Thermodynamics of Restricted Boltzmann Machines and Related Learning Dynamics

Aurélien Decelle, Giancarlo Fissore, Cyril Furtlehner

INRIA, LRI, Université Paris-Saclay
TAU team

Restricted Boltzmann Machines (RBM)

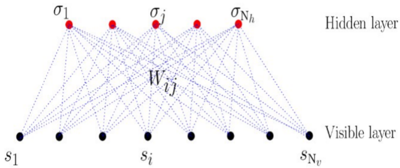
Task: modeling high-dimensional probability distributions of empirical data

Solution: we can use a Restricted Boltzmann Machine (RBM), a neural-network based model

Problem: neural networks are "black boxes"

The RBM as a bipartite spin-glass

$$H(s,\sigma) = -\sum_{i,j} s_i W_{ij} \sigma_j + \sum_{i=1}^{N_v} \eta_i s_i + \sum_{j=1}^{N_h} \theta_j \sigma_j \qquad p(s,\sigma) = \frac{e^{-H(s,\sigma)}}{Z}$$



Learn W_{ij} (Maximum Likelihood)

↓

Sample the visible layer s

↓

Linearized mean-field equations

Mean-field equation for the visible layer of the RBM:

$$m_i^v = sigm\left(\eta_i + \sum_j w_{ij} m_j^h - \sum_j w_{ij}\right) \qquad \left(\mathbf{m}^v = \langle \mathbf{s} \rangle, \mathbf{m}^h = \langle \sigma \rangle\right)$$

Expanding over Singular Value Decomposition (SVD) components:

$$w_{ij} = \sum_\alpha w_\alpha u_{i,\alpha} v_{j,\alpha} \qquad m_\alpha^v = \sum_i u_{i,\alpha} m_i^v$$

$$\Downarrow$$

$$m_\alpha^v \simeq \frac{1}{4} w_\alpha m_\alpha^h$$

Magnetizations related to strong w_α are amplified

Dynamics & statistical ensemble

- Dynamical evolution

$$\frac{dw_\alpha}{dt} = \langle s_\alpha \sigma_\alpha \rangle_{data} - \langle s_\alpha \sigma_\alpha \rangle_{model}, \quad s_\alpha = \sum_i s_i u_{i,\alpha}$$

- We need to define a statistical ensemble

$$w_{ij} = \sum_{\alpha=1}^{K} w_\alpha u_{i,\alpha} v_{j,\alpha} + r_{ij}$$

w_α: singular values

$u_{i,\alpha} v_{j,\alpha}$: singular vectors components

r_{ij}: gaussian noise

Note: we average with respect to u_i, v_j and the noise r_{ij} keeping s_α, σ_α fixed.

Non-linear mean-field

- Thouless-Anderson-Palmer (TAP) free energy - **"numerical"**

$$
\begin{aligned}
F_{TAP}(\mathbf{m}^v, \mathbf{m}^h) = &+ S(\mathbf{m}^v) + S(\mathbf{m}^h) \\
&- \sum_i \eta_i m_i^v - \sum_j \theta_j m_j^h - \sum_{i,j} w_{ij} m_i^v m_j^h \\
&+ \sum_{i,j} \frac{w_{ij}^2}{2} \left(1 - {m_i^v}^2\right) \left(1 - {m_j^h}^2\right)
\end{aligned}
$$

- Replica symmetry framework - **"theoretical"**

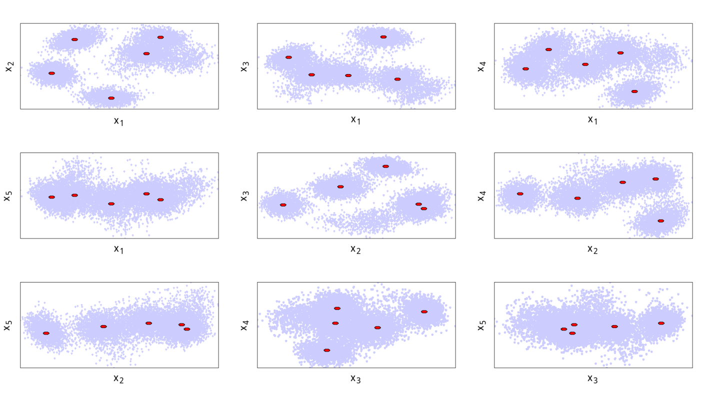
$$
\begin{aligned}
m_\alpha^v &= \left(w_\alpha m_\alpha^h - \eta_\alpha\right)\left(1 - q_\alpha^v\right) \\
m_\alpha^h &= \left(w_\alpha m_\alpha^v - \theta_\alpha\right)\left(1 - q_\alpha^h\right)
\end{aligned}
$$

$$m_\alpha^v = E_{u,v,r}\left(\langle s_\alpha \rangle\right) \qquad m_\alpha^h = E_{u,v,r}\left(\langle \sigma_\alpha \rangle\right)$$

q_α^v, q_α^h: spin-glass order parameters

Clustering interpretation

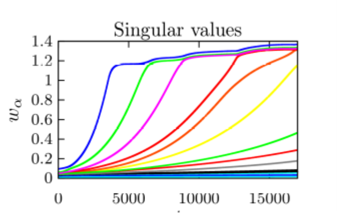
Data get clustered in the singular space, and the fixed point solutions of the mean-field equations serve as centroids

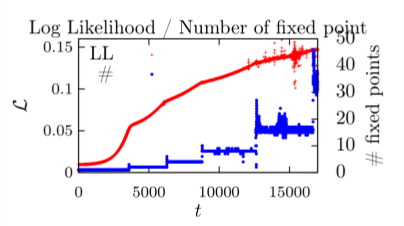


(a) Samples from the training set and fixed points (in red) are plotted with respect to the strongest directions in the singular space

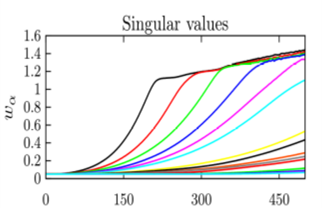
Non-linear dynamics

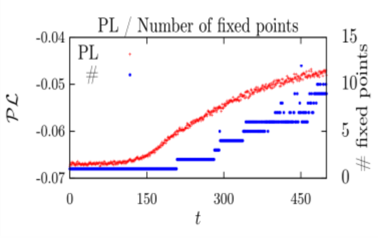
Theoretical





Experimental





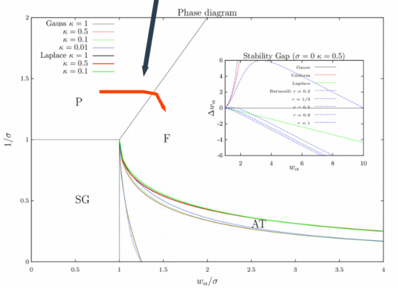
Phase diagram

$$\text{Statistical ensemble:} \quad W_{ij} = \sum_{\alpha=1}^{K} w_\alpha u_i^\alpha v_j^\alpha + r_{ij}$$

Decomposition over K "eigenmodes"

Gaussian noise

Learning trajectory



Control parameters:

$\frac{1}{\sigma} \rightarrow$ noise ("temperature")

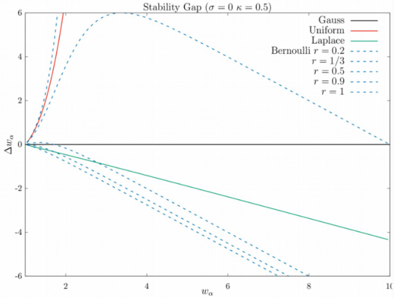
$\frac{w_\alpha}{\sigma} \rightarrow$ "ferromagnetic coupling" along "eigenvector" α

Ferromagnetic phase

$$W_{ij} = \sum_{\alpha=1}^{K} w_\alpha u_i^\alpha v_j^\alpha + r_{ij}$$

$$\text{Relative kurtosis: } \mathrm{k} = \mathrm{E}\left[\left(\frac{\mathrm{X}-\mu}{\sigma}\right)^4\right] - 3 = \frac{\mu_4}{\sigma^4} - 3$$

Kurtosis	k < 0	k = 0	**k > 0**
Gap	Δw > 0	Δw = 0	**Δw < 0**
Distribution	Bernoulli	Gaussian	**Laplace**
Structure	Metastable states	Unimodal Dominant state	**Compositional phase** (possibly)



Conclusion

Outcomes:

- comprehensive theoretical description of the model, both in linear and non-linear regimes
- precise characterization of the learning dynamics (and definition of a deterministic learning trajectory)
- assessment of the role and importance of the fixed point solutions of the mean-field equations
- clustering interpretation of the training process
- characterization of the statistical properties of the weights of the model

Perspectives:

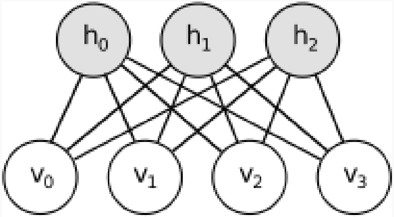
- introducing symmetries: translational (and rotational) invariance
- dealing with lossy datasets

Thank you!

Overview of the RBM model

Definition of the model

RBM model: a neural network structured as a a bipartite graph



Specifically:

- a layer of hidden units h_j and a layer of visible units v_i are present
- data are represented as configurations of the visible layer
- there are not connections among units in the same layer
- we restrict our treatment to the case of binary units $h_i, v_i = 0, 1$

RBM training

- The probability of a visible configuration is given by

$$P(\mathbf{v}) = \sum_{\mathbf{h}} P(\mathbf{h}, \mathbf{v}) = \frac{e^{-F_c(\mathbf{v})}}{Z}, \qquad Z = \sum_{\mathbf{v}} e^{-F_c(\mathbf{v})}$$

- We want to maximize $P(\mathbf{v})$ for the samples belonging to the training set

$\Longrightarrow$ gradient ascent over the log-likelihood $\log P(\mathbf{v})$

Update rule

$$\Delta \mathbf{W} = \alpha \left(\langle \mathbf{v}\mathbf{h}^T \rangle_{data} - \langle \mathbf{v}\mathbf{h}^T \rangle_{model} \right)$$

Problem: the term $\langle \cdot \rangle_{model}$ is intractable

Best approximate algorithm: *persistence contrastive divergence* (PCD), a Monte Carlo based method

The RBM and Statistical Physics

The **RBM model** is mapped to a **Statistical Physics** model by the definition of an *energy function*

$$E(\mathbf{h}, \mathbf{v}) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{i,j} v_i w_{ij} h_j$$

$$P(\mathbf{h}, \mathbf{v}; \mathbf{W}) = \frac{e^{-E(\mathbf{h},\mathbf{v})}}{Z}$$

This let us borrow **analytical** and **algorithmic tools** from statistical physics! In particular mean-field methods.

Remark: w_{ij} are the links connecting visible and hidden units and serve as parameters of the model

Extended Mean Field (EMF) training

- The log-likelihood can be expressed as

$$\log P(\mathbf{v}) = \log \frac{e^{-F_c(\mathbf{v})}}{Z} = -\overbrace{F_c(\mathbf{v})}^{\textbf{tractable}} - \underbrace{\log Z}_{\textbf{intractable}}$$

- $F = \log Z$ is the *free energy* of the system and it can be approximated exploiting a high-temperature expansion[1]

New update rule

$$\Delta \mathbf{W} = \alpha \left(\langle \mathbf{v}\mathbf{h}^T \rangle_{data} - \underbrace{\frac{\partial F_{TAP}(\tilde{\mathbf{m}}^v, \tilde{\mathbf{m}}^h)}{\partial w_{ij}}}_{\textbf{tractable}} \right)$$

[1] **A. Georges**, **J. S. Yedidia**,
"How to expand around mean-field theory using high-temperature expansions",
Journal of Physics A: Mathematical and General, **Volume 24**, **Number 9**, **1991**.

EMF training

Introducing the inverse temperature β

$$P(\mathbf{h}, \mathbf{v}) = \frac{e^{-\beta E(\mathbf{h},\mathbf{v})}}{Z}$$

High-T expansion

Setting $\beta \to 0$ a **tractable** *effective free energy* depending on the magnetizations is obtained: $F_{TAP} = F_{TAP}(\mathbf{m}^v, \mathbf{m}^h)$

Its minimization gives an approximation to F:

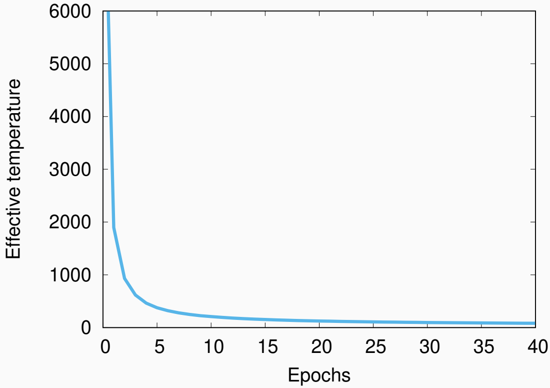
$$F \simeq F_{TAP}(\tilde{\mathbf{m}}^v, \tilde{\mathbf{m}}^h), \qquad \left. \frac{dF_{TAP}}{d\mathbf{m}} \right|_{\tilde{\mathbf{m}}^v, \tilde{\mathbf{m}}^h} = 0 \tag{1}$$

- magnetizations: $\mathbf{m}^v = \langle \mathbf{v} \rangle, \mathbf{m}^h = \langle \mathbf{h} \rangle$
- $\tilde{\mathbf{m}}^v, \tilde{\mathbf{m}}^h$ are found by iterating to a fixed point the equations given by constraint (1)

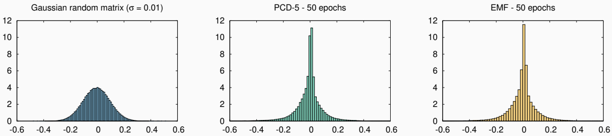
Effective temperature

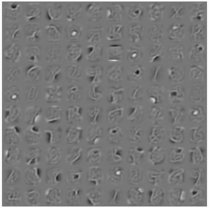
In the context of a RBM the high-T expansion is substituted by a **weak-couplings expansion** (w_{ij} small) and an *effective temperature* is defined:

$$T_{eff} = \frac{1}{Var(\mathbf{W})}$$

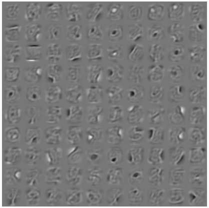


Comparison of PCD and EMF trainings

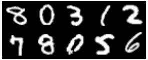




(a) PCD features



(b) EMF features



(c) MNIST

(d) PCD

(e) EMF

Learning dynamics are independent on the training procedure

Singular Value Decomposition (SVD)

SVD is the generalization of eigenmodes decomposition to rectangular matrices

$$\mathbf{W} = \mathbf{U}\Sigma\mathbf{V}^T$$

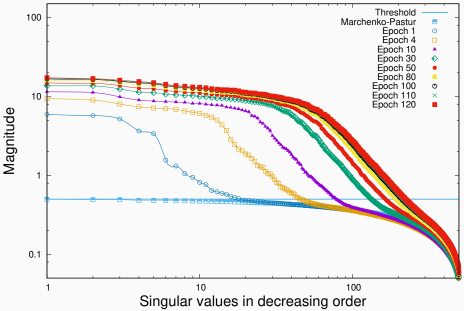
where:

- $\mathbf{U}$ is an orthogonal matrix whose columns are the *left singular vectors* $\mathbf{u}_\alpha$
- $\mathbf{V}$ is an orthogonal matrix whose columns are the *right singular vectors* $\mathbf{v}_\alpha$
- Σ is a diagonal matrix whose elements are the singular values σ_α

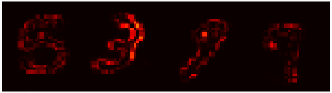
Remark

Singular vectors $\mathbf{u}_\alpha$ can be visualized in pixel space

Characterization of the modes



(a) "filtered" samples: only the first 100 modes are retained



(b) boundary adjustments: the remaining 400 modes

A basic statistical characterization

$$w_{ij} = \underbrace{\sum_{\alpha \in bulk} \sigma_\alpha u_{i,\alpha} v_{j,\alpha}}_{random \to r_{ij}} + \sum_{\alpha \in outliers} \sigma_\alpha u_{i,\alpha} v_{j,\alpha}$$

Updates dynamics

Introducing a time variable t we write

$$w_{ij}(t) = \sum_{\alpha} \sigma_{\alpha}(t) \mu_{i,\alpha}(t) \nu_{j,\alpha}(t) \tag{2}$$

and taking the continuous limit of the learning equations we obtain

$$\frac{dw_{ij}}{dt} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \tag{3}$$

$$\frac{da_i}{dt} = \langle v_i \rangle_{data} - \langle v_i \rangle_{model} \tag{4}$$

$$\frac{db_j}{dt} = \langle h_j \rangle_{data} - \langle h_j \rangle_{model} \tag{5}$$

Linearized dynamics

Introducing time t

$$w_{ij}(t) = \sum_{\alpha} \sigma_\alpha(t) u_{i,\alpha}(t) v_{j,\alpha}(t)$$

Assuming Gaussian distributions for visible and hidden nodes (with σ_v, σ_h):

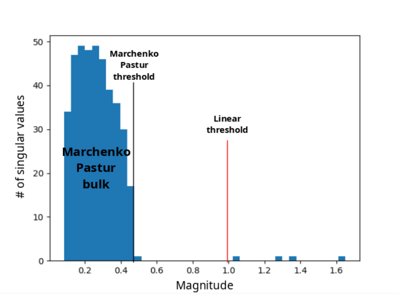
$$\frac{d\sigma_\alpha}{dt} = \sigma_h^2 \sigma_\alpha \left(\langle v_\alpha^2 \rangle_{data} - \frac{\sigma_v^2}{1 - \sigma_v^2 \sigma_h^2 \sigma_\alpha^2} \right)$$

By linear stability analysis we can find the stable fixed points

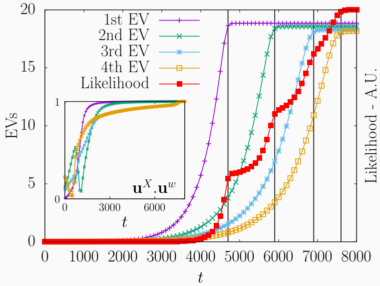
$$\sigma_\alpha^2 = \begin{cases} \frac{\langle v_\alpha^2 \rangle_{data} - \sigma_v^2}{\sigma_v^2 \sigma_h^2 \langle v_\alpha^2 \rangle_{data}} & \langle v_\alpha^2 \rangle_{data} > \sigma_v^2 \\ 0 & \langle v_\alpha^2 \rangle_{data} < \sigma_v^2 \end{cases}$$

Linear dynamics

Time evolution of the singular values ("eigenvalues") in the linear model:



(a) Empirical distribution of the singular values (at the beginning, weights are random and the distribution is Marchenko-Pastur)



(b) Time evolution of the strongest singular values. The strengthening of a singular value determines an increase in the likelihood of the training data

Expansion over SVD basis

$$\left(\frac{d\mathbf{W}}{dt}\right)_{\alpha\beta} = \sum_{ij} \mu_{i,\alpha} \frac{dw_{ij}}{dt} \nu_{j,\beta}$$
$$= \delta_{\alpha,\beta} \frac{d\sigma_\alpha}{dt} + (1 - \delta_{\alpha\beta}) \left(\sigma_\alpha \Omega^h_{\alpha\beta} + \sigma_\beta \Omega^v_{\beta\alpha}\right) \tag{6}$$

where we have defined the generators of rotations in both μ_α and ν_α bases

$$\Omega^v_{\alpha\beta}(t) = \frac{d\mu^T_\alpha}{dt} \mu_\beta \tag{7}$$
$$\Omega^h_{\alpha\beta}(t) = \frac{d\nu^T_\alpha}{dt} \nu_\beta \tag{8}$$

Off-diagonal variations are related to the basis rotations, while the diagonal dynamics correspond to eigenvalues changes.

Update equations in SVD basis

Projecting the full learning equations on the SVD basis we obtain

$$\left(\frac{d\mathbf{W}}{dt}\right)_{\alpha\beta} = \langle v_\alpha h_\beta \rangle_{data} - \langle v_\alpha h_\beta \rangle_{model} \tag{9}$$

$$\left(\frac{d\mathbf{a}}{dt}\right)_{\alpha} = \langle v_\alpha \rangle_{data} - \langle v_\alpha \rangle_{model} \tag{10}$$

$$\left(\frac{d\mathbf{b}}{dt}\right)_{\alpha} = \langle h_\alpha \rangle_{data} - \langle h_\alpha \rangle_{model} \tag{11}$$

with

$$v_\alpha = \sum_i v_i \mu_{i,\alpha}\,, \qquad h_\alpha = \sum_j h_j \nu_{j,\alpha} \tag{12}$$

Naive mean-field free energy

$$
\begin{aligned}
F(\mathbf{m}^v, \mathbf{m}^h) &= \frac{1}{2}\sum_{i=1}^{N}(1+m_i^v)\log(1+m_i^v) + (1-m_i^v)\log(1-m_i^v) \\
&+ \frac{1}{2}\sum_{j=1}^{M}(1+m_j^h)\log(1+m_j^h) + (1-m_j^h)\log(1-m_j^h) \\
&- \sum_{i,j} w_{ij} m_i^v m_j^h + \sum_{i=1}^{N} a_i m_i^v + \sum_{j=1}^{M} b_j m_j^h \\
&\simeq \frac{1}{2}\sum_{i=1}^{N}(m_i^v)^2 + \frac{1}{2}\sum_{j=1}^{M}\left(m_j^h\right)^2 - \sum_{ij} w_{ij} m_i^v m_j^h \\
&+ \sum_{i=1}^{N} a_i m_i^v + \sum_{j=1}^{M} b_j m_j^h \qquad (13)
\end{aligned}
$$

Non-linear mean-field

- Thouless-Anderson-Palmer (TAP) free energy

$$
\begin{aligned}
F_{TAP}(\mathbf{m}^v, \mathbf{m}^h) = & + S(\mathbf{m}^v) + S(\mathbf{m}^h) \\
& - \sum_i \eta_i m_i^v - \sum_j \theta_j m_j^h - \sum_{i,j} w_{ij} m_i^v m_j^h \\
& + \sum_{i,j} \frac{w_{ij}^2}{2} \left(1 - {m_i^v}^2\right) \left(1 - {m_j^h}^2\right)
\end{aligned} \tag{14}
$$

- Replica symmetry framework

$$
\begin{aligned}
m_\alpha^v &= \left(\sigma_\alpha m_\alpha^h - a_\alpha\right)\left(1 - q_\alpha^v\right) \\
m_\alpha^h &= \left(\sigma_\alpha m_\alpha^v - b_\alpha\right)\left(1 - q_\alpha^h\right)
\end{aligned}
$$

$$
m_\alpha^v = E_{u,v,r}\left(\langle v_\alpha \rangle\right) \qquad m_\alpha^h = E_{u,v,r}\left(\langle h_\alpha \rangle\right)
$$

Gaussian approximation

$$\text{cov}(\mathbf{m^v}, \mathbf{m^h}) = \begin{pmatrix} \frac{\sigma_h^{-2}}{\sigma_v^{-2}\sigma_h^{-2} - \mathbf{W}\mathbf{W}^\intercal} & \mathbf{W}\frac{1}{\sigma_v^{-2}\sigma_h^{-2} - \mathbf{W}^\intercal\mathbf{W}} \\ \mathbf{W}^\intercal \frac{1}{\sigma_v^{-2}\sigma_h^{-2} - \mathbf{W}\mathbf{W}^\intercal} & \frac{\sigma_h^{-2}}{\sigma_v^{-2}\sigma_h^{-2} - \mathbf{W}\mathbf{W}^\intercal} \end{pmatrix} \tag{15}$$

$$\Downarrow$$

$$\langle v_\alpha h_\beta \rangle_{data} = \sigma_h^2 \sigma_\beta \langle v_\alpha v_\beta \rangle_{data} = \sigma_h^2 \sigma_\beta \, \text{cov}(v_\alpha, v_\beta) \tag{16}$$

$$\Downarrow$$

$$\frac{d\sigma_\alpha}{dt} = \sigma_h^2 \sigma_\alpha \left(\langle v_\alpha^2 \rangle_{data} - \frac{\sigma_v^2}{1 - \sigma_v^2 \sigma_h^2 \sigma_\alpha^2} \right) \tag{17}$$

Linear stability

$$\sigma_\alpha^2 = \begin{cases} \frac{\langle v_\alpha^2 \rangle_{data} - \sigma_v^2}{\sigma_v^2 \sigma_h^2 \langle v_\alpha^2 \rangle_{data}} & \langle v_\alpha^2 \rangle_{data} > \sigma_v^2 \\ 0 & \langle v_\alpha^2 \rangle_{data} < \sigma_v^2 \end{cases} \tag{18}$$

We see how the evolution of the singular values in the linear regime is driven by the SVD modes of the training data. The strongest modes, those above the threshold σ_v^2, are selected and learnt while the modes below threshold are damped.

Quenched mean-field equations

Statistical Physics kicks in! The **Replica trick** is used to get the mean-field equations for the non-linear regime (in Replica Symmetry setting)

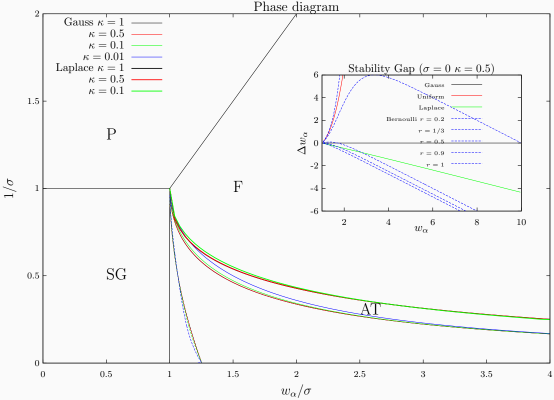
$$m_\alpha^v = \left(\sigma_\alpha m_\alpha^h - a_\alpha\right)\left(1 - q_\alpha^v\right)$$
$$m_\alpha^h = \left(\sigma_\alpha m_\alpha^v - b_\alpha\right)\left(1 - q_\alpha^h\right)$$

$$m_\alpha^v = E_{u,v,r}\left(\langle v_\alpha \rangle\right) \qquad m_\alpha^h = E_{u,v,r}\left(\langle h_\alpha \rangle\right)$$

where q_α^v, q_α^h are spin-glass order parameters

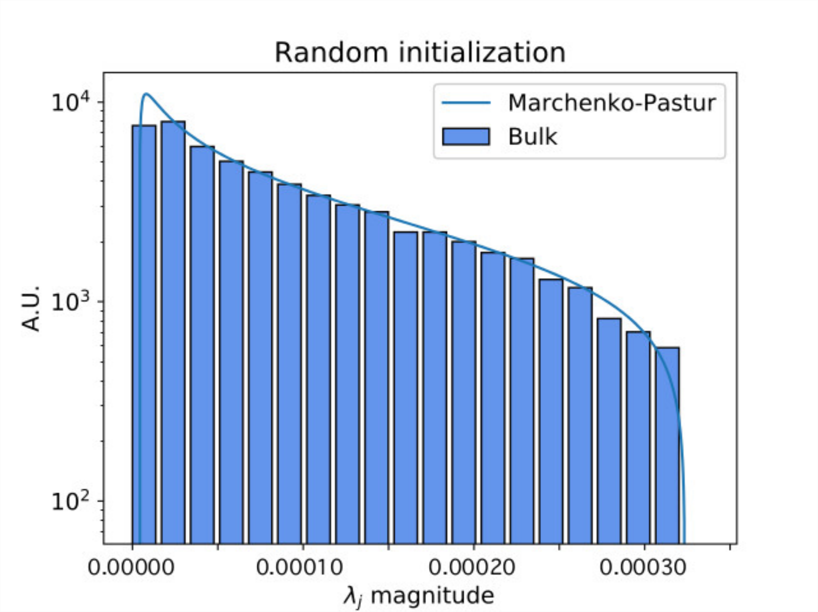
Note: averages are taken with respect to u_i, v_j and the noise r_{ij}. The specific realization of the weights is not important, just their distribution is.

From the mean-field equations we can compute the phase diagram of the model, a more complete description with respect to the stability analysis of the linear case:

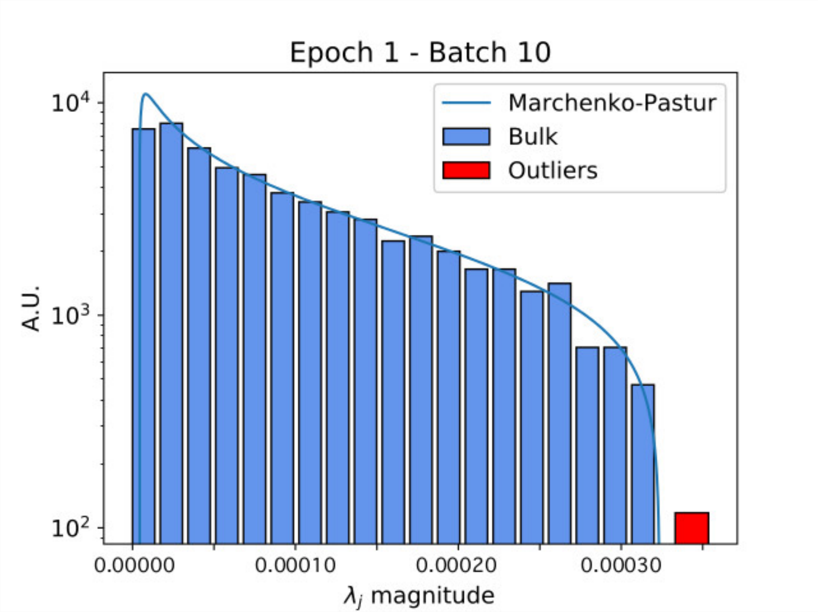


SVD analysis

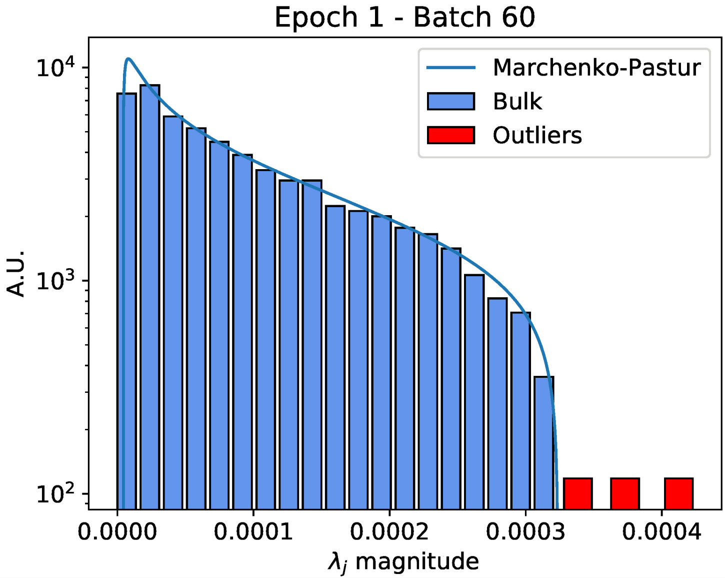
Singular values evolution



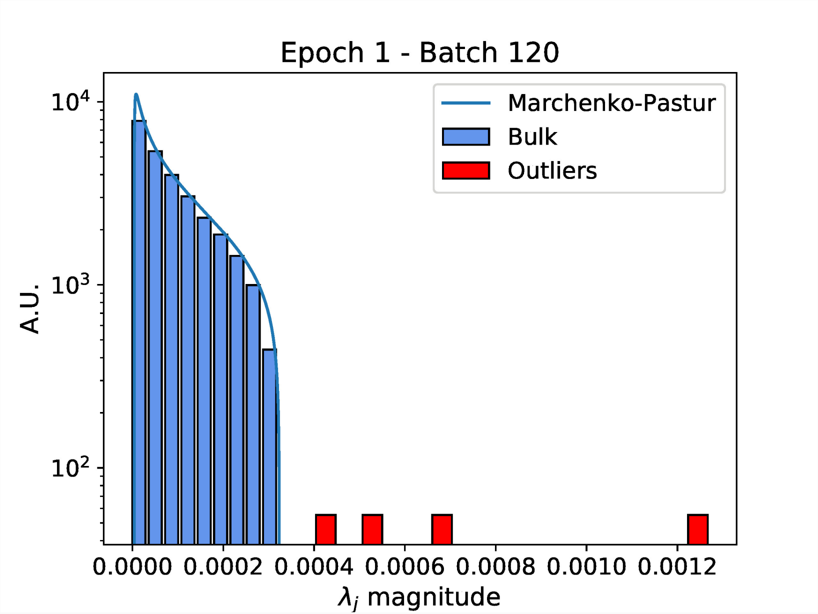
Singular values evolution



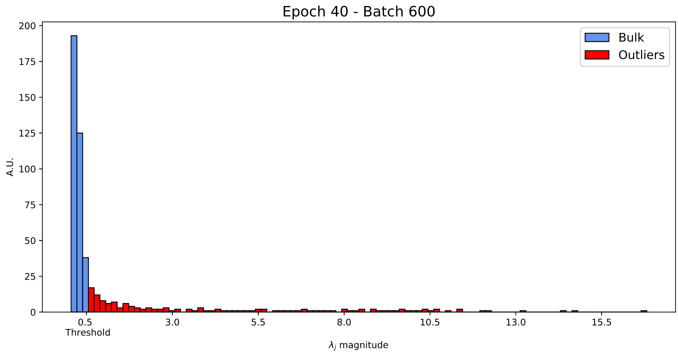
Singular values evolution



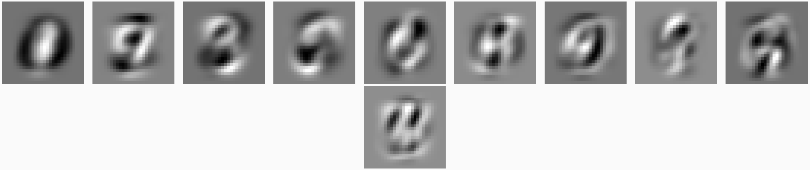
Singular values evolution



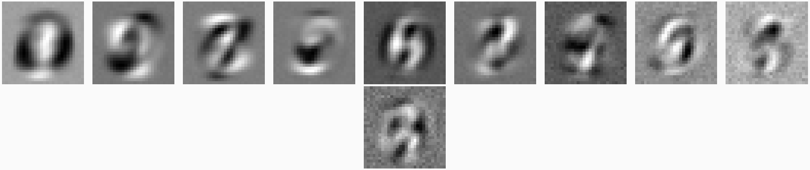
Singular values evolution



SVD modes



(a) SVD modes extracted from the training set



(b) The first 10 SVD modes of a RBM trained for 1 epoch