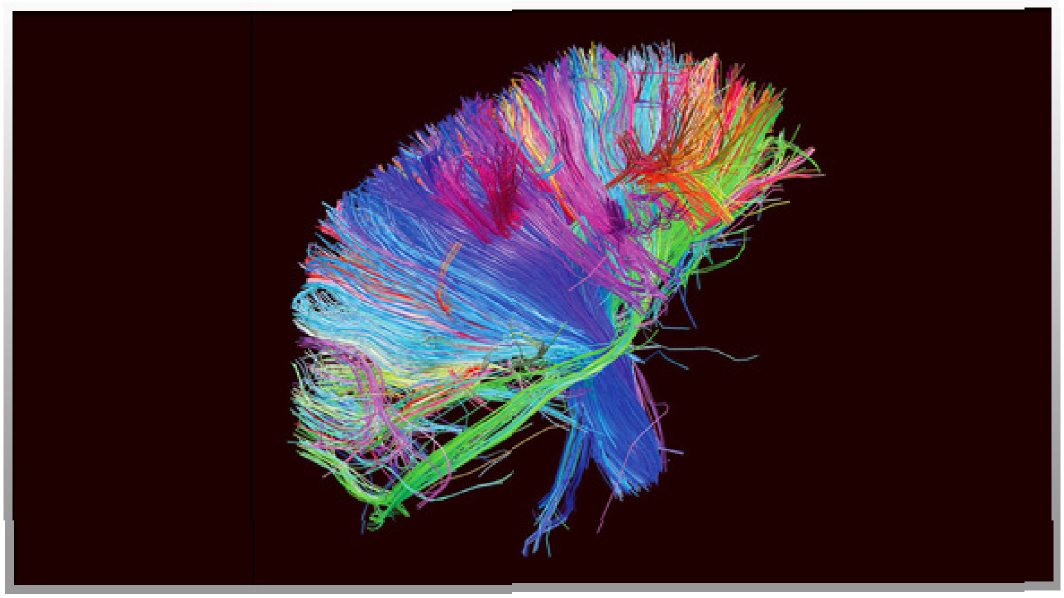
Neural Networks

Pedro Jesús Jódar Siles



Abstract

In the present work I have made a bibliographic study of Hopfield model and some variations that can improve its performance. I have summarized the motivation behind the model and I have characterized some of its main properties, namely capacity and stability of the stored pattern. These two parameters has been checked numerically. Additionally, an introduction to learning methodology inspired in Hebb rule has been included.

Contents

1 Introduction.

The study of thought and perception has been, and continues to be, a field of enormous interest. In addition to its special relevance from a humanistic perspective, it also arouses interest in more applied areas. Traditionally, the understanding of cognitive phenomena has been used to improve the treatment of psychiatric diseases. However, in recent years a growing interest has arisen in fields related to engineering. A highly active field is neuromorphic engineering, which is devoted to developing algorithms and devices that replicate biological behaviors for practical purposes. Neuroscience is an area where the interests of professionals with very diverse backgrounds and interests intersect.

Perhaps one of the first intellectuals who hypothesized that brain is related to thought was Descartes. In the *Homine* section of *Meditationes* Descartes proposed that perception appears as a result of the connection of sensory organs with a processing unit via nerves. The brain process information and the result is then transmitted to the executing organs via nerves. This scheme has remained valid until our days. However, it has an important limitation: it does not propose the physical mechanism that allows the transmission of information. This question remained unanswered until 1791, when Galvani proposed that information was transported by electric currents.

To wait for the next important milestone we have to move to the discovery of the neuron by Ramón y Cajal, more than one century after. This discovery was revolutionary, for the first time the basic element of the system was known. However, it has been shown that most phenomena observed in the brain cannot be simply described as a sum of the behaviour of independent neurons, emergent behaviours are crucial to understand the phenomena observed in the nervous system.

In order to study how the human being interacts with the environment the study of psychology is also necessary. In this area, behaviorism deserves special attention. This school treats the brain as a black box that receives inputs and returns outputs. The object of study is the behavior resulting from different stimuli. However, in 1949, Donald Hebb managed to show that an analogy between learning processes and synaptic connection could be established. This discovered lead to the creation of biopsychology, a new field that bridges psychology with biology, .

Using the rules of Hebbian learning Hopfield managed to create a neural network of associative memory. To develop this model, he used the same rules defined by Ising years earlier to describe the behavior of a ferromagnetic materials. This connection increased the interest of theoretical physicists in the field of neural networks.

In the present work we will explore the behavior of Hopfield network. We will first describe the biological motivation of the model, we will continue with a psychological motivation and them we will describe the Hopfield model itself and a variation of this model, the Boltzmann network.

1.1 Biological Motivation.

1.1.1 Anatomy

The structure of the nervous system broadly corresponds to the structure proposed by Descartes. Sensory organs send information to the cortex, this information is previously managed by the thalamus, the cortex processes the information and sends it back to the thalamus to be to the thalamus for transmission to the motor organs. In addition to these structures, there are auxiliary organs (hypothalamus, reticular formation, nigrostriatal formation, cerebellum, hippocampus and colliculus) that contribute to processing by performing specific tasks.

When the cortex is studied in detail regions with a specific neuronal density or thickness can be distinguished; these are cortical areas. Each cortical area perform a set of specific functions. Cortical areas can be subsequently separated into more simple structures where microcolumns (strongly connected neurons assembled in narrow arrays) are the most simple ones. In an upper level we can observe modules, complex structures that manage specific spectrum of the signals. These later structures do not have well defined anatomic properties [1] .

1.1.2 Synapses.

In complex system we are interested in understanding the mechanisms that generate patterns observed in a macroscopic-or mesoscopic- scale. In neuroscience the mechanism that governs interactions between neurons is synapses, namely inside neurons information travels as an electrical current and the in the region between neurons information is transferred using chemical components called neurotransmitters. Synapsis starts with a neurotransmitter that arrives at the surface of a neuron. It joins an ion channel it generate a conformational change that let sodium cations travel to the interior of the cell. This change alters the gradient of concentration at both sides of the cell generating a difference of potential. This difference of potential open potassium channel. This induces a change of the concentration gradient that counterbalances the difference of potential. However for 0.5 ms the membrane is polarized and the electric current travels through the axon. At the end of this structure the change of potential let neurotransmitters exit the neuron. So, information is transferred to another neuron. The advantage of this process is that not every neuron recognize all neurotransmitters, so, only certain connections can be established.

This process was modeled in great detail by Hodgkin and Huxley [2]. However, to describe macroscopic phenomena we can use the simplified model defined by McCulloch and Pitts [3]. Hopfield already stated that the nature of the agents is secondary to describe the emergent behaviour, interaction is the key ingredient [4]. In particular we will only be concerned about a neuron being spiking or not.

$$s_i^t = \begin{cases} 1 & , \\ 0 & . \end{cases} \tag{1}$$

$$\sigma_i^t = \begin{cases} 1 & , \\ -1 & . \end{cases} \tag{2}$$

s (or σ) will take the value 1 is the neuron is firing and 0 otherwise (or -1 in the case of σ). To describe the dynamics of the process we will take a initial state of a set of neurons and a evolution operator ($\chi(t)$). The post-synaptic potential induced by neuron j in neuron i can be described by the following expression:

$$v_{ij} = J_{ij} \int d\tau \chi S_j(t - \tau_{ij} - \tau). \tag{3}$$

J_{ij} describes the amplitude of the effect in i due to the action of j, τ is the lag caused by processes inside the neuron and τ_{ij} is the time information takes to travel from neuron i to neuron j. Previous expression can be simplified, resulting:

$$v_{ij} = J_{ij} S_j(t - \tau_{ij}). \tag{4}$$

However, neural system have a stochastic nature that is not currently present in the system. To introduce this ingredient we will assume that the system will behave as before with a given probability p (experimental studies show that it takes a value close to 0.6[5]) , otherwise no event takes place:

$$v_{ij} = \begin{cases} J_{ij} S_j(t - \tau_{ij}) & p < 0.6 \\ 0 & p > 0.6 \end{cases} \tag{5}$$

Lastly the potential will be compared with the threshold, so, if it superior to this value neuron will fire:

$$v_{ij} = \begin{cases} 1 & si\, t - t_r < \tau_r\ o\, v_{ij} > \theta_i \\ 0 & si\, t - t_r > \tau_r\ o\, v_{ij} < \theta_i \end{cases} \tag{6}$$

1.2 Psychological inspiration.

Behaviourism assumes that all responses observed in animals are either a results of external stimuli or a result of the subject pass. This link between stimuli and response is established due to a process called conditioning. If you show an animal two stimulus, one associated to a certain response (no-conditioned stimuli) and a new stimuli that is not linked to any response (conditioned stimuli), after having shown the two stimuli together a number of times the subject will response to the conditioned stimuli as it does to the no-conditioned one. This variation is called classic conditioning. There are more complex processes where several conditioned stimuli are shown at the same time, it is called parallel conditioning.

To model the relation among stimulus and response we can use the following expression:

$$S^r = H(\sum_k J^k S^k - \theta) \tag{7}$$

Where H is Heaviside function, J shows the influence of the stimulus in the response, S^r is 1 if the stimulus is present and 0 otherwise, θ is a threshold. In the case of classic conditioning only two stimuli are found:

$$S^r = H(J^i \cdot S^i + J^C S^C - \theta) \tag{8}$$

Where J^i represents the influenced of the not-conditioned stimulus in the response triggered by itself and J^c quantifies the influence of the conditioned stimulus in the response of the conditioned stimulus. Hebb identified S^r and S^c with the states of two neurons, so, J is identified with the coupling J_{ij}. In a classic conditioning process the first magnitude will remain constant while the second one will grow with time. The variation can be described by the following formula:

$$\Delta J^C = \varepsilon S^r S^c \tag{9}$$

For a system of neurons synaptic weight are assumed to change in the following manner:

$$\Delta J_{ij} = \varepsilon J_i J_j \tag{10}$$

The hypothesis of Hebb is frequency condensed in the phrase "neurons that fire together, wire together", synapses coupling is strengthen after each spike. He proposed that synapses is the mechanism that enables memory. Although many models back this idea no experimental data has been able to prove the hypothesis [6].

2 Hopfield model.

The model developed in 1982 by Hopfield is a standard in the field. Its simplicity and versatility turn it a perfect choice to serve as a base for more complex approximations. He understood that neural network behave as a physical system where energy is minimized. The Hamiltonian proposed is:

$$H = -\frac{1}{2}\sum_{i \neq j} J_{ij}\sigma_i\sigma_j \tag{11}$$

Where J_{ij} describe the connection between neuron i and j ; and σ_i σ_j is the state of neuron i and j. The state of minimum energy is achieved when all connected neurons are in the same state. In Hopfield network input and output network coincide, it is a single layer network with no hidden nodes.

To study the evolution of the system we will start with this algorithm:

- A neuron is chosen at random
- The variation of energy (ΔE) between the current system and the system will the state of the neuron changed is calculated.
- If ΔE is negative the state of the neuron is changed.
- Go back to the first step

When only one neuron is changed at each step the algorithm is called asynchronous and when all neurons are updated each step the algorithms are synchronous. The first type of algorithm is encouraged as they have computational advantages and they introduce delays among neurons.

One of the most important variations of Hopfield's model is the Boltzmann machine. This algorithm was proposed by Ackley Hinton and Senojwki in 1985[7]. It introduces two differences with respect to Hopfield's: i) the dynamic is stochastic ii)Hidden layers can appear. From an evolutionary viewpoint introducing noise in the dynamics is highly beneficial as it prevents the systems to be trapped in a local minimum. To model the evolution of the system we use Metropolis algorithm [8]. This algorithm can be summarized as:

- A neuron is chosen at random
- We calculate the transition probability ω between the states that originate with the chosen neuron at its current state and the state that originates when the state of the neuron is changed
- From an uniform distribution defined between [0,1] we generate a random number (p). If $\omega > p$ the state of the state of the neuron is changed.
- Go back to first step

To close the algorithm we should define the transition probability. Prior to defining this value we should specify the probability of finding the system in a particular state. As temperature is fixed the probability of finding a state will be dropped from a Gibbs distribution [9]:

$$P = \frac{1}{Z} e^{-\frac{E}{kT}}, \tag{12}$$

The transition probability is described by the following expression

$$\omega = e^{-\frac{\Delta E}{kT}} \tag{13}$$

This transition probability satisfies local detailed balanced, so the dynamics of the system remains invariant under a temporal inversion and the arrival to the stationary state is satisfied. When $\Delta E > 0$ $\omega > 1$, so the state of the neuron is changed with independence of the value of p.

2.1 Associative memory.

The interesting property of neural network is that it can remember a set of states, so once the Hamiltonian is set the system will converge to one of the desired patterns. The pattern can be defined by setting the constants J_{ij}. To define its value I will start by defining the overlap (M), a parameter that quantifies the similitude of a pattern $(\sigma_1, ...\sigma_N)$ and the desired equilibrium state of the system $(\xi_1, ...\xi_N)$:

$$M^\mu = \frac{1}{N} \sum_{i}^{N} \xi_i^\mu \sigma_i \tag{14}$$

$\frac{1}{N}$ is a normalization constant used to ensure that M^μ is found in the interval [-1, 1]. This parameter is the basis of a Hamiltonian proposed:

$$H(I) = -\frac{1}{2N} \sum_{\mu=1}^{P} (M^\mu)^2 = -\frac{1}{2N} \sum \sum (\sum_{\mu} \xi_i^\mu \xi_j^\mu) \sigma_i \sigma_j \tag{15}$$

where P is the number of patterns introduced in the system. From these equations it can be obtained that J_{ji} takes the following value:

$$J_{ji} = \frac{1}{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \tag{16}$$

To avoid convergence problems we set $J_{ii} = 0$. As it occurred in the previous model, $J_{ij} = 0$ are symmetric under the exchange of j by i. The system will acquire the state of minimum energy when overlap is minimum, i.e. $M^{mu} = -\frac{1}{N}$. Up to this point we have studied stable points, up to this point stability is not ensured. To study stability we make use of the following parameter:

$$x_i(I^{\mu 0}) = \xi_i^{\mu_0} h_i(I^{\mu_0}) > 0 \tag{17}$$

When $x > 0$ the steady state is stable. In the case that an external is applied, understand as a perturbation.

$$x_i(I^{\mu_0}) = 1 + \frac{1}{N} \sum_{\mu \neq \mu_0}^{P} \sum_{j}^{N} \xi_i^{\mu} \xi_i^{\mu_0} \xi_j^{\mu} \xi_j^{\mu_0} \tag{18}$$

The first term is the signal and the second term is noise. The pattern remains stable when noise is inferior to 1. To achieve maximum stability the second term should vanish, patterns that fulfill this condition are called orthogonal. At first place we will take completely random pattern, so the probability that a neuron takes a given state can be described as:

$$P(\xi_i^{\mu}) = \frac{1}{2}(\delta(\xi_i^{\mu} - 1) + \delta(\xi_i^{\mu} + 1)) \tag{19}$$

We arrive at a problem that is equivalent to a random walk, in each step the walker should choose whether to change the state of a neuron or not. The properties of this types of processes are well known [10]. In particular, for $t \to \infty$ fluctuations become Gaussian and dispersion converges to $\sqrt{NP}$. So, the process will be stable when:

$$1 \geq \frac{\sqrt{NP}}{N} \tag{20}$$

Equation 19 assumes that the average of spins is 0. However, this property is not always valid. In neural systems activity rate are considerably smaller than 0.5. Besides, when Hopfield model is used for image recognition this hypothesis also fail, background tends to take most of the image so the average is usually smaller than 0. In section 4 calculation will be generalized for these scenarios.

Another limitation to the previous calculation is that the patterns introduced (both in the study of biological systems and in Machine Learning applications) don't have to be orthogonal. To determine the capacity of a network where non-orthogonal are included we take one pattern and invert the state of one of its neurons. The system evolves and we check if the final state matches the initial one. If the initial pattern is recovered the state of other neuron is changed and the same procedure is repeated until the network is no longer able to recover the initial state. At this point:

$$x_r = \xi_i^1 \frac{1}{N} \sum_{j=1}^{N} (\xi_i^1 \xi_j^1) \sigma_j(I^{1,R}) = \left(1 - \frac{2R}{N}\right) \tag{21}$$

The probability of obtaining this state will be given by a Gaussian:

$$P(x_i^n) = \frac{1}{\sqrt{2\pi \langle (x^n)^2 \rangle}} exp[-\frac{(x_i^n)^2}{2\langle (x^n)^2 \rangle}], \tag{22}$$

where variance takes the following value:

$$\langle (x^n)^2 \rangle = \frac{P}{N} = \alpha \tag{23}$$

Now we have to determine the value of x that turns noise bigger than signal. This value depends on the error rate that is considered acceptable, a standard choice is P=0.5 [5] [11]. In this case:

$$\alpha \approx \frac{(1-2R/N)^2}{2\mathrm{log(N)}} \tag{24}$$

Maximum capacity will be the limit of α when $R \to 0$.

$$P_c = \frac{\mathrm{N}}{4\mathrm{logN}} \tag{25}$$

In this calculations we are performing linear approximations however dynamics have not-linear features. In the case that patterns are sufficiently deformed the network won't be able to recover the expected state, even in the case that P were inferior to P_c. In fact, a range of attraction were patterns recover the desired state can be defined. This parameter is called the basin of attraction and in Hopfield model it takes the following value[12]:

$$d = \frac{N}{2P} \tag{26}$$

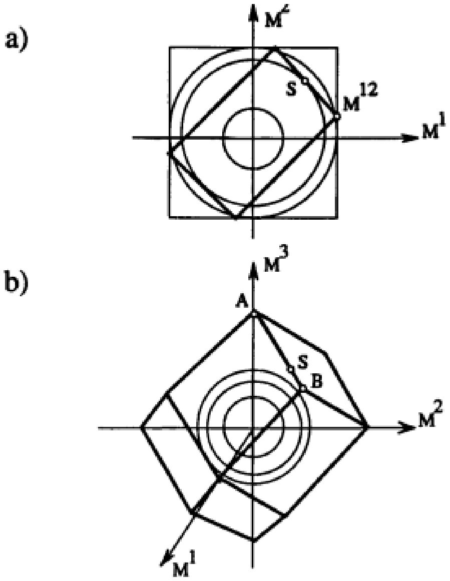


Figure 1: In the image it is shown a representation of the points that satisfy $H = -\frac{N}{2}\sum_{\mu}(M^{\mu}(I))^2 = C$

Another inconvenient that can prevent the system from recovering the expected behaviour are spurious states.Even when all states are orthogonal undesired states may minimize energy. This phenomena can be explained using the visualization of the Hamiltonian shown in fig. 1.

In a Boltzmann machine the number of these states becomes maximum when temperature tends to 0[13], so Hopfield network will always have more spurious states than a Hopfield machine.

3 Boltzmann Network

3.1 Comparative with Ising Model

The analogy of Ising network and Boltzmann machine is clear and it becomes specially important when noise is understand as temperature, the bigger the temperature the stronger the noise. I will describe Ising model to explore the similarities in detail

Ising model is the simplest description of a ferromagnetic material. The system is composed by a square lattice where each site is occupied by a particle described by its spin. The simplest case is that of atoms with spin 1/2, in this scenario atoms can only take two values, spin up or spin down, described by +1 and -1 (for simplicity). The role of spins is equivalent to that of neurons in Hopfield model. In Ising model each spin is under a magnetic field that contains two contributions: i) A magnetic field that comes from external sources h^{ext} ii) a magnetic field induced by other spins. The contribution of each spin in the induced magnetic field is linear. So, the magnetic field experienced by spin i is expressed as follows:

$$h_i = \sum_j \omega_{ij} S_j + h^{ext}, \tag{27}$$

Where ω_{ij} represents the intensity of the interaction among spins. The Hamiltonian results as:

$$H = -\frac{1}{2} \sum_i h_i S_i = \frac{1}{2} \sum_{ij} \omega_{ij} S_i S_j - h^{ext} \sum_j S_j \tag{28}$$

ω_{ij} play the role of synaptic weights and h^{exp} acts as a potential threshold. When temperature is sufficiently high thermal fluctuations can prevent the system from reaching the minimum energy state. Entropy forces compete with electromagnetic forces and the equilibrium state is achieved at a compromise between these two effects. To include fluctuations in the system we can use the strategy introduced by Glauber:

$$S_i = \begin{cases} 1 & \text{with probability} g(h_i) \\ -1 & \text{with probability} 1 - g(h_i) \end{cases} \tag{29}$$

where

$$g(h_i) = \frac{1}{1 + e^{-2\beta h_i}} \tag{30}$$

3.2 Mean field solution

The system cannot be solved in general. However, for many cases the mean field solution may be sufficient. This approximation assumes that the field h_i can be substituted by its mean value,

$$\langle h_i \rangle = \sum_j \omega_{ij} \langle S_i \rangle + \langle h^{ext} \rangle \tag{31}$$

To obtain $\langle h_i \rangle$ we start with the partition function

$$Z = Tr_S exp(-\beta H(S_i)) \tag{32}$$

Where Tr_S is the sum over all possible states. By rearranging terms H_0 results:

$$H_0 = -\frac{1}{2N} \sum_{\mu=1}^{p} (\sum_i S_i \xi_i^\mu)^2 + \frac{p}{2} \tag{33}$$

We define $\alpha = \frac{p}{N}$ as the percentage of patterns we want to store in the network. Then we take the thermodynamic limit, namely $N \to \infty$.
By adding an extra term to the Hamiltonian we arrive at:

$$H = H_0 - \sum_\mu h^\mu \sum_i \xi_i^\mu S_i \tag{34}$$

We will later turn $h^\mu = 0$, but this extra parameter is necessary for the calculation. We introduce the Hamiltonian in the partition function:

$$Z = e^{\frac{-\beta p}{2}} Tr_S exp\left[\frac{\beta}{2N} \sum_\mu (\sum_i S_i \xi_i^\mu)^2 + \beta \sum_\mu h^\mu \sum_i \xi_i^\mu S_i\right] \tag{35}$$

The summation can be approximated by an integral. For convenience we add the auxiliary variable p and rename $a = \frac{2\beta}{N}$ and $b^\mu = \beta \sum_i S_i \xi_i^\mu$. By introducing these changes we arrive at:

$$Z = e^{\frac{-\beta p}{2}} (\frac{\beta N}{2\pi})^{\frac{p}{2}} Tr_S \prod_\mu \int dm^\mu e^{-\frac{1}{2}\beta N (m^\mu)^2 + \beta(m^\mu + h^\mu)\sum_i \xi_i^\mu}. \tag{36}$$

Then we use the vectorial form and make use of the identity $2cosh(x) = e^x + e^{-x}$, so:

$$Z = (\frac{\beta N}{2\pi})^{\frac{p}{2}} \int dm e^{-\beta N f(\beta, m)} \tag{37}$$

where

$$f(\beta, m) = \frac{1}{2}\alpha^2 + \frac{1}{2}m^2 - \frac{1}{\beta N} \sum_i log(2cosh(\beta(m+h)xi_i)) \tag{38}$$

m, h, ξ_i are vectors. If we take the limit $N \to \infty$ we can make use of the saddle point method (it is explained in appendix A). For eq. 27 we obtain:

$$-\frac{1}{N} \log Z = \beta min f(\beta, m) \tag{39}$$

As $F = -\frac{1}{b} \log(Z)$ is free energy, we know that:

$$\frac{F}{N} = min f(\beta, m) \tag{40}$$

so we recover the minimum value of free energy normalized by the number of unit. Minimizing f.

$$0 = \frac{\partial f}{\partial m^\mu} = m^\mu - \frac{1}{N}\sum_i \xi_i^\mu \tanh(\beta(m+h)\cdot\xi_i) \tag{41}$$

We arrive at a system of p not linear equations with p unknown, but the system is self-averaged, so we get at:

$$m^\mu = \langle\langle \xi_i^\mu \tanh(\beta(m+h)\cdot\xi_i)\rangle\rangle \tag{42}$$

where $\langle\langle\cdots\rangle\rangle$ means average over a random distribution of patterns. Similarly:

$$f(\beta, m) = +\frac{1}{2}m^2 - \frac{1}{\beta}\langle\langle \log(2\cosh(\beta(m+h)\xi_i))\rangle\rangle \tag{43}$$

We can extract a physical meaning out of the state of minimum energy. If we derivative Gibbs free energy F with respect to h^μ:

$$\frac{\partial F}{\partial h^\mu} = -\sum_i \langle S_i\rangle \xi_i^\mu \tag{44}$$

or equivalently

$$\frac{\partial F}{\partial h^\mu} = N\frac{\partial f}{\partial h^\mu} = -N\langle\langle \xi_i^\mu \tanh(\beta(m+h)\cdot\xi_i)\rangle\rangle = -Nm^\mu \tag{45}$$

So

$$m^\mu = c1N\sum_i \xi_i^\mu \langle S_i\rangle \tag{46}$$

The minimum energy state is the average of the overlap of our configuration with the pattern μ To arrive at previous expression we have made use of h^μ. However, this parameter is no longer needed, so we can make it equal 0. Therefore, we arrive at the following expression:

$$m^\mu = \langle\langle \xi_i^\mu \tanh(\beta(m)\cdot\xi_i)\rangle\rangle \tag{47}$$

Previous equations has a large number of solutions. The simplest and most important ones are the the state where the system collapses into one of the initial patterns, namely $m = (s, 0, 0, 0, 0, ...)$. If we introduce this proposal in the equation obtained we arrive at:

$$s = \tanh(\beta s) \tag{48}$$

This equation is identical to the one found in Ising model. It implies that for $T < 1$ the state of memory is stable. The fraction overlap obtained is shown in image 3.2.
The equation obtained only have a non trivial solution at $\beta = \frac{1}{T} > 1$.
There are more complex states that can fulfil the required condition. These are the spurious states that have been mentioned before and in the mean field model they appear as a combination of various patterns. The simplest one take the following expression:

$$m = (s, s, s, s, 0, ...) \tag{49}$$

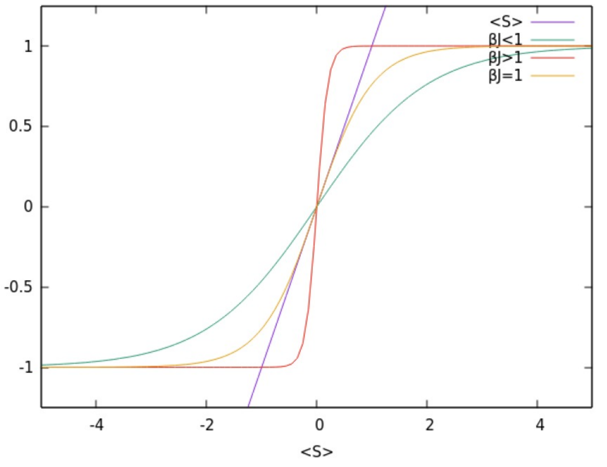


Figure 2:

with n not null input, where all them take the value s. If we introduce this solution in the mean field equation

$$m^\mu = \langle\langle \xi_i^\mu \tanh(\beta s \sum_{i=1}^{n} \xi_i)\rangle\rangle \tag{50}$$

we observed that this expression vanishes when $\mu > n$ as $\langle\langle \xi_{mu} \xi_\mu \rangle\rangle = 0$ for $\mu \neq \nu$, additionally

$$s = \frac{\langle\langle z \tanh \beta s z \rangle\rangle}{n} \tag{51}$$

where z is a random variable:

$$z = \sum_{i=1}^{n} \xi^\mu \tag{52}$$

that follows a binomial distribution. This system has solutions for every n if $T < 1$.

However, not all solutions are stable. To check if m produces a stable state we need the eigenvalues of the matrix A to be greater than zeros:

$$A_{\mu\nu} = \frac{\partial^2 f}{\partial m^\mu \partial m^\nu} \tag{53}$$

This requirement is only fulfill when n is odd and $T < 1$.

4 Improvements to Hopfield model

Previous calculations assumes that both states of the neurons are equal likely. However, this approximation is rarely observed. In biological neural networks the spiking rate is considerably smaller than 50%[14]. Besides, image rarely have 50 % of pixels in color and 50% in white. To increase performance we can introduce a change in the system, so, orthogonal patterns can have an average of a. In this case equation 19 turns into:

$$P(\xi) = \frac{1}{2}(1+a)\delta(\xi-1) + \frac{1}{2}(1-a)\delta(\xi+1) \tag{54}$$

If equation 14 is used to calculate the overlap it takes the value a.

$$\langle\langle \xi_i^\mu \xi_i^\nu \rangle\rangle = \delta^{\mu\nu} + a^2(1-\delta^{\mu\nu}) \tag{55}$$

Where $\langle\langle\ \rangle\rangle$ shows the average over all possible patterns. In the language of Ising model it means that the system has an intrinsic magnetization, this effect can be modelled as an external field (equivalent to an external magnetic field in Ising model). Having all this results in consideration a new Hamiltonian can be defined:

$$H = \frac{-1}{2N}\sum_{ij} J_{ij}s_is_j + \theta\sum_i s_is_j \tag{56}$$

$\theta = \sum_{j\neq i} J_{ij}$ is the effect of the previously mentioned magnetic field. Synaptic weights are modified, so now they take the following expression:

$$J_{ji} = \frac{1}{N}\sum_{\mu=1}^{P}(\xi_i^\mu - a)(\xi_j^\mu - a) \tag{57}$$

$a = \sum_{i=1}^{N}\sum_{\mu=1}^{P}\xi_i^\mu$. In the case a=0 we recover the original expression. Now, as $a \neq 0$, noise is no longer centered at 0 so the probability of a pattern being destabilized is bigger, magnetization among patterns should be null and this can only be modified by redefining magnetization:

$$m^\mu = \frac{1}{N}\sum_i \langle(\xi_i^\mu - a)S_i\rangle \tag{58}$$

$$m^\mu = m_\nu\delta^{\mu\nu} \tag{59}$$

The capacity now results:

$$P_c = \frac{N}{2a|\ln(a)|} \tag{60}$$

As it occurs for orthogonal patterns the value of P_c for a fix value of a is proportional to N. However, the system contains an additional improvement. For $a > 0.99$ the value of P_c is bigger than the one obtained for orthogonal patterns, this result is extremely helpful. As it occurs for in the case of $a = 0$ the behaviour of the network depends on the temperature. In specific a phase transition at T_c is observed. Now T_c takes the following value:

$$T_c(p) \approx p\frac{(1-\theta/p)}{|\ln(1-\theta/p)|} \tag{61}$$

However, there is an extra constrain that we are not considering. When the total number of neurons spiking and in repose are equal in all parameters only a the following number of pattern can be established:

$$\binom{N}{Na} \tag{62}$$

This number grows significantly faster than the capacity, so, it does impose any extra limitation to the capacity of the network.

In some cases using neurons that take the values 1,0 (in stead of 1,-1) can be useful. The Hamiltonian proposed in expression 11 can be adapted to the new scenario. To do so, we should apply a new normalization that counterbalances the effects produced by the change in the type of neuron. This model is described by following equations:

$$\begin{aligned}
H(S) &= -\frac{1}{2}\sum_{i,j} J_{i,j}s_is_j + \sum_i \theta_i s_i \\
J_{ij} &= \frac{1}{a(1-a)N}\sum_{\mu=1}^{P}(\xi_i^\mu - a)(\xi_j^\mu - a) \\
\theta &= \frac{1}{2}\sum_j J_{i,j}
\end{aligned} \tag{63}$$

Magnetization should be substituted by next equation:

$$m^\mu(s) = \frac{1}{a(1-a)N}\sum_i(\xi_i^\mu - a)(s_i - \frac{1}{2}) \tag{64}$$

The only difference between this expression and the ones previously used is the factor 1/2 and the normalization constant. However, both provide the same information about the system. This modification increases the capacity of the network but it also produces more spurious states. When a grows this latter effect loses importance, but the probability of recovering a combination of various patterns remains high[14].

5 Learning.

The algorithms developed so far only recognizes patterns that have been introduced in the Hamiltonian. The next step is to create a network that can learn new patterns. To achieve this goal J_{ij} should be updated. The simplest model that can make the network learn was proposed by Gelperin and Hopfield [15]. They propose a algorithm that mimic classic conditioning. The initial value of J_{ij} is given by Hopfield model and this values are updated using Hebbian rules:

$$\Delta J_{ij} = \varepsilon \sigma_i \sigma_{j-} \qquad \varepsilon > 0 \tag{65}$$

The theory does not fix a value of ε. Its value will depend on the problem, a value of ε will make J evolve fast but it can lead to numerical errors as the system will be unable to detect many aspects of the problem. In the other hand if is small the system will learn slow, so, it may not be able to achieve the desired state. To obtain valid results we will impose the following conditions:

1. Weak restrictions: J_{ij} should be normalized to L

   $$\sum_j |J_{ij}|^2 = L \tag{66}$$

   Where L is a constant that can take any value , however a standard practice is to take L=N. This condition is called strict weak condition. However, it does not allow a change of the values of J, so, to make use of Hebbian rule we should substitute it by the so called "loose" condition

   $$\sum_j |J_{ij}|^2 < L \tag{67}$$

2. Strong condition: when ε is small all the parameters of the system have importance, so, patterns stop being stable fast. Using the stability condition described in eq. 17 we arrive at:

   $$\varepsilon > \sqrt{\frac{\langle \delta J^2 \rangle}{N}} \tag{68}$$

   Best performance is achieved when ε takes the minimum value compatible with the strong condition. The optimum value is:

   $$\varepsilon = \frac{eL}{N} \tag{69}$$

   Where e is a constant that depends on the algorithm used. For reversible processes e=3 and for irreversible processes e=3.33

To close the problem we have defined two simple algorithms that let us observe the evolution of synaptic weights. The first algorithms let the network evolve without any restriction until the weight satisfy the weak restriction. At this point the value of the weight becomes frozen until the learning rule make J_{ij} satisfy the weak condition again. The effects observed in this algorithm resemble those of long term memory. As, patterns a subject is first exposed to are better remained in memory[16]. An alternative is to let the system explore the forbidden regions. This model forgets the first patterns when it learns new ones, the process is similar to short-term memory. Short term memory can remember up to 7 patterns, an artificial neural network needs at least 500 neurons. This number matches the amount of neurons in a microcolumns. Although this may lead to believe that short term memory is a function that is carried out in microcolumns recent studies has shown that it is a highly complex process where many different areas in the brain are involved. However, sometimes patterns are stored in specific areas.

5.1 Widrow-Hoff algorithm

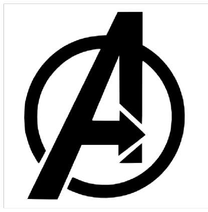
To increase learning capacity Widrow-Hoff developed an algorithm that managed to reduce the effect of noise. To do so he included an extra term in Hopfield Hatmiltonian

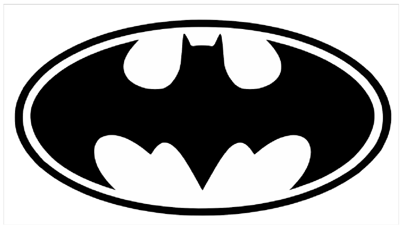
$$J_{ij}(\mu) = J_{ij}(\mu - 1) + \frac{1}{N}\xi_i^\mu \xi_j^\mu - \frac{1}{N}\sum_k J_{ik}(\mu - 1)\xi_k^\mu \xi_j^\mu \tag{70}$$

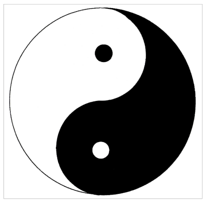
This rule mimics classic learning where external stimuli are silenced. Although this algorithm offers better performance for machine learning problems it is too artificial to be used to describe real learning processes. [17] [18] [19] [5] [11]

6 Results

To obtain the following results I have made used of two types of patterns: random and manually chosen. The following images have been manually included:







We have picked icons as they are simple and they can be described with a small number of neurons. In particular we have made use of a network with 196 neurons. The mean value of neurons for each of these four cases are:

a1=0,265
a2=0,565
a3=-0,255
a4=0,32

The average of the four images is 0.22. These four patterns are distorted to check if the neural network can return them. To establish a well defined criteria a pattern is assumed to be recovered when the overlap overlap is superior to 90%. To distort the images I use a homemade code that changes the state of a neuron with a probability given by the user. We start by changing the state of the neurons with a probability of 5% and we increase the probability by 5 % until we reach a maximum value of 65%.
The network defined in section 4 is able to recognize the image even when the percentage of deformation is 65%. However, the basic network worked correctly in a smaller region, it turns unable to recognize images with a percentage of change superior to 40%. The

percentage of deformation is not the only important parameter the overlap among images is also crucial to obtain bigger or smaller basin of attraction.
Another interesting parameter is the number of iterations the network needs to recover a given pattern. The system converges fast, in less than 1000 Montecarlo steps the system has already reached its steady state. To check the stationary of the state we have made use of two parameters: magnetization and energy. However, both parameters stabilize at the same time. The evolution is shown in figure 3.

6.1 Capacity

To further characterize the network we are going to use capacity. We take random pattern and we add new ones until the system can no longer learn them. The patterns are generated using a random number that describes the probability of a pixel being white or black. This allow us to fix the mean value.
We introduce images in batches of 5 until we reach a maximum of 70 patterns. After a new batch is introduced we check the number of images the network can remember. When the system has stored less than 50 images it can remember all the images. Once it reaches this point the number of images that can be remember by the network diminishes. The capacity obtained is superior to the one defined analytically, which is of 17 patterns.
Spurious patterns have not played a significant role in the dynamics. This may be caused by the reduced size of the network. The number of spurious states grow exponentially with the number of neurons. [20] [14]

6.2 Effect of temperature

When the system is in the absolute zero the system can only evolve towards a state with smaller energy, so as shown in figure 5, the state reaches the state of minimum energy or equivalently maximum magnetization. When temperature increases levels that were previously forbidden become available, so, evolution turns more erratic. Energy won't follow a increasing state, in fact, as we observe energy decreases. States with a smaller value of magnetization are encouraged from an entropic point of view, so, when temperature increases they became the most likely steady states. The evolution of magnetization with temperature is observed in image 6. We can see that there are discrepancies between the simulations and the theory. Three possible sources of error have been identified a) Simulations may not have reached the steady state, to make sure that this is not the source of error we repeated the simulation with ten times more montecarlo steps and we did not observe any discrepancy b) Surface error: in the calculation we have assumed mean field approximation is valid, so surface effects are neglected. When the transition is studied using Landau- Ginzburg theory a extra term appears, however this may play a more relevant role in smaller networks, in the case of the study this coefficient should be smaller than 0.05. [21]. So it is not sufficiently big to explain the discrepancy observed c) Network topology: it has been observed that

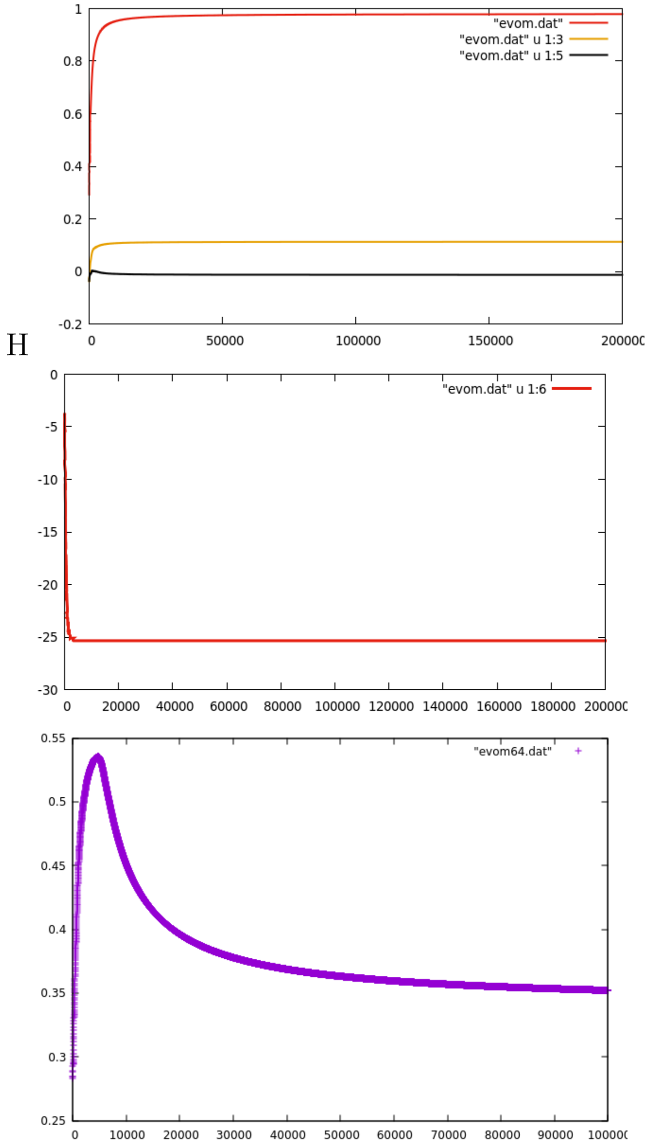


Figure 3: Top: evolution of the overlap when the input pattern is one of desired states with 35% of the pixels changed . Center : evolution of the energy in the previous condition Bottom: evolution of the magnetization in a network with 64 patterns saved. The input pattern corresponds to one of the 64 states with 10 % of the pixels changed. In x-axis number of Montecarlo steps is shown and in y-axis overlap is plotted

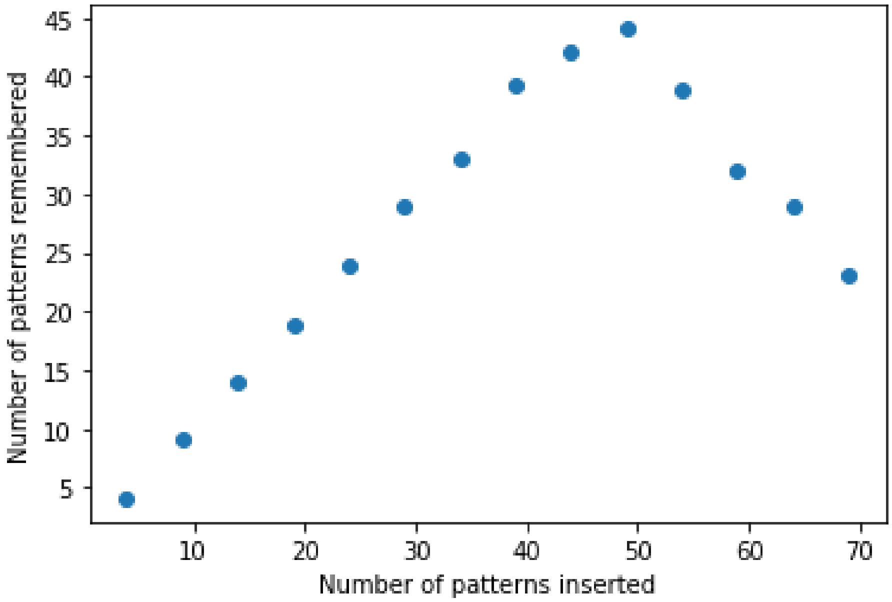


Figure 4: In the x axis the number of patterns is introduced. In y-axis we plot the number of remembered patterns

differences in the shape of connections can lead to an error of even 20% for network of the size used. So, this may be main cause of error.

Behaviour around the critical point follows a power law. To check this relation we perform a linear regression. We observe a linear behaviour in a double logarithm scale with a $r^2 = 0.977$. Although, this coefficient is low to provide an accurate fix we recover an exponent with only 10% of error. Discrepancies are caused by finite size effects. Calculations have been repeated with a network where only one parameter has been stored. The mean value of the pixels pattern is 0.1 and 0.9. In these cases the behaviour is slightly more linear but we cannot recover the value of the exponent or the critical temperature with greater accuracy. So, the main source of error are finite size effects.

Lastly we have check the behaviour of the improved Hopfield model, in this case the interval of temperature where the network can recover information is even smaller. Critical temperature is reached for a value of 0.25.

$$\log(M) = (-0.55 \pm 0.03)\log(T) + (-0.804 \pm 0.022) \tag{71}$$

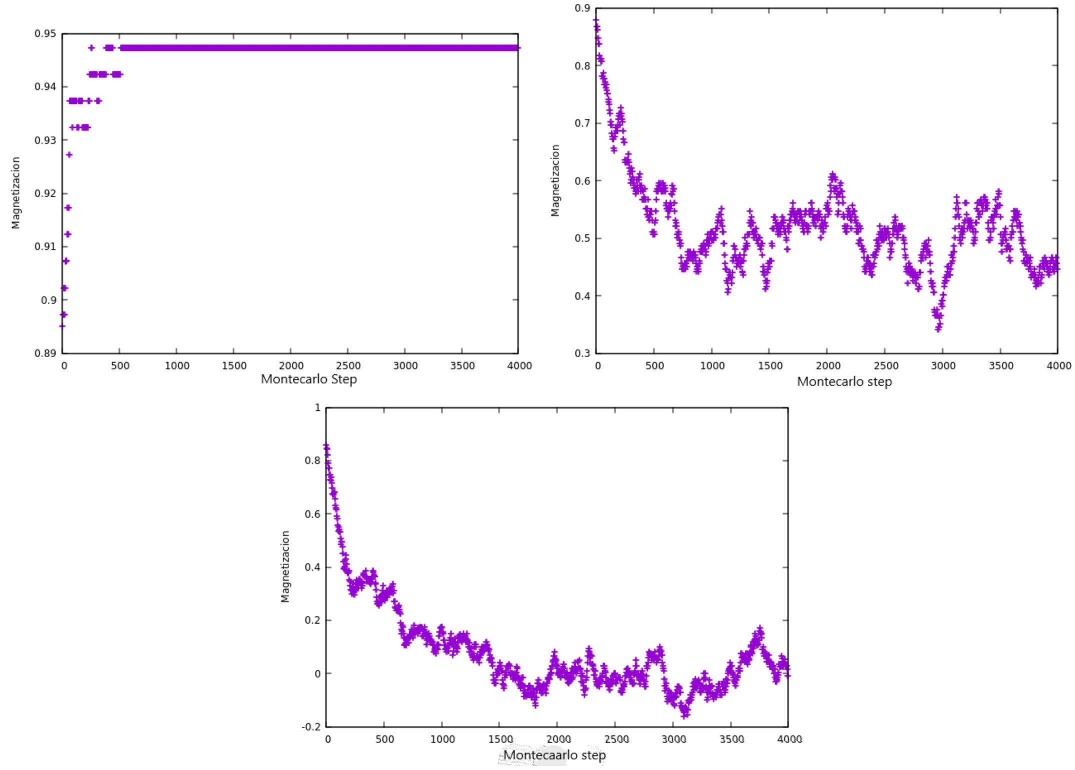


Figure 5: In x-axis the number of steps is plotted, in y-axis magnetization is shown. The upper left image corresponds to a simulation with a temperature of 0.01, the upper right image corresponds to a simulation with a temperature of 0.7. The image in the bottom corresponds to a simulation with a temperature of 1.2

7 Conclusions

In the present job we have studied Hopfield's model and two of its variants, the one proposed by Amit and the Boltzmann Machine. We first motivate the study of this model and continued by describing the model and some of its key properties. Then, we could improve the description of the system by including noise. We characterized the phase transition observed when temperature turns the patterns introduced unstable. We end the bibliographic section by describing an improved model and some basic learning algorithms .
In the second part of the work we check the performance of a Hopfield network. We observed that the basin of attraction of the model, specially of the improved version, were considerably big. We could recover images even when the state of the 65% of its neurons where changed. We calculated the maximum number of pattern that a network can store and we end the work by characterizing the phase transition. In this section we observed that for a network with 196 neurons the mean field approximation introduces a significant error in the results obtained.

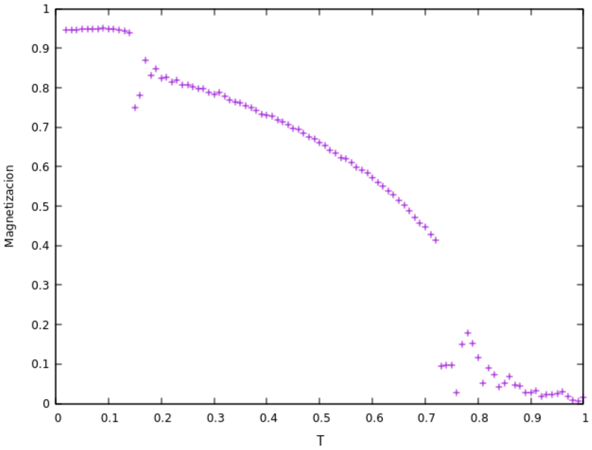
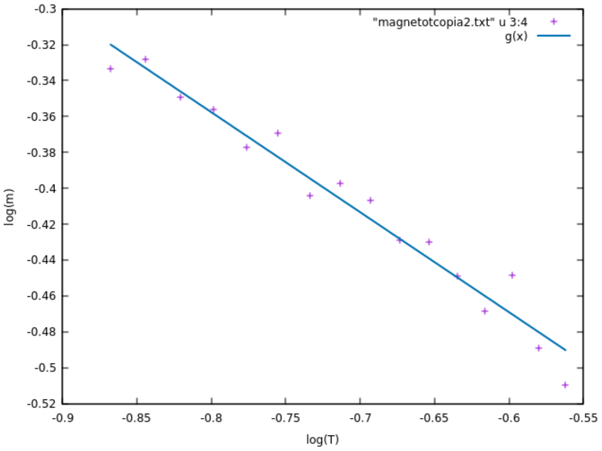


Figure 6: In x-axis temperature is plotted, in y-axis magnetization is shown



A Saddle point method

Saddle point method will be explained with an example. Assume an integral with the following expression:

$$I = \sqrt{N} \int dx e^{-Ng(x)}. \tag{72}$$

If we span g around its minimum and we only take the first not-null coefficient we arrive at:

$$I = \sqrt{N} \int dx exp(g(x_0) + \frac{1}{2} g''(x_0)(x - x_0)^2). \tag{73}$$

This is a Gaussian integral

$$I = \sqrt{N} e^{Ng(x_0)} \sqrt{\frac{2\pi}{Ng''(x_0)}} = e^{Ng(x_0)} \sqrt{\frac{2\pi}{g''(x_0)}} \tag{74}$$

Taking logarithms at both sides:

$$-\frac{1}{N}\log(I) = g(x_0) + \frac{1}{2N}(log(\sqrt{\frac{2\pi}{g''(x_0)}})) = g(x_0) \tag{75}$$

When N tends to infinity. So we simply have to find x_0.

References

[1] M. F. Casanova and I. Opris, *Recent Advances in the Modular structure of the cortex.* Springer, 2015.

[2] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of physiology*, vol. 117, no. 4, p. 500, 1952.

[3] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[4] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[5] P. Peretto and P. Pierre, *An introduction to the modeling of neural networks*, vol. 2. Cambridge University Press, 1992.

[6] C. R. Gallistel and L. D. Matzel, "The neuroscience of learning: beyond the hebbian synapse," *Annual review of psychology*, vol. 64, pp. 169–200, 2013.

[7] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for boltzmann machines," *Cognitive science*, vol. 9, no. 1, pp. 147–169, 1985.

[8] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The journal of chemical physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[9] R. K. Pathria, *Statistical mechanics.* Elsevier, 2016.

[10] C. W. Gardiner *et al.*, *Handbook of stochastic methods*, vol. 3. springer Berlin, 1985.

[11] J. A. Hertz, *Introduction to the theory of neural computation.* CRC Press, 2018.

[12] Y. Baram, "Orthogonal patterns in binary neural networks," 1988.

[13] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Storing infinite numbers of patterns in a spin-glass model of neural networks," *Physical Review Letters*, vol. 55, no. 14, p. 1530, 1985.

[14] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Information storage in neural networks with low levels of activity," *Physical Review A*, vol. 35, no. 5, p. 2293, 1987.

[15] A. Gelperin, J. Hopfield, and D. Tank, "The logic of limax learning," in *Model neural networks and behavior*, pp. 237–261, Springer, 1985.

[16] B. B. Murdock Jr, “The serial position effect of free recall.,” *Journal of experimental psychology*, vol. 64, no. 5, p. 482, 1962.

[17] D. Jaeger and R. Jung, *Encyclopedia of computational neuroscience*. 2015.

[18] H. Sompolinsky, “Statistical mechanics of neural networks,” *Physics Today*, vol. 41, no. 12, pp. 70–80, 1988.

[19] M. Mézard, J. Nadal, and G. Toulouse, “Solvable models of working memories,” *Journal de physique*, vol. 47, no. 9, pp. 1457–1462, 1986.

[20] M. V. Tsodyks and M. V. Feigel’man, “The enhanced storage capacity in neural networks with low activity level,” *EPL (Europhysics Letters)*, vol. 6, no. 2, p. 101, 1988.

[21] J. Rudnick, H. Guo, and D. Jasnow, “Finite-size scaling and the renormalization group,” in *Current Physics–Sources and Comments*, vol. 2, pp. 47–67, Elsevier, 1988.