Revisiting the training of RBMs

Beatriz Seoane

Theoretical Physics, UCM Madrid

UNIVERSIDAD COMPLUTENSE MADRID

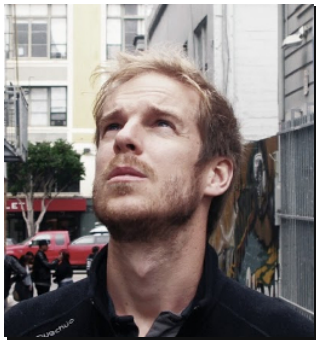
Comunidad de Madrid

talento cm Programa de atracción de talento investigador Comunidad de Madrid

Beatriz Seoane

Theoretical Physics, UCM Madrid

In collaboration with



Aurélien Decelle

(UCM)

Cyril Furtlehner

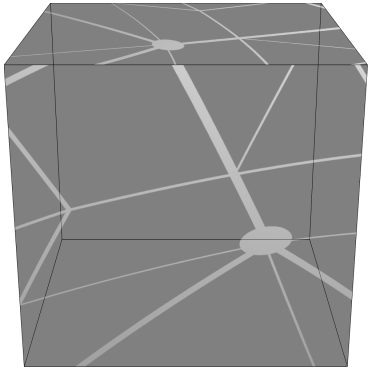
(Tau team, Université Paris-Saclay)

Nicolas Bereux

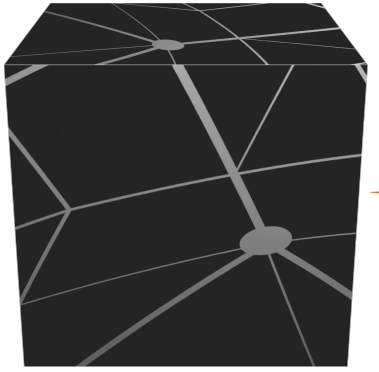
Student Paris-Saclay

Introduction : generative approach

training



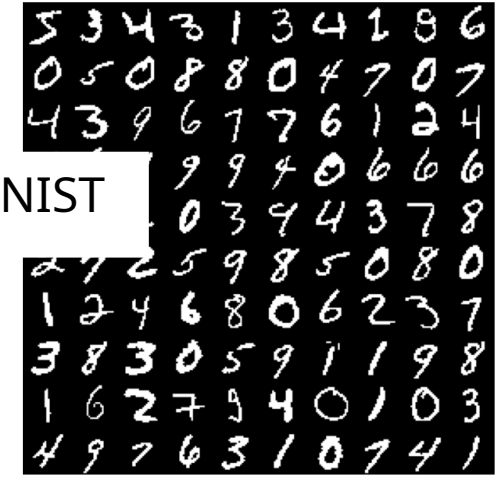
generating



- **Energy based models (RBMs, Generative Convnets)**
- Generative Adverarial Network (GAN)
- Variational AutoEncoder (VAE)

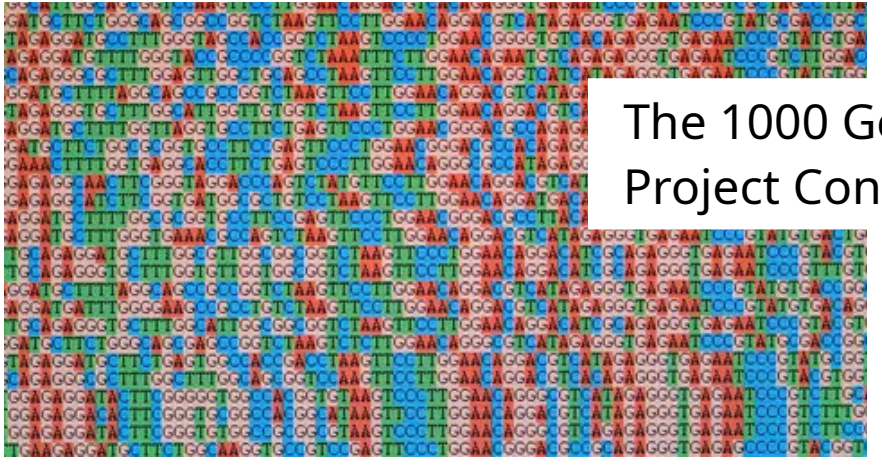
Introduction : generative approach

MNIST

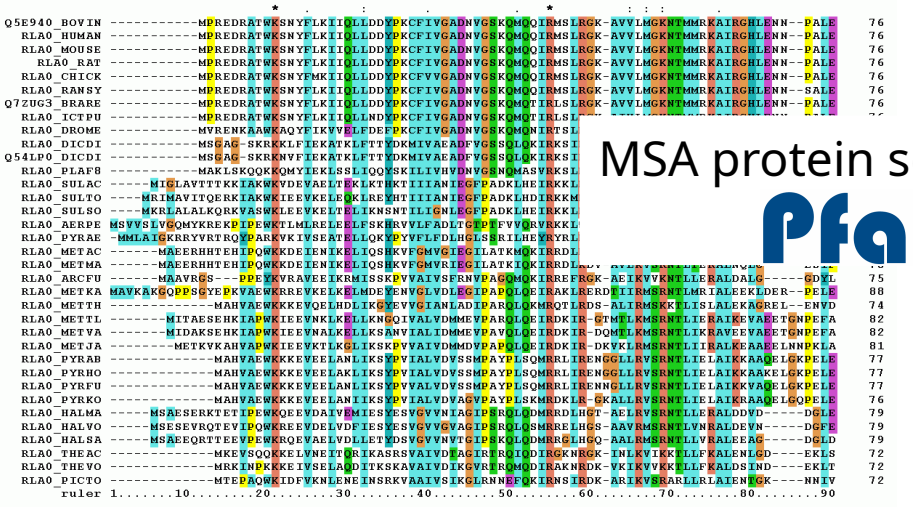


CELEBA

The 1000 Genomes Project Consortium



MSA protein sequences



Pfam

Energy based models (EBMs)

- Assign an energy $E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})$ → Unnormalized probability

$$p(\boldsymbol{v}, \boldsymbol{h}) \propto e^{-E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}$$

Observed

latent

parameters

- Learning : adjust the parameters so that the dataset configurations are typical configurations of the model.
- Maximize the likelihood:

$$L = \prod_{m=1}^{M} p\left(v^{(m)}\right)$$

Gradient ascent $\nabla_\theta L$

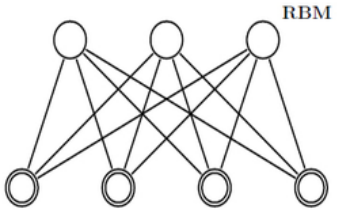
- Problem $Z = \sum_{\boldsymbol{v}, \boldsymbol{h}} p(\boldsymbol{v}, \boldsymbol{h})$ Is in generally impossible to compute exactly

We need Monte Carlo Sampling

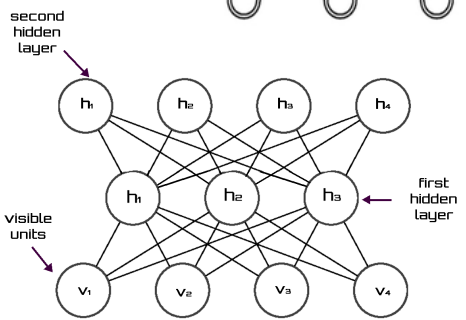
Energy based models (EBMs)

$$E(v, h; \theta)$$

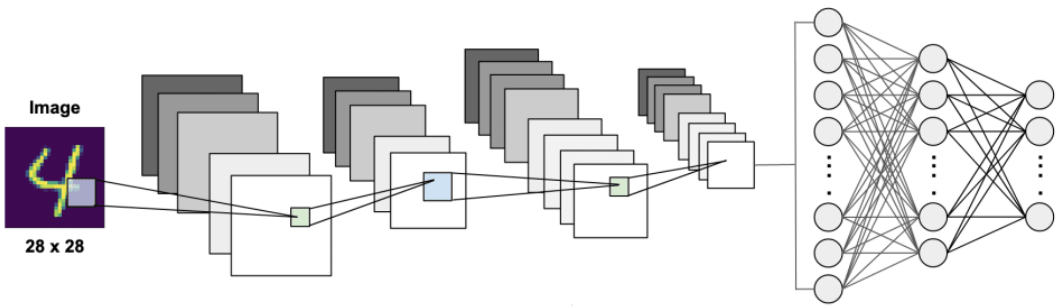
- **Restricted Boltzmann Machines**



- Deep Boltzmann Machines



- Generative ConvNets



Why Restricted Boltzmann Machines (RBMs)?

In this talk we focus on RBM for the following reason:

- It is an *Ising model*: a canonical model for statistical physicists
- We can write explicitly the probability distribution which give us tools to study it
- It relies on a simple shallow neural network that can be looked into: can it be used to extract dataset features (interpretability) ?
- It can model complex dataset : it can be shown in the binary case that it can overfit anything…

Reasons to not use it:

- The training can be (very) capricious for complicated reasons
- Generate new data can be long … (How long?)
- So far, no implementation of a convolutional RBM has been made usable in practice
- …

RBM (biased) Historical events

Machine Learning aspects

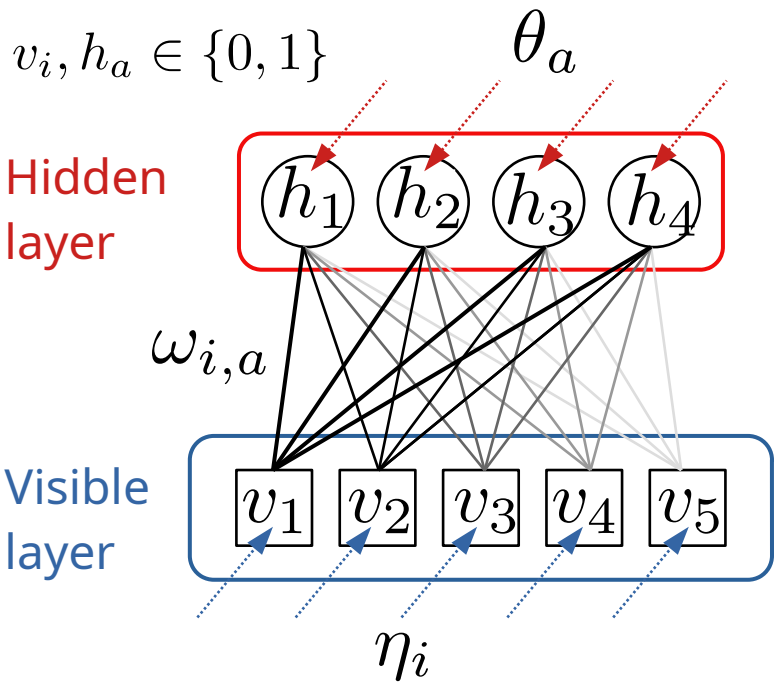
- It was introduced by Smolenski and popularized by Hinton ~80/90
  → introducing the Contrastive Divergence it was proved to be “practical”
- It was getting popular as a pre-training tools for deep-network
- Then, the interest for RBMs slowly decreases around ~2010
- The rise of GAN/VAE alternative achieve to out-faschion RBM

Statistical Physics aspects

- Many works on RBM in ~2010 anaylising the phase diagram of RBM
- Works on the learning dynamics
- Works on message-passing learning algorithm
- …

Restricted Boltzmann Machine

Smolensky (1989)



Energy of a configuration

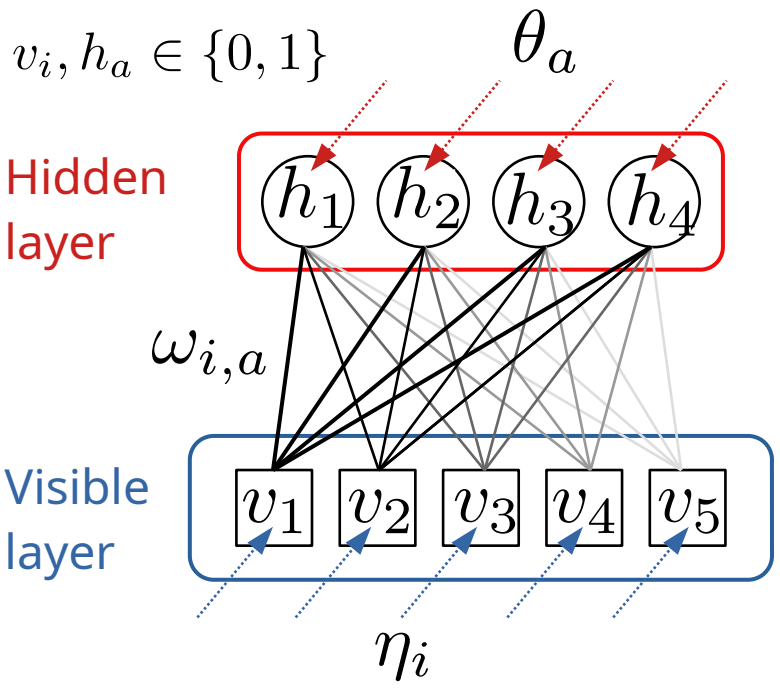
$$E[\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\theta}] = -\sum_{ia} v_i w_{ia} h_a - \sum_i \eta_i v_i - \sum_a \theta_a h_a$$

Visible : **data**

Hidden : “Neurons” → **features extracted**
(interpretability!)

Restricted Boltzmann Machine

Smolensky (1989)



Energy of a configuration

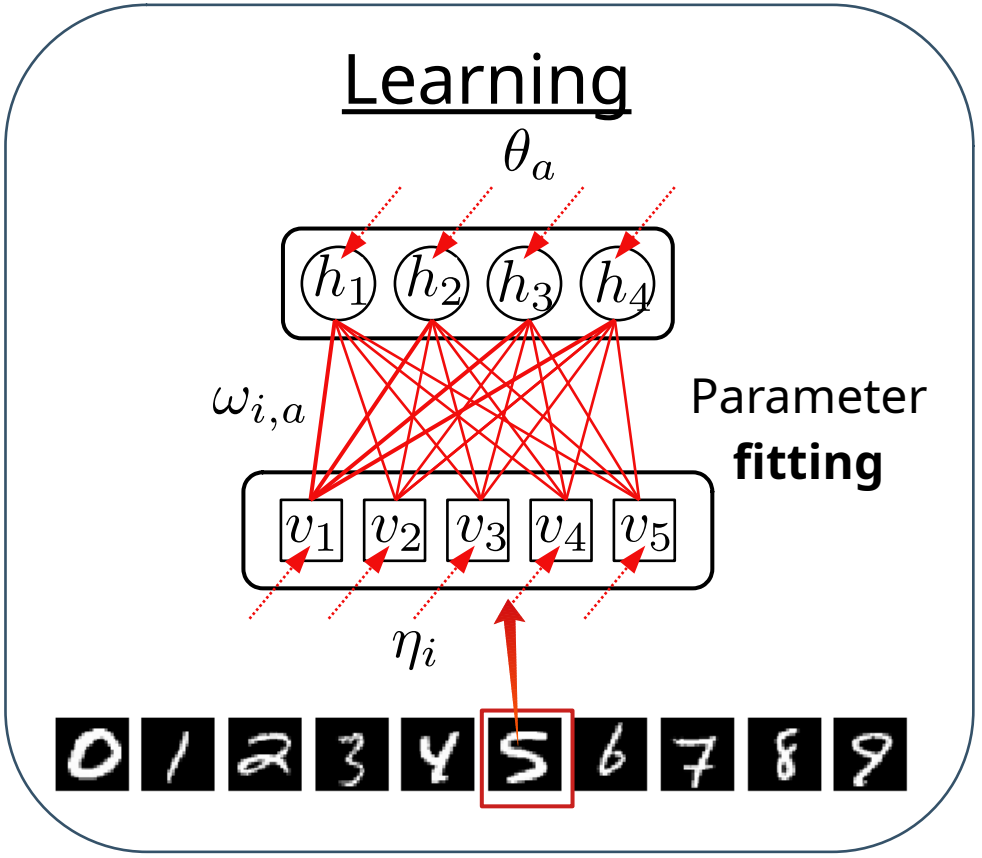
$$E[\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\theta}] = -\sum_{ia} v_i w_{ia} h_a - \sum_i \eta_i v_i - \sum_a \theta_a h_a$$

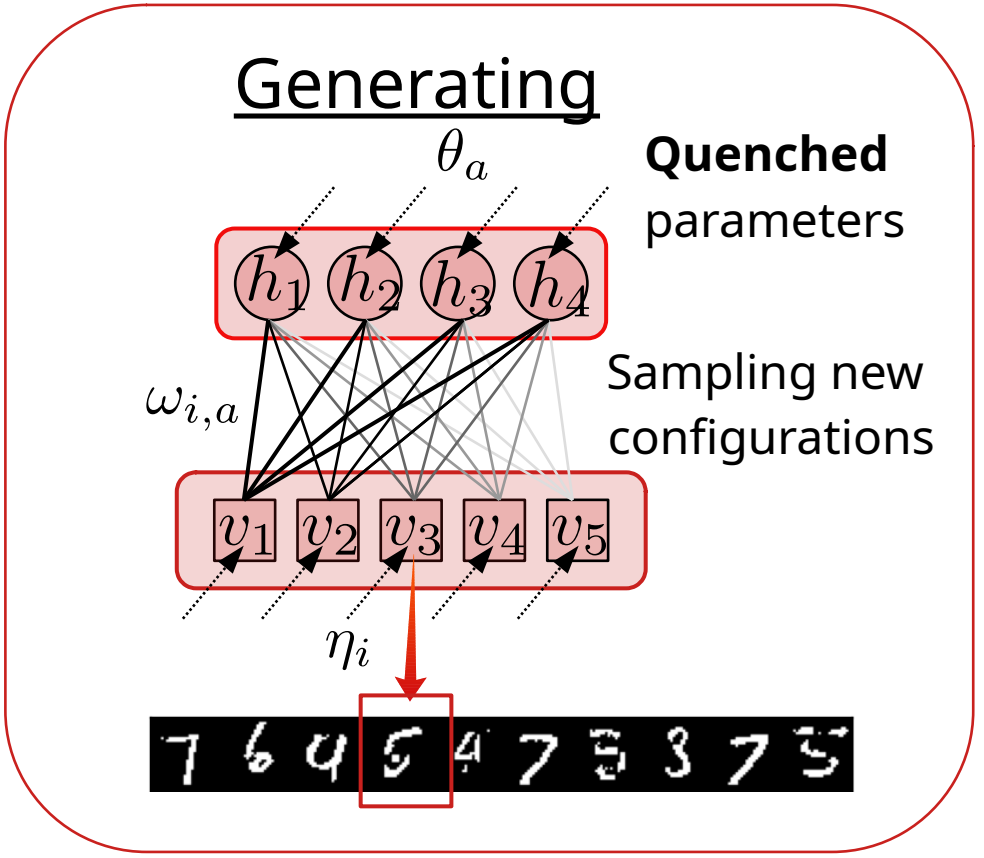
Visible : **data**

Hidden : “Neurons” → **features extracted**
(interpretability!)

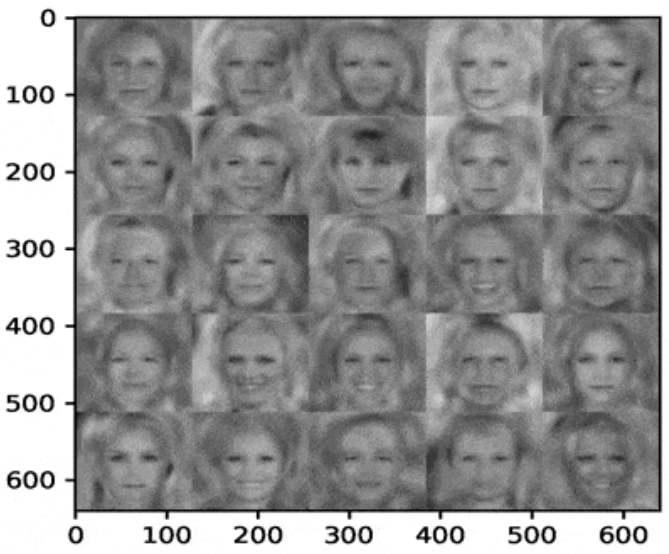
Sequencing context: Tubiana, Cocco, Monasson, *Elife (2019)*; Shimagaki, Weigt, *PRE (2019)*, Yelmen *et al.*, *PLoS genetics (2021)* ; *Bravi* et al. *PloS CB (2021); Bravi* et al. *Cell Systems (2021);* ⇒ *Tubiana’s talk*

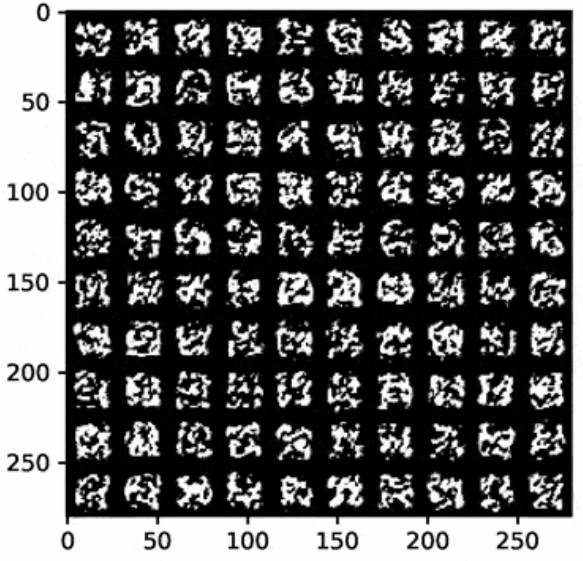
Training vs. Sampling and RBM





Gibbs sampling of a trained RBM





Learning an RBM

- Gibbs equilibrium distribution

$$p[\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\theta}] = \frac{\exp(-E[\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\theta}])}{Z} \quad \text{with } Z = \sum_{\{\boldsymbol{v}, \boldsymbol{h}\}} e^{-E[\boldsymbol{v}, \boldsymbol{h}]}$$

- Dataset $S = \{\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(M)}\}$ are the **typical samples** of *p(v)*

- Maximize the log-likelihood $\mathcal{L}(\boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\theta}|S) = \sum_{m=1}^{M} \ln p(\boldsymbol{v} = v^{(m)}|\boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\theta})$

- Gradient ascent $\frac{\partial \mathcal{L}}{\partial w_{ia}} = \langle v_i h_a \rangle_{\mathcal{D}} - \langle v_i h_a \rangle_{\mathcal{H}}$

$$\frac{\partial \mathcal{L}}{\partial \eta_i} = \langle v_i \rangle_{\mathcal{D}} - \langle v_i \rangle_{\mathcal{H}} \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial \theta_a} = \langle h_a \rangle_{\mathcal{D}} - \langle h_a \rangle_{\mathcal{H}}$$

On the interpretability (I)

- One can extract the point correlations of our data up to any order !

Once we integrate out the hidden variables: $E(v;\theta) = -\log\left(\sum_{\boldsymbol{h}} e^{-E[\boldsymbol{v},\boldsymbol{h};\boldsymbol{w},\boldsymbol{\eta},\boldsymbol{\theta}]}\right)$

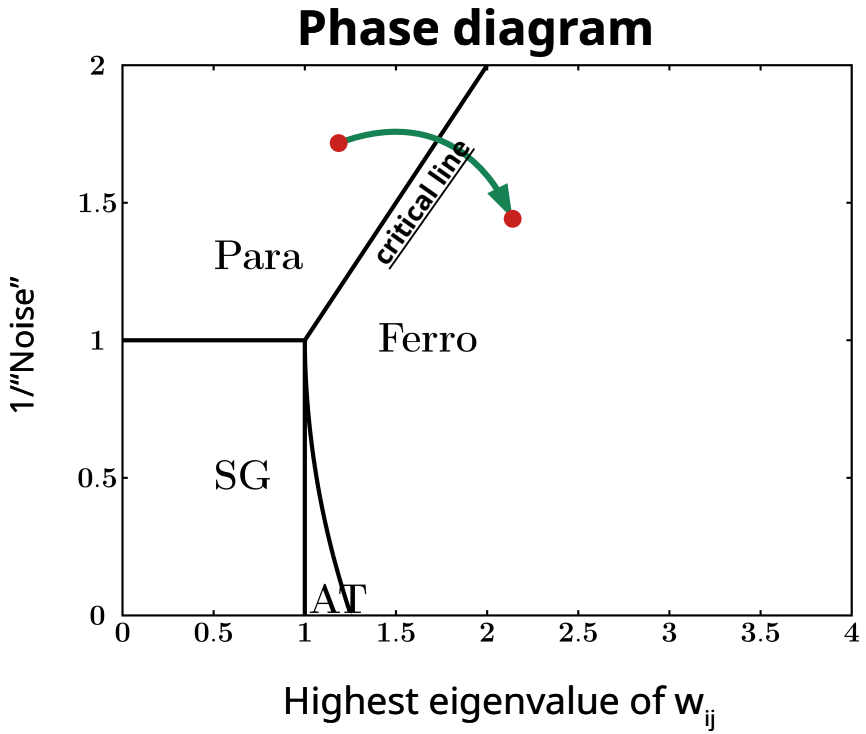
$$E(v) = -E_0 - \sum_i H_i v_i - \sum_{i,j} J^{(2)}_{i,j} v_i v_j - \sum_{i,j,k} J^{(3)}_{ijk} v_i v_j v_k \cdots - \sum_{j_1 \cdots j_n} J^{(n)}_{j_1 \cdots j_n} v_{j_1} \cdots v_{j_n} - \cdots$$

Effective model for the problem

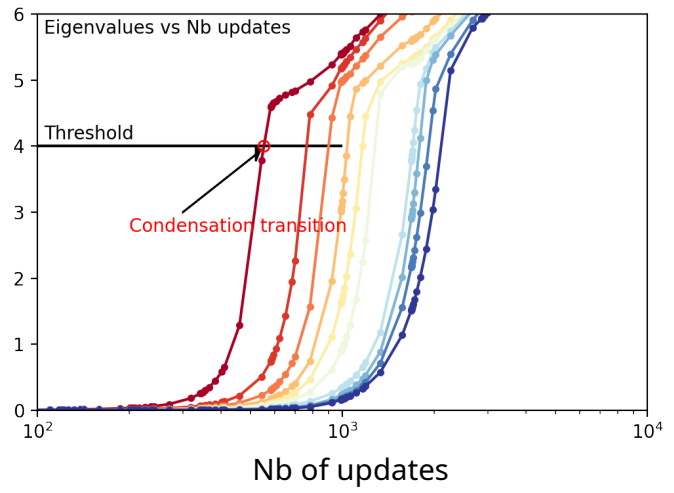
On the interpretability (II)

- It can be shown that when the RBM starts to learn features of the data, the system suffers a **phase order transition** from a paramagnetic phase to a ferromagnetic phase

[Decelle, Fissore, Furtlehner J. of stat phys (2018)]



Para: high temperature (low variance)
Ferro: strong eigenmode - low noise

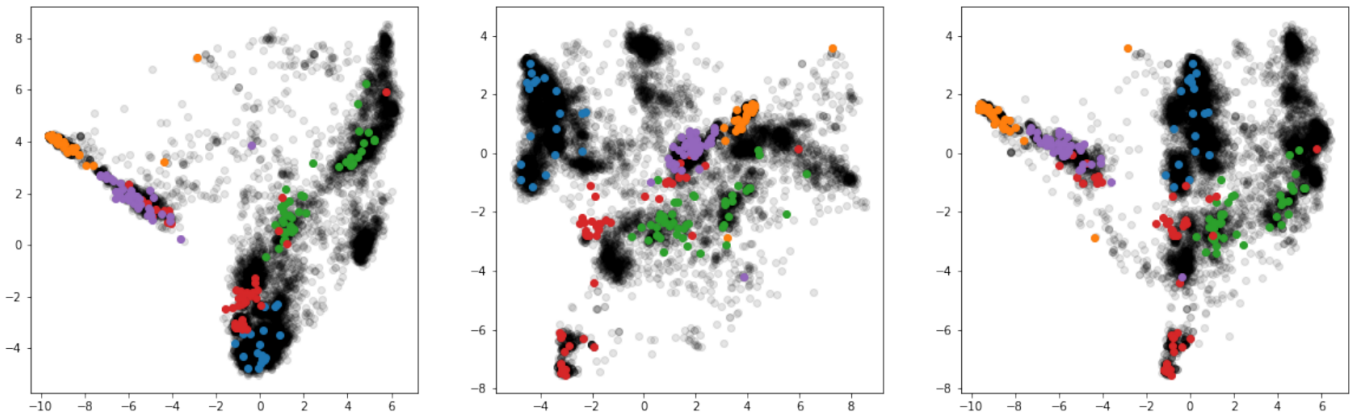


On the interpretability (II)

- The eigenvectors $\boldsymbol{w}_i$ of matrix $\boldsymbol{\omega}$ align with the important directions of the dataset:

$$m_i = \langle \boldsymbol{v} \cdot \boldsymbol{w}_i \rangle_{\mathcal{D}} \neq 0$$

[Decelle, Fissore, Furtlehner J. of stat phys (2018)]



CPF protein

Work in collaboration with R. Vanderhaegen, A. Decelle, A. Carbone

On the gradient

$$\langle f(\boldsymbol{v},\boldsymbol{h})\rangle_{\mathcal{D}} = M^{-1}\sum_{m}\sum_{\{\boldsymbol{h}\}} f(\boldsymbol{v}^{(m)},\boldsymbol{\tau})p(\boldsymbol{h}|\boldsymbol{v}^{(m)})$$

$$\frac{\partial \mathcal{L}}{\partial w_{ia}} = \langle v_i h_a\rangle_{\mathcal{D}} - \langle v_i h_a\rangle_{\mathcal{H}}$$

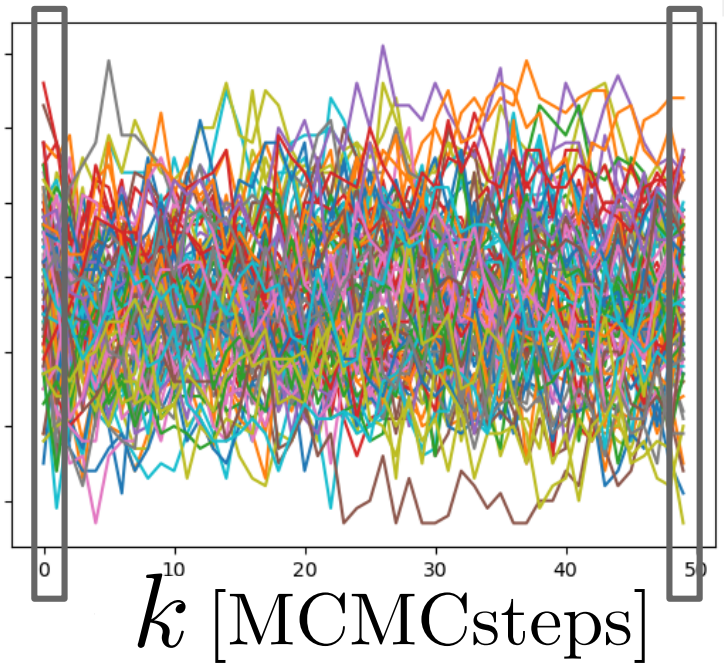
$$\frac{\partial \mathcal{L}}{\partial \eta_i} = \langle v_i\rangle_{\mathcal{D}} - \langle v_i\rangle_{\mathcal{H}}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_a} = \langle h_a\rangle_{\mathcal{D}} - \langle h_a\rangle_{\mathcal{H}}$$

Monte Carlo

N_s parallel Markov chains

initialization



Measure ⇒ r.h.s gradient

On the gradient

$\langle \cdots \rangle_{\mathcal{H}}$ R.h.s gradient will be correctly computed if the simulations thermalize

$$k \sim n\tau_{\text{mixing}}$$

$$\frac{\partial \mathcal{L}}{\partial w_{ia}} = \langle v_i h_a \rangle_{\mathcal{D}} - \langle v_i h_a \rangle_{\mathcal{H}}$$

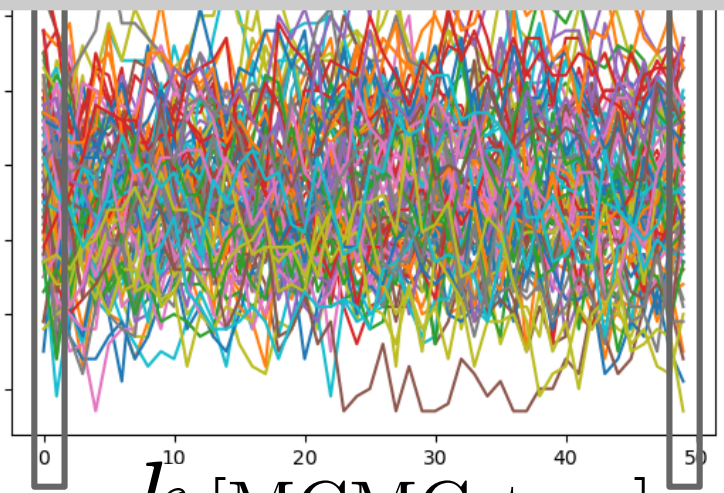
$$\frac{\partial \mathcal{L}}{\partial \eta_i} = \langle v_i \rangle_{\mathcal{D}} - \langle v_i \rangle_{\mathcal{H}}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_a} = \langle h_a \rangle_{\mathcal{D}} - \langle h_a \rangle_{\mathcal{H}}$$

Monte Carlo

N_s parallel Markov chains

initialization



Measure ⇒ r.h.s gradient

On the gradient

$\langle \cdots \rangle_{\mathcal{H}}$ R.h.s gradient will be correctly computed if the simulations thermalize

$$k \sim n\tau_{\text{mixing}}$$

$$\frac{\partial \mathcal{L}}{\partial w_{ia}} = \langle v_i h_a \rangle_{\mathcal{D}} - \langle v_i h_a \rangle_{\mathcal{H}}$$

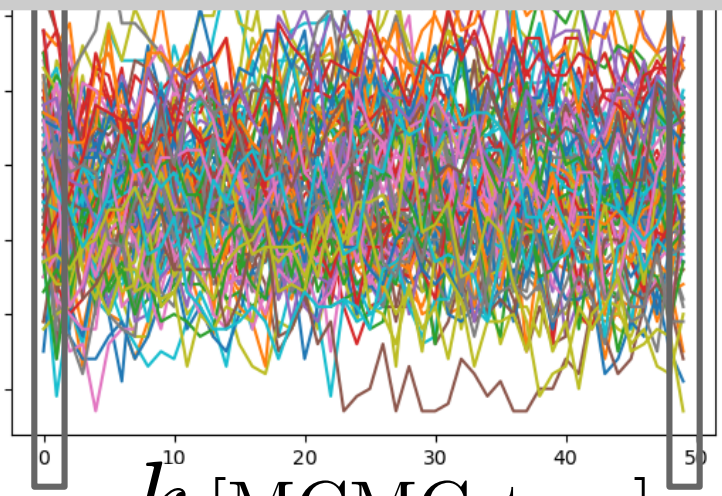
$$\frac{\partial \mathcal{L}}{\partial \eta_i} = \langle v_i \rangle_{\mathcal{D}} - \langle v_i \rangle_{\mathcal{H}}$$

$$\frac{\partial \mathcal{L}}{\partial \quad} = \langle h \quad \rangle_{\mathcal{D}} - \langle h \quad \rangle_{\mathcal{H}}$$

Monte Carlo

N_s parallel Markov chains

initialization



Let's choose good starting point and approximate with k~O(10) steps

- Contrastive divergence (CD) [*Hinton (2002)]* **Init – dataset**
- Persistence contrastive divergence (PCD) [*Tieleman (2008)]* **Init – previous last point**
- Mean field (TAP) *[Gabrié, Tramel, and Krzakala (2015)]*

Or

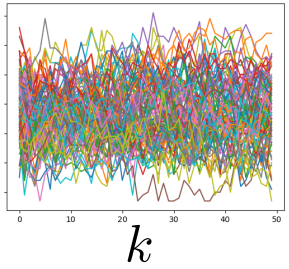
- Simulated annealing, Parallel Tempering, …

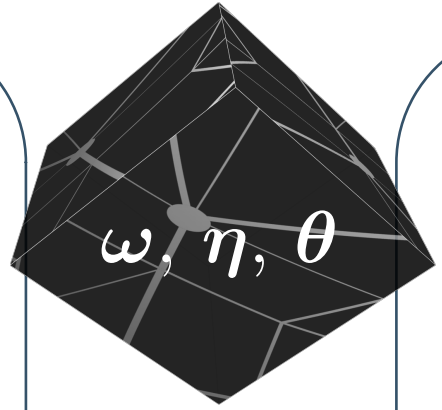
Equilibrium vs. Non-eq. regimes

Decelle, Furtlehner, Seoane
ArXiv:2105.13889
(NeurIPS2021)
See Decelle's poster

Training

∇LL



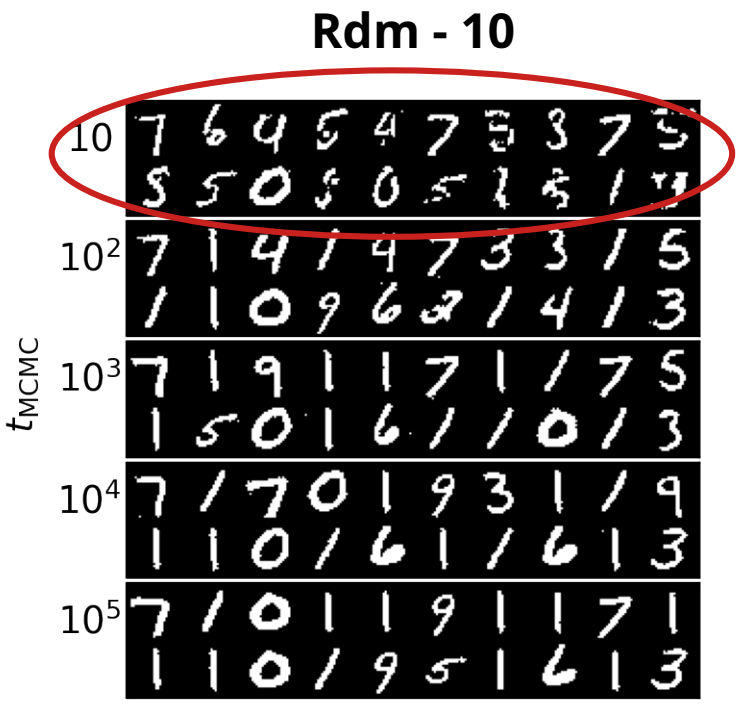


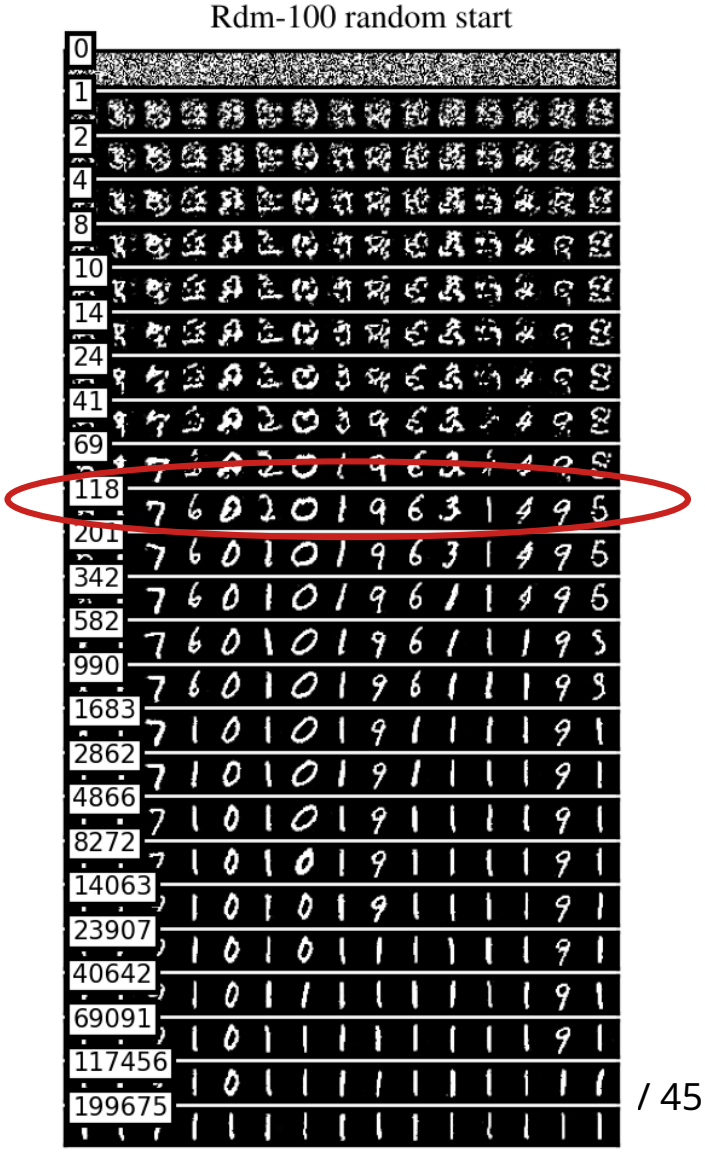
Sampling

- **Non-equilibrium** $k < t_{\text{therm}}$
- **Equilibrium** $k > t_{\text{therm}}$

- **"Learns the dynamics"**
- **Learns a good model for the data**

Non-equilibrium regime



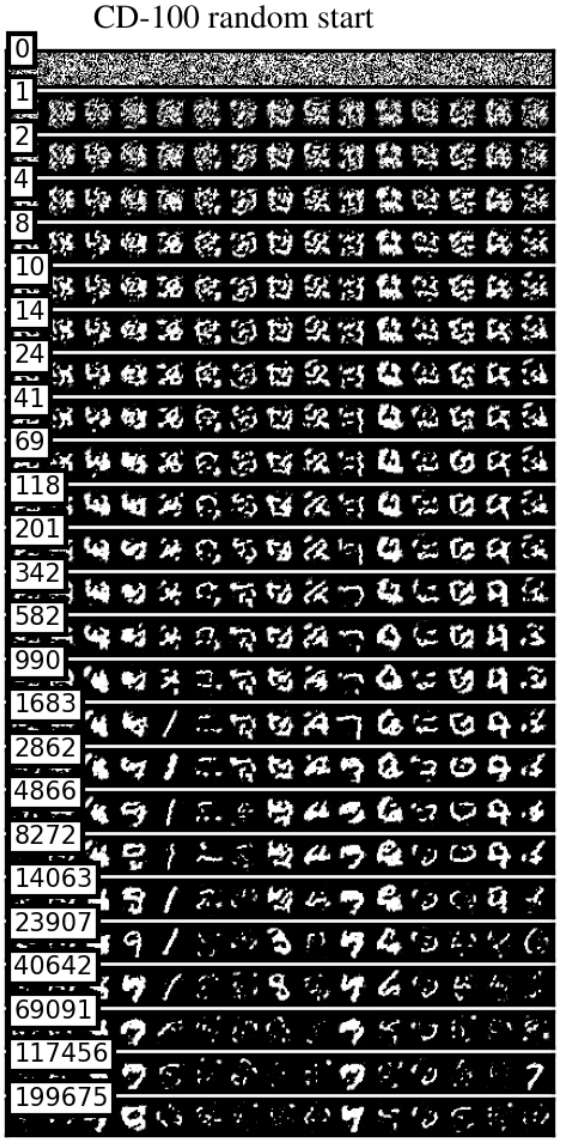


Non-equilibrium regime

Training using CD: chain initialisation at the dataset

If we sample the RBM from:

- random **NOTHING**

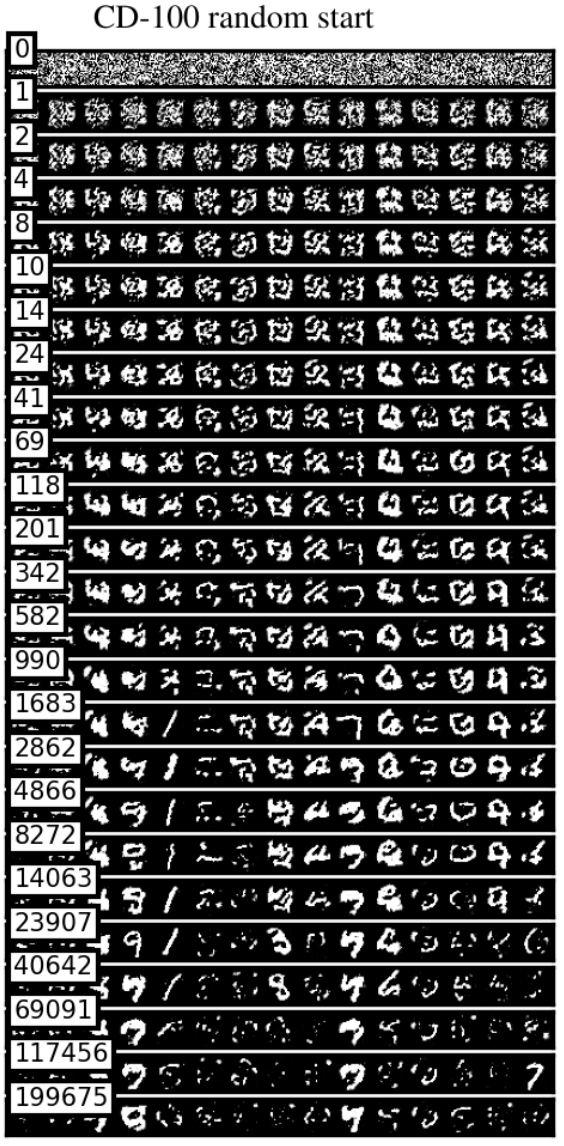


Non-equilibrium regime

Training using CD: chain initialisation at the dataset

If we sample the RBM from:

- random **NOTHING**
- the dataset we **do not get anything new**



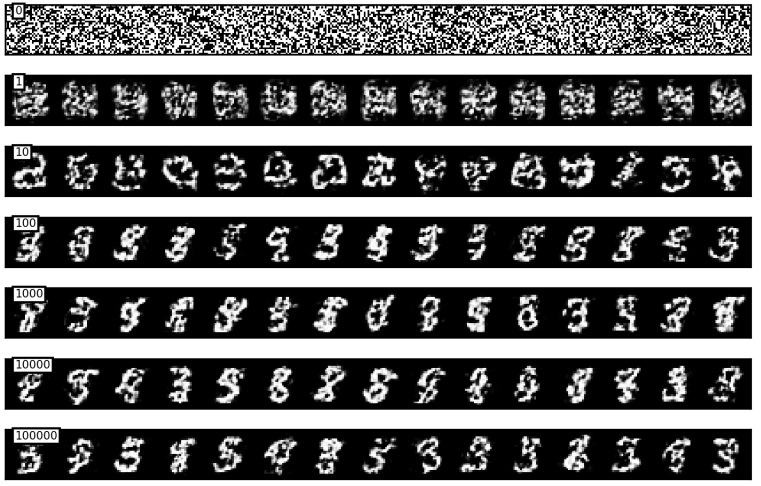
Non-equilibrium regime

Training using Mean Field dynamics

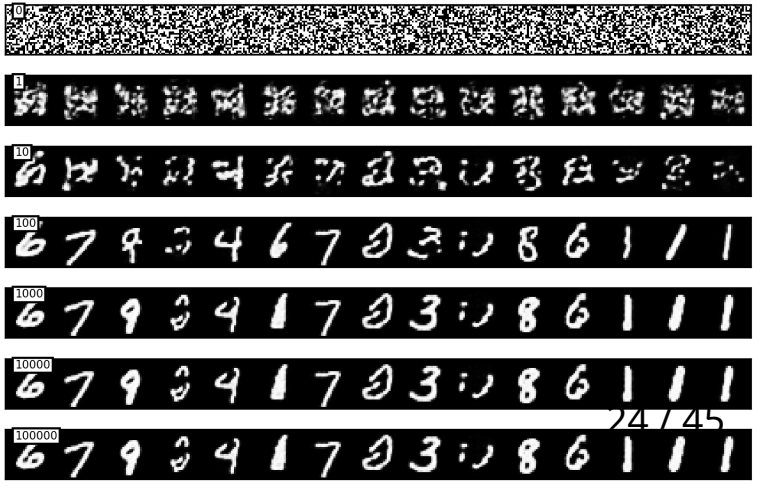
If we sample the RBM :

- Heat bath dynamics **NOTHING**
- With MF **good data at t~k and it does not change anymore**

With HB MCMC dynamics

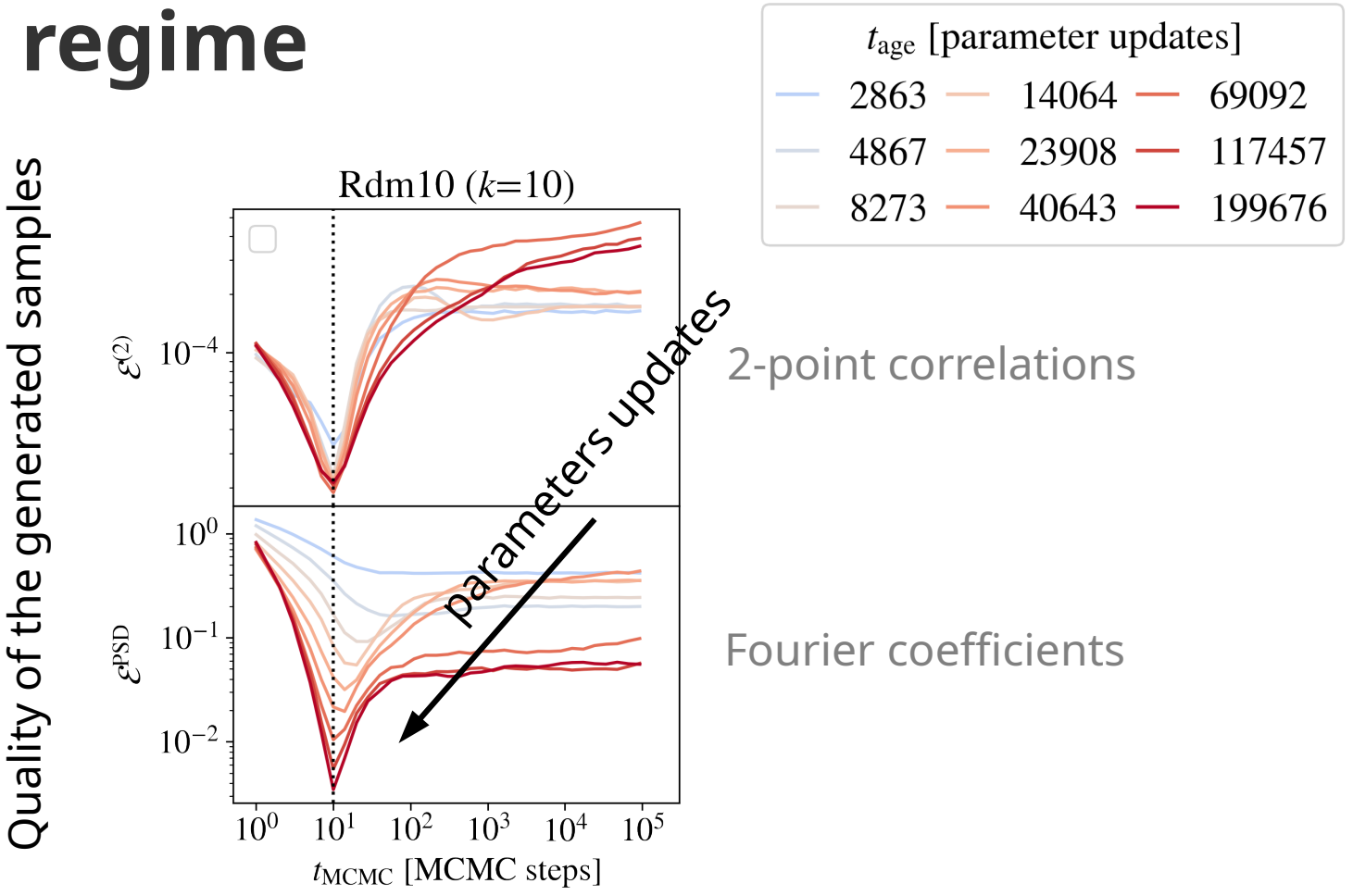


With MF dynamics



Non-equilibrium regime

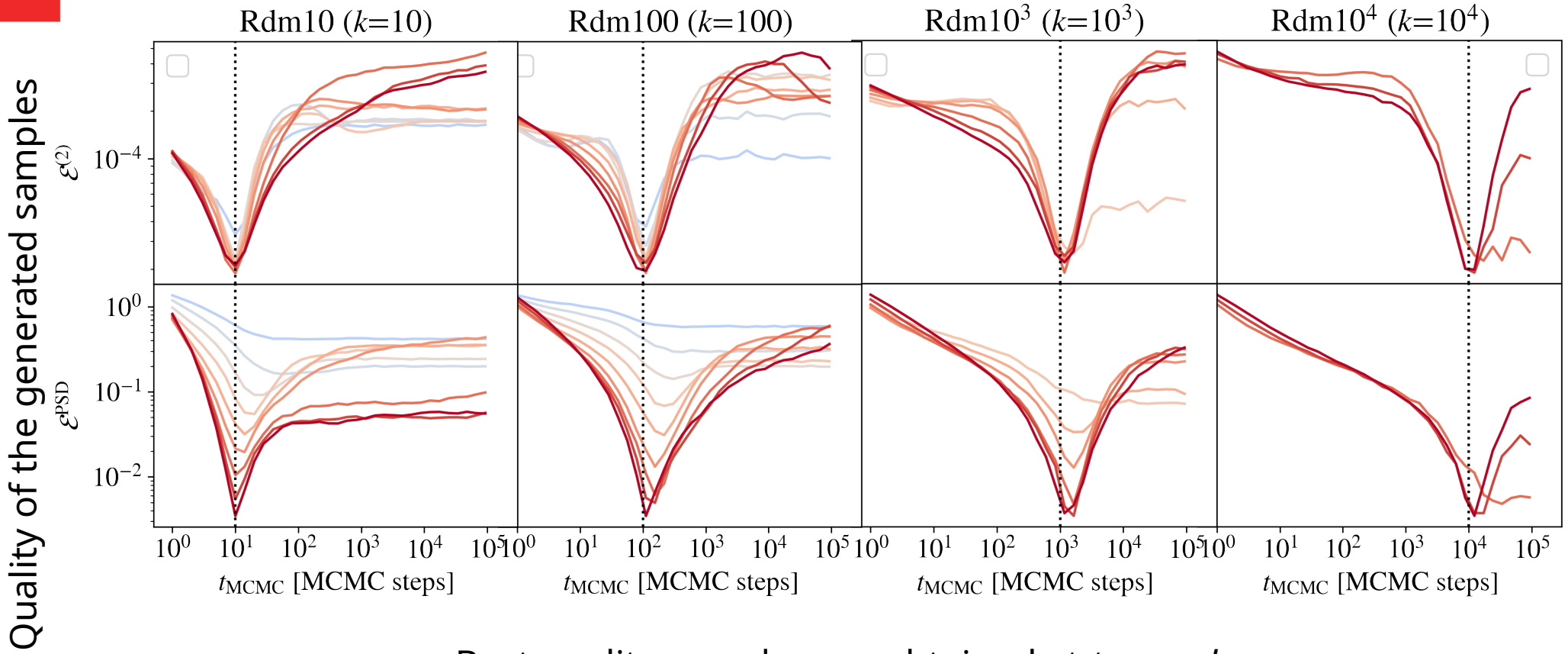
Decelle, Furtlehner, Seoane
arXiv:2105.13889



2-point correlations

Fourier coefficients

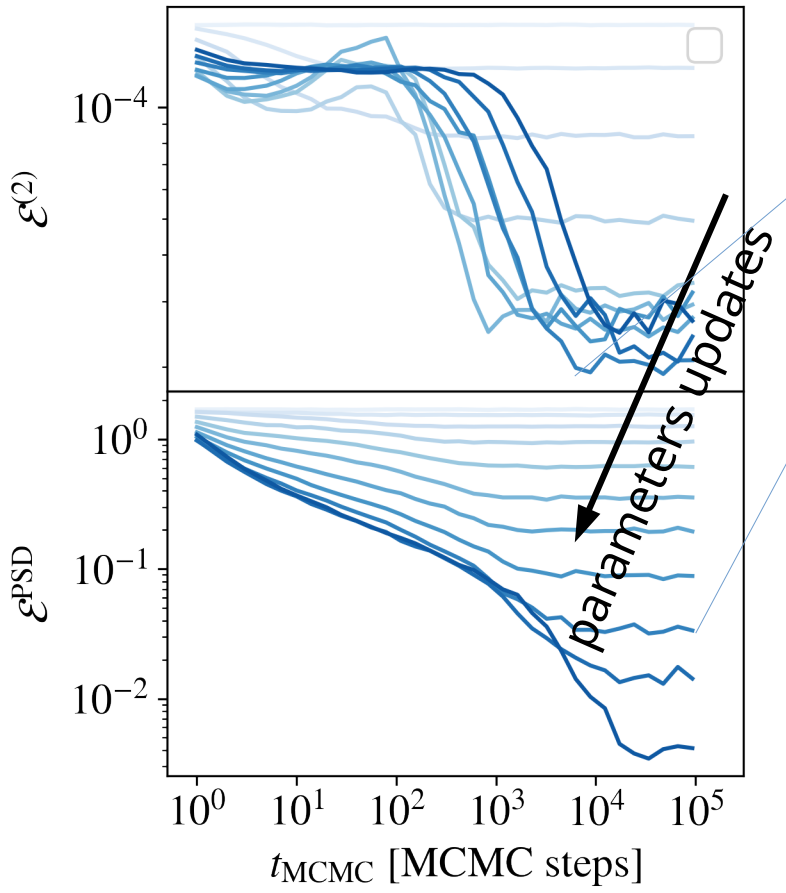
Non-equilibrium regime : generation

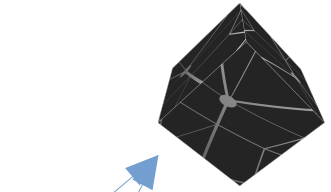


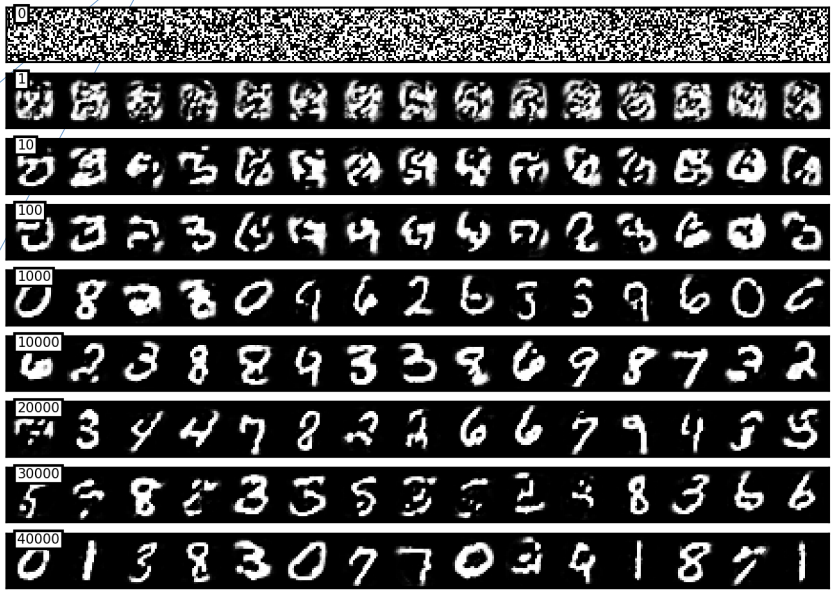
Best quality samples are obtained at $t_{\mathrm{MCMC}} \sim k$

Decelle, Furtlehner, Seoane
arXiv:2105.13889

Equilibrium regime







Dynamics are much faster

Equilibrium vs. Non-eq. regimes

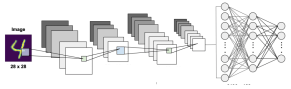
Decelle, Furtlehner, Seoane
ArXiv:2105.13889
(NeurIPS2021)
See Decelle's poster

Non-equilibrium
$k < t_{\text{therm}}$

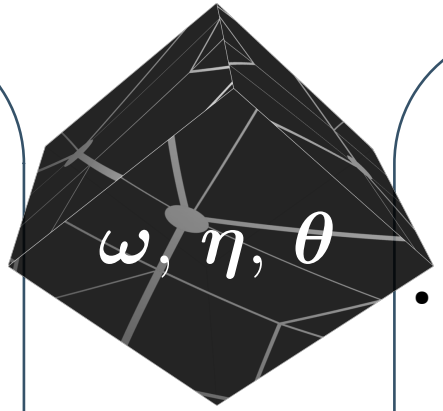
- "Learns the dynamics"
- **Advantage:** Optimal for data generation

  → random noise initialization

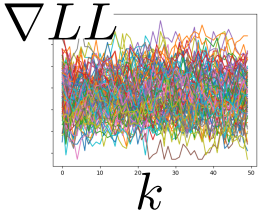
  Nijkamp, Hill, Han, Wu, Zhu. NIPS 2019, AAAI 2020.

  BM: Muntoni, Pagnani, Martin Weigt, Zamponi (2021)
- **Drawbacks:**
  - Unpredictable if not controlled
  - not a good model for the data
  - Extremely slow dynamics



Training



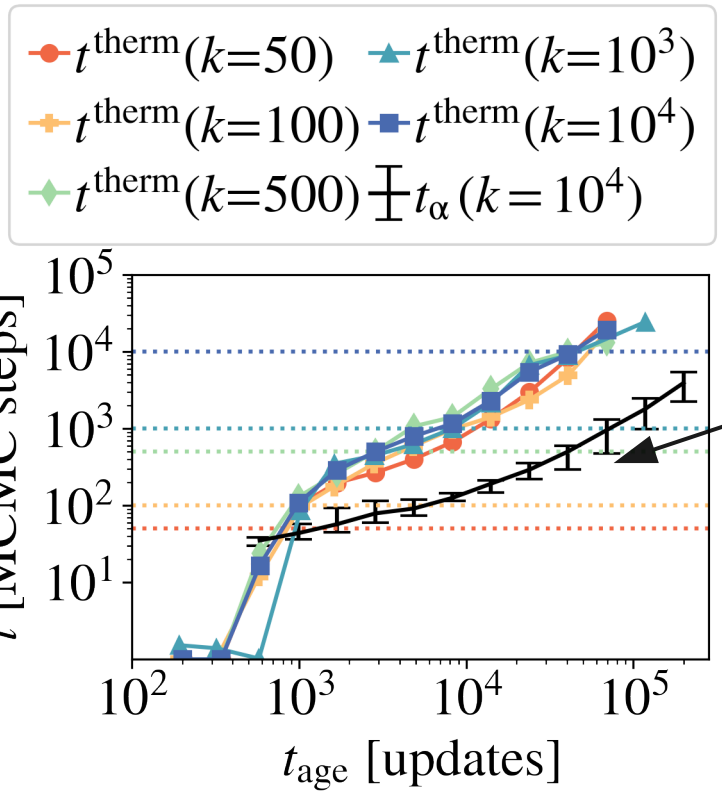
Equilibrium
$k > t_{\text{therm}}$

- Learns the (unnormalized) prob. Distribution of the data
- **Advantage:**
  - Fits a good model for the data
  - Sampling is stable
- **Drawbacks:**
  - Very slow training: need large *k*
  - Generating new configurations can become prohibitive

Equilibration: how long? Easy : MNIST

$$C_m^{\text{eq}}(t) = \frac{1}{N_v}\sum_i (m_i(t)-m)(m_i(0)-m)$$

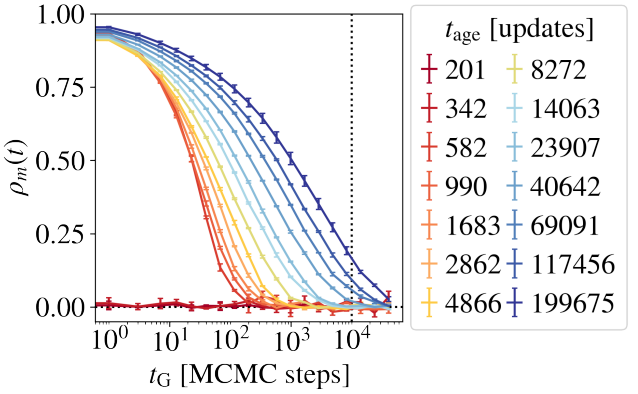
$$\rho_m(t) = \frac{C_m(t)}{C_m(0)} \sim A\exp(-t/t_\alpha)$$



Mixing time

- Grows with the learning updates
- Always over 50

$$t^{\text{therm}} \sim n\tau_{\text{mixing}}$$



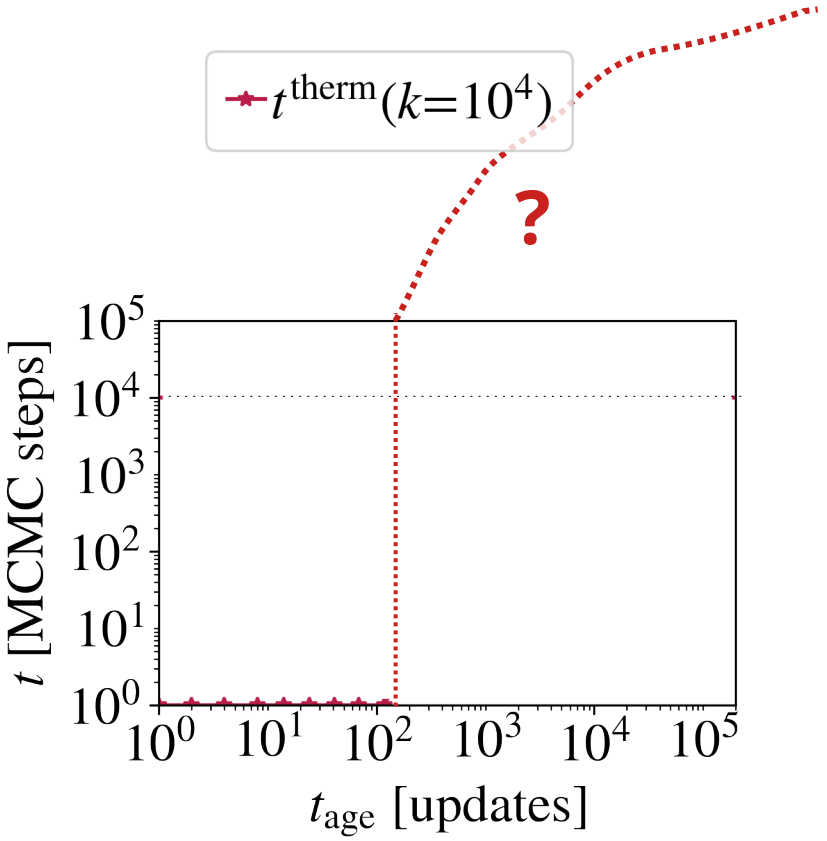
Literature k~10

Always out-of-equilibrium regime !

Decelle, Furtlehner, Seoane

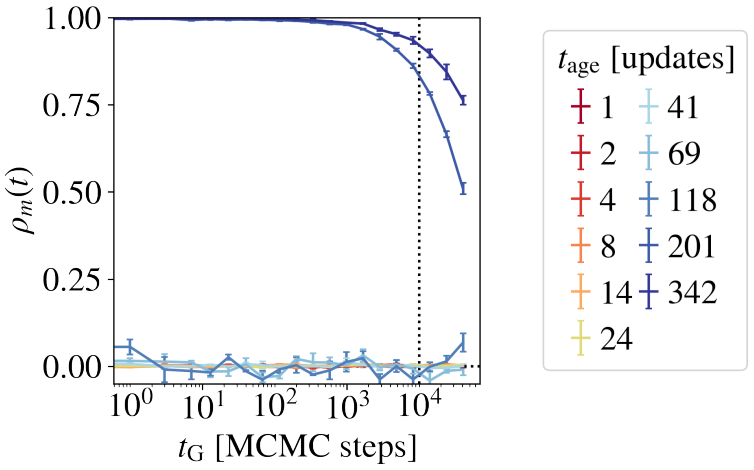
arXiv:2105.13889

Equilibration: how long? Hard : GENE



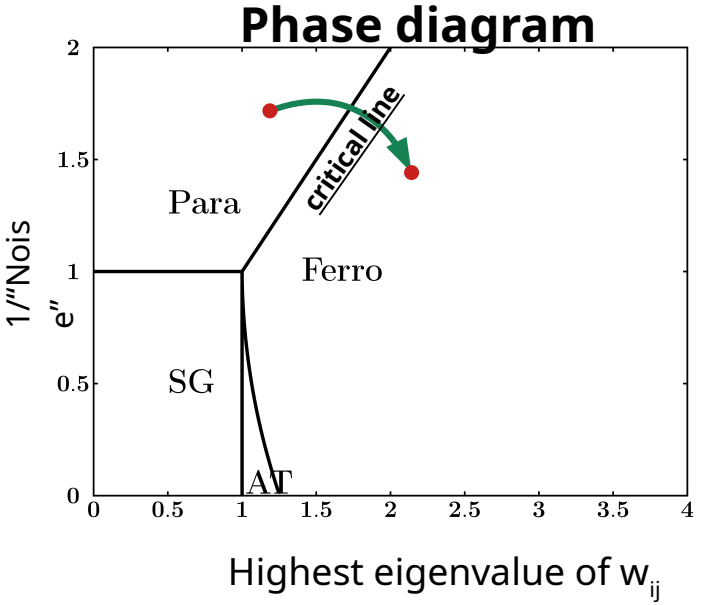
The thermalisation time jumps suddenly beyond 10^5 MCMC steps.

The equilibrium regime is beyond our reach...



What does it happen?

$$E[\boldsymbol{v}, \boldsymbol{h}] = -\sum_{ia} v_i \boldsymbol{w}_{ia} h_a - \sum_i \eta_i v_i - \sum_a \theta_a h_a$$



$\boldsymbol{\omega}_i$: *i*-th eigenvector of the W matrix

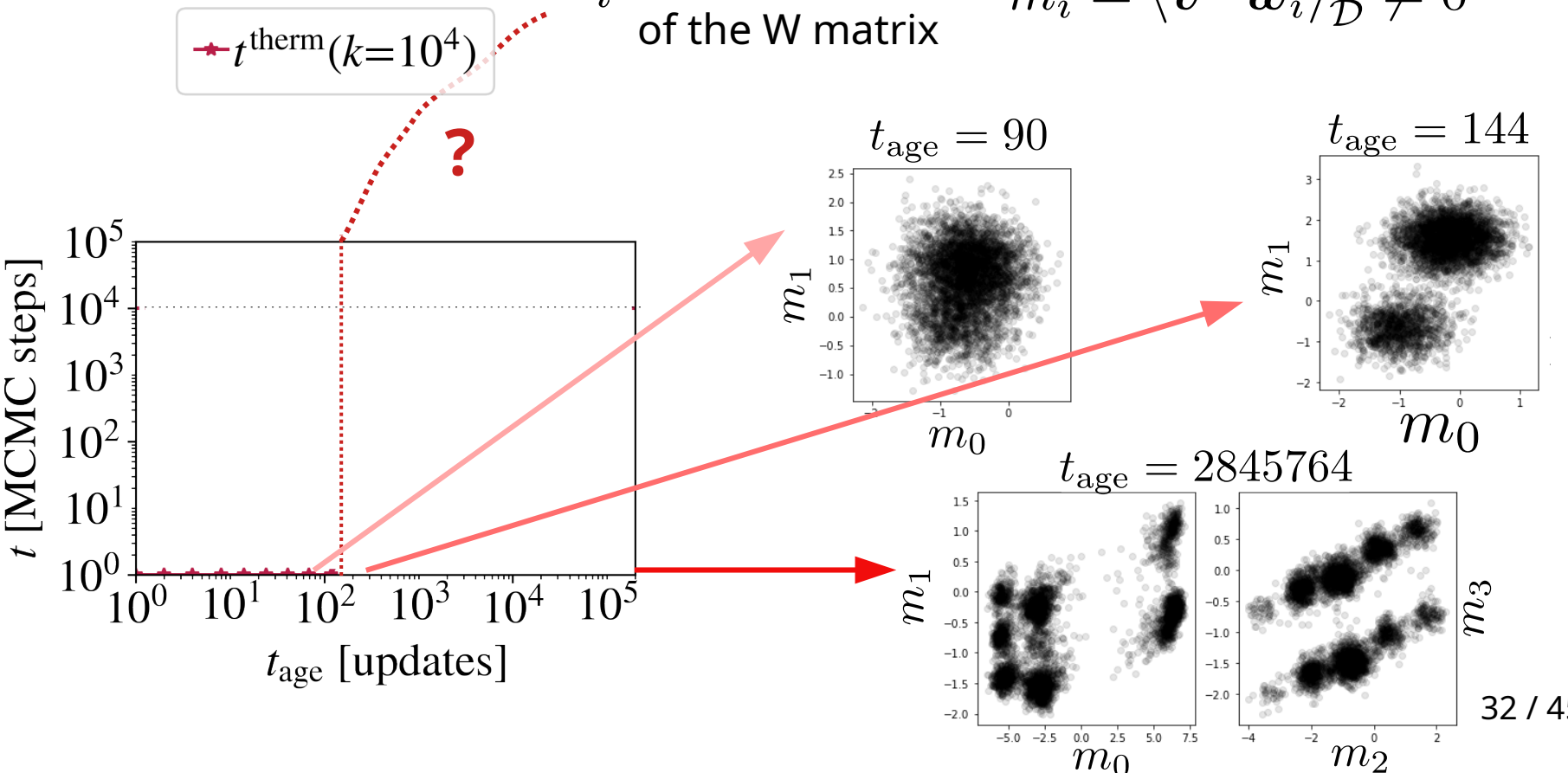
$$m_i = \langle \boldsymbol{v} \cdot \boldsymbol{w}_i \rangle_{\mathcal{D}} \neq 0$$

What does it happen?

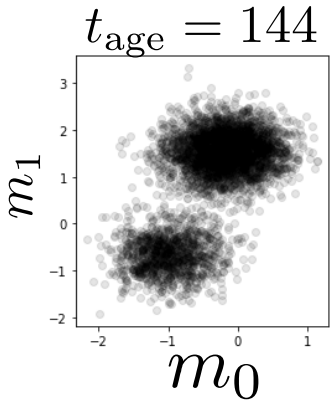
$$E[\boldsymbol{v},\boldsymbol{h}] = -\sum_{ia} v_i \boldsymbol{w}_{ia} h_a - \sum_i \eta_i v_i - \sum_a \theta_a h_a$$

$\boldsymbol{\omega}_i$: *i*-th eigenvector of the W matrix

$$m_i = \langle \boldsymbol{v} \cdot \boldsymbol{w}_i \rangle_{\mathcal{D}} \neq 0$$



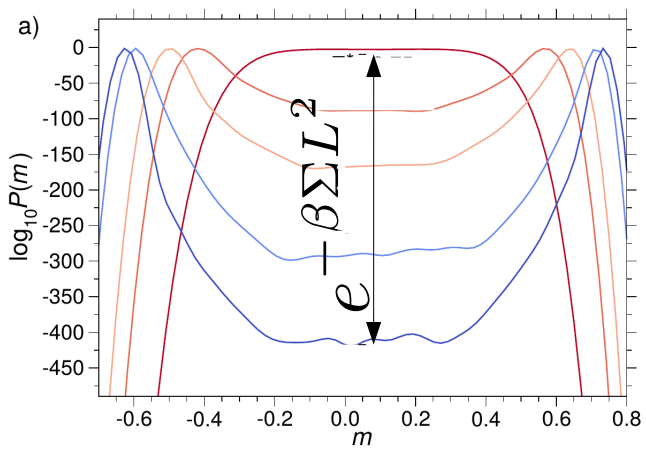
What does it happen?

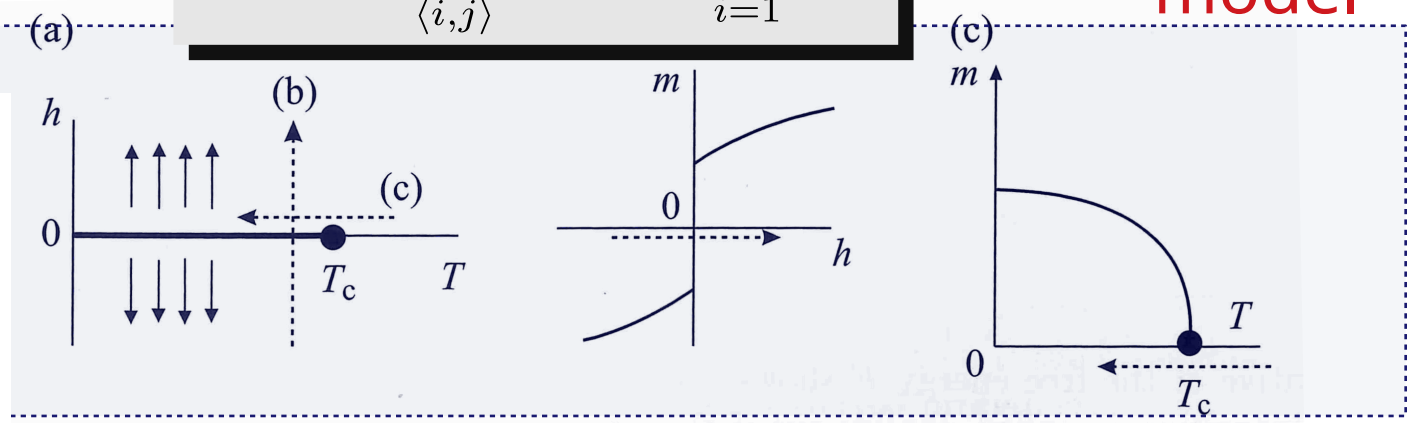


As learning advances we start to have metastable states

$$\mathcal{H} = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j - h \sum_{i=1}^{N} \sigma_j ,$$

Ising model

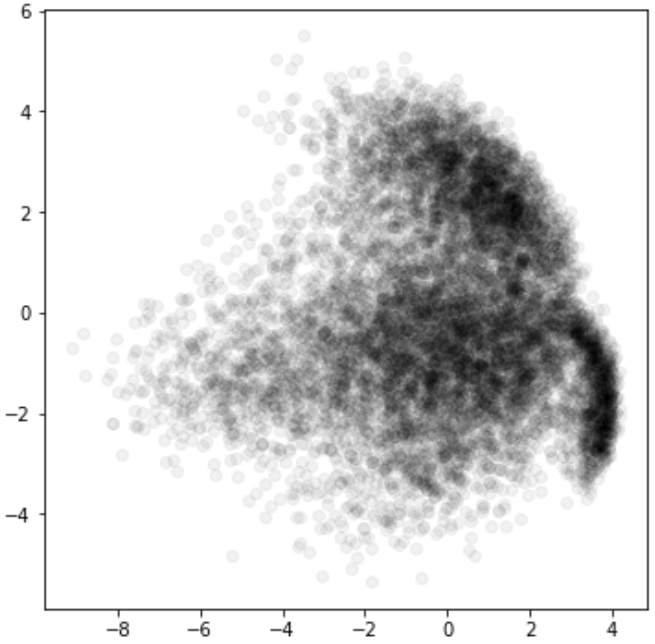




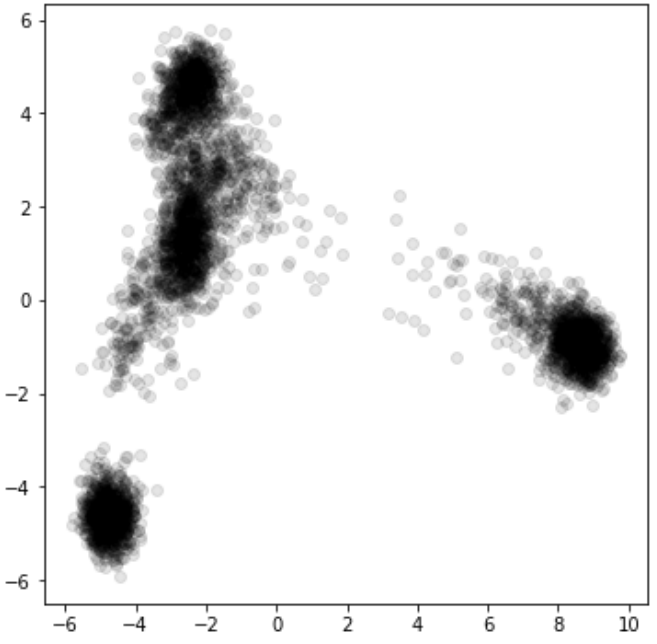
Structured datasets

We do the PCA of the data and project the data long the first 2 eigenvectors

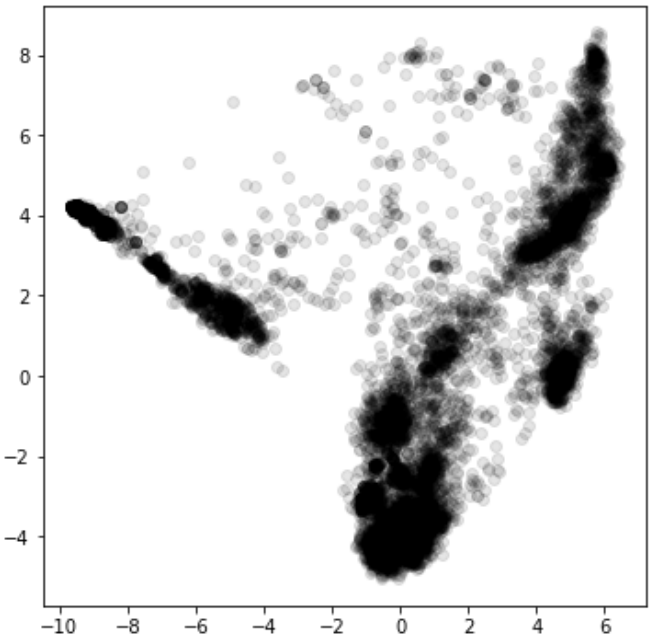
MNIST



GENE



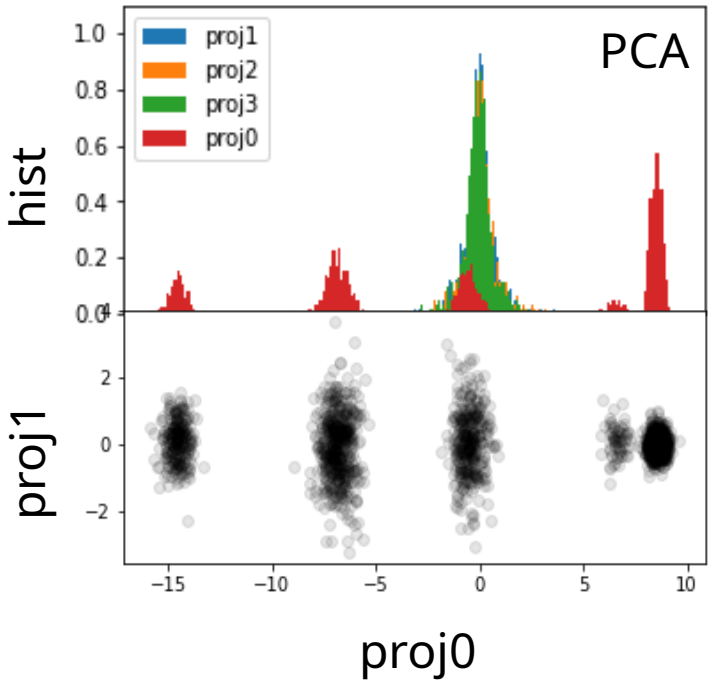
PROTEIN FAMILY



Clusterized datasets

Step back : high dimensional clusters along 1D

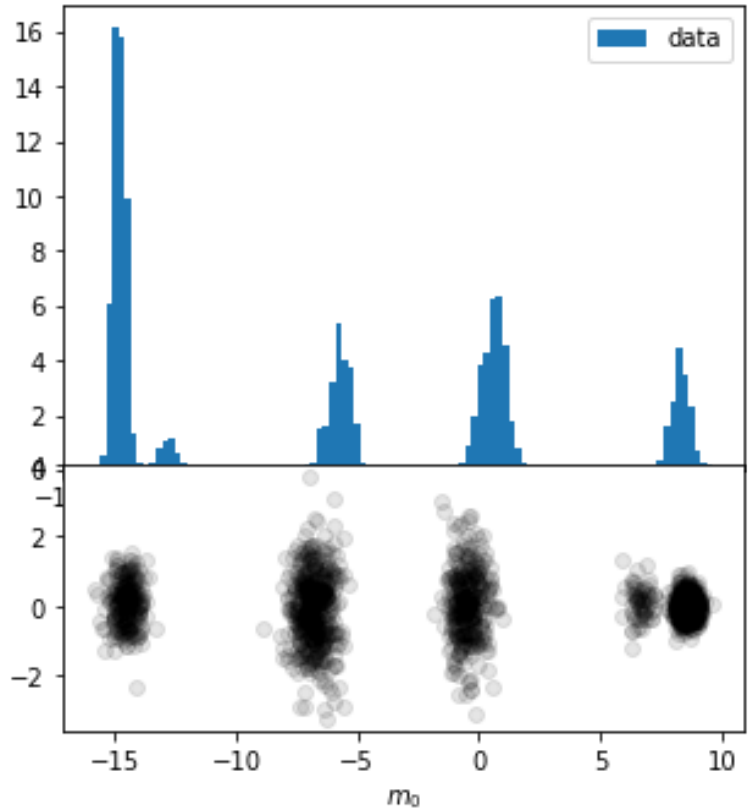
We feed the RBM with points belonging to different clusters (in high dimensions) but Separated only in one



- Standard RBM training procedure fails completely to fit such dataset
- Yet, this simple low dimensional dataset can be trained analytically [Decelle, Furtlehner, PRL 2021]
  - ➔ We can have a perfect model ω, θ, η to test the biased sampling

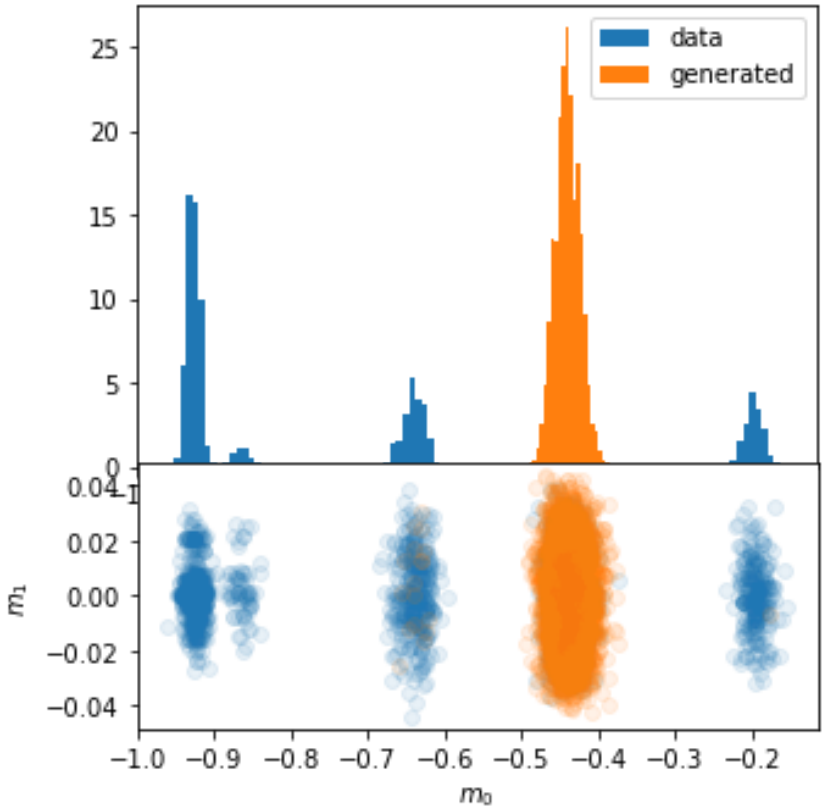
[Bereux, Decelle, Furtlehner, Seoane, *In preparation*]

Problems of the standard MCMC sampling



Projection of the first eigenvector
of the W matrix (normalized)

Problems of the standard MCMC sampling



Projection of the first eigenvector
of the W matrix (normalized)

The Tethered Monte Carlo approach (I)

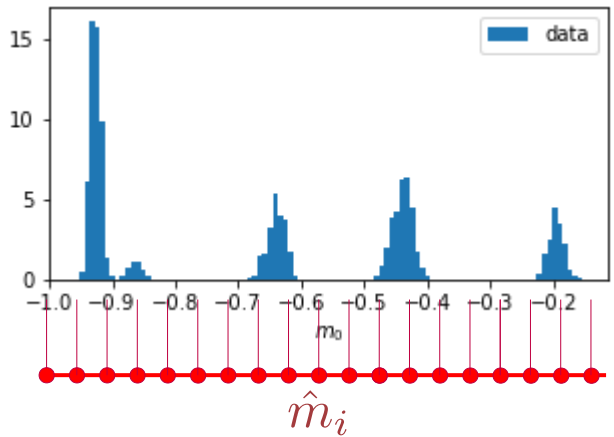
[Fernandez, Martin-Mayor, Yllanes - Nuclear physics (2009), Martin-Mayor, Seoane, Yllanes, Journal of Statistical Physics (2011), Fernández, Martín-Mayor, Seoane, Verrocchio, PRL (2012)]

$$Z = \sum_{\boldsymbol{v},\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})} = \sqrt{\frac{N}{2\pi}} \int_{-\infty}^{\infty} d\hat{m} \sum_{\boldsymbol{v},\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})} e^{-N(\hat{m}-m_0(\boldsymbol{v}))^2/2} = \int_{-\infty}^{\infty} d\hat{m} e^{-N\Omega(\hat{m})}$$

$$\langle O(\boldsymbol{v},\boldsymbol{h})\rangle = \frac{\sum_{\boldsymbol{v},\boldsymbol{h}} O e^{-E(\boldsymbol{v},\boldsymbol{h})}}{\sum_{\boldsymbol{v},\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})}} = \int_{-\infty}^{\infty} d\hat{m} \, \langle O \rangle_{\hat{m}} \, e^{-N\Omega(\hat{m})}$$

$$\langle O(\boldsymbol{v},\boldsymbol{h})\rangle_{\hat{m}} = \frac{\sum_{\boldsymbol{v},\boldsymbol{h}} O \; \omega(\hat{m},\boldsymbol{v},\boldsymbol{h})}{\sum_{\boldsymbol{v},\boldsymbol{h}} \omega(\hat{m},\boldsymbol{v},\boldsymbol{h})}$$

$$\omega(\hat{m},\boldsymbol{v},\boldsymbol{h}) = e^{-E(\boldsymbol{v},\boldsymbol{h})} e^{-N(\hat{m}-m_0(\hat{v}))^2/2}$$



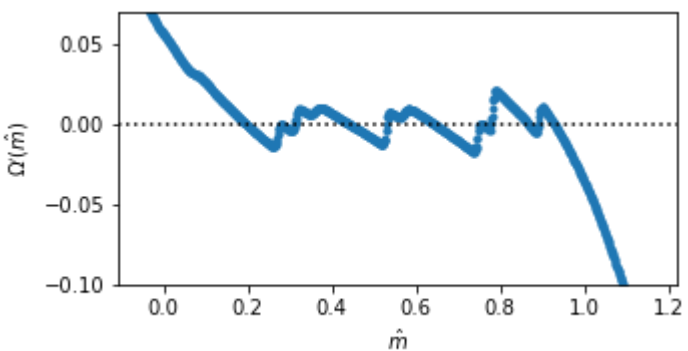
- Run *K* simulations at $\hat{m}_i$, with *i*=1,...,*K* fixed
- We break the metastability: fast thermalisation

The Tethered Monte Carlo approach (II)

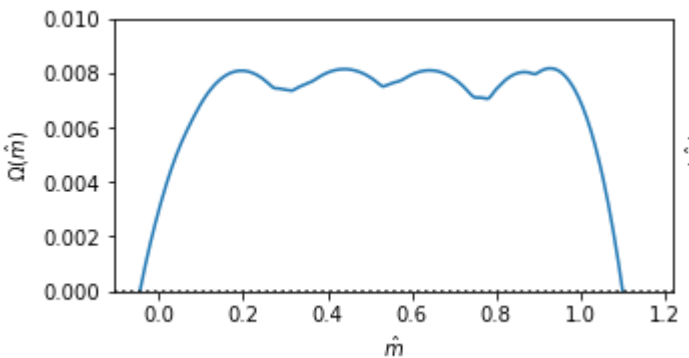
[Fernandez, Martin-Mayor, Yllanes - Nuclear physics (2009), Martin-Mayor, Seoane, Yllanes, Journal of Statistical Physics (2011), Fernández, Martín-Mayor, Seoane, Verrocchio, PRL (2012)]

$$\frac{d\Omega}{d\hat{m}} = \langle \hat{m} - m_0(\boldsymbol{v}) \rangle_{\hat{m}}$$

1) Compute Ω'

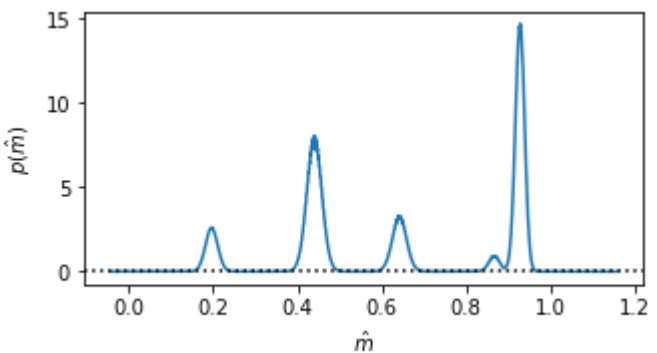


2) Integrate it numerically



3) Extract

$$p(\hat{m}) = \frac{e^{-N\Omega(\hat{m})}}{\int_{-\infty}^{\infty} e^{-N\Omega(\hat{m})}}$$

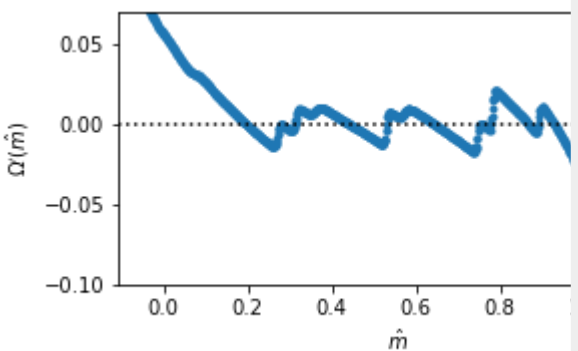


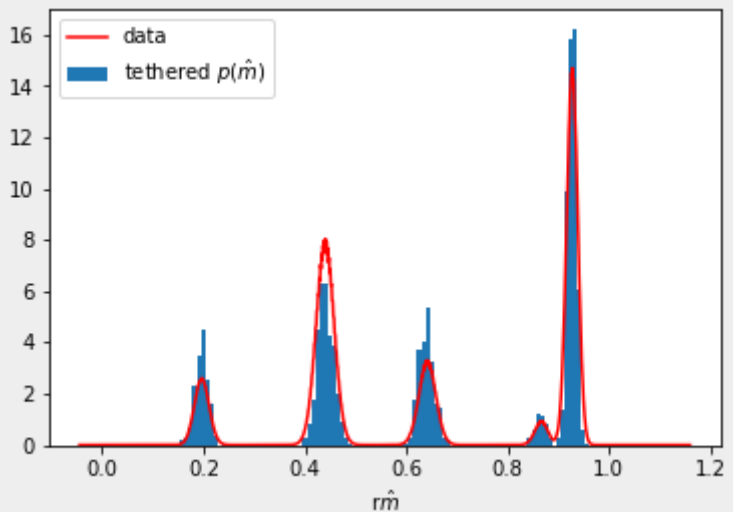
The Tethered Monte Carlo approach (II)

[Fernandez, Martin-Mayor, Yllanes - Nuclear physics (2009), Martin-Mayor, Seoane, Yllanes, Journal of Statistical Physics (2011), Fernández, Martín-Mayor, Seoane, Verrocchio, PRL (2012)]

$$\frac{d\Omega}{d\hat{m}} = \langle \hat{m} - n$$

1) Compute Ω'

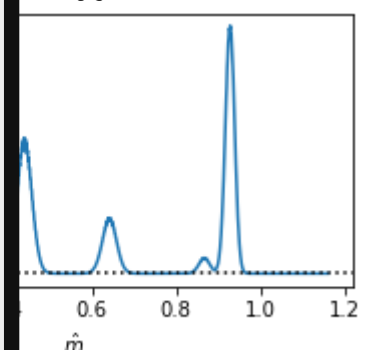




Once having $p(m)$, we can use inverse sampling for sample generation!

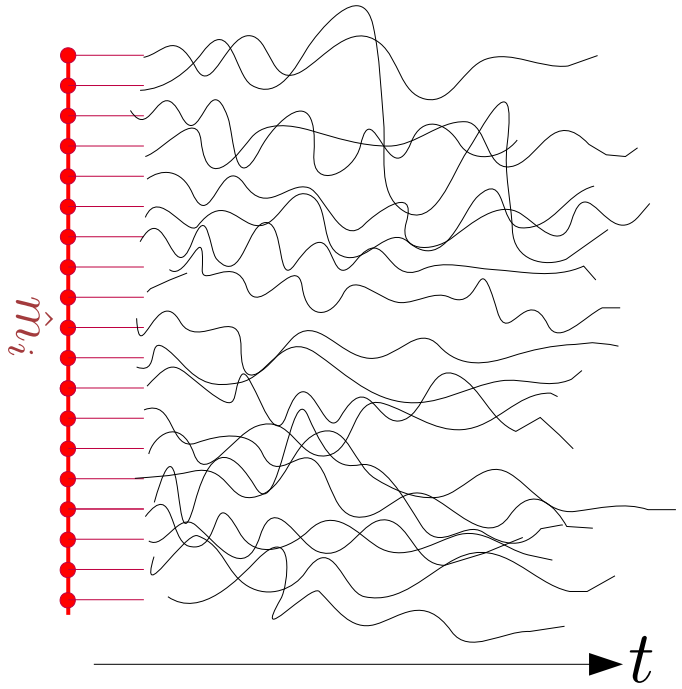
xtract

$$\frac{e^{-N\Omega(\hat{m})}}{\int_{-\infty}^{\infty} e^{-N\Omega(\hat{m})}}$$

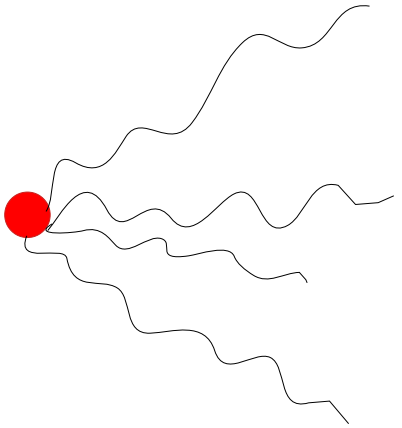


[Bereux, Decelle, Furtlehner, Seoane, *In preparation*]

Learning with TMCMC



$$\frac{\partial \mathcal{L}}{\partial w_{ia}} = \langle v_i h_a \rangle_{\mathcal{D}} - \langle v_i h_a \rangle_{\mathcal{H}}$$
$$\frac{\partial \mathcal{L}}{\partial \eta_i} = \langle v_i \rangle_{\mathcal{D}} - \langle v_i \rangle_{\mathcal{H}}$$
$$\frac{\partial \mathcal{L}}{\partial \theta_a} = \langle h_a \rangle_{\mathcal{D}} - \langle h_a \rangle_{\mathcal{H}}$$



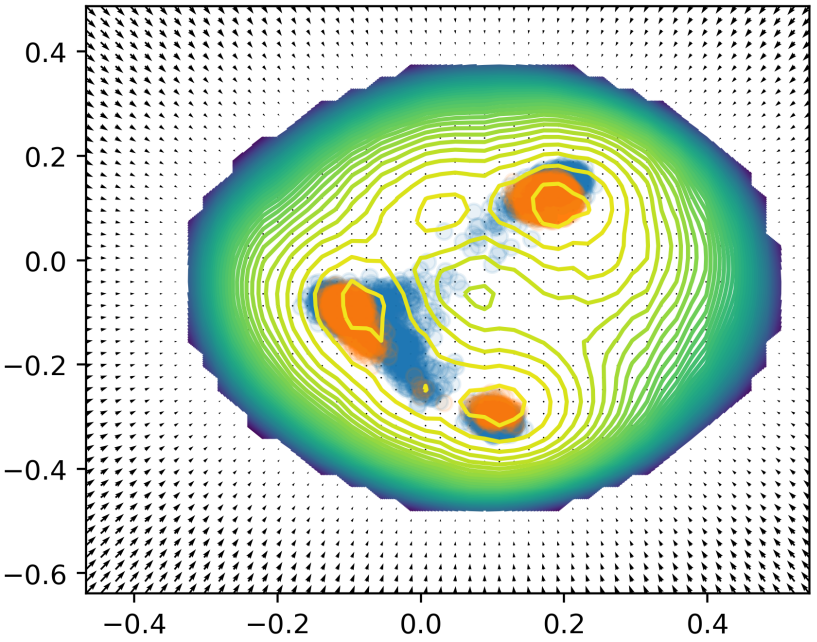
$$\langle O(\boldsymbol{v}, \boldsymbol{h}) \rangle_{\mathcal{H}} = \int_{-\infty}^{\infty} d\hat{m} \, \langle O \rangle_{\hat{m}} \, e^{-N\Omega(\hat{m})}$$

Learning with TMCMC

Generalization to higher number of conserved observables is straigtforward...

$$\omega(\hat{m}_1, \hat{m}_2, \boldsymbol{v}, \boldsymbol{h}) = e^{-E(\boldsymbol{v},\boldsymbol{h})} e^{-N(\hat{m}_2 - m_0(\hat{v}))^2/2} e^{-N(\hat{m}_2 - m_1(\hat{v}))^2/2}$$

$$\Omega'(\hat{m}) \to \boldsymbol{\nabla}\Omega = \left(\langle \hat{m}_0 - m_0 \rangle_{\hat{m}_0, \hat{m}_1} \langle \hat{m}_1 - m_1 \rangle_{\hat{m}_0, \hat{m}_1} \right)$$



Conclusions

Decelle, Furtlehner, Seoane
ArXiv:2105.13889

- RBM have a major advantage in terms of interpretability of the extracted patterns, but training is very unstable following the standard recipes.
- Instability is a consequence of the nonequilibrium sampling during the sampling and can be controlled and taken in advance to generate good samples with short trainings.
- In order to fit a good model for the data, the sampling during the learning must equilibrate:
  - Datasets without structure : mixing time grows with Nb. Updates
  - Structured datasets: thermalisation is hampered by coexistence of states → biased sampling

Parameters MNIST

- Number of hidden nodes: $N_h = 500$
- Learning rate: $\alpha = 0.01$
- Minibatch size: $n_{mb} = 500$
- no ℓ_2 regularization of momemtum.
- The gradient is centered according to [1]
- The visible biases are initialized to match the empirical frequency of the training dataset:

$$\eta_i = \log\left(\frac{\bar{m}_i}{1-\bar{m}_i}\right) \text{ where } \bar{m}_i = \frac{1}{M}\sum_m s_i^{(m)} \tag{1}$$

- The number of MC chains used for the negative term was always equal to n_{mb}
- The number of MC steps for the negative chains is indicated by the variables t_{GL} and can vary.

RBM: learning and phase transition

We can confirm experimentally that the divergence of the mixing time correspond to the 2nd order phase transition