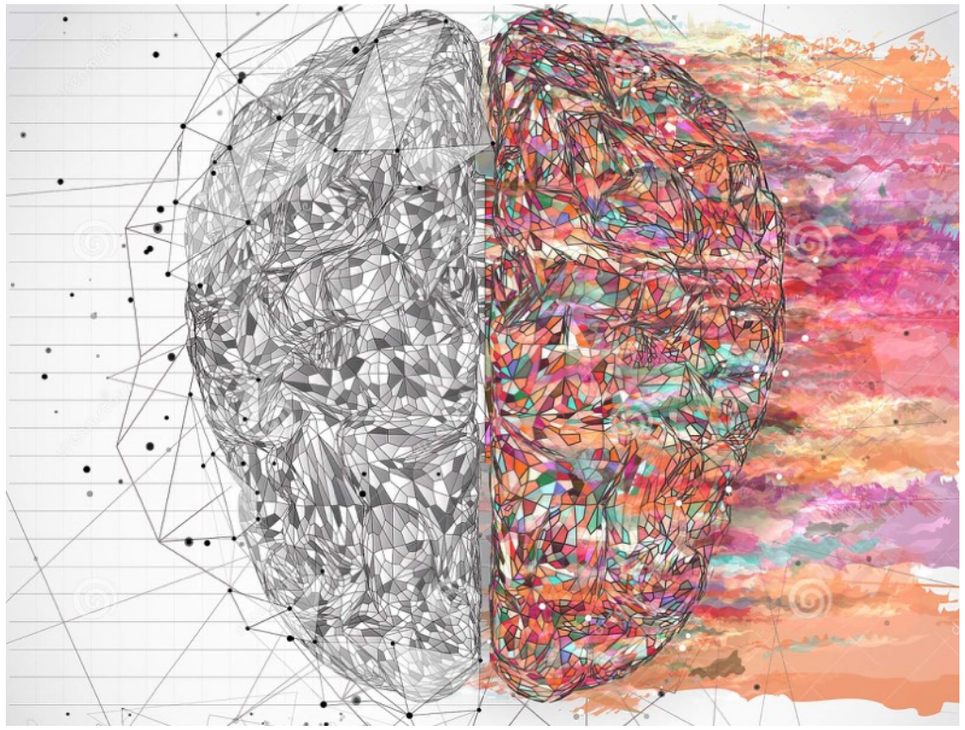
The early bird regularizes better

Elena Agliari

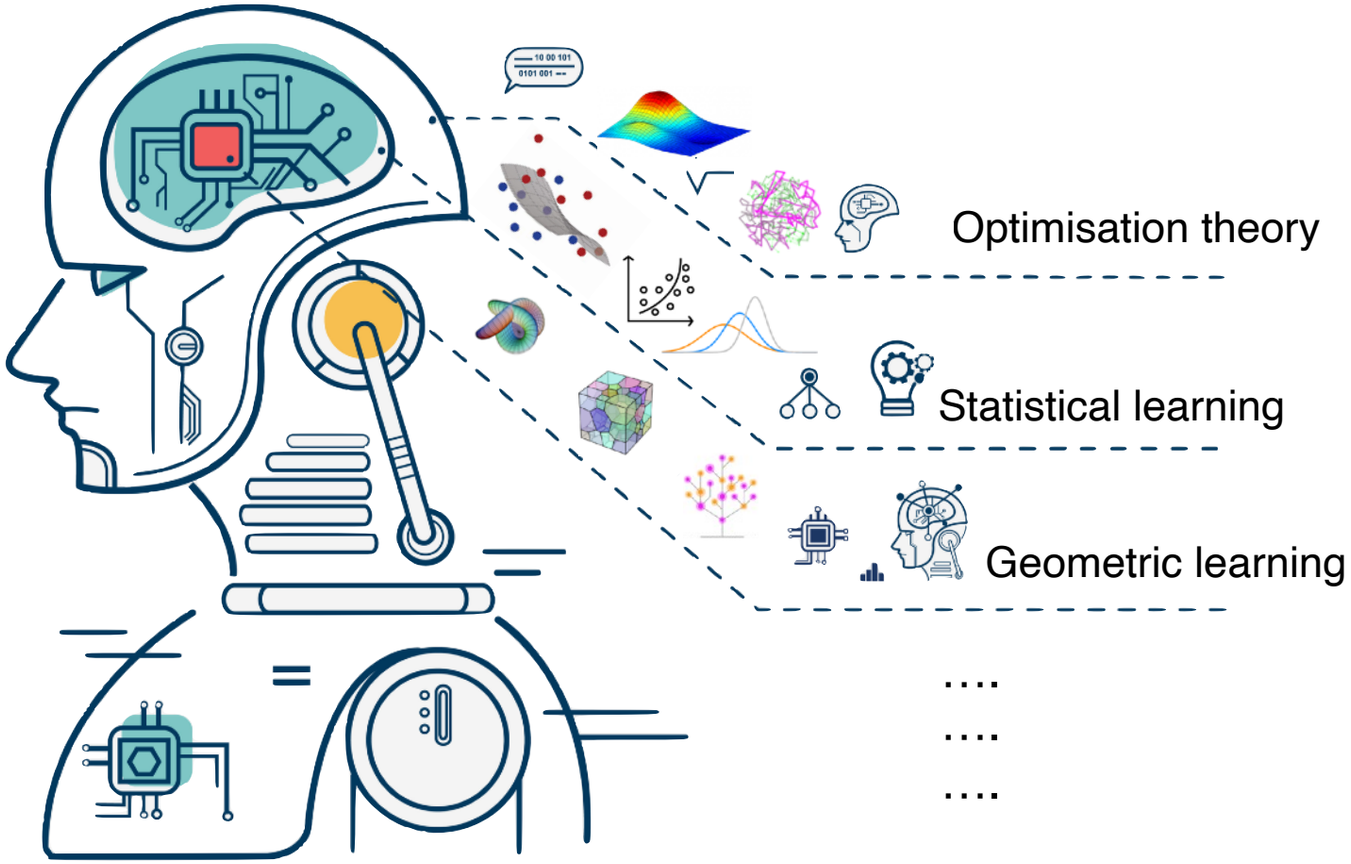
Dipartimento di Matematica, Sapienza Università di Roma
Istituto Nazionale di Alta Matematica, Roma

Joint work with:
Adriano Barra, Salento
Alberto Fachechi, Sapienza
Linda Albanese, Salento
Miriam Aquaro, Sapienza
Francesco Alemanno, Bologna



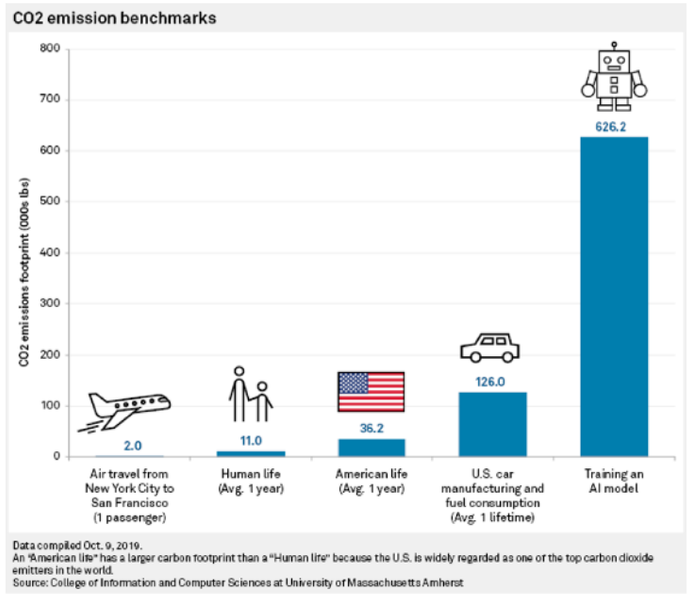
A Day on Statistical Physics for Machine Learning
Dipartimento di Matematica — Università Tor Vergata
15 September 2023

Hard sciences for machine learning



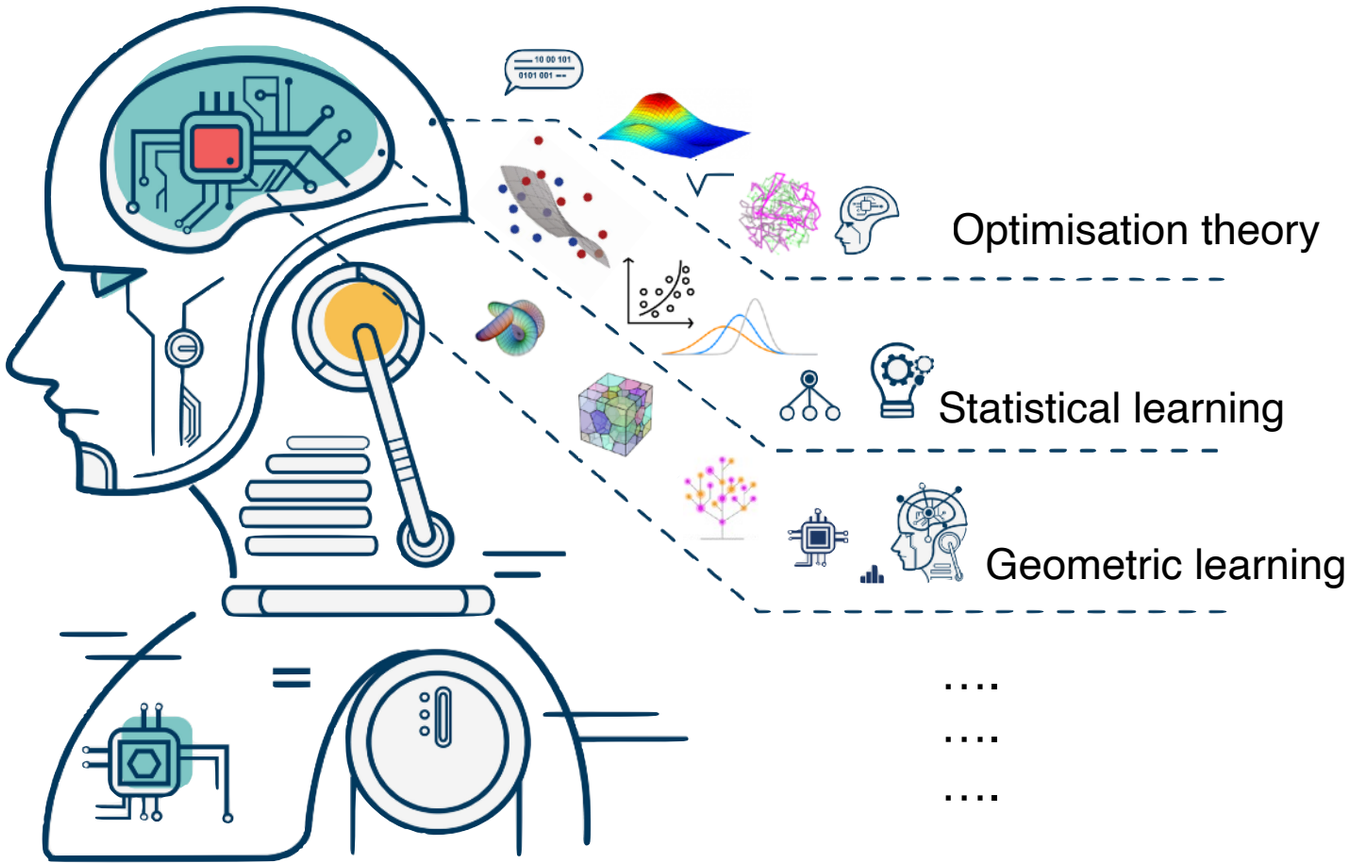
Mathematical control
- design optimal architecture a priori
- estimate hyperparameters
- check dataset size
- ....

→ reduce AI carbon footprint!

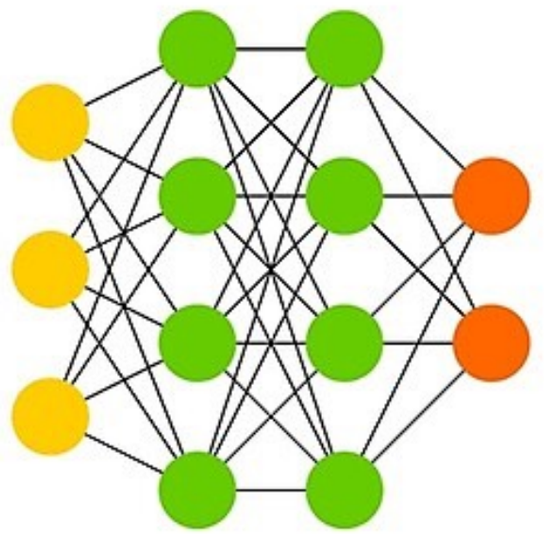


*transformer (213M parameters)
w/ neural architecture search

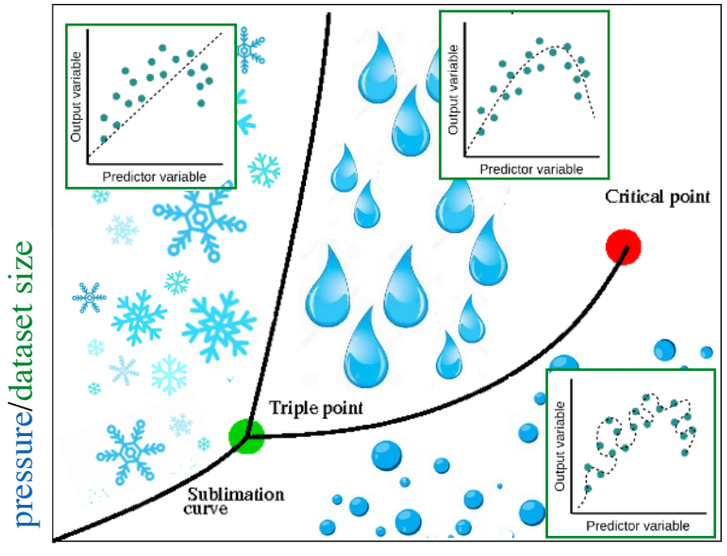
Hard sciences for machine learning



Statistical physics perspective



Information-processing capability as emerging collective behavior

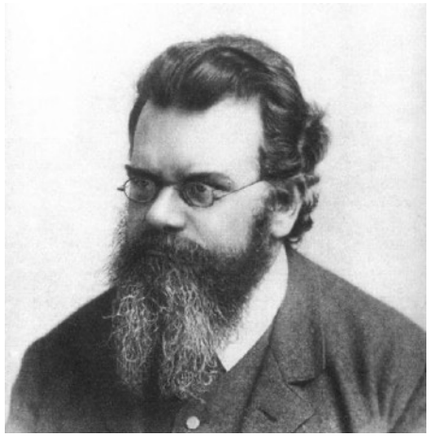


The Hopfield model

Mimics *retrieval* capabilities

Pattern recognition
Pattern reconstruction
Denoising
Content addressable memory

CAPTCHA

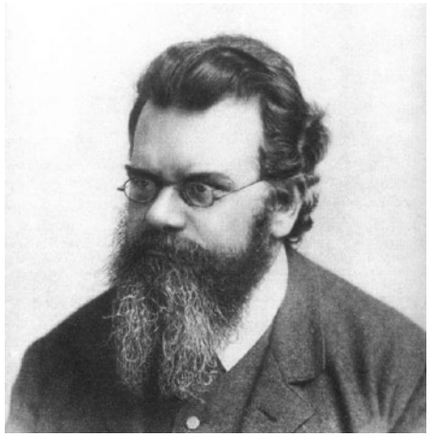


The Hopfield model

Mimics *retrieval* capabilities

Pattern recognition
Pattern reconstruction
Denoising
Content addressable memory

CAPTCHA

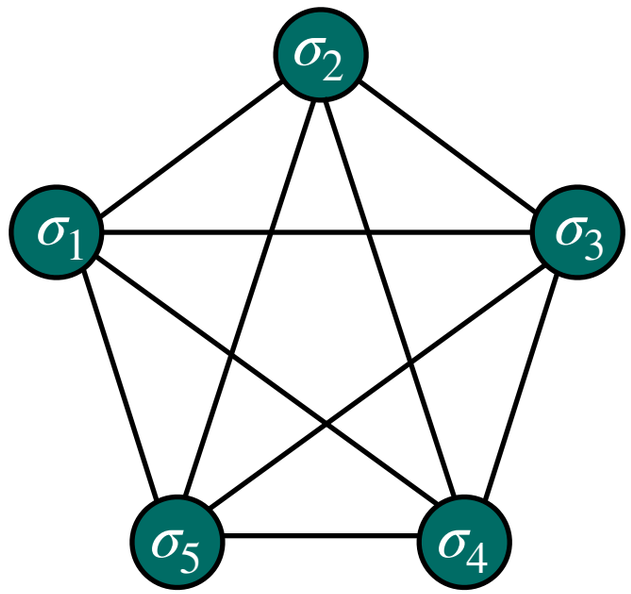


The Hopfield model

[Hopfield - PNAS '82; Pastur, Figotin - Theor. Math. Phys. '78]

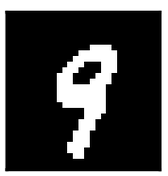
Set of N binary neurons $\boldsymbol{\sigma} \in \{-1, +1\}^N$ with pair-wise interaction strengths $\boldsymbol{J} \in \mathbb{R}^{N \times N}$

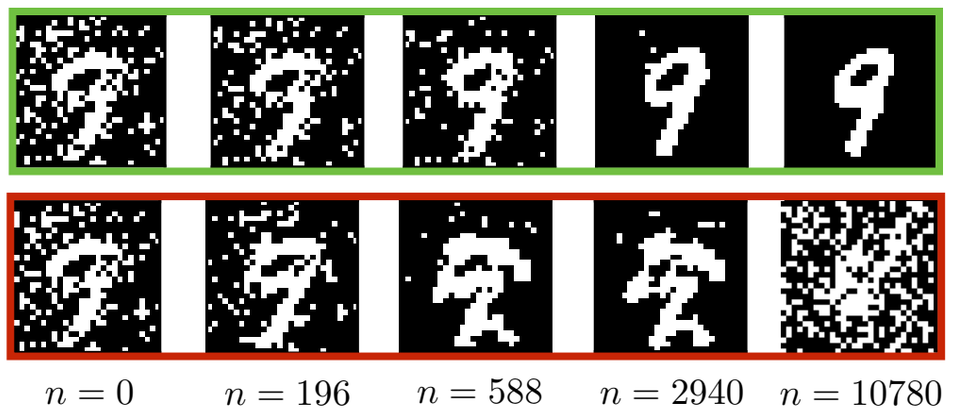
Neural dynamics $\sigma_i^{(n+1)} = \text{sign}\left[\sum_j J_{ij} \sigma_j^{(n)} + T\zeta_i \right]$, with $T \in \mathbb{R}^+$, ζ_i iid r.v.



Task: reconstruct K binary vectors (patterns) of length N
$\boldsymbol{\xi} := \{\boldsymbol{\xi}^\mu\}_{\mu=1,\dots,K} \in \{-1, +1\}^{N \times K}$

$\Rightarrow$ choose $\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{\xi})$ s.t. each pattern is (δ, ϵ)-stable for $\delta, \epsilon > 0$

e.g., $\boldsymbol{\xi}^\mu =$ 



The Hopfield model

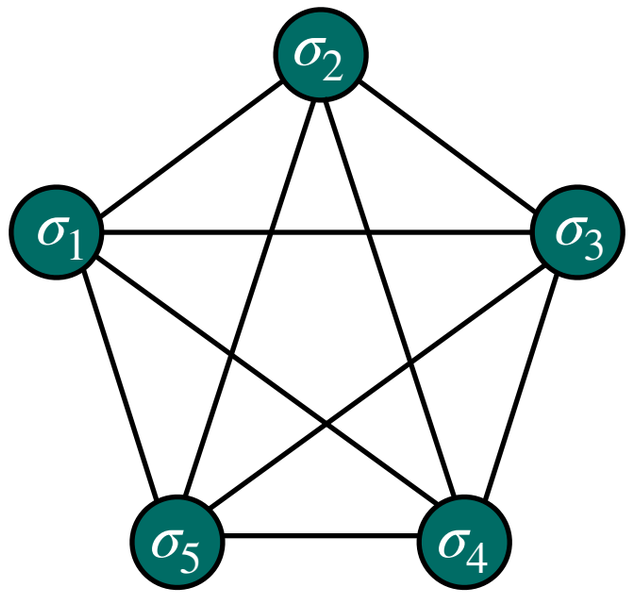
[Hopfield - PNAS '82; Pastur, Figotin - Theor. Math. Phys. '78]

Set of N binary neurons $\boldsymbol{\sigma} \in \{-1, +1\}^N$ with pair-wise interaction strengths $\boldsymbol{J} \in \mathbb{R}^{N \times N}$

Neural dynamics $\sigma_i^{(n+1)} = \text{sign}\left[\sum_j J_{ij} \sigma_j^{(n)} + T\zeta_i \right]$, with $T \in \mathbb{R}^+$, ζ_i iid r.v.

Task: reconstruct K binary vectors (patterns) of length N
$\boldsymbol{\xi} := \{\boldsymbol{\xi}^\mu\}_{\mu=1,\dots,K} \in \{-1, +1\}^{N \times K}$

$\Rightarrow$ choose $\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{\xi})$ s.t. each pattern is (δ, ϵ)-stable for $\delta, \epsilon > 0$



Gibbs measure is reversible w.r.t. neural dynamics

$\rho_{N,\boldsymbol{J},\beta}(\boldsymbol{\sigma}) := \dfrac{e^{-\beta \mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma})}}{\mathscr{Z}_{N,\boldsymbol{J},\beta}}$, with $\mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma}) = -\sum_{i,j=1}^{N} \sum_{\mu=1}^{K} \sigma_i \; J_{ij} \; \sigma_j$

$\Rightarrow$ choose $\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{\xi})$ s.t. $\boldsymbol{\xi}^\mu = \underset{\boldsymbol{\sigma} \in \{-1,+1\}^N}{\text{argmin}} \; \mathscr{H}_{N,\boldsymbol{J}}(\boldsymbol{\sigma})$ for $\mu = 1,\dots,K$

Hebb's rule $J_{ij}(\boldsymbol{\xi}) := (1 - \delta_{ij}) \sum_{\mu=1}^{N} \xi_i^\mu \xi_j^\mu$

AGS solution

[Amit, Gutfreund, Sompolinsky - Phys. Rev. A '87]

$$\mathscr{H}_{N,K,\boldsymbol{\xi}}(\boldsymbol{\sigma}) = -\frac{1}{2N}\sum_{\substack{i,j=1\\ i\neq j}}^{N}\sum_{\mu=1}^{K}\sigma_i\ \xi_i^\mu\xi_j^\mu\ \sigma_j$$

Patterns $\xi_i^\mu \sim_{iid} \mathscr{R},\ \forall i,\mu$

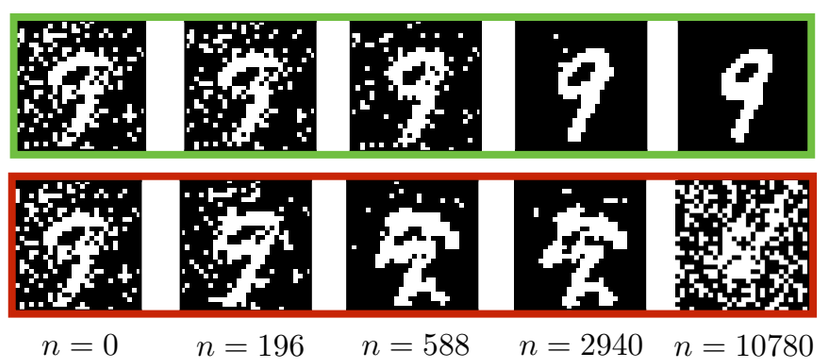
Magnetization $m_\mu := \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i,$

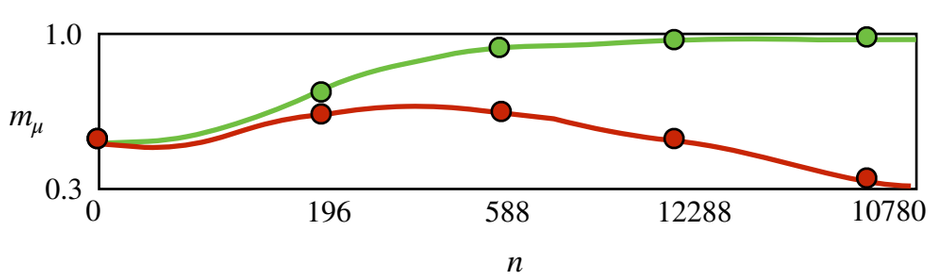
Overlap $q_{ab} := \frac{1}{N}\sum_{i=1}^{N}\sigma_i^{(a)}\sigma_i^{(b)}$

Noise T

Load $\alpha := \lim_{N\to\infty}\frac{K}{N}$

e.g., $\boldsymbol{\xi}^\mu =$ 





AGS solution

[Amit, Gutfreund, Sompolinsky - Phys. Rev. A '87]

$$\mathscr{H}_{N,K,\boldsymbol{\xi}}(\boldsymbol{\sigma}) = -\frac{1}{2N}\sum_{\substack{i,j=1\\ i\neq j}}^{N}\sum_{\mu=1}^{K}\sigma_i\ \xi_i^\mu\xi_j^\mu\ \sigma_j$$

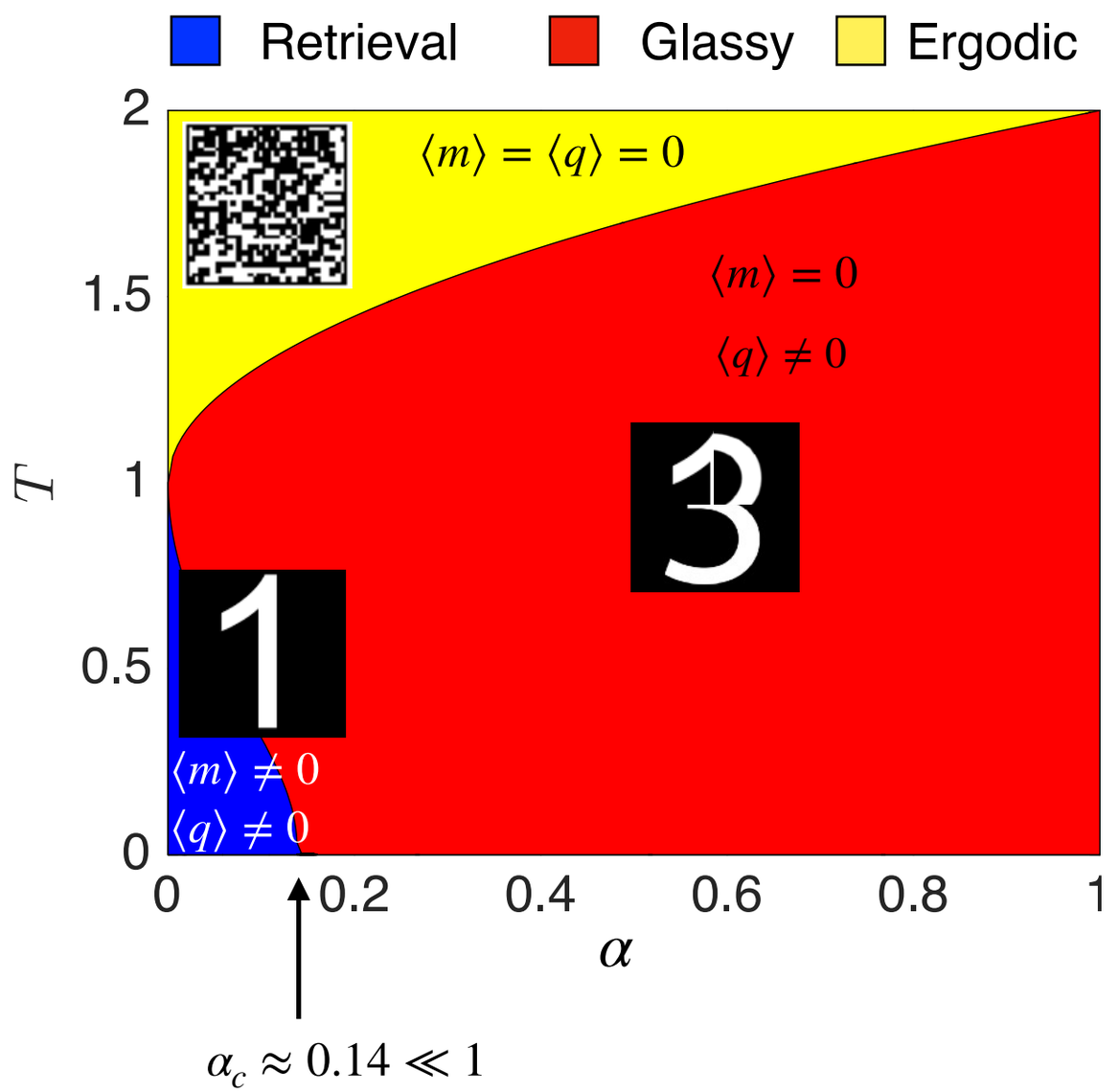
Patterns $\xi_i^\mu \sim_{iid} \mathscr{R},\ \forall i,\mu$

Magnetization $m_\mu := \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i,$

Overlap $q_{ab} := \frac{1}{N}\sum_{i=1}^{N}\sigma_i^{(a)}\sigma_i^{(b)}$

Noise T

Load $\alpha := \lim_{N\to\infty}\frac{K}{N}$



Revise Hebb's rule to enhance storing
(e.g., Kohonen '72, Personnaz '85, Kanter&Sompolinsky '86, Opper '89, Dotsenko '91, Plakhov '94, Zamponi et al. '22)

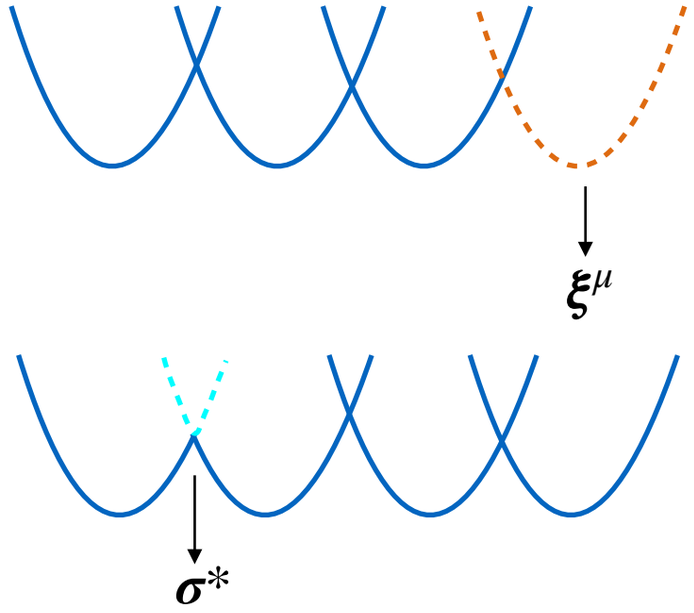
possibly including iterative protocols
(e.g., Hopfield '83, Hopfield '84, Parisi '86, Sherrington '95, van Hemmen '97, Marinari '18)

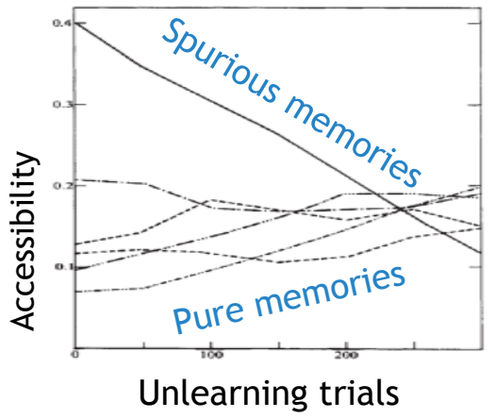
Revise Hebb's rule by iterative protocols

Storing rule $$J_{ij}^{(\mu)} \leftarrow J_{ij}^{(\mu-1)} + \xi_i^\mu \xi_j^\mu$$

*Un*storing rule $$J_{ij}^{(n)} \leftarrow J_{ij}^{(n-1)} - \sigma_i^* \sigma_j^*$$

[Hopfield et al. - Nature Lett. '83]



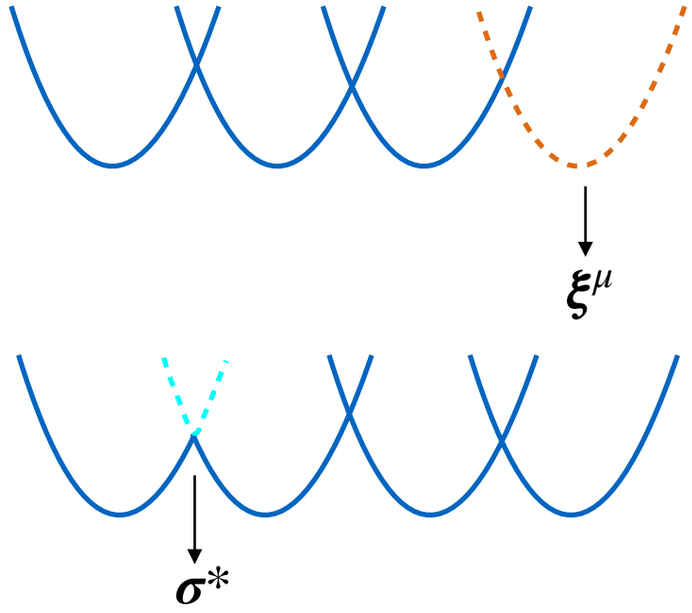


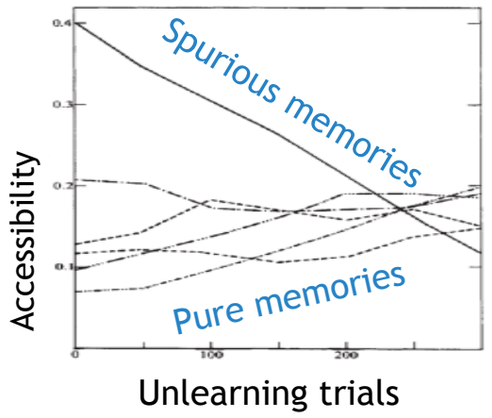
Revise Hebb's rule by iterative protocols

Storing rule $$J_{ij}^{(\mu)} \leftarrow J_{ij}^{(\mu-1)} + \xi_i^\mu \xi_j^\mu$$

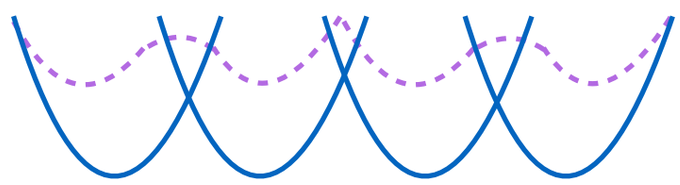
*Un*storing rule $$J_{ij}^{(n)} \leftarrow J_{ij}^{(n-1)} - \sigma_i^* \sigma_j^*$$

[Hopfield et al. - Nature Lett. '83]





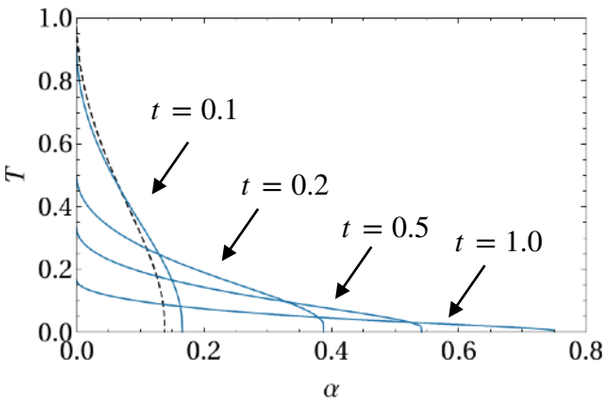
Analytical unlearning rule $$J_{ij}^{(n)} \leftarrow J_{ij}^{(n-1)} - \varphi_i(\boldsymbol{\sigma}^*)\varphi_j(\boldsymbol{\sigma}^*)$$



$$J_{ij}(t) = \frac{1}{N} \sum_{\mu,\nu} \xi_i^\mu \left(1 + t\boldsymbol{C}\right)^{-1}_{\mu\nu} \xi_j^\nu$$

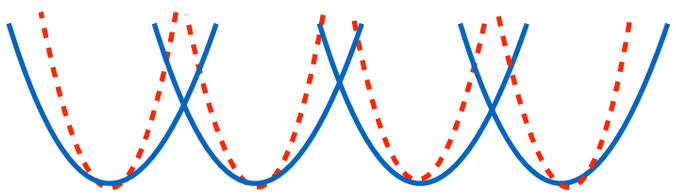
where $C_{\mu\nu} := \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \xi_j^\mu$ pattern correlation matrix

[Dotsenko et al. - J. Phys. A '91, Plakhov - IEEE Patt. Recogn. '94]



Revise Hebb's rule by iterative protocols

Remotion&Reinforcement

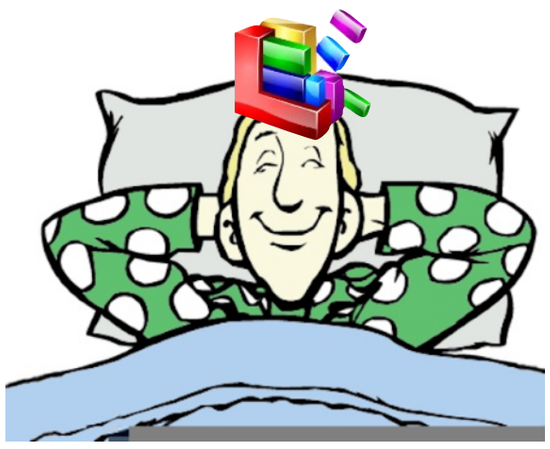


$$J_{ij}^{(n)} \leftarrow J_{ij}^{(n-1)} + \frac{1}{1+n}[J^{(n-1)} - (J^{(n-1)})^2]$$

$$J_{ij}(t) = \frac{1}{N}\sum_{\mu,\nu} \xi_i^\mu \left(\frac{1+t}{1+t\boldsymbol{C}}\right)_{\mu\nu} \xi_j^\nu$$

where $C_{\mu\nu} := \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \xi_j^\mu$ pattern correlation matrix

[Fachechi, Agliari, Barra - Neur. Net. '19]



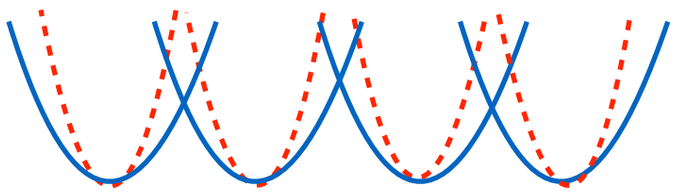
$$\underbrace{(1+t)}_{\text{reinforcement (SW)}} \cdot \underbrace{(\boldsymbol{I}+t\boldsymbol{C})^{-1}}_{\text{remotion (REM)}}$$

t "sleeping time"

[Crick, Mitchinson - Nature '83; Stickgold - Nature '05; Diekelmann, Born - Nature Rev. Neurosc. '10]

Revise Hebb's rule by iterative protocols

Remotion&Reinforcement



$$J_{ij}^{(n)} \leftarrow J_{ij}^{(n-1)} + \frac{1}{1+n}[J^{(n-1)} - (J^{(n-1)})^2]$$

$$J_{ij}(t) = \frac{1}{N}\sum_{\mu,\nu} \xi_i^\mu \left(\frac{1+t}{1+t\boldsymbol{C}}\right)_{\mu\nu} \xi_j^\nu$$

where $C_{\mu\nu} := \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \xi_j^\mu$ pattern correlation matrix

[Fachechi, Agliari, Barra - Neur. Net. '19]

$$J_{ij}(t) = \frac{1}{N}\sum_{\mu,\nu} \xi_i^\mu \left(\frac{1+t}{1+t\boldsymbol{C}}\right)_{\mu\nu} \xi_j^\nu$$

$t \to 0$: Hebb's rule

$t \to \infty$: Kohonen rule '72, Kanter-Sompolinsky '86

“Dreaming” Hopfield network

$$\mathscr{H}^{(\mathrm{DHN})}_{N,K,\boldsymbol{\xi},t}(\boldsymbol{\sigma}) = -\frac{1}{2N}\sum_{i,j=1}^{N,N}\sum_{\mu,\nu=1}^{K,K}\sigma_i\xi_i^{\mu}\left(\frac{1+t}{\boldsymbol{I}+t\boldsymbol{C}}\right)_{\mu,\nu}\xi_j^{\nu}\sigma_j$$

Goal: explicit expression for quenched pressure $\quad \mathscr{A}_{\beta,\alpha,t} := \lim_{N\uparrow\infty}\frac{1}{N}\mathbb{E}\log\mathscr{Z}_{\beta,N,\boldsymbol{\xi},t}$

Interpolating technique

$$\mathscr{H}_N(\boldsymbol{\sigma};\boldsymbol{J}) \to \tilde{\mathscr{H}}_N(\boldsymbol{\sigma};\boldsymbol{J},\boldsymbol{J}') = s\ \mathscr{H}_N(\boldsymbol{\sigma};\boldsymbol{J}) + (1-s)\ \mathscr{H}_N^{easy}(\boldsymbol{\sigma};\boldsymbol{J}'),\ s \in [0,1]$$

$\tilde{\mathscr{H}}_N(\boldsymbol{\sigma};\boldsymbol{J},\boldsymbol{J}')$ yields $\tilde{\mathscr{Z}}_{N,\beta}(\boldsymbol{J},\boldsymbol{J}'),\ \tilde{\rho}_{N,\beta}(\boldsymbol{\sigma};\boldsymbol{J},\boldsymbol{J}')$

$$\mathscr{A}_N = \tilde{\mathscr{A}}_N(1) = \tilde{\mathscr{A}}_N(0) + \int_0^1 \frac{d\tilde{\mathscr{A}}_N(s)}{ds}\bigg|_{s=s'} ds'$$

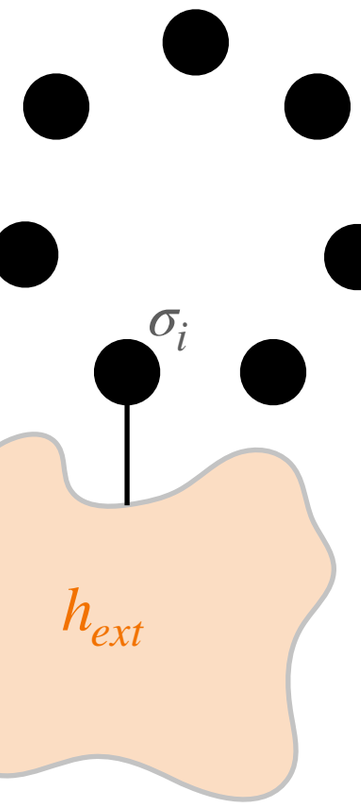
$$= \tilde{\mathscr{A}}_N(0) + f[\bar{m},\bar{q},\dots] + \int_0^1 g[(m-\bar{m})^2,(q-\bar{q})^2,\dots,s']\big|_{s=s'} ds'$$

$$= \mathscr{A}_N^{trial} + \int_0^1 R_N(s)\ ds$$

$\downarrow$ $N \to \infty$

$\downarrow$ ansatz (e.g., k-RSB)

0



[Guerra - Fields Inst. Comm. ’01]

Hopfield w/reinforcement&remotion

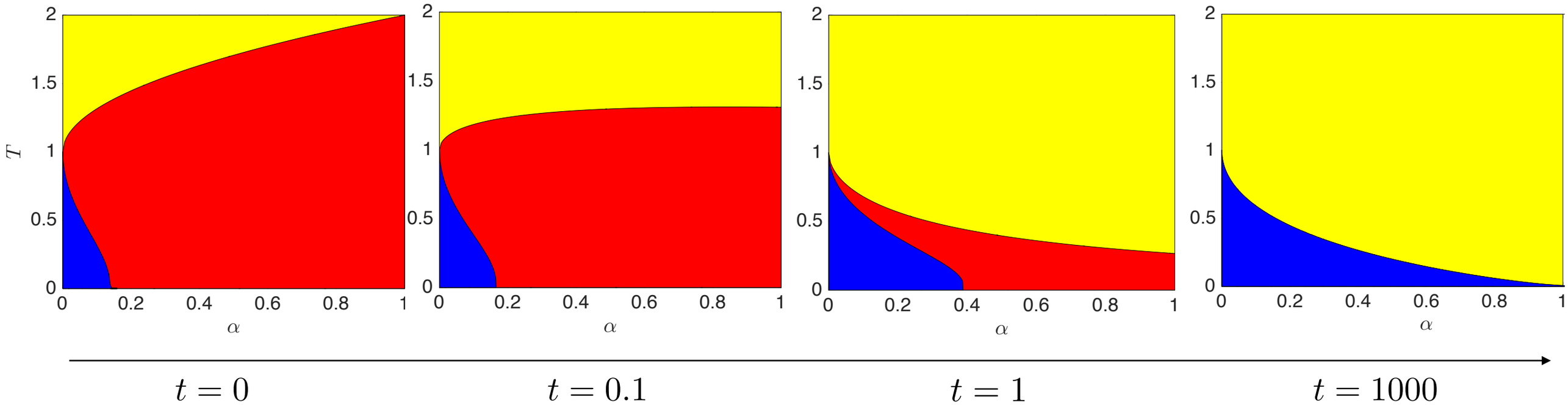
$$\mathscr{H}^{(\mathrm{DHM})}_{N,K,\boldsymbol{\xi},t}(\boldsymbol{\sigma}) = -\frac{1}{2N}\sum_{i,j=1}^{N,N}\sum_{\mu,\nu=1}^{K,K}\sigma_i\xi_i^{\mu}\left(\frac{1+t}{\boldsymbol{I}+t\boldsymbol{C}}\right)_{\mu,\nu}\xi_j^{\nu}\sigma_j$$

Goal: explicit expression for quenched pressure $\quad \mathscr{A}_{\beta,\alpha,t} := \lim_{N\uparrow\infty}\frac{1}{N}\mathbb{E}\log\mathscr{Z}_{\beta,N,\boldsymbol{\xi},t}$

Hopfield w/reinforcement&remotion

$$\mathscr{H}^{(\mathrm{DHM})}_{N,K,\boldsymbol{\xi},t}(\boldsymbol{\sigma}) = -\frac{1}{2N}\sum_{i,j=1}^{N,N}\sum_{\mu,\nu=1}^{K,K}\sigma_i\xi_i^{\mu}\left(\frac{1+t}{\boldsymbol{I}+t\boldsymbol{C}}\right)_{\mu,\nu}\xi_j^{\nu}\sigma_j$$

Goal: explicit expression for quenched pressure $\quad \mathscr{A}_{\beta,\alpha,t} := \lim_{N\uparrow\infty}\frac{1}{N}\mathbb{E}\log\mathscr{Z}_{\beta,N,\boldsymbol{\xi},t}$



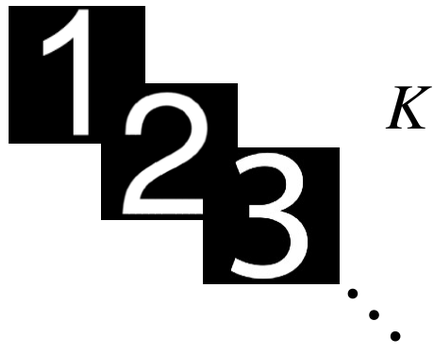
- Retrieval region gets wider
- Spin-glass collapses
- Ergodic region gets wider

Allocate more information with the same resources

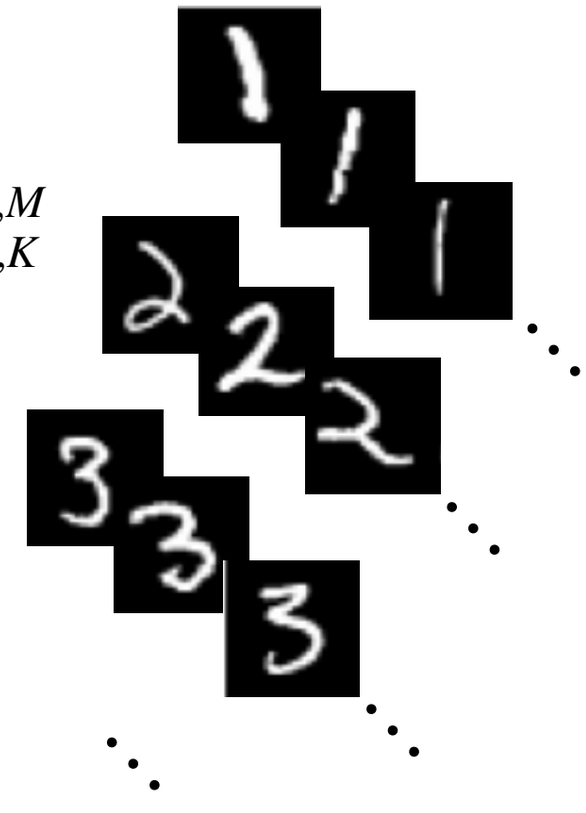
[Agliari, Alemanno, Barra, Fachechi - Neur. Net. ’19, J. Stat. ’19]

Replace *archetypes* by a sample of *examples*

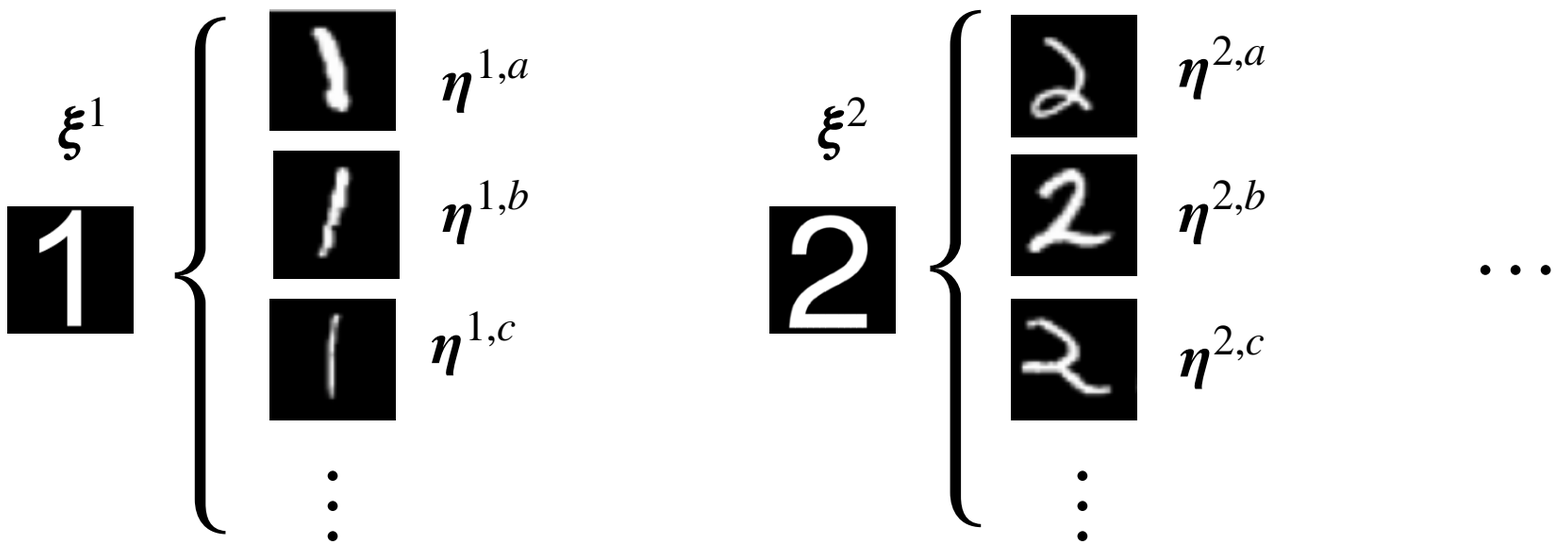
K "archetypes" $\{\boldsymbol{\xi}^{\mu}\}_{\mu=1,\dots,K}$



$K \times M$ "blurred" examples $\{\boldsymbol{\eta}^{\mu,a}\}_{\substack{a=1,\dots,M \\ \mu=1,\dots,K}}$

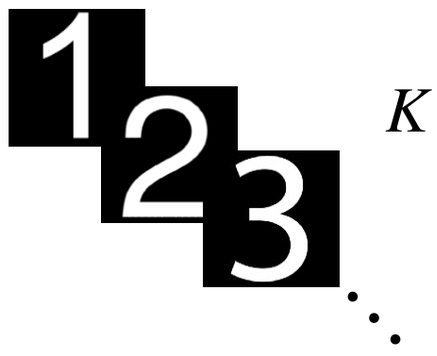


$$\eta_i^{\mu,a} = \xi_i^{\mu}\,\chi_i^{\mu,a} \qquad \text{with } \mathscr{P}(\chi_i^{\mu,a} = \pm 1) = \frac{1 \pm r}{2},\ r \in [0,1]$$

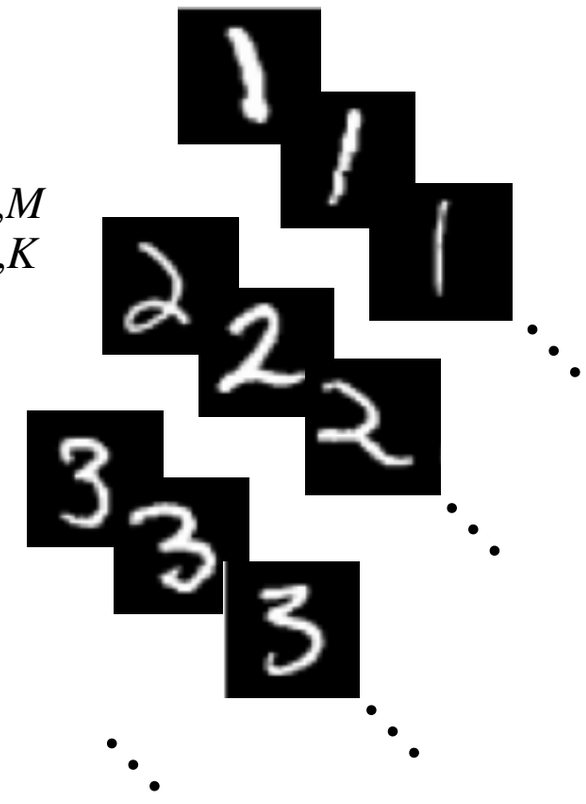


Replace *archetypes* by a sample of *examples*

K "archetypes" $\{\boldsymbol{\xi}^{\mu}\}_{\mu=1,\dots,K}$



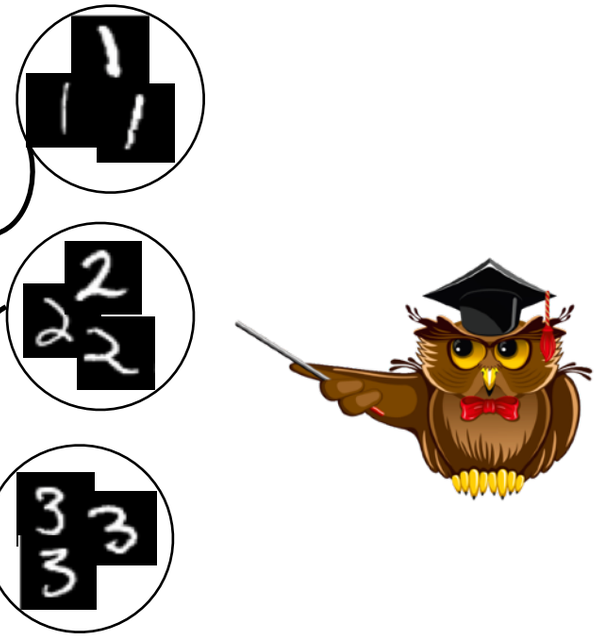
$K \times M$ "blurred" examples $\{\boldsymbol{\eta}^{\mu,a}\}_{\substack{a=1,\dots,M \\ \mu=1,\dots,K}}$



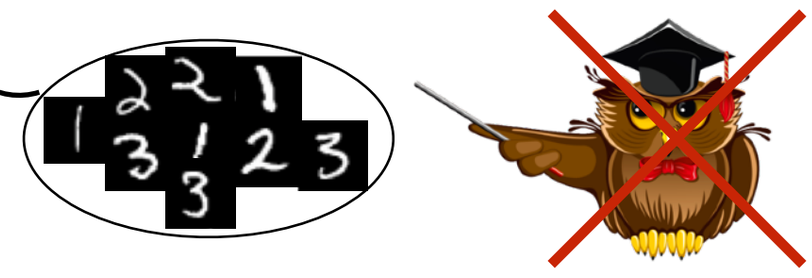
$$\eta_i^{\mu,a} = \xi_i^{\mu}\chi_i^{\mu,a} \qquad \text{with } \mathscr{P}(\chi_i^{\mu,a} = \pm 1) = \frac{1 \pm r}{2},\ r \in [0,1]$$

$$J_{ij}(\boldsymbol{\xi}) \propto \sum_{\mu=1}^{K} \xi_i^{\mu}\xi_i^{\mu}$$

$$J_{ij}^{(sup)} \propto \sum_{\mu=1}^{K}\left(\sum_{a=1}^{M}\eta_i^{\mu,a}\right)\left(\sum_{b=1}^{M}\eta_i^{\mu,b}\right)$$



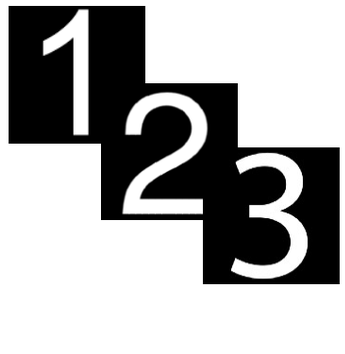
$$J_{ij}^{(unsup)} \propto \sum_{\mu=1}^{K}\sum_{a=1}^{M}\eta_i^{\mu,a}\eta_j^{\mu,a}$$



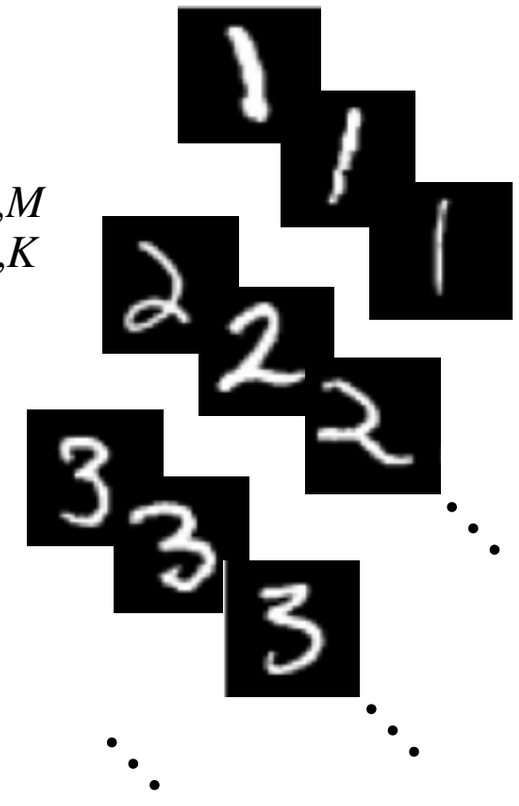
[Agliari, Alemanno, Barra, De Marzo - Neur. Net. '22]

Replace *archetypes* by a sample of *examples*

K "archetypes" $\{\boldsymbol{\xi}^{\mu}\}_{\mu=1,\dots,K}$



$K \times M$ "blurred" examples $\{\boldsymbol{\eta}^{\mu,a}\}_{\substack{a=1,\dots,M \\ \mu=1,\dots,K}}$



$$\eta_i^{\mu,a} = \xi_i^{\mu}\,\chi_i^{\mu,a} \qquad \text{with } \mathscr{P}(\chi_i^{\mu,a} = \pm 1) = \frac{1 \pm r}{2},\ r \in [0,1]$$

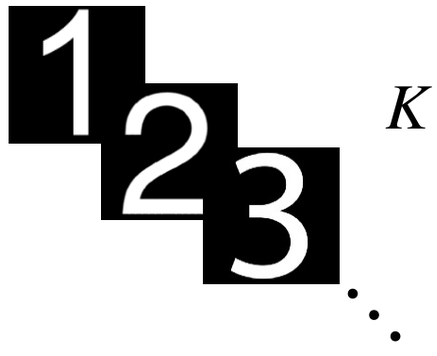
$$J_{ij}^{(sup)}(\boldsymbol{\eta}) \propto \sum_{\mu,\nu=1}^{K} \sum_{a,b=1}^{M} \eta_i^{\mu,a} \left(\frac{1+t}{\boldsymbol{I} + t\boldsymbol{C}(\bar{\boldsymbol{\eta}})} \right)_{\mu\nu} \eta_j^{\nu,b} = \frac{1}{N} \sum_{\mu,\nu=1}^{K} \bar{\eta}_i^{\mu} \left(\frac{1+t}{\boldsymbol{I} + t\boldsymbol{C}(\bar{\boldsymbol{\eta}})} \right)_{\mu\nu} \bar{\eta}_j^{\nu}$$

$$J_{ij}(\boldsymbol{\xi}) \propto \sum_{\mu,\nu=1}^{K} \xi_i^{\mu} \left(\frac{1+t}{\boldsymbol{I} + t\boldsymbol{C}(\boldsymbol{\xi})} \right)_{\mu\nu} \xi_j^{\nu}$$

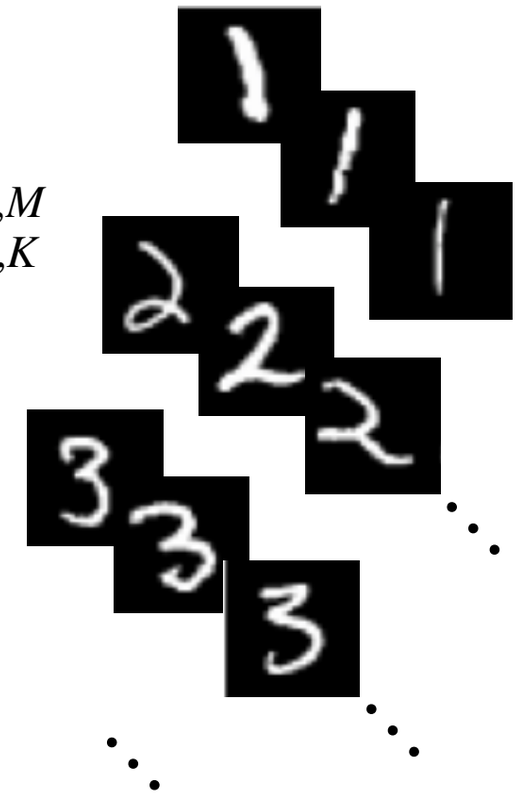
$$J_{ij}^{(unsup)}(\boldsymbol{\eta}) \propto \sum_{(\mu,a)=(1,1)}^{(K,M)} \sum_{(\nu,b)=(1,1)}^{(K,M)} \eta_i^{\mu,a} \left(\frac{1+t}{\boldsymbol{I} + t\boldsymbol{C}(\boldsymbol{\eta})} \right)_{(\mu,a),(\nu,b)} \eta_j^{\nu,b}$$

Replace *archetypes* by a sample of *examples*

K "archetypes" $\{\boldsymbol{\xi}^{\mu}\}_{\mu=1,\dots,K}$



$K \times M$ "blurred" examples $\{\boldsymbol{\eta}^{\mu,a}\}_{\substack{a=1,\dots,M \\ \mu=1,\dots,K}}$



$$\eta_i^{\mu,a} = \xi_i^{\mu}\,\chi_i^{\mu,a} \qquad \text{with } \mathscr{P}(\chi_i^{\mu,a} = \pm 1) = \frac{1 \pm r}{2},\ r \in [0,1]$$

$$J_{ij}^{(sup)}(\boldsymbol{\eta}) \propto \sum_{\mu,\nu=1}^{K} \sum_{a,b=1}^{M} \eta_i^{\mu,a} \left(\frac{1+t}{\boldsymbol{I} + t\boldsymbol{C}(\bar{\boldsymbol{\eta}})} \right)_{\mu\nu} \eta_j^{\nu,b} = \frac{1}{N} \sum_{\mu,\nu=1}^{K} \bar{\eta}_i^{\mu} \left(\frac{1+t}{\boldsymbol{I} + t\boldsymbol{C}(\bar{\boldsymbol{\eta}})} \right)_{\mu\nu} \bar{\eta}_j^{\nu}$$

$$J_{ij}(\boldsymbol{\xi}) \propto \sum_{\mu,\nu=1}^{K} \xi_i^{\mu} \left(\frac{1+t}{\boldsymbol{I} + t\boldsymbol{C}(\boldsymbol{\xi})} \right)_{\mu\nu} \xi_j^{\nu}$$

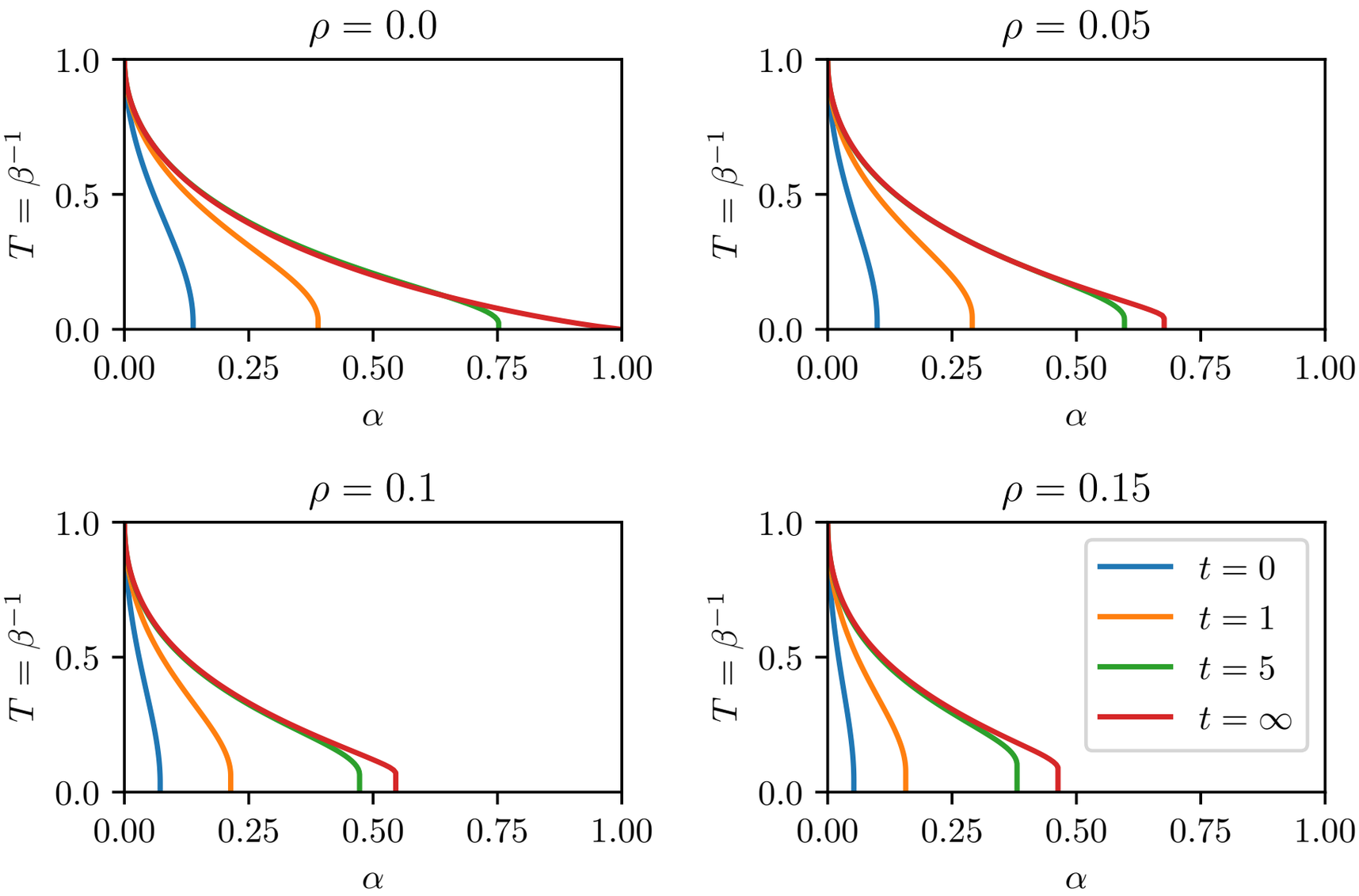
$$\mathscr{H}_{N,K,\boldsymbol{\eta},t}^{(\mathrm{DHN})}(\boldsymbol{\sigma}) = -\frac{1}{2NM^2} \sum_{i,j=1}^{N,N} \sum_{\mu,\nu=1}^{K,K} \sum_{a,b=1}^{M} \sigma_i \eta_i^{\mu,a} \left(\frac{1+t}{\boldsymbol{I} + t\boldsymbol{C}(\bar{\boldsymbol{\eta}})} \right) \eta_j^{\nu,b} \sigma_j$$

T, α, t

$\rho(M, r) := \dfrac{1 - r^2}{Mr^2}$ Dataset “entropy”

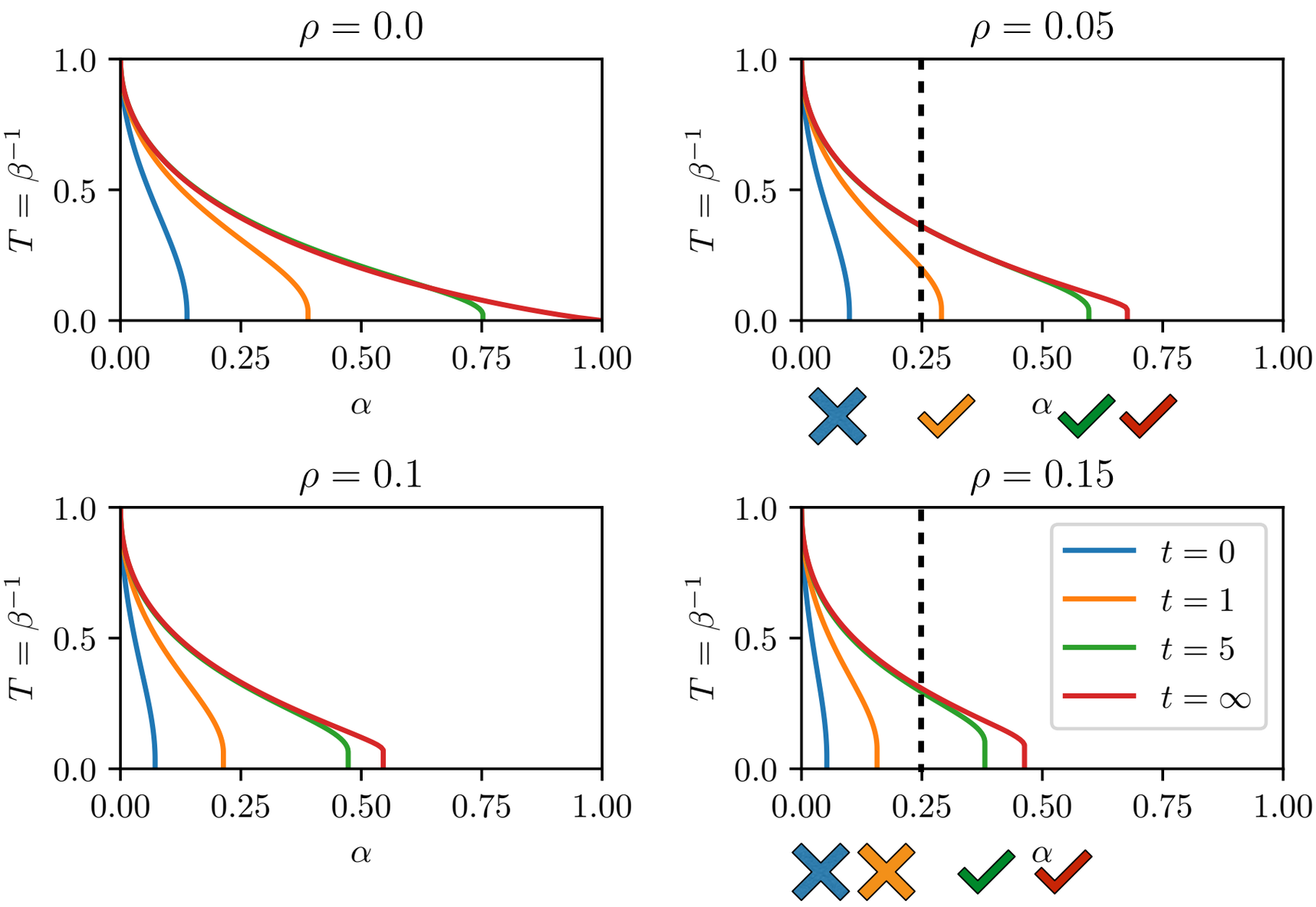
T, α, t

$\rho(M, r) := \dfrac{1 - r^2}{Mr^2}$ Dataset “entropy”



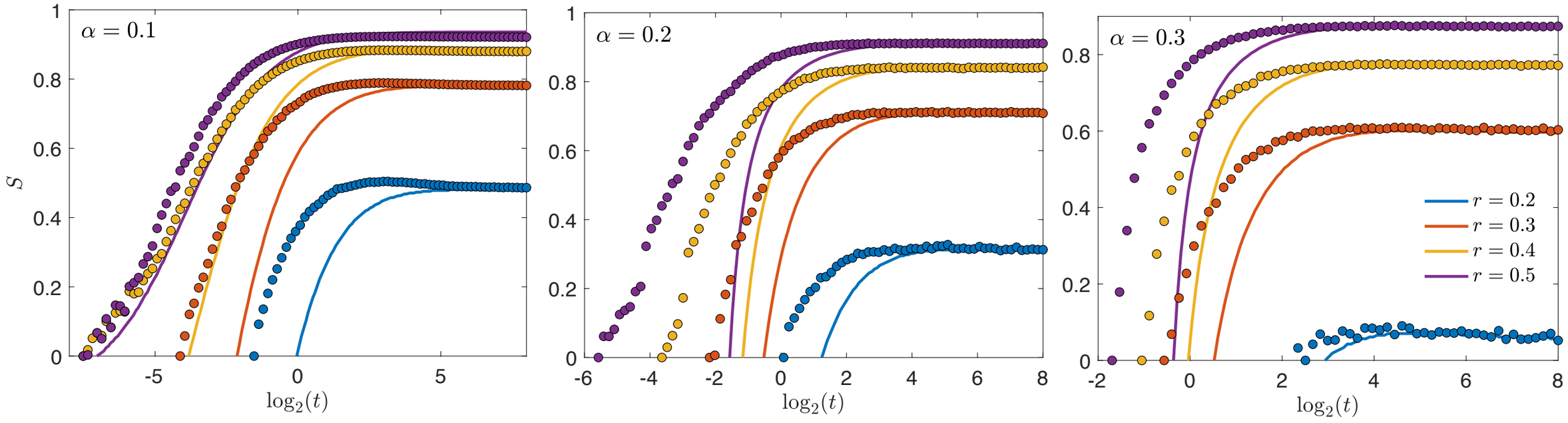
T, α, t

$\rho(M, r) := \dfrac{1 - r^2}{Mr^2}$ Dataset “entropy”



$T = 0,\ \alpha = 0.25$

$t = 1 \to t = 5$

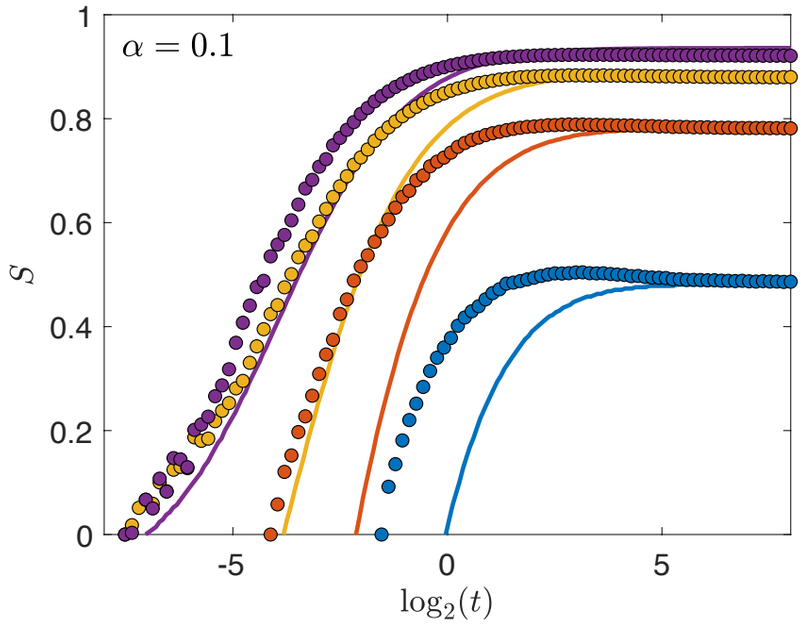


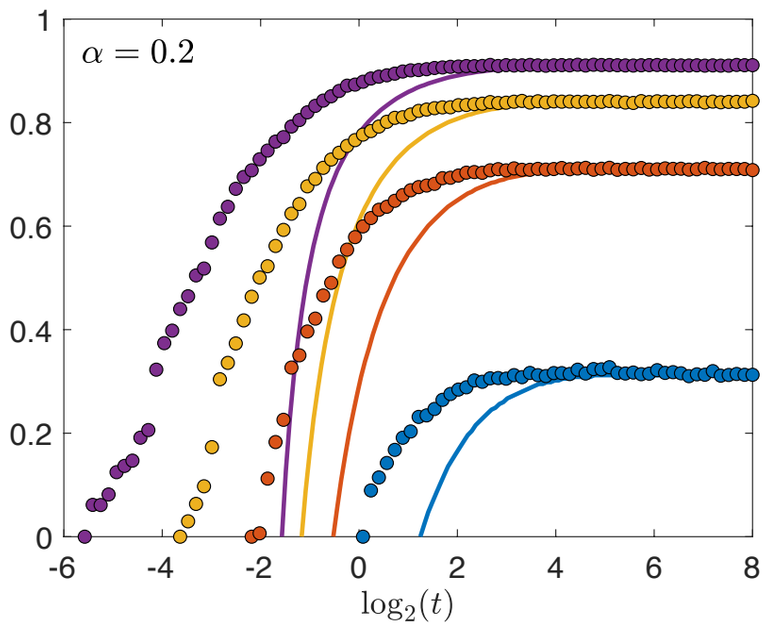
$M_c(m_\times, t)$ lowest number of examples for success ($m_\mu > m_\times$)

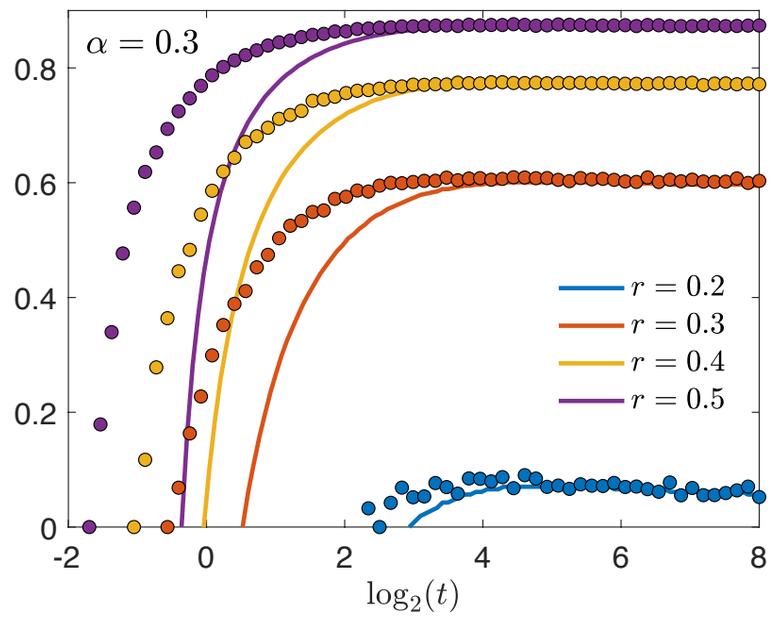
$$S(m_\times, t) = 1 - \frac{M_c}{\max_{t \in \mathbb{R}^+} M_c}$$ data “saving”

[Aquaro, Alemanno, Kanter, Durante, Barra, Agliari, submitted]

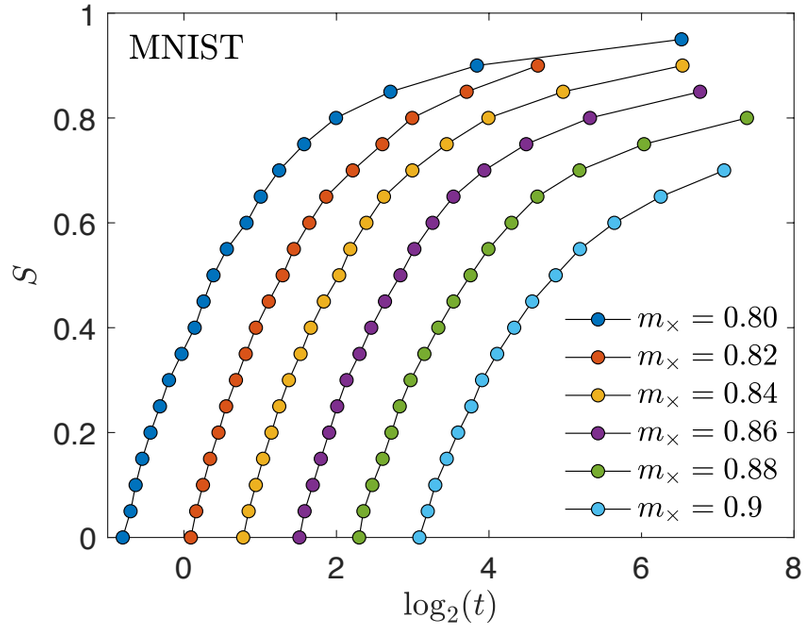
Structureless

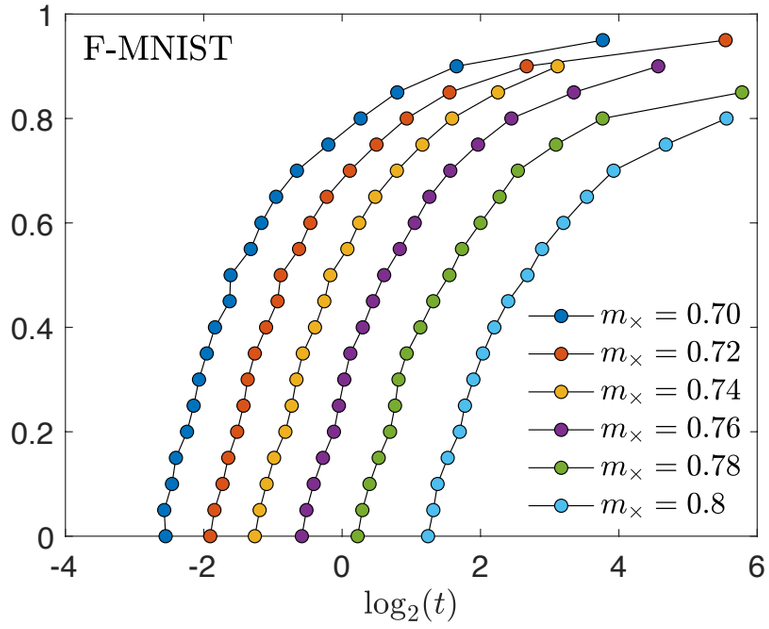


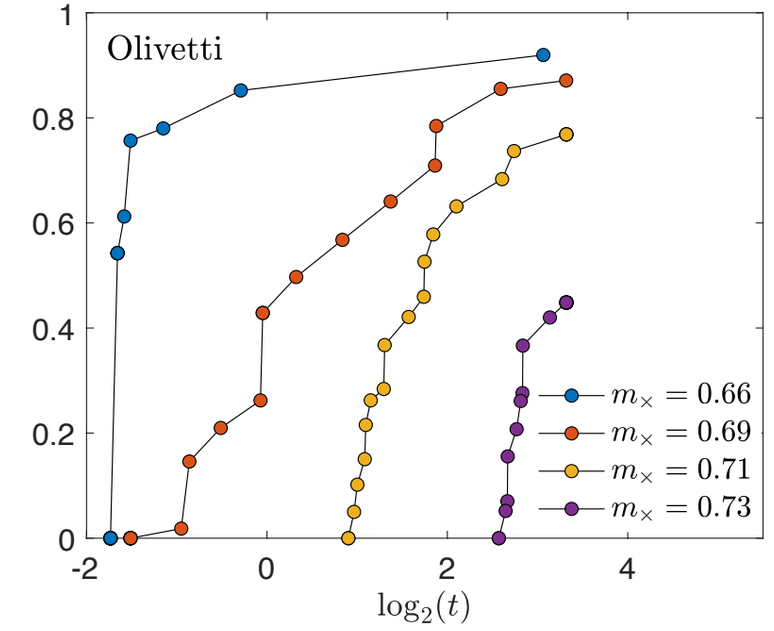


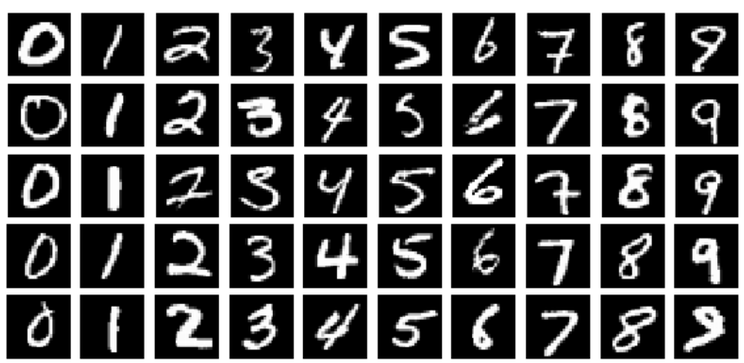


Structured

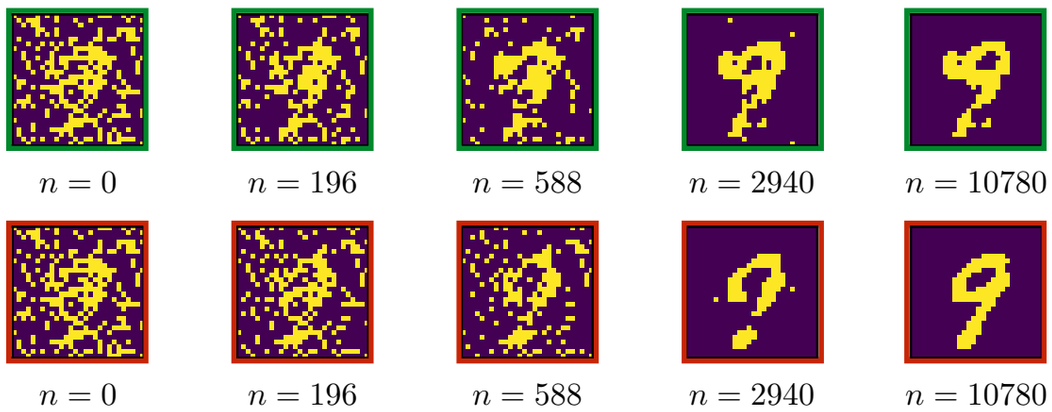






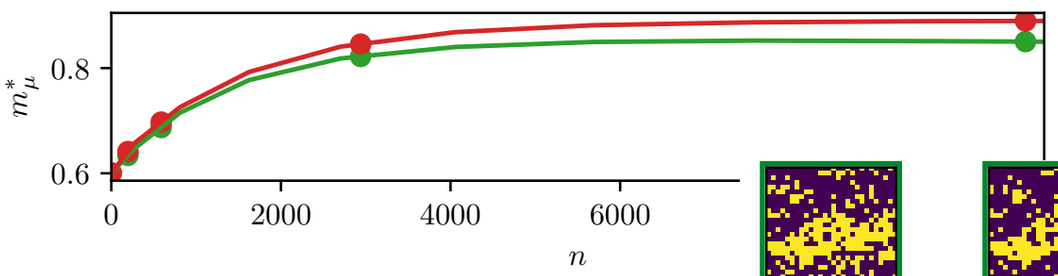


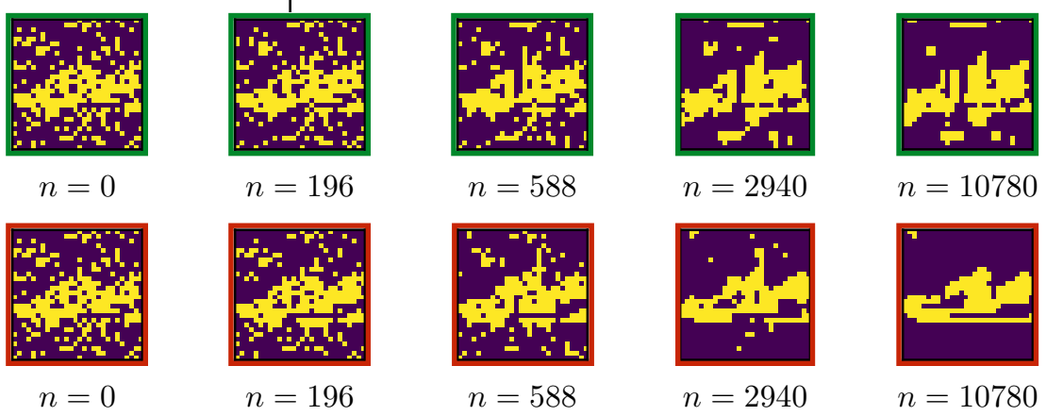
[Aquaro, Alemanno, Kanter, Durante, Barra, Agliari, submitted]



$t_a = 1,\ M_a = 5000$

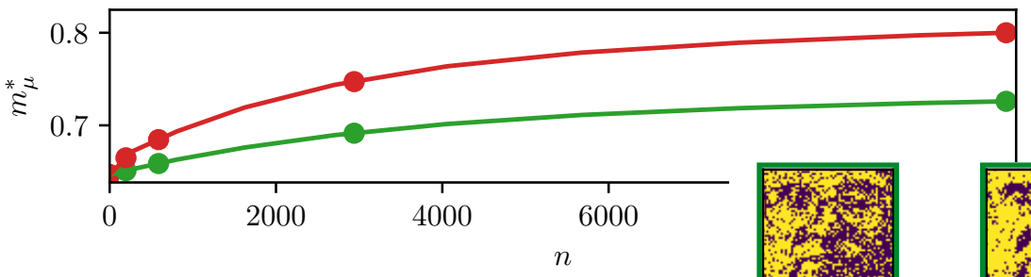
$t_b = 200,\ M_b = 0.65 M_a$

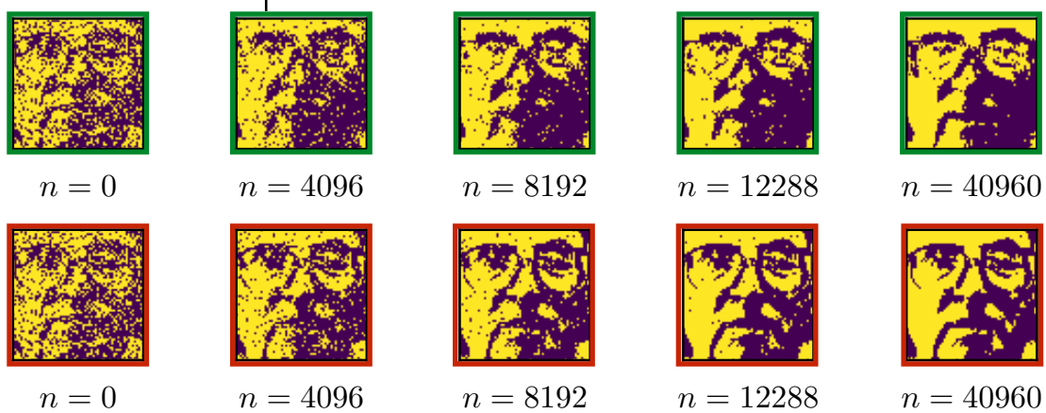




$t_a = 0.25,\ M_a = 6000$

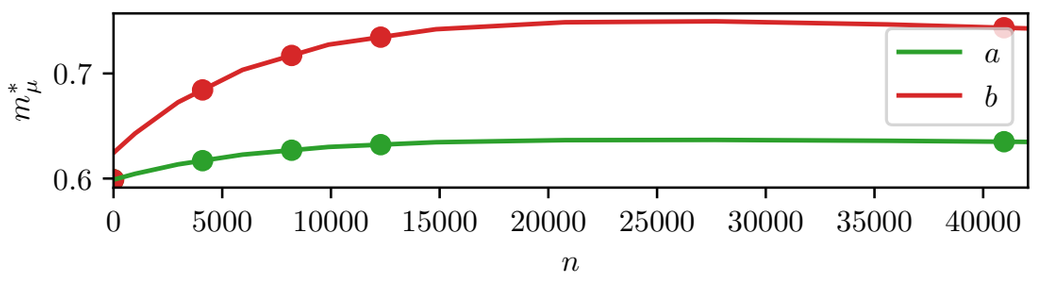
$t_b = 10,\ M_b = 0.65 M_a$





$t_a = 0.25,\ M_a = 320$

$t_b = 100,\ M_b = 0.65 M_a$



[Aquaro, Alemanno, Kanter, Durante, Barra, Agliari, submitted]

Why should we ever wake up?

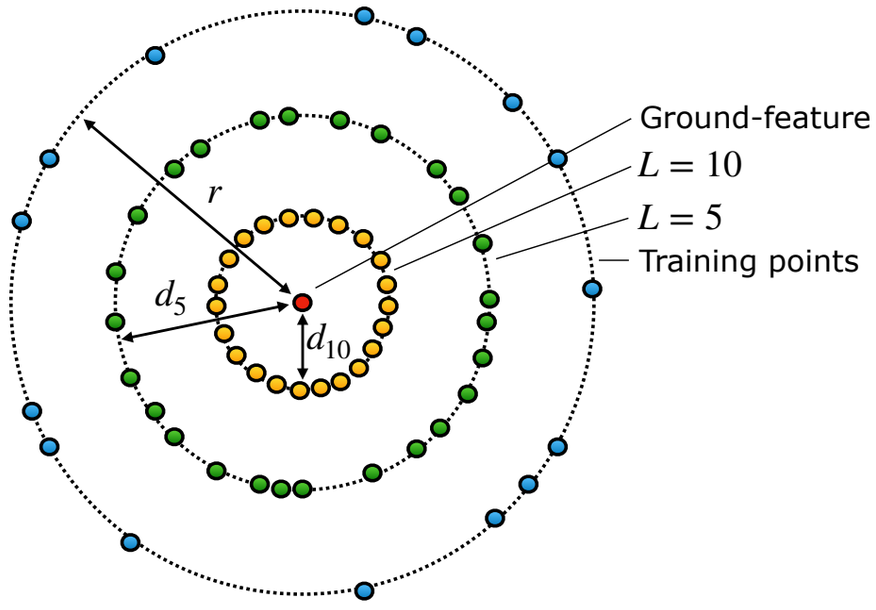
[Kleinfeld, Pendergraft - Biophys. J. ’87]

[van Hemmen, Ioffe, Kühn, Vaas - Physica A ’90]

[Benedetti, Ventura, Marinari, Ruocco, Zamponi - J. Chem. Phys. ’22]

A qualitative picture

$K \times M$ "blurred" examples $\{\boldsymbol{\eta}^{\mu,a}\}_{\mu=1,\dots,K}^{a=1,\dots,M}$ generated from a set of K (unknown) archetypes $\{\boldsymbol{\xi}^{\mu}\}_{\mu=1,\dots,K}$



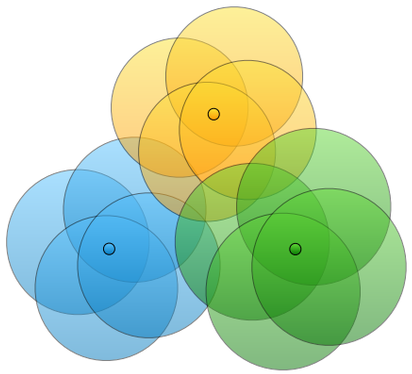
$$\bar{\boldsymbol{\xi}}_L^{\mu} = \operatorname{sgn} \sum_{\ell=1}^{L} \boldsymbol{\xi}^{\mu,a_\ell}$$

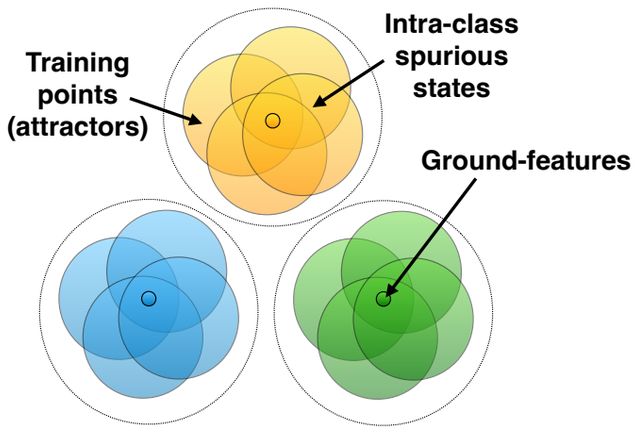
$$r \leq \frac{\bar{\boldsymbol{\xi}}_L^{\mu} \cdot \boldsymbol{\xi}^{\mu}}{N} \underset{L \gg 1}{\rightarrow} 1$$

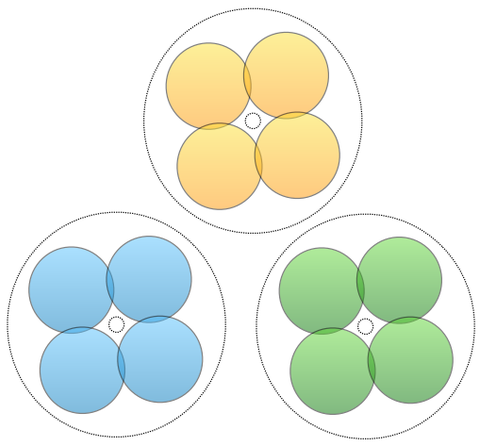
Primary attractors (examples)

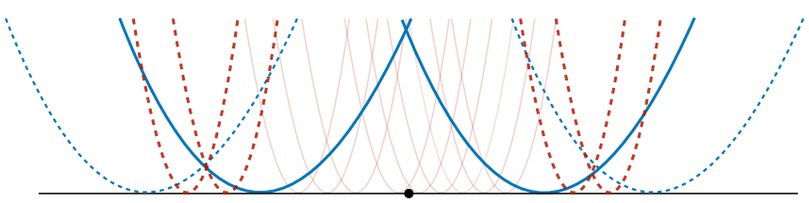
Secondary attractors (spurious states)

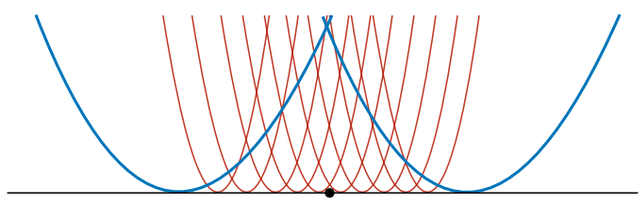
Inter-class spurious attractors

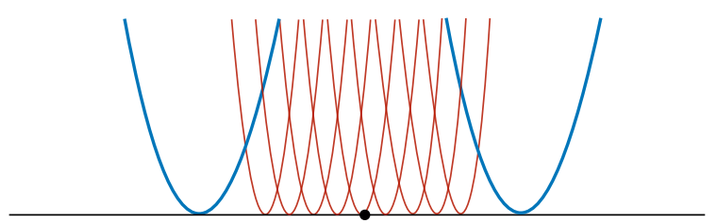












An analytical formulation

Look for a set of couplings $\{J_{ij}\}_{i,j=1,...,N}$ s.t. any pattern $\boldsymbol{\xi}^\mu$ is a fixed point, for $\mu = 1,...,K$

$$\xi_i^\mu = \text{sign}\left(\sum_j J_{ij}\xi_j^\mu + h_i \right)$$

with $T = 0$ and $h_i \in \mathbb{R}$ are external fields.

Stronger condition to guarantee a finite basin of attraction

$$\xi_i^\mu \left(\sum_j J_{ij}\xi_j^\mu + h_i \right) > \kappa \rightarrow \sum_j J_{ij}\xi_j^\mu + h_i = \gamma \xi_i^\mu,\ \gamma \geq \kappa,$$

with κ, γ finite constants.

[Gardner - J Phys A '86; Gardner, Derrida J Phys A '86]

[Personnaz, Guyon, Dreyfus - J Phys Lett '85]

$$\mathscr{L}_{\boldsymbol{\xi}}(\boldsymbol{J},\boldsymbol{h}) = \frac{1}{2N}\sum_{i,\mu}\left(\sum_j \xi_j^\mu J_{ij} + h_i - \gamma\xi_i^\mu \right)^2 + \underbrace{\frac{1}{2N}\sum_{j,\mu}\left(\sum_i \xi_i^\mu J_{ij} + h_j - \gamma\xi_j^\mu \right)^2}_{\text{symmetrize}} + \underbrace{\epsilon_J \sum_{i,j} J_{ij}^2}_{\text{regularize}} + \underbrace{\epsilon_h \sum_i h_i^2}_{\text{regularize}}$$

[Agliari, Alemanno, Aquaro, Fachechi, submitted]

$$\frac{d\boldsymbol{J}}{d\tau} = -\nabla_{\boldsymbol{J}}\mathscr{L}_{\boldsymbol{\xi}}(\boldsymbol{J},\boldsymbol{h})$$

$$\frac{d\boldsymbol{h}}{d\tau} = -\nabla_{\boldsymbol{h}}\mathscr{L}_{\boldsymbol{\xi}}(\boldsymbol{J},\boldsymbol{h})$$

$$\boldsymbol{J}_{\epsilon_J=t^{-1}}(\tau\to\infty) \to \boldsymbol{J}_t^{(DHN)}$$

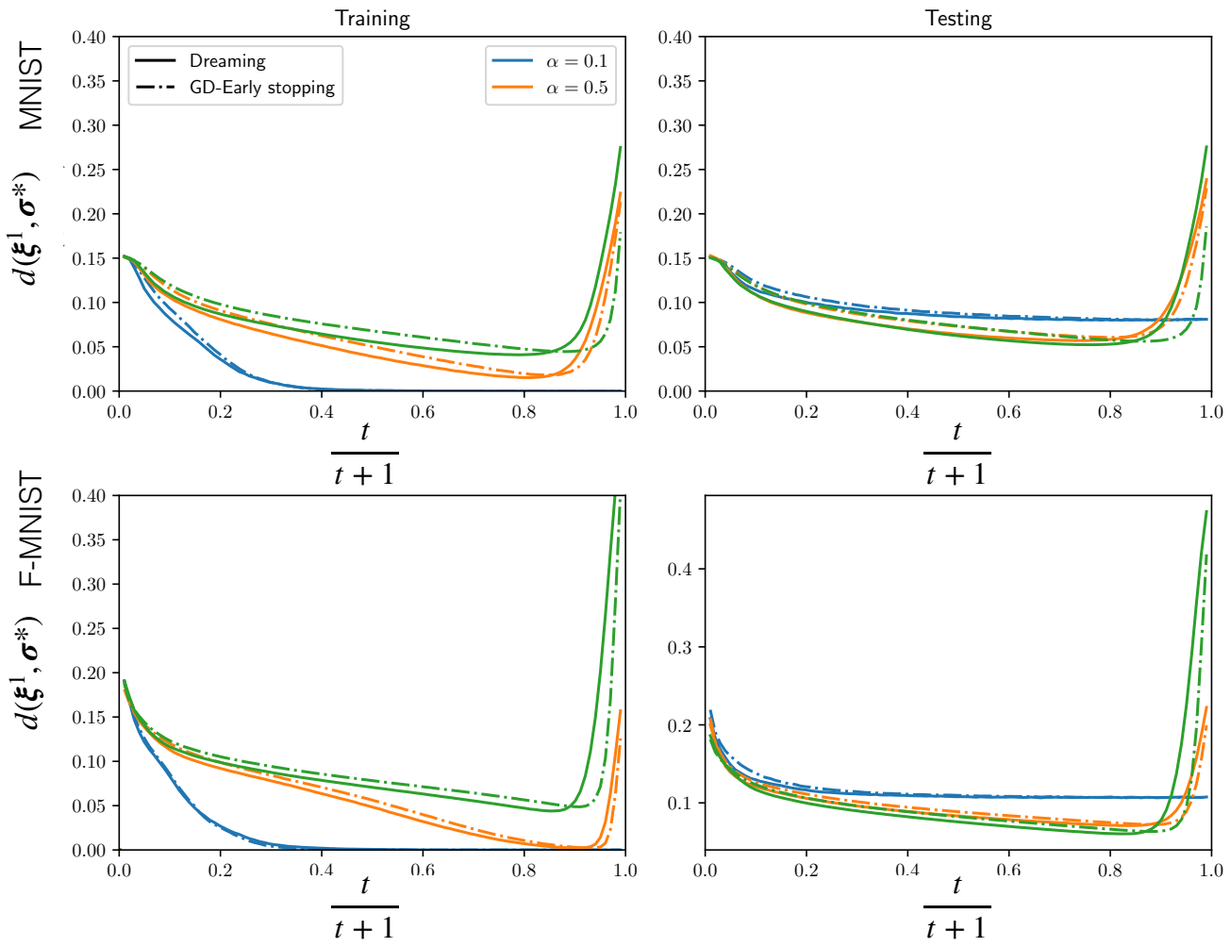
$$\tau^*(\epsilon_J) = \underset{\tau}{\operatorname{argmin}} \left\|\boldsymbol{J}_{\epsilon_J=0,\gamma}(\tau) - \boldsymbol{J}_{\epsilon_J,\gamma}(\tau\to\infty)\right\|_F$$

$$\to \tau^* = A\log(1+tB)+C$$

$$\boldsymbol{J}_{\epsilon_J,\epsilon_h}(\tau\to\infty) = \frac{1}{N}\hat{\boldsymbol{\xi}}\left(\frac{1}{\boldsymbol{C}+\epsilon_J\boldsymbol{I}}\right)\hat{\boldsymbol{\xi}}^T$$

$$\boldsymbol{h}_{\epsilon_J,\epsilon_h}(\tau\to\infty) = \frac{(\boldsymbol{I}-\boldsymbol{J})\bar{\boldsymbol{\xi}}}{1+\epsilon_h}$$

$$\text{with } \bar{\xi}_i := \frac{1}{N}\sum_\mu \xi_i^\mu, \quad \hat{\xi}_i^\mu := \xi_i^\mu - \left(1-\sqrt{\frac{\epsilon_h}{\epsilon_h+1}}\right)\bar{\xi}_i, \quad \boldsymbol{C} = \frac{1}{N}\hat{\boldsymbol{\xi}}^T\hat{\boldsymbol{\xi}}$$



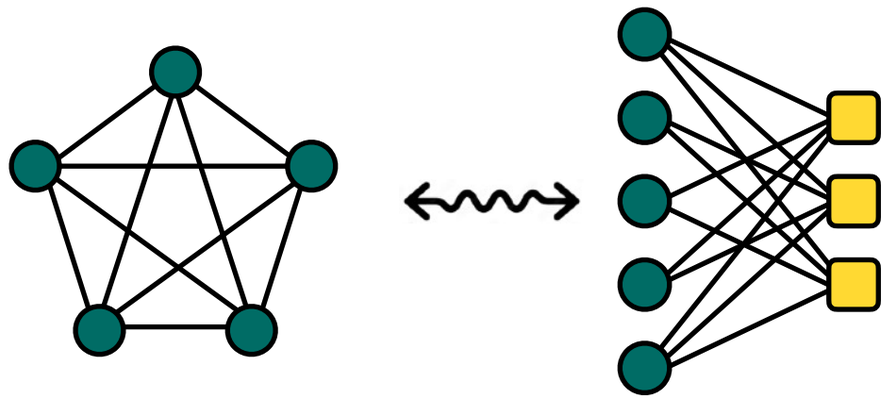
[Girosi, Jones, Poggio - Neur. Com. ’95]

[Zhang, Yu - Ann. Stat. ’05]

Hopfield network / Restricted Boltzmann machines equivalence

$$\boldsymbol{\sigma} \in \{-1, +1\}^N$$

$$\boldsymbol{\xi} \in \{-1, +1\}^{N \times K}$$



$$\boldsymbol{\sigma} \in \{-1, +1\}^N$$

$$z \sim \mathcal{N}(0, \boldsymbol{I}\beta^{-1})$$

$$\boldsymbol{W} = \boldsymbol{\xi}$$

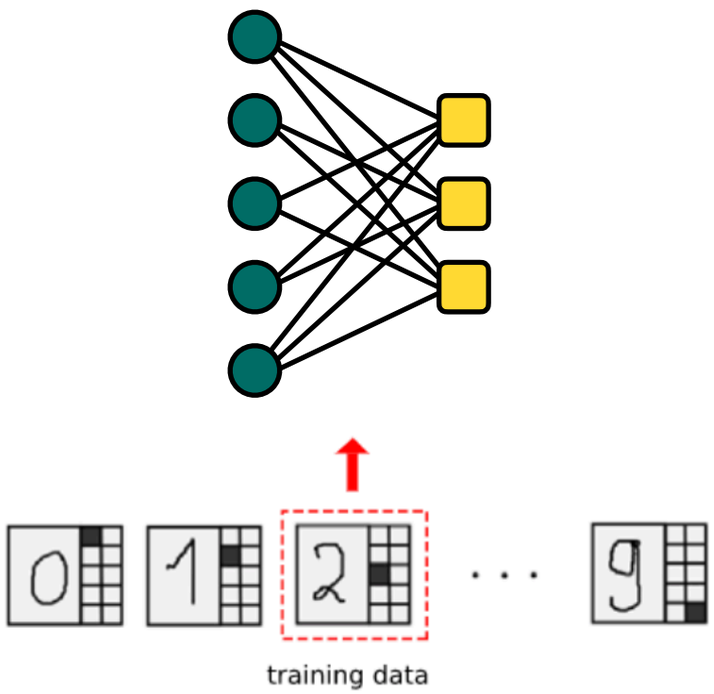
Hubbard-Stratonovich transformation

$$\mathcal{Z}^{(\mathrm{HN})}_{\beta,N,K}(\boldsymbol{\xi}) = \sum_{\boldsymbol{\sigma}} e^{\frac{\beta}{2N}\sum_{i,j,\mu}\xi_i^\mu \xi_j^\mu \sigma_i \sigma_j} = \sum_{\boldsymbol{\sigma}} \int \prod_{\mu} dz_\mu e^{-\frac{\beta z_\mu^2}{2}} e^{\frac{\beta}{\sqrt{N}}\sum_{i,\mu}\sigma_i \xi_i^\mu z_\mu} = \mathcal{Z}^{(\mathrm{RBM})}_{\beta,N,K}(\boldsymbol{\xi})$$

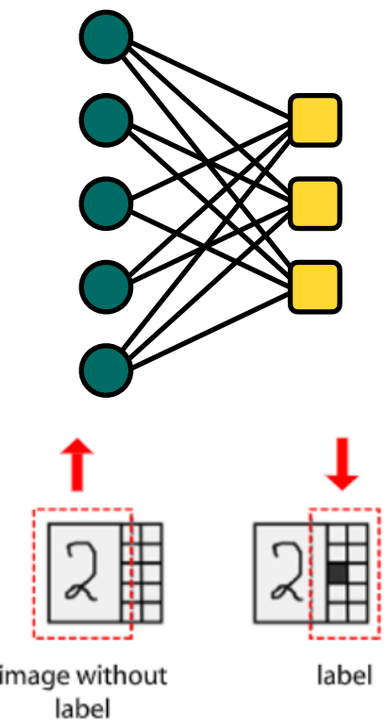
Gaussian integration

[Agliari, Barra, Galluzzi, Guerra, Moauro - Phys. Rev. Lett ’12]
[Barra, Bernacchia, Santucci, Contucci - Neur. Netw. ’12]
[Mézard - Phys. Rev. E ’17]
[Tubiana, Monasson - Rhys. Rev. Lett. ’17]
[Barra, Genovese, Sollich, Tantari - Rhys. Rev. E ’18]
[Decelle, Fissore, Furtlehner - J. Stat. Phys. ‘18]
[Smart, Zilman - ICLR ’21]

learning with labels



classification



$$\mathcal{S} = \{\boldsymbol{\eta}^{\mu,a}\}_{a=1,\dots M}^{\mu=1,\dots,K},\ \boldsymbol{\eta}^{\mu,a} \to z_\nu^{(\mu)} = \delta_{\mu,\nu}$$

$$q(\boldsymbol{\sigma}, \boldsymbol{x}) = \sum_{\mu,a} \delta_{\boldsymbol{\eta}^{\mu,a},\boldsymbol{\sigma}} \delta_{\boldsymbol{z}^{(\mu)},\boldsymbol{z}}$$

$$\rho(\boldsymbol{\sigma}, \boldsymbol{z}; \boldsymbol{W}) \propto \exp\Big[-\frac{\beta z_\mu^2}{2} + \frac{\beta}{\sqrt{N}} \sum_{i,\mu} \sigma_i W_i^\mu z_\mu \Big]$$

$$\Delta W_i^\mu \propto \Big(\langle \sigma_i z_\mu \rangle_{\text{clamped}} - \langle \sigma_i z_\mu \rangle_{\text{free}} \Big), \quad \forall (i,\mu) \in (N \times K)$$

$\boldsymbol{W} = \bar{\boldsymbol{\eta}}$ For structureless datasets: fixed point

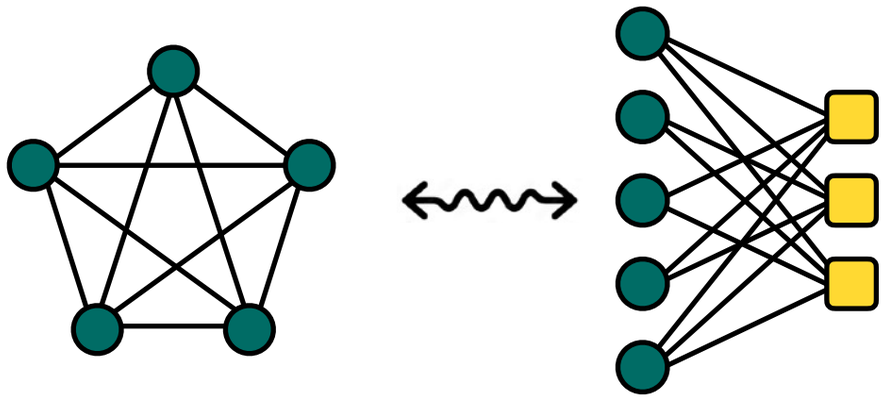
For structured datasets: effective initialisations

[Agliari, Leonelli, Marullo - Appl. Math. Comput. ’22]

Hopfield network / Restricted Boltzmann machines equivalence

$$\boldsymbol{\sigma} \in \{-1, +1\}^N$$

$$\boldsymbol{\xi} \in \{-1, +1\}^{N \times K}$$



$$\boldsymbol{\sigma} \in \{-1, +1\}^N$$

$$\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}\beta^{-1})$$

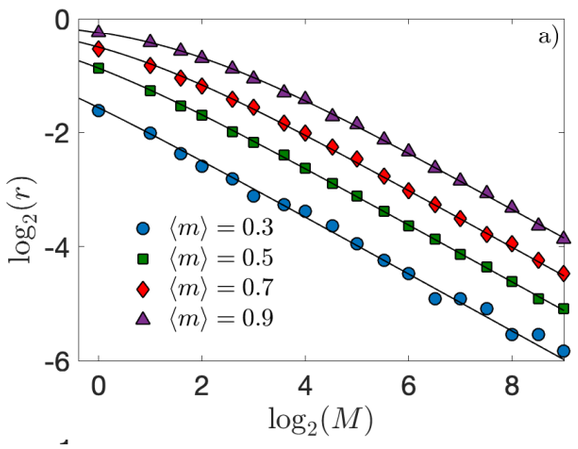
$$\boldsymbol{W} = \boldsymbol{\xi}$$

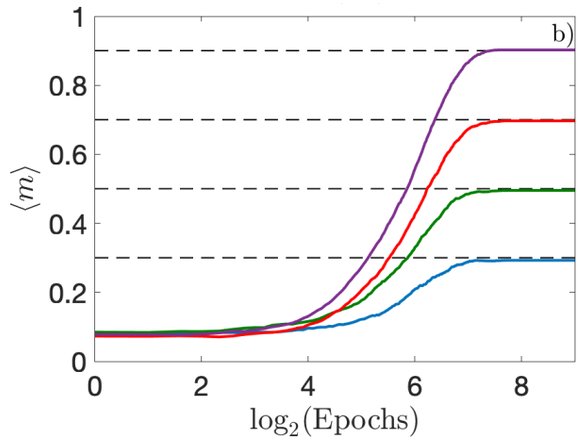
$$\{\boldsymbol{\eta}^{\mu,a}\}_{a=1,\dots,M}^{\mu=1,\dots,K} \in \{-1, +1\}^{N \times K}$$

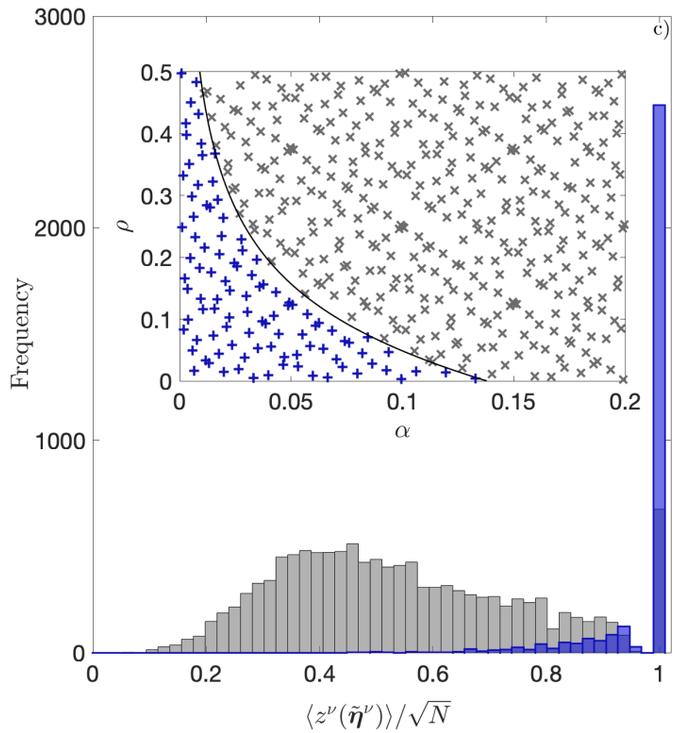
$$\bar{\boldsymbol{\eta}} = \frac{1}{M} \sum_a \boldsymbol{\eta}^a$$

$$\boldsymbol{W} = \bar{\boldsymbol{\eta}}$$

$$\mathcal{Z}^{(\mathrm{HN})}_{\beta,N,K}(\boldsymbol{\eta}) = \sum_{\boldsymbol{\sigma}} e^{\frac{\beta}{2N} \sum_{i,j,\mu} \bar{\eta}_i^\mu \bar{\eta}_j^\mu \sigma_i \sigma_j} = \sum_{\boldsymbol{\sigma}} \int \prod_\mu dz_\mu e^{-\frac{\beta z_\mu^2}{2}} e^{\frac{\beta}{\sqrt{N}} \sum_{i,\mu} \sigma_i \bar{\eta}_i^\mu z_\mu} = \mathcal{Z}^{(\mathrm{RBM})}_{\beta,N,K}(\boldsymbol{\eta})$$







[Agliari, Alemanno, Aquaro, Barra, Kanter - Europhys. Lett. Perspective '23]

“Dreaming” Boltzmann machine

BM with $\sigma_i \in \{-1,+1\}$, $z_\mu \sim \mathcal{N}[0,T(1+t)]$

$$\mathscr{H}^{(\mathrm{DBM})}_{N,K,\boldsymbol{W},t}(\boldsymbol{\sigma},\boldsymbol{z}) = -\frac{1}{\sqrt{N}}\sum_{i,\mu}^{N,K} W_{i\mu}\sigma_i z_\mu + \frac{t}{1+t}\sum_{\mu<\nu}^{K,K} C_{\mu\nu} z_\mu z_\nu,$$

equivalent to “Dreaming” Hopfield network as $W_{i\mu} = \xi^\mu_i$ for any i,μ

$$\mathscr{Z}^{(\mathrm{DBM})}_{\beta,N,K,\boldsymbol{\xi},t} = \sum_{\boldsymbol{\sigma}}\int\prod_\mu dz_\mu e^{-\frac{\beta z_\mu^2}{2(1+t)}} e^{-\beta\mathscr{H}^{(\mathrm{DBM})}_{N,K,\boldsymbol{\xi},t}(\boldsymbol{\sigma},\boldsymbol{z})} = \sum_{\boldsymbol{\sigma}} e^{-\beta\mathscr{H}^{(\mathrm{DHN})}_{N,K,\boldsymbol{\xi},t}(\boldsymbol{\sigma})} = \mathscr{Z}^{(\mathrm{DHN})}_{\beta,N,K,\boldsymbol{\xi},t}$$

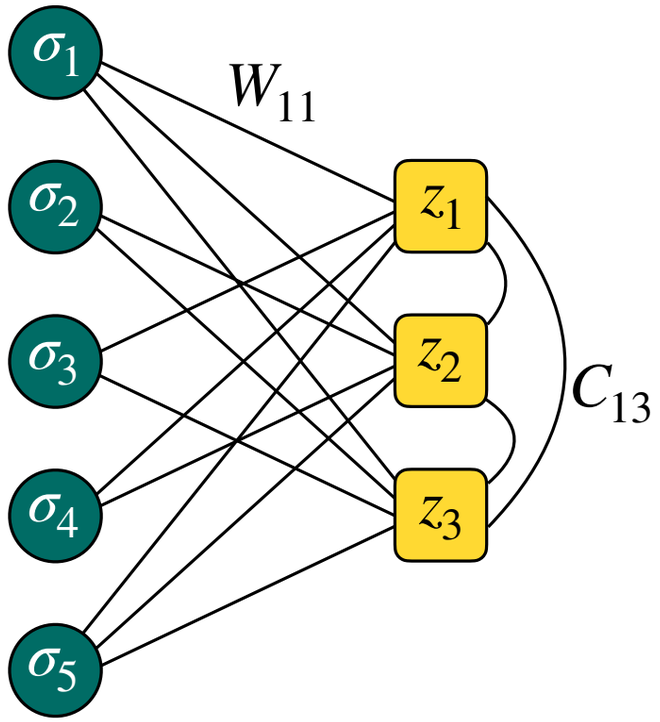
Gradient descent over KL divergence

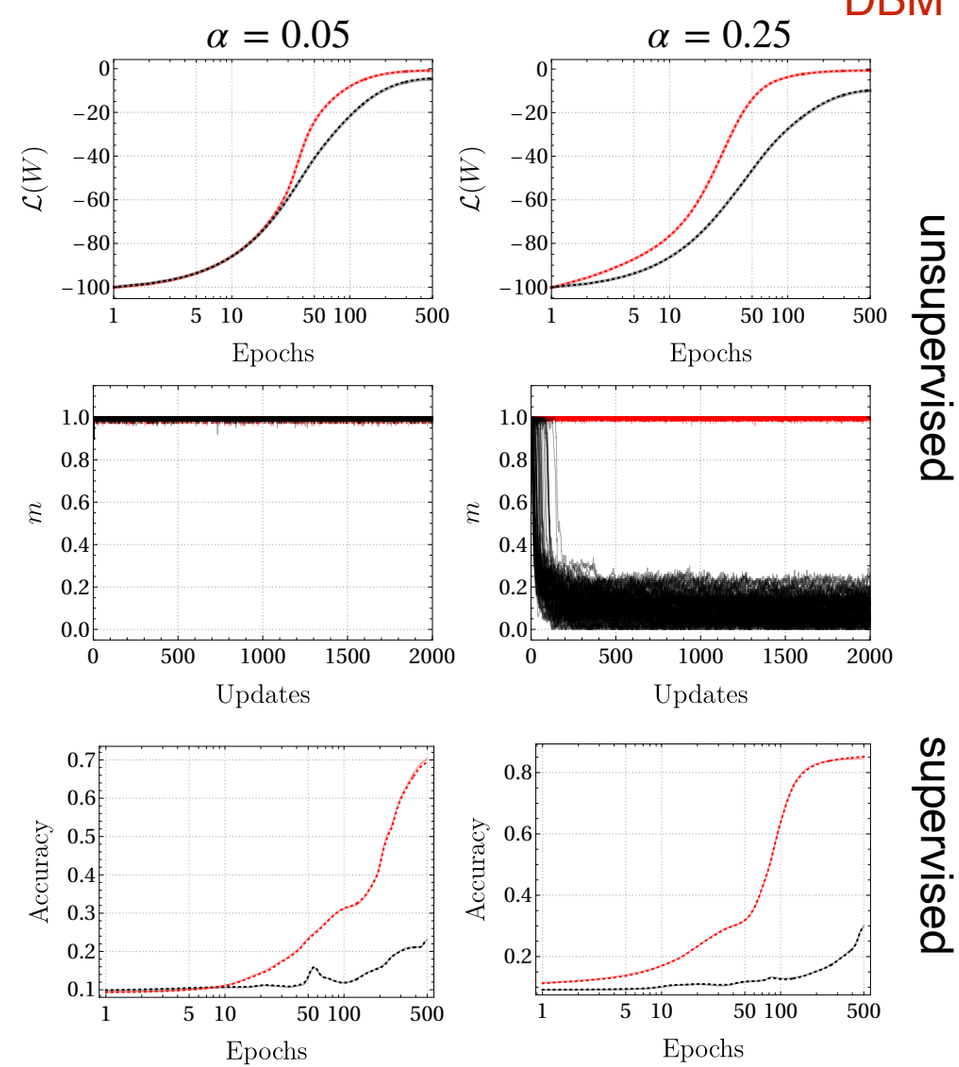
$$\Delta W_{j\rho} = \epsilon\beta\left[\langle\sigma_j z_\rho\rangle_+ - \langle\sigma_j z_\rho\rangle_- - \frac{N}{2}\frac{t}{1+t}\sum_{\mu\nu}\frac{\partial C_{\mu\nu}}{\partial W_{j\rho}}\left(\langle z_\mu z_\nu\rangle_+ - \langle z_\mu z_\nu\rangle_-\right)\right]$$

$$\Delta t = \epsilon\beta\frac{N}{2(1+t)^2}\sum_{\mu\nu}\left(1-\boldsymbol{C}\right)_{\mu\nu}\left(\langle z_\mu z_\nu\rangle_+ - \langle z_\mu z_\nu\rangle_-\right)$$

$\langle\cdot\rangle_+$ clamped average
$\langle\cdot\rangle_-$ free average

[Agliari et al. Neur. Netw. ’21, IEEE Trans Neural Netw Learn Syst. ’22]





THANKS

Regularisation and Early-Stopping

$$\frac{d\boldsymbol{J}}{d\tau} = -\nabla_{\boldsymbol{J}}\mathscr{L}_{\boldsymbol{\xi}}(\boldsymbol{J},\boldsymbol{h})$$

$$\frac{d\boldsymbol{h}}{d\tau} = -\nabla_{\boldsymbol{h}}\mathscr{L}_{\boldsymbol{\xi}}(\boldsymbol{J},\boldsymbol{h})$$

Natural basis: eigenvectors of $\hat{\boldsymbol{C}} = \hat{\boldsymbol{\xi}}\hat{\boldsymbol{\xi}}^T \in \mathbb{R}^{N\times N}$

$$J_{ab}(\tau) \underset{\tau\gg 1}{\sim} \begin{cases} 0 & \text{if} \quad a \neq b \\ \dfrac{\lambda_a(1+\epsilon_J)}{\lambda_a+\epsilon_J} & \text{if} \quad a = b \end{cases}$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$\epsilon_J \to 0 \ (t \to \infty)$

$$\boldsymbol{J}^{(KS)} = \boldsymbol{\xi}^T\boldsymbol{C}^{-1}\boldsymbol{\xi}$$

$$\begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

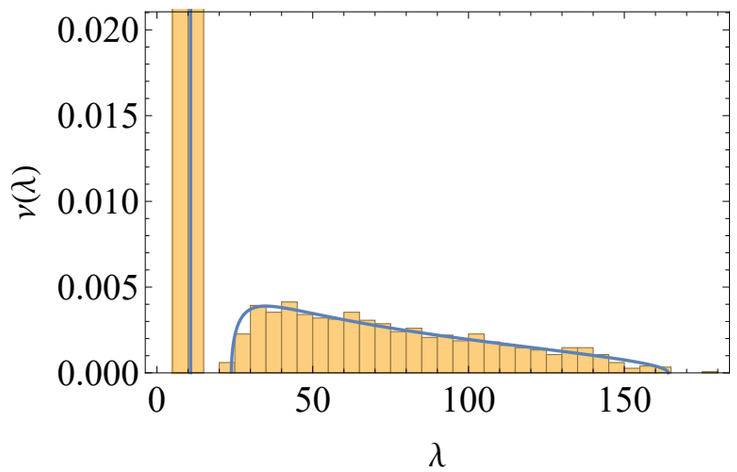
$\epsilon_J \to \infty \ (t \to 0)$

$$\boldsymbol{J}^{(Hebb)} = \boldsymbol{\xi}^T\boldsymbol{\xi}$$

$$J_{aa}(\tau) \underset{\epsilon_J=0}{=} 1-(1-\lambda_a)e^{-2\tau\lambda_a}$$

$$\lambda_1 \leq \lambda_2 \ldots \leq \underbrace{\lambda_{N-\ell+1} \leq \ldots \leq \lambda_N}_{\ell \text{ fastest}}$$

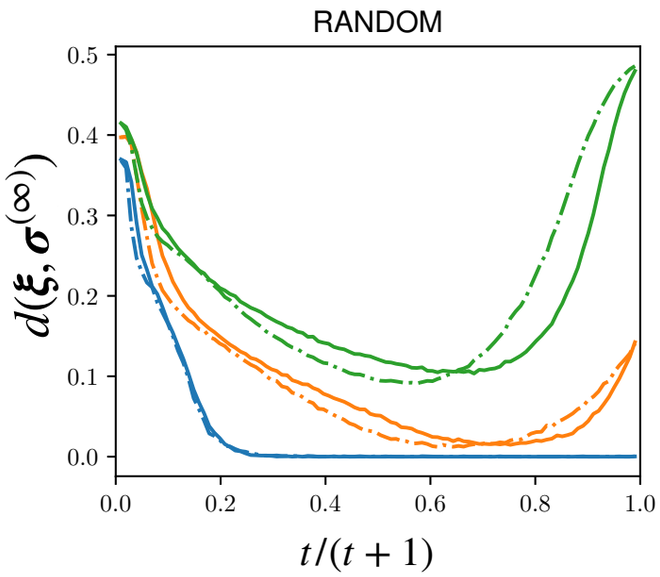
Finite stopping-time τ^* corresponding to λ^*

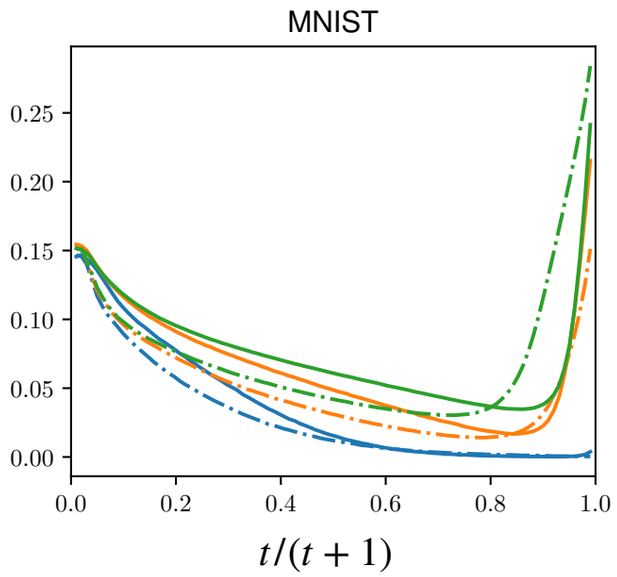


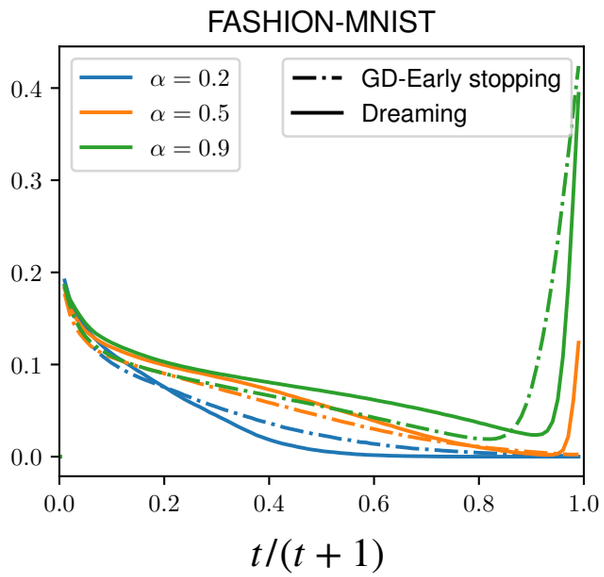
$$\lambda_{N-K+1} = \frac{MK}{N}[r^2(1-\alpha^{-1/2})^2 + (1-r^2)]$$

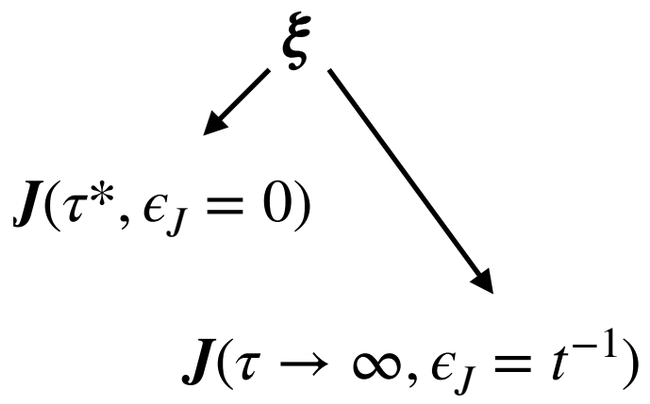
with K, M, r determinable by $\hat{\boldsymbol{C}}$'s moments

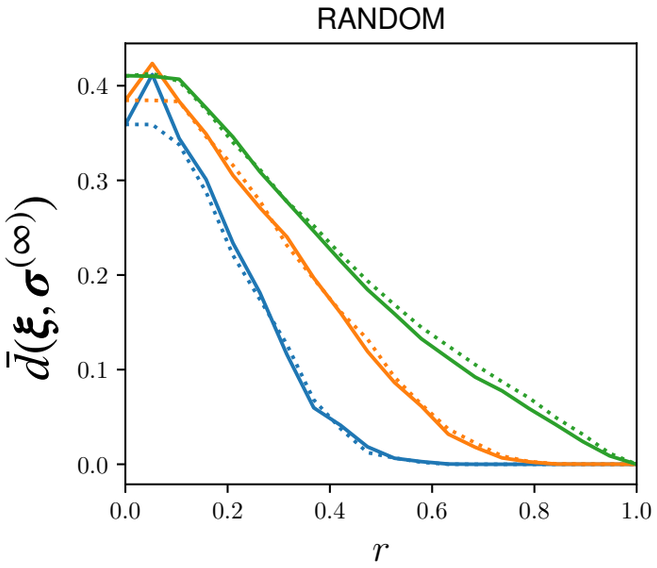
Numerical experiments

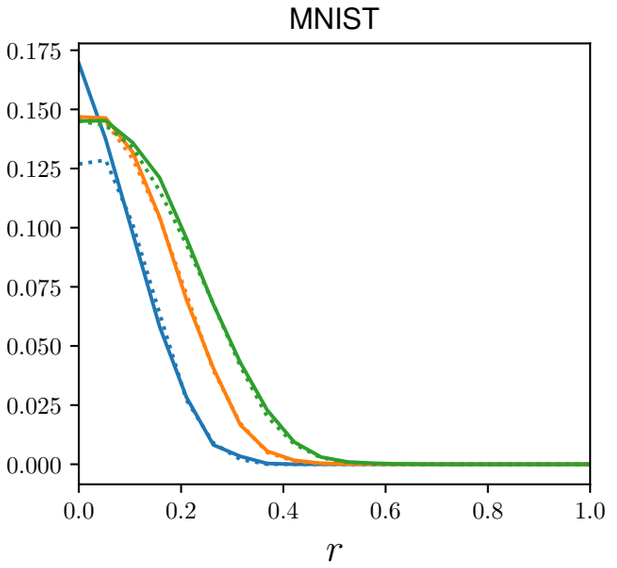


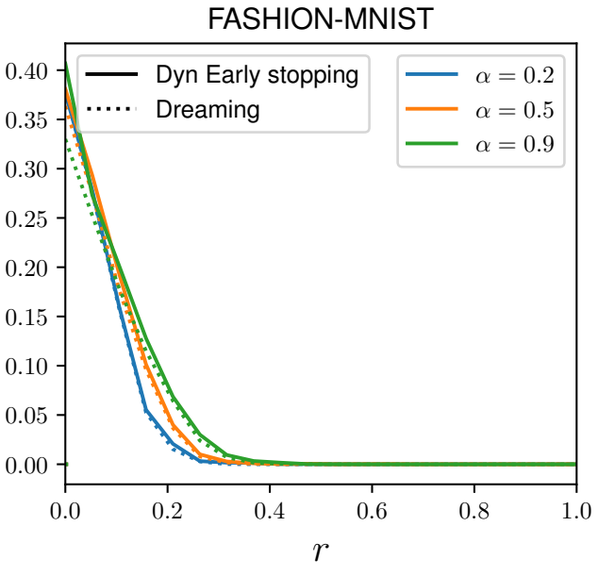


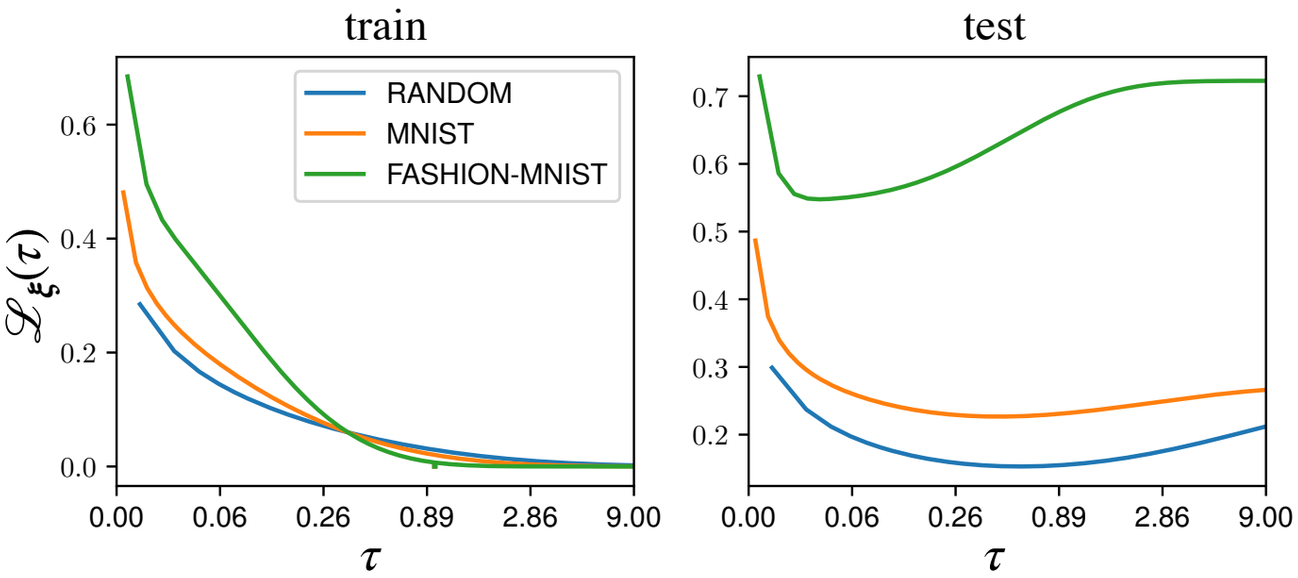












$\epsilon_J = 0,\ K/N = 0.25$

$r = 0.8$

Numerical experiments

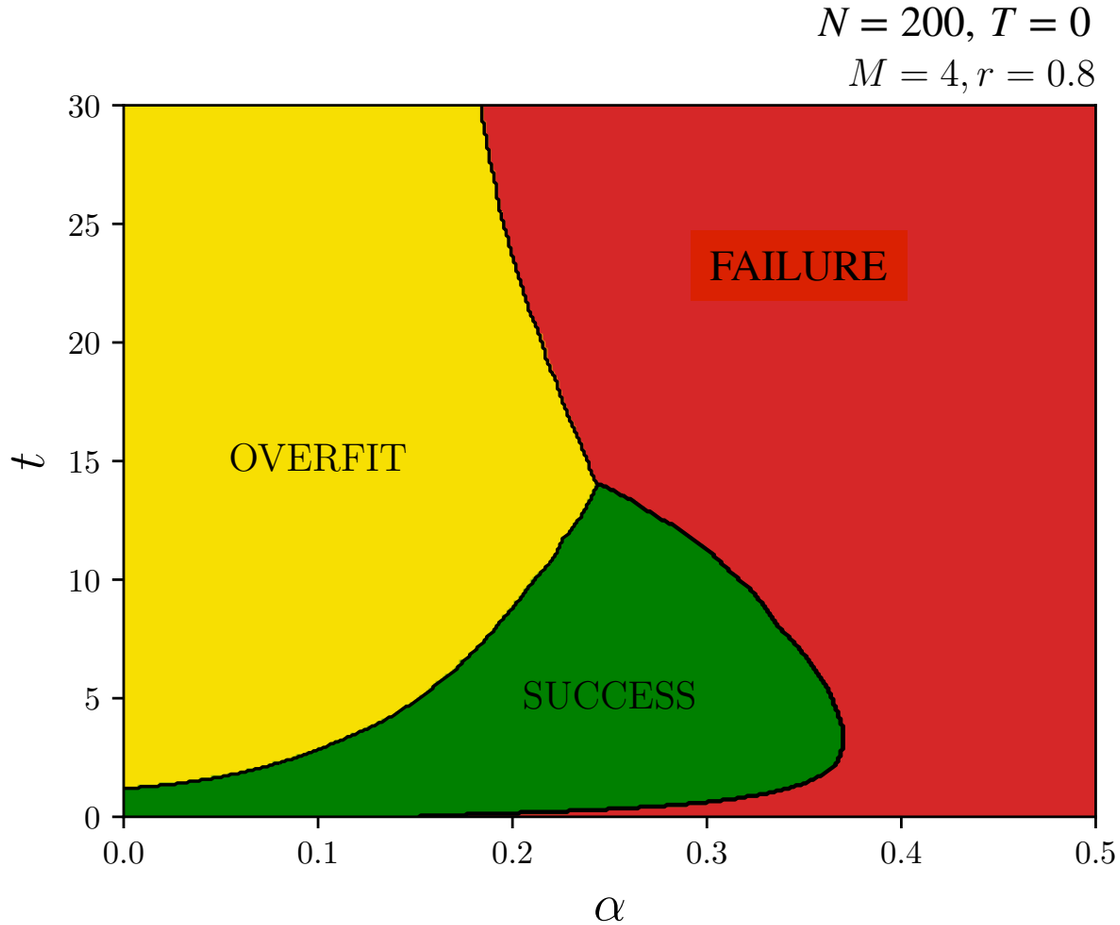
Dataset

$K \times M$ "blurred" examples $\{\boldsymbol{\eta}^{\mu,a}\}_{\mu=1,\dots,K}^{a=1,\dots,M}$

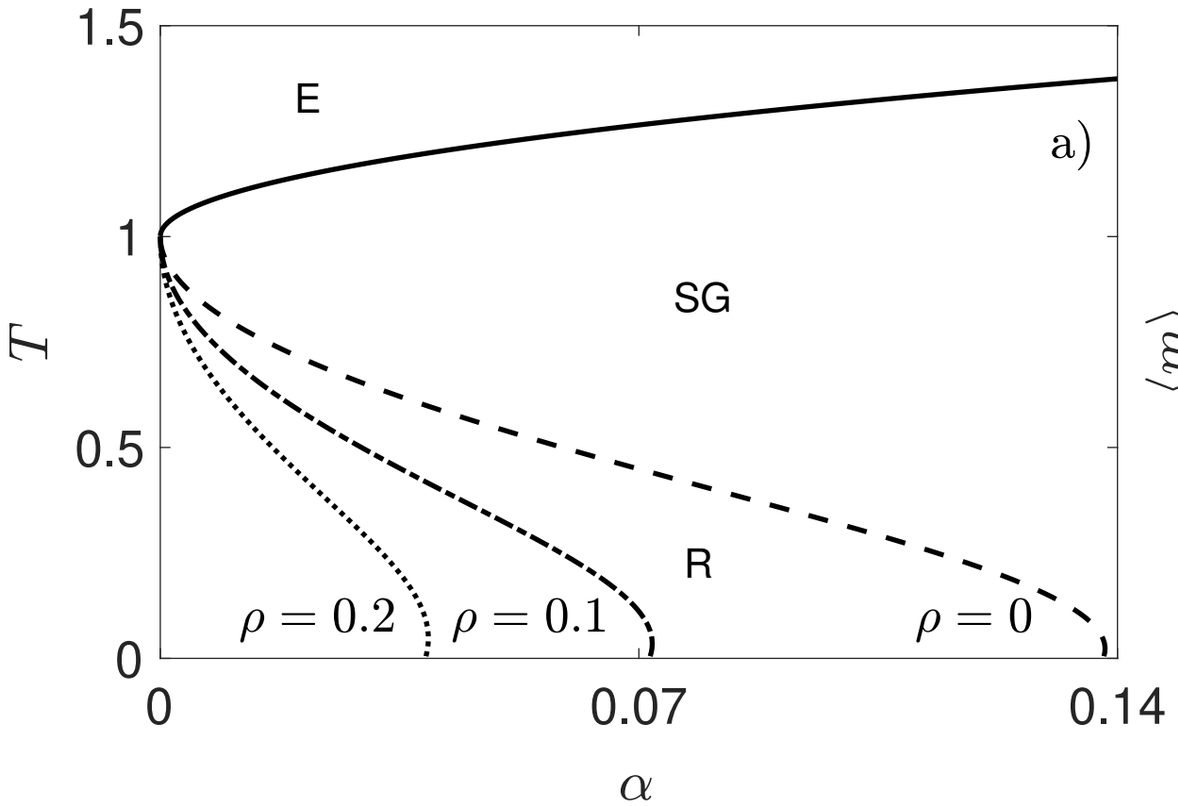
Success: `if` $m_{\mu,a} > r$ `and` $m_{\mu} > n_{\mu,a}$

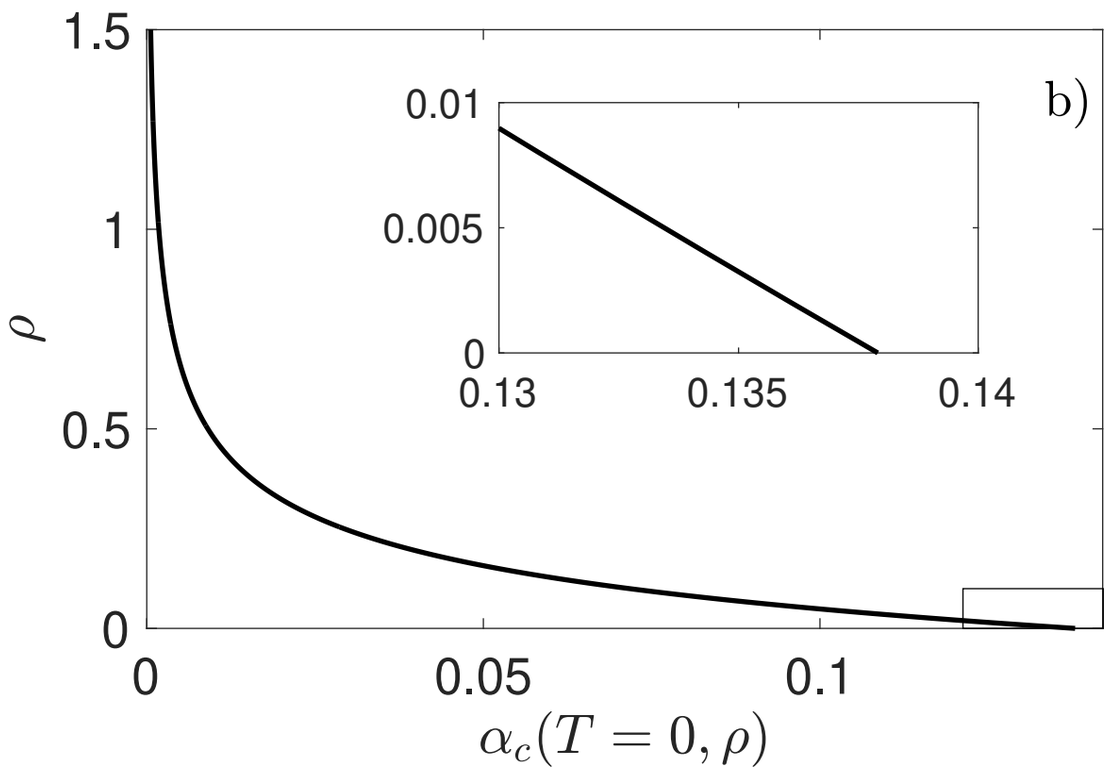
Overfit: `elseif` $n_{\mu,a} > r^2$ `and` $n_{\mu,a} > m_{\mu}$

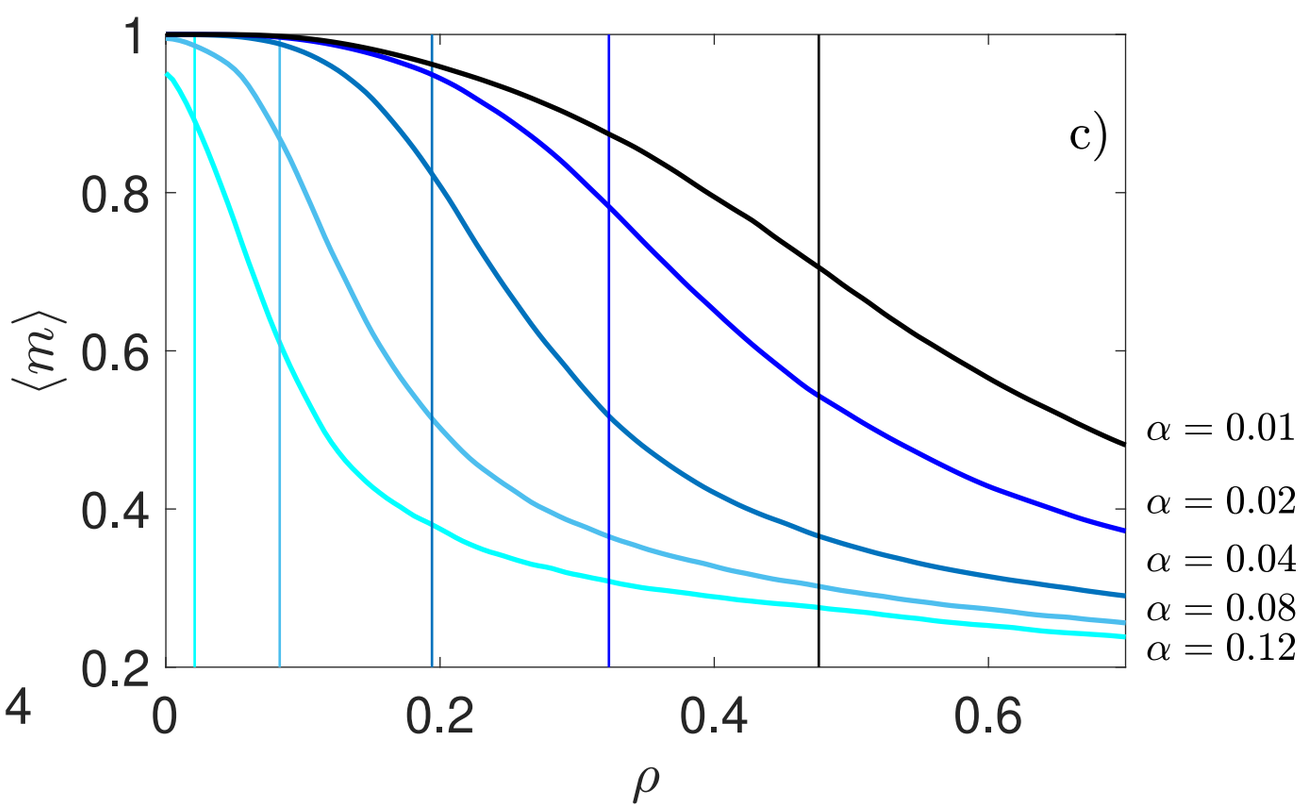
Failure: `else`

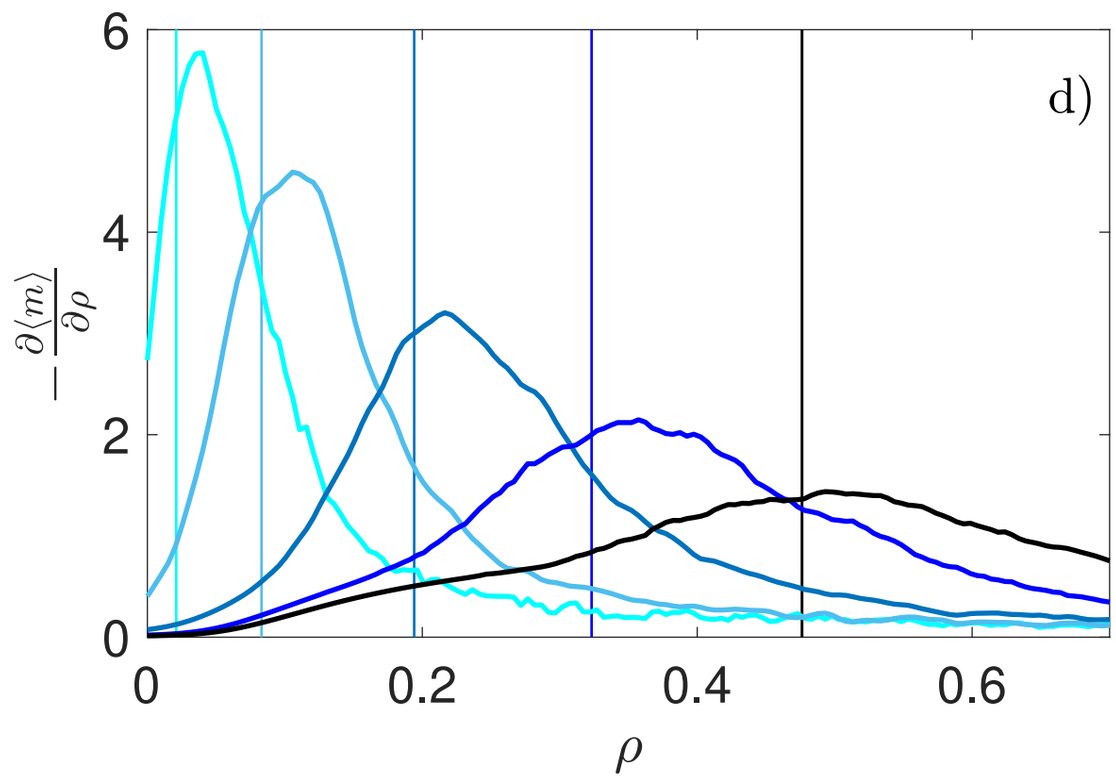


[Agliari, Alemanno, Aquaro, Barra, Fachechi, in preparation]


E
SG
R
a)
$\rho = 0.2$
$\rho = 0.1$
$\rho = 0$
T
α
1.5
1
0.5
0
0
0.07
0.14


b)
ρ
$\alpha_c(T = 0, \rho)$
0.01
0.005
0
0.13
0.135
0.14
1.5
1
0.5
0
0
0.05
0.1


c)
$\langle m \rangle$
ρ
$\alpha = 0.01$
$\alpha = 0.02$
$\alpha = 0.04$
$\alpha = 0.08$
$\alpha = 0.12$
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6


d)
$-\frac{\partial \langle m \rangle}{\partial \rho}$
ρ
6
4
2
0
0
0.2
0.4
0.6

Learning Hebb's rule

Hopfield network and blurred dataset

Does Hebbian learning work also for imperfect (unlabelled) training data?

K "archetypes" $\{\boldsymbol{\xi}^{\mu}\}_{\mu=1,\dots,K}$
$K \times M$ "blurred" examples $\{\boldsymbol{\eta}^{\mu,a}\}_{\mu=1,\dots,K}^{a=1,\dots,M}$

$$\eta_i^{\mu,a} = \xi_i^{\mu}\chi_i^{\mu,a} \qquad \text{with } \mathscr{P}(\chi_i^{\mu,a} = +1) = 1 - \mathscr{P}(\chi_i^{\mu,a} = -1) = p \in [1/2,1]$$

Hebb's rule $$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{K}\sum_{a=1}^{M}\eta_i^{\mu,a}\eta_j^{\mu,a}$$

$$\mathscr{H}_{N,K,M}(\boldsymbol{\sigma};\boldsymbol{\chi},\boldsymbol{\xi}) = -\frac{1}{2N}\sum_{a=1}^{M}\sum_{\mu=1}^{K}\Big(\sum_{i=1}^{N}\xi_i^{\mu}\chi_i^{\mu,a}\sigma_i\Big)^2 .$$

Additional parameters

M	sample size
$r := \langle\chi\rangle = (2p-1)$	sample quality

Stat-mech solution

$$\mathscr{H}_{N,K,M}(\boldsymbol{\sigma};\boldsymbol{\chi},\boldsymbol{\xi}) = -\frac{1}{2N}\sum_{a=1}^{M}\sum_{\mu=1}^{K}\Big(\sum_{i=1}^{N}\xi_i^{\mu}\chi_i^{\mu,a}\sigma_i\Big)^2 .$$

Def. Intensive quenched free-energy

$$f_{N,K,M}(\beta) := -\frac{1}{N\beta}\mathbb{E}\ln Z_{N,K,M}(\beta;\boldsymbol{\chi},\boldsymbol{\xi}),$$

where $\mathbb{E} := \mathbb{E}_{\chi}\mathbb{E}_{\xi}$ and $Z_{N,K,M}(\beta;\boldsymbol{\chi},\boldsymbol{\xi}) = \sum_{\sigma}^{2^N}\exp\Big[\frac{\beta}{2N}\sum_{a=1}^{M}\sum_{\mu=1}^{K}\Big(\sum_{i=1}^{N}\xi_i^{\mu}\chi_i^{\mu,a}\sigma_i\Big)^2\Big]$.

Def. Mattis magnetizations

$$n_{\mu,a} := \frac{1}{N}\sum_{i=1}^{N}\xi_i^{\mu}\chi_i^{\mu,a}\sigma_i,\quad \mu = 1,...,K,\ a = 1,...,M; \qquad\qquad m_{\mu} := \frac{1}{N}\sum_{i=1}^{N}\xi_i^{\mu}\sigma_i,\quad \mu = 1,...,K\,.$$

Def. Two-replica overlap

$$q_{12} := \frac{1}{N}\sum_{i=1}^{N}\sigma_i^{(1)}\sigma_i^{(2)} \qquad\qquad p_{12} := \frac{1}{K}\sum_{\mu=1}^{K}z_{\mu}^{(1)}z_{\mu}^{(2)}$$

Hp. Replica symmetry & Thermodynamic limit

RS $n_{1,a}, q_{12}, p_{12}$ self-average at $\bar{n}, \bar{q}, \bar{p}$, respectively

M fixed, finite; $K, N \to \infty$ s.t. $\alpha := \lim_{N\to\infty} K/N$ finite

As $M \gg 1$, recalling $r = (2p-1)$

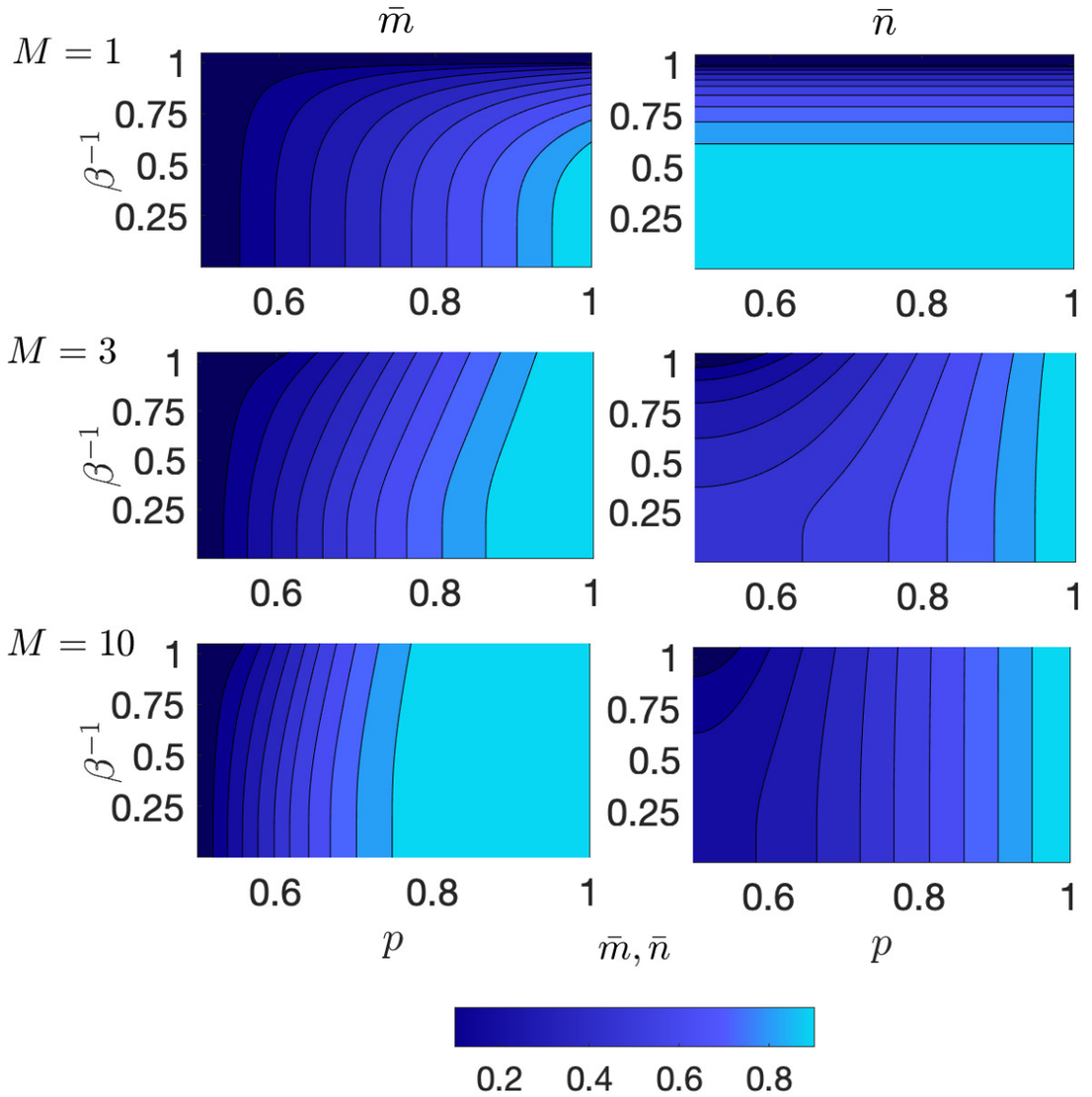
$$\bar{n} = \frac{\bar{m} r}{1 - \beta(1-\bar{q})(1-r^2)}$$

Being $Z \sim \mathcal{N}(0,1)$, under noise rescaling $\beta \to \dfrac{\beta}{r^2 + \beta(1-\bar{q})\left(1-r^2\right)}$

$$\bar{m} = \mathbb{E}_Z \tanh\left[\beta \bar{m} M + Z\beta\sqrt{M\frac{1-r^2}{r^2}\bar{m}^2 + \frac{\alpha\bar{p}}{r^4\beta}M}\right]$$

$$\bar{q} = \mathbb{E}_Z \tanh^2\left[\beta \bar{m} M + Z\beta\sqrt{M\frac{1-r^2}{r^2}\bar{m}^2 + \frac{\alpha\bar{p}}{r^4\beta}M}\right]$$

$$\bar{p} = \frac{\beta\bar{q}}{[1-\beta(1-\bar{q})]^2}$$



$$\bar{n} = \frac{\bar{m} r}{1 - \beta(1 - \bar{q})(1 - r^2)}$$

$$\bar{m} = \mathbb{E}_Z \tanh \left[\beta \bar{m} M + Z\beta \sqrt{M \frac{1 - r^2}{r^2} \bar{m}^2 + \frac{\alpha \bar{p}}{r^4 \beta} M} \right]$$

- Archetype stability

signal carried by $\bar{m}$ and *two* sources of (slow) noise

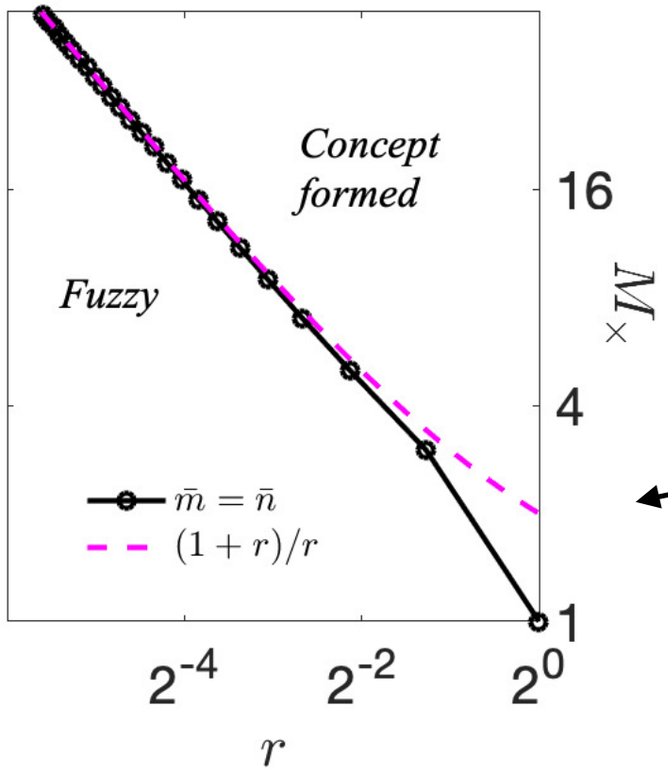
i. **pattern interference**

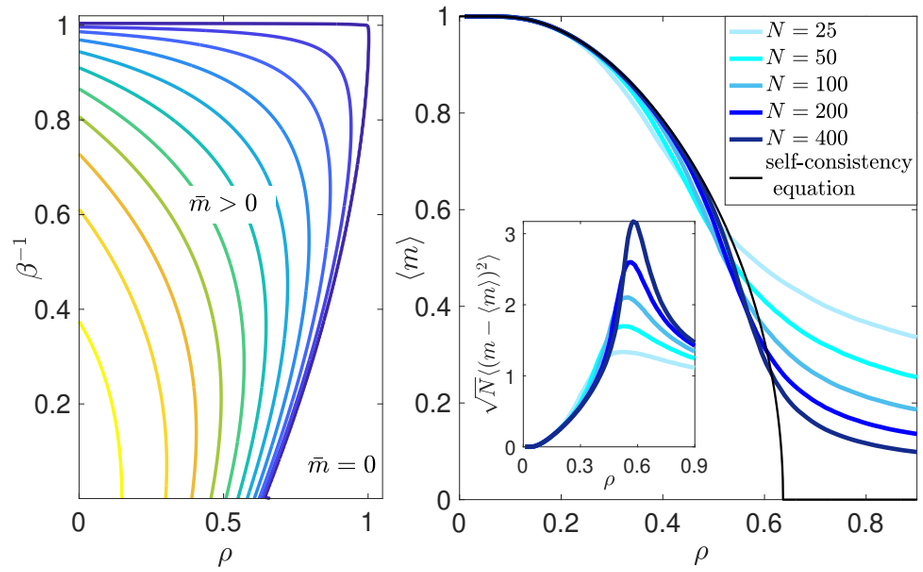
ii. **example interference**

$$\rightarrow \quad M > M_c \sim r^{-4}$$

- Archetype vs Example retrieval

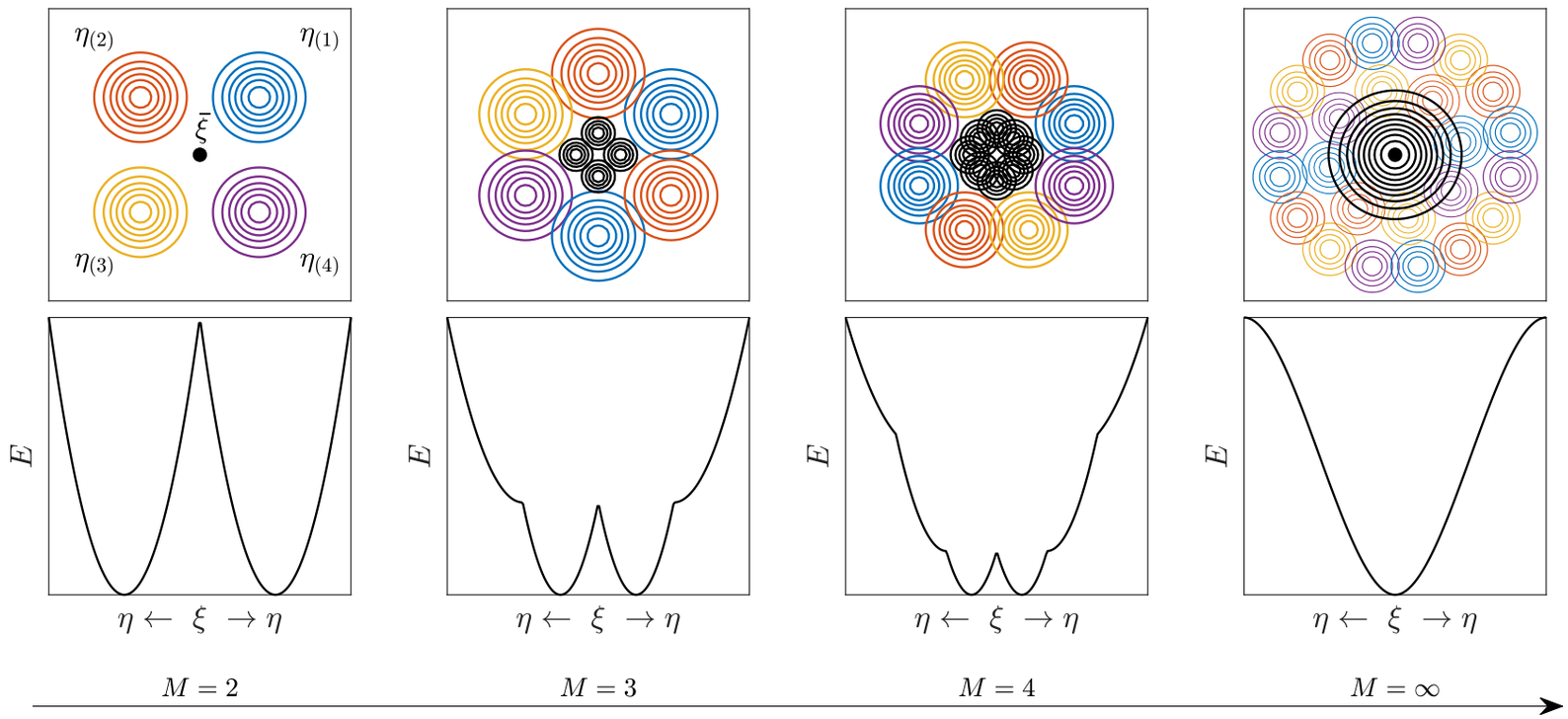
$$\bar{m} > \bar{n} \quad \rightarrow \quad M > M_\times \sim \frac{1 + r}{r}$$





$\rho := \alpha/(Mr^4)$ tuneable parameter (ruling the overall slow noise) able to trigger a phase transition

$\rho_c = 2/\pi$ critical transition



$\mathcal{S} = \{\boldsymbol{\eta}^{\mu,a}\}_{\mu=1,\dots,K}^{a=1,\dots,M}$ Training set

$\mathcal{Z} = \{z^{\nu}\}_{\mu=1,\dots,K}$ Set of one-hot vectors of length K ($z_{\mu}^{\nu} = \delta_{\mu,\nu}$)

$$\mathscr{H}_{N,K}(\boldsymbol{\sigma}, z; \boldsymbol{w}) = -\frac{1}{\sqrt{N}} \sum_{\mu=1}^{K} \sum_{i=1}^{N} \sigma_i w_i^{\mu} z_{\mu}$$

$$\mathscr{P}_{N,K}(\boldsymbol{\sigma}, z; \boldsymbol{w}) = \frac{1}{Z_{N,K}(\boldsymbol{w})} e^{-\beta \mathscr{H}_{N,K}(\boldsymbol{\sigma}, z; \boldsymbol{w})}$$

$$\Delta w_i^{\mu} \propto \Big(\langle \sigma_i z_{\mu} \rangle_{\text{clamped}} - \langle \sigma_i z_{\mu} \rangle_{\text{free}} \Big), \quad \forall (i,\mu) \in (N \times K)$$ Supervised learning rule

``clamped'' average is evaluated over the pairs $(\boldsymbol{\sigma}_E, z_E) \in \mathcal{S} \times \mathcal{Z}$, e.g. $\boldsymbol{\sigma}_E = \boldsymbol{\eta}^{\nu,1}, z_E = z^{\nu}$

``free'' average is sampled via a single step of Gibbs dynamics:
select $(\boldsymbol{\sigma}_E, z_E) \in \mathcal{S} \times \mathcal{Z}$, then using the Gibbs-chain $z_E \to \boldsymbol{\sigma}_{\text{free}} \to z_{\text{free}}$ obtain the state $(\boldsymbol{\sigma}_{\text{free}}, z_{\text{free}})$

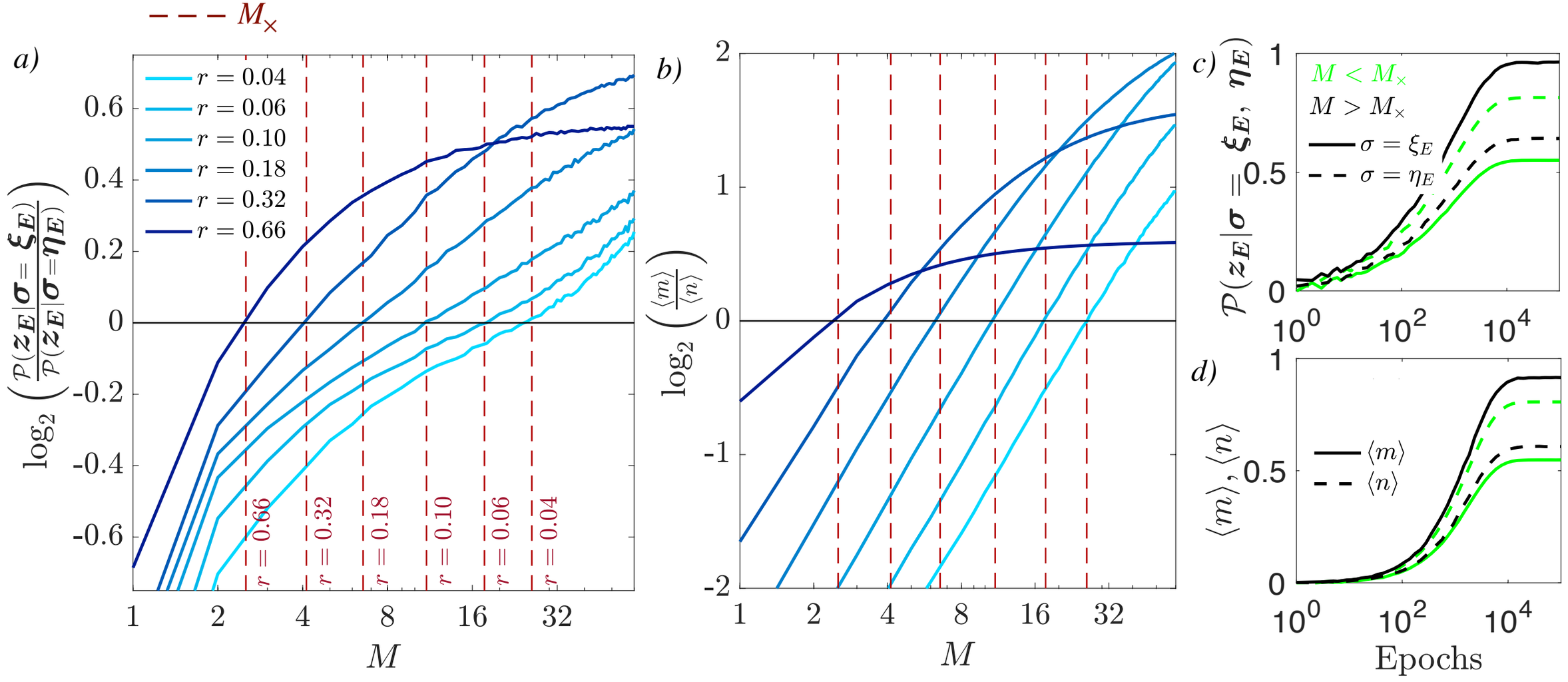
Classifier $\boldsymbol{\sigma} = \boldsymbol{\xi}^\nu, \boldsymbol{\eta}^{\nu,a} \mapsto z^\nu$

Performance measure: $\log \left[\dfrac{\mathscr{P}(z_E | \boldsymbol{\sigma} = \boldsymbol{\xi}_E)}{\mathscr{P}(z_E | \boldsymbol{\sigma} = \boldsymbol{\eta}_E)} \right]$

Generative model $z^\nu \mapsto \boldsymbol{\sigma} = \boldsymbol{\xi}^\nu, \boldsymbol{\eta}^{\nu,a}$

Performance measure: $\log \dfrac{\langle m_\nu \rangle_{z^\nu}}{\langle n_\nu \rangle_{z^\nu}}$

As training is running, as long as $M < M_\times$, the saturation values for $\mathscr{P}(z_E | \boldsymbol{\sigma} = \boldsymbol{\eta}_E)$ and $\langle n \rangle$ are larger than those obtained for $\mathscr{P}(z_E | \boldsymbol{\sigma} = \boldsymbol{\xi}_E)$ and $\langle m \rangle$; the opposite holds when $M > M_\times$

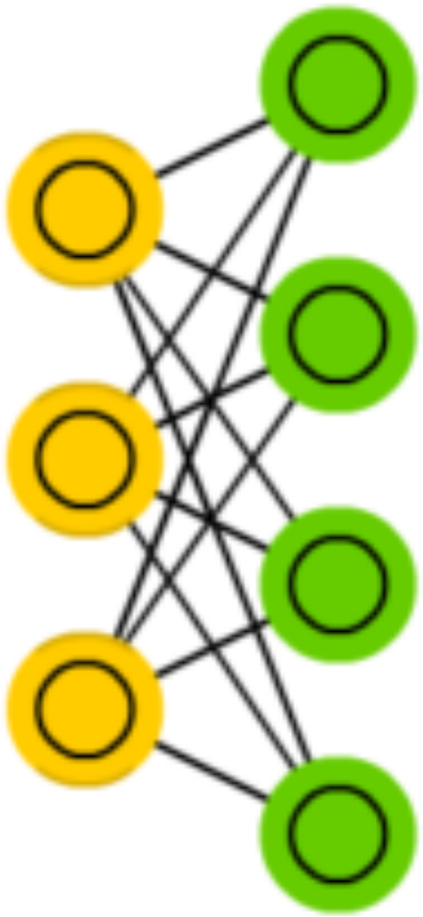


EA, F. Alemanno, A. Barra, G. Di Marzo, *The emergence of a concept in shallow neural networks* (2021)

Restricted Boltzmann machines

$\boldsymbol{v} \in \{-1, +1\}^{N_v}$

visible layer



$\boldsymbol{h} \in \{-1, +1\}^{N_h}$

hidden layer

Two layers composed of $N = N_v + N_h$ binary neurons

$\boldsymbol{s} = (\boldsymbol{v}, \boldsymbol{h}) \in \{-1, +1\}^{N_v+N_h}$

$\{J_{ij}\}$, with $i, j \in \{1,...,N\}$ interaction matrix (symmetric, zero eye)

$\{\vartheta_i\}$, with $i \in \{1,...,N\}$ bias vector

Theor. The partition function of the RR algorithm, given by

$$Z_{\beta,N,P,t}(\boldsymbol{\xi}) = \sum_{\{\boldsymbol{\sigma}\}} e^{-\beta H^{(RR)}_{N,P,t}(\boldsymbol{\sigma}|\boldsymbol{\xi})} = \sum_{\{\boldsymbol{\sigma}\}} \exp\left[\frac{\beta}{2N}\sum_{i,j=1}^{N,N}\sum_{\mu,\nu=1}^{P,P}\xi_i^\mu\xi_j^\nu\left(\frac{1+t}{\mathbb{I}+tC}\right)_{\mu,\nu}\sigma_i\sigma_j\right]$$

where $C_{\mu,\nu} := \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\xi_i^\nu$

can be represented in Gaussian integral form as

$$Z_{\beta,N,P,t}(\boldsymbol{\xi}) = \sum_{\{\sigma\}}\int\Big(\prod_{\mu=1}^{P}d\mu(z_\mu)\Big)\Big(\prod_{i=1}^{N}d\mu(\phi_i)\Big)\exp\left(\sqrt{\frac{\beta}{N}(t+1)}\sum_{\mu,i}^{P,N}z_\mu\xi_i^\mu\sigma_i + i\sqrt{\frac{t}{N}}\sum_{\mu,i}^{P,N}z_\mu\xi_i^\mu\phi_i\right)$$

that is equivalent to the partition function of a tripartite spin-glass with
intermediate layer made of a set of real neurons $\{z_\mu\}_{\mu=1,\dots,P}$ with $z_\mu \sim \mathcal{N}[0,1]$
external layers made, respectively, of a set of Boolean neurons $\{\sigma_i\}_{i=1,\dots,N}$ and of a set of imaginary neurons with magnitude $\{\phi_i\}_{i=1,\dots,N}$, being $\phi_i \sim \mathcal{N}[0,1]$.

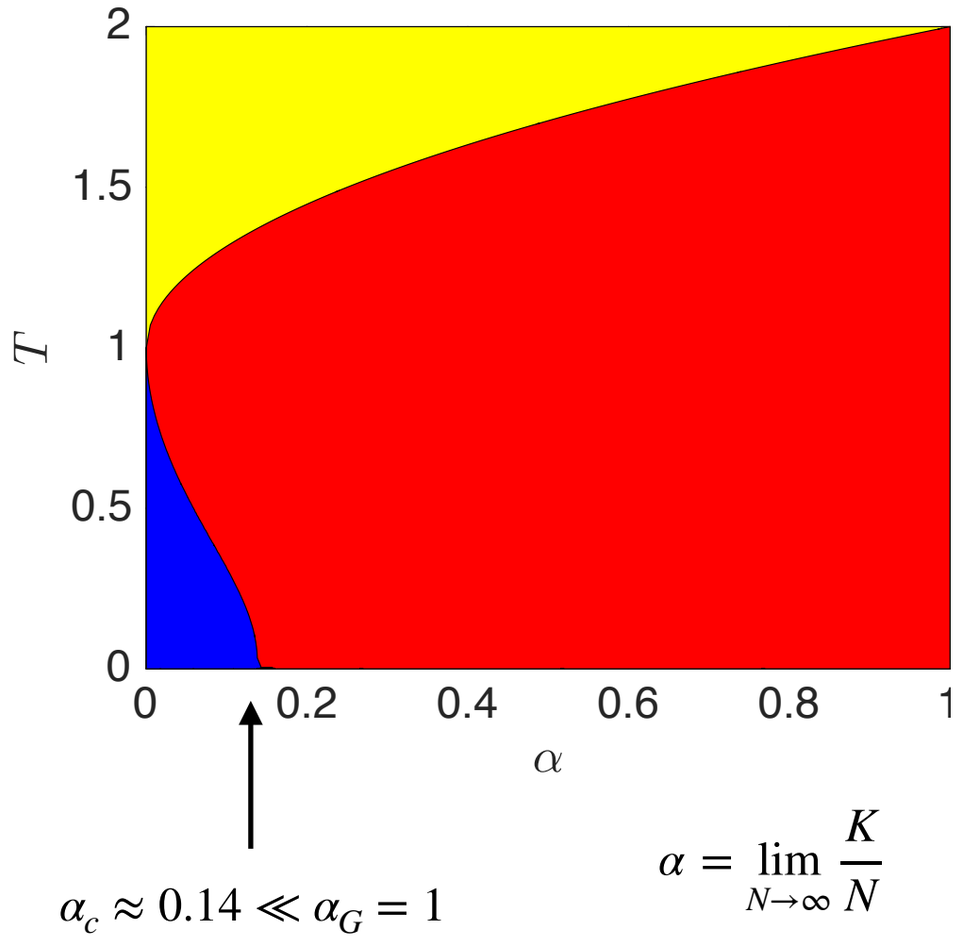
Dreaming Hopfield network

Hopfield network with “wide” retrieval region

Maximise the retrieval region to enhance storing

For symmetric neural networks capacity upper bounded by α_G =1

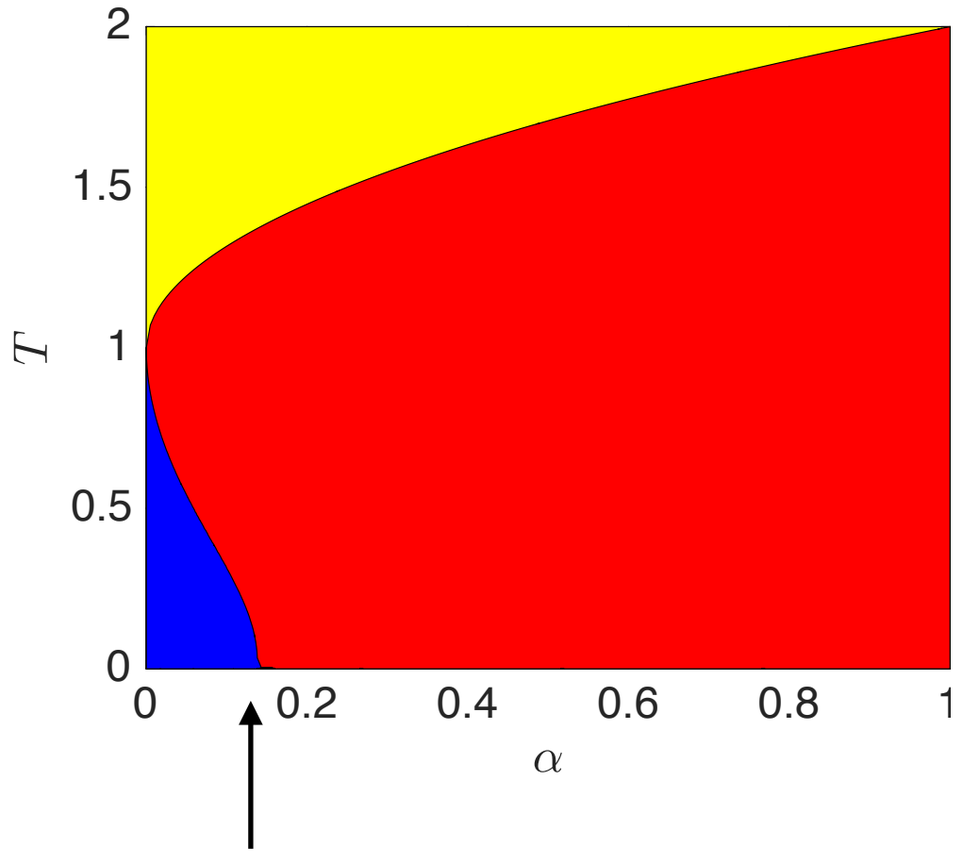
E. Gardner - J. Phys. A (1988)



Hopfield network with “wide” retrieval region

Maximise the retrieval region to enhance storing

For symmetric neural networks capacity upper bounded by α_G =1
E. Gardner - J. Phys. A (1988)



Projection rule (decorrelation prescription)
Kohonen - IEEE Trans. Comp. ’72
Personnaz - J. Phys. Lett. ’85
Kanter&Sompolinsky - Phys. Rev. A ‘86

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{K} \xi_i^\mu \xi_j^\mu \rightarrow \sum_{\mu,\nu=1}^{K} \xi_i^\mu C_{\mu\nu}^{-1} \xi_j^\nu$$

$$C_{\mu\nu} := N^{-1}\sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu$$

Interpreted as an iterative rule accounting for learning (patterns) and unlearning (spurious states)
Hopfield et al. - Nature Lett. ’83
Opper - Europhys. Lett. ’89
Plakhov - IAPR ’94

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{K} \xi_i^\mu \xi_j^\mu \rightarrow \left[J_{ij} - \epsilon \langle h_i h_j \rangle_{t'} \right]_{t'=0}^{t} \rightarrow J_{ij}(t) = \sum_{\mu,\nu=1}^{K} \xi_i^\mu \, (\boldsymbol{I}_K + t\boldsymbol{C})_{\mu,\nu}^{-1} \, \xi_j^\nu$$

Reinforcement&Removal (RR) algorithm

$$\frac{dJ}{dt} = \frac{1}{1+t}(J - J^2)$$

$$J_{ij}(t) = \frac{1}{N} \sum_{\mu,\nu=1}^{K,K} \xi_i^\mu \xi_j^\nu \left(\frac{1+t}{\boldsymbol{I}_K + t\boldsymbol{C}} \right)_{\mu\nu}$$

with $C_{\mu\nu} = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \xi_j^\nu$ the pattern correlation matrix

$$\mathscr{H}_{N,K}^{(RR)}(\boldsymbol{\sigma}; \boldsymbol{\xi}, t) = -\frac{1}{2N} \sum_{i,j=1}^{N,N} \sum_{\mu,\nu=1}^{K,K} \xi_i^\mu \xi_j^\nu \left(\frac{1+t}{\boldsymbol{I}_K + t\boldsymbol{C}} \right)_{\mu,\nu} \sigma_i \sigma_j$$

Goal: thermodynamic limit of the quenched-averaged intensive pressure

$$f(\beta, \alpha, t) := -\lim_{N\uparrow\infty} \frac{1}{N\beta} \mathbb{E} \ln Z_{N,K}(\beta, \boldsymbol{\xi}, t)$$

being $Z_{N,K}(\boldsymbol{\xi}, t) = \sum_{\boldsymbol{\sigma}} e^{-\beta \mathscr{H}_{N,K}^{(RR)}(\boldsymbol{\sigma};\boldsymbol{\xi},t)}$

A. Fachechi, EA, A. Barra, *Dreaming neural networks: forgetting spurious memories and reinforcing pure ones*, Neur. Netw. (2019)

$$-\beta f^{RS}(\beta,\alpha,t) = \log 2 - \frac{\beta m^2}{2(1+t)}\Big(1+\frac{t}{\Delta}\Big) - \frac{(1+t)(\Delta-1)}{2t}\beta\bar{q} + \mathbb{E}_\eta \log\cosh\Big[\frac{\beta}{\Delta}(m+\sqrt{\alpha\bar{p}}\eta)\Big] - \frac{\log\Delta}{2} - \frac{\alpha\beta\bar{p}t}{2(1+t)\Delta}$$

$$-\frac{\alpha}{2}\Big(\log[1-\beta(1+t)(\bar{q}-\bar{q}')] + \frac{\bar{q}'\beta^2(1+t)}{1-\beta(1+t)(\bar{q}-\bar{q}')}\Big) - \frac{(1+t)(1-\Delta)\beta}{2t\Delta} - \frac{\alpha\beta^2}{2}\bar{p}(\bar{q}-\bar{q}')\,.$$

$$\begin{aligned}
m &= \frac{1+t}{\Delta+t}\mathbb{E}_\eta \tanh\Big[\frac{\beta}{\Delta}(m+\sqrt{\alpha\bar{p}}\eta)\Big],\\
\bar{p} &= \frac{\bar{q}'(1+t)^2}{[1-\beta(1+t)(\bar{q}-\bar{q}')]^2},\\
\Delta &= 1+\frac{\alpha t}{1-\beta(1+t)(\bar{q}-\bar{q}')},\\
\bar{q}' &= \bar{q}+\frac{t}{\beta(1+t)\Delta}-\frac{1}{\Delta^2}\mathbb{E}_\eta \cosh^{-2}\Big[\frac{\beta}{\Delta}(m+\sqrt{\alpha\bar{p}}\eta)\Big],\\
\bar{q}\Delta^2 &= 1-\frac{t\Delta}{\beta(1+t)}+\frac{\alpha\bar{p}t^2-m^2t(t+2\Delta)}{(1+t)^2}-\frac{2\alpha\beta\bar{p}t}{(1+t)\Delta}\mathbb{E}_\eta \cosh^{-2}\Big[\frac{\beta}{\Delta}(m+\sqrt{\alpha\bar{p}}\eta)\Big].
\end{aligned}$$

Critical surface delimiting the ergodic region

$$\begin{aligned}
\beta_c &= \frac{1}{1+t}\Big[\frac{\Delta^2}{1+\sqrt{\alpha}}+t\Delta\Big],\\
\Delta &= 1+\sqrt{\alpha}(1+\sqrt{\alpha})t.
\end{aligned}$$

Let the network… dream!

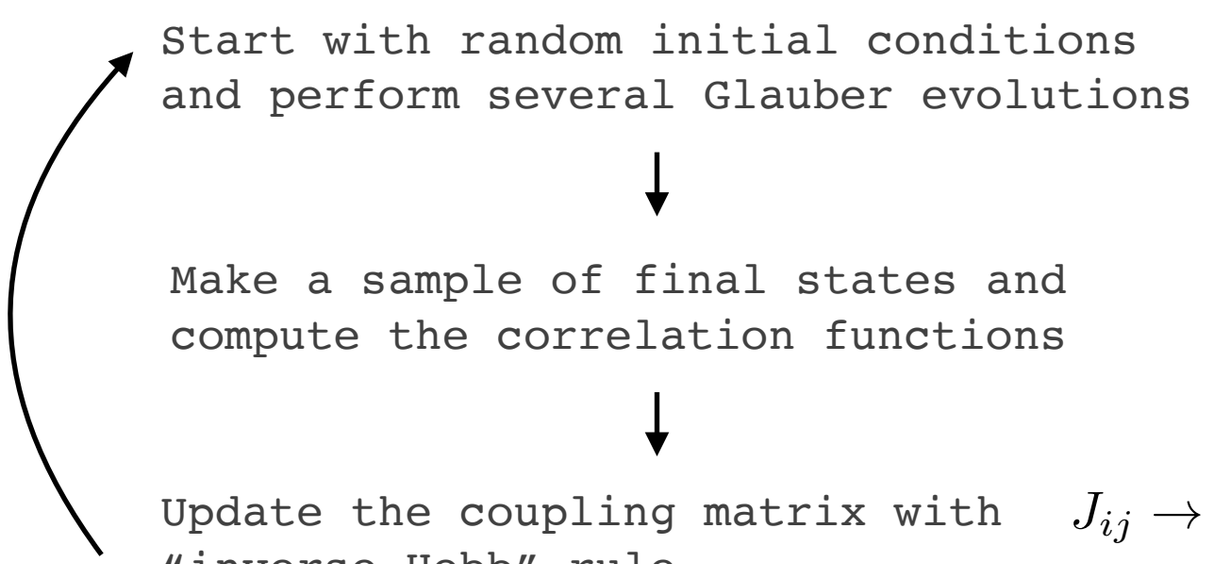
F. Crick, G. Mitchinson, *The function of dream sleep*, Nature (1983).

«We propose that the function of dream sleep (more properly rapid-eye movement or REM sleep) is to remove certain undesirable modes of interaction in networks of cells in the cerebral cortex. We postulate that this is done in REM sleep by a ***reverse learning mechanism****, so that the trace in the brain of the unconscious dream is weakened, rather than strengthened, by the dream.»*

REM sleeping is associated to the **removal** of unwanted attractors (*unlearning*)

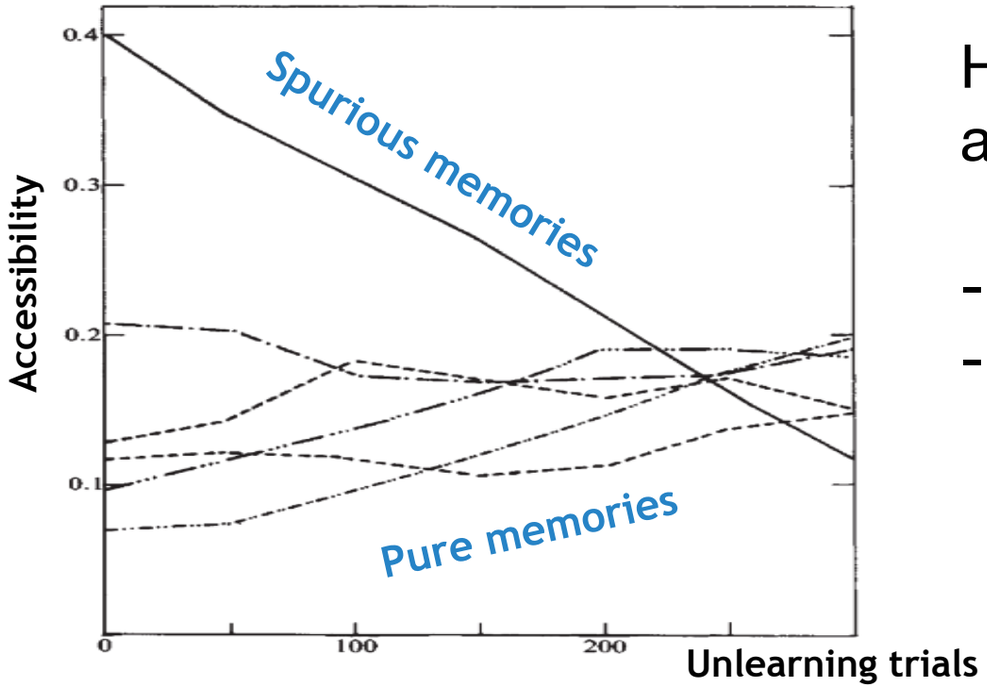
J.J. Hopfield, D.I. Feinstein, R.G. Palmer, *Unlearning has a stabilizing effect in collective memories*, Nature (1983).

Unlearning rule aiming to increase the energy of non pure attractors (they get less stable)



$$J_{ij} \to J_{ij} - \frac{\epsilon}{N}\langle \sigma_i \sigma_j \rangle_{\text{sample}}$$

$$\epsilon \ll 1$$

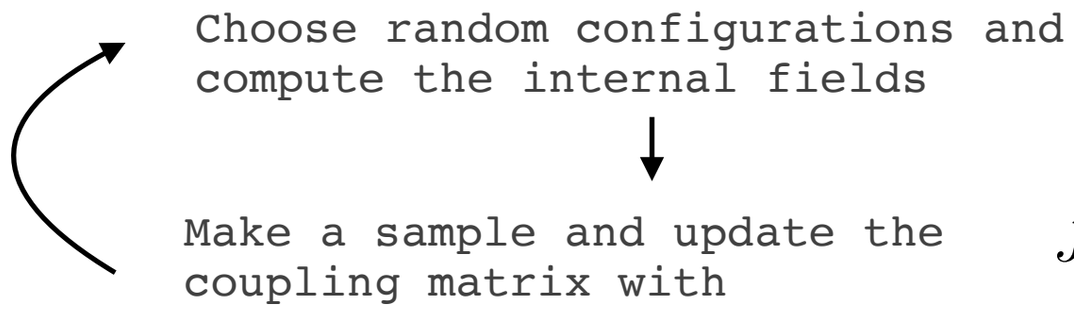


Hopfield's unlearning succeeds in reducing the accessibility of spurious memories, yet…

- Too much unlearning can also destroy pure memories
- No analytical treatment

J.L. Van Hemmen, L.B. Ioffe, R. Kühn, M. Vaas, *Increasing the efficiency of a neural network through unlearning*, Phys. A (1990).

A. Y. Plakhov, S.A. Semenov, *The modified unlearning procedure for enhancing storage capacity in Hopfield network*, IEEE Trans. (1992).



$$J_{ij} \to J_{ij} - \frac{\epsilon}{N}\langle h_i h_j \rangle_{\text{sample}}$$

In the limit of large *t*, the coupling matrix converges to the **projector matrix *P*** → Analytically treatable (but no longer local)

$$P_{ij} = \frac{1}{N}\sum_{\mu,\nu=1}^{P,P} \xi_i^\mu (C^{-1})_{\mu\nu} \xi_j^\mu,$$

$$\text{where } C_{\mu\nu} = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu$$

The unlearning strength ϵ identifies a temporal scale

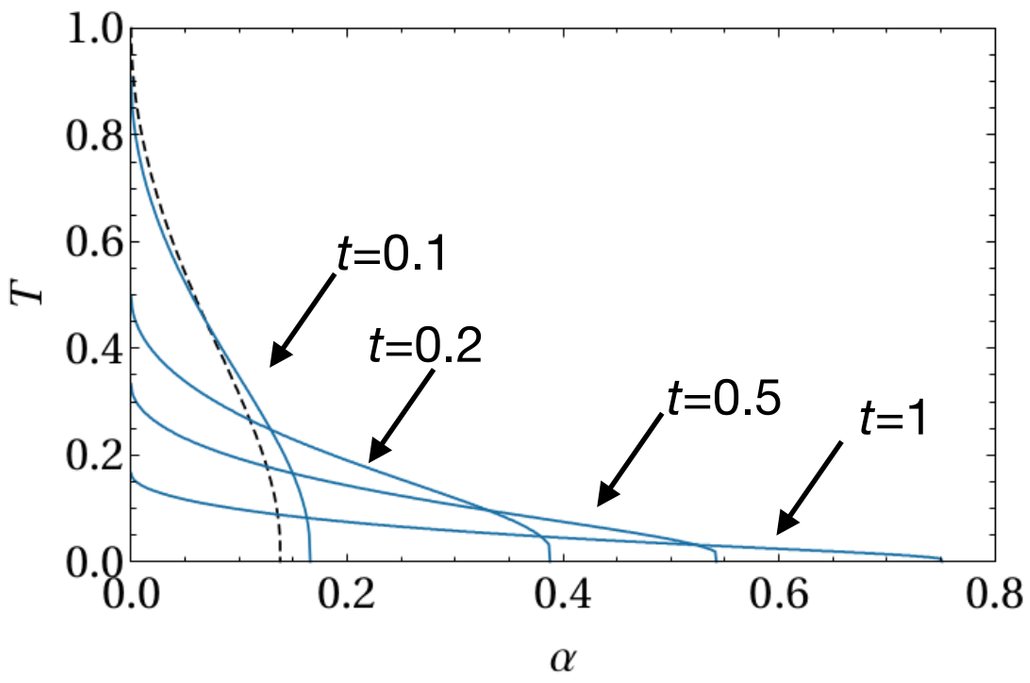
Taking the continuous time limit *(dt ~ ε)*, the unlearning procedure is described by a the differential equation

$$\frac{dJ}{dt} = -J^2 \qquad \text{with solution} \qquad J_{ij}(t) = \frac{1}{N} \sum_{\mu,\nu=1}^{P,P} \xi_i^\mu \xi_j^\nu (1 + tC)^{-1}_{\mu\nu}$$

where $C_{\mu\nu} = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \xi_j^\nu$ is the pattern correlation matrix

V. Dotsenko, N.D. Yarunin, E.A. Dorotheyev, Statistical mechanics of Hopfield-like neural networks with modified interactions, J. Phys. A (1991).

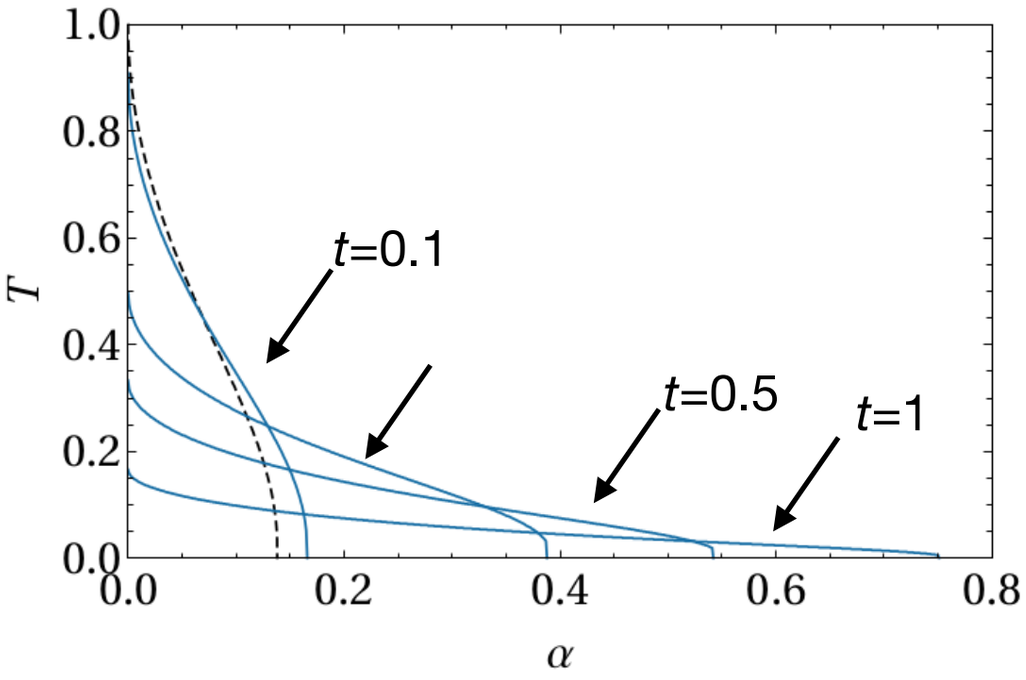
Analytical solution of the model in the replica-symmetric regime by means of replica trick technique.



The critical capacity at zero temperature grows with *t*, but the area retrieval region vanishes in the large unlearning time limit

Critical load ($t = \infty$) ≈ 1.07

V. Dotsenko, N.D. Yarunin, E.A. Dorotheyev, Statistical mechanics of Hopfield-like neural networks with modified interactions, J. Phys. A (1991).



Too much unlearning destroys pure memories
With this unlearning rule all couplings tend to zero in the large unlearning-time limit

The area of the retrieval region vanishes in the large unlearning-time limit

Intuitively…
The unlearning algorithm does not distinguish between pure and spurious memories
→ Too much unlearning destroys also the former.

Mathematically…
The generalized kernel in the coupling matrix equals the zero matrix in the large unlearning time limit → All synaptic strengths vanish.

Biologically…
The function of sleeping (REM sleep + Slow Wave Sleep) is to weaken fictitious synaptic strengths *and* consolidate relevant ones (Walker 2009, Born 2010…).

A. Fachechi, E. Agliari, A. Barra, *Dreaming neural networks: forgetting spurious memories and reinforcing pure ones*, Neur. Net. (2019).

Introduce a regularizator

$$(1+tC)^{-1}_{\mu\nu} \to \left(\frac{1+t}{1+tC}\right)_{\mu\nu}$$

$$J_{ij}(t) = \frac{1}{N}\sum_{\mu,\nu=1}^{P,P} \xi_i^\mu \xi_j^\nu \left(\frac{1+t}{1+tC}\right)_{\mu\nu}$$

The resulting model accounts for both **removal and consolidation**!

Argument 1: Look at the evolution process leading to *J*(t)

$$\frac{dJ}{dt} = \frac{1}{1+t}(J - J^2)$$

Or, in discrete sleeping time…

$$dt \to \epsilon,$$
$$t \to k\epsilon,$$
$$\dot{J} \to \epsilon^{-1}[J(k+1) - J(k)],$$

$$\boldsymbol{J}(k+1) = \boldsymbol{J}(k) + \underbrace{\frac{\epsilon}{1+\epsilon k}}\left[\boldsymbol{J}(k) - \boldsymbol{J}(k)^2\right]$$

effective sleep strength depends on the sleep session *k*

$$J(k+1) = J(k) + \frac{\epsilon}{1+\epsilon k}\Big[J(k) - J(k)^2\Big]$$

with this prescription $||\mathbf{J}(k) - \mathbf{P}|| \xrightarrow{k\to\infty} 0$ under operator norm

Projector (or pseudo-inverse)

$$P_{ij} = \frac{1}{N}\sum_{\mu,\nu=1}^{P,P} \xi_i^\mu (C^{-1})_{\mu\nu} \xi_j^\mu,$$

$$\text{where } C_{\mu\nu} = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu$$

Operator norm

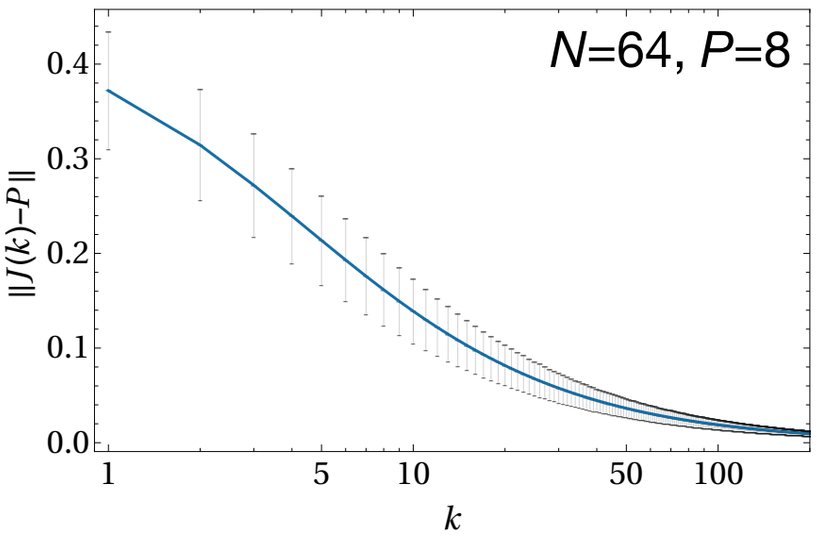
$$||\mathbf{A}|| = \sqrt{\max\{a | a \in \sigma(\mathbf{A}^T\mathbf{A})\}}$$

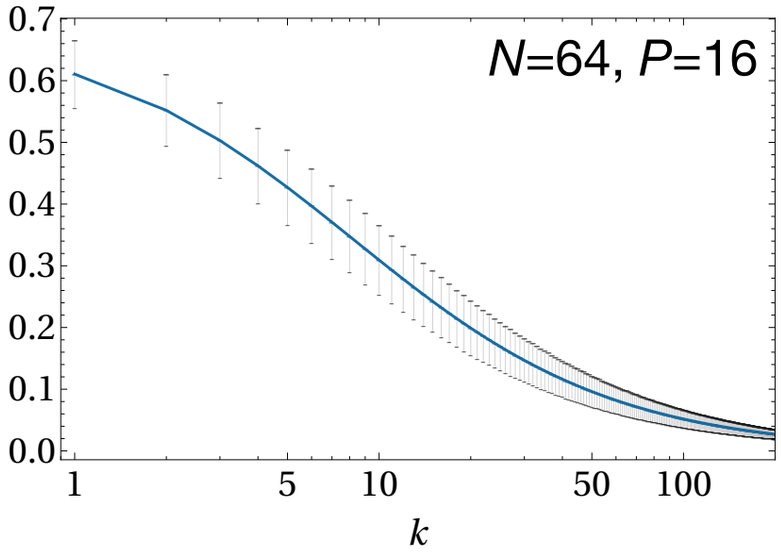
Sketch of the proof

Introduce a matrix ***G*** such that $J_{ij}(k) = \frac{1}{N}\sum_{\mu,\nu=1}^{P} \xi_i^\mu \xi_j^\nu G_{\mu\nu}(k)$

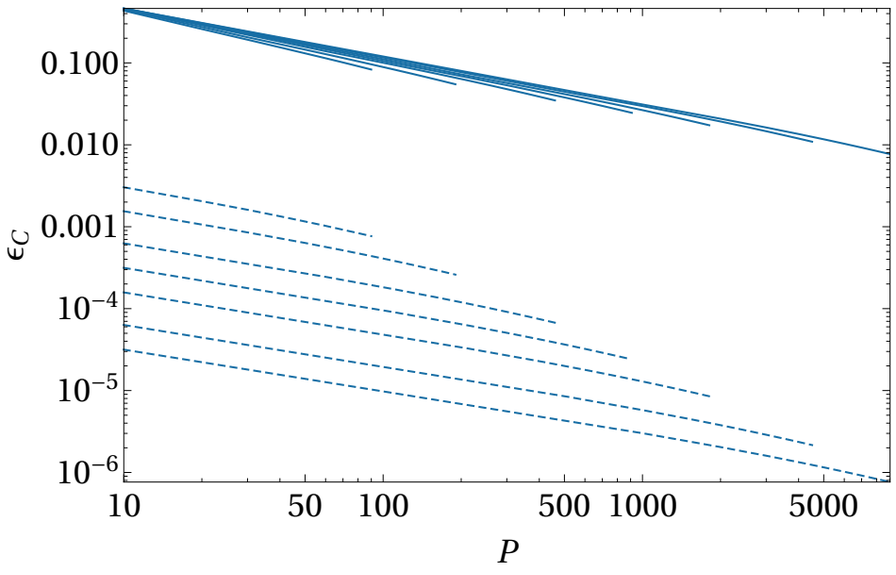
then, prescription recast as $\mathbf{G}(k+1) = \left(1 + \frac{\epsilon}{1+\epsilon k}\right)\mathbf{G}(k) - \frac{\epsilon}{1+\epsilon k}\mathbf{G}(k)\mathbf{C}\mathbf{G}(k)$, with initial condition $G(0) = \mathbb{I}$

- ||***C***|| ≥ 1
- ***G***(k) commutes with ***C*** $\forall k$
- ***G***(*k*) are invertible for *k*≥0 and $\varepsilon > \varepsilon_c$
- There exists finite, real $c_k \geq$ ||***C*** ***G***(*k*)||. The sequence is therefore upper bounded by $\bar{c} = \max_k c_k$
- The unlearning algorithm converges to the stationary solution ***G***(∞) = ***C***$^{-1}$ in the sense defined by the operator norm
- Critical value for ε to ensure convergence $\varepsilon_c = (||\mathbf{C}||-1)^{-1}$





Average over 500 different realizations of patterns, ε=0.5



ε_c higher than Plakhov et al.'s and decreases slower with *N*

A. Fachechi, E. Agliari, A. Barra, *Dreaming neural networks: forgetting spurious memories and reinforcing pure ones*, Neur. Net. (2019).

Introduce a regularizator

$$(1+tC)^{-1}_{\mu\nu} \to \left(\frac{1+t}{1+tC}\right)_{\mu\nu}$$

$$J_{ij}(t) = \frac{1}{N}\sum_{\mu,\nu=1}^{P,P} \xi_i^\mu \xi_j^\nu \left(\frac{1+t}{1+tC}\right)_{\mu\nu}$$

The resulting model accounts for both **removal and consolidation**!

Argument 1: Look at the evolution process leading to *J*(t)

$$\frac{dJ}{dt} = \frac{1}{1+t}(J - J^2)$$

Or, in discrete sleeping time…

$$dt \to \epsilon,$$
$$t \to k\epsilon,$$
$$\dot{J} \to \epsilon^{-1}[J(k+1) - J(k)],$$

$$\boldsymbol{J}(k+1) = \boldsymbol{J}(k) + \underbrace{\frac{\epsilon}{1+\epsilon k}}\left[\boldsymbol{J}(k) - \boldsymbol{J}(k)^2\right]$$

effective sleep strength depends on the sleep session *k*

A. Fachechi, E. Agliari, A. Barra, *Dreaming neural networks: forgetting spurious memories and reinforcing pure ones*, Neur. Net. (2019).

Introduce a regularizator

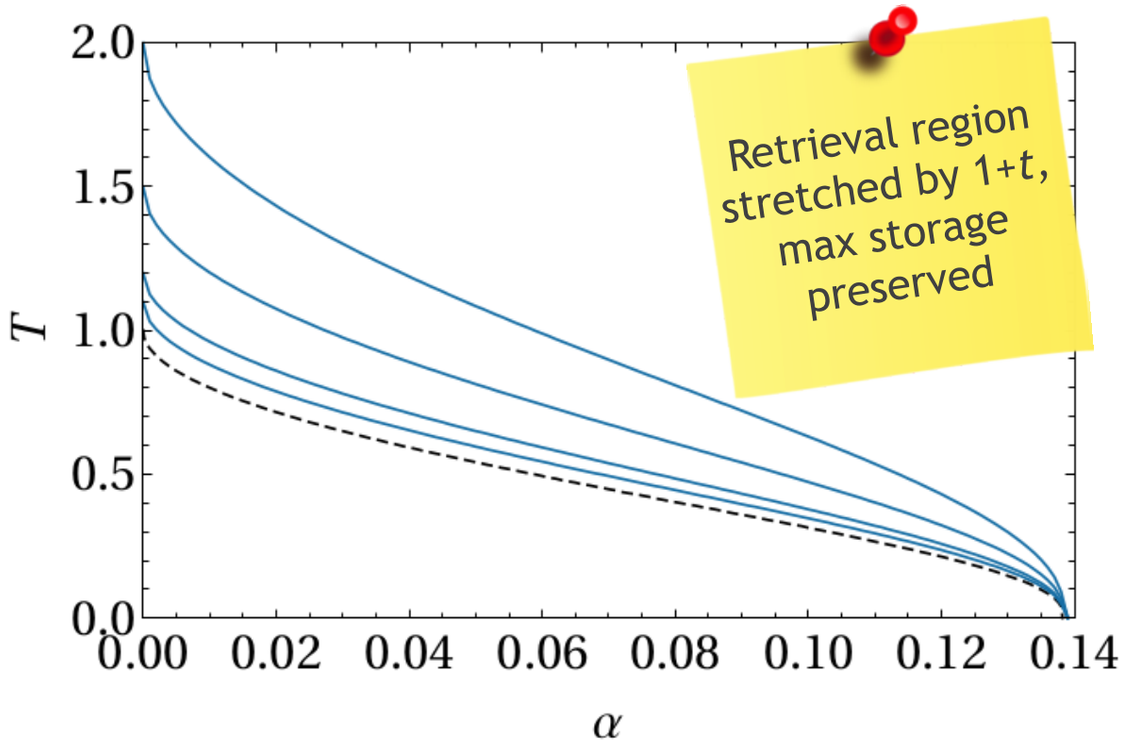
$$(1+tC)^{-1}_{\mu\nu} \to \left(\frac{1+t}{1+tC}\right)_{\mu\nu}$$

$$J_{ij}(t) = \frac{1}{N}\sum_{\mu,\nu=1}^{P,P} \xi_i^\mu \xi_j^\nu \left(\frac{1+t}{1+tC}\right)_{\mu\nu}$$

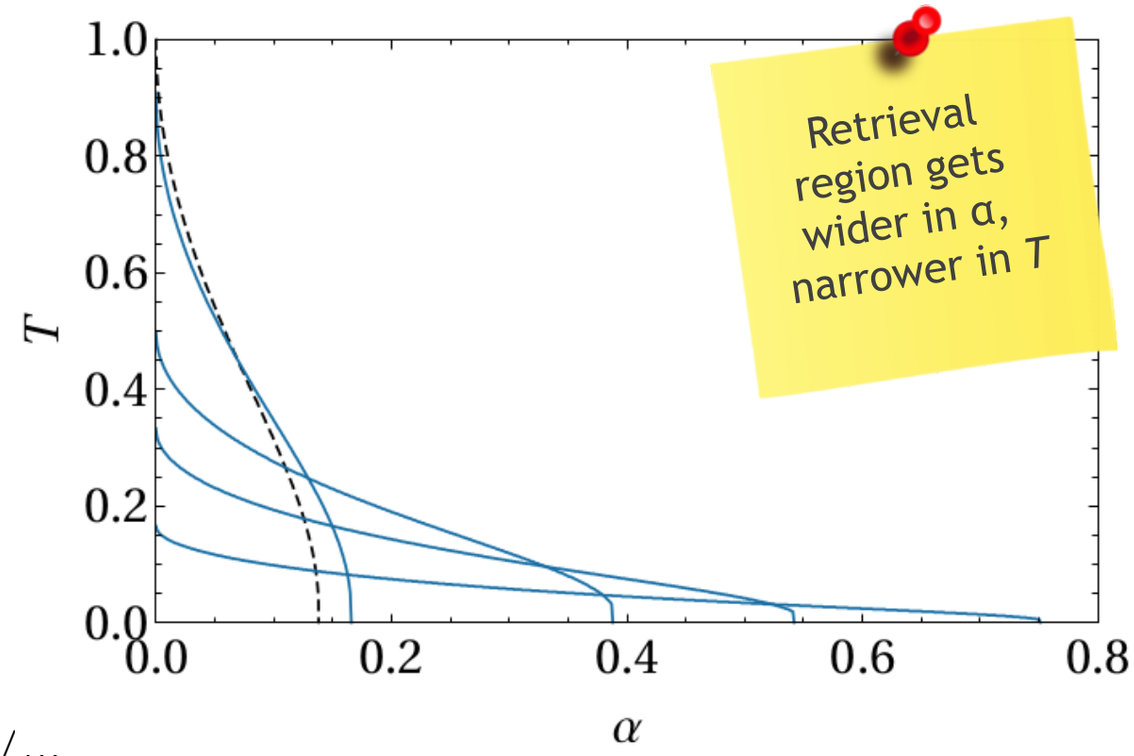
The resulting model is accounts for both **removal and consolidation**!

Argument 2: consider the statistical mechanics of two separated models

$$H^{(1)}_{N,P} = -\frac{1}{2}\sum_{\mu=1}^{P}\sum_{i,j=1}^{N,N} \xi_i^\mu \xi_j^\nu (1+t)\sigma_i\sigma_j$$



$$H^{(2)}_{N,P} = -\frac{1}{2}\sum_{\mu,\nu=1}^{P,P}\sum_{i,j=1}^{N,N} \xi_i^\mu \xi_j^\nu (\mathbb{I}+tC)^{-1}_{\mu\nu}\sigma_i\sigma_j$$



The model is analytically treatable, even in the high-load regime!

Self-consistent equations for the order parameters obtained (rigorously) under replica-symmetry hypothesis.

Order parameters

m → assess retrieval of pattern ξ^1 (measure overlap between neural configuration and pattern)

$\boldsymbol{q}$, $\boldsymbol{p}$ → assess glassiness in the system (measure overlap between neural configurations of two replicas)

Replica symmetry ansatz

$$
\begin{aligned}
m_1^\alpha &= m \quad \forall \alpha, \\
q_{\alpha\beta} &= Q\delta_{\alpha\beta} + q(1-\delta_{\alpha\beta}), \\
p_{\alpha\beta} &= P\delta_{\alpha\beta} + p(1-\delta_{\alpha\beta}),
\end{aligned}
$$

Useful quantity

$$\Delta := 1 + \alpha\beta t(1+t)^{-1}(P-p)$$

The model is analytically treatable, even in the high-load regime!

Self-consistent equations for the order parameters obtained (rigorously) under replica-symmetry hypothesis.

$$m = \frac{1+t}{\Delta+t}\int Dx \tanh\left[\frac{\beta}{\Delta}(m+\sqrt{\alpha p}x)\right]$$

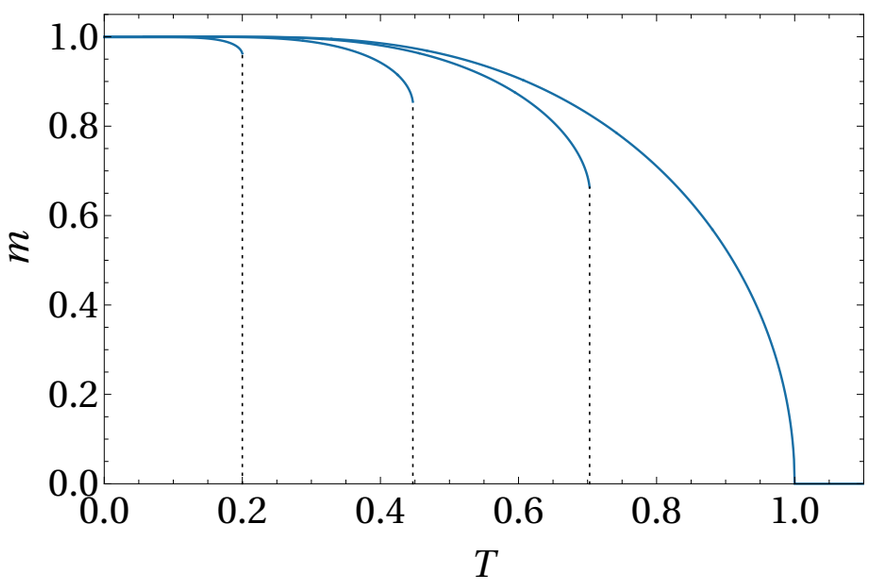
$$p = \frac{q(1+t)^2}{[1-\beta(1+t)(Q-q)]^2}$$

$$\Delta = 1+\frac{\alpha t}{1-\beta(1+t)(Q-q)}$$

$$q = Q+\frac{t}{\beta(1+t)\Delta}-\frac{1}{\Delta^2}\int Dx \cosh^{-2}\left[\frac{\beta}{\Delta}(m+\sqrt{\alpha p}x)\right]$$

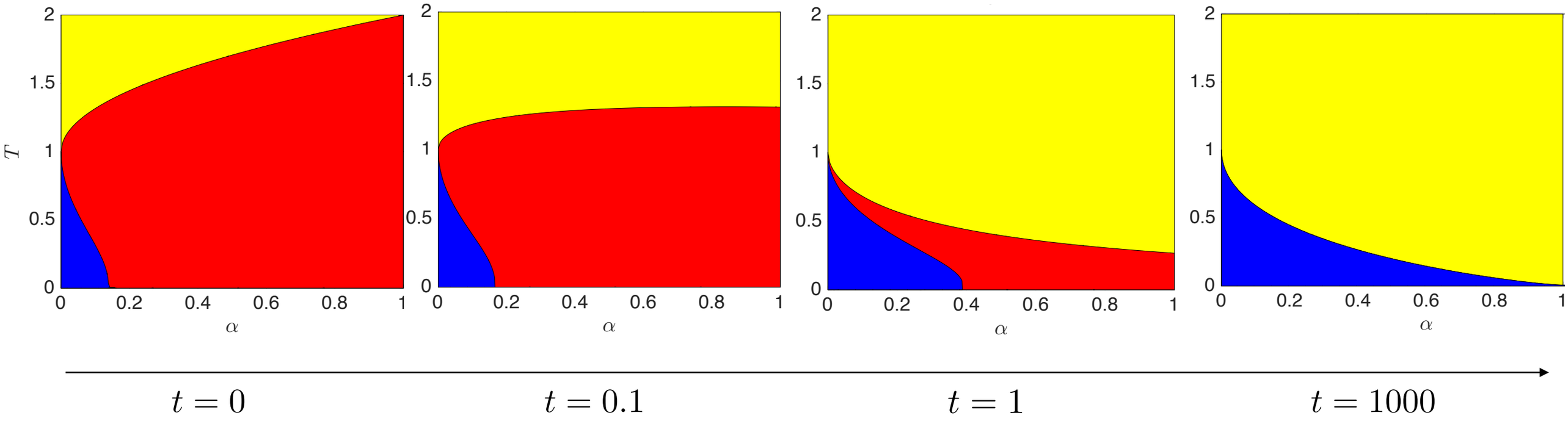
$$Q\Delta^2 = 1-\frac{t\Delta}{\beta(1+t)}+\frac{\alpha p t^2}{(1+t)^2}-\frac{m^2 t(t+2\Delta)}{(1+t)^2}-\frac{2\alpha\beta p t}{(1+t)\Delta}\int Dx \cosh^{-2}\left[\frac{\beta}{\Delta}(m+\sqrt{\alpha p}x)\right]$$

By evaluating these observables as functions of the parameters (T, α) we build the phase diagram, focusing on the retrieval region



Mattis magnetization for t = 1000 as a function of temperature, for various storage capacity values (α=0, 0.05, 0.2 and 0.5)

E. Agliari, F. Alemanno, A. Barra, A. Fachechi, *Dreaming neural networks: rigorous results*, J. Stat. Phys. (2019)



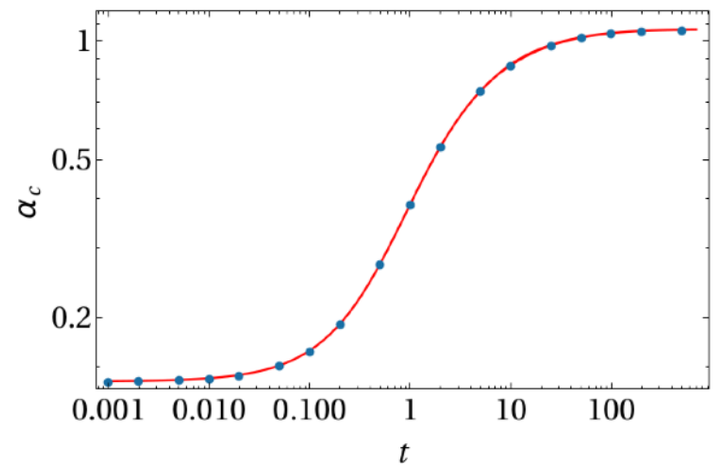
Retrieval region gets wider

Spin-glass (where slow noise, i.e., spurious states, prevail) region collapses

Ergodic (where fast noise, i.e., purely random states, prevail) region gets wider

Ergodic line changes concavity

Critical load ($t = \infty$) ≈ 1.07

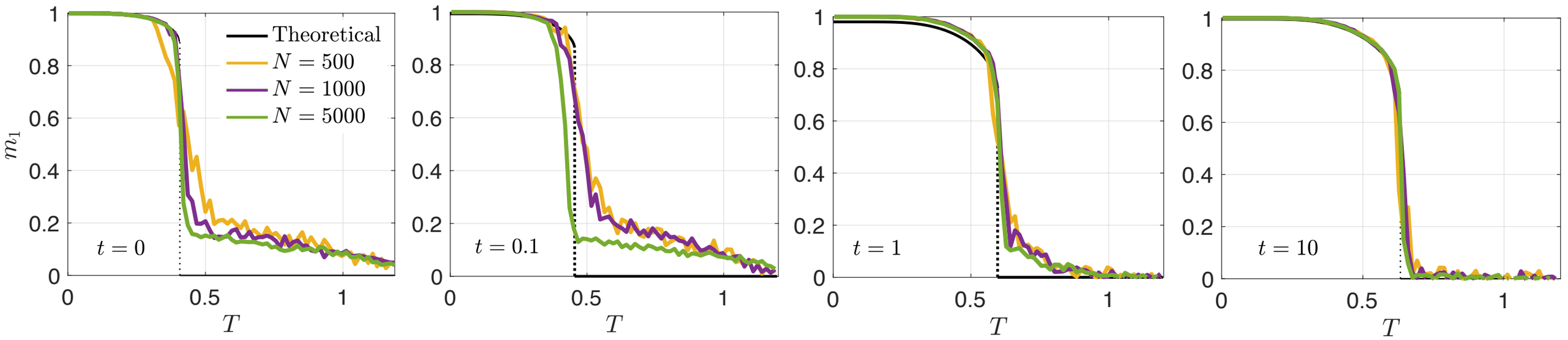


Perfect retrieval is accomplished!

A technical note: the underlying mathematics is rigorous (at replica-symmetry resolution)

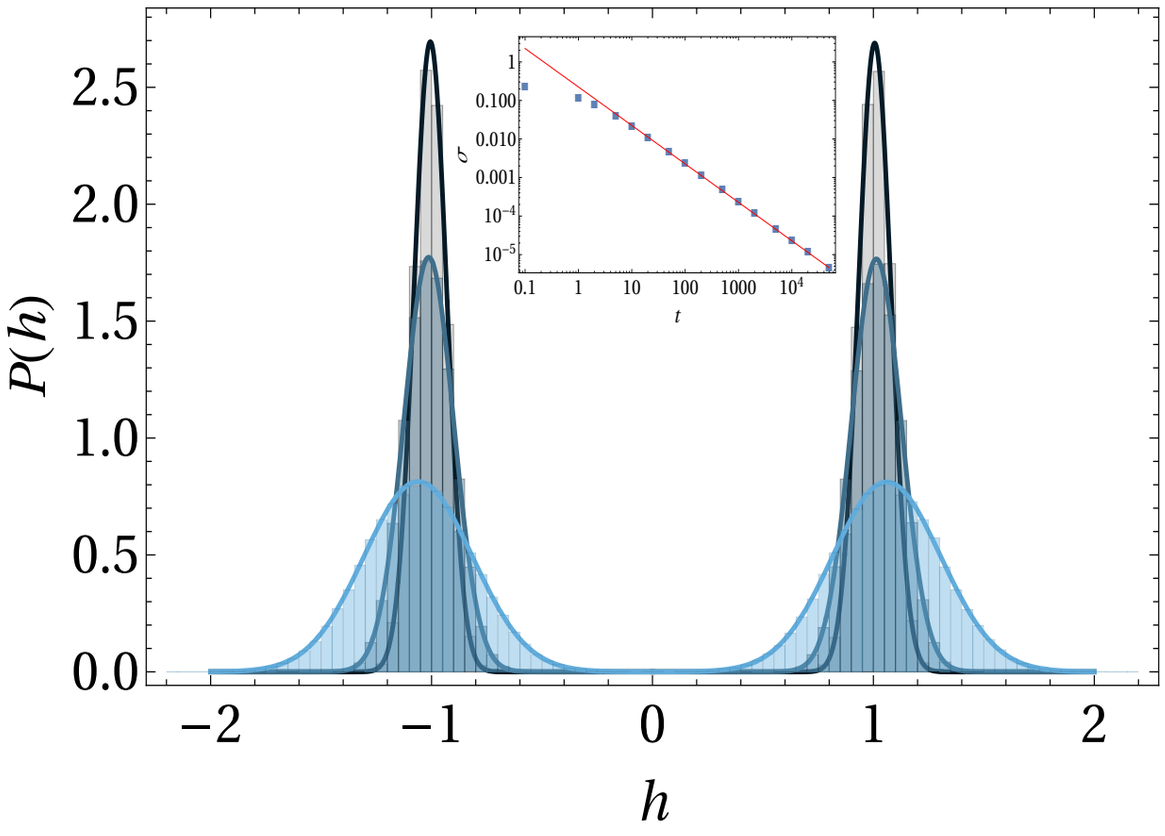
Comparison with numerics

Check hp underlying analytics (thermodynamic limit, replica symmetry)



Average values for the Mattis magnetization m_1 corresponding to the retrieved pattern ξ^1 obtained from numerical simulations for fixed α=0.08 and M=10

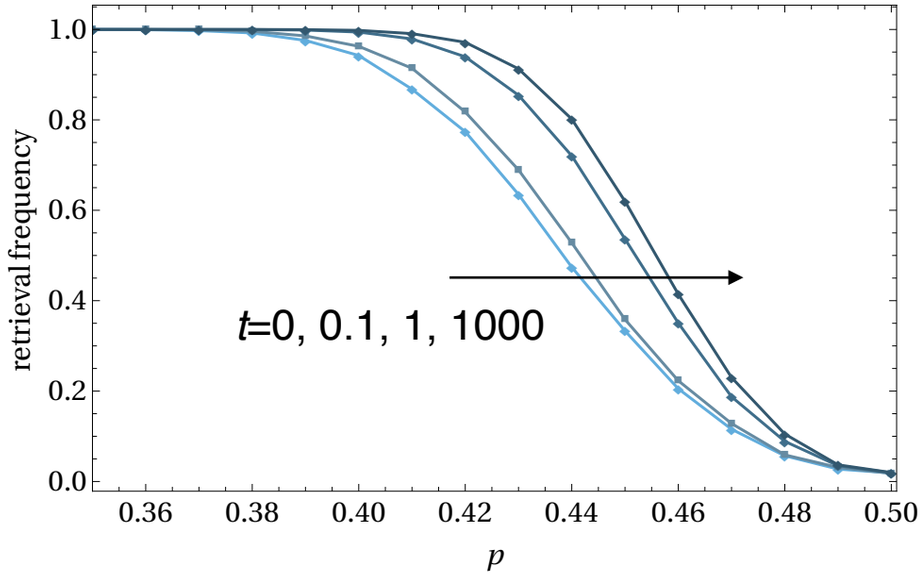
Check robustness of attraction basins



$$h_i = \frac{1}{N}\sum_{j=1}^{N}\sum_{\mu\nu}\xi_i^\mu \xi_j^\nu (1+t)(1+tC)^{-1}_{\mu\nu}\sigma_j$$

$\sigma_{P(h)} \sim t^{-1}$ → Sleep as a regularizer in the internal field distributions

p = misalignment probability of the initial state
200 different stochastic evolutions for each of the 200 pattern realizations



More techanical details

Definition 1 *The Hamiltonian of the reinforcement&removal algorithm reads as:*

$$H_{N,P}^{(RR)}(\sigma|\xi,t) = -\frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{\mu=1}^{P}\sum_{\nu=1}^{P}\xi_i^\mu\xi_j^\nu\left(\frac{1+t}{\mathbb{I}+tC}\right)_{\mu,\nu}\sigma_i\sigma_j, \tag{1}$$

where ξ^1, the pattern candidate to be retrieved, has binary entries $\xi_i^1 \in \{-1,+1\}$ drawn from $P(\xi_i^\mu = +1) = P(\xi_i^\mu = -1) = \frac{1}{2}$, while the remaining $P-1$ patterns $\{\xi^\mu\}_{\mu=2,\dots,P}$, have i.i.d. standard Gaussian entries $\xi_i^\mu \sim \mathcal{N}[0,1]$, and the correlation matrix $\boldsymbol{C}$ is defined as

$$C_{\mu,\nu} := \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\xi_i^\nu.$$

Definition 2 *Being $\beta \in \mathbb{R}^+$ a parameter tuning the level of* fast noise *in the network (with the physical meaning of inverse temperature), the partition function of the model (1) is introduced as*

$$Z_{N,P}(\sigma|\xi,t) = \sum_{\{\sigma\}} e^{-\beta H_{N,P}^{(RR)}(\sigma|\xi,t)} = \sum_{\{\sigma\}}\exp\left[\frac{\beta}{2N}\sum_{i,j=1}^{N,N}\sum_{\mu,\nu=1}^{P,P}\xi_i^\mu\xi_j^\nu\left(\frac{1+t}{\mathbb{I}+tC}\right)_{\mu,\nu}\sigma_i\sigma_j\right]. \tag{2}$$

Definition 3 *Denoting with $\mathbb{E}_\xi$ the average over the quenched patterns, for a generic function $O(\sigma,\xi)$ of the neurons and the couplings, we can define the Boltzmann $\omega(O(\sigma,\xi))$ and the quenched $\langle O(\sigma,\xi)\rangle$ averages as*

$$\omega\left(O(\sigma,\xi)\right) = \frac{\sum_{\{\sigma\}} O(\sigma,\xi)e^{-\beta H_{N,P}^{(RR)}(\sigma|\xi,t)}}{Z_{N,P}(\sigma|\xi,t)},$$

$$\langle O(\sigma,\xi)\rangle = \mathbb{E}_\xi\omega\left(O(\sigma,\xi)\right).$$

Definition 4 *Once introduced the partition function $Z_{N,P}(\sigma|\xi,t)$, we can define the infinite volume limit of the intensive quenched free-energy $F_N(\alpha,\beta,t)$ and of the intensive quenched pressure $A(\alpha,\beta,t)$ associated to the model (1) as*

$$-\beta F(\alpha,\beta,t) = A(\alpha,\beta,t) = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\ln Z_{N,P}(\sigma|\xi,t).$$

Remark 1 *For the sake of mathematical convenience, we take solely the pattern candidate for retrieval (i.e. the* signal*) to be Boolean, while all the remaining ones (acting as* slow noise *on the retrieval) are chosen as Gaussian. "Universality" as long as the distribution function of the quenched patterns generating the slow noise satisfies*

$$\exists L > 0 \,|\, p > 2 \implies |\mathbb{E}_\xi \xi^p| < L, \mathbb{E}_\xi \xi = 0, \mathbb{E}_\xi \xi^2 = 1.$$

Lemma 1 *The partition function (2) can be represented in Gaussian integral form as*

$$Z_{N,P}(\sigma|\xi,t) = \sum_{\{\sigma\}} \int \Big(\prod_{\mu=1}^{P} d\mu(z_\mu)\Big)\Big(\prod_{i=1}^{N} d\mu(\phi_i)\Big) \exp\left(\sqrt{\frac{\beta}{N}(t+1)}\sum_{\mu,i}^{P,N} z_\mu \xi_i^\mu \sigma_i + i\sqrt{\frac{t}{N}}\sum_{\mu,i}^{P,N} z_\mu \xi_i^\mu \phi_i\right), \qquad (3)$$

that is equivalent to the partition function of a tripartite spin-glass where the intermediate layer is made of a set of real neurons $\{z_\mu\}_{\mu=1,\dots,P}$ with $z_\mu \sim \mathcal{N}[0,1], \forall\mu$, while the external layers are made, respectively, of a set of Boolean neurons $\{\sigma_i\}_{i=1,\dots,N}$ and of a set of imaginary neurons with magnitude $\{\phi\}_{i=1,\dots,N}$, being $\phi_i \sim \mathcal{N}[0,1], \forall i$.

Definition 5 *Once expressed the partition function (2) in its integral representation (3), we can introduce the related tripartite spin-glass Hamiltonian as*

$$H_{N,P} = \frac{a}{\sqrt{N}} \sum_{i=1}^{N} \sum_{\mu=1}^{P} z_\mu \xi_i^\mu k_i, \tag{4}$$

where we introduced the "multi-spin" $k_i = \sigma_i + b\phi_i$ *and where*

$$a = \sqrt{\beta(t+1)}, \quad b = i\sqrt{\frac{t}{\beta(t+1)}}. \tag{5}$$

The cost function (4) and the one associated to the original model (1) share the same partition function and therefore exhibit the same thermodynamic.

Definition 6 *The natural order parameters for the neural network model (1) are the overlaps* q_{ab} *and* p_{ab} *between the k's and the z's variables, respectively, as functions of two replicas (a,b) of the system, and the generalized Mattis overlap* m_1*, namely*

$$\begin{aligned} q_{ab} &:= \tfrac{1}{N} \textstyle\sum_{i=1}^{N} k_i^{(a)} k_i^{(b)}, \\ p_{ab} &:= \tfrac{1}{P} \textstyle\sum_{\mu \geq 2} z_\mu^{(a)} z_\mu^{(b)}, \\ m_1 &:= \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \xi_i^1 k_i. \end{aligned}$$

Remark 2 *RS approximation* $\rightarrow$ *order-parameters do not fluctuate in the thermodynamic limit,*

$$\begin{aligned} q_{ab} &\overset{\text{RS}}{\rightarrow} W\delta_{ab} + q(1-\delta_{ab}), \\ p_{ab} &\overset{\text{RS}}{\rightarrow} X\delta_{ab} + p(1-\delta_{ab}), \\ m_1 &\overset{\text{RS}}{\rightarrow} m. \end{aligned}$$

Theorem 1 *In the infinite volume limit, the replica symmetric free energy related to the neural network (1) can be expressed in terms of the natural order parameters of the theory as*

$$\begin{aligned} A_{RS}(\alpha,\beta,t) &= \log 2 - \frac{\beta m^2}{2(1+t)}\Big(1+\frac{t}{\Delta}\Big) - \frac{(1+t)(\Delta-1)}{2t}\beta W + \mathbb{E}_\eta \log\cosh\Big[\frac{\beta}{\Delta}(m+\sqrt{\alpha p}\eta)\Big] - \frac{\log\Delta}{2} - \frac{\alpha\beta p t}{2(1+t)\Delta} \\ &- \frac{\alpha}{2}\Big(\log[1-\beta(1+t)(W-q)] + \frac{q\beta^2(1+t)}{1-\beta(1+t)(W-q)}\Big) - \frac{(1+t)(1-\Delta)\beta}{2t\Delta} - \frac{\alpha\beta^2}{2}p(W-q). \end{aligned} \tag{6}$$

Proposition 1 *Using the standard variational principle $\vec{\nabla} A_{RS} = 0$ on the free energy (6), namely by extremizing the latter over the order parameters, we obtain the following set of self-consistent equations*

$$\begin{aligned} m &= \frac{1+t}{\Delta+t}\mathbb{E}_\eta \tanh\Big[\frac{\beta}{\Delta}(m+\sqrt{\alpha p}\eta)\Big], \\ p &= \frac{q(1+t)^2}{[1-\beta(1+t)(W-q)]^2}, \\ \Delta &= 1+\frac{\alpha t}{1-\beta(1+t)(W-q)}, \\ q &= W + \frac{t}{\beta(1+t)\Delta} - \frac{1}{\Delta^2}\mathbb{E}_\eta \cosh^{-2}\Big[\frac{\beta}{\Delta}(m+\sqrt{\alpha p}\eta)\Big], \\ W\Delta^2 &= 1 - \frac{t\Delta}{\beta(1+t)} + \frac{\alpha p t^2 - m^2 t(t+2\Delta)}{(1+t)^2} - \frac{2\alpha\beta p t}{(1+t)\Delta}\mathbb{E}_\eta \cosh^{-2}\Big[\frac{\beta}{\Delta}(m+\sqrt{\alpha p}\eta)\Big]. \end{aligned}$$

The proof is based on Guerra's interpolating techniques…

Definition 6 *Being $s \in [0,1]$ an interpolating parameter, $\{\eta_i\}_{i\in(1,\dots,N)}$ a set of N i.i.d. Gaussian variables, $\{\lambda_\mu\}_{\mu\in(2,\dots,P)}$ a set of $P-1$ i.i.d. Gaussian variables, and the scalars C_1, C_2, C_3, C_4, C_5 to be set a posteriori, we use as interpolating pressure the following quantity*

$$\begin{aligned}\mathcal{A}(s) &= \frac{1}{N}\mathbb{E}_{\xi,\eta,\lambda}\ln\sum_{\sigma}\int d\mu(z,\phi)\exp\Big[\sqrt{s}\frac{a}{\sqrt{N}}\sum_{i,\mu\geq 2} z_\mu \xi_i^\mu k_i + \sqrt{s}\frac{a}{\sqrt{N}}\sum_i z_1\xi_i^1 k_i \\ &+ \sqrt{1-s}\Big(C_1\sum_i^N \eta_i k_i + C_2\sum_{\mu\geq 2}\lambda_\mu z_\mu\Big) + \frac{1-s}{2}\Big(C_3\sum_{\mu\geq 2} z_\mu^2 + C_4\sum_i k_i^2 + C_5 a\sum_i \xi_i^1 k_i\Big)\Big].\end{aligned} \tag{7}$$

When $s=1$ we recover the original model, namely $A(\alpha,\beta,t) = \lim_{N\to\infty}\mathcal{A}(s=1)$, while for $s\to 0$ we are left with a one-body problem, and, consequently, the probabilistic structure of $\mathcal{A}(s=0)$ is more tractable.

Proposition 2 *The infinite volume limit of the quenched free energy related to the model (1) can be obtained by using the Fundamental Theorem of Calculus as*

$$A(\alpha,\beta,t) = \lim_{N\to\infty}\mathcal{A}(s=1) = \lim_{N\to\infty}\left(\mathcal{A}(s=0) + \int_0^1 \frac{d\mathcal{A}(s)}{ds}\,ds\right).$$

Performing calculations and setting the free scalars $C_{1,..,5}$ as

$$C_1^2 = a^2\alpha p, \quad C_2^2 = a^2 q, \quad C_3 = a^2(W-q), \quad C_4 = a^2\alpha(X-p), \quad C_5 = 2ma, \tag{8}$$

one gets that under RS $d_s\mathcal{A}(s)$ is independent of s and calculations straightforwardly yields to we finally get (6).

Theorem 2 *The ergodic region of the model defined by the cost function (**??**) is delimited by the following critical surface in the (α, β, t) space of the tunable parameters*

$$\begin{aligned} \beta_c &= \frac{1}{1+t}\Big[\frac{\Delta^2}{1+\sqrt{\alpha}} + t\Delta\Big], \\ \Delta &= 1+\sqrt{\alpha}(1+\sqrt{\alpha})t. \end{aligned}$$

Remark 3 *At $t = 0$, where the model reduces to Hopfield's scenario, the critical surface correctly collapses over the Amit-Gutfreund-Sompolinsky critical line $\beta_c = (1+\sqrt{\alpha})^{-1}$, but in the large t limit the ergodic region collapses to the axis $T = 0$: we speculate that this may have a profound implications, namely that the ergodic region -during the sleep state- invades the spin-glass region, the latter being de facto suppressed.*

The proof is again based on Guerra's interpolating techniques...

Definition 7 *The centered and rescaled overlap fluctuations θ_{lm} and ρ_{lm} are introduced as*

$$\begin{aligned} \theta_{lm} &= \sqrt{N}\big[q_{lm} - \delta_{lm}W - (1-\delta_{lm})q\big] \\ \rho_{lm} &= \sqrt{P}\big[p_{lm} - \delta_{lm}X - (1-\delta_{lm})p\big]. \end{aligned}$$

Remark 4 *As we will address the problem of the overlap fluctuations in the ergodic region, the signal is absent, thus there is no need to introduce a rescaled Mattis order parameter.*

Proposition 2 *We introduce the $r-$replicated interpolating free energy $\mathcal{A}_J^r(s)$, where we further added a source field J, coupled to an observable O (that is a smooth function of the neurons of the r-replicas) as*

$$\mathcal{A}_J^r(s) = \mathbb{E}_{\xi,\eta,\lambda} \ln \sum_{\sigma_R} \int d\mu(z_R, \phi_R) \exp \Big[\sqrt{s} \frac{a}{\sqrt{N}} \sum_{l=1}^{r} \sum_{i,\mu} z_\mu^{(l)} \xi_i^\mu k_i^{(l)} + J\hat{O} \tag{9}$$

$$+ \sqrt{1-s} \Big(C_1 \sum_{l=1}^{r} \sum_{i} \eta_i k_i^{(l)} + C_2 \sum_{l=1}^{r} \sum_{\mu} \lambda_\mu z_\mu^{(l)} \Big) + \frac{1-s}{2} \Big(C_3 \sum_{l=1}^{r} \sum_{\mu} (z_\mu^{(l)})^2 + C_4 \sum_{l=1}^{r} \sum_{i} (k_i^{(l)})^2 \Big) \Big]. \tag{10}$$

where k_i and the interpolation constants $C_{1,2,3,4}$ are the same given before.

One can verify that

$$\langle O(s) \rangle = \frac{\partial \mathcal{A}_J^r(s)}{\partial J}\Big|_{J=0}, \qquad \partial_s \langle O(s) \rangle = \frac{\partial (\partial_s \mathcal{A}_J^r)}{\partial J}\Big|_{J=0}. \tag{11}$$

Therefore, in order to evaluate the fluctuations of O we need to evaluate first $\partial_s \mathcal{A}_J^r$ and, by a routine calculation, we get

$$\partial_s \mathcal{A}_J^r = \frac{1}{2} \sqrt{\alpha} \beta (1+t) \sum_{l,m=1}^{r} \Big[\langle g_{l,m} \rangle - \langle g_{l,m+r} \rangle \Big], \qquad g_{l,m} = \theta_{lm} \rho_{lm}. \tag{12}$$

Proposition 3 *Given O as a smooth function of r replica overlaps $(q_1, \dots, q_r)$ and $(p_1, \dots, p_r)$, the following streaming equation holds:*

$$d_\tau \langle O \rangle = \frac{1}{2} \sum_{a,b}^{r} \langle O \cdot g_{a,b} \rangle - r \sum_{a=1}^{r} \langle O \cdot g_{a,r+1} \rangle + \frac{r(r+1)}{2} \langle O \cdot g_{r+1,r+2} \rangle - \frac{r}{2} \langle O \cdot g_{r+1,r+1} \rangle, \tag{13}$$

where we used the operator d_τ defined as

$$d_\tau = \frac{1}{\beta(1+t)\sqrt{\alpha}} \frac{d}{ds}, \tag{14}$$

in order to simplify calculations and presentation.

Definition 8 *The centered and rescaled overlap fluctuations $\xi_{l,m}$ and $\rho_{l,m}$ are introduced as*

$$\begin{aligned}\xi_{l,m} &= \sqrt{N}\,[q_{l,m} - \delta_{l,m}W - (1-\delta_{l,m})q] \\ \rho_{l,m} &= \sqrt{P}\big[p_{l,m} - \delta_{l,m}X - (1-\delta_{l,m})p\big].\end{aligned}$$

Posing $Y(s) = \langle \xi_{12}\rho_{12}\rangle_s + \frac{\langle \xi_{12}^2\rangle_0}{\langle \rho_{12}^2\rangle_0}\langle \rho_{12}\rangle_s$ we get $d_\tau Y = Y^2$, whose solution evaluated at $\tau = \beta(1+t)\sqrt{\alpha}s$, $s = 1$ is given by

$$\begin{aligned}Y(s=1) &= \tfrac{W}{1-\beta(1+t)W(1+\sqrt{\alpha})}, \\ W\Delta^2 &= 1 - \tfrac{t\Delta}{\beta(1+t)}, \\ \Delta &= 1 + \tfrac{\alpha t}{1-\beta(1+t)W}.\end{aligned} \tag{15}$$

Since we are interested in obtaining the critical temperature for ergodicity breaking, where fluctuations (in this case Y) grow arbitrarily large we check where the denominator at the r.h.s. of eq. (15) vanishes.

Theorem 2 *The ergodic region of the model defined by the cost function (1) is delimited by the following critical surface in the (α, β, t) space of the tunable parameters*

$$\begin{aligned}\beta_c &= \frac{1}{1+t}\Big[\frac{\Delta^2}{1+\sqrt{\alpha}} + t\Delta\Big], \\ \Delta &= 1 + \sqrt{\alpha}(1+\sqrt{\alpha})t.\end{aligned}$$

Remark 4 *At $t = 0$, where the model reduces to Hopfield's scenario, the critical surface correctly collapses over the Amit-Gutfreund-Sompolinsky critical line $\beta_c = (1+\sqrt{\alpha})^{-1}$, but -in the deep sleep phase (i.e. in the large t limit)- the ergodic region collapses to the axis $T = 0$: the ergodic region -during the sleep state- invades the spin-glass region, the latter being de facto suppressed.*

Dreaming Boltzmann machine

Boltzmann Machines with “large” hidden-layer size

$$\mathscr{H}_{N,K}^{(RR-HN)}(\boldsymbol{\sigma};\boldsymbol{\xi},t) = -\frac{1}{2N}\sum_{i,j=1}^{N,N}\sum_{\mu,\nu=1}^{K,K}\xi_i^\mu\,\xi_j^\nu\left(\frac{1+t}{\boldsymbol{I}_K+t\boldsymbol{C}}\right)_{\mu,\nu}\sigma_i\,\sigma_j$$

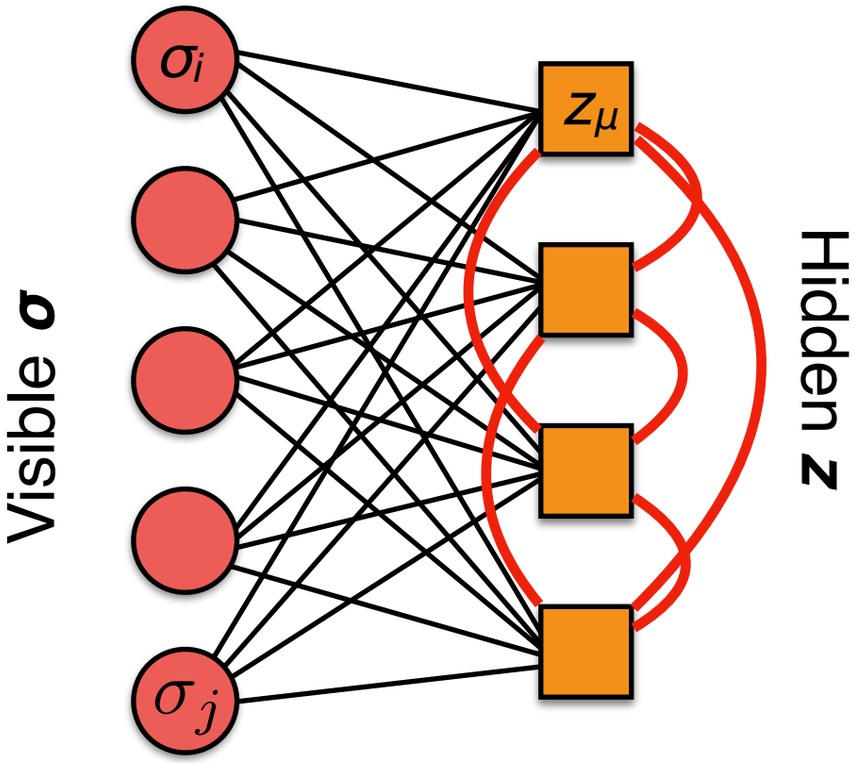
HN with Reinforcement&Removal

$$\mathscr{H}_{N,K}^{(RR-BM)}(\boldsymbol{\sigma},\boldsymbol{z};\boldsymbol{w},t) = -\frac{1}{\sqrt{N}}\sum_{i,\mu} w_i^\mu\sigma_i z_\mu + \frac{t}{2(1+t)}\sum_{\mu,\nu} C_{\mu,\nu}\;z_\mu z_\nu$$

BM with Reinforcement&Removal

$$C_{\mu\nu} := \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\xi_j^\nu$$

$$z_\mu \sim \mathscr{N}(0,[\beta/(1+t)]^{-1})$$



EA, F. Alemanno, A. Barra, A. Fachechi,
Dreaming Neural Networks for learning (2021)

Boltzmann Machines with “large” hidden-layer size

$$\mathscr{H}_{N,K}^{(RR-HN)}(\boldsymbol{\sigma};\boldsymbol{\xi},t) = -\frac{1}{2N}\sum_{i,j=1}^{N,N}\sum_{\mu,\nu=1}^{K,K}\xi_i^\mu\,\xi_j^\nu\left(\frac{1+t}{\boldsymbol{I}_K+t\boldsymbol{C}}\right)_{\mu,\nu}\sigma_i\,\sigma_j$$

HN with Reinforcement&Removal

$$\mathscr{H}_{N,K}^{(RR-BM)}(\boldsymbol{\sigma},\boldsymbol{z};\boldsymbol{w},t) = -\frac{1}{\sqrt{N}}\sum_{i,\mu} w_i^\mu\sigma_i z_\mu + \frac{f(t)}{2[1+f(t)]}\sum_{\mu,\nu} P_{\mu,\nu}\,z_\mu z_\nu$$

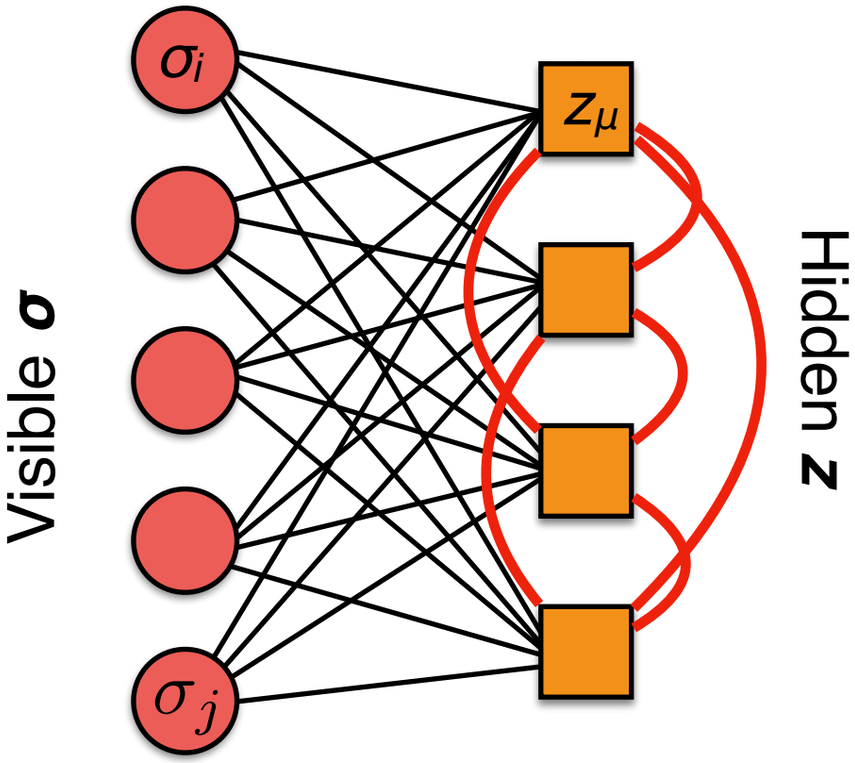
BM with Reinforcement&Removal

P empirical Pearson correlation matrix

$$P_{\mu\nu} = \frac{\frac{1}{N}\sum_{i=1}^{N}(W_{i\mu}-\bar{W}_{i\mu})(W_{i\nu}-\bar{W}_{i\nu})}{\sqrt{(\frac{1}{N}\sum_{i=1}^{N}W_{i\mu}^2-\bar{W}_{i\mu}^2)(\frac{1}{N}\sum_{i=1}^{N}W_{i\nu}^2-\bar{W}_{i\nu}^2)}}$$

$f(t)$ s.t. $f(0)=0$ and $[\boldsymbol{I}_K+f(t)\boldsymbol{P}]$ is pos. def. (e.g., $f(t)=|t|$)

$$z \sim \mathscr{N}(0,\boldsymbol{I}_K(\beta/[1+f(t)])^{-1})$$



EA, F. Alemanno, A. Barra, A. Fachechi,
Dreaming Neural Networks for learning (2021)

Boltzmann Machines with “large” hidden-layer size

$$\mathscr{H}^{(RR-HN)}_{N,K}(\boldsymbol{\sigma};\boldsymbol{\xi},t) = -\frac{1}{2N}\sum_{i,j=1}^{N,N}\sum_{\mu,\nu=1}^{K,K}\xi_i^\mu\,\xi_j^\nu\left(\frac{1+t}{\boldsymbol{I}_K+t\boldsymbol{C}}\right)_{\mu,\nu}\sigma_i\,\sigma_j$$

HN with Reinforcement&Removal

$$\mathscr{H}^{(RR-BM)}_{N,K}(\boldsymbol{\sigma},\boldsymbol{z};\boldsymbol{w},t) = -\frac{1}{\sqrt{N}}\sum_{i,\mu} w_i^\mu\sigma_i z_\mu + \frac{f(t)}{2[1+f(t)]}\sum_{\mu,\nu} P_{\mu,\nu}\,z_\mu z_\nu$$

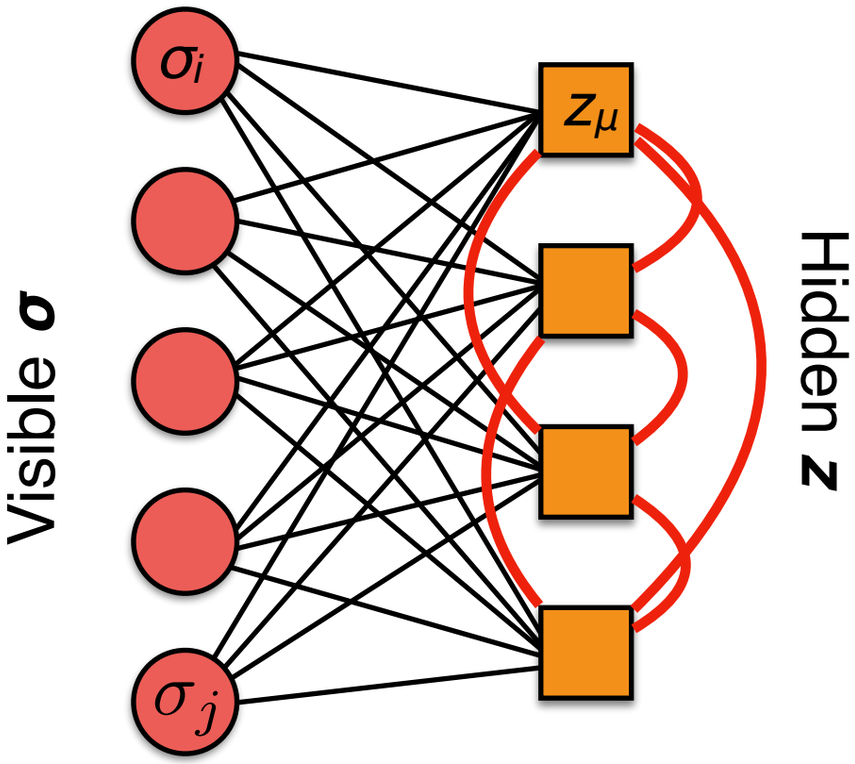
BM with Reinforcement&Removal

P empirical Pearson correlation matrix

$$P_{\mu\nu} = \frac{\frac{1}{N}\sum_{i=1}^{N}(W_{i\mu}-\bar{W}_{i\mu})(W_{i\nu}-\bar{W}_{i\nu})}{\sqrt{(\frac{1}{N}\sum_{i=1}^{N}W_{i\mu}^2-\bar{W}_{i\mu}^2)(\frac{1}{N}\sum_{i=1}^{N}W_{i\nu}^2-\bar{W}_{i\nu}^2)}}$$

$f(t)$ s.t. $f(0)=0$ and $[\boldsymbol{I}_K+f(t)\boldsymbol{P}]$ is pos. def. (e.g., $f(t)=|t|$)

$$z \sim \mathscr{N}(0,\boldsymbol{I}_K(\beta/[1+f(t)])^{-1})$$



By gradient descent over KL divergence

$$\Delta w_{j,\rho} = \epsilon\left[\langle\sigma_j z_\rho\rangle_+ - \langle\sigma_j z_\rho\rangle_- - \frac{N}{2}\frac{f(t)}{1+f(t)}\sum_{\mu,\nu}\frac{\partial P_{\mu,\nu}}{\partial w_{j,\rho}}\left(\langle z_\mu z_\nu\rangle_+ - \langle z_\mu z_\nu\rangle_-\right)\right]$$

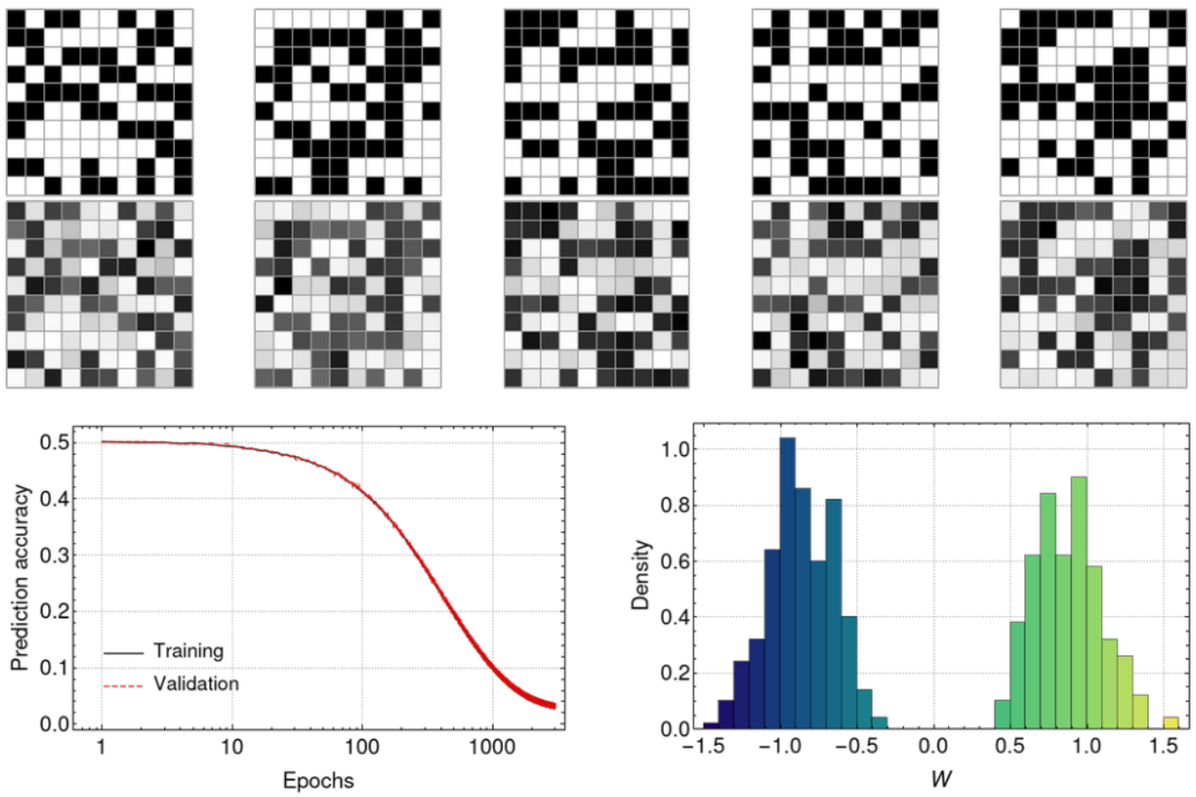
$$\Delta t = \epsilon\frac{df}{dt}\frac{N}{2(1+f(t))^2}\sum_{\mu,\nu}(\boldsymbol{I}_K-\boldsymbol{P})_{\mu,\nu}\left(\langle z_\mu z_\nu\rangle_+ - \langle z_\mu z_\nu\rangle_-\right)$$

$\langle\cdot\rangle_+$ clamped average

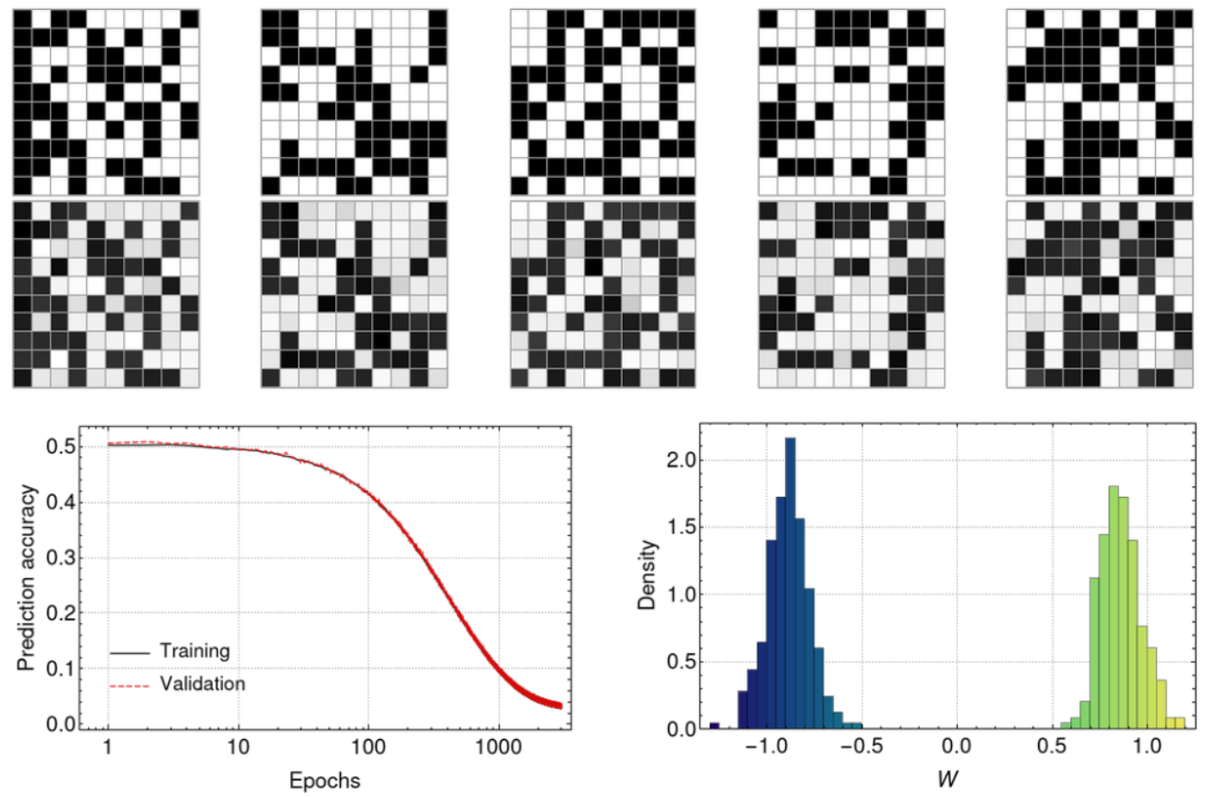
$\langle\cdot\rangle_-$ free average

Sub-critical regime ($N = 100,\ K = 5,\ T = 0.2$)

RBM

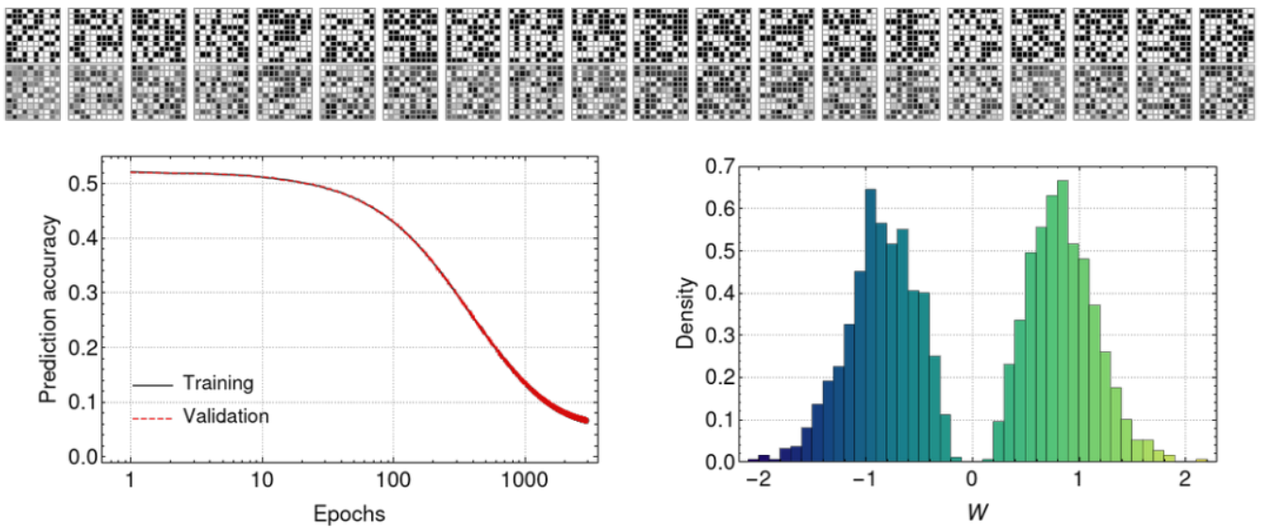


R&R-BM

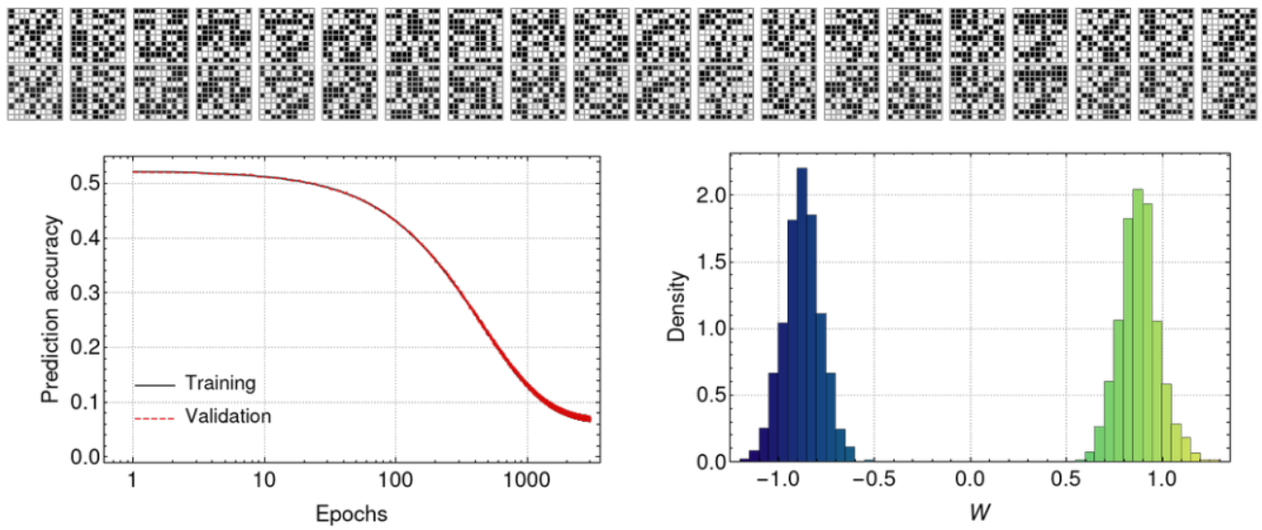


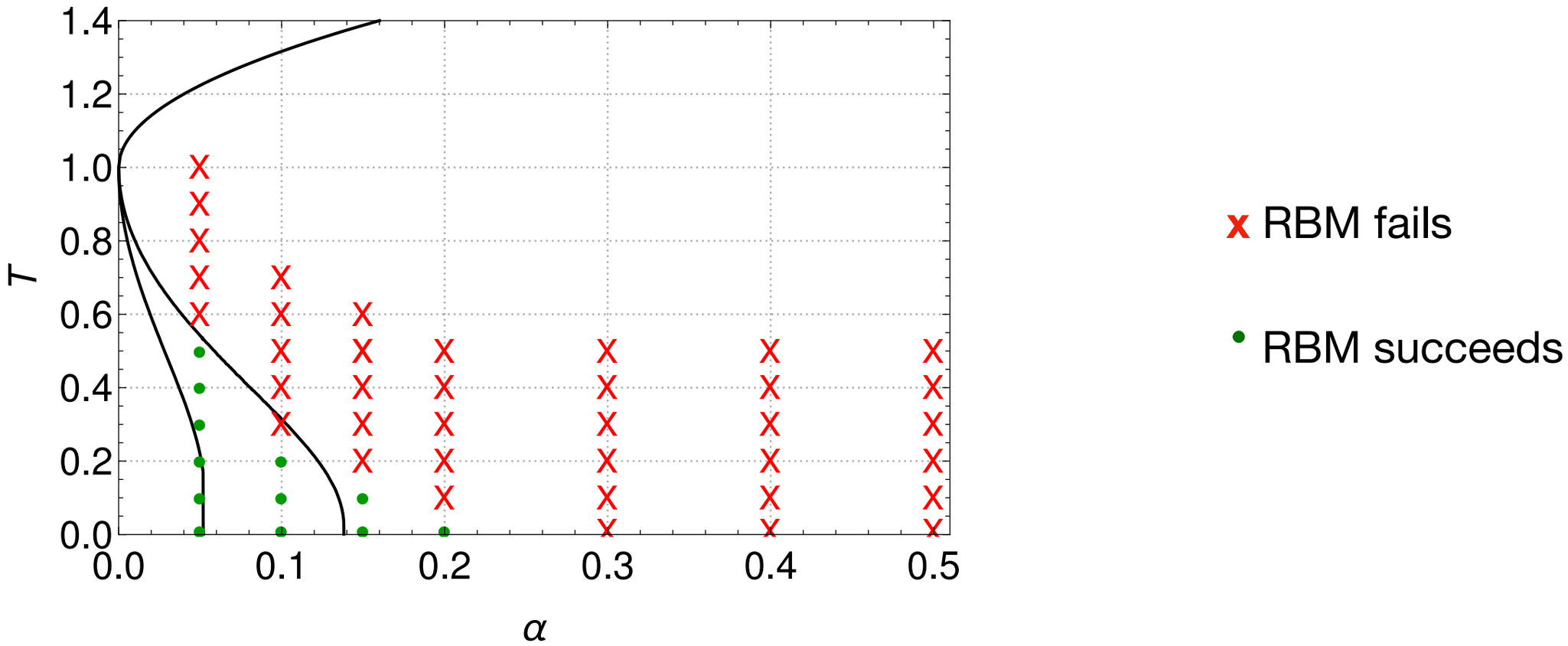
Super-critical regime ($N = 100,\ K = 20,\ T = 0.2$)

RBM

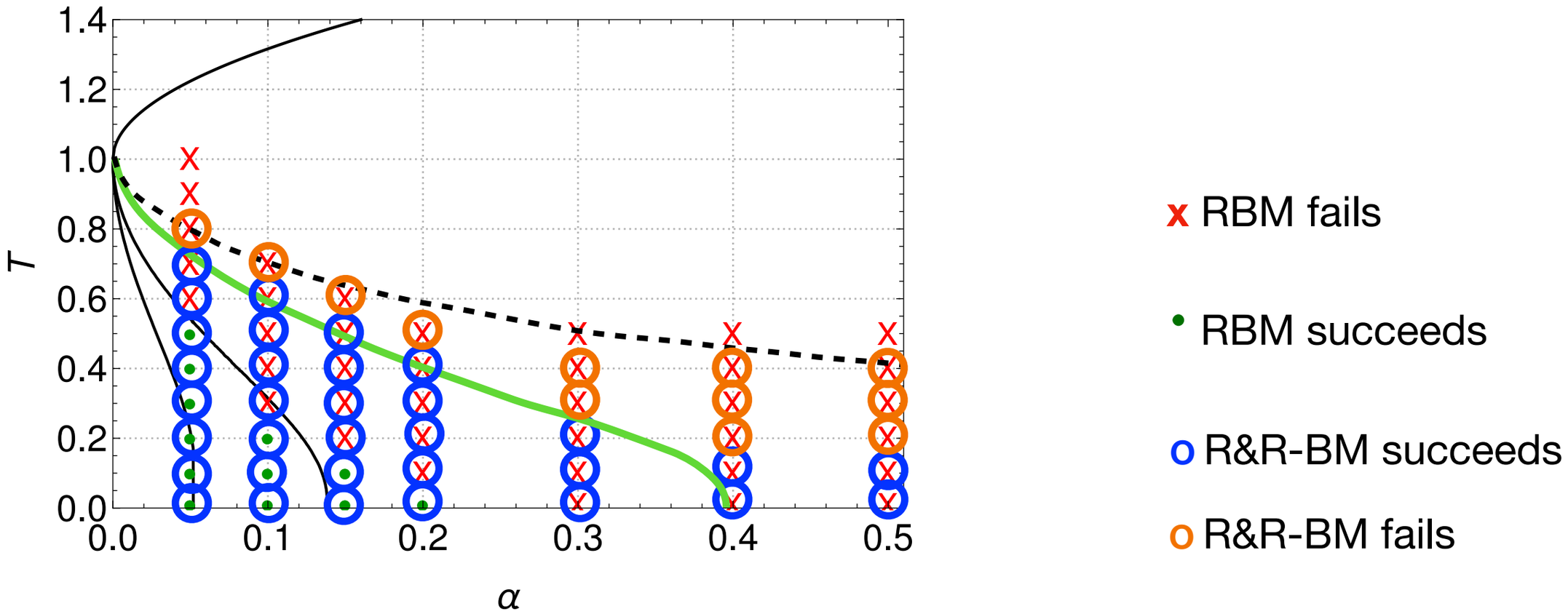


R&R-BM

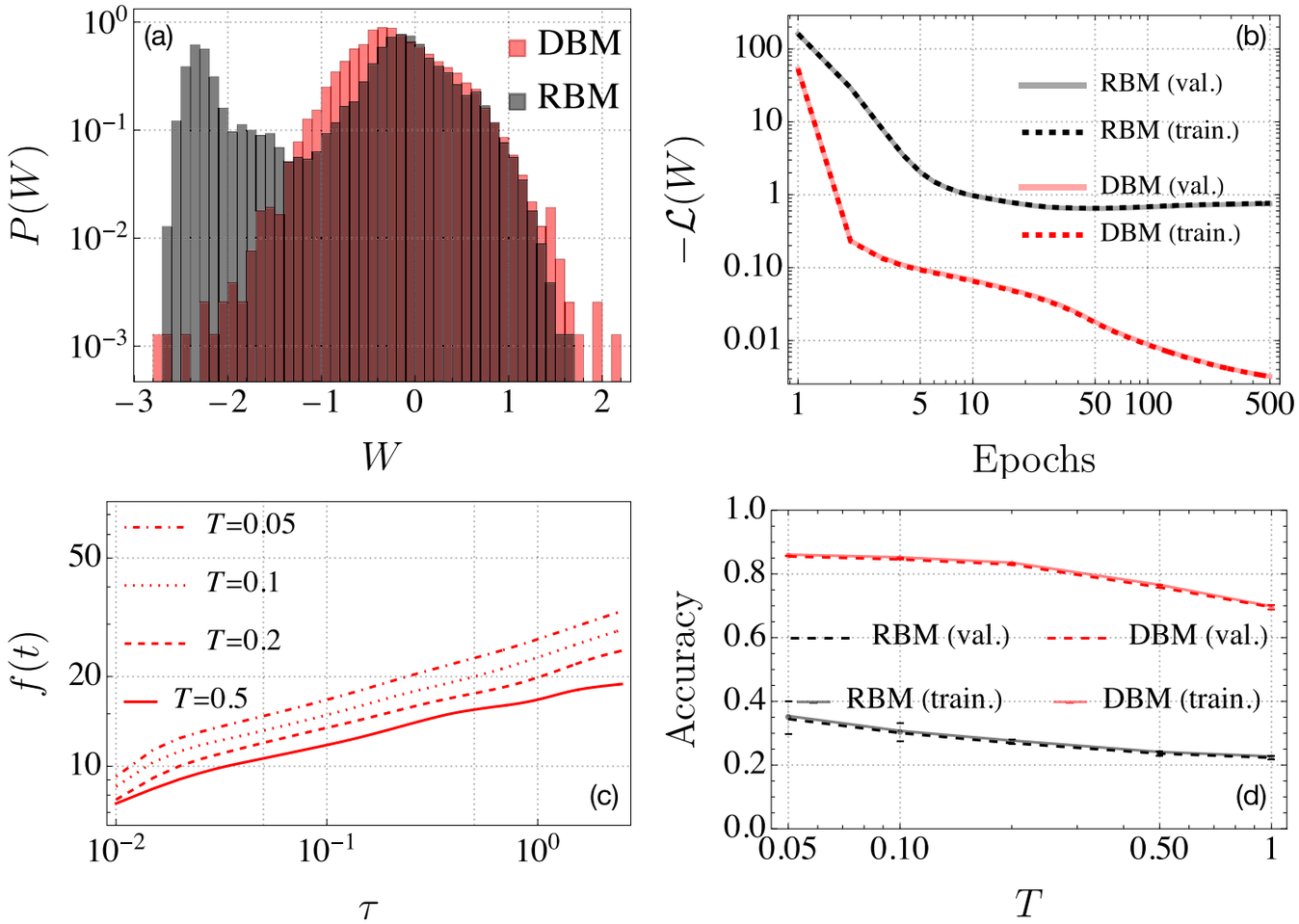




N= 100, 200, 400, CD-1, ϵ = 0.0005 T , p= 0.05 example noise, M= 600 examples per pattern (M_t= 500 for training, M_v= 100 for validation), M_b= 200 minibatch size, τ = 3000 epochs, network parameter initial condition $w_{i,\mu} \sim \mathcal{N}(0,1)$, supervised.

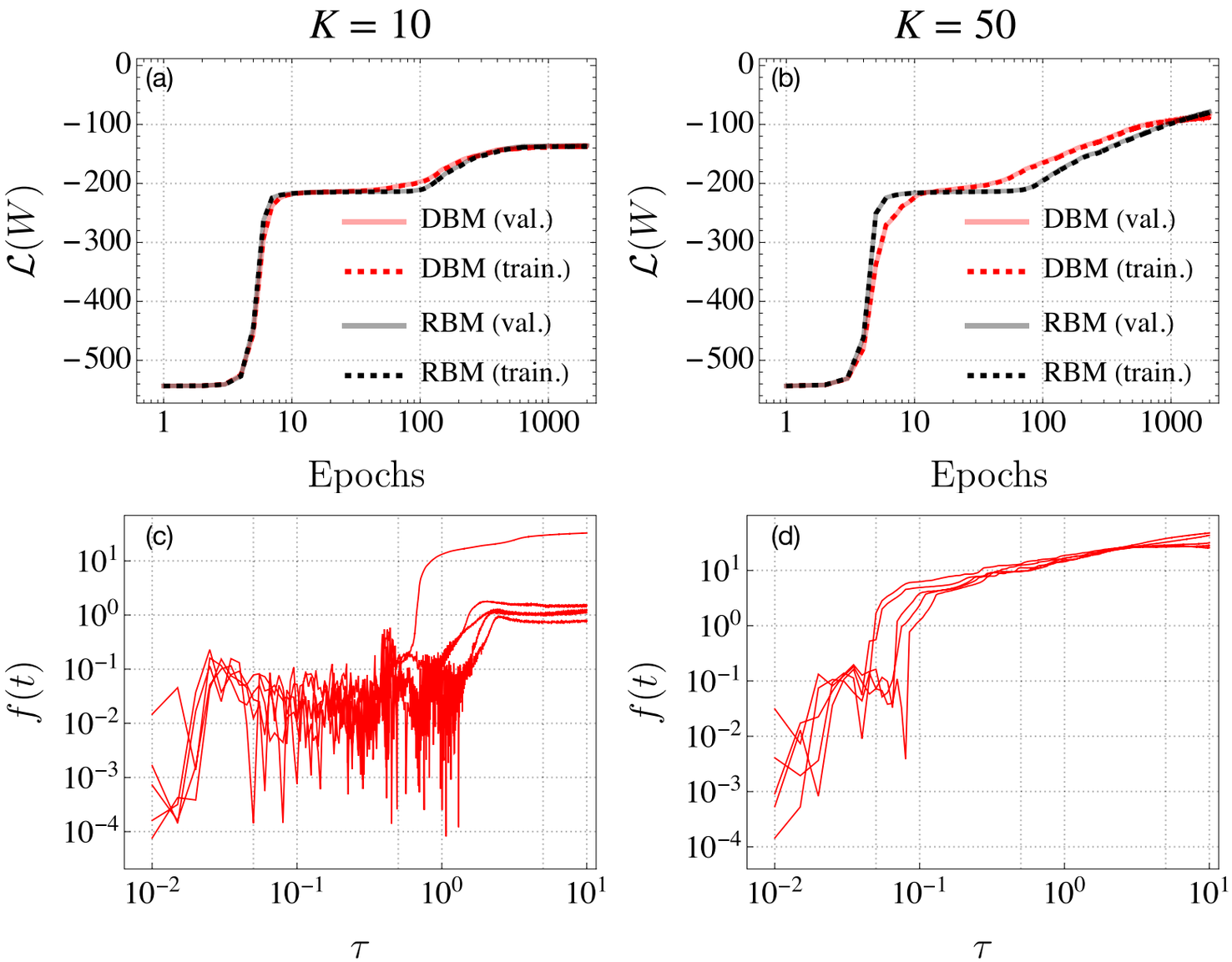


N= 100, 200, 400, CD-1, ϵ = 0.0005 T, p= 0.05 examples noise, M= 600 examples per pattern (M_t= 500 for training, M_v= 100 for validation), M_b= 200 minibatch size, τ = 3000 epochs, network parameter initial condition $w_{i,\mu} \sim \mathcal{N}(0,1)$.



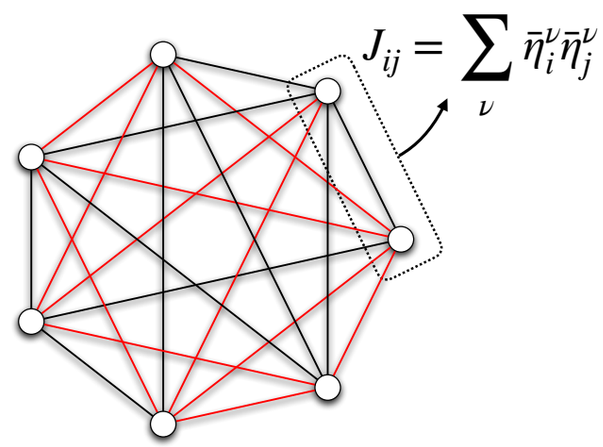
Results for the MNIST dataset with supervised training and random initialization of the weights. (a) the pseudo-(log)likelihood, averaged over five different realizations of (b) training procedure, (c) dreaming time $f(t)$ as a function of the learning time τ, and (d) the final accuracy as a function of the temperature, computed after 100 updates of the whole network. (e) and (f) First four digits, respectively, for the RBM and the DBM. The learning parameters are $\epsilon = 0.005T$, the size of the minibatch is fixed to 200 examples for epoch.

The graphical representation of the RBM weights is surrounded by a dark halo, whose origin lies in the fact that none of the examples exhibits activated pixels in that area. This means that these sites would contribute positively to the Mattis magnetizations, thus increasing the signal for all the hidden units. This does not happen for the DBM where entries corresponding to the external part of patterns are randomly oriented; hence, their contribution to the signal is lower w.r.t. the RBM case.

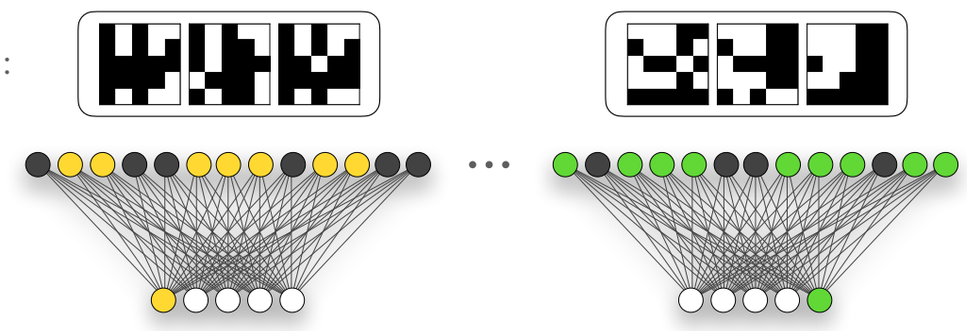


Learning results for the MNIST dataset with unsupervised training and random initialization of the weights. The plots show the results for the pseudo-(log)likelihood (a) and (b) averaged over five different realizations of the training procedure, and the dreaming time as a function of the learning time (c) and (d) for each of the five different learning sessions. The initial condition is $W_{i\mu} \sim \mathcal{N}(0,0.1)$. The learning strength is $\epsilon = 0.005T$, the temperature is $T = 1$, the size of the minibatch is fixed to 200 examples, and the number of epochs is 2000.

t as a free parameter: it smoothly interpolates between the RBM and the DBM models; thus, the minima landscape of the objective function will be modified in a continuous way



Packets of examples :



$\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_N)$

$\boldsymbol{z} = (z_1, \cdots, z_K)$