Report 99-005

The Storage Capacity of the Hopfield Model and Moderate Deviation Principles

Matthias Löwe

ISSN: 1389–2355

THE STORAGE CAPACITY OF THE HOPFIELD MODEL AND MODERATE DEVIATION PRINCIPLES

MATTHIAS LÖWE

ABSTRACT. This note relates the storage capacity of the Hopfield model of neural networks to the existence of a moderate deviation principle for the empirical correlation of the patterns. This moderate deviation principle is satisfied under a certain condition on the moment generating function of these correlations which on the other hand can be verified in many cases by GHS- and FKG-type inequalities. Examples of such situations will be given.

1. INTRODUCTION AND THE BASIC SETUP

The Hopfield model is the simplest and best-studied model of a neural network. Originally introduced by Pastur and Figotin [FP77] as a so-called frustrated system, it received most of its attention by its reinterpretation by Hopfield [Ho82] as a very simple model for the brain. Its closeness to the spin-glass models, in particular to the Sherrington-Kirkpatrick model, evoked the physicists' interest in the model and led to a number of papers that claimed to rigorously "solve" the model (see, e.g. [AGS87]). Unfortunately, their techniques which go under name "replica method" (see e.g. [MPV87] for an survey over these techniques) not only involve some mathematically completely unjustified operations (such as interchanging different limits) but also introduce objects which for a mathematician are hard to understand (e.g. the largest eigenvalue of a symmetric $N \times N$ matrix when N goes to zero). So, though the Hopfield model has been extensively studied yet the number of mathematically clean results is limited, has basically been found in the last decade and is – with a few exceptions – restricted to the case where the so called patterns are chosen to be i.i.d. a case which is not very close to a realistic situation. The corresponding results have been proven in a number of papers by Bovier, Gayrard, partially in collaboration with Picco – for an exhaustive and very readable survey see [BP98] and especially [BG98] therein and all the references given there (e.g. [BG96a], [BG97a], [BG97b]) – and the fundamental paper by Talagrand [T98].

In this little note we will treat Hopfield models with correlated patterns and prove a result on their storage capacity. The question of the storage capacity has been asked for Hopfield models with i.i.d. patterns in different ways and by different authors. The definition of storage capacity we use in this note (which in a way is the most fundamental one) has been considered for i.i.d. patterns by McEliece et al. [MPRV87]. Rigorous proofs can be found in the overview paper by Petritis [P96] and extensions are due to Burshtein [Bu94]. Another, more liberal, definition of storage capacity (which also allows minor errors in the reconstruction of the stored

Date: February 1, 1999.

1991 *Mathematics Subject Classification.* 82C32, 82B44, 60K35.

Key words and phrases. Hopfield model, neural networks, storage capacity, moderate deviations, large deviations, spin glasses.

patterns) has basically been introduced by Amit et al. [AGS87] and rigorously treated by Newman [N88]. Improvements of his results are due to Loukianova [Lou94] and Talagrand [T95], [T98]. Both notions of storage capacity have been treated for weakly dependent patterns produced by a Markov chain in [Lö99a].

To be more specific let us define the Hopfield model. First of all we choose two numbers $N, M \in \mathbb{N}$ which will denote the number of spins or "neurons" and the number of so-called patterns, respectively. Note that $M = M(N)$ may and actually will depend on N. We shall write M and thus drop its dependency on N whenever there is no danger of confusion. The random function

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

denotes the so-called Hamiltonian of the Hopfield model, which is a function of the spin configuration $\sigma \in \{-1,+1\}^N$. This function is random as the variables $\xi_i^\mu \in \{-1,+1\}$ with ξ_i^μ denoting the ith component of the μth pattern are chosen randomly. In most of the papers on the Hopfield model it is generally assumed that the ξ_i^μ are i.i.d. unbiased random variables, i. e., that at given system size N, the family of random variables $\{\, \xi_i^\mu : i \in \{1,\dots,N\}, \mu \in \{1,\dots,M(N)\} \,\}$ is independent with

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

for all i and μ. Here and in the following $\mathbb{P}$ stands for the distribution of the $(\xi_i^\mu)_{i,\mu}$ while we denote by $\mathbb{E}$ expectations with respect to $\mathbb{P}$.
The case of independent but biased patterns, i.e. the case where (2) is violated, has been treated e.g. in [Lö99b].
In this paper we will consider the case where the ξ_i^μ may be correlated in such a way that a special condition on the four-point-correlation function (see Assumption 2.1 below) is fulfilled and the marginal distributions of the ξ_i^μ are still unbiased, i.e. (2) is still satisfied. For correlated, biased patterns some additional term has to be added to the Hamiltonian (1) to make the model work. As with this correction term neither the results nor the techniques differ very much from the ones presented here (for a discussion we refer the reader to [Lö99b]), we rather prefer not to treat the case of unbiased patterns in any length.

Whenever convenient, we shall write ξ for the $(N \times M)$-matrix consisting of the $(\xi_i^\mu)_{i,\mu}$, while $\xi_i = (\xi_i^1, \dots, \xi_i^M)$ and $\xi^\mu = (\xi_1^\mu, \dots, \xi_N^\mu)$, respectively, stand for the ith row and the μth column of this matrix, respectively.

The spin variables are assumed to be independent with an unbiased a priori distribution P, i. e.,

$$\mathrm{P}(\sigma_i = +1) = \mathrm{P}(\sigma_i = -1) = \frac{1}{2}$$

for all $i \in \mathbb{N}$.

The Hopfield model may now either be identified with the Hamiltonian (1) or, equivalently, with the Gibbs measure at temperature $1/\beta \in (0,\infty)$ with respect to the Hamiltonian (1), i. e.,

$$\varrho_{N,\beta}(\sigma) = 2^{-N} \exp\{-\beta H_N(\sigma)\}/\Sigma_{N,\beta}, \qquad \sigma \in \{-1,+1\}^N, \tag{3}$$

where the so-called partition function

$$\Sigma_{N,\beta} = \frac{1}{2^N} \sum_{\sigma \in \{-1,+1\}^N} \exp\{-\beta H_N(\sigma)\} \tag{4}$$

is the normalisation which makes $\varrho_{N,\beta}$ a probability measure.

The idea behind this setup is the following. Suppose for the moment that $M \equiv 1$. Then $H_N(\sigma)$ clearly has two minima at $\sigma_i = \xi_i^1$ for all i or $\sigma_i = -\xi_i^1$ for all i. Now, for any M, if the ξ_i^μ are chosen as i.i.d. random variables, by the the Central Limit Theorem

$$\frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu = \delta_{\mu,\nu} + \mathcal{O}(\frac{1}{\sqrt{N}}) \tag{5}$$

for each choice of μ and ν (fixed), suggesting that for $M(N)$ not growing too fast (as a function of N) still the ξ^μ are minima of the Hamiltonian and thus ground states of the system. So, if we interpret the ξ^μ's as stored information in a brain consisting of the neurons σ_i we might be able to even reconstruct noised information by a stochastic retrieval dynamics, such as the Monte-Carlo dynamics, which favours states of low H_N-value. (And, indeed, such a Monte-Carlo dynamics at temperature $1/\beta$ has $\varrho_{N,\beta}$ as its invariant measure.)

As this heuristics and, in particular (5), is not available for correlated patterns, several authors doubted that the Hopfield model in the current setup would be able to store any increasing number $M(N)$ of correlated patterns. In [Lö99a] we have been able to show that this is indeed the case, provided the variables ξ_i^μ are either correlated in μ or in i and independent in the other variable and that the correlations stems from a one dimensional Markov chain. Unfortunately, the proofs found there are not easy to transfer to any other situation (for example to the relevant and interesting situation where the ξ_i^μ for every fixed μ describe a picture, thus a two dimensional random field) since it heavily exploits the martingale structure of one dimensional Markov chains.

In this note we will show that under a condition which is easy to verify in many important examples and which is closely related to a so-called moderate deviation principle for the empirical four point correlations, also correlated patterns may be stored in the Hopfield model described above, provided we use the notion of storage introduced by McEliece et. al. [MPRV87].

This little note has two further sections: To be able to describe our result in Section 2 we first will specify the notion of storage we have in mind, then introduce our central assumption (Assumption 2.1) which is closely related to a moderate deviation principle for the patterns and finally give the actual result. Section 3 contains the proof together with a list of the most important examples.

Acknowledgement: I am thankful to Peter Eichelsbacher for bringing reference [Wu95] to my attention. I also would like to express my gratitude to Wim Senden for helping me with the technical details.

2. The Notion of Storage Capacity and the Main Result

In this section we will mainly state our result on the storage capacity of the Hopfield model with correlated patterns.
First let us first briefly explain the concept of storage we are dealing with. The idea behind it is that a possible retrieval dynamics is a Monte–Carlo dynamics at zero temperature working as follows: Choose a site i at random. Flip the spin σ_i, if flipping lowers the energy (the Hamiltonian) and let the spin σ_i unchanged otherwise. On a more formal level we define the gradient dynamics T on the energy landscape on $\{-1,+1\}$ induced by H_N via

$$T_i : \sigma_i \mapsto \operatorname{sgn}(\sum_{j\neq i} \sigma_j J_{ij})$$

where sgn is the sign function. The map T is then defined by $T(\sigma) := (T_i(\sigma_i))_i$. We will call a configuration $\sigma = (\sigma_i)_{i\leq N}$ stable if it is a fixed point of T, i.e.

$$\sigma_i = \operatorname{sgn}(\sum_{j\neq i} \sigma_j J_{ij}) \quad \text{for all } i = 1, \dots, N$$

which means that σ is a local minimum of the Hamiltonian. The storage capacity in this concept is defined as the asymptotics of the greatest number of patterns $M := M(N)$ such that all the patterns ξ^ν are stable in the above sense almost surely or with probability converging to one. (Here and in the following the notion "almost surely" refers to the probability measure on the space of all sequences of patterns (of infinite length) while in "probability converging to one" the convergence is with $N \to \infty$). Note that this concept is a very natural way to define "storage capacity", since that the stored information is stable under the retrieval dynamics is in some sense "the least we would expect".

Let us quickly mention another approach to storage capacity which is due to Amit, Gutfreund and Sompolinsky [AGS85] and has rigorously been analysed by Newman [N88]. It takes into consideration that we possibly are willing to tolerate small errors in the restoration of the patterns. So we are satisfied, if the retrieval dynamics converges to a configuration which is not too far away from the original patterns. Thus in this concept a pattern ξ^ν is called stable, if it is close to a local minimum of the Hamiltonian, or, in other words, if it is surrounded by a sufficiently high energy barrier. Technically speaking we will call ξ^ν stable if there exist $\varepsilon > 0$ and $\delta > 0$ such that

$$\inf_{\sigma\in S_\delta(\xi^\nu)} H_N(\sigma) \geq H_N(\xi^\nu) + \varepsilon N. \tag{6}$$

Here the set $S_\delta(\xi^\nu)$, the infimum is taken over, is the Hamming sphere of radius δN centred in ξ^ν. Again we will use the notion of storage capacity for the maximal number $M(N)$ of patterns such that (6) holds true for all ξ^ν almost surely.

Before stating our main result we have to make one central assumption. To describe the consequences of this assumption we have to dwell a bit on large an moderate deviation theory.

In general, if X_n is any sequence of random variables, we say that it obeys a large deviation principle (LDP) with speed ε_n^{-1} and rate function $I(\cdot)$, if $\lim_{n\to\infty} \varepsilon_n = 0$,

and

$$0 \not\equiv I(\cdot) \leq \infty$$

is lower semi-continuous, the level sets $\{x : I(x) \leq L\}$ are compact, and if for all Borel sets A the following inequalities hold

$$-\inf_{x\in A^\circ} I(x) \leq \liminf_{n\to\infty} \varepsilon_n \log P(X_n \in A^\circ) \leq \limsup_{n\to\infty} \varepsilon_n \log P(X_n \in \overline{A}) \leq -\inf_{x\in\overline{A}} I(x).$$

Here A° and $\overline{A}$, respectively, denote the interior and the closure of A, respectively.
The best understood example of an LDP goes back already to Cramér [Cr37]. He showed that the sequence $X_n = \frac{1}{n}\sum_{i=1}^n Y_i$ for an i.i.d. sequence Y_i with $EY_i = 0$ and $Ee^{tY_1} < \infty$ for all t obeys an LDP with speed n and rate function $I(\cdot)$ which is the Legendre transform of Ee^{tY_1}:

$$I(x) = \sup_{t\in\mathbb{R}}[tx - Ee^{tY_1}].$$

As a matter of fact when "changing the scale $\frac{1}{n}$" in this example to $\frac{1}{\sqrt{n}\psi(n)}$ for an increasing function $\psi(n)$ with $\psi(n) \to \infty$ but $\psi(n)/\sqrt{n} \to 0$ well still have an LDP but this time with speed $\psi^2(n)$ and rate function $t^2/2\sigma^2$ (where $\sigma^2 = EY_1^2$). As this reflects both, the closeness to the large deviations as well as the closeness to the Central Limit Theorem (CLT), the latter LDP is often referred to as moderate deviation principle (MDP).
Observe also that in the MDP above the rate functions is independent of the speed, i.e. the rate function is the same of the whole range for possible ψ functions, while it is different for $\psi = \sqrt{n}$. Therefore an LDP with rate function that does not depend on the speed is commonly called an MDP (not only for sequences of i.i.d. variables). For a general and very readable survey over the huge field of large deviation theory we refer the reader to the book of Dembo and Zeitouni [DZ98].

The following assumption is closely related to the existence of an MDP.

Assumption 2.1. *Henceforth we shall assume the following for the "four-point-correlation"- functions:*

$$X_N^{\mu,i} := \sum_{j=1}^{N} \sum_{1=\nu\neq\mu}^{M(N)} \xi_i^\mu \xi_i^\nu \xi_j^\mu \xi_j^\nu - m_{\mu,i}^N \tag{7}$$

($1 \leq i \leq N$ and $1 \leq \mu \leq N$ fixed), where

$$m_{\mu,i}^N := \mathbb{E}\sum_{j=1}^{N} \sum_{1=\nu\neq\mu}^{M(N)} \xi_i^\mu \xi_i^\nu \xi_j^\mu \xi_j^\nu.$$

Consider

$$\Lambda_N^{\mu,i}(t) := \frac{1}{N(M(N)-1)} \log \int_{\mathbb{R}} \exp(-tX_N^{\mu,i}) d\mathbb{P} \tag{8}$$

and assume that there is a function $\Lambda^{\mu,i} : \mathbb{R} \to (-\infty, \infty]$ which is finite in a neighbourhood of zero such that

$$\Lambda_N^{\mu,i}(t) \to \Lambda^{\mu,i}(t) \qquad \forall t \in \mathbb{R}. \tag{9}$$

Moreover assume that $\Lambda_N^{\mu,i}(t)$ *and* $\Lambda^{\mu,i}(t)$ *are twice differentiable and that*

$$\left(\Lambda_N^{\mu,i}\right)''(t) \to \left(\Lambda^{\mu,i}\right)''(t) \qquad \forall t \in [0,\delta] \tag{10}$$

and that this convergence is uniform in t *for some* $\delta > 0$ *and uniform in* i, μ. *Here we define*

$$\left(\Lambda^{\mu,i}\right)''(0) := \left(\Lambda^{\mu,i}\right)''_+(0) := \lim_{t\to 0^+} \frac{1}{t}\left(\left(\Lambda^{\mu,i}\right)'(t) - \left(\Lambda^{\mu,i}\right)'_+(0)\right).$$

Additionally, suppose that

$$\frac{\log N}{N}|m_{\mu,i}^N - m_{\mu,i}| \to 0. \tag{11}$$

Finally assume that

$$\inf_{i,\mu} m_{\mu,i} \geq 0$$

for all N, *together with*

$$\sup_{i,\mu}\left(\Lambda^{\mu,i}\right)''(0) \leq V < \infty$$

for all N *and some* V.

Remark 2.2. *When analyzing the proof of Theorem 2.3 below we find the following:*

a) *If indeed,* $\inf_{i,\mu} m_{\mu,i} > 0$ *condition (11) is obsolete. Moreover we also might find a result for* $\inf_{i,\mu} m_{\mu,i} < 0$ *if it is not too big in absolute value. As it is quite hard to think of a situation where this might occur, we didn't qnatify this statement.*

b) *Actually also the uniformity requirement in (10) is a bit too strong. Indeed much less is required, e.g. that there is a seuqnce of numbers* $\Delta(N,i,\mu)$ *such that*

$$\left|\left(\Lambda_N^{\mu,i}\right)''(t) - \left(\Lambda^{\mu,i}\right)''(t)\right| \leq \Delta(N,i,\mu) \to 0$$

uniformly in some interval $t \in [0,\delta)$ *and such that*

$$N^{-1-\epsilon}\sum_{i=1}^{N}\Delta(N,i,\mu) \to 0$$

for all $\epsilon > 0$ *uniformly in* μ *or*

$$N^{-2-\epsilon}\sum_{\mu=1}^{M}\sum_{i=1}^{N}\Delta(N,i,\mu) \to 0$$

for all $\epsilon > 0$. *This condition will be easier to check in some examples.*

With these definitions our result concerning the storage capacity for correlated patterns reads as follows:

Theorem 2.3. *Assume the random matrix* ξ *fulfils Assumption 2.1 and suppose that* $M(N) = \frac{N}{\gamma \log N}$.
Then there exist positive numbers $c_1 > c_2 > c_3 > 0$ *such that the following assertions hold true:*

1. *If* $\gamma > c_1$

$$P(\liminf_{N\to\infty}(\cap_{\mu=1}^{M(N)} T\xi^\mu = \xi^\mu)) = 1$$

i.e. the patterns are almost surely stable.

2. *If* $\gamma > c_2$

$$P((\cap_{\mu=1}^{M(N)} T\xi^\mu = \xi^\mu)) = 1 - R_N$$

with $\lim_{N\to\infty} R_N = 0$, *i.e. all the patterns are stable with probability converging to one.*

3. *If* $\gamma > c_3$ *for every fixed* $\mu = 1, \dots, M(N)$

$$P(T\xi^\mu = \xi^\mu) = 1 - R_N$$

with $\lim_{N\to\infty} R_N = 0$, *i.e. every fixed pattern is stable with probability converging to one.*

Remark 2.4. *The above theorem basically states that the storage capacity of Hopfield models with correlated patterns fulfilling Assumption 2.1 qualitatively behaves like that of Hopfield models where the patterns are chosen i.i.d. Although Assumption 2.1 seems a bit technical, the proof of Theorem 2.3 to come in the next section will show that a condition similar to Assumption 2.1 is actually needed. Moreover we will see that it is satisfied in a variety of important examples.*

3. Proof of the Theorem and Examples

In this section we will give the proof of Theorem 2.3 and some examples where Assumption 2.1 is fulfilled.
We will substantially make use of that Assumption 2.1 is indeed closely related to an MDP. This observation is due to Wu [Wu95] in a more general setting. He also gives additional conditions that imply a CLT.

Proof of Theorem 2.3. Observe that — according to the definition of the dynamics T and the Hopfield model — for any $1 \le \nu \le M$ the pattern ξ^ν is stable if and only if

$$\xi_i^\nu = sgn(\sum_{j=1}^{N} \xi_j^\nu J_{ij}) = sgn(\sum_{j=1}^{N} \sum_{\mu=1}^{M(N)} \xi_j^\nu \xi_i^\mu \xi_j^\mu)$$

for all $i = 1, \dots, N$. Therefore ξ^ν is stable if and only if

$$\sum_{j=1}^{N} \sum_{\mu=1}^{M(N)} \xi_i^\nu \xi_j^\nu \xi_i^\mu \xi_j^\mu \ge 0$$

for all $i = 1, \dots, N$. Now the sum on the right hand side contains one deterministic positive summand (for $\mu = \nu$), such that the sum actually has a tendency to stay positive. As this deterministic part has size N the pattern ξ^ν is stable if and only if

$$\sum_{j=1}^{N} \sum_{1=\mu\neq\nu}^{M} \xi_i^\nu \xi_j^\nu \xi_i^\mu \xi_j^\mu \ge -N$$

for all $i = 1, \dots, N$.

Hence for all $t \geq 0$

$$\begin{aligned} & P(\xi^\nu \text{ is not stable }) \\ \leq & \sum_{i=1}^{N} P\Big(\sum_{j=1}^{N}\sum_{1=\mu\neq\nu}^{M} \xi_i^\nu \xi_j^\nu \xi_i^\mu \xi_j^\mu \leq -N\Big) \\ \leq & \sum_{i=1}^{N} e^{-tN}\mathbb{E}\left(\exp\left(-t\sum_{j=1}^{N}\sum_{1=\mu\neq\nu}^{M} \xi_i^\nu \xi_j^\nu \xi_i^\mu \xi_j^\mu\right)\right) \end{aligned} \tag{12}$$

where we have applied the exponential Chebyshev–Markov inequality. Centering the sum in the exponential on the right gives

$$\begin{aligned} P(\xi^\nu \text{ is not stable }) \leq & \sum_{i=1}^{N} e^{-t(N+m_{\nu,i}^N)}\mathbb{E}\left(\exp\left(-t(\sum_{j=1}^{N}\sum_{1=\mu\neq\nu}^{M} \xi_i^\nu \xi_j^\nu \xi_i^\mu \xi_j^\mu - m_{\nu,i}^N)\right)\right) \\ = & \sum_{i=1}^{N} e^{-t(N+m_{\nu,i}^N)}\mathbb{E}\big(\exp(-tX_N^{\nu,i})\big) \end{aligned} \tag{13}$$

In order to estimate $\mathbb{E}\big(\exp(-tX_N^{\nu,i})\big)$ we will now make use of Assumption 2.1. Indeed, as $EX_N^{\nu,i} = 0$ we have by the twice differentiability of $\Lambda_N^{\nu,i}$ together with

$$\Lambda_N^{\nu,i}(0)' = \mathbb{E}X_N^{\nu,i} = 0$$

that

$$\Lambda_N^{\nu,i}(s) = \int_0^s \big(\Lambda_N^{\nu,i}\big)''(x)(s-x)dx$$

for all $0 \leq s < \delta$. Similar for $\Lambda^{\nu,i}(s)$:

$$\Lambda^{\nu,i}(s) = \big(\Lambda^{\nu,i}\big)'_+(0) + \int_0^s \big(\Lambda^{\nu,i}\big)''(x)(s-x)dx$$

for all $0 \leq s < \delta$. Now as a consequence of Assumption 2.1

$$\big(\Lambda^{\nu,i}\big)'(x) - \big(\Lambda^{\nu,i}\big)'_+(0) = \int_0^x \big(\Lambda^{\nu,i}\big)''(s)ds = \lim_{N\to\infty}\int_0^x \big(\Lambda_N^{\nu,i}\big)''(s)ds = \big(\Lambda^{\nu,i}\big)'(x).$$

Thus $\big(\Lambda^{\nu,i}\big)'_+(0) = 0$ and therefore

$$\begin{aligned} & \sup_{0\leq x<\delta}\frac{1}{x^2}\big|\Lambda_N^{\nu,i}(x) - \Lambda^{\nu,i}(x)\big| \\ \leq & \sup_{0\leq x<\delta}\frac{1}{x^2}\int_0^x \Big|\big(\Lambda_N^{\nu,i}\big)''(y) - \big(\Lambda^{\nu,i}\big)''(y)\Big|(x-y)dy \\ \leq & \sup_{0\leq x<\delta}\frac{1}{2}\Big|\big(\Lambda_N^{\nu,i}\big)''(x) - \big(\Lambda^{\nu,i}\big)''(x)\Big|, \end{aligned} \tag{14}$$

where the expression on the right hand side converges to zero. Note that under the uniformity assumption of Assumption 2.1 the above convergence is also uniform in i and ν.

This estimate will be crucial for the rest of the proof. We use it with $x = t/a(N)$ for some fixed t and for some sequence $a(N)$ with $a(N) \to \infty$ and $a(N)/\sqrt{NM} \to 0$.

Then the above estimate implies by Taylor expansion of $\Lambda^{\nu,i}$ that

$$\begin{aligned}&\lim_{N\to\infty}\sup_{i,\nu}\left|a^2(N)\Lambda_N^{\nu,i}(t/a(N))-\frac{1}{2}\left(\Lambda^{\nu,i}\right)''(0)t^2\right|\\ \leq\ &\lim_{N\to\infty}\left|a^2(N)\Lambda^{\nu,i}(t/a(N))-\frac{1}{2}\left(\Lambda^{\nu,i}\right)''(0)t^2\right|=0\end{aligned}$$

implying that

$$\mathbb{E}\exp\left(\frac{-t}{a(N)}X_N^{\nu,i}\right)=\exp\left(\frac{1}{2}t^2\frac{NM(N)}{a^2(N)}\left(\Lambda^{\nu,i}\right)''(0)\right)(1+o(1))$$

where due to the uniformity assumption in Assumption 2.1 the $o(1)$ is uniform in i and ν.

Plugging this into (13) gives by replacing t by $\frac{t}{a(N)}$

$$\begin{aligned}&P(\xi^\nu \text{ is not stable })\\ \leq\ &\sum_{i=1}^N\exp\left(-\frac{t}{a(N)}(N+m_{\nu,i}^N)\right)\exp\left(\frac{1}{2}t^2\frac{NM(N)}{a^2(N)}\left(\Lambda^{\nu,i}\right)''(0)\right)(1+o(1))\\ \leq\ &\sum_{i=1}^N\exp\left(-\frac{t}{a(N)}(N+m_{\nu,i})+\frac{t}{a(N)}(m_{\nu,i}-m_{\nu,i}^N)\right)\exp\left(\frac{1}{2}t^2\frac{NM(N)}{a^2(N)}V\right)(1+o(1))\end{aligned}$$

where again we have made use of Assumption 2.1 and the $o(1)$ term is uniform in i and ν.

Now choosing the essentially optimal $t=1/\gamma$ and $a(N)=M(N)$ we first see that with the ansatz (15) below is admissible in the sense that $a(N)\to 0$ but $a(N)/\sqrt{NM}\to 0$ and moreover that

$$P(\xi^\nu \text{ is not stable })\leq\sum_{i=1}^N\exp\left(-\frac{1}{2}\frac{N}{MV}\right)\exp\left(\frac{1}{VM}(m_{\nu,i}-m_{\nu,i}^N)\right)(1+o(1)).$$

The ansatz

$$M(N)=\frac{N}{\gamma\log N}\tag{15}$$

for some positive constant γ yields

$$\begin{aligned}&P(\xi^\nu \text{ is not stable })\\ \leq\ &\sum_{i=1}^N\exp\left(-\frac{1}{2}\frac{\gamma\log N}{V}\right)\exp\left(\gamma(\log N/N)(m_{\nu,i}-m_{\nu,i}^N)\right)(1+o(1))\\ \leq\ &N\exp\left(-\frac{1}{2}\frac{\gamma\log N}{V}\right)(1+o(1))\end{aligned}\tag{16}$$

The choice $\gamma>2V$ gives

$$P(\xi^\nu \text{ is not stable })\leq N^{1-\frac{\gamma}{2V}}\to 0,$$

which is part three of the theorem.

For parts one and two observe that by (16) and the ansatz (15)

$$P(\exists\nu:\xi^\nu \text{ is not stable })\leq\frac{N^2}{\gamma\log N}\exp\left(-\frac{1}{2}\frac{\gamma\log N}{V}\right)(1+o(1)).$$

So choosing $\gamma \geq 4V$ yields

$$P(\exists \nu : \xi^\nu \text{ is not stable }) \to 0$$

which is part two of the theorem.
Finally the choice of $\gamma > 6V$ gives that

$$P(\exists \nu : \xi^\nu \text{ is not stable }) \leq N^{-\kappa}$$

for some $\kappa > 1$. As $N^{-\kappa}$ is summable (over N) the Borel-Cantelli Lemma eventually proves part one of the theorem.

□

Remark 3.1. a) *Observe that the theorem gives precise bounds on the constants c_1, c_2 and c_3 occurring in Theorem 2.3. Note also that if $\inf_{\nu,i} m_{\nu,i}$ is strictly greater than zero these bounds can be improved.*
b) *Check that Remark 2.2., in particular part b) of it, apply.*

To see whether the conditions of Assumption 2.1 are ever fulfilled let us start with the very basic example of i.i.d. patterns where all the calculations can be done by hand.

Example 3.2. *Independent patterns*
Assume that the matrix ξ consists of i.i.d. entries obeying (2). Then by independence for all ν, i and N

$$\Lambda_N^{\mu,i}(t) = \frac{1}{N(M(N)-1)} \log \mathbb{E} \exp\Big(-t \sum_{j=1}^{N} \sum_{1 \leq \mu \neq \nu}^{M} \xi_i^\mu \xi_i^\nu \xi_j^\mu \xi_j^\nu\Big) = \log\cosh(t).$$

Hence $\Lambda_N^{\mu,i}$ converges (and indeed ist identical) to $g(t) := \log\cosh(t)$ as well as all its derivatives converge to the corresponding derivatives of $\log\cosh(t)$. Taking moreover into account that

$$\mathbb{E} \sum_{j=1}^{N} \sum_{1 \leq \mu \neq \nu}^{M} \xi_i^\mu \xi_i^\nu \xi_j^\mu \xi_j^\nu = 0$$

for all ν, i and N and that

$$\frac{d^2}{dt^2} \log\cosh(t)\Big|_{t=0} = 1$$

Theorem 2.3 not only yields that the Hopfield model can store $\frac{N}{\gamma \log N}$ i.i.d. patterns but also gives the constants $c_1 = 6$, $c_2 = 4$, and $c_3 = 2$, respectively, for the cases one, two, and three, in Theorem 2.3, respectively. These results agree with those obtained by McEliece et al. [MPRV87] *and* [P96] *and have basically been shown to be optimal by Bovier* [Bo99].

For the other examples we heavily exploit the following theorem from [Wu95, Theorem1.4].

Theorem 3.3. *If for some $\delta > 0$ the functions $\left(\Lambda_N^{\mu,i}\right)'$ are all concave (or convex) on $[0, \delta)$ and if $\Lambda_N^{\mu,i}$ is twice continuously differentiable on $[0, \delta)$ and if moreover*

$$\left(\Lambda_N^{\mu,i}\right)''(0) \to \left(\Lambda_N^{\mu,i}\right)''(0) \tag{17}$$

uniformly in μ and i as N tends to infinity, then (8) and (9) are satisfied.

Remark 2.2 b) applies correspondingly.
The reason why this theorem is extremely helpful is (as was also already pointed out by Wu) that the concavity can be achieved by a GHS inequality while (17) usually follows from a FKG-type inequality (in the examples below we will follow [Wu95, Proof of Lemma 3.2], the interested reader will find the basic ideas there).

If we now introduce correlations among the patterns we might – from a mathematical point of view – of course, correlate each ξ_i^μ to each ξ_j^ν in some strange fashion and look whether Assumption 2.1 is still fulfilled. Taking into account, however, that the vector $(\xi_i^\mu)_{i=1,\dots,N}$ is supposed to describe an image or at least some information to be stored (which is different for different ν) there are two reasonable ways to correlate the patterns. The correlation of the ξ_i^μ in μ is called *sequential* correlation and may be reasonable e.g. when storing films, while the correlation in i is referred to as *spatial* correlation, which may be a reasonable model when storing images.
The following two example are basically covered by [Lö99a, Theorems 2.1,2.2]

Example 3.4. *Spatial Markov Chains*
Consider sequences of spatially correlated patterns $(\xi_i^\mu)_{i=1,\dots,N,\mu=1,\dots,M(N)}$ where the correlation stems form a one dimensional Markov chain. More precisely we will assume that the random variables $(\xi_i^\mu)_{i\in\mathbb{N},\mu\in\mathbb{N}}$ are independent for different μ and for fixed μ form a Markov chain in i with initial distribution

$$P(\xi_1^\mu = x_1^\mu,\ \mu = 1,\dots,M) = 2^{-M} \quad \text{for all } x_1^\mu \in \{-1,1\} \tag{18}$$

and transition probabilities

$$\begin{aligned} & P(\xi_i^\mu = x_i^\mu | \xi_j^\nu = x_j^\nu, j = 1,\dots,i-1, \nu = 1,\dots,M) \\ = \ & P(\xi_i^\mu = x_i^\mu | \xi_{i-1}^\mu = x_{i-1}^\mu) = Q(x_{i-1}^\mu, x_i^\mu). \end{aligned} \tag{19}$$

Here Q denotes a symmetric 2×2 matrix with entries

$$Q = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}$$

where $0 < p < 1$ (note that $p = \frac{1}{2}$ is the case of independent patterns). Because of the independence and identical distribution of the patterns ξ^μ

$$\begin{aligned} \Lambda_N^{\nu,i}(t) &= \frac{1}{N(M(N)-1)} \log \mathbb{E} \exp\Big(-t \sum_{j=1}^N \sum_{1\le\mu\neq\nu}^M \xi_i^\mu \xi_i^\nu \xi_j^\mu \xi_j^\nu\Big) \\ &= \frac{1}{N} \log \mathbb{E} \exp\Big(-t \sum_{j=1}^N \xi_i^1 \xi_j^1 \xi_i^2 \xi_j^2\Big). \end{aligned}$$

As the whole situation is completely symmetric under the flipping ξ_i^2 into the direction of ξ_i^1 we may assume that $\xi_i^1 \xi_i^2 = 1$. Hence the computation above boils down to the computation of

$$\Lambda_N^{\nu,i}(t) = \frac{1}{N} \log \mathbb{E} \exp\Big(-t \sum_{j=1}^N \xi_j^1 \xi_j^2\Big)$$

(which does not any longer depend on ν). Observe that $(Y_j) := (\xi_j^1 \xi_j^2)$ is a Markov chain in j with transition matrix

$$\overline{Q} = \begin{pmatrix} q & 1-q \\ 1-q & q \end{pmatrix}$$

and with the fixed "starting" point $Y_i = 1$. As $q := p^2 + (1-p)^2 \geq 1/2$ and $Y_i = 1$ also Y_j has a tendency to be $+1$ rather than -1 and therefore (as is easily checked)

$$\mathbb{E}\sum_{j=1}^{N} \xi_j^1 \xi_j^2 \geq 0$$

which has been part of Assumption 2.1. Now as its is not too difficult (basically using Perron-Froebenius theory) to show that

$$\frac{1}{N}\log \mathbb{E}\exp\left(-t\sum_{j=1}^{N}\xi_j^1\xi_j^2\right) \to \log\lambda_1(t) \tag{20}$$

where $\lambda_1(t)$ is the largest eigenvalue of

$$\overline{Q}(t) = \begin{pmatrix} qe^{-t} & (1-q)e^{-t} \\ (1-q)e^{t} & qe^{t} \end{pmatrix}$$

we could try to verify Assumption 2.1 directly (the calculations could partially follow those in [Lö99a]*). However, note that the limit $\lambda(\cdot)$ in (20) always exists due to Varadhan's Lemma, because the chain Y_j obeys an LDP. On the other hand it follows easily from the presentation of a Markov chain as a random field (see e.g.* [Ge88, Chapter3]*) and the GHS inequality for Bernoulli random fields (as derived in* [GHS70], [EMN76], [EN78]*) that the chain Y_i fulfils the GHS inequalities. Following* [Wu95, Lemma3.2] *and applying* [E85, pp.167] *these imply the concavity of the Λ's. For the same reasons it also fulfils the FKG inequality (*[FKG71]*, e.g.). They imply that $\Lambda''(0) = \sum_{j=0}^{\infty}\mathbb{E}Y_0Y_j$ and that (see* [Wu95, Lemma3.2]*) $\left(\Lambda_N^{\mu,i}\right)''(0) = \sum_{j=1}^{N}\mathbb{E}Y_iY_j$ converges to $\Lambda''(0)$. As this convergence is uniform for all points which are outside a strip (of width growing slowly with n) along the boundary (which is negligible compared to the other points) indeed the assumptions of Theorem 3.3 are fulfilled. Hence following Theorem 2.3 we obtain a storage capacity of $N/\gamma\log N$ patterns. However, to obtain bounds for γ which are of the right order, more subtleties such as exploiting the Martingale structure of a Markov chain, are needed.*

Example 3.5. *Sequential Markov Chains*
One could as well consider patterns that are only sequentially correlated by a Markov chain. So more precisely, assume that the random variables $(\xi_i^\mu)_{i\in\mathbb{N},\mu\in\mathbb{N}}$ are independent for different i and for fixed i form a Markov chain in μ with initial distribution

$$P(\xi_i^1 = x_i^1,\ i = 1,\dots,N) = 2^{-N} \quad \text{for all } x_i^1 \in \{-1,1\} \tag{21}$$

and transition probabilities

$$\begin{aligned} & P(\xi_i^\mu = x_i^\mu | \xi_j^\nu = x_j^\nu, j = 1,\dots,N, \nu = 1,\dots,\mu-1) \\ = \ & P(\xi_i^\mu = x_i^\mu | \xi_i^{\mu-1} = x_i^{\mu-1}) = Q(x_i^{\mu-1}, x_i^\mu) \end{aligned} \tag{22}$$

where Q is as above. Note that, although the Hamiltonian of the model works differently in the lower indices i and the upper indices μ, our definition of storage capacity, however does not see a big difference between the model in this example and the model in the example above. So, indeed the case of the sequentially correlated patterns can be treated just like the case in Example 3.4 by interchanging the rôle of i and μ. It should be mentioned that a minor difference could be observed when taking into account also the expectations of the Y_i, as has been done in [Lö99a].

The following example we were not able to treat in [Lö99a]. To understand why we consider it important, recall that the usual setup for an image is that of a random field. In the next example we will try to store M independent images in the Hopfield model. Avoiding the possibly complicated notation of random fields we restrict ourselves to one of the simplest cases: the ferromagnetic Ising model. Intuitively speaking we will try to store M independent black and white pictures in the Hopfield model where a black pixel is more likely to sit next to a black pixel than a white one (and vice versa).

Example 3.6. *Independent Ising models*
Consider patterns $(\xi_i^\mu)_{i\in\mathbb{N},\mu\in\mathbb{N}}$ that are independent of different μ and for fixed μ are distributed according to the distribution of a d-dimensional Ising model at inverse temperature β which is not the critical temperature, with zero external field and free boundary conditions (otherwise the patterns might have a bias). That means (for simplicity) we assume that N is a d'th power and that for each fixed μ the spins $(\xi_i^\mu)_i$ are distributed according to

$$\mathbb{P}(\xi^\mu) = \frac{\exp\big(\beta \sum_{<i,j>} \xi_i^\mu \xi_j^\mu\big)}{Z_N(\beta)}. \tag{23}$$

Here

$$Z_N(\beta) = \sum_\sigma \exp\big(\beta \sum_{<i,j>} \sigma_i \sigma_j\big)$$

is the so called partition function of the Ising model, $\beta \in [0,\infty)$ is its inverse temperature, and the summation $\sum_{<i,j>}$ is taken over all neighbouring pairs of indices i,j in $\mathbb{Z}^d$ that sit inside the box of side length $N^{1/d}$ centred at the origin. Again due to the independence and identical distribution of the patterns ξ^μ

$$\begin{aligned}\Lambda_N^{\nu,i}(t) &= \frac{1}{N(M(N)-1)} \log \mathbb{E}\exp\big(-t \sum_{j=1}^N \sum_{1\le\mu\neq\nu}^M \xi_i^\mu \xi_i^\nu \xi_j^\mu \xi_j^\nu\big) \\ &= \frac{1}{N} \log \mathbb{E}\exp\big(-t \sum_{j=1}^N \xi_i^1 \xi_j^1 \xi_i^2 \xi_j^2\big).\end{aligned}$$

Again we might flip the images until we may assume that $\xi_i^1 \xi_i^2 = 1$. Hence we are again left to treat

$$\Lambda_N^i(t) = \frac{1}{N} \log \mathbb{E}\exp\big(-t \sum_{j=1}^N \xi_j^1 \xi_j^2\big).$$

(where the upper index ν has been skipped since obviously this quantity does not depend on ν) under the restriction that $\xi_i^1 = \xi_i^2 = 1$. The existence of the limit as N approaches infinity of the Λ_N^i again can be proved by large deviation arguments (using e.g. [Co86], or [Ol88]) as the vectors $((\xi_j^1, \xi_j^2))_j$ may be regarded as a two dimensional random field with nearest neighbour interaction. On the other hand considering the partition function of this model with external field $h = (h_j)_{j=1,\dots N}$

$$\tilde{Z}_N(\beta, h) := \sum_{\xi^1,\xi^2} \exp\big(\beta \sum_{<k,j>} \xi_k^1 \xi_j^1 + \beta \sum_{<k,j>} \xi_k^2 \xi_j^2 + \sum_{j=1}^N h_j(\xi_j^1 + \xi_j^2)\big)$$

we see that clearly

$$\tilde{Z}_N(\beta, h) = (Z_N(\beta, h))^2$$

where

$$Z_N(\beta, h) = \sum_{\sigma} \exp\Big(\beta \sum_{<k,j>} \sigma_k \sigma_j + \sum_{j=1}^{N} h_j \sigma_j\Big).$$

Therefore $\frac{1}{N} \log \tilde{Z}_N(\beta)$ *satisfies*

$$\frac{\partial^3}{\partial h_j \partial h_k \partial h_l} \frac{1}{N} \log \tilde{Z}_N(\beta) \leq 0$$

for each choice of the indices j, k, l *as also* $\frac{1}{N} \log Z_N(\beta)$ *satisfies these so-called GHS inequalities. Moreover as the system has component-wise ferromagnetic nearest neighbour interaction the FKG-inequalities follow along the lines of* [FKG71]. *Finally by the famous Onsager formulas the Ising model away from the critical temperature has exponentially decaying and hence summable correlations which readily implies the summability of the correlations in our model. So by following the arguments in Example 3.4, i.e. using again* [Wu95, Lemma3.2], *which is based on the GHS and the FKG inequality, together with* [E85, pp.167] *we see that the conditions of Theorem 3.3 are satisfied. Thus Theorem 2.3 is true and also in this case the Hopfield model has a storage capacity of* $N/\gamma \log N$ *patterns for some constant* γ.

Of course, we might also try to store more complicated random fields in the Hopfield model. As can be seen in the above examples this will always work as we can make sure that the conditions of Theorem 3.3 are satisfied, e.g. by employing [Wu95, Lemma3.2]. We are also firmly convinced, but have not checked the details, that along the same lines we could show that the Hopfield model can store $N/\gamma \log N$ independent Curie-Weiss models.

References

[AGS87] D.J. Amit, G. Gutfreund, H. Sompolinsky; Statistical mechanics of neural networks near saturation; Ann. Phys. 173, 30-67 (1987)

[Bo99] A. Bovier; Sharp upper bounds for perfect retrieval in the Hopfield model; Preprint, WIAS Berlin to appear in J. Appl. Prob. 36 (1999)

[BG92] A. Bovier, V. Gayrard; Rigorous bounds on the storage capacity of the dilute Hopfield model; J. Stat. Phys. 69, 597-627 (1992)

[BG96a] A. Bovier, V. Gayrard; An almost sure large deviation principle for the Hopfield model, Ann. Probab. 24, 1444–1475 (1996)

[BG97a] A. Bovier, V. Gayrard; An almost sure Central Limit Theorem for the Hopfield Model Markov Processes Relat. Fields 3, 151-173 (1997)

[BG97b] A. Bovier, V. Gayrard; The retrieval phase of the Hopfield model: A rigorous analysis of the overlap distribution, Probab. Theory Related Fields 107, 61–98 (1997)

[BG98] A.Bovier, V.Gayrard; Hopfield models as a generalized mean field model, Preprint, in "Mathematics of spin glasses and neural networks", A. Bovier, P.Picco (eds.), "Progress in Probability", Birkhäuser (1998)

[BGP94] A.Bovier, V.Gayrard, P.Picco; Gibbs states for the Hopfield model in the regime of perfect memory; Prob. Th. Rel. Fields 100, 329-363 (1994)

[BGP95] A.Bovier, V.Gayrard, P.Picco; Large deviation principles for the Hopfield model and the Kac-Hopfield model; Prob. Th. Rel. Fields 101, 511-546 (1995)

[BP98] Mathematical Aspects of Spin Glasses and Neural Networks, A. Bovier, P.Picco (eds.), "Progress in Probability", Birkhäuser, Boston (1998)

[Bu94] D. Burshtein; Nondirect convergence radius and number of iterations of the Hopfield associative memory; IEEE Trans. Inf. Th. 40, 838-847, (1994)

[Co86] F. Comets; Grandes deéviations pour chanps Gibbs sur $\mathbb{Z}^d$, C.R.A.S, Serie I Math. 303, 511-517 (1986)

[Cr37] Cramér, H.: On a new limit theorem in the teory of probability; Colloquium of the theory of probability, Hermann Paris (1937)

[DZ98] A. Dembo, O. Zeitouni; Large deviations techniques and applications, 2nd edition, Applications of Mathematics 38, Springer Verlag, New York (1998)

[E85] R.S. Ellis;Entropy, large deviations and statistical mechanics; Grundlehren der Math. Wissenschaften 271, Springer, Berlin (1985)

[EMN76] R.S. Ellis, J. L. Monroe, C. Newman; The GHS and other correlation inequalities for even ferromagnets, Comm. Math. Phys 46, 167-182 (1976)

[EN78] R.S. Ellis, C. Newman; Necessary and sufficient conditions for the GHS inequality with applications to analysis and probability Trans. Am. Math. Soc. 237, (1978)

[FP77] L.A. Pastur, A.L. Figotin; Exactly soluble model of a spin-glas; Sov. J. of Low Temprature Phys. 3(6), 378-383 (1977)

[FKG71] C.M. Fortuin, J. Ginibre, P.W. Kasteleyn; Correlation inequlities for partially ordered sets, Comm. Math. Phys. 22, 89-103 (1971)

[Ge88] H.O. Georgii; Gibbs measures and phase transition, de Gruyter Studies in Mathematics 9, Walter de Gruyter, New York (1988)

[GHS70] R.B. Griffith, C.A. Hurst, S.Shermann; Concavity of magnetization of an Ising ferromagnet in a positive magnetic field, J. Math. Phys. 11, 790–795 (1970)

[Ho82] J.J. Hopfield; Neural networks and physical systems with emergent collective computational abilities; Proc. nat. Acad. Sci. USA 79, 2554-2558 (1982)

[Lou94] D. Loukianova; Capacité de mémoire dans le modèle de Hopfield; C.R. Acad. Sci. Paris 318, 157-160 (1994)

[Lö97] M. Löwe; On the storage capacity of the Hopfield model; "Mathematics of spin glasses and neural networks", A. Bovier, P.Picco (eds.), "Progress in Probability", Birkhäuser (1997)

[Lö99a] M. Löwe; On the storage capacity of Hopfield models with weakly correlated patterns, Preprint, Universität Bielefeld, to appear in: Annals of Appl. Probability

[Lö99b] M. Löwe; On the storage capacity of Hopfield models with biased patterns, Preprint, Universität Bielefeld, to appear in: IEEE. Inf. Theory

[MPRV87] R. McEliece, E. Posner, E. Rodemich. S. Venkatesh; The capacity of the Hopfield associative memory; IEEE Inf. Th. 33, 461-482 (1987)

[MPV87] M. Mezard, G. Parisi, M.A. Virasoro; Spin Glass Theory and beyond, World Scientific Lecture Notes in Physics 9, World Scientific Publishing Co., Teaneck, NJ (1987)

[N88] C. Newman; Memory capacity in neural networks; Neural Networks 1, 223-238 (1988)

[Ol88] S. Olla; Large deviations for Gibbs random fields; Prob. Th. rel. Fields 77, 343-357 (1988)

[P96] D. Petritis; Thermodynamic formalism of neural computing; Nonlinear Phenomena of Complex Systems 2, 86–146, Kluwer Acad. Publ., Dordrecht (1986)

[T95] M. Talagrand; Résultats rigoureux pour le modèle de Hopfield, C.R. Acad. Sci. 321, I, 309-312,(1995)

[T98] M. Talagrand; Rigorous Results of the Hopfield Model with Many Patterns, Prob. Theory Rel. Fields 110, 177-276 (1998)

[Wu95] Wu L.M.; Moderate Deviations for Dependent Random Variables Related to CLT, Annals of Prob. 23, 420-445 (1995)

(Matthias Löwe) EURANDOM, P.O. BOX 513, 5600 MB EINDHOVEN, THE NETHERLANDS
E-mail address, Matthias Löwe: lowe@eurandom.tue.nl