M A D

NYU | COURANT INSTITUTE OF MATHEMATICAL SCIENCES

MATHEMATICS OF DEEP LEARNING

JOAN BRUNA , CIMS + CDS, NYU, SPRING'18

Lecture 2: Geometric Stability in Euclidean Domains: The Scattering Transform and beyond

LECTURE OVERVIEW

- Geometric Stability: definitions and priors

- The Scattering Transform.

- Stability Properties of Scattering Transforms

PART I: GEOMETRY OF DATA

high-dimensional space

SETUP: BINARY IMAGE CLASSIFICATION

high-dimensional space

$$f(x) = p(y = 1|x)$$

$$y = \{0, 1\}$$

LINEARIZATION

- We want to obtain a representation $\Phi(x)$ such that

$$\hat{f}(x) = p(y = 1|x) = \sigma(a^\top \Phi(x) + b) \ .$$

is a good approximation of $f(x)$.

$$\sigma(z) = (1 + e^{-z})^{-1}$$

- Thus $f(x)$ is approximately *linearized* by $\Phi(x)$:

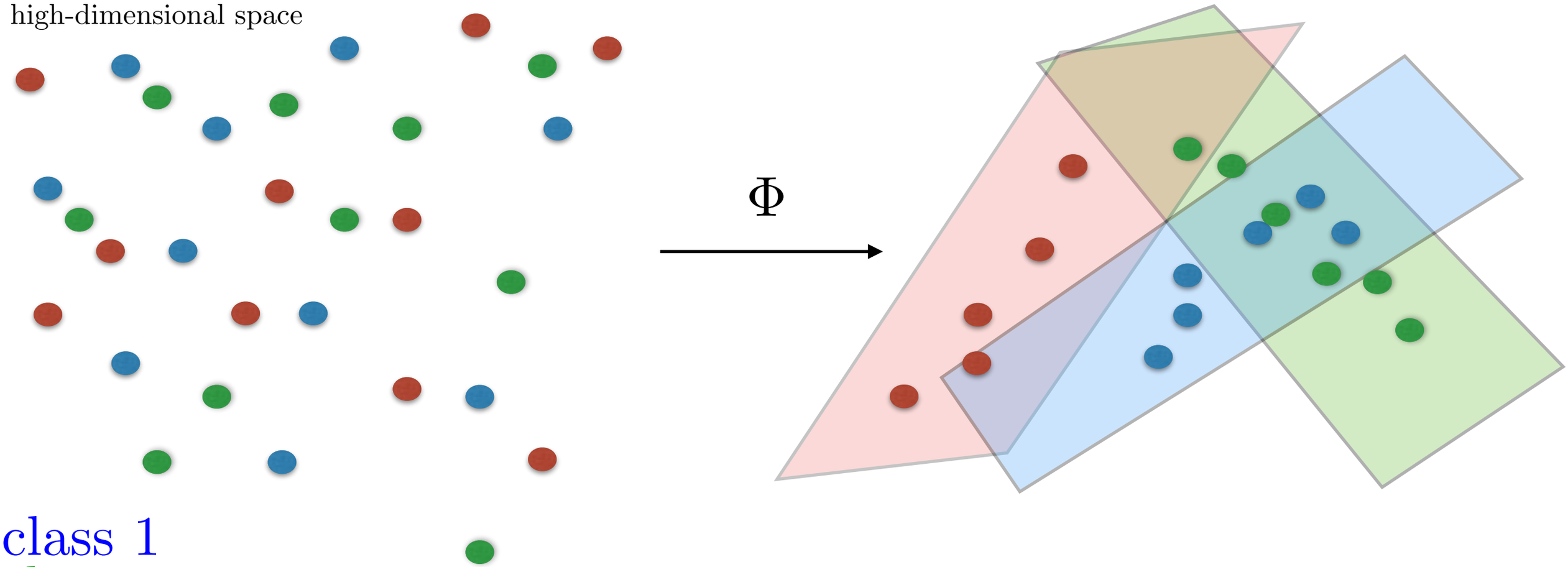
$$f(x) \approx \sigma(a^\top \Phi(x) + b) \ .$$

- In particular, we would like to have

$$a^\top (\Phi(x) - \Phi(x')) = 0 \Rightarrow f(x) = f(x') \ .$$

  - thus level sets of f(x) should be mapped to hyperplanes by Φ.

LINEARIZATION



In order to beat the curse of dimensionality, we need features that **linearize intra-class variability** and **preserve inter-class variability** ***using prior information.***

INVARIANCE AND SYMMETRY

- A global symmetry is an operator $\varphi \in Aut(\Omega)$ that leaves f invariant:

$$\forall\ x \in \Omega\ ,\ \ f(\varphi(x)) = f(x)\ .$$

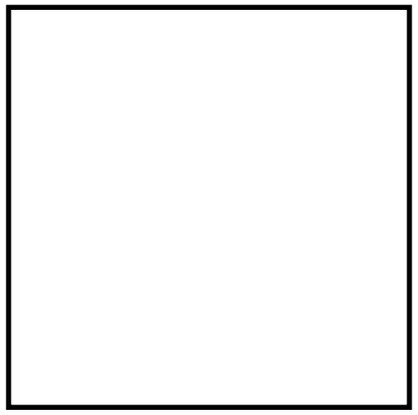
- They can be absorbed by Φ to varying degrees:

Invariants: $\Phi(\varphi(x)) = \Phi(x)$ for each x.
Equivariants: $\Phi(\varphi(x)) = \varphi(\Phi(x))$ for each x, φ.

- What are those symmetries? How to impose them on Φ without breaking discriminability?

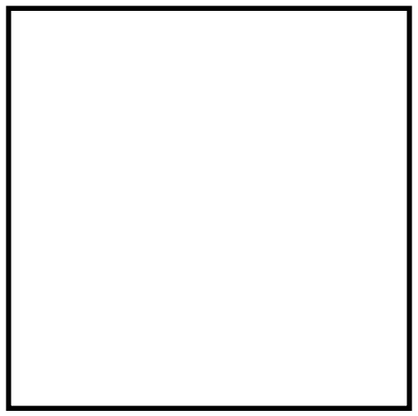
DISCRETE SYMMETRIES

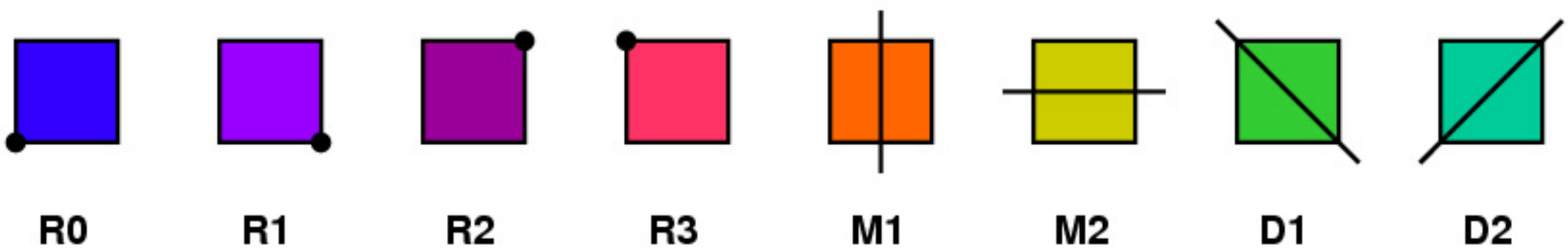
➤ Which transformations leave this square unchanged?



DISCRETE SYMMETRIES

➤ Which transformations leave this square unchanged?



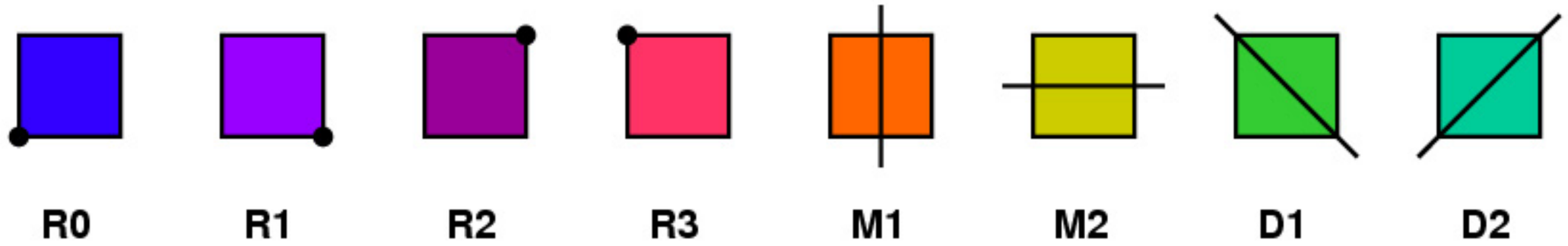


➤ They form a *Group*.

(from http://www.cs.umb.edu/~eb/)

DISCRETE SYMMETRIES

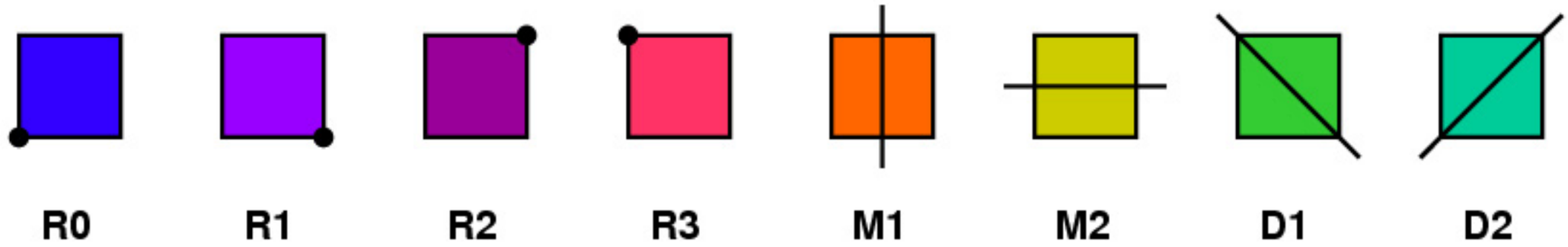
- Which transformations leave this square unchanged?



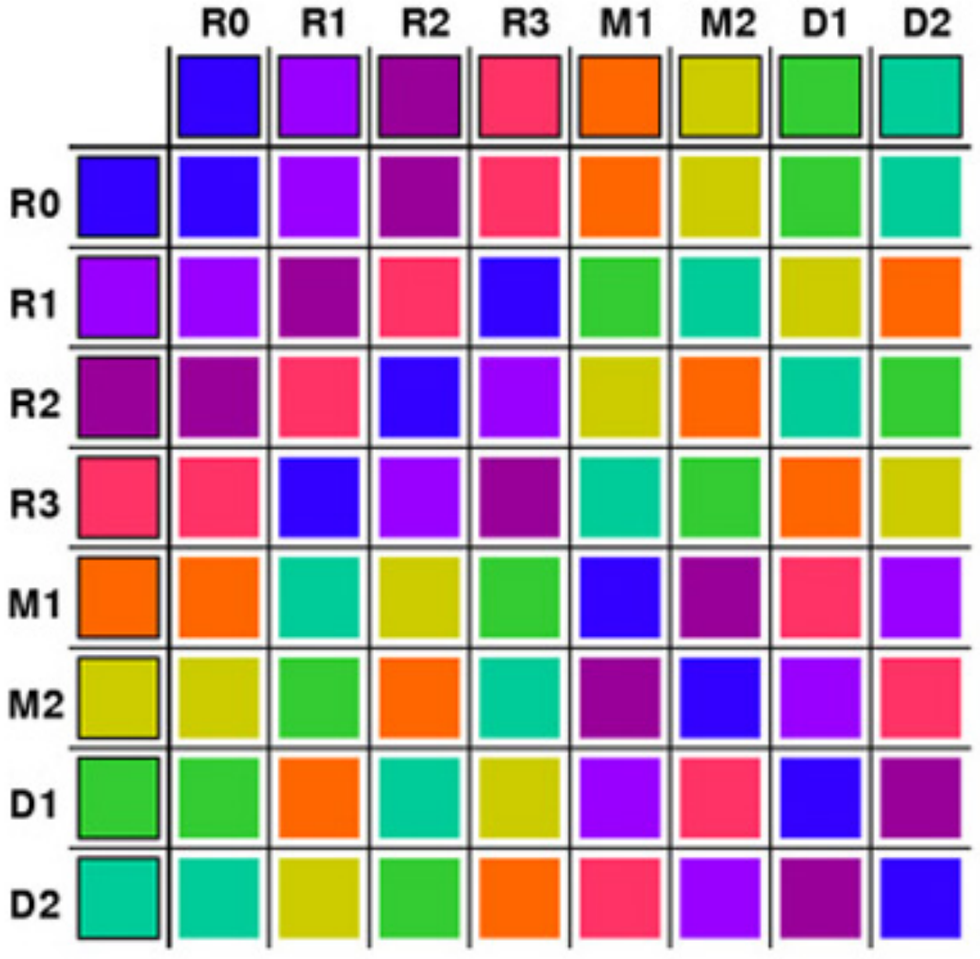
- The set of all symmetries forms a *group* G :
  - group operation:
    $$\forall\ g_1, g_2 \in G, \quad g_1 \cdot g_2 \in G\ .$$
  - identity element:
    $$\exists e \in G\ s.t.\ g \cdot e = e \cdot g = g\ \ \forall\, g \in G\ .$$
  - inverse: $\forall\ g \in G\ \exists\, g^{-1} \in G\ s.t.\ g \cdot g^{-1} = e\ .$

DISCRETE SYMMETRIES

- Which transformations leave this square unchanged?



- Discrete groups are completely characterized by their multiplication table:



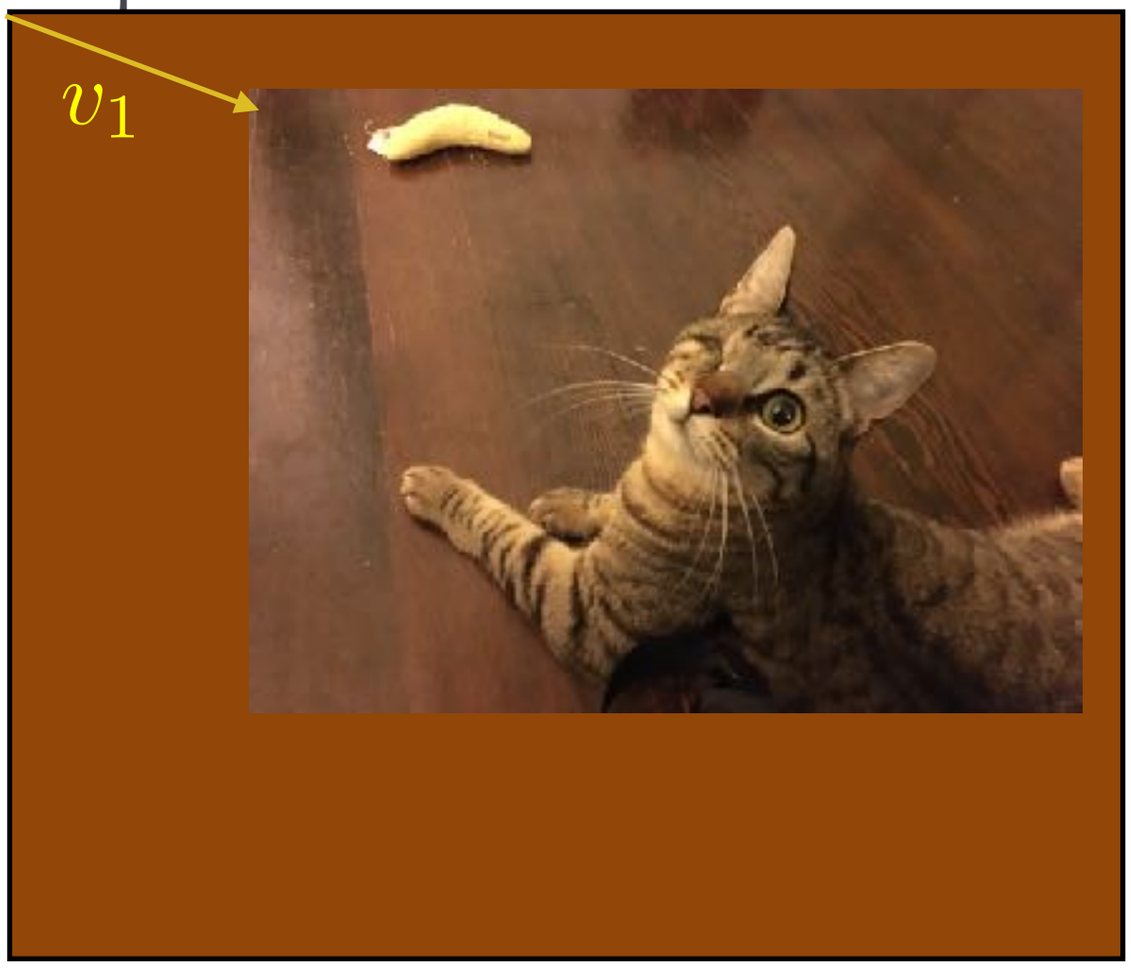
(from http://www.cs.umb.edu/~eb/)

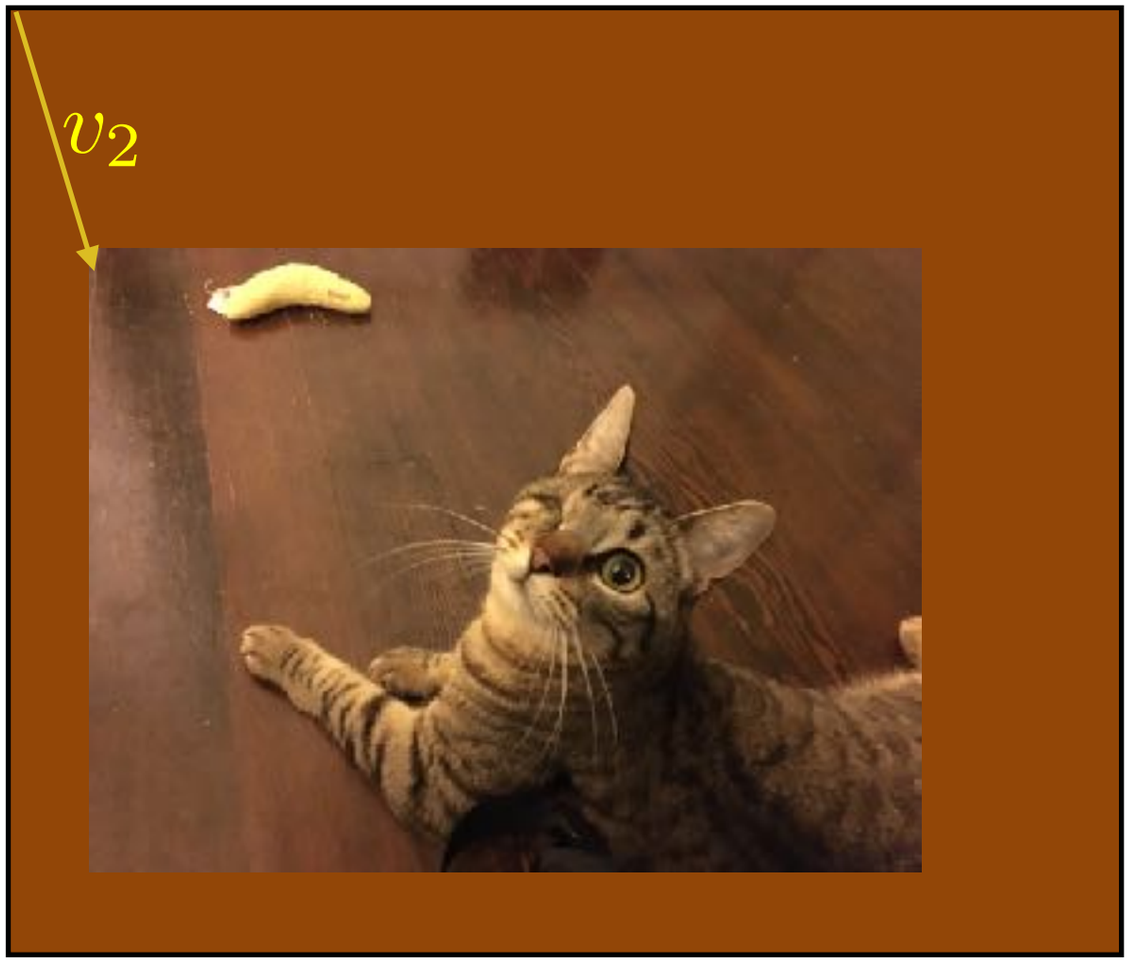
RIGID TRANSFORMATION SYMMETRIES

- Which symmetries are we likely to find in image recognition problems?

RIGID TRANSFORMATION SYMMETRIES

- Which symmetries are we likely to find in image recognition problems?

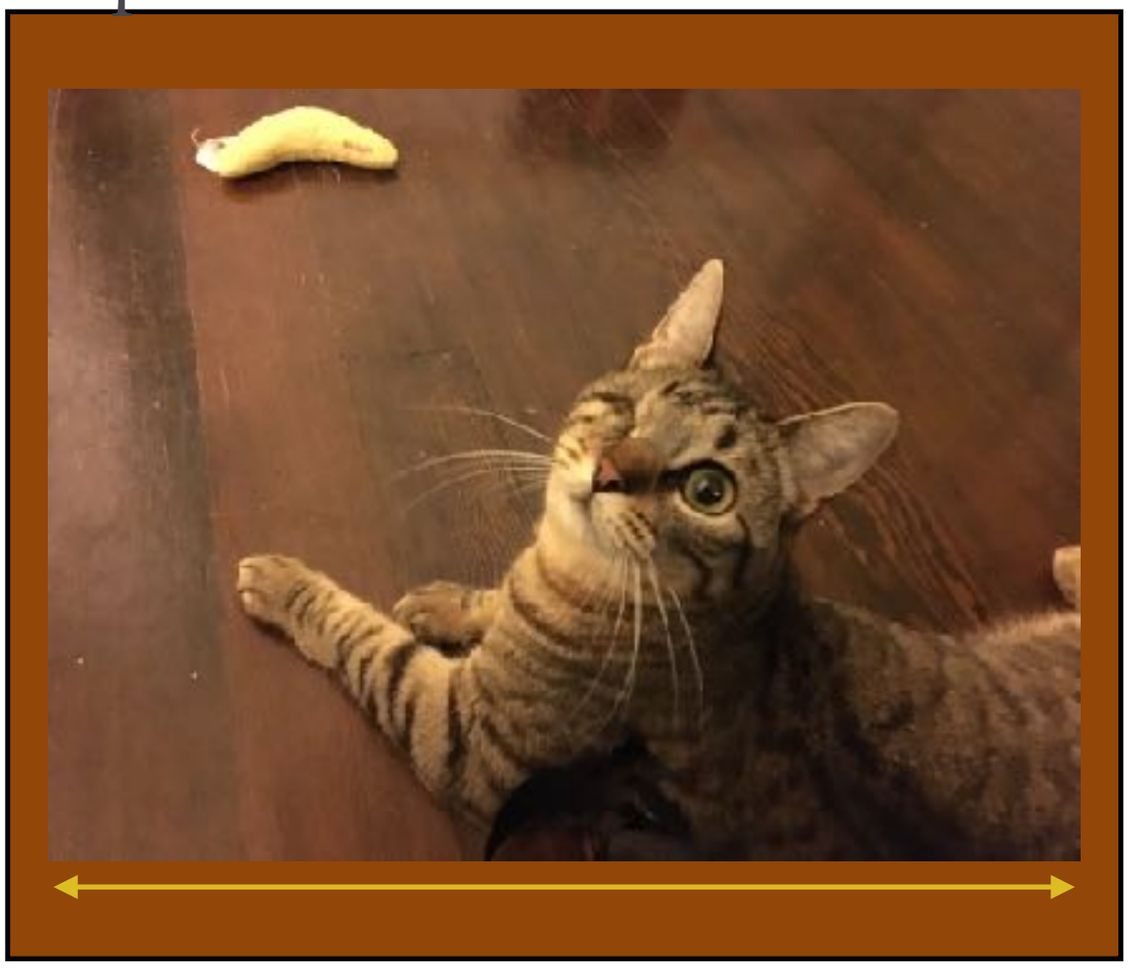


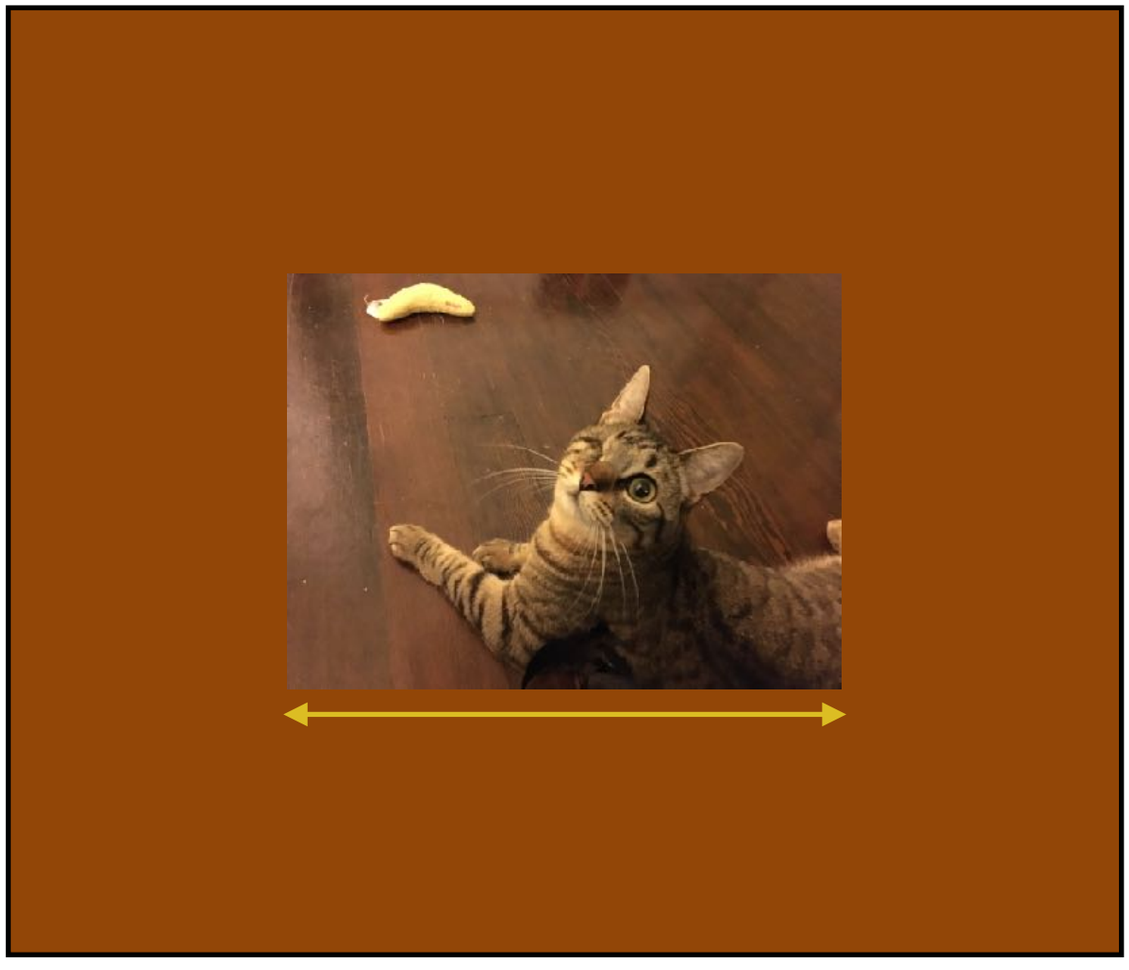


Translations: $\{\varphi_v\,; v \in \mathbb{R}^2\}$, with $\varphi_v(x)(u) = x(u - v)$.

RIGID TRANSFORMATION SYMMETRIES

- Which symmetries are we likely to find in image recognition problems?

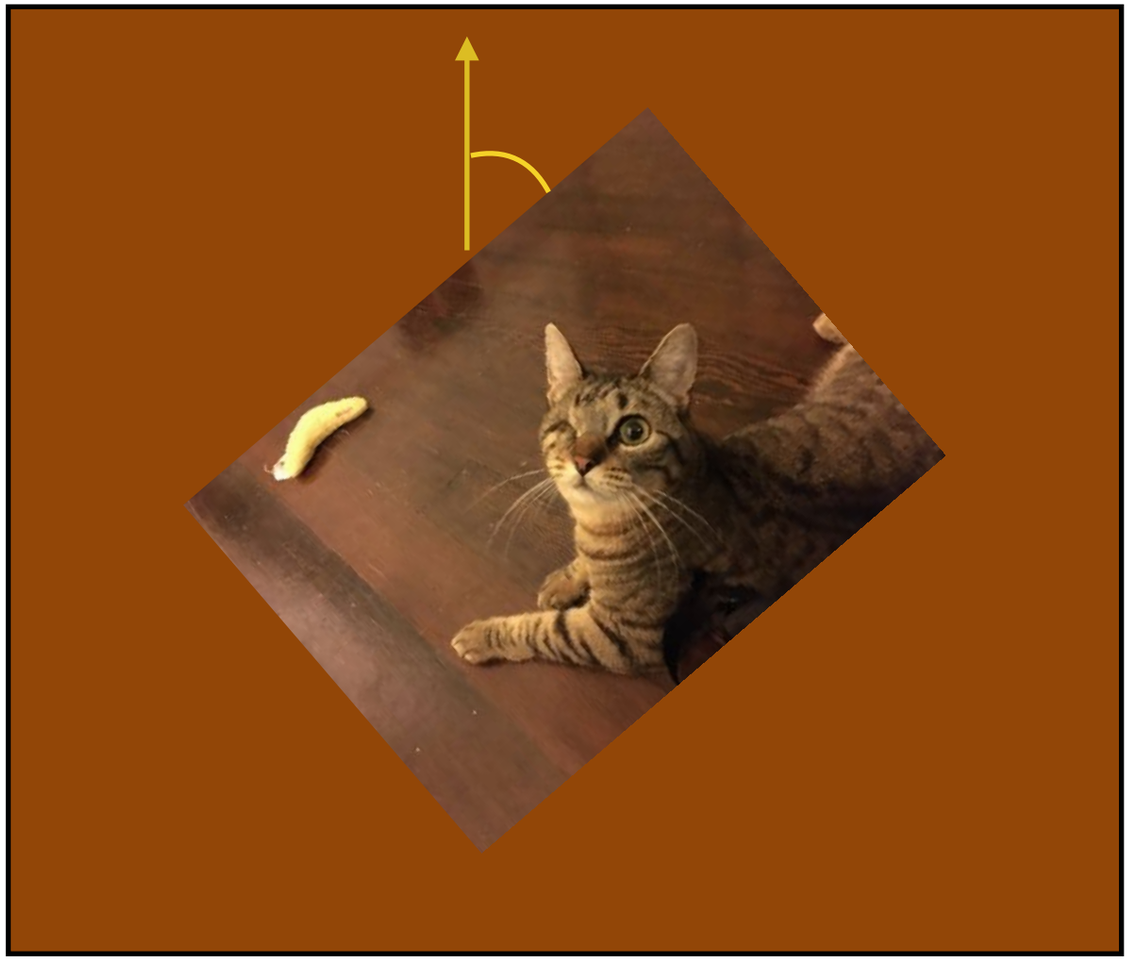




Dilations: $\{\varphi_s\,;s\in\mathbb{R}_+\}$, with $\varphi_s(x)(u)=s^{-1}x(s^{-1}u)$.

RIGID TRANSFORMATION SYMMETRIES

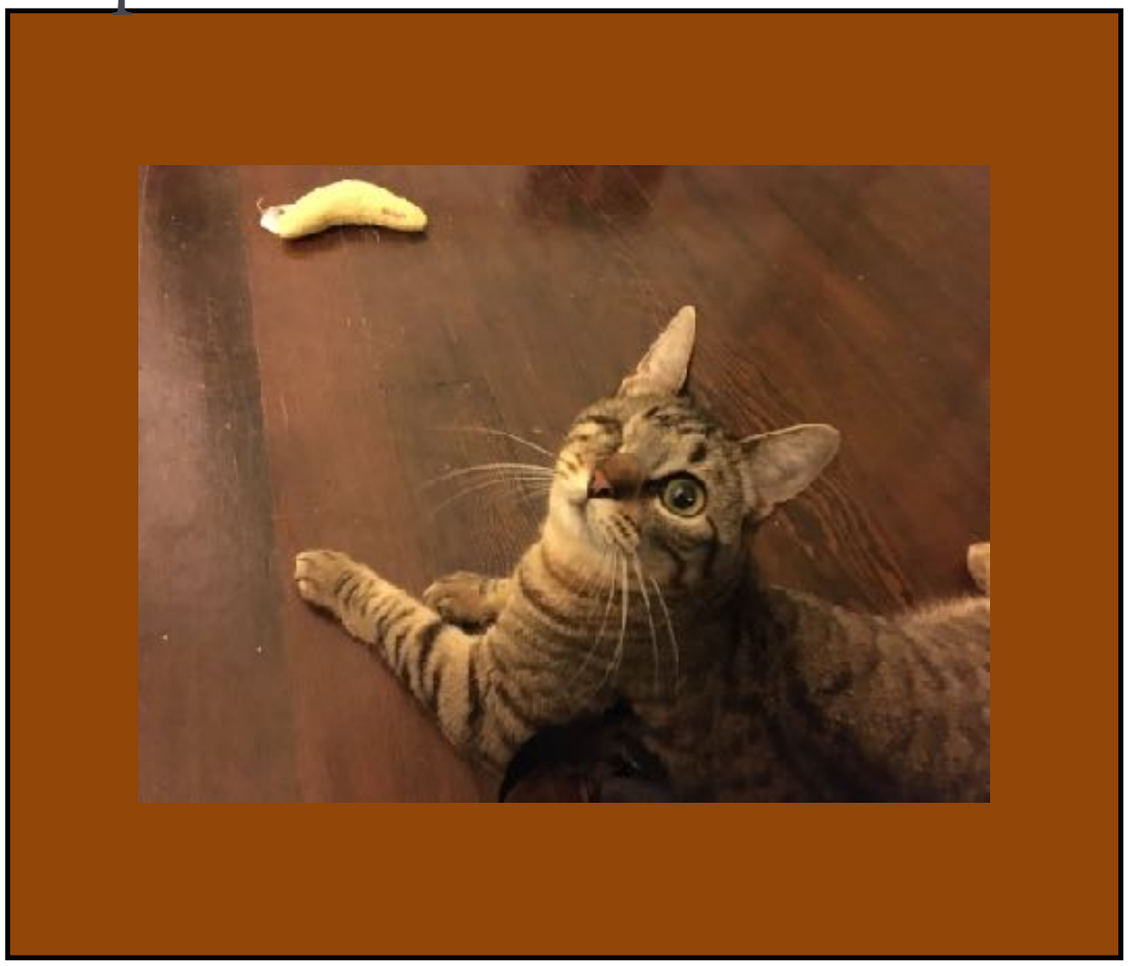
- Which symmetries are we likely to find in image recognition problems?

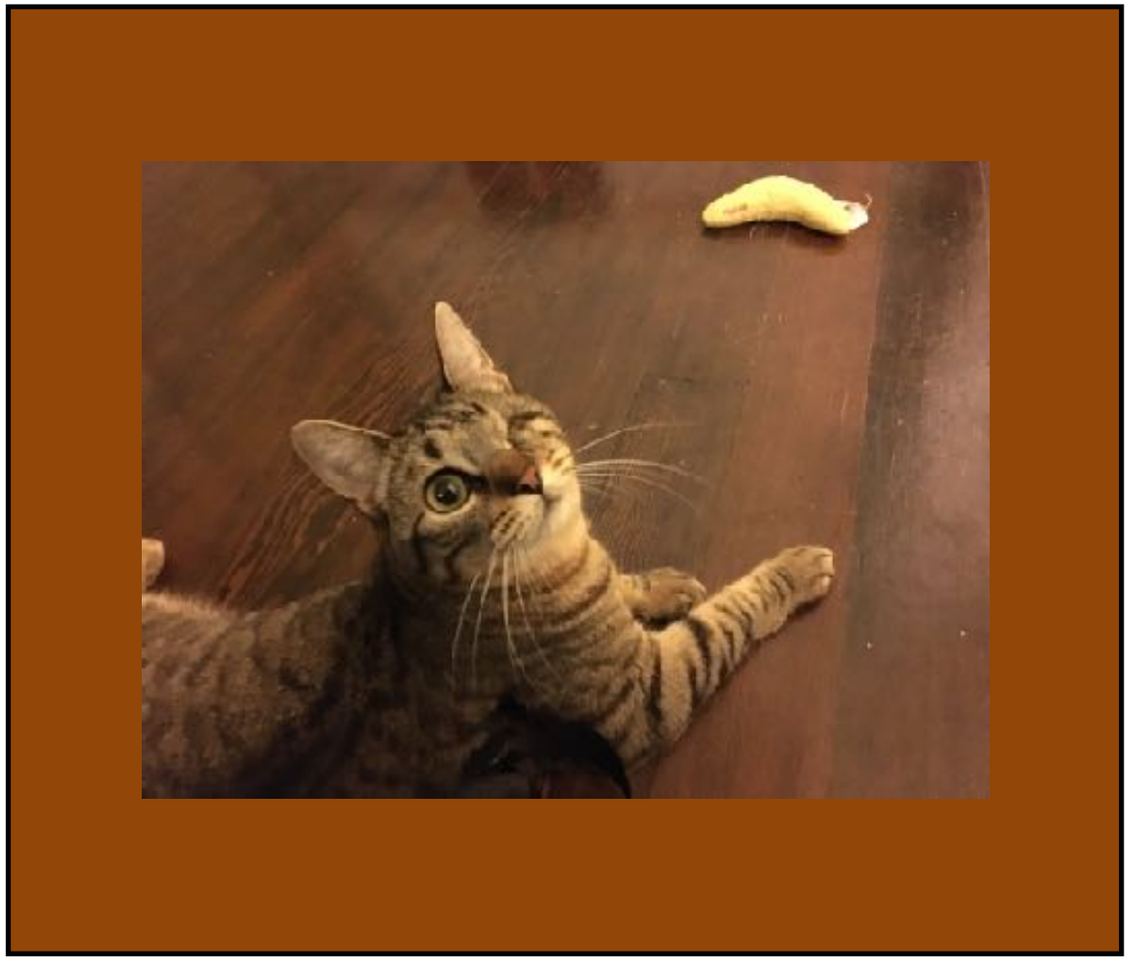


Rotations: $\{\varphi_\theta\,;\theta \in [0, 2\pi)\}$, with $\varphi_\theta(x)(u) = x(R_\theta u)$.

RIGID TRANSFORMATION SYMMETRIES

- Which symmetries are we likely to find in image recognition problems?





Mirror symmetry: $\{e\,, M\}$, with $Mx(u_1, u_2) = x(-u_1, u_2)$.

RIGID TRANSFORMATION SYMMETRIES

- We can combine all these transformations into a single group, the Affine Group $\mathbf{Aff}(\mathbb{R}^2)$.
- It has 6 degrees of freedom; in the representation

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \mapsto \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

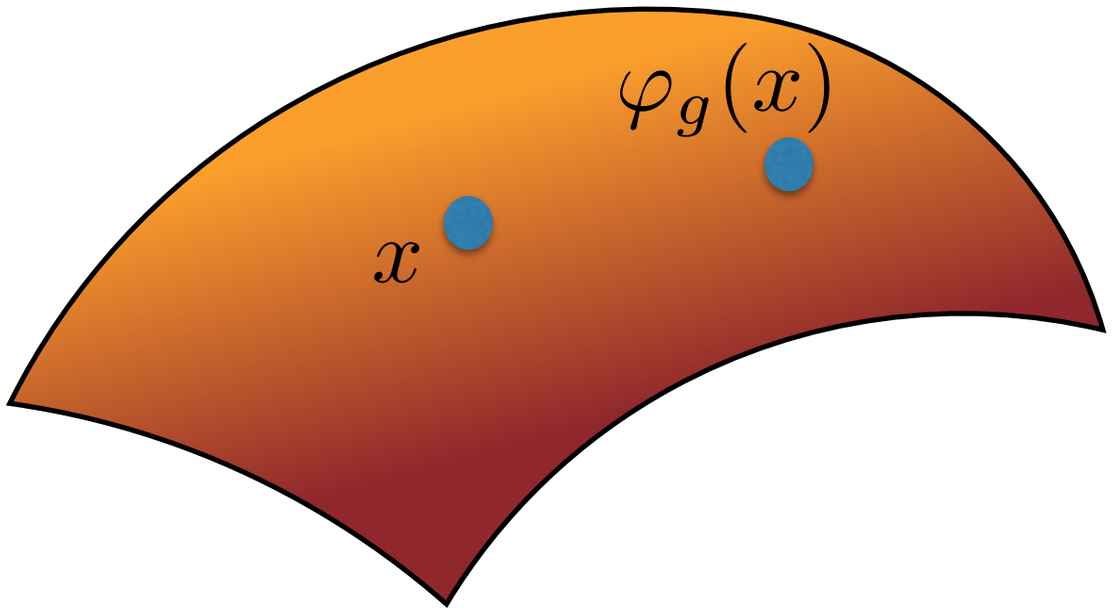
$$g = (v_1, v_2, a_1, a_2, a_3, a_4)$$

- Note that this is in general a *non-commutative* group.
- For some groups, we might only observe partial invariance (e.g. rotation and dilation).
- In speech, the underlying group modeling time-frequency shifts is the *Heisenberg* group.

INVARIANT REPRESENTATIONS

- Given a transformation group G and an input x, the *action* of G onto x is called an *orbit*:

$$G \cdot x = \{\varphi_g(x)\,, g \in G\}$$

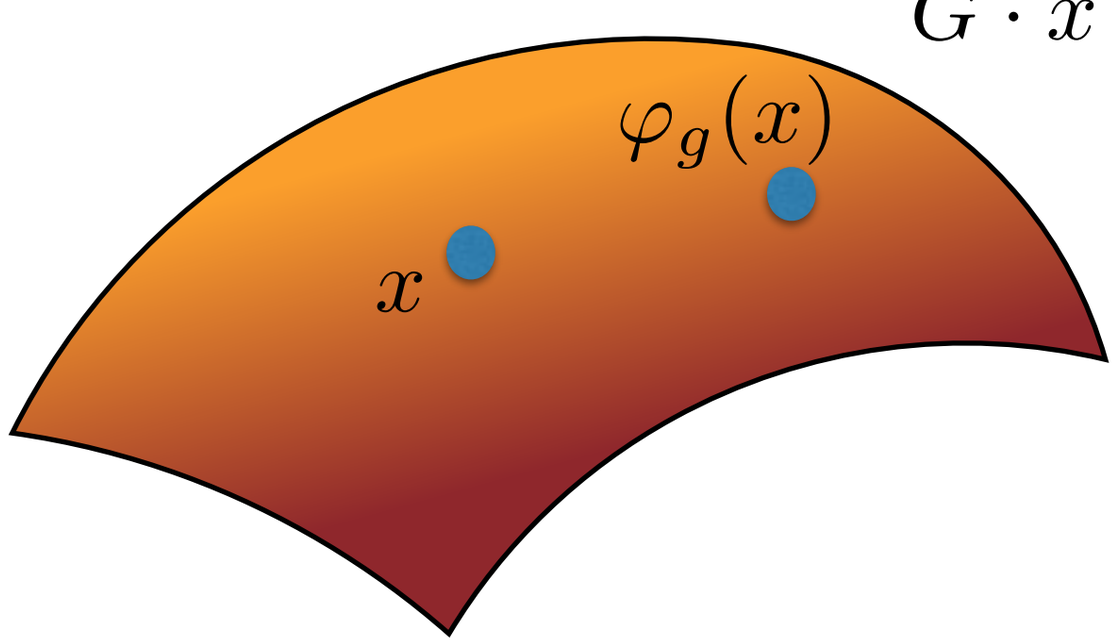


- Impact on the learning task?
- Since our estimator is linear in $\Phi(x)$, $\Phi(G \cdot x)$ should be “flat”.

INVARIANT REPRESENTATIONS

- Problem?

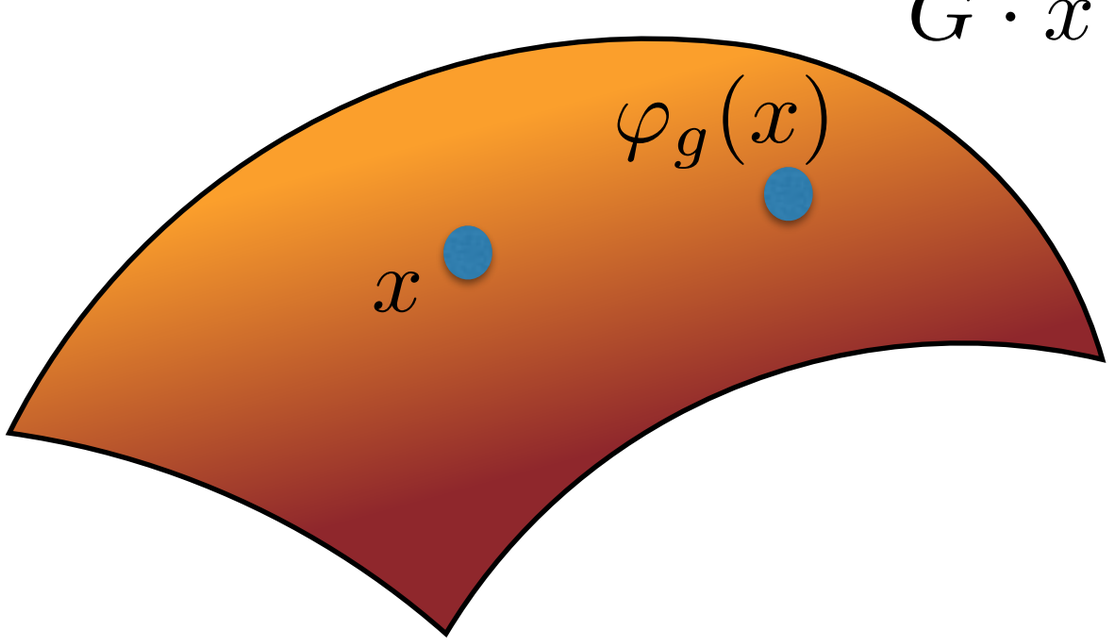
$$G \cdot x = \{\varphi_g(x)\,, g \in G\}$$



INVARIANT REPRESENTATIONS

- Problem? A 6-dimensional curvy space looks flat in a high-dimensional space.

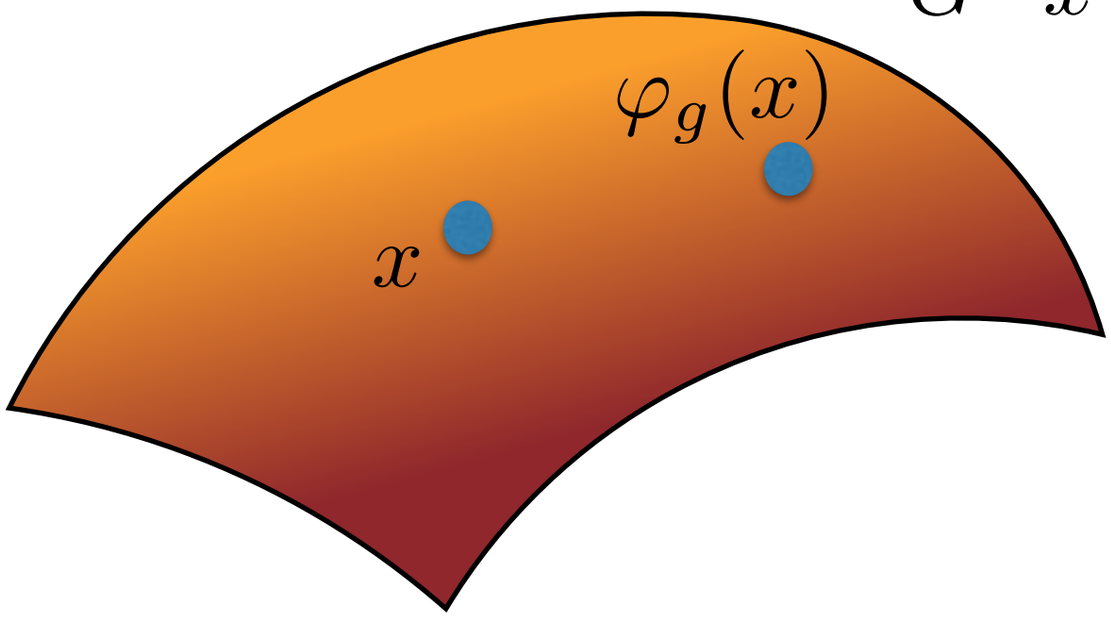
$$G \cdot x = \{\varphi_g(x), g \in G\}$$



INVARIANT REPRESENTATIONS

- Problem? A 6-dimensional curvy space looks flat in a high-dimensional space.
- Group symmetries are not sufficient to beat the curse of dimensionality.

$$G \cdot x = \{\varphi_g(x)\,, g \in G\}$$



FROM INVARIANCE TO STABILITY

- Symmetry is a very strict criteria. Can we relax it?

FROM INVARIANCE TO STABILITY

- Symmetry is a very strict criteria. Can we relax it?

- Although image and audio recognition does not have high-dimensional symmetry groups, it is *stable* to local deformations.

$$x \in L^2(\mathbb{R}^m) \ , \ \tau : \mathbb{R}^m \to \mathbb{R}^m \text{ diffeomorphism}$$

$$x_\tau = \varphi_\tau(x) \ , \ x_\tau(u) = x(u - \tau(u))$$

φ_τ is a change of variables: (think of x_τ as adding noise to the pixel *locations* rather than to the pixel values)

FROM INVARIANCE TO STABILITY

- Informally, if $\|\tau\|$ measures the amount of deformation, many recognition tasks satisfy

$$\forall\ x, \tau,\ \ |f(x) - f(x_\tau)| \lesssim \|\tau\|$$

FROM INVARIANCE TO STABILITY

- Informally, if $\|\tau\|$ measures the amount of deformation, many recognition tasks satisfy

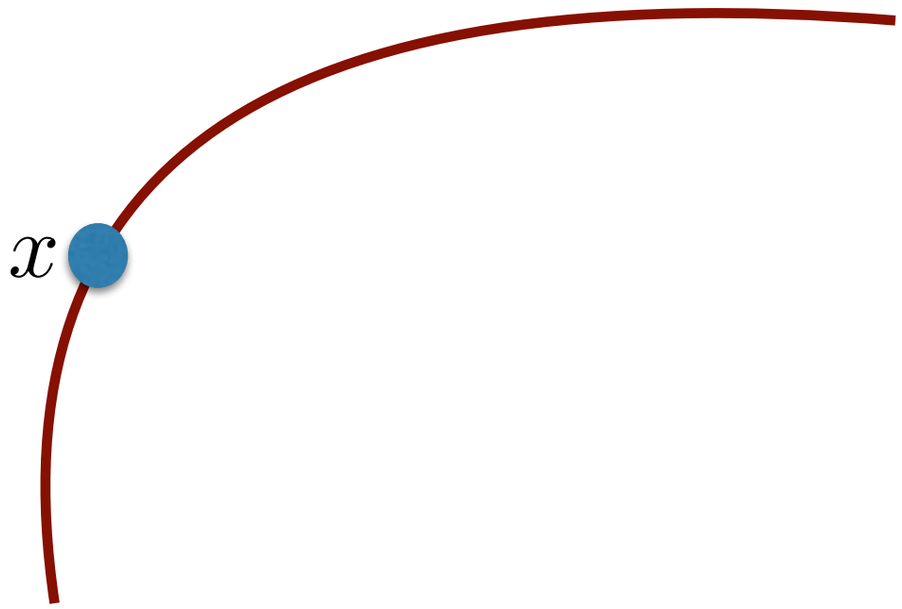
$$\forall\ x, \tau,\ \ |f(x) - f(x_\tau)| \lesssim \|\tau\|$$

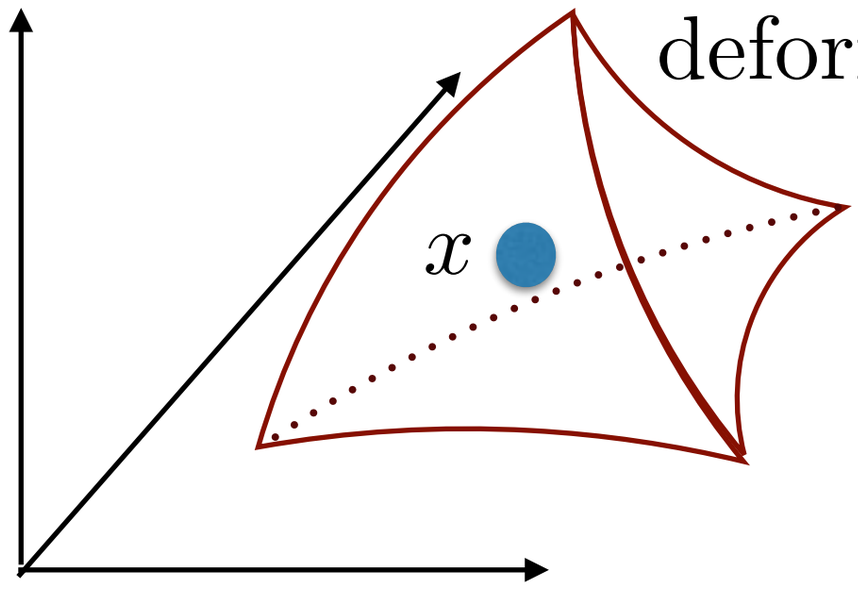
- If our representation is stable, then

$$\forall\ x, \tau,\ \ \|\Phi(x) - \Phi(x_\tau)\| \leq C\|\tau\| \implies |\hat{f}(x) - \hat{f}(x_\tau)| \leq \tilde{C}\|\tau\|$$

FILLING THE SPACE WITH DEFORMATIONS

symmetry group: low dimension





deformations fill the space

GEOMETRIC STABILITY CONDITION

- We introduced the stability condition

$$\forall\ x, \tau,\ \ \|\Phi(x) - \Phi(x_\tau)\| \lesssim \|\tau\|\ .$$

- If we fix the 'template' x and consider the mapping

$$F\ :\ \tau \mapsto \Phi(x_\tau)$$

the previous condition becomes

$$\|F(\tau) - F(0)\| \leq C\|\tau\|\ ,$$

thus F is Lipschitz with respect to the deformation metric $\|\tau\|$ *uniformly* on x.

TRANSFORMATION GROUPS

- We discussed about "universal" transformation groups acting on images, audio and video:

Translations: $\{\varphi_v\,; v \in \mathbb{R}^2\}$, with $\varphi_v(x)(u) = x(u - v)$.

Dilations: $\{\varphi_s\,; s \in \mathbb{R}_+\}$, with $\varphi_s(x)(u) = s^{-1}x(s^{-1}u)$.

Rotations: $\{\varphi_\theta\,; \theta \in [0, 2\pi)\}$, with $\varphi_\theta(x)(u) = x(R_\theta u)$.

- Systematic approach to obtain representations invariant to these groups?

ONE-PARAMETER UNITARY GROUPS

- A particularly simple example is given by *continuous one-parameter unitary transformations*:

Definition: A one-parameter unitary group $\{\varphi_t \in Aut(\Omega)\}_{t\in\mathbb{R}}$ satisfies

1. $\forall\ t, s,\ \varphi_{s+t} = \varphi_t\varphi_s\ ,$
2. $\lim_{s\to t} \|\varphi_s - \varphi_t\| = 0,$
3. $\forall\ t \in \mathbb{R}\,,\ x \in \Omega\ ,\ \|\varphi_t x\| = \|x\|\ .$

ONE-PARAMETER UNITARY GROUPS

- A particularly simple example is given by *continuous one-parameter unitary transformations*:

Definition: A one-parameter unitary group $\{\varphi_t \in Aut(\Omega)\}_{t\in\mathbb{R}}$ satisfies

1. $\forall\ t, s,\ \varphi_{s+t} = \varphi_t\varphi_s\ ,$
2. $\lim_{s\to t} \|\varphi_s - \varphi_t\| = 0,$
3. $\forall\ t \in \mathbb{R}\,,\ x \in \Omega\ ,\ \|\varphi_t x\| = \|x\|\ .$

- In particular, these are Abelian groups.
  - Rotations and Translations are 1-parameter unitary groups
  - Dilations can be made unitary: $\varphi_s x(u) = s^{1/2} x(su)\ .$

STONE'S THEOREM

Theorem: Suppose Ω is a Hilbert space. There is a one-to-one correspondence between self-adjoint operators on Ω and one-parameter unitary groups of $Aut(\Omega)$.
Given $\{\varphi_t\}_{t\in\mathbb{R}}$, there exists A self-adjoint such that $\forall\, t\ ,\ \varphi_t = e^{itA}$. Conversely, if A is self-adjoint, the family $\{e^{itA}\}_t$ is a one-parameter unitary group.

Remark: In finite dimensions, we define the matrix exponential e^A, $A \in \mathbb{C}^{n\times n}$, as $e^A := \sum_{k\geq 0} \frac{A^k}{k!}$.

FOURIER TRANSFORM DEFROST

Definition The Fourier transform of a function $x \in L^2(\mathbb{R})$ is defined as

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

[Main Properties]:

- Linear: $z = \alpha x + \beta y \Longrightarrow \hat{z} = \alpha \hat{x} + \beta \hat{y}$.
- Parseval identity: $\|\hat{x}\| = \|x\|$, $\langle x, y \rangle = \langle \hat{x}, \hat{y} \rangle$.
- Inverse Fourier transform: $x(u) = \int \hat{x}(\omega) e^{i\omega u} d\omega$.
- Translation: $y(u) = x(u - u_0) \Longrightarrow \hat{y}(\omega) = e^{i\omega u_0} \hat{x}(\omega)$.
- Dilation: $y(u) = x(su)$ for $s > 0 \Longrightarrow \hat{y}(\omega) = s^{-1} \hat{x}(s^{-1}\omega)$.

STONE THEOREM, FOURIER AND GLOBAL INVARIANTS

- Translations are simultaneously diagonalized by Fourier atoms.

STONE THEOREM, FOURIER AND GLOBAL INVARIANTS

- Translations are simultaneously diagonalized by Fourier atoms.
- The Stone theorem formalizes the fact that a collection of "nice" commuting operators simultaneously diagonalizes (in a complex basis):

$$A = V^* \mathrm{diag}(\lambda_1, \ldots, \lambda_n) V$$

- Unitary condition implies that eigenvalues are unitary complex numbers.

STONE THEOREM, FOURIER AND GLOBAL INVARIANTS

- Translations are simultaneously diagonalized by Fourier atoms.
- The Stone theorem formalizes the fact that a collection of "nice" commuting operators simultaneously diagonalizes (in a complex basis):

$$A = V^* \mathrm{diag}(\lambda_1, \ldots, \lambda_n) V$$

  - Unitary condition implies that eigenvalues are unitary complex numbers.
- What happens on larger Abelian (commuting) groups?
  - Factorization of Abelian groups into one-parameter groups (eg translations in R2)

$$G = G_1 \times G_2 \times \ldots G_l$$

STONE THEOREM, FOURIER AND GLOBAL INVARIANTS

- Translations are simultaneously diagonalized by Fourier atoms.
- The Stone theorem formalizes the fact that a collection of "nice" commuting operators simultaneously diagonalizes (in a complex basis):

$$A = V^* \mathrm{diag}(\lambda_1, \ldots, \lambda_n) V$$

  - Unitary condition implies that eigenvalues are unitary complex numbers.
- What happens on larger Abelian (commuting) groups?
  - Factorization of Abelian groups into one-parameter groups (eg translations in R2)

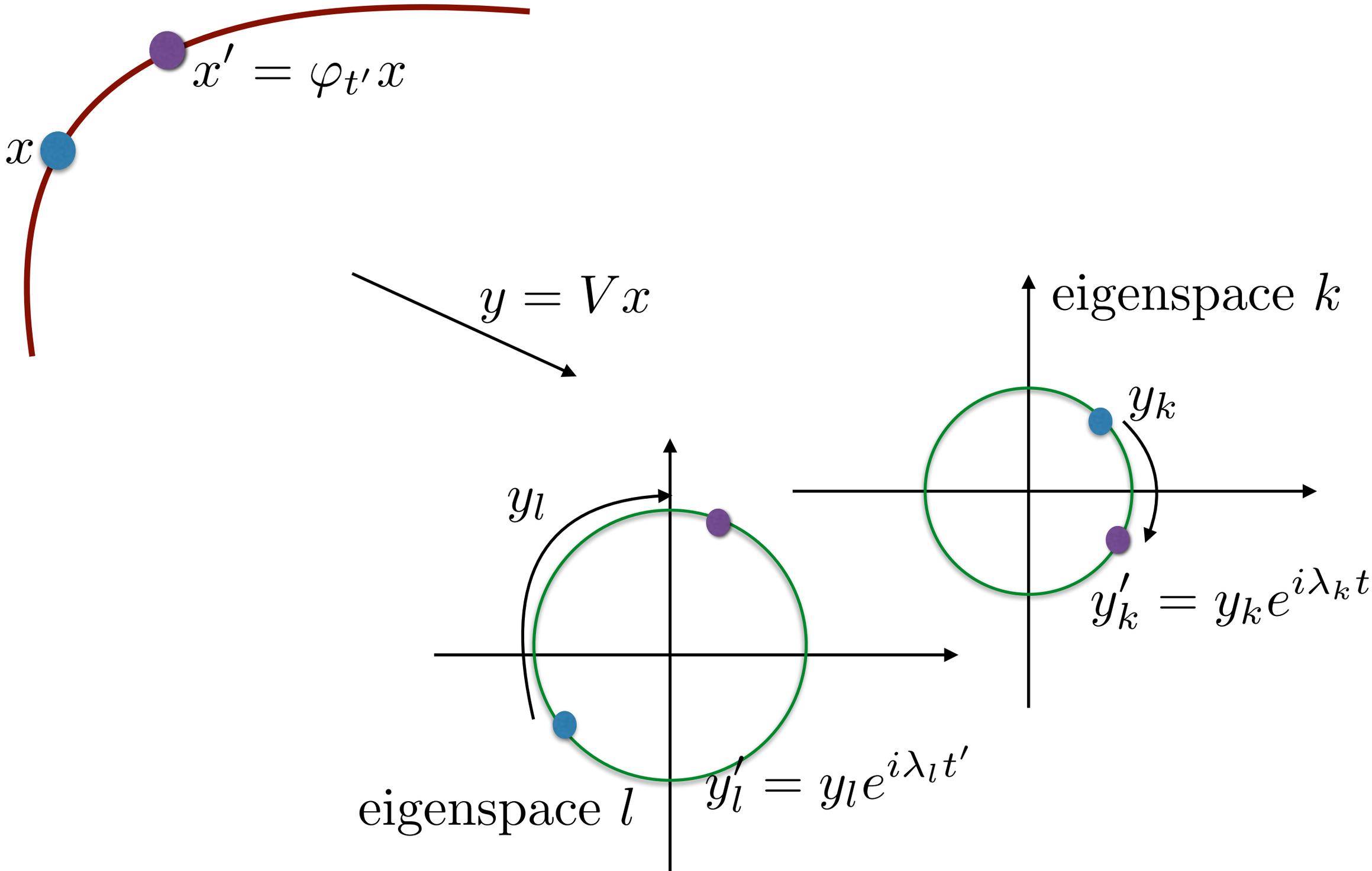
$$G = G_1 \times G_2 \times \ldots G_l$$

- Q: How to obtain global invariants in that case?

STONE THEOREM, FOURIER AND GLOBAL INVARIANTS

$\{\varphi_t x\}_{t\in\mathbb{R}}$ one-parameter group

$$A = V^* \mathrm{diag}(\lambda_1, \ldots, \lambda_n) V$$



STONE THEOREM, FOURIER AND GLOBAL INVARIANTS

- Thus $\Phi(x) = |Vx|$ satisfies

$$\forall\ x, t\ ,\ \ \Phi(\varphi_t(x)) = \Phi(x)\ .$$

STONE THEOREM, FOURIER AND GLOBAL INVARIANTS

- Thus $\Phi(x) = |Vx|$ satisfies

$$\forall\ x, t\ ,\ \ \Phi(\varphi_t(x)) = \Phi(x)\ .$$

- Indeed,

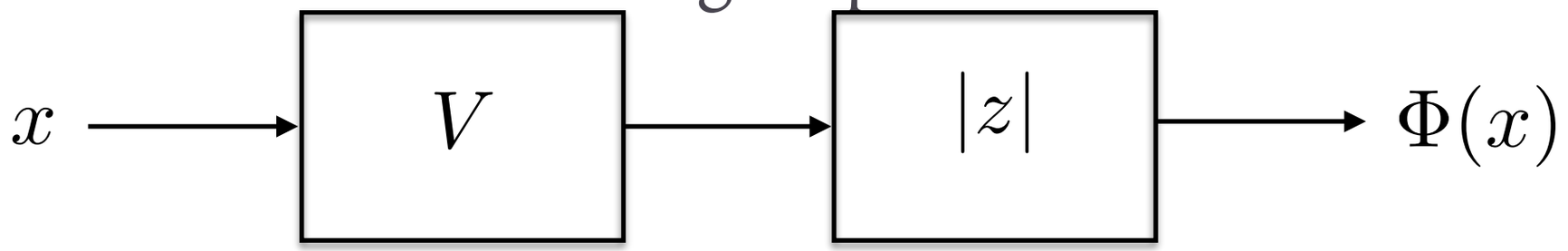
$$A = V^* \mathrm{diag}(\lambda_1, \ldots, \lambda_n) V \implies e^{itA} = V^* \mathrm{diag}(e^{it\lambda_1}, \ldots, e^{it\lambda_n}) V\ .$$

$$\begin{aligned} V\varphi_t x = V e^{itA} x &= V V^* \mathrm{diag}(e^{it\lambda_1}, \ldots, e^{it\lambda_n}) V x \\ &= \mathrm{diag}(e^{it\lambda_1}, \ldots, e^{it\lambda_n}) V x \end{aligned}$$

$$\text{thus } \Phi(\varphi_t x) = |V\varphi_t x| = |Vx|\ .$$

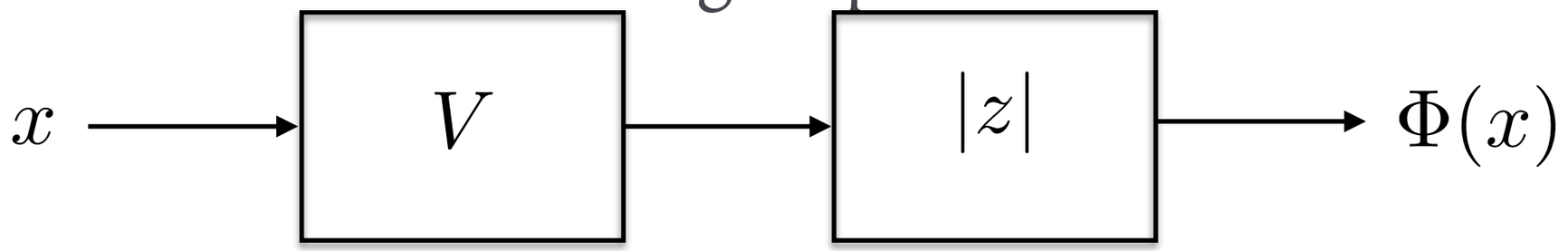
LIMITS OF GROUP DIAGONALISATION

- A shallow (1 layer) network is thus sufficient to achieve invariance to commutative group transformations:



LIMITS OF GROUP DIAGONALISATION

- A shallow (1 layer) network is thus sufficient to achieve invariance to commutative group transformations:



- However, this architecture has a number of shortcomings.

LIMITS OF GROUP DIAGONALISATION

- Non-commutative Groups:

Proposition: If $G = \{\varphi_t\}_t$ is non-commutative, then there is no basis V that diagonalises simultaneously all φ_t.

LIMITS OF GROUP DIAGONALISATION

- Non-commutative Groups:

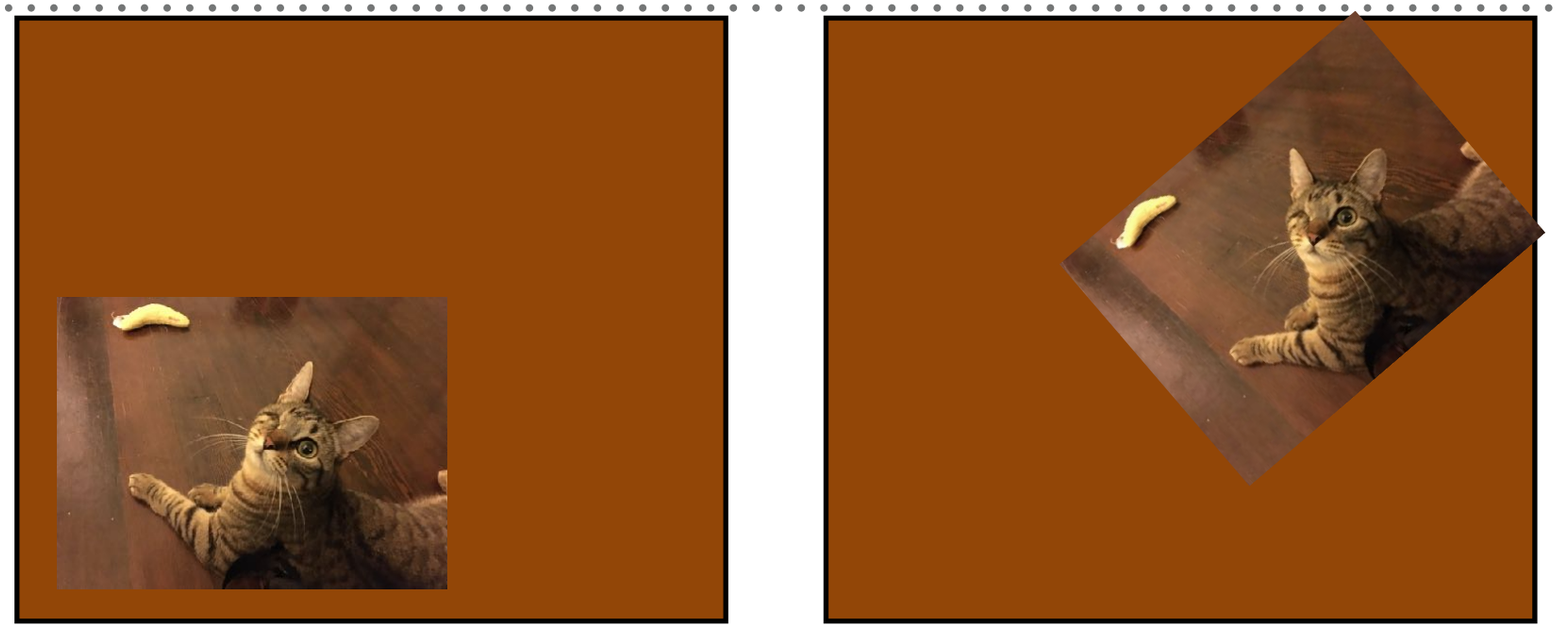
Proposition: If $G = \{\varphi_t\}_t$ is non-commutative, then there is no basis V that diagonalises simultaneously all φ_t.

Square matrices A and B commute

$\Updownarrow$

A and B share the same eigenvectors.

EXAMPLE: THE ROTO-TRANSLATION GROUP



Roto-translation group: $\{\varphi_{v,\theta}\,; v \in \mathbb{R}^2,\ \theta \in [0, 2\pi)\}$.

$$\varphi_{v,\theta} \ : \ u \mapsto R_\theta(u - v) \ .$$

$$\begin{aligned}
\varphi_{v',\theta'} \cdot \varphi_{v,\theta} u &= R_{\theta'}(\varphi_{v,\theta} u - v') = R_{\theta'}(R_\theta u - R_\theta v - v') \\
&= R_{\theta'} R_\theta u - (R_{\theta'} R_\theta v + R_{\theta'} v') \\
&= R_{\theta+\theta'}\,(u - (v + R_{-\theta} v'))
\end{aligned}$$

$$\text{Thus } (v', \theta') \cdot (v, \theta) = (v + R_{-\theta} v', \theta + \theta')$$

- We will see later how to deal with such groups.

LIMITS OF GROUP DIAGONALISATION

- Stable to deformations?

LIMITS OF GROUP DIAGONALISATION

- Stable to deformations?

- The diagonalisation ensures that $\Phi(\varphi_t x) = \Phi(x)\ \forall\, t, x$, but we have no control outside the group $\{\varphi_t\}_t$ in general.

LIMITS OF GROUP DIAGONALISATION

- Stable to deformations?

- The diagonalisation ensures that $\Phi(\varphi_t x) = \Phi(x)\ \forall\, t, x$, but we have no control outside the group $\{\varphi_t\}_t$ in general

- To evaluate stability, we first need to quantify the amount of deformation.
- Also, we need the notion of **scale:** in many applications, we are interested in *local* invariance rather than *global* group invariance.

DEFORMATION METRIC

Assume $\tau : \mathbb{R}^d \to \mathbb{R}^d$ differentiable, and denote

$$\varphi_\tau x(u) := x(u - \tau(u)) \ .$$

$\|\nabla\tau(u)\|$: operator norm of Jacobian of τ at u.

If $\|\nabla\tau\|_\infty = \sup_u \|\nabla\tau(u)\| < 1$,
then φ_τ is invertible, and it defines a diffeomorphism.

We consider the following deformation cost:

$$\|\tau\| := 2^{-J}\|\tau\|_\infty + \|\nabla\tau\|_\infty \ .$$

DEFORMATION METRIC

We consider the following deformation cost:

$$\|\tau\| := 2^{-J}\|\tau\|_\infty + \|\nabla\tau\|_\infty \ .$$

Scale J controls how much we pay for absolute displacements

Stability criterion: $\forall\ \|x\| = 1, \tau,\ \|\Phi(x) - \Phi(x_\tau)\| \leq C\|\tau\|$.

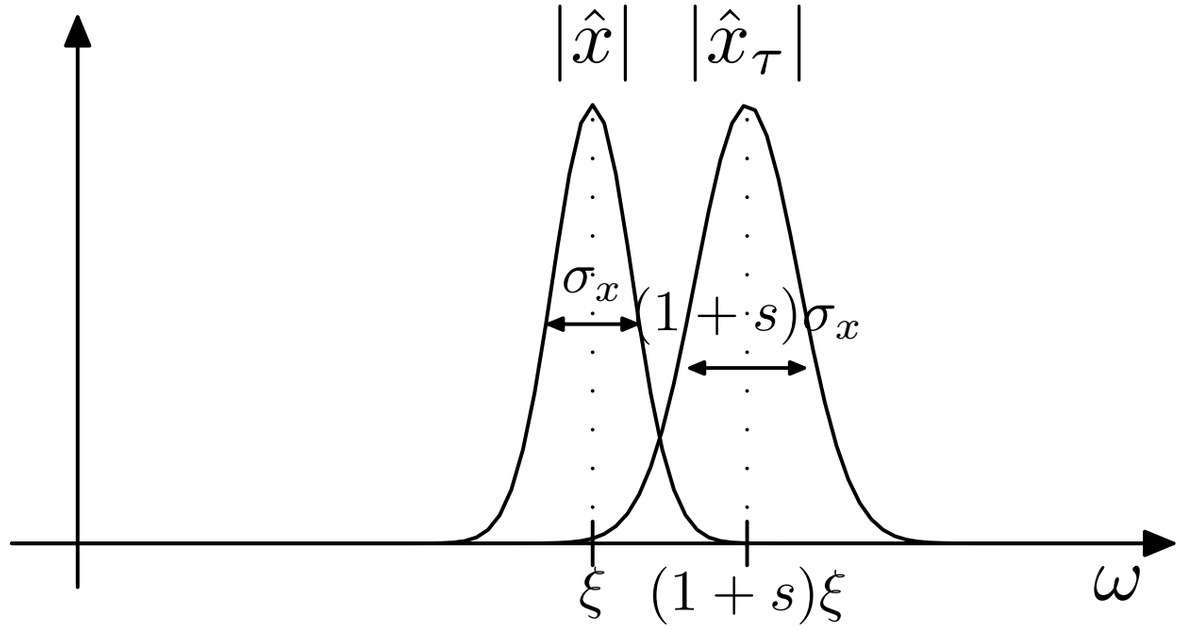
We can define similar metrics for diffeomorphisms associated with other transformation groups (e.g. rotation).

SHALLOW INVARIANTS ARE UNSTABLE

- Consider a lowpass window $h(u)$ of bandwidth σ_h and $x(u) = h(u)e^{i\xi u}$.

(bandwidth: $\sigma_h^2 = \int |\hat{h}(\omega)|^2 |\omega|^2 d\omega$.)

- Consider a deformation of the form $\varphi_\tau x(u) = x((1+s)u)$ with $s \ll 1$.

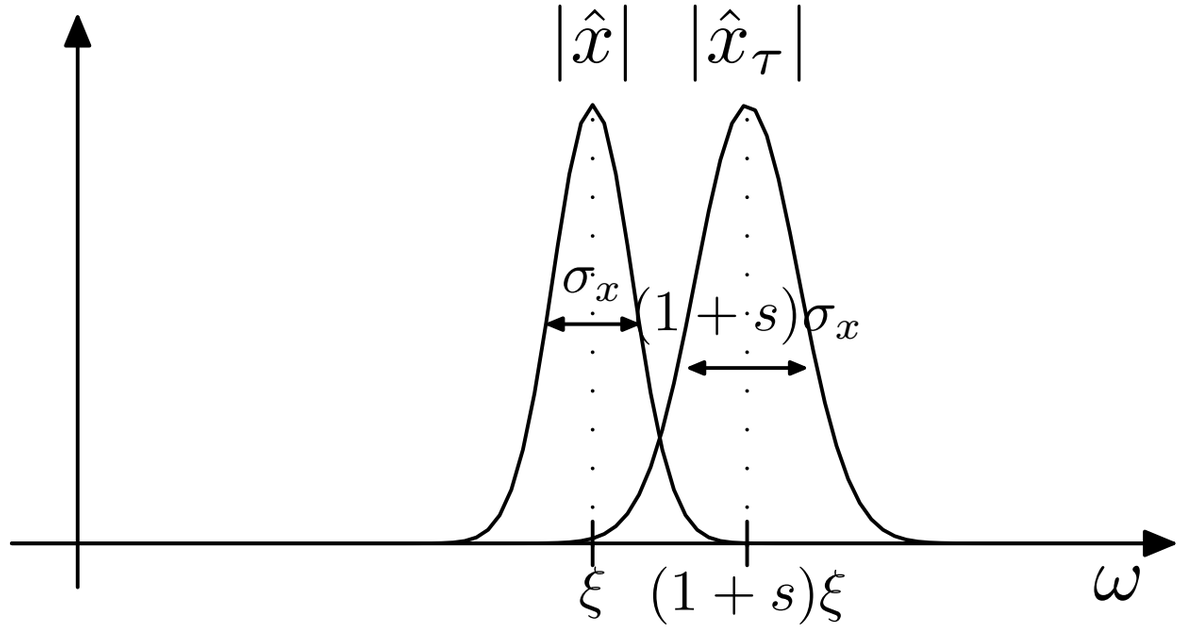


SHALLOW INVARIANTS ARE UNSTABLE

- Consider a lowpass window $h(u)$ of bandwidth σ_h and $x(u) = h(u)e^{i\xi u}$.

(bandwidth: $\sigma_h^2 = \int |\hat{h}(\omega)|^2 |\omega|^2 d\omega$.)

- Consider a deformation of the form $\varphi_\tau x(u) = x((1+s)u)$ with $s \ll 1$.



If $(1+s)\xi - \xi = s\xi \gg \sigma_h(2+s)$
(central frequency separation $\gg$ bandwidth)

$$\Longrightarrow \big\| |\hat{x}| - |\widehat{\varphi_\tau x}| \big\| \sim \|x\|$$

SHALLOW INVARIANTS ARE UNSTABLE

- Fourier Modulus is therefore unstable: high-frequency information spans a large linear subspace as soon as there is non-rigid deformation.

SHALLOW INVARIANTS ARE UNSTABLE

- Fourier Modulus is therefore unstable: high-frequency information spans a large linear subspace as soon as there is non-rigid deformation.

- Similarly, we can obtain a translation-invariant representation with the signal auto-correlation:

$$R_x(v) = \int x(u)x^*(u+v)du$$

$$\left(\|R_x - R_y\| = \|\hat{R}_x - \hat{R}_y\| = \||\hat{x}|^2 - |\hat{y}|^2\|\right)$$

- This suffers from the same problem as Fourier.

SHALLOW INVARIANTS ARE UNSTABLE

- Fourier Modulus is therefore unstable: high-frequency information spans a large linear subspace as soon as there is non-rigid deformation.

- Similarly, we can obtain a translation-invariant representation with the signal auto-correlation:

$$R_x(v) = \int x(u)x^*(u+v)du$$

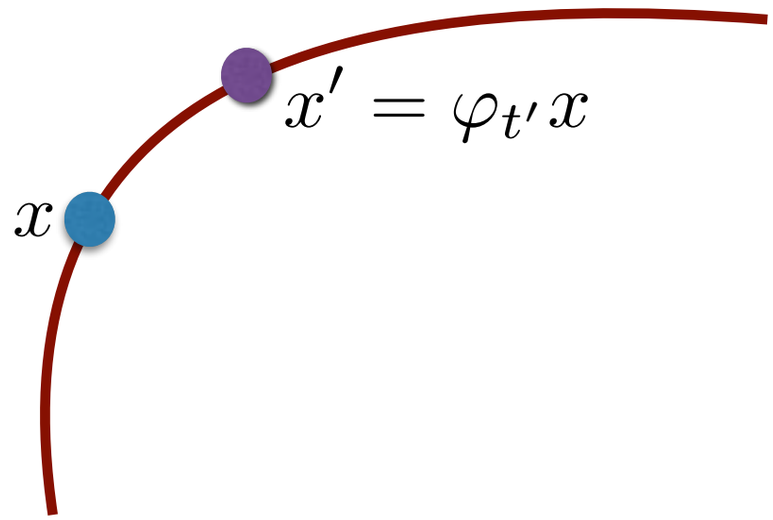
$$\left(\|R_x - R_y\| = \|\hat{R}_x - \hat{R}_y\| = \||\hat{x}|^2 - |\hat{y}|^2\|\right)$$

  - This suffers from the same problem as Fourier.

- How to fix it?

LOCAL INVARIANTS AND CONVOLUTION

- Local translation invariance:

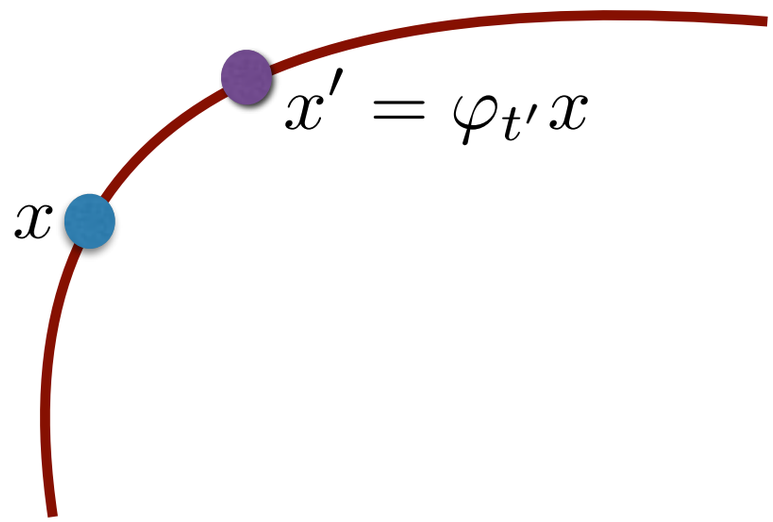


$$\|\Phi(x)-\Phi(\varphi_v x)\| \leq C2^{-J}\|v\| \text{ , or}$$

$$\forall\ v, \|x\|=1\ ,\ \frac{\|\Phi(x)-\Phi(\varphi_v x)\|}{\|v\|} \leq C2^{-J}\ .$$

LOCAL INVARIANTS AND CONVOLUTION

- Local translation invariance:



$$\|\Phi(x) - \Phi(\varphi_v x)\| \leq C 2^{-J} \|v\| \ , \text{or}$$

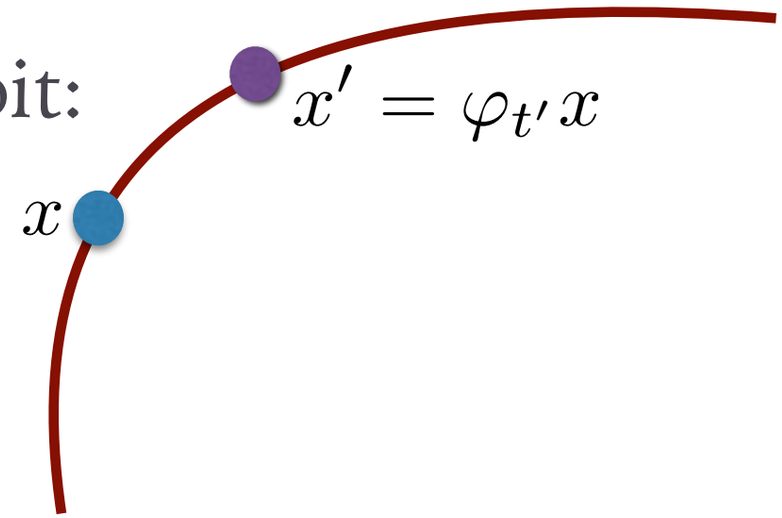
$$\forall\ v, \|x\| = 1\ , \ \frac{\|\Phi(x) - \Phi(\varphi_v x)\|}{\|v\|} \leq C 2^{-J}\ .$$

- So, we want to *smooth* along the orbits.
- Local averaging within the translation orbit:

$$\Phi(x) = 2^{-dJ} \int_v \phi(2^{-J} v) \varphi_v x dv\ , \ \left(\int \phi(v) dv = 1, \phi \geq 0 \right)\ .$$

LOCAL INVARIANTS AND CONVOLUTION

- Local averaging within the translation orbit:



- In coordinates, it becomes

$$\Phi(x)=2^{-dJ}\int_v \phi(2^{-J}v)\varphi_v x dv \ , \quad \left(\int \phi(v)dv=1, \phi\geq 0\right) \ .$$

$$\Phi(x)(u)=\int \phi_J(v)x(u-v)dv = x*\phi_J(u) \ , \text{with}$$

$$\phi_J(v)=2^{-Jd}\phi(2^{-J}v)$$

LOCAL AVERAGE AND STABILITY

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

- Not surprising, since this operator removes the problematic high-frequencies.
- Are there other linear operators with the same property?

AVERAGE AND UNIQUENESS

➤ The only linear, translation-invariant operator is the average:

$$\forall\ v\ ,\ \Phi(x) = \Phi(\varphi_v x) \Longrightarrow \Phi(x) = \frac{1}{|G|}\int \Phi(\varphi_v x)dv$$

$$\Longrightarrow \Phi(x) = \Phi\left(\frac{1}{|G|}\int \varphi_v x dv\right) = \Phi\left(\frac{1}{|G|}\int x(u)du\right)\ .$$

➤ And a similar argument can be used locally.

FROM AVERAGES TO WAVELETS

➤ Low-pass information is insufficient:

> The SIFT method originally consists in a keypoint detection phase, using a Differences of Gaussians pyramid, followed by a local description around each detected keypoint. The keypoint detection computes local maxima on a scale space generated by isotropic gaussian differences, which induces invariance to translations, rotations and

FROM AVERAGES TO WAVELETS

- Low-pass information is insufficient:

> The SIFT method originally consists in a keypoint detection phase, using a Differences of Gaussians pyramid, followed by a local description around each detected keypoint. The keypoint detection computes local maxima on a scale space generated by isotropic gaussian differences, which induces invariance to translations, rotations and

- Thus, we must capture high-frequency.
- These new measurements must involve a non-linearity.

FROM AVERAGES TO WAVELETS

- Low-pass information is insufficient:

> The SIFT method originally consists in a keypoint detection phase, using a Differences of Gaussians pyramid, followed by a local description around each detected keypoint. The keypoint detection computes local maxima on a scale space generated by isotropic gaussian differences, which induces invariance to translations, rotations and

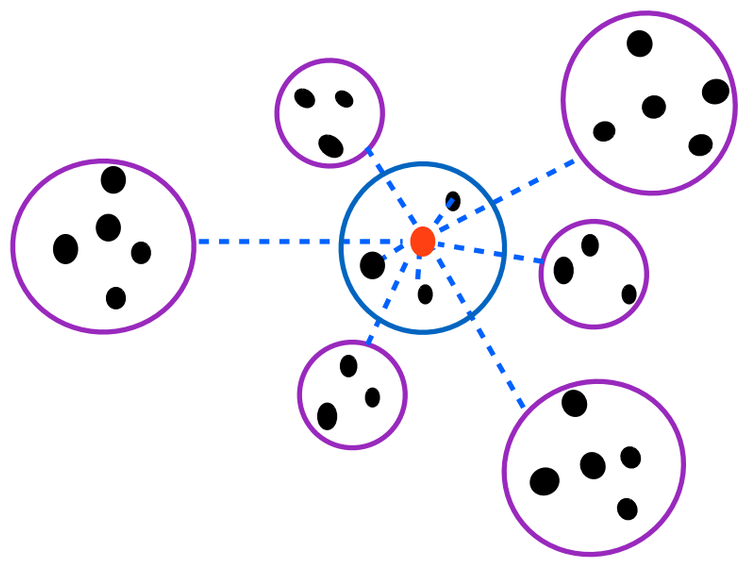
- Thus, we must capture high-frequency.
- These new measurements must involve a non-linearity.
- We want them to preserve stability to deformations.
- And we want them to preserve inter-class variability.

SCALE SEPARATION

Interactions of d variables of X: pixels, particules, agents...

- Markov: each variable interacts only with its neighbours and ignores large scale interactions.

Factorisation into multscale interactions



Multiscale regroupments reduce the number of interactions

from d to $O(\log d)$

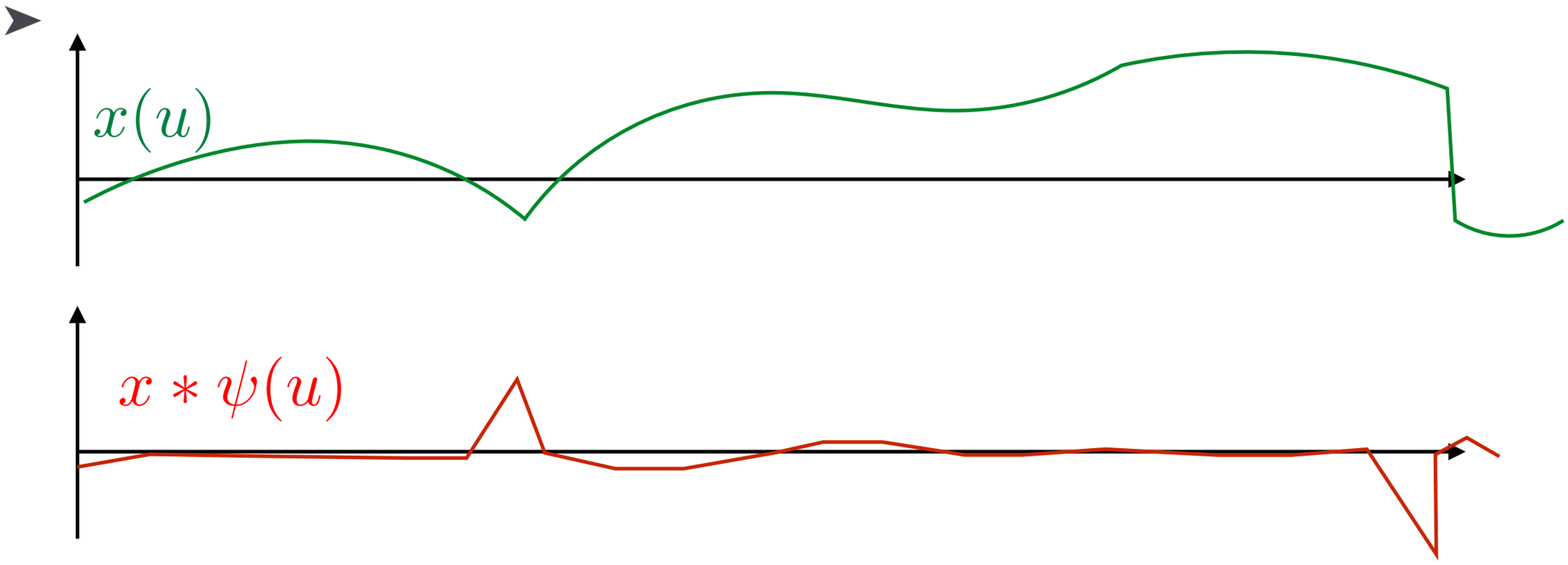
$\Rightarrow$ wavelet transforms and deep convolutional networks

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 $l < k$)

If $x(u)$ is piece-wise smooth, then $x * \psi(u)$ is mostly zero



WAVELETS

- The local average $x * \phi$ is a "blurry" version of x, whereas
- $x * \psi$ carries the details lost by the blurring.

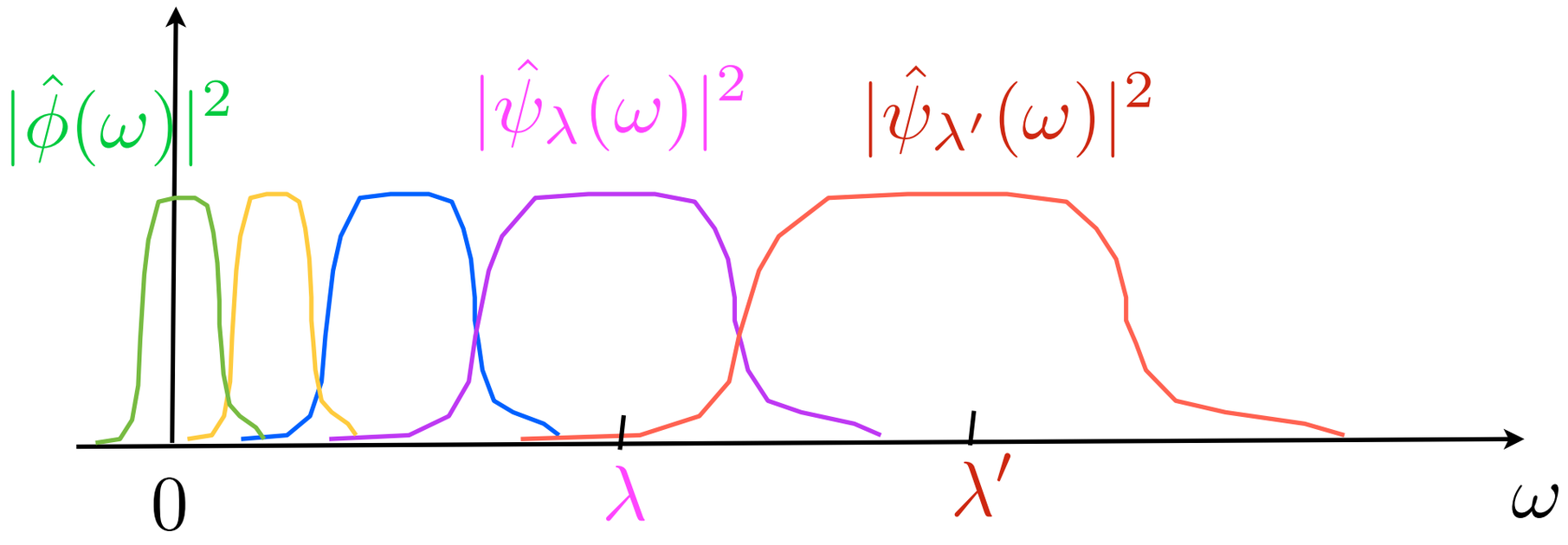
WAVELETS

- The local average $x * \phi$ is a "blurry" version of x, whereas
- $x * \psi$ carries the details lost by the blurring.
- The details are relative to a given resolution. How to obtain a decomposition that captures details at *all* resolutions?

WAVELETS

- The local average $x * \phi$ is a blurry version of x, whereas
- $x * \psi$ carries the details lost by the blurring.
- The details are relative to a given resolution. How to obtain a decomposition that captures details at *all* resolutions?

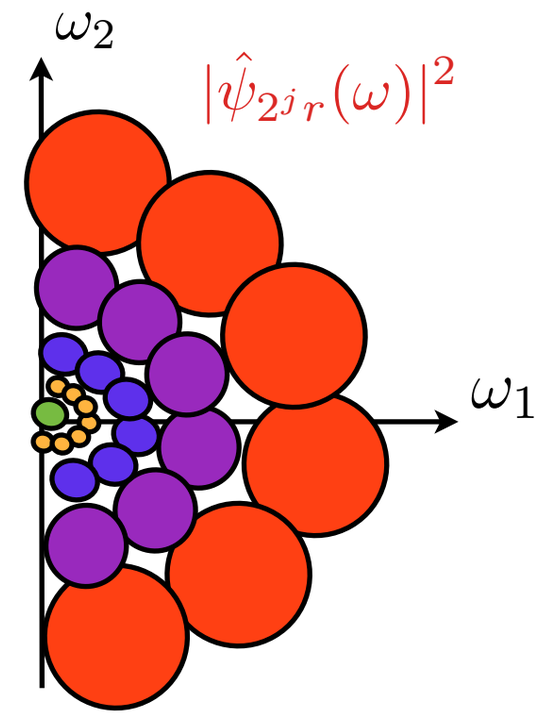
- Dilated wavelets: $\psi_j(u) = 2^{-j}\psi(2^{-j}u)\,,\ j \in \mathbb{Z}$



LITTLEWOOD-PALEY WAVELET FILTER BANKS

- For images, dilated and rotated wavelets:

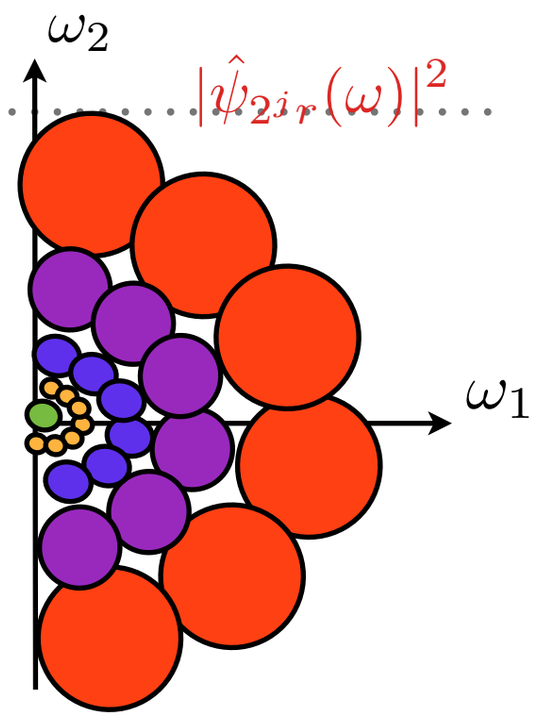
$$\psi_\lambda(u) = 2^{-j/2}\psi(2^{-j}ru) \ , \ \text{with } \lambda = 2^j r$$



LITTLEWOOD-PALEY WAVELET FILTER BANKS

- For images, dilated and rotated wavelets:

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



$$Wx = \{x \star \phi(u) \ , \ x \star \psi_\lambda(u)\}_{\lambda \in \Lambda} \qquad x \star \psi(u) = \int x(v)\psi(u-v)dv \ .$$

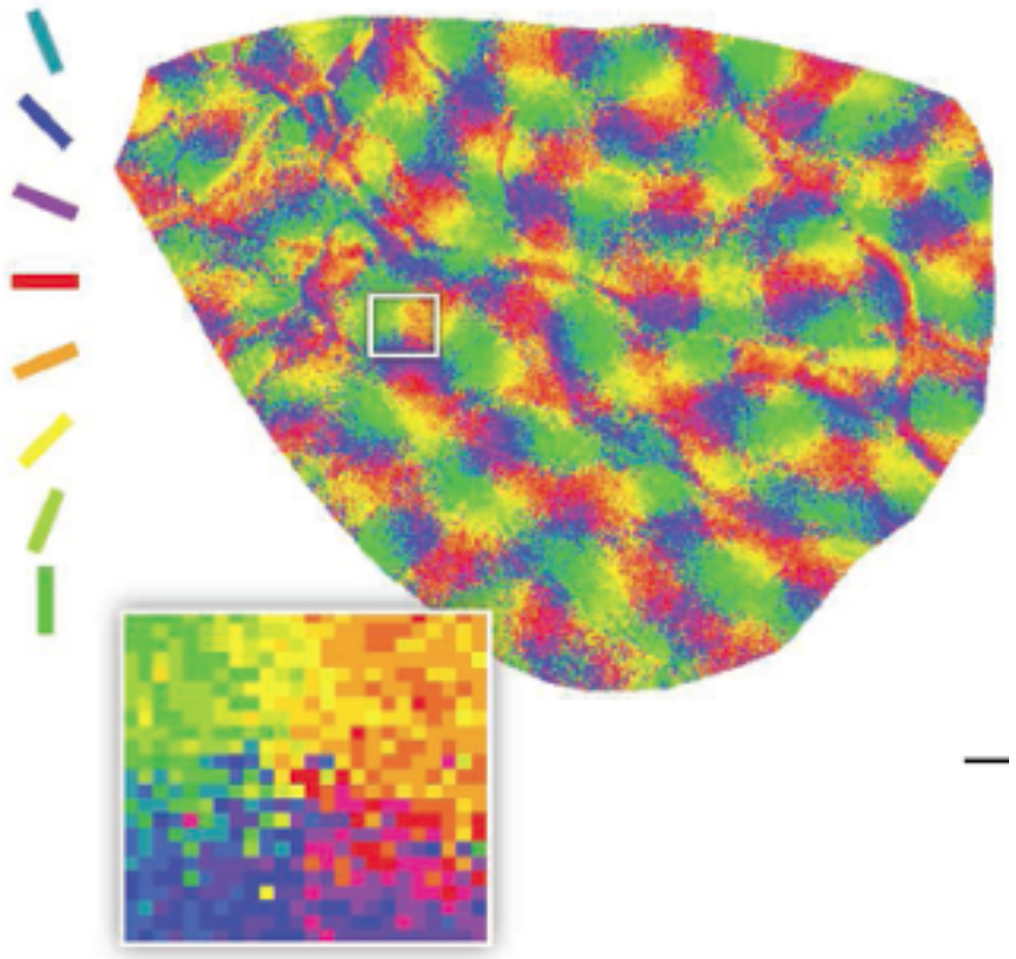
Theorem (Littlewood-Paley): If there exists $\delta > 0$ such that

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

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

WAVELETS IN VISION

- *V1* Model of Simple and Complex cells: First layer of processing is selective in orientation, scale and position.



- cells are organized in *pinwheels*. (more on that later).

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 $\Phi(x) = \{x * \psi_\lambda\}_\lambda$?

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 equivariance:

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

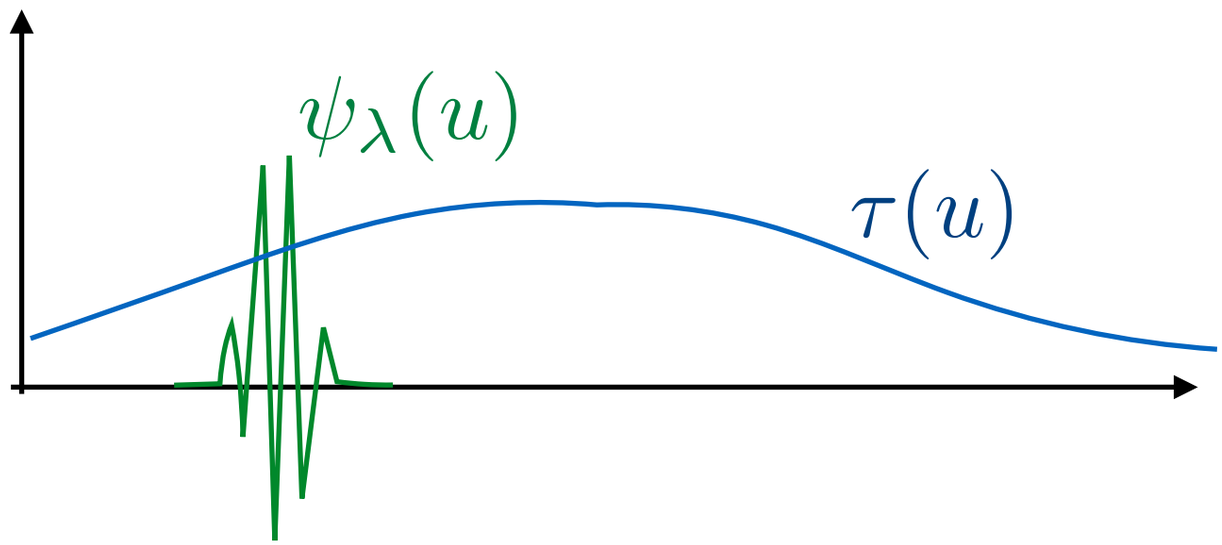
$$\|W_J\varphi_\tau - \varphi_\tau W_J\| \leq C(J\|\nabla\tau\|_\infty + \|H\tau\|_\infty)\ .$$

($H\tau$: Hessian of τ)

WAVELETS AND DEFORMATIONS

➤ Qualitative idea behind this result:

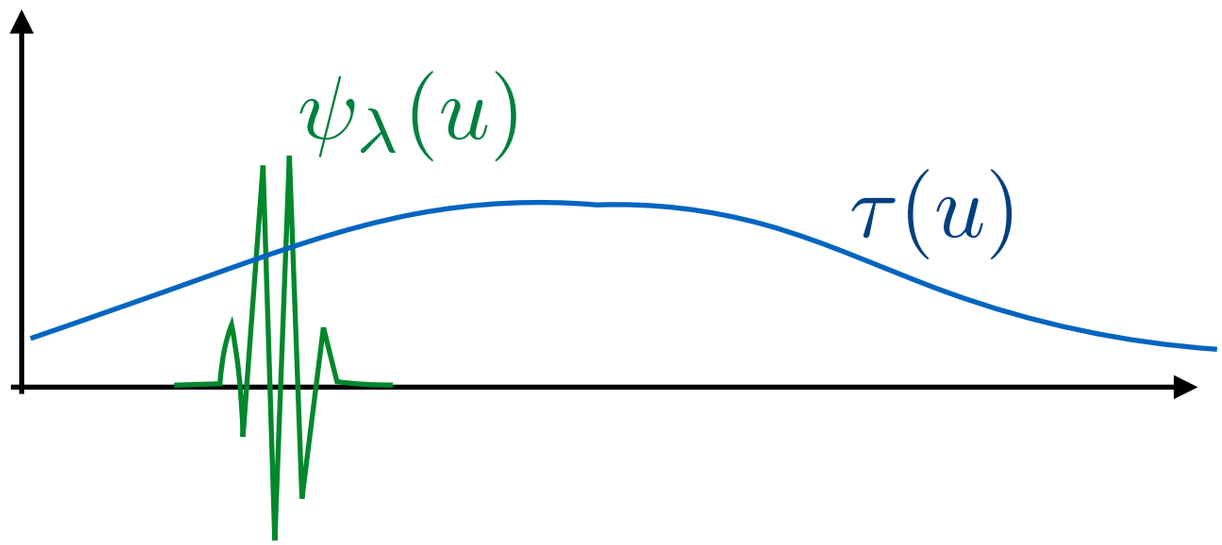
Each ψ_λ only "sees" the part of the deformation τ that intersects its support.



WAVELETS AND DEFORMATIONS

➤ Qualitative idea behind this result:

Each ψ_λ only "sees" the part of the deformation τ that intersects its support.



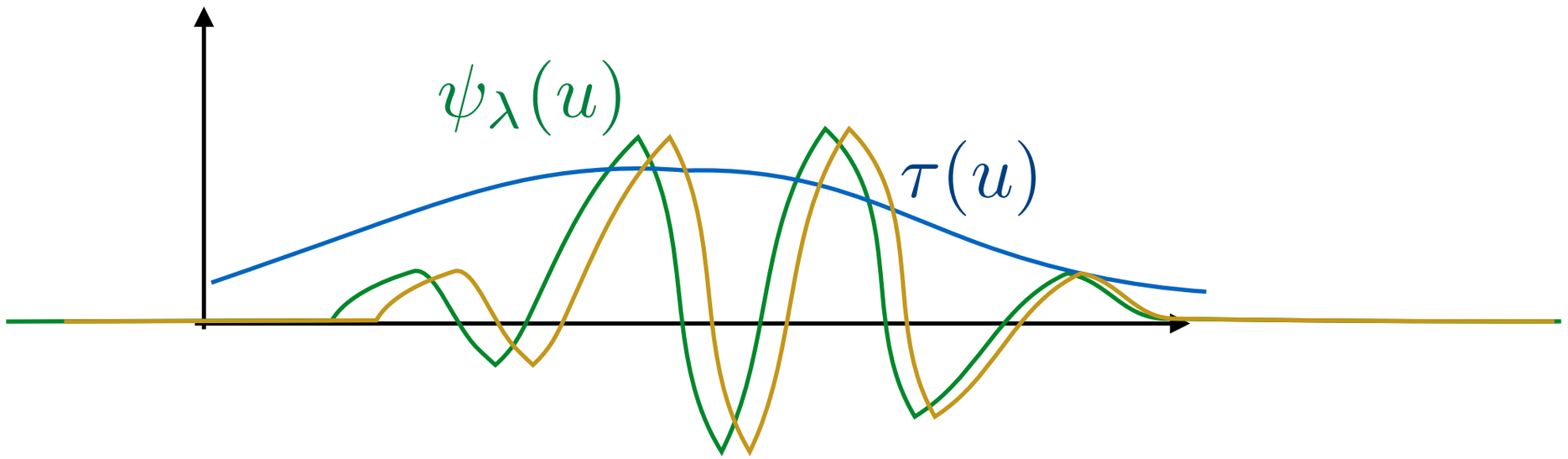
For small scales, ψ_λ has small support, and for u, v within that support, because τ is smooth, $|\tau(v) - \tau(u)| \sim 2^{-j}|\nabla\tau|_\infty$.

Thus $|(\varphi_\tau x) * \psi_\lambda(u) - x * \psi_\lambda(u - \tau(u))| \sim |\nabla\tau|_\infty$.

WAVELETS AND DEFORMATIONS

➤ Qualitative idea behind this result:

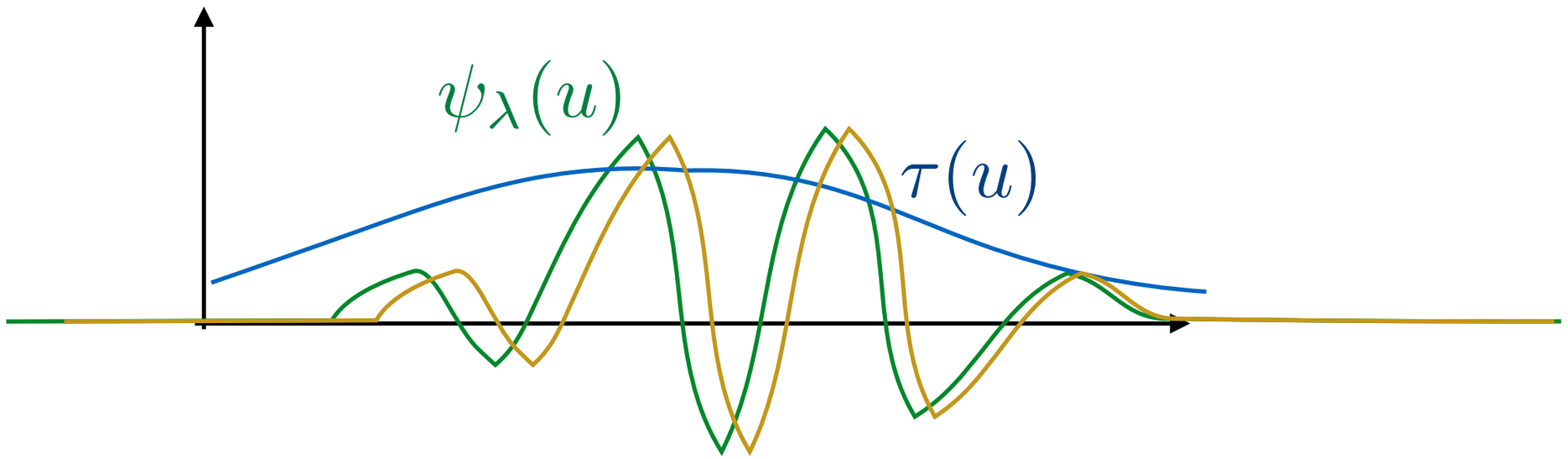
Each ψ_λ only "sees" the part of the deformation τ that intersects its support.



WAVELETS AND DEFORMATIONS

➤ Qualitative idea behind this result:

Each ψ_λ only "sees" the part of the deformation τ that intersects its support.

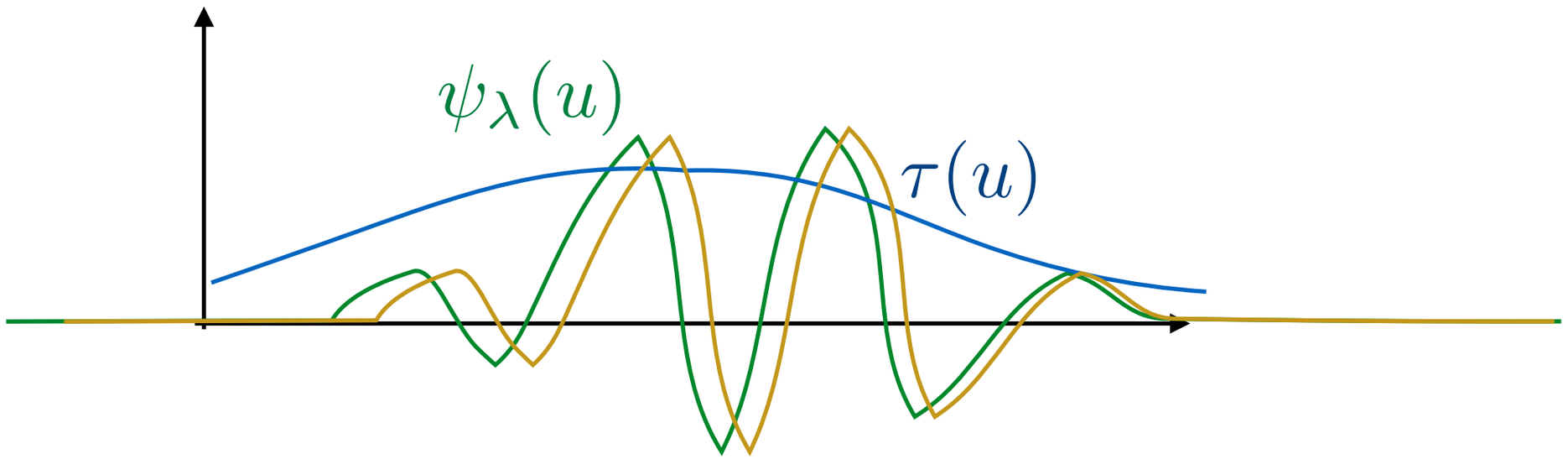


For large scales, ψ_λ is itself smooth, thus $|\varphi_\tau(x * \psi_\lambda) - (\varphi_\tau x) * \psi_\lambda| \sim \|\nabla\tau\|_\infty$.

WAVELETS AND DEFORMATIONS

➤ Qualitative idea behind this result:

Each ψ_λ only "sees" the part of the deformation τ that intersects its support.

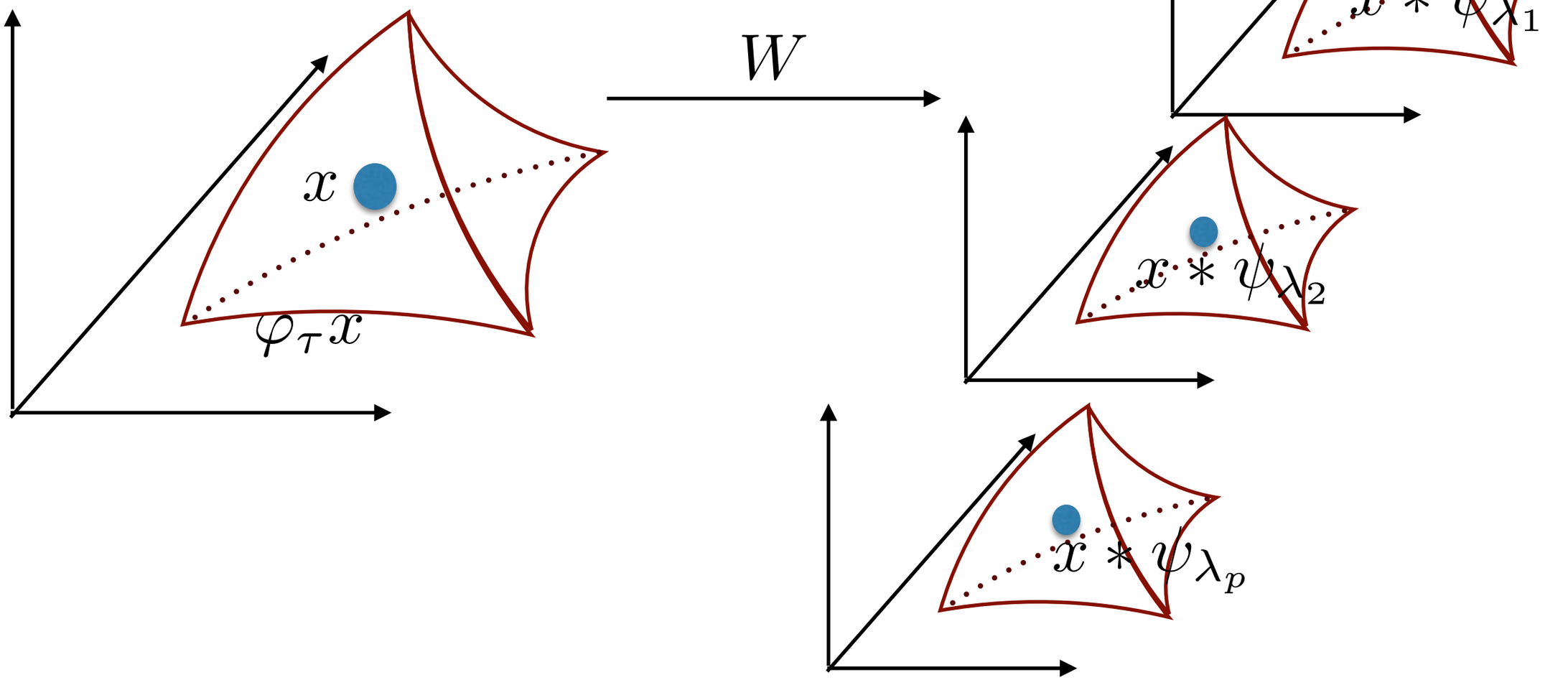


For large scales, ψ_λ is itself smooth, thus
$|\varphi_\tau(x * \psi_\lambda) - (\varphi_\tau x) * \psi_\lambda| \sim \|\nabla\tau\|_\infty$.

And, most importantly, wavelet separates scales
(so errors do not accumulate)

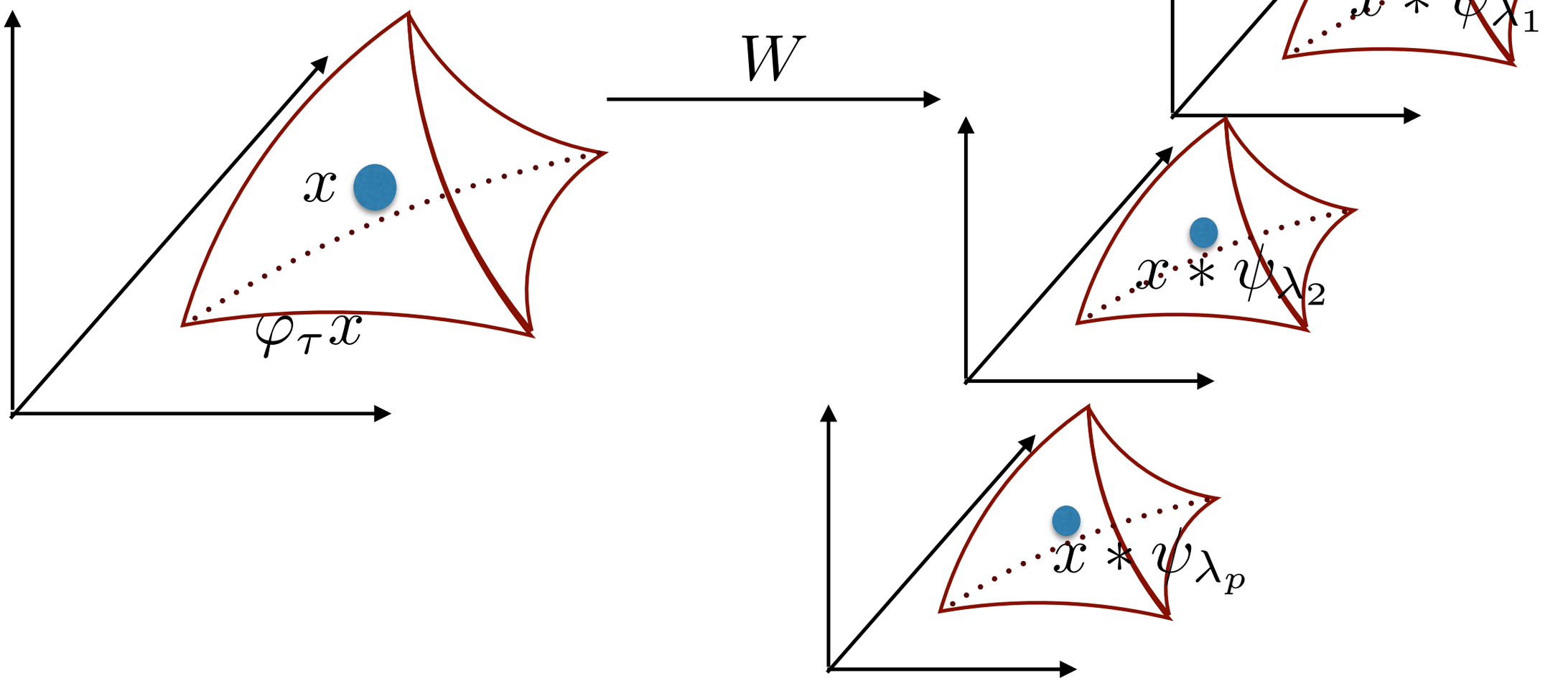
WAVELETS AND NON-LINEARITIES

- The commutation property says that deformations in the input are approximately mapped to deformations in the wavelet domain:



WAVELETS AND NON-LINEARITIES

- The commutation property says that deformations in the input are approximately mapped to deformations in the wavelet domain:



- We want to extract again stable measurements: *need non-linear operator.*

CHARACTERIZATION OF STABLE NON-LINEARITIES

➤Preserve additive stability:

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

CHARACTERIZATION OF STABLE NON-LINEARITIES

➤Preserve additive stability:

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

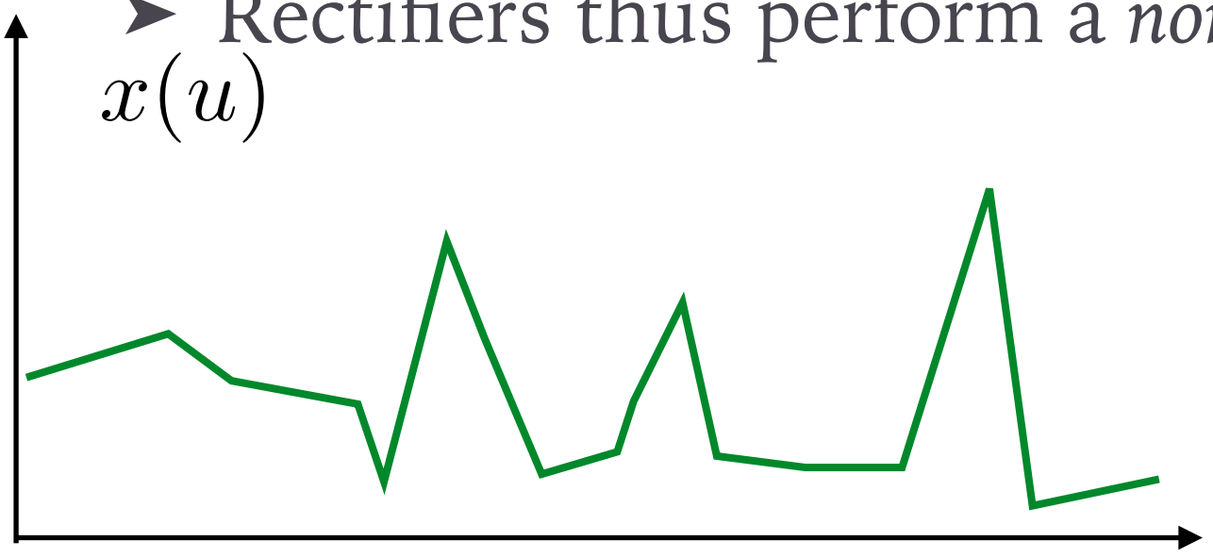
➤Preserve geometric stability: It is sufficient to commute with diffeomorphisms.

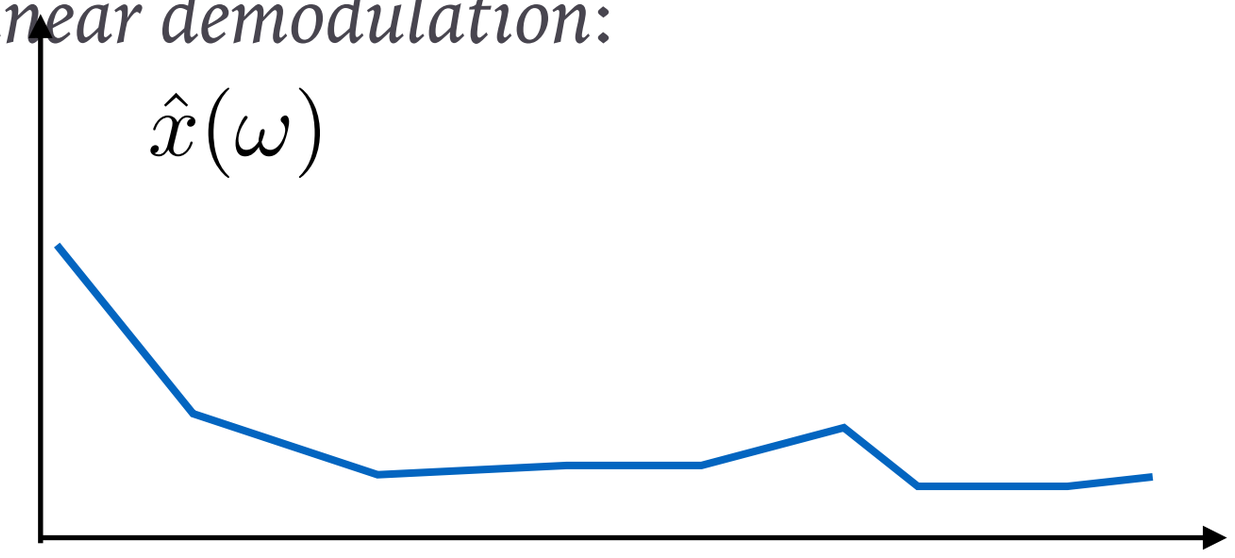
Theorem: If M is non-expansive operator in L^2 such that $\varphi_\tau M = M\varphi_\tau$ for all τ, then M is point-wise:

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

UNDERSTANDING THE EFFECT OF NONLINEARITIES

➤ Rectifiers thus perform a *non-linear demodulation*:

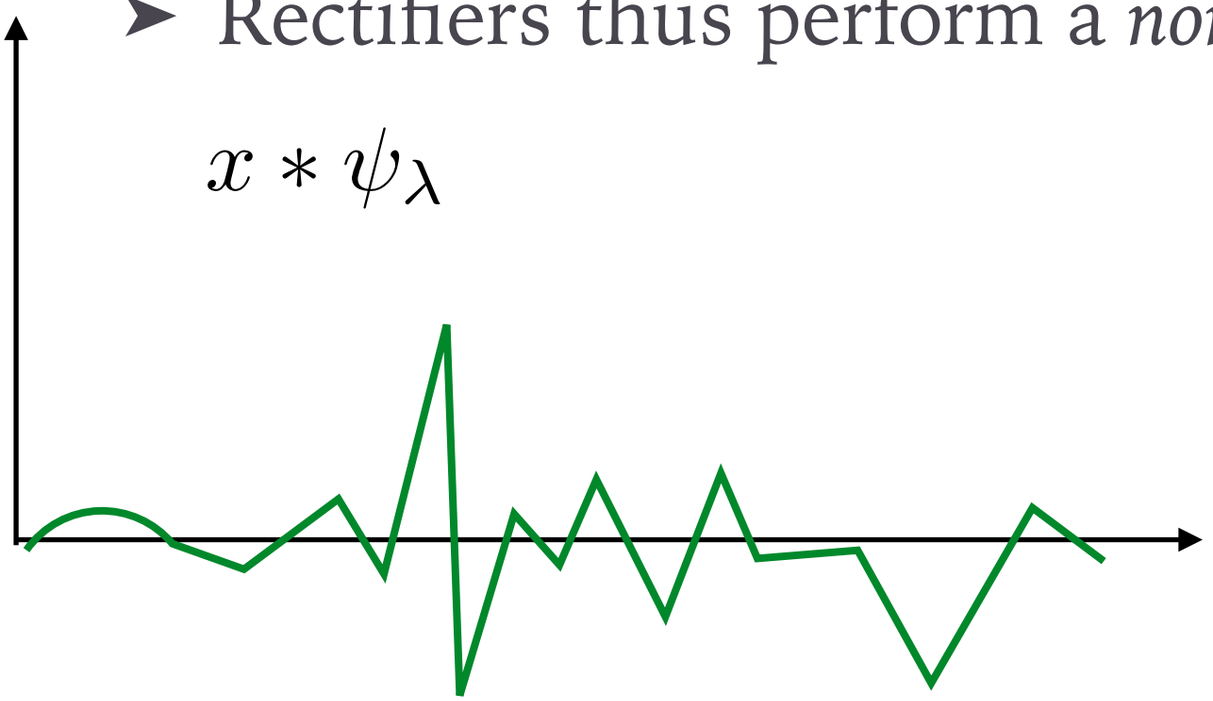




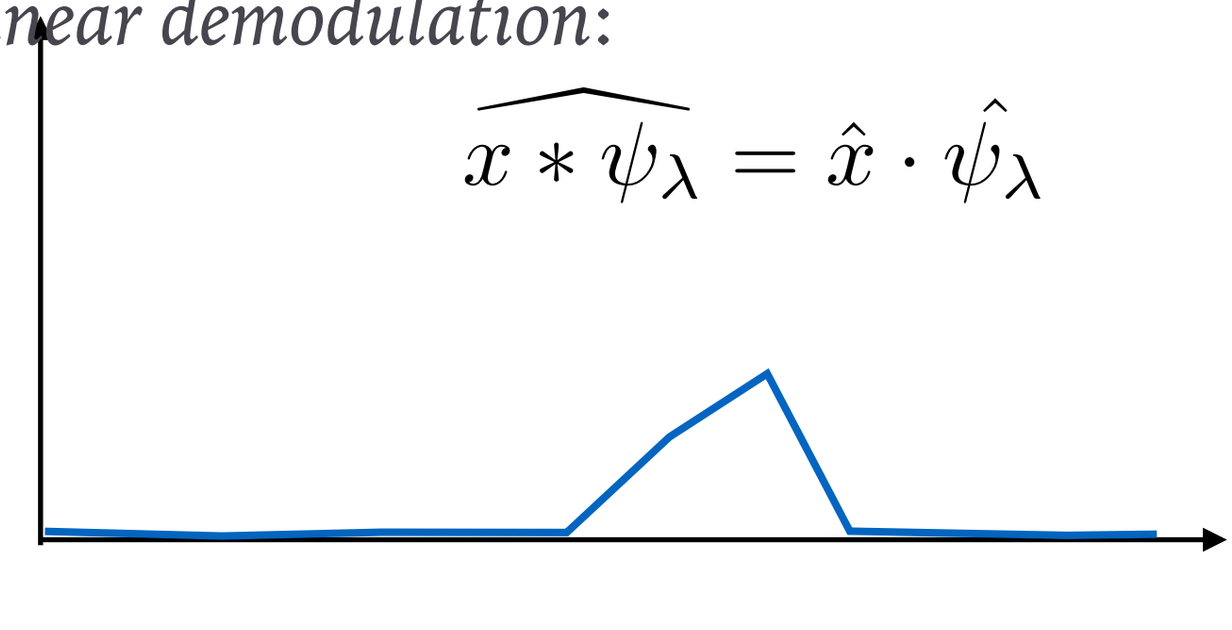
UNDERSTANDING THE EFFECT OF NONLINEARITIES

- Rectifiers thus perform a *non-linear demodulation*:

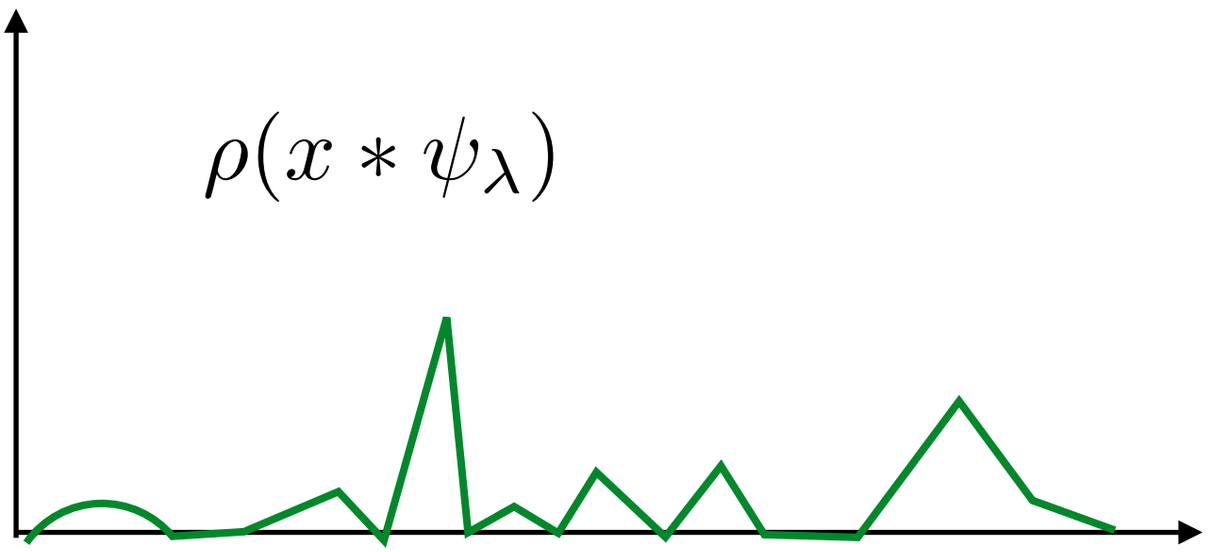
$x * \psi_\lambda$



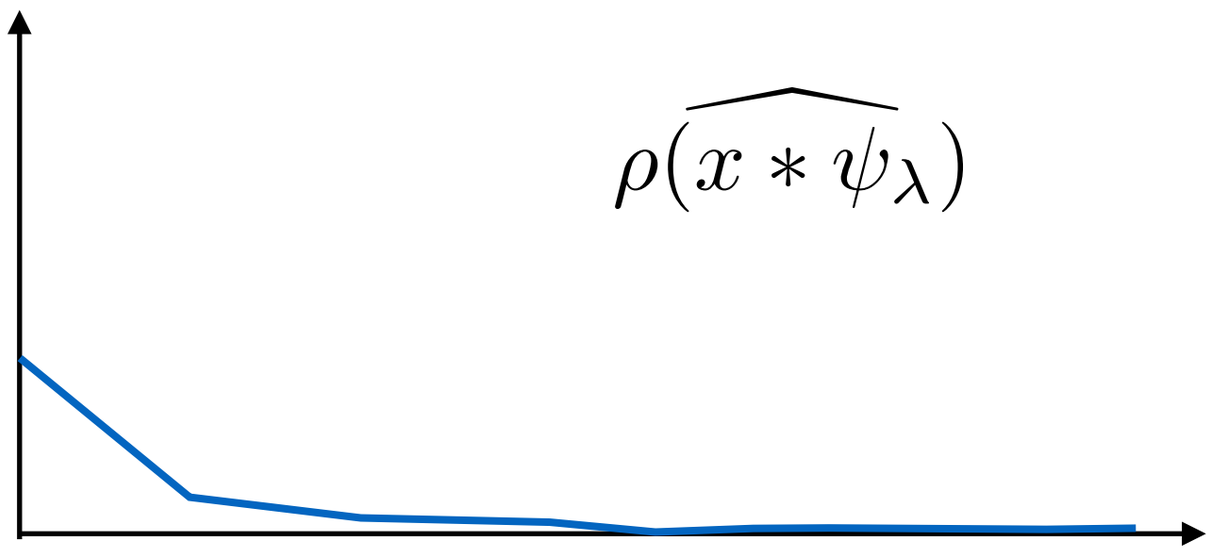
$\widehat{x * \psi_\lambda} = \hat{x} \cdot \hat{\psi}_\lambda$



$\rho(x * \psi_\lambda)$



$\widehat{\rho(x * \psi_\lambda)}$



sometimes called the *envelope*

CHOICE OF POINTWISE NONLINEARITY

- Full rectification $\rho(z) = |z|$ preserves energy:
  - When the wavelet is complex, it produces smoother envelopes (thus more stable features).

- Half rectification (ReLU) $\rho(z) = \max(z, 0)$ captures half the energy, and it also creates *sparsity*.
  - We will see that this is important to perform *detection*.

- Sigmoid nonlinearity $\rho(z) = (1 + e^{-z})^{-1}$.
  - It is not homogeneous
  - Saturating regimes are problematic for learning via back propagation in deep models.
- Other: "Leaky" ReLU [MSR'14]: parametrized half-rectifier, ELU , etc.

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 $\mathcal{U}_J(x) = \{x \star \phi_J \, , \, |x \star \psi_\lambda|\}_{\lambda \in \Lambda_J} \, .$
- Recover lost information:
  - Point-wise, non-expansive non-linearities: maintain stability.
  - Complex modulus maps energy towards low-frequencies.
- Cascade the "recovery" operator:

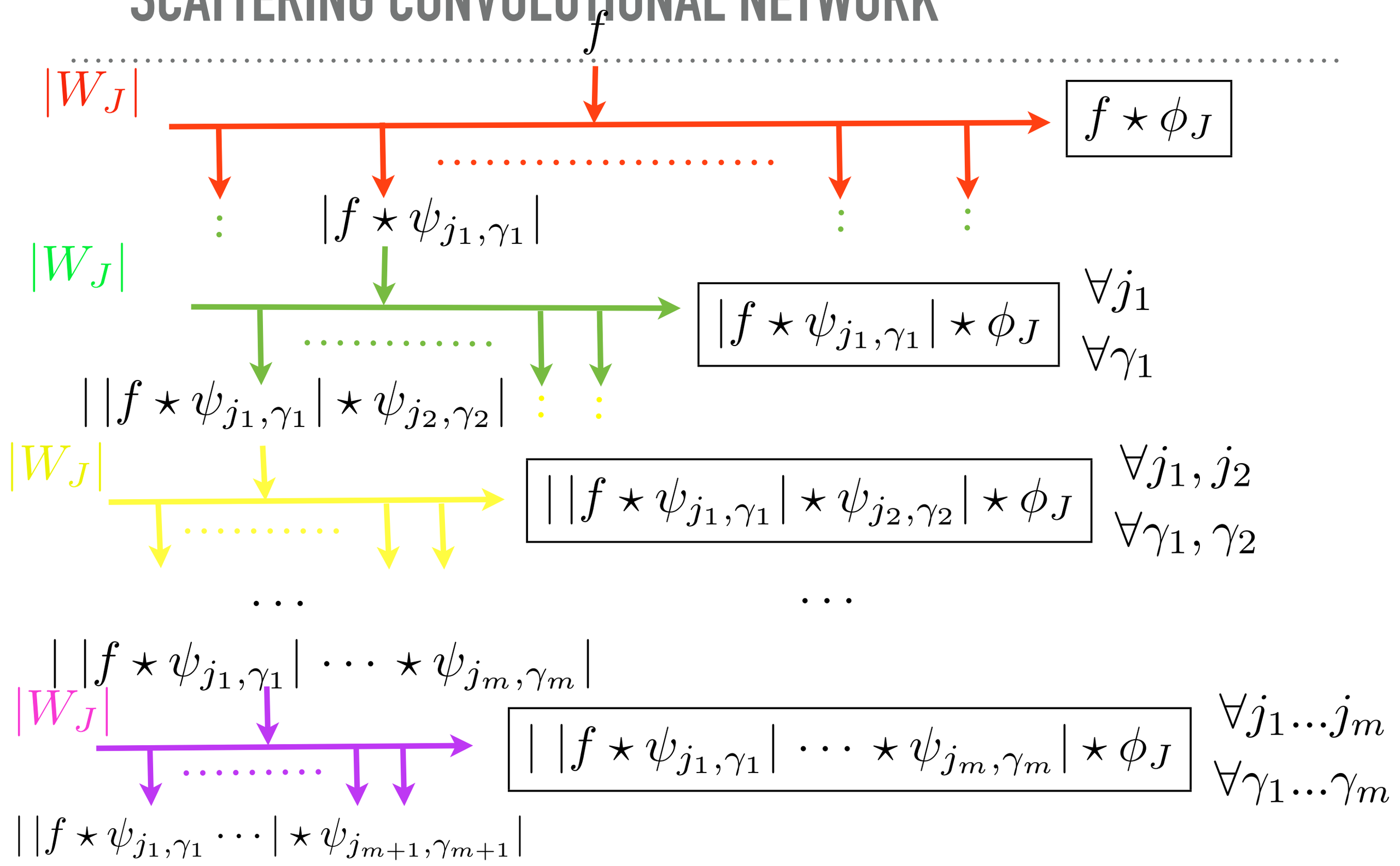
$$\mathcal{U}_J^2(x) = \{x \star \phi_J \, , \, |x \star \psi_\lambda| \star \phi_J, \, ||x \star \psi_\lambda| \star \psi_{\lambda'}|\}_{\lambda, \lambda' \in \Lambda_J} \, .$$

$$p = (\lambda_1, \ldots, \lambda_m) \; :$$

- Scattering coefficient along a path

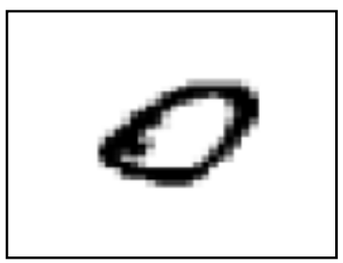
$$S_J[p]x(u) = |||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \ldots | \star \psi_{\lambda_m}| \star \phi_J(u) \; .$$

SCATTERING CONVOLUTIONAL NETWORK



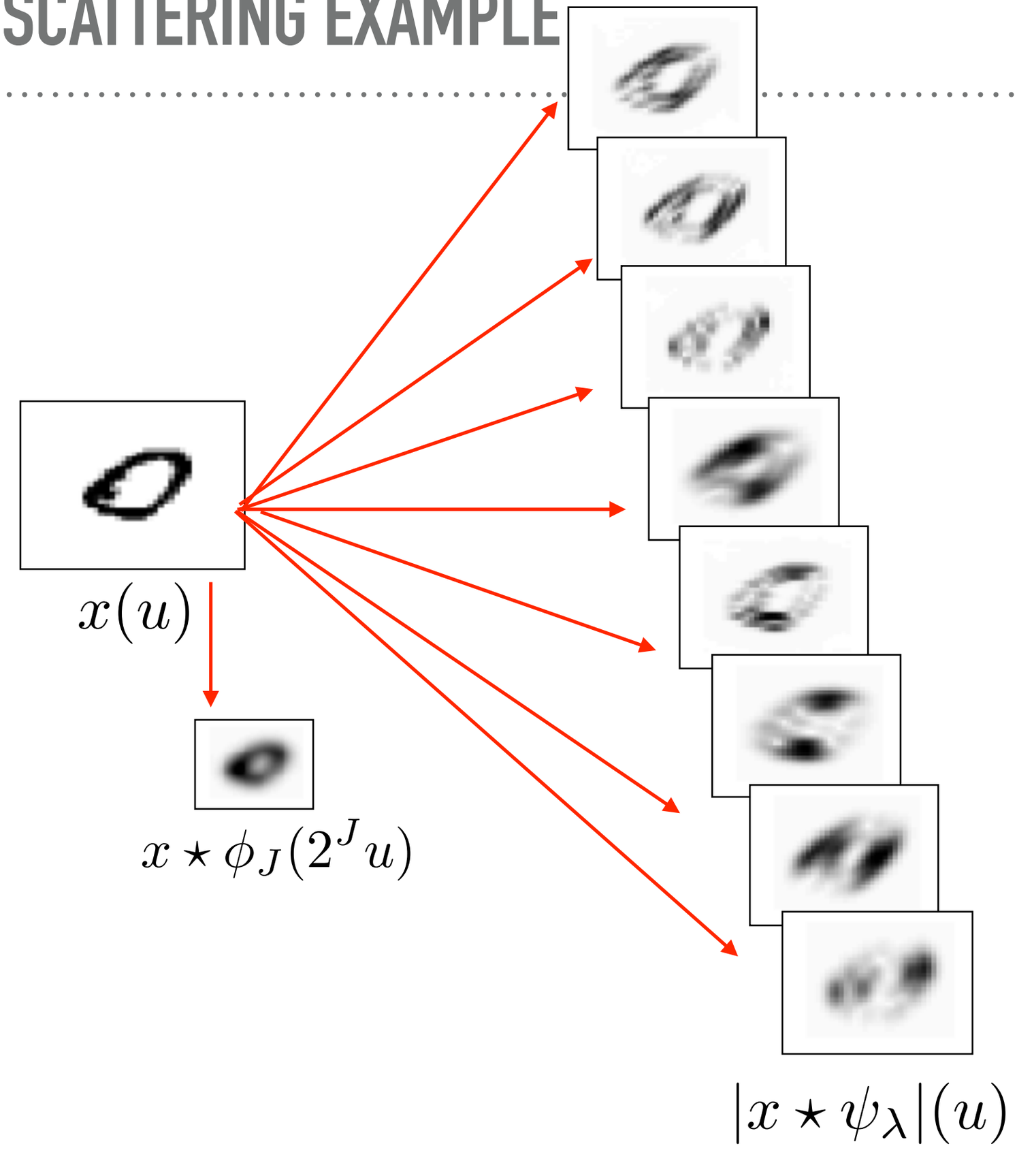
Cascade of contractive operators.

SCATTERING EXAMPLE

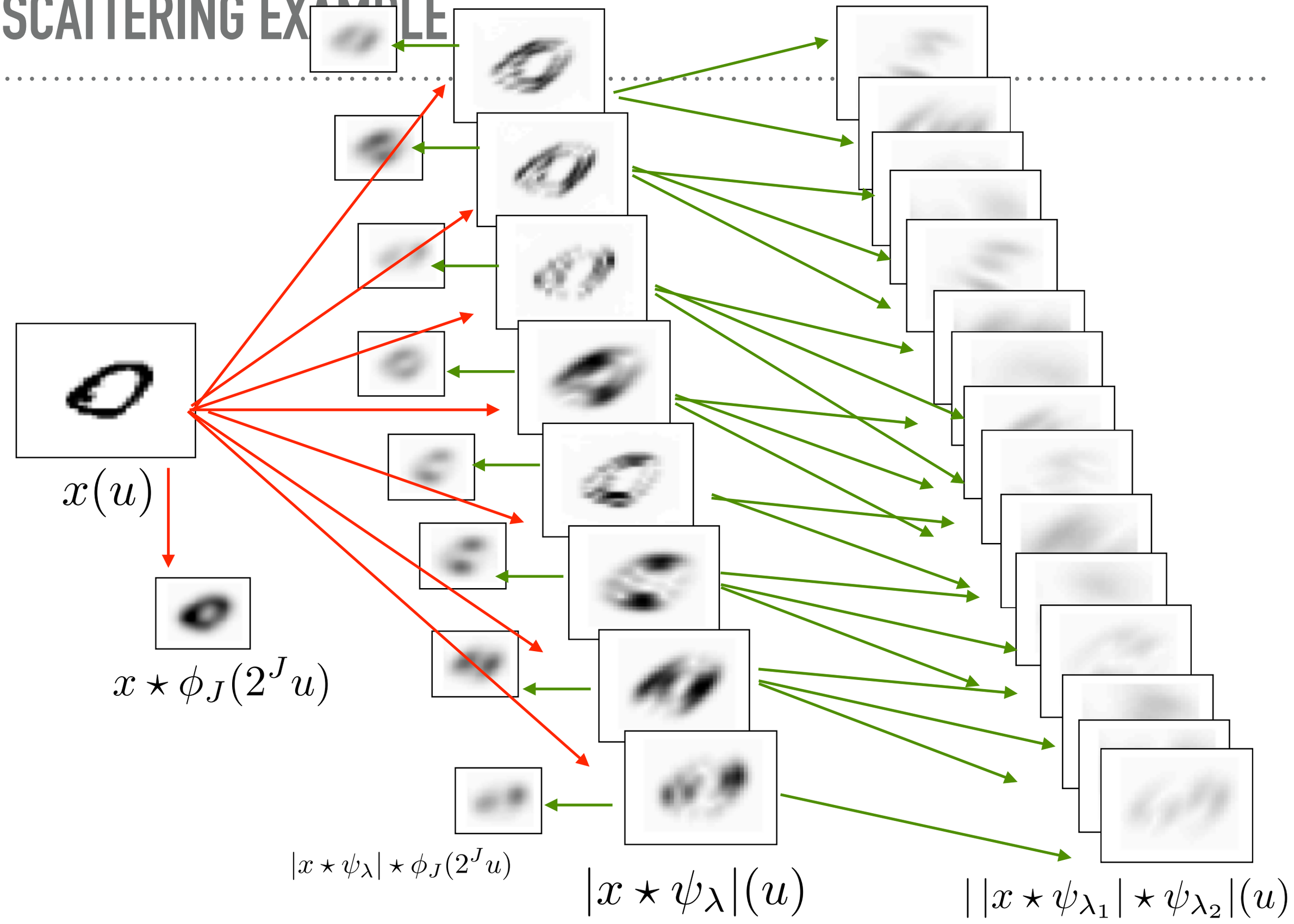


$x(u)$

SCATTERING EXAMPLE



SCATTERING EXAMPLE



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 $(1+JL)$.

MULTI-RESOLUTION WAVELETS

At each scale j, we consider a low-pass *scaling filter* h and band-pass filters g_θ, $\theta \in [1, \dots, L]$.

Wavelets and the bluring kernel are obtained at each j by cascading these filters:

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

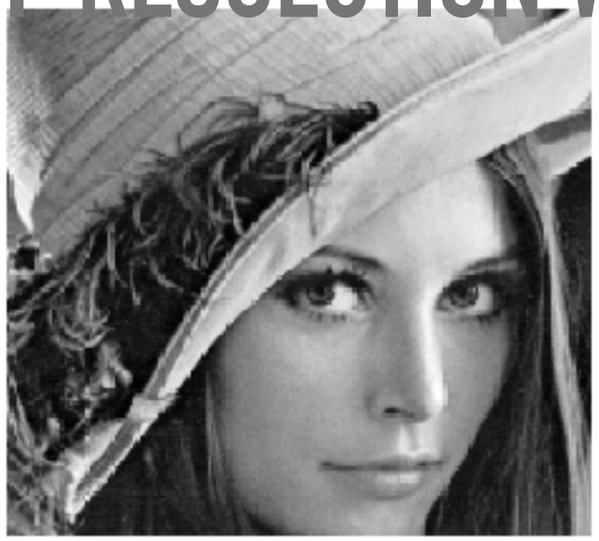
Decompositions are obtained by cascading fine-to-coarse:

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

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

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

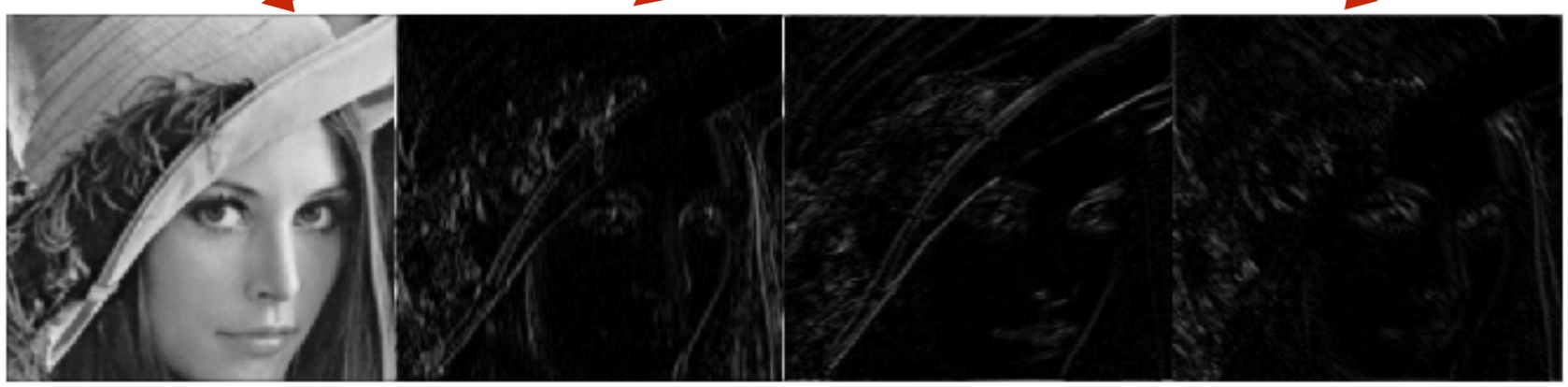
SCATTERING WITH MULTI-RESOLUTION WAVELETS

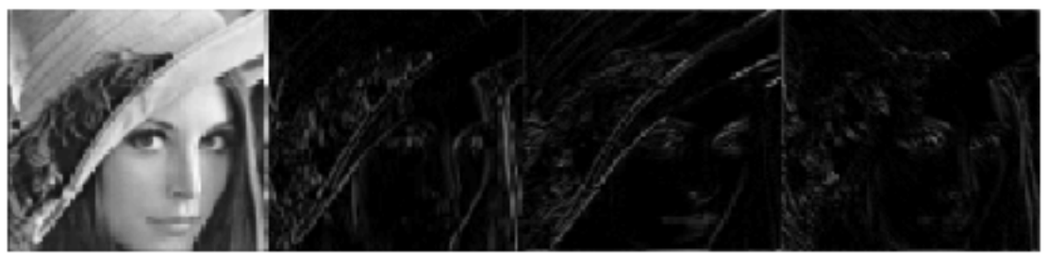


x

$x \star h$

$|x \star g_\theta|$

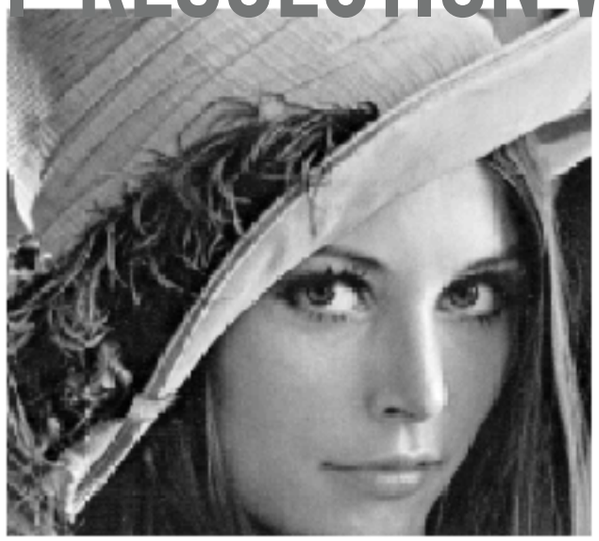




SCATTERING WITH MULTI-RESOLUTION WAVELETS

#layers:

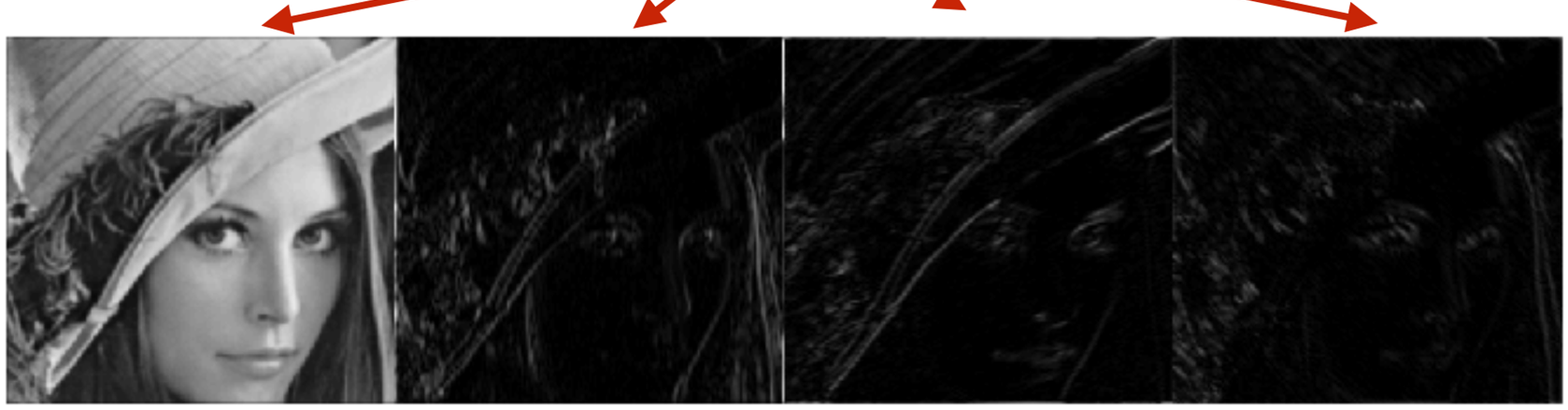
maximum scale

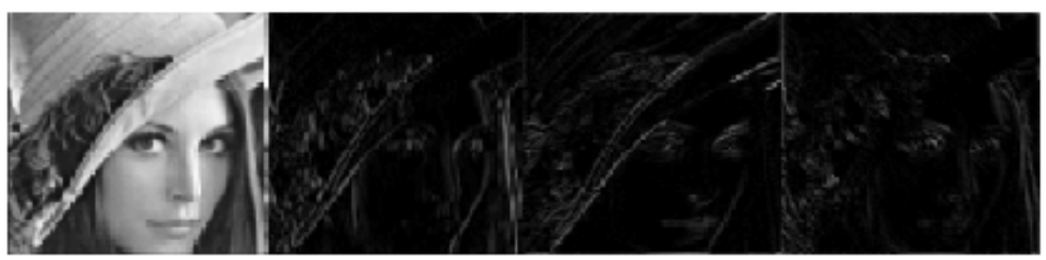


x

$x \star h$

$|x \star g_\theta|$





SCATTERING CONSERVATION OF ENERGY

➤ Additive stability and 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:

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

$$|x \star \psi_{j,\theta}|$$

- Repeat formula on each output

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

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

$\forall\ m$

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

INTERPRETATION

- Unitary Wavelet decomposition preserves energy:

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

$$|x \star \psi_{j,\theta}|$$

- Repeat formula on each output

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

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

$\forall\ m$

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

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

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

- 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: ??

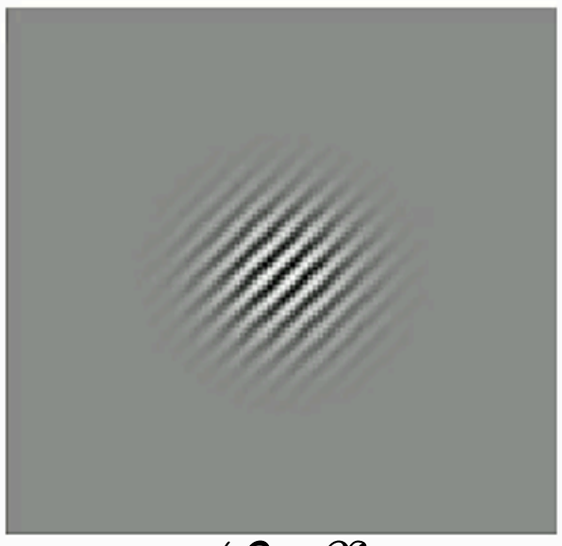
SCATTERING GEOMETRIC STABILITY

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

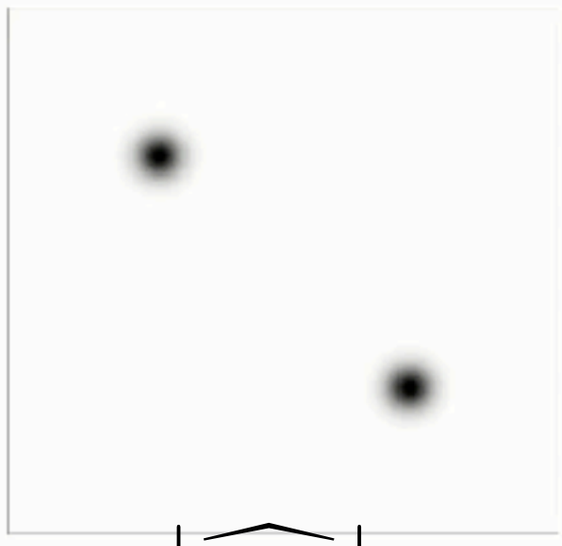
➤Geometric Stability:

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

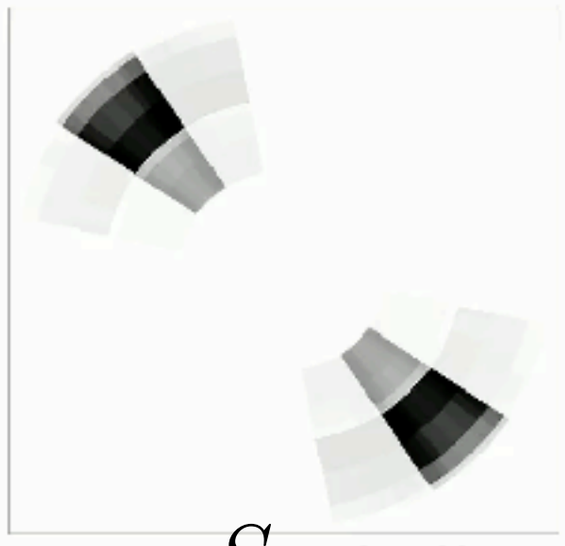
$$\|S_J \varphi_\tau x - S_J x\| \leq Cm\|x\|(2^{-J}|\tau|_\infty + \|\nabla\tau\|_\infty + \|H\tau\|_\infty) \ .$$



$\varphi_\tau x$



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



$S_J \varphi_\tau x$

INTERPRETATION

- Denote

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

- We know that

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

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

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

$$S_J = \{A_J, A_J M W_J, A_J M W_J M W_J, \dots\}$$

-

INTERPRETATION

- Denote

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

- We know that

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

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

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

$$S_J = \{A_J, A_J M W_J, A_J M W_J M W_J, \dots\}$$

-

INTERPRETATION

- Each order contributes separately:

$$\|S_J - \hat{S}_J\varphi_\tau\|^2 = \|A_J - \hat{A}_J\varphi_\tau\|^2 + \|A_J MW_J - A_J MW_J\varphi_\tau\|^2 + \dots$$

INTERPRETATION

- Each order contributes separately:

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

- Let us inspect a generic term:

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

$$(U_J = MW_J)$$

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

INTERPRETATION

- Each order contributes separately:

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

- Let us inspect a generic term:

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

$$(U_J = MW_J)$$

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

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

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

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

DISCRIMINABILITY AND SPARSITY

- Typical non-linearities are contractive:

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

DISCRIMINABILITY AND SPARSITY

- Typical non-linearities are contractive:

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

x, x'

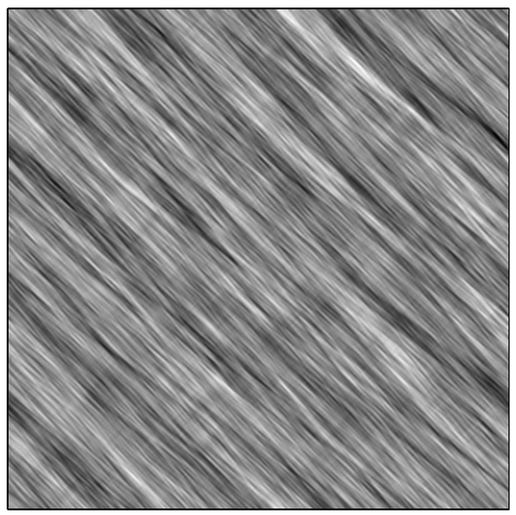
- However, if are sparse, this inequality is an equality in most of the signal domain.

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

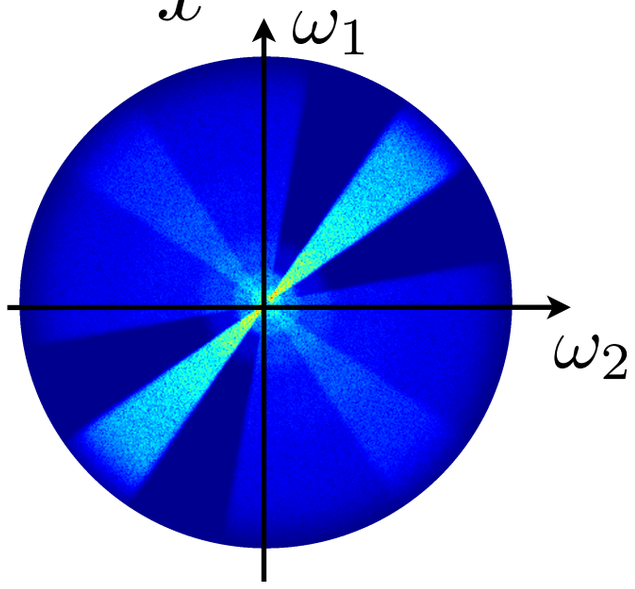
IMAGE EXAMPLES

Images ..... Fourier ..... Wavelet Scattering .....

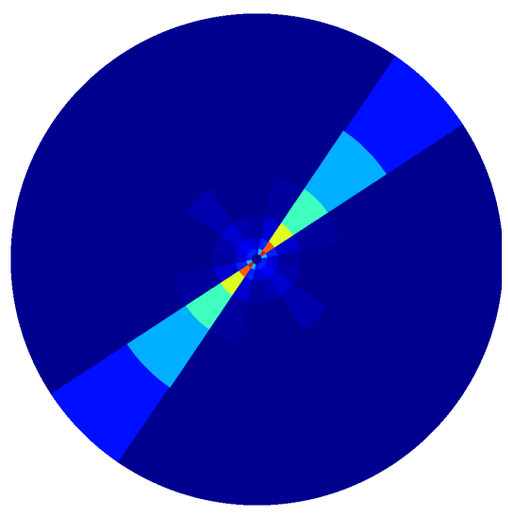
x



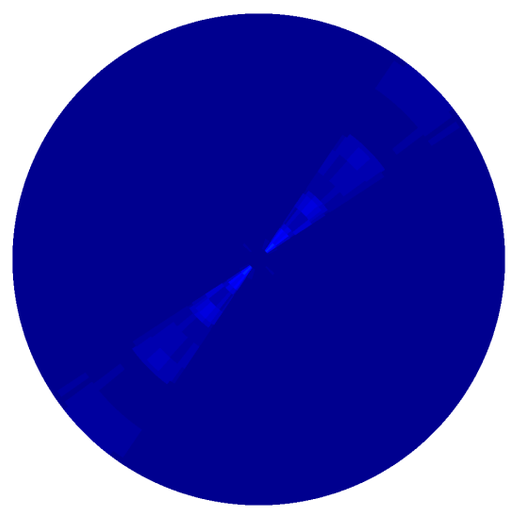
$\hat{x}$

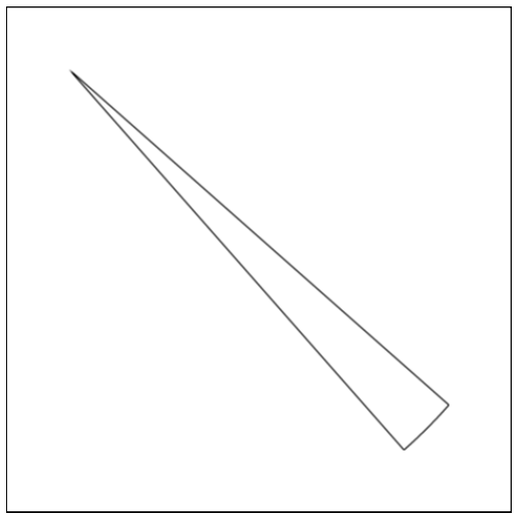


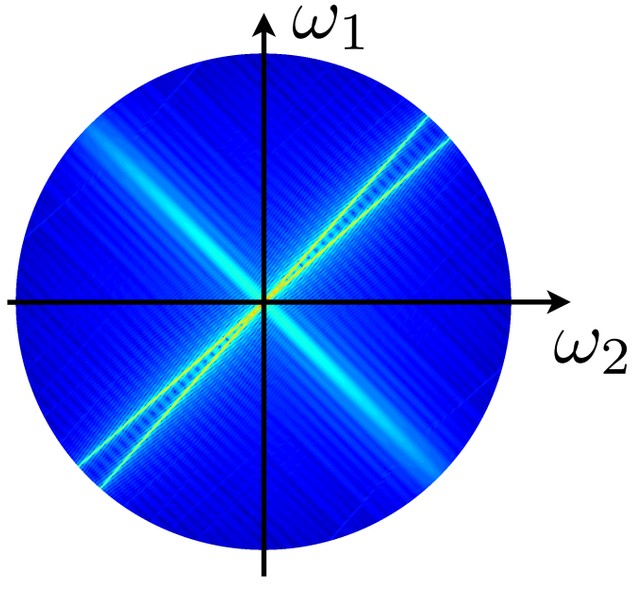
$|x \star \psi_{\lambda_1}| \star \phi_J$

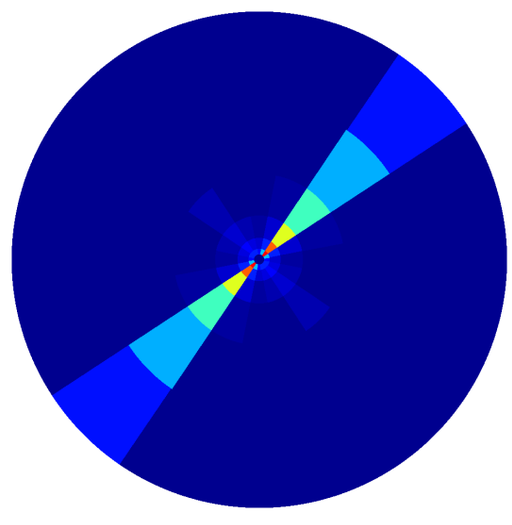


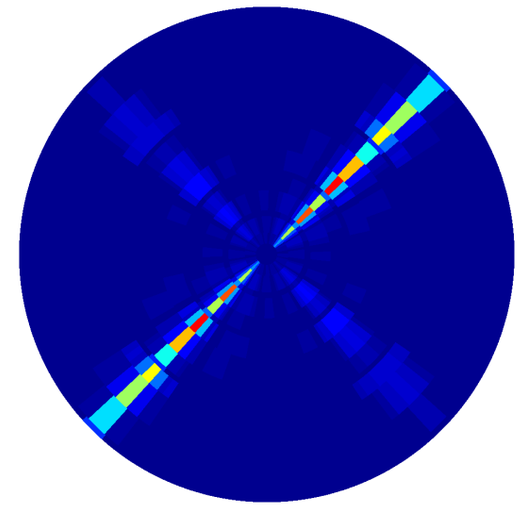
$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_J$







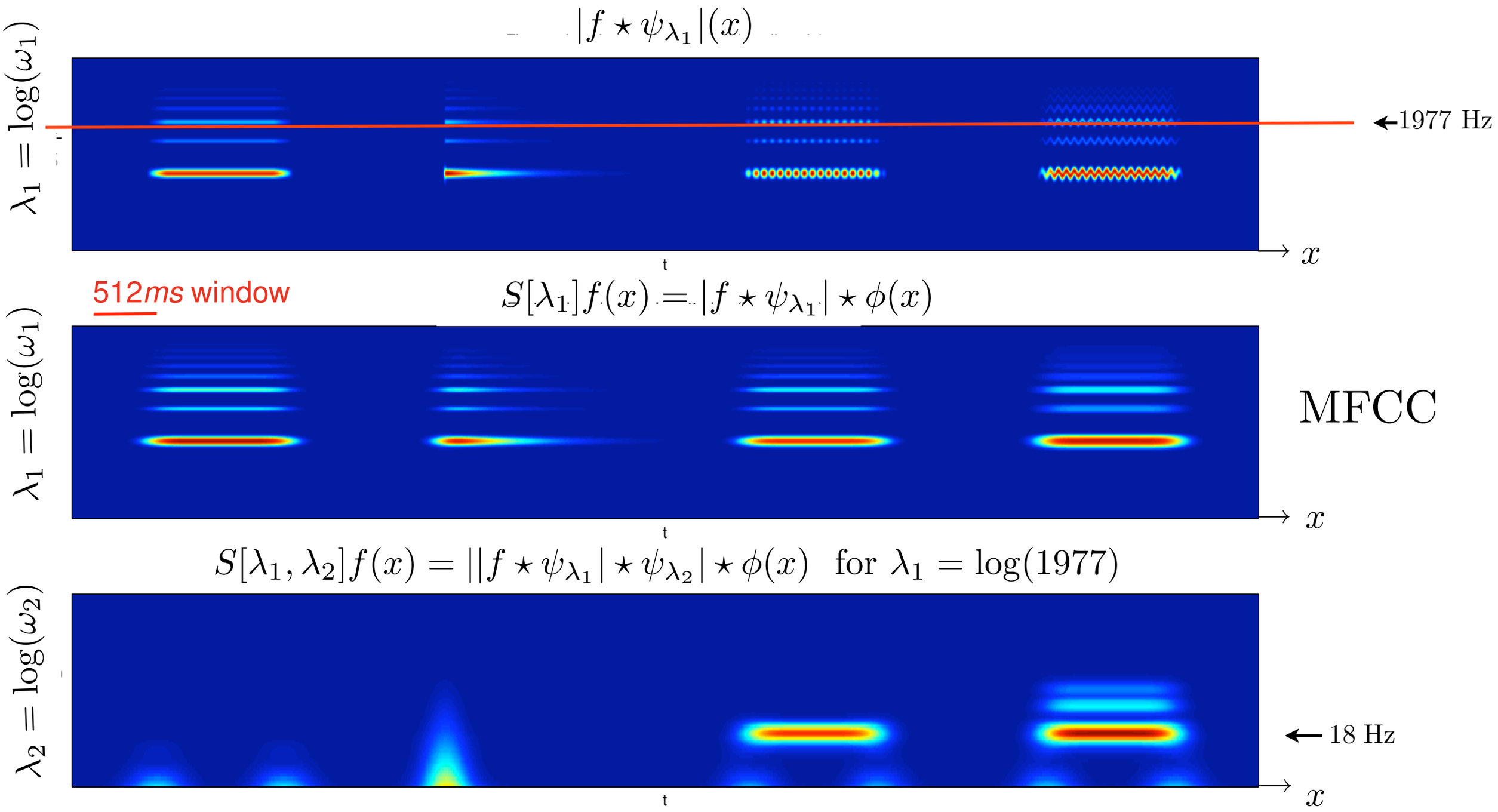




window size = image size

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 i

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

$$S^1$$

- 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 i

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

$$S^1$$

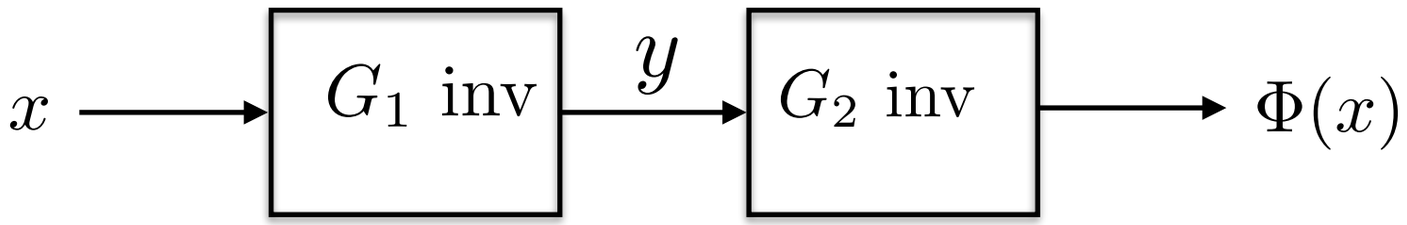
- Each S^1 acts on images along a different coordinate:

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

JOINT VERSUS SEPARABLE INVARIANCE

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

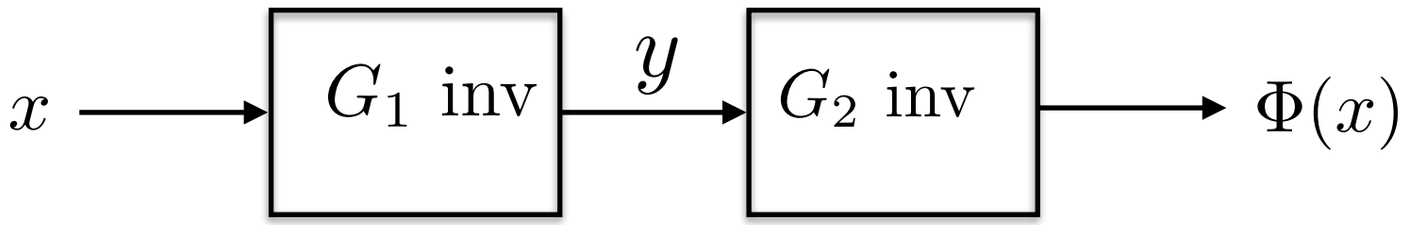
$$G = G_1 \times G_2$$



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

not capture

many things

(u_1, u_2)

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

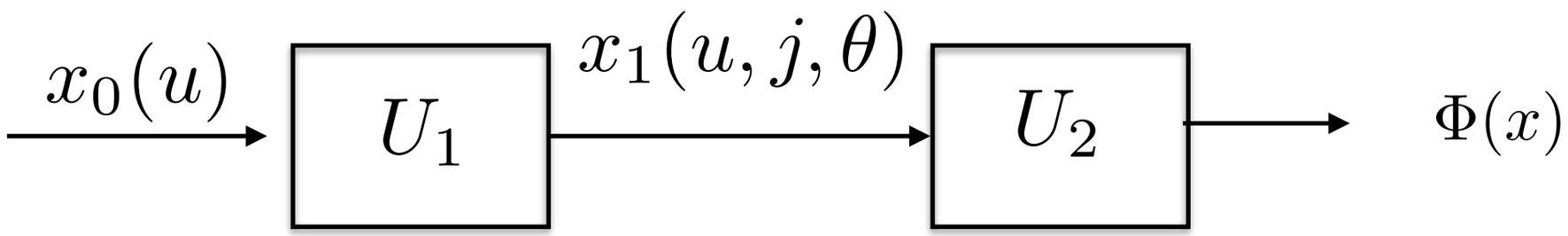
Definition: Let G be a group equipped with a Haar measure $d\mu$, acting on Ω, and $h \in L^1(G)$. The group convolution $x \star_G h$ is defined as

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

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.

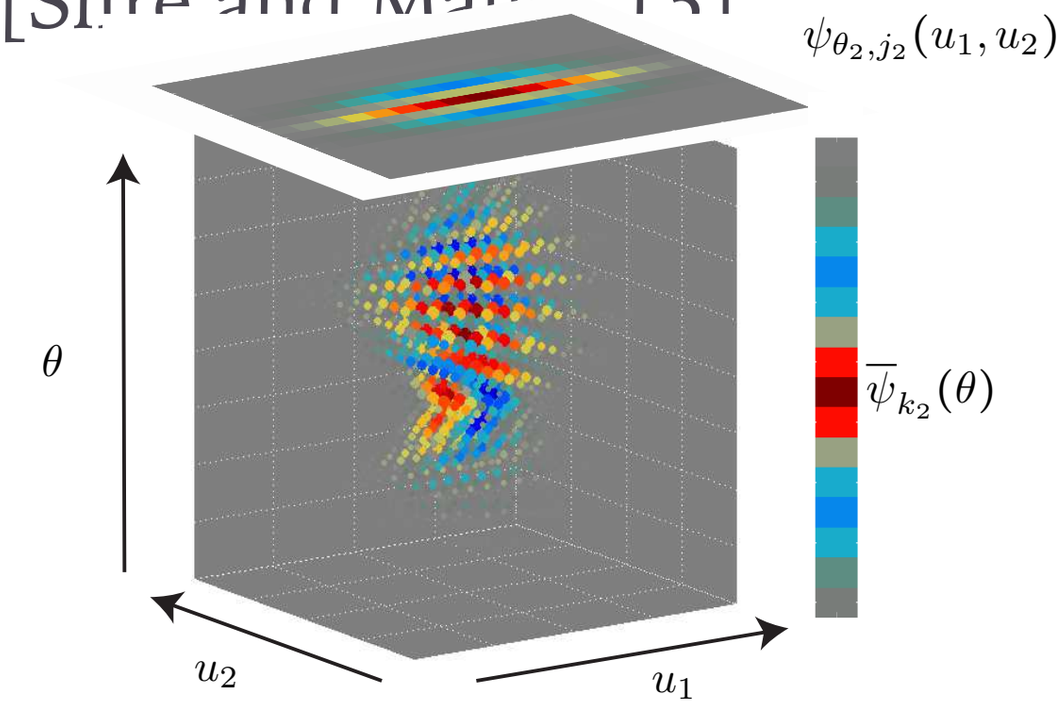


- We then adapt the second wavelet operator to its input joint variability structure.

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

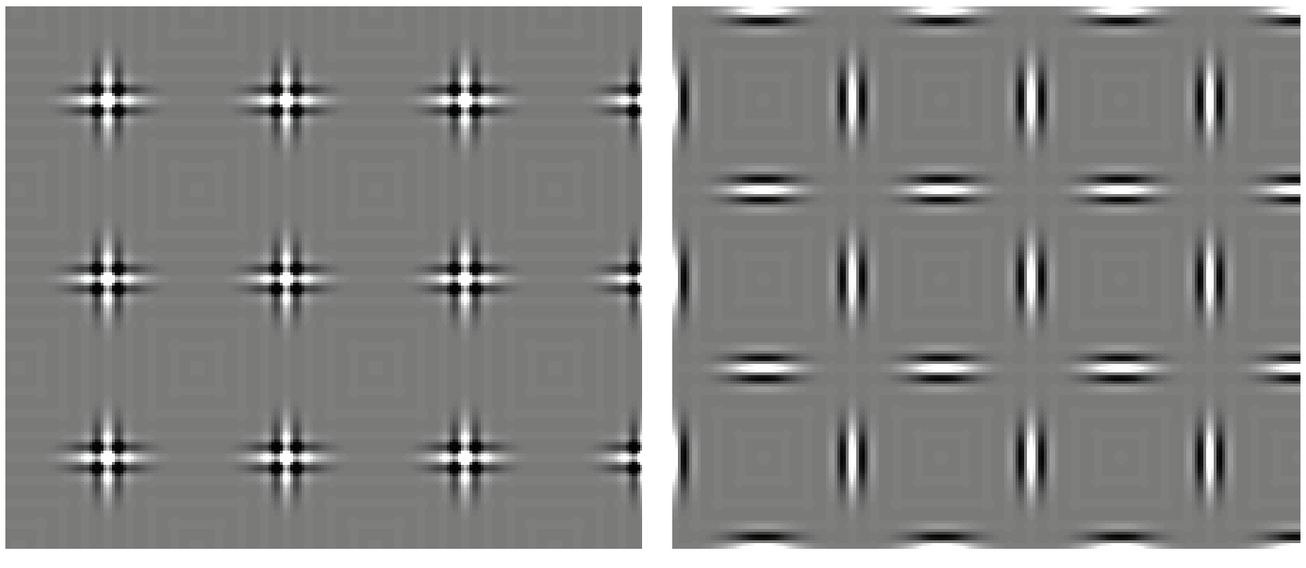
EXAMPLE: ROTO-TRANSLATION SCATTERING

- [Sifre and Mallat'13]



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

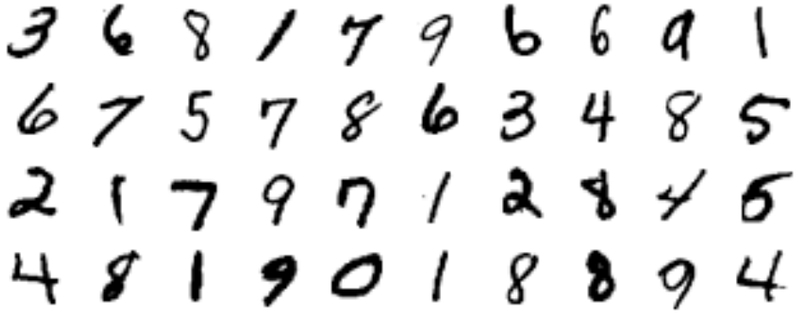
$$\Psi_\lambda(u,\theta) = \psi_{\lambda_1}(u)\psi_{\lambda_2}(\theta)$$

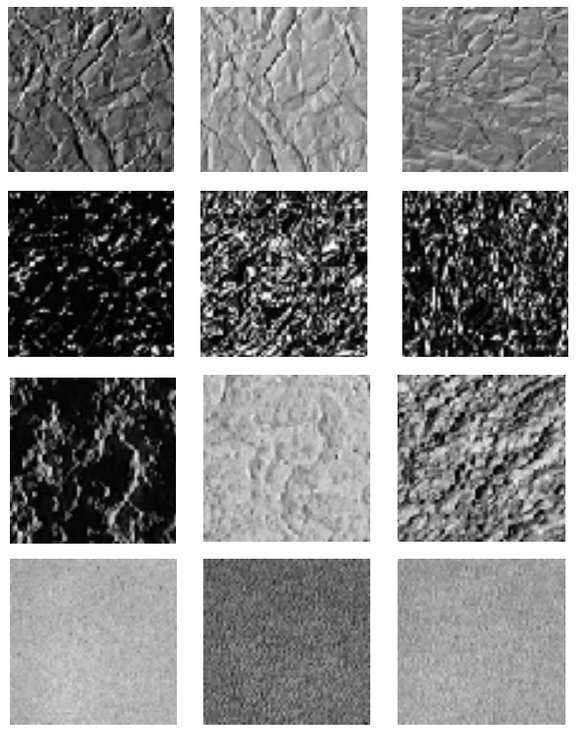


example of patterns that are discriminated by joint scattering but not with separable scattering.

CLASSIFICATION WITH SCATTERING

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



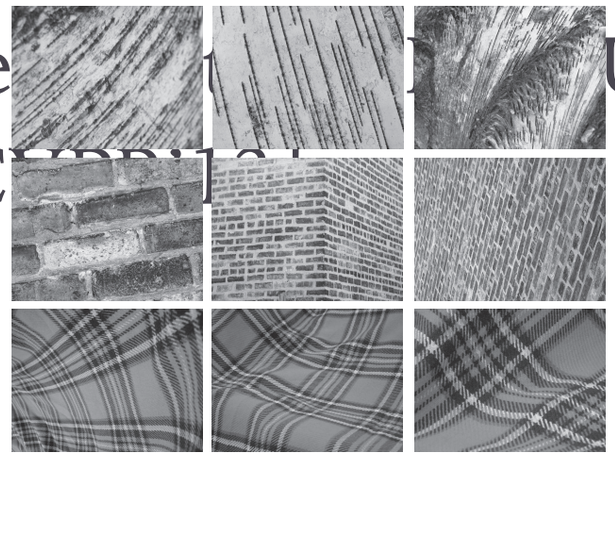


  - Texture (CUREt) [Pami'13]

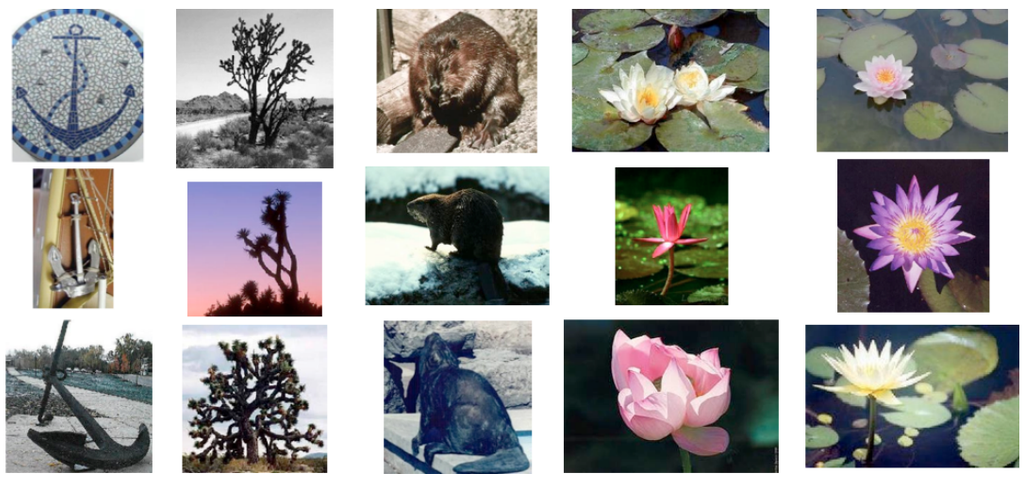
- Music Genre Classification (GTZAN) [IEEE Acoustic '13]

CLASSIFICATION WITH SCATTERING

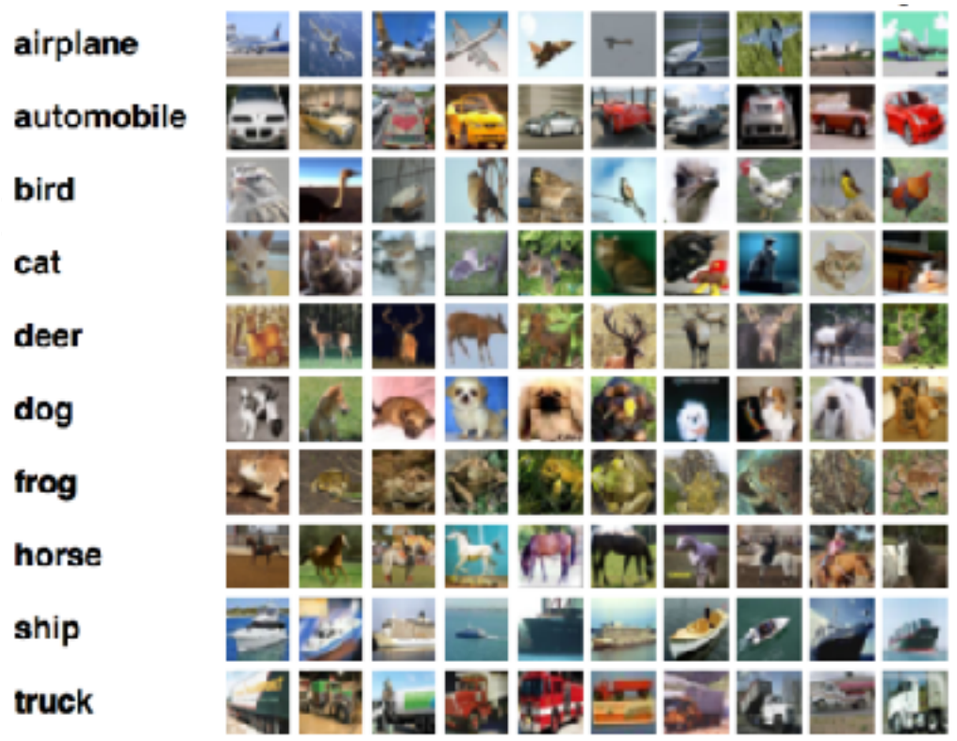
- Joint Scattering Improves Performance:
  - More complicate[illegible]UIUC,UMD)
    [Sifre&Mallat, C[illegible]]



- Small-mid scale Object Recognit[illegible]5]



0



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.