Approximation Theory and Expressivity I

Gitta Kutyniok

(Technische Universität Berlin and University of Tromsø)

Banach Center – Oberwolfach Graduate Seminar:
Mathematics of Deep Learning
Polish Academy of Sciences, Będlewo, November 17 – 23, 2019

Viewpoint of Approximation Theory

Main Research Goal

Some Questions:

- Which architecture to choose for a particular application?
- What is the expressive power of a given architecture?
- What effect has the depth of a neural network in this respect?
- What is the complexity of the approximating neural network?
- What are suitable function spaces to consider?
- Can deep neural networks beat the curse of dimensionality?

Main Research Goal

Some Questions:

- Which architecture to choose for a particular application?
- What is the expressive power of a given architecture?
- What effect has the depth of a neural network in this respect?
- What is the complexity of the approximating neural network?
- What are suitable function spaces to consider?
- Can deep neural networks beat the curse of dimensionality?

Mathematical Problem:

Under which conditions on a neural network Φ and an activation function ϱ can every function from a prescribed function class $\mathcal{C}$ be arbitrarily well approximated, i.e.

$$\|R_\varrho(\Phi) - f\| \leq \varepsilon, \quad \text{for all } f \in \mathcal{C}.$$

Function Approximation in a Nutshell

Goal: Given

- a function class $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$,
- a function system $(\varphi_i)_{i \in I} \subseteq L^2(\mathbb{R}^d)$.

Measure the suitability of $(\varphi_i)_{i \in I}$ for uniformly approximating functions from $\mathcal{C}$.

Function Approximation in a Nutshell

Goal: Given

- a function class $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$,
- a function system $(\varphi_i)_{i\in I} \subseteq L^2(\mathbb{R}^d)$.

Measure the suitability of $(\varphi_i)_{i\in I}$ for uniformly approximating functions from $\mathcal{C}$.

Definition: The *error of best N-term approximation* of some $f \in \mathcal{C}$ is given by

$$\|f - f_N\|_{L^2(\mathbb{R}^d)} := \inf_{I_N \subset I, \#I_N = N, (c_i)_{i\in I_N}} \|f - \sum_{i\in I_N} c_i \varphi_i\|_{L^2(\mathbb{R}^d)}.$$

Function Approximation in a Nutshell

Goal: Given

- a function class $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$,
- a function system $(\varphi_i)_{i \in I} \subseteq L^2(\mathbb{R}^d)$.

Measure the suitability of $(\varphi_i)_{i \in I}$ for uniformly approximating functions from $\mathcal{C}$.

Definition: The *error of best N-term approximation* of some $f \in \mathcal{C}$ is given by

$$\|f - f_N\|_{L^2(\mathbb{R}^d)} := \inf_{I_N \subset I, \# I_N = N, (c_i)_{i \in I_N}} \|f - \sum_{i \in I_N} c_i \varphi_i\|_{L^2(\mathbb{R}^d)}.$$

The largest $\gamma > 0$ such that

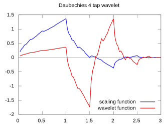
$$\sup_{f \in \mathcal{C}} \|f - f_N\|_{L^2(\mathbb{R}^d)} = O(N^{-\gamma}) \qquad \text{as } N \to \infty$$

determines the *optimal (sparse) approximation rate* of $\mathcal{C}$ by $(\varphi_i)_{i \in I}$.

The Wavelet Transform

Definition for $L^2(\mathbb{R})$: Let $\varphi \in L^2(\mathbb{R})$ be a scaling function and $\psi \in L^2(\mathbb{R})$ a wavelet. Then the associated wavelet system is defined by

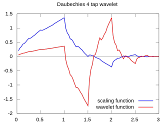
$$\{\varphi(x-m) : m \in \mathbb{Z}\} \cup \{2^{j/2}\,\psi(2^j x - m) : j \geq 0, m \in \mathbb{Z}\}.$$



The Wavelet Transform

Definition for $L^2(\mathbb{R})$: Let $\varphi \in L^2(\mathbb{R})$ be a scaling function and $\psi \in L^2(\mathbb{R})$ a wavelet. Then the associated wavelet system is defined by

$$\{\varphi(x-m) : m \in \mathbb{Z}\} \cup \{2^{j/2}\,\psi(2^j x - m) : j \geq 0, m \in \mathbb{Z}\}.$$

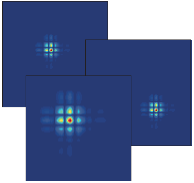


Definition for $L^2(\mathbb{R}^2)$: A wavelet system is defined by

$$\{\varphi^{(1)}(x-m) : m \in \mathbb{Z}^2\} \cup \{2^j\psi^{(i)}(2^j x - m) : j \geq 0, m \in \mathbb{Z}^2, i = 1,2,3\},$$

where

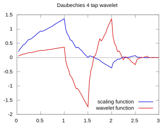
$$\varphi^{(1)}(x) = \varphi(x_1)\varphi(x_2) \quad \text{and} \quad \begin{aligned} \psi^{(1)}(x) &= \varphi(x_1)\psi(x_2), \\ \psi^{(2)}(x) &= \psi(x_1)\varphi(x_2), \\ \psi^{(3)}(x) &= \psi(x_1)\psi(x_2). \end{aligned}$$



The Wavelet Transform

Definition for $L^2(\mathbb{R})$: Let $\varphi \in L^2(\mathbb{R})$ be a scaling function and $\psi \in L^2(\mathbb{R})$ a wavelet. Then the associated wavelet system is defined by

$$\{\varphi(x-m) : m \in \mathbb{Z}\} \cup \{2^{j/2}\psi(2^j x - m) : j \geq 0, m \in \mathbb{Z}\}.$$

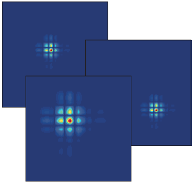


Definition for $L^2(\mathbb{R}^2)$: A wavelet system is defined by

$$\{\varphi^{(1)}(x-m) : m \in \mathbb{Z}^2\} \cup \{2^j\psi^{(i)}(2^j x - m) : j \geq 0, m \in \mathbb{Z}^2, i = 1, 2, 3\},$$

where

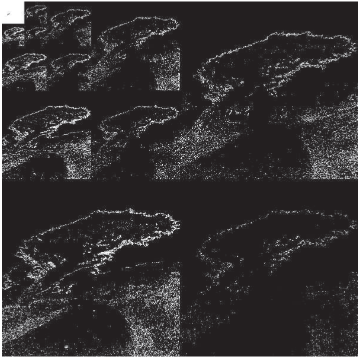
$$\varphi^{(1)}(x) = \varphi(x_1)\varphi(x_2) \quad \text{and} \quad \begin{aligned} \psi^{(1)}(x) &= \varphi(x_1)\psi(x_2), \\ \psi^{(2)}(x) &= \psi(x_1)\varphi(x_2), \\ \psi^{(3)}(x) &= \psi(x_1)\psi(x_2). \end{aligned}$$



Wavelet Transform (JPEG2000):

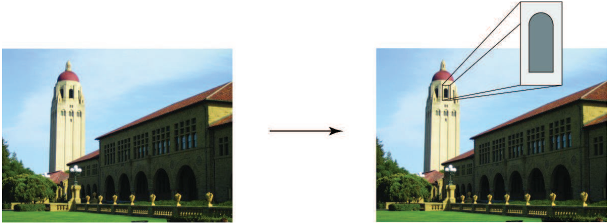
$$f \mapsto ((\langle f, \varphi_m \rangle)_m, (\langle f, \psi_{j,m,i} \rangle)_{j,m,i}).$$

Application of the Wavelet Transform

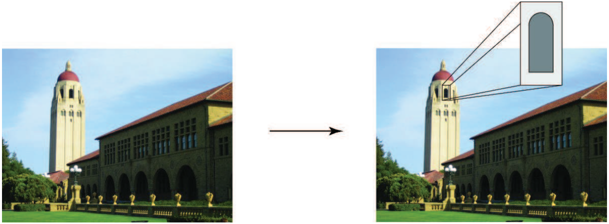


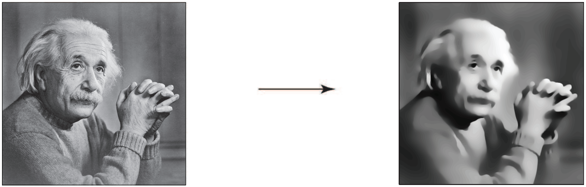
What is an Image?

What is an Image?



What is an Image?





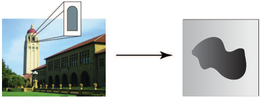
Fitting Model

Definition (Donoho; 2001):

The set of cartoon-like functions $\mathcal{E}^2(\mathbb{R}^2)$ is defined by

$$\mathcal{E}^2(\mathbb{R}^2) = \{f \in L^2(\mathbb{R}^2) : f = f_0 + f_1 \cdot \chi_B\},$$

where $\emptyset \neq B \subset [0,1]^2$ simply connected with C^2-boundary and bounded curvature, and $f_i \in C^2(\mathbb{R}^2)$ with supp $f_i \subseteq [0,1]^2$ and $\|f_i\|_{C^2} \leq 1$, $i = 0, 1$.

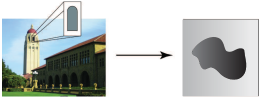


Fitting Model

Definition (Donoho; 2001):
The set of cartoon-like functions $\mathcal{E}^2(\mathbb{R}^2)$ is defined by

$$\mathcal{E}^2(\mathbb{R}^2) = \{f \in L^2(\mathbb{R}^2) : f = f_0 + f_1 \cdot \chi_B\},$$

where $\emptyset \neq B \subset [0,1]^2$ simply connected with C^2-boundary and bounded curvature, and $f_i \in C^2(\mathbb{R}^2)$ with supp $f_i \subseteq [0,1]^2$ and $\|f_i\|_{C^2} \leq 1$, $i = 0, 1$.



Theorem (Donoho; 2001):
Let $(\psi_\lambda)_\lambda \subseteq L^2(\mathbb{R}^2)$. Allowing only polynomial depth search, we have the following optimal behavior for $f \in \mathcal{E}^2(\mathbb{R}^2)$:

$$\|f - f_N\|_2 \asymp N^{-1} \quad \text{as } N \to \infty.$$

What can Wavelets do?

Problem:

- For $f \in \mathcal{E}^2(\mathbb{R}^2)$, wavelets only achieve $\|f - f_N\|_2^2 \asymp N^{-1}$, $N \to \infty$.
- Isotropic structure of wavelets:

$$\{2^j \psi(\begin{pmatrix} 2^j & 0 \\ 0 & 2^j \end{pmatrix} x - m) : j \geq 0, m \in \mathbb{Z}^2\}.$$

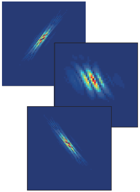
- Wavelets cannot sparsely represent cartoon-like functions.

Intuitive explanation:

Shearlets

Shearlets (K, Labate; 2006):

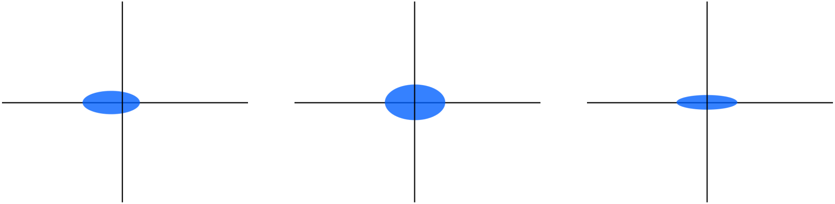
$$A_j := \begin{pmatrix} 2^j & 0 \\ 0 & 2^{j/2} \end{pmatrix}, \qquad S_k := \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}, \quad j,k \in \mathbb{Z}.$$



Then

$$\psi_{j,k,m} := 2^{\frac{3j}{4}} \psi(S_k A_j \cdot - m).$$

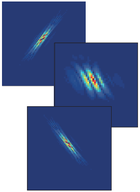
Notice: *$x \mapsto S_k A_j x - m$ is an affine-linear map!*



Shearlets

Shearlets (K, Labate; 2006):

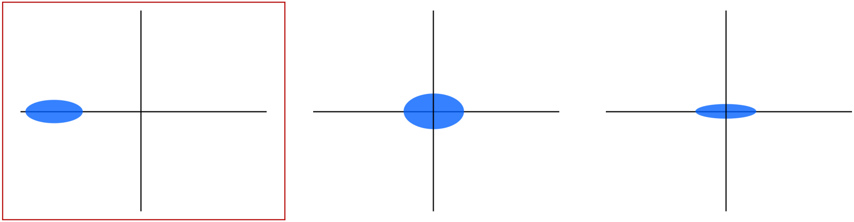
$$A_j := \begin{pmatrix} 2^j & 0 \\ 0 & 2^{j/2} \end{pmatrix}, \qquad S_k := \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}, \quad j,k \in \mathbb{Z}.$$



Then

$$\psi_{j,k,m} := 2^{\frac{3j}{4}} \psi(S_k A_j \cdot - m).$$

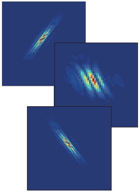
Notice: *$x \mapsto S_k A_j x - m$ is an affine-linear map!*



Shearlets

Shearlets (K, Labate; 2006):

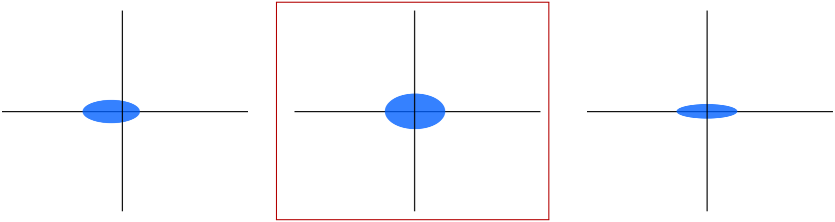
$$A_j := \begin{pmatrix} 2^j & 0 \\ 0 & 2^{j/2} \end{pmatrix}, \qquad S_k := \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}, \quad j,k \in \mathbb{Z}.$$



Then

$$\psi_{j,k,m} := 2^{\frac{3j}{4}} \psi(S_k A_j \cdot -m).$$

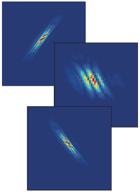
Notice: *$x \mapsto S_k A_j x - m$ is an affine-linear map!*



Shearlets

Shearlets (K, Labate; 2006):

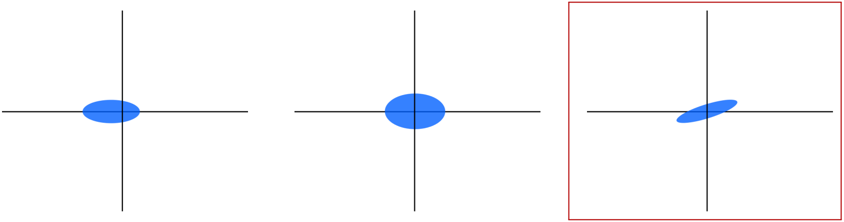
$$A_j := \begin{pmatrix} 2^j & 0 \\ 0 & 2^{j/2} \end{pmatrix}, \qquad S_k := \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}, \quad j,k \in \mathbb{Z}.$$



Then

$$\psi_{j,k,m} := 2^{\frac{3j}{4}} \psi(S_k A_j \cdot - m).$$

Notice: *$x \mapsto S_k A_j x - m$ is an affine-linear map!*



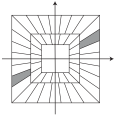
(Cone-adapted) Shearlet Systems

Definition (K, Labate; 2006):
The (cone-adapted) shearlet system $\mathcal{SH}(\phi, \psi, \tilde{\psi})$ generated by $\phi \in L^2(\mathbb{R}^2)$ and $\psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$ is the union of

$$\{\phi(\cdot - m) : m \in \mathbb{Z}^2\},$$

$$\{2^{3j/4}\psi(S_k A_{2^j} \cdot - m) : j \geq 0, |k| \leq \lceil 2^{j/2} \rceil, m \in \mathbb{Z}^2\},$$

$$\{2^{3j/4}\tilde{\psi}(\tilde{S}_k \tilde{A}_{2^j} \cdot - m) : j \geq 0, |k| \leq \lceil 2^{j/2} \rceil, m \in \mathbb{Z}^2\}.$$



(Cone-adapted) Shearlet Systems

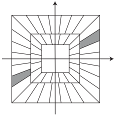
Definition (K, Labate; 2006):

The (cone-adapted) shearlet system $\mathcal{SH}(\phi, \psi, \tilde{\psi})$ generated by $\phi \in L^2(\mathbb{R}^2)$ and $\psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$ is the union of

$$\{\phi(\cdot - m) : m \in \mathbb{Z}^2\},$$

$$\{2^{3j/4}\psi(S_k A_{2^j} \, \cdot \, - m) : j \geq 0, |k| \leq \lceil 2^{j/2} \rceil, m \in \mathbb{Z}^2\},$$

$$\{2^{3j/4}\tilde{\psi}(\tilde{S}_k \tilde{A}_{2^j} \, \cdot \, - m) : j \geq 0, |k| \leq \lceil 2^{j/2} \rceil, m \in \mathbb{Z}^2\}.$$



Theorem (K, Labate, Lim, Weiss; 2006):

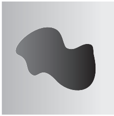
For $\psi, \tilde{\psi}$ classical shearlets, $\mathcal{SH}(\phi, \psi, \tilde{\psi})$ is a Parseval frame for $L^2(\mathbb{R}^2)$:

$$\|f\|_2^2 = \sum_{\sigma \in \mathcal{SH}(\phi, \psi, \tilde{\psi})} |\langle f, \sigma \rangle|^2 \quad \text{for all } f \in L^2(\mathbb{R}^2).$$

Optimally Sparse Approximation

Theorem (K, Lim; 2011):

Let $\phi, \psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$ be compactly supported, and let $\hat{\psi}$, $\hat{\tilde{\psi}}$ satisfy certain decay condition. Then $\mathcal{SH}(\phi, \psi, \tilde{\psi})$ provides an optimally sparse approximation of $f \in \mathcal{E}^2(\mathbb{R}^2)$, i.e.,



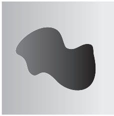
$$\|f - f_N\|_2 \lesssim N^{-1}(\log N)^{\frac{3}{2}} \quad \text{as } N \to \infty.$$

Optimally Sparse Approximation

Theorem (K, Lim; 2011):

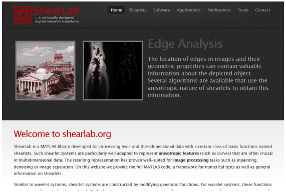
Let $\phi, \psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$ be compactly supported, and let $\hat{\psi}$, $\hat{\tilde{\psi}}$ satisfy certain decay condition. Then $\mathcal{SH}(\phi, \psi, \tilde{\psi})$ provides an optimally sparse approximation of $f \in \mathcal{E}^2(\mathbb{R}^2)$, i.e.,

$$\|f - f_N\|_2 \lesssim N^{-1}(\log N)^{\frac{3}{2}} \quad \text{as } N \to \infty.$$



2D&3D (parallelized) Fast Shearlet Transform (`www.ShearLab.org`):

- Matlab *(K, Lim, Reisenhofer; 2013)*
- Julia *(Loarca; 2017)*
- Python *(Look; 2018)*
- Tensorflow *(K, Loarca; 2019)*



Function Approximation in a Nutshell

Goal: Given

- a function class $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$,
- a function system $(\varphi_i)_{i\in I} \subseteq L^2(\mathbb{R}^d)$.

Measure the suitability of $(\varphi_i)_{i\in I}$ for uniformly approximating functions from $\mathcal{C}$.

Definition: The *error of best N-term approximation* of some $f \in \mathcal{C}$ is given by

$$\|f - f_N\|_{L^2(\mathbb{R}^d)} := \inf_{I_N \subset I, \#I_N = N, (c_i)_{i\in I_N}} \|f - \sum_{i\in I_N} c_i \varphi_i\|_{L^2(\mathbb{R}^d)}.$$

The largest $\gamma > 0$ such that

$$\sup_{f\in\mathcal{C}} \|f - f_N\|_{L^2(\mathbb{R}^d)} = O(N^{-\gamma}) \qquad \text{as } N \to \infty$$

determines the *optimal (sparse) approximation rate* of $\mathcal{C}$ by $(\varphi_i)_{i\in I}$.

Expressivity of Deep Neural Networks

Goals

General Question:

Let f belong to a function class, and let $\mathcal{C}$ be a class of neural networks.

Which complexity does a neural network $\Phi \in \mathcal{C}$, which approximates f up to ε, need to have?

Goals

General Question:

Let f belong to a function class, and let $\mathcal{C}$ be a class of neural networks.

Which complexity does a neural network $\Phi \in \mathcal{C}$, which approximates f up to ε, need to have?

Complexity:

We measure complexity of a neural network Φ by

$$M(\Phi) := \sum_{\ell=1}^{L} \|A_\ell\|_0 + \|b_\ell\|_0,$$

i.e., the number of weights (edges), where $\|\cdot\|_0$ is the number of non-zero entries.

Universality Results

Universality of Shallow Neural Networks

Remark:

Assume ϱ is a polynomial of degree q. Then $\varrho(Ax + b)$ is also a polynomial of degree q, hence $R_\varrho(\Phi)$ is also a polynomial of degree $\leq L \cdot q$. Hence in this case $C(\mathbb{R}^d)$ cannot be well approximated.

Universality of Shallow Neural Networks

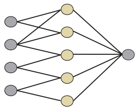
Remark:

Assume ϱ is a polynomial of degree q. Then $\varrho(Ax+b)$ is also a polynomial of degree q, hence $R_\varrho(\Phi)$ is also a polynomial of degree $\leq L \cdot q$. Hence in this case $C(\mathbb{R}^d)$ cannot be well approximated.

Universal Approximation Theorem (Cybenko, 1989)(Hornik, 1991):

Let $\varrho : \mathbb{R} \to \mathbb{R}$ be continuous, but not a polynomial. Also, fix $d \geq 1$, $L = 2$, $N_L \geq 1$ and a compact set $K \subseteq \mathbb{R}^d$. Then, for any continuous $f : \mathbb{R}^d \to \mathbb{R}^{N_L}$ and every $\varepsilon > 0$, there exist $M, N \in \mathbb{N}$ and $\Phi \in \mathcal{NN}_{d,M,N,2}$ with

$$\sup_{x \in K} |R_\varrho(\Phi)(x) - f(x)| \leq \varepsilon.$$



Proof: ...on the board!

General Statement:

"Every continuous function on a compact set can be arbitrarily well approximated with a neural network with one single hidden layer."

General Approximation Power of Neural Networks

"Universal Network Theorem" (Maiorov and Pinkus, 1999):
There exists an activation function $\varrho : \mathbb{R} \to \mathbb{R}$ such that for any $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ compact, $f : K \to \mathbb{R}$ continuous, and any $\varepsilon > 0$, there exists $M, N \in \mathbb{N}$ (only dependent on d) and $\Phi \in \mathcal{NN}_{d,M,N,3}$ with

$$\sup_{x \in K} |R_\varrho(\Phi)(x) - f(x)| \leq \varepsilon.$$

General Approximation Power of Neural Networks

"Universal Network Theorem" (Maiorov and Pinkus, 1999):
There exists an activation function $\varrho : \mathbb{R} \to \mathbb{R}$ such that for any $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ compact, $f : K \to \mathbb{R}$ continuous, and any $\varepsilon > 0$, there exists $M, N \in \mathbb{N}$ (only dependent on d) and $\Phi \in \mathcal{NN}_{d,M,N,3}$ with

$$\sup_{x \in K} |R_\varrho(\Phi)(x) - f(x)| \leq \varepsilon.$$

The weights can be arbitrarily huge!

Non-Exhaustive List of Expressivity Results

Approximation by NNs with one Single Hidden Layer:

- Bounds in terms terms of nodes and sample size (Barron; 1993, 1994).
- Localized approximations (Chui, Li, and Mhaskar; 1994).
- Fundamental lower bound on approximation rates (DeVore, Oskolkov, and Petrushev; 1997), (Candès; 1998).
- Approximation using specific rectifiers (Cybenko; 1989).
- Approximation of specific function classes (Mhaskar and Micchelli; 1995), (Mhaskar; 1996).

Approximation by NNs with Multiple Hidden Layers:

- Approximation with sigmoidal rectifiers (Hornik, Stinchcombe, and White; 1989).
- Approximation of continuous functions (Funahashi; 1998).
- Relation between one and multi layers (Eldan and Shamir; 2016), (Mhaskar and Poggio; 2016).
- Approximation by DDNs versus best M-term approximations by wavelets (Shaham, Cloninger, and Coifman; 2017).
- Complexity of approximation with ReLU networks (Yarotzky; 2017).
- Phase diagram of approximation rates (Yarotsky and Zhevnerchuk; 2019).
- Nonlinear Approximation and (Deep) ReLU Networks (Daubechies, DeVore, Foucart, Hanin, and Petrova; 2019).

Lower Bounds for Approximation

Vapnik-Chervonenkis Dimension

Definition: Let X be a set, $S \subset X$, and let $H \subseteq \{h : X \to \{0,1\}\}$ be a set of binary valued maps on X. We define

$$H_{|S} := \{h_{|S} : h \in H\},$$

which, in words, is the *restriction of the function class H to S*. The *VC dimension* of H is now defined as

$$\mathrm{VCdim}(H) := \sup\left\{ m \in \mathbb{N} : \sup_{|S| \leq m} |H_{|S}| = 2^m \right\}.$$

Intuition:

- This is a tool for understanding the classification capabilities of a function class.
- The VC dimension of H is the largest integer m such that there exists a set $S \subset X$ containing only m points such that $H_{|S}$ has the maximum possible cardinality given by 2^m.

Vapnik-Chervonenkis Dimension

Definition: Let X be a set, $S \subset X$, and let $H \subseteq \{h : X \to \{0,1\}\}$ be a set of binary valued maps on X. We define

$$H_{|S} := \{h_{|S} : h \in H\},$$

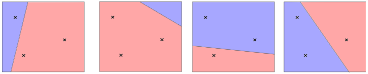
which, in words, is the *restriction of the function class H to S*. The *VC dimension* of H is now defined as

$$\mathrm{VCdim}(H) := \sup\left\{ m \in \mathbb{N} : \sup_{|S| \leq m} |H_{|S}| = 2^m \right\}.$$

Example: Let $X = \mathbb{R}^2$ and $h = \chi_{\mathbb{R}^+}$ and

$$H = \left\{ h_{\theta,t} := h\left(\left\langle \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, \bullet - t \right\rangle \right) \;\Big|\; \theta \in [-\pi,\pi], t \in \mathbb{R}^2 \right\}.$$

Then H is the set of all linear classifiers. If S contains 3 points in general position, then $|H_{|S}| = 8$. On the other hand, 4 points cannot be shattered by H.



Vapnik-Chervonenkis Dimension

Definition: Let X be a set, $S \subset X$, and let $H \subseteq \{h : X \to \{0,1\}\}$ be a set of binary valued maps on X. We define

$$H_{|S} := \{h_{|S} : h \in H\},$$

which, in words, is the *restriction of the function class H to S*. The *VC dimension* of H is now defined as

$$\mathrm{VCdim}(H) := \sup \left\{ m \in \mathbb{N} : \sup_{|S| \leq m} |H_{|S}| = 2^m \right\}.$$

Theorem (Anthony, Bartlett; 2009): Let ρ be piecewise polynomial with p pieces of degree at most ℓ, $h = \chi_{\mathbb{R}^+}$, and for $N, M, d \in \mathbb{N}$ we define

$$H_{N,M,d,L} := \{h \circ \Phi \ : \ \Phi \in \mathcal{NN}_{d,M,N,L}\}.$$

Then

$$\mathrm{VCdim}(H_{N,M,d,L}) = \mathcal{O}(ML \log_2(M) + ML^2).$$

Sparse Connectivity and More

Key Questions:

- How well can functions be approximated by neural networks with few non-zero weights?
  - Can we derive a lower bound on the necessary number of weights?
  - Can we construct neural networks which attain this bound?
- Are neural networks as good approximators as wavelets and shearlets?

Rate Distortion Theory

Definition:

- Let $d \in \mathbb{N}, \Omega \in \mathbb{R}^d$ and $\mathcal{C} \subset L^2(\Omega)$. For any $l \in \mathbb{N}$

$$\mathcal{E}^l = \{E : \mathcal{C} \to \{0,1\}^l\}$$

  is called the set of binary encoders of length l and

$$\mathcal{D}^l = \{D : \{0,1\}^l \to L^2(\Omega)\}$$

  is called the set of binary decoders of length l.

Rate Distortion Theory

Definition:

- Let $d \in \mathbb{N}, \Omega \in \mathbb{R}^d$ and $\mathcal{C} \subset L^2(\Omega)$. For any $l \in \mathbb{N}$
$$\mathcal{E}^l = \{E : \mathcal{C} \to \{0,1\}^l\}$$
is called the set of binary encoders of length l and
$$\mathcal{D}^l = \{D : \{0,1\}^l \to L^2(\Omega)\}$$
is called the set of binary decoders of length l.
- A pair $(E, D) \in \mathcal{E}^l \times \mathcal{D}^l$ achieves distortion $\varepsilon > 0$ over $\mathcal{C}$, if
$$\sup_{f \in \mathcal{C}} \|D(E(f)) - f\|_{L^2} \leq \varepsilon.$$

Rate Distortion Theory

Definition:

- Let $d \in \mathbb{N}, \Omega \in \mathbb{R}^d$ and $\mathcal{C} \subset L^2(\Omega)$. For any $l \in \mathbb{N}$

$$\mathcal{E}^l = \{E : \mathcal{C} \to \{0,1\}^l\}$$

  is called the set of binary encoders of length l and

$$\mathcal{D}^l = \{D : \{0,1\}^l \to L^2(\Omega)\}$$

  is called the set of binary decoders of length l.
- A pair $(E, D) \in \mathcal{E}^l \times \mathcal{D}^l$ achieves distortion $\varepsilon > 0$ over $\mathcal{C}$, if

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

- For $\varepsilon > 0$, the minimal code length $L(\varepsilon, \mathcal{C})$ is

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

  The optimal exponent $\gamma^*(\mathcal{C})$ is

$$\gamma^*(\mathcal{C}) := \inf\{\gamma \in \mathbb{R} : L(\varepsilon, \mathcal{C}) = O(\varepsilon^{-\gamma})\}.$$

Optimal Exponent

Example: ...on the board!

Theorem:
For $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$, the optimal $N-$term approximation rate is given by

$$N^{-\frac{1}{\gamma^*(\mathcal{C})}}.$$

A Fundamental Lower Bound

Theorem (Bölcskei, Grohs, K, and Petersen; 2017):
Let $d \in \mathbb{N}$, $\varrho : \mathbb{R} \to \mathbb{R}$, $c > 0$ and $\mathcal{C} \subset L^2(\mathbb{R}^d)$. Let

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

be such that all weights of $Learn(\varepsilon, f)$ can be encoded with $-c \log_2(\varepsilon)$ bits. Moreover

$$\sup_{f \in \mathcal{C}} \|f - R_\varrho(Learn(\varepsilon, f))\| < \varepsilon.$$

Then, for all $\gamma < \gamma^*(\mathcal{C})$

$$\sup_{\varepsilon \in (0, \frac{1}{2})} \varepsilon^\gamma \sup_{f \in \mathcal{C}} M(Learn(\varepsilon, f)) = \infty.$$

Proof: ...on the board!

A Fundamental Lower Bound

Theorem (Bölcskei, Grohs, K, and Petersen; 2017):
Let $d \in \mathbb{N}$, $\varrho : \mathbb{R} \to \mathbb{R}$, $c > 0$ and $\mathcal{C} \subset L^2(\mathbb{R}^d)$. Let

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

be such that all weights of $Learn(\varepsilon, f)$ can be encoded with $-c \log_2(\varepsilon)$ bits. Moreover

$$\sup_{f \in \mathcal{C}} \| f - R_\varrho(Learn(\varepsilon, f)) \| < \varepsilon.$$

Then, for all $\gamma < \gamma^*(\mathcal{C})$

$$\sup_{\varepsilon \in (0,\frac{1}{2})} \varepsilon^\gamma \sup_{f \in \mathcal{C}} M(Learn(\varepsilon, f)) = \infty.$$

Proof: ...on the board!

What happens for $\gamma = \gamma^*(\mathcal{C})$?

Optimality

Goal:

- How well can functions be approximated by neural networks with few non-zero weights?
  - Can we derive a lower bound on the necessary number of weights?
  - Can we construct neural networks which attain this bound?
- Are neural networks as good approximators as wavelets and shearlets?

Strategy:

- Consider general (affine) systems including wavelets, shearlets, etc.
- Mimic the N-term approximation concept with deep neural networks.

Affine Systems

Definition:

Let $d \in \mathbb{N}, (A_j)_{j \in \mathbb{N}} \subseteq GL(\mathbb{R}^d), \psi_1, ...\psi_S \in L^2(\mathbb{R}^d)$ be compactly supported. Then we define affine systems as

$$\{det(A_j)^{\frac{1}{2}} \psi_s(A_j x - b) | s = 1, ...S, b \in \mathbb{Z}^d, j \in \mathbb{N}\}.$$

Examples:

- Wavelet systems
- Shearlet systems
- ...

Memory-Optimal Neural Networks

Theorem (Bölcskei, Grohs, K, and Petersen; 2017):
Let $\Omega \subseteq \mathbb{R}^d$ be bounded and $(\phi_i)_i \subseteq L^2(\Omega)$ be an affine system with ψ_s, $1 \leq s \leq S$ defined as before. Further, let $\rho : \mathbb{R} \to \mathbb{R}$ be an activation function. Assume that there exists $\varepsilon > 0$ such that, for all $D, \varepsilon > 0$ and s, there exists $\Phi_{D,\varepsilon} \in \mathcal{NN}_{d,C,2C,L}$ with

$$\|\psi_s - R_\rho(\Phi_{D,\varepsilon})\|_{L^2([-D,D]^d)} \leq \varepsilon \quad \text{for some } C > 0.$$

Let $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$. Then, if $\varepsilon > 0, M \in \mathbb{N}, f \in L^2(\Omega) \cap \mathcal{C}$ such that there exists $(d_i)_{i=1}^M$ with

$$\|f - \sum_{i=1}^{M} d_i \phi_i\| \leq \varepsilon,$$

then there exists a deep neural network Φ with $O(M)$ edges such that

$$\|f - R_\rho(\Phi)\| \leq 2\varepsilon.$$

This produces memory-optimal deep neural networks.

Proof: ...on the board!

Again: Memory-Optimal Neural Networks

Corollary: Assume an affine system $(\phi_i)_i \subset L^2(\mathbb{R}^d)$ satisfies:

- For each i, there exists a neural network Φ_i with at most $C > 0$ edges such that $\varphi_i = R_\rho(\Phi_i)$.
- There exists $\tilde{C} > 0$ such that, for all $f \in \mathcal{C} \subset L^2(\mathbb{R}^d)$ with

$$\|f - \sum_{i=1}^{M} f_i \phi_i\| \leq \tilde{C} M^{-\frac{1}{\gamma^*(\mathcal{C})}}.$$

Then every $f \in \mathcal{C}$ can be approximated up to an error of ε by a neural network with only $O(\varepsilon^{-\gamma^*(\mathcal{C})})$ edges.

Recall: If a neural network stems from a fixed learning procedure **Learn**, then, for all $\gamma < \gamma^*(\mathcal{C})$, there does not exist $C > 0$ such that

$$\sup_{f \in \mathcal{C}} M(\mathbf{Learn}(\varepsilon, f)) \leq C\varepsilon^{-\gamma} \qquad \text{for all } \varepsilon > 0.$$

Proof: ...on the board!

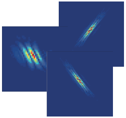
Road Map

General Approach:

(1) Determine a class of functions $\mathcal{C} \subseteq L^2(\mathbb{R}^2)$.

(2) Determine an associated representation system with the following properties:

- The elements of this system can be realized by a neural network with controlled number of edges.
- This system provides optimally sparse approximations for $\mathcal{C}$.



Road Map

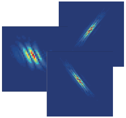
General Approach:

(1) Determine a class of functions $\mathcal{C} \subseteq L^2(\mathbb{R}^2)$.
$\rightsquigarrow$ *Cartoon-like functions!*

(2) Determine an associated representation system with the following properties:

- The elements of this system can be realized by a neural network with controlled number of edges.

- This system provides optimally sparse approximations for $\mathcal{C}$.



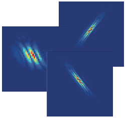
Road Map

General Approach:

(1) Determine a class of functions $\mathcal{C} \subseteq L^2(\mathbb{R}^2)$.
$\rightsquigarrow$ *Cartoon-like functions!*

(2) Determine an associated representation system with the following properties:
$\rightsquigarrow$ *Shearlets!*

- The elements of this system can be realized by a neural network with controlled number of edges.
- This system provides optimally sparse approximations for $\mathcal{C}$.



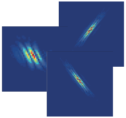
Road Map

General Approach:

(1) Determine a class of functions $\mathcal{C} \subseteq L^2(\mathbb{R}^2)$.
$\rightsquigarrow$ *Cartoon-like functions!*

(2) Determine an associated representation system with the following properties:
$\rightsquigarrow$ *Shearlets!*

- The elements of this system can be realized by a neural network with controlled number of edges.
- This system provides optimally sparse approximations for $\mathcal{C}$.
  $\rightsquigarrow$ *This has been proven!*



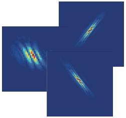
Road Map

General Approach:

(1) Determine a class of functions $\mathcal{C} \subseteq L^2(\mathbb{R}^2)$.
$\rightsquigarrow$ *Cartoon-like functions!*

(2) Determine an associated representation system with the following properties:
$\rightsquigarrow$ *Shearlets!*

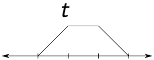
- The elements of this system can be realized by a neural network with controlled number of edges.
  $\rightsquigarrow$ *Still to be analyzed!*
- This system provides optimally sparse approximations for $\mathcal{C}$.
  $\rightsquigarrow$ *This has been proven!*



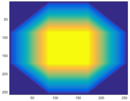
Networks which approximate ψ_s

Wavelet generators (LeCun; 1987), (Shaham, Cloninger, Coifman; 2017):

- Assume activation function $\rho(x) = \max\{x, 0\}$ (ReLUs).
- Define
$$t(x) := \rho(x) - \rho(x-1) - \rho(x-2) + \rho(x-3).$$



  $\rightsquigarrow$ *t can be constructed with a 2 layer network.*
- Observe that
$$\phi(x_1, x_2) := \rho(t(x_1) + t(x_2) - 1)$$

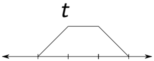


  yields a 2D bump function.
- Summing up shifted versions of ϕ yields a function ψ_s with so-calledvanishing moments.
  $\rightsquigarrow$*ψ can be realized by a 3 layer neural network.*

Networks which approximate ψ_s

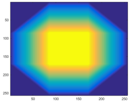
Wavelet generators (LeCun; 1987), (Shaham, Cloninger, Coifman; 2017):

- Assume activation function $\rho(x) = \max\{x, 0\}$ (ReLUs).
- Define
$$t(x) := \rho(x) - \rho(x-1) - \rho(x-2) + \rho(x-3).$$



$\rightsquigarrow$ *t can be constructed with a 2 layer network.*

- Observe that
$$\phi(x_1, x_2) := \rho(t(x_1) + t(x_2) - 1)$$



yields a 2D bump function.

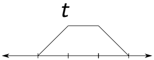
- Summing up shifted versions of ϕ yields a function ψ_s with so-calledvanishing moments.

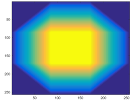
$\rightsquigarrow$$\psi$ *can be realized by a 3 layer neural network.*

This cannot yield differentiable functions ψ!

Networks which approximate ψ_s

Wavelet generators (LeCun; 1987), (Shaham, Cloninger, Coifman; 2017):

- Assume activation function $\rho(x) = \max\{x, 0\}$ (ReLUs).
- Define
$$t(x) := \rho(x) - \rho(x-1) - \rho(x-2) + \rho(x-3).$$


$\rightsquigarrow$ *t can be constructed with a 2 layer network.*
- Observe that
$$\phi(x_1, x_2) := \rho(t(x_1) + t(x_2) - 1)$$

yields a 2D bump function.
- Summing up shifted versions of ϕ yields a function ψ_s with so-calledvanishing moments.
$\rightsquigarrow$*ψ can be realized by a 3 layer neural network.*

Idea: Use a smoothed version of a ReLU.
$\rightsquigarrow$ *Leads to appropriate shearlet generators!*

Optimal Approximation

Theorem (Bölcskei, Grohs, K, and Petersen; 2017): Let ρ be an admissible smooth rectifier, and let $\varepsilon > 0$. Then there exist $C_\varepsilon > 0$ such that, for all cartoon-like functions f and $N \in \mathbb{N}$, we can construct a neural network $\Phi \in \mathcal{NN}_{3,O(N),2,\rho}$ satisfying

$$\|f - \Phi\|_{L^2(\mathbb{R}^2)} \leq C_\varepsilon N^{-1+\varepsilon}.$$

This is the optimal rate; hence the first bound is sharp!

Optimal Approximation

Theorem (Bölcskei, Grohs, K, and Petersen; 2017): Let ρ be an admissible smooth rectifier, and let $\varepsilon > 0$. Then there exist $C_\varepsilon > 0$ such that, for all cartoon-like functions f and $N \in \mathbb{N}$, we can construct a neural network $\Phi \in \mathcal{NN}_{3,O(N),2,\rho}$ satisfying

$$\|f - \Phi\|_{L^2(\mathbb{R}^2)} \leq C_\varepsilon N^{-1+\varepsilon}.$$

This is the optimal rate; hence the first bound is sharp!

Function classes which are optimal representable by shearlets, etc.
are also optimally approximated
by memory-efficient neural networks with a parallel architecture!

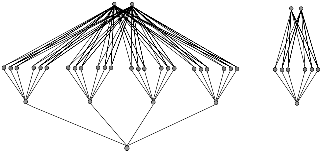
Some Numerics

Typically weights are learnt by backpropagation. This raises the following question:

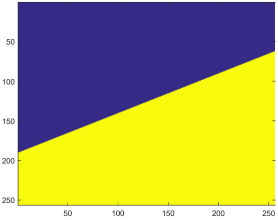
Does this lead to the optimal sparse connectivity?

Our setup:

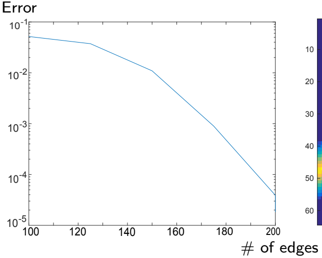
- Fixed network topology with ReLUs.
- Specific functions to learn.
- Learning through SGD.
- Analyze the learnt subnetworks.
- Analysis of the connection between approximation error and number of edges.

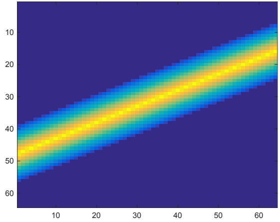


Numerical Experiments (with ReLUs & Backpropagation)



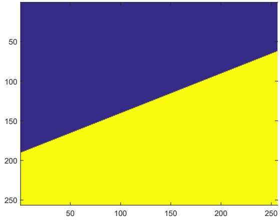
Linear Singularity



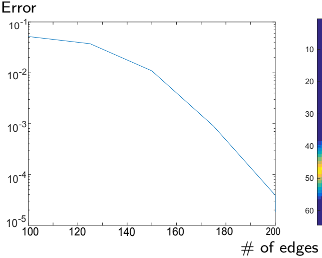


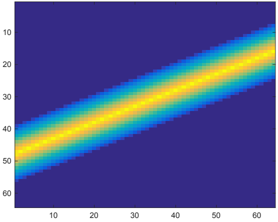
Subnetworks: Ridgelets!

Numerical Experiments (with ReLUs & Backpropagation)

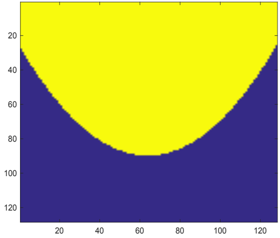


Linear Singularity

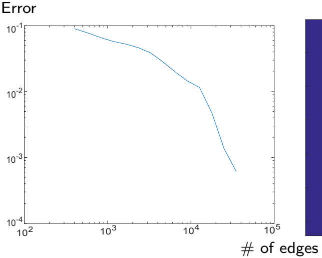


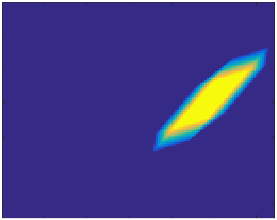


Subnetworks: Ridgelets!



Curvilinear Singularity





Subnetworks: $\approx$ Shearlets!

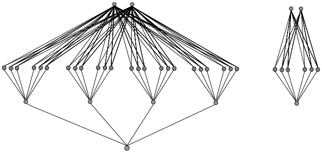
Some Numerics

Typically weights are learnt by backpropagation. This raises the following question:

Does this lead to the optimal sparse connectivity?

Our setup:

- Fixed network topology with ReLUs.
- Specific functions to learn.
- Learning through SGD.
- Analyze the learnt subnetworks.
- Analysis of the connection between approximation error and number of edges.



Sparse Connectivity and More

We now answered the following questions:

- How well can functions be approximated by neural networks with few non-zero weights?
  - Can we derive a lower bound on the necessary number of weights?
  - Can we construct neural networks which attain this bound?
- Are neural networks as good approximators as wavelets and shearlets?

Impact of Depth

Impact of Depth

Theorem (Eldan, Shamir; 2016):

"There exists a simple (approximately radial) function on $\mathbb{R}^d$, expressible by a 3-layer neural network of width polynomial in the dimension d, which cannot be arbitrarily well approximated by 2-layer networks, unless their width is exponential in d."

Remark:

- It shows that depth – even if increased by 1 – can be exponentially more valuable than width for standard feedforward neural networks.
- Key idea of proof:
  - Approximating radial function: First the squared norm function, then the univariate function acting on the norm $\rightsquigarrow$ Easy with 3 layers!
  - But approximating radial functions with 2-layers $\rightsquigarrow$ Difficult!

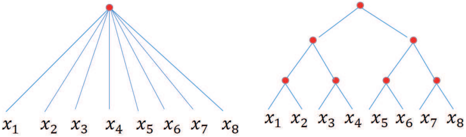
Impact of Depth

Theorem (Eldan, Shamir; 2016):

"There exists a simple (approximately radial) function on $\mathbb{R}^d$, expressible by a 3-layer neural network of width polynomial in the dimension d, which cannot be arbitrarily well approximated by 2-layer networks, unless their width is exponential in d."

Theorem (Mhaskar, Liao, Poggio; 2017):

"Deep (hierarchical) networks can approximate the class of compositional functions $f(x_1, \ldots x_n) = h_1(h_2(h_3(x_1, x_2), h_4(x_3, x_4)), \ldots)$ with the same accuracy as shallow networks but with exponentially lower number of (training) parameters."



Conclusions

What to take Home...?

Deep Learning:

- Impressive performance also for mathematical problem settings such as inverse problems and partial differential equations.
- Theoretical foundation of neural networks in large parts missing: Expressivity, Learning, Generalization, and Explainability.

Expressivity of Deep Neural Networks:

- One part of the error of statistical learning theory.
- Numerous settings can be considered such as special function classes, activation functions, ...
- Desired properties are
  - controlled complexity,
  - optimality,
  - beating the curse of dimensionality.
- Neural networks are as powerful approximators as classical systems such as wavelets, shearlets, ...

THANK YOU!

References available at:

www.math.tu-berlin.de/∼kutyniok

Code available at:

www.ShearLab.org