MATH 6397 - Mathematics of Data Science

Instructor: Demetrio Labate

January 30, 2023

What is data science?

In a way, applied and numerical mathematics has always been about "data science".

Classical problems from applied and numerical mathematics:

1. how to predict a pattern?
2. how to recover a signal from measurements?
3. how to interpolate or fit the data?
4. ...

The hallmark of applied and numerical mathematics is the reliance on **models,** e.g., a differential equation, a function space, a representation system, that is invoked as a way to explain the data.

What is data science?

The current notion of data science by contrast in centered around the notion of **learning**: How can we explain data and predict patterns by learning from examples?

Model-based approach

- Can be applied to multiple operational problems
- Low data calibration
- High computational cost in general
- Requires domain knowledge

Learning-based approach

- Retraining needed when operational conditions change
- Careful data calibration
- Low computational cost once trained
- No domain knowledge needed

What is data science?

One day, there was a fire in a wastebasket in the office of the Dean of Research. In rushed a physicist, a chemist, and a data scientist.
The physicist immediately starts to work on how much energy would have to be removed from the fire to stop combustion.
The chemist works on which reagent would have to be added to the fire to prevent oxidation.
While they are busy with their calculations, the data scientist is setting fires to all the other wastebaskets in the office.
"What are you doing?" the others inquire. The data scientist replies: "Well, to solve the problem, you obviously need a larger sample size."

`[www.analyticsvidhya.com/blog/2015/12/hilarious-jokes-videos-statistics-data-science/]`

What is data science? Shall we ignore models?

There are good reasons for integrating model-based (also called physics-based) and learning-based methods.

- Learning-based algorithms typically require many training samples - a situation which is not always feasible.
- Learning-based methods, e.g., deep neural networks, are often not interpretable.
- While AI was born with the goal to emulate the human brain, it is known that human learning is not purely based on learning-by-examples.

Further, model-based ideas play an important role in the design and interpretation of deep learning architectures and other machine learning schemes.

Course Outline

1. Mathematics of signal processing
   - 1.1 Classical signal representations
   - 1.2 Fourier series
   - 1.3 Wavelets
   - 1.4 Shearlets (omit)
   - 1.5 Scattering Transform
2. Mathematics of machine learning
   - 2.1 Geometry of high dimensional data
   - 2.2 Statistical learning theory
   - 2.3 Support Vector Machines
   - 2.4 Convolutional Neural Networks

References:

- *The Mathematics of Signal Processing*, by Damelin and Miller
- *Foundations of Data Science*, by Blum, Hopcroft and Kannan
- *Foundations of Machine Learning*, by Mohri, Rostamizadeh and Talwalkar
- *Deep Learning with PyTorch*, by Stevens, Antiga and Viehmann

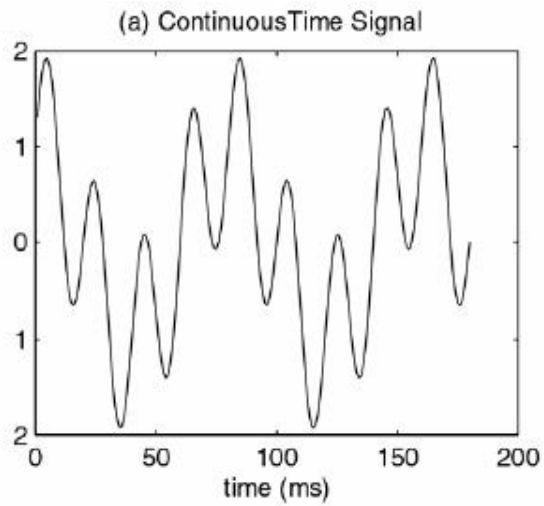
Part I
Mathematics of Signal Processing

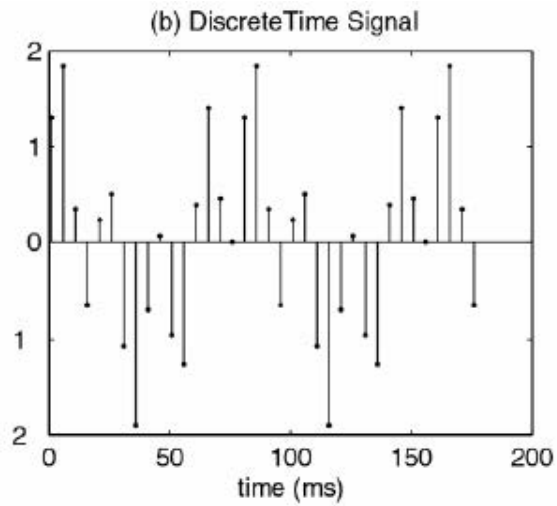
Signals and systems

In classical signal processing, a **signal** is a function conveying information about the state of a physical system.

Examples:
A speech signal is a function of time, a photographic image is a brightness function of two space variables.

Continuous-time or analog signals are represented by a continuous independent variable.
Discrete-time signals, which typically arise by sampling continuous-time signals, are represented by a discrete variable.





Signals and systems

Classical signal processing adopts the formalism of **Hilbert spaces**.
A continuous-time signal is a function $f \in L^2([0, a])$, for some $a > 0$.
A (continuous) image is a function $f \in L^2([0, 1]^2)$
A discrete-time signal is a function $f \in \ell^2$

Often, we consider signal transformations of the form

$$y = Tx$$

where x is the **input signal**, y is the **output signal** and T is a linear operator modeling the mapping the output signal to the input signal.

For instance, T can be used to model a communication channel. In this case, the problem of interest is *how to recover the input signal x given an altered version y of x.*

Signals and systems

A **signal restoration** problem is an inverse problem aiming at recovering a signal x from the observation

$$y = Tx + \epsilon$$

where T is a linear operator modeling the communication channel or the signal acquisition system and ϵ is a **noise** term.

If $T = I$, the task of recovering x from $y = x + \epsilon$ is called a **signal denoising** problem.

Several algorithms have been proposed to recover the signal x from its corrupted version y, including both model-based and learning-based methods

Signal Processing

A fundamental idea in signal processing and harmonic analysis is to use **function representations**

Suppose that any function f in a function space (e.g., a Hilbert space) can be expressed as a superposition of a simple, easily generated collections of functions

$$\{e_k : k \in K\}$$

so that

$$f = \sum_{k \in K} c_k(f)\, e_k.$$

Then a linear transformation T on f can be broken down into simpler operations on the elementary functions e_k:

$$Tf = \sum_{k \in K} c_k\, Te_k$$

Signal Processing

The three typical steps of a classical **signal processing** process are:

1. **Analysis.** It decomposes a signal into basis components.

$$f \mapsto \{c_k(f), k \in K\} \quad \text{e.g., } c_k(f) = \langle f, e_k \rangle$$

2. **Processing.** It modifies (some of) the basis components of the signal that were obtained through the analysis.

$$\tilde{c}_k = T c_k$$

3. **Synthesis.** It reconstitutes the signal from its modified components.

$$\tilde{f} = \sum_k \tilde{c}_k \, e_k$$

1.1 Fourier Series

Fourier series

Jean Baptiste Joseph Fourier (1768-1830) was a French mathematician, physicist and engineer.

Around 1808, he was trying to solve the heat equation (which he discovered), and was able to compute solutions by expressing them as superpositions of an infinite number of sinusoidal waves.

$$f(t) \sim \sum_k a_k \cos(kt) + \sum_k b_k \sin(kt)$$

He made the claim – seemingly preposterous at his time – that any function (in the interval $[0, 2\pi]$), continuous or discontinuous, could be represented as a linear combination of functions $\sin kt$, $\cos kt$.

Fourier series

Fourier claim about trigonometric representations was literally incorrect but 'morally' true.

The complex exponentials

$$e_k(t) = \frac{1}{\sqrt{2\pi}} e^{ikt}, \quad k \in \mathbb{Z}$$

form an orthonormal basis (ONB) in the Hilbert space $L^2([0, 2\pi])$ with inner product

$$\langle f, g \rangle = \int_0^{2\pi} f(t)\, \overline{g}(t)\, dt.$$

Hence, any $f \in L^2([0, 2\pi])$ satisfies

$$f = \sum_{k \in \mathbb{Z}} \langle f, e_k \rangle\, e_k,$$

where convergence is understood in the sense of L^2 convergence.

Fourier series

The rigorous study of the convergence properties of Fourier series requires careful analysis.

Theorem. *Suppose f has the following properties*

- *$f(t)$ is periodic with period 2π,*
- *$f(t)$ is continuous on $[0, 2\pi]$,*
- *$f'(t)$ is piecewise continuous on $[0, 2\pi]$.*

Then the Fourier series of f

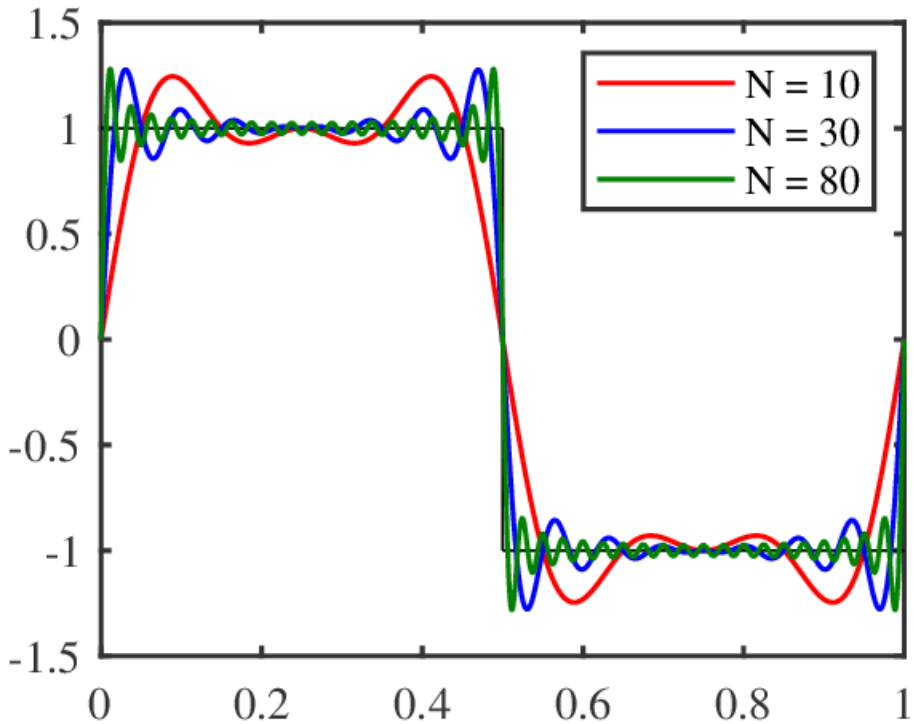
$$\sum_{k\in\mathbb{Z}} a_k e^{ikt}, \qquad a_k = \frac{1}{2\pi}\int_0^{2\pi} f(t)e^{-ikt}\,dt$$

converges pointwise and uniformly.

Remark. Pointwise convergence of a Fourier series does not hold in general. If f has a discontinuity at t_0, then the Fourier series does not converge uniformly at t_0.

Fourier series

Gibbs phenomenon: The Fourier sums overshoot at a jump discontinuity and that this overshoot does not die out as more terms are added to the sum.



This behavior reflects the difficulty inherent in approximating a discontinuous function by a finite series of continuous sine and cosine waves

Fourier series

General principle: **decay vs. regularity**

The decay of the Fourier coefficients of a function at infinity is controlled by the smoothness of that function.
Smooth functions have rapidly decaying Fourier coefficients, resulting in the rapid convergence of the Fourier series.
By contrast, discontinuous functions have slowly decaying Fourier coefficients (causing the Fourier series to converge very slowly).

Fourier series

Riemann-Lebesgue Lemma *Suppose f is piece-wise continuous in the interval $[0, 2\pi]$. Then the Fourier coefficients of f satisfy*

$$\lim_{k\to\infty} a_k = 0.$$

Proof. If $f = \chi_{[a,b]}$, then, by direct integration,

$$\lim_{k\to\infty} a_k = \lim_{k\to\infty} \frac{1}{2\pi} \int_a^b e^{-ikt}\, dt = \lim_{k\to\infty} \frac{e^{ikb} - e^{ika}}{2\pi ik} = 0.$$

The proof follows by observing that any $f \in L^1([0, \pi])$ can be approximated using simple functions (finite linear combinations of characteristic functions) and that simple functions are dense in $L^1([0, \pi])$. □

Remark. The argument also shows that the Fourier coefficients of the characteristic functions decay as $O(\frac{1}{k})$.

Fourier series

Proposition *Let $f \in C^1([0, 2\pi])$. Then the Fouries coefficients a_k of f satisfy*

$$|a_k| = o(\tfrac{1}{k}) \qquad \textit{(decay is faster than } \tfrac{1}{k}\textit{)}$$

Proof. Integration by parts gives that

$$\begin{aligned} a_k = \frac{1}{2\pi}\int_0^{2\pi} f(t)e^{-ikt}\,dt &= -f(t)\frac{e^{-ikt}}{i2\pi k}\Big|_0^{2\pi} + \frac{1}{2\pi}\int_0^{2\pi} f'(t)\frac{e^{-ikt}}{ik}\,dt \\ &= -\frac{i}{2\pi k}\int_0^{2\pi} f'(t)e^{-ikt}\,dt, \end{aligned}$$

showing that $ka_k = -\frac{i}{2\pi}\int_0^{2\pi} f'(t)e^{-ikt}\,dt$.
By the Riemann-Lebesgue Lemma, the RHS converges to 0 as $k \to \infty$. $\square$

Fourier series

The same argument holds for higher order derivatives.

Proposition *Let $f \in C^n([0, 2\pi])$. Then the Fourier coefficients a_k of f satisfy*

$$|a_k| = o(\tfrac{1}{k^n})$$

Proposition *Let $f \in C^\infty([0, 2\pi])$. Then the Fouries coefficients a_k of f satisfy*

$$\lim_{k \to \infty} k^n |a_k| = 0, \quad \text{for all } n.$$

Sparse representations

For practical applications, it is useful to approximate functions using finite-length expansions.

For instance, we can approximate $f \in L^2([0, 2\pi])$ using the truncated Fourier series

$$\sum_{|k| \leq M} a_k e^{ikt},$$

where the coefficients a_k are the Fourier coefficients of f.

In general, given a representation system $\{\psi_i, i \in I\}$, we can compute the N term approximation of $f \in X \subset L^2([0, 2\pi])$ as

$$f_N = \sum_{N \text{ terms}} c_i(f)\, \psi_i.$$

How shall we assess the *approximation properties* of the representation system $\{\psi_i, i \in I\}$?

Sparse representations

We measure the N-term approximation error on the function f

$$E_N(f) = \|f - f_N\|^2,$$

which is clearly a non-increasing function of N.

If $E_N(f)$ decays rapidly as N increases, for all $f \in X$, we say that the representation system $\{\psi_i : i \in I\}$ is **sparse** or **compressible** in X.

Hence most of the information of $f \in X$ can be recovered using a relatively small number of representation terms.

Sparse representations

If a signal f is highly regular, that is, $f \in C^n([0, 2\pi])$ with large n, the Fourier coefficients of f have rapid decay.
This implies that if we approximate f with an N-term Fourier approximation f_N, then the approximation error

$$\|f - f_N\|^2 \leq C\, N^{-2n}$$

has rapid decay as $N \to \infty$.

However, if f has a discontinuity, the Fourier coefficients a_k of f only decay as $O(k^{-1})$ in which case

$$\|f - f_N\|^2 \leq C\, N^{-1}$$

This estimate holds for functions of bounded variation $BV([0, 2\pi])$.

The results above show that Fourier series have limitations when dealing with discontinuous signals.

Sparse representations

Why Sparse Representations?

• **Data Compression.** If f has a sparse representation with respect to a basis $\{\psi_i : i \in I\}$, then it is sufficient to keep only a "small" number of representation coefficients to have a "sufficiently good" approximation of f.
This is useful to store or to transmit information.

$$s(t) = \sum_{\mu \in M} c_\mu \psi_\mu \rightarrow \quad \text{(compress)} \quad \rightarrow \{c_\mu\}_{\mu \in M_N}$$

$$\rightsquigarrow \quad \text{(transmit)} \quad \rightsquigarrow s_N(t) = \sum_{\mu \in M_N} c_\mu \psi_\mu$$

Examples: MP3 (audio); JPEG, JPEG2000 (image); MPEG (video)

Sparse representations

Why Sparse Representations?

Sparsity goes beyond compression. Capturing the sparse representation of a signal entails capturing its essential features.

- **Denoising.** Let s be a signal corrupted by noise:

$$s(t) = f(t) + n(t).$$

By representing s with respect to a sparse representation

$$s(t) = \sum_i c_i(s)\,\psi_i(t)$$

most of the information of f is concentrated in a few coefficients. By discarding the other coefficients, most of the noise is also removed.

- Also: **Feature Extraction, Inverse problems, . . .**

1.2 Wavelets

Wavelet representations

One major problem about Fourier series is that trigonometric functions are *not local*.

Wavelets were introduced to address this issue.

A wavelet basis is constructed by taking *dilated and translated copies* of an appropriate 'mother' function $\psi \in L^2(\mathbb{R})$.
Define the unitary operators of $L^2(\mathbb{R})$

- **Translation operator.** $T_y f(x) = f(x - y)$, $y \in \mathbb{R}$
- **Dilation operator.** $D_a f(x) = a^{1/2} f(ax)$, $a > 0$

A wavelet system has the form

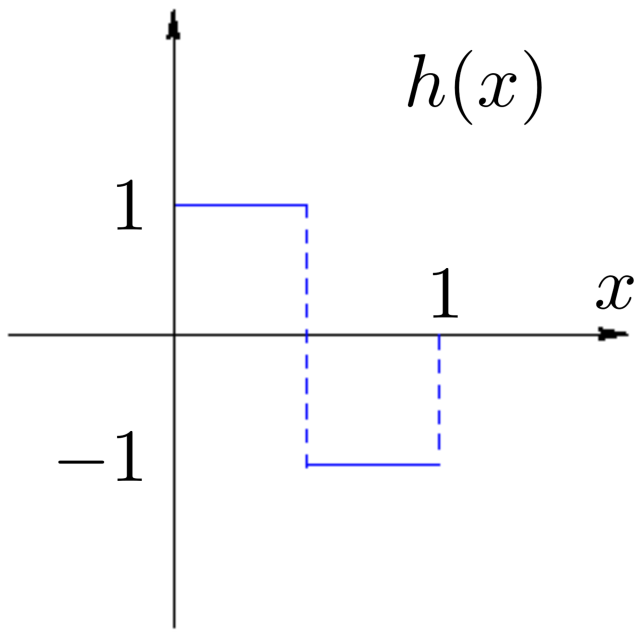
$$\Psi = \{\psi_{j,k}(x) = D_2^j T_k \psi(x) = 2^{j/2}\, \psi(2^j x - k) : \ j, k \in \mathbb{Z}\} \subset L^2(\mathbb{R})$$

Note: $\|\psi_{j,k}\| = \|\psi\|$, for each j, k.

Wavelet representations - The Haar wavelet

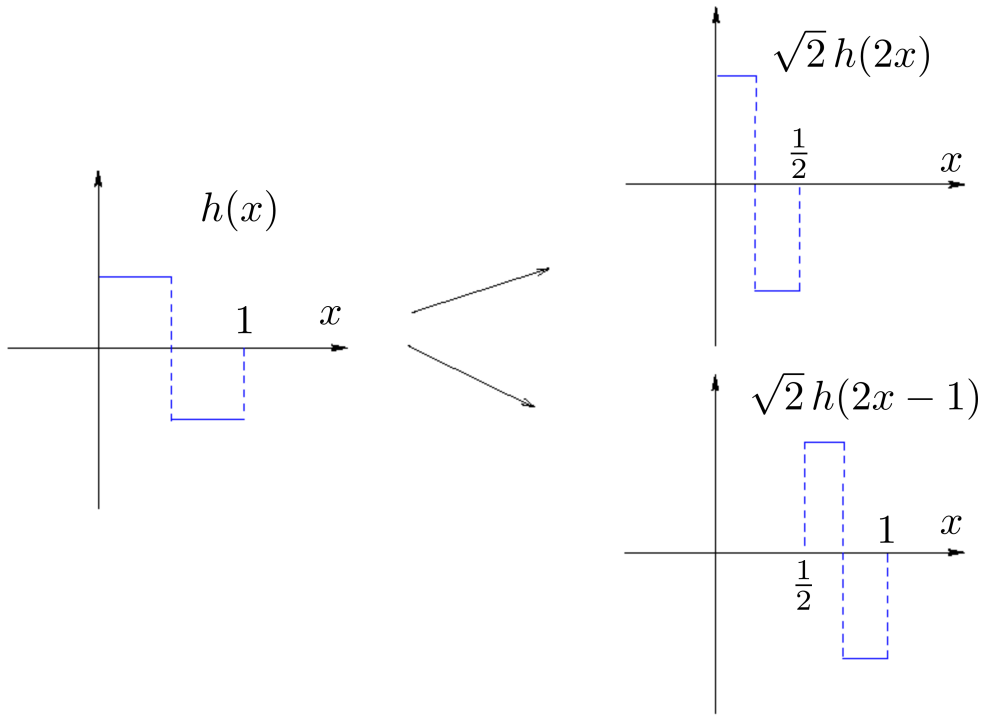
Example: *Haar wavelet system* (1910)

The Haar wavelet is: $h(x) = \begin{cases} 1 & \text{if } 0 \leq x < \frac{1}{2} \\ -1 & \text{if } \frac{1}{2} < x < 1 \\ 0 & \text{otherwise} \end{cases}$



Wavelet representations - The Haar wavelet

The Haar system $\{h_{j,k}(x) = 2^{j/2}\, h(2^j x - k) : \ j, k \in \mathbb{Z}\}$ is an orthonormal system of $L^2(\mathbb{R})$.



- Each $h_{j,k}$ is supported on an interval $I_{j,k}$ of size 2^{-j}.
- $I_{j,k} \cap I_{j,k'} = \varnothing$ if $k \neq k'$. Hence $\langle h_{j,k}, h_{j,k'} \rangle = 0$.
- If $j < j'$, $h_{j,k}$ and $h_{j',k'}$ can only overlap on $I_{j',k'}$ where $h_{j,k}$ is constant. Hence $\langle h_{j,k}, h_{j',k'} \rangle = 0$.

Wavelet representations - The Haar wavelet

Theorem. The Haar system $\{h_{j,k} : j,k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$. For any $f \in L^2(\mathbb{R})$:

$$f = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \langle f, h_{j,k}\rangle \, h_{j,k},$$

with convergence in L^2 sense

$$\|f\|^2 = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} |\langle f, h_{j,k}\rangle|^2$$

Interpretation: The *Haar coefficients* $\langle f, h_{j,k}\rangle$ measure the energy content of f at **location** $2^{-j}k$ and **scale** 2^{-j}.

Remark. The Haar system is an unconditional basis for L^p, $1 < p < \infty$ [Paley, 1932]. It is a conditional basis for L^1. (cf. *A basis theory primer,* by C. Heil)

Wavelet representations - The Haar wavelet

Proof. I consider the alternate Haar system

$$B = \{T_k \chi_{[0,1]} : k \in \mathbb{Z}\} \cup \{h_{j,k} = D_2^j T_k h : j \geq 0, k \in \mathbb{Z}\}$$

Since I have already shown orthogonality, I only need to prove completeness. For that, I will show that there is no $f \in L^2(\mathbb{R})$ which is orthogonal to all elements of B.
Since

$$\langle f, \chi_{[0,1]} \rangle = \int_0^1 f = \int_0^{1/2} f + \int_{1/2}^1 f = 0$$

and

$$\langle f, h \rangle = \int_0^{1/2} f - \int_{1/2}^1 f = 0,$$

by adding and subtracting we have that

$$\int_0^{1/2} f = \int_{1/2}^1 f = 0$$

Wavelet representations - The Haar wavelet

Proof (continued)

Applying the same ideas for general j, k, it follows that

$$\int_{I_{j,k}} f = 0 \quad \text{for any } I_{j,k} = [\tfrac{k}{2^j}, \tfrac{k+1}{2^j}]$$

Given $x \in \mathbb{R}$, for each $j \in \mathbb{N}$ there is a dyadic interval $J_j(x) = I_{j,k_j(x)}$ such that $x \in J_j(x)$.
Note that $|J_j(x)| = 2^{-j}$ so that $\lim_{j\to\infty} J_j(x) = \{x\}$.
By the Lebesgue Differentiation Theorem, for almost all $x \in \mathbb{R}$,

$$f(x) = \lim_{j\to\infty} \frac{1}{|J_j(x)|} \int_{J_j(x)} f(u)\, du = 0$$

This shows that $f = 0$ a.e. $\square$

Wavelet representations - MRA

There is a great number of examples of wavelet systems

$$\psi_{j,k}(x) = D_2^j \, T_k \, \psi(x) = 2^{j/2} \, \psi(2^j \, x - k)$$

Multiresolution Analysis (MRA) [Mallat, Meyer, 1989] provides a general method to construct orthonormal wavelet bases, even with additional properties such as regularity, decay, support.

The mother *wavelet* ψ can be chosen to be a *well-localized* function, i.e., ψ has rapid decay both in $\mathbb{R}$ (the 'time' domain) and $\widehat{\mathbb{R}}$ (the 'frequency' domain).

Note that the Haar wavelet is not well-localized since it is discontinuous (hence its Fourier transform decay a $O(\frac{1}{\omega})$).

Wavelet representations - MRA

The idea of multiresolution analysis is that dyadic wavelet bases naturally divide $L^2(\mathbb{R})$ into subspaces with spaces resolution levels. If we define closed subspaces

$$W_j = \overline{\text{span}}\{D_2^j T_k \psi\}_{k\in\mathbb{Z}}, \qquad j \in \mathbb{Z}$$

and let Q_j denote the orthogonal projection of $L^2(\mathbb{R})$ onto W_j, then we can write any $f \in L^2(\mathbb{R})$ as

$$f = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \langle f, \psi_{j,k}\rangle \, \psi_{j,k} = \sum_{j\in\mathbb{Z}} Q_j f$$

Interpretation: The space W_j is generated by functions $D_2^j T_k \psi$ that all have the same 'detail size.'

Wavelet representations - MRA

Next, we define closed subspaces

$$V_j = \overline{\text{span}}\{D_2^i T_k \psi\}_{i<j,k\in\mathbb{Z}}, \qquad j \in \mathbb{Z}$$

and let P_j denote the orthogonal projection of $L^2(\mathbb{R})$ onto V_j.
In some sense $P_j f$ is an **approximation** to f at 'resolution level j.'
We move from a resolution level to the next one by adding *details* from W_j via the operator Q_j:

$$P_{j+1} f = P_j f + Q_j f.$$

By definition of orthonormal basis, $P_j f$ converges to f as j increases

$$f = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \langle f, \psi_{j,k}\rangle \, \psi_{j,k} = \lim_{j\to\infty} \sum_{i<j} \sum_{k\in\mathbb{Z}} \langle f, \psi_{i,k}\rangle \, \psi_{i,k}$$

and we accomplish this by adding information with finer and finer details $W_j f$ as j increases.

Wavelet representations - MRA

Definition (Multiresolution Analysis). A multiresolution analysis (MRA) for $L^2(\mathbb{R})$ is a sequence $\{V_j\}_{j\in\mathbb{Z}}$ of closed subspaces of $L^2(\mathbb{R})$ such that:

1. $V_j \subset V_{j+1}$ for each $j \in \mathbb{Z}$,
2. $V_{j+1} = D_2(V_j)$ for each $j \in \mathbb{Z}$,
3. $\cup_{j\in\mathbb{Z}} V_j$ is dense in $L^2(\mathbb{R})$,
4. $\cap_{j\in\mathbb{Z}} V_j = \{0\}$,
5. there exists a function $\phi \in V_0$ such that $\{T_k\phi\}_{k\in\mathbb{Z}}$ is an orthonormal basis of V_0.

ϕ is called a **scaling function** of the MRA.

Remarks. Statements 1-5 above are not independent. Statement 4 can be shown to be implied by the other statements.
One can create a more general definition of MRA by requiring only that $\{T_k\phi\}_{k\in\mathbb{Z}}$ be a Riesz basis or a frame for V_0.

Wavelet representations - MRA

The Proposition below follows directly from the MRA definition.

Proposition. Suppose that $\{V_j\}_{j\in\mathbb{Z}}$is an MRA for $L^2(\mathbb{R})$ and let P_j denote the orthogonal projection of $L^2(\mathbb{R})$ onto V_j. Then the following statements hold.

1. $V_j = D_2^j(V_0) = \{f(2^j x) : f \in V_0\}$ for each $j \in \mathbb{Z}$.
2. $\{D_2^j T_k \phi\}_{k\in\mathbb{Z}}$ is an ONB of V_j.
3. V_0 is shift invariant and V_j is 2^{-j}-shift invariants.
4. $\lim_{j\to\infty} P_j f = f$ in $L^2(\mathbb{R})$, for every $f \in L^2(\mathbb{R})$.
5. $\lim_{j\to-\infty} P_j f = 0$ in $L^2(\mathbb{R})$, for every $f \in L^2(\mathbb{R})$.

According to the Proposition, the spaces V_j in an MRA are completely determined by the base space V_0. Therefore, if we want to build an MRA then we can focus on the space V_0 and the scaling function ϕ.

Wavelet representations - Haar MRA

Example: Haar MRA. While in the direct construction above we have begun with a wavelet and the wavelet system that it generates, here we start with the Haar scaling function and show how the Haar wavelet is produced from this MRA.

The base space V_0 for the Haar MRA is the space of all step functions in $L^2(\mathbb{R})$ that are constant on intervals $[k, k+1)$:

$$V_0 = \left\{ \sum_k c_k \chi_{[k,k+1]} : c_k \in \ell^2(\mathbb{Z}) \right\} \subset L^2(\mathbb{R})$$

We define $V_j = D_2^j(V_0)$ (property 2 of the MRA).
Let us verify the other properties of the MRA .

- Clearly $V_1 = D_2(V_0) \subset V_0$ since V_1 is the space of step functions that are constant on intervals $[k/2, (k+1)/2)$. Since $V_j = D_2^j(V_0)$ is the space of step functions that are constant on intervals $[k/2^j, (k+1)/2^j)$, the nestedness requirement $V_j \subset V_{j+1}$ is satisfied (property 1 of the MRA).

Wavelet representations - Haar MRA

- If we set $\phi = \chi_{[0,1]}$, $\{T_k\phi : k \in \mathbb{Z}\}$ is an ONB of V_0, hence ϕ is the scaling function of the MRA (property 5 of the MRA).
- Suppose that $f \in L^2(\mathbb{R})$ belongs to every subspace V_j. Then f must be constant on every interval $[k/2^j, (k+1)/2^j)$ for all $j \in \mathbb{Z}$. In particular, f is constant on $[0, 2^j)$ for every $j \in \mathbb{N}$, which implies f is constant on $[0, \infty)$ and similarly it is constant on $(\infty, 0]$. Since $f \in L^2(\mathbb{R})$, this implies that $f = 0$. This shows property 4 of the MRA.

Wavelet representations - Haar MRA

- To show that $\cup_{j\in\mathbb{Z}} V_j$ is dense in $L^2(\mathbb{R})$, I will show that the projection $P_j f$ of f onto V_j converges to f as $j \to \infty$. In fact,

$$P_j f = \sum_k \langle f, D_2^j T_k \phi \rangle \, D_2^j T_k \phi = \sum_k c_{k,j} \, \chi_{[2^{-j}k, 2^{-j}(k+1))}$$

where

$$c_{k,j} = 2^j \int_{2^{-j}k}^{2^{-j}(k+1)} f(x)\, dx$$

is the average of f over the interval $[2^{-j}k, 2^{-j}(k+1))$.

Wavelet representations - Haar MRA

To explain the connection with the Haar wavelet, define

$$\psi(x) = \chi_{[0,1/2)}(x) - \chi_{[1/2,1)}(x) = \phi(2x) - \phi(2x-1)$$

and set

$$W_0 = \overline{\text{span}}\{T_k\psi\}_{k\in\mathbb{Z}}$$

It is easy to see that $\{T_k\psi\}_{k\in\mathbb{Z}}$ is an ONB of W_0.
In addition, ψ and ϕ are orthogonal so that the spaces V_0 and W_0 are also orthogonal.
The definition of ψ shows that $W_0 \subset V_1$, hence

$$V_0 \oplus W_0 = \{f + g : f \in V_0, g \in W_0\} \subset V_1.$$

In fact, we do have an equality as, by direct calculation, one that any $h \in V_1$ can be written as

$$h = \sum_k c_k D_2 T_k \phi = \sum_k a_k T_k \phi + \sum_k b_k T_k \psi$$

Wavelet representations - Haar MRA

Next, let

$$W_j = \overline{\text{span}}\{D_2^j T_k \psi\}_{k \in \mathbb{Z}}.$$

The spaces W_j are orthogonal and, for any $j > 0$ we have

$$V_{j+1} = V_j \oplus W_j.$$

Hence, iterating we have

$$V_{j+1} = V_0 \oplus W_0 \oplus \cdots \oplus W_j.$$

Wavelet representations - Haar MRA

Since $\{T_k\phi\}_{k\in\mathbb{Z}}$ is an ONB for V_0 and $\{D_2^j T_k\psi\}_{k\in\mathbb{Z}}$ is an ONB of W_j, then we can write

$$P_{j+1}f = \sum_{k\in\mathbb{Z}} \langle f, T_k\phi\rangle\, T_k\phi + \sum_{m=0}^{j}\sum_{k\in\mathbb{Z}} \langle f, D_2^m T_k\psi\rangle\, D_2^m T_k\psi$$

Since $P_j f \to f$, we therefore have

$$f = \sum_{k\in\mathbb{Z}} \langle f, T_k\phi\rangle\, T_k\phi + \sum_{m=0}^{\infty}\sum_{k\in\mathbb{Z}} \langle f, D_2^m T_k\psi\rangle\, D_2^m T_k\psi,$$

which is to the Haar wavelet expansion I derived above.

By writing $V_{j+1} = V_{-n} \oplus W_{-n} \oplus \cdots \oplus W_0 \oplus \cdots \oplus W_j$ and observing that $P_{-n}f \to 0$ and $n \to \infty$, then we can similarly show that the Haar system $\{D_2^j T_k\psi : k, j \in \mathbb{Z}\}$ is an ONB for $L^2(\mathbb{R})$.

Wavelet representations - approximation spaces

In the MRA, the orthogonal projection onto the spaces V_j is

$$P_j f = \sum_{m \le j} \sum_{k \in \mathbb{Z}} \langle f, \phi_{m,k} \rangle \, \phi_{m,k},$$

where $\phi_{m,k}(t) = 2^{m/2} \phi(2^m t - k)$. The coefficients

$$\begin{aligned} \langle f, \phi_{m,k} \rangle &= \int_{\mathbb{R}} f(t) \, 2^{m/2} \overline{\phi(2^m t - k)} dt \\ &= \int_{\mathbb{R}} f(t) \, 2^{m/2} \overline{\phi(2^m (t - 2^{-m} k))} dt \\ &= f * \tilde{\phi}_m(2^{-m} k) \qquad (\tilde{\phi}_m(t) = 2^{m/2} \overline{\phi(-2^m t)}) \end{aligned}$$

measure the energy content of f at scale 2^{-m} and location $2^{-m} k$.

In the Haar case:

$$\langle f, \phi_{m,k} \rangle = \int_{I_{m,k}} f(t) \, dt, \quad I_{m,k} = [2^{-m} k, 2^{-m}(k+1)]$$

Remark: convolution

Formally, the **convolution** of two real or complex functions f and g on $\mathbb{R}^d$ is defined as the integral

$$(f * g)(x) = \int_{\mathbb{R}^d} f(y)\, g(x-y)\, dy = \int_{\mathbb{R}^d} g(y)\, f(x-y)\, dy$$

The convolution of f and g exists if f and g are both in $L^1(\mathbb{R}^d)$, in which case $f * g \in L^1(\mathbb{R}^d)$. The proof of this fact follows from Fubini-Tonelli's theorem.

Also, if $f \in L^1(\mathbb{R}^d)$ and $g \in L^p(\mathbb{R}^d)$, $1 \leq p \leq \infty$, then $f * g \in L^p(\mathbb{R}^d)$ and

$$\|f * g\|_{L^p} \leq \|f\|_{L^1} \|g\|_{L^p}$$

Remark: convolution

Some useful properties:

- If f and g have compact support, then $f * g$ has also compact support.
- If f and g have rapid decay at infinity, then $f * g$ have also rapid decay at infinity.
- The convolution commutes with translations, that is

$$T_y(f * g) = (T_y f) * g = f * (T_y g)$$

- Convolution theorem: the Fourier transform maps convolutions into pointwise products

$$(f * g)^{\wedge}(\xi) = \hat{f}(\xi)\,\hat{g}(\xi)$$

Remark: convolution

In electrical engineering, the operation of convolution is associated with the study of **linear time-invariant systems** (LTI).
The output of an LTI system is given by the convolution of an input signal x with the impulse response h of the system:

$$y(t) = (x * h)(t).$$

The non-rigorous argument often presented in engineering classes is that, using the delta distribution, any signal can be expressed as

$$x(t) = \int x(y)\,\delta(y - x)\,dy$$

If L is a linear operator (and δ is even), then

$$y(t) = (Lx)(t) = \int x(y)\,L(\delta(t - y))\,dy = \int x(y)\,L(\delta(y - t))\,dy$$

Translation-covariance next gives that $LT_t\delta = T_tL\delta$, hence

$$y(t) = (Lx)(t) = \int x(y)\,h(t - y)\,dy$$

where $h = L\delta$.

Remark: convolution

A rigorous argument that linear translation-covariant operators can be represented by convolution follows by the **Schwartz kernel theorem** (L Schwartz 1954).

Theorem

Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ be open sets. A linear map $K : C_0^\infty(Y) \mapsto \mathcal{D}'(X)$ is continuous if and only if it is generated by a distribution kernel $k \in \mathcal{D}'(X \times Y)$, that is, informally,

$$K\phi(x) = \int_Y k(x, y)\, \phi(y)\, dy$$

Moreover, the kernel k is uniquely determined by the mapping K.

Distributions, Ch. 15, Duistermaat and Kolk, Birkhäuser 2010.
On the theory of kernels of Schwartz, L. Ehrenpreis, Proceedings of the American Mathematical Society Vol. 7, No. 4 (Aug., 1956), p. 713-718
The Schwartz kernel theorem for the tempered distributions on the Heisenberg group, Y. Oka, Hokkaido Math. J. Vol. 44 (2015) p. 425–439

Remark: convolution

Clearly, if a linear continuous operator K is a convolution operator $K\phi = h * \phi$, for some h, then $T_{x'} K\phi = K T_{x'}\phi$ for any $x' \in X$.

Conversely, assume that $T_{x'} K\phi = K T_{x'}\phi$ for any $x' \in X$. Then

$$\begin{aligned} \int_{\mathbb{R}^n} k(x - x', y)\, \phi(y)\, dy &= \int_{\mathbb{R}^n} k(x, y)\, \phi(y - x')\, dy \\ &= \int_{\mathbb{R}^n} k(x, y + x')\, \phi(y)\, dy \end{aligned}$$

This implies that $k(x - x', y) = k(x, y + x')$ for any $x, x', y \in \mathbb{R}^n$. Hence (changing variables $y \to y - x'$) $k(x, y) = k(x - x', y - x')$ for any $x, x', y \in \mathbb{R}^n$. Hence, by defining $h(z) = k(0, -z)$

$$\begin{aligned} K\phi(x) = \int_{\mathbb{R}^n} k(x, y)\, \phi(y)\, dy &= \int_{\mathbb{R}^n} k(0, y - x)\, \phi(y)\, dy \\ &= \int_{\mathbb{R}^n} h(x - y)\, \phi(y)\, dy \\ &= (h * \phi)(x) \end{aligned}$$

Remark: convolution

The operation of convolution extends to sequences.
For real or complex valued sequences f and g on $\mathbb{Z}$, the **discrete convolution** is defined as

$$(f * g)[n] = \sum_{k \in \mathbb{Z}} f[k]\, g[n-k] = \sum_{k \in \mathbb{Z}} f[n-k]\, g[k].$$

For finite sequences, we have

$$(f * g)[n] = \sum_{k=0}^{M} f[k]\, g[n-k]$$

and the commutative property does not hold.

However, if a sequence g_N is periodic with period N, then, for functions f such that $f * g_N$ exists, the convolution is also periodic.
This leads to the notion of **periodic convolution**.
Cf. *Discovering Transforms: A Tutorial on Circulant Matrices, Circular Convolution, and the Discrete Fourier Transform*, Bassam Bamieh.
https://arxiv.org/abs/1805.05533

Wavelet representations - Haar MRA

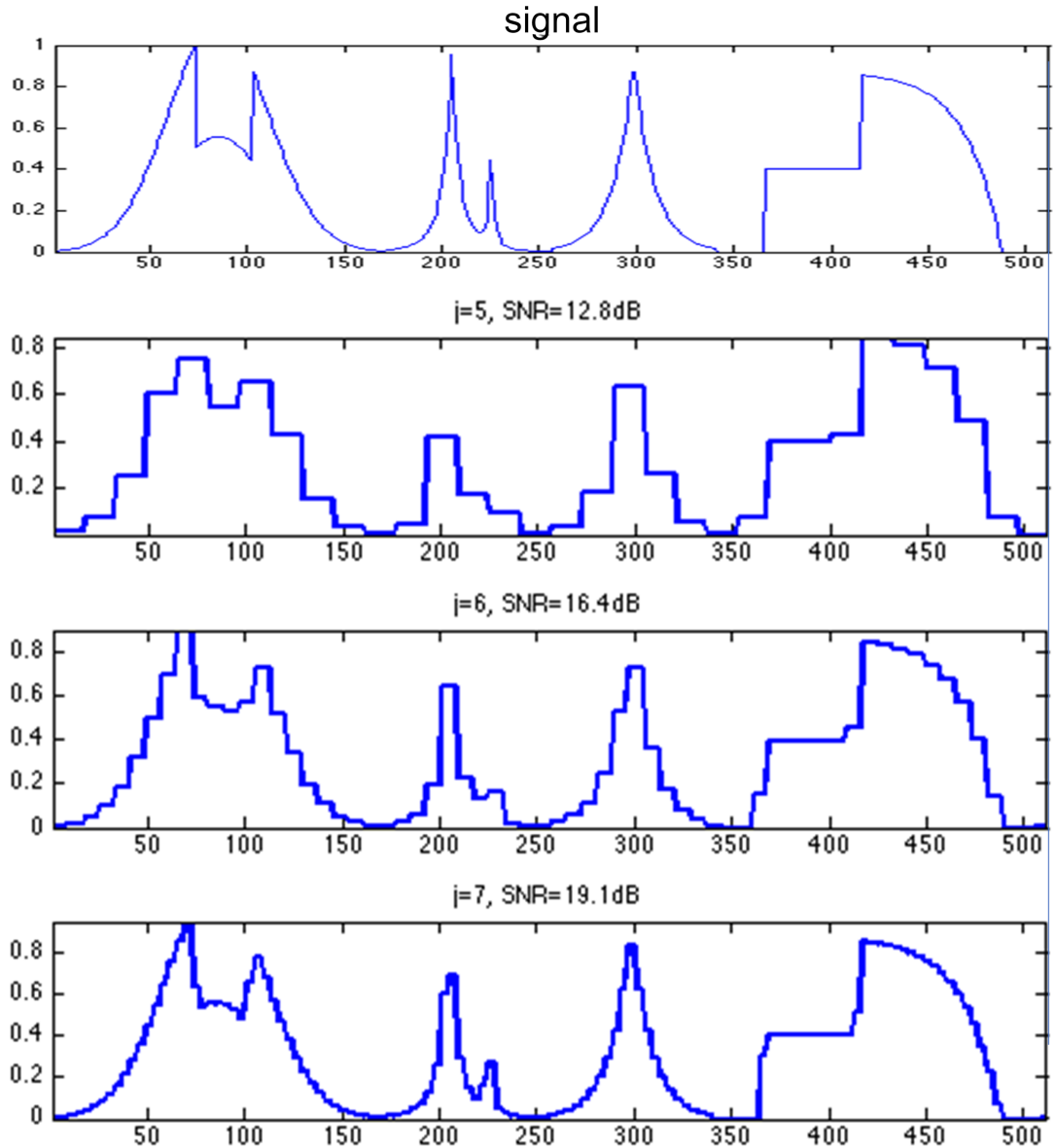


Figure: Haar approximations at different resolution levels

Wavelet representations - Shannon MRA

The scaling function of the Shannon MRA is

$$\phi(x) = \frac{\sin \pi x}{\pi x},$$

whose Fourier transform is $\hat{\phi}(\xi) = \chi_{[-\frac{1}{2},\frac{1}{2}]}(\xi)$.

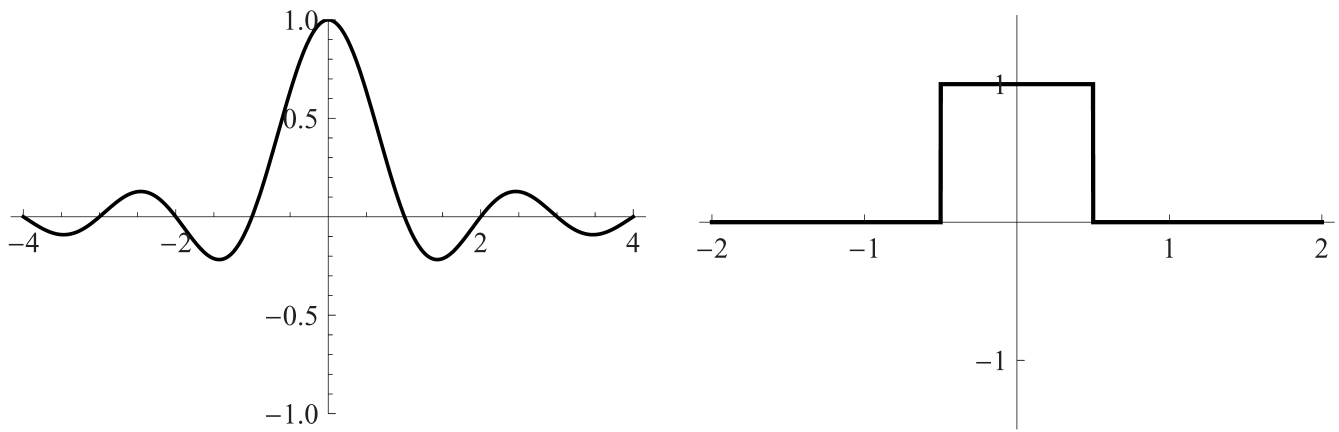


Figure: Left: Shannon scaling function. Right: Its Fourier transform.

Wavelet representations - Shannon MRA

Hence (recalling that $(T_k\phi)^\wedge(\xi) = e^{-2\pi ik\xi}\hat{\phi}(\xi)$)

$$V_0 = \overline{\text{span}}\{T_k\phi : k \in \mathbb{Z}\} = \{f \in L^2(\mathbb{R}) : \text{supp}(\hat{f}) \subset [-\tfrac{1}{2}, \tfrac{1}{2}]\}$$

Dilating, the space V_j contains the functions in $L^2(\mathbb{R})$ that are bandlimited to $[-2^{j-1}, 2^{j-1}]$, that is

$$V_j = \{f \in L^2(\mathbb{R}) : \text{supp}(\hat{f}) \subset [-2^{j-1}, 2^{j-1}]\}$$

It is clear that $V_j \subset V_{j+1}$.
It is a simple consequence of Fourier analysis that $\cup_{j\in\mathbb{Z}} V_j = L^2(\mathbb{R})$ and $\cap_{j\in\mathbb{Z}} V_j = \{0\}$.

This shows that the subspaces $\{V_j\}_{j\in\mathbb{Z}}$ form an MRA.

Wavelet representations - Shannon MRA

Let $\psi(x) = \frac{\sin 2\pi x}{\pi x} - \frac{\sin \pi x}{\pi x}$. Then $\hat{\psi} = \chi_{[-1,1]\setminus[-\frac{1}{2},\frac{1}{2}]}$.
The detail spaces of the Shannon MRA are

$$\begin{aligned} W_j &= \overline{\text{span}}\{D_2^j T_k \psi : k \in \mathbb{Z}\} \\ &= \{f \in L^2(\mathbb{R}) : \text{supp}(\hat{f}) \subset [-2^j, 2^j] \setminus [-2^{j-1}, 2^{j-1}]\} \end{aligned}$$

As for the Haar MRA, the spaces W_j are mutually orthogonal and we have

$$V_{j+1} = V_j \oplus W_j \quad \text{and} \quad V_{j+1} = V_0 \oplus W_0 \oplus \cdots \oplus W_j.$$

Similarly, we have that

$$\{T_k \phi\}_{k\in\mathbb{Z}} \cup \{D_2^j T_k \psi : j \geq 0, k \in \mathbb{Z}\} \text{ and } \{D_2^j T_k \psi : k, j \in \mathbb{Z}\}$$

are each ONBs for $L^2(\mathbb{R})$

Wavelet representations - Shannon MRA

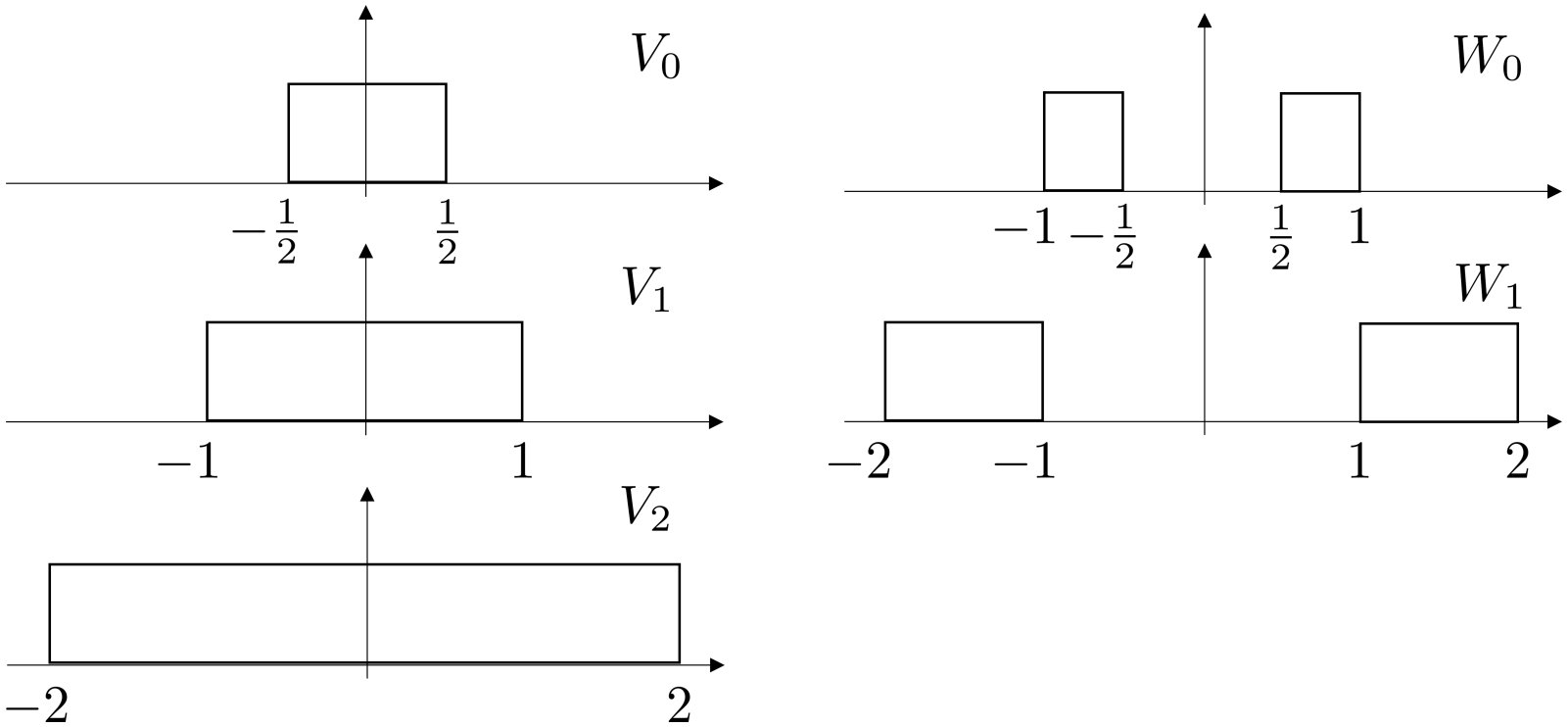


Figure: Illustration of Shannon MRA (Fourier domain).

Wavelet representations - MRA constructions

The Haar and Shannon wavelets are useful to illustrate the idea of multiresolution analysis.

However they are not the best examples for most practical applications.

- The Haar wavelet is discontinuous.
- The Fourier transform of the Shannon is discontinuous.

As a result, neither one is *well-localized*.

The Haar and Shannon examples presented above are not representative of the power of the MRA approach.

Wavelet representations - MRA constructions

• How can we construct a scaling function so that the resulting wavelet has desirable properties such as *compact support*, *regularity*, *vanishing moments*, ...?

The key to using an MRA to construct a wavelet orthonormal basis is the scaling function ϕ.
The scaling function determines V_0, hence V_j. These spaces determine the detail spaces W_j and ultimately the wavelet ψ.

There is a well developed and rather involved mathematical theory to construct ONBs of wavelets with desirable properties .

I will sketch the main ideas.

Wavelet representations - MRA constructions

Definition. A function $\phi \in L^2(\mathbb{R})$ is **refinable** if there exists a sequence of scalars (c_k) such that the series $\sum_{k\in\mathbb{Z}} c_k \phi(2x-k)$ converges in $L^2(\mathbb{R})$ and

$$\phi(x) = \sum_{k\in\mathbb{Z}} c_k \phi(2x-k) \qquad \text{(refinement equation)}$$

The scalars c_k are called the *refinement coefficients*.

Example: Haar scaling function satisfies $\phi(x) = \phi(2x) - \phi(2x-1)$

Proposition. Suppose $\phi \in L^2(\mathbb{R})$ is refinable with refinement coefficients $(c_k) \in \ell^2$. Then

$$\hat{\phi}(\xi) = m_0(\tfrac{\xi}{2})\,\hat{\phi}(\tfrac{\xi}{2}) \quad \text{a.e.,}$$

where $m_0(\xi) = \frac{1}{2}\sum_k c_k e^{-2\pi i k\xi}$

The 1-periodic function m_0 is called the *symbol* (or low-pass filter) of the refinement equation.

Wavelet representations - MRA constructions

Proof (sketch).

$$\phi = \sum_k 2^{-1/2} c_k D_2 T_k \phi$$

Hence

$$\begin{aligned}
\hat{\phi}(\xi) &= \sum_k 2^{-1/2} c_k \, (D_2 T_k \phi)^\wedge(\xi) \\
&= \sum_k 2^{-1/2} c_k \, D_{1/2} M_{-k} \hat{\phi}(\xi) \qquad (M_y f(x) = e^{2\pi i x y} f(x)) \\
&= \sum_k 2^{-1/2} c_k \, 2^{-1/2} e^{-2\pi i k \frac{\xi}{2}} \hat{\phi}(\tfrac{\xi}{2}) \\
&= \frac{1}{2} \left(\sum_k c_k \, e^{-2\pi i k \frac{\xi}{2}} \right) \hat{\phi}(\tfrac{\xi}{2})
\end{aligned}$$

Wavelet representations - MRA constructions

Given a refinable function, we construct an MRA as follows.

Theorem. Assume $\phi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ is a refinable function with refinement coefficients (c_k) and $\{T_k\phi : k \in \mathbb{Z}\}$ is an orthonormal sequence. If we set

$$V_0 = \overline{\text{span}}\{T_k\phi : k \in \mathbb{Z}\} \quad \text{and} \quad V_j = D_2^j V_0, j \in \mathbb{Z},$$

then $\{V_j\}_{j\in\mathbb{Z}}$ is an MRA of $L^2(\mathbb{R})$.
If we set

$$\psi(x) = \sum_k (-1)^{k-1}\overline{c_{1-k}}\,\phi(2x-k)$$

or, equivalently,

$$\hat{\psi}(\xi) = m_1(\tfrac{\xi}{2})\,\hat{\phi}(\tfrac{\xi}{2}) \quad \text{where} \quad m_1(\xi) = e^{-2\pi ik\xi}\overline{m_0(\xi+\tfrac{1}{2})}$$

then $\{D_2^j T_k\psi : j,k \in \mathbb{Z}\}$ is an ONB of $L^2(\mathbb{R})$.

m_1 is called the high-pass filter of the MRA.

Wavelet representations - MRA constructions

The properties of the ONB wavelet basis $\{D_2^j T_k \psi : j, k \in \mathbb{Z}\}$ can be completely determined by appropriately choosing the refinement coefficients (c_k) or the corresponding symbol m_0.

- A wavelet is **compactly supported** iff the corresponding (c_k) is a finite sequence.
- A wavelet has p **vanishing moments** iff $m_0^{(n)}(\frac{1}{2}) = 0$ for $n = 0, \ldots, p-1$.

Definition. A wavelet ψ has p **vanishing moments** if

$$\int_{\mathbb{R}} t^k \, \psi(t) \, dt = 0, \quad \text{for } k = 0, \ldots, p-1.$$

Regularity implies vanishing moments.
Proposition. If $\psi \in C^r(\mathbb{R})$ and $|\psi(x) \leq C \in (1 + |x|)^{-r-1-\epsilon}$ for some $\epsilon > 0$, then ψ has r vanishing moments.

Wavelet representations - MRA constructions

Having p vanishing moments means that ψ is orthogonal to any polynomial up to degree $p-1$.
This implies that, if f is regular and ψ has sufficiently many vanishing moments, then the wavelet coefficients

$$\langle f, \psi_j, k \rangle = \int_{\mathbb{R}} f(t) 2^{1/2} \psi(2^j t - k)\, dt$$

are small at fine scale 2^j.

Explanation: If f is locally C^k, then over a small interval it is well approximated by its Taylor polynomial of degree k. If the number of vanishing moments is $p > k$, then the wavelet coefficients are negligible at fine scales.

Wavelet representations - MRA constructions

Daubechies in 1988 was the first to construct compactly supported ON wavelets with some degree of smoothness.

Daubechies wavelets are chosen to have the **highest number p of vanishing moments**, (this does not imply the best smoothness) **for given support width (number of coefficients)** $2p$.

Note: there is no compactly supported ON wavelet with infinitely many vanishing moments or C^∞ regularity.
There are two naming schemes in use, DN using the length or number of taps, and dbA referring to the number of vanishing moments.

- D2 = db1 = Haar wavelet
- D4 = db2 = Daubechies wavelet with filter length 4 and 2 vanishing moments.

Daubechies wavelets do not have an explicit form of the scaling and wavelet functions (except than the Haar wavelet = D2); in fact, they are not possible to write down in closed form.

Wavelet representations - MRA constructions

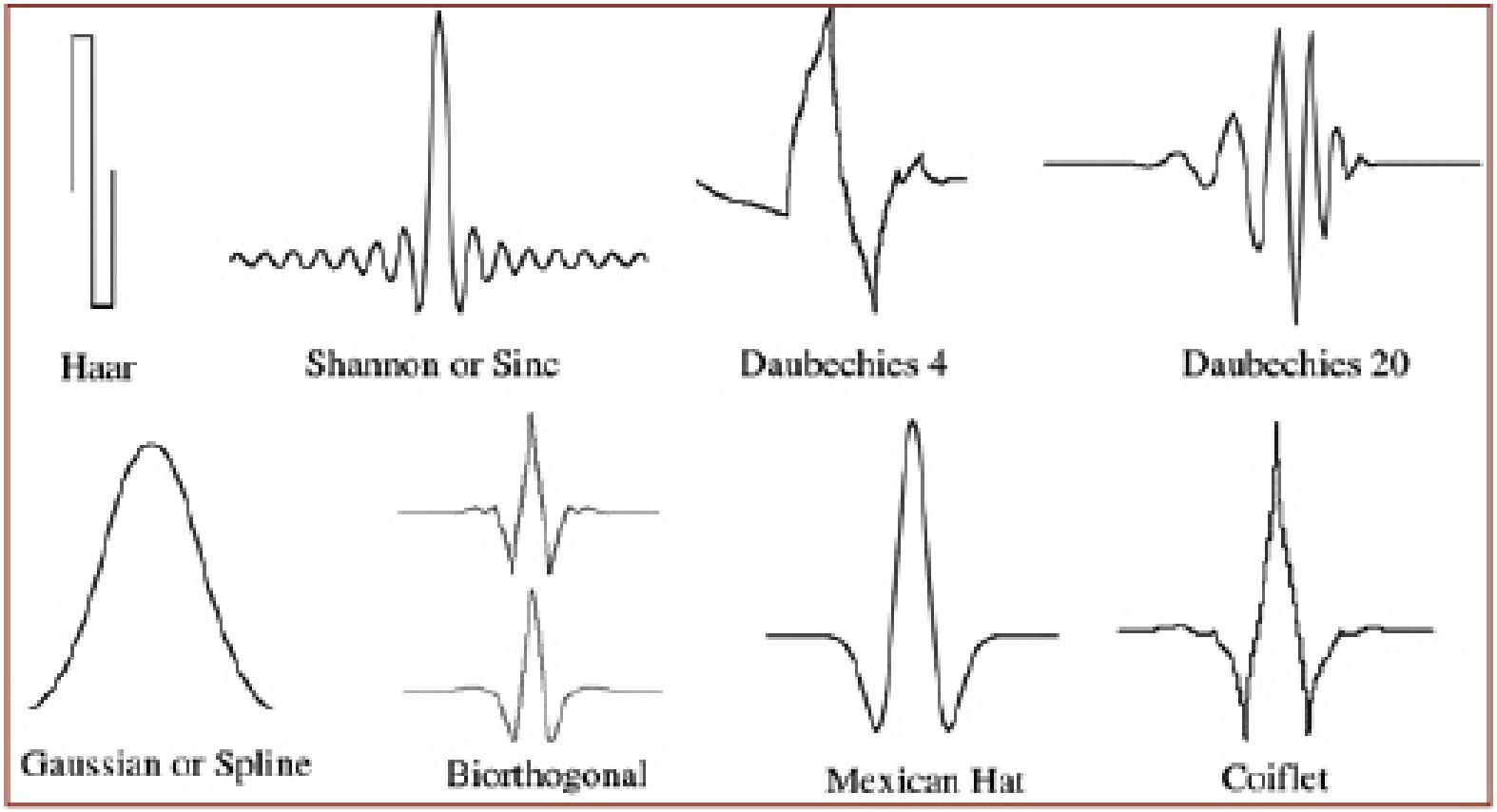


Figure: Examples of wavelets.

Wavelet representations - Approximations

To quantify approximation properties of ON wavelet bases, it is useful to introduce a notion of **non–linear approximation**.

• The M-term non–linear approximation of $f \in L^2$ in a basis $\{\psi_i\}$ is obtained by taking the *largest* M coefficients of a representation:

$$\tilde{f}_M = \sum_{i \in \mathcal{I}_M} \langle f, \psi_i \rangle \, \psi_i, \quad |\mathcal{I}_M| = M,$$

$\mathcal{I}_M$ is the set of indices of **the M largest coefficients** $|\langle f, \psi_i \rangle|$**.**

$\tilde{f}_M$ is obtained by **thresholding:** $\mathcal{I}_M = \{i : |\langle f, \psi_i \rangle| > T(M)\}$.

• The non-linear approximation contrasts with the **linear approximation** of f which is obtained by keeping only the first M coefficients of its expansion:

$$\tilde{f}_M^{(lin)} = \sum_{i=1}^{M} \langle f, \psi_i \rangle \, \psi_i.$$

Wavelet representations - Approximations

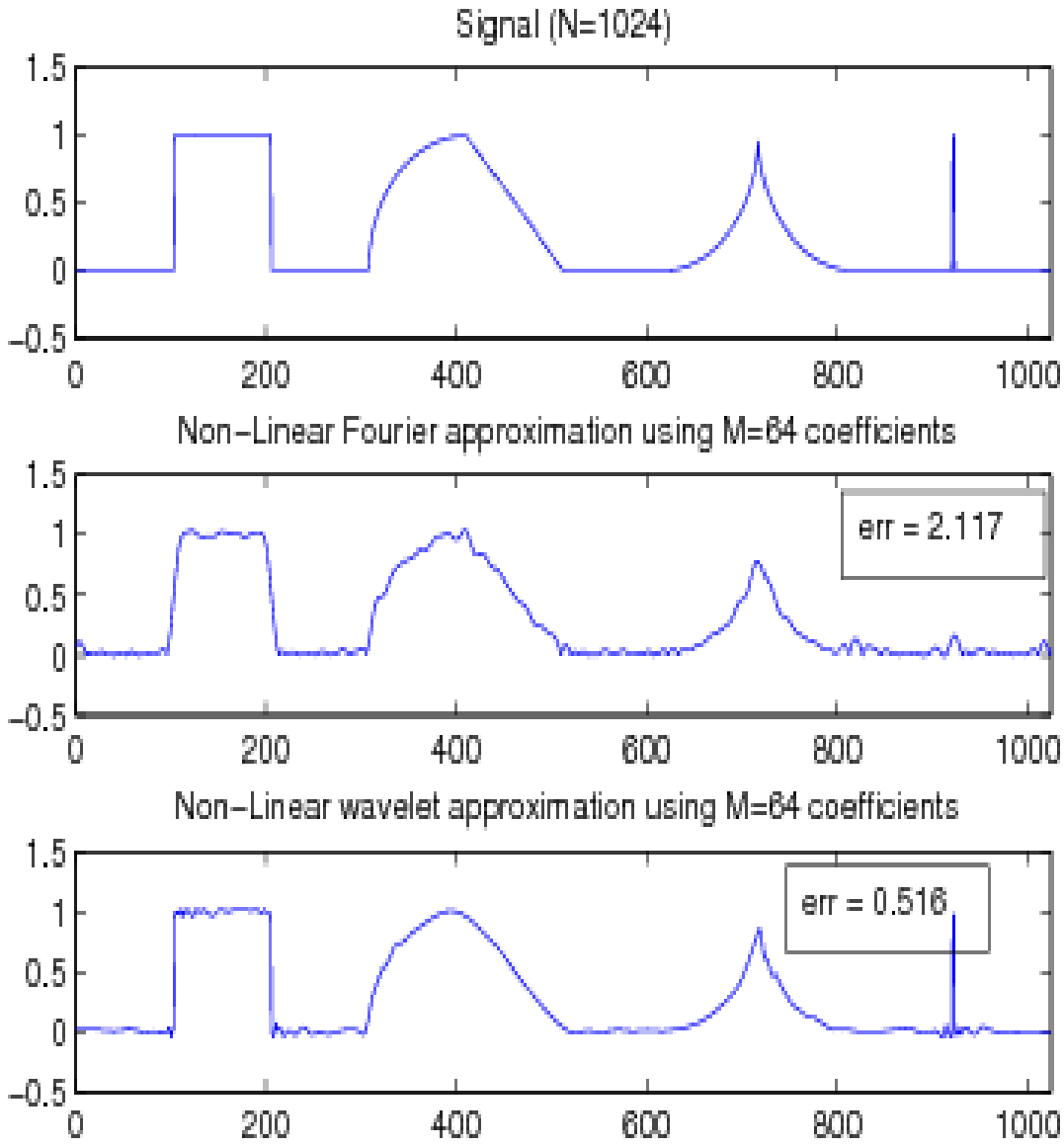
- The **non–linear approximation error** is

$$E^{(n)}(f; M) = \|f - \tilde{f}_M\|^2 = \sum_{i \notin \mathcal{I}_M} |\langle f, \psi_i \rangle|^2.$$

The error depends on the decay rate of the sorted coefficients.

Let us compare Fourier vs. wavelet nonlinear approximation error $E^{(n)}(f; M) = \|f - \tilde{f}_M\|$ for piece-wise regular functions.

Wavelet representations - Approximations



Wavelet representations - Approximations

The numerical example suggests that wavelets provide a *sparser* representation than Fourier series for functions with discontinuities.

In general, if f is $C^{\alpha}(\mathbb{R})$, $\alpha \geq 1$, apart from finitely many discontinuities, the nonlinear approximation error satisfies:

- **Fourier** approximation error:

$$E^{(F)}(f; N) = \|f - \tilde{f}_N^{(F)}\|^2 \leq C\, N^{-1}, \quad N \to \infty$$

- **Wavelet** approximation error:

$$E^{(W)}(f; N) = \|f - \tilde{f}_N^{(W)}\|^2 \leq C\, N^{-2\alpha}, \quad N \to \infty$$

Wavelets do not 'feel' the discontinuity and provide the **optimal approximation** error rate for univariate functions with discontinuities (including BV).

Wavelet representations - Higher dimensions

The theory and numerical example show that wavelets provide a *sparse* representation for *piecewise smooth signals*.

In fact, they provide optimally sparse representations for this class of signals.

Unfortunately, the multi-D situation is more complicated, and (conventional) wavelets do not work as well, even though their approximation properties outperform Fourier methods.

Wavelet representations - Higher dimensions

The simplest way to extend the wavelet construction to $L^2(\mathbb{R}^2)$ is by using a **tensor product**

$$\{\psi_{j,k}(x) = \psi_{j_1,k_1}(x_1)\,\psi_{j_2,k_2}(x_2) : j = (j_1, j_2) \in \mathbb{Z}^2, k = (k_1, k_2) \in \mathbb{Z}^2\}$$

This idea leads to a MRA construction based on separable wavelets whose elements are products of function dilated *at the same scale*.

Give a MRA $\{V_j\}_{j\in\mathbb{Z}}$ of $L^2(\mathbb{R})$, a separable 2-dimensional MRA is composed of tensor product spaces

$$V_j^2 = V_j \otimes V_j, \quad j \in \mathbb{Z}.$$

If ϕ is a scaling of function of the 1-dimensional MRA $\{V_j\}_{j\in\mathbb{Z}}$, using the observation that $\{\phi_{j,m}\}_{m\in\mathbb{Z}}$ is an ONB of V_j, a straightforward argument shows that

$$\{\phi_{j,k}^2(x_1, x_2) = \phi_{j,k_1}(x_1)\phi_{j,k_2}^2(x_2) : k = (k_1, k_2) \in \mathbb{Z}^2\}$$

is an ONB of V_j^2.

Wavelet representations - Higher dimensions

To build a 2-dimensional MRA, we write

$$V_{j+1}^2 = V_j^2 \oplus W_j^2 = (V_j \otimes V_j) \oplus W_j^2 \tag{1}$$

where W_j^2, the orthogonal complement to V_j^2 in V_{j+1}^2 is the detail space at scale j.
Since $V_{j+1} = V_j \oplus W_j$, it follows that

$$\begin{aligned} V_{j+1}^2 &= V_{j+1} \otimes V_{j+1} \\ &= (V_j \oplus W_j) \otimes (V_j \oplus W_j) \\ &= (V_j \otimes V_j) \oplus (V_j \otimes W_j) \oplus (W_j \otimes V_j) \oplus (W_j \otimes W_j) \end{aligned} \tag{2}$$

Combining (1) and (2) we see that

$$W_j^2 = (V_j \otimes W_j) \oplus (W_j \otimes V_j) \oplus (W_j \otimes W_j)$$

Wavelet representations - Higher dimensions

Since $\{\phi_{j,m}\}_{m\in\mathbb{Z}}$ is an ONB of V_j and $\{\psi_{j,m}\}_{m\in\mathbb{Z}}$ is an ONB of W_j, it follows that

$$\{\phi_{j,k_1}(x_1)\psi_{j,k_2}(x_2), \psi_{j,k_1}(x_1)\phi_{j,k_2}(x_2), \psi_{j,k_1}(x_1)\psi_{j,k_2}(x_2)\}_{j,k_1,k_2\in\mathbb{Z}}$$

is an ONB of $L^2(\mathbb{R}^2)$

It is called a **separable 2D wavelet basis**.

Note: a 2d separable wavelet basis of $L^2(\mathbb{R}^2)$ has 3 generators.

Wavelet representations - Higher dimensions

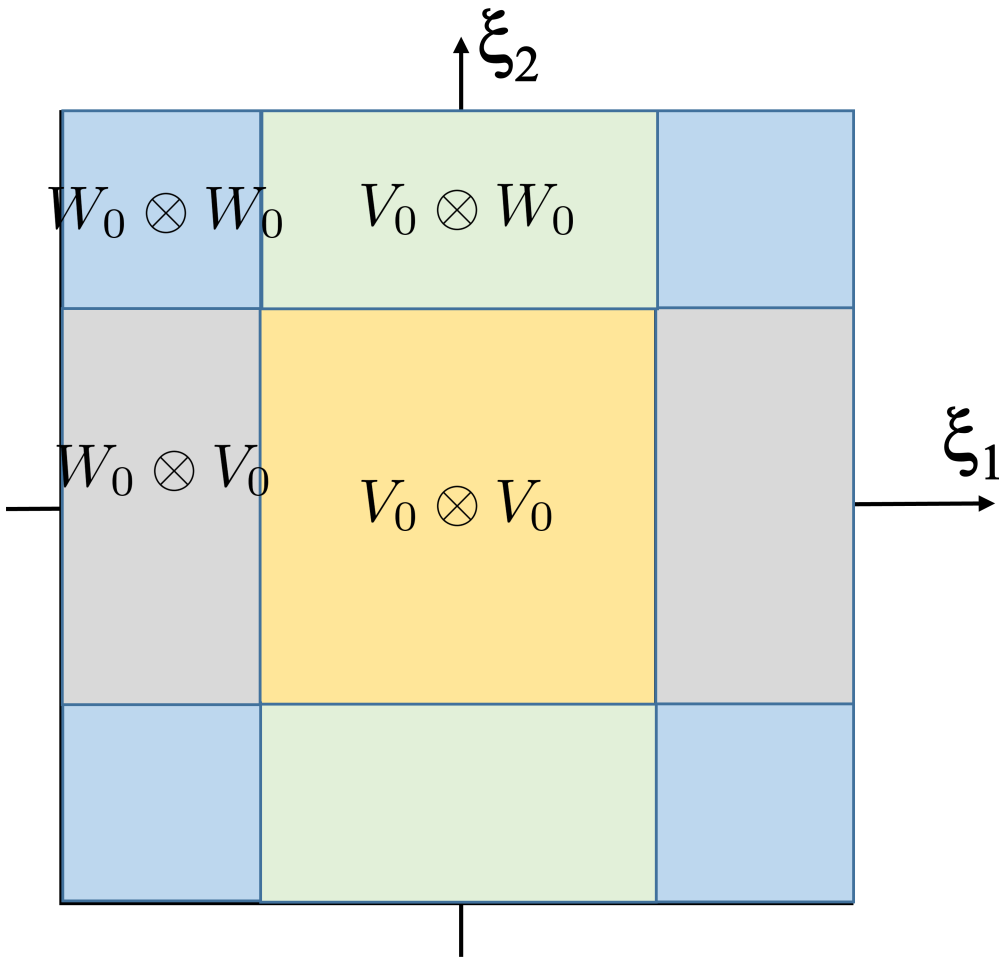


Figure: 2d Shannon decomposition.

$$V_1^2 = (V_0 \otimes V_0) \oplus (V_0 \otimes W_0) \oplus (W_0 \otimes V_0) \oplus (W_0 \otimes W_0)$$

Wavelet representations - Higher dimensions

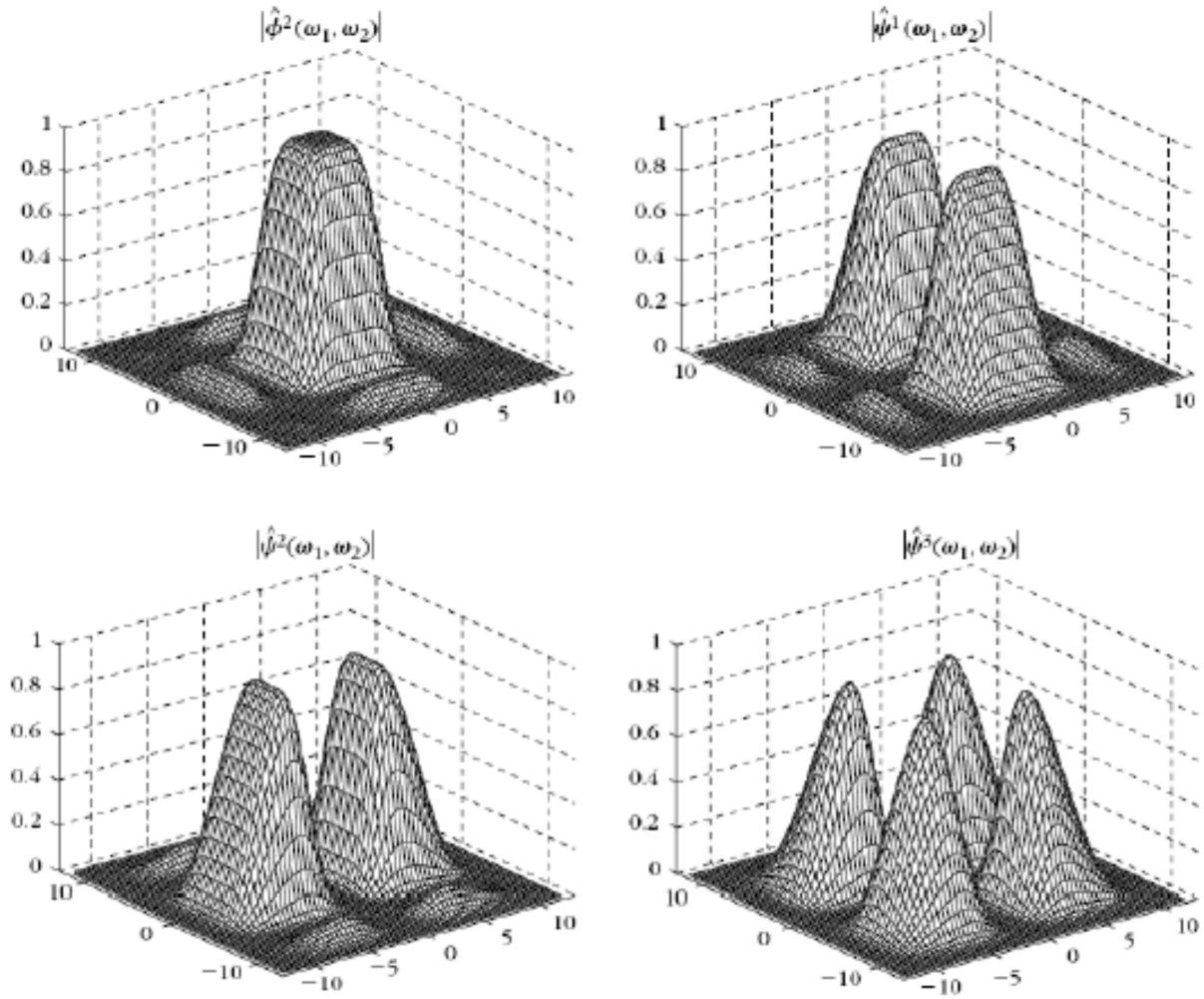
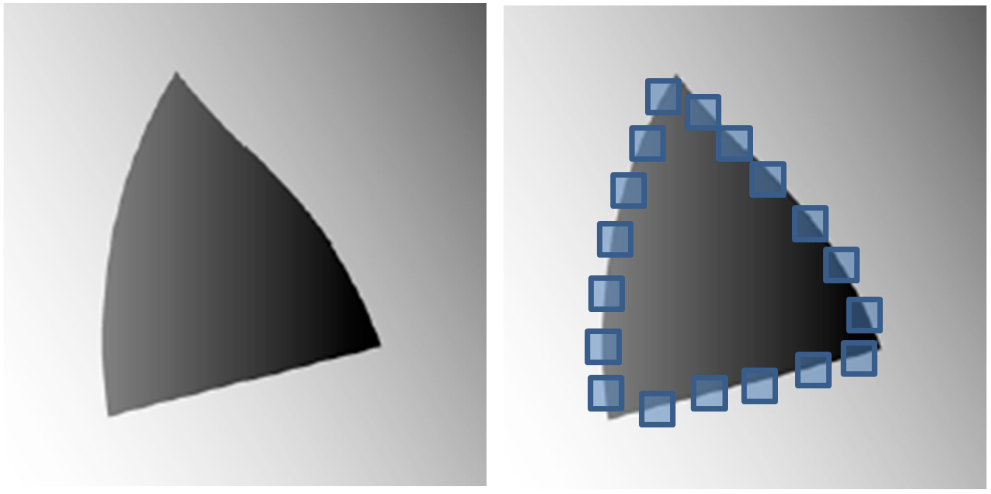


Figure: 2d separable Meyer wavelets in Fourier domain.

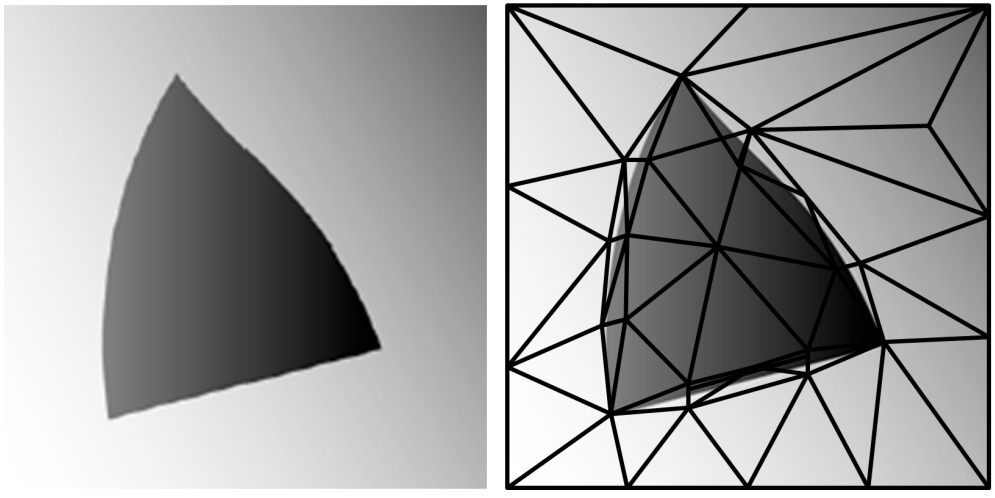
Wavelet representations - Higher dimensions

Separable wavelets are **sub-optimal**.

The problem is that, while wavelets are 'optimal' in handling *point–discontinuities*, in higher dimensions there are other kind of discontinuities, e.g., discontinuities along lines and surfaces.



Many wavelet coefficients are needed to represent discontinuous curves efficiently.



Using adapted triangles one can provide a sparse Approximations, if the curve is regular.

Wavelet representations - Higher dimensions

Consider a *cartoon-like image*, i.e., $f \in C^2(\mathbb{R}^2)$ apart from C^2 edges as in the numerical example.

The estimate of the nonlinear approximation error gives:

- **Fourier** approximation error:

$$E^{(F)}(f;N) = \|f - \tilde{f}_N^{(F)}\|^2 \leq C\, N^{-1/2}, \quad N \to \infty$$

- **Wavelet** approximation error:

$$E^{(W)}(f;N)\|f - \tilde{f}_N^{(W)}\|^2 \leq C\, N^{-1}, \quad N \to \infty$$

- **Theoretical optimal** approximation error [Donoho, 2001]:

$$E^{(T)}(f;N) = \|f - \tilde{f}_N\|^2 \leq C\, N^{-2}, \quad N \to \infty$$

Wavelet representations

Wavelets bibliography:

1. I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.
2. D. L. Donoho, Sparse components of images and optimal atomic decomposition, Constr. Approx. 17 (2001), 353–382.
3. S. Mallat, *A Wavelet Tour of Signal Processing. The Sparse Way*, 3rd Edition, Academic Press, San Diego 2008.

1.3 Shearlets

Wavelet representations - Higher dimensions

Separable wavelets are unable to capture the geometry of edges discontinuities efficiently. To improve upon separable wavelets, one needs analyzing functions with **improved directional sensitivity**.

A number of methods were introduced:

- Ridgelets *(Candès and Donoho; 1999)*
- Complex wavelets *(Kingsbury; 2001)*
- Curvelets *(Candès and Donoho; 2002)*
- Contourlets *(Do and Vetterli; 2002)*
- Wavelets with composite dilations *(Guo, Labate, Lim, Weiss, and Wilson, 2004)*
- Shearlets *(Guo, Kutyniok, and Labate; 2005)*
- Bandlets *(LePennec and Mallat; 2005)*

Wavelet representations - Shearlets

To define a directional version of the wavelet representation in dimensions $n = 2$, we consider an affine-like system of the form

$$\{\psi_{j,\ell,k} = 2^{\frac{3}{4}j}\,\psi(M_{j,\ell}x - k) : j, \ell \in \mathbb{Z}, k \in \mathbb{Z}^2\},$$

where $M_{j,\ell} = \begin{pmatrix} 2^{2j} & 2^j\ell \\ 0 & 2^j \end{pmatrix}$

Note that $M_{j,\ell} = B^\ell A^j$ where

- $B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ is the **shear matrix**;
- $A = \begin{pmatrix} 2 & 0 \\ 0 & \sqrt{2} \end{pmatrix}$ is the **anisotropic dilation matrix**.

This system is called a **shearlet system.**

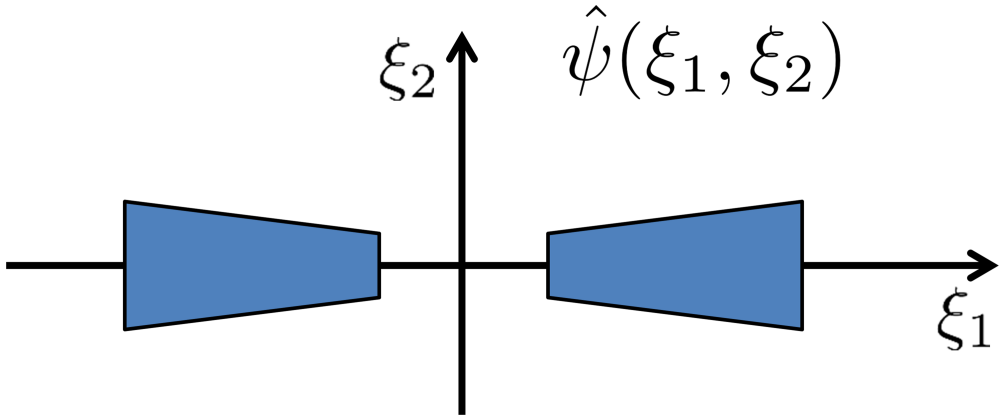
Wavelet representations - Shearlets

We can construct a **well-localized** generator by choosing ψ as:

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

where

- ψ_1 is a wavelet with $\hat{\psi}_1 \in C^\infty(\mathbb{R})$ and $\text{supp}\hat{\psi}_1 \subset [-2, -\frac{1}{2}] \cup [\frac{1}{2}, 2]$,
- ψ_2 satisfies $\hat{\psi}_2 \in C^\infty(\mathbb{R})$ and supp $\hat{\psi}_2 \subset [-1, 1]$.



Wavelet representations - Shearlets

With this choice of generator, the shearlet system

$$\{\psi_{j,\ell,k} = |\det A|^{j/2}\, \psi(B^{\ell}\, A^{j} x - k) : j, \ell \in \mathbb{Z}, k \in \mathbb{Z}^2\},$$

is a **Parseval frame** for $L^2(\mathbb{R}^2)$

That is, for all $f \in L^2(\mathbb{R}^2)$

$$\|f\|^2 = \sum_{j\in\mathbb{Z}} \sum_{\ell\in\mathbb{Z}} \sum_{k\in\mathbb{Z}^2} |\langle f, \psi_{j,\ell,k}\rangle|^2$$

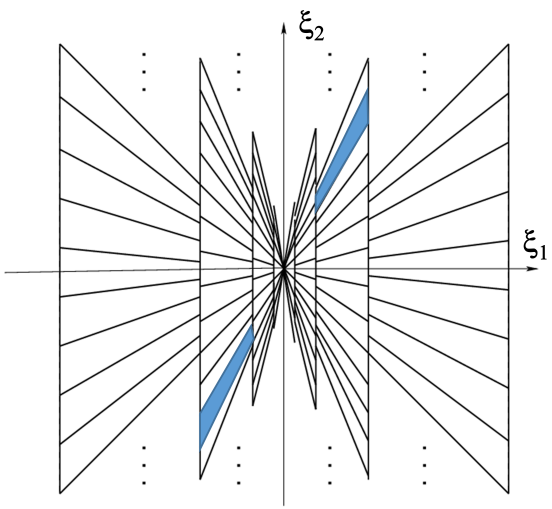
Each element $\psi_{j,\ell,k}$ is associated with a **scale** 2^{-j}, **location** $2^{-j}\, k$ and **orientation** $2^{-j}\ell$.

Wavelet representations - Shearlets

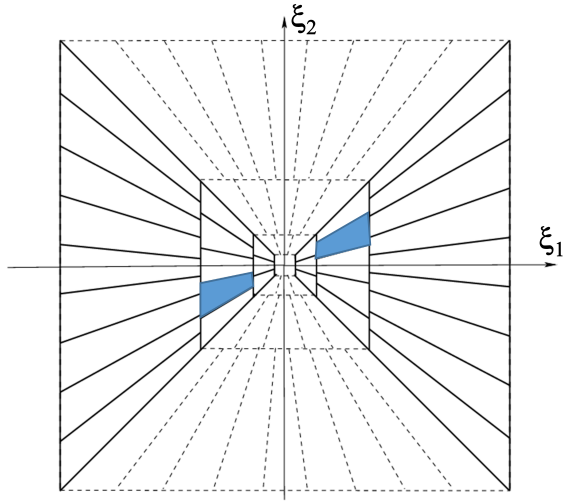
In the Fourier domain, the elements of the shearlet system

$$\hat{\psi}_{j,\ell,k}(\xi) = 2^{-3j/4}\, e^{2\pi i \xi A^{-j} B^{-\ell} k}\, \hat{\psi}_1(2^{-j}\,\xi_1)\, \hat{\psi}_2(2^{j/2}\,\frac{\xi_2}{\xi_1} - \ell)$$

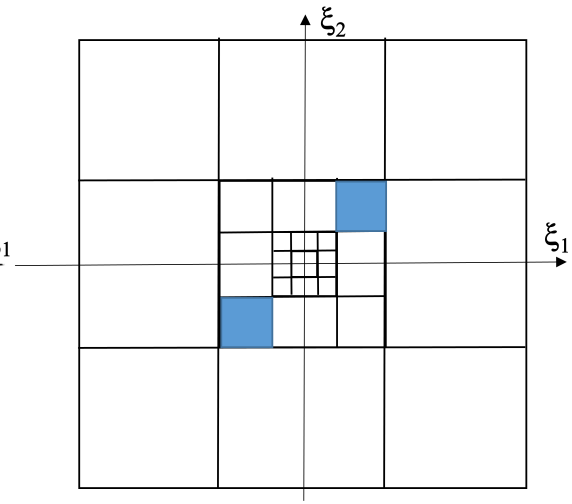
are supported on trapezoids, at various **scales** 2^j, with **orientations** controlled by ℓ.



shearlet frequency tiling



shearlet frequency tiling



wavelet frequency tiling

Wavelet representations - Shearlet approximations

Because of their elongated supports and directionality, shearlets are sparser than wavelets in approximating functions with edge discontinuities

Theorem [Guo, Labate 2006] Let $\tilde{f}_N$ be the approximation of a cartoon-like function f obtained by taking the N largest coefficient in the shearlet expansion. Then:

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

Up to the log-like factor, this is the **optimal** approximation rate.

The curvelets by Candès and Donoho [2002] and the compactly supported shearlet frames of Kutyniok and Lim [2011] have similar sparsity properties.

Wavelet representations - Shearlet approximations

Here is a heuristic (non-rigorous) argument to compare shearlets and shearlets approximations representation:

Consider an image f on $[0,1]^2$, which is smooth apart from a discontinuity along a smooth edges.

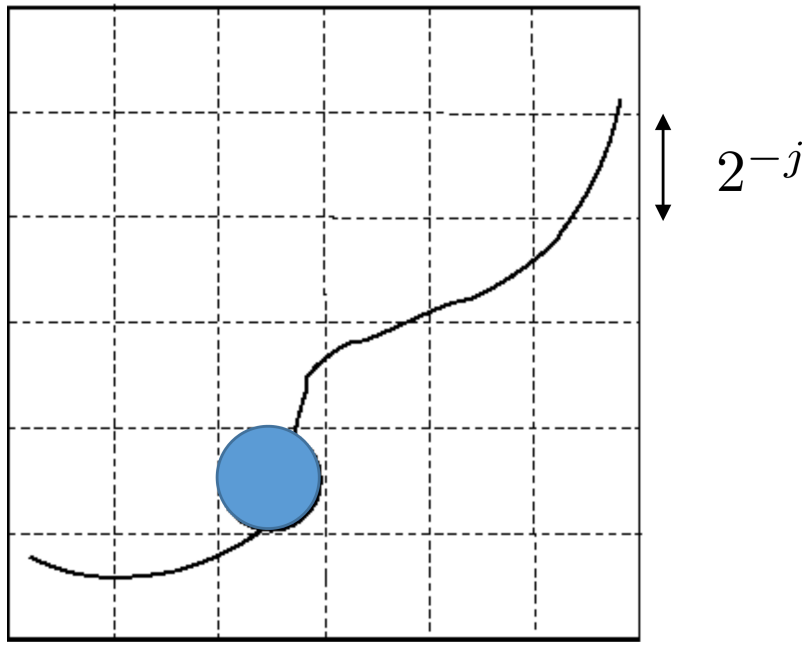
Consider its *wavelet representation*

$$f = \sum_{j} \sum_{k_1=1}^{2^j} \sum_{k_1=1}^{2^j} \langle f, \psi_{j,k} \rangle \, \psi_{j,k}$$

Since f is smooth outside of the edge, essentially all significant coefficients are those associated with the edge.

Wavelet representations - Shearlet approximations

At scale 2^{-j}, there are 2^{2j} wavelet coefficients $\langle f, \psi_{j,k} \rangle$, and $O(2^j)$ of them intersect the edge.

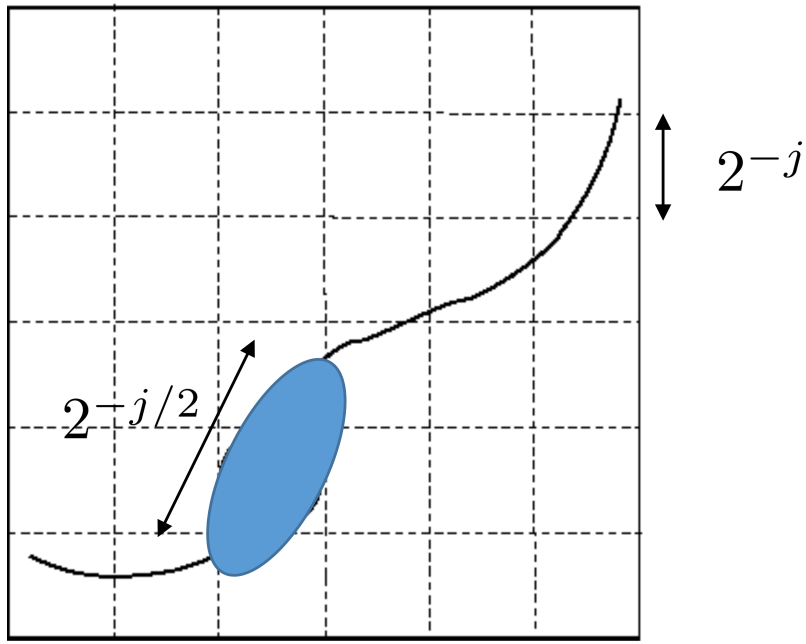


Since $|c_{j,k}| = |\langle f, \psi_{j,k} \rangle| \leq C\, 2^{-j}$, then the $N-$th largest coefficients $|c_{j,k}|_{(N)}$ is bounded by $C\, N^{-1}$ and

$$\|f - f_N^W\|^2 \leq \sum_{m>N} |c_{j,k}|^2_{(m)} \leq C\, N^{-1}$$

Wavelet representations - Shearlet approximations

Let us repeat the calculation for shearlets.
At scale 2^{-j}, there are $O(2^{j/2})$ shearlet s $\psi_{j,\ell,k}$ tangent to the edge. The other elements have negligible impact (at fine scales).



Since $|c_{j,\ell,k}| = |\langle f, \psi_{j,\ell,k}\rangle| \leq C\, 2^{-3j/4}$, the $N-$th largest coefficients $|c_{j,\ell,k}|_{(N)}$ is bounded by $C\, N^{-3/2}$ and

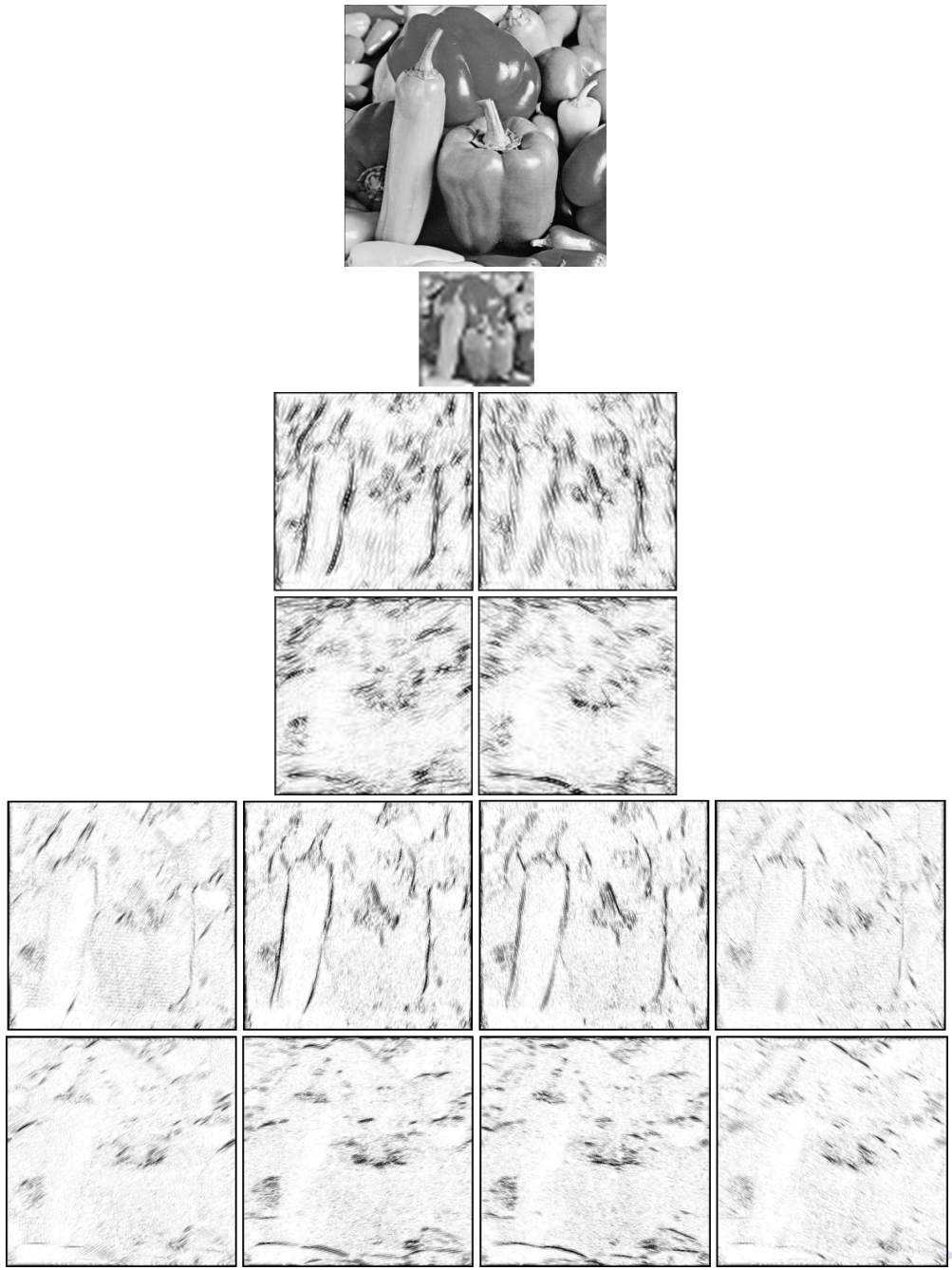
$$\|f - f_N^S\|^2 \leq \sum_{m>N} |c_{j,\ell,k}|^2_{(m)} \leq C\, N^{-2}$$

Wavelet representations - Shearlet approximations

In general, shearlets provide:

- Optimal sparsity for functions with edge discontinuities (similar to curvelets)
- Affine mathematical structure. They are generated from the action of dilations, translations and shear transformations on a single function.
- Generalizations to 3D and similar optimal sparsity property [Guo, Labate, 2010]. However, numerical implementation is computationally intensive in 3D.
- Because of sparsity and directional sensitivity, shearlet representations are useful in image processing applications including image denoising and enhancement, feature detection and inpainting.

Wavelet representations - Shearlet decomposition



Wavelet representations - Shearlets

Shearlet-based image enhancement.

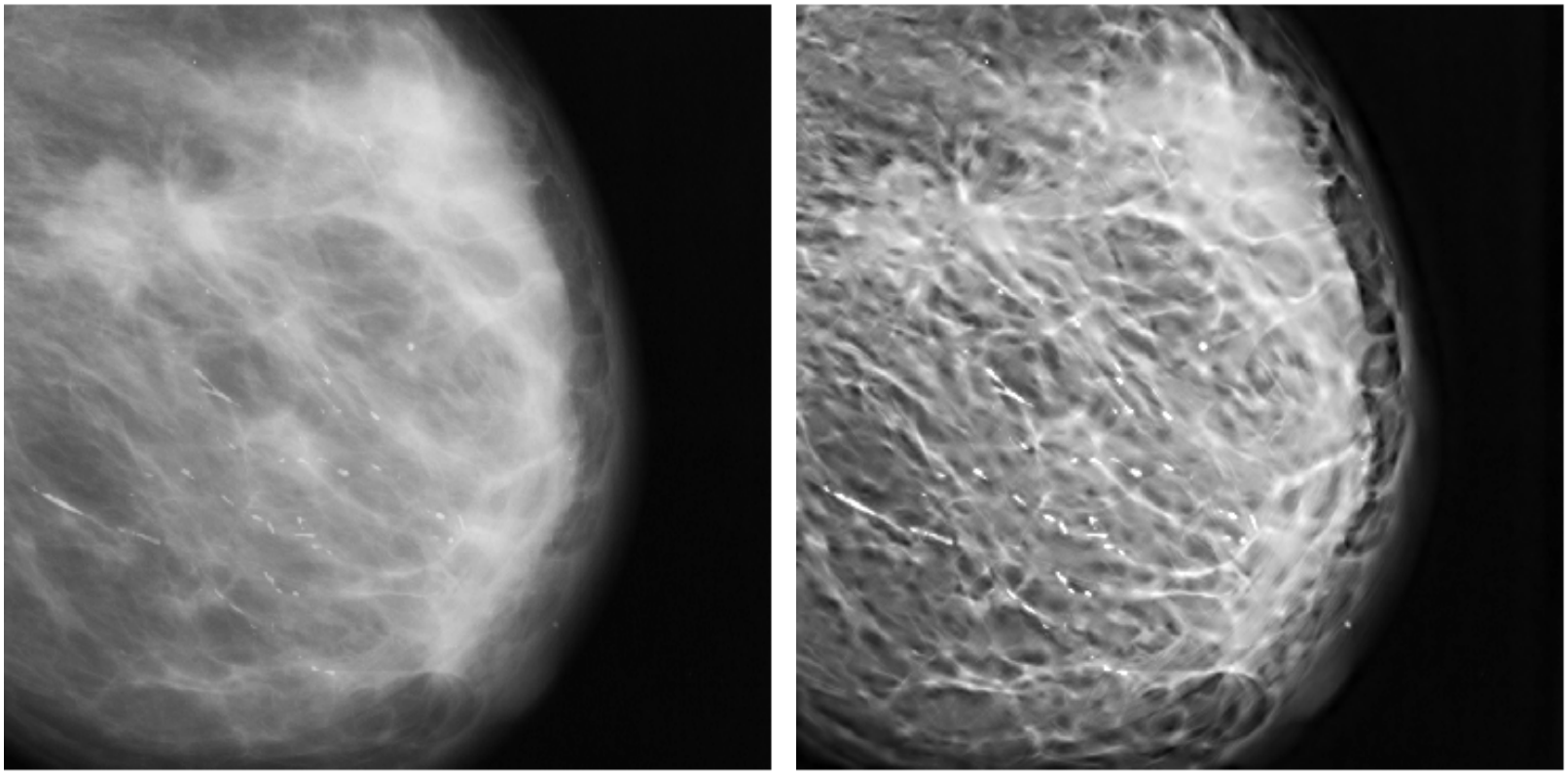


Figure: From left to right. Original mammogram. Enhanced mammogram using a shearlet-based routine.

Wavelet representations - Shearlets

Shearlet-based edge detection.

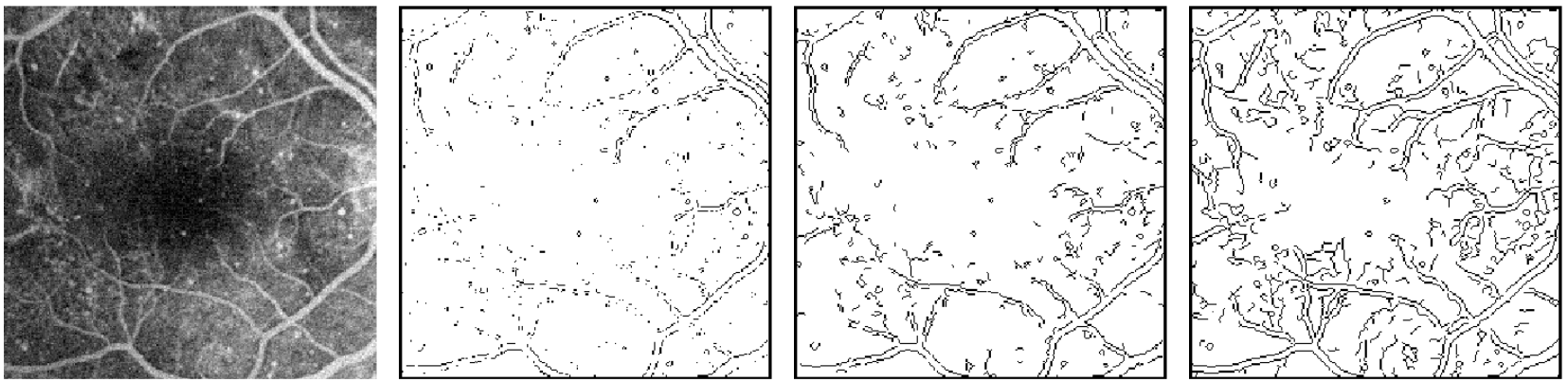


Figure: Comparison of edge detection methods on a retina image. From left to right: Original noisy image (PSNR=24.58 dB), Prewitt (FOM=0.15), Canny (wavelet) (FOM=0.27), shearlet-based algorithm (FOM=0.45). The Figure Of Merit (FOM) measures how close is the reconstruction to the true edge map.

Wavelet representations - Shearlet approximations

Shearlets bibliography:

1. P. Grohs, Optimally Sparse Data Representations, in *Applied and Numerical Harmonic Analysis*, pp. 199-248, Birkhauser, 2015.
2. G. Kutyniok and D. Labate, *Shearlets: Multiscale analysis for multivariate data*, Birkhauser, 2012.
3. K. Guo, D. Labate, W. Lim, G. Weiss, and E. Wilson, Wavelets with composite dilations and their MRA properties, Appl. Comput. Harmon. Anal., 20 , pp. 231-249 (2006).
4. K. Guo and D. Labate, Optimally sparse multidimensional representation using shearlets, SIAM J Math. Anal., 39 pp. 298-318 (2007)
5. K. Guo, and D. Labate, The construction of smooth Parseval frames of shearlets, Math. Model. Nat. Phenom. 8(1) p. 82-105 (2013)

1.4 Wavelet Scattering Transform

Wavelet scattering transform

The **wavelet scattering transform** [Mallat,2012, Mallat & Bruna,2013] was introduced to compute function representations targeted to problems of pattern recognition.

As I will show below, this transform can be implemented using a multilayered structure consisting of **convolutional filters** and **nonlinearities**, which is reminiscent of convolutional neural networks.

Unlike neural networks though where the coefficients or filters are learned during training, here the **filters are fixed**.
In fact, this is a transform mapping $f \in L^(\mathbb{R}^2)$ into appropriate representation coefficients

Wavelet scattering transform

Recall that the wavelet transform

$$\mathcal{W}_\psi : f \mapsto \mathcal{W}_\psi f(j,k) = (f * \tilde{)}\psi_j(2^{-j}k)$$

(where $\tilde{\psi}_j(\cdot) = 2^{j/2}\overline{\psi(-2^j\cdot)}$), maps a function f into a set of wavelet coefficients at multiple scales and location.

The wavelet coefficients encode relevant information of f, hence, can be applied as features for problems of pattern recognition, e.g., classification).

Key features of the wavelet scattering transform

- It extracts **locally translation invariant**, **stable** features.
- It is implemented through a cascade of **wavelet filters** and **modulus operators** over multiple layers (cf. multi-layer convolution network)

Invariant features

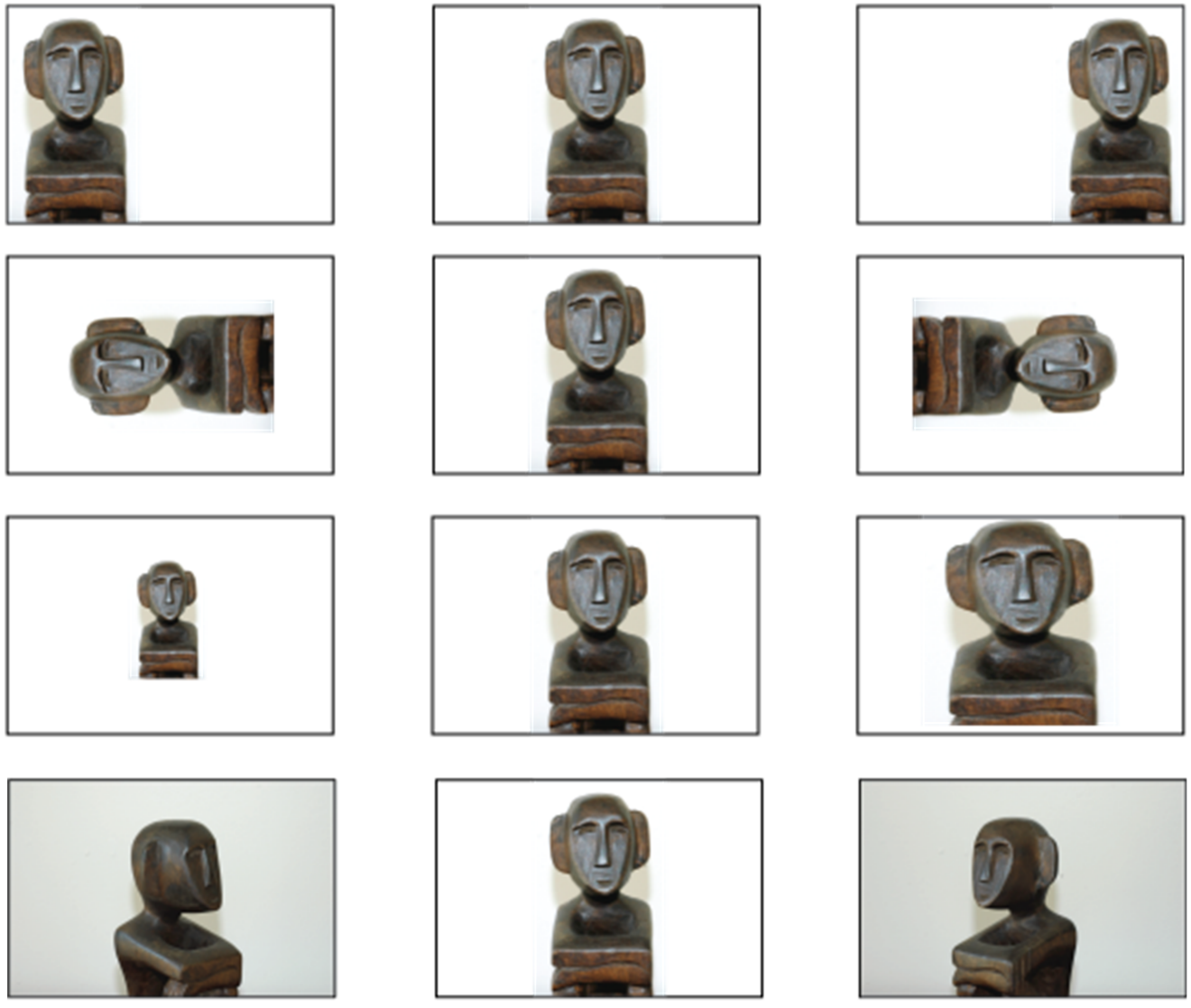


Figure: Object detection and retrieval algorithms: one major challenge is to handle variations in position, angle, scale and viewpoint.

Invariant and covariant features

Definition. An operator Φ on $L^2(\mathbb{R}^d)$ is **translation-invariant** if

$$\Phi(T_y f) = \Phi(f), \quad \forall f \in L^2(\mathbb{R}^d), \forall y \in \mathbb{R}^d;$$

Φ is **translation-covariant** if

$$\Phi(T_y f) = T_y \Phi(f), \quad \forall f \in L^2(\mathbb{R}^d), \forall y \in \mathbb{R}^d.$$

Proposition: The Fourier transform modulus $\Phi(f) = |\hat{f}|$ is translation invariant.

Proof. For any $y \in \mathbb{R}^d$, we have

$$T_y f(x) = f(x - y), \quad (T_y f)^\wedge(\omega) = e^{2\pi i \omega y} \hat{f}(\omega).$$

Hence

$$|(T_y f)^\wedge(\omega)| = |\hat{f}|.$$

Invariant features

Deformations are more difficult to handle.

We can model a **deformation** of f as

$$L_\tau f(x) = f(x - \tau(x)),$$

where τ is a differentiable map.

As a result of the dependence on x of τ, the Fourier transform modulus is **not deformation-invariant.**

For example, if $\tau(x) = -\alpha x$, $0 < \alpha < 1$ (scaling deformation), then

$$(L_\tau f)^\wedge(\omega) = \frac{1}{1+\alpha}\,\hat{f}\left(\frac{\omega}{1+\alpha}\right)$$

which may create instability at higher frequency (it increases the high frequency components of f).

Invariant features

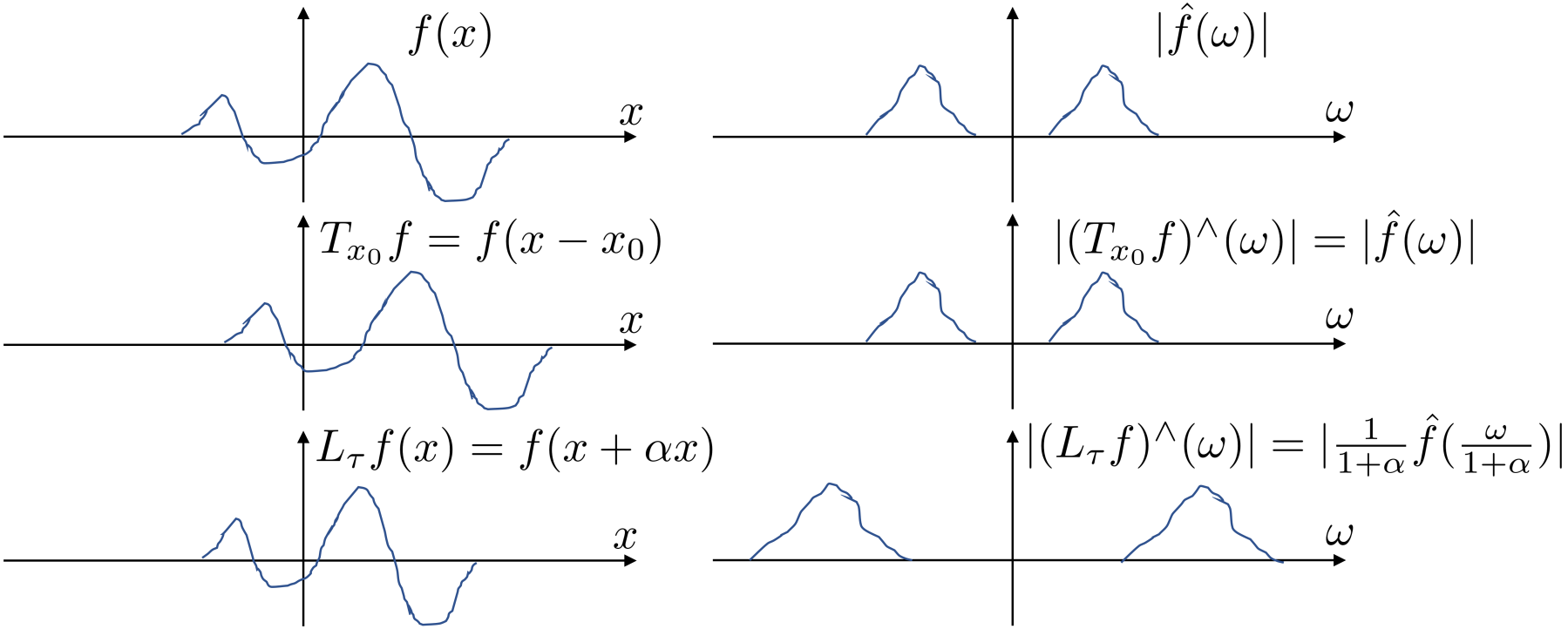


Figure: Impact of translation and scaling deformation.

Invariant features

The distance between 1 and $1-\tau$ over any compact subset S of $\mathbb{R}^d$ is defined as

$$d_S(1, 1-\tau) = \sup_{x \in S} |\tau(x)| + \sup_{x \in S} |\nabla \tau(x)|,$$

where $|\tau(x)|$ is the Euclidean norm on $\mathbb{R}^d$ and $|\nabla \tau(x)|$ measures the deformation amplitude at x.

Definition. A translation invariant operator Φ is **Lipschitz continuous** to the action of C^1 diffeomorphisms if, for any compact $S \subset \mathbb{R}^d$, there exits a constant C such that, for all $f \in L^2(\mathbb{R}^d)$ supported in S and all $\tau \in C^1(\mathbb{R}^d)$,

$$\|\Phi(f) - \Phi(L_\tau f))\| \leq C \|f\| \left(\sup_{x \in S} |\nabla \tau(x)| \right)$$

Wavelet scattering transform

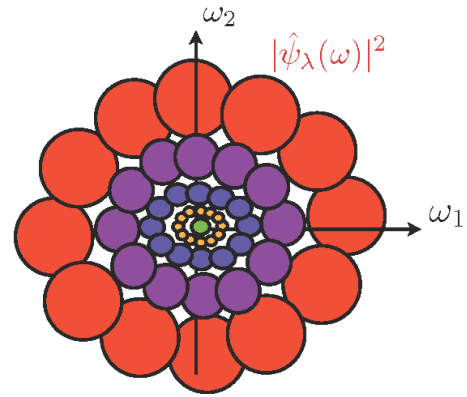
How to build scattering wavelets?

We start with two-dimensional multiscale directional wavelets.

Littlewood-Paley wavelets include both dilations and rotations by elements in a finite rotation group G.

$$\psi_\lambda(x) = 2^j \psi(2^j r x), x \in \mathbb{R}^2$$

with $\lambda = 2^j r$, $j \in \mathbb{Z}, r \in G$.



$$W[\lambda]f(x) = f * \psi_\lambda(x) = \int f(u)\, \psi_\lambda(x-u)\, du$$

By the properties of convolution, the wavelet transform is **translation covariant**:

$$W[\lambda](T_y f)(x) = W[\lambda]f(x-y) = T_y W[\lambda]f(x)$$

Wavelet scattering transform

Let $\Lambda_J = \{\lambda = 2^j r : j \geq J, r \in G\}$ and ϕ be a scaling function such that

$$|\hat{\phi}_J(\omega)|^2 + \sum_{\lambda \in \Lambda_J} |\hat{\psi}_\lambda(\omega)|^2 = 1$$

The **Littlewood-Paley wavelet transform** is defined as

$$\mathcal{W}_J f = \{f * \phi_J, f * \psi_\lambda, \lambda \in \Lambda_J\}$$

Here $f * \phi_J$ covers the low-frequency range which is not covered by the elements $f * \psi_\lambda$, $\lambda \in \Lambda_J$.

Because we can choose a mother wavelet that is regular and localized, the wavelet transform $\mathcal{W}_J$ is Lipschitz-continuous under the action of diffeomorphisms. However, a wavelet transform is not invariant to translations.
The goal is to build translation-invariant coefficients while maintaining stability under actions of diffeomorphisms.

Wavelet scattering transform

Basic idea: to compute translation invariant wavelet coefficients, which remain stable to the action of diffeomorphisms without losing high frequency information.

The first step is to compute wavelet coefficients

$$W[\lambda]f(x) = f * \psi_\lambda(x) \qquad \text{(Lipschitz-continuous)}$$

To achieve translation invariance, we take $U[\lambda]f = |f * \psi_\lambda|$ and then integrate

$$\int U[\lambda]f(x)\,dx = \int |f * \psi_\lambda(x)|\,dx \qquad \text{(translation invariant)}.$$

However, this operation removes the high frequencies of $|f * \psi_\lambda|$.
To recover these frequencies, we take $U[\lambda]f * \psi_{\lambda'} = |f * \psi_\lambda| * \psi_{\lambda'}$.
To achieve translation invariance, again we take

$$\int U[\lambda']U[\lambda]f(x)\,dx = \int ||f * \psi_\lambda| * \psi_{\lambda'}(x)|\,dx$$

We repeat this process.

Wavelet scattering transform

Definition. An ordered sequence $p = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ with $\lambda_k \in 2^{\mathbb{Z}} \times G$ is called a **path**. The empty path is $p = \varnothing$.
For $f \in L^2(\mathbb{R}^2)$ and $\lambda \in 2^{\mathbb{Z}} \times G$, let

$$U[\lambda]f = |f * \psi_\lambda|.$$

A **scattering propagator** is path-ordered product of $U[\lambda]$ operators:

$$U[p] = U[\lambda_m] \cdots U[\lambda_2]U[\lambda_1]$$

with $U[\varnothing] = I$.

The scattering propagator of f is a cascade of convolutions and modulus operators

$$U[p]f = |\cdots|f * \psi_{\lambda_1}| * \psi_{\lambda_2}| \cdots * \psi_{\lambda_m}|$$

Note: each $U[\lambda]$ filters the frequency component in the band associated with ψ_λ and maps it to lower frequencies through the modulus operator.

Wavelet scattering transform

Definition. Let $J \in \mathbb{Z}$ and P_J be a set of finite paths $p = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ with $\lambda_k \in \Lambda_J \times G$. A **windowed scattering transform** is defined for all $p \in P_J$ as

$$S_J[p]f(x) = U[p]f * \phi_J(x) = \int U[p]f(u)\,\phi_J(x-u)\,du$$

Hence

$$S_J[p]f(x) = |\cdots|f * \psi_{\lambda_1}| * \psi_{\lambda_2}|\cdots * \psi_{\lambda_m}| * \phi_J(x)$$

The convolution with ϕ_J localizes the windowed scattering transform over spatial domains of size proportional to 2^J.

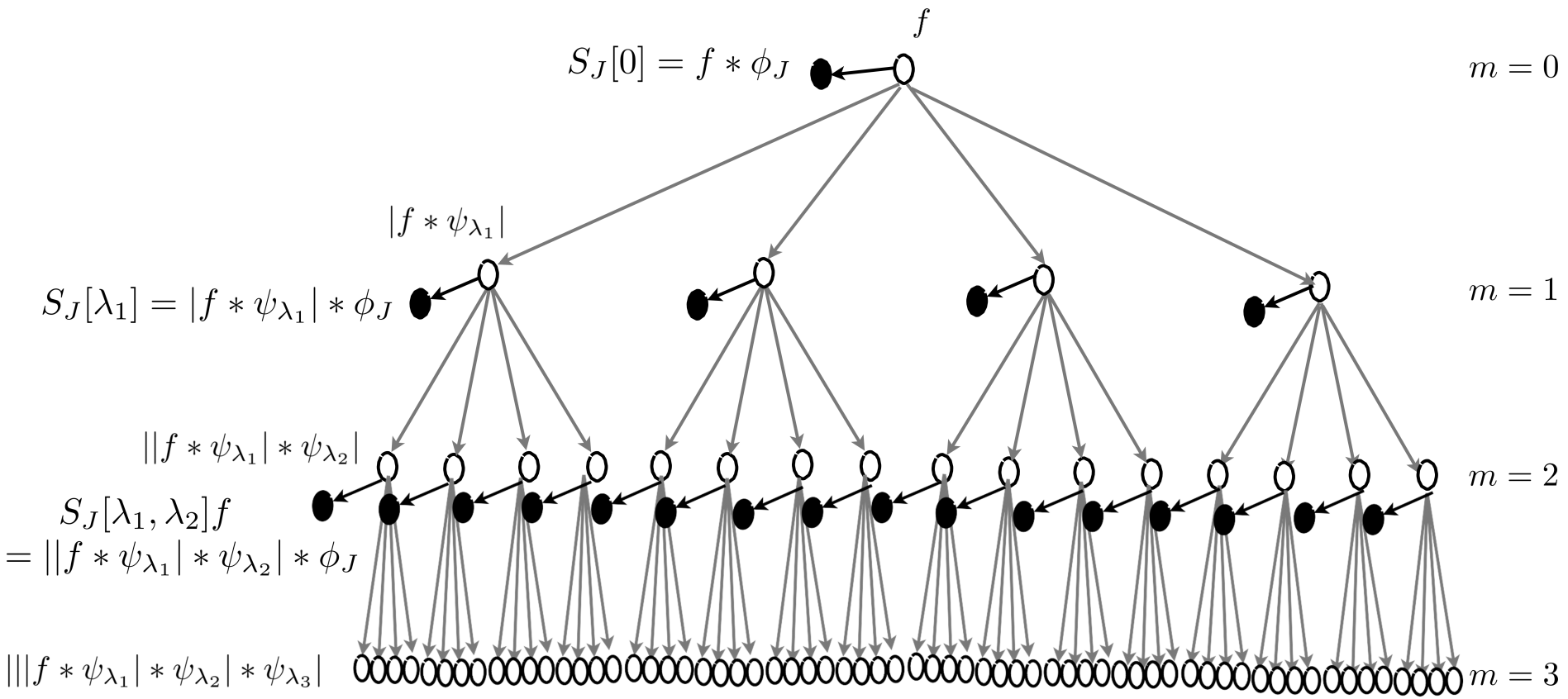
We define a countable family of functions indexed by P_J

$$S_J[P_J]f = \{S_J[p]f : p \in P_J\}$$

Scattering Convolution Network

The wavelet scattering transform is obtained by recursively applying convolutions with the low pass filter ϕ_J to each $U[p]f$ along paths of length $m \leq m_{max}$, starting with $U[0]f = f$.

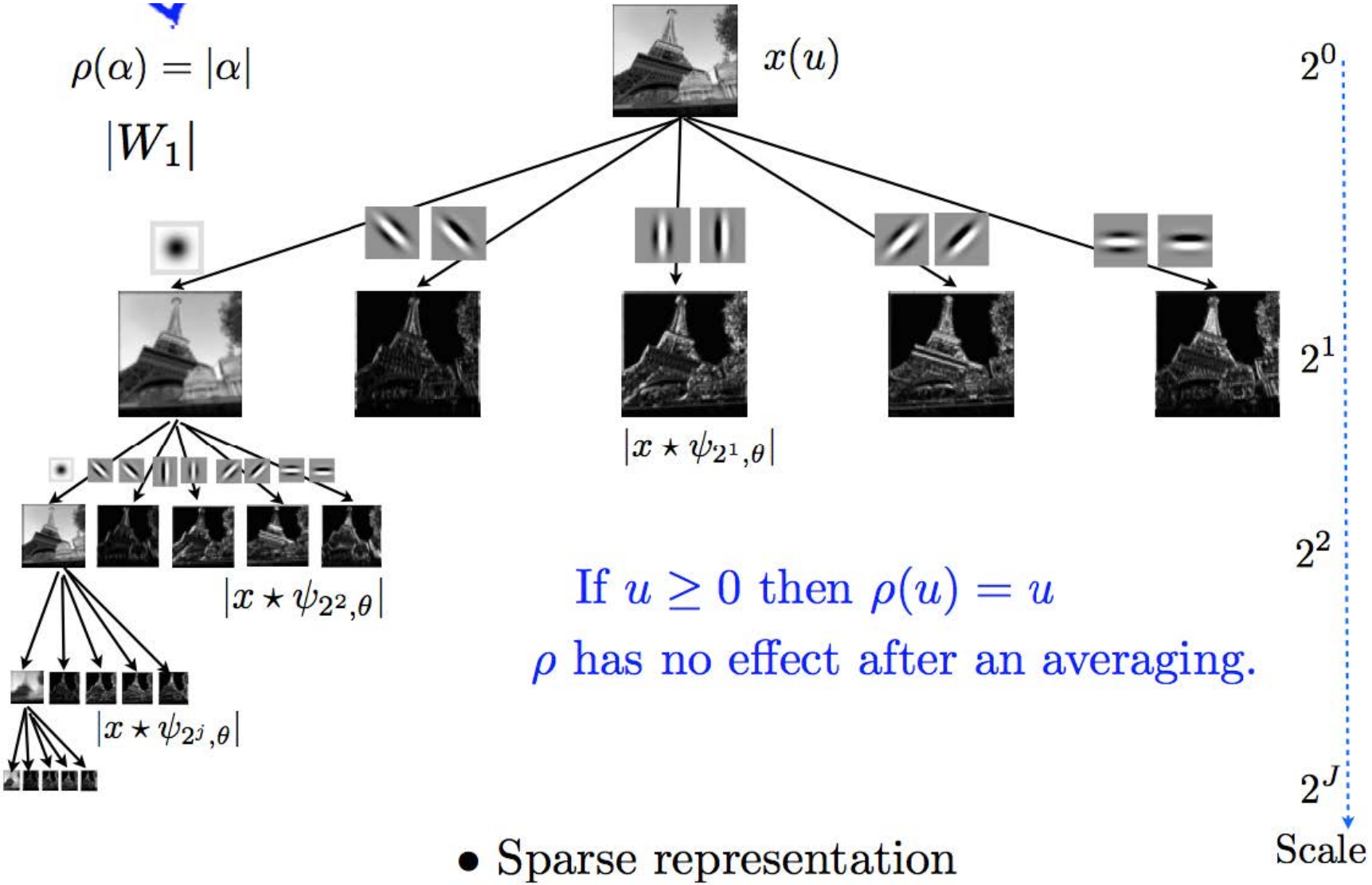
This results in the **Scattering Convolution Network**:



Scattering Convolution Network

- While the Scattering Convolution Network share the hierarchical structure of a Convolutional Neural Network, the filters are predefined wavelet filters. They are not learned.
- A scattering network outputs coefficients $S_J[p]$ at all layers $m = 0, 1, \ldots, m_{max}$.
- Scattering coefficients are locally translation invariant and stable to deformations.
- The scattering coefficients are interpretable and theoretically justified

Scattering Convolution Network



If $u \geq 0$ then $\rho(u) = u$

ρ has no effect after an averaging.

- Sparse representation

Wavelet scattering transform

Mathematical properties of the wavelet scattering transform.

- **Contractivity.** For every $f, g \in L^2(\mathbb{R}^2)$,

$$\|S_J[P_J]f - S_J[P_J]g\| \leq \|f - g\|.$$

- **Energy conservation.** For every $f \in L^2(\mathbb{R}^2)$ and for appropriate wavelets,

$$\|S_J f\|^2 = \sum_m \sum_{p \in \Lambda_J^m} \|S_J[p]f\|^2 = \|f\|^2$$

- **Stability to deformation.** Let $L_\tau f(x) = f(x - \tau(x))$ with $\|\nabla\tau\|_\infty < 1$ then, for $J > \log \frac{\|\tau\|_\infty}{\|\nabla\tau\|_\infty}$,

$$\|S_J f - S_J(L_\tau f)\| \leq C\, m_{max}\, \|f\|\, \|\nabla\tau\|_\infty$$

Wavelet scattering transform

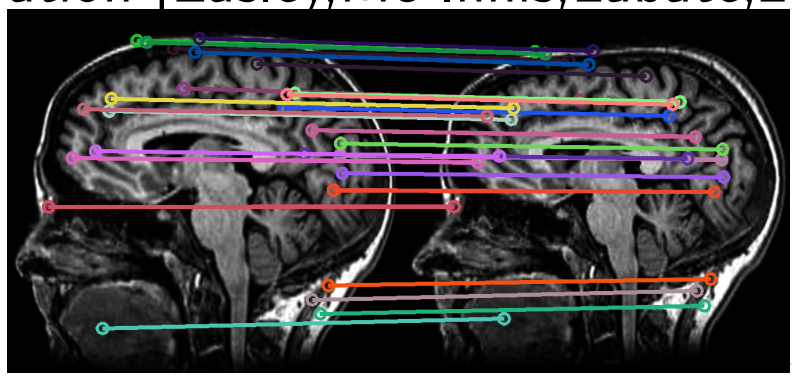
The Wavelet Scattering Transform is useful to generates image features

- invariant to local translations and stable to small deformations;
- Other invariances, e.g., rotation and affine invariance, can be built into this approach [Sifre, Mallat, 2014].

Multiple applications including:

- texture classification [Sifre,Mallat,2014]

- image registration [Easley,Mc-Innis,Labate,2015]



Wavelet scattering transform

Wavelet scattering transform bibliography

1. J. Bruna, S. Mallat, Invariant scattering convolution networks. IEEE transactions on pattern analysis and machine intelligence, 35(8) pp. 1872-1886 (2013).
2. S. Mallat, Group Invariant Scattering, Communications on Pure and Applied Mathematics, Vol. 65, pp. 1331-1398 (2012).
3. Scattering transform resources (S. Mallat): `https://www.di.ens.fr/data/scattering/`
4. `https://www.kymat.io/` Python package that implements wavelet scattering, leveraging the PyTorch framework.