STATISTICAL MECHANICS OF NEURAL NETWORKS: THE HOPFIELD MODEL AND THE KAC-HOPFIELD MODEL [#]

Anton Bovier [1]

Weierstraß–Institut
für Angewandte Analysis und Stochastik
Mohrenstraße 39, D-10117 Berlin, Germany

Véronique Gayrard[2]

Centre de Physique Théorique - CNRS
Luminy, Case 907
F-13288 Marseille Cedex 9, France

Abstract: We survey the statistical mechanics approach to the analysis of neural networks of the Hopfield type. We consider both models on complete graphs (mean-field), random graphs (dilute model), and on regular lattices (Kac-model). We try to explain the main ideas and techniques, as well as the results obtained by them, without however going into too much technical detail. We also give a short history of the main developments in the mathematical analysis of these models over the last 20 years.

Keywords: Hopfield model, mean field theory, Kac-models, neural networks, Gibbs measures, large deviations, replica symmetry

[#] Work partially supported by the Commission of the European Union under contract CHRX-CT93-0411

[1] e-mail: bovier@wias-berlin.de

[2] e-mail: gayrard@cpt.univ-mrs.fr

1. Introduction

Large inhomogeneous interactive networks play an increasing rôle in many technologically relevant areas of modern technology such as communication networks, processor networks, and neural networks. As the sizes of available systems is increasing rapidly, it becomes more and more important to gain analytical control over the functioning of such systems. As is evident from many of the contributions to this volume, a probabilistic approach to such systems is most promising, and in particular methods coming from the theory of large deviations and statistical mechanics appear to provide a natural approach to tackle such questions.

In the present review we will focus on a class of models coming from the theory of neural networks and that generalize what is know as the Hopfield model [Ho] of an autoassociative memory. These models have been heavily investigated over the last 20 years both on the level of theoretical physics and, more recently, of rigorous mathematics. They are thus well suited to explain the use of the formalism of statistical mechanics and thermodynamics in the context of disordered networks, and the purpose of these notes is to give an overview of the results achieved so far in this line.

Our aim in this text is to be as understandable as possible also to the non-expert. Thus, while we will present results that are proven rigorously, we will be somewhat informal, explaining and paraphrasing results occasionally rather than stating theorems in a technically precise form. In the same spirit, proofs will not be given but only the main ideas explained. We hope to provide in this way an easily readable text that could serve as a first introduction to the field. A good source for technically more detailed expositions is the recent collection of reviews [BP]. Good reviews on the physical and biological aspects of the field are for instance [A,HKP,GM,MR,DHS].

Let us begin with a short presentation of the original Hopfield model. This model is intended to describe in a very simplified way the interaction of neurons in the cortex in the process of retrieval of memorized information[1]. Each neuron, i, is supposed to have essentially two states, active ("firing") and passive ("non-firing") and is thus represented by the binary variable $\sigma_i \in \{-1, 1\}$. Let N be the total number of neurons in the system. It is known that neurons communicate with each other by sending ("firing") electrical impulses over the connecting *dendrites*. The firing intensity depends on the activation state of the neuron and a given neuron will change its state with time according to the information received from the other neurons. The point now is that the connections between different pairs of neuron do not have identical properties. Rather, the response of neuron i to neuron j depends on a variable J_{ij}, and the collection of these variables represents the memorized information in the network. Note that J_{ij} may represent both the fact whether or not there exists

[1] We should note that the dynamic considered here do not really model the behaviour of actual biological neural networks in detail. A model that reflects more of the neural reality was studied for instance by Turova [Tu].

a dendrite connecting neuron i with neuron j, and the properties of this connection ("synaptic efficiency"), if it exists. The original model of Hopfield assumes full connection, i.e. each neuron is connected to each other, but we will discuss variants of the model where each neuron is only connected to a small (random or deterministic) fraction of the others. Given the set of values J_{ij}, the time evolution of the system is now modeled by a (discrete or continuous time) Markov process where the transition rates for neuron i to change its state at time t depend on the state of the other neurons through the variable $h_i(t) \equiv \sum_{j=1}^N J_{ij}\sigma_j(t)$. More precisely, in the discrete time case the process is described as follows: We start with some initial configuration $\sigma_i(0) = \sigma_i$, $i = 1, \dots, N$. At time $t = n$ we first select a neuron i at random and set

$$\sigma_i(n) = \begin{cases} +1, & \text{with probability } p(\sigma_i(n-1), h_i(n-1)) \\ -1, & \text{with probability } 1 - p(\sigma_i(n-1), h_i(n-1)) \end{cases} \tag{1.1}$$

(Such a dynamic is called "asynchronous", as opposed to a "synchronous" dynamic, where all neurons change their state simultaneously at the same instant. We stick to the asynchronous case here).

Finally, we have to choose the functions $p(\sigma_i, h_i)$. One possibility would be a *deterministic* dynamic[2]where

$$p(\sigma_i, h_i) = \begin{cases} 1, & \text{if } h_i \geq 0 \\ 0, & \text{if } h_i < 0 \end{cases} \tag{1.2}$$

Hopfield observed that if the dendritic couplings are symmetric[3], i.e. $J_{ij} = J_{ji}$, then this deterministic dynamic follows the gradients of the function

$$H_N(\sigma)[J] \equiv -\frac{1}{2N} \sum_{i,j=1}^N J_{ij}\sigma_i\sigma_j \tag{1.3}$$

(the choice of the factor $\frac{1}{2N}$ is of course arbitrary at this point but will become clear soon). The crucial point of this observation is that $H_N(\sigma)[J]$ looks like the Hamiltonian of a *mean field* model for a spin system with inhomogeneous interaction J_{ij}. More precisely, since we will see that the dendritic interactions will be modeled by a collection of random variables and that they will be allowed to take both positive and negative values, this spin system will be qualified as a so-called "spin glass". This observation in Hopfield's 1982 paper certainly sparked the growing interest of the statistical mechanics community in models of neural networks.

[2] For a more extensive discussion of the dynamic of the Hopfield model in general, see the paper by Malyshev and Spieksma [MS].

[3] It has been argued frequently that this symmetry assumption is unrealistic in the context of biological systems. One should therefore not forget that there are many systems that cannot be treated immediately with the methods we describe in this paper.

Once a Hamiltonian is seen to appear, it is quite natural to introduce a non-deterministic dynamic in such a way that the corresponding invariant measure is the Gibbs measure for this Hamiltonian with inverse temperature β. This can be achieved e.g. by choosing

$$p(\sigma_i, h_i) = \frac{e^{\beta h_i/2}}{2\cosh(\beta h_i/2)} \tag{1.4}$$

In fact, with this choice the Markov chain defined in (1.1) is reversible with respect to the Gibbs measure

$$\mu_N(\sigma)[J] \equiv \frac{e^{-\beta H_N(\sigma)[J]}}{Z_{N,\beta}[J]} \tag{1.5}$$

This is the reason why the statistical mechanics problem of studying the Gibbs measures for H_N is relevant for the analysis of the dynamic of the Hopfield neural network. We will come back to the relation between dynamic and equilibrium statistical mechanics later in more detail.

So far we have not said much about the dentritic efficiencies J_{ij}. Hopfield's intention was to model an autoassociative memory. That is to say, the dynamic of the network should allow to associate initial conditions σ ("presented images") to given, previously memorized images, called "patterns" and conventionally denoted $\xi^\mu \in \{-1,1\}^N$, if the presented image is in some sense close to a given patterns. To do this, the dendritic efficiencies should be chosen as a function of a set of patterns $\xi^1, \xi^2 \dots, \xi^M$ that one wants to store. There are many elaborate ways of choosing J as a function of these patterns, but Hopfield's choice was the old Hebb's learning rule

$$J_{ij} = \sum_{\mu=1}^{M} \xi_i^\mu \xi_j^\mu \tag{1.6}$$

With such a choice, one wants to answer the basic question: For which values of the parameter β and for which values of M does the above Markov chain function as a memory, i.e. when does $\sigma(0) \sim \xi^\mu$, for some μ, imply $\sigma(t) \to \xi^\mu$, as $t \uparrow \infty$, or at least $\sigma(t) \sim \xi^\mu$ for "most of the time" (at least as long as t is not astronomically large). Of course the answer to this question may depend on the specific patterns stored. In principle, we would want an affirmative answer for all possible patterns, but some reflection shows that this will hardly be possible. On the other hand, if N is large, there are enormously many patterns, and we may be willing to accept that for a small subset of patterns that we are not likely to select, the memory does not work. To make such statements precise, it is natural to construct a probabilistic model for the possible choice of the patterns. The most simple one is to assume that all patterns are chosen independently and each possible pattern has the same probability to be chosen. This leads to the assumption that ξ_i^μ are independent identically distributed random variables and $I\!P[\xi_i^\mu = \pm 1] = 1/2$. This is again Hopfield's choice and we will be mostly concerned with this situation here. But note that considerable work has been done for different choices of the distribution of the patterns (see e.g. [L1] and references therein).

The Hopfield model (the above setup will be sometimes called the "standard Hopfield model") has the special feature that it can be considered as a "mean-field model" in a very specific sense. Namely, we can introduce the so called "overlap parameters"

$$m_N^\mu(\sigma)[\xi] \equiv \frac{1}{N}\sum_{i=1}^N \xi_i^\mu \sigma_i, \quad \mu = 1,\dots,M \tag{1.7}$$

Then the Hamiltonian turns out to be a simple function of these parameters only

$$H_N(\sigma)[\xi] = -\frac{N}{2}\sum_{\mu=1}^M \left(m_N^\mu(\sigma)[\xi]\right)^2 \equiv -\frac{N}{2}\left\|m_N(\sigma)[\xi]\right\|_2^2 \tag{1.8}$$

The overlap parameters play the rôle of "macroscopic variables", or order parameters, in this model. The general principle of the thermodynamic formalism is to deduce from a probabilistic description of the microscopic variables (the σ_i in our case) the *deterministic* laws (both dynamical and analytical) of the *macroscopic* observables. The specific form of H_N helps greatly to make this program feasible. However, contrary to conventional "mean field models", there are two essential difficulties we have to deal with here: First, the macroscopic observables are *random* functions of the microscopic ones, and second in the situation we are most interested in, the number of macroscopic variables, M, depends on the size, N, of the system, and tends to infinity. This is due to the fact that in the memory context, one of our main questions is how many patterns a network can store! Hopfield observed in numerical simulations that the maximal number scales like $M(N) = \alpha N$, with $\alpha \approx 0.14$ (in the case of deterministic gradient dynamic).

The first part of this review will be devoted to these mean field models. As we said earlier, the assumption of full connectivity of the network is frequently unrealistic. One would thus like to study models with some geometric structure. We will discuss two variants of the Hopfield model that take this into account:

Dilute Hopfield model: This model can be viewed as a Hopfield model on a random graph. Consider independent random variables $\epsilon_{ij} = \epsilon_{ji}$, $i \geq j$, that take values in $\{0,1\}$, with $I\!P[\epsilon_{ij} = 1] = p(N)$, where we may allow $p(N)$ to go to zero as N tends to infinity. The Hamiltonian for this model is

$$H_N(\sigma)[\xi,\epsilon] = -\frac{1}{2p(N)N}\sum_{i,j=1}^N \epsilon_{ij}\sum_{\mu=1}^M \xi_i^\mu \xi_j^\mu \sigma_i\sigma_j \tag{1.9}$$

Kac-Hopfield model: Here we assume the neurons to be located on the vertices of a finite subset Λ of some lattice $\mathbb{Z}^d$. We choose a deterministic function $J_\gamma(i)$ and set

$$H_{\Lambda,\gamma}(\sigma)[\xi] = -\frac{1}{2}\sum_{i,j=1}^N J_\gamma(i-j)\sum_{\mu=1}^M \xi_i^\mu \xi_j^\mu \sigma_i\sigma_j \tag{1.10}$$

Here γ is a small parameter and the function $J_\gamma(i)$ is supposed to be of the form $J_\gamma(i) = \gamma^d J(\gamma i)$, with $J(x)$ some fixed function that has compact support or decays rapidly (e.g. $J(x) = e^{-|x|}$, or $J(x)$ the indicator function of the unit cube). The Kac-Hopfield model is suited to investigate properties of the system on a mesoscopic scale which makes it particularly interesting. It has received only little attention so far, although it was introduced already in 1978 by Figotin and Pastur [FP3]. We will discuss in detail some recent progress in this field.

The remainder of this paper is organized as follows. In Section 2 we introduce a class of generalized random mean field models, discuss the basic questions to be studied and introduce the main mathematical tools used for their analysis. In Section 3 we present the results obtained in this way for the Hopfield model. In Section 4 we review some results obtained for the dilute and the Kac-Hopfield models. Finally, in Section 5, we give a brief historical outline of the main developments that have lead to our present status of understanding of these models.

2. Mean field models: Thermodynamic formalism

2.0 Generalized random mean field models. In this section we want to discuss the thermodynamic formalism for a class of models that somewhat generalize the Hopfield model [BG5]. We consider a configuration space $\mathcal{S}_N = \{-1,1\}^N$, and a family of linear random maps $\xi^\mu : \mathcal{S}_N \to I\!R$, $\mu \in I\!N$, defined on some abstract probability space $(\Omega, \mathcal{F}, I\!P)$. Define

$$m_N^\mu(\sigma)[\xi] \equiv \frac{1}{N}(\xi^\mu, \sigma) \tag{2.1}$$

where in coordinates $(\xi^\mu, \sigma) \equiv \sum_{i=1}^N \xi_i^\mu \sigma_i$. Let E_M be a real valued function on $I\!R^M$. We will assume that E_M is non-negative, convex and essentially smooth [Ro] (i.e. its gradient diverges on the boundary of its domain). We can now define a Hamiltonian

$$H_{N,M}(\sigma)[\xi] \equiv -\frac{N}{2} E_M\left(m_N(\sigma)[\xi]\right) \tag{2.2}$$

where $E_M\left(m_N(\sigma)[\xi]\right) \equiv E_M\left(m_N^1(\sigma)[\xi], \dots, m_N^M(\sigma)[\xi]\right)$. It is clear that the standard Hopfield model is an example of such a system, but neither the dilute Hopfield model not the Kac-Hopfield model fall into this class. An important feature in these models is that there are two parameters, N and M, which both should be thought of as large. In fact, we will be interested in the asymptotic behaviour as N tends to infinity with $M = M(N)$ a given function that tends to infinity. The most interesting case is $M(N) = \alpha N$.

Given such a Hamiltonian, we can define the finite volume Gibbs measures by

$$\mu_{\beta,N,M}(d\sigma)[\xi] = \frac{e^{-\beta H_{N,M}(\sigma)[\xi]}}{Z_{\beta,N,M}[\xi]} \prod_{i=1}^N q(d\sigma_i) \tag{2.3}$$

where q is the a priori distribution on $\{-1,1\}$ of the single spin, e.g. $q(\pm 1) = 1/2$. We write $I\!E_\sigma$ for the expectation with respect to this a priori measure. The *partition function* $Z_{\beta,N,M}[\xi]$ is given by

$$Z_{\beta,N,M}[\xi] = I\!E_\sigma e^{-\beta H_{N,M}(\sigma)[\xi]} \tag{2.4}$$

In models of our type the finite volume Gibbs measures can be reconstructed from the distribution of the overlap parameters,

$$\mathcal{Q}_{\beta,N,M}(m)[\xi] \equiv \mu_{\beta,N,M}\left(\{m_N(\sigma)[\xi] = m\}\right) \tag{2.5}$$

We call $\mathcal{Q}_{\beta,N,M}$ the *induced measures.*

Remark: We would like to stress at this point that while the Gibbs and induced measures are equivalent in finite volume, this is not necessarily true if one passes to the infinite volume limits. The reason is that the natural topology on the spin space in the infinite volume limit is the product topology, and the same is true on the space $I\!R^\infty$ of overlaps. But the maps $m_N : \mathcal{S}_N \to I\!R^M$ are not uniformly continuous with respect to these topologies as $N \uparrow \infty$. This point has to be kept in mind when discussing infinite volume limits.

The measures $\mathcal{Q}_{\beta,N,M}$ can be expected to have large deviation properties. In fact, if M is kept fixed as $N \uparrow \infty$, this is not very difficult to verify and has been the object of intensive study in the early 80's [vHvEC,vH1,GK,vHGHK,vH2,AGS2,JK,vEvHP].

The study of the induced measures can be seen both as a tool to get insights in the Gibbs measures and as an end in itself. Namely, we might say that the knowledge of the overlaps of a spin configuration with all the stored patterns gives sufficient criteria for the recognition of a pattern in the retrieval process.

2.1 Mean field dynamics. For fixed N and M and fixed ξ, we denote by $\mathcal{W}_{N,M}[\xi]$ the set

$$\mathcal{W}_{N,M}[\xi] \equiv \left\{x \in I\!R^M \left| \exists_{\sigma\in\mathcal{S}_N} : x = m_N(\sigma)[\xi]\right.\right\} \tag{2.6}$$

We now define the (discrete time) Markov chain $x_N(n)$, $n \in I\!N$ with state space $\mathcal{W}_{N,M}$ with transition matrix

$$P(y|x) = \begin{cases} K\left[\frac{\mathcal{Q}_{\beta,N,M}(x\pm 2\xi_i/N)[\xi]}{\mathcal{Q}_{\beta,N,M}(x)[\xi]}\right]^{1/2}, & \text{if } y = x \pm 2\xi_i/N \text{ for some } i \\ 1 - K\sum_{i=1}^N \sum_{s=\pm 1}\left[\frac{\mathcal{Q}_{\beta,N,M}(x+s\xi_i/N)[\xi]}{\mathcal{Q}_{\beta,N,M}(x)[\xi]}\right]^{1/2}, & \text{if } x = y \\ 0, & \text{else} \end{cases} \tag{2.7}$$

where K must be chosen such that $P(y|x) \geq 0$ for all $x, y \in \mathcal{W}_{N,M}$. Obviously, this chain is reversible w.r.t. $\mathcal{Q}_{\beta,N,M}$. Note that this dynamics is *not* identical to the dynamics induced on the overlaps by a Markovian Glauber (spin-flip) dynamic reversible w.r.t. the Gibbs measures (this induced dynamics is generally not Markovian), but it is trying to imitate this by a Markov process as closely as possible. It seems reasonable to expect that the two dynamics actually have quite similar long-time behaviour, but as far as we know nothing is known rigorously about this issue. In any case, we find it certainly interesting to study this mean-field dynamics by itself. It is suggestive to define

$$F_{\beta,N,M}(x)[\xi] \equiv -\frac{1}{\beta N} \ln \mathcal{Q}_{\beta,N,M}(x)[\xi] \tag{2.8}$$

The above process can then be seen as some kind of random walk (on $I\!R^M$) in a landscape described by $F_{\beta,N,M}$. One would expect the process to prefer to stay in the "valleys", i.e. the minima of $F_{\beta,N,M}$ and to take long times to get from one valley to another. This intuition is mainly based on the following simple but fundamental observation. Let $\tau_x^y > 0$ denote the first time that the process $X_N(t)$ starting at point y at time $t = 0$ hits the point x at some later moment.

Lemma 2.1: *Let $x_N(t)$ be a Markov process with state space $\mathcal{W}_{N,M}$ that is reversible with respect to the measure $\mathcal{Q}_{\beta,N,M}$. Let P denote the law of this process. Then*

$$P\left[\tau_x^y < \tau_x^x\right] = e^{\beta N[F_{\beta,N,M}(x) - F_{\beta,N,M}(y)]} P\left[\tau_y^x < \tau_y^y\right] \tag{2.9}$$

for any $x, y \in \mathcal{W}_{N,M}$.

This lemma says that the probability that the process starting from x hits y *before* it returns to x is related to the probability of doing the opposite, namely to hit x before y when starting in y by a factor $e^{\beta N[F_{\beta,N,M}(x) - F_{\beta,N,M}(y)]}$. In particular, if x is in a valley and y on a mountain, than the probability to reach y from x without an intermediate return to x is very small, namely of the order $\exp(-\beta N)$. It is not difficult to deduce from Lemma 2.1 that in such a case, the expected time to reach y from x is at least of order $\exp\left(\beta N[F_{\beta,N,M}(y) - F_{\beta,N,M}(x)]\right)$, i.e. exponentially large.

This leads to the general picture that any local minimum of the function $F_{\beta,N,M}$ surrounded by walls of height δ should trap the process for times of order $\exp(\beta N \delta)$. Moreover, the process spends most of its time near the deepest minima of $F_{\beta,N,M}$. This picture is clear and well understood in the case of one single order parameter. The corresponding analysis of the long time behaviour was performed in [CGOV]; recently, a more precise analysis that also computes the polynomial corrections to the exponential asymptotic was given by Eckhoff [Eck]. In higher dimensions, we are not aware of any systematic analysis of the situation. The problem here is that on the one hand, one-dimensional methods, based on exact solutions of the finite difference equations cannot be applied. On the other hand, other techniques (see. e.g. [OS]) require a finite state space, while in our situation the number of points in the state space depends on the large parameter. A first analysis of the problem of exit times from domains containing a local minimum in the case of finite, N-independent dimension is carried out in [E1]. The expected behaviour on the level of the exponential asymptotic is confirmed. To deal with the case where the dimension M increases with N will require considerably more work and more precise estimates. Such an analysis is under way.

If we take the picture outlined above for granted, we see that the system functions in the desired way as a memory, if for large N, the function $F_{\beta,N,M}$ has its deepest minima near the points $m_N(\xi^\mu)$. This motivates why we want to check under what conditions on the parameters such a statement is true.

2.2 Large deviations Above we have motivated why the functions $F_{\beta,N,M}(y)$ are of interest in understanding the long time dynamics of our model. However, their definition in (2.8) is not very convenient, mainly because they can only be defined on the discrete, N-dependent sets $\mathcal{W}_{N,M}$. It is thus suitable to define the smoothed out version of this function on all of $\mathbb{R}^M$ by

$$f_{\beta,N,M,\rho}(x)[\xi] \equiv -\frac{1}{\beta N} \ln \mathcal{Q}_{\beta,N,M}\left(B_\rho(x)\right)[\xi] \tag{2.10}$$

where $B_\rho(x)$ denotes the ball in $\mathbb{R}^M$ of radius ρ centered at x. In the case $M < \infty$ independent of N, we can immediately ask whether

$$\lim_{\rho\downarrow 0} \lim_{N\uparrow\infty} f_{\beta,N,M,\rho}(x)[\xi] \equiv f_{\beta,M}(x)[\xi] \tag{2.11}$$

exists. If so, the family of measures $\mathcal{Q}_{\beta,N,M}$ satisfies a large deviation principle with rate function $f_{\beta,M}(x)[\xi]$. It is not hard to establish that under our assumptions on the energy function E and under mild assumptions on the distribution of ξ, this limit will exist for almost all ξ and, moreover, the limit will be independent of ξ.

Can such a result be extended to the case where M depends on N? The real question here is how we want to extend it. The problem here is that the domains where the functions $f_{\beta,N,M(N),\rho}(x)[\xi]$ are defined now again depend on N. An obvious way out is to study the distributions of only finitely many of the overlap parameters at a time, i.e. to fix any finite set $I \subset \mathbb{N}$ and to define functions

$$f^I_{\beta,N,M,\rho}(x)[\xi] \equiv -\frac{1}{\beta N} \ln \mathcal{Q}_{\beta,N,M}\left(B^I_\rho(x)\right)[\xi] \tag{2.12}$$

where now $x \in \mathbb{R}^I$ and $B^I_\rho(x)$ is a ball in $\mathbb{R}^I$. We can now ask whether for all I these functions converge to a limit as $N \uparrow \infty$ and $\rho \downarrow 0$. This has been proven to be the case in the standard Hopfield model under the condition that $\lim_{N\uparrow\infty} M(N)/N = 0$ in a rather recent paper by Bovier and Gayrard [BG3]. The rate functions are again almost surely independent of ξ and are given in [BG3] in terms of some (rather horrible) variational formula.

While such a result is esthetically appealing, it is in some sense not very satisfying for understanding the functioning of the network. For one thing, the case $M = o(N)$ is not so interesting (for we expect the memory to function with a much higher load), for the other we are in a way more interested in knowing how the rate function looks like then to know that it exists. Also, for the study of the Gibbs measures themselves, we need to have large deviation estimates in the ℓ_2 sense and not in the product topology.

There are two ways out of this difficulty. The first is to avoid the use of large deviation techniques altogether and to make use of what is called the "Hubbard-Stratonovich transformation". This technique only works in the case where E_M is a purely quadratic function, and thus in particular in the case of the standard Hopfield model, where it was used in the original papers of Figotin and Pastur [FP1,FP2] for the first time. It is based on the simple observation that

$$e^{x^2/2} = \frac{1}{\sqrt{2\pi}} \int dz e^{-z^2/2+xz} \tag{2.13}$$

By this trick one can show, if $E_M(m) = \frac{1}{2}\|m\|_2^2$, that the measure

$$\widetilde{\mathcal{Q}}_{\beta,N,M}[\xi] \equiv \mathcal{Q}_{\beta,N,M}[\xi] \star \mathcal{N}\left(0, (\beta N)^{-1} \mathbb{1}\right) \tag{2.14}$$

(where $\star$ denotes the convolution and $\mathcal{N}(0, a\mathbb{1})$ the M-dimensional gaussian measure with mean zero and covaraince matrix $a\mathbb{1}$) is absolutely continuous with respect to the M-dimensional Lebesgue measure with density

$$\frac{d\widetilde{\mathcal{Q}}_{\beta,N,M}[\xi](z)}{d^M z} = \frac{e^{-\beta N \Phi_{\beta,N,M}(z)[\xi]}}{Z_{\beta,N,M}[\xi]} \tag{2.15}$$

where

$$\Phi_{\beta,N,M}(z)[\xi] \equiv \frac{\|z\|_2^2}{2} - \frac{1}{\beta N} \sum_{i=1}^{N} \ln\cosh\left(\beta(\xi_i, z)\right) \tag{2.16}$$

and $(\xi_i, z) \equiv \sum_{\mu=1}^{M} \xi_i^\mu z_\mu$. The measures $\widetilde{\mathcal{Q}}_{\beta,N,M}[\xi]$ can be studied using the Laplace method (in a space of growing dimension) by studying the function $\Phi_{\beta,N,M}(z)[\xi]$. The information obtained can be used to deduce properties of the induced measures and the Gibbs measures.

While this approach is very elegant and simple, the limitation to purely quadratic functions is rather annoying and makes it appear to be some cheap trick. Therefore we have devised in [BG5] an alternative approach based on large deviation ideas that works in more generality.

Let us define the function

$$\Phi_{\beta,N,M}(x) = -E_M(x) + (x, \nabla E_M(x)) - \mathcal{L}_{\beta,N,M}(\nabla E_M(x)) \tag{2.17)}$$

with

$$\begin{aligned} \mathcal{L}_{\beta,N,M}(t) &\equiv -\frac{1}{\beta N} \ln I\!\!E_\sigma e^{\beta(t, m_N(\sigma)[\xi])} \\ &= -\frac{1}{\beta N} \sum_{i=1}^{N} \ln\cosh\left(\beta(t, \xi_i)\right) \end{aligned} \tag{2.18}$$

Then

Theorem 2.1:

(i) Let x^* be a point in $I\!R^M$ such that for some $\rho_0 > 0$, for all $x, x' \in B_{\rho_0}(x^*)$, $\|\nabla E_M(x) - \nabla E_M(x')\|_2 < c\|x - x'\|_2$. Then, for all $0 < \rho < \rho_0$

$$\frac{1}{N} \log Z_{N,\rho}(x^*) \leq -\Phi_{\beta,N,M}(x^*) + \frac{1}{2}c\rho^2 \tag{2.19}$$

(ii) Let x^* be the position of a local extremum of $\Phi_{\beta,N,M}[\xi]$. Then,

$$\frac{1}{N} \log Z_{N,\rho}(x^*) \geq -\Phi_{\beta,N,M}(x^*) + \tfrac{1}{N} \log(1 - \tfrac{1}{\rho^2 N} \Delta\mathcal{L}_{\beta,N,M}(\nabla E_M(x^*))) \tag{2.20}$$

The proof of this theorem can be found in [BG5].

Remark: Note that in the case where E_M is quadratic, the function defined in (2.17) is the same as the one bearing the same name that appears in (2.15). This is no coincidence.

2.3 Laplace method. Theorem 2.1 is just made to suffice to control concentration properties of the measures $\mathcal{Q}_{\beta,N,M}$ "near" the minima of the function $\Phi_{\beta,N,M}$, when N, but also M becomes large. Here what will be "near" will depend explicitly on the speed of divergence of M. The idea of the proof is very simple: Let $\mathcal{A} \subset I\!R^M$ be such that the absolute minimum of $\Phi_{\beta,N,M}$ is contained in it together with a sufficiently large neighborhood to use the lower bound (2.20) to bound its probability from below. Then cover the complement of $\mathcal{A}$ with sufficiently small balls (small enough to make the error term $c\rho^2$ in (2.19) unimportant). Use the upper bound to estimate the probabilities and sum up. Compare the two contributions. If $\mathcal{A}$ wins, the measure concentrates on $\mathcal{A}$. It will win if the volume of the level sets of $\Phi_{\beta,N,M}$ do not grow too fast.

A precise version of the result is the following Theorem:

Theorem 2.2: *Let $\mathcal{A} \subset I\!R^\infty$ be a set such that for all N sufficiently large the following holds:*

(i) There is $n \in \mathcal{A}$ such that for all $m \in \mathcal{A}^c$,

$$\Phi_{\beta,N,M}(m) - \Phi_{\beta,N,M}(n) \geq C\alpha \tag{2.21}$$

for $C > 0$ sufficiently large.

(ii) $\Delta\mathcal{L}_{N,M}(\nabla E_M) \leq KM$ for some $K < \infty$, and $B_{K\sqrt{\alpha}}(n) \subset \mathcal{A}$. Assume further that Φ satisfies a tightness condition, i.e. there exists a constant, a, sufficiently small (depending on C), such that for all $r > C\alpha$

$$\ell\left(\{m \,|\, \Phi_{\beta,M,N}(m) - \Phi_{\beta,M,N}(n) \leq r\}\right) \leq r^{M/2} a^M M^{-M/2} \tag{2.22}$$

where $\ell(\cdot)$ denotes the Lebesgue measure. Then there is $L > 0$ such that

$$\mathcal{Q}_{\beta,N,M}\left(\mathcal{A}^c\right) \leq e^{-L\beta M} \tag{2.23}$$

and in particular

$$\lim_{N\uparrow\infty} \mathcal{Q}_{\beta,N,M}\left(\mathcal{A}\right) = 1 \tag{2.24}$$

Remark: Condition (2.22) is verified, e.g., if Φ is bounded from below by a quadratic function.

2.4 Symmetry. We have seen that under reasonable conditions we can show that the induced measures concentrate on a small neighborhood of the set where the function Φ is not by more than αC larger than its absolute minimum. The most interesting situation for us is when this set consists of several disjoint connected components. This corresponds to the case of "phase transitions" or, in the memory context, memorization of several patterns. If such a situation occurs we would like to be able to compare the relative masses of the individual connected components in order to decide which are the most important, viz. most stable ones. Due to the symmetry of the disorder variables (e.g. in the standard Hopfield case where their distribution is invariant under the permutation of all the indices i and μ), one can also expect that the set $\mathcal{A}$ breaks up into connected subsets $\mathcal{A}_k$ whose masses have the same probability distribution, in particular we may have

$$I\!\!E \ln \mathcal{Q}_{\beta,N,M}(\mathcal{A}_k) = I\!\!E \ln \mathcal{Q}_{\beta,N,M}(\mathcal{A}_l) \tag{2.25}$$

for all indices k, l. In such a situation we would like to assert that with overwhelming probability, the random quantities $\ln \mathcal{Q}_{\beta,N,M}(\mathcal{A}_k)$ for different indices k differ from each other only by terms small compared to N. Thinking about this problem one soon realizes that in the case $\alpha > 0$, it cannot be solved by elementary means, simply because the errors in our large deviation estimates will always be bigger than the random fluctuations we want to control. This difficulty was responsible for the fact that this problem remained unsolved for quite some time. In [BGP3] the crucial idea to use concentration of measure techniques that give exponential estimates on fluctuations was introduced and used to prove such a result in the Hopfield model. Since then, the proof was refined and streamlined, mainly due to the use of a new general theorem on measure concentration that was proven by Michel Talagrand [T1] (presumably in view of applications to the Hopfield model). Using his theorem, we proved in [BG5] the following general theorem that is valid for a large class of the models we discuss here.

Theorem 2.3: *Let ξ_i^μ be bounded i.i.d. random variables with mean zero and variance 1. Let $\mathcal{A}_k$ be a family of sets that verify (2.25) and for which for some constant $c < \infty$*

$$\mathcal{Q}_{\beta,N,M}(\mathcal{A}_k) \geq e^{-c\beta N} \tag{2.26}$$

with probability greater than $1 - e^{-M}$. *Then there is a finite constant* $0 < C < \infty$ *such that for all* $1/C > \epsilon > 0$, *for any* k, l,

$$\mathbb{P}\left[\tfrac{1}{\beta N}\left|\ln \mathcal{Q}_{\beta,N,M}(\mathcal{A}_k) - \ln \mathcal{Q}_{\beta,N,M}(\mathcal{A}_l)\right| \geq C\epsilon^{1/4} + x\right] \leq Ce^{-M} + Ce^{-\frac{x^2\epsilon^2 N}{C}} \tag{2.27}$$

We see that this theorem implies the desired result by choosing ϵ and x of the form $N^{-\delta}$ with e.g. $\delta = 1/5$. Let us also note that in the case of the standard Hopfield model, one can get somewhat sharper estimates (see [BG4]).

We will not give the proof of this theorem which can be found in [BG5]. Very roughly, it consists of showing that the quantities $\frac{1}{\beta N}\ln \mathcal{Q}_{\beta,N,M}(\mathcal{A})$ are Lipshitz continuous as functions of the NM random variables ξ_i^μ, and then to use Talagrand's general theorem on the concentration properties of Lipshitz functions of bounded i.i.d. random variables.We stress that this is an extremely powerful tool in the analysis of fluctuations of random systems which has not received sufficient appreciation yet.

3. Some specific results

In the previous section we explained how to reduce the analysis of the induced measures to the study of the explicitly calculable random function Φ. A lot of our efforts have to go into this analysis for specific models. Our specific results so far concern the cases where $E_M(x) = \frac{1}{p}\|x\|_p^p$ for integer $p \geq 2$, spin space $\mathcal{S} = \{-1, 1\}$ and i.i.d. bounded random variables ξ_i^μ. In this case, the results can be summarized by saying that if the corresponding mean field free energy for the single pattern model has two degenerate minima $\pm m^*$, then, if $M = \alpha N$ with α sufficiently small (typically we need $\sqrt{\alpha} < \gamma(m^*)^2$, and γ small enough), then the lowest minima of Φ are located in the balls of radius $c\gamma m^*$ around the points $\pm m^* e^\mu$, $\mu = 1, \dots, M$, with overwhelming probability.

The most elegant and general strategy to prove such a statement is to first consider the averaged function $\mathbb{E}\Phi_N(x)$. One shows quite easily that its lowest minima are located precisely at the points $\pm m^* e^\mu$, and with some more work one also gets bounds on the growth of the function away from these minima. Only this part of the analysis depends on the details of the the function E. Then one studies the fluctuations of $\Phi_N(x)$ around its average. Roughly speaking, one arrives at showing that with probability close to one these are uniformly in x of order at most $\sqrt{\alpha}$. The technique used here relies on concentration of measure estimates together with "chaining", a technique well known for instance from the analysis of the regularity properties of stochastic processes.

We will get estimates on fluctuations uniformly inside balls $B_R(x) \equiv \left\{x' \in \mathbb{R}^M \mid \|x - x'\|_2 \leq R\right\}$ of radius R centered at the point $x \in \mathbb{R}^M$.

Proposition 3.1: *Assume $\alpha \leq 1$. Let $\{\xi_i^\mu\}_{i=1,\dots,N;\mu=1,\dots,M}$ be i.i.d. random variables taking values in $[-1,1]$ and satisfying $\mathbb{E}\xi_i^\mu = 0$, $\mathbb{E}(\xi_i^\mu)^2 = 1$. For any $R < \infty$ and $x_0 \in \mathbb{R}^M$ and $x_0 \in \{sm^* e^\mu, s = \pm 1, \mu = 1, \dots, M\}$, we have:*

i) for $p = 2$ and $\beta < 11/10$, there exist finite numerical constants C, K such that

$$
\begin{aligned}
\mathbb{P}\Bigg[\sup_{x \in B_R(x_0)} & \left| \frac{1}{\beta N} \sum_{i=1}^{N} \{\mathbb{E} \ln\cosh(\beta(\xi_i, x)) - \ln\cosh(\beta(\xi_i, x))\} \right| \\
& \geq C\sqrt{\alpha} R(m^* + R) + C\alpha m^* + 4\alpha^3(m^* + R)\Bigg] \\
& \leq \ln\left(\tfrac{R}{\alpha^3}\right) e^{-\alpha N} + e^{-\alpha^2 N}
\end{aligned}
\tag{3.1}
$$

ii) For $\beta \geq 11/10$,

$$
\begin{aligned}
\mathbb{P}\Bigg[\sup_{x \in B_R(x_0)} & \left| \frac{1}{\beta N} \sum_{i=1}^{N} \{\mathbb{E} \ln\cosh(\beta(\xi_i, x)) - \ln\cosh(\beta(\xi_i, x))\} \right| \\
& \geq C\sqrt{\alpha} R(R + \|x_0\|_2) + C\alpha + 4\alpha^3\Bigg] \leq \ln\left(\tfrac{R}{\alpha^3}\right) e^{-\alpha N} + e^{-\alpha^2 N}
\end{aligned}
\tag{3.2}
$$

Remark: The proof of this proposition can be found in [BG5].

Having control over fluctuations, the problem remains to study the behaviour of the average function, $\mathbb{E}\Phi_{\beta,N}$. Note that while this function is independent of N, it is still a function of M variables and not entirely trivial to study. Fortunately, in the case of the standard Hopfield model, it is easy to see that the points $\pm m^* e^\mu$, are absolute minima of this function, and it is also not hard to see that they are the only ones. However, to get strict lower bounds is already a non-trivial matter that so far demands considerable work. We proved the following result in [BG5]:

Proposition 3.2: *Assume that ξ_i^μ are i.i.d. with $\mathbb{P}[\xi_i^\mu = \pm 1] = 1/2$. Then, for all $\beta > 1$, there exists a strictly positive constant $C(\beta)$ such that*

$$
\mathbb{E}\Phi_{\beta,N,M}(x) - \Phi_{\beta,N,M}(e^1 m^*(\beta)) \geq C(\beta) \inf_{s,\mu} \|x - s e^\mu m^*(\beta)\|_2^2 \tag{3.3}
$$

The infima are over $s \in \{-1, +1\}$ and $\mu = 1, \dots, M$.

With these ingredients one can now show easily that the induced measures are supported on small balls around the points $\pm e^\mu m^*$, provided α is sufficiently small. We emphasize that the only place where very specific properties of the model enter, and where work has to be done to generalize is Proposition 3.2.

The replica symmetric solution. The large deviation approach outlined above clearly can only be expected to yield in some sense "qualitative" results, when $\alpha > 0$. These are in nice agreement with the predictions obtained with the replica method, but there appeared to be little hope that the precise numerical predictions of that method could be reproduced. Almost surprisingly, however, it turned out quite recently that at least some of the exact results of the replica approach can be recovered by rigorous methods. The key additional idea here is to use the so-called "cavity method", which is nothing but induction over the size of the volume. This method had been used by physicists (see [MPV]) both in the SK-model and in the Hopfield model as an alternative device to derive the predictions of the replica method. The original implementation of this approach involves numerous uncontrolled approximations, and as such is no more rigorous than the replica method itself. However, with enough courage one might hope that a rigorous version of this method could be derived. The idea to do this appears first in a paper by Pastur and Shcherbina [PS] in the context of the SK-model, and later in a paper by the same authors and Tirozzi [PST] for the Hopfield model. These papers provide conditional results that link the validity of the replica symmetric solution to self-averaging properties of some order parameter, without showing that this property was ever satisfied (there are also some steps in the chain of arguments that are not easy to verify). This basic idea was reconsidered in a recent paper by Talagrand [T2]. Carrying the induction method through with full control on all error terms (which he controlled in turn by induction), and using the a priori estimates on the distribution of the overlaps obtained earlier, he succeeded in proving that there exists a non-trivial domain of the parameters for which the replica symmetric solution of [AGS] can be proven to be correct. Subsequently, we gave a different proof of this result, and some more consequences of it in [BG7] and we find it instructive to give a brief outline of this approach (which takes up more closely some of the ideas in [PST]).

Suppose we wanted to construct, instead of the measure on the overlaps the original Gibbs measures on the spin variables. Since the topology which we consider is the product topology on the spins, to control the measures it is enough to consider any finite subset $I \subset I\!N$ and to compute the probability that $\sigma_i = s_i$, for all $i \in I$. We assume that $\Lambda \supset I$, and for notational simplicity we put $|\Lambda| = N + |I|$.

Without loss of generality it suffices to consider the measures $\mu^{(1,1)}_{\Lambda,\beta,\rho}$ that are the Gibbs measures conditioned s.t. $m_\Lambda(\sigma) \subset B_\rho(m^* e^1)$. Here $\rho = c\sqrt{\alpha}/m^*$ is such that the induced measure concentrate on this set. We are interested in the probabilities

$$\mu^{(1,1)}_{\Lambda,\beta,\rho}[\omega]\left(\{\sigma_I = s_I\}\right) \equiv \frac{I\!E_{\sigma_{\Lambda\setminus I}} e^{\frac{1}{2}\beta|\Lambda|\left\|m_\Lambda(s_I,\sigma_{\Lambda\setminus I})\right\|_2^2} 1\!\!1_{\{m_\Lambda(s_I,\sigma_{\Lambda\setminus I})\in B^{(1,1)}_\rho\}}}{I\!E_{\sigma_I} I\!E_{\sigma_{\Lambda\setminus I}} e^{\frac{1}{2}\beta|\Lambda|\left\|m_\Lambda(\sigma_I,\sigma_{\Lambda\setminus I})\right\|_2^2} 1\!\!1_{\{m_\Lambda(s_I,\sigma_{\Lambda\setminus I})\in B^{(1,1)}_\rho\}}} \tag{3.4}$$

Note that $\|m_I(\sigma)\|_2 \leq \sqrt{M}$. Now we can write

$$m_\Lambda(\sigma) = \frac{N}{|\Lambda|} m_{\Lambda\setminus I}(\sigma) + \frac{|I|}{|\Lambda|} m_I(\sigma) \tag{3.5}$$

Then

$$\begin{aligned}
\mathbb{1}_{\{m_\Lambda(s_I,\sigma_{\Lambda\setminus I})\in B_\rho^{(1,1)}\}} &\leq \mathbb{1}_{\{m_{\Lambda\setminus I}(\sigma)\in B_{\rho_+}^{(1,1)}\}} \\
\mathbb{1}_{\{m_\Lambda(s_I,\sigma_{\Lambda\setminus I})\in B_\rho^{(1,1)}\}} &\geq \mathbb{1}_{\{m_{\Lambda\setminus I}(\sigma)\in B_{\rho_-}^{(1,1)}\}}
\end{aligned} \tag{3.6}$$

where $\rho_\pm \equiv \rho \pm \frac{\sqrt{M}|I|}{N}$. For all practical purposes the distinction between ρ, ρ_-, and ρ_+ plays no rôle whatsoever and we will ignore it for the purpose of this review. If we introduce the Laplace-transforms of the measures $\mathcal{Q}$ and $\widetilde{\mathcal{Q}}$

$$\mathcal{L}_{N,\beta,\rho}^{(\mu,s)}[\omega](t) \equiv \int e^{(t,x)} d\mathcal{Q}_{N,\beta,\rho}^{(\mu,s)}[\omega](x)\,, \quad t \in I\!R^{M(N)} \tag{3.7}$$

and

$$\widetilde{\mathcal{L}}_{N,\beta,\rho}^{(\mu,s)}[\omega](t) \equiv \int e^{(t,x)} d\widetilde{\mathcal{Q}}_{N,\beta,\rho}^{(\mu,s)}[\omega](x)\,, \quad t \in I\!R^{M(N)} \tag{3.8}$$

it is not very hard to show (see [BG7] for details) that, with probability tending to 1 rapidly, one has

(i)

$$\begin{aligned}
\mu_{\Lambda,\beta,\rho}^{(1,1)}[\omega]\left(\{\sigma_I = s_I\}\right) = &\frac{\mathcal{L}_{\Lambda/I,\beta,\rho}^{(1,1)}[\omega](\beta'|I|m_I(s_I))}{2^{|I|} I\!E_{\sigma_I} \mathcal{L}_{\Lambda/I,\beta,\rho}^{(1,1)}[\omega](\beta'|I|m_I(\sigma_I))} \\
&+ O(N^{-1/4})
\end{aligned} \tag{3.9}$$

and alternatively

(ii)

$$\begin{aligned}
\mu_{\Lambda,\beta,\rho}^{(1,1)}[\omega]\left(\{\sigma_I = s_I\}\right) = &\frac{\widetilde{\mathcal{L}}_{\Lambda/I,\beta,\rho}^{(1,1)}[\omega](\beta'|I|m_I(s_I))}{2^{|I|} I\!E_{\sigma_I} \widetilde{\mathcal{L}}_{\Lambda/I,\beta,\rho}^{(1,1)}[\omega](\beta'|I|m_I(\sigma_I))} \\
&+ O\left(e^{-O(M)}\right)
\end{aligned} \tag{3.10}$$

where $\beta' \equiv \frac{N}{|\Lambda|}\beta$. Thus the computation of the marginals of the Gibbs measures is reduced to the computation of the Laplace transforms of the induced measures at the random points $t = \sum_{i\in I} s_i \xi_i$, or, in other words, to that of the distribution of the random variables (ξ_i, m), $i \in I$.

Now it is physically very natural that the law of the random variables (ξ_i, m) should determine the Gibbs measures completely. The point is that in a mean field model, the distribution of the spins in a finite set I is determined entirely in terms of the effective mean fields produced by the rest of the system that act on the spins σ_i. These fields are precisely the (ξ_i, m). In a "normal" mean field situation, the mean fields are constant almost surely with respect to the Gibbs measure. In the

Hopfield model with subextensively many patterns, this will also be true, as m will be concentrated near one of the values $\pm m^* e^\mu$ (see [BGP1]). In that case (ξ_i, m) will depend only in a local and very explicit form on the disorder, and the Gibbs measures will inherit this property. In a more general situation, the local mean fields may have a more complicated distribution, in particular they may not be constant under the Gibbs measure, and the question is how to determine this. The approach of the *cavity method* (see e.g. [MPV]) as carried out by Talagrand [T2] consists in deriving this distribution by induction over the volume. [PST] also followed this approach, using however the assumption of "self-averaging" of the order parameter to control errors. Our approach consists in using the detailed knowledge obtained on the measures $\widetilde{\mathcal{Q}}$, and in particular the local convexity to determine a priori the form of the distribution; induction will then only be used to determine the remaining few parameters.

Let us write $\mathbb{E}_{\Phi_N}$ for the expectation with respect to the measures $\widetilde{\mathcal{Q}}_{\Lambda\setminus I,\beta,h}[\omega]$ conditioned on B_ρ and we set $\bar{Z} \equiv Z - \mathbb{E}_{\Phi_N} Z$. We will write $\mathbb{E}_{\xi_I}$ for the expectation with respect to the family of random variables ξ_i^μ, $i \in I$, $\mu = 1, \ldots, M$.

The first step in the computation of our Laplace transform consists in centering, i.e. we write

$$\mathbb{E}_{\Phi_N} e^{\sum_{i\in I} \beta s_i(\xi_i, Z)} = e^{\sum_{i\in I} \beta s_i(\xi_i, \mathbb{E}_{\Phi_N} Z)} \mathbb{E}_{\Phi_N} e^{\sum_{i\in I} \beta s_i(\xi_i, \bar{Z})} \tag{3.11}$$

The most difficult part of the entire analysis is to show that (in a suitable regime of the parameters β, α)

$$\mathbb{E}_{\Phi_N} e^{\sum_{i\in I} \beta s_i(\xi_i, \bar{Z})} \approx e^{\beta^2 \mathbb{E}_{\Phi_N} \|\bar{Z}\|_2^2 \sum_{i\in I} \beta s_i^2} \tag{3.12}$$

i.e. that the centered variables $(\xi_i, \bar{Z})$ are asymptotically independent gaussians with variance $\mathbb{E}_{\Phi_N} \|\bar{Z}\|_2^2$. In our approach, this relies essentially on the fact, proven in [BG4], that in a certain domain of parameters the function Φ is strictly convex on the support of our measures (with large probability), from which (3.12) is easily deduced using the so-called *Brascamp-Lieb inequalities* [BL,HS]. We cannot enter into the details of the proof here and refer the interested reader to [BG7]. Since $s_i^2 = 1$, we see that assuming (3.12), the second term in (3.11) is actually without importance at the moment and the only quantities we need to control are the random variables $(\xi_i, \mathbb{E}_{\Phi_N} Z)$. These are obviously random variables with mean value zero and variance $\|\mathbb{E}_{\Phi_N} Z\|_2$. Moreover, the variables $(\xi_i, \mathbb{E}_{\Phi_N} Z)$ and $(\xi_j, \mathbb{E}_{\Phi_N} Z)$ are uncorrelated for $i \neq j$. Now $\mathbb{E}_{\Phi_N} Z$ has one macroscopic component, namely the first one, while all others are expected to be small. It is thus natural to expect that for large N these variables will actually be close to a sum of a Bernoulli variable $\xi_i^1 \mathbb{E}_{\Phi_N} Z_1$ plus independent gaussians with variance $T_N \equiv \sum_{\mu=2}^M [\mathbb{E}_{\Phi_N} Z_\mu]^2$, and it is indeed possible, although far from trivial, to prove this. For the details see [BG7]. At this stage we fully control the distribution of our random variables up to three unknown parameters $m_1(N) \equiv \mathbb{E}_{\Phi_N} Z_1$, $T_N \equiv \sum_{\mu=2}^M [\mathbb{E}_{\Phi_N} Z_\mu]^2$ and $U_N \equiv \mathbb{E}_{\Phi_N} \|\bar{Z}\|_2^2$. What we have to show is that these quantities converge almost surely and that the limits satisfy the equations of the replica symmetric solution of Amit, Gutfreund and Sompolinsky [AGS].

The proof of this fact relies finally on induction over N, and we will not present any of the details here.

Proposition 3.3: *There exists an open set of the parameters α, β for which the following holds: For any finite $I \subset I\!N$*

$$\mu_{\Lambda,\beta,\rho}^{(1,1)}\left(\{\sigma_I = s_I\}\right) \stackrel{\mathcal{D}}{\to} \prod_{i\in I} \frac{e^{\beta s_i\left[m_1\bar{\xi}_i^1 + g_i\sqrt{\alpha r}\right]}}{2\cosh\left(\beta\sigma_i\left[m_1\bar{\xi}_i^1 + g_i\sqrt{\alpha r}\right]\right)} \tag{3.13}$$

where the convergence holds in law with respect to the measure $I\!P$. $\{g_i\}_{i\in I\!N}$ is a family of i.i.d. standard normal random variables and $\{\bar{\xi}_i^1\}_{i\in I\!N}$ are independent Bernoulli random variables, independent of the g_i and having the same distribution as the variables ξ_i^1. Moreover the constants r, m_1, q are nonzero solutions of the system of equations

$$\begin{aligned} m_1 &= \int d\mathcal{N}(g)\tanh(\beta(m_1 + \sqrt{\alpha r}g)) \\ q &= \int d\mathcal{N}(g)\tanh^2(\beta(m_1 + \sqrt{\alpha r}g)) \\ r &= \frac{q}{(1-\beta+\beta q)^2} \end{aligned} \tag{3.14}$$

Remark: Equations (3.14) determine the replica symmetric solution of [AGS]. The domain of parameters where our proof works is essentially bounded by the three lines $\alpha = 0$, $\alpha \leq c(m^*)^4$, $\beta \geq c\alpha$. The last curve is due to our convexity requirement. This curve does not seem optimal, as the results of Talagrand (that are not exactly the same as ours) are obtained on a larger domain.

Remark: Note that the Gibbs measures converge in law, and not almost surely, if $\alpha > 0$. This may appear as a rather unusual feature, however it should be seen as rather natural in the context of strongly disordered systems. An extensive discussion of this issue is to be found in [NS] and also in [BG7].

4. Beyond mean field: dilute and Kac models

From the point of view of statistical mechanics, the models discussed so far represent convenient simplifications, but miss a crucial feature of the local structure of realistic systems. In fact, in realistic models of statistical mechanics the interaction between the microscopic components of the models depends rather strongly on their distance, and the most common models allow interactions only between nearest neighbors. In the context of networks, what represent a realistic modeling of the network geometry is certainly less clear, however, it is obvious that the idealization of a complete connectivity without any dependence of some kind of "distance" is realistic at best in rather small systems. E.g., in the brain the number of neurons is of the order of 10^{10}, while each neuron is directly connected to at most about $10^5 - 10^6$ others. One may describe such situations in various ways that will certainly depend on the details of the system studied. I will not enter into these modeling problems, but just mention two extreme cases: First, one may conceive the underlying network as a *random graph*, modeled by a family of i.i.d. random variables ϵ_{ij} that take the values 0 and 1 with probabilities p and $1-p$, respectively. Note that it may be suitable to allow p to depend on the system size N. The Hopfield model on such a random graph is then simply defined by the Hamiltonian

$$H_N(\sigma)[\xi,\epsilon] \equiv \frac{1}{2pN} \sum_{i,j=1}^{N} \sum_{\mu=1}^{M} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j \epsilon_{ij} \tag{4.1}$$

As long as one is not interested in very dilute graphs, more precisely as long as one requires $\lim_{N\uparrow\infty} pN = \infty$, and as long as one stays in the "retrieval phase", meaning here that $M \leq \alpha pN$, this model exhibits a nice homogenization property, i.e. the Hamiltonian (4.1) is uniformly very close to the one obtained by averaging over the variables ϵ. Indeed, it was proven in [BG2] that

$$\sup_{\sigma\in\mathcal{S}^N} \left| H_{N,M}[\xi,\epsilon](\sigma) - I\!E_\epsilon \left[H_{N,M}[\xi,\epsilon](\sigma) \right] \right| \leq cN\sqrt{\frac{M}{xN}} \tag{4.2}$$

with probability tending to one. Here $I\!E_\epsilon$ denotes the expectation w.r.t. the random variables ϵ_{ij}.

Thus for small load the dilute model behaves like the model on the homogeneous graph, while the analysis in the strong load case seems hopelessly difficult. More interesting things happen if we study instead of the random graph a regular lattice. A particularly nice situation arises when we consider the basic lattice $\mathbb{Z}^d$ and introduce long-range interactions. To be specific, e.g. choose the function $J_\gamma(i-j) \equiv \gamma J\left(\gamma^d |i-j|\right)$, and

$$J(x) = \begin{cases} 1, & \text{if } \|x\|_\infty \leq 1/2 \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

(Note that other choices for the function $J(x)$ are possible. They must satisfy the conditions $J(x) \geq 0$, $\int dx J(x) = 1$, and must decay rapidly to zero on a scale of order unity.

The interaction between two spins at sites i and j will be chosen as

$$-\frac{1}{2} \sum_{\mu=1}^{M(\gamma)} \xi_i^\mu \xi_j^\mu J_\gamma(i-j) \sigma_i \sigma_j \tag{4.4}$$

and the Hamiltonian will be, for a finite subset $\Lambda \subset \mathbb{Z}^d$

$$H_{\gamma,\Lambda}(\sigma)[\xi] = -\frac{1}{2} \sum_{i,j\in\Lambda} \sum_{\mu=1}^{M(\gamma)} \xi_i^\mu \xi_j^\mu J_\gamma(i-j) \sigma_i \sigma_j \tag{4.5}$$

Note that we anticipate, again, that the number of patterns can only be proportional to the local connectivity, i.e. γ^{-d}.

Such Kac-models have a long tradition in statistical mechanics. They were introduced in the sixties by M. Kac [KUH] to provide a microscopic model for a rigorous derivation of the van der Waals-Maxwell theory of the liquid vapor transition. Recently, there was an upsurge of interest in these models mainly in connection with static and dynamic aspects of phase separation problems. Kac models are also natural candidates to study disordered systems, and in particular the question to what extend results in mean field theories are relevant for finite dimensional system. Indeed, the Kac version of the Hopfield model had been introduced already by Figotin and Pastur [FP3]. More recently, a number of different disordered Kac models have been studied, notably the site diluted model [Bod], and a spin glass [B].

In the Kac model one defines finite volume measures with boundary conditions σ_{Λ^c} by assigning to each $\sigma_\Lambda \in \mathcal{S}_\Lambda$ the mass

$$\mu_{\beta,\gamma,\Lambda}^{\sigma_{\Lambda^c}}[\xi](\sigma_\Lambda) \equiv \frac{1}{Z_{\beta,\gamma,\Lambda}^{\sigma_{\Lambda^c}}[\xi]} e^{-\beta[H_{\gamma,\Lambda}[\xi](\sigma_\Lambda)+W_{\gamma,\Lambda}[\xi](\sigma_\Lambda,\sigma_{\Lambda^c})]} \tag{4.6}$$

where $Z_{\beta,\gamma,\Lambda}^{\sigma_{\Lambda^c}}[\xi]$ is the *partition function* and

$$W_{\gamma,\Lambda}[\xi](\sigma_\Lambda,\sigma_{\Lambda^c}) = -\sum_{i\in\Lambda}\sum_{j\in\Lambda^c}\sum_{\mu=1}^{M(\gamma)} \xi_i^\mu \xi_j^\mu J_\gamma(i-j)\sigma_i\sigma_j \tag{4.7}$$

represents the influence of the boundary conditions.

The classical type of result in Kac models is the statement that the large deviation rate function converges to the convex hull of that of the corresponding mean field model ("Lebowitz-Penrose theorem" [LP]). On this level, there is still no effect either of the boundary conditions nor of the dimensionality of the lattice visible. In our case, define

$$m_\Lambda(\sigma) \equiv \frac{1}{|\Lambda|}\sum_{i\in\Lambda}\xi_i\sigma_i \tag{4.8}$$

and set, as in (2.12)

$$f_{\gamma,\beta,\Lambda,M(\gamma),\rho}^{I}(x)[\xi] \equiv -\frac{1}{\beta|\Lambda|}\ln \mu_{\beta,\gamma,\Lambda}^{\sigma_{\Lambda^c}}[\xi]\left(\{m_\Lambda^I(\sigma)\in B_\rho(x)\}\right) \tag{4.9}$$

In [BGP2] we proved the following Theorem:

Theorem 4.1: *Assume that $M(\gamma)$ satisfies $\lim_{\gamma\downarrow 0} M(\gamma) = +\infty$ and $\lim_{\gamma\downarrow 0}\gamma M(\gamma) = 0$. Then, for any β, and any finite subset I, almost surely,*

$$\lim_{\rho\downarrow 0}\lim_{\gamma\downarrow 0}\lim_{\Lambda\uparrow \boldsymbol{Z}} f_{\gamma,\beta,\Lambda,M(\gamma),\rho}^{I}(x)[\xi] = conv F_\beta^{Hopf,I}(x) \tag{4.10}$$

where $F_\beta^{Hopf,I}(x)$ is the rate function of the Hopfield model and convF denotes the convex hull of the function F.

Much more interesting than such a global large deviation result, however, is the possibility in Kac models to study the distribution of "local overlap profiles". By this we mean the following: Let us introduce a scale ℓ much bigger than one and much smaller than $1/\gamma$. Consider blocks $b_x \subset \Lambda$ defined e.g. by $b_x \equiv \{i \in \mathbb{Z}^d | \, \|i - \ell x\|_\infty \leq \ell/2\}$. We will abuse notation and identify b_x with its label x. We then set

$$m_x(\sigma) \equiv \frac{1}{\ell^d} \sum_{i \in x} \xi_i \sigma_i \tag{4.11}$$

The key point is that with large probability the Hamiltonian of out model is close to a function of the variables $m_x(\sigma)$.

To see this, let us introduce the function

$$E^\ell_{\gamma,\Lambda}(m) \equiv -\frac{1}{2}(\gamma\ell)^d \sum_{(x,y)\in\Lambda\times\Lambda} J_{\gamma\ell}(x-y)(m(x), m(y)) \tag{4.12}$$

We write

$$H_{\gamma,\Lambda}(\sigma_\Lambda) = \gamma^{-d} E^\ell_{\gamma,\Lambda}(m_\ell(\sigma)) + \Delta H^\ell_{\gamma,\Lambda}(\sigma_\Lambda) \tag{4.13}$$

We have exhibited a γ^{-d} factor in front of $E^\ell_{\gamma,\Lambda}(m_\ell(\sigma))$ to make clear the scaling involved in the problem.

The following lemma is the basic result to control the block spin approximation.

Lemma 4.2: *There exists a finite constant c such that for all $\delta > 0$*

$$\begin{aligned} &\mathbb{P}\left[\sup_{\sigma\in\mathcal{S}_\Lambda} \frac{\gamma^d}{|\Lambda|} |\Delta H_\Lambda(\sigma)| \geq \gamma\ell(\gamma)(c+\delta) + c\gamma M(\gamma)\right] \\ &\leq 16 e^{-\delta \frac{|\Lambda|}{\gamma^d}} \end{aligned} \tag{4.14}$$

The proof of (4.14) can be found in [BGP2], for the case $d = 1$, but one readily sees that it holds in arbitrary dimensions. Let us mention the important fact that since the parameter $M(\gamma)$, $\ell(\gamma)$ and $L(\gamma)$ are chosen in such a way that $\alpha(\gamma) \equiv \gamma M(\gamma) \downarrow 0$, $\gamma\ell(\gamma) \downarrow 0$ and $\gamma L(\gamma) \downarrow 0$, it follows from (2.9) and (2.10) that with $\mathbb{P}$-probability very close to one the errors of the block spin approximations is of order a small parameter times the volume (expressed in the *macroscopic* unit). This will allow us to control only the Gibbs-probability of cylindrical events that have a basis with a uniformly bounded diameter. The main problem is to obtain estimates for the infinite volume Gibbs measure.

Technically, a great deal of the difficulties in the study of Kac models arise from the problem of controlling the effect of the error terms ΔH. However, to understand what is basically going on, it is useful to ignore them for the time being, and this is the attitude we will take in this review.

To be precise, we define the block approximation of the Gibbs measure

$$\bar{\mu}_{\Lambda,\gamma,\beta}(\sigma)[\xi] \equiv \frac{1}{\bar{Z}_{\Lambda,\gamma,\beta}[\xi]} I\!\!E_\sigma \exp\left(-\beta E^{\ell}_{\gamma,\Lambda}(m(\sigma))\right) \tag{4.15}$$

(We do not put any boundary conditions as we are only discussing qualitative features; in [BGP2] the interested reader will find an extensive discussion on this matter). Now since E depends only on $m(\sigma)$, one can as in the mean field models give a rather explicit representation of the measure induced on the local overlaps, i.e.

$$\bar{\mu}_{\Lambda,\gamma,\beta}(\{m_x(\sigma) = m_x, x \in \Lambda\})[\xi] = \frac{1}{\bar{Z}_{\Lambda,\gamma,\beta}[\xi]} \exp\left(-\beta E^{\ell}_{\gamma,\Lambda}(m)\right) I\!\!E_\sigma \prod_{x\in\Lambda} \mathbb{1}_{\{m_x(\sigma)=m_x\}} \tag{4.16}$$

Using that $(m_x, my) = -\frac{1}{2}\|m_x - m_y\|_2^2 + \frac{1}{2}\|m_x\|_2^2 + \frac{1}{2}\|m_y\|_2^2$, this can be re-written in the form (we ignore boundary terms or think of periodic boundary conditions)

$$\bar{\mu}_{\Lambda,\gamma,\beta}(\{m_x(\sigma) = m_x, x \in \Lambda\})[\xi] = \frac{1}{\bar{Z}_{\Lambda,\gamma,\beta}[\xi]} \exp\left(-\ell^d \beta \mathcal{F}^{\ell}_{\Lambda,\gamma}(m)\right) \tag{4.17}$$

with

$$\mathcal{F}^{\ell}_{\Lambda,\gamma}(m) \equiv \frac{1}{4}\sum_{x,y\in\Lambda} J_{\gamma\ell}(x-y)\|m_x - m_y\|_2^2 + \sum_{x\in L} f_{x,\beta}(m_x) \tag{4.18}$$

where

$$f_{x,\beta}(m_x) = \frac{1}{2}\|m_x\|_2^2 - \frac{1}{\beta\ell^d}\ln I\!\!E_\sigma \mathbb{1}_{\{m_x(\sigma)=m_x\}} \tag{4.19}$$

is nothing but the free energy functional of the normal Hopfield model in the set x.

The main feature here is that (formally), this representation gives a large deviation type representation for the law of the local overlaps with ℓ playing the rôle of the rate function. In other, more physically inspired word, integrating out the spin variables for fixed values of the local overlaps, we obtain a new model with local spin space $[-1,1]^M$ at inverse temperature $\beta\ell^d$ (i.e. at very low temperature) with an attractive interaction of range $1/\gamma\ell$ and with a local a priori spin distribution proportional to $e^{-\beta\ell^d f_{x,\beta}}$. It is important to note that as random variables the $f_{x,\beta}$ are independent. As a consequence, we may expect that the typical overlaps will tend to minimize the functional $\mathcal{F}$. Now, the local part of this functional wants the local overlap to minimize $f_{x,\beta}$, while the quadratic part wants to align all overlaps. If β, ℓ and M are such that we are in the situation of our results from Section 3, we know that $f_{x,\beta}$ has $2M$ "lowest" minima which are almost degenerate. However, as these minima are not totally degenerate, but show some random fluctuations, these will create some local bias towards one of them. The main question of interest is then on which scale the competition between the random fluctuations and the quadratic interaction will equilibrate. This situation is similar to the low temperature random field Ising model.

Rigorous results on this question, for the full model, have been obtained only in the one dimensional case in [BGP97]. They concern the situation where $lim_{\gamma\downarrow 0}\gamma M(\gamma)=0$. essentially, the results can be summarized as follows:

Quasitheorem: *Assume that* $\lim_{\gamma\downarrow 0}\gamma M(\gamma)=0$. *Then there is a scale* $L\ll\gamma^{-1}$ *such that with* $I\!\!P$*-probability tending to one (as* $\gamma\downarrow 0$*) the following holds:*

(i) In any given macroscopic finite volume in any configuration that is "typical" with respect to the infinite volume Gibbs measure, for "most" blocks r*,* $m_L(r,\sigma)$ *is very close to one of the values* $\pm a(\beta)e^{\mu}$ *(we will say that* $m_L(u,\sigma)$ *is "close to equilibrium").*

(ii) In any macroscopic volume Δ *that is small compared to* γ^{-1}*, in a typical configuration, there is at most one connected subset* J *(called a "jump") with* $|J|\sim\frac{1}{\gamma L}$ *on which* m_L *is not close to equilibrium. Moreover, if such a jump occurs, then there exist* (s_1,μ_1) *and* (s_2,μ_2)*, such that for all* $u\in\Delta$ *to the left of* J*,* $m_L(u,\sigma)\sim s_1a(\beta)e^{\mu_1}$ *and for all* $u\in\Delta$ *to the right of* J*,* $m_L(u,\sigma)\sim s_2a(\beta)e^{\mu_2}$

For a more precise formulation, and the, unfortunately quite tedious proofs, we refer the reader to the original paper [BGP4].

5. Historical remarks

We would like to give a brief account of the main developments in the study of the Hopfield model that have lead to our present status of knowledge. This account will certainly be biased, and we excuse ourselves in advance for omissions and oversights which will reflect only our own limited state of knowledge. In particular, we will essentially concentrate only on the mathematically rigorous results and mention others only in as far as this is indispensable for the understanding.

The early history of the model goes back to the roots of spin glass theory. In spin glass models, the basic idea was to replace the deterministic Rudderman-Kittel-Koruda-Yoshida (RKKY) interaction that is of the form $\frac{\cos(k_F\cdot(x-y))}{|x-y|^3}$, where the Fermi-momentum k_F is incommensurate with the lattice vectors, by some random interaction (of short or long range type) that would be easier to treat and would capture the supposed main feature of the rather irregular sign-changing RKKY-interaction. The choice of independent random J_{ij} that was made in the Edwards-Anderson [EA] and the Sherrington-Kirkpatrick [SK] models proved difficult, and so people tried other, possibly simpler solutions. Mattis [Ma] proposed $J_{ij}=\epsilon_i\epsilon_j$, with independent $\epsilon_i=\pm 1$. This was soon found too trivial, for a gauge transformation could immediately reduce the system to the original ferromagnetic Ising model. Luttinger [Lu] proposed the simple amendment $J_{ij} = \epsilon_i^1 \epsilon_j^1 + \epsilon_i^2 \epsilon_j^2$, which at least was not totally trivial (although this model has some particular features that make it almost trivial: It factors into two ferromagnetic Ising models on two random sublattices with no interaction between the two subsystems). In 1977, Figotin and Pastur [FP1] generalized this model in several ways: instead of two summands, they allowed an arbitrary number p, there were additional weights a_k for each summand, and the distribution of the random variables was allowed to be more general. They also considered, in a separate paper [FP2], the quantum version of this model. However, the number of terms p was kept a fixed, finite parameter. Using the Hubbard-Stratonovich transformation, they got a fairly complete description of the main features of the model that was mathematically essentially rigorous. In a later paper [FP3] they also introduced the Kac version of their model and proved that the free energy converges to the mean-field free energy in the Lebowitz-Penrose limit. These papers all appeared in Soviet journals and seem to have received very little attention; at least we did not find them quoted in the Hopfield literature until the early 90's.

In 1982 John Hopfield introduced the same type of models in the context of neural networks [Ho]. However, there was one notable difference: Since Hopfield was interested in the memory capacity of his model, he investigated (numerically) the behaviour of models with various sizes and various numbers, p, of stored patterns. He observed the striking phenomenon that the number of patterns that could be stored is proportional to the size, and that there is a sharp critical ratio of about 0.14 above which the networks no longer retrieves the stored information.

Very soon after the publication of Hopfield's paper the investigation of the statistical mechanics of the Hopfield model by physicists from spin glass theory started. One can class these investigations into two groups: The first was based on the non-rigorous replica method that had just been successfully applied in the SK-model by Parisi and co-authors. Here, Amit, Gutfreund and Sompolinsky [AGS2] obtained a strikingly complete picture of the properties of the model as function of the temperature and the load parameter $\alpha = M/N$ that explained in a quantitatively precise way the numerical results of Hopfield. This success lead to an enormous and continuing development which we cannot follow in this brief note. The second was the approach to these models on the basis of large deviation theory and is the precursor of the work exposed in this review. This work was largely rigorous (or could be made so) mathematically, but limited to the case $M < \infty$ independent of N, already studied in [FP1]. Some of the more important contributions are [vHvEC,vH1,GK,vHGHK,vH2,AGS1,JK,vEvHP]. The development culminated in the rather general paper on large deviations in such systems by Comets [Co].

The next step to extend this approach was taken by Koch and Piasko [KP] (see also [vEvH]).

They succeeded to extend the large deviation approach to the case there $M \sim \ln N$. This paper again did not receive the attention it deserved. The next step in the direction of increasing N was taken only in 1992. In two papers by Koch [K] and Shcherbina and Tirozzi [ST], it was shown that the free energy in the general case $M = o(N)$ converges to that of the Curie-Weiss model (more precisely, Koch showed the convergence of the average free energy, [ST] the convergence in probability, while [BG2] observed that the proof of Koch could be easily modified to give the almost sure convergence). Notably, in these papers the large deviations techniques were abandoned in favor of the original Hubbard-Stratonovich approach used in [FP1]. In these papers, the question of Gibbs states was not touched. This problem was tackled in [BGP1] where under the same assumption $M = o(N)$ the limiting induced measures were constructed. The full extension of the finite M results to this situation was completed with the large deviation principle only in [BG3]. The paper [BGP1] contained already first results for the case $M = \alpha N$, with small α. The complete proof that in this case there exist (at least) one limiting Gibbs measure for each pattern was only given some time later in [BGP3]; this picture was further cleaned in the paper [BG4].

Another development started in the paper by Pastur et al. [PST] in 92. This was an attempt to obtain the results of the replica method via a rigorous application of what is know as the cavity method, i.e. induction over the volume. This attempt was partially successful. They found that the validity of the replica symmetric solution appeared to be linked to the self-averaging of the Edwards-Anderson order parameter, but neither could be established in any non-trivial regime of parameters. There appeared to be also some gaps in the arguments of the proof. This idea was taken up in 1996 by Talagrand who actually proved by induction that the replica symmetric solution holds in some region of the β, α plane. Following this, another proof of this fact was given in [BG5] that used some convexity results established in [BG4] and the Brascamp-Lieb inequalities. In [BG7] this was extended to a systematic analysis of the structure of the limiting Gibbs states.

A somewhat related development concerns the central limit theorem for the overlap distribution in one of the extremal states of the Hopfield model. This problem was apparently first considered in a paper by Gentz [G1] in 1995. Here the case M bounded was solved. Using techniques from [BG4] in [G2,G3] the condition on M could be relaxed to $p = o\left(\sqrt{N}\right)$. Finally, using Brascamp-Lieb inequalities and the convexity results from [BG4], the CLT could be established under just the hypothesis $M = o(N)$ in [BG6].

Let us also mention a somewhat independent line of research that concerns just the structure of the local minima of the Hamiltonian itself. This started essentially with a paper by McEliece et al. [MPRV] there it was argued that all patterns ξ^μ should be local minima of the Hamiltonian if $M(N) < \frac{N}{2\ln N}$. This was proven rigorously (with a slightly worse constant than 2) by Martinez [Mar]. In 1988 Newman [N1] proved that local minima *near* the patterns surrounded by extensive energy barriers exist at least as long as $M(N) \leq \alpha_c N$, with $\alpha_c \geq 0.055$. The lower bound on α_c was subsequently improved by Loukianova [Lo1] to 0.071 and Talagrand [T2] to some unspecified value. Similar results were also proved for variants of the Hopfield model: Newman [N1] himself treated the model with p-spin interactions, the q-state Potts-Hopfield model was considered by Ferrari, Martinez and Picco [FMP], the dilute Hopfield model by the present authors [BG1]. The only rigorous results on the Hopfield model with correlated patterns are also estimates of this kind and were obtained recently by Löwe [L1]. Finally we mention more detailed results on the domains of attraction of these minima by Komlos and Füredi [KF] and in more refined form by Burshtein [Bu]. A notoriously difficult problem is to get converse results, i.e. to show that beyond a certain α_c, there are no minima in a certain neighborhood of the patterns. There is only one quite recent result on this question due to Loukianova [Lo2] who could show that for for all $\alpha > 0$, there is a $r(\alpha) > 0$ such that the balls of radius $r(\alpha)$ around each pattern are free of local minima. However, the estimate on $r(\alpha)$ obtained is quite poor, in fact it was only shown that $\liminf_{\alpha\uparrow\infty} r(\alpha) \geq 0.05$

Finally, let us point to extensions ton non-mean field models. The Kac-version of the model had already been introduced in 1980 by Pastur and Figotin [PF3]. They also showed that with finitely many patterns, the free energy of this model converged to that of the Curie-Weiss model. Apparently there was no work on this model until 1994 when we proved with Picco [BGP2] the Lebowitz-Penrose theorem under the condition that $\gamma M(\gamma) \downarrow 0$.

References

[A] D.J. Amit, "Modelling brain function", Cambridge University Press, Cambridge (1989).

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

[AGS2] D.J. Amit, H. Gutfreund and H. Sompolinsky, "Statistical mechanics of neural networks near saturation", Ann. Phys. **173**, 30-67 (1987).

[B] A. Bovier, "The Kac version of the Sherrington-Kirkpatrick model at high temperatures", submitted to J. Stat. Phys. (1997).

[BG1] A. Bovier and V. Gayrard, "Rigorous bounds on the storage capacity for 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**, 643-664 (1993).

[BG3] A. Bovier and V. Gayrard, "An almost sure large deviation principle for the Hopfield model", Ann. Probab **24**, 1444-1475 (1996).

[BG4] A. Bovier and V. Gayrard, "The retrieval phase of the Hopfield model, A rigorous analysis of the overlap distribution", Prob. Theor. Rel. Fields **107**, 61-98 (1995).

[BG5] A. Bovier and V. Gayrard, "Hopfield models as generalized random mean field models", in "Mathematical aspects of spin glasses and neural networks", A.Bovier and P. Picco, Eds., Progress in Probability Vol. 41, Birkhäuser, Boston, (1997).

[BG6] A. Bovier and V. Gayrard, "An almost sure central limit theorem in the Hopfield model", Markov Proc. Rel. Fields (1997).

[BG7] A. Bovier and V. Gayrard, "Metastates in the Hopfield model in the replica symmetric regime", submitted to MPAG (1997).

[BGP1] A. Bovier, V. Gayrard, and P. Picco, "Gibbs states of the Hopfield model in the regime of perfect memory", Prob. Theor. Rel. Fields **100**, 329-363 (1994).

[BGP2] A. Bovier, V. Gayrard, and P. Picco, "Large deviation principles for the Hopfield model and the Kac-Hopfield model", Prob. Theor. Rel. Fields **101**, 511-546 (1995).

[BGP3] A. Bovier, V. Gayrard, and P. Picco, "Gibbs states of the Hopfield model with extensively many patterns", J. Stat. Phys. **79**, 395-414 (1995).

[BGP4] A. Bovier, V. Gayrard, and P. Picco, "Distribution of overlap profiles in the one-dimensional Kac-Hopfield model", Commun. Math. Phys. **186**, 323-379 (1997).

[BL] H.J. Brascamp and E.H. Lieb, "On extensions of the Brunn-Minkowski and Pékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation", J. Funct. Anal. **22**, 366-389 (1976).

[BP] A. Bovier and P. Picco (Eds.), "Mathematical aspects of spin glasses and neural networks", Progress in Probability Vol. 41, Birkhäuser, Boston (1997).

[Bod] Th. Bodineau, "Interface for a one-dimensional random Kac potential", preprint (1996).

[Bu] D. Burshtein, "Nondirect convergence radius and number of iterations in the Hopfield autoassociative memory", IEEE-Trans.Inform. Theory **40**, 838-847 (1994)

[CGOV] M. Cassandro, A. Galves, E. Olivieri, and M.E. Vares, "Metastable behaviour of stochstic dynamics: A pathwise approach", J. Stat. Phys. **35**,603-634 (1984).

[Co] F. Comets, "Large deviation estimates for a conditional probability distribution. Applications to random Gibbs measures", Probab. Theor. Rel. Fields **80**, 407-432 (1989).

[DHS] E. Domany, J.L. van Hemmen, and K. Schulten (Eds.), "Models of neural networks", Springer Verlag, Berlin (1991).

[E1] M. Eckhoff, "Scharfe Tunnelabschätzungen im Curie-Weiss und Hopfield Modell" Diploma thesis, TU-Berlin, 1996.

[EA] S.F. Edwards and P.W. Anderson, "Theory of spin glasses", J. Phys. **F 5**, 965-974 (1975).

[vEvHP] A.C.D. van Enter, J.L. van Hemmen and C. Pospiech, "Mean-field theory of random- site q-state Potts models", J. Phys. **A 21**, 791-801 (1988).

[FMP] P. Ferrari, S. Martínez, and P. Picco, "A lower bound on 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 spin systems", Theor. Math. Phys. **51**, 564-569 (1982).

[Ge1] B. Gentz, "An almost sure central limit theorem for the overlap parameters in the Hopfield model", Stochastic Process. Appl. **62**, 243-262 (1996).

[Ge2] B. Gentz, "A central limit theorem for the overlap in the Hopfield model", Ann. Probab. **24**, 1809-1841 (1996).

[Ge3] B. Gentz, "A central limit theorem for the overlap in the Hopfield model", Ph. D. Thesis, University of Zürich (1996).

[GK] D. Grensing and K. Kühn, "On classical spin-glass models", J. Physique **48**, 713-721 (1987).

[GM] E. Golez and S. Martínez, "Neural and automata networks", Kluwer Academic Publ., Dodrecht (1990).

[HKP] J. Hertz, A. Krogh, and R. Palmer, "Introduction to the theory of neural computation", Addison-Wesley, Redwood City (1991).

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

[HS] B. Helffer and J. Sjöstrand, "On the correlation for Kac-like models in the convex case", J.

Stat. Phys. **74**, 349-409 (1994).

[JK] J. Jedrzejewski and A. Komoda, "On equivalent-neighbour, random-site models of disordered systems", Z. Phys. **B 63**, 247-257 (1986).

[vH1] J.L. van Hemmen, "Equilibrium theory of spin-glasses: mean-field theory and beyond", in "Heidelberg colloquium on spin glasses", Eds. J.L. van Hemmen and I.Morgenstern, LNP 192 Springer, Berlin-Heidelberg-New York (1983).

[vH2] J.L. van Hemmen, "Spin glass models of a neural network", Phys. Rev. A **34**, 3435-3445 (1986).

[vHGHK] J.L. van Hemmen, D. Grensing, A. Huber and R. Kühn, "Elementary solution of classical spin-glass models, Z. Phys. **B 65**, 53-63 (1986).

[vHvE] J.L. van Hemmen and A.C.D.van Enter, "Chopper model for pattern recognition", Phys.Rev. **A 34**, 2509-2512 (1986).

[vHvEC] J.L. van Hemmen, A.C.D. van Enter, and J. Canisius. "On a classical spin-glass model", Z. Phys. **B 50**, 311-336 (1983).

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

[KF] Komlos and Füredi, "Convergence results in an autoassociative memory model", Neural Networks **1**, 239-250 (1988).

[KP] H. Koch and J. Piasko, "Some rigorous results on the Hopfield neural network model", J. Stat. Phys. **55**, 903-928 (1989).

[KUH] M. Kac, G. Uhlenbeck, and P.C. Hemmer, "On the van der Waals theory of vapour-liquid equilibrium. I. Discussion of a one-dimensional model" J. Math. Phys. **4**, 216-228 (1963); "II. Discussion of the distribution functions" J. Math. Phys. **4**, 229-247 (1963); "III. Discussion of the critical region", J. Math. Phys. **5**, 60-74 (1964).

[LP] J. Lebowitz and O. Penrose, "Rigorous treatment of the Van der Waals Maxwell theory of the liquid-vapour transition", J. Math. Phys. **7**, 98-113 (1966).

[L1] M. Löwe, "On the storage capacity of the Hopfield model", in "Mathematical aspects of spin glasses and neural networks", A.Bovier and P. Picco, Eds., Progress in Probability Vol. 41, Birkhäuser, Boston, (1997).

[Lo1] D. Loukianova, "Capacité de mémoire dans le modèle de Hopfield", C.R.A.S. Paris **t. 318, Série 1**, 157-160 (1994).

[Lo2] D. Loukianova, "Lower bounds on the restitution error in the Hopfield model", Prob. Theor.

Rel. Fields **107**, 161-176 (1997).

[Lu] J.M. Luttinger, “Exactly Soluble Spin-Glass Model”, Phys.Rev. Lett. **37**, 778-782 (1976).

[Mar] S. Martinez, “Introduction to neural networks”, preprint, Temuco, (1992).

[Ma] D.C. Mattis, “Solvable spin system with random interactions”, Phys. Lett. **56A**, 421-422 (1976).

[McE] 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).

[MR] B. Müller and J. Reinhardt, “Neural networks: an introduction”, Springer Verlag, Berlin (1990).

[MS] V.A. Malyshev and F.M. Spieksma, “Dynamics of binary neural networks with a finite number of patterns”. Part 1: General picture of the asynchronous zero temperature dynamics”, MPEJ **3**, 1-36 (1997).

[N] Ch.M. Newman, “Memory capacity in neural network 00models: Rigorous results”, Neural Networks **1**, 223-238 (1988).

[NS] Ch.M. Newman and D.L. Stein, “Non-mean-field behaviour in realistic spin glasses”, Phys. Rev. Lett. **76**, 515-518 (1996); “Spatial inhomogeneity and thermodynamic chaos”, Phys. Rev. Lett. **76**, 4821-4824 (1996); “Thermodynamic chaos and the structure of short range spion glasses”, in “Mathematical aspects of spin glasses and neural networks”, A.Bovier and P. Picco, Eds., Progress in Probability Vol. 41, Birkhäuser, Boston, 1997.

[OS] E. Olivieri and E. Scoppolla, “Markov chains with exponentially small transition probabilities: First exit problem from a general domain. I. The reversible case”, J. Stat. Phys. **79**, 613 (1995).

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

[PST] L. Pastur, M. Shcherbina, and B. Tirozzi, “The replica symmetric solution without the replica trick for the Hopfield model”, J. Stat. Phys. **74**, 1161-1183 (1994).

[Ro] R.T. Rockafellar, “Convex analysis”, Princeton University Press, Princeton (1970).

[SK] D. Sherrington and S. Kirkpatrick, “Solvable model of a spin glass”, Phys. Rev. Lett. **35**, 1792-1796 (1972).

[ST] M. Shcherbina and B. Tirozzi, “The free energy for a class of Hopfield models”, J. Stat. Phys. **72**, 113-125 (1992).

[T1] M. Talagrand, “A new look at independence”, Ann. Probab. **24**, 1-34 (1996).

[T2] M. Talagrand, “Rigorous results for the Hopfield model with many patterns”, preprint (1996), to appear in Probab. Theor. Rel. Fields.

[Tu] T. Turova, “Analysis of a biologically plausible neural network via an hourglass model”, Markov Proc. Rel. Fields **2**, 487-510 (1996).