let $K = B_{d,p,\nu}$ be the unit ball in $\mathcal{W}^{d,p}([0,1]^\nu)$, and $\mathcal{F} = L^p([0,1]^\nu)$. Then $h_W \geq CW^{-d/\nu}$ for some constant $C(d,p,\nu)$.

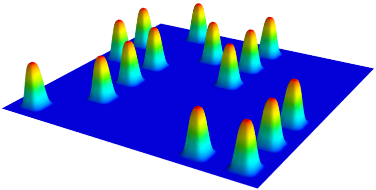
Sketch of proof for $p = \infty$ – beginning

Fix some $\phi \in C^\infty(\mathbb{R}^\nu)$ such that $\phi(\mathbf{x}) = 0$ if $|\mathbf{x}| > \frac{1}{2}$.

For a given $N \in \mathbb{N}$, consider the grid $G_N = \{\frac{1}{N}, \frac{2}{N}, \ldots, \frac{N}{N}\}^\nu \subset [0,1]^\nu$. Note that $|G_N| = N^\nu$.

Consider the map $\Phi_N : [-1,1]^{G_N} \to \mathcal{W}^{d,p}([0,1]^\nu)$ that places rescaled, shifted and weighted functions ϕ ("spikes") at the grid points:

$$\Phi_N(\{c_\mathbf{n}\}_{\mathbf{n}\in G_N}) = CN^{-d} \sum_{\mathbf{n}\in G_N} c_\mathbf{n}\phi(N(\cdot - \mathbf{n}))$$



(In this example, $c_\mathbf{n} \in \{0,1\}$)

If C is small enough, then $\Phi_N([-1,1]^{G_N}) \subset B_{d,\infty,\nu}$ for any N

Sketch of proof – continued

Lemma (Borsuk-Ulam antipodality theorem)

Suppose that g maps continuously the n-dimensional sphere S^n to $\mathbb{R}^n$. Then there exist $\mathsf{x} \in S^n$ such that $g(\mathsf{x}) = g(-\mathsf{x})$.

Exercise: prove for $n = 1$.

Sketch of proof – end

Let $\mathcal{D} = \partial([-1,1]^{G_N})$, then $\mathcal{D} \cong S^{N^\nu - 1}$.

Consider the map $g = P_W \circ \Phi_N$ on $\mathcal{D}$. By Borsuk-Ulam, if $W \leq N^\nu - 1$, then there exists $\mathbf{x} \in \mathcal{D}$ such that $P_W(\Phi_N(\mathbf{x})) = P_W(\Phi_N(-\mathbf{x}))$.

Then,

$$\begin{aligned}
\sup_{f\in K} \|f - M_W(P_W(f))\|_\infty \geq\ & \max\Big(\|\Phi_N(\mathbf{x}) - M_W(P_W(\Phi_N(\mathbf{x})))\|_\infty, \\
& \|\Phi_N(-\mathbf{x}) - M_W(P_W(\Phi_N(-\mathbf{x})))\|_\infty\Big) \\
\geq\ & \tfrac{1}{2}\|\Phi_N(\mathbf{x}) - \Phi_N(-\mathbf{x})\|_\infty \\
=\ & CN^{-d}\|\phi\|_\infty.
\end{aligned}$$

Taking $N \sim W^{1/\nu}$, we get $\sup_{f\in K} \|f - M_W(P_W(f))\| \geq CW^{-d/\nu}$. $\square$

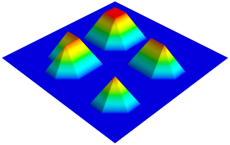
The upper bound

Proposition

1. *Let $K = B_{d,\infty,\nu}$ be the unit ball in $\mathcal{W}^{d,\infty}([0,1]^\nu)$. Then $h_W \le CW^{-d/\nu}$.*
2. *The bound can be attained with* linear *maps P_W, M_W.*

Proof.

Take $\phi \in C_0(\mathbb{R}^\nu), 0 \le \phi \le 1$, such that the spikes $\{\phi(N(\cdot - \mathbf{n}))\}_{\mathbf{n}\in G_N}$ form a *partition of unity*:



$$\sum_{\mathbf{n}\in G_N} \phi(N(\mathbf{x}-\mathbf{n})) \equiv 1, \quad \mathbf{x} \in [0,1]^\nu.$$

Let:

$$P_W(f) = \{D^{\mathbf{k}} f(\mathbf{n})\}_{\mathbf{n}\in G_N, |\mathbf{k}|\le d-1} \in \mathbb{R}^{cN^\nu}$$

$$M_W(\{w_{\mathbf{k}}(\mathbf{n})\}_{\mathbf{n}\in G_N, |\mathbf{k}|\le d-1}) = \sum_{\mathbf{n}\in G_N} \phi(N(\mathbf{x}-\mathbf{n})) \sum_{|\mathbf{k}|\le d-1} \frac{w_{\mathbf{k}}(\mathbf{n})}{\mathbf{k}!}(\mathbf{x}-\mathbf{n})^{\mathbf{k}}$$

The upper bound – continued

Exercise: P_W is continuous on K

Claim: $\|f - M_W(P_W(f))\|_\infty \le CN^{-d}$

$$
\begin{aligned}
|f(\mathbf{x}) - M_W(P_W(f))| &= \Big| \sum_{\mathbf{n}\in G_N} \phi(N(\mathbf{x}-\mathbf{n}))\Big[f(\mathbf{x}) - \sum_{|\mathbf{k}|\le d-1} \frac{D^{\mathbf{k}} f(\mathbf{n})}{\mathbf{k}!}(\mathbf{x}-\mathbf{n})^{\mathbf{k}}\Big]\Big| \\
&\le \sum_{\substack{\text{finitely many } \mathbf{n}\in G_N: \\ |\mathbf{n}-\mathbf{x}|_\infty < c/N}} \Big| f(\mathbf{x}) - \sum_{|\mathbf{k}|\le d-1} \frac{D^{\mathbf{k}} f(\mathbf{n})}{\mathbf{k}!}(\mathbf{x}-\mathbf{n})^{\mathbf{k}}\Big| \\
&\le CN^{-d}
\end{aligned}
$$

(by a Taylor remainder bound)

Since $W = cN^\nu$, we get $\|f - M_W(P_W(f))\|_\infty \le CW^{-d/\nu}$ $\square$

Example: Fourier series approximation

$$f(\mathbf{x}) \sim \sum_{\mathbf{n}\in\mathbb{Z}^\nu} c_{\mathbf{n}} e^{2\pi i \mathbf{n}\cdot\mathbf{x}}, \qquad \mathbf{x}\in[0,1]^\nu$$

(= a shallow NN with fixed hidden layer weights and activation sin)

Linear parameter assignment P_W and reconstruction M_W:

$$c_{\mathbf{n}} = \int_{[0,1]^\nu} f(\mathbf{x}) e^{-2\pi i \mathbf{n}\cdot\mathbf{x}} d\mathbf{x}, \qquad \widetilde{f}_W(\mathbf{x}) = \sum_{\mathbf{n}\in\mathbb{Z}^\nu : |\mathbf{n}|\le N} c_{\mathbf{n}} e^{2\pi i \mathbf{n}\cdot\mathbf{x}}, \quad N \sim W^{1/\nu}$$

- Using the abstract theory we expect:

$$\|\widetilde{f}_W - f\|_\infty \gtrsim W^{-d/\nu} \sim N^{-d}$$

- Compare with classical Jackson theorem for $\nu = 1$:

$$\|\widetilde{f}_W - f\|_\infty \le C \frac{\ln N}{N^d} \omega\left(\tfrac{1}{N}\right),$$

where ω is the modulus of continuity of $\frac{d^d f}{dx^d}$

Do neural networks achieve optimal approximation rates?

For ReLU networks, we'll show:

- Yes – for deep networks
- No – for shallow networks

Efficient approximation by deep ReLU networks

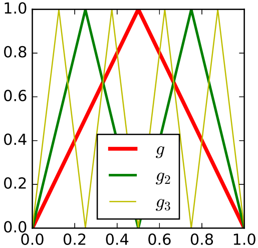
Consider increasingly complex f's:

$$
\begin{array}{rc}
 & f(x) = x^2 \\
xy = \frac{1}{4}((x+y)^2 - (x-y)^2) & \Bigg\Downarrow \\
 & \text{General } \textbf{polynomial } f \\
\text{local Taylor expansion} & \Bigg\Downarrow \\
 & \text{General } \textbf{smooth } f
\end{array}
$$

The sawtooth functions (Telgarsky, arXiv:1602.04485)

The "tooth" function

$$g(x) = \begin{cases} 2x, & x < \frac{1}{2} \\ 2(1-x), & x \geq \frac{1}{2} \end{cases}$$
$$= 2(x)_+ - 4(x-0.5)_+ + 2(x-1)_+$$



Iterated "sawtooth" functions with 2^{m-1} "teeth":

$$g_m(x) = \underbrace{g \circ g \circ \cdots \circ g}_{m}(x)$$

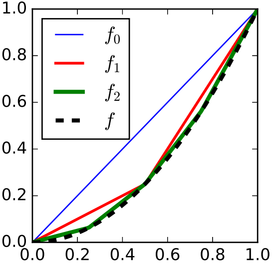
Efficient implementation of $f(x) = x^2$ (arxiv:1610.01145)

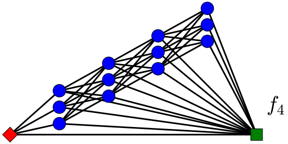
Let

$$\widetilde{f}_m(x) = x - \sum_{k=1}^{m} \frac{g_k(x)}{2^{2k}}$$

Then

$$\|\widetilde{f}_m(x) - x^2\|_{C[0,1]} = \frac{1}{2^{2m+2}}$$





Extension to polynomials

Multiplication reduces to squaring thanks to polarization identity:

$$xy = \frac{1}{4}((x+y)^2 - (x-y)^2)$$

Exercise: A fixed polynomial on a bounded domain can be implemented with accuracy ϵ using a ReLU network with $O(\log(1/\epsilon))$ layers, neurons and connections.

Extension to Sobolev balls

Let $K = B_{d,p=\infty,\nu}$ (the Sobolev unit ball).

We look for P_W, M_W such that

$$\sup_{f\in K} \|f - M_W(P_W(f))\|_\infty < \epsilon \tag{1}$$

Theorem

Eq.(1) can be fulfilled with linear maps P_W, M_W, where M_W is implemented by a ReLU network with $W = O(\epsilon^{-\nu/d}\log(1/\epsilon))$ weights and $O(\log(1/\epsilon))$ layers.

Sketch of proof: follow the proof of the upper bound $h_W = O(W^{-d/\nu})$; approximate Taylor polynomials by ReLU subnetworks.

Extension to analytic functions

Let f be (real) analytic in a neighborhood of $[a, b] \subset \mathbb{R}$.

Exercise (cf. Liang & Srikant, arxiv:1610.04161) $\|f - \widetilde{f}\|_{C[a,b]} < \epsilon$ can be achieved with $\widetilde{f}$ implemented by a ReLU network with $O(\log^2(1/\epsilon))$ layers and connections.

Counting linear pieces in $\widetilde{f}$

Let $\widetilde{f} : [0, 1] \to \mathbb{R}$ be implemented by a ReLU network with L hidden layers and U neurons. Then $\widetilde{f}$ is piecewise linear on $[a, b]$. Let M denote the number of pieces.

Lemma (Telgarsky, arXiv:1602.04485)

$M \leq (2U)^L$

Proof. By induction. For $n \leq L$, suppose that $[0, 1]$ can be divided into N_n intervals $[a_{n,k}, b_{n,k}]_{k=1}^{N_n}$ such that the outputs of all neurons of all layers $< n$ are affine functions (without breakpoints) on these intervals. In particular, $N_1 = 1$ and $[a_{1,1}, b_{1,1}] = [0, 1]$.

Consider the action of the n'th layer on one $[a_{n,k}, b_{n,k}]$. Each neuron in this layer can create at most one breakpoint in this interval. Therefore, $N_{n+1} \leq (U_n + 1)N_n$, where U_n is the number of neurons in the n'th layer. So, $N_{L+1} \leq (U_1 + 1)(U_2 + 1) \cdots (U_L + 1) \leq (2U)^L$. $\square$

Proposition

To approximate $f(x) = x^2$ on $[0, 1]$ with uniform accuracy ϵ, a ReLU network with L hidden layers requires at least $\frac{1}{2}(8\epsilon)^{-1/(2L)}$ computation units and weights.

Proof. If $\widetilde{f}$ is linear on $[a, b]$, then $\max_{x\in[a,b]} |\widetilde{f}(x) - x^2| \geq \frac{(b-a)^2}{8}$.

By counting lemma, if the network has U neurons, then we can find such an interval of linearity with $b - a \geq (2U)^{-L}$. Therefore $\epsilon \geq \frac{(2U)^{-2L}}{8}$, and then $U \geq \frac{1}{2}(8\epsilon)^{-1/(2L)}$. □

Conclusion: To approximate $f(x) = x^2$, fixed-depth ReLU networks require a faster complexity growth ($\gtrsim \epsilon^{-1/(2L)}$) than arbitrary-depth ones ($O(\log(1/\epsilon))$)

Vapnik-Chervonenkis (VC) dimension: overview

VC-dimension: characterizes expressiveness of classifiers

Our goal: examine VC-dimension of networks and related models

Sources:

- (main) M. Anthony, P. Bartlett, Neural Network Learning: Theoretical Foundations, 1999. Chapters 3, 6 – 8
- M. Raginsky, Vapnik-Chervonenkis classes

The growth function

H: some family of maps $X \to \{0,1\}$
(e.g., all neural networks of given architecture with thresholded output)

$H|_S$: restrictions of maps $f \in H$ to a subset $S \subset X$

The growth function:

$$\Pi_H(m) = \sup_{S \subset X, |S|=m} \big| H|_S \big|$$

Exercise: Compute the growth function ($X = \mathbb{R}$):

1. $H = \{f_{a,b}\}_{a,b \in \mathbb{R}}$; $f_{a,b}(x) = \mathrm{sgn}(ax+b)$ (where $\mathrm{sgn}(x) := \mathbf{1}_{(0,+\infty)}(x)$)
2. $H = \{f_{a,b}\}_{a<b}$; $f_{a,b}(x) = \mathbf{1}_{[a,b]}(x)$
3. $H = \{f_a\}_{a \in \mathbb{R}}$; $f_a(x) = \mathrm{sgn}(\sin(ax))$

VC-dimension

$S \subset X$ is **shattered** by H: $H|_S$ implements all possible $2^{|S|}$ maps $S \to \{0, 1\}$

VC-dimension:

$$\begin{aligned}\mathrm{VCdim}(H) &= \sup\{m : |S| = m \text{ and } S \text{ is shattered by } H\} \\ &= \sup\{m : \Pi_H(m) = 2^m\}\end{aligned}$$

Exercise: $\Pi_H(m) = 2^m$ for all $m \le \mathrm{VCdim}(H)$
Exercise: Compute VCdim for families H from the previous exercise. Show that $\mathrm{VCdim}(\{\mathrm{sgn}(\sin(ax))\}) = \infty$.

The Sauer-Shelah lemma (good exposion: Wikipedia)

By definition, the growth function Π_H determines $\mathrm{VCdim}(H)$

Conversely, $\mathrm{VCdim}(H)$ restricts Π_H:

Theorem (Sauer-Shelah)

$$\Pi_H(m) \le \sum_{k=0}^{\mathrm{VCdim}(H)} \binom{m}{k}$$

$$\binom{a}{b} := \begin{cases} \frac{a!}{b!(a-b)!}, & a \ge b \\ 0, & a < b \end{cases}$$

Theorem (Pajor)

H shatters at least $\left|H|_S\right|$ subsets of S (including $\emptyset$).

Exercise: Pajor $\implies$ Sauer-Shelah (use that S has $\sum_{k=0}^{d}\binom{|S|}{k}$ subsets of size $\le d$ and then there must be at least one large shattered subset)

Proof of Pajor theorem

Let $\mathcal{F} = H|_S$. Proof by induction on $|\mathcal{F}|$. The base of induction: $|\mathcal{F}| = 1$, then H shatters $\emptyset$.

Let us prove theorem for given $\mathcal{F}$ assuming it holds for smaller sizes. Take some $\mathbf{x} \in S$ such that both $\mathcal{F}_0 = \{f \in \mathcal{F} : f(\mathbf{x}) = 0\}$ and $\mathcal{F}_1 = \{f \in \mathcal{F} : f(\mathbf{x}) = 1\}$ are nonempty.

We have $\mathcal{F} = \mathcal{F}_0 \sqcup \mathcal{F}_1, |\mathcal{F}| = |\mathcal{F}_0| + |\mathcal{F}_1|$. By induction assumption, theorem holds for $\mathcal{F}_0$ and $\mathcal{F}_1$. Let

$$A_k = \{Q \subset S : Q \text{ is shattered by } \mathcal{F}_k\}, \quad k = 0, 1,$$

then $|A_0| + |A_1| \geq |\mathcal{F}_0| + |\mathcal{F}_1| = |\mathcal{F}|$. Note that if $Q \in A_0$ or $Q \in A_1$, then $\mathbf{x} \notin Q$.

Let

$$A = (A_0 \cup A_1) \cup \{Q \cup \{\mathbf{x}\} : Q \in A_0 \cap A_1\}$$

Then $|A| = |A_0| + |A_1|$, and any $Q \in A$ is shattered by $\mathcal{F}$. $\square$

A more convenient bound on the growth function

Lemma

For $m \geq d \geq 1$,

$$\sum_{k=0}^{d} \binom{m}{k} \leq \left(\frac{em}{d}\right)^d$$

Proof:

$$\sum_{k=0}^{d} \binom{m}{k} \leq \left(\frac{m}{d}\right)^d \sum_{k=0}^{d} \binom{m}{k} \left(\frac{d}{m}\right)^k \leq \left(\frac{m}{d}\right)^d \left(1 + \frac{d}{m}\right)^m \leq \left(\frac{em}{d}\right)^d$$

Corollary: If $\mathrm{VCdim}(H) = d$, then

$$\Pi_H(m) \begin{cases} = 2^m, & m \leq d \\ \leq \left(\frac{em}{d}\right)^d, & m > d \end{cases}$$

In particular, $\Pi_H(m)$ grows exponentially for $m \leq d$, but polynomially for $m > d$.

The simple perceptron model

Simple perceptron: $X = \mathbb{R}^\nu$, $H = \{\mathrm{sgn}(f_{\mathbf{w},h})\}_{\mathbf{w}\in\mathbb{R}^\nu, h\in\mathbb{R}}$, where $f_{\mathbf{w},h}(\mathbf{x}) = \mathbf{w}^t\mathbf{x} - h$, i.e.

$$\mathrm{sgn}(f_{\mathbf{w},h}(\mathbf{x})) = \begin{cases} 1, & \mathbf{w}^t\mathbf{x} - h > 0 \\ 0, & \text{otherwise} \end{cases}$$

Theorem

1. $\Pi_H(m) = 2\sum_{k=0}^{\nu} \binom{m-1}{k}$
2. $\mathrm{VCdim}(H) = \nu + 1$

Exercise: 1) $\Longrightarrow$ 2)

Proof: step 1 – Topological reduction

$\mathrm{CC}(A)$: number of connected components in the set A

Lemma

Let $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \mathbb{R}^{\nu+1}$. *Define*

$$\begin{aligned} P_i &= \{(\mathbf{w}, h) \in \mathbb{R}^{\nu} : f_{\mathbf{w},h}(\mathbf{x}_i) = 0\} \\ &= \{(\mathbf{w}, h) \in \mathbb{R}^{\nu+1} : \mathbf{w}^t \mathbf{x}_i - h = 0\} \end{aligned}$$

Then

$$\big|H|_S\big| = \mathrm{CC}(\mathbb{R}^{\nu+1} \setminus \cup_{i=1}^m P_i)$$

Sketch of proof. Each connected component corresponds to an element of $H|_S$, so $\big|H|_S\big| \le \mathrm{CC}(\mathbb{R}^{\nu+1} \setminus \cup_{i=1}^m P_i)$.

Moreover, an element of $H|_S$ corresponds to only one connected component since the sets $\{(\mathbf{w}, h) \in \mathbb{R}^{\nu+1} : \pm f_{\mathbf{w},h}(\mathbf{x}_i) > 0\}$ are convex and have a convex intersection. $\square$

Proof: step 2 – Combinatorics

Let $\widetilde{\mathbf{x}} = (\mathbf{x}, -1)$ and $\widetilde{\mathbf{w}} = (\mathbf{w}, h)$, then we can write

$$P_i = \{\widetilde{\mathbf{w}} \in \mathbb{R}^{\nu+1} : \widetilde{\mathbf{w}}^t \widetilde{\mathbf{x}}_i = 0\}$$

Assume $\{\widetilde{\mathbf{x}}_i\}_{i=1}^m$ are in *general position*, i.e. any subset of up to $\nu + 1$ points are linearly independent.

Define $C(m, \nu) := \mathrm{CC}(\mathbb{R}^{\nu+1} \setminus \cup_{i=1}^m P_i)$

Lemma

$$C(m+1, \nu) = C(m, \nu) + C(m, \nu - 1)$$

Proof: When we add a new hyperplane P_{m+1}, the number of CC in $\mathbb{R}^{\nu+1} \setminus \cup_{i=1}^m P_i$ is increased by the number of CC in $P_{m+1} \setminus \cup_{i=1}^m P_i$. □

Proof: step 2 – Combinatorics (cont-d)

Exercise: $C(m,0) \equiv C(1,\nu) \equiv 2$

$$
\begin{aligned}
C(m,\nu) &= C(m-1,\nu) + C(m-1,\nu-1) \\
&= C(m-2,\nu) + 2C(m-2,\nu-1) + C(m-2,\nu-2) \\
&= \dots \\
&= C(1,\nu) + \binom{m-1}{1} C(1,\nu-1) + \binom{m-1}{2} C(1,\nu-2) + \\
&\quad + \dots + \binom{m-1}{\nu} C(1,0) \\
&= 2\sum_{k=0}^{\nu} \binom{m-1}{k}
\end{aligned}
$$

$\square$

Computation of VCdim(Perceptron): Summary

1. *(topology)* Reduce computation of the growth function Π_H to computation of $\mathrm{CC}(\mathbb{R}^{\nu+1} \setminus \cup_{i=1}^m P_i)$
2. *(combinatorics)* Compute $\mathrm{CC}(\mathbb{R}^{\nu+1} \setminus \cup_{i=1}^m P_i)$
3. Compute VCdim via Π_H

An alternative computation

Exercise: Give an alternative proof that VCdim(Perceptron) $= \nu + 1$:

- Show that the perceptron shatters the set $\{\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_\nu\}$ and hence VCdim $\geq \nu + 1$
- Show that VCdim $\leq \nu + 1$ as follows. Suppose that $|S| > \nu + 1$, then the vectors $\widetilde{\mathbf{x}}_i$ are linearly dependent and some $\widetilde{\mathbf{x}}_k$ can be linearly expressed through the others, e.g. $\widetilde{\mathbf{x}}_{|S|} = \sum_{i=1}^{|S|-1} a_i \widetilde{\mathbf{x}}_i$. Then, if

$$\operatorname{sgn}(f_{\widetilde{\mathbf{w}}}(\mathbf{x}_i)) = \begin{cases} 1, & a_i > 0 \\ 0, & a_i \leq 0 \end{cases}$$

  for $i = 1, \ldots, |S| - 1$, then $\operatorname{sgn}(f_{\widetilde{\mathbf{w}}}(\mathbf{x}_{|S|})) = 1$, i.e. S is not shattered.

Dependence of VCdim on the number of parameters?

Perceptron: VCdim equals the number of degrees of freedom in the perceptron (i.e., $\nu + 1$)

Deep nets: VCdim scales as the product of the number of parameters and depth

Deep networks

Existing results for deep ReLU and piecewise linear networks[2]:

$$cWL\log(W/L) \le \mathrm{VCdim}(W,L) \le CWL\log W,$$

where

- W: total weights; L: depth; c, C: global constants
- $\mathrm{VCdim}(W,L)$: largest VC-dimension of a piecewise linear network with W parameters and L layers

Proofs:

- Upper bound: bounding the growth function Π_H
- Lower bound: an explicit construction ("bit-extraction technique")

The methods extend to more general models (piecewise polynomial activations, general arithmetic networks, etc.)[3]

[2] P. Bartlett et al., Nearly-tight VC-dimension bounds for piecewise linear neural networks, arXiv:1703.02930

[3] Anthony-Bartlett, Ch.8

Proof of the upper bound: main ideas

- *(topology)* Π_H can be upper bounded by counting connected components in various intersections of level sets of f, where $H = \{\mathrm{sgn}(f)\}$
- *(combinatorics)* For ReLU and piecewise polynomial networks, the weight space $\mathbb{R}^W$ can be split into subsets corresponding to polynomial computational branches
- *(algebraic geometry)* In a polynomial branch, apply bounds on the number of CC in *algebraic sets*.

Polynomial dependence on the weights

Exercise: Consider a neural network $y = f(\mathbf{x}, \mathbf{w})$ of depth L, where the activation function is piecewise polynomial with degree at most d. Then, in each smooth computational branch, $f(\mathbf{x}, \cdot)$ for fixed $\mathbf{x}$ is a polynomial in $\mathbf{w}$ of degree not greater than:

$$\begin{cases} L, & d = 1 \text{ (e.g., ReLU)} \\ (d+1)^L, & d \geq 1 \end{cases}$$

Algebraic sets: $\cap_{k=1}^{N} \{\mathbf{w} : f_k(\mathbf{w}) = 0\}$ with polynomial f_k

Semi-algebraic sets: $\cap_{k=1}^{N} \{\mathbf{w} : f_k(\mathbf{w}) (= \text{or} >) 0\}$ with polynomial f_k

Algebraic geometry

Theorem (Oleinik-Petrovsky, Milnor, Thom,...)

Let $f : \mathbb{R}^W \to \mathbb{R}$ be a polynomial of degree l. Then the number of connected components of $\{\mathbf{w} \in \mathbb{R}^W : f(\mathbf{w}) = 0\}$ is no more than $l^{W-1}(l+2)$.

Exercise: Let $f(\mathbf{w}) = \sum_{k=1}^{W}(w_k - 1)^2(w_k - 2)^2 \cdots (w_k - l/2)^2$. How many CC's does the set $\{\mathbf{w} : f(\mathbf{w}) = 0\}$ have?

Related, but simpler results:

Proposition (from main theorem of algebra)

Let $f : \mathbb{R} \to \mathbb{R}$ be a polynomial of degree l. Then the number of roots $\{w \in \mathbb{R} : f(w) = 0\}$ is no more than l.

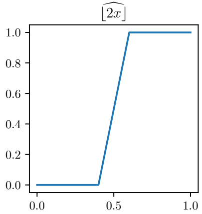
Theorem (Bézout)

Consider two algebraic curves in $\mathbb{R}^2$ defined as the zero sets of polynomials $f, g : \mathbb{R}^2 \to \mathbb{R}$. Then they intersect at no more than $\deg(f) \cdot \deg(g)$ points.

The bit extraction technique (Bartlett-Maiorov-Meir '98)

Given a number $a = 0.b_1 b_2 \ldots$, the bits $b_1, b_2, \ldots \in \{0, 1\}$ can be extracted by a deep ReLU network:

$$\begin{tikzcd}
 & a = 0.b_1 b_2 \ldots \arrow[dl, "\lfloor 2a \rfloor"'] \arrow[d, "2a - \lfloor 2a \rfloor"] \\
b_1 & a_1 = 0.b_2 b_3 \ldots \arrow[dl, "\lfloor 2a_1 \rfloor"'] \arrow[d, "2a_1 - \lfloor 2a_1 \rfloor"] \\
b_2 & a_2 = 0.b_3 b_4 \ldots \arrow[dl, "\lfloor 2a_2 \rfloor"'] \arrow[d, "2a_2 - \lfloor 2a_2 \rfloor"] \\
\ldots & \ldots
\end{tikzcd}$$



Proof of the lower bound

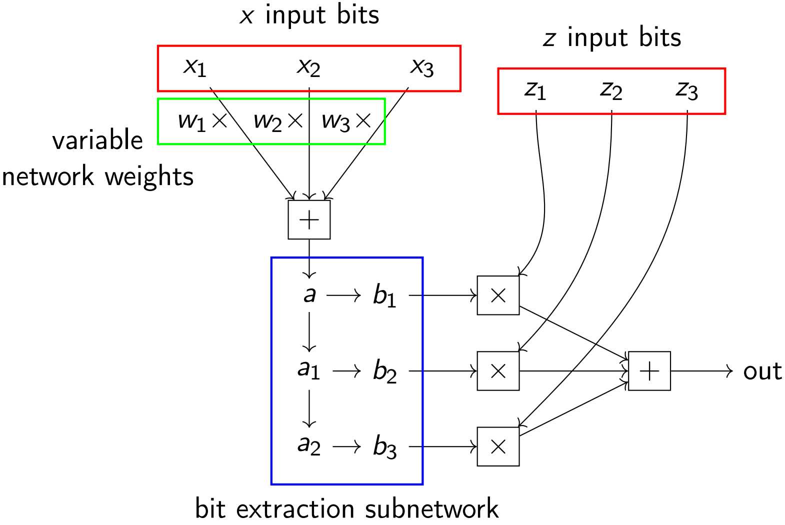
A ReLU network with W weights and L layers that has VCdim $\geq cWL$ (i.e., asymptotically almost maximally expressive):

- Use bit expansion of real numbers: $a = 0.b_1 b_2 \dots b_N$ with $b_n \in \{0, 1\}$
- Construct a finite network that maps $0.b_1 b_2 \dots \mapsto (b_1, 0.b_2 b_3 \dots)$ (i.e., $a \mapsto (\lfloor 2a \rfloor, 2a - \lfloor 2a \rfloor)$)
- By stacking, construct a depth-$O(N)$ network extracting all bits: $a \mapsto (b_1, b_2, \dots, b_N)$
- Extend this to a network $\mathcal{N}_1$ with M inputs that computes $a = \sum_{m=1}^{M} w_m x_m$ in the first layer, and then extracts the digits of a

Proof continued

- Construct a finite network multiplying numbers from the set $\{0, 1\}$
- Construct the final network $\mathcal{N}$ by adding to $\mathcal{N}_1$ a subnetwork with N binary inputs $z_1, \ldots, z_N$ that computes $y = \sum_{n=1}^{N} b_n z_n$.[4]
- Observe: when $\mathbf{x} = \mathbf{e}_m$ and $\mathbf{z} = \mathbf{e}_n$, $\mathcal{N}$ computes the n'th bit of w_m
- $\mathcal{N}$ shatters the set $\{(\mathbf{x}, \mathbf{z}) = (\mathbf{e}_m, \mathbf{e}_n)\}_{m,n=1}^{M,N}$ of size MN (by choosing arbitrary bit expansions of the weights $w_1, \ldots, w_M$)
- $\mathcal{N}$ has size $O(N + M)$ and depth $O(N)$; choose $M \sim N$ to get VCdim $\geq cWL$

[4] If we want connections to be only between neighboring layers, then we can ensure the size of $\mathcal{N}_1$ is increased only by $O(N)$ if we compress $z = \sum_{n=1}^{N} 2^{-n} z_n$ and then reconstruct $z_1, z_2, \ldots$ from z as before.


x input bits
x_1 x_2 x_3
$w_1\times$ $w_2\times$ $w_3\times$
variable
network weights
z input bits
z_1 z_2 z_3
$+$
$a \to b_1$
$a_1 \to b_2$
$a_2 \to b_3$
$\times$
$\times$
$\times$
$+$
out
bit extraction subnetwork

VCdim: application in Statistical Learning Theory[5]

Ground truth: $y_* = y_*(\mathbf{x}) \in \{-1, 1\}$, where $\mathbf{x} \in X$ and has a particular probability distribution

A **classifier**: $g : X \to \{-1, 1\}$, belong to a family H

Risk of a classifier: $R(g) = \mathbb{E}(g(\mathbf{x}) \neq y_*(\mathbf{x}))$

Empirical risk: $R_n(g) = \frac{1}{n}\sum_{k=1}^n \mathbf{1}_{g_n(\mathbf{x}_k)\neq y_*(\mathbf{x}_k)}$, where $\mathbf{x}_k$ are randomly and independently sampled

Theorem (V.-C.). For any δ, with probability at least $1-\delta$,

$$R(g) \leq R_n(g) + 2\sqrt{2\frac{\log \Pi_H(2n) + \log \frac{2}{\delta}}{n}}, \quad \forall g \in H$$

Corollary: For $n > \mathrm{VCdim}(H)/2$,

$$R(g) \leq R_n(g) + 2\sqrt{2\frac{\mathrm{VCdim}(H) \log \frac{2en}{\mathrm{VCdim}(H)} + \log \frac{2}{\delta}}{n}}, \quad \forall g \in H$$

[5] http://www.econ.upf.edu/~lugosi/mlss_slt.pdf

The Kolmogorov(-Arnold) superposition theorem

Theorem (Kolmogorov '57)

There exist $d(2d+1)$ univariate functions $\phi_{ij} \in C[0,1]$ such that any $f \in C([0,1]^d)$ can be represented in the form

$$f(x_1, \dots, x_d) = \sum_{i=1}^{2d+1} \chi_i \Big(\sum_{j=1}^{d} \phi_{ij}(x_j) \Big)$$

with some $\chi_i \in C[0,1]$ depending on f.

- "Every multivariate continuous operation reduces to univariate ones and sums"
- *Exact representation*, not approximation
- The internal functions ϕ_{ij} are very complex "activations"
- Valid only for *continuous* functions – no analog for *smooth* functions

Hilbert's 13'th problem

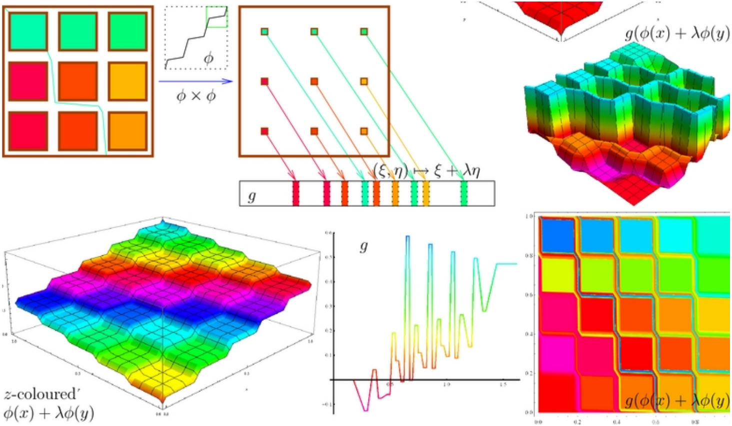
K(-A)ST resolves the "continuous" variant of Hilbert's 13th problem:

Problem: Is it possible to express the roots of general 7th-degree univariate polynomial using continuous (variant: algebraic) functions of two variables?

- The algebraic variant of the problem is still open
- Algebraic function: expressible as a root of a polynomial equation
- 7th-degree equation is known to be solvable by algebraic functions of three variables (by reduction to equation $x^7 + ax^3 + bx^2 + cx + 1 = 0$)

K(-A)ST: proof idea

A clever hierarchical replacement of d-dimensional connectedness by one-dimensional one



(From Dror Bar-Natan, https://www.math.toronto.edu/~drorbn/Talks/Fields-0911/Hilbert13th.html)

Relevance to conventional neural networks?

Girosi-Poggio '89: "Representation properties of networks: **Kolmogorov's theorem is irrelevant**"// Neural Computation, 1(4), 465-469.

Kůrková '91: "**Kolmogorov's theorem is relevant**"// Neural Computation 3(4), 617-622

Relevance to conventional neural networks?

Maiorov-Pinkus '99: 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 two hidden layer networks

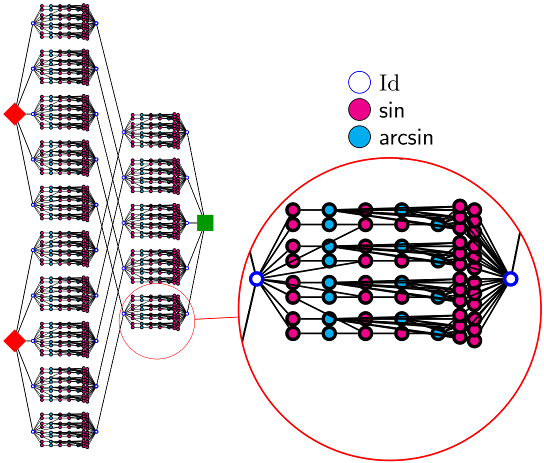
$$\widetilde{f}(\mathbf{x}) = \sum_{i=1}^{6d+3} a_i \sigma\Big(\sum_{j=1}^{3d} c_{ij} \sigma(\mathbf{w}^{ij} \cdot \mathbf{x} + \theta_{ij}) + \gamma_i \Big)$$

Note: *finitely many* neurons, but a *very complex* activation function σ

arXiv:2102.10911: a similar result for a finite network with activations $\sin, \arcsin$. The complexity is hidden in the weights.

A $\{\sin, \arcsin\}$-superexpressive architecture for $d = 2$ inputs

Any function $f \in C([0,1]^2)$ can be approximated with arbitrary accuracy by this fixed-size network:



Model complexity: the information theory approach[6]

Let F be a set in a metric space.

Covering number $\mathcal{N}_\epsilon(F)$: the smallest number of ϵ-balls covering F

Packing number $\mathcal{M}_\epsilon(F)$: the largest cardinality of a subset of F with elements separated by distance $\geq \epsilon$

Exercise: For any F,

$$\mathcal{M}_{2\epsilon}(F) \leq \mathcal{N}_\epsilon(F) \leq \mathcal{M}_\epsilon(F)$$

[6] A. Kolmogorov, V. Tikhomirov, ϵ-Entropy and ϵ-Capacity of Sets In Functional Spaces (or in Russian)

ϵ-**entropy** of F in a metric space:

$$\mathcal{H}_\epsilon(F) = \log_2 \mathcal{N}_\epsilon(F)$$

ϵ-**capacity** of F in a metric space:

$$\mathcal{C}_\epsilon(F) = \log_2 \mathcal{M}_\epsilon(F)$$

Remark: By previous exercise, $\mathcal{C}_{2\epsilon}(F) \le \mathcal{H}_\epsilon(F) \le \mathcal{C}_\epsilon(F)$

Exercise: Show that $\frac{\mathcal{H}_\epsilon([0,1]^\nu)}{\log_2(1/\epsilon)}$ and $\frac{\mathcal{C}_\epsilon([0,1]^\nu)}{\log_2(1/\epsilon)}$ have finite limits as $\epsilon \to 0$, and find these limits.

Model expressiveness vs. stored information

General principle: Suppose we have a parametric model approximating the elements of F. Then, to be able to achieve the accuracy ϵ for each $f \in F$, the model must contain at least $\mathcal{H}_\epsilon(F)$ bits of information. If the model is implemented as a Boolean circuit, it must contain at least $\mathcal{H}_\epsilon(F) - 1$ elementary unary or binary logical operations.

Corollary[7]**:** Suppose F is approximated by a neural network with a fixed architecture and a fixed bitwise representation of the weights. Then, to approximate each $f \in F$ with accuracy ϵ, the total number of bits in all the weights must be not less than $\mathcal{H}_\epsilon(F)$.

[7] Petersen-Voigtlaender '18, '19

Kolmogorov-Tikhomirov '59: Let $F = B_{\nu,d,\infty}$ be a Sobolev ball in $\mathcal{W}^{d,\infty}([0,1]^\nu)$. Consider $B_{\nu,d,\infty}$ with the distance defined by the norm $\|\cdot\|_\infty$. Then

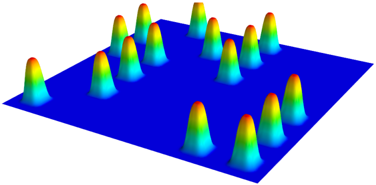
$$\mathcal{H}_\epsilon(B_{\nu,d,\infty}) \asymp \epsilon^{-\nu/d} \quad (\epsilon \to 0),$$

where $f \asymp g$ means that $cf \le g \le Cf$, for some constants $c, C > 0$.

Corollary: Any model providing uniform approximation accuracy ϵ on $B_{\nu,d,\infty}$ must contain at least $\asymp \epsilon^{-\nu/d}$ bits of information. If the model is implemented as a Boolean circuit, it must contain $\gtrsim \epsilon^{-\nu/d}$ elementary logical operations.

Proof of the lower bound

Use the grid of spike functions (again)



Fix some $\phi \in C^\infty(\mathbb{R}^\nu)$ such that $\phi(\mathbf{x}) = 0$ if $|\mathbf{x}| > \frac{1}{2}$, and $\phi(0) = 0$.
Consider the grid $G_N = \{\frac{1}{N}, \frac{2}{N}, \ldots, \frac{N}{N}\}^\nu \subset [0,1]^\nu$, with $|G_N| = N^\nu$.
For any assignment $\mathbf{c} : G_N \to \{0,1\}, \mathbf{c} = \{c_\mathbf{n}\}_{\mathbf{n}\in G_N}$, set

$$f_\mathbf{c} = CN^{-d} \sum_{\mathbf{n}\in G_N} c_\mathbf{n} \phi(N(\cdot - \mathbf{n}))$$

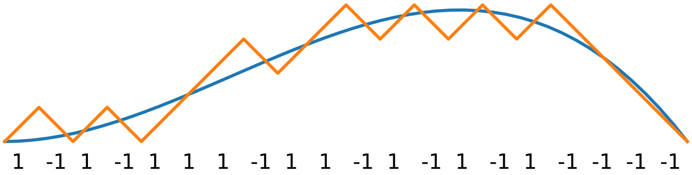
If C is small enough, then $f_\mathbf{c} \in B_{\nu,d,\infty}$ for all N and $\mathbf{c}$, and for $\mathbf{c} \neq \mathbf{c}'$ we have $\|f_\mathbf{c} - f_{\mathbf{c}'}\|_\infty \geq CN^{-d}$. Then, with $\epsilon = CN^{-d}$, capacity

$$\mathcal{C}_\epsilon(B_{\nu,d,\infty}) \geq \log_2(2^{N^\nu}) = N^\nu = C'\epsilon^{-\nu/d}$$

The upper bound: sketch of proof for $\nu = d = 1$

$B_{\nu=1,d=1,p=\infty}$: Lipschitz functions $(|f| \leq 1, |f'| \leq 1)$

Approximate f using a piecewise linear $\widetilde{f}$ with N nodes, with slopes ± 1



- Number of approximations: $O(N2^N)$
- Accuracy: $\epsilon \sim \frac{1}{N}$
- Hence $\mathcal{H}_\epsilon(B_{1,1,\infty}) \lesssim \log_2(N2^N) \sim N \sim \epsilon^{-1}$

For a higher smoothness d: use a smoother, Taylor-expansion-based approximation

Summary: Boolean vs. arithmetic complexity

Network size complexity of implementing ϵ-approximation of $f \in B_{\nu,d,\infty}$:

- As a Boolean circuit: $\gtrsim \epsilon^{-\nu/d}$
- As a ReLU neural network in a "classical mode": $O(\epsilon^{-\nu/d} \log 1/\epsilon)$
- As a neural network with special, complex acivation function: $O(1)$

Non-polynomial activations: Pfaffian functions

Challenge: How to estimate expressiveness of networks with non-(piecewise)-polynomial activations (e.g., logistic $x \mapsto e^x/(1+e^x)$)?

For VCdim bounds, how to bound the number of CC's?

A related question: Given an elementary function $f(x)$, can we bound the number of its roots (where $f(x) = 0$) based on some "complexity" of the description of f (like we did it for polynomials)?

Answer: the theory of *Pfaffian functions*[8]

[8] A. Khovansky, Fewnomials (Малочлены), 1991

Pfaffian functions: definition

A **Pfaffian chain** of analytic functions $f_1, \dots, f_l : U \subset \mathbb{R}^n \to \mathbb{R}$:

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

where P_{ij} are polynomials of degree $\leq \alpha$.

A **Pfaffian function**:

$$f(\mathbf{x}) = P(\mathbf{x}, f_1(\mathbf{x}), \dots, f_l(\mathbf{x})),$$

where P is a polynomial of degree β.

Pfaffian complexity: (α, β, l).

Important fact: all elementary function are Pfaffian on sutable domains

Exercise: The following functions are Pfaffian:

- polynomials on $U = \mathbb{R}^d$
- e^x on $\mathbb{R}$
- $\ln x$ on $\mathbb{R}_+$
- $\arcsin x$ on $(-1, 1)$

The function $\cos x$ is **not** Pfaffian on $\mathbb{R}$, but it is Pfaffian on any bounded interval (A, B), with complexity depending on $B - A$

Exercise: $\cos x$ is Pfaffian on $(-\pi, \pi)$ via the chain

$$\tan \frac{x}{2} \longrightarrow \cos^2 \frac{x}{2} \longrightarrow \cos x$$

Operations with Pfaffian functions

Exercise:

- Sums and products of Pfaffian functions f, g with a common domain U are Pfaffian
- If the domain of a Pfaffian function f includes the range of a Pfaffian function g, then the composition $f \circ g$ is Pfaffian on the domain of g

The complexity of the resulting functions $f + g, fg, f \circ g$ is determined by the complexity of the functions f, g.

Standard activations as (piecewise-)Pfaffian functions

Call a function $\sigma : \mathbb{R} \to \mathbb{R}$ **piecewise Pfaffian** if its domain can be divided into **finitely many** intervals on which σ is Pfaffian.

Most practical activations are Pfaffian or piecewise Pfaffian on $\mathbb{R}$, 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)

Main result

We call a solution $\mathbf{x} \in \mathbb{R}^d$ of a system $f_1(\mathbf{x}) = \ldots = f_d(\mathbf{x}) = 0$ *nondegenerate* if the respective Jacobi matrix $\frac{\partial f_i}{\partial x_j}(\mathbf{x})$ is nondegenerate.

Theorem (Khovanskii). Let $f_1, \ldots, f_d$ be Pfaffian d-variable functions on a domain $U \subset \mathbb{R}^d$ with a common Pfaffian chain of length l and respective degrees (α, β_i). Then the number of nondegenerate solutions of the system $f_1(\mathbf{x}) = \ldots = f_d(\mathbf{x}) = 0$ is bounded by

$$2^{l(l-1)/2}\beta_1 \cdots \beta_d \Big(\min(d, l)\alpha + \sum_{i=1}^{d} \beta_i - d + 1 \Big)^l.$$

Proof idea: use a generalized Rolle's lemma and bound the number of common zeros of the functions f_k by the number of common zeros of suitable polynomials (in a larger number of variables). The latter number can then be upper bounded using the classical Bézout theorem.

Some implications

Fixed-size networks with (piecewise-)Pfaffian activations:

- have a **finite** VC dimension (in contrast to, e.g., $\sin(ax)$)
- **cannot** approximate arbitrary continuous functions with arbitrary accuracy

Pfaffian functions and Betti numbers

Betti numbers $b_k(S), k = 0, 1, \ldots$ of a topological space S: numbers of "topological defects/holes" in S

$b_0(S)$: number of connected components in S

$b_0(S) \leq B(S) := \sum_k b_k(S)$ ("total number of defects")

Pfaffian set: $\cap_k \{\mathbf{x} \in U : f_k(\mathbf{x}) = 0\}$ with Pfaffian f_k
Semi-Pfaffian set: $\cap_k \{\mathbf{x} \in U : f_k(\mathbf{x})(= \text{or} >)0\}$ with Pfaffian f_k

Theorem (Zell '99)

Let S be a compact semi-Pfaffian set in $U \subset \mathbb{R}^n$, given on a compact Pfaffian set of dimension n', defined by s sign conditions on Pfaffian functions. If all the functions defining S have complexity at most (α, β, l), then

$$B(S) \leq s^{n'} 2^{l(l-1)/2} O\big((n\beta + \min(n, l)\alpha)^{n+l}\big)$$

Topological expressiveness of neural networks[9]

$S_{\mathcal{N}} : \{\mathbf{x} \in \mathbb{R}^n : f_{\mathcal{N}}(\mathbf{x}) > 0\}$

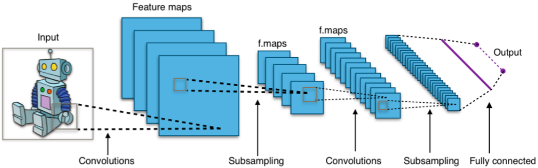
UPPER AND LOWER BOUNDS ON THE GROWTH OF $B(S_{\mathcal{N}})$ FOR NETWORKS WITH h HIDDEN UNITS, n INPUTS, AND l HIDDEN LAYERS. THE BOUND IN THE FIRST ROW IS A WELL-KNOWN RESULT AVAILABLE IN [26]

Inputs	**Layers**	**Activation function**	**Bound**
		Upper bounds	
n	3	threshold	$O(h^n)$
n	3	arctan	$O((n+h)^{n+2})$
n	3	polynomial, degree r	$\frac{1}{2}(2+r)(1+r)^{n-1}$
1	3	arctan	h
n	any	arctan	$2^{h(2h-1)}O((nl+n)^{n+2h})$
n	any	tanh	$2^{(h(h-1))/2}O((nl+n)^{n+h})$
n	any	polynomial, degree r	$\frac{1}{2}(2+r^l)(1+r^l)^{n-1}$
		Lower bounds	
n	3	any sigmoid	$(\frac{h-1}{n})^n$
n	any	any sigmoid	2^{l-1}
n	any	polynomial, deg. $r \geq 2$	2^{l-1}

[9] M. Bianchini, F. Scarselli, On the Complexity of Neural Network Classifiers: A Comparison Between Shallow and Deep Architectures, 2014

Expressiveness: future directions?

Practical neural networks work with complex multi-dimensional data



https://en.wikipedia.org/wiki/Convolutional_neural_network

Existing abstract approaches (VC dimension, approximation theory, etc.) do not quite fit these applications

The challenges:

- Describe relevant and mathematically natural spaces of dependencies?
- Explore the limits (infinitely deep/wide networks, infinite domain resolution, etc.)
- Explore particular structures (convnets, hierarchical models, etc.)