GIBBS STATES OF THE HOPFIELD MODEL WITH EXTENSIVELY MANY PATTERNS[#]

Anton Bovier [1]

Institut für Angewandte Analysis und Stochastik
Mohrenstrasse 39, D-10117 Berlin, Germany

Véronique Gayrard[2] **and Pierre Picco**[3]

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

Abstract: We consider the Hopfield model with $M(N) = \alpha N$ patterns, where N is the number of neurons. We show that if α is sufficiently small and the temperature sufficiently low, then there exist disjoint Gibbs states for each of the stored patterns, almost surely with respect to the distribution of the random patterns. This solves a problem left open in previous work [BGP1]. The key new ingredient is a self averaging result on the free energy functional. This result has considerable additional interest and some consequences are discussed. A similar result for the free energy of the Sherrington-Kirkpatrick model is also given.

Keywords: Hopfield model, Gibbs states, self-averaging, spin glasses

---

[#] Work partially supported by the Commission of the European Communities under contract No. SC1-CT91-0695

[1] e-mail: bovier@iaas-berlin.d400.de

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

[3] e-mail: picco@cpt.univ-mrs.fr

1. Introduction

Recently, considerable progress has been made towards a rigorous understanding of some of the main thermodynamic properties of the so-called Hopfield model [Ho]. This model had been introduced first by Figotin and Pastur [FP1,FP2] as a simple soluble model of a spin glass, but has enjoyed, after its re-interpretation as a model for an autoassociative memory by Hopfield [Ho] an enormous success. Notably, the application of the *replica-method*, familiar to theoretical physicists for many years from work in particular on the Sherrington-Kirkpatrick model [SK, MPV], by Amit et al [AGS] has allowed for the first time for an analytical reproduction of earlier findings from numerical simulations. In spite of the success of this method, it is, we hope not only from the point of view of mathematics, somewhat unsatisfactory as it involves a number of ad hoc procedures which cannot, up to now, be interpreted within the framework of rigorous mathematics. Moreover, this method computes various quantities in a fictitious replica-space which makes the *physical* interpretation of what is going on somewhat awkward; in particular, this method can at best compute certain quenched averages of correlations functions, but is intrinsically inadequate to obtain results that are *typically* (in the sense of the probabilistic term *almost sure*) true in a given fixed realization of the disorder.

Over the last year, however, some mathematically rigorous results on this model have been obtained (for a summary see e.g.[BG2]), albeit under fairly stringent conditions on the parameters of the model, notably the ratio $\alpha(N)$ of the number $M(N)$ of stored *patterns* to the system size N. Under the condition that this ratio tends to zero as N tends to infinity, the complete set of all limiting Gibbs measures could be constructed [BGP1]. While these results are already quite difficult to obtain, it is clear that the more interesting things should happen in a regime where $M(N)$ is proportional to N. In [BGP1] some fairly weak results concerning the Gibbs states could be proven, but they fell somewhat short of what one would like to have. In particular, no procedure that would even assure the existence of limiting Gibbs measures in this situation had been found. Beyond that, there are only very few results: One, due to Shcherbina and Tirozzi [ST] asserts that the *free energy* of the model is *self averaging* in the sense that its variance is of the order of the inverse system size. Another result, due to Pastur, Shcherbina and Tirozzi [PST] states that the mean field equations obtained from the replica trick (without replica symmetry breaking) are exact, provided the Edwards-Anderson order parameter is self-averaging. Unfortunately, only if $\alpha = 0$ or at high temperatures is it possible to verify this assumption.

In this paper we prove, for the first time, the existence of limiting Gibbs measures associated to any of the stored patterns or finite, albeit very small, α. We repose heavily on the results from [BGP1], but add, as we shall see, a crucial new ingredient: this is an improved self-averaging estimate on the large deviation rate function (free energy functional). Although in its derivation we

use many of the ideas from [ST], our estimates are, and for our purposes have to be, much sharper. Related, but different bounds have also been proven in [BGP2].

Before we explain our results in detail, let us give precise definitions of the model and the quantities we will deal with. We also refer to [BGP1] for more details.

Let us describe the Hopfield model. We set $\Lambda \equiv \{1, \ldots, N\}$ and $\mathcal{S}_\Lambda = \{-1, 1\}^N$ the space of functions $\sigma : \Lambda \to \{-1, 1\}$. We call σ a *spin configuration* on Λ. 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 $\mathcal{B}_\Lambda$ and $\mathcal{B}$ the corresponding Borel sigma algebras. We will define a random Hamiltonian function on the spaces $\mathcal{S}_\Lambda$ 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 $I\!P(\xi_i^\mu = 1) = I\!P(\xi_i^\mu = -1) = \frac{1}{2}$. The Hopfield Hamiltonian on $\mathcal{S}_\Lambda$ is then given by

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

For $\eta \in I\!N$, we denote by $\mathcal{G}^\eta_{N,\beta,h}[\omega]$ the random probability measure on $(\mathcal{S}_\Lambda, \mathcal{B}(\mathcal{S}_\Lambda))$ that assigns to each $\sigma \in \mathcal{S}_\Lambda$ the mass

$$\mathcal{G}^\eta_{N,\beta,h}[\omega](\sigma) \equiv \frac{1}{Z^\eta_{N,\beta,h}[\omega]} e^{-\beta H_N[\omega](\sigma) + \beta h \sum_{i \in \Lambda} \xi_i^\eta[\omega] \sigma_i} \tag{1.2}$$

where $Z^\eta_{N,\beta,h}[\omega]$ is a normalizing factor usually called *partition function*. The reason for the introduction of these measure and the *magnetic field* term h will become apparent later; for a more detailed discussion on the definition and construction of limiting Gibbs measures in mean field models, see [BGP1].

The quantity

$$f^\eta_{N,\beta,h}[\omega] \equiv -\frac{1}{\beta N} \ln Z^\eta_{N,\beta,h}[\omega] \tag{1.3}$$

is called the *free energy*. $\mathcal{G}^\eta_{N,\beta,h}[\omega]$ is called a *finite volume Gibbs state with magnetic field*. Note that the Hamiltonian can be written in terms of the *overlap parameters*

$$m^\mu_N[\omega](\sigma) \equiv \frac{1}{N} \sum_{i=1}^N \xi_i^\mu[\omega] \sigma_i, \quad \mu = 1, \ldots, M \tag{1.4}$$

in the form

$$H_N[\omega](\sigma) = -\frac{N}{2} \sum_{\mu=1}^M \left(m^\mu_N[\omega](\sigma)\right)^2 \equiv -\frac{N}{2} \left\| m_N(\sigma) \right\|_2^2 \tag{1.5}$$

This suggests to introduce the distribution, $\mathcal{Q}^\eta_{N,\beta,h}[\omega]$, of these parameters under the Gibbs measures, i.e.

$$\mathcal{Q}^\eta_{N,\beta,h}[\omega](m) \equiv \mathcal{G}^\eta_{N,\beta,h}[\omega]\left(\{m_N(\sigma) = m\}\right) \tag{1.6}$$

The measures $\mathcal{Q}^{\eta}_{N,\beta,h}[\omega](m)$ on $(I\!R^M, \mathcal{B}(I\!R^M))$ are called *induced measures.*

The following notation is taken from [BGP1]. For $\delta > 0$, we write $a(\delta,\beta)$ for the largest solution of the equation

$$\delta a = \tanh(\beta a) \tag{1.7}$$

We denote by $\|\cdot\|$ 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, for $\nu \in I\!N$ and $s \in \{-1,+1\}$

$$B^{(\nu,s)}_{\rho} \equiv \left\{ x \in I\!R^{I\!N} \,\middle|\, \|x - sa(1-2\sqrt{\alpha},\beta)e^{\nu}\| \leq \rho \right\} \tag{1.8}$$

where e^{ν} denotes the ν-th unit vector in $I\!R^{I\!N}$. With this notation we can announce the following theorem

Theorem 1: *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 all ω,*

$$\lim_{h\downarrow 0} \lim_{N\uparrow\infty} \mathcal{Q}^{\eta}_{N,\beta,h}[\omega]\left(B^{(\eta,+1)}_{\rho}\right) = 1 \tag{1.9}$$

In [BGP1] it had been proven that, under the same hypothesis,

$$\lim_{N\uparrow\infty} \mathcal{Q}_{N,\beta,h=0}[\omega]\left(B_{\rho}\right) = 1 \tag{1.10}$$

where

$$B_{\rho} \equiv \bigcup_{(\nu,s)\in I\!N\times\{-1,+1\}} B^{(\nu,s)}_{\rho} \tag{1.11}$$

is the union of all the balls appearing in Theorem 1. The crucial difference between that result and our new one is that this time we can *select different* limits by adding an arbitrarily small bias in terms of the magnetic field aligned to one of the patterns. To appreciate the difference between these results, notice that from Theorem 1 it follows in particular that the finite dimensional marginal distributions possess limit points that clearly distinguish the selected pattern; to be precise, let $I \subset I\!N$ denote some finite set of positive integers, let $I\!R^I$ denote the finite dimensional space generated by the vectors e^{μ}, with $\mu \in I$, and let Π_I be the orthogonal projector from $I\!R^{M(N)}$ (for any N such that $I \subset \{1,\ldots,M(N)\}$) onto $I\!R^I$. We can introduce the marginal measures on $I\!R^I$ as

$$\mathcal{Q}^{\eta,I}_{N,\beta,h}[\omega] \equiv \mathcal{Q}^{\eta}_{N,\beta,h}[\omega] \circ \Pi_I^{-1} \tag{1.12}$$

Then, (1.9) implies in particular that

$$\lim_{h\downarrow 0} \lim_{N\uparrow\infty} \mathcal{Q}^{\eta,I}_{N,\beta,h}[\omega]\left(\Pi_I B^{(\eta,+1)}_{\rho}\right) = 1 \tag{1.13}$$

Therefore, if $\eta \in I$, the limiting marginal is concentrated on on a $|I|$-dimensional ball around the vector e^η. If we had only (1.10), we would get instead of (1.13) only

$$\lim_{h\downarrow 0}\lim_{N\uparrow\infty} \mathcal{Q}^{\eta,I}_{N,\beta,h}[\omega]\left(\Pi_I B_\rho\right) = 1 \tag{1.14}$$

from which it is not possible to conclude that there exists any finite I for which the corresponding marginal distribution is *not* concentrated on a ball around the origin!

Remark: In the discussion above we have supposed, of course that the balls $B^{(\eta,s)}_\rho$ are disjoint. As we have already pointed out in [BGP1], since $a(\beta,\delta) \sim (\beta-\delta)$ for $(\beta-\delta)$ small, Theorem 1 allows to choose ρ such that this is the case as long as $\beta > 1/(1-c\alpha^{1/4})$. This should be compared with the predictions of Amit et al [AGS] that the "Mattis phase" is bounded by a line $\beta = 1/(1-c\alpha^{1/2})$. The exponent $1/4$ in our bound is in fact due to estimates that are most likely not optimal and should thus not be taken too seriously.

Let us explain the main issue in the proof of Theorem 1. In [BGP1], it has been shown that (with probability tending rapidly to 1 as $N\uparrow\infty$)

$$\mathcal{Q}^{(\eta)}_{N,\beta,h}[\omega]\left(B^c_\rho\right) \leq e^{-cN} \tag{1.15}$$

for some positive constant c, provided that ρ is as large as demanded. Thus, for fixed large N, almost all of the total mass is concentrated on the union of the $2M(N)$ balls $B^{(\eta,s)}_\rho$. The question is then how this mass is distributed over the individual balls: We set

$$F^{(\eta,s)}_{N,\beta,\rho} \equiv -\beta^{-1}\frac{1}{N}\ln \mathcal{Q}_{N,\beta,h=0}[\omega]\left(B^{(\eta,s)}_\rho\right) \tag{1.16}$$

Clearly, the measure is sharply concentrated on the ball for which this quantity takes is minimal value. If for $h=0$ for different η these quantities differ only by terms that tend to zero as $N\uparrow\infty$, then, by adding an arbitrarily small magnetic field aligned on one of the patterns, the corresponding $F^{(\eta,signh)}$ can be tuned to be the minimal value and the measure is concentrated on the corresponding ball. In [BGP1] it was proven that these differences could only be of the order of M/N, which is sufficient in the case $\lim_{N\uparrow\infty} M(N)/N = 0$, but useless if $\lim_{N\uparrow\infty} M(N)/N = \alpha > 0$.

Here we will show that the quantities $F^{(\eta,s)}_{N,\beta,\rho}$ satisfy a strong self-averaging condition. Note that they can be naturally regarded as 'local free energies', associated with the particular state labelled (η,s). The crucial estimate is contained in the following

Proposition 1.1: *Assume that* $\lim \frac{M(N)}{N} = \alpha > 0$. *Let* $\rho < 1$. *Then, for all* $n < \infty$ *there exists* $\tau_n < \infty$, *such that for all* $\tau \geq \tau_n$, *and for* N *large enough,*

$$I\!P\left[\sup_{(\eta,s)}\left|F^{(\eta,s)}_{N,\beta,\rho} - I\!E F^{(\eta,s)}_{N,\beta,\rho}\right| \geq \tau(\ln N)^{3/2}N^{-\frac{1}{2}}\right] \leq N^{-n+1} \tag{1.17}$$

The proof of this proposition will be given in Section 2. Since (1.15) has already been obtained in [BGP1], the proof of Theorem 1, assuming Proposition 1.1, is actually easy. We will give it here:

Proof: (of Theorem 1) Let us introduced the (non-normalized) restricted partition functions

$$Z^{\eta}_{N,\beta,h}[\omega]\left(B^{(\mu,s)}_{\rho}\right) \equiv \frac{1}{2^N}\sum_{\sigma\in\mathcal{S}_N} e^{-\beta H_N(\sigma)+\beta h\sum_{i\in\Lambda}\xi^{\eta}_i\sigma_i}\mathbb{1}_{\left\{\|m_N(\sigma)-se^{\mu}a(\beta)\|_2^2\leq\rho\right\}} \tag{1.18}$$

Notice first that these quantities are easily compared to the corresponding ones in zero magnetic field (we consider only the case h positive):

$$Z^{\eta}_{N,\beta,h}[\omega]\left(B^{(\eta,+1)}_{\rho}\right) \geq e^{\beta Nh(a(\beta)-\sqrt{\rho})}Z_{N,\beta,h=0}[\omega]\left(B^{(\eta,+1)}_{\rho}\right) \tag{1.19}$$

and for $(\mu,s)\neq(\eta,+1)$

$$Z^{\eta}_{N,\beta,h}[\omega]\left(B^{(\mu,s)}_{\rho}\right) \leq e^{+\beta Nh\sqrt{\rho}}Z_{N,\beta,h=0}[\omega]\left(B^{(\eta,+1)}_{\rho}\right) \tag{1.20}$$

Now by Proposition 1.1, with probability greater than, say, $1-N^{-10}$, *all* of the quantities $Z_{N,\beta,h=0}[\omega]\left(B^{(\eta,+1)}_{\rho}\right)$ satisfy

$$e^{-\beta N\mathbb{E}F_{N,\beta,\rho}-\tau_{11}\sqrt{N(\ln N)^3}} \leq Z_{N,\beta,h=0}[\omega]\left(B^{(\eta,+1)}_{\rho}\right) \leq e^{-\beta N\mathbb{E}F_{N,\beta,\rho}+\tau_{11}\sqrt{N(\ln N)^3}} \tag{1.21}$$

Here we have written $\mathbb{E}F_{N,\beta,\rho}$ instead of $\mathbb{E}F^{(\mu,s)}_{N,\beta,\rho}$ to make manifest that, by symmetry, these averaged quantities do not, of course, depend on the indices (μ,s) if the magnetic field is zero. Obviously, again with probability grater than $1-N^{-10}$,

$$\begin{aligned}\mathcal{Q}^{\eta}_{N,\beta,h}[\omega]\left(B^{(\eta,+1)}_{\rho}\right) &= \frac{Z^{\eta}_{N,\beta,h}\left(B^{(\eta,+1)}_{\rho}\right)}{Z^{\eta}_{N,\beta,h}\left(B^{(\eta,+1)}_{\rho}\right)+\sum_{(\mu,s)\neq(\eta,+1)}Z^{\eta}_{N,\beta,h}\left(B^{(\mu,s)}_{\rho}\right)+Z^{\eta}_{N,\beta,h}\left(B^{c}_{\rho}\right)} \\ &\geq \frac{Z^{\eta}_{N,\beta,h}\left(B^{(\eta,+1)}_{\rho}\right)}{Z^{\eta}_{N,\beta,h}\left(B^{(\eta,+1)}_{\rho}\right)\left(1+2Me^{-\beta hN(a(\beta)-2\sqrt{\rho})+2\tau_{11}\sqrt{N(\ln N)^3}}+e^{-cN}\right)}\end{aligned} \tag{1.22}$$

where (1.15), (1.19), (1.20) and (1.21) were used to obtain the second line of (1.22). From here Theorem 1 follows by an application of the first Borel-Cantelli lemma. $\Diamond\Diamond$

In the next section we derive self-averaging properties of large deviation rate functions and prove in particular Proposition 1.1. The actual technical estimates that will be used in the proof are even more consequential and in a final Section 3 we discuss some of these as well as open problems.

2. Self averaging of rate functions

The main new technical result of the present paper is a refined self-averaging estimate on the large deviation rate function. Let us set, for $\tilde{m} \in I\!R^{M(N)}$,

$$F_{N,\beta,\rho}(\tilde{m}) \equiv -\beta^{-1} \frac{1}{N} \ln \left(\mathcal{Q}_{N,\beta} \left[\| m_\Lambda - \tilde{m} \|_2^2 \leq \rho \right] \right) \tag{2.1}$$

(We set $h = 0$ in this section in order not to overcharge notations. The reader will convince himself that all results apply to the case with finite h with some minimal modifications). In [BGP2] an estimate on the probability of deviations of this quantity from its mean value has been proven that turned out to be useful in the case of $\rho \sim M(N)^{-1}$. Here we will in fact be interested in much large values of ρ, typically of the order of some power of α. On the other hand, we will *not* this time need an exponential estimate. Thus, although in principle we follow the same strategy in deriving our bound, there will be a number of important differences.

The main technical result of this section is the following proposition.

Proposition 2.1: *Assume that* $\lim \frac{M(N)}{N} = \alpha > 0$. *Let* $\rho < 1$, $\|\tilde{m}\|_2$ *bounded. Then, for all* $n < \infty$ *there exists* $\tau_n < \infty$, *such that for all* $\tau \geq \tau_n$, *and for* N *large enough,*

$$I\!P \left[|F_{N,\beta,\rho}(\tilde{m}) - I\!E F_{N,\beta,\rho}(\tilde{m})| \geq \tau (\ln N)^{3/2} N^{-\frac{1}{2}} \right] \leq N^{-n} \tag{2.2}$$

Proposition 1.1 from the last section is of course an immediate corollary of this proposition, which is in fact far more general. Thus all that is left is to prove Proposition 2.1.

Proof: (of Proposition 2.1) For technical reasons that will become clear later, we will consider instead of $F^{Hopf}_{N,\beta}(\tilde{m})$ a slightly modified quantity in which the characteristic function $\mathbb{1}_{\{\|m_\Lambda - \tilde{m}\|_2^2 \leq \rho\}}$ is replaced by a smooth version of this function. We let $\chi_{\delta,\rho}(x)$ be a family of infinitely differentiable functions satisfying:

(1) $\chi_{\delta,\rho}(x) \geq 0$,

(2) $\left| \frac{\frac{d}{dx}\chi_{\delta,\rho}(x)}{\chi_{\delta,\rho}(x)} \right| \leq \delta^{-1}$,

(3) $\mathbb{1}_{\{|x| \leq \rho\}} \leq \chi_{\rho,\delta}(x) \leq \mathbb{1}_{\{|x| \leq \rho + \delta\}}$,

(4) $\ln \chi_{\delta,\rho}(x)$ is a concave function of x (where we use the convention $\ln 0 \equiv -\infty$).

We will see that the parameter δ can be chosen as $\delta = O(1/N)$, so that this modification makes no difference whatsoever.

Let us now define

$$Z_N(\tilde{m}) \equiv Z_{N,\beta,\rho,\delta}(\tilde{m}) \equiv \frac{1}{2^N} \sum_{\sigma \in \mathcal{S}_N} e^{-\beta H_N(\sigma)} \chi_{\delta,\rho} \left(\| m_N(\sigma) - \tilde{m} \|_2^2 \right) \tag{2.3}$$

and

$$f_N(\tilde{m}) \equiv -\beta^{-1} \ln Z_N(\tilde{m}) \tag{2.4}$$

We now introduce the decreasing sequence of sigma-algebras $\mathcal{F}_k$ that are generated by the random variables $\{\xi_i^\mu\}_{i \geq k}^{\mu \in I\!N}$ and the corresponding martingale difference sequence

$$\tilde{f}_N^{(k)}(\tilde{m}) \equiv I\!E\left[f_N(\tilde{m}) \middle| \mathcal{F}_k\right] - I\!E\left[f_N(\tilde{m}) \middle| \mathcal{F}_{k+1}\right] \tag{2.5}$$

Notice that we have the identity

$$f_N(\tilde{m}) - I\!E f_N(\tilde{m}) \equiv \sum_{k=1}^{N} \tilde{f}_N^{(k)}(\tilde{m}) \tag{2.6}$$

Let us recall that this construction was first introduced by V.V. Yurinskii [Yu] and employed in the context of spin-glasses and the Hopfield model by Pastur, Shcherbina and Tirozzi [PS,ST].

Our aim is to use an exponential Markov inequality for martingales. This requires in particular bounds on the conditional Laplace transforms of the martingale differences. Namely, we clearly have that

$$\begin{aligned} I\!P\left[\left|\sum_{k=1}^{N} \tilde{f}_N^{(k)}(\tilde{m})\right| \geq Nz\right] &\leq 2 \inf_{t \in I\!R} e^{-|t|Nz} I\!E \exp\left\{t \sum_{k=1}^{N} \tilde{f}_N^{(k)}(\tilde{m})\right\} \\ &= 2 \inf_{t \in R} e^{-|t|Nz} I\!E\left[I\!E\left[\ldots I\!E\left[e^{t\tilde{f}_N^{(1)}(\tilde{m})} \middle| \mathcal{F}_2\right] e^{t\tilde{f}_N^{(2)}(\tilde{m})} \middle| \mathcal{F}_3\right] \ldots e^{t\tilde{f}_N^{(N)}(\tilde{m})} \middle| \mathcal{F}_{N+1}\right] \end{aligned} \tag{2.7}$$

Therefore, *if* we can show that, for some function $\mathcal{L}^{(k)}(t)$, $\ln I\!E\left[e^{t\tilde{f}_N^{(k)}(\tilde{m})} \middle| \mathcal{F}_{k+1}\right] \leq \mathcal{L}^{(k)}(t)$, uniformly in $\mathcal{F}_{k+1}$, then we obtain that

$$I\!P\left[\left|\sum_{k=1}^{N} \tilde{f}_N^{(k)}(\tilde{m})\right| \geq Nz\right] \leq 2 \inf_{t \in I\!R} e^{-|t|Nz + \sum_{k=1}^{N} \mathcal{L}^{(k)}(t)} \tag{2.8}$$

To bound the conditional Laplace transforms, we introduce

$$H_N^{(k)}(\sigma) = -\frac{1}{2N} \sum_{\mu} \sum_{\substack{i,j \\ i,j \neq k}} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j \tag{2.9}$$

$$R_N^{(k)}(\sigma) = -\frac{1}{N} \sum_{\mu} \sum_{\substack{i \\ i \neq k}} \xi_i^\mu \xi_k^\mu \sigma_i \sigma_k \tag{2.10}$$

and

$$\tilde{H}_N^{(k)}(\sigma, u) = H_N^{(k)}(\sigma) + u R_N^{(k)}(\sigma) \tag{2.11}$$

We also define the $M(N)$-dimensional vectors

$$m_N^{(k)}(\sigma, u) \equiv \frac{1}{N}\left(\sum_{\substack{i \\ i \neq k}} \xi_i \sigma_i + u \xi_k \sigma_k\right) \tag{2.12}$$

Naturally, we set

$$Z_N^{(k)}(\tilde{m},u) \equiv \frac{1}{2^N} \sum_{\sigma \in \mathcal{S}_N} e^{-\beta \hat{H}_N^{(k)}(\sigma,u)} \chi_{\delta,\rho}\left(\| m_N^{(k)}(\sigma,u) - \tilde{m} \|_2^2 \right) \tag{2.13}$$

and finally

$$f_N^{(k)}(\tilde{m},u) = -\beta^{-1} \left(\ln Z_N^{(k)}(\tilde{m},u) - \ln Z_N^{(k)}(\tilde{m},0) \right) \tag{2.14}$$

Since for the remainder of the proof, $\tilde{m}$ as well as N will be fixed values, to simplify our notations we will write $f_k(u) \equiv f_N^{(k)}(\tilde{m},u)$. Notice that

$$\tilde{f}_N^{(k)}(\tilde{m}) = I\!E\left[f_k(1) | \mathcal{F}_k \right] - I\!E\left[f_k(1) | \mathcal{F}_{k+1} \right] \tag{2.15}$$

To bound the Laplace transform, we use that, for all $x \in I\!R$,

$$e^x \leq 1 + x + \frac{1}{2} x^2 e^{|x|} \tag{2.16}$$

so that

$$I\!E\left[e^{t \tilde{f}_N^{(k)}(\tilde{m})} \middle| \mathcal{F}_{k+1} \right] \leq 1 + \frac{1}{2} t^2 \, I\!E\left[\left(\tilde{f}_N^{(k)}(\tilde{m}) \right)^2 e^{|t \tilde{f}_N^{(k)}(\tilde{m})|} \middle| \mathcal{F}_{k+1} \right] \tag{2.17}$$

Our strategy will be to use a rather poor *uniform* bound on $\tilde{f}_N^{(k)}(\tilde{m})$ in the exponent but to prove a better estimate on the remaining conditioned expectation of the square. The uniform estimate has been obtained in [BGP2]. We repeat its derivation for the convenience of the reader.

We note first that, alternatively to (2.15), we have also

$$\tilde{f}_N^{(k)}(\tilde{m}) = \int_0^1 du \left(I\!E\left[f_k'(u) \middle| \mathcal{F}_k \right] - I\!E\left[f_k'(u) \middle| \mathcal{F}_{k+1} \right] \right) \tag{2.18}$$

But

$$f_k'(u) = \mathcal{E}_{k,u} \left(\frac{1}{N} \sum_\mu \sum_{i \neq k} \xi_i^\mu \xi_k^\mu \sigma_i \sigma_k + \frac{2}{\beta N} \frac{\chi_{\rho,\delta}'(\| m_N(\sigma,u) - \tilde{m} \|_2^2)}{\chi_{\rho,\delta}(\| m_N(\sigma,u) - \tilde{m} \|_2^2)} \sum_\mu \left(m_N^\mu(\sigma,u) - \tilde{m}^\mu \right) \xi_k^\mu \sigma_k \right) \tag{2.19}$$

where $\mathcal{E}_{k,u}$ denotes the expectation w.r.t. the probability measure

$$\frac{1}{Z_N^{(k)}(\tilde{m},u)} \chi_{\rho,\delta}(\| m_N(\sigma,u) - \tilde{m} \|_2^2) e^{-\beta \hat{H}_N^{(k)}(\sigma,u)} d\sigma \tag{2.20}$$

By some trivial manipulations (see [BGP2] we get

$$|f_k'(u)| \leq \left| \sum_\mu \tilde{m}^\mu \xi_k^\mu \right| + \mathcal{E}_{k,u} \left[\left| \sum_\mu \left(m_N^\mu(\sigma,u) - \tilde{m}^\mu \right) \xi_k^\mu \sigma_k \right| \right] \left(1 + \tfrac{2}{\beta N \delta} \right) + \tfrac{M}{N} \tag{2.21}$$

Now clearly

$$\left|\sum_{\mu}\tilde{m}^{\mu}\xi_k^{\mu}\right| \leq \sum_{\mu}|\tilde{m}^{\mu}| = \|\tilde{m}\|_1 \leq \sqrt{M}\|\tilde{m}\|_2 \tag{2.22}$$

Similarly, for σ in the support of $\mathcal{E}_{k,u}$,

$$\left|\sum_{\mu}\left(m_N^{\mu}(\sigma,u)-\tilde{m}^{\mu}\right)\xi_k^{\mu}\sigma_k\right| \leq \sqrt{M}\|m_N(\sigma,u)-\tilde{m}\|_2 \leq \sqrt{M(\rho+\delta)} \tag{2.23}$$

From this and using assumption (2) on the function χ, we see that

$$|f_k'(u)| \leq \|\tilde{m}\|_1 + \sqrt{M}\sqrt{\rho+\delta}\left(1+\tfrac{2}{\beta N\delta}\right) + \tfrac{M}{N} \tag{2.24}$$

We will now choose $\delta = \frac{2}{\beta N}$ so that we get effectively the bound

$$|f_k'(u)| \leq \|\tilde{m}\|_1 + 2\sqrt{M\rho} \tag{2.25}$$

Using this bound and (2.18) to estimate $\tilde{f}_N^{(k)}(\tilde{m})$ and inserting the result in (2.17), we get that

$$I\!E\left[e^{t\tilde{f}_N^{(k)}(\tilde{m})}\middle|\mathcal{F}_{k+1}\right] \leq 1+\frac{1}{2}t^2e^{2|t|(\|\tilde{m}\|_1+2\sqrt{M\rho})}I\!E\left[\left(\tilde{f}_N^{(k)}(\tilde{m})\right)^2\middle|\mathcal{F}_{k+1}\right] \tag{2.26}$$

Of course we could also use (2.25) to bound the expectation of the sqare in (2.26), but due to the presence of the $\sqrt{M}$ in that bound, this could at best be useful for very small ρ. We will need, however, a bound for finite ρ and thus will have to be more careful.

We will now use (2.15) to write (recall that $\mathcal{F}_k$ are defined in such a way that $\mathcal{F}_k \supset \mathcal{F}_{k+1}$)

$$\begin{aligned}
I\!E\left[\left(\tilde{f}_N^{(k)}(\tilde{m})\right)^2\middle|\mathcal{F}_{k+1}\right] &= I\!E\left[\left(I\!E\left[f_k(1)-I\!E[f_k(1)|\mathcal{F}_{k+1}]\,\middle|\mathcal{F}_k\right]\right)^2\middle|\mathcal{F}_{k+1}\right]\\
&\leq I\!E\left[I\!E\left[\left(f_k(1)-I\!E[f_k(1)|\mathcal{F}_{k+1}]\right)^2\middle|\mathcal{F}_k\right]\middle|\mathcal{F}_{k+1}\right]\\
&= I\!E\left[\left(f_k(1)-I\!E[f_k(1)|\mathcal{F}_{k+1}]\right)^2\middle|\mathcal{F}_{k+1}\right]\\
&= I\!E\left[\left(f_k(1)\right)^2\middle|\mathcal{F}_{k+1}\right]-\left(I\!E[f_k(1)|\mathcal{F}_{k+1}]\right)^2\\
&\leq I\!E\left[\left(f_k(1)\right)^2\middle|\mathcal{F}_{k+1}\right]
\end{aligned} \tag{2.27}$$

At this point it is important to notice that $f_k(u)$ is a *concave* function of u. Note that condition (4) was imposed on $\chi_{\rho,\delta}$ to ensure this fact. Since moreover $f_k(0)=0$, this implies that $|f_k(1)| \leq \max\left(|f_k'(0)|,|f_k'(1)|\right)$. Moreover, from (2.19) and the fact that neither $\tilde{H}_N(\sigma,0)$ nor $m_N(\sigma,0)$ depend on the spin σ_k, we see that $f_k'(0)=0$, so that we get simply

$$I\!E\left[\left(\tilde{f}_N^{(k)}(\tilde{m})\right)^2\middle|\mathcal{F}_{k+1}\right] \leq I\!E\left[\left(f_k'(1)\right)^2\middle|\mathcal{F}_{k+1}\right] \tag{2.28}$$

Let us use the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ and (2.19) to see that

$$\begin{aligned}\left(f_k'(1)\right)^2 \leq\ & 2\left(\mathcal{E}_{k,1}\left(\frac{1}{N}\sum_{\mu}\sum_{i\neq k}\xi_i^\mu\xi_k^\mu\sigma_i\sigma_k\right)\right)^2 \\ & + 2\left(\mathcal{E}_{k,1}\left(\frac{2}{\beta N}\frac{\chi'_{\rho,\delta}(\|m_N(\sigma)-\tilde{m}\|_2^2)}{\chi_{\rho,\delta}(\|m_N(\sigma)-\tilde{m}\|_2^2)}\sum_{\mu}\left(m_N^\mu(\sigma)-\tilde{m}^\mu\right)\xi_k^\mu\sigma_k\right)\right)^2\end{aligned} \tag{2.29}$$

Bounding the square of the expectation by the expectation of the square, we get from this

$$\begin{aligned}I\!E\left[\left(f_k'(1)\right)^2\left|\mathcal{F}_{k+1}\right.\right] \leq\ & 2I\!E\left[\mathcal{E}_{k,1}\left(\frac{1}{N}\sum_{\mu}\sum_{i\neq k}\xi_i^\mu\xi_k^\mu\sigma_i\sigma_k\right)^2\left|\mathcal{F}_{k+1}\right.\right] \\ & + 2I\!E\left[\mathcal{E}_{k,1}\left(\tfrac{2}{\beta N\delta}\sum_{\mu}\left(m_N^\mu(\sigma)-\tilde{m}^\mu\right)\xi_k^\mu\sigma_k\right)^2\left|\mathcal{F}_{k+1}\right.\right]\end{aligned} \tag{2.30}$$

Now we use the following crucial trick: $\mathcal{E}_{k,1}$ is in fact independent of k, and therefore the expectations conditioned on $\mathcal{F}_{k+1}$ are the same if the index k inside it is replaced by any of the indices $j \in \{1,\ldots,k\}$. This allows us to replace (2.30) by

$$\begin{aligned}I\!E\left[\left(f_k'(1)\right)^2\left|\mathcal{F}_{k+1}\right.\right] \leq\ & 2I\!E\left[\mathcal{E}_{k,1}\left(\frac{1}{k}\sum_{j=1}^{k}\frac{1}{N^2}\sum_{\mu}\sum_{\nu}\sum_{i\neq j}\sum_{l\neq j}\xi_i^\mu\xi_l^\nu\xi_j^\mu\xi_j^\nu\sigma_i\sigma_l\right)\left|\mathcal{F}_{k+1}\right.\right] \\ & + 2I\!E\left[\left(\tfrac{2}{\beta N\delta}\right)^2\frac{1}{k}\sum_{j=1}^{k}\sum_{\mu}\sum_{\nu}\xi_j^\mu\xi_j^\nu\mathcal{E}_{k,1}\left(m_N^\mu(\sigma)-\tilde{m}^\mu\right)\left(m_N^\nu(\sigma)-\tilde{m}^\nu\right)\left|\mathcal{F}_{k+1}\right.\right]\end{aligned} \tag{2.31}$$

Let us define the random $M \times M$-matrices $B^{(k)}$ with elements

$$B_{\mu\nu}^{(k)} \equiv \frac{1}{k}\sum_{j=1}^{k}\xi_j^\mu\xi_j^\nu \tag{2.32}$$

Note that these matrices are measurable w.r.t. the sigma algebra $\mathcal{F}\backslash\mathcal{F}_{k+1}$. We will write $b_k \equiv \left\|B^{(k)}\right\|$ for the norms of these matrices.

We can write (2.31) in the form

$$\begin{aligned}& I\!E\left[\left(f_k'(1)\right)^2\left|\mathcal{F}_{k+1}\right.\right] \\ & \leq 2I\!E\left[\sum_{\mu,\nu}\frac{1}{k}\sum_{j=1}^{k}\xi_j^\mu\xi_j^\nu\mathcal{E}_{1,k}\left(\frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i-\frac{1}{N}\xi_j^\mu\sigma_j\right)\left(\frac{1}{N}\sum_{l=1}^{N}\xi_l^\nu\sigma_l-\frac{1}{N}\xi_j^\nu\sigma_j\right)\left|\mathcal{F}_{k+1}\right.\right] \\ & + 2\left(\tfrac{2}{\beta N\delta}\right)^2 I\!E\left[\mathcal{E}_{k,1}\left((m_N(\sigma)-\tilde{m}),B^{(k)}(m_N(\sigma)-\tilde{m})\right)\left|\mathcal{F}_{k+1}\right.\right]\end{aligned} \tag{2.33}$$

In the last line the $(\cdot,\cdot)$ denotes the scalar product in $I\!R^M$. For the second term in (2.33) we get immediately the bound

$$\begin{aligned} & 2\left(\tfrac{2}{\beta N\delta}\right)^2 I\!E\left[\mathcal{E}_{k,1}\left((m_N(\sigma)-\tilde{m}), B^{(k)}(m_N(\sigma)-\tilde{m})\right)\Big|\mathcal{F}_{k+1}\right] \\ & \leq 2\rho\left(\tfrac{2}{\beta N\delta}\right)^2 I\!E b_k \end{aligned} \tag{2.34}$$

where we have used that $B^{(k)}$ is measurable with respect to $\mathcal{F}\backslash\mathcal{F}_{k+1}$ to replace the conditional expectation by the expectation.

The first term in (2.33) is more complicated. We rewrite it as

$$\begin{aligned} & 2I\!E\left[\sum_{\mu,\nu}\frac{1}{k}\sum_{j=1}^{k}\xi_j^\mu\xi_j^\nu\mathcal{E}_{1,k}\left(m_N^\mu(\sigma)m_N^\nu(\sigma)\right)\Big|\mathcal{F}_{k+1}\right] \\ & + 2I\!E\left[\sum_{\mu,\nu}\frac{1}{k}\sum_{j=1}^{k}\xi_j^\mu\xi_j^\nu\frac{1}{N^2}\xi_j^\mu\xi_j^\nu\Big|\mathcal{F}_{k+1}\right] \\ & - 4I\!E\left[\sum_{\mu,\nu}\frac{1}{k}\sum_{j=1}^{k}\xi_j^\mu\xi_j^\nu\mathcal{E}_{1,k}\left(\frac{1}{N^2}\sum_{i=1}^{N}\xi_i^\mu\sigma_i\xi_j^\nu\sigma_j\right)\Big|\mathcal{F}_{k+1}\right] \\ & \equiv I + II - III \end{aligned} \tag{2.35}$$

We deal with the three terms separately. First,

$$II = 2\frac{M^2}{N^2} \tag{2.36}$$

On the other hand

$$\begin{aligned} III &= 4\frac{M}{N}I\!E\left[\sum_{\mu}\frac{1}{kN}\sum_{j=1}^{k}\sum_{i=1}^{N}\mathcal{E}_{1,k}\left(\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j\right)\Big|\mathcal{F}_{k+1}\right] \\ &= 4\frac{M}{N}I\!E\left[\mathcal{E}_{1,k}\left(\frac{1}{k}\sum_{j=1}^{k}\xi_j\sigma_j\,,\,\frac{1}{N}\sum_{i=1}^{N}\xi_i\sigma_i\right)\Big|\mathcal{F}_{k+1}\right] \end{aligned} \tag{2.37}$$

Using the Schwarz inequality, one obtains therefore that

$$|III| \leq 4\frac{M}{N}I\!E\left[\mathcal{E}_{1,k}\left\|\frac{1}{k}\sum_{j=1}^{k}\xi_j\sigma_j\right\|_2\|m_N(\sigma)\|_2\Big|\mathcal{F}_{k+1}\right] \tag{2.38}$$

But

$$\left\|\frac{1}{k}\sum_{j=1}^{k}\xi_j\sigma_j\right\|_2^2 = \frac{1}{k^2}\sum_{\mu}\sum_{j=1}^{k}\sum_{l=1}^{k}\sigma_j\xi_j^\mu\xi_l^\mu\sigma_l \leq \left\|A^{(k)}\right\| \tag{2.39}$$

where $A^{(k)}$ is the $k \times k$-matrix with elements $A^{(k)}_{jl} \equiv \frac{1}{k}\sum_\mu \xi^\mu_j \xi^\mu_l$. It is easy to verify by simple algebraic manipulations that $\left\|A^{(k)}\right\| = \left\|B^{(k)}\right\|$. Since on the other hand

$$\mathcal{E}_{1,k}\left\|m_N(\sigma)\right\|_2 \leq (\sqrt{\rho} + \|\tilde{m}\|_2) \tag{2.40}$$

this combines to

$$|III| \leq 4\frac{M}{N}I\!E\sqrt{\left\|B^{(k)}\right\|}(\sqrt{\rho} + \|\tilde{m}\|_2) \leq 4\frac{M}{N}\sqrt{I\!Eb_k}(\sqrt{\rho} + \|\tilde{m}\|_2) \tag{2.41}$$

We are left with the term I. We write

$$\begin{aligned}
I =& 2I\!E\left[\sum_{\mu,\nu}\frac{1}{k}\sum_{j=1}^k \xi^\mu_j\xi^\nu_j\mathcal{E}_{1,k}\left((m^\mu_N(\sigma)-\tilde{m}^\mu)(m^\nu_N(\sigma)-\tilde{m}^\nu)\right)\big|\mathcal{F}_{k+1}\right] \\
&+ 4I\!E\left[\sum_{\mu,\nu}\frac{1}{k}\sum_{j=1}^k \xi^\mu_j\xi^\nu_j\mathcal{E}_{1,k}\left((m^\mu_N(\sigma)-\tilde{m}^\mu)\tilde{m}^\nu\right)\big|\mathcal{F}_{k+1}\right] \\
&+ 2I\!E\left[\sum_{\mu,\nu}\frac{1}{k}\sum_{j=1}^k \xi^\mu_j\xi^\nu_j\tilde{m}^\mu\tilde{m}^\nu\big|\mathcal{F}_{k+1}\right] \\
=& 2I\!E\left[\mathcal{E}_{1,k}\left((m_N(\sigma)-\tilde{m})\,,\,B^{(k)}(m_N(\sigma)-\tilde{m})\right)\big|\mathcal{F}_{k+1}\right] \\
&+ 4I\!E\left[\mathcal{E}_{1,k}\left((m_N(\sigma)-\tilde{m})\,,\,B^{(k)}\tilde{m}\right)\big|\mathcal{F}_{k+1}\right] \\
&+ 2I\!E\left[\frac{1}{k}\sum_{j=1}^k\left(\sum_\mu \xi^\mu_j\tilde{m}^\mu\right)^2\right]
\end{aligned} \tag{2.42}$$

Using the Schwartz inequality for the first two terms, we obtain from (2.42) the bound

$$I \leq 2\rho I\!Eb_k + 4\sqrt{\rho}\|\tilde{m}\|_2 + 2\|\tilde{m}\|_2^2 \tag{2.43}$$

Combining all these estimates with (2.34) and choosing as before $\delta = 2/(\beta N)$ gives that

$$I\!E\left[\left(f'_k(1)\right)^2\big|\mathcal{F}_{k+1}\right] \leq \|\tilde{m}\|_2^2 + 2\tfrac{M^2}{N^2} + 4\sqrt{\rho}\|\tilde{m}\|_2 + 4\rho I\!Eb_k + 4\tfrac{M}{N}\sqrt{I\!Eb_k}(\|\tilde{m}\|_2 + \sqrt{\rho}) \tag{2.44}$$

Collecting our bounds and inserting them into (2.26) we have

$$\begin{aligned}
&I\!E\left[e^{t\tilde{f}^{(k)}_N(\tilde{m})}\big|\mathcal{F}_{k+1}\right] \\
&\leq 1 + \tfrac{1}{2}t^2e^{2|t|(\|\tilde{m}\|_1+2\sqrt{M\rho})}\left(\|\tilde{m}\|_2^2 + 2\tfrac{M^2}{N^2} + 4\sqrt{\rho}\|\tilde{m}\|_2 + 4\rho I\!Eb_k + 4\tfrac{M}{N}\sqrt{I\!Eb_k}(\|\tilde{m}\|_2+\sqrt{\rho})\right) \\
&\leq \exp\left\{\tfrac{1}{2}t^2e^{2|t|(\|\tilde{m}\|_1+2\sqrt{M\rho})}\left(\|\tilde{m}\|_2^2 + 2\tfrac{M^2}{N^2} + 4\sqrt{\rho}\|\tilde{m}\|_2 + 4\rho I\!Eb_k + 4\tfrac{M}{N}\sqrt{I\!Eb_k}(\|\tilde{m}\|_2+\sqrt{\rho})\right)\right\}
\end{aligned} \tag{2.45}$$

This is a uniform bound on $\mathcal{L}^{(k)}(t)$ so that

$$\begin{aligned}\mathbb{E}\exp\left\{t\sum_{k=1}^{N}\tilde{f}_N^{(k)}(\tilde{m})\right\} \leq \exp\Bigg\{\tfrac{1}{2}t^2 e^{2|t|(\|\tilde{m}\|_1+2\sqrt{M\rho})}\Bigg(N\left(\|\tilde{m}\|_2^2+2\tfrac{M^2}{N^2}+4\sqrt{\rho}\|\tilde{m}\|_2\right)\\ +\sum_{k=1}^{N}\left(4\rho\mathbb{E}b_k+4\tfrac{M}{N}\sqrt{\mathbb{E}b_k}(\|\tilde{m}\|_2+\sqrt{\rho})\right)\Bigg)\Bigg\}\end{aligned} \tag{2.46}$$

All we still need is a bound on the expectation of the b_k. But this follows easily from the estimates on the norms of such matrices proven, for instance in [ST, BG1]. We will use the bounds on the traces of powers of such matrices proven in [BG1] to deduce

Lemma 2.2: *Let $B^{(k)}$ denote the $M \times M$ matrices with elements defined in (2.32). Then*

$$\mathbb{E}\left\|B^{(k)}\right\| \leq \begin{cases} 2\frac{M}{k}+2e\sqrt{\frac{M}{k}} & ,\text{if } k \leq M \\ 2+2e\sqrt{\frac{M}{k}} & ,\text{ if } k \geq M \end{cases} \tag{2.47}$$

From this lemma it follows that

$$\begin{aligned}\sum_{k=1}^{N}\mathbb{E}b_k &\leq 2\sum_{k=1}^{M}M/k+2e\sum_{k=1}^{N}\sqrt{M/k}+2(N-M)\\ &\leq c(M\ln M+N+\sqrt{MN})\end{aligned} \tag{2.48}$$

and

$$\sum_{k=1}^{N}\sqrt{\mathbb{E}b_k} \leq c'N \tag{2.49}$$

for some numerical constants c and c'. Using these estimates we get

$$\begin{aligned}&\mathbb{E}\exp\left\{t\sum_{k=1}^{N}\tilde{f}_N^{(k)}(\tilde{m})\right\}\\ &\leq \exp\left\{\tfrac{1}{2}t^2e^{2|t|(\|\tilde{m}\|_1+2\sqrt{M\rho})}N\left[\|\tilde{m}\|_2^2+2\tfrac{M^2}{N^2}+4\sqrt{\rho}\|\tilde{\|}_2+4c\rho\tfrac{M}{N}\ln N+4c'\tfrac{M}{N}(\|\tilde{m}\|_2+\sqrt{\rho})\right]\right\}\end{aligned} \tag{2.50}$$

Let us remark that we will use this bound only for $\tilde{m}$ with bounded ℓ^2-norm and for ρ and M/N much smaller than 1. Thus (2.50) takes on the simple form

$$\mathbb{E}\exp\left\{t\sum_{k=1}^{N}\tilde{f}_N^{(k)}(\tilde{m})\right\} \leq \exp\left\{ct^2e^{c'|t|\sqrt{M}}N(1+\rho\tfrac{M}{N}\ln N)\right\} \tag{2.51}$$

for (new) constants c and c'. Eq. (2.51) can now be used together with (2.7) to derive a variety of bounds by suitable choices of t. Note that the presence of the term $e^{|t|\sqrt{M}}$ restricts the useful values of t essentially to the intervall $[0, M^{-1/2}]$, so that in particular no exponential estimates can

be obtained (but see [BGP2] where this is done for $\rho \sim 1/M$). But for our present purposes we will not need this. In fact, the most convenient bounds for us are derived by choosing $t = n\frac{\ln N}{zN}$. This yields (we put $M/N = \alpha$) that

$$I\!P\left[|f_N(\tilde{m}) - I\!E f_N(\tilde{m})| \geq Nz\right] \leq N^{-n} \exp\left\{c\frac{(\ln N)^2 n^2(1+\rho\alpha\ln N)}{z^2 N} N^{\frac{c'\sqrt{M/N}}{z\sqrt{N}}}\right\} \tag{2.52}$$

If $z\sqrt{N}$ is sufficiently large, e.g. $z\sqrt{N} = \tau(\ln N)^{3/2}$ then for arbitrary n, the argument of the exponential function converges to 0 as $N \uparrow \infty$. From this the statement of Proposition 2.1 follows immediately for the non-normalized quantities $f_N(\tilde{m})$ (which, by looking at the proof of Theorem 1, is in fact all we would really need). The reader might worry whether the same estimate holds also for the logarithm of the normalizing factor, i.e. the free energy itself. We recall that in [ST] only the vanishing of the variance of the free energy was proven. To obtain our sharper estimates, we should in principle repeat our proof with $\tilde{m} = 0$ and $\rho = \infty$. Doing this naively, we would run into trouble. However, note that we can of course always write

$$Z_{N,\beta} = Z^{<}_{N,\beta} + Z^{>}_{N,\beta} \tag{2.53}$$

where

$$Z^{<}_{N,\beta} \equiv \frac{1}{2^N} \sum_{\sigma\in\mathcal{S}_N} e^{-\beta H_N(\sigma)} \mathbb{1}_{\left\{\|m_N(\sigma)\|_2^2 \leq 2\right\}} \tag{2.54}$$

and

$$Z^{>}_{N,\beta} \equiv \frac{1}{2^N} \sum_{\sigma\in\mathcal{S}_N} e^{-\beta H_N(\sigma)} \mathbb{1}_{\left\{\|m_N(\sigma)\|_2^2 > 2\right\}} \tag{2.55}$$

But $\|m_N(\sigma)\|_2^2 \leq \|A\|$, where A is the $N \times N$-matrix with elements $A_{ij} = \frac{1}{N}\sum_{\mu=1}^{N} \xi_i^\mu \xi_j^\mu$. This matrix has obviously the same norm as the matrix $B^{(N)}$ defined above (check!) so that the estimates on the norm of these random matrices from [ST] or [BG1] can be used. It follows in particular that this norm is less than two with probability at least $1 - e^{-N^{1/6}}$! Therefore,

$$I\!P\left[Z^{>}_{N,\beta} = 0\right] \geq 1 - e^{-N^{1/6}} \tag{2.56}$$

Since on the other hand the deviation of $\ln Z^{<}_{N,\beta}[\omega]$ from its mean is easily shown to satisfy the bound (2.2), we obtain the statement of the proposition. $\diamondsuit$

3. Discussion and conclusions

The result on the strong self-averaging property of the rate function that is contained in Proposition 2.1 is quite interesting beyond the fact that it allows to prove Theorem 1. Let us note before all that curiously enough, although we have such strong estimates on the fluctuations of the local free energies about their mean, nothing is known concerning the *convergence* of the means themselves as $N \uparrow \infty$, as soon as $\alpha > 0$. This is certainly quite curious, but, as we have seen, not necessarily very disturbing.

The result stated in Proposition 1.1 reflects a very high degree of symmetry among the patterns. For $\alpha = 0$, the free energy functional has its absolute minima very precisely at the points $\pm e^{\mu} a(\beta)$ (the "Mattis states") with the value fixed at that of the Curie-Weiss model. As α increases, the positions of these minima shift in a continuous, and probably somewhat random, fashion away from these points, but, surprisingly enough, the value of this function at all these minima remains strictly the same. Somehow, although the function changes randomly in a different way near each of the Mattis states, the profoundness of the ensuing minimal values is kept the same to an astonishing degree of precision. Note that this fact remains valid well beyond the value of α for which we know that the absolute minima are near the Mattis states. This suggests that, if, as expected, the 'ordered phase' of the Hopfield model disappears, this happens in such a way that for some very precise value of α (depending however on β) *all* the minima near the Mattis states cease to be *absolute* minima, while somewhere else the new absolute minima appear. This scenario is to be contrasted with the other imaginable picture in which first a competition arises between the Mattis states in the course of which some remain absolute minima while others turn metastable. In such a scenario, which we can now exclude, the existence of limiting Gibbs states would in fact have been doubtful if not unlikely.

It may be of interest in this context to make some remarks on the self-averaging properties of the free energy in the Sherrington-Kirkpatrick [SK] model of a spin glass. We recall that the Hamiltonian of this model is given by

$$H_N(\sigma) = -\frac{1}{\sqrt{N}} \sum_{i,j=1}^{N} J_{ij} \sigma_i \sigma_j \tag{3.1}$$

where $\{J_{ij}\}_{i \leq j \in I\!N \times I\!N}$ is a family of independent Gaussian random variables with mean zero and variance one. Pastur and Shcherbina [PS] have proven that in this model

$$I\!E \left[\left(F_{N,\beta} - I\!E F_{N,\beta} \right)^2 \right] \leq \frac{c}{N} \tag{3.2}$$

and that therefore the difference between the free energy and its mean tends to zero in probability as $N \uparrow \infty$. Using the techniques of Section 2, it is actually very easy to improve this result and to show that in fact

Proposition 3.1: *In the Sherrington-Kirkpatrick model, for all $\beta > 0$, and for all $\infty > z \geq 0$*

$$I\!P\left[|F_{N,\beta} - I\!E F_{N,\beta}| \geq z\right] \leq \exp\left(-N\frac{z^2}{5}\right) \tag{3.3}$$

if N is sufficiently large.

Proof: We write $\ell_1, \ell_2, \ldots, \ell_{N(N+1)/2}$ for an arbitrarily chosen fixed arrangement of the 'links' (ij) with $i \leq j$. If $\ell_k = (ij)$, we set $J_{\ell_k} \equiv J_{ij}$ and $\sigma_{\ell_k} \equiv \sigma_i \sigma_j$. We denote by $\mathcal{F}_k$ the sigma-algebra generated by the random variables $\{J_{\ell_m}\}_{m \geq k}$.

With the same notations as in Section 2, just suppressing the $\tilde{m}$, this allows us to write that

$$F_{N,\beta} - I\!E F_{N,\beta} = \frac{1}{N} \sum_{k=1}^{N(N+1)/2} \tilde{f}_N^{(k)} \tag{3.4}$$

Thus we see that the exponential bound in Proposition 3.1 will follow from a suitable bound on the conditional Laplace transform of $\tilde{f}_N^{(k)}$. The necessary analogue of (2.11) is

$$\tilde{H}_N^{(k)}(\sigma, u) \equiv -\frac{1}{\sqrt{N}} \sum_{n \neq k} J_{\ell_n} \sigma_{\ell_n} - u\frac{1}{\sqrt{N}} J_{\ell_k} \sigma_{\ell_k} \tag{3.5}$$

This yields that this time

$$f_k'(u) = \frac{1}{\sqrt{N}} J_{\ell_k} \mathcal{E}_{k,u} \sigma_{\ell_k} \tag{3.6}$$

Trivially, here

$$|f_k'(u)| \leq \frac{1}{\sqrt{N}} |J_{\ell_k}| \tag{3.7}$$

Therefore, using (2.16) we get

$$\begin{aligned} I\!E\left[e^{t\tilde{f}_N^{(k)}} \middle| \mathcal{F}_{k+1}\right] &\leq 1 + \frac{2}{N} t^2 I\!E\left[J_{\ell_k}^2 e^{2|t||J_{\ell_k}|/\sqrt{N}}\right] \\ &\leq 1 + \frac{2}{N} t^2 e^{2t^2/N}\left(1 + 4t^2/N\right) \\ &\leq \exp\left(\frac{2}{N} t^2 e^{2t^2/N}\left(1 + 4t^2/N\right)\right) \end{aligned} \tag{3.8}$$

Thus we obtain:

$$\begin{aligned} I\!P\left[\left|\sum_{k=1}^{N(N+1)/2} \tilde{f}_N^{(k)}\right| \geq Nz\right] &\leq 2 \inf_{t \in I\!R} \exp\left(-|t|Nz + Nt^2 e^{2t^2/N}\left(1 + 4t^2/N\right)\right) \\ &\leq 2e^{-z^2 N/5} \end{aligned} \tag{3.9}$$

where the last expression holds if z^2/N is small enough.$\diamondsuit$

This implies in particular the almost sure convergence to zero of $F_{N,\beta} - I\!E F_{N,\beta}$. This does not imply the almost sure convergence of the free energy since it is not know that the mean of the free energy converge below the critical temperature $\beta^{-1} = 1$.

It may be surprising that Proposition 3.1 gives an estimate in the SK model that is much sharper than what we get in the Hopfield model, while its proof is considerably simpler. The crucial property responsible for this fact is the independence of the two-spin couplings, which does not hold in the Hopfield case.

References

[AGS] 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 results on the thermodynamics of the dilute Hopfield model", J. Stat. Phys. **69**: 597-627 (1993).

[BG2] A. Bovier and V. Gayrard, "Rigorous results on the Hopfield model of neural networks", to appear in Resenhas do IME-USP **2** (1994).

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

[BGP2] A. Bovier, V. Gayrard, and P. Picco, "Large deviation principles for the Hopfield model and the Kac-Hopfield model", submitted to Prob. Theor. Rel. Fields (1994).

[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).

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

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

[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).

[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).

[Yu] V.V. Yurinskii, "Exponential inequalities for sums of random vectors", J. Multivariate Anal. **6**: 473-499 (1976).