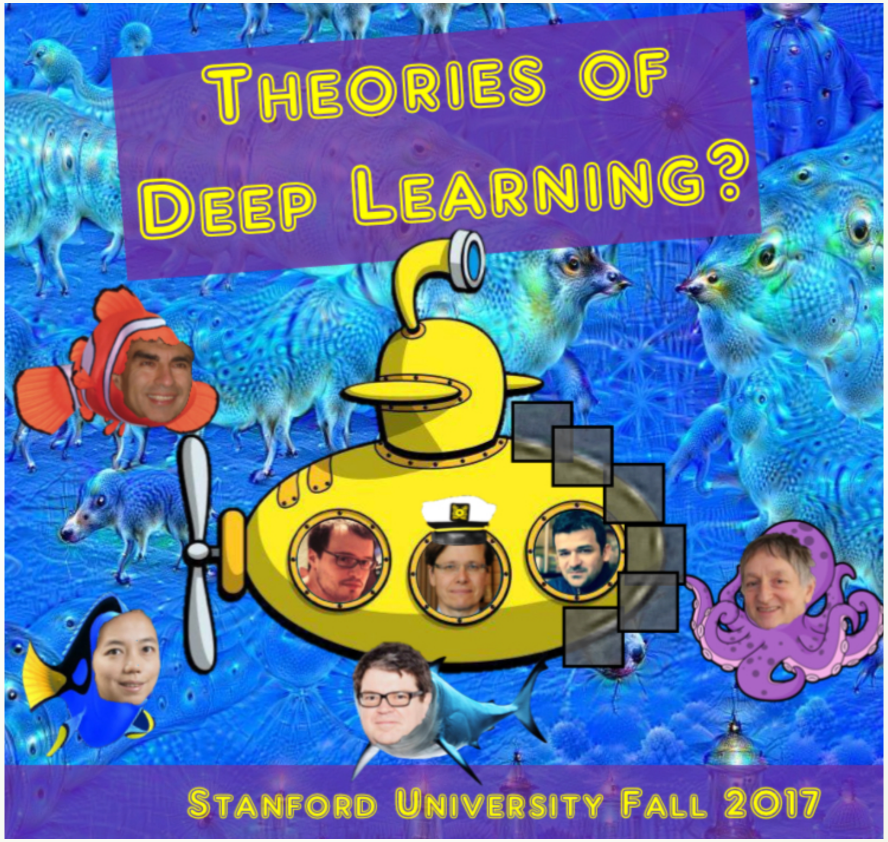
Harmonic Analysis of Deep Convolutional Networks

1

Yuan YAO
HKUST
Based on Mallat and Bolcskei talks etc.

Acknowledgement



A following-up course at HKUST: https://deeplearning-math.github.io/

High Dimensional Natural Image Classification

- High-dimensional $x = (x(1), ..., x(d)) \in \mathbb{R}^d$:

- **Classification:** estimate a class label $f(x)$ given n sample values $\{x_i\,,\, y_i = f(x_i)\}_{i \le n}$

Image Classification $\quad d = 10^6$

Huge variability inside classes

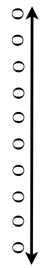
Find invariants

Curse of Dimensionality

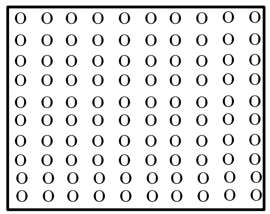
- Analysis in high dimension: $x \in \mathbb{R}^d$ with $d \geq 10^6$.

- Points are far away in high dimensions d:

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



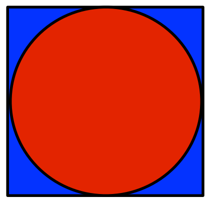
  - 100 points for $[0,1]^2$



  - need 10^d points over $[0,1]^d$

    impossible if $d \geq 20$

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



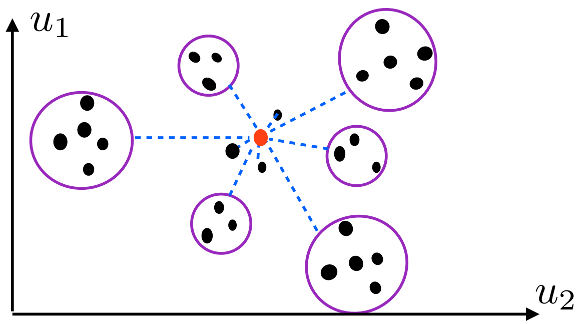
points are concentrated in 2^d corners!

$\Rightarrow$ Euclidean metrics are not appropriate on **raw data**.

A Blessing from Physical world? Multiscale "compositional" sparsity

- Variables $x(u)$ indexed by a low-dimensional u: time/space... pixels in images, particles in physics, words in text...

- Mutliscale interactions of d variables:



From d^2 interactions to $O(\log^2 d)$ multiscale interactions.

- Multiscale analysis: wavelets on groups of symmetries.
  hierarchical architecture.

Learning as an Approximation

- To estimate $f(x)$ from a sampling $\{x_i\ ,\ y_i = f(x_i)\}_{i \leq M}$

  we must build an M-parameter approximation f_M of f.

- Precise sparse approximation requires some "regularity".

- For binary classification $f(x) = \begin{cases} 1 & \text{if } x \in \Omega \\ -1 & \text{if } x \notin \Omega \end{cases}$

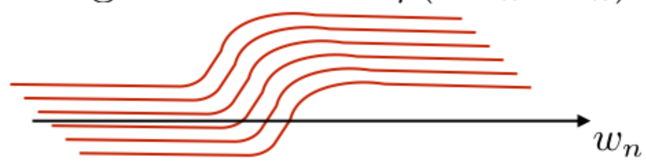
$$f(x) = \mathrm{sign}(\tilde{f}(x))$$

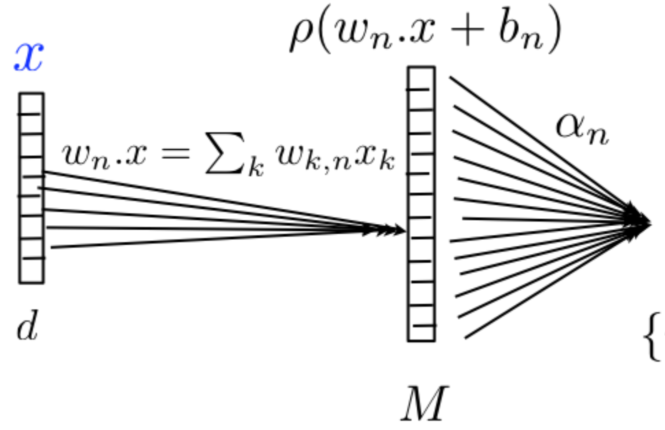
  where $\tilde{f}$ is potentially regular.

- What type of regularity ? How to compute f_M ?

1 Hidden Layer Neural Networks

One-hidden layer neural network: ridge functions $\rho(x.w_n + b_n)$





$$f_M(x) = \sum_{n=1}^{M} \alpha_n \, \rho(w_n.x + b_n)$$

$\{w_{k,k}\}_{k,n}$ and $\{\alpha_n\}_n$ are learned

non-linear approximation.

Cybenko, Hornik, Stinchcombe, White

Theorem: For "resonnable" bounded $\rho(u)$

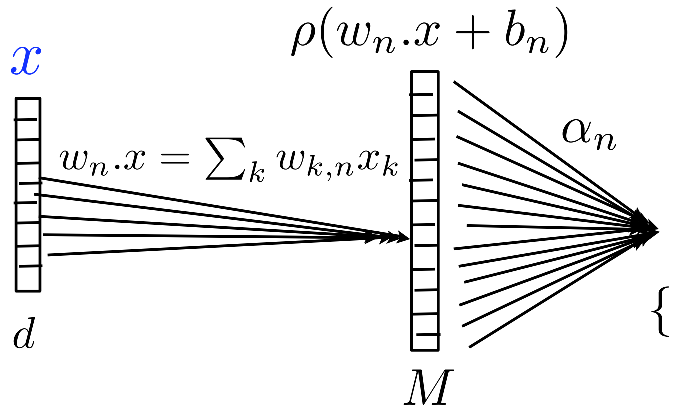
and appropriate choices of $w_{n,k}$ and α_n:

$$\forall f \in \mathbb{L}^2[0,1]^d \qquad \lim_{M \to \infty} \|f - f_M\| = 0 \ .$$

No big deal: curse of dimensionality still there.

1 Hidden Layer Neural Networks

One-hidden layer neural network:



$$f_M(x) = \sum_{n=1}^{M} \alpha_n \, \rho(w_n.x + b_n)$$

$\{w_{k,k}\}_{k,n}$ and $\{\alpha_n\}_n$ are learned

non-linear approximation.

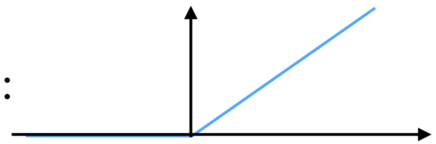
Fourier series: $\rho(u) = e^{iu}$

$$f_M(x) = \sum_{n=1}^{M} \alpha_n \, e^{i w_n . x}$$

For nearly all ρ: essentially same approximation results.

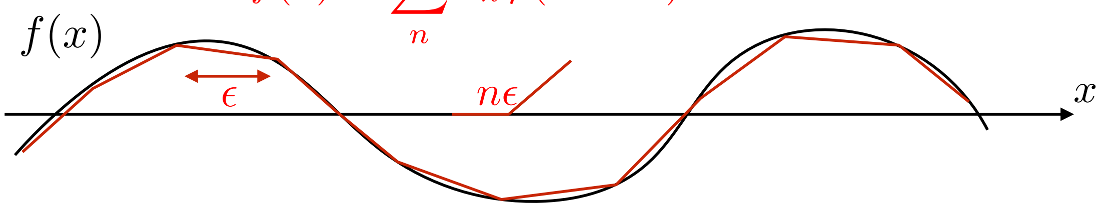
Piecewise Linear Approximation

- Piecewise linear approximation:



$\rho(u) = \max(u, 0)$

$$\tilde{f}(x) = \sum_n a_n \, \rho(x - n\epsilon)$$



If f is Lipschitz: $|f(x) - f(x')| \leq C\,|x - x'|$

$$\Rightarrow \quad |f(x) - \tilde{f}(x)| \leq C\,\epsilon.$$

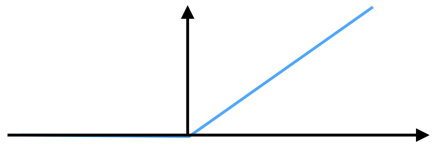
Need $M = \epsilon^{-1}$ points to cover $[0,1]$ at a distance ϵ

$$\Rightarrow \|f - f_M\| \leq C\,M^{-1}$$

Linear Ridge Approximation

- Piecewise linear ridge approximation: $x \in [0,1]^d$

$$\tilde f(x) = \sum_n a_n \, \rho(w_n . x - n\epsilon)$$



$\rho(u) = \max(u, 0)$

If f is Lipschitz: $|f(x) - f(x')| \leq C \, \|x - x'\|$

Sampling at a distance ϵ:

$$\Rightarrow \quad |f(x) - \tilde f(x)| \leq C \, \epsilon .$$

need $M = \epsilon^{-d}$ points to cover $[0,1]^d$ at a distance ϵ

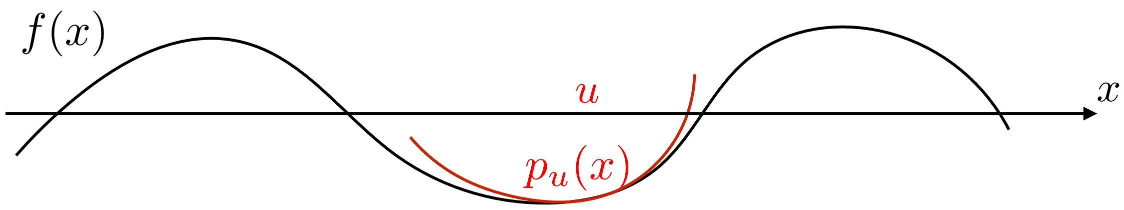
$$\Rightarrow \; \|f - f_M\| \leq C \, M^{-1/d}$$

Curse of dimensionality!

Approximation with Regularity

- What prior condition makes learning possible ?
- Approximation of regular functions in $\mathbf{C}^s[0,1]^d$:

$$\forall x, u \quad |f(x) - p_u(x)| \leq C\,|x-u|^s \quad \text{with} \quad p_u(x) \quad \text{polynomial}$$



$$|x-u| \leq \epsilon^{1/s} \quad \Rightarrow \quad |f(x) - p_u(x)| \leq C\,\epsilon$$

Need $M^{-d/s}$ point to cover $[0,1]^d$ at a distance $\epsilon^{1/s}$

$$\Rightarrow \|f - f_M\| \leq C\,M^{-s/d}$$

- Can not do better in $\mathbf{C}^{\mathbf{s}}[0,1]^d$, not good because $s \ll d$.

Failure of classical approximation theory.

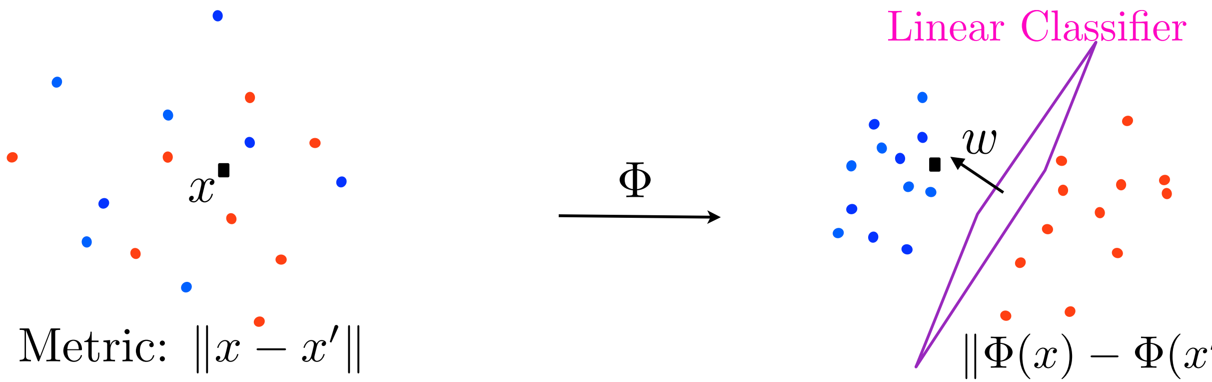
Kernel Learning

Change of variable $\Phi(x) = \{\phi_k(x)\}_{k \leq d'}$

to nearly linearize $f(x)$, which is approximated by:

$$\tilde{f}(x) = \underbrace{\langle \Phi(x)\,,\, w \rangle}_{\textbf{1D projection}} = \sum_k w_k\, \phi_k(x)\ .$$

Data: $x \in \mathbb{R}^d$ $\Phi(x) \in \mathbb{R}^{d'}$



Metric: $\|x - x'\|$ $\|\Phi(x) - \Phi(x')\|$

- How and when is possible to find such a Φ ?
- What "regularity" of f is needed ?

Increase Dimensionality

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

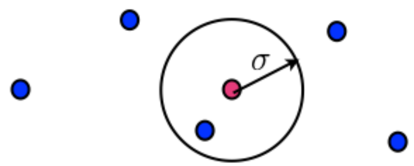
$\Rightarrow$ Choose Φ increasing dimensionality !

Problem: generalisation, overfitting.

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

$\Phi(x)$ is of dimension $d' = \infty$

If σ is small, nearest neighbor classifier type:



Spirit in Fisher's Linear Discriminant Analysis

Reduction of Dimensionality

- Discriminative change of variable $\Phi(x)$:

$$\Phi(x) \neq \Phi(x') \ \text{ if } \ f(x) \neq f(x')$$

$$\Rightarrow \quad \exists \tilde{f} \ \text{ with } \ f(x) = \tilde{f}(\Phi(x))$$

- If $\tilde{f}$ is Lipschitz: $|\tilde{f}(z) - \tilde{f}(z')| \leq C \, \|z - z'\|$

$$z = \Phi(x) \quad \Leftrightarrow \quad |f(x) - f(x')| \leq C \, \|\Phi(x) - \Phi(x')\|$$

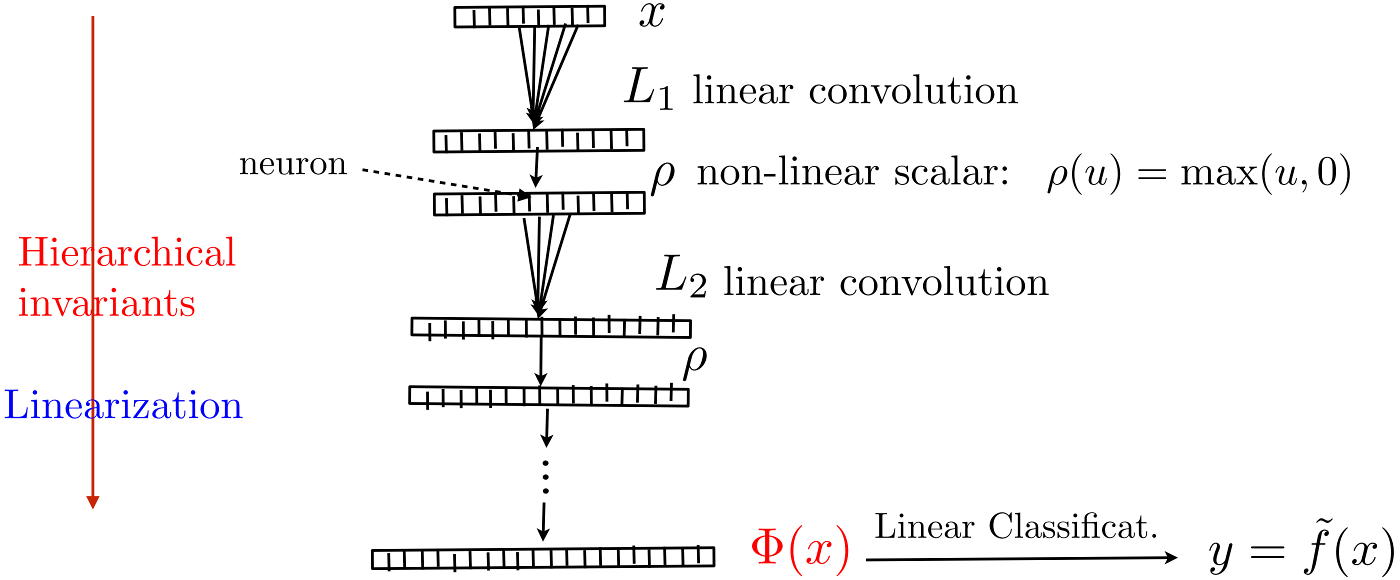
Discriminative: $\|\Phi(x) - \Phi(x')\| \geq C^{-1} \, |f(x) - f(x')|$

- For $x \in \Omega$, if $\Phi(\Omega)$ is bounded and a low dimension d'

$$\Rightarrow \|f - f_M\| \leq C \, M^{-1/d'}$$

Deep Convolution Neworks

- The revival of neural networks: *Y. LeCun*

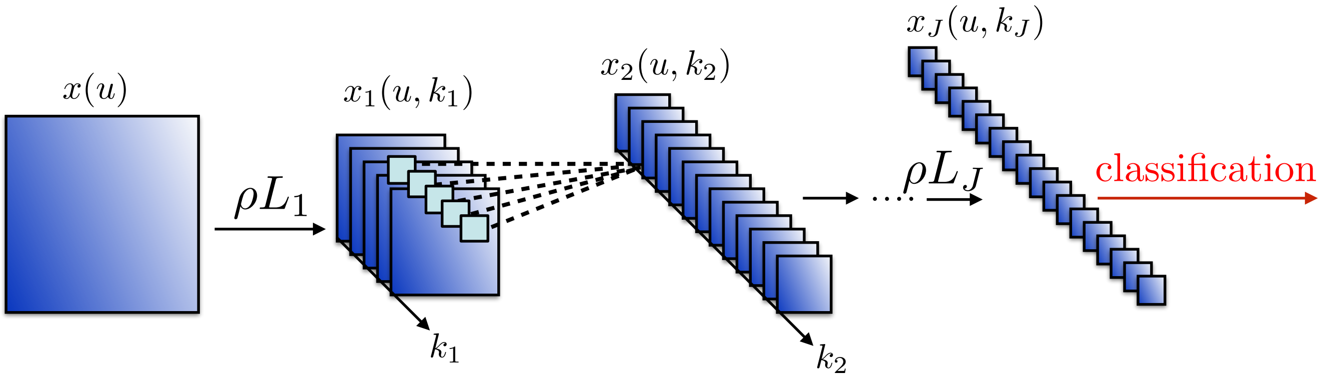


Optimize L_j with architecture constraints: over 10^9 parameters

Exceptional results for *images, speech, language, bio-data...*

Why does it work so well ? **A difficult problem**

Deep Convolutional Networks



$$x_j = \rho \, L_j \, x_{j-1}$$

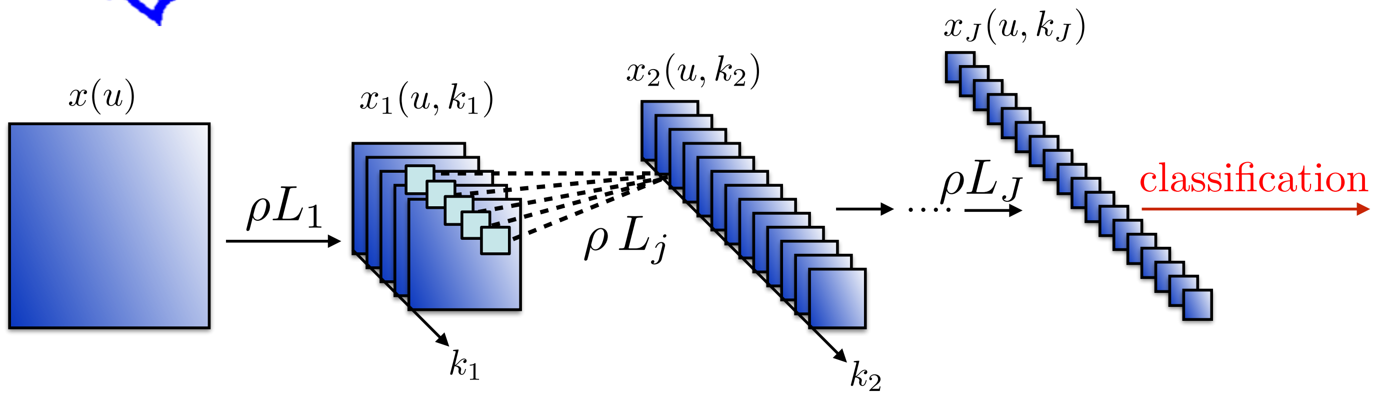
- L_j is a linear combination of convolutions and subsampling:

$$x_j(u, k_j) = \rho\Big(\underbrace{\sum_k}_{\text{sum across channels}} x_{j-1}(\cdot, k) \star h_{k_j,k}(u) \Big)$$

- ρ is contractive: $|\rho(u) - \rho(u')| \leq |u - u'|$

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

Many Questions



- Why convolutions ? Translation covariance.
- Why no overfitting ? Contractions, dimension reduction

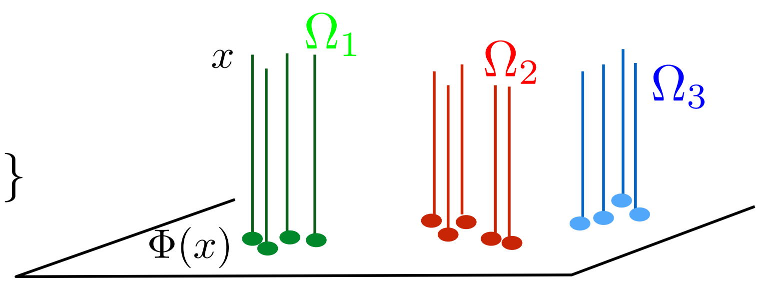
- Why hierarchical cascade ?
- Why introducing non-linearities ?
- How and what to linearise ?
- What are the roles of the multiple channels in each layer ?

Linear Dimension Reduction

Classes

Level sets of $f(x)$

$$\Omega_t = \{x \ : \ f(x) = t\}$$



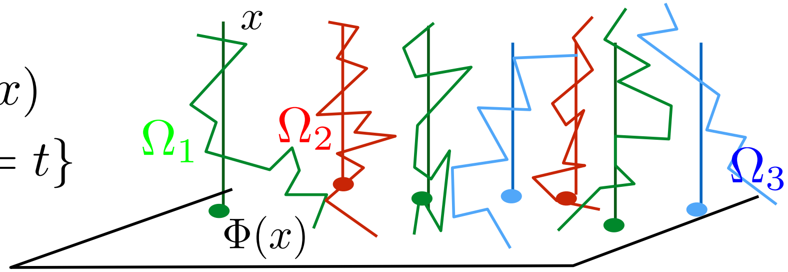
If level sets (classes) are parallel to a linear space
then variables are eliminated by linear projections: *invariants*.

Linearise for Dimensionality Reduction

Classes

Level sets of $f(x)$

$$\Omega_t = \{x \ : \ f(x) = t\}$$

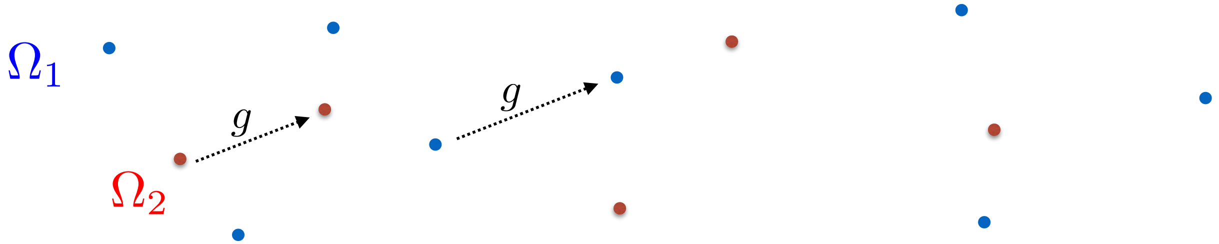


- If level sets Ω_t are not parallel to a linear space
  - Linearise them with a change of variable $\Phi(x)$
  - Then reduce dimension with linear projections
- Difficult because Ω_t are high-dimensional, irregular, known on few samples.

Level Set Geometry: Symmetries

- Curse of dimensionality $\Rightarrow$ not local but global geometry

  Level sets: classes, characterised by their global symmetries.



- A symmetry is an operator g which preserves level sets:

$$\forall x \quad , \quad f(g.x) = f(x) \ : \ \text{global}$$

If g_1 and g_2 are symmetries then $g_1.g_2$ is also a symmetry

$$f(g_1.g_2.x) = f(g_2.x) = f(x)$$

Groups of symmetries

- $G = \{$ all symmetries $\}$ is a group: unknown

$$\forall (g, g') \in G^2 \quad \Rightarrow g.g' \in G$$

Inverse: $\quad \forall g \in G \quad , \quad g^{-1} \in G$

Associative: $\quad (g.g').g'' = g.(g'.g'')$

If commutative $\; g.g' = g'.g \;$: Abelian group.

- Group of dimension n if it has n generators:

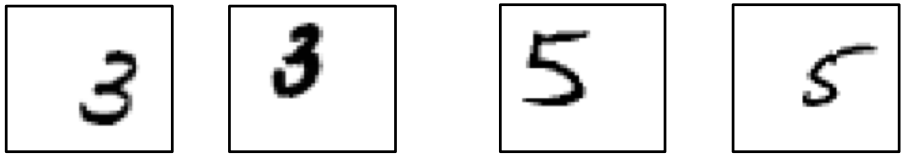
$$g = g_1^{p_1}\, g_2^{p_2} \ldots g_n^{p_n}$$

- Lie group: infinitely small generators (Lie Algebra)

Translation and Deformations

- Digit classification:

$x(u)$ $\quad$ $x'(u) = x(u - \tau(u))$

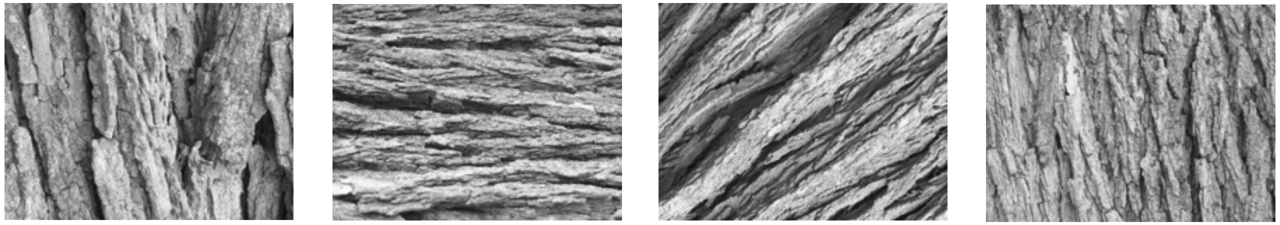
Ω_3  Ω_5

- Globally invariant to the translation group: small
- Locally invariant to small diffeomorphisms: huge group

Video of Philipp Scott Johnson

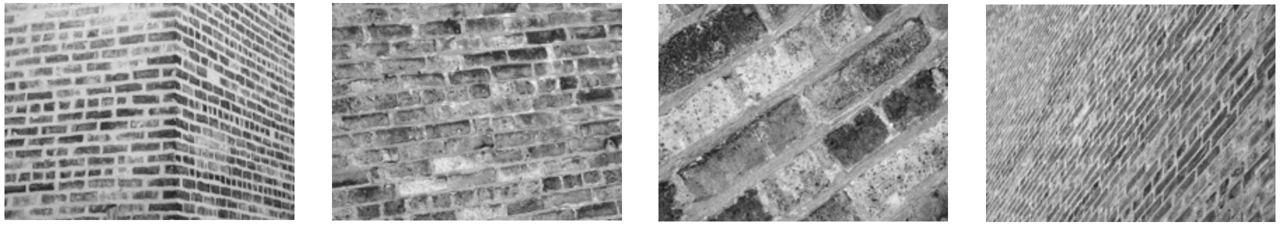
Rotation and Scaling Variability

- Rotation and deformations



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

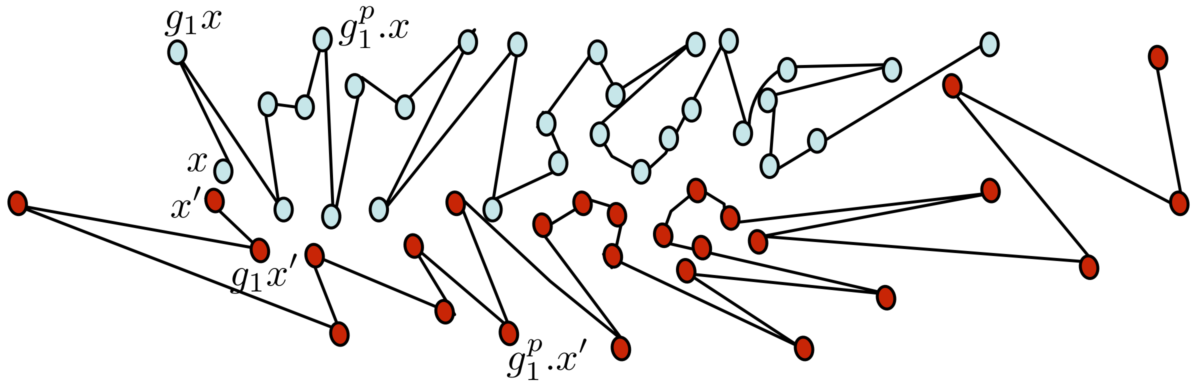
- Scaling and deformations



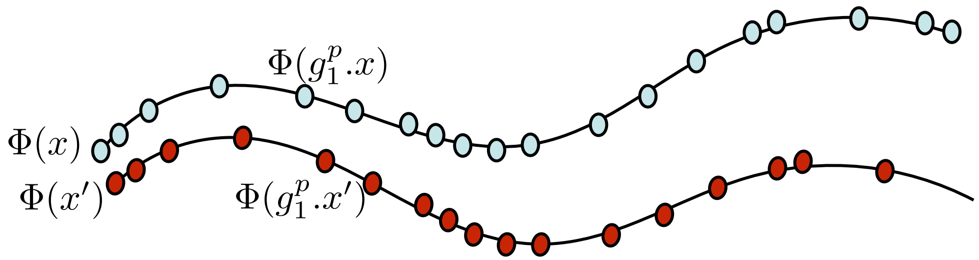
Group: $\mathbb{R} \times \mathrm{Diff}(\mathbb{R})$

Linearize Symmetries

- A change of variable $\Phi(x)$ must linearize the orbits $\{g.x\}_{g \in G}$



- Linearise symmetries with a change of variable $\Phi(x)$

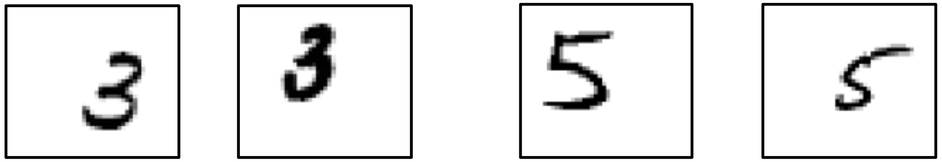


- Lipschitz: $\forall x, g \; : \; \|\Phi(x) - \Phi(g.x)\| \leq C \, \|g\|$

Translation and Deformations

- Digit classification:

$x(u)$ $x'(u)$



- Globally invariant to the translation group
- Locally invariant to small diffeomorphisms

Linearize small diffeomorphisms:
$\Rightarrow$ Lipschitz regular

Video of Philipp Scott Johnson

Translations and Deformations

- Invariance to translations:

$$g.x(u) = x(u - c) \quad \Rightarrow \quad \Phi(g.x) = \Phi(x) \ .$$

- Small diffeomorphisms: $g.x(u) = x(u - \tau(u))$

  Metric: $\|g\| = \|\nabla\tau\|_\infty$ maximum scaling

  Linearisation by Lipschitz continuity

$$\|\Phi(x) - \Phi(g.x)\| \leq C \, \|\nabla\tau\|_\infty \ .$$

- Discriminative change of variable:

$$\|\Phi(x) - \Phi(x')\| \geq C^{-1} \, |f(x) - f(x')|$$

Fourier Deformation Instability

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

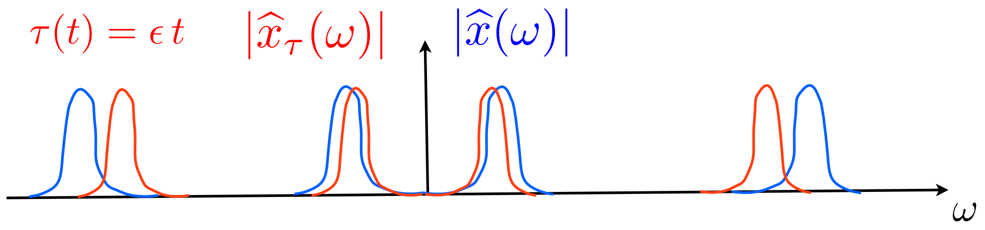
$$x_c(t) = x(t-c) \quad \Rightarrow \quad \hat{x}_c(\omega) = e^{-ic\omega}\, \hat{x}(\omega)$$

The modulus is invariant to translations:

$$\Phi(x) = |\hat{x}| = |\hat{x}_c|$$

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

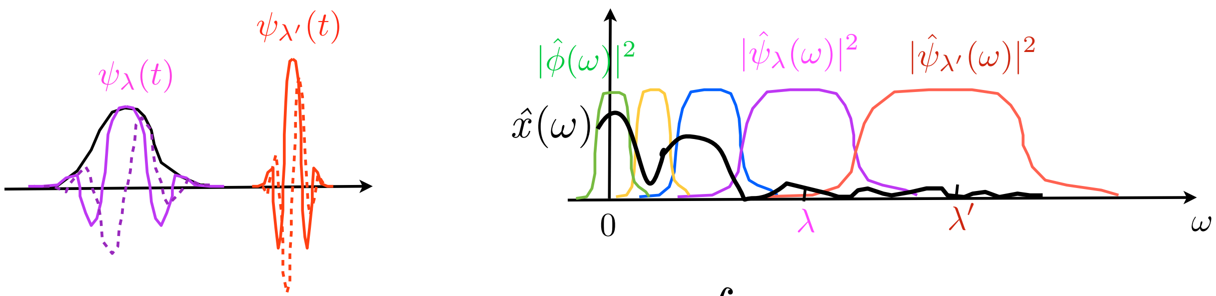
$$\big|\, |\hat{x}_\tau(\omega)| - |\hat{x}(\omega)| \,\big| \quad \text{is big at high frequencies}$$



$$\Rightarrow \quad \|\,|\hat{x}| - |\hat{x}_\tau|\,\| \gg \|\nabla\tau\|_\infty\, \|x\|$$

Wavelet Transform

- Complex wavelet: $\psi(t) = \psi^a(t) + i\,\psi^b(t)$
- Dilated: $\psi_\lambda(t) = 2^{-j}\,\psi(2^{-j}t)$ with $\lambda = 2^{-j}$.



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

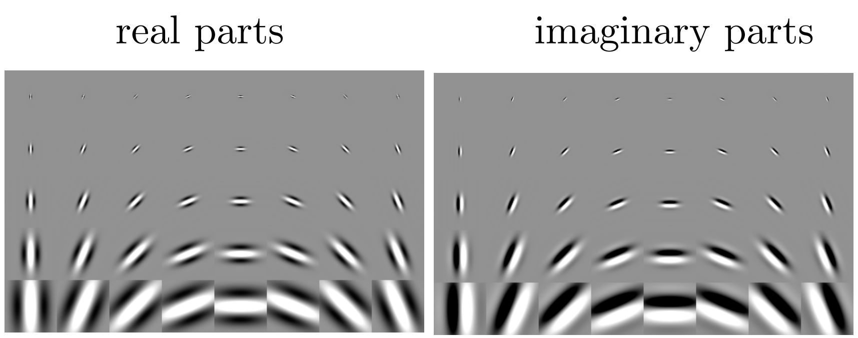
$$Wx = \begin{pmatrix} x \star \phi(t) \\ x \star \psi_\lambda(t) \end{pmatrix}_{t,\lambda}$$

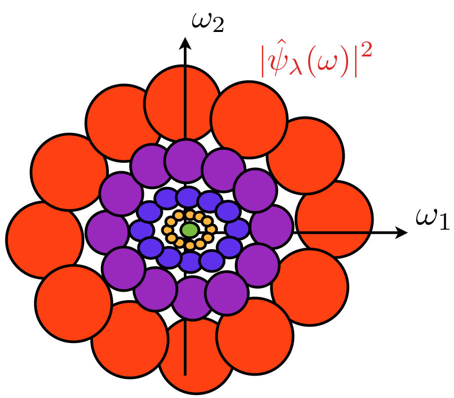
Unitary: $\|Wx\|^2 = \|x\|^2$.

Image Wavelet Transform

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

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





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

  Unitary: $\|Wx\|^2 = \|x\|^2$.

Why Wavelets ?

- Wavelets are uniformly stable to deformations:

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

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

- Wavelets separate multiscale information.

- Wavelets provide sparse representations.

Why Wavelets?

- Wavelets are uniformly **stable to deformations**

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

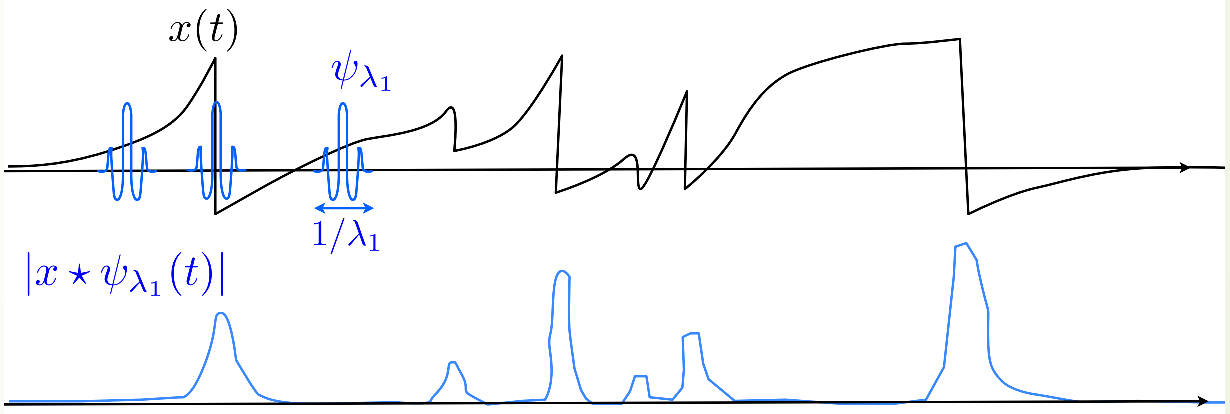
$$\|\psi_\lambda - \psi_{\lambda,\tau}\| \leq C \sup_t |\nabla\tau(t)| \ .$$

- Wavelets are **sparse** representations of functions
- Wavelets separate **multiscale** information
- Wavelets can be locally **translation invariant**

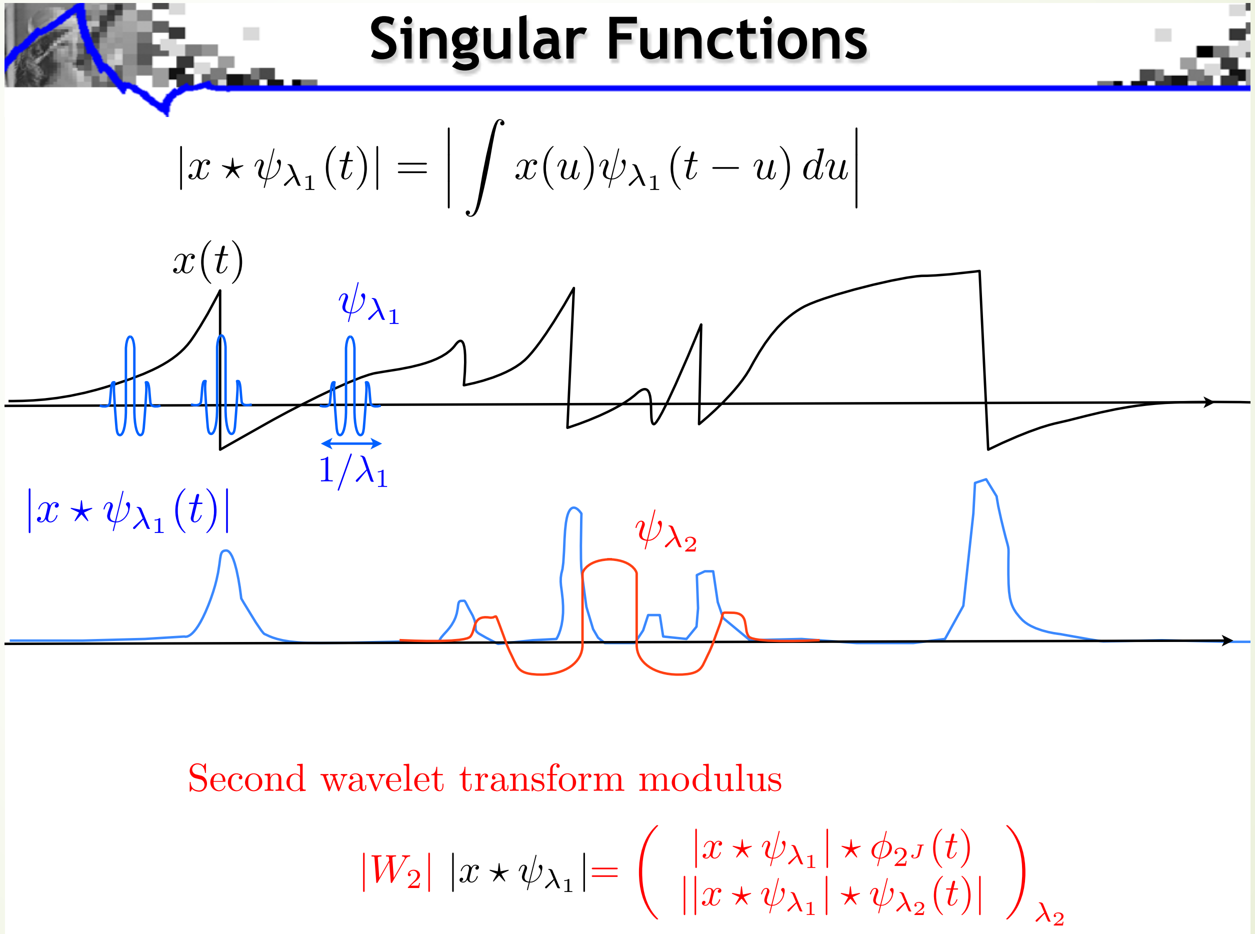
Sparsity of Wavelet Transforms

Singular Functions

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

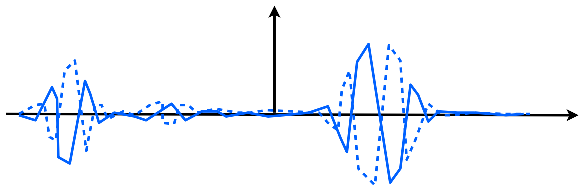


Singularity is preserved in multiscale transform



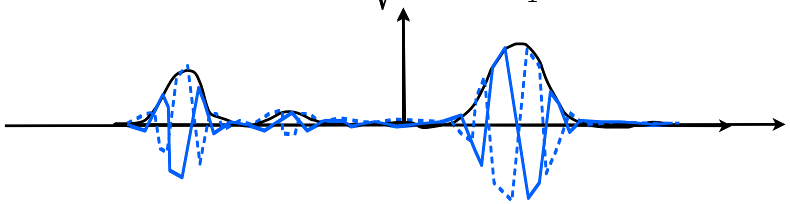
Wavelet Translation Invariance

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



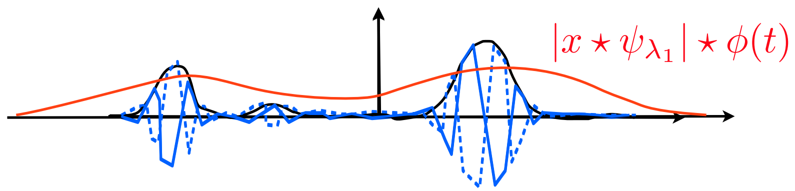
Wavelet Translation Invariance

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



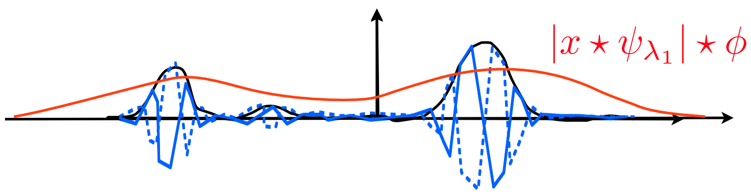
- The modulus $|x \star \psi_{\lambda_1}|$ is a regular envelop

Wavelet Translation Invariance



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

Wavelet Translation Invariance



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

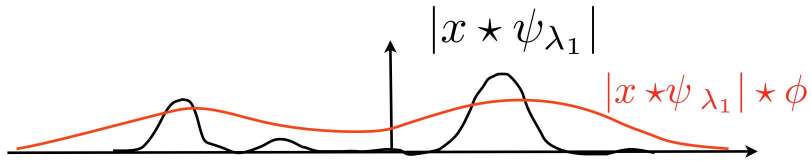
- The average $|x \star \psi_{\lambda_1}| \star \phi(t)$ is invariant to small translations relatively to the support of ϕ.

- Full translation invariance at the limit:

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

but few invariants.

Recovering Lost Information



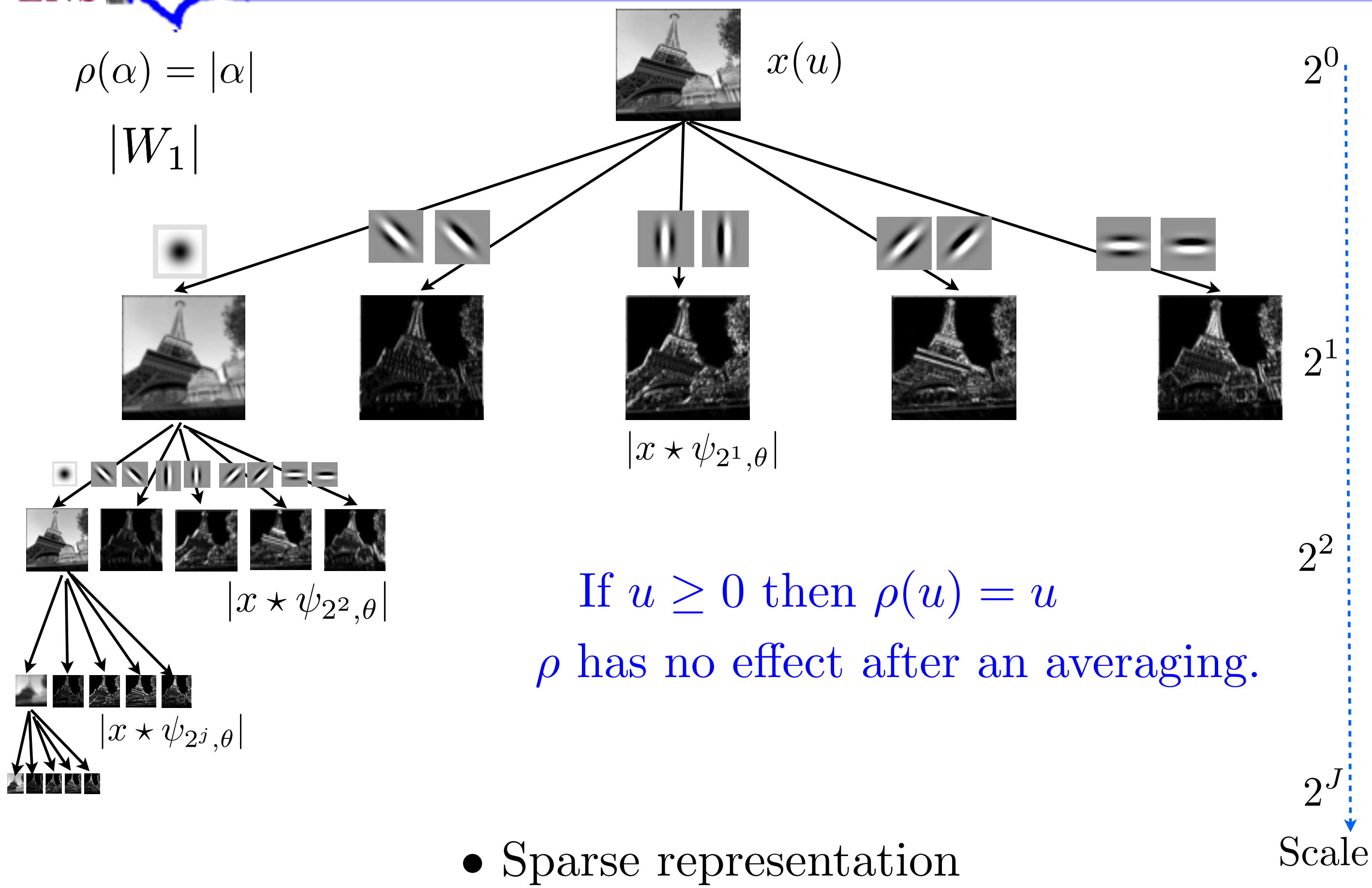
- The high frequencies of $|x \star \psi_{\lambda_1}|$ are in wavelet coefficients:

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

- Translation invariance by time averaging the amplitude:

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

Wavelet Filter Bank



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

ρ has no effect after an averaging.

- Sparse representation

Contraction

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

$$\rho(u) = |u|$$

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

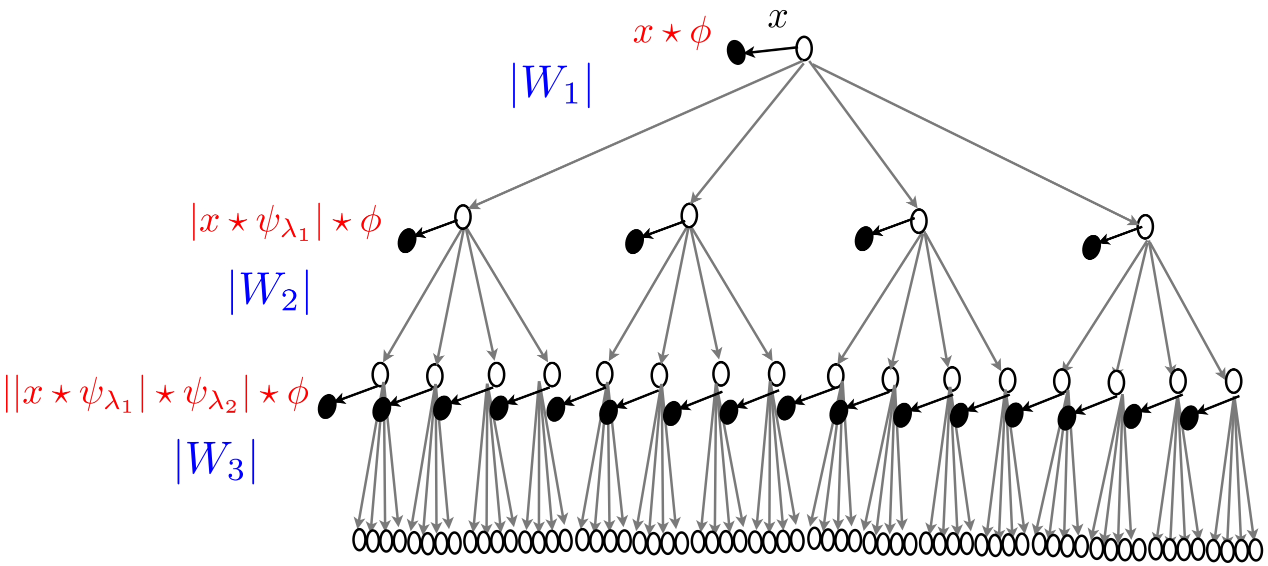
- it is contractive $\||W|x - |W|y\| \leq \|x - y\|$

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

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

Wavelet Scattering Network

Cascade of Contractions



- Cascade of contractive operators

$$\| |W_k| x - |W_k| x' \| \le \|x - x'\| \quad \text{with} \quad \| |W_k| x \| = \|x\| \ .$$

Stability of Wavelet Scattering Transform

Scattering Properties

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

Theorem: *For appropriate wavelets, a scattering is*

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

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

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

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

$\Rightarrow$ linear discriminative classification from $\Phi x = Sx$