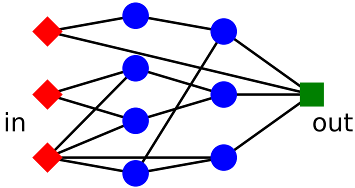
Expressiveness of neural networks

Dmitry Yarotsky

Feedforward neural networks



Implements a map $y = \widetilde{f}(\mathbf{x}, \mathbf{W})$ (or $y = \widetilde{f}(\mathbf{x})$ if $\mathbf{W}$ is fixed)

- $\mathbf{x} = (x_1, \ldots, x_\nu) \in \mathbb{R}^\nu$: 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)

The ReLU activation function

Exercise: Why does σ need to be nonlinear?

ReLU (Rectified Linear Unit):

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

Exercise (equivalence of piecewise linear activation functions)

- Suppose that $f : \mathbb{R}^\nu \to \mathbb{R}$ is implemented by a NN with some 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 piecewise-linear activation function σ_1, f can be implemented by a σ_1-NN.

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)

Approximation with one-hidden-layer networks

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

A one-hidden-layer network:

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

Network complexity: $\sim N$

Approximation in L^p norms

In a nutshell: Given f, find $\widetilde{f}$ with small $\|f - \widetilde{f}\|$

L^p norms:

$$\|f\|_p = \begin{cases} \left(\int_\Omega |f(\mathbf{x})|^p d\mathbf{x}\right)^{1/p}, & 1 \le p < \infty \\ \max_{\mathbf{x}\in\Omega} |f(\mathbf{x})|, & p = \infty \end{cases}$$

Exercise: Show that $\|f\|_\infty = \lim_{p\to+\infty} \|f\|_p$

$$L^p(\Omega) = \{f : \|f\|_p < \infty\}$$

Exercise: $L^p(\Omega)$ is a Banach space (normed + metrically complete)

Uniform approximation on compact sets

Let $f : \mathbb{R}^\nu \to \mathbb{R}$

A *compact* set in $\mathbb{R}^\nu$: bounded + closed

Approximation on compact sets: for any compact $K \subset \mathbb{R}^\nu$ and $\epsilon > 0$ find $\widetilde{f}$ such that $\|f - \widetilde{f}\|_\infty < \epsilon$

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) $\Longrightarrow$ 2)

The nontrivial part: 2) $\Longrightarrow$ 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

Weierstrass and Stone-Weierstrass theorems

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 $C([-a, a]^\nu, \mathbb{R})$

Special case of ReLU σ**:** approximate f by a linear spline $\widetilde{f}$ and write

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

Special case: $\sigma \in C^\infty(\mathbb{R})$

Proposition

Suppose $\sigma \in C^\infty(a,b)$ and σ is not a polynomial on (a,b). Then there exists $x_0 \in (a,b)$ such that all derivatives $\frac{d^n\sigma}{dx^n}(x_0) \neq 0, n = 0, 1, \dots$
(**Exercise***: prove it.)

Then, any monomial $(x - x_0)^n$ can be approximated by expressions $\sum_{k=0}^n c_k \sigma(w_k x + h_k)$:

$$\begin{aligned}
\sigma(x_0 + w(x - x_0)) &= \sigma(x_0) + o(1) \\
\frac{1}{w}(\sigma(x_0 + w(x - x_0)) - \sigma(x_0)) &= \tfrac{d\sigma}{dx}(x_0)(x - x_0) + o(1) \\
\frac{1}{w^2}(\sigma(x_0 + 2w(x - x_0)) - 2\sigma(x_0 + w(x - x_0)) + \sigma(x_0)) &= \tfrac{d^2\sigma}{dx^2}(x_0)(x - x_0)^2 + o(1) \\
&\dots
\end{aligned}$$

where $w \to 0$

Remark: this approximation uses small weights w_k and large c_k

General nonpolynomial $\sigma \in C(\mathbb{R})$

Suppose that some x^m cannot be approximated by $\sum_n c_n \sigma(w_n \cdot + h_n)$

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 $\sum_n c_n \sigma(w_n \cdot + h_n)$

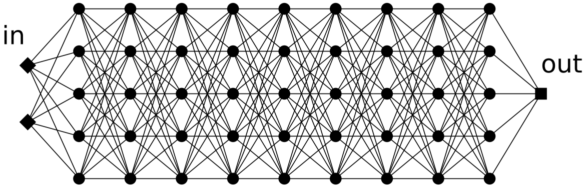
From the argument for smooth σ_ϕ :

$$\frac{d^m \sigma_\phi}{dx^m}(0) x^m = 0,$$

i.e. $\frac{d^m \sigma_\phi}{dx^m}(0) = 0$. By shifting ϕ, $\frac{d^m \sigma_\phi}{dx^m}(x) = 0$ for all x, i.e. σ_ϕ is a polynomial of degree $< m$. Hence σ is a polynomial of degree $< m$.

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*: a connected component of $\widetilde{f}^{-1}(\mathbf{a})$ for some $\mathbf{a} \in \mathbb{R}^\nu$

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 level set of $\widetilde{f}$ that is bounded 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 level sets are contained in S

Suppose $\mathbf{x} \in \widetilde{f}^{-1}(\mathbf{a})$ is not in S, then, 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 $\widetilde{f}_j$ be the action of first j hidden layers:

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

$$\mathrm{ReLU}(z_1, \ldots, z_\nu) = ((z_1)_+, \ldots, (z_\nu)_+)$$

Then $\mathrm{ReLU}^{-1}(\widetilde{f}_k(\mathbf{x}))$ contains an infinite ray R.

Then $\widetilde{f}^{-1}(\mathbf{a}) \supset (A_k A_{k-1} \cdots A_1)^{-1} R \ni \mathbf{x}$, and $(A_k A_{k-1} \cdots A_1)^{-1} R$ is unbounded since $A_k A_{k-1} \cdots A_1 : \mathbb{R}^\nu \to \mathbb{R}^\nu$. (**Remark:** this wouldn't be true for $H > \nu$.) $\square$

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.
- What are the minimal networks widths for other activation functions?

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.

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 $f \in C_0^\infty(\Omega)$, 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 $f(x) \equiv 1$. Then $f \in \mathcal{W}^{0,\infty}([0,1])$ in the sense of Def.2, but not Def.1.

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 $d-1$.

Parametric approximations

Suppose we want to approximate functions from a set $K \subset \mathcal{F}$, where $\mathcal{F}$ is a normed space (e.g., $\mathcal{F} = C([0,1])$).

A *parametric approximation with W parameters*: $M_W : \mathbb{R}^W \to \mathcal{F}$.
(Example: a neural network with W weights. The weights are varied, the architecture is fixed.)

But this is too general:
Exercise: Let K be <u>compact</u>. Then, for $W = 1$, there is a smooth maps M_W such that $K \subset \overline{M_W(\mathbb{R})}$.

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

A *parameter assignment map* $P_W : K \to \mathbb{R}^W$. (Example: network weight assignment.)

Full approximation pipeline: given $f \in K$,

$$f \mapsto M_W(P_W(f))$$

Key requirement: parameter assignment P_W is continuous
(Remark: no assumption on M_W)

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

Optimal approximation

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

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

Key result: Let K be a ball in $\mathcal{W}^{d,p}([0,1]^\nu)$. Then $h_W \asymp W^{-d/\nu}$.

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 $p = \infty$.

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:

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

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

Lemma (Borsuk-Ulam antipodality theorem)

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

Exercise: prove for $n = 1$.

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

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

Then,
$\sup_{f \in K} \|f - M_W(P_W(f))\|_\infty \ge \frac{1}{2}\|\Phi_N(\mathbf{x}) - \Phi_N(-\mathbf{x})\|_\infty = CN^{-d}\|\phi\|_\infty$.

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

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:

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

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

Exercise: P_W is continuous on K

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

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

(by a Taylor remainder bound)

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

Do neural networks achieve optimal approximation rates?

For ReLU networks, we'll show:

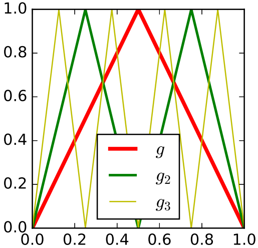
- Yes – for deep networks
- No – for shallow networks

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”:

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

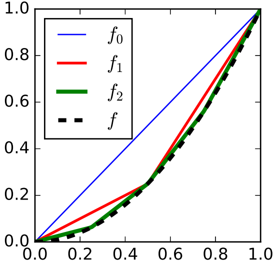


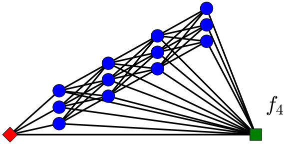
Let

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

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)

$M \le (2U)^L$

Proof. By induction. For $n \le L$, 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 kinks). 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 kink in this interval. Therefore, $N_{n+1} \le (U_n + 1)N_n$, where U_n is the number of neurons in the n'th layer. So, $N_{L+1} \le (U_1+1)(U_2+1)\cdots(U_L+1) \le (2U)^L$. □

Slow approximation of $f(x) = x^2$ by fixed-depth ReLU networks

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 $U \geq \frac{1}{2}(8\epsilon)^{-1/(2L)}$. $\square$

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-Chevonenkis (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

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

$S \subset X$ 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 $m \leq \mathrm{VCdim}(H)$

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)

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

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

Theorem (Pajor)

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

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

Proof of Pajor theorem

Let $\mathcal{F} = H|_S$. Proof by induction in $|\mathcal{F}|$. 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.

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 $|A_0| + |A_k| \geq |\mathcal{F}_0| + |\mathcal{F}_1| = |\mathcal{F}|$. 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 $S = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathbb{R}^{\nu+1}$. *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 $\big|H|_S\big| \leq \mathrm{CC}(\mathbb{R}^{\nu+1} \setminus \cup_{i=1}^m P_i)$.

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 $P_{m+1} \setminus \cup_{i=1}^m P_i$. $\square$

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

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

$\square$

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

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

- Show that the perceptron shatters the set $\{\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_\nu\}$ and hence $\mathrm{VCdim} \geq \nu + 1$
- Show that $\mathrm{VCdim} \leq \nu + 1$ 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

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

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

Deep networks

Existing results for deep ReLU and piecewise linear networks[1]:

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

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.)[2]

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

[2] 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*.

Let $H = \{\mathrm{sgn}(f) | f : \mathbb{R}^W \times X \to \mathbb{R}\}$. Then

$$\Pi_H(m) = \sup_{S:|S|=m} \big|H|_S\big| \leq \sup_{S=\{\mathbf{x}_1,\ldots,\mathbf{x}_m\}} \mathrm{CC}(\mathbb{R}^W \setminus \cup_{i=1}^m P_i),$$

where $P_i = \{\mathbf{w} \in \mathbb{R}^W : f(\mathbf{x}_i, \mathbf{w}) = h_i\}$.

How to count these CC?

Solution set component bounds

A set G of functions $f:\mathbb{R}^W\to\mathbb{R}$ has *solution set component bound (SSCB)* B if for any $1\le k\le W$ and any $f_1,\ldots,f_k\subset G$ that have regular zero-set intersections[3] we have

$$\mathrm{CC}\left(\cap_{i=1}^k\{\mathbf{w}\in\mathbb{R}^W: f_i(\mathbf{w})=0\}\right)\le B.$$

Theorem

Let F be a family of smooth functions $f:\mathbb{R}^W\times X\to\mathbb{R}$ and $H=\{\mathrm{sgn}(f): f\in F\}$. Suppose F is closed under addition of constants and $G=\{\mathbf{w}\mapsto f(\mathbf{w},\mathbf{x})|\mathbf{x}\in X\}$ has a SSCB B. Then

$$\Pi_H(m)\le B\sum_{k=1}^W\binom{m}{k}\le B\Big(\frac{em}{W}\Big)^W$$

for $m\ge W$.

[3] Some nondegeneracy assumption

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.

Application to solution set components bound

Proposition

For any l, the set of degree l polynomials defined on $\mathbb{R}^W$ has solution set components bound $B = 2(2l)^W$.

Proof: Given k degree-l polynomials $f_1, \ldots, f_k$, set $f = \sum_{n=1}^k f_n^2$. Then

$$\cap_{i=1}^k \{\mathbf{w} \in \mathbb{R}^W : f_i(\mathbf{w}) = 0\} = \{\mathbf{w} \in \mathbb{R}^W : f(\mathbf{w}) = 0\}.$$

Therefore, B can be upper bounded by using above theorem with degree $2l$.

Exercise: Let $H_{L,W,d}$ be the family of neural networks of a fixed architecture that has L layers, W weights, purely polynomial activation functions of degree d, and the threshold sgn at the output. Show that $\mathrm{VCdim}(H_{L,W,d}) \leq CWL \ln(d+1)$ with some universal constant C.

The "bit extraction technique" (Bartlett, Maiorov, Meir (1998))

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: $a = 0.a_1 a_2 \dots a_N$ with $a_n \in \{0, 1\}$
- Construct a finite network that maps $0.a_1 a_2 \ldots \mapsto (a_1, 0.a_2 a_3 \ldots)$ (i.e., $a \mapsto (\lfloor 2a \rfloor, 2a - \lfloor 2a \rfloor)$)
- By stacking, construct a depth-$O(N)$ network extracting all bits: $a \mapsto (a_1, a_2, \ldots, a_N)$
- 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

The "bit extraction technique" (cont-d)

- 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 $y = \sum_{n=1}^{N} a_n z_n$. 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
- 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 $\mathrm{VCdim} \geq cWL$

Non-polynomial activations: Pfaffian functions

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

Bounds on CC's are available for *Pfaffian functions*[4]

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

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

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

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).

---

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

Properties of Pfaffian functions

Exercise: The logistic function is Pfaffian

General properties:

- The set of Pfaffian functions is closed under arithmetic operations and compositions
- Elementary functions are Pfaffian on suitable domains (e.g. $\cos x$ is Pfaffian on $(-\pi, \pi)$ via the chain $\tan \frac{x}{2} \longrightarrow \cos^2 \frac{x}{2} \longrightarrow \cos x$)

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

Betti numbers $b_k(S), k = 0, 1, \dots$ 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")

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[5]

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

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}

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

Fastest approximations with ReLU nets[6]

Assume $f \in C([0,1]^\nu)$, characterized by modulus of continuity:

$$\omega_f(r) = \max\{|f(\mathbf{x}) - f(\mathbf{y})| : |\mathbf{x} - \mathbf{y}| \le r\}$$

Let $\widetilde{f}_W$ be a ReLU neural network approximation with W weights

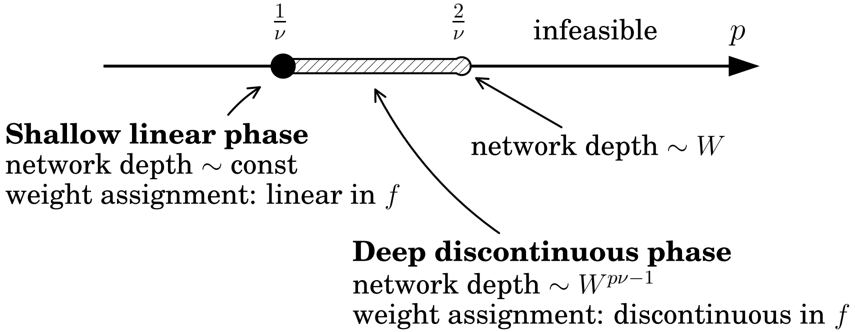
For which p can we achieve the convergence rate

$$\boxed{\|f - \widetilde{f}_W\|_\infty = O(\omega_f(O(W^{-p}))} \ ?$$

[6] D. Yarotsky, Optimal approximation of continuous functions by very deep ReLU networks, arXiv:1802.03620

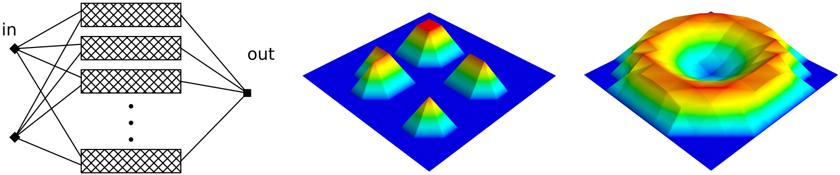
The answer: a phase diagram

$$\|f - \widetilde{f}_W\|_\infty = O(\omega_f(O(W^{-p}))$$



The shallow linear phase ($p = \frac{1}{\nu}$)

Approximation can be formed by a linear combination of “spikes” and implemented by a fixed-depth network consisting of $O(W)$ parallel blocks



The weight assignment is linear and continuous in f

Beyond the linear phase

From general results on VC dims and continuous parametric approximation[7]:

- Rates with $p > \frac{2}{\nu}$ are infeasible
  Let S be the $N \times \cdots \times N$ grid, then $|S| = N^\nu$ and S can be shattered if $\frac{1}{N} \gtrsim W^{-p}$. But we know that VCdim $\lesssim LW \lesssim W^2$, so $W^{p\nu} \lesssim \text{VCdim} \lesssim W^2$
- Rates with $p > \frac{1}{\nu}$ are infeasible if weight assignment is continuous in f
  Special case of the optimal rate $W^{-d/\nu}$ with $d = 1$
- Rates with $p > \frac{1}{\nu}$ are infeasible by networks of depth $\lesssim W^{p\nu - 1}$
  Follows by trying to shatter the grid and using VCdim $\lesssim LW$

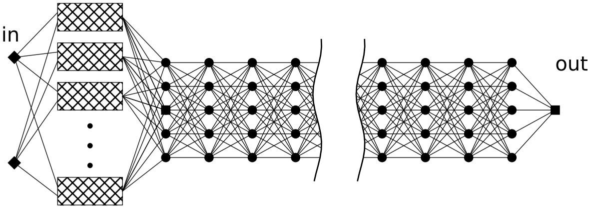
[7] Goldberg & Jerrum '95, DeVore et al. '89, Bartlett et al. '17

Existence of the deep discontinuous phase

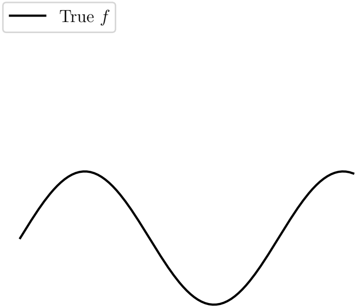
Theorem

For any $p \in (\frac{1}{\nu}, \frac{2}{\nu}]$, the approximation rate can be achieved using architectures of depth $L = O(W^{p\nu-1})$

- *For $p = \frac{2}{\nu}$: use fully-connected architectures of constant width $2\nu + 10$*
- *For $p \in (\frac{1}{\nu}, \frac{2}{\nu})$: use parallel shallow architectures stacked with fully-connected architectures of width $3^\nu(2\nu + 10)$ and depth $W^{p\nu-1}$*

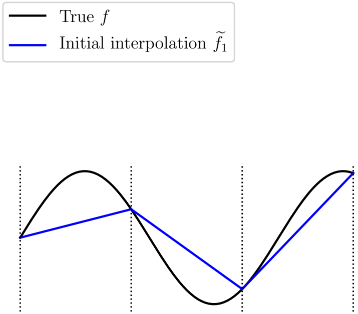


The full approximation is the sum of two parts:



The full approximation is the sum of two parts:

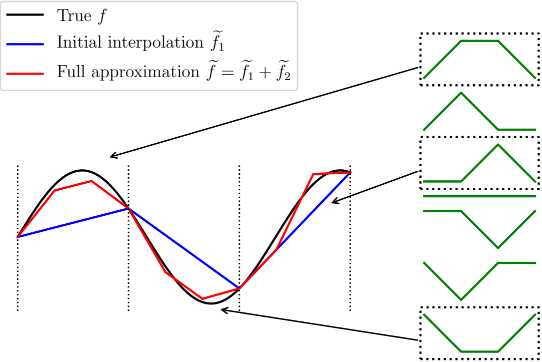
- $\widetilde{f}_1$: piecewise-linear interpolation of f on the length scale $\frac{1}{N} \sim W^{-1/\nu}$



Proof ideas: two-scales approximation

The full approximation is the sum of two parts:

- $\widetilde{f}_1$: piecewise-linear interpolation of f on the length scale $\frac{1}{N} \sim W^{-1/\nu}$
- $\widetilde{f}_2$: *discrete* approximation on the smaller length scale $\frac{1}{M} \sim W^{-p}$
  - Use a *finite set* of candidate shapes
  - In each patch of size $\frac{1}{N}$, fit one of the shapes to $f - \widetilde{f}_1$



Proof ideas: network implementation

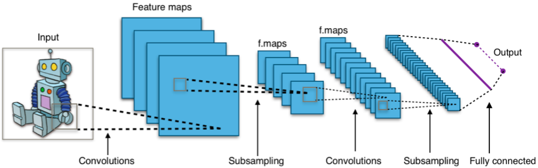
- Encode and store the $\widetilde{f}_2$ shape in each patch using a single network weight

 $\longleftrightarrow b = 0.102$

- When computing $\widetilde{f}_2(\mathbf{x})$, use the bit extraction technique to recover the shape from the special weight

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

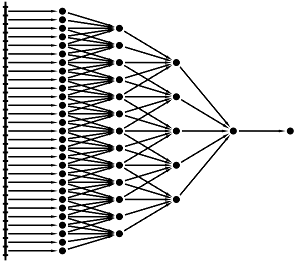
The challenges:

- 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.)

Example: a universal approximation theorem for maps on infinite-dimensional spaces[8]

Theorem

A map $f : L^2(\mathbb{R}^\nu) \to \mathbb{R}$ is a limit point of convnets with donsampling if and only if f is continuous in the norm topology.



[8] D. Yarotsky, Universal approximations of invariant maps by neural networks, arXiv:1804.10306