Course on

High Dimensional

Data Analysis

Stéphane Mallat

École Normale Supérieure
www.di.ens.fr/data/scattering

High Dimensional Data

• Tremendous increase of data acquistion: audio, images, video medical/biological data, industrial processes, social networks...

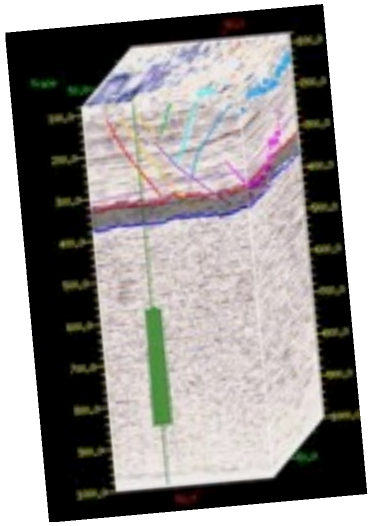
$$\text{Numerical data: } x \in \mathbb{R}^d \text{ with } d \geq 10^6$$

• Automatic analysis becomes critical for industries, science medecine, Internet search, new services.

Audio

Video phones

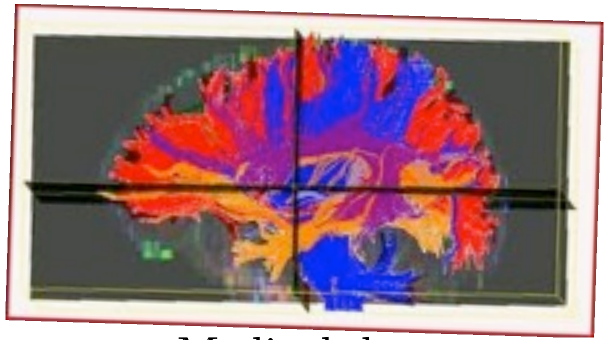
Electrocardiogram
Multiple Sensors



Seismic data

Satellite images

HD Television



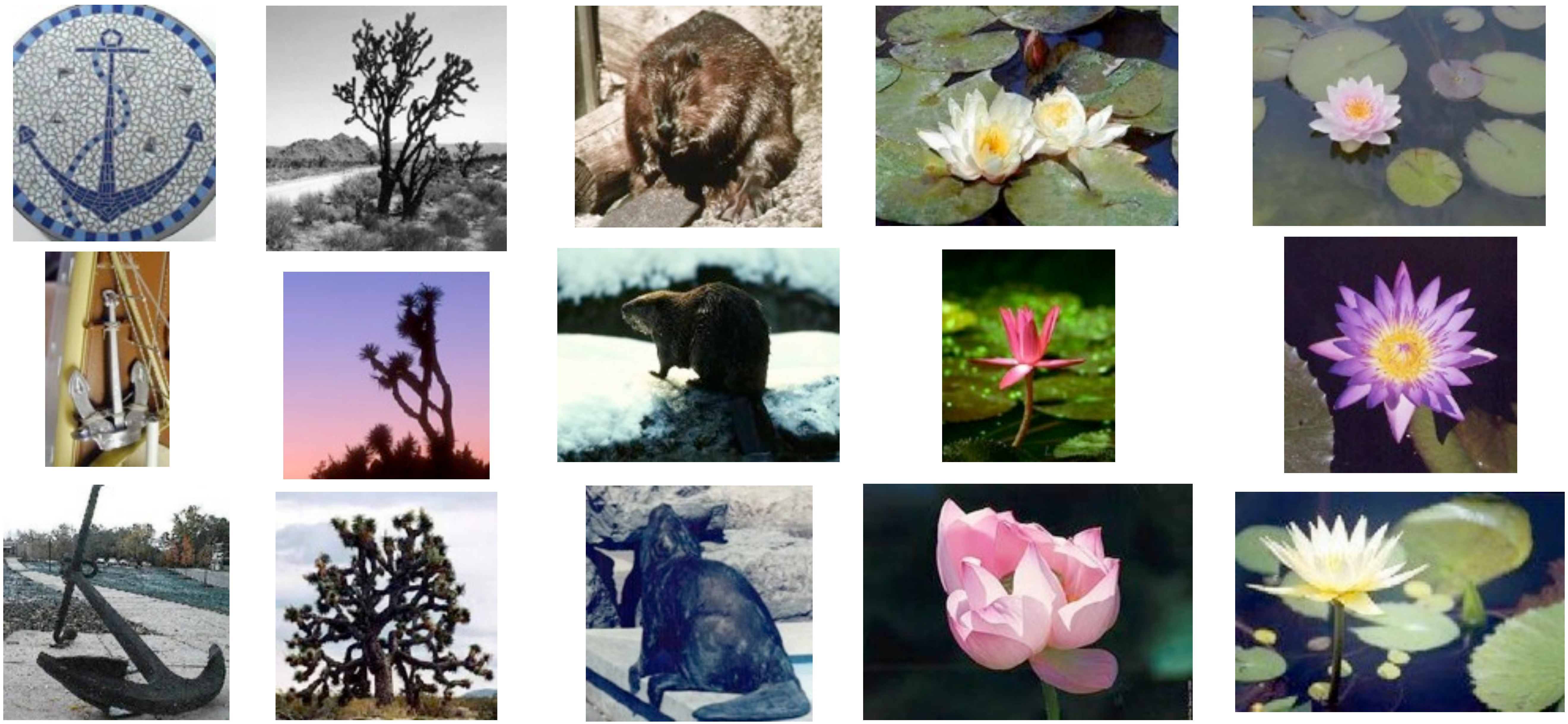
Medical data

• Needs geometry, harmonic analysis, probability and statistics.

High Dimensional Classification

CalTech 101

Anchor | Joshua Tree | Beaver | Lotus | Water Lily



- Considerable variability in each class.
- Euclidean distances are meaningless
- Need to find **discriminative invariants**.

Problems/Datasets in Computer Vision

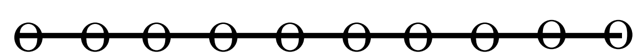
- Imagenet
  - 14,000,000 images (1,000,000 with bounding box annotations)
  - 20000 categories

Curse of Dimensionality

- Analysis in high dimension: $x \in \mathbb{R}^d$ with $d \geq 10^6$.

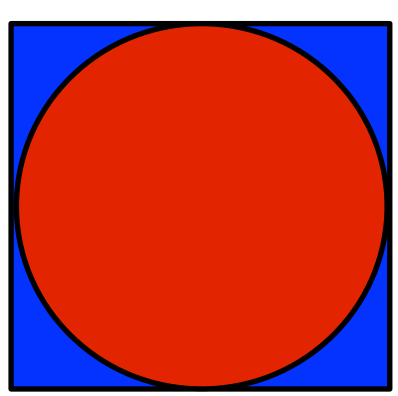
- Points are far away in high dimensions d:

  - 10 points cover $[0, 1]$ at a distance 10^{-1}



  - 10^d points cover $[0, 1]^d$ at a distance 10^{-1}.

$$\lim_{d \to \infty} \frac{\text{volume sphere of radius r}}{\text{volume } [0, r]^d} = 0$$



: nearly all points are in the 2^d corners!

$\Rightarrow$ there is typically no close data point in high dimension.

Classification

- Classification problems:

find the label $y(x) \in \{1, ..., K\}$ for a data vector $x \in \mathbb{R}^d$

$$d \geq 10^6 \text{ and } 2 \leq K \leq 10^4$$

$$\text{Training samples: } \Big\{(x_i, y_i)\Big\}_{i \leq N} \text{ with } 10 \leq N/K \leq 10^3$$

- An interpolation problem:

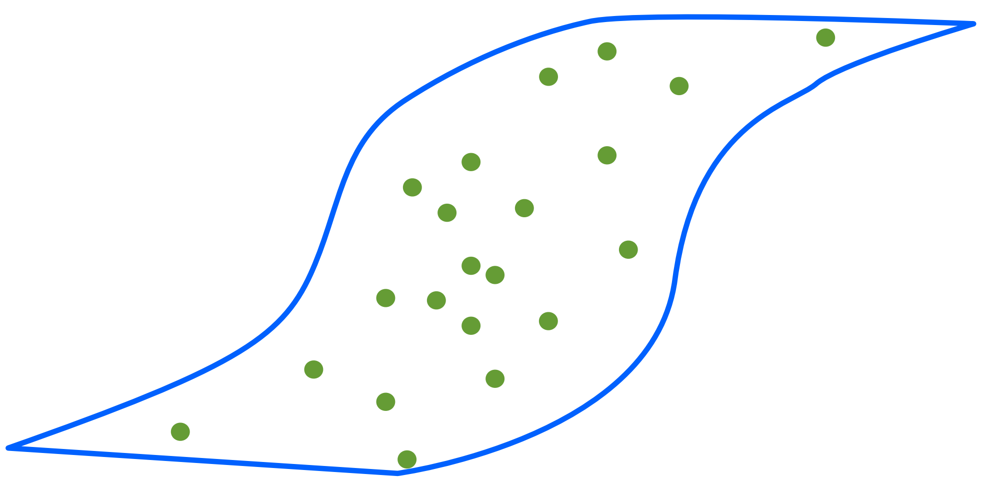
find a good approximation $\tilde{y}(x)$ of $y(x)$, with $\Big\{\tilde{y}(x_i) = y_i\Big\}_i$

- Piecewise constant interpolation: nearsest neighbor classifier

$$\tilde{y}(x) = y_j \text{ if } x_j = \arg\min_{x_i} \|x - x_i\|$$

Low-Dimensional Manifold

- The curse of dimensionality is not a problem if signals belong to low-dimensional manifolds:

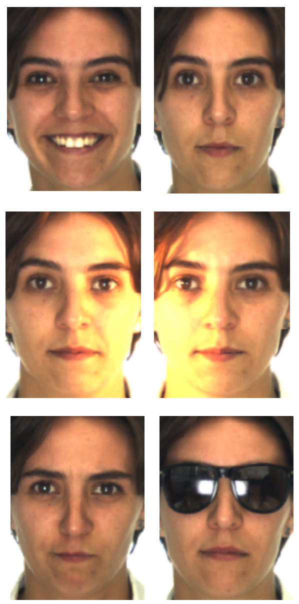


$\Rightarrow$ Euclidean distances provide local similarity measures

- Manifold learning: find intrinsic coordinates by diagonalizing the graph Laplacian.

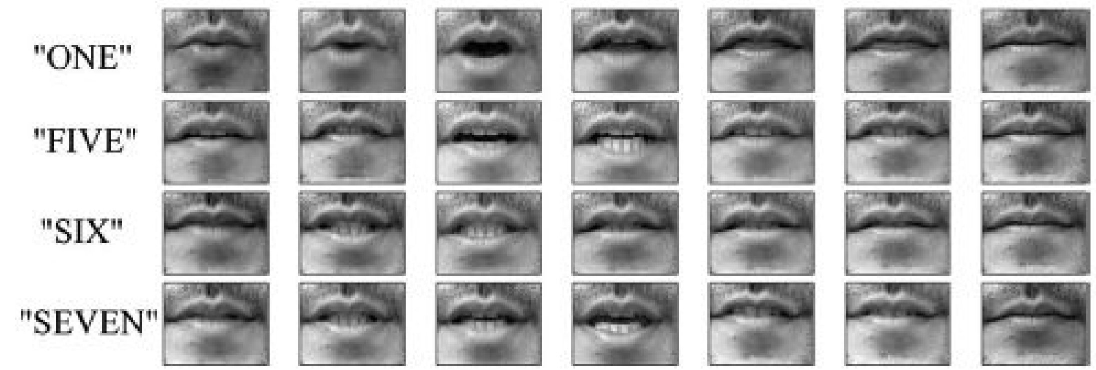
Low Dimensional Data

- Face variations

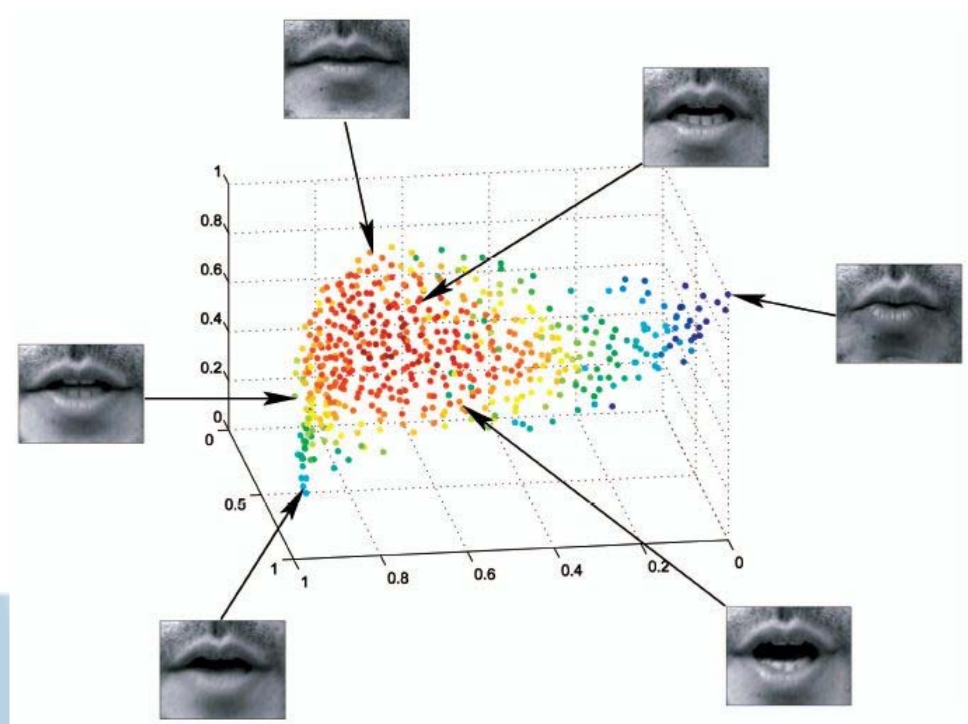


- Rigid motions

- Lips motion

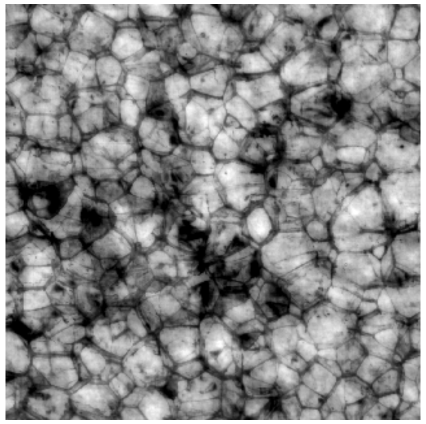


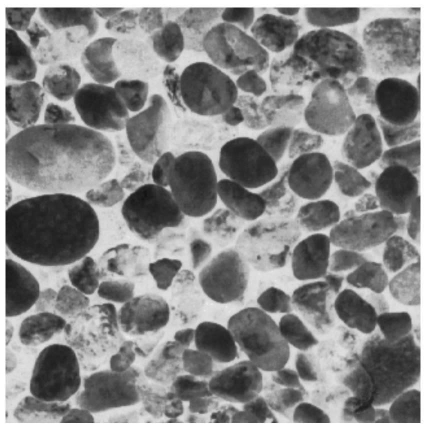
- Identify the manifold where the data lies.

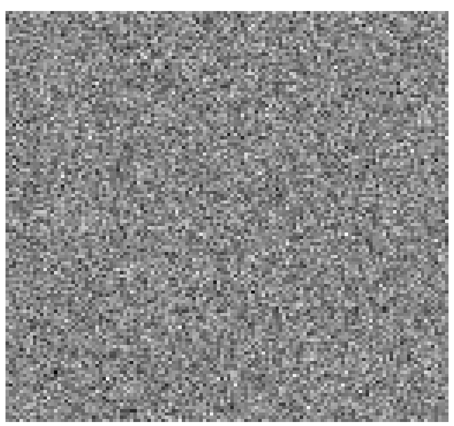


High-Dimensional Data

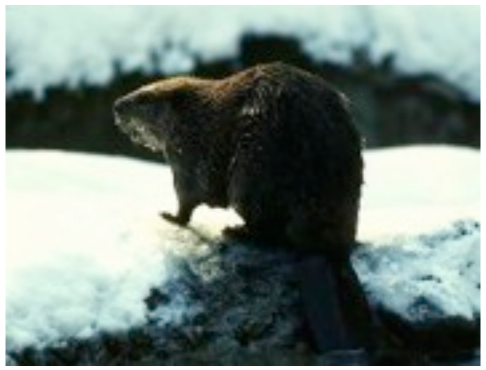
Textures





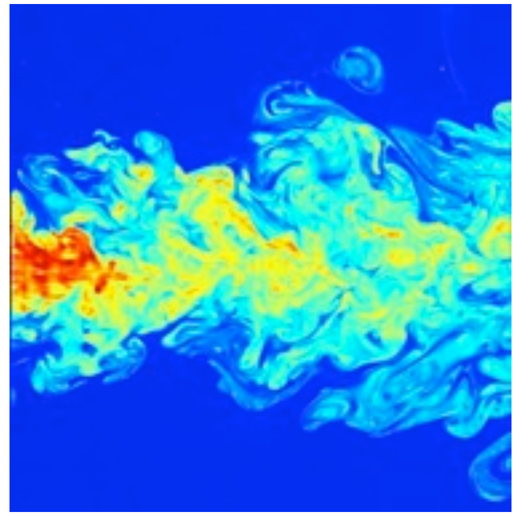


Beaver



Financial time-series

Turbulences



- Need to eliminate irrelevant variability: compute invariants.

Linear Classifier

- Classifications can be reduced to multiple binary classifications

Training samples: $\{(x_i, y_i)\}_i$

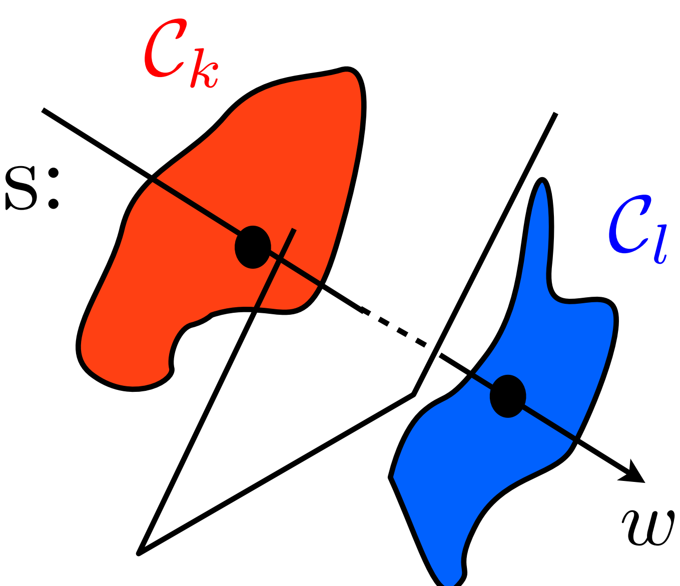
Representation

Supervised linear classification

$$\underset{\mathbb{R}^d}{x} \longrightarrow \boxed{\Phi} \longrightarrow \underset{\mathbb{R}^{d'}}{\Phi x} = \{\phi_n(x)\}_{n \leq d'} \longrightarrow \boxed{\langle \Phi x, w \rangle + b \overset{?}{\geq} 0} \longrightarrow \text{class}$$

$d' \gg d$

Hyperplance separation between pairs of classes:



$$f(x) = \langle \Phi x, w \rangle + b = \sum_n w_n \, \phi_n(x) + b$$

- (1) How to optimize (w, b) to minimize "errors" ?

SVM: $f(x)$ depends on kernel values $K(x, x_i) = \langle \Phi(x), \Phi(x_i) \rangle$.

- (2) How to define Φ to get linear discriminative invariants ?

Increase Dimensionality

Proposition: There exists a hyperplane separating any two subsets of N points $\{\Phi x_i\}_i$ in dimension $d' > N+1$ if $\{\Phi x_i\}_i$ are not in an affine subspace of dimension $< N$.

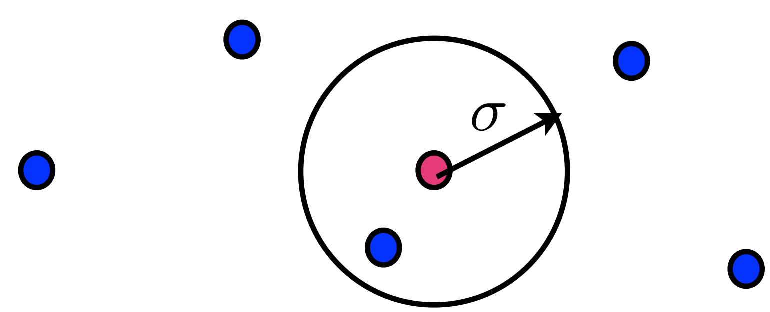
$\Rightarrow$ Choose Φ increasing dimensionality !

Problem: generalisation.

Example: Gaussian kernel $K(x', x) = \exp\left(\frac{-\|x - x'\|^2}{2\sigma^2}\right)$

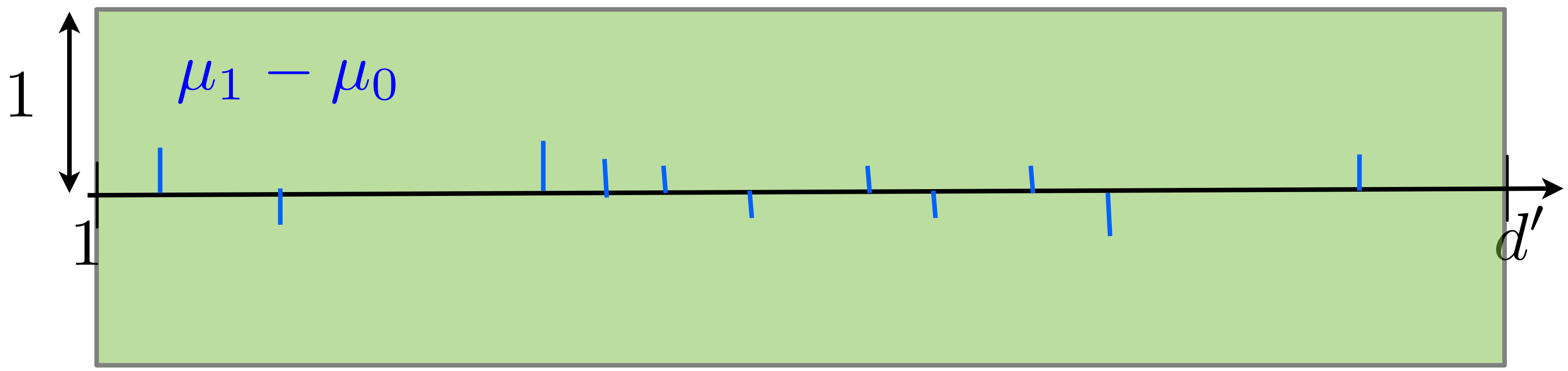
$K(x', x) = \langle \Phi(x'), \Phi(x) \rangle$ where $\Phi x \in \mathcal{H}$ inifinite dimensional.

If σ is small, nearest neighbor classifier type:



Weak and Rare Feature Selection

- Two clasees X_0 and X_1, $\mu_0 = \mathbb{E}(\Phi X_1)$ and $\mu_1 = \mathbb{E}(\Phi X_0)$.
- Normalize X mixture of X_0 and X_1: $\mathrm{Var}(\phi_n(X)) = 1$.
- Rare: $\mu_1 - \mu_0 \in \mathbb{R}^{d'}$ is sparse. Weak: $\|\mu_1 - \mu_0\|_\infty \ll 1$.



Donoho & Jin
Tsybakov

$$f(x) = \sum_n w_n \, \phi_n(x) - \frac{\mu_0 + \mu_1}{2}.$$

$\Rightarrow$ feature selection: $(w_n)_n$ should be sparse

- Find w so that $f(X_1) \approx 1$ and $f(X_0) \approx -1$:

linear invariant which discriminates the two clases.

Optimal Representation

Find an operator Φ which:

- eliminate useless variability for classification

$\Rightarrow$ **reduce dimensionality**

- can yield many possible invariants as linear combinations

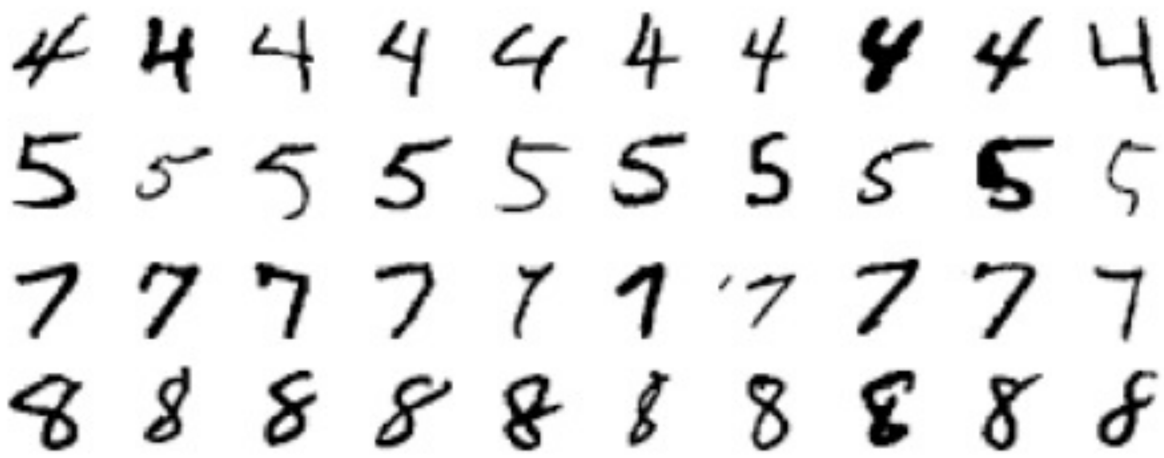
$$\sum_n w_n \, \phi_n(x)$$

depending upon the class of x.

need enough invariants $\Rightarrow$ **increase dimensionality**.

Translations and Deformations

- Patterns are translated and deformed



Invariance to Translations

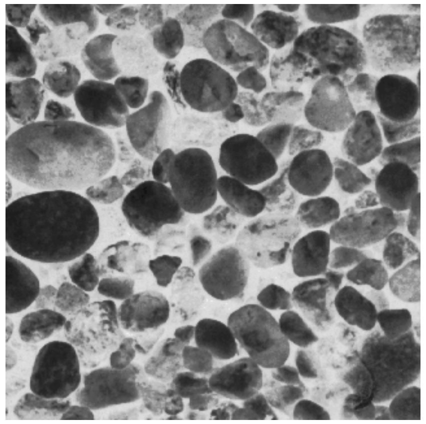
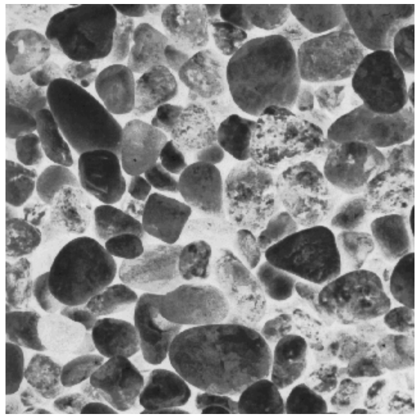
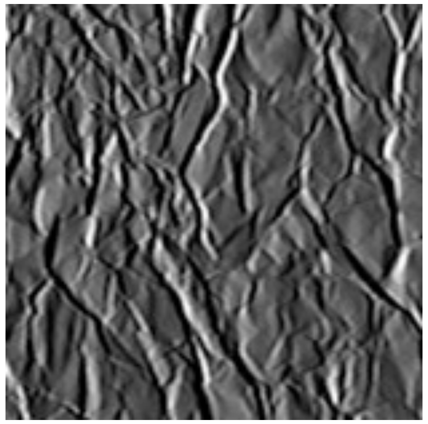
Two dimensional group: $\mathbb{R}^2$

Deformations are actions of diffeomorphisms: infinite group.

Each digit is invariant to a specific set of small deformations

- Textures are stationary (translation invariant) processes

with deformations

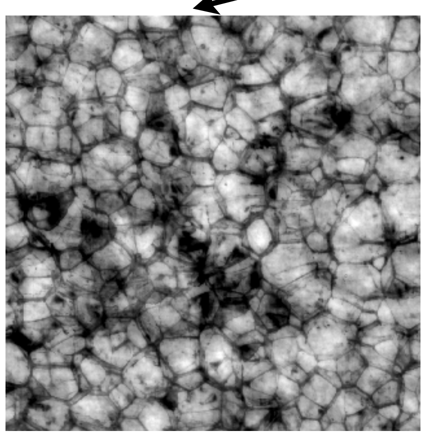
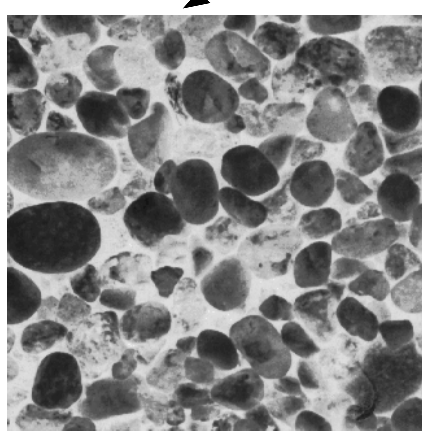
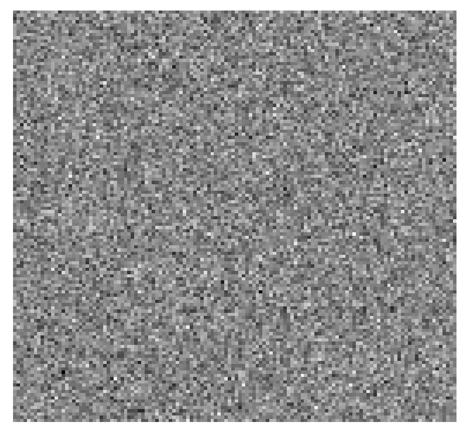
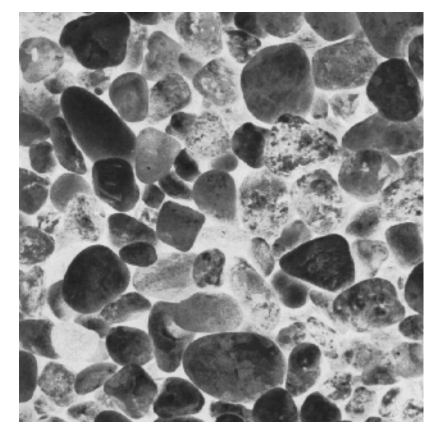
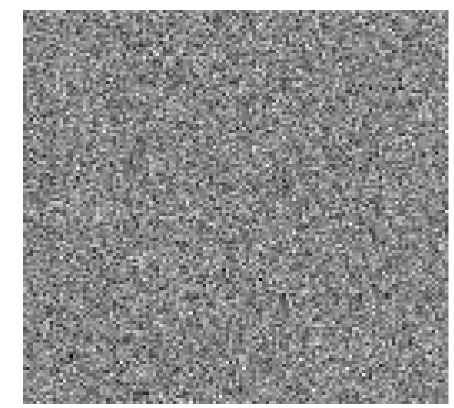
    

Texture Discrimination

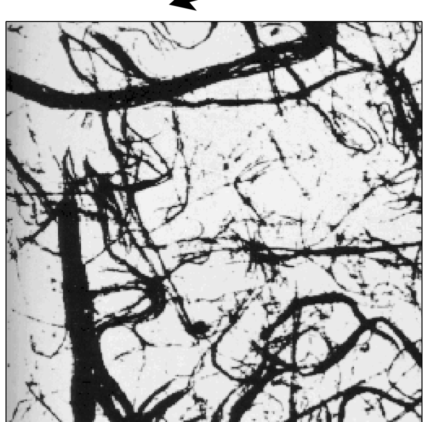
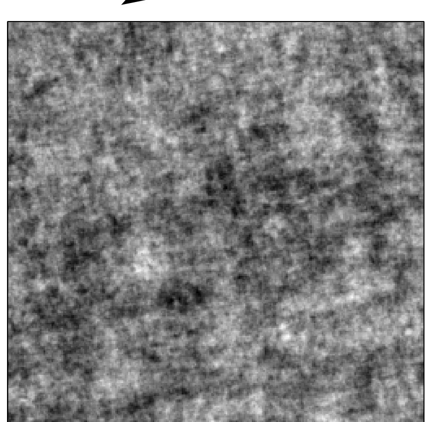
• Textures are realizations of high-dimensional stationary processes, which are typically not Gaussian or Markovian.

• Second order moment invariants: $\mathbb{E}\Big(X(t)\,X(t-m)\Big)$ estimated from 1 realization with weak ergodicity conditions.

same second order moments

same second order moments: not discriminative.

• Use higher order moments ?
Estimators have a large variance
$\Rightarrow$ not sufficiently invariant.

Audio Textures

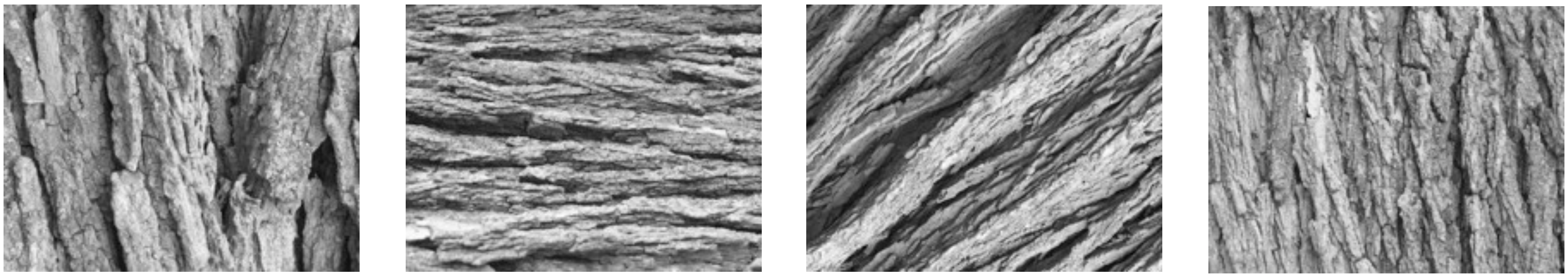
J. McDermott textures

same second order moments

- Natural Sounds (1s) Original Gaussian model
  - Hammer
  - Water
  - Applause

Rotation and Scaling Variability

- Rotation and deformations

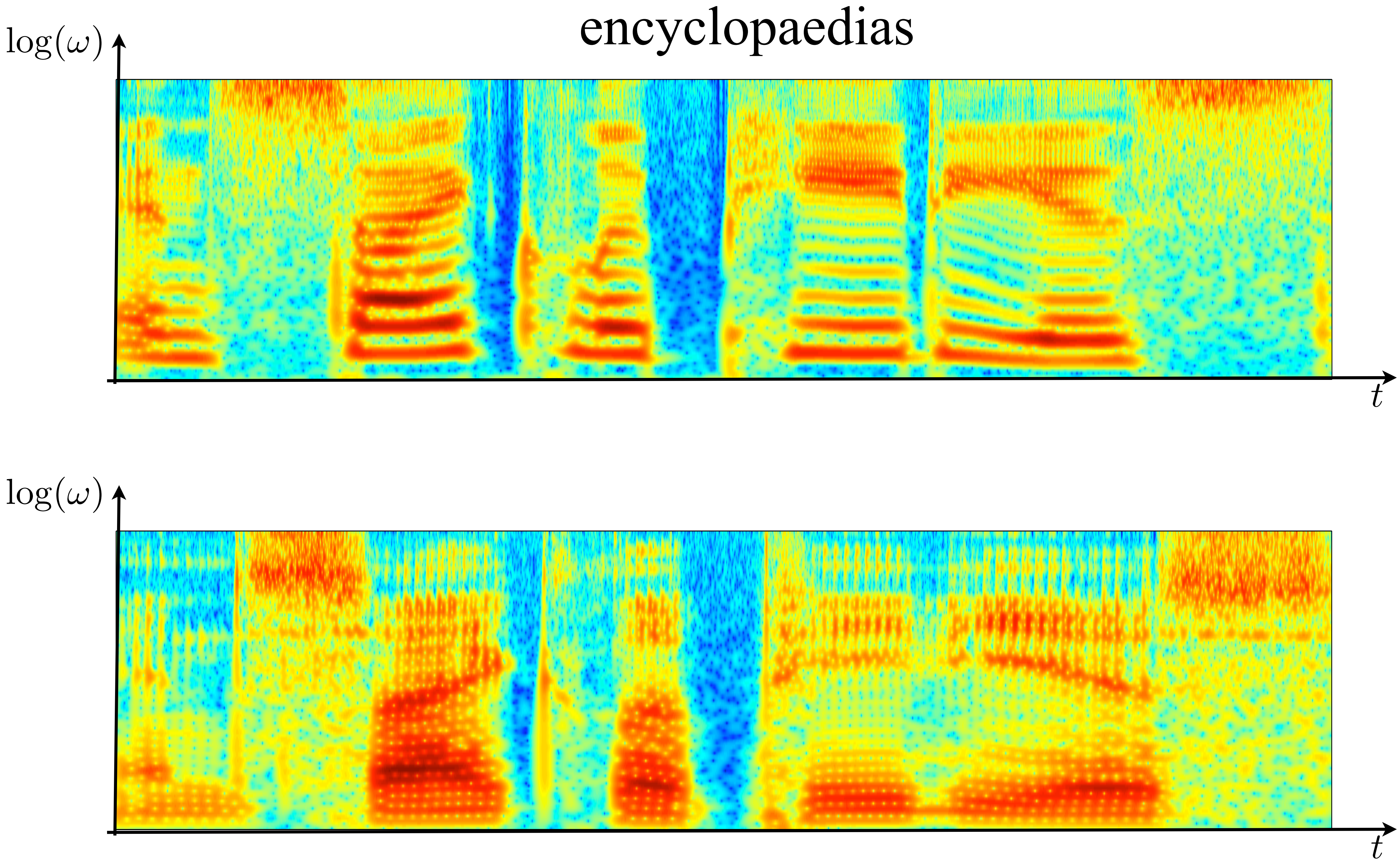


Group: $SO(2) \times \mathrm{Diff}(SO(2))$

- Scaling and deformations

Group: $\mathbb{R} \times \mathrm{Diff}(\mathbb{R})$

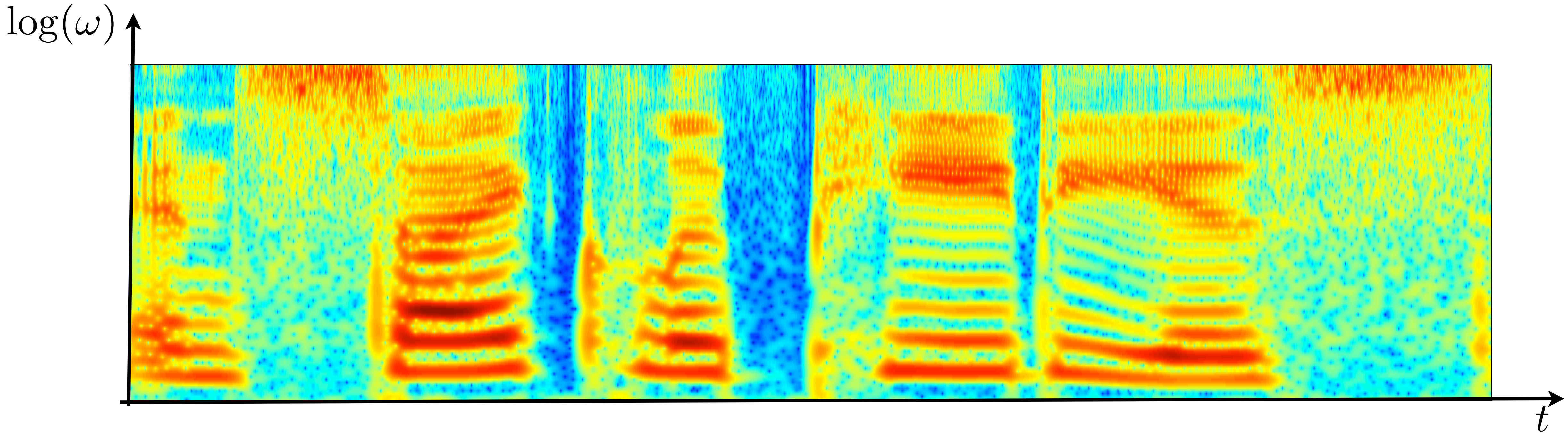
Frequency Transpositions



H : Heisenberg group of "time-frequency" translations

Frequency Transpositions

Time and frequency translations and deformations:



- Frequency transposition invariance is needed

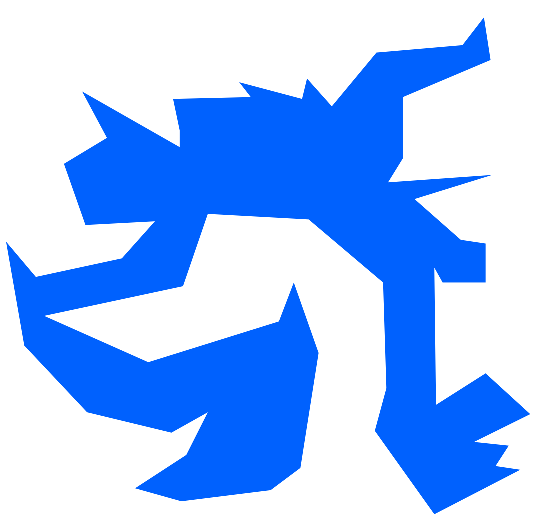
  for speech recognition not for locutor recognition.

Classification with Invariants

Representation — Supervised linear classification

$$\underset{\mathbb{R}^d}{x} \longrightarrow \boxed{\Phi} \longrightarrow \underset{d' \gg d}{\overset{\mathbb{R}^{d'}}{\Phi x}} \longrightarrow \boxed{\langle \Phi x, w \rangle + b \overset{?}{\geq} 0} \longrightarrow \text{class}$$

- Φ may be defined from prior knowledge on data.
- **Unsupervised learning** of Φ from unlabeled examples $\{x_i\}$: requires to model a very high dimension distribution in $\mathbb{R}^d$.

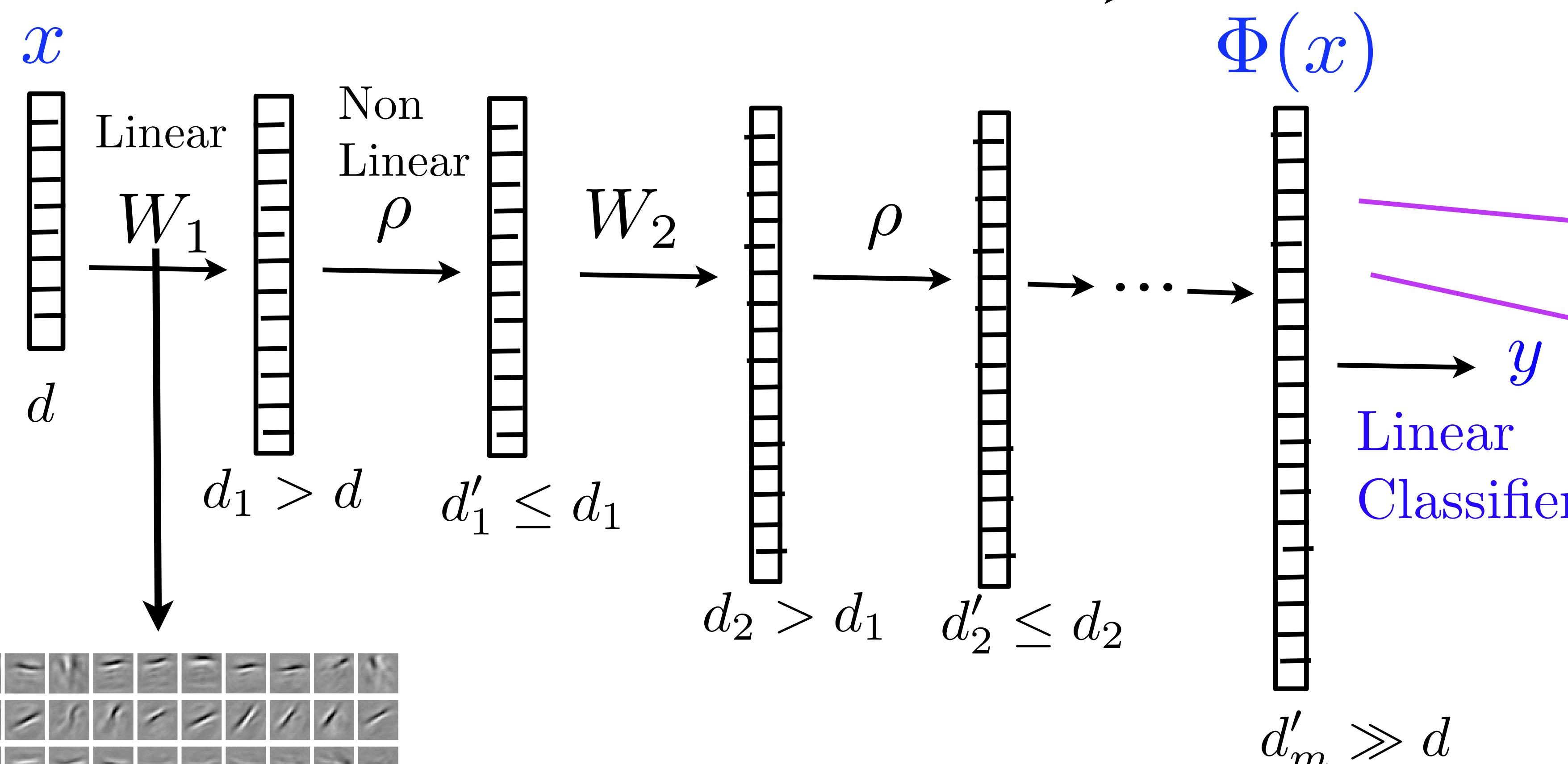


Dimension is too high for:
- Gaussian mixture models
- Graphical models

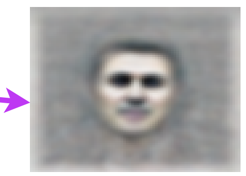
Deep Neural Networks

J. Hinton, Y. LeCun "State of the art results"

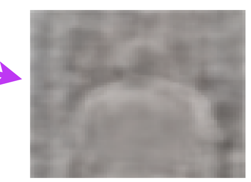
Hierarchical invariance



Le, Ranzato, Ng et. al.:
unsupervised training
1 billion variables

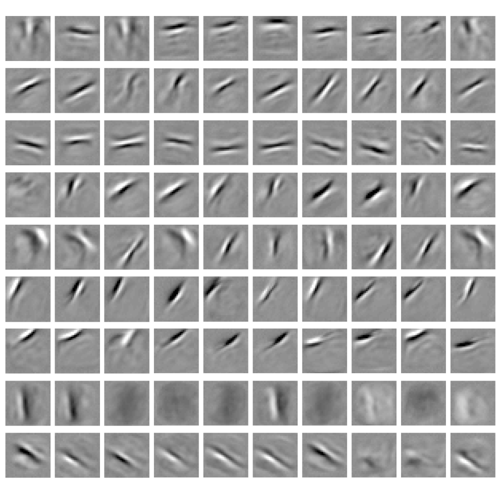


Face detector



Body detector

"Grand-mother cells"



Wavelets

Hinton, Bengio, Ranzato et. al.:
usupervised learning with sparsity

Why and how does it work ?

Overview

- **Part I:**

  Invariants and stability to diffeomorphisms

  Scattering and deep neural networks

- **Part II:**

  Limit scattering transform

  Expected scattering of stationary processes

  Texture discrimination and synthesis

- **Part III:**

  Multifractal random processes

  Scattering on Lie Groups

  Unsupervised learning of representations

Translations and Deformations

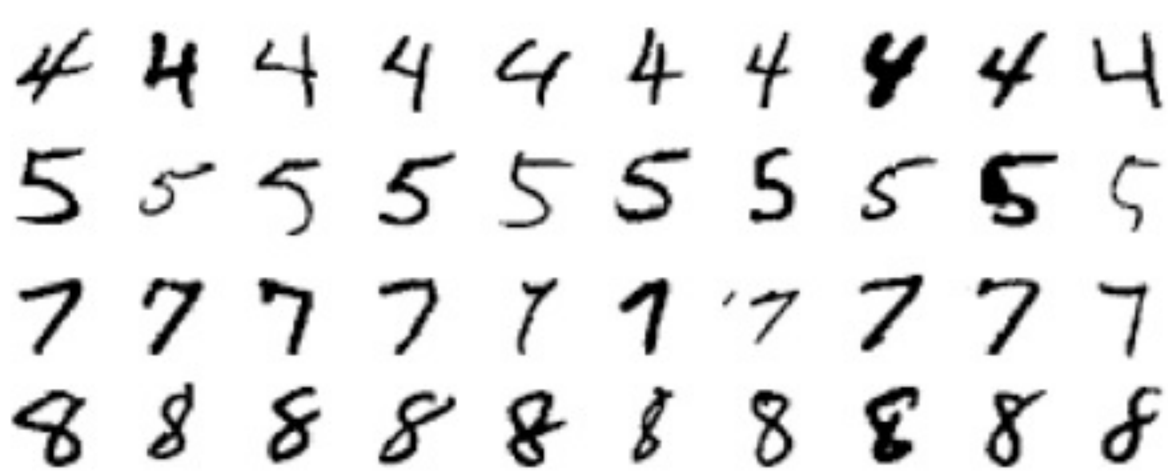
- Patterns are translated and deformed (class dependent)

Need invariance to translations: two dimensional group $\mathbb{R}^2$

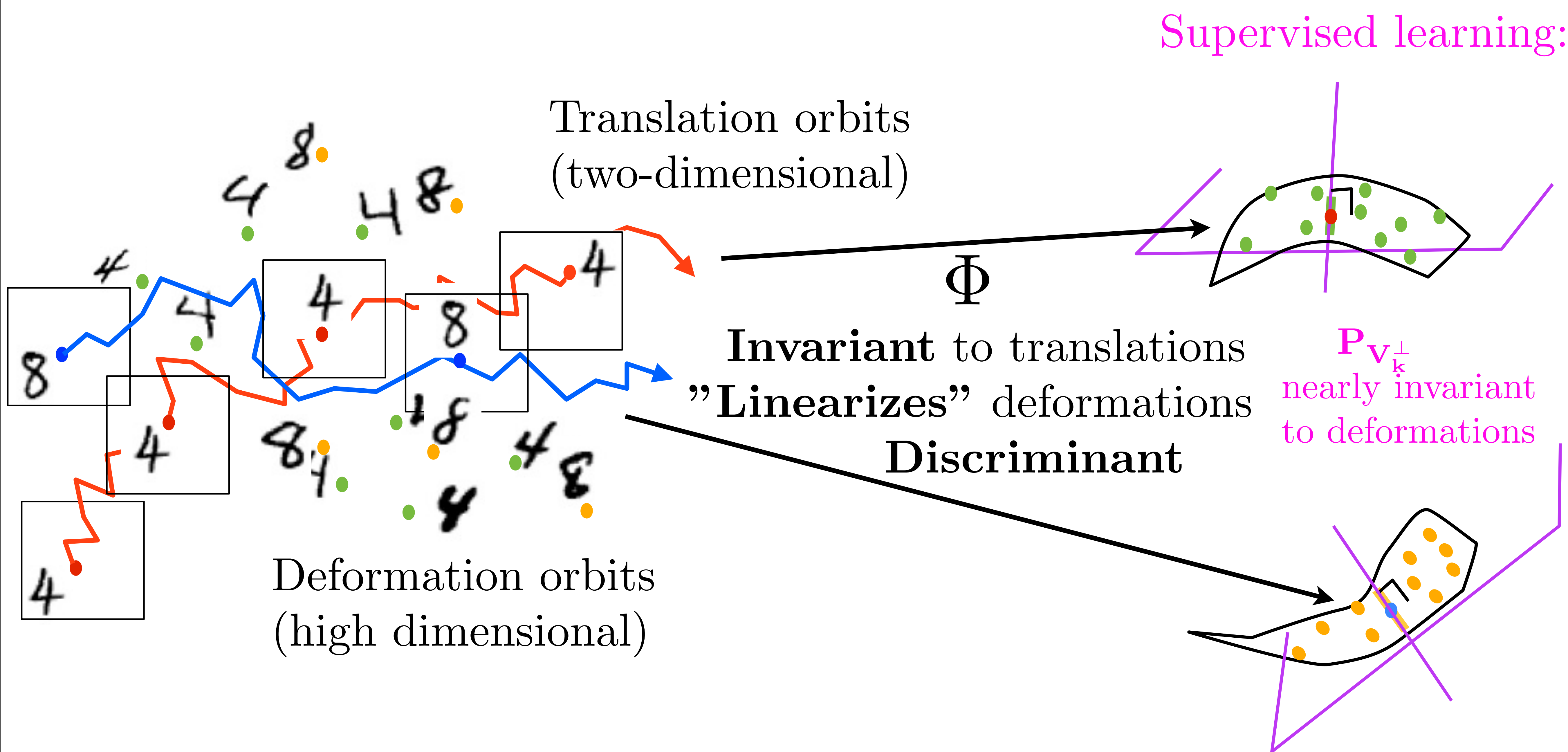
Need class dependent invariance to small deformations:

belong to diffeomorphism group.

Digit recognition



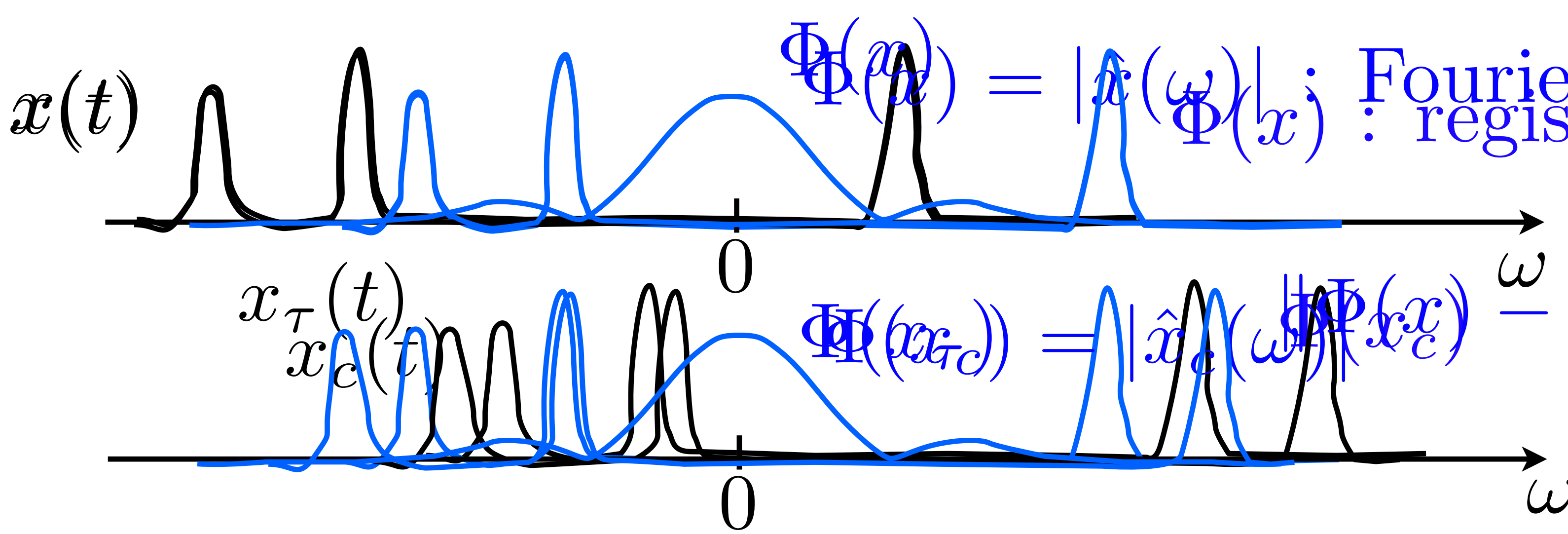
Translation and Deformations



Stable Translation Invariants

- **Invariance** to translations $x_c(t) = x(t-c)$

$$\forall c \in \mathbf{R} \quad , \quad \Phi(x_c) = \Phi(x) \ .$$



Not stable

$$\|\Phi(x) - \Phi(x_\tau)\| \gg \sup_t |\tau'(t)|\,\|x\|$$

Fourier invariants
are not stable either.

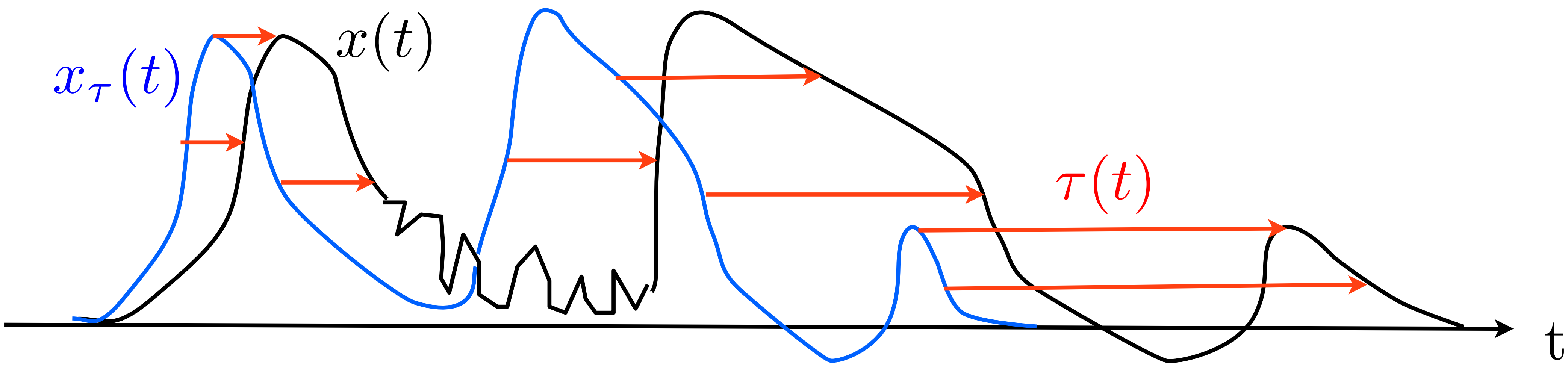
- **Lipschitz stable** to diffeomorphisms $x_\tau(t) = x(t - \tau(t))$

small deformations of $x \implies$ small modifications of $\Phi(x)$

$$\forall \tau \quad , \quad \|\Phi(x_\tau) - \Phi(x)\| \leq C \underbrace{\sup_t |\nabla \tau(t)|\,\|x\|}_{\text{diffeomorphism metric}} \ .$$

Estimation of Deformations

- Deformation appear in many statistical problems:



- We do not want to compute $\tau(t)$.
- Need to be "sensitive" to $\tau(t)$.

- **Lipschitz stable** to diffeomorphisms $x_\tau(t) = x(t - \tau(t))$

$$\forall \tau \quad , \quad \|\Phi(x_\tau) - \Phi(x)\| \leq C \sup_t |\tau'(t)| \, \|x\| \ .$$

Fourier Translation Invariance

- Fourier transform $\hat{x}(\omega) = \int x(t)\, e^{-i\omega t}\, dt$ invariance:

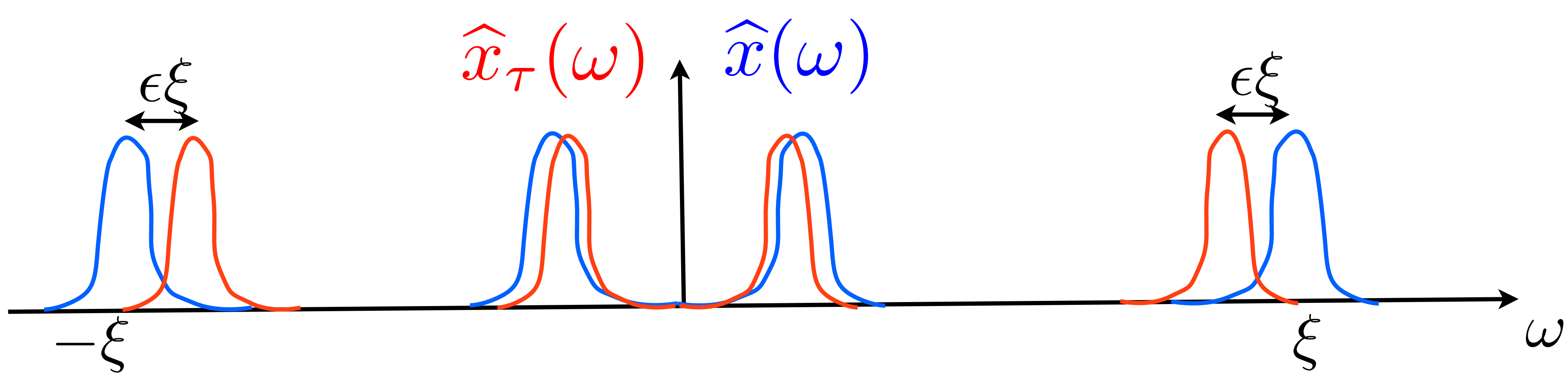
$$\text{if } x_c(t) = x(t-c) \text{ then } |\widehat{x_c}(\omega)| = |\hat{x}(\omega)|$$

- Instabilites to small deformations $x_\tau(t) = x(t - \tau(t))$:

$$\big|\, |\hat{x}_\tau(\omega)| - |\,\hat{x}(\omega)|\, \big| \;\text{ is big at high frequencies}$$

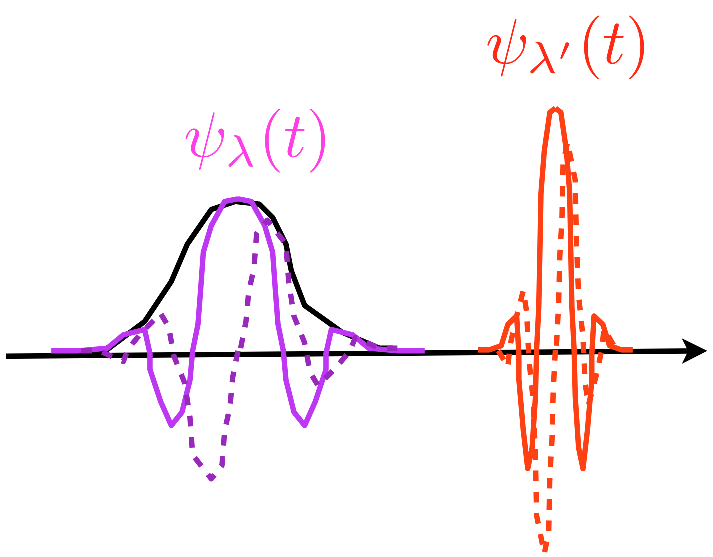
Example: If $\tau(t) = \epsilon t$ then $x_\tau(t) = x((1-\epsilon)t)$

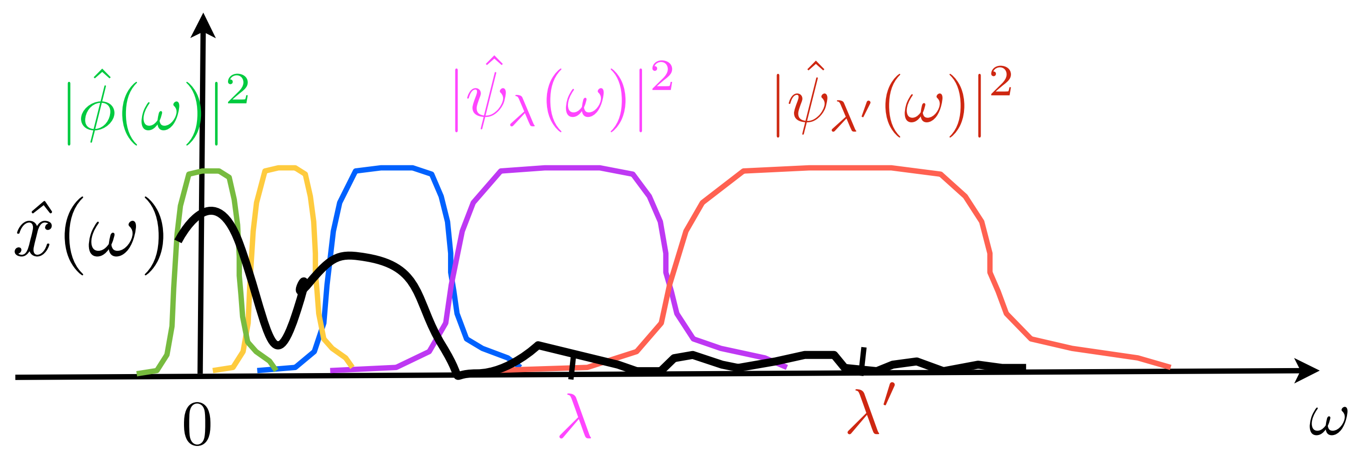
$$\Rightarrow \quad \widehat{x}_\tau(\omega) = (1-\epsilon)^{-1}\, \widehat{x}((1-\epsilon)^{-1}\omega)$$



Wavelet Transform

- Complex analytic wavelet: $\psi(t) = \psi^a(t) + i\,\psi^b(t)$
- Dilated: $\psi_\lambda(t) = \alpha^{-j}\,\psi(\alpha^{-j}t)$ with $\lambda = \alpha^{-j}$.





- Wavelet transform: $x \star \psi_\lambda(t) = \int x(u)\,\psi_\lambda(t-u)\,du$

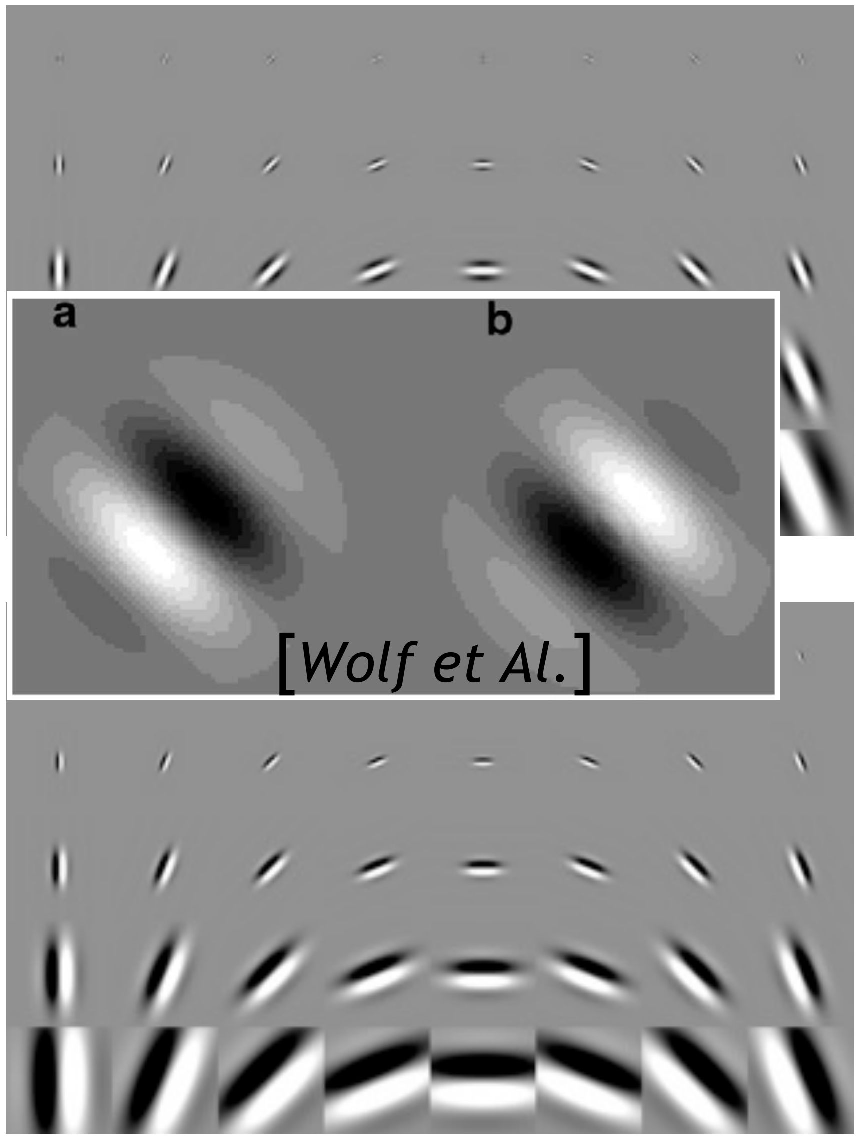
$$W x = \left(\begin{array}{c} x \star \phi(t) \\ x \star \psi_\lambda(t) \end{array} \right)_{t,\lambda}$$

Image Wavelet Transform

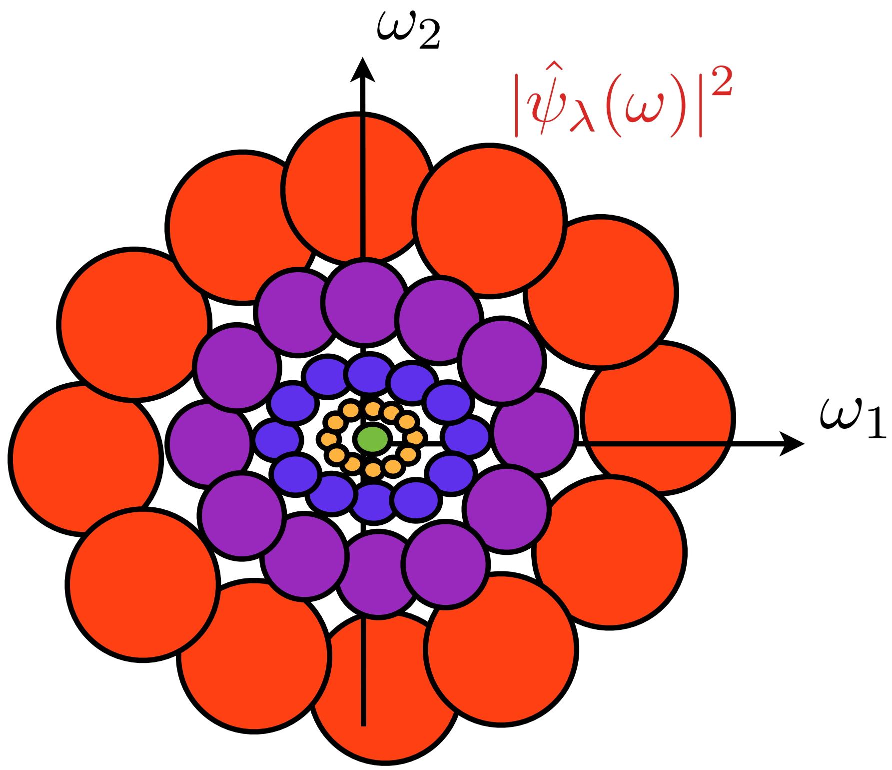
- Complex wavelet: $\psi(t) = \psi^a(t) + i\,\psi^b(t) \quad , \quad t = (t_1, t_2)$

  rotated and dilated: $\psi_\lambda(t) = 2^{-j}\,\psi(2^{-j} r t)$ with $\lambda = (2^j, r)$

real parts



imaginary parts



- Wavelet transform: $Wx = \begin{pmatrix} x \star \phi(t) \\ x \star \psi_\lambda(t) \end{pmatrix}_{t,\lambda}$

Wavelet Tight Frames in L2

Functions in $\mathbf{L^2}(\mathbb{R}^d)$: $\|x\|^2 = \int |x(t)|^2\, dt < \infty$

$$Wx = \begin{pmatrix} x \star \phi(t) \\ x \star \psi_\lambda(t) \end{pmatrix}_{t,\lambda}$$

Proposition: *(Littlewood-Paley)*

The wavelet transform is a tight frame for $x \in \mathbf{L^2}(\mathbb{R}^d)$

$$\|Wx\|^2 = \|x \star \phi\|^2 + \sum_\lambda \|x \star \psi_\lambda\|^2 = \|x\|^2$$

if and only if for almost all ω.

$$|\hat{\phi}(\omega)|^2 + \frac{1}{2} \sum_\lambda \left(|\hat{\psi}_\lambda(\omega)|^2 + |\hat{\psi}_\lambda(-\omega)|^2 \right) = 1$$

Why Wavelets ?

- The wavelet dictionary $\{\psi_\lambda(t-u)\}_{t,\lambda}$ is translation invariant.

- Wavelets are uniformly stable to deformations:

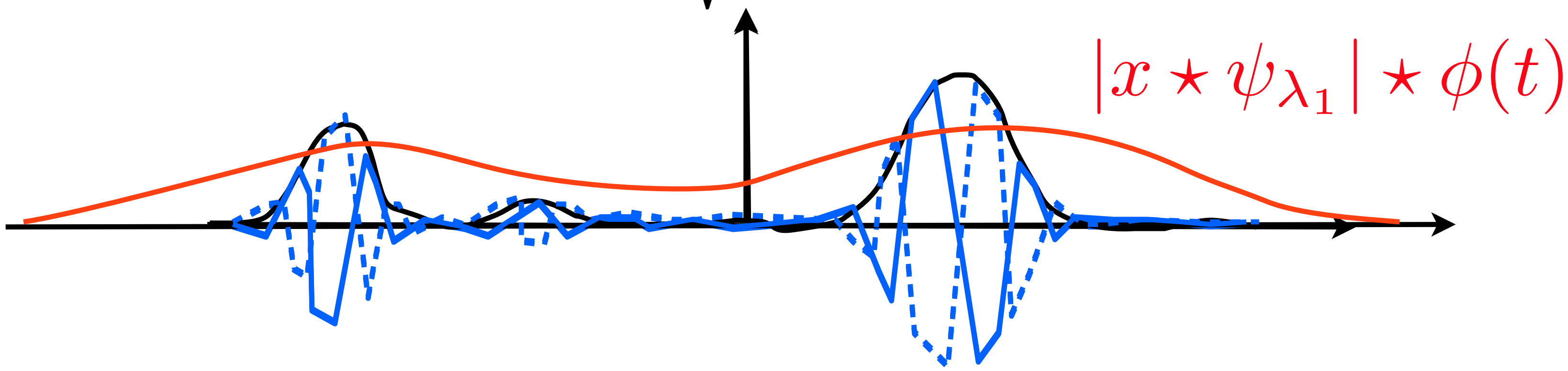
  if $\psi_{\lambda,\tau}(t) = \psi_\lambda(t-\tau(t))$ then

$$\|\psi_\lambda - \psi_{\lambda,\tau}\| \leq C \sup_t |\nabla\tau(t)| \ .$$

Wavelet Translation Invariance

$$x \star \psi_{\lambda_1}(t) = x \star \psi^a_{\lambda_1}(t) + i\, x \star \psi^b_{\lambda_1}(t)$$

$$|x \star \psi_{\lambda_1}(t)| = \sqrt{|x \star \psi^a_{\lambda_1}(t)|^2 + |x \star \psi^b_{\lambda_1}(t)|^2} \quad \text{pooling}$$



- The modulus $|x \star \psi_{\lambda_1}|$ is a regular envelop
- The average $|x \star \psi_{\lambda_1}| \star \phi(t)$ is invariant to small translations relatively to the support of ϕ.
- Full translation invariance at the limit:

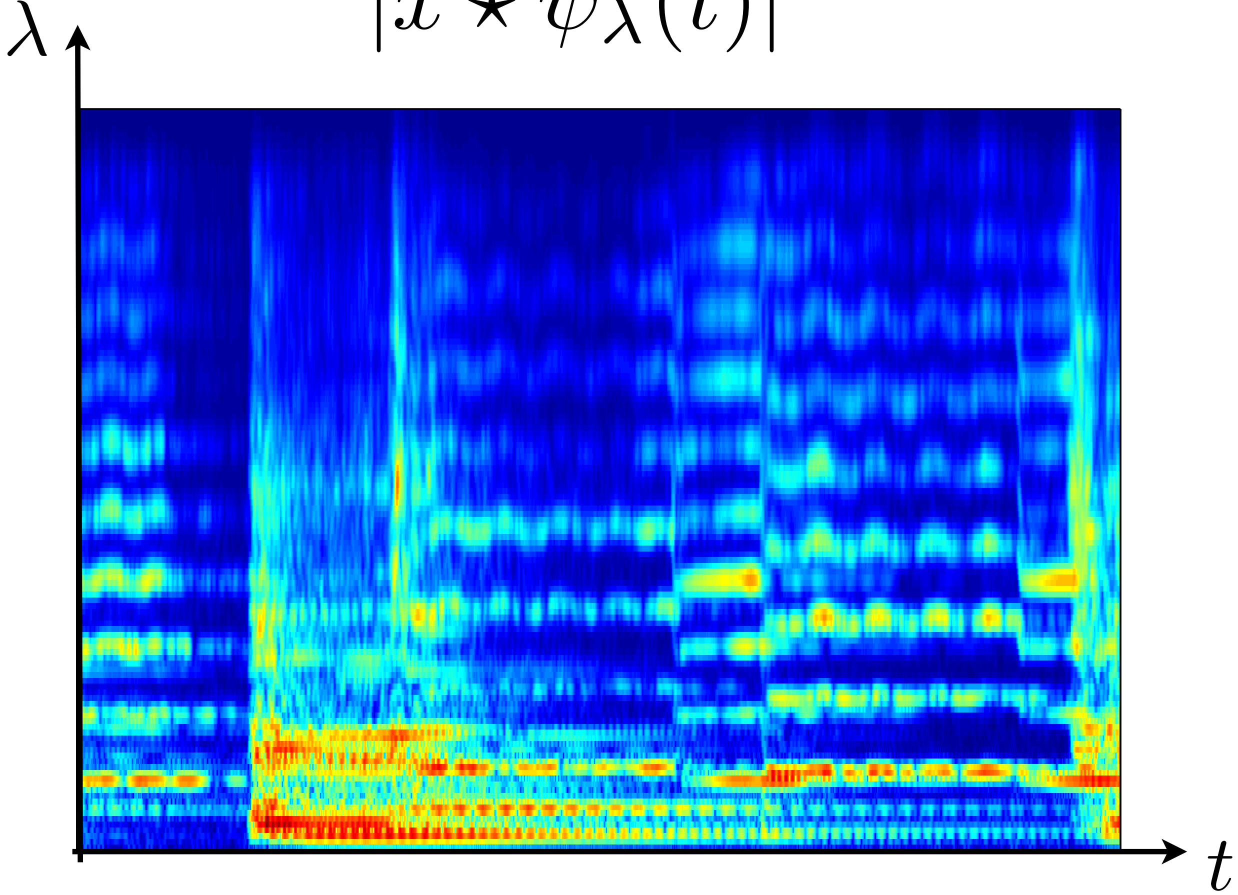
$$\lim_{\phi \to 1} |x \star \psi_{\lambda_1}| \star \phi(t) = \int |x \star \psi_{\lambda_1}(u)|\, du = \|x \star \psi_{\lambda_1}\|_1$$

but few invariants.

Wavelet Stabilization

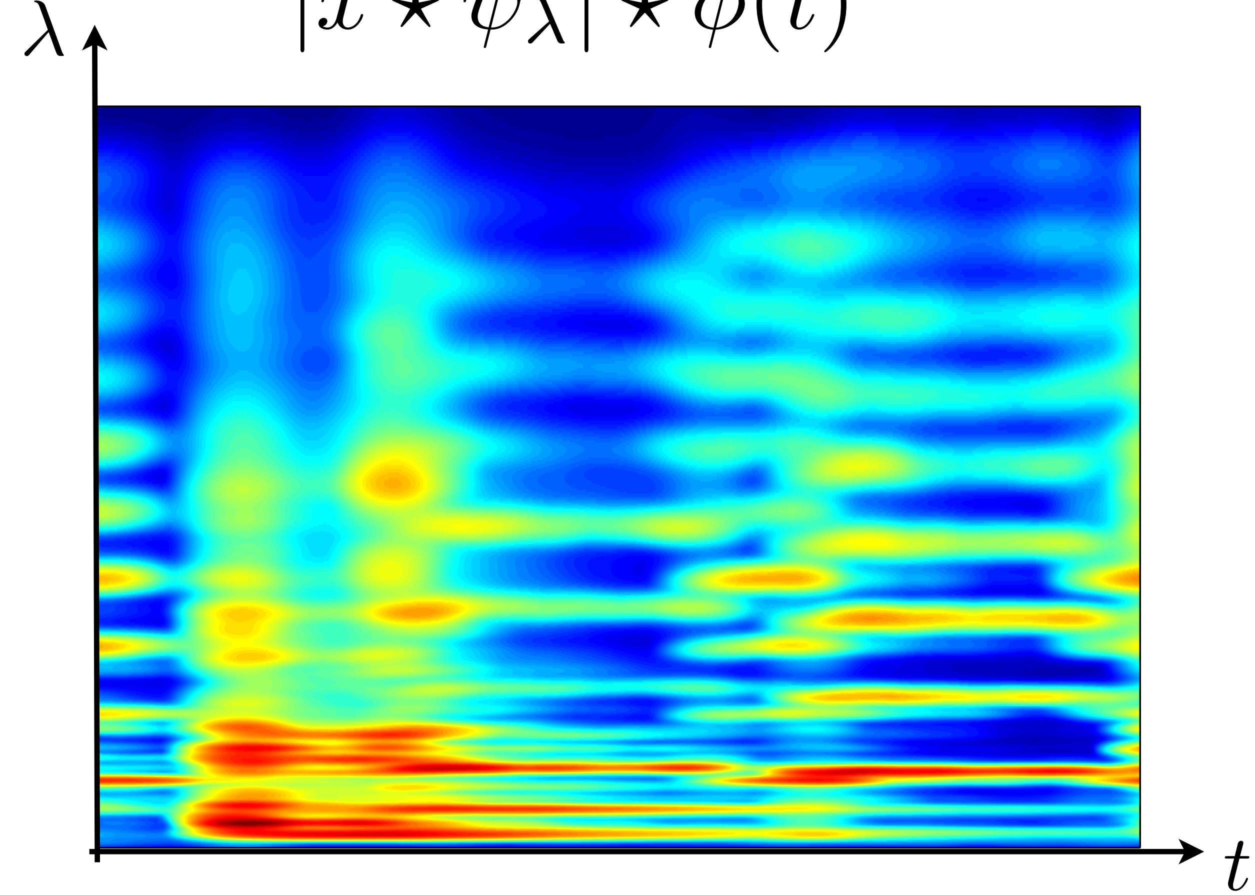
Wavelet time-frequency

$$|x \star \psi_\lambda(t)|$$



Time averaging on **370ms**

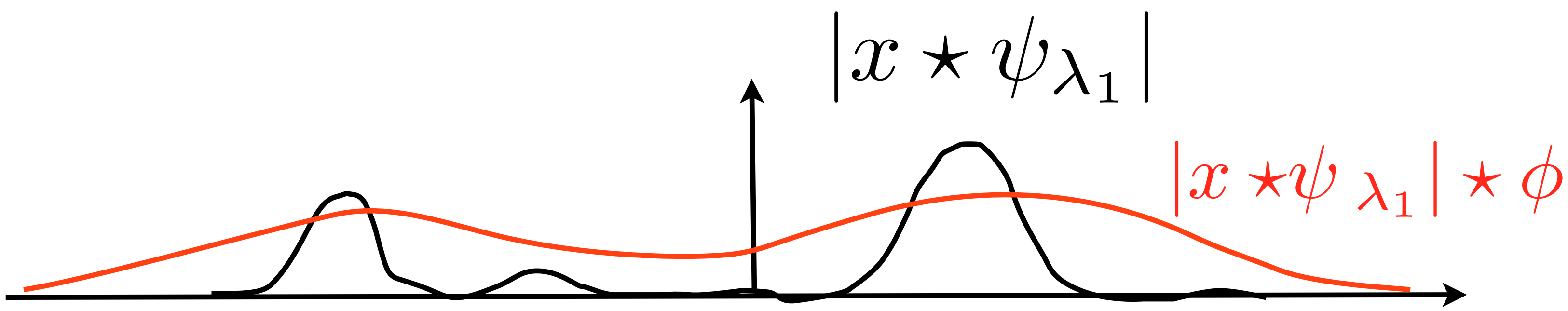
$$|x \star \psi_\lambda| \star \phi(t)$$



Locally invariant to translations and stable to deformations

but loss of information.

Recovering Lost Information



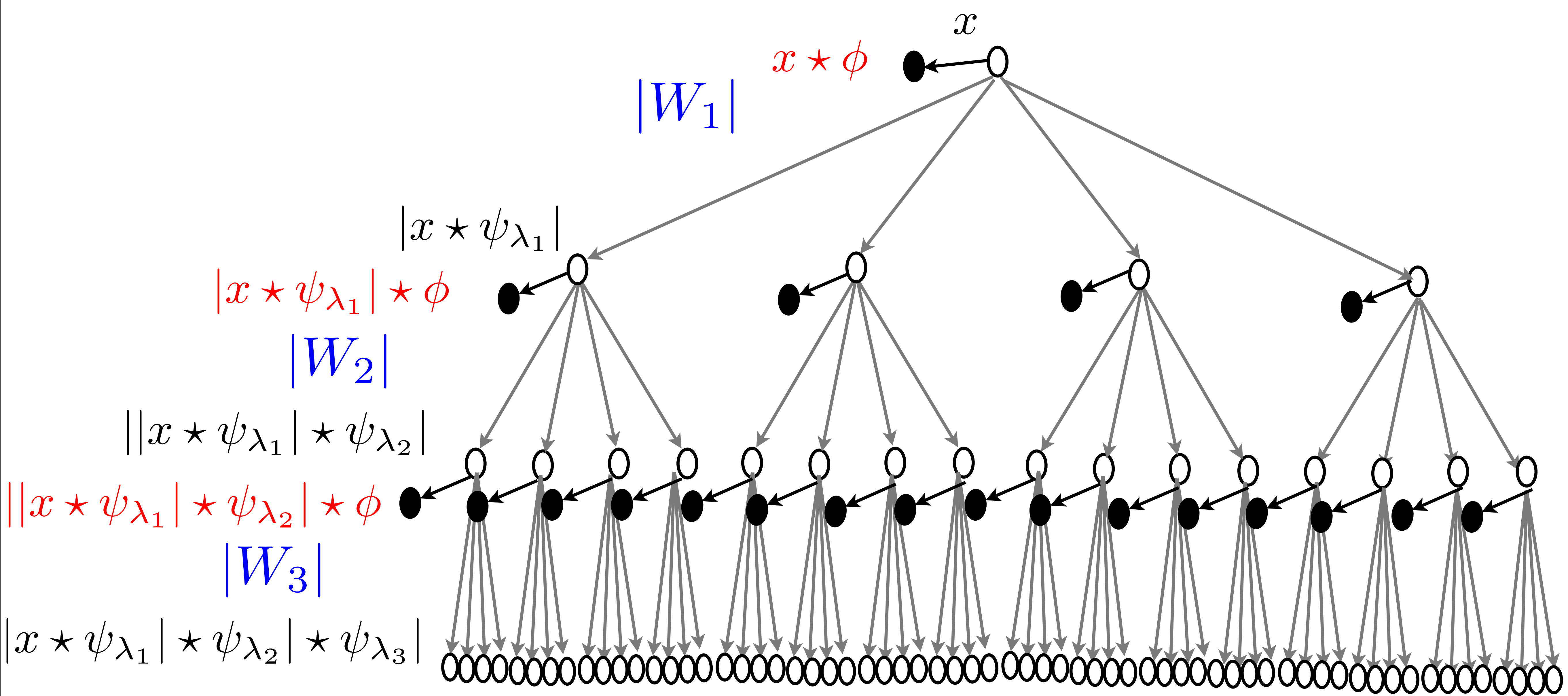
- The high frequencies of $|x \star \psi_{\lambda_1}|$ are in wavelet coefficients:

$$W|x \star \psi_{\lambda_1}| = \left(\begin{array}{c} |x \star \psi_{\lambda_1}| \star \phi(t) \\ |x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(t) \end{array} \right)_{t,\lambda_2}$$

- Translation invariance by time averaging the amplitude:

$$\forall \lambda_1 , \lambda_2 , \quad ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t)$$

Deep Convolution Network



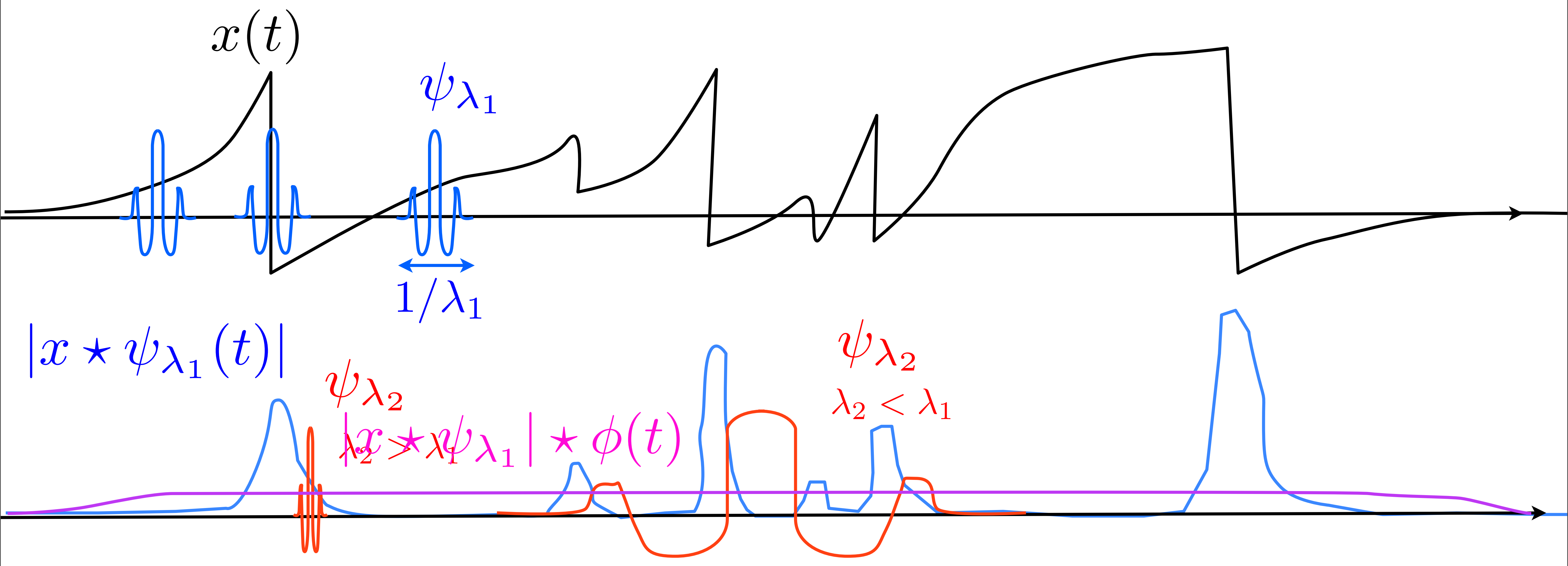
Scattering Vector

Network ouptut:

$$
Sx = \left(\begin{array}{c} x \star \phi\,(u) \\ |x \star \psi_{\lambda_1}| \star \phi(u) \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(u) \\ |||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi(u) \\ \ldots \end{array} \right)_{u,\lambda_1,\lambda_2,\lambda_3,\ldots}
$$

Singular Functions

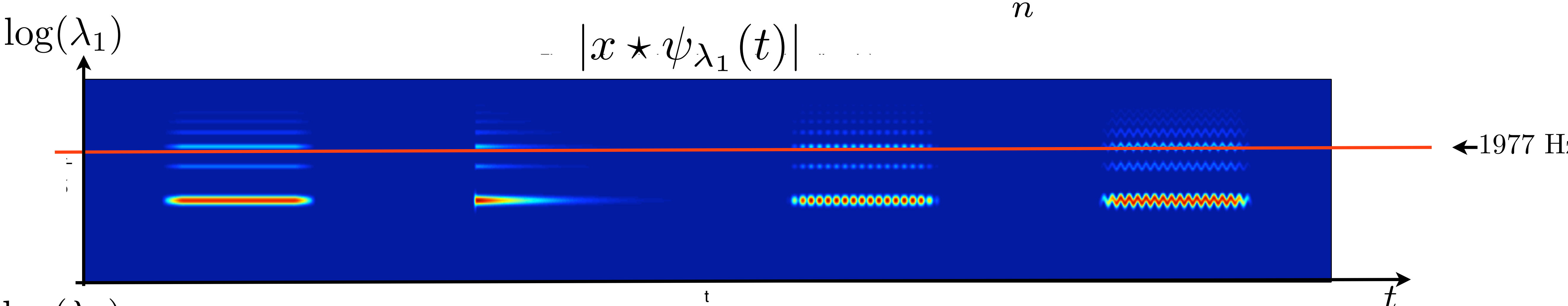
$$|x \star \psi_{\lambda_1}(t)| = \left| \int x(u)\psi_{\lambda_1}(t-u)\,du \right|$$



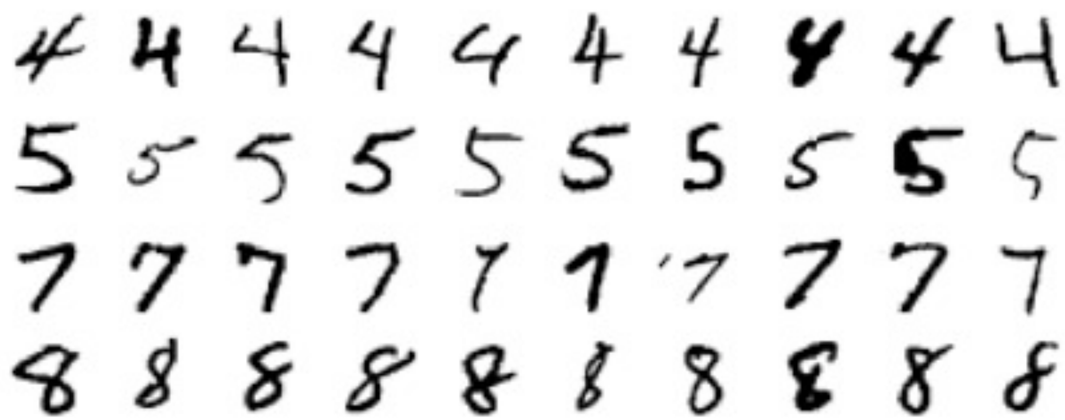
$$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(t)| \approx 0 \quad \text{if } \lambda_2 > \lambda_1$$

Amplitude Modulation

$$x_i(t) = a_i(t)\Big(c \star h(t)\Big) \quad \text{with} \quad c(t) = \sum_n \delta(t - nT)\ .$$



Translations and Deformations



- **Invariance** to translations $x_c(t) = x(t-c)$

$$\forall c \in \mathbf{R} \quad , \quad \Phi(x_c) = \Phi(x) \ .$$

Fourier invariant: $\Phi(x) = |\hat{x}(\omega)| = \left| \int x(t)\, e^{-it\omega}\, d\omega \right|$

- **Lipschitz stable** to diffeomorphisms $x_\tau(t) = x(t - \tau(t))$

$$\forall \tau \quad , \quad \|\Phi(x_\tau) - \Phi(x)\| \leq C \underbrace{\sup_t |\nabla \tau(t)|}_{\text{diffeomorphism metric}} \|x\| \ .$$

Fourier Failure

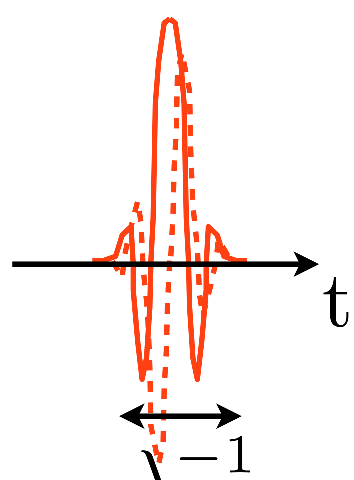
Wavelet Transform

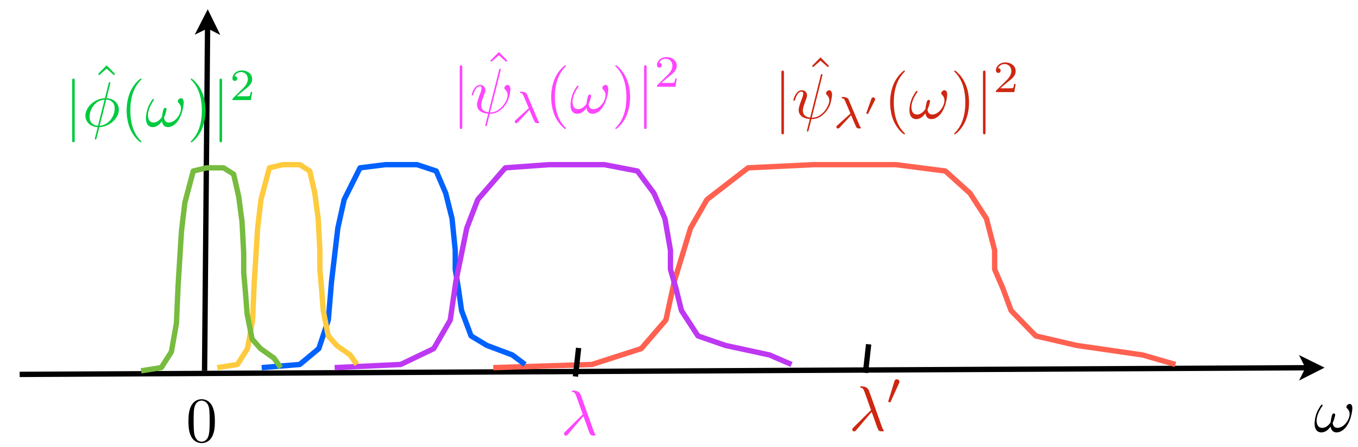
- Complex analytic wavelet: $\psi(t) = \psi^a(t) + i\,\psi^b(t)$

- For $t \in \mathbb{R}$, dilated

$$\psi_\lambda(t) = 2^j \psi(2^j t)$$

with $\lambda = 2^j$

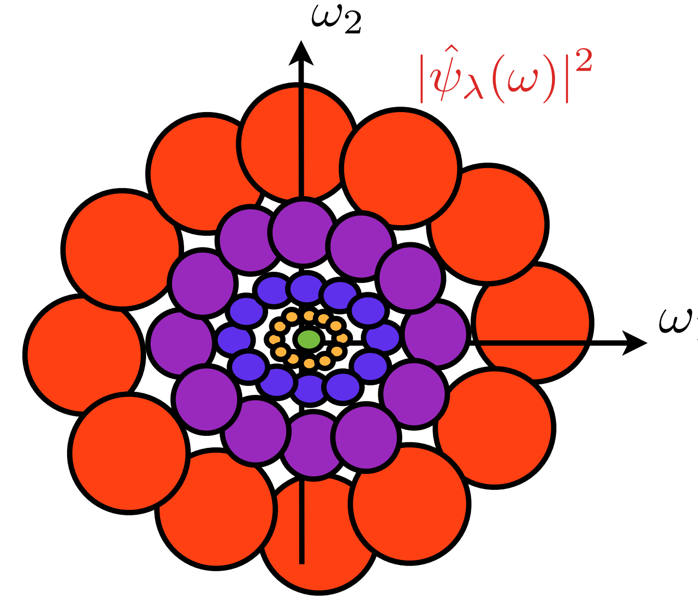




- For $t \in \mathbb{R}^2$, dilated and rotated

$$\psi_\lambda(t) = 2^j\,\psi(2^j r t)$$

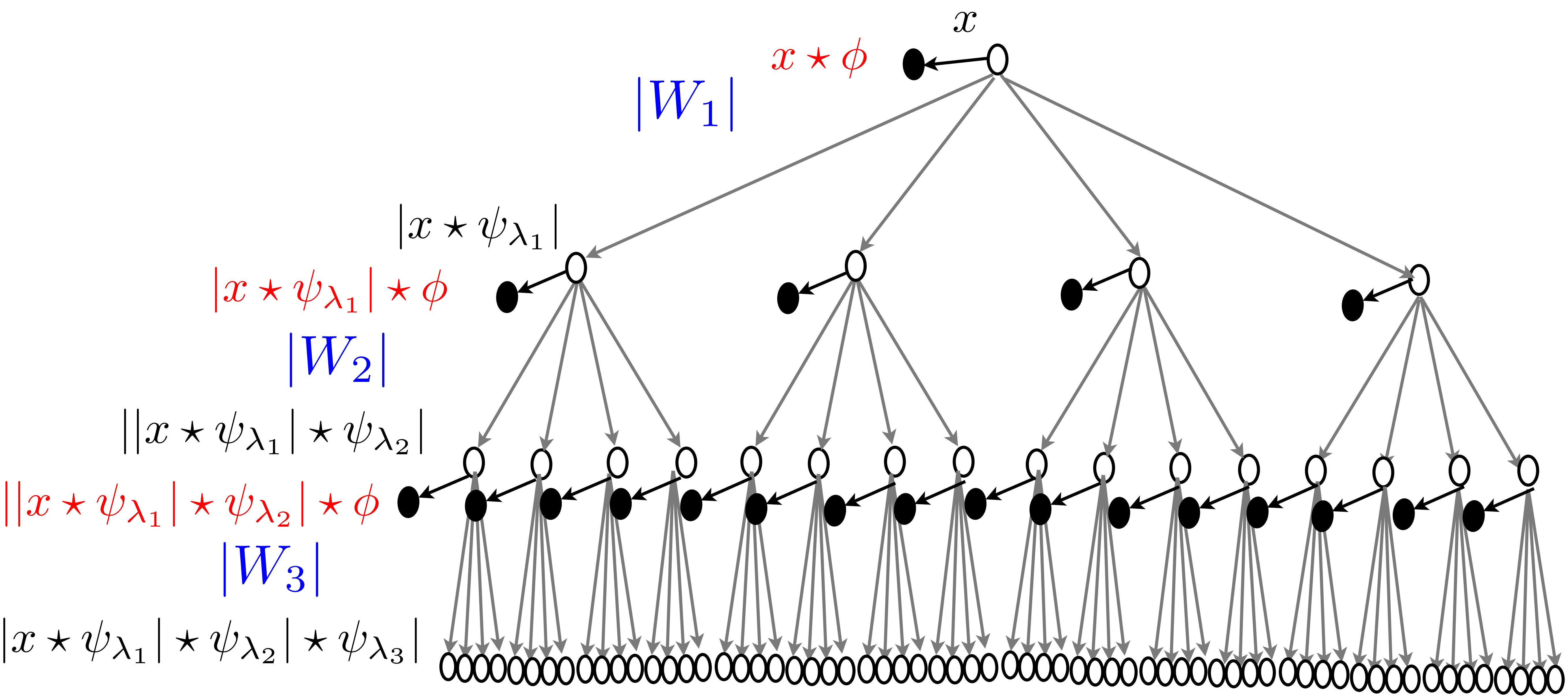
with $\lambda = (2^j, r)$



Wavelet transform $Wx = \Big(x \star \phi(t)\,,\, x \star \psi_\lambda(t)\Big)_\lambda$

Tight frame $\|Wx\|^2 = \|x \star \phi\|^2 + \sum_\lambda \|x \star \psi_\lambda\|^2 = \|x\|^2$

Local Scattering Transform



Scattering Properties

$$Sx = \begin{pmatrix} x \star \phi\,(u) \\ |x \star \psi_{\lambda_1}| \star \phi(u) \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(u) \\ |||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi(u) \\ \dots \end{pmatrix}_{u,\lambda_1,\lambda_2,\lambda_3,\dots}$$

$$\|Sx\|^2 = \sum_{m=0}^{\infty} \sum_{\lambda_1,\dots,\lambda_m} \Big\| |||x \star \psi_{\lambda_1}| \star \dots| \star \psi_{\lambda_m}| \star \phi \Big\|^2$$

Theorem: *For appropriate wavelets, a scattering is*

contractive $\|Sx - Sy\| \leq \|x - y\|$

preserves norms $\|Sx\| = \|x\|$

stable to deformations $x_\tau(t) = x(t - \tau(t))$

$$\|Sx - Sx_\tau\| \leq C \sup_t |\nabla \tau(t)| \, \|x\|$$

Contraction

$$Wx = \begin{pmatrix} x \star \phi(t) \\ x \star \psi_\lambda(t) \end{pmatrix}_{t,\lambda}$$ is linear and $\|Wx\| = \|x\|$

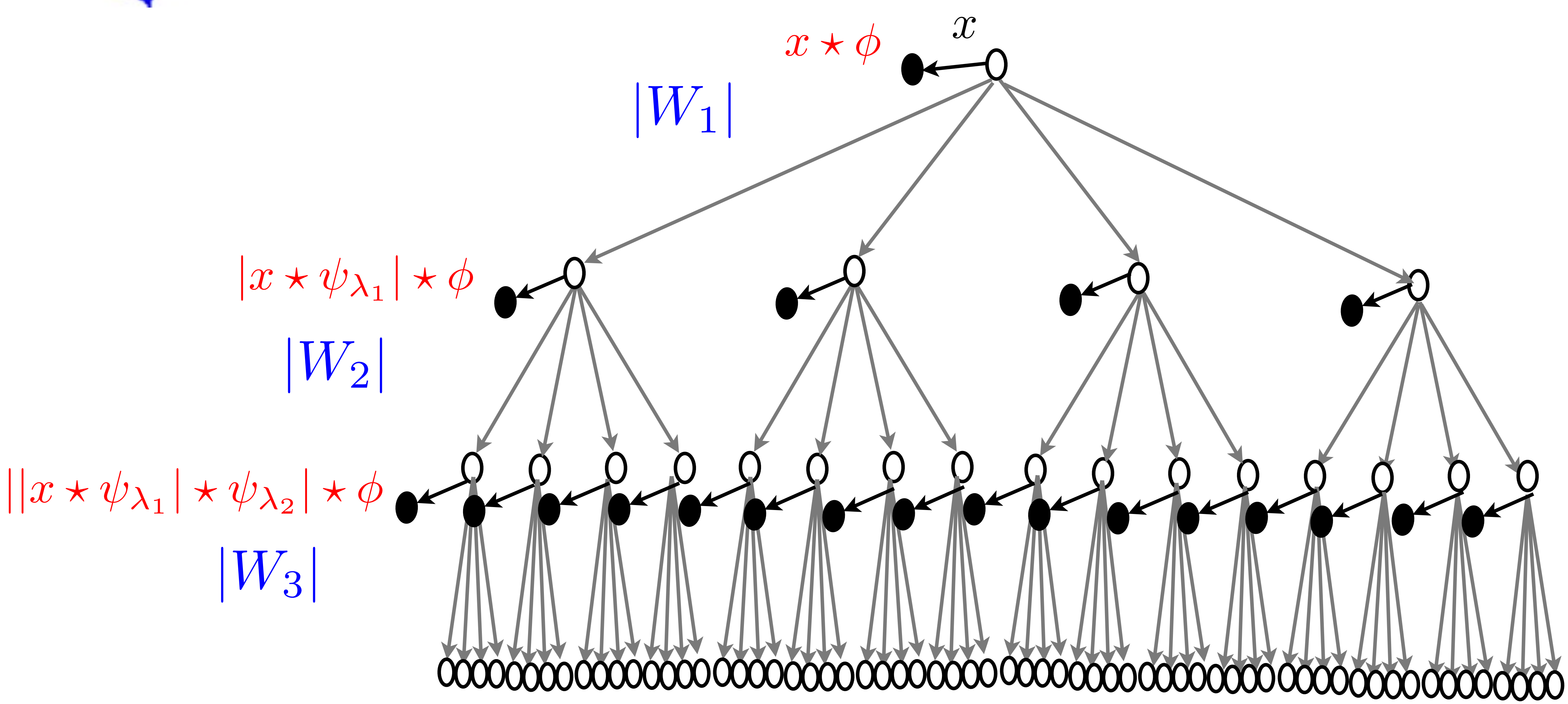
$$|W|x = \begin{pmatrix} x \star \phi(t) \\ |x \star \psi_\lambda(t)| \end{pmatrix}_{t,\lambda}$$ is non-linear

- it is contractive $\||W|x - |W|y\| \leq \|x - y\|$

because for $(a, b) \in \mathbb{C}^2$ $||a| - |b|| \leq |a - b|$

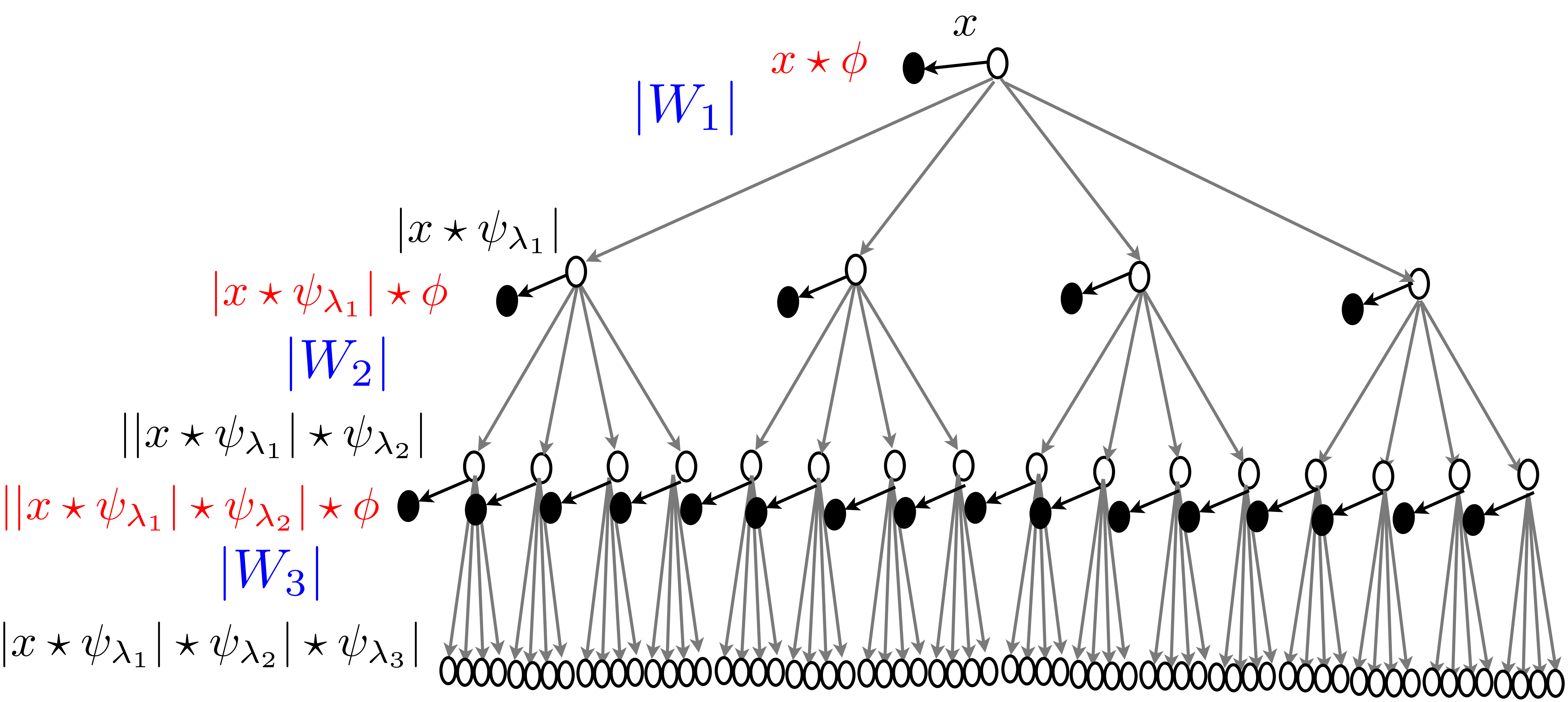
- it preserves the norm $\||W|x\| = \|x\|$

Scattering Contraction



- S is contractive because product of contractive operators.

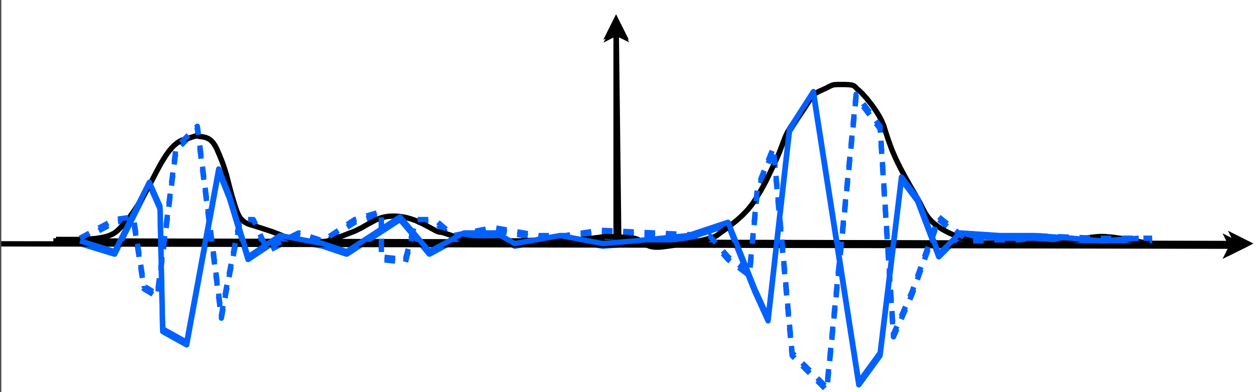
Scattering Energy Conservation

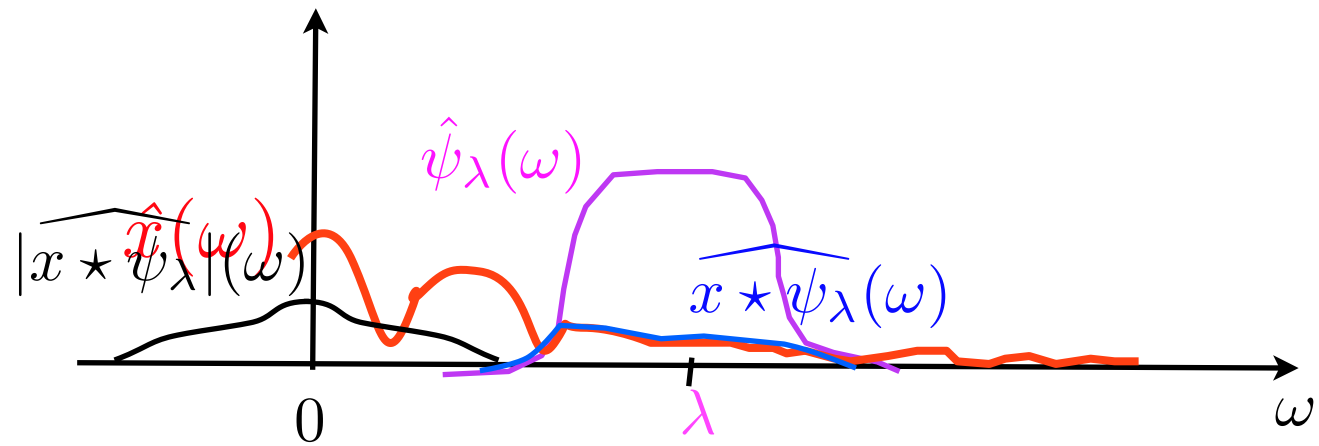


- S preserves the norm because inner layer energy converge to zero as the depth increases.

Modulus «Demodulation»

$$x \star \psi_{\lambda_1}(t) = x \star \psi^a_{\lambda_1}(t) + i\, x \star \psi^b_{\lambda_1}(t)$$





- The modulus $|x \star \psi_{\lambda_1}|$ is a regular lower frequency envelop

Modulus shift wavelet coefficient energy to low frequencies.

Lipschitz Stability to Deformations

Wavelet transforms "nearly commute" with deformations:

$$D_\tau x(t) = x(t - \tau(t))$$

Commutator operator:

$$[W, D_\tau] = W\, D_\tau - D_\tau\, W$$

Lemma :

$$\|\,[W, D_\tau]\,\| \leq C \sup_t |\nabla\tau(t)|\ .$$

and $\|\,[|W|, D_\tau]\,\| \leq \|\,[W, D_\tau]\,\|$

because modulus commutes with diffeomorphisms.

Fourier versus Scattering

Frequencies $\omega = m\xi$

$$e^{im\xi t} x(t) = e^{i\xi t} \ldots e^{i\xi t} e^{i\xi t} x(t)$$

Countable frequency set

Local Fourier:

$$\int e^{im\xi u}\, x(u)\, \phi(t-u) du$$

$\hat{\phi}(\omega)$ in $[-\xi, \xi]$

Fourier transform:

$$\hat{x}(\omega) = \int e^{i\omega u}\, x(u)\, du$$

$\hat{\delta}(\omega) = 1$

Frequency set: $\mathbb{R}$

$\phi \longrightarrow 1$

$\xi \longrightarrow 0$

Paths $p = (\lambda_1, \lambda_2 ..., \lambda_m)$

$$|\, |x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \ldots | \star \psi_{\lambda_m}(t)|$$

Countable path set for $\lambda_k = 2^{j_k} \geq \xi$

Local scattering:

$$\int |\, |x \star \psi_{\lambda_1}| ... | \star \psi_{\lambda_m}(u)|\, \phi(t-u)\, du\ .$$

Scattering transform:

$$\overline{S}x(p) = \mu_p^{-1} \int |\, |x \star \psi_{\lambda_1}| ... | \star \psi_{\lambda_m}(u)|\, du$$

$\overline{S}\delta(p) = 1$

Path set $\mathcal{P} \sim \mathbb{Z}^{\mathbb{N}} \sim \mathbb{R}$

Scattering Transform

Theorem S *converges weakly to* $\overline{S}$ *when* ϕ *goes to* 1

There exists a measure $d\mu$ *on* $\mathcal{P}$ *such that*

$$\forall x \in \mathbf{L^2}(\mathbf{R}) \quad , \quad \overline{S}x(p) \in \mathbf{L^2}(\mathcal{P}, d\mu)$$

$$\int_{\mathcal{P}} |\overline{S}x(p)|^2 \, d\mu(p) < \infty \ .$$

We know that $\|Sx\|^2 = \|x\|^2$ and $\lim_{\phi \to 1} S = \overline{S}$

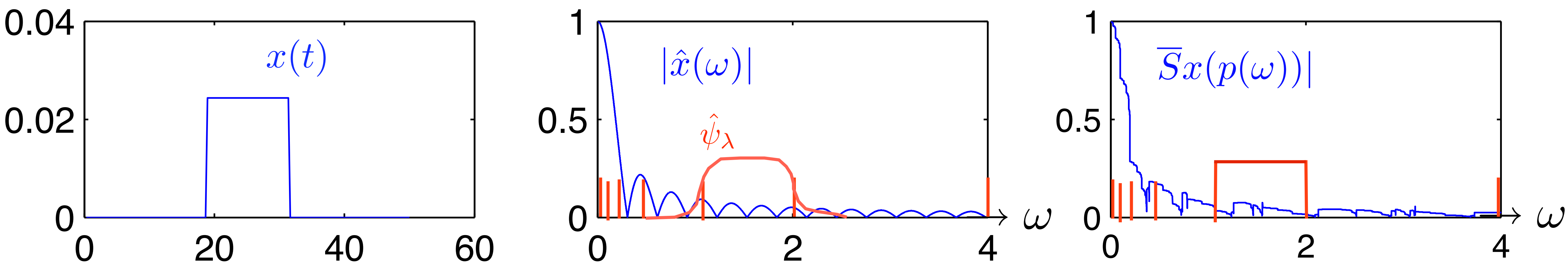
Conjecture: $$\int_{\mathcal{P}} |\overline{S}x(p)|^2 \, d\mu(p) = \|x\|^2 \ .$$

Frequency to Paths Mapping

$$\mathbb{R} \longrightarrow \mathcal{P}$$

$$\omega \longrightarrow p(\omega) \quad \text{with } d\mu(p(\omega)) = d\omega$$

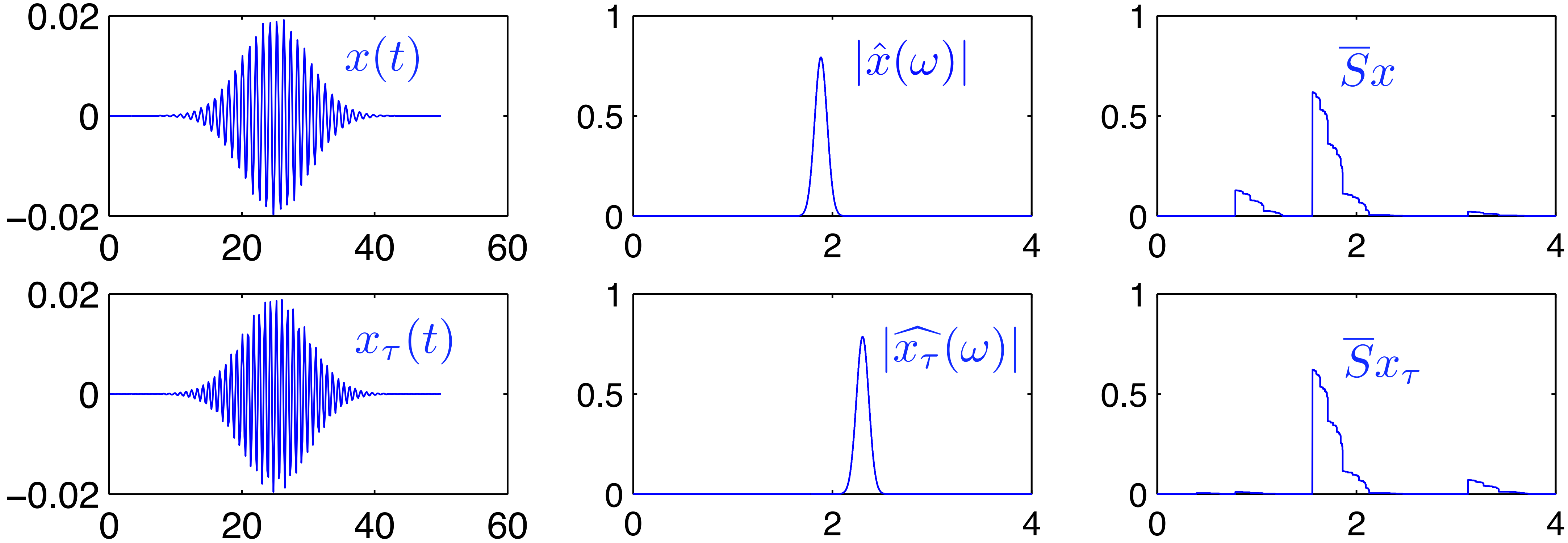
$$\omega \longrightarrow \overline{S}x(p(\omega))$$



$$\int |\hat{x}(\omega)|^2 \, |\hat{\psi}(2^j\omega)|^2 \, d\omega = \int_{2^j\pi}^{2^{j+1}\pi} |\overline{S}x(p(\omega))|^2 \, d\omega$$

Scattering Integral Examples

$$x_\tau(t) = x(t - \tau(t)) \quad \text{with} \quad \tau(t) = \epsilon\, t\ .$$



$\hat{\psi}_\lambda$

$$\frac{\|\,|\hat{x}| - |\hat{x}_\tau|\,\|}{\|x\|\,\|\tau'\|_\infty} = 13 \qquad \frac{\|\overline{S}x - \overline{S}x_\tau\|_{\mathcal{P}}}{\|x\|\,\|\tau'\|_\infty} = 1.4$$

Fourier transforms maps regularity and decay and vice-versa.

What notion of regularity defined by the scattering decay ?

Depends on the sparsity/geometry of wavelet coefficients.

Image Scattering Transforms

Image

$x(t)$

$t = (t_1, t_2)$

Fourier Modulus

$|\hat{x}(\omega)|$

$\omega = (\omega_1, \omega_2)$

Scattering $\phi(t) = 1$

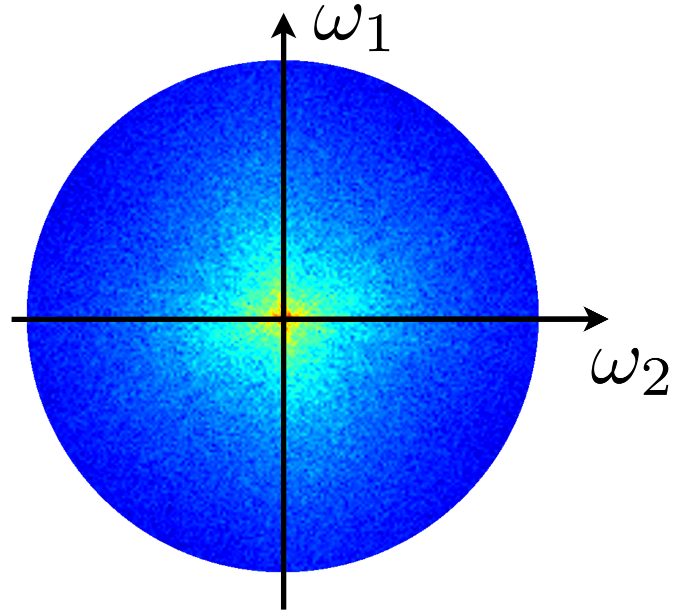
$|x \star \psi_{\lambda_1}| \star \phi$ $\quad$ $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi$

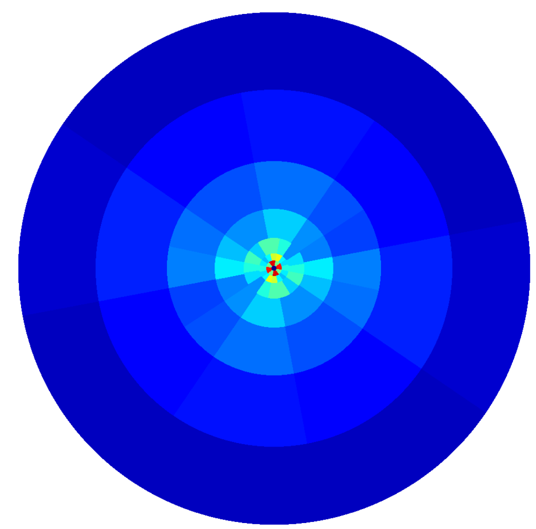
$\|x \star \psi_{\lambda_1}\|_1$ $\quad$ $\||x \star \psi_{\lambda_1}| \star \psi_{2^{j_2}}\|_1$

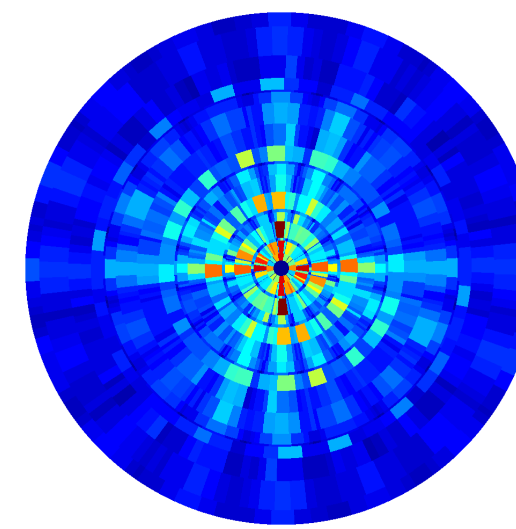
$\lambda_1 = 2^{j_1}\, r_{\theta_1}$

$\lambda_1 = 2^{j_1}\, r_{\theta_1}$

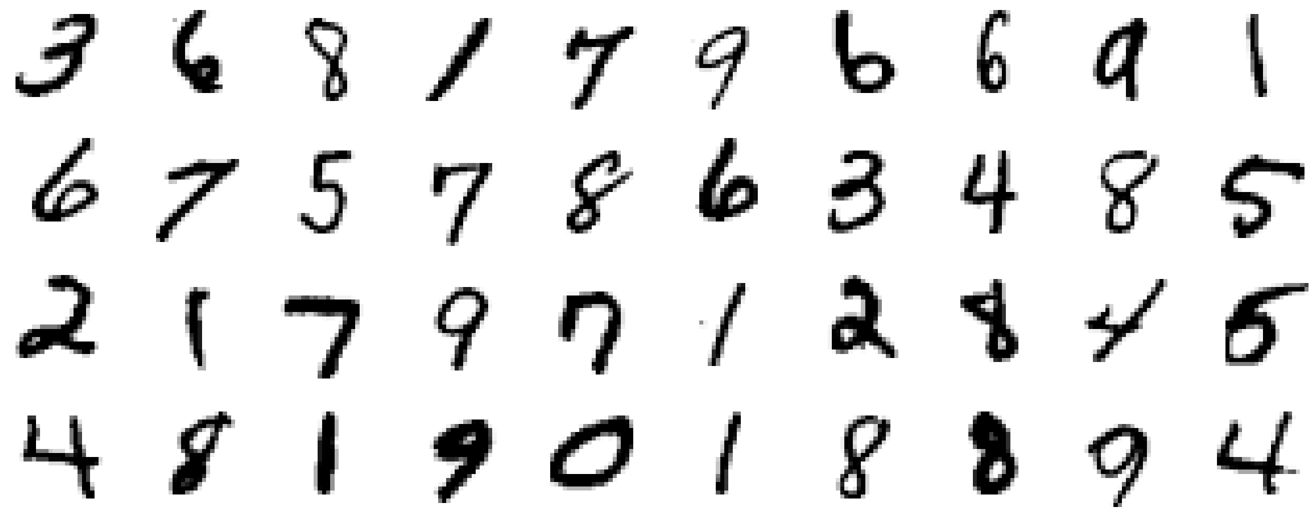
$\lambda_2 = 2^{j_2}\, r_{\theta_2}$







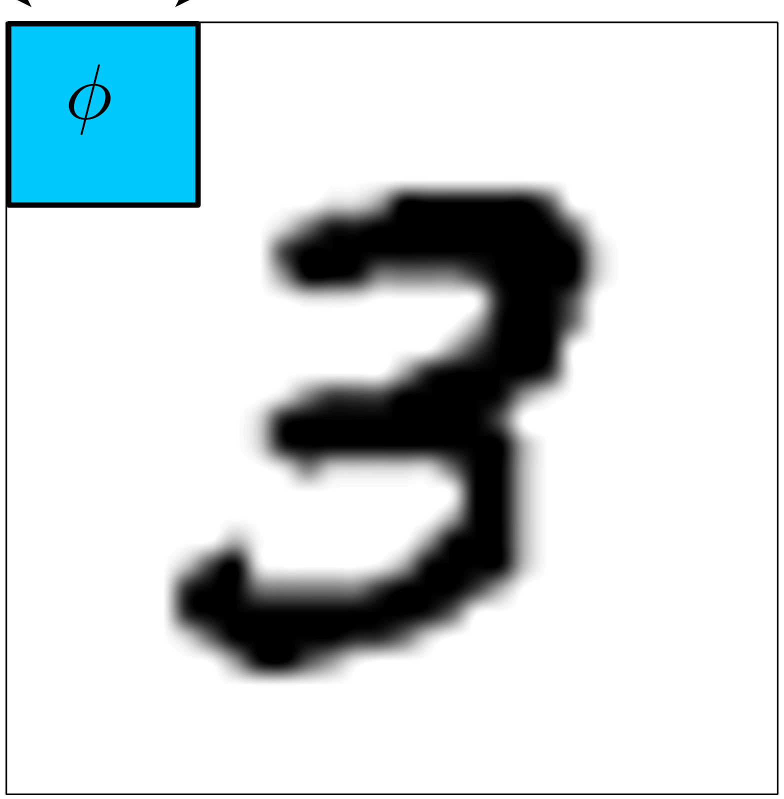
Digit Classification: MNIST



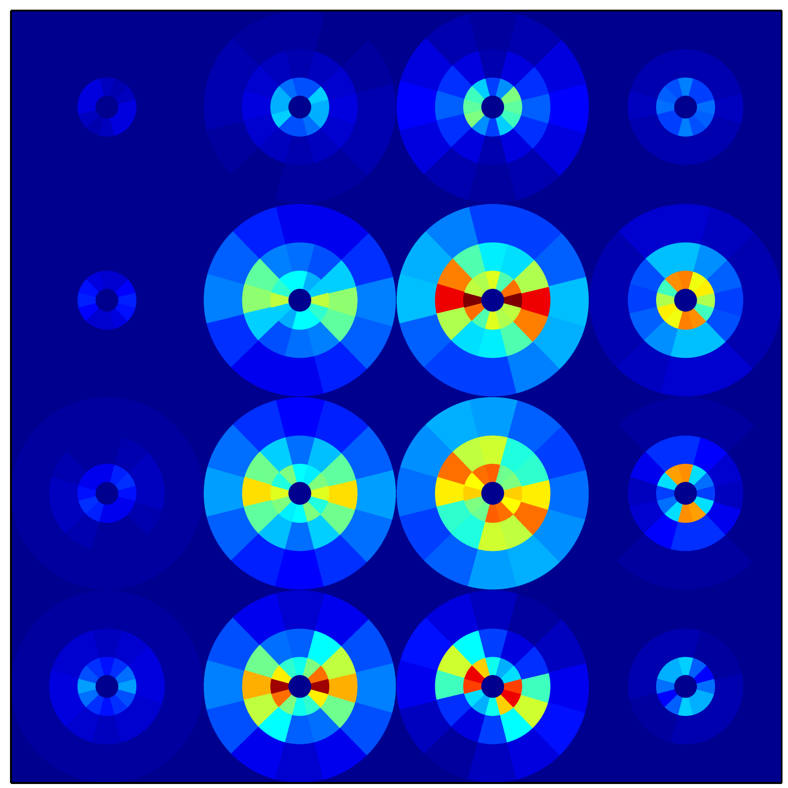
Digit Classification: MNIST

Second order Scattering Sx:

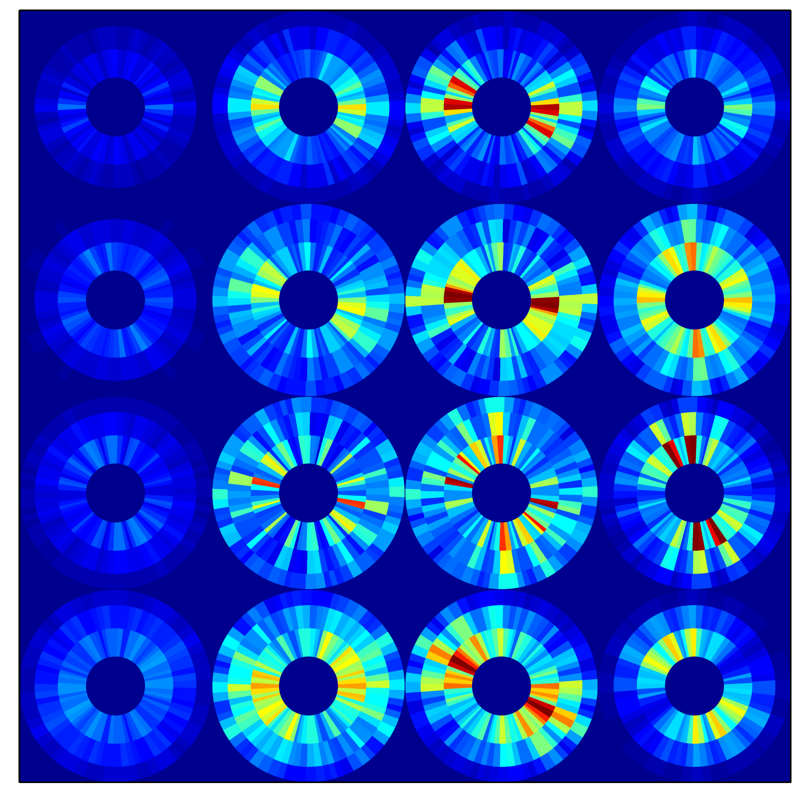
2^J $\quad x(t)$



$|x \star \psi_{\lambda_1}| \star \phi(2^J n)$

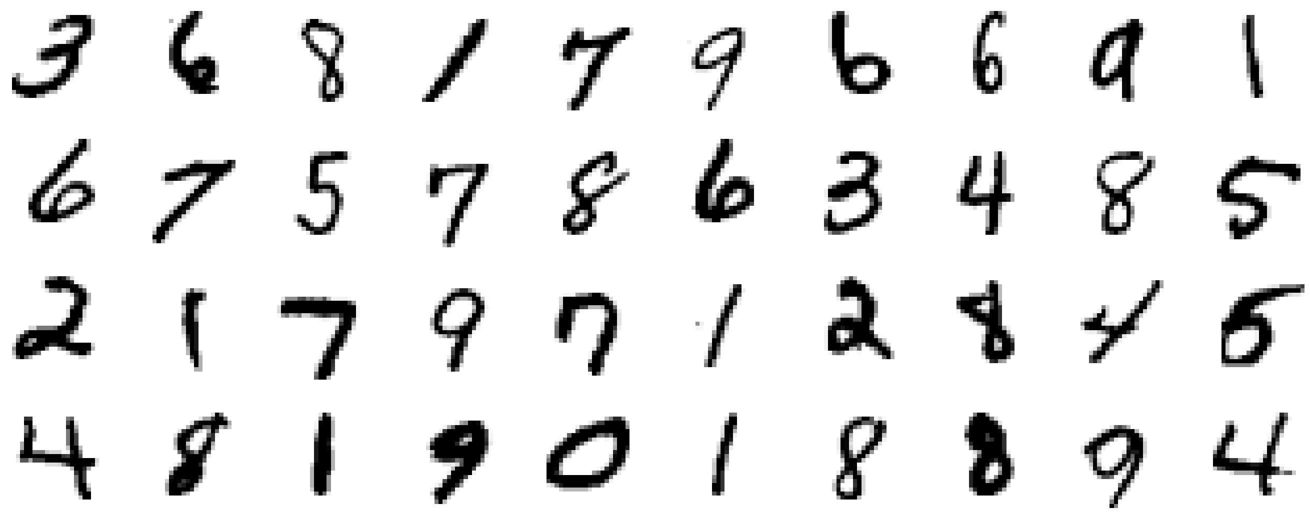


$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(2^J n)$



Digit Classification: MNIST

Joan Bruna



$$x \longrightarrow \boxed{Sx} \longrightarrow \boxed{\text{Supervised Linear Classifier: PCA/SVM}} \longrightarrow y$$

Classification Errors

Training size	Conv. Net.	Scattering
300	7.2%	**4.4**%
5000	1.5%	**1.0**%
20000	0.8%	**0.6**%
60000	0.5%	**0.4**%

LeCun et. al.

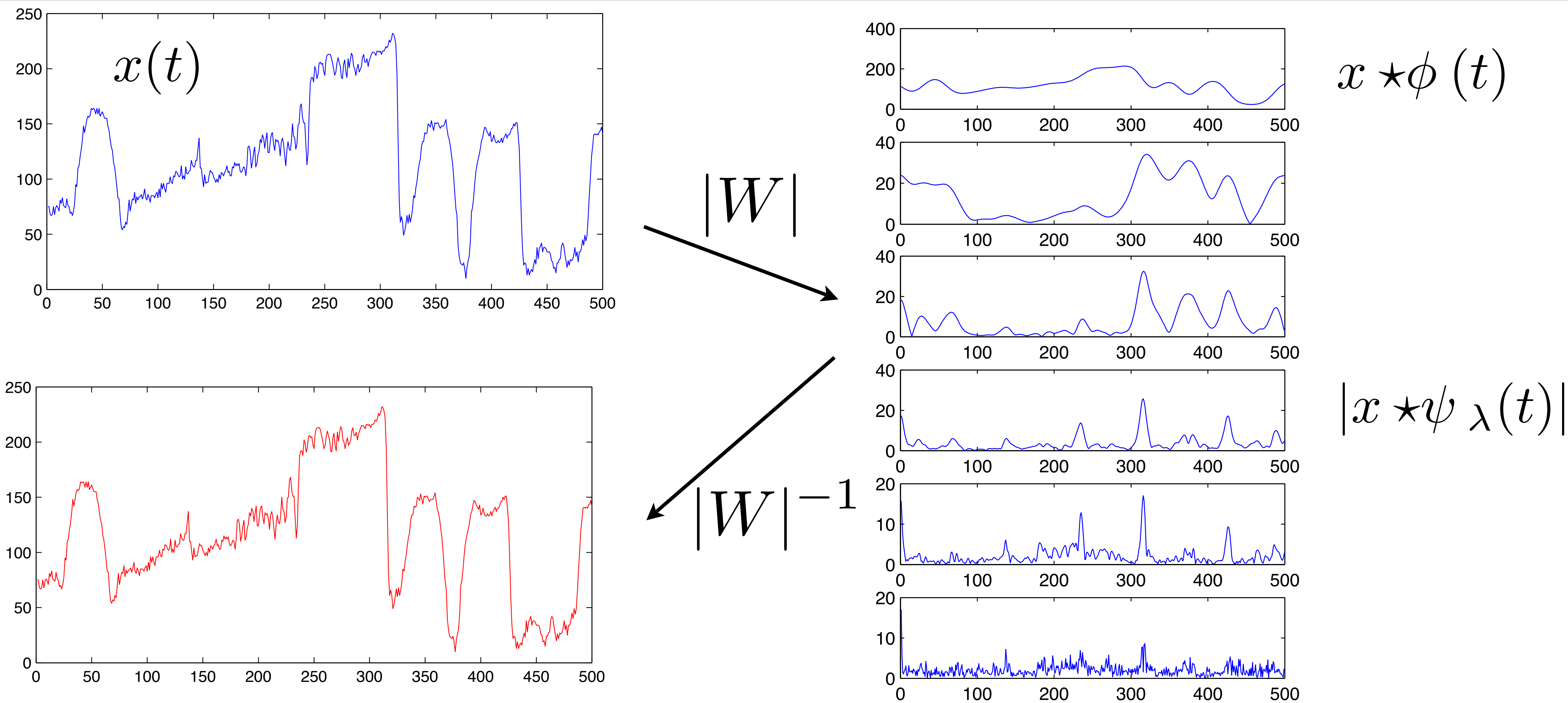
Scattering Inversion: Phase Recovery

$$Wx = \left\{ x \star \phi \,,\, x \star \psi_\lambda \right\}_\lambda$$ is linear and unitary.

Theorem For appropriate wavelets *I. Waldspurger*

$$|W|x = \left\{ x \star \phi \,,\, |x \star \psi_\lambda| \right\}_\lambda$$

is invertible and the inverse is continuous.



Scattering Inversion: Phase Recovery

I. Waldspurger

Theorem For appropriate wavelets

$$|W|x = \Big\{ x \star \phi \,,\, |x \star \psi_\lambda| \Big\}_\lambda$$

is invertible and the inverse is continuous.

Inverse scattering:

$$x$$

$$\uparrow |W|^{-1}$$

$$\Big\{ \underline{x \star \phi} \,,\, |x \star \psi_{\lambda_1}| \Big\}_{\lambda_1}$$

$$\uparrow |W|^{-1}$$

$$\Big\{ \underline{|x \star \psi_{\lambda_1}| \star \phi} \,,\, ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \Big\}_{\lambda_1,\lambda_2}$$

$$\uparrow |W|^{-1}$$

Propagation of errors

$$\Big\{ \underline{||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi} \,,\, |||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \Big\}_{\lambda_1,\lambda_2,\lambda_3}$$

$$\uparrow |W|^{-1}$$

$$\Big\{ \underline{|||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi} \,,\, ||||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \psi_{\lambda_4}| \Big\}_{\lambda_1,\lambda_2,\lambda_3,\ldots}$$

Audio Reconstruction

J. Anden

Original audio signal x

Reconstruction from Sx for an averaging window ϕ of 1 s

from 1st layer coefficients $|x \star \psi_{\lambda_1}| \star \phi$

adding 2nd layer coefficients $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi$

Expected Scattering Transform

- If $X(t)$ is a stationary process then

  $||X \star \psi_{\lambda_1}| \star ...| \star \psi_{\lambda_m}(t)|$ is also stationary.

Scattering :

$$SX(t) = \begin{pmatrix} X \star \phi(t) \\ |X \star \psi_{\lambda_1}| \star \phi(t) \\ ||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t) \\ |||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi(t) \\ ... \end{pmatrix}_{\lambda_1,\lambda_2,\lambda_3,...}$$

- When $\phi \to 1$ with "appropriate" ergodicity conditions"

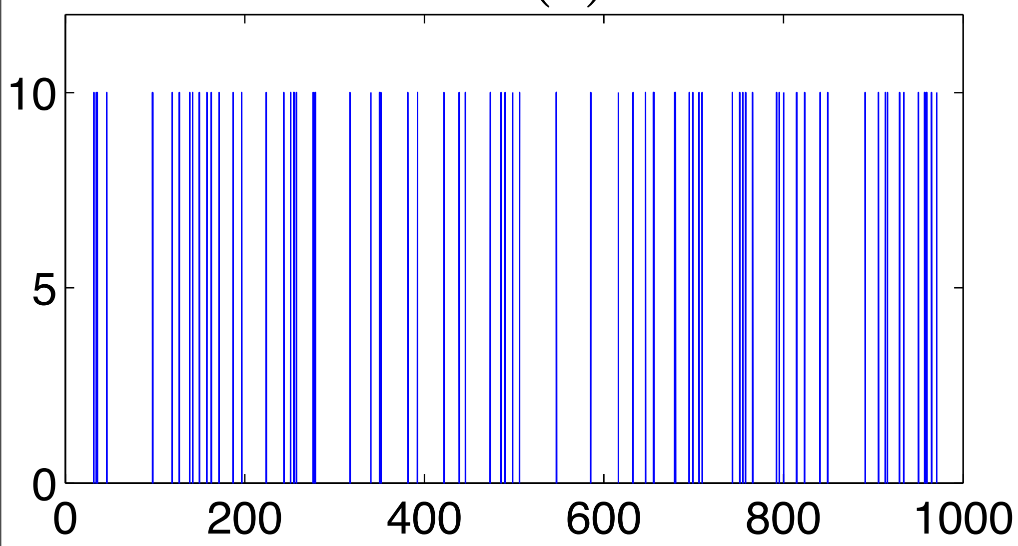
$SX(t)$ may converge to the expected scattering transform:

$$\overline{S}X = \begin{pmatrix} E(X) \\ E(|X \star \psi_{\lambda_1}|) \\ E(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) \\ E(|||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}|) \\ ... \end{pmatrix}_{\lambda_1,\lambda_2,\lambda_3,...}$$

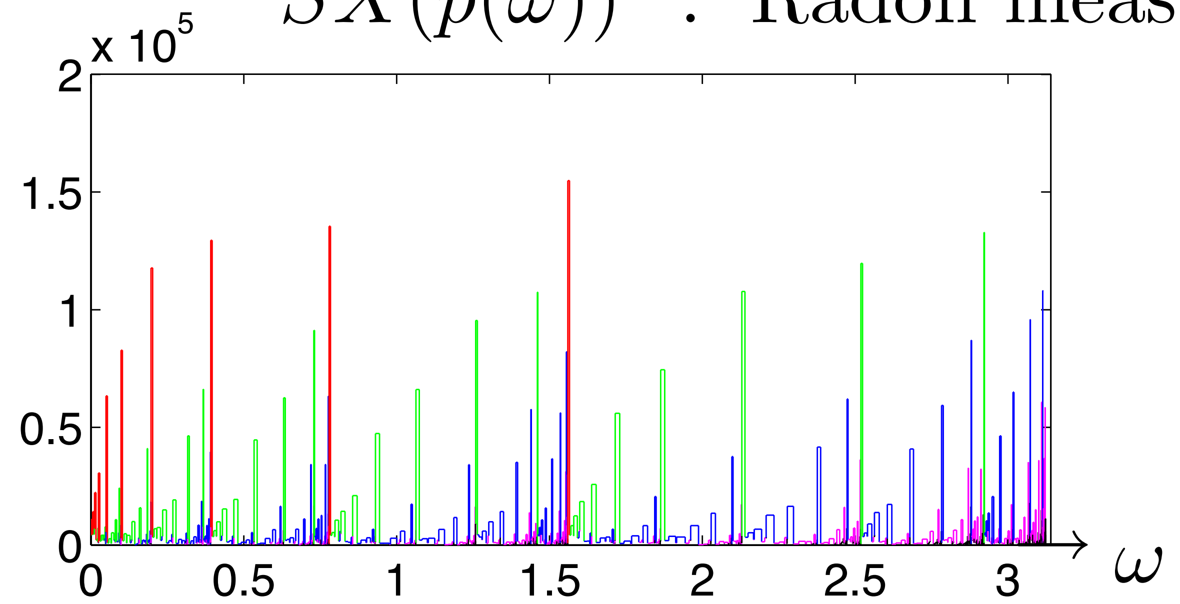
Scattering White Noises

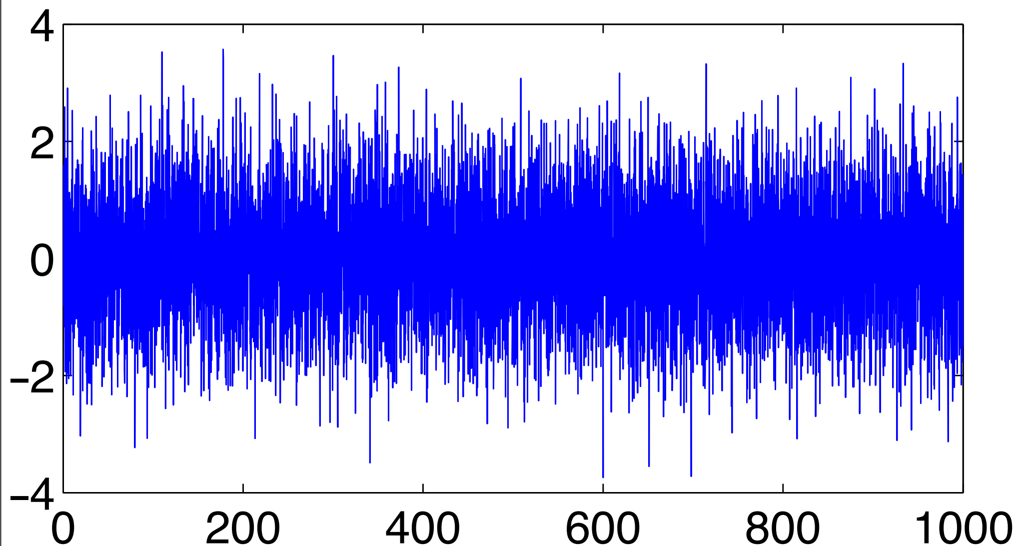
Constant Fourier power spectrum: $\hat{R}_X(\omega) = \sigma^2$.

Bernoulli $X(t)$

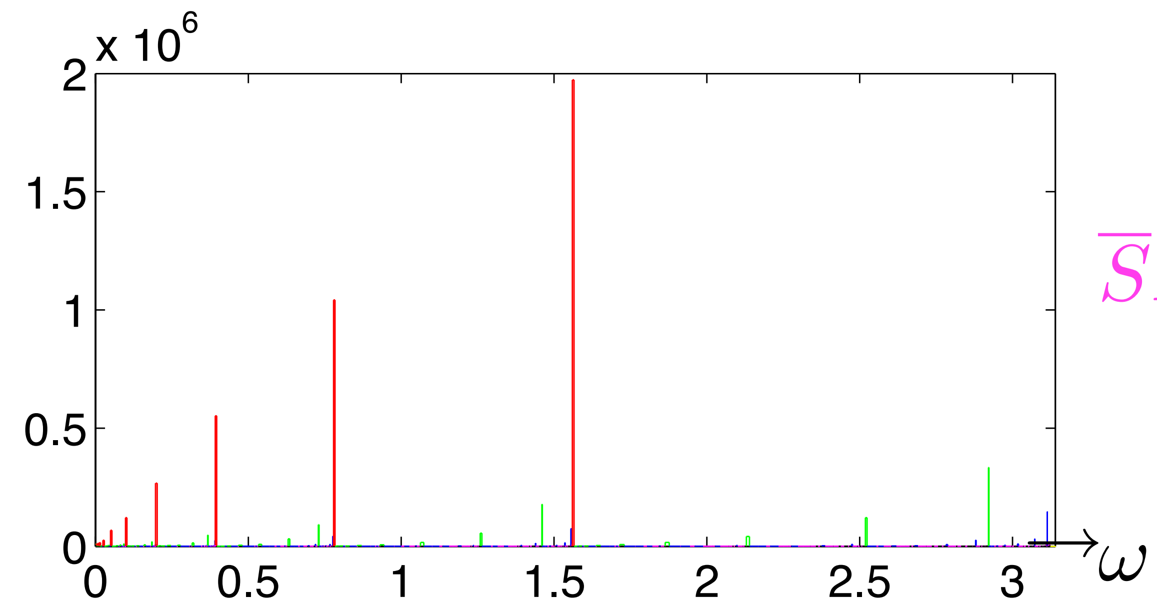


$\overline{S}X(p(\omega))^2$: Radon measure





Gaussian White



$\overline{S}X(\lambda_1)$

$\overline{S}X(\lambda_1, \lambda_2)$

$\overline{S}X(\lambda_1, \lambda_2, \lambda_3)$

$\overline{S}X(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$

Wavelet Tight Frames in L2

Functions in $\mathbf{L^2}(\mathbb{R}^d)$: $\|x\|^2 = \int |x(t)|^2\, dt < \infty$

Wavelet transform: $Wx = \begin{pmatrix} x \star \phi(t) \\ x \star \psi_\lambda(t) \end{pmatrix}_{t,\lambda}$

Proposition: *(Littlewood-Paley)*

The wavelet transform is a tight frame for $x \in \mathbf{L^2}(\mathbb{R}^d)$

$$\|Wx\|^2 = \|x \star \phi\|^2 + \sum_\lambda \|x \star \psi_\lambda\|^2 = \|x\|^2$$

if and only if for almost all ω.

$$|\hat{\phi}(\omega)|^2 + \frac{1}{2} \sum_\lambda \left(|\hat{\psi}_\lambda(\omega)|^2 + |\hat{\psi}_\lambda(-\omega)|^2 \right) = 1$$

Wavelet Frames of Processes

Stationary processes $X(t)$ with $\mathbb{E}(|X(t)|^2) < \infty$.

Wavelet transform: $WX = \begin{pmatrix} \mathbb{E}(X) \\ X \star \psi_\lambda(t) \end{pmatrix}_{t,\lambda}$

Proposition: *(Littlewood-Paley)*

The wavelet transform preserves the variance of stationary X

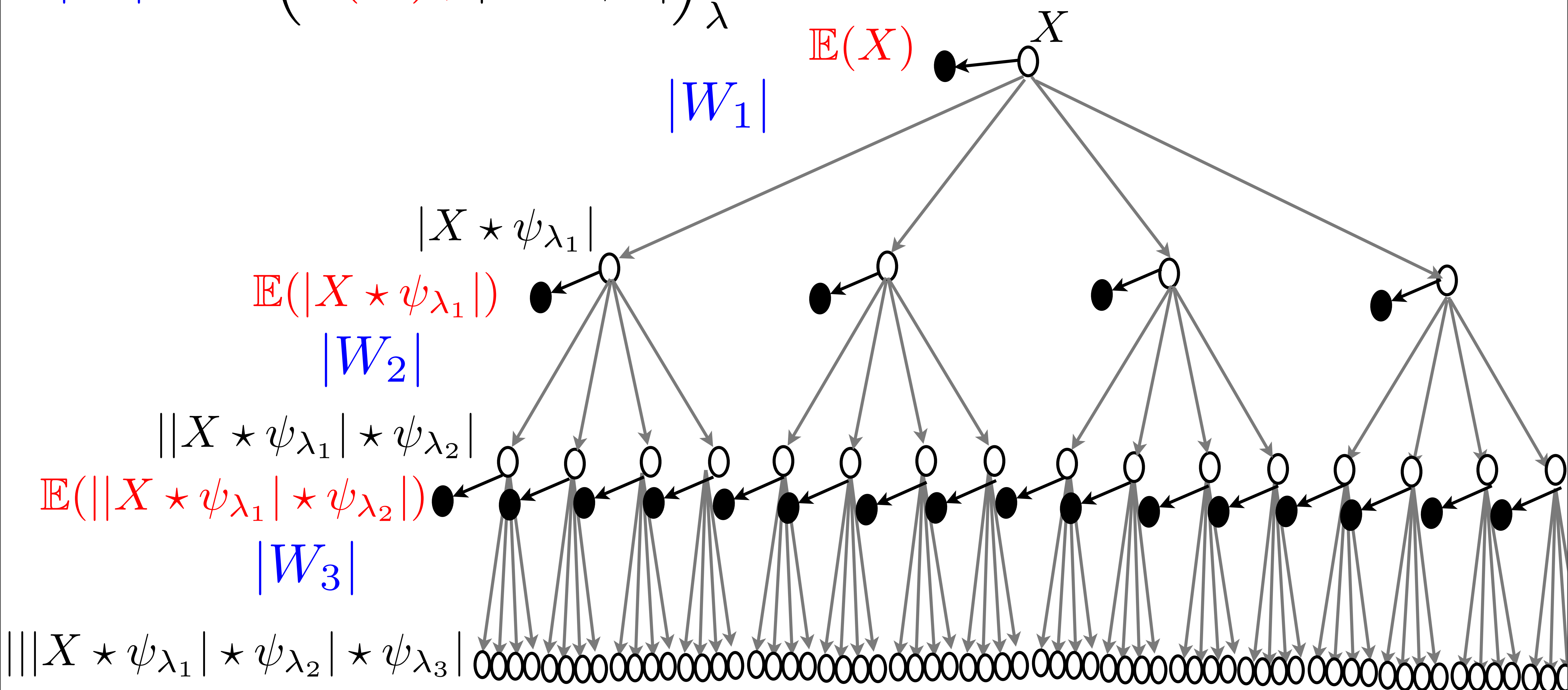
$$\mathbb{E}(X)^2 + \sum_\lambda \mathbb{E}(|X \star \psi_\lambda|^2) = \mathbb{E}(|X|^2)$$

if and only if for almost all ω.

$$\frac{1}{2} \sum_\lambda \Big(|\hat{\psi}_\lambda(\omega)|^2 + |\hat{\psi}_\lambda(-\omega)|^2 \Big) = 1$$

Expected Scattering Transform

$$|W|X = \Big(\mathbb{E}(X)\,,\ |X \star \psi_\lambda|\Big)_\lambda$$



- S preserves is contractive because each $|W_k|$ are contractive

Expected Scattering Transform

$X(t)$ stationary process:

$$\overline{S}X = \begin{pmatrix} E(X) \\ E(|X \star \psi_{\lambda_1}|) \\ E(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) \\ E(|||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}|) \\ \dots \end{pmatrix}_{\lambda_1,\lambda_2,\lambda_3,\dots}$$

$$\|\overline{S}X\|^2 = \mathbb{E}(X)^2 + \sum_{m=1}^{\infty} \sum_{\lambda_1,\dots,\lambda_m} \mathbb{E}\Big(|||X \star \psi_{\lambda_1}| \star \dots| \star \psi_{\lambda_m}|\Big)^2$$

Theorem: *A scattering is*

contractive $\|\overline{S}X - \overline{S}Y\|^2 \leq E(|X-Y|^2)$

stable to stationary deformations $X_\tau(t) = X(t-\tau(t))$

$$\|\overline{S}X - \overline{S}X_\tau\| \leq C \sup_t |\nabla\tau(t)|\, E(|X|^2)^{1/2} \;.$$

Textures with Same Spectrum

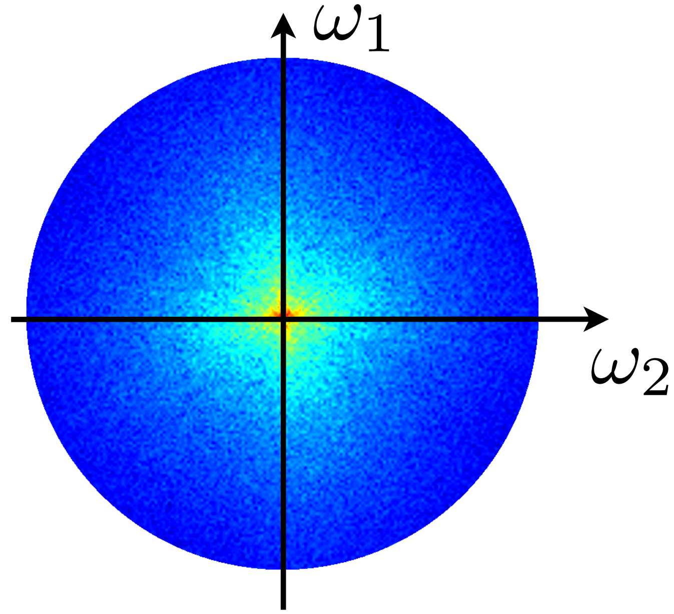
$X(t)$
stationary process

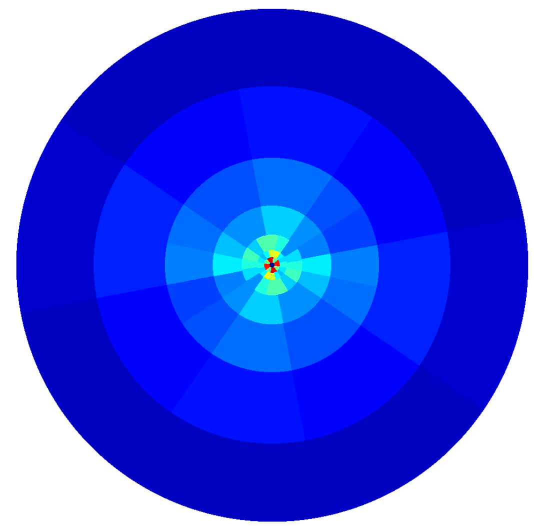
$\hat{R}_X(\omega)$
Power Spectrum

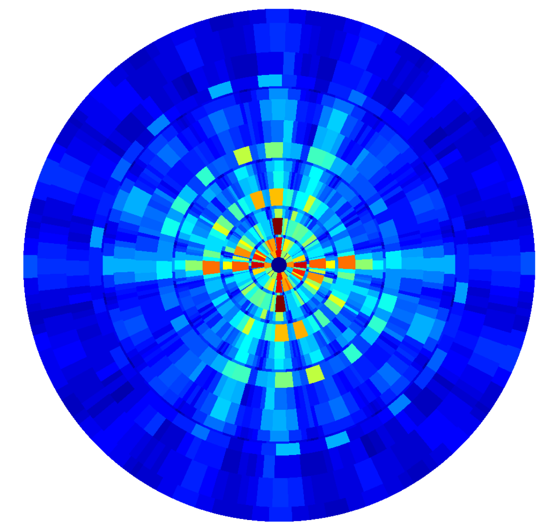
Estimated Expected Scattering

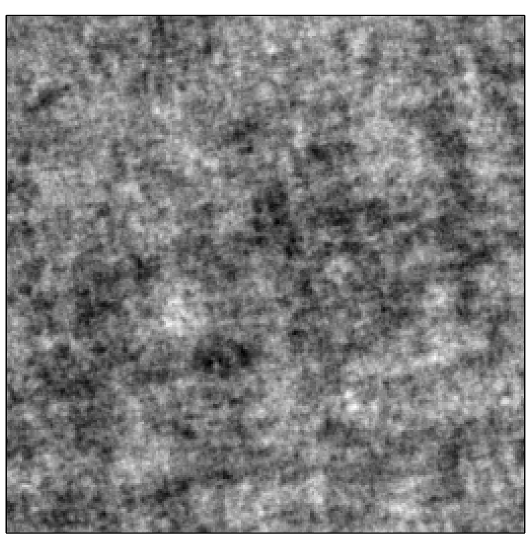
$|X \star \psi_{\lambda_1}| \star \phi$

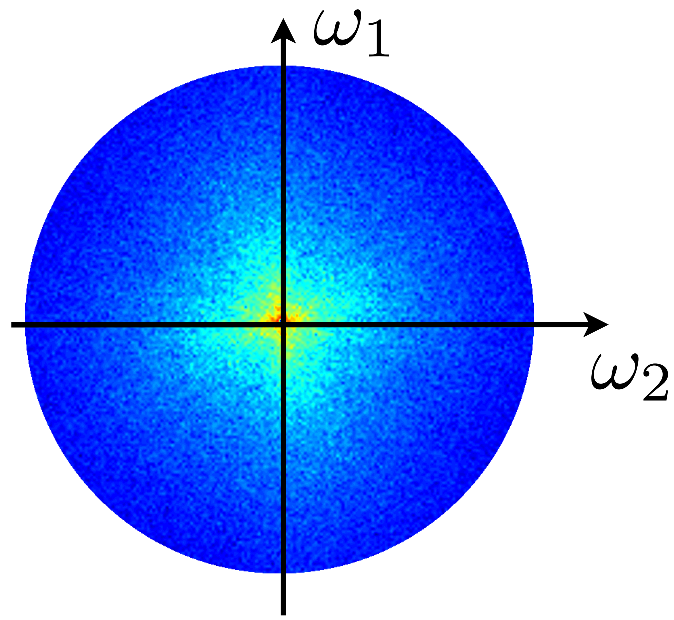
$||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi$

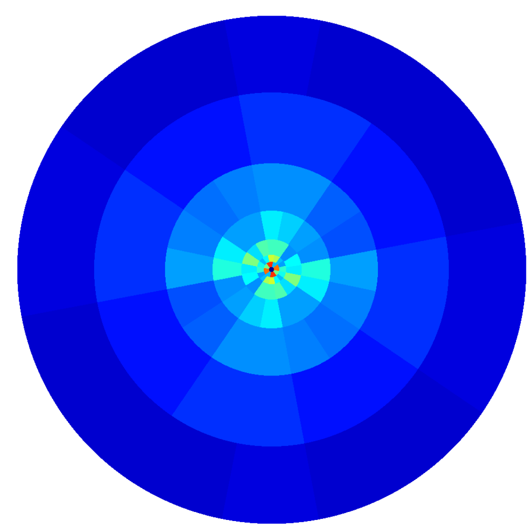


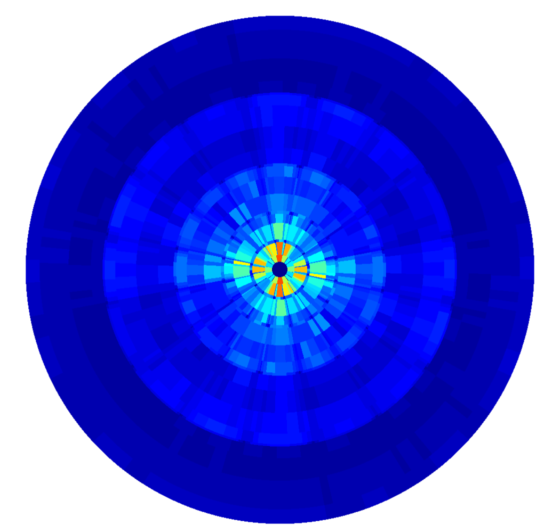






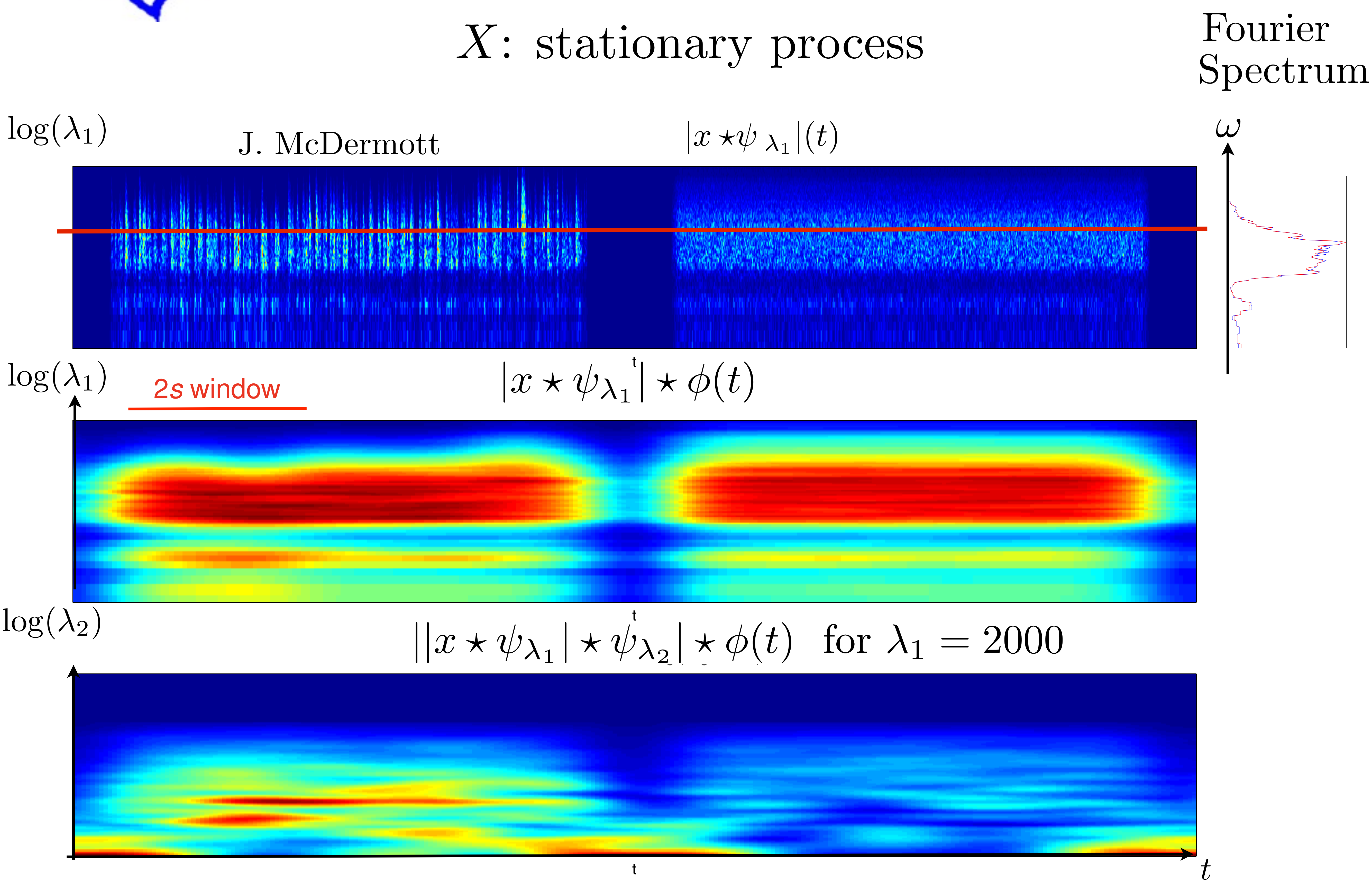






Sounds with Same Spectrum

X: stationary process



Mean-Square Consistency

- Empricial average scattering coefficients

$$|||X \star \psi_{\lambda_1}| \star \ldots| \star \psi_{\lambda_m}| \star \phi$$

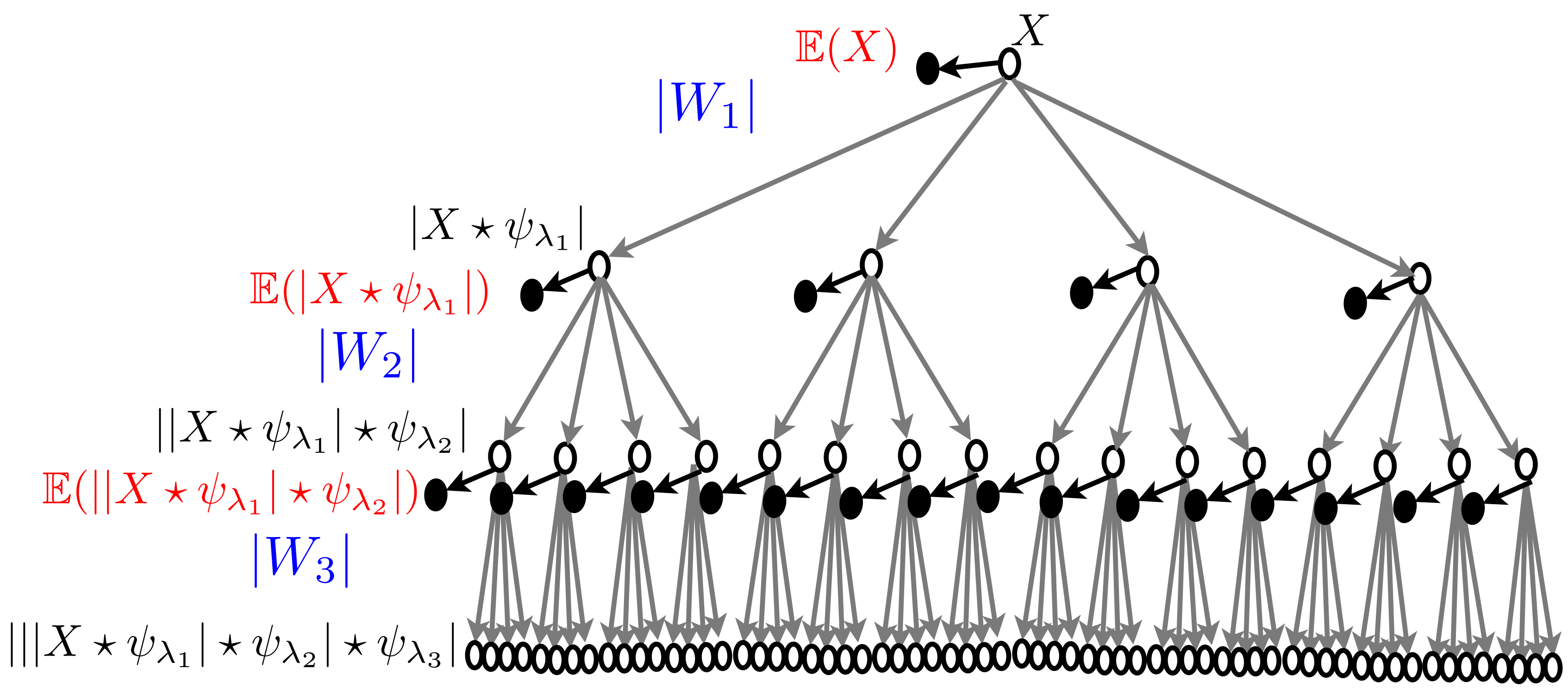
are unbiased estimators of

$$\mathbb{E}\Big(|||X \star \psi_{\lambda_1}| \star \ldots| \star \psi_{\lambda_m}|\Big)$$

- A scattering is mean-square consistent if

$$\lim_{\phi \to 1} \sum_{m=0}^{\infty} \sum_{\lambda_1,\ldots,\lambda_m} \mathbb{E}\Big(|||X \star \psi_{\lambda_1}| \star \ldots| \star \psi_{\lambda_m}| \star \phi - \mathbb{E}(|||X \star \psi_{\lambda_1}| \star \ldots| \star \psi_{\lambda_m}|)\Big)^2 = 0$$

Expected Scattering Transform



Theorem For any stationary X, equivalent propositions:

(i) The scattering transform is mean-square consistent.

(ii) $\|\overline{S}X\|^2 = E(|X|^2)$

(iii) $$\lim_{m \to \infty} \sum_{\lambda_1, \ldots, \lambda_m} \mathbb{E}\Big(|||X \star \psi_{\lambda_1}| \ldots \star \psi_{\lambda_m}|\Big)^2 = 0$$

- Numerically always verified but not proved.

Representation of Random Processes

- An expected scattering is a non-complete representation

$$\overline{S}X = \left(\begin{array}{rcl} E(X) & = & E(U_0 X) \\ E(|X \star \psi_{\lambda_1}|) & = & E(U_1 X) \\ E(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) & = & E(U_2 X) \\ E(|||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}|) & = & E(U_3 X) \\ \ldots & & \end{array} \right)_{\lambda_1,\lambda_2,\lambda_3,\ldots}$$

Theorem *(Boltzmann)* The distribution $p(x)$ which satisfies

$$\int_{\mathbb{R}^N} U_m x \; p(x)\, dx = E(U_m X)$$

and maximizes the entropy $-\int p(x) \, \log p(x)\, dx$

can be written: $\quad p(x) = \frac{1}{Z} \exp\Big(\sum_{m=1}^{\infty} \lambda_m \, . \, U_m x \Big)$

Synthesis from Second Order

J. McDermott textures

Joakim Anden *Joan Bruna*

- Maximum entropy estimation of $X(t)$:
  - Gaussian model from 2nd order moments (N power spectrum coefficients)
  - Scattering model 1st & 2nd orders ($(\log_2 N)^2$ coefficients)

- Original jackhammer
- Gaussian model
- Scattering model

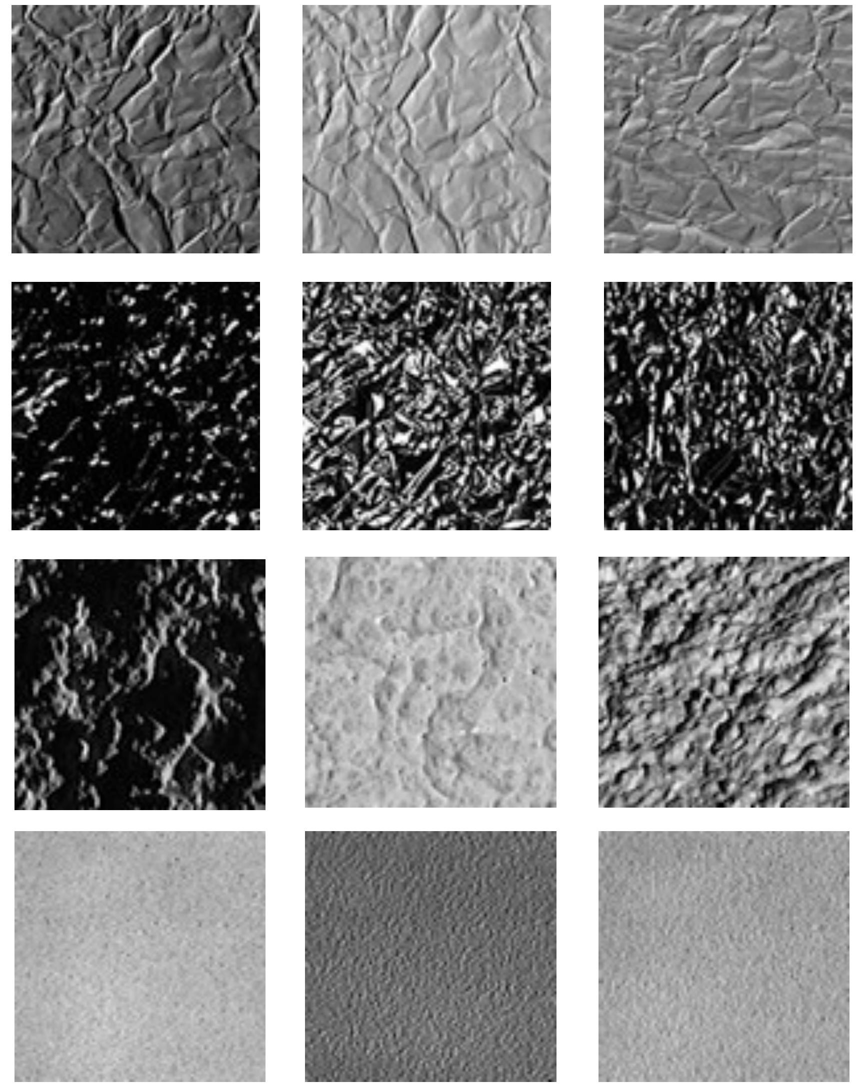
- Original water
- Gaussian model
- Scattering model

- Original applause
- Gaussian model
- Scattering model

Classification of Textures

CUREt database

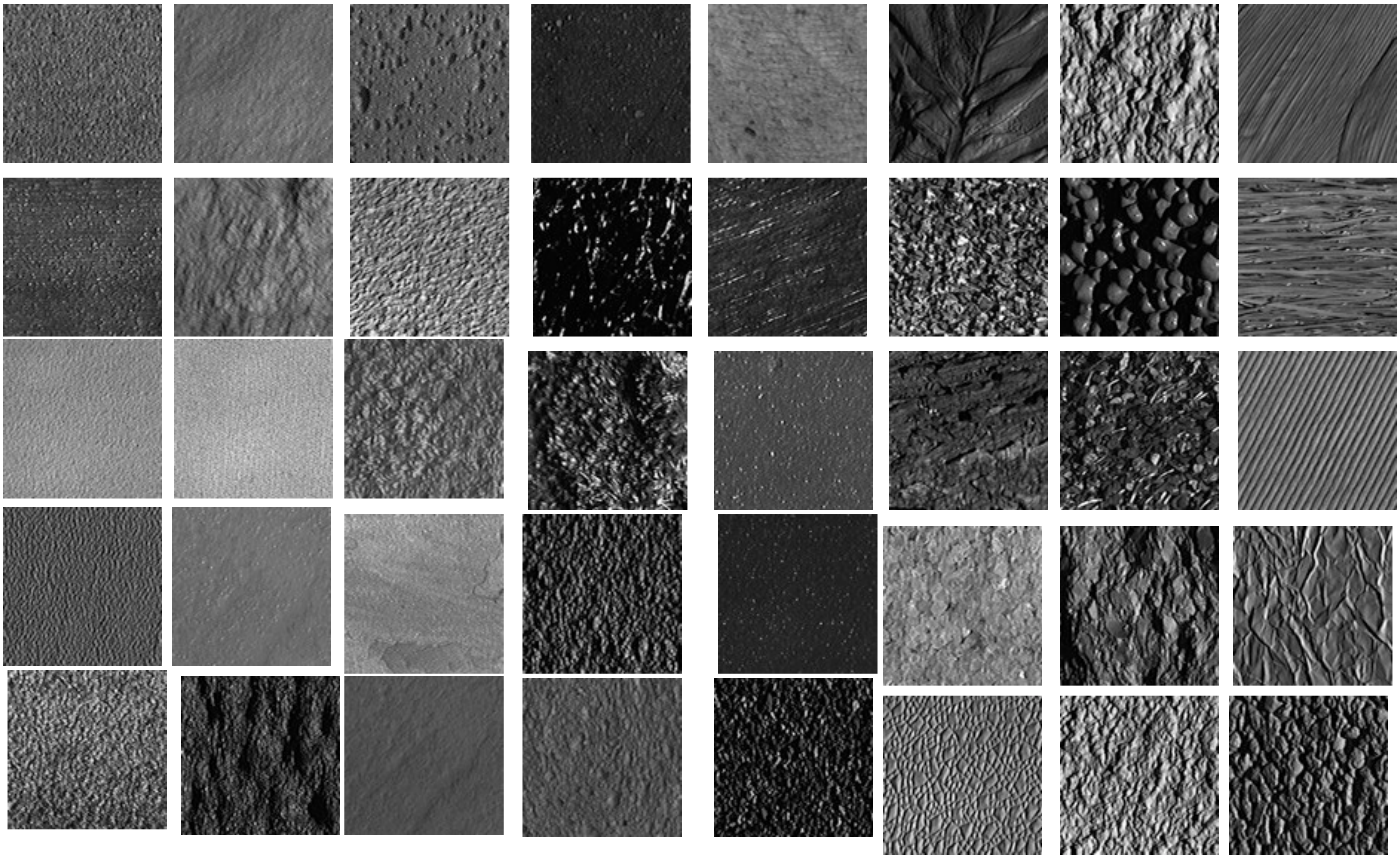
61 classes



Rotations and illumination variations.

Classification of Textures

40 classes of CureT

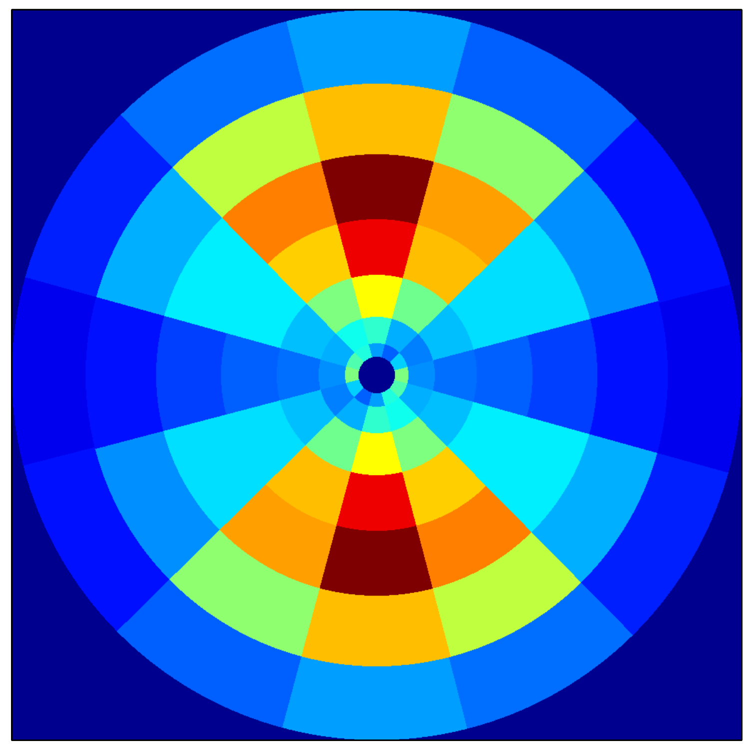


Classification of Textures

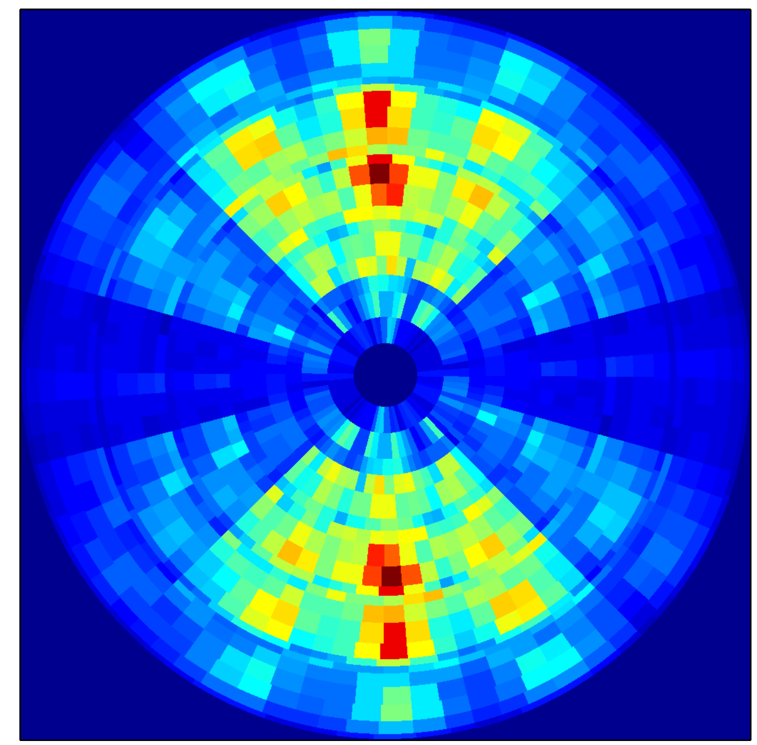
Expected Scattering
estimated with $\phi = 1$

X

$|X \star \psi_{\lambda_1}| \star \phi$



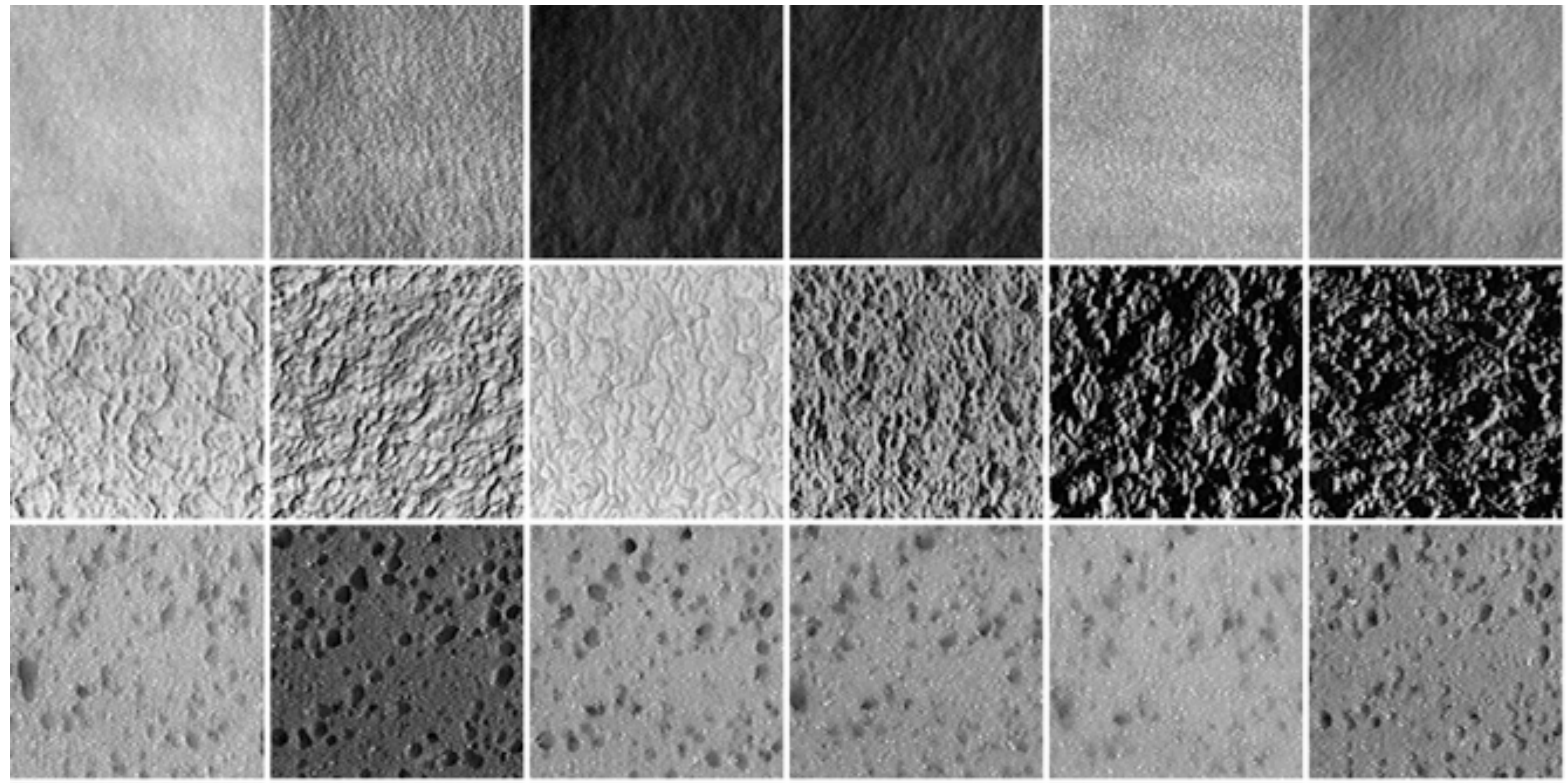
$||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi$



Classification of Textures

J. Bruna

CUREt database
61 classes



$$x \longrightarrow \boxed{Sx} \longrightarrow \boxed{\text{Supervised Linear Classifier: PCA/SVM}} \longrightarrow y$$

Training per class	Fourier Spectr.	Histogr. Features	Scattering
46	1%	1%	**0.2** %

Self-Similar Processes

- If $X(t)$ has stationary increments then $X \star \psi_{2^j}(t)$ is stationary

- If $\mathbb{E}(|X(t) - X(t-\tau)|) < \infty$ then for all $(j_1, ..., j_m)$

$$\mathbb{E}\Big(|||X \star \psi_{2^{j_1}}| \star ...| \star \psi_{2^{j_m}}|\Big) < \infty. \Rightarrow \quad \overline{S}X \text{ exists.}$$

- Self-similarity: $X(st) \equiv s^H X(t)$

  and $X(t)$ has stationary increments.

$$\Rightarrow \ \mathbb{E}(|X \star \psi_{2^j}|^q) = 2^{jHq}\,\mathbb{E}(|X \star \psi|^q) \ .$$

Examples: Fractional Brownian motions, Levy stable processes

Scattering Fractals

J.Bruna, E.Bacry, J.F.Muzy

$$X(st) \equiv s^H \, X(t)$$

- First order scattering coefficients

$$\overline{S}X(2^{j_1}) = \mathbb{E}(|X \star \psi_{j_1}|) = \mathbb{E}(|X \star \psi|) \, 2^{Hj_1}$$

Not sufficient to discriminate different selsimilar processes.
Avoid high order moments: numerical instabilities.

- Normalized second order scattering

$$\widetilde{S}X(2^{j_1}, 2^{j_2}) = \frac{E(||X \star \psi_{2^{j_1}}| \star \psi_{2^{j_2}}|)}{E(|X \star \psi_{2^{j_1}}|)}$$

Proposition If X has stationary increments and self-similar:

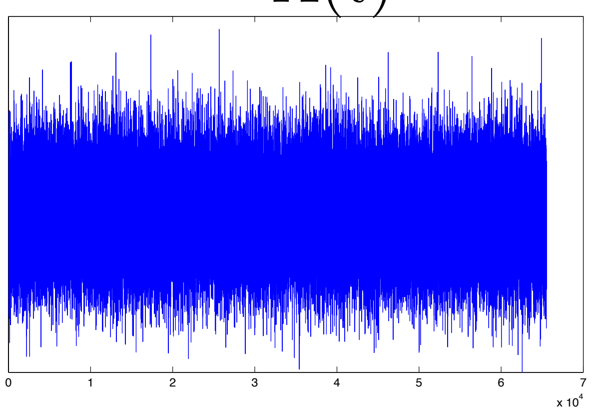
$$\widetilde{S}X(2^{j_1}, 2^{j_2}) = \widetilde{S}X(2^{j_1 - j_2}) \; .$$

Fractional Brownian Scattering

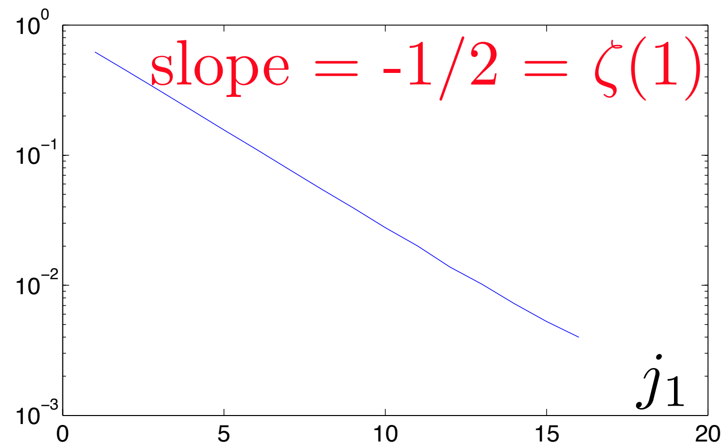
Proposition: For fractional Brownian motion and noise

$$\widetilde{S}X(2^{j_1}, 2^{j_2}) = \frac{E(||X \star \psi_{2^{j_1}}| \star \psi_{2^{j_2}}|)}{E(|X \star \psi_{2^{j_1}}|)} \sim 2^{-(j_2 - j_1)/2}$$

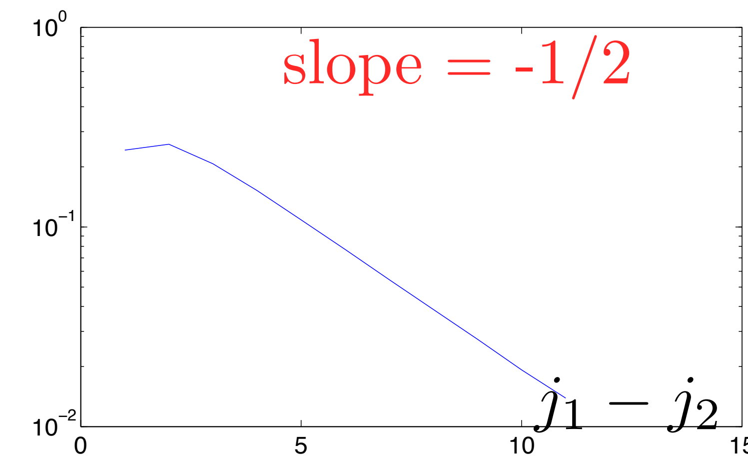
Gaussian white noise
X(t)



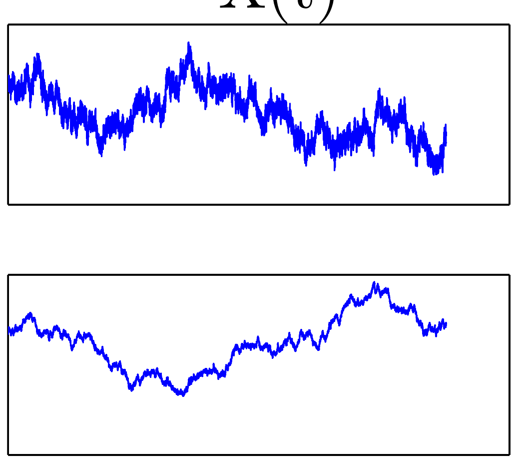
$\log \overline{S}X(2^{j_1}) \sim \zeta(1)\, j_1$



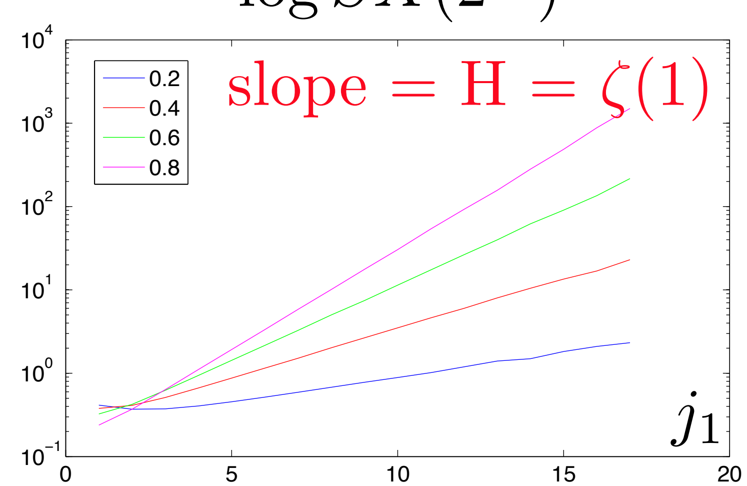
$\log \overline{S}X(2^{j_1}, 2^{j_1}) \sim -(j_2 - j_1)/2$



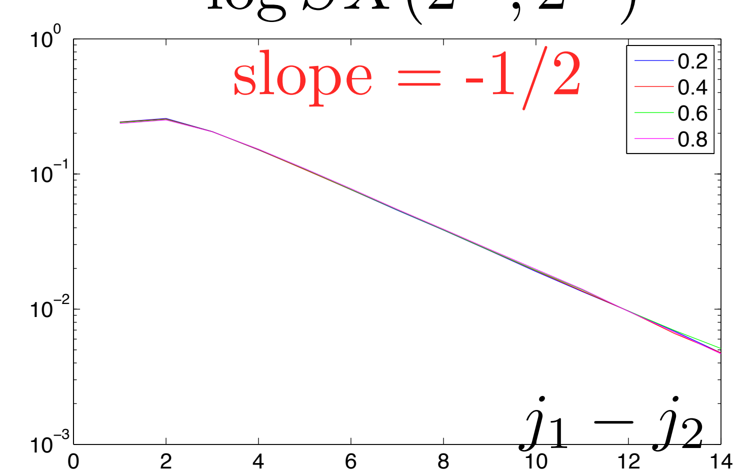
Fractional Brownian
X(t)



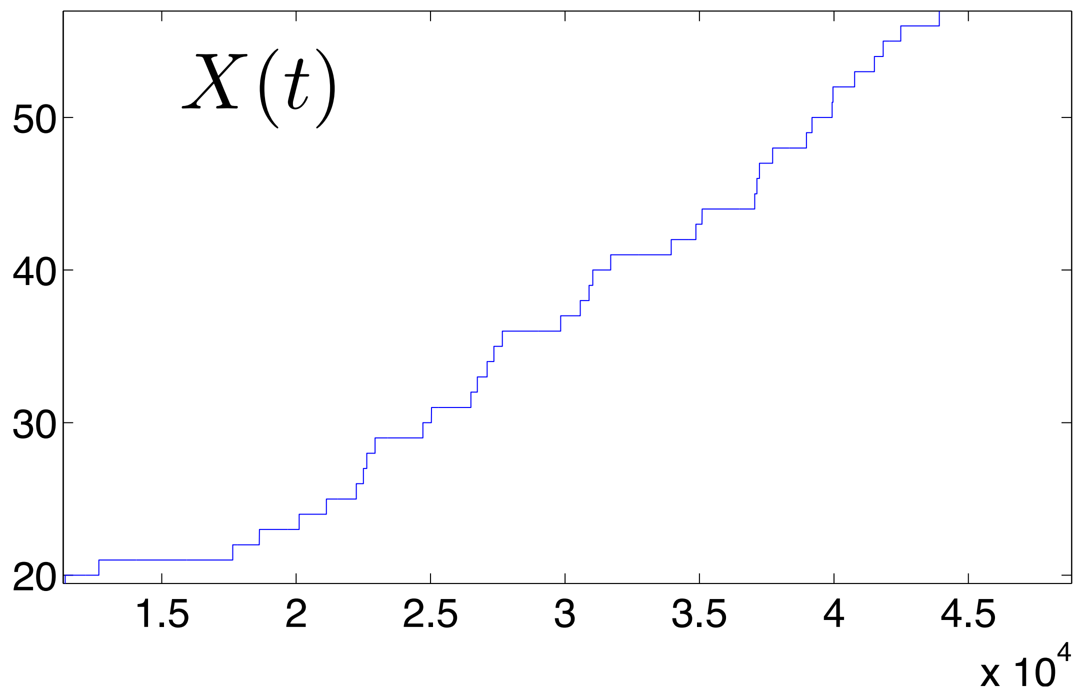
$\log \overline{S}X(2^{j_1})$

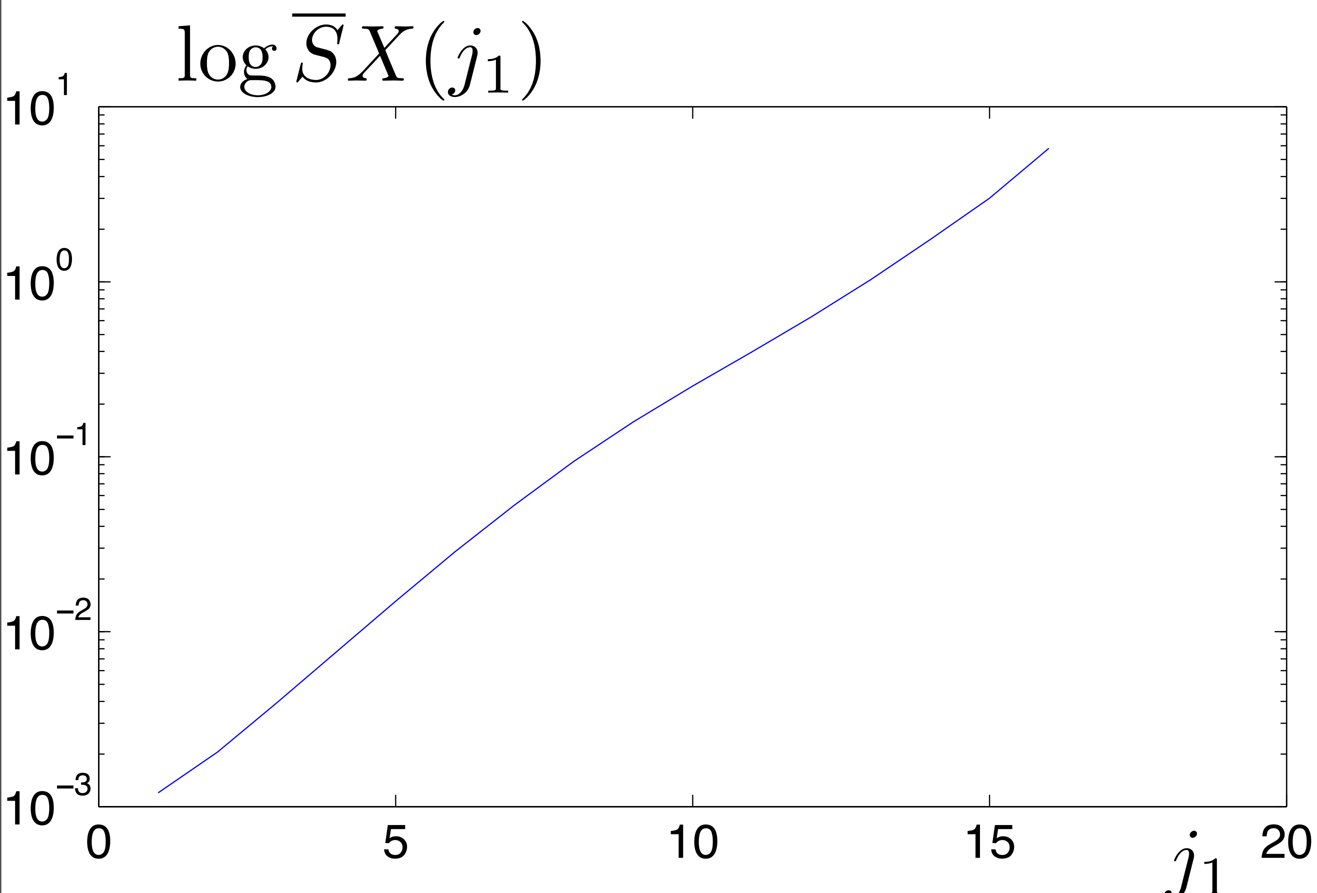


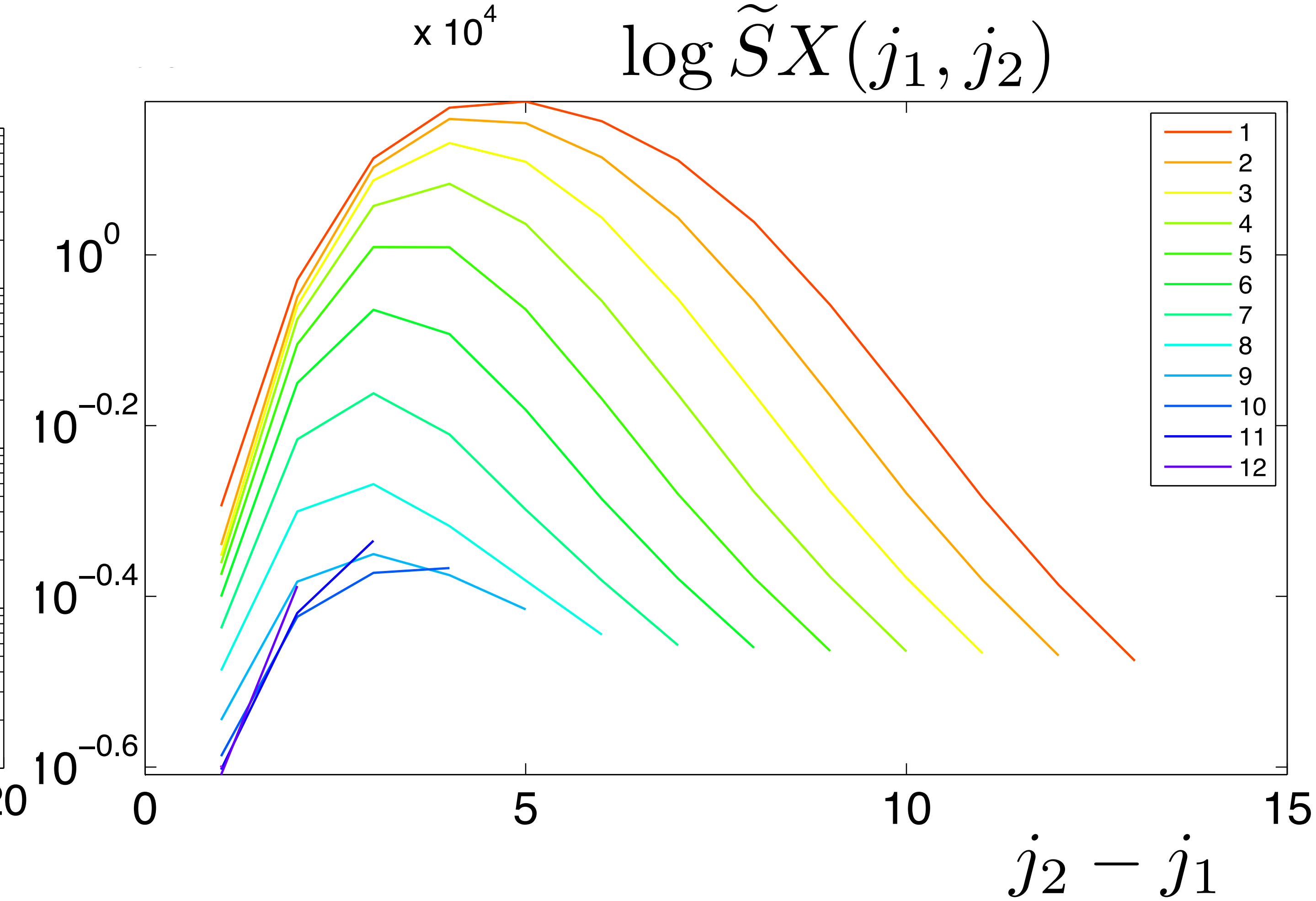
$\log \widetilde{S}X(2^{j_1}, 2^{j_2})$



Poisson Process



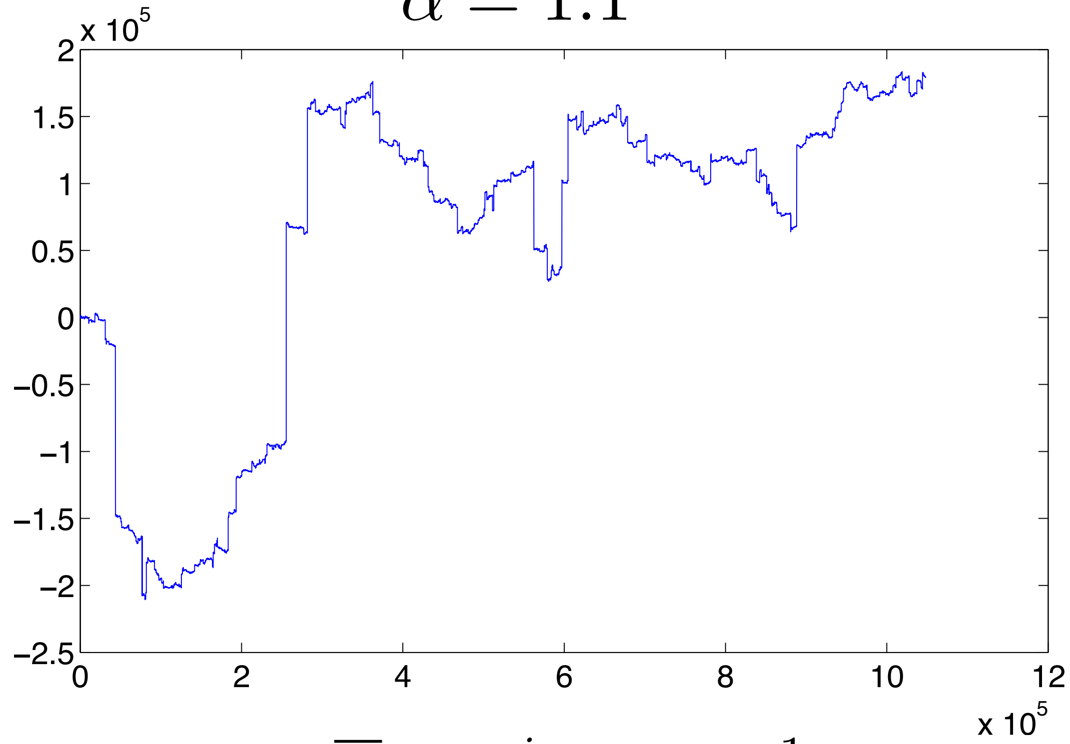




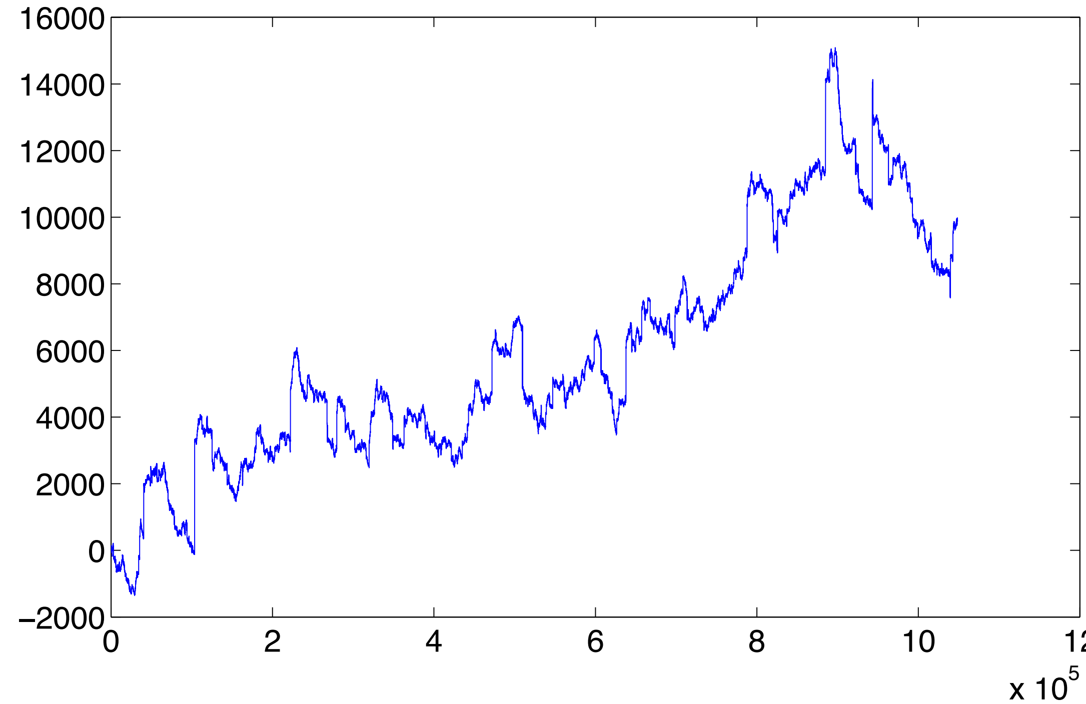
Scattering Stable Levy Measures

$$X(t)$$

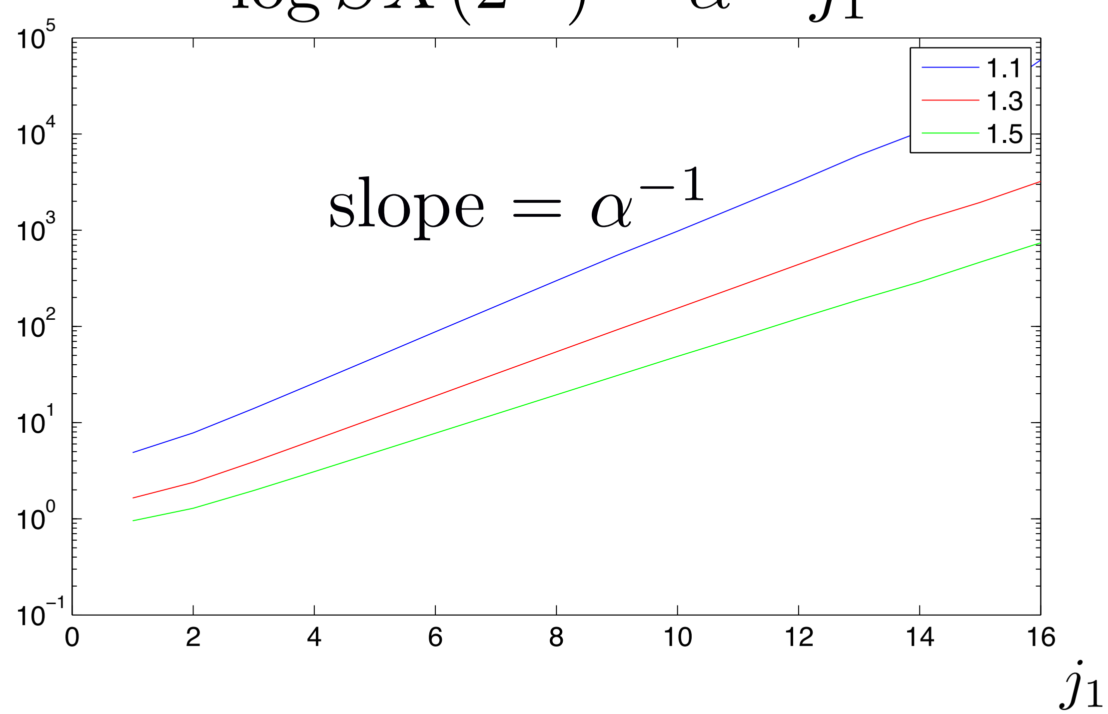
$\alpha = 1.1$



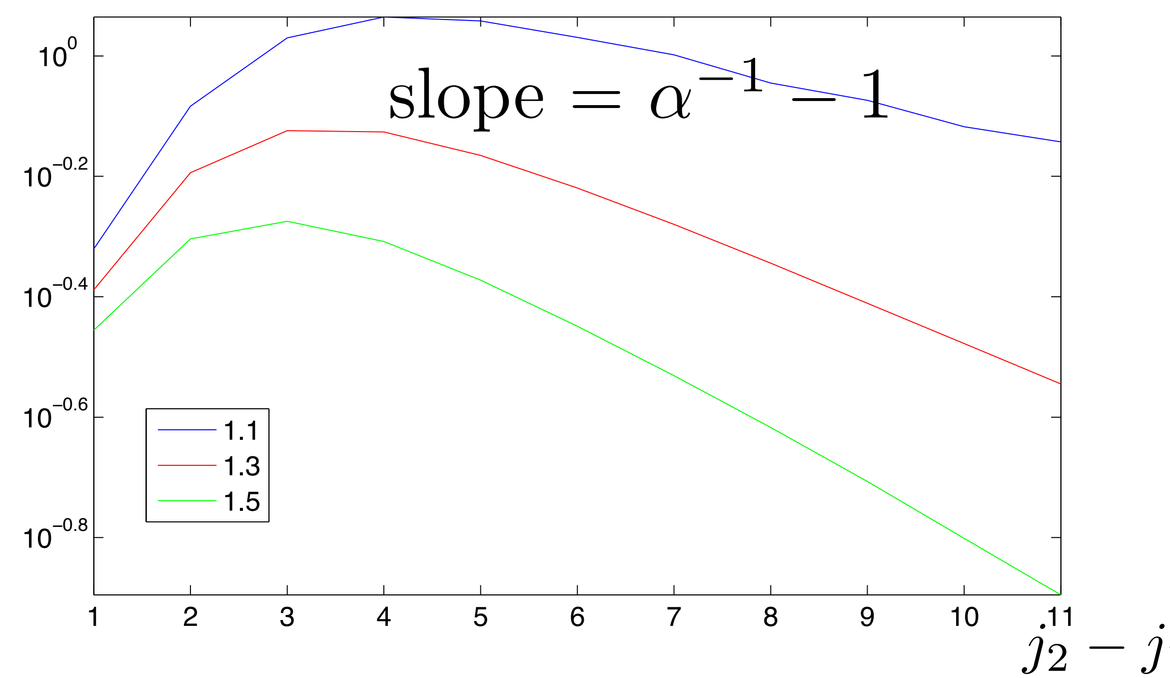
$\alpha = 1.5$



$\log \overline{S}X(2^{j_1}) \sim \alpha^{-1} j_1$



$\log \widetilde{S}X(2^{j_1}, 2^{j_2}) \sim (\alpha^{-1} - 1)(j_2 - j_1)$



$\alpha = 2$: Brownian motion.

Scattering Multifractals

- Stochastic self-similarity: $X(st) \equiv A_s\, X(t)$

  where A_s is a random variable independant of X and

$$E(|A_s|^q) \sim s^{\zeta(q)}$$

  and $X(t)$ has stationary increments.

- A_s is constant for fractional Brownians and Levy Stable:

$$\Rightarrow \quad \zeta(q) = \zeta(1)\, q\ .$$

- A_s is a log-normal random variable for Mandelbrot cascades.

Proposition If X has stationary increments and self-similar:

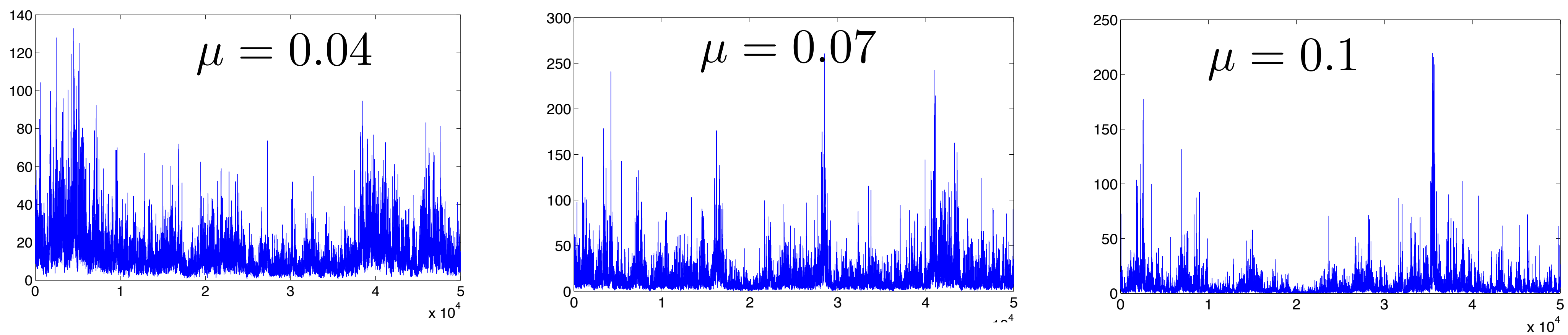
$$\widetilde{S}X(2^{j_1}, 2^{j_2}) = \widetilde{S}X(2^{j_1 - j_2})\ .$$

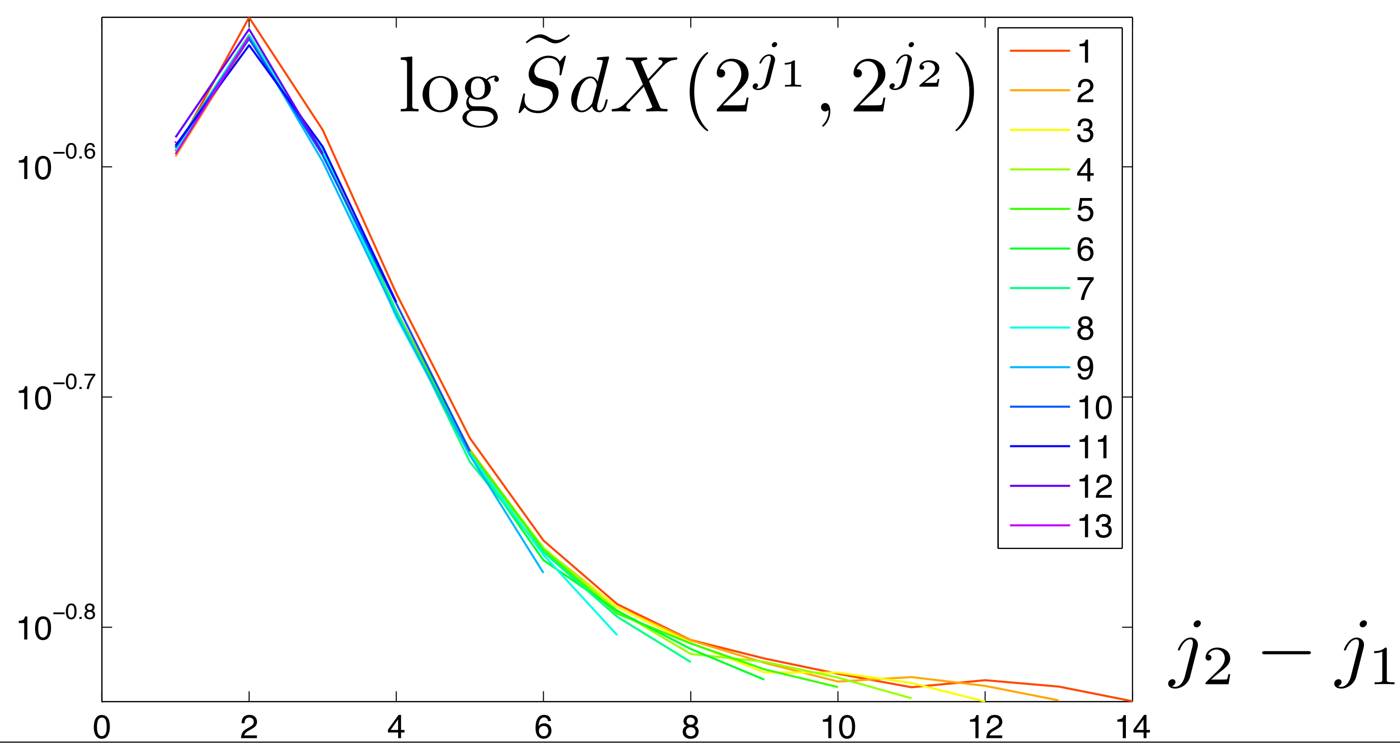
Mandelbrot Cascades

Barral, Mandelbrot

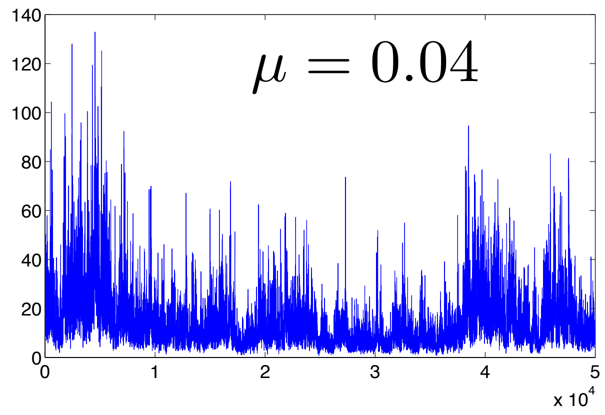
- Stationary log normal random measure $dX(t)$ obtained as multiscale products of log-normal random variables.

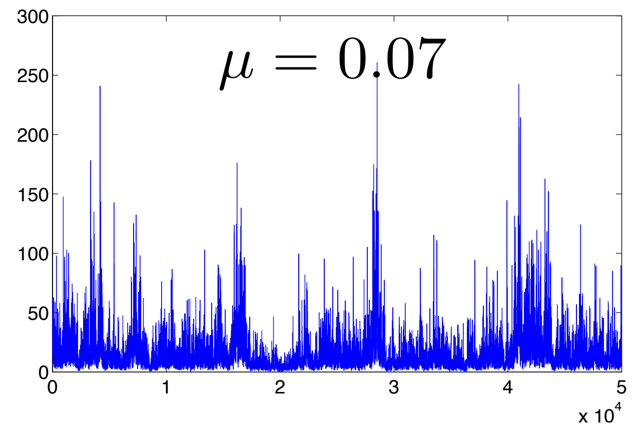
$$\zeta(q) = \left(1 + \frac{\mu}{2}\right) q - \frac{\mu}{2}\, q^2$$

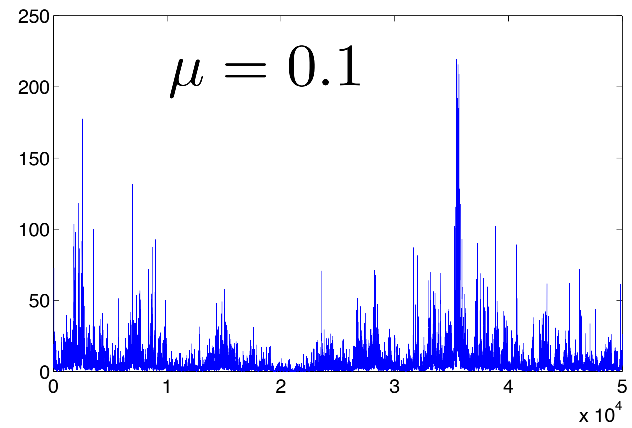




Scattering Mandelbrot Cascades





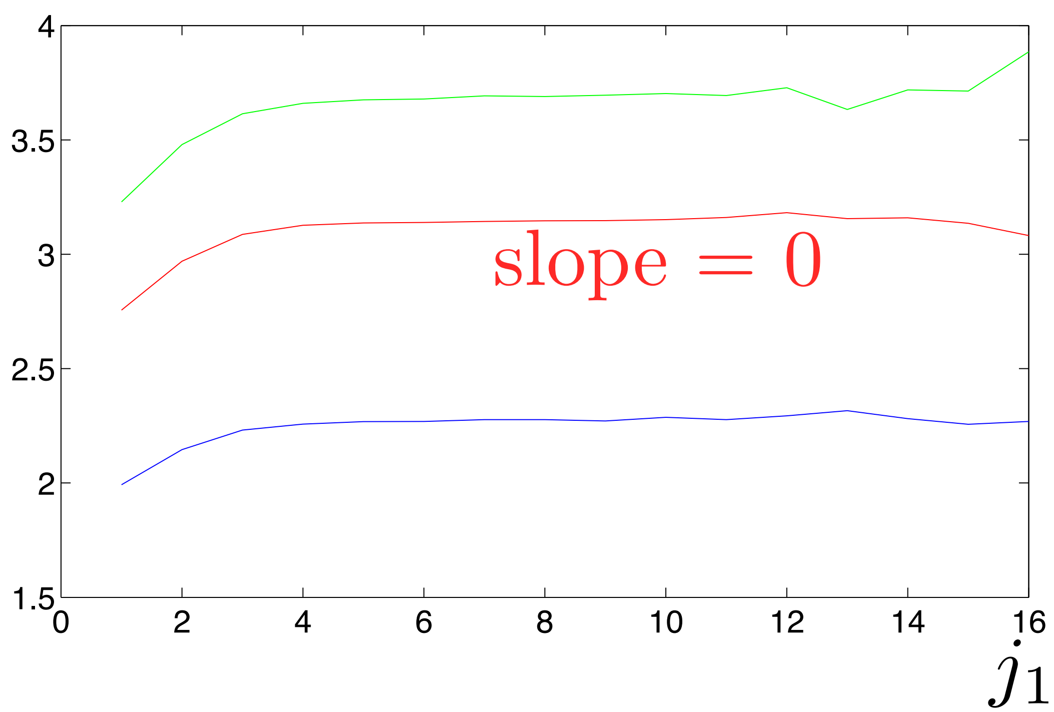


J. Bruna, E. Bacry, J.F. Muzy

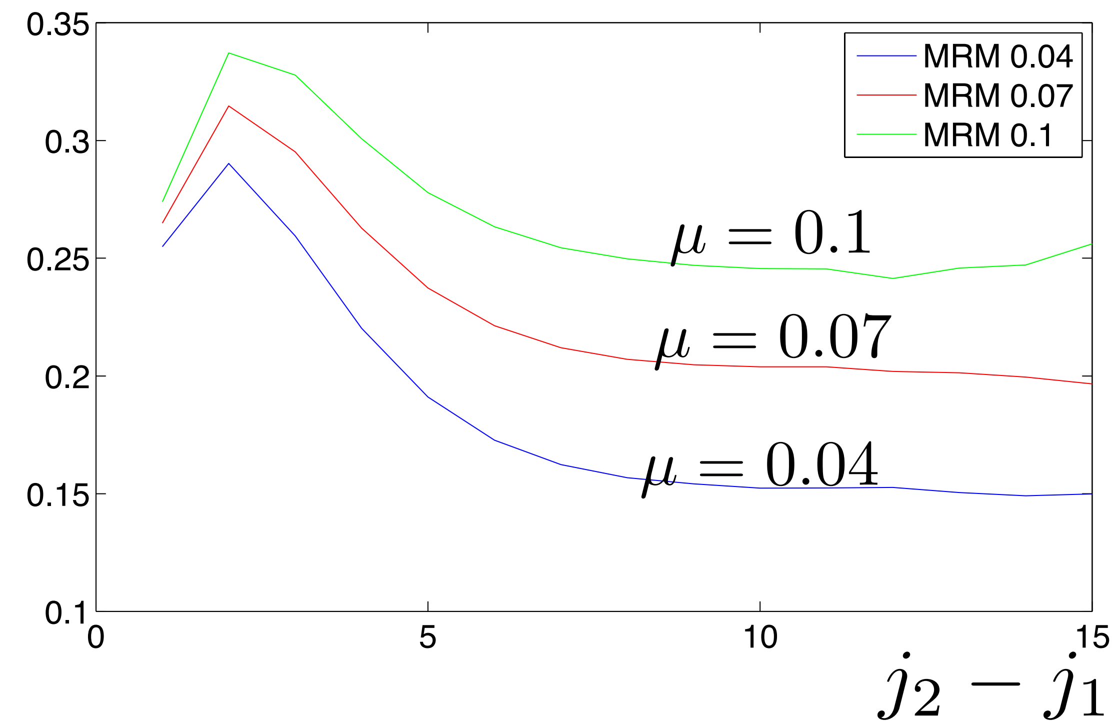
Theorem: Mandelbort Random Measures dX satisfy:

$$\lim_{j_2 - j_1 \to \infty} \widetilde{S} dX(2j_1, 2^{j_2}) = C_2\, \mu \ .$$

$$\log \overline{S} dX(2^{j_1}) \sim C_1$$

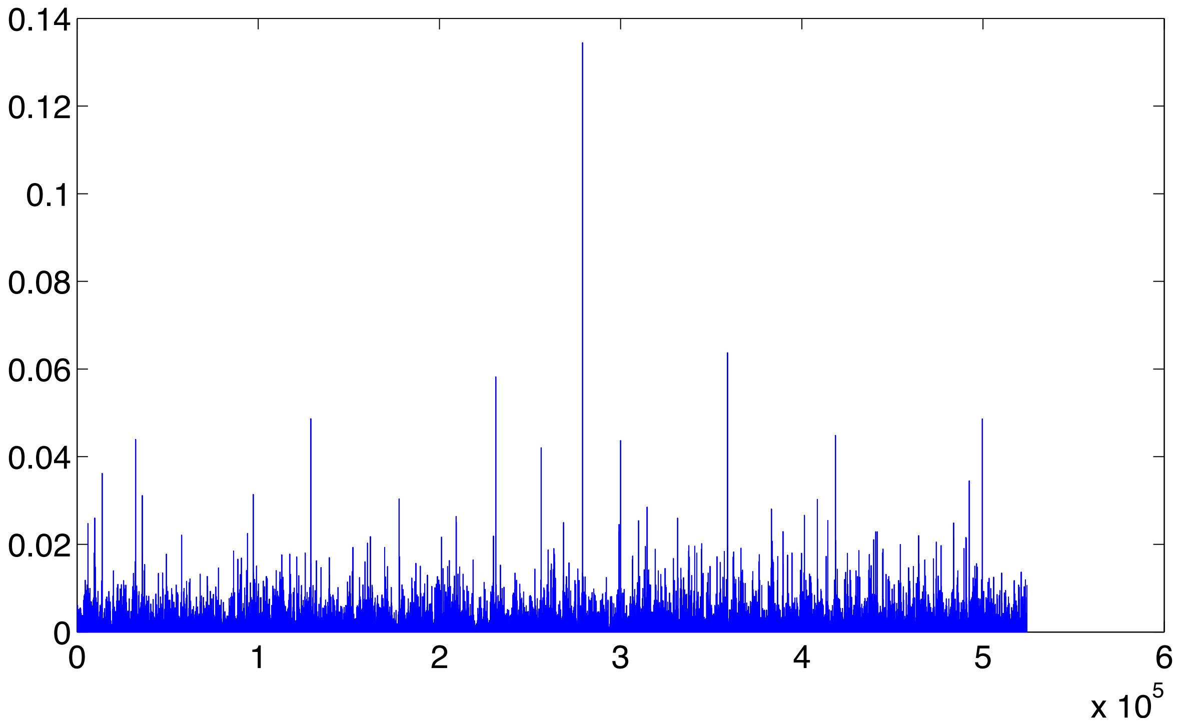


$$\log \widetilde{S} dX(2^{j_1}, 2^{j_2})$$

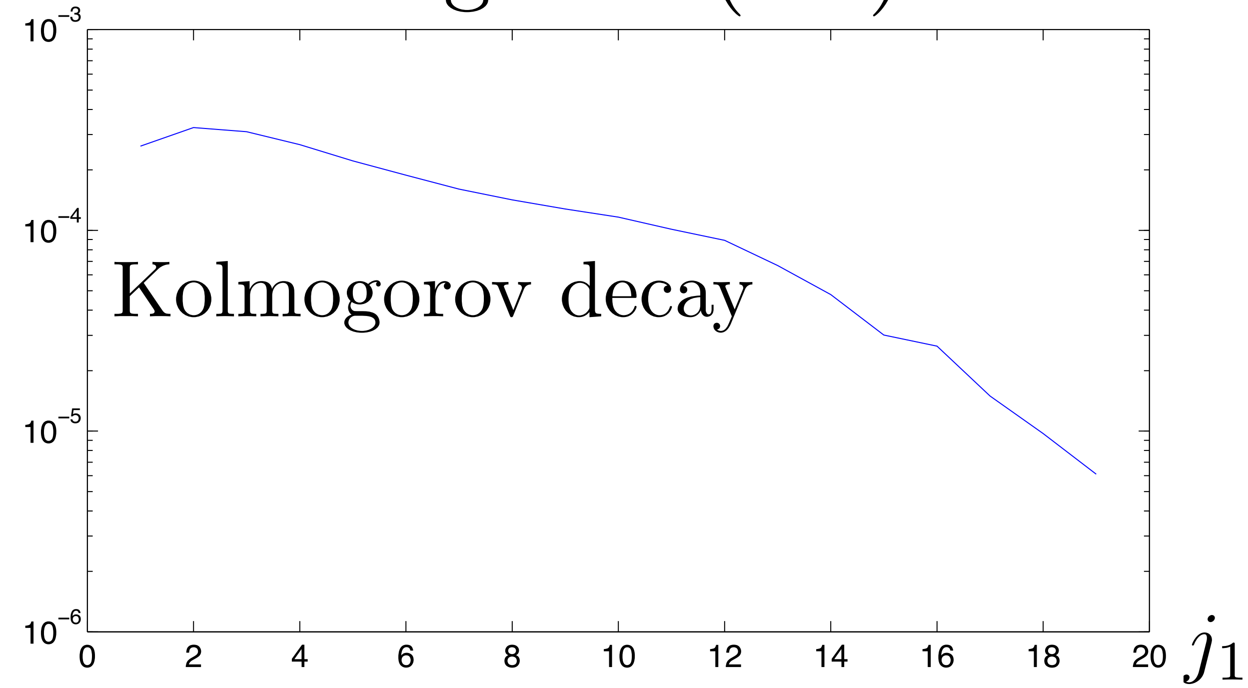


Scattering Turbulence

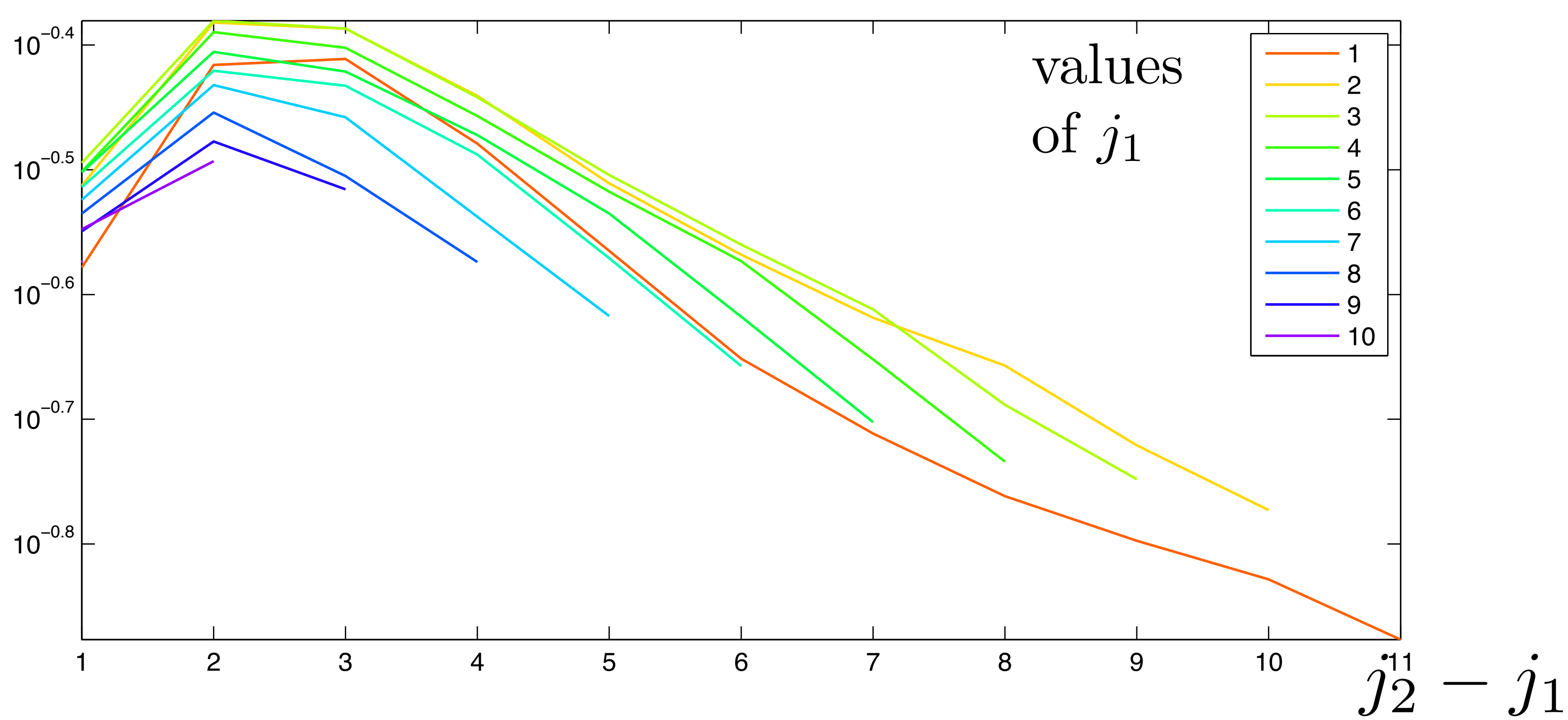
$dX(t)$: dissipation energy



$\log \overline{S} dX(2^{j_1})$



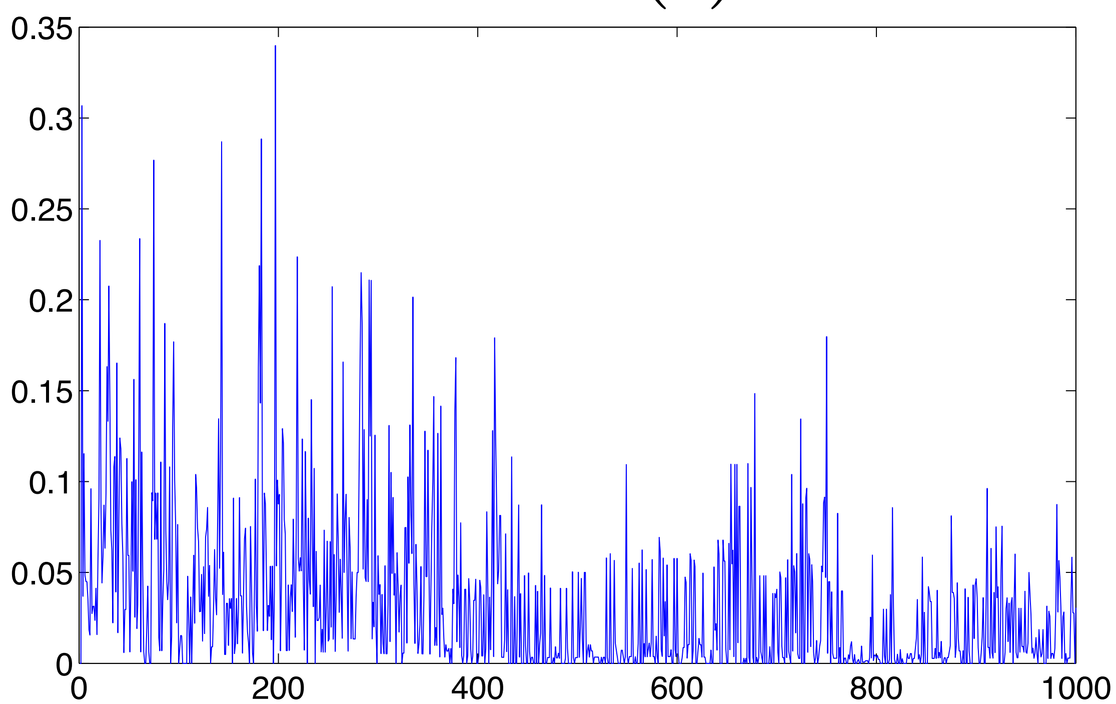
$\log \widetilde{S}(2^{j_1}, 2^{j_2})$



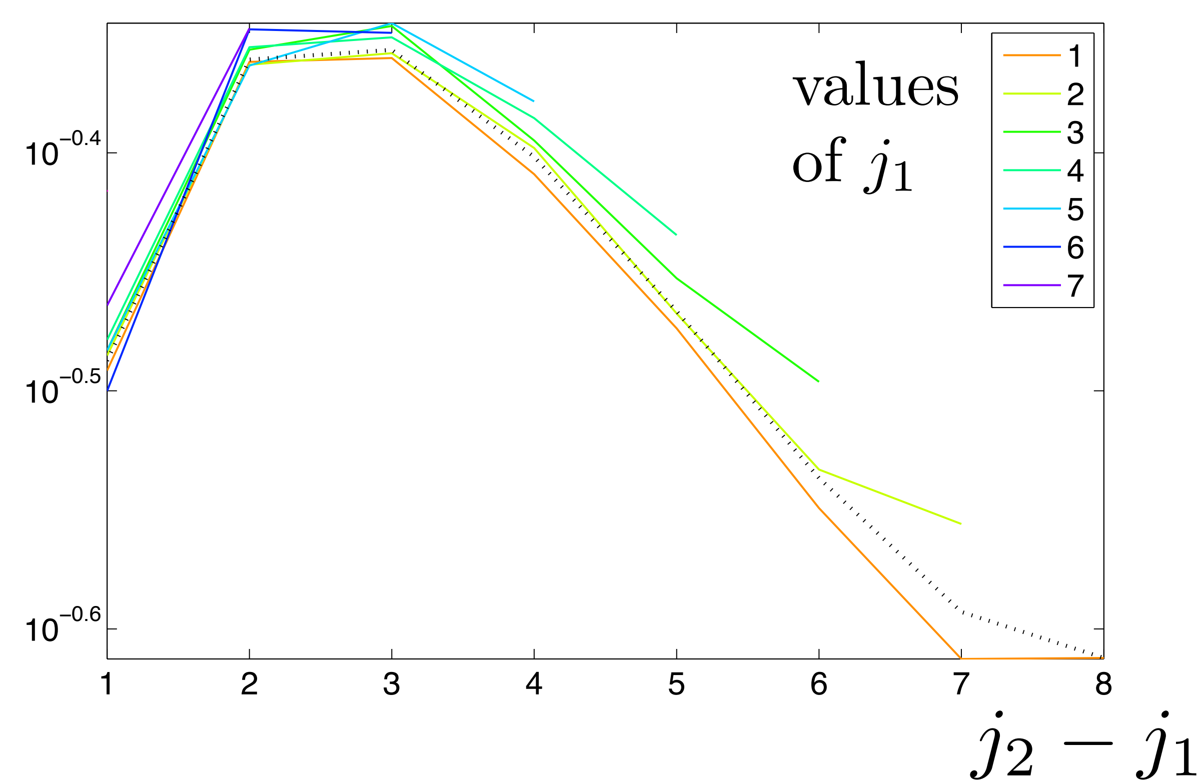
Financial Time Series

1 Trading day of German Bund.

$dX(t)$

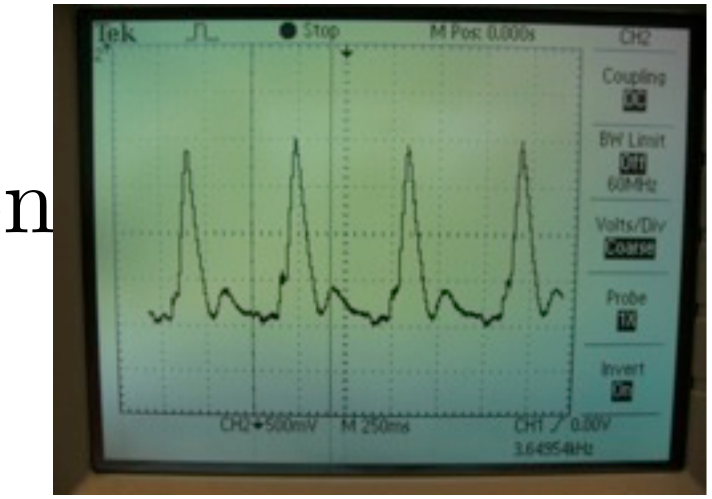


$\log \widetilde{S}(2^{j_1}, 2^{j_2}) \approx F(j_2 - j_1)$



Fetal Heart Rate Variability

P. Abry, J. Anden, V. Chudacek, M. Doret, R. Talmon



- Fetal heart rate monitoring gives information on the stress level of babies before delivery.

- Recording over 30 minutes before delivery $x \in \mathbb{R}^{10^4}$

- Locally stationary over 2 minutes: 10^3 points.

- Build a scattering representation Sx

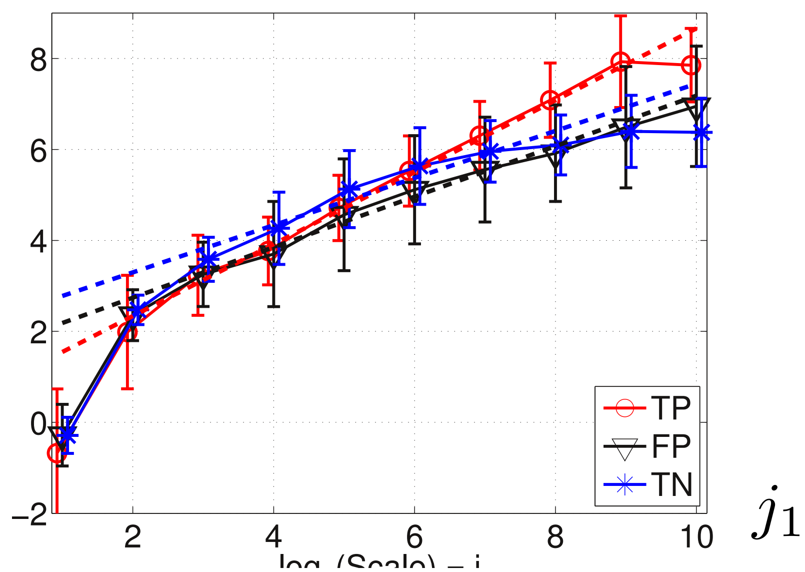
Fetal Heart Rate Variability

P. Abry, J. Anden, V. Chudacek, M. Doret

Fetal heart rate monitoring gives information on the stress level of babies before delivery.

"Fractal behavior"

$$\log \overline{S}X(2^{j_1})$$

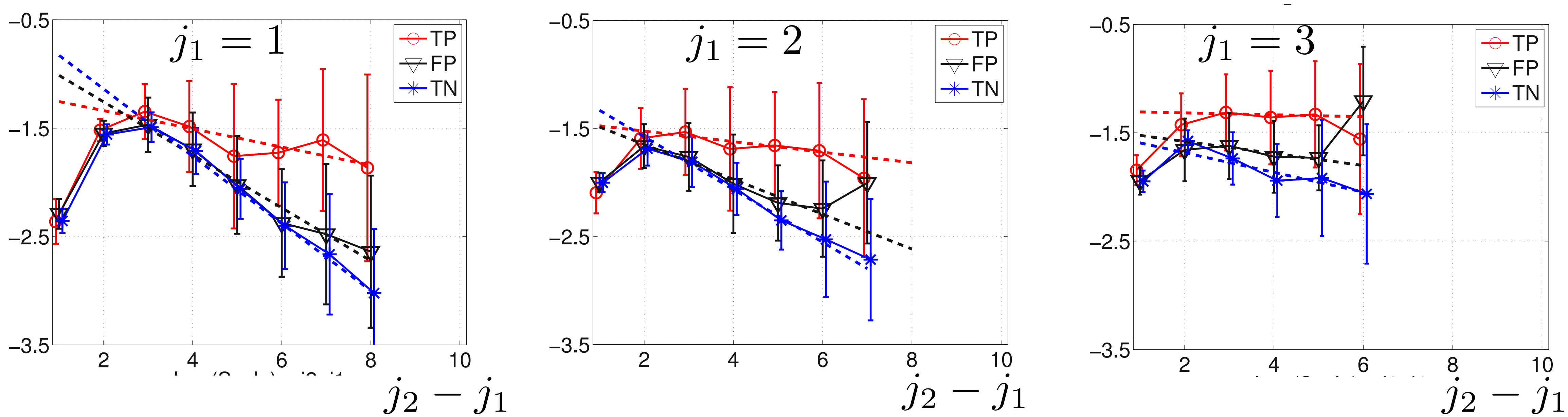


True Unhealthy

False Healthy

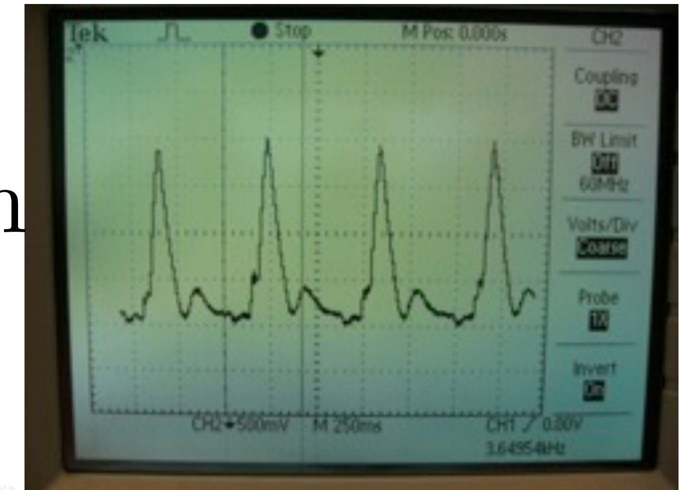
True Healty

$$\log \tilde{S}X(2^{j_1}, 2^{j_2}) \neq F(j_2 - j_1) \;\Rightarrow\; \text{Not self-similar}$$



Fetal Heart Rate Variability

P. Abry, J. Anden, V. Chudacek, M. Doret, R. Talmon

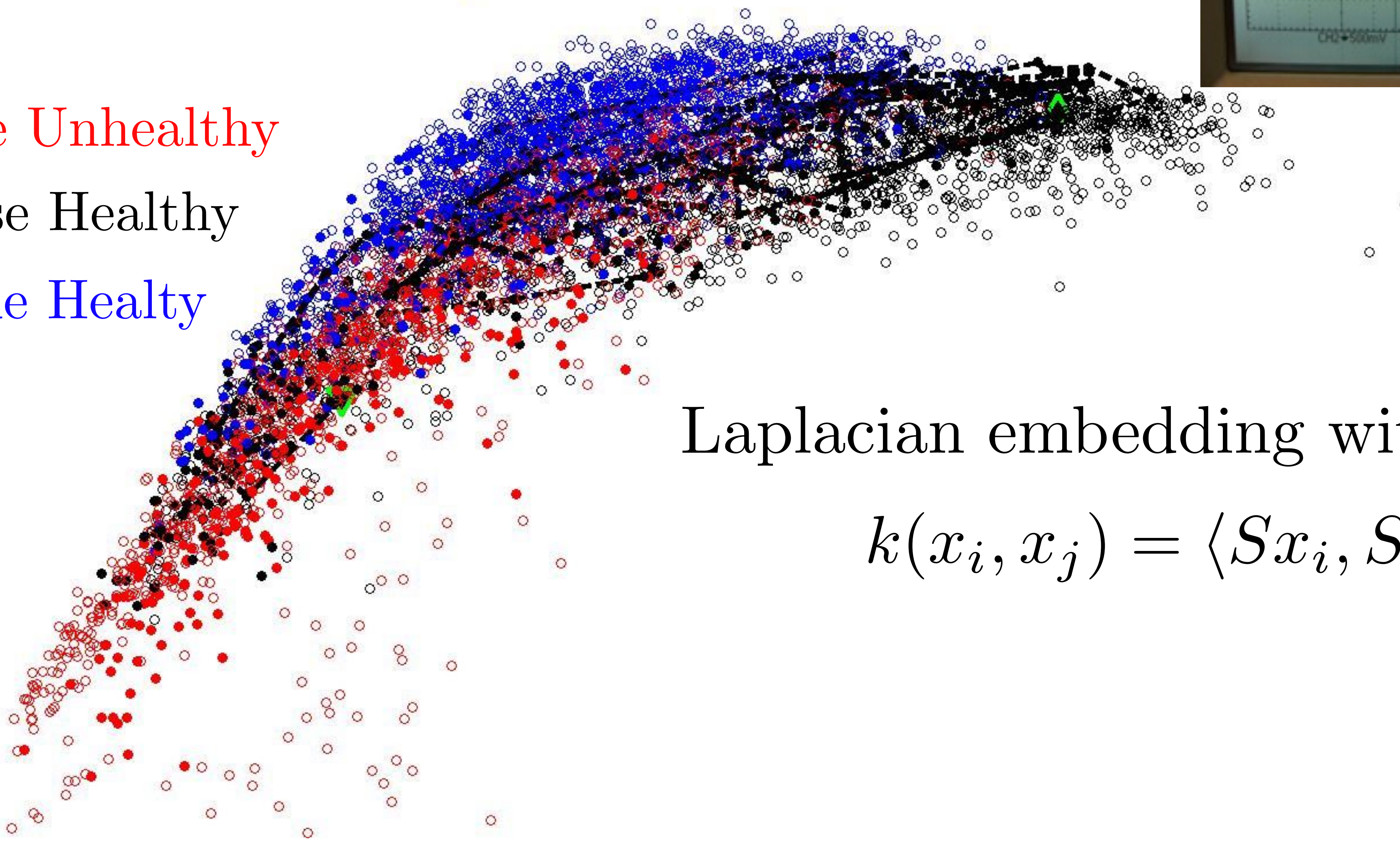


- Fetal heart rate monitoring gives information on the stress level of babies before delivery.

True Unhealthy

False Healthy

True Healty



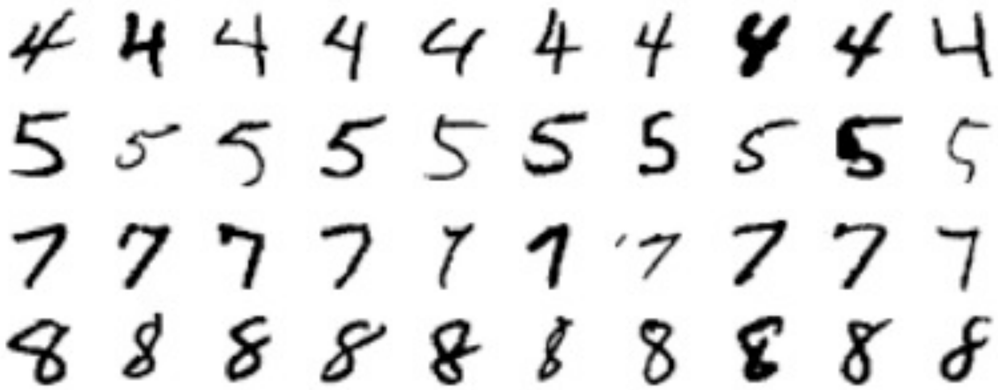
Laplacian embedding with

$$k(x_i, x_j) = \langle Sx_i, Sx_j \rangle$$

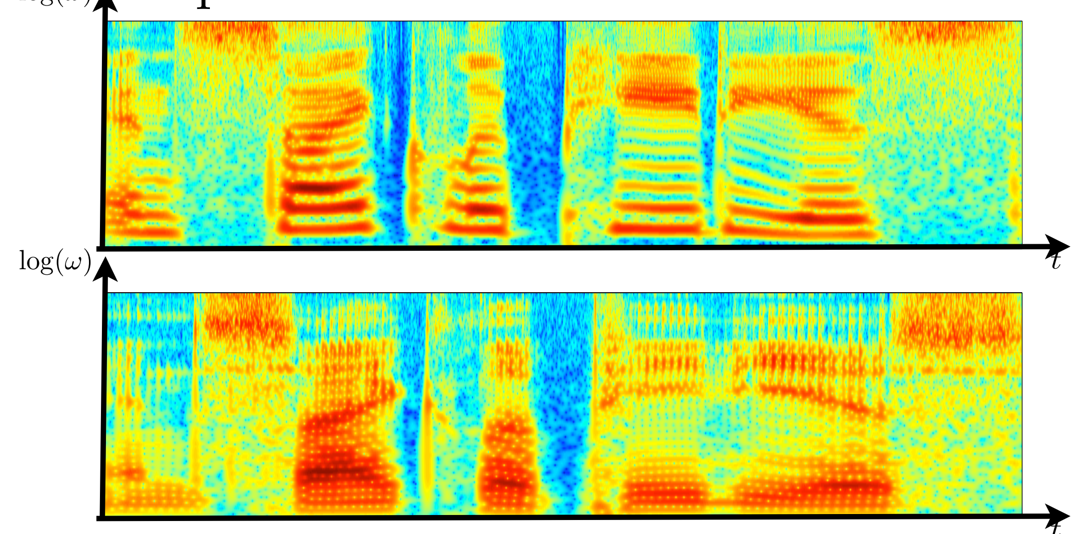
Part III: Adapted Invariants

- How to represent high-dimensional data $x \in \mathbb{R}^d$ for classification ?
  $d \geq 10^6$
- Need to compute **discriminative invariants.**

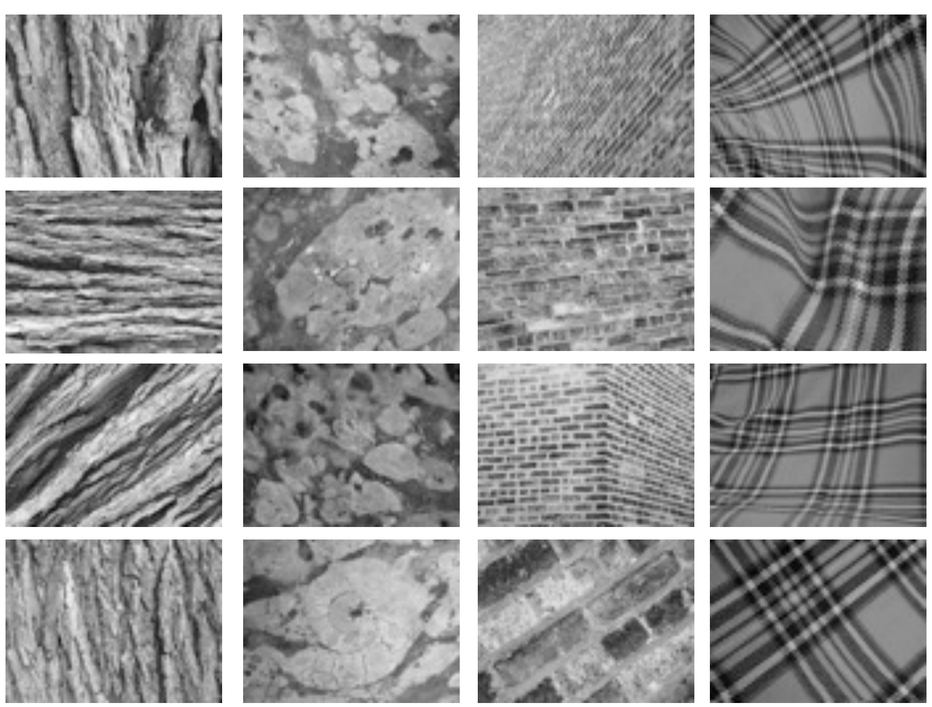
MNIST digit classification



Speech and Music classification

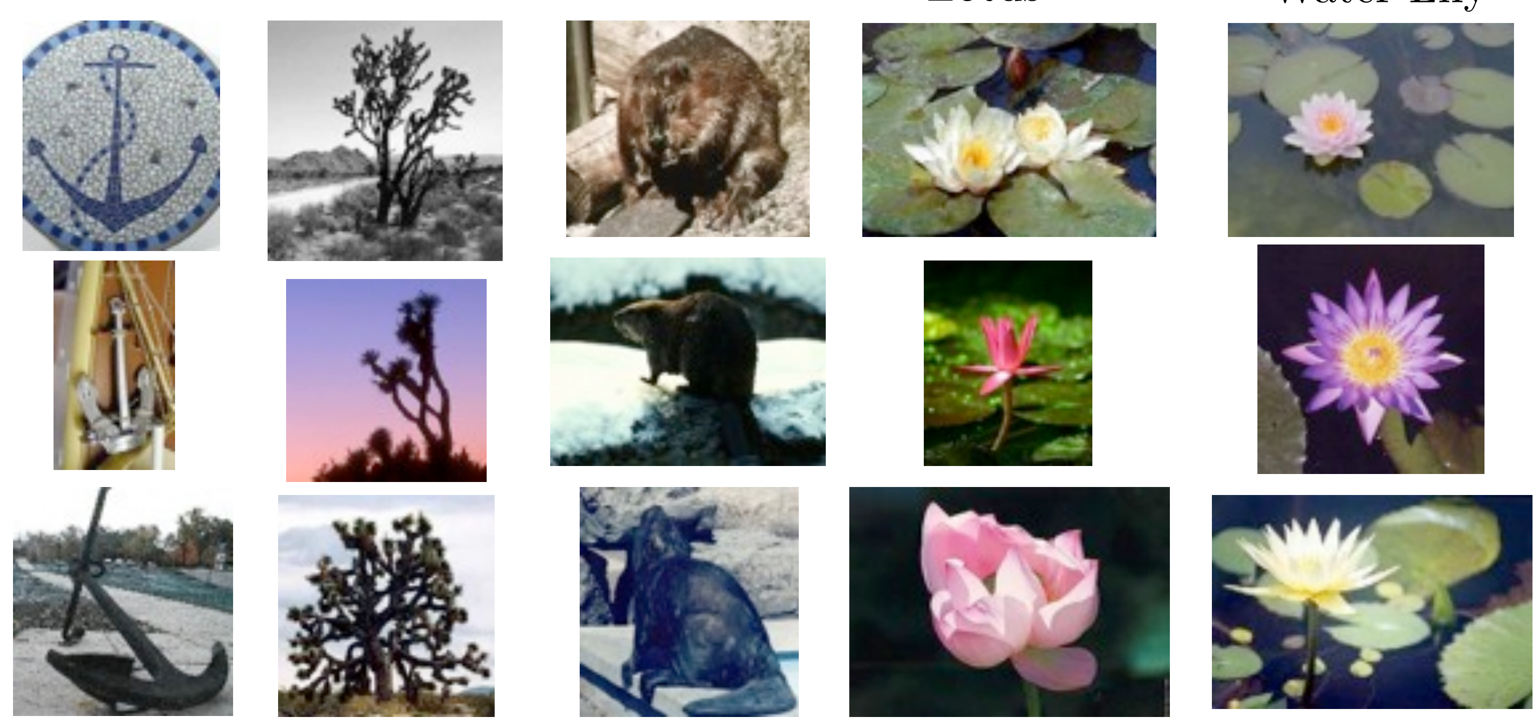


Texture classification



CalTech 101

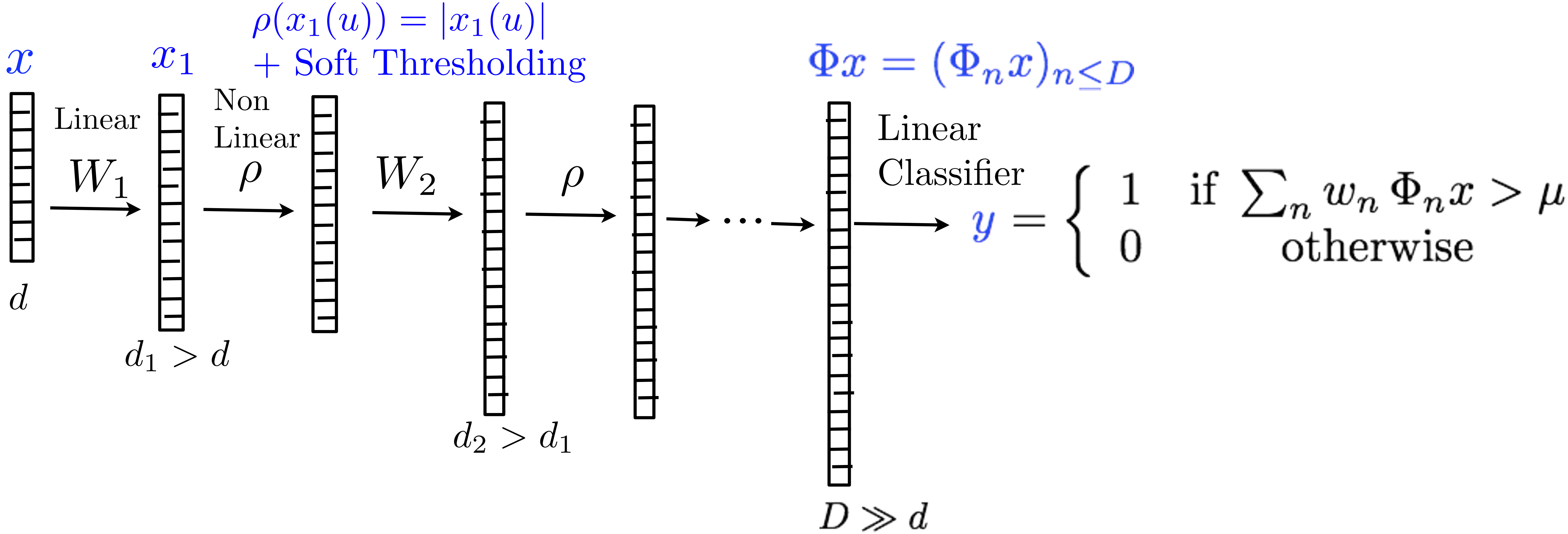
Anchor | Joshua Tree | Beaver | Lotus | Water Lily



Deep Neural Network Classifiers

J. Hinton, Y. LeCun "State of the art results"

Hierarchical invariance



- Deep network algorithms learn the W_k with sparsity.

Why does it work ?

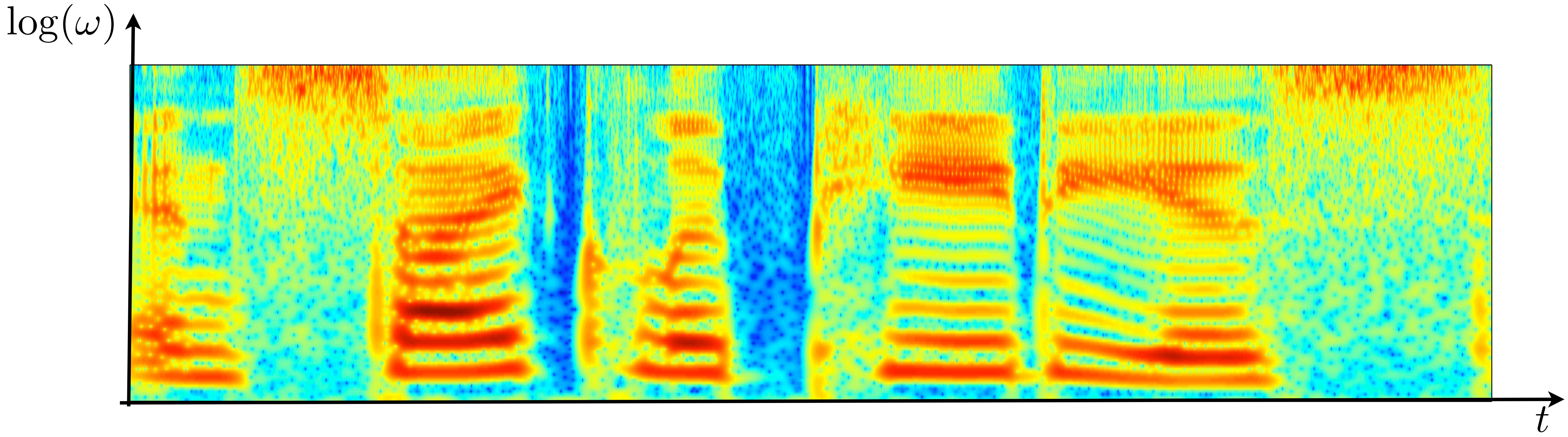
Overview

- Invariance to a Lie group action and stability to diffeomorphisms
  - Translation and frequency transpositions
  - Translations and rotations
  - Tranvariance to translation-rotations and scaling

- Unsupervised learning of unknown variability sources

Frequency Transpositions

Time and frequency translations and deformations:



- Frequency transposition invariance is needed

  for speech recognition not for locutor recognition.

Transposition Invariance

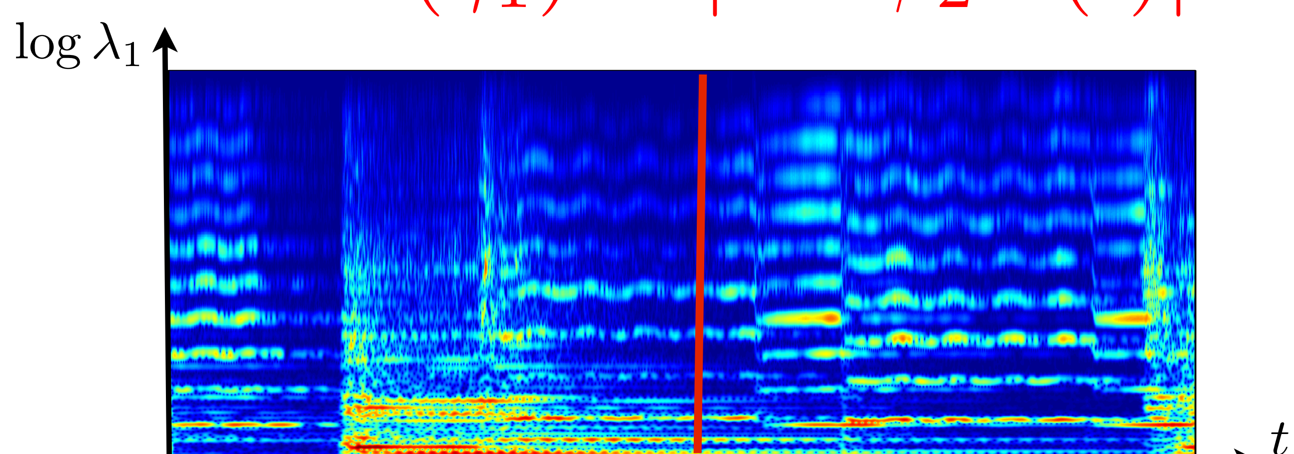
J.Anden

- Frequency transposition is a common source of variability
- Transposition $\Leftrightarrow$ translation and deformations in $\log \lambda_1$
- Invariance with a "frequency scattering" along $\log \lambda_1$

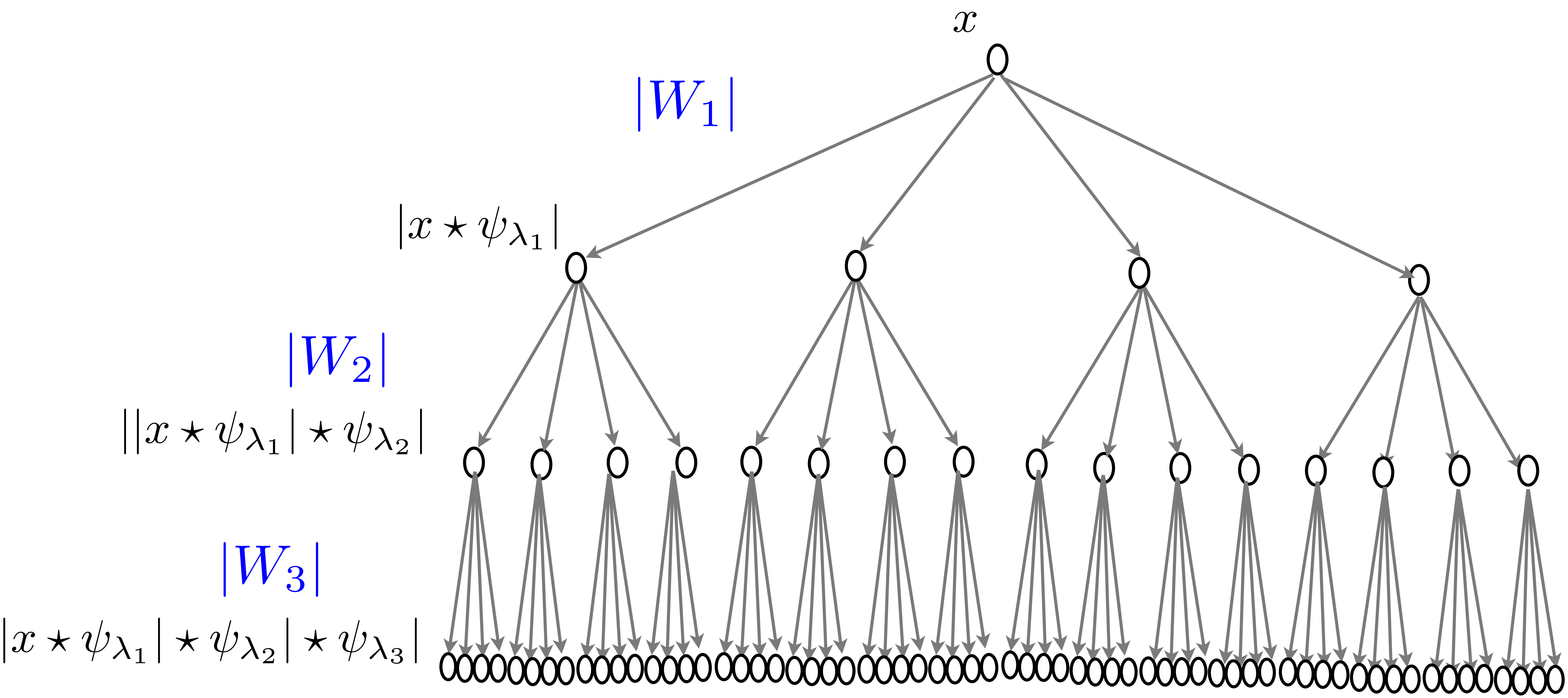
$$x(t) \longrightarrow \boxed{\text{Scattering along } t} \longrightarrow \boxed{\text{Scattering along } \log \lambda_1} \longrightarrow \Phi x$$

Scattering along log frequency $\gamma_1 = \log_2 \lambda_1$:

$$z(\gamma_1) = |x \star \psi_{2^{\gamma_1}}(t)|$$



Non-Averaged Scattering



Genre Classification (GTZAN)

J.Anden

- GTZAN: music genre classification (jazz, rock, classical, ...) 10 classes and 30 seconds tracks.
- Each frame is classified using a Gaussian kernel SVM.

$$T = 370 \text{ ms}$$

Feature Set	Error (%)
Δ-MFCC (32 ms)	19.3
Time Scat., m = 1	17.9
Time Scat., m = 2	12.3
Time & Frequency Scat., m=2	10.3

Phone Classification (TIMIT)

J.Anden

- Training on 3696 phrases (139868 phones) and and testing on 192 phrases (7201 phones)
- Each phone is classified using a Gaussian kernel SVM.

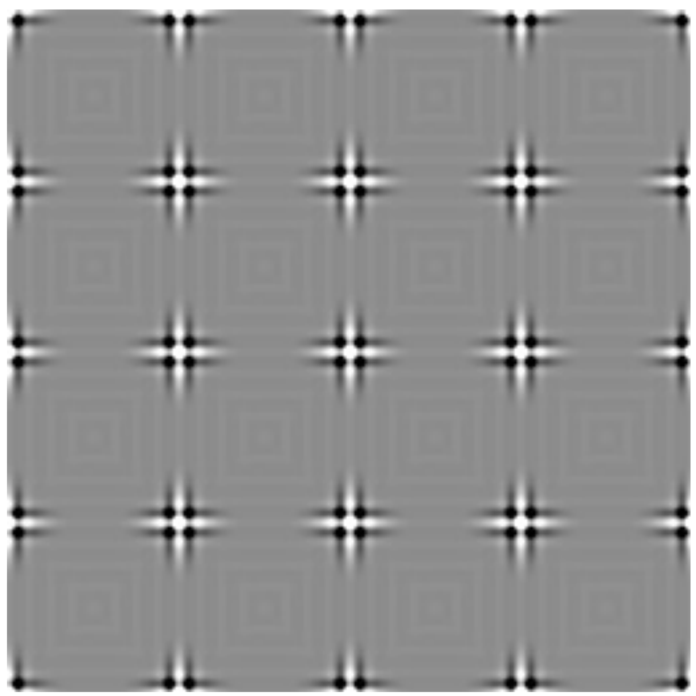
$T = 32$ ms

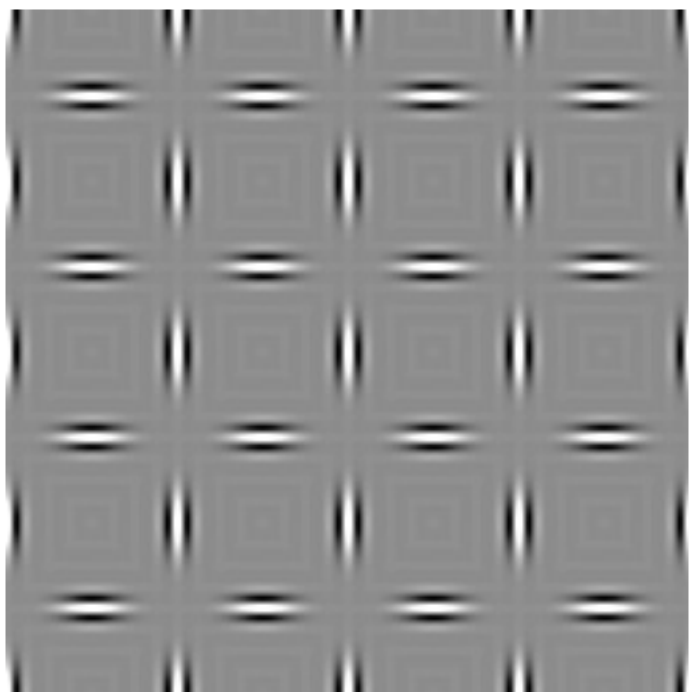
Feature Set	Error (%)
Δ-MFCC (32 ms)	19.3
State of the art (excl. scattering)	16.7
Time Scat., m = 1	18.5
Time Scat., m = 2	17.7
Time & Freq. Scat., m = 2	16.5

Joint versus Separable Invariants

• Separable cascade of invariants loose joint distributions.

• Separable rotation and translation invariants can not discriminate:





⇒ need to build invariant on the joint roto-translation group.

Roto-Translation Group

- Roto-translation group $G = \{g = (r,t) \in SO(2) \times \mathbb{R}^2\}$

$$(r,t)\,.\,x(u) = x(r^{-1}(u-t))$$

- Group multiplication:

$$(r',t')\,.\,(r,t) = (r'r\,,\,r't+t') \ \ : \ \text{not commutative.}$$

- Inverse: $(r,t)^{-1} = (r^{-1}, -r^{-1}t)$.

- An averaging invariant is convolution on $\mathbf{L^2}(G)$: $x(g) = x(r,t)$

for translations : $x \star \phi(t) = \int_{\mathbb{R}^2} x(t')\, \phi(t-t')\, dt'$

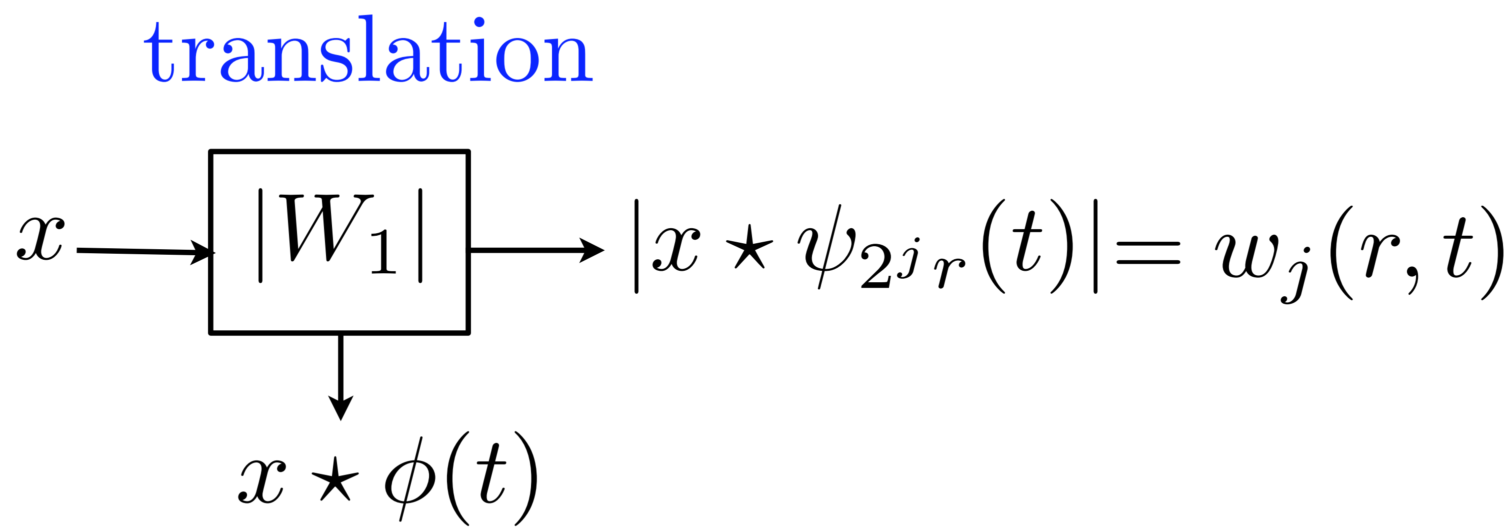
for roto-translations : $x \circledast \overline{\phi}(g) = \int_G x(g')\, \overline{\phi}(g'^{-1} g)\, dg'$

- Roto-translation Haar measure : $dg = dt\, d\theta$ (rotation angle θ)

Scattering on a Lie Group

L. Sifre

- First layer: wavelet transform along the translation group.



Scattering on a Lie Group

L. Sifre

- How to define a wavelet transform of $x(r,t) \in \mathbf{L^2}(G)$?
- One can define separable complex wavelets $\overline{\psi}_{\lambda_2}(r,t) \in \mathbf{L^2}(G)$

$$W_2 x = \left(\begin{array}{c} x \circledast \overline{\phi}(r,t) \\ x \circledast \overline{\psi}_{\lambda_2}(r,t) \end{array} \right)_{\lambda_2, r, t} \quad \text{is tight frame of } \mathbf{L^2}(G).$$

$$x \circledast \overline{\psi}_\lambda(g) = \int_G x(g')\, \overline{\psi}_\lambda(g'^{-1} g)\, dg'$$

$$\|x\|^2 = \int_G |x(g)|^2 dg = \|x \circledast \phi\|^2 + \sum_{\lambda_2} \|x \circledast \psi_{\lambda_2}\|^2$$

Scattering on a Lie Group

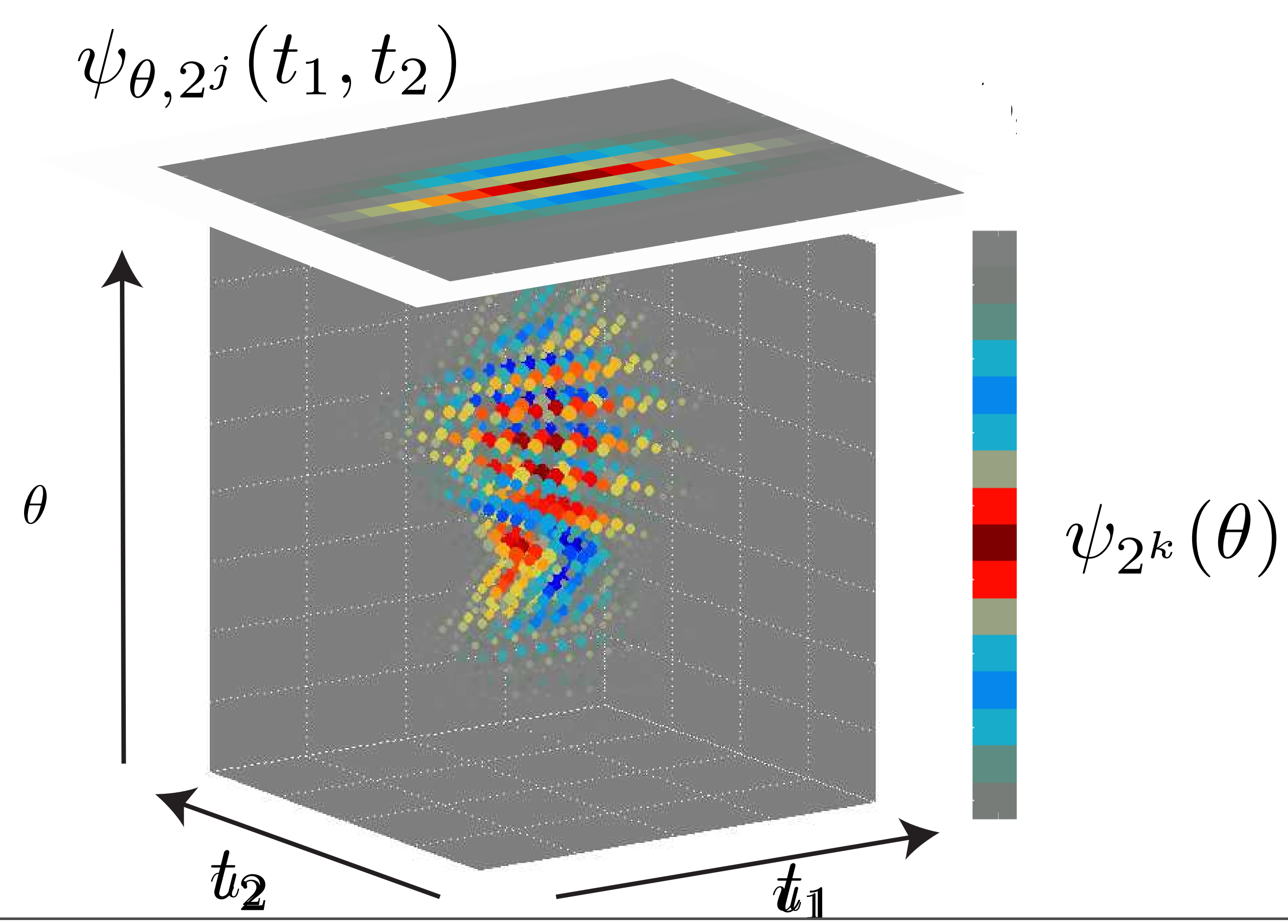
L. Sifre

- Separable wavelet: $t = (t_1, t_2)$

$$\overline{\psi}_\lambda(r_\theta, t_1, t_2) = \overline{\psi}_{2^k}(\theta)\ \psi_{\alpha,2^j}(t_1, t_2) \quad \text{with } \lambda = (2^k, 2^j, \alpha)$$

$$x \circledast \overline{\psi}_\lambda(g) = \int_G x(g')\, \overline{\psi}_\lambda(g'^{-1} g)\, dg' \quad \text{with} \quad g = (r, t)$$

$$x \circledast \overline{\psi}_\lambda(r_\theta, t) = \int_0^{2\pi} \left(\int_{\mathbb{R}^2} x(t', \theta')\, \psi_{\theta,2^j}(r_{-\theta'}(t - t')) \right) \overline{\psi}_{2^k}(\theta - \theta') d\theta' dt'$$



Scattering on a Lie Group

L. Sifre

- A roto-translation scattering applies

$$|W_2|x = \begin{pmatrix} x \circledast \overline{\phi}(r,t) \\ |x \circledast \overline{\psi}_{\lambda_2}(r,t)| \end{pmatrix}_{\lambda_2, r, t} \quad \text{and } |W_m| = |W_2| \text{ for } m \geq 2.$$

translation

scalo-roto-translation

roto-translation

+ renormalization

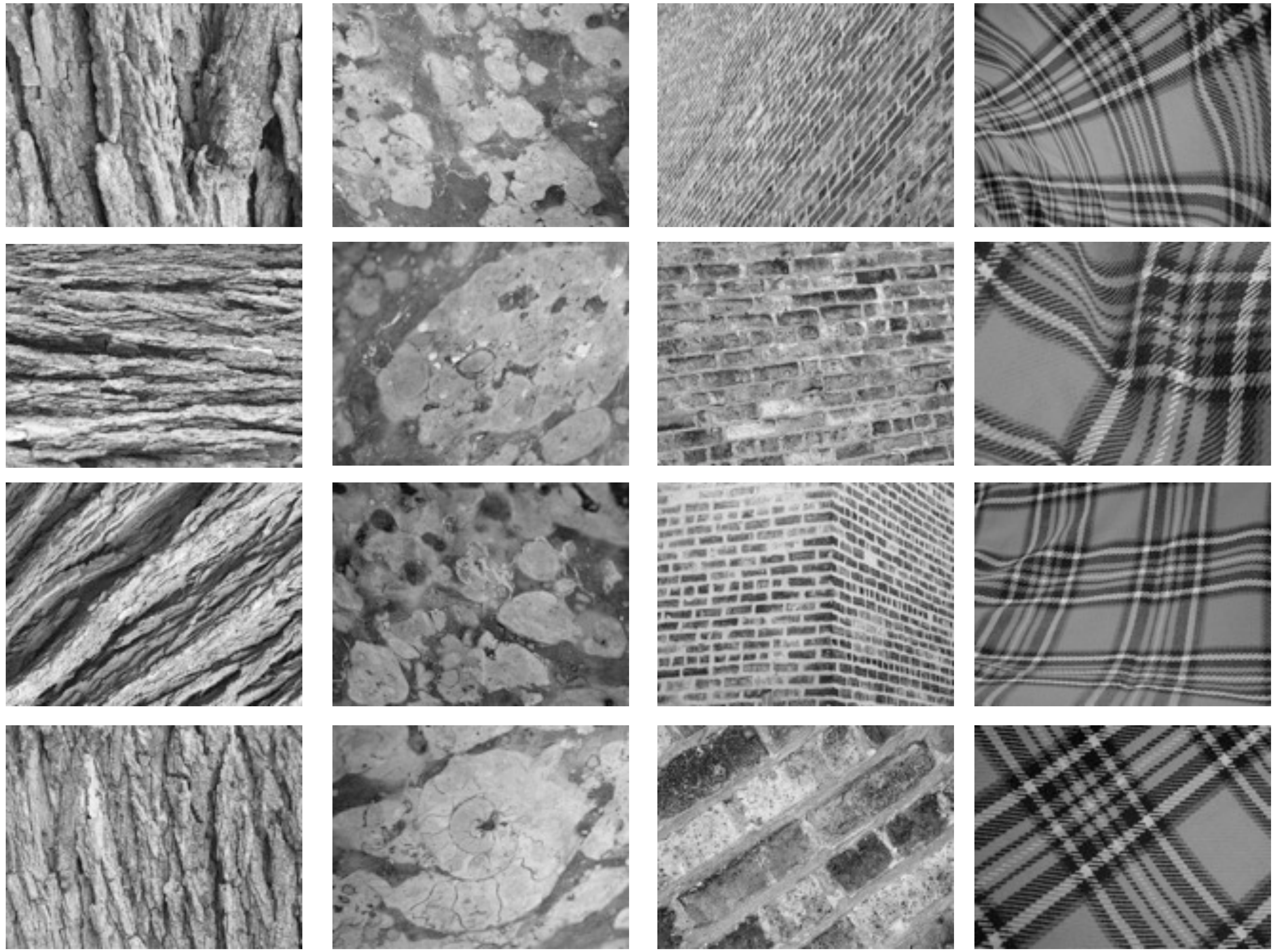
$$x \longrightarrow \boxed{|W_1|} \longrightarrow |x \star \psi_{2^j r}(t)| = w_j(2^j r, t) \longrightarrow \boxed{|W_2|} \longrightarrow |w_j \circledast \overline{\psi}_{\lambda_2}(2^j r, t)|$$

$$x \star \phi(t) \qquad\qquad w_j \circledast \overline{\phi}(2^j r, t)$$

Rotation and Scaling Invariance

Laurent Sifre

UIUC database:
25 classes



Scattering classification errors

$x \rightarrow Sx \rightarrow$ Supervised Linear Classifier: PCA/SVM $\rightarrow y$

Training	Translation	Transl + Rotation	+ Scaling
20	20 %	2%	**0.6**%

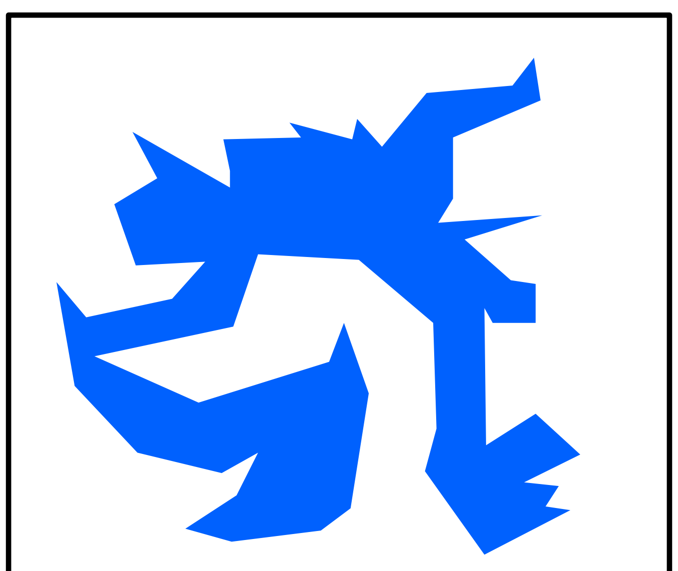
Learning Representations

Unsupervised Learning Representation

Learn with labeled examples $\{(x_i, y_i)\}_i$

$$x \in \mathbb{R}^d \longrightarrow \boxed{\Phi} \longrightarrow \Phi x \in \mathbb{R}^D \longrightarrow \boxed{\text{Linear Supervised Classifier}} \longrightarrow \text{class}$$

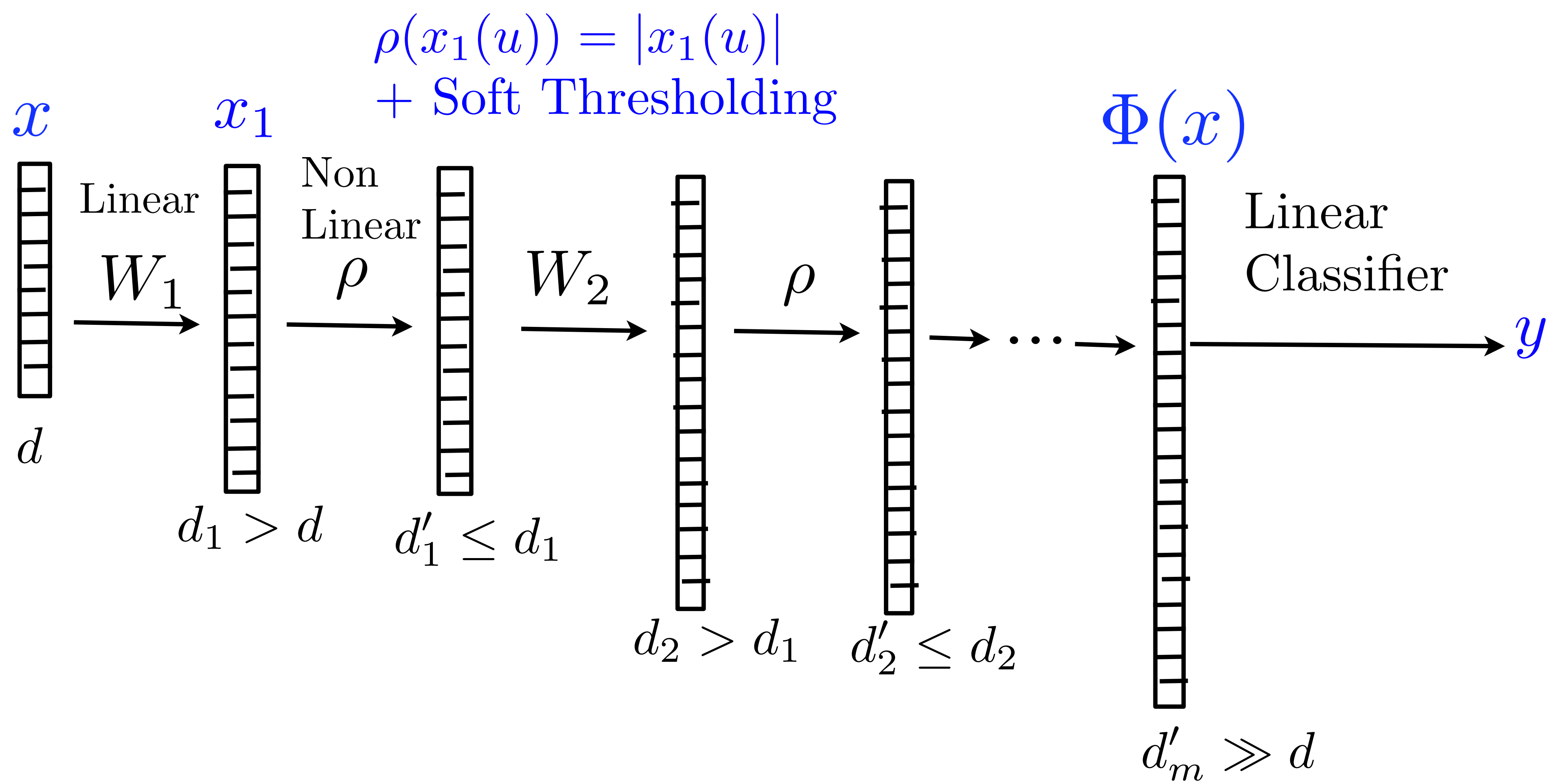
- **Unsupervised learning** of Φ from unlabeled examples $\{x_i\}$:
  - model the $\{x_i\}_i$ as realization of a random vector $X \in \mathbb{R}^d$
  - adapt Φ to the high-dimensional distribution $p(x)$ of X



but we can not estimate $p(x)$...

Scattering Generalization

- Towards general deep networks:



Generalized Scattering

Initialize $X_0 = X \in \mathbb{R}^N$

- Define $W_1 x = \Big(\langle x, \theta_n \rangle \Big)_{n \leq N_1}$ from $\mathbb{R}^N \to \mathbb{R}^{N_1}$ or $\mathbb{C}^{N_1}$

  Tight frame: $\sum_n |\langle x, \theta_n \rangle|^2 = \|x\|^2 \;\Leftrightarrow\; W_1^* W_1 = Id$

$$
\begin{aligned}
X_1 &= \Big| W_1 \Big(X_0 - E(X_0) \Big) \Big| \\
&= \Big(\Big| \langle X_0 - E(X_0), \theta_n \rangle \Big| \Big)_n
\end{aligned}
$$

Examples:

Wavelet transform $W_1 X = \Big(\sum_n X(n) \;,\; X \star \psi_{2^j}(n) \Big)_{1 \leq j \leq \log_2 N}$
with $N_1 = N \log_2 N + 1$

Identity $W_1 = I$ with $N_1 = N$.

Generalized Scattering

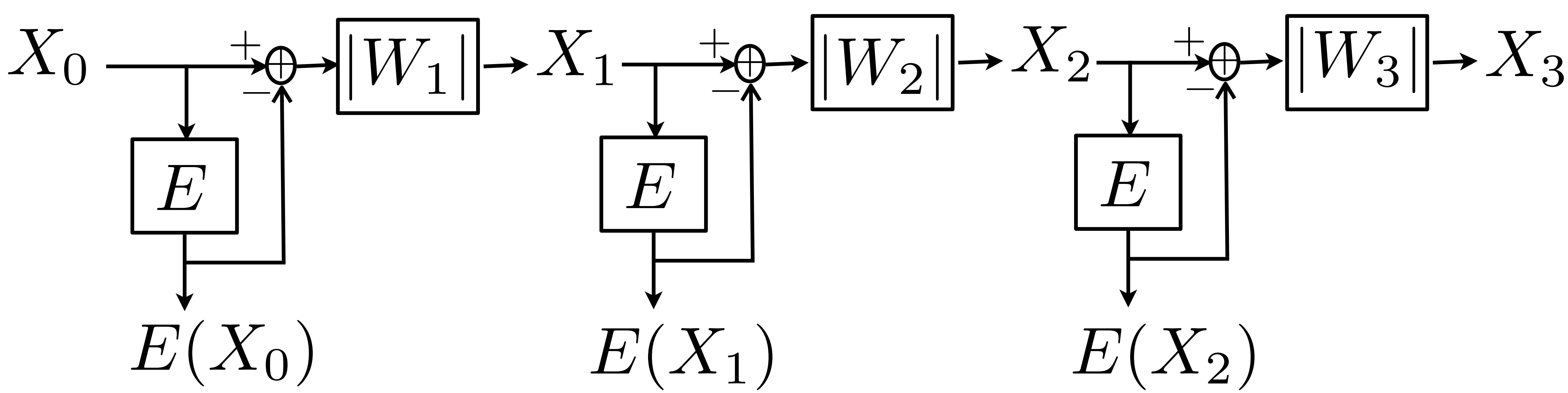
- Iteratively compute $X_m \in \mathbb{R}^{N_m}$: HOW ?

  Define $W_m x = \Big(\langle x, \theta_n^m \rangle\Big)_{n \leq N_m}$ from $\mathbb{R}^{N_{m-1}} \to \mathbb{R}^{N_m}$ or $\mathbb{C}^{N_m}$

  Tight frame: $\sum_n |\langle x, \theta_n^m \rangle|^2 = \|x\|^2 \Leftrightarrow W_m^* W_m = Id$

$$
\begin{aligned}
X_m &= \Big| W_m \Big(X_{m-1} - E(X_{m-1}) \Big) \Big| \\
&= \Big(\Big| \langle X_{m-1} - E(X_{m-1}), \theta_n^m \rangle \Big| \Big)_n
\end{aligned}
$$

- Expected scattering transform: $\overline{S} X = \{E(X_m)\}_{m \in \mathbb{N}}$



Revisit Expected Scattering

For wavelet transforms

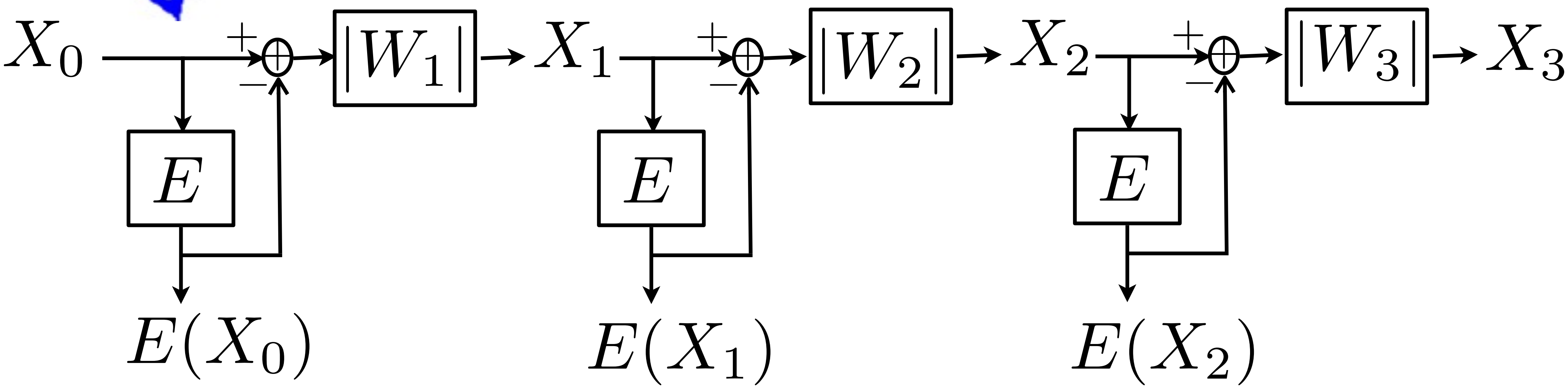
Initialize $X_0 = X$

$$
\begin{aligned}
X_1 &= |W_1(X_0 - E(X_0))| \\
X_2 &= |W_2(X_1 - E(X_1))| \\
X_3 &= |W_3(X_2 - E(X_2))| \\
&\ldots
\end{aligned}
$$

- Expected scattering: $\overline{S}X = \Big(E(X_m)\Big)_{m \in \mathbb{N}}$

$$
\overline{S}X = \begin{pmatrix} E(X) \\ E(|X \star \psi_{\lambda_1}|) \\ E(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) \\ E(|||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}|) \\ \ldots \end{pmatrix}_{\lambda_1, \lambda_2, \lambda_3, \ldots}
$$

Scattering Properties



$$\overline{S}X = \Big(E(X_m)\Big)_{m\in\mathbb{N}} \quad \text{and} \quad \|\overline{S}X\|^2 = \sum_{m\in\mathbb{N}} |E(X_m)|^2$$

- Since W_m is a tight frame operator $\|W_m x\| = \|x\|$ and

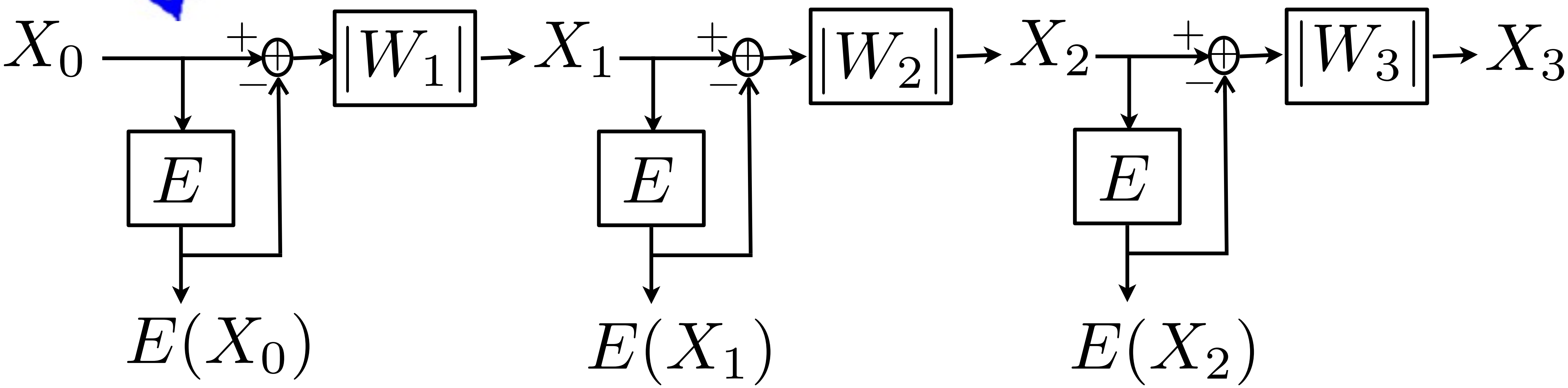
$$\||W_m|x - |W_m y|\| \leq \|x - y\|$$

Theorem: *I. Waldspurger*

$$\|\overline{S}X - \overline{S}Y\| \leq E(\|X - Y\|^2)$$

$$\|\overline{S}X\| = E(\|X\|^2)$$

Scattering Properties



$$\overline{S}X = \Big(E(X_m)\Big)_{m\in\mathbb{N}} \quad \text{and} \quad \|\overline{S}X\|^2 = \sum_{m\in\mathbb{N}} |E(X_m)|^2$$

Theorem: The empirical average estimation $\widehat{E(X_m)}$ of $E(X_m)$ from P realizations of X satisfies

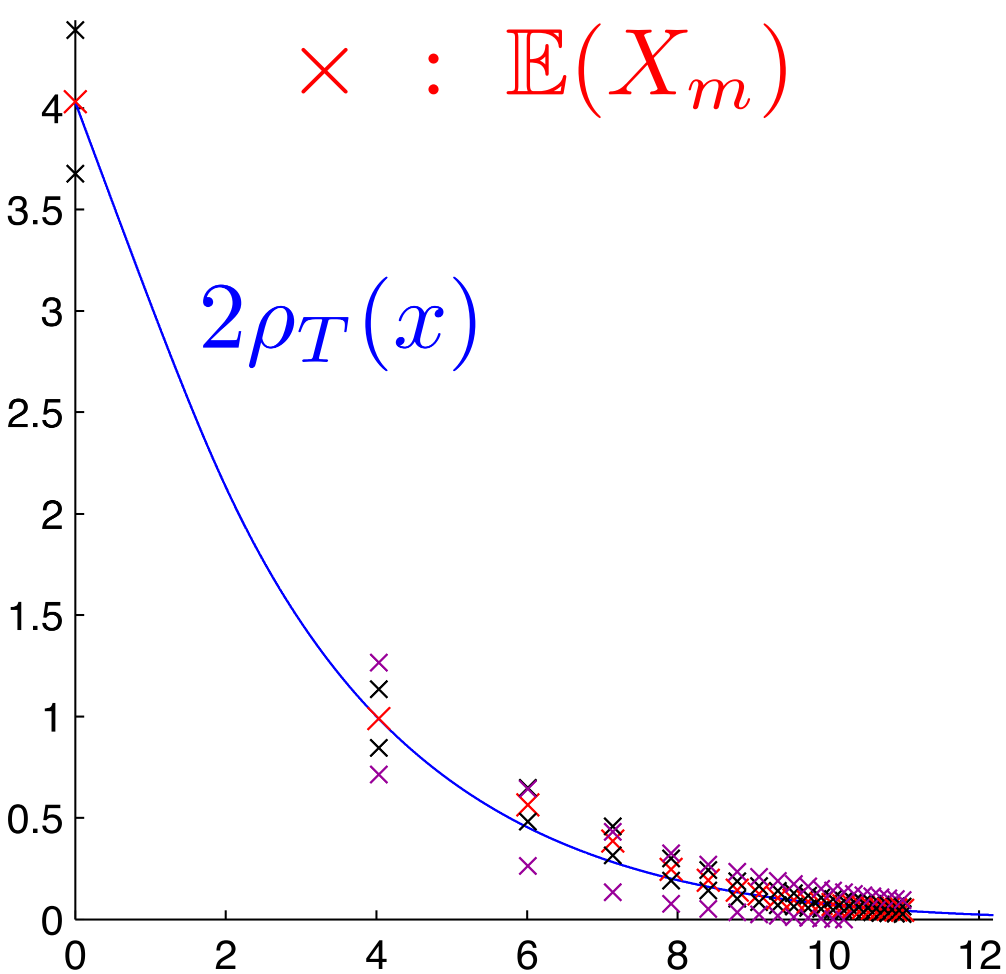
$$E\Big(\|\widehat{E(X_m)} - E(X_m)\|^2\Big) \leq C\, m \frac{E(\|X\|^2)}{P} \ .$$

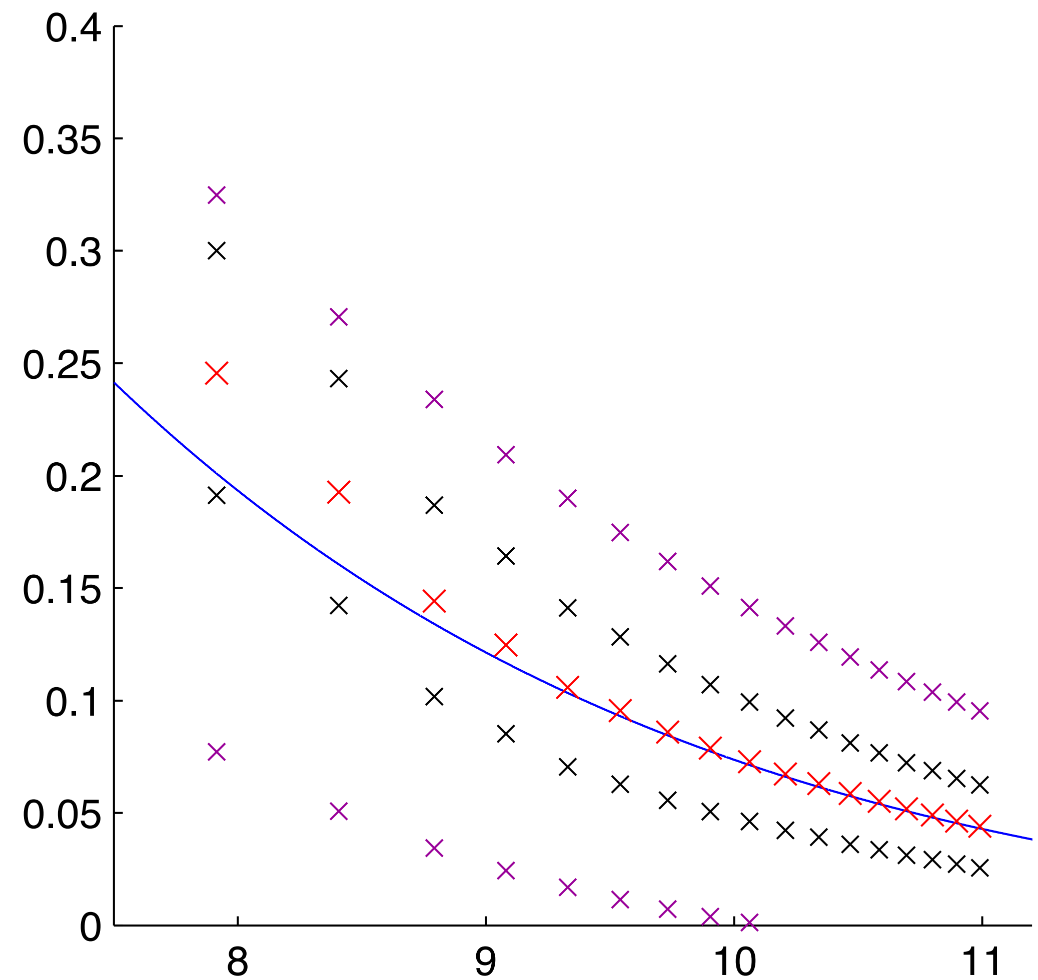
Almost Soft Thresholding

Theorem Let $\rho_T(x) = \max(|x| - T, 0)$. *I. Waldspurger*

If for all m, $W_m = I$ in $\mathbb{R}$ then

$$\mathbb{E}(X_m) \sim 2\,\rho_{T_m}(X) \;\text{ with } T_m = \textstyle\sum_{n=0}^{m} \mathbb{E}(X_n).$$





Representation of Random Processes

- Expected scattering: $\overline{S}X = \Big(E(X_m)\Big)_{m\in\mathbb{N}} = \Big(E(U_m X)\Big)_{m\in\mathbb{N}}$

Theorem *(Boltzmann)* The distribution $p(x)$ which satisfies

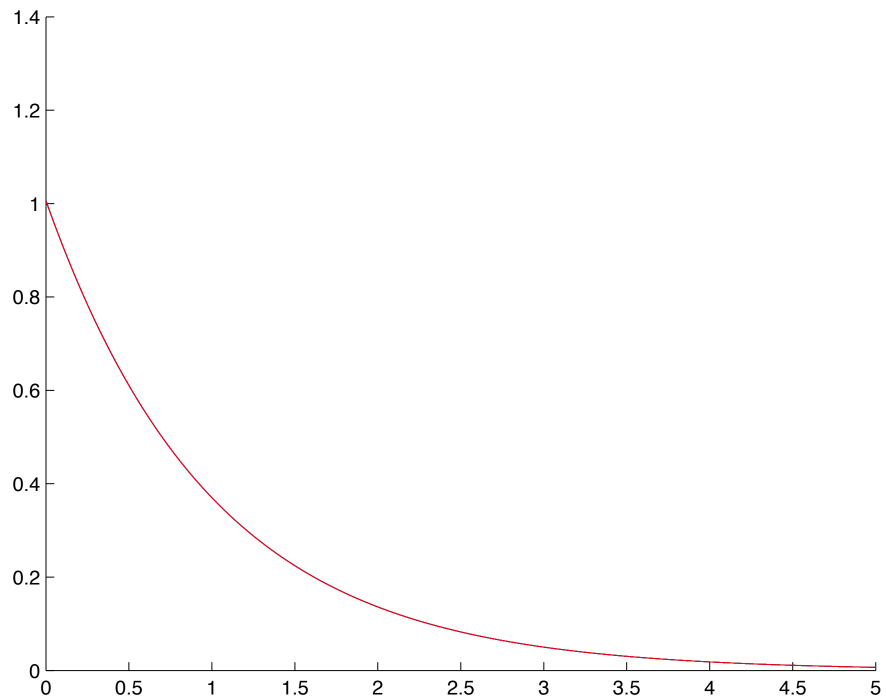
$$\int_{\mathbb{R}^N} U_m x \; p(x)\, dx = E(U_m X)$$

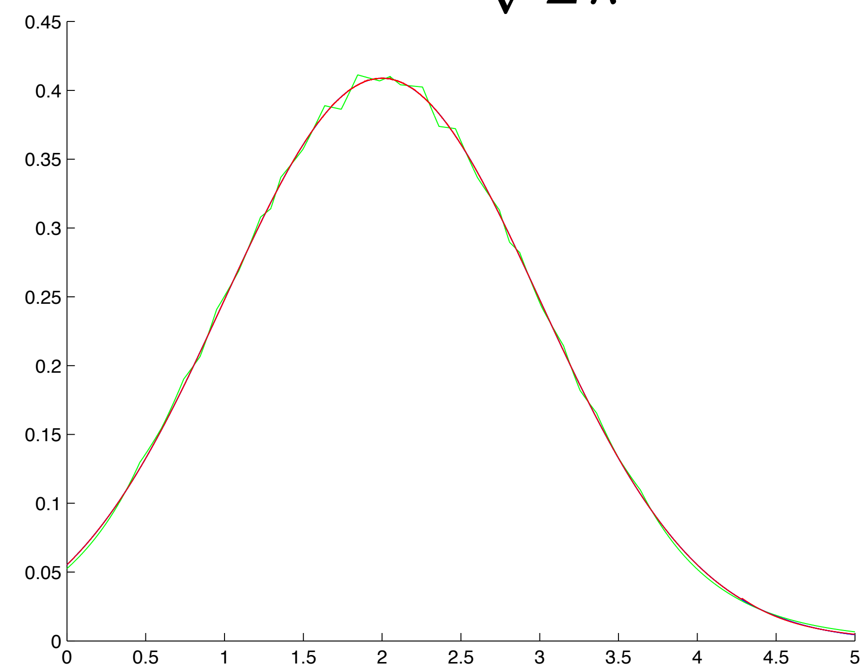
and maximizes the entropy $-\int p(x)\,\log p(x)\,dx$

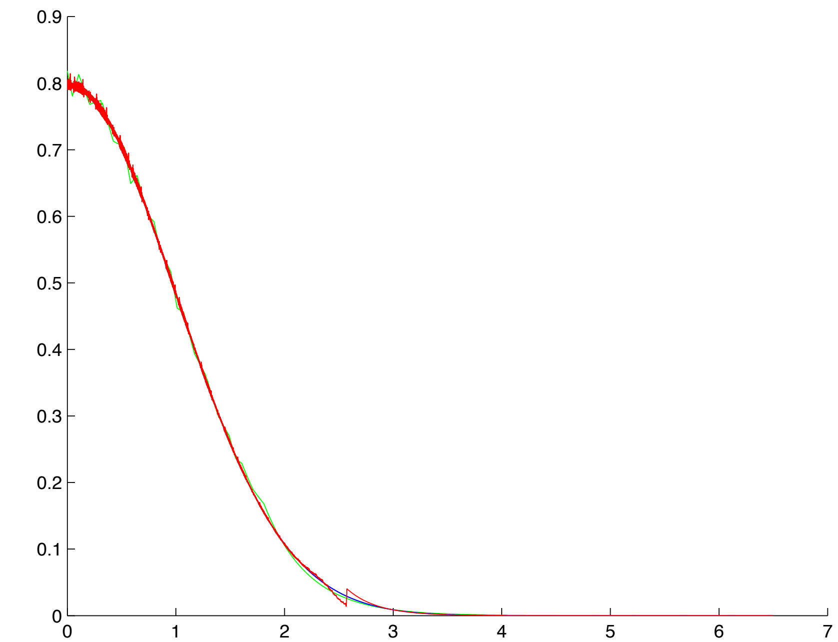
can be written: $\quad p(x) = \dfrac{1}{Z}\exp\Big(\sum_{m=1}^{\infty} \lambda_m \,.\, U_m x\Big)$

Approximation of Distributions

$$p(x) = \frac{e^{-|x|}}{2} \qquad p(x) = \frac{e^{-|x-2|^2/2}}{\sqrt{2\pi}} \qquad p(x) = \frac{e^{-|x|^2/2}}{\sqrt{2\pi}}$$

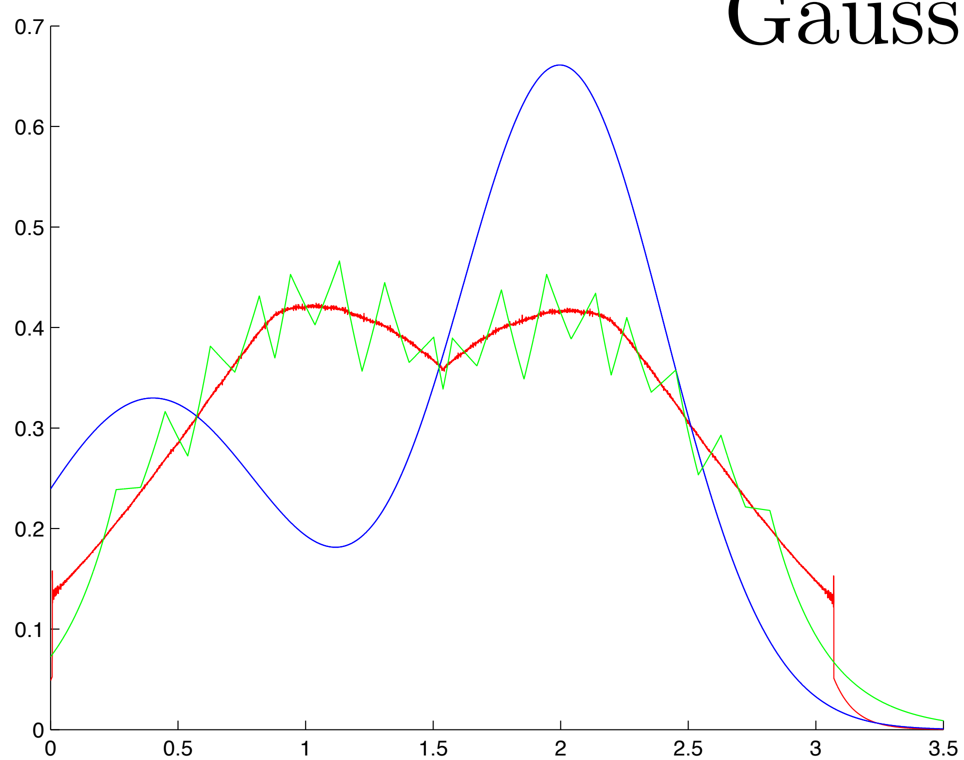


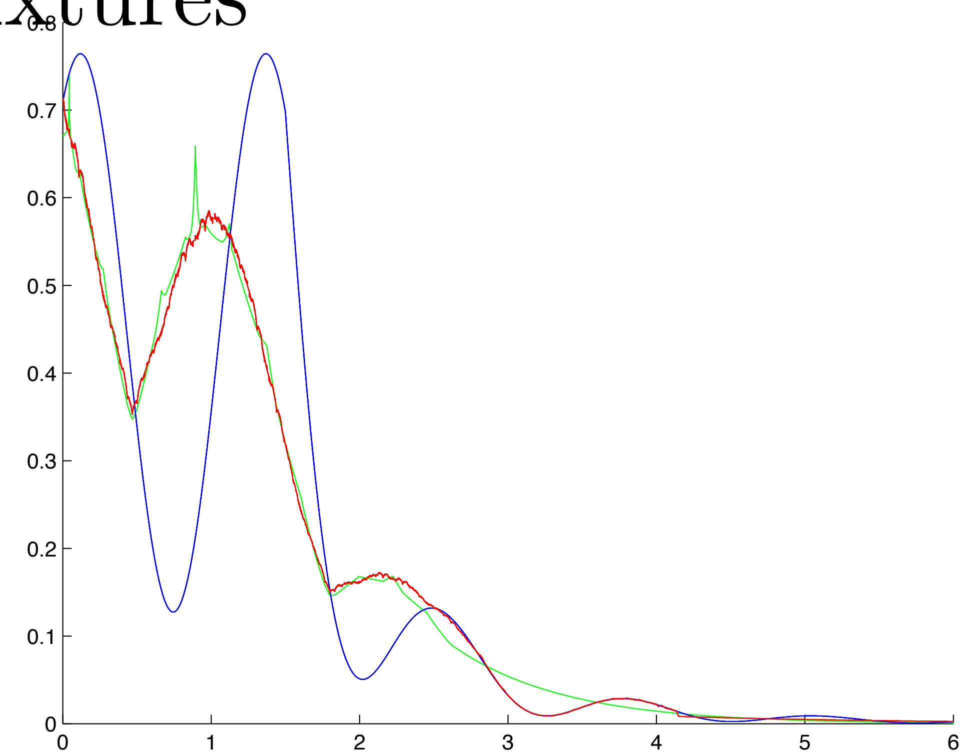




- Converges numerically for all $p(x) = C\,(1+x)^k$ for $k \geq 1$.

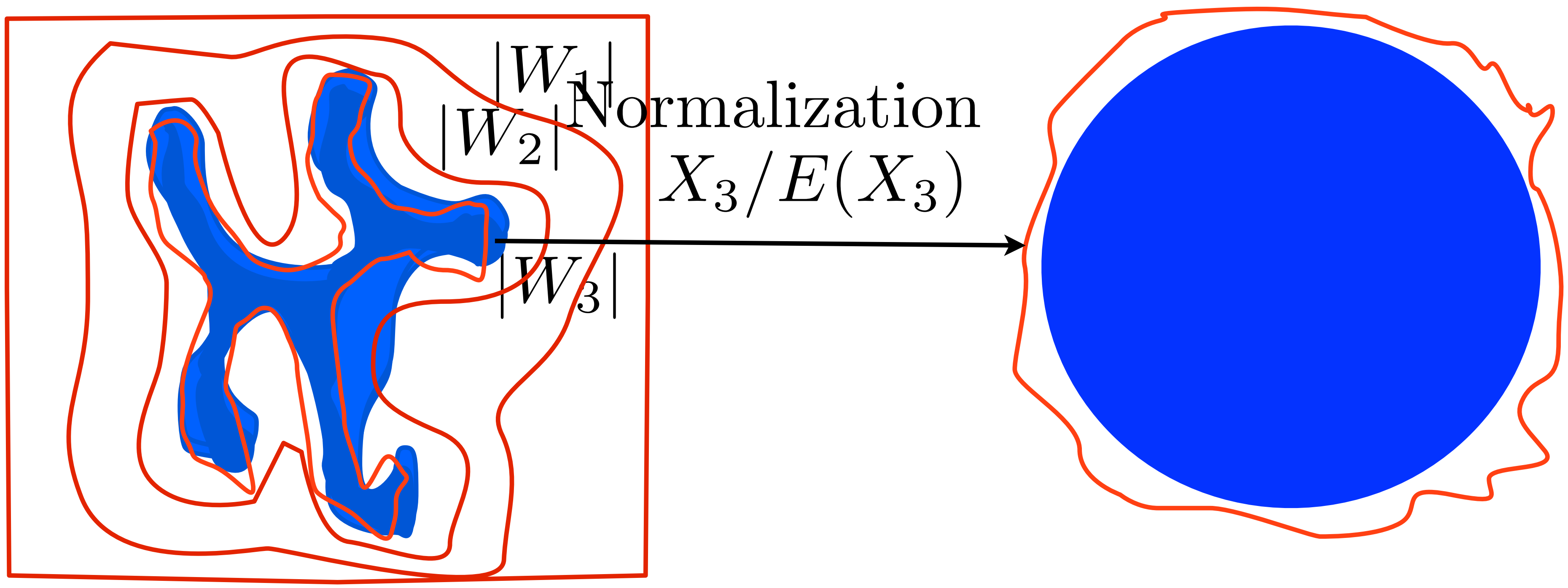
Gaussian Mixtures





Optimized Space Contraction

- A generalized scattering progressively contracts the space
- For classification, we need to squeeze the space while minimizing the data volume reduction



Proposition: The data volume reduction at layer m is

$$E(\|X_{m-1} - E(X_{m-1})\|^2) - E(\|X_m - E(X_m)\|^2) = \|E(X_m)\|^2$$

$$\Rightarrow \text{ for all } m \text{ minimize } \|E(X_m)\| \,.$$

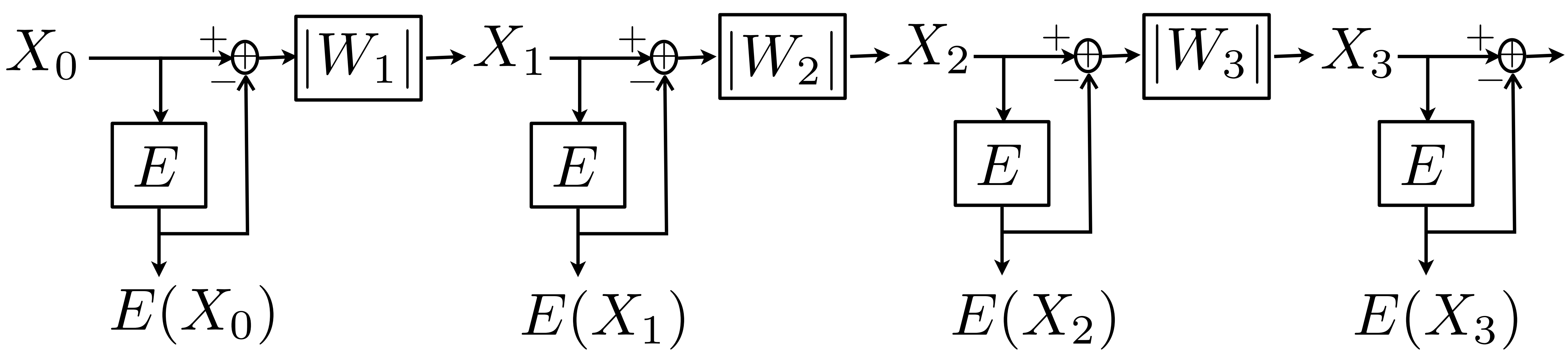
Sparse Layerwise Learning

$$X_m = |W_m(X_{m-1} - E(X_{m-1})| \quad \text{with } W_m^* W_m = Id.$$

- Given $X_{m-1} - E(X_{m-1})$ we compute W_m by minimizing

$$\|E(X_m)\| = \Big\| \underbrace{E\Big(|W_m(X_{m-1} - E(X_{m-1})|\Big)}_{\mathbf{l^1} \text{ norm across realizations}} \Big\|$$

$\Rightarrow W_m$ defines a sparse representation of $X_{m-1} - E(X_{m-1})$
Sparse dictionary learning problem.

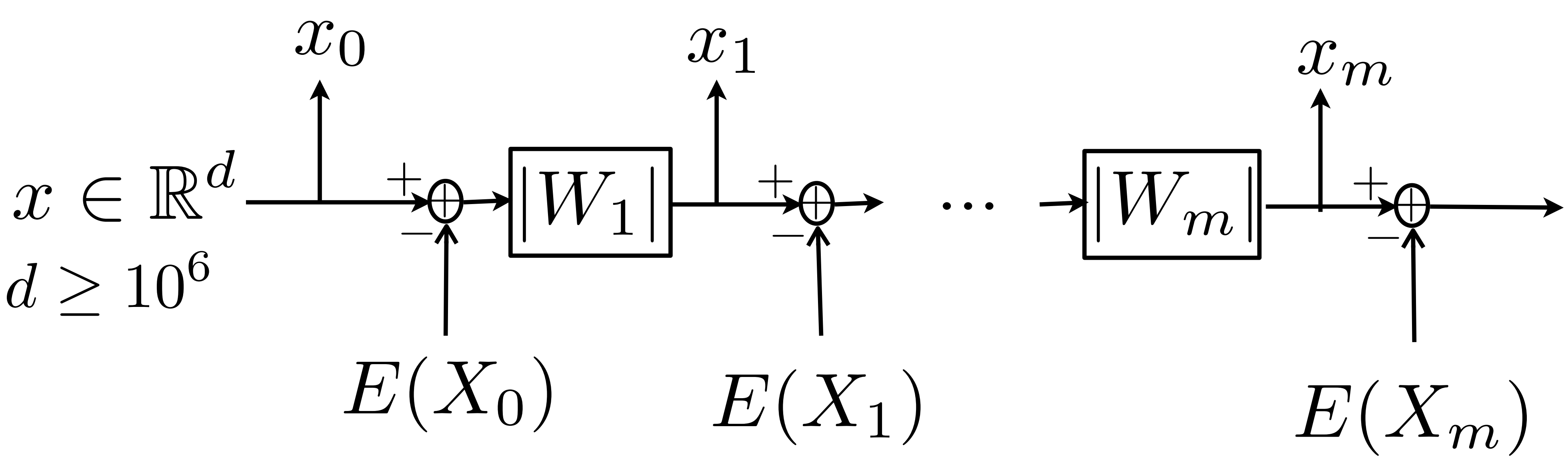


Determinist Scattering Transform

- Given $\overline{S}X = \Big(\mathbb{E}(X_m)\Big)_{m\in\mathbb{N}}$

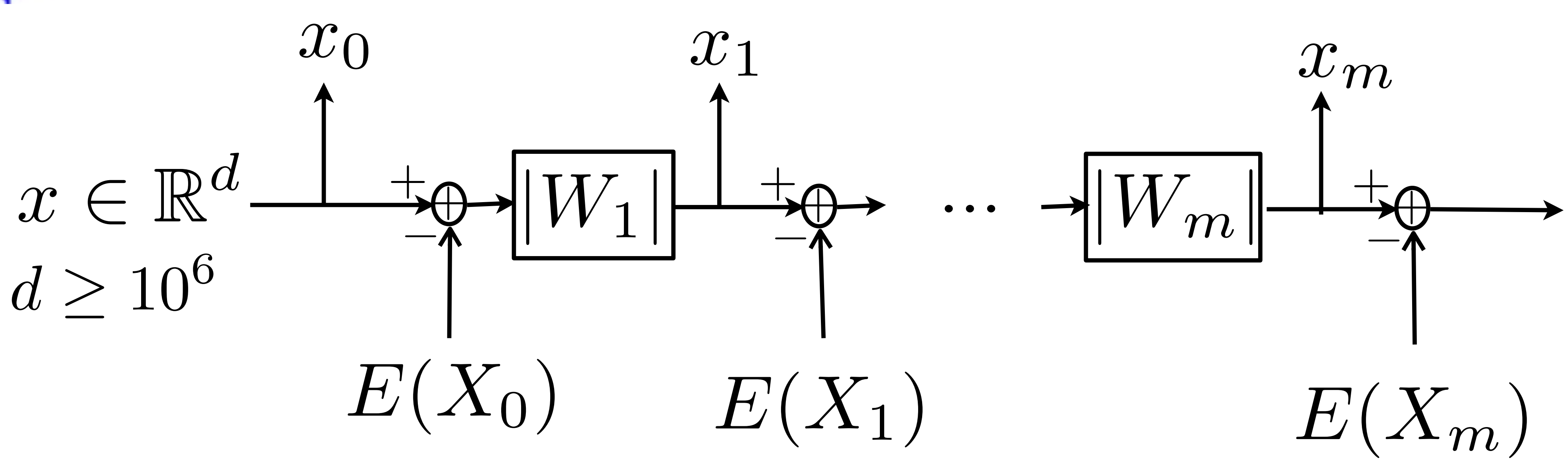
$$\text{Initialize } x_0 = x$$

$$\forall m \quad x_m = \Big|W_m\Big(x_{m-1} - E(X_{m-1})\Big)\Big|$$

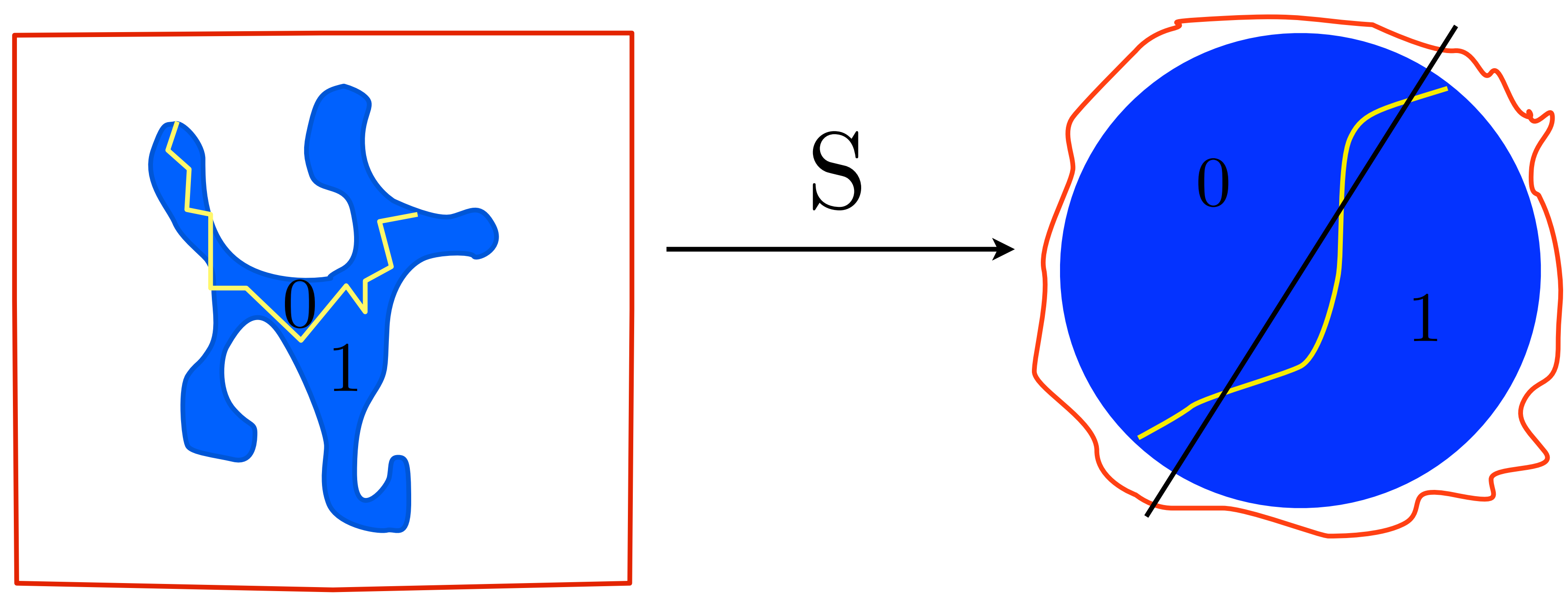


- Scattering transform $Sx = \Big(x_m\Big)_{m\in\mathbb{N}}$

Supervised Linear Classifiers



Binary classification
$y(x) = 0$ or 1



- Which models to evaluate the classification loss ?

Conclusion

- High dimensional classification algorithms have considerably improved in the last few years with many applications.

- Opportunity to develop different non-parametric approaches to modeling and estimation of stochastic processes.

- **Papers and Softwares:** www.di.ens.fr/data/scattering