Fundamental Limits of Deep Neural Network Learning

Helmut Bőlcskei

Chair for Mathematical Information Science

Dept. ITET & Dept. Math.

ETH zürich

April 2019

Many thanks to D. Perekrestenko

joint work with P. Grohs, R. Gül, D. Elbrächter, G. Kutyniok, D. Perekrestenko, P. Petersen, and V. Vlačić

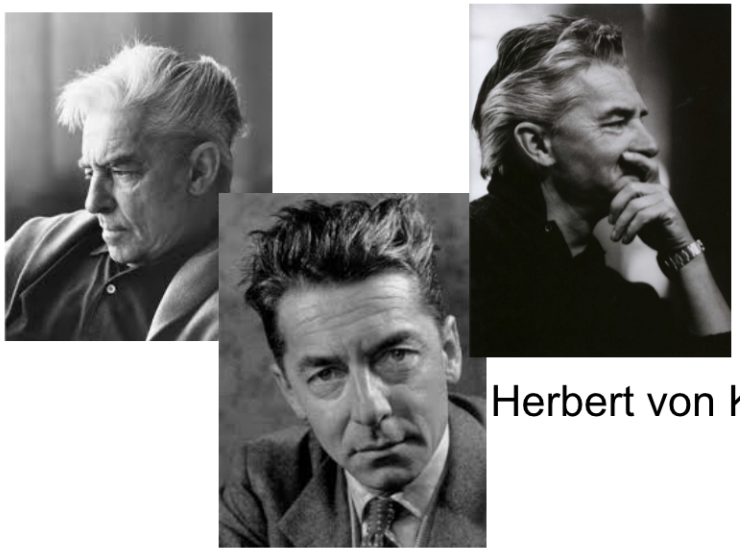
Disclaimer

This document contains images obtained by routine Google searches. Some of these images may be subjected to copyright. They are included here for educational noncommercial purposes and are considered to be covered by the doctrine of **Fair Use**.

It is not feasible to give full scholarly credit to the creators of these images. We hope that they can be satisfied with the positive role they are playing in the educational process.

Classification

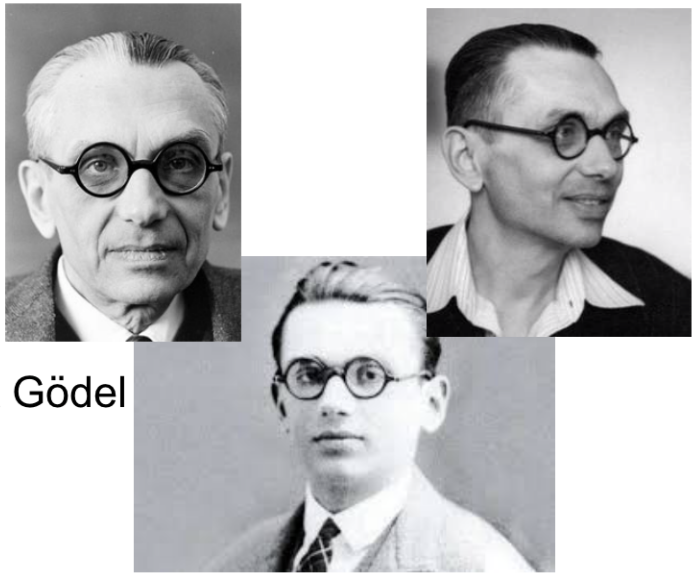
Classification



Herbert von Karajan

John von Neumann

Pyotr I. Tchaikovsky



Kurt Gödel

Describing the content of an image

CNNs generate sentences describing the content of an image [*Vinyals et al., 2015*]

Describing the content of an image

CNNs generate sentences describing the content of an image [*Vinyals et al., 2015*]

"Carlos *."*

Describing the content of an image

CNNs generate sentences describing the content of an image [*Vinyals et al., 2015*]

"Carlos Kleiber *."*

Describing the content of an image

CNNs generate sentences describing the content of an image [*Vinyals et al., 2015*]

"Carlos Kleiber conducting the .”

Describing the content of an image

CNNs generate sentences describing the content of an image [*Vinyals et al., 2015*]

"Carlos Kleiber conducting the Vienna Philharmonic's *."*

Describing the content of an image

CNNs generate sentences describing the content of an image [*Vinyals et al., 2015*]

"Carlos Kleiber conducting the Vienna Philharmonic's New Year's Concert ."

Describing the content of an image

CNNs generate sentences describing the content of an image [*Vinyals et al., 2015*]

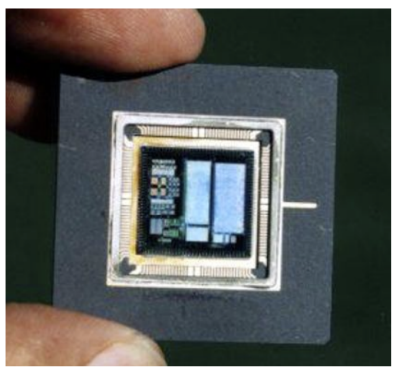
"Carlos Kleiber conducting the Vienna Philharmonic's New Year's Concert 1989."

Man vs. Machine 0:1

IBM's Deep Blue vs. Garry Kasparov, 1997

Computational power
480 chess chips

Brute force search: 200M positions per second



Slide credit: T. Hofmann, D-INFK, ETH Zurich

Go!

CNNs beat Go-champion Lee Se-dol [Silver et al., 2016]

Man vs. Machine 0:2

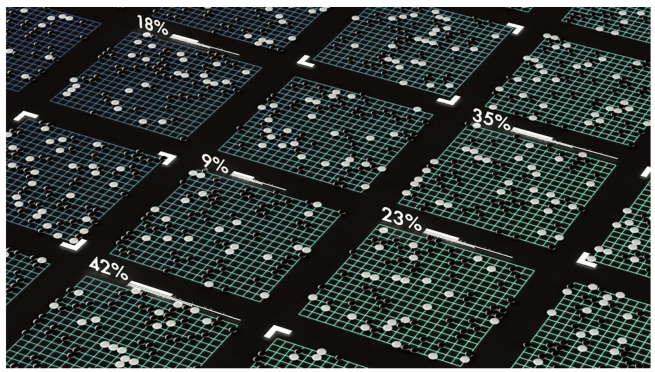
Lee Se-dol vs. AlphaGo, March 2017

Ke Jie vs. AlphaGo, May 2017

Power consumption: approx. 1MW or more than 50'000 times more than human brain

Game of Go not amenable to brute force search (10^{170} pos.)

Pattern recognition to evaluate positions + **self-play** for training



Slide credit: T. Hofmann, D-INFK, ETH Zurich

Feature extraction and classification

input: $f =$

non-linear mapping

feature vector $\Phi(f)$

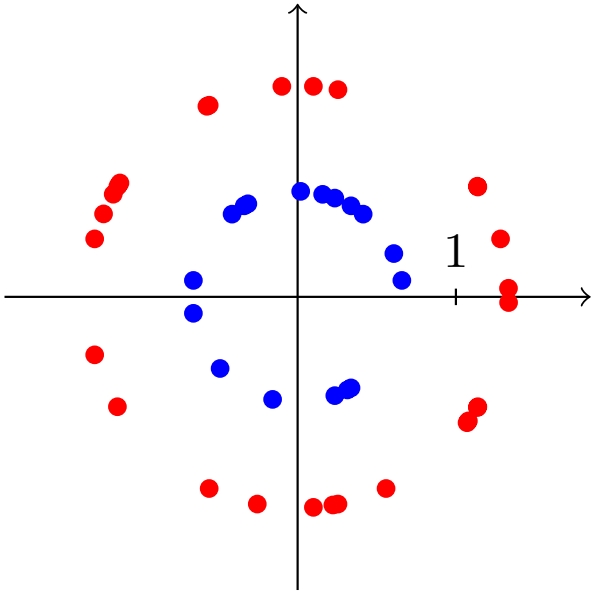
scattering networks

linear classifier

output: $\begin{cases} \langle w, \Phi(f) \rangle > 0, & \Rightarrow \text{Gödel} \\ \langle w, \Phi(f) \rangle < 0, & \Rightarrow \text{von Neumann} \end{cases}$

Why non-linear mappings?

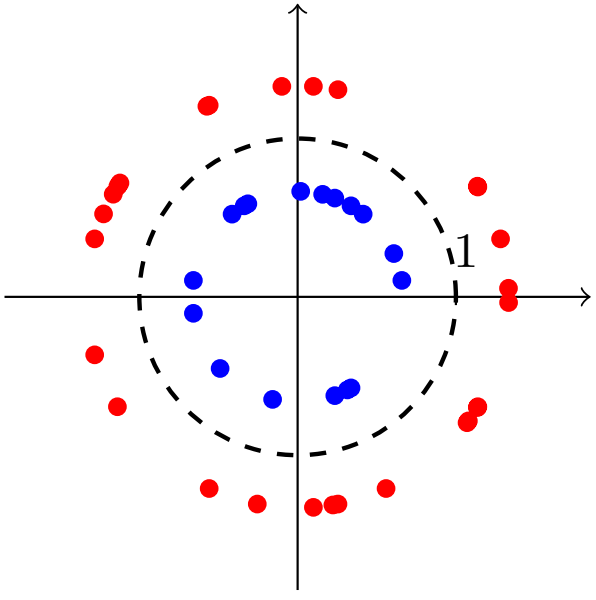
Task: Separate two categories of data through a **linear** classifier



- : $\langle w, f \rangle > 0$
- : $\langle w, f \rangle < 0$

Why non-linear mappings?

Task: Separate two categories of data through a **linear** classifier

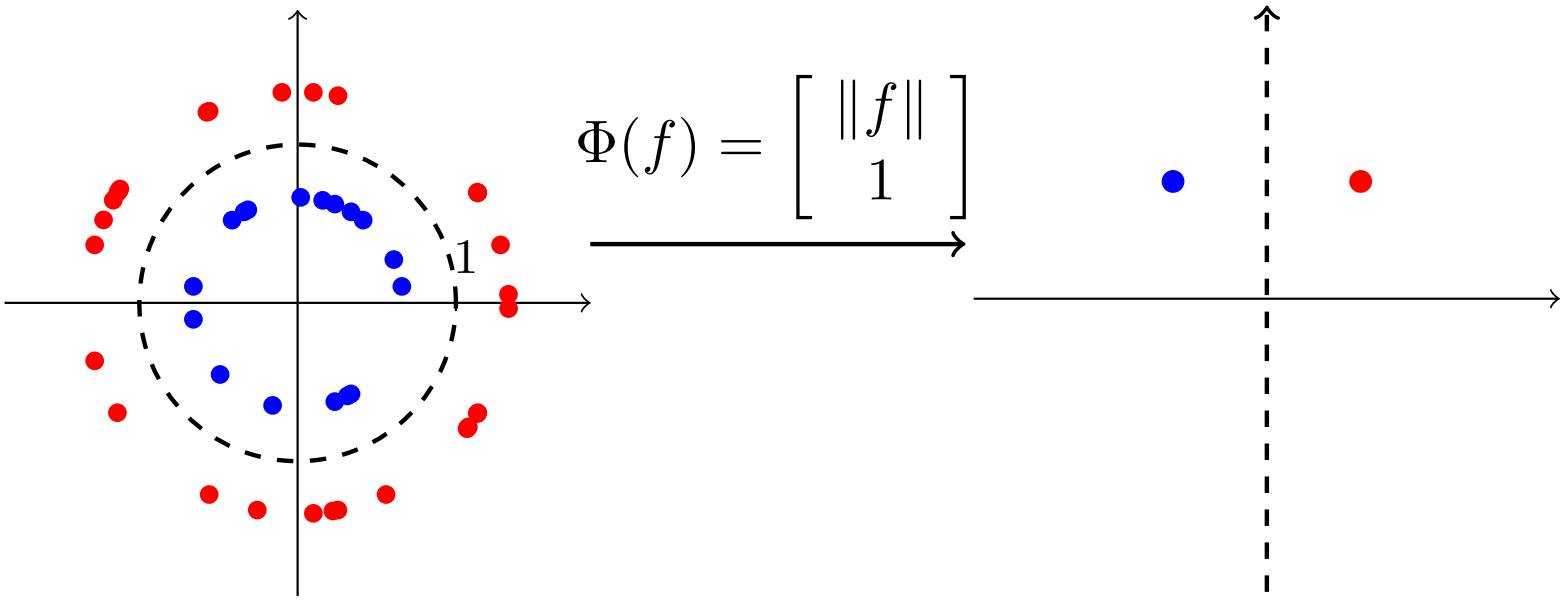


- : $\langle w, f \rangle > 0$
- : $\langle w, f \rangle < 0$

not possible!

Why non-linear mappings?

Task: Separate two categories of data through a **linear** classifier



- $\bullet$: $\langle w, f \rangle > 0$
- $\bullet$: $\langle w, f \rangle < 0$

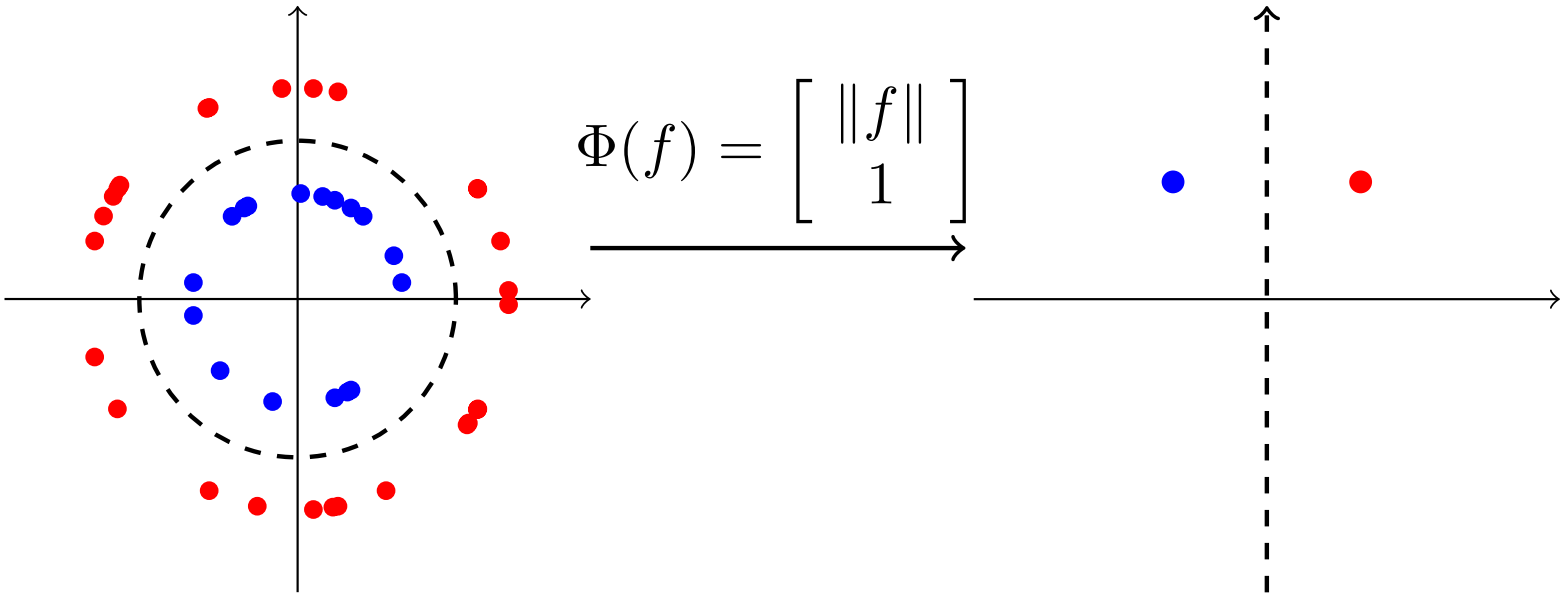
not possible!

- $\bullet$: $\langle w, \Phi(f) \rangle > 0$
- $\bullet$: $\langle w, \Phi(f) \rangle < 0$

possible with $w = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

Why non-linear mappings?

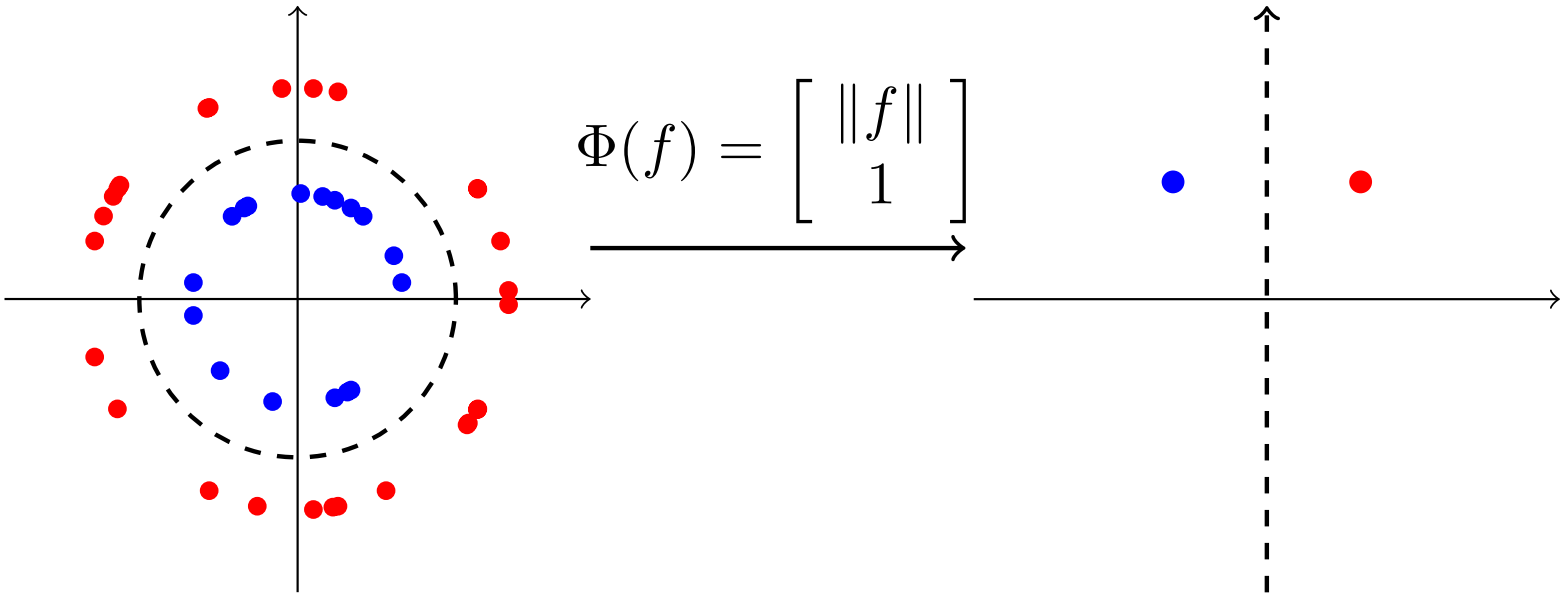
Task: Separate two categories of data through a **linear** classifier



$\Rightarrow \Phi$ is **invariant** to angular component of the data

Why non-linear mappings?

Task: Separate two categories of data through a **linear** classifier



$\Rightarrow \Phi$ is **invariant** to angular component of the data

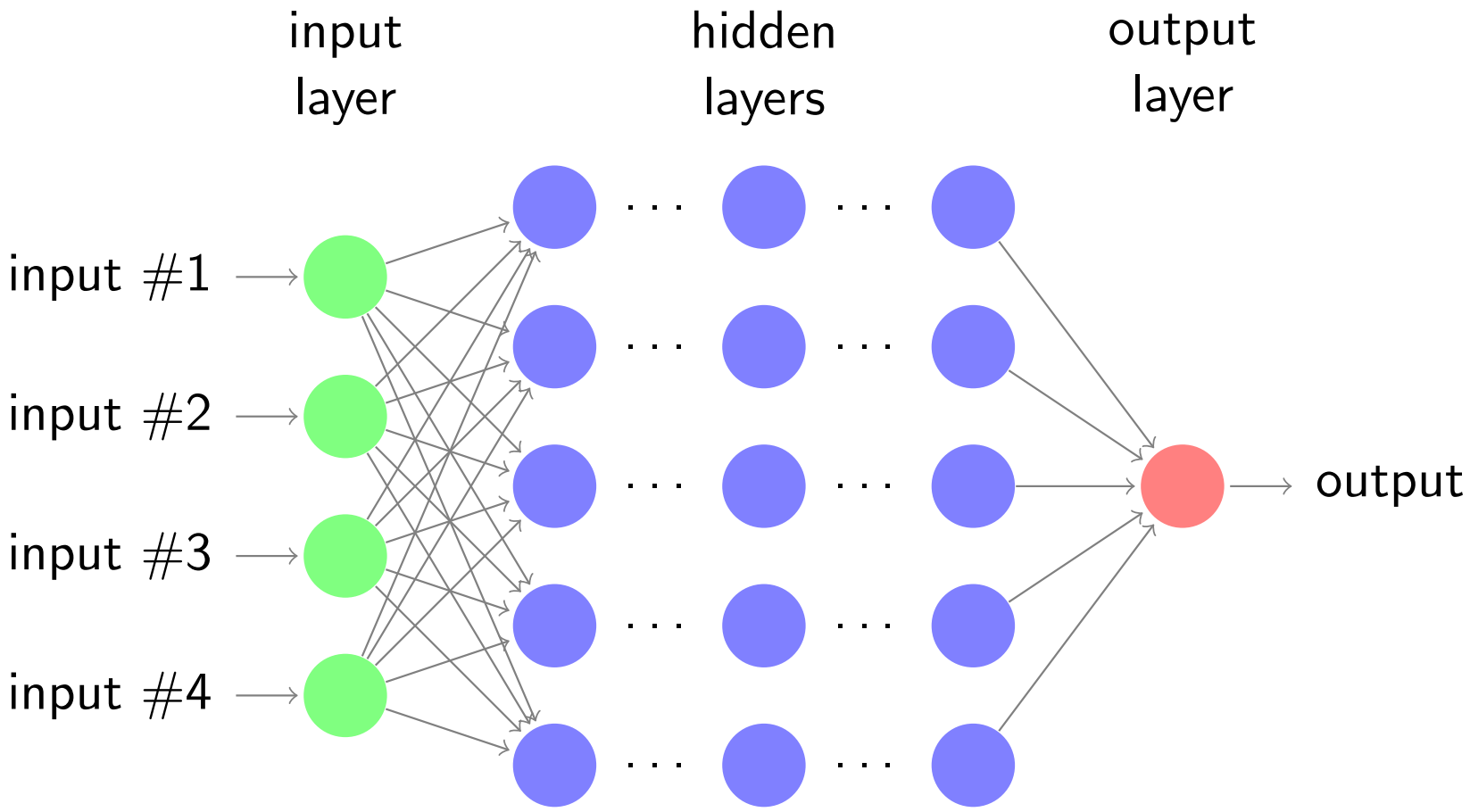
$\Rightarrow$ **Linear separability** in feature space!

Translation invariance

Handwritten digits from the MNIST database [LeCun & Cortes, 1998]

Feature vector should be invariant to spatial location
$\Rightarrow$ translation invariance

Neural networks



Concatenation of affine mappings
and non-linearities

Neural networks

Let $L, N_0, N_1, \ldots, N_L \in \mathbb{N}$, $L \geq 2$.

- **Affine maps**: $W_\ell = A_\ell x + b_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$, $\ell \in \{1, 2, \ldots, L\}$
- **Network connectivity**: $\mathcal{M}(\Phi)$ – total number of non-zero parameters in W_ℓ
- **Depth of network** or **number of layers**: $\mathcal{L}(\Phi) := L$
- **Width of network**: $\mathcal{W}(\Phi) := \max_{\ell=0,\ldots,L} N_\ell$
- **Maximum absolute value of weights in the network**: $\mathcal{B}(\Phi) := \max_{\ell=1,\ldots,L} \{\|A_\ell\|_\infty, \|b_\ell\|_\infty\}$
- **Non-linearity** or **activation function**: ρ acts component-wise

Neural networks

A map $\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ given by

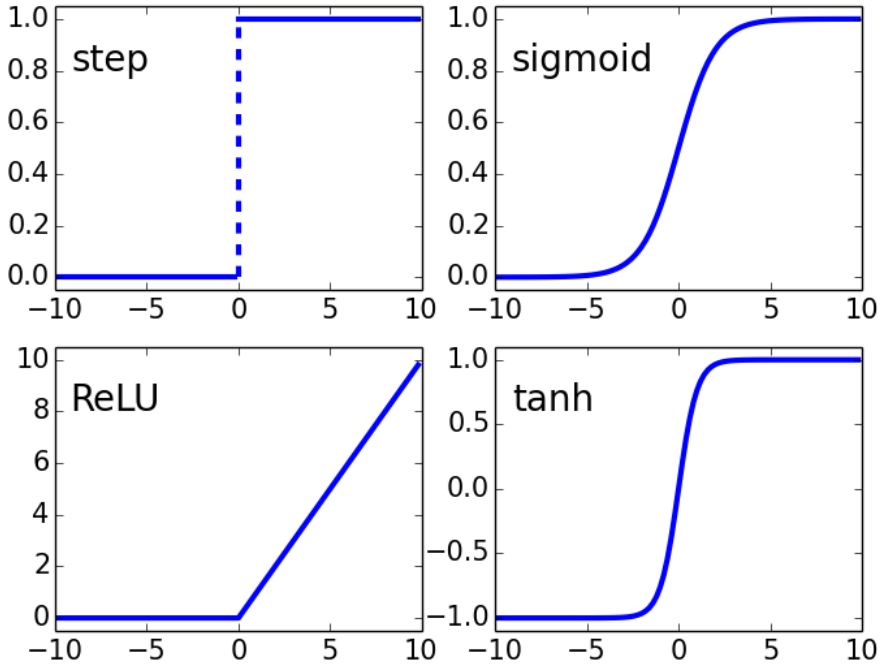
$$\Phi(x) := W_L(\rho\,(W_{L-1}(\rho\,(\ldots\rho\,(W_1(x))))))$$

is called a **neural network (NN)**.

Class of networks $\Phi : \mathbb{R}^d \to \mathbb{R}^{N_L}$ with no more than L layers, connectivity no more than M, input dimension d, output dimension N_L, and non-linearity ρ is denoted by $\mathcal{NN}^{\rho}_{L,M,d,N_L}$.

Commonly used non-linearities

- **Step-function:** Simplified model of biological neuron, hard to train
- **Sigmoid:** Smooth, easy to train, vanishing gradient problem
- **Hyperbolic tangent:** Smooth, easy to train, gradient "stronger" than for sigmoid, but still vanishing gradient problem
- **Rectified Linear Unit (ReLU):** Computationally cheap, dying ReLU problem



Course outline

- This course is about the **fundamental limits** of deep neural network learning.
- We assume an **optimal learning algorithm** and access to **infinite amounts of data**.
- Want to understand **fundamental limits** in representing functional relationships $\Phi(x)$ (learned in practice) in the form

$$\Phi(x) = W_L(\rho\,(W_{L-1}(\rho\,(\ldots\rho\,(W_1(x))))))$$

The universal approximation theorem

Theorem (Cybenko, 1989, Hornik, 1991, Pinkus, 1999*)

Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ compact, $f : \Omega \to \mathbb{R}$ continuous, $\rho : \mathbb{R} \to \mathbb{R}$ continuous and not a polynomial, and let $\varepsilon > 0$. Then, there exists a two-layer NN Φ with non-linearity ρ, such that

$$\|\Phi - f\|_\infty < \varepsilon.$$

** – no bound on width (could grow exponentially in ε)*

Theorem (Lu et al., 2017)**

For any Lebesgue-integrable function $f : \mathbb{R}^n \to \mathbb{R}$ and any $\varepsilon > 0$, there exists a ReLU network of width $n + 4$, such that

$$\int_{\mathbb{R}^n} |f(x) - \Phi(x)| < \varepsilon.$$

*** – no bound on depth (could grow exponentially in ε)*

Approximation-theoretic results for single hidden layer

- Approximation **error bounds** for smooth functions in terms of no. of neurons: [Barron, 1993,1994]
- Non-existence of **localized approximations**: [Chui et al., 1994]
- Lower **bounds on approximation rates**: [DeVore et al., 1996]
- Optimal approximation of **smooth functions**: [Mhaskar and Micchelli, 1995]

Approximation-theoretic results for multiple hidden layers

- **Universal approximation results** for general functions: [Hornik et al., 1989, Mhaskar, 1993]
- **Universal approximation results** for functions together with their derivatives: [Nguyen-Thien and Tran-Cong, 1999]
- **Deep networks can perform better than single-hidden layer networks** for certain approximation tasks: [Chui et al., 1994]
- Existence of **functions** which, albeit expressible through a **small 3-layer network**, can only be represented through **very large two-layer networks**: [Eldan and Shamir, 2016]

Approximation-theoretic results for multiple hidden layers

- Existence of **functions** that can be realized through **relatively small deep networks** but require **exponentially larger shallow networks**: [Cohen et al., 2016]
- Deep neural networks can **break the curse of dimensionality** in the approximation of the solution of certain PDEs: [Jentzen, Schwab, Grohs et al., 2018]
- **Relation** between M**-term approximation rates** of functions that are sparse in wavelet frames **and** M**-edge approximation rates**: [Shaham, Cloninger, and Coifman, 2018]

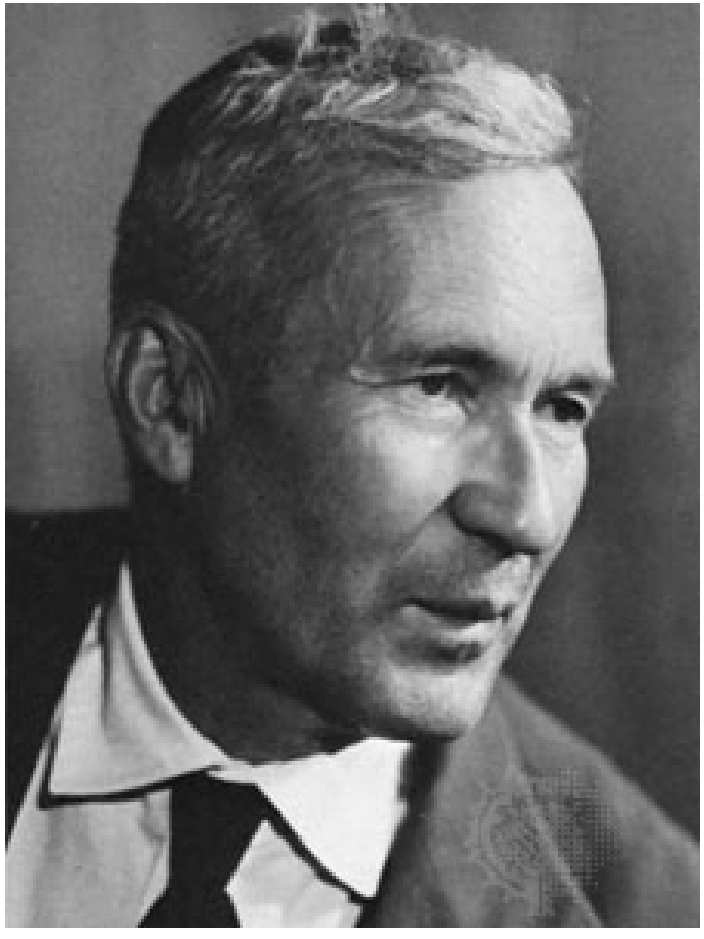
A systematic framework

- Need for a **systematic framework**
- Consider **approximation of functions from a given function class** $\mathcal{C}$, e.g., Besov spaces, modulation spaces, smooth functions, ...
- How does **"complexity" of a network** approximating all elements of $\mathcal{C}$ to within prescribed accuracy depend on **complexity of** $\mathcal{C}$?

Andrey Nikolayevich Kolmogorov, 1903-1987

Main contributions

- **Probability theory:** Established its mathematical foundation, Kolmogorov axioms, Chapman-Kolmogorov equation
- **Statistics:** Kolmogorov-Smirnov test
- **Algorithmic information theory:** Kolmogorov complexity
- **Classical mechanics:** Kolmogorov-Arnold-Moser theorem



Main results

- Deep networks provide **exponential approximation accuracy** for a wide range of functions such as the squaring operation, multiplication, polynomials, sinusoidal functions, and even one-dimensional oscillatory textures and fractal functions.

- Deep neural networks can **learn optimally vastly different function classes** such as affine systems, Gabor systems, and smooth functions.

- This universality is afforded by a **concurrent invariance property of deep networks to time-shifts, scalings, and frequency-shifts**.

Outline of the course

- Approximation of basic functions, namely x^2, polynomials, and sinusoids
- Approximation of function classes
- Quantifying approximation quality and relation to network complexity
- Affine systems
- Gabor systems and Wilson systems
- Oscillatory textures and the Weierstrass function
- Impossibility results and the case for depth

Approximation of x^2

Proposition (Squaring)

There exists a constant $C > 0$ such that for all $\varepsilon \in (0, 1/2)$, there is a network $\Phi_\varepsilon \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying $\mathcal{L}(\Phi_\varepsilon) \leq C \log(\varepsilon^{-1})$, $\mathcal{W}(\Phi_\varepsilon) = 4$, $\mathcal{B}(\Phi_\varepsilon) \leq 4$, $\Phi_\varepsilon(0) = 0$, and

$$\|\Phi_\varepsilon(x) - x^2\|_{L^\infty([0,1])} \leq \varepsilon.$$

Approximation of x^2

Finite width

Poly-log in ε^{-1}

Proposition (Squaring)

There exists a constant $C > 0$ such that for all $\varepsilon \in (0, 1/2)$, there is a network $\Phi_\varepsilon \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying $\mathcal{L}(\Phi_\varepsilon) \leq C \log(\varepsilon^{-1})$, $\mathcal{W}(\Phi_\varepsilon) = 4$, $\mathcal{B}(\Phi_\varepsilon) \leq 4$, $\Phi_\varepsilon(0) = 0$, and

$$\|\Phi_\varepsilon(x) - x^2\|_{L^\infty([0,1])} \leq \varepsilon.$$

Approximation of x^2 – Proof

Consider the function $g : [0,1] \to [0,1]$,

$$g(x) = \begin{cases} 2x, & \text{if } x < \frac{1}{2}, \\ 2(1-x), & \text{if } x \geq \frac{1}{2}, \end{cases}$$

along with the "sawtooth" functions given by the s-fold compositions

$$g_s := \underbrace{g \circ g \circ \cdots \circ g}_{s}, \quad s \geq 2,$$

and set $g_0(x) := x, g_1(x) := g(x)$.

Approximation of x^2 – Proof

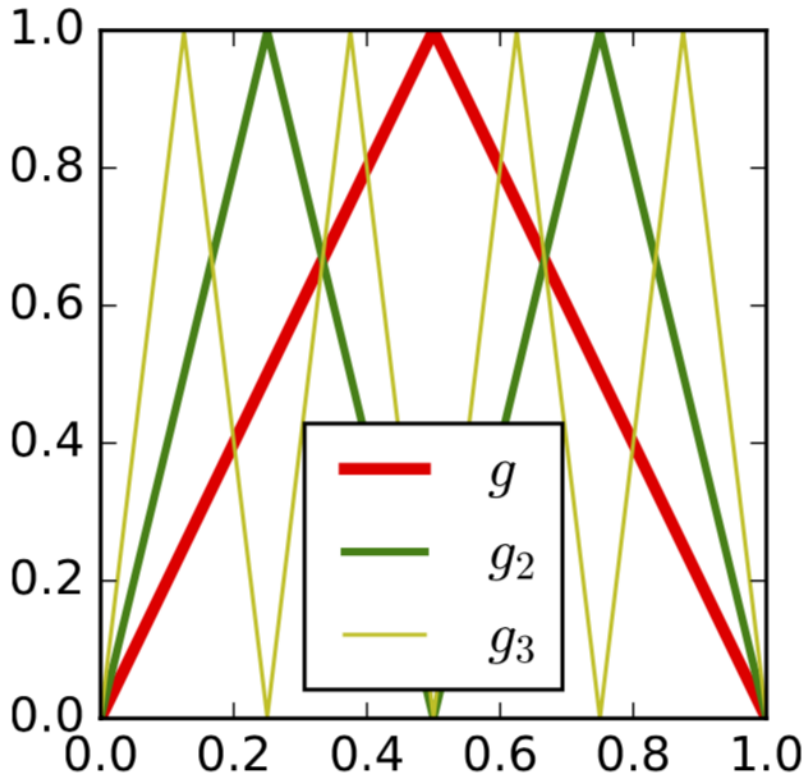
Let f_m be the piecewise linear interpolation of $f(x) = x^2$ with $2^m + 1$ uniformly spaced "knots" according to

$$f_m\left(\frac{k}{2^m}\right) = \left(\frac{k}{2^m}\right)^2, \quad k = 0, \ldots, 2^m, \quad m \in \mathbb{N}_0.$$

f_m approximates f according to

$$\|f_m(x) - x^2\|_{L^\infty[0,1]} \leq 2^{-2m-2}.$$

Approximation of x^2 – Proof



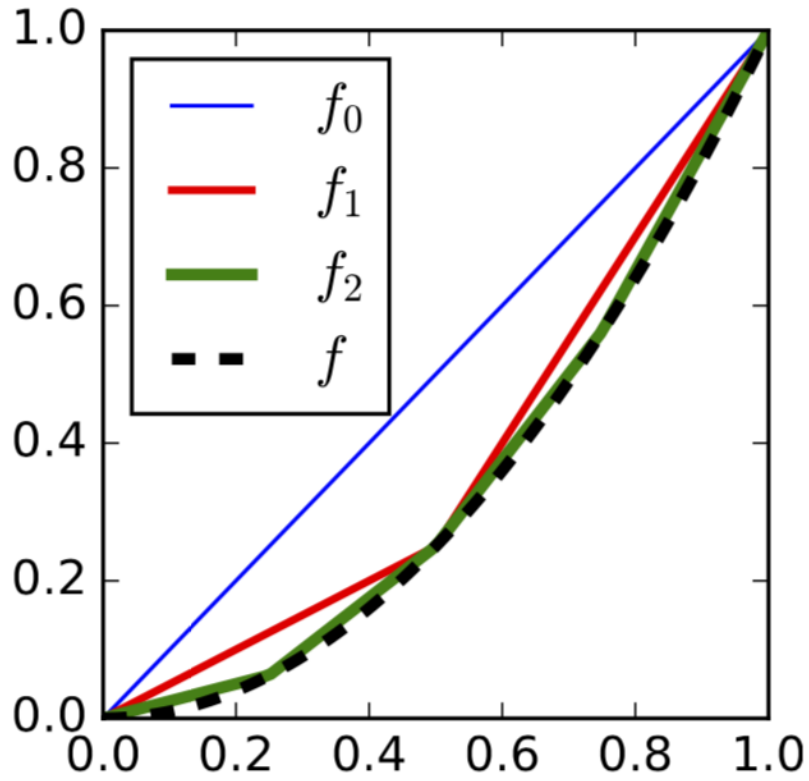


Image credit: [Yarotsky, 2017]

Refinement of interpolation

Refine interpolation by going from f_{m-1} to f_m through adjustment with a sawtooth function according to

$$f_m(x) = f_{m-1}(x) - \frac{g_m(x)}{2^{2m}}.$$

This leads to overall representation

$$f_m(x) = x - \sum_{s=1}^{m} \frac{g_s(x)}{2^{2s}}.$$

Approximation of x^2 – Proof

$$g(x) = 2\rho(x) - 4\rho(x - 1/2) + 2\rho(x - 1),$$

$$g_m(x) = g(g_{m-1}(x))$$

$$g_m = 2\rho(g_{m-1}) - 4\rho(g_{m-1} - 1/2) + 2\rho(g_{m-1} - 1),$$

and since $f_m = \rho(f_m), \forall m \in \mathbb{N}_0$, we have

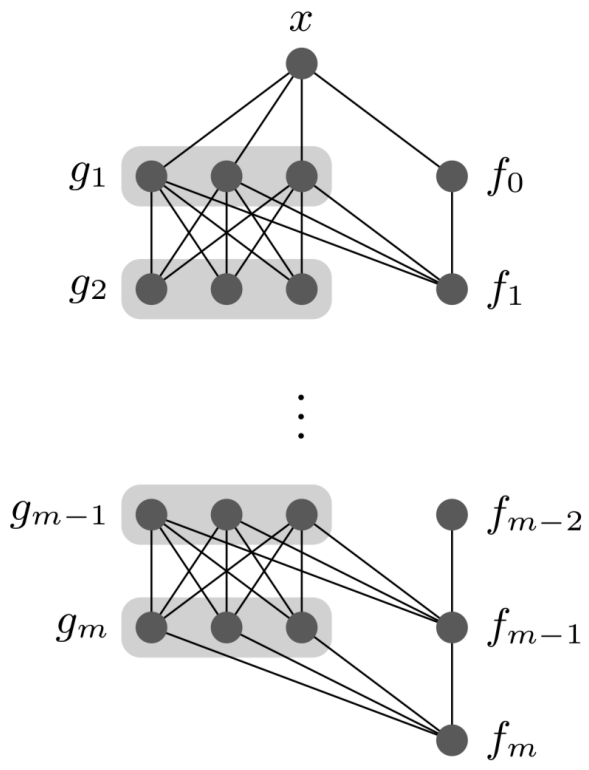


Image credit: [Schwab & Zech, 2017]

$$f_m = \rho(f_{m-1}) - 2^{-2m}\Big(2\rho(g_{m-1}) - 4\rho(g_{m-1} - 1/2) + 2\rho(g_{m-1} - 1)\Big).$$

Approximation of x^2 – Proof

Rewriting as a composition of affine linear maps and the ReLU nonlinearity

$$\begin{pmatrix} g_m \\ f_m \end{pmatrix} = W_1 \left(\rho \left(W_2 \begin{pmatrix} g_{m-1} \\ f_{m-1} \end{pmatrix} \right) \right),$$

and iterating yields

$$\begin{pmatrix} g_m \\ f_m \end{pmatrix} = W_1 \left(\rho \left(W_2 \left(\ldots \left(W_1 \left(\rho \left(W_2 \begin{pmatrix} x \\ x \end{pmatrix} \right) \right) \right) \right) \right) \right).$$

Summary: f_m realized through an $\mathcal{NN}_{m+1,13m,1,1}$ of width 4 and with $\mathcal{B}(.) \leq 4$. Statement follows from $\varepsilon_m = 2^{-2m-2}$ and hence $\log(1/\varepsilon_m) = 2m + 2$.

Relation to Yarotsky, 2017

- **No skip connections** needed.
- **Explicitly specified width**.
- Result applies to **arbitrary domains**.
- **Weights of network scale no faster than polynomial in the size of the domain**, crucial for optimality later on.

Exponential approximation accuracy

- Approximating network has **finite width** and **depth scaling poly-log** in $1/\varepsilon$.
- Owing to

$$\mathcal{M}(\Phi) \leq \mathcal{L}(\Phi)\mathcal{W}(\Phi)(\mathcal{W}(\Phi)+1),$$

  we have

$$\varepsilon \leq 2^{-(\mathcal{M}(\Phi))^{1/p}}.$$

- Finite width combined with poly-log (in $1/\varepsilon$) depth yields **exponential error decay in connectivity**.

Approximation of multiplication operation

Main idea: Write xy as

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

and realize $\frac{1}{2}\big((x+y)^2 - x^2 - y^2\big)$ as a linear combination of squaring networks.

Proposition (Multiplication operation)

There exists a constant $C > 0$ such that for all $D \in \mathbb{R}_+$ and $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,2,1}$ satisfying $\mathcal{L}(\Phi_{D,\varepsilon}) \leq C \log(\lceil D \rceil^2 \varepsilon^{-1})$, $\mathcal{W}(\Phi_{D,\varepsilon}) \leq 12$, $\mathcal{B}(\Phi_{D,\varepsilon}) \leq \max\{4, 2\lceil D \rceil^2\}$, $\Phi_{D,\varepsilon}(0, x) = \Phi_{D,\varepsilon}(x, 0) = 0$, for all $x \in \mathbb{R}$, and

$$\|\Phi_{D,\varepsilon}(x, y) - xy\|_{L^\infty([-D,D]^2)} \leq \varepsilon.$$

Approximation results for polynomials

Main idea: Realize arbitrary powers x^k through composition of squaring and multiplication networks and arbitrary polynomials by taking weighted linear combinations of powers of x.

Proposition (Polynomial approximation)

There exists a constant $C > 0$ such that for all $m \in \mathbb{N}$, $A \in \mathbb{R}_+$, $p_m(x) = \sum_{i=0}^m a_i x^i$ with $\max_{i=0,\dots,m} |a_i| = A$, $D \in \mathbb{R}_+$, and $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{p_m,D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying $\mathcal{L}(\Phi_{p_m,D,\varepsilon}) \leq Cm(\log(\lceil A \rceil) + \log(\varepsilon^{-1}) + m\log(\lceil D \rceil) + \log(m))$, $\mathcal{W}(\Phi_{p_m,D,\varepsilon}) \leq 16$, $\mathcal{B}(\Phi_{p_m,D,\varepsilon}) \leq \max\{A, 8\lceil D \rceil^{2m-2}\}$, and

$$\|\Phi_{p_m,D,\varepsilon} - p_m\|_{L^\infty([-D,D])} \leq \varepsilon.$$

In contrast to [Yarotsky, 2017] width of this network does not scale in the degree of the polynomial.

Universal approximation

Theorem (Weierstrass approximation theorem)

Let $[a, b] \subseteq \mathbb{R}$ *and* $f \in C([a, b])$. *Then, for every* $\varepsilon > 0$, *there exists a polynomial* π *such that*

$$\|f - \pi\|_{L^\infty([a,b])} \leq \varepsilon.$$

- **Every continuous function** on a closed interval can be **approximated** to within arbitrary accuracy by a **deep ReLU network of finite width**.
- This yields a **variant of the universal approximation theorem** for finite-width deep ReLU networks, but **result is not quantitative**.

Smooth functions

Definition

For $D \in \mathbb{R}_+$, let the set $\mathcal{S}_D \subseteq C^\infty([-D, D], \mathbb{R})$ be given by

$$\mathcal{S}_D = \left\{ f \in C^\infty([-D, D], \mathbb{R}) \colon \|f^{(n)}(x)\|_{L^\infty([-D,D])} \leq n!,\ n \in \mathbb{N}_0 \right\}$$

Main idea: Use Chebyshev interpolation combined with the polynomial approximation network.

Proposition (Smooth functions)

There exist a constant $C > 0$ and a polynomial π such that for all $D \in \mathbb{R}_+$, $f \in \mathcal{S}_D$, and $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{f,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying $\mathcal{L}(\Psi_{f,\varepsilon}) \leq C\lceil D \rceil (\log(\varepsilon^{-1}))^2$, $\mathcal{W}(\Psi_{f,\varepsilon}) \leq 23$, $\mathcal{B}(\Psi_{f,\varepsilon}) \leq \max\{1/D, \lceil D \rceil\}\pi(\varepsilon^{-1})$, and

$$\|\Psi_{f,\varepsilon} - f\|_{L^\infty([-D,D])} \leq \varepsilon.$$

Approximation results for sinusoidal functions

Main idea: Taylor series approximation of one period and periodic extension through "sawtooth" function.

Theorem (Cosine approximation)

There exists a constant $C > 0$ *such that for every* $a, D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, *there is a network* $\Psi_{a,D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ *satisfying* $\mathcal{L}(\Psi_{a,D,\varepsilon}) \leq C((\log(1/\varepsilon))^2 + \log(\lceil aD \rceil))$, $\mathcal{W}(\Psi_{a,D,\varepsilon}) \leq 16$, $\mathcal{B}(\Psi_{a,D,\varepsilon}) \leq C$, *and*

$$\|\Psi_{a,D,\varepsilon} - \cos(a\,\cdot)\|_{L^\infty([-D,D])} \leq \varepsilon.$$

Approximation of $\cos(ax)$ – Proof

- Thanks to the Taylor theorem with remainder in Lagrange form and with $N_\varepsilon := \lceil 2\pi^2 e \log(2/\varepsilon) \rceil$, we have, for all $x \in [0,1]$

$$\left| \cos(2\pi x) - \sum_{n=0}^{N_\varepsilon} \frac{(-1)^n}{(2n)!} (2\pi x)^{2n} \right| \leq \frac{(2\pi)^{4N_\varepsilon+2}}{(2N_\varepsilon+1)!} \leq \frac{\varepsilon}{2}.$$

- By the polynomial approximation theorem, there hence exists a network $\Phi_{\varepsilon/2}$ such that

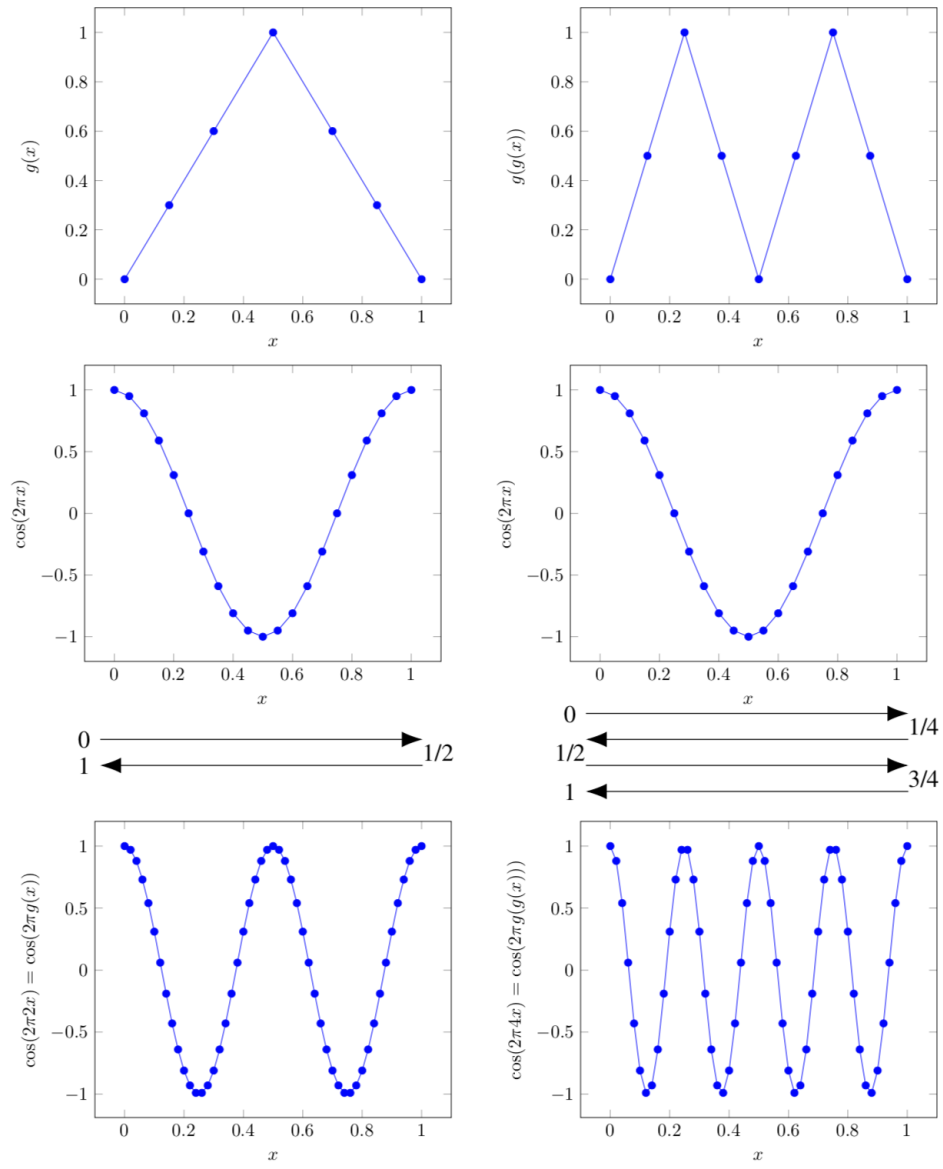
$$\|\Phi_{\varepsilon/2} - \cos(2\pi \cdot)\|_{L^\infty([0,1])} \leq \varepsilon.$$

Approx. of $\cos(2\pi a x)$ using iterated "sawtooth" functions

$x \mapsto \cos(2\pi x)$ is 1-periodic and even. Recall the "sawtooth" functions $g_s : [0,1] \to [0,1]$ and note that

$$\cos(2\pi 2^s x) = \cos(2\pi g_s(x)).$$

This "periodization trick" avoids coefficients of magnitude a^{2N_ε}, coming from Taylor polynomial for $\cos(ax)$.



Approximation of $\cos(ax)$ – Proof

For every $a > 0$, there exists a $C_a \in (1/2, 1]$ such that $a/(2\pi) = C_a 2^{\lceil \log(a) - \log(2\pi) \rceil}$. It then follows that

$$\begin{aligned}
&\Big\| \Phi_{\varepsilon/2} \Big(g_{\lceil \log(a) - \log(2\pi) \rceil}(C_a |x|) \Big) \\
&- \cos\big(2\pi g_{\lceil \log(a) - \log(2\pi) \rceil}(C_a |x|)\big) \Big\|_{L^\infty([-1,1])} \\
&= \Big\| \Phi_{\varepsilon/2} \Big(g_{\lceil \log(a) - \log(2\pi) \rceil}(C_a |x|) \Big) - \cos(ax) \Big\|_{L^\infty([-1,1])} \leq \varepsilon.
\end{aligned}$$

Extension to arbitrary domains $[-D, D]$ through scaling arguments.

Deep neural network approximation of signal classes

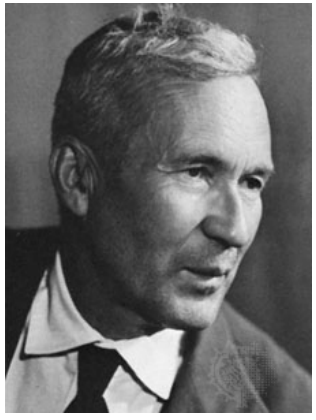
- **So far** approximation of **individual functions**.
- Now **approximation of entire function classes** $\mathcal{C}$.

Main goal

Establish a **relationship** between **complexity of** $\mathcal{C}$ and **complexity** of corresponding approximating **networks**.

- Consider networks with quantized weights.
- **Network complexity** measured in terms of **number of bits** needed to store **network topology and quantized weights**.

Asymptotic min-max rate distortion theory



A. N. Kolmogorov

D. Donoho

Definition

Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, and consider compact $\mathcal{C} \subset L^2(\Omega)$.

Encoders and decoders:

$$\mathfrak{E}^\ell := \left\{ E : \mathcal{C} \to \{0,1\}^\ell \right\} \qquad \mathfrak{D}^\ell := \left\{ D : \{0,1\}^\ell \to L^2(\Omega) \right\}$$

Optimal exponent

Definition

Minimax code length:

$$L(\varepsilon, \mathcal{C}) := \min \Big\{ \ell \in \mathbb{N} : \exists (E, D) \in \mathfrak{E}^\ell \times \mathfrak{D}^\ell : \sup_{f \in \mathcal{C}} \| D(E(f)) - f \|_{L^2(\Omega)} \leq \varepsilon \Big\}$$

Optimal exponent:

$$\gamma^*(\mathcal{C}) := \sup \Big\{ \gamma \in \mathbb{R} : L(\varepsilon, \mathcal{C}) \in \mathcal{O}\Big(\varepsilon^{-1/\gamma} \Big), \varepsilon \to 0 \Big\}$$

- $\gamma^*(\mathcal{C})$ quantifies "description complexity of function class $\mathcal{C}$
- Larger $\gamma^*(\mathcal{C}) \Rightarrow$ smaller growth rate $\Rightarrow$ smaller memory requirements for storing signals $f \in \mathcal{C}$

Kolmogorov-Tikhomirov ε-entropy

- $B_\varepsilon(f_0) = \{g : \|g - f_0\|_{L^2} \leq \varepsilon\}$ is the ball of radius ε around f_0.
- ε-net for $\mathcal{C}$ is a finite collection of points $(f_i)_{i=1}^N$ in L^2 such that

$$\mathcal{C} \subset \bigcup_{i=1}^{N} B_\varepsilon(f_i).$$

- $N(\varepsilon, \mathcal{C})$ is the minimum possible cardinality of any such ε-net.
- Kolmogorov-Tikhomirov ε-entropy of $\mathcal{C}$ is $H_\varepsilon(\mathcal{C}) = \log_2 N(\varepsilon, \mathcal{C})$.

Kolmogorov-Tikhomirov ε-entropy cont'd

- Binary encoding of ball centers yields distortion upper-bounded by ε with $L(\varepsilon, \mathcal{C}) = \lceil \log_2 N(\varepsilon, \mathcal{C}) \rceil$.
- When $\mathcal{C}$ is not a finite set, then $H_\varepsilon(\mathcal{C}) \to \infty$ as $\varepsilon \to 0$.
- In many interesting cases $H_\varepsilon(\mathcal{C}) \asymp \varepsilon^{-1/\alpha}$ or $H_\varepsilon(\mathcal{C}) \asymp \varepsilon^{-1/\alpha} \log(\varepsilon^{-1})^\beta$ for some $\alpha, \beta > 0$.
- Optimal exponent

$$\gamma^*(\mathcal{C}) := \sup \left\{ \gamma \in \mathbb{R} : L(\varepsilon, \mathcal{C}) \in \mathcal{O}\left(\varepsilon^{-1/\gamma}\right), \varepsilon \to 0 \right\}$$

  is hence a **crude measure of growth**.

Nonlinear approximation through dictionaries

Definition (Best M-term approximation rate)

Given $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, a function class $\mathcal{C} \subset L^2(\Omega)$, and a dictionary $\mathcal{D} = (\varphi_i)_{i \in I} \subset L^2(\Omega)$, the supremal $\gamma > 0$ so that

$$\sup_{f \in \mathcal{C}} \inf_{\substack{I_M \subseteq I, \\ \#I_M = M, (c_i)_{i \in I_M}}} \left\| f - \sum_{i \in I_M} c_i \varphi_i \right\|_{L^2(\Omega)} \in \mathcal{O}(M^{-\gamma}), \ M \to \infty,$$

will be denoted by $\gamma^*(\mathcal{C}, \mathcal{D})$ and referred to as *best M-term approximation rate of $\mathcal{C}$ in the dictionary $\mathcal{D}$*.

Function classes and dictionaries

- Function classes $\mathcal{C}$ typically studied in the literature include unit balls in Lebesgue, Sobolev, or Besov spaces.

- Dictionaries: Wavelets, ridgelets, curvelets, shearlets, parabolic molecules, α-molecules, Gabor frames, Wilson frames, local cosine bases, and wave atoms.

Hardness of approximation

- $\gamma^*(\mathcal{C}, \mathcal{D})$ quantifies **how well** function class $\mathcal{C}$ can be approximated in dictionary $\mathcal{D}$.
- **Larger** $\gamma^*(\mathcal{C}, \mathcal{D})$ means **better** approximation.
- For given $\mathcal{C}$, is there a **fundamental limit** on $\gamma^*(\mathcal{C}, \mathcal{D})$ when one is allowed to vary over $\mathcal{D}$?
- Every dense (and countable) $\mathcal{D}$ results in $\gamma^*(\mathcal{C}, \mathcal{D}) = \infty$.
- However, identifying and storing the optimal set of participating elements in $\mathcal{D}$ is **practically infeasible** as it requires
  - **searching an infinite set**
  - **infinitely many bits** to **store** the corresponding **indices**

Effective best M-term approximation [Donoho, 1993]

- **Restrict search** for the M elements in $\mathcal{D}$ participating in the best M-term representation to the first $\pi(M)$ elements, with π a polynomial.

- **Require** that the **coefficients** c_i in the best M-term approximation $f_M = \sum_{i \in I_M} c_i \varphi_i$ be **uniformly bounded** so that they can be quantized and stored with a finite number of bits.

Formal definition

Definition (Effective best M-term approximation rate)

Given $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, a function class $\mathcal{C} \subset L^2(\Omega)$, and a dictionary $\mathcal{D} = (\varphi_i)_{i \in I} \subset L^2(\Omega)$, the supremal $\gamma > 0$ so that there exist a polynomial π and a constant $D > 0$ such that

$$\sup_{f \in \mathcal{C}} \inf_{\substack{I_M \subset \{1,2,\dots,\pi(M)\}, \\ \#I_M = M, (c_i)_{i \in I_M}}} \left\| f - \sum_{i \in I_M} c_i \varphi_i \right\|_{L^2(\Omega)} \in \mathcal{O}(M^{-\gamma}), \ M \to \infty,$$

where $\max_{i \in I_M} |c_i| \leq D$, will be denoted $\gamma^{*,\mathrm{eff}}(\mathcal{C}, \mathcal{D})$ and referred to as *effective best M-term approximation rate of $\mathcal{C}$ in the dictionary $\mathcal{D}$*.

Optimal representability by dictionaries

Theorem (Donoho, 1993, Grohs, 2015)

Let $d \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^d$. The effective best M-term approximation rate of the function class $\mathcal{C} \subset L^2(\Omega)$ in the dictionary $\mathcal{D} \subset L^2(\Omega)$ satisfies

$$\gamma^{*,\mathrm{eff}}(\mathcal{C}, \mathcal{D}) \leq \gamma^*(\mathcal{C}).$$

Definition (Optimal representability of a function class by $\mathcal{D}$)

Let $d \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^d$. If the effective best M-term approximation rate of the function class $\mathcal{C} \subset L^2(\Omega)$ in the dictionary $\mathcal{D} \subset L^2(\Omega)$ satisfies

$$\gamma^{*,\mathrm{eff}}(\mathcal{C}, \mathcal{D}) = \gamma^*(\mathcal{C}),$$

we say that the function class $\mathcal{C}$ is **optimally representable** by $\mathcal{D}$.

Back of the envelope calculation

Polynomial-depth search and bounded coefficients lead to Kolmogorov-optimal approximation

- Need $M \log(\pi(M)) = \mathcal{O}(M \log(M))$ bits to represent indices of participating dictionary elements.
- Coefficients c_i quantized by rounding to integer multiples of $\lceil M^{-\alpha} \rceil$ for some $\alpha \Rightarrow \mathcal{O}(M \log(M))$ bits to represent quantized coefficients.

Back of the envelope cont'd

- Encoder-decoder pair reconstructing f from $\mathcal{O}(M \log(M))$ bits with $\varepsilon \approx M^{-\gamma^{*,\mathsf{eff}}(\mathcal{C},\mathcal{D})}$
- For $\mathcal{D}$ optimally representing $\mathcal{C}$, we have $\gamma^{*,\mathsf{eff}}(\mathcal{C},\mathcal{D}) = \gamma^*(\mathcal{C})$
- Resulting code length

$$M \log(M) = \varepsilon^{-1/\gamma^*(\mathcal{C})} \log(\varepsilon^{-1/\gamma^*(\mathcal{C})}) = \mathcal{O}(\varepsilon^{-1/\gamma^*(\mathcal{C})})$$

Function classes and their optimal exponents

Class	F	optimal dictionary	$\gamma^*(\mathcal{C})$
L^2-Sobolev	W_2^m	Fourier or Wavelet	m
L^p-Sobolev	W_p^m	Wavelet	m
Hölder	C^α	Wavelet	α
Bump Algebra	$B_{1,1}^1$	Wavelet	1
Bounded Variation	BV	Haar	1
Segal Algebra	S	Wilson	$1/2$
Besov	* $B_{p,q}^s$	Wavelet	s
Modulation	** M_p	Wilson	$(-1/2+1/p)^{-1}$

* $q > (s+1/2)^{-1}$

** $1 \leq p < 2$

Approximation with deep neural networks

- We develop the **new concept** of **best M-weight approximation** through deep neural networks
- **Neural network** interpreted as an **encoder** and evaluated in Kolmogorov-Donoho framework
- Need to **encode** network **topology** and **quantized weights**
- Need to **control quantization-induced error**

Best M-weight approximation

Definition (Best M-weight approximation rate)

Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, and $\mathcal{C} \subset L^2(\Omega)$ be a function class. The supremal $\gamma > 0$ so that

$$\sup_{f \in \mathcal{C}} \inf_{\Phi \in \mathcal{NN}_{\infty, M, d, 1}} \|f - \Phi\|_{L^2(\Omega)} \in \mathcal{O}(M^{-\gamma}),\ M \to \infty,$$

is referred to as *best M-weight approximation rate of $\mathcal{C}$ by neural networks* and will be denoted $\gamma^*_{\mathcal{NN}}(\mathcal{C})$.

- Infimum over all networks with no more than M weights and arbitrary depth L, in particular over **all possible network topologies**
- Best M-weight approximation rate **benchmarks all learning algorithms** that map an f and an $\varepsilon > 0$ to a neural network

Effective best M-weight approximation rate

- For dictionaries, polynomial depth search constraint made to allow encoding of indices with $\mathcal{O}(M \log(M))$ bits.
- Tree-like structure of network automatically guarantees that nonzero network weight positions can be encoded with $\mathcal{O}(M \log(M))$ bits.
- Total number of weights in network can not exceed $\mathcal{L}(\Phi)\mathcal{W}(\Phi)(\mathcal{W}(\Phi)+1) \Rightarrow \mathcal{O}(M^3)$ possibilities for locations of M nonzero weights.
- Encoding of locations of M nonzero weights requires $\log(\binom{M^3}{M}) = \mathcal{O}(M \log(M))$ bits.

Effective best M-weight approximation rate

- Network layout, i.e., L and the N_k can be encoded with $\mathcal{O}(M \log(M))$ bits.
- Inspired by results in first part of the course, we impose $\mathcal{L}(\Phi) = \mathcal{O}(\log(\varepsilon^{-1}))$.
- Since we are interested in approximation error decaying as $M^{-\gamma}$, this suggests to have $\mathcal{L}(\Phi)$ grow poly-logarithmically in $\log(M)$.

Weight growth conditions

- In dictionary approximation weights were assumed uniformly bounded.
- More generous condition in neural network approximation, will allow weights to grow polynomially in M.
- Can convert networks with weight growth polynomial in M to networks with bounded weights at expense of depth increase, but depth scaling remains poly-log in M.

In summary, we have the concept of best M-weight approximation subject to polylogarithmic depth and polynomial weight growth.

Effective best M-weight approximation rate

Definition (Effective best M-weight approximation rate)

Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, and $\mathcal{C} \subset L^2(\Omega)$ be a function class. The supremal $\gamma > 0$ so that there is a polynomial π such that

$$\sup_{f \in \mathcal{C}} \inf_{\Phi \in \mathcal{NN}^{\pi(M)}_{\pi(\log(M)),M,d,1}} \|f - \Phi\|_{L^2(\Omega)} \in \mathcal{O}(M^{-\gamma}),\ M \to \infty,$$

is referred to as *effective best M-weight approximation rate of $\mathcal{C}$ by neural networks* and will be denoted by $\gamma^{*,\mathsf{eff}}_{\mathcal{NN}}(\mathcal{C})$.

Optimal representability by neural networks

Theorem

Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ be bounded, and $\mathcal{C} \subset L^2(\Omega)$. Then, we have

$$\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) \leq \gamma^*(\mathcal{C}).$$

Definition (Optimal representability of a function class by NNs)

For $d \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^d$ bounded, we say that the function class $\mathcal{C} \subset L^2(\Omega)$ is **optimally representable by neural networks** if

$$\gamma^{*,\mathrm{eff}}_{\mathcal{NN}}(\mathcal{C}) = \gamma^*(\mathcal{C}).$$

Quantization of weights

Definition

Let $m \in \mathbb{N}$ and $\varepsilon \in (0, 1/2)$. The network Φ is said to have (m, ε)-quantized weights if all its weights are elements of $2^{-m\lceil \log(\varepsilon^{-1})\rceil}\mathbb{Z} \cap [-\varepsilon^{-m}, \varepsilon^{-m}]$.

Lemma

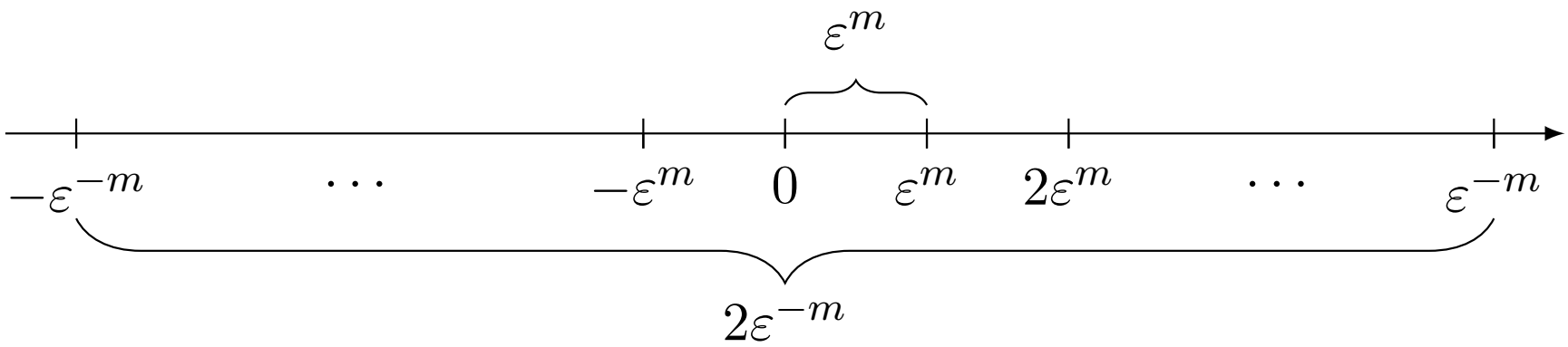
Let $B, k \in \mathbb{N}$, $\Omega \subseteq [-B, B]^d$, $\varepsilon \in (0, 1/2)$, and $M \leq \varepsilon^{-k}$. Further, let $\Phi \in \mathcal{NN}_{L,M,d,d'}$ with $\mathcal{B}(\Phi) \leq \varepsilon^{-k}$ and let $m \in \mathbb{N}$ be such that

$$m \geq 3kL + \log(\max\{1, B\}).$$

Then, there exists a network $\tilde{\Phi} \in \mathcal{NN}_{L,M,d,d'}$ with (m, ε)-quantized weights satisfying

$$\sup_{x \in \Omega} \|\Phi(x) - \tilde{\Phi}(x)\|_\infty \leq \varepsilon.$$

Quantization back of the envelope



- # of pieces of size ε^m: $\frac{2\varepsilon^{-m}}{\varepsilon^m} = 2\varepsilon^{-2m}$
- # of possible weight values: $2\varepsilon^{-2m} + 1$
- # of bits needed to encode weight:
  $\log(2\,\varepsilon^{-2m} + 1) \leq \log(3\,\varepsilon^{-2m}) \leq 2m \log(\varepsilon^{-1}) + 2$

Summary: # of bits needed to encode (m, ε)-quantized weight is $\mathcal{O}(m \log(1/\varepsilon))$.

Minimum connectivity growth rate

Proposition

Let $\Omega \subset \mathbb{R}^d$, $\mathcal{C} \subset L^2(\Omega)$, and let π be a polynomial. Further, let

$$\Psi : \left(0, \frac{1}{2}\right) \times \mathcal{C} \to \mathcal{NN}_{\infty,\infty,d,d'}$$

be a map such that for every $f \in \mathcal{C}$, $\varepsilon \in (0, 1/2)$, the network $\Psi(\varepsilon, f)$ has $(\lceil \pi(\log(\varepsilon^{-1})) \rceil, \varepsilon)$-quantized weights and satisfies

$$\sup_{f \in \mathcal{C}} \|f - \Psi(\varepsilon, f)\|_{L^2(\Omega)} \leq \varepsilon.$$

Then,

$$\sup_{f \in \mathcal{C}} \mathcal{M}(\Psi(\varepsilon, f)) \notin \mathcal{O}\left(\varepsilon^{-1/\gamma}\right), \ \varepsilon \to 0, \quad \text{for all } \gamma > \gamma^*(\mathcal{C}).$$

Connectivity growth rate

- Connectivity growth rate of networks achieving uniform approximation error ε must exceed $\left(\varepsilon^{-1/\gamma^*(\mathcal{C})}\right)$, $\varepsilon \to 0$.
- For connectivity growing according to $\left(\varepsilon^{-1/\gamma}\right)$, $\varepsilon \to 0$ with $\gamma > \gamma^*(\mathcal{C})$, we have a strong converse.

Proposition

Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ be bounded, π a polynomial, and $\mathcal{C} \subset L^2(\Omega)$. Then, for all $C > 0$ and $\gamma > \gamma^*(\mathcal{C})$, we have

$$\sup_{f \in \mathcal{C}} \inf_{\Phi \in \mathcal{NN}^{\pi(M)}_{\pi(\log(M)),M,d,1}} \|f - \Phi\|_{L^2(\Omega)} \geq C M^{-\gamma},$$

for infinitely many $M \in \mathbb{N}$.

Function classes and their optimal exponents

Class	F	optimal dictionary	$\gamma^*(\mathcal{C})$
L^2-Sobolev	W_2^m	Fourier or Wavelet	m
L^p-Sobolev	W_p^m	Wavelet	m
Hölder	C^α	Wavelet	α
Bump Algebra	$B_{1,1}^1$	Wavelet	1
Bounded Variation	BV	Haar	1
Segal Algebra	S	Wilson	$1/2$
Besov	* $B_{p,q}^s$	Wavelet	s
Modulation	** M_p	Wilson	$(-1/2+1/p)^{-1}$

* $q > (s+1/2)^{-1}$

** $1 \leq p < 2$

Transitioning from dictionaries to neural networks

- We build a theory for **transferring optimal approximation results** for dictionaries to optimal approximation results for neural networks.
- For given $\mathcal{C}$ and associated $\mathcal{D}$, we establish conditions guaranteeing the existence of a neural network with connectivity $O(M)$ that achieves the same uniform error over $\mathcal{C}$ as best M-term approximation.
- Leads to a **characterization of function classes** $\mathcal{C}$ that are **optimally representable by neural networks.**

Effective representability of dictionaries by neural networks

Definition

Let $d \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$, and $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$ be a dictionary. Then, $\mathcal{D}$ is said to be **effectively representable by neural networks**, if there exists a bivariate polynomial π such that for all $i \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$, there is a neural network $\Phi_{i,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying $\mathcal{M}(\Phi_{i,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(i))$, $\mathcal{B}(\Phi_{i,\varepsilon}) \leq \pi(\varepsilon^{-1}, i)$, and

$$\|\varphi_i - \Phi_{i,\varepsilon}\|_{L^2(\Omega)} \leq \varepsilon.$$

Optimality transfer

Central result

Optimality of a representation system $\mathcal{D}$ for a signal class $\mathcal{C}$ combined with effective representability of $\mathcal{D}$ by neural networks implies optimal representability of $\mathcal{C}$ by neural networks.

Optimal dictionaries

Affine dictionaries (e.g. wavelets, ridgelets, curvelets, shearlets, α-shearlets) are optimally representable by neural networks.

Affine dictionaries

Definition (Affine dictionary)

Consider the compactly supported functions

$$g_s := \sum_{k=1}^{r} c_k^s f(\cdot - b_k), \quad s = 0, \dots, S.$$

We define the **affine dictionary** $\mathcal{D} \subset L^2(\Omega)$ corresponding to $(g_s)_{s=0}^S$ according to

$$\mathcal{D} := \Big\{ g_s^{j,e} := \Big(|\det(A_j)|^{\frac{1}{2}} g_s(A_j \cdot - \delta e) \Big) \big|_\Omega : s \in \{1, 2, \dots, S\}, e \in \mathbb{Z}^d, \\ j \in \mathbb{N}, \text{ and } g_s^{j,e} \neq 0 \Big\} \cup \{ g_0^e := g_0(\cdot - \delta e)\big|_\Omega : e \in \mathbb{Z}^d \text{ and } g_0^e \neq 0 \},$$

and refer to f as the **generator (function) of $\mathcal{D}$.**

Includes wavelets, ridgelets, curvelets, shearlets, α-shearlets, and more generally α-molecules.

Invariance to Affine Transformations

Proposition

Let $f \in L^p(\mathbb{R}^d)$. Assume that there exists a bivariate polynomial π_1 such that for all $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\|f - \Phi_{D,\varepsilon}\|_{L^p([-D,D]^d)} \leq \varepsilon,$$

with $\mathcal{M}(\Phi_{D,\varepsilon}) \leq \pi_1(\log(\varepsilon^{-1}), \log(\lceil D \rceil))$. Then, there exists a bivariate polynomial π_2 such that for all A, e, E, and $\eta \in (0, 1/2)$, there is a network $\Psi_{A,e,E,\eta} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\left\| |\det(A)|^{\frac{1}{p}} f(A \cdot - e) - \Psi_{A,e,E,\eta} \right\|_{L^p([-E,E]^d)} \leq \eta,$$

with $\mathcal{M}(\Psi_{A,e,E,\eta}) \leq \pi_2(\log(\eta^{-1}), \log(\lceil E\|A\|_\infty + \|e\|_\infty \rceil))$. Polynomial weight growth inherited as well.

Invariance to Affine Transformations cont'd

Proposition

Let $f \in L^p(\mathbb{R}^d)$. Assume that there is a network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\|f - \Phi_{D,\varepsilon}\|_{L^p([-D,D]^d)} \leq \varepsilon,$$

with $\mathcal{M}(\Phi_{D,\varepsilon}) \leq \pi_1(\log(\varepsilon^{-1}), \log(\lceil D \rceil))$. Then, there exists a polynomial π_2 such that for all c_i, b_i, E, and $\eta \in (0, 1/2)$, there is a network $\Psi_{c,b,E,\eta} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\left\| \sum_{i=1}^{r} c_i f(\cdot - b_i) - \Psi_{c,b,E,\eta} \right\|_{L^p([-E,E]^d)} \leq \eta,$$

with $\mathcal{M}(\Psi_{c,b,E,\eta}) \leq \pi_2(\log(\eta_c^{-1}), \log(E_b))$, where $E_b = \lceil E + \max_{i=1,\ldots,r} \|b_i\|_\infty \rceil$ and $\eta_c = \eta / \max\{1, \sum_{i=1}^{r} |c_i|\}$. Polynomial weight growth inherited as well.

Optimal representation

Theorem

Let $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$ be an affine dictionary with generator function f. Assume that there exists a network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\|f - \Phi_{D,\varepsilon}\|_{L^2([-D,D])} \leq \varepsilon,$$

with $\mathcal{M}(\Phi_{D,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(\lceil D \rceil))$ and $\mathcal{B}(\Phi_{D,\varepsilon}) \leq \pi(\varepsilon^{-1}, D)$. Assume furthermore that there exist $a, c > 0$ such that

$$\sum_{k=1}^{j-1} |\det(A_k)| \geq c\|A_j\|_\infty^a, \quad \text{for all } j \in \mathbb{N},\ j \geq 2.$$

Then, $\mathcal{D}$ is effectively representable by neural networks.

Optimal representation

Theorem

Let $\mathcal{D} = (\varphi_i)_{i\in\mathbb{N}} \subset L^2(\Omega)$ be an affine dictionary that is effectively representable by neural networks. Then, we have

$$\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) \geq \gamma^{*,\mathit{eff}}(\mathcal{C},\mathcal{D})$$

for all function classes $\mathcal{C} \subseteq L^2(\Omega)$.

In particular, if $\mathcal{C}$ is optimally representable by $\mathcal{D}$, i.e., $\gamma^{*,\mathit{eff}}(\mathcal{C},\mathcal{D}) = \gamma^*(\mathcal{C})$, then

$$\gamma^*(\mathcal{C}) \geq \gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) \geq \gamma^{*,\mathit{eff}}(\mathcal{C},\mathcal{D}) = \gamma^*(\mathcal{C})$$

and hence $\mathcal{C}$ is optimally representable by neural networks, i.e., $\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) = \gamma^*(\mathcal{C})$.

Spline wavelets

Definition

Let $N_1 := \chi_{[0,1]}$ and for $m \in \mathbb{N}$, define

$$N_{m+1} := N_1 * N_m.$$

We refer to N_m as the **univariate cardinal B-spline of order** m.

Lemma (Spline approximation)

Let $m \in \mathbb{N}$. Then, there exist a constant $C > 0$ and a polynomial π such that for all $D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, there is a neural network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying

$$\|\Phi_{D,\varepsilon} - N_m\|_{L^\infty([-D,D])} \leq \varepsilon,$$

with $\mathcal{M}(\Phi_{D,\varepsilon}) \leq C(\log(\varepsilon^{-1}) + \log(\lceil D \rceil))$ and $\mathcal{B}(\Phi_{D,\varepsilon}) \leq \pi(D)$.

Spline approximation – Proof idea

- The proof is based on the following representation

$$N_m(x) = \frac{1}{m!} \sum_{k=0}^{m+1} (-1)^k \binom{m+1}{k} \rho((x-k)^m).$$

- Next – use network that approximates polynomials.

Spline wavelet systems

Theorem (Chui & Wang, 1992)

Let $m \in \mathbb{N}$ and consider the m-th order spline

$$\psi_m(x) = \frac{1}{2^{m-1}} \sum_{j=0}^{2m-2} (-1)^j N_{2m}(j+1) \frac{d^m}{dx^m} N_{2m}(2x-j),$$

with support $[0, 2m-1]$. The set

$$\begin{aligned} \mathcal{W}_m := &\{\psi_{k,n}(x) = 2^{k/2}\psi_m(2^k x - n) : n \in \mathbb{Z}, k \in \mathbb{N}_0\} \cup \\ &\{\phi_n(x) = N_m(x-n) : n \in \mathbb{Z}\} \end{aligned}$$

is a countable complete orthonormal wavelet basis in $L^2(\mathbb{R})$.

Optimality

Theorem

Let $\Omega \subset \mathbb{R}$ be bounded and $\mathcal{D} = (\varphi_i)_{i \in \mathbb{N}} \subset L^2(\Omega)$ a spline wavelet system. Then, all function classes $\mathcal{C} \subseteq L^2(\Omega)$ that are optimally representable by $\mathcal{D}$, are optimally representable by neural networks.

Gabor systems

Definition (Gabor systems)

Let $d \in \mathbb{N}$, $f \in L^2(\mathbb{R}^d)$, and $x, \xi \in \mathbb{R}^d$. Define the translation operator $T_x \colon L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ according to

$$T_x f(t) := f(t - x)$$

and the modulation operator $M_\xi \colon L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d, \mathbb{C})$ as

$$M_\xi f(t) := e^{2\pi i \langle \xi, t \rangle} f(t).$$

Let $\Omega \subseteq \mathbb{R}^d$, $\alpha, \beta > 0$, and $g \in L^2(\mathbb{R}^d)$. The Gabor system $\mathcal{G}(g, \alpha, \beta, \Omega) \subseteq L^2(\Omega)$ is defined as

$$\mathcal{G}(g, \alpha, \beta, \Omega) := \left\{ M_\xi T_x g \big|_\Omega \colon (x, \xi) \in \alpha\mathbb{Z}^d \times \beta\mathbb{Z}^d \right\}.$$

Gabor systems

- We again build on **invariance results**.
- Suppose that **generator function** g is **well approximated** by neural networks satisfying growth conditions on connectivity and weight sizes.
- **Invariance to time-shifts**: Follows by realizing that **time-shift can be incorporated into first network layer**.
- **Invariance to frequency-shifts**: Follows from result on **cosine approximation network** with periodicity and sawtooth **twist** guaranteeing that **weights** of approximating polynomial **do not depend on frequency**. Also uses result on **multiplication network**.

Effective representability

Theorem

Let $g \in L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$, and let $\mathcal{G}(g, \alpha, \beta, \Omega)$ be the corresponding Gabor system. Suppose that there is a network $\Phi_{D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,d,1}$ satisfying

$$\|g - \Phi_{D,\varepsilon}\|_{L^\infty([-D,D]^d)} \leq \varepsilon,$$

with $\mathcal{M}(\Phi_{D,\varepsilon}) \leq \pi(\log(\varepsilon^{-1}), \log(\lceil D \rceil))$ and $\mathcal{B}(\Phi_{D,\varepsilon}) \leq \pi(\varepsilon^{-1}, D)$. Then, $\mathcal{G}(g, \alpha, \beta, \Omega)$ is effectively representable by neural networks.

Optimality

Theorem

Let $g \in L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$, and let $\mathcal{G}(g, \alpha, \beta, \Omega)$ be a Gabor system that is effectively representable by neural networks. Then, for all function classes $\mathcal{C} \subseteq L^2(\Omega)$,

$$\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) \geq \gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{G}(g, \alpha, \beta, \Omega)).$$

In particular, if $\mathcal{C}$ is optimally representable by $\mathcal{G}(g, \alpha, \beta, \Omega)$, i.e., $\gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{G}(g, \alpha, \beta, \Omega)) = \gamma^*(\mathcal{C})$, then

$$\gamma^*(\mathcal{C}) \geq \gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) \geq \gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{G}(g, \alpha, \beta, \Omega)) = \gamma^*(\mathcal{C}),$$

and $\mathcal{C}$ is optimally representable by neural networks, i.e., $\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) = \gamma^*(\mathcal{C})$.

Wilson systems

Definition (Wilson system)

Let $g \in L^2(\mathbb{R})$. We define the **Wilson system with generator** g to be the family

$$\psi_{k,\ell}(x) := \begin{cases} g(x-k), & \text{if } \ell = 0; \\ \sqrt{2}\cos(2\pi\ell x)g\left(x-\frac{k}{2}\right), & \ell \in \mathbb{N}, \ell + k \text{ even}; \\ \sqrt{2}\sin(2\pi\ell x)g\left(x-\frac{k}{2}\right), & \ell \in \mathbb{N}, \ell + k \text{ odd}. \end{cases}$$

When g is such that the **Wilson system** constitutes a basis for $L^2(\mathbb{R})$, we call $(\psi_{k,\ell})_{k\in\mathbb{Z},\,\ell\in\mathbb{N}}$ a **Wilson basis**.

Applications of Wilson systems

- Wilson bases are **unconditional bases** for **modulation spaces** [Feichtinger and Gröchenig, 1989], including Bessel potential spaces, the Segal algebra S_0, and the Schwartz space.
- Wilson bases are **optimal dictionaries** for modulation spaces [Gröchenig, 2000], [Donoho, 1993].
- occur in **pseudo-differential calculus**

Basic definitions

Definition (Modulation spaces (Feichtinger and Gröchenig, 1989))

For p with $1 \leq p < \infty$, the modulation space $M_p(\mathbb{R})$ is the space of functions $f : \mathbb{R} \to \mathbb{R}$ satisfying

$$\|f\|_{M_p} = \Big(\int_{\mathbb{R}} \Big(\int_{\mathbb{R}} |S_g f(x,y)|^p dx \Big) dy \Big)^{1/p} < \infty,$$

where

$$S_g f(x,y) := \int_{\mathbb{R}} f(t) g(t-x) e^{-2\pi i y t} dt$$

denotes the short-time Fourier transform (STFT).

Wilson basis with compactly supported generator

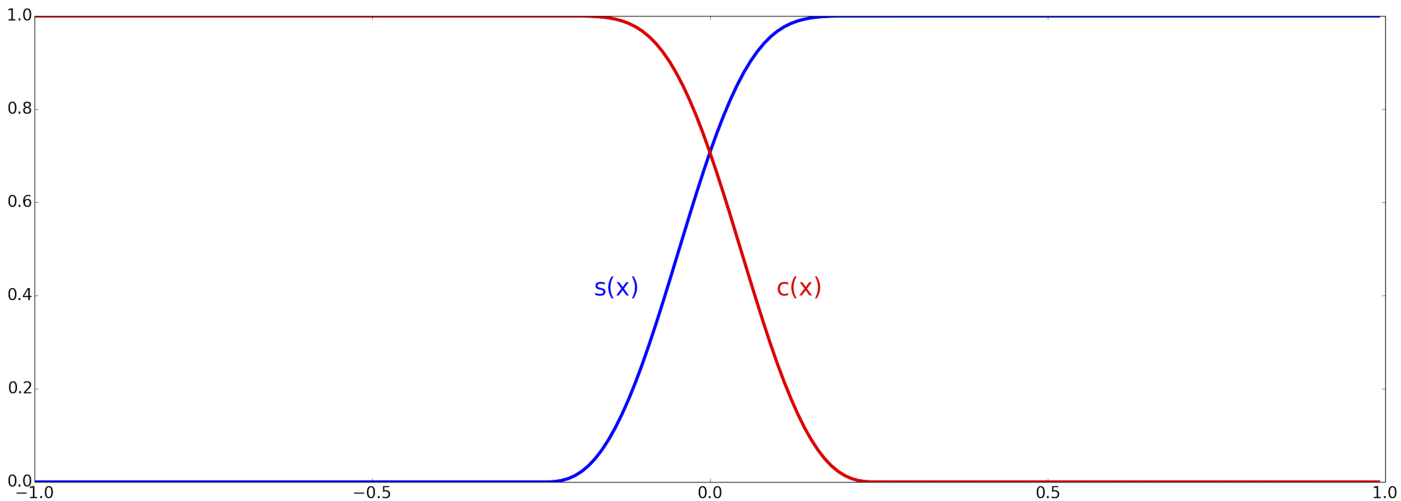
Consider the generator function

$$g(x) = \sqrt{2}s\Big(x + \frac{1}{4}\Big)c\Big(x - \frac{1}{4}\Big),$$

$$s(x) = \sin(\theta(x) + \pi/4),\ c(x) = \cos(\theta(x) + \pi/4)$$

and

$$\theta(x) = \begin{cases} \frac{\pi}{4}, & \text{if } x \geq \frac{1}{4}, \\ -\frac{\pi}{4}, & \text{if } x \leq -\frac{1}{4}, \\ 96\pi x^5 - 20\pi x^3 + \frac{15\pi}{8}x, & \text{else.} \end{cases}$$



Effective representability for Wilson systems

Theorem

Let $g \in L^2(\mathbb{R})$ be compactly supported and bounded, let $\mathcal{W}(g)$ be the corresponding Wilson system, and assume that there is a neural network $\Phi_\varepsilon \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying $\mathcal{M}(\Phi_\varepsilon) \leq \pi(\log(\varepsilon^{-1}))$, $\mathcal{B}(\Phi_\varepsilon) \leq \pi(\varepsilon^{-1})$, and

$$\|g - \Phi_\varepsilon\|_{L^\infty(\mathbb{R})} \leq \varepsilon.$$

Then, $\mathcal{W}(g)$ is effectively representable by neural networks.

The entire Wilson system can be well approximated by neural networks.

Optimal representation by neural networks

Theorem

Let $g \in L^2(\mathbb{R})$ be such that $\mathcal{W}(g)$ is effectively representable by neural networks. Then,

$$\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) \geq \gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{W}(g)).$$

In particular, if $\mathcal{C}$ is optimally representable by $\mathcal{W}(g)$, i.e., $\gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{W}(g)) = \gamma^*(\mathcal{C})$, then $\mathcal{C}$ is optimally representable by neural networks, i.e., $\gamma^{*,\mathit{eff}}_{\mathcal{NN}}(\mathcal{C}) = \gamma^*(\mathcal{C})$.

Oscillatory textures

Definition

Let the sets $\mathcal{F}_{D,a}, D, a \in \mathbb{R}_+$, be given by

$$\mathcal{F}_{D,a} = \left\{ \cos(ag)h \colon g, h \in \mathcal{S}_D \right\}.$$

Definition (Smooth functions)

For $D \in \mathbb{R}_+$, let the set $\mathcal{S}_D \subseteq C^\infty([-D, D], \mathbb{R})$ be given by

$$\mathcal{S}_D = \left\{ f \in C^\infty([-D, D], \mathbb{R}) \colon \|f^{(n)}(x)\|_{L^\infty([-D,D])} \leq n!,\ n \in \mathbb{N}_0 \right\}$$

- **Hard to approximate** for a large due to combination of rapidly oscillating cosine and warping function g.
- **Best known approximation rate** is **low-order polynomial** by [Demanet and Ying, 2007] using wave atom dictionaries.

Weierstrass function

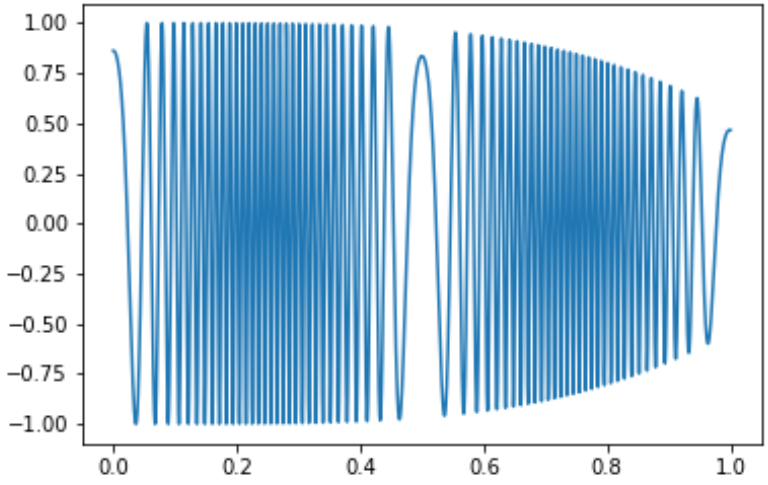
Definition

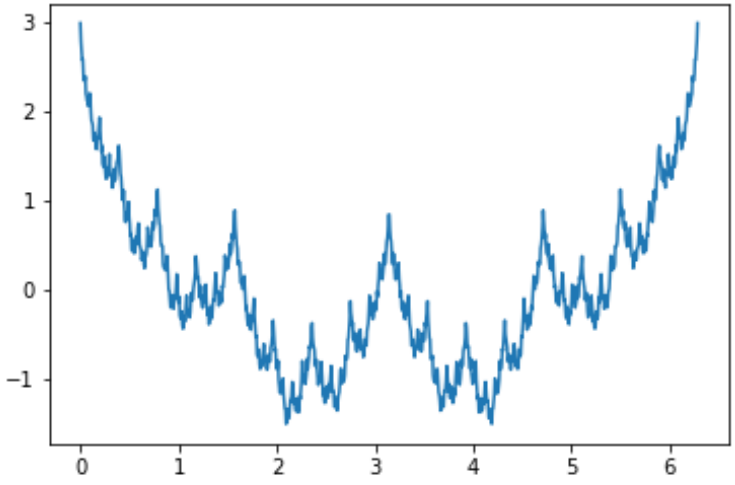
Weierstrass function

$$W_{p,a}(x) = \sum_{k=0}^{\infty} p^k \cos(a^k \pi x), \quad \text{for } p \in (0, 1/2),\ a \in \mathbb{R}_+,\ \text{with } ap \geq 1.$$

- A fractal function which is **continuous everywhere but differentiable nowhere**.
- Classical methods exploit **Hölder smoothness** and achieve **polynomial approximation rates**.

Oscillatory textures and Weierstrass function





Left: A function in $\mathcal{F}_{1,100}$. Right: The function $W_{\frac{1}{\sqrt{2}},2}$.

No known approximation algorithms achieving exponential accuracy.
$\Rightarrow$ **These functions are "hard" to approximate**.

Oscillatory textures

Proposition

Let $f \in \mathcal{F}_{D,a}$. There exists a network $\Gamma_{f,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying

$$\|f - \Gamma_{f,\varepsilon}\|_{L^\infty([-D,D])} \leq \varepsilon,$$

with $\mathcal{L}(\Gamma_{f,\varepsilon}) \leq C\lceil D \rceil (\log(\varepsilon^{-1}) + \log(\lceil a \rceil))^2$, $\mathcal{W}(\Gamma_{f,\varepsilon}) \leq 23$, and $\mathcal{B}(\Gamma_{f,\varepsilon}) \leq \max\{1/D, \lceil D \rceil\}\, \pi((\varepsilon/\lceil a \rceil)^{-1})$.

Neural networks approximate functions in $\mathcal{F}_{D,a}$ with exponential accuracy.

Weierstrass function

Proposition

There exists a neural network $\Psi_{p,a,D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ *satisfying*

$$\|\Psi_{p,a,D,\varepsilon} - W_{p,a}\|_{L^\infty([-D,D])} \leq \varepsilon,$$

with $\mathcal{L}(\Psi_{p,a,D,\varepsilon}) \leq C((\log(1/\varepsilon))^3 + (\log(1/\varepsilon))^2 \log(\lceil a \rceil) + \log(1/\varepsilon)\log(\lceil D \rceil))$, $\mathcal{W}(\Psi_{p,a,D,\varepsilon}) \leq 20$, *and* $\mathcal{B}(\Psi_{p,a,D,\varepsilon}) \leq C$.

Neural networks approximate the Weierstrass function with exponential accuracy.

The case for depth

For periodic functions, finite-width deep networks require asymptotically—in the function's "highest frequency"—smaller connectivity than finite-depth wide networks.

This statement is then extended to sufficiently smooth non-periodic functions, thereby establishing the benefit of depth for a wide class of functions.

ReLU networks realize sawtooth functions

Definition

Let $k \in \mathbb{N}$. A function $f : \mathbb{R} \to \mathbb{R}$ is called k-sawtooth if it is piecewise linear with no more than k pieces, i.e., its domain $\mathbb{R}$ can be partitioned into k intervals such that f is linear on each interval.

Lemma (Telgarsky, 2015)

Every $\Phi \in \mathcal{NN}_{\infty,\infty,1,1}$ *is* $(2\mathcal{W}(\Phi))^{\mathcal{L}(\Phi)}$*-sawtooth.*

Measure of non-linearity

Definition

For a u-periodic function $f \in C(\mathbb{R})$, we define

$$\xi(f) := \sup_{\delta \in [0,u)} \inf_{c,d \in \mathbb{R}} \|f(x) - (cx + d)\|_{L^\infty([\delta, \delta+u])}.$$

$\xi(f)$ measures the error incurred by the best linear approximation of f on any segment of length equal to the period of f; it can hence be interpreted as quantifying the non-linearity of f.

Finite-depth networks

Finite-depth networks with width scaling poly-logarithmically in the "highest frequency" of the periodic function to be approximated can not achieve arbitrarily small approximation error.

Impossibility result

Proposition

Let $f \in C(\mathbb{R})$ be a non-constant u-periodic function, $L \in \mathbb{N}$, and π a polynomial. Then, there exists an $a \in \mathbb{N}$ such that for every network $\Phi \in \mathcal{NN}_{L,\infty,1,1}$ with $\mathcal{W}(\Phi) \leq \pi(\log(a))$,

$$\|f(a\,\cdot) - \Phi\|_{L^\infty([0,u])} \geq \xi(f) > 0.$$

- $\xi(\cos) > 0$
- Approximation of $f(x) = \cos(ax)$ at arbitrarily small error with finite-depth networks requires faster than poly-logarithmic connectivity growth in a.

But deep networks can do this

Theorem (Cosine approximation)

There exists a constant $C > 0$ *such that for every* $a, D \in \mathbb{R}_+$, $\varepsilon \in (0, 1/2)$, *there is a network* $\Psi_{a,D,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ *satisfying* $\mathcal{L}(\Psi_{a,D,\varepsilon}) \leq C((\log(1/\varepsilon))^2 + \log(\lceil aD \rceil))$, $\mathcal{W}(\Psi_{a,D,\varepsilon}) \leq 16$, $\mathcal{B}(\Psi_{a,D,\varepsilon}) \leq C$, *and*

$$\|\Psi_{a,D,\varepsilon} - \cos(a\,\cdot)\|_{L^\infty([-D,D])} \leq \varepsilon.$$

Smooth functions

Theorem (Frenzen et al., 2010)

Let $f \in C^3([a,b])$ and consider a piecewise linear approximation of f on $[a,b]$ that is accurate to within ε in the $L^\infty([a,b])$-norm. The minimal number of linear pieces required to accomplish this scales according to

$$s(\varepsilon) \sim \frac{c}{\sqrt{\varepsilon}},\ \varepsilon \to 0, \quad \textit{where } c = \frac{1}{4}\int_a^b \sqrt{|f''(x)|}dx.$$

Impossibility result for smooth function

Theorem

Let $f \in C^3([a,b])$ *with* $\int_a^b \sqrt{|f''(x)|}dx > 0$, $L \in \mathbb{N}$, *and* π *a polynomial. Then, there exists* $\varepsilon > 0$ *such that for every network* $\Phi \in \mathcal{NN}_{L,\infty,1,1}$ *with* $\mathcal{W}(\Phi) \leq \pi(\log(\varepsilon^{-1}))$,

$$\|f - \Phi\|_{L^\infty([a,b])} > \varepsilon.$$

Any function that is at least three times continuously differentiable cannot be approximated by finite-depth networks with connectivity scaling poly-logarithmically in the inverse of the approximation error.

But deep networks can do this

Definition

For $D \in \mathbb{R}_+$, let the set $\mathcal{S}_D \subseteq C^\infty([-D, D], \mathbb{R})$ be given by

$$\mathcal{S}_D = \left\{ f \in C^\infty([-D, D], \mathbb{R}) \colon \|f^{(n)}(x)\|_{L^\infty([-D,D])} \leq n!,\ n \in \mathbb{N}_0 \right\}$$

Proposition (Smooth functions)

There exist a constant $C > 0$ and a polynomial π such that for all $D \in \mathbb{R}_+$, $f \in \mathcal{S}_D$, and $\varepsilon \in (0, 1/2)$, there is a network $\Psi_{f,\varepsilon} \in \mathcal{NN}_{\infty,\infty,1,1}$ satisfying $\mathcal{L}(\Psi_{f,\varepsilon}) \leq C\lceil D \rceil (\log(\varepsilon^{-1}))^2$, $\mathcal{W}(\Psi_{f,\varepsilon}) \leq 23$, $\mathcal{B}(\Psi_{f,\varepsilon}) \leq \max\{1/D, \lceil D \rceil\}\pi(\varepsilon^{-1})$, and

$$\|\Psi_{f,\varepsilon} - f\|_{L^\infty([-D,D])} \leq \varepsilon.$$

Weights can be converted to $\leq const.$ at poly-log increase in depth and width $\Rightarrow$ overall connectivity still poly-log in ε^{-1}.

References

H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen,
Optimal approximation with sparsely connected deep neural networks, SIAM Journal on Mathematics of Data Science, Vol. 1, No. 1, pp. 8-45, 2019.

P. Grohs, D. Perekrestenko, D. Elbrächter, and H. Bölcskei,
Deep neural network approximation theory, IEEE Transactions on Information Theory, invited paper, 2019.

Yours truly