-Decelle, Fissore, Furtlehner J. of Stat. Physics 2018: themodynamics
-Decelle, Furtlehner Chinese physics B, 2021: RBM & Stat. Phys.
-Decelle, Furtlehner, Seoane ArXiv:2105.13889 (NeurIPS 2021) “Generation”
-Decelle, Furtlehner, Rosset, Seoane PRE 2023, Interpretability

The Restricted Boltzmann Machine:

- Phase diagram, generation and interpretability

Aurélien Decelle

Theoretical Physics, UCM Madrid

Seminar Outline

- Introduction to the Restricted Boltzmann Machine (RBM)
- Training of RBMs
- Statistical physics →
    - Linear regime
    - Phase diagram
    - Mixing time and clustering

What can Stat. Phys do for you

Two approaches are possible:

→ what Machine Learning can do for physics

Higgs challenge

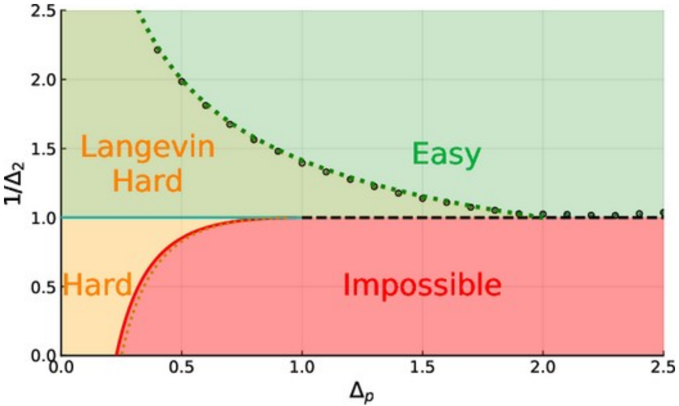
the HiggsML challenge

May to September 2014

When High Energy Physics meets Machine Learning

→ **<u>what (statistical) physics can do for Machine Learning</u>**

Gradient descent behavior



Manelli et al. 2018

Broad vision of Machine Learning

Machine Learning tasks are often categorized in three categories

- Supervised Learning
- Unsupervised Learning
- ~~Reinforcement Learning~~

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

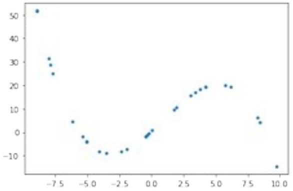
$$\{\boldsymbol{x}_m\}_{m=1,\dots,M} \qquad \{y_m\}_m$$

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 $f_{\theta*}(x_m) = y_m$

Broad vision of Machine Learning

Machine Learning tasks are often categorized in three categories

- Supervised Learning
- **Unsupervised Learning**
- ~~Reinforcement Learning~~

A dataset of M elements in dimension N

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

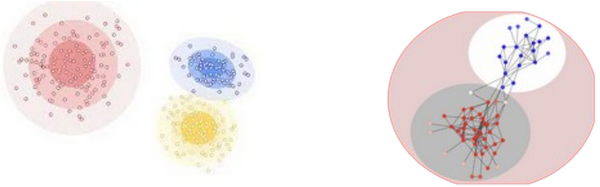
Then, in most settings we want to learn a probability distribution matching the empirical one

Example of generative models

$$x =$$

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

Examples of clustering



Generative model

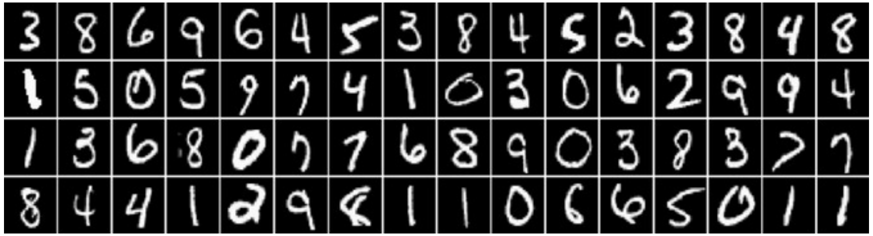
We can generate new “data”, after training a model on a given dataset

This person doesn’t exist

Energy based models

Hinton, Hopfield, LeCun, Bengio

- Dataset $$X = \left\{x^{(1)}, \ldots, x^{(M)}\right\}$$



Empirical

Model

$$p_{\text{data}}(x) \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Boltzmann distribution

$$E_{\boldsymbol{\theta}}(\boldsymbol{x})$$

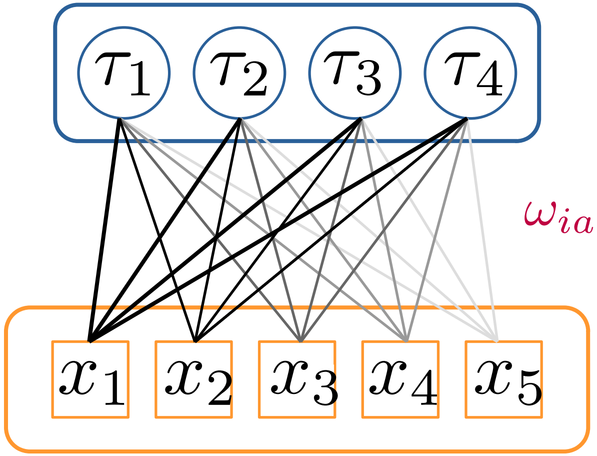
Learning : adjust the parameters so that the dataset configurations are typical configurations of the model.

Define the Energy function using latent variables

-Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory.*

The Restricted Boltzmann Machine (RBM)

$$\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\tau}) = -\sum_{ia} x_i w_{ia} \tau_a - \sum_i \eta_i x_i - \sum_a \theta_a \tau_a$$



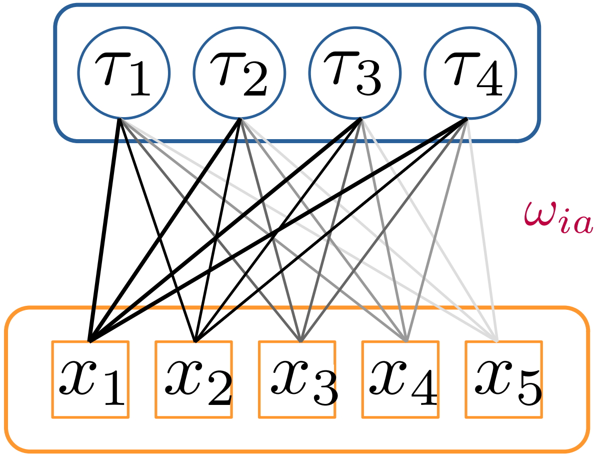
$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{\tau}} e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\tau})}}{Z_{\boldsymbol{\theta}}} = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Define the Energy function using latent variables

-Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory.*

The Restricted Boltzmann Machine (RBM)

$$\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\tau}) = -\sum_{ia} x_i w_{ia} \tau_a - \sum_i \eta_i x_i - \sum_a \theta_a \tau_a$$



$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{\tau}} e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\tau})}}{Z_{\boldsymbol{\theta}}} = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}}{Z_{\boldsymbol{\theta}}}$$

Le Roux and Bengio. Neural computation (2008)

$$E_{\boldsymbol{\theta}}(\boldsymbol{x}) = -\log\left(\sum_{\boldsymbol{\tau}} e^{-\mathcal{E}_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\tau})}\right)$$

Effective model for the RBM can encode higher order correlations!

$$= -E_0 - \sum_i h_i s_i - \sum_{i,j} J^{(2)}_{i,j} s_i s_j - \sum_{i,j,k} J^{(3)}_{ijk} s_i s_j s_k \cdots - \sum_{j_1\cdots j_n} J^{(n)}_{j_1\cdots j_n} s_{j_1}\cdots s_{j_n} - \cdots$$

RBMs are simple, yet powerful

- It is basically an Ising model in its discrete version
- It can model complex dataset (e.g. images or other real dataset)
- It is simple enough to be analyzed theoretical and to be "interpreted"
  → cf Alfonso Navas
- → Ideal playground for physicist:

  - Monasson's group: Tubiana, Roussel, Fernandez de Cossio, … Phase diagram, dynamics
  - Tanaka, Yasuda, Belief Propagation
  - H. Huang, one synapse RBM
  - Barra, Agliara, Tantari et al, phase diagram and equivalent with Hopfield
  - Other contributions see talks of thursday/friday

  …

Training a RBM

Gibbs equilibrium distribution

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

Dataset $S = \{\boldsymbol{s}^{(1)}, \cdots, \boldsymbol{s}^{(M)}\}$

make them the **typical samples** of *p*

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

$$\frac{\partial \mathcal{L}}{\partial w_{ia}} = \boxed{\langle s_i \tau_a \rangle_{\mathcal{D}}} - \boxed{\langle s_i \tau_a \rangle_{\mathcal{H}}}$$

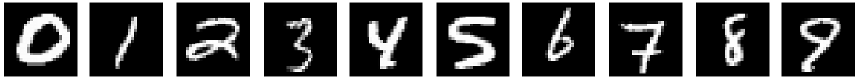
HARD: Monte-Carlo Markov-Chain

EASY!

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

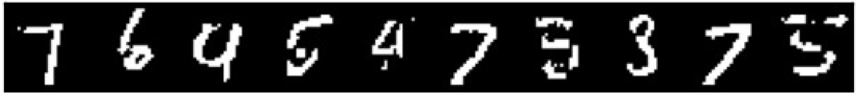
Training process

Given some data:



1. Compute the positive term $\langle s_i \tau_a \rangle_{\mathcal{D}}$
2. Compute the negative term using Mont-Carlo $\langle s_i \tau_a \rangle_{\mathcal{H}}$
3. Update the weights

When the training is done, what can you do ?
→ generate new (fake) data ! using Monte-Carlo.



Gif time (on your own device)

MNIST

LinkMNIST

FacesBW

LinkFaces

CelebA

LinkCelebA

Phase Diagram of the model

- Before focusing on the learning we can try to understand the recall properties of the model
- It allows to understand the effect of the training on the mixing time of the chain
- And how the features are related to the dataset at the beginning of the learning.

Preamble: dynamics of Gaussian model

- We can study the learning dynamics of the very simple case of the Gaussian-Gaussian RBM.
- We first decompose the weight matrix to diagonalize the Gaussian measure

$$w_{ij} = \sum_\alpha u_i^\alpha w_\alpha v_j^\alpha$$

- Then we can project the gradient on each element

Dynamics of the modes

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

Dynamics of the eigenvectors

$$\Omega^{u,v}_{\alpha\beta} = (1-\delta_{\alpha\beta}) \left(\frac{w_\beta - w_\alpha}{w_\alpha + w_\beta} \mp \frac{w_\beta + w_\alpha}{w_\alpha - w_\beta} \right) \boxed{\langle s_\alpha s_\beta \rangle_{\text{Data}}}$$

Correlation matrix projected on the eigenmodes of W

Preambule: dynamics of Gaussian model

- We can study the learning dynamics of the very simple case of the Gaussian-Gaussian RBM.
- We first decompose the weight matrix to diagonalize the Gaussian measure
- Th

In the linear regime:

→ the eigenvectors of the weigth matrix aligned with those of the PCA of the dataset

→ the eigenmodes are expressed if the signal is higher than the intrinsic noise

Dynamics

$$\frac{\quad}{dt} = w_\alpha \sigma_h^2 \left(\langle s_\alpha^2 \rangle_{\text{Data}} - \frac{\quad}{1 - \sigma_v^2 \sigma_h^2 w_\alpha^2} \right)$$

Dynamics of the eigenvectors

$$\Omega_{\alpha\beta}^{u,v} = (1 - \delta_{\alpha\beta}) \sigma_h^2 \left(\frac{w_\beta - w_\alpha}{w_\alpha + w_\beta} \mp \frac{w_\beta + w_\alpha}{w_\alpha - w_\beta} \right) \boxed{\langle s_\alpha s_\beta \rangle_{\text{Data}}}$$

Correlation matrix projected on the eigenmodes of W

Rank-K signals

Instead of taking the “usual” path of studying the RBM in the independent weight approximation (quite unlikely for the learning pb).

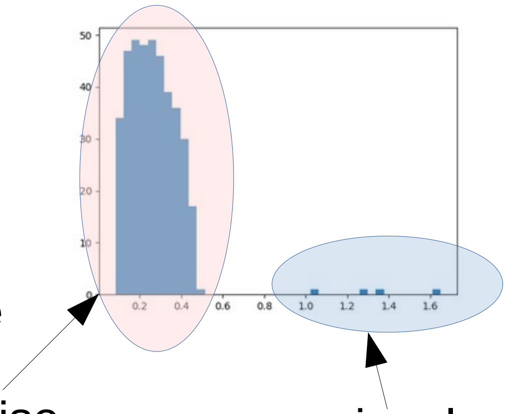
→ We consider a rank-K decomposition of the matrix

$$w_{ia} \approx \sum_{\alpha=1}^{K} u_i^\alpha w_\alpha v_a^\alpha + r_{ia}$$

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

Rank K decomposition plus random noise

Spectrum



noise

signals

Order parameters

The magnetization along a mode α

$$m_\alpha \sim E_{u,v,r}\Big(\langle s_\alpha\rangle\Big) \sim \sum_i u_i^\alpha \langle s_i\rangle$$

$$\bar{m}_\alpha \sim E_{u,v,r}\Big(\langle \tau_\alpha\rangle\Big) \sim \sum_a v_a^\alpha \langle \tau_a\rangle$$

The overlap of the system

$$q_{\mu\nu} \sim E_{u,v,r}\Big(\langle s_i^\mu s_i^\nu\rangle\Big)$$

$$\bar{q}_{\mu\nu} \sim E_{u,v,r}\Big(\langle \tau_a^\mu \tau_a^\nu\rangle\Big),$$

Self-consistent equations

$$m_\alpha = (w_\alpha \bar{m}_\alpha - \theta_\alpha)(1-q)$$

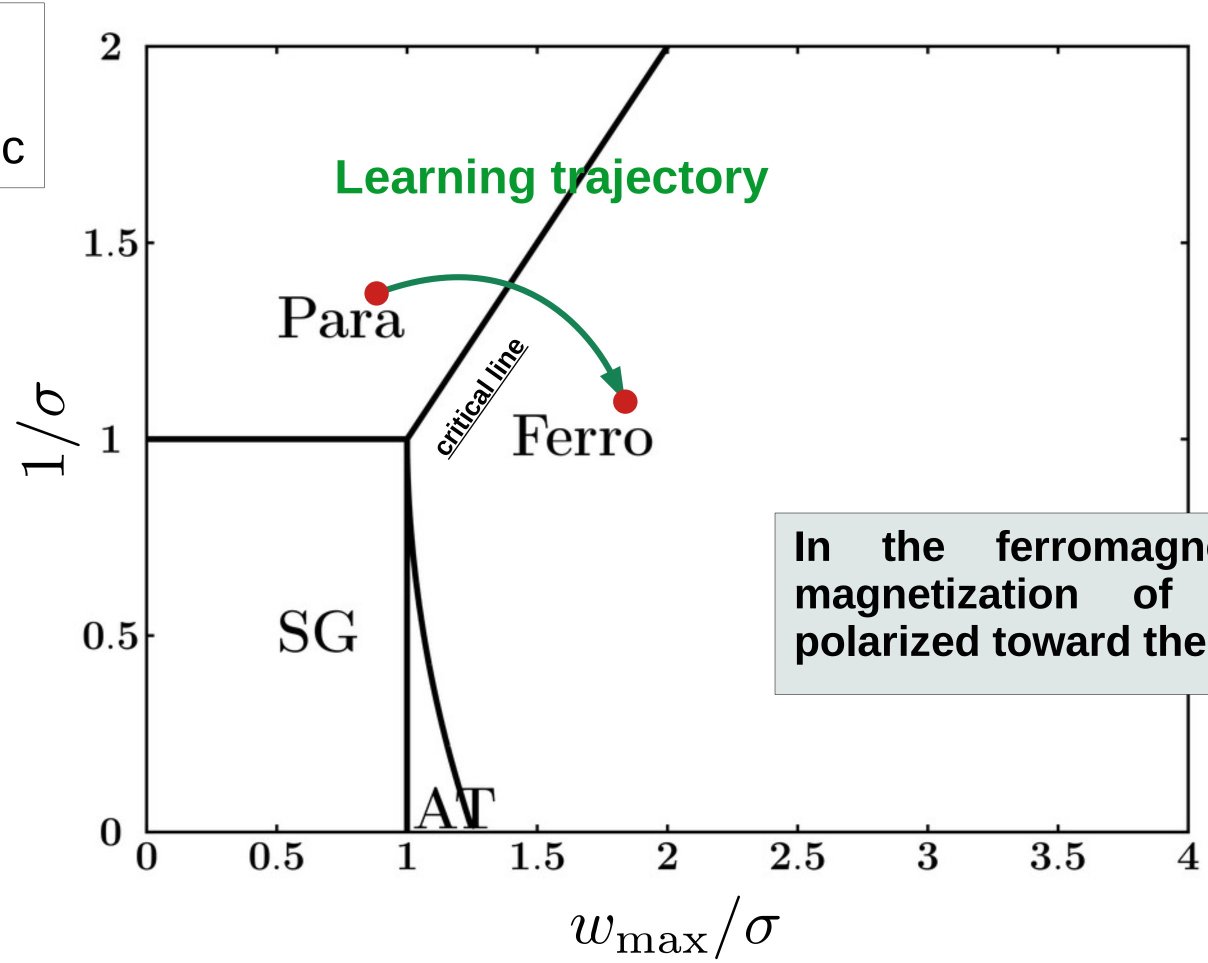
$$\bar{m}_\alpha = (w_\alpha m_\alpha - \eta_\alpha)(1-\bar{q}),$$

$$q = \int dy \frac{e^{-y^2/2}}{\sqrt{2\pi}} \tanh^2(\sqrt{\bar{v}}\kappa^{-1/4} y),$$

$$\bar{q} = \int dy \frac{e^{-y^2/2}}{\sqrt{2\pi}} \tanh^2(\sqrt{v}\kappa^{1/4} y),$$

Phase Diagram

SG= Spin Glass
Para= Paramagnetic
Ferro= Ferromagnetic



In the ferromagnetic phase, the magnetization of the system is polarized toward the eigenmodes of w!

Application of the theory

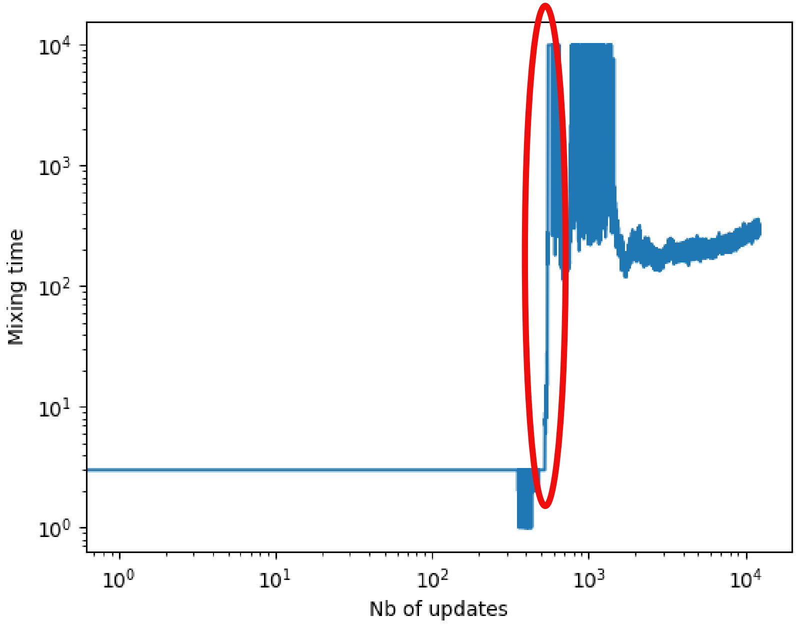
The mean-field theory is in general not correct for long training time, still it is very useful:

1- to understand problems that occur in the training and
2- to design new tools to investigate the trained machine

I - Mixing time and training problem

It is in generally accepted that training RBM can be hard. The main problem is related to the **Monte Carlo estimates** when computing the gradient

Mixing time as a function of the training time (MNIST)

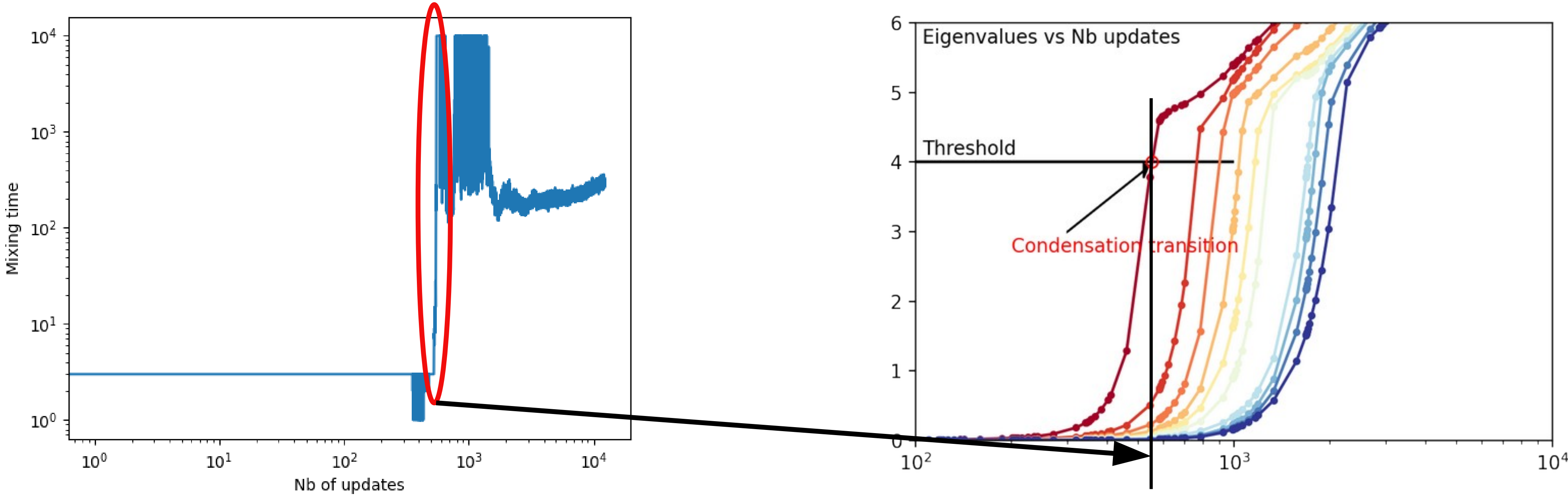


Huge jump !

I - Mixing time and training problem

It is in generally accepted that training RBM can be hard. The main problem is related to the **Monte Carlo estimate** when computing the gradient

Mixing time as a function of the training time (MNIST)



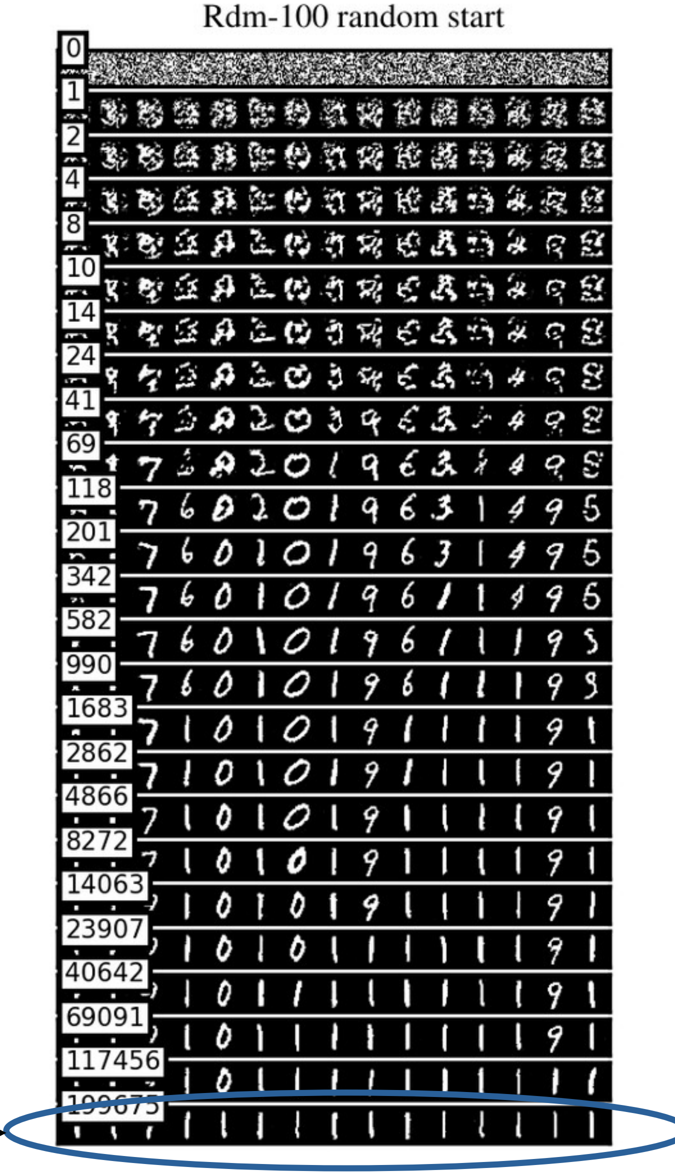
It corresponds to the 2nd order phase transition of the Phase Diagram

Consequence on the training

It is usual to use a very small number of Monte Carlo steps to train RBMs

→ the machine generally end up in a regime where the MC estimates do not coincide with the true thermodynamics one and thus the generated data can be quite bad.

Example on the right on a trained machine with 100 MC steps at each update. After many MC steps we find all ones !



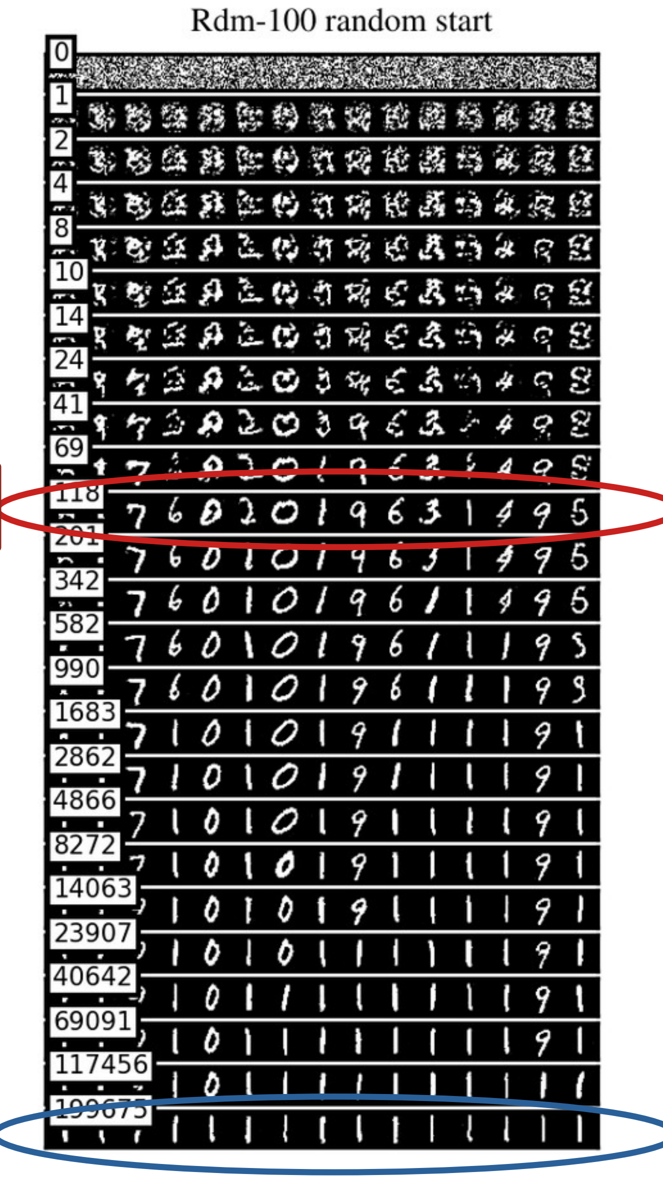
Very biased samples

Consequence on the training

It is usual to use a very small number of Monte Carlo steps to train RBMs

→ the machine generally end up in a regime where the MC estimate do not coincide with the true thermodynamics one and thus the generated data can be quite bad.

But there exists a sweet spot that correspond to reproducing exactly the same dynamics.



Memory

***k* = 100**

Very biased samples

II – Following the learning trajectory

- We have seen that the training undergoes 2nd order phase transition. The theory says that it can undergo severals.
- At each phase transition, the distribution splits into several modes.

→ by following the learning trajectory, we can follow the creation of modes!

How to follow the modes ? **Using the mean-field theory - Plefka expansion**

Plefka, J Phys A, 1982
Gabrié et al, NeurIps 2015

II – Following the learning trajectory

How to follow the modes ? **Using the mean-field theory - Plefka expansion**

Example for the mean-field Ising model: the magnetization respect the self-consistent eq.

$$m_i = \tanh(\textstyle\sum_j J_{ij} m_j + a_i)$$

The solution correspond to minima of the free energy (modes of the distribution)

Plefka, J Phys A, 1982
Gabrié et al, NeurIps 2015

II – Following the learning trajectory

Case of Ising

$m_i = \tanh(\sum_j J_{ij} m_j + a_i)$

For RBM: Mean field self-consistent equations

$$m_i^v = \mathrm{sig}\left(\sum_a w_{ia} m_a^h + \theta_i\right)$$

$$m_a^h = \mathrm{sig}\left(\sum_i w_{ia} m_i^v + \eta_a\right)$$

In the MF regime, it corresponds to local maximum of the probability distribution.

II – Following the learning trajectory

Mean field iterations

$$m_j^h[t+1] \leftarrow \mathrm{sigm}\left[b_j + \sum_i W_{ij} m_i^v[t] \text{[redacted]} \right]$$

$$m_i^v[t+1] \leftarrow \mathrm{sigm}\left[a_i + \sum_j W_{ij} m_j^h[t+1] \text{[redacted]} \right]$$

Gabrié et al 2015
Decelle et al 2023

II – Following the learning trajectory

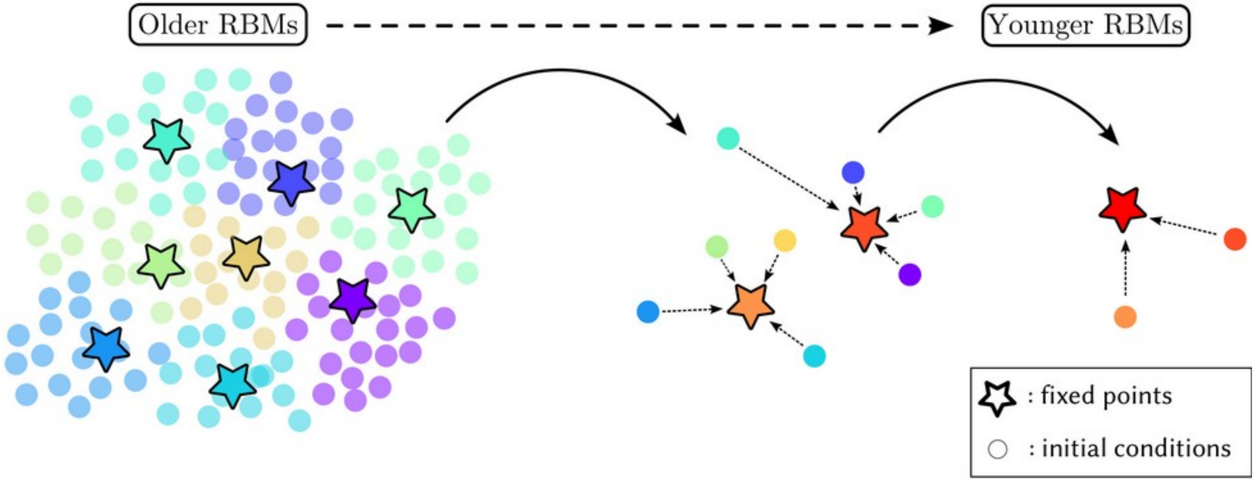
Mean field iterations (at second order)

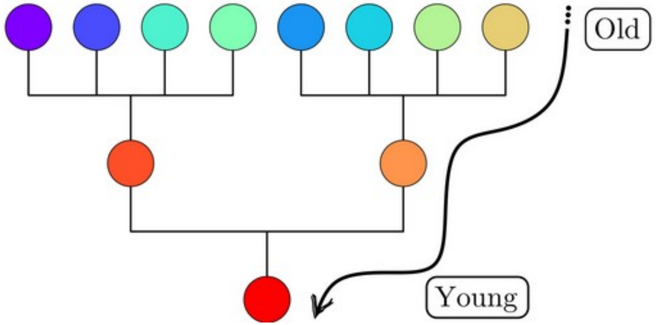
$$m_j^h[t+1] \leftarrow \mathrm{sigm}\left[b_j + \sum_i W_{ij} m_i^v[t] - W_{ij}^2 \left(m_j^h[t] - \frac{1}{2} \right) \left(m_i^v[t] - (m_i^v[t])^2 \right) \right]$$

$$m_i^v[t+1] \leftarrow \mathrm{sigm}\left[a_i + \sum_j W_{ij} m_j^h[t+1] - W_{ij}^2 \left(m_i^v[t] - \frac{1}{2} \right) \left(m_j^h[t+1] - (m_j^h[t+1])^2 \right) \right]$$

Gabrié et al 2015
Decelle et al 2023

II – Following the learning trajectory



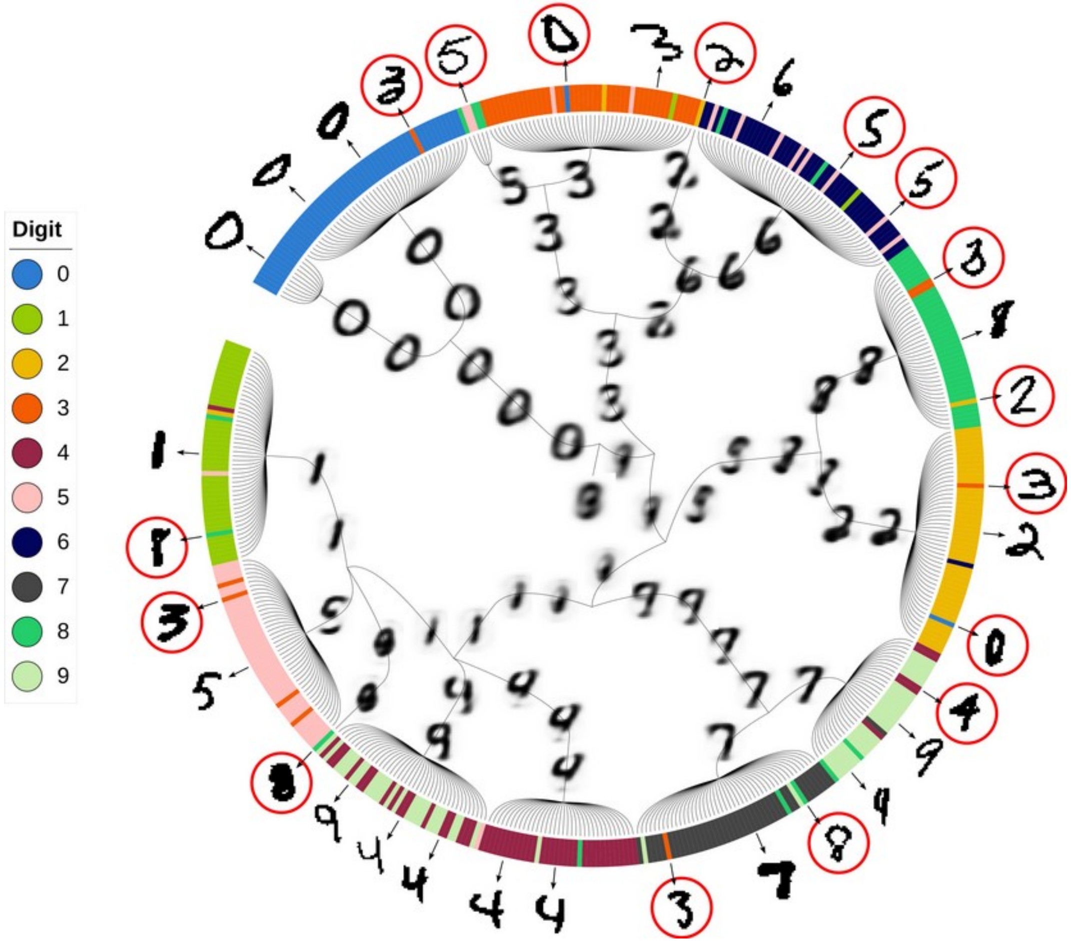


Starting from older trained models:

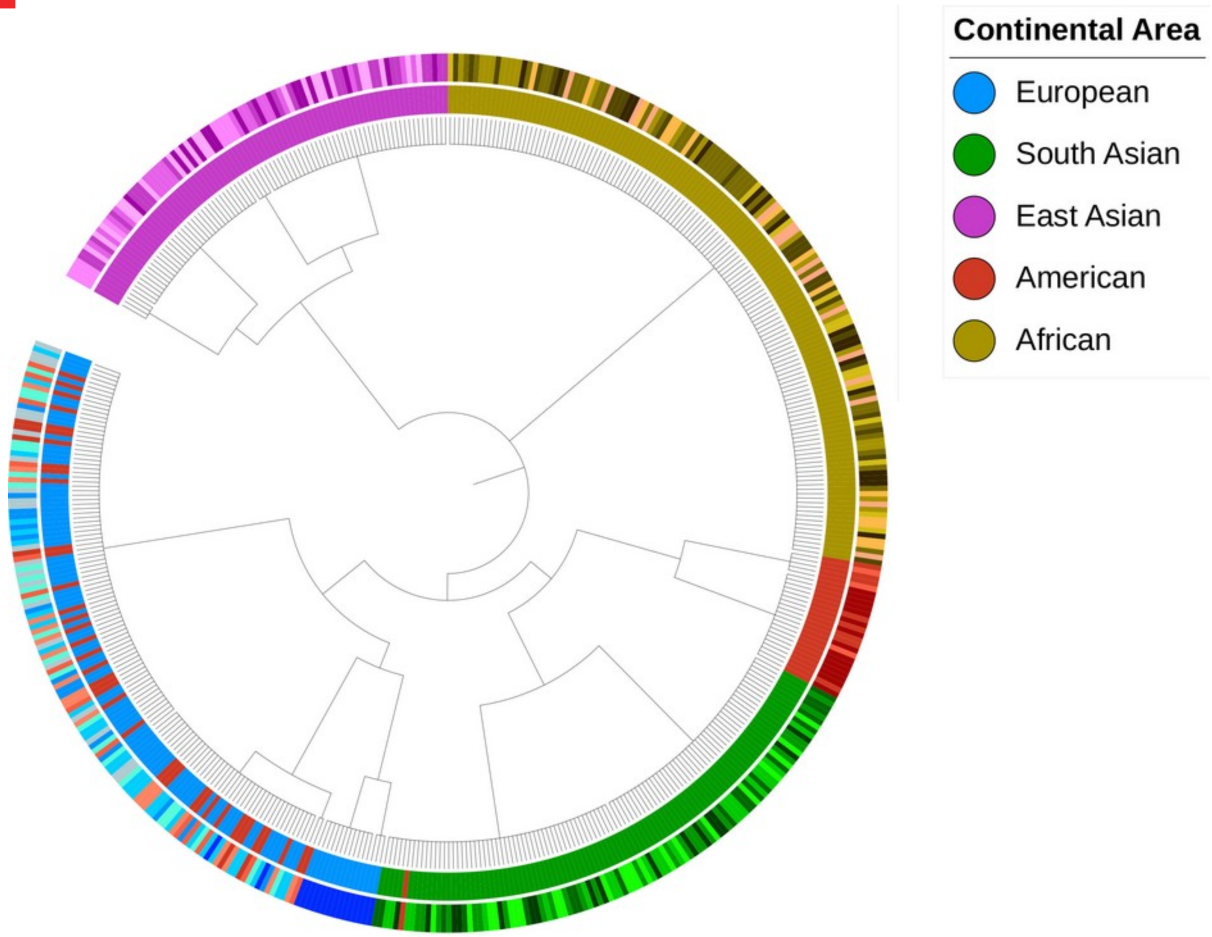
1- we find the mean-field fixed points associated to the datapoint

2- we follow the evolution of these fps when going to “younger” models.

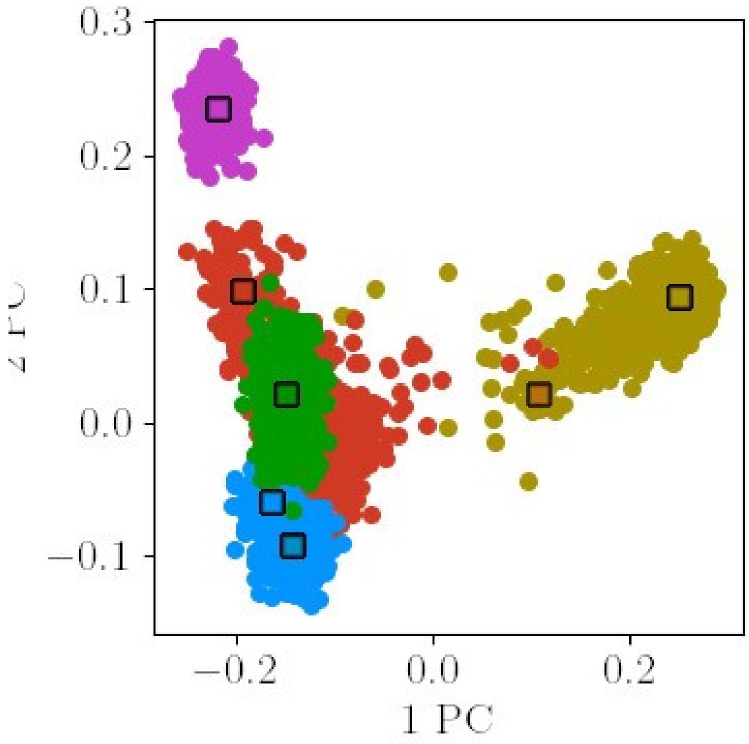
II – Following the learning trajectory



II – Following the learning trajectory



Population genetics dataset



Summary

- RBM difficulties lie mainly in the misunderstood of Monte Carlo Markov Chain
- It can model real dataset with accuracy
- A perfect playground for physicists:
  - Rich phase diagram
  - Complex learning dynamics
  - Yet it is simple enough for analytical computations

Main challenge:
→ understanding the learning behavior
→ understanding the relation between the learned features and the dataset

Acknowledgments

Acknowledgments:

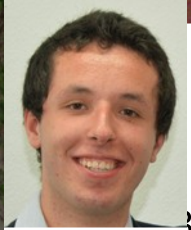
Aurélien Decelle
Giovanni Catania
Nicolas Béreux
Alfonso Navas
Lorenzo Rosset

UCM

Cyril Furtlehner (Paris-Saclay)

Javier Moreno Gordo (Uex)

Elisabeth Agoritsas (Geneve)



talento cm Programa de atracción de talento investigador Comunidad de Madrid

Santander

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

AGENCIA ESTATAL DE INVESTIGACIÓN

ANR AGENCE NATIONALE DE LA RECHERCHE