Learning Physics from Data with Multiscale Interaction Models

Stéphane Mallat,

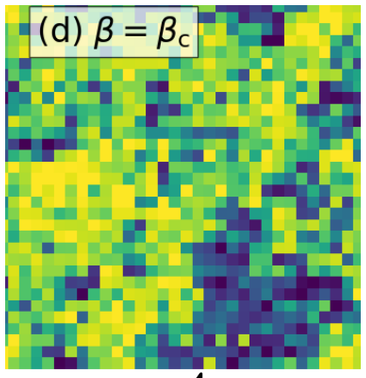
Collège de France
École Normale Supérieure

Learning Physics and Probabilities

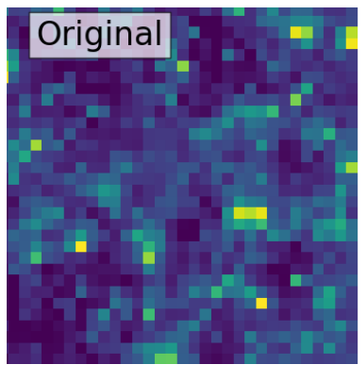
- Unsupervised learning: estimate $p(x)$ for $x \in \mathbb{R}^d$ from $\{x_i\}_i$
  Physical systems at equilibrium: learn the energy $U(x)$

$$p(x) = \mathcal{Z}^{-1}\ e^{-U(x)}$$

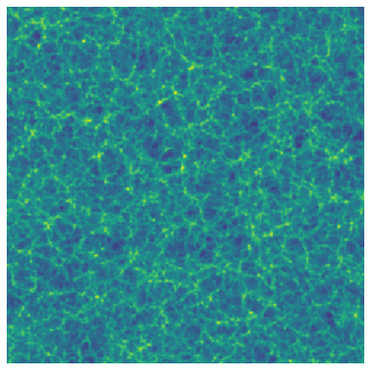
Curse of dimensionality unless strong priors



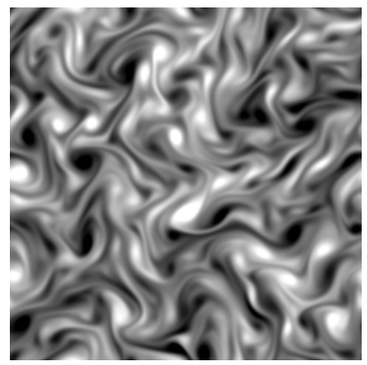
φ^4



weak lensing



Matter density



Turbulences

Learn from few samples

Applications:

- Understand the physics from the energy $U(x)$.
- Data generation by sampling $p(x)$

Learning Physics and Probabilities

- Unsupervised learning: estimate $p(x)$ for $x \in \mathbb{R}^d$ from $\{x_i\}_i$
  Physical systems at equilibrium: learn the energy $U(x)$

$$p(x) = \mathcal{Z}^{-1}\ e^{-U(x)}$$

Three Problems:

- Find $\Phi(x) = \{\Phi_k(x)\}_k$ which can linearly approximate $U(x)$:

$$U(x) \approx \theta^T\, \Phi(x) \qquad \text{(Ansatz)}$$

$$\Leftrightarrow\ \ p_\theta(x) = \mathcal{Z}_\theta^1\, e^{-\theta^T \Phi(x)} \text{ closely approximates } p(x).$$

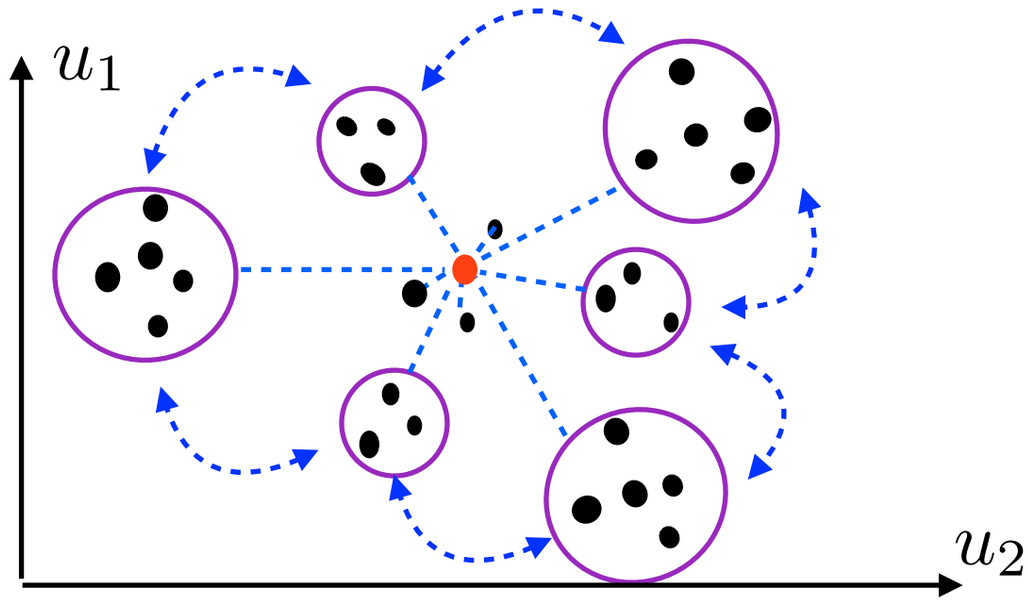
- Optimise θ to minimise the approximation error.
- Sample $p_\theta(x)$ to generate new data x.

Difficult if the energy $U(x)$ is non convex: usually the case.

A solution: scale separation and renormalisation group.

Multiscale Interactions
Dimension Reduction

Interaction of d variables $x(u)$: pixels, particles, agents...



Regroupement of d interactions in $O(\log d)$ multiscale terms.

Fast multipole algorithms.

Multiscale Interactions Claims

1. Long range spatial dependencies are captured by short range dependencies over wavelet coefficients : reduction of dimension of $\Phi(x)$

2. Can specify $\Phi(x)$ from interaction energies across scales, with the *renormalisation group*, and which are "nearly" convex.

3. Interaction energies across scales are non-linear, but can be partly linearised with a modulus or a ReLU by eliminating phases to define $\Phi(x)$ (similar to deep networks).

Overview

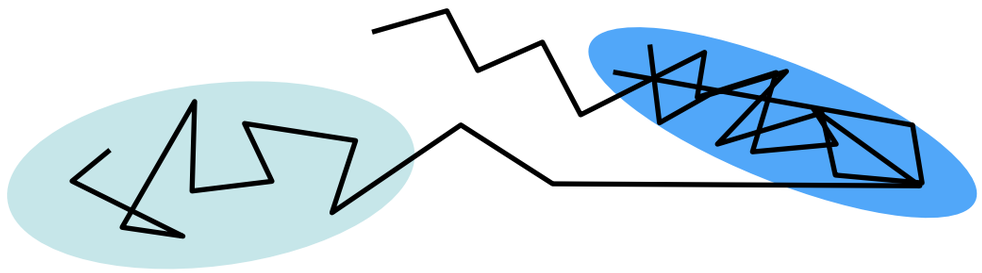
- **I:** Sampling, max likelihood and score matching estimations
- **II:** Renormalisation group and energy interactions across scales
- **III:** Scattering-Gaussian models of scale interactions: turbulences

I. Sampling by Langevin (or MCMC)

- Computing a sample x of p_θ with a Langevin dynamics

$$x_{t+1} - x_t = \epsilon \nabla_x \log p_\theta(x_t) + \sqrt{2\epsilon}\, z_t \quad \text{where} \quad z_t \sim \mathcal{N}(0, Id)$$

Converges but very slow if $-\log p_\theta(x)$ is non-convex in x.



Exploration / Exploitation

Exponential convergence if the energy is strictly convex,

The rate is the inverse of the condition number of $-\nabla_x^2 \log p_\theta(x)$.
(Hessian of energy)

Pb: we want $p_\theta \approx p$, but $-\nabla_x^2 \log p$ is usually not positive,
or badly conditioned.

Optimisation of Parameters θ

- Estimated by minimising Kullback-Liebler divergences

$$KL(p||p_\theta) = \int p(x) \log \frac{p(x)}{p_\theta(x)} dx \; = \mathbb{E}_p[\log p(x) - \log p_\theta(x)]$$

- Optimise θ by gradient descent (likelihood ascent)

$$\theta_{t+1} - \theta_t = \epsilon \nabla_\theta \mathbb{E}_p(\log p_\theta)$$

- For $-\log p_\theta(x) = \log \mathcal{Z}_\theta + \theta^T \Phi(x)$ the optimisation is convex with $\nabla_\theta \mathbb{E}_p(\log p_\theta) = \mathbb{E}_p(\Phi(x)) - \mathbb{E}_{p_\theta}(\Phi(x))$: moment matching which requires to compute samples of p_θ : expensive.

Score Matching

$$-\log p_\theta(x) = \log \mathcal{Z}_\theta + \theta^T \Phi(x)$$

- The score matching eliminates the influence of $\mathcal{Z}_\theta$

  Minimisation of the Fisher Information divergence:

$$FI(p\|p_\theta) = \mathbb{E}_p[\|\nabla_x \log p_\theta(x) - \nabla_x \log p(x)\|^2]$$

  Quadratic function of θ:

$$FI(p\|p_\theta) = \mathbb{E}_p[\frac{1}{2}\|\theta^T \nabla_x \Phi(x)\|^2 - \theta^T \Delta_x \Phi(x)]$$

Calculated from data, no need to sample p_θ : fast algorithm.

Score Matching

$$-\log p_\theta(x) = \log \mathcal{Z}_\theta + \theta^T \Phi(x)$$

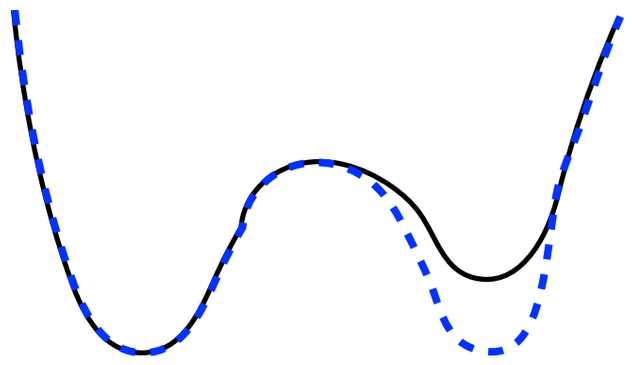
- The score matching eliminates the influence of $\mathcal{Z}_\theta$

  Minimisation of the Fisher Information divergence:

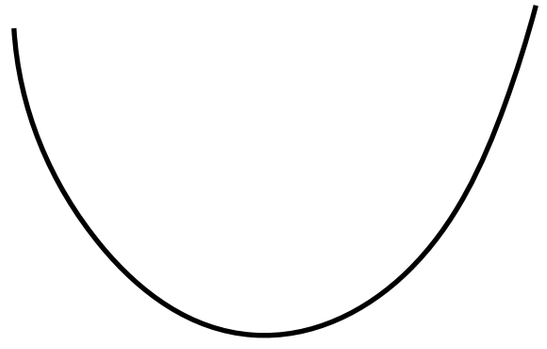
$$FI(p\|p_\theta) = \mathbb{E}_p[\|\nabla_x \log p_\theta(x) - \nabla_x \log p(x)\|^2]$$

Theorem: If $-\log p_\theta(x)$ and $-\log p(x)$ are strictly convex then score matching estimation is identical to maximum likelihood.

Non convex energy



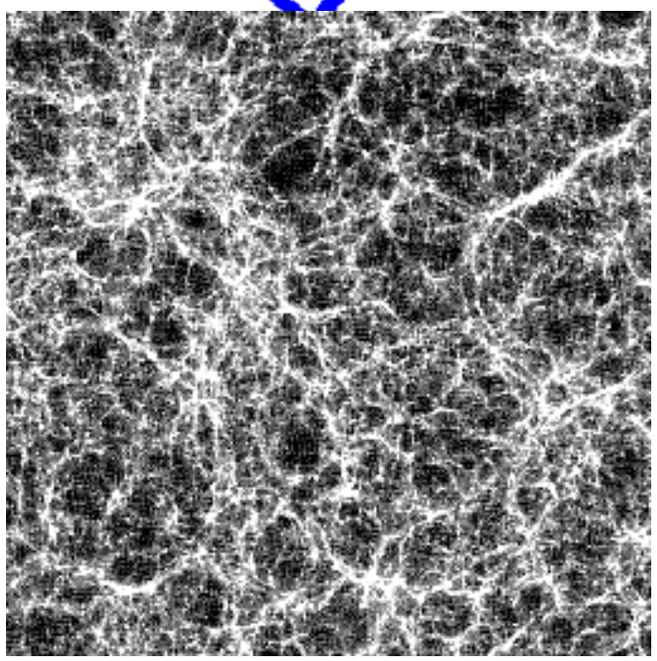
Convex energy



Eliminating constants does not allow to estimate local minima levels

II. Renormalisation Group

Kadanoff, Wilson 1970



x

- Given $p(x) = \mathcal{Z}^{-1} e^{-U(x)}$ compute

  $p(x_j)$ at all scales 2^j: regular evolution.

  $x_{j-1} \to x_j$: average, subsample, normalise.

x_{j-1}

x_j

x_{j+1}

x_J

- Ansatz : $p(x_j) = \mathcal{Z}_j^{-1} e^{-\theta_j^T \Phi(x_j)}$ how to find it ?

$\theta_j = f(\theta_{j-1})$: RG equation on couplings

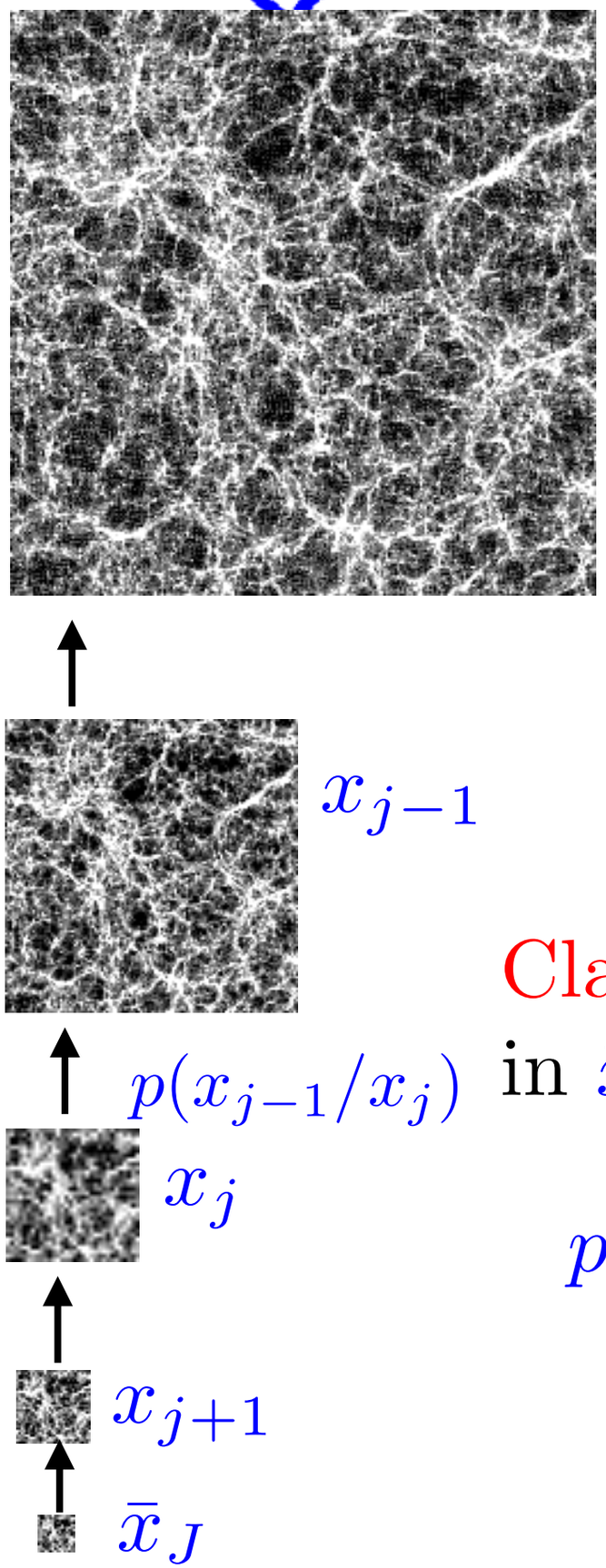
At phase transitions: $\theta_j = \theta_{j-1}$.

- Calculated by decomposing x_{j-1} into $(x_j, \overline{x}_j)$:

$$\Rightarrow\ p(x_j) = \int p(x_{j-1})\, d\bar{x}_j\ = \int p(x_j)\, p(\bar{x}_j|x_j)\, d\bar{x}_j$$

Wilson: approximation with Gaussian integration in $\bar{x}_j$

Inverse Renormalization Group



T. Marchand, M. Ozawa, G. Biroli

- **Pb:** estimate physical energies from data.

Coarse to fine:

$$p(x_{j-1}) = p(x_j)\, p(\bar{x}_j/x_j)$$

with normalised variances $\bar{x}_j \to \bar{x}_j/\sigma_j$

$$\Rightarrow\ p(x) = p(x_J) \prod_{j=1}^{J} p(\bar{x}_j/x_j)$$

Claims 1,2: $p(\bar{x}_j|x_j)$ has short range dependencies in $\bar{x}_j$, and its energy is "nearly" convex in $\bar{x}_j$.

$p(x_J)$ may be non-convex but low-dimensional.

It is usually Gaussian.

Wavelet Orthogonal Bases

Decompose each x_{j-1} into $(x_j, \bar{x}_j)$: in what basis ?

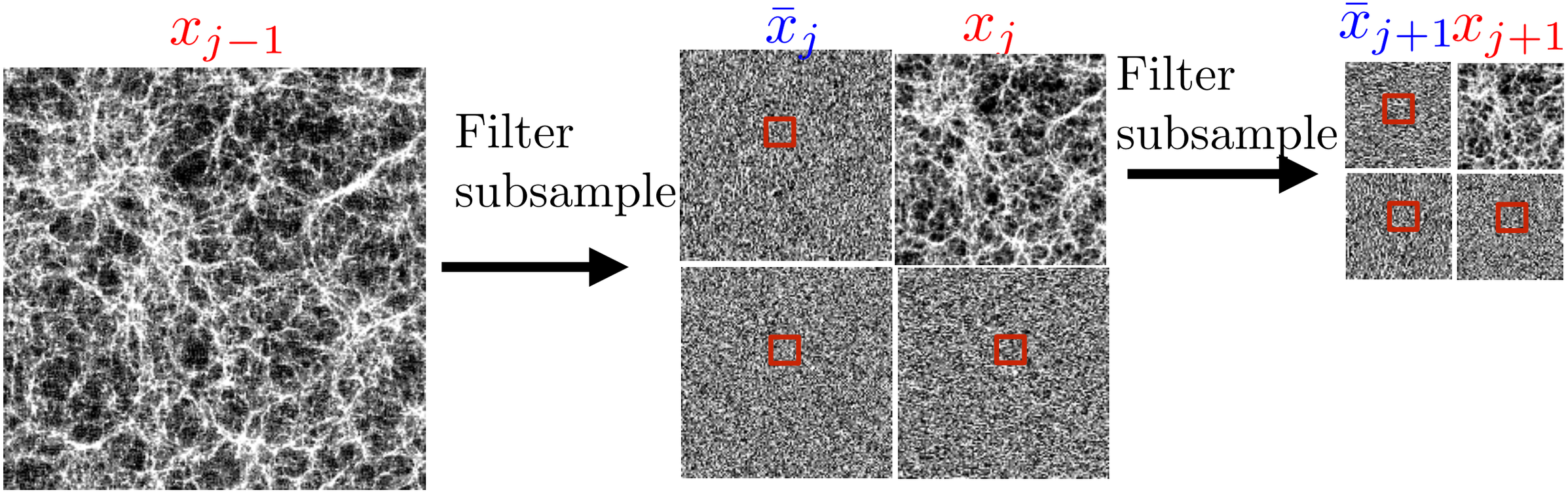
translated, and dilated local wavelets

$$\psi^m_{j,n}(u) = 2^{-dj/2}\,\psi^m(2^{-j}u - n)$$

$\left\{\psi^m_{j,n}\right\}_{m,j,n}$ can define an orthonormal basis of $\mathbf{L^2}(\mathbb{R}^d)$

$$\overline{x}_j(n,m) = \langle x, \psi^m_{j,n}\rangle = x * \psi^m_j(2^j n)$$

Fast wavelet transform: $p(\bar{x}_j|x_j)$ has short dependencies



Wavelet Orthogonal Bases

Decompose each x_{j-1} into $(x_j, \bar{x}_j)$: in what basis ?

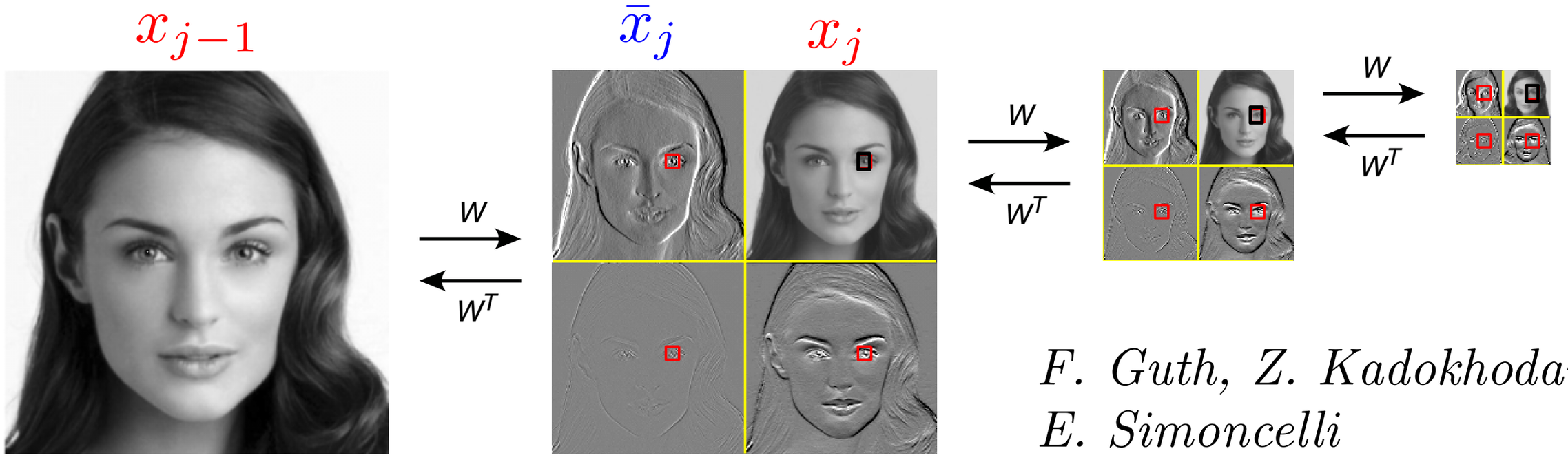
translated, and dilated local wavelets

$$\psi^m_{j,n}(u) = 2^{-dj/2}\,\psi^m(2^{-j}u - n)$$

$\left\{\psi^m_{j,n}\right\}_{m,j,n}$ can define an orthonormal basis of $\mathbf{L^2}(\mathbb{R}^d)$

$$\overline{x}_j(n,m) = \langle x, \psi^m_{j,n}\rangle = x * \psi^m_j(2^j n)$$

Fast wavelet transform: $p(\bar{x}_j|x_j)$ has short dependencies



F. Guth, Z. Kadokhodaie
E. Simoncelli

Wavelet Conditional RG Models

T. Marchand, M. Ozawa, G. Biroli

$$p(x) = p(x_J) \prod_{j=1}^{J} p(\overline{x}_j / x_j)$$

Define parameterised models of each $p(\overline{x}_j / x_j)$

$$p_{\bar{\theta}_j}(\overline{x}_j / x_j) = \bar{\mathcal{Z}}_j^{-1} \, e^{-\bar{\theta}_j^T \Phi(x_{j-1})}$$

with same Φ but different $\bar{\theta}_j$ at all scales

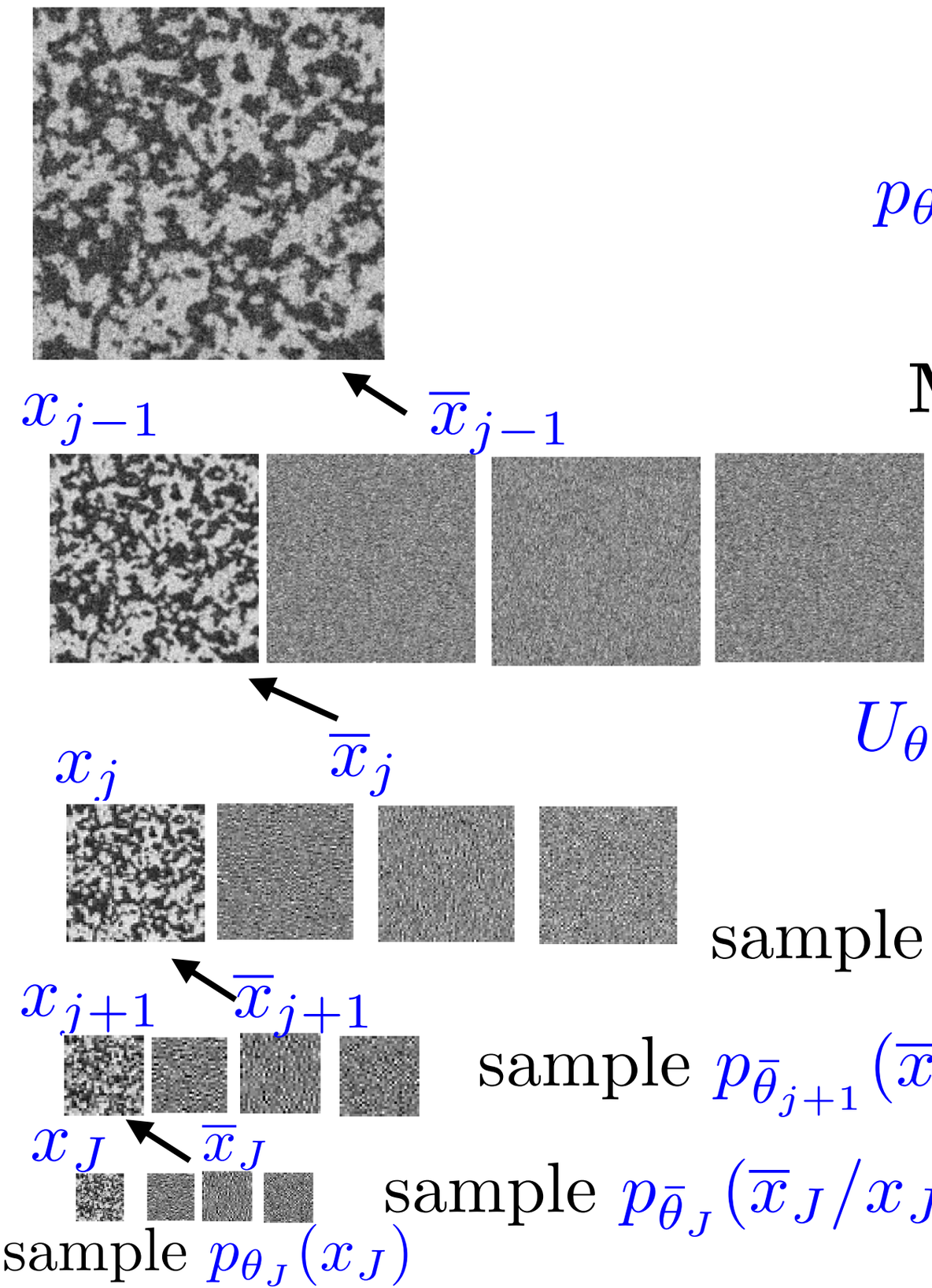
Claims 1,2:

1. Short range dependencies: $\bar{\theta}_j$ has few non-zero parameters
2. Convexity of $\bar{\theta}_j^T \Phi(x_{j-1})$ in $\bar{x}_j$ and well conditioned Hessian.

$\Rightarrow$ fast sampling of $p_{\bar{\theta}_j}$ and fast optimisation of $\bar{\theta}_j$.

Multiscale Sampling & Energy

T. Marchand, M. Ozawa, G. Biroli



$$p_\theta(x) = p_{\theta_J}(x_J) \prod_{j=1}^{J} p_{\bar{\theta}_j}(\overline{x}_j/x_j)$$

Multiscale energy estimation

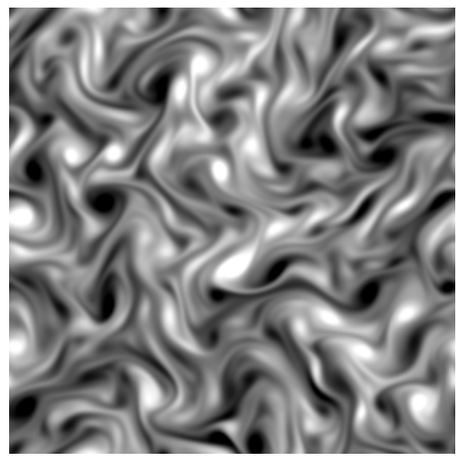
$$p_\theta(x) = \mathcal{Z}_\theta^{-1} e^{-U_\theta(x)}$$

$$U_\theta(x) = \theta_J^T \Phi(x_J) + \sum_{j=1}^{J} \bar{\theta}_j^T \Phi(x_{j-1})$$

Multiscale Gaussian Models

$$p_\theta(x) = \mathcal{Z}_\theta^{-1}\, e^{-\frac{1}{2} x^T\, \theta\, x} \;= \mathcal{Z}_\theta^{-1}\, e^{-\theta^T \Phi(x)} \text{ with } \Phi(x) = xx^T$$

Turubulent Flow:

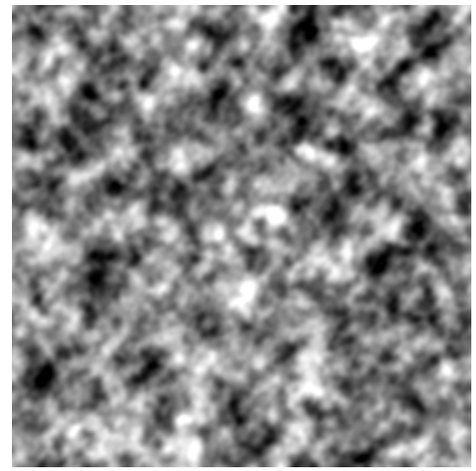


L

At convergence $\nabla_x^2 \log p_\theta(x) = \theta = K \geq 0$

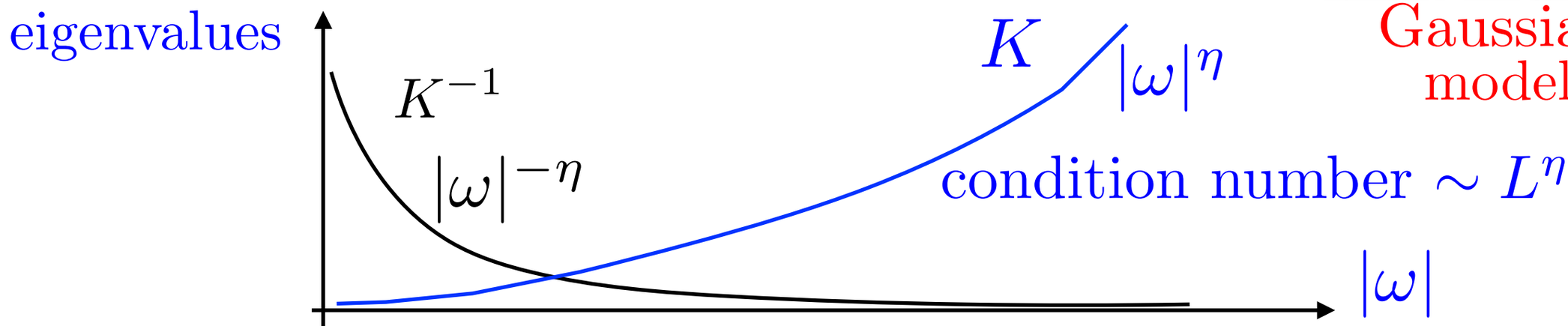
where K^{-1} is the process covariance.

Stationarity $\Rightarrow$ K^{-1} is Toeplitz diagonalised in Fourier



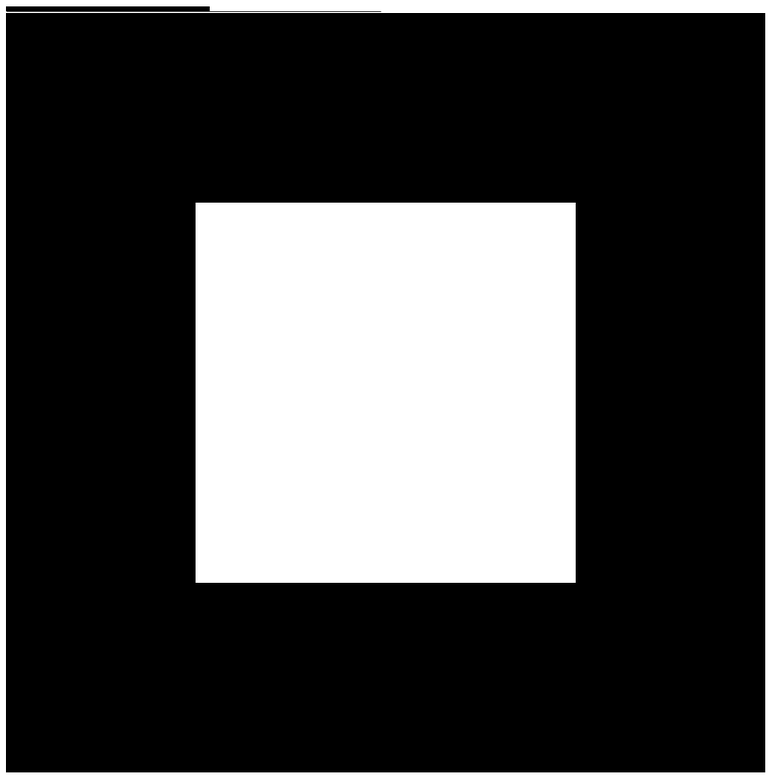
Gaussian model

Power spectrum = covariance eigenvalues,
has a power-lay decay for multiscale fields.
$\Rightarrow$ long range dependencies.



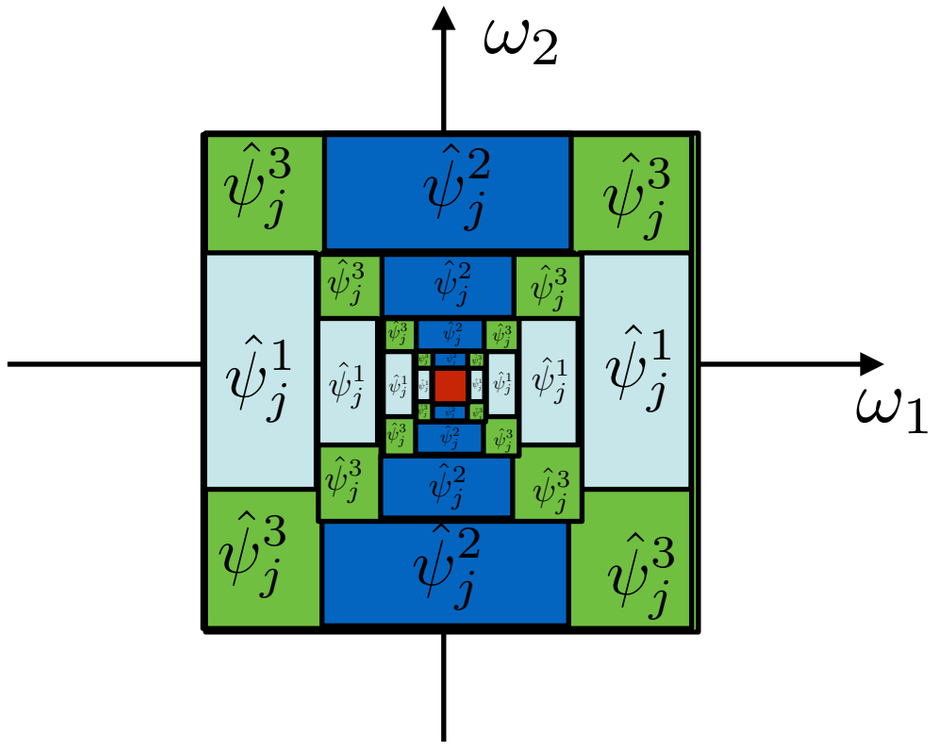
Wavelet Transform in Fourier

Orthogonal wavelets decomposes in different frequency bands



Wavelet coefficients

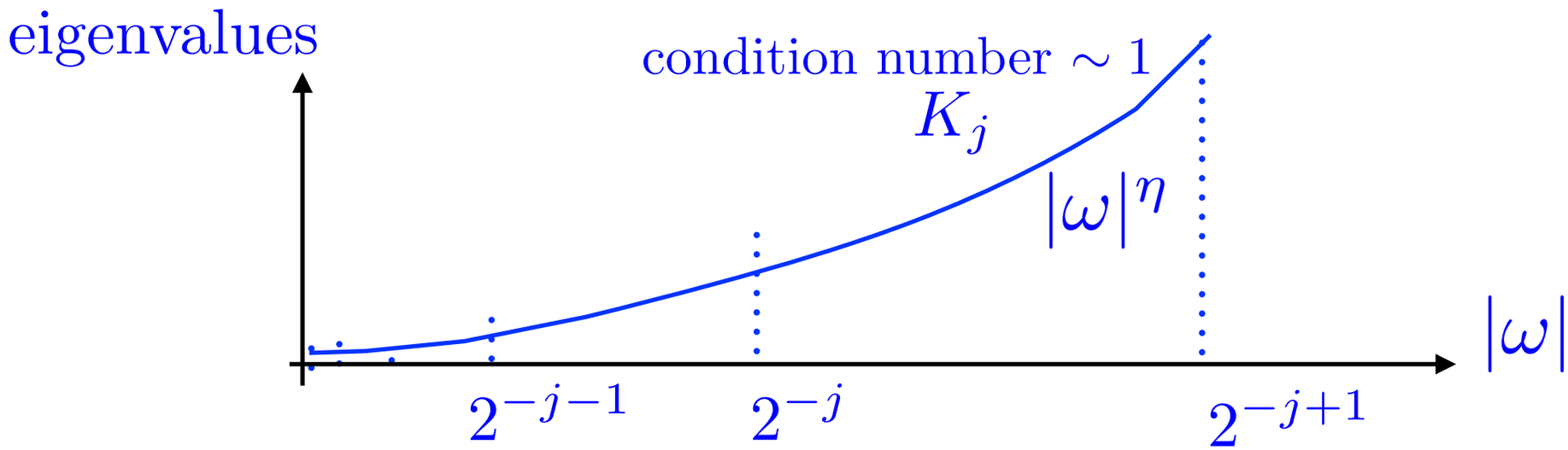
Frequency (Fourier) domain



Well Conditioned Hessian

- For fields having a power spectrum with power law decay

$$\nabla^2 \log p(\bar{x}_j|x_j) = K_j = \bar{\theta}_j$$



Wavelet representation of singular operators

Theorem: For Gaussian models of a spectrum with power-law decay each K_j is a band matrix with a condition number ~ 1.

$\Rightarrow$ fast sampling of $p_{\bar{\theta}_j}(\overline{x}_j/x_j)$ with Langevin (or MCMC).

and short range dependencies: low-dimensional model.

Potential Energies in Physics

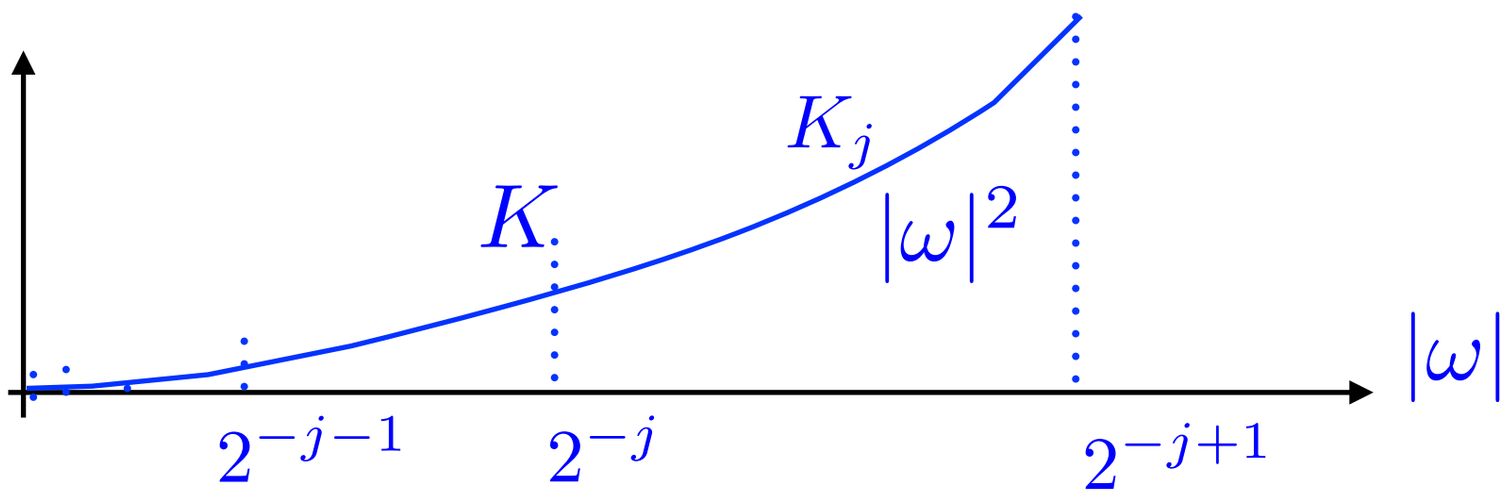
$$p(x) = \mathcal{Z}^{-1}\, e^{-U(x)} \quad \text{with } U(x) = \tfrac{1}{2} x^T K x + V(x)$$

where $K = \beta\, \Delta$ is the kinetic energy

$V(x)$ is the non-convex potential, creating scale interactions

$\nabla^2 \log p(x) = K + \nabla^2 V(x)$ is usually not convex.

The non-convexity appears at lower frequencies



$\nabla^2 \log p(\bar{x}_j | x_j) = K_j + \nabla^2 \bar{V}_j(x_{j-1})$ is "often" positive.

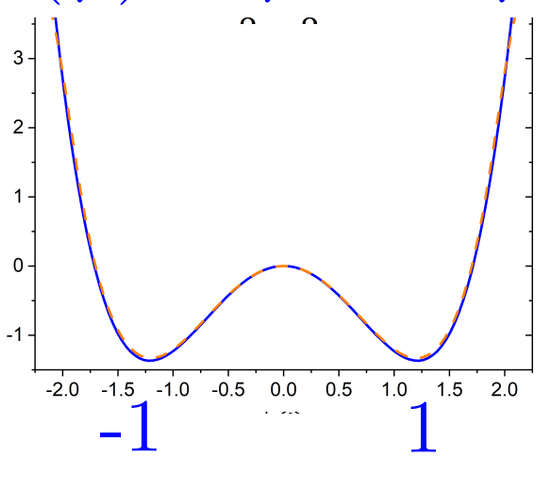
because K_j is strongly positive.

Scalar Potential Models

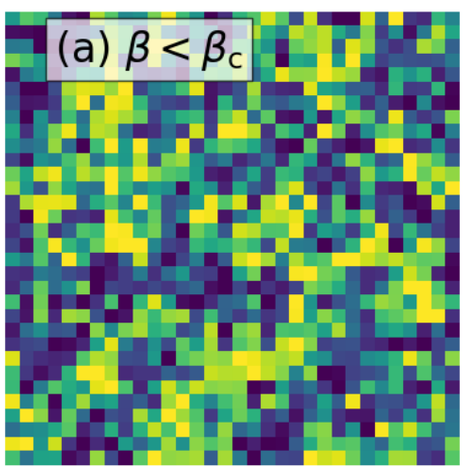
$$\varphi^4 \text{ model : } U(x) = \frac{\beta}{2} x^T \Delta x + \sum_n V(x_n)$$

non-convex scalar potential

$$V(\varphi) = \varphi^4 - \alpha\varphi^2$$

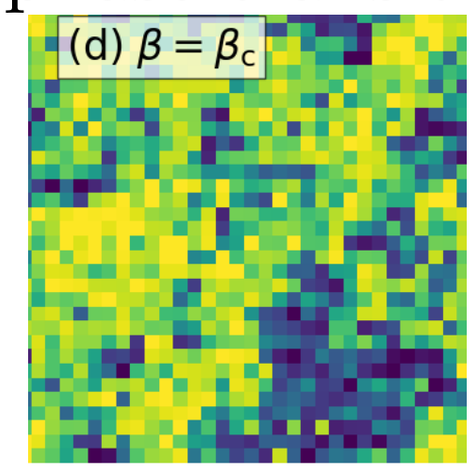


-1 1



$\beta < \beta_c$

phase transition



$\beta = \beta_c$

- Polynomial parametric model:

$$V(\varphi) = \sum_k \theta_k \, P_k(\varphi) \quad \Rightarrow \quad U(x) = \theta^T \Phi(x)$$

$$\text{with } \Phi(x) = \left(xx^T \, , \sum_n P_k(x_n) \right)_k$$

- Stationarity $\Rightarrow$ $U(x) = \tilde{\theta} * \tilde{\Phi}(x)$

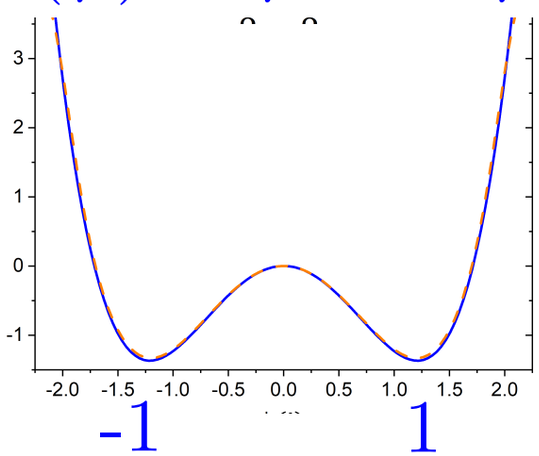
  Convolutional scalar potential Ansatz

Scalar Potential Models: No Convexity

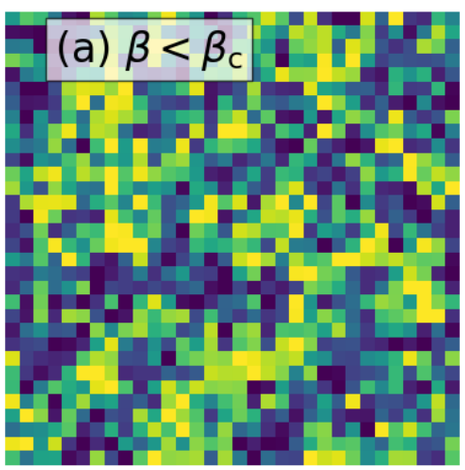
$$\varphi^4 \text{ model} : U(x) = \frac{\beta}{2} x^T \Delta x + \sum_n V(x_n)$$

scalar potential

$$V(\varphi) = \varphi^4 - \alpha\varphi^2$$

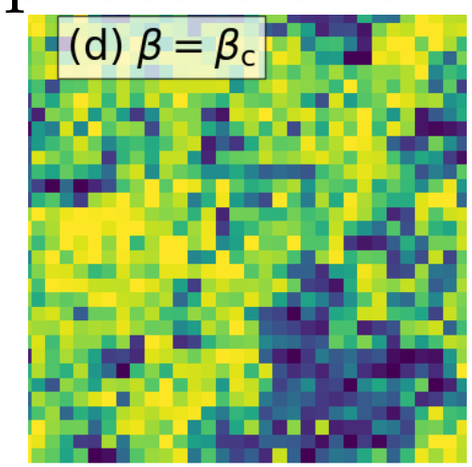


-1 1



$\beta < \beta_c$

phase transition

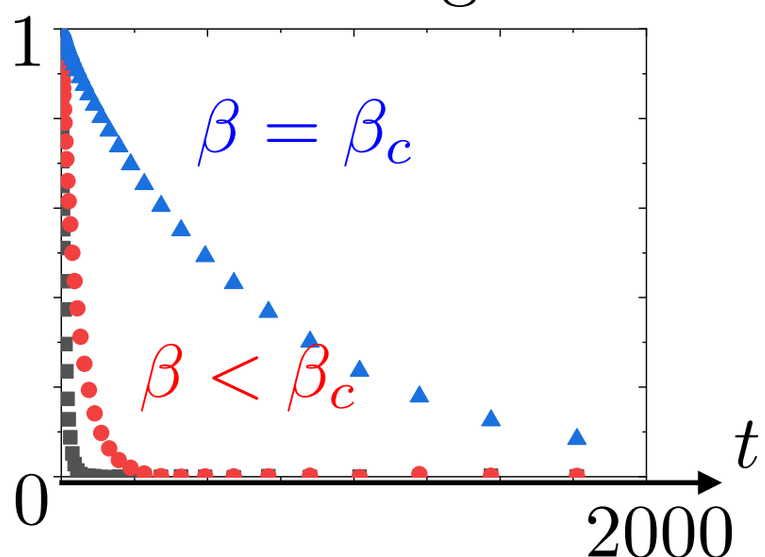


$\beta = \beta_c$

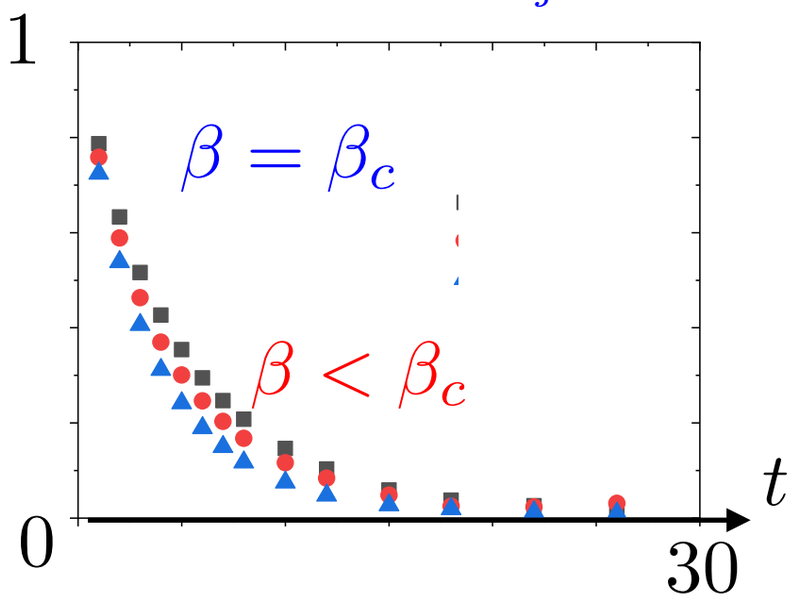
- Mixing time for sampling by MCMC Metropolis:

Sampling $p(x)$

"Critical slowing down"



Sampling all $p_{\bar{\theta}_j}(\bar{x}_j|x_j)$

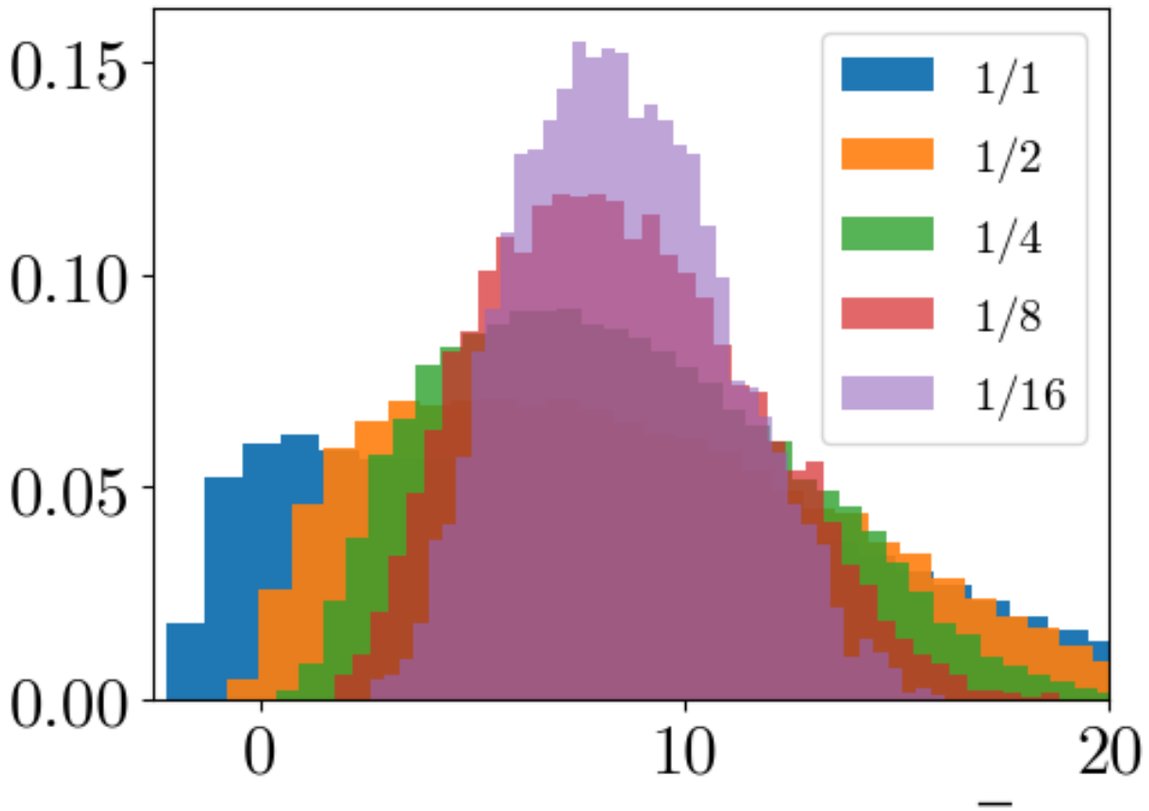


Convexity and score matching

J. Bruna, F. Guth, E. Lempereur

The interaction energies $\bar{\theta}_j \star \Phi(\bar{x}_{j-1})$ of all $p_{\bar{\theta}_j}(x_j|x_j)$ are "nearly" convex with well conditioned Hessians:

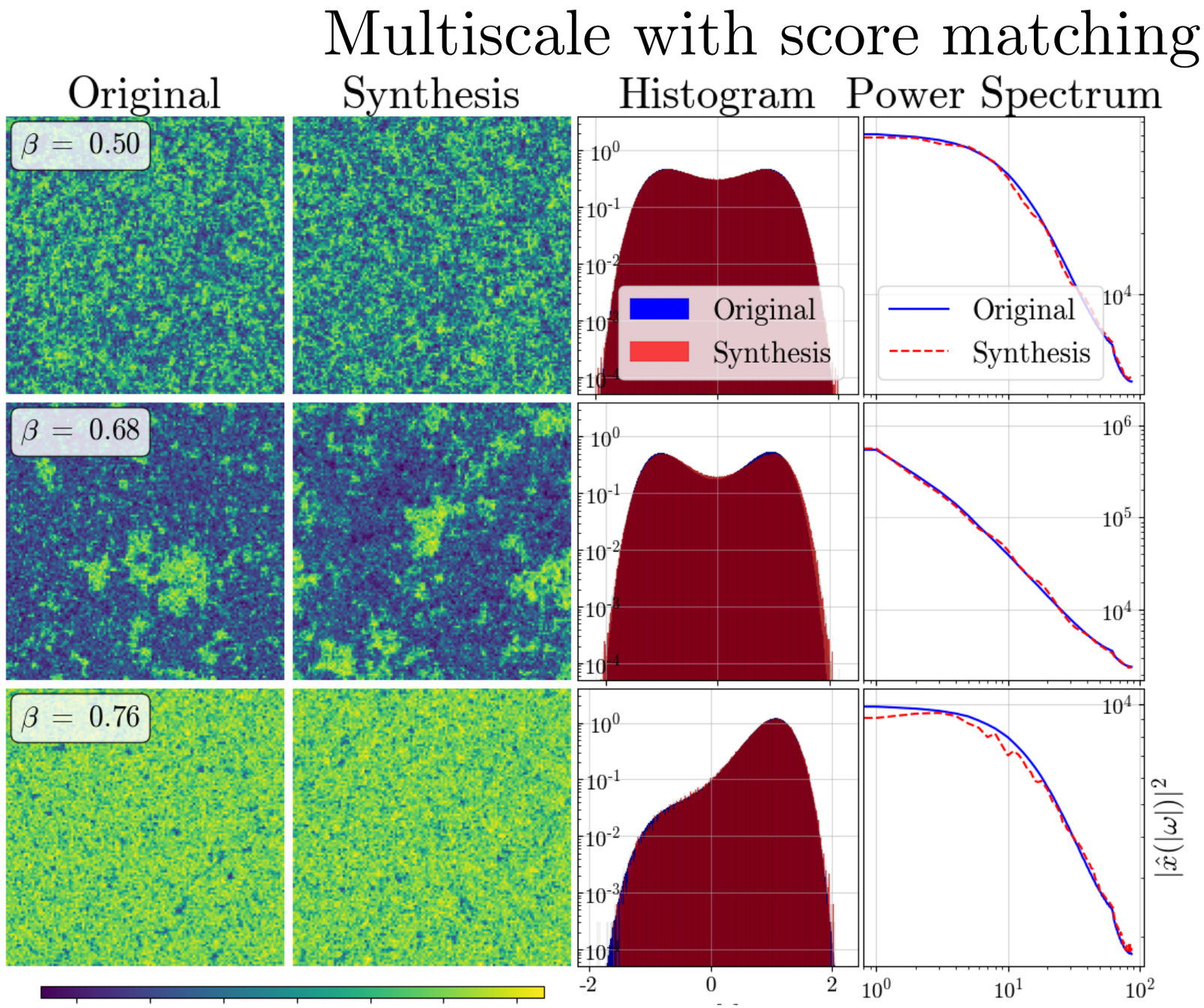
Histograms of Hessian eigenvalues



Convexity and score matching

J. Bruna, F. Guth, E. Lempereur

The interaction energies $\bar{\theta}_j \star \Phi(\bar{x}_{j-1})$ of all $p_{\bar{\theta}_j}(x_j|x_j)$ are "nearly" convex with well conditioned Hessians:



Modeling Weak Lensing

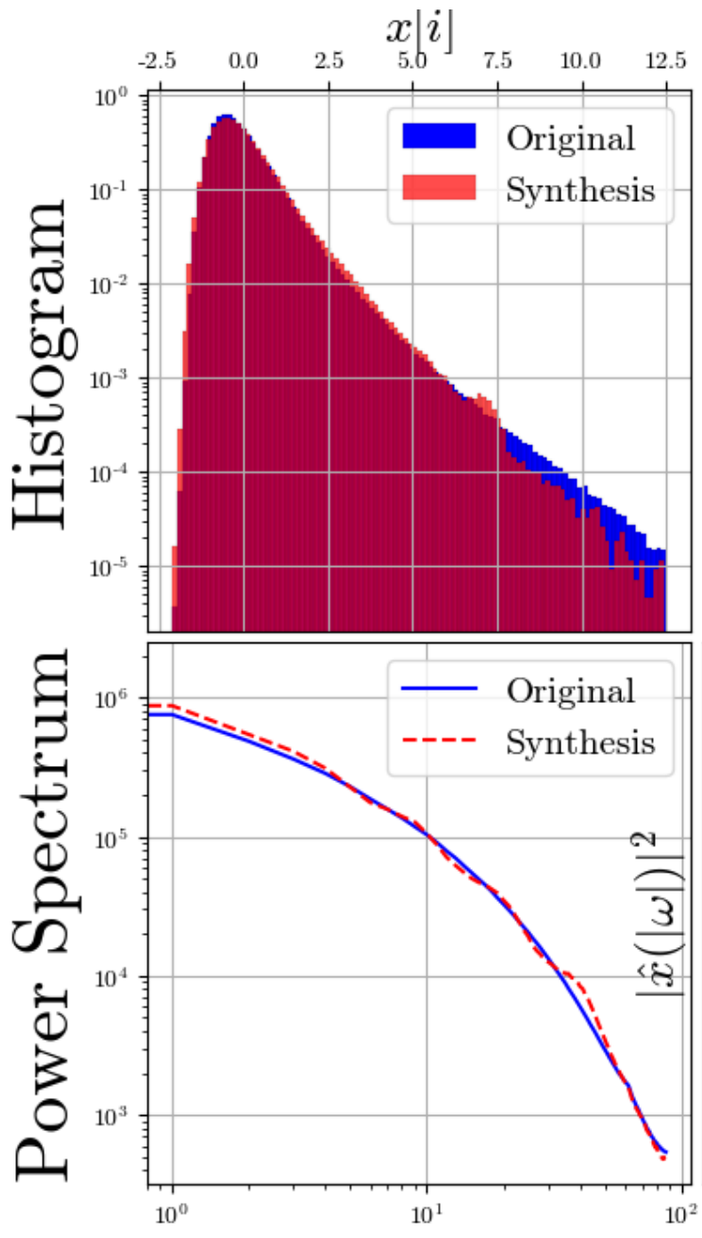
J. Bruna, F. Guth, E. Lempereur

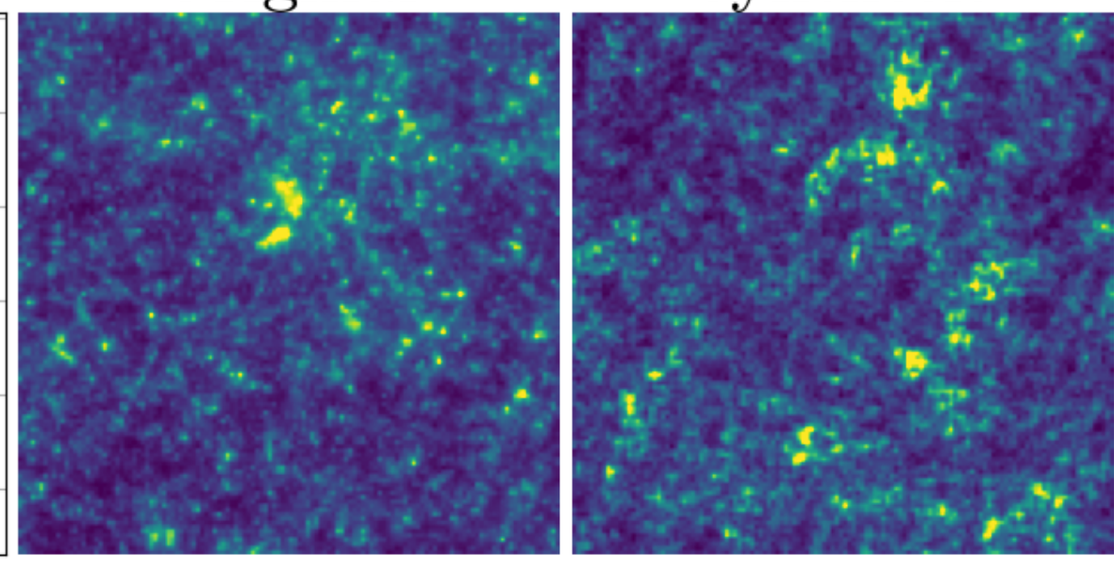
Multiscale convolutional scalar potential models of weak lensing

$$U_\theta(x) = \theta_J^T \Phi(x_J) + \sum_{j=1}^{J} \bar{\theta}_j^T \Phi(x_{j-1})$$

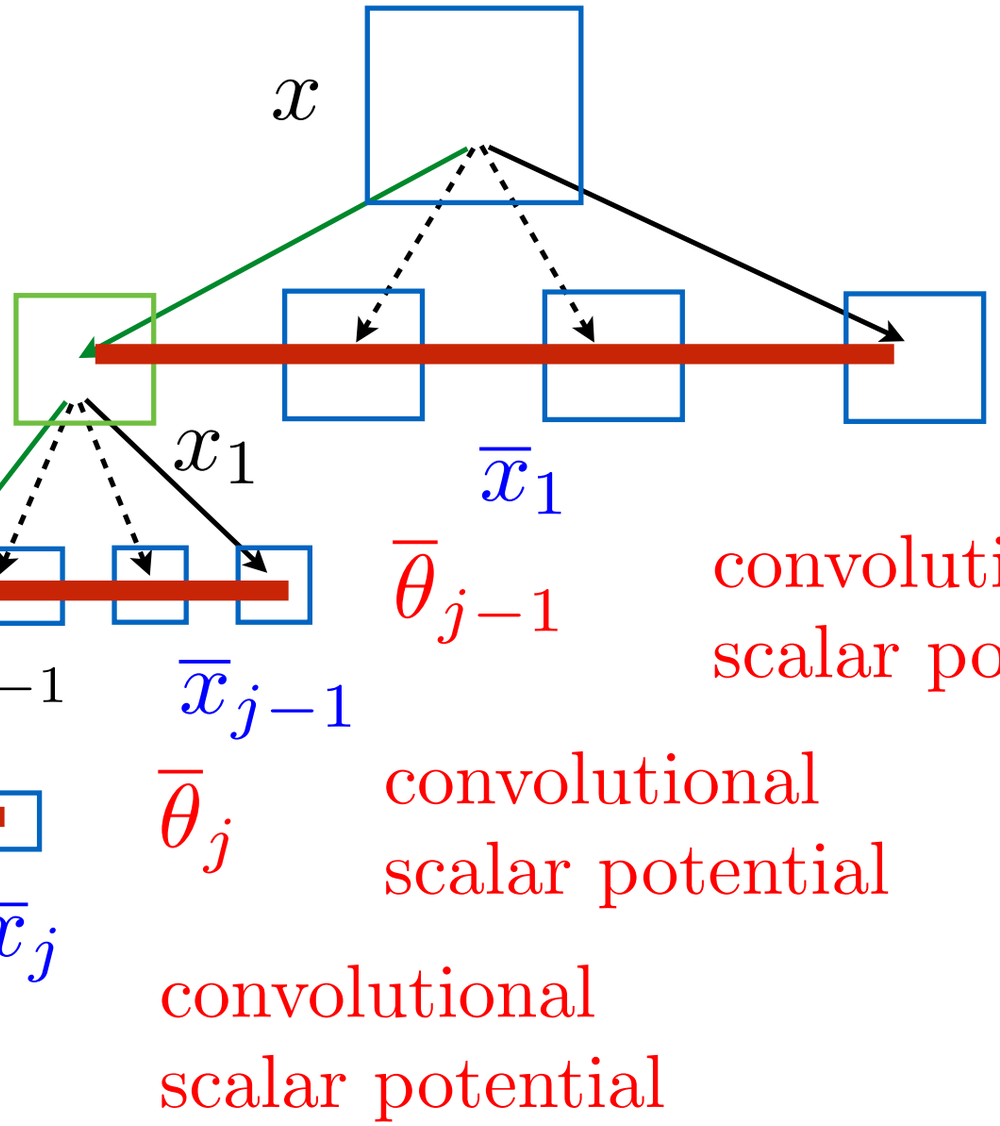
Euclid mission

Columbia lensing simulations





Wavelet Conditional RG



normalisation
interaction model

normalisation
interaction

normalisation
interaction

normalisation
interaction

convolutional
scalar potential

$\overline{\theta}_{j-1}$

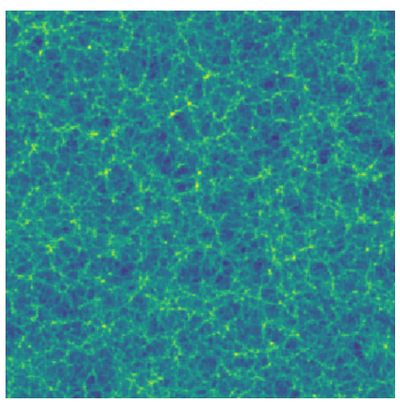
convolutional
scalar potential

$\overline{\theta}_j$

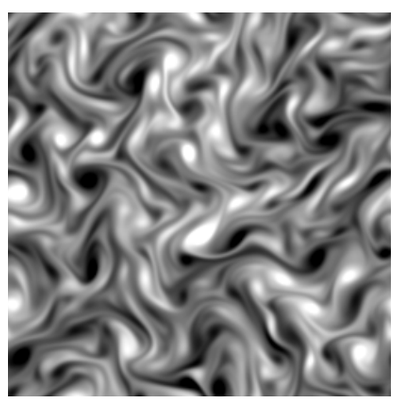
convolutional
scalar potential

Not sufficient to model

Matter density

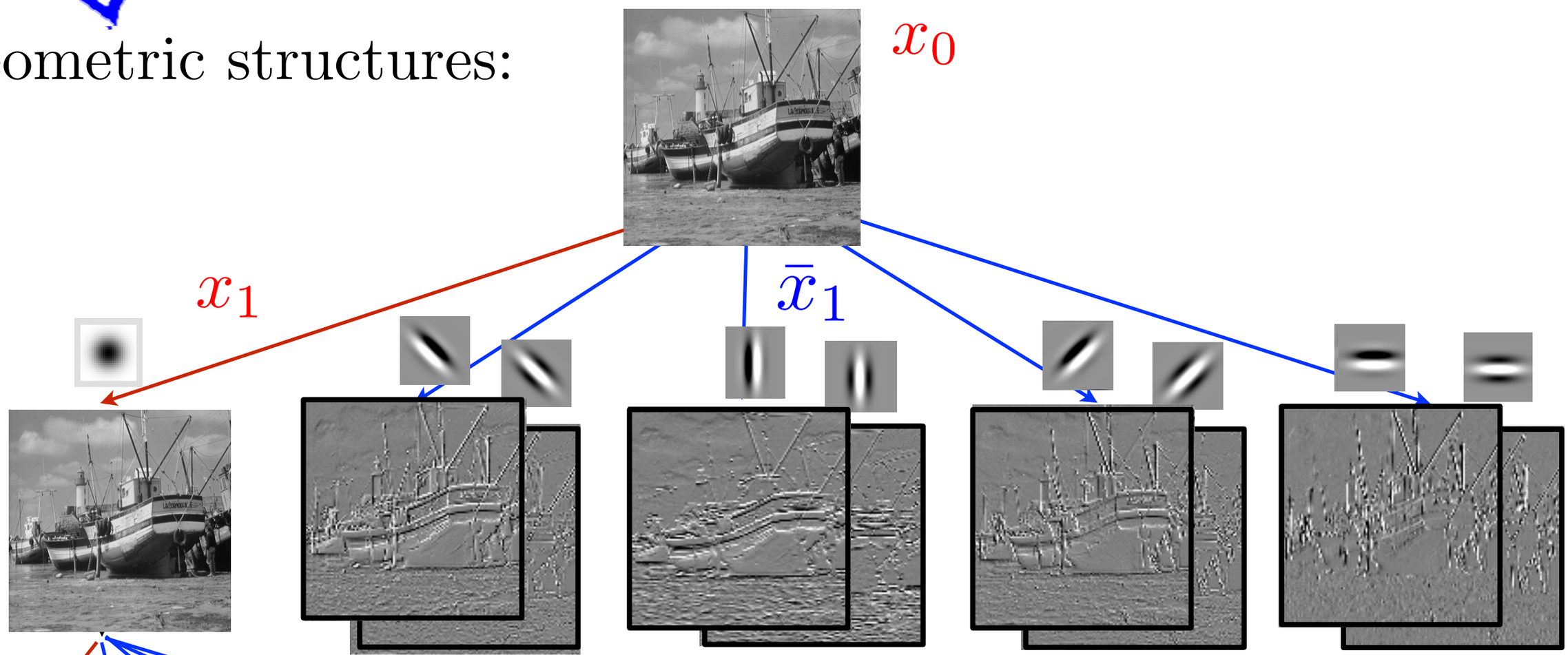


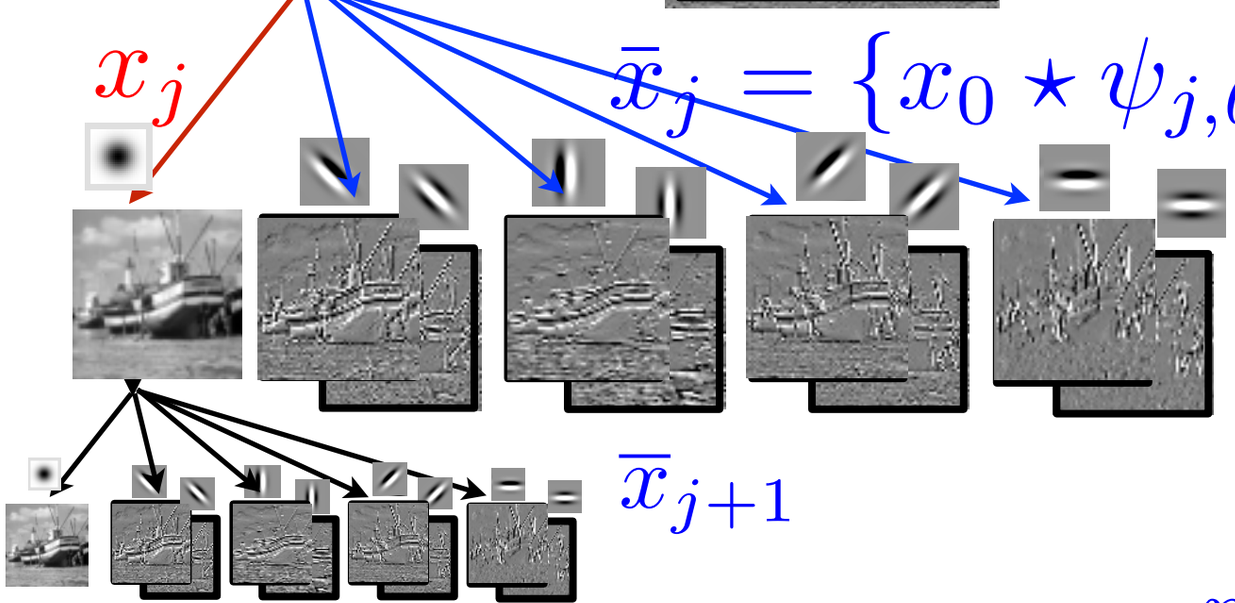
Turbulences



III- Modeling Scale Dependencies

Geometric structures:





- Sparse wavelet coefficients

- Strongly dependant across scales and angles.

$$p(\overline{x}_j / x_j) = p(\overline{x}_j / \overline{x}_\ell \,,\, \ell > j)$$

Modeling Scale Dependencies

Must model the energy of $p(\bar{x}_j / x_j) = p(\bar{x}_j / \bar{x}_{j'}, j' > j)$

Linear models $\theta^T \Phi$ are matching moments: $\mathbb{E}_{p_\theta}(\Phi(x)) = \mathbb{E}_p(\Phi(x))$

What moments should we choose ?

- Gaussian model: $\Phi(x) = x\, x^T \;\Leftrightarrow\; \Phi(x) = Wx\,(Wx)^T$

in a wavelet basis $Wx = (\bar{x}_j)_j$.

For x stationary, the $\overline{x}_j = x \star \psi_j^m$ are not correlated across scales:

$$\mathbb{E}\Big(\overline{x}_j(u)\ \overline{x}_{j'}(u-\tau)\Big) \approx 0 \quad \text{if} \quad j \neq j'$$

because their phase oscillate at different rate since

ψ_j^m and $\psi_{j'}^{m'}$ are supported in different frequency bands.

$\Rightarrow$ Gaussian models do not capture scale dependencies.

Scattering-Gaussian Models

- Eliminates phases with modulus: $(x\,,\,|Wx|) = (x, |\bar{x}_j|)_j$
- Scattering-Gaussian model: *Etienne Lempereur*

$$p_\theta(x) = \mathcal{Z}_\theta{}^{-1}\, e^{-(x,|Wx|)\,\theta\,(x,|Wx|)^T}$$

$$p_\theta = \mathcal{Z}_\theta{}^{-1}\, e^{-\theta^T \Phi(x)} \quad \text{with} \quad \Phi(x) = (x\,,\,|Wx|)\,(x\,,\,|Wx|)^T$$

$$\Rightarrow \quad \Phi(x) = \begin{pmatrix} x(u)\,x(u-\tau) & x(u)\,|\bar{x}_j(u-\tau)| \\ x(u)\,|\bar{x}_j(u-\tau)| & |\bar{x}_j(u)|\,|\bar{x}_{j'}(u-\tau)| \end{pmatrix}_{j,j',\tau}$$

Large matrix because modulus have long range correlations.

Scattering-Gaussian Models

- Eliminates phases with modulus: $(x\,,\,|Wx|) = (x, |\bar{x}_j|)_j$

- Scattering-Gaussian model: *Etienne Lempereur*

$$p_\theta(x) = \mathcal{Z}_\theta{}^{-1}\, e^{-(x,|Wx|)\,\theta\,(x,|Wx|)^T}$$

- Wavelet: $W(x\,,\,|Wx|)$ eliminates long range correlations

$p_\theta = \mathcal{Z}_\theta{}^{-1}\, e^{-\tilde{\theta}^T\,\widetilde{\Phi}(x)}$ with $\widetilde{\Phi}(x) = (Wx\,,\,W|Wx|)\,(Wx\,,\,W|Wx|)^T$

and $\tilde{\theta} = W\theta W^T$ of reduced dimension

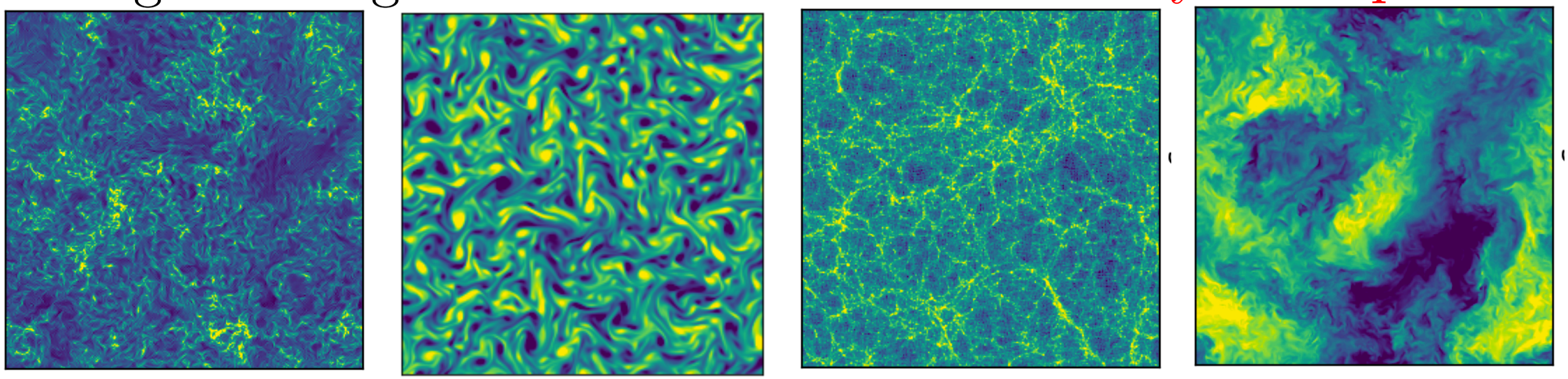
- $\tilde{\theta}$ and $\widetilde{\Phi}(x)$ are reduced to $O(\log^3 d)$ diagonal coefficients:

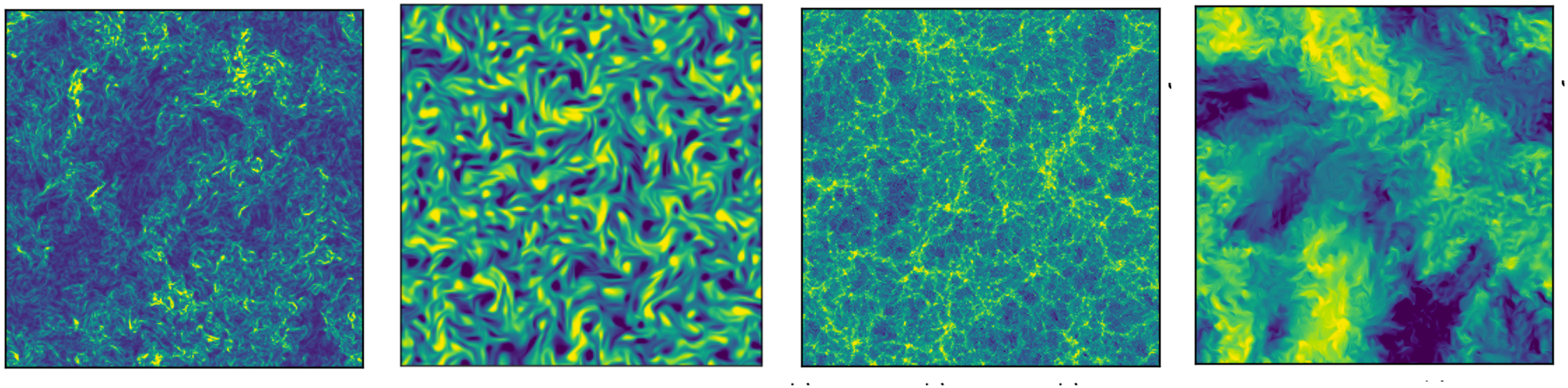
$$\widetilde{\Phi}(x) = \begin{pmatrix} |\bar{x}_j(u)|^2 & \bar{x}_j(u)\,|\bar{x}_{j'}(u)| \\ \bar{x}_j(u)\,|\bar{x}_{j'}(u)| & |\bar{x}_j| \star \psi_k(u)\,|\bar{x}_{j'}| \star \psi_k(u) \end{pmatrix}_{j,j',k}$$

Generation from Scattering Covar.

E. Allys, F. Boulanger, A. Brochard, J. Bruna, S. Chen, B. Ménard, R. Morel, G. Rochette, S. Zhang

Original images of dimension $d = 5\,10^4$: only 1 sample



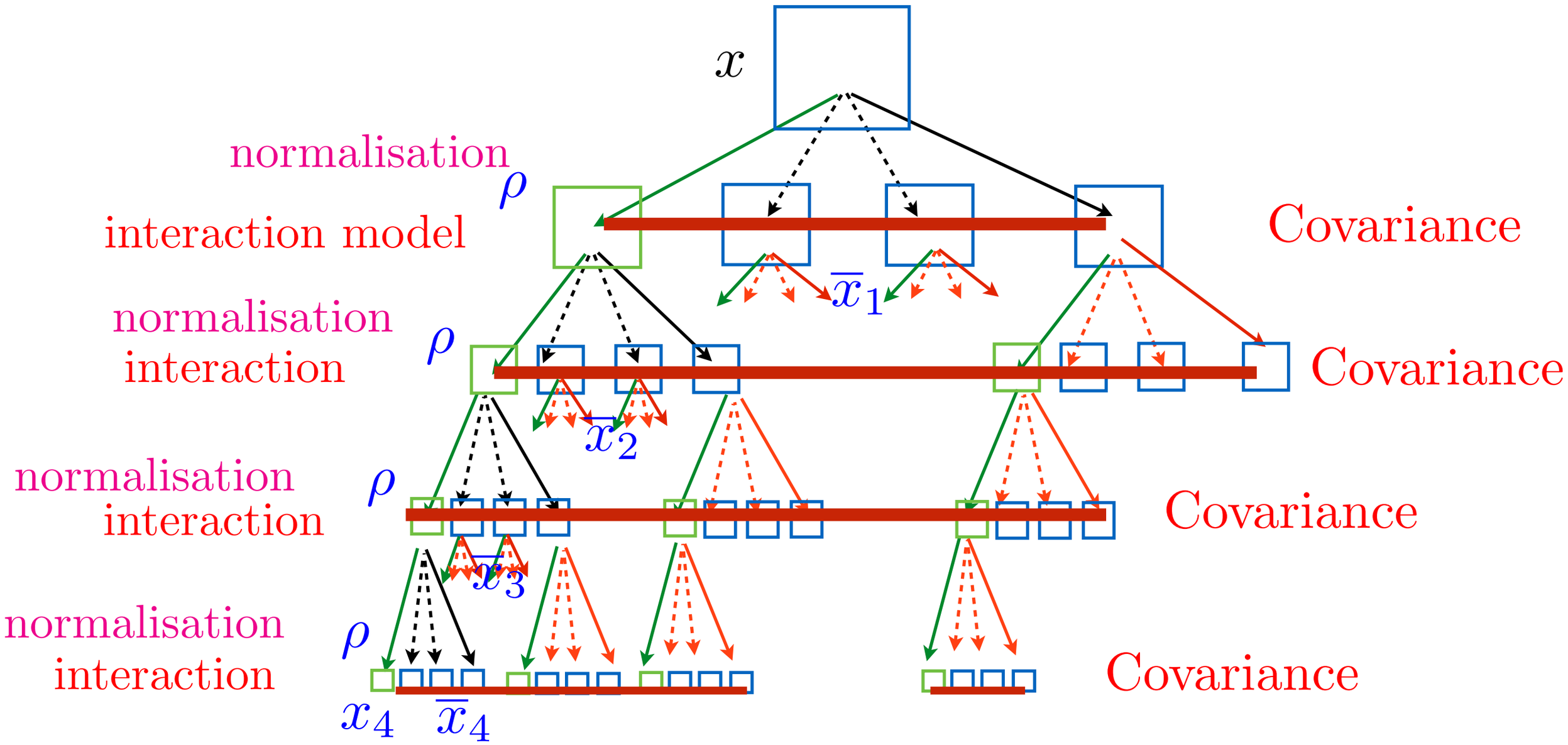


Generated with models having 500 parameters from 1 sample

Reproduces moments of order 3 (bispectrum) and 4 (trispectrum)

Scattering-Gaussian Model

$$W\,\rho\,W x \quad \text{with } \rho(\alpha) = (\alpha\,,\,|\alpha|)$$



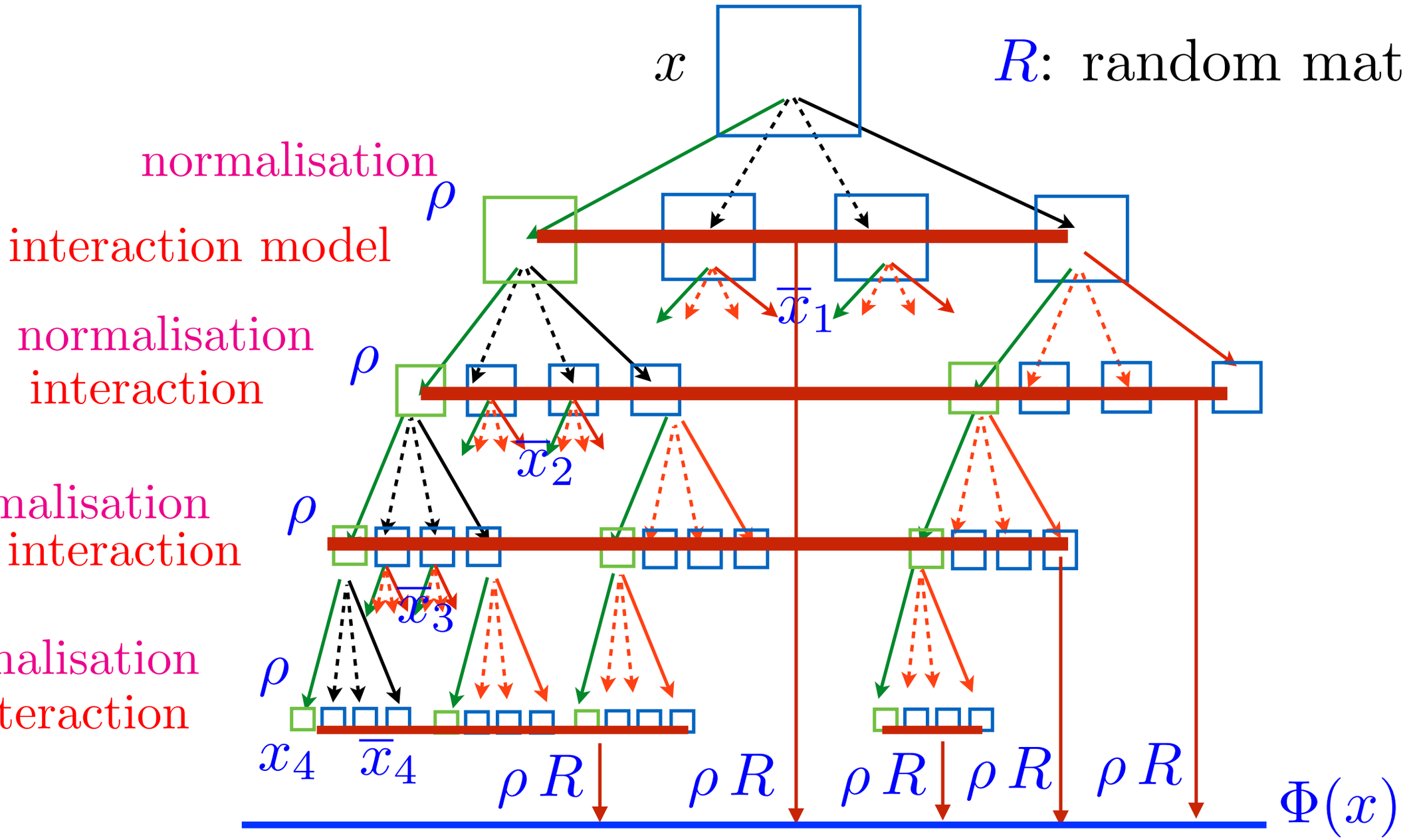
We can replace covariances by random projections and ReLU ρ

A low-dimensional deep networks

Deep Scattering Network

$W\,\rho\,W x$ with $\rho(\alpha) = (\alpha\,,\,|\alpha|)$

R: random matrix



$\Phi(x) = \rho\,R\,W\,\rho W x$: last hidden network layer

$-\log p_\theta(x) = \theta^T \Phi(x) + \log \mathcal{Z}_\theta$: learn only last linear filter

Conclusion

- Build models from interaction energies across scales : wavelet conditional renormalisation group.

- When are interaction energies convex (or nearly) ?

- *Scattering-Gaussian models:* interaction energies captured by wavelet modulus correlations across scales: turbulence, active matter…

Wavelet Conditional Renormalisation Group : Phys. Rev. X,
T. Marchand, M. Ozawa, G. Biroli, S.M.

Conditionally Strongly Log Concave Generative Models, ICML 2023
J. Bruna, E. Lempereur, F. Guth. S. M.

Scattering Spectra Models for Physics, arXiv:2306.17210
S. Cheng, R. Morel, E. Allys, B. Menard, S. M.