*Journal of Statistical Physics, Vol. 79, Nos. 1/2, 1995*

# Gibbs States of the Hopfield Model with Extensively Many Patterns

**Anton Bovier,[1] Véronique Gayrard,[2] and Pierre Picco[2]**

*Received May 9, 1994; final October 24, 1994*

We consider the Hopfield model with $M(N)=\alpha N$ patterns, where $N$ is the number of neurons. We show that if $\alpha$ 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. 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.

**KEY WORDS:** Hopfield model; Gibbs states; self-averaging; spin glasses.

## 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.[8] This model had been introduced first by Figotin and Pastur[6,7] as a simple soluble model of a spin glass, but has enjoyed, after its reinterpretation as a model for an autoassociative memory by Hopfield,[8] 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[13,10] by Amit *et al.*,[1] has allowed for the first time 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,

[1] Weierstrass-Institut für Angewandte Analysis und Stochastik, D-10117 Berlin, Germany. E-mail: bovier@iaas-berlin.d400.de.

[2] Centre de Physique Théorique-CNRS, Luminy, Case 907, F-13288 Marseille Cedex 9, France. E-mail: gayrard@cpt.univ-mrs.fr; picco@cpt.univ-mrs.fr.

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., ref. 2), 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.[4] 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 ref. 4 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,[14] 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 *et al.*,[12] 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 with any of the stored patterns or finite, albeit very small, $\alpha$. We rest heavily on the results from ref. 4, 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 ref. 11, our estimates are, and for our purposes have to be, much sharper. Related, but different bounds have also been proven in ref. 5.

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 ref. 4 for more details.

Let us describe the Hopfield model. We set $\Lambda \equiv \{1,..., N\}$ and $\mathcal{S}_\Lambda = \{-1, 1\}^N$ the space of functions $\sigma: \Lambda \to \{-1, 1\}$. We call $\sigma$ a *spin configuration* on $\Lambda$. We shall write $\mathcal{S} \equiv \{-1, 1\}^{\mathbb{N}}$ for the space of half-infinite sequences equipped with the product topology of discrete topology on

$\{-1, 1\}$. We denote by $\mathscr{B}_\Lambda$ and $\mathscr{B}$ the corresponding Borel sigma algebras. We will define a random Hamiltonian function on the spaces $\mathscr{S}_\Lambda$ as follows. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be an abstract probability space. Let $\xi \equiv \{\xi_i^\mu\}_{i,\mu \in \mathbb{N}}$ be a two-parameter family of independent, identically distributed random variables on this space such that $\mathbb{P}(\xi_i^\mu = 1) = \mathbb{P}(\xi_i^\mu = -1) = 1/2$. The Hopfield Hamiltonian on $\mathscr{S}_\Lambda$ 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 $\eta \in \mathbb{N}$, we denote by $\mathscr{G}_{N,\beta,h}^\eta[\omega]$ the random probability measure on $(\mathscr{S}_\Lambda, \mathscr{B}(\mathscr{S}_\Lambda))$ that assigns to each $\sigma \in \mathscr{S}_\Lambda$ the mass

$$\mathscr{G}_{N,\beta,h}^\eta[\omega](\sigma) \equiv \frac{1}{Z_{N,\beta,h}^\eta[\omega]} \exp\left\{-\beta H_N[\omega](\sigma) + \beta h \sum_{i \in \Lambda} \xi_i^\eta[\omega]\, \sigma_i\right\} \tag{1.2}$$

where $Z_{N,\beta,h}^\eta[\omega]$ is a normalizing factor usually called the *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 ref. 4.

The quantity

$$f_{N,\beta,h}^\eta[\omega] \equiv -\frac{1}{\beta N} \ln Z_{N,\beta,h}^\eta[\omega] \tag{1.3}$$

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

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

in the form

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

This suggests that we introduce the distribution $\mathscr{Q}_{N,\beta,h}^\eta[\omega]$ of these parameters under the Gibbs measures, i.e.,

$$\mathscr{Q}_{N,\beta,h}^\eta[\omega](m) \equiv \mathscr{G}_{N,\beta,h}^\eta[\omega](\{m_N(\sigma) = m\}) \tag{1.6}$$

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

The following notation is taken from ref. 4. 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 $\|\cdot\|_2$ the $l^2$-norm on $\mathbb{R}^{\mathbb{N}}$. Given that $\lim_{N\uparrow\infty}[M(N)/N] = \alpha$, we set, for fixed $\beta$, for $\nu \in \mathbb{N}$, and $s \in \{-1, +1\}$,

$$B_\rho^{(\nu,s)} \equiv \{x \in \mathbb{R}^{\mathbb{N}} \mid \|x - sa(1 - 2\sqrt{\alpha}, \beta)\, e^\nu\|_2 \leqslant \rho\} \tag{1.8}$$

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

**Theorem 1.** There exists $\alpha_0 > 0$ such that if $\lim[M(N)/N] = \alpha$, with $\alpha \leqslant \alpha_0$, then, for all $\beta > 1 + 3\sqrt{\alpha}$, if $\rho^2 > C[a(1 - 2\sqrt{\alpha}, \beta)]^{3/2}\, \alpha^{1/8}\, |\ln \alpha|^{1/4}$, for almost all $\omega$,

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

In ref. 4 it had been proven that, under the same hypothesis,

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

where

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

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 $I \subset \mathbb{N}$ denote some finite set of positive integers, let $\mathbb{R}^I$ denote the finite-dimensional space generated by the vectors $e^\mu$, with $\mu \in I$, and let $\Pi_I$ be the orthogonal projector from $\mathbb{R}^{M(N)}$ (for any $N$ such that $I \subset \{1,\ldots, M(N)\}$) onto $\mathbb{R}^I$. We can introduce the marginal measures on $\mathbb{R}^I$ 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

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

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

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

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 ref. 4, since $a(\beta, \delta) \sim (\beta - \delta)$ for $(\beta - \delta)$ small, Theorem 1 allows us to choose $\rho$ 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.*[1] that the "Mattis phase"[3] is bounded by a line $\beta = 1/(1 - c\alpha^{1/2})$ [see the curve $T = T_c(\alpha)$ in Fig. 2 and the last equation in Section 5 therein]. 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 ref. 4 it has been shown that (with probability tending rapidly to 1 as $N \uparrow \infty$)

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

for some positive constant $c$, provided that $\rho$ 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

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

Clearly, the measure is sharply concentrated on the ball for which this quantity takes its minimal value. If for $h = 0$ for different $\eta$ 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 $F^{(\eta, \operatorname{sign} h)}$ can be tuned to be the minimal value and the measure

[3] We use this name for the parameter region in which, in the words of Amit *et al.*, "the retrieval FM states are global minima." Note that these are, again in their wording, "Mattis-like, but in the case of finite $\alpha$, $m$ is less than 1 even at $T = 0$." Note also that in this case, the "random overlaps (*with the other patterns*) will be of order $O(1/\sqrt{N})$," which is in agreement with our results. We prefer thus the name "Mattis phase" for this region, rather than "ferromagnetic" or "retrieval phase," which may be misleading.

is concentrated on the corresponding ball. In ref. 4 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 labeled $(\eta, s)$. The crucial estimate is contained in the following

**Proposition 1.1.** Assume that $\alpha$ and $\rho$ satisfy the hypothesis of Theorem 1. Then, for all $n<\infty$ there exists $\tau_n<\infty$ such that for all $\tau\geqslant\tau_n$ and for $N$ large enough,

$$\mathbb{P}[\sup_{(\eta,s)} |F^{(\eta,s)}_{N,\beta,\rho}-\mathbb{E}F^{(\eta,s)}_{N,\beta,\rho}|\geqslant\tau(\ln N)^{3/2}N^{-1/2}]\leqslant N^{-n+1} \tag{1.17}$$

The proof of this proposition will be given in Section 2. Since (1.15) has already been obtained in ref. 4, the proof of Theorem 1, assuming Proposition 1.1, is actually easy. We will give it here:

*Proof of Theorem 1.* Let us introduce the (nonnormalized) restricted partition functions

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

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

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

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

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

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

$$\begin{aligned}&\exp\{-\beta N\mathbb{E}F_{N,\beta,\rho}-\tau_{11}[N(\ln N)^3]^{1/2}\}\\ &\quad\leqslant\frac{Z_{N,\beta,h=0}[\omega](B^{(\eta,+1)}_{\rho})}{Z_{N,\beta,h=0}[\omega]}\\ &\quad\leqslant\exp\{-\beta N\mathbb{E}F_{N,\beta,\rho}+\tau_{11}[N(\ln N)^3]^{1/2}\}\end{aligned} \tag{1.21}$$

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 $(\mu, s)$ if the magnetic field is zero. Obviously, again with probability greater than $1-N^{-10}$,

$$\begin{aligned}
&\mathcal{Q}^{\eta}_{N,\beta,h}[\omega](B^{(\eta,+1)}_{\rho})\\
&\quad=\frac{Z^{\eta}_{N,\beta,h}(B^{(\eta,+1)}_{\rho})}{Z^{\eta}_{N,\beta,h}(B^{(\eta,+1)}_{\rho})+\sum_{(\mu,s)\neq(\eta,+1)}Z^{\eta}_{N,\beta,h}(B^{(\mu,s)}_{\rho})+Z^{\eta}_{N,\beta,h}(B^{c}_{\rho})}\\
&\quad\geqslant\left[1+2Me^{-\beta hN[\alpha(\beta)-2\rho]+2\bar{c}_{11}[N(\ln N)^3]^{1/2}}+e^{-CN}\right]^{-1} \qquad (1.22)
\end{aligned}$$

where (1.15) and (1.19)–(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. ▌

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 $\tilde{m}\in\mathbb{R}^{M(N)}$,

$$F_{N,\beta,\rho}(\tilde{m})\equiv-\beta^{-1}\frac{1}{N}(\mathcal{Q}_{N,\beta}[\,\|m_N(\sigma)-\tilde{m}\|_2\leqslant\rho]) \qquad (2.1)$$

For technical reasons that will become clear later, we will not consider directly $F_{N,\beta,\rho}(\tilde{m})$, but a slightly modified quantity in which the characteristic function $\mathbb{1}_{\{\|m_N-\tilde{m}\|_2\leqslant\rho\}}$ is replaced by a smooth version of this function. We let $\chi_{\rho,\delta}(x)$ be a family of infinitely differentiable functions satisfying:

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

(2) $|(d/dx)\,\chi_{\rho,\delta}(x)|\leqslant 2\delta^{-1}$.

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

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

Let us now define

$$\tilde{Z}_{N,\beta,\rho,\delta}(\tilde{m}) \equiv \frac{1}{2^N} \sum_{\sigma \in \mathcal{S}_N} e^{-\beta H_N(\sigma)} \chi_{\rho,\delta}(\|m_N(\sigma) - \tilde{m}\|_2) \tag{2.2}$$

and

$$\tilde{F}_{N,\beta,\rho}(\tilde{m}) \equiv -\frac{1}{\beta N} \ln \tilde{Z}_{N,\beta,\rho}(\tilde{m}) \tag{2.3}$$

(We set $h = 0$ in this section in order not to overload the notations. The reader can convince herself or himself that all results apply to the case with finite $h$ with some minimal modifications.)

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

**Lemma 2.1.** Assume that $\rho$ is as in Theorem 1, and $\delta < a(\beta) - 2\rho$. Then, for all $\omega$ for which (1.15) holds,

$$|\tilde{F}_{N,\beta,\rho}(sa(\beta)\, e^{\eta}) - F^{(\eta,s)}_{N,\beta,\rho}| \leqslant \frac{1}{N} \tag{2.4}$$

*Proof.* We have that

$$|\tilde{F}_{N,\beta,\rho}(sa(\beta)\, e^{\eta}) - F^{(\eta,s)}_{N,\beta,\rho}| \leqslant \frac{1}{\beta N} \ln \left( \frac{Z_{N,\beta}(B^{(\eta,s)}_{\rho+\delta})}{Z_{N,\beta}(B^{(\eta,s)}_{\rho})} \right) \tag{2.5}$$

For $\delta \leqslant a(\beta) - 2\rho$, the annulus between $\rho$ and $\rho + \delta$ is contained in $B^c_\rho$ and thus the numerator differs only by an exponentially small term from the denominator in (2.5), if $\rho$ is chosen large enough so that the Gibbs measures are concentrated on $B_\rho$. This can be proven by some slight modifications of the estimates in ref. 4, in particular the proof of part (ii) of Lemma 4.2 in that paper. We will not give the details here. ▮

**Remark.** Of course the analog of Lemma 2.1 holds in many different situations. The crucial point is that we should consider the mass of a region in which much of the total mass is concentrated.

**Remark.** The restriction of the statement of the lemma to the subspace of $\omega$'s on which we can prove (1.10) is of course irrelevant, since it is only there that the conclusion of Theorem 1 holds. We will use Lemma 2.1 in the course of this section, but since the mass of the complement of this subspace is much smaller than all the probabilities we estimate here, this will make no difference for our estimates, and we will not make this explicit, to avoid overloading our notations.

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

**Proposition 2.2.** Assume that $\lim[M(N)/N]=\alpha$ and $\rho$ are as in Theorem 1, and let $\tilde{m}=\pm a(\beta)\, e^{\mu}$. Then, for all $n<\infty$ there exists $\tau_n<\infty$ such that for all $\tau \geqslant \tau_n$ and for $N$ large enough,

$$\mathbb{P}[\,|\tilde{F}_{N,\beta,\rho}(\tilde{m})-\mathbb{E}\tilde{F}_{N,\beta,\rho}(\tilde{m})| \geqslant \tau(\ln N)^{3/2}\, N^{-1/2}] \leqslant N^{-n} \tag{2.6}$$

Proposition 1.1 from the last section is of course an immediate corollary of this proposition together with Lemma 2.1. Thus all that is left is to prove Proposition 2.2.

*Proof of Proposition 2.2.* Let us set $f_N(\tilde{m}) \equiv N\tilde{F}_{N,\beta,\rho}(\tilde{m})$.

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

$$\tilde{f}_N^{(k)}(\tilde{m}) \equiv \mathbb{E}[f_N(\tilde{m})\,|\,\mathscr{F}_k]-\mathbb{E}[f_N(\tilde{m})\,|\,\mathscr{F}_{k+1}] \tag{2.7}$$

Notice that we have the identity

$$f_N(\tilde{m})-\mathbb{E}f_N(\tilde{m}) \equiv \sum_{k=1}^{N} \tilde{f}_N^{(k)}(\tilde{m}) \tag{2.8}$$

Let us recall that this construction was first introduced by Yurinskii[15] and employed in the context of spin-glasses and the Hopfield model by Pastur, Shcherbina, and Tirozzi.[11,14]

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

$$\begin{aligned}
&\mathbb{P}\left[\left|\sum_{k=1}^{N} \tilde{f}_N^{(k)}(\tilde{m})\right| \geqslant Nz\right] \\
&\quad \leqslant 2 \inf_{t \in \mathbb{R}} \exp(-|t|\, Nz)\, \mathbb{E}\exp\left\{t \sum_{k=1}^{N} \tilde{f}_N^{(k)}(\tilde{m})\right\} \\
&\quad = 2 \inf_{t \in \mathbb{R}} \exp(-|t|\, Nz)\, \mathbb{E}[\mathbb{E}[\cdots \mathbb{E}[\exp\{t\tilde{f}_N^{(1)}(\tilde{m})\}\,|\,\mathscr{F}_2] \\
&\qquad \times \exp\{t\tilde{f}_N^{(2)}(\tilde{m})\}\,|\,\mathscr{F}_3]\cdots \exp\{t\tilde{f}_N^{(N)}(\tilde{m})\}\,|\,\mathscr{F}_{N+1}]
\end{aligned} \tag{2.9}$$

Therefore, *if* we can show that, for some function $\mathscr{L}^{(k)}(t)$, $\ln \mathbb{E}[\exp\{t\tilde{f}_N^{(k)}(\tilde{m})\}\,|\,\mathscr{F}_{k+1}] \leqslant \mathscr{L}^{(k)}(t)$, uniformly in $\mathscr{F}_{k+1}$, then we obtain that

$$\mathbb{P}\left[\left|\sum_{k=1}^{N} \tilde{f}_N^{(k)}(\tilde{m})\right| \geqslant Nz\right] \leqslant 2 \inf_{t \in \mathbb{R}} \exp\left[-|t|\, Nz+\sum_{k=1}^{N} \mathscr{L}^{(k)}(t)\right] \tag{2.10}$$

To bound the conditional Laplace transforms, we introduce first, for $u \in [0, 1]$, the $M$-dimensional vectors

$$m_N^{(k)}(\sigma, u) \equiv \frac{1}{N}\left(\sum_{\substack{i \\ i \neq k}} \xi_i \sigma_i + u \xi_k \sigma_k\right) \tag{2.11}$$

and define

$$\tilde{H}_N^{(k)}(\sigma, u) = -\frac{N}{2} \|m_N^{(k)}(\sigma, u)\|_2^2 \tag{2.12}$$

Note that $\tilde{H}_N^{(k)}(\sigma, 1) = H_N(\sigma)$, while $\tilde{H}_N^{(k)}(\sigma, 0)$ does not depend on $\xi_k$. Naturally, we set

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

and finally

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

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

$$\tilde{f}_N^{(k)}(\tilde{m}) = \mathbb{E}[f_k(1) \mid \mathcal{F}_k] - \mathbb{E}[f_k(1) \mid \mathcal{F}_{k+1}] \tag{2.15}$$

To bound the Laplace transform, we use that, for all $x \in \mathbb{R}$,

$$e^x \leqslant 1 + x + \tfrac{1}{2} x^2 e^{|x|} \tag{2.16}$$

so that

$$\mathbb{E}[e^{t \tilde{f}_N^{(k)}(\tilde{m})} \mid \mathcal{F}_{k+1}] \leqslant 1 + \tfrac{1}{2} t^2 \mathbb{E}[(\tilde{f}_N^{(k)}(\tilde{m}))^2 e^{|t \tilde{f}_N^{(k)}(\tilde{m})|} \mid \mathcal{F}_{k+1}] \tag{2.17}$$

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. A simple computation shows that

$$\begin{aligned} f'_k(u) = \mathcal{E}_{k,u}\Bigg(&\sum_\mu \xi_k^\mu \sigma_k m_N^{(k),\mu}(\sigma, u) \\ &+ \frac{1}{\beta N} \frac{\chi'_{\rho,\delta}(\|m_N^{(k)}(\sigma, u) - \tilde{m}\|_2)}{\chi_{\rho,\delta}(\|m_N^{(k)}(\sigma, u) - \tilde{m}\|_2)} \sum_\mu \frac{(m_N^{(k),\mu}(\sigma, u) - \tilde{m}^\mu)}{\|m_N^{(k)}(\sigma, u) - \tilde{m}\|_2} \xi_k^\mu \sigma_k\Bigg) \end{aligned} \tag{2.18}$$

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

$$\frac{1}{Z_N^{(k)}(\tilde{m}, u)} \chi_{\rho,\delta}(\|m_N^{(k)}(\sigma, u) - \tilde{m}\|_2)\, e^{-\beta \tilde{H}_N^{(k)}(\sigma,u)}\, d\sigma \tag{2.19}$$

We see that $f'_k(0) = 0$, and thus, as $f_k(0) = 0$ and $f_k(u)$ is a concave function of $u$, $|f_k(1)| \leqslant |f'_k(1)|$. Now

$$\begin{aligned}
&\mathscr{E}_{k,1} \left| \frac{1}{\beta N} \frac{\chi'_{\rho,\delta}(\|m_N(\sigma) - \tilde{m}\|_2)}{\chi_{\rho,\delta}(\|m_N(\sigma) - \tilde{m}\|_2)} \sum_\mu \frac{m_N^\mu(\sigma) - \tilde{m}^\mu}{\|m_N(\sigma) - \tilde{m}\|_2} \xi_k^\mu \sigma_k \right| \\
&\qquad \leqslant \sup_\sigma \frac{\|m_N(\sigma) - \tilde{m}\|_1}{\|m_N(\sigma) - \tilde{m}\|_2} \frac{2}{\beta\delta N} \left( \frac{Z_{N,\beta}(B_{\rho+\delta}^{(\eta,s)})}{Z_{N,\beta}(B_\rho^{(\eta,s)})} - 1 \right) \\
&\qquad \leqslant \sqrt{M} \frac{2}{\beta\delta N}
\end{aligned} \tag{2.20}$$

where we have used Lemma 2.1 for the last inequality. Thus, using the bound

$$\mathscr{E}_{k,1} \left| \sum_\mu \xi_k^\mu m_N^\mu(\sigma) \right| \leqslant \left| \sum_\mu \xi_k^\mu \tilde{m}^\mu \right| + \sqrt{M}(\rho + \delta) \leqslant \|\tilde{m}\|_1 + \sqrt{M}(\rho + \delta) \tag{2.21}$$

we get

$$|f'_k(1)| \leqslant \|\tilde{m}\|_1 + \sqrt{M} \left( (\rho + \delta) + \frac{2}{\beta\delta N} \right) \tag{2.22}$$

We will now choose $\delta = 2/\beta N \rho$, so that we get effectively the bound

$$|f'_k(1)| \leqslant \|\tilde{m}\|_1 + 2\sqrt{M}\,\rho \tag{2.23}$$

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

$$\mathbb{E}[e^{t\tilde{f}_N^{(k)}(\tilde{m})} \,|\, \mathscr{F}_{k+1}] \leqslant 1 + \tfrac{1}{2} t^2 e^{2|t|(\|\tilde{m}\|_1 + 2\sqrt{M}\rho)} \mathbb{E}[(\tilde{f}_N^{(k)}(\tilde{m}))^2 \,|\, \mathscr{F}_{k+1}] \tag{2.24}$$

Of course we could also use (2.23) to bound the expectation of the square in (2.24), but due to the presence of the $\sqrt{M}$ in that bound, this would not be very useful.

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

$$\begin{aligned}\mathbb{E}[(\tilde{f}_N^{(k)}(\tilde{m}))^2 \mid \mathscr{F}_{k+1}] &= \mathbb{E}[\{\mathbb{E}[f_k(1) - \mathbb{E}[f_k(1) \mid \mathscr{F}_{k+1}] \mid \mathscr{F}_k]\}^2 \mid \mathscr{F}_{k+1}] \\ &\leqslant \mathbb{E}[\mathbb{E}[\{f_k(1) - \mathbb{E}[f_k(1) \mid \mathscr{F}_{k+1}]\}^2 \mid \mathscr{F}_k] \mid \mathscr{F}_{k+1}] \\ &= \mathbb{E}[\{f_k(1) - \mathbb{E}[f_k(1) \mid \mathscr{F}_{k+1}]\}^2 \mid \mathscr{F}_{k+1}] \\ &= \mathbb{E}[(f_k(1))^2 \mid \mathscr{F}_{k+1}] - \{\mathbb{E}[f_k(1) \mid \mathscr{F}_{k+1}]\}^2 \\ &\leqslant \mathbb{E}[(f_k(1))^2 \mid \mathscr{F}_{k+1}] \leqslant \mathbb{E}[(f'_k(1))^2 \mid \mathscr{F}_{k+1}] \end{aligned} \tag{2.25}$$

Let us use the fact that $(a+b)^2 \leqslant 2a^2 + 2b^2$ and (2.18) to see that

$$\begin{aligned}(f'_k(1))^2 \leqslant{}& 2\left\{\mathscr{E}_{k,1}\left(\sum_\mu \xi_k^\mu \sigma_k m_N^\mu(\sigma)\right)\right\}^2 \\ &+ 2\left\{\sum_\mu \mathscr{E}_{k,1}\left(\frac{1}{\beta N}\frac{\chi'_{\rho,\delta}(\|m_N(\sigma)-\tilde{m}\|_2)}{\chi_{\rho,\delta}(\|m_N(\sigma)-\tilde{m}\|_2)}\frac{(m_N^\mu(\sigma)-\tilde{m}^\mu)}{\|m_N(\sigma)-\tilde{m}\|_2}\xi_k^\mu\sigma_k\right)\right\}^2\end{aligned} \tag{2.26}$$

Using the Schwarz inequality, we get from this that

$$\begin{aligned}(f'_k(1))^2 \leqslant{}& 2\mathscr{E}_{k,1}\left\{\sum_\mu \xi_k^\mu \sigma_k m_N^\mu(\sigma)\right\}^2 \\ &+ 2\mathbb{E}\left[\mathscr{E}_{k,1}\left(\frac{1}{\beta N}\frac{\chi'_{\rho,\delta}(\|m_N(\sigma)-\tilde{m}\|_2)}{\chi_{\rho,\delta}(\|m_N(\sigma)-\tilde{m}\|_2)}\right)^2\right. \\ &\times \left.\mathscr{E}_{k,1}\frac{\{\sum_\mu \xi_k^\mu(m_N^\mu(\sigma)-\tilde{m}^\mu)\}^2}{\|m_N(\sigma)-\tilde{m}\|_2^2}\right| \mathscr{F}_{k+1}\Bigg]\end{aligned} \tag{2.27}$$

The expectation of $\chi'/\chi$ is bounded using Lemma 2.1 as in (2.20) and just gives a factor $\rho^2$, with the previous choice of $\delta$. To deal with the first term, we use the following crucial trick: $\mathscr{E}_{k,1}$ is in fact independent of $k$, and therefore the expectations conditioned on $\mathscr{F}_{k+1}$ are the same if the index $k$ inside it is replaced by any of the indices $j \in \{1,..., k\}$. This allows us to derive from (2.27)

$$\begin{aligned}&\mathbb{E}[(f'_k(1))^2 \mid \mathscr{F}_{k+1}] \\ &\quad\leqslant 2\mathbb{E}\left[\mathscr{E}_{k,1}\left(\frac{1}{k}\sum_{j=1}^k \sum_\mu \sum_\nu \xi_j^\mu \xi_j^\nu m_N^\nu(\sigma)\, m_N^\mu(\sigma)\, m_N^\mu(\sigma)\right)\middle| \mathscr{F}_{k+1}\right] \\ &\qquad + 2\rho^2\mathbb{E}\left[\mathscr{E}_{k,1}\frac{1}{k}\sum_{j=1}^k \frac{\sum_\mu \sum_\nu \xi_j^\mu \xi_j^\nu (m_N^\mu(\sigma)-\tilde{m}^\mu)(m_N^\nu(\sigma)-\tilde{m}^\nu)}{\|m_N(\sigma)-\tilde{m}\|_2^2}\middle| \mathscr{F}_{k+1}\right]\end{aligned} \tag{2.28}$$

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

$$B_{\mu\nu}^{(k)} \equiv \frac{1}{k} \sum_{j=1}^{k} \xi_j^\mu \xi_j^\nu \tag{2.29}$$

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

We can write (2.28) in the form

$$\begin{aligned}\mathbb{E}[(f_k'(1))^2 \mid \mathcal{F}_{k+1}] \leqslant{}& 2\mathbb{E}[\mathscr{E}_{1,k}((m_N(\sigma), B^{(k)} m_N(\sigma))) \mid \mathcal{F}_{k+1}] \\ &+ 2\rho^2 \mathbb{E}\left[\mathscr{E}_{1,k} \frac{((m_N(\sigma) - \tilde{m}), B^{(k)}(m_N(\sigma) - \tilde{m}))}{\|m_N(\sigma) - \tilde{m}\|_2^2} \middle| \mathcal{F}_{k+1}\right]\end{aligned} \tag{2.30}$$

Here $(\cdot, \cdot)$ denotes the scalar product in $\mathbb{R}^M$. The second summand is immediately seen to be bounded by $2\rho^2 \mathbb{E} b_k$, while for the first we write

$$\begin{aligned}&2\mathbb{E}[\mathscr{E}_{1,k}((m_N(\sigma), B^{(k)} m_N(\sigma))) \mid \mathcal{F}_{k+1}] \\ &\quad = 2\mathbb{E}\left[\sum_{\mu,\nu} \frac{1}{k} \sum_{j=1}^{k} \xi_j^\mu \xi_j^\nu \mathscr{E}_{1,k}((m_N^\mu(\sigma) - \tilde{m}^\mu)(m_N^\nu(\sigma) - \tilde{m}^\nu)) \mid \mathcal{F}_{k+1}\right] \\ &\qquad + 4\mathbb{E}\left[\sum_{\mu,\nu} \frac{1}{k} \sum_{j=1}^{k} \xi_j^\mu \xi_j^\nu \mathscr{E}_{1,k}((m_N^\mu(\sigma) - \tilde{m}^\mu)\, \tilde{m}^\nu) \mid \mathcal{F}_{k+1}\right] \\ &\qquad + 2\mathbb{E}\left[\sum_{\mu,\nu} \frac{1}{k} \sum_{j=1}^{k} \xi_j^\mu \xi_j^\nu \tilde{m}^\mu \tilde{m}^\nu \mid \mathcal{F}_{k+1}\right] \\ &\quad = 2\mathbb{E}[\mathscr{E}_{1,k}((m_N(\sigma) - \tilde{m}), B^{(k)}(m_N(\sigma) - \tilde{m})) \mid \mathcal{F}_{k+1}] \\ &\qquad + 4\mathbb{E}[\mathscr{E}_{1,k}((m_N(\sigma) - \tilde{m}), B^{(k)} \tilde{m}) \mid \mathcal{F}_{k+1}] \\ &\qquad + 2\mathbb{E}\left[\frac{1}{k} \sum_{j=1}^{k} \left(\sum_{\mu} \xi_j^\mu \tilde{m}^\mu\right)^2\right] \\ &\quad \leqslant [2(\rho + \delta)^2 + 4\, \|\tilde{m}\| (\rho + \delta)]\, \mathbb{E} b_k + 2\, \|\tilde{m}\|_2^2\end{aligned} \tag{2.31}$$

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

Collecting our bounds and inserting them into (2.24), we have (ignoring the difference between $\rho$ and $\rho + \delta$)

$$\begin{aligned}&\mathbb{E}[e^{t\tilde{\mathcal{F}}_N^{(k)}(\tilde{m})} \mid \mathcal{F}_{k+1}] \\ &\quad \leqslant 1 + \tfrac{1}{2} t^2 e^{2\,|t|(\|\tilde{m}\|_1 + 2\sqrt{M}\rho)} [2\, \|\tilde{m}\|_2^2 + (2\rho^2 + 4\, \|\tilde{m}\|\, \rho)\, \mathbb{E} b_k] \\ &\quad \leqslant \exp\{\tfrac{1}{2} t^2 e^{2\,|t|(\|\tilde{m}\|_1 + 2\sqrt{M}\rho)} [2\, \|\tilde{m}\|_2^2 + (2\rho^2 + 4\, \|\tilde{m}\|\, \rho)\, \mathbb{E} b_k]\}\end{aligned} \tag{2.32}$$

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

$$\begin{aligned}\mathbb{E}\exp\left\{t\sum_{k=1}^{N}\tilde{f}_N^{(k)}(\tilde{m})\right\}\\ \leqslant\exp\left\{\tfrac{1}{2}t^2e^{2\,|t|(\|\tilde{m}\|_1+2\sqrt{M}\,\rho)}\left[N2\,\|\tilde{m}\|_2^2+(2\rho^2+4\,\|\tilde{m}\|\,\rho)\sum_{k=1}^{N}\mathbb{E}b_k\right]\right\}\end{aligned}\tag{2.33}$$

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 refs. 14 and 2. We will use the bounds on the traces of powers of such matrices proven in ref. 2 to deduce

**Lemma 2.2.** Let $B^{(k)}$ denote the $M\times M$ matrices with elements defined in (2.29). Then

$$\mathbb{E}\,\|B^{(k)}\|\leqslant\begin{cases}2\dfrac{M}{k}+2e\left(\dfrac{M}{k}\right)^{1/2} & \text{if}\quad k\leqslant M\\[2ex] 2+2e\left(\dfrac{M}{k}\right)^{1/2} & \text{if}\quad k\geqslant M\end{cases}\tag{2.34}$$

From this lemma it follows that

$$\begin{aligned}\sum_{k=1}^{N}\mathbb{E}b_k&\leqslant 2\sum_{k=1}^{M}M/k+2e\sum_{k=1}^{N}(M/k)^{1/2}+2(N-M)\\ &\leqslant c[M\ln M+N+(M/N)^{1/2}]\end{aligned}\tag{2.35}$$

for some numerical constant $c$. Using these estimates, we get

$$\begin{aligned}\mathbb{E}\exp\left\{t\sum_{k=1}^{N}\tilde{f}_N^{(k)}(\tilde{m})\right\}\\ \leqslant\exp\left\{\frac{1}{2}t^2e^{2\,|t|(\|\tilde{m}\|_1+2\sqrt{M}\,\rho)}N\left[2\,\|\tilde{m}\|_2+4c''\rho\,\|\tilde{m}\|_2^2\right.\right.\\ \left.\left.+c(2\rho^2+4\,\|\tilde{m}\|\,\rho)\frac{M}{N}\ln N\right]\right\}\end{aligned}\tag{2.36}$$

Let us remark that we will use this bound only for $\tilde{m}$ with bounded $l^2$-norm and for $\rho$ and $M/N$ much smaller than 1. Thus (2.36) takes on the simple form

$$\mathbb{E}\exp\left\{t\sum_{k=1}^{N}\tilde{f}_N^{(k)}(\tilde{m})\right\}\leqslant\exp\left\{ct^2e^{c'\,|t|\,\sqrt{M}}N\left(1+\rho\frac{M}{N}\ln N\right)\right\}\tag{2.37}$$

for (new) constants $c$ and $c'$. Equation (2.37) can now be used together with (2.9) 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 interval $[0, M^{-1/2}]$, so that in particular no exponential estimates can be obtained. But for our present purposes we will not need this. In fact, the most convenient bounds for us are derived by choosing $t = n(\ln N)/zN$. This yields (we put $M/N = \alpha$) that

$$\mathbb{P}[\,|f_N(\tilde{m}) - \mathbb{E}f_N(\tilde{m})| \geqslant Nz] \\ \leqslant N^{-n} \exp\left\{c\,\frac{(\ln N)^2\, n^2(1 + \rho\alpha \ln N)}{z^2 N}\, N^{c'(M/N)^{1/2}\, zN^{1/2}}\right\} \tag{2.38}$$

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 nonnormalized 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 ref. 14 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

$$Z_{N,\beta} = Z^{<}_{N,\beta} + Z^{>}_{N,\beta} \tag{2.39}$$

where

$$Z^{<}_{N,\beta} \equiv \frac{1}{2^N} \sum_{\sigma \in \mathcal{S}_N} e^{-\beta H_N(\sigma)} \mathbb{1}_{\{\|m_N(\sigma)\|_2 \leqslant 2\}} \tag{2.40}$$

and

$$Z^{>}_{N,\beta} \equiv \frac{1}{2^N} \sum_{\sigma \in \mathcal{S}_N} e^{-\beta H_N(\sigma)} \mathbb{1}_{\{\|m_N(\sigma)\|_2 > 2\}} \tag{2.41}$$

But $\|m_N(\sigma)\|_2^2 \leqslant \|A\|$, where $A$ is the $N \times N$ matrix with elements $A_{ij} = (1/N) \sum_{\mu=1}^{N} \xi_i^\mu \xi_j^\mu$. This matrix has obviously the same norm as the matrix $B^{(N)}$ defined above, so that the estimates on the norm of these random matrices from ref. 14 or ref. 2 can be used. It follows in particular that this norm is less than two with probability at least $1 - e^{-N^{1/6}}$. Therefore,

$$\mathbb{P}[Z^{>}_{N,\beta} = 0] \geqslant 1 - e^{-N^{1/6}} \tag{2.42}$$

Since on the other hand the deviation of $\ln Z^{<}_{N,\beta}[\omega]$ from its mean is easily shown to satisfy the bound (2.6), we obtain the statement of the proposition. ▌

## 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 us to prove Theorem 1. Let us note 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 $\alpha$ 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 $\alpha$ 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 $\alpha$ (depending, however, on $\beta$) *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[13] model of a spin glass. We recall that the Hamiltonian of this model is given by

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

where $\{J_{ij}\}_{i<j\in\mathbb{N}\times\mathbb{N}}$ is a family of independent Gaussian random variables with mean zero and variance one. Pastur and Shcherbina[11] have proven that in this model

$$\mathbb{E}[(F_{N,\beta}-\mathbb{E}F_{N,\beta})^2]\leqslant\frac{c}{N} \tag{3.2}$$

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 shown that in fact the following holds.

**Proposition 3.1.** In the Sherrington–Kirkpatrick model, for all $\beta>0$ and for all $\infty>z\geqslant 0$

$$\mathbb{P}[\,|F_{N,\beta}-\mathbb{E}F_{N,\beta}|\geqslant z]\leqslant 2\exp(-Nz^2) \tag{3.3}$$

*Proof.* The basic idea of the proof of this proposition is the same as the one used in the Hopfield model. However, to get the optimal constant in the exponent in (3.3) we use the following additional trick, which is specific for Gaussian $J_{ij}$. Namely, we may represent the Gaussian variables $J_{ij}$ as sums of independent copies $J^\mu_{ij}$ in the form

$$J_{ij}=\frac{1}{\sqrt{K}}\sum_{\mu=1}^{K}J^\mu_{ij} \tag{3.4}$$

for arbitrarily chosen $K$. Let us introduce an arbitrary enumeration of the $KN(N-1)/2$ independent random variables $J^\mu_{ij}$ and write for them $J(1)$, $J(2),\dots, J(KN(N-1)/2)$.

We denote by $\mathscr{F}_k$ the sigma algebra generated by the random variables $\{J(m)\}_{m\geqslant k}$.

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

$$F_{N,\beta}-\mathbb{E}F_{N,\beta}=\frac{1}{N}\sum_{k=1}^{KN(N-1)/2}\tilde f_N^{(k)} \tag{3.5}$$

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 analog of (2.12) is

$$\begin{aligned}\tilde H_N^{(k)}(\sigma,u)\equiv&-\frac{1}{(KN)^{1/2}}\sum_{(i,j,\mu)\neq(i(k),j(k),\mu(k))}J^\mu_{ij}\sigma_i\sigma_j\\&-u\frac{1}{(KN)^{1/2}}J^{\mu(k)}_{i(k),j(k)}\sigma_{i(k),j(k)}\end{aligned} \tag{3.6}$$

This yields that this time

$$f'_k(u) = \frac{1}{(NK)^{1/2}} J(k)\, \mathscr{E}_{k,u} \sigma_{i(k)} \sigma_{j(k)} \tag{3.7}$$

Trivially, here

$$|f'_k(u)| \leqslant \frac{1}{(KN)^{1/2}} |J(k)| \tag{3.8}$$

Let us use this time that

$$e^x \leqslant 1 + x + \frac{x^2}{2} + \frac{|x|^3}{6} e^{|x|} \tag{3.9}$$

Therefore, using (2.16), we get

$$\begin{aligned} \mathbb{E}[e^{t\tilde{f}_N^{(k)}} \,|\, \mathscr{F}_{k+1}] \leqslant 1 + \frac{t^2}{2} \mathbb{E}[(\tilde{f}_N^{(k)})^2 \,|\, \mathscr{F}_{k+1}] \\ + \frac{|t|^3}{6} \mathbb{E}[\,|\tilde{f}_N^{(k)}|^3\, e^{|t|\,|\tilde{f}_N(k)|} \,|\, \mathscr{F}_{k+1}] \end{aligned} \tag{3.10}$$

To bound the first expectation, we can proceed as in (2.25) and just note that by convexity,

$$|f_k(1)| \leqslant \max(|f'_k(0)|, |f'_k(1)|) \leqslant \frac{1}{(KN)^{1/2}} |J(k)|$$

Thus

$$\mathbb{E}[(\tilde{f}_N^{(k)})^2 \,|\, \mathscr{F}_{k+1}] \leqslant \frac{1}{KN} \tag{3.11}$$

For the second term we can be less careful and use just that, by (2.15),

$$|\tilde{f}_N^{(k)}| \leqslant 2\, |f_k(1)| \leqslant \frac{2}{(KN)^{1/2}} |J(k)|$$

Thus

$$\begin{aligned} \mathbb{E}[\,|\tilde{f}_N^{(k)}|^3\, e^{|t|\,|\tilde{f}_N(k)|} \,|\, \mathscr{F}_{k+1}] &\leqslant \frac{8}{(KN)^{3/2}} \mathbb{E}\, |J|^3\, e^{|t|\,|J|/(KN)^{1/2}} \\ &\leqslant \frac{C}{(KN)^{3/2}} e^{ct^2/(KN)} \end{aligned} \tag{3.12}$$

for some universal numerical constants $c$, $C$. Thus we obtain

$$\begin{aligned}\mathbb{P}\left[\left|\sum_{k=1}^{KN(N-1)/2} \tilde{J}_N^{(k)}\right| \geqslant Nz\right] &\leqslant 2 \inf_{t \in \mathbb{R}} \exp\left[-|t|\, Nz + \frac{(N-1)\, t^2}{4} + C\left(\frac{N}{K}\right)^{1/2} |t|^3\, e^{ct^2/(KN)}\right] \\ &\leqslant 2 \exp\left[-N\frac{N}{N-1} z^2 + Cz^3\left(\frac{N}{K}\right)^{1/2} e^{4z^2/(KN)}\right] \qquad (3.13)\end{aligned}$$

Since (3.13) holds for arbitrary $K$, this proves the proposition. ▌

**Remark.** Let us note that Proposition 3.1 can also be derived by applying a concentration estimate that is given, for instance, in ref. 9, p. 21, Eq. (1.6); as pointed out there, however, the proof of that inequality (with the sharp constant) employs more sophisticated techniques of stochastic calculus. Our proof, being fairly elementary, may thus still be useful. We thank M. Talagrand for having brought this to our attention.

Proposition 3.1 implies in particular the almost sure convergence to zero of $F_{N,\beta} - \mathbb{E}F_{N,\beta}$. This does not imply the almost sure convergence of the free energy, since it is not known that the mean of the free energy converges 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.

## ACKNOWLEDGMENT

This work was partially supported by the Commission of the European Communities under contract SC1-CT91-0695.

## REFERENCES

1. D. J. Amit, H. Gutfreund, and H. Sompolinsky, Statistical mechanics of neural networks near saturation, *Ann. Phys.* **173**:30–67 (1987).
2. A. Bovier and V. Gayrard, Rigorous results on the thermodynamics of the dilute Hopfield model, *J. Stat. Phys.* **69**:627 (1993).
3. A. Bovier and V. Gayrard, Rigorous results on the Hopfield model of neural networks, *Resenhas IME-USP* **2**:161–172 (1994).
4. A. Bovier, V. Gayrard, and P. Picco, Gibbs states of the Hopfield model in the regime of perfect memory, *Prob. Theory Related Fields* **100**:329–363 (1994).

5. A. Bovier, V. Gayrard, and P. Picco, Large deviation principles for the Hopfield model and the Kac–Hopfield model, *Prob. Theory Related Fields* (1995), to appear.
6. L. A. Pastur and A. L. Figotin, Exactly soluble model of a spin glass, *Sov. J. Low Temp. Phys.* **3**(6):378–383 (1977).
7. L. A. Pastur and A. L. Figotin, On the theory of disordered spin systems, *Theor. Math. Phys.* **35**:403–414 (1978).
8. J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci. USA* **79**:2554–2558 (1982).
9. M. Ledoux and M. Talagrand, *Probability in Banach Spaces* (Springer, Berlin, 1991).
10. M. Mézard, G. Parisi, and M. A. Virasoro, *Spin-Glass Theory and Beyond* (World Scientific, Singapore, 1988).
11. 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).
12. 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).
13. D. Sherrington and S. Kirkpatrick, Solvable model of a spin glass, *Phys. Rev. Lett.* **35**:1792–1796 (1972).
14. M. Shcherbina and B. Tirozzi, The free energy for a class of Hopfield models, *J. Stat. Phys.* **72**:113–125 (1992).
15. V. V. Yurinskii, Exponential inequalities for sums of random vectors, *J. Multivariate Anal.* **6**:473–499 (1976).