On the Storage Capacity of Hopfield Models with Spatially Correlated Patterns

Matthias Löwe
Universität Bielefeld
Fakultät für Mathematik
Postfach 100131
33501 Bielefeld
email: loewe@mathematik.uni-bielefeld.de

Keywords: Hopfield model, neural networks, storage capacity, Markov chains, large deviations

AMS Subject Classification: 82C32, 82B44, 60K35

Abstract

We analyze the storage capacity of the Hopfield model with spatially correlated patterns ξ_i^ν (i.e. the patterns ξ^ν themselves are independent but have correlated components ξ_i^ν). We show that the standard Hopfield model of neural networks with N neurons can store $\frac{N}{\gamma \log N}$ or αN semantically correlated patterns (depending on which notion of storage is used), provided that the correlation comes from a homogeneous Markov chain. The quantity γ increases with an increasing correlation of the patterns, while α decreases with large correlations. This shows that the standard Hopfield can store an increasing number of spatially correlated patterns. The proofs use adaption of the methods developed for the case of i.i.d. or semantically correlated patterns to the situation of spatial correlation together with an estimate of the determinant of a symmetric matrix with trace zero.

1 Introduction

The study of neural networks has been a major topic of research in mathematics, physics, and computer science during the past 15 years.

Basically a model of a neural network consists of a labeled and possibly oriented graph $G = (V, E)$ together with a set S with $card(S) \geq 2$ to describe the set of neurons (by V), the synapses connecting these neurons (by E), and the activity of each of the neurons (by a variable $\sigma_i \in S$ for $i \in V$). Moreover, the information to be stored is encoded in so-called patterns $\xi^\mu, \mu = 1, \ldots, M(N)$ each of the ξ^μ itself being a sequence of $\xi_i^\mu \in S, i = 1, \ldots, N$. Finally, to make the neural net capable of adapting to different sequences of patterns, we have to introduce a set of variables $J_{ij}, ij \in E$ called the synaptic efficacies and describing the strength of interaction between the neurons at sites i and j. It is commonly assumed that the variable J_{ij} is measurable with respect to the set $\{\xi_i^\mu, \xi_j^\mu, \mu = 1, \ldots, M(N)\}$ (so-called *locality* of the weights J_{ij}). The basic idea now behind this setup is to define a Hamiltonian $H_N(\sigma)$ on S^N, such that the Monte-Carlo dynamics governed by this Hamiltonian eventually converges to one of $M(N)$ states close to the original patterns.

One of the classical and best understood examples of a neural network is the so-called Hopfield model [Ho82]. Although originally introduced by Pastur and Figotin [FP77] as a simplified model of a spin–glass, this model achieved much of its interest by its reinterpretation as an auto-associative memory by Hopfield and may therefore by right be called the Hopfield model. Here the graph G is the complete graph K_N on N vertices, $S = \{-1, +1\}$, corresponding to that a neuron may be either switched "on" or "off" and the weights J_{ij} are given by "Hebb's learning rule", i.e. the formula

$$J_{ij} := \frac{1}{N} \sum_{\mu=1}^{M(N)} \xi_i^\mu \xi_j^\mu.$$

One of the most important advances due to Hopfield [Ho82] has been to understand that this setup corresponds to a Hamiltonian H_N given by

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

Note that (1) may be rewritten in the very convenient form

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

where $||m_N(\sigma)||_2^2$ is the l_2-norm of the so-called overlap vector

$$m_N(\sigma) := (m_N^\mu(\sigma))_{\mu=1,\ldots,M(N)} = (\frac{1}{N} \sum_{i=1}^{N} \sigma_i \xi_i^\mu)_{\mu=1,\ldots,M(N)}.$$

Note that $m_N(\sigma)$ may be regarded as an index for how much a configuration is correlated to one of the given patterns (large absolute values of a component of $m_N(\sigma)$

corresponding to large correlations). Also observe that (2) makes it plausible that – at least for $M := M(N)$ small enough – the minima of H_N are located close to the patterns ξ^μ (notice that this is trivially fulfilled, if the patterns are orthogonal, i.e. if

$$m_N(\xi^\mu) = \delta_{\mu,\nu}).$$

Indeed, much of the recent work on the Hopfield model can be summarized under the aspect to make precise and mathematically verify this last statement. To this end in most of the papers it has been assumed that the ξ_i^μ are unbiased i.i.d. random variables, i.e. that

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

for all i and μ independently of all the other ξ_j^ν.

Already Hopfield [Ho82] by computer simulations discovered that in this case there is a value $\alpha_c > 0$ such that if $M(N) \leq \alpha_c N$ almost all patterns are memorized, whereas for $M(N) > \alpha_c N$ the Hopfield model tends to "forget" all of the patterns and therefore one may rightfully speak about a storage capacity of the Hopfield model. The numerical value of α_c found by Hopfield was close to 0.14. This finding with a similar value for α_c has been supported by non-rigorous analytical computations (including the notorious replica–trick) by Amit et al. in [AGS85] and [AGS87] (for the current state of the art concerning the replica–method see [KS94]). The first rigorous result concerning the storage capacity of the Hopfield model were by McEliece et al. [MPRV87] showing that if one focuses on exact reproduction of the patterns the memory capacity of the Hopfield model is only in the order of $const.\frac{N}{\log N}$ where the constant ranges between 1/2 and 1/6 and mainly depends on whether one is interested in "almost sure" results or in results holding with "probability converging to one" (see the corresponding results in Section 2, also see the survey paper by Petritis [P95]).

On the other Newman in [N88] showed rigorously that, if small errors are tolerated, the Hopfield model is able to successfully retrieve a number of patterns M proportional to the number of neurons N, and his (lower) bound on the constant α_c was $\alpha_c \geq 0.056$. This bound recently has been improved by Loukianova [Lou94] to $\alpha_c \geq 0.071$ and Talagrand [T95], [T96] to $\alpha_c \geq 0.08$ by a refined use of Newman's technique. But to rigorously verify Hopfield's value for α_c and to give a good picture of what happens above this bound still remains a major open problem and only very few results on this subject are available (see [Lou95] and [T96]).

However, nowadays the Hopfield model in the regime where the patterns are chosen independently and their number is "small" compared to the number of neurons is well understood (also in the case of non-zero temperature). The corresponding analysis has been carried out in a series of papers by Bovier and Gayrard, partially in collaboration with Picco [BGP94], [BGP95a], [BGP95b],[BG95a], [BG95b]. Another milestone certainly has been set in a very recent paper by Talagrand [T96], especially what concerns the high-temperature phase and the validity of the so-called replica–symmetric solution in the Hopfield model. For a comprehensive review over the quick development in this area during the last few years and a particularly nice proof of the validity of the replica-symmetric solution we refer the reader to [BG96]. In this context, of course, a result due to Gentz [Gen95] should be mentioned where a central limit theorem for the overlap parameter in this model is proved.

Now in most realistic situations the patterns are not at all independent (see e.g [Mi88]). Indeed, there are at least two sensible ways to introduce correlations among the patterns. One is to consider spatial correlation, i.e. to choose the patterns correlated in i but independent in μ, which may be interesting when e.g. thinking about the patterns as images to be stored. The other way is to choose sequentially or semantically correlated patterns, which means that the dependency now enters via μ only. This situation might be useful as a very simple model for patterns with some sort of causal relations, as e.g. films (for a Hopfield model with deterministic sequences of patterns also see the work of Sandmeier [S96]).

In both situations the picture is less clear than in the "i.i.d. case". As a matter of fact until recently hardly any rigorous results on the Hopfield model with correlated patterns have been available and while some authors ([Mo93], [TL93]) argued that correlations can only increase the storage capacity of a neural network – which is a reasonable conjecture from the point of view of information content (a sequence of dependent patterns contains less information than a sequence of independent ones) – others claimed a decrease of the storage capacity up to the extreme case that the Hopfield model can not store any unbounded number of correlated patterns (see e.g. [HK91]).

In [Lö96a] we were able to settle the question of an extensive storage capacity of the Hopfield model and to show that the Hopfield model indeed can store $N/(\gamma \log N)$ (for an appropriate constant γ) or even $\alpha_c N$ correlated patterns – depending on which notion of storage is used – provided that the correlation is semantical and comes from a homogeneous Markov chain. The lower bounds on the constants $1/\gamma$ and α_c we were able to get were decreasing with an increasing strength of the correlation (actually [Lö96a] also contains an analysis of a class of generalized Hopfield models showing the effect that neural net should be able to store at least as much correlated patterns as uncorrelated ones).

The situation of spatially correlated patterns might be even more interesting but definitely is also more challenging due to the fact that Hamiltonian is linear in μ but quadratic in i. In this paper we mainly give a description of the memory capacity of the Hopfield model with spatially correlated patterns, again provided that the correlations stem from a homogeneous Markov chain (corresponding to one dimensional images). We show that the Hopfield model can store $N/(\gamma \log N)$ or even $\alpha_c N$ spatially correlated patterns (depending on the notion of storage capacity). Again the lower bounds on the constants $1/\gamma$ and α_c are decreasing with an increasing strength of the correlation and as a matter of fact the γ we get is the same as in the case of a sequence of semantically correlated patterns coming from a Markov chain with the same transition matrix (in Section 3 we will explain why this is not at all an surprise). The methods of proof we use rely on similar ideas as the proofs in [Lö96a], together with a new estimate on the determinant of a quadratic form with trace zero.

This paper has two further sections. Section 2 contains the basic setup including a definition of the two notions of storage capacity we use and a description of the homogeneous Markov chain we have in mind. Moreover we state our main results. Section 3 is devoted to the proofs. Especially, the proof of Theorem 2.3 is a little involved and we prepare it by two auxiliary lemmas which may interesting in their own right.

Acknowledgement: The author thanks the WIAS for its hospitality. He also thanks Holger Knöpfel for discussions on the analytical problems.

2 The Setup and the Main Results

In this section we will mainly state our results on the storage capacity of the Hopfield model with spatially correlated patterns.
Before let us first briefly explain the two different concepts we are dealing with. To this end let us first recall the definition of the Hopfield Hamiltonian with $M := M(N)$ patterns

$$H_N(\sigma) = -\frac{1}{2}\sum_{i,j=1}^{N} J_{ij}\sigma_i\sigma_j \tag{3}$$

where

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

and $\sigma_i \in \{-1, 1\}$.
The idea behind the first notion of storage capacity 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

$$T : \sigma_i \mapsto sgn(\sum_{\substack{j=1\\ j\neq i}}^{N} \sigma_j J_{ij})$$

where sgn is the sign function. Slightly overloading the notation T we will also write $T(\sigma)$ and mean $(T(\sigma_i))_i$. We will call a configuration $\sigma = (\sigma_i)_{i\leq N}$ stable if it is a fixed point of T, i.e.

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

which means that σ is a local minimum of the Hamiltonian. The storage capacity in this concept is defined as 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).

The other approach to storage capacity is due to Amit, Gutfreund and Sompolinsky [AGS85] and has rigorously been analyzed 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

$$\inf_{\sigma\in S_\delta(\xi^\nu)} H_N(\sigma) \geq H_N(\xi^\nu) + \varepsilon N. \tag{4}$$

Here the set $S_\delta(\xi^\nu)$, the infimum is taken over, is the Hamming sphere of radius δN centered in ξ^ν. Again we will use the notion of storage capacity for the maximal number $M(N)$ of patterns such that (4) holds true for all ξ^ν almost surely.

Before we can state our result we now have to describe the form of spatial correlations we are going to study. So let us 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

$$P(\xi_1^\mu = x_1^\mu,\ \mu = 1,\dots,M) = 2^{-M} \quad \text{for all } x_1^\mu \in \{-1,1\} \text{ and all } \mu = 1,\dots,M \tag{5}$$

and transition probabilities

$$\begin{aligned} & P(\xi_i^\mu = x_i^\mu | \xi_j^\nu = x_j^\nu, j = 1,\dots,i-1, \nu = 1,\dots,M) \\ = \ & P(\xi_i^\mu = x_i^\mu | \xi_{i-1}^\mu = x_{i-1}^\mu) = Q(x_{i-1}^\mu, x_i^\mu). \end{aligned} \tag{6}$$

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). Here we use the notation and setup of a Markov chain to stress the similarity with the case of semantically correlated patterns and since it is more convenient. However, let us mention that this setup is equivalent to that of a one dimensional homogeneous Markov random field (see e.g. [Geo88]), which maybe somewhat closer to the usual modelling of an image.

With this definition our results concerning the storage capacity for spatially correlated patterns read as follows:

Theorem 2.1 *Assume the random patterns ξ^μ fulfill (5) and (6) and suppose that $M(N) = \frac{N}{\gamma \log N}$.*
Then the following assertions hold true:

1. *If $\gamma > \frac{3(p^2+(1-p)^2)}{p(1-p)}$*
$$P(\liminf_{N\to\infty}(\cap_{\mu=1}^{M(N)} T\xi^\mu = \xi^\mu)) = 1$$
i.e. the patterns are almost surely stable.

2. *If $\gamma \geq \frac{2(p^2+(1-p)^2)}{p(1-p)}$*
$$P((\cap_{\mu=1}^{M(N)} T\xi^\mu = \xi^\mu)) = 1 - R_N$$
with $\lim_{N\to\infty} R_N = 0$, i.e. all the patterns are stable with probability converging to one.

3. *If $\gamma > \frac{p^2+(1-p)^2}{p(1-p)}$ 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.

Note that $\gamma \to \infty$ as $p \to 0$ or $p \to 1$ (i.e. the bound on the storage capacity decreases) and that for $p = 1/2$ the bound coincides with the well-known bounds for the Hopfield model with i.i.d. patterns (see [MPRV87], [P95]).

The above theorem basically states that – provided M is small enough – the patterns are fixed points of the gradient dynamics and hence are recognized if one starts with them. But just recalling patterns if they are presented without errors is not quite what can be be called an associative memory. What we would like to have is that even if a pattern is corrupted by a certain percentage of noise the gradient dynamics is able to retrieve this pattern. The following theorem shows that with a smaller number of patterns also noised patterns can be successfully reconstructed, if the noise is less than 50 percent (which is the best rate one could hope for, since the Hopfield model is symmetric under spin flip and therefore with an noise of more than 50 percent the negative images become stable).

Theorem 2.2 *Let $r \in [0, \frac{1}{2})$ and for each $\mu = 1, \ldots, M(N)$ let $\tilde{\xi^\mu}$ be an element of the Hamming sphere of radius rN centered at ξ^μ (where, to avoid useless complications in the notations, we assume that rN is an integer). Assume that $M(N) = (1-2r)^2 \frac{N}{\gamma \log N}$. Then:*

1. *If $\gamma > \frac{3(p^2+(1-p)^2)}{p(1-p)}$*
$$P(\liminf_{N\to\infty}(\cap_{\mu=1}^{M(N)} T\tilde{\xi^\mu} = \xi^\mu)) = 1$$
i.e. the noised patterns are almost surely attracted to the centre of the sphere.

2. *If $\gamma \geq \frac{2(p^2+(1-p)^2)}{p(1-p)}$*
$$P((\cap_{\mu=1}^{M(N)} T\tilde{\xi^\mu} = \xi^\mu)) = 1 - R_N$$
with $\lim_{N\to\infty} R_N = 0$, i.e. with probability converging to one all the noised patterns are attracted to the centre of the sphere.

3. *If $\gamma > \frac{(p^2+(1-p)^2)}{p(1-p)}$ for every fixed $\nu = 1, \ldots, M$*
$$P(T\tilde{\xi^\mu} = \xi^\mu) = 1 - R_N$$
with $\lim_{N\to\infty} R_N = 0$, i.e. with probability converging to one every fixed of the noised patterns is attracted to the centre of the sphere.

Let us now turn to the second notion of storage capacity. We will see, that if small errors are tolerated, the Hopfield model indeed can store a number of spatially correlated patterns M proportional to the number of neurons N. The critical α_c will depend on the strength of the correlation and decrease with an increasing correlation.

Theorem 2.3 *Suppose that the random patterns fulfill (5) and (6). There exists an $\alpha_c > 0$ (depending on p) such that if $M(N) \leq \alpha_c N$, then there are $\varepsilon > 0$ and $0 < \delta < 1/2$ such that*

$$P\left(\liminf_{N\to\infty}(\cap_{\mu=1}^{M(N)} \cap_{\sigma\in S_\delta(\xi^\mu)} (H_N(\sigma) \geq H_N(\xi^\mu) + \varepsilon N))\right) = 1$$

where $S_\delta(\xi^\mu)$ is the Hamming sphere of radius δN centered in ξ^ν.

x

3 Proofs

Proof of Theorem 3.1: Note that Theorem 2.1 deals with the first notion of storage capacity. Other than the second one this notion is defined via the so–called local field $\sum_j J_{ij}\sigma_j$ which (other than the Hamiltonian of the Hopfield model) is linear in *both* j and μ. Since therefore the model "can not differentiate" between upper and lower indices, the proof can be basically carried out following the ideas of the proof of Theorem 3.1 in [Lö96a]. However, to avoid confusion of the reader and to keep the paper self–contained we will give the main steps of the proof here.

Fix $1 \leq \nu \leq M(N)$. Following the definition of the dynamics T and the definition of stability introduced in Section 2 the pattern ξ^ν is stable if and only if

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

for all $i = 1, \ldots, N$, or, in other words, if

$$\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{\mu=1}^{M(N)} \xi_i^\nu \xi_j^\nu \xi_i^\mu \xi_j^\mu \geq 0$$

for all $i = 1, \ldots, N$ (where we use the convention $sgn(0) = 1$ which does not really influence our calculations since the probability that the above sum equals zero vanishes for large N).
Hence by bounding the probability of a union of events by the sum of the corresponding probabilities and the exponential Chebyshev–Markov inequality we get for all $t \geq 0$

$$\begin{aligned} P(\xi^\nu \text{ is not stable }) &\leq \sum_{i=1}^{N} P(\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{\substack{\mu=1\\ \mu\neq\nu}}^{M(N)} \xi_i^\nu \xi_j^\nu \xi_i^\mu \xi_j^\mu \leq -N) \\ &\leq \sum_{i=1}^{N} e^{-tN} E\left(\exp(-t(\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{\substack{\mu=1\\ \mu\neq\nu}}^{M(N)} \xi_i^\nu \xi_j^\nu \xi_i^\mu \xi_j^\mu))\right) \\ &= \sum_{i=1}^{N} e^{-tN} \left(E\left(\exp(-t(\sum_{\substack{j=1\\ j\neq i}}^{N} \xi_i^1 \xi_j^1 \xi_i^2 \xi_j^2))\right)\right)^{M-1} \end{aligned} \tag{7}$$

Now putting $Y_i := \xi_i^1 \xi_i^2$ and calculating the expectation in (7) for a fixed index i leads to

$$E(\exp(-t\sum_{\substack{j=1\\ j\neq i}}^{N} \xi_i^1 \xi_i^2 \xi_j^1 \xi_j^2)) = E(\exp(-t\sum_{\substack{j=1\\ j\neq i}}^{N} Y_i Y_j))$$

$$= \sum_{y_j=-1,+1} P(Y_1 = y_1)P(Y_2 = y_2|Y_1 = y_1)\cdots P(Y_i = y_i|Y_{i-1} = y_{i-1}) \times$$

$$
\begin{aligned}
&\times \quad P(Y_{i+1}=y_{i+1}|Y_i=y_i)\cdots P(Y_N=y_N|Y_{N-1}=y_{N-1})\exp(-t\sum_{\substack{j=1\\ j\neq i}}^{N} y_i y_j)\\
&= \quad \frac{1}{2}\sum_{\substack{y_j=-1,+1,\\ j\neq i}} R(y_1,y_2)e^{-ty_1}\cdots R(y_{i-1,},1)e^{-ty_{i-1}}\times\\
&\qquad\qquad \times R(1,y_{i+1})e^{-ty_{i+1}}\cdots R(y_{N-1},y_N)e^{-ty_N} \qquad (8)\\
&+ \quad \frac{1}{2}\sum_{\substack{y_j=-1,+1,\\ j\neq i}} R(y_1,y_2)e^{ty_1}\cdots R(y_{i-1,},-1)e^{ty_{i-1}}\times\\
&\qquad\qquad \times R(-1,y_{i+1}e^{ty_{i+1}}\cdots R(y_{N-1},y_N)e^{ty_N}
\end{aligned}
$$

where R is the transition matrix for the Y_j, i.e.

$$R=\begin{pmatrix} p^2+(1-p)^2 & 2p(1-p)\\ 2p(1-p) & p^2+(1-p)^2\end{pmatrix}=:\begin{pmatrix} q & 1-q\\ 1-q & q\end{pmatrix}.$$

Due to the symmetry of R and the Y_j the two sums on the right hand side in (8) agree such that

$$
\begin{aligned}
&E(\exp(-t\sum_{\substack{j=1,\\ j\neq i}}^{N} Y_iY_j))\\
&= \sum_{\substack{y_j=-1,+1,\\ j\neq i}} R(y_1,y_2)e^{-ty_1}\cdots R(y_{i-1,},1)e^{-ty_{i-1}}\times R(1,y_{i+1})e^{-ty_{i+1}}\cdots R(y_{N-1},y_N)e^{-ty_N}\\
&= \sum_{\substack{y_1=-1,1,\\ y_N=-1,1}} \Pi_L^{i-1}(y_1,1)\Pi_R^{N-i}(1,y_N)
\end{aligned}
$$

where

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

and

$$\Pi_R:=\begin{pmatrix} qe^{-t} & (1-q)e^{t}\\ (1-q)e^{-t} & qe^{t}\end{pmatrix}.$$

Observe that $(\Pi_L)^t=\Pi_R$, implying that Π_L and Π_R have the same eigenvalues which are

$$\lambda_1=q\cosh(t)+\sqrt{1-2q+q^2\cosh^2(t)} \qquad (9)$$

and

$$\lambda_2=q\cosh(t)-\sqrt{1-2q+q^2\cosh^2(t)}.$$

Now

$$
\begin{aligned}
&\sum_{y_1=-1,1,y_N=-1,1} \Pi_L^{i-1}(y_1,1)\Pi_R^{M-i}(1,y_N)\\
&= \quad \left(\begin{pmatrix}1\\1\end{pmatrix}\Pi_L^{i-1}\begin{pmatrix}1\\0\end{pmatrix}\right)\times\left(\begin{pmatrix}1\\0\end{pmatrix}\Pi_R^{N-i}\begin{pmatrix}1\\1\end{pmatrix}\right)\le\lambda_1^{N-1}.
\end{aligned}
$$

Note that the last inequality does not trivially follow from the fact that λ_1 is the largest eigenvalue of Π_L and Π_R. As a matter of fact the inequality can be seen by observing that from e.g. bringing Π_R into diagonal form we get

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \Pi_R^{N-i} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{ad-bc}(a(d-b)\lambda_1^{N-i} + b(a-c)\lambda_2^{N-i})$$

with

$$\begin{aligned} a = \sqrt{1-2q+q^2\cosh^2(t)} - q\sinh(t) \quad b &= (1-q)e^t \\ c = (1-q)e^{-t} \qquad d &= q\sinh(t) - \sqrt{1-2q+q^2\cosh^2(t)}. \end{aligned}$$

Now $2q \geq 1$ and $t \geq 0$ implying that

$$(1-q)(1-e^{-2t}) \leq 2qe^{-t}\sinh(t) = q(1-e^{-2t})$$

and therefore

$$(1-q)^2 + q^2\sinh^2(t) \leq q^2\sinh^2(t) + 2q(1-q)e^{-t}\sinh(t) + (1-q)^2e^{-2t}.$$

Hence

$$\sqrt{1-2q+q^2\cosh^2(t)} - q\sinh(t) - (1-q)e^{-t} \leq 0$$

(since $\sqrt{1-2q+q^2\cosh^2(t)} = \sqrt{(1-q)^2+q^2\sinh^2(t)}$) such that

$$b(a-c) \leq 0.$$

Moreover

$$ad-bc = -\sqrt{1-2q+q^2\cosh^2(t)}(\sqrt{1-2q+q^2\cosh^2(t)} - q\sinh(t)) \leq 0$$

implying that

$$\frac{b(a-c)}{ad-bc} \geq 0.$$

Hence

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \Pi_R^{N-i} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \leq \lambda_1^{N-i}$$

and similarly

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} \Pi_L^{i-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \leq \lambda_1^{i-1}.$$

Since the above estimate

$$\sum_{y_1=-1,1,y_N=-1,1} \Pi_L^{i-1}(y_1,1)\Pi_R^{M-i}(1,y_N) \leq \lambda_1^{N-1}$$

holds true independently of the choice of i we obtain

$$\begin{aligned} P(\exists\nu : \xi^\nu \text{ is not stable }) &\leq MP(\xi^1 \text{ is not stable }) \\ &\leq NMe^{-tN}\lambda_1^{(M-1)(N-1)}. \end{aligned}$$

Expanding the root in (9) using $\sqrt{(1+x)} \leq 1 + \frac{x}{2}$ yields

$$\lambda_1 \leq q\cosh(t) + (1-q) + \frac{q^2\sinh^2(t)}{2(1-q)}.$$

Moreover recalling that

$$\cosh(t) = 1 + \frac{t^2}{2} + \mathcal{O}(t^4)$$

and

$$\sinh(t) = t + \mathcal{O}(t^3)$$

leads to

$$\lambda_1 \leq 1 + t^2\frac{q}{2(1-q)} + \mathcal{O}(t^4) \leq \exp(t^2\frac{q}{2(1-q)})(1+\mathcal{O}(t^4)).$$

Hence we arrive at

$$\begin{aligned} & P(\exists\nu : \xi^\nu \text{ is not stable }) \\ \leq\ & MN\exp(-tN + t^2MN\frac{q}{2(1-q)})(1+\mathcal{O}(t^4))^{MN} \end{aligned}$$

where the last factor on the right hand side converges to 1, if t goes to zero fast enough (and then e.g. is bounded by 2 for all t small enough). Choosing $t = \frac{1-q}{qM}$ gives

$$P(\exists\nu : \xi^\nu \text{ is not stable }) \leq MN\exp(-\frac{1-q}{2q}\frac{N}{M})(1+\mathcal{O}(t^4))^{MN}.$$

So if $M = \frac{N}{\gamma\log N}$ the last factor on the right hand side can be bounded by $\exp(const.\frac{(\log N)^4}{N^2})$ which is converging to one and therefore the choice of $\gamma > \frac{6q}{1-q} = \frac{3(p^2+(1-p)^2)}{p(1-p)}$ leads to the converging series $\sum\frac{1}{N^\kappa\log N}$ for an $\kappa > 1$ and thus proves part one of the theorem by the Borel–Cantelli Lemma. The choice of $\gamma > \frac{4q}{1-q} = \frac{2(p^2+(1-p)^2)}{p(1-p)}$ yields

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

and therefore part two of the theorem.

Part three of the theorem follows by observing that for any fixed ν

$$\begin{aligned} & P(\xi^\nu \text{ is not stable }) \\ \leq\ & N\exp(-tN + t^2MN\frac{q}{2(1-q)})(1+\mathcal{O}(t^4))^{MN} \end{aligned}$$

and then continuing just as above. □

Remark 3.1 *The attentive reader may have noticed that we only made use of the Markov property of our patterns in order to estimate the moment generating function, or more precisely to obtain a bound of the form*

$$E\left(\exp(-t(\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{\substack{\mu=1\\ \mu\neq\nu}}^{M(N)}\xi_i^\nu\xi_j^\nu\xi_i^\mu\xi_j^\mu))\right) \leq \exp(const MNt^2).$$

for t small enough, i.e. that the moment generating function "looks Gaussian" in a neighborhood of zero. This is, of course, not only true for Markov chains (e.g. a similar bound might be shown for appropriate higher dimensional random fields) and indeed it is an easy exercise to prove a corresponding theorem for sequences of patterns admitting a bound of the above form. However, the reason why we stick to Markov chains here is, that we will need this assumption to prove Theorem 2.3, anyway. Moreover in the case of Markov chains one can easily detect the qualitative behavior of the bound on the storage capacity.

The proof of Theorem 2.2 is basically a rerun of the above proof and therefore will be skipped here.

Let us start the proof of Theorem 2.3 with two auxiliary lemmas. The first of them collects two purely analytical statements.

Lemma 3.2 *1. For all $0 < p < 1$ and all $t \in \mathbb{R}$*

$$p\exp(-2(1-p)t) + (1-p)\exp(2pt) \leq \cosh((1+|2p-1|)t) \tag{10}$$

2. Let $x_i \in \mathbb{R}, i = 1, \ldots, N$ with $\sum_{i=1}^N x_i = 0$ and $\sum_{i=1}^N x_i^2 \leq 1$. Then

$$\prod_{i=1}^N (1-x_i) \geq 1 - \sum_{i=1}^N x_i^2. \tag{11}$$

Proof:

1. Note that for any $t \in \mathbb{R}$, and any $1/2 \leq p < 1$ we have

   $$2pe^{-2t} + 2(1-p) \leq 1 + e^{-4pt}$$

   (since the functions agree in 0 and the derivative (with respect to t) of the left hand side can be bounded by the derivative of the right hand side). Therefore by some simple manipulations

   $$pe^{-2(1-p)t} + (1-p)e^{2pt} \leq \cosh((1+2p)t).$$

   Hence by symmetry arguments for all $t \in \mathbb{R}$, and any $0 < p < 1$

   $$pe^{-2(1-p)t} + (1-p)e^{2pt} \leq \cosh((1+|2p-1|)t).$$

2. Note that $K := \{x \in \mathbb{R}^N : \sum_{i=1}^N x_i = 0 \ \sum_{i=1}^N x_i^2 \leq 1\}$ is a compact subset of $\mathbb{R}^N$ and therefore the function

   $$f(x) = \prod_{i=1}^N (1-x_i) + \sum_{i=1}^N x_i^2.$$

   assumes its minimum in some $a \in K$. We have to show that $f(a) \geq 1$. If $a = 0$ we are done.
   If $a \neq 0$ then, since $a \in K$, there are $j \neq k$ with $a_j \neq a_k$.

Define a' by

$$a'_i = \begin{cases} a_i & i \neq j,k \\ \frac{a_j+a_k}{2} & i = j,k. \end{cases}$$

Note that $a' \in K$ by convexity. Moreover

$$\begin{aligned} f(a') &= (\prod_{\substack{i=1 \\ i\neq j,k}}^{N}(1-a_i))(1-\frac{a_j+a_k}{2})^2 + \sum_{\substack{i=1 \\ i\neq j,k}}^{N} a_i^2 + \frac{1}{2}(a_j+a_k)^2 \\ &= \prod_{i=1}^{N}(1-a_i) + \frac{1}{4}(\prod_{\substack{i=1 \\ i\neq j,k}}^{N}(1-a_i))(a_j-a_k)^2 + \sum_{i=1}^{N} a_i^2 - \frac{1}{2}(a_j-a_k)^2 \\ &= f(a) + \frac{1}{4}(\prod_{\substack{i=1 \\ i\neq j,k}}^{N}(1-a_i))(a_j-a_k)^2 - \frac{1}{2}(a_j-a_k)^2. \end{aligned}$$

Since $f(a') \geq f(a)$

$$(\prod_{\substack{i=1 \\ i\neq j,k}}^{N}(1-a_i)) \geq 2$$

implying that

$$\begin{aligned} f(a) &\geq 2(1-a_j)(1-a_k) + a_j^2 + a_k^2 \\ &= 1 + (1-(a_j+a_k))^2 \geq 1. \end{aligned}$$

□

Note that part two of the above lemma can and actually will be used for giving a lower bound on the determinant of $Id - A$ where A is some symmetric matrix with trace 0.

The second lemma shows how to center a quadratic form of a homogeneous Markov chain on $\{-1,+1\}$.

Lemma 3.3 *Assume that $(Y_i)_{i\in\mathbb{N}}$ is a homogeneous Markov chain on $\{-1,+1\}$ with transition matrix*

$$R = \begin{pmatrix} q & 1-q \\ 1-q & q \end{pmatrix}.$$

starting in equilibrium, i.e.

$$P(Y_1 = +1) = P(Y_1 = -1) = \frac{1}{2}.$$

Let $A = (a_{i,j})$ be a symmetric $N \times N$ matrix with $a_{i,i} = 0$ for all $i = 1,\ldots,N$ Define

$$\overline{Y}_i = \begin{cases} Y_i - (2q-1)Y_{i-1} & \text{if } i \geq 2 \\ Y_1 & \text{otherwise.} \end{cases} \tag{12}$$

Then

$$\sum_{i,j=1}^{N} a_{i,j}Y_iY_j = \sum_{i,j=1}^{N} b_{i,j}\overline{Y}_i\overline{Y}_j + 2\sum_{1\leq i<j\leq N}(2q-1)^{j-i}a_{i,j} \tag{13}$$

where

$$b_{i,j} = \sum_{k=0}^{j-i-1} \sum_{l=0}^{N-j} (2q-1)^{k+l} a_{i+k,j+l}. \tag{14}$$

for $i < j$, $b_{i,j} = b_{j,i}$ for $i > j$ and finally $b_{i,i} = 0$ for all i.

Proof: First of all note that due to the symmetry $a_{i,j} = a_{j,i}$ and $b_{i,j} = b_{j,i}$ together with $a_{i,i} = b_{i,i} = 0$ for all i we only have to show that

$$\sum_{1 \le i < j \le N} a_{i,j} Y_i Y_j = \sum_{1 \le i < j \le N} b_{i,j} \overline{Y}_i \overline{Y}_j + \sum_{1 \le i < j \le N} (2q-1)^{j-i} a_{i,j}$$

This will be done by induction. For $N = 2$ we have

$$\begin{aligned} &\sum_{1 \le i < j \le N} a_{i,j} Y_i Y_j = a_{1,2} Y_1 Y_2 = \\ &a_{1,2} \overline{Y}_1 \overline{Y}_2 + (2q-1) a_{1,2} Y_1 Y_1 = \\ &b_{1,2} \overline{Y}_1 \overline{Y}_2 + (2q-1) a_{1,2} = \sum_{1 \le i < j \le N} b_{i,j} \overline{Y}_i \overline{Y}_j + (2q-1) a_{1,2}. \end{aligned}$$

Now suppose (3.3) was already shown for $N-1$. To prove it for N we calculate

$$\begin{aligned} &\sum_{1 \le i < j \le N} a_{i,j} Y_i Y_j \\ = &\sum_{1 \le i < j \le N-1} a_{i,j} Y_i Y_j + \sum_{1 \le i < N} a_{i,N} Y_i Y_N \\ = &\sum_{1 \le i < j \le N-1} a_{i,j} Y_i Y_j + \sum_{1 \le i < N} a_{i,N} Y_i \overline{Y}_N + (2q-1) \sum_{1 \le i < N-1} a_{i,N} Y_i Y_{N-1} + (2q-1) a_{N-1,N} \\ = &\sum_{1 \le i < j \le N-1} a^{(1)}_{i,j} Y_i Y_j + \sum_{1 \le i < N} b_{i,N} \overline{Y}_i \overline{Y}_N + (2q-1) a_{N-1,N} \end{aligned} \tag{15}$$

where we have set

$$a^{(1)}_{i,j} = \begin{cases} a_{i,j} & \text{if } 1 \le i < j \le N-2 \\ a_{i,N-1} + (2q-1) a_{i,n} & \text{if } 1 \le i < j = N-1. \end{cases}$$

Applying the induction hypotheses to first summand in (15) leads to

$$\begin{aligned} &\sum_{1 \le i < j \le N-1} a^{(1)}_{i,j} Y_i Y_j \\ = &\sum_{1 \le i < j \le N-1} \sum_{k=0}^{j-i-1} \sum_{l=0}^{N-j-1} (2q-1)^{k+l} a^{(1)}_{i+k,j+l} \overline{Y}_i \overline{Y}_j + \sum_{1 \le i < j \le N-1} (2q-1)^{j-i} a^{(1)}_{i,j} \\ = &\sum_{1 \le i < j \le N-1} \Big(\sum_{k=0}^{j-i-1} \sum_{l=0}^{N-j-2} (2q-1)^{k+l} a_{i+k,j+l} \\ &\qquad + \sum_{k=0}^{j-i-1} (2q-1)^{k+N-1} (a_{i+k,N-1} + (2q-1) a_{i+k,N}) \Big) \overline{Y}_i \overline{Y}_j \end{aligned}$$

$$
\begin{aligned}
&+ \sum_{1\le i<j\le N-2} (2q-1)^{j-i} a_{i,j} + \sum_{1\le i<N-1} (2q-1)^{N-i-1}(a_{i,N-1} + (2q-1)a_{i,N}) \\
&= \sum_{1\le i<j\le N-1} b_{i,j}\overline{Y}_i\overline{Y}_j \\
&+ \sum_{1\le i<j\le N-2} (2q-1)^{j-i} a_{i,j} + \sum_{1\le i<N-1} (2q-1)^{N-i-1}(a_{i,N-1} + (2q-1)a_{i,N})
\end{aligned}
$$

such that together with the above calculations indeed

$$
\sum_{1\le i<j\le N} a_{i,j} Y_i Y_j = \sum_{1\le i<j\le N} b_{i,j}\overline{Y}_i\overline{Y}_j + \sum_{1\le i<j\le N} (2q-1)^{j-i} a_{i,j}.
$$

□

Now we are ready to prove Theorem 2.3.

Proof of Theorem 2.3: The main steps of the proof consist of a centering of the patterns as prepared in Lemma 3.3 and by replacing them by appropriate Gaussian random variables. Although the second part of this basic idea is fairly standard in the context of storage capacity estimates (see e.g. [N88], [BG92]) in our situation the computations become technically quite involved (especially we will make use of Lemma 3.2 to be able to treat the moment generating function of the resulting quadratic form of Gaussian random variables).

We set

$$
h_N(\sigma,\delta) := \inf_{\sigma'\in S_\delta(\sigma)} H_N(\sigma').
$$

First of all observe that

$$
\begin{aligned}
&P\left(\{\cap_{\nu=1}^{M(N)}(h_N(\xi^\nu,\delta) \ge H_N(\xi^\nu)+\varepsilon N)\}^c\right) \\
\le &\sum_{J:|J|=\delta N} \sum_{\nu=1}^{M(N)} P\left(H_N(\xi^\nu_J) - H_N(\xi^\nu) \le \varepsilon N\right)
\end{aligned}
$$

where ξ^ν_J denotes a configuration differing from ξ^ν exactly in the coordinates J and δ is chosen in such a way that δN is an integer.
Let us keep ν fixed in the sequel (without loss of generality we choose $\nu = 1$) and note that

$$
H_N(\xi^1_J) - H_N(\xi^1) = \frac{1}{N}\sum_{\mu\neq 1} \sum_{\substack{(i\in J, j\notin J)\\(i\notin J, j\in J)}} \xi^1_i\xi^1_j\xi^\mu_i\xi^\mu_j + 2\delta(1-\delta)N.
$$

Thus by the exponential Chebyshev-Markov inequality for any $t \ge 0$

$$
\begin{aligned}
P\left(H_N(\xi^1_J) - H_N(\xi^1) \le \varepsilon N\right) &\le e^{-t\varepsilon' N} E\left(\exp\left(-\frac{t}{N}\sum_{\mu\neq 1}\sum_{\substack{(i\in J, j\notin J)\\(i\notin J, j\in J)}} \xi^1_i\xi^1_j\xi^\mu_i\xi^\mu_j\right)\right) \\
&= e^{-t\varepsilon' N} E\left(\exp\left(-\frac{t}{N}\sum_{\substack{(i\in J, j\notin J)\\(i\notin J, j\in J)}} \xi^1_i\xi^1_j\xi^2_i\xi^2_j\right)\right)^{M-1} \qquad (16)
\end{aligned}
$$

where we have set $\varepsilon' = -\varepsilon + 2\delta(1-\delta)$.
Our main goal is now to estimate the expectation on the right hand side of (16). To this end note that the exponent is a genuine quadratic form in Markovian random variables as treated in Lemma 3.3. Indeed, putting

$$Y_i := \xi_i^1 \xi_i^2$$

and

$$a_{i,j} := \begin{cases} 1 & \text{if } i \in J, j \notin J \\ 1 & \text{if } i \notin J, j \in J \\ 0 & \text{otherwise} \end{cases}$$

(note that this especially implies that $a_{i,i} = 0$) we obtain that

$$E\left(\exp\left(-\frac{t}{N}\sum_{\substack{(i\in J, j\notin J)\\(i\notin J, j\in J)}} \xi_i^1\xi_j^1\xi_i^2\xi_j^2\right)\right) = E\left(\exp\left(-\frac{t}{N}\sum_{i,j} a_{i,j}Y_iY_j\right)\right).$$

Observe that Y_i is a Markov chain on the set $\{-1,+1\}$ (in i) with transition matrix R given by

$$R = \begin{pmatrix} p^2+(1-p)^2 & 2p(1-p) \\ 2p(1-p) & p^2+(1-p)^2 \end{pmatrix} =: \begin{pmatrix} q & 1-q \\ 1-q & q \end{pmatrix}$$

(notice that $q \geq 1/2$). Hence we are in the situation of Lemma 3.3. Therefore

$$E\left(\exp\left(-\frac{t}{N}\sum_{i,j} a_{i,j}Y_iY_j\right)\right) =$$
$$\exp\left(-\frac{2t}{N}\sum_{1\leq i<j\leq N}(2q-1)^{j-i}a_{i,j}\right) E\left(\exp\left(-\frac{t}{N}\sum_{i,j} b_{i,j}\overline{Y}_i\overline{Y}_j\right)\right) \qquad (17)$$

where the $\overline{Y}_i$ and $b_{i,j}$ are defined as in (12) and (14) resp. To estimate the right hand side of (17) observe that due to the symmetry of the $a_{i,j}$ we have that $2\sum_{1\leq i<j\leq N}(2q-1)^{j-i}a_{i,j} = \sum_{1\leq i,j\leq N}(2q-1)^{|j-i|}a_{i,j}$ and that

$$2\left(\frac{2q-1}{2(1-q)}\right)^2(1-(2q-1)^{\delta N})(1-(2q-1)^{(1-\delta)N}) \leq \sum_{1\leq i,j\leq N}(2q-1)^{|j-i|}a_{i,j} \leq \frac{N}{q}. \qquad (18)$$

The upper bound in (18) is derived by simply setting all the $a_{i,j} = 1$ and is just to show that the term $\exp\left(-\frac{2t}{N}\sum_{1\leq i<j\leq N}(2q-1)^{j-i}a_{i,j}\right)$ is at most of the same order as the expectation on the right hand side of (17) (which for example may fail in the Hopfield model with biased patterns if the Hamiltonian is not appropriately normalized, see [Lö96b] and then would be an index that the model is not chosen correctly). The lower bound in (18) is due to the fact the number of $a_{i,j}$ that are equal to one is independent of the choice of J such that $\sum_{1\leq i,j\leq N}(2q-1)^{|j-i|}a_{i,j}$ becomes minimal for sets J which are maximally connected, e.g. $J = \{1,\ldots,\delta N\}$. Hence for N large enough

$$\exp\left(-\frac{2t}{N}\sum_{1\leq i<j\leq N}(2q-1)^{j-i}a_{i,j}\right) \leq \exp\left(-\frac{t}{N}\left(\frac{2q-1}{2(1-q)}\right)^2\right).$$

To estimate the expectation on the right hand side of (17) we employ Lemma 3.2. First of all let us agree on the following notation: With E_I (where $I \subset \{1,\dots,N\}$) we denote the integration with respect to those random variables Y_i (resp. $\overline{Y}_i$) with $i \in I$. With these notations we will be able to prove that

$$E\left(\exp\left(-\frac{t}{N}\sum_{1\le i,j\le N} b_{i,j}\overline{Y}_i\overline{Y}_j\right)\right) \le E\left(\exp\left(\frac{2qt}{N}\sum_{1\le i,j\le N} b_{i,j}z_iz_j\right)\right), \qquad (19)$$

where the z_i are i.i.d. standard Gaussian random variables.
Indeed,

$$\begin{aligned}
& E\left(\exp\left(-\frac{t}{N}\sum_{i,j} b_{i,j}\overline{Y}_i\overline{Y}_j\right)\right) = E\left(\exp\left(-\frac{2t}{N}\sum_{1\le i<j\le N} b_{i,j}\overline{Y}_i\overline{Y}_j\right)\right) \\
= \;& E_{\{1,\dots,N-1\}}\left(\exp\left(-\frac{2t}{N}\sum_{1\le i<j\le N-1} b_{i,j}\overline{Y}_i\overline{Y}_j\right)\right) E_{\{N\}}\left(\exp\left(-\frac{2t}{N}\sum_{i=1}^{N-1} b_{i,N}\overline{Y}_i\overline{Y}_N\right)\right) \\
= \;& E_{\{1,\dots,N-1\}}\left(\exp\left(-\frac{2t}{N}\sum_{1\le i<j\le N-1} b_{i,j}\overline{Y}_i\overline{Y}_j\right)\right) \\
& \times\left(q\exp\left(-4(1-q)\frac{t}{N}\sum_{i=1}^{N-1} b_{i,N}\overline{Y}_i\right) + (1-q)\exp\left(4q\frac{t}{N}\sum_{i=1}^{N-1} b_{i,N}\overline{Y}_i\right)\right) \\
\le \;& E_{\{1,\dots,N-1\}}\left(\exp\left(-\frac{2t}{N}\sum_{1\le i<j\le N-1} b_{i,j}\overline{Y}_i\right)\right)\times\cosh\left(\frac{4tq}{N}\sum_{i=1}^{N-1} b_{i,N}\overline{Y}_i\right) \\
\le \;& E_{\{1,\dots,N-1\}}\left(\exp\left(-\frac{2t}{N}\sum_{1\le i<j\le N-1} b_{i,j}\overline{Y}_i\right)\right)\times\exp\left(2\frac{q^2t^2}{N^2}(\sum_{i=1}^{N-1} b_{i,N}\overline{Y}_i)^2\right)
\end{aligned}$$

where we have used the inequality (10) together with

$$\cosh(x) \le \exp(x^2/2).$$

Recalling the well-known identity

$$\exp(\frac{1}{2}x^2) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp(xy-\frac{1}{2}y^2)dy$$

yields

$$\begin{aligned}
& E\left(\exp\left(-\frac{t}{N}\sum_{i,j} b_{i,j}\overline{Y}_i\overline{Y}_j\right)\right) \\
\le \;& E_{\{1,\dots,N-1\}}\left(\exp\left(-\frac{2t}{N}\sum_{1\le i<j\le N-1} b_{i,j}\overline{Y}_i\right)\right) E_{z_N}\exp\left(\frac{4qt}{N}\sum_{i=1}^{N-1} b_{i,N}\overline{Y}_i z_N\right)
\end{aligned}$$

where E_{z_N} denotes integration with respect to the Gaussian random variable z_N. Note that we have replaced $\overline{Y}_N$ by z_N and also observe that this replacement was independent of all the other $\overline{Y}_i$, $i \ne N$ and just relied on the fact that $\overline{Y}_N$ was appropriately centered

such that Lemma 3.2.1 could be applied. Since this is true for all the $\overline{Y}_i$,too, the other replacements can be carried out along the same lines just be rearranging the order of integration appropriately, such that (19) indeed is true.

So we have arrived at

$$E\left(\exp\left(-\frac{t}{N}\sum_{1\leq i,j\leq N}a_{i,j}Y_iY_j\right)\right)\leq\exp\left(-\frac{t}{N}\left(\frac{2q-1}{2(1-q)}\right)^2\right)E\left(\exp\left(\frac{2qt}{N}\sum_{1\leq i,j\leq N}b_{i,j}z_iz_j\right)\right). \tag{20}$$

Note that the expectation on the right hand side of (20) is only defined if t is small enough, i.e. if $Id-\frac{2qt}{N}B$ is a positive definite matrix (here Id denotes the $N\times N$ identity matrix and B is the quadratic form given by $B=(b_{i,j})_{1\leq i,j\leq N}$). In this case

$$E\left(\exp\left(\frac{2qt}{N}\sum_{i,j}b_{i,j}z_iz_j\right)\right)=\frac{1}{\sqrt{\det(Id-\frac{2qt}{N}B)}}.$$

To estimate this determinant denote by $(\beta_i)_{1\leq i\leq N}$ the (real) eigenvalues of B. Then, on one hand, since Id and B are trivially simultaneously diagonalizable,

$$\det(Id-\frac{2qt}{N}B)=\prod_{i=1}^{N}(1-\frac{2qt}{N}\beta_i),$$

on the other hand

$$\sum_{i=1}^{N}\beta_i=0,$$

since all the diagonal entries of B are zero.
Now let us assume that t is so small that even

$$\frac{4t^2q^2}{N^2}\sum_{i=1}^{N}\beta_i^2\leq 3/4. \tag{21}$$

An immediate application of (11) yields that

$$\prod_{i=1}^{N}(1-\frac{2qt}{N}\beta_i)\geq 1-\frac{4q^2t^2}{N^2}\sum_{i=1}^{N}\beta_i^2$$

implying that

$$\frac{1}{\sqrt{\det(Id-\frac{2qt}{N}B)}}\leq\frac{1}{\sqrt{1-\frac{4q^2t^2}{N^2}\sum_{i=1}^{N}\beta_i^2}}.$$

And since

$$\frac{1}{\sqrt{1-x}}\leq e^x$$

for $0\leq x\leq 3/4$ we obtain

$$\frac{1}{\sqrt{\det(Id-\frac{2qt}{N}B)}}\leq\exp\left(\frac{4q^2t^2}{N^2}\sum_{i=1}^{N}\beta_i^2\right).$$

Finally, observe that due to the symmetry of B

$$\sum_{i=1}^{N} \beta_i^2 = tr(B^2) = \sum_{i=1}^{N} \sum_{j=1}^{N} b_{i,j}^2.$$

Since

$$\sum_{i,j=1}^{N} a_{i,j} = 2\delta(1-\delta)N^2$$

we arrive at

$$\sum_{i=1}^{N} \sum_{j=1}^{N} b_{i,j}^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=0}^{j-i-1} \sum_{l=0}^{N-j} (2q-1)^{k+l} a_{i+k,j+l} \leq 2 \left(\frac{1}{2(1-q)} \right)^2 \delta(1-\delta)N^2. \quad (22)$$

So altogether for $t \geq 0$ and N large enough

$$\begin{aligned} & P\left(H_N(\xi_J^1) - H_N(\xi^1) \leq \varepsilon N\right) \leq \\ & e^{-t\varepsilon' N} \exp\left(-\frac{tM}{N} \left(\frac{2q-1}{2(1-q)} \right)^2 \right) \exp\left(8q^2t^2 \left(\frac{1}{2(1-q)} \right)^2 \delta(1-\delta)M \right). \quad (23) \end{aligned}$$

The proof now could in principle be finished following the ideas of the proofs of the corresponding theorems in [N88], [BG92], or [Lö96a]. However, to take into account the minor changes to be made and for the sake of completeness let us also give the final arguments here. First of all note that the above bound (23) is uniform in J with $|J| = \delta N$. Hence by setting $M = \alpha N$ and estimating the term $\binom{N}{\delta N}$ by Stirling's formula we arrive at

$$\begin{aligned} & \sum_{J:|J|=\delta N} \sum_{\nu=1}^{M(N)} P\left(H_N(\xi_J^\nu) - H_N(\xi^\nu) \leq \varepsilon N\right) \\ \leq \; & \alpha N \inf_{t^* \geq t \geq 0} \exp\left((-\delta \log \delta - (1-\delta)\log(1-\delta))N \right) \\ & \times \exp\left(-t \left(\varepsilon' N + \alpha \left(\frac{2q-1}{2(1-q)} \right)^2 \right) + 8t^2q^2 \left(\frac{1}{2(1-q)} \right)^2 \delta(1-\delta)\alpha N \right). \end{aligned}$$

Now choose

$$t = t^* := \frac{(1-q)^2}{4\alpha q^2}$$

and

$$\alpha \leq \sqrt{\frac{\delta(1-\delta)}{6}} \frac{1-q}{q} \quad (24)$$

then t^* is admissible in the sense that (21) is fulfilled. Indeed, together with (22) we obtain

$$\frac{4t^2q^2}{N^2} \sum_{i=1}^{N} \beta_i^2 \leq 2 \frac{4q^2(1-q)^4}{16q^4\alpha^2N^2} \frac{1}{2(1-q)^2} \delta(1-\delta)N^2 = \frac{(1-q)^2\delta(1-\delta)}{8q^2\alpha^2} \leq \frac{3}{4}.$$

Moreover by choosing $\varepsilon > 0$ small (e.g. $\varepsilon = \delta(1-\delta)$ such that also $\varepsilon' = \delta(1-\delta)$ we get that

$$\begin{aligned}
&\sum_{J:|J|=\delta N} \sum_{\nu=1}^{M(N)} P\left(H_N(\xi_J^\nu) - H_N(\xi^\nu) \leq \varepsilon N\right) \\
\leq\ & \alpha N \exp\left((-\delta\log\delta - (1-\delta)\log(1-\delta))N\right) \\
& \times \exp\left(\frac{1}{8}\frac{N\delta(1-\delta)}{\alpha}\frac{(1-q)^2}{q^2} - \frac{\alpha(2q-1)}{2(1-q)}\right) \\
\leq\ & \alpha N \exp\left((-\delta\log\delta - (1-\delta)\log(1-\delta))N\right) \exp\left(\frac{\sqrt{6}}{8} N\sqrt{\delta(1-\delta)}\frac{(1-q)}{q}\right) \quad (25)
\end{aligned}$$

As it is quickly checked that for each positive constant C there is an interval $[0, r]$ (depending on C, of course) such that

$$C\sqrt{\delta(1-\delta)} \geq -\delta\log\delta - (1-\delta)\log(1-\delta)$$

for all $\delta \in [0, r]$, the above exponent becomes negative if we choose δ small enough. This completes the proof of the theorem. □

Let us finally comment a little on the result of Theorem 2.3. Observe that the bound on the moment generating function in (25) as well as the bound on α in (24) depends on p mainly via the factor $\frac{1-q}{q}$ which converges to zero for p close to one or close to zero and therefore large correlations can only deteriorate the bounds for α (allowing smaller $\alpha's$ only) for large correlations. Due to the many estimates in the proof of Theorem 2.3 this is, of course, in no way a proof that the storage capacity decreases with an increasing correlation (only our bounds do), but it might either indicate that the Hopfield model has problems to store patterns with large correlations or it just shows that our estimates get worse for large q (i.e. very large or very small p). (Independent of whether the first part of the last statement is true the second one is probably right, which e.g. can be seen from the fact that our bound on

$$P\left(H_N(\xi_J^\nu) - H_N(\xi^\nu) \leq \varepsilon N\right)$$

is independent of the special choice of J and just depends on its size; on the other hand we are firmly convinced that improving the estimates in this point may, of course, increase the numerical value of the storage capacity (if we were after such a number), but would not change the qualitative behavior in p).
However, as already mentioned in Section 1, a decrease of storage capacity (when the correlation increases) would not be totally unexpected due to the way the Hopfield model is assumed to work.

Finally let us mention that we are firmly convinced that the storage abilities of the Hopfield model are not limited to the case where the correlations come from a Markov chain (at least the cases of exponentially decaying correlations or even summable correlations should be tractable). Our method, however, seems to be. The reader may notice that even for the (interesting) case of a two dimensional Markov random field (Ising model) the centred variables do not only depend on the two-point correlations any longer, such that new methods to treat the corresponding moment generating function have to be invented.

References

[AGS85] D.J. Amit, G. Gutfreund, H. Sompolinsky; Spin-glass models of neural; Phys. Rev. A 32, 1007-1018 (1985)

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

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

[BG95a] A.Bovier, V.Gayrard; An almost sure large deviation principle for the Hopfield model, WIAS-preprint 146, to appear in Ann. Prob. (1996)

[BG95b] A.Bovier, V.Gayrard; The retrieval phase of the Hopfield model, WIAS-preprint 161, to appear in Prob. Th. Rel. Fields (1996)

[BG96] A.Bovier, V.Gayrard; Hopfield models as a generalized mean field model, Preprint, to appear in "Mathematics of spin glasses and neural networks", A. Bovier, P.Picco (eds.), "Progress in Probability", Birkhäuser (1997)

[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)

[BGP95a] 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)

[BGP95b] A.Bovier, V.Gayrard, P.Picco; Gibbs states for the Hopfield model with extensively many patterns; J. Stat. Phys. 79, 395-414 (1995)

[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)

[Gen95] B. Gentz; A central limit theorem for the overlap parameter in the Hopfield model; Preprint; Universität Zürich (1995), to appear in Ann. Prob.

[Geo88] H.-O. Georgii; Gibbs measures and phase transition; de Gruyter Studies in Mathematics 9 (1988)

[HK91] L. van Hemmen, R. Kühn, Collective phenomena in neural networks; in Models of neural networks, E. Domany, L. v. Hemmen, R. Schulte (eds.), Springer, Berlin (1991)

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

[KS94] R. Kühn, H.Steffan; Replica symmetry breaking in attractor neural network models; Z. Phys. B 95, 249–260 (1994)

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

[Lou95] D. Loukianova; Two rigorous bounds in the Hopfield model of associative memory; Preprint, to appear in Prob. Th. Rel. Fields (1995)

[Lö96a] M. Löwe; On the storage capacity of Hopfield models with weakly correlated patterns, Preprint, Universität Bielefeld, submitted (1996)

[Lö96b] M. Löwe; On the storage capacity of the Hopfield model with biased patterns, Preprint, Universität Bielefeld, submitted (1996)

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

[Mi88] Y. Miyashita, Neuronal correlate of visual associative long term memory in the primate temporal cortex, Nature 335, 817–819 (1988)

[Mo92] R. Monasson; Properties of neural networks storing spatially correlated patterns; J. Phys. A: Math. Gen. 335, 3701-3720 (1992)

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

[P95] D. Petritis; Thermodynamic formalism of neural computing; Preprint, Université de Rennes I (1995)

[S96] M. Sandmeier; On the storage capacity of neural networks with temporal association; Preprint, Universität Bielefeld (1996)

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

[T96] M. Talagrand; Rigorous Results of the Hopfield Model with Many Patterns; Preprint, Paris (1996)

[TL93] W. Tarkowski, M. Lewenstein; Storage of sets of correlated data in neural network memories, J. Phys. A: Math. Gen. 26, 2453- 2469 (1993)