On the Storage Capacity of the Hopfield Model with Biased 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 introduce a form of the Hopfield model that is able to store an increasing number of biased i.i.d. patterns (it is well known that the standard Hopfield model fails to work properly in this context). We show that this new form of the Hopfield model with N neurons can store $\frac{N}{\gamma \log N}$ or αN biased patterns (depending on which notion of storage is used). The quantity γ increases with an increasing bias of the patterns, while α decreases when the bias gets large.

1 Introduction

Neural networks, especially the so–called Hopfield model, have been in the centre of interest in the recent probabilistic and physical literature. Originally introduced by Pastur and Figotin [FP77] as a simplified model of a spin glass the Hopfield model connects the ideas of neural computing with those of statistical mechanics. However, it was its re-interpretation as a neural network by Hopfield [Ho82] that created the interest in the model.

The basic idea behind this model is to choose the information to be stored – which usually is referred to as patterns – as the local minima of an appropriate Hamiltonian (this is going to be made more formal in the next section) on the set $\{1,\ldots,N\}$ which is thought of as the set of neurons. Hence the retrieval dynamics (usually a Monte–Carlo dynamics, possibly at zero temperature) will eventually converge to the stored patterns. Since, according to the other rules of neural networks, this is possible only up to a certain accuracy depending on the number of patterns to be stored, one may introduce the notion of a storage capacity of the Hopfield model as the critical number of patterns (depending on N, of course) up to which the model can successfully reconstruct these patterns.

To be more precise, at zero temperature there are two distinct definitions of storage capacity: The first one introduces the gradient (or zero temperature Monte–Carlo) dynamics on the energy landscape on $\{1,\ldots,N\}$ associated with the Hamiltonian of the system and requires that the patterns are fixed points of this dynamics (and thus minima of the Hamiltonian). The second one going back to a seminal paper of Newman [N88] defines the notion of storage capacity more liberally requiring that a pattern is surrounded by an energy barrier of height proportional to N implying that it is close to a local minimum of the Hamiltonian.

It is well known and first was announced by McEliece et al. [MPRV87] that the first definition leads to a storage capacity of at least $\frac{N}{\gamma\log N}$ (where the possible values of γ range between 2 and 6 and basically depend on whether we are interested in results with high probability or in results for almost all patterns) for the Hopfield model with N neurons and randomly and independently chosen unbiased patterns (also see [M92], [B94], and the recent survey paper by Petritis [P95]). On the other hand the essential progress of Newman's work ([N88]) was that he was able to give a proof for the non-rigorous result of Amit, Gutfreund and Sompolinsky [AGS87] showing that the Hopfield model allows to store even αN patterns (for an $\alpha > 0$ small enough), if small errors are tolerated. The value for α obtained by Newman was $\alpha = 0.056$, which has been recently improved by Loukianova [L94] to $\alpha = 0.071$ by a refined large deviations analysis. The latest result concerning the critical storage capacity is due to Talagrand [T95b],[T96] who has given a bound of $\alpha = 0.08$, obtained via an improvement of Loukianova's idea. Nevertheless, it is still an open question, whether the prediction of $\alpha = 0.14$ in [AGS87] on the basis of the non-rigorous replica method and computer simulations can be mathematically justified.

The other important source of interest in the Hopfield model is that it may serve as a simplified spin glass model, where the complexity of the random environment can be controlled by the number of stored patterns. As a matter of fact, if this number equals one the Hopfield model is (by a simple change of variables) equivalent to the classical Curie-Weiss model of a ferromagnet and thus the interactions are purely deterministic. Quite differently in the region where the number of patterns roughly equals the number of neurons the model is expected to behave similar to the Sherrington–Kirkpatrick model of a spin glass. Motivated by such considerations one has achieved a fairly complete picture of the small α-regime (where still $\alpha = \frac{M}{N}$) of the Hopfield model with i.i.d. unbiased patterns in the last few years. The results mainly were obtained in a series of papers by Bovier and Gayrard, partially in collaboration with Picco [BGP94], [BGP95a], [BGP95b],[BG95a],[BG95b]. Another milestone certainly has been 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 we refer the reader to [BG96].

If the patterns are not chosen to be i.i.d. and unbiased the picture is less clear. To our knowledge the only rigorous result concerning such sequences of patterns is given in [Lö96] where it is shown that weak semantical correlations of the patterns do not destroy the extensive storage capacity of the Hopfield model (although the numerical values for the number of patterns that can be store are somewhat lower than for i.i.d. patterns and decrease when the correlation become large). In this paper we will analyze the Hopfield model with biased patterns. Such a situation is, of course, interesting in its own right, since unbiasedness of the stored data is a simplification that usually is not fulfilled in realistic situations (for example when storing images that are in average more black than white). Moreover, our results may be considered as a step towards showing the universality of the Hopfield model. Finally, the case of biased patterns often has served as some sort of a toy model for the case of spatially correlated data (which are a natural model e.g. for the storage of images) (see e.g. [HK91], [FI91]) which we will analyze elsewhere.

As has already been remarked in [HK91] the "classical" Hopfield model is not at all able to store an increasing amount of biased patterns. This is due to the fact that a Hamiltonian with biased patterns tends to favor ferromagnetic states (all neurons $\sigma_i = +1$ or all $\sigma_i = -1$) and hence for large numbers of biased patterns there are only two stable states (not related to the patterns anymore, of course), or in terms of a signal–to–noise analysis, the signal of a single pattern becomes deterministically smaller than the noise created by the bias of the other patterns. However, this problem is readily overcome by just centering the patterns in the Hamiltonian. Note that this is the most natural way to handle the problem (although e.g. [HK91] propose to use the so-called pseudo–inverse rule, which not only contradicts the rules of neural computing in being non-local, but also seems to be inappropriate to our setup, since it was designed to eliminate correlations among the patterns (which our patterns do not have anyway)).

With this setup the results we obtain strongly resemble the results of [Lö96]: The Hopfield model is shown to be able to store a number of biased patterns in the same order of magnitude as the number of unbiased patterns that can be stored (no matter whether the "dynamic" or the "static" (the latter being the one due to Newman) notion of storage is used). The storage capacities we obtain are below those for independent patterns (though similar techniques are employed) and the capacities decrease when the bias becomes larger.

We organize this paper in the following way: Section 2 will contain the basic setup, especially the type of patterns we consider and the definition of the Hopfield model we have in mind. In Section 3 we give our results concerning the storage capacity of the Hopfield model with biased patterns. We analyze both the dynamic and the static notion of storage. Finally, Section 4 contains the proofs.

2 The Hopfield Model

Recall that each neural network consists of a set of neurons and a set of synapses which can appropriately be modeled as the vertices and edges of a (usually directed) (hyper-)graph. In the case of the Hopfield model this graph has a fairly simple structure, since it is given by the complete graph on N (labeled) vertices K_N. Each of the "neurons" (i.e. each of the vertices of the K_N) σ_i may assume one of two possible values: +1 (on) or -1 (off). Hence the state of the network can be described by a vector of plus and minus ones of length N.

To describe the "synaptic efficacies", i.e. the strength of the interaction between we first of all have to introduce the information to be stored, called the patterns. So throughout the rest of the paper we consider a sequence $(\xi^\nu)_{\nu\in\mathbb{N}}$ of patterns, each of them being itself a sequence of plus and minus ones. So altogether we have an $\mathbb{N}\times\mathbb{N}$ array $((\xi_i^\nu)_{i\in\mathbb{N}})_{\nu\in\mathbb{N}} \in \{-1,+1\}^{\mathbb{N}\times\mathbb{N}}$. In this paper we will also assume that the ξ_i^ν are randomly chosen and that the random variables ξ_i^ν are i.i.d. with

$$P(\xi_i^\nu = 1) = p \quad \text{and} \quad P(\xi_i^\nu = -1) = 1-p \tag{1}$$

(for a model where the patterns are not i.i.d. see [Lö96]).

The synaptic efficacies J_{ij} connecting neurons i and j now is defined by Hebb's learning rule

$$J_{ij} = \sum_{\nu=1}^{M(N)} \overline{\xi_i^\nu}\,\overline{\xi_j^\nu},$$

where $M(N)$ denotes the number of patterns to be stored with N neurons, (note that $M(N)$ can be regarded as an index for the complexity of the neural network on one hand and for the complexity of the random environment $\sum_{\nu=1}^{M(N)} \overline{\xi_i^\nu}\,\overline{\xi_j^\nu}$ on the other). The $\overline{\xi_i^\nu}$ are the centered patterns ξ_i^ν, i.e.

$$\overline{\xi_i^\nu} = \xi_i^\nu - (2p-1)$$

(and we choose such a centring since it can be easily checked that the Hopfield model is not able to store any increasing amount of biased patterns that are not centred (see e.g. [HK91])). Note that in the case of unbiased patterns (i.e. for $p=\frac{1}{2}$) this choice agrees with the usual definition of the Hopfield model (see e.g. [P95]) and thus our setup can be regarded as a generalization of the standard Hopfield model.

It was one of the main achievements of Hopfield [Ho82] to observe that the synaptic efficacies always induce a Hamiltonian or energy function on the state space of the neurons. More precisely in our context the Hamiltonian of the Hopfield model on N neurons is defined as

$$H_N(\sigma) = -\frac{1}{2N} \sum_{i,j=1}^{N} \sigma_i \sigma_j J_{ij} = -\frac{1}{2N} \sum_{i,j=1}^{N} \sum_{\nu=1}^{M(N)} \sigma_i \sigma_j \overline{\xi_i^\nu}\, \overline{\xi_j^\nu}, \quad (2)$$

The idea behind the choice of this Hamiltonian is that it is in a way the "simplest" reasonable function of the so-called overlap vector given by

$$m_N(\sigma) := (m_N^\nu(\sigma))_{\nu \in \mathbb{N}} := (\frac{1}{N} \sum_{i=1}^{N} \overline{\xi_i^\nu} \sigma_i)_{\nu \in \mathbb{N}}$$

which measures how much a configuration σ agrees with one of the given patterns. More precisely we have

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

where $|| \cdot ||_2$ denotes the norm in l_2.

Since the retrieval dynamics of the Hopfield model is assumed to be a Monte-Carlo dynamics on the energy landscape given by H_N (which has the Gibbs-measure as its invariant measure) it is convenient to identify the Hopfield model itself with the Gibbs measure associated with $H_N(\sigma)$, i.e.

$$\varrho_{N,\beta}(\sigma) := \frac{e^{-\beta H_N(\sigma)}}{Z_{N,\beta}}$$

where β denotes the inverse temperature and $Z_{N,\beta}$ is the normalizing constant

$$Z_{N,\beta} := \sum_{\sigma \in \{-1,1\}^N} e^{-\beta H_N(\sigma)}.$$

Note that (at least for β large enough) $\varrho_{N,\beta}(\sigma)$ favors configurations with low energy, and the hope is that for $M(N)$ small enough these configurations agree with or at least are close to the stored patterns.

This motivates the following two definitions of storage capacity. For the first consider the Hopfield Hamiltonian (2) with $M := M(N)$ patterns and note that it can be expressed in terms of the so called local fields (stressing that in physical terms the Hopfield model is a generalized mean-field model)

$$h_i := \sum_{j=1}^{N} \sigma_j J_{ij}.$$

The gradient (or zero temperature Monte-Carlo) dynamics T is supposed to work in the following way:

Pick a site i at random and flip the spin σ_i if that lowers the energy of the configuration; otherwise stay with that spin. Mathematically this corresponds to a mapping T on the space $\{-1,+1\}$ given by

$$T : \sigma_i \mapsto sgn(\sum_{j=1}^{N} \sigma_j J_{ij}) \tag{3}$$

where sgn is the sign function. We call a configuration $\sigma = (\sigma_i)_{i\leq N}$ stable if it is a fixed point of T, i.e.

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

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 Newman [N88]. It takes into account small errors we are willing to accept in the restoration of the patterns (with the idea to increase the storage capacity, of course). 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.

As already mentioned in the introduction and has been proved in [MPRS87] and [N88], resp., (also see [P95]) the Hopfield model works well as an associative memory if the patterns are chosen i.i.d. and unbiased, i.e.$P(\xi_i^\nu = 1) = P(\xi_i^\nu = -1) = 1/2$ independently of all the other ξ_j^μ and if $M(N) < \frac{N}{6\log N}$ and $M(N) \leq \alpha N$ for a small α, resp., depending on which definition of storage is used.

3 Results

In this section we will give a lower bound on the number of biased patterns of the form (1) that can be stored in the Hopfield model (2).

Let us first of all analyze the storage capacity for the "dynamic" notion of storage.

Theorem 3.1 *Assume the random patterns ξ^ν fulfill (1) and $M(N) = \frac{N}{\gamma\log N}$. Then for the Hopfield model (2) the following assertions hold true:*

1. *If* $\gamma > \frac{3}{8p^2(1-p)^2}$

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

i.e. the patterns are almost surely stable.

2. *If* $\gamma > \frac{1}{4p^2(1-p)^2}$

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

with $\lim_{N\to\infty} R_N = 0$.

3. *If* $\gamma > \frac{1}{8p^2(1-p)2}$ *for every fixed* $\nu = 1, \ldots, M$

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

with $\lim_{N\to\infty} R_N = 0$.

Here T is the mapping given by (3).

Theorem 3.1 in other words states that the patterns are fixed points of the gradient dynamics and hence are recognized if one starts with them. However, just recalling patterns if they are presented without errors can hardly be called an associative memory. Of course, we would expect that even if a pattern is corrupted by a certain percentage of noise the gradient dynamics is able to retrieve this pattern in spite of the noise. The following theorem shows that also noised patterns can be successfully reconstructed by just applying the retrieval dynamics once (so-called direct convergence; note that for the Hopfield model with unbiased patterns Burshtein [Bu94] obtained that the number of patterns that can be reconstructed by several steps of the retrieval dynamics (non-direct convergence) equals the number of patterns that can be stored such that each of them is a local-minimum).

Theorem 3.2 *Let $r \in [0, \frac{1}{2})$ and for each $\nu = 1, \ldots, M(N)$ let $\tilde{\xi^\nu}$ be an element of the Hamming sphere of radius rN centered at ξ^ν. Assume the random patterns ξ^ν fulfill (1) and $M(N) = (1-2r)^2 \frac{N}{\gamma \log N}$.*
Then for the Hopfield model (2) the following assertions hold true:

1. *If* $\gamma > \frac{3}{8p^2(1-p)^2}$

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

i.e. the noised patterns are almost surely attracted.

2. *If* $\gamma > \frac{1}{4p^2(1-p)^2}$

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

with $\lim\limits_{N\to\infty} R_N = 0$.

3. *If* $\gamma > \frac{1}{8p^2(1-p)2}$ *for every fixed* $\nu = 1, \ldots, M$

$$P(T\tilde{\xi^\nu} = \xi^\nu) = 1 - R_N$$

with $\lim_{N\to\infty} R_N = 0$.

Remark 3.3 *Note that the estimates of the above Theorems for $p = \frac{1}{2}$ (the unbiased case) agree with the results in the standard Hopfield model. It may of course be true that the estimates can be improved in some respects. Note however, that our bound on the storage capacity of the Hopfield model with biased patterns is a decreasing function in the bias of the patterns (which in Theorems 3.1 and 3.2 appears as a factor $\frac{1}{p^2(1-p)^2} \to \infty$ for $p \to 0$ or $p \to 1$) .*

We now give a result on the storage capacity of the Hopfield model with biased patterns provided that Newman's concept of storage is used. It turns out that a bias does not destroy the storage abilities of the Hopfield model and that it can store "extensively many" patterns (i.e. M(N) grows like αN), although the critical α decreases to zero when the bias get s large.

Theorem 3.4 *Suppose that the random patterns fulfill (1). 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 for the standard Hopfield model (2)*

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

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

Note that these results resemble the results of the Hopfield model with semantically correlated patterns obtained in [Lö96].

4 Proofs

In this section we will give the proofs of the theorems stated above.

Proof of Theorem 3.1: Observe that — according to the definition of the dynamics T and the Hopfield model — for any $1 \leq \nu \leq M$ the pattern ξ^ν is stable if and only if

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

for all $i = 1, \ldots, N$. In other words ξ^ν is stable, if

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

for all $i = 1, \ldots, N$ (with the convention $sgn(0) = 1$ which does not influence the calculations since the probability that the above sum equals zero vanishes with N getting large).
Hence for all $t \geq 0$

$$\begin{aligned} & P(\xi^\nu \text{ is not stable }) \qquad (5) \\ \leq\ & NP(\sum_{j=1}^N \sum_{\mu=1}^{M(N)} \xi_1^\nu \xi_j^\nu \overline{\xi_1^\mu}\,\overline{\xi_j^\mu} \leq 0) \end{aligned}$$

$$\leq \; NE\left(\exp(-t(\sum_{j=1}^{N}\sum_{\mu=1}^{M(N)} \xi_1^\nu \xi_j^\nu \overline{\xi_1^\mu}\,\overline{\xi_j^\mu}))\right) \tag{6}$$

$$= \; NE\left(\exp(-t\xi_1^1\xi_2^1\overline{\xi_1^1}\,\overline{\xi_2^1}))\right)^N \left(E\left(\exp(-t\xi_1^1\xi_2^1\overline{\xi_1^2}\,\overline{\xi_2^2}))\right)\right)^{N\,(M(N)-1)}$$

where in the first line we have used the identical distribution of the spins, for the second inequality we have applied the exponential Chebyshev–Markov inequality and in the last line we have made use of the independence of the variables and the identical distribution of the patterns (e.g. by choosing without loss of generality $\nu = 1$).

Let us now estimate the two expectations occuring in the last line of (5). First note that the variable $\xi_1^1\overline{\xi_1^1}$ (as well as $\xi_2^1\overline{\xi_2^1}$) assumes values $2(1-p)$ and $2p$ with probability p and $1-p$, respectively. Thus

$$\xi_1^1\overline{\xi_1^1}\xi_2^1\overline{\xi_2^1} = \begin{cases} 4(1-p)^2 & \text{with probability} \quad p^2 \\ 4p^2 & \text{with probability} \quad (1-p)^2 \\ 4p(1-p) & \text{with probability} \quad 2p(1-p). \end{cases}$$

Hence

$$\begin{aligned} & E\left(\exp(-t\xi_1^1\xi_2^1\overline{\xi_1^1}\,\overline{\xi_2^1}))\right) \\ = \; & p^2\exp(-4t(1-p)^2) + (1-p)^2\exp(-4tp^2) + 2p(1-p)\exp(-4tp(1-p)). \end{aligned}$$

Expanding the exponentials up to the first order gives

$$E\left(\exp(-t\xi_1^1\xi_2^1\overline{\xi_1^1}\,\overline{\xi_2^1}))\right) = 1 - 16tp^2(1-p)^2 + \mathcal{O}(t^2) \leq \exp(-16tp^2(1-p)^2 + \mathcal{O}(t^2)).$$

The second expectation is treated similarly. By observing that

$$\overline{\xi_1^2\xi_2^2} = \begin{cases} 4(1-p)^2 & \text{with probability} \quad p^2 \\ 4p^2 & \text{with probability} \quad (1-p)^2 \\ -4p(1-p) & \text{with probability} \quad 2p(1-p). \end{cases}$$

we obtain

$$\begin{aligned} & E\left(\exp(-t\xi_1^1\xi_2^1\overline{\xi_1^2}\,\overline{\xi_2^2}))\right) = \\ & E(p^2\exp(-4t(1-p)^2\xi_1^1\xi_2^1) + (1-p)^2\exp(-4tp^2\xi_1^1\xi_2^1) + 2p(1-p)\exp(4tp(1-p)\xi_1^1\xi_2^1). \end{aligned}$$

Since $|\xi_1^1\xi_2^1| = 1$ we (again by expanding the exponentials) arrive at

$$E\left(\exp(-t\xi_1^1\xi_2^1\overline{\xi_1^2}\,\overline{\xi_2^2}))\right) \leq \exp(\frac{t^2}{2}16p^2(1-p)^2 + \mathcal{O}(t^3)).$$

Putting these estimates together gives

$$\begin{aligned} & P(\xi^\nu \text{ is not stable }) \\ \leq \; & N\exp(-16N(tp^2(1-p)^2 + \mathcal{O}(t^2)) + 16N(M(N)-1)(\frac{t^2}{2}p^2(1-p)^2 + \mathcal{O}(t^3))). \end{aligned}$$

Choosing $t = \frac{1}{M(M)}$ yields

$$P(\xi^\nu \text{ is not stable }) \leq N \exp\left(-8\frac{N}{M(N)}p^2(1-p)^2 + \mathcal{O}(\frac{N}{M^2}) + \mathcal{O}(\frac{N^2}{M^3})\right).$$

Note that the $\mathcal{O}$-terms in the above expression become negligible for our choice of $M(N)$ which is $M(N) = \frac{N}{\gamma \log N}$ for some positive and finite constant γ.

Finally observe that, if we just want $P(\xi^\nu \text{ is not stable }) \to 0$ for every fixed ν (which is part 3 of Theorem 3.1) it suffices if we choose

$$\gamma > \frac{1}{8p^2(1-p)^2}.$$

For the other two parts observe that the bounds on $P(\xi^\nu \text{ is not stable })$ do not depend on ν and therefore

$$P(\exists \nu : \xi^\nu \text{ is not stable }) \leq M(N)N \exp\left(-8\frac{N}{M(N)}p^2(1-p)^2 + \mathcal{O}(\frac{N}{M^2}) + \mathcal{O}(\frac{N^2}{M^3})\right).$$

Thus $P(\exists \nu : \xi^\nu \text{ is not stable }) \to 0$ for $\gamma \geq \frac{1}{4p^2(1-p)^2}$.
Finally if we choose $\gamma \geq \frac{3}{8p^2(1-p)^2}$ the probabilities for a non-stable ξ^ν become summable and hence the almost sure result (part 1 of Theorem 3.1) follows by a standard Borel–Cantelli argument. □

Proof of Theorem 3.2: This proof is basically a rerun of the arguments leading to the proof of Theorem 3.1. Therefore we will just sketch it.

Following the lines of the proof of Theorem 3.1 we obtain for any $\nu = 1, \ldots, M(N)$ and any $t \geq 0$ that

$$P((T\tilde{\xi^\nu})_i = -\xi_i^\nu) \leq \exp(-tN16p^2(1-p)^2(1-2r) + 8t^2 M(N)Np^2(1-p)^2)$$

(where we have already dropped the $\mathcal{O}(\frac{N}{M^2})$ and $\mathcal{O}(\frac{N^2}{M^3})$ terms).
Hence

$$\begin{aligned}&P(\cup_{\nu=1}^{M(N)}(T\tilde{\xi^\nu})_i = -\xi_i^\nu) \leq \\ &M(N)N\exp(-16tN(1-2r) + 8t^2M(N)Np^2(1-p)^2).\end{aligned}$$

Optimizing in t and applying the Borel–Cantelli Lemma again (if necessary) proves the theorem. □

Proof of Theorem 3.4: In this proof we will again use large deviation estimates (in the sense of upper bounds). The idea leading to these estimates – basically to replace the ξ_i^μ by appropriate Gaussian random variables – is rather standard in the framework of the Hopfield model with independent patterns (see e.g [N88] or [BG92]). However, in our case due to the assymetry of the ξ_i^μ a new estimate is needed. Moreover, the "leading term" is no longer deterministic.

We set

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

Then

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

where ξ^ν_J denotes a configuration differing from ξ^ν exactly in the coordinates J (and for convenience we have chosen δ in such a way that δN is an integer).
Observing that again without loss of generality we may assume that $\nu = 1$ and that

$$
\begin{aligned}
& H_N(\xi^1_J) - H_N(\xi^1) \\
= \; & -\frac{1}{2N} \sum_{\mu=1}^{M(N)} \sum_{i,j=1}^{N} (\xi^1_{J,i}\xi^1_{J,j} - \xi^1_i\xi^1_j))\overline{\xi^\mu_i}\,\overline{\xi^\mu_j} \\
= \; & \frac{1}{N} \sum_{\mu=1}^{M(N)} \sum_{\substack{i\in J, j\notin J \\ i\notin J, j\in J}} \xi^1_i\xi^1_j\overline{\xi^\mu_i}\,\overline{\xi^\mu_j} \\
= \; & \frac{2}{N} \sum_{\mu=1}^{M(N)} \sum_{i\in J, j\notin J} \xi^1_i\xi^1_j\overline{\xi^\mu_i}\,\overline{\xi^\mu_j}
\end{aligned}
$$

we again can estimate the probability of interest with the help of the exponential Chebyshev–Markov inequality. Thus for all $t \geq 0$

$$
\begin{aligned}
& P\left(H_N(\xi^\nu_J) - H_N(\xi^\nu) \leq \varepsilon N\right) \\
= \; & P\left(\frac{1}{N}\sum_{\mu=1}^{M} \sum_{i\in J, j\notin J} \xi^1_i\xi^1_j\overline{\xi^\mu_i}\,\overline{\xi^\mu_j} \leq \varepsilon/2N\right) \\
\leq \; & e^{t\varepsilon N/2} E\left(\exp\left(-\frac{t}{N}\sum_{\mu=1}^{M} \sum_{i\in J, j\notin J} \xi^1_i\xi^1_j\overline{\xi^\mu_i}\,\overline{\xi^\mu_j}\right)\right)
\end{aligned}
$$

and we are again left with estimating the expectation of an exponential.

To this end observe that due to the independence and identical distribution of the spins ξ^μ_i

$$
\begin{aligned}
& E\left(\exp\left(-\frac{t}{N}\sum_{\mu=1}^{M} \sum_{i\in J, j\notin J} \xi^1_i\xi^1_j\overline{\xi^\mu_i}\,\overline{\xi^\mu_j}\right)\right) = \\
& E\left(\exp\left(-\frac{t}{N} \sum_{i\in J, j\notin J} \xi^1_i\xi^1_j\overline{\xi^1_i}\,\overline{\xi^1_j}\right)\right) \times E\left(\exp\left(-\frac{t}{N} \sum_{i\in J, j\notin J} \xi^1_i\xi^1_j\overline{\xi^2_i}\,\overline{\xi^2_j}\right)\right)^{M-1}.
\end{aligned}
$$

To treat the first term on the right hand side above just notice that $\xi_i^1 \overline{\xi_i^1} \geq 2\min\{p,(1-p)\}$ and therefore

$$E\left(\exp\left(-\frac{t}{N}\sum_{i\in J, j\notin J} \xi_i^1\xi_j^1\overline{\xi_i^1}\,\overline{\xi_j^1}\right)\right) \leq \exp(-4\delta(1-\delta)tN(\min\{p,(1-p)\})^2).$$

(Observe that, of course, one could expect $-\frac{t}{N}\sum_{i\in J,j\notin J}\xi_i^1\xi_j^1\overline{\xi_i^1}\,\overline{\xi_j^1}$ to behave like its expectation which is $16t\delta(1-\delta)Np^2(1-p)^2$ rather than $4t\delta(1-\delta)N(\min\{p,(1-p)\})^2$. Such a behavior could again be analyzed by exponential Chebyshev inequalities or by using some sort of concentration of measure techniques as in [T95a]. On the other hand every such estimate has to allow deviations from the mean that are large enough such that the bounds are correct for all the possible $\binom{N}{\delta N}$ choices of I. For most p we then do not gain very much over $4t\delta(1-\delta)N(\min\{p,(1-p)\})^2$ (especially since the deviations become large as δ becomes small and a reasonable value of δ is at most 0.01. Moreover we may notice that for the most interesting cases $p\to 0$, $p\to 1$, or $p=1/2$ the given bounds agree).

For the second expectation we rewrite

$$E\left(\exp\left(-\frac{t}{N}\sum_{i\in J, j\notin J} \xi_i^1\xi_j^1\overline{\xi_i^2}\,\overline{\xi_j^2}\right)\right) = E_J E_{J^c}\left(\exp\left(-\frac{t}{N}\sum_{i\in J}\sum_{j\in J^c} \xi_i^1\xi_j^1\overline{\xi_i^2}\,\overline{\xi_j^2}\right)\right)$$

where for $I\subseteq\{1,\ldots,N\}$ the symbol E_I denotes integration with respect to those ξ_i with $i\in I$. Putting

$$Z := \sum_{i\in J}\xi_i^1\overline{\xi_i^2},$$

using the independence of the coordinates in J^c and that for every i

$$\xi_i^1\overline{\xi_i^2} = \begin{cases} 2(1-p) & \text{with probability } \; p^2 \\ -2(1-p) & \text{with probability } \; p(1-p) \\ 2p & \text{with probability } \; p(1-p) \\ -2p & \text{with probability } \; (1-p)^2. \end{cases}$$

we arrive at

$$\begin{aligned} &E_J E_{J^c}\left(\exp\left(-\frac{t}{N}\sum_{i\in J}\sum_{j\in J^c}\xi_i^1\xi_j^1\overline{\xi_i^2}\,\overline{\xi_j^2}\right)\right) = \\ &E_J\prod_{j\in J^c}E_{\{j\}}\left(\exp\left(-\frac{t}{N}\xi_j^1\overline{\xi_j^2}Z\right)\right) = \qquad (7)\\ &E_J\prod_{j\in J^c}\left[p^2\exp(-2\frac{t}{N}(1-p)Z) + p(1-p)\exp(2\frac{t}{N}(1-p)Z)\right.\\ &\left.+p(1-p)\exp(2\frac{t}{N}pZ) + (1-p)^2\exp(-2\frac{t}{N}pZ)\right]. \end{aligned}$$

Now for any $x\in\mathbb{R}$, and any $1/2\leq p<1$ we have

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

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

$$pe^{-2(1-p)x} + (1-p)e^{2px} \le \cosh((1+2p)x).$$

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

$$pe^{-2(1-p)x} + (1-p)e^{2px} \le \cosh((1+|2p-1|)x).$$

So finally

$$\begin{aligned} & p^2 \exp(-2(1-p)x) + p(1-p)\exp(2(1-p)x) + p(1-p)\exp(2px) + (1-p)^2\exp(-2px) \\ \le\; & \cosh((1+|2p-1|)x). \end{aligned}$$

Applying this inequality to the right hand side of (7) we obtain

$$\begin{aligned} & E_J \prod_{j\in J^c} \Big[p^2 \exp(-2\frac{t}{N}(1-p)Z) + p(1-p)\exp(2\frac{t}{N}(1-p)Z) \\ + \; & p(1-p)\exp(2\frac{t}{N}pZ) + (1-p)^2 \exp(-2\frac{t}{N}pZ) \Big] \\ \le \; & E_J \prod_{j\in J^c} \cosh((1+|2p-1|)\frac{t}{N}Z) \\ \le \; & E_J \prod_{j\in J^c} \exp(\frac{1}{2}(1+|2p-1|)^2 \frac{t^2}{N^2} Z^2) \\ = \; & E_J \prod_{j\in J^c} E_{z_j} \exp(z_j(1+|2p-1|)\frac{t}{N}Z) \end{aligned}$$

where z_j are i.i.d. Gaussian random variables with expectation 0 and variance 1, E_{z_j} denotes integration with respect to z_j, and we have used that

$$\cosh(x) \le e^{\frac{1}{2}x^2}$$

as well as

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

Denoting by E_{z_I} integration with respect to those z_j with $j \in I$ we get that

$$\begin{aligned} & E_J E_{J^c} \left(\exp\left(-\frac{t}{N} \sum_{i\in J}\sum_{j\in J^c} \xi_i^1 \xi_j^1 \overline{\xi_i^2}\, \overline{\xi_j^2} \right)\right) \\ \le \; & E_{z_{J^c}} E_J \exp\left(\frac{t}{N}(1+|2p-1|) \sum_{i\in J}\sum_{j\in J^c} z_j \xi_i^1 \overline{\xi_i^2} \right) \\ = \; & E_{z_J^c} \prod_{i\in J} E_{\{i\}} \left(\exp\left(\frac{t}{N} \xi_i^1 \overline{\xi_i^2} \tilde{Z} \right)\right) \end{aligned}$$

where we have set

$$\tilde{Z} := \sum_{j\in J^c} (1+|2p-1|) z_j.$$

Following the lines above

$$E_{\{i\}}\left(\exp\left(\frac{t}{N}\xi_i^1\overline{\xi_i^2}\tilde{Z}\right)\right) \le E_{z_i}\left(\exp\left(\frac{t}{N}(1+|2p-1|)z_i\tilde{Z}\right)\right)$$

and thus

$$\begin{aligned} & E\left(\exp\left(-\frac{t}{N}\sum_{i\in J, j\notin J}\xi_i^1\xi_j^1\overline{\xi_i^2}\,\overline{\xi_j^2}\right)\right) \\ \le\ & E_z\left(\exp\left(\frac{t}{N}(1+|2p-1|)^2\sum_{i\in J, j\notin J} z_i z_j\right)\right) \\ =\ & E_z\left(\exp\left(t(1+|2p-1|)^2\sqrt{\delta(1-\delta)}\sum_{i\in J, j\notin J}\frac{z_i}{\sqrt{\delta N}}\frac{z_j}{\sqrt{(1-\delta)N}}\right)\right). \end{aligned}$$

The rest of the proof now could be carried out similar to [N88], proof of Theorem 3.1. We just finish the proof here to keep the paper self-contained. So, note that the expectation on the right hand side above is an expectation of a (very simple) quadratic form of the Gaussian random vector (z', z'') where

$$z' := \sum_{i\in J}\frac{z_i}{\sqrt{\delta N}}$$

and

$$z'' := \sum_{j\in J^c}\frac{z_j}{\sqrt{(1-\delta)N}}.$$

Thus the expectation is finite only if t is small enough ($t \le \frac{1}{\sqrt{\delta(1-\delta)}(1+|2p-1|)^2}$) and it that case

$$\begin{aligned} & E_z\left(\exp\left(t(1+|2p-1|)^2\sqrt{\delta(1-\delta)}\sum_{i\in J, j\notin J}\frac{z_i}{\sqrt{\delta N}}\frac{z_j}{\sqrt{(1-\delta)N}}\right)\right) \\ =\ & \frac{1}{\sqrt{1-t^2\delta(1-\delta)(1+|2p-1|)^4}}. \end{aligned}$$

Thus putting things together and setting $M = \alpha N$ we arrive at

$$\begin{aligned} & P\left(H_N(\xi_J^\nu) - H_N(\xi^\nu) \le \varepsilon N\right) \\ \le\ & \inf_{t^*\ge t\ge 0}\exp\left(Nt\frac{\varepsilon}{2} - 4Nt\delta(1-\delta)\min\{p^2,(1-p)^2\} - \log\left(1-t^2\delta(1-\delta)(1+|2p-1|)^4\right)\frac{\alpha}{2}N\right) \end{aligned}$$

where $t^* = \frac{1}{(1+|2p-1|)^2\sqrt{\delta(1-\delta)}}$.
Finally by Stirling's formula and the above estimate

$$\sum_{J:|J|=\delta N}\sum_{\nu=1}^{M(N)} P\left(H_N(\xi_J^\nu) - H_N(\xi^\nu) \le \varepsilon N\right)$$

$$\begin{aligned}
&\le M(N)\binom{N}{\delta N}\inf_{t^*\ge t\ge 0}\exp\left(Nt\frac{\varepsilon}{2}-4Nt\delta(1-\delta)\min\{p^2,(1-p)^2\}\right)\times\\
&\times \exp\left(-\log\left(1-t^2\delta(1-\delta)(1+|2p-1|)^4\right)\frac{\alpha}{2}N\right)\\
&\le \alpha N\inf_{t^*\ge t\ge 0}\exp((-\delta\log\delta-(1-\delta)\log(1-\delta))N)\times\\
&\times \exp\left(Nt\frac{\varepsilon}{2}-4Nt\delta(1-\delta)\min\{p^2,(1-p)^2\}-\log\left(1-t^2\delta(1-\delta)(1+|2p-1|)^4\right)\frac{\alpha}{2}N\right)
\end{aligned}$$

and we have to find an admissible t (i.e. $0\le t\le t^*$) and values of δ and α such that the above exponent becomes negative. This can be seen by noticing that for any $t\ge 0$ with $t^2\delta(1-\delta)(1+|2p-1|)^4\le 3/4$

$$\frac{1}{\sqrt{1-t^2\delta(1-\delta)(1+|2p-1|)^4}}\le \exp\left(t^2\delta(1-\delta)(1+|2p-1|)^4\right).$$

and hence the above exponent can be estimated by

$$\begin{aligned}
&(-\delta\log\delta-(1-\delta)\log(1-\delta))N+t\frac{\varepsilon}{2}N-\\
-\;&4Nt\delta(1-\delta)\min\{p^2,(1-p)^2\}-\log(1-t^2\delta(1-\delta)(1+|2p-1|)^4)\frac{\alpha}{2}N\le\\
\le\;&(-\delta\log\delta-(1-\delta)\log(1-\delta))N+t\frac{\varepsilon}{2}N-\\
-\;&4Nt\delta(1-\delta)\min\{p^2,(1-p)^2\}+t^2\delta(1-\delta)(1+|2p-1|)^4)\alpha N
\end{aligned}$$

if $t\le t^{**}:=\sqrt{\frac{3}{4\delta(1-\delta)(1+|2p-1|)^4}}$. Choosing ε sufficiently small the exponent is minimized by a t which is close to

$$t_{\min}=\frac{1}{\alpha}\frac{2\min\{p^2,(1-p)^2\}}{(1+|2p-1|)^4}$$

Observe that $t_{\min}\le t^{**}$ if

$$\alpha\ge\sqrt{\delta(1-\delta)}\frac{4\min\{p^2,(1-p)^2\}}{\sqrt{3}(1+|2p-1|)^2}. \tag{8}$$

On the other hand inserting $t_{\min}$ into the the above exponent and choosing ε sufficiently small gives

$$\begin{aligned}
&(-\delta\log\delta-(1-\delta)\log(1-\delta))N+t_{\min}\frac{\varepsilon}{2}N-4t_{\min}N\delta(1-\delta)\min\{p^2,(1-p)^2\}\\
+\;&t_{\min}^2\delta(1-\delta)(1+|2p-1|)^4\alpha N \qquad (9)\\
\le\;&(-\delta\log\delta-(1-\delta)\log(1-\delta))N-\gamma N\delta(1-\delta)\frac{4\min\{p^4,(1-p)^4\}}{(1-|2p-1|)^4}\frac{1}{\alpha}
\end{aligned}$$

with $\gamma<1$ and close to 1 (as ε becomes small). The right hand side of this inequality becomes negative when δ and α become small appropriately. To check whether this can be done in agreement with (8) we insert

$$\alpha=\sqrt{\delta(1-\delta)}\frac{4\min\{p^2,(1-p)^2\}}{\sqrt{3}(1+|2p-1|)^2}$$

into the right hand side of (9) and obtain

$$\left(-\frac{\sqrt{3}\gamma}{(1+|2p-1|)^2}\min\{p^2,(1-p)^2\}\sqrt{\delta(1-\delta)}-\delta\log\delta-(1-\delta)\log(1-\delta)\right)N. \tag{10}$$

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 and e.g. α as the right hand side of (8). This completes the proof of the theorem. □

Remark 4.1 *Observe that the bound on the moment generating function in (10) as well as the bound on* α *in (9) depends on p mainly via the factor* $\min\{p^2,(1-p)^2\}$ *(the other terms containing p are bounded from above and away from 0) which converges to zero for p close to one or close to zero and therefore can only deteriorate the bounds for* α *(allowing smaller* α'*s only) for a large bias. Although we do not claim to have rigorously proven that the storage capacity decreases with a large bias (indeed such a result may be quite difficult to obtain, since it would requires e.g. some uniformity statement in the estimates) we just emphasize that Theorem 3.4 is in good agreement with the picture of Theorem 3.1 and 3.2 where the other notion of storage is used.*

References

[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 Prob. Th. Rel. Fields (1996)

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

[BG96] A.Bovier, V.Gayrard; Hopfield models as a generalized eman 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)

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

[FI91] M.V. Feigel'man, L.B. Ioffe, Hierarchical Organiztion of Memory; in "Models of neural networks", E. Domany, L. v. Hemmen, R. Schulte (eds.), Physics of Neural Networks, Springer, Berlin (1991)

[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)

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

[L94] D. Loukianova; Capacité de mémoire dans le modèle de Hopfield; C.R. Acad. Sci. Paris 318, 157-160 (1994)

[Lö96] M. Löwe; On the storage capacity of Hopfield models with weakly correlated 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)

[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)

[T95a] M. Talagrand; Concentration of measure and isoperimetric inequalities in product spaces; Publ. Math. I.H.E.S. 81, 73-205 (1995)

[T95b] 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)