RIGOROUS RESULTS ON THE HOPFIELD MODEL OF NEURAL NETWORKS[1]

Anton Bovier[2]
Institut für Angewandte Analysis und Stochastik
Mohrenstrasse 39, D-10117 Berlin, Germany

and

Véronique Gayrard[2]
Centre de Physique Théorique, CNRS
Luminy, Case 907, F-13288 Marseille Cedex 9, France

Abstract: We review some recent rigorous results in the theory of neural networks, and in particular on the thermodynamic properties of the Hopfield model. In this context, the model is treated as a Curie-Weiss model with random interactions and large deviation techniques are applied. The tractability of the random interactions depends strongly on how the number, M, of stored patterns scales with the size, N, of the system. We present an exact analysis of the thermodynamic limit under the sole condition that $M/N \downarrow 0$, as $N \uparrow \infty$, i.e. we prove the almost sure convergence of the free energy to a non-random limit and the a.s. convergence of the measures induced on the overlap parameters. We also present results on the structure of local minima of the Hopfield Hamiltonian, originally derived by Newman. All these results are extended to the Hopfield model defined on dilute random graphs.

[1] Invited talk presented by A.B. at the 5^o CLAPEM, São Paulo, 1993
[2] Research supported in part by the Commission of the European Communities under contract No. SC1-CT91-0695

I. Introduction

In this lecture we review some results on a disordered mean field spin system that has over the last decade attracted, under the name of "Hopfield model", considerable attention in the context of modelling of cognitive phenomena in neural networks such as the brain, and in particular has been used as the prototypical model for autoassociative memory. Our main point of view here will be, however, that of statistical mechanics of disordered systems and we will only comment on the interpretation of the thermodynamical properties of this model in terms of memory. For a more detailed exposition of these aspects, see e.g. the book by Amit [A].

Let us first describe this model. Let $\mathcal{S}_N = \{-1,1\}^N$ be the space of functions $\sigma : \{1,\dots,N\} \to \{-1,1\}$. We call σ a *spin configuration* on the set $\{1,\dots,N\}$, and $\sigma_i \in \{-1,+1\}$ the spin at (neural state of) the vertex (neuron) i. We shall write $\mathcal{S} \equiv \{-1,1\}^{I\!N}$ for the space of half infinite sequences equipped with the product topology of discrete topology on $\{-1,1\}$. We denote by $\mathcal{B}_N$ and $\mathcal{B}$ the corresponding Borel sigma algebras. We will define a random Hamiltonian function on the spaces $\mathcal{S}_N$ as follows. Let $(\Omega, \mathcal{F}, I\!P)$ be an abstract probability space. Let $\xi \equiv \{\xi_i^\mu\}_{i,\mu\in I\!N}$ be a two-parameter family of independent, identically distributed random variables on this space such that $I\!P(\xi_i^\mu = 1) = I\!P(\xi_i^\mu = -1) = \frac{1}{2}$. For a given non-decreasing integer valued function $M : I\!N \to I\!N$ we denote by $\mathcal{F}_N$ the sub-sigma algebra generated by the random variables $\{\xi_i^\mu\}_{1\leq i\leq N}^{1\leq\mu\leq M(N)}$. We will occasionally denote this sub-family of random variables by $\xi|_N$. A vector $\xi^\mu \equiv \{\xi_i^\mu\}_{i=1,\dots,N}$ is a particular random state of the system (neural network) and often called a 'pattern'. The Hopfield Hamiltonian on $\mathcal{S}_N$ is then given by

$$H_N(\sigma) = -\frac{1}{2N} \sum_{i,j=1}^{N} \sum_{\mu=1}^{M(N)} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j \tag{1.1}$$

The history of this Hamiltonian is quite interesting. The simplest version of it, where $M(N) \equiv 1$, was proposed by Mattis [Ma] as a simple model of a disordered magnet, but it was of course immediately realized that such a model is entirely trivial and differs from a ferromagnet only by a gauge transformation $\sigma_i \to \sigma_i' = \xi_i \sigma_i$ of the spins. Luttinger [Lu] proposed a less trivial variant with $M(N) \equiv 2$ as a model of a spin glass and finally in 1977, Figotin and Pastur [FP1] proposed a fairly large class of models, which included the above with arbitrary, but fixed M, again as soluble model of a spin glass. Their paper, and two follow-ups [FP2, FP3] contain a very nice and detailed analysis, including the quantum and the Kac version of the model. As a spin glass model, this remained, however, somewhat unsatisfactory, and this may be the reason why these papers remained largely ignored (in fact, I only became aware of them when meeting Pastur during this very conference in São Paulo!!!). Five years later, Hopfield [Ho] apparently unaware of this previous work, proposed the above Hamiltonian, this time, however, as a model for autoassociative memory, and notably with $M(N)$ a possibly non-constant function of N. His work was inspired by

Hebb's learning rule [He] and arose from a dynamics point of view (a related model had already been propsed by Little [Li] in 1974). It was however the fact that the thermodynamic properties of his model show an immediate bearing on the memory properties of the associated dynamical model that sparked the interest of the physics and notably the spin-glass community in this new field. The findings of Figotin and Pastur for the case of bounded M were recovered in papers by Amit, Gutfreund and Sompolinski [AGS1] and van Hemmen [vH], and interpreted in the new context as perfect functioning of the memory. However, more interesting phenomena were discovered if M was allowed to grow with N; a seminal paper by Amit, Gutfreund and Sompolinsky [AGS2] analysed this case using the method of replicas and the idea of replica symmetry breaking, developed by Parisi et al. [P] in the study of the Sherrington-Kirkpatrick spin glass model. They discovered that if M was chosen as $M(N) = \alpha N$, several phase transitions occurred as the value of the parameter α increased. In particular, for $\alpha > \alpha_c \approx 0.14$ the model would enter into what they interpreted as a genuine spin glass phase. In terms of memory, this phase was interpreted as a breakdown of memory and the critical parameter α_c is called the memory capacity of the system. These findings were also confirmed by numerous numerical investigations.

The rich structure of this model thus invites a more rigorous mathematical investigation. Here it may be seen as an advantage over, say, the classical spin glass models, that the function $M(N)$ provides a parameter that allows to tune the model from an essentially trivial ferromagnetic situation to complex spin glass like behaviour. In spite of that, progress on the mathematical level has been fairly slow and is still lagging considerably behind the heuristic understanding provided by the 1985 paper of Amit et al. [AGS2]. In this paper we try to summarize part of the few rigorously established results. These results are essentially of two types: One, originally due to Newman [N], concerns the structure of the local minima of the Hamiltonian only and thus is immediately relevant for the a noiseless gradient dynamics. The second concerns the actual thermodynamics at finite temperature and is relevant for noisy dynamics, which will have to be employed if spurious 'false' minima are to be avoided. This turns out, however, to bring about serious new difficulties. The next two sections will be devoted to these two situations, respectively.

Many variants of the classical Hamiltonian (1.1) have been proposed over the years to more appropriately reflect particular model situations. They involve the modification of the state space of a single neuron to accommodate more that two values (Potts-Hopfield model [Ka,ES,FMP,G], modification of the a-priori distribution of the patterns to accommodate asymmetries (biased model [BM]) or correlations, and many others. A variant that we want to highlight here is the so-called dilute model, where not all neurons are interconnected, but where this connection is described by a underlying, pattern-independent (dilute) random graph. While this appears an important point for the modelling of actual neural networks, it turns out that many of the results for fully connected model carry over, suitably modified, to this situation. As we will come back to this model, we give here a precise definition.

The Hamiltonian of the dilute Hopfield model is given by

$$H_N(\sigma) = -\frac{1}{Np}\sum_{\substack{i,j\\ i\neq j}} \epsilon_{ij} \sum_{\mu=1}^{m} \xi_i^\mu \xi_j^\mu \sigma_i\sigma_j \tag{1.2}$$

where $p = I\!E(\epsilon_{ij}) > 0$. For given N, the ϵ_{ij} for $i, j \in \{1, \dots, N\}$ form a family of N^2 independent, identically distributed random variables with common distribution such that $I\!P(\epsilon_{ij} = 1) = 1 - I\!P(\epsilon_{ij} = 0) = p(N)$. The precise dependence of these random variables on N can be set up in various ways (see e.g. [BG2]) but this will not be an issue here. We just notice that these variables of course describe a random graph process with edge density $p(N)$. It will be of interest to allow $p(N)$ to decrease with N and in particular to see how fast it may be allowed to decrease in order to maintain certain properties.

II. Thermodynamics of the Hopfield model

In this chapter we review some of the results on the thermodynamics of the Hopfield model. To do this, let us briefly introduce the thermodynamic formalism.

For $\eta \in I\!N$, we denote by $\mathcal{G}^\eta_{N,\beta,h}$ the random probability measure (finite volume Gibbs measure) on $(\mathcal{S}_N, \mathcal{B}(\mathcal{S}_N))$ that assigns to each $\sigma \in \mathcal{S}_N$ the mass

$$\mathcal{G}^\eta_{N,\beta,h}(\sigma) \equiv \frac{1}{Z^\eta_{N,\beta,h}} e^{-\beta H_N(\sigma) - \beta h \sum_{i\in\Lambda} \xi_i^\eta \sigma_i} \tag{2.1}$$

where $Z^\eta_{N,\beta,h}$ is a normalizing factor usually called *partition function.* The quantity

$$f^\eta_{N,\beta,h} \equiv -\frac{1}{\beta N} \ln Z^\eta_{N,\beta,h} \tag{2.2}$$

is called the free energy. Note that all these quantities are $\mathcal{F}_N$-measurable random variables. The parameter β is the inverse temperature and h is called a *magnetic field* aligned on the pattern ξ^η. The purpose of thermodynamics is to identify and characterize the nature of these measures in the limit as N tends to infinity. In particular, one asks the question what happens with the limiting measures when the parameter h is taken to zero. In the case where this procedure leads to measures depending on the index η and the sign of h, we speak of 'coexistence' of several infinite volume measures for the zero-field model, or of a first order phase transition. If these distinct measures are in one-to-one correspondence with the original patterns, this phenomenon can also be interpreted in the sense of a functioning memory, due to the fact that the Gibbs states furnish in fact the invariant measures for the retrieval dynamics of the memory. This situation is expected to take place at low temperature (and not too large M). Another possibility, expected (and in fact essentially proven by Tirozzi and Scaccitelli [TS]) at high temperature is uniqueness, i.e. the $h \downarrow 0$ limits of all the infinite volume Gibbs measures should coincide. This clearly has the interpretation of no memory. A more interesting breakdown of memory is expected even at low temperature, if the number of patterns grows too rapidly with N: in this situation one may still expect non-uniqueness but no simple correspondence between the set of extremal Gibbs measures and the patterns.

A remark is in order concerning the above definition of the Gibbs measures, and in particular concerning the magnetic field term. In principle, we might just set $h = 0$ for the finite volume measures and consider the limits as $N \uparrow \infty$. If this is done in the usual *ferromagnetic* Curie-Weiss model, one finds a limit which is, depending whether β is larger or smaller than 1, a delta measure concentrated at 0 or a symmetric mixture of delta-measures concentrated at $+a(\beta)$ and $-a(\beta)$ (for some $a(\beta) > 0$, see below), resp., where the latter two represent the actual extremal (clustering) Gibbs measures of the model (which can be obtained as limits with added positive or negative magnetic fields). In the present disordered situation, such a procedure would encounter additional difficulties. Namely, the corresponding sequence of measures would not be expected to converge at all, and only suitably chosen *random* subsequences would converge to specific limit points. One can then ask the question whether all extremal measures can be obtained as limit points in this way, and for the situation we will treat below, there are arguments that make this plausible, although this has not been proven. Adding the magnetic fields as done above is a very convenient tool to circumvent this difficulty. What it does is, in fact, to give infinitely more weight to one specific extremal measure and so to favour convergence to this particular limit, no matter how small h is. However, it should be noticed that in order for such a scheme to work, we need to be able to guess correctly what these extremal measures should look like, which, as we will see works for M not growing too fast. If M grows faster and the system gets into a spin glass phase, no such information is (yet) available, and this makes such a procedure impracticable.

An important observation is that the value of the measure $\mathcal{G}^{\eta}_{N,\beta,h}(\sigma)$ does depend on σ only through the quantities

$$m^{\mu}_N(\sigma) \equiv \frac{1}{N}\sum_{i=1}^{N} \xi^{\mu}_i \sigma_i, \quad \mu = 1, \ldots, M \tag{2.3}$$

called *overlap parameters*, since the Hamiltonian may be written in the form

$$H_N(\sigma) = -N \sum_{\mu=1}^{M} \left(m^{\mu}_N(\sigma)\right)^2 \tag{2.4}$$

This suggests to define the random map

$$\begin{aligned} \mathcal{M}_N : \mathcal{S}_\Lambda &\to I\!R^M \\ \sigma &\to \mathcal{M}_N(\sigma) \equiv \left(m^1_N(\sigma), \ldots, m^M_N(\sigma)\right) \end{aligned} \tag{2.5}$$

and the measures $\mathcal{Q}^\eta_{N,\beta,h}$ on $(I\!R^M, \mathcal{B}(I\!R^M))$ that are induced by $\mathcal{G}^\eta_{N,\beta,h}$ through the map $\mathcal{M}_N$, i.e.

$$\mathcal{Q}^\eta_{N,\beta,h} \equiv \mathcal{G}^\eta_{N,\beta,h} \circ \mathcal{M}_N^{-1} \tag{2.6}$$

Since

$$\mathcal{G}^\eta_{N,\beta,h}(\sigma) = \frac{1}{\left|\mathcal{M}_N^{-1}(\mathcal{M}_N(\sigma))\right|} \mathcal{Q}^\eta_{N,\beta,h}(\mathcal{M}_N(\sigma)) \tag{2.7}$$

these induced measures determine the original measures uniquely. Thus it suffices to determine the limiting induced measures. It turns out that a complete solution to this problem is possible as long as $M(N)/N \downarrow 0$ as $N \uparrow \infty$. Namely the following theorem has been proven by Bovier, Gayrard and Picco [BGP]:

Theorem 1: [BGP] *Assume that M is non-decreasing and satisfies $\lim_{N\uparrow\infty} \frac{M(N)}{N} = 0$. Let $a^\pm(\beta)$ denote the largest (resp. smallest) solution of $a = \tanh(\beta a)$. Then, for all $\beta \geq 0$,*

$$\lim_{h\to 0\pm} \lim_{N\uparrow\infty} \mathcal{Q}^\eta_{N,\beta,h} = \delta^\infty_{a^\pm(\beta)e^\eta}, \quad a.s. \tag{2.8}$$

where the limits are understood in the sense of weak convergence of probability distributions; $\delta^\infty_{a^\pm(\beta)e^\eta}$ denotes the Dirac-measure concentrated on $a^\pm(\beta)e^\eta$ and ϵ^η is the η-th unit vector in $I\!R^{I\!N}$. Moreover,

$$\lim_{N\uparrow\infty} f_{N,\beta} = f_\beta^{CW} \equiv \min_{y\in I\!R} \left(\frac{y^2}{2} - \frac{1}{\beta}\ln\cosh(\beta y)\right), \quad a.s. \tag{2.9}$$

Under stronger hypothesis on the growth of M, this theorem has been proven before in a slightly weaker form (i.e. rather then considering the limiting measures themselves, generally only the expectation values of the overlap parameters were studied): For the case of bounded M, a proof was given first by Figotin and Pastur [FP1, FP2], and later reproduced, with more or less mathematical rigour, in papers by Amit et al. [AGS1], von Hemmen [vH], etc. Later, Koch and Piasko [KP], using a method due to Grensing and Kühn [GK] (who, as we note in passing, were also interested in models for disordered magnets and apparently at first quite unaware of the neural network aspects of the models they discussed) obtained a result for unbounded M, however under the rather strong hypothesis that $M(N) < \frac{\ln N}{\ln 2}$. This result was extended to the Potts-Hopfield model and presented in the form of Theorem 1 by Gayrard [G]. In 1992, two quite different approaches were presented to get results on the free energy under weaker hypothesis on M. One, due to Shcherbina and Tirozzi [ST] proved (2.9) with convergence in probability, while another, much simpler one due to Koch [K], proved convergence of the mean, but could, as was noted by Bovier and Gayrard [BG2], easily be modified to yield the almost sure convergence. In fact, the basic idea used in [K] furnished the starting point for the proof of Theorem 1 in [BGP].

The conditions in Theorems 1 are certainly optimal for the conclusions to hold. They represent in a certain sense an ideal situation for memorization. As M is allowed to be proportional to N this situation is expected to change in that the Gibbs measures are no longer expected to be concentrated on configurations that have exactly overlap 1 (or -1) with one of the original patterns. However, Theorem 1 should be structurally stable in the sense that for small enough α and low enough temperature, the Gibbs states of the model can be seen as small perturbations of the previous ones. A weak version of such a result was proven in [BGP]. To state it, we first need some notations:

For $\delta > 0$, we denote by $a(\delta, \beta)$ the largest solution of the equation

$$\delta a = \tanh(\beta a) \tag{2.10}$$

Let $\|\cdot\|$ be the ℓ^2-norm on $I\!R^{I\!N}$. Given that $\lim_{N\uparrow\infty} \frac{M(N)}{N} = \alpha$, we set, for fixed β,

$$B_\rho^{(\nu,s)} \equiv \left\{x \mid \|x - sa(1-2\sqrt{\alpha}, \beta)e^\nu\| < \rho\right\} \tag{2.11}$$

Finally, put

$$B_\rho \equiv \bigcup_{(\nu,s)\in I\!N\times\{-1,+1\}} B_\rho^{(\nu,s)} \tag{2.12}$$

Theorem 2: *There exists $\alpha_0 > 0$ such that if $\lim \frac{M(N)}{N} = \alpha$, with $\alpha \leq \alpha_0$, then, for all $\beta > 1 + 3\sqrt{\alpha}$, if $\rho^2 > C\left(a(1-2\sqrt{\alpha}, \beta)\right)^{3/2} \alpha^{1/8} |\ln \alpha|^{1/4}$, for almost surely,*

$$\lim_{N\uparrow\infty} \mathcal{Q}_{N,\beta,h=0}\left(B_\rho\right) = 1 \tag{2.13}$$

The set B_ρ is a union of disjoint balls as long as $\beta > \frac{1}{1-c\alpha^{1/4}}$ (The power 1/4 is probably not optimal and due to technical problems in the proofs; we would expect this result with a power 1/2). In this case, one would naturally expect that the *extremal* Gibbs measures are concentrated on these individual balls, that is would really be perturbations of the measures in the $\alpha = 0$ case. Unfortunately, we have so far no rigorous argument to proof this.

We will not give the proofs of these Theorems here, as they are quite involved, but only indicate a broad outline. The first step in the proofs consist of slightly smoothening out the induced measures $\mathcal{Q}^\eta_{N,\beta,h}$ by convoluting them with a normal distribution of variance $1/(\beta N)$. While this does not change the limiting measures, the resulting measures have a density with respect to M-dimensional Lebesgue measure and, moreover, their density can be written in quite explicit form as

$$Q^\eta_{N,\beta,h}(x) = \frac{\exp\left(-\beta N \Phi^\eta_{N,\beta,h}(x)\right)}{\int d^M x \, \exp\left(-\beta N \Phi^\eta_{N,\beta,h}(x)\right)} \tag{2.14}$$

with

$$\Phi^\eta_{N,\beta,h}(x) = \frac{1}{2}(x - he^\eta, x - he^\eta) - \frac{1}{\beta N}\sum_{i=1}^N \ln\cosh\left(\beta(\xi x)_i\right) \tag{2.15}$$

Thus, we are almost in the standard situation for an application of Laplace's method, except that

(a) the function Φ is random and

(b) the dimension M of the underlying space depends on our large parameter N.

As a matter of fact, if M remains bounded, problem (b) does not exist, and problem (a) is almost non-existent, as Φ converges to a non-random limit (by the law of large numbers). These points have been noted and exploited already in [FP1]. For unbounded M our proof is pushing Laplace's methods beyond its immediate domain of applicability 'by hand', but this requires the growth conditions on $M(N)$. In fact, we show that under the condition of Theorem 1 (i.e. if $M = o(N)$), the position and values of the *absolute* minima of the function Φ are asymptotically non-random and that problem (b) is harmless. The proof of Theorem 2 relies on the fact that even for $M = \alpha N$ with small enough α we can localize approximately the absolute minima, but with much less precision. We expect, however, that these estimates can still be improved.

Let us note that the statements of Theorem 1 can be generalized in particular to the dilute Hopfield model with Hamiltonian (1.2). Namely

Theorem 3: *The conclusions of Theorem 1 hold for the dilute Hopfield model (1.2) if the dilution rate $p(N)$ and the number of patterns $M(N)$ satisfy the conditions*

(i) $p(N)N \uparrow \infty$ *as* $N \uparrow \infty$ *and*

(ii) $\frac{M(N)}{p(N)N} \downarrow 0$ *as* $N \uparrow \infty$.

The conditions on the dilution rate in Theorem 3 is presumably the weakest possible for the result in this sharp form. The basic result in [BG2] that allows to prove Theorem 2 is a bound that states that with large probability the Hamiltonian (1.2) and its mean with respect to the dilution (i.e. the ϵ_{ij} which is of course nothing but the original Hopfield Hamiltonian (1.1)) are close to each other in the sense that their difference is of order $o(N)$, *uniformly* in the $\sigma \in S_N$, provided hypothesis (i) and (ii) hold. For the precise statement, see [BG2].

Before closing this section, let us mention one more recent result by Pastur, Shcherbina and Tirozzi [PST]. They consider the so-called Edwards-Anderson parameter,

$$q_N \equiv \frac{1}{N}\sum_{i=1}^{N}\left[\mathcal{G}^{\eta}_{N,\beta,h}(\sigma_i)\right]^2 \tag{2.16}$$

Their result can be paraphrased by saying that *if* the variance of q_N tends to zero as $N \uparrow \infty$, then the order-parameters of the model are those given by the simple-minded use of the "replica-method" (see [AGS2]). This result is analogous to the one obtained for the Sherrington-Kirkpatrick model by Pastur and Shcherbina [PS]. The problem is of course to determine whether the assumption on

q_N is verified. From our Theorem 1, it follows that this is the case if $\alpha = 0$, and from [ScT] this is known to be true at high temperature. In general, at low temperature, one cannot expect this 'self-averaging' to hold.

Summarizing, we see that the low-temperature properties of the Hopfield model for $\alpha > 0$ still remain to be analyzed from a mathematical point of view; Theorem 2 is a first step into this direction. In the next section we discuss some results concerning at least the structure of the Hamiltonian function in this regime.

III. Bounds on the storage capacity

The results on the thermodynamics in the last section concern the true stable states of the dynamics of the infinite system at finite temperature (noise). If one is interested in functioning of the memory on some long, but not infinite time-scale, this may not necessarily be the relevant issue, and it definitely is not the relevant issue for a deterministic gradient dynamics. Newman [N] therefore in 1987 considered the following question: For which range of the parameter α is there a correspondence between the patterns and the *local* minima of the Hamiltonian in the sense that each pattern is surrounded by an energy barrier of extensive height? Clearly, for the gradient dynamics this condition means that starting not too far from a stored pattern, the system will remain close to this pattern for all times; and even if noise is added, this should remain true for a rather long, though finite time.

Newman's result has been generalized to the Potts-version of the Hopfield model in [FMP] and to the dilute model in [BG1]. We give a precise formulation of it in this latter context.

We define on the space of spin configurations the usual Hamming distance,

$$d(\sigma, \sigma') \equiv \frac{1}{2}[N - (\sigma, \sigma')], \tag{3.1}$$

that is the number of components of the spins σ and σ' that disagree. For any σ and any number $\delta \in [0, 1]$ we denote by $\mathcal{S}(\sigma, \delta)$ the sphere of radius δN centered at σ, i.e.

$$\mathcal{S}(\sigma, \delta) \equiv \left\{\sigma' \middle| d(\sigma, \sigma') = [\delta N]\right\}, \tag{3.2}$$

Let us set

$$h_N(\sigma, \delta) \equiv \min_{\sigma' \in \mathcal{S}(\sigma, \delta)} H_N(\sigma') \tag{3.3}$$

We will say that there exists an energy barrier of height ϵN centered at ξ^μ, if for some $\delta \in (0, 1/2)$,

$$h_N(\xi^\mu, \delta) \geq H_N(\xi^\mu) + \epsilon N \tag{3.4}$$

Then

Theorem 4: [BG1] *Suppose* $p \geq c\sqrt{\frac{\ln N}{N}}$. *Then there exists* $\alpha_c \geq 0$, *such that if* $M \leq \alpha_c p N$, *then there exists* $\epsilon > 0$ *and* $0 < \delta < 1/2$ *such that*

$$\liminf_{N\uparrow\infty} \inf_{0\leq\mu\leq m} \left(h_N(\xi\mu, \delta) - H_N(\xi^\mu) - \epsilon N\right) \geq 0 \quad a.s \tag{3.5}$$

Moreover, $\alpha_c \approx\geq (16(\ln(2\sqrt{8(1+a)})\ln(2))^{-1}$, *where*

(i) $a \approx 0$ *if* $\frac{p^2 N}{\ln N} \uparrow \infty$

(ii) $a < \frac{1}{2}$ *if* $c^2 >\approx 7$, *and*

(iii) $a = 1 + \frac{2}{c}$ *otherwise.*

In the particular case $p \equiv 1$ this theorem was first proven by Newman [N]. It is possible to get from the proof more precise information on the relation between ϵ, δ and α. In particular, it is possible to extract from it that local minima are located precisely at the original patterns if $M < c\frac{N}{\ln N}$, and they are located a distance $o(N)$ from the patterns if $M = o(N)$. The first statement was known from earlier work of McEliece et al. [MPRV] and the second agrees with the zero-temperature version of Theorem 1. It should be noticed that the α_c in Theorem 4 is much larger than that of Theorem 3, and that they are not supposed (even ideally) to coincide.

Newman also showed that these minima are not the only ones, but that there exist many others, associated to 'mixtures' of the original patterns, in accordance with prior non-rigorous results of Amit et al. [AGS2]. On the other hand, an exhaustive enumeration of *all* local minima is still missing, as is a complete analysis of the depth of all those minima. Both information are needed to analyse the finite temperature properties of the model. Also, for a more detailed analysis of the dynamics and the various time-scales that could appear, such information is required.

Let us remark finally that so far nobody has been able to prove a converse of Newman's Theorem, that is to show that of α exceeds a critical value, then (3.5) is false. Numerical results appear to imply this with $\alpha_c \approx 0.14$, and it would be interesting to get a better idea for what is happening at this threshold.

The proof of Theorem 4 is based on quite standard large deviation estimates and actually rather straightforward, at least in the case $p = 1$. For general $p < 1$, it requires some further, non-trivial probabilistic bounds on the largest eigenvalues of all submatrices E_I defined as

$$(E_I)_{ij} = \begin{cases} \epsilon_{ij} & \text{if } i \in I \text{ and } j \in I^c \\ \epsilon_{ij} & \text{if } i \in I^c \text{ and } j \in I \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

uniformly for all subsets I of cardinality $|I| = \delta N$, namely that with large probability, $\max_I \|E_I\| \leq$

$c\delta(1-\delta)pN$. In [BG1] this was shown to hold under the condition $p \geq \sqrt{\frac{\ln N}{N}}$, but is is not clear that this is the optimal condition. Note in comparison that Theorem 2 requires only that $pN \uparrow \infty$!

Aknowledgements: A.B. would like to thank the organizers of the 5 CLAPEM, and in particular Pablo Ferrari and Nelson Tanaka, for the invitation to São Paulo and the members and staff of the Instituto de Matemáticas e Estastistica da Universidade de São Paulo for their warm hospitality. He also thanks Leonid Pastur for many interesting discussions.

References

[A] D.J. Amit, "Modelling brain: the world of attractor neural networks", Cambridge University Press, Cambridge, 1989.

[AGS1] D.J. Amit, H. Gutfreund and H. Sompolinsky, "Spin-glass models of neural networks", Phys. Rev. **A 32**, 1007-1018 (1985).

[AGS2] D.J. Amit, H. Gutfreund and H. Sompolinsky, "Storing infinite numbers of patterns in a spin glass model of neural networks", Phys. Rev. Letts. **55**, 1530-1533 (1985).

[BG1] A. Bovier and V. Gayrard, "Rigorous bounds on the storage capacity of the dilute Hopfield model", J. Stat. Phys. **69**, 597-627 (1992).

[BG2] A. Bovier and V. Gayrard, "Rigorous results on the thermodynamics of the dilute Hopfield model", J. Stat. Phys. **72**, 79-112 (1993).

[BM] D. Bollé and F. Mallezie, J. Phys. A **22**, 4409 (1989).

[ES] D. Ederfield and D. Sherrington, "Spin glass, ferromagnetic and mixed phase in the disordered Potts model", J. Phys. C **16**, L971 (1983).

[FMP] P. Ferrari, S. Martinez and P. Picco, "A lower bound for the memory capacity in the Potts-Hopfield model", J. Stat. Phys. **66**, 1643-1651 (1992).

[FP1] L.A. Pastur and A.L. Figotin, "Exactly soluble model of a spin glass", Sov. J. Low Temp. Phys. **3(6)**, 378-383 (1977).

[FP2] L.A. Pastur and A.L. Figotin, "On the theory of disordered spin systems", Theor. Math. Phys. **35** 403-414 (1978).

[FP3] L.A. Pastur and A.L. Figotin, "Infinite range limit for a class of disordered systems", Theor. Math. Phys. **51** 564-569 (1982). (1978).

[G] V. Gayrard, "The thermodynamic limit of the q-state Potts-Hopfield model with infinitely many patterns", *J. Stat. Phys.*__68__, 977-1011 (1992).

[He] D.O. Hebb, "The organization of behaviour", Wiley, New York, 1949.

[vH] J.L. van Hemmen, "Spin glass models of a neural network", Phys. Rev. A **34**, 3435-3445 (1986).

[Ho] J.J. Hopfield,"Neural networks and physical systems with emergent collective computational abilities", Proc. Natl. Acad. Sci. USA **79**, 2554-2558 (1982).

[K] H. Koch, "A free energy bound for the Hopfield model", J. Phys. A **26**, L353-L355 (1993).

[Ka] I. Kanter, Potts-glass model of neural networks, Phys. Rev. A **37**, 2739 (1988).

[KP] H. Koch and J. Piasko, "Some rigorous results on the Hopfield neural network model", J. Stat. Phys. **55**, 903 (1989).

[KPa] J. Komlós and R. Paturi, "Convergence results in a autoassociative memory model", Neural Networks **1**,239-250 (1988).

[Li] W.A. Little, The existence of persistent states in the brain, Math. Biosciences **19**, 101-119 (1974).

[Lu] J. Luttinger, Phys. Rev. Lett. **37**, 778 (1976).

[Ma] D.C. Mattis, Phys. Lett. A **56**, 421 (1976).

[MPRV] R.J. McEliece, E.C. Posner, E.R. Rodemich and S.S. Venkatesh, "The capacity of the Hopfield associative memory", IEEE Trans. Inform. Theory **33**, 461-482 (1987).

[MPV] M. Mézard, G. Parisi and M.A. Virasoro, "Spin-glass theory and beyond", *World Scientific*, Singapore (1988).

[N] Ch.M. Newman, "Memory capacity in neural network models: Rigorous lower bounds", Neural Networks **1**, 223-238 1988).

[PS] L.A. Pastur and M.V. Shcherbina, "Absence of self-averaging of the order parameter in the Sherrington-Kirkpatrick model", J. Stat. Phys. **62**, 1-19 (1991).

[PST] L.A. Pastur, M.V. Shcherbina and B. Tirozi, "The replica symmetric solution without replica trick for the Hopfield model", preprint, to appear in J. Stat. Phys. (1993).

[ScT] E. Scacciatelli and B. Tirozzi, "Fluctuation of the free energy in the Hopfield model", J. Stat. Phys. **67**, 981-1008 (1992).

[ST] M.V. Sherbina and B. Tirozzi, "The free energy for a class of Hopfield models", J. Stat. Phys. **72**, 113 (1992).