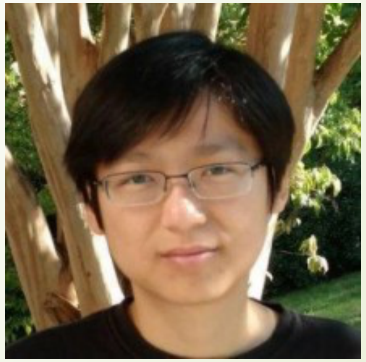
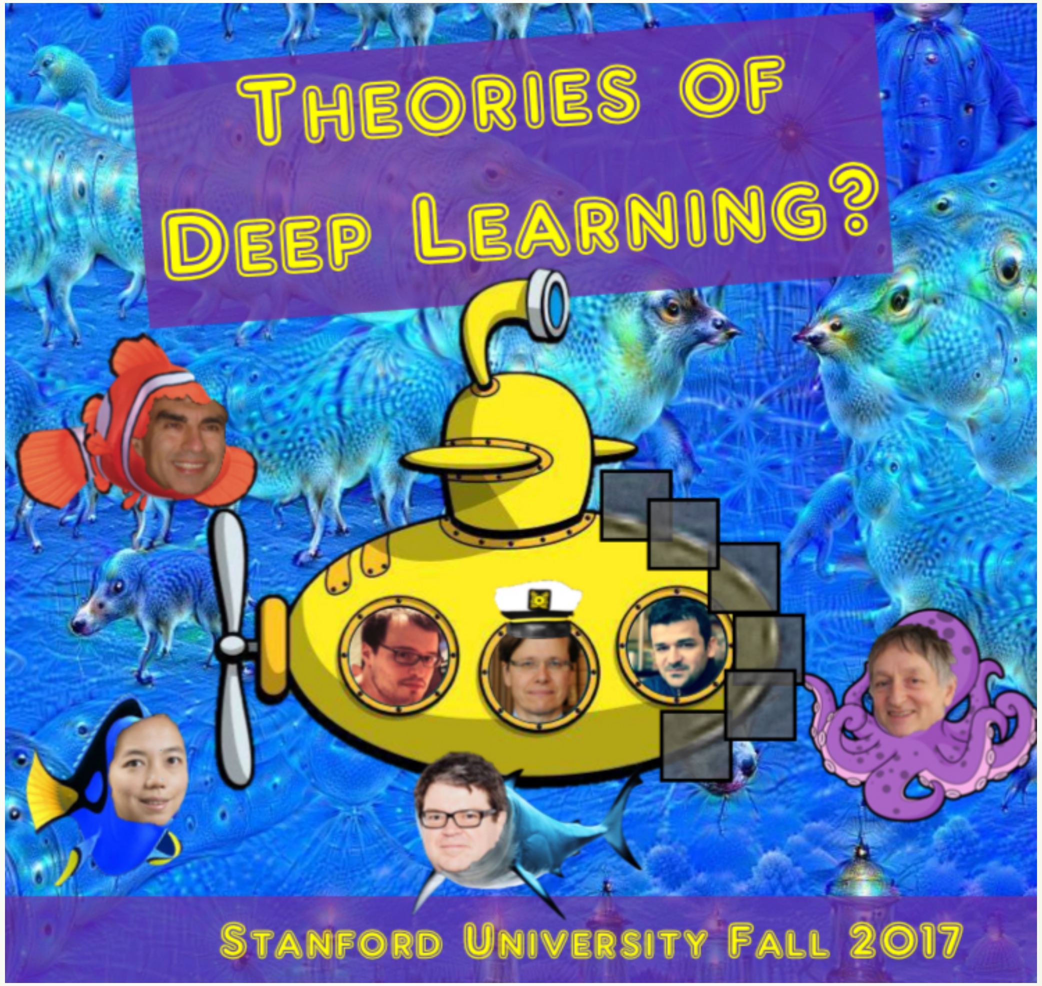
Symmetry and Network Architectures

Yuan YAO
HKUST
Based on Mallat, Bolcskei, Cheng talks etc.



Acknowledgement



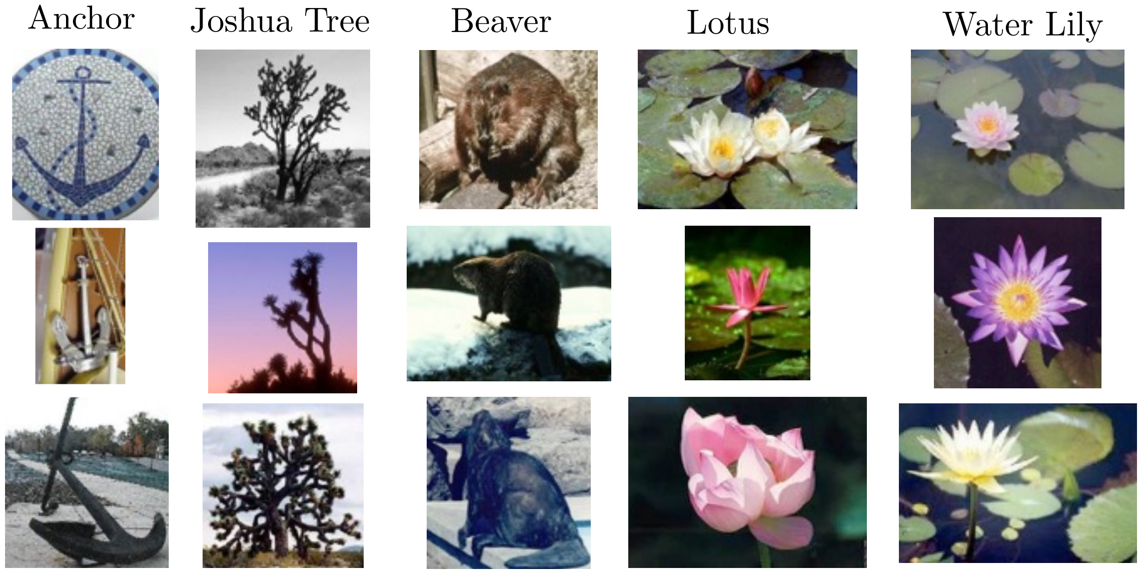
A following-up course at HKUST: https://deeplearning-math.github.io/

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 $\{x_i \, , \, y_i = f(x_i)\}_{i \leq n}$

Image Classification $d = 10^6$



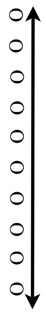
Huge variability inside classes

Find invariants

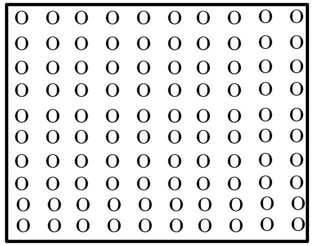
Curse of Dimensionality

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

- Points are far away in high dimensions d:



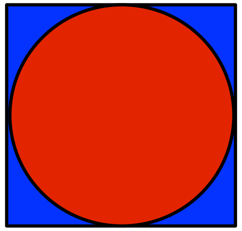
  - 10 points cover $[0,1]$ at a distance 10^{-1}
  - 100 points for $[0,1]^2$



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

    impossible if $d \geq 20$

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



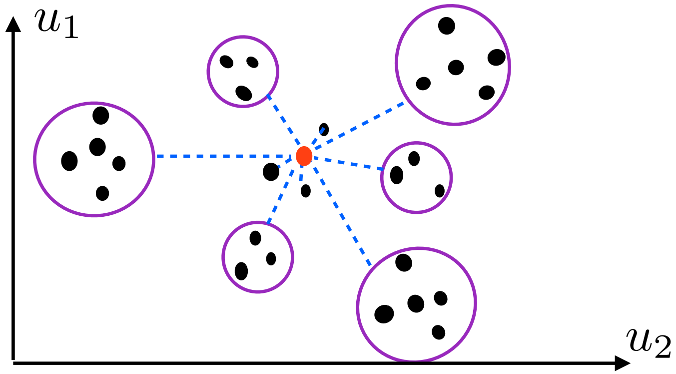
points are concentrated in 2^d corners!

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

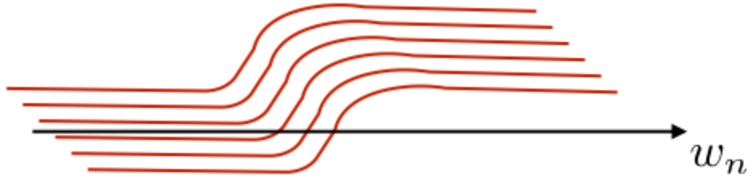
$$f(x) = \mathrm{sign}(\tilde{f}(x))$$

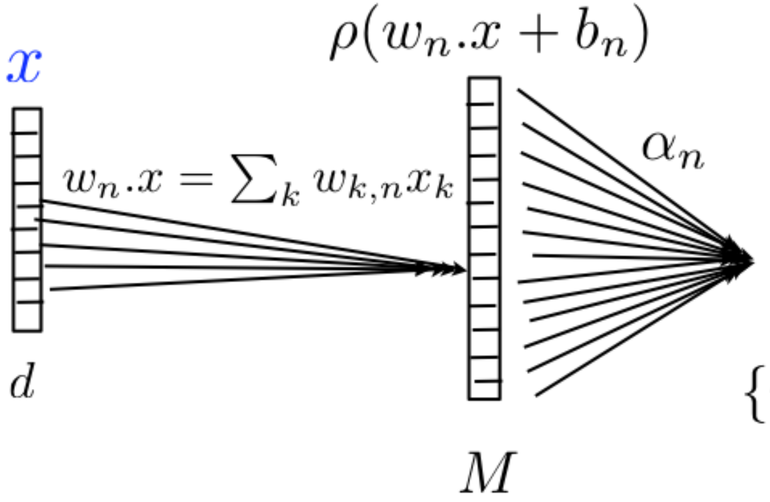
  where $\tilde{f}$ is potentially regular.

- 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 $\rho(u)$

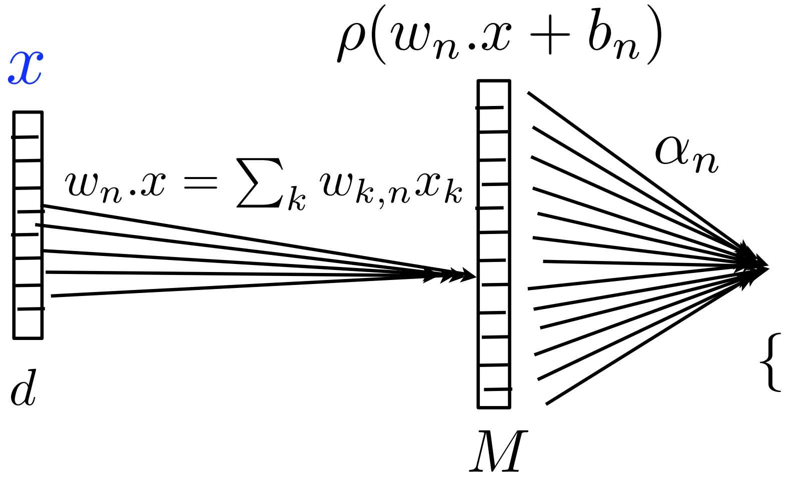
and appropriate choices of $w_{n,k}$ and α_n:

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

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.

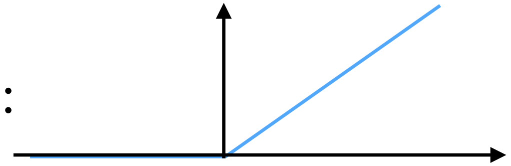
Fourier series: $\rho(u) = e^{iu}$

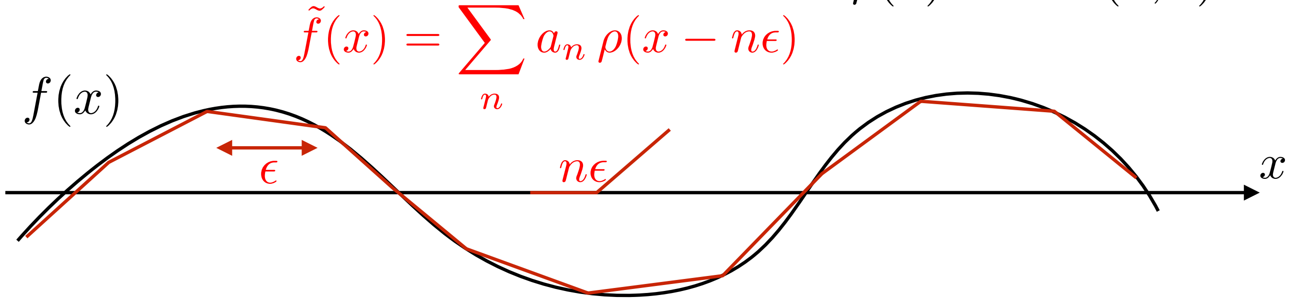
$$f_M(x) = \sum_{n=1}^{M} \alpha_n \, e^{i w_n . x}$$

For nearly all ρ: essentially same approximation results.

Piecewise Linear Approximation

- Piecewise linear approximation:





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

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

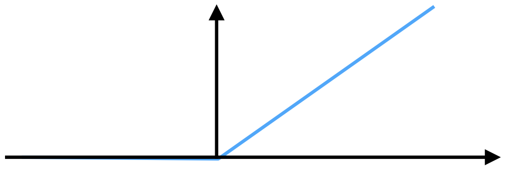
Need $M = \epsilon^{-1}$ points to cover $[0, 1]$ at a distance ϵ

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

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

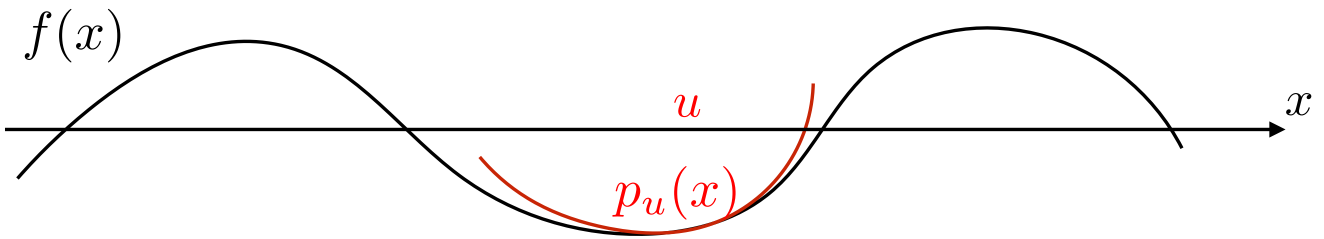
$$\Rightarrow \; \|f - f_M\| \leq C \, M^{-1/d}$$

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 $\mathbf{C}^s[0,1]^d$, not good because $s \ll d$.

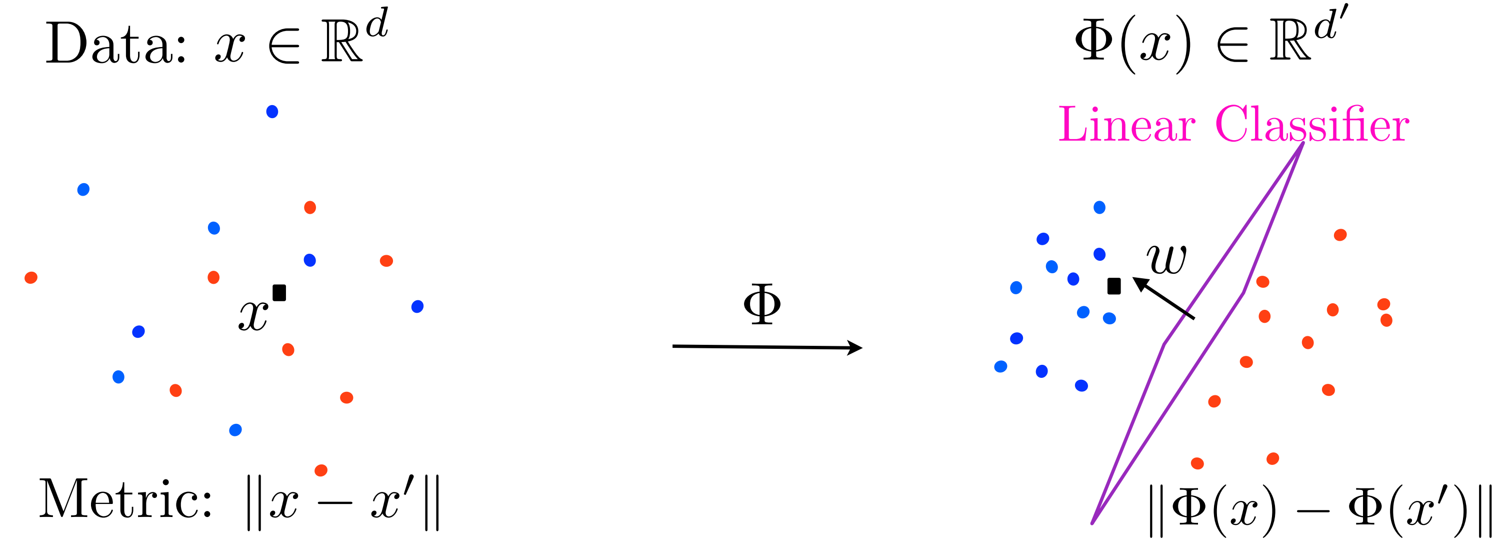
Failure of classical approximation theory.

Kernel Learning

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

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

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



- 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 $< N$.

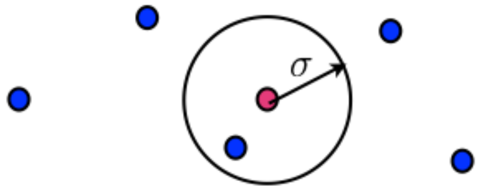
$\Rightarrow$ Choose Φ increasing dimensionality !

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 $\Phi(x)$:

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

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

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

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

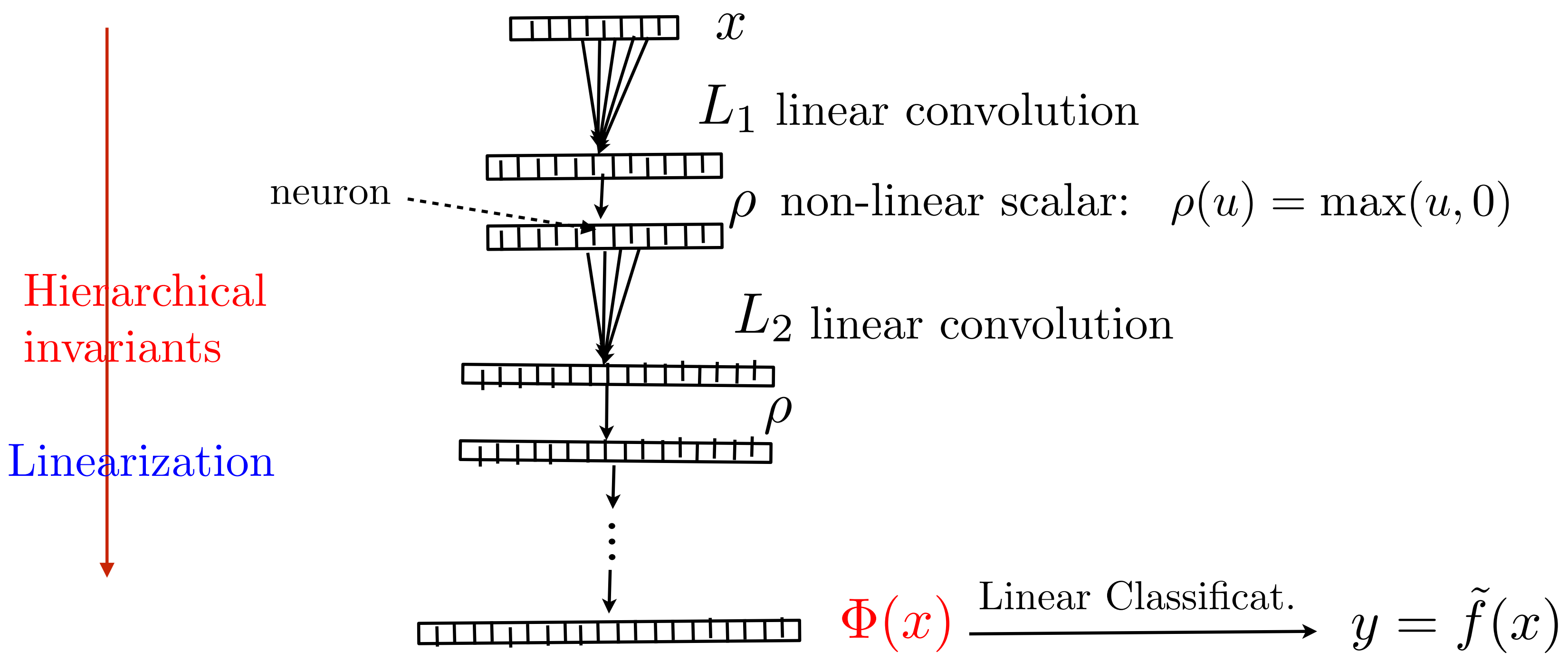
Discriminative: $\|\Phi(x) - \Phi(x')\| \geq C^{-1} \, |f(x) - f(x')|$

- 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*

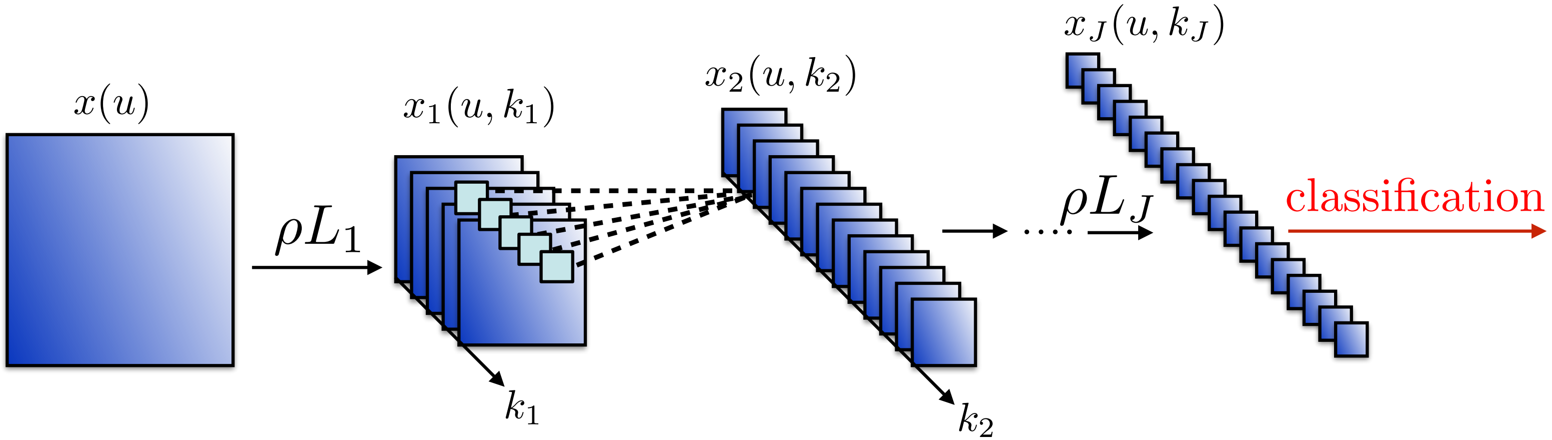


Optimize L_j with architecture constraints: over 10^9 parameters

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

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

Deep Convolutional Networks



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

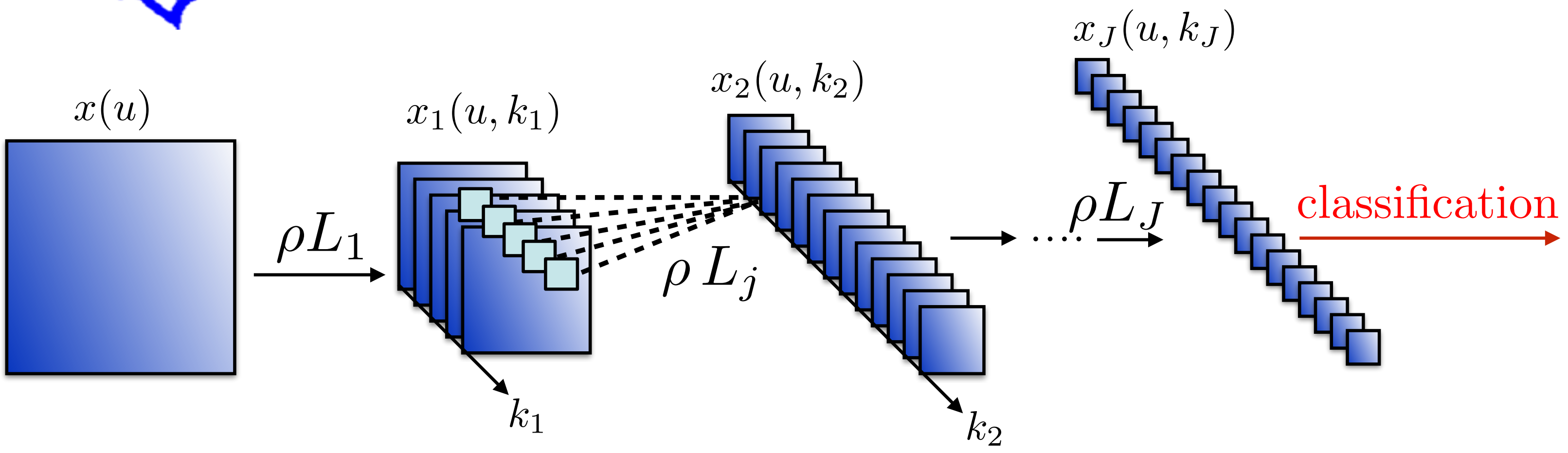
- L_j is a linear combination of convolutions and subsampling:

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

- 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

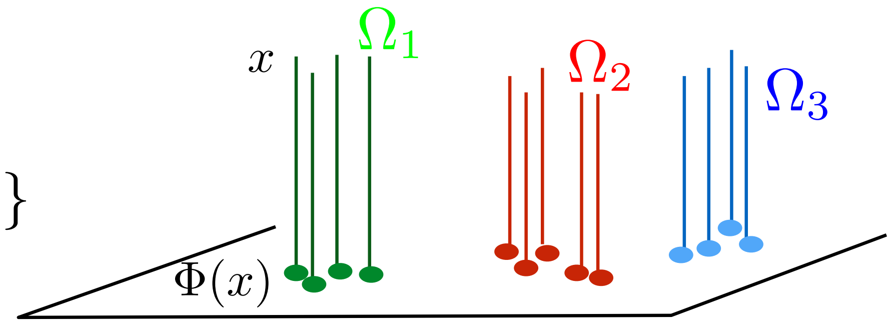
- Why hierarchical cascade ?
- Why introducing non-linearities ?
- How and what to linearise ?
- What are the roles of the multiple channels in each layer ?

Linear Dimension Reduction

Classes

Level sets of $f(x)$

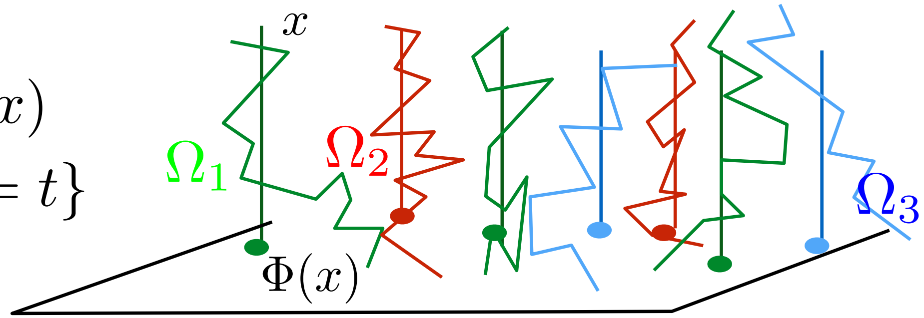
$$\Omega_t = \{x \ : \ f(x) = t\}$$



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

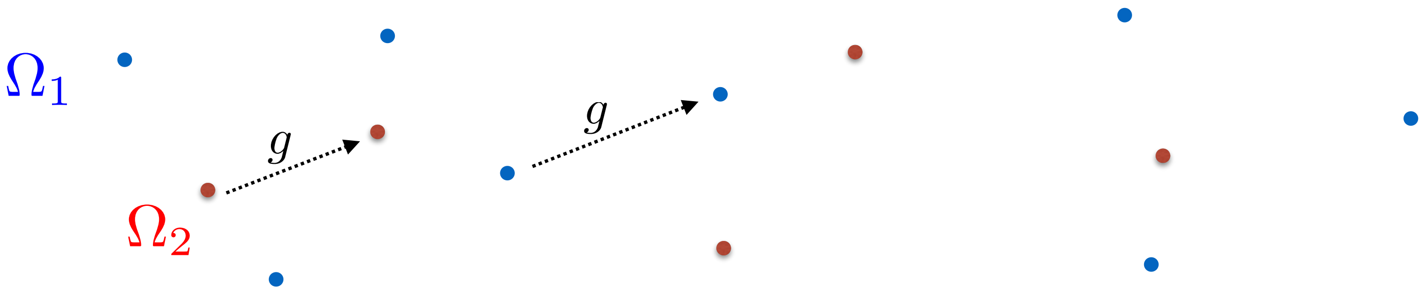


- If level sets Ω_t are not parallel to a linear space
  - Linearise them with a change of variable $\Phi(x)$
  - Then reduce dimension with linear projections

- Difficult because Ω_t are high-dimensional, irregular, known on few samples.

Level Set Geometry: Symmetries

- Curse of dimensionality ⇒ 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: $\forall g \in G \quad , \quad g^{-1} \in G$

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

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

- Group of dimension n if it has n generators:

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

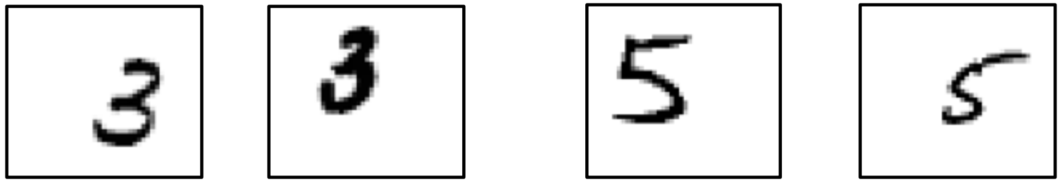
- Lie group: infinitely small generators (Lie Algebra)

Translation and Deformations

- Digit classification:

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

Ω_3



Ω_5

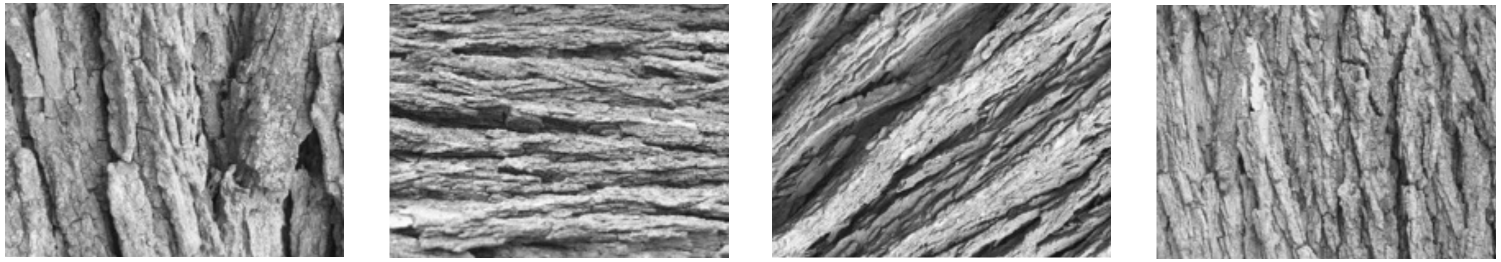
- Globally invariant to the translation group: small
- Locally invariant to small diffeomorphisms: huge group

Video of Philipp Scott Johnson

https://www.youtube.com/watch?v=nUDIoN-_Hxs

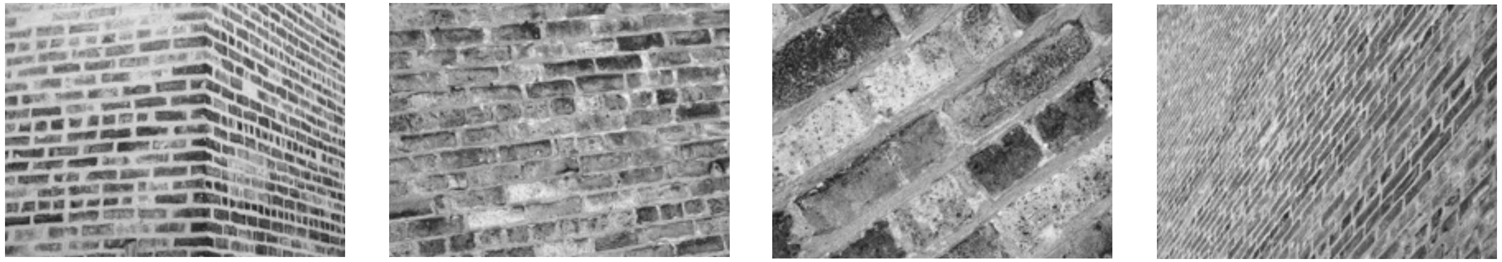
Rotation and Scaling Variability

- Rotation and deformations



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

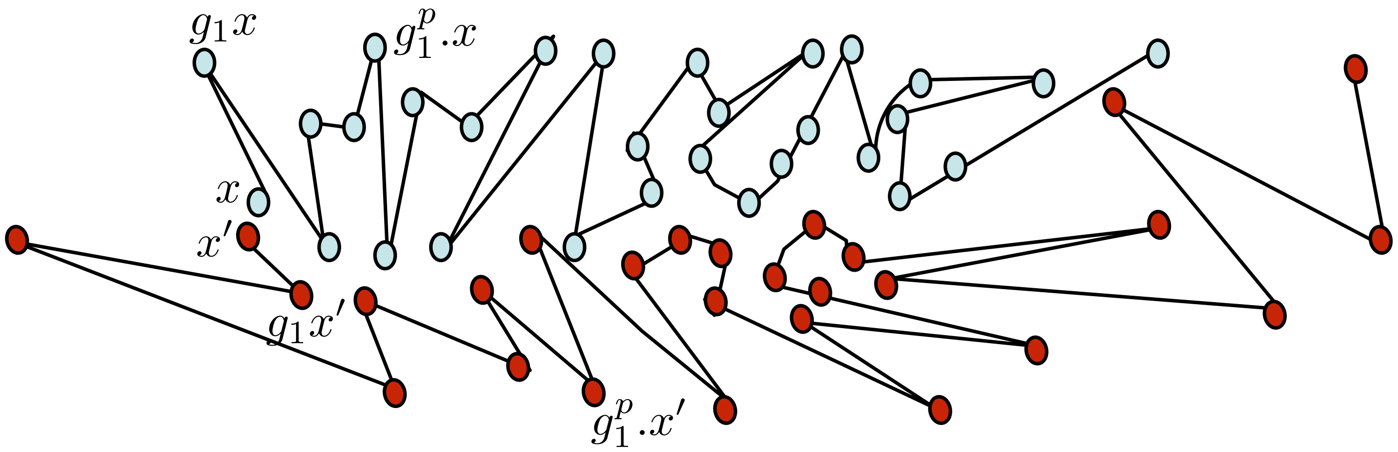
- Scaling and deformations



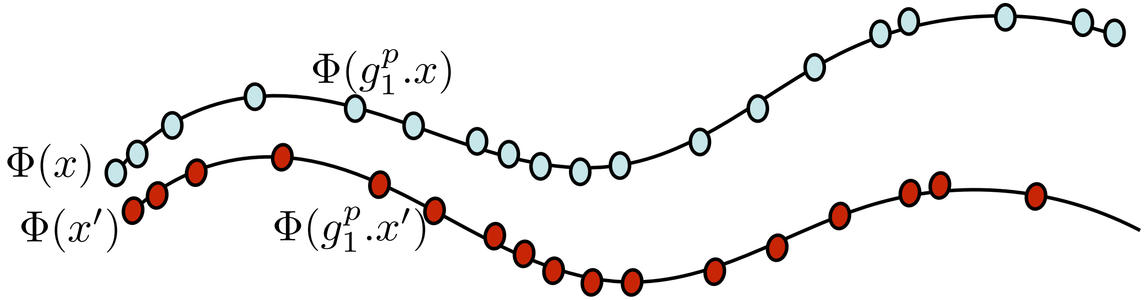
Group: $\mathbb{R} \times \mathrm{Diff}(\mathbb{R})$

Linearize Symmetries

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



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

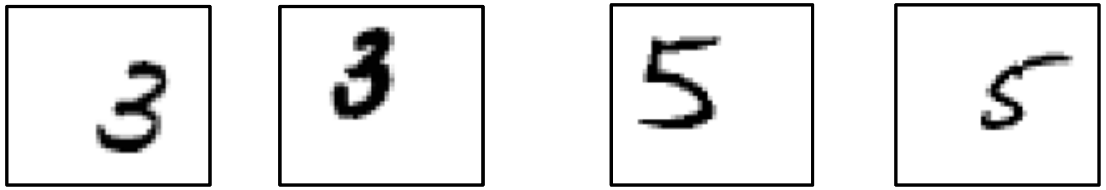


- Lipschitz: $\forall x, g \;:\; \|\Phi(x) - \Phi(g.x)\| \leq C\,\|g\|$

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

https://www.youtube.com/watch?v=nUDIoN-_Hxs

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

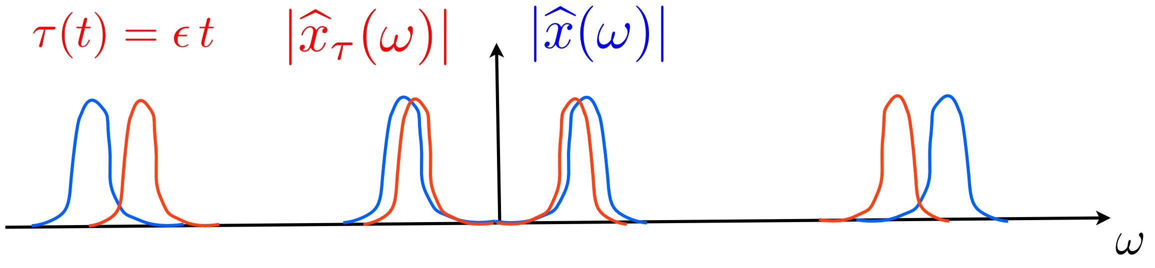
$$x_c(t) = x(t-c) \quad \Rightarrow \quad \hat{x}_c(\omega) = e^{-ic\omega}\, \hat{x}(\omega)$$

The modulus is invariant to translations:

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

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

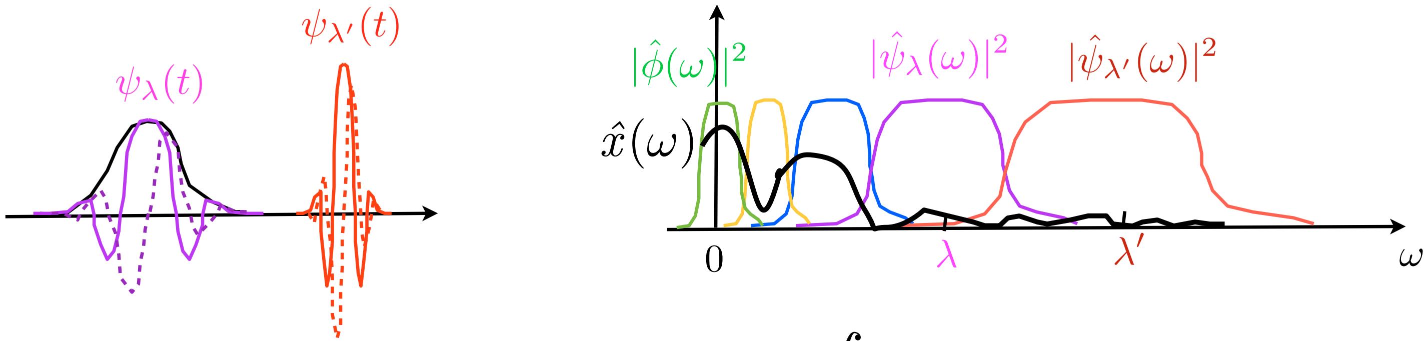
$\big|\, |\hat{x}_\tau(\omega)| - |\hat{x}(\omega)| \,\big|$ is big at high frequencies



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

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

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

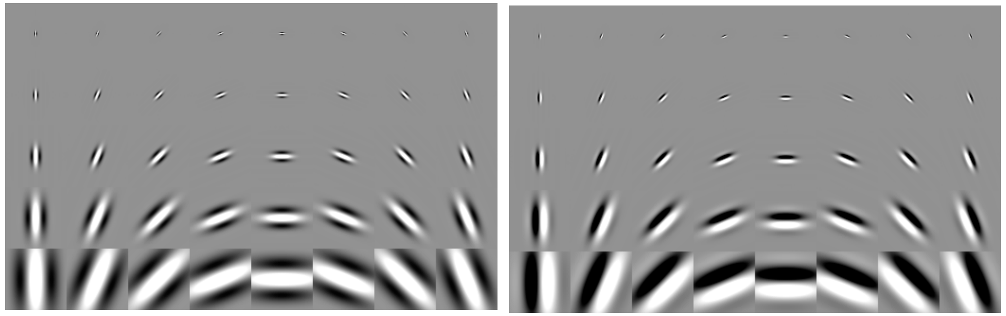
Unitary: $\|Wx\|^2 = \|x\|^2$.

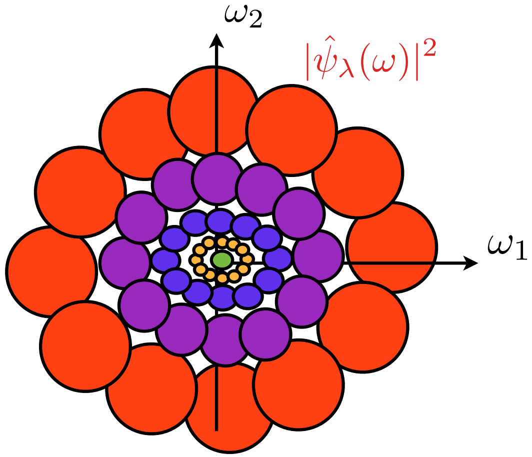
Image Wavelet Transform

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

real parts imaginary parts





- 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 (complex band limited) are uniformly **stable to deformations**

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

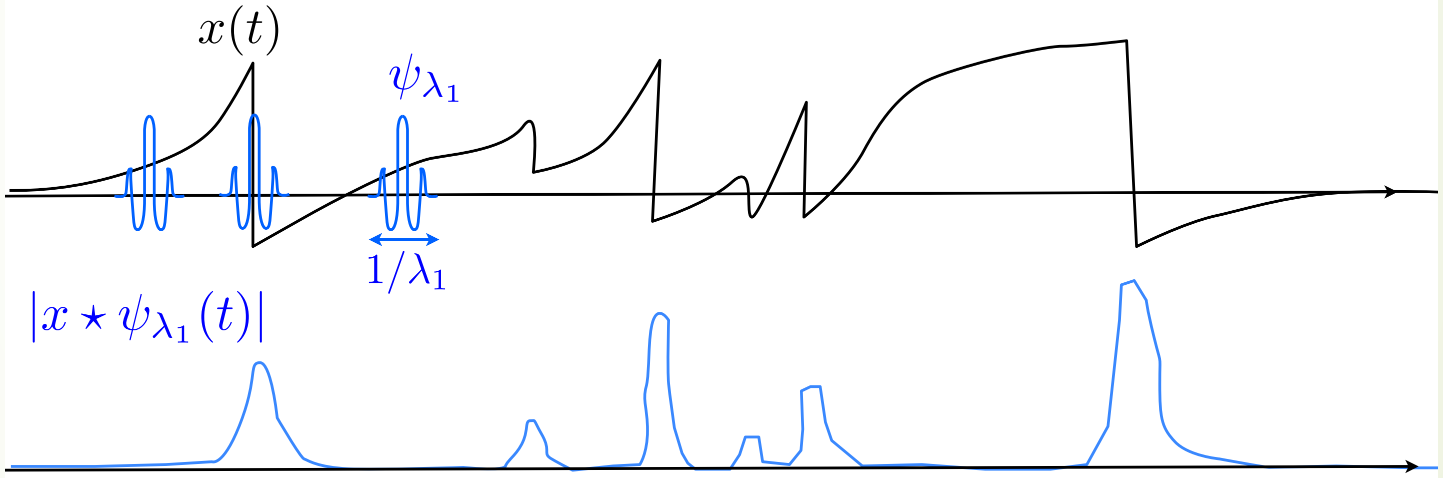
$$\|\psi_\lambda - \psi_{\lambda,\tau}\| \leq C \sup_t |\nabla\tau(t)| \;.$$

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

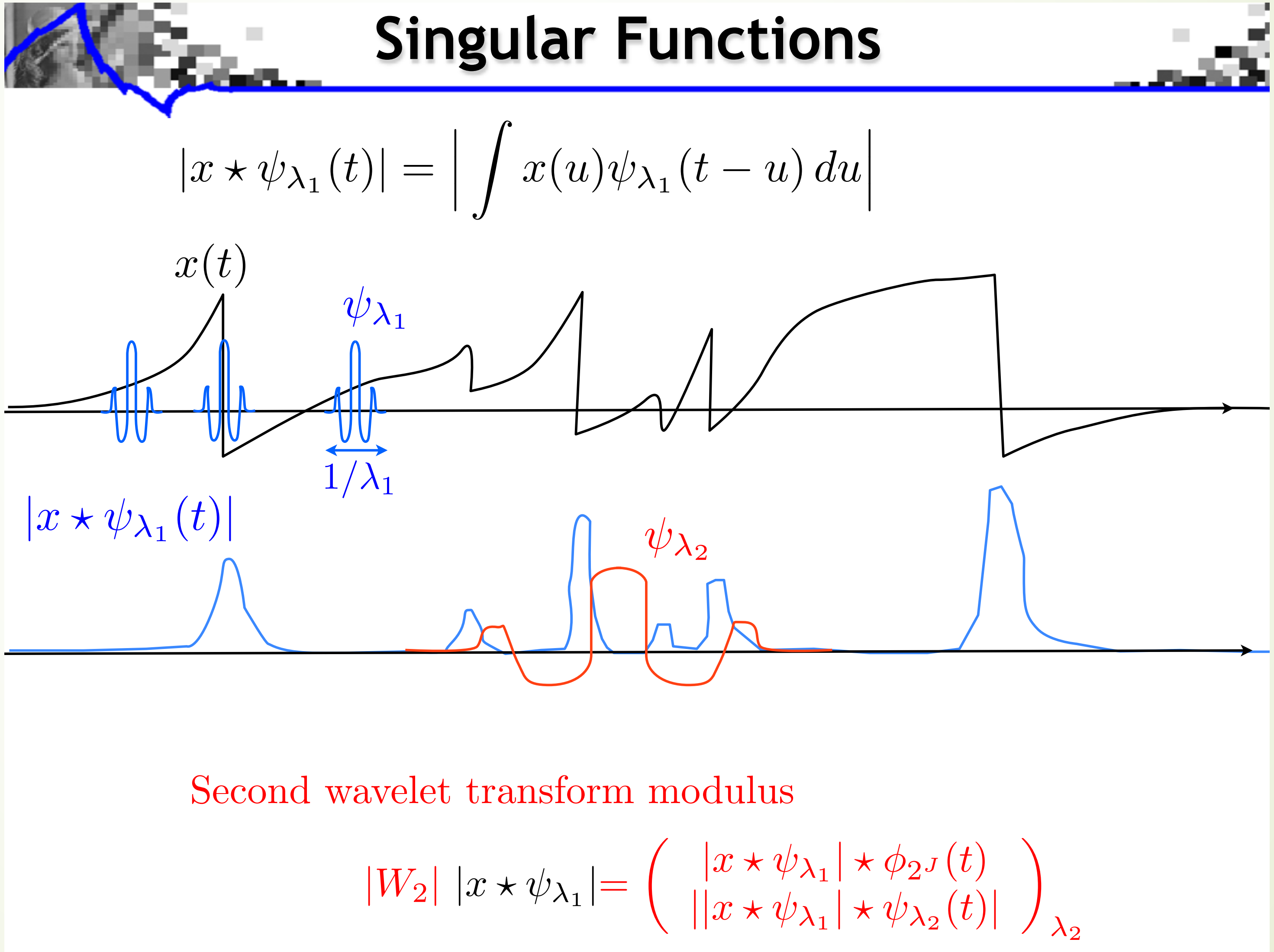
Sparsity of Wavelet Transforms

Singular Functions

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

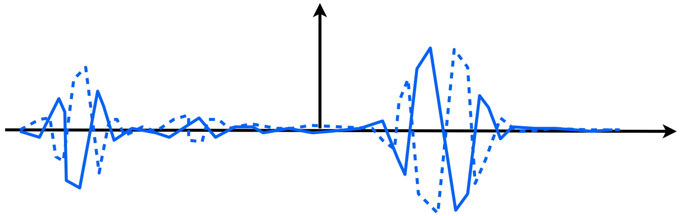


Singularity is preserved in multiscale transform



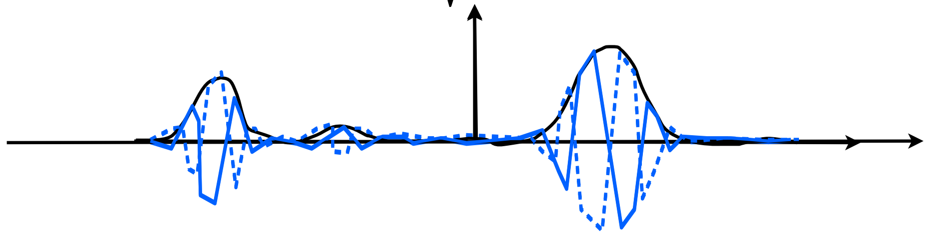
Wavelet Translation Invariance

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



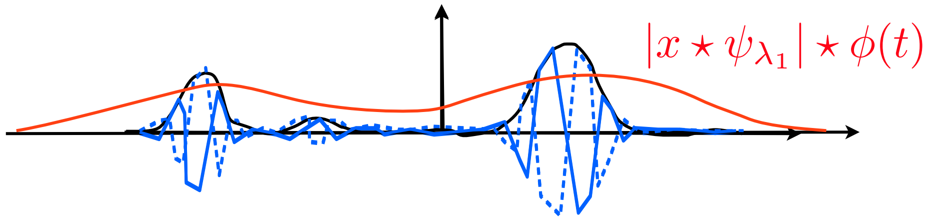
Wavelet Translation Invariance

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



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

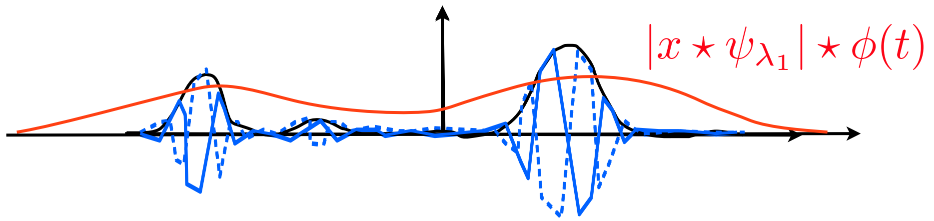
Wavelet Translation Invariance



- 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

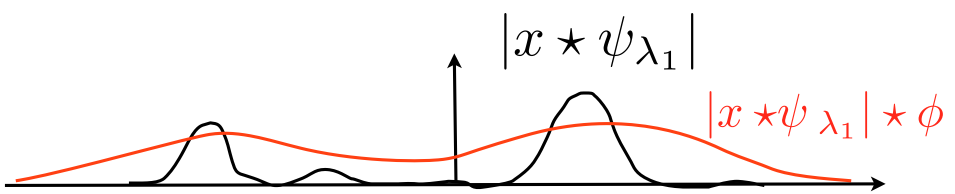


- The modulus $|x \star \psi_{\lambda_1}|$ is a regular envelop
- The average $|x \star \psi_{\lambda_1}| \star \phi(t)$ is invariant to small translations relatively to the support of ϕ.
- Full translation invariance at the limit:

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

but few invariants.

Recovering Lost Information



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

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

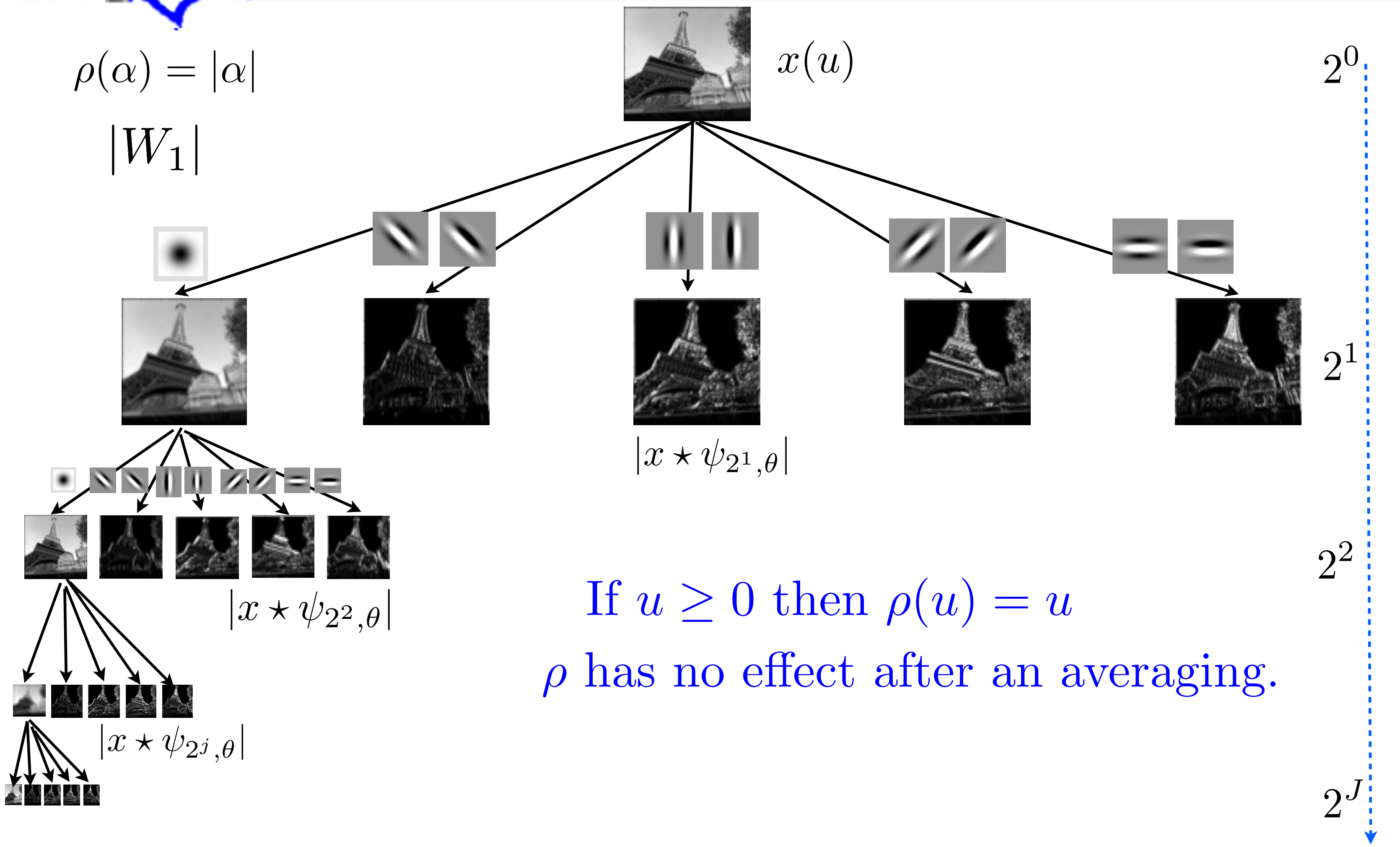
- Translation invariance by time averaging the amplitude:

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

Wavelet Filter Bank

$\rho(\alpha) = |\alpha|$

$|W_1|$



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

ρ has no effect after an averaging.

- Sparse representation

Contraction

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

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

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

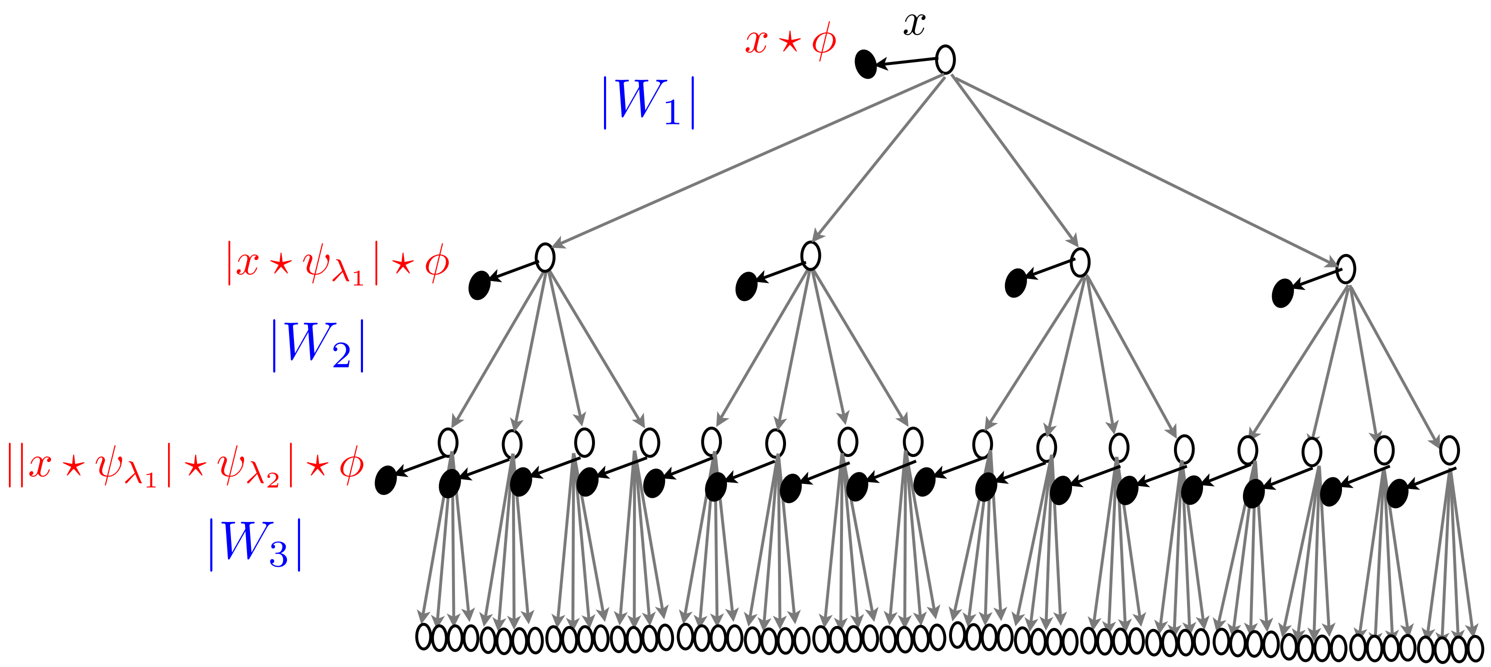
- it is contractive $\||W|x - |W|y\| \leq \|x - y\|$

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

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

Stability of Wavelet Scattering Transform

Scattering Properties

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

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 $\Phi x = Sx$

Summary: Wavelet Scattering Net

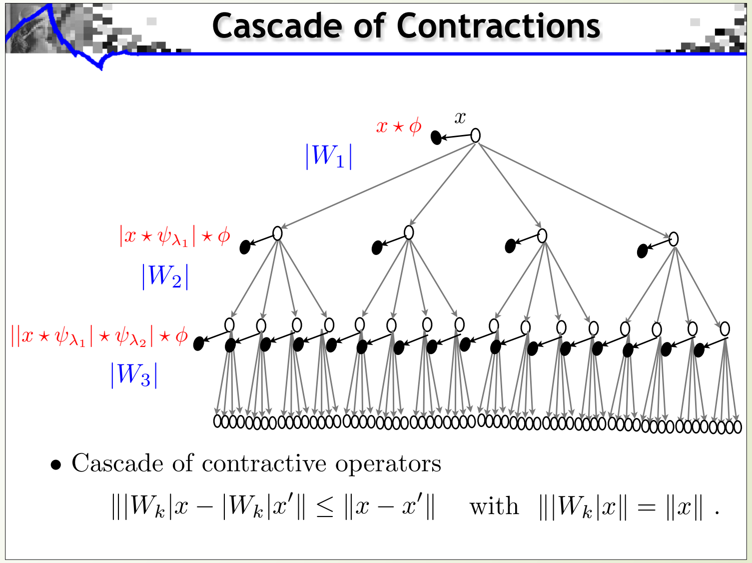
- Architechture:
  - Convolutional filters: band-limited wavelets
  - Nonlinear activation: modulus (Lipschitz)
  - Pooling: L1 norm as averaging
- Properties:
  - A Multiscale Sparse Representation
  - Norm Preservation (Parseval's identity):

$$\|Sx\| = \|x\|$$

  - Contraction:

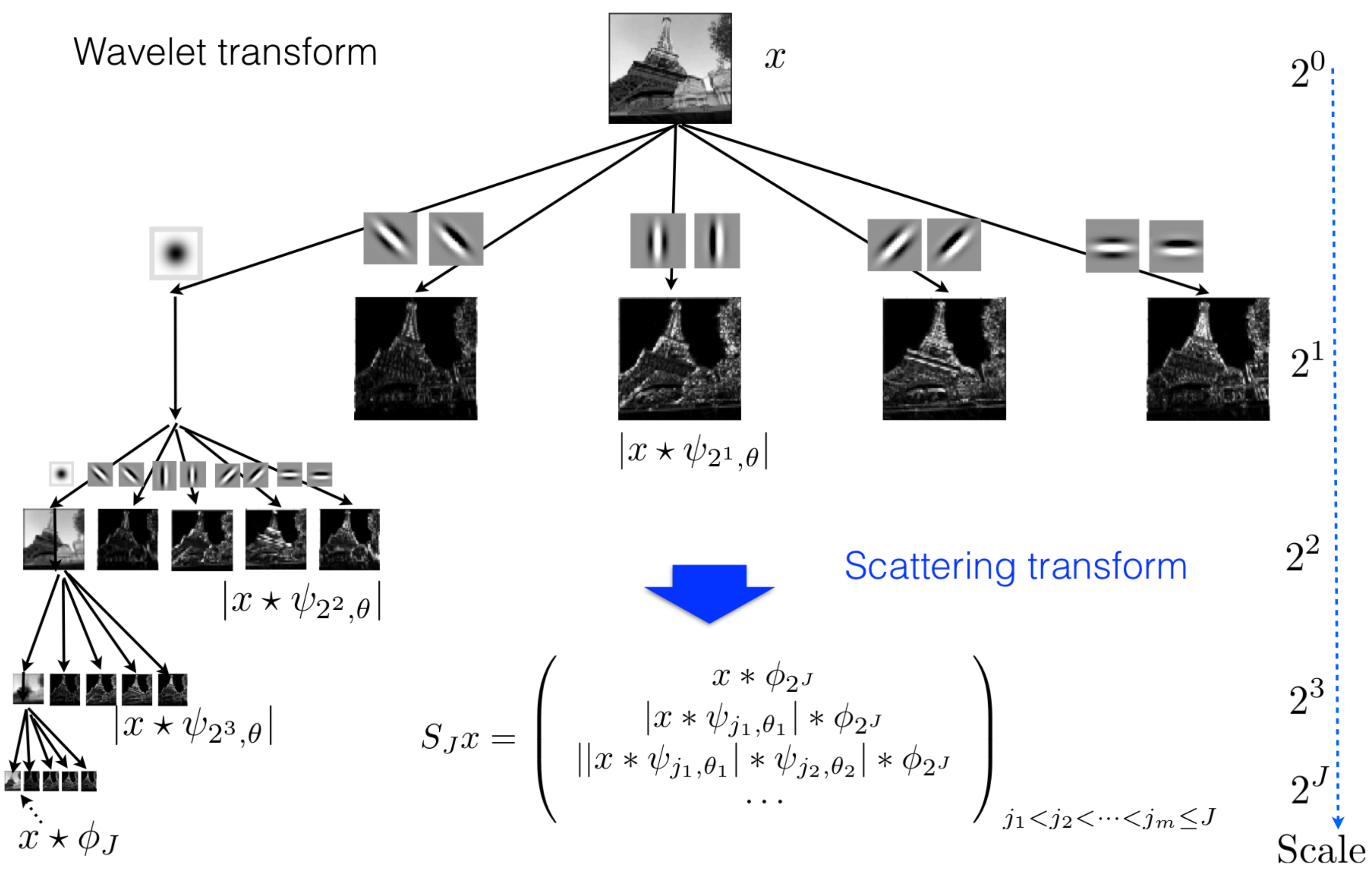
$$\|Sx - Sy\| \leq \|x - y\|$$

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



Scattering Networks

[Mallat '12]



$$S_J x = \begin{pmatrix} x * \phi_{2^J} \\ |x * \psi_{j_1,\theta_1}| * \phi_{2^J} \\ ||x * \psi_{j_1,\theta_1}| * \psi_{j_2,\theta_2}| * \phi_{2^J} \\ \cdots \end{pmatrix}_{j_1 < j_2 < \cdots < j_m \leq J}$$

Scattering Networks

[Mallat '12]

Stability of scattering representations

- Non-expansive mapping

$$\|S_J x - S_J y\| \leq \|x - y\|$$

- Deformation insensitivity

$$D_\tau x(u) = x(u - \tau(u)), \quad \|S_J D_\tau x - S_J x\| \leq C(\tau, J)\|x\|$$

No fitting,
Thus no overfitting!

Group Invariants/Stability

- Translation Invariance:
  - 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:

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

- Stable Small Deformations:

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

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

Applications and extensions:

- Invertibility/completeness of representation [Waldspurger et al. '12]
- Extension to signals on graphs [Chen et al. '14] [Cheng et al. '16]
- With general family of filters [Bolcskei et al. '15] [Czaja et al. '15]

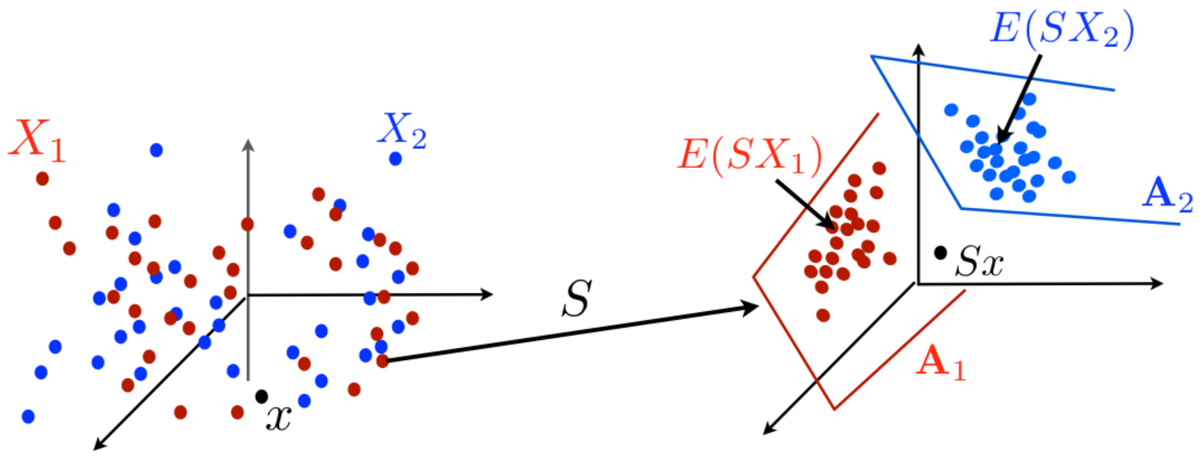
Feature Extraction

Linearized Classification

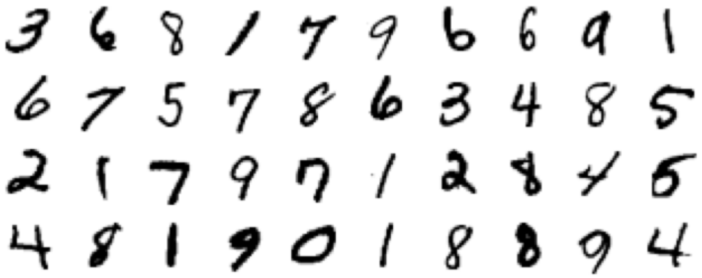
Joan Bruna

- Each class X_k is represented by a scattering centroid $E(SX_k)$

  Affine space model $\mathbf{A}_k = E(SX_k) + \mathbf{V}_k$. computed with PCA.

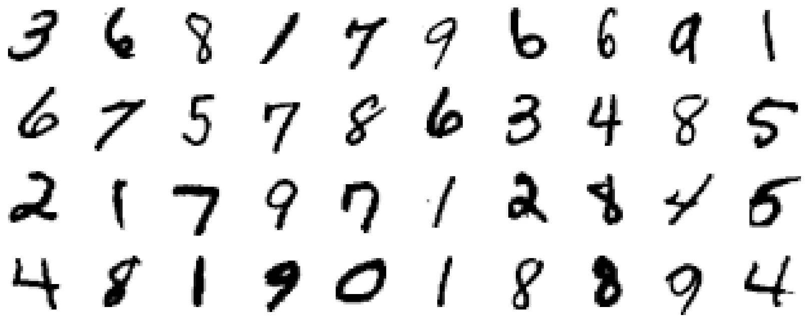


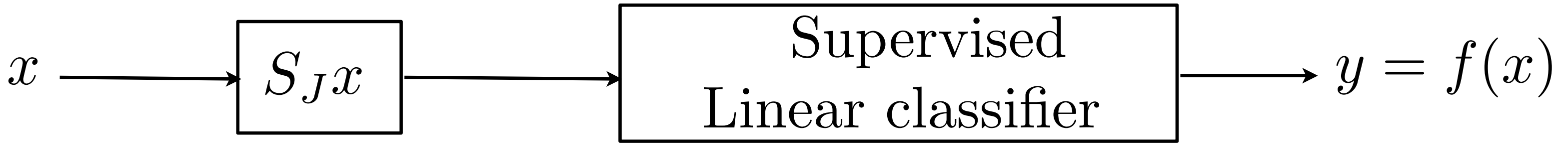
MNIST data basis:



Digit Classification: MNIST

Joan Bruna





Invariants to translations
Linearises small deformations
No learning

Invariants to specific deformations
Separates different patterns

Classification Errors

Training size	Conv. Net.	Scattering
50000	0.4%	0.4%

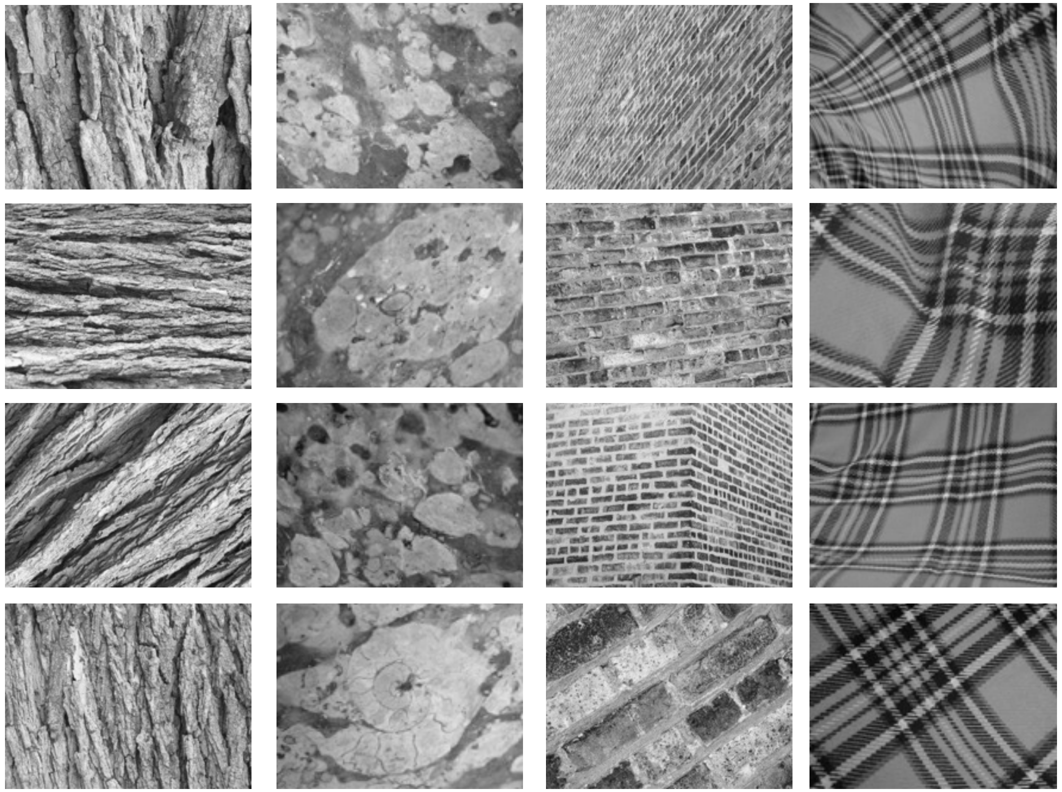
LeCun et. al.

Other Invariants?
Cross-channel pooling!

Rotation and Scaling Invariance

Laurent Sifre

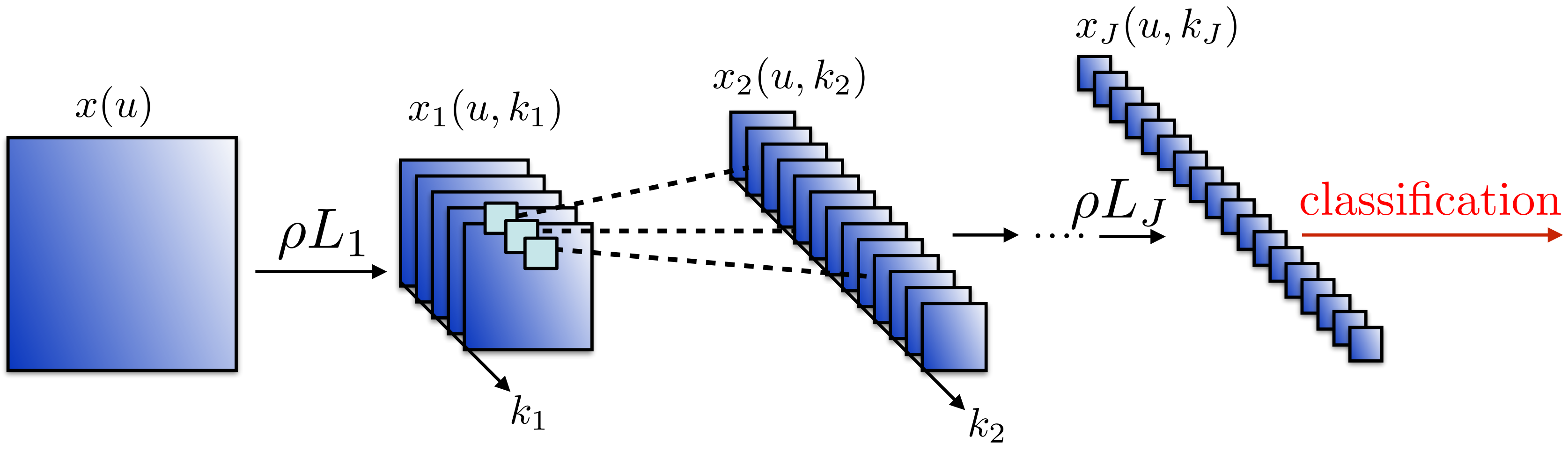
UIUC database:
25 classes



Scattering classification errors

Training	Scat. Translation
20	20 %

Deep Convolutional Trees



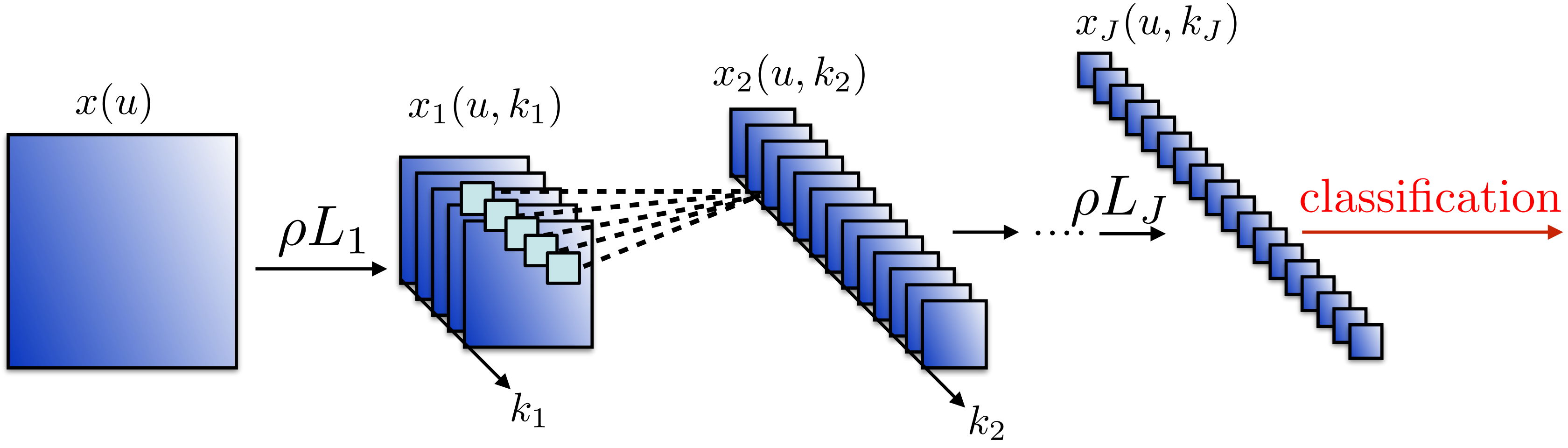
$$x_j = \rho\, L_j\, x_{j-1}$$

L_j is composed of convolutions and subs samplings:

$$x_j(u, k_j) = \rho\Big(x_{j-1}(\cdot, k) \star h_{k_j,k}(u)\Big)$$

No channel communication: what limitations ?

Deep Convolutional Networks



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

- L_j is a linear combination of convolutions and subsampling:

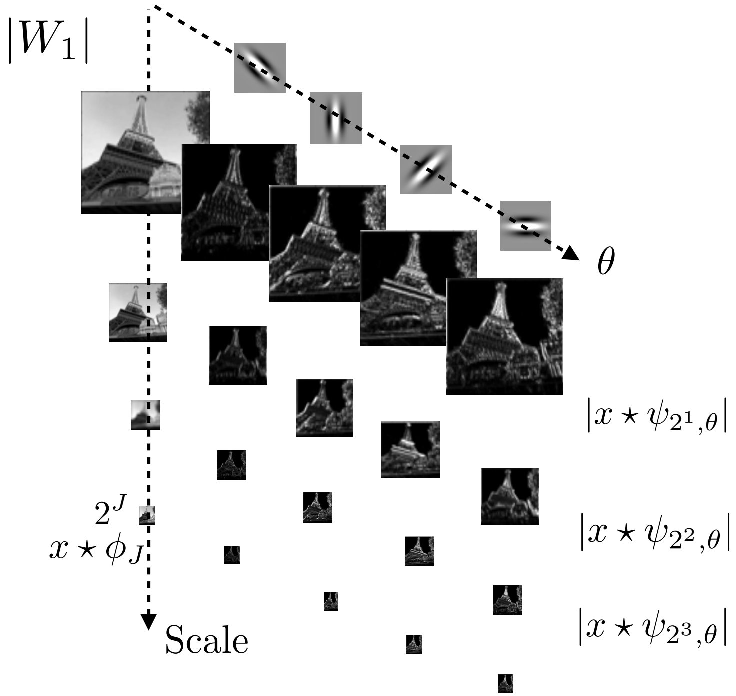
$$x_j(u, k_j) = \rho\Big(\underbrace{\sum_k}_{\text{sum across channels}} x_{j-1}(\cdot, k) \star h_{k_j,k}(u)\Big)$$

What is the role of channel connections ?

Linearize other symmetries beyond translations.

Rotation Invariance

- Channel connections linearize other symmetries.



- Invariance to rotations are computed by convolutions along the rotation variable θ with wavelet filters.

  $\Rightarrow$ invariance to rigid mouvements.

Wavelet Transform on a Group

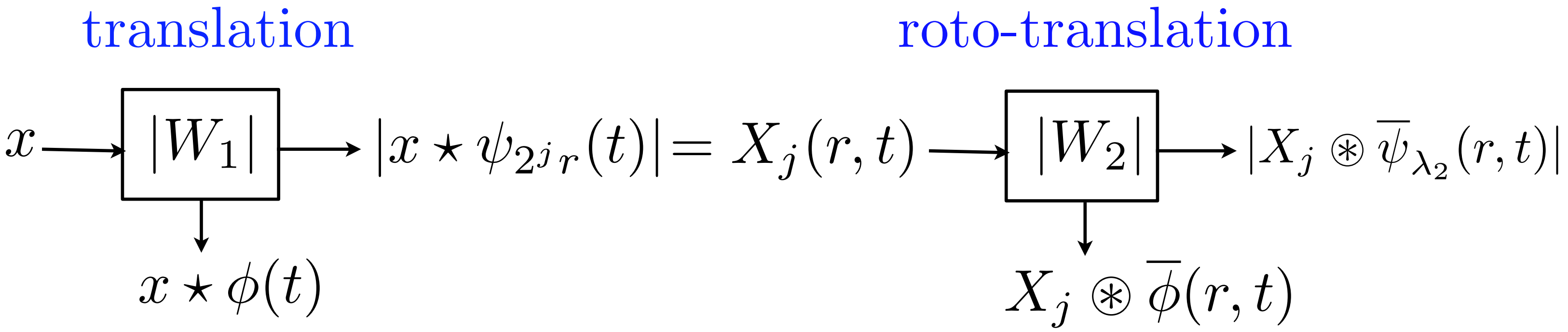
Laurent Sifre

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

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

- Averaging on G: $\quad X \circledast \overline{\phi}(g) = \int_G X(g') \, \overline{\phi}(g'^{-1} g) \, dg'$

- Wavelet transform on G: $\quad W_2 X = \begin{pmatrix} X \circledast \overline{\phi}(g) \\ X \circledast \overline{\psi}_{\lambda_2}(g) \end{pmatrix}_{\lambda_2, g}$.



Wavelet Transform on a Group

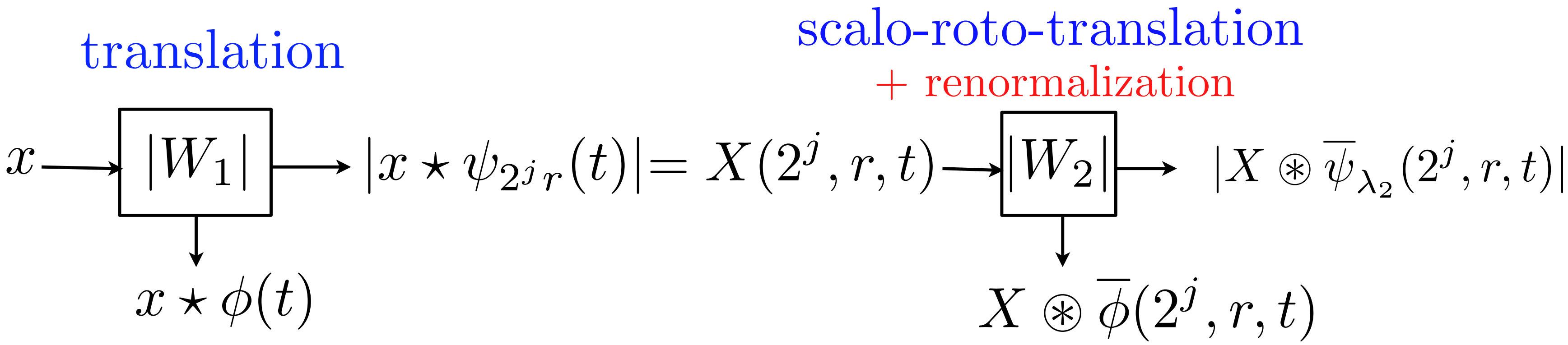
Laurent Sifre

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

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

- Averaging on G: $\quad X \circledast \overline{\phi}(g) = \int_G X(g')\, \overline{\phi}(g'^{-1} g)\, dg'$

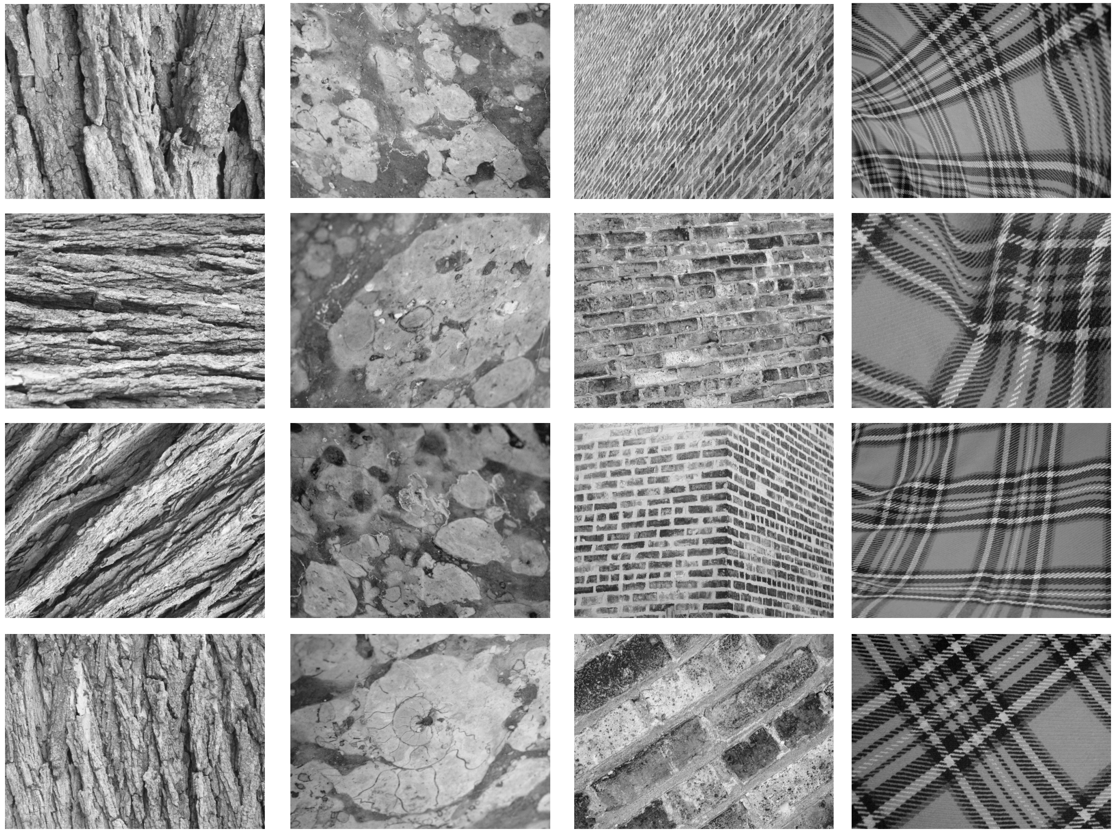
- Wavelet transform on G: $\quad W_2 X = \begin{pmatrix} X \circledast \overline{\phi}(g) \\ X \circledast \overline{\psi}_{\lambda_2}(g) \end{pmatrix}_{\lambda_2, g}$.



Rotation and Scaling Invariance

Laurent Sifre

UIUC database:
25 classes



Scattering classification errors

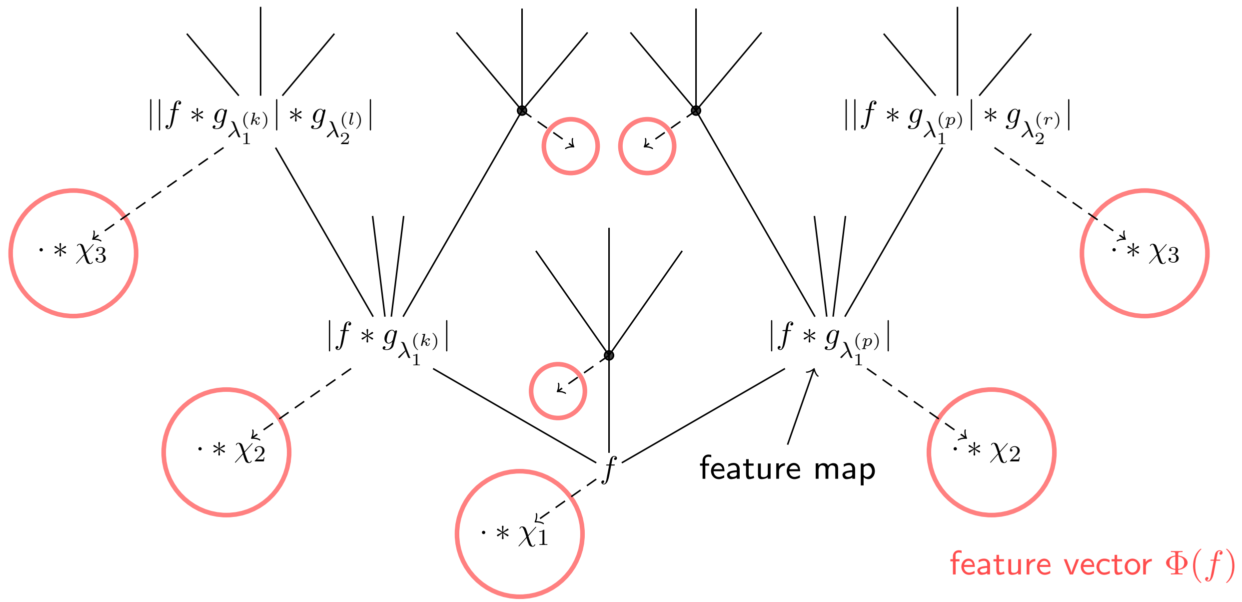
Training	Translation	Transl + Rotation	+ Scaling
20	20 %	2%	**0.6%**

Wiatowski-Bolcskei'15

- Scattering Net by Mallat et al. so far
  - Wavelet Linear filter
  - Nonlinear activation by modulus
  - Average pooling
- Generalization by Wiatowski-Bolcskei'15
  - Filters as frames
  - Lipschitz continuous Nonlinearities
  - General Pooling: Max/Average/Nonlinear, etc.

Generalization of Wiatowski-Bolcskei'15

Scattering networks ([*Mallat, 2012*], [*Wiatowski and HB, 2015*])



General scattering networks guarantee [*Wiatowski & HB, 2015*]

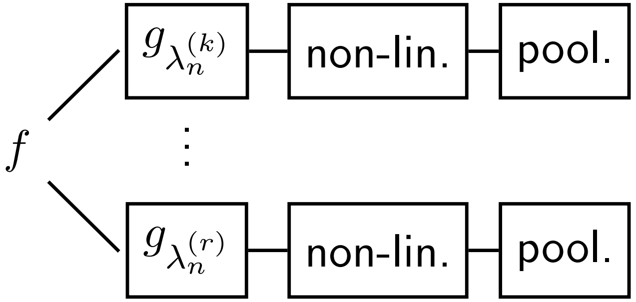
- (vertical) **translation invariance**
- **small deformation sensitivity**

essentially irrespective of filters, non-linearities, and poolings!

Wavelet basis -> filter frame

Building blocks

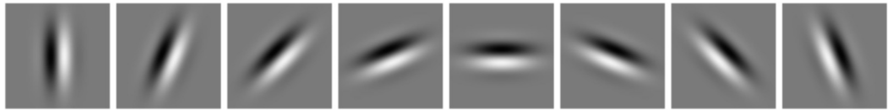
Basic operations in the n-th network layer



Filters: Semi-discrete frame $\Psi_n := \{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$

$$A_n \|f\|_2^2 \leq \|f * \chi_n\|_2^2 + \sum_{\lambda_n \in \Lambda_n} \|f * g_{\lambda_n}\|^2 \leq B_n \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$

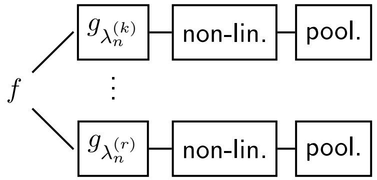
e.g.: Structured filters



Frames: random or learned filters

Building blocks

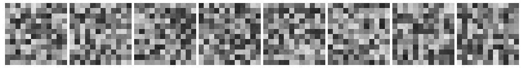
Basic operations in the n-th network layer



Filters: Semi-discrete frame $\Psi_n := \{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$

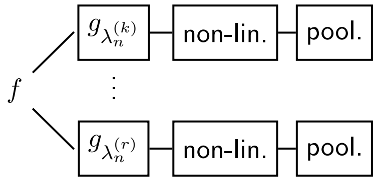
$$A_n \|f\|_2^2 \leq \|f * \chi_n\|_2^2 + \sum_{\lambda_n \in \Lambda_n} \|f * g_{\lambda_n}\|^2 \leq B_n \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$

e.g.: Unstructured filters



Building blocks

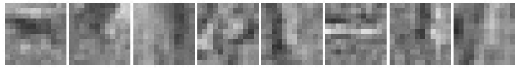
Basic operations in the n-th network layer



Filters: Semi-discrete frame $\Psi_n := \{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$

$$A_n \|f\|_2^2 \leq \|f * \chi_n\|_2^2 + \sum_{\lambda_n \in \Lambda_n} \|f * g_{\lambda_n}\|^2 \leq B_n \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$

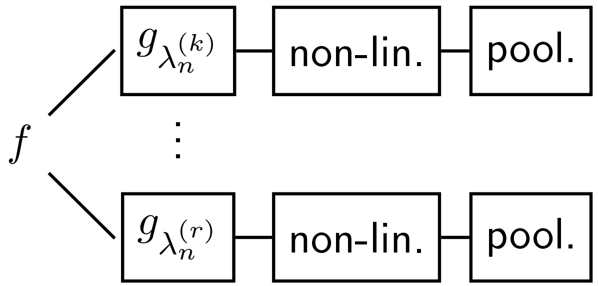
e.g.: Learned filters



Nonlinear activations

Building blocks

Basic operations in the n-th network layer



Non-linearities: Point-wise and Lipschitz-continuous

$$\|M_n(f) - M_n(h)\|_2 \le L_n \|f - h\|_2, \quad \forall\, f, h \in L^2(\mathbb{R}^d)$$

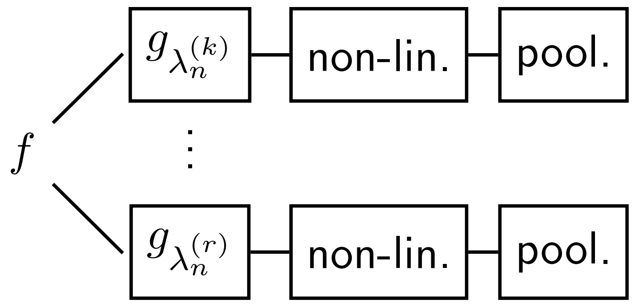
$\Rightarrow$ Satisfied by virtually **all** non-linearities used in the **deep learning literature**!

ReLU: $L_n = 1$; modulus: $L_n = 1$; logistic sigmoid: $L_n = \frac{1}{4}$; ...

Pooling

Building blocks

Basic operations in the n-th network layer



Pooling: In continuous-time according to

$$f \mapsto S_n^{d/2} P_n(f)(S_n \cdot),$$

where $S_n \geq 1$ is the **pooling factor** and $P_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ is R_n-Lipschitz-continuous

$\Rightarrow$ **Emulates** most **poolings** used in the **deep learning literature**!

e.g.: Pooling by **sub-sampling** $P_n(f) = f$ with $R_n = 1$

e.g.: Pooling by **averaging** $P_n(f) = f * \phi_n$ with $R_n = \|\phi_n\|_1$

Vertical translation invariance

Theorem (Wiatowski and HB, 2015)

Assume that the filters, non-linearities, and poolings satisfy

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall\, n \in \mathbb{N}.$$

Let the pooling factors be $S_n \geq 1$, $n \in \mathbb{N}$. *Then,*

$$|||\Phi^n(T_t f) - \Phi^n(f)||| = \mathcal{O}\left(\frac{\|t\|}{S_1 \dots S_n}\right),$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, $n \in \mathbb{N}$.

The condition

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall\, n \in \mathbb{N},$$

is **easily satisfied** by **normalizing** the filters $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$.

Vertical translation invariance

Theorem (Wiatowski and HB, 2015)

Assume that the filters, non-linearities, and poolings satisfy

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall\, n \in \mathbb{N}.$$

Let the pooling factors be $S_n \geq 1$, $n \in \mathbb{N}$. *Then,*

$$|||\Phi^n(T_t f) - \Phi^n(f)||| = \mathcal{O}\left(\frac{\|t\|}{S_1 \dots S_n}\right),$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, $n \in \mathbb{N}$.

$\Rightarrow$ Features become **more invariant** with **increasing** network **depth**!

Vertical translation invariance

Theorem (Wiatowski and HB, 2015)

Assume that the filters, non-linearities, and poolings satisfy

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall\, n \in \mathbb{N}.$$

Let the pooling factors be $S_n \geq 1$, $n \in \mathbb{N}$. *Then,*

$$|||\Phi^n(T_t f) - \Phi^n(f)||| = \mathcal{O}\left(\frac{\|t\|}{S_1 \dots S_n}\right),$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, $n \in \mathbb{N}$.

Full translation invariance: If $\lim_{n\to\infty} S_1 \cdot S_2 \cdot \ldots \cdot S_n = \infty$, then

$$\lim_{n\to\infty} |||\Phi^n(T_t f) - \Phi^n(f)||| = 0$$

Philosophy behind invariance results

Mallat's "horizontal" translation invariance [*Mallat, 2012*]:

$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \qquad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d$$

- features become invariant in every network layer, but needs $J \to \infty$
- applies to wavelet transform and modulus non-linearity without pooling

"Vertical" translation invariance:

$$\lim_{n\to\infty} |||\Phi^n(T_t f) - \Phi^n(f)||| = 0, \qquad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d$$

- features become more invariant with increasing network depth
- applies to general filters, general non-linearities, and general poolings

Non-linear deformations

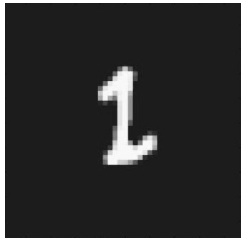
Non-linear deformation $(F_\tau f)(x) = f(x - \tau(x))$, where $\tau : \mathbb{R}^d \to \mathbb{R}^d$

For "small" τ:

Non-linear deformations

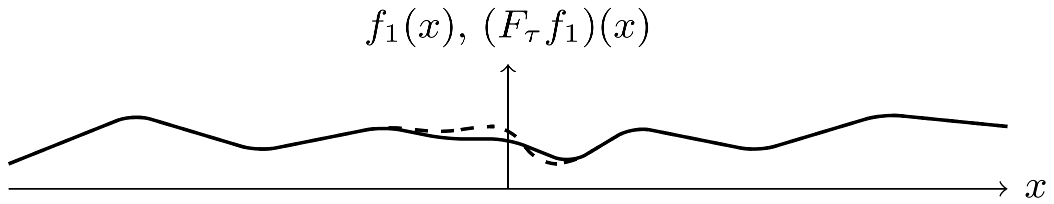
Non-linear deformation $(F_\tau f)(x) = f(x - \tau(x))$, where $\tau : \mathbb{R}^d \to \mathbb{R}^d$

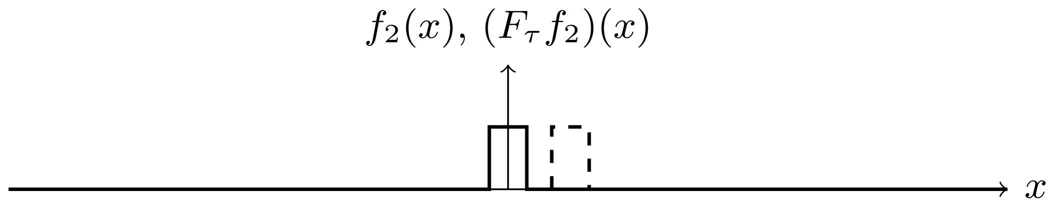
For "large" τ:



Deformation sensitivity for signal classes

Consider $(F_\tau f)(x) = f(x - \tau(x)) = f(x - e^{-x^2})$





For given τ the amount of deformation induced can depend drastically on $f \in L^2(\mathbb{R}^d)$

Wiatowski-Bolcskei'15 Deformation Stability Bounds

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound [*Mallat, 2012*]:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- The signal class H_W and the corresponding norm $\|\cdot\|_W$ depend on the mother wavelet (and hence the network)

Our deformation sensitivity bound:

$$|||\Phi(F_\tau f) - \Phi(f)||| \leq C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- The signal class $\mathcal{C}$ (band-limited functions, cartoon functions, or Lipschitz functions) is independent of the network

Wiatowski-Bolcskei'15 Deformation Stability Bounds

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound [*Mallat, 2012*]:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \le C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- Signal class description complexity implicit via norm $\|\cdot\|_W$

Our deformation sensitivity bound:

$$|||\Phi(F_\tau f) - \Phi(f)||| \le C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- Signal class description complexity explicit via $C_{\mathcal{C}}$
  - L-band-limited functions: $C_{\mathcal{C}} = \mathcal{O}(L)$
  - cartoon functions of size K: $C_{\mathcal{C}} = \mathcal{O}(K^{3/2})$
  - M-Lipschitz functions $C_{\mathcal{C}} = \mathcal{O}(M)$

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound [*Mallat, 2012*]:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- The bound depends explicitly on higher order derivatives of τ

Our deformation sensitivity bound:

$$|||\Phi(F_\tau f) - \Phi(f)||| \leq C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- The bound implicitly depends on derivative of τ via the condition $\|D\tau\|_\infty \leq \frac{1}{2d}$

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound [*Mallat, 2012*]:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \le C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- The bound is *coupled* to horizontal translation invariance

$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d$$

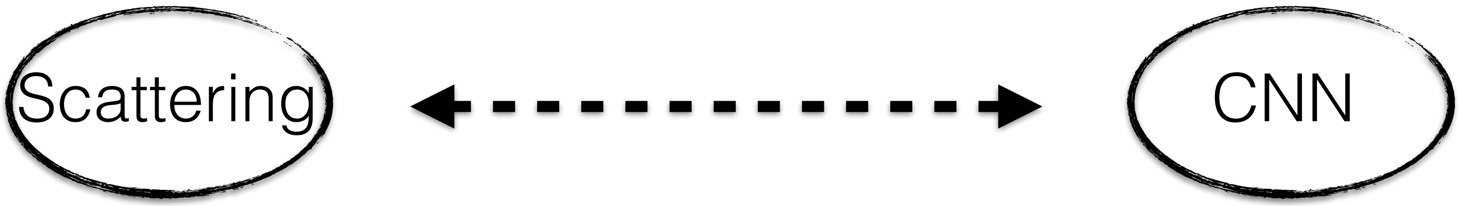
Our deformation sensitivity bound:

$$|||\Phi(F_\tau f) - \Phi(f)||| \le C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- The bound is *decoupled* from vertical translation invariance

$$\lim_{n\to\infty} |||\Phi^n(T_t f) - \Phi^n(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d$$

What is in between?



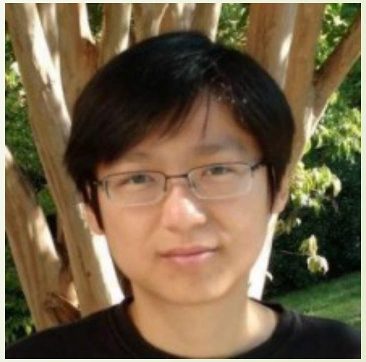
- No training until the classifier
- No parameters in the convolutional layers
- Most “control” of regularity and robustness
- Strong performance and explainable features

- Fully trained by large volume of data
- Lots of parameters (largest model capacity)
- Least “control” of regularity and robustness
- Best performance but not explainable

Decomposed Convolutional Filters (DCF)

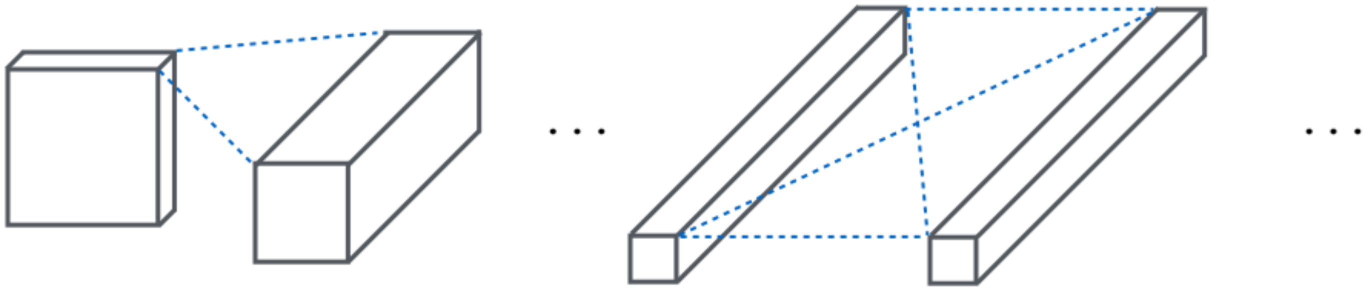
Xiuyuan Cheng et al.

https://arxiv.org/abs/1802.04145



Decomposition of Convolutional Filters

$$x^{(0)} \mapsto x^{(1)} \mapsto \cdots \mapsto x^{(l-1)} \mapsto x^{(l)} \mapsto \cdots$$



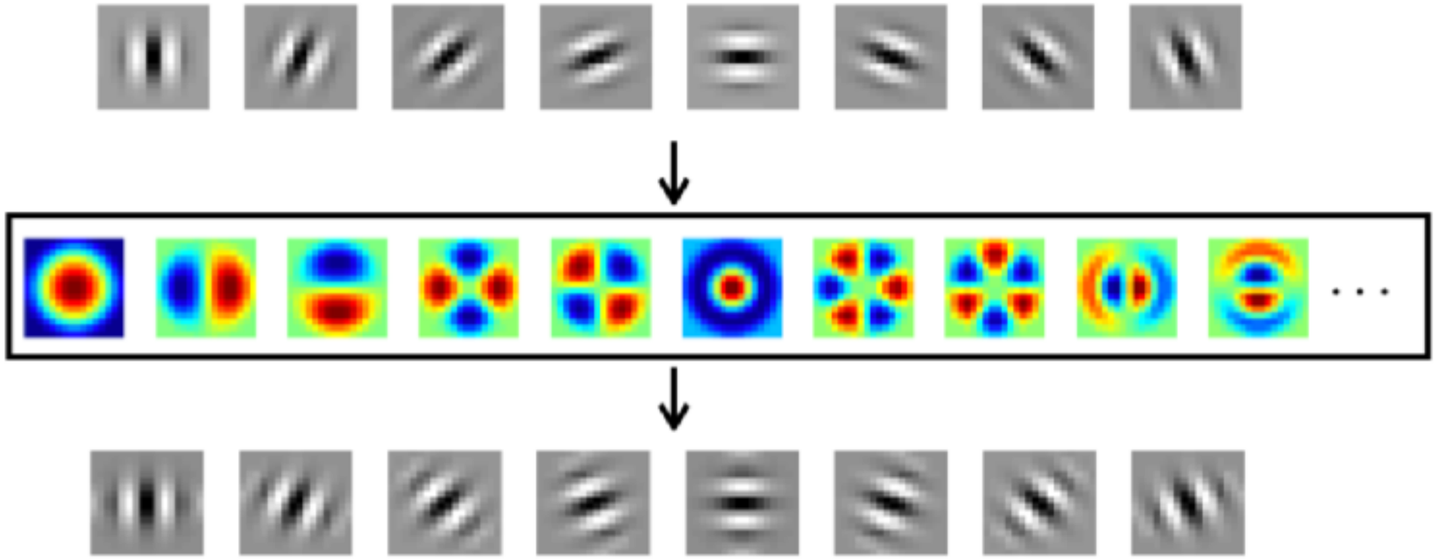
The mapping in a convolutional layer

$$x^{(l)}(u,\lambda) = \sigma\left(\sum_{\lambda'} \int W^{(l)}_{\lambda',\lambda}(v')x^{(l-1)}(u+v',\lambda')dv' + b^{(l)}(\lambda)\right)$$

Decomposition of Convolutional Filters

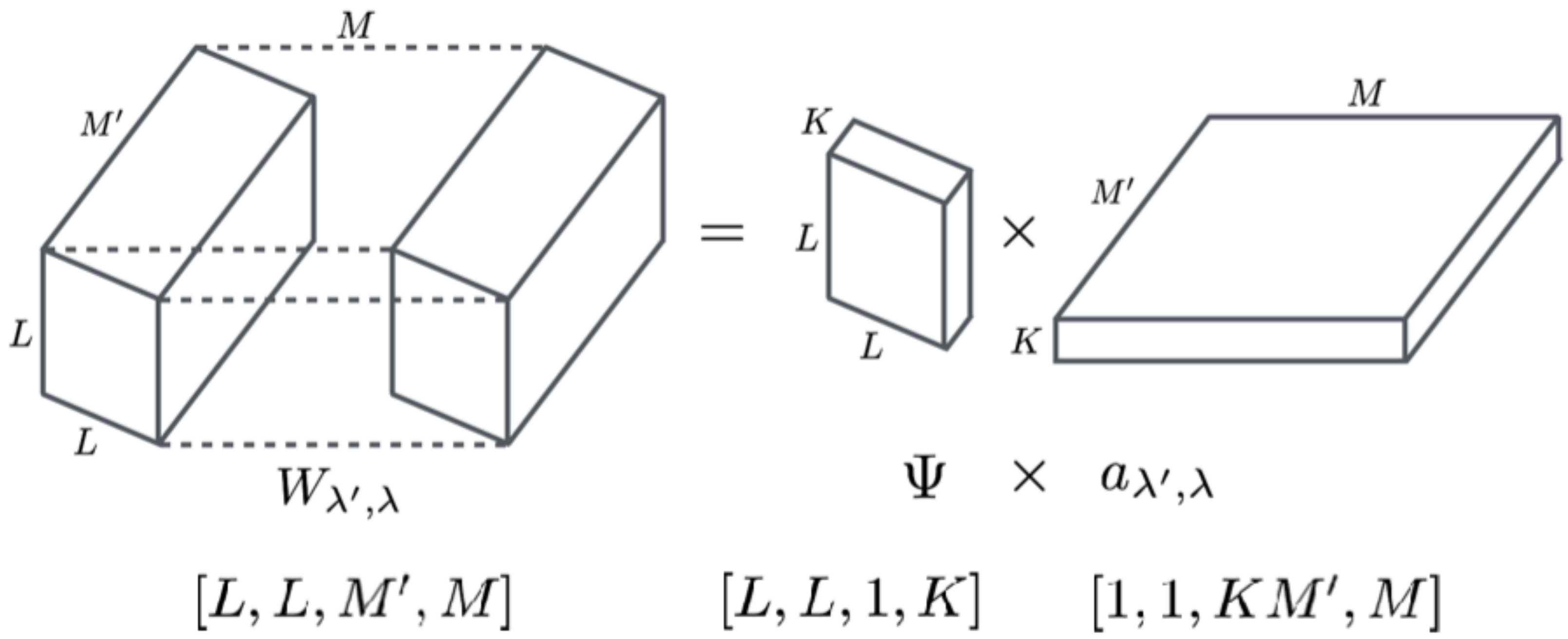
Introducing bases ψ_k

$$W_{\lambda',\lambda}(u) = \sum_{k=1}^{K} (a_{\lambda',\lambda})_k \psi_k(u).$$



Decomposition of Convolutional Filters

- Filters viewed in tensors



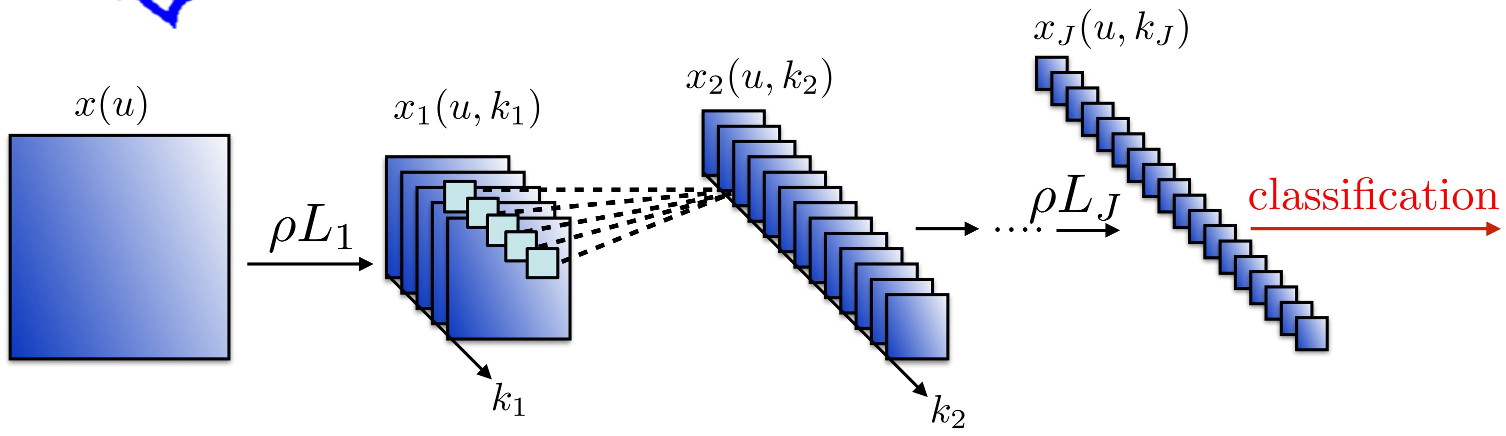
- Psi prefixed, a trained from data

Reduction in the Number of Parameters

- Number of parameters
  - Regular conv layer: $L \times L \times M' \times M$
  - DCF layer: $K \times M' \times M$
- Forward-pass computation
  - Regular conv layer: $M'W^2 \cdot M(1 + 2L^2)$
  - DCF layer: $M'W^2 \cdot 2K(L^2 + M)$

A factor of $\frac{K}{L^2}$ *!*

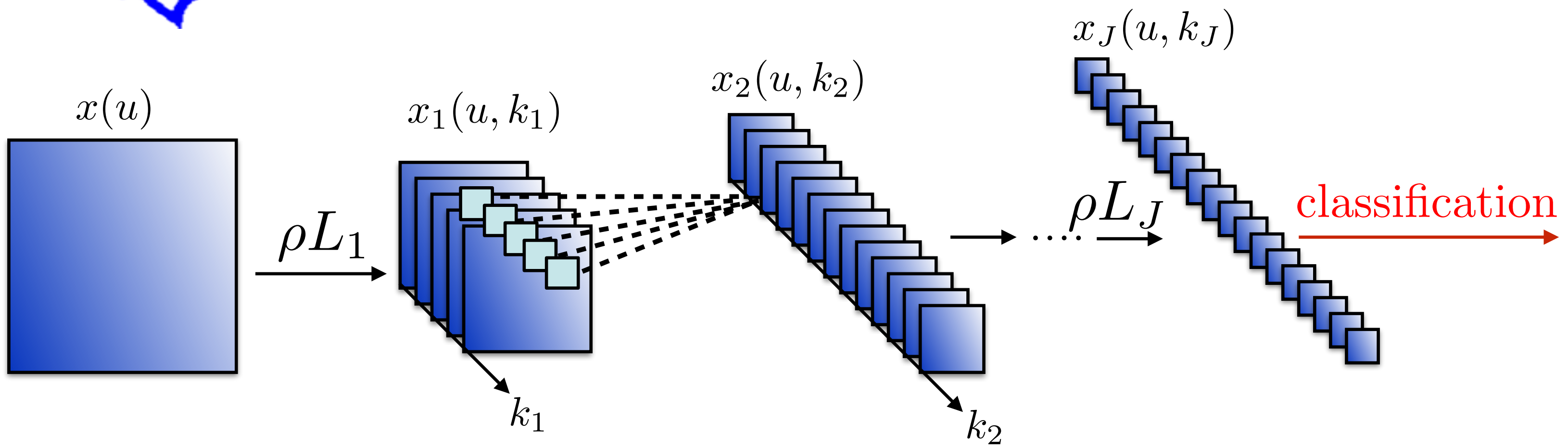
Deep Convolutional Networks



- The convolution network operators L_j have many roles:
  - Linearize non-linear transformations (symmetries)
  - Reduce dimension with projections
  - Memory storage of « characteristic » structures

- Difficult to separate these roles when analyzing learned networks

Open Problems



- Can we recover symmetry groups from the matrices *Lj* ?
- What kind of groups ?
- Can we characterise the regularity of *f(x)* from these groups ?
- Can we define classes of high-dimensional « regular » functions that are well approximated by deep neural networks ?
- Can we get approximation theorems giving errors depending on number of training exemples, with a fast decay ?

Group Invariant and Equivariant Networks

Cohen, Welling, https://arxiv.org/abs/1602.07576

Sannai, Takai, Cordonnier, https://arxiv.org/abs/1903.01939v2

Definition 2.1. Let G be a group and X and Y two sets. We assume that G acts on X (resp. Y) by $g \cdot x$ (resp. $g * y$) for $g \in G$ and $x \in X$ (resp. $y \in Y$) . We say that a map $f\colon X \to Y$ is

- *G-invariant* if $f(g \cdot x) = f(x)$ for any $g \in G$ and any $x \in X$,
- *G-equivariant* if $f(g \cdot x) = g * f(x)$ for any $g \in G$ and any $x \in X$.

Group Convolution Neural Network

[Cohen, Welling, https://arxiv.org/abs/1602.07576]

$$[f * \psi^i](x) = \sum_{y \in \mathbb{Z}^2} \sum_{k=1}^{K^l} f_k(y)\psi_k^i(x - y)$$

$$[f \star \psi](g) = \sum_{h \in G} \sum_{k} f_k(h)\psi_k(g^{-1}h).$$

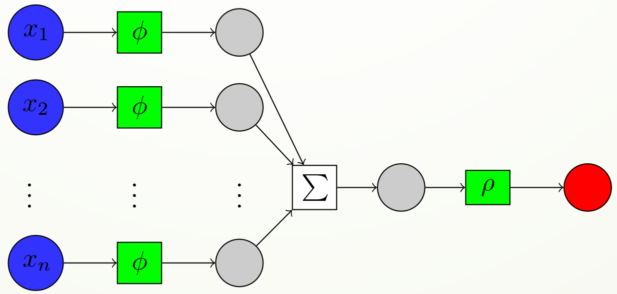
Permutation Invariant Functions

When $G = S_n$ and the actions are induced by permutation, we call G-invariant (resp. G-equivariant) functions as *permutation invariant* (resp. *permutation equivariant*) functions.

Theorem 3.1 ([28] Kolmogorov-Arnold's representation theorem for permutation actions)**.** *Let $K \subset \mathbb{R}^n$ be a compact set. Then, any continuous S_n-invariant function $f\colon K \longmapsto \mathbb{R}$ can be represented as*

$$f(x_1, \dots, x_n) = \rho\left(\sum_{i=1}^{n} \phi(x_i)\right) \quad (1)$$

for some continuous function $\rho\colon \mathbb{R}^{n+1} \to \mathbb{R}$. Here, $\phi\colon \mathbb{R} \to \mathbb{R}^{n+1}; x \mapsto (1, x, x^2, \dots, x^n)^\top$.



Permutation Equivariant Functions

Proposition 4.1. *A map* $F\colon \mathbb{R}^n \to \mathbb{R}^n$ *is* S_n*-equivariant if and only if there is a* $\mathrm{Stab}(1)$*-invariant function* $f\colon \mathbb{R}^n \to \mathbb{R}$ *satisfying* $F = (f, f \circ (1\ 2), \ldots, f \circ (1\ n))^\top$. *Here,* $(1\ i) \in S_n$ *is the transposition between* 1 *and* i.

Corollary 4.1 (Representation of $\mathrm{Stab}(1)$-invariant function)**.** *Let* $K \subset \mathbb{R}^n$ *be a compact set, let* $f\colon K \longrightarrow \mathbb{R}$ *be a continuous and* $\mathrm{Stab}(1)$*-invariant function. Then,* $f(\boldsymbol{x})$ *can be represented as*

$$f(\boldsymbol{x}) = f(x_1, \ldots, x_n) = \rho\left(x_1, \sum_{i=2}^{n} \phi(x_i)\right),$$

for some continuous function $\rho\colon \mathbb{R}^{n+1} \longrightarrow \mathbb{R}$. *Here,* $\phi\colon \mathbb{R} \to \mathbb{R}^n$ *is similar as in Theorem 3.1.*

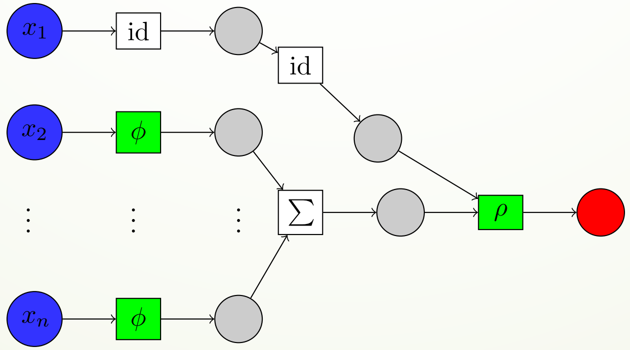


Diagram 3: A neural network approximating the $\mathrm{Stab}(1)$-invariant function f

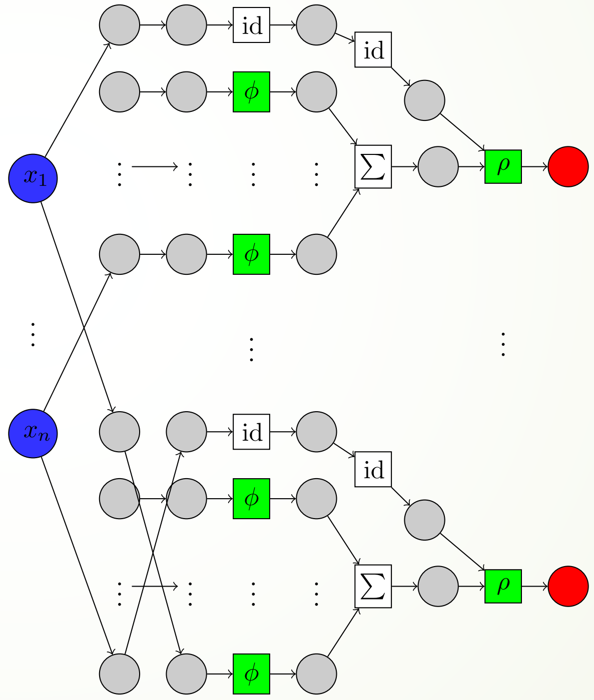


Diagram 2: A neural network approximating S_n-equivariant map F

Thank you!