Addressing the Curse of Dimensionality with Convolutional Neural Networks

Joan Bruna

UC Berkeley —> Courant Institute, NYU

collaborators:

Ivan Dokmanic (UIUC), Stephane Mallat (ENS, France)

Pablo Sprechmann (NYU), Yann LeCun (NYU)

Berkeley
UNIVERSITY OF CALIFORNIA

NYU | COURANT INSTITUTE OF MATHEMATICAL SCIENCES

From Sanja Fisler:

(inspired by R.Urtasun ICLR'16)

Computer Vision: what Neural Networks can do for me.

Machine Learning: what can I do for Neural Networks.

From Sanja Fisler:

(inspired by R.Urtasun ICLR'16)

Computer Vision: what Neural Networks can do for me.

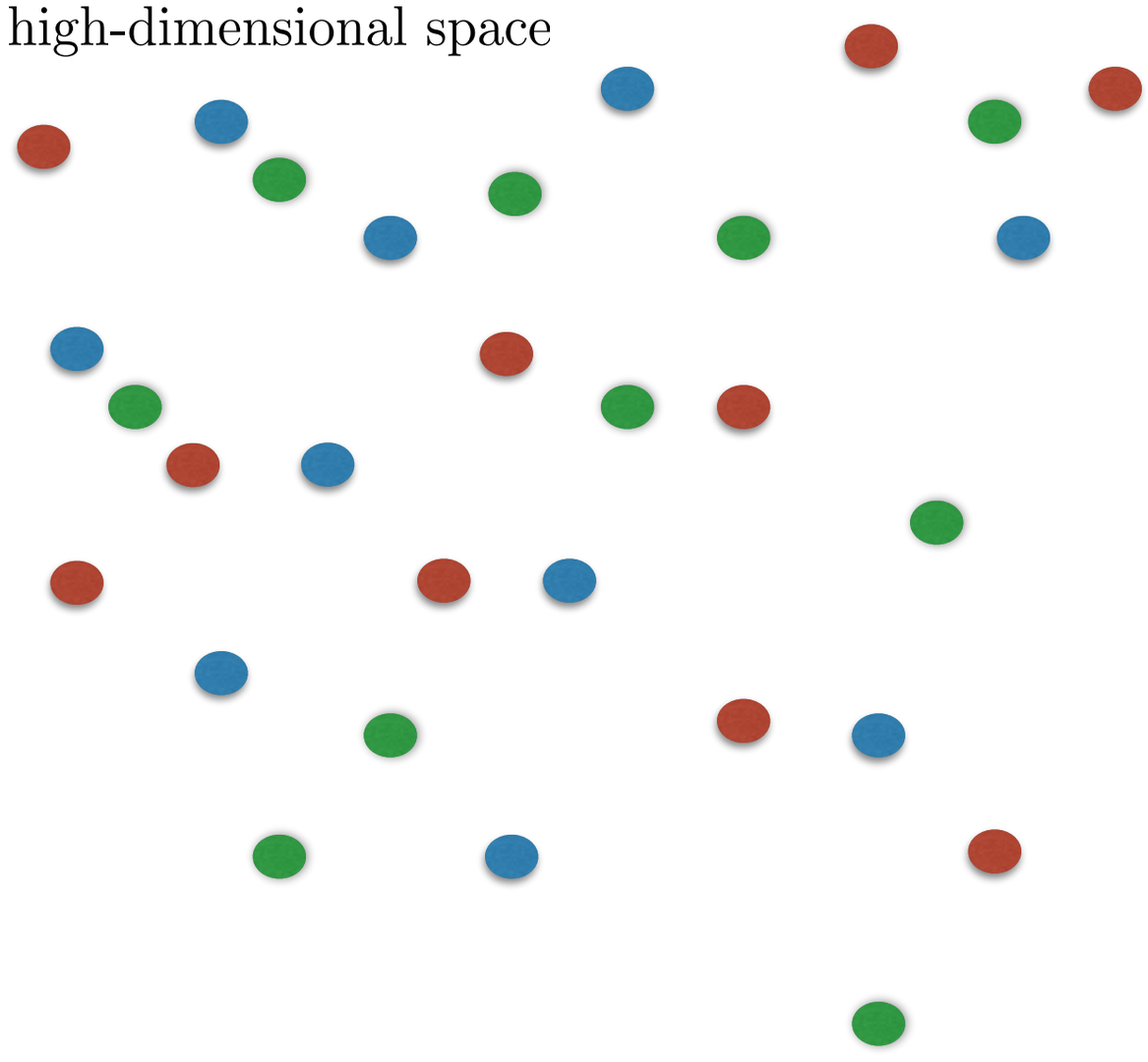
Machine Learning: what can I do for Neural Networks.

Theory??

Curse of Dimensionality

- Images, videos, audio and text are instances of high-dimensional data.

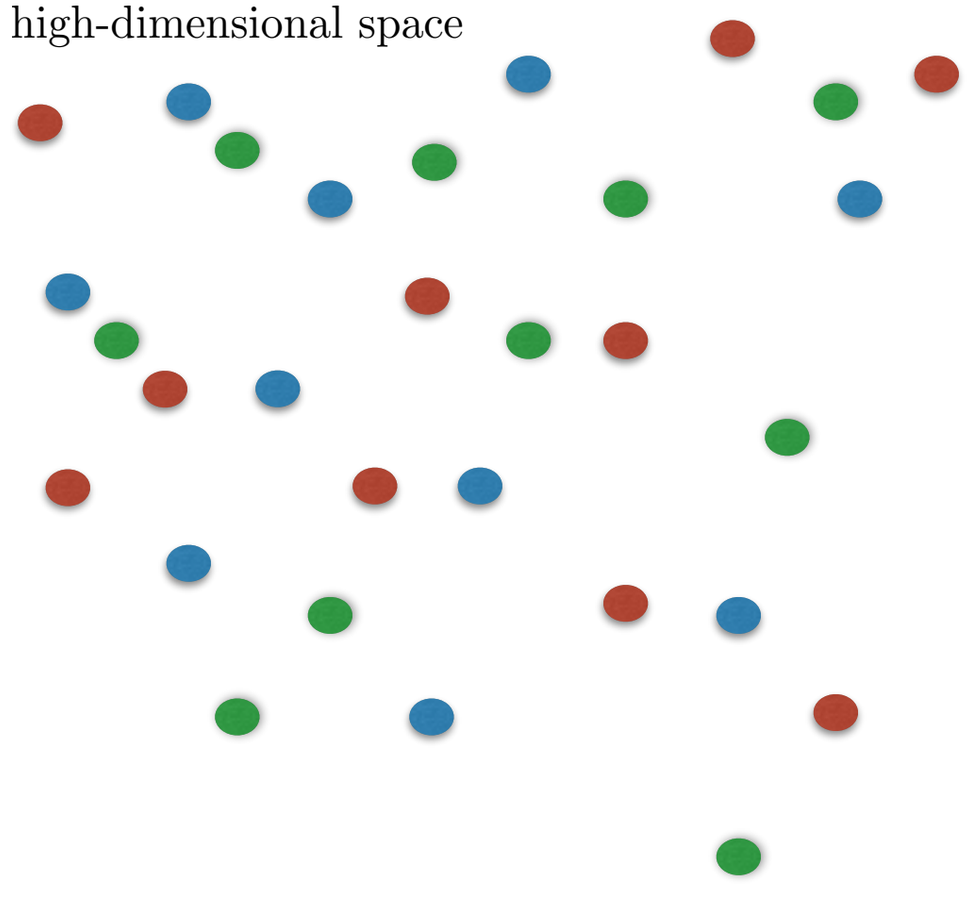
$$x \in \mathbb{R}^d \ , \ d \sim 10^6 \ .$$



Curse of Dimensionality

- Learning as a High-dimensional interpolation problem:

$$\textbf{Observe } \{(x_i, y_i = f(x_i))\}_{i \leq N} \textbf{ for some unknown } f : \mathbb{R}^d \to \mathbb{R}.$$

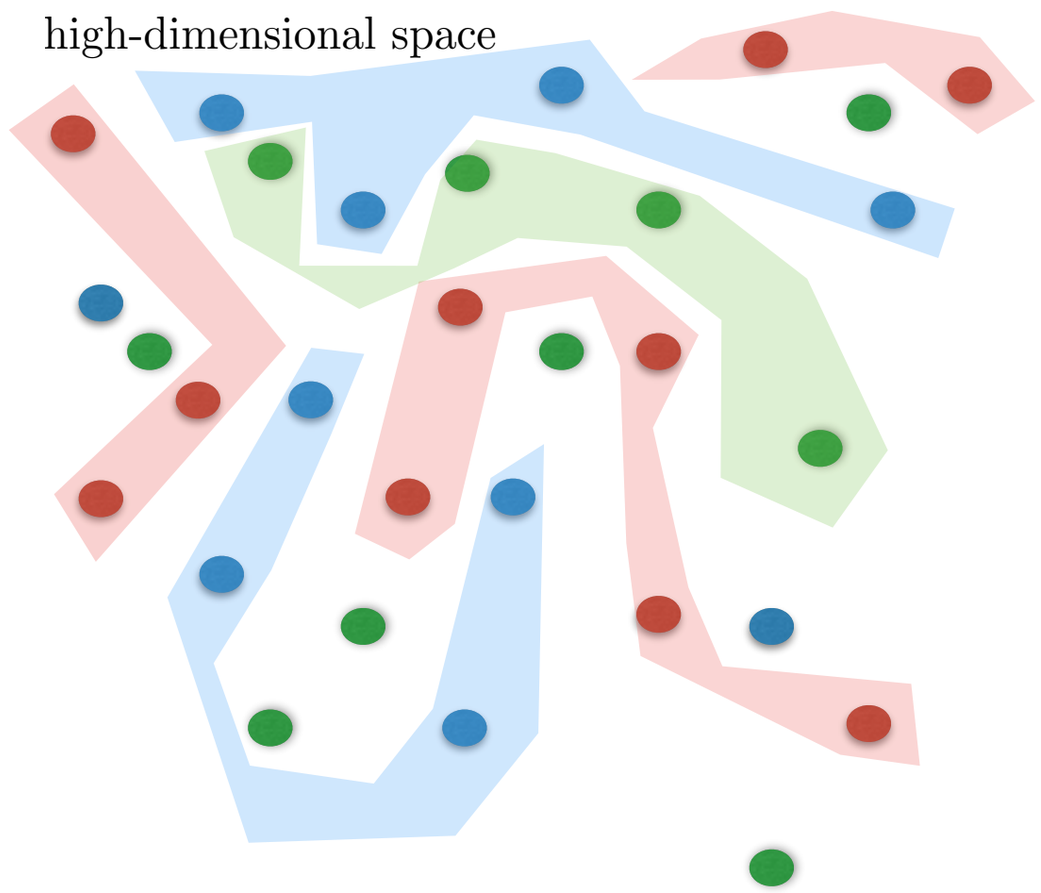


Curse of Dimensionality

- Learning as a High-dimensional interpolation problem:

$$\text{Observe } \{(x_i, y_i = f(x_i))\}_{i \leq N} \text{ for some unknown } f : \mathbb{R}^d \to \mathbb{R}.$$

- **Goal:** Estimate $\hat{f}$ from training data.



Curse of Dimensionality

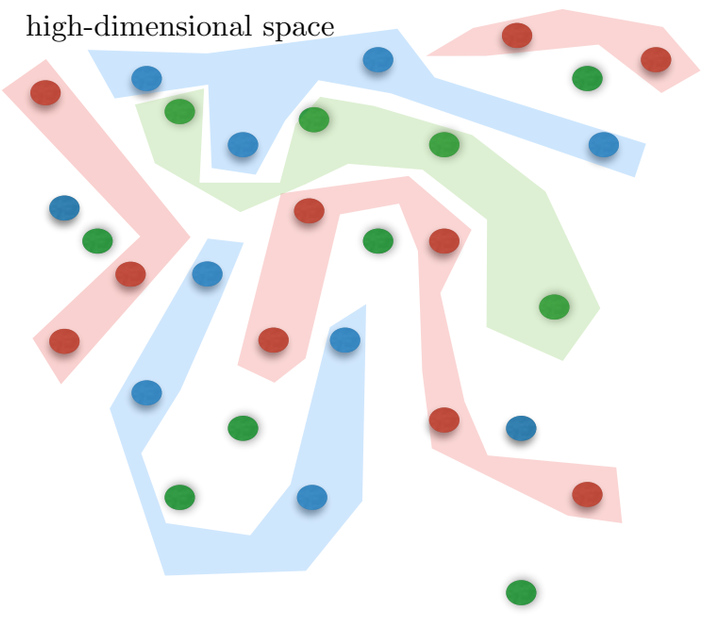
- Learning as a High-dimensional interpolation problem:

Observe $\{(x_i, y_i = f(x_i))\}_{i \leq N}$ for some unknown $f : \mathbb{R}^d \to \mathbb{R}$.

- **Goal:** Estimate $\hat{f}$ from training data.
- "Pessimist" view:

Assume as little as possible on f.

e.g. f is Lipschitz: $|f(x) - f(x')| \leq L\|x - x'\|$.



Curse of Dimensionality

- Learning as a High-dimensional interpolation problem:

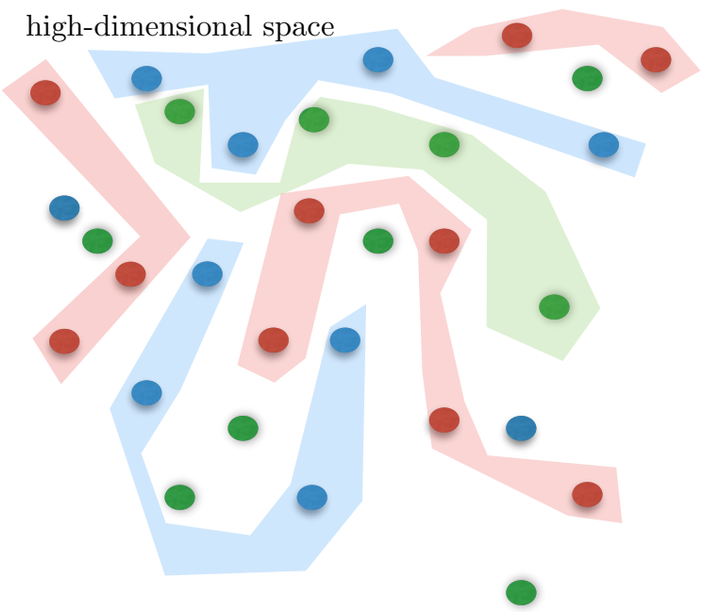
Observe $\{(x_i, y_i = f(x_i))\}_{i \leq N}$ for some unknown $f : \mathbb{R}^d \to \mathbb{R}$.

- **Goal:** Estimate $\hat{f}$ from training data.
- "Pessimist" view:
-

$$\text{Assume as little as possible on } f.$$

e.g. f is Lipschitz: $|f(x) - f(x')| \leq L\|x - x'\|$.

Q: How many points do we need to observe to guarantee error $|\hat{f}(x) - f(x)| < \epsilon$?



Curse of Dimensionality

- Learning as a High-dimensional interpolation problem:

Observe $\{(x_i, y_i = f(x_i))\}_{i \leq N}$ for some unknown $f : \mathbb{R}^d \to \mathbb{R}$.

- **Goal:** Estimate $\hat{f}$ from training data.
- "Pessimist" view:

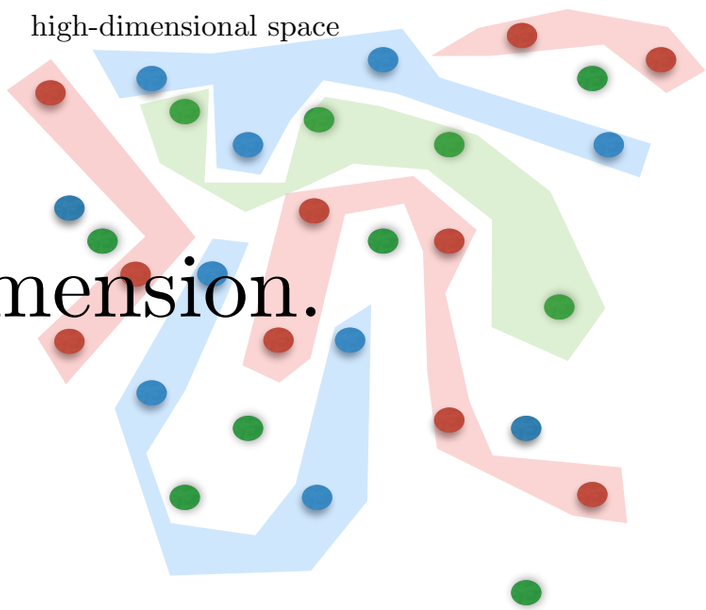
Assume as little as possible on f.

- e.g. f is Lipschitz: $|f(x) - f(x')| \leq L\|x - x'\|$.

- Q: How many points do we need to observe to guarantee error $|\hat{f}(x) - f(x)| < \epsilon$?

N should be $\sim \epsilon^{-d}$.

We pay an exponential price on the input dimension.



Curse of Dimensionality

- Therefore, in order to beat the curse of dimensionality, it is necessary to make assumptions about our data and exploit them in our models.

Curse of Dimensionality

- Therefore, in order to beat the curse of dimensionality, it is necessary to make assumptions about our data and exploit them in our models.

- Invariance and local stability perspective:

x

Supervised learning: $(x_i, y_i)_i$, $y_i \in \{1, K\}$ labels.

$f(x) = p(y \mid x)$ satisfies $f(x_\tau) \approx f(x)$ if $\{x_\tau\}_\tau$ is a high-dimensional family of deformations of x

x_τ

Unsupervised learning: $(x_i)_i$.

Density $f(x) = p(x)$ also satisfies $f(x_\tau) \approx f(x)$.

ill-posed Inverse Problems

- Consider the following linear problem:

$$y = \Gamma x + w \quad , \; x, y, w \in L^2(\mathbb{R}^d) \; , \text{ where}$$

Γ is singular
x is drawn from a certain distribution (e.g. natural images)
w independent of x

- Examples:
  - Super-Resolution, Inpainting, Deconvolution.
- Standard Regularization route:

$$\hat{x} = \arg\min_x \|\Gamma x - y\|^2 + \mathcal{R}(x) \; .$$

- Q: How to leverage training data? $\mathcal{R}(x)$ (typically) convex
- Q: Is this formulation always appropriate to estimate images?

Regularization and Learning

- The inverse problem requires a high-dimensional model for $p(x \mid y)$
- Underlying probabilistic model for regularized least squares is

$$p(x \mid y) \propto \mathcal{N}(\Gamma x - y, \mathbf{I}) e^{-\mathcal{R}(x)} \ .$$

- Suppose x_1, x_2 are such that $p(x_1|y) = p(x_2|y)$.
- Since $-\log p(x|y)$ is convex, it results that

$$p(\alpha x_1 + (1-\alpha)x_2|y) \geq p(x_1|y) \ , \ \alpha \in [0,1] \ .$$

Regularization and Learning

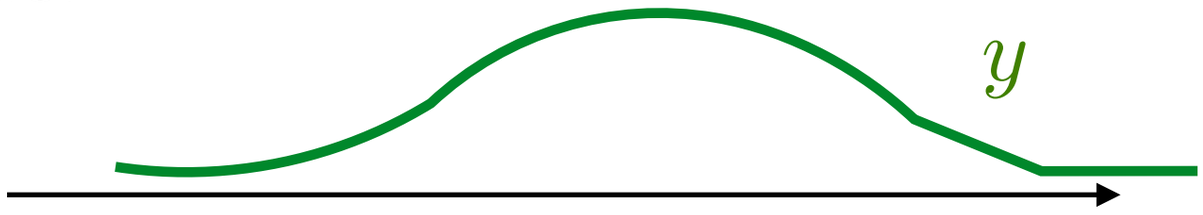
- The inverse problem requires a high-dimensional model for $p(x \mid y)$
- Underlying probabilistic model for regularized least squares is

$$p(x \mid y) \propto \mathcal{N}(\Gamma x - y, \mathbf{I}) e^{-\mathcal{R}(x)} \ .$$

- Suppose x_1, x_2 are such that $p(x_1|y) = p(x_2|y)$.
- Since $-\log p(x|y)$ is convex, it results that

$$p(\alpha x_1 + (1-\alpha)x_2|y) \geq p(x_1|y) \ , \ \alpha \in [0,1] \ .$$

- Jitter model



Regularization and Learning

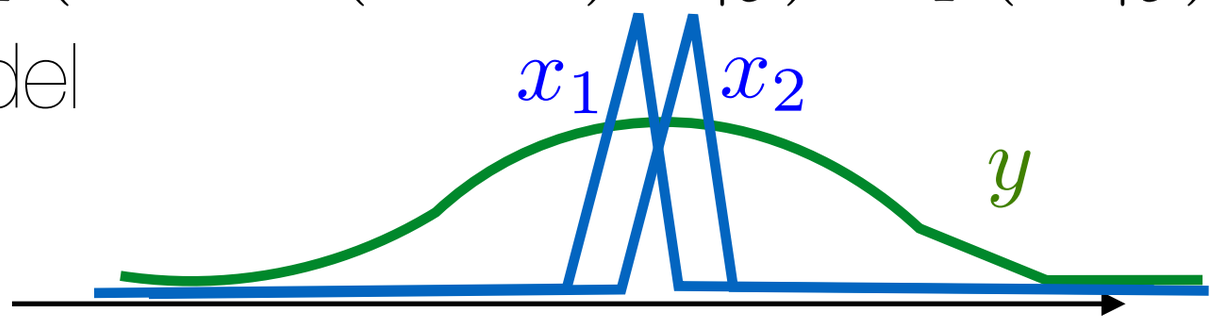
- The inverse problem requires a high-dimensional model for $p(x \mid y)$
- Underlying probabilistic model for regularized least squares is

$$p(x \mid y) \propto \mathcal{N}(\Gamma x - y, \mathbf{I}) e^{-\mathcal{R}(x)} \ .$$

- Suppose x_1, x_2 are such that $p(x_1|y) = p(x_2|y)$.
- Since $-\log p(x|y)$ is convex, it results that

$$p(\alpha x_1 + (1-\alpha)x_2|y) \geq p(x_1|y) \ , \ \alpha \in [0,1] \ .$$

- Jitter model



Regularization and Learning

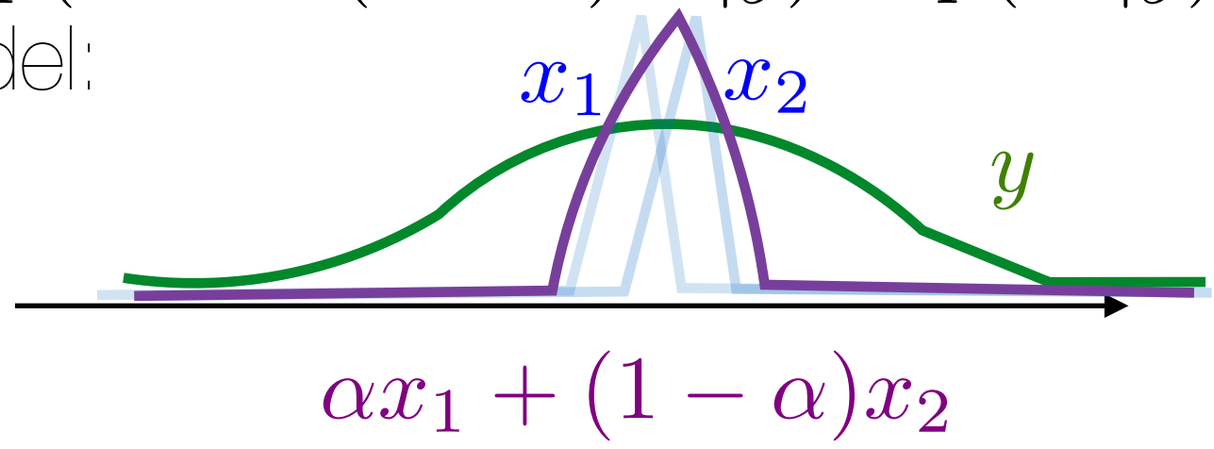
- The inverse problem requires a high-dimensional model for $p(x \mid y)$
- Underlying probabilistic model for regularized least squares is

$$p(x \mid y) \propto \mathcal{N}(\Gamma x - y, \mathbf{I}) e^{-\mathcal{R}(x)} \ .$$

- Suppose x_1, x_2 are such that $p(x_1|y) = p(x_2|y)$.
- Since $-\log p(x|y)$ is convex, it results that

$$p(\alpha x_1 + (1-\alpha)x_2|y) \geq p(x_1|y) \ , \ \alpha \in [0,1] \ .$$

- Jitter model:



- The conditional distribution of images is not well modeled only with Gaussian noise. Non-convexity in general.

Plan

- CNNs and the curse of dimensionality on image/audio regression.
- Multi scale wavelet scattering convolutional networks.
- Generative Models using CNN sufficient statistics.
- Applications to inverse problems: Image and Texture Super-Resolution

joint work with P. Sprechmann (NYU), Yann LeCun (NYU/FAIR)
Ivan Dokmanic (UIUC), S. Mallat (ENS) and M. de Hoop (Rice)

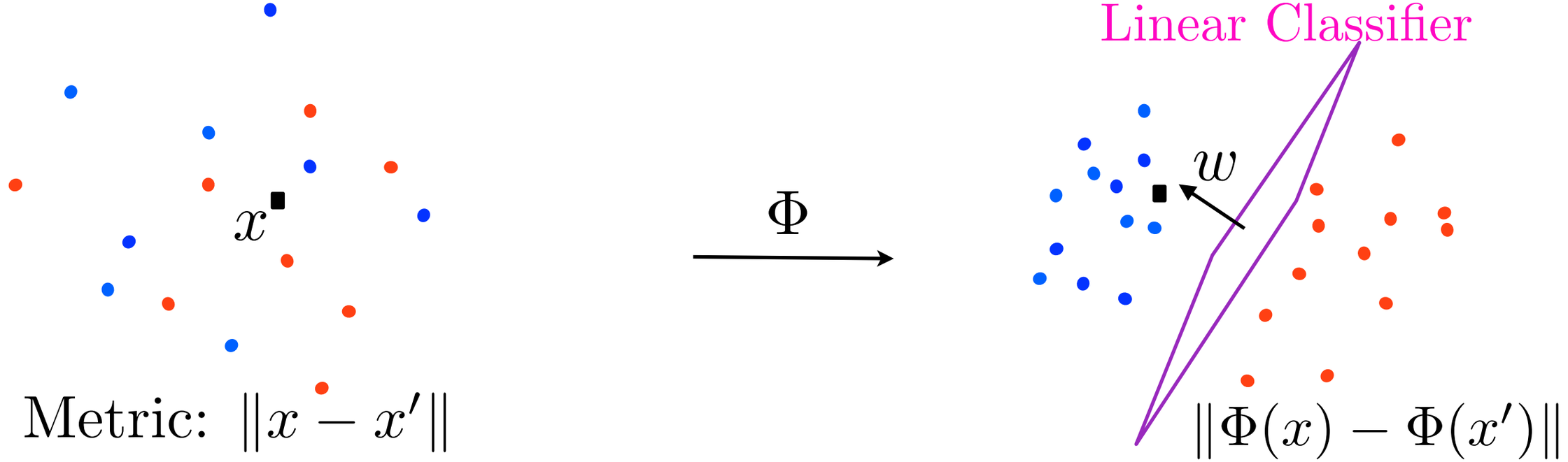
Learning, features and Kernels

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

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

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

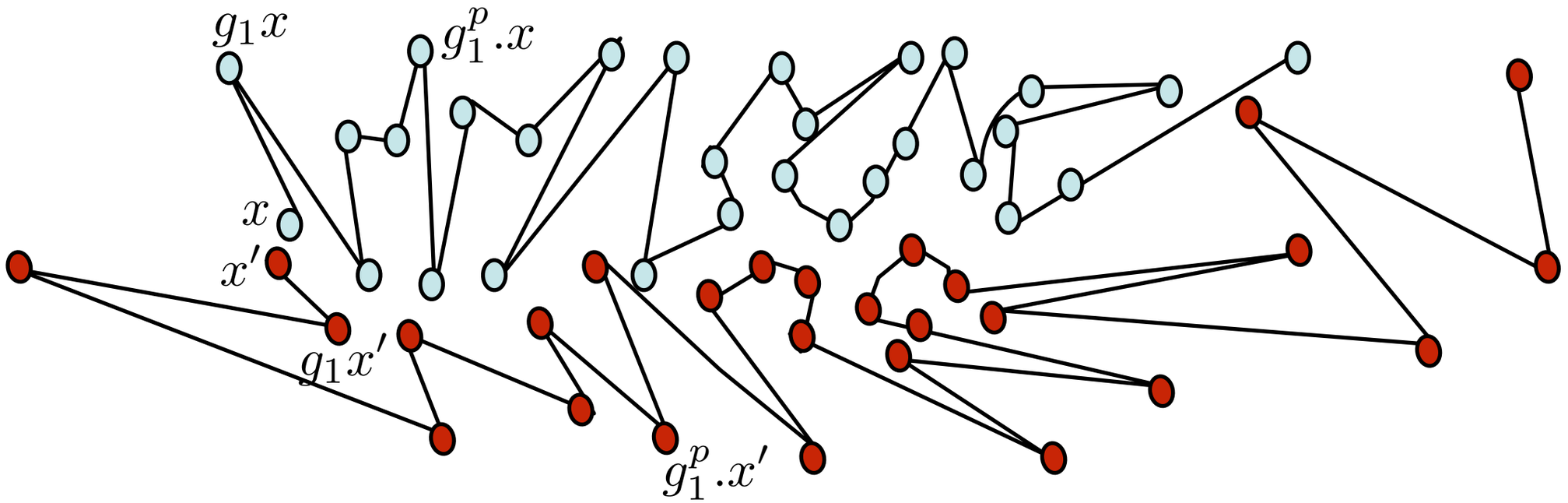
Data: $x \in \mathbb{R}^d$ $\qquad\qquad$ $\Phi(x) \in \mathbb{R}^{d'}$



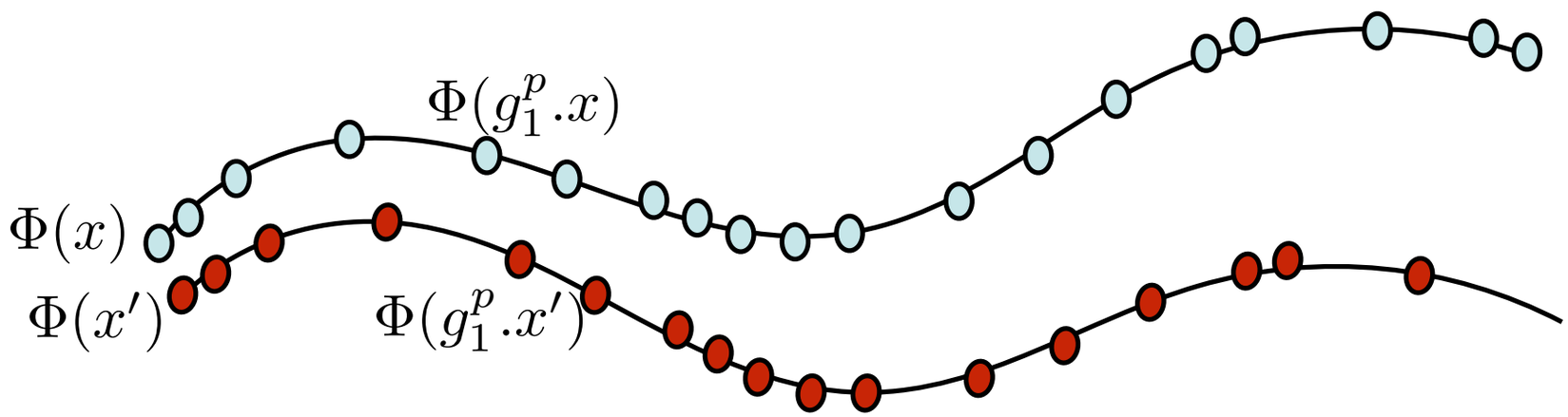
- How and when is possible to find such a Φ ?
- What "regularity" of f is needed ?

Feature Representation Wish-List

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

Geometric Variability Prior

$x(u)$, u : pixels, time samples, etc. $\tau(u)$, : deformation field

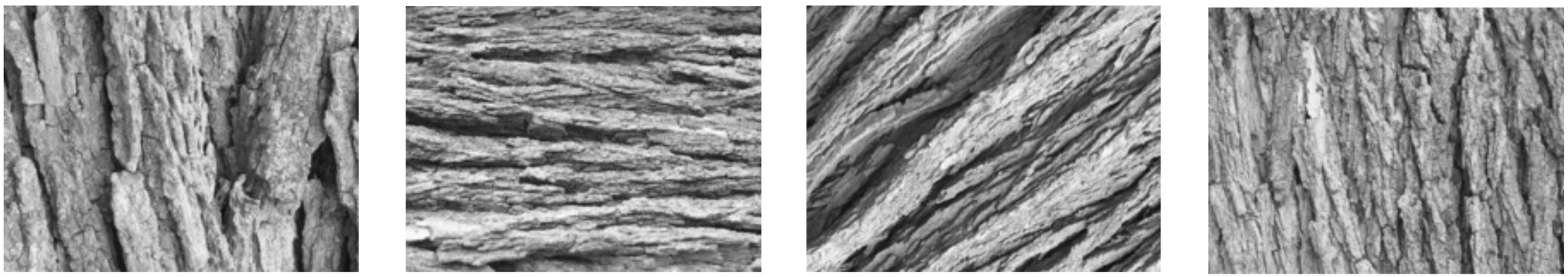
$$L_\tau(x)(u) = x(u - \tau(u)) \text{ : warping}$$

Video of Philipp Scott Johnson

- Deformation prior: $$\|\tau\| = \lambda \sup_u |\tau(u)| + \sup_u |\nabla \tau(u)| \ .$$
  - Models change in point of view in images
  - Models frequency transpositions in sounds
  - Consistent with local translation invariance

Rotation and Scaling Variability

- Rotation and deformations



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

- Scaling and deformations

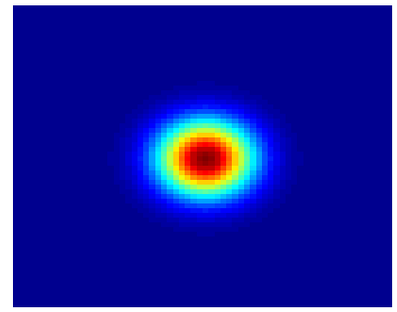
Group: $\mathbb{R} \times \text{Diff}(\mathbb{R})$

Geometric Variability Prior

- Blur operator: $Ax = x * \phi \ , \ \phi\text{: local average}$
  - The only **linear** operator A stable to deformations

$$\|AL_\tau x - Ax\| \leq \|\tau\| \|x\| \ .$$

[Bruna'12]



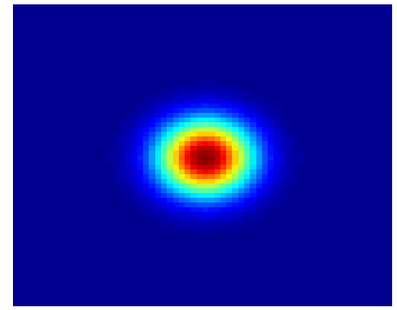
$\phi(u)$

Geometric Variability Prior

- Blur operator: $Ax = x * \phi \ , \ \phi$: local average
  - The only **linear** operator A stable to deformations

$$\|AL_\tau x - Ax\| \leq \|\tau\| \|x\| \ .$$

[Bruna'12]



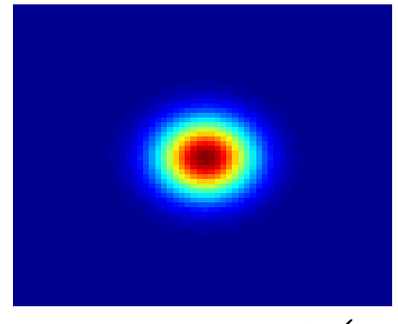
$\phi(u)$

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

Geometric Variability Prior

- Blur operator: $Ax = x * \phi \ , \ \phi\text{: local average}$
  - The only **linear** operator A stable to deformations:

$$\|AL_\tau x - Ax\| \leq \|\tau\| \|x\| \ .$$



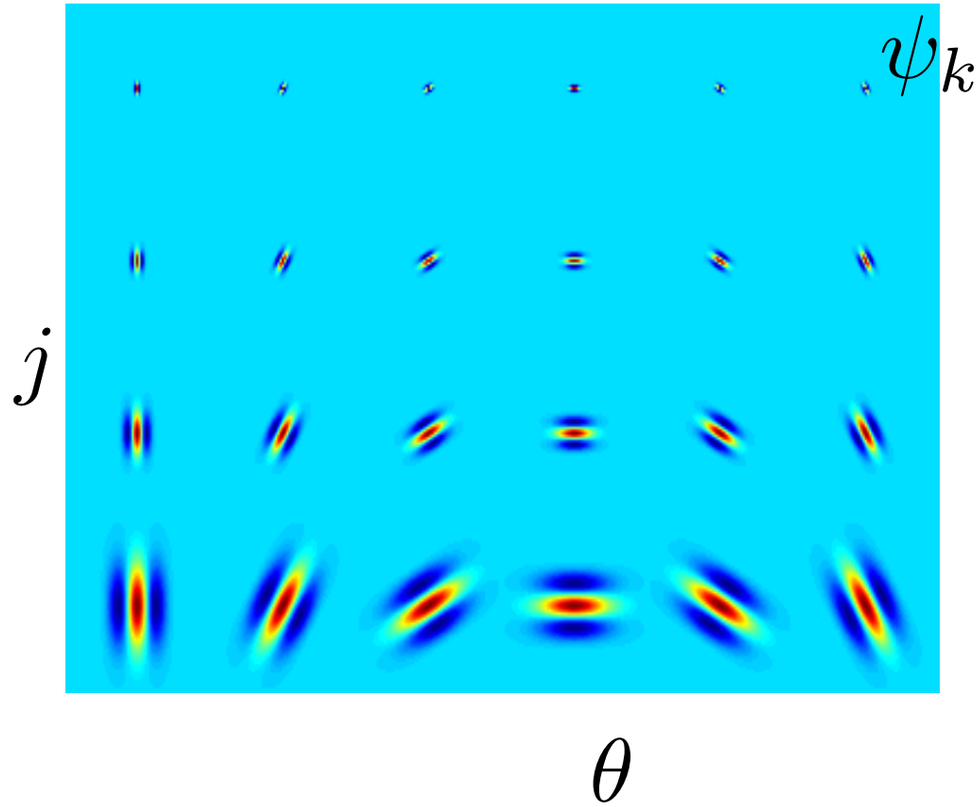
$\phi(u)$

[Bruna'12]

- Wavelet filter bank: $Wx = \{x * \psi_k\} \ , \ \psi_k(u) = 2^{-j}\psi(2^{-j}R_\theta u)$

ψ: spatially localized band-pass filter.

W recovers information lost by A.



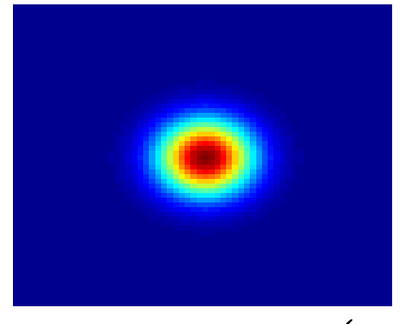
Geometric Variability Prior

- Blur operator: $Ax = x * \phi \ , \ \phi$: local average

  – The only **linear** operator A stable to deformations:

$$\|AL_\tau x - Ax\| \leq \|\tau\| \|x\| \ .$$

[Bruna'12]



$\phi(u)$

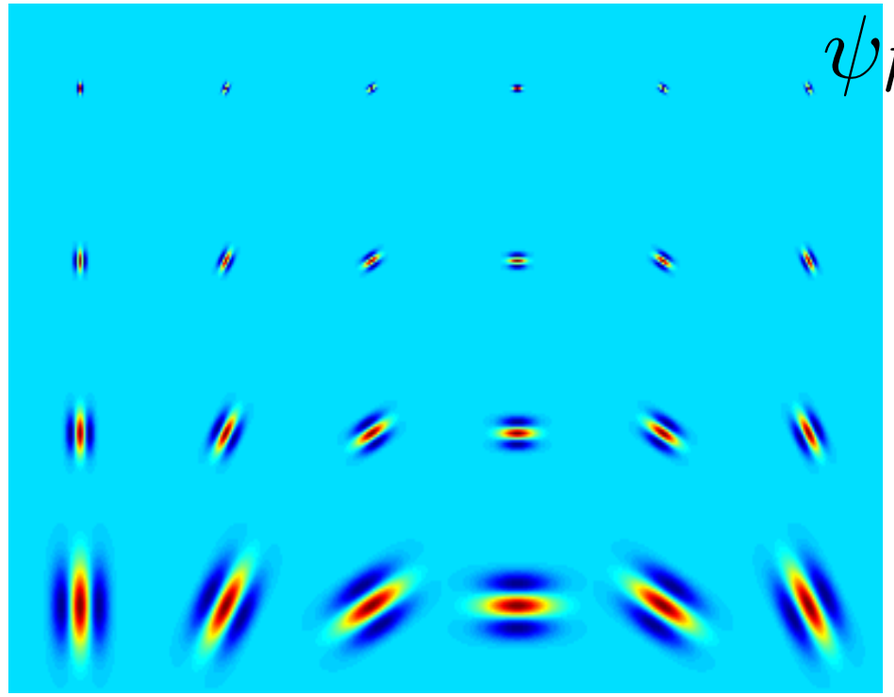
- Wavelet filter bank: $Wx = \{x * \psi_k\} \ , \ \psi_k(u) = 2^{-j}\psi(2^{-j}R_\theta u)$

ψ: spatially localized band-pass filter.

W recovers information lost by A.

W *nearly* commutes with deformations:

$\|WL_\tau - L_\tau W\| \leq C|\nabla\tau|_\infty$



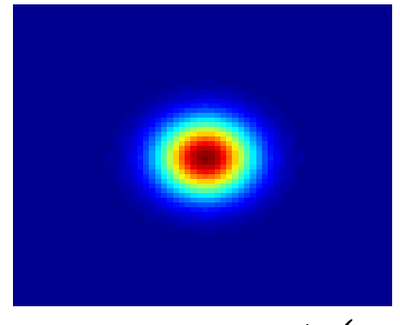
Geometric Variability Prior

- Blur operator: $Ax = x * \phi \ , \ \phi$: local average

  – The only **linear** operator A stable to deformations:

$$\|AL_\tau x - Ax\| \leq \|\tau\| \|x\| \ .$$

[Bruna'12]



$\phi(u)$

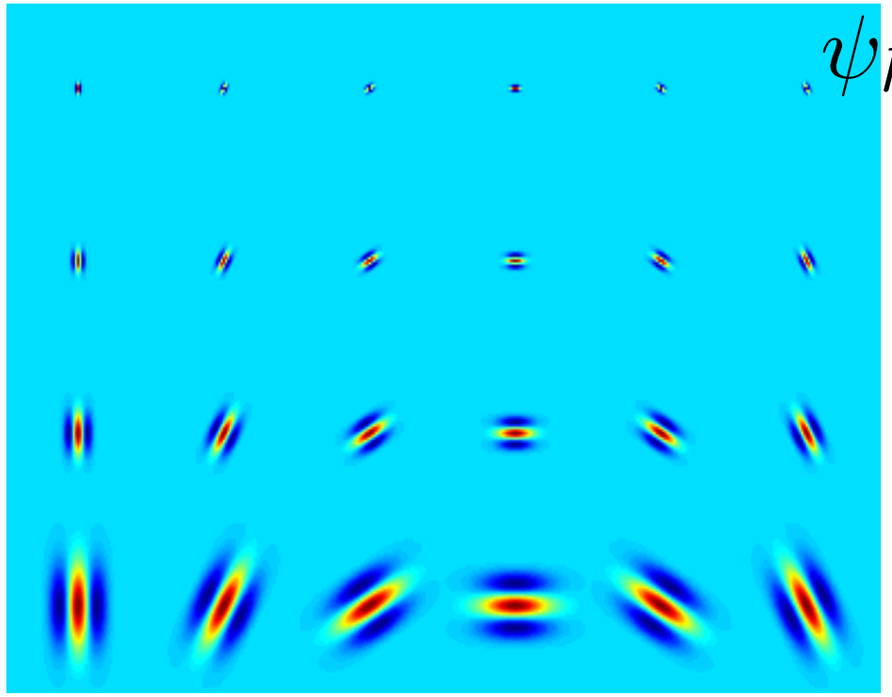
- Wavelet filter bank: $Wx = \{x * \psi_k\} \ , \ \psi_k(u) = 2^{-j}\psi(2^{-j}R_\theta u)$

ψ: spatially localized band-pass filter.

W recovers information lost by A.

W *nearly* commutes with deformations:

$\|WL_\tau - L_\tau W\| \leq C|\nabla\tau|_\infty$

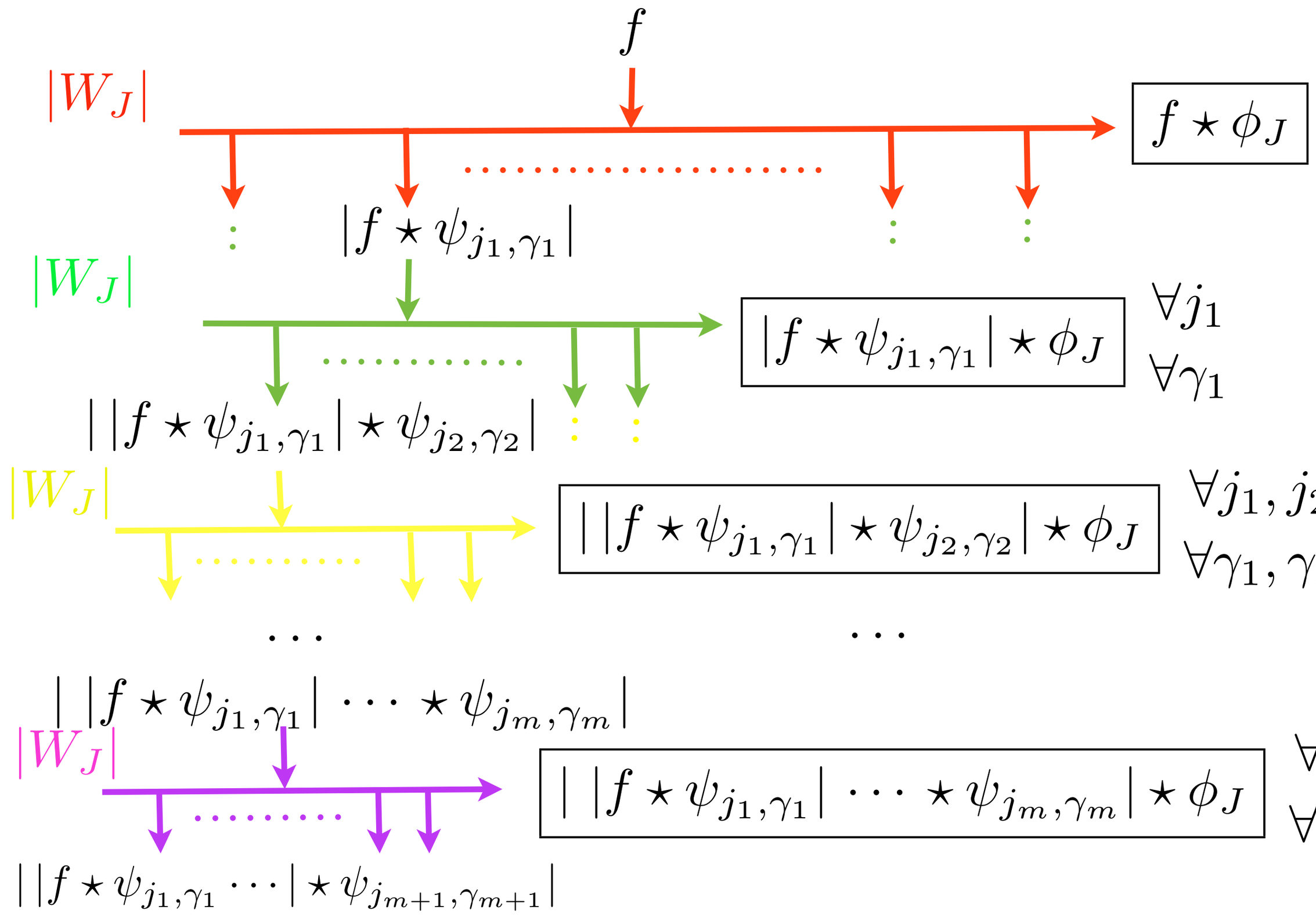


- Point-wise non-linearity $\rho(x) = |x|$

  – **Commutes** with deformations: $\rho L_\tau x = L_\tau \rho x$ [Bruna'12]

  – **Demodulates** wavelet coefficients, preserves energy.

Scattering Convolutional Network



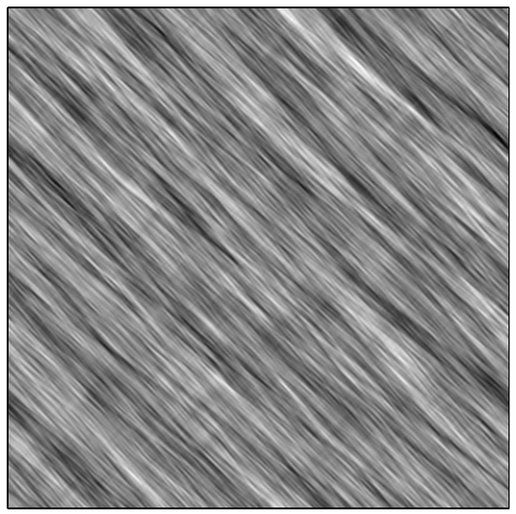
Cascade of contractive operators.

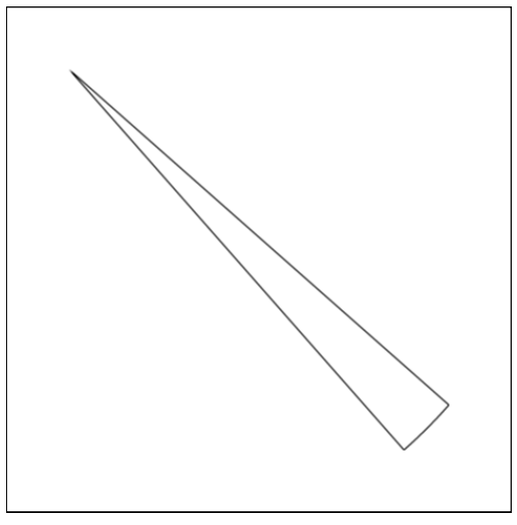
Image Examples

[Bruna, Mallat, '11,'12]

Images

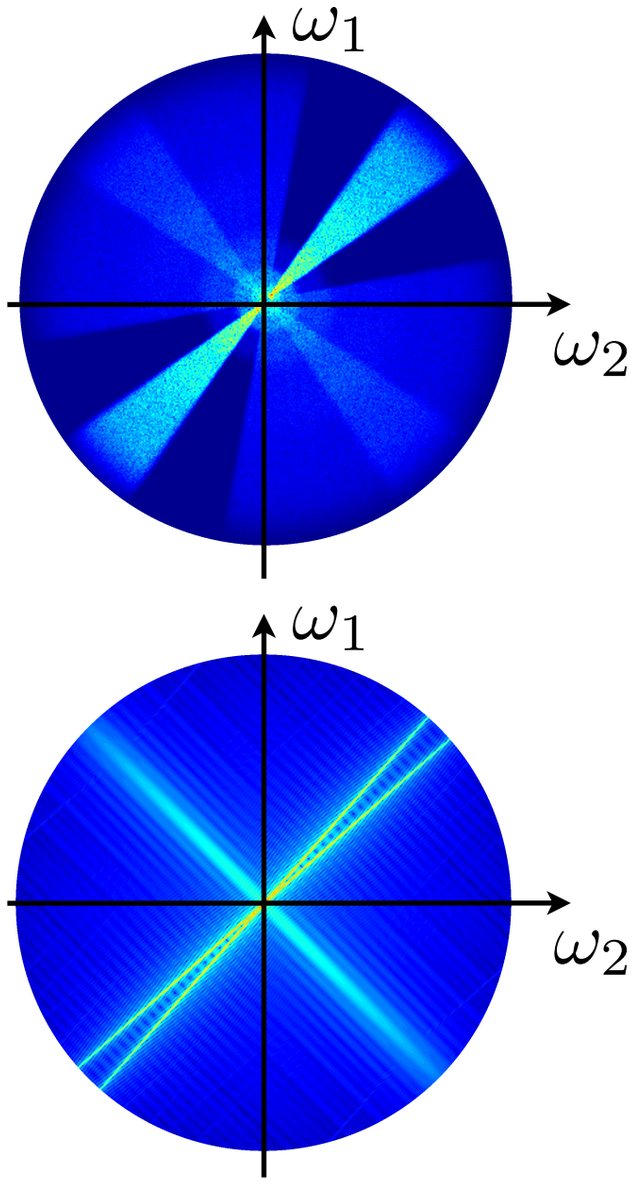
f





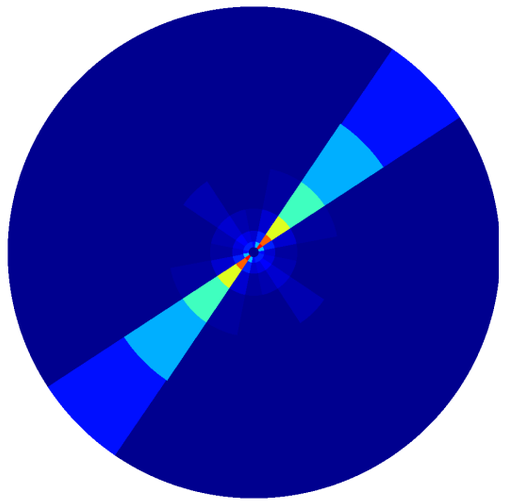
Fourier

$\hat{f}$

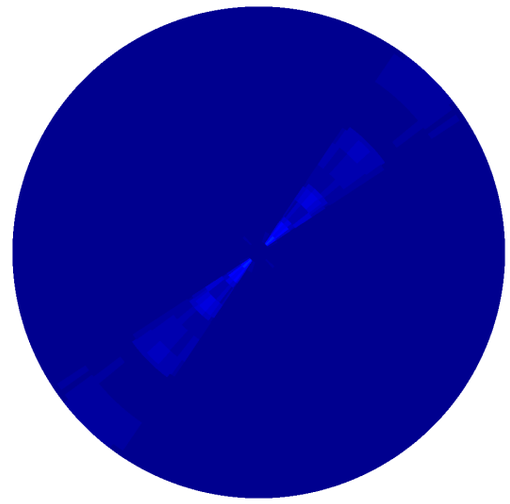


Wavelet Scattering

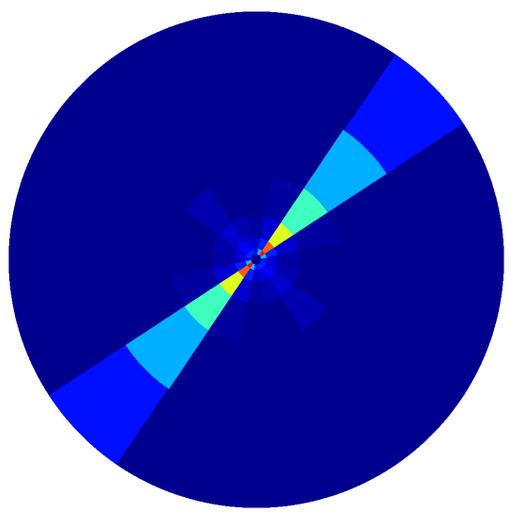
$|f \star \psi_{\lambda_1}| \star \phi$

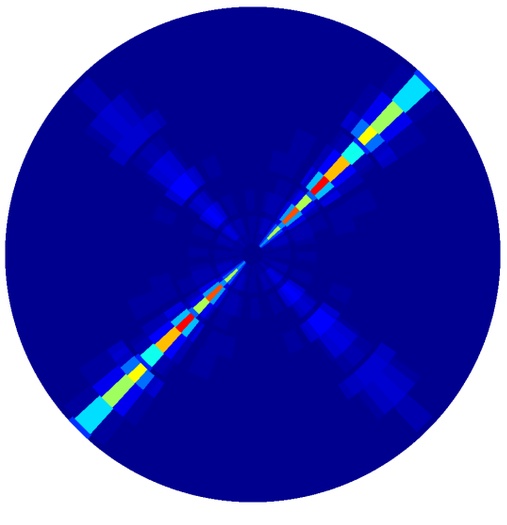


$||f \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi$



SIFT





window size = image size

Scattering Stability

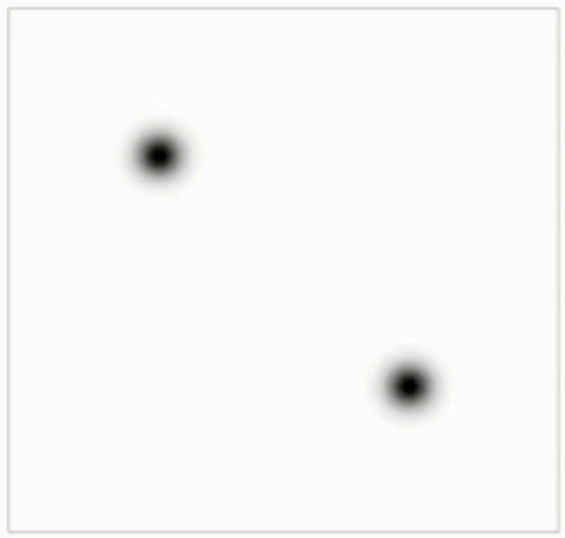
Theorem: [Mallat '10] With appropriate wavelets, S_J is stable to additive noise,

$$\|S_J(x+n) - S_J x\| \leq \|n\| \ ,$$

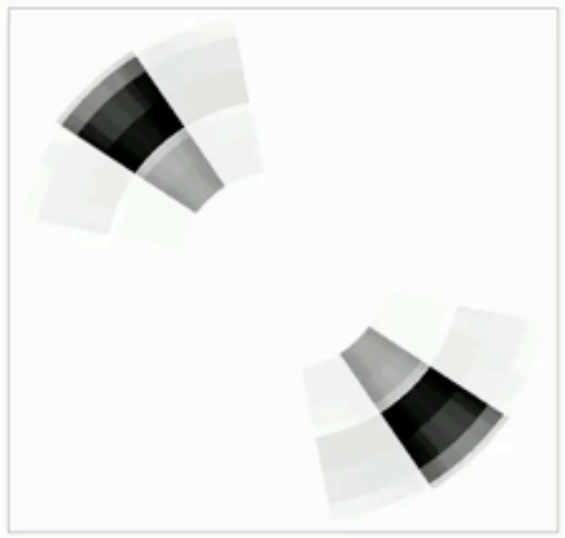
unitary, $\|S_J x\| = \|x\|$, and stable to deformations

$$\|S_J x_\tau - S_J x\| \leq C\|x\|\|\nabla\tau\| \ .$$

x_τ



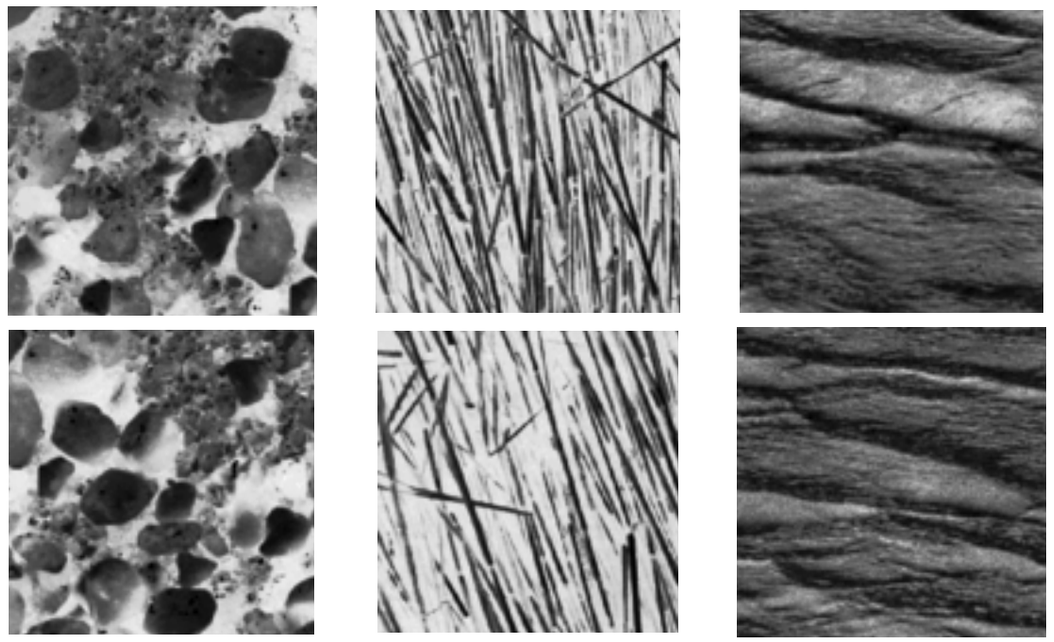
$|\widehat{x_\tau}|$



$S_J x_\tau$

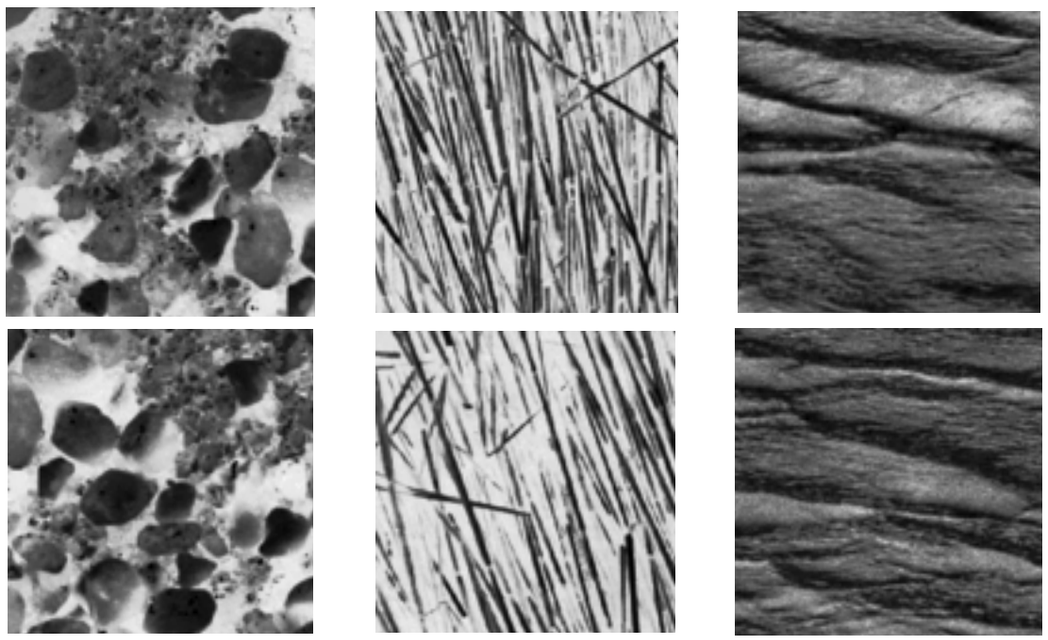
Representation of Stationary Processes

$x(u)$: realizations of a stationary process $X(u)$ (not Gaussian)



Representation of Stationary Processes

$x(u)$: realizations of a stationary process $X(u)$ (not Gaussian)



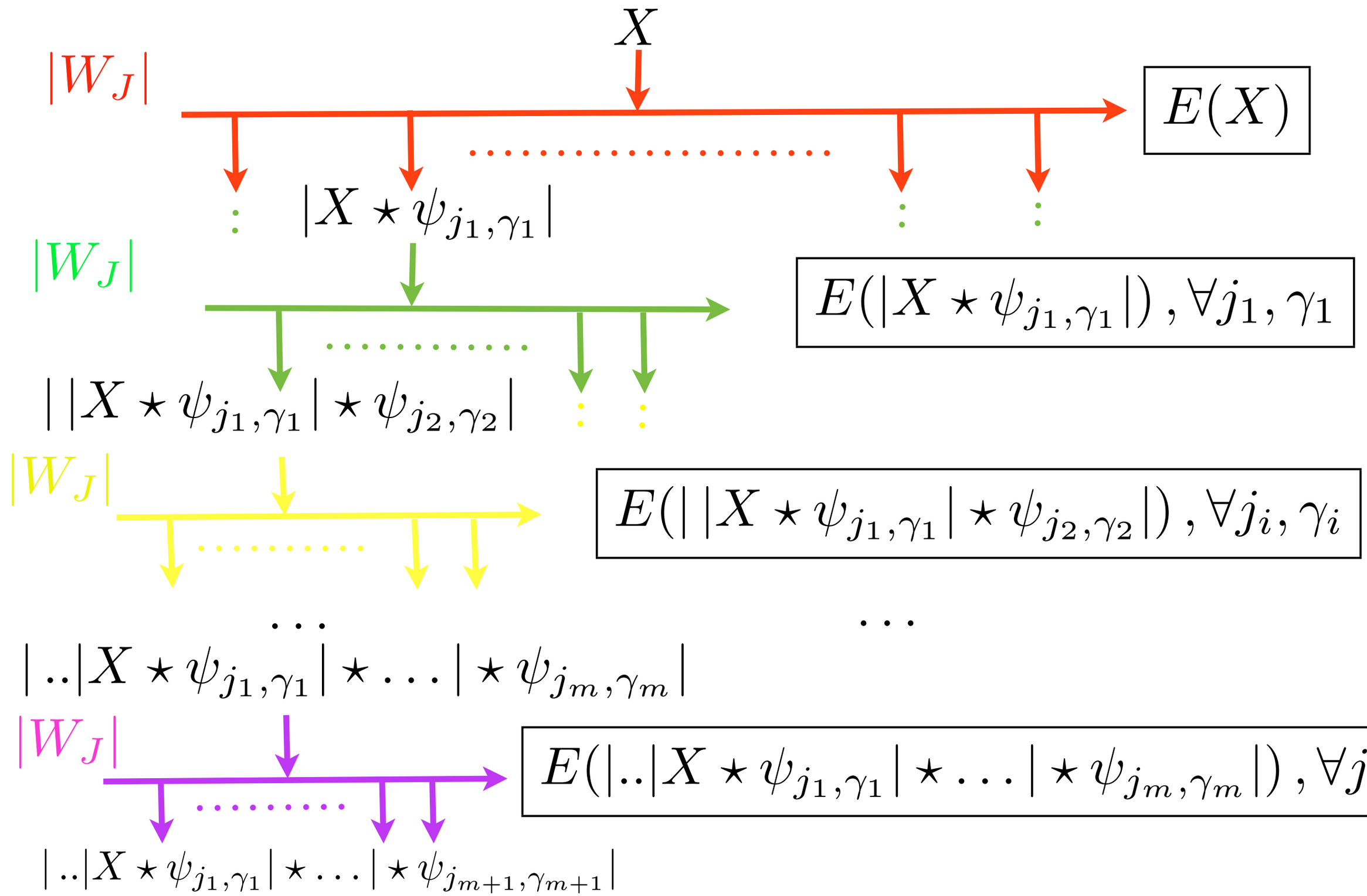
$$\Phi(X) = \{E(f_i(X))\}_i$$

Estimation from samples $x(n)$: $\widehat{\Phi}(X) = \left\{ \frac{1}{N} \sum_n f_i(x)(n) \right\}_i$

Discriminability: need to capture high-order moments

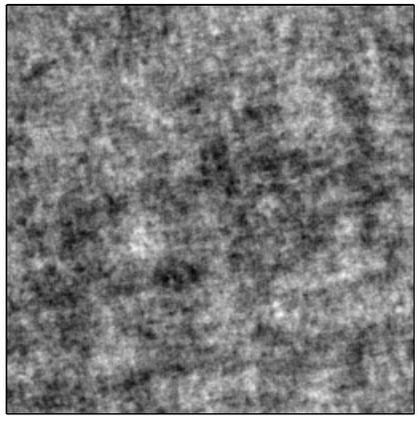
Stability: $E(\|\widehat{\Phi}(X) - \Phi(X)\|^2)$ small

Scattering Moments

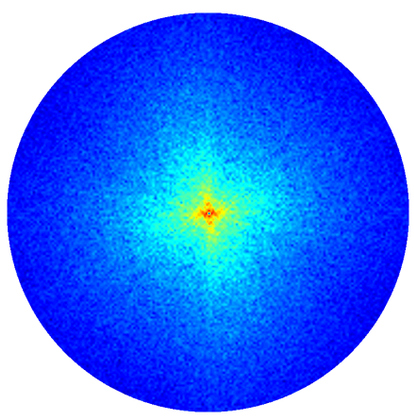


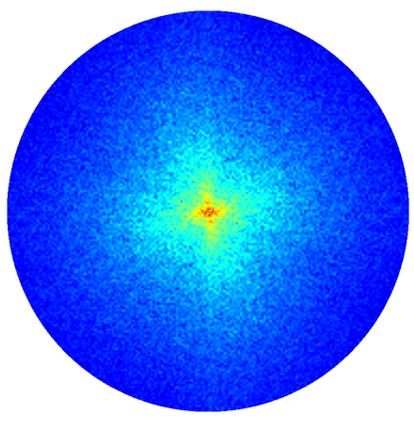
Properties of Scattering Moments

- Captures high order moments: [Bruna, Mallat, '11,'12]



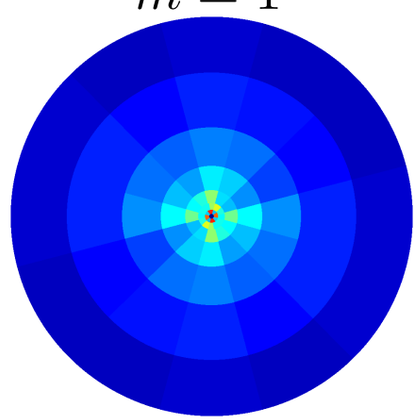
Power Spectrum

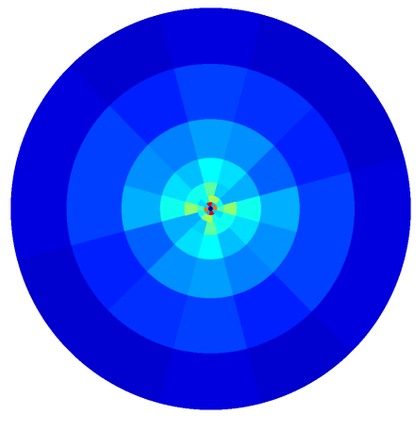




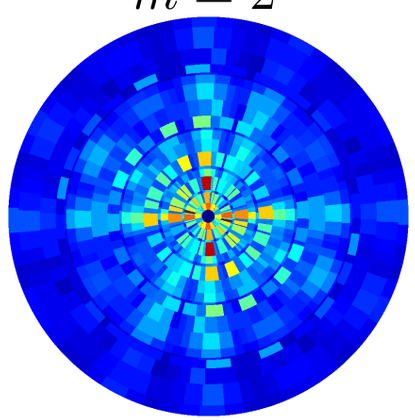
$S_J[p]X$

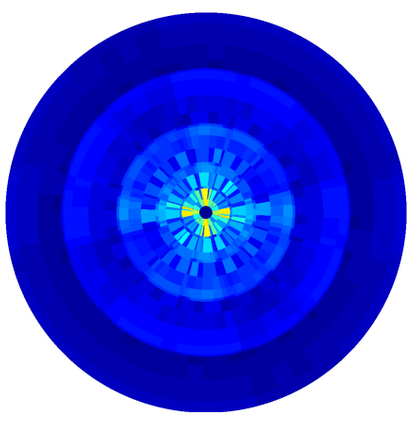
$m = 1$





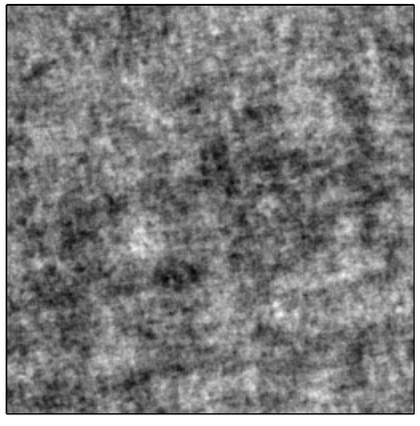
$m = 2$



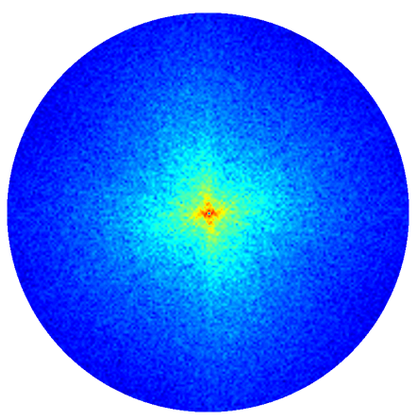


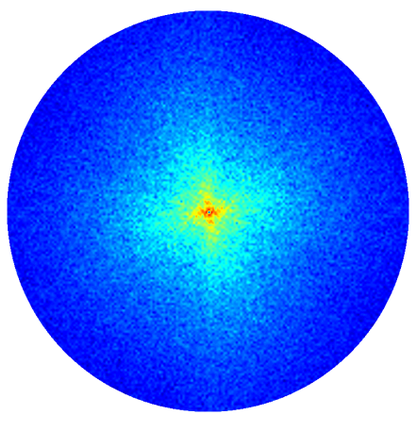
Properties of Scattering Moments

- Captures high order moments: [Bruna, Mallat, '11,'12]



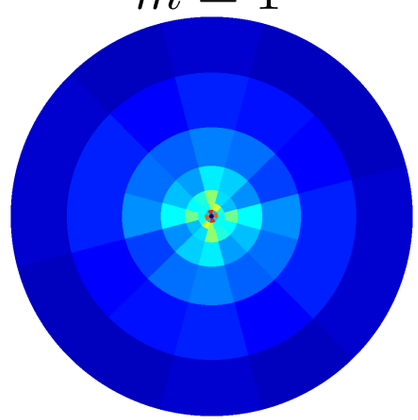
Power Spectrum

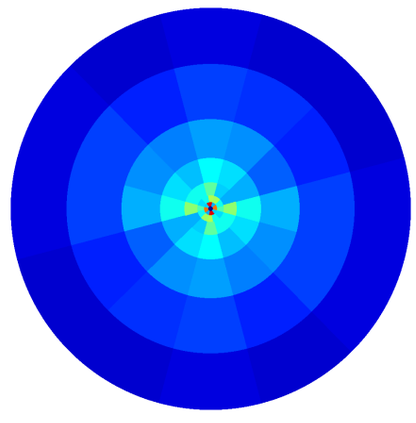




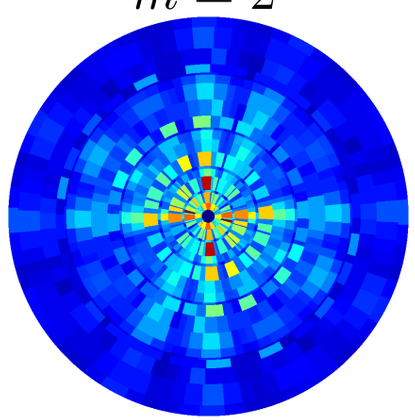
$S_J[p]X$

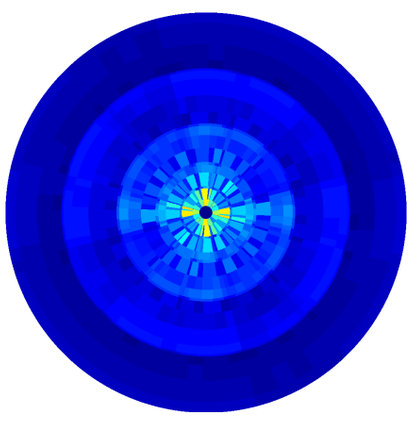
$m = 1$





$m = 2$





- Cascading non-linearities is ***necessary*** to reveal higher-order moments.

Consistency of Scattering Moments

Theorem: **[B'15]** If ψ is a wavelet such that $\|\psi\|_1 \leq 1$, and $X(t)$ is a linear, stationary process with finite energy, then

$$\lim_{N \to \infty} E(\|\hat{S}_N X - SX\|^2) = 0 \ .$$

Consistency of Scattering Moments

Theorem: **[B'15]** If ψ is a wavelet such that $\|\psi\|_1 \leq 1$, and $X(t)$ is a linear, stationary process with finite energy, then

$$\lim_{N \to \infty} E(\|\hat{S}_N X - SX\|^2) = 0 \ .$$

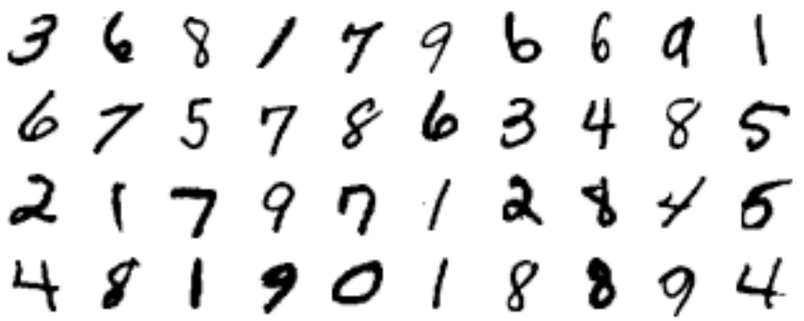
Corollary: If moreover $X(t)$ is bounded, then

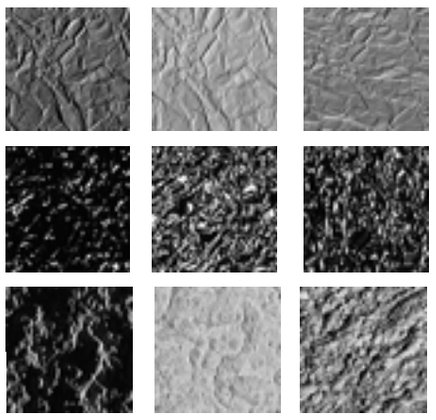
$$E(\|\hat{S}_N X - SX\|^2) \leq C \frac{|X|_\infty^2}{\sqrt{N}} \ .$$

- Although we extract a growing number of features, their global variance goes to 0.
- No variance blow-up due to high order moments.
- Adding layers is critical (here depth is log(N)).

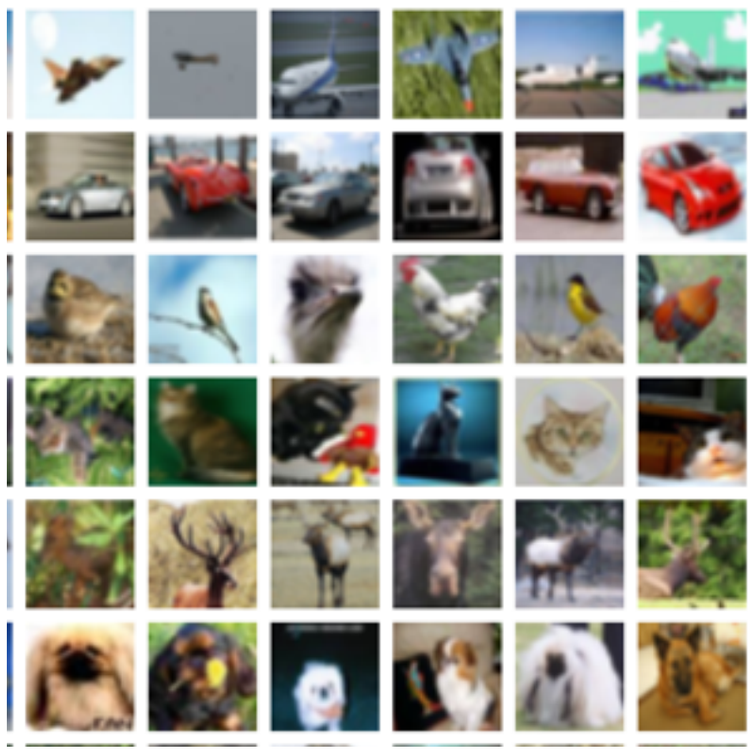
Classification with Scattering

- State-of-the art on pattern and texture recognition:
  - MNIST [Pami'13]





  - Texture (CUREt) [Pami'13]



- Object Recognition:
  - 17% error on Cifar-10 [*Oyallon, Mallat, CVPR'15*] using better second layer wavelets that recombine channels.
  - How good is that? *Not good enough.* Best models do <5%.

Scattering Limitations

- Modeling intraclass variability as geometric variability/ texture is not sufficient
  - Adapting wavelets to the dataset, even to the class.

- No simple mechanism to reduce the number of feature maps without learning
  - Scattering construction increases feature maps at exponential rate.

- Extend group formalism to understand behavior in deeper layers.
  - Although extensions to rototranslation group yield state-of-the-art results in general object recognition ["Group Equivariant CNNs", Cohen & Welling, ICML'16].

Signal and Texture Recovery Challenge

$$S_J x = \{x * \phi_J, |x * \psi_{j_1}| * \phi_J, ||x * \psi_{j_1}| * \psi_{j_2}| * \phi_J, \dots\}_{j_i \leq J}$$

- [Q1] Given $S_J x$ computed with m layers, under what conditions can we recover x (up to global symmetry)? Using what algorithm? As a function of the localization scale J ?

$$\overline{S} X = \{E(X), E(|X * \psi_{j_1}|), E(||X * \psi_{j_1}| * \psi_{j_2}|), \dots\}$$

- [Q2] Given SX, how can we characterize interesting processes? How to sample from such distributions?

Sparse Signal Recovery

Theorem [B,M'14]: Suppose $x_0(t) = \sum_n a_n \delta(t - b_n)$ with $|b_n - b_{n+1}| \geq \Delta$, and $S_J x_0 = S_J x$ with $m = 1$ and $J = \infty$. If ψ has compact support, then

$$x(t) = \sum_n c_n \delta(t - e_n) \ , \ \text{with } |e_n - e_{n+1}| \gtrsim \Delta \ .$$

Sparse Signal Recovery

Theorem [B,M'14]: Suppose $x_0(t) = \sum_n a_n \delta(t - b_n)$ with $|b_n - b_{n+1}| \geq \Delta$, and $S_J x_0 = S_J x$ with $m = 1$ and $J = \infty$. If ψ has compact support, then

$$x(t) = \sum_n c_n \delta(t - e_n) \ , \ \text{with } |e_n - e_{n+1}| \gtrsim \Delta \ .$$

- Sx essentially identifies sparse measures, up to log spacing factors.
- Here, sparsity is encoded in the measurements themselves.
- In 2D, singular measures (ie curves) require $m = 2$ to be well characterized.

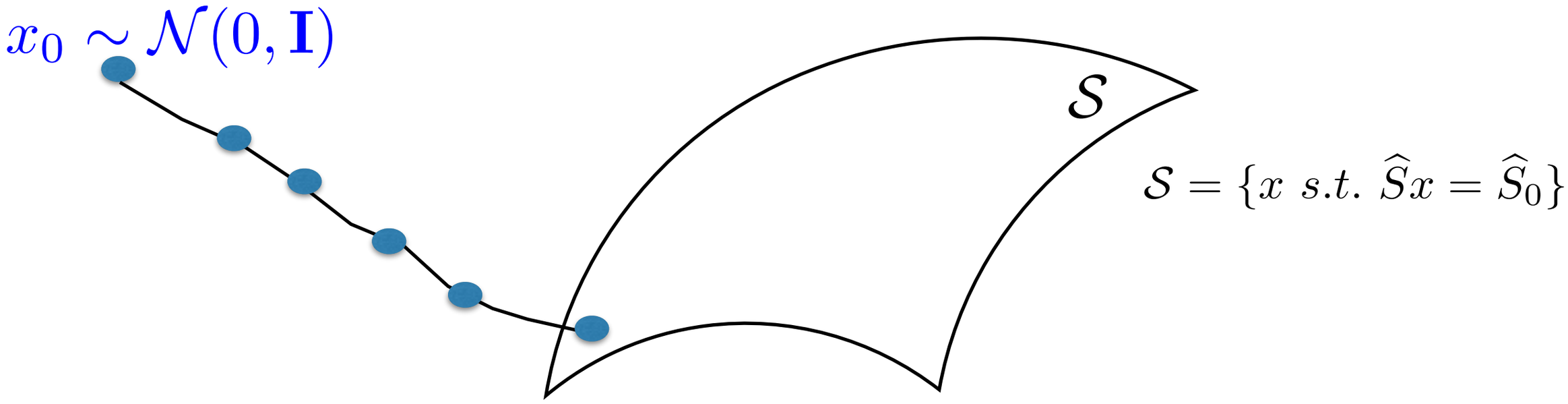
Oscillatory Signal Recovery

Theorem [B,M'14]: Suppose $\widehat{x_0}(\xi) = \sum_n a_n \delta(\xi - b_n)$ with $|\log b_n - \log b_{n+1}| \geq \Delta$, and $S_J x = S_J x_0$ with $m = 2$ and $J = \log N$. If $\widehat{\psi}$ has compact support $K \leq \Delta$, then

$$\widehat{x}(\xi) = \sum_n c_n \delta(\xi - e_n) \ , \ \text{with } |\log e_n - \log e_{n+1}| \gtrsim \Delta \ .$$

- Oscillatory, lacunary signals are also well captured with the **same** measurements.
- It is the opposite set of extremal points from previous result.

Scattering Reconstruction Algorithm

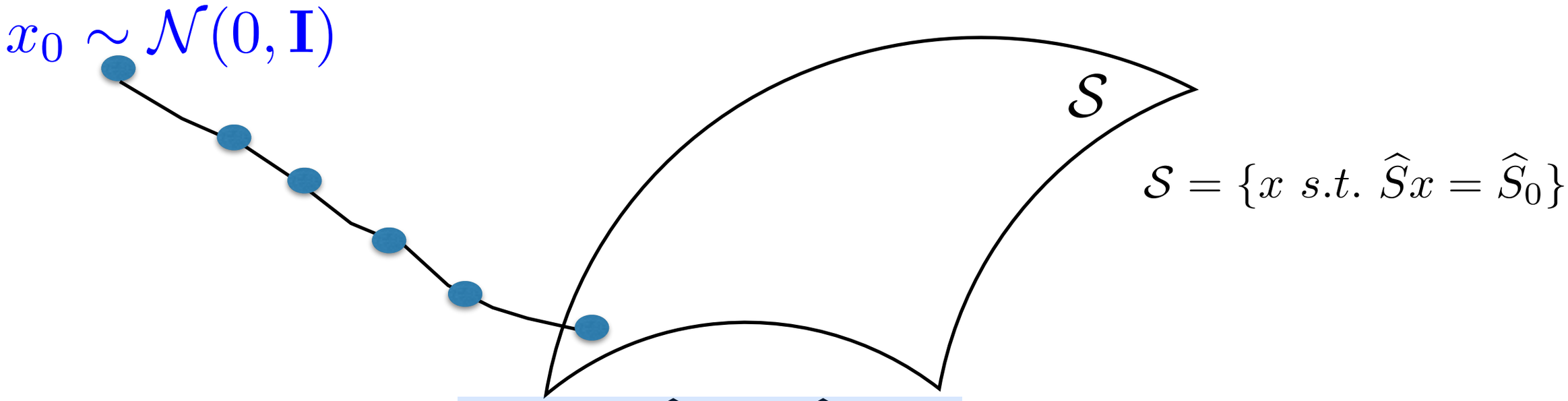


$$\min_x \|\widehat{S}x - \widehat{S}_0\|^2$$

- Non-linear Least Squares, non-convex optimization.
  - Levenberg-Marquardt gradient descent:

$$x_{n+1} = x_n - \gamma(D\widehat{S}x_n)^\dagger(\widehat{S}x_n - \widehat{S}_0)$$

Scattering Reconstruction Algorithm



- Non-linear Least Squares.
  - Levenberg-Marquardt gradient descent:

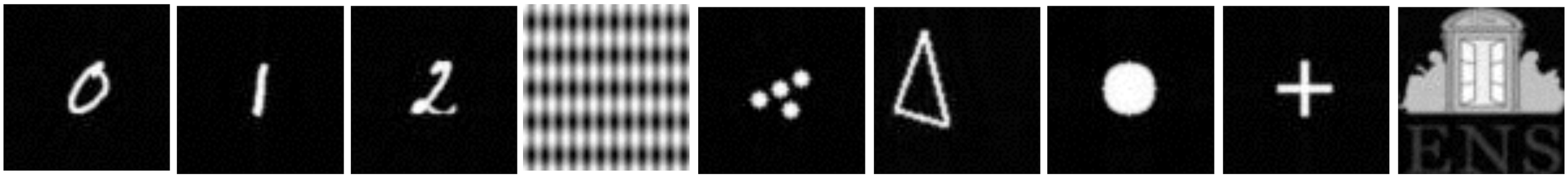
$$x_{n+1} = x_n - \gamma(D\widehat{S}x_n)^\dagger(\widehat{S}x_n - \widehat{S}_0)$$

- (Weak) Global convergence guarantees using complex wavelets:

$$D\hat{S}x \text{ is full rank for } m = 2 \text{ if } x \text{ compact support.}$$

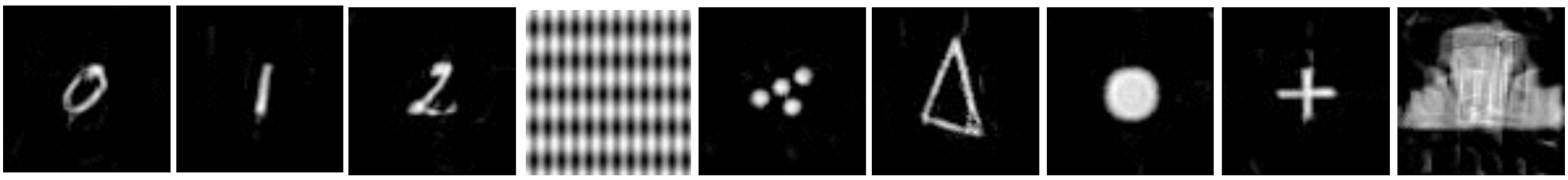
Sparse Shape Reconstructions

Original images of N^2 pixels:



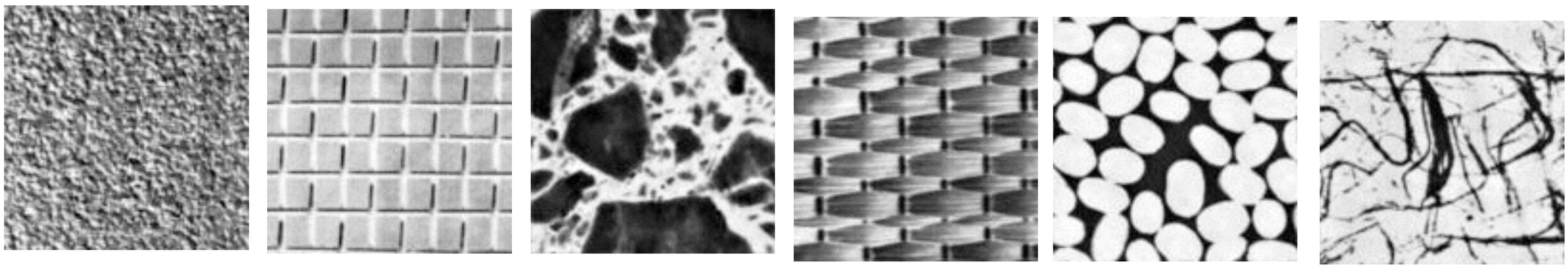
$m = 1$, $2^J = N$: reconstruction from $O(\log_2 N)$ scattering coeff.

$m = 2$, $2^J = N$: reconstruction from $O(\log_2^2 N)$ scattering coeff.

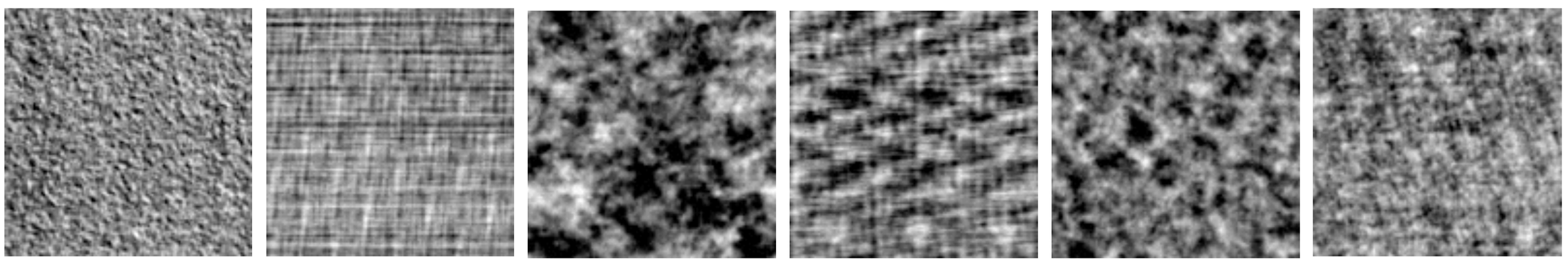


Ergodic Texture Reconstruction

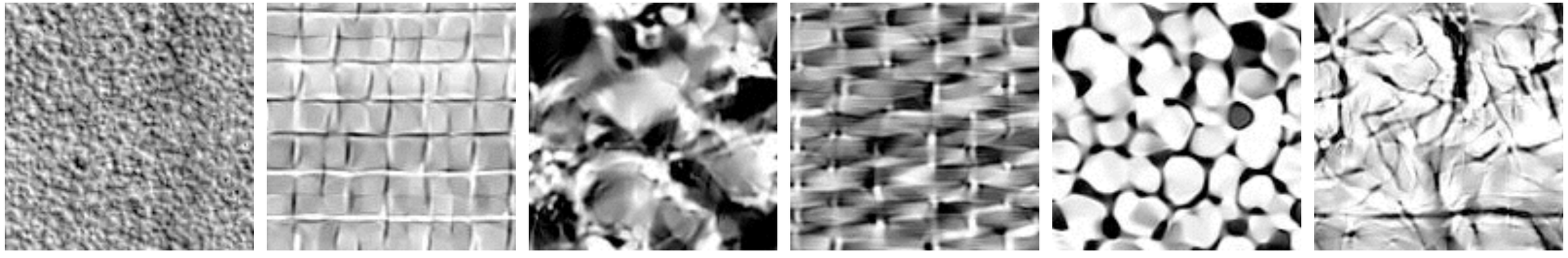
Original Textures



Gaussian process model with same second order moments



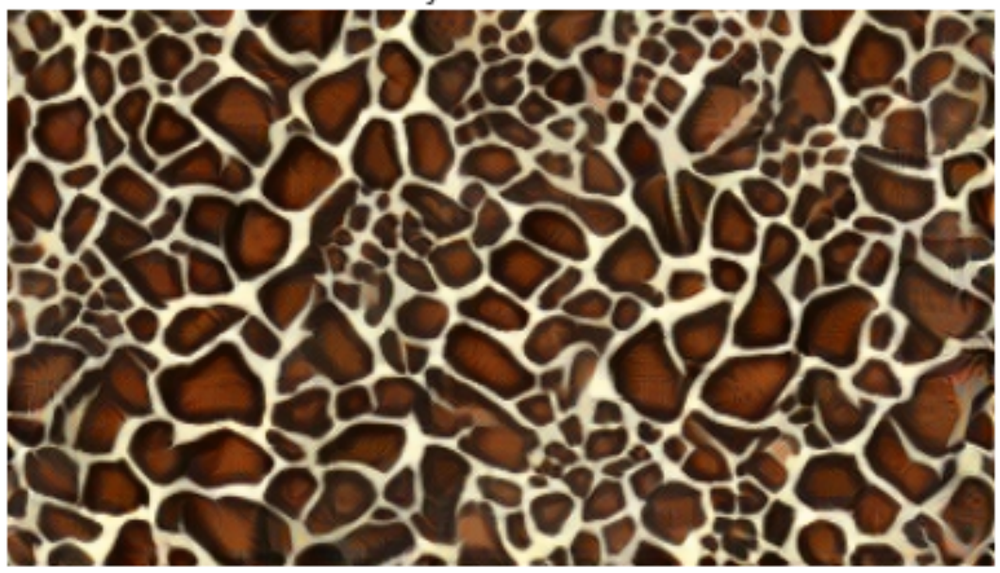
$m = 2$, $2^J = N$: reconstruction from $O(\log_2^2 N)$ scattering coeff.



Ergodic Texture Reconstruction using CNNs

- Results using a deep CNN network from [Gathys et al, NIPS'15]
- Uses a much larger feature vector than scattering ($O((\log N)^2)$).

Synthesised



Source

Synthesised

Source

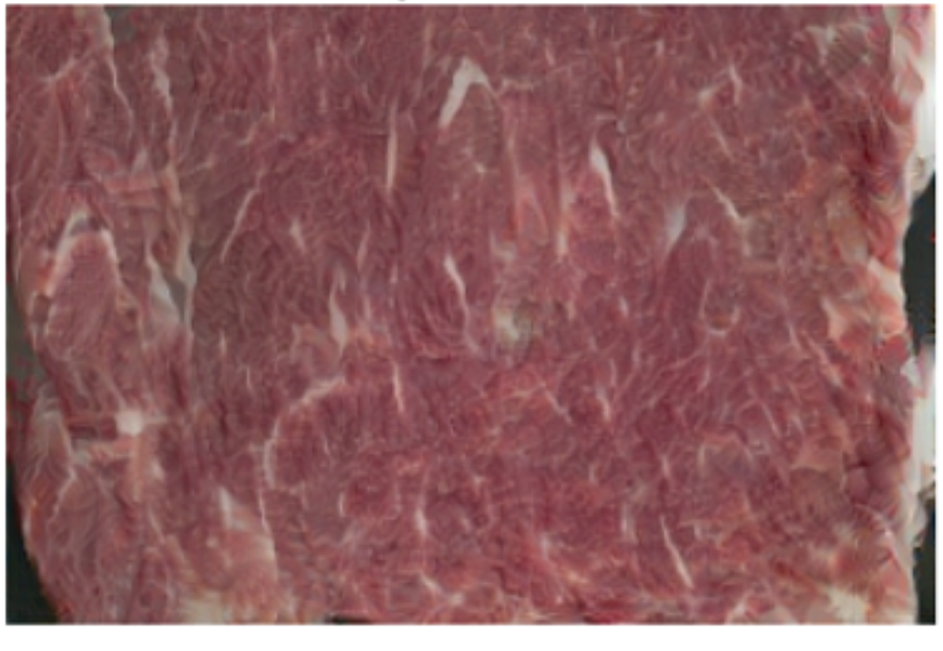
Ergodic Texture Reconstruction using CNNs

- Results using a deep CNN network from [Gathys et al, NIPS'15]
- Uses a much larger feature vector than scattering ($O((\log N)^2)$).

Synthesised

Source

Synthesised



Source

Application: Super-Resolution

x

F

y

- Best Linear Method: Least Squares estimate (linear interpolation):

$$\hat{y} = (\widehat{\Sigma}_x^{\dagger}\widehat{\Sigma}_{xy})x$$

Application: Super-Resolution

x

F

y

- Best Linear Method: Least Squares estimate (linear interpolation):

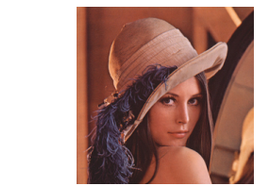
$$\hat{y} = (\hat{\Sigma}_x^{\dagger}\hat{\Sigma}_{xy})x$$

- State-of-the-art Methods:
  - Dictionary-learning Super-Resolution
  - CNN-based: Just train a CNN to regress from low-res to high-res.
  - They optimize cleverly a fundamentally unstable metric criterion:

$$\Theta^* = \arg\min_{\Theta} \sum_i \|F(x_i, \Theta) - y_i\|^2 \quad , \; \hat{y} = F(x, \Theta^*)$$

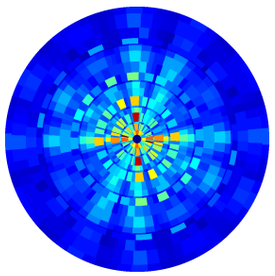
Scattering Approach

- Relax the metric:



x

S



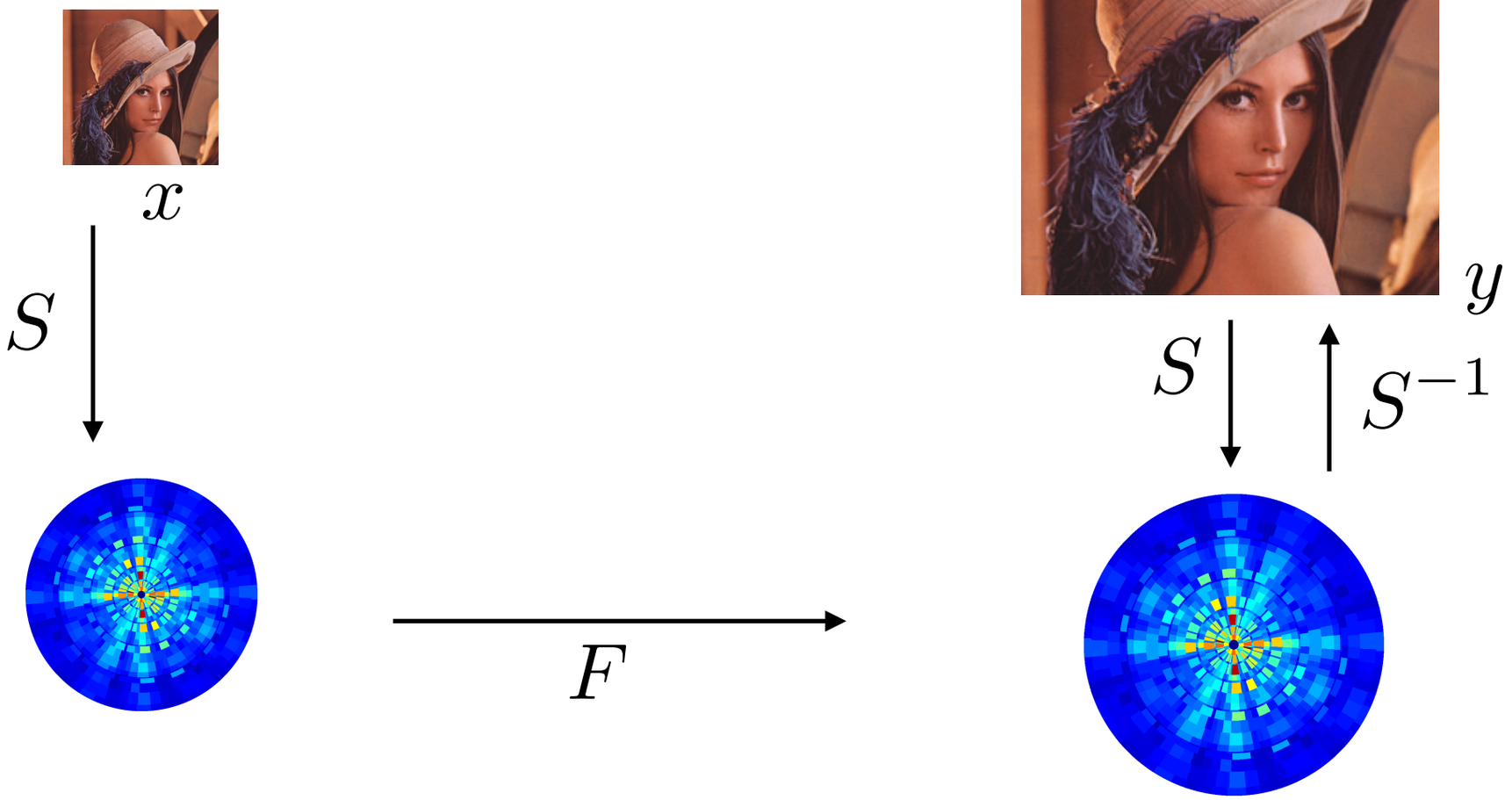
F

y

S S^{-1}

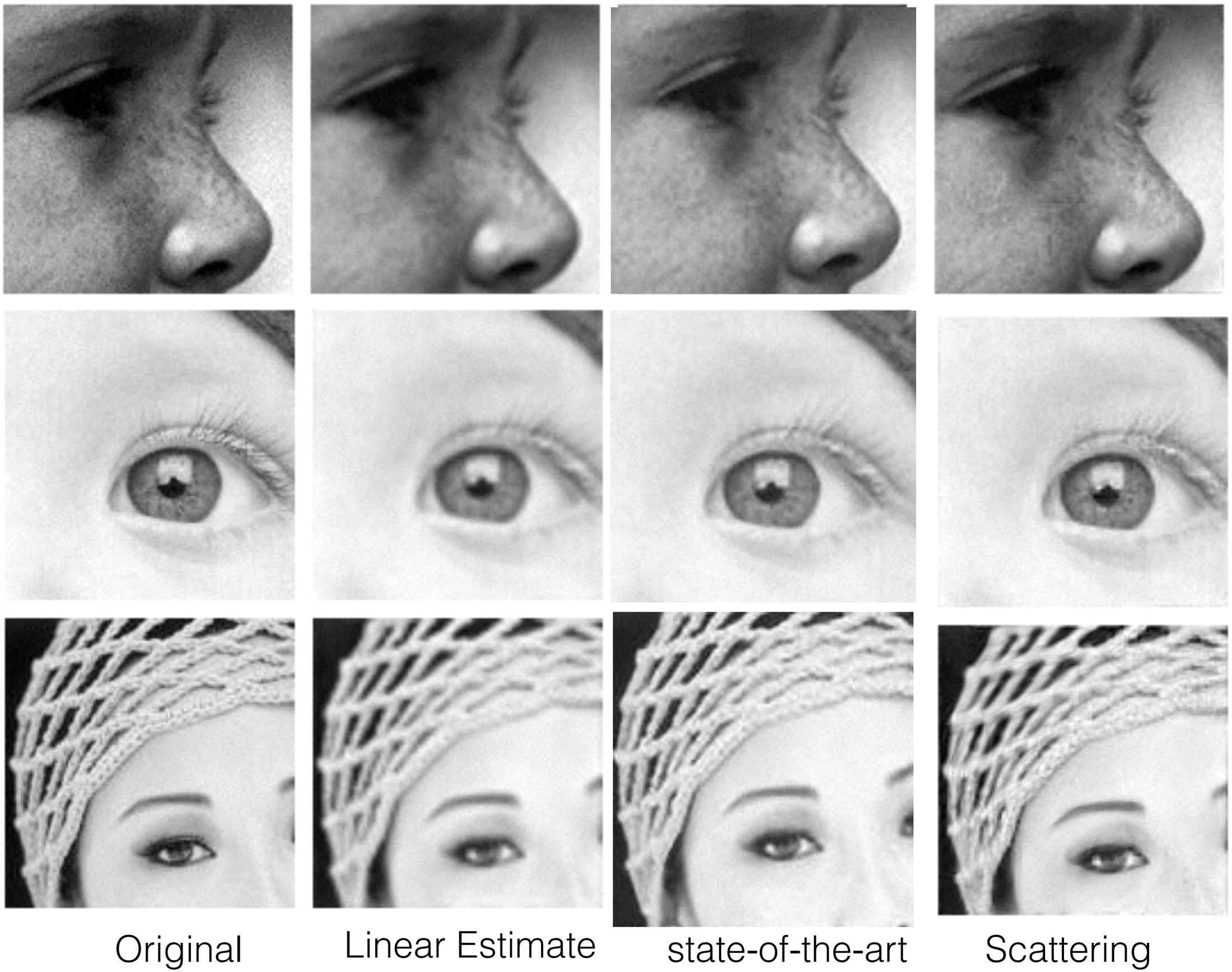
Scattering Approach

- Relax the metric:



  - Start with simple linear estimation on scattering domain.
  - Deformation stability gives more approximation power in the transformed domain via locally linear methods.
  - The method is not necessarily better in terms of PSNR!

Some Numerical Results



Original

Linear Estimate

state-of-the-art

Scattering

Some Numerical Results

Original

Best
Linear Estimate

state-of-the-art

Scattering
Estimate

Some Numerical Results

Original

Best
Linear Estimate

state-of-the-art

Scattering
Estimate

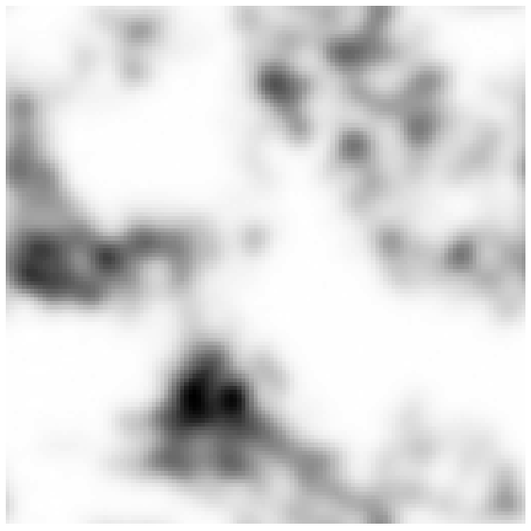
Sparse Spike Super-Resolution

(with I. Domanic (ENS/UIUC), S. Mallat)

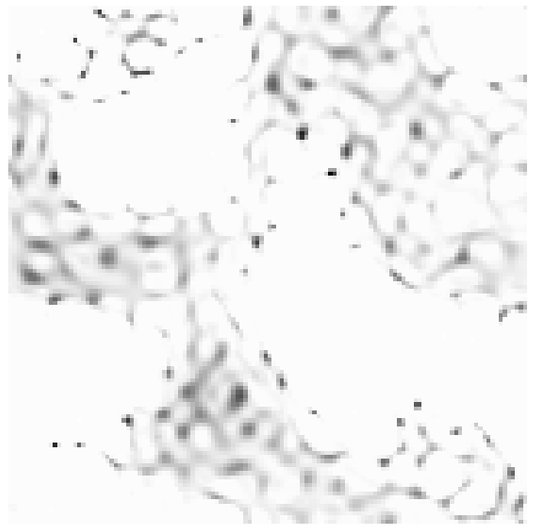
Examples with Cox Processes (inhomogeneous Poisson point processes)

Cox realization

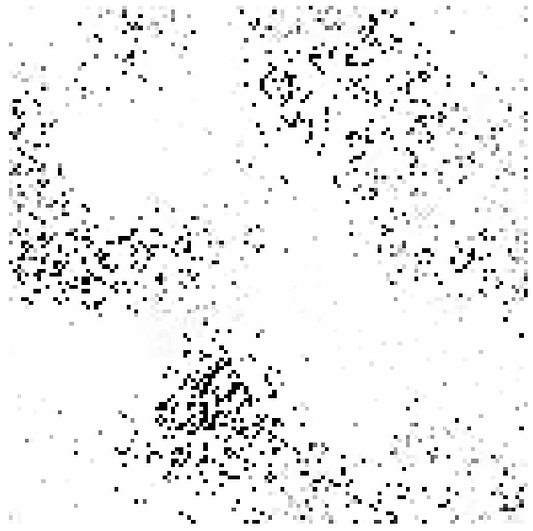
Measurements



ℓ_1 minimization

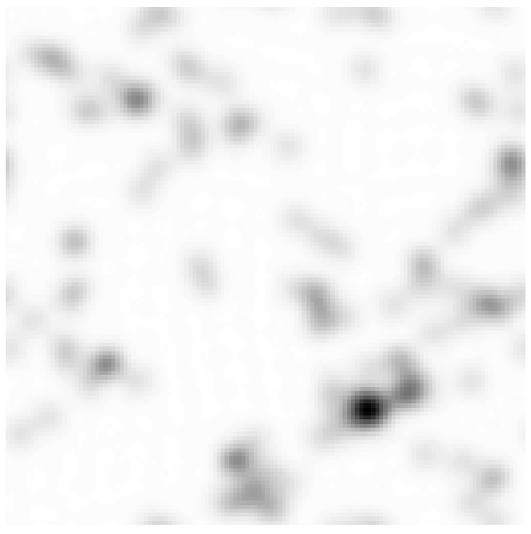


Scattering reconstruction



Cox realization

Measurements



ℓ_1 minimization

Scattering reconstruction

Conclusions and Open Problems

- CNNs: Geometric encoding with built-in deformation stability.
  - Equipped to break curse of dimensionality.
- This statistical advantage is useful both in supervised and unsupervised learning.
  - Gibbs CNN distributions are stable to deformations.
  - Exploited in high-dimensional inverse problems.
- Challenges Ahead:
  - Decode geometry learnt by CNNs: role of higher layers?
  - CNNs and unsupervised learning: we need better inference.
  - Optimization in CNNs: exploit layerwise/convolutional model.
  - Non-Euclidean domains (text, genomics, n-body dynamical systems)

Thank you!