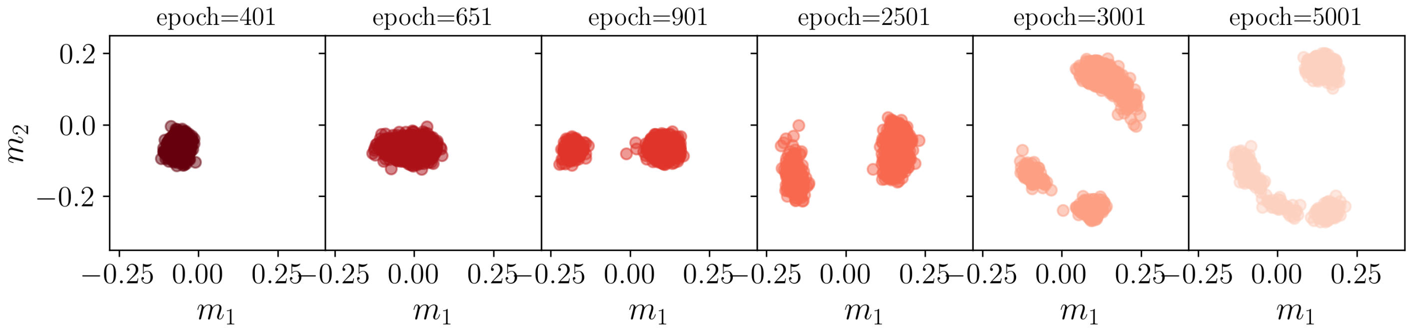
Cascade of phase transitions in the training of Energy-based models



Aurélien Decelle

from Universidad Complutense de Madrid

Joint work with
Dimitrios Bachtis @ ENS Paris
Giulio Biroli @ ENS Paris
Beatriz Seoane @ Complutense Madrid

UNIVERSIDAD COMPLUTENSE MADRID
Comunidad de Madrid
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Cofinanciado por la Unión Europea
AGENCIA ESTATAL DE INVESTIGACIÓN

Unsupervised learning and Energy-based model

Energy based model is a class of unsupervised models where the distribution is given by

$$p(\boldsymbol{s}) = \frac{1}{Z} \exp\left(-\mathcal{H}[\boldsymbol{s}; \boldsymbol{\theta}]\right)$$

In such model, the learning is typically done by maximizing the likelihood w.r.t. $\boldsymbol{\theta}$

$$\nabla_{\boldsymbol{\theta}} \mathcal{L} = \langle \nabla_{\boldsymbol{\theta}} \mathcal{H} \rangle_{\mathrm{data}} - \langle \nabla_{\boldsymbol{\theta}} \mathcal{H} \rangle_{\mathrm{p}}$$

The difficulty lies in the computation of the average w.r.t. the model which is usually done by performing Monte Carlo estimation

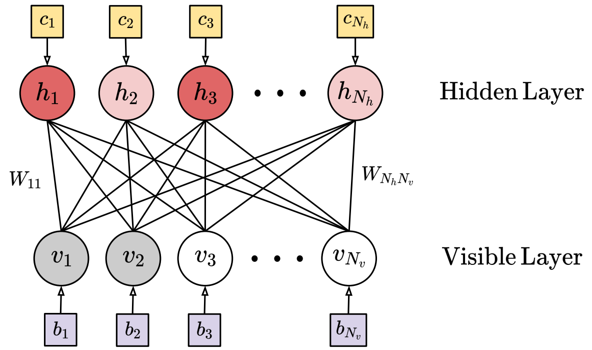
Related works and setting

In order to pursue analytical computation we will restrict ourselves to the case of the Restricted Boltzmann Machine

$$\mathcal{H} = -\sum_{ia} v_i w_{ia} h_a - \sum_i v_i b_i - \sum_a h_a c_a$$

$$v_i = \{\pm 1\} \text{ or } \{0,1\}$$

$$h_a = \{0,1\} \text{ or Gaussian}$$



More generally, It has been shown in some generative models

1 – How the phase space of RBM can exhibit spontaneous broken symmetry

2 – How perfectly trained diffusion model undergoes several phase transition during sample generation

We show theoretically and numerically the phase transition occurring in the learning of RBM

Refs.
- Decelle et al. 2018
- Biroli et al. 2024
- Ambrogoni et al. 2023

Theoretical setting

We consider a simple bimodal artificial dataset which we learn with an RBM.
We can compute the gradient in the infinite size limit.

$$\frac{\partial \mathcal{L}}{\partial w_{ia}} \approx \frac{dw_{ia}}{dt} = \epsilon \left[\langle s_i h_a \rangle_{\mathrm{data}} - \langle s_i h_a \rangle_{\mathrm{p}} \right]$$

$$= \epsilon \left[\xi_i \sum_k w_k \xi_k m - N_{\mathrm{v}} h^* \tanh(h^* w_i) \right]$$

$\boldsymbol{\xi}$: preferred direction of the dataset
m : width of the modes of the dataset
$\boldsymbol{w}$: weights of the RBM
$\mathrm{h}^* = \frac{1}{N} \sum_k w_k \tanh(w_k h^*)$

We see that we can project the gradient onto $\boldsymbol{\xi}$ to obtain a simpler form to solve.

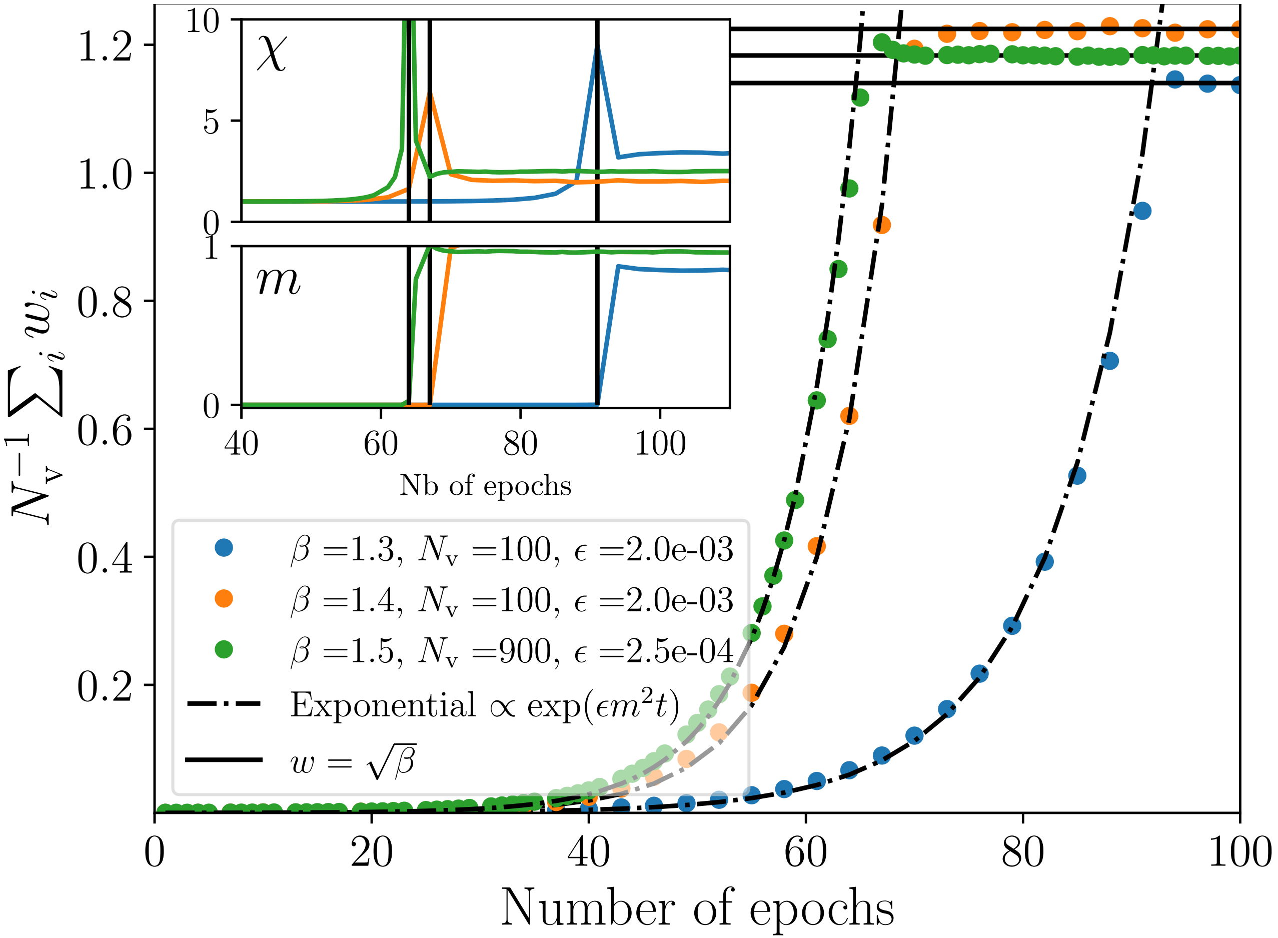
Three stages:

1. The positive/data term triggers the learning in the direction of the defined by the two lumps.
2. The weights are growing and undergoes a phase transition with a diverging susceptibility
3. The negative term counter-balance the data to cancel the gradient adjusting the lumps

Results

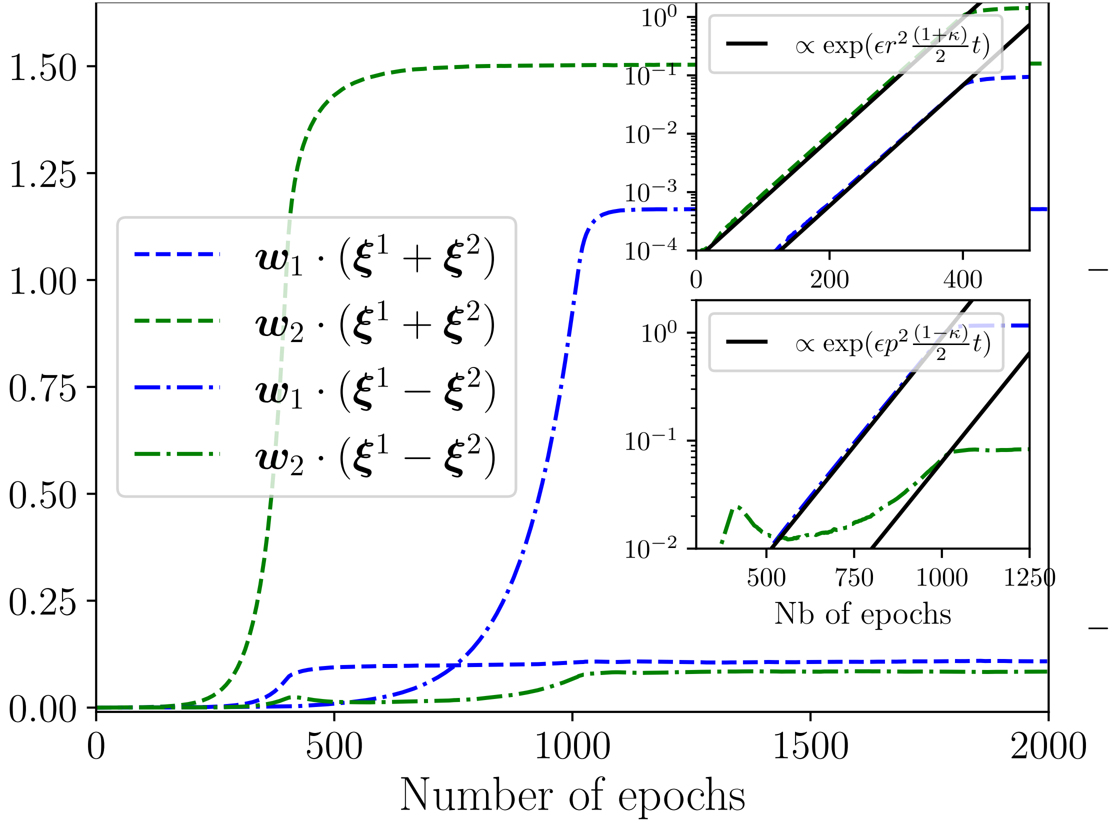
$m = \tanh(\beta m)$

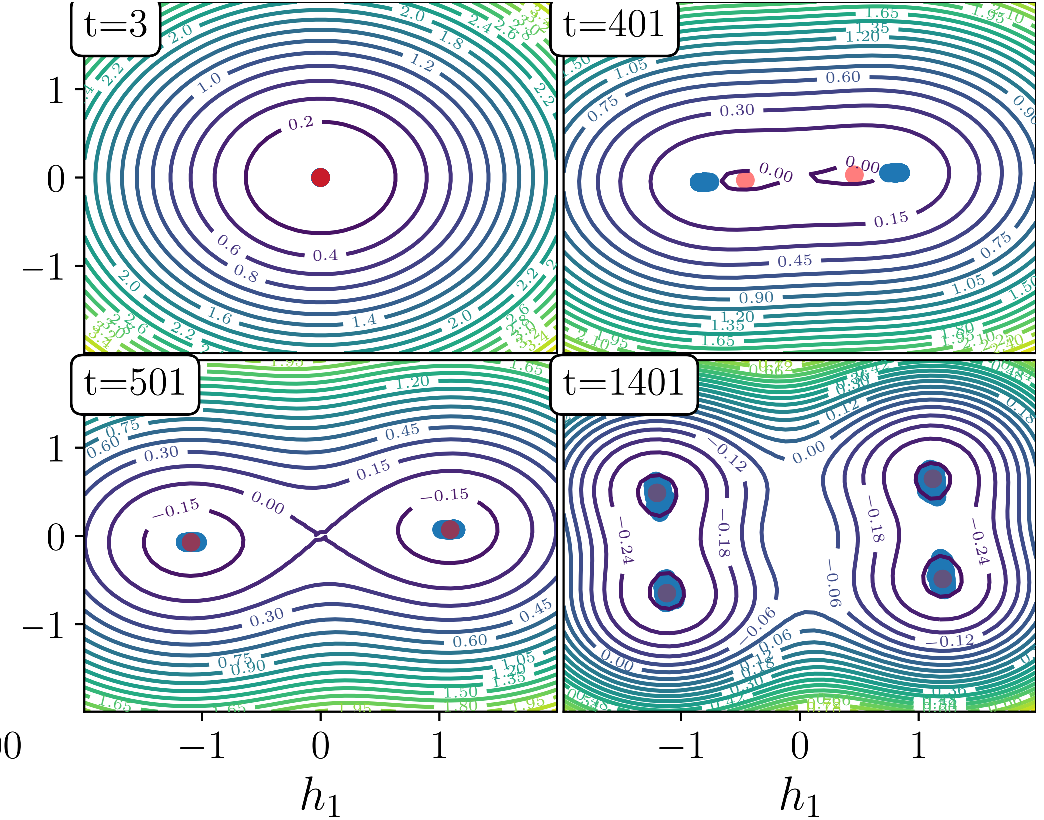
With two modes to learn
→ learning curve $\boldsymbol{w}(t)$



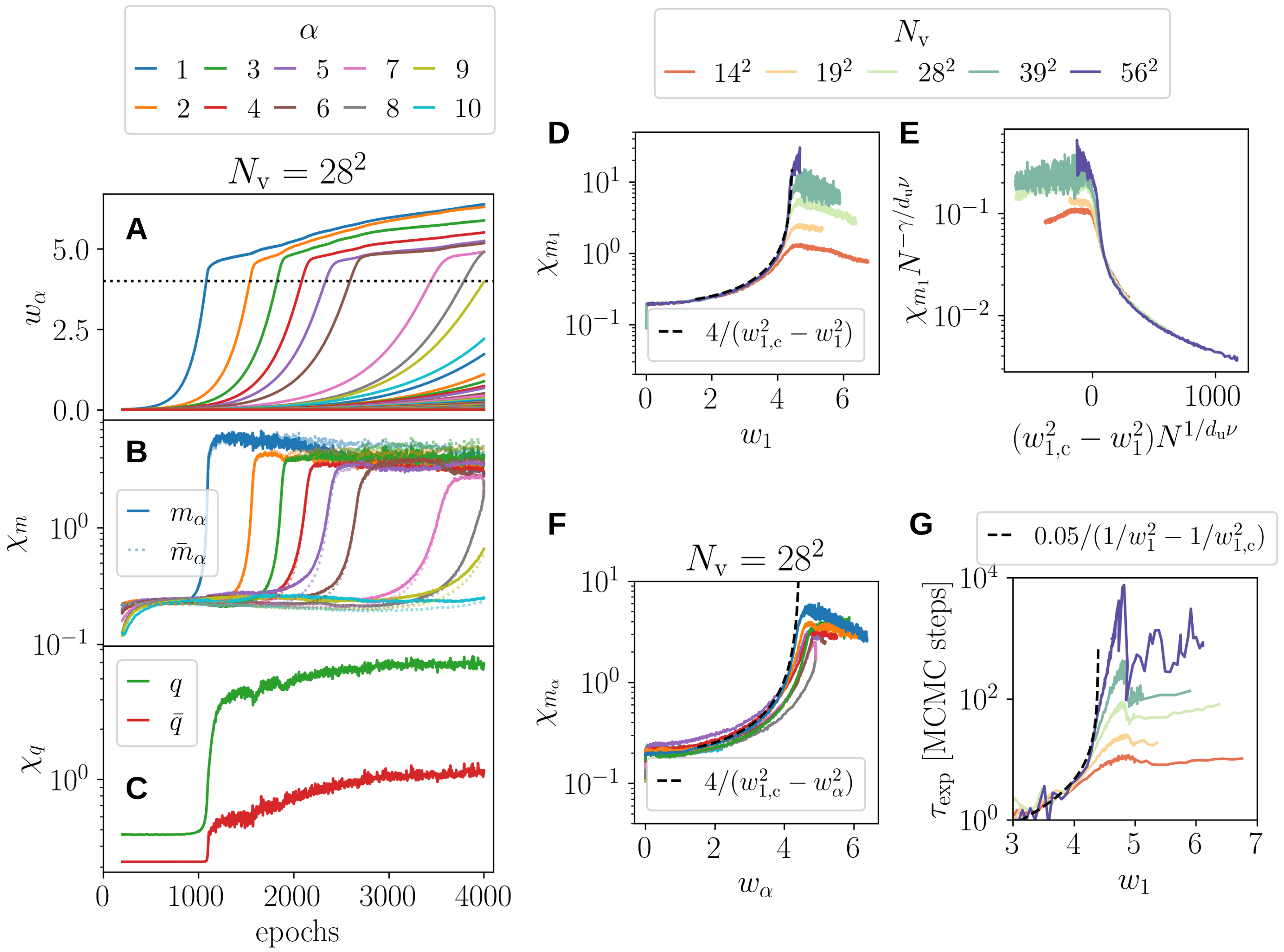
Results

With four non-orthogonal clusters, two learned directions $\boldsymbol{\xi}^1 + \boldsymbol{\xi}^2$ and $\boldsymbol{\xi}^1 - \boldsymbol{\xi}^2$

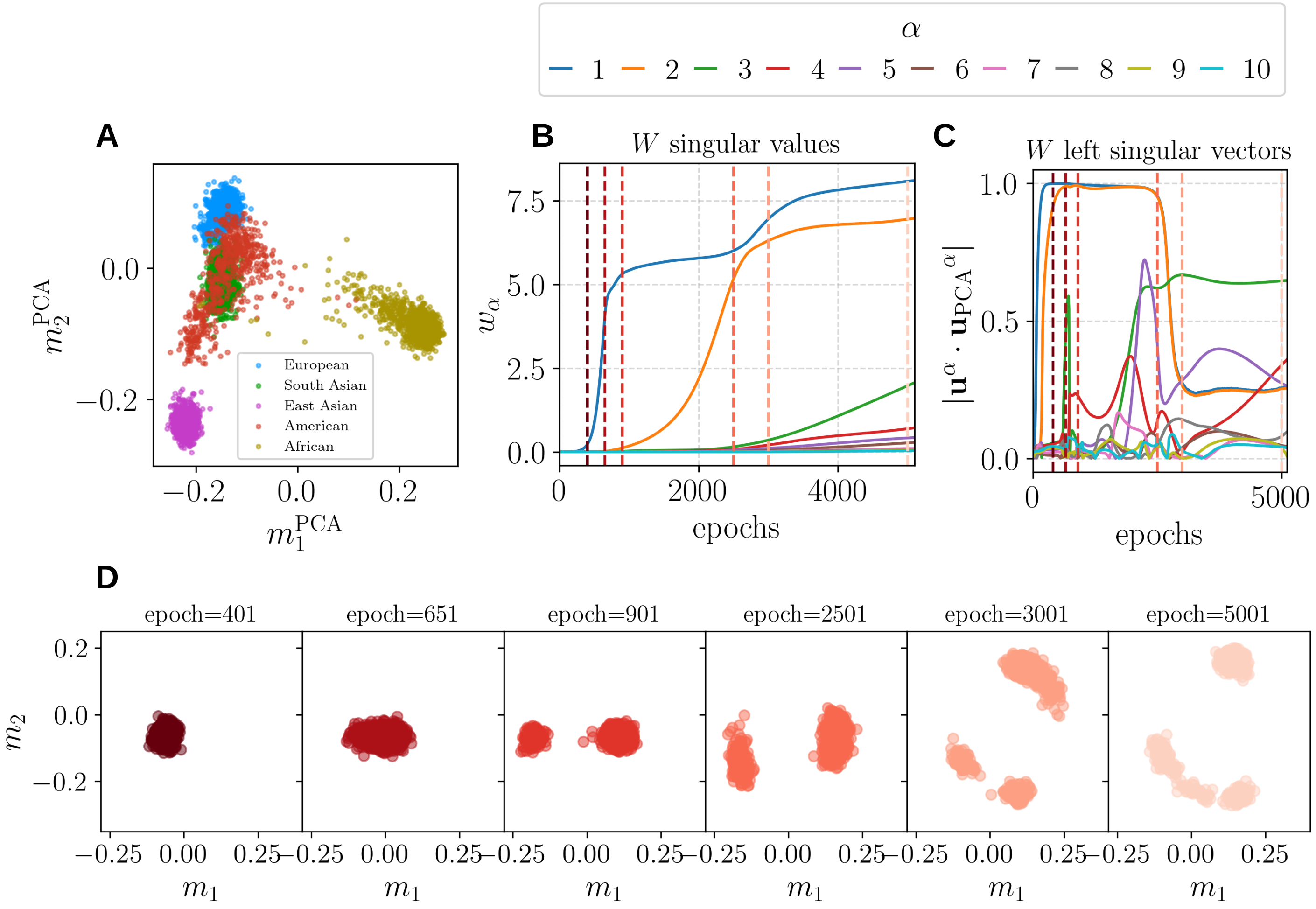




Numerical results on MNIST



Numerical results on Genetic



Conclusions

- We can characterize precisely the learning of RBMs theoretically in a simple setting
- The learning trajectory passes through several phase transitions which results in sharp sudden grows of the mixing time of the model
- Numertical experiment confirm the nature of the phase transition and its links to the Principal Component Analysis of the dataset
- These results should probably extend to generic Energy Based model
- We can not control the late training time in more complex situations