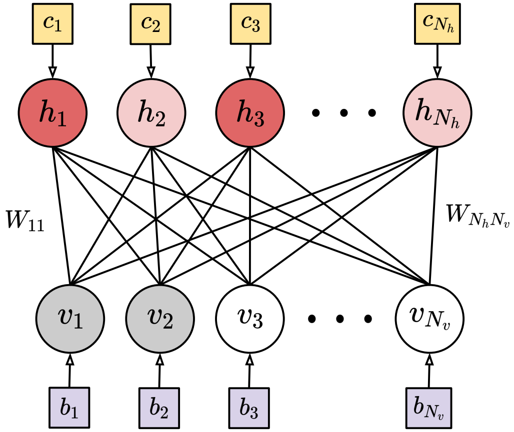
How phase transitions shape the learning of complex data in the Resctricted Boltzmann Machine



Bachtis et al NeurIPS '24
Rosset et al PRE '23

Aurélien Decelle – Universidad Complutense de Madrid

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

Madrid's group in Machine Learning

Giovanni Catania   Beatriz Seoane   AD   Lorenzo Rosset   Alfonso Navas Gomez

<u>Paris-Saclay</u>
PhD student
Nicolas Béreux

<u>Paris-Saclay</u>
Cyril Furtlehner

<u>ENS-Paris</u>
Tony Bonnaire
Giulio Biroli
Dimitrios Bachtis

Machine Learning and generative model

Generative modelling is a quite common task when dealing with, for instance, Bayesian inference

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{Z(\mathcal{D})}$$

To infer/learn the parameters $\boldsymbol{\theta}$ of some problem, we need to define the likelihood.
→ the likelihood is a generative model

Machine Learning and generative model

Some approaches taken by modern Machine Learning is to
→ use very « expressive » but not related to the data distribution (e.g. neural networks)
→ bypass the need to compute the likelihood (e.g. Generative Adversarial Networks)

Others such as Diffusion Models (see Biroli's talk) are based upon a different setting.

Supervised vs Unsupervised

Machine Learning tasks are often categorized in three categories

- **Supervised Learning**
- Unsupervised Learning

A dataset of M elements in dimension N, with labels (a class or real value)

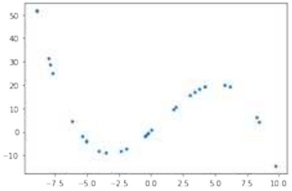
$$\{\boldsymbol{x}_m\}_{m=1,\dots,M} \qquad \{y_m\}_m$$

Example of classification

$x =$ 

$y =$ "cats"

Example of regression

$(x, y) =$ 

In both cases, we are looking to find the parameters of some function f that manage to predict the correct answer $f_{\theta^*}(x_m) = y_m$

Supervised vs Unsupervised

Machine Learning tasks are often categorized in three categories

- Supervised Learning
- Unsupervised Learning

A dataset of M elements in dimension N $\longrightarrow$ $\{\boldsymbol{x}_m\}_{m=1,\ldots,M}$

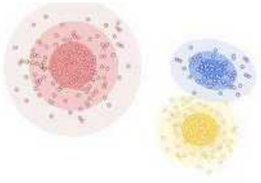
Then, in most settings we want to **learn a probability distribution** matching the empirical one

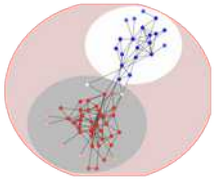
Example of generative models

$$x = $$

$$\hat{x} \sim p_{\boldsymbol{\theta}^*}(x)$$

Examples of clustering



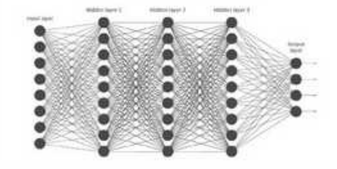


Example of generative modelling

This models is part of what is usually called “Unsupervised learning” or Generative model.
Its “purpose” is to learn a probability distribution based on a dataset.
Examples:

<u>pictures of celebrity</u>

Learning

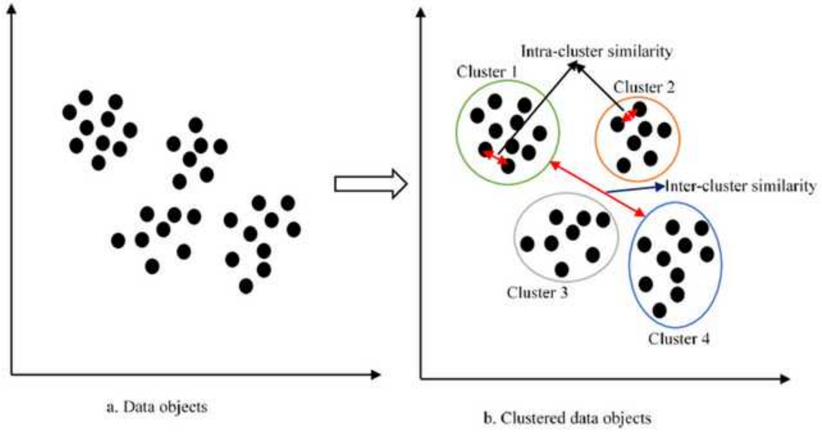


Generating

I do not exist

<u>automatic clustering</u>

Define a <u>mixture model</u>: distribution **P**
→ learns its parameters



Non-exhaustive list of generative models

- **Energy-based models**
- Variational AutoEncoder
- Generative Adversarial Models
- Autoregressive models
- Diffusion models

Nicely related to

→ Ising model and Inverse Ising problem
→ Boltzmann distribution
→ tracktable energy function can be used
→ the Hopfield model

Before...

Before I had to excuse myself for dealing with Boltzmann Machine when spealing about Machine Learning

Brief recall

Hopfield model : associative memory model → it recalls « planted » patterns

- introduced by Hopfield in ‘82
- AGS ‘85 replica theory to recall when storing an extensive number of patterns
- Dreaming mechanism to increase the recall regime (Dotsenko Kanter, Sompolinski ~90’, later Barra, Agliari et al. ~2019)
- More recently : Modern Hopfield model with exponential capacity

The Restricted Boltzmann Machine : generative model → it can generate new complex samples

- Smolensky ‘86, then popularized by Hinton with contrastive divergence ~2000
- It was use to extract features and train deep NN in ~2000→ 2010
- Re-discovered by physicists ~2010 : Barra, Agliari, Monasson, …
- Roots for energy-based models

The Restricted Boltzmann Machine from Hopfield to Hinton (and back?)

Recall on the Hopfield model (will be useful later) – we consider discrete spins $s_i = \pm 1$

$$\mathcal{H} = -\frac{1}{N}\sum_{i} J_{ij} s_i s_j = -\frac{1}{2N}\sum_{\mu}\left(\sum_{i} s_i \xi_i^{\mu}\right)^2$$

$$J_{ij} = \sum_{\mu} \xi_i^{\mu} \xi_j^{\mu}$$

$$\xi_i^{\mu} = \pm 1 \text{ with } p = 1/2 \text{ and } \mu = 1, \ldots, P = \alpha N$$

The RBM : from Hopfield to Hinton (and back?)

$$\mathcal{H} = -\frac{1}{2N}\sum_\mu \left(\sum_i s_i \xi_i^\mu\right)^2$$

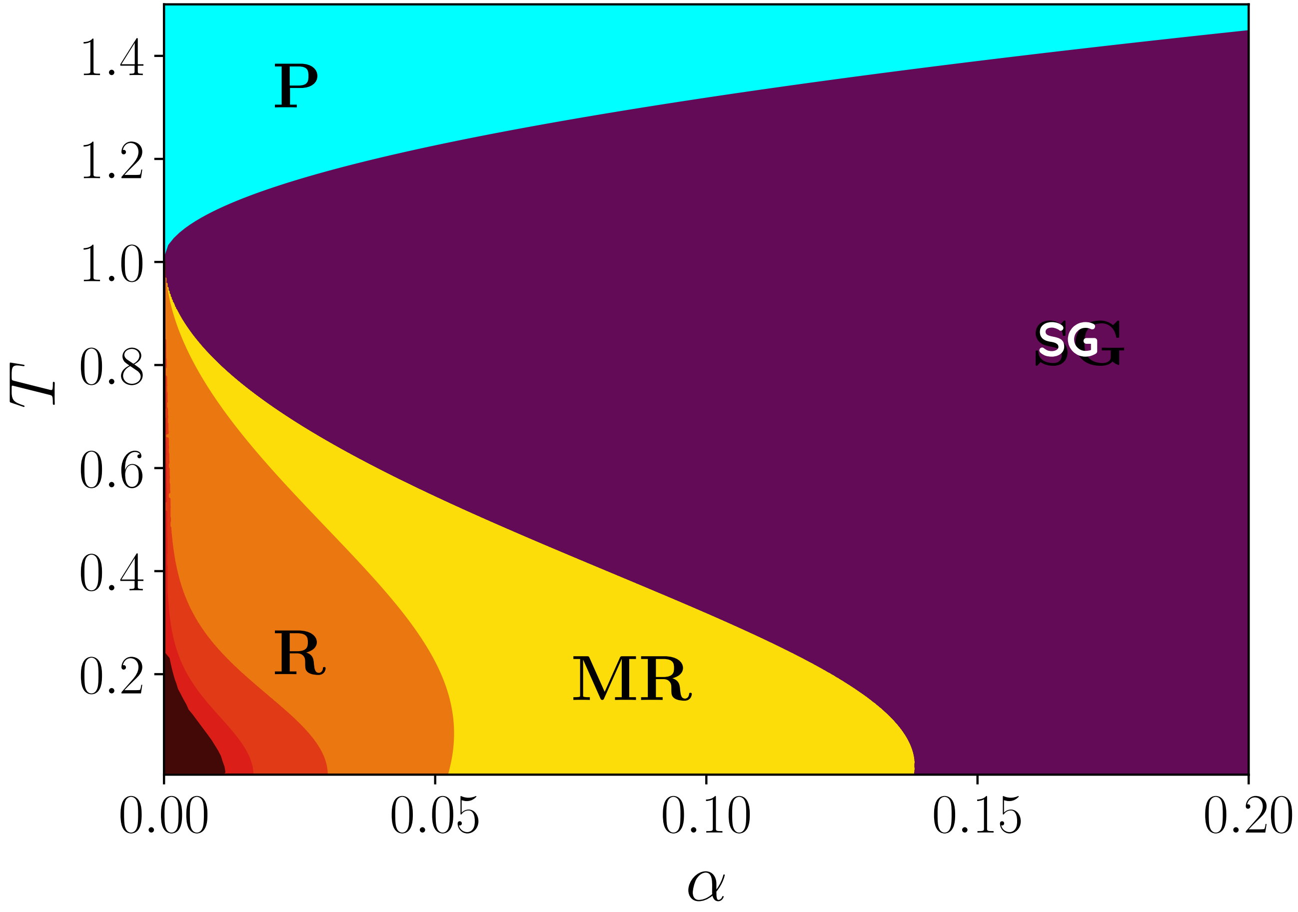
<u>Order parameters</u>

$$m = \frac{1}{N}\sum_i \xi_i^\mu \langle s_i \rangle$$

$$q = \frac{1}{N}\sum_i \mathbb{E}_{\boldsymbol{\xi}}\left[\langle s_i^a\rangle\langle s_i^b\rangle\right]$$

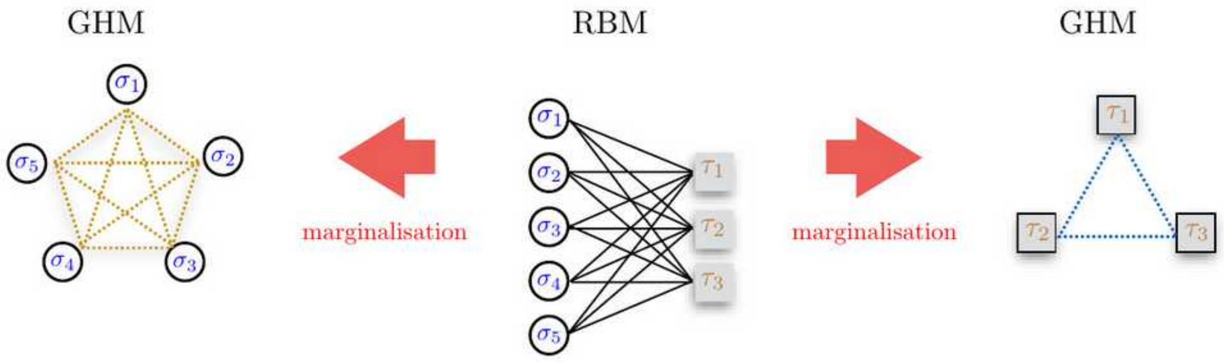
Different phases

- P : Paramagnetic $q,m=0$
- R : Recall $q,m\neq 0$
- SG : Spin Glass $q\neq 0, m=0$
- MR : Metastable Recall



From Hopfield to Bipartite architecture

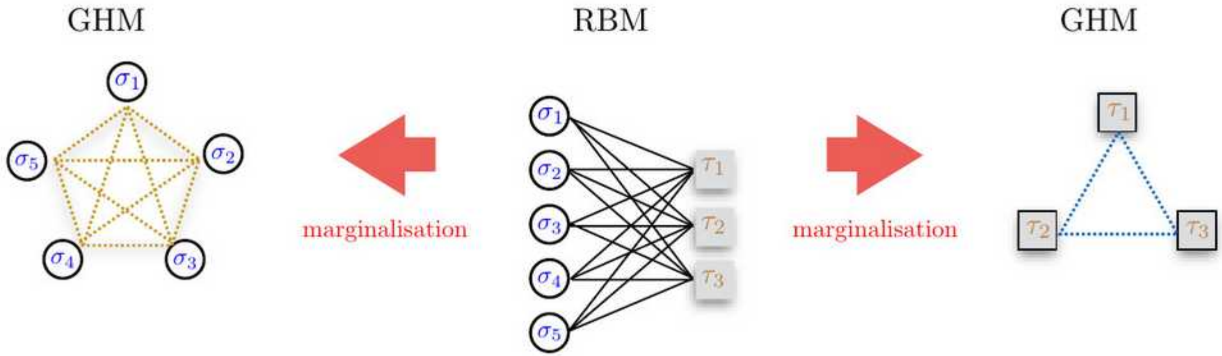
$$
\begin{aligned}
p(\boldsymbol{s}) &= \frac{1}{Z}\exp\left(-\beta\mathcal{H}[\boldsymbol{s}]\right) \\
&= \frac{1}{Z}\int d\boldsymbol{\tau}\exp\left(-\boldsymbol{\tau}^2/2+\sqrt{\frac{\beta}{N}}\sum_{i,\mu}s_i\xi_i^\mu\tau_\mu\right) \longrightarrow \mathcal{H}_{\mathrm{RBM}}[\boldsymbol{s},\boldsymbol{\tau}] = -\textstyle\sum_{i\mu}s_iw_{i\mu}\tau_\mu \\
&= \int d\boldsymbol{\tau}\,p_{\mathrm{RBM}}[\boldsymbol{s},\boldsymbol{\tau}]
\end{aligned}
$$



Credit to Barra et al. 2018

From Hopfield to Bipartite architecture

$$\mathcal{H}_{\mathrm{RBM}}[\boldsymbol{s},\boldsymbol{\tau}] = -\sum_{i\mu} s_i w_{i\mu} \tau_\mu$$



We have a bipartite system between
→ s_i a layer of discrete binary spins
→ τ_μ a layer of continuous gaussian variables

$$p(s_i|\boldsymbol{\tau}) = \tanh\left(\sum_\mu \xi_i^\mu \tau_\mu\right)$$

$$p(\tau_\mu|\boldsymbol{s}) \propto \exp\left[\left(x - \sum_i \xi_i^\mu s_i\right)^2 /2\right]$$

linear response from the hidden layer

Credit to Barra et al. 2018

From Hopfield to Bipartite architecture

$$\mathcal{H}_{\mathrm{RBM}}[\boldsymbol{s}, \boldsymbol{\tau}] = -\sum_{i\mu} s_i w_{i\mu} \tau_\mu$$

What if we change the nature of the hidden nodes ? $\tau_a = \pm 1$

$$p(s_i|\boldsymbol{\tau}) = \tanh\left(\sum_\mu \xi_i^\mu \tau_\mu\right)$$

$$p(\tau_\mu|\boldsymbol{s}) = \tanh\left(\sum_i \xi_i^\mu s_i\right)$$

Distribution on the spins :

$$p(\boldsymbol{s}) = \frac{1}{Z} \exp\left(\sum_\mu \left[\sum_i \xi_i^\mu s_i\right]^2 - A \sum_\mu \left[\sum_i \xi_i^\mu s_i\right]^4 + \mathcal{O}(s^6)\right)$$

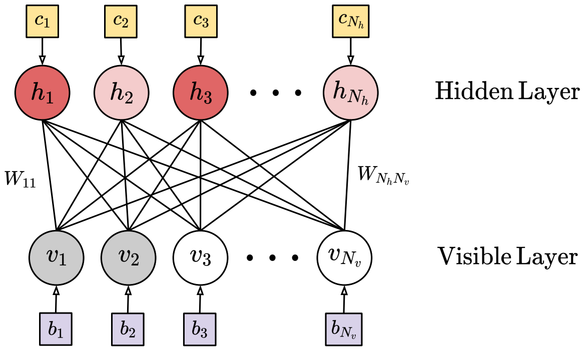
With non-linear response (not Gaussian), we can fit higher order statistics

The Restricted Boltzmann Machine

Discrete $s_i, \tau_a = \pm 1$ or $\{0,1\}$
Weights : $\{\boldsymbol{w}, \boldsymbol{a}, \boldsymbol{b}\}$

$$\mathcal{H}[\boldsymbol{s}, \boldsymbol{\tau}] = -\sum_{i,a} s_i w_{ia} \tau_a - \sum_i a_i s_i - \sum_a b_a \tau_a$$

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



The training is usually done by maximizing the likelihood

$$\mathcal{L} = \frac{1}{M} \sum_m \log\left(p(\boldsymbol{s}^{(m)})\right) - \log Z$$

$$\frac{dw_{ia}}{dt} \sim \frac{\partial \mathcal{L}}{\partial w_{ia}} = \langle s_i \tau_a \rangle_{\text{data}} - \boxed{\langle s_i \tau_a \rangle_{\mathcal{H}}}$$

<u>Curse of Monte Carlo</u>

The Restricted Boltzmann Machine

Challenges :

- Practical training aspects : Monte Carlo problem
- Learning dynamics
- Landscape of the learned Machine

Mean-Field approach

In the small-weight regime – typically at the beginning of the learning -, we can try to describe the probability distribution on a set of uncoupled variables.
→ typically naive-MF, MF approximation etc

$$p_{\text{indep}}(\boldsymbol{s},\boldsymbol{\tau}) = \prod_i p_i(s_i) \prod_a(\tau_a) \propto \prod_i e^{h_i s_i} \prod_a e^{h_a \tau_a}$$

$$\{h_i, h_a\} = \text{argmin} D_{KL}(p_{RBM}||p_{\text{indep}})$$

$$\longrightarrow$$

$$m_i = \tanh\left(\sum_a w_{ia} m_a + a_i\right)$$

$$m_a = \tanh\left(\sum_i w_{ia} m_i + b_a\right)$$

<u>**Singular Values Eqs**</u>
(in the linear regime)

$$\boldsymbol{m}^{(vis)} = \boldsymbol{W}\boldsymbol{m}^{(hid)}$$

$$\boldsymbol{m}^{(hid)} = \boldsymbol{W}^T\boldsymbol{m}^{(vis)}$$

The paramagnetic fixed point is unstable for $\lambda_{max} = 1$

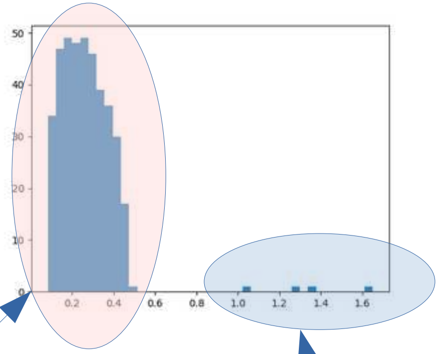
Mean-Field approach

In the linear regime, the properties of the RBM is dominated by the spectral properties or $\boldsymbol{W}$

Singular Values Eqs
(in the linear regime)

$$\boldsymbol{m}^{(vis)} = \boldsymbol{W}\boldsymbol{m}^{(hid)}$$
$$\boldsymbol{m}^{(hid)} = \boldsymbol{W}^T\boldsymbol{m}^{(vis)}$$

Spectrum



noise

signals

Consider the low-rank model matrix

$$w_{ia} = \sum_{\alpha=1}^{K} u_i^\alpha w_\alpha \bar{u}_a^\alpha + r_{ia}$$

$$r_{ia} \sim \mathcal{N}(0, \sigma)$$

$\boldsymbol{r}$,$\boldsymbol{u}$,$\bar{\boldsymbol{u}}$: quenched average
w_α are fixed

Mean-Field approach

$$L = \sqrt{N_h N_v}$$

$$s_\alpha = \sum_i s_i u_i^\alpha$$

$$\tau_\alpha = \sum_a \tau_a v_a^\alpha$$

Order parameters

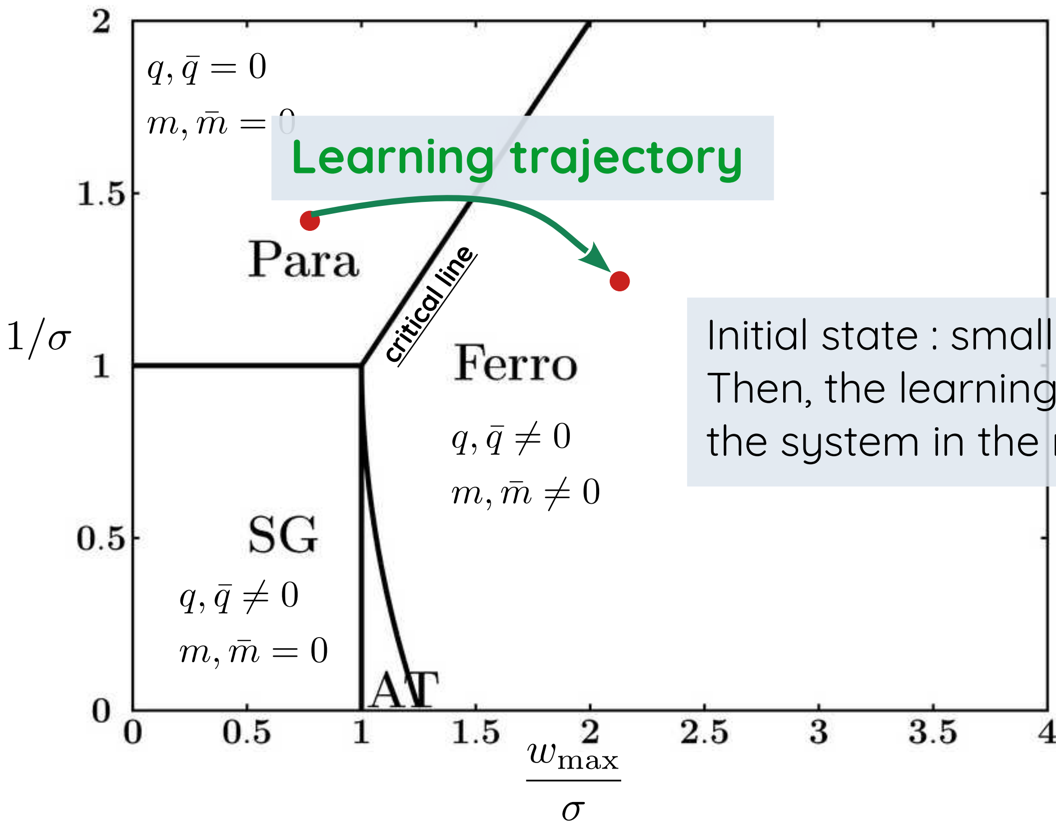
$$m_\alpha \sim \frac{1}{\sqrt{L}} \mathbb{E}_{\boldsymbol{u},\bar{\boldsymbol{u}},\boldsymbol{r}}\left(\langle \sigma_\alpha \rangle\right)$$

$$\bar{m}_\alpha \sim \frac{1}{\sqrt{L}} \mathbb{E}_{\boldsymbol{u},\bar{\boldsymbol{u}},\boldsymbol{r}}\left(\langle s_\alpha \rangle\right)$$

$$Q_{12} \sim \mathbb{E}_{\boldsymbol{u},\bar{\boldsymbol{u}},\boldsymbol{r}}\left(\langle \tau_a^1 \tau_a^2 \rangle\right)$$

$$\bar{Q}_{12} \sim \mathbb{E}_{\boldsymbol{u},\bar{\boldsymbol{u}},\boldsymbol{r}}\left(\langle s_j^1 s_j^2 \rangle\right)$$

Hopfield Recall-like



Initial state : small weight (para)
Then, the learning dynamics drives the system in the recall phase

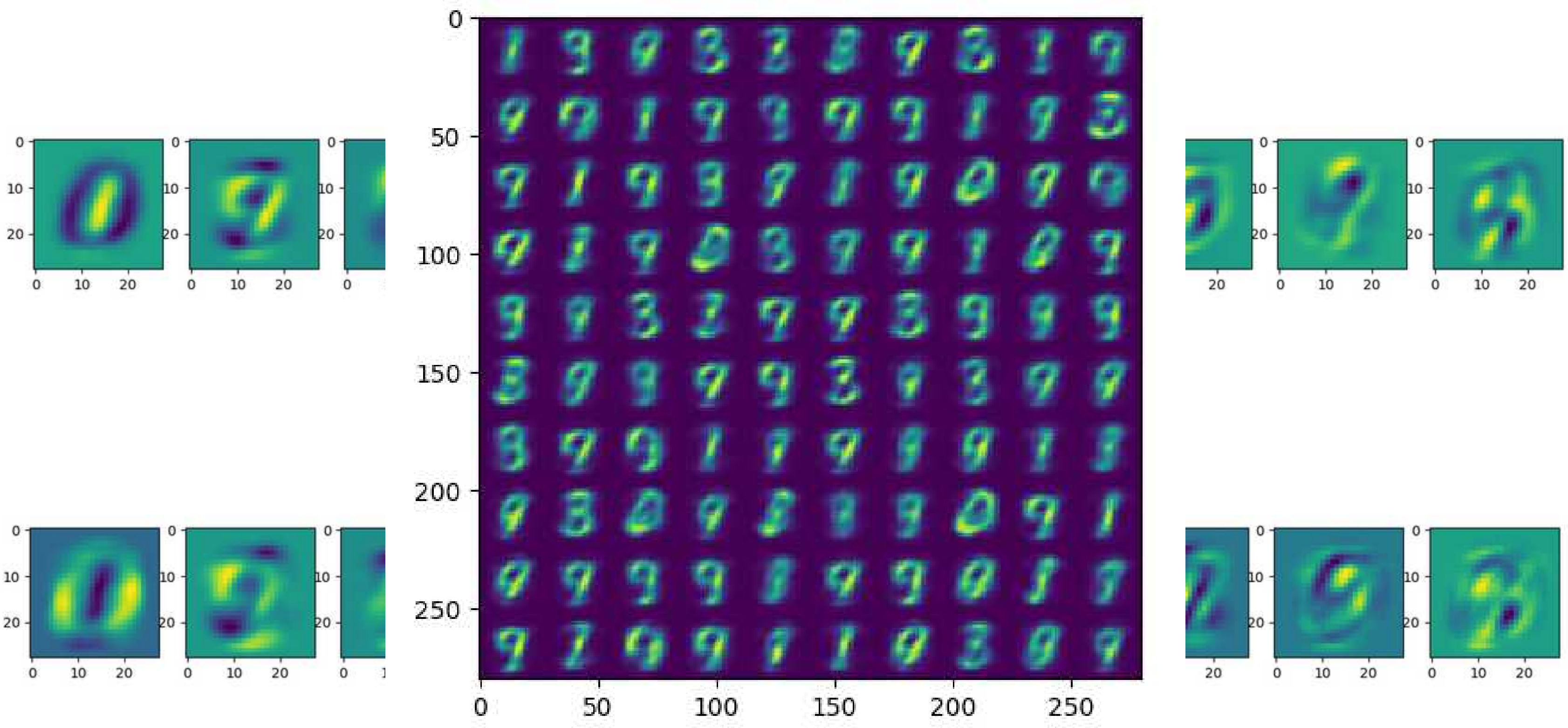
Linear learning dynamics

We can confirm this picture first by computing the gradient in the linear regime, in the SVD spa ce of the $\boldsymbol{W}$ matrix

$$\frac{dw_\alpha}{dt} = w_\alpha\Big(\langle s_\alpha^2\rangle_{\mathrm{Data}} - 1\Big)$$

$$\bar{u}^\alpha \frac{d\boldsymbol{u}^\beta}{dt} = \Omega^u_{\alpha\beta} = (1-\delta_{\alpha\beta})\left(\frac{w_\beta - w_\alpha}{w_\alpha + w_\beta} - \frac{w_\beta + w_\alpha}{w_\alpha - w_\beta}\right)\langle s_\alpha s_\beta\rangle_{\mathrm{Data}}$$

Empirical evidence on MNIST



Theory of learning dynamics
with a simplified model of data

We can solve the gradient equation in a simplified case
→ we consider one Gaussian hidden node
→ the dataset is generated by a Curie Weiss model in the low temperature regime

$$p_{\mathrm{CW}}(\boldsymbol{s}) = \tfrac{1}{Z}\exp\left(\beta \textstyle\sum_{i<j} \xi_i \xi_j s_i s_j\right)$$

$$p_{\mathrm{HF}}(\boldsymbol{s}, \tau) = \tfrac{1}{Z}\exp\left(\textstyle\sum_i s_i \tau w_i - \tfrac{\tau^2 N}{2}\right)$$

$\boldsymbol{\xi}$ a preferred direction

The gradient for the weight matrix is given by

$$\begin{aligned}\frac{dw_i}{dt} &= \langle s_i \tau \rangle_{\mathrm{data}} - \langle s_i \tau \rangle_{\mathcal{H}} \\ &= N^{-1} \sum_j \langle s_i s_j \rangle_{\mathrm{data}} w_j - N^{-1} \sum_j \langle s_i s_j \rangle_{\mathcal{H}} w_j\end{aligned}$$

Theory of learning dynamics
with a simplified model of data

The gradient for the weight matrix is given by

$$\begin{aligned}\frac{dw_i}{dt} &= \langle s_i \tau \rangle_{\mathrm{data}} - \langle s_i \tau \rangle_{\mathcal{H}} \\ &= N^{-1} \sum_j \langle s_i s_j \rangle_{\mathrm{data}} w_j - N^{-1} \sum_j \langle s_i s_j \rangle_{\mathcal{H}} w_j\end{aligned}$$

The correlation of the dataset is given by

$$\langle s_i s_j \rangle_{\mathrm{data}} = \xi_i \xi_j m^2 \text{ where } m = \tanh(\beta m)$$

The correlation of the RBM is given by

$$\langle s_i s_j \rangle_{\mathcal{H}} = \tanh(h^* w_i) \tanh(h^* w_j) \text{ where } h^* = \tfrac{1}{N} \textstyle\sum_k w_k \tanh(h^* w_k)$$

Theory of learning dynamics
with a simplified model of data

The correlation of the dataset is given by

$$\langle s_i s_j \rangle_{\text{data}} = \xi_i \xi_j m^2 \text{ where } m = \tanh(\beta m)$$

The correlation of the RBM is given by

$$\langle s_i s_j \rangle_{\mathcal{H}} = \tanh(h^* w_i) \tanh(h^* w_j) \text{ where } h^* = \tfrac{1}{N} \textstyle\sum_k w_k \tanh(h^* w_k)$$

$\langle s_i s_j \rangle_{\mathcal{H}} \approx 0$ for small $\boldsymbol{W}$

$$\Rightarrow \quad \frac{dw_i}{dt} = \tfrac{1}{N} \xi_i \textstyle\sum_j \xi_j w_j m^2$$

We can project the equations on

$$u_i = N^{-1/2} \xi_i$$

$$U_{\boldsymbol{\xi}} = \sum_i u_i w_i$$

Theory of learning dynamics
with a simplified model of data

$\frac{dU_{\boldsymbol{\xi}}}{dt} = U_{\boldsymbol{\xi}} m^2 \Rightarrow U_{\boldsymbol{\xi}}(t) = U_{\boldsymbol{\xi}}(0)\exp(m^2 t)$ Exponential growth in the direction of $\boldsymbol{\xi}$

As the weights grow, we can for instance monitor the suceptibility of the model

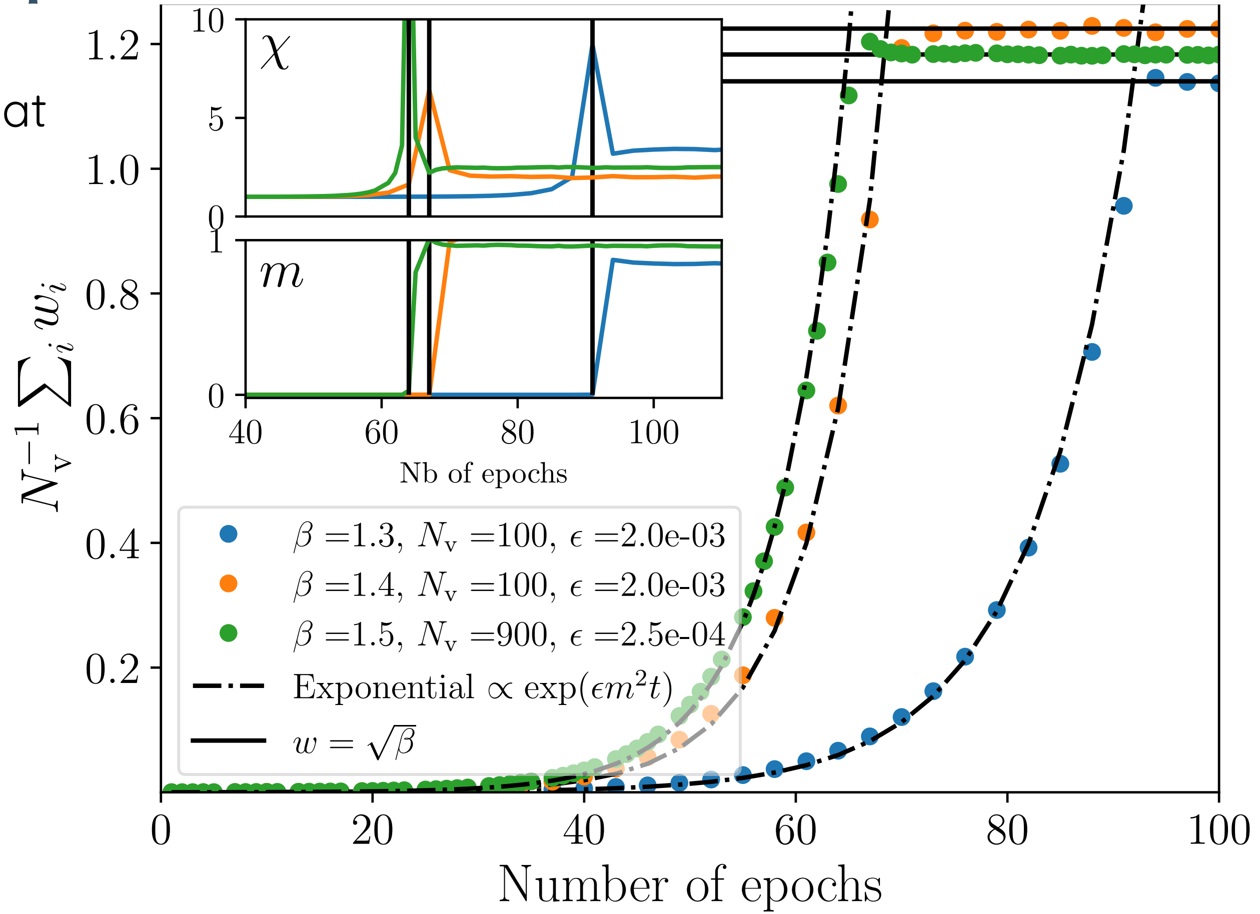
$$\chi = \sum_{i,j} \xi_i \xi_j \langle s_i s_j \rangle_{\mathcal{H}} \approx \left(\xi_i s_i\right)^2 \frac{1}{N(1-\sum_i w_i^2/N)}$$

It diverges as $\frac{1}{N}\sum_i w_i^2 \sim 1$

→ signal of a phase transition, the magnetization departs from zero, the critical exponent associated to the sucesptibility is $\gamma = 1$

Theory of learning dynamics
with a simplified model of data

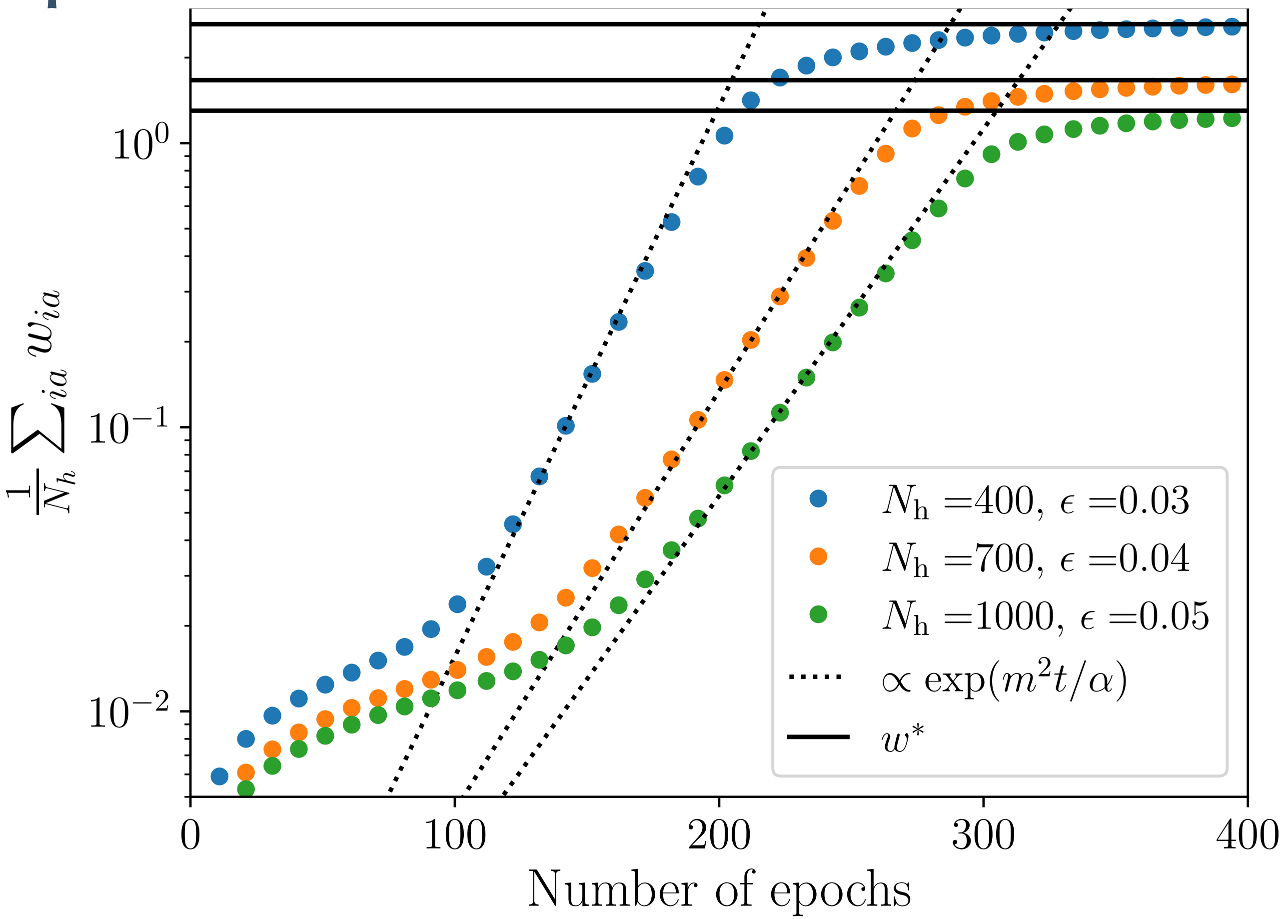
At late learning time, we can show that
→ the orthogonal directions to $\boldsymbol{\xi}$ are suppressed
→ $w_i = w\,\xi_i$ and $w = \sqrt{\beta}$



Theory of learning dynamics
with a simplified model of data

We can also analyze the dynamics in the case of a binary-binary RBM

$$\alpha = \frac{N_h}{N}$$



Theory of learning dynamics
with a simplified model of data

It is actually possible to study a problem with two correlated patterns

$$p(\boldsymbol{s}) = \frac{1}{Z} \exp\left(\frac{\beta}{2} \sum_a \left[\sum_i \xi_i^a s_i \right]^2 \right)$$

$$\boldsymbol{\xi}^1 = \boldsymbol{\eta}^1 + \boldsymbol{\eta}^2$$
$$\boldsymbol{\xi}^2 = \boldsymbol{\eta}^1 - \boldsymbol{\eta}^2$$

$$\eta_i^1 = \begin{cases} \pm 1 \text{ if } 1 \leq i \leq N^{\frac{1+\kappa}{2}} \\ 0 \text{ otherwise} \end{cases}$$

$$\eta_i^2 = \begin{cases} \pm 1 \text{ if } N^{\frac{1+\kappa}{2}} + 1 \leq i \leq N \\ 0 \text{ otherwise} \end{cases}$$

Theory of learning dynamics
with a simplified model of data

$$\boldsymbol{\xi}^1 = \boldsymbol{\eta}^1 + \boldsymbol{\eta}^2$$
$$\boldsymbol{\xi}^2 = \boldsymbol{\eta}^1 - \boldsymbol{\eta}^2$$

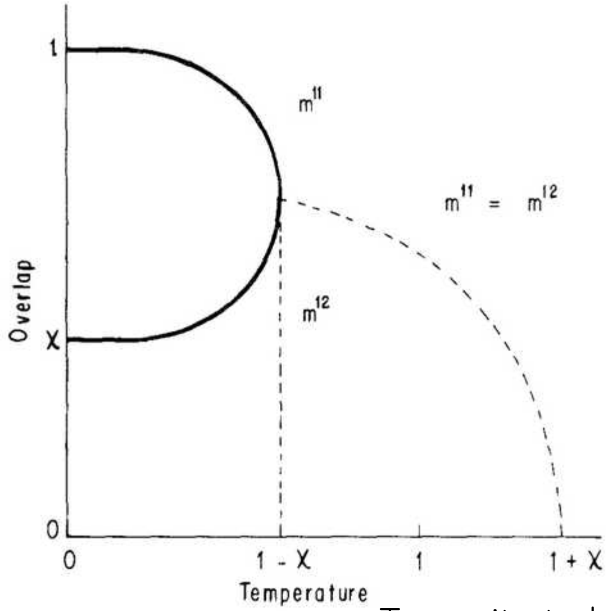
Phase diagram of such model :

$$m_1 = \frac{1+\kappa}{2}\tanh\left(\beta(m_1+m_2)\right) + \frac{1-\kappa}{2}\tanh\left(\beta(m_1-m_2)\right)$$
$$m_2 = \frac{1+\kappa}{2}\tanh\left(\beta(m_1+m_2)\right) - \frac{1-\kappa}{2}\tanh\left(\beta(m_1-m_2)\right)$$

At $T_1 = 1+\kappa$, magnetization along the direction $\boldsymbol{\xi}^1$

$m = \frac{1+\kappa}{2}\tanh\left(\beta 2m\right)$

At $T_2 = 1\text{-}\kappa$, $m_1 \neq m_2$



Tamarit et al. 1990

Theory of learning dynamics
with a simplified model of data

We can decompose the correlation function of the dataset upon the SVD

$$\langle s_i s_j \rangle_{\mathcal{D}} = r^2 \eta_i^1 \eta_j^1 + p^2 \eta_i^2 \eta_j^2$$

where $r = \tanh(\beta(m^+ + m^-))$

$p = \tanh(\beta(m^+ - m^-))$

$$m^+ = \max(m_1, m_2) \text{ and } m^- = \min(m_1, m_2)$$

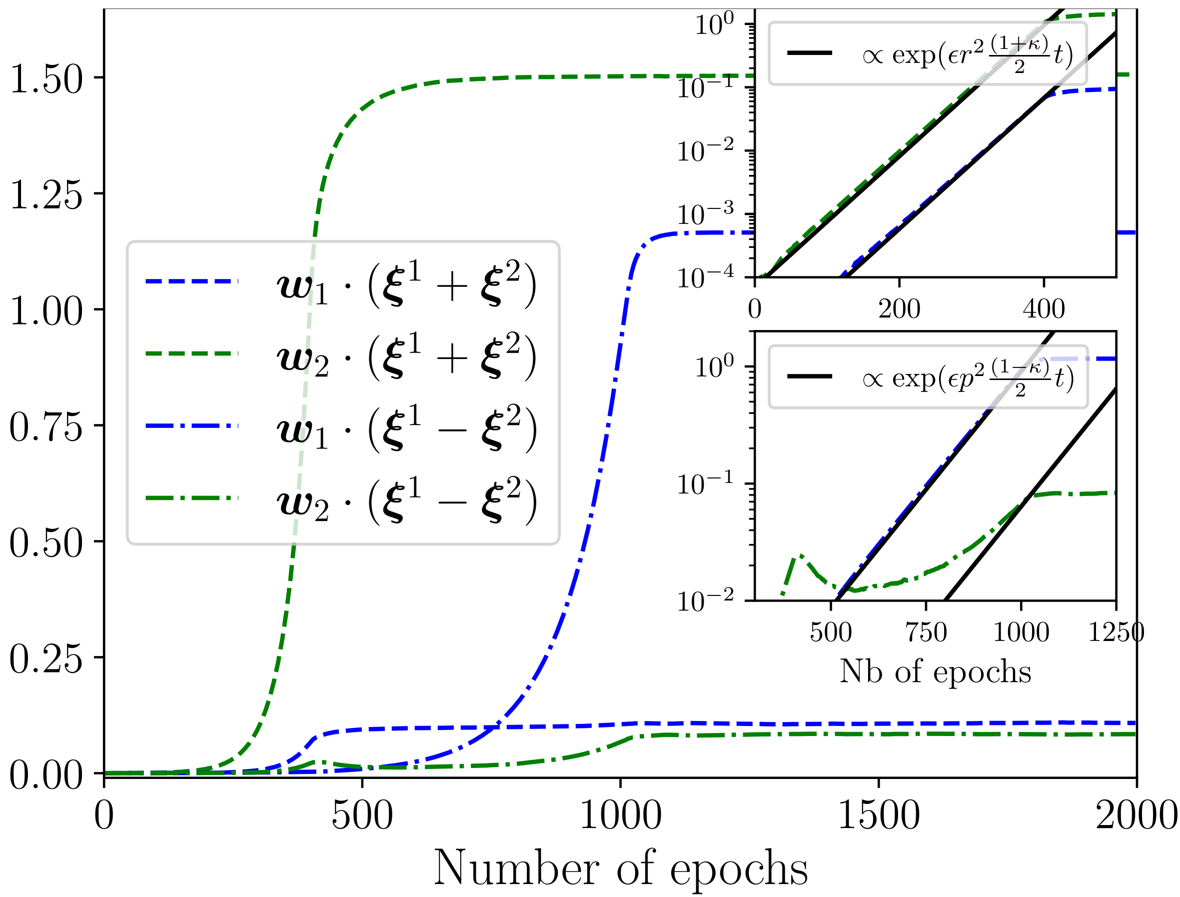
$$m_1 = \frac{1+\kappa}{2} \tanh\left(\beta(m_1 + m_2)\right) + \frac{1-\kappa}{2} \tanh\left(\beta(m_1 - m_2)\right)$$
$$m_2 = \frac{1+\kappa}{2} \tanh\left(\beta(m_1 + m_2)\right) - \frac{1-\kappa}{2} \tanh\left(\beta(m_1 - m_2)\right)$$

Theory of learning dynamics
with a simplified model of data

We can control the growth into each direction :

$$\boldsymbol{w}^a(t) = \frac{z^a}{\sqrt{\left(\frac{1+\kappa}{2}\right)}} e^{r^2\left(\frac{1+\kappa}{2}\right)t}\boldsymbol{\eta}^1 + \frac{\tilde{z}^a}{\sqrt{\left(\frac{1-\kappa}{2}\right)}} e^{p^2\left(\frac{1-\kappa}{2}\right)t}\boldsymbol{\eta}^2$$

$a = 1, 2$

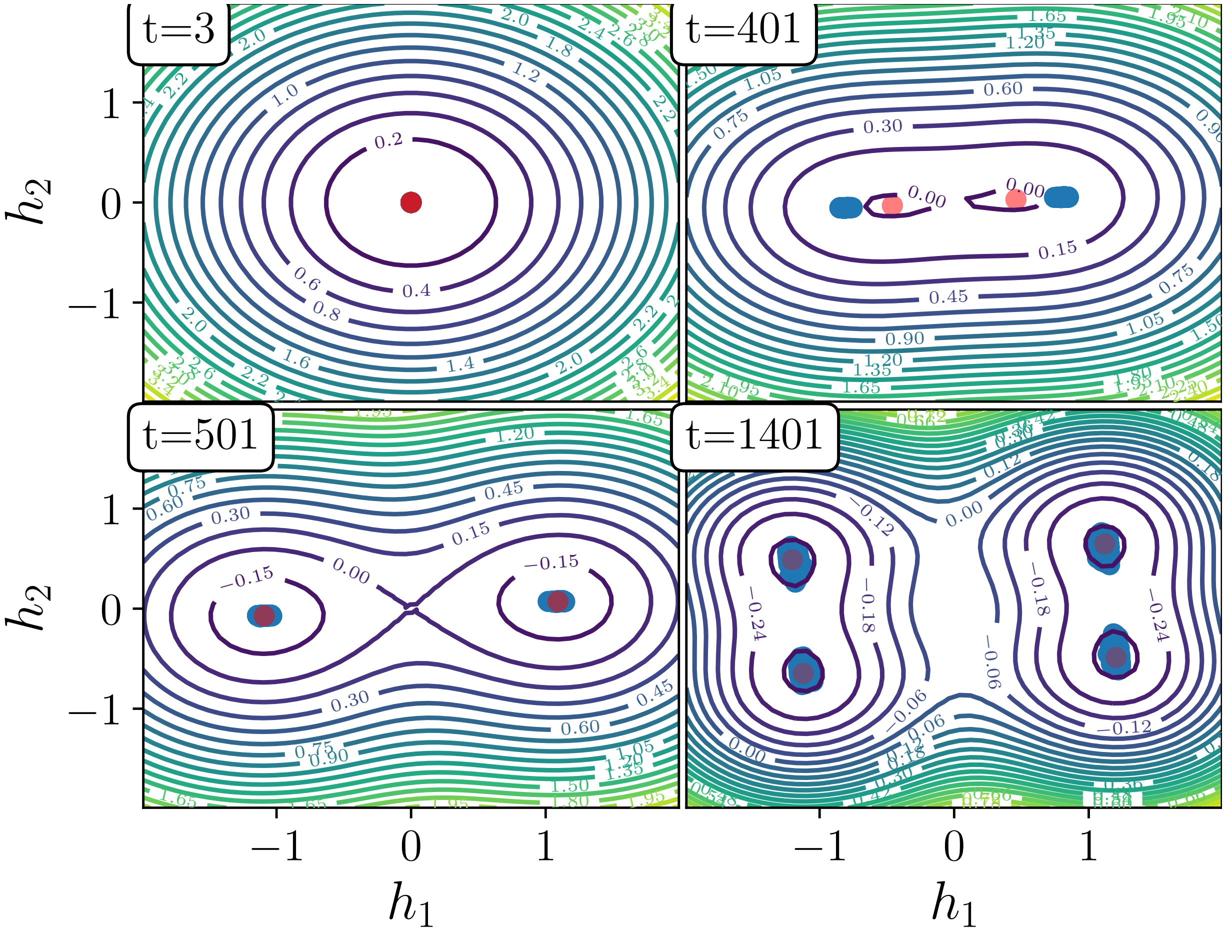


Theory of learning dynamics
with a simplified model of data

We can control the growth into each direction :

$$\boldsymbol{w}^a(t) = \frac{z^a}{\sqrt{\left(\frac{1+\kappa}{2}\right)}} e^{r^2\left(\frac{1+\kappa}{2}\right)t}\boldsymbol{\eta}^1 + \frac{\tilde{z}^a}{\sqrt{\left(\frac{1-\kappa}{2}\right)}} e^{p^2\left(\frac{1-\kappa}{2}\right)t}\boldsymbol{\eta}^2$$

$a = 1, 2$



Free energy as a function of h_1, h_2

Numerical evidence

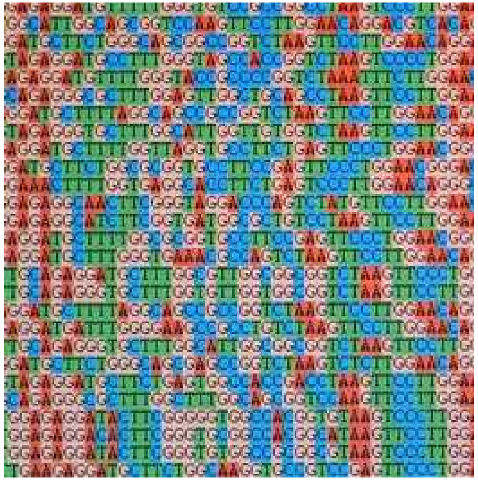
Are those transitions observed in this simple regime meanningful ?
We the behavior of several training on various dataset :
→ MNIST
→ genetic dataset
→ CelebA

MNIST

28x28 pixels
10.000 samples

CelebA

128x128 pixels
30.000 samples



The 1000 Genomes
Project Consortium

805 bases
4500 samples

Numerical evidence

What do we want to observe :

→ phase transition as the eigenvalues pass a certain threshold

→ critical exponent $\chi = -\frac{\partial^2 \beta F}{\partial h^2} \sim_{T \to T_c} (T - T_c)^{-\gamma}$ (variance of the order parameter)

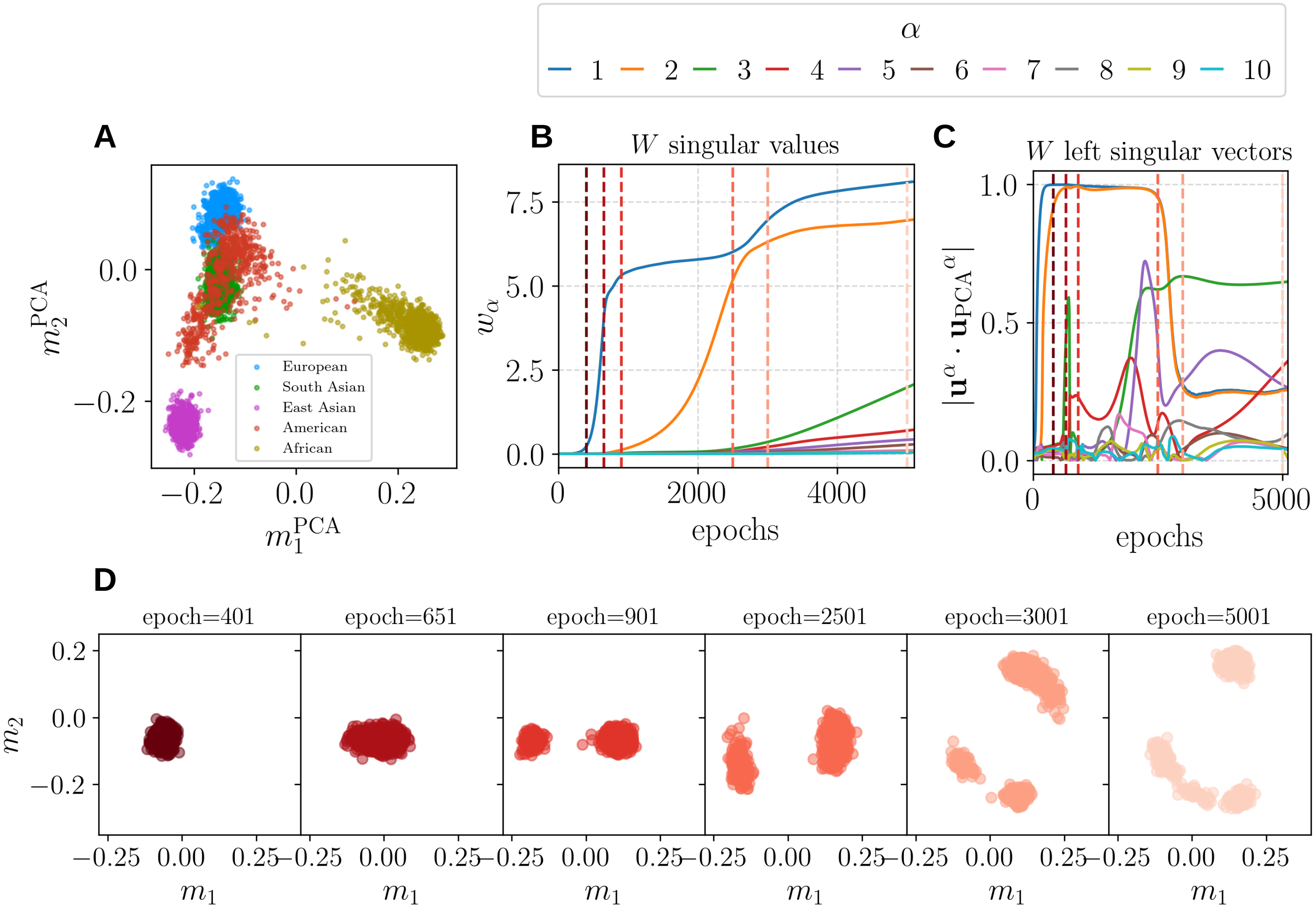
→ relaxation time $\tau_{\mathrm{exp}} \sim_{T \to T_c} N^z$

→ hysteresis ? … first order transition *in field*

Here, we will use binary {0,1} variables, for which the phase transition threshold is $\lambda = 4$

DNA dataset

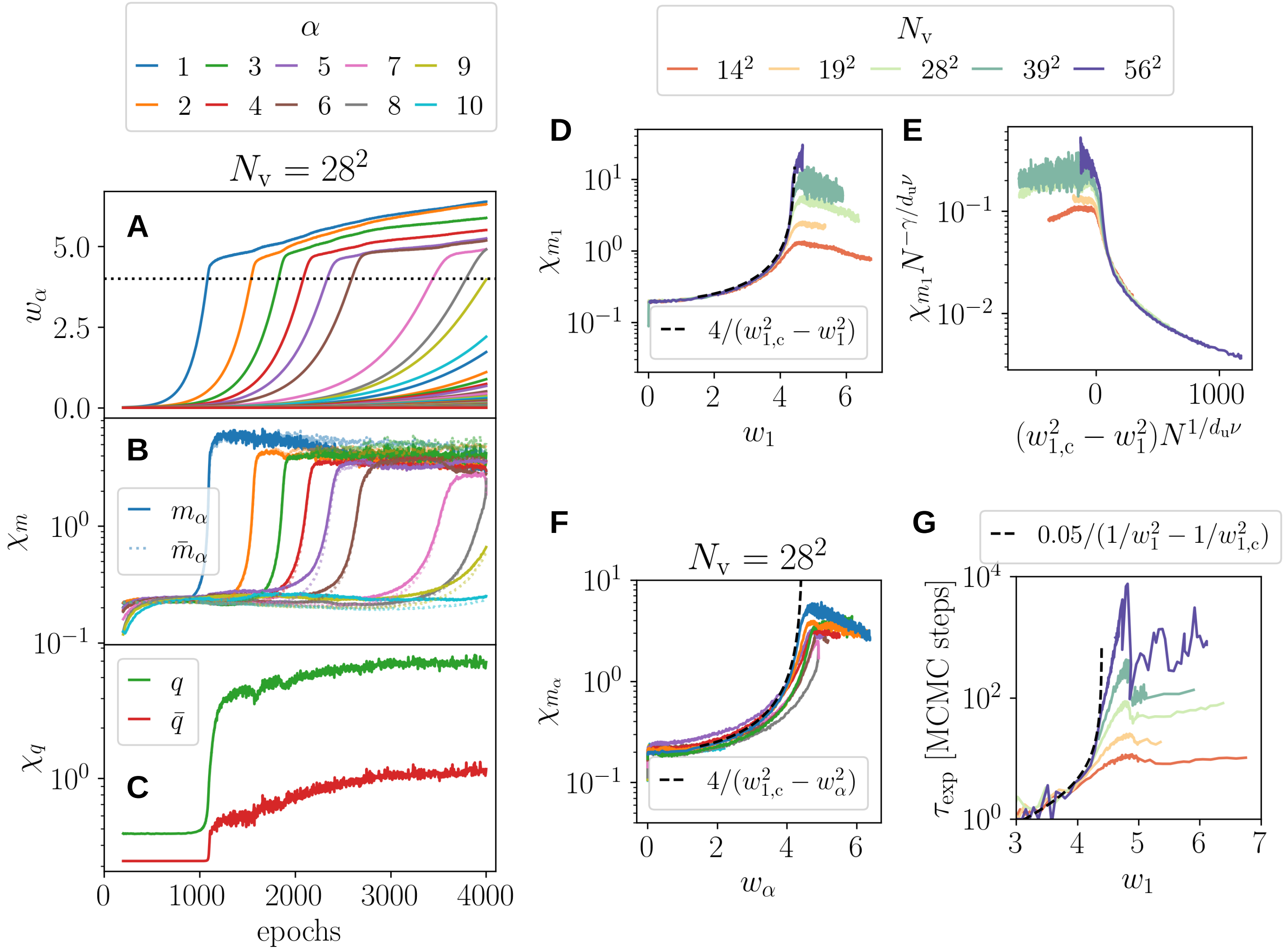
Numerical evidence



Samples →

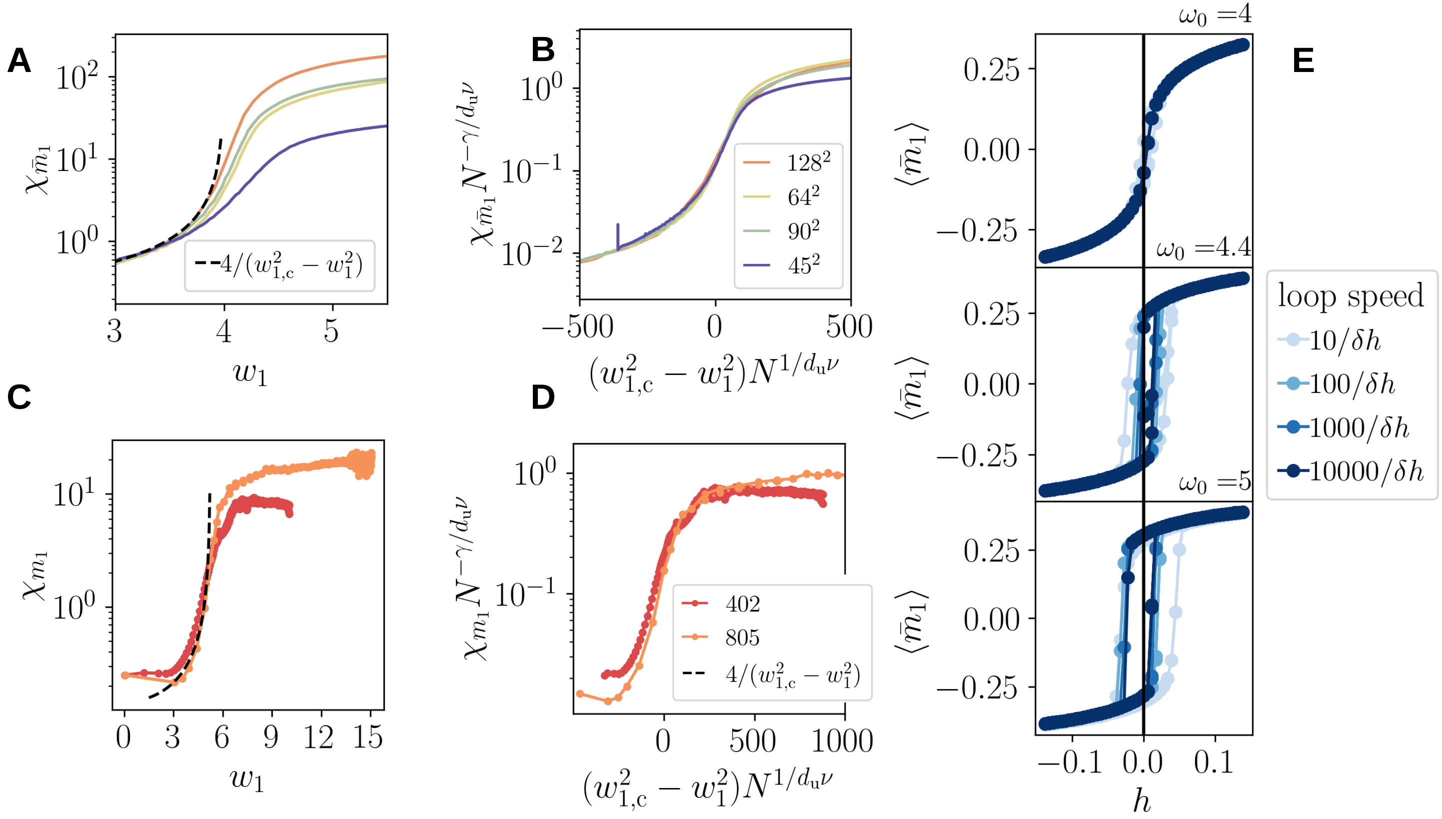
MNIST dataset

Numerical evidence



CelebA dataset

Numerical evidence

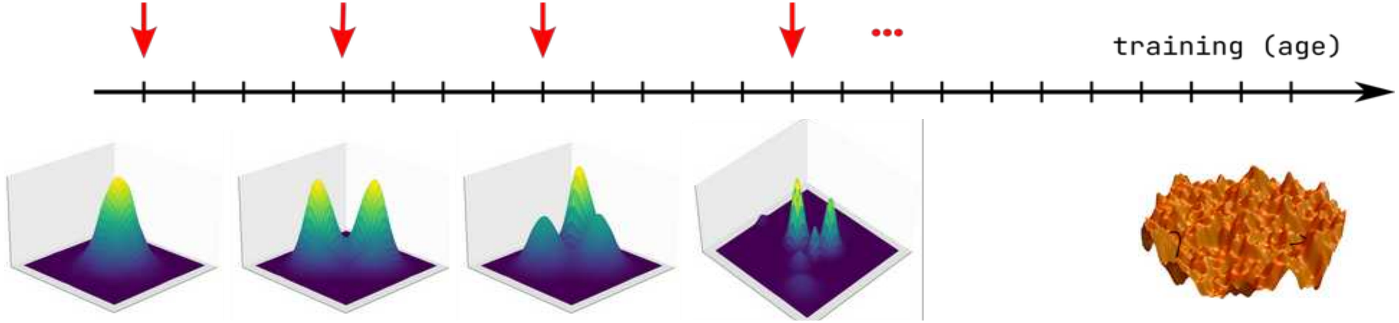


How is that useful for ?

- Be carefull to the relaxation time (now you know) !
- Monitor the learning of the model
- Do these phenomena happen in other generative models (e.g. Diffusion) ?
- **You might want to use the « cascade of phase transition » to « understand » the model**

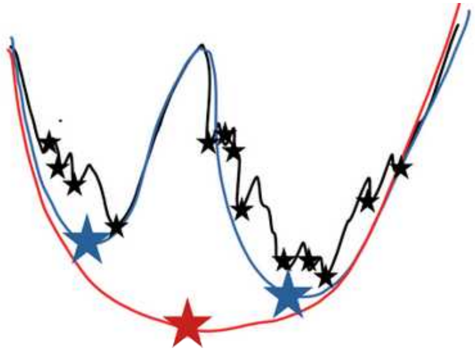
Hierarchical carving

Now that we know that the landscape is shaped by a sequence of phase transitions, we can try to use them to explore what the RBM is learning.



Decelle, A., Rosset, L., & Seoane, B. PRE (2023)

We follow the maximum of the prob. Dist. using the Mean-Field equations



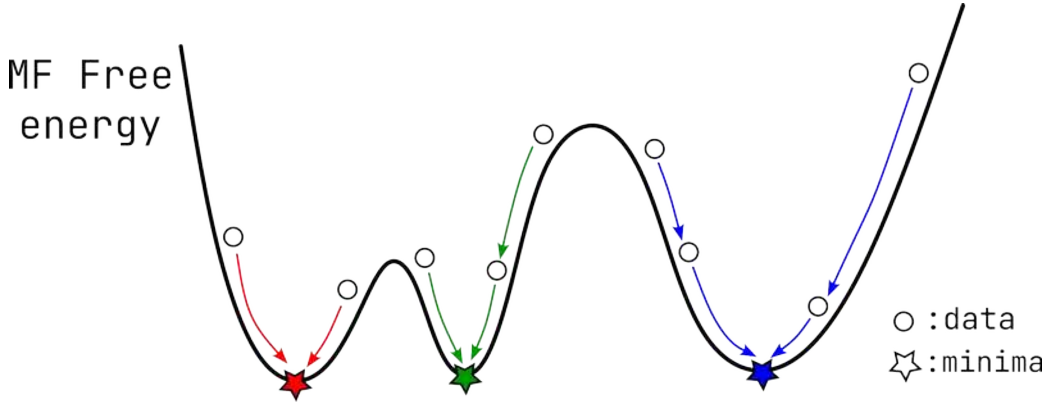
Lorenzo Rosset

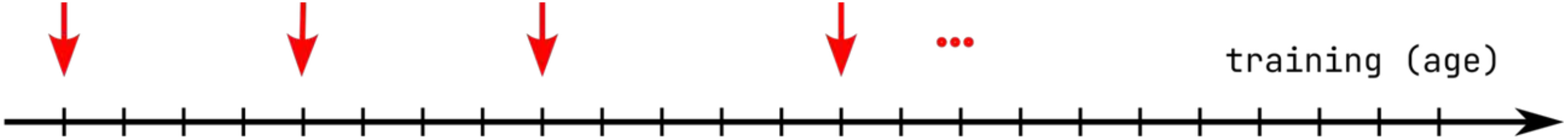
- Approximate the Free energy
  (Plefka’82,Plefka, Georges & Yedidia’91)

  * Gabrié et al. (2015), Tramel et al. (2018), Maillard et al. (2019)
  * Decelle, Rosset, Seoane (2023)

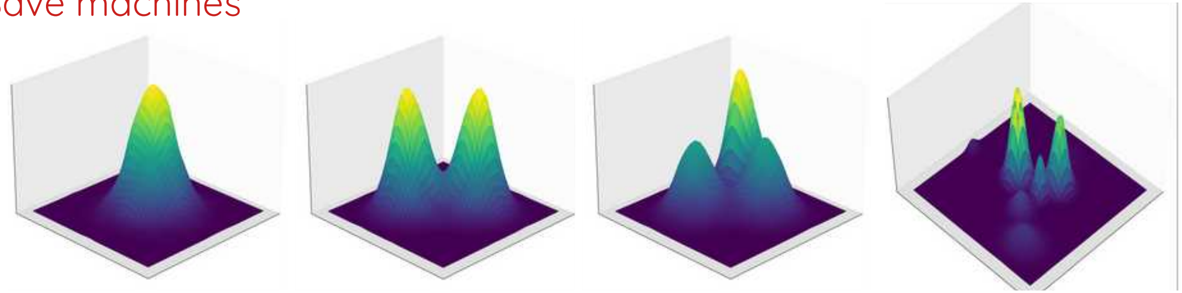
- Identify the nearest minima (TAP)

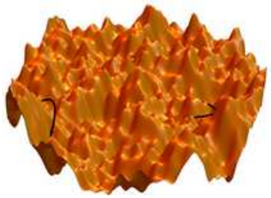
- Use the minima for hierarchical clustering





Save machines





More and more structure
are added to the model

Hierarchical carving

We can approximate the free energy using the Plefka expansion (small coupling)

$$\Gamma(\boldsymbol{m}^{(vis)}, \boldsymbol{m}^{(hid)}) \approx S(\boldsymbol{m}^{(vis)}, \boldsymbol{m}^{(hid)}) - \sum_i a_i m_i - \sum_a b_a m_a$$
$$- \sum_{ia} \left[w_{ia} m_i m_a + \frac{w_{ia}^2}{2} (m_i - m_i^2)(m_a - m_a^2) \right]$$

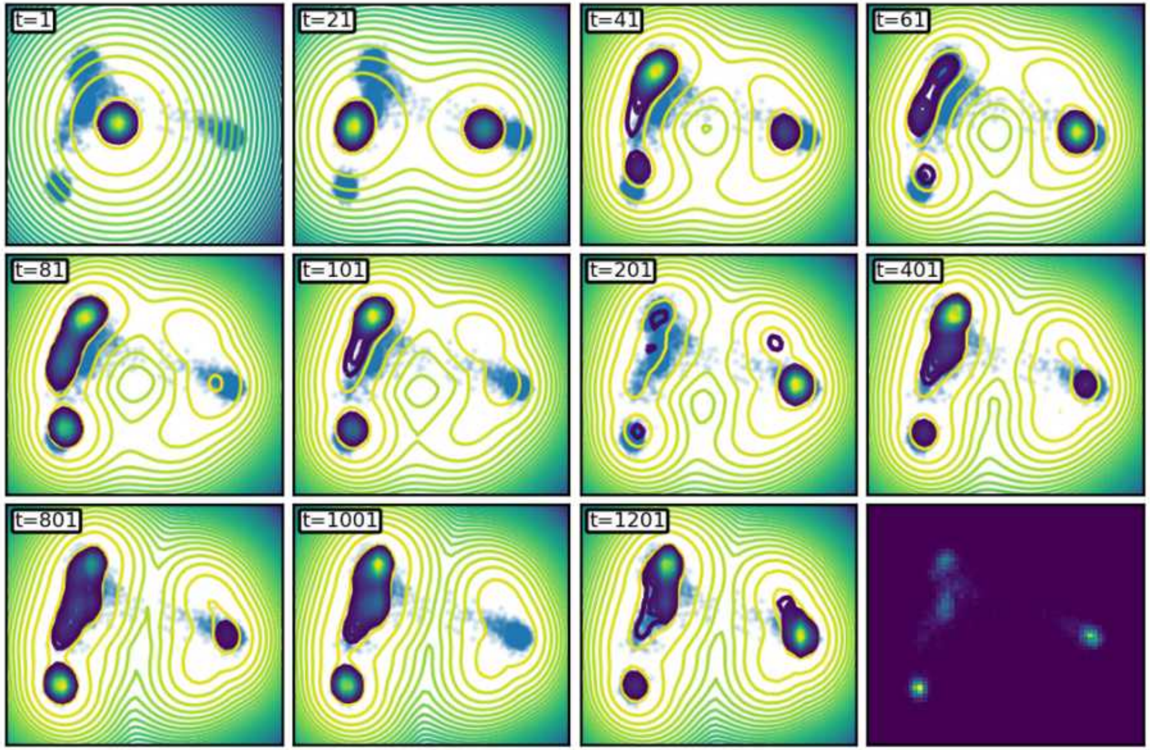
Self-consistent eqs.
$$\begin{cases} m_i = \mathrm{sigm}(a_i + \sum_a w_{ia} m_a - \sum_a w_{ia}^2 \left(m_i - \frac{1}{2} \right) (m_a - m_a^2)) \\ m_a = \mathrm{sigm}(b_a + \sum_i w_{ia} m_i - \sum_i w_{ia}^2 \left(m_a - \frac{1}{2} \right) (m_i - m_i^2)) \end{cases}$$

$$\mathrm{sigm(x)} = (1 + \exp(-\mathrm{x}))^{-1}$$

Gabrié et al, Neurips 2015

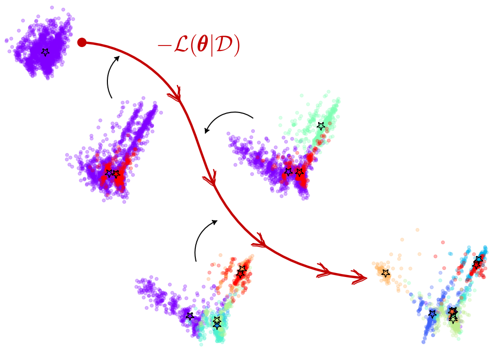
Example on the genetic dataset

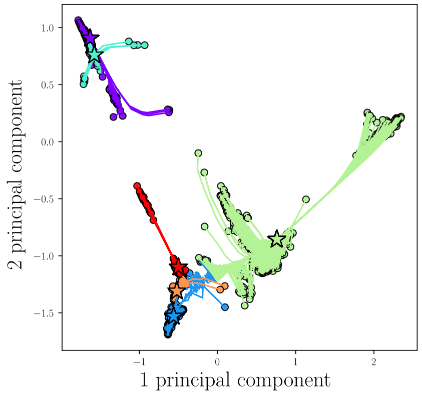
- 5008 sequences of mutated or not (0/1) genes (samples)
- 805 genes (variables)



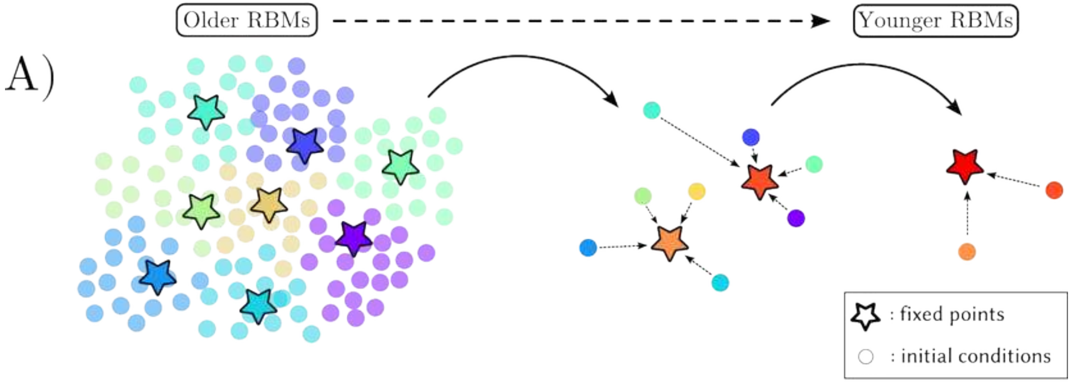
Tree reconstruction of the minima

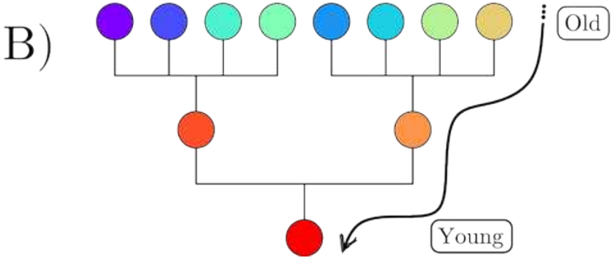
Learning Trajectory





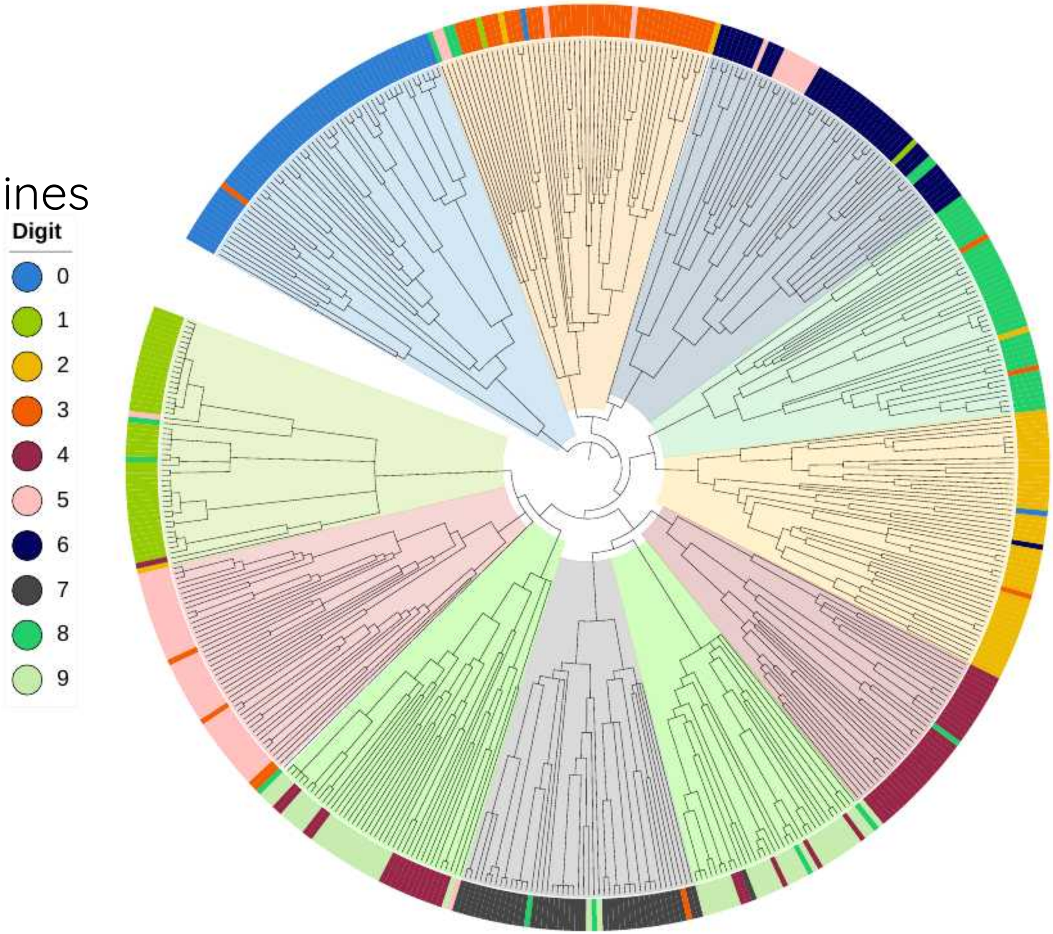
Building a hierarchical tree from it !

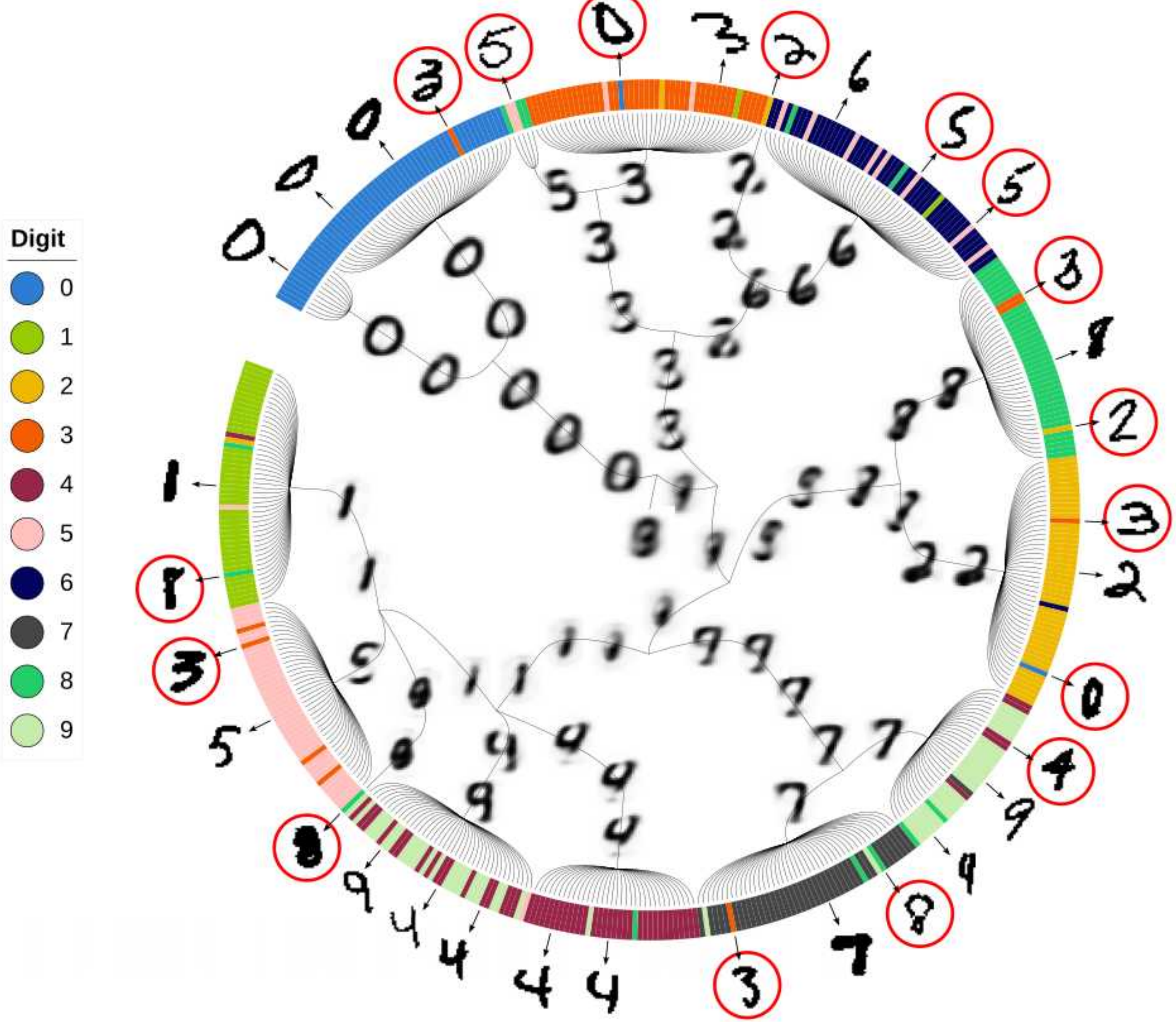




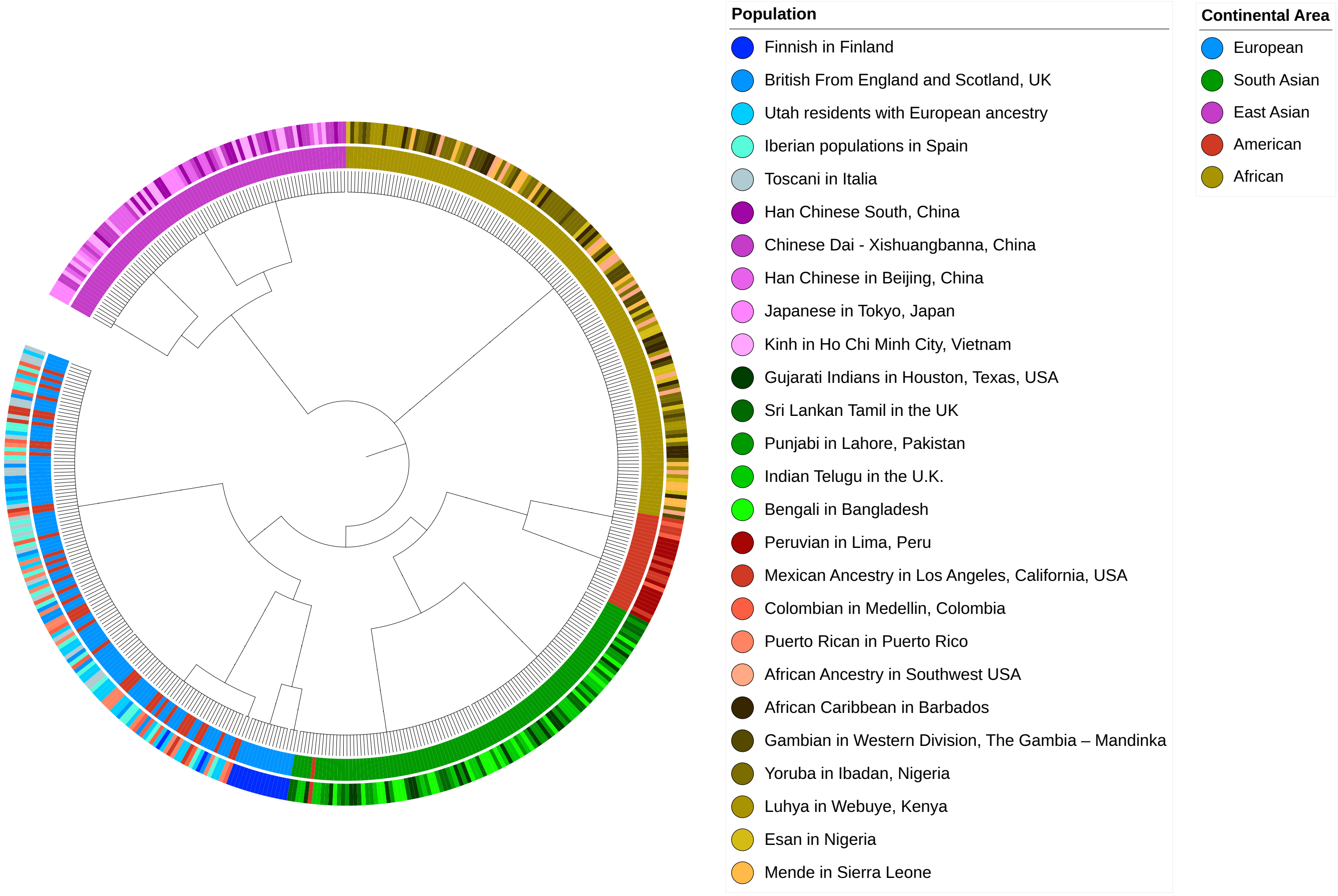
On MNIST

Toward the center: older and older machines
The leafs: dataset




Digit
0
1
2
3
4
5
6
7
8
9

On DNA dataset



Conclusion

- RBMs undergoe phase transition at the beginning of the learning
- We can associate mean-field critical exponents to this transition
- Concrete effect : the relaxation time diverges
  → strong constraints on the Monte Carlo estimation
- Possible application : Hierarchical shattering of the landscape as the learning goes