Linear and Nonlinear Schemes for Forward Model Reduction and Inverse Problems

Olga Mula (TU Eindhoven)

Cemracs 2023

2023-07-17

Agenda

1. Elements of approximation theory
    - Linear and Nonlinear Approximation
    - Nonlinear approximation with Neural Networks
2. Forward Problem: Reduced Order Modelling of parametrized PDEs
    - Linear and Nonlinear MOR
    - Role of geometry
3. Inverse Problems
    - Optimal linear and nonlinear algorithms for State Estimation
    - Role of Geometry
4. Hands-on session with Agustin Somacal

Material

- Slides: `www.olgamula.com`
- Notebook:

  `https://github.com/agussomacal/ROMHighContrast/tree/NonLinearROM`
- Lecture Notes: **Inverse Problems: A Deterministic Approach using Physics-Based Reduced Models**, OM.

Part I.1

Elements of approximation theory

Linear and Nonlinear Approximation

Approximation of a function

Let $(V, \|\cdot\|)$ be a Banach/Hilbert space.

Examples: $V = \mathbb{R}^d, L^p(\Omega), W^{k,p}(\Omega)$, with $\Omega \subset \mathbb{R}^d$.

The goal of approximation is to replace a target function $u \in V$ by a simpler function (easy to work with in practice).

An approximation is searched in a set of functions $V_n \subset V$, where n is related to some complexity measure, typically the number of parameters.

Approximation of a function

We distinguish:

- linear approximation when V_n is a finite-dimensional linear space

$$V_n = \text{span}\{v_i\}_{i=1}^{n} = \left\{ \sum_{i=1}^{n} c_i v_i \ : \ c_i \in \mathbb{R} \right\}$$

where the $\{v_i\}_{i=1}^{n}$ form a basis of V_n.

Examples of v_i: polynomials, trigo. polynomial, fixed knot splines...

Approximation of a function

We distinguish:

- linear approximation when V_n is a finite-dimensional linear space

$$V_n = \mathrm{span}\{v_i\}_{i=1}^n = \left\{ \sum_{i=1}^n c_i v_i \ : \ c_i \in \mathbb{R} \right\}$$

  where the $\{v_i\}_{i=1}^n$ form a basis of V_n.

  Examples of v_i: polynomials, trigo. polynomial, fixed knot splines...

- nonlinear approximation when V_n is a nonlinear set. Examples:
  - n-term approximation

$$V_n = \left\{ \sum_{i=1}^n c_i v_i \ : \ c_i \in \mathbb{R},\ v_i \in \mathcal{D} \right\}$$

    where $\mathcal{D} = \{v_i\}_{i=1}^\infty$ is a dictionary of functions

Examples: rational functions, free-knot splines, neural networks, tensor networks...

Approximation of a function

We distinguish:

- linear approximation when V_n is a finite-dimensional linear space

$$V_n = \operatorname{span}\{v_i\}_{i=1}^n = \left\{ \sum_{i=1}^n c_i v_i \ : \ c_i \in \mathbb{R} \right\}$$

  where the $\{v_i\}_{i=1}^n$ form a basis of V_n.

  Examples of v_i: polynomials, trigo. polynomial, fixed knot splines...

- nonlinear approximation when V_n is a nonlinear set. Examples:
  - n-term approximation

$$V_n = \left\{ \sum_{i=1}^n c_i v_i \ : \ c_i \in \mathbb{R},\ v_i \in \mathcal{D} \right\}$$

    where $\mathcal{D} = \{v_i\}_{i=1}^\infty$ is a dictionary of functions

  - nonlinear parametric manifold:

$$V_n = \{\mathrm{D}(c) \ : \ c \in \mathbb{R}^n\}$$

    with some given nonlinear map $\mathrm{D} : \mathbb{R}^n \to V$.

Examples: rational functions, free-knot splines, neural networks, tensor networks...

Encoding

We take the view that nonlinear approximation methods depending on n parameters are built on two mappings:

- An encoder mapping
$$\mathrm{E} : V \to \mathbb{R}^n$$
which when given $u \in V$ chooses n parameters $\mathrm{E}(u) \in \mathbb{R}^n$ to represent u.

- A decoder mapping
$$\mathrm{D} : \mathbb{R}^n \to V$$
which maps a vector $c \in \mathbb{R}^n$ back into V and is used to build the approximation of u.

  The set
$$V_n := \mathrm{Im}(\mathrm{D}) = \{\mathrm{D}(c) \ : \ c \in \mathbb{R}^n\} \subset V$$
is viewed as a parametric manifold.

Remark: The encoder-decoder scheme describes a process of dimensionality reduction.

Examples of E, D

- **Linear orthogonal projection:**
  - V Hilbert, $V_n = \mathrm{span}\{v_i\}_{i=1}^n$, $\{v_i\}$ ONB
  - Choose $\mathrm{E}(u) = (\langle u, v_i \rangle)_{i=1}^n$.
  - Choose $\mathrm{D}(c) = \sum_{i=1}^n c_i v_i \in V_n$
  - $u \approx \mathrm{D}(\mathrm{E}(u)) = \sum_{i=1}^n \langle u, v_i \rangle \, v_i = P_{V_n} u$
- **Compressed sensing:**
  - $V = \mathbb{R}^N$ with $N \gg 1$.
  - Σ_k^N: Vectors of $\mathbb{R}^N$ with at most k non zero entries.
  - Task: Approximate $u \in \Sigma_k^N$ from m observations $\Phi_1^T u, \dots, \Phi_m^T u$ (with $1 \leq m \leq N$).
  - Choose $\mathrm{E}(u) = \Phi^T u$ with $\Phi = [\Phi_1 | \dots | \Phi_m]$
  - Choose $\mathrm{D}(c) = \arg\min_{v \in \mathbb{R}^N} \{ \|v\|_{\ell_1} \; : \; \Phi^T v = c\} \; \in \Sigma_{\tilde{k}}^N$
  - $u \approx \mathrm{D}(\mathrm{E}(u)) \in \Sigma_{\tilde{k}}^N$
- **Neural Networks** (see later on).

Error of best approximation

Let $(V, \|\cdot\|)$ be a normed space, V_n a given approximation set.

For a given target function $u \in V$, the error of best approximation

$$e_n(u) = \inf_{v \in V_n} \|u - v\|$$

quantifies the best we can expect from V_n.

Fundamental problems in approximation

For a sequence $(V_n)_{n\geq 1}$ of sets of growing complexity:

- (universality) Does $e_n(u) \to 0$ as $n \to \infty$ for all $u \in V$?
- (expressivity) For a certain class of functions $\mathcal{K} \subset V$, determine how fast $e_n(u) \to 0$, or determine the complexity

$$n = n(\varepsilon, \mathcal{K}) \quad \text{s.t.} \quad e_n(u) \leq \varepsilon.$$

Typically, we search for

$$e_n(u) \leq C\gamma(n)^{-1}$$

where $C \geq 1$, γ is a stricly increasing growth function, and

$$n(\varepsilon, u) \geq \gamma^{-1}(\varepsilon / C).$$

Fundamental problems in approximation

- (approximation classes) Characterize the class of functions for which a certain convergence type is achieved, e.g.,

  $$\mathcal{A}^\gamma = \{u \in V \ : \ \sup_{n\geq 1} \gamma(n) e_n(u) < +\infty\}$$

  for some growth function γ.

- (algorithm) Construct an approximation $u_n \in V_n$ such that

  $$\|u - u_n\| \leq C e_n(u) = C \inf_{v \in V_n} \|u - v\|$$

  with C independent of n.

  Algorithms depend on the available information, e.g., given by observations such as point evaluations, or equations satisfied by the functions.

Optimal approximation of a model class

Suppose now that we want to approximate all functions u from a compact subset $\mathcal{K} \subset V$ (model class).

The quality of an encoding procedure with n parameters can be measured as:

- In the worst case sense:

$$\mathcal{E}^{\mathsf{wc}}(\mathcal{K}; (\mathrm{E}, \mathrm{D})) := \max_{v \in \mathcal{K}} \| v - \mathrm{D}(\mathrm{E}(v)) \|$$

- On average:

$$\mathcal{E}^{\mathsf{av}}(\mathcal{K}; (\mathrm{E}, \mathrm{D})) := \int_{\mathcal{K}} \| v - \mathrm{D}(\mathrm{E}(v)) \| \mathrm{d}\rho(v)$$

and we can thus define a notion of n-width

$$\inf_{\substack{(\mathrm{E},\mathrm{D}) \\ \mathrm{E}: V \to \mathbb{R}^n \\ \mathrm{D}: \mathbb{R}^n \to V}} \quad \mathcal{E}^{\star}(\mathcal{K}; (\mathrm{E}, \mathrm{D})), \quad \star = \{\mathsf{wc}, \mathsf{av}\}$$

by optimizing the choice of (E, D) under specific restrictions.

Notions of n-widths

$$\inf_{(\mathrm{E},\mathrm{D})} \max_{v \in \mathcal{K}} \|v - \mathrm{D}(\mathrm{E}(v))\|.$$

We distinguish:

- **Approximation numbers** $a_n(\mathcal{K})$: Both E, D are linear:

$$a_n(\mathcal{K}) := \inf_{\substack{L \text{ linear} \\ \mathrm{rank}(L) = n}} \max_{v \in \mathcal{K}} \|v - L(v)\|,$$

Notions of n-widths

$$\inf_{(\mathrm{E},\mathrm{D})} \max_{v\in\mathcal{K}} \|v - \mathrm{D}(\mathrm{E}(v))\|.$$

We distinguish:

- **Approximation numbers** $a_n(\mathcal{K})$: Both E, D are linear:

$$a_n(\mathcal{K}) := \inf_{\substack{L \text{ linear} \\ \operatorname{rank}(L)=n}} \max_{v\in\mathcal{K}} \|v - L(v)\|,$$

- **Kolmogorov** n**-width** $d_n(\mathcal{K})$: Only D is linear

$$d_n(\mathcal{K}) := \inf_{\dim(V_n)=n} \operatorname{dist}(\mathcal{K}, V_n) = \inf_{\dim(V_n)=n} \sup_{u\in\mathcal{K}} \inf_{v\in V_n} \|u - v\|_V$$

We have

$$d_n(\mathcal{K}) \leq a_n(\mathcal{K}),$$

and equality holds in Hilbert spaces.

$d_n(\mathcal{K})$ measures how well the set $\mathcal{K}$ can be approximated (uniformly) by an n-dimensional space.

Useful in forward MOR.

Notions of n-widths

$$\inf_{(\mathrm{E},\mathrm{D})} \max_{v\in\mathcal{K}} \|v - \mathrm{D}(\mathrm{E}(v))\|.$$

We distinguish:

- **Approximation numbers** $a_n(\mathcal{K})$: Both E, D are linear:
- **Kolmogorov *n*-width** $d_n(\mathcal{K})$: Only D is linear
- **Sensing numbers** $s_n(\mathcal{K})$: Only E is linear:

$$s_n(\mathcal{K}) := \inf_{\mathrm{D},\lambda_1,\dots,\lambda_n} \max_{u\in\mathcal{K}} \|u - \mathrm{D}(\lambda_1(u),\dots,\lambda_n(u))\|$$

where the inf runs over all D maps and $\lambda_i \in V'$.
Useful in inverse problems.

Notions of n-widths

$$\inf_{(\mathrm{E},\mathrm{D})} \max_{v\in\mathcal{K}} \|v - \mathrm{D}(\mathrm{E}(v))\|.$$

We distinguish:

- **Approximation numbers** $a_n(\mathcal{K})$: Both E, D are linear:
- **Kolmogorov** n**-width** $d_n(\mathcal{K})$: Only D is linear
- **Sensing numbers** $s_n(\mathcal{K})$: Only E is linear:
- **Manifold width** $\delta_n(\mathcal{K})$: Both E, D are nonlinear.

Notions of n-widths

$$\inf_{(\mathrm{E},\mathrm{D})} \max_{v\in\mathcal{K}} \|v - \mathrm{D}(\mathrm{E}(v))\|.$$

We distinguish:

- **Approximation numbers** $a_n(\mathcal{K})$: Both E, D are linear:
- **Kolmogorov** ***n*****-width** $d_n(\mathcal{K})$: Only D is linear
- **Sensing numbers** $s_n(\mathcal{K})$: Only E is linear:
- **Manifold width** $\delta_n(\mathcal{K})$: Both E, D are nonlinear.
- **Stable manifold width** $\delta_n^{\text{cont},\gamma}(\mathcal{K})$: for robustness against noise perturbations,

$$\delta_n^{\text{cont},\gamma}(\mathcal{K}) := \inf_{(\mathrm{E},\mathrm{D})} \sup_{u\in\mathcal{K}} \|u - \mathrm{D}(\mathrm{E}(u))\|,$$

where E, D are seached among Lipschitz cont. mappings,

$$\|\mathrm{E}(u) - \mathrm{D}(v)\|_{\mathbb{R}^n} \le \gamma\|u - v\|_V, \quad \|\mathrm{D}(c) - \mathrm{D}(q)\|_V \le \gamma\|c - q\|_{\mathbb{R}^n}.$$

This concept is studied in [DHM89, CDPW21].

Notions of n-widths

$$\inf_{(\mathrm{E},\mathrm{D})} \max_{v\in\mathcal{K}} \|v - \mathrm{D}(\mathrm{E}(v))\|.$$

We distinguish:

- **Approximation numbers** $a_n(\mathcal{K})$: Both E, D are linear:
- **Kolmogorov** ***n*****-width** $d_n(\mathcal{K})$: Only D is linear
- **Sensing numbers** $s_n(\mathcal{K})$: Only E is linear:
- **Manifold width** $\delta_n(\mathcal{K})$: Both E, D are nonlinear.
- **Stable manifold width** $\delta_n^{\mathrm{cont},\gamma}(\mathcal{K})$: for robustness against noise perturbations,
- We have:

$$a_n(\mathcal{K}) \geq \{d_n(\mathcal{K}), s_n(\mathcal{K})\} \geq \delta_n(\mathcal{K}),$$
$$a_n(\mathcal{K}) \geq \delta_n^{\mathrm{cont},\gamma}(\mathcal{K}) \geq \delta_n(\mathcal{K})$$

Optimal linear approximation: Kolmogorov widths

The Kolmogorov n-width of $\mathcal{K}$ is defined as

$$d_n(\mathcal{K}) := \inf_{\dim(V_n)=n} \operatorname{dist}(\mathcal{K}, V_n) = \inf_{\dim(V_n)=n} \sup_{u\in\mathcal{K}} \inf_{v\in V_n} \|u - v\|_V$$

where the infimum runs over all linear subspaces V_n of dimension n.

If $\mathcal{K}$ is equipped with a probability measure $\pi \in \mathcal{P}(\mathcal{K})$, we can define a weighted Kolmogorov n-width,

$$d_n^{(p,\pi)} := \inf_{\dim(V_n)=n} \left(\int_{\mathcal{K}} \inf_{v\in V_n} |u - v\|_V^p \mathrm{d}\pi(u) \right)^{1/p}$$

If V is a Hilbert space, $p = 2$ and π the push-forward measure of a random variable $U \in L^2(\Omega, V)$, this is equivalent to

$$\inf_{\dim(V_n)=n} \int_{\mathcal{K}} \|u(\omega, \cdot) - P_{V_n} u(\omega, \cdot)\|_V^2 \mathrm{d}\omega$$

and an optimal space V_n is given by **Singular Value Decomposition**.

Bounds for Kolmogorov n-width

For $V = L^p(\Omega)$, $\Omega = [0,1]^d$, $1 \leq p \leq \infty$, and $\mathcal{K}$ the unit ball of $W^{k,p}(\Omega)$:

$$d_n(\mathcal{K}) \sim n^{-k/d}$$

and optimal performance is obtained by fixed-knot splines (with degree adapted to the regularity).

We observe:

- The curse of dimensionality: Rate degrades as d increases.
- The blessing of smoothness: Improvement of the rate of approximation when k increases.
- Few results beyond the classical smoothness classes. However, some crucial results exist for $\mathcal{K}$ generated by parametric PDEs: we will discuss them in the MOR part.

Bounds for manifold n-width

For $V = L^p(\Omega)$, $\Omega = [0,1]^d$, $1 \leq p \leq \infty$, and $\mathcal{K}$ the unit ball of $W^{k,p}(\Omega)$ or Besov spaces $B_p^k(L^\tau)$ which compactly embed in L^p, it holds

$$\delta_n^{\text{cont},\gamma}(\mathcal{K}) \sim n^{-k/d}$$

and optimal performance is obtained by, e.g., **free-knot splines** or best n-term approximation with a dictionary of tensor products of dilated splines. [DHM89]

We observe:

- We cover a larger class of sets than with linear methods.
- For approximation with continuous parameters, the curse of dimension is unavoidable. What happens for discontinuous parameters?
- Extra regularity does not help: curse of dimension even for $\mathcal{K} \subset \mathcal{C}^\infty(\Omega)$.
- These rates do not explain behavior of neural networks. ⇒ Need to consider different approximation classes.

Part I.2

Elements of Approximation Theory

The class V_n of Neural Networks

Let us avoid alchemy

To avoid that Machine Learning and DNN become the alchemy of our century, Ali Rahimi asked in 2017 for more rigor, and more theory in his Test-of-Time award's speach. There is still much to do.

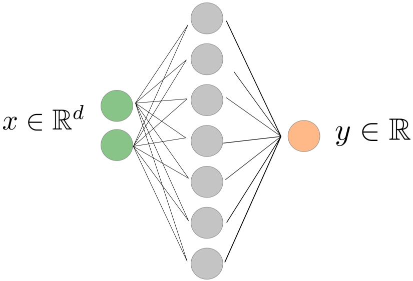
Shallow Neural Networks

A shallow neural network (with one hidden layer of width m) is a function

$$f : \mathbb{R}^d \to \mathbb{R}$$

$$x \mapsto f(x) := a^T \sigma(Ax + b) = \sum_{i=1}^{m} a_i \sigma \left(\sum_{j=1}^{d} A_{ij} x_j + b_i \right)$$

where $a \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times d}$, and σ is a given nonlinear function (activation function).



Remark that we can view NN as a nonlinear decoder:

$$\mathrm{D} : \mathbb{R}^n \to V = \mathbb{F}(\mathbb{R}^d, \mathbb{R})$$

$$c = \{a, A, b\} \mapsto \mathrm{D}(c) = f$$

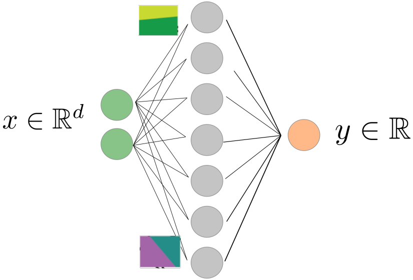
Shallow Neural Networks

Classical piecewise polynomial activation functions:

- ReLU function $\sigma(t) = \max\{0, t\}$
- RePU(p) function $\sigma(t) = \max\{0, t\}^p$

ReLU and RePU networks produce free-knot splines: they are a piecewise polynomial approximation on a free partition of $\mathbb{R}^d$ determined by m hyperplanes:

$$H_i = \{x \in \mathbb{R}^d \ : \ w_i^T x + b_i = 0\}, \quad w_i = (A_{ij})_{j=1}^d \in \mathbb{R}^d$$



Shallow Neural Networks

Universal approximation property:

The set

$$V_n := \text{span}\{x \in \mathbb{R}^d \mapsto \sigma(w^T x + b) \ : \ w \in \mathbb{R}^d,\, b \in \mathbb{R}\}$$

is dense in $\mathcal{C}([0,1]^d)$ if and only if σ is not a polynomial.

Some historical papers on this topic: Cybenko [Cyb89], Hornik [Hor91], Pinkus et al. [LLPS93].

Width vs. depth and the ability to create oscillations

With the universal approximation property, we have addressed the question:

"Can I approximate any continuous function by a neural network?"

And the theorem's answer is

YES provided that the activation functions are not polynomials.

But how much complexity (how many terms) do we need to approximate a function to a given accuracy ε? Can the complexity be decreased by leveraging depth?

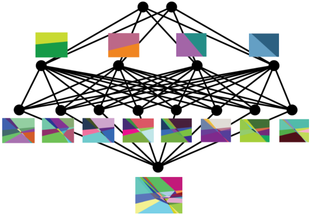
Deep Neural Networks

$$\begin{aligned} f : \mathbb{R}^d &\to \mathbb{R} \\ x &\mapsto f(x) := T_L \circ \sigma \circ T_{L-1} \circ \sigma \circ \cdots \circ T_1 \circ \sigma \circ T_0(x) \end{aligned}$$

with $T_\ell : \mathbb{R}^{m_\ell} \to \mathbb{R}^{m_{\ell+1}}$ an affine map

$$T_\ell(x) = A_\ell x + b_\ell, \quad 0 \le \ell \le L$$

and $(m_1, \dots, m_L) \in \mathbb{N}^L$ with $m_0 = d$, $m_{L+1} = 1$ (or $m_{L+1} = \tilde{d}$).



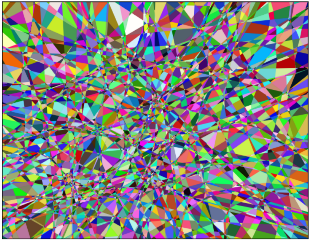


Figure: Evolution of linear regions in a DNN with a 2d input (from [HR19])

For a ReLU or REPU(p) activation function σ, the number of piecewise domains grows exponentially with the depth L. (free-knot spline)

Width vs. depth and the ability to create oscillations

Frensen, Sasao and Butler prove the following theorem [FSB10]:

Let $f \in \mathcal{C}^3([a, b])$ and set

$$\kappa := \frac{1}{4} \int_a^b \sqrt{|f''(x)|} \mathrm{d}x$$

Let $\mathrm{CPA}_\varepsilon(f)$ be

$$\mathrm{CPA}_\varepsilon(f) := \{g : [a, b] \to \mathbb{R} \text{ cont. and piecewise affine s.t. } \|f - g\|_\infty \leq \varepsilon\}$$

The smallest number of segments for a $g \in \mathrm{CPA}_\varepsilon(f)$ scales as

$$s(\varepsilon) \sim \frac{\kappa}{\sqrt{\varepsilon}} \quad \text{as } \varepsilon \to 0.$$

Complexity of a NN to create that many oscillations?

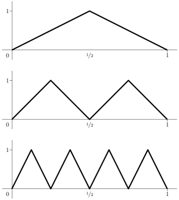
Width vs. depth and the ability to create oscillations

Composition creates exponentially many oscillations, addition only linearly many.

Consider the sawtooth function

$$f(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1/2 \\ 2-2x & \text{if } 1 \leq x \leq 1 \end{cases}$$

$$= 2\mathrm{ReLU}(x) - 4\mathrm{ReLU}(x-1/2)$$



The m-fold composition $f_m = f \circ \cdots \circ f$ is a function in $[0,1]$ with $2^m - 1$ oscillations.

A function with m scaled and translated copies has only m oscillations.

Therefore

$$s(\varepsilon) \sim \frac{\kappa}{\sqrt{\varepsilon}} \quad \Rightarrow \quad m \sim \begin{cases} \ln(\kappa\varepsilon^{-1/2}) & \text{folds} \\ \kappa\varepsilon^{-1/2} & \text{shallow (translated copies)} \end{cases}$$

Parametric manifolds V_n based on neural networks

Let $\Phi_{L,m}$ be the class of neural networks with depth L and widths $m = (m_1, \dots, m_L)$.

We define

$$V_n := \{v \in \Phi_{L,m} \ : \ L \in \mathcal{L},\ m \in \mathcal{M}_L,\ \mathrm{compl}(v) \leq n\}$$

where $\mathrm{compl}(v)$ is a complexity measure, $\mathcal{L} \subset \mathbb{N}$ is the set of possible depths and $\mathcal{M}_L \subset \mathbb{N}^L$ the set of possible widths.

Two typical classes of architectures:

- Fixed depth L and free width:

$$\mathcal{L} = \{L\}, \quad \mathcal{M}_L = \{(W, \dots, W) \ : \ W \in \mathbb{N}\}$$

- Free depth and fixed width W:

$$\mathcal{L} = \mathbb{N}, \quad \mathcal{M}_L = \{(W, \dots, W)\}$$

Complexity $\mathrm{compl}(v)$

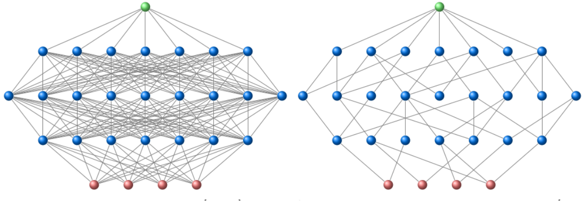
For a function v in the class $\Phi_{L,m}$ of neural networks with depth L and widths $m = (m_1, \dots, m_L)$, different measures of complexity:

- number of parameters (fully connected networks):

$$\mathrm{compl}_F(v) = \sum_{\ell=0}^{L} m_\ell m_{\ell+1} + m_{\ell+1} \sim W^2 L \text{ for } m_\ell \sim W$$

- number of non-zeros parameters (sparsely connected networks)

$$\mathrm{compl}_S(v) = \sum_{\ell=0}^{L} \|A_\ell\|_0 + \|b_\ell\|_0$$



Structured sparsity can be imposed (convolutional NN, recurrent NN...) or sparsity pattern can be considered as a free parameter (algorithmic challenge).

Link to best approximation and training

Suppose $V = L^2(\Omega)$ is our ambient space, and suppose we take

$$V_n = \Phi_{L,m}, \quad L \text{ fixed}, \; m = (W, \dots, W)$$

as our approximation set.

For a given target function $u \in V$, the error of best approximation is

$$\begin{aligned} e_n(u) &= \inf_{v \in \Phi_{L,m}} \|u - v\|_{L^2(\Omega)} \\ &= \inf_{(A_0,b_0),\dots,(A_L,b_L)} \int_\Omega |u(x) - T_L \circ \sigma T_{L-1} \circ \cdots \circ \sigma \circ T_0(x)|^2 \mathrm{d}x \end{aligned}$$

quantifies the best we can expect from V_n.

In practice, we throw N random points $x_i \in \Omega$, and we optimize

$$e_n(u) \approx \min_{(A_0,b_0),\dots,(A_L,b_L)} \frac{1}{N} \sum_{i=1}^{N} |u(x_i) - T_L \circ \sigma T_{L-1} \circ \cdots \circ \sigma \circ T_0(x_i)|^2$$

⇒ Questions on optimization and generalization (not covered here).

Deep neural networks approximation theory

Many recent results on the expressivity of deep neural networks:

- Approximation classes of DNN (free depth and fixed width) are larger than those of shallow NN (fixed depth and free width) [DDF+22].
- Emulation: DNN are as expressive as many classical approximation families (polynomials, free-knot splines...).
- They achieve (near to) optimal performance for functions from classical smoothness classes (isotropic and anisotropic Sobolev, Besov, analytic functions...).

  Example: For functions $u \in \mathcal{K} = W^{k,\infty}((0,1)^d)$, ReLU networks achieve

  $$\delta_n^{\mathrm{cont},\gamma}(\mathcal{K})_{L^\infty} := \inf_{\mathrm{E},\mathrm{D}} \sup_{u\in\mathcal{K}} \|u - \mathrm{D}(\mathrm{E}(u))\|,$$

  with continuous parameter selection.
- DNN approximate efficiently functions beyond classical smoothness classes (Takagi functions, discontinuous functions, fractals...).

Deep neural networks approximation theory

A few surprises:

- In [LSYZ21], it was proven that for functions in $\mathcal{K} = W^{k,\infty}((0,1)^d)$, ReLU networks with free depth can achieve

  $$e_n(u)_{L^\infty} \lesssim n^{-p} \quad \text{for arbitrary } p \leq 2k/d.$$

  However, we said that

  $$\delta_n^{\text{cont},\gamma}(\mathcal{K}) \sim n^{-k/d}.$$

  Therefore a rate $p > k/d$ can only be achieved with discontinuous parameter selection. Also, it requires more than $\mathcal{O}(\log_2(\varepsilon^{-1}))$ dofs to achieve accuracy ε.

- **Theory-to-practice gap:** No matter how high the theoretically possible approximation rate may be to approximate a given function with a DNN, one requires in practice an exponential quantity of samples. [GV21]

- **Personal interpretation:** We need to study DNN approximation for more specific model classes classes, and derive more closed forms to overcome the curse of dimension, and truly justify the use of NN from the approximation point of view.

Conclusions

Linear approximation:

- V_n is linear
- Kolmogorov n-width $d_n(\mathcal{K})$: measures optimal linear approx.
- Parametric PDEs: $d_n(\mathcal{K})$ decays exponentially fast for elliptic/parabolic problems.

Nonlinear approximation:

- V_n is nonlinear. Usually generated by E, D.
- Manifold n-width $\delta_n(\mathcal{K})$: measures optimal nonlinear approx.
- $\delta_n^{\text{cont}}(\mathcal{K}) \sim n^{-k/d}$ for classical regularity sets. Curse of dimension is unavoidable.
- Neural Networks: emulation, oscillations, efficient beyond classical smoothness classes but no complete picture yet.

References I

A. Cohen, R. Devore, G. Petrova, and P. Wojtaszczyk, *Optimal stable nonlinear approximation*, Foundations of Computational Mathematics (2021), 1–42.

G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Mathematics of control, signals and systems **2** (1989), no. 4, 303–314.

Ingrid Daubechies, Ronald DeVore, Simon Foucart, Boris Hanin, and Guergana Petrova, *Nonlinear approximation and (deep) relu networks*, Constructive Approximation **55** (2022), no. 1, 127–172.

R. A. DeVore, R. Howard, and C. Micchelli, *Optimal nonlinear approximation*, Manuscripta mathematica **63** (1989), 469–478.

C. L Frenzen, T. Sasao, and J. T. Butler, *On the number of segments needed in a piecewise linear approximation*, Journal of Computational and Applied mathematics **234** (2010), no. 2, 437–446.

P. Grohs and F. Voigtlaender, *Proof of the theory-to-practice gap in deep learning via sampling complexity bounds for neural network approximation spaces*, arXiv preprint arXiv:2104.02746 (2021).

References II

K. Hornik, *Approximation capabilities of multilayer feedforward networks*, Neural networks **4** (1991), no. 2, 251–257.

B. Hanin and D. Rolnick, *Complexity of linear regions in deep networks*, International Conference on Machine Learning, PMLR, 2019, pp. 2596–2604.

M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, *Multilayer feedforward networks with a nonpolynomial activation function can approximate any function*, Neural networks **6** (1993), no. 6, 861–867.

J. Lu, Z. Shen, H. Yang, and S. Zhang, *Deep network approximation for smooth functions*, SIAM Journal on Mathematical Analysis **53** (2021), no. 5, 5465–5506.