Estimation of High-Dimensional Densities

Joan Bruna, Stéphane Mallat,

École Normale Supérieure

High-Dimensional Density Estimation

- Estimation $\tilde{p}(x)$ of a probability density $p(x)$ for $x \in \mathbb{R}^d$ given n realizations $\{x_i\}_{i \leq n}$ of a random vector X.

- $p(x)$ is the space $\mathbf{C^1}(\mathbb{R}^d)$ of Lipschitz functions if

$$\text{then at best } \mathbb{E}(\|p - \tilde{p}\|_2^2) = O(n^{\frac{-2}{d+2}})$$

- If $d > 10$ then n must be huge: impossible.

$d = 10^6$

$Turbulence \;\; x(u)$

Problem:

Find regularity properties which can break the curse of dimensionality.

Markov Hypothesis

- Markov hypothesis: local conditional dependence

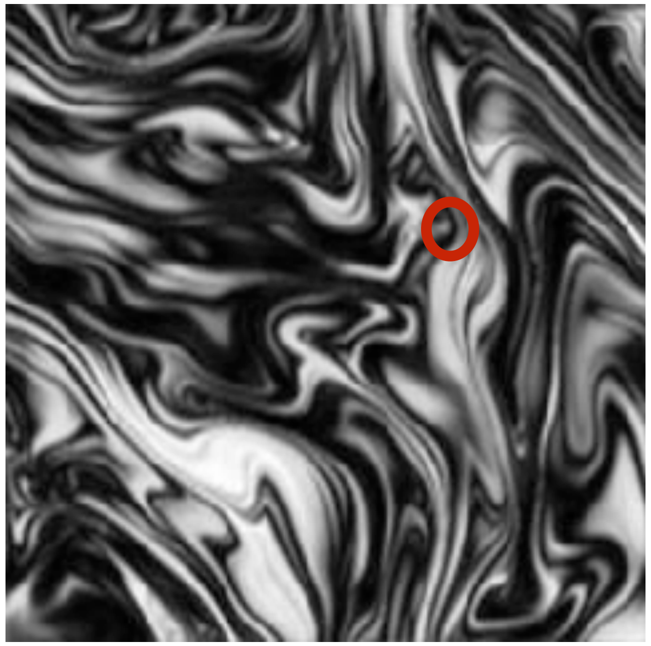
$$p\Big(x(u) \Big/ x(u')\,,\; u' \neq u\Big) = p\Big(x(u) \Big/ x(u')\,,\; u' \in \mathcal{N}_u\Big)$$

- Hammersely-Clifford theorem proves that

$$\log p(x) = \beta_0 + \sum_{k=1}^{K} \phi_k(x(u)\,,\; u \in C_k)$$

separation over small cliques of neighbour variables
of conditionnaly independent components.

- Problem: Markov hypothesis often not valid



Gibbs Distributions

Approximation of $p(x)$ conditioned on K moments $\mathbb{E}_p(\phi_k(x))$

by $\tilde{p}$ which maximizes the entropy $H_{\tilde{p}} = -\int \tilde{p}(x) \log \tilde{p}(x)\, dx$

Theorem [*Canonical Gibbs*] If $\tilde{p}(x)$ satisfies

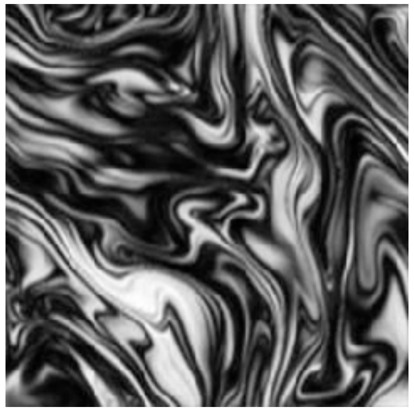
$$\forall k \leq K \quad , \quad \mathbb{E}_{\tilde{p}}(\phi_k(x)) = \int_{\mathbb{R}^N} \phi_m(x)\ \tilde{p}(x)\, dx = \mathbb{E}_p(\phi_k(x))$$

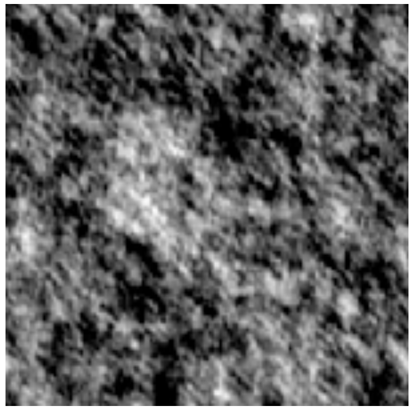
and maximizes $H_{\tilde{p}} = -\int \tilde{p}(x) \log \tilde{p}(x)\, dx$ then

$$\log \tilde{p}(x) = \beta_0 + \sum_{k=1}^{K} \beta_k\, \phi_k(x) \qquad \text{(separation)}$$

Kolmogorov

Gaussian





Problems:

- How to choose the ϕ_k to approximate p ?
- Can not compute the β_k.

  $\phi_k(x)$ are quadratic $\Leftrightarrow \tilde{p}$ is Gaussian

Key Ideas

We want $\log p(x) \approx \log \tilde{p}(x) = \beta_0 + \sum_{k=1}^{K} \beta_k \, \phi_k(x)$: separation

$\Rightarrow$ the regularity of the ϕ_k is defined by the regularity of p

- Regularity of $p(x)$ defined by diffeomorphism groups acting on x

- Separations are scale separations (not Markov) $\Rightarrow$ wavelets

- $H_{\tilde{p}} \geq H_p$ and if $H_{\tilde{p}} = H_p$ then $\tilde{p} = p$

  The ϕ_k should minimize the maximum entropy $H_{\tilde{p}}$

  Obtained with sparsity and intersections of $\mathbf{l^1}$ balls

- Approximate the canonical $\tilde{p}$ by a microcanonical distribution

- Implemented by a deep convolutional network

Lipschitz Regularity on a Group

- Group G of operators acting on x with a metric.

- An $f(x)$ is in $\mathbf{C^1}(G)$ of Lipschitz functions for the action of G

$$\forall (g,x) \in G \times \mathbb{R}^d \ , \ |f(x) - f(g.x)| \leq C \, \mathrm{dist}(g, Id)$$

The usual Lipschitz space is $\mathbf{C^1}(\mathbb{R}^\mathbf{d})$: $g.x = x - g$ for $g \in \mathbb{R}^d$.

$\mathrm{dist}(\mathrm{d}, \mathrm{Id}) = \|g\|$

- Lipschitz continuity to spatial diffeomorphisms: deformations

Images $x(u) \in \mathbf{L^2}(\mathbb{R}^2)$ $\quad g.x(u) = x(g(u))$ for $g \in \mathrm{Diff}(\mathbb{R}^2)$

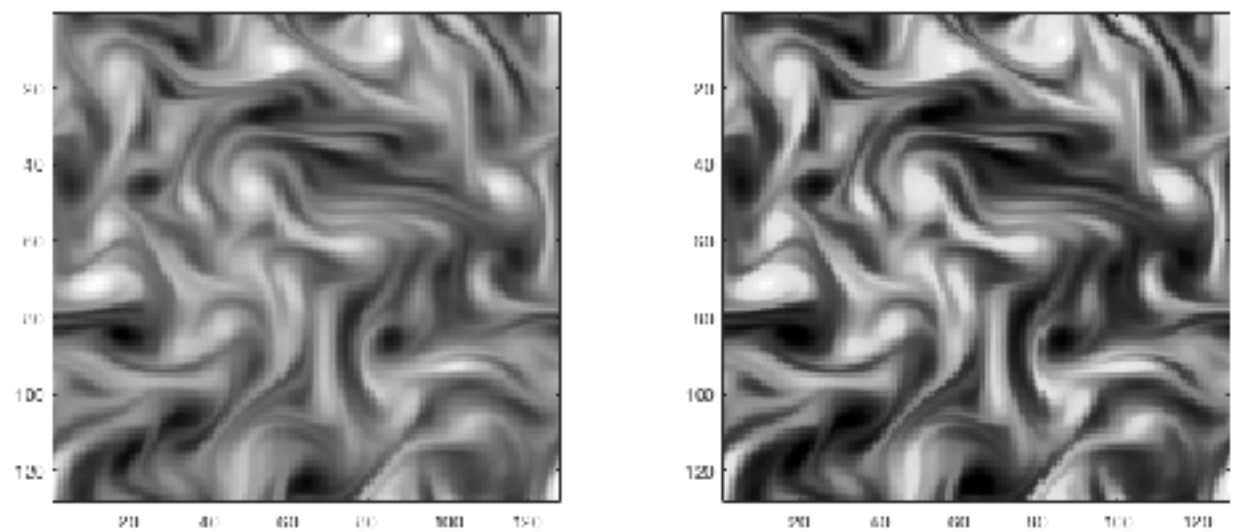
Weak topology: $\mathrm{dist}(g, Id) = \|\nabla g\|_\infty$

$\Rightarrow \quad |f(x) - f(g.x)| \leq C \, \|\nabla g\|_\infty \quad \Rightarrow$ translation invariance

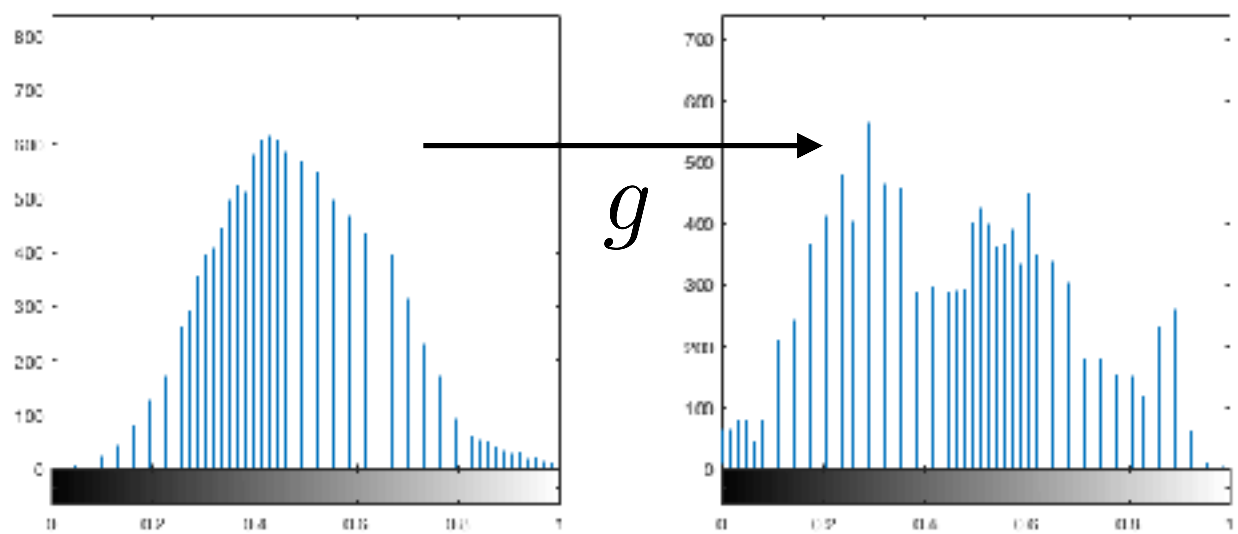
Amplitude Diffeomorphisms

- Amplitude deformation of $x(u) \in \mathbf{L^2}(\mathbb{R}^2)$ with $g \in \mathrm{Diff}(\mathbb{R})$

$$g.x(u) = g(x(u))$$



histogram of $x(u)$ histogram of $g.x(u)$

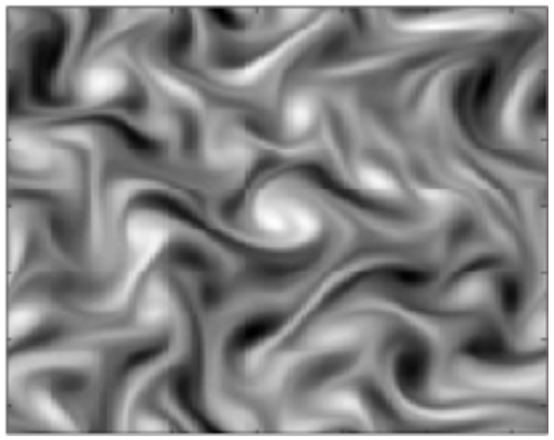


Amplitude-Space Deformations

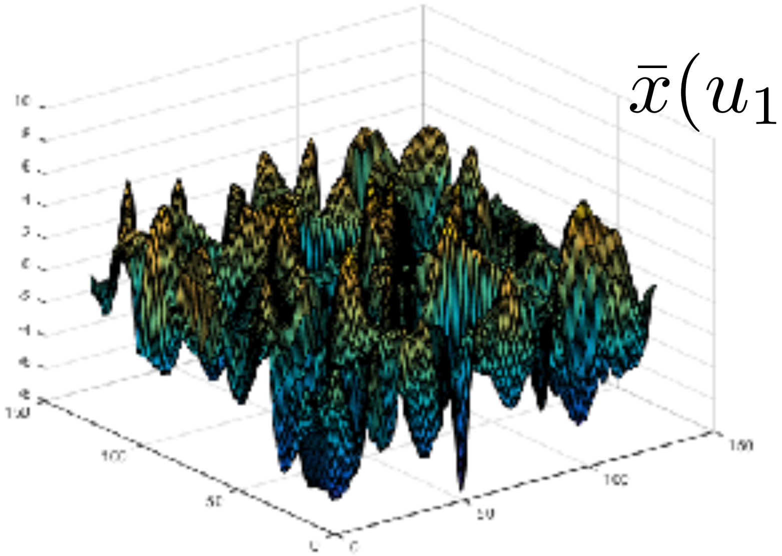
- The action of $\mathrm{Diff}(\mathbb{R}^3)$ on x deforms the $3D$ measure

$$\bar{x}(u_1, u_2, u_3) = \delta(u_3 - x(u_1, u_2))$$

$x(u_1, u_2)$



$\Longleftrightarrow$

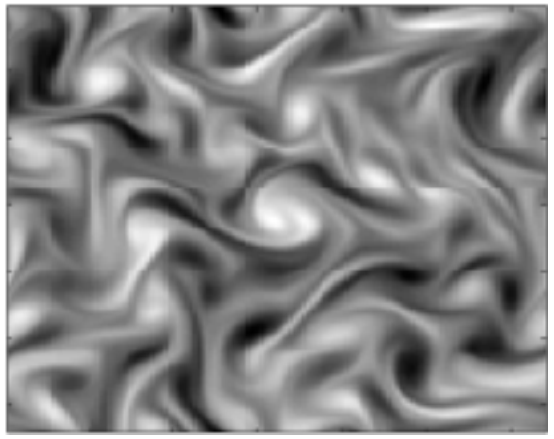
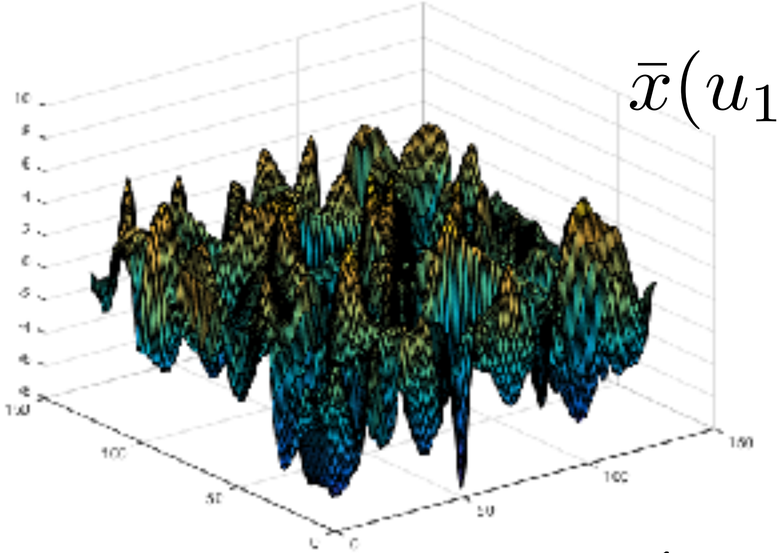


$\bar{x}(u_1, u_2, u_3)$

Amplitude-Space Deformations

- The action of Diff($\mathbb{R}^3$) on x deforms the $3D$ measure

$$\bar{x}(u_1, u_2, u_3) = \delta(u_3 - x(u_1, u_2))$$

$x(u_1, u_2)$  $\Longleftrightarrow$  $\bar{x}(u_1, u_2, u_3)$

- Image classification functions are typically in $\mathbf{C^1}(\mathrm{Diff}(\mathbb{R}^3))$

Video of Philipp Scott Johnson

Lipschitz Approximations

- We want to approximate $\log p$ in $\mathbf{C^1}(\mathrm{Diff}(\mathbb{R}^3))$ with

$$\log \tilde{p}(x) = \sum_{k=0}^{K-1} \beta_k \phi_k(x) = \langle \Phi(x), \beta \rangle$$

$\log \tilde{p} \in \mathbf{C^1}(\mathrm{Diff})$ if Φ is in $\mathbf{C^1}(\mathrm{Diff}(\mathbb{R}^3))^{\mathrm{K}}$ with

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

How can we build such Φ ?

Marginal Distributions

Cramer-Wold theorem

- A stationary density p of X is characterised by the $1D$ marginals of $X \star \psi_\alpha(u)$ for all $\psi_\alpha \in \mathbb{R}^d$

$\Rightarrow$ choose a "large" family of $\{\psi_\alpha\}_\alpha$ *Mumford, Zhu*

estimate the distribution of $X \star \psi_\alpha(u)$ with a histogram

$\tilde{p}$: maximum entropy conditioned to these histogram values

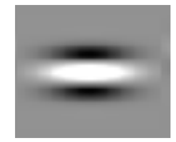
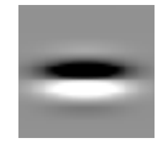
A bit too optimistic...

- To approximate $\log p$ is in $\underbrace{\mathbf{C^1}(\mathrm{Diff}(\mathbb{R}^2))}^{\text{spatial deformations}}$ we need that

$$\forall \alpha \ , \quad \|u \,.\, \nabla \psi_\alpha(u)\|_1 \leq C$$

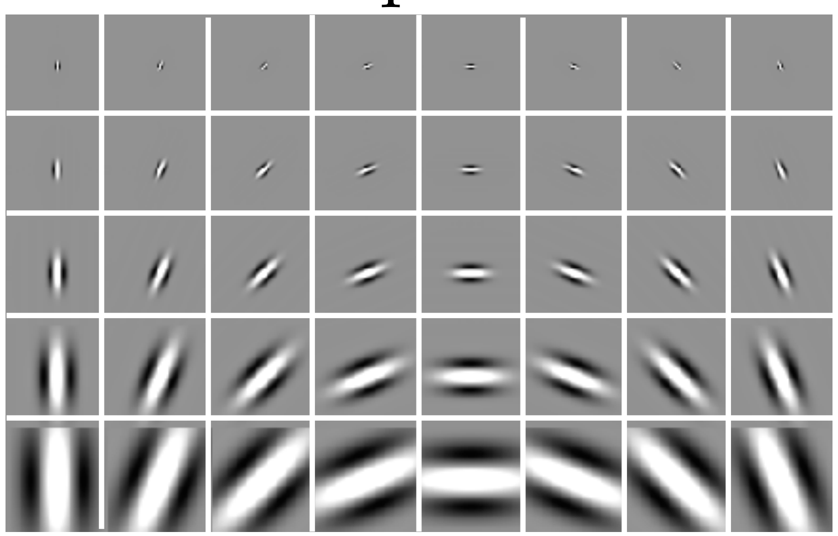
dilated filters: scale separation

Scale separation with Wavelets

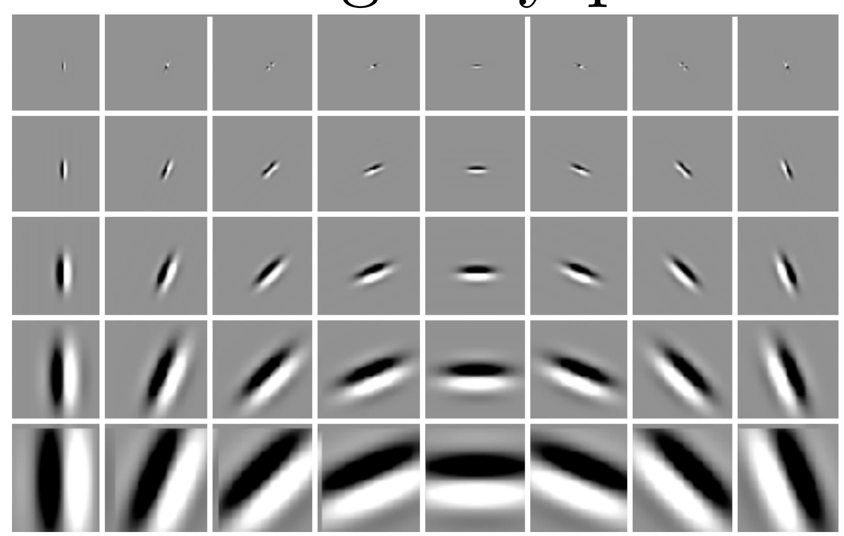
- Wavelet filter $\psi(u)$:  $+\,i$ 

rotated and dilated: $\psi_{2^j,\theta}(u) = 2^{-j}\,\psi(2^{-j} r_\theta u)$

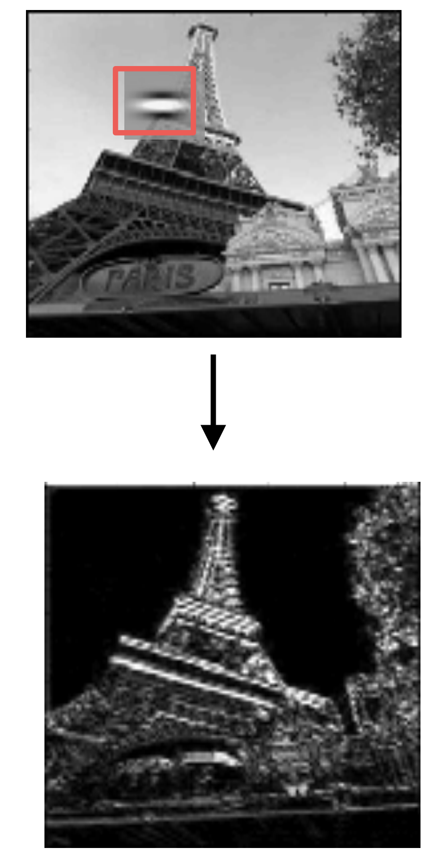
real parts



imaginary parts



$x(u)$

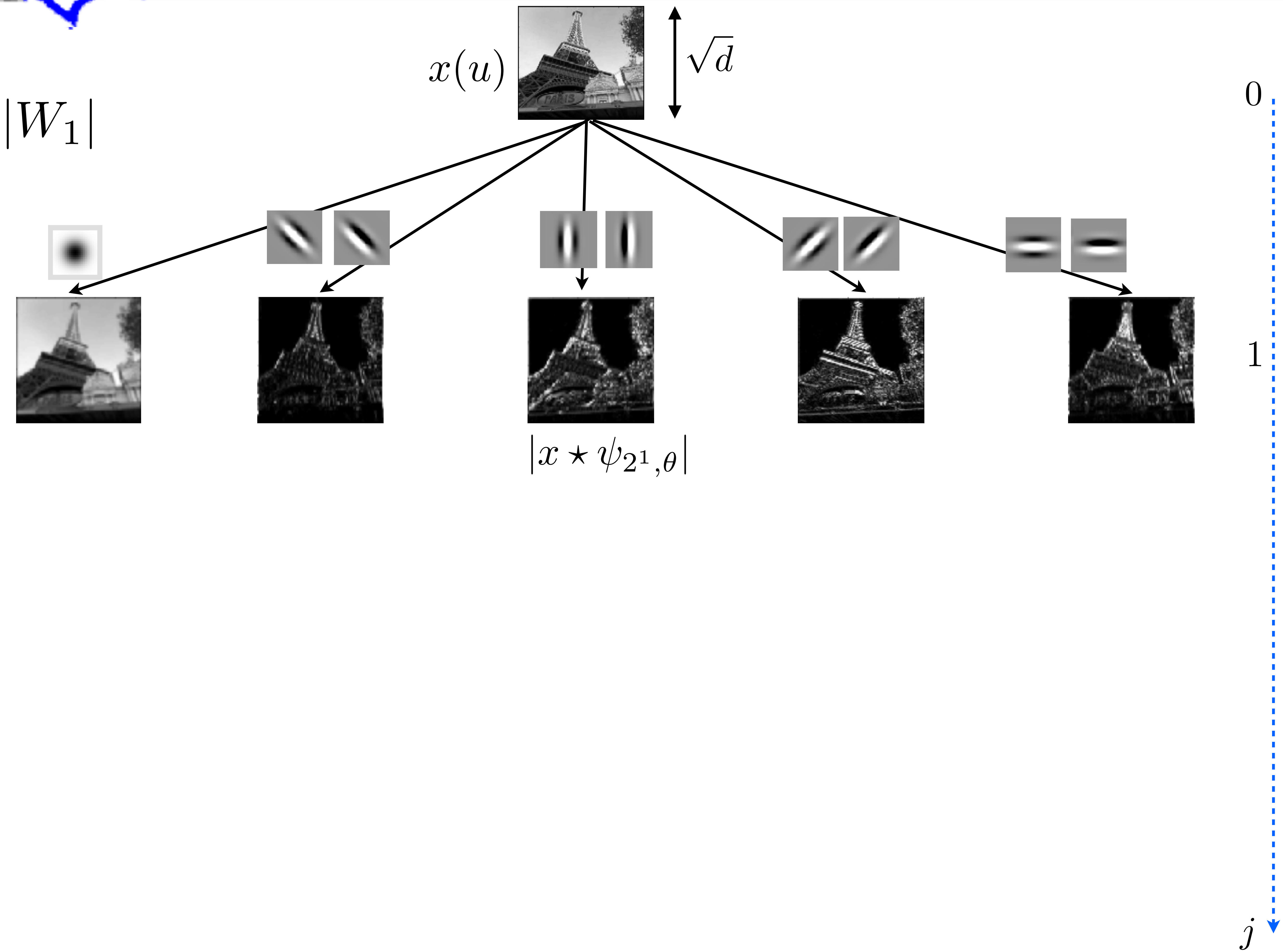


$$x \star \psi_{2^j,\theta}(u) = \int x(v)\,\psi_{2^j,\theta}(u-v)\,dv$$

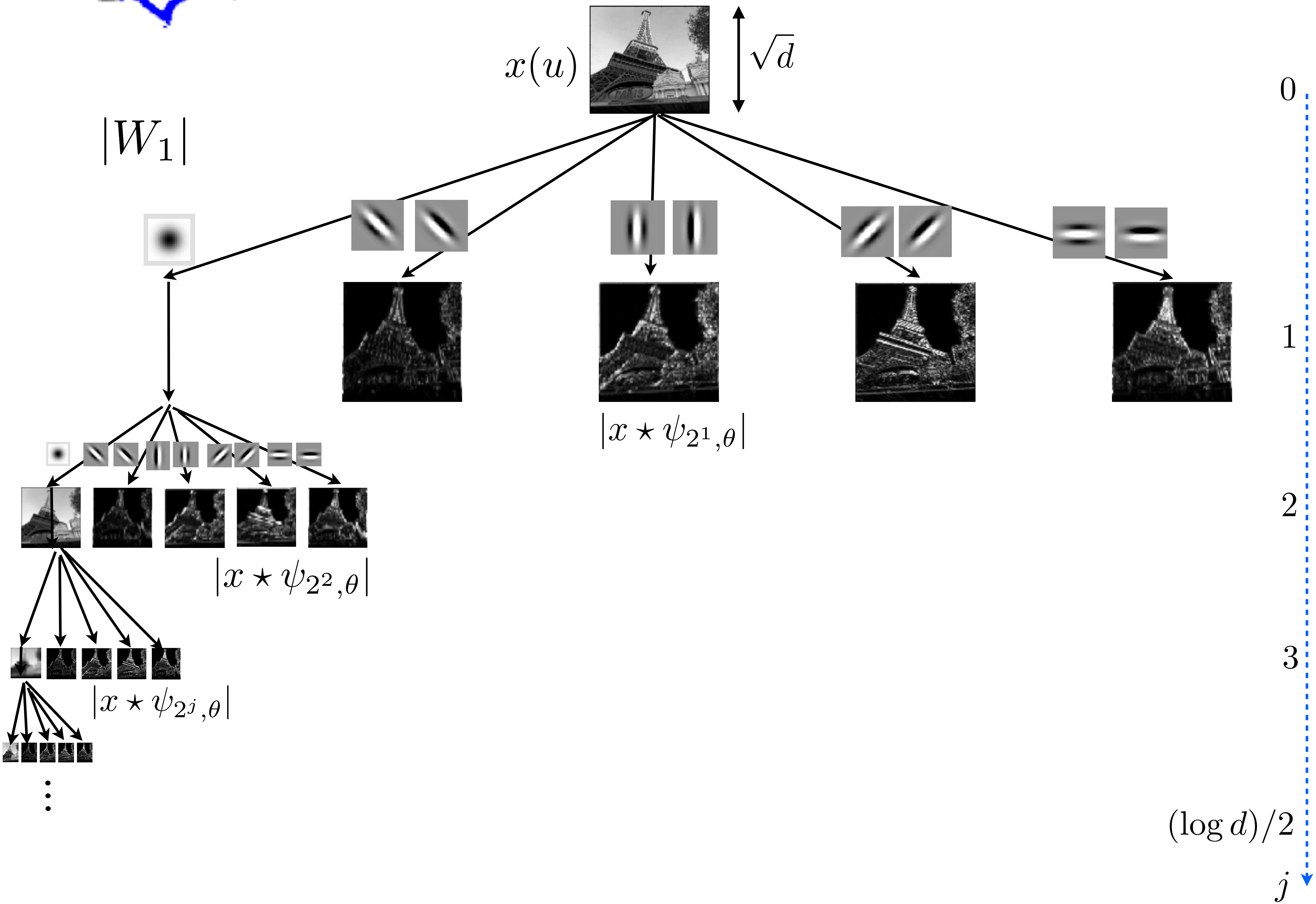
- Wavelet transform: $Wx = \begin{pmatrix} d^{-1}\sum_u x(u) \\ x \star \psi_{2^j,\theta}(u) \end{pmatrix}_{j,\theta}$: average
: higher frequencies

Preserves norm: $\|Wx\|^2 = \|x\|^2$.

Fast Wavelet Filter Bank



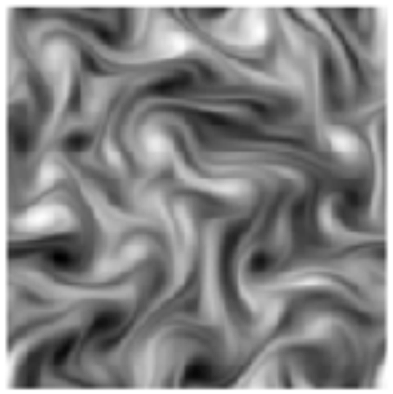
Wavelet Filter Bank

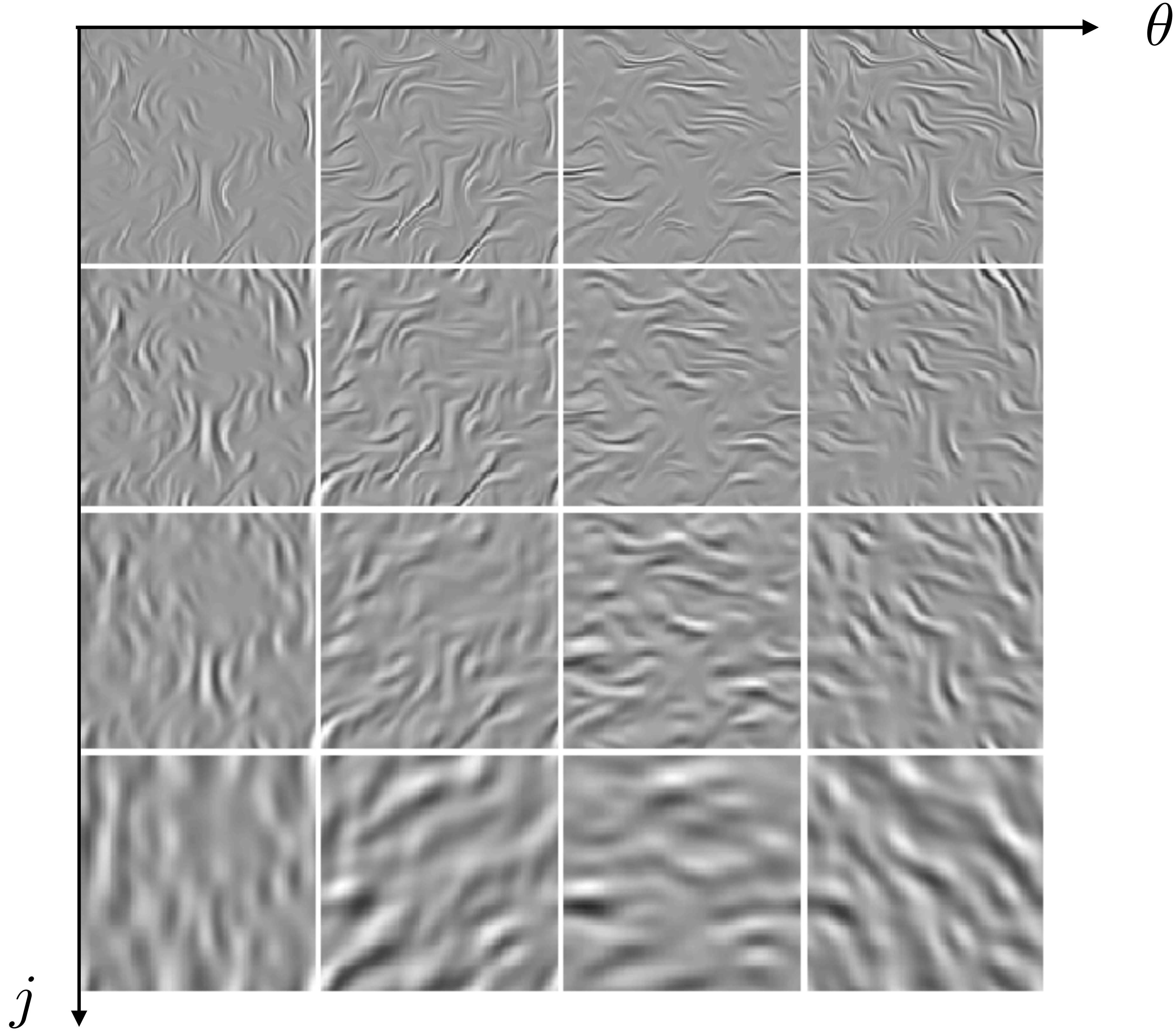


Wavelet transform

$x \star \psi_\lambda$ (real part) : $\lambda = (2^j, \theta)$

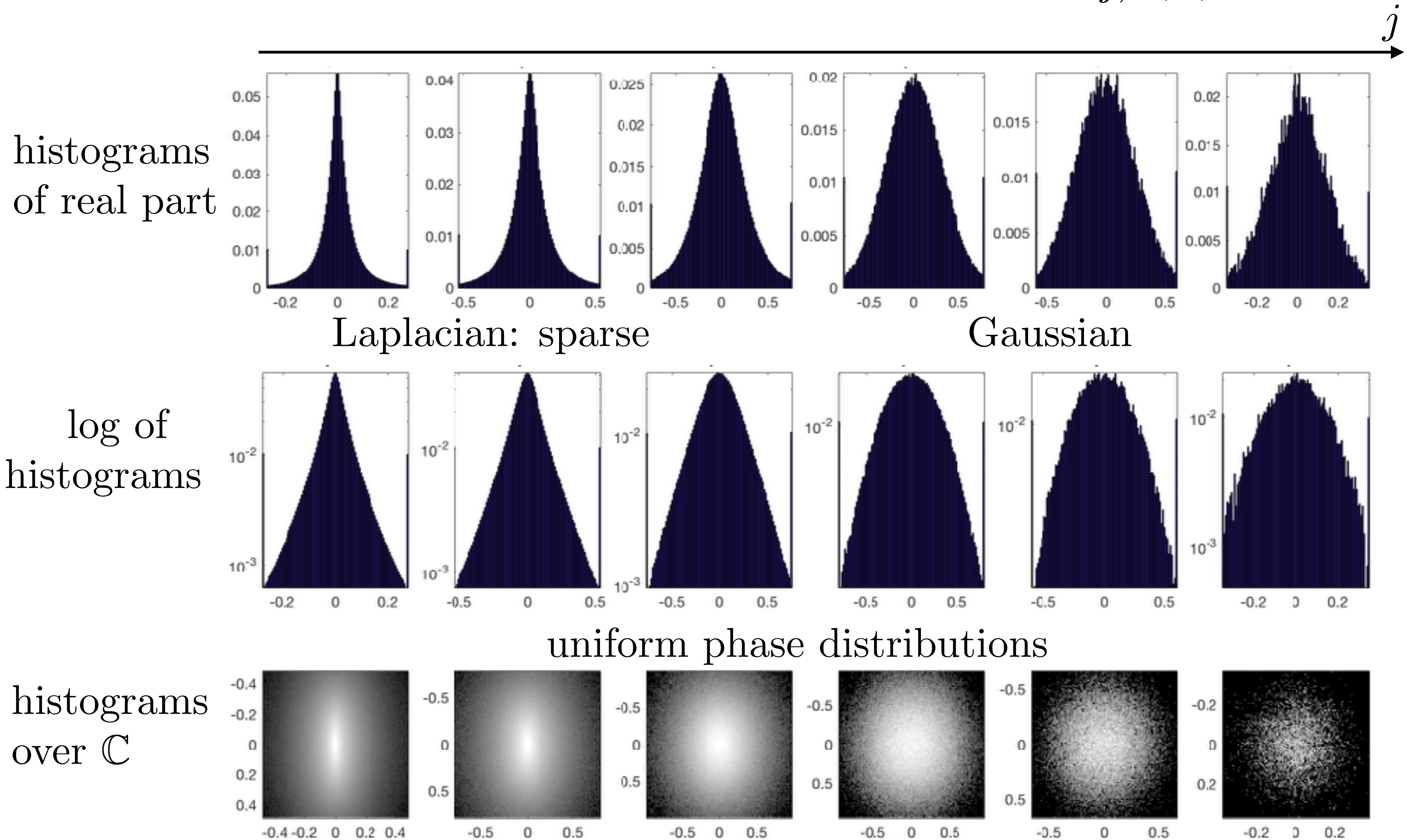
x





Wavelet Transform Marginals

Marginal distribution of wavelet coeffs $X \star \psi_{j,\theta}(u)$



Sparsity and Gaussianization

- If $X \star \psi_\lambda(u)$ has a Laplacian density $\alpha\, e^{-\beta|y|}$ then

$$\|X \star \psi_\lambda\|_1 = \sum_u |X \star \psi_\lambda(u)|$$

is a sufficient statistics of maximum entropy models.

- If $X \star \psi_\lambda(u)$ has a Gaussian density $\alpha\, e^{-\beta|y|^2}$ then

$$\|X \star \psi_\lambda\|_2^2 = \sum_u |X \star \psi_\lambda(u)|^2$$

is a sufficient statistics of maximum entropy models.

Wavelet Maximum Entropy

- Wavelet model

$$\Phi(x) = \Big\{ \sum_u x(u) \quad , \quad \sum_u |x \star \psi_{j,\theta}(u)| \, , \, \sum_u |x \star \psi_{j,\theta}(u)|^2 \Big\}_{(j,\theta)}$$

- Separates scales j and angles θ
- Markovian along u over cliques of size $\sim 2^j$ for each j, θ

- Canonical max entropy distribution conditioned by $\mathbb{E}_p(\Phi(x))$

$$\log \tilde{p}(x) = \langle \Phi(x), \beta \rangle + \beta_0 \ .$$

Problem: computing β is too expensive

$\Rightarrow$ microcanonical approximation of $\tilde{p}$

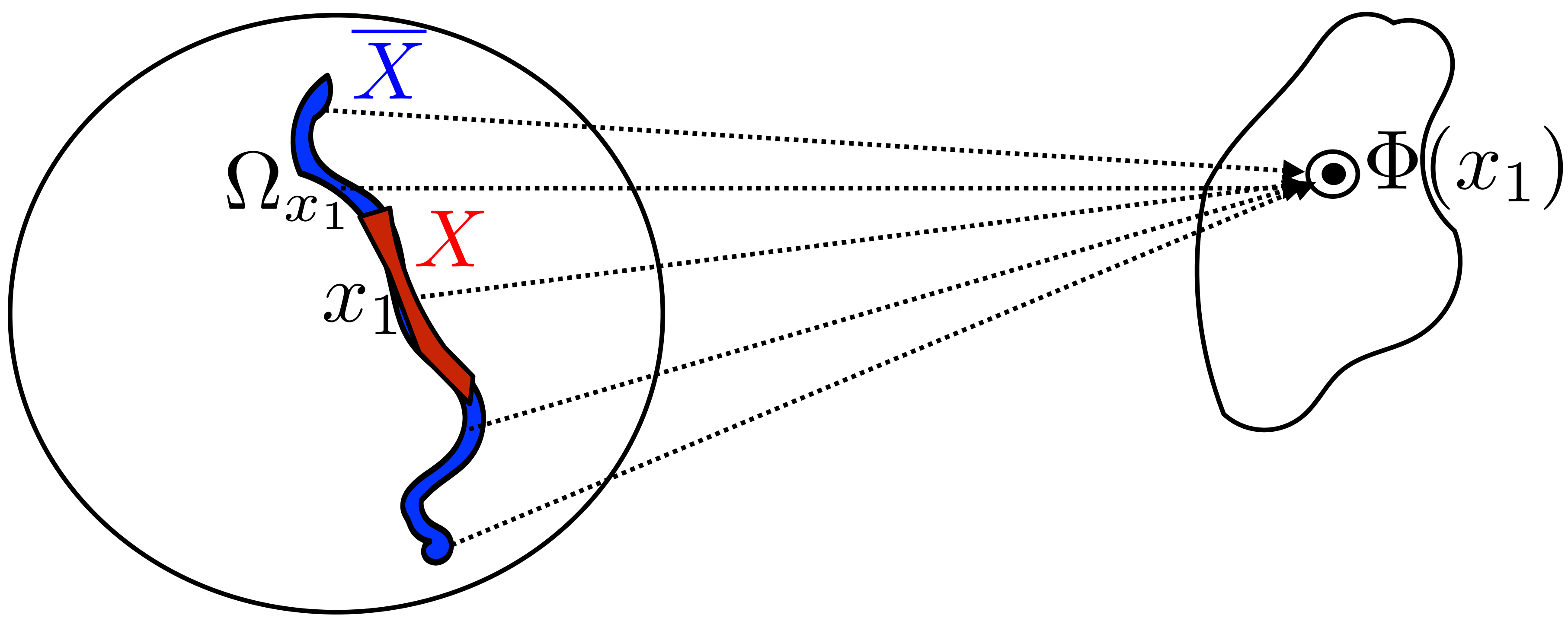
Ergodic Microcanonical Model

Only $n = 1$ realisation x_1 of X is known

Microcanonical set: $\Omega_{x_1} = \{x : \|\Phi x - \Phi x_1\| \leq \epsilon\}$

Microcanonical model $\bar{p}$: maximum entropy supported in Ω_{x_1}

$\Rightarrow$ uniform in Ω_{x_1} if bounded set.



Ergodicity: $\text{Prob}\Big(|\Phi X - \mathbb{E}(\Phi X)| < \epsilon\Big) \underset{d \to \infty}{\longrightarrow} 1 \Rightarrow \Phi x_1 \approx \mathbb{E}(\Phi X)$

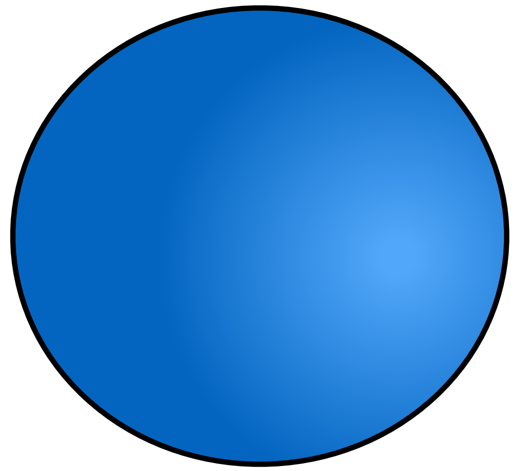
Gibbs conjecture: conditioning on Φx_1 or on $\mathbb{E}(\Phi X)$

converges to the same Gibbs measure when d goes to ∞.

Uniform Distribution on Balls

- Sphere in $\mathbb{R}^d$ $\Phi x = d^{-1}\|x\|_2^2 = d^{-1}\sum_{k=1}^{d}|x(k)|^2$

Ω_x

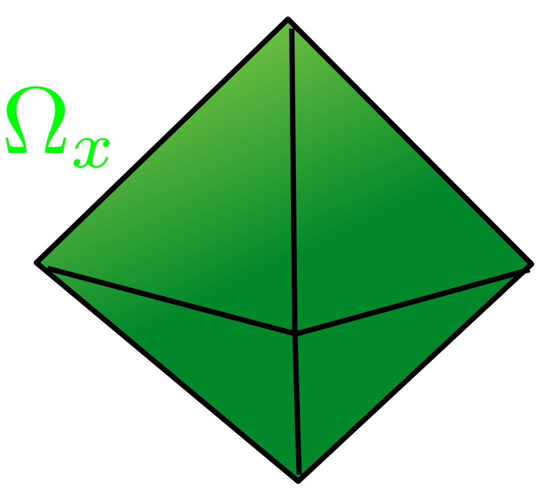


Borel 1914
Diaconis, Freedman 1987

$$\overline{X}(1), ..., \overline{X}(d) \underset{d\to\infty}{\longrightarrow} \text{.i.i.d Gaussian} \quad \sim e^{-u^2/2\sigma^2}$$

- Simplex in $\mathbb{R}^d$ $\Phi x = d^{-1}\|x\|_1 = d^{-1}\sum_{k=1}^{d}|x(k)| = \mu$

Ω_x



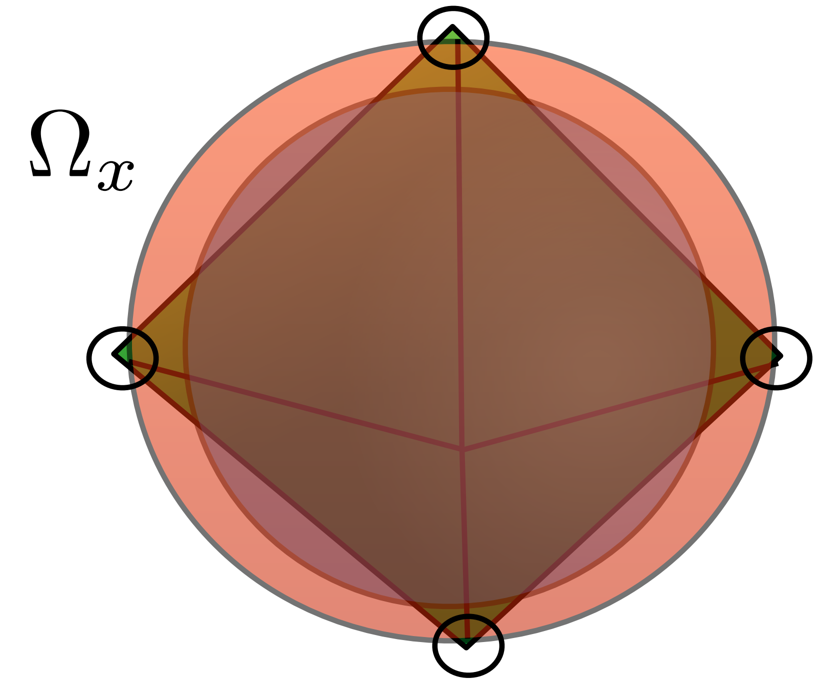
Diaconis, Freedman 1987

$$\overline{X}(1), ..., \overline{X}(d) \underset{d\to\infty}{\longrightarrow} \text{i.i.d Exponential} \quad \sim e^{-\lambda|u|}$$

Intersection of Sphere/Simplex

- Intersection of a Sphere and a Simplex in $\mathbb{R}^d$

$$\Phi x = (\|x\|_1 \, , \, \|x\|_2^2)$$



Chatterjee 2015

- If d goes to ∞ then $\overline{X}(1), ..., \overline{X}(d)$ converges to:

a canonical Gibbs: $e^{-\alpha|x|-\beta|x|^2}$ if $r = \|x\|_2/\|x\|_1 < 2$

- Gaussian if $r = \sqrt{\pi/2}$

- Laplacian if $r = \sqrt{2}$

a singular sparse distribution if $r > 2$

Gibbs Hypothesis

Theorem *(H. Georgii)*

If $\Phi x = \sum_u U x(u)$ where Ux has a bounded range for $u \in \mathbb{Z}^d$

If the macro canonical distribution exists and converges

to a unique Gibbs measure when d goes to ∞

then the microcanonical model converges to the same measure

for a weak topology.

Proof: large deviation principle

- Sample max entropy $\overline{X}$ in Ω_{x_1}: $\|\Phi\overline{X} - \Phi x_1\| \leq \epsilon$

  Algorithm:

  Initialized with X_0 Gaussian white noise

  Iteratively reduce $\|\Phi X_n - \Phi x_1\|^2$ with gradient descent

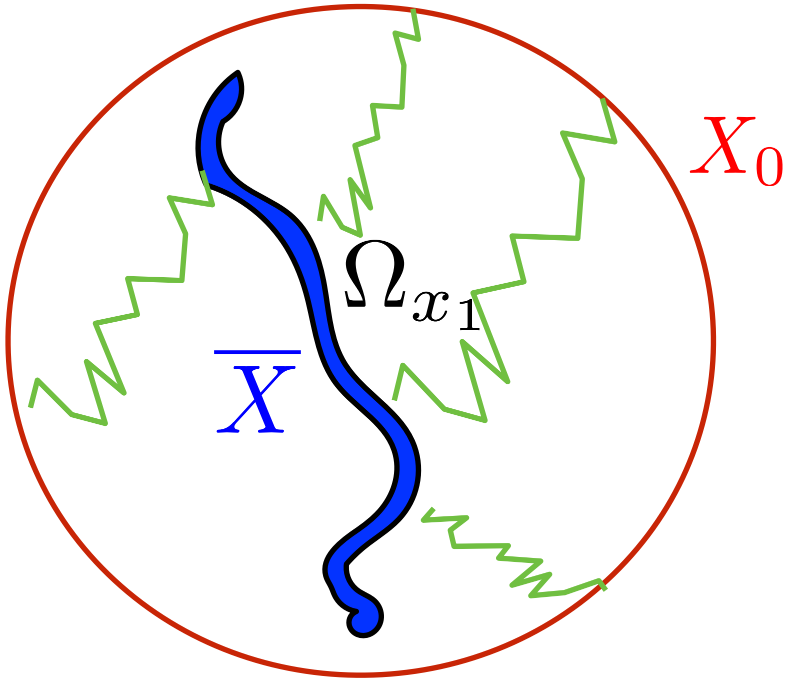
- Proof of convergence to a stationary process X_∞

  The algorithm defines a transport of measure.

Math problems:

- No proof on maximum entropy

- Entropy lower bounds depend upon the Jacobian of Φ...



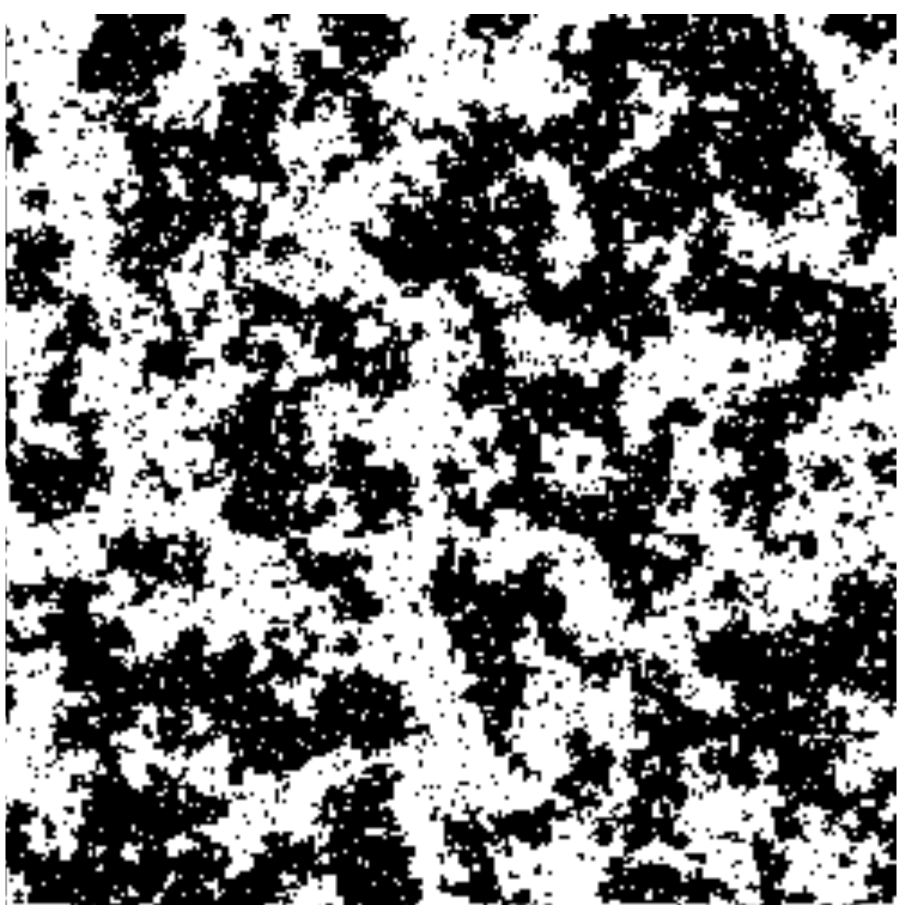
Ising at Critical Temperature

$$x(u) \in \{0,1\} \qquad p(x) = Z^{-1} \exp\Big(\frac{1}{T} \sum_{(u,u') \in C_I} x(u)\, x(u')\Big)$$

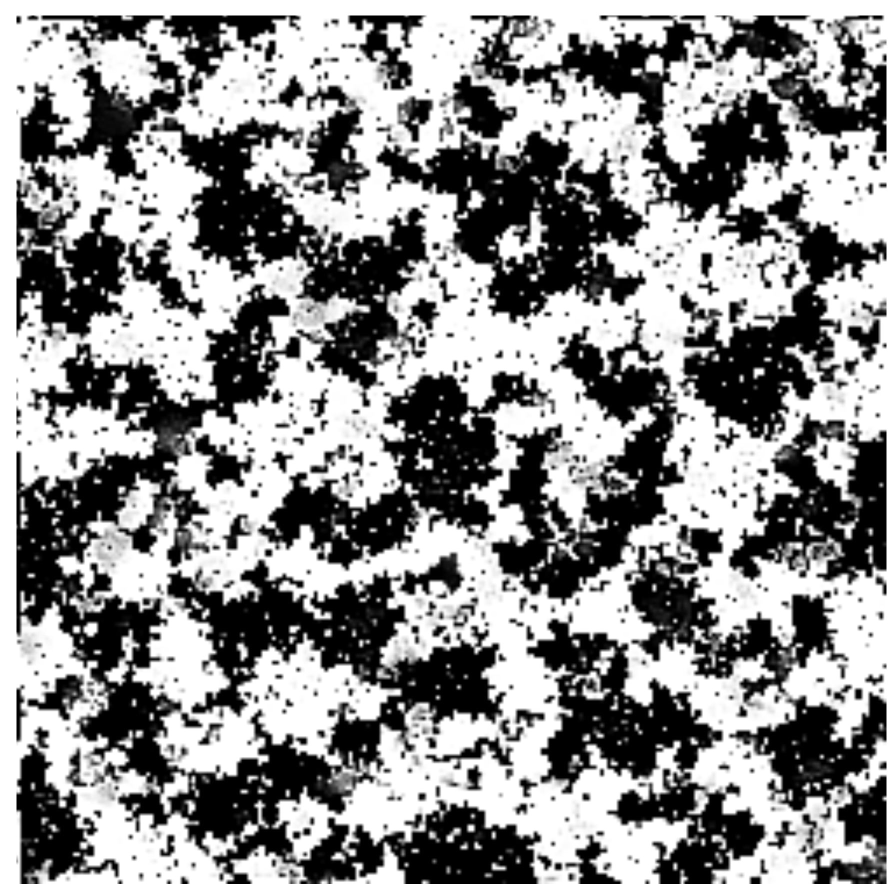
$$\Phi(x) = \Big\{ d^{-1} \sum_u x(u) \quad , \quad \|x \star \psi_\lambda\|_1 \, , \, \|x \star \psi_\lambda\|_2^2 \Big\}_\lambda$$

$T = T_{\text{critic}} + \epsilon$

Realization x_1 of X



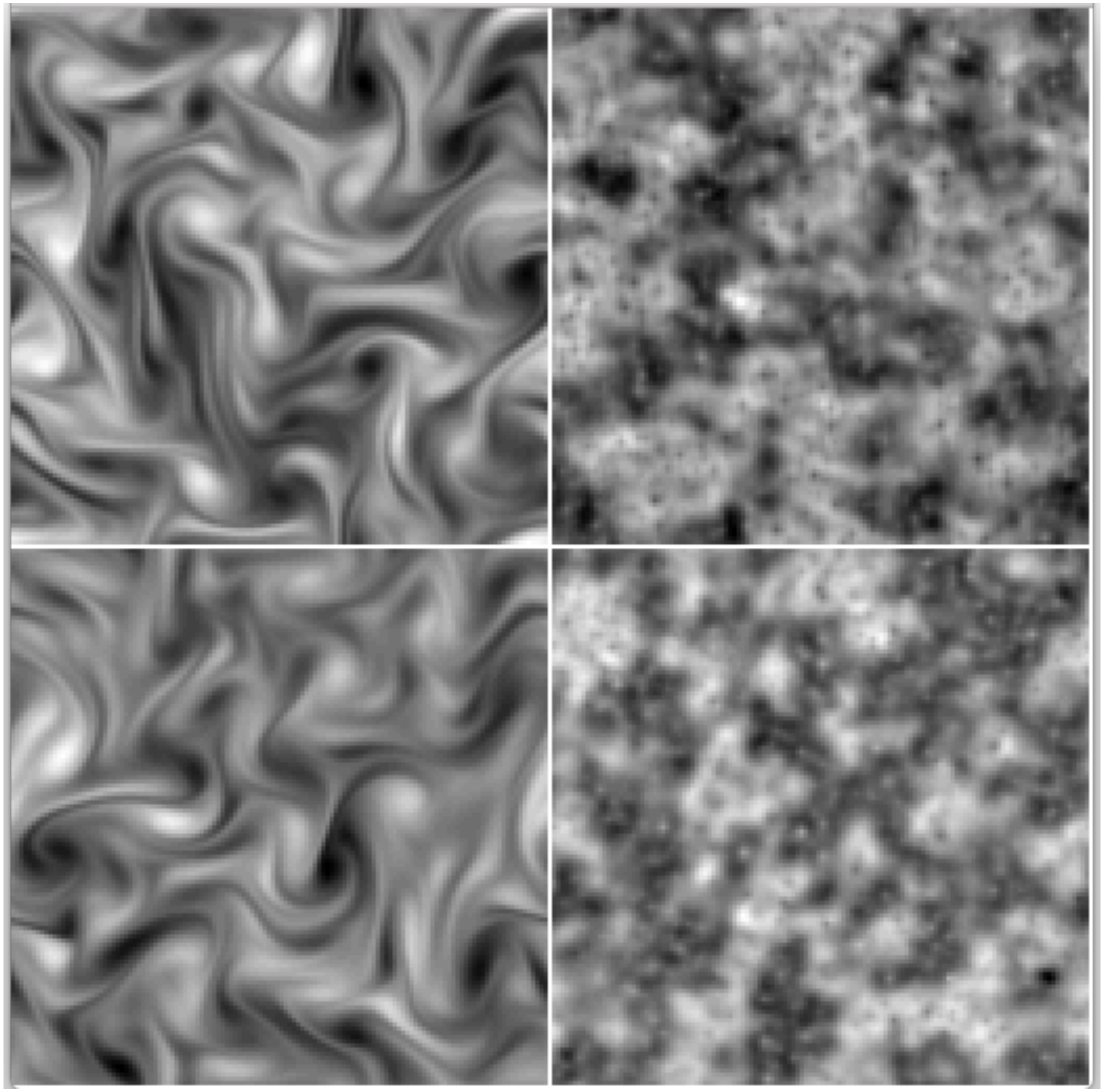
Microcanonical X_∞



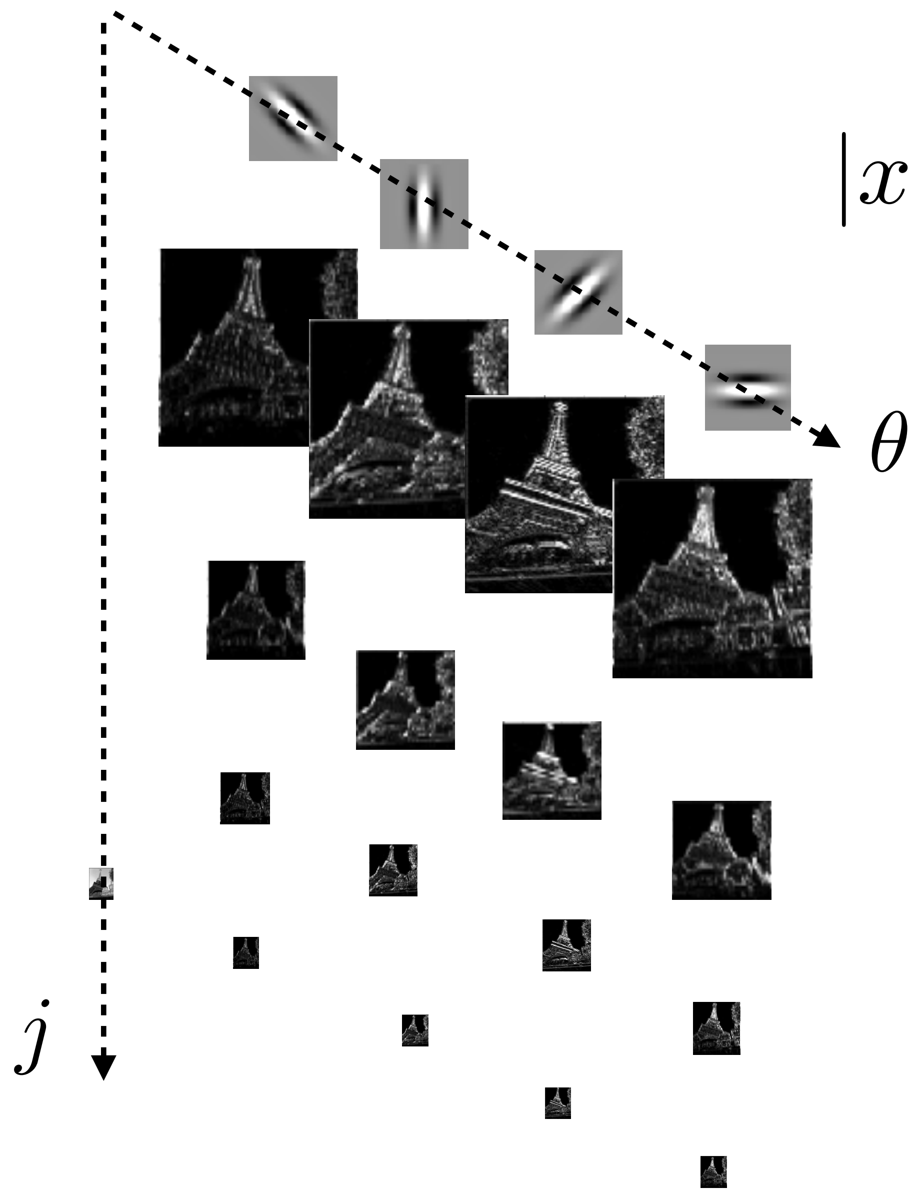
Microcanonical Reconstruction

$$\Phi(x) = \left\{ d^{-1} \sum_u x(u) \quad , \quad \|x \star \psi_\lambda\|_1 \, , \, \|x \star \psi_\lambda\|_2^2 \right\}_\lambda$$

Realization x_1 of X — Microcanonical X_∞



Wavelet Model



$$|x \star \psi_{j,\theta}(u)| = w(u, j, \theta)$$

$$\Phi(x) = \Big\{ \sum_u x(u) \quad , \quad \sum_u |x \star \psi_{j,\theta}(u)| \, , \, \sum_u |x \star \psi_{j,\theta}(u)|^2 \Big\}_{(j,\theta)}$$

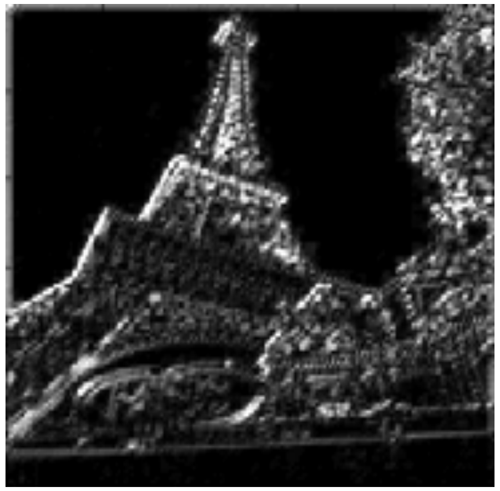
"Conditional independence" may be violated along u , θ, j

Higher Order Wavelet Coefficients

Loss of information:

$$\|x \star \psi_{\lambda_1}\|_1 = \sum_u |x \star \psi_{\lambda_1}(u)|$$

eliminates all variations of $|x \star \psi_{\lambda_1}(u)|$ along u

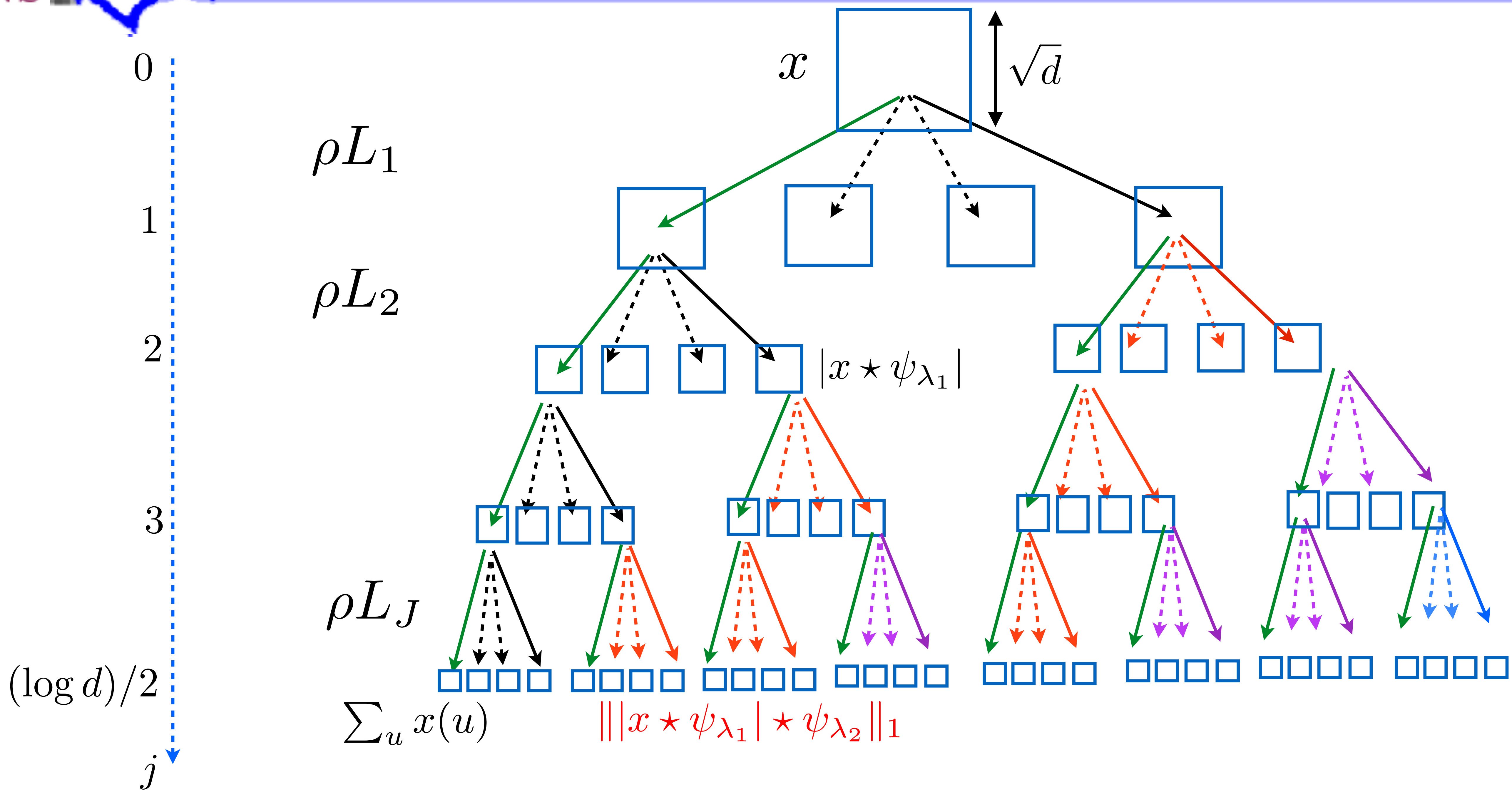


Lipschitz to diffeomorphisms:

recover them as wavelet coefficients of $|x \star \psi_{\lambda_1}(u)|$

$$|W_2|\ |x \star \psi_{\lambda_1}| = \begin{pmatrix} \sum_u |x \star \psi_{\lambda_1}(u)| \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(u)| \end{pmatrix}_{\lambda_2}$$

Wavelet Scattering Network



$$\Phi = |W_{\log d/2}| \cdots |W_2|\,|W_1|$$

$$\Phi x = \Big\{ \| | | x \star \psi_{\lambda_1} | \star \psi_{\lambda_2} \star \ldots | \star \psi_{\lambda_m} \|_1 \Big\}_{\lambda_k}$$

Scattering Properties

$$\Phi x = \begin{pmatrix} \sum_u x(u) \\ \|x \star \psi_{\lambda_1}\|_1 \\ \||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}\|_1 \\ \|||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}\|_1 \\ \cdots \end{pmatrix}_{\lambda_1,\lambda_2,\lambda_3,\ldots} = \ldots |W_3|\,|W_2|\,|W_1|\,x$$

$$\|W_k x\| = \|x\| \quad \Rightarrow \quad \||W_k x| - |W_k x'|\| \leq \|x - x'\|$$

Lemma: If $g \in \mathrm{Diff}(\mathbb{R}^2)$ then

$$\|[W_k, g]\| = \|W_k g - g W_k\| \leq C\, \|\nabla g\|_\infty$$

Theorem: *For appropriate wavelets, a scattering is*

contractive $\|\Phi x - \Phi y\| \leq \|x - y\|$: *in* $\mathbf{C^1}(\mathbf{L^2}(\mathbb{R}^2))$

preserves norms $\|\Phi x\| = \|x\|$

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

Energy conservation

$$\|x\| = \|\Phi x\| \quad \Rightarrow \quad \|x \star \psi_{\lambda_1}\|_2^2 = \|\Phi\, |x \star \psi_{\lambda_1}|\|^2$$

$$\|x \star \psi_{\lambda_1}\|_2^2 = \sum_{m=2}^{\infty} \sum_{\lambda_2,\ldots,\lambda_m} \|||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \ldots| \star \psi_{\lambda_m}\|_1^2$$

All $\mathbf{L^2}$ norms are derived from $\mathbf{L^1}$ norms.

Non-negligible $\mathbf{L^1}$ norms appear at order 1 and 2:

$$\Phi(x) = \Big\{ \sum_u x(u) \;\; , \;\; \|x \star \psi_{\lambda_1}\|_1 \, , \, \||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}\|_1 \Big\}_{\lambda 1, \lambda_2}$$

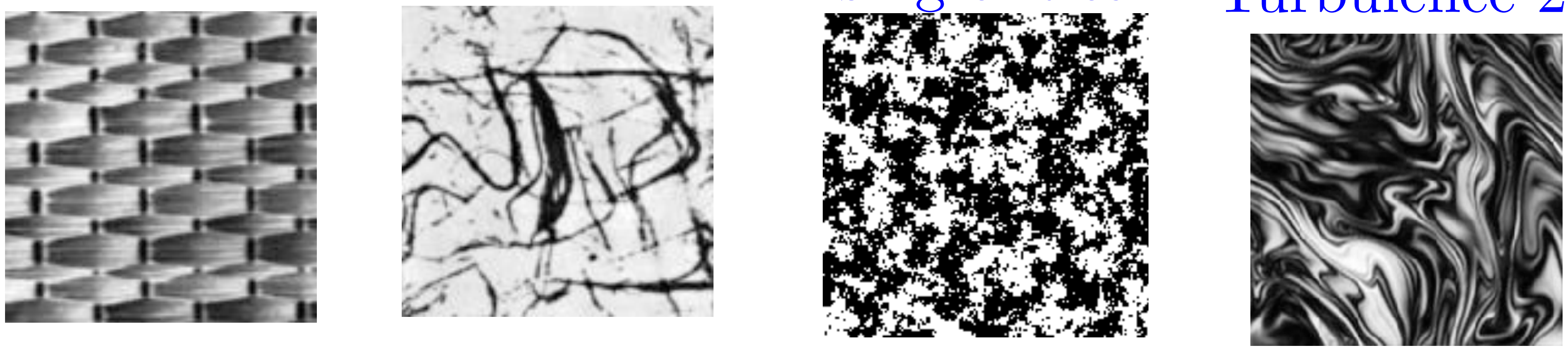
If $x \in \mathbb{R}^d$ then $\Phi x \in \mathbb{R}^{O(\log^2 d)}$

Texture Reconstructions

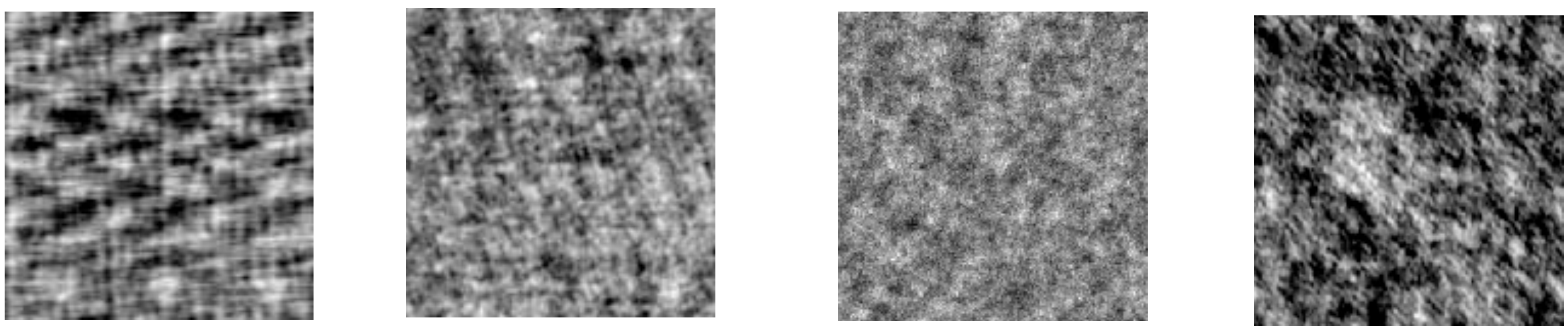
Texture of d pixels

Ising-critical

Turbulence 2D

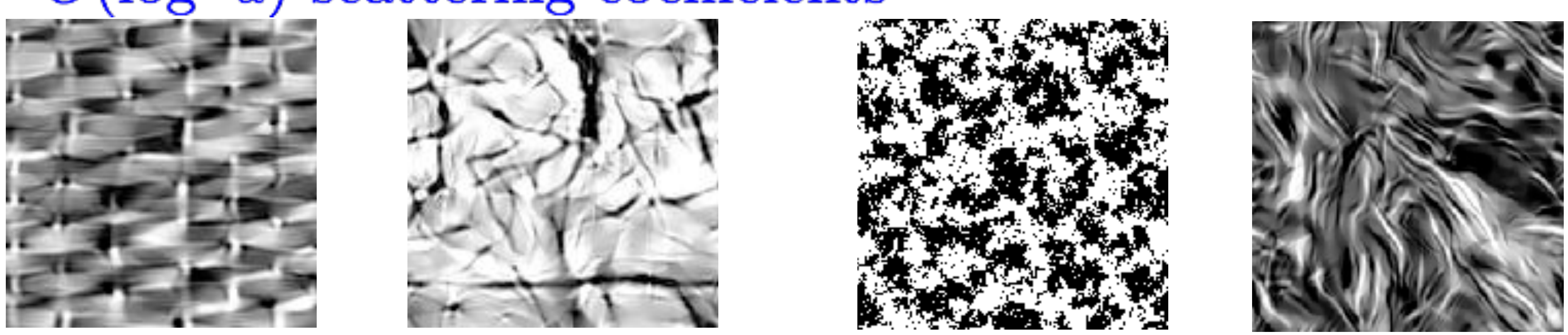


Gaussian process model with d second order moments



Reconstructions from $\|X \star \psi_{\lambda_1}\|_1$ and $\||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}\|_1$

$O(\log^2 d)$ scattering coefficients

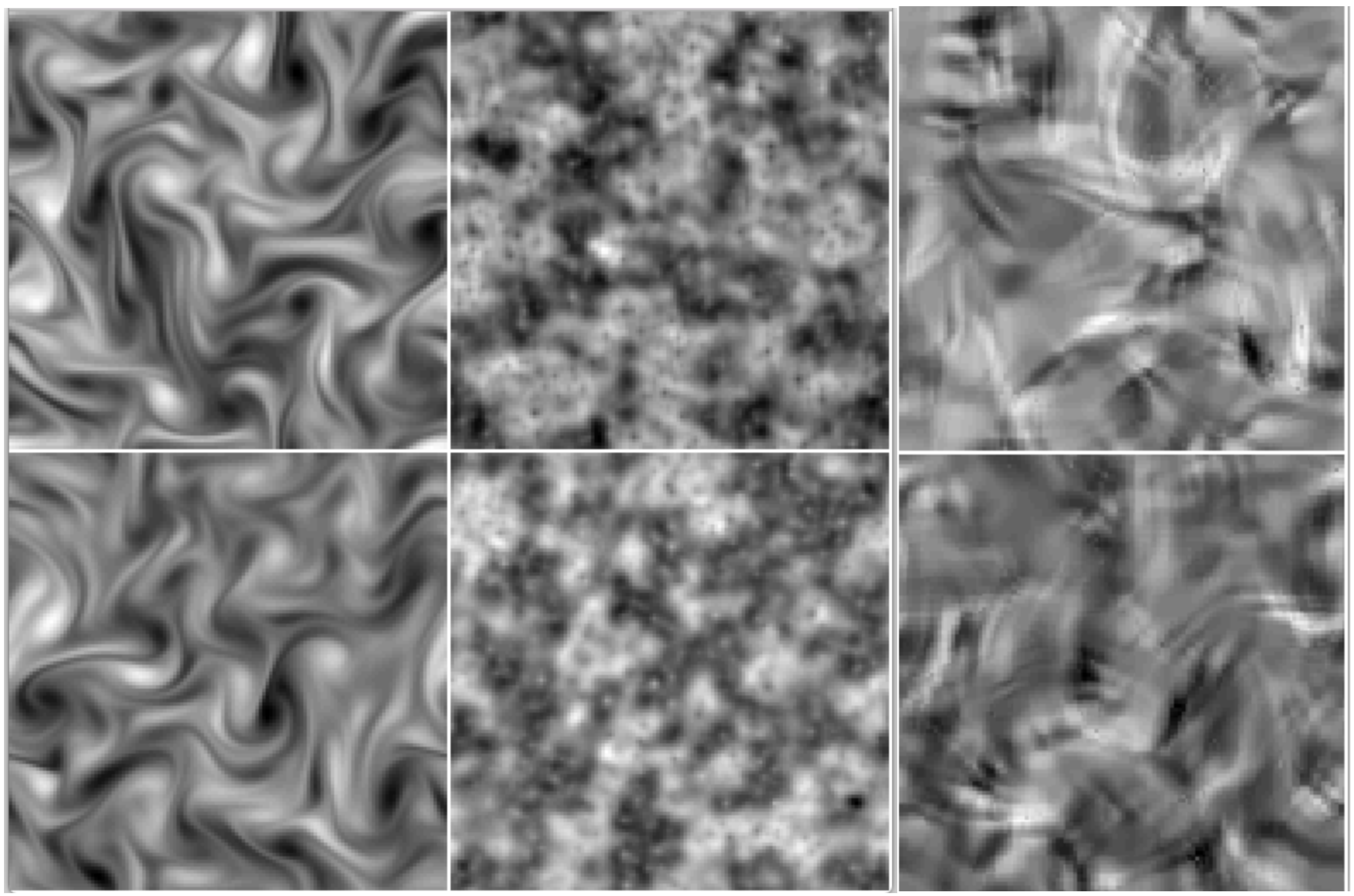


Microcanonical Reconstructions

$$\Phi(x) = \Big\{ \sum_u x(u) \ \ , \ \ \|x \star \psi_{\lambda_1}\|_1 \, , \, \||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}\|_1 \Big\}_{\lambda 1, \lambda_2}$$

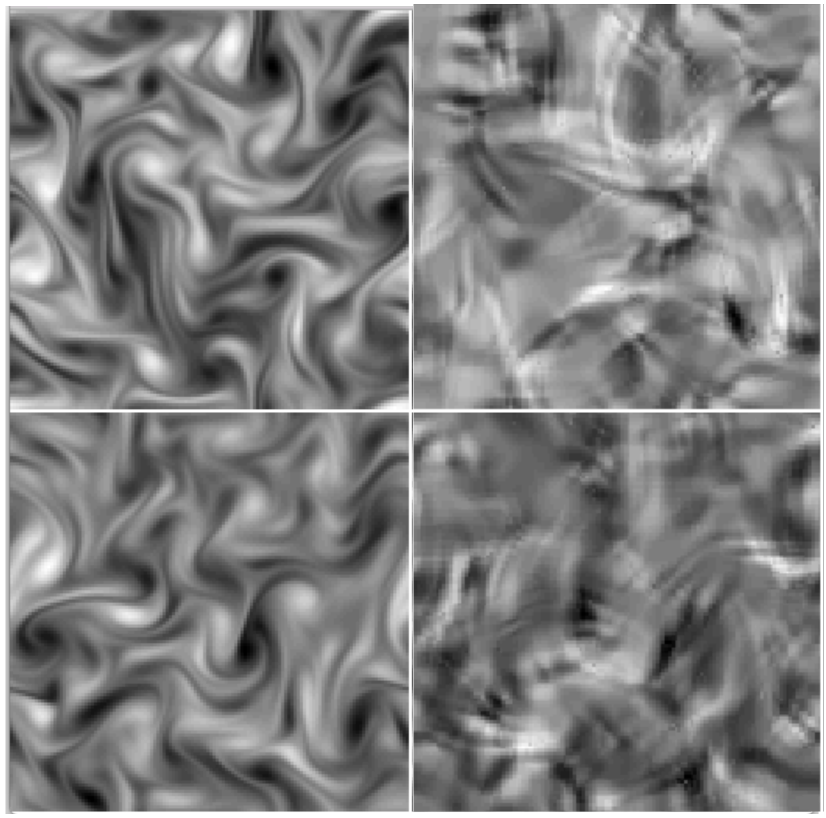
Microcanonical X_∞

Realization x_1 of X order 1 order 2



Must further reduce entropy

Reduction of Model Entropy



- Scattering model of too high entropy

$$|x \star \psi_{\theta,j}(u)| = w(u, \theta, j)$$

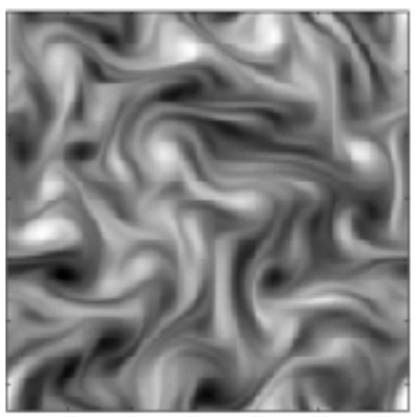
- not sparse at intermediate scales 2^j but not Gaussian

- joint dependance in $(u, \theta) \Rightarrow$ wavelet transforms in (u, θ)

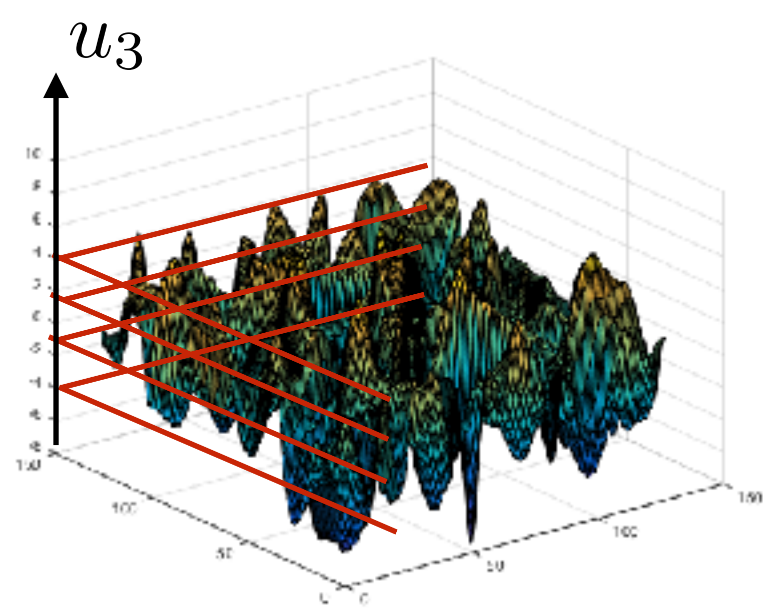
- dependence on amplitude values ?

3D Scattering for Amplitude

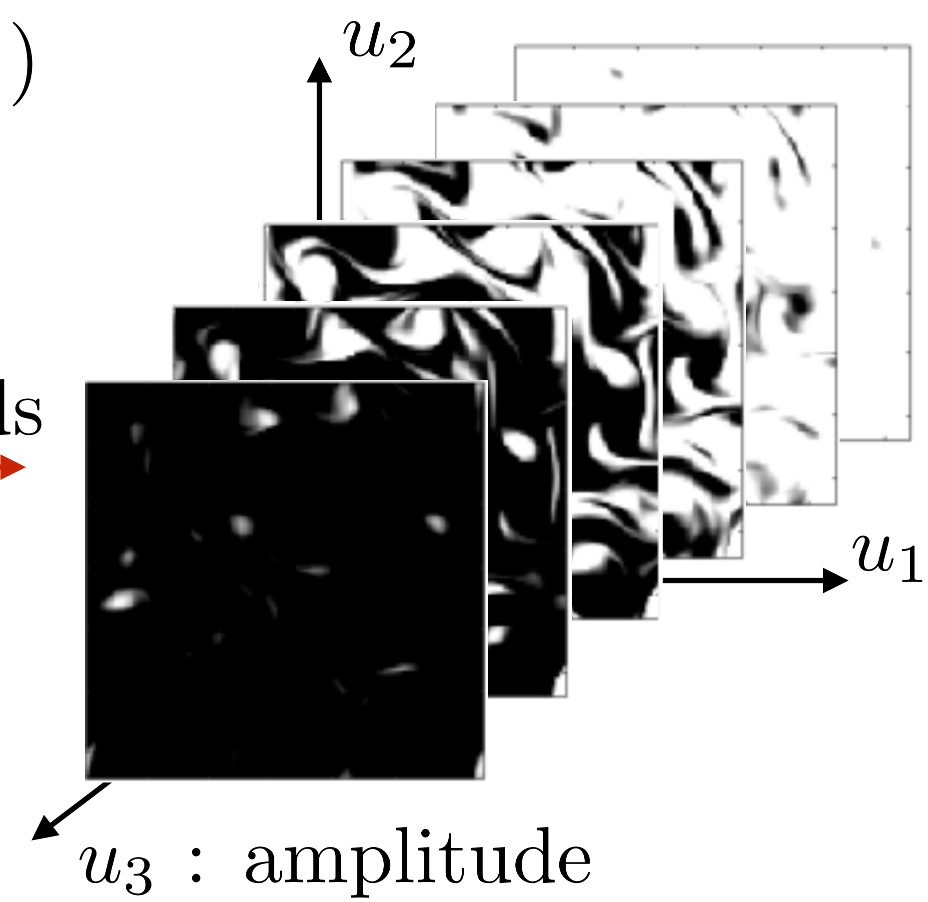
$x(u)$



$$\bar{x}(u_1,u_2,u_3) = \delta(u_3 - x(u_1,u_2))$$



sigmoids $\longrightarrow$



We want Φ in $\mathbf{C^1}(\mathrm{Diff}(\mathbb{R}^{\mathbf{3}}))$

$3D$ wavelets: $\psi_\lambda(u_1,u_2,u_3) = 2^{-2j}\psi(2^{-j}r_\theta(u_1,u_2))\,2^{-\ell}\psi(2^{-\ell}u_3)$

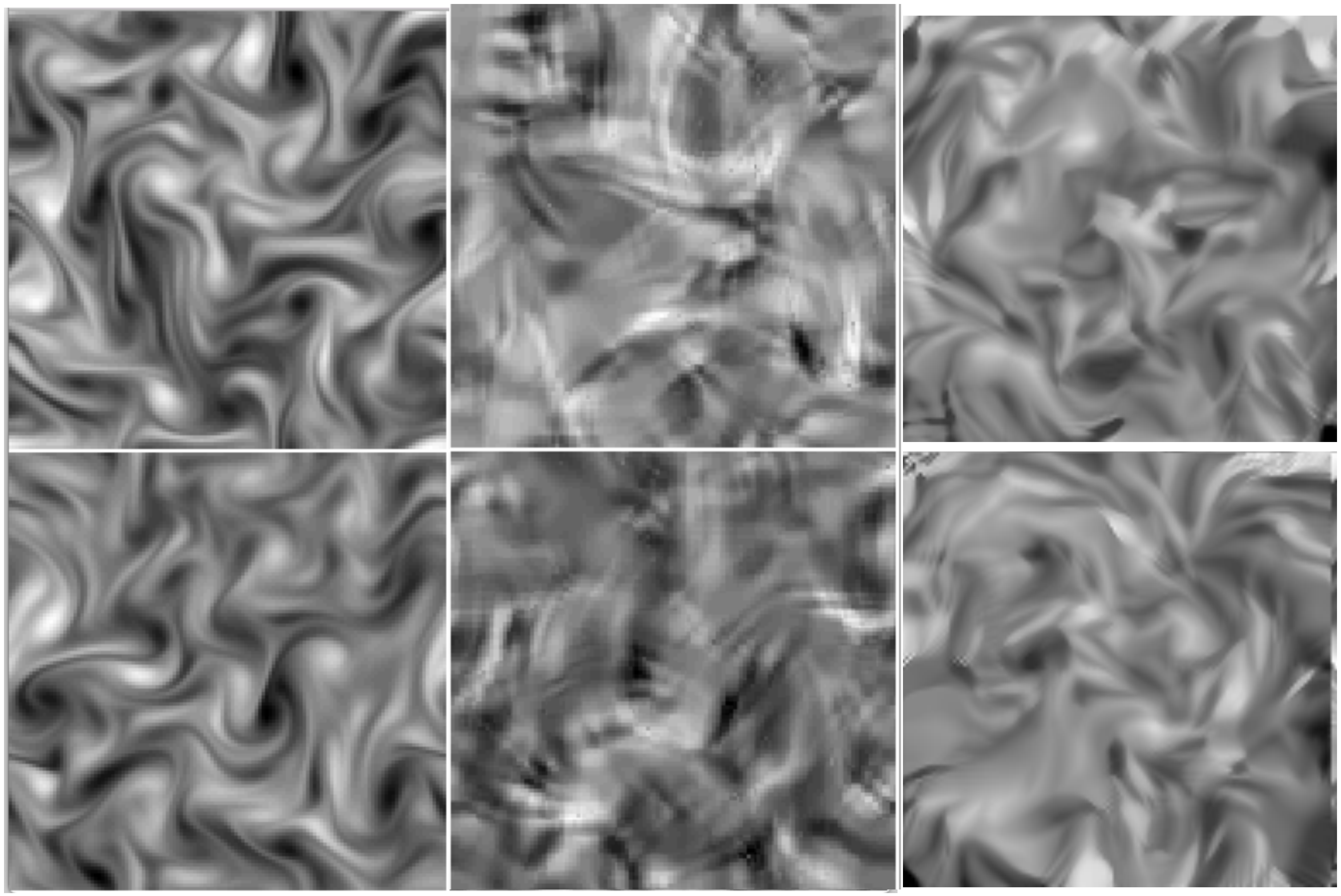
Joint dependance on amplitude and spatial geometry

$$\Phi\bar{x} = \begin{pmatrix} \sum_u \bar{x}(u) \\ \|\bar{x}\star\psi_{\lambda_1}\|_1 \\ \||x\star\psi_{\lambda_1}|\star\psi_{\lambda_2}\|_1 \\ \cdots \end{pmatrix}_{\lambda_1,\lambda_2,\ldots}$$

Wavelet coefficients are much more sparse at intermediate scales

3D Scattering Models

Preliminary results



Conclusions

- Regularity in high dimension as regularity to action of diffeomorphisms on different groups

- Long range dependence: variable separation through scales

- Entropy reduction with sparsity: **L1** geometry