Renormalisation Group and Machine Learning

The Wavelet Inverse Renormalisation Group

Giulio Biroli
Ecole Normale Supérieure
Paris

S. Mallat DI-ENS & CdF

M. Ozawa ENS

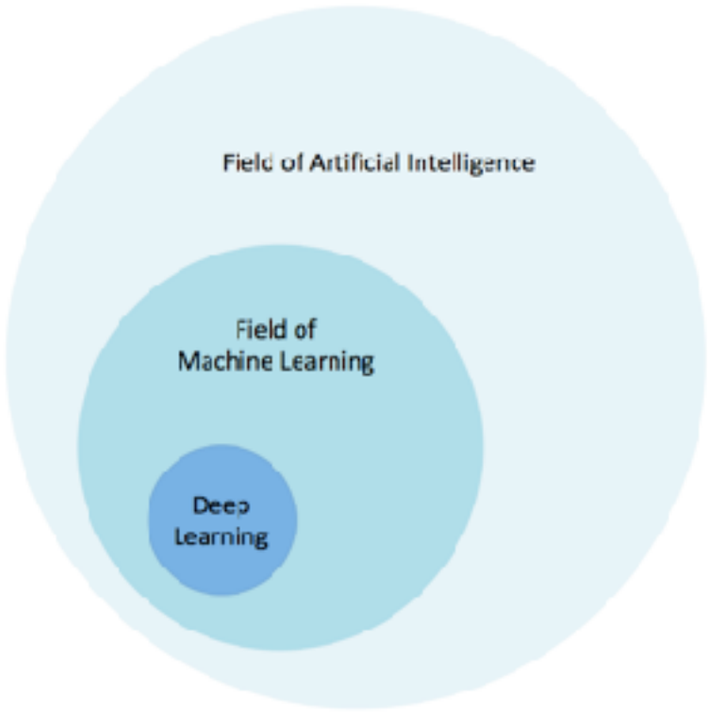
T. Marchand ENS

ENS Département de Physique École normale supérieure

PR[AI]RIE PaRis Artificial Intelligence Research InstitutE

SIMONS FOUNDATION

A revolution in Data Science



Starting from ~2010

- Vision
- Translation
- Text generation
- Self-driving vehicles
- …

Supervised Learning

Generative Models

Models & Data

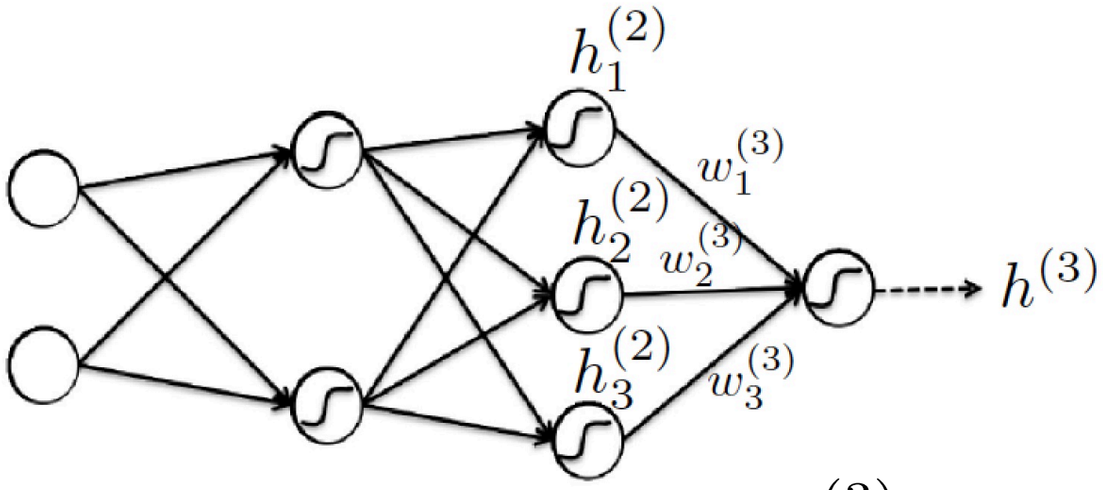
A lot of data & a lot of parameters

Image classification

image ~ 1000 numbers
training data ~ 100000 images
adjustable parameters ~ 100 millions

- Fitting in very high-D with a lot of parameters
- Approximate very high-D probability distributions from a lot of data

The input-output function of deep neural networks is highly non-linear and *hierarchical*



Simple input-output relation through a single neuron

$$h^{(3)} = g(w_1^{(3)} h_1^{(2)} + w_2^{(3)} h_2^{(2)} + w_3^{(3)} h_3^{(2)} - w_4^{(3)})$$

$g(x)$ Simple non-linear function

Questions & Riddles

Escaping the curse of "random landscapes"

Why the dynamics of training lead to good minima?

Escaping the curse of dimensionality

Why learning with so "few parameters" is possible? $1 \ll \mathcal{N} \ll e^D$

Challenges for mathematicians, computer scientists and statistical physicists

- Structure of data & structure of networks
- Towards a constructive approach

  Pioneered by S. Mallat (the scattering transform)

Renormalisation Group & Machine Learning

Renormalisation Group: hierarchical coarse grain of the probability distribution from small to large scale. From small scale properties to large scale physics.

Deep neural networks also work in a hierarchical way and learn small scale structure at the first layers and hierarchically go to global features in the last layer.

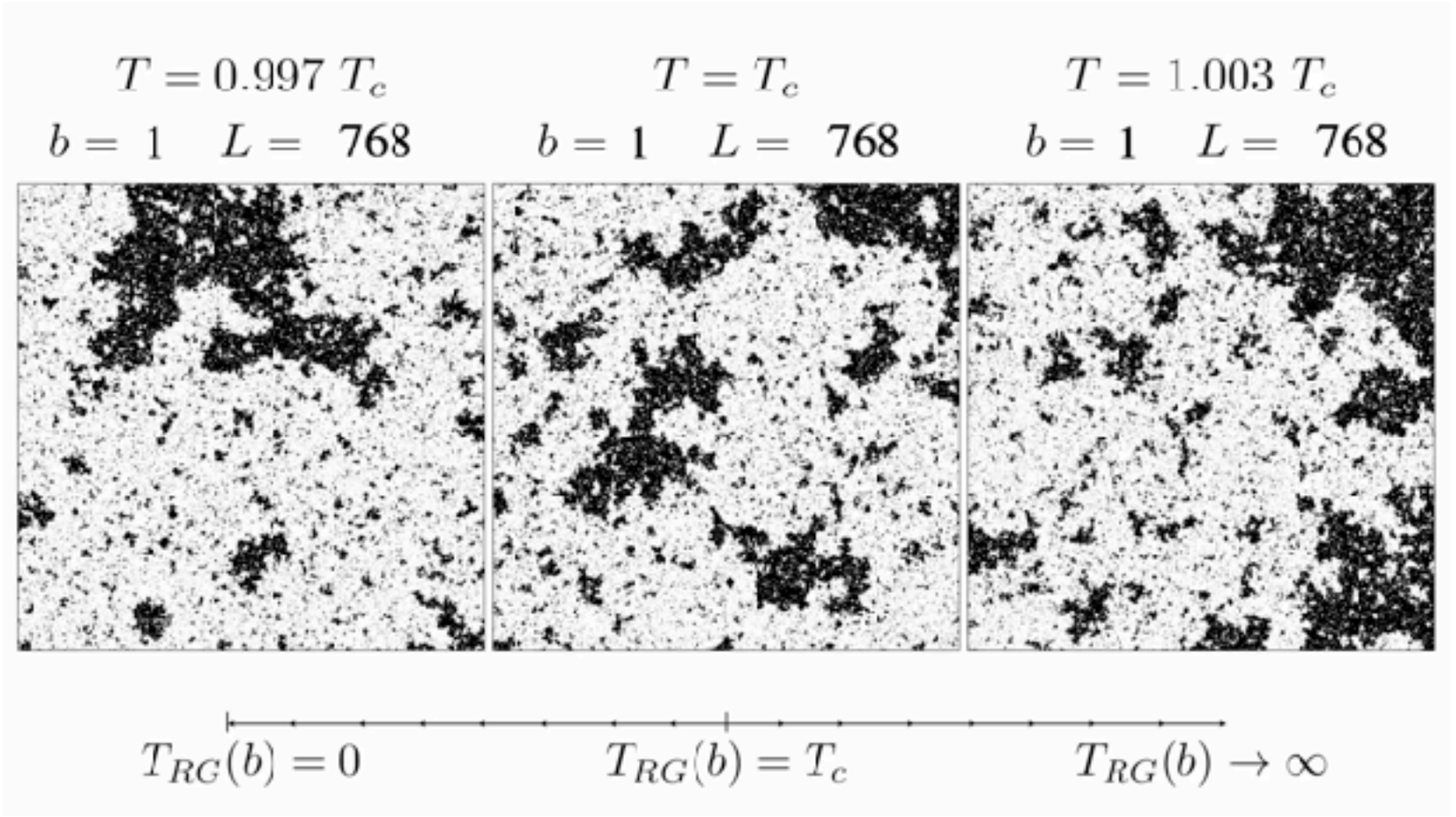
Several works have proposed and investigated this analogy in the last 8 years

Our contribution

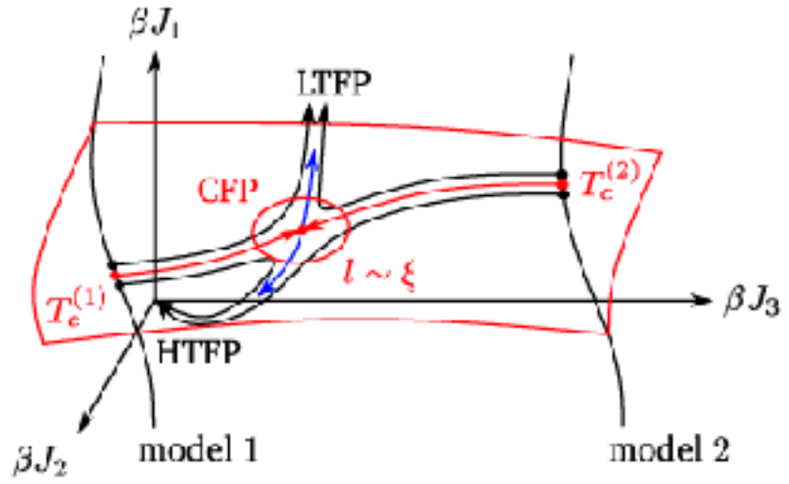
- Focus on images from physical models (statphys & cosmology)
- Using the underlying data structure (from physics) to construct a method - *the wavelet inverse RG* - to approximate the probability distribution
- Address the questions outlined before in a specific context and using RG results as a guide

An intermezzo on renormalisation group

RG for
the Ising Model



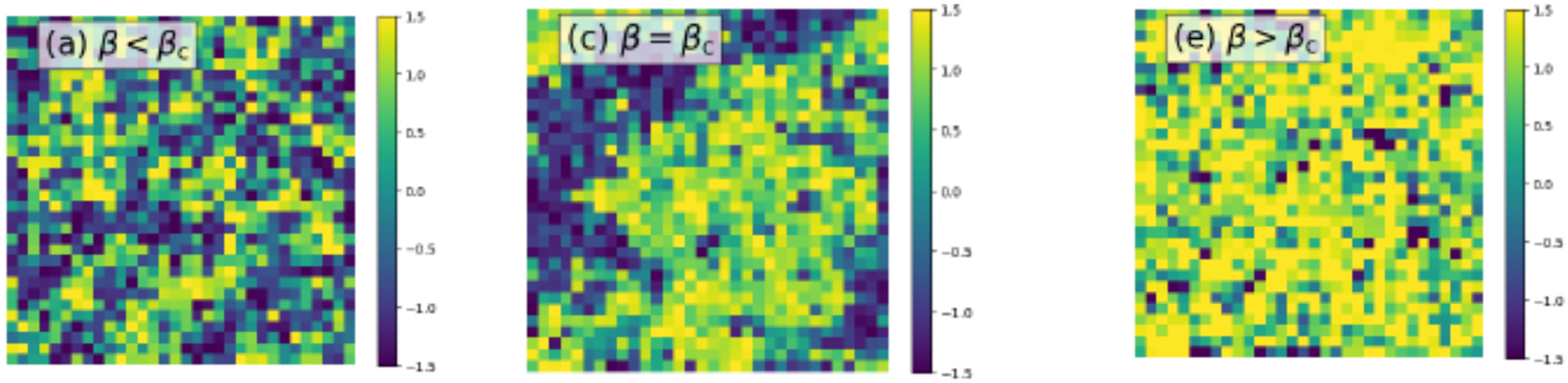
- Integrate out the “fast” (or local) degrees of freedom and rescale
- RG leads to a flow in the space of models (or energy functions)
- Second order phase transition associated to non-trivial fixed points



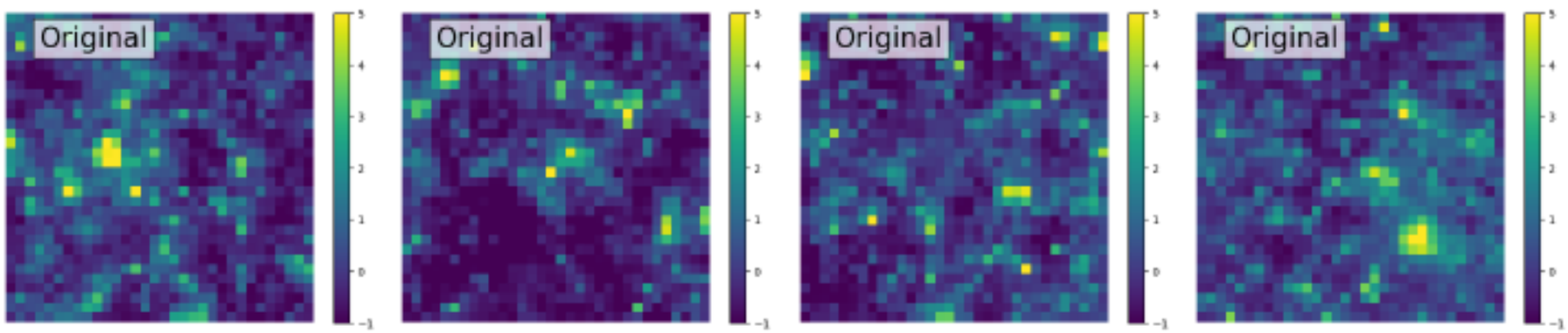
Data and Systems

- Discretized 2D field theory for the ferromagnetic phase transition

$$H(\varphi) = -\frac{1}{2}\sum \varphi(i)\Delta_{ij}\varphi(j) + \sum V(\varphi(i))$$



- Cosmological data (weak lensing maps from the Columbia group)



Data: 32 by 32 images, ~10000-70000 images (~250 parameters)

approximate faithfully the probability distribution
obtain the Hamiltonian

Wavelet Inverse RG

- Reconstruct the probability distribution scale by scale from coarse to fine scales

Usual RG (coarse graining)

$$\ell_{j-1} = 2^{j-1} \rightarrow \ell_j = 2^j$$

Wavelet decomposition

$$\varphi_j(i) = L\varphi_{j-1}(i) \quad \psi_j(i) = G\varphi_{j-1}(i)$$

RG: Integrate out the “fast” microscopic variables

Inverse RG

$$\ell_j = 2^j \rightarrow \ell_{j-1} = 2^{j-1}$$

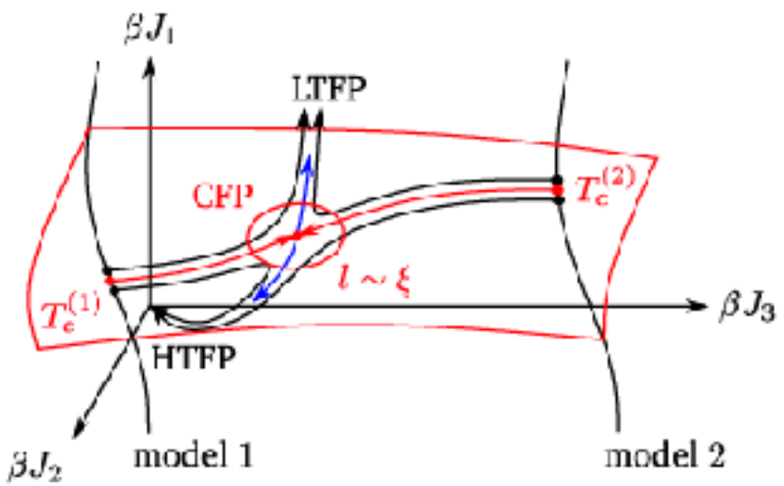
$$\varphi_{j-1}(i) = L^T\varphi_j(i) + G^T\psi_j(i)$$

Inverse RG: add the “fast” microscopic variables and their distribution (and rescale)

One step of WI-RG: $$p^A_{j-1}(\varphi_{j-1}) = p^A_j(\psi_j|\varphi_j)p^A_j(\varphi_j)$$

known

estimate from data



Model for the conditional probability

Usual RG flow: model (or Ansatz) for the energy function over which "project" the flow

$$\mathcal{H}_{j-1}(\varphi_{j-1}) = -\frac{1}{2}\sum \varphi_{j-1}(i) K^A_{j-1}(i-i')\varphi_{j-1}(i') + \sum V^A_{j-1}(\varphi_{j-1}(i)) \quad \mathcal{H}_{j-1} \to \mathcal{H}_j$$

$$p_j(\psi_j|\varphi_j) = p_{j-1}(\varphi_{j-1})/p_j(\varphi_j) \simeq e^{-\mathcal{H}_{j-1}(\varphi_{j-1})+\mathcal{H}_j(\varphi_j)}$$

$$p_j(\psi_j|\varphi_j) \simeq e^{-(\mathcal{H}_{j-1}(\varphi_{j-1})-\mathcal{H}_{j-1}(\varphi_{j-1})|_{\psi_j=0})-\mathcal{H}_{j-1}(\varphi_{j-1})|_{\psi_j=0}+\mathcal{H}_j(\varphi_j)}$$

Model $\quad p^A_j(\psi_j|\varphi_j) = e^{-\overline{\mathcal{H}}_{j-1}(\varphi_{j-1})+F_j(\varphi_j)}$

$$\overline{\mathcal{H}}_{j-1}(\varphi_{j-1}) = \mathcal{H}_{j-1}(\varphi_{j-1}) - \mathcal{H}_{j-1}(\varphi_{j-1})|_{\psi_j=0} \quad F_j(\varphi_j) = -\ln \int d\psi_j e^{-\overline{\mathcal{H}}_{j-1}(\varphi_{j-1})}$$

Model for the conditional probability

Usual RG flow: model (or Ansatz) for the energy function over which "project" the flow

$$\mathcal{H}_{j-1}(\varphi_{j-1}) = -\frac{1}{2}\sum \varphi_{j-1}(i) K^A_{j-1}(i-i')\varphi_{j-1}(i') + \sum V^A_{j-1}(\varphi_{j-1}(i)) \quad \mathcal{H}_{j-1} \to \mathcal{H}_j$$

Model $\quad p^A_j(\psi_j|\varphi_j) = e^{-\overline{\mathcal{H}}_{j-1}(\varphi_{j-1}) + F_j(\varphi_j)}$

$$\overline{\mathcal{H}}_{j-1}(\varphi_{j-1}) = \mathcal{H}_{j-1}(\varphi_{j-1}) - \mathcal{H}_{j-1}(\varphi_{j-1})|_{\psi_j=0} \qquad F_j(\varphi_j) = -\ln \int d\psi_j e^{-\overline{\mathcal{H}}_{j-1}(\varphi_{j-1})}$$

estimate from data

All is needed to get a generative model

$$p^A(\varphi) = \left[\prod p^A_j(\psi_j|\varphi_j)\right] p^A_J(\varphi_J)$$

To reconstruct the microscopic energy $\quad \mathcal{H} = \sum \left(\overline{\mathcal{H}}_{j-1} - F_j\right) + \mathcal{H}_J$

Model $\quad F_j(\varphi_j) = -\mathcal{H}_{j-1}(\varphi_{j-1})|_{\psi_j=0} + \mathcal{H}_j(\varphi_j) + \sum \tilde{V}^A_j(\varphi_j(i))$

New term needed for cosmological data

$$\mathcal{H}(\varphi) = -\frac{1}{2}\sum \varphi(i) K^A_0(i-i')\varphi(i') + \sum V^A_0(\varphi(i)) + \sum_j \sum_i \tilde{V}_j{}^A(\varphi_j(i))$$

Estimation from data

$$p_j^A(\psi_j|\varphi_j) = e^{-\overline{\mathcal{H}}_{j-1}(\varphi_{j-1})+F_j(\varphi_j)} \longrightarrow p_j^A(\psi_j|\varphi_j) = e^{-\sum_l g_l^{(j)} U_l(\varphi_{j-1})+F_j(\varphi_j)}$$

To obtain the parameters: minimise the Kullback-Leibler divergence $D_{KL}(p_j^{true}||p_j^A)$

Gradient descent: $g_l^{(j)}(t+1) - g_l^{(j)}(t) = \eta\left(\langle U_l(\phi_{j-1})\rangle_{p_j^A} - \langle U_l(\phi_{j-1})\rangle_{p_j^{true}}\right)$

From Montecarlo simulation

From data

To obtain the approximated free-energy: regress the true free-energy to the model one

$$F_j(\varphi_j) = -\ln \int d\psi_j e^{-\overline{\mathcal{H}}_{j-1}(\varphi_{j-1})} \xrightarrow[\text{regression}]{} F_j^A(\varphi_j)$$

From Montecarlo simulation

Key aspect: always work on the "fast" degrees of freedom (wavelet field)

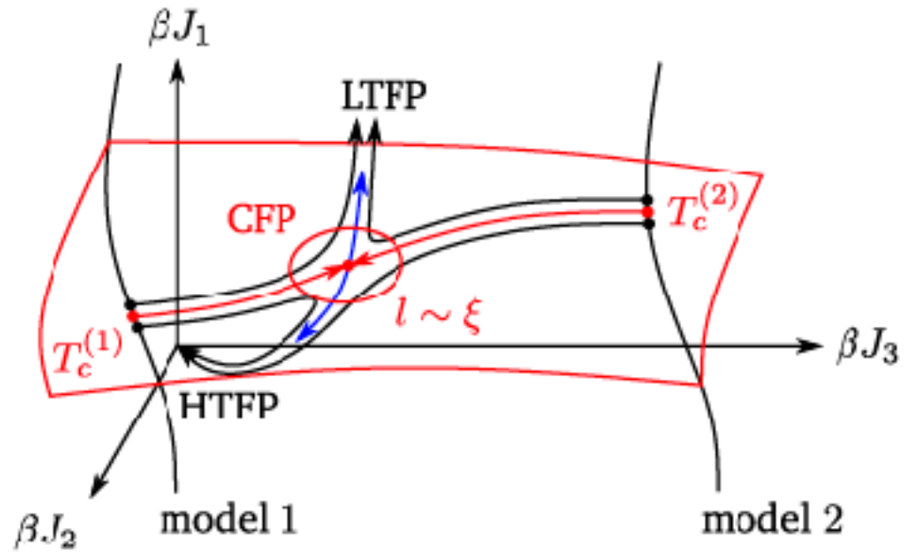
- Estimate the conditional probability scale by scale
- Montecarlo simulation only on wavelet field
- Determine the couplings associated to the wavelet field

→ Fast
Well-conditioned

Rigorous analysis in the Gaussian case

Similarities & differences with RG

- Reverse RG flow: from coarse to fine constructing the model for the microscopic probability distribution scale by scale



- As in RG treatment: always work on "fast" degrees of freedom at the scale at hand. Crucial to get a fast and stable method.

- The Ansatz is on the conditional probability: contrary to RG where is on the energy function at a given scale. Open the way to more general expressions for the energy, and to treat more general images.

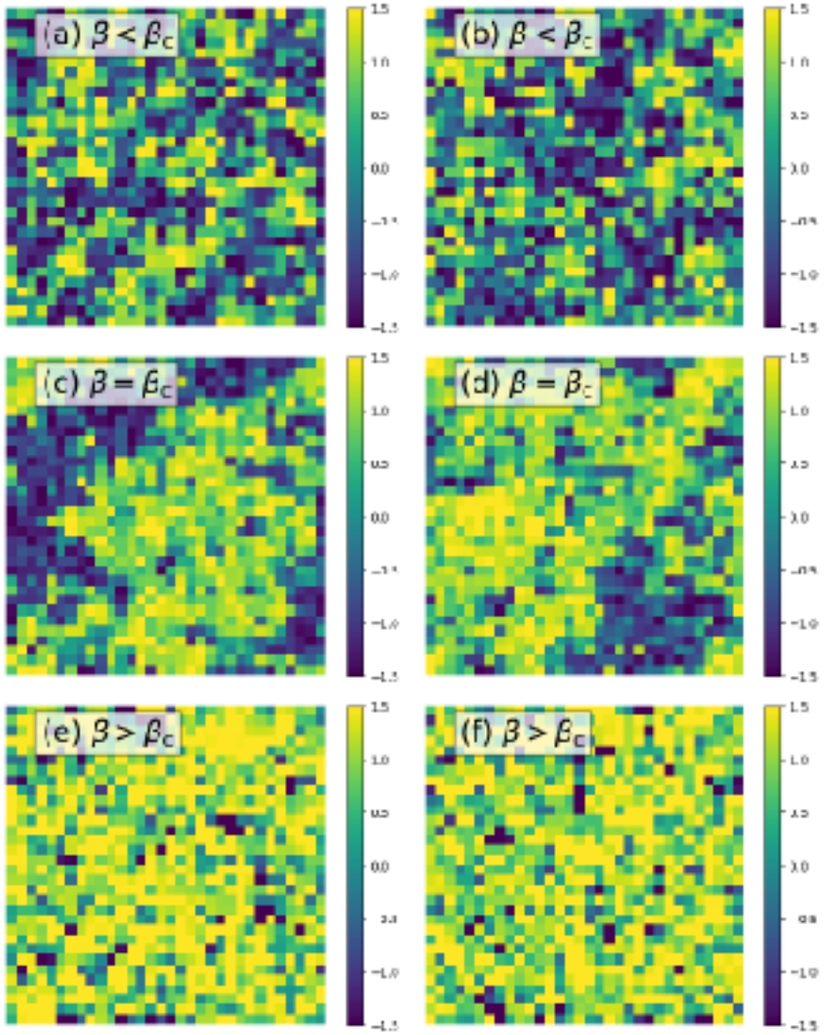
Numerical Applications I

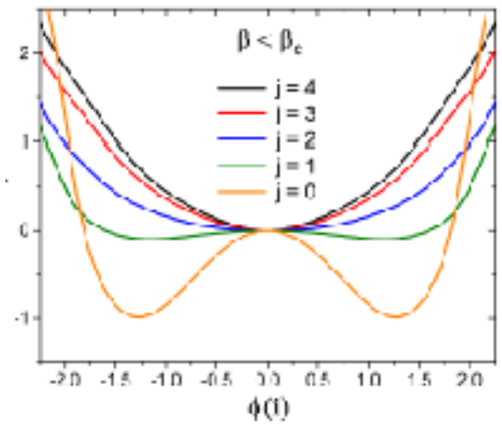
Discretized 2D field theory for the ferromagnetic phase transition

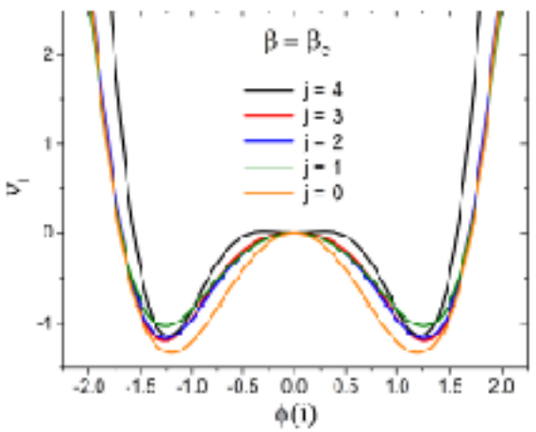
True

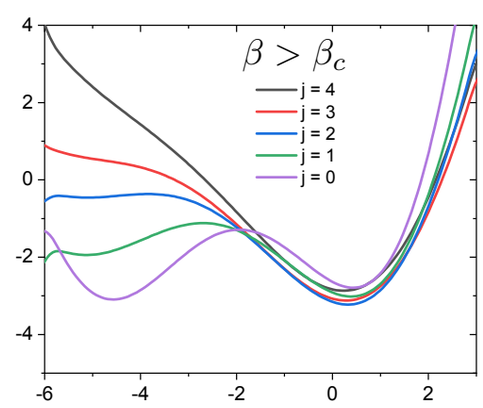
Generated

Inverse RG Flow

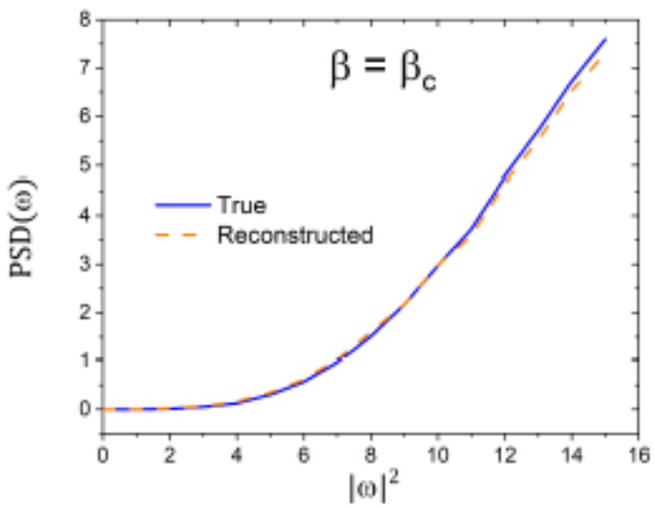


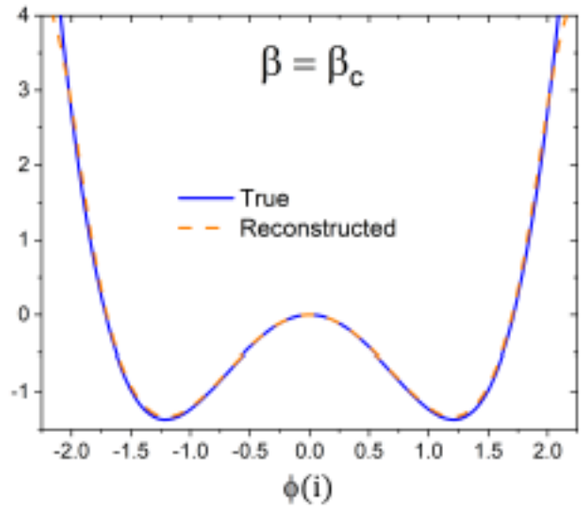






Comparison true & reconstructed

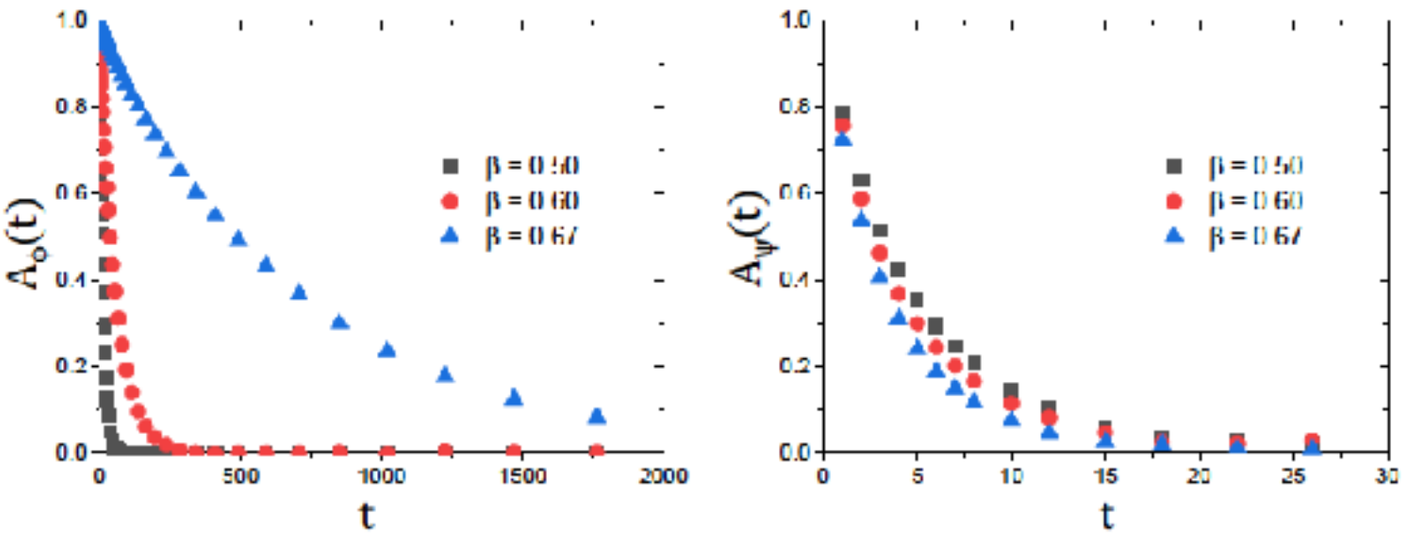




Numerical Applications I

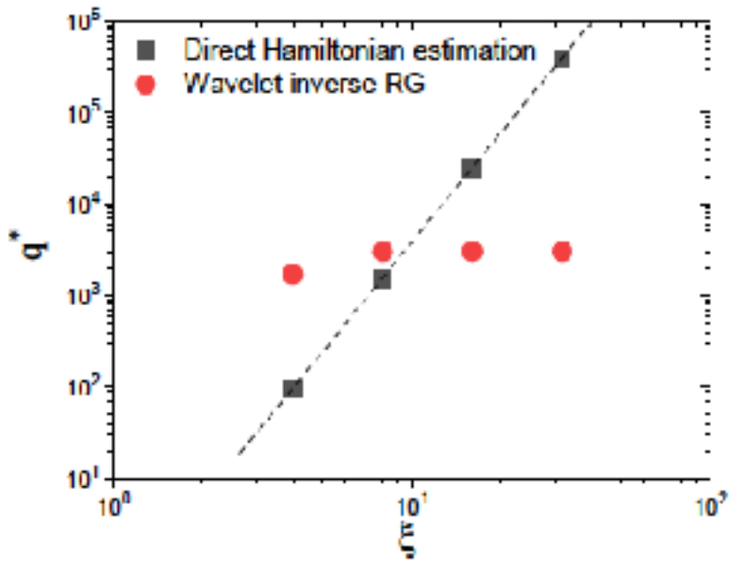
Discretized 2D field theory for the ferromagnetic phase transition

Critical slowing down avoided
Montecarlo on wavelet field is fast at each scale



Normalized time-dependent correlation functions

Estimation by gradient descent is numerically stable

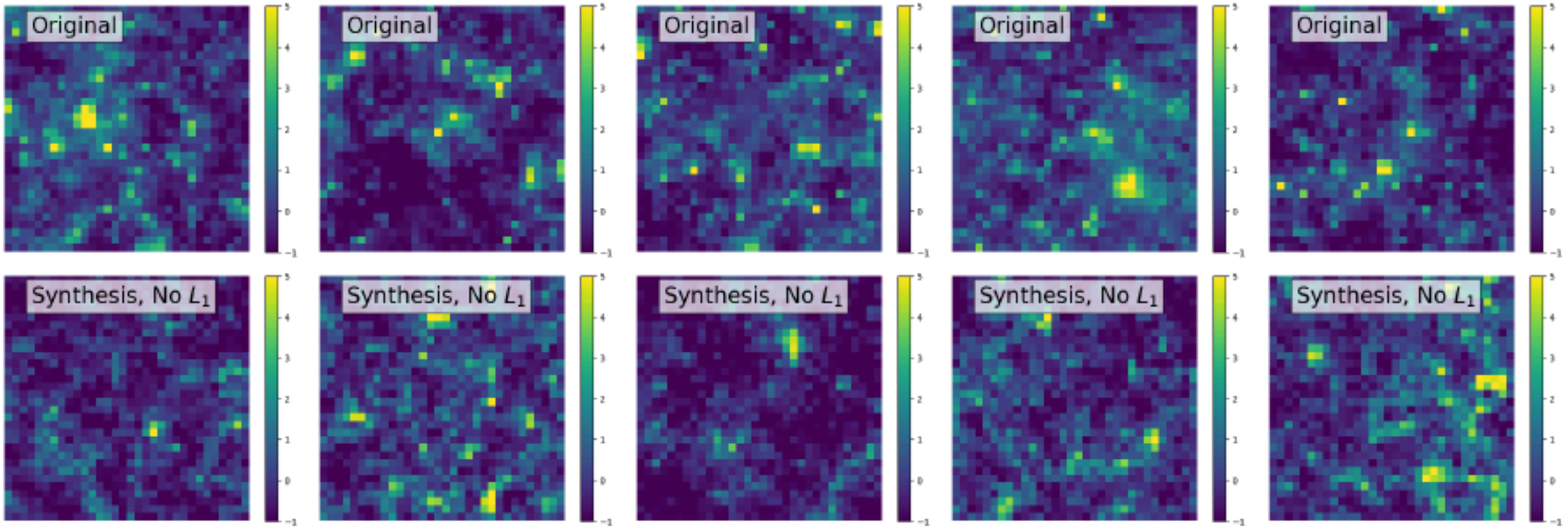


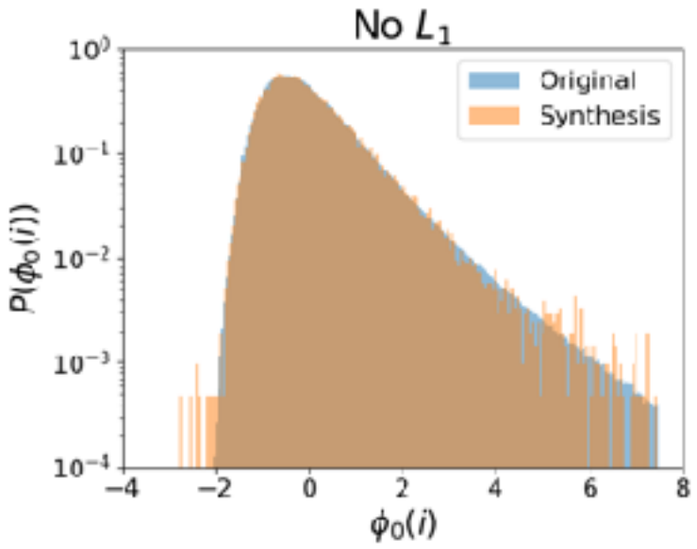
Iterations to reach a prescribed accuracy

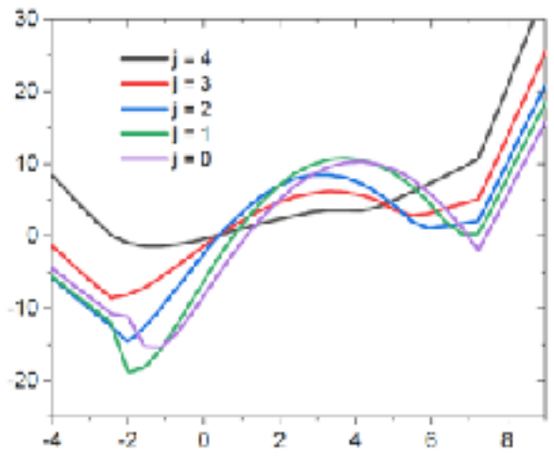
- The Wavelet Inverse RG is a well-conditioned method
- It can generates typical samples in times NlogN even at criticality

Numerical Applications II

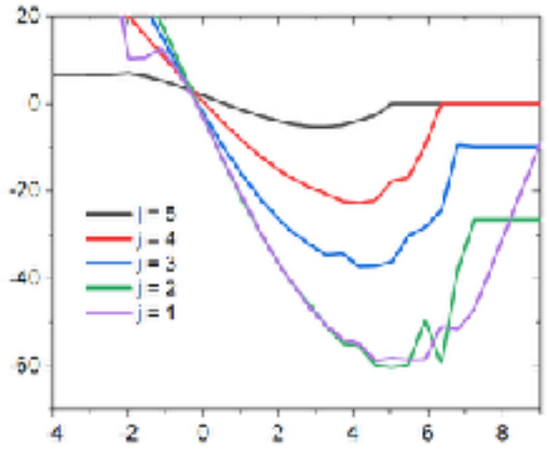
- Cosmological data (weak lensing maps from the Columbia group)







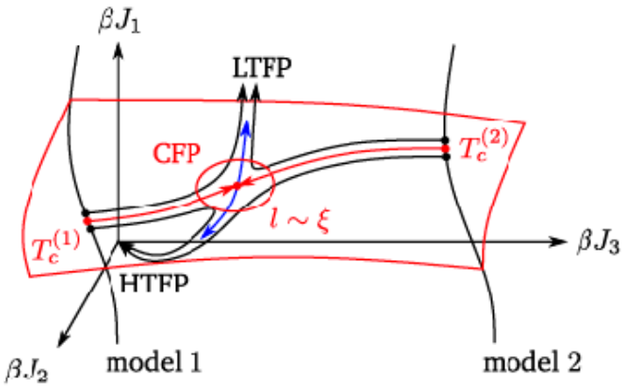
$V_j^A(\varphi_j)$



$\tilde{V}_j^A(\varphi_j)$

Remarks & Perspectives

- *Wavelet Inverse RG*. Using the property of the data and locality scale by scale



Hierarchical method, avoiding all the “curses”

Fast and stable estimation and with not exponentially many parameters

- Data lie in a restricted part of the space, and WIRG only approximates the probability distribution there.

See for NNs Goldt, Mézard, Krzakala, Zdeborová, PRX 2020.

- Key ingredient: model for the conditional probability. More general Ansatz are needed for more structured images (cf. Mallat’s scattering transform).
  Next challenge: images from out of equilibrium processes from statistical physical.