An Analysis of the Hopfield Network

Jasper Chen, Physics 212 Final Project

Autumn 2024

1. Introduction

Spin glasses model phases of matter that have locally frustrated interactions [1, 2], and thus provide us with a tool to study different optimization problems where many constraints have to be simultaneously satisfied. In this project, I reproduced some spin-glass calculations for key results on the Hopfield network[3]. The majority of calculations are dedicated towards classic results reported in [4, 5].

2. The Generalized Hopfield Network

The Hopfield network serves as a model of associative memory and is defined by a fully connected network of spins $S_{i=1,\dots,N} = \pm 1$ and their interactions W_{ij}. In order to store a set of p binary patterns $\{\xi^{\mu=1,2,\dots,p}_{i=1,2,\dots,N}\}$ in the spins, we prescribe the weights $W_{ij} = N^{-1}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu$. This prescription constitutes the training stage of the Hopfield network, which contrasts with the popular deep learning practice today, where we algorithmically search for the optimal weights via gradient descent and other optimization methods. After the weights are prescribed, we then initialize the network with an arbitrary set of spin states and update the network as follows until convergence:

$$S_i(t+1) = \mathrm{sgn}\left(\sum_{j=1,j\neq i}^{N} W_{ij}S_j - b_i\right) \tag{1}$$

where b_i is a local bias term analogous to the magnetic field h in the Ising model. If we then define the energy of the network as follows, similar to the Ising model, we have

$$H = -\frac{1}{2}\sum_{i\neq j} W_{ij}S_iS_j - \sum_{i=1}^{N} b_iS_i = -\frac{1}{2N}\sum_{i\neq j}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu S_iS_j - \sum_{i=1}^{N} b_iS_i \tag{2}$$

Then, the update rule serves to always either decrease the energy or keep it the same, because the rule is always aligning the i-th spin with the effective field it feels. We can, in fact, generalize this update rule and introduce some randomness to it, which will give us a notion of temperature in this model. Let ΔE_i be the change in energy when the i-th spin is flipped, then the spin is flipped with probability $1/(1+e^{\beta \Delta E_i})$, where β is the inverse temperature. Then, the original formulation of the Hopfield network can be seen as the $\beta \to \infty$ limit of this *generalized Hopfield network*. We can then model the network as a spin glass and assign it the Boltzmann weight $e^{-\beta H}$ after its energy has equilibriated after an arbitrarily large number of updates.

As with the Ising model, this model also has the $1 \to -1$ symmetry, so that two configurations that are flipped versions of each other have the same probability in our statistical mechanics model. A question we can ask is: given p patterns, how well can a network of size N memorize them? I have implemented a Monte Carlo simulation on two networks, $N = 100$ and $N = 200$ as a preliminary investigation. The Hopfield networks are fed with a range of patterns, and then were tasked with retrieving each of these patterns with κN bits flipped. The errors of the retrievals were evaluated with the Hamming distances between the convergent pattern and the underlying pattern to be retrieved. Doing this sampling repeatedly, I have obtained Fig. 1. We can clearly see the $1 \to -1$ symmetry in these plots. Additionally, as we might expect, the error increases as p and κ increase. Although this experiment provides us with some insight into the workings of the Hopfield network, can we be more quantitative about its description? This is the aim of the following sections.

3. Mean-Field Analysis of the Hopfield Network

We can first investigate the mean-field behavior of the model. Considering the single-spin Hamiltonian, we can isolate out an effective field at each spin site as follows:

$$H_i = -\left(\frac{1}{N} \sum_{j=1, j\neq i}^{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu S_j \right) S_i = -h_i S_i \tag{3}$$

Note that in constrast to the Ising model, the effective field here necessarily depends on i because the weights are expressed in terms of ξ_i^μ. We can then

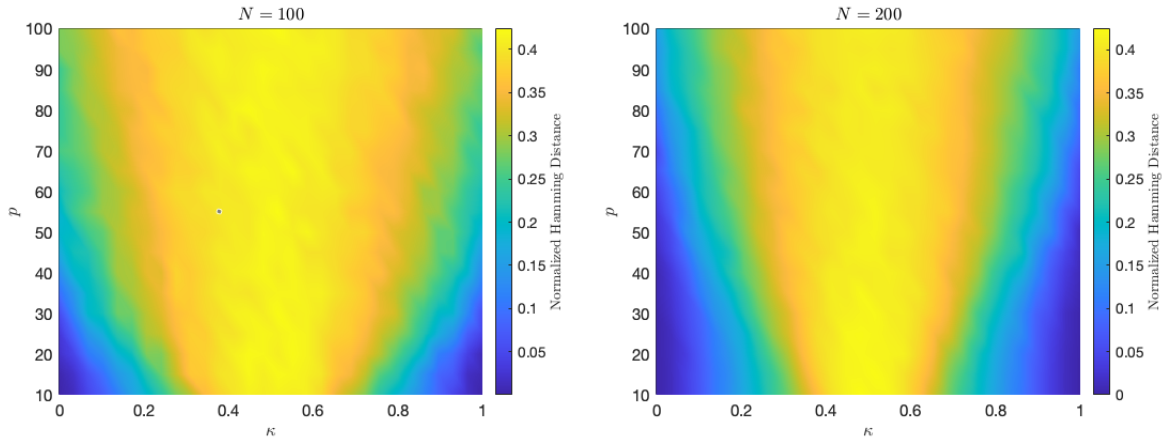

Figure 1: Hamming distance errors of retrieving perturbed patterns.

compute the thermal average of a spin:

$$\langle S_i \rangle = \frac{\sum_{S_i=\pm 1} S_i \; e^{\beta h_i S_i}}{\sum_{S_i=\pm 1} e^{\beta h_i S_i}} = \tanh(\beta h_i) \tag{4}$$

Now, if we make the mean-field assumption that fluctuations in S_i are small, we may make the approximation that $S_j \approx \langle S_j \rangle$ in the sum of h_i:

$$h_i \approx -\frac{1}{N} \sum_{j=1, j\neq i}^{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \langle S_i \rangle \;\Rightarrow\; \langle S_i \rangle = \tanh\left(\frac{\beta}{N} \sum_{j=1, j\neq i}^{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \langle S_j \rangle \right) \tag{5}$$

This result provides a natural definition for an order parameter describing the network — not the average magnetization but rather the average *overlap*: $m^\nu = \sum_j \xi_j^\nu \langle S_j \rangle / N$. Using these definitions, we then have

$$\langle S_i \rangle = \tanh(\beta \mathbf{m} \cdot \boldsymbol{\xi}_i) \tag{6}$$

where $\mathbf{m} = (m^1, m^2, \ldots, m^p)$ and $\boldsymbol{\xi}_i = (\xi_i^1, \xi_i^2, \ldots, \xi_i^p)$. This equation is simple but does not provide a lot of information about the learning capacity of the network. However, it does tell us that we need to study the quantity $\mathbf{m}$ instead.

4. Free Energy of the Hopfield Network

We turn to the thermodynamics of the network for more theoretical descriptions of its behaviors. Our goal is to evaluate the free energy $f =$

$(N\beta)^{-1}\langle \log Z\rangle_d$, and find the stable states of the system. Here, $\langle\cdot\rangle_d$ denotes average over the quenched disorder $\{\xi_i^\nu\}$.

We first study the limit where $N\to\infty$ but p stays finite, so that $\alpha = p/N$ goes to zero. Here, we follow Ref. [4] and evaluate

$$f = -\lim_{N\to\infty}(N\beta)^{-1}\left\langle \log \sum_{S_i=\pm 1} \exp\left(\frac{\beta}{2N}\sum_{i\neq j}^{N}\sum_{\nu=1}^{p}\xi_i^\nu\xi_j^\nu S_i S_j\right)\right\rangle_d \tag{7}$$

We first decouple the sum over i and j by completing the square

$$f = -\lim_{N\to\infty}(N\beta)^{-1}\left\langle \log e^{-\beta p/2}\sum_{S_i=\pm 1}\left[\exp\frac{\beta}{2N}\sum_{\nu=1}^{p}\left(\sum_{i=1}^{N}\xi_i^\nu S_i\right)^2\right]\right\rangle_d \tag{8}$$

$$= -\lim_{N\to\infty}(N\beta)^{-1}\left\langle \log e^{-\beta p/2}\sum_{S_i=\pm 1}\prod_{\nu=1}^{p}\left[\exp\frac{\beta}{2N}\left(\sum_{i=1}^{N}\xi_i^\nu S_i\right)^2\right]\right\rangle_d \tag{9}$$

Using the integral identity (sometimes called the Hubbard-Stratonovich transform)

$$\exp(\lambda a^2) = \int \frac{dz}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}+\sqrt{2\lambda}az} \tag{10}$$

we can then transform this expression into

$$f = -\lim_{N\to\infty}(N\beta)^{-1}\left\langle \log e^{-\beta p/2}\sum_{S_i=\pm 1}\left[\int\prod_{\nu=1}^{p}\frac{dz}{\sqrt{2\pi}}e^{-(z^\nu)^2/2+\sqrt{\beta/N}\sum_{i=1}^{N}\xi_i^\nu S_i z^\nu}\right]\right\rangle_d \tag{11}$$

Rescaling the variable $z^\nu = \sqrt{N\beta}m^\nu$, and expanding out the product over ν, we obtain sums inside the exponential of the integrand

$$f = -\lim_{N\to\infty}(N\beta)^{-1}\left\langle \log e^{-\beta p/2}(N\beta)^{p/2}\times \right.$$
$$\left.\sum_{S_i=\pm 1}\left[\int\prod_{\nu=1}^{p}\frac{dm^\nu}{\sqrt{2\pi}}e^{-\sum_{\nu=1}^{p}N\beta(m^\nu)^2/2}\prod_{i=1}^{N}e^{\beta\sum_{\nu=1}^{p}\xi_i^\nu S_i m^\nu}\right]\right\rangle_d \tag{12}$$

Now we perform the sum over S_i, we obtain hyperbolic cosines over the decoupled sum. Introducing vector notations for $\mathbf{m} = m^{\nu=1,\ldots,p}$ and $\boldsymbol{\xi}_i =$

$\xi_i^{\nu=1,\ldots,p}$, the free energy is then given by

$$f = -\lim_{N\to\infty}(N\beta)^{-1}\Bigg\langle \log e^{-\beta p/2}(N\beta)^{p/2} \times \Bigg[\int \prod_{\nu=1}^{p}\frac{dm^\nu}{\sqrt{2\pi}} e^{-N\beta\mathbf{m}^2/2}\prod_{i=1}^{N}[2\cosh(\beta\boldsymbol{\xi}_i\cdot\mathbf{m})]\Bigg]\Bigg\rangle_d \quad (13)$$

We rewrite $\prod_{i=1}^{N} 2\cosh(\beta\boldsymbol{\xi}_i\cdot\mathbf{m}) = e^{\sum_i \log 2\cosh(\beta\boldsymbol{\xi}_i\cdot\mathbf{m})}$. Here we want to use the saddle point approximation to perform this integration, but special notice should be paid to the fact that $\log 2\cosh(\beta\boldsymbol{\xi}_i\cdot\mathbf{m}) > 0$ for all $\boldsymbol{\xi}_i$ and $\mathbf{m}$. If this term becomes too large, the saddle point approximation does not apply. Their dot product scales with p, which is kept finite in this calculation, so we may proceed:

$$f = -\lim_{N\to\infty}(N\beta)^{-1}\Bigg\langle -\frac{\beta p}{2} + \frac{p}{2}\log\beta N + \log\int\prod_{\nu=1}^{p}\frac{dm^\nu}{\sqrt{2\pi}} e^{-N\beta\mathbf{m}^2/2+\sum_i \log 2\cosh(\beta\boldsymbol{\xi}_i\cdot\mathbf{m})}\Bigg]\Bigg\rangle_d \quad (14)$$

Let us then denote the minima of the integrand as $\mathbf{m}^*$, and define the function

$$g(\mathbf{m}) = -\frac{\beta\mathbf{m}^2}{2} + \frac{1}{N}\sum_{i=1}^{N}\log 2\cosh(\beta\mathbf{m}\cdot\boldsymbol{\xi}_i) \quad (15)$$

The saddle-point approximation then reads

$$\int\prod_{\nu=1}^{p}\frac{dm^\nu}{\sqrt{2\pi}} e^{-Ng(\mathbf{m})} \approx \frac{e^{-Ng(\mathbf{m}^*)}}{\sqrt{\det(-NA)}} \quad (16)$$

where A is the $p\times p$ Hessian matrix evaluated at $\mathbf{m}^*$

$$A^{\mu\nu} = \left.\frac{\partial^2 g}{\partial m^\mu \partial m^\nu}\right|_{\mathbf{m}=\mathbf{m}^*} \quad (17)$$

Substituting in this result and taking the thermodynamic limit, we have

$$f = -\lim_{N\to\infty}\left\langle \frac{P}{2N} + \frac{p\log\beta N}{\beta N} - \frac{g(\mathbf{m}^*)}{\beta} - \frac{\log\sqrt{\det(-NA)}}{\beta N}\right\rangle_d \quad (18)$$

$$f(\mathbf{m}^*) = \langle|\mathbf{m}^*|^2/2 - \beta^{-1}\log 2\cosh(\beta\mathbf{m}^*\cdot\boldsymbol{\xi}))\rangle_d \quad (19)$$

where we used the law of large numbers to make the replacement $\frac{1}{N}\sum_{i=1}^{N} \to \langle\cdot\rangle_d$ in $g(\mathbf{m})$. The quenched disorder $\{\xi_i^\nu\}$ is said to be *self-averaging*. Also, note that $\log\sqrt{\det(-NA)} \sim p\log N$, so the logarithm makes sure the last term vanishes in the thermodynamic limit.

5. Solving the mean-field equations

It may be odd that we have introduced a rather suggestive nomenclature for the arbitrary Gaussian variable m^ν we introduced in the integral transform above, but we can make it clear as follows.

By differentiating the free energy we find the self-consistent equation describing the free energy ground states

$$\nabla f(\mathbf{m})\Big|_{\mathbf{m}=\mathbf{m}^*} = \langle -\mathbf{m} + \boldsymbol{\xi}\tanh(\beta\mathbf{m}\cdot\boldsymbol{\xi})\rangle_d\Big|_{\mathbf{m}=\mathbf{m}^*} = \mathbf{0} \tag{20}$$

or

$$\mathbf{m}^* = \langle\boldsymbol{\xi}\tanh(\beta\mathbf{m}^*\cdot\boldsymbol{\xi})\rangle_d \tag{21}$$

But this equation is the same as the mean-field magnetization 6 with $\boldsymbol{\xi}_i$ multiplied and then averaged over disorder. This shows us that $\mathbf{m}$ is the same as we had defined it in the mean-field analysis!

To actually solve this equation, though, we have to prescribe a probability distribution to the disorder ξ_i^ν. For simplicity, we can simply take the uniform distribution, so that each ξ_i^ν is identically and independently distributed with probability $p(\xi_i^\nu = 1) = p(\xi_i^\nu = -1) = 1/2$. This gives us $\langle\xi_\nu\rangle_d = 0$ and $\langle\xi_\mu\xi_\nu\rangle_d = \delta_{\mu\nu}$. We then expand the expression for f and solve the equation perturbatively for $\mathbf{m}^*$

$$f(\mathbf{m}) = \langle\mathbf{m}^2/2 - \beta^{-1}\log 2\cosh(\beta\mathbf{m}\cdot\boldsymbol{\xi}))\rangle_d \tag{22}$$

$$\approx \mathbf{m}^2/2 - \beta^{-1}\langle\log 2 + (\beta\mathbf{m}\cdot\boldsymbol{\xi})^2/2\rangle_d \tag{23}$$

From our prescribed distribution for ξ, we can calculate

$$\langle(\mathbf{m}\cdot\boldsymbol{\xi})^2\rangle_d = \left\langle\sum_{\mu=1}^{N}\sum_{\nu=1}^{N} m^\mu m^\nu\xi_\mu\xi_\nu\right\rangle_d = \sum_{\mu=1}^{N}\sum_{\nu=1}^{N} m^\mu m^\nu\langle\xi_\mu\xi_\nu\rangle_d$$

$$= \sum_{\mu=1}^{N}\sum_{\nu=1}^{N} m^\mu m^\nu\delta_{\mu\nu} = \mathbf{m}^2 \tag{24}$$

Hence, the free energy to leading order is given by

$$f(\mathbf{m};\beta) = -\beta^{-1}\log 2 + (1-\beta)\mathbf{m}^2/2 \tag{25}$$

When $\beta < 1$ or $T > 1$, this function has one and only one global minimum at $\mathbf{m}^* = 0$, with $f(\mathbf{m}^*) = -\beta^{-1}\log 2$. This solution represents disordered states where the network is unable to memorize any of the patterns. Then, we also observe a phase transition at $\beta = 1$ where $\mathbf{m}^* = 0$ becomes unstable. Hence, we conclude that $\beta > 1$ corresponds to a disordered phase where there are no macroscopic overlaps with the prescribed patterns and therefore no useful memorization happens.

To understand the detailed structure of the ground states $\mathbf{m}^*$, we need to expand the mean-field equation for $\mathbf{m}$ as well.

$$\begin{aligned}
m^\nu &= \langle \xi_\nu \tanh(\beta \mathbf{m}\cdot\boldsymbol{\xi})\rangle_d && (26)\\
&\approx \left\langle \xi_\nu \beta \mathbf{m}\cdot\xi - \xi_\nu \frac{(\beta\mathbf{m}^*\cdot\boldsymbol{\xi})^3}{3}\right\rangle_d && (27)\\
&= \beta\sum_{\mu=1}^{p} m^\mu \langle \xi_\nu\xi_\mu\rangle_d - \frac{\beta^3}{3}\sum_{\mu_1=1}^{p}\sum_{\mu_2=1}^{p}\sum_{\mu_3=1}^{p} m^{\mu_1}m^{\mu_2}m^{\mu_3}\langle \xi_\nu\xi_{\mu_1}\xi_{\mu_2}\xi_{\mu_3}\rangle_d && (28)\\
&= \beta\sum_{\mu=1}^{p} m^\mu \delta_{\mu\nu} - \frac{\beta^3}{3}\sum_{\mu_1=1}^{p}\sum_{\mu_2=1}^{p}\sum_{\mu_3=1}^{p} m^{\mu_1}m^{\mu_2}m^{\mu_3}\langle \xi_\nu\xi_{\mu_1}\xi_{\mu_2}\xi_{\mu_3}\rangle_d && (29)
\end{aligned}$$

To evaluate the expected value $\langle \xi_\nu\xi_{\mu_1}\xi_{\mu_2}\xi_{\mu_3}\rangle_d$, we note that this value is only non-zero if ν equals one of the μ_i, and the other two are the same. However, if we evaluate this sum, we would count the special case where $\nu = \mu_1 = \mu_2 = \mu_3$ three times. Hence, accounting for the overcounting, we arrive at

$$\begin{aligned}
m^\nu &= \beta m^\nu - \frac{\beta^3}{3}(3m^\nu\mathbf{m}^2 - 2(m^\nu)^3) && (30)\\
m^\nu &= \beta m^\nu + \frac{2\beta^3}{3}(m^\nu)^3 - \beta^3 m^\nu \mathbf{m}^2 && (31)
\end{aligned}$$

From this expression for m^ν, along with the expression for f, we can identify a symmetry where the f is invariant and the m^ν equations remain satisfied under permutation of $\{m^\nu\}$ and sign flips $m^\nu \to -m^\nu$. Therefore we can restrict ourselves to studying solutions $\mathbf{m} = (m^1,\ldots,m^n,0,\ldots,0)$, where $m^i > 0$. For simplicity, we first consider these solutions where all non-zero

components have the same value m_0. We can show the pertinence of these so-called *symmetric* solutions by dividing the equation for m^ν by m^ν:

$$1 = \beta + \frac{2\beta^3}{3}(m^\nu)^3 - \beta^3 \mathbf{m}^2 \tag{32}$$

$$\Rightarrow \frac{2}{3}(m^\nu)^2 = \beta^{-2}(\beta^{-1} - 1) + \mathbf{m}^2 \tag{33}$$

which shows that m^ν is independent of ν at least near $\mathbf{m} \sim 0$ (stable at $\beta < 1$). Therefore, these symmetric solutions exist for $\beta < 1$ and are the only minima of f around the critical $\beta_c = 1$. If $n = 1$, these symmetric solutions represent overlap with one particular pattern and are called *Mattis states*. If $n > 1$, these solutions represent equal mixtures of n different patterns. To get a more detailed picture, we can study the behavior of the free energy with these states.

The expansion 25 suffices to describe the behavior of the system around $\beta_c = 1$. Plugging in these symmetric solutions yields

$$f_n = \frac{n}{2}(1 - \beta)m_0^2 - \frac{\log 2}{\beta} \tag{34}$$

To calculate the value of m_0 near β_c, we use the expansion 31. The calculation is essentially the same as 31, but with $m^{\mu_1} = m^{\mu_2} = m^{\mu_3} = m_0$.

$$m_0 = \beta m_0 - \frac{\beta^3 m_0^3}{3}(3n - 2) \tag{35}$$

or equivalently (ignoring the trivial solution)

$$\frac{3n-2}{3}\beta^3 m_0^3 = \beta - 1 \;\Rightarrow\; m_0^2 = \frac{3(\beta - 1)}{\beta^3(3n - 2)} \tag{36}$$

Using this expression for m_0, the free energy is

$$f_n = \frac{n}{2}(1 - \beta)\frac{3(\beta - 1)}{\beta^3(3n - 2)} = -\frac{3n(1 - \beta)^2}{2\beta^3(3n - 2)} \tag{37}$$

The ground state energy near β_c is therefore given by the $n = 1$ Mattis states

$$f_1 = -\frac{3(1 - \beta)^2}{2\beta^3},\; m_0^{(1)} = \frac{3(\beta - 1)}{\beta^3} \tag{38}$$

We would also like to study the ground states near $\beta \to \infty$, and to do this, we take the limits of equations 19 and 21 as $\beta \to \infty$. We use the fact that

$$\lim_{\beta\to\infty} \cosh(\beta x) = \lim_{\beta\to\infty} \frac{e^{\beta x} + e^{-\beta x}}{2} \approx \begin{cases} e^{\beta x}/2 & x > 0 \\ e^{-\beta x}/2 & x < 0 \end{cases} \tag{39}$$

to conclude

$$\lim_{\beta\to\infty} \beta^{-1} \log 2\cosh(\beta \mathbf{m}\cdot\boldsymbol{\xi}) = \left.\begin{cases} \mathbf{m}\cdot\boldsymbol{\xi} & \mathbf{m}\cdot\boldsymbol{\xi} > 0 \\ -\mathbf{m}\cdot\boldsymbol{\xi} & \mathbf{m}\cdot\boldsymbol{\xi} < 0 \end{cases}\right\} = |\mathbf{m}\cdot\boldsymbol{\xi}| \tag{40}$$

which in turn gives

$$\lim_{\beta\to\infty} f = \mathbf{m}^2/2 - \langle|\mathbf{m}\cdot\boldsymbol{\xi}|\rangle_d = nm_0^2/2 \tag{41}$$

Note that this average vanishes because there is only one power of ξ in the average.

Similarly, we can take the $\beta \to \infty$ limit for $\mathbf{m}$ as well, which yields

$$\lim_{\beta\to\infty} \mathbf{m} = \lim_{\beta\to\infty} \langle \boldsymbol{\xi} \tanh(\beta \mathbf{m}\cdot\boldsymbol{\xi})\rangle_d = \langle \boldsymbol{\xi}\ \mathrm{sgn}_0(\mathbf{m}\cdot\boldsymbol{\xi})\rangle_d \tag{42}$$

where $\mathrm{sgn}_0(x)$ is the modified sign function so that $\mathrm{sgn}_0(0) = 0$. To evaluate this average, we can isolate out a component, say the first component, without loss of generality due to the solution's symmetry.

$$m_0 = \langle \xi_1 \mathrm{sgn}_0(\mathbf{m}\cdot\boldsymbol{\xi})\rangle_d = \langle \xi_1 \mathrm{sgn}_0(m_0 \textstyle\sum_{\nu=1}^n \xi_\nu)\rangle_d = \langle \xi_1 \mathrm{sgn}_0(\sum_{\nu=1}^n \xi_\nu)\rangle_d \tag{43}$$

where the positive value m_0 does not affect the value of the sign function. Taking the average over only ξ_1 gives

$$m_0 = \frac{1}{2}\langle \mathrm{sgn}_0(1 + \textstyle\sum_{\nu=2}^n \xi_\nu)\rangle_d - \frac{1}{2}\langle \mathrm{sgn}_0(-1 + \sum_{\nu=2}^n \xi_\nu)\rangle_d \tag{44}$$

By symmetry, we must have $\langle \mathrm{sgn}_0(1+\sum_{\nu=2}^n \xi_\nu)\rangle_d = -\langle \mathrm{sgn}_0(-1+\sum_{\nu=2}^n \xi_\nu)\rangle_d$, so that $m_0 = \langle \mathrm{sgn}_0(1 + \sum_{\nu=2}^n \xi_\nu)\rangle_d$.

For readability, we denote $\sum_{\nu=2}^n \xi_\nu = z_n$. We can carry out the averaging over disorder for odd and even n separately. Let us first consider the odd case. Here, the sum $\sum_{\nu=2}^n \xi_\nu$ can only take on values in $\{\ldots -4, -2, 0, 2, 4\}$. Hence,

$$m_0^{\mathrm{odd}} = -1 \times \mathcal{P}(z_n \le -2) \ + \ 1 \times \mathcal{P}(z_n \ge 0) \tag{45}$$

By symmetry again, we must have

$$P(z_n \leq -2) = P(z_n \geq 2) \tag{46}$$

Now, by properties of a discrete probability distribution, we further observe that

$$\mathcal{P}(z_n \geq 0) = \mathcal{P}(z_n = 0) + \mathcal{P}(z_n \geq 2) \tag{47}$$

Combining these two observations, we have

$$m_0^{\text{odd}} = \mathcal{P}(z_n = 0) = \frac{1}{2^{n-1}} \binom{n-1}{(n-1)/2} \tag{48}$$

The calculation for the even case is similar, except now the sum z_n can take on values $\{\dots, -3, -1, 1, 3, \dots\}$:

$$\begin{aligned} m_0^{\text{even}} &= -1 \times \mathcal{P}(z_n \leq -3) + 0 \times \mathcal{P}(z_n = -1) + 1 \times \mathcal{P}(z_n \geq 1) && (49)\\ &= -1 \times \mathcal{P}(z_n \geq 3) + 1 \times \mathcal{P}(z_n \geq 1) && (50)\\ &= \mathcal{P}(z_n = 1) && (51)\\ &= \frac{1}{2^{n-1}} \binom{n-1}{(n-2)/2} && (52)\\ &= \frac{1}{2^n} \binom{n}{n/2} && (53) \end{aligned}$$

Plugging these expressions for m_0 into the free energy expression 41 yields the plot in Fig. 2. From this result, we conclude that the ground states of free energy for symmetric states near $\beta \to \infty$ are *also* the Mattis states with $n = 1$, as in the case of near $\beta_c = 1$. Since we can expect asymmetric states to have higher energy than the symmetric states, as they appear only at a temperature lower than $\beta_c = 1$, we can also conclude that the Mattis states are the ground states of the Hopfield network at both $\beta = 1$ and $\beta = \infty$. However, the system was shown by equation 25 to only have a phase transition at $\beta_c = 1$, we can also conclude that the Mattis states are the ground states for $\beta \in [1, \infty)$, albeit new local minima consisting of asymmetric solutions do appear at lower temperature.

6. The Infinite p Limit of the Hopfield Network

This analysis above is in the limit that $N \to \infty$ but p stays finite. It therefore informs us how noise in the network dynamics impacts its ability to memorize patterns. However, we are also interested in how the network's memory

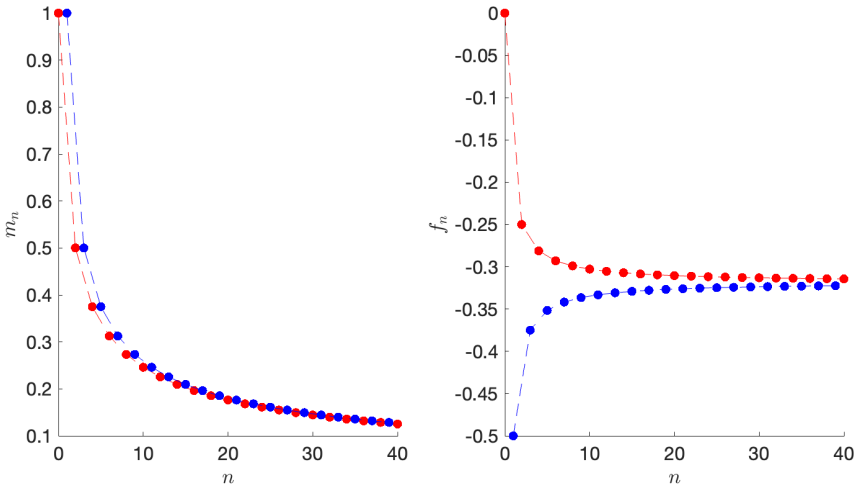

Figure 2: The values of m_n and f_n near $\beta \to \infty$. Colored in blue are the odd states, and colored in red are the even states.

capability varies with the number of patterns to be memorized! Quantitatively, we would like to study the $p \to \infty$ limit but with $\alpha = p/N$ staying finite. The free energy calculation in this limit is significantly more complicated, involving using the replica trick and is beyond the scope of this project. However, a very detailed, step-by-step calculation is found in [6] So we simply quote the result.

$$f_\alpha = \frac{1}{2}(\mathbf{m}^<)^2 + \frac{1}{2}\alpha\left(\beta^{-1}\log[1-\beta(1-q)] + \frac{(1-\beta)(1-q)}{1-\beta(1-q)} + \beta r(1-q)\right) - (\sqrt{2\pi}\beta)^{-1}\langle \int_{-\infty}^{\infty} dz\ e^{-z^2/2}\log 2\cosh\beta[\sqrt{\alpha r}z + \mathbf{m}^< \cdot \boldsymbol{\xi}^<]\rangle_d \quad (54)$$

where $\mathbf{m}^<$ and $\boldsymbol{\xi}^<$ represent the $s < p$ patterns that have macroscopic overlaps with the network states. This is because as $p \to \infty$, there can be a macroscopic number of patterns that do not have meaningful overlap with the network states. To account for this, the new order parameter $r = \langle \alpha^{-1}\sum_{\mu>s}^{N\alpha}(m^\mu)^2\rangle_d$. The other new order parameter emerging from the calculation is $q = \sum_{i=1}^{N}\langle\langle S_i\rangle^2\rangle_d$. With this free energy, we can draw the complete phase diagram of the Hopfield network in the β-α plane. Also note how we recover the finite p analysis for $\alpha = 0$.

Here, we focus on the $\beta = \infty$ analysis. As with the finite p case, we can obtain the equations for m^μ by differentiating f

$$\frac{\partial f_\alpha}{\partial m^\mu} = m^\mu - (\sqrt{2\pi}\beta)^{-1}\langle \beta \xi^\mu \int_{-\infty}^{\infty} dz\ e^{-z^2/2}\ \tanh\beta[\sqrt{\alpha r}z + \mathbf{m}^< \cdot \boldsymbol{\xi}^<]\rangle_d \quad (55)$$

for $\mu = 1, 2, \ldots, s$, which gives us

$$\mathbf{m} = (\sqrt{2\pi})^{-1}\langle \boldsymbol{\xi}^< \int_{-\infty}^{\infty} dz\ e^{-z^2/2}\ \tanh\beta[\sqrt{\alpha r}z + \mathbf{m}^< \cdot \boldsymbol{\xi}^<]\rangle_d \quad (56)$$

Similarly, we can do so for the other two order parameters.

$$0 = \frac{\partial f_\alpha}{\partial q} = \frac{\alpha}{2[(1-\beta(1-q)]} + \frac{\alpha(1-\beta)}{2[1-\beta(1-q)]^2} - \frac{\alpha\beta}{2}r \quad (57)$$

$$= \frac{\alpha\beta q}{2[1-\beta(1-q)]^2} - \frac{\alpha\beta}{2} \quad (58)$$

$$\Rightarrow r = \frac{q}{[1-\beta(1-q)]^2} \quad (59)$$

and

$$0 = \frac{\partial f_\alpha}{\partial r} = \frac{\alpha\beta}{2}(1-q) - \int_{-\infty}^{\infty} dz\ \frac{e^{-z^2/2}}{2}\sqrt{\frac{\alpha}{r}}z\langle\tanh\beta[\sqrt{\alpha r}z + \mathbf{m}^< \cdot \boldsymbol{\xi}^<]\rangle_d \quad (60)$$

$$\Rightarrow \sqrt{\alpha r}\beta(1-q) = \int_{-\infty}^{\infty} dz\ e^{-z^2/2}z\langle\tanh\beta[\sqrt{\alpha r}z + \mathbf{m}^< \cdot \boldsymbol{\xi}^<]\rangle_d \quad (61)$$

Integrating by parts and cancelling out constants from both sides, we find that

$$q = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dz\ e^{-z^2/2}\langle\tanh^2\beta[\sqrt{\alpha r}z + \mathbf{m}^< \cdot \boldsymbol{\xi}^<]\rangle_d \quad (62)$$

From the finite p analysis above, we know that the ground states of the Hopfield network are the Mattis states $m^\mu = m_0\delta_{\mu\nu}$. Here, we focus on these states as well. As $\beta \to \infty$, we can approximate tanh as the sign function, and therefore we obtain

$$m_0 = (\sqrt{2\pi})^{-1}\int_{-\infty}^{\infty} dz\ e^{-z^2/2}\ \langle\xi\ \mathrm{sgn}(\sqrt{\alpha r}z + m_0)\rangle_d \quad (63)$$

$$= \frac{1}{2\sqrt{2\pi}}\int_{-\infty}^{\infty} dz\ e^{-z^2/2}\ [\mathrm{sgn}(\sqrt{\alpha r}z + m_0) - \mathrm{sgn}(\sqrt{\alpha r}z - m_0)] \quad (64)$$

Here the integrand is only non-zero on the interval $[-m/\sqrt{\alpha r}, m/\sqrt{\alpha r}]$, on which it is identically 2. Therefore

$$m_0 = \frac{1}{\sqrt{2\pi}} \int_{-m_0/\sqrt{\alpha r}}^{m_0/\sqrt{\alpha r}} dz\ e^{-z^2/2} \tag{65}$$

$$= \sqrt{\frac{2}{\pi}} \int_{0}^{m_0/\sqrt{\alpha r}} dz\ e^{-z^2/2} \tag{66}$$

$$= \mathrm{erf}(m_0/\sqrt{2\alpha r}) \tag{67}$$

To obtain a solution for m_0, we also need the equation for q and r. We first rewrite $\tanh^2$ in the equation for q and then take the limit $\beta \to \infty$, treating the sharply-peaked integrand as a delta function:

$$q = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dz\ e^{-z^2/2} \langle 1 - \cosh^{-2} \beta[\sqrt{\alpha r} z + m_0 \xi] \rangle_d \tag{68}$$

$$= 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dz\ e^{-z^2/2} \langle \cosh^{-2} \beta[\sqrt{\alpha r} z + m_0 \xi] \rangle_d \tag{69}$$

$$= 1 - \sqrt{\frac{2}{\pi}} \int_{-\infty}^{\infty} dz\ e^{-z^2/2} \cosh^{-2} \beta[\sqrt{\alpha r} z - m_0] \tag{70}$$

$$= 1 - \sqrt{\frac{2}{\alpha \pi r}} \int_{-\infty}^{\infty} dx\ e^{\frac{-(x+m_0)^2}{2\alpha r}} \cosh^{-2} \beta x \tag{71}$$

$$\approx 1 - \sqrt{\frac{2}{\alpha \beta^2 \pi r}} e^{-\frac{m_0^2}{2\alpha r}} \tag{72}$$

which after some algebra gives

$$r = \left(1 - \sqrt{\frac{2}{\alpha \pi r}} e^{-\frac{m_0^2}{2\alpha r}}\right)^{-2} \tag{73}$$

We can then numerically solve this system of equations for m_0 using the fixed point iteration method. This computation yields the following results in Fig. 3, where we can observe a discontinuous phase transition at $\alpha \approx 0.14$, where m_0 sharply drops to zero from a state of finite memorization. This is consistent with Hopfield's observation in his 1982 paper [3]. The behavior r correspondingly shoots up sharply, indicating spurious overlaps overtaking the macroscopic overlaps.

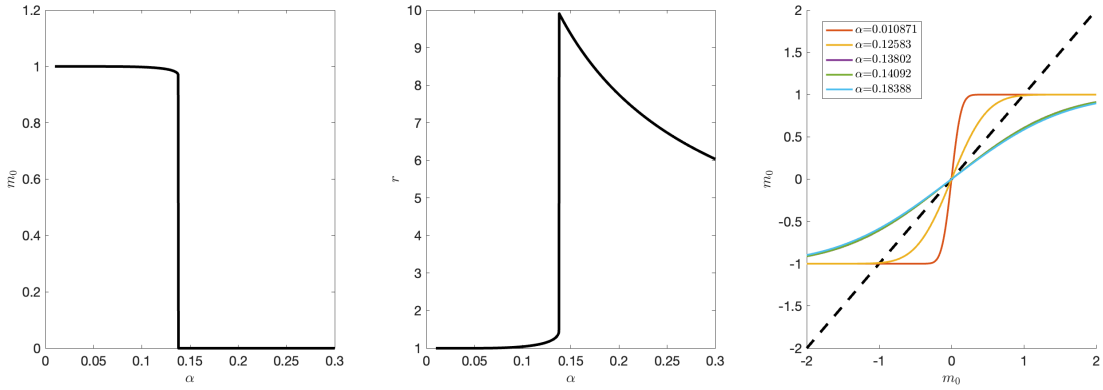

Figure 3: Numerical solutions for m_0 (left), r (middle), and the plots for the self-consistent equation for m_0 (right) at $\beta = \infty$.

This concludes our investigation into the Hopfield network, although much remains to be studied and the ideas perhaps applied to the puzzles around deep neural networks today.

Code. The code written for this project can be found at this Github repository.

References

1. Castellani, T. & Cavagna, A. Spin-glass theory for pedestrians. *Journal of Statistical Mechanics: Theory and Experiment* **2005,** P05012 (2005).
2. Mézard, M. Spin glasses and optimization in complex systems. *Europhysics News* **53,** 15–17 (2022).
3. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* **79,** 2554–2558 (1982).
4. Amit, D. J., Gutfreund, H. & Sompolinsky, H. Spin-glass models of neural networks. *Physical Review A* **32,** 1007–1018 (1, 1985).
5. Amit, D. J., Gutfreund, H. & Sompolinsky, H. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters* **55,** 1530 (1985).

6. Dotsenko, V. *An introduction to the theory of spin glasses and neural networks* (World Scientific, 1995).