Stat 212b: Topics in Deep Learning
Lecture 5

Joan Bruna
UC Berkeley

Berkeley
UNIVERSITY OF CALIFORNIA

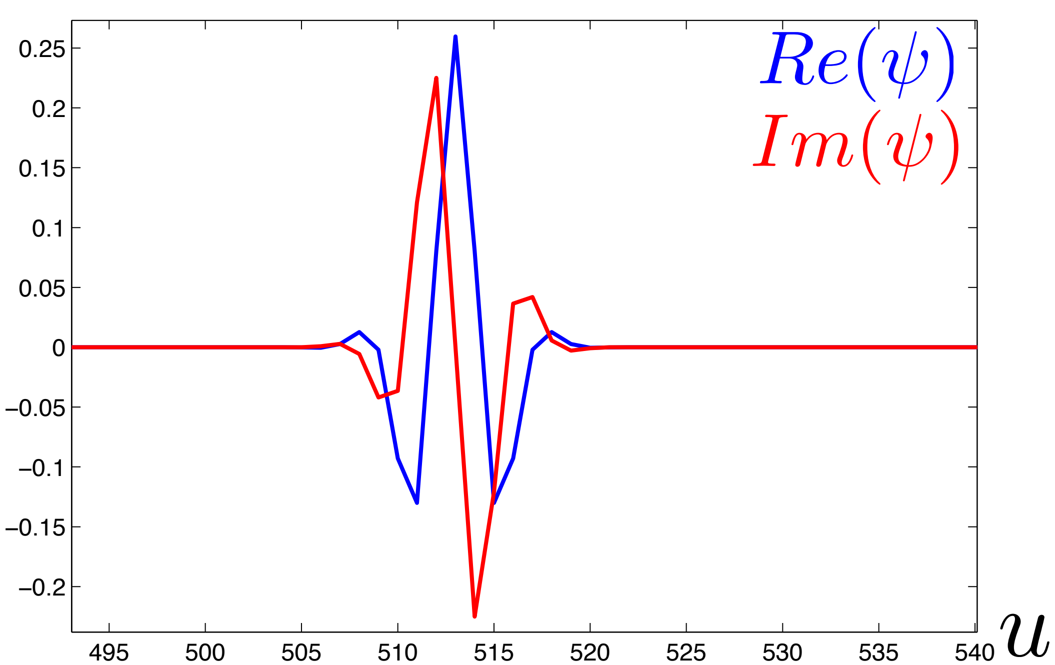
Review: Wavelets

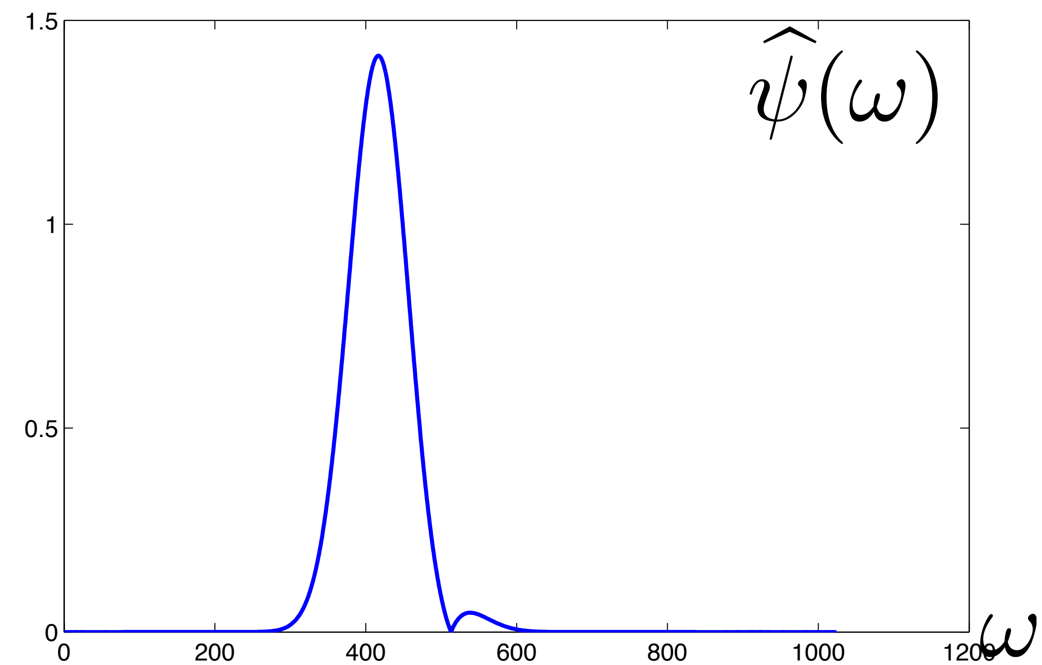
- ψ : bandpass (ie oscillating) signal, well localized in space and frequency.
- At least one vanishing moment: $\int \psi(u)du = 0$

(we say that ψ has k vanishing moments if $\int \psi(u)u^l du = 0$ for $l < k$)

- Can be real or complex. $\psi = \psi_r + i\psi_i$

Ex: Morlet wavelet

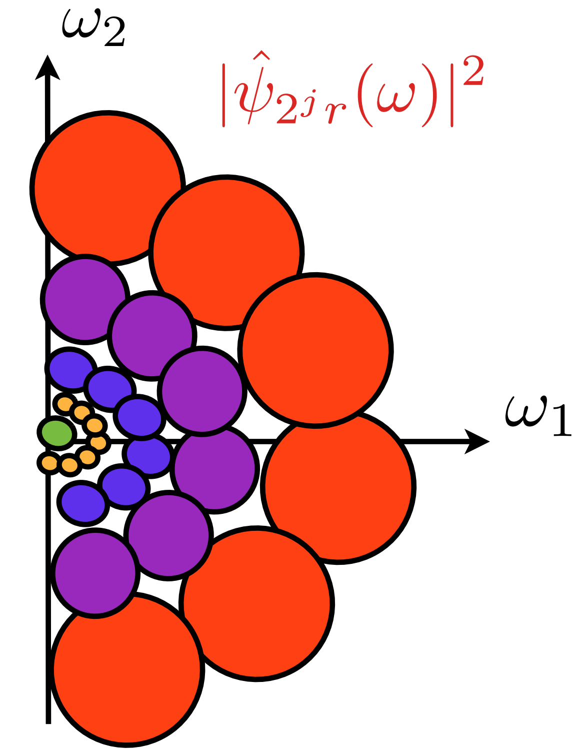




Review: Littlewood-Paley Wavelet Filter Banks

- For images, dilated and rotated wavelets:

$$\psi_\lambda(u) = 2^{-j/2}\psi(2^{-j}ru) \ , \ \text{with } \lambda = 2^j r$$



- Wavelet transform convolutional filter bank:

$$Wx = \{x \star \phi(u) \, , \, x \star \psi_\lambda(u)\}_{\lambda \in \Lambda}$$

$$x \star \psi(u) = \int x(v)\psi(u-v)dv \ .$$

Theorem (Littlewood-Paley): If there exists $\delta > 0$ such that

$$\forall \omega > 0 \ , \ 1 - \delta \leq |\hat{\phi}(\omega)|^2 + \frac{1}{2}\sum_\lambda |\hat{\psi}(\lambda^{-1}\omega)|^2 \leq 1 \ ,$$

then $\forall\, x \in L^2$, $(1-\delta)\|x\|^2 \leq \|Wx\|^2 \leq \|x\|^2$.

Review: Wavelets and Deformations

- We saw before that a blurring kernel is nearly invariant to deformations:

Proposition: The local averaging $\Phi(x) = x * \phi_J$ satisfies $\forall \ \|x\| = 1 \in L^2 \ , \ \tau \ , \ \|\Phi(x) - \Phi(\varphi_\tau x)\| \leq C \|\tau\| \ .$

- What about the wavelet operator $\Phi(x) = \{x * \psi_\lambda\}_\lambda$?
  - We don't have local invariance, but we have a form of local covariance:

Proposition [Mallat]: For each $\delta > 0$ there exists $C > 0$ such that for all J and all $\tau \in C^2$ with $\|\nabla \tau\|_\infty \leq 1 - \delta$ we have

$$\|W_J \varphi_\tau - \varphi_\tau W_J\| \leq C(J \|\nabla \tau\|_\infty + \|H\tau\|_\infty) \ .$$

($H\tau$: Hessian of τ)

Review: Characterization of stable non-linearities

- Preserve additive stability:

$$\|Mx - Mx'\| \leq \|x - x'\| \ . \qquad M \text{ non-expansive} \ .$$

- Preserve geometric stability: It is sufficient to commute with diffeomorphisms.

Theorem: If M is non-expansive operator in L^2 such that $\varphi_\tau M = M \varphi_\tau$ for all τ, then M is point-wise:

$$Mx(u) = \rho(x(u)) \ .$$

- Since we want to smooth orbits, we may choose a point-wise nonlinearity that reduces oscillations:

$$\rho(z) = |z| \text{ or } \rho(z) = \max(0, z)$$

Objectives

- Scattering Representations
  - Main Properties
  - Main Limitations
  - Extensions: Joint rigid scattering.
- Convolutional Neural Networks
  - From fixed groups to adaptive templates

Separable Scattering Operators

- Local averaging kernel: $x \star \phi_J$
  - locally translation invariant
  - stable to additive and geometric deformations
  - loss of high-frequency information.

Separable Scattering Operators

- Local averaging kernel: $x \star \phi_J$
  - locally translation invariant
  - stable to additive and geometric deformations
  - loss of high-frequency information.
- Recover lost information: $\mathcal{U}_J(x) = \{x \star \phi_J \,,\, |x \star \psi_\lambda|\}_{\lambda \in \Lambda_J}$.
  - Point-wise, non-expansive non-linearities: maintain stability.
  - Complex modulus maps energy towards low-frequencies.

Separable Scattering Operators

- Local averaging kernel: $x \star \phi_J$
  - locally translation invariant
  - stable to additive and geometric deformations
  - loss of high-frequency information.
- Recover lost information: $\mathcal{U}_J(x) = \{x \star \phi_J \, , \, |x \star \psi_\lambda|\}_{\lambda \in \Lambda_J} \, .$
  - Point-wise, non-expansive non-linearities: maintain stability.
  - Complex modulus maps energy towards low-frequencies.
- Cascade the "recovery" operator:

$$\mathcal{U}_J^2(x) = \{x \star \phi_J \, , \, |x \star \psi_\lambda| \star \phi_J , \, ||x \star \psi_\lambda| \star \psi_{\lambda'}|\}_{\lambda, \lambda' \in \Lambda_J} \, .$$

Separable Scattering Operators

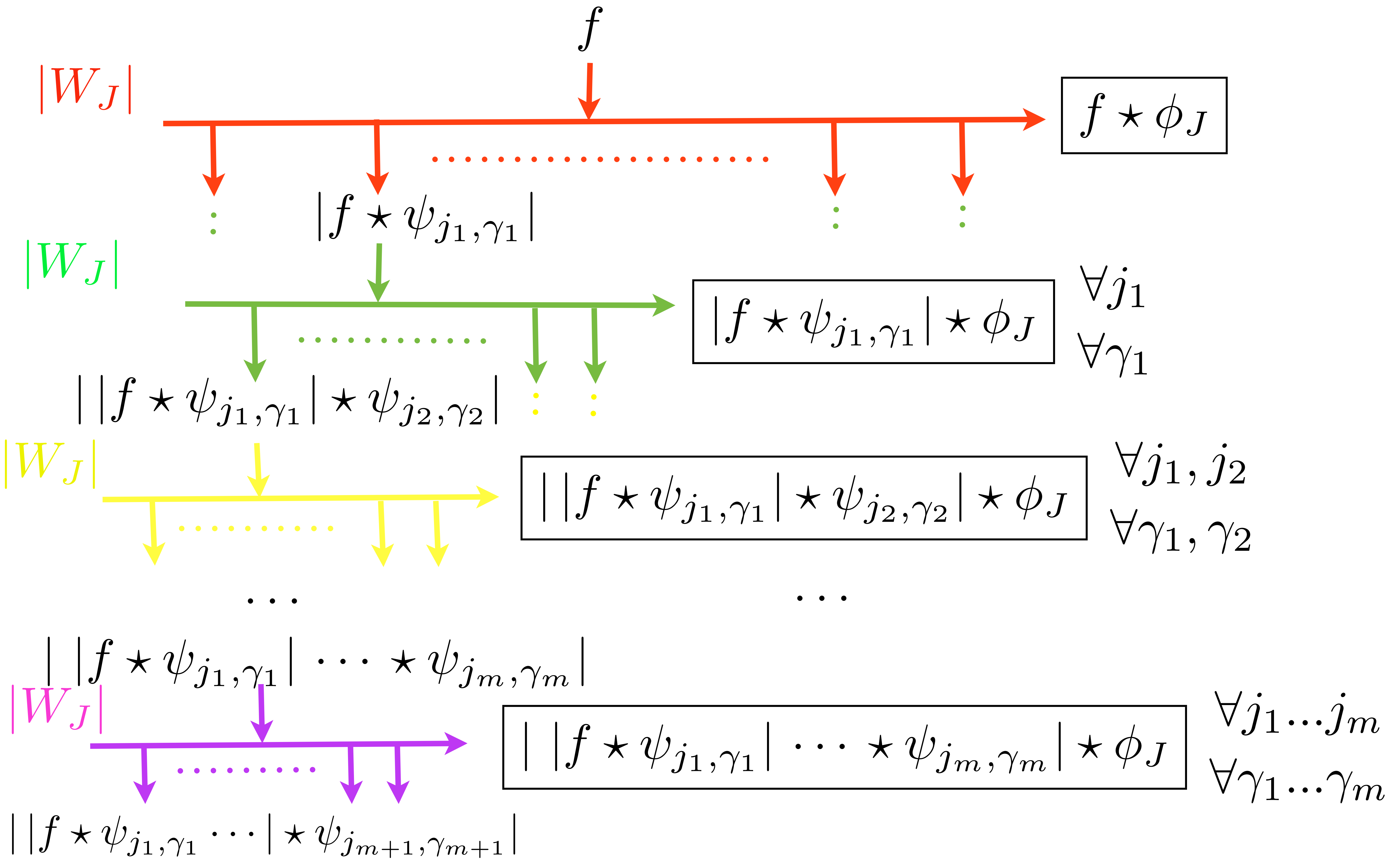
- Local averaging kernel: $x \star \phi_J$
  - locally translation invariant
  - stable to additive and geometric deformations
  - loss of high-frequency information.
- Recover lost information: $\mathcal{U}_J(x) = \{x \star \phi_J \,,\, |x \star \psi_\lambda|\}_{\lambda \in \Lambda_J}$.
  - Point-wise, non-expansive non-linearities: maintain stability.
  - Complex modulus maps energy towards low-frequencies.
- Cascade the "recovery" operator:

$$\mathcal{U}_J^2(x) = \{x \star \phi_J \,,\, |x \star \psi_\lambda| \star \phi_J ,\, ||x \star \psi_\lambda| \star \psi_{\lambda'}|\}_{\lambda, \lambda' \in \Lambda_J} \ .$$

- Scattering coefficient along a path $p = (\lambda_1, \dots, \lambda_m)$:

$$S_J[p]x(u) = |||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \dots | \star \psi_{\lambda_m}| \star \phi_J(u) \ .$$

Scattering Convolutional Network

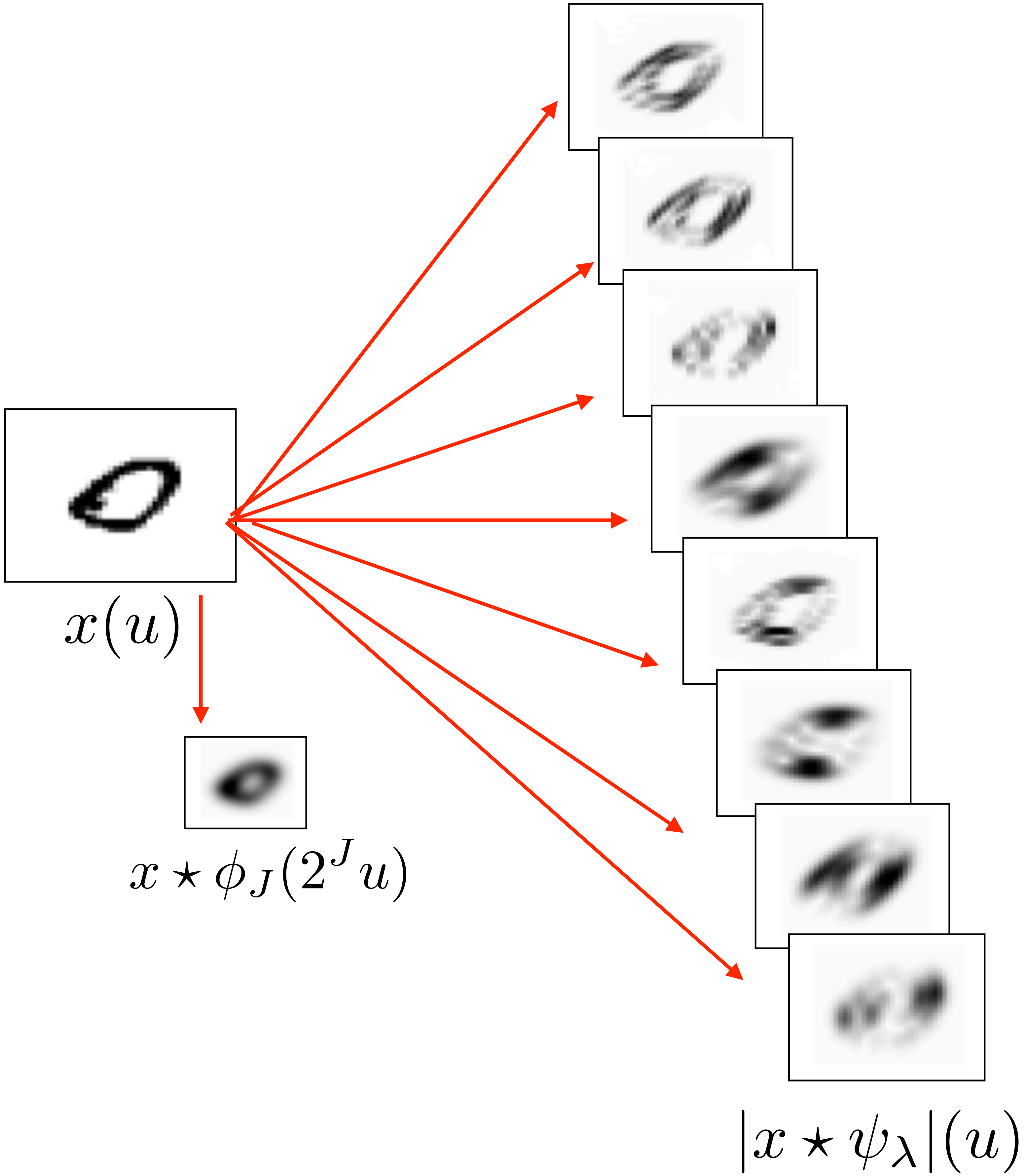


Cascade of contractive operators.

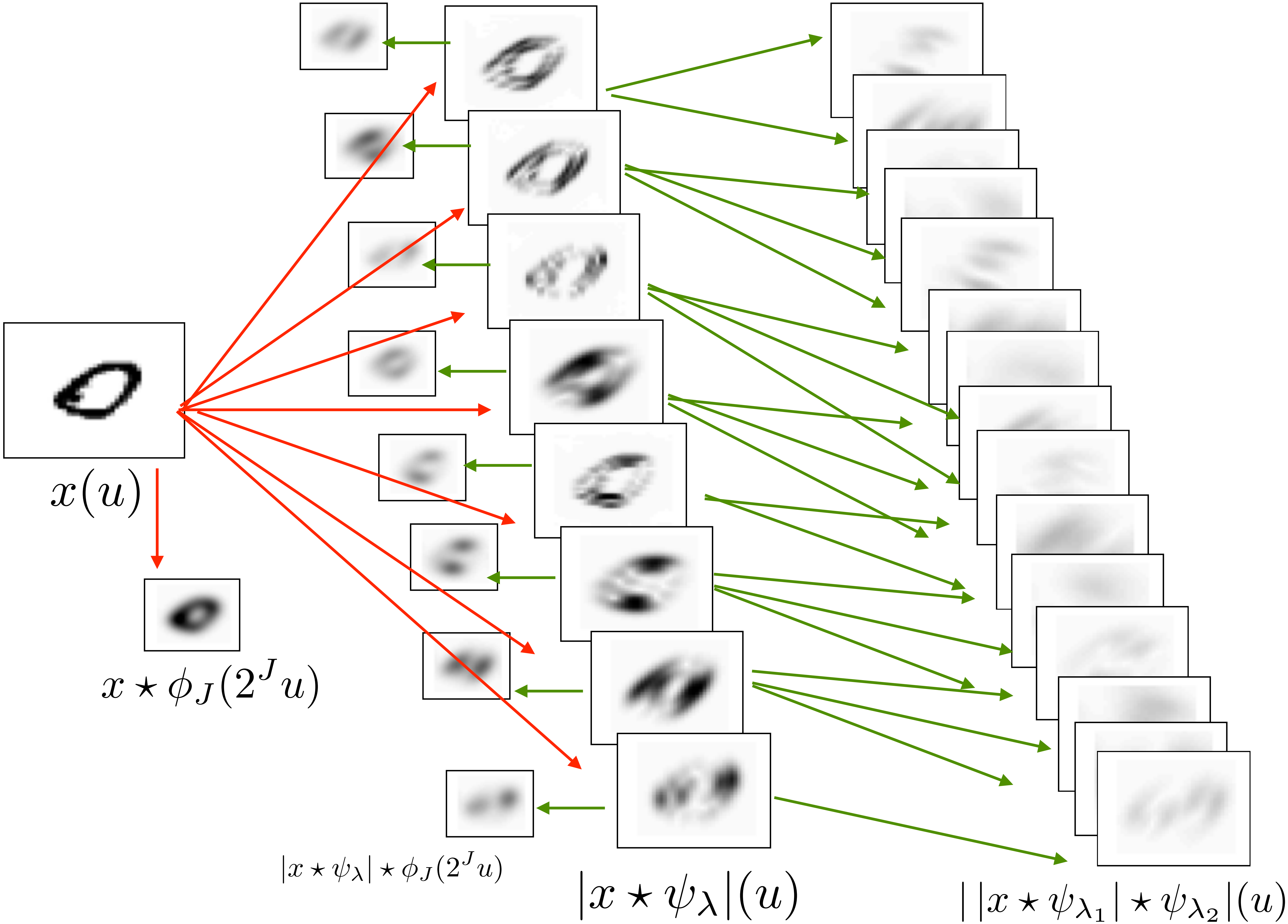
Scattering Example

$x(u)$

Scattering Example

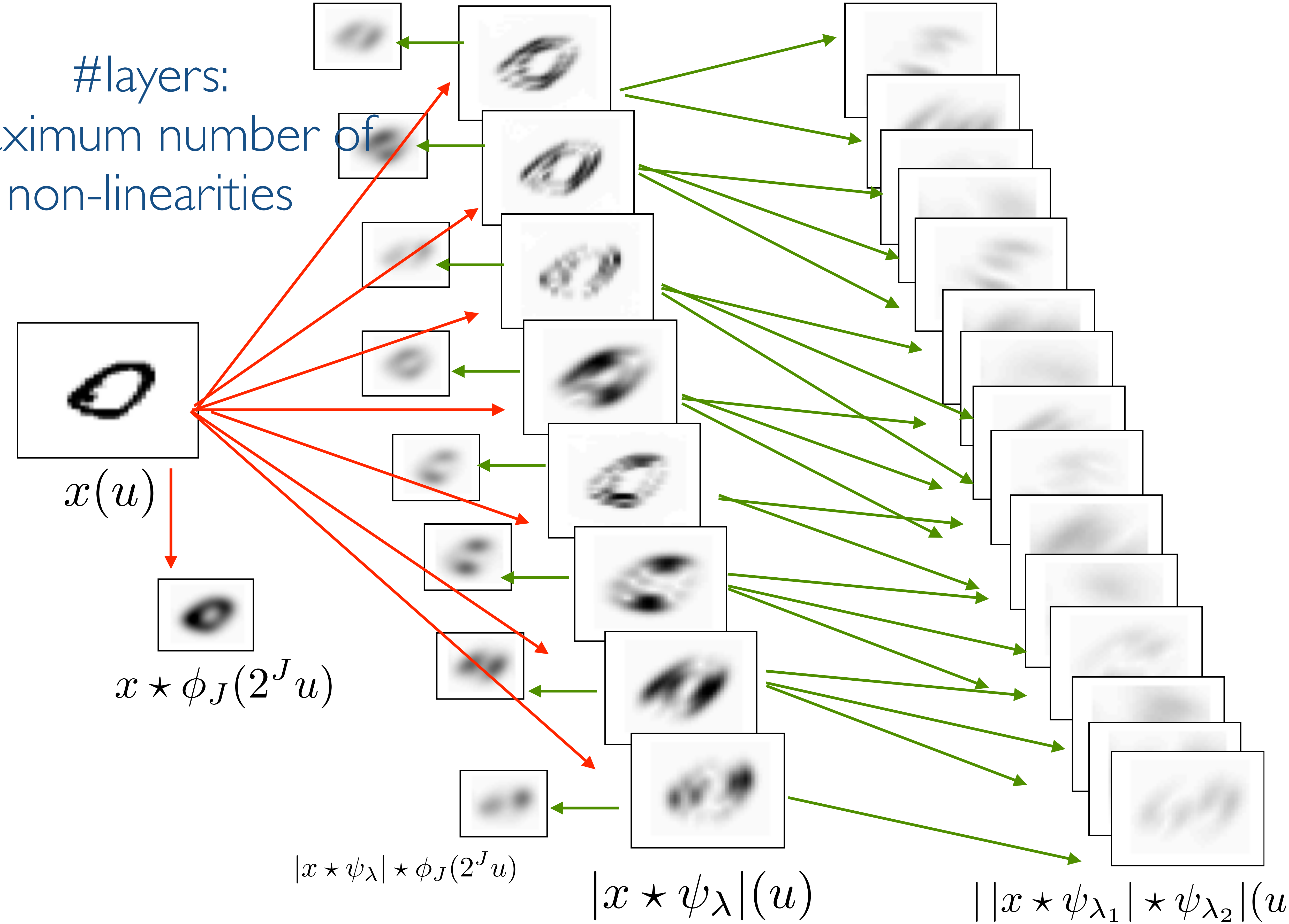


Scattering Example



Scattering Example

#layers:
maximum number of
non-linearities



Scattering with Multi-Resolution Wavelets

- We have considered a collection $\psi_{j,\theta}$ of oriented and dilated wavelets, and a translation co-variant wavelet decomposition operator:

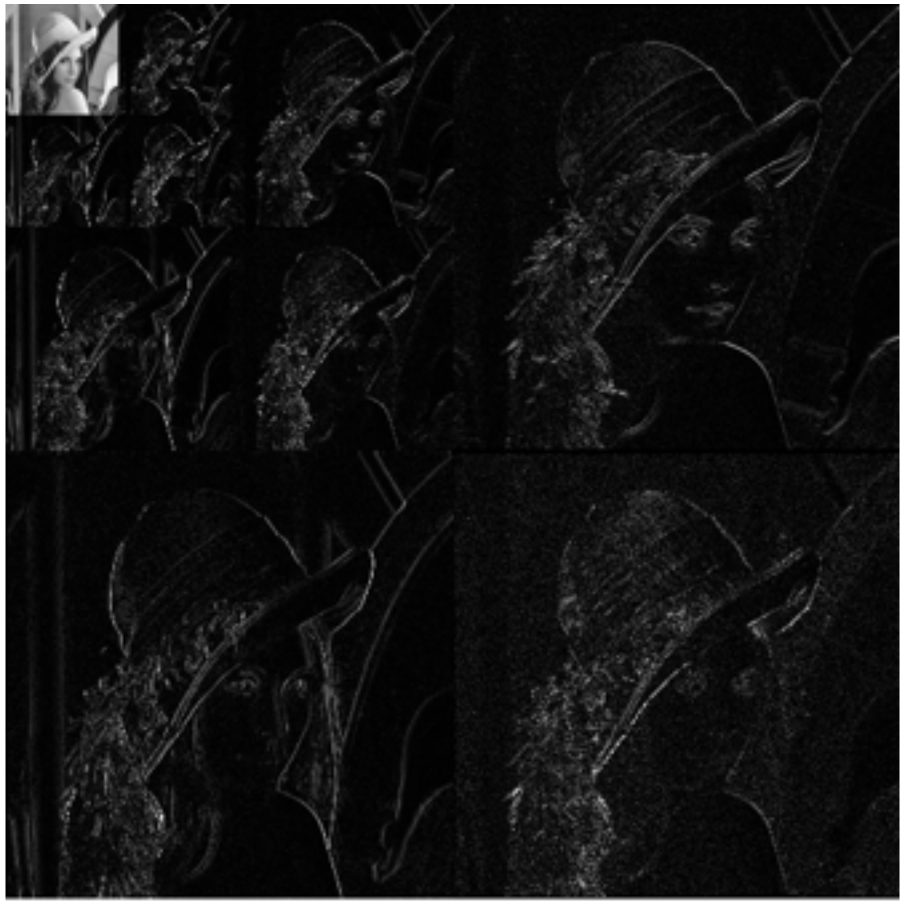
$$Wx = \{x \star \phi_J, x \star \psi_{j,\theta}\}$$

- With *J* scales and *L* orientations, the redundancy is (1+*JL*).

Scattering with Multi-Resolution Wavelets

- With J scales and L orientations, the redundancy is $(1+JL)$.
- This is in contrast with *orthogonal* wavelet transforms, used for compression and (suboptimally) for denoising.

$$x \in \mathbb{R}^N \to Wx \in \mathbb{R}^N \; . \qquad W^T W = Id$$



example of orthogonal wavelet decomposition

- A very efficient algorithm exists using filter cascades with MultiResolution Analysis.

Multi-Resolution Wavelets

- At each scale j, we consider a low-pass *scaling filter* h and band-pass filters g_θ, $\theta \in [1, \ldots, L]$.
- Wavelets and the bluring kernel are obtained at each j by cascading these filters:

$$\phi_j = \phi_{j-1} \star h_j \quad \psi_{j,\theta} = \phi_{j-1} \star g_{j,\theta} \ .$$

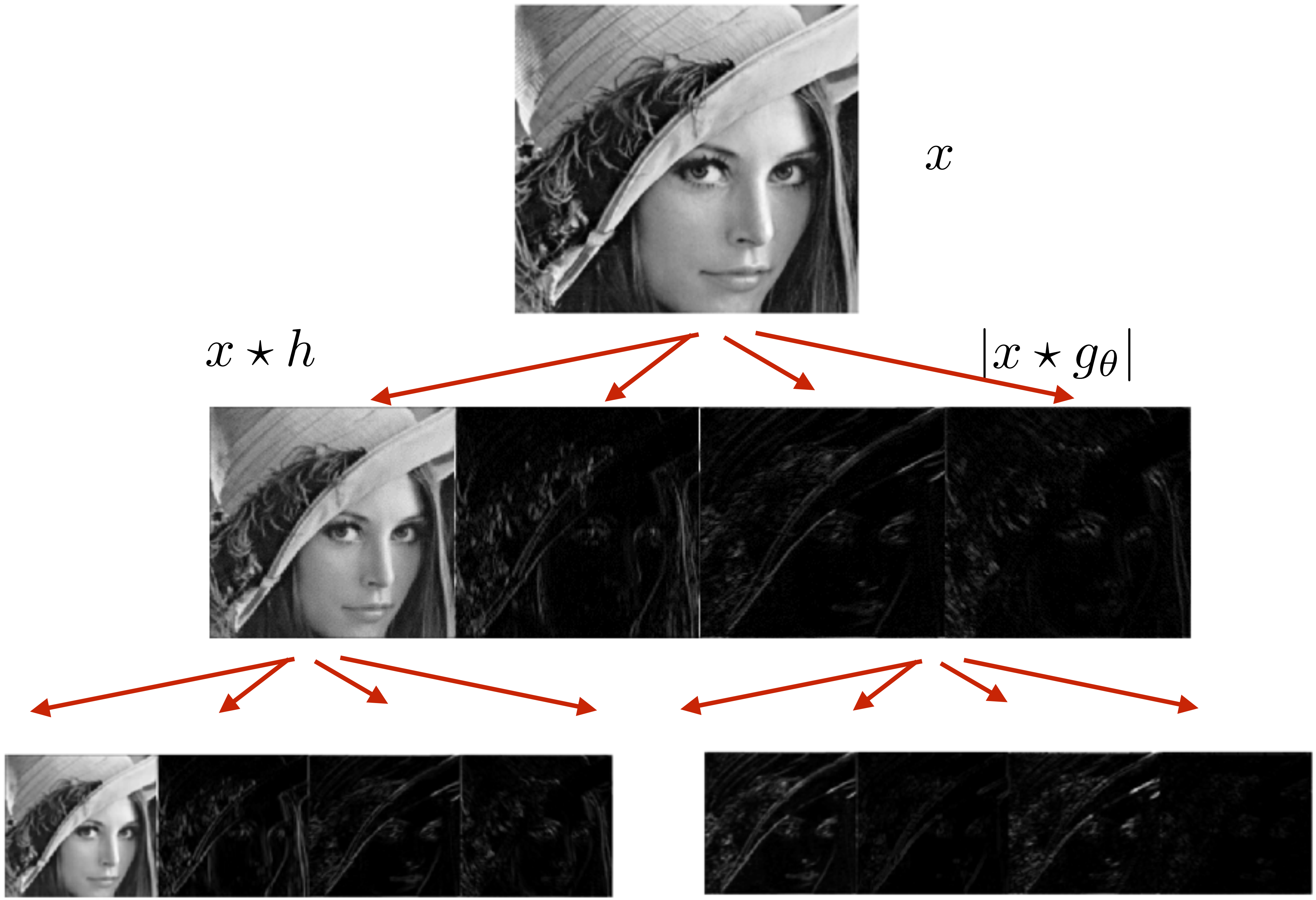
- Decompositions are obtained by cascading fine-to-coarse:

$$x \star \phi_j(u) = (x \star \phi_{j-1}) \star h_j(u) \quad , \quad x \star \psi_{j,\theta}(u) = (x \star \phi_{j-1}) \star g_{j,\theta}(u) \ .$$

- Downsampling (or "*stride*") adaptive to signal smoothness:

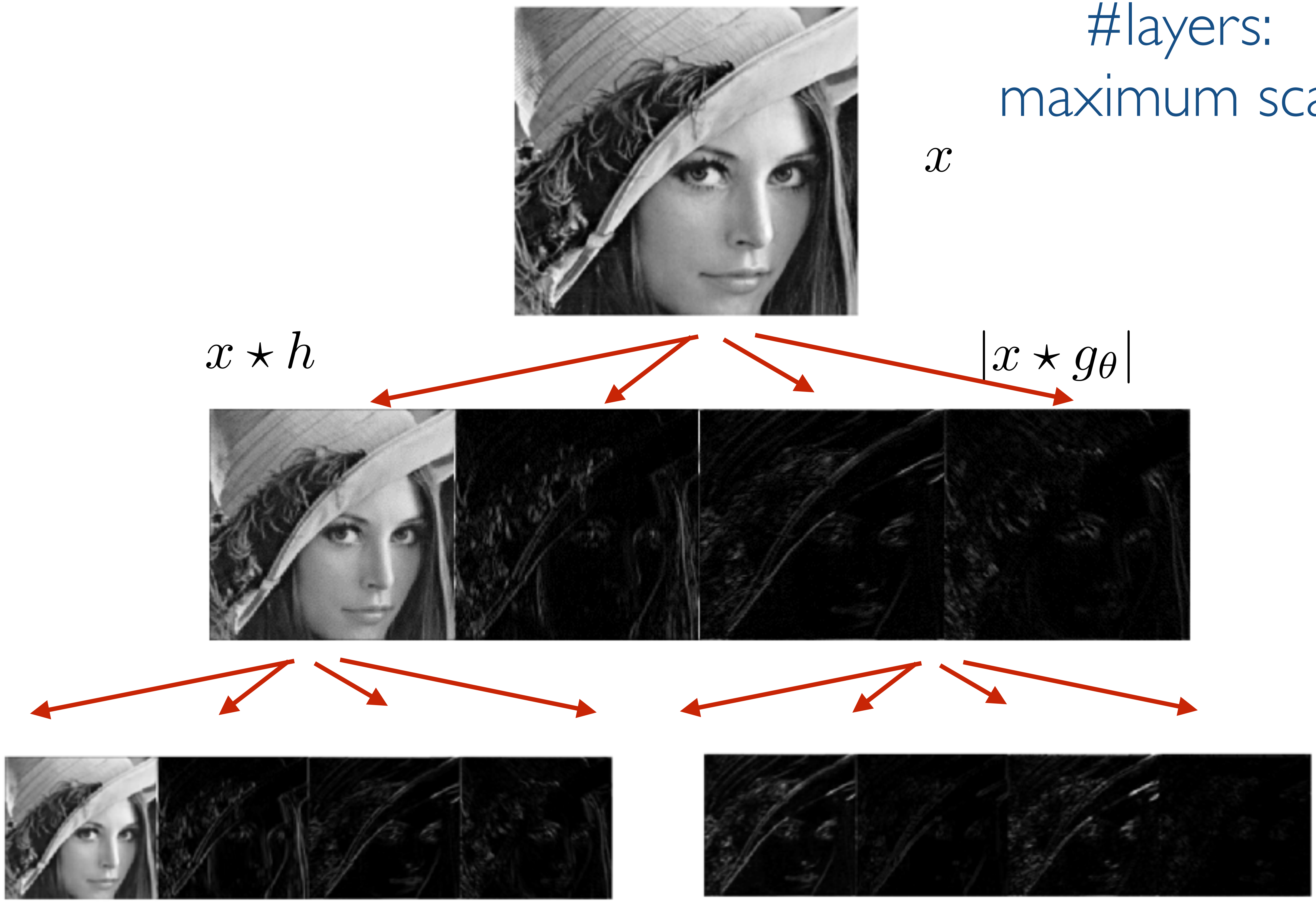
$$x \star \phi_j(u) = (x \star \phi_{j-1}) \star h(2u) \quad , \quad x \star \psi_{j,\theta}(u) = (x \star \phi_{j-1}) \star g_\theta(2u) \ .$$

Scattering with Multi-Resolution Wavelets



Scattering with Multi-Resolution Wavelets

#layers:
maximum scale



Scattering Conservation of Energy

Theorem *(Mallat)*: For appropriate wavelets, the scattering representation is contractive, $\|S_J x - S_J x'\| \leq \|x - x'\|$,
and unitary, $\|S_J x\| = \|x\|$.

$$\|S_J x\|^2 = \sum_{p \in \mathcal{P}_J} \|S_J[p]x\|^2$$

Scattering Conservation of Energy

Theorem *(Mallat)*: For appropriate wavelets, the scattering representation is contractive, $\|S_J x - S_J x'\| \leq \|x - x'\|$, and unitary, $\|S_J x\| = \|x\|$.

$$\|S_J x\|^2 = \sum_{p \in \mathcal{P}_J} \|S_J[p]x\|^2$$

- In practice, the transform is limited to a finite number of layers m_{max}. This result shows residual error converges to 0.
- The result requires complex wavelets (ie, not real).

Interpretation

- Unitary Wavelet decomposition preserves energy:

$$\|x\|^2 = \|x \star \phi_J\|^2 + \sum_{j \le J, \theta} \|x \star \psi_{j,\theta}\|^2 \ .$$

- Repeat formula on each output $|x \star \psi_{j,\theta}|$:

$$\||x \star \psi_{j,\theta}|\|^2 = \||x \star \psi_{j,\theta}| \star \phi_J\|^2 + \sum_{j_2 \le J, \theta_2} \||x \star \psi_{j,\theta}| \star \psi_{j_2,\theta_2}\|^2 \ .$$

$$\|x\|^2 = \|S_J[0]x\|^2 + \sum_{|p|=1} \|S_J[p]x\|^2 + \sum_{|p|=2} \||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}\|^2$$

$\forall \ m$

$$\|x\|^2 = \sum_{|p|<m} \|S_J[p]x\|^2 + \sum_{|p|=m} \|||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \dots \psi_{\lambda_m}\|^2$$

Interpretation

- Unitary Wavelet decomposition preserves energy:

$$\|x\|^2 = \|x \star \phi_J\|^2 + \sum_{j \le J, \theta} \|x \star \psi_{j,\theta}\|^2 \ .$$

- Repeat formula on each output $|x \star \psi_{j,\theta}|$:

$$\||x \star \psi_{j,\theta}|\|^2 = \||x \star \psi_{j,\theta}| \star \phi_J\|^2 + \sum_{j_2 \le J, \theta_2} \||x \star \psi_{j,\theta}| \star \psi_{j_2,\theta_2}\|^2 \ .$$

$$\|x\|^2 = \|S_J[0]x\|^2 + \sum_{|p|=1} \|S_J[p]x\|^2 + \sum_{|p|=2} \||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}\|^2$$

$\forall\ m$

$$\|x\|^2 = \sum_{|p|<m} \|S_J[p]x\|^2 + \sum_{|p|=m} \|||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \dots \psi_{\lambda_m}\|^2$$

Interpretation

- Result amounts to proving that

$$\lim_{m\to\infty} \sum_{|p|=m, j_i \leq J} \| | | x \star \psi_{\lambda_1} | \star \ldots | \star \psi_{\lambda_m} | \|^2 = 0 \ .$$

- *Fact*: Every time we apply the (complex) wavelet modulus, we push energy towards the low frequencies.
- Result is obtained by formally proving this fact.

Interpretation

- Result amounts to proving that

$$\lim_{m\to\infty} \sum_{|p|=m, j_i \le J} \| |||x \star \psi_{\lambda_1}| \star \ldots | \star \psi_{\lambda_m}| \|^2 = 0 \ .$$

- Fact: Every time we apply the (complex) wavelet modulus, we push energy towards the low frequencies.
- Result is obtained by formally showing this fact.

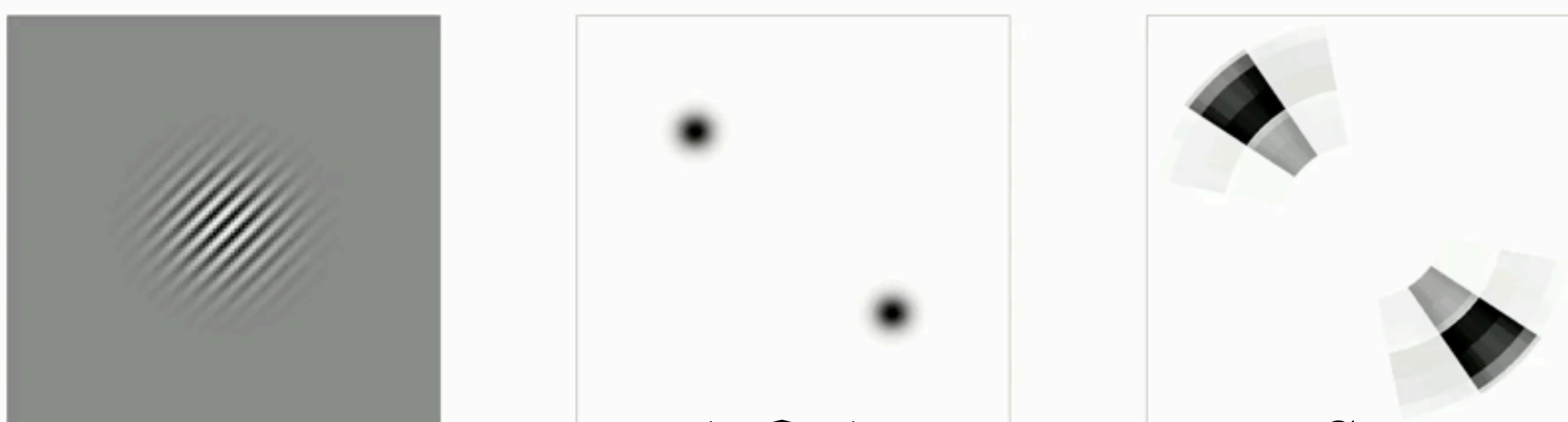
- It requires a non-linearity that produces smooth envelopes:
  - complex wavelets OK
  - real wavelets: ??

Scattering Geometric Stability

- Geometric Stability: $$\|S_J x\|^2 = \sum_{p \in \mathcal{P}_J} \|S_J[p]x\|^2$$

Theorem (Mallat'10): There exists C such that for all $x \in L^2(R^d)$ and all m, the m-th order scattering satisfies

$$\|S_J \varphi_\tau x - S_J x\| \leq C m \|x\| (2^{-J} |\tau|_\infty + \|\nabla \tau\|_\infty + \|H\tau\|_\infty) \ .$$



$\varphi_\tau x$ $\quad$ $|\widehat{\varphi_\tau x}|$ $\quad$ $S_J \varphi_\tau x$

Interpretation

- Denote

$$A_J x = x \star \phi_J \qquad W_J x = \{x \star \psi_\lambda\}_\lambda \qquad M x = |x|$$

- We know that

$$\|A_J - A_J \varphi_\tau\| \leq C(2^{-J}|\tau|_\infty + |\nabla\tau|_\infty)$$

$$\|W_J \varphi_\tau - \varphi_\tau W_J\| \leq C(J|\nabla\tau|_\infty + |H\tau|_\infty)$$

$$M\varphi_\tau = \varphi_\tau M$$

$$([A, B] = AB - BA \ : \ \text{Commutator})$$

- $S_J = \{A_J, A_J M W_J, A_J M W_J M W_J, \ldots\}$

Interpretation

- Denote

$$A_J x = x \star \phi_J \qquad W_J x = \{x \star \psi_\lambda\}_\lambda \qquad Mx = |x|$$

- We know that

$$\|A_J - A_J \varphi_\tau\| \leq C(2^{-J}|\tau|_\infty + |\nabla \tau|_\infty)$$

$$\|W_J \varphi_\tau - \varphi_\tau W_J\| \leq C(J|\nabla \tau|_\infty + |H\tau|_\infty)$$

$$M\varphi_\tau = \varphi_\tau M$$

$$([A, B] = AB - BA \ : \ \text{Commutator})$$

- $S_J = \{A_J, A_J M W_J, A_J M W_J M W_J, \ldots\}$

Interpretation

- Each order contributes separately:

$$\|S_J - S_J\varphi_\tau\|^2 = \|A_J - A_J\varphi_\tau\|^2 + \|A_J M W_J - A_J M W_J \varphi_\tau\|^2 + \ldots$$

Interpretation

- Each order contributes separately:

$$\|S_J - S_J\varphi_\tau\|^2 = \|A_J - A_J\varphi_\tau\|^2 + \|A_J M W_J - A_J M W_J \varphi_\tau\|^2 + \ldots$$

- Let us inspect a generic term:

$$\|A_J \underbrace{MW_J MW_J \ldots MW_J}_{k \text{ times}} - A_J \underbrace{MW_J MW_J \ldots MW_J}_{k \text{ times}} \varphi_\tau\|$$

$$(U_J = MW_J)$$

$$\|A_J U_J^k - A_J U_J^k \varphi_\tau\| \leq \|A_J U_J^k - A_J U_J^{k-1}\varphi_\tau U_J\| + \|A_J U_J^{k-1}\varphi_\tau U_J - A_J U_J^k \varphi_\tau\|$$

Interpretation

- Each order contributes separately:

$$\|S_J - S_J\varphi_\tau\|^2 = \|A_J - A_J\varphi_\tau\|^2 + \|A_J M W_J - A_J M W_J \varphi_\tau\|^2 + \dots$$

- Let us inspect a generic term:

$$\|A_J \underbrace{MW_J MW_J \dots MW_J}_{k \text{ times}} - A_J \underbrace{MW_J MW_J \dots MW_J}_{k \text{ times}} \varphi_\tau\|$$

$$(U_J = MW_J)$$

$$\|A_J U_J^k - A_J U_J^k \varphi_\tau\| \leq \|A_J U_J^k - A_J U_J^{k-1}\varphi_\tau U_J\| + \|A_J U_J^{k-1}\varphi_\tau U_J - A_J U_J^k \varphi_\tau\|$$

$$\leq \|A_J U_J^{k-1} - A_J U_J^{k-1}\varphi_\tau\| + \|A_J U_J^{k-1}[\varphi_\tau, U_J]\|$$

$$\leq \|A_J U_J^{k-1} - A_J U_J^{k-1}\varphi_\tau\| + \|[\varphi_\tau, U_J]\|$$

$$\leq k\|[\varphi_\tau, U_J]\| + \|A_J - A_J\varphi_\tau\| \leq k\|[\varphi_\tau, W_J]\| + \|A_J - A_J\varphi_\tau\|$$

Scattering Discriminability

- For appropriate wavelets, the information is preserved at each layer:

Theorem: (Waldspurger) For appropriate wavelets, the operator $Ux = \{x \star \phi_J, |x \star \psi_j|\}_{j \leq J}$ is injective.

Scattering Discriminability

- For appropriate wavelets, the information is preserved at each layer:

Theorem: (Waldspurger) For appropriate wavelets, the operator $Ux = \{x \star \phi_J, |x \star \psi_j|\}_{j \leq J}$ is injective.

- Very different situation than Fourier modulus (why?)

Scattering Discriminability

- For appropriate wavelets, the information is preserved at each layer:

Theorem: (Waldspurger) For appropriate wavelets, the operator $Ux = \{x \star \phi_J, |x \star \psi_j|\}_{j \leq J}$ is injective.

- Very different situation than Fourier modulus (why?)
- The representation is highly redundant.
- However, the inverse is unstable for large J: we might be contracting too much in general.
- How to prevent that?

Discriminability and Sparsity

- Typical non-linearities are contractive:

$$\|\rho(x) - \rho(x')\| \leq \|x - x'\|$$

Discriminability and Sparsity

- Typical non-linearities are contractive:

$$\|\rho(x) - \rho(x')\| \leq \|x - x'\|$$

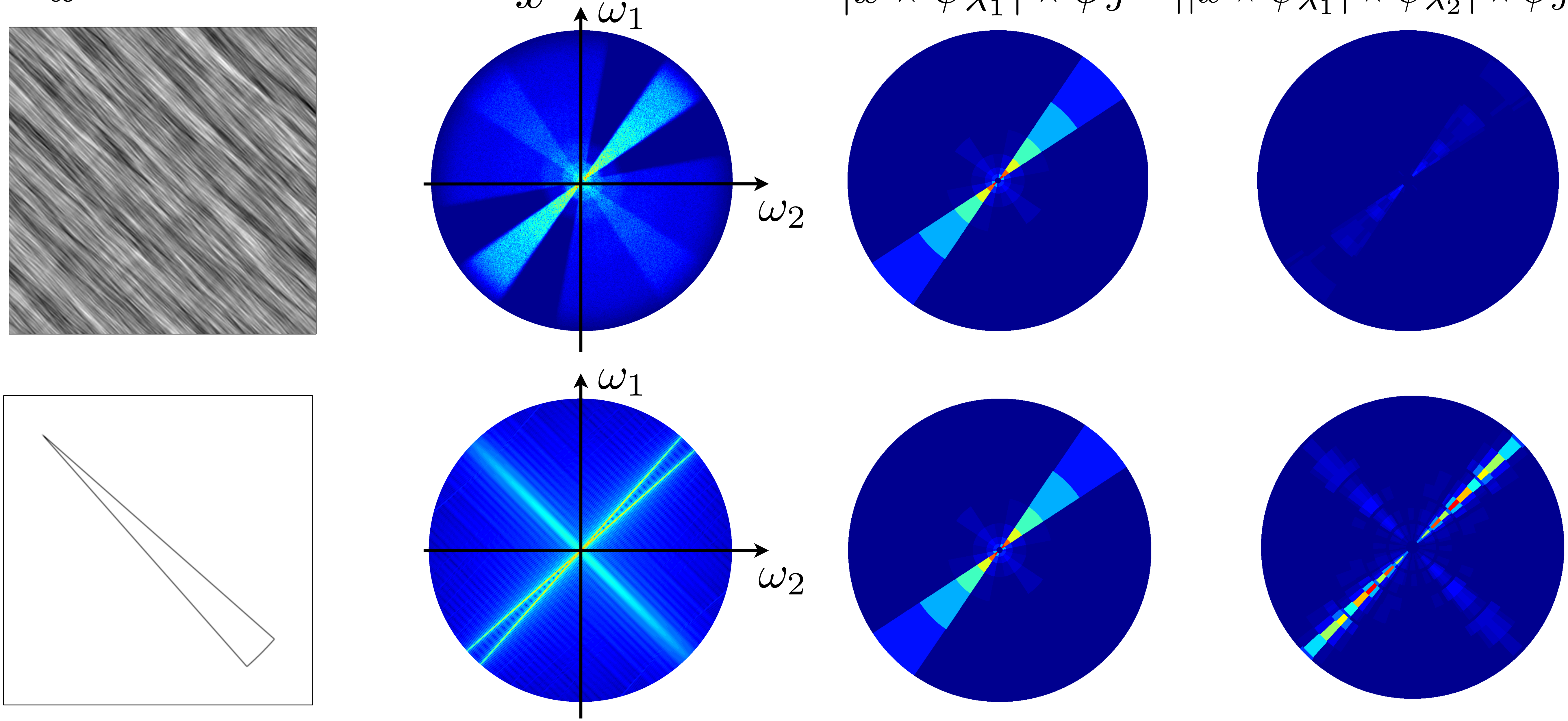
- However, if x, x' are sparse, this inequality is an equality in most of the signal domain.

- Thus sparsity is a means to control and prevent excessive contraction of different signal classes.

Image Examples

Images | Fourier | Wavelet Scattering

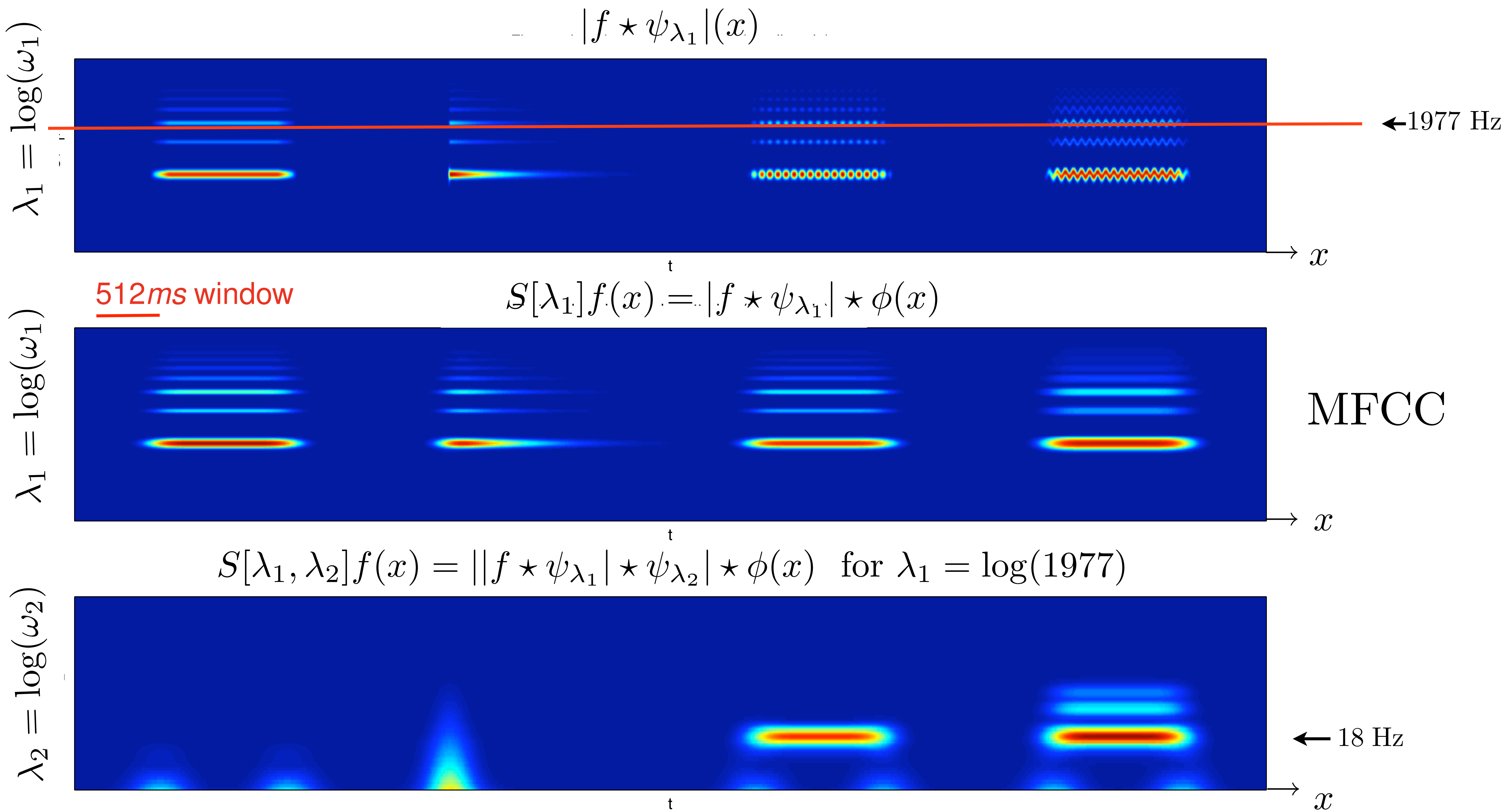
x | $\hat{x}$ | $|x \star \psi_{\lambda_1}| \star \phi_J$ | $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_J$



window size = image size

Sound Examples

(courtesy J. Anden)



Limitations of Separable Scattering

- No feature dimensionality reduction
  - The number of features increases exponentially with depth and polynomially with scale.

Limitations of Separable Scattering

- No feature dimensionality reduction
  - The number of features increases exponentially with depth and polynomially with scale.

- We are indirectly assuming that each wavelet band is deformed independently
  - We cannot capture the *joint* deformation structure of feature maps
  - Loss of discriminability.

Limitations of Separable Scattering

- No feature dimensionality reduction
  - The number of features increases exponentially with depth and polynomially with scale.

- We are indirectly assuming that each wavelet band is deformed independently
  - We cannot capture the *joint* deformation structure of feature maps
  - Loss of discriminability.

- The deformation model is rigid and non-adaptive
  - We cannot adapt to each class
  - Wavelets are hard to define *a priori* on high-dimensional domains.

Joint versus Separable Invariance

- Suppose we simply want stable translation invariance.
- Two-dimensional translation group in a periodic domain:

$$G \cong (\mathbb{R}/([0,N]))^2 = S^1 \times S^1 \cong \mathbb{T}^2$$

- Each S^1 acts on images along a different coordinate:

$$\varphi_a^1 x(u_1, u_2) = x(u_1 - a, u_2) \;,\; \varphi_a^2 x(u_1, u_2) = x(u_1, u_2 - a)$$

Joint versus Separable Invariance

- Suppose we simply want stable translation invariance.
- Two-dimensional translation group in a periodic domain:

$$G \cong (\mathbb{R}/([0,N]))^2 = S^1 \times S^1 \cong \mathbb{T}^2$$

- Each S^1 acts on images along a different coordinate:

$$\varphi_a^1 x(u_1, u_2) = x(u_1 - a, u_2)\ ,\ \varphi_a^2 x(u_1, u_2) = x(u_1, u_2 - a)$$

Joint versus Separable Invariance

- So we could just consider one-dimensional (stable) translation invariant representations and compose:

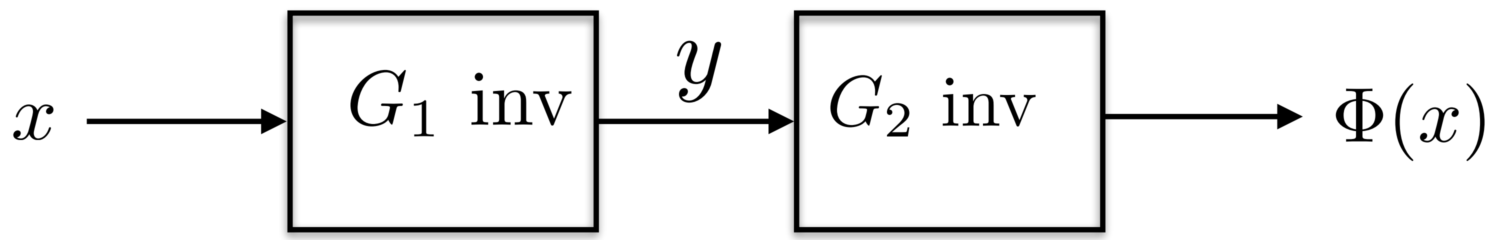
$$G = G_1 \times G_2$$

$$x \longrightarrow \boxed{G_1 \ \text{inv}} \xrightarrow{y} \boxed{G_2 \ \text{inv}} \longrightarrow \Phi(x)$$

Joint versus Separable Invariance

- So we could just consider one-dimensional (stable) translation invariant representations and compose:

$$G = G_1 \times G_2$$



- If for each u_2, $x(\cdot, u_2) \mapsto \Phi_1(x)(\cdot, u_2)$ is G_1 invariant then $\Phi_1(\varphi^1 x) = \Phi_1(x)$ for all x and $\varphi^1 \in G_1$
- If for each λ, $y(\lambda, \cdot) \mapsto \Phi_2(y)(\lambda, \cdot)$ is G_2 invariant then $\Phi_2(\varphi^2 y) = \Phi_2(y)$ for all y and $\varphi^2 \in G_2$

Joint versus Separable Invariance

- Thus, if Φ_1 is G_1 invariant and G_2 covariant, and Φ_2 is G_2 invariant, then $\Phi = \Phi_2 \circ \Phi_1$ satisfies

$$\forall\ \varphi \in G,\ \varphi = \varphi^1 \varphi^2,\ \varphi^i \in G_i$$

$$\Phi(\varphi x) = \Phi_2 \Phi_1(\varphi^1 \varphi^2 x) = \Phi_2 \Phi_1(\varphi^2 x) = \Phi_2 \varphi^2 \Phi_1(x) = \Phi_2 \Phi_1(x) = \Phi(x)$$

Joint versus Separable Invariance

- Thus, if Φ_1 is G_1 invariant and G_2 covariant, and Φ_2 is G_2 invariant, then $\Phi = \Phi_2 \circ \Phi_1$ satisfies

$$\forall\ \varphi \in G,\ \varphi = \varphi^1 \varphi^2,\ \varphi^i \in G_i$$

$$\Phi(\varphi x) = \Phi_2 \Phi_1(\varphi^1 \varphi^2 x) = \Phi_2 \Phi_1(\varphi^2 x) = \Phi_2 \varphi^2 \Phi_1(x) = \Phi_2 \Phi_1(x) = \Phi(x)$$

- So we achieve further invariance by composing partial invariances.

Joint versus Separable Invariance

- Thus, if Φ_1 is G_1 invariant and G_2 covariant, and Φ_2 is G_2 invariant, then $\Phi = \Phi_2 \circ \Phi_1$ satisfies

$$\forall\ \varphi \in G,\ \varphi = \varphi^1 \varphi^2,\ \varphi^i \in G_i$$

$$\Phi(\varphi x) = \Phi_2 \Phi_1(\varphi^1 \varphi^2 x) = \Phi_2 \Phi_1(\varphi^2 x) = \Phi_2 \varphi^2 \Phi_1(x) = \Phi_2 \Phi_1(x) = \Phi(x)$$

- So we achieve further invariance by composing partial invariances.
- Is there a problem here?

Joint versus Separable Invariance

- The factorization does not capture the joint action of G$_l$ along the domain (u_1, u_2).
- We are invariant to *too many* things.

Wavelet Covariants

- If we replace input image by first layer output:

$$\rho(x_0 \star \psi_{j,\theta})(u) = x_1(u, j, \theta)$$

Let $\tilde{x}_0 = R_\alpha x_0$ be a rotation of α degrees.

$$\rho(\tilde{x}_0 \star \psi_{j,\theta})(u) = x_1(R_\alpha u, j, \theta + \alpha)$$

- Similarly, roto-translation acts on x_1 by rotating and translating spatial coordinates and translating orientation coordinates

Let $\tilde{x}_0 = \varphi_{(v,\alpha)} x_0$ be a roto-translation with parameters (v, α).

$$\rho(\tilde{x}_0 \star \psi_{j,\theta})(u) = x_1(\varphi_v R_\alpha u, j, \theta + \alpha)$$

- So we can replace convolutions over translation by convolutions over roto-translations.

Group Convolutions

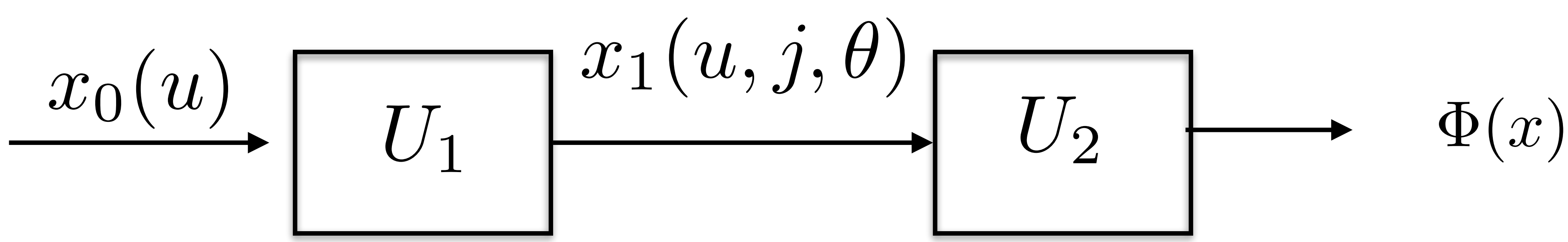
Definition: Let G be a group equipped with a Haar measure $d\mu$, acting on Ω, and $h \in L^1(G)$. The group convolution $x \star_G h$ is defined as

$$x \star_G h(u) = \int_G h(g) x(\varphi_g u) d\mu(g) \quad , \ x \in L^2(\Omega) \ .$$

- If $x = x_1(u, j, \theta)$ and G are roto-translations, these convolutions recombine different orientation channels.

Joint Scattering

- We start by *lifting* the image with spatial wavelet convolutions: stable and covariant to roto-translations.

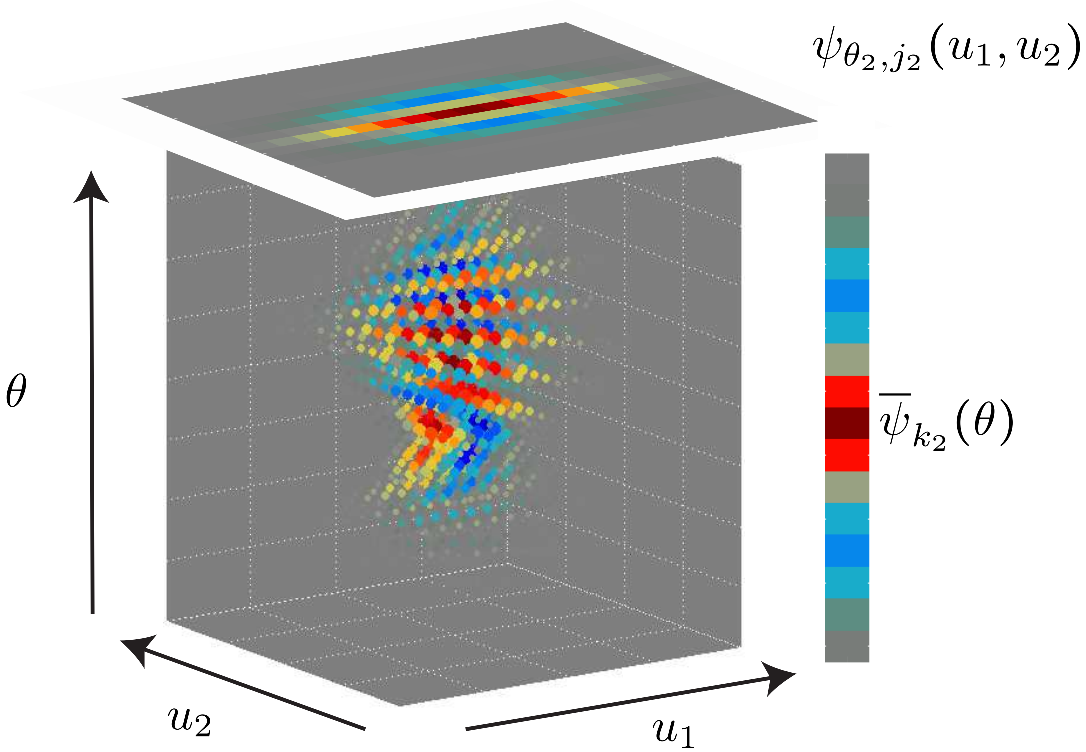


- We then adapt the second wavelet operator to its input joint variability structure.

- More discriminability.
- Requires defining wavelets on more complicated domains

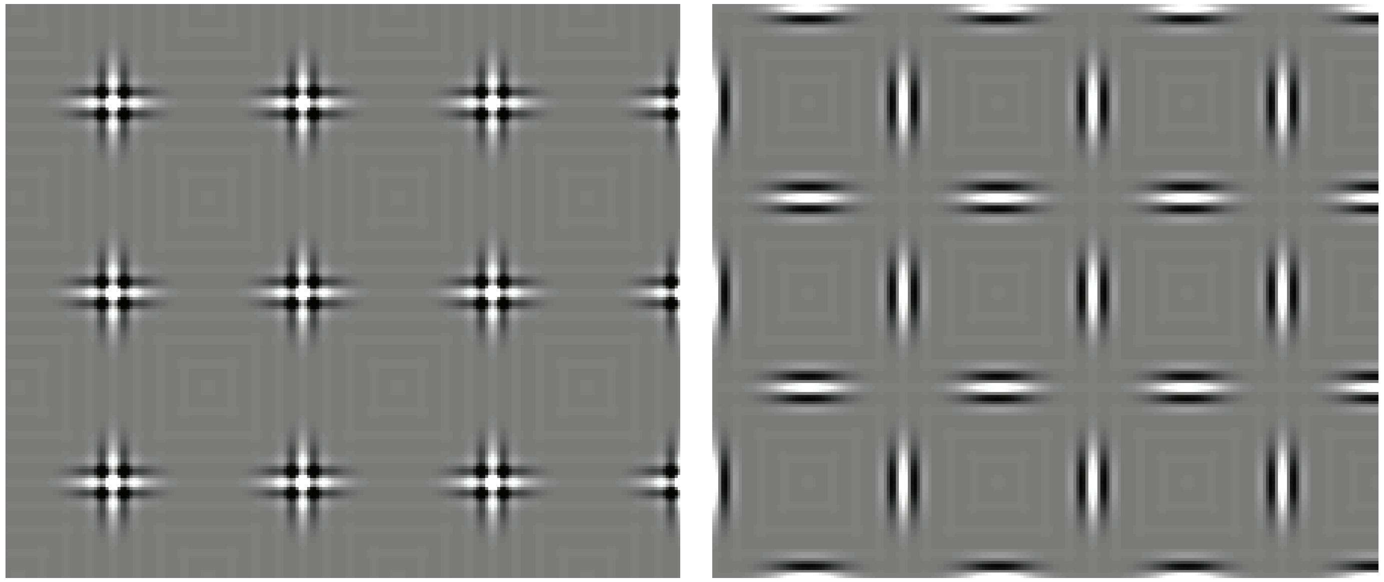
Example: Roto-Translation Scattering

- [Sifre and Mallat'13]



second layer wavelets constructed by a separable product on spatial and rotational wavelets

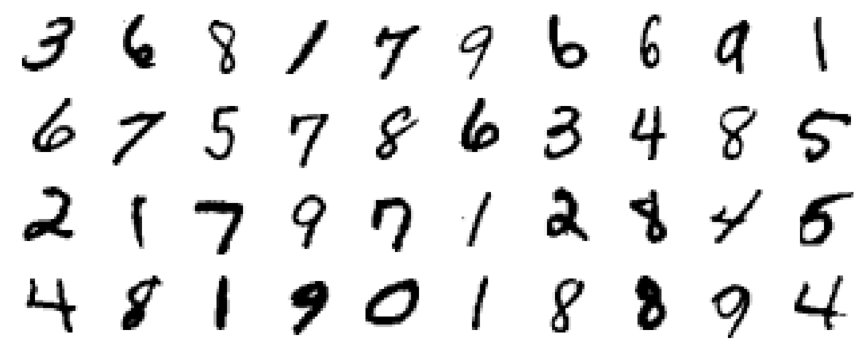
$$\Psi_\lambda(u, \theta) = \psi_{\lambda_1}(u)\psi_{\lambda_2}(\theta)$$



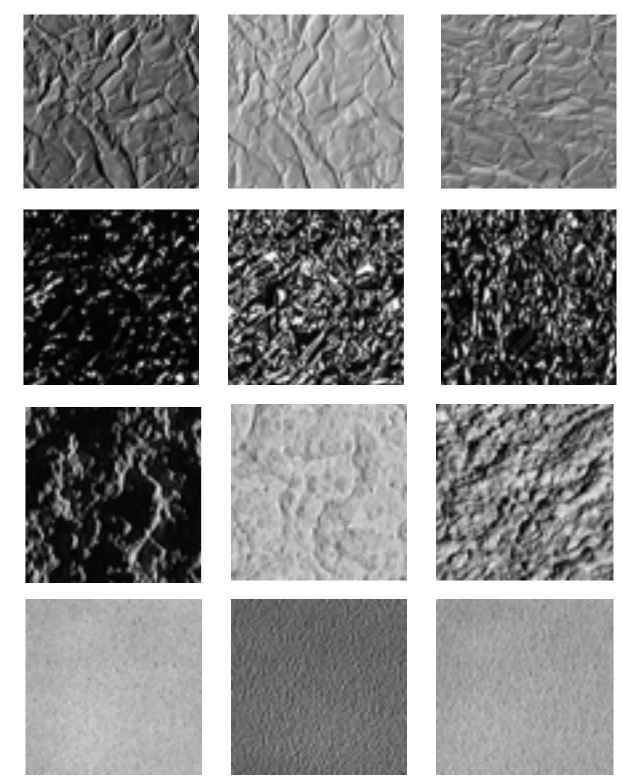
example of patterns that are discriminated by joint scattering but not with separable scattering.

Classification with Scattering

- State-of-the art on pattern and texture recognition using separable scattering followed by SVM:
  - MNIST, USPS [Pami'13]



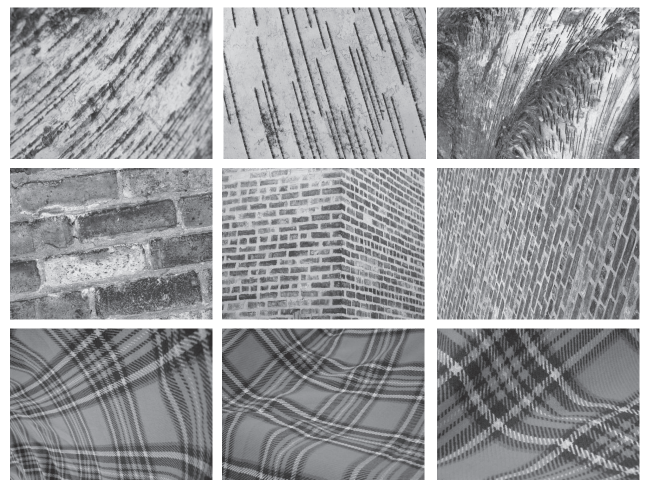
  - Texture (CUREt) [Pami'13]



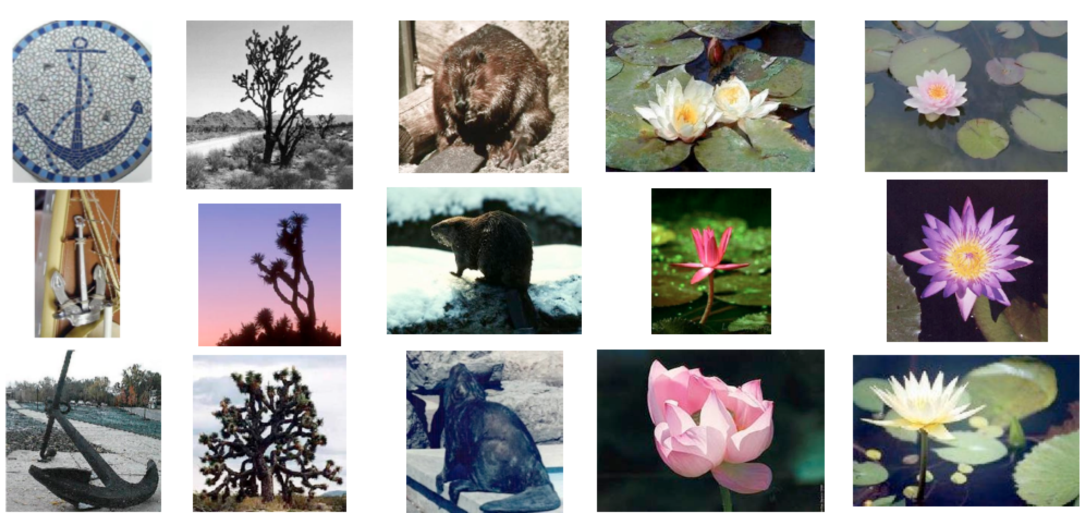
  - Music Genre Classification (GTZAN) [IEEE Acoustic '13]

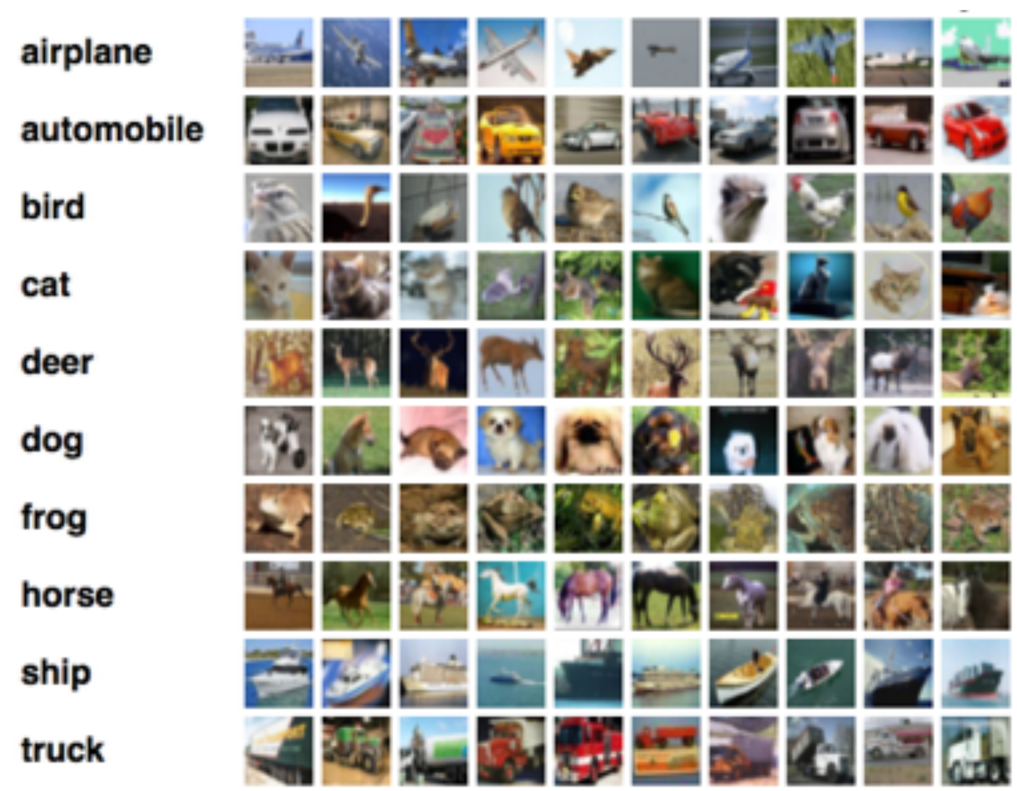
Classification with Scattering

- Joint Scattering Improves Performance:
  - More complicated Texture (KTH,UIUC,UMD) [Sifre&Mallat, CVPR'13]



  - Small-mid scale Object Recognition (Caltech, CIFAR) [Oyallon&Mallat, CVPR'15]
    - ~17% error on Cifar-10





Limitations of Joint Scattering

- Variability from physical world expressed in the language of transformation groups and deformations
  - However, there are not many possible groups: essentially the affine group and its subgroups.

- As a new wavelet layer is introduced, we create new coordinates, but we do not destroy existing coordinates
  - Hard to scale: dimensionality reduction is needed.
  - Wavelet design complicated beyond roto-translation groups.
- Beyond physics, many deformations are class-specific and not small.
  - Learning filters from data rather than designing them.

From Scattering to CNNs

- Given $x(u,\lambda)$ and a group G acting on both u and λ, we defined wavelet convolutions over G as

$$x \star_G \psi_{\lambda'}(u,\lambda) = \int_v \int_\alpha \psi_\lambda(R_{-\alpha}(u-v))x(v,\alpha)dvd\alpha$$

- In discrete coordinates,

$$x \star_G \psi_{\lambda'}(u,\lambda) = \sum_v \sum_\alpha \overline{\psi}_{\lambda'}(u-v,\alpha,\lambda)x(v,\alpha)$$

- Which in general is a convolutional tensor.