Classification by Invariant Scattering

Joan Bruna, Stéphane Mallat

Centre de Mathématiques Appliquées
Ecole Polytechnique

Anestis Birthday
March 2011

Classification by Invariant Scattering

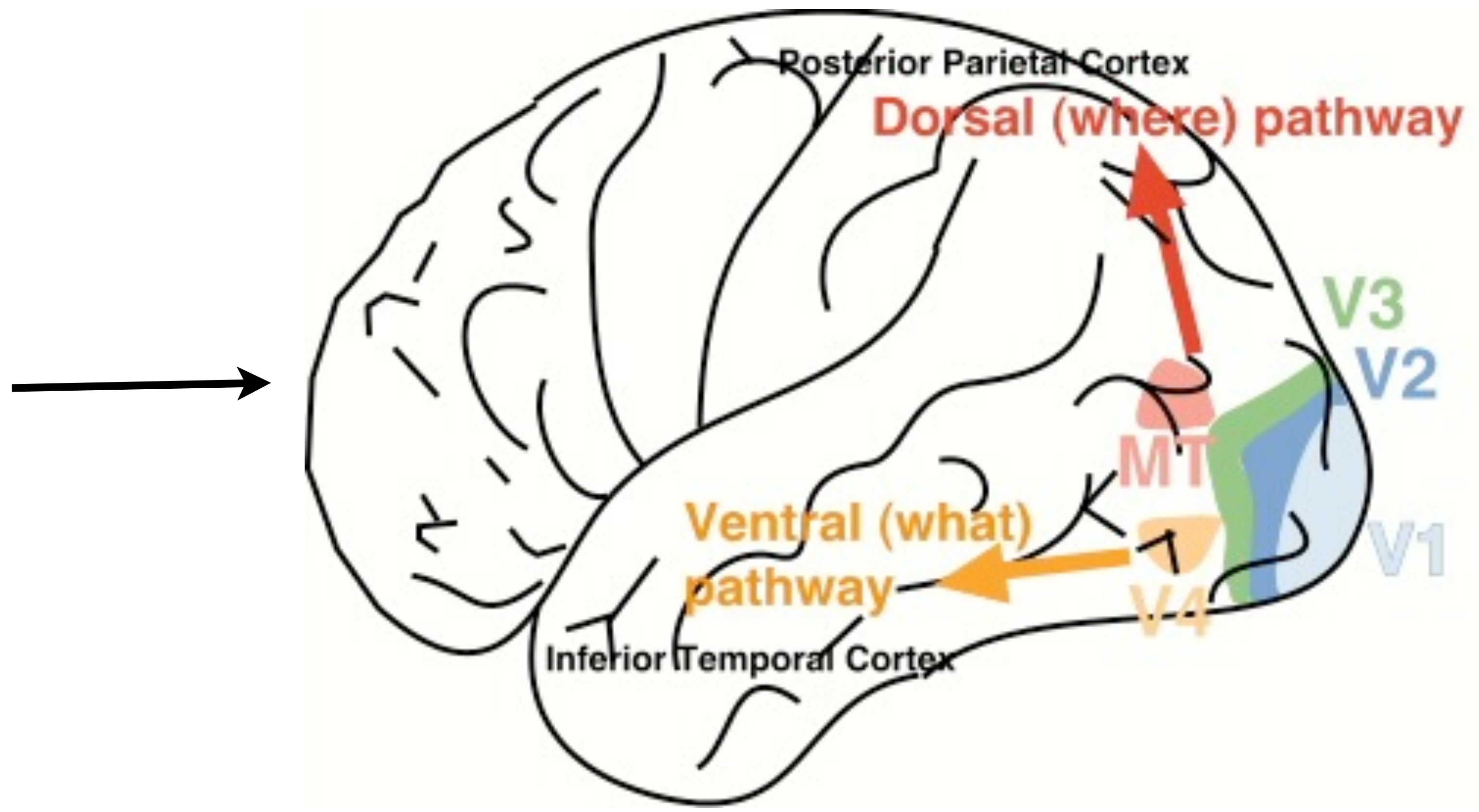
Joan Bruna, Stéphane Mallat

Centre de Mathématiques Appliquées
Ecole Polytechnique

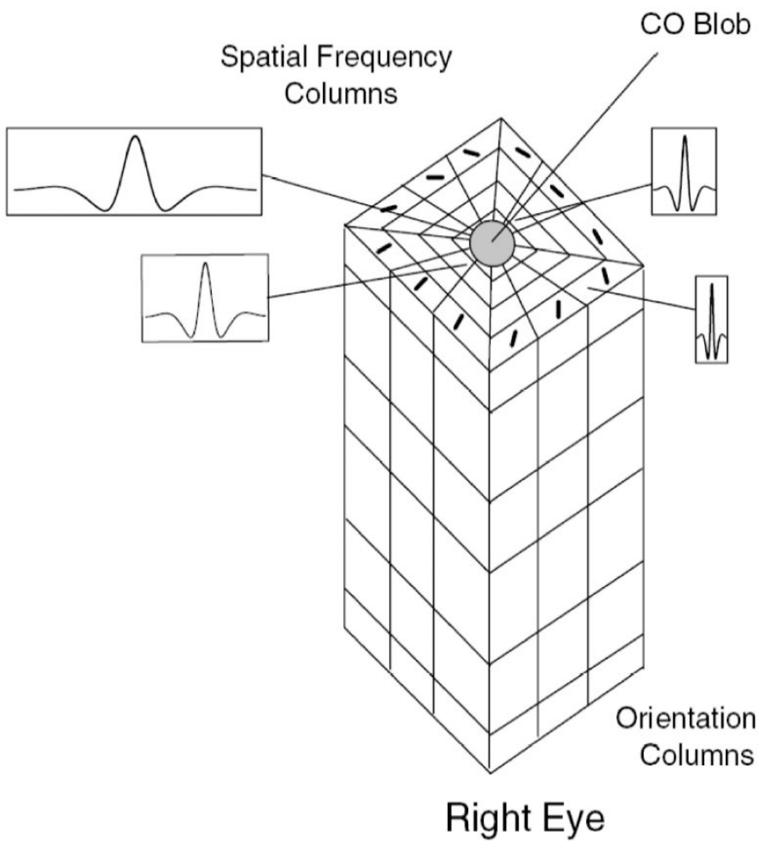
CMAP

Anestis Birthday
March 2011

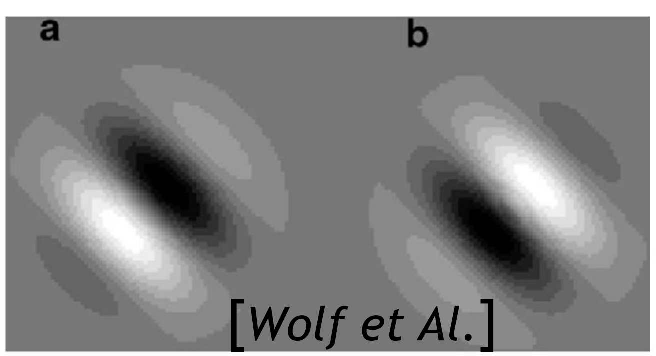
The Best Image Classifier



Hypercolumns in V1:
directional wavelets



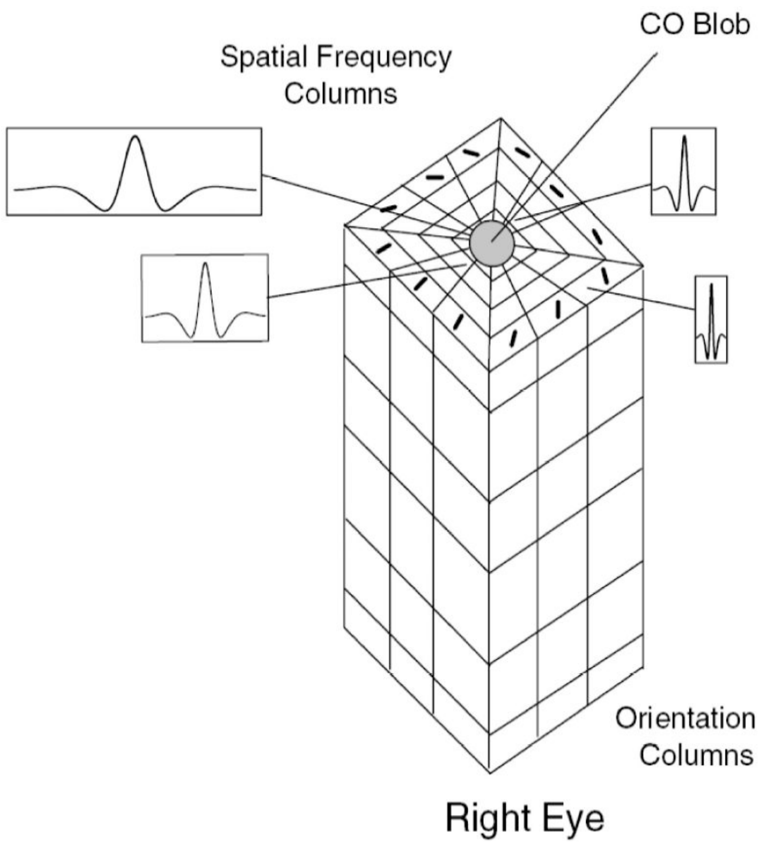
Simple cells Gabor linear models



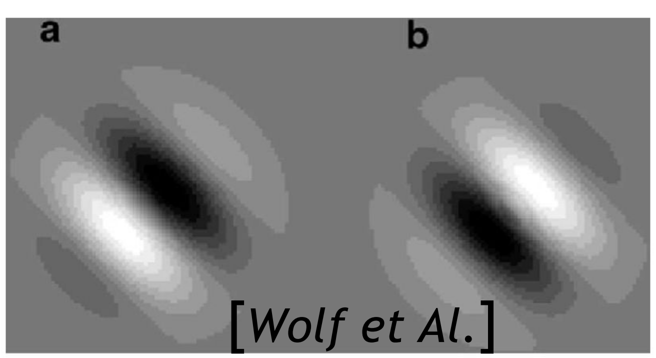
$$\psi(x) = \theta(x) e^{i\xi x}$$

Psychophysics of Vision

Hypercolumns in V1:
directional wavelets



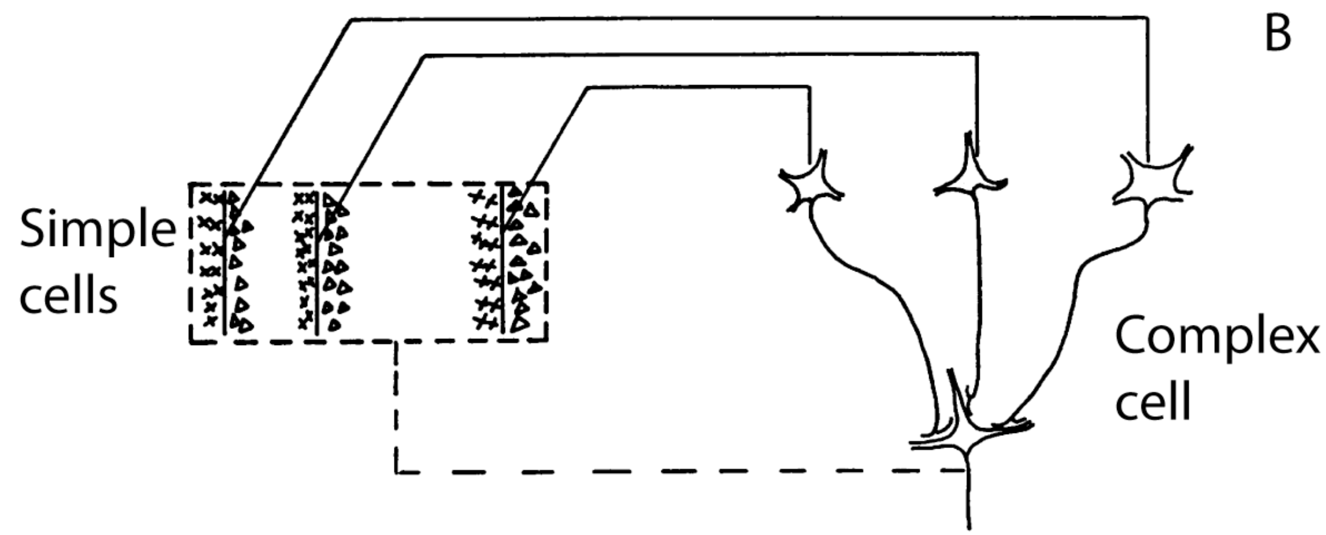
Simple cells Gabor linear models



$$\psi(x) = \theta(x)e^{i\xi x}$$

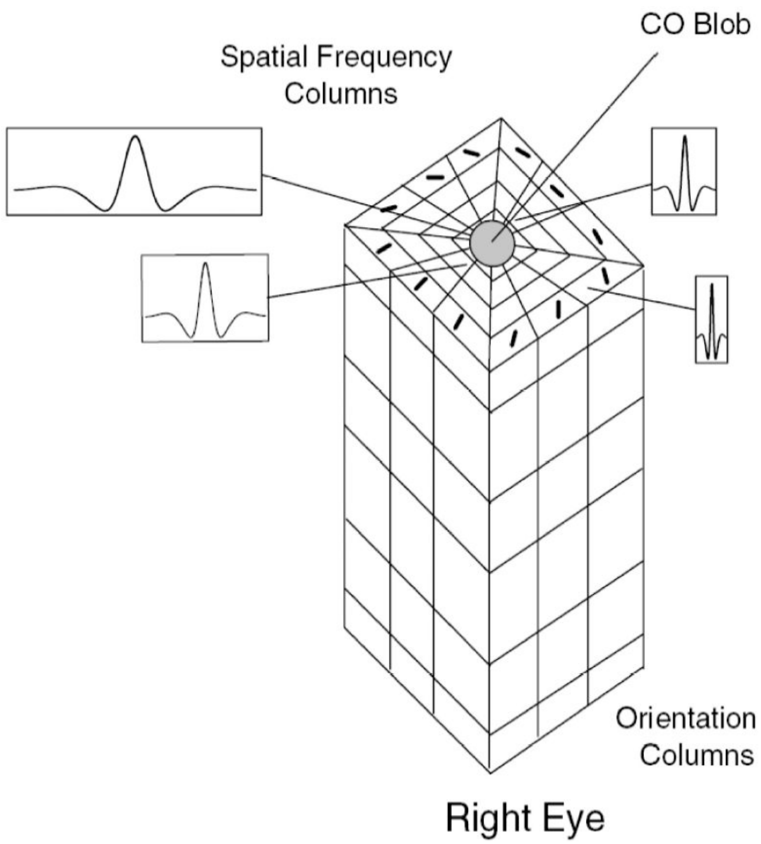
Complex Cells

- Non-linear
- Large receptive fields
- Some forms of invariance

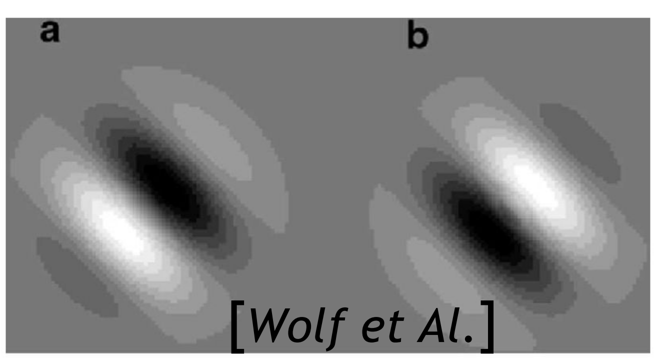


Psychophysics of Vision

Hypercolumns in V1:
directional wavelets



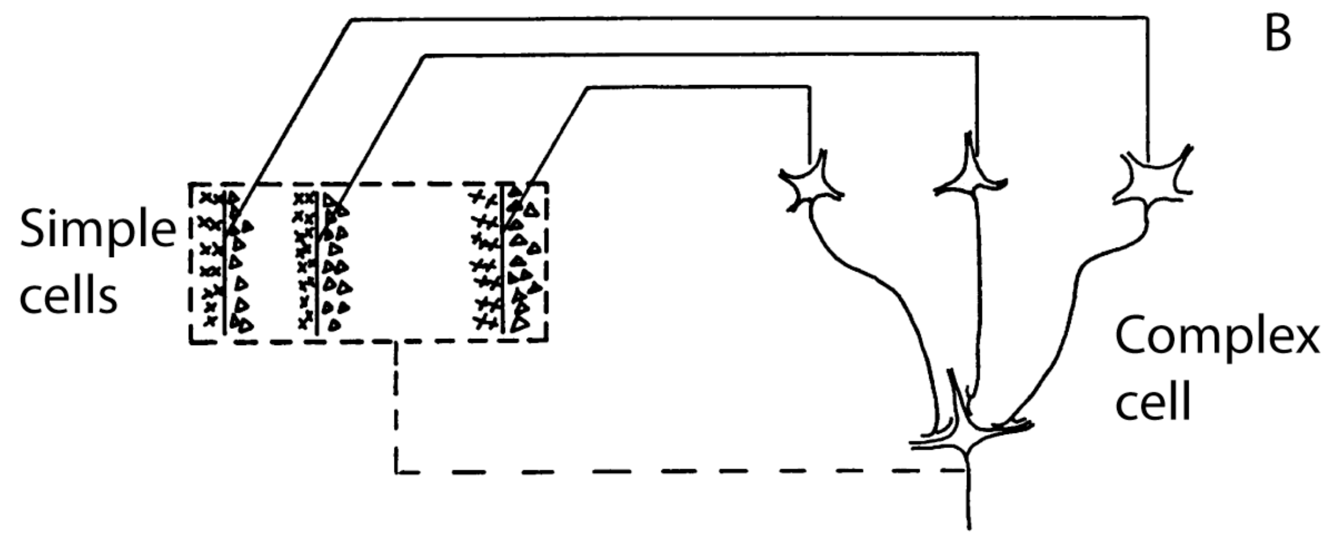
Simple cells Gabor linear models



$$\psi(x) = \theta(x)e^{i\xi x}$$

Complex Cells

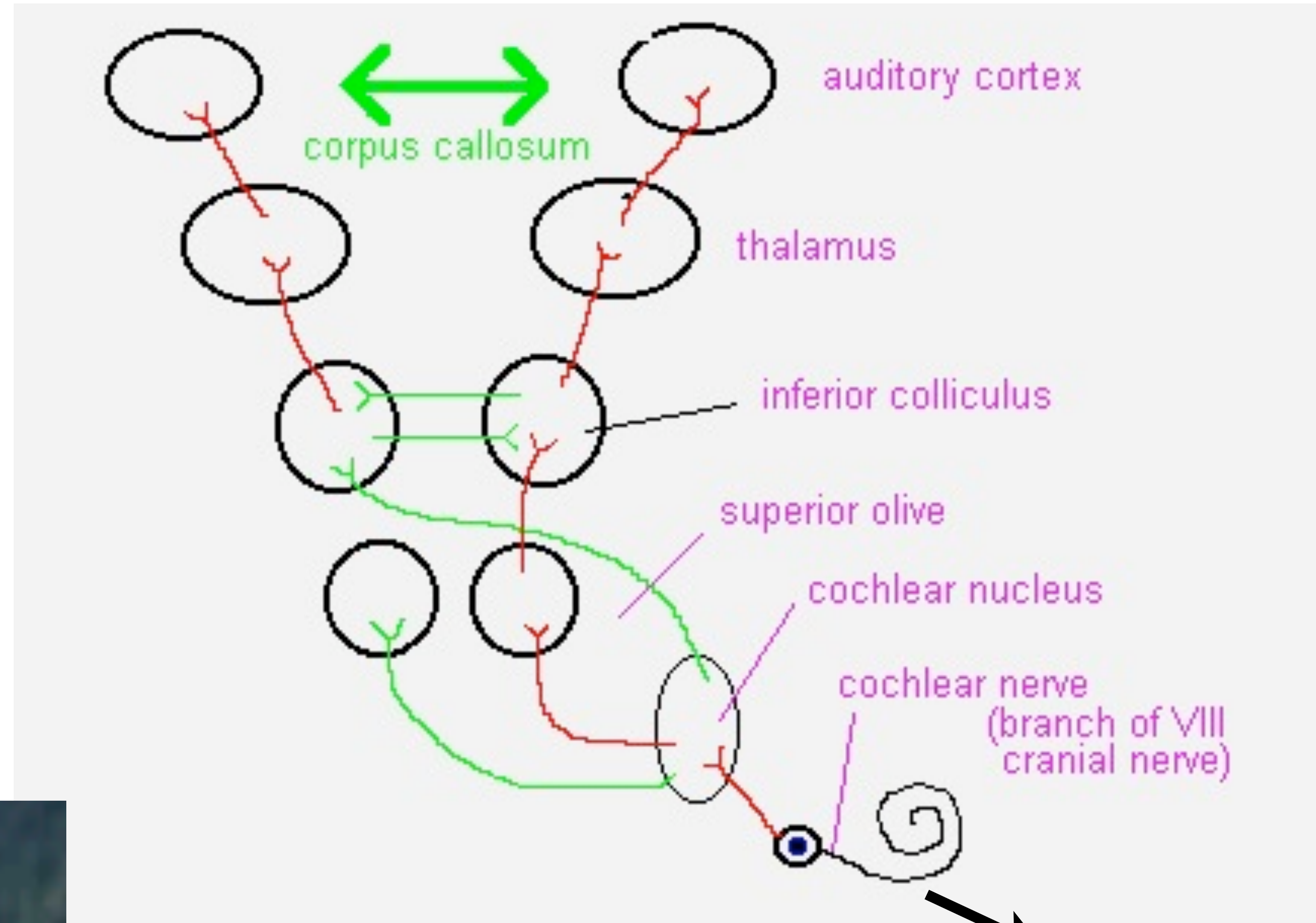
- Non-linear
- Large receptive fields
- Some forms of invariance



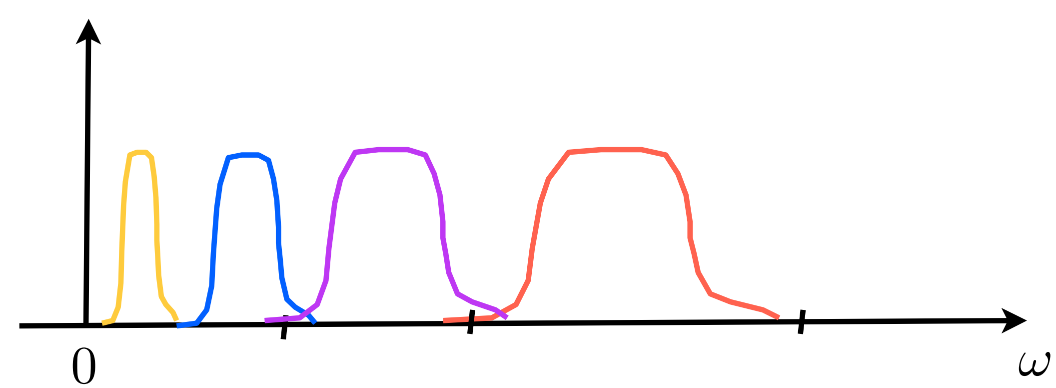
«What» Pathway towards V4:

- More specialized invariance
- «Grand mother cells»

Audio Psychophysics



Cochlea:
dilated wavelet filters



Metric for Classification

- Classification requires finding a metric to compare signals, with:
  - small distances $d(f,g)$ within a class
  - large distances $d(f,g)$ across classes.

- If one finds a representation $\Phi(f)$ such that

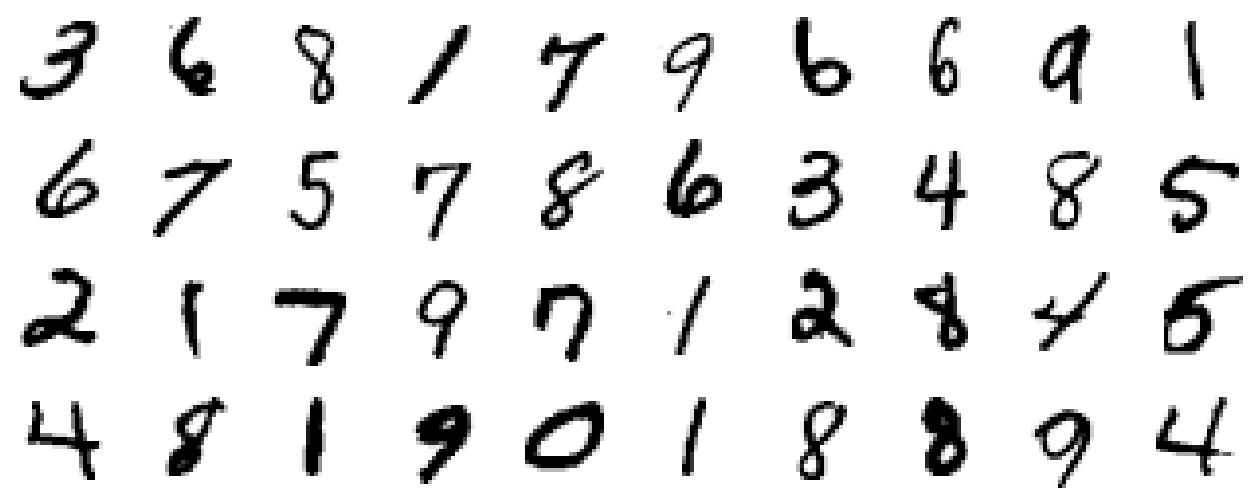
  $$d(f,g) = \|\Phi(f) - \Phi(g)\| \quad \text{(kernel metric)}$$

  then the classification may be linearized (SVM, PCA,...).

- Is there an appropriate kernel metric, which Φ ?
- Should it increase dimensionality ?

Perceptual Distance

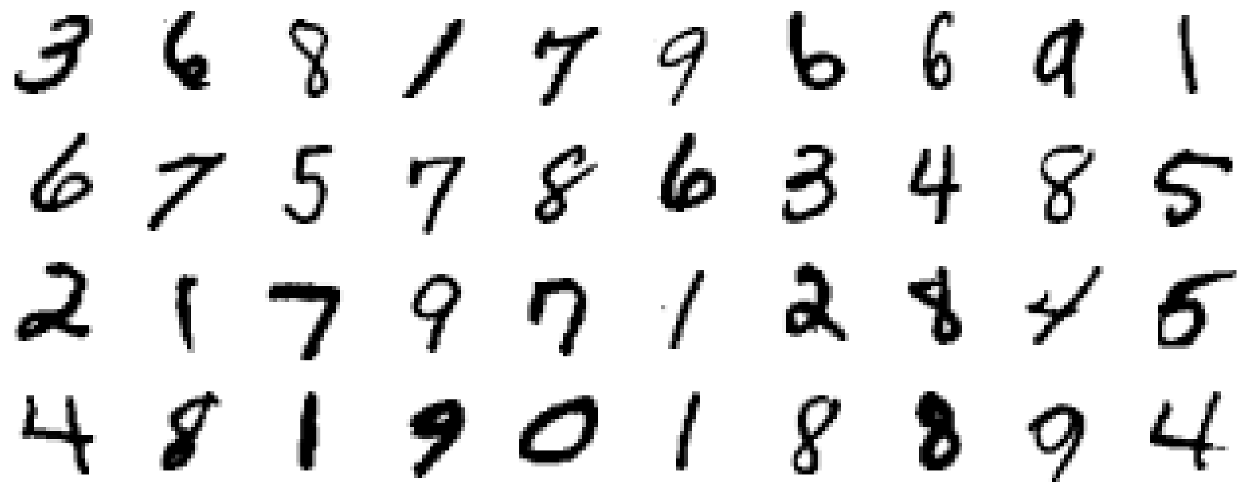
- Invariant to translation or scaling.
- Stable to elastic deformations.

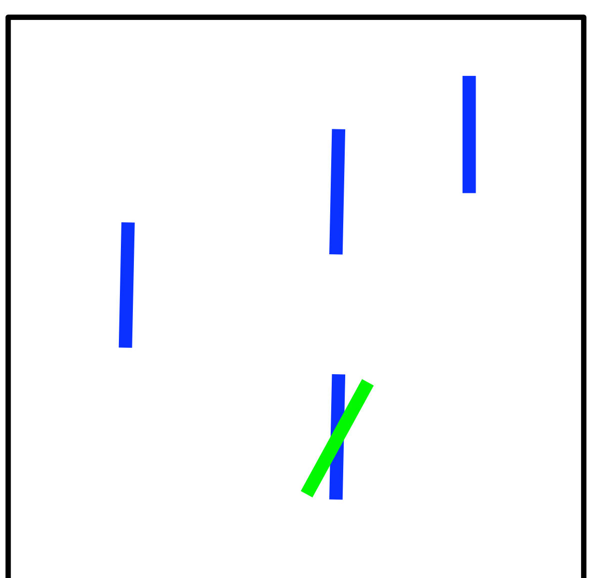


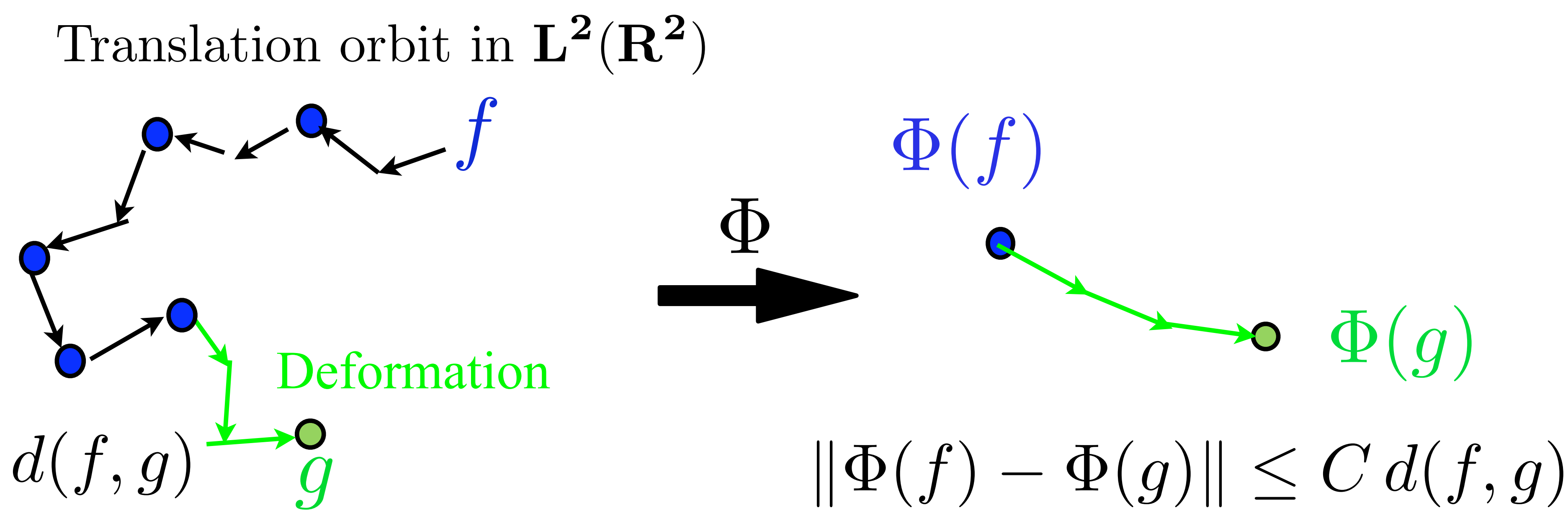
Perceptual Distance

- Invariant to translation or scaling.

- Stable to elastic deformations.







Distance from Representations

- Distance: $\|\Phi(f) - \Phi(g)\|$.

Invariance to groups of operators $\{D_\tau\}_\tau$ such as rigid translations $D_\tau f(x) = f(x-\tau)$:

$$\Phi(D_\tau f) = \Phi(f) \text{ if } \tau = cst, \; weak\ property.$$

Lipschitz continuity to deformations $D_\tau f(x) = f(x - \tau(x))$

$$\tau(x) \approx \tau(x_0) + \nabla\tau(x_0)(x - x_0)$$

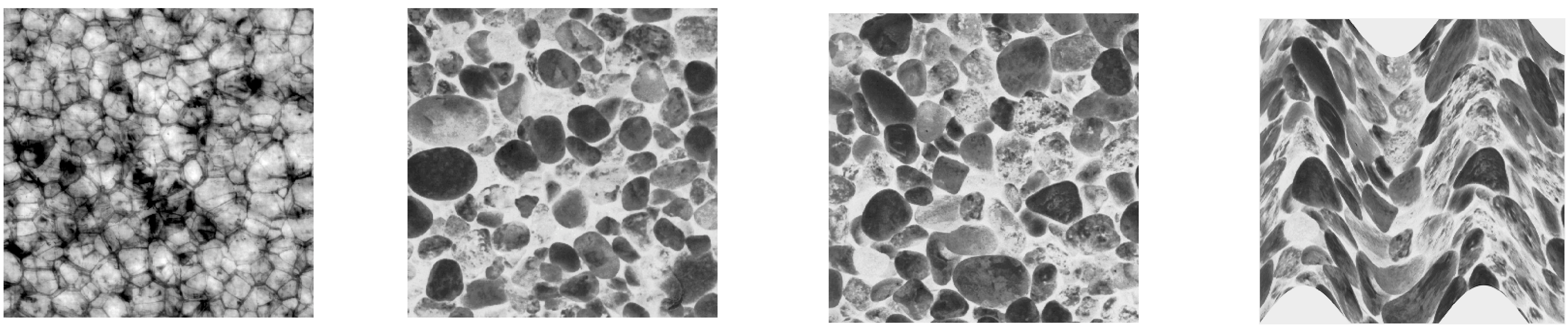
$$\|\Phi(f) - \Phi(D_\tau f)\| \leq C\,\|f\|\|\nabla\tau\|_\infty \; .$$

Linearizes local deformations.

Texture Discrimination

- A texture F is stationary but typically non-Gaussian and non-Markovian process.

Textures with same power spectrum



- We want to find Φ so that:
  - Invariance: $\Phi(F) = E\{\Phi(F)\}$ with probability 1.
  - Lipschitz continuity to random deformations:

$$E\{\|\Phi(F) - \Phi(D_\tau F)\|^2\} \leq C\, E\{|F|^2\}\, E\{\|\nabla\tau\|_\infty^2\} \ .$$

Overview

- Failures of Fourier and wavelet representations.
- Invariance and continuity through scattering space contraction.
- Representation of stationary processes
- Scattering PCA classification of patterns and textures
- Learning invariance and contraction for classification.

Deformation Instability of Fourier

- Elastic deformation $D_\tau f(x) = f(x - \tau(x))$ with $|\nabla\tau| < 1$.

- The Fourier modulus is translation invariant:

  If $\tau(x) = cst$ then $|\widehat{D_\tau f}(\omega)| = |\hat{f}(\omega)|$ $: \Phi(f) = |\hat{f}|$.

- High frequencies are not Lipschitz continuous to deformations:

  If $\tau(x) \neq cst$ then $\tau(x) \approx \tau(x_0) + \nabla\tau(x_0) \cdot (x - x_0)$ affine.

  If $\hat{f}(\omega)$ has energy at high frequencies ξ:

$$\Rightarrow \|\, |\widehat{D_\tau f}| - |\hat{f}|\, \| \sim \|\nabla\tau \cdot \xi\|_\infty$$

Sparsity and Discriminability

- Modulus reduces discriminability for non-sparse signals:

  $\delta(x)$ and e^{ix^2} have same Fourier modulus (constant).

- In an orthonormal basis $\mathcal{B} = \{g_m\}_{m \in \mathbf{Z}}$, for any f:

$$\Big\{h \ : \ |\langle h, g_m\rangle| = |\langle f, g_m\rangle|\Big\}$$

  has a dimension equal to the number of non-zero $\langle f, g_m\rangle$.

- The loss of discriminability with a modulus is small

  for classes of sparse signals in $\mathcal{B}$ (Kolmogorov entropy).

Wavelet Transforms

- In $\mathbf{L^2}(\mathbf{R})$, dilated wavelets: $\psi_j(x) = a^{-j}\psi(a^{-j}x)$ with $a > 1$.

- In $\mathbf{L^2}(\mathbf{R^2})$, $x = (x_1, x_2)$, dilated and rotated wavelets:

$$\psi_{j,\gamma}(x) = 2^{-2j}\,\psi(2^{-j}R_\gamma x) \quad \text{where} \quad R_\gamma \text{ is a rotation by } \gamma.$$

- Wavelet transform of f for all $\gamma \in \Gamma$ and $2^j < 2^J$

$$W_J f(x) = \begin{pmatrix} f \star \phi_J(x) \\ f \star \psi_{j,\gamma}(x) \end{pmatrix}_{j<J,\gamma\in\Gamma}$$

where $\phi_J(x)$ filters lower frequencies: $\int \phi_J(x)\,dx = 1$.

Wavelet Contraction/Unitary

Proposition: A wavelet transform is contractive

$$\|W_J f\|^2 = \int \Big(|f \star \phi_J(x)|^2 + \sum_{j<J, \gamma \in \Gamma} |f \star \psi_{j,\gamma}(x)|^2 \Big)\, dx \leq \|f\|^2$$

if and only if for almost all $\omega \in \mathbf{R}^d$

$$|\hat{\phi}_J(\omega)|^2 + \frac{1}{2} \sum_{j<J,\gamma} \Big(|\hat{\psi}_{j,\gamma}(\omega)|^2 + |\hat{\psi}_{j,\gamma}(-\omega)|^2 \Big) \leq 1$$

and unitary if it is an equality.

Wavelet Contraction/Unitary

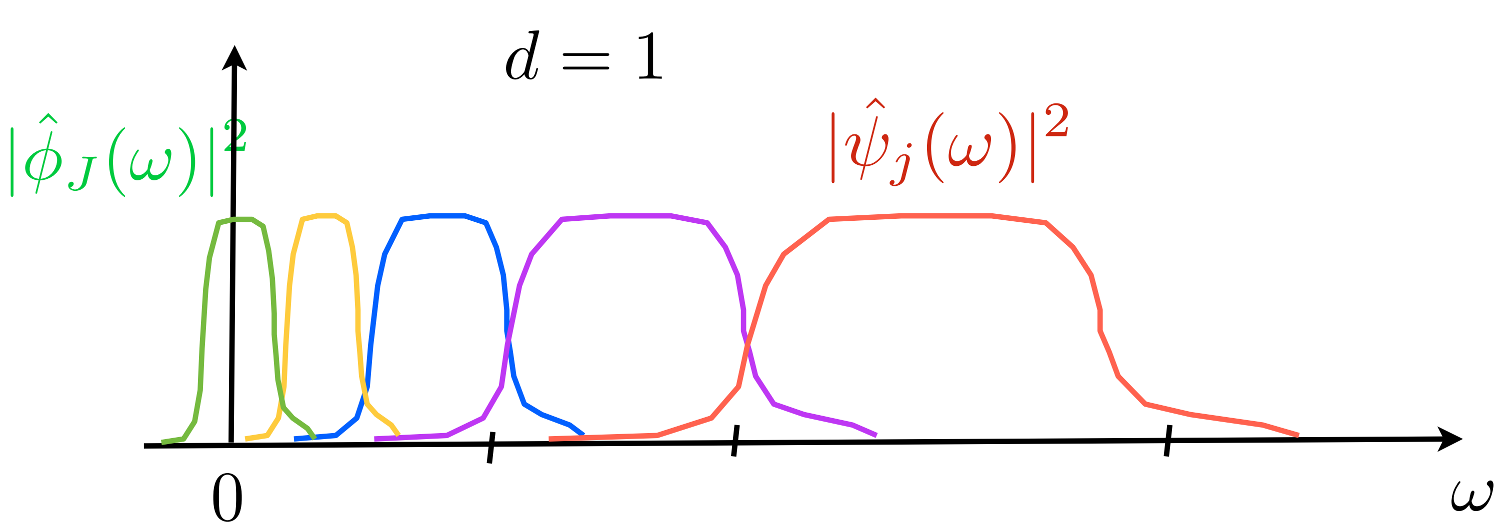
Proposition: A wavelet transform is contractive

$$\|W_J f\|^2 = \int \Big(|f \star \phi_J(x)|^2 + \sum_{j<J, \gamma \in \Gamma} |f \star \psi_{j,\gamma}(x)|^2 \Big)\, dx \leq \|f\|^2$$

if and only if for almost all $\omega \in \mathbf{R}^d$

$$|\hat{\phi}_J(\omega)|^2 + \frac{1}{2} \sum_{j<J,\gamma} \Big(|\hat{\psi}_{j,\gamma}(\omega)|^2 + |\hat{\psi}_{j,\gamma}(-\omega)|^2 \Big) \leq 1$$

and unitary if it is an equality.



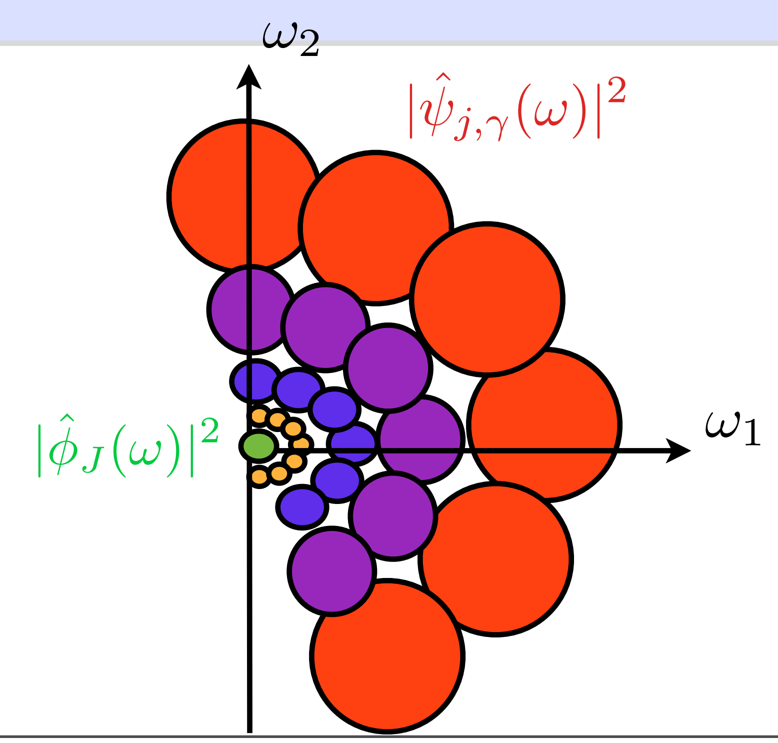


Image and Audio Descriptors

- How to build invariant descriptors from wavelet coefficients ?

- If f is translated then $f \star \psi_{j,\gamma}$ is translated

Image and Audio Descriptors

- How to build invariant descriptors from wavelet coefficients ?
- If f is translated then $f \star \psi_{j,\gamma}$ is translated
- $|f \star \psi_{j,\gamma}|$ is almost invariant to translations by $\tau \ll 2^j$.

Image and Audio Descriptors

- How to build invariant descriptors from wavelet coefficients ?

- If f is translated then $f \star \psi_{j,\gamma}$ is translated

- $|f \star \psi_{j,\gamma}|$ is almost invariant to translations by $\tau \ll 2^j$.

- $|f \star \psi_{j,\gamma}| \star \phi_J$ is almost invariant to translations by $\tau \ll 2^J$.

Image and Audio Descriptors

- How to build invariant descriptors from wavelet coefficients ?

- If f is translated then $f \star \psi_{j,\gamma}$ is translated

- $|f \star \psi_{j,\gamma}|$ is almost invariant to translations by $\tau \ll 2^j$.

- $|f \star \psi_{j,\gamma}| \star \phi_J$ is almost invariant to translations by $\tau \ll 2^J$.

- **Problem:** Important loss of information by averaging.

- Can we recover information that remains locally invariant ?

Scattering Operators

High frequencies are removed from $|f \star \psi_{j_1,\gamma_1}| \star \phi_J$.

Scattering Operators

High frequencies are removed from $|f \star \psi_{j_1,\gamma_1}| \star \phi_J$.

Recovered with fine scale wavelet coefficients:

$$|f \star \psi_{j_1,\gamma_1}| \star \psi_{j_2,\gamma_2} \quad \text{for} \quad 2^{j_2} < 2^J \ .$$

Local invariance by removing the phase and averaging:

$$||f \star \psi_{j_1,\gamma_1}| \star \psi_{j_2,\gamma_2}| \star \phi_J \ .$$

Co-occurrence at scales 2^{j_1}, 2^{j_2} and directions γ_1, γ_2.

Scattering Operators

High frequencies are removed from $|f \star \psi_{j_1,\gamma_1}| \star \phi_J$.

Recovered with fine scale wavelet coefficients:

$$|f \star \psi_{j_1,\gamma_1}| \star \psi_{j_2,\gamma_2} \quad \text{for} \quad 2^{j_2} < 2^J \ .$$

Local invariance by removing the phase and averaging:

$$||f \star \psi_{j_1,\gamma_1}| \star \psi_{j_2,\gamma_2}| \star \phi_J \ .$$

Co-occurrence at scales 2^{j_1}, 2^{j_2} and directions γ_1, γ_2.

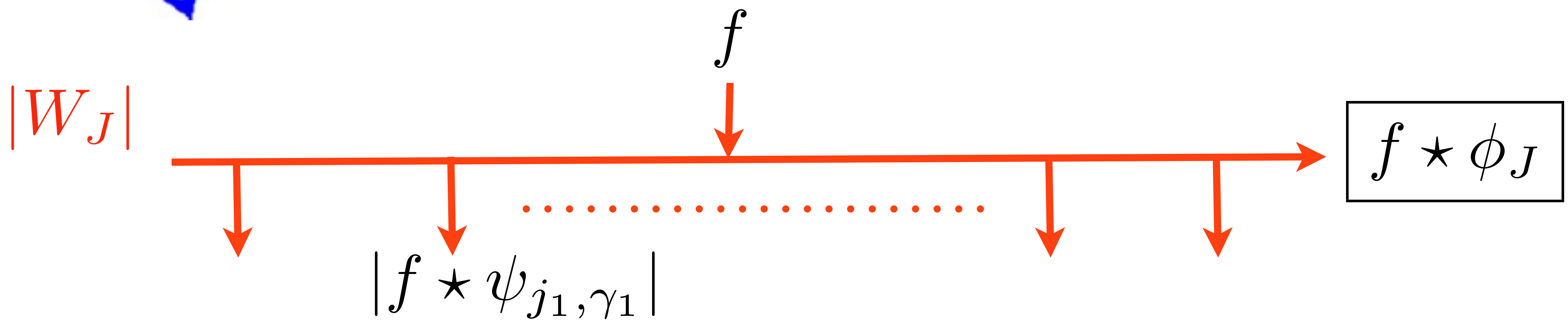
Lost high frequencies recovered with wavelets coefficients...

Scattering Cascade

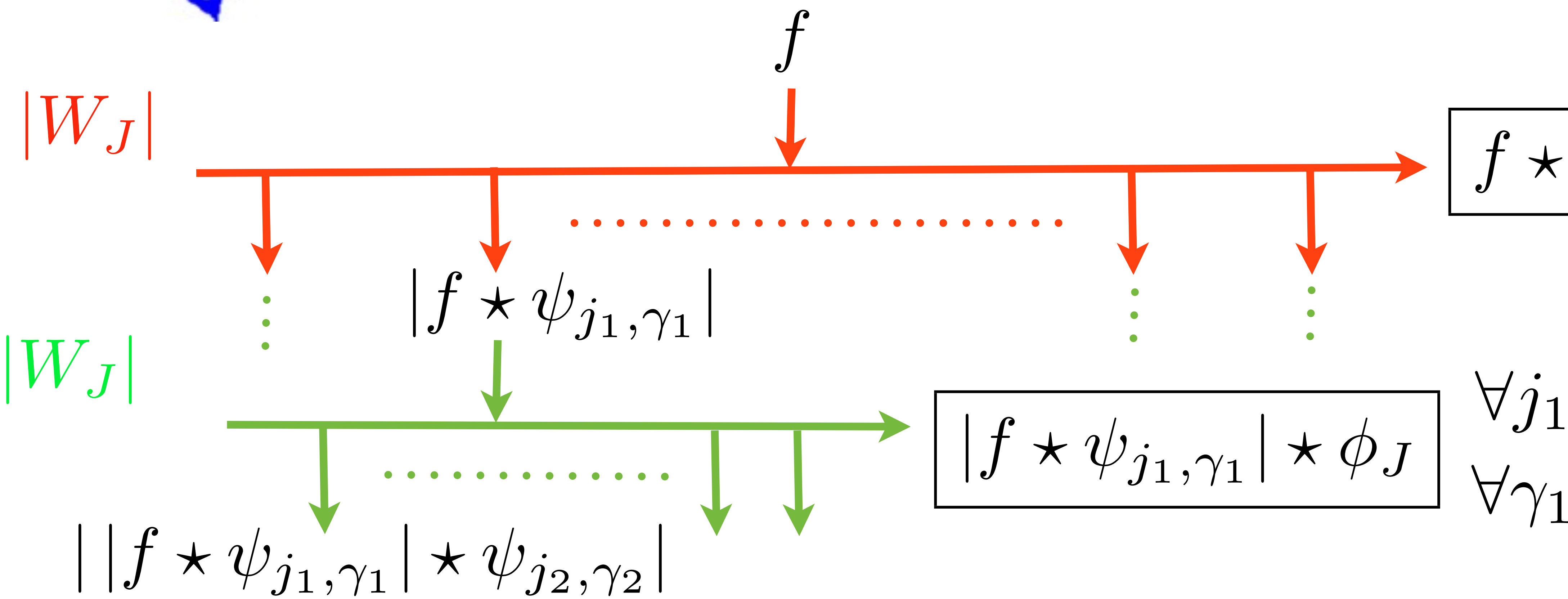
Scattering Cascade

f

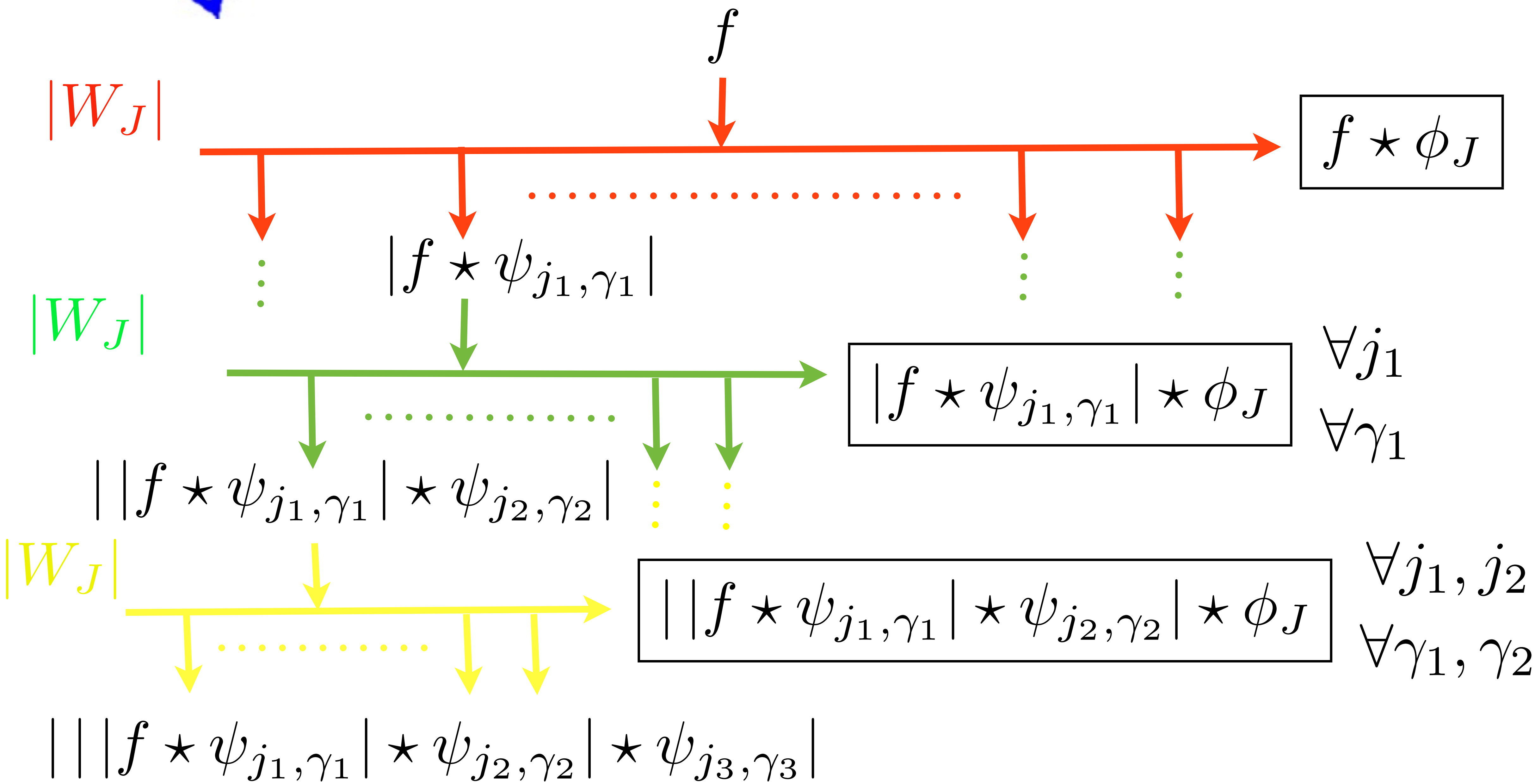
Scattering Cascade



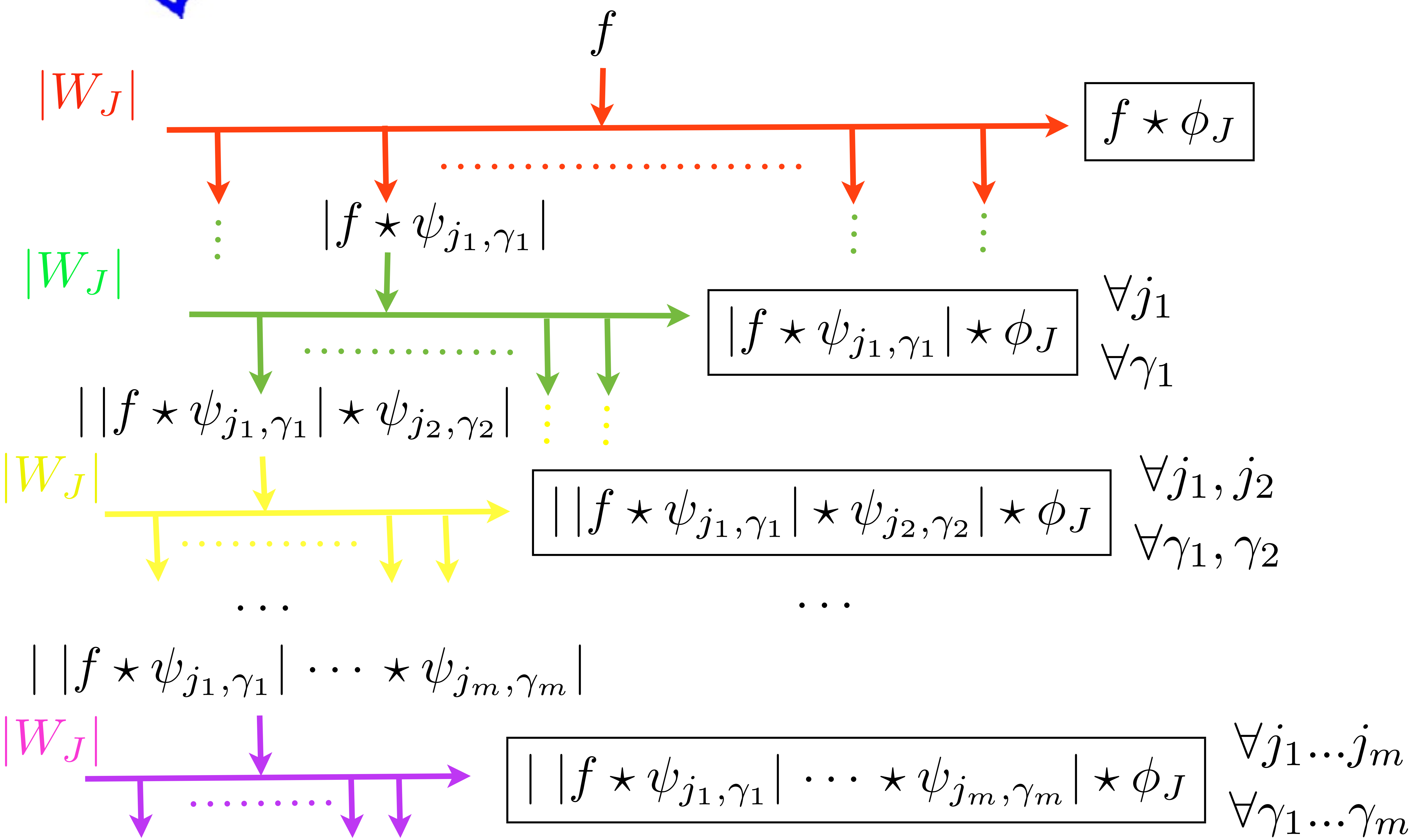
Scattering Cascade



Scattering Cascade



Scattering Cascade



Cascade of contractive wavelet and modulus operators.

Scattering Representation

$$S_J f(x) = \begin{pmatrix} f \star \phi_J(x) \\ |f \star \psi_{j_1,\gamma_1}| \star \phi_J(x) \\ ||f \star \psi_{j_1,\gamma_1}| \star \psi_{j_2,\gamma_2}| \star \phi_J(x) \\ \cdots \\ |\,|f \star \psi_{j_1,\gamma_1}| \cdots \star \psi_{j_m,\gamma_m}| \star \phi_J(x) \end{pmatrix}_{\substack{\forall j_1 \ldots j_m \\ \forall \gamma_1 \ldots \gamma_m}}$$

Scattering Representation

$$S_J f(x) = \begin{pmatrix} f \star \phi_J(x) \\ |f \star \psi_{j_1,\gamma_1}| \star \phi_J(x) \\ ||f \star \psi_{j_1,\gamma_1}| \star \psi_{j_2,\gamma_2}| \star \phi_J(x) \\ \cdots \\ |\,|f \star \psi_{j_1,\gamma_1}| \cdots \star \psi_{j_m,\gamma_m}| \star \phi_J(x) \end{pmatrix}_{\substack{\forall j_1 \ldots j_m \\ \forall \gamma_1 \ldots \gamma_m}}$$

Euclidean norm: $|S_J f(x)|^2$

Scattering norm: $\|S_J f\|^2 = \int |S_J f(x)|^2 \, dx$

Scattering Representation

$$S_J f(x) = \begin{pmatrix} f \star \phi_J(x) \\ |f \star \psi_{j_1,\gamma_1}| \star \phi_J(x) \\ ||f \star \psi_{j_1,\gamma_1}| \star \psi_{j_2,\gamma_2}| \star \phi_J(x) \\ \dots \\ |\,|f \star \psi_{j_1,\gamma_1}| \cdots \star \psi_{j_m,\gamma_m}| \star \phi_J(x) \end{pmatrix}_{\substack{\forall j_1 \dots j_m \\ \forall \gamma_1 \dots \gamma_m}}$$

Euclidean norm: $|S_J f(x)|^2$

Scattering norm: $\|S_J f\|^2 = \int |S_J f(x)|^2 \, dx$

Contractive because cascade of contractive operators:

$$\|S_J f - S_J g\| \leq \|f - g\|.$$

Scattering Energy Conservation

Theorem: For appropriate complex wavelets

$$\lim_{m\to\infty} \sum_{\substack{(j_1\cdots j_m)\in\mathbf{Z}^m \\ (\gamma_1\cdots\gamma_m)\in\Gamma^m}} \| \, | \, | f \star \psi_{j_1,\gamma_1} | \cdots | \star \psi_{j_m,\gamma_m} | \, \|^2 = 0$$

and $$\|S_J f\|^2 = \|f\|^2 \ .$$

Computational Complexity

- If $f(n)$ is of size N

Compute only $S_J f(2^J n)$: $2^{-2J} N$ scattering vectors.

For K directional wavelets and $2^{2J} = N$:

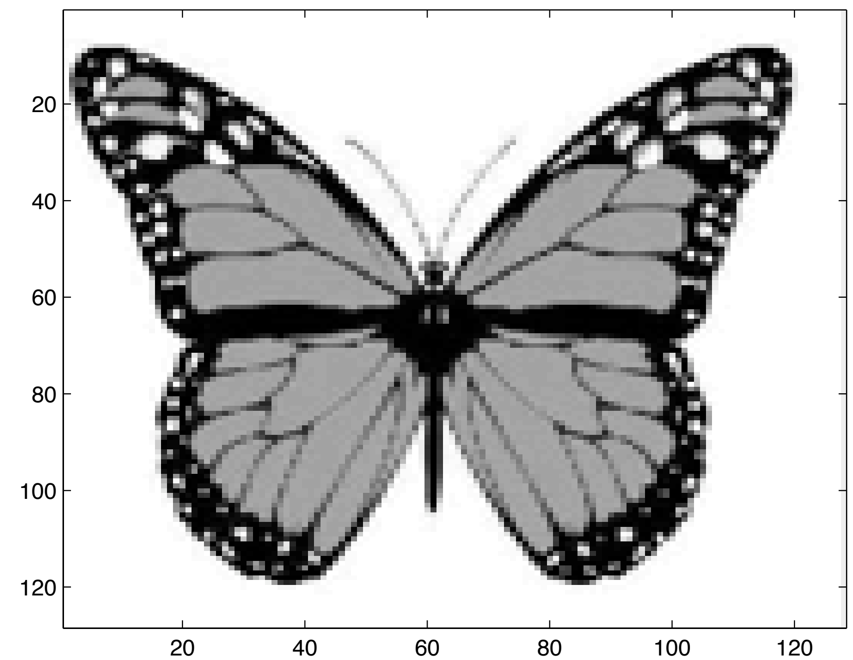
$$O(K^m (\log N)^m) \text{ scattering coefficients}$$

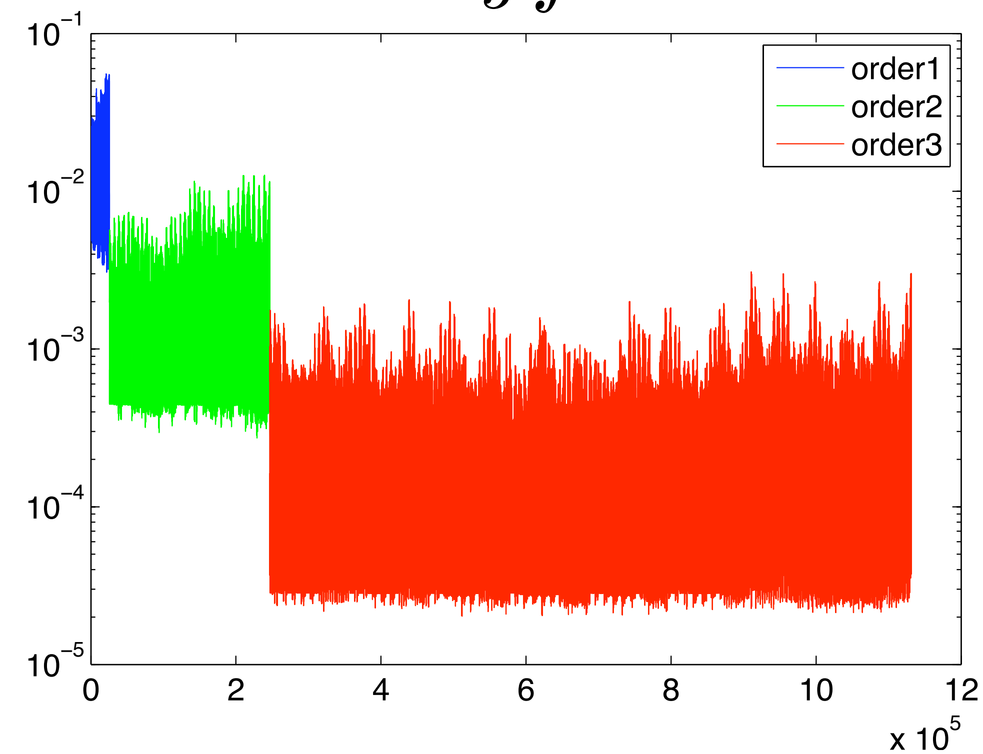
computed with $O(N \log N)$ operations.

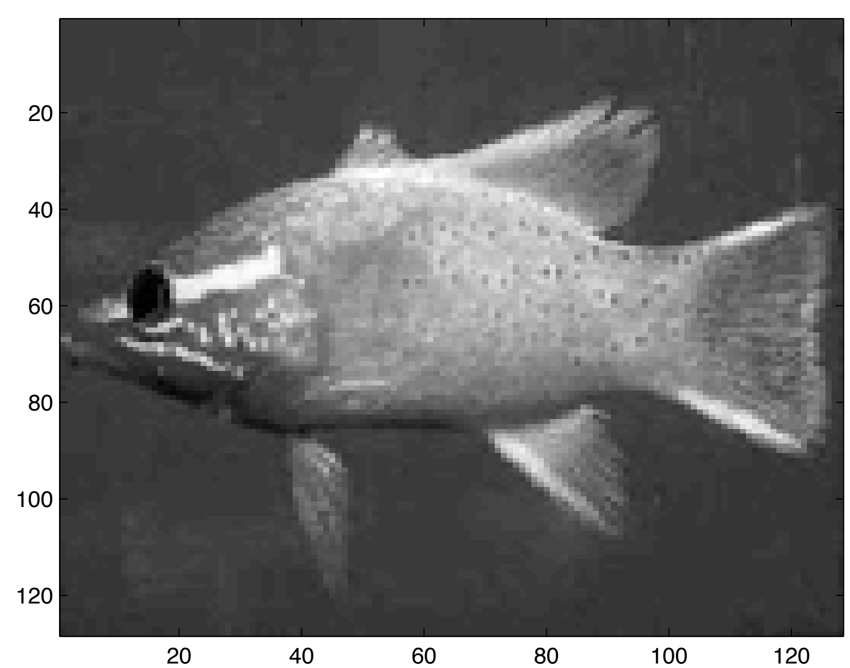
For images: $K = 6$, $m \leq 3$.

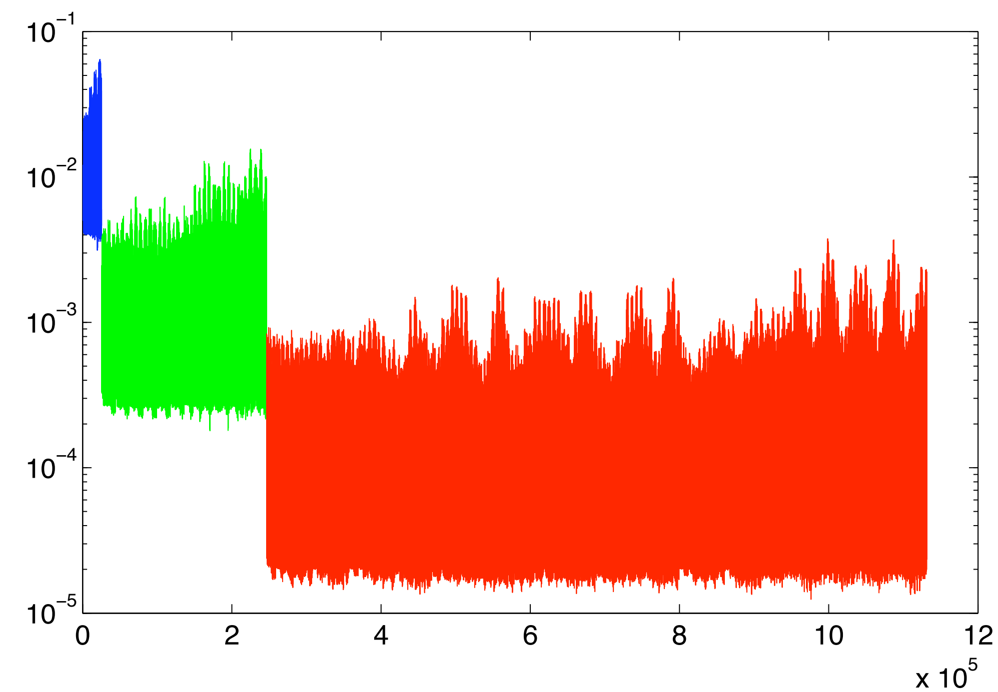
Scattering Examples

f $\qquad\qquad$ $S_J f$









Translation Invariance

- When 2^J goes to ∞ coefficients converge to $\mathbf{L^1}$ norms:

$$\lim_{J\to\infty} 2^{dJ} \,|\,|f \star \psi_{j_1,\gamma_1}|...\star \psi_{j_m,\gamma_m}| \star \phi_J(x) = \int |\,|f \star \psi_{j_1,\gamma_1}|...\star \psi_{j_m,\gamma_m}(u)|\, du.$$

Theorem: $\lim_{J\to\infty} \|S_J f - S_J g\|$ converges and

if $D_\tau f(x) = f(x-\tau)$ is a translation then

$$\lim_{J\to\infty} \|S_J f - S_J (D_\tau f)\| = 0\ .$$

Continuity to Deformations

Theorem If $D_\tau f(x) = f(x - \tau(x))$ with $\|\nabla\tau\|_\infty < 1$

then for $J > \log \frac{\|\tau\|_\infty}{\|\nabla\tau\|_\infty}$

$$\|S_J f - S_J(D_\tau f)\| \leq C\, m\, \|f\| \,\log\Big(\frac{\|\tau\|_\infty}{\|\nabla\tau\|_\infty}\Big)\, \|\nabla\tau\|_\infty$$

Proof: $\|S_J - S_J D_\tau\| \leq \|D_\tau S_J - S_J\| + \|[S_J\,,\, D_\tau]\|$

$$\|[S_J\,,\, D_\tau]\| \leq C\, m\, \|[W_J\,,\, D_\tau]\|$$

$$\|[W_J\,,\, D_\tau]\| \leq C\, \log\left(\frac{\|\tau\|_\infty}{\|\nabla\tau\|_\infty}\right)\, \|\nabla\tau\|_\infty$$

Scattering Stationary Processes

If $F(x)$ is a stationary process then

$$E\{S_J F(x)\} = \begin{pmatrix} E\{F\} \\ E\{|F \star \psi_{j_1,\gamma_1}|\} \\ \cdots \\ E\{|\,|F \star \psi_{j_1,\gamma_1}| \cdots \star \psi_{j_m,\gamma_m}|\} \end{pmatrix}_{\substack{\forall j_1 \ldots j_m \\ \forall \gamma_1 \ldots \gamma_m}}$$

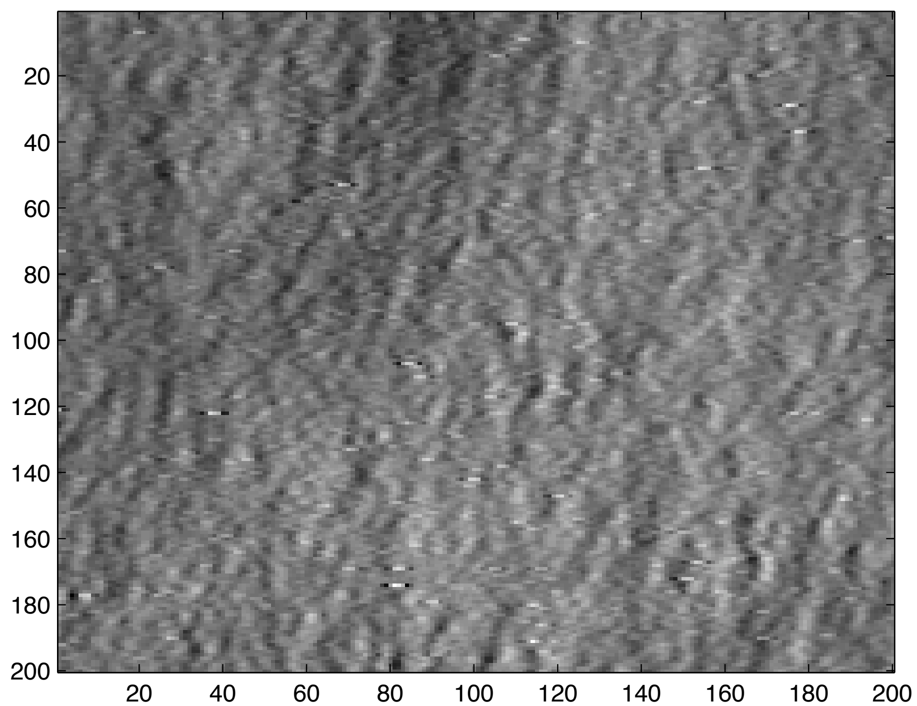
does not depend upon J and x.

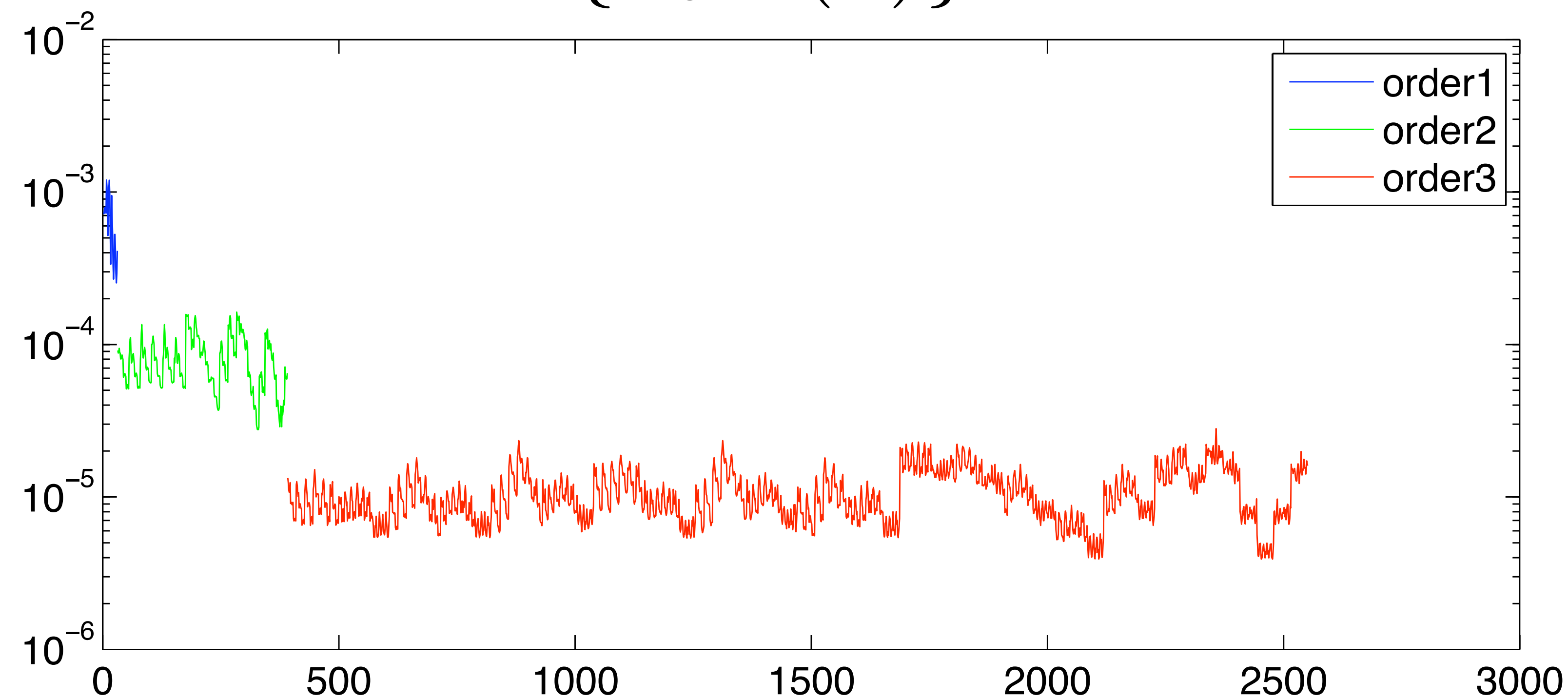
Theorem: For appropriate complex wavelets

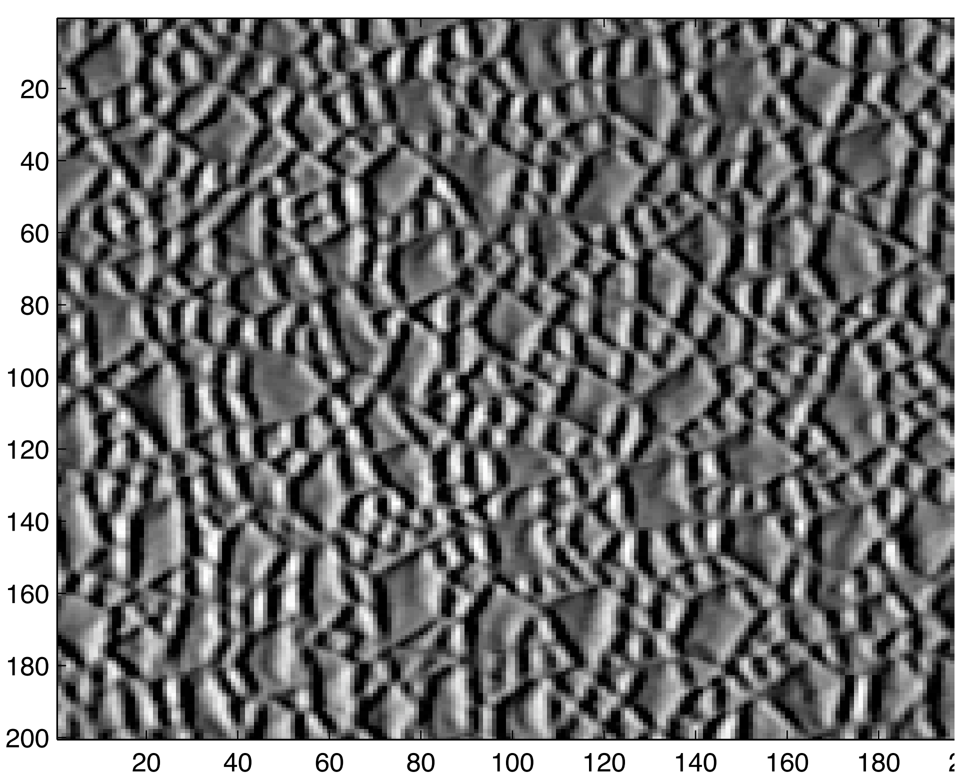
$$E\{|S_J F(x)|^2\} = E\{|F(x)|^2\} \; .$$

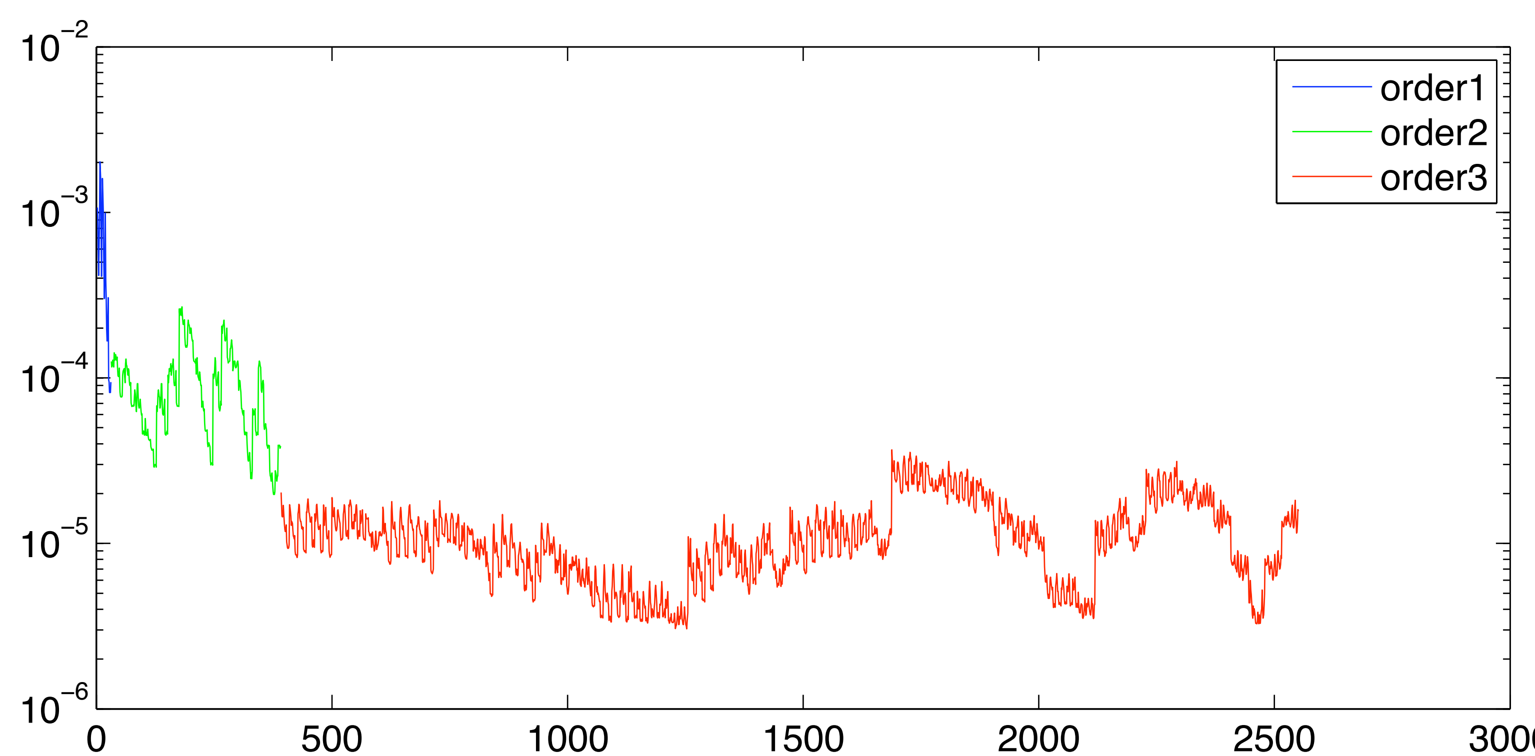
Scattering of Stationary Processes

F | $E\{S_J F(x)\}$





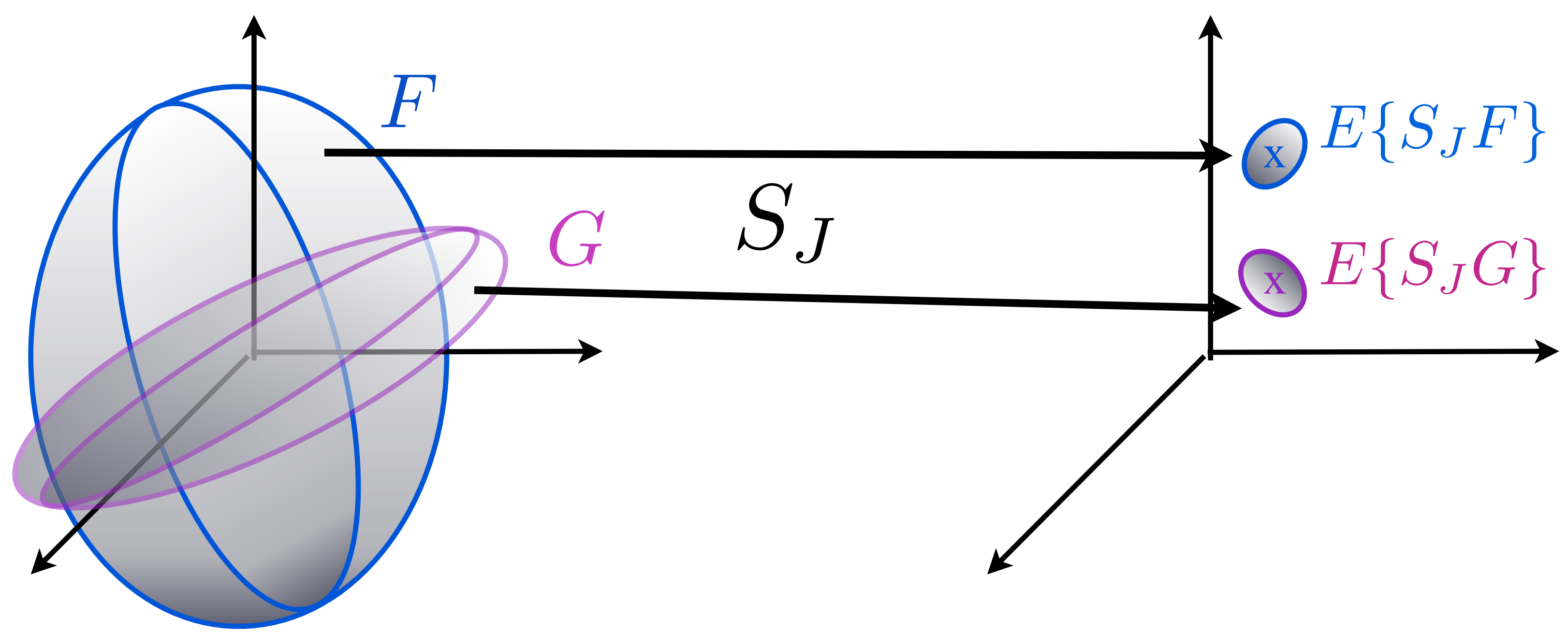




Scattering Stationary Processes

Conjecture: for a wide class of "ergodic" stationary processes

$$\lim_{J\to\infty} E\{|S_J F - E\{S_J\, F\}|^2\} = 0 \quad \text{exponentialy}$$

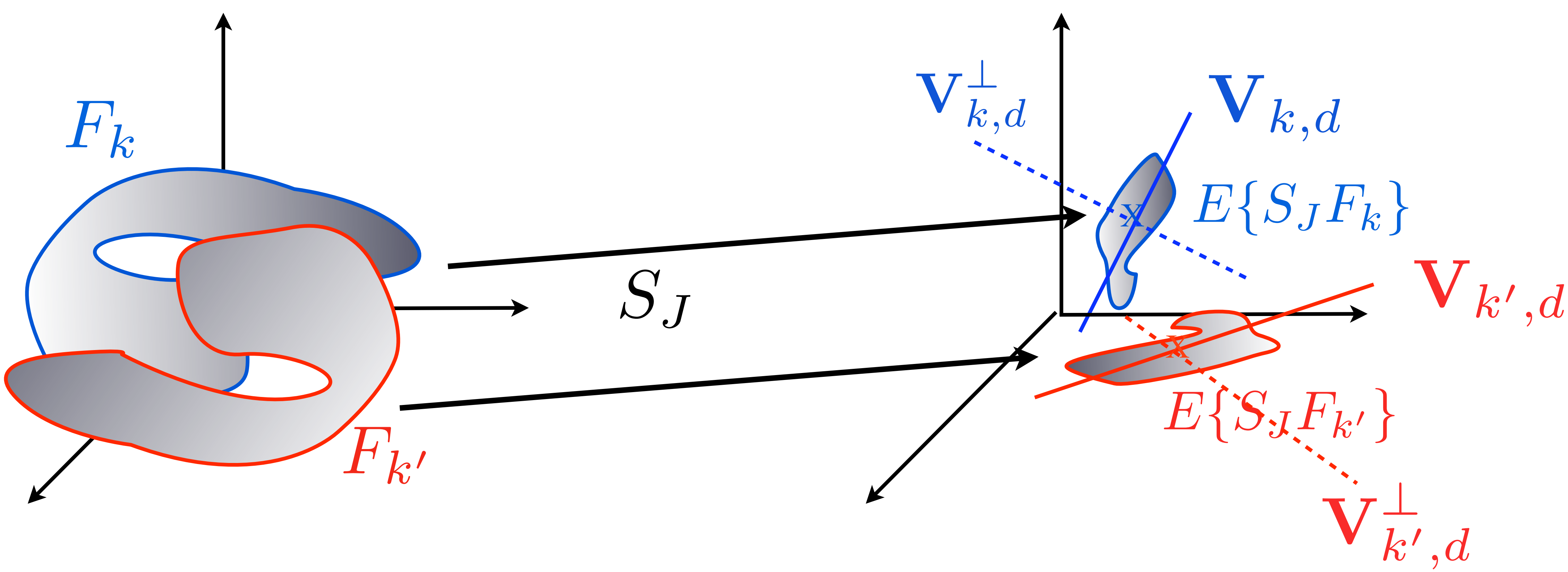


Theorem: for $J > \log \frac{\|\tau\|_\infty}{\|\nabla\tau\|_\infty}$

$$E\{|S_J F - S_J(D_\tau F)|^2\} \leq C\, m^2\, E\{|F|^2\}\, E\{\log\Big(\frac{\|\tau\|_\infty}{\|\nabla\tau\|_\infty}\Big)^2\; \|\nabla\tau\|_\infty^2\}$$

Classification : *Joan Bruna*

- K classes corresponding to K (non stationary) processes $\{F_k\}_{k \leq K}$
- Each class is represented by the centroid $E\{S_J F_k\}$
- Intra class variance reduction by eliminating the space $\mathbf{V}_{d,k}$ of the d principal variance components (main deformations):



Scattering PCA Classification

- **PCA** calculation of the d dimensional spaces $\mathbf{V}_{k,d}$ of maximum variability of $S_J F_k - E\{S_J F_k\}$ from training samples of F_k

- **Classification** of a signal f :

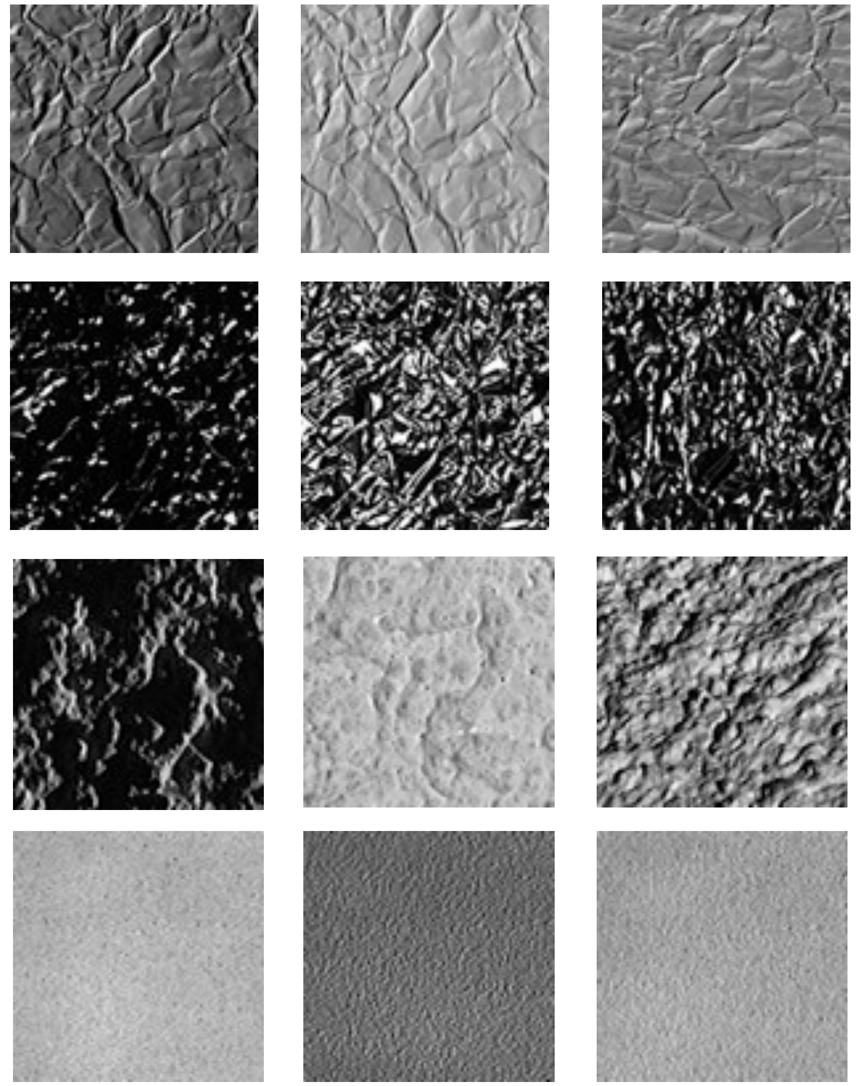
$$k(f) = \arg \min_{1 \leq k \leq K} \|P_{\mathbf{V}_{k,d}^{\perp}}(S_J f - E\{S_J F_k\})\|^2$$

- **Cross-validation:**
  - d : dimension of the variability reduction.
  - J : maximum scattering scale.

- Class per class classification models: not discriminative.

Classification of Textures

61 classes



Rotations and illumination variations.

Scattering $J = \log_2 N$

Training per class	PCA $m = 2$	SVM $m = 2$	LBP	Mark. Rand. 8
23	**0.9**%		18.23%	22.43%
46	**0.09**%	1.7%	3.96%	2.46%

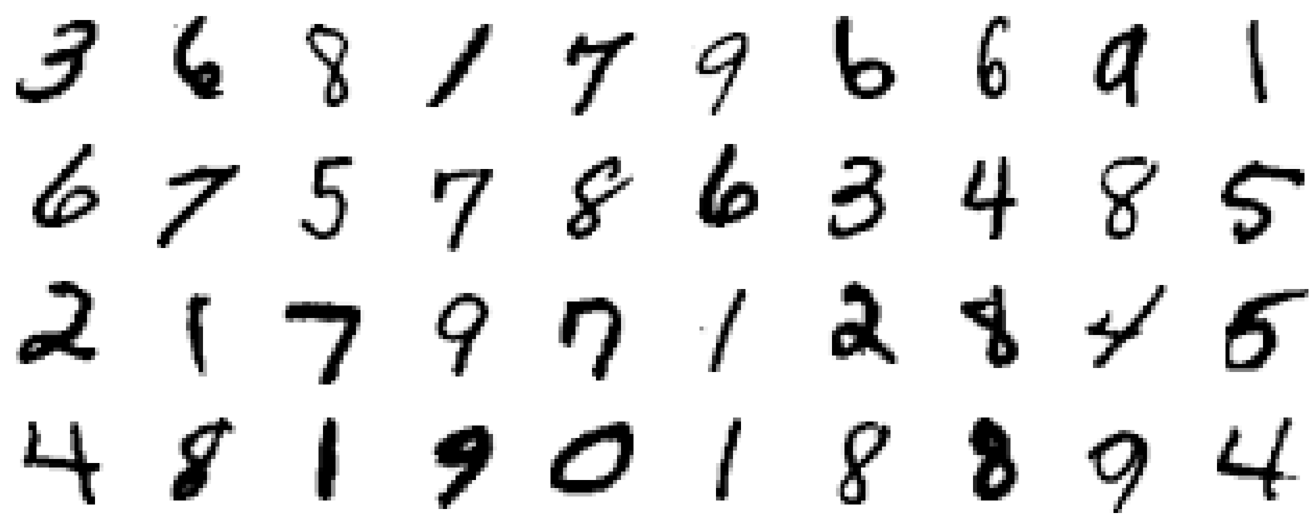
Non-Gaussian Process Characterization

- Usual approaches use high order moments: bad estimators. Does not work for image textures.

- Can characterize non-gaussian processes with first and second order moments of scattering vectors.

- Scattering estimation of multifractal properties without moments *(Bacry, Duvernet)*

Digit Classification: MNIST



Scattering with $J = 3$

Training	Conv. Net.	PCA	SVM
		$m = 2$	$m = 2$
300	7.18	**6.05**	21.5
1000	3.21	**2.39**	3.06
2000	2.53	**1.71**	1.87
5000	1.52	**1.22**	1.54
10000	**0.85**	1.17	1.15
20000	**0.76**	1.4	0.96
40000	**0.65**	0.78	0.85
60000	**0.53**	0.77	0.7

Combined Scattering

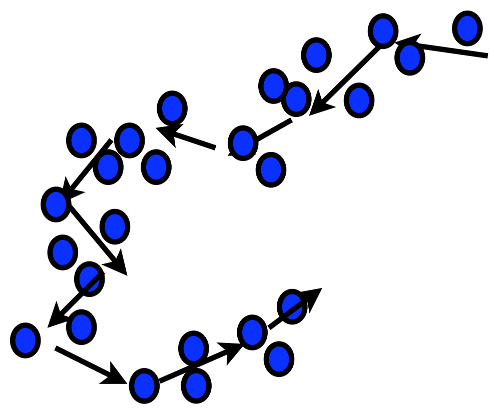
- Translation invariance is not sufficient for complex classes.

- Non-linear class variability need to be further reduced:

$$f \longrightarrow \boxed{S_J^{\text{Trans}}} \longrightarrow \boxed{S_{J'}^G} \longrightarrow \ldots$$

- Scattering $S_{J'}^G$ over a compact Lie group G with iterated wavelet tranforms over $\mathbf{L^2}(G)$ (instead of $\mathbf{L^2}(\mathbf{R^2})$) and cascaded with modulus operators.

Lie Algebra Learning



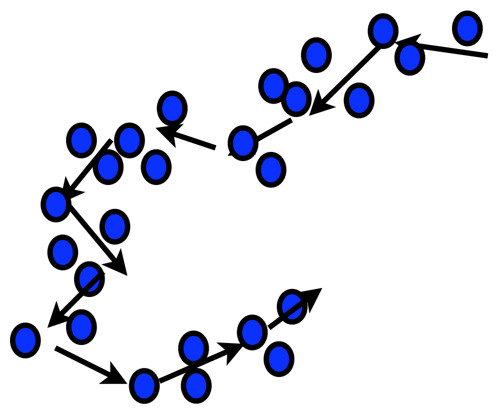
- Non-linear space contraction along the Lie Algebra to reduce variability within classes.

- Estimating a Lie Algebra with a Haar wavelet basis:

- Scattering by cascading the Haar wavelet transform.

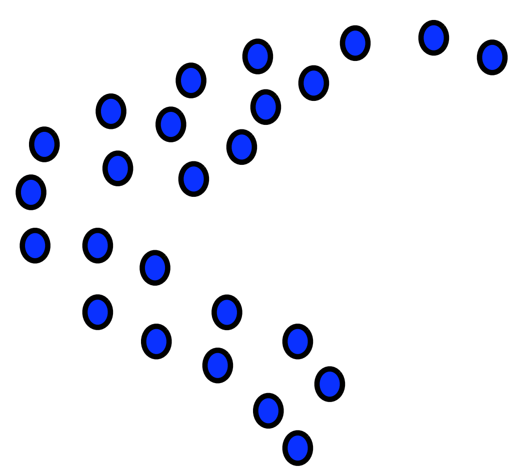
Lie Algebra Learning

- Find the local directions of variability in classes: a Lie Algebra, no manifold model.



- Non-linear space contraction along the Lie Algebra to reduce variability within classes.

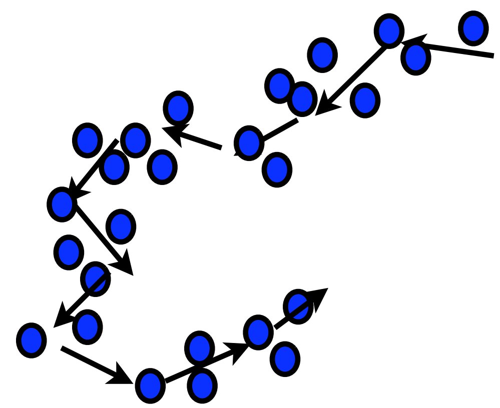
- Estimating a Lie Algebra with a Haar wavelet basis:



- Scattering by cascading the Haar wavelet transform.

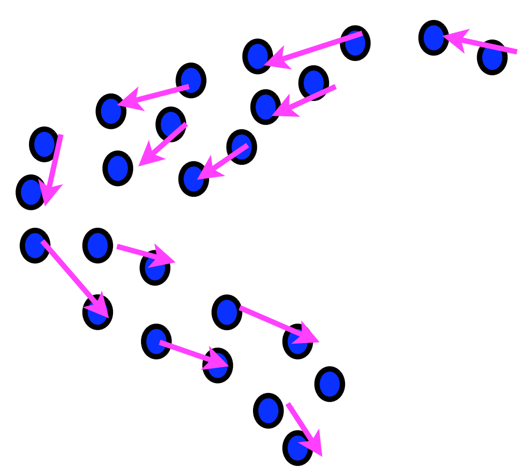
Lie Algebra Learning

- Find the local directions of variability in classes: a Lie Algebra, no manifold model.



- Non-linear space contraction along the Lie Algebra to reduce variability within classes.

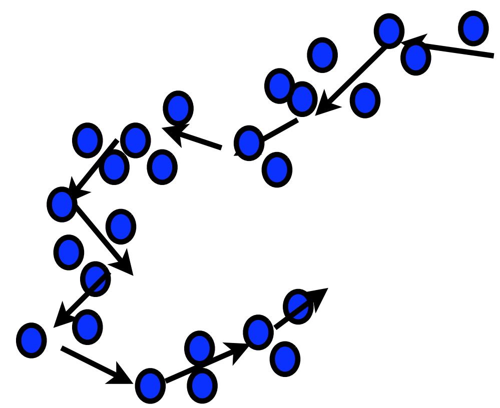
- Estimating a Lie Algebra with a Haar wavelet basis:



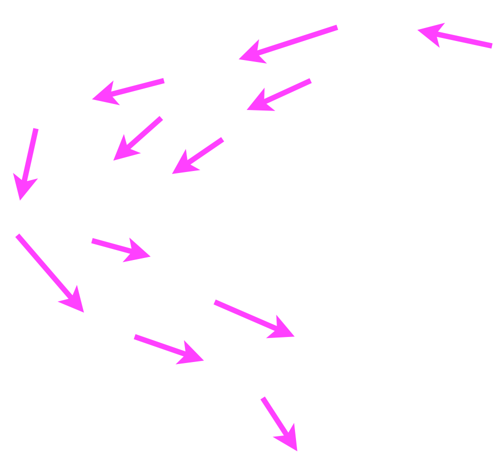
- Scattering by cascading the Haar wavelet transform.

Lie Algebra Learning

- Find the local directions of variability in classes: a Lie Algebra, no manifold model.



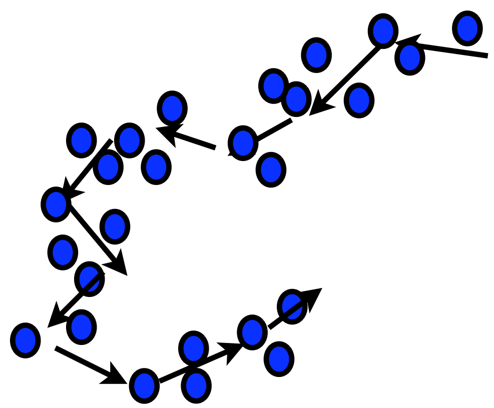
- Non-linear space contraction along the Lie Algebra to reduce variability within classes.
- Estimating a Lie Algebra with a Haar wavelet basis:



- Scattering by cascading the Haar wavelet transform.

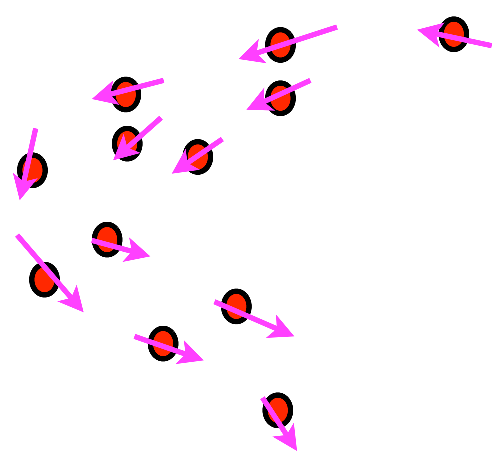
Lie Algebra Learning

- Find the local directions of variability in classes: a Lie Algebra, no manifold model.



- Non-linear space contraction along the Lie Algebra to reduce variability within classes.

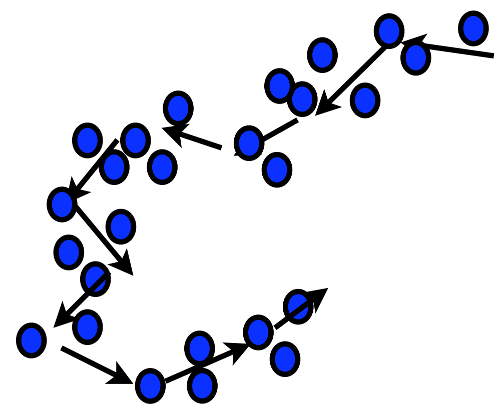
- Estimating a Lie Algebra with a Haar wavelet basis:



- Scattering by cascading the Haar wavelet transform.

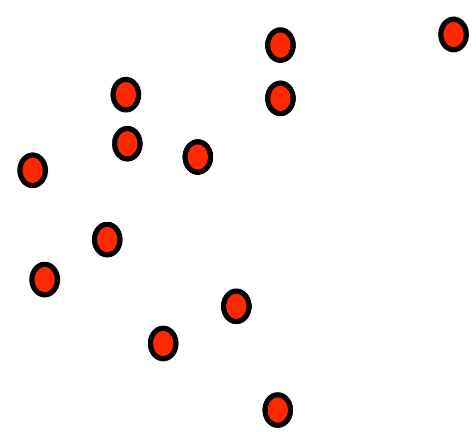
Lie Algebra Learning

- Find the local directions of variability in classes: a Lie Algebra, no manifold model.



- Non-linear space contraction along the Lie Algebra to reduce variability within classes.

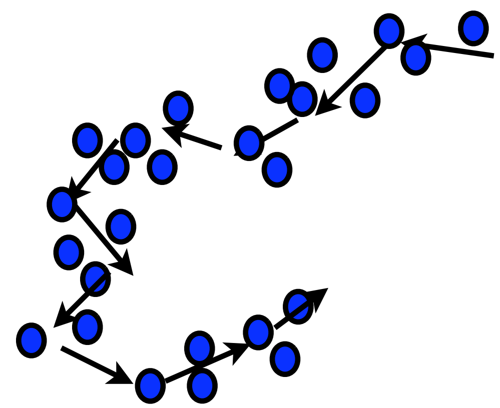
- Estimating a Lie Algebra with a Haar wavelet basis:



- Scattering by cascading the Haar wavelet transform.

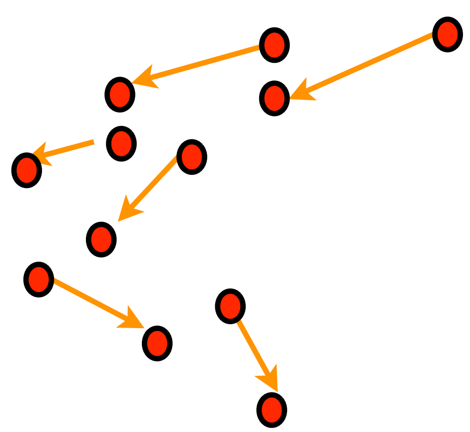
Lie Algebra Learning

- Find the local directions of variability in classes: a Lie Algebra, no manifold model.



- Non-linear space contraction along the Lie Algebra to reduce variability within classes.

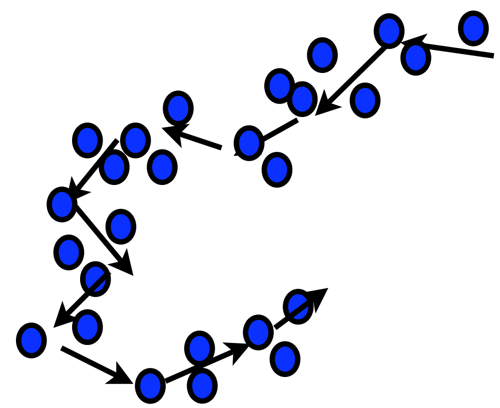
- Estimating a Lie Algebra with a Haar wavelet basis:



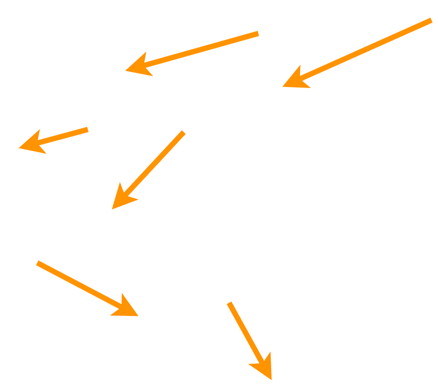
- Scattering by cascading the Haar wavelet transform.

Lie Algebra Learning

- Find the local directions of variability in classes: a Lie Algebra, no manifold model.



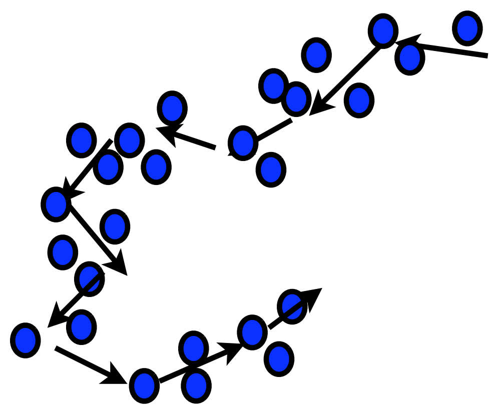
- Non-linear space contraction along the Lie Algebra to reduce variability within classes.
- Estimating a Lie Algebra with a Haar wavelet basis:



- Scattering by cascading the Haar wavelet transform.

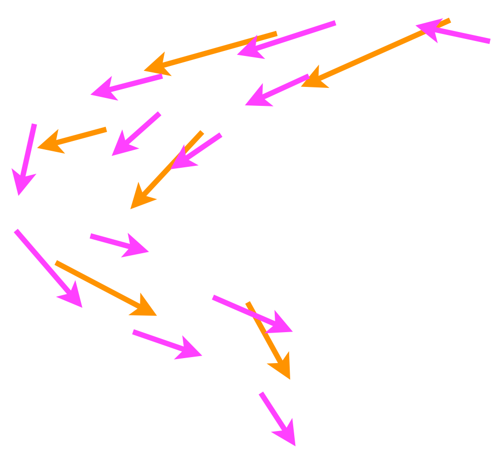
Lie Algebra Learning

- Find the local directions of variability in classes: a Lie Algebra, no manifold model.



- Non-linear space contraction along the Lie Algebra to reduce variability within classes.

- Estimating a Lie Algebra with a Haar wavelet basis:



- Scattering by cascading the Haar wavelet transform.

Conclusion

- Multiscale scattering contractions yield invariants that are Lipschitz continuous to deformations.
- Iterative scattering contractions is effective for high dimensional non-discriminative classification.
- New representation of stationary processes to explore.
- Iterative filter bank contractions seem to exist in audio cortex.
- An approach to understand some biological architectures ?
- **Papers/softwares:** www.cmap.polytechnique.fr/scattering