2 Associative Memory and the Hopfield Model

The Sherrington–Kirkpatrick model is a mean-field spin glass with random couplings. A natural next step is to ask whether one can choose the couplings in a more structured way, so that a set of prescribed spin configurations become stable low-energy states. This leads to the Hopfield model of associative memory.

The basic idea is simple: one wishes to store a set of binary patterns

$$\xi_i^\mu = \pm 1, \qquad i = 1, \dots, N, \quad \mu = 1, \dots, P,$$

in a network of N Ising variables $S_i = \pm 1$, by choosing the couplings J_{ij} so that the stored patterns are attractors of the dynamics. The network should then be able to recover a full pattern from partial or noisy information.

2.1 Hebb rule

The simplest prescription for constructing the couplings is the Hebb rule:

$$\boxed{J_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu, \qquad i \neq j, \qquad J_{ii} = 0.}$$

This rule strengthens the coupling between sites that tend to be aligned in the stored patterns, and weakens or reverses it when they tend to be anti-aligned.

The corresponding Hopfield Hamiltonian is

$$H = -\frac{1}{2} \sum_{i \neq j} J_{ij} S_i S_j.$$

The deterministic zero-temperature dynamics is

$$S_i(t+1) = \operatorname{sgn}\left(\sum_j J_{ij} S_j(t) \right),$$

which lowers the energy as long as the couplings are symmetric.

2.2 Overlap order parameters

To describe retrieval of the stored memories, we introduce the overlaps

$$\boxed{m_\mu = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu S_i.}$$

If the current state S_i coincides with memory μ, then $m_\mu = 1$, while the overlaps with the other random memories are small.

Using the definition of J_{ij}, we may rewrite the energy as

$$H = -\frac{1}{2N}\sum_{\mu=1}^{P}\sum_{i\neq j}\xi_i^\mu \xi_j^\mu S_i S_j.$$

Up to an additive constant, this becomes

$$\boxed{H = -\frac{N}{2}\sum_{\mu=1}^{P} m_\mu^2.}$$

Thus the energy is lowered when one or more overlaps become nonzero.

At finite temperature, the partition function is

$$Z = \sum_{\{S_i\}} \exp\left[\frac{\beta N}{2}\sum_{\mu=1}^{P} m_\mu^2\right].$$

2.3 Mean-field free energy

Introducing Hubbard–Stratonovich variables for the overlaps, one finds

$$Z \propto \int \prod_{\mu=1}^{P} dm_\mu \; \exp\left[-N\beta f(\{m_\mu\})\right],$$

with

$$f(\{m_\mu\}) = \frac{1}{2}\sum_{\mu=1}^{P} m_\mu^2 - \frac{1}{\beta N}\sum_{i=1}^{N} \ln\left[2\cosh\left(\beta\sum_{\mu=1}^{P}\xi_i^\mu m_\mu\right)\right].$$

For random unbiased patterns, the site average may be replaced by an average over the pattern variables:

$$\boxed{f(\{m_\mu\}) = \frac{1}{2}\sum_{\mu} m_\mu^2 - \frac{1}{\beta}\left\langle \ln\left[2\cosh\left(\beta\sum_{\mu}\xi^\mu m_\mu\right)\right]\right\rangle_\xi.}$$

This is the natural free energy functional for the memory overlaps.

2.4 Landau expansion near the critical point

We first consider the case in which the number of potentially condensed memories is finite, $P = O(1)$, and expand the free energy for small m_μ. Using

$$\ln\cosh x = \frac{x^2}{2} - \frac{x^4}{12} + O(x^6),$$

with

$$x = \beta\sum_{\mu}\xi^\mu m_\mu,$$

we obtain

$$f = -\frac{\ln 2}{\beta} + \frac{1}{2}\sum_\mu m_\mu^2 - \frac{\beta}{2}\left\langle \left(\sum_\mu \xi^\mu m_\mu\right)^2 \right\rangle + \frac{\beta^3}{12}\left\langle \left(\sum_\mu \xi^\mu m_\mu\right)^4 \right\rangle + \cdots .$$

Since the ξ^μ are independent, random, and unbiased,

$$\left\langle \left(\sum_\mu \xi^\mu m_\mu\right)^2 \right\rangle = \sum_\mu m_\mu^2,$$

and

$$\left\langle \left(\sum_\mu \xi^\mu m_\mu\right)^4 \right\rangle = \sum_\mu m_\mu^4 + 6\sum_{\mu<\nu} m_\mu^2 m_\nu^2.$$

Hence

$$\boxed{f = -\frac{\ln 2}{\beta} + \frac{1-\beta}{2}\sum_\mu m_\mu^2 + \frac{\beta^3}{12}\sum_\mu m_\mu^4 + \frac{\beta^3}{2}\sum_{\mu<\nu} m_\mu^2 m_\nu^2 + O(m^6).}$$

The quadratic coefficient changes sign at

$$\boxed{T_c = 1.}$$

2.5 Selection of one memory below T_c

To determine which ordered state is preferred below T_c, suppose that n memories condense equally:

$$m_1 = \cdots = m_n = m, \qquad m_{n+1} = \cdots = 0.$$

Then the Landau free energy becomes

$$f_n = -\frac{\ln 2}{\beta} + \frac{n(1-\beta)}{2}m^2 + \frac{\beta^3}{12}n(3n-2)m^4 + \cdots .$$

Minimizing with respect to m gives

$$m^2 = \frac{3(\beta-1)}{\beta^3(3n-2)}.$$

Substituting back into the free energy yields

$$f_n^{\min} = -\frac{\ln 2}{\beta} - \frac{3n(\beta-1)^2}{4\beta^3(3n-2)}.$$

This is minimized for $n = 1$. Therefore just below the transition the system selects a state with a single nonzero overlap:

$$\boxed{\text{below } T_c, \text{ one memory is selected.}}$$

This is the basic mean-field picture of associative recall.

2.6 Finite loading: $P = \alpha N$

To discuss the storage capacity of the network, one must consider the regime in which the number of stored memories grows extensively with the size of the system:

$$P = \alpha N,$$

with α fixed as $N \to \infty$.

This is qualitatively different from the case $P = O(1)$. Even if only one memory is to be retrieved macroscopically, the remaining $P - 1$ memories contribute a finite amount of crosstalk noise. The problem therefore becomes analogous to a spin glass with one distinguished ordered direction.

One condensed pattern. We assume that one memory, say $\mu = 1$, is retrieved with a macroscopic overlap

$$m_1 \equiv m = O(1),$$

while the overlaps with all the other patterns remain small:

$$m_\mu = \frac{1}{N} \sum_i \xi_i^\mu \langle S_i \rangle = O(N^{-1/2}), \qquad \mu \geq 2.$$

Although each of these non-condensed overlaps is small, there are $O(N)$ of them, so their combined effect is finite.

Local field decomposition. The local field acting on spin S_i is

$$h_i = \sum_j J_{ij} S_j = \frac{1}{N} \sum_{\mu=1}^{P} \sum_j \xi_i^\mu \xi_j^\mu S_j = \sum_{\mu=1}^{P} \xi_i^\mu m_\mu.$$

Separating the condensed pattern from the rest,

$$h_i = \xi_i^1 m + \sum_{\mu=2}^{P} \xi_i^\mu m_\mu.$$

The first term is the desired signal; the second is the crosstalk from the other stored memories.

Gaussian crosstalk approximation. For $\mu \geq 2$, the quantities m_μ are random and of order $N^{-1/2}$. Since the ξ_i^μ are independent random signs, the sum over $\mu \geq 2$ may be treated, by the central limit theorem, as a Gaussian random variable:

$$\sum_{\mu=2}^{P} \xi_i^\mu m_\mu \;\longrightarrow\; z\sqrt{\alpha r}, \qquad z \sim \mathcal{N}(0,1),$$

where r is the variance generated by the non-condensed overlaps. Thus the local field is approximated by

$$\boxed{h_i = \xi_i^1 m + z\sqrt{\alpha r}.}$$

The task is now to determine m, the Edwards–Anderson parameter q, and the variance r self-consistently.

Thermal average of a single spin. Given a local field h_i, the thermal average of S_i is

$$\langle S_i \rangle = \tanh(\beta h_i) = \tanh\bigl[\beta(\xi_i^1 m + z\sqrt{\alpha r})\bigr].$$

Averaging over the random sign $\xi_i^1 = \pm 1$ and the Gaussian variable z, we obtain the equation for the overlap with the condensed pattern:

$$m = \frac{1}{N}\sum_i \xi_i^1 \langle S_i \rangle = \Bigl\langle \xi^1 \tanh\bigl[\beta(\xi^1 m + z\sqrt{\alpha r})\bigr] \Bigr\rangle_{\xi^1, z}.$$

Because the distribution of z is symmetric, this reduces to

$$\boxed{m = \int Dz \ \tanh\bigl[\beta(m + \sqrt{\alpha r}\, z)\bigr],}$$

where

$$Dz = \frac{dz}{\sqrt{2\pi}} e^{-z^2/2}.$$

Spin-glass order parameter. The appropriate analogue of the Edwards–Anderson parameter is

$$q = \frac{1}{N}\sum_i \langle S_i \rangle^2.$$

Using the same effective-field distribution, this becomes

$$q = \Bigl\langle \tanh^2\bigl[\beta(\xi^1 m + z\sqrt{\alpha r})\bigr] \Bigr\rangle_{\xi^1, z} = \boxed{\int Dz \ \tanh^2\bigl[\beta(m + \sqrt{\alpha r}\, z)\bigr].}$$

Determination of r. It remains to express the crosstalk variance r in terms of q. For a non-condensed pattern $\mu > 1$, write

$$h_i = h_i^{(\mu)} + \xi_i^\mu m_\mu.$$

Expanding

$$\langle S_i \rangle = \tanh(\beta h_i)$$

to first order in m_μ, one finds

$$\langle S_i \rangle \approx \tanh(\beta h_i^{(\mu)}) + \beta\bigl[1 - \tanh^2(\beta h_i^{(\mu)})\bigr]\xi_i^\mu m_\mu.$$

Multiplying by ξ_i^μ, summing over i, and using

$$q = \frac{1}{N}\sum_i \langle S_i \rangle^2,$$

gives

$$m_\mu = \eta_\mu + \beta(1 - q)m_\mu, \qquad \eta_\mu = \frac{1}{N}\sum_i \xi_i^\mu \tanh(\beta h_i^{(\mu)}).$$

Hence

$$m_\mu = \frac{\eta_\mu}{1-\beta(1-q)}.$$

Since $\overline{\eta_\mu^2} = q/N$, it follows that

$$\overline{m_\mu^2} = \frac{q/N}{[1-\beta(1-q)]^2}.$$

Summing over the $P = \alpha N$ non-condensed patterns, the crosstalk variance is

$$\alpha r = \alpha N\, \overline{m_\mu^2},$$

so

$$\boxed{r = \frac{q}{[1-\beta(1-q)]^2}.}$$

Final self-consistency equations. Collecting the results, the replica-symmetric mean-field equations in the finite-loading regime are

$$\boxed{m = \int Dz\ \tanh\bigl[\beta(m + \sqrt{\alpha r}\, z)\bigr],}$$

$$\boxed{q = \int Dz\ \tanh^2\bigl[\beta(m + \sqrt{\alpha r}\, z)\bigr],}$$

$$\boxed{r = \frac{q}{[1-\beta(1-q)]^2}.}$$

These equations have a simple interpretation:

- m measures the signal, i.e. the overlap with the retrieved memory;
- q measures the degree of freezing of the spins;
- r measures the variance of the crosstalk noise produced by all the other memories.

The finite-loading problem is therefore a competition between the signal m and the noise $\sqrt{\alpha r}$. The storage capacity is determined by the point at which the crosstalk overwhelms the signal and the retrieval solution disappears.

Remark. When $\alpha \to 0$, the crosstalk vanishes and we recover the simpler low-loading theory in which the Hopfield model undergoes an ordinary mean-field ordering transition at $T_c = 1$. For finite α, the spin-glass order parameter q and the variance r are essential, even arbitrarily close to the transition.

2.7 Capacity near the critical temperature

Close to the retrieval transition, both m and q are small. Expanding

$$\tanh x = x - \frac{x^3}{3} + \cdots, \qquad \tanh^2 x = x^2 + \cdots,$$

we obtain

$$0 = (\beta - 1)m - \beta^3 \alpha r\, m - \frac{\beta^3}{3} m^3 + \cdots,$$

$$q = \beta^2 (m^2 + \alpha r) + \cdots.$$

For small q,

$$r = \frac{q}{(\beta - 1)^2} + \cdots.$$

Eliminating r, one finds

$$q = \frac{\beta^2 (\beta - 1)^2}{(\beta - 1)^2 - \alpha \beta^2}\, m^2.$$

This is only physical if the coefficient is positive, which requires

$$(\beta - 1)^2 - \alpha \beta^2 > 0.$$

Thus retrieval is possible only when

$$\boxed{\alpha < \alpha_c(T) = \frac{(\beta - 1)^2}{\beta^2} = (1 - T)^2 \qquad (T \lesssim 1).}$$

Therefore, already close to the phase transition, the network has a finite storage capacity:

$$\boxed{\alpha_c(T) \sim (1 - T)^2.}$$

As the temperature approaches $T_c = 1$, the capacity vanishes continuously.

2.8 Zero-temperature limit

At zero temperature the dynamics becomes deterministic, and the local field at site i may be written as

$$h_i = \sum_j J_{ij} S_j.$$

Assuming memory $\mu = 1$ is retrieved, the local field can be decomposed into a signal and a crosstalk term:

$$h_i = \xi_i^1 m + z\sqrt{\alpha r}, \qquad z \sim \mathcal{N}(0, 1).$$

The zero-temperature mean-field equation is therefore

$$m = \int Dz\ \operatorname{sgn}(m + \sqrt{\alpha r}\, z) = \operatorname{erf}\left(\frac{m}{\sqrt{2\alpha r}} \right).$$

Introducing

$$x = \frac{m}{\sqrt{2\alpha r}},$$

we have

$$m = \mathrm{erf}(x).$$

At $T = 0$, one also finds

$$C \equiv \lim_{\beta\to\infty} \beta(1-q) = \sqrt{\frac{2}{\pi\alpha r}}\, e^{-x^2}, \qquad r = \frac{1}{(1-C)^2}.$$

Eliminating C and r, one arrives at the standard equation

$$\boxed{\mathrm{erf}(x) = \sqrt{2\alpha}\, x + \frac{2x}{\sqrt{\pi}} e^{-x^2}.}$$

Equivalently,

$$\boxed{\alpha(x) = \frac{1}{2}\left[\frac{\mathrm{erf}(x)}{x} - \frac{2}{\sqrt{\pi}} e^{-x^2}\right]^2.}$$

The zero-temperature capacity is obtained by maximizing this function:

$$\boxed{\alpha_c = \max_x \alpha(x) \approx 0.138.}$$

Thus the number of retrievable random memories scales as

$$\boxed{P_c \approx 0.138\, N.}$$

2.9 Physical interpretation

The Hopfield model provides a concrete example of a structured spin glass: the couplings are chosen so that a prescribed set of low-energy states is encoded in the network. For low loading, the network undergoes a mean-field ordering transition at $T_c = 1$, below which one of the memories is selected. For finite loading $P = \alpha N$, crosstalk from the other stored memories produces spin-glass-like noise, and this limits the capacity of the network.

Thus the Hopfield model interpolates naturally between:

- a retrieval phase, in which one memory is macroscopically recalled,
- and a glassy regime, in which the accumulated frustration from many stored memories destroys associative recall.

In this sense the Hopfield model is both a neural-network model and a particularly structured mean-field spin glass.