The Statistical Mechanics of Associative Memories

Mathematics for Artificial Intelligence, Instituto Superior Técnico

Paulo Duarte Mourão

Department of Mathematics - Sapienza University of Rome

March 2026

The quintessential example - the Hopfield model

The Hopfield model brings together different major scientific areas. It can be described as

- a recurrent neural network;
- a spin system;
- an associative memory.

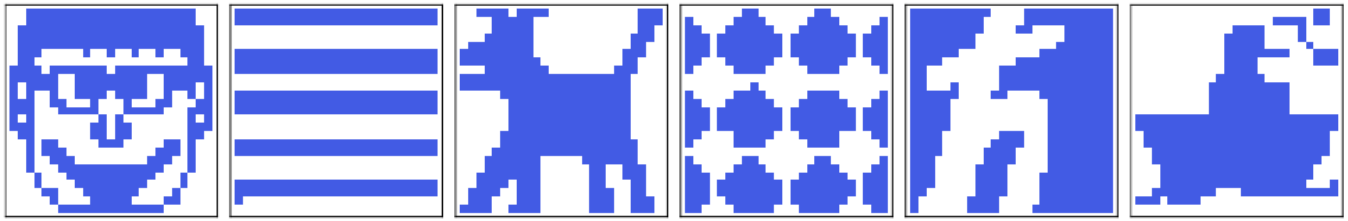


Figure 3.2: Set of $P = 6$ patterns stored in a Hopfield network of $N = 625$ spins. Patterns are black and while images: the network is dealing with digital storage of information Coolen et al. (2005).

Part I

Outline

Outline

Motivation

Why do we need Statistical Mechanics in the first place?

- Classical equations are generally not easily treatable for a high number of interacting bodies N (the system with $N = 3$ and gravitational interaction is already chaotic).

Motivation

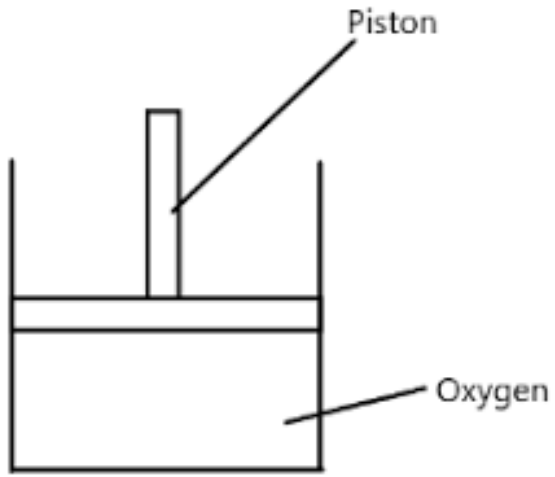
Why do we need Statistical Mechanics in the first place?

- Classical equations are generally not easily treatable for a high number of interacting bodies N (the system with $N = 3$ and gravitational interaction is already chaotic).
- Statistical Mechanics proposes leveraging statistics to derive macroscopic behaviors from microscopic properties of very large systems. This is done by studying instead probability distributions over a given set of possible microstates.

Thermodynamic systems in equilibrium

We will need:

- A set of possible microstates $\Omega = \{\omega_i\}_{i \in I}$.
- A function $S \colon \Omega \to \mathbb{R}$ to be maximized (i.e. the entropy).
- A (set of) constraint(s) $V \colon P(\Omega) \to \mathbb{R}$.



A natural choice of entropy

A natural choice of entropy is the so-called Shannon entropy

$$S = -k_B \sum_{i \in I} p_i \ln p_i \tag{1}$$

where $p_i := P(\omega_i)$, for each $i \in I$. We shall also take $k_B = 1$.

Two possibilities for constraints

Let us consider an energy function

$$\Omega \to \mathbb{R}$$
$$\omega_i \mapsto E_i$$

In the **microcanonical ensemble**, the energy E is kept fixed, and we have

$$p_i = \begin{cases} \frac{1}{N_i} & \text{if } E_i = E \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

with N_i denoting the number of states available at energy E_i.

Two possibilities for constraints

Let us consider an energy function

$$\Omega \to \mathbb{R}$$
$$\omega_i \mapsto E_i$$

In the **canonical ensemble**, instead the average energy $\overline{E} = \sum_i p_i E_i$ is kept fixed and we have

$$p_i = \frac{e^{-\beta E_i}}{Z_N(\beta)}, \quad Z_N(\beta) = \sum_{i \in I} e^{-\beta E_i}, \tag{3}$$

for some constant β. The above is known as the **Gibbs measure**, and Z as the **partition function**.

Lower values of the energy are more likely. How much depends on how low β is.

Gibbs measure

The previous maximization problem may be restated (via Lagrange multipliers) as minimizing

$$F_N = \overline{E} - \beta^{-1} S \tag{4}$$

with $\beta = 1/T$ acting as the multiplier.

Gibbs measure

The previous maximization problem may be restated (via Lagrange multipliers) as minimizing

$$F_N = \overline{E} - \beta^{-1} S \tag{4}$$

with $\beta = 1/T$ acting as the multiplier.

Indeed, adding the normalization constraint

$$\mathcal{L}[p] = \sum_i p_i E_i + \beta^{-1} \sum_i p_i \ln p_i - \lambda \sum_i p_i . \tag{5}$$

Gibbs measure

The previous maximization problem may be restated (via Lagrange multipliers) as minimizing

$$F_N = \overline{E} - \beta^{-1} S \tag{4}$$

with $\beta = 1/T$ acting as the multiplier.

Indeed, adding the normalization constraint

$$\mathcal{L}[p] = \sum_i p_i E_i + \beta^{-1} \sum_i p_i \ln p_i - \lambda \sum_i p_i. \tag{5}$$

Differentiating gives

$$\frac{\partial \mathcal{L}}{\partial p_i} = E_i + \beta^{-1}(\ln p_i + 1) - \lambda. \tag{6}$$

and so

$$\frac{\partial \mathcal{L}}{\partial p_i} = 0 \iff \ln p_i = -\beta E_i + \beta\lambda - 1 \iff p_i = e^{\beta\lambda - 1} e^{-\beta E_i} \tag{7}$$

Gibbs measure

Furthermore, we have

$$\begin{aligned} F_N(\beta) &= \sum_i p_i \left(E_i + \beta^{-1} \ln p_i\right) \\ &= \sum_i p_i \left(E_i - E_i - \beta^{-1} \ln Z_N(\beta)\right) \\ &= -\beta^{-1} \ln Z_N(\beta). \end{aligned}$$

Which should look familiar from probabilities if we look carefully.

The free energy is all you need

Both the partition function and the free energy encode the knowledge we have about the system. For instance

$$\beta \frac{\partial F_N(\beta)}{\partial \beta} = \frac{\sum_{i \in I} e^{-\beta E_i} E_i}{Z_N(\beta)} = \sum_i p_i E_i = \overline{E} \tag{8}$$

They are nothing other than the moment generating function and the cumulant generating function, respectively. Furthermore, the free energy density $f_N(\beta) := F_N(\beta)/N$ should have a limit as $N \to \infty$.

$$f(\beta) := \lim_{N \to \infty} f_N(\beta) \tag{9}$$

The strategy

If we can write the partition function, and consequently the free energy, constrained to an appropriate order parameter $\boldsymbol{m}$

$$Z_N(\beta) = \int d\boldsymbol{m} Z_N^c(\boldsymbol{m};\beta) = \int d\boldsymbol{m} e^{-Nf^c(\boldsymbol{m};\beta)+\mathcal{O}(1)} \tag{10}$$

Then saddle-point approximation gives

$$f(\beta) = -\lim_{N\to\infty} \frac{\ln Z_N(\beta)}{\beta N} = \min_{\boldsymbol{m}} f^c(\boldsymbol{m}). \tag{11}$$

Outline

A quick detour: spins

- Spin is an intrinsic form of angular momentum carried by elementary particles (although without actually "spinning" in the usual sense).
- Ampère's Law: an electric current creates a magnetic field around it.

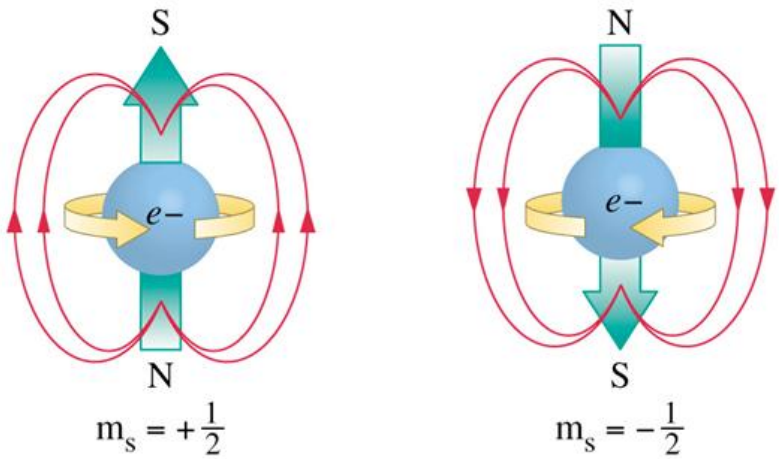


Figure: Magnetic fields resulting from electron spins.[1]

[1] Image from General Chemistry, 3rd edition, by Hill and Petrucci

The Curie-Weiss model

We now consider a state space of N spins that can take the values ± 1, meaning we have

$$\Omega = \{\pm 1\}^N \tag{12}$$

The Curie-Weiss model is a **mean-field approximation** model defined by the Hamiltonian

$$H_N^{\mathrm{CW}}(\boldsymbol{\sigma}) = -\frac{1}{2N} \sum_{i,j=1}^{N} \sigma_i \sigma_j - h \sum_{i=1}^{N} \sigma_i \tag{13}$$

for each state $\boldsymbol{\sigma} \in \Omega$, and $h \in \mathbb{R}$ the intensity of the external field.

The Curie-Weiss model

Starting from the partition function

$$Z_N^{\mathrm{CW}}(\beta) = \sum_{\boldsymbol{\sigma}\in\{-1,+1\}^N} e^{-\beta H_N^{\mathrm{Cw}}(\boldsymbol{\sigma})}, \tag{14}$$

we insert the order parameter

$$m_N(\boldsymbol{\sigma}) = \frac{1}{N}\sum_{i=1}^{N}\sigma_i \tag{15}$$

via the Fourier representation of the delta distribution

$$\delta\left(m - \frac{1}{N}\sum_{i=1}^{N}\sigma_i\right) = \frac{N}{2\pi}\int_{-\infty}^{+\infty} d\hat{m}\,\exp\left[i\hat{m}\left(m - \frac{1}{N}\sum_{i=1}^{N}\sigma_i\right)\right]. \tag{16}$$

The Curie-Weiss model

This yields

$$Z_N^{\mathrm{CW}}(\beta) = \sum_{\boldsymbol{\sigma}} \int dm \delta \left(m - \frac{1}{N} \sum_{i=1}^{N} \sigma_i \right) \exp \left[\frac{N}{2} \beta m^2 + N \beta h m \right]$$

The Curie-Weiss model

This yields

$$Z_N^{\mathrm{CW}}(\beta) = \sum_{\boldsymbol{\sigma}} \int dm \delta \left(m - \frac{1}{N} \sum_{i=1}^{N} \sigma_i \right) \exp \left[\frac{N}{2} \beta m^2 + N \beta h m \right]$$

$$= \frac{N}{2\pi} \sum_{\boldsymbol{\sigma}} \int dm d\hat{m} \exp \left[N \left(\frac{\beta}{2} m^2 + \beta h m + i \hat{m} m \right) - i \hat{m} \sum_{i=1}^{N} \sigma^i \right]$$

The Curie-Weiss model

This yields

$$
\begin{aligned}
Z_N^{\mathrm{CW}}(\beta) &= \sum_{\boldsymbol{\sigma}} \int dm \delta \left(m - \frac{1}{N} \sum_{i=1}^{N} \sigma_i \right) \exp \left[\frac{N}{2} \beta m^2 + N \beta h m \right] \\
&= \frac{N}{2\pi} \sum_{\boldsymbol{\sigma}} \int dm d\hat{m} \exp \left[N \left(\frac{\beta}{2} m^2 + \beta h m + i \hat{m} m \right) - i \hat{m} \sum_{i=1}^{N} \sigma^i \right] \\
&= \frac{N}{2\pi} \int dm d\hat{m} \exp \left[N \left(\frac{\beta}{2} m^2 + \beta h m + i \hat{m} m \right) \right] \sum_{\boldsymbol{\sigma}} \prod_{i=1}^{N} e^{i \hat{m} \sigma_i}
\end{aligned}
$$

The Curie-Weiss model

This yields

$$
\begin{aligned}
Z_N^{\mathrm{CW}}(\beta) &= \sum_{\boldsymbol{\sigma}} \int dm \delta\left(m - \frac{1}{N}\sum_{i=1}^{N}\sigma_i\right) \exp\left[\frac{N}{2}\beta m^2 + N\beta h m\right] \\
&= \frac{N}{2\pi}\sum_{\boldsymbol{\sigma}} \int dm d\hat{m} \exp\left[N\left(\frac{\beta}{2}m^2 + \beta h m + i\hat{m}m\right) - i\hat{m}\sum_{i=1}^{N}\sigma^i\right] \\
&= \frac{N}{2\pi}\int dm d\hat{m} \exp\left[N\left(\frac{\beta}{2}m^2 + \beta h m + i\hat{m}m\right)\right] \sum_{\boldsymbol{\sigma}}\prod_{i=1}^{N} e^{i\hat{m}\sigma_i} \\
&= \frac{N}{2\pi}\int dm d\hat{m} \exp\left[N\left(\frac{\beta}{2}m^2 + \beta h m + i\hat{m}m\right)\right] \prod_{i=1}^{N}\sum_{\sigma_i=\pm 1} e^{i\hat{m}\sigma_i}
\end{aligned}
$$

.

The Curie-Weiss model

This yields

$$
\begin{aligned}
Z_N^{\mathrm{CW}}(\beta) &= \sum_{\boldsymbol{\sigma}} \int dm \delta\left(m - \frac{1}{N}\sum_{i=1}^{N}\sigma_i\right) \exp\left[\frac{N}{2}\beta m^2 + N\beta h m\right] \\
&= \frac{N}{2\pi}\sum_{\boldsymbol{\sigma}} \int dm d\hat{m} \exp\left[N\left(\frac{\beta}{2}m^2 + \beta h m + i\hat{m}m\right) - i\hat{m}\sum_{i=1}^{N}\sigma^i\right] \\
&= \frac{N}{2\pi}\int dm d\hat{m} \exp\left[N\left(\frac{\beta}{2}m^2 + \beta h m + i\hat{m}m\right)\right] \sum_{\boldsymbol{\sigma}}\prod_{i=1}^{N} e^{i\hat{m}\sigma_i} \\
&= \frac{N}{2\pi}\int dm d\hat{m} \exp\left[N\left(\frac{\beta}{2}m^2 + \beta h m + i\hat{m}m\right)\right] \prod_{i=1}^{N}\sum_{\sigma_i=\pm 1} e^{i\hat{m}\sigma_i} \\
&= \frac{N}{2\pi}\int dm d\hat{m} \exp\left[N\left(\frac{\beta}{2}m^2 + \beta h m + i\hat{m}m + \ln 2\cos\hat{m}\right)\right].
\end{aligned}
$$

The Curie-Weiss model

Therefore, we get

$$f^{\mathrm{CW}}(\beta) = -\lim_{N\to\infty} \frac{1}{\beta N} \ln Z_N^{\mathrm{CW}}(\beta)$$
$$= \min_m \left[-\frac{m^2}{2} - hm - i\beta^{-1}\hat{m}m - \beta^{-1} \ln 2\cos\hat{m} \right]$$

The Curie-Weiss model

Therefore, we get

$$f^{\mathrm{CW}}(\beta) = -\lim_{N\to\infty} \frac{1}{\beta N} \ln Z_N^{\mathrm{CW}}(\beta)$$
$$= \min_m \left[-\frac{m^2}{2} - hm - i\beta^{-1}\hat{m}m - \beta^{-1} \ln 2\cos\hat{m} \right]$$

The saddle-point equations are then

$$\left.\frac{\partial f^{\mathrm{CW}}(\beta)}{\partial m}\right|_{\substack{m=m^* \\ \tilde{m}=\tilde{m}^*}} = 0 \iff m^* + h = -i\beta^{-1}\hat{m}^* \tag{17}$$

$$\left.\frac{\partial f^{\mathrm{CW}}(\beta)}{\partial \hat{m}}\right|_{\substack{m=m^* \\ \tilde{m}=\tilde{m}^*}} = 0 \iff m^* = -i\tan\hat{m}^* \tag{18}$$

The Curie-Weiss model

Therefore, we get

$$f^{\mathrm{CW}}(\beta) = -\lim_{N\to\infty} \frac{1}{\beta N} \ln Z_N^{\mathrm{CW}}(\beta)$$
$$= \min_m \left[-\frac{m^2}{2} - hm - i\beta^{-1}\hat{m}m - \beta^{-1} \ln 2\cos\hat{m} \right]$$

The saddle-point equations are then

$$\left.\frac{\partial f^{\mathrm{CW}}(\beta)}{\partial m}\right|_{\substack{m=m^* \\ \tilde{m}=\tilde{m}^*}} = 0 \iff m^* + h = -i\beta^{-1}\hat{m}^* \tag{17}$$

$$\left.\frac{\partial f^{\mathrm{CW}}(\beta)}{\partial \hat{m}}\right|_{\substack{m=m^* \\ \tilde{m}=\tilde{m}^*}} = 0 \iff m^* = -i\tan\hat{m}^* \tag{18}$$

Combining both gives

$$m^* = \tanh\left(\beta\left(m^* + h\right)\right) \tag{19}$$

The Curie-Weiss model

Given the simplicity of the saddle point equation in m, it can also be used to replace $\hat{m} = i\beta\,(m + h)$ into constrained the free energy, yielding the sometimes called pseudo (constrained) free energy

$$\tilde{f}^{\mathrm{CW,c}}(m;\beta) = \frac{m^2}{2} - \beta^{-1} \ln 2 \cosh \beta\,(m + h) \tag{20}$$

which has the same extrema as the true constrained free energy.

The Curie-Weiss model

The next picture[6] shows clearly possible solutions for the minima of the free energy when $h = 0$ ($\beta = 1$ is a phase transition!)

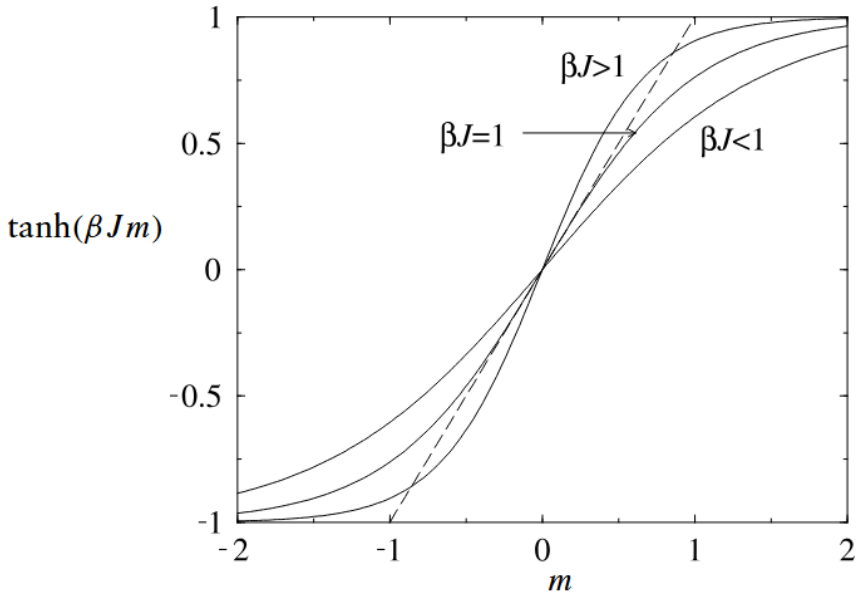


Figure 20.2 Graphical solution of the Curie–Weiss equation (20.43), viz. $m = \tanh(\beta J m)$. Solid lines show the function $\tanh(\beta J m)$ for different values of βJ. Dashed: the diagonal which this function must intersect. For $\beta J < 1$, $m = 0$ is the only solution; for larger values of βJ, there are two additional solutions m_β and $-m_\beta$.

Outline

The McCulloch-Pitts neuron

The MP neuron is a model of an artificial neuron that takes in binary inputs $\{S_i = 0, 1\}_{i=1}^{N}$ and outputs a binary response $y = 0, 1$ given by

$$y = \theta\left(\sum_{i=1}^{N} J_i S_i - U^*\right) \tag{21}$$

with θ being the Heaviside step-function.

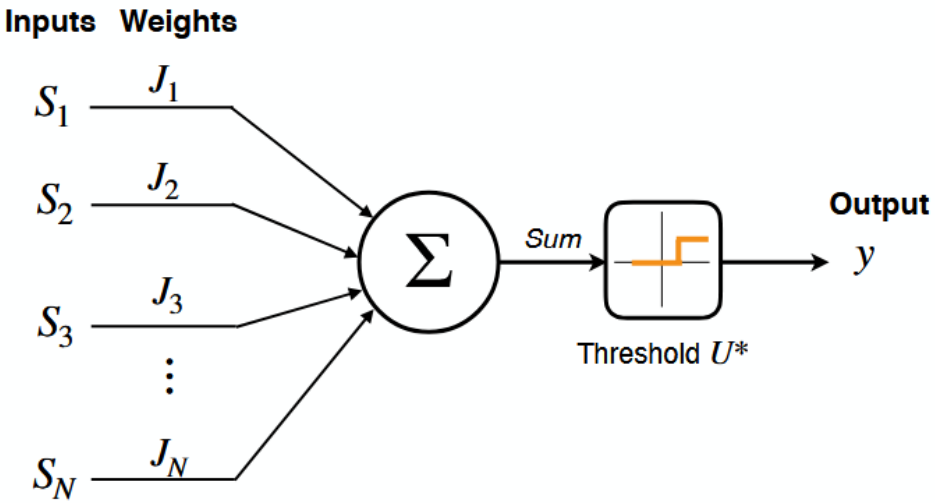


Figure 1.9: Schematic representation of a MP neuron.

Recurrent Neural Networks (RNNs)

Fully-connected recurrent neural networks are described by a family of N MP neurons $\{S_i = 0, 1\}_{i=1}^{N}$ and an update rule given by

$$S_i(t+1) = \theta\left(\sum_{j=1}^{N} J_{ij} S_j(t) - U_i^*(t)\right) \tag{22}$$

And to simulate the effect of noise in the system, we consider the U_i^* to be a random variable

$$U_i^*(t) = U_i - \frac{1}{2} T z_i(t), \tag{23}$$

with $\mathbb{E}[z_i(t)] = 0$, $\mathbb{E}\left[z_i(t)^2\right] = 1$ and $T > 0$.

From RNNs to Spin Systems

Let us apply the transformation

$$\sigma_i(t) = 2S_i(t) - 1, \tag{24}$$

then the update rule becomes (notice that $\text{sgn}(x) = 2\theta(x) - 1$)

$$\sigma_i(t+1) = \text{sgn}\left(\sum_{j=1}^{N} J_{ij}\sigma_j(t) + h_i + Tz_i(t)\right), \tag{25}$$

with $h_i = \sum_j J_{ij} - 2U_i$. But how far does the analogy *really* go?

Emergence of the Gibbs measure

Theorem 1

There exists a choice of noises z_i such that, for any symmetric interaction ($J_{ij} = J_{ji}$) without self-interactions ($J_{ii} = 0$), the dynamics (25) updated sequentially has the equilibrium distribution given by

$$p(\boldsymbol{\sigma}) = Z^{-1} e^{-\frac{H(\boldsymbol{\sigma})}{T}} \tag{26}$$

$$Z = \sum_{\boldsymbol{\sigma} \in \{\pm 1\}^N} e^{-\frac{H(\boldsymbol{\sigma})}{T}} \tag{27}$$

with

$$H(\boldsymbol{\sigma}) = -\frac{1}{2} \sum_{i,j=1}^{N} J_{ij} \sigma_i \sigma_j - \sum_{i=1}^{N} h_i \sigma_i. \tag{28}$$

Emergence of the Gibbs measure

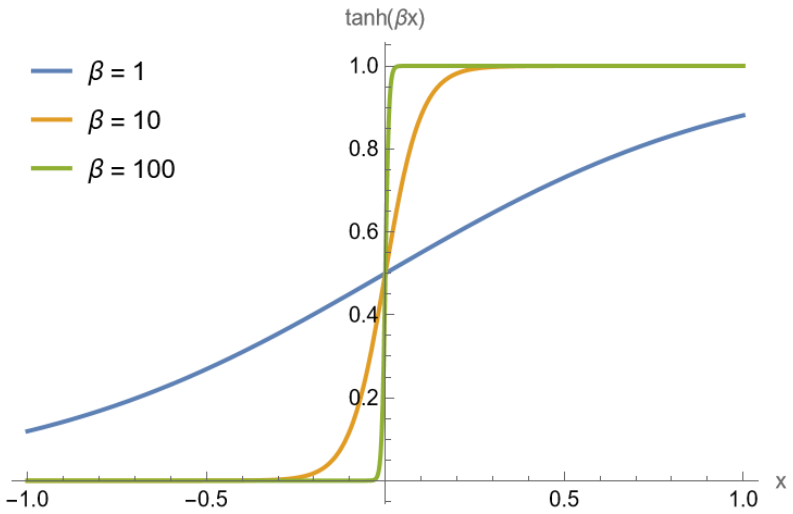
If the $z_i(t)$'s are defined according probability distribution function given by

$$f_{z_i(t)}(x) = 1 - \tanh^2(x). \tag{29}$$

It can be shown that this is equivalent to writing the dynamics (25) as

$$\mathbb{P}\left[\sigma_i(t+1) = \pm\sigma_i(t)\right] = \frac{1}{2}\left[1 \pm \tanh\left(\frac{E_i\left(\boldsymbol{\sigma}(t)\right)}{T}\right)\right] \tag{30}$$

with $E_i(\boldsymbol{\sigma}) = \sum_j \sigma_i(t) J_{ij} \sigma_j(t) + h_i \sigma_i(t)$.



Emergence of the Gibbs measure

The update rule (30) can also be written as

$$\mathbb{P}\left[\sigma_i(t+1) \neq \sigma_i(t)\right] = \frac{1}{1 + e^{\frac{\Delta E_i}{T}}} \tag{31}$$

which corresponds to the Glauber Monte-Carlo dynamics for the canonical ensemble defined by the Hamiltonian

$$H(\boldsymbol{\sigma}) = \sum_{ij} \sigma_i(t) J_{ij} \sigma_j(t) + \sum_i h_i \sigma_i(t), \tag{32}$$

where ΔE_i is the energy difference resulting from swapping spin i.

Emergence of the Gibbs measure

And so for this choice of noises or system converges to the Boltzmann-Gibbs distribution

$$p(\boldsymbol{J}, \boldsymbol{\sigma}) = \frac{e^{-\beta H_{N,\boldsymbol{h},H}(\boldsymbol{J},\boldsymbol{\sigma})}}{Z_{N,\boldsymbol{h},H}(\beta, \boldsymbol{J})} \tag{33}$$

in equilibrium with

$$Z_{N,\boldsymbol{h},H}(\beta, \boldsymbol{J}) = \sum_{\boldsymbol{\sigma} \in \{\pm 1\}^N} e^{-\beta H_{N,\boldsymbol{h},H}(\boldsymbol{J},\boldsymbol{\sigma})}. \tag{34}$$

Outline

Back to the Curie-Weiss model

The Curie-Weiss model is a **mean-field approximation** model defined, with no external field, by the Hamiltonian

$$H_N^{\mathrm{CW}}(\boldsymbol{\sigma}) = -\frac{1}{2N}\sum_{i,j=1}^{N}\sigma_i\sigma_j \tag{35}$$

Generalizing Curie-Weiss

One way to think of the Hamiltonian for the Curie-Weiss model (35) is that it "incentivizes" the spins to align themselves with the configuration $\boldsymbol{\eta} = (1, \dots 1)$. In fact, the overall magnetization m_N may be written as

$$m_N(\boldsymbol{\sigma}; \boldsymbol{\eta}) = \frac{1}{N} \sum_{i=1}^{N} \eta_i \sigma_i \tag{36}$$

Generalizing Curie-Weiss

What if we picked a different configuration, say $\xi \in \{-1,1\}^N$? Then we could define a new "generalized" Curie-Weiss Hamiltonian

$$H_N^{\mathrm{CW}}(\boldsymbol{\sigma},\boldsymbol{\xi}) = -\frac{N}{2} m_N^2(\boldsymbol{\sigma};\boldsymbol{\xi}) - hNm_N(\boldsymbol{\sigma};\boldsymbol{\xi}), \tag{37}$$

or, expanding,

$$H_N^{\mathrm{CW}}(\boldsymbol{\sigma};\boldsymbol{\xi}) = -\frac{1}{2N}\sum_{i,j=1}^{N} \xi_i\xi_j\sigma_i\sigma_j - h\sum_{i=1}^{N}\xi_i\sigma_i, \tag{38}$$

Generalizing Curie-Weiss

We can see that this "generalization" of the Curie-Weiss model is actually equivalent, if we do a very simple change of variables $\tilde{\sigma}_i = \xi_i \sigma_i$. Then

$$H_N^{\mathrm{CW}}(\boldsymbol{\sigma}; \boldsymbol{\xi}) = -\frac{1}{2N} \sum_{i,j=1}^{N} \xi_i \xi_j \sigma_i \sigma_j - h \sum_{i=1}^{N} \xi_i \sigma_i$$

$$= -\frac{1}{N} \sum_{i,j=1}^{N} \tilde{\sigma}_i \tilde{\sigma}_j - h \sum_{i=1}^{N} \tilde{\sigma}_i$$

Since the $\tilde{\boldsymbol{\sigma}}$'s also live in $\{-1, 1\}^N$, the systems are equivalent. The only difference is the spins now align themselves with $\boldsymbol{\xi}$.

From Curie-Weiss to Hopfield

Now what if we add more of these patterns? For instance, we can consider K patterns $\boldsymbol{\xi}^{\mu}$, with $1 \leqslant \mu \leqslant K$ and write

$$H_{N,K}(\boldsymbol{\sigma};\boldsymbol{\xi}) = -\frac{N}{2}\sum_{\mu=1}^{K} m_N^2(\boldsymbol{\sigma};\boldsymbol{\xi}^{\mu}) - hNm_N(\boldsymbol{\sigma};\boldsymbol{\xi}^1) \tag{39}$$

or, once again expanded,

$$H_{N,K}(\boldsymbol{\sigma};\boldsymbol{\xi}) = -\frac{1}{2N}\sum_{\mu=1}^{K}\sum_{i,j=1}^{N}\xi_i^{\mu}\xi_j^{\mu}\sigma_i\sigma_j - h\sum_{i=1}^{N}\xi_i^1\sigma_i \tag{40}$$

This is the Hamiltonian of the Hopfield model! [1, 2]

Part II

Outline

Outline

From Curie-Weiss to Hopfield

Now what if we add more of these patterns? For instance, we can consider K patterns $\boldsymbol{\xi}^{\mu}$, with $1 \leqslant \mu \leqslant K$ and write

$$H_{N,K}(\boldsymbol{\sigma};\boldsymbol{\xi}) = -\frac{N}{2}\sum_{\mu=1}^{K} m_N^2(\boldsymbol{\sigma};\boldsymbol{\xi}^{\mu}) - hNm_N\left(\boldsymbol{\sigma};\boldsymbol{\xi}^{1}\right) \tag{41}$$

or, once again expanded,

$$H_{N,K}(\boldsymbol{\sigma};\boldsymbol{\xi}) = -\frac{1}{2N}\sum_{\mu=1}^{K}\sum_{i,j=1}^{N}\xi_i^{\mu}\xi_j^{\mu}\sigma_i\sigma_j - h\sum_{i=1}^{N}\xi_i^{1}\sigma_i \tag{42}$$

This is the Hamiltonian of the Hopfield model! [1, 2]

Why is this harder?

- There are two main reasons why the Hopfield model is generally (at least for big enough K) more complicated than the Curie-Weiss model: disorder and frustration.
- These two properties are characteristic of a wide array of models, known as **disordered systems**.

Frustration

Looking at the interaction matrix for the Hopfield Model

$$J_{ij}^{\mathrm{H}} = \frac{1}{N} \sum_{\mu=1}^{K} \xi_i^\mu \xi_j^\mu \tag{43}$$

we see that there is a non-zero probability of having connections that do not agree with a certain pattern.

Frustration

- The matrix $\boldsymbol{J}$ tells us whether the system is trying to align ($J_{ij} > 0$) or oppose ($J_{ij} < 0$) the spins i and j.
- Crucially, in the Hopfield Model, contrary to the Curie-Weiss model, not all interactions can be satisfied.

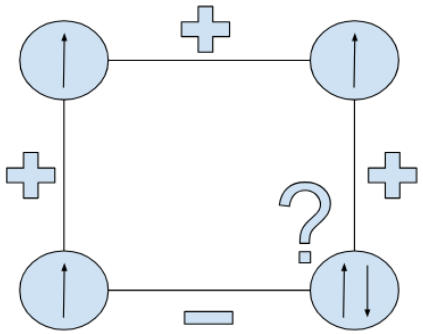


Figure 2.9: A very simple example of a frustrated system. The spins tend to be parallel when they interact with a positive coupling and anti-parallel when the interaction is negative. Obviously, not all the conditions can be met simultaneously, meaning that interaction is frustrated.

Disorder

The partition function of the Hopfield model is given by

$$Z_{N,K}(\beta;\boldsymbol{\xi}) = \sum_{\boldsymbol{\sigma}\in\{\pm 1\}^N} e^{-\beta H_{N,K}(\boldsymbol{\sigma};\boldsymbol{\xi})} \tag{44}$$

which depends on the choice of patterns, and so does the free energy.

Calculating these quantities would generally be very difficult....

Disorder

The solution is then to consider the $\boldsymbol{\xi}$'s as being drawn with respect to a certain probability distribution, known as the **disorder**, for instance

$$\mathbb{P}\left(\xi_i^\mu = \pm 1\right) = \frac{1}{2} \tag{45}$$

and compute the so called **quenched** free energy density with respect to this distribution

$$\overline{f}(\beta) = \mathbb{E}_{\boldsymbol{\xi}} f\left(\beta; \boldsymbol{\xi}\right) \tag{46}$$

Minimization of the quenched free energy

The quantity we are interested in is therefore

$$\overline{f}(\beta) = -\lim_{N\to\infty} \frac{1}{\beta N} \mathbb{E}_{\boldsymbol{\xi}} \ln Z_{N,K}(\beta, \boldsymbol{\xi}) \tag{47}$$

Our goal (now more complicated) is to find order-parameters $\boldsymbol{m}$ that self-average in the thermodynamic limit such that (if (47) exists)

$$\overline{f}(\beta) = \min_{\boldsymbol{m}} \lim_{N\to\infty} \overline{f}^{c}_{N,K}(\boldsymbol{m}) \tag{48}$$

still holds.

Guerra's interpolation

Given an interpolating Hamiltonian H_N^t, with $H_N^1 = H_N$ and H_N^0 an easier model, we compute the free energy via the Fundamental Theorem of Calculus

$$f^1 = f^0 + \int_0^1 dt \frac{d}{dt} f^t \tag{49}$$

The replica trick

The replica trick

- Step 1: Calculate $\langle Z_N^n \rangle$ for $n \in \mathbb{N}$.

The replica trick

- Step 1: Calculate $\langle Z_N^n \rangle$ for $n \in \mathbb{N}$.
- Step 2: Conveniently forget that n is an integer and compute

$$\langle \ln Z_N \rangle = \lim_{n \to 0} \frac{1}{n} \ln \langle Z_N^n \rangle . \tag{50}$$

The replica trick

- Step 1: Calculate $\langle Z_N^n \rangle$ for $n \in \mathbb{N}$.
- Step 2: Conveniently forget that n is an integer and compute

$$\langle \ln Z_N \rangle = \lim_{n \to 0} \frac{1}{n} \ln \langle Z_N^n \rangle . \tag{50}$$

- Step 3: Conveniently forget...

The replica trick

- Step 1: Calculate $\langle Z_N^n \rangle$ for $n \in \mathbb{N}$.
- Step 2: Conveniently forget that n is an integer and compute

$$\langle \ln Z_N \rangle = \lim_{n \to 0} \frac{1}{n} \ln \langle Z_N^n \rangle \,. \tag{50}$$

- Step 3: Conveniently forget...the order of the limits:

$$\lim_{N \to \infty} \lim_{n \to 0} \frac{1}{nN} \ln \langle Z_N^n \rangle = \lim_{n \to 0} \lim_{N \to \infty} \frac{1}{nN} \ln \langle Z_N^n \rangle \tag{51}$$

Outline

The Hopfield model

We recall the Hamiltonian for the Hopfield model

$$H_{N,K}(\boldsymbol{\sigma};\boldsymbol{\xi}) = -\frac{1}{2N}\sum_{\mu=1}^{K}\sum_{i,j=1}^{N}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j \tag{52}$$

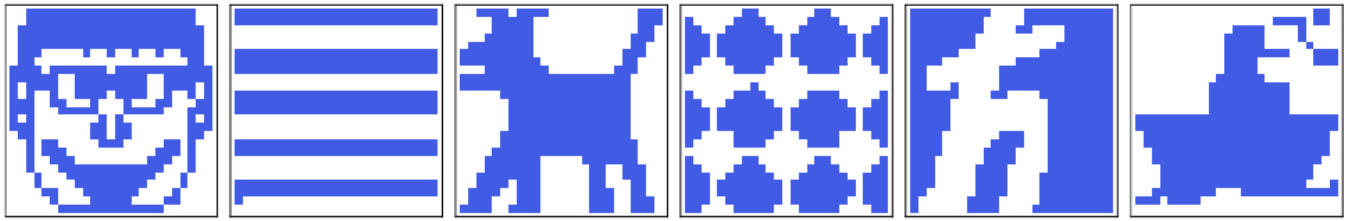


Figure 3.2: Set of $P = 6$ patterns stored in a Hopfield network of $N = 625$ spins. Patterns are black and while images: the network is dealing with digital storage of information Coolen et al. (2005).

The free energy

For finite K, the pseudo constrained free energy of the Hopfield model is

$$\tilde{\tilde{f}}^{c}(\beta) = \frac{1}{2}\sum_{\mu=1}^{K}\left(m_{\mu}^{*}\right)^{2} - \beta^{-1}\mathbb{E}_{\boldsymbol{\xi}} \ln 2\cosh\left(\beta\sum_{\mu=1}^{K} m_{\mu}^{*}\xi^{\mu}\right) \tag{53}$$

which reduces to that of the Curie-Weiss model when retrieving only a single pattern.

$$\tilde{f}^{\mathrm{CW,c}}(m;\beta) = \frac{m^{2}}{2} - \beta^{-1}\ln 2\cosh\left(\beta m\right) \tag{54}$$

Memory retrieval in the Hopfield model

If the model is within the retrieval region, and the initial state has enough overlap with one of the patterns, the model successfully retrieves the stored pattern.

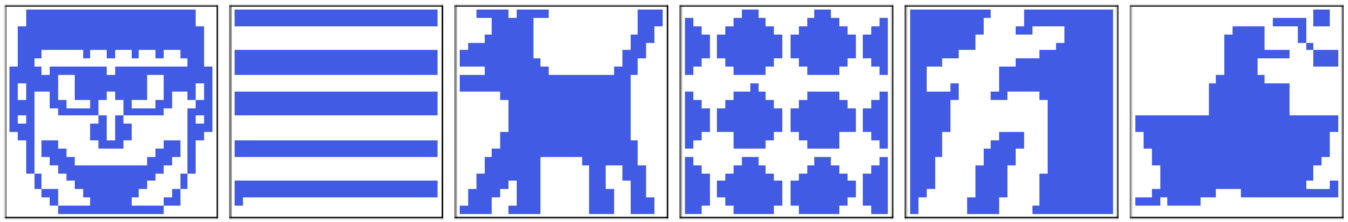


Figure 3.2: Set of $P = 6$ patterns stored in a Hopfield network of $N = 625$ spins. Patterns are black and while images: the network is dealing with digital storage of information Coolen et al. (2005).

Memory retrieval in the Hopfield model

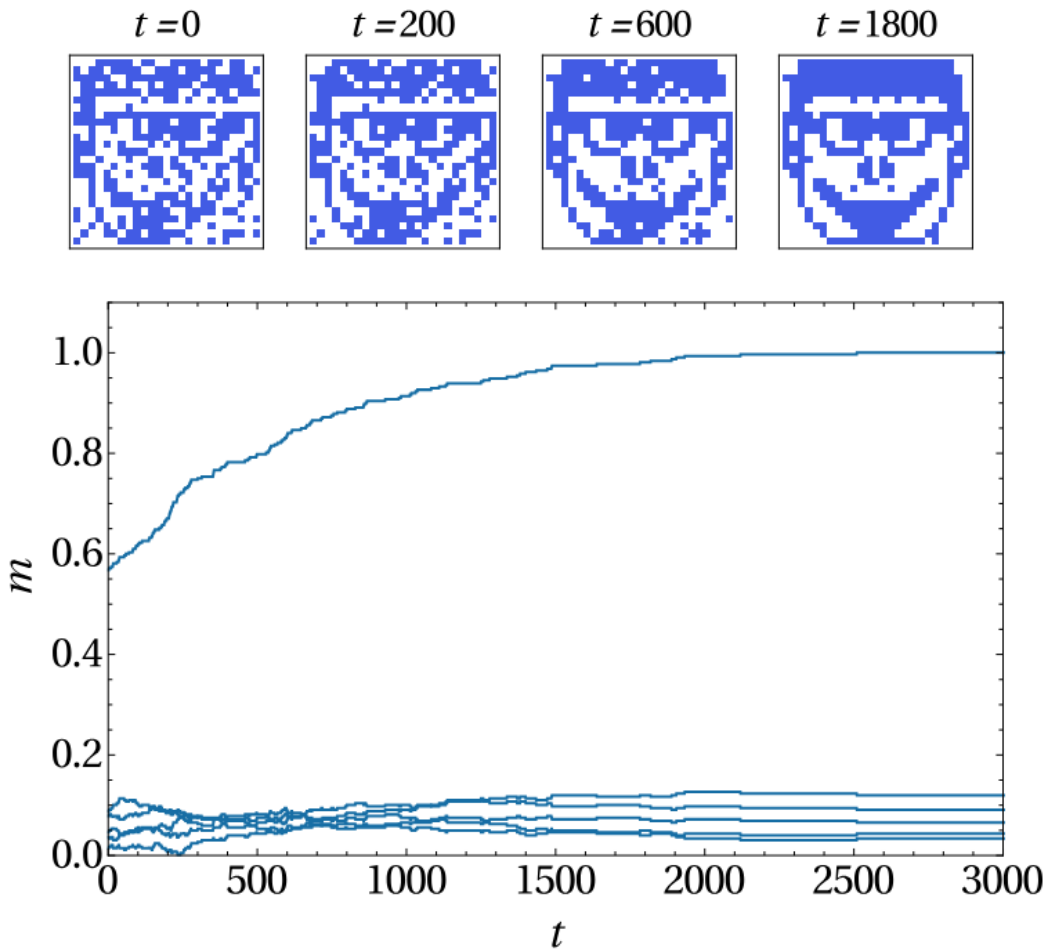


Figure 3.3: Example of a sucessfull pattern reconstruction. The Hopfield network is made of $N = 625$ neurons and $P = 6$ patterns are allocated in the synaptic matrix. Starting from a corrupted version of one of the patterns, the network is able to retrieve the associated pattern. We observe that, among the six Mattis magnetizations that quantify the retrieval of the six stored patterns, just one out of them grows up to one and its corresponding pattern is indeed retrieved by the network.

Spurious states in the Hopfield model

Frustration in the Hopfield model manifests itself in the form of so-called **spurious states**, i.e. equilibrium states of the model that are not original patterns, but instead mixtures, for instance

$$\tilde{\boldsymbol{\xi}} = \operatorname{sgn}\left(\boldsymbol{\xi}^1 + \boldsymbol{\xi}^2 + \boldsymbol{\xi}^3\right). \tag{55}$$

The amount of spurious states grows exponentially in K and so eventually takes over the dynamics of the system. The maximum capacity of the Hopfield model turns out to be

$$K \lesssim \alpha_c N, \tag{56}$$

with $\alpha_c \sim 0.138$, in the thermodynamic limit.

Spurious states in the Hopfield model

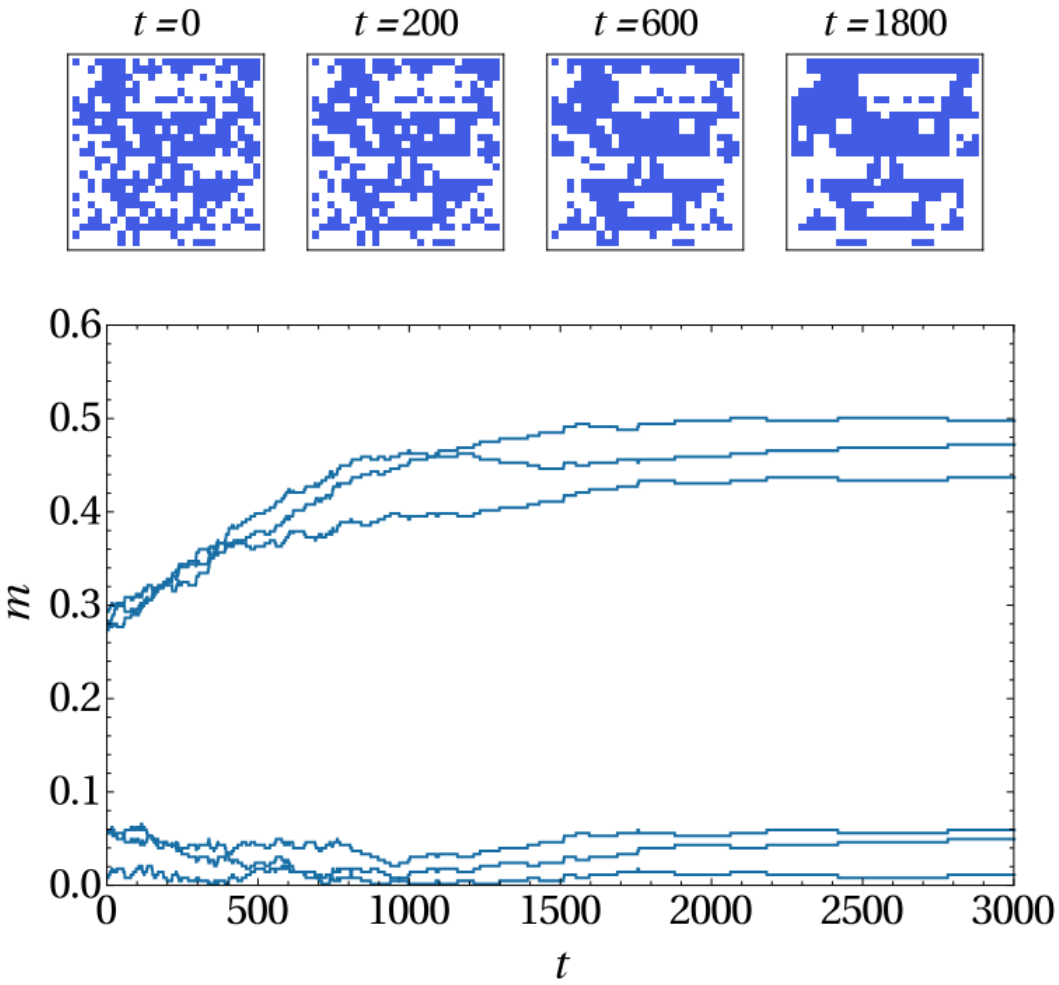


Figure 3.4: Example of dynamics reconstruction ending in a spurious state. The Hopfield network is the same as in Fig. 3.3. In this example, it is possible to observe that several (three) Mattis magnetizations raise sensibly over the noise due to the finite size effects and, correspondingly, the network has not been able to properly retrieve a single pattern, rather obtaining a useless mixture of the stored patterns.

The free energy

If we instead take

$$\lim_{N\to\infty} \frac{K}{N} = \alpha \tag{57}$$

The free energy

If we instead take

$$\lim_{N\to\infty} \frac{K}{N} = \alpha \tag{57}$$

then the replica-symmetric pseudo free energy for the recovery of ℓ patterns is given by

$$\begin{aligned}
\tilde{\tilde{f}}^{c}_{\mathsf{RS}}(\beta,\alpha) = &\frac{1}{2}\sum_{\mu=1}^{\ell}\left(m^{*}_{\mu}\right)^{2} + \frac{\alpha\beta}{2}p(1-q) \\
&+ \frac{\alpha}{2\beta}\left(\beta + \ln\left(1-\beta(1-q)\right) - \frac{q\beta}{1-\beta(1-q)}\right) \\
&- \beta^{-1}\int \frac{dz\, e^{-\frac{z^{2}}{2}}}{\sqrt{2\pi}}\mathbb{E}_{\boldsymbol{\xi}} \ln 2\cosh\left(\beta\sum_{\mu=1}^{\ell} m^{*}_{\mu}\xi^{\mu} + \beta z\sqrt{\alpha p}\right)
\end{aligned} \tag{58}$$

Order parameters

To obtain the high load free energy, we need, apart from magnetizations, the replica overlaps

$$m_{\mu,N}(\boldsymbol{\xi},\boldsymbol{\sigma}) = \frac{1}{N}\sum_{i=1}^{N}\xi_i^{\mu}\sigma_i, \quad \mu = 1,\dots,K \tag{59}$$

$$q_N(\boldsymbol{\sigma}^{(1)},\boldsymbol{\sigma}^{(2)}) = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^{(1)}\sigma_i^{(2)} \tag{60}$$

Phase diagram

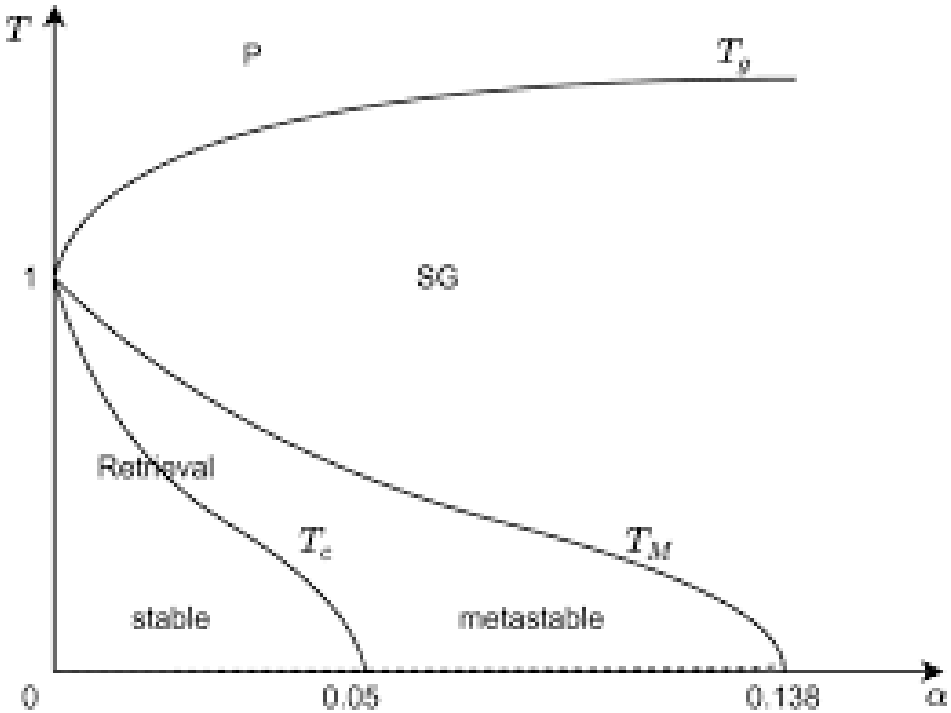


Figure: Phase diagram of the Hopfield model, first obtained by D. Amit, H. Gutfrend and H. Sompolinsky [3].

Outline

Dataset noise

The first generalization we consider is dataset noise. Here, the patterns are replaced by a noisy dataset (the elements of which we call examples)

$$\{\boldsymbol{\xi}^{\mu}\}_{\mu=1,\dots,K} \longrightarrow \{\boldsymbol{\eta}_a^{\mu}\}_{a=1,\dots,M}^{\mu=1,\dots,K}, \tag{61}$$

with M denoting the size of the dataset and, for each $\mu = 1, \dots, K$, $a = 1, \dots, M$ and $i = 1, \dots, N$, we have

$$\eta_{ia}^{\mu} = \xi_i^{\mu} \chi_{ia}^{\mu} \tag{62}$$

$$P\left(\chi_{ia}^{\mu} = \pm 1\right) = \frac{1 \pm r}{2}, \tag{63}$$

with $r = (0, 1]$ denoting the quality of the dataset.

Two kinds of learning

In supervised learning, we assume to have access to the pattern labels corresponding to each example and so we can store the averages

$$J_{ij}^{\text{sup}} \propto \sum_{\mu=1}^{K} \sum_{a,b=1}^{M} \eta_{ia}^{\mu} \eta_{jb}^{\mu}, \quad i,j = 1, \dots, N. \tag{64}$$

On the other hand, in unsupervised learning, the best we can do is

$$J_{ij}^{\text{unsup}} \propto \sum_{\mu=1}^{K} \sum_{a=1}^{M} \eta_{ia}^{\mu} \eta_{ja}^{\mu}, \quad i,j = 1, \dots, N \tag{65}$$

In the case of supervised learning, it can be shown that the influence of the noise depends only on r and M through what we call the entropy:

$$\rho := \frac{1 - r^2}{Mr^2}. \tag{66}$$

The L-directional associative memory

An L-directional associative memory is composed of L independent Hopfield networks (known as modules), with the corresponding state denoted by $(\sigma_i^k)_{i=1,\dots,N}^{k=1,\dots,L}$, and the Hamiltonian is

$$J_{ij}^{\ell k} = \frac{1}{\mathcal{R}NM^2} \sum_{\mu=1}^{K} \sum_{a,b=1}^{M} g_{\ell k} \eta_{ia}^{\mu} \eta_{jb}^{\mu}, \quad i,j = 1,\dots,N, \tag{67}$$

where $\mathcal{R} = r^2 + \frac{1-r^2}{M}$ is the second raw moment of the noise.

Tackling disentanglement

The goal with disentanglement is, in the most simple case, to input a mixture of L patterns in every module and get back one of its constituents per module

$$\left(\boldsymbol{\sigma}^{\text{mix}}, \ldots, \boldsymbol{\sigma}^{\text{mix}}\right) \longrightarrow \left(\boldsymbol{\xi}^{1}, \ldots, \boldsymbol{\xi}^{L}\right), \tag{68}$$

for a mixture $\boldsymbol{\sigma}^{\text{mix}} = \text{sgn}\left(\boldsymbol{\xi}^{1} + \ldots + \boldsymbol{\xi}^{L}\right)$. In order to achieve this, we take $g_{\ell k} = \delta_{\ell k} - \lambda\left(1 - \delta_{\ell k}\right)$, with $0 \leq \lambda < 0.5$.

A detailed study of disentanglement in this system for $L = 3$ (known as three-directional associative memory or TAM) was published in Physica A [4].

Order parameters

The order parameters must now also be generalized from the Hopfield model. For each $\mu = 1, \dots, K$ and $\ell, k = 1, \dots, L$, we have

$$m_\mu^\ell(\boldsymbol{\eta}, \boldsymbol{\sigma}) = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \sigma_i^\ell \tag{69}$$

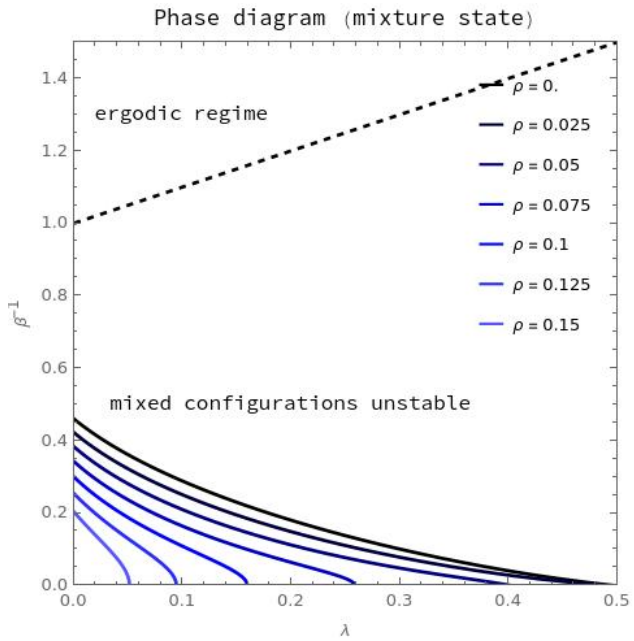
$$q_{12}^{\ell k}(\boldsymbol{\sigma}^{(1)}, \boldsymbol{\sigma}^{(2)}) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^{(1),\ell} \sigma_i^{(2),k} \tag{70}$$

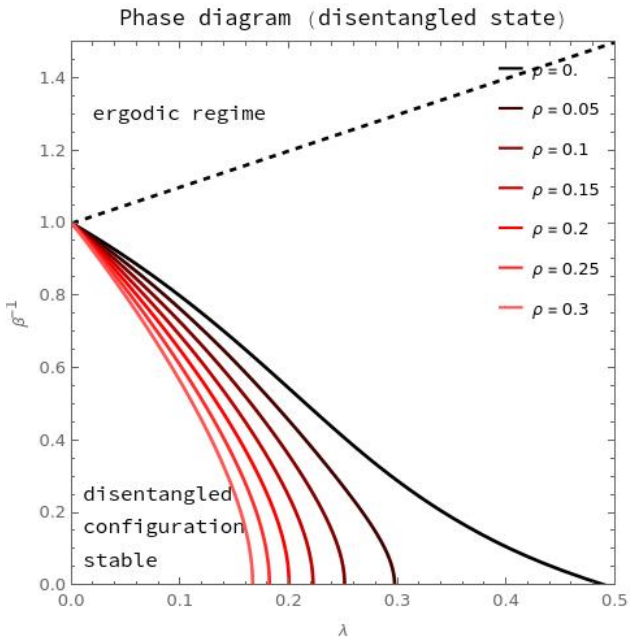
$$q_{11}^{\ell k}(\boldsymbol{\sigma}) = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^\ell \sigma_i^k \tag{71}$$

And the entropy now becomes the entropy per layer.

Self-consistency equations

The Statistical Mechanics framework presented before allowed us to get self-consistency equations for the order parameters of this system, which we solved in the low-load (i.e. finite number of patterns K in the thermodynamic limit $N \to \infty$).

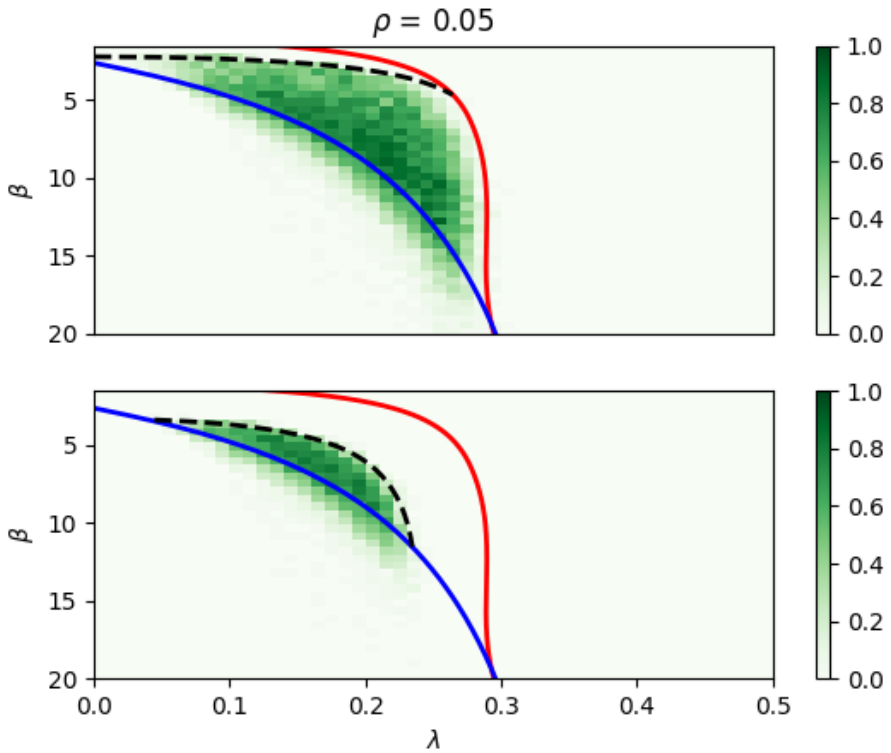




Prediction of disentanglement

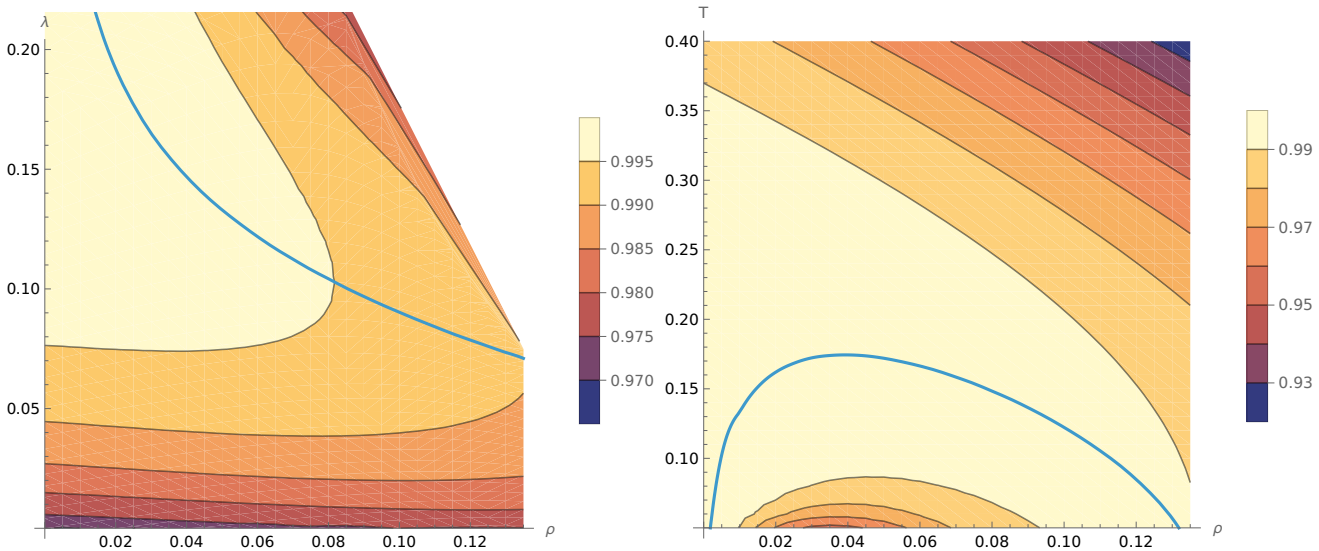
We observe, in the phase diagrams, that there's a region of state space where mixtures are unstable but disentangled states are stable.

Our hypothesis, shown to be valid through Monte-Carlo experiments, is that disentanglement is possible in this region.

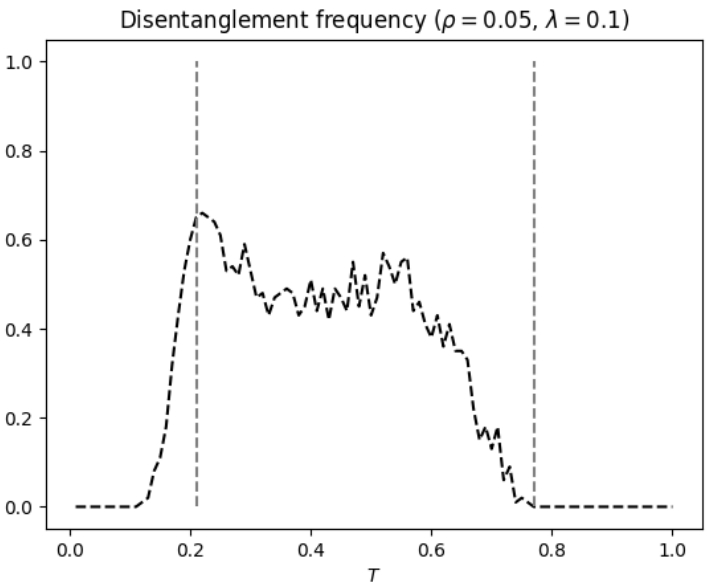


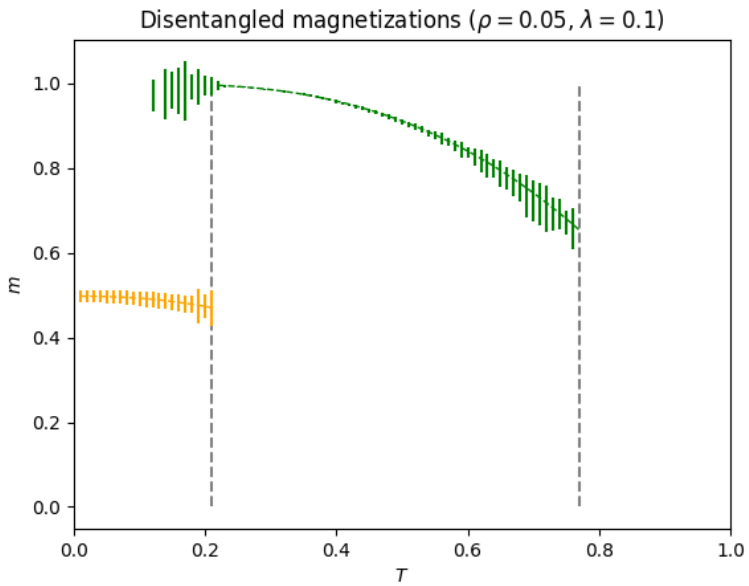
Optimal parameters

By maximizing retrieval magnetizations (i.e. the magnetizations of the output) within the disentanglement region, we also obtained optimal control parameters for each value of the dataset entropy per module ρ.

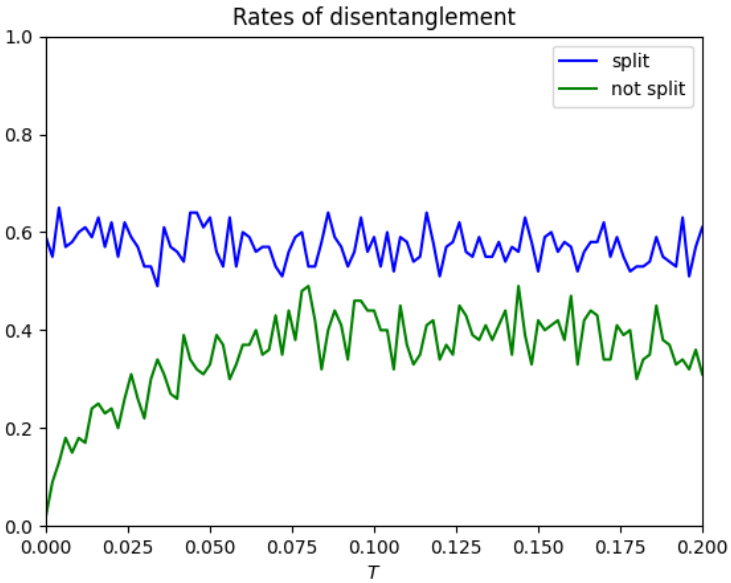


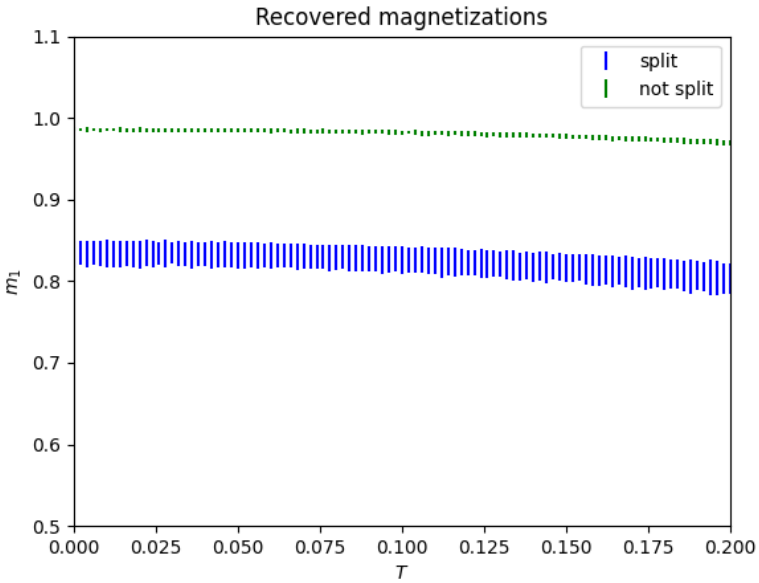
More experimental evidence





Splitting


Rates of disentanglement
split
not split
1.0
0.8
0.6
0.4
0.2
0.0
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
T


Recovered magnetizations
split
not split
1.1
1.0
0.9
0.8
0.7
0.6
0.5
m_1
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
T

Outline

Prof. Elena Agliari's talk (April 24-th)

Using the dreaming model

$$J_{ij}^{D} = \frac{1}{N} \sum_{\mu,\nu=1}^{K} \xi_i^{\mu} \left(\frac{1+t}{\mathbb{1} + t\boldsymbol{C}} \right)_{\mu\nu} \xi_j^{\nu}, \tag{72}$$

with

$$C_{ab}^{\mu\nu} = \frac{1}{N} \sum_{i=1}^{N} \xi_i^{\mu} \xi_i^{\nu}. \tag{73}$$

to boost learning from unsupervised datasets.

Outline

Bibliography

[1] W. Little. "The existence of persistent states in the brain". In: *Mathematical Biosciences* 19.1 (1974), pp. 101–120.

[2] J. J. Hopfield. "Neural networks and physical systems with emergent collective computational abilities". In: *Proceedings of the National Academy of Sciences* 79.8 (1982), pp. 2554–2558.

[3] D. J. Amit, H. Gutfreund, and H. Sompolinsky. "Storing Infinite Numbers of Patterns in a Spin-Glass Model of Neural Networks". In: *Phys. Rev. Lett.* 55 (14 Sept. 1985), pp. 1530–1533.

[4] E. Agliari, A. Fachechi, and P. D. Mourão. "The beneficial role of noises for disentanglement tasks in modular Hebbian networks". In: *Physica A: Statistical Mechanics and its Applications* (2025), p. 131134.

[6] C. A.C.C., R. Kuehn, and P. Sollich. *Theory of Neural Information Processing Systems*. Oxford University Press, 2005.

Other resources

- M. Mezard, G. Parisi, and M. A. Virasoro. *Spin Glass Theory And Beyond: An Introduction To The Replica Method And Its Applications*. World Scientific Publishing Company, 1987.
- M. Talagrand. *Mean Field Models for Spin Glasses*. Springer, 2010.
- A. Bovier. *Statistical Mechanics of Disordered Systems: A Mathematical Perspective*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2006.

My e-mail: paulo.duartemourao@uniroma1.it

Thank you to my advisors and collaborators!




Replicated partition function

We recall that our goal is to compute

$$\overline{f}(\beta,\alpha) = \lim_{\substack{N\to\infty \\ K/N\to\alpha}} -\frac{1}{\beta N}\mathbb{E}_{\boldsymbol{\xi}} \ln Z_{N,K}(\beta;\boldsymbol{\xi}) \tag{74}$$

with

$$Z_{N,K}(\beta;\boldsymbol{\xi}) = \sum_{\boldsymbol{\sigma}\in\Omega} e^{-\beta H_{N,K}(\boldsymbol{\sigma};\boldsymbol{\xi})} = \sum_{\boldsymbol{\sigma}\in\Omega} e^{\frac{\beta}{2N}\sum_{\mu=1}^{K}\sum_{i,j=1}^{N}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j} \tag{75}$$

Then

$$Z_{N,K}(\beta;\boldsymbol{\xi})^n = \sum_{\boldsymbol{\sigma}^1,\dots,\boldsymbol{\sigma}^n\in\Omega} e^{\frac{\beta}{2N}\sum_{r=1}^{n}\sum_{\mu=1}^{K}\sum_{i,j=1}^{N}\xi_i^\mu\xi_j^\mu\sigma_i^{(r)}\sigma_j^{(r)}} \tag{76}$$

Signal vs noise

Taking ℓ to be the number of patterns recovered, we have

$$Z_{N,K}(\beta;\boldsymbol{\xi})^n = \sum_{\boldsymbol{\sigma}^1,\dots,\boldsymbol{\sigma}^n\in\Omega} z_{\text{signal}}\left(\{\boldsymbol{\sigma}\}\right) \times z_{\text{noise}}\left(\{\boldsymbol{\sigma}\}\right) \tag{77}$$

with

$$z_{\text{signal}}\left(\{\boldsymbol{\sigma}\}\right) = e^{\frac{\beta}{2N}\sum_{a=1}^{n}\sum_{\mu=1}^{\ell}\sum_{i,j=1}^{N}\xi_i^\mu\xi_j^\mu\sigma_i^a\sigma_j^a} \tag{78}$$

$$z_{\text{noise}}\left(\{\boldsymbol{\sigma}\}\right) = e^{\frac{\beta}{2N}\sum_{a=1}^{n}\sum_{\mu=\ell+1}^{K}\sum_{i,j=1}^{N}\xi_i^\mu\xi_j^\mu\sigma_i^a\sigma_j^a}. \tag{79}$$

Furthermore, the disorder is independent between both, and the goal is to apply the replica trick to the disorder average of the non-recovered patterns, which we denote $\mathbb{E}'_{\boldsymbol{\xi}}$

$$\mathbb{E}'_{\boldsymbol{\xi}} Z_{N,K}(\beta;\boldsymbol{\xi})^n = \sum_{\boldsymbol{\sigma}^1,\dots,\boldsymbol{\sigma}^n\in\Omega} z_{\text{signal}}\left(\{\boldsymbol{\sigma}\}\right) \times \mathbb{E}_{\boldsymbol{\xi}} z_{\text{noise}}\left(\{\boldsymbol{\sigma}\}\right) \tag{80}$$

The signal part

For the signal part, we insert the order parameters corresponding to the magnetizations, as we did for the Curie-Weiss model

$$
\begin{aligned}
Z_{\text{signal}}(\{\boldsymbol{\sigma}\}) &= \exp\left[\frac{\beta}{2N}\sum_{a=1}^{n}\sum_{\mu=1}^{\ell}\sum_{i,j=1}^{N}\xi_i^\mu\xi_j^\mu\sigma_i^a\sigma_j^a\right] \\
&= \left(\frac{N}{2\pi}\right)^{n\ell}\int\left[\prod_{a,\mu=1}^{n,\ell}dm_\mu^a\,\delta\left(m_\mu^a-\frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i^a\right)\right] \\
&\quad\times\exp\left[\frac{\beta N}{2}\sum_{a=1}^{n}\sum_{\mu=1}^{\ell}\left(m_\mu^a\right)^2\right] \\
&= \left(\frac{N}{2\pi}\right)^{n\ell}\int\left[\prod_{a,\mu=1}^{n,\ell}dm_\mu^a\,d\hat{m}_\mu^a\exp\left[iNm_\mu^a\hat{m}_\mu^a-i\hat{m}_\mu^a\sum_{i=1}^{N}\xi_i^\mu\sigma_i^a\right]\right] \\
&\quad\times\exp\left[\frac{\beta N}{2}\sum_{a=1}^{n}\sum_{\mu=1}^{\ell}\left(m_\mu^a\right)^2\right]
\end{aligned}
$$

The Hubbard-Stratonovich Transform

To treat the noise part, we make use of the Hubbard-Stratonovich transform

$$e^{\frac{x^2}{2}} = \sqrt{\frac{1}{2\pi}} \int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2}+zx} \tag{81}$$

to get

$$z_{\text{noise}}\left(\{\boldsymbol{\sigma}\}\right) = \exp\left[\frac{\beta}{2N}\sum_{a=1}^{n}\sum_{\mu=\ell+1}^{K}\sum_{i,j=1}^{N}\xi_i^\mu\xi_j^\mu\sigma_i^a\sigma_j^a\right]$$

$$= \int \prod_{\substack{a=1,\dots,n \\ \mu=\ell+1,\dots,K}} \frac{dz_\mu^a}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(z_\mu^a\right)^2 + \sqrt{\frac{\beta}{N}}\sum_{i=1}^{N}\xi_i^\mu\sigma_i^a z_\mu^a\right]$$

Average over the disorder

Averaging over the disorder now gives, with $\mathcal{D}\boldsymbol{z} = \prod_{a,\mu} dz_\mu^a \, e^{-\left(z_\mu^a\right)^2/2}/\sqrt{2\pi}$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}} z_{\text{noise}}\left(\{\boldsymbol{\sigma}\}\right) &= \int \mathcal{D}\boldsymbol{z} \prod_{i,\mu} \mathbb{E}_{\boldsymbol{\xi}} \exp\left[\sqrt{\frac{\beta}{N}} \sum_{a=1}^{n} \xi_i^\mu \sigma_i^a z_\mu^a\right] \\
&= \int \mathcal{D}\boldsymbol{z} \prod_{i,\mu} \exp\left[\ln\cosh\left(\sqrt{\frac{\beta}{N}} \sum_{a=1}^{n} \sigma_i^a z_\mu^a\right)\right] \\
&= \left(\int \mathcal{D}\tilde{\boldsymbol{z}} \exp\left[\frac{\beta}{2N} \sum_{a,b=1}^{n} \sum_{i=1}^{N} z^a z^b \sigma_i^a \sigma_i^b + \mathcal{O}\left(\frac{1}{N}\right)\right]\right)^{K-\ell}
\end{aligned}$$

and $\mathcal{D}\tilde{\boldsymbol{z}} = \prod_a dz^a \, e^{-(z^a)^2/2}/\sqrt{2\pi}$.

Inserting the overlaps

We now insert the overlaps, again via delta distributions. Denoting $\tilde{K} := K - \ell$, we have

$$\mathbb{E}_{\boldsymbol{\xi}} z_{\text{noise}}(\{\boldsymbol{\sigma}\}) \approx \left(\frac{N}{2\pi}\right)^{n^2-n} \prod_{\substack{a,b=1 \\ a\neq b}}^{n} \int dQ_{ab}\, \delta\left(Q_{ab} - \frac{1}{N}\sum_{i=1}^{N} \sigma_i^a \sigma_i^b\right)$$

$$\times \int \mathcal{D}\tilde{\boldsymbol{z}} \exp\left[\frac{\beta}{2}\sum_{a,b=1}^{n} z^a z^b Q_{ab}\right]$$

$$= \left(\frac{N}{2\pi}\right)^{n^2-n} (2\pi)^{\frac{n^2}{2}} \int d\boldsymbol{Q} \exp\left[-\frac{\tilde{K}}{2} \ln\det\left(\mathbb{1} - \beta \boldsymbol{Q}\right)\right]$$

$$\times \int d\hat{\boldsymbol{Q}} \exp\left[i \sum_{\substack{a,b=1 \\ a\neq b}}^{n} \left(N Q_{ab}\hat{Q}_{ab} - \sum_{i=1}^{N} \hat{Q}_{ab}\sigma_i^a \sigma_i^b\right)\right]$$

Replicated partition function

Putting everything together and reparameterizing $\hat{Q}_{ab} \to i\frac{\alpha\beta^2}{2}P_{ab}$, we get

$$\mathbb{E}'_{\boldsymbol{\xi}} Z_{N,K}(\beta;\boldsymbol{\xi})^n \propto \int d\boldsymbol{m}d\hat{\boldsymbol{m}}d\boldsymbol{Q}d\boldsymbol{P} \exp\left[N\sum_{a,\mu=1}^{n,\ell}\left(\frac{\beta}{2}\left(m_\mu^a\right)^2 + i m_\mu^a \hat{m}_\mu^a\right)\right]$$

$$\times \exp\left[-\frac{\tilde{K}}{2}\ln\det\left(\mathbb{1}-\beta\boldsymbol{Q}\right) - \frac{\alpha\beta^2}{2}N\sum_{\substack{a,b=1\\a\neq b}}^{n} Q_{ab}P_{ab}\right]$$

$$\times \sum_{\boldsymbol{\sigma^1},\ldots,\boldsymbol{\sigma}^n\in\Omega}\prod_{i=1}^{N}\exp\left[-\sum_{a=1}^{n}\sum_{\mu=1}^{\ell} i\hat{m}_\mu^a\xi_i^\mu\sigma_i^a + \frac{\alpha\beta^2}{2}\sum_{\substack{a,b=1\\a\neq b}}^{n} P_{ab}\sigma_i^a\sigma_i^b\right]$$

Replicated partition function

Furthermore, assuming the magnetizations self-average, the last product in the i-index factorizes, meaning that we get

$$\mathbb{E}'_{\boldsymbol{\xi}} Z_{N,K}(\beta;\boldsymbol{\xi})^n \propto \int d\boldsymbol{m} d\hat{\boldsymbol{m}} d\boldsymbol{Q} d\boldsymbol{P} \exp\left[N \sum_{a,\mu=1}^{n,\ell} \left(\frac{\beta}{2}\left(m_\mu^a\right)^2 + i m_\mu^a \hat{m}_\mu^a\right)\right]$$
$$\times \exp\left[N\left(-\frac{\tilde{K}}{2N} \ln\det\left(\mathbb{1} - \beta \boldsymbol{Q}\right) - \frac{\alpha\beta^2}{2} \sum_{\substack{a,b=1 \\ a\neq b}}^{n} Q_{ab} P_{ab}\right)\right]$$
$$\times \left(\sum_{\sigma^{\mathbf{1}},\dots,\sigma^n=\pm 1} \exp\left[-\sum_{a=1}^{n}\sum_{\mu=1}^{\ell} i \hat{m}_\mu^a \xi^\mu \sigma^a + \frac{\alpha\beta^2}{2} \sum_{\substack{a,b=1 \\ a\neq b}}^{n} P_{ab}\sigma^a\sigma^b\right]\right)^N$$

Taking the thermodynamic limit

Furthermore, assuming the magnetizations self-average, the last product in the i-index factorizes, meaning that we get

$$\lim_{N\to\infty} -\frac{1}{\beta N}\mathbb{E}'_{\boldsymbol{\xi}} Z_{N,K}(\beta;\boldsymbol{\xi})^n = \min_{\boldsymbol{m},\hat{\boldsymbol{m}},\boldsymbol{Q},\boldsymbol{P}} f(\boldsymbol{m},\hat{\boldsymbol{m}},\boldsymbol{Q},\boldsymbol{P}) \tag{82}$$

with

$$f = -\frac{1}{2}\sum_{a,\mu=1}^{n,\ell}\left(m_\mu^a\right)^2 - i\beta^{-1}\sum_{a,\mu=1}^{n,\ell} m_\mu^a \hat{m}_\mu^a + \frac{\alpha\beta}{2}\sum_{\substack{a,b=1\\a\neq b}}^{n} Q_{ab}P_{ab} + \frac{\alpha}{2\beta}\ln\det\left(\mathbb{1}-\beta\boldsymbol{Q}\right)$$

$$-\beta^{-1}\ln\sum_{\boldsymbol{\sigma}}\exp\left[-\sum_{a=1}^{n}\sum_{\mu=1}^{\ell} i\hat{m}_\mu^a\xi^\mu\sigma^a + \frac{\alpha\beta^2}{2}\sum_{\substack{a,b=1\\a\neq b}}^{n} P_{ab}\sigma^a\sigma^b\right]$$

Taking the thermodynamic limit

By minimizing already with respect to $\boldsymbol{m}$, we get $\hat{m}^a_\mu = i\beta m^a_\mu$, which gives us

$$\lim_{N\to\infty} -\frac{1}{\beta N}\mathbb{E}'_{\boldsymbol{\xi}} Z_{N,K}(\beta;\boldsymbol{\xi})^n = \min_{\boldsymbol{m},\boldsymbol{Q},\boldsymbol{P}} \tilde{f}(\boldsymbol{m},\boldsymbol{Q},\boldsymbol{P}) \tag{83}$$

with

$$\tilde{f} = \frac{1}{2}\sum_{a,\mu=1}^{n,\ell}\left(m^a_\mu\right)^2 + \frac{\alpha\beta}{2}\sum_{\substack{a,b=1\\a\neq b}}^{n} Q_{ab}P_{ab} + \frac{\alpha}{2\beta}\ln\det\left(\mathbb{1}-\beta\boldsymbol{Q}\right)$$

$$-\beta^{-1}\ln\sum_{\boldsymbol{\sigma}}\exp\left[\beta\sum_{a=1}^{n}\sum_{\mu=1}^{\ell} m^a_\mu\xi^\mu\sigma^a + \frac{\alpha\beta^2}{2}\sum_{\substack{a,b=1\\a\neq b}}^{n} P_{ab}\sigma^a\sigma^b\right]$$

Replica symmetry ansatz

To be able to proceed with the computations and take the limit $n \to 0$, we must make assumptions on the structure of the matrices $\boldsymbol{m}$, $\hat{\boldsymbol{m}}$, $\boldsymbol{P}$ and $\boldsymbol{Q}$. The so-called replica-symmetry ansatz assumes complete independence on the choice of replicas, meaning that

$$m_\mu^a = m_\mu \tag{84}$$

$$Q_{ab} = \delta_{ab} + (1 - q)\delta_{ab} \tag{85}$$

$$P_{ab} = p \tag{86}$$

Replica symmetry ansatz

From this ansatz, we get, for the first and second terms

$$\lim_{n\to 0} \frac{1}{n} \sum_{a,\mu=1}^{n,\ell} \left(m_\mu^a\right)^2 = \sum_{\mu=1}^{\ell} \left(m_\mu\right)^2 \tag{87}$$

$$\lim_{n\to 0} \frac{1}{n} \sum_{\substack{a,b=1 \\ a\neq b}}^{n} Q_{ab} P_{ab} = -pq \tag{88}$$

Replica symmetry ansatz

For the third term, we have that

$$(\mathbb{1} - \beta \boldsymbol{Q})_{ab} = (1 - \beta)\delta_{ab} + (1 - \beta q)(1 - \delta_{ab}) \tag{89}$$

whose eigenvalues are

$$\lambda_1 = 1 - \beta(1 - q) - \beta q n \tag{90}$$

$$\lambda_2 = 1 - \beta(1 - q) \tag{91}$$

with multiplicities 1 and $n - 1$, respectively.

Replica symmetry ansatz

We therefore have

$$\begin{aligned}\log\det\left(\mathbb{1}-\beta\boldsymbol{Q}\right) &= \ln(1-\beta(1-q)-\beta qn)+(n-1)\ln(1-\beta(1-q))\\ &= n\ln(1-\beta(1-q))-n\frac{\beta q}{1-\beta(1-q)}+\mathcal{O}(n^2)\end{aligned}$$

and so

$$\lim_{n\to 0}\log\det\left(\mathbb{1}-\beta\boldsymbol{Q}\right)=\ln(1-\beta(1-q))-\frac{\beta q}{1-\beta(1-q)} \tag{92}$$