Memory-Optimal Deep Neural Networks

Gitta Kutyniok

(Technische Universität Berlin)

joint with

Helmut Bölcskei (ETH Zürich)

Philipp Grohs (Universität Wien)

Philipp Petersen (Technische Universität Berlin)

CoSIP Intense Course on Deep Learning
Technische Universität Berlin, November 29 – December 1, 2017

Deep Neural Networks

Deep neural networks have recently shown impressive results in a variety of real-world applications.

Some Examples...

- Image classification (ImageNet).
  $\rightsquigarrow$ *Hierarchical classification of images.*
- Games (AlphaGo).
  $\rightsquigarrow$ *Success against world class players.*
- Speech Recognition (Siri).
  $\rightsquigarrow$ *Current standard in numerous cell phones.*

Deep Neural Networks

Deep neural networks have recently shown impressive results in a variety of real-world applications.

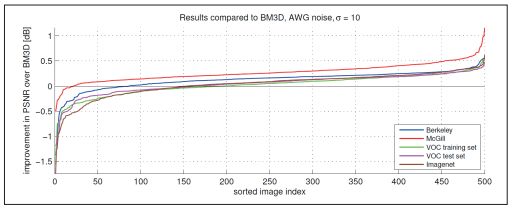
Some Examples...

- Image classification (ImageNet).
  $\rightsquigarrow$ *Hierarchical classification of images.*
- Games (AlphaGo).
  $\rightsquigarrow$ *Success against world class players.*
- Speech Recognition (Siri).
  $\rightsquigarrow$ *Current standard in numerous cell phones.*

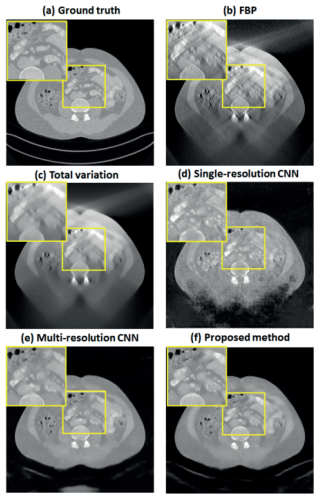
Very few theoretical results explaining their success!

Further Applications of Deep Neural Networks

Some Examples from Areas in Mathematics...

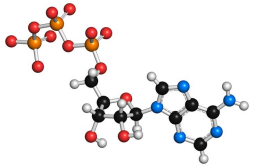


- Imaging Sciences.
  $\rightsquigarrow$ *Image denoising (Burger, Schuler, Harmeling; 2012).*



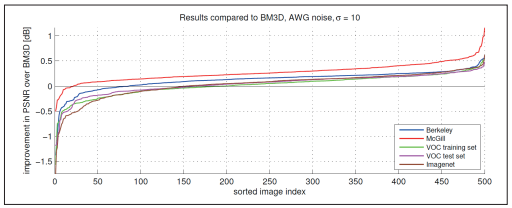
- Inverse Problems.
  $\rightsquigarrow$ *Limited-angle tomography (Gu, Ye; 2017).*

- PDE Solvers.
  $\rightsquigarrow$ *Schrödinger equation (Rupp, Tkatchenko, Müller, von Lilienfeld; 2012).*

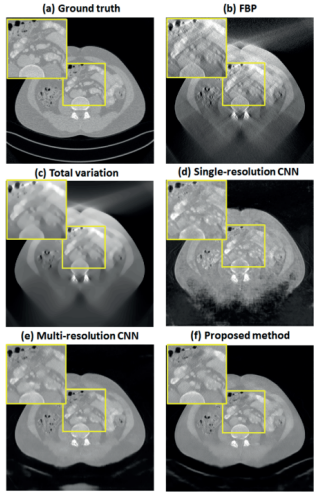


Further Applications of Deep Neural Networks

Some Examples from Areas in Mathematics...

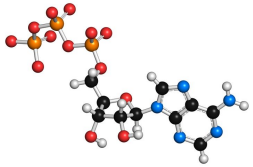


- Imaging Sciences.
  $\rightsquigarrow$ *Image denoising (Burger, Schuler, Harmeling; 2012).*



- Inverse Problems.
  $\rightsquigarrow$ *Limited-angle tomography (Gu, Ye; 2017).*

- PDE Solvers.
  $\rightsquigarrow$ *Schrödinger equation (Rupp, Tkatchenko, Müller, von Lilienfeld; 2012).*



Deep, Deep Trouble:

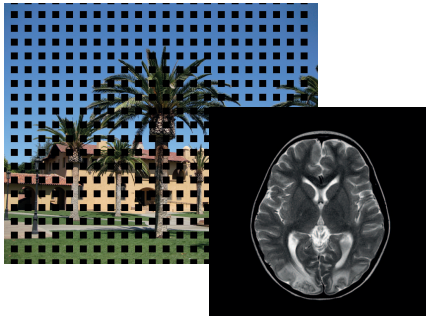
Deep Learnings Impact on Image Processing, Mathematics, and Humanity

Michael Elad (CS, Technion), SIAM News

Inverse Problems and Deep Neural Networks

Examples:

- Denoising.
- Inpainting.
- Magnetic Resonance Tomography.
- ...



Generalized Tikhonov Regularization:

Given an ill-posed inverse problem $Kx = y$, where $K : X \to Y$, an approximate solution $x^\alpha \in X$, $\alpha > 0$, can be determined by

$$\min_{\tilde{x}} \|K\tilde{x} - y\|^2 + \alpha \mathcal{P}(\tilde{x}).$$

Deep Learning Approaches:

- Pure deep learning.
- Combine the "physics" of the problem with deep learning; often by invoking the adjoint operator K^*.

Fundamental Questions

Three Problem Complexes:

- Expressibility.
  - How powerful is the network architecture?
  - Can it indeed represent the correct functions for the learning procedures?

  $\rightsquigarrow$ *Applied Harmonic Analysis, Approximation Theory, ...*
- Learning.
  - Why does the stochastic gradient descent algorithm produce anything reasonable?
  - Which local minima are acceptable?

  $\rightsquigarrow$ *Optimization, Optimal Control, ...*
- Generalization.
  - Does the resulting network generalize?
  - Which network architectures (depth, ...) are optimal for this task?

  $\rightsquigarrow$ *Statistics, Learning Theory, ...*

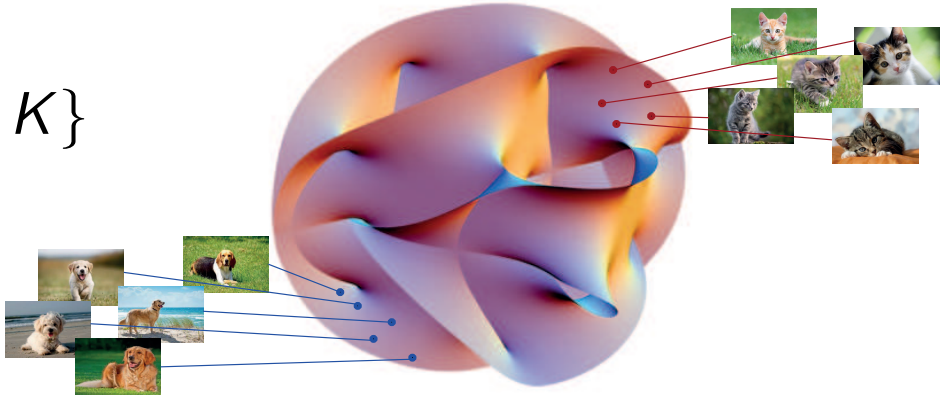
Expressibility

Main Task:

Deep neural networks approximate highly complex functions typically based on given sampling points.

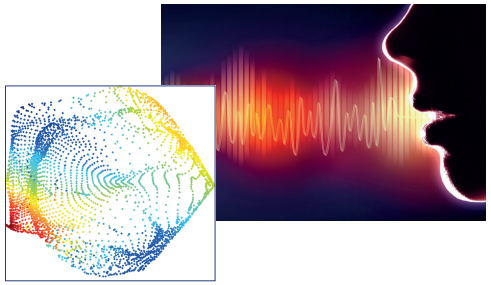
- Image Classification:

$$f : \mathcal{M} \to \{1, 2, \ldots, K\}$$



- Speech Recognition:

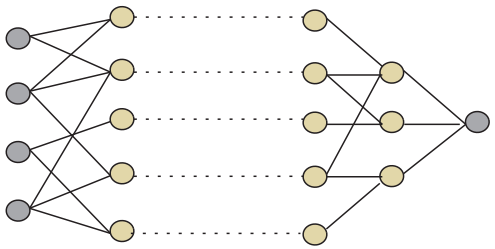
$$f : \mathbb{R}^{S_1} \to \mathbb{R}^{S_2}$$



Sparsely Connected Deep Neural Networks

Key Problem:

- Deep neural networks employed in practice often consist of hundreds of layers.
- Training and operation of such networks pose formidable computational challenge.



$\rightsquigarrow$ *Employ deep neural networks with sparse connectivity!*

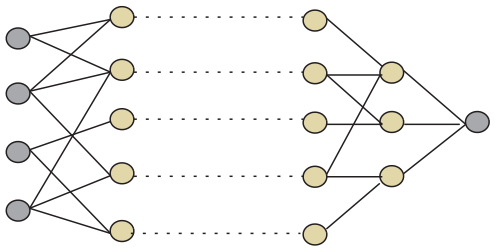
Example of Speech Recognition:

- Typically speech recognition is performed in the cloud (e.g. SIRI).
- New speech recognition systems (e.g. Android) can operate offline and are based on a sparsely connected deep neural network.

Sparsely Connected Deep Neural Networks

Key Problem:

- Deep neural networks employed in practice often consist of hundreds of layers.
- Training and operation of such networks pose formidable computational challenge.



$\rightsquigarrow$ *Employ deep neural networks with sparse connectivity!*

Example of Speech Recognition:

- Typically speech recognition is performed in the cloud (e.g. SIRI).
- New speech recognition systems (e.g. Android) can operate offline and are based on a sparsely connected deep neural network.

Key Challenge for Memory Efficient DNNs:

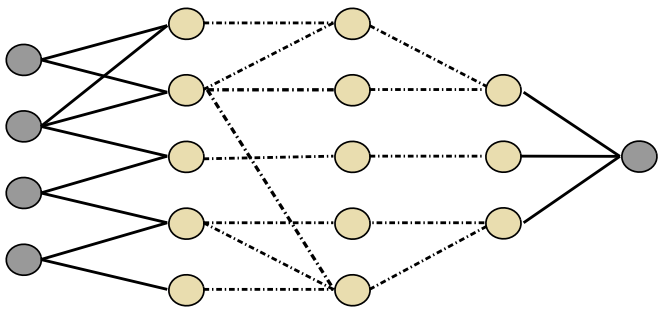
Approximation accuracy $\leftrightarrow$ *Complexity of approximating DNN in terms of sparse connectivity*

Deep Neural Networks and Approximation

Neural Networks

Definition:

Assume the following notions:



- $d \in \mathbb{N}$: Dimension of input layer.
- L: Number of layers.
- N: Number of neurons.
- $\rho : \mathbb{R} \to \mathbb{R}$: (Non-linear) function called *rectifier*.
- $W_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$, $\ell = 1, \ldots, L$: Affine linear maps.

Then $\Phi : \mathbb{R}^d \to \mathbb{R}^{N_L}$ given by

$$\Phi(x) = W_L \rho(W_{L-1} \rho(\ldots \rho(W_1(x))), \quad x \in \mathbb{R}^d,$$

is called a *(deep) neural network (DNN)*.

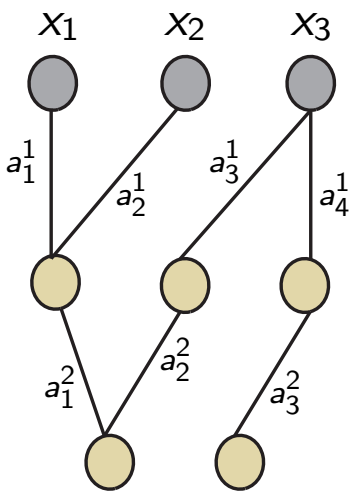
Sparse Connectivity

Remark: The affine linear map W_ℓ is defined by a matrix $A_\ell \in \mathbb{R}^{N_{\ell-1} \times N_\ell}$ and an affine part $b_\ell \in \mathbb{R}^{N_\ell}$ via

$$W_\ell(x) = A_\ell x + b_\ell.$$

$$A_1 = \begin{pmatrix} a_1^1 & a_2^1 & 0 \\ 0 & 0 & a_3^1 \\ 0 & 0 & a_4^1 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} a_1^2 & a_2^2 & 0 \\ 0 & 0 & a_3^2 \end{pmatrix}$$

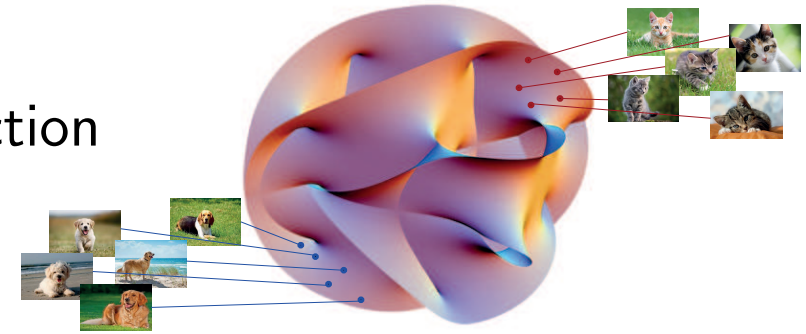


Definition: We write $\Phi \in \mathcal{NN}_{L,M,d,\rho}$, where M number of edges with non-zero weight. A DNN with small M is called *sparsely connected*.

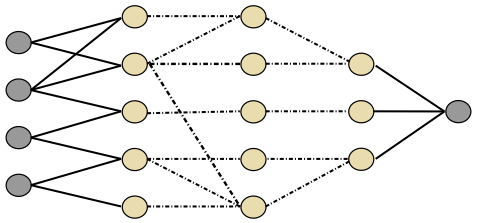
Training of Deep Neural Networks

High-Level Set Up:

- Given are (random) samples of a function such as $f : \mathcal{M} \to \{1, 2, \ldots, K\}$.



- Given an architecture of a deep neural network, i.e., a choice of d, L, $(N_\ell)_{\ell=1}^L$, and ρ.
  Sometimes selected entries of the matrices $(A_\ell)_{\ell=1}^L$, i.e., weights, are set to zero at this point.



- Learn the affine-linear functions $(W_\ell)_{\ell=1}^L = (A_\ell \cdot + b_\ell)_{\ell=1}^L$, i.e., the weights and offsets, yielding the network

$$\Phi : \mathbb{R}^d \to \mathbb{R}^{N_L}, \quad \Phi(x) = W_L \rho(W_{L-1} \rho(\ldots \rho(W_1(x))).$$

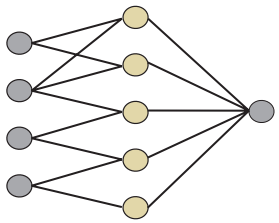
This is often done by backpropagation, a special case of gradient descent.

Goal: $\Phi \approx f$

The Start

Universal Approximation Theorem (Cybenko, 1989)(Hornik, 1991) (Pinkus, 1999):
Let $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ compact, $f : K \to \mathbb{R}$ continuous, $\rho : \mathbb{R} \to \mathbb{R}$ continuous and not a polynomial. Then, for each $\varepsilon > 0$, there exist $N \in \mathbb{N}$, $a_k, b_k \in \mathbb{R}, w_k \in \mathbb{R}^d$ such that

$$\|f - \sum_{k=1}^{N} a_k \rho(\langle w_k, \cdot \rangle - b_k)\|_\infty \leq \varepsilon.$$

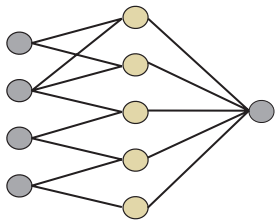


Interpretation: Every continuous function can be approximated up to an error of $\varepsilon > 0$ with a neural network with a single hidden layer and with $O(N)$ neurons.

The Start

Universal Approximation Theorem (Cybenko, 1989)(Hornik, 1991) (Pinkus, 1999):
Let $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ compact, $f : K \to \mathbb{R}$ continuous, $\rho : \mathbb{R} \to \mathbb{R}$ continuous and not a polynomial. Then, for each $\varepsilon > 0$, there exist $N \in \mathbb{N}$, $a_k, b_k \in \mathbb{R}, w_k \in \mathbb{R}^d$ such that

$$\|f - \sum_{k=1}^{N} a_k \rho(\langle w_k, \cdot \rangle - b_k)\|_\infty \leq \varepsilon.$$



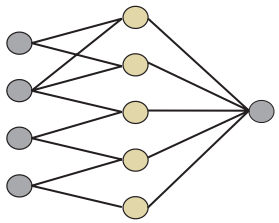
Interpretation: Every continuous function can be approximated up to an error of $\varepsilon > 0$ with a neural network with a single hidden layer and with $O(N)$ neurons.

What is the connection between ε and N?

The Start

Universal Approximation Theorem (Cybenko, 1989)(Hornik, 1991) (Pinkus, 1999):
Let $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ compact, $f : K \to \mathbb{R}$ continuous, $\rho : \mathbb{R} \to \mathbb{R}$ continuous and not a polynomial. Then, for each $\varepsilon > 0$, there exist $N \in \mathbb{N}$, $a_k, b_k \in \mathbb{R}, w_k \in \mathbb{R}^d$ such that

$$\|f - \sum_{k=1}^{N} a_k \rho(\langle w_k, \cdot \rangle - b_k)\|_\infty \leq \varepsilon.$$



Interpretation: Every continuous function can be approximated up to an error of $\varepsilon > 0$ with a neural network with a single hidden layer and with $O(N)$ neurons.

What is the connection between ε and N?

Approximation accuracy $\leftrightarrow$ Complexity of approximating DNN?

Function Approximation in a Nutshell

Goal: Given $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$ and $(\varphi_i)_{i\in I} \subseteq L^2(\mathbb{R}^d)$. Measure the suitability of $(\varphi_i)_{i\in I}$ for uniformly approximating functions from $\mathcal{C}$.

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

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

The largest $\gamma > 0$ such that

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

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

Function Approximation in a Nutshell

Goal: Given $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$ and $(\varphi_i)_{i\in I} \subseteq L^2(\mathbb{R}^d)$. Measure the suitability of $(\varphi_i)_{i\in I}$ for uniformly approximating functions from $\mathcal{C}$.

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

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

The largest $\gamma > 0$ such that

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

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

Approximation accuracy $\leftrightarrow$ *Complexity of approximating system in terms of sparsity*

Example: Wavelets

Definition (1D): Let $\phi \in L^2(\mathbb{R})$ be a scaling function and $\psi \in L^2(\mathbb{R})$ be a wavelet. Then the associated wavelet system is defined by

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

Example: Wavelets

Definition (1D): Let $\phi \in L^2(\mathbb{R})$ be a scaling function and $\psi \in L^2(\mathbb{R})$ be a wavelet. Then the associated wavelet system is defined by

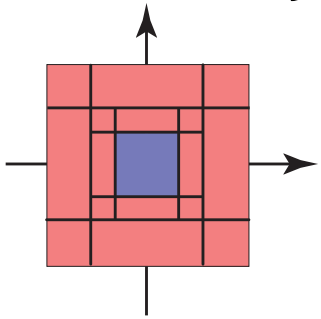
$$\{\phi(x-m) : m \in \mathbb{Z}\} \cup \{2^{j/2}\,\psi(2^j x - m) : j \geq 0, m \in \mathbb{Z}\}.$$

Definition (2D): A wavelet system is defined by

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

where

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



Example: Wavelets

Definition (1D): Let $\phi \in L^2(\mathbb{R})$ be a scaling function and $\psi \in L^2(\mathbb{R})$ be a wavelet. Then the associated wavelet system is defined by

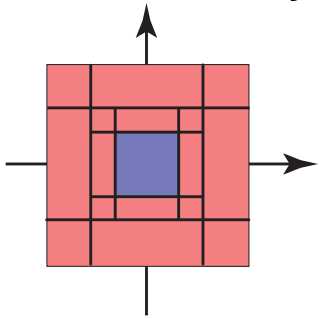
$$\{\phi(x-m) : m \in \mathbb{Z}\} \cup \{2^{j/2}\,\psi(2^j x - m) : j \geq 0, m \in \mathbb{Z}\}.$$

Definition (2D): A wavelet system is defined by

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

where

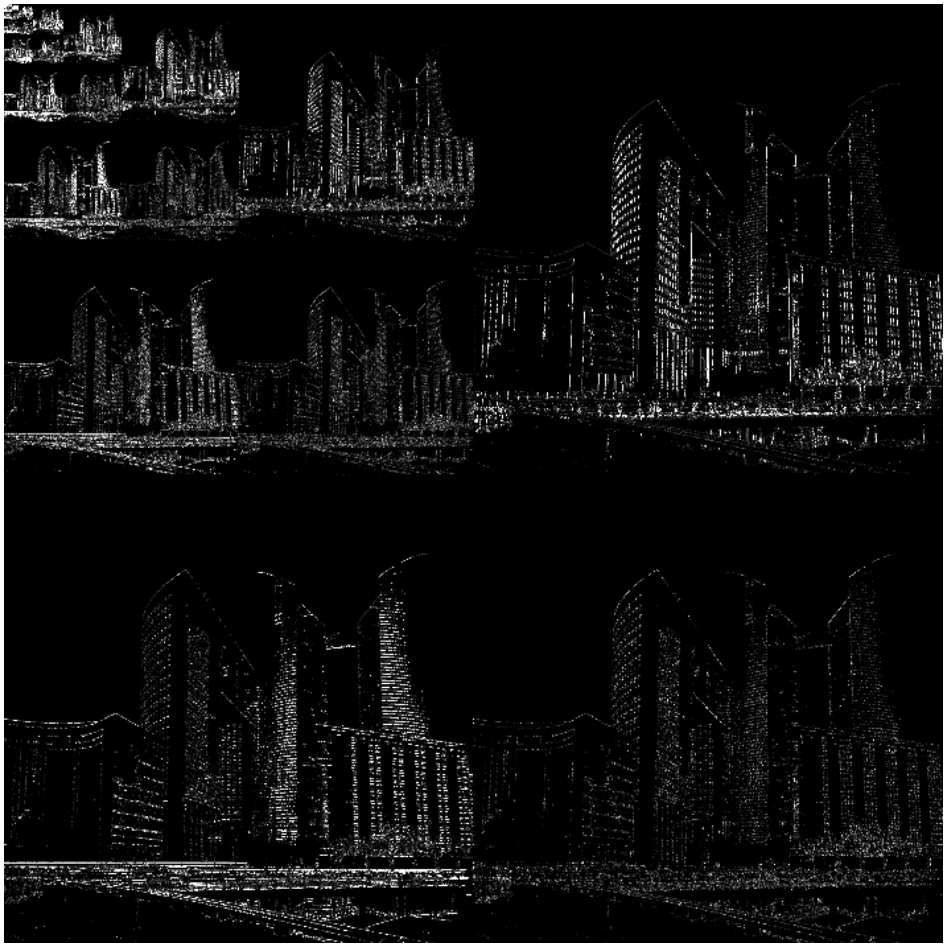
$$\phi^{(1)}(x) = \phi(x_1)\phi(x_2) \quad \text{and} \quad \begin{aligned} \psi^{(1)}(x) &= \phi(x_1)\psi(x_2), \\ \psi^{(2)}(x) &= \psi(x_1)\phi(x_2), \\ \psi^{(3)}(x) &= \psi(x_1)\psi(x_2). \end{aligned}$$



Approximation of some f with a wavelet ONB $(\psi_\lambda)_{\lambda \in \Lambda}$:

$$f_M = \sum_{\lambda \in \Lambda_M} c_\lambda \psi_\lambda, \quad \text{where } c_\lambda = \langle f, \psi_\lambda \rangle .$$

Wavelet Decomposition: JPEG2000



Wavelet Decomposition: JPEG2000

Original

25% Compression

5% Compression

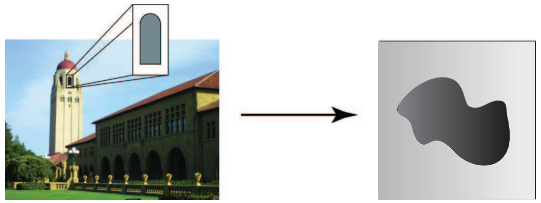
Fitting Model for Images

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



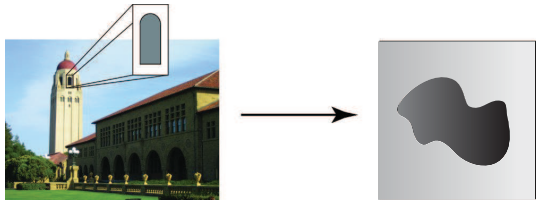
Fitting Model for Images

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



Theorem:

Given a wavelet orthonormal basis $(\psi_\lambda)_{\lambda\in\Lambda} \subseteq L^2(\mathbb{R}^2)$, the decay rate of the L^2-error of best M-term approximation of $f \in \mathcal{E}^2(\mathbb{R}^2)$ is

$$\|f - f_M\|_2 \asymp M^{-\frac{1}{2}}, \quad M \to \infty, \quad \text{where } f_M = \sum_{\lambda\in\Lambda_M} c_\lambda \psi_\lambda.$$

But this is not the optimal rate (Donoho; 2001)!

Function Approximation in a Nutshell

Goal: Given $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$ and $(\varphi_i)_{i\in I} \subseteq L^2(\mathbb{R}^d)$. Measure the suitability of $(\varphi_i)_{i\in I}$ for uniformly approximating functions from $\mathcal{C}$.

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

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

The largest $\gamma > 0$ such that

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

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

Approximation accuracy $\leftrightarrow$ *Complexity of approximating system in terms of sparsity*

Approximation with Sparse Deep Neural Networks

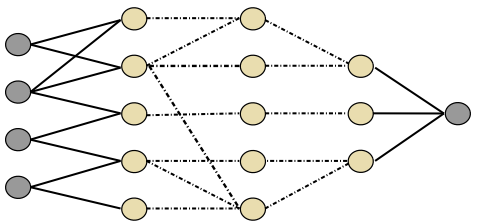
Complexity: Sparse connectivity = Few non-zero weights.

Definition: Given $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$ and fixed ρ. For each $M \in \mathbb{N}$ and $f \in \mathcal{C}$, we define

$$\Gamma_M(f) := \inf_{\Phi \in \mathcal{NN}_{\infty,M,d,\rho}} \|f - \Phi\|_{L^2(\mathbb{R}^d)}.$$

The largest $\gamma > 0$ such that

$$\sup_{f \in \mathcal{C}} \Gamma_M(f) = O(M^{-\gamma}) \qquad \text{as } M \to \infty$$



then determines the *optimal (sparse) approximation rate* of $\mathcal{C}$ by $\mathcal{NN}_{\infty,\infty,d,\rho}$.

Approximation with Sparse Deep Neural Networks

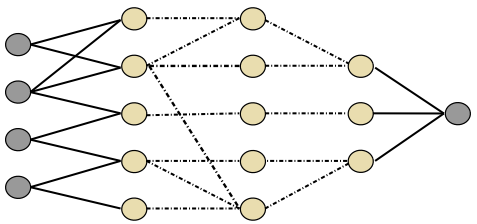
Complexity: Sparse connectivity = Few non-zero weights.

Definition: Given $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$ and fixed ρ. For each $M \in \mathbb{N}$ and $f \in \mathcal{C}$, we define

$$\Gamma_M(f) := \inf_{\Phi \in \mathcal{NN}_{\infty,M,d,\rho}} \|f - \Phi\|_{L^2(\mathbb{R}^d)}.$$

The largest $\gamma > 0$ such that

$$\sup_{f \in \mathcal{C}} \Gamma_M(f) = O(M^{-\gamma}) \qquad \text{as } M \to \infty$$



then determines the *optimal (sparse) approximation rate* of $\mathcal{C}$ by $\mathcal{NN}_{\infty,\infty,d,\rho}$.

Approximation accuracy $\leftrightarrow$ Complexity of approximating DNN in terms of sparse connectivity

Non-Exhaustive List of Previous 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).
- Lower bounds on the sparsity in terms of number of neurons (Schmitt; 1999).
- Approximation using specific rectifiers (Cybenko; 1989).
- Approximation of specific function classes (Mhaskar and Micchelli; 1995), (Mhaskar; 1996).

Non-Exhaustive List of Previous 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).
- Lower bounds on the sparsity in terms of number of neurons (Schmitt; 1999).
- 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).
- Approximation of functions together and their derivatives (Nguyen-Thien and Tran-Cong; 1999).
- 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).

Goal for Today

Challenges and Contributions:

(1) How many edges do we need for a certain accuracy?
$\rightsquigarrow$ *Conceptual lower bound on the number of edges of the DNN, which each learning algorithm has to obey!*

(2) Are there DNNs which are optimally sparsely connected?
$\rightsquigarrow$ *Sharpness of the bound by explicit construction of optimal DNNs!*

(3) But what happens in practise using ReLUs and backpropagation?
$\rightsquigarrow$ *Success of certain network topologies to reach optimal bound!*

A Lower Bound on Sparse Connectivity

Rate Distortion Theory

To use information theoretic arguments, we require the following notions from information theory:

Definition:
Let $d \in \mathbb{N}$. Then, for each $\ell \in \mathbb{N}$, we let

$$\mathfrak{E}^\ell := \{E : L^2(\mathbb{R}^d) \to \{0,1\}^\ell\}$$

denote the set of *binary encoders with length* ℓ and

$$\mathfrak{D}^\ell := \{D : \{0,1\}^\ell \to L^2(\mathbb{R}^d)\}$$

the set of *binary decoders of length* ℓ.

Rate Distortion Theory

Definition (continued):

For arbitrary $\varepsilon > 0$ and $\mathcal{C} \subset L^2(\mathbb{R}^d)$, the *minimax code length* $L(\varepsilon, \mathcal{C})$ is given by

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

and the *optimal exponent* $\gamma^*(\mathcal{C})$ is defined by

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

Interpretation:

The optimal exponent $\gamma^*(\mathcal{C})$ describes the dependence of the code length on the required approximation quality.

$\gamma^*(\mathcal{C})$ is a measure of the complexity of the function class!

Rate Distortion Theory

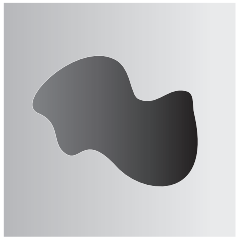
Examples for optimal exponents:

- Let $\mathcal{C} \subset B^s_{p,q}(\mathbb{R}^d)$ be bounded. Then we have

$$\gamma^*(\mathcal{C}) = \frac{d}{s}.$$

- (Donoho; 2001) Let $\mathcal{C} = \mathcal{E}^2(\mathbb{R}^2)$, the class of cartoon-like functions. Then we have

$$\gamma^*(\mathcal{C}) \geq 1.$$



A Fundamental Lower Bound

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

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

satisfy that, for each $f \in \mathcal{C}$ and $0 < \varepsilon < 1$:

(1) Each weight of $\textbf{Learn}(\varepsilon, f)$ can be encoded with $< -c \log_2(\varepsilon)$ bits,

(2) and

$$\sup_{f \in \mathcal{C}} \|f - \textbf{Learn}(\varepsilon, f)\|_{L^2(\mathbb{R}^d)} \leq \varepsilon.$$

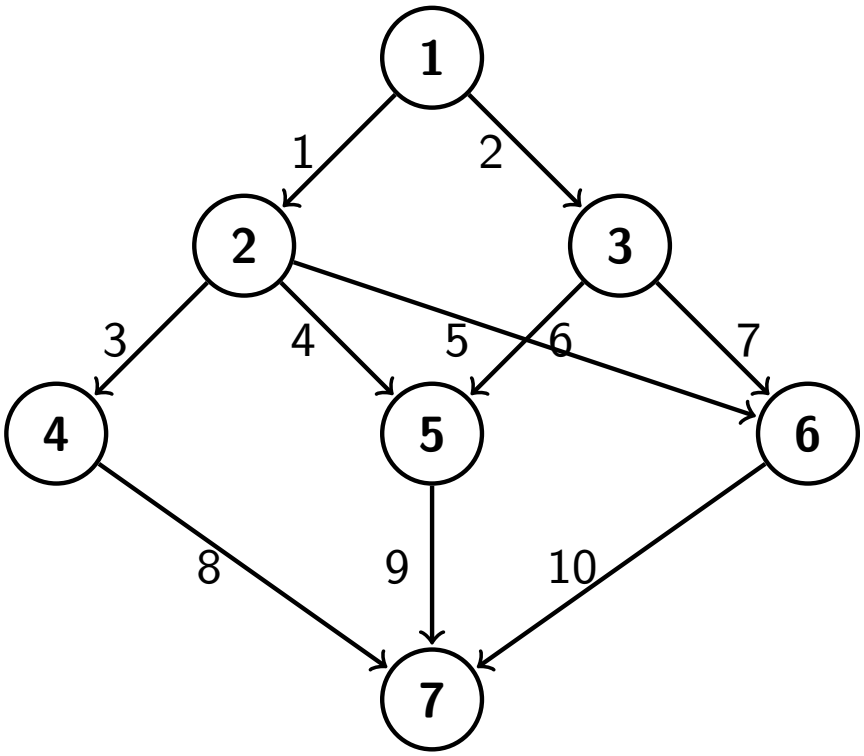
Then, for all $\gamma < \gamma^*(\mathcal{C})$,

$$\varepsilon^\gamma \sup_{f \in \mathcal{C}} \mathcal{M}(\textbf{Learn}(\varepsilon, f)) \to \infty \qquad \text{as } \varepsilon \to 0,$$

where $\mathcal{M}(\textbf{Learn}(\varepsilon, f))$ denotes the number of non-zero weights in $\textbf{Learn}(\varepsilon, f)$.

Idea of Proof

Every network with M edges can be encoded in a bit string of length $O(M)$.



Encode:

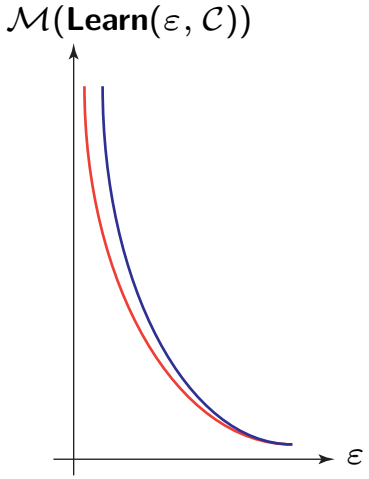
- # layers,
- # neurons in each layer,
- for each neuron in chronological order # the number of children,
- for each neuron in chronological order the indices of children,
- in chronological order the weights of edges.

A Fundamental Lower Bound

Some implications and remarks:

- 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

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



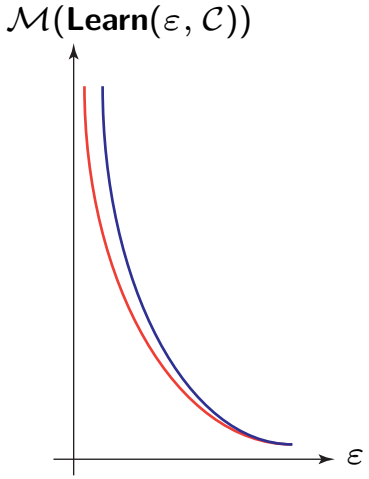
$\Rightarrow$ There exists a fundamental lower bound on the number of edges.

A Fundamental Lower Bound

Some implications and remarks:

- 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

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



$\Rightarrow$ There exists a fundamental lower bound on the number of edges.

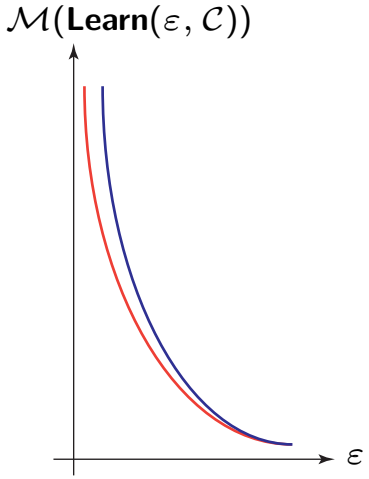
What happens for $\gamma = \gamma^*(\mathcal{C})$?

A Fundamental Lower Bound

Some implications and remarks:

- 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

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



$\Rightarrow$ There exists a fundamental lower bound on the number of edges.

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

- This bound is in terms of the edges, hence the sparsity of the connectivity, not the neurons. However, the number of neurons is always bounded up to a uniform constant by the number of edges.

Optimally Sparse Deep Neural Networks

DNNs and Representation Systems, I

Question:

Can we exploit approximation results with representation systems?

DNNs and Representation Systems, I

Question:

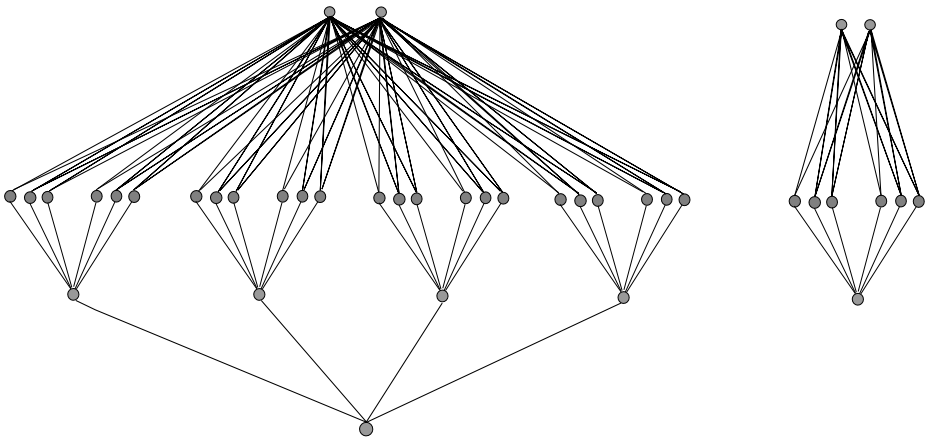
Can we exploit approximation results with representation systems?

Observation: Assume a system $(\varphi_i)_{i\in I} \subset L^2(\mathbb{R}^d)$ satisfies:

- For each $i \in I$, there exists a neural network Φ_i with at most $C > 0$ edges such that $\varphi_i = \Phi_i$.

Then we can construct a network Φ with $O(M)$ edges with

$$\Phi = \sum_{i\in I_M} c_i \varphi_i, \quad \text{if } |I_M| = M.$$



DNNs and Representation Systems, II

Corollary: Assume a system $(\varphi_i)_{i\in I} \subset L^2(\mathbb{R}^d)$ satisfies:

- For each $i \in I$, there exists a neural network Φ_i with at most $C > 0$ edges such that $\varphi_i = \Phi_i$.
- There exists $\tilde{C} > 0$ such that, for all $f \in \mathcal{C} \subset L^2(\mathbb{R}^d)$, there exists $I_M \subset I$ with
$$\|f - \sum_{i\in I_M} c_i\varphi_i\| \leq \tilde{C} M^{-1/\gamma^*(\mathcal{C})}.$$

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.

Proof:

- There exists a network Φ with $O(M)$ edges with $\Phi = \sum_{i\in I_M} c_i\varphi_i$.
- Set $\varepsilon = \tilde{C} M^{-1/\gamma^*(\mathcal{C})}$ and solve for the number of edges M, yielding
$$M = O(\varepsilon^{-\gamma^*(\mathcal{C})}).$$

DNNs and Representation Systems, II

Corollary: Assume a system $(\varphi_i)_{i\in I} \subset L^2(\mathbb{R}^d)$ satisfies:

- For each $i \in I$, there exists a neural network Φ_i with at most $C > 0$ edges such that $\varphi_i = \Phi_i$.
- There exists $\tilde{C} > 0$ such that, for all $f \in \mathcal{C} \subset L^2(\mathbb{R}^d)$, there exists $I_M \subset I$ with
$$\|f - \sum_{i\in I_M} c_i\varphi_i\| \leq \tilde{C} M^{-1/\gamma^*(\mathcal{C})}.$$

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

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

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 $\mathcal{C}$.

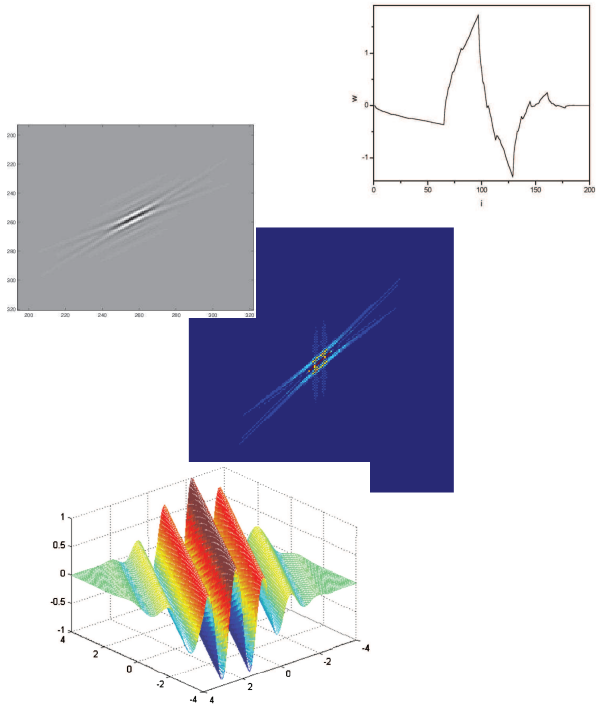
DNNs have as much approximation power as most classical systems!

Applied Harmonic Analysis

Representation systems designed by Applied Harmonic Analysis concepts have established themselves as a standard tool in applied mathematics, computer science, and engineering.

Examples:

- Wavelets.
- Ridgelets.
- Curvelets.
- Shearlets.
- ...



Key Property:

Fast Algorithms combined with Sparse Approximation Properties!

Affine Transforms

Building Principle:

Many systems from applied harmonic analysis such as

- wavelets,
- ridgelets,
- shearlets,

constitute affine systems:

$$\{|\det A|^{d/2}\psi(A\cdot -t) : A \in G \subseteq GL(d),\ t \in \mathbb{Z}^d\}, \quad \psi \in L^2(\mathbb{R}^d).$$

Affine Transforms

Building Principle:

Many systems from applied harmonic analysis such as

- wavelets,
- ridgelets,
- shearlets,

constitute affine systems:

$$\{|\det A|^{d/2}\psi(A\cdot -t) : A \in G \subseteq GL(d),\ t \in \mathbb{Z}^d\}, \quad \psi \in L^2(\mathbb{R}^d).$$

Realization by Neural Networks:

The following conditions are equivalent:

(i) $|\det A|^{d/2}\psi(A\cdot -t)$ can be realized by a neural network Φ_1.

(ii) ψ can be realized by a neural network Φ_2.

Also, Φ_1 and Φ_2 have the same number of edges up to a constant factor.

What are Shearlets?

Key Ideas of the Shearlet Construction

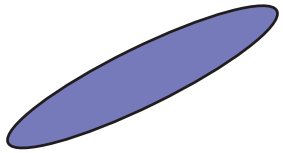
Wavelet versus Shearlet Approximation:

Key Ideas of the Shearlet Construction

Wavelet versus Shearlet Approximation:

Parabolic scaling ('width $\approx$ length^2'):

$$A_{2^j} = \begin{pmatrix} 2^j & 0 \\ 0 & 2^{j/2} \end{pmatrix}, \quad j \in \mathbb{Z}.$$

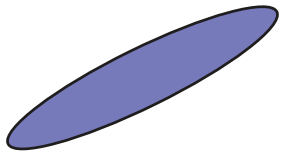


Key Ideas of the Shearlet Construction

Wavelet versus Shearlet Approximation:

Parabolic scaling ('width $\approx$ length^2'):

$$A_{2^j} = \begin{pmatrix} 2^j & 0 \\ 0 & 2^{j/2} \end{pmatrix}, \quad j \in \mathbb{Z}.$$



Orientation via shearing:

$$S_k = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}, \quad k \in \mathbb{Z}.$$

Advantage:

- Shearing leaves the digital grid $\mathbb{Z}^2$ invariant.
- Uniform theory for the continuum and digital situation.

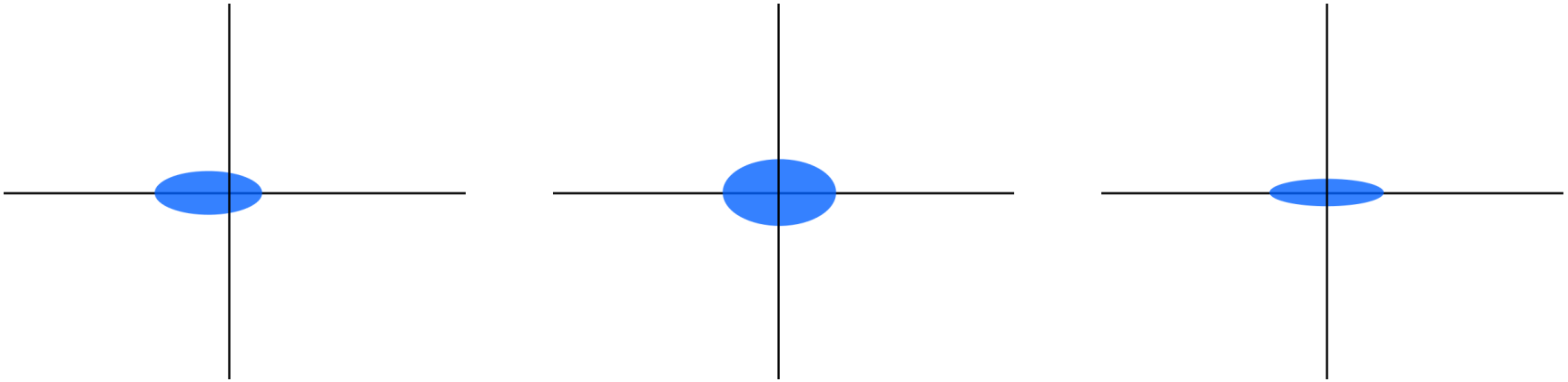
Shearlet Systems

Affine systems:

$$\{|\det M|^{1/2}\psi(M\,\cdot\,-m) : M \in G \subseteq GL_2,\ m \in \mathbb{Z}^2\}.$$

Definition (K, Labate; 2006):
For $\psi \in L^2(\mathbb{R}^2)$, the associated shearlet system is defined by

$$\{2^{\frac{3j}{4}}\psi(S_k A_{2^j}\cdot - m)\ :\ j,k \in \mathbb{Z}, m \in \mathbb{Z}^2\}.$$



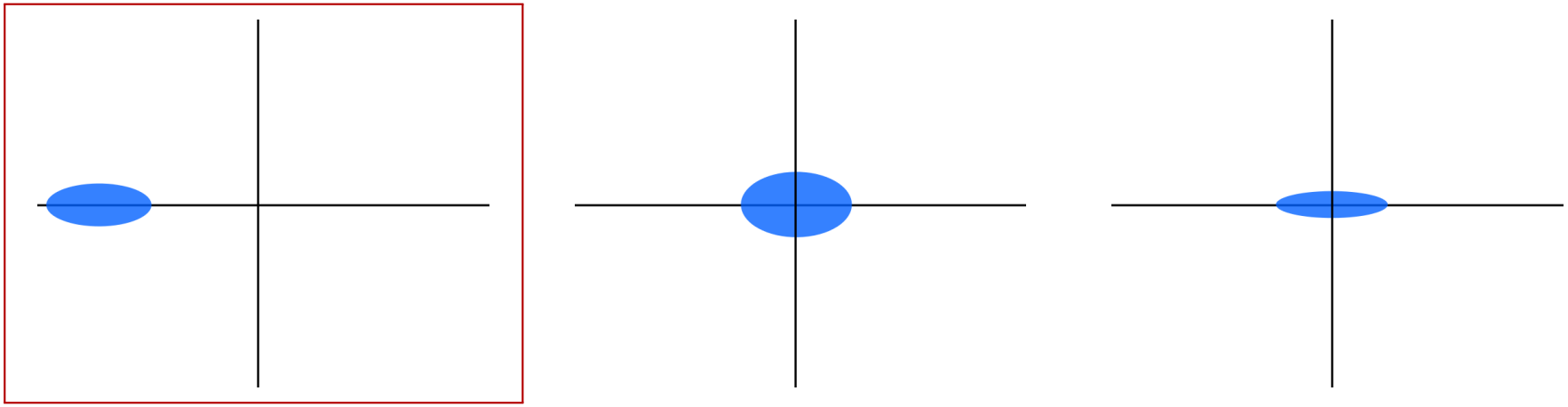
Shearlet Systems

Affine systems:

$$\{|\det M|^{1/2}\psi(M\,\cdot\,-m) : M \in G \subseteq GL_2,\ m \in \mathbb{Z}^2\}.$$

Definition (K, Labate; 2006):
For $\psi \in L^2(\mathbb{R}^2)$, the associated shearlet system is defined by

$$\{2^{\frac{3j}{4}}\psi(S_k A_{2^j}\cdot - m)\ :\ j,k \in \mathbb{Z}, m \in \mathbb{Z}^2\}.$$



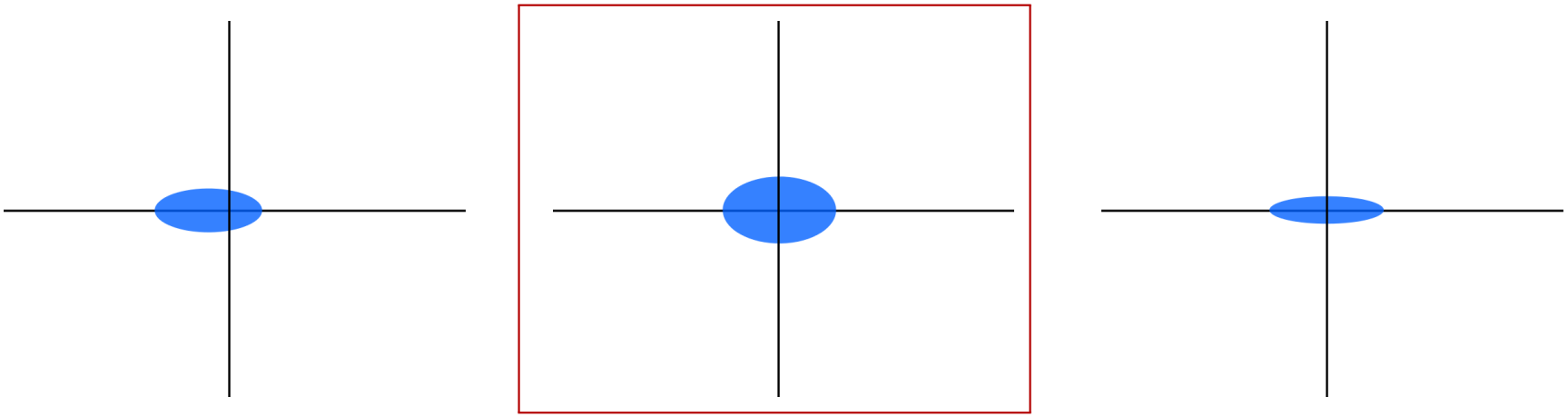
Shearlet Systems

Affine systems:

$$\{|\det M|^{1/2}\psi(M\,\cdot\,-m) : M \in G \subseteq GL_2,\ m \in \mathbb{Z}^2\}.$$

Definition (K, Labate; 2006):
For $\psi \in L^2(\mathbb{R}^2)$, the associated shearlet system is defined by

$$\{2^{\frac{3j}{4}}\psi(S_k A_{2^j}\cdot - m)\ :\ j,k \in \mathbb{Z}, m \in \mathbb{Z}^2\}.$$



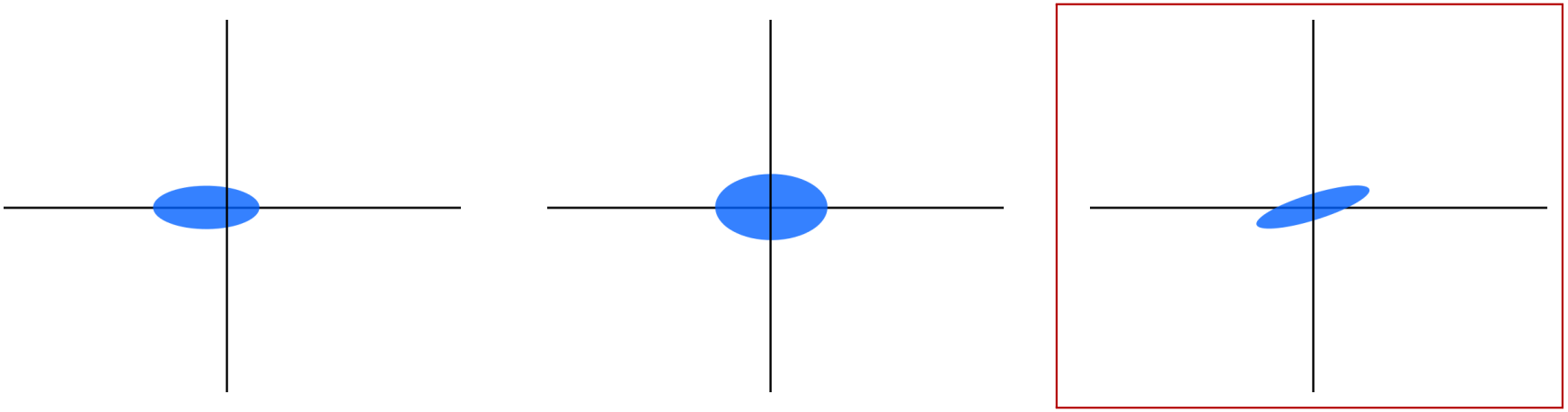
Shearlet Systems

Affine systems:

$$\{|\det M|^{1/2}\psi(M\,\cdot\,-m) : M \in G \subseteq GL_2,\ m \in \mathbb{Z}^2\}.$$

Definition (K, Labate; 2006):
For $\psi \in L^2(\mathbb{R}^2)$, the associated shearlet system is defined by

$$\{2^{\frac{3j}{4}}\psi(S_k A_{2^j}\cdot - m)\ :\ j,k \in \mathbb{Z}, m \in \mathbb{Z}^2\}.$$



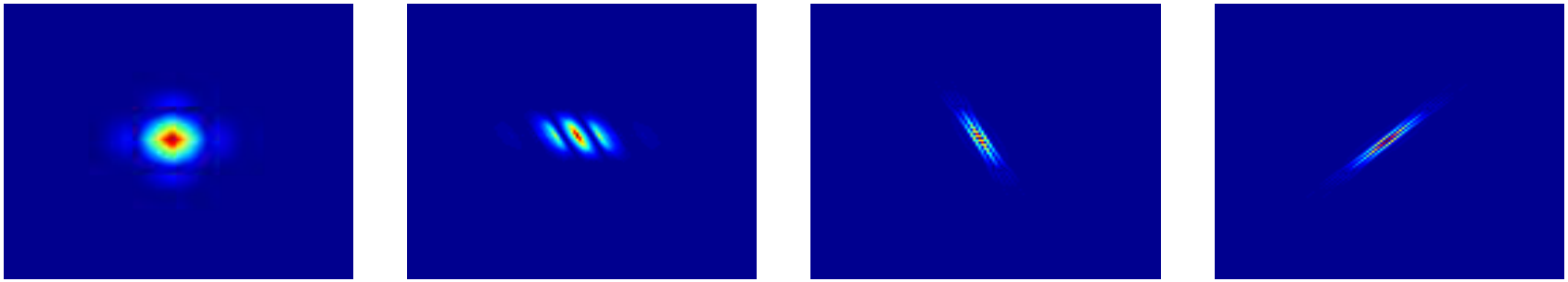
Shearlet Systems

Affine systems:

$$\{|\det M|^{1/2}\psi(M\,\cdot\,-m) : M \in G \subseteq GL_2,\ m \in \mathbb{Z}^2\}.$$

Definition (K, Labate; 2006):
For $\psi \in L^2(\mathbb{R}^2)$, the associated shearlet system is defined by

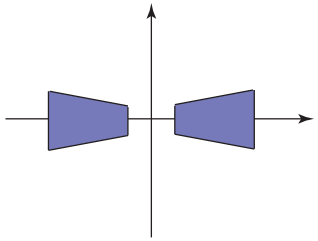
$$\{2^{\frac{3j}{4}}\psi(S_k A_{2^j}\,\cdot\,-m)\ :\ j,k \in \mathbb{Z}, m \in \mathbb{Z}^2\}.$$



Example of Classical (Band-Limited) Shearlet

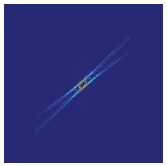
Let $\psi \in L^2(\mathbb{R}^2)$ be defined by

$$\hat{\psi}(\xi) = \hat{\psi}(\xi_1, \xi_2) = \hat{\psi}_1(\xi_1)\, \hat{\psi}_2(\tfrac{\xi_2}{\xi_1}),$$



where

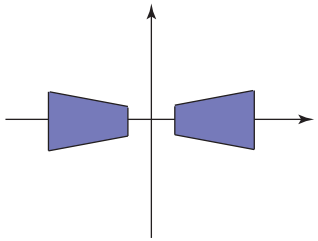
- ψ_1 wavelet, $\mathrm{supp}(\hat{\psi}_1) \subseteq [-2, -\frac{1}{2}] \cup [\frac{1}{2}, 2]$ and $\hat{\psi}_1 \in C^\infty(\mathbb{R})$.
- $\mathrm{supp}(\hat{\psi}_2) \subseteq [-1, 1]$ and $\hat{\psi}_2 \in C^\infty(\mathbb{R})$.



Example of Classical (Band-Limited) Shearlet

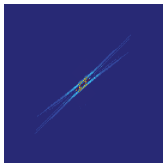
Let $\psi \in L^2(\mathbb{R}^2)$ be defined by

$$\hat{\psi}(\xi) = \hat{\psi}(\xi_1, \xi_2) = \hat{\psi}_1(\xi_1)\, \hat{\psi}_2(\tfrac{\xi_2}{\xi_1}),$$

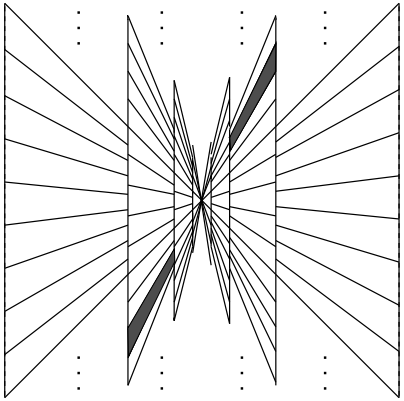


where

- ψ_1 wavelet, $\mathrm{supp}(\hat{\psi}_1) \subseteq [-2, -\frac{1}{2}] \cup [\frac{1}{2}, 2]$ and $\hat{\psi}_1 \in C^\infty(\mathbb{R})$.
- $\mathrm{supp}(\hat{\psi}_2) \subseteq [-1, 1]$ and $\hat{\psi}_2 \in C^\infty(\mathbb{R})$.



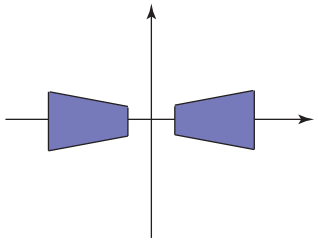
Induced tiling of Fourier domain:



Example of Classical (Band-Limited) Shearlet

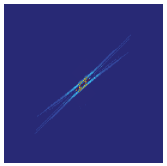
Let $\psi \in L^2(\mathbb{R}^2)$ be defined by

$$\hat{\psi}(\xi) = \hat{\psi}(\xi_1, \xi_2) = \hat{\psi}_1(\xi_1)\, \hat{\psi}_2(\tfrac{\xi_2}{\xi_1}),$$

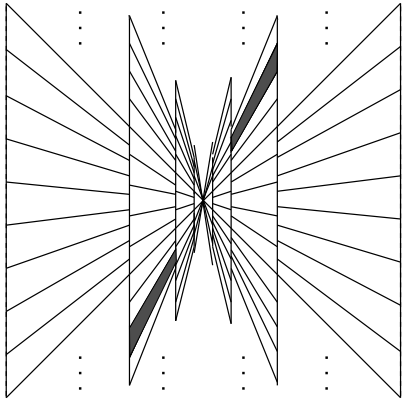


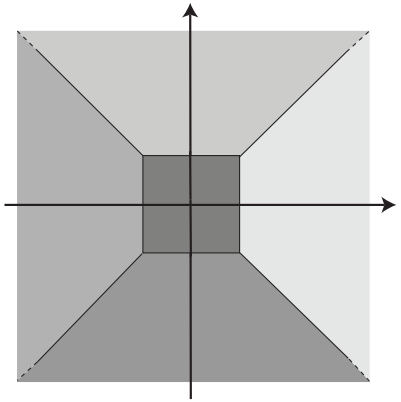
where

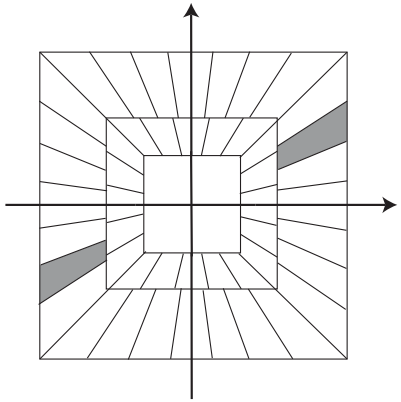
- ψ_1 wavelet, $\mathrm{supp}(\hat{\psi}_1) \subseteq [-2, -\frac{1}{2}] \cup [\frac{1}{2}, 2]$ and $\hat{\psi}_1 \in C^\infty(\mathbb{R})$.
- $\mathrm{supp}(\hat{\psi}_2) \subseteq [-1, 1]$ and $\hat{\psi}_2 \in C^\infty(\mathbb{R})$.



Induced tiling of Fourier domain:





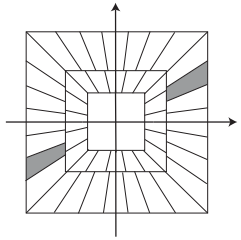


(Cone-adapted) Discrete Shearlet Systems

Definition (K, Labate; 2006):

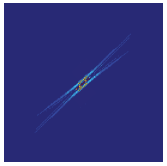
The (cone-adapted) discrete shearlet system $\mathcal{SH}(c;\phi,\psi,\tilde{\psi})$, $c>0$, generated by $\phi \in L^2(\mathbb{R}^2)$ and $\psi,\tilde{\psi} \in L^2(\mathbb{R}^2)$ is the union of

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



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

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

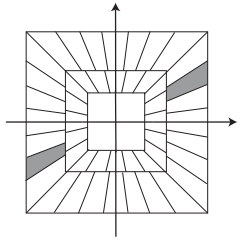


(Cone-adapted) Discrete Shearlet Systems

Definition (K, Labate; 2006):

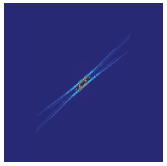
The (cone-adapted) discrete shearlet system $\mathcal{SH}(c;\phi,\psi,\tilde{\psi})$, $c>0$, generated by $\phi \in L^2(\mathbb{R}^2)$ and $\psi,\tilde{\psi} \in L^2(\mathbb{R}^2)$ is the union of

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



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

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



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

For $\psi,\tilde{\psi}$ classical shearlets, $\mathcal{SH}(1;\phi,\psi,\tilde{\psi})$ is a Parseval frame for $L^2(\mathbb{R}^2)$:

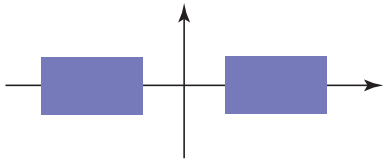
$$A\|f\|_2^2 \leq \sum_{\sigma \in \mathcal{SH}(1;\phi,\psi,\tilde{\psi})} |\langle f,\sigma\rangle|^2 \leq B\|f\|_2^2 \quad \text{for all } f \in L^2(\mathbb{R}^2)$$

holds for $A = B = 1$.

Compactly Supported Shearlets

Theorem (Kittipoom, K, Lim; 2012):

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 there exists c_0 such that $\mathcal{SH}(c; \phi, \psi, \tilde{\psi})$ forms a shearlet frame with controllable frame bounds for all $c \leq c_0$.

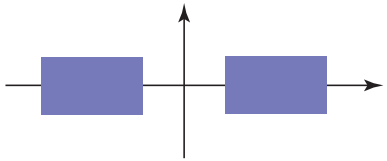


Remark: Exemplary class with $B/A \approx 4$.

Compactly Supported Shearlets

Theorem (Kittipoom, K, Lim; 2012):

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 there exists c_0 such that $\mathcal{SH}(c; \phi, \psi, \tilde{\psi})$ forms a shearlet frame with controllable frame bounds for all $c \leq c_0$.

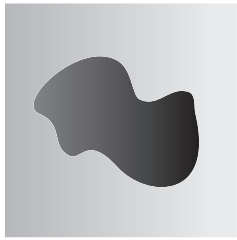


Remark: Exemplary class with $B/A \approx 4$.

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.,

$$\|f - f_N\|_2^2 \leq C \cdot N^{-2} \cdot (\log N)^3, \quad N \to \infty.$$



Recent Approaches to Fast Shearlet Transforms

www.ShearLab.org:

- Separable Shearlet Transform *(Lim; 2009)*
- Digital Shearlet Transform *(K, Shahram, Zhuang; 2011)*
- 2D&3D (parallelized) Shearlet Transform *(K, Lim, Reisenhofer; 2014)*

Additional Code:

- Filter-based implementation *(Easley, Labate, Lim; 2009)*
- Fast Finite Shearlet Transform *(Häuser, Steidl; 2014)*
- Shearlet Toolbox 2D&3D *(Easley, Labate, Lim, Negy; 2014)*

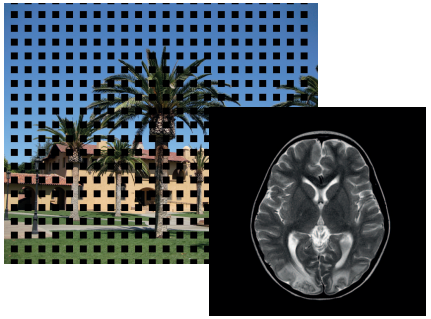
Theoretical Approaches:

- Adaptive Directional Subdivision Schemes *(K, Sauer; 2009)*
- Shearlet Unitary Extension Principle *(Han, K, Shen; 2011)*
- Gabor Shearlets *(Bodmann, K, Zhuang; 2013)*

Application to Inverse Problems

Examples:

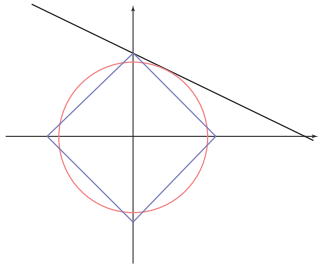
- Denoising.
- Feature Extraction.
- Inpainting.
- Magnetic Resonance Tomography.
- ...



Sparse Regularization:

Given an ill-posed inverse problem $Kx = y$, where $K : X \to Y$ and x is known to be sparsely representable by a shearlet frame $(\sigma_\eta)_\eta$, an approximate solution $x^\alpha \in X$, $\alpha > 0$, can be determined by

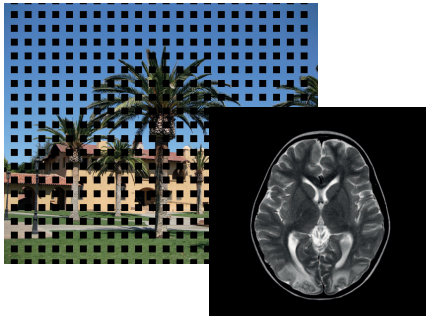
$$\min_{\tilde{x}} \|K\tilde{x} - y\|^2 + \alpha \|(\langle \tilde{x}, \sigma_\eta \rangle)_\eta\|_1.$$



Application to Inverse Problems

Examples:

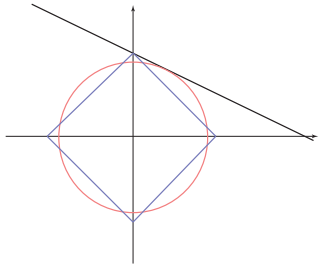
- Denoising.
- Feature Extraction.
- Inpainting.
- Magnetic Resonance Tomography.
- ...



Sparse Regularization:

Given an ill-posed inverse problem $Kx = y$, where $K : X \to Y$ and x is known to be sparsely representable by a shearlet frame $(\sigma_\eta)_\eta$, an approximate solution $x^\alpha \in X$, $\alpha > 0$, can be determined by

$$\min_{\tilde{x}} \|K\tilde{x} - y\|^2 + \alpha \|(\langle \tilde{x}, \sigma_\eta \rangle)_\eta\|_1.$$

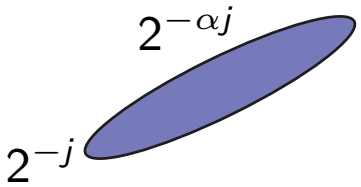


Extension to α-Shearlets

Main Idea:

- Introduction of a parameter $\alpha \in [0,1]$ to measure the amount of anisotropy.
- For $j \in \mathbb{Z}$, define

$$A_{\alpha,j} = \begin{pmatrix} 2^j & 0 \\ 0 & 2^{\alpha j} \end{pmatrix}.$$



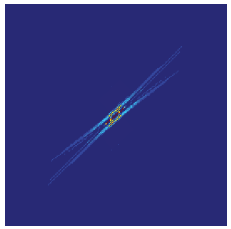
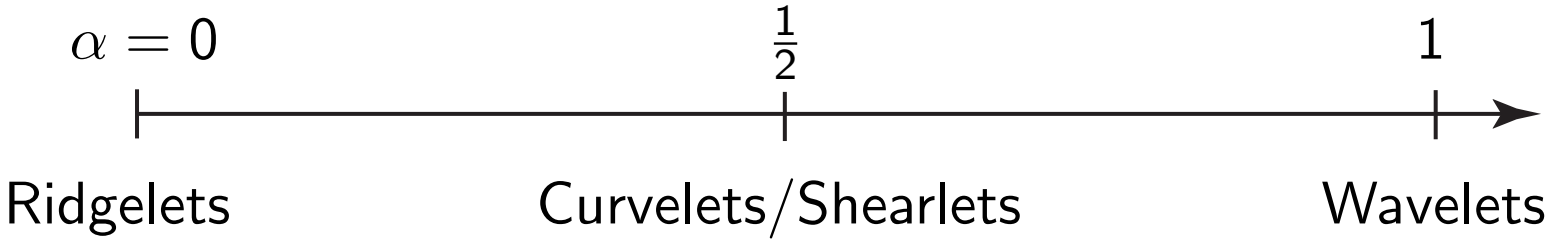


Illustration:



α-Shearlets

Definition (Grohs, Keiper, K, and Schäfer; 2016)(Voigtlaender; 2017):
For $c \in \mathbb{R}^+$ and $\alpha \in [0,1]$, the *cone-adapted α-shearlet system* $\mathcal{SH}_\alpha(\phi, \psi, \tilde{\psi}, c)$ generated by $\phi, \psi, \tilde{\psi} \in L^2(\mathbb{R}^2)$ is defined by

$$\mathcal{SH}_\alpha(\phi, \psi, \tilde{\psi}, c) := \Phi(\phi, c, \alpha) \cup \Psi(\psi, c, \alpha) \cup \tilde{\Psi}(\tilde{\psi}, c, \alpha),$$

where

$$\begin{aligned}
\Phi(\phi, c, \alpha) &:= \{\phi(\cdot - m) : m \in c\mathbb{Z}^2\}, \\
\Psi(\psi, c, \alpha) &:= \{2^{j(1+\alpha)/2}\psi(S_k A_{\alpha,j} \cdot - m) : j \geq 0, |k| \leq \lceil 2^{j(1-\alpha)} \rceil, \\
&\qquad\qquad m \in c\mathbb{Z}^2, k \in \mathbb{Z}^2\}, \\
\tilde{\Psi}(\tilde{\psi}, c, \alpha) &:= \{2^{j(1+\alpha)/2}\widetilde{\psi}(S_k^T \tilde{A}_{\alpha,j} \cdot - m) : j \geq 0, |k| \leq \lceil 2^{j(1-\alpha)} \rceil, \\
&\qquad\qquad m \in c\mathbb{Z}^2, k \in \mathbb{Z}^2\}.
\end{aligned}$$

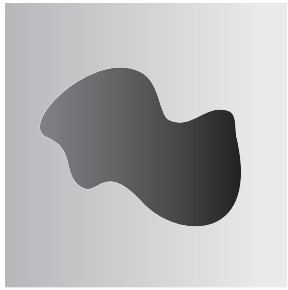
Cartoon-Like Functions

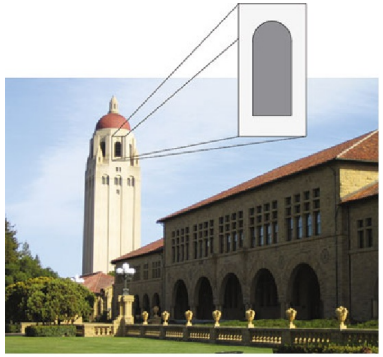
Definition (Donoho; 2001)(Grohs, Keiper, K, and Schäfer; 2016):
Let $\alpha \in [\frac{1}{2}, 1]$ and $\nu > 0$. We then define the class of *α-cartoon-like functions* by

$$\mathcal{E}^{\frac{1}{\alpha}}(\mathbb{R}^2) = \{f \in L^2(\mathbb{R}^2) : f = f_1 + \chi_B f_2\},$$

where $B \subset [0,1]^2$ with $\partial B \in C^{\frac{1}{\alpha}}$, and the functions f_1 and f_2 satisfy $f_1, f_2 \in C_0^{\frac{1}{\alpha}}([0,1]^2)$, $\|f_1\|_{C^{\frac{1}{\alpha}}}, \|f_2\|_{C^{\frac{1}{\alpha}}}, \|\partial B\|_{C^{\frac{1}{\alpha}}} < \nu$.

Illustration:





Optimal Sparse Approximation with α-Shearlets

Theorem (Grohs, Keiper, K, and Schäfer; 2016)(Voigtlaender; 2017):
Let $\alpha \in [\frac{1}{2}, 1]$, let $\phi, \psi \in L^2(\mathbb{R}^2)$ be sufficiently smooth and compactly supported, and let ψ have sufficiently many vanishing moments. Also set $\widetilde{\psi}(x_1, x_2) := \psi(x_2, x_1)$ for all $x_1, x_2 \in \mathbb{R}$.
Then there exists some $c^* > 0$ such that, for every $\varepsilon > 0$, there exists a constant $C_\varepsilon > 0$ with

$$\|f - f_N\|_{L^2(\mathbb{R}^2)} \leq C_\varepsilon N^{-\frac{1}{2\alpha}+\varepsilon} \qquad \text{for all } f \in \mathcal{E}^{\frac{1}{\alpha}}(\mathbb{R}^2),$$

where f_N is a best N-term approximation with respect to $\mathcal{SH}_\alpha(\varphi, \psi, \widetilde{\psi}, c)$ and $0 < c < c^*$.

This is the (almost) optimal sparse approximation rate!

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 $\mathcal{C} \subseteq L^2(\mathbb{R}^2)$.
$\rightsquigarrow$ *α-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 $\mathcal{C} \subseteq L^2(\mathbb{R}^2)$.
$\rightsquigarrow$ *α-Cartoon-like functions!*

(2) Determine an associated representation system with the following properties:
$\rightsquigarrow$ *α-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 $\mathcal{C} \subseteq L^2(\mathbb{R}^2)$.
$\rightsquigarrow$ *α-Cartoon-like functions!*

(2) Determine an associated representation system with the following properties:
$\rightsquigarrow$ *α-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 $\mathcal{C} \subseteq L^2(\mathbb{R}^2)$.
$\rightsquigarrow$ *α-Cartoon-like functions!*

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

Construction of Generators

Next Task: Realize sufficiently smooth functions with sufficiently many vanishing moments with a neural network.

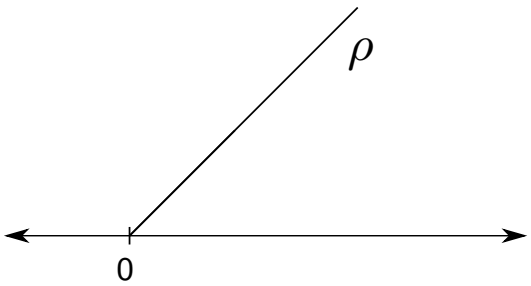
Construction of Generators

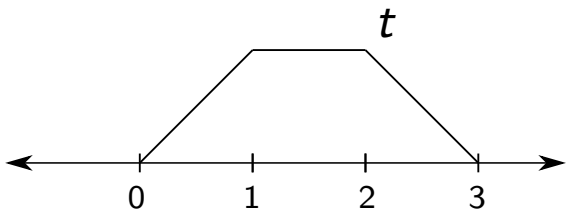
Next Task: Realize sufficiently smooth functions with sufficiently many vanishing moments with a neural network.

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

- Assume rectifiers $\rho(x) = \max\{x, 0\}$ (ReLUs).
- Define

$$t(x) := \rho(x) - \rho(x-1) - \rho(x-2) + \rho(x-3).$$





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

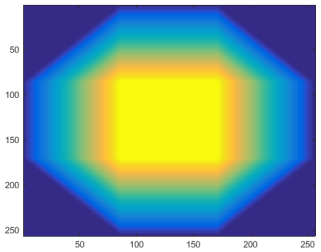
Construction of Wavelet Generators

Construction by (Shaham, Cloninger, and Coifman; 2017) continued:

- Observe that

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

  yields a 2D bump function.

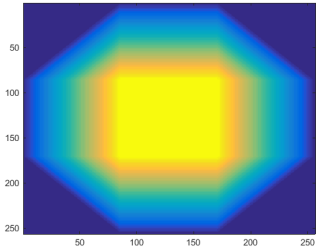


- Summing up shifted versions of ϕ yields a function ψ with vanishing moments.
- Then ψ can be realized by a 3 layer neural network.

Construction of Wavelet Generators

Construction by (Shaham, Cloninger, and Coifman; 2017) continued:

- Observe that
$$\phi(x_1, x_2) := \rho(t(x_1) + t(x_2) - 1)$$
yields a 2D bump function.



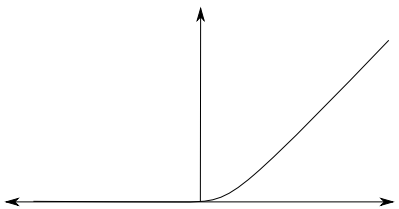
- Summing up shifted versions of ϕ yields a function ψ with vanishing moments.
- Then ψ can be realized by a 3 layer neural network.

This cannot yield differentiable functions ψ!

New Class of Rectifiers

Definition (Bölcskei, Grohs, K, Petersen; 2017):
Let $\rho : \mathbb{R} \to \mathbb{R}^+$, $\rho \in C^\infty(\mathbb{R})$ satisfy

$$\rho(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ x & \text{for } x \geq K, \end{cases}$$

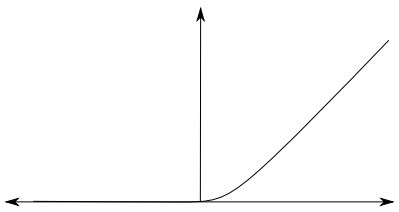


for some constant $K > 0$. Then we call ρ an *admissible smooth rectifier*.

New Class of Rectifiers

Definition (Bölcskei, Grohs, K, Petersen; 2017):
Let $\rho : \mathbb{R} \to \mathbb{R}^+$, $\rho \in C^\infty(\mathbb{R})$ satisfy

$$\rho(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ x & \text{for } x \geq K, \end{cases}$$



for some constant $K > 0$. Then we call ρ an *admissible smooth rectifier*.

Construction of 'good' generators:
Let ρ be an admissible smooth rectifier, and define

$$\begin{aligned} t(x) &:= \rho(x) - \rho(x-1) - \rho(x-2) + \rho(x-3), \\ \phi(x) &:= \rho(t(x_1) + t(x_2) - 1). \end{aligned}$$

This yields smooth bump functions ϕ, and thus smooth functions ψ with many vanishing moments.
$\rightsquigarrow$ *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 the following holds:
For all $f \in \mathcal{E}^{\frac{1}{\alpha}}(\mathbb{R}^2)$ and $N \in \mathbb{N}$, there exists $\Phi \in \mathcal{NN}_{3,O(N),2,\rho}$ such that

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

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 the following holds:
For all $f \in \mathcal{E}^{\frac{1}{\alpha}}(\mathbb{R}^2)$ and $N \in \mathbb{N}$, there exists $\Phi \in \mathcal{NN}_{3,O(N),2,\rho}$ such that

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

This is the optimal approximation rate:

Theorem (Grohs, Keiper, K, Schäfer; 2016): We have

$$\gamma^*(\mathcal{E}^{\frac{1}{\alpha}}(\mathbb{R}^2)) = 2\alpha.$$

Different Choice of Rectifiers

Similar results can be obtained if ρ is not smooth, but a sigmoidal function.

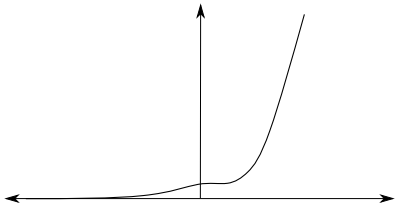
Definition (Cybenko; 1989): A function $\rho : \mathbb{R} \to \mathbb{R}$ is called a *sigmoidal function of order $k \geq 2$*, if

$$\lim_{x \to -\infty} \frac{1}{x^k} \rho(x) = 0, \quad \lim_{x \to \infty} \frac{1}{x^k} \rho(x) = 1$$

and, for all $x \in \mathbb{R}$,

$$|\rho(x)| \leq C(1 + |x|)^k,$$

where $C \geq 1$ is a constant.

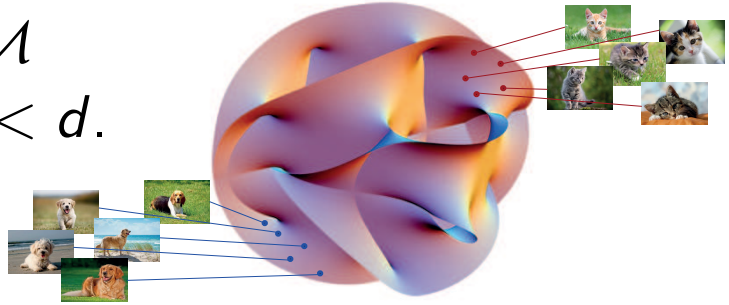


Key Idea:
Use a result by (Chui, Mhaskar; 1994) about approximation of B-splines with DNNs.

Functions on Manifolds

Situation:

We now consider $f : \mathcal{M} \subseteq \mathbb{R}^d \to \mathbb{R}$, where $\mathcal{M}$ is an immersed submanifold of dimension $m < d$.



Road Map for Extension of Results:

- Construct atlas for $\mathcal{M}$ by covering it with open balls.
- Obtain a smooth partition of unity of $\mathcal{M}$.
- Represent any function on $\mathcal{M}$ as a sum of functions on $\mathbb{R}^m$.

$\rightsquigarrow$ *We require function classes $\mathcal{C}$ which are invariant with respect to diffeomorphisms and multiplications by smooth functions such as $\mathcal{E}^{\frac{1}{\alpha}}(\mathbb{R}^2)$.*

Theorem (Bah, Keiper, and K; 2017):

"Deep neural networks are optimal for the approximation of piecewise smooth functions on manifolds."

Finally some Numerics...

Approximation by Learned Networks

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

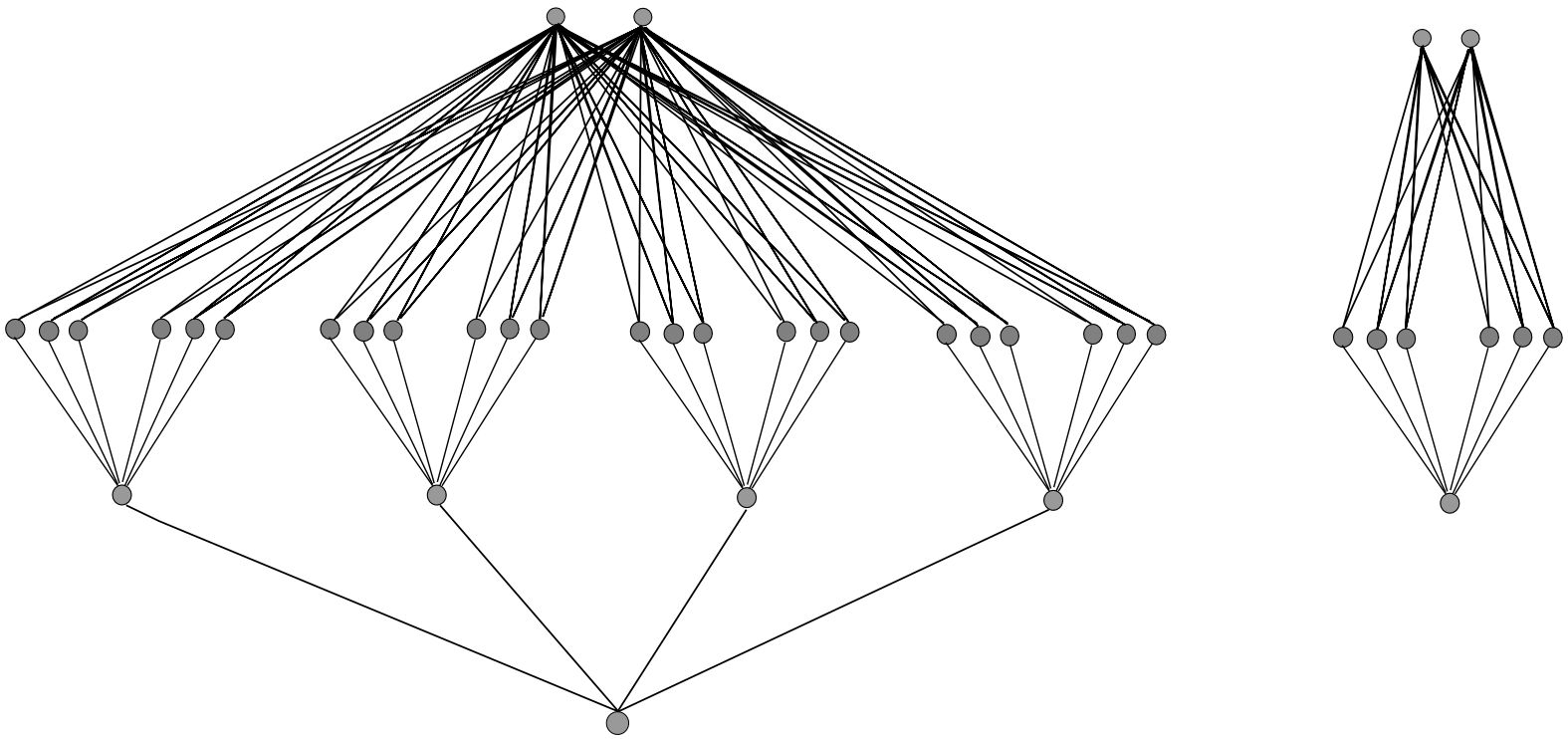
Does this lead to the optimal sparse connectivity?

Our setup:

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

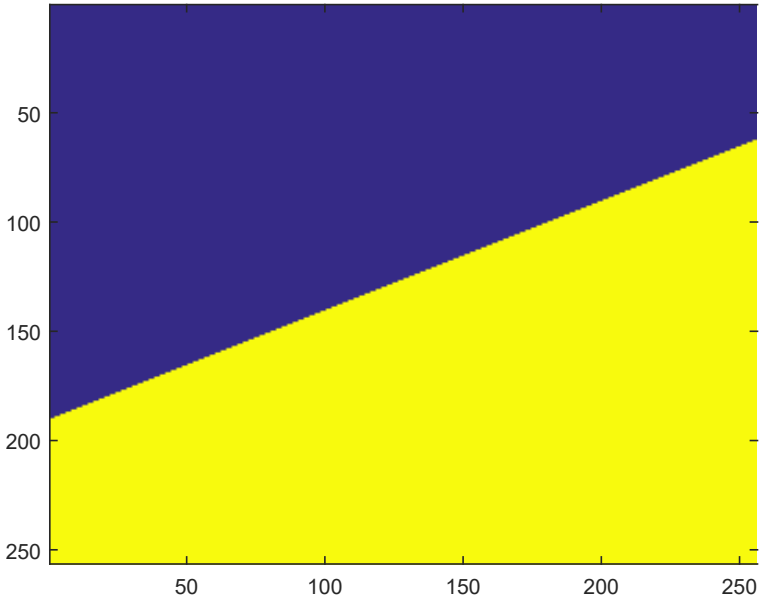
Chosen Network Topology

Topology inspired by previous construction of functions:



Example: Function 1

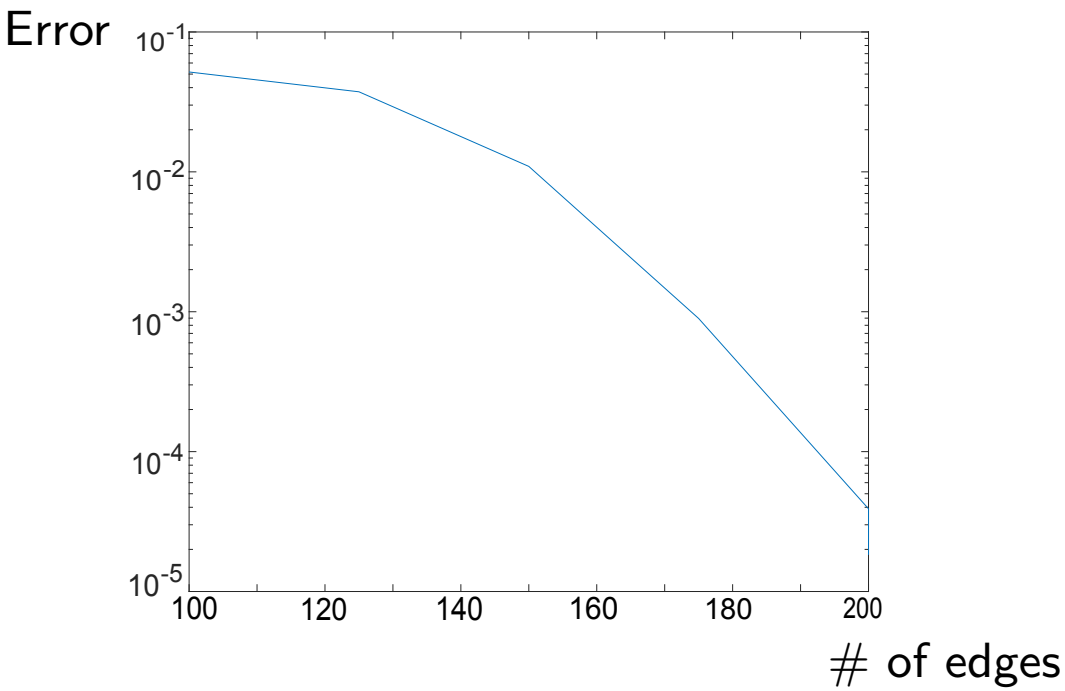
We train the network using the following function:



Function 1: Linear Singularity

Training with Function 1

Approximation Error:

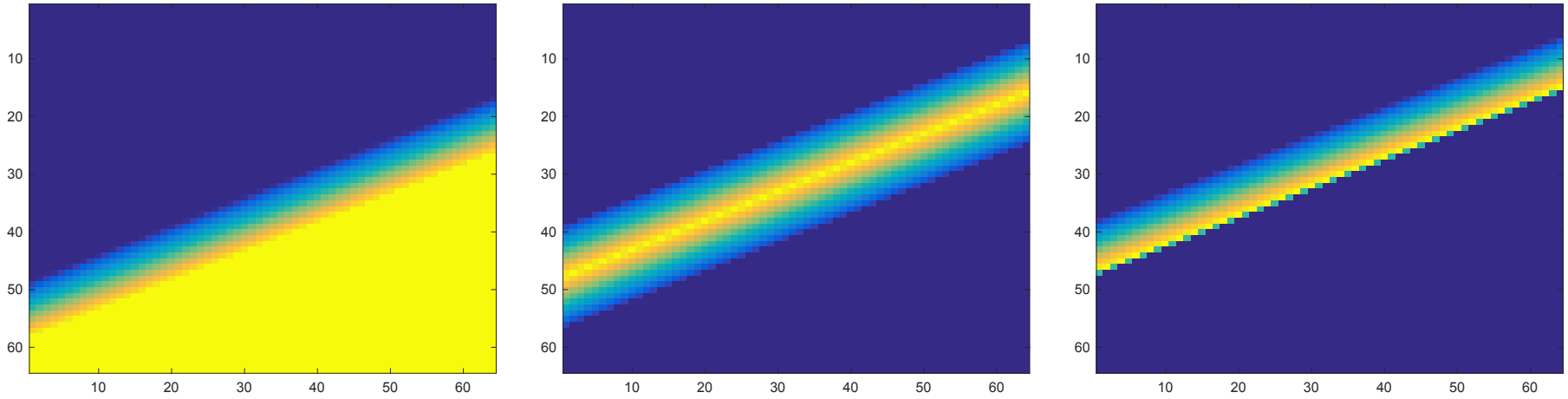


Observation: The decay is exponential. This is expected if the network is a sum of 0-shearlets, which are ridgelets.

Training with Function 1

The network with fixed topology naturally admits subnetworks.

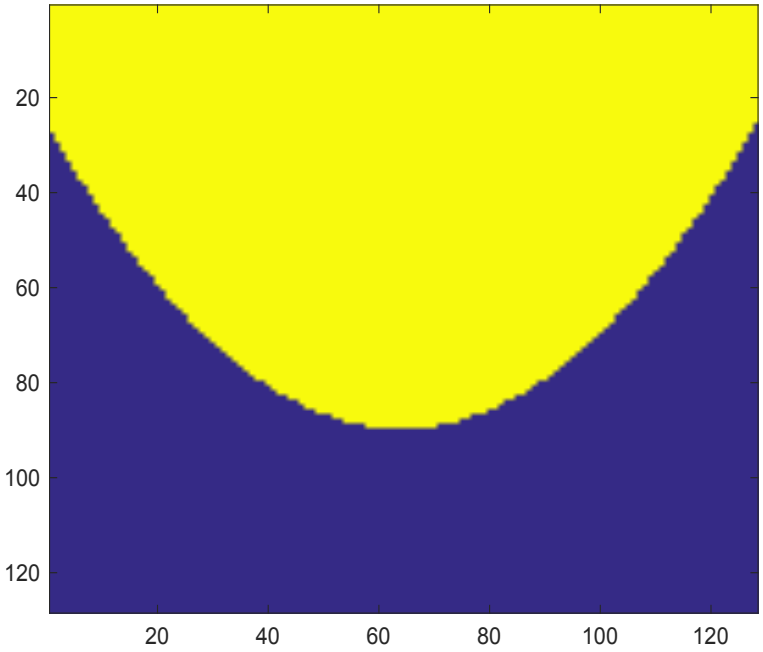
Examples of Subnetworks:



These have indeed the shape of ridgelets!

Example: Function 2

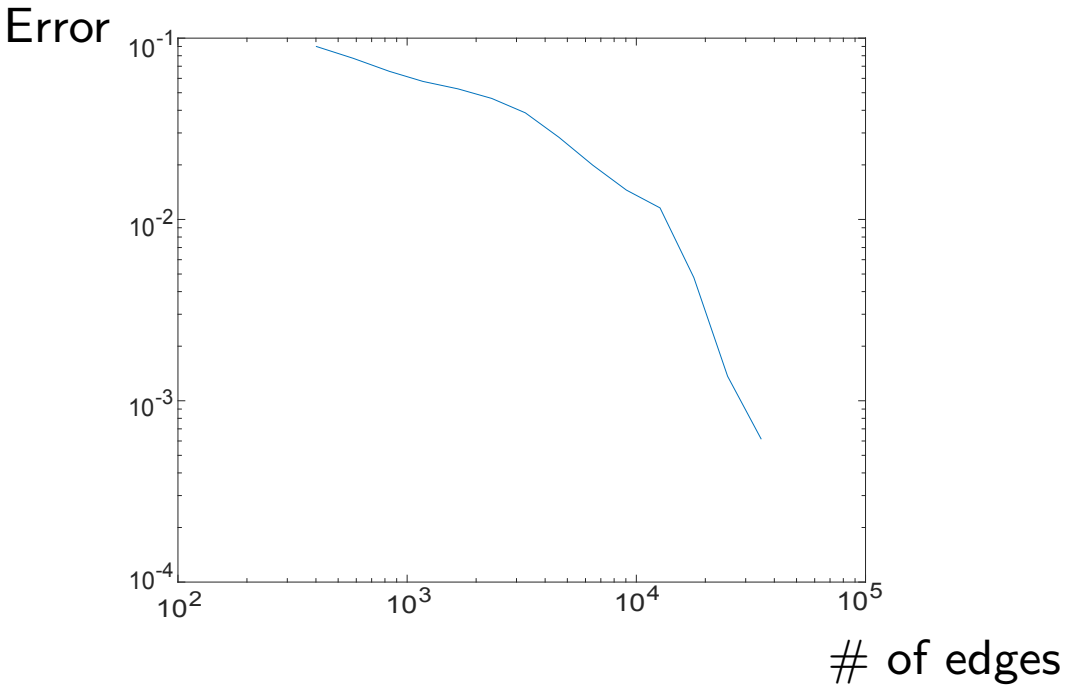
We train the network using the following function:



Function 2: Curvilinear Singularity

Training with Function 2

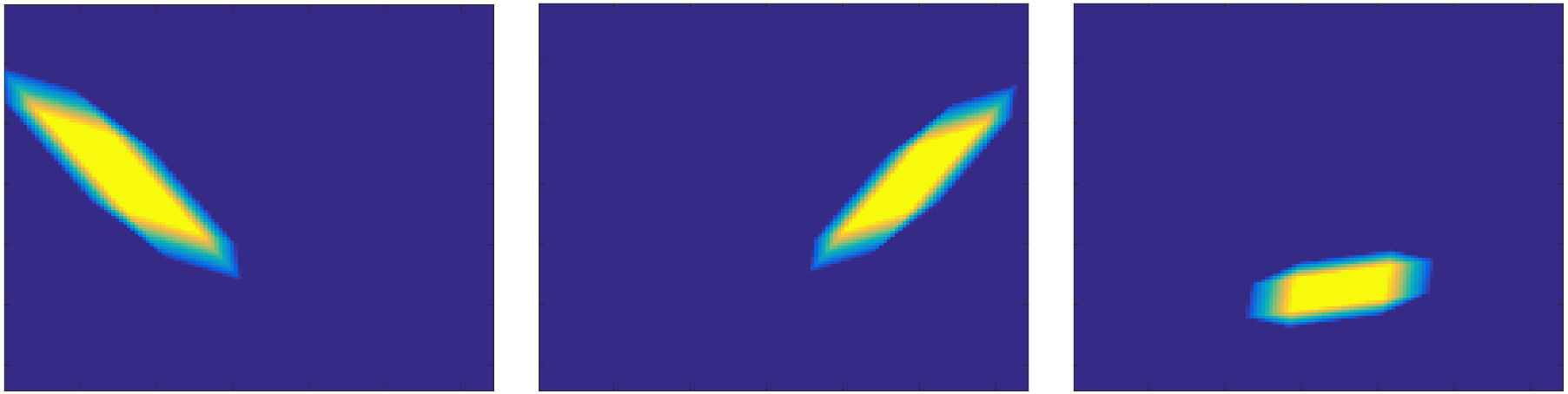
Approximation Error:



Observation: The decay is of the order M^{-1}. This is expected if the network is a sum of $\frac{1}{2}$-shearlets.

Training with Function 2

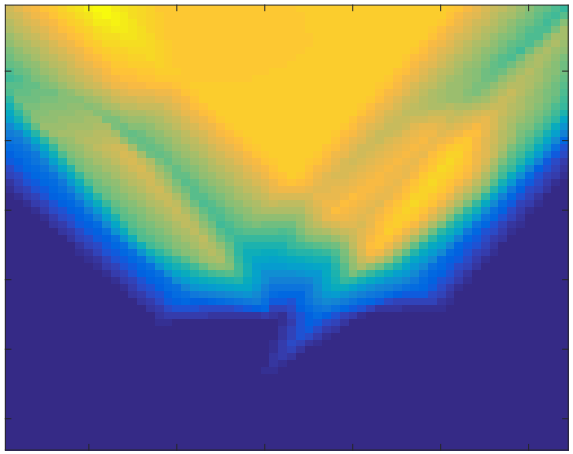
Examples of Subnetworks:



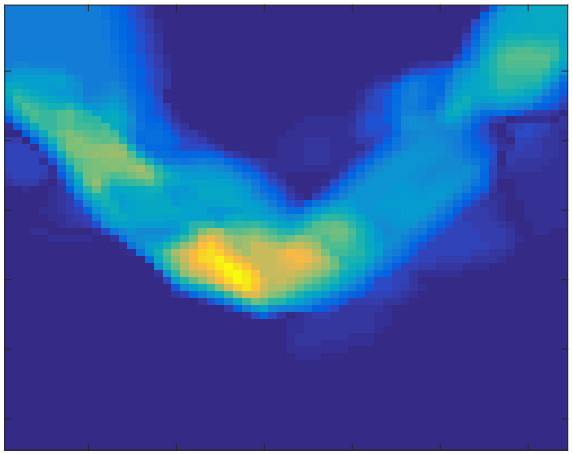
These seem to be indeed anisotropic!

Training with Function 2

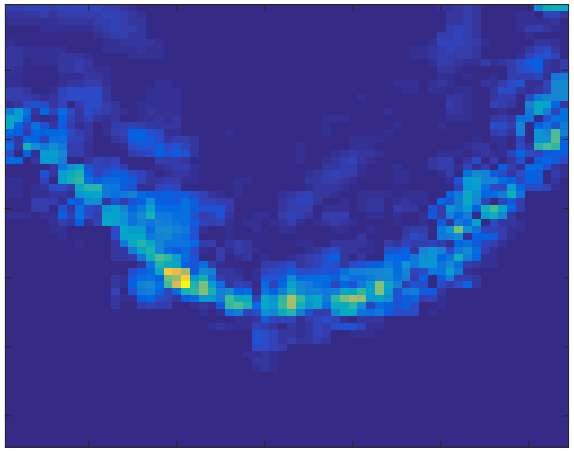
Form of Approximation:



Largest Elements



Medium Sized Elements



Smallest Elements

The learnt neural network has a multiscale behavior!

Let's conclude...

What to take Home...?

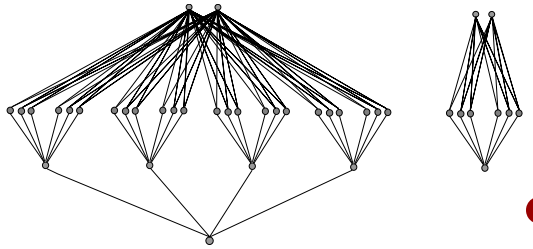
Approximation accuracy $\leftrightarrow$ *Complexity of approximating DNN in terms of sparse connectivity*

Our Contributions:

(1) How many edges do we need for a certain accuracy?
$\rightsquigarrow$ *Conceptual lower bound on the number of edges of the DNN, which each learning algorithm has to obey!*

(2) Are there DNNs which are optimally sparsely connected?
$\rightsquigarrow$ *Sharpness of the bound by explicit construction of optimal DNNs using approximation by representation systems!*

(3) But what happens in practise using ReLUs and backpropagation?
$\rightsquigarrow$ *Success of certain network topologies to reach optimal bound!*
$\rightsquigarrow$ *Learning of optimal multiscale systems from applied harmonic analysis!*



THANK YOU!

References available at:

www.math.tu-berlin.de/∼kutyniok

Our Blog/Database for the Mathematical Theory for Deep Learning:

www.DeepMath.org