HEBBIAN NEURAL NETWORKS

The most famous example is HN, but we will see also other examples.
Start with a basic introduction on HN where I will recall its bio inspiration
Next, I will present some variations on the theme evidencing advantages and drawbacks
Next I will present an analytical method to tackle these models which is based on Guerra's interpolation.
This method suggests a bridge between associative memories and simple models for ML, specifically BM
I will then review BMs and leverage the knowledge available for HNs to get some strategies for an effective training.
Finally, I will provide a brief overview on modern architecture for hebbian networks, specifically dense and modular.
Our journey will span roughly 80 years.

Many excellent textbooks and reviews available.
For instance,

[B] Bovier, "Statistical Mechanics of Disordered Systems", CUP 2006

[CKS] Coolen, Kühn, Sollich, "Theory of Neural Information processing Systems", OUP 2005

[BG] Bovier, Gayrard, "Hopfield model as generalized random mean-field models", book chapter 1997

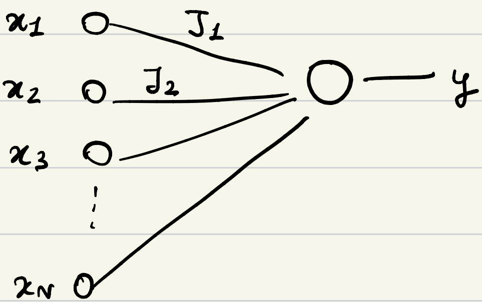
[P] Picco, AHP 1996

MC-CULLOCH - PITTS NEURON '43

(neurophysiologist - logician)

Mimics the threshold mechanism underlying the activation of a biological neuron

Schematically



$x_i \in \{0, 1\} \quad i = 1, \dots, N$

$y \in \{0, 1\}$

$J_i \in \mathbb{R}$

$$U = \sum_{k=1}^{N} J_k x_k$$

$$y = \Theta(U - U^*) \quad , \quad U^* \in \mathbb{R} \quad \text{firing threshold}$$

"Static" version of the Rosenblatt perceptron, where I keep synaptic weights quenched. Even if we are not interested in how the weights should be designed to reach a certain task it is important to recall that, if we interpret $\underline{x} = (x_1, \dots, x_N)$ as an input and $y \in \{0,1\}$ as an output this kind of object is very limited in the sense that the input-output relations

$M: \Omega \subseteq \{0,1\}^N \to \{0,1\}$

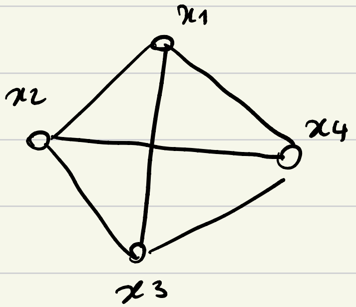
that can be recovered are only those linearly separable whatever the choice for the weights

$\exists \underline{B} \in \mathbb{R}^N$ st $\forall \underline{x} \in \Omega$

$M(\underline{x}) = \Theta(\underline{B} \cdot \underline{x})$, that $\Omega = \Omega^+ \cup \Omega^-$

st. $B\underline{x} > 0 \ (<0) \quad \forall x \in \Omega^+ \ (\Omega^-)$

On the other hand if we combine MP neurons together we can obtain a network of MP neurons whose info processing capabilities is by far increased.



$$U_i^{(t)} = \sum_{k=1}^{N} J_{ki} \, x_i^{(t)}$$

$$x_i^{(t+1)} = \Theta(U_i^{(t)} - U_i^*)$$

Given the network structure, the update of neuron i implies a revision of the signal acting on the adjacent neurons, therefore it is suitable to look at this process dynamically and introduce a discrete time unit

$U_i^{(t)} - U_i^{*} + \frac{1}{2} T z_i^{(t)}$ $\qquad T \in \mathbb{R}_0^+$

- biological meaningful $\qquad z_i^{(t)} \sim_{iid} \rho_z \qquad z$
- computationally convenient

ρ_z : $\quad \mathbb{E}[z] = 0 \qquad \mathbb{E}[\cdot] = \int \cdot \; \rho_z(z)\, dz$

$\quad\quad \mathbb{E}[z^2] = 1$

$$\underline{x} \in \{0, 1\}^N \to \underbrace{\underline{\sigma} \in \{-1, +1\}^N}_{\text{SM tools}}$$

$$x_i^{(t)} = \frac{1}{2}\left[\sigma_i^{(t)} + 1\right]$$

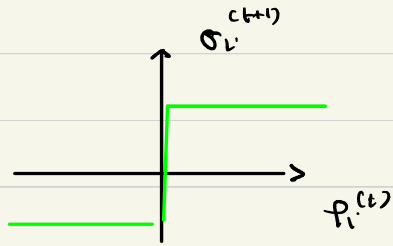
$$\boxed{\sigma_i^{(t+1)} = \mathrm{sgn}\left[\varphi_i^{(t)} + T z_i^{(t)}\right]}$$

$$\varphi_i^{(t)} \equiv \varphi_i(\underline{\sigma}^{(t)}) = \underbrace{\sum_{k=1}^{N} J_{ki}\, \sigma_i^{(t)}}_{\text{internal field}} + \underbrace{h_i}_{\text{external field}}$$

$$h_i = -2U_i^{*} + \sum_k J_{ki}$$

$T = 0$ noiseless (deterministic) dynamics

$\varphi_i^{(t)} > 0 \to \sigma_i^{(t+1)} = +1$



NEURONAL DYNAMICS

[CKS]

$T > 0$ noisy (stochastic) dynamics

$T \to \infty \qquad Med(Z) = 0 \qquad \sigma_i^{(t)} \sim Rad(0) \quad \forall i, t$

<u>Def</u> $\quad X \sim Rad(p)$

$$P(x = +1) = \frac{1+p}{2} = 1 - P(x = -1), \quad p \in [-1, +1]$$

<u>Hp</u> $\quad \rho_Z$ symmetric

$$g_Z(z) := \int_{-\infty}^{z} \rho_Z(u)\, du$$

$$\text{Prob}\left[\sigma_i^{(t+1)} = \pm 1\right] = g\left(\pm \frac{\varphi_i^{(t)}}{T}\right)$$

HW

<u>Hp</u> $\quad g_Z(z) = \frac{1}{2}\left[1 + \tanh(z)\right]$

$$\rho_Z(z) = \frac{1}{2}\left[1 - \tanh^2(z)\right]$$

$$\mathcal{N}(0,1) \longrightarrow \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right]$$

sequential (asynch) $\begin{cases} \text{select } i \\ \text{Prob}\left[\sigma_i^{(t+1)} = \pm 1\right] = g\left(\pm \varphi_i^{(t)}/T\right) \end{cases}$ ←

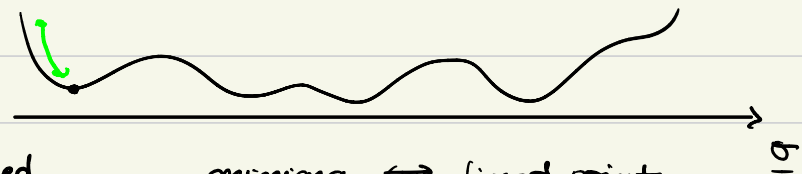
parallel (asynch.) $\text{Prob}\left[\underline{\sigma}^{(t)}\right] = \prod_i g\left(\sigma_i^{(t)} \varphi_i^{(t)} / T\right)$

<u>Prop</u> If $J = J^T$, $\text{diag}(J) \geq 0$, $\underline{h}$ stationary

$$L_N(\underline{\sigma}; J, h) = -\frac{1}{2} \underline{\sigma}^T J \underline{\sigma} - \underline{h}^T \underline{\sigma} \quad \text{Lyapunov function}$$

$$= -\frac{1}{2} \sum_{i,j} \sigma_i \sigma_j J_{ij} - \sum_i h_i \sigma_i$$

$\Delta L_N \leq 0$



$L(\underline{\sigma})$ lower-bounded minima ⟷ fixed points

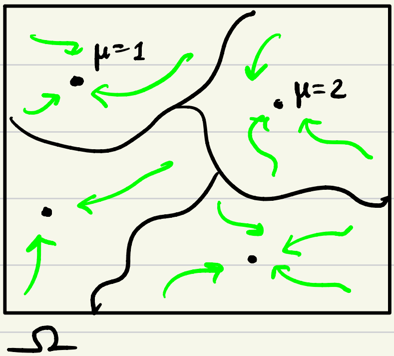
$\exists\, t^* \in \mathbb{R}$ st $L(\sigma^{(t+1)}) = L(\underline{\sigma}^{(t)}) \quad \forall\, t \geq t^*$ [CKS]

input $\underline{\sigma}^{(0)}$ → ☐ → output $\underline{\sigma}^{(\infty)}$ $\quad \underline{\sigma}^{(\infty)}(\underline{\sigma}^{(0)})$

Design the machine $(J, \underline{h})$ st a suitable input-output relation is reproduced.

ATTRACTOR NEURAL NETWORKS

this kind of mechanism justifies the name of this class of NNs



Project config. on a 2D space

Detect fixed points and denote them as $\underline{\xi}^\mu$ with $\mu=1,\dots,K$

$\underline{\xi}^\mu \in \{-1,+1\}^N$

We can also distinguish the related attraction basins denoted as $\mathcal{B}_\mu$, that is

$$\mathcal{B}_\mu \ni \underline{\sigma}^{(0)} \rightarrow \underline{\sigma}^{(\infty)}(\underline{\sigma}^{(0)}) = \underline{\xi}^\mu$$ doomed to end up

Typical tasks as pattern reconstruction/denoising/retrieval

i.e. $\underline{\sigma}^{(0)}$ corrupted / partial version of $\underline{\xi}^\mu$

Associative memories

To this purpose, we need to design a network s.t.

- host many attractors
- non-zero basins
- trainable (easy algorithm, fixed points suitably located)

task : retrieval

$\underline{\xi}^\mu$ meant as vectors encoding information

(eg b/w picture $\rightarrow$ entries as pixels)

to be retrieved starting with a suitable input,

therefore $\underline{\xi}^\mu$ must correspond to the minima of $L_N(\underline{\sigma}; J, \underline{h})$

and this requires that J_{ij}'s must depend on the $\underline{\xi}$'s

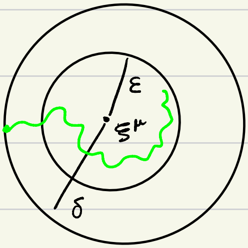
Def Retrieval

we say that a n.n. with $J = J(\underline{\xi})$ retrieves a pattern $\underline{\xi}^\mu$

if $\underline{\xi}^\mu$ is (δ, ε)-stable, $\varepsilon, \delta > 0$

$$d_H(\underline{\sigma}^{(0)}, \underline{\xi}^\mu) \le \delta \quad \text{then} \quad d_H(\underline{\sigma}^{(t)}, \underline{\xi}^\mu) < \varepsilon$$

A fixed point is just a $(\bar{0}, 0)$-stable state



Crucial $\quad J = J(\{\underline{\xi}_\mu\}_{\mu=1,\dots,k})$

Simple task: $\underline{\xi}^\mu$'s perfectly available $\rightarrow$ do not need training routines but choose J based on some bio-inspired approach or by construction requiring that the task is fulfilled.

– bio-inspired $\rightarrow$ $\boxed{\text{Hebb's rule}}$ '49 (the organization of behavior)

theoretical postulate of which later found exp. evidence "neurons that fire..."

$$\uparrow \circ \overset{+}{\longrightarrow} \circ \uparrow$$

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

time scale

synapses are much slower than neurons

$J^{(0)} = 0$ tabula rasa (null matrix)

$$J_{ij}^{(0)} \rightarrow J_{ij}^{(1)} = J_{ij}^{(0)} + \xi_i^1 \xi_j^1$$

$$J_{ij}^{(1)} \rightarrow J_{ij}^{(2)} = J_{ij}^{(1)} + \xi_i^2 \xi_j^2$$

additive

$J = J^T$ not biological : from this point we will abandon the biological pathway

– local

– simple :)

Does this rule work?

Two short answers $\rightarrow$ it is worth going on

- $\underline{\xi}^{\nu} \perp \underline{\xi}^{\mu}$ $\quad \mu \neq \nu$ $\qquad \frac{1}{N} (\underline{\xi}^{\nu})^{T} \underline{\xi}^{\mu} = \delta_{\mu\nu}$ $\quad k \leq N$

$J_{ij} = \frac{1}{N} (\underline{\xi} \cdot \underline{\xi}^{T})_{ij}$ makes Lyapunov function lin. ext. wrt N $\qquad (\underline{h} = \underline{0})$

$L_N(\underline{\sigma}; J) \geq L_N(\pm \underline{\xi}^{\mu}; J)$ HW

$\underline{\sigma} = \underline{\xi}^{\mu}$ $\quad \mu = 1, \dots, k$ F.P.

- $k = 1$

$J_{ij} = \frac{1}{N} \xi_i^1 \xi_j^1$ $\qquad \underline{h} = \underline{0}$ $\qquad \xi_i^{\mu} \sim_{iid} \text{Rad}(0)$

$L_N(\underline{\sigma}; J) = -\frac{1}{2N} \sum_{ij} \sigma_i \xi_i^1 \xi_j^1 \sigma_j$ $\qquad$ Mattis model '76

$\tilde{\sigma}_i = \sigma_i \xi_i^1$

$= -\frac{1}{2N} \sum_{ij} \tilde{\sigma}_i \tilde{\sigma}_j$ $\qquad$ CW model

$\geq L_N(\underline{\xi}^1; J)$ $\qquad$ global minima

in analogy with FM system $\rightarrow$ magnetization

<u>Def</u> Mattis magnetization

$$m_1 := \frac{1}{N} \sum_i \xi_i^1 \sigma_i \in [-1, +1]$$

to be extended as $m_{\mu} := \frac{1}{N} \sum_i \xi_i^{\mu} \sigma_i$

$$L_N(\underline{\sigma}; J) = -\frac{N}{2} m_1^2 \geq -\frac{N}{2} \qquad m_1 = \pm 1$$

energy function

- computational approach $\rightarrow$ $\boxed{\text{kohonen's projection rule}}$ '70s

no biological bias

J st $\quad \sum_j J_{ij} \xi_j^\mu = \lambda \xi_i^\mu \qquad \lambda > 0 \quad \forall i$

$$\Rightarrow \quad \underline{\sigma}^{(0)} = \underline{\xi}^\mu$$
$$\underline{\sigma}^{(1)} = \mathrm{sgn}(J \cdot \underline{\sigma}^{(0)}) = \mathrm{sgn}(\lambda \underline{\xi}^\mu) = \underline{\xi}^\mu$$

basically we are asking $\underline{\xi}^\mu$ eigenvector for J

in particular, if we set $\lambda = 1$

$$J_{ij} = \frac{1}{N} \sum_{\mu\nu} \xi_i^\mu C^{-1}_{\mu\nu} \xi_j^\nu \qquad \text{HW}$$

$$C_{\mu\nu} := \frac{1}{N} \sum_{i=1}^N \xi_\nu^\mu \xi_i^\nu \qquad K \le N$$

J idempotent $\qquad$ HW

$$\sum_j J_{ij} \sigma_j = \frac{1}{N} \sum_j \sum_\mu \xi_i^\mu \underbrace{\left(\sum_\nu C^{-1}_{\mu\nu} \xi_j^\nu \right)}_{\eta_j^\mu} \sigma_j$$

project σ_j onto η_j^μ

now $\eta_j^\mu \perp \xi^\nu$ for $\mu \neq \nu$, therefore this projection filters out the component of σ_j that is $\perp$ to $\underline{\xi}^\mu$

In the following adopt $J = \frac{1}{N} \xi \cdot \xi^T$

Look the set of N neurons as a system of N interacting units evolving according to the evolution rule introduced before, where

$L(\underline{\sigma}; J)$ plays as $E(\sigma; J)$

→ basic ingredients of the Hopfield model

Before proceeding it is worth spending a few words on the hystorical context where these models and algorithms were introduced

In fact, a wide variety of models had been proposed to account for pattern recognition, even before the seminal work by Hopfield in '82.

Nakano '71

Kohonen & Amari (1972) who introduced the idea of building associative memories with recurrent neural networks, where one creates specific stable microscopic network states by manipulation of the interaction strengths, that is, the synapses.

This approach was, in turn, inspired by the work of Caianiello '61

Approximately in the same years (1974) Little introduced concepts from stat. phys. of magnetic systems (such as temperature) into the study of recurrent networks.

Pastur & Figotin '78

NOISELESS CASE

Given a set of memories $\{\underline{\xi}^\mu\}_{\mu=1,\dots,K}$ where each entry $\xi_i^\mu \sim \text{Rad}(0)$
ask whether each of the K is a fixed point and, if so,
under which conditions in terms of the system parameters $(N, K, \underline{h}=\underline{0})$
we also want to estimate the width of the attraction basins.

stability of $\underline{\xi}^\mu$

$$\sigma_i^{(t+1)} = \text{sgn}\left(\varphi_i^{(t)}\right) \longrightarrow \xi_i^\mu = \text{sgn}\left(\varphi_i(\underline{\xi}^\mu)\right) \text{ fulfilled}$$

recast

$$\xi_i^\mu \cdot \varphi_i(\underline{\xi}^\mu) > 0$$

popular approach — based on scaling argument
- physics literature / engineering : S2N pseudo-heuristic
- math " : large-deviation theory

↳ its validity needs to be checked for $i=1,\dots,N$ and for the K patterns

$\mu=1$ target pattern wlg

$$\xi_i^1 \varphi_i(\underline{\xi}^1) = \frac{\xi_i^1}{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \sum_\mu \xi_i^\mu \xi_j^\mu \xi_j^1 =$$

$$= \underbrace{\frac{N-1}{N}}_{\substack{\mu=1 \\ \text{signal}}} + \underbrace{\frac{1}{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \sum_{\mu>1} \xi_i^1 \xi_j^1 \xi_i^\mu \xi_j^\mu}_{\mu>1 \text{ noise}}$$

$$\underset{N\gg1}{\approx} 1 \qquad\qquad \underset{N\gg1}{\approx} \sqrt{\frac{K}{N}}$$

slow-noise
that in addition to fast-noise acts directly on the neuronal dynamics (temperature)
It acts on the energy landscape making it more rugged and slowing down the dynamics as we will see.

we anticipate that $\frac{K}{N}$ play a crucial role in the interplay

- estimate attraction basin $\qquad \underline{\sigma} = \underline{\xi}^\mu \rightarrow \text{sgn}(\underline{\xi}^1 + \underline{\xi}^2 + \underline{\xi}^3)$
- check stability of different configurations eg. spurious $\qquad K$ finite

Review by Ricco '96

Perfect retrieval, i.e. $\underline{\sigma} = \underline{\xi}^{\mu}$ is a fixed point, $(\delta, 0)$-stable

McEliece '87 (LDT) $\quad K = \mathcal{O}\left(\frac{N}{\log N}\right)$ as $N \uparrow \infty$

$\delta < \frac{1}{2}$ normalized H-distance $\quad$ mismatched entries in the initial datum is strictly smaller than $N/2$

where the constant varies according to some details (eg existence of some exceptional memories, synch. vs asynch. dynamics, convergence in prob $\frac{1}{4}$ vs almost sure $1/6$ convergence)

In particular, when K is finite while $N \uparrow \infty$, retrieval is perfectly ensured for each pattern.

Relax perfect retrieval, i.e. $(0, \varepsilon)$ - stable

Newman '88 $\qquad K = \Theta(N) \quad$ as $N \uparrow \infty$

Loukianova '94 $\qquad$ (numerical estimate for $\varepsilon \gtrsim 0.015$)

Again we can be more explicit in the relation and to this purpose it is useful to introduce the following

<u>Def</u> load of the netw. is the ratio between the amount of stored info and of resources

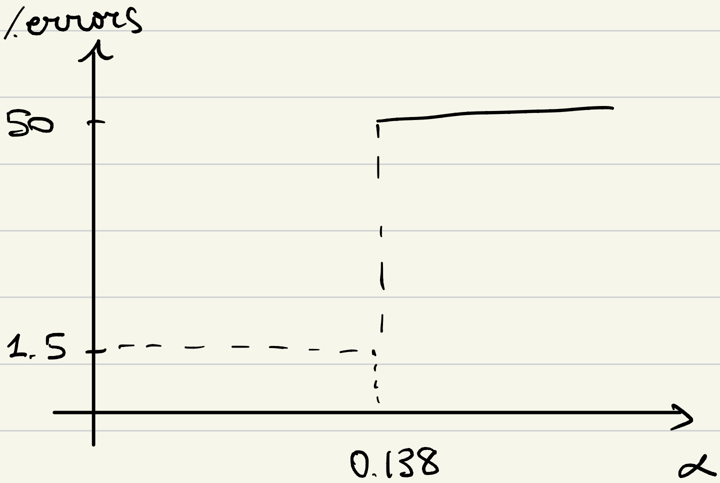
$$\alpha := \lim_{N \uparrow \infty} \frac{K}{N} \qquad \begin{cases} \text{Low load} \\ \text{High load} \end{cases}$$

$\alpha_c \approx 0.056$

$\alpha_c \approx 0.071 \qquad \alpha_c \approx 0.138 \quad$ AGS '85

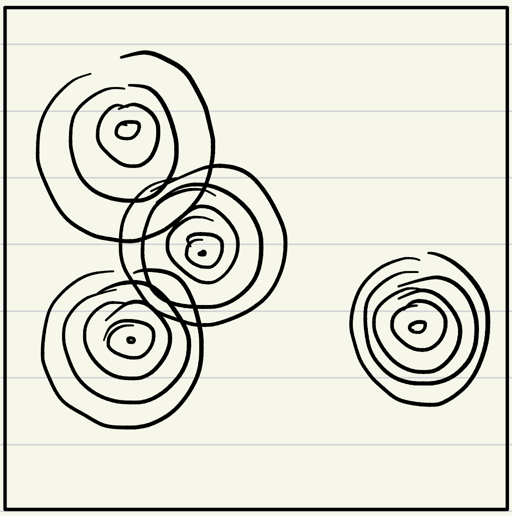
Black-out scenario

too large load



When α is relatively large, slow noise can prevail over the signal and we abruptly loose our capacity of retrieving.

We can look at Hebb's rule as an iterative rule s.t. at each step a new minimum in the energy landscape is generated.



As K grows attraction basins start to overlap, saddle point an local minima emerge and the energy landscape gets more and more rugged.

As K exceeds a certain threshold, config. $\underline{\sigma} = \underline{\xi}^{\mu}$ are no longer minima.

Now minima correspond to config. that are not useful for our purposes ($m = 0$)

To summarize:

When α gets too large, patterns $\underline{\xi}^\mu$ are densely packed in the embedding space $\{-1,+1\}^N$ and, as a result, they cease to be global minima and eventually to be minima at all.

→ revisions in the Hebbian have been introduced to mitigate this effect.

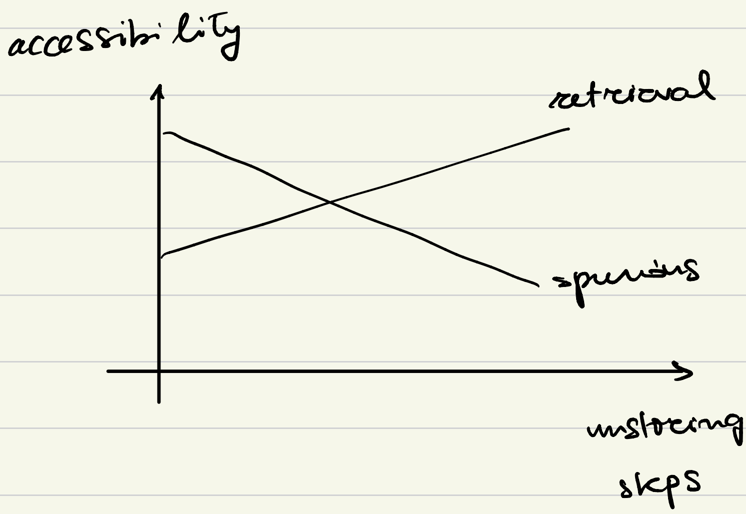
Examples inspired by the numerical work by Hopfield '83

Hebb: iterative storing rule

$$J_{ij} \to J_{ij} + \xi_i^\mu \xi_j^\mu \quad \mu\text{-th step}$$

Sample σ^* → reverse storing mechanism

$$J_{ij} \to J_{ij} - \sigma_i^* \sigma_j^*$$



Plakhov '94

Dotsenko et al '91

$$J_{ij}^{(n)} \to J_{ij}^{(n+1)} = J_{ij}^{(n)} - \varphi_i(\underline{\sigma}^*)^{(n)} \varphi_j(\underline{\sigma}^*)^{(n)} \frac{\varepsilon_n}{N}$$

unlearn. rate

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

$$C = \frac{\xi^T \xi}{N} \qquad C_{\mu\nu} = \frac{\sum_i \xi_i^\mu \xi_i^\nu}{N}$$

as $t \to \infty$

minima are shallow $\to$ saturate to $\alpha = 1$ but fragile

Fachechi et al '19 $\quad J_{ij}^{(n)} \to J_{ij}^{(n+1)} = J_{ij}^{(n)} + \dfrac{\varepsilon}{1+\varepsilon n} \left[J^{(n)} - (J^{(n)})^2 \right]$

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

Rationale : mimicking mechanisms occurring in mammal's brain during sleep.

$$\dot{J} = \frac{1}{1+t} \, (J - J^2)$$

$\downarrow$ consolidation

$\hookrightarrow$ remotion

$t \to 0$ Hebb's rule

$t \to \infty$ Kohonen's projection

(Kanter & Sompolinsky '86)

ε = unlearning strength (to be set small enough to secure convergence of algorithm)

NOISY DYNAMICS

$$\sigma_i^{(t+1)} = \text{sgn}\left[\varphi_i^{(t)} + T z_i^{(t)} \right] \quad i = 1, \dots, N \text{ finite}$$

$T \in \mathbb{R}^+$

$z_i^{(t)} \sim_{iid} \rho_z(z)$

$\beta := 1/T$

$$g_z(z) = \frac{1}{2}\left[1 + \tanh(z)\right]$$

$$\text{Prob}\left[\sigma_i^{(t+1)}\right] = g\left(\sigma_i^{(t+1)} \varphi_i^{(t)} \cdot \beta\right) = \frac{1}{2}\left[1 + \sigma_i^{(t+1)} \tanh\left(\beta \varphi_i(\underline{\sigma}^{(t)})\right)\right]$$

stochastic contribution $\to$ probabilistic approach

$$p^{(t+1)}(\underline{\sigma}) = \sum_{\underline{\sigma}' \in \Omega} W[\underline{\sigma} | \underline{\sigma}'] \, p^{(t)}(\underline{\sigma}')$$

$\Omega = \{-1, +1\}^N$

$|\Omega| = 2^N$

$$W[\underline{\sigma} | \underline{\sigma}'] \in [0, 1]^{2^N \times 2^N}$$

Stochastic matrix

$$\sum_{\underline{\sigma} \in \Omega} W[\underline{\sigma} | \underline{\sigma}'] = 1$$

Rows sum to 1

Transfer probability matrix $\underline{\sigma}' \to \underline{\sigma}$

As W is homogeneous, given $p^{(0)}(\underline{\sigma})$,

$$p^{(1)} = W p^{(0)} \to p^{(2)} = W^2 p^{(0)} \to \dots \to p^{(n)} = W^n p^{(0)}$$

Markov process

- irreducible (all states communicate)
- aperiodic (sequential dynamics with random selection)

$\Big\}$ ergodic

✓ This process is also homogeneous (W independent of time) — homogeneity not required for ergodicity.

✓ Ergodicity would also require positive recurrence

Ergodicity $\exists$ path of finite length n s.t.

$$W[\underline{\sigma} | \underline{\sigma}']^n > 0 \quad \forall\, \underline{\sigma}, \underline{\sigma}' \in \Omega$$

any state can be reached from any other state in a finite number of steps, whatever the initial state.

N state MC

ergodicity $\Rightarrow$ $\exists !$ stationary $\left[\text{i.e. } p^{(\infty)}(\underline{\sigma}) = \sum_{\underline{\sigma}'} W[\underline{\sigma} | \underline{\sigma}'] p^{(\infty)}\right]$
distribution $p^{(\infty)}$
s.t. for all $\underline{\sigma}, \underline{\sigma}' \in \Omega$ $\quad \lim_{t \to \infty} p^{(t)}(\underline{\sigma} | \underline{\sigma}') = p^{\infty}(\underline{\sigma})$

in particular, $p^{(\infty)}$ is independent of starting point.

$W p^{(\infty)} = p^{(\infty)}$ (right) eigenstate left invariant

2^N linear equations, coupled

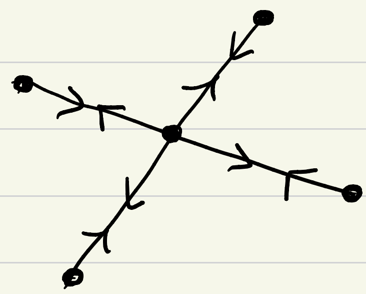
$p^{(\infty)}(\underline{\sigma})$, $\underline{\sigma} \in \Omega$

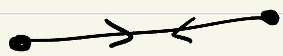
short-cut : reversible / equilibrium

$$W[\underline{\sigma}|\underline{\sigma}']\, p^{(\infty)}(\underline{\sigma}') = W[\underline{\sigma}'|\underline{\sigma}]\, p^{(\infty)}(\underline{\sigma}) \quad \forall \underline{\sigma}, \underline{\sigma}' \in \Omega$$

detailed balance

$$p^{(t+1)}(\underline{\sigma}) - p^{(t)}(\underline{\sigma}) = \underbrace{\sum_{\underline{\sigma}' \neq \underline{\sigma}} \overbrace{\left\{ \underbrace{W[\underline{\sigma}|\underline{\sigma}']\, p^{(t)}(\underline{\sigma}')}_{\text{in}} - \underbrace{W[\underline{\sigma}'|\underline{\sigma}]\, p^{(t)}(\underline{\sigma})}_{\text{out}} \right\}}^{\text{db}}}_{\text{stationariety}}$$





$$W[\underline{\sigma}|\underline{\sigma}'] = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{1}{2} \left[1 + \underline{\sigma}' \tanh\left(\beta \varphi_i(\underline{\sigma}')\right)\right] \delta_{\underline{\sigma}, \underline{\sigma}'} + \frac{1}{2} \left[1 - \underline{\sigma}' \tanh\left(\beta \varphi_i(\underline{\sigma}')\right)\right] \delta_{F_i \underline{\sigma}, \underline{\sigma}'} \right\}$$

$F_i \underline{\sigma} := (\sigma_1, \dots, -\sigma_i, \dots, \sigma_N)$

$\frac{1}{N} \to$ uniform distribution for selecting the neuron to be updated

$\underline{\sigma}' \to \frac{1}{N}$

$\to \text{Prob}[\sigma_i] = \frac{1}{2}\left[1 + \sigma_i \tanh(\beta \varphi_i)\right]$

<u>Prop</u> For sequential dynamics, no self-interactions $J_{ii} = 0 \ \forall i$

$$\boxed{\underline{J} = \underline{J}^T \iff \text{db}}$$

If db, then $p^{(\infty)}(\underline{\sigma}) \propto \exp\left[-\mathcal{H}_N(\sigma; J)\beta\right]$

with $\mathcal{H}_N(\sigma; J) = -\frac{1}{2}\sum_{i,j} J_{ij}\sigma_i\sigma_j - h\sum_i \sigma_i$

$\equiv L_N(\sigma; J)$

Recovered equilibrium stat. mechanics framework

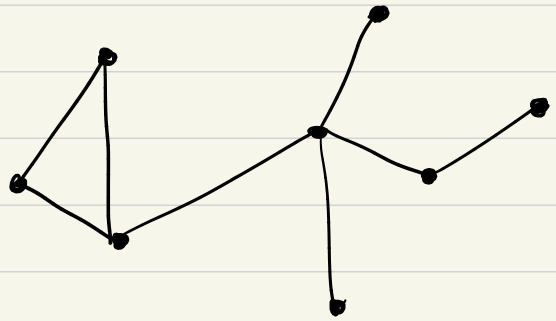
$\to$ leverage tools available

Remarks

- Consistency with Lyapunov : small energy config. are more likely
- Noiseless case $\rightarrow$ fixed point / ergodicity breakdown
- Noisy case $\rightarrow$ ergodicity

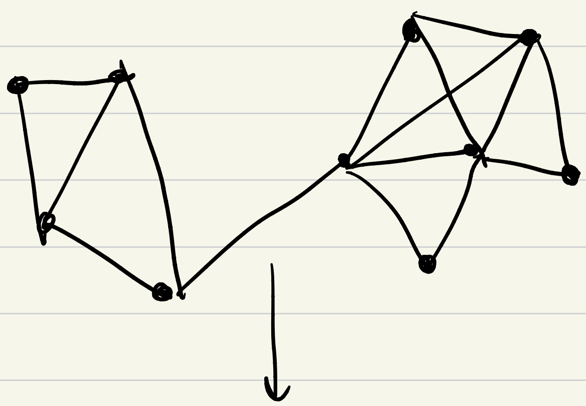
How can I ensure that a target configuration is reached and retained even in a noisy system?

Asymptotically $N \rightarrow \infty$



By adding nodes the topology can be such that there are modules densely connected which on the other hand are loosely connected each other.

As a result, as $N \rightarrow \infty$ the prob. of moving from one module to the other gets vanishing.



bottle-necks

3

W transition matrix

Spectral expansion $\rightarrow$ asymptotic degeneration of Perron-Frobenius eigenvalue

$\rightarrow$ loose independence on starting configuration

Otherwise stated, as $N \to \infty$ positive recurrence no longer holds

I want that the target configuration can be reached and retained. By ergodicity breaking I can ensure that a target config. is, in fact, (δ, ε)-stable that is, it remains confined in an accessible set of configurations within a radius ε. This ball is surrounded by energy barriers that cannot be overcome. Then, if I want to retrieve a different pattern, I have to reinitialize the system.

Remark

$$P_{N,\beta}(\underline{\sigma}; J) = \exp\left[-\beta H_N(\underline{\sigma}; J)\right]$$

Can also be recovered following an information-driven approach MEA (Jaynes '57)

THE HOPFIELD MODEL

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

$$\mathcal{H}_{N,h}(\underline{\sigma}; \underline{\xi}) = -\frac{1}{N} \sum_{i<j} \xi_i^\mu \sigma_i \sigma_j \xi_j^\mu - \sum_\mu \sum_i h_\mu \xi_i^\mu \sigma_i$$

$\xi_i^\mu \sim \mathrm{Rad}(0)$, $\xi_i^\mu \in \{-1, +1\}$ $\to$ synthetic, structureless dataset
build up a theory where statistical properties are controllable

$$P_{N,\beta,h}(\underline{\sigma}; \underline{\xi}) = \frac{1}{Z_{N,\beta,h}(\underline{\xi})} \exp\left(-\beta \mathcal{H}_{N,\beta,h}(\underline{\sigma}; \underline{\xi})\right)$$

To lighten the notation: drop the subscript, but retain N which is dropped only when referring to the TDL.

Def Expectations

$$\omega_N(f) = \sum_{\underline{\sigma}} P_N(\underline{\sigma}; \underline{\xi}) \cdot f(\underline{\sigma}) \qquad \text{thermal average}$$

$$\mathbb{E}[\cdot] = \sum_{\underline{\xi}} (\cdot) \qquad \text{average of disorder}$$

$$\langle \cdot \rangle_N = \mathbb{E}\, \omega_N(\cdot)$$

Def Mattis magnetization, or pattern overlap

$$m_\mu = \frac{1}{N} \sum_{i=1}^N \sigma_i \xi_i^\mu \in [-1, +1] \qquad \mu = 1, \dots, k$$

$\underline{m} = (m_1, m_2, \ldots, m_K)^T$ order parameter

An order parameter is a measurable physical quantity that allows us to differentiate between different phases of a system.

Here $\underline{m}^2 = 1$ pure retrieval state

$\underline{m} = 0$ disordered state

$$\mathcal{H}_N(\underline{\sigma};\underline{\xi}) = -\frac{1}{2N}\sum_{i,j}\sum_{\mu}\xi_i^\mu\sigma_i\xi_j^\mu\sigma_j + \underbrace{\frac{1}{2N}\sum_{i}\sum_{\mu}\xi_i^\mu\xi_i^\mu\sigma_i\sigma_i}_{\text{diagonal correction}}$$

$$-\sum_{\mu,i} h_\mu \xi_i^\mu \sigma_i$$

$$= -\frac{1}{2N}\sum_{\mu}(Nm_\mu)(Nm_\mu) + \frac{1}{2N}\sum_{i,\mu} 1 - \sum_{\mu} h_\mu N m_\mu$$

$$= -\frac{N}{2}\underline{m}^2 + \frac{1}{2N}\cdot K\cdot N - N\,\underline{h}^T\underline{m}$$

$$= -\frac{N}{2}\underline{m}^2 + \frac{K}{2} - N\,\underline{h}^T\cdot\underline{m}$$

- Hamiltonian scales linearly with the size, as it should
- diagonal corrections negligible in the thermodynamic limit and as long as $\frac{K}{N} \to 0$, while if $\frac{K}{N}$ remains finite they provide a constant contribution

Def Load $\alpha := \lim_{N\to\infty} \frac{k}{N}$ $\begin{cases} \alpha = 0 & \text{low-load} \\ \alpha > 0 & \text{high-load} \end{cases}$

goal $\langle \underline{m} \rangle$ vs β, α, h

$\to$ phase diagram $\to$ a priori setting

Distinguish phases when the system behaves qualitatively different
Specifically, I am interested in a phase where the machine can be exploit for retrieval purposes.

goal Explicit expression (in the control parameters) of the free-energy density

$$f_N(\underline{\xi}) := -\frac{1}{N\beta} \log Z_N(\underline{\xi})$$

$$f_N := -\frac{1}{N\beta} \mathbb{E} \log Z_N(\underline{\xi})$$

interested in the typical behaviour, not the one for a specific realization

$$f := \lim_{N\to\infty} f_N$$

why f is a key quantity?

- probabilistic point of view: $\log Z_N$ plays as a cumulant generating function, derivatives of which yield mean, variance, etc of the observable conjugated eg. $\frac{d}{dh_\mu} f_N = -\langle m_\mu \rangle_N$ HW $\chi = \frac{1}{T}(\langle M^2\rangle - \langle M\rangle^2)$

$\frac{d^2}{dh_\mu^2} f_N = -N\beta \, \mathrm{Var}_N(m) \propto \chi = \frac{dm}{dh}$

When the system does not encompass any external field we can always introduce an auxiliary one to be set to zero

$$\langle m_\mu \rangle = \lim_{h' \to 0} \left. \frac{df_N}{dh_\mu} \right|_{h_\mu = h'}$$

- physical viewpoint

the free-energy is a thermodynamical potential

this can be determined for a given configuration of the system $\underline{\sigma}$ or, by applying a coarse-graining, for a given value of the M. magnetization $\underline{m}$

As the system spontaneously evolves the free energy decreases as a result of the first and the second principles of TD. Therefore, the expectation of m can be obtained by applying an extremization procedure over f.

$$\langle \underline{m} \rangle = \underset{\underline{m}}{\mathrm{argmin}} \, f(\underline{m})$$

Techniques (not exhaustive list)

- Saddle point / Laplace method
- Large deviation principle — Guerra, Bovier, Picco, Talagrand — 1990s
- Random matrix theory — Shcherbina, Pastur — 1990s
- Martingale — Tirozzi

INTERPOLATION TECHNIQUE

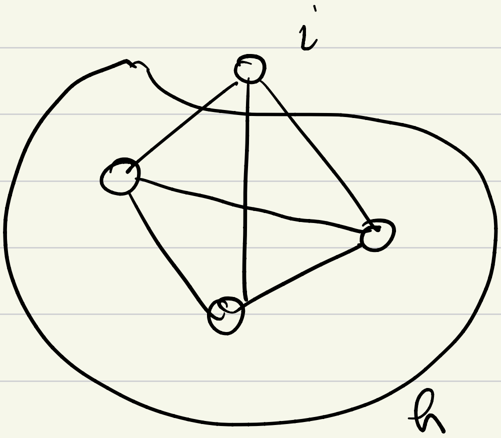
Guerra '01 Fields Inst. Comm

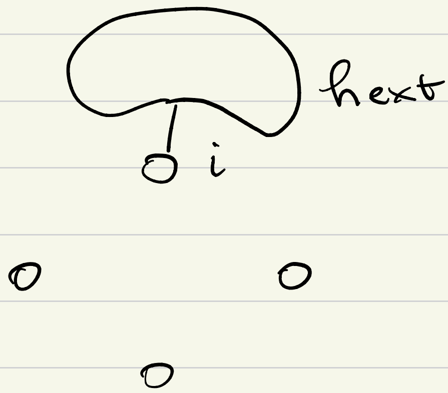
$$\mathcal{H}_N(\underline{\sigma}; J) \longrightarrow \tilde{\mathcal{H}}_N(\underline{\sigma}; J, J'; t) =$$

$$= t\,\mathcal{H}_N(\underline{\sigma}; J) + (1-t)\,\mathcal{H}_N^{easy}(\underline{\sigma}; J')$$

$t \in [0,1]$ s.t. $t = 1 \to$ original model

$t = 0 \to$ "easy" model





$$\tilde{\mathcal{H}} \to \tilde{Z}, \tilde{F}, \dots$$

$$F_N = \tilde{F}(t=1) = \tilde{F}(t=0) + \int_0^1 \frac{d\tilde{F}}{dt}\Big|_{t=t'} dt'$$

$$= \underbrace{\tilde{F}(t=0) + \phi_1(w(m), \dots)}_{F^{trial}} + \underbrace{\int_0^1 \phi_2\left[(w(m)-m)^2, \dots\right] dt'}_{remainder}$$

$$\xrightarrow[RS]{N\to\infty} 0$$

$$\tilde{H}_N(\underline{\sigma};\underline{\xi},\underline{M}) = t\underbrace{\left(-\frac{N}{2}\sum_\mu m_\mu^2\right)}_{\mathcal{H}_N(\underline{\sigma};\xi)} + (1-t)\underbrace{\left(-N\sum_\mu m_\mu M_\mu\right)}_{\sum_\mu\sum_i \sigma_i \xi_i^\mu M_\mu}$$

superposition of K fields, each pointing in the direction of the related pattern and tuned by M

$\underline{M} = (M_1, \dots, M_K)$ to be set a posteriori

M_μ mimics m_μ

$\sum_\mu \xi_i^\mu M_\mu$ mimics the field acting on σ_i

$$\varphi_i(\underline{\sigma}) = \sum_j J_{ij}\sigma_j = \frac{1}{N}\sum_{j\mu}\xi_i^\mu\xi_j^\mu\sigma_j = \sum_\mu \xi_i^\mu m_\mu$$

$$\tilde{A}(t=0) = \frac{1}{N}\log\tilde{Z}(t=0)$$

$$= \frac{1}{N}\log\sum_{\underline{\sigma}} e^{+\beta \cancel{N}\sum_\mu\sum_i \frac{\xi_i^\mu\sigma_i M_\mu}{\cancel{N}}}$$

$$= \frac{1}{N}\log\sum_{\underline{\sigma}}\prod_{i=1}^N e^{\beta\sum_\mu \xi_i^\mu M_\mu\sigma_i}$$

$$= \frac{1}{N}\log\prod_{i=1}^N\sum_{\sigma_i=\pm1} e^{\beta(\underline{\xi}_i\cdot\underline{M})\sigma_i}$$

$$= \frac{1}{N}\log\prod_{i=1}^N 2\cosh\left(\beta(\underline{\xi}_i\cdot\underline{M})\right)$$

$$= \frac{1}{N} \left[\log 2^N + \sum_i \log \left[\beta (\underline{\xi}_i \cdot \underline{M}) \right] \right] =$$

$$= \log 2 + \underbrace{\frac{1}{N} \sum_i \log \left[\beta (\underline{\xi}_i \cdot \underline{M}) \right]}_{\overline{\log [\beta (\underline{\xi} \cdot \underline{M})]}} \qquad \underline{\xi}_i \in \{-1, +1\}^K$$

$$\xrightarrow[\substack{N \to \infty \\ \frac{K}{N} \to 0}]{} \log 2 + \mathbb{E} \log \left[\beta (\underline{\xi}^T \cdot \underline{M}) \right]$$ law of large numbers

$$\tilde{A} = \frac{1}{N} \log \tilde{Z}$$

$$\frac{d\tilde{A}}{dt} = \frac{1}{N} \frac{1}{\tilde{Z}} \frac{d\tilde{Z}}{dt} = \frac{1}{N} \frac{1}{\tilde{Z}} \frac{d}{dt} \sum_{\underline{\sigma}} e^{-\beta \tilde{\mathcal{H}}(\underline{\sigma}; \underline{\xi}, \underline{M})} =$$

$$= \frac{1}{\cancel{N}} \frac{1}{\tilde{Z}} \sum_{\underline{\sigma}} \left(\beta \frac{\cancel{N}}{2} \underline{m}^2 - \beta \cancel{N} \underline{m} \cdot \underline{M} \right) e^{-\beta \tilde{\mathcal{H}}(\underline{\sigma}; \underline{\xi}, \underline{M})}$$

$$= \beta \sum_{\sigma} \left[\left(\frac{\underline{m}^2}{2} - \underline{m} \cdot \underline{M} \right) \frac{e^{-\beta \tilde{\mathcal{H}}}}{\tilde{Z}} \right]$$

$$= \beta \left[\tilde{\omega} \left(\frac{\underline{m}^2}{2} \right) - \tilde{\omega}(\underline{m}) \cdot \underline{M} \right]$$

$$= \frac{\beta}{2}\left[\tilde{w}(\underline{m}^2) - 2\underline{M}\cdot\tilde{w}(\underline{m}) + \underline{M}^2 - \underline{M}^2 \right]$$

$$= \frac{\beta}{2}\left[\tilde{w}(\underline{m}-\underline{M})^2 - \underline{M}^2 \right]$$

$$\downarrow$$

$$\tilde{w}(\underline{m}^2 + \underline{M}^2 - 2\underline{m}\underline{M})$$

$$= \tilde{w}(\underline{m}^2) + \underline{M}^2 - 2\tilde{w}(\underline{m})\underline{M}$$

$$\rightarrow \frac{d\tilde{A}}{dt} = \frac{\beta}{2}\tilde{w}(\underline{m}-\underline{M})^2 - \frac{\beta}{2}\underline{M}^2$$

$$A = \tilde{A}(t=1) = \tilde{A}(t=0) + \int_0^1 \frac{d\tilde{A}}{dt}\Big|_{t=t'} dt'$$

$$= \log 2 + \mathbb{E}\log\cosh\left[\beta\,\underline{\xi}\cdot\underline{M}\right] - \frac{\beta}{2}\underline{M}^2 + \int_0^1 \frac{\beta}{2}\tilde{w}(\underline{m}-\underline{M})^2 dt$$

Dealing with integral $\rightarrow$ see Agliari, Fachechi, Marullo '20

Low load : rigorous treatment is feasible

Exploit self-average of $\underline{m}$ $\rightarrow$ see [B]

written simply almost surely $\lim_{N\to\infty} P_N(\underline{m}) = \delta(\underline{m}-\underline{\bar{m}})$, with $\underline{\bar{m}} = \omega(\underline{m})$

Extremize w.r.t. M_μ

$$\frac{dA}{dM_\mu} = 0 \rightarrow \mathbb{E}\, \xi^\mu \tanh\left[\beta(\underline{\xi}\cdot\underline{M})\right] = M_\mu$$

$M_\mu = \omega(m_\mu)$ must fulfil this equation

Consistency check

$$\mathcal{H} + \sum_\mu h_\mu m_\mu \rightarrow A = \log 2 + \mathbb{E}\log\cosh\left[\beta(\underline{\xi}\cdot\underline{M}) + \beta(\underline{\xi}\cdot\underline{h})\right] - \frac{\beta M^2}{2}$$

$$\lim_{\underline{h}'\to 0} \left.\frac{dA}{dh_\mu}\right|_{\underline{h}=\underline{h}'} = \mathbb{E}\tanh\left[\beta\,\underline{\xi}\cdot\underline{M}\right]\cdot\xi_\mu = \omega(m_\mu) = M_\mu$$

Another heuristic argument [CKS]

$$\text{Prob}\left[\sigma_i \rightarrow -\sigma_i\right] = \frac{1}{2}\left[1 - \sigma_i \tanh\left(\beta \varphi_i(\underline{\sigma})\right)\right]$$

Ask for stationariety

$$1 = \sigma_i \tanh\left(\beta\varphi_i(\underline{\sigma})\right)$$

$$\sigma_i = \tanh\left[\beta\sum_j J_{ij}\sigma_j\right]$$

$$\xi_i^1\sigma_i = \xi_i^1 \tanh\left[\frac{\beta}{N}\sum_j\sum_\mu \xi_i^\mu\xi_j^\mu\sigma_j\right]$$

$$\underbrace{\frac{1}{N}\sum_i \xi_i^1\sigma_i}_{m_1} = \frac{1}{N}\sum_i \xi_i^1 \tanh\left[\beta\sum_\mu \xi_i^\mu m_\mu\right] = \overline{\xi^1 \tanh\left[\beta\underline{\xi}\,\underline{m}\right]}$$

$$\rightarrow \mathbb{E}\left[\xi^1\tanh\left(\beta\underline{\xi}\,\underline{m}\right)\right]$$

$$\underline{m} = \mathbb{E}\left[\underline{\xi}\tanh\left(\beta\underline{\xi}\cdot\underline{m}\right)\right]$$

$\underline{m} = \mathbb{E}_{\xi} \tanh[\beta\, \underline{\xi} \cdot \underline{m}]$ self-consistency eq.

Let's now inspect their solutions. For this purpose, we recall that, from S2N, we know that, at $T=0$, a configuration $\underline{\sigma}^{(n)}$ that is a symmetric, odd mixture of the patterns $\underline{\xi}^1, \dots, \underline{\xi}^n$ (wlg we can consider the first n patterns), namely

$$\sigma^{(n)} = \mathrm{sgn}\,(\underline{\xi}^1 + \dots + \underline{\xi}^n)$$

is stable. The related magnetization is

$$m_\mu(\underline{\sigma}^{(n)}) = \frac{1}{N} \sum_i \sigma_i^{(n)} \xi_i^\mu \quad \text{for } \mu = 1, \dots, K$$

In the limit of large N they are in the form

$$\underline{m}(\underline{\sigma}^{(n)}) \underset{N\to\infty}{=} m^{(n)} (\underbrace{1, \dots, 1}_{n}, \underbrace{0, \dots, 0}_{K-n})$$ where $m^{(n)}$ measures the extent of the overlap

For instance:

$n=1$ $\quad m^{(1)}(1, 0, \dots, 0) \qquad m^{(1)} = 1$

$n=3$ $\quad m^{(3)}(1, 1, 1, 0, \dots, 0) \qquad m^{(3)} = 1/2$

$$m_\mu = \frac{1}{N} \sum_i \sigma_i^{(3)} \xi_i^\mu \xrightarrow[N\to\infty]{\mathbb{E}} \begin{cases} 0 & \mu > 3 \\ \frac{1}{2} & \mu \le 3 \end{cases}$$

law of large numbers

$$\mathrm{sign}(\xi_i^1 + \xi_i^2 + \xi_i^3)\, \xi_i^1 = \left[+1 \cdot (1 - 1/4) - 1 \cdot 1/4\right] \cdot 2$$

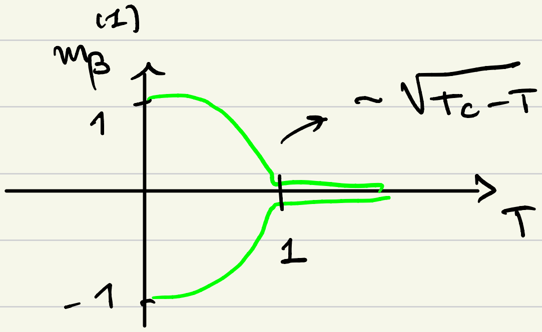
At $T>0$ $\rightarrow$ $\underline{m} = m_\beta^{(n)} (1, \dots, 1, 0, \dots, 0)$

$n=1$

let's plug this expression into SC

• $\mu = 1$ $\quad m_\beta^{(1)} = \mathbb{E}\, \cancel{\xi^1} \tanh\left(\beta \cancel{\xi^1} \cdot m_\beta^{(1)}\right) = \tanh\left(\beta m_\beta^{(1)}\right)$

SC of CW model $\quad m = \tanh(\beta m)$



• $\mu > 1$ $\quad 0 = \mathbb{E}\, \xi^{\mu>1} \tanh\left(\beta \xi^1 \cdot m_\beta^{(1)}\right) = \mathbb{E}[\ ]\,\mathbb{E}[\ ] = 0$

$\exists\, T_c = 1$ s.t. $\begin{cases} T > T_c & \underline{m} = \underline{0} \quad \text{unique solution} \\ T < T_c & \underline{m} \neq \underline{0} \quad \text{many admissible solution} \end{cases}$

this is true in general

and can be proved by showing that $\underline{m}^2 \leq \beta\, \underline{m}^2$ HW

$$\underline{m}^2 = \sum_\mu m_\mu \cdot m_\mu = \sum_\mu m_\mu \, \mathbb{E}\, \xi^\mu \tanh\left[\beta(\underline{\xi}\cdot\underline{m})\right]$$

$$= \mathbb{E}\left(\underline{\xi}\cdot\underline{m}\right)\tanh\left[\beta(\underline{\xi}\cdot\underline{m})\right] = \mathbb{E}\,|\underline{\xi}\cdot\underline{m}|\tanh\left[\beta|\underline{\xi}\cdot\underline{m}|\right]$$

$$\underset{\tanh(x)\le x}{\le} \beta\,\mathbb{E}\left(|\underline{\xi}\cdot\underline{m}|^2\right) = \beta\,\mathbb{E}\sum_{\mu\nu}\xi^\mu\xi^\nu m_\mu m_\nu$$

$$= \beta\sum_{\mu\nu} m_\mu m_\nu\,\mathbb{E}\left[\xi^\mu\xi^\nu\right] \xrightarrow[N\to\infty]{} \beta\sum_{\mu\nu} m_\mu m_\nu \delta_{\mu\nu}$$

$$= \beta\underline{m}^2$$

$$m(T=1) = 0$$

$\Rightarrow \underline{m}^2 \le \beta\underline{m}^2$ As $\beta < 1$ the only solution is $\underline{m} = \underline{0}$

We can expand around $\underline{m} \simeq \underline{0}$ for $\beta - 1 = \varepsilon \ll 1$

$\rightarrow m_\mu \propto \sqrt{\underline{m}^2 - \varepsilon}$ (consistent with the critical behaviour of the CW model as anticipated)

$$m_\beta^{(n)} \sim \frac{\sqrt{\varepsilon}}{\sqrt{n}}$$ As $n\uparrow$ the overlap progressively decreases

Now, the solutions of the SC eqs. are only candidate eq. states, since we also need to ask for stability

$$f(\underline{m}) = \frac{1}{2}\underline{m}^2 - T\,\mathbb{E}\log\cosh(\beta\,\underline{\xi}\cdot\underline{m})$$

$$\frac{df}{d\underline{m}} = 0 \longrightarrow \underline{m}^* \quad \text{(gradient)}$$

$$H_{\mu\nu}(\underline{m}^*) := \left.\frac{\partial^2 f(\underline{m})}{\partial m_\mu \partial m_\nu}\right|_{m=m^*}$$

m^* stable iff $H(\underline{m}^*)$ positive definite (i.e. positive eigenvalues)

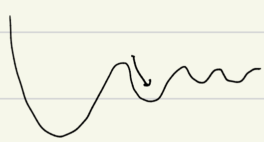
up to three distinct eigenvalues

$n=1$ always stable for $T<1$

$n=2$ stable for $T>1$, but there only $\underline{m}=\underline{0}$ is solution (saddle point)

n odd stable for $T<T_c^{(n)}$

$$T_c^{(3)} \simeq 0.461$$
$$T_c^{(5)} \simeq 0.385 \quad \Big\downarrow n$$
$$T_c^{(7)} \simeq 0.345$$



Beneficial role of the temperature

$$f(m^{(k)}) < f(m^{(k+2)}) \quad k \text{ even}$$

HIGH-LOAD HOPFIELD MODEL

$\alpha \in \mathbb{R}^+$

$\alpha = 0 \rightarrow \alpha > 0$ richer phenomenology, eg. from S2N: even at $T=0$ the load can be too large to allow retrieval

$\alpha = 0$ Simple

$\alpha > 0$ Complex

change of paradigms

Oversimplifying: simple system $\rightarrow$ energy landscape is relatively smooth (e.g. CW)

complex system $\rightarrow$ it hosts a # minima that grows "fast" with N (e.g. SK)

Intuition: $J_{ij}^{(SK)} \sim \mathcal{N}(0, N^{-1})$

$$\left(\frac{1}{N}\xi \cdot \xi^T\right)_{ij} \xrightarrow{d} J_{ij}^{(SK)}$$

INTERMEZZO: SK MODEL

Instructive to first face this model which constitutes a paradigmatic case of complex system ($\to$ # minima grows with system size)

$$\mathcal{H}_N^{SK}(\underline{\sigma}; J) := -\frac{1}{2\sqrt{N}} \sum_{i,j} J_{ij}\sigma_i\sigma_j \qquad J_{ij} \sim \mathcal{N}(0,1)$$

Diagonal terms

$$\frac{1}{2\sqrt{N}} \underbrace{\sum_i J_i}_{\mathcal{N}(0,N)}$$

- sqrt normalization

preliminary: Wick - Isserlis theorem

let $x \sim \mathcal{N}(0, \nu^2)$ and let $\mathcal{C}^1 \ni g : \mathbb{R} \to \mathbb{R}$

and let $\lim_{x\to\pm\infty} g(x) P_\nu(x) = 0$

then $\mathbb{E}[g(x)\cdot x] = \mathbb{E}[x^2]\mathbb{E}[g'(x)]$

$$\langle \mathcal{H}_N^{SK}(\underline{\sigma}; J)\rangle = \mathbb{E} \frac{\sum_{\underline{\sigma}} \mathcal{H}_N^{SK} e^{-\beta\mathcal{H}_N^{SK}}}{\sum_{\underline{\sigma}} e^{-\beta\mathcal{H}_N^{SK}}} =$$

$$= -\frac{1}{2\sqrt{N}} \sum_{i,j} \underbrace{\mathbb{E}\, J_{ij} \frac{\sum_{\underline{\sigma}} \sigma_i\sigma_j e^{-\beta\mathcal{H}_N^{SK}}}{\sum_{\underline{\sigma}} e^{-\beta\mathcal{H}_N^{SK}}}}_{\text{Wick-Isserlis}} =$$

$$= -\frac{1}{2\sqrt{N}} \sum_{i,j} \underbrace{\mathbb{E}[J_{ij}^2]}_{=1\ \smile} \mathbb{E}\, \partial_{J_{ij}} \frac{\sum_{\underline{\sigma}} \sigma_i\sigma_j e^{-\beta\mathcal{H}_N^{SK}}}{\sum_{\underline{\sigma}} e^{-\beta\mathcal{H}_N^{SK}}} =$$

$\frown$

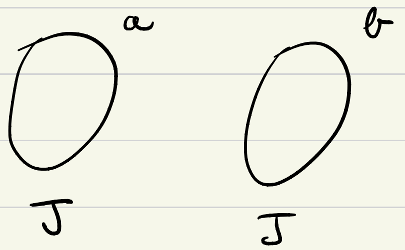
$$= -\frac{1}{2\sqrt{N}}\sum_{i,j}\left[\frac{\sum_{\underline{\sigma}} \sigma_i\sigma_j \frac{\beta}{2\sqrt{N}}\sigma_i\sigma_j\, e^{-\beta H_N^{SK}}}{Z} - \frac{\sum_{\underline{\sigma}} \sigma_i\sigma_j\, e^{-\beta H_N^{SK}} \sum_{\underline{\sigma}} \frac{\beta}{2\sqrt{N}}\sigma_i\sigma_j\, e^{-\beta H_N^{SK}}}{Z^2}\right]$$

$$= -\frac{\beta}{4N}\sum_{i,j}\left[\underbrace{\sum_{\underline{\sigma}}\frac{1\cdot e^{-\beta H_N^{SK}}}{Z}}_{1} - \frac{\sum_{\underline{\sigma}^{(a)}}\sigma_i^{(a)}\sigma_j^{(a)}\, e^{-\beta H_N^{SK,(a)}}\cdot\sum_{\underline{\sigma}^{(b)}}\sigma_i^{(b)}\sigma_j^{(b)}\, e^{-\beta H_N^{SK,(b)}}}{Z^2}\right]$$

$$= -\frac{\beta}{4N}\left[N^2 - \mathbb{E}\sum_{\underline{\sigma}^{(a)}}\sum_{\underline{\sigma}^{(b)}}\frac{\overbrace{\sum_i \sigma_i^{(a)}\sigma_i^{(b)}}^{N q_{ab}}\;\overbrace{\sum_j \sigma_j^{(a)}\sigma_j^{(b)}}^{N q_{ab}}\; e^{-\beta\left[H_N^{SK,(a)}+H_N^{SK,(b)}\right]}}{Z^2}\right]$$

Def: overlap

$$q_{ab} := \frac{1}{N}\sum_i \sigma_i^{(a)}\sigma_i^{(b)} \in [-1,+1]$$



physical meaning: proximity measure between two configurations, $\underline{\sigma}^{(a)}$, $\underline{\sigma}^{(b)}$, sampled from the same prob. distrib.

$$= -\frac{\beta N}{4}\left[1 - \langle q_{ab}^2\rangle\right] \propto N$$

- Quenched vs Annealed averages

Two different averages over J

$$f^{a}_{N,\beta,h} = -\frac{1}{\beta N}\,\mathbb{E}\log Z_{N,\beta,h}(J)$$

$$f^{A}_{N,\beta,h} = -\frac{1}{\beta N}\log \mathbb{E} Z_{N,\beta,h}(J)$$

$\mathbb{E}\sum_{\underline{\sigma}}$ $\rightarrow$ $\underline{\sigma}$, J live on the same timescale

$\sum_{\underline{\sigma}}$ then $\mathbb{E}$ $\rightarrow$ $\underline{\sigma}$ faster degree of freedom

First, relaxation of the σ's occurs while the J's are frozen (quenched)

Q: synapses are slow degrees of freedom wrt neurons
~ days ~ ms

$$f^{A}_{N,\beta,h} \xrightarrow[N\to\infty]{} -\frac{\beta}{4} - \frac{1}{\beta}\log 2 - \frac{1}{\beta}\log\cosh(\beta h) \quad \text{HW}$$

Quenched free energy for SK by interpolation

Hereafter drop the superscript Q

$$\mathcal{H}_N^{SK}(\sigma; J) = -\frac{1}{\sqrt{N}} \sum_{i<j} J_{ij}\sigma_i\sigma_j \qquad J_{ij} \sim \mathcal{N}(0,1)$$

$$\tilde{\mathcal{H}}(\underline{\sigma}; J, \eta) = -\frac{\sqrt{t}}{\sqrt{N}} \sum_{i<j} J_{ij}\sigma_i\sigma_j - \sqrt{1-t}\, C \sum_i \eta_i \sigma_i$$

- sqrt convenient (Wick - Isserlis)
- $\eta_i \sim \mathcal{N}(0,1)$ $\quad i = 1, \dots, N$
- C must be related with q

$$\tilde{A}_N = \frac{1}{N}\mathbb{E}\log \tilde{Z}_N \qquad \mathbb{E} \to \text{wrt } J \text{ and } \eta$$

$$A_N = \frac{\mathbb{E}}{N}\log Z_N = \tilde{A}_N(1) = \tilde{A}_N(0) + \int_0^1 \left.\frac{\partial \tilde{A}}{\partial t}\right|_{t=s} ds$$

$$\tilde{A}_N(0) = \frac{1}{N}\mathbb{E}\log\left[\sum_{\underline{\sigma}} \exp\left(\beta C \sum_i \eta_i\sigma_i\right)\right]$$

$$= \frac{1}{N}\mathbb{E}\log\left[\prod_{i=1}^{N} \sum_{\sigma_i=\pm 1} \exp\left(\beta C \eta_i \sigma_i\right)\right]$$

$$= \frac{1}{N} \mathbb{E} \sum_i \log \left[2\cosh(\beta c \eta_i) \right] =$$

$$= \log 2 + \mathbb{E} \log \cosh(\beta c \eta)$$

$$\frac{d\tilde{A}_N}{dt} = \frac{1}{N} \frac{\partial}{\partial t} \mathbb{E} \log \left[\sum_{\underline{\sigma}} \exp \left\{ \beta \frac{\sqrt{t}}{\sqrt{N}} \sum_{i<j} J_{ij} \sigma_i \sigma_j + \beta \sqrt{1-t}\, c \sum_i \eta_i \sigma_i \right\} \right]$$

$$= \frac{\beta}{N} \mathbb{E} \frac{1}{\tilde{Z}} \left[\sum_{\underline{\sigma}} \left\{ \frac{1}{2\sqrt{t}} \frac{1}{\sqrt{N}} \sum_{i<j} J_{ij} \sigma_i \sigma_j - \frac{1}{2\sqrt{1-t}} c \sum_i \eta_i \sigma_i \right\} \right]$$

$$= \frac{\beta}{N} \left[\frac{1}{2\sqrt{t}} \frac{1}{\sqrt{N}} \underbrace{\mathbb{E} \tilde{\omega} \left(\sum_{i<j} J_{ij} \sigma_i \sigma_j \right)}_{1} - \frac{1}{2\sqrt{1-t}} c \underbrace{\mathbb{E} \tilde{\omega} \left(\sum_i \eta_i \sigma_i \right)}_{2} \right]$$

$$1 \quad \mathbb{E} \tilde{\omega} \left(\sum_{i<j} J_{ij} \sigma_i \sigma_j \right) =$$

$$\sum_{i<j} \mathbb{E} \left[J_{ij} \tilde{\omega}(\sigma_i \sigma_j) \right] \underset{\downarrow \atop \text{WE}}{=} \sum_{i<j} \underbrace{\mathbb{E} \left[J_{ij}^2 \right]}_{= 1} \mathbb{E} \left[\partial_{J_{ij}} \tilde{\omega}(\sigma_i \sigma_j) \right]$$

$$\sum_{i<j} \mathbb{E} \left[\frac{\sum_{\underline{\sigma}} \sigma_i \sigma_j \frac{\beta \sqrt{t}}{\sqrt{N}} \sigma_i \sigma_j e^{-\beta \tilde{H}}}{\tilde{Z}} - \frac{\sum_{\sigma} \sigma_i \sigma_j e^{-\beta \tilde{H}} \sum_{\underline{\sigma}} \sigma_i \sigma_j \frac{\beta \sqrt{t}}{\sqrt{N}} e^{-\beta \tilde{H}}}{\tilde{Z}^2} \right]$$

$$\rightarrow \frac{d\tilde{A}_N}{dt} = \frac{\beta}{N}\left[\frac{1}{2\cancel{\sqrt{t}}\sqrt{N}}\frac{\beta\cancel{\sqrt{t}}}{2\sqrt{N}}\left(N(N-1) - \langle q_{ab}^2\rangle N^2\right)\right.$$

$$\left. - \frac{c}{2\cancel{\sqrt{1-t}}} NC\cancel{\sqrt{1-t}}\,\beta\left(1-\langle q_{ab}\rangle\right)\right]$$

$$\underset{N\to\infty}{=} \frac{\beta^2}{\cancel{N}}\left[\frac{1}{4\cancel{N}}\cancel{N^2}\left(1-\langle q_{ab}^2\rangle\right) - \frac{c^2\cancel{N}}{2}\left(1-\langle q_{ab}\rangle\right)\right]$$

$$= \frac{\beta^2}{4}\left[1-\langle q_{ab}^2\rangle - 2c^2 + 2c^2\langle q_{ab}\rangle\right]$$

$$\langle (q_{ab}-\bar{q})^2\rangle = \qquad \bar{q} \text{ constant}$$

$$\langle q_{ab}^2 - 2\bar{q}q_{ab} + \bar{q}^2\rangle =$$

$$= \langle q_{ab}^2\rangle - 2\bar{q}\langle q_{ab}\rangle + \bar{q}^2$$

by choosing $c^2 = \bar{q}$

$$= \frac{\beta^2}{4}\left[-\langle (q_{ab}-\bar{q})^2\rangle + 1 - 2\bar{q} + \bar{q}^2\right]$$

$$= \frac{\beta^2}{4}\left[-\langle (q_{ab}-\bar{q})^2\rangle + (1-\bar{q})^2\right]$$

$$\to A = \log 2 + \mathbb{E} \log\cosh [\beta \sqrt{\bar q}\, \eta] +$$

$$+ \frac{\beta^2}{4} (1-\bar q)^2 - \int_0^1 \frac{\beta^2}{4} \langle (q_{ab} - \bar q)^2 \rangle_t \, dt$$

$\bar q$ $\quad \lim_{N\to\infty} P_N(q) = \delta(q-\bar q)$

(hp $\langle \cdot \rangle_t$ indep. of t) analogously to what done for HM-LL regime we neglect integral as $N\to\infty$

$$A = \log 2 + \mathbb{E} \log\cosh [\beta \sqrt{\bar q}\, \eta] + \frac{\beta^2}{4} (1-\bar q)^2$$

$$\frac{\partial A}{\partial \bar q} = 0 \qquad -\frac{\beta^2}{2}(1-\bar q) + \mathbb{E} \tanh [\beta \sqrt{\bar q}\,\eta] \frac{1}{2\sqrt{\bar q}} \beta\eta = 0$$

$$\bar q = 1 - \int \frac{\eta}{\beta\sqrt{\bar q}} \tanh [\beta \sqrt{\bar q}\, \eta] \frac{e^{-\eta^2/2}}{\sqrt{2\pi}} d\eta$$

by part

$$= 1 - \left[\frac{1}{\beta\sqrt{\bar q}} \tanh [\beta\sqrt{\bar q}\,\eta] \frac{e^{-\frac{\eta^2}{2}}}{\sqrt{2\pi}} \right|_{-\infty}^{+\infty}$$

$$- \int \frac{e^{-\eta^2/2}}{\sqrt{2\pi}} \left[1 - \tanh^2 [\beta\sqrt{\bar q}\,\eta]\right] d\eta =$$

$$= \int_{-\infty}^{+\infty} \tanh^2 \left[\beta \sqrt{\bar{q}}\, \eta \right] \frac{e^{-\eta^2/2}}{\sqrt{2\pi}}\, d\eta$$

Expand around $q \simeq 0$

$$q \simeq \int_{-\infty}^{+\infty} \frac{d\eta}{\sqrt{2\pi}}\, e^{-\frac{\eta^2}{2}} \left(\beta \sqrt{\bar{q}}\, \eta \right)^2 = \beta^2 \bar{q} \rightarrow \beta_c = 1$$

HW : $J_{ij} = (1 - \delta_{ij}) \left[\frac{J_0}{N} + \frac{J}{\sqrt{N}} z_{ij} \right] \quad z_{ij} \sim \mathcal{N}(0,1)$

preparatory for HN

HIGH-LOAD HOPFIELD MODEL

lemmas on universality of quenched noise

- Carmona & Hu '06 : Any symmetric probability distribution $P(J)$ with finite moments can be chosen for J_{ij} without modifying the free energy of the system apart from corrections vanishing in the TDL (e.g. Rademacher)

- Genovese '22 : Same result but for bipartite SGs
- Agliari et al '17 : When interested in the retrieval of one pattern, say $\underline{\xi}^1$, we can replace the remaining $K-1$ by Gaussian vectors, without modifying the free energy of the system in the TDL.

$$\underline{\xi}^1 \in \{-1,+1\}^N \qquad \underline{\xi}^{\mu>1} \sim \mathcal{N}(0, I_N)$$

$\mu=1$ marked condensed

$\mu>1$ unmarked uncondensed $\rightarrow$ contribute to slow noise

$$Z_{N,\beta}(\underline{\xi}) = \sum_{\underline{\sigma}} \exp\Big\{ \overbrace{\frac{\beta}{2N}\sum_{i,j}\sigma_i\sigma_j\xi_i^1\xi_j^1}^{\text{FM-like}} + \overbrace{\frac{\beta}{2N}\sum_{i,j}\sum_{\mu>1}\sigma_i\sigma_j\xi_i^\mu\xi_j^\mu}^{\text{difficult}} \Big\}.$$

$$\exp\Big\{-\frac{\beta}{2N}\sum_{i,\mu}(\xi_i^\mu)^2\Big\} \rightarrow \text{diagonal}^*, \text{ neglected and resumed in the end}$$

$$\overset{HS}{=} \sum_{\underline{\sigma}} \exp\Big\{\frac{\beta N}{2} m_1^2\Big\} \cdot \int \prod_{\mu>1} d\mu(z_\mu) \exp\Big\{\sqrt{\frac{\beta}{N}}\sum_{\mu>1}\sum_i \xi_i^\mu \sigma_i z_\mu\Big\} \cdots$$

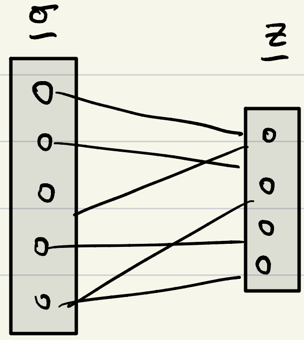
$$d\mu(z_\mu) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{z_\mu^2}{2}\Big)$$

$$* \quad \frac{1}{N}\mathbb{E}\log Z = \frac{1}{N}\mathbb{E}\Big[-\frac{\beta}{2N}\sum_{i,\mu}(\xi_i)^2\Big]$$

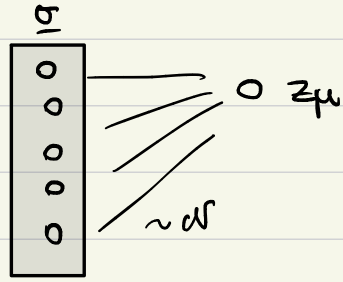
$$= -\frac{1}{N}\frac{\beta}{2\cancel{N}}\cancel{N}\cdot K = -\frac{\beta\alpha}{2}$$

Remark

- I can always treat the contribution $\mu=1$ as a FM one, by applying a Mattis transformation (I can apply it only once for $\mu=1$)

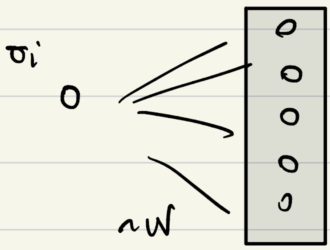


- The first contribution shall be treated analogously to the low-load case



- Linearization → Bipartite hybrid spin-glass

The sum acting on each neuron involve $O(N)$ neurons, therefore, I expect that TLC kicks in and the field distribution stabilizes over a Gaussian. Of course MF is mandatory for the application of the interpolation.



$\mathcal{N}(\mu_x, \sigma_x^2)$ $\mathcal{N}(\mu_y, \sigma_y^2)$

$z = x \cdot y$; ρ = coeff. corr.

$\mathbb{E}[z] = \mu_x \mu_y + \rho\, \sigma_x \sigma_y$

$\mathrm{Var}[z] = \mu_x^2 \sigma_y^2 + \mu_y^2 \sigma_x^2 + \sigma_x^2 \sigma_y^2 (1+\rho^2)$

$$\tilde{H}_N(\sigma; \xi, \eta, \theta) =$$

$$= \underbrace{-t \frac{N}{2} m_1^2 - (1-t)\beta \psi N m_1}_{\text{FM-like}}$$

$$\underbrace{- \frac{\sqrt{t}}{\sqrt{N}} \sum_{i,\mu} \xi_i^\mu \sigma_i z_\mu - \sqrt{1-t}\left[A \sum_i \eta_i \sigma_i + B \sum_\mu \theta_\mu z_\mu\right]}_{\text{Bipartite SG}}$$

$\eta_i, \theta_\mu \sim \mathcal{N}(0, 1)$ to mimic the statistic of the internal fields

$\psi, A, B, C \in \mathbb{R}$ to be determined a posteriori

By looking at $\tilde{\mathcal{H}}$ we can anticipate the order parameters

$\psi \to m_1$

$A, B, C \to$ overlap

- m_1
- replica overlap $q_{ab} := \frac{1}{N} \sum_i \sigma_i^{(a)} \sigma_i^{(b)} \in [-1, +1]$

It plays as a proximity measure

- replica overlap for Gaussian neurons $p_{ab} := \frac{1}{K} \sum_\mu z_\mu^{(a)} z_\mu^{(b)} \in \mathbb{R}$

$$A^{RS} = -\frac{\alpha\beta}{2} + \log 2 + \mathbb{E} \log\cosh \left[\beta \bar{m} \xi^1 + \sqrt{\alpha\beta\bar{p}}\, \eta \right] - \frac{\alpha}{2} \log(1-\beta(1-\bar{q}))$$

$$+ \frac{\alpha\beta\bar{q}}{2(1-\beta(1-\bar{q}))} - \frac{\bar{m}^2\beta}{2} + \frac{\alpha\beta\bar{p}}{2}(\bar{q}-1)$$

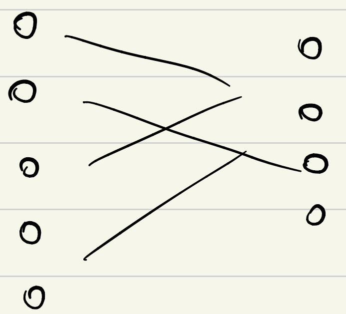
where $\bar{m}$, $\bar{p}$ and $\bar{q}$ must be extremal, namely, they must fulfil

$$\begin{cases} \dfrac{\partial A^{RS}}{\partial \bar{m}} = 0 \\ \dfrac{\partial A^{RS}}{\partial \bar{q}} = 0 \\ \dfrac{\partial A^{RS}}{\partial \bar{p}} = 0 \end{cases} \rightarrow \begin{cases} \bar{m} = \mathbb{E}\left[\xi^1 \tanh(\beta \bar{m} \xi_1 + \sqrt{\alpha\beta\bar{p}}\, \eta) \right] \\ \bar{q} = \mathbb{E}\left[\tanh^2(\beta \bar{m} \xi_1 + \sqrt{\alpha\beta\bar{p}}\, \eta) \right] \\ \bar{p} = \dfrac{\beta\bar{q}}{[1-\beta(1-\bar{q})]^2} \end{cases}$$

Remark

- As expected, we recover the sc eqs. originally found by AGS in '85 by replica trick.
- There is some analogies with low-load solution
  $m = \mathbb{E}[\xi \tanh(\xi m \beta)]$ $q = \int d\mu(z) \tanh^2(\beta\sqrt{q}\, z)$
- $\sqrt{\alpha\beta\bar{p}}$ plays as a noise as it is multiplied by a r.v. η.
  In fact, this is tuned by α which is the control parameter for the slow noise
- $\bar{p}$ corresponds to an algebraic expression, that is, we can substitute it with an expression in terms of β and $\bar{q}$. Namely, we do not have a SC eq. for $\bar{p}$ consistently with the fact that $\bar{p}$ stems from auxiliary variables (it is not an intrinsic order param.)

Resume Hopfield's model and emphasize that the HS transformation basically highlights a thermodynamic equivalence (meaning that their partition functions are equiv.) between the HN and a bipartite hybrid SG



This structure is reminiscent of RBMs where binary (gaussian) spins are interpreted as visible (hidden) neurons

$\underline{\sigma} \in \{-1,+1\}^N$ $\quad \underline{z} \sim \mathcal{N}(0, I_N \beta^{-1})$

$\downarrow \qquad\qquad \downarrow$

$\underline{v} \qquad\qquad \underline{h}$

$$Z^{(HN)}_{N,\beta}(\xi) = \sum_{\underline{\sigma} \in \{-1,+1\}^N} e^{\frac{\beta}{2N} \underline{\sigma}^T (\xi \xi^T) \underline{\sigma}} = \int \sum_{\underline{v} \in \{-1,+1\}^N} e^{\frac{\beta}{\sqrt{N}} \underline{v}^T \xi \underline{v}} \frac{e^{-\beta \underline{h}^2/2}}{\sqrt{2\pi/\beta}} d\underline{h}$$

$$= Z^{(RBM)}_{N,\beta}(\xi)$$

HW! prove it in the presence of external field

$$\int \sum_{v} e^{\beta/\sqrt{N}\, \underline{v}^T \xi \underline{h}} \frac{e^{-\beta \underline{h}^2/2}}{\sqrt{2\pi/\beta}} e^{\beta \underline{\theta}^T \underline{v}} e^{\beta \underline{\lambda}^T \underline{h}} d\underline{h}$$

$$= \sum_{\underline{v}} \prod_{\underline{\mu}} e^{\beta \underline{\theta}^T \cdot \underline{v}} e^{\frac{\beta}{2} \left(\frac{\underline{\xi}_\mu^T \underline{v}}{\sqrt{N}} + \lambda_\mu \right)^2} =$$

$$= \sum_{\underline{v}} e^{\beta \underline{\theta}^T \cdot \underline{v} + \frac{\beta}{2} \frac{\underline{v}^T \xi \cdot \xi^T \underline{v}}{N} + \frac{\beta}{2} \underline{\lambda}^2 + \underbrace{\beta \underline{v}^T \xi^T \underline{\lambda}}_{\sum_\mu \sum_i \xi_\mu^i \lambda_\mu v_i}}$$

the equivalence can be generalized to the case

binary / binary
gaussian / gaussian $\Big\}$ Barra, Sollich et al '18

(in general the retrieval region shrinks as you move away from the binary / gaussian case) – holds for all degrees of freedom (z, σ, ξ)

→ In the following I will use the most convenient framework according to the case.

I stress that the equivalence, at this stage, is a formal one I mean, it does not encompass any of the learning features characterizing the RBM, it only considers a trained RBM where, as an effect of training, weights have been quenched as ξ_i^μ and, in this setting, the RBM exhibits retrieval capabilities in the retrieval region outlined when investigating the HN.

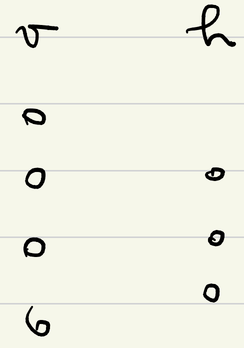
(static picture over the weights)

In the next, I will try to show that this equivalence is not only formal. To this goal I will first review the RBM.

RESTRICTED BOLTZMANN MACHINES

RBMs are shallow, two-layer, nets

The architecture of an RBM is shown here and it is made up of a set of visible units v_i, $i=1,\dots,N$ (so called because they usually collect observed data and a set of hidden units h_μ, $\mu=1,\dots,K$ connected by weight $W_{i\mu}$. External field / bias $\theta_{i,\mu}$



The state of the units in an RBM are determined via the BG distribution $\quad P(\underline{v},\underline{h};W,\theta) = \dfrac{1}{Z_\beta(W,\theta)} \exp\left[-\beta \mathcal{H}(\underline{v},\underline{h})\right]$

where, in the simplest case

$$\mathcal{H}(\underline{v},\underline{h}) = -\sum_i \theta_i v_i - \sum_\mu \theta_\mu h_\mu - \sum_{i,j} W_{i\mu} v_i h_\mu$$

$\underline{W} \in \mathbb{R}^{N\times K}$, $\underline{\theta} \in \mathbb{R}^{N+K}$ are treated as parameters

Name BM → prob. distr.

Adjective R → topology

they constitute a paradigmatic model for ML

they can be trained in sup or unsup. way, depending on the task (e.g. dimensionality reduction, classification, generation, regression). Training here means properly setting W and θ s.t. the model distribution does reproduce unknown prob. distrib. over datasets provided as in input.

Hystorical note: introduced in '86 by Smolensky under the name harmonium, but the stat. mech. approach finally prevailed over information-theory driven one.

Interest in RBMs

RBMs, much as like the HN, is a simple model as for the kind of affordable tasks, but they play as an harmonic oscillator to start the investigation of ML.

Mézard '17
Tubiana, Cocco, Monasson '19
Decelle '17, '18, '20s ....
Roudi '20s ...
...

- Role as building blocks of a class of deep NNs
- Paradigmatic models for learning that allow us to inspect foundational issues, such as the role of the nature of units (eg. binary, continuous) and of the related activation functions (eg ReLu, linear ..) → see Roudi et al. '21 '25
- RBMs have been shown to be universal approximators of discrete prob. distrib., given a sufficient # of hidden nodes. This result due to Le Roux and Bengio '08 was built on Cybenko's theorem '89 and later generalizations.

unknown

At the end of the day (ultimately) the task is to learn an joint probability distribution $q(x,y)$ where x plays as an input (query) and y plays as the output, or, in the case of unsupervised task we learn a probability $q(x)$, think of generative tasks.

We want the model distribution $p(\underline{v},\underline{h})$ to mimic this distribution.

In the following, to simplify the treatment, I will assume that $\underline{v}$, $\underline{h}$ are both binary

- supervised

$(\underline{x}^{(i)}, y)$ $\quad i = 1, \dots, M$ $\quad$ training dataset

eg. MNIST classification

$\underline{x} \in [0,1]^{28} \times [0,1]^{28}$ $\quad$ square picture, size 28 pixel each pixel gray scale

$y \in \{0, \dots, 9\}$

$(\underline{x}, y) \sim_{iid} q(\underline{x}, y)$

recast the embedding space to $\{-1,+1\}^N \otimes \{-1,+1\}^K$

to compare $q(\underline{x}, y)$ with $p(\underline{v}, \underline{h}; W, \theta)$

pictures are flattened and binarized (thresholded and mapped to ± 1)

$\underline{x} \rightsquigarrow \underline{v} \in \{-1, +1\}^N \quad N = 784$

$y \rightsquigarrow \underline{h} \in \{-1, +1\}^K \quad K = 10$

one-hot vector notation $\left.\begin{array}{l} h_\mu = 1 \\ h_{\nu \neq \mu} = -1 \end{array}\right\} \to y = \mu$

- unsupervised, eg MNIST generative

$\underline{x} \rightsquigarrow \underline{v}$

$\underline{h}$ used latent variables

Having recast $q(x, y)$ into $q(\underline{v}, \underline{h})$ we can introduce a suitable

Performance measure: KL distance / cross entropy

$$D(q \| P) = \sum_{\underline{v}, \underline{h}} q(\underline{v}, \underline{h}) \log_2 \frac{q(\underline{v}, \underline{h})}{p(\underline{v}, \underline{h})}$$

For unsupervised tasks we should first marginalize $p(\underline{v}, \underline{h})$ over $\underline{h}$ to get $p(\underline{v})$, whence

$$D(q \| P) = \sum_{\underline{v}} q(\underline{v}) \log_2 \frac{q(\underline{v})}{p(\underline{v})}$$

Let's recall $p(\underline{v},\underline{h}) = \dfrac{e^{-\beta E(\underline{v},\underline{h})}}{Z}$

$$E(\underline{v},\underline{h}) = -\sum_{i,\mu} v_i W_{i\mu} h_\mu - \sum_i \theta_i v_i - \sum_\mu \theta_\mu h_\mu$$

$P(\underline{v},\underline{h}\,;W,\theta)$ parametrized by W and θ that need to be estimated s.t. $D(q\|P)$ is minimized.

Gradient descent

$$\Delta J_{ij} = -\varepsilon \frac{\partial D}{\partial J_{ij}} \qquad \Delta\theta_i = -\varepsilon \frac{\partial D}{\partial \theta_i}$$

where ε is the learning rate or steps.

By applying this updating rules $D \downarrow$

We will derive explicitly these updating rules, which will clearly involve a training dataset

$\{(\underline{x}^{(i)}, y^{(i)})\}_{i=1,\dots,M}$ or $\{\underline{x}^{(i)}\}_{i=1,\dots,M}$ according to the case

Before moving to calculations, let me conclude to overview by recalling that, by implementing these learning rules we will reach estimates $\hat{W}$, $\hat{\theta}$.

Supervised case

$$\hat{\underline{W}}, \hat{\underline{\theta}} \quad \text{s.t.} \quad p(\underline{v},\underline{h}; \hat{J}, \hat{\theta}) \simeq q(\underline{v},\underline{h})$$

$$x_{test} \to \underline{v}' \to \hat{h}(\underline{v}') \ \text{r.v.}$$

$$\text{Prob}[\underline{h} \mid \underline{v} = \underline{v}'] = p(\underline{h} \mid \underline{v}' ; \hat{W}, \hat{\underline{\theta}})$$

Unsuper.

$$\hat{W}, \hat{\underline{\theta}} \quad \text{st.} \quad p(\underline{v}; \hat{W}, \hat{\theta}) \simeq q(\underline{v})$$

$$\text{Prob}[v] = \sum_h p(\underline{v},\underline{h}; \hat{w}, \hat{\theta})$$

Grand-mother cell

the one-hot vector notation implies that the hidden layer is made on a # of neurons corresponding to the # of stored memories. Also, it implies that we have a specific hidden neuron for each of the stored memories. This setting basically reproduces the so-called grand-mother cell theory. (50's) According to this theory, we have an hypothetical neuron that represents a complex, but specific, complex or object. It activates when a person sees, hears, or otherwise sensibly discriminates a specific entity, such as their grandmother.

learning rules for sup.

$$p(\underline{v},\underline{h};\underbrace{W,\theta}_{\lambda}) = \frac{e^{-\beta E(v,h;W,\theta)}}{Z_\beta(W,\theta)} = \frac{Z_\beta(v,h;W,\theta)}{Z_\beta(W,\theta)}$$

In the following I will adopt a lighter notation omitting dependences if not ambiguous.

$$\frac{\partial}{\partial\lambda} D(q\|P) = -\frac{1}{\log 2}\sum_{\underline{v},\underline{h}} q(\underline{v},\underline{h})\frac{\partial}{\partial\lambda}\log P(\underline{v},\underline{h})$$

$$= -\frac{1}{\log 2}\sum_{\underline{v},\underline{h}} q(\underline{v},\underline{h})\frac{\partial}{\partial\lambda}\left[\log Z(\underline{v},\underline{h}) - \log Z\right] =$$

$$= -\frac{1}{\log 2}\sum_{\underline{v},\underline{h}} q(\underline{v},\underline{h})\left[-\beta\frac{\partial E}{\partial\lambda} - \sum_{\underline{v}',\underline{h}'}\frac{e^{-\beta E(\underline{v}',\underline{h}')}}{Z}(-\beta)\frac{\partial E}{\partial\lambda}\right]$$

$$= \frac{\beta}{\log 2}\left[\underbrace{\sum_{\underline{v},\underline{h}} q(\underline{v},\underline{h})\frac{\partial E}{\partial\lambda}}_{\mathbb{E}_q\left[\frac{\partial E}{\partial\lambda}\right]} - \underbrace{\sum_{\underline{v},\underline{h}} q(\underline{v},\underline{h})}_{1}\underbrace{\sum_{\underline{v}',\underline{h}'}\frac{\partial E}{\partial\lambda}p(\underline{v}',\underline{h}')}_{\mathbb{E}_p\left[\frac{\partial E}{\partial\lambda}\right]}\right]$$

when $\lambda = W_{i\mu}$ $\rightarrow$ $\frac{\partial E}{\partial\lambda} = -v_i h_\mu$

$\lambda = \theta_\mu$ $\rightarrow$ $\frac{\partial E}{\partial\theta_\mu} = -h_\mu$

$$\Delta W_{i\mu} = + \frac{\varepsilon\beta}{\log 2} \Big[\mathbb{E}_q [v_i h_\mu] - \mathbb{E}_p [v_i h_\mu] \Big]$$

$$\Delta \theta_\mu = \frac{\varepsilon\beta}{\log 2} \Big[\mathbb{E}_q [h_\mu] - \mathbb{E}_p [h_\mu] \Big]$$

$$\Delta \theta_i = \frac{\varepsilon\beta}{\log 2} \Big[\mathbb{E}_q [v_i] - \mathbb{E}_p [v_i] \Big]$$

Learning means that $\mathbb{E}_q[\cdot] \approx \mathbb{E}_p[\cdot]$

interpretable and reasonable condition

$\Rightarrow$ the model can capture 1st and 2nd order correlations (at least within a certain threshold level)

$$E(\underline{s}) = - \sum_{i<j} J_{ij} s_i s_j - \sum_{i<j<k} J_{ijk} s_i s_j s_k$$ not necessarily restricted

ΔJ_{ij} ..

ΔJ_{ijk} ..

$\rightarrow$ higher-order correlations among neurons (MEA)

$\rightarrow$ higher accuracy and control

Practical implementation : $\mathbb{E}_q \rightarrow$ empirical average over training dataset

$\mathbb{E}_p \rightarrow$ MCMC

Learning rules for unsupervised

$$p(\underline{v}; W, \theta) = \frac{Z_\beta(v; W, \theta)}{Z_\beta(W, \theta)}$$

$$= \frac{\sum_{\underline{h}} e^{-\beta E(v, h)}}{Z_\beta(W, \theta)}$$

$$\frac{\partial D(q\|p)}{\partial \lambda} = -\frac{1}{\log 2} \sum_{\underline{v}} q(\underline{v}) \frac{\partial}{\partial \lambda}\left[\log Z(\underline{v}) - \log Z\right]$$

$$= \frac{+\beta}{\log 2} \sum_{\underline{v}} q(\underline{v}) \left[\sum_{\underline{h}} \underbrace{\frac{1}{Z(\underline{v})} e^{-\beta E}}_{p(\underline{h}|\underline{v})} \frac{\partial E}{\partial \lambda} - \sum_{\underline{v}', \underline{h}'} p(\underline{v}', \underline{h}') \frac{\partial E}{\partial \lambda} \right]$$

$$= \frac{\beta}{\log 2} \left[\underbrace{\sum_{\underline{v}, \underline{h}} q(\underline{v}) p(\underline{h}|\underline{v}) \frac{\partial E}{\partial \lambda}} - \mathbb{E}_p\left[\frac{\partial E}{\partial \lambda}\right] \right]$$

$$= \frac{\beta}{\log 2} \left\{ \mathbb{E}_{p|q}\left[\frac{\partial E}{\partial \lambda}\right] - \mathbb{E}_p\left[\frac{\partial E}{\partial \lambda}\right] \right\}$$

Again the rule can be generalized to more complex energy models.
Again the rule implies a MCMC and an empirical evalution.
Again the rule is perfectly interpretable.
The first and the second moments (or even higher) exhibit the same values irrespective of whether the visible layer was left free or was quenched as sampled from the target distrib.
That is, in this sense, there is no difference between visible config. sampled from p and from q → consistency with generative tasks.

Remarks

We might have obtained analogous learning rules by maximizing the log-likelihood $\mathcal{L}(\lambda \mid \underbrace{\{(\underline{x}^{(i)}, y)\}_{i=1,\dots,M}}_{D_i})$ of $\{\begin{matrix}\text{sup}\\ \text{unsup}\end{matrix}$

the parameter λ with respect to the given dataset.

$$p(\{D_i\}_{i=1}^{M};\lambda) = \prod_i p(D_i;\lambda) = \exp\left[\sum_i \log p(D_i;\lambda)\right]$$

$$\mathcal{L}(\lambda \mid \{D_i\}) = \frac{1}{M}\sum_{i=1}^{M} \log p(D_i;\lambda) \qquad \hat{\lambda} = \arg\max_{\lambda} \mathcal{L}(\lambda \mid \{D_i\})$$

In fact, it is easy to prove that maximizing the average log-likelihood of the model's parameters with respect to the data is asymptotically equivalent (in the # of data points) to minimizing the KL divergence between the model and data distributions → HW: check it

$$\mathbb{E}_{data} \qquad \mathbb{E}_p$$

$$\Rightarrow \quad \Delta w_{i\mu} = \eta\,(\underbrace{\langle v_i h_\mu \rangle_{data}}_{\substack{\text{positive}\\ \text{phase}}} - \underbrace{\langle v_i h_\mu \rangle_{model}}_{\substack{\text{negative}\\ \text{phase}}}) \rightarrow \text{intractable}$$

This challenge has led to extensive research on approx methods for computing the average in the negative phase. These approx can be numerical methods that accelerate sampling, analytical techniques for estimating expectations, or hybrid approaches that combine both. Generally, numerical approx rely on various forms of sampling from a distribution, while analytical methods are based on MF approx.

CONTRASTIVE DIVERGENCE

In practice, generating unbiased samples is extremely time-consuming because, after each weight update, the RBM must reach equilibrium before reliable samples can be drawn. Mixing time is often prohibitively long.

→ alternative sampling techniques, eg. n-step CD

Instead of minimizing the KL divergence between p_{data} and p_{model}, as in standard ML estimation, CD_n minimizes the so called contrastive divergence, defined as the difference between two KL divergences:

$D_{KL}(p_{data} \| p) - D_{KL}(p_n \| p)$

where p_n represents the distribution of a Markov chain initialized from the data distribution and run for n alternating Gibbs sampling steps.

when $n \to \infty$, $p_n \to p$ and $D_{KL}(p_n \| p) \to 0$ and we recover the standard ML objective. Taking the gradient of the current objective function:

$$\Delta W_{i\mu} \propto \left(\overset{\mathbb{E}_0}{\langle v_i h_\mu \rangle_{data}} - \overset{\mathbb{E}_n}{\langle v_i h_\mu \rangle_{p_n}} \right)$$

as the expectation of the exact model distrib. cancel out.

The CD_n algorithm is relatively simple and computationally efficient.

However, the theoretical properties remain not fully understood.

- the quality of the gradient estimate is highly dependent on the dataset and on the number of steps n
- the gradient can be weakly or strongly biased, potentially leading to suboptimal learning outcomes

Decelle et al '22

To mitigate these issues → several variants, e.g. PCD

Tieleman '08

In general, the fixed point of CD differ from those of ML and thus CD is a biased algorithm. However, the bias is generally small, since CD converges typically very near an ML optimum; this small bias can be eliminated by running ML for a few iterations after CD, i.e. using CD as an initialisation strategy for ML.

While training algorithms are crucial for successful RBM learning their mathematical description is hard. Statistical physics has been used to understand the training dynamics in RBM (Decelle et al '17, '18) and to develop strategies to improve sampling and training efficiency.
However, useful information can be extracted from theoretical approaches that identify regions in parameter space where learning is facilitated, as well as limitations to the model's expressive power and trainability.

SUPERVISED AND UNSUP. HEBB

Now we resume the HN ⟺ RBM duality and notice that it does not encompass any learning protocol as it is based on the perfect knowledge of the reality that we want to capture.

6

My goal now is to relax this hp and make the equivalence more tight.

$\{\underline{\xi}^\mu\}_{\mu=1,\dots,K} \rightarrow \{\underline{\xi}^{\mu,A}\}_{\substack{\mu=1,\dots,K \\ A=1,\dots,M}}$ prone to error

"archetype" "examples"

where $\underline{\xi}^{\mu,A} = \underline{\xi}^\mu \odot \underline{\chi}^{\mu,A}$ $\quad \chi_i^{\mu,A} \underset{i.i.d}{\sim} \text{Rad}(r)$, $r \in [0,1]$

$$P(\chi_i^{\mu,A} = \pm 1) = \frac{1 \pm r}{2}$$ (Hadamard product)

M quantity, r quality Fontanari '90

two protocols $\begin{cases} \text{supervised} \\ \text{unsupervised} \end{cases}$

- supervised : μ is disclosed by a teacher
  → group examples belonging to the same class μ.

$$J_{ij}^{sup} = \sum_\mu \underbrace{\frac{\sum_a \xi_i^{\mu,a}}{M}}_{\substack{\text{estimator} \\ \text{for } \xi_i^\mu}} \frac{\sum_b \xi_i^{\mu,b}}{M} = \sum_\mu \bar{\xi}_i^\mu \bar{\xi}_j^\mu$$

- unsupervised : μ is veiled

$$J_{ij}^{unsup} = \frac{1}{M} \sum_\mu \sum_a \xi_i^{\mu,a} \xi_j^{\mu,a}$$

I treat each examples of the KM as a single pattern

pixel-to-pixel correlation averaged over examples

first focus on supervised case where the analogy
RBM – HN is stricter (know how many hidden neurons)

Agliari, Barra, Kanter '22
Alemanno Aquaro

Control parameters : $\underbrace{M, r}, \beta, \alpha$

$$\rho := \frac{1-r^2}{M r^2}$$

$$H(\xi_i^\mu \mid \{\xi_i^{\mu,A}\}_{A=1,\dots,M}) \underset{\uparrow}{=} f(\rho) \text{ monotonically increasing}$$

$$\text{Prob}\left[\text{sgn}\left(\sum_A \chi_i^{\mu,A}\right) = -1\right] \underset{M>>1}{\simeq} \left[1 - \text{erf}\,\frac{1}{\sqrt{2\rho}}\right]$$

$\rho \to 0$ recover standard Hopfield

when $r=1$ each example is a perfect copy of the archetype
when $M \to \infty$ (and r is finite) by the strong law of large numbers

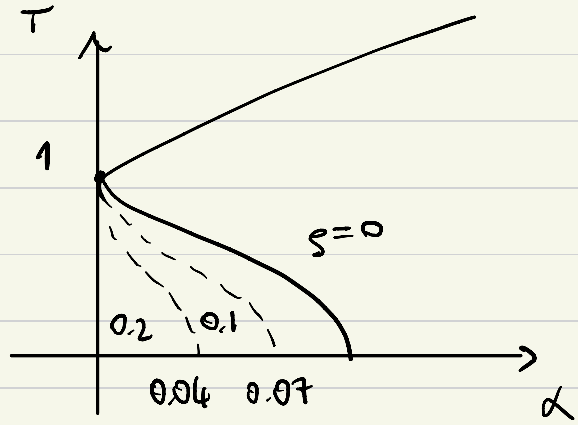
$$\bar{\xi}_i^\mu \xrightarrow[a.s.]{} \xi_i^\mu$$

Order parameters :
$$m_\mu = \frac{1}{N} \sum_i \xi_i^\mu \sigma_i$$

$$n_\mu = \frac{1}{r(1+\rho)} \frac{1}{NM} \sum_{i,A} \xi_i^{\mu A} \sigma_i$$

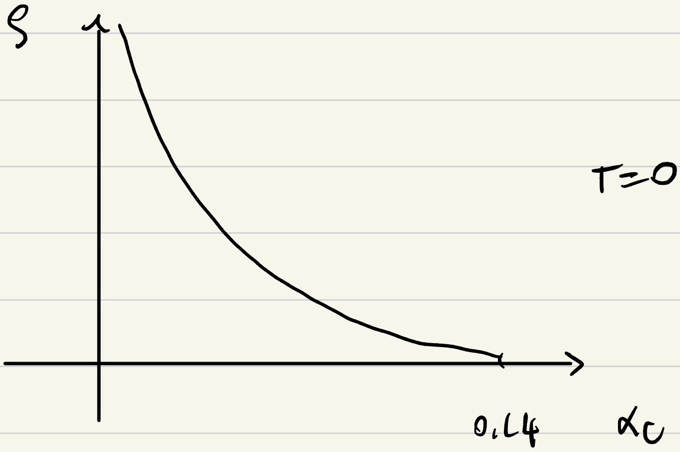
$$q_{12} = \frac{1}{N} \sum_i \sigma_i^{(1)} \sigma_i^{(2)}$$

$$\mathcal{H}_N(\underline{\sigma}; \xi) \propto \underline{n}^2$$



$\rho \uparrow$ retrieval region shrinks

(RS)



T=0

Get back to RBM

$$\bar{q}(\underline{\sigma},\underline{z}) = \sum_{\mu,A} \delta_{\underline{\xi}^{\mu,A},\underline{\sigma}} \cdot \delta_{\underline{z}^{(\mu)},\underline{z}}$$ empirical joint distribution

Kronecker deltas

$\delta_{\underline{\xi}^{\mu,A},\underline{\sigma}} \rightarrow$ non zero if $\xi_i^{\mu,A} = \sigma_i \quad i=1,\dots,N$

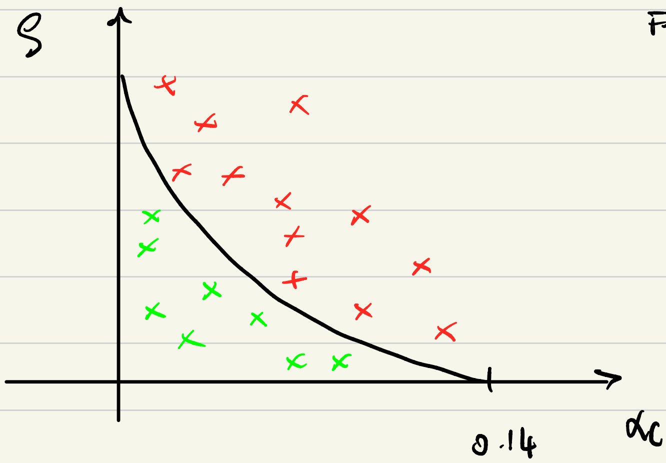
$$\underline{z}^{(\mu)} = (0,0,\dots,\underset{\mu}{1},0,\dots,0)$$

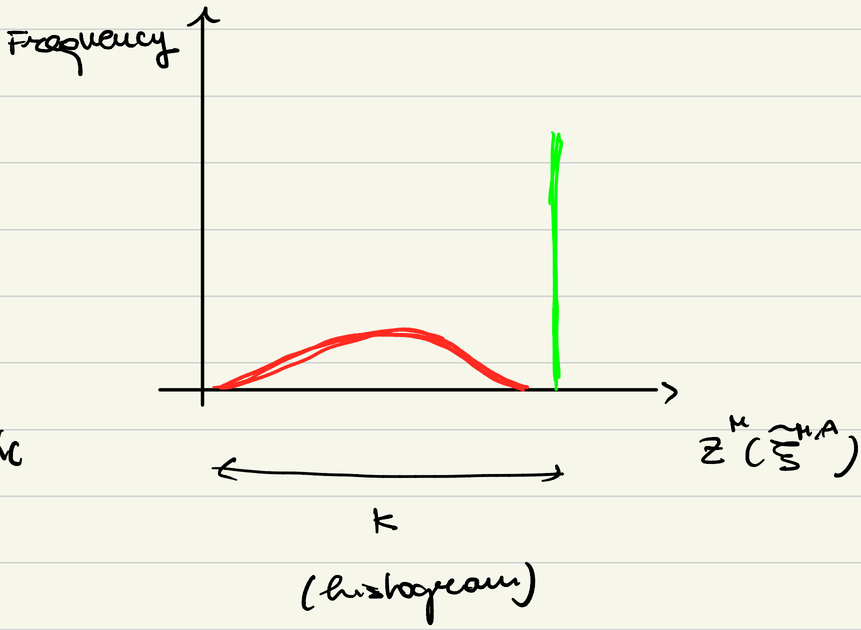
Gradient descent over KL

$W^{(\infty)} = \bar{\xi}$ if examples in different classes are $\perp$ (at least in the average) but we could anticipate this ...

$$Z_\beta^{(sup)}(\{\underline{\xi}^{\mu,A}\}) = e^{+\frac{\beta}{2N}\sum_{ij}\sum_{\mu}\bar{\xi}_i^{\mu}\bar{\xi}_j^{\mu}\sigma_i\sigma_j}$$

$$= \int d\mu(z)\, e^{\frac{\beta}{\sqrt{N}}\sum_{i\mu}\sigma_i\bar{\xi}_i^{\mu} z_\mu}\, d\underline{z}$$

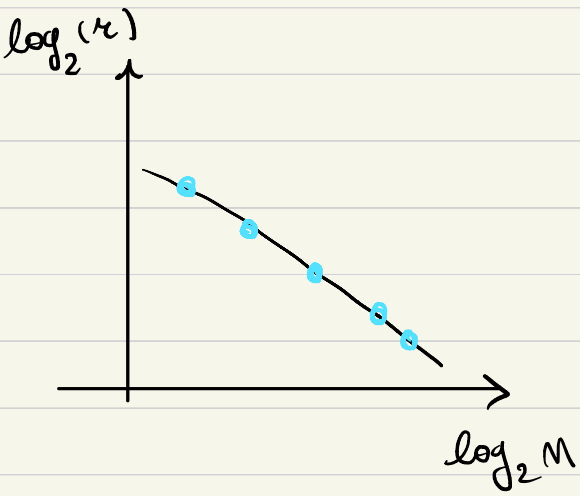




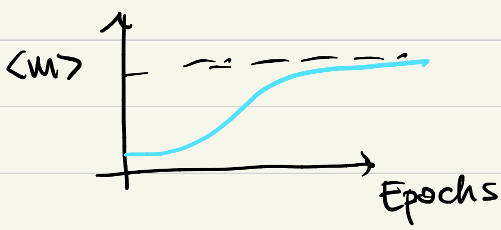
$\{\underline{\xi}^{\mu,A}\} \rightarrow W$

$\tilde{\xi}^{\mu,A} \rightarrow z^\mu \ \& \ z^{\nu\neq\mu}$

Set $\langle m_1\rangle \xrightarrow{SC} r, M$



— values of r and M such that $\langle m_1\rangle$ takes a certain value



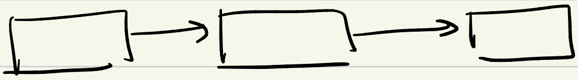
Set r and M $\rightarrow \langle m_1\rangle$

Remarks

- Any useful result beyond the toy setting of Rad. patterns?

For structured data $W=\underline{\xi}$ is no longer a fixed point for GD but effective pre-training (Smart - Zilman '21 - several decodes gain)

- Pre-process to make them orthogonal

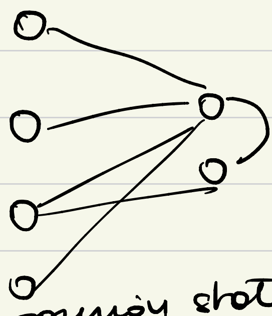


$\{\underline{\xi}^\mu\} \xrightarrow[GS]{} \{\tilde{\underline{\xi}}^\mu\}$ $\qquad J^{GS} = J^{K}$

- K. projection and the dreaming hebbian rule were useful to deal with correlated data and an effect evidenced in the HN was that they lower the likelihood of spurious states

Is there a related BM? I would call it DBM.

$$\mathcal{H}^{DBM}(\underline{\sigma};\underline{z};W,t) = -\frac{1}{\sqrt{N}}\sum_{i\mu} W_{i\mu}\,\sigma_i z_\mu + \frac{t}{1+t}\sum_{\mu<\nu} C_{\mu\nu}\, z_\mu z_\nu$$



$z_\mu \sim \mathcal{N}\left(0, \frac{1+t}{\beta}\right) \quad \mu = 1, \dots, k$

$\left.\begin{matrix}\Delta W\\ \Delta t\end{matrix}\right\}$ GD

Inhibition among hidden nodes avoids simultaneous activation of hidden neurons hence spurious states

- Task : minimize distance between $\underline{\sigma}$ and $\underline{\xi}^\mu$ for $\mu=1,\dots,k$

Suitable loss $(\underline{\sigma} - \underline{\xi}^\mu)^2 \propto -2\,\underline{\sigma}\cdot\underline{\xi}^\mu$ (Euclidean distance L2 norm)

$$L_\mu^{\pm} = \frac{1}{2N}\|\underline{\sigma} \pm \underline{\xi}^\mu\|_2^2 = 1 \pm m_\mu \quad \rightarrow \mathcal{H}(\sigma; J) = -N\sum_\mu (1 - L_\mu^+ L_\mu^-)$$

"MODERN" ARCHITECTURES : HYPERGRAPHS AND MODULAR GRAPHS

↳ Ramsauer et al. popularized referring to generalized Hopfield nets where neurons can be continuous variables and their activation function can take a polynomial or even exp. shape.
Highlight application in the context of transformers
→ upsurge of interest in Hopfield-like nets.

Here I am actually referring to HN where the embedding structure also plays a role and we want to leverage its topology to make the network able to afford more complex tasks.

In contrast with complete graphs → enhanced performance
→ different class of tasks

HYPERGRAPHS

the architecture includes hyperedges which connect two or more nodes

MEA : higher-order interactions → higher accuracy in capturing experimental data

BM : " → learn high-order correlations in the reality that we want to capture

$$\mathcal{H}_N^{(P)}(\underline{\sigma};\xi) = -\sum_{i_1 \dots i_p} J_{i_1 \dots i_p}\, \sigma_{i_1} \cdots \sigma_{i_p} =$$

$$= -\underbrace{\frac{1}{P!\,N^{P-1}}}_{\substack{\text{suitable}\\ \text{normalization}}} \underbrace{\sum_{i_1 \dots i_p}}_{\substack{\text{runs over all}\\ p\text{-tuples}}} \sum_{\mu} \xi_{i_1}^{\mu} \cdots \xi_{i_p}^{\mu}\, \sigma_{i_1} \cdots \sigma_{i_p}$$

(ordered → factor $\frac{1}{P!}$ normalizes over permutations)

this kind of nets : extensively investigated since the 80's
→ rapid overview, useful to recall different names by which they appeared in the literature

- Peretto & Niez '86 (multi-connected : starting from biology, they included all higher-order interactions
  $\to k_c \propto \#$ synopsis)
- Baldi & Venkatesh '87 (higher-order) $k_c \sim N^{p-1}$
- Gardner '87 $k_c \sim N^{p-1}/\log N$ perfect storage
- Abbott & Arian '87 (generalized networks)
  $$\mathcal{H} = -N^{1-p/2} \sum_\mu F\left[\frac{1}{\sqrt{N}} \sum_i \sigma_i \xi_i^\mu\right]$$
  with F polynomial of degree p : $F(x) = \sum_k A_k x^k$
- Bovier & Niederhauser '01 (p-spin interactions)
  Krotov & Hopfield '16
- Agliari et al '20s
  Demircigil et al '17
  Ramsauer et al '20

Super-linear performance $k_c \sim N^{p-1}$ ($p=2$ recover HN)

Expected : # stored patterns grows with # synapses

Scaling : # stored patterns $\propto \frac{1}{N}$ # synapses $= \frac{1}{N}\binom{N}{p}$

Intuition: as the size of the tensor J grows we have access to a larger number of weights where we can spread the information on patterns.

Proved at a rigorous level, here S2N

$\varphi_i^{(p)}(\underline{\xi}^1)\cdot\xi^1_{i_1} = S + R$

Focus on target pattern $\mu = 1$ and neuron i_1, wlg

Check stability $\underline{\sigma} = \underline{\xi}^1$

$$S = \frac{1}{p!\,N^{p-1}} \sum_{i_2 \dots i_p}^{N} \xi^1_{i_1}\xi^1_{i_2}\dots\xi^1_{i_p}\left(\xi^1_{i_1}\dots\xi^1_{i_p}\right) \sim 1$$

$$R = \frac{1}{p!\,N^{p-1}} \sum_{\mu=2}^{K} \sum_{i_2 \dots i_p}^{N} \left(\xi^1_{i_1}\xi^1_{i_2}\dots\xi^1_{i_p}\right)\left(\xi^\mu_{i_1}\xi^\mu_{i_2}\xi^\mu_{i_2}\dots\xi^\mu_{i_p}\right)$$

$$\sim \sqrt{\frac{K}{N^{p-1}}} \qquad \text{RW of length } K\cdot N^p$$

$\left[\lim_{N\to\infty} \frac{K}{N^{p-1}} = 0\right]$

$K \sim o(N^{p-1})$ low-load regime

$K \sim \Theta(N^{p-1})$ high-load regime

Check the robustness of stability vs corruption with the starting point now dealing with a fraction of flipped spins is a bit involved (# flips is even, field left invariant)

We apply an additional noise

$\sigma_i^{(0)} = \xi_i^\mu \rightarrow \xi_i^\mu + w\, z_{i\mu}$ where $z_{i\mu} \sim \mathcal{N}(0,1)$, $w \in \mathbb{R}$

$S \sim 1$

$$R \sim \sum_{n=1}^{p} \frac{w^n}{\sqrt{N^{n-1}}} + w^p \sqrt{\frac{K}{N^{p-1}}}$$

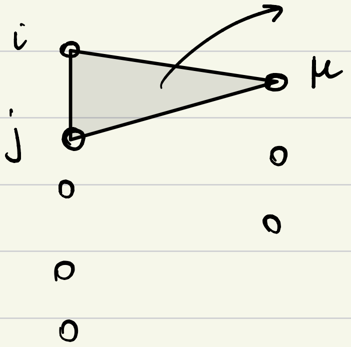
Retrieval is possible as long as $w \sim 1$, independently of K up to $K \sim \Theta(N^{p-1})$

We notice that, as long as p is even, HS transformation can be applied to these higher-order networks

$$\mathcal{Z}^{(p)}(\xi) = \sum_{\sigma} \prod_{\mu=1}^{K} \int d\mu(z_\mu) \, \frac{e^{-z_\mu^2/2}}{\sqrt{2\pi}} \, e^{\beta N^{\frac{2-p}{2}} \left(\prod_{j=1}^{p/2} \sum_{i_j} \sigma_{i_j} \xi_{i_j}^{\mu} \right) z_\mu}$$

$\to$ higher-order also in the RBM

e.g. $p = 4$ $\qquad \xi_i^\mu \xi_j^\mu = W_{ij\mu}$



$$\begin{array}{ccc} \text{HN} & & \text{RBM} \\ p & \longrightarrow & \frac{p}{2} + 1 \\ & & v \quad h \end{array}$$

$$\underbrace{\xi_{i_1}^\mu \cdots \xi_{i_{p/2}}^\mu}_{\substack{\text{weight of the plaquette} \\ i_1, \dots, i_{p/2}, \mu}} \longrightarrow \xi_{i_1}^\mu \cdots \xi_{i_{p/2}}^\mu + \underbrace{w z}_{\text{learning noise}} \qquad z \sim \mathcal{N}(0,1)$$

$S \sim 1$

$R \sim w N^{\frac{1}{2} - p/4} + w^2 \sqrt{K} N^{1/2 - \frac{p}{2}}$ $\qquad$ neglecting sub-leading contributions

Non trivial scaling, featuring an interplay between p, w, K, N

$K \sim N^a$ $\qquad$ $W \sim N^b$ $\qquad$ $a, b \in \mathbb{R}^+$

i) $a = p-1$ (high-load) $\qquad$ $b \leq 0$

ii) $a < p-1$, $a > 1$ (low-load) $\qquad$ $b \leq \frac{p-1-a}{4}$

iii) $a \leq 1$ $\qquad$ $b \leq \frac{p-2}{4}$

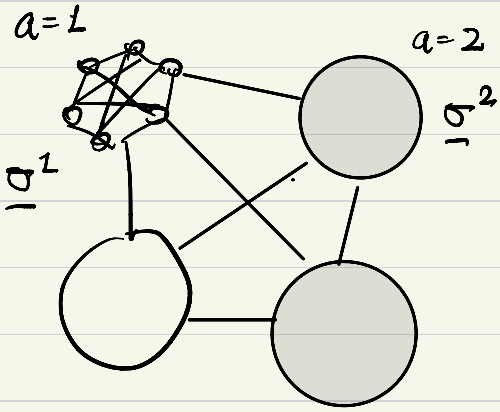
Interplay between the load and the affordable noise

If we accept to downsize the load then we can cope with a noise scaling with N and diverging in TDL.

Intuitive argument: if $a < p-1$, we can enjoy a redundant representation of patterns because the info encoded by patterns is spread over a large number of weights.

→ We can leverage redundancy to mitigate the effects of slow noise (moving from $k \sim N$ to $k \sim N^{p-1}$) and of learning noise.

Modular networks



Several modules / layer $a = 1, \dots, L$

$$\underline{\sigma}^{(a)} = (\sigma_1^a, \sigma_2^a, \dots, \sigma_{N_a}^a) \quad a = 1, \dots, L$$

$$\underline{\sigma}^{(a)} \in \{-1, +1\}^{N_a}$$

In general, K is the same for any layer

N_a size layer dependent

$$\{\underline{\xi}^{\mu,a}\}_{\mu=1,\dots,K} \quad \text{pattern set layer dependent}$$

$$\underline{\xi}^{\mu,a} \in \{-1, +1\}^{N_a} \quad \text{for } \begin{array}{l} a = 1, \dots, L \\ \mu = 1, \dots, K \end{array}$$

$$\mathcal{H}_N^L(\underline{\sigma}; g, \xi) = -\frac{1}{\sqrt{N_a N_b}} \sum_\mu \sum_{a,b} \sum_{i,j} g_{ab}\, \sigma_i^a \sigma_j^b \underbrace{\xi_i^{\mu,a} \xi_j^{\mu,b}}_{\text{Hebbian interaction}}$$

- $\frac{1}{\sqrt{N_a N_b}}$ normalization factor to keep $\mathcal{H}$ lin. ext. in size

  (hp: $\frac{N_a}{N_b} \xrightarrow[N_a \to \infty]{} \gamma$, γ finite, $\forall a, b$)

- $m_\mu^a = \frac{1}{N_a} \sum_i \sigma_i^a \cdot \xi_i^{\mu,a}$ Mattis magnetization

- $\mathcal{H}_N^L(\underline{\sigma}; g, \xi) = -\sqrt{N_a N_b} \sum_\mu \sum_{a,b} m_\mu^a\, g_{ab}\, m_\mu^b$

- $g \in \mathbb{R}^{L \times L}$

  $g_{ab} > 0 \rightarrow \underline{m}^a \cdot \underline{m}^b$ max

  $< 0 \rightarrow \underline{m}^a \cdot \underline{m}^b$ min

TASK: MIXTURE DISENTANGLEMENT

Agliari, Alessandrelli, Barra, Centonze, Ricci-Tersenghi '25

A net of nets can exhibit capacities that exceed the sum of the single net?

For $L=3$ we design

$$\mathcal{H}_N^{L=3}(\underline{\sigma}; g, \xi) = -N \sum_{\mu} \sum_{a,b} g_{ab}\, m_\mu^a m_\mu^b$$

$$g_{ab} = \begin{cases} 1 & a=b \rightarrow (\underline{m}^a)^2 \ \text{max} \\ -\lambda & a \neq b \rightarrow \underline{m}^a \cdot \underline{m}^b \ \text{min} \end{cases}$$

→ each layer is maximally aligned with a given pattern

→ two layers are orthogonal

$\lambda < \frac{1}{2}$ for energy sign

SM → SC equations

Usually this is the final point because then it is just a matter of solving them numerically

Here we can as well solve the SCs but this would not provide an answer to our question

In fact, we are wondering whether this particular state $\underline{\sigma}^{(1+2+3)}$ belongs to the attraction basin of $\underline{\sigma}^{(1,2,3)}$

Of course, we can rely on MC simulations, but we also want an analytical estimate.

We envisage two ways

- solve SC eqs fixed point iteration method initializing the state $\underline{\sigma}^{(\underline{1}+\underline{2}+\underline{3})}$ and checking if the detected solution is the disentangled one.

Subtle: not perfectly stable, sensitive on initial condition (when I implement some perturbation)

and the path followed to find the solution may not necessarily be the one suggest by gibbs dynamics

- study the Hessian sign

  condition on λ, α, β such that

  $\underline{\sigma}^{(1+2+3)}$ unstable & $\underline{\sigma}^{(1,2,3)}$ stable

  ensures the existence of a region, in the space of control parameters where the machine can in principle work - this is an upper bound for the disentanglement region.

  Subtle: I do not know whether the instability of the former finally leads to the target state.

- MCMC

Overall the three methods are consistent