J Stat Phys (2017) 168:288–299
DOI 10.1007/s10955-017-1806-y

On a Model of Associative Memory with Huge Storage Capacity

Mete Demircigil[1] · Judith Heusel[2] · Matthias Löwe[2] · Sven Upgang[2] · Franck Vermet[3]

Received: 7 March 2017 / Accepted: 2 May 2017 / Published online: 19 May 2017

Abstract In Krotov et al. (in: Lee (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., Red Hook, 2016) Krotov and Hopfield suggest a generalized version of the well-known Hopfield model of associative memory. In their version they consider a polynomial interaction function and claim that this increases the storage capacity of the model. We prove this claim and take the "limit" as the degree of the polynomial becomes infinite, i.e. an exponential interaction function. With this interaction we prove that model has an exponential storage capacity in the number of neurons, yet the basins of attraction are almost as large as in the standard Hopfield model.

Keywords Neural networks · Associative memory · Hopfield model · Exponential inequalities

✉ Franck Vermet
Franck.Vermet@univ-brest.fr

Mete Demircigil
mete.demircigil@ens.fr

Judith Heusel
judith.heusel@uni-muenster.de

Matthias Löwe
maloewe@uni-muenster.de

Sven Upgang
s.upgang@uni-muenster.de

1 Département de Mathématiques et Applications, Ecole Normale Supérieure, 45 rue d'Ulm, 75005 Paris, France

2 Fachbereich Mathematik und Informatik, University of Münster, Einsteinstraße 62, 48149 Münster, Germany

3 Laboratoire de Mathématiques de Bretagne Atlantique, UMR CNRS 6205, Université de Bretagne Occidentale, 6, avenue Victor Le Gorgeu, CS 93837, 29238 Brest Cedex 3, France

1 Introduction

Neural networks and associative memories have been a highly active research area in computer science, physics and probability theory for more than thirty years. The standard model of an associative memory was developed in the seminal paper [6]. His model is based on N neurons, each of which can only take the values ± 1. Each pair of these neurons is connected and thus interacts. We want to store a set of input data $(\xi^\mu)_{\mu=1}^M$, so called patterns or images, where M may and will depend on the system size N, in the model. Each of the patterns ξ^μ is itself a bit string of length N, hence $\xi^\mu = (\xi_i^\mu)_{i=1}^N$ where $\xi_i^\mu \in \{-1, +1\}$ for all i and μ. The strength at which two neurons i and j interact depends on the images and is given by the so-called synaptic efficacy

$$J_{ij} = \sum_{\mu=1}^{M} \xi_i^\mu \xi_j^\mu .$$

With this set of (J_{ij})'s we associate a dynamics or updating rule $T = (T_i)_{i=1}^N$ on $\{-1, +1\}^N$ such that

$$T_i(\sigma) := \operatorname{sgn}\left(\sum_{j=1}^{N} J_{ij}\sigma_j\right) \qquad \sigma = (\sigma_i) \in \{-1, +1\}^N$$

and the indices i are either updated uniformly at random or in a given order. One of the patterns (ξ^μ) is considered to be stored, if and only if it is stable under the (retrieval) dynamics T, i.e. if and only if $T_i(\xi^\mu) = \xi_i^\mu$ for all $i = 1, \ldots, N$.

The central question is now: How many patterns can be stored in the above model? Of course, this sensitively depends on the way we choose these patterns. Much in agreement with the choice of messages in information theory in most of the test scenarios for associative memories the patterns are chosen independent and identically distributed (even though, other choices may be considered as well, see e.g. [10], [9], [11] or [12]). More precisely, it is assumed that

$$\mathbb{P}(\xi_i^\mu = 1) = \mathbb{P}(\xi_i^\mu = -1) = \frac{1}{2} \quad \text{for all } i \text{ and } \mu$$

and that all $(\xi_i^\mu)_{i,\mu}$ are i.i.d. Under these assumptions it was shown in [13] that the Hopfield model described above can store $M = C\frac{N}{\log N}$ patterns with $C < \frac{1}{2}$, if we require that a fixed (but arbitrary) pattern is stable, and $C < \frac{1}{4}$ if we ask for stability of all patterns simultaneously (also see [3] for a matching upper bound). However, non-rigorous computations involving the so-called replica trick, suggest that we may even achieve a capacity of $M = \alpha N$ if we allow for a small fraction of errors in the retrieval and $\alpha < 0.138$ (see [1], [2]). This prediction was mathematically confirmed (yet with smaller values of α) in [14], [8], and [15].

In a very recent contribution Krotov and Hopfield suggest a generalized version of the Hopfield model (see [7]). There, the authors replace the updating dynamics $T = (T_i)$ by a more general, still asynchronous one, namely:

$$T_i(\sigma) := \operatorname{sgn}\Big[\sum_{\mu=1}^{M}\Big(F\Big(1 \cdot \xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma_j\Big) - F\Big((-1)\cdot\xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma_j\Big)\Big)\Big] \tag{1}$$

where $F : \mathbb{R} \to \mathbb{R}$ is some smooth function. The case of $F(x) = x^2$ reduces to the standard Hopfield model with the quadratic crosstalk-terms introduced above. Indeed, in this case the

argument in the sign-function is given by the difference

$$\begin{aligned}
&\sum_{\mu=1}^{M} F\big(1\cdot\xi_i^\mu+\sum_{j\neq i}\xi_j^\mu\sigma_j\big)-F\big((-1)\cdot\xi_i^\mu+\sum_{j\neq i}\xi_j^\mu\sigma_j\big)\\
&\quad=\sum_{\mu=1}^{M}1+2\sum_{j\neq i}\xi_i^\mu\xi_j^\mu\sigma_j+\left(\sum_{j\neq i}\xi_j^\mu\sigma_j\right)^2-1+2\sum_{j\neq i}\xi_i^\mu\xi_j^\mu\sigma_j-\left(\sum_{j\neq i}\xi_j^\mu\sigma_j\right)^2\\
&\quad=4\sum_{\mu=1}^{M}\sum_{j\neq i}\xi_i^\mu\xi_j^\mu\sigma_j
\end{aligned}$$

and its sign is of course the same as that of $\sum_{\mu=1}^{M}\sum_{j\neq i}\xi_i^\mu\xi_j^\mu\sigma_j$, therefore the dynamics are equal.

The reason for this more general choice of the interaction function F is the following insight: in the standard Hopfield model there is an energy associated with the dynamics T (i.e. the energy of a configuration decreases by an application of T). This energy of a spin configuration σ is given by $H(\sigma)=-\frac{1}{N}\sum_\mu\sum_{i,j}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j$ and it decreases "too slowly" as the configuration σ approaches pattern to allow for a superlinear storage capacity

The authors in [7] therefore in particular analyze what they call "polynomial interaction" (or, as they put it, energy) functions, i.e. $F(x)=x^n$. Krotov and Hopfield state the following assertion

Theorem 1 ([7], formulas (5) and (6))

1. *In the generalized Hopfield model with interaction function $F(x)=x^n$ one can store up to $M=\alpha_n N^{n-1}$ patterns, if small errors in the retrieval are tolerated.*
2. *In the same model, one can store $M=\frac{N^{n-1}}{c_n\log N}$ patterns for $c_n>2\,(2n-3)!!$, if one wants a fixed pattern to be a fixed point of the dynamics T introduced in (1) with a probability converging to 1.*

A proof of this theorem could probably be rather involved. This is due to the fact that the energy function of the model described in Theorem 1 is of a polynomial form. As a matter of fact, up to normalization, the energy of the model in these cases is $H(\sigma)=\sum_\mu P(m^\mu)$, where $m^\mu:=\sum_i\xi_i^\mu\sigma_i$ is the overlap of the configuration σ with the μ'th pattern and P is a polynomial, or, in other words, with $F(x)=x^n$ and n even

$$F\left(1\cdot\xi_i^\mu+\sum_{j\neq i}\xi_j^\mu\sigma_j\right)-F\left((-1)\cdot\xi_i^\mu+\sum_{j\neq i}\xi_j^\mu\sigma_j\right)$$

consists of many summands of the form $\xi_i^\mu\left(\sum_{j\neq i}\xi_j^\mu\sigma_j\right)^m$ where m is an even number smaller than n. There is, however, a closely related model, where the above statement can be proven. To this end consider the dynamics $\widehat{T}=(\widehat{T}_i)$ on $\{-1,+1\}^N$ defined by

$$\widehat{T}_i(\sigma):=\operatorname{sgn}\left(\sum_{1=j_1,\dots j_{n-1}}^{N}\sigma_{j_1}\cdots\sigma_{j_{n-1}}W_{i,j_1\dots j_{n-1}}\right). \tag{2}$$

with

$$W_{i_1,\ldots,i_n} = \frac{1}{N^{n-1}} \sum_{\mu=1}^{M} \xi_{i_1}^{\mu} \xi_{i_2}^{\mu} \cdots \xi_{i_n}^{\mu}. \tag{3}$$

Then the following theorem can be shown

Theorem 2 *1. In the generalized Hopfield model with dynamics $\widehat{T}$ defined in (2) and (3) one can store up to $M = \alpha_n N^{n-1}$ patterns, if small errors in the retrieval are tolerated.*

2. In the same model, one can store $M = \frac{N^{n-1}}{c_n \log N}$ patterns for $c_n > 2\,(2n-3)!!$, if one wants a fixed pattern to be a fixed point of the dynamics $\widehat{T}$ with a probability converging to 1.

While part 1 of the above theorem was proved in [14], we will give a proof of the second part (including a computation of the constants c_n) in Sect. 2 below. Note that the thermodynamics of this model was analyzed in [4].

More interesting than these polynomial models is, however, the question what happens if we formally send n to infinity. One could conjecture from above that this would lead to an interaction function of the form e^x, on the one hand and to a super-polynomial storage capacity, on the other. This is indeed what we will show. Actually, we will even prove slightly more: In general, one could imagine that an increase in capacity goes to expense of associativity, such that in the extreme case, one could store 2^N patterns but none of them has a positive radius of attraction. We will show that this is not the case for our model: The dynamics is even able to repair an amount of random errors of order N.

Theorem 3 *Consider the generalized Hopfield model with the dynamics described in (1) and interaction function F given by $F(x) = e^x$. For a fixed $0 < \alpha < \log(2)/2$ let $M = \exp(\alpha N) + 1$ and let $\xi^1, \ldots, \xi^M$ be M patterns chosen uniformly at random from $\{-1, +1\}^N$. Moreover fix $\varrho \in [0, 1/2)$. For any μ and any $\widetilde{\xi}^{\mu}$ taken uniformly at random from the Hamming sphere with radius ϱN centered in ξ^{μ}, $\mathcal{S}(\xi^{\mu}, \varrho N)$, where ϱN is assumed to be an integer, it holds that*

$$\mathbb{P}\left(\exists \mu\ \exists i : T_i\left(\widetilde{\xi}^{\mu}\right) \neq \xi_i^{\mu}\right) \to 0,$$

if α is chosen in dependence of ϱ such that

$$\alpha < \frac{I(1-2\varrho)}{2}$$

with $I : x \mapsto \frac{1}{2}\left((1+x)\log(1+x) + (1-x)\log(1-x)\right)$.

Remark 1 Note that Theorem 3 in particular implies that

$$\mathbb{P}\left(\exists \mu\ \exists i : T_i\left(\xi^{\mu}\right) \neq \xi_i^{\mu}\right) \to 0$$

as $N \to \infty$, i.e. with a probability converging to 1, all the patterns are fixed points of the dynamics. In this case we can even take $\alpha < \frac{I(1)}{2}$.

Remark 2 Theorem 3 can be proven analogously if the configuration $\widetilde{\xi}^{\mu}$ is drawn uniformly at random from a Hamming ball $B(\xi^{\mu}, \rho N)$. Indeed, the probability of correcting an arbitrary pattern of the sphere $S(\xi^{\mu}, \rho N)$ can be used as a bound for the probability of correcting an arbitrary pattern of lower spheres.

Theorem 2 and Theorem 3 will be proven in the following section.

2 Proofs

We start with the proof of Theorem 2.

Proof of Theorem 2 As already mentioned the first part of the Theorem has already been proven in [14].

We are interested in bounding the following probability

$$\mathbb{P}(\exists i \leq N : \widehat{T}_i(\xi^1) \neq \xi_i^1) = \mathbb{P}\Big(\exists i \leq N : -\sum_{\mu=2}^{M} \xi_i^1 \xi_i^\mu \Big(\sum_j \xi_j^1 \xi_j^\mu\Big)^{n-1} > N^{n-1}\Big).$$

With the exponential Chebyshev inequality and the independence of the random variables (ξ_i^μ) for different μ it follows that

$$\mathbb{P}(\exists i \leq N : \widehat{T}_i(\xi^1) \neq \xi_i^1) \leq N e^{-tN^{n-1}} \mathbb{E}\Big[\exp\big(-t\,\xi_i^1 \xi_i^2 \big(\sum_j \xi_j^1 \xi_j^2\big)^{n-1}\big)\Big]^{M-1}.$$

For the moment generating function we condition on the values of $\xi_i^1 \xi_i^2$ to get the upper bound

$$\mathbb{P}(\exists i \leq N : \widehat{T}_i(\xi^1) \neq \xi_i^1) \leq N e^{-tN^{n-1}} \mathbb{E}\left[\cosh\left(t\left(\sum_j \xi_j^1 \xi_j^2\right)^{n-1}\right)\right]^{M-1}.$$

The random variables $(\xi_j^1 \xi_j^2)_j$ are i.i.d. and distributed like ξ_1^1. Define $m = \frac{1}{\sqrt{N}} \sum_j \xi_j^1 \xi_j^2$ and write the expectation as the sum over all possible values x of m:

$$\mathbb{E}\left[\cosh\left(t\left(\sum_j \xi_j^1 \xi_j^2\right)^{n-1}\right)\right] = \sum_x \cosh\left(tN^{\frac{n-1}{2}} x^{n-1}\right) \cdot \mathbb{P}(m = x).$$

The sum is over all $x \in \{0, \pm\frac{1}{\sqrt{N}}, \ldots, \pm\sqrt{N}\}$. First we want to eliminate the tail events and use the fact that the probability vanishes fast enough if we restrict x away from zero. To this end we fix $\beta > \frac{1}{2}$ and split the sum at $\log(N)^\beta$. Additionally observe that x cannot grow faster than $\sqrt{N}$ and set $t = a_n/M$ for $a_n > 0$ independent of N. Thus

$$\begin{aligned}
&\sum_{x:\log(N)^\beta < |x| \leq \sqrt{N}} \cosh\left(tN^{\frac{n-1}{2}} x^{n-1}\right) \cdot \mathbb{P}(m = x) \\
&\quad \leq 2\cosh\left(tN^{n-1}\right) \mathbb{P}(m > \log(N)^\beta) \leq 2\,\exp\left(tN^{n-1}\right) \exp\left(-\frac{1}{2}\log(N)^{2\beta}\right) \\
&\quad = 2\,\exp\left(\left[a_n c_n - \frac{1}{2}\log(N)^{2\beta-1}\right]\log(N)\right).
\end{aligned}$$

Here we used a standard large deviations estimate and $\cosh(z) \leq \exp(|z|)$. This part of the moment generating function converges to zero for $N \to \infty$ because for N large enough the term in brackets can be bounded by a negative value.

For the critical values of m we use the inequality $\cosh(z) \leq \exp(\frac{z^2}{2})$ for all z and write the exponential function in its Taylor expansion:

$$\sum_{x:|x|\leq\log(N)^\beta} \cosh\left(tN^{\frac{n-1}{2}}x^{n-1}\right)\mathbb{P}(m=x) \leq \sum_{x:|x|\leq\log(N)^\beta} e^{\frac{t^2}{2}N^{n-1}x^{2(n-1)}}\mathbb{P}(m=x)$$
$$= \sum_{x:|x|\leq\log(N)^\beta}\left(1+\frac{t^2}{2}N^{n-1}x^{2(n-1)}+\sum_{k=2}^{\infty}\frac{1}{2^k}\frac{(t^2N^{n-1}x^{2(n-1)})^k}{k!}\right)\mathbb{P}(m=x).$$

The distribution of m converges to a standard normal distribution and its moments are bounded by the moments of the latter. For $l \in \mathbb{N}$ let κ_{2l} be the $2l$-th moment of a standard normal distribution. The sum of probabilities can be bounded by one. So for N large enough, $t = a_n/M$, and $M = \text{const.}\frac{N^{n-1}}{\log N}$ we derive the following upper bound for the moment generating function:

$$\begin{aligned}
&\sum_{x:|x|\leq\log(N)^\beta} \cosh\left(tN^{\frac{n-1}{2}}x^{n-1}\right)\mathbb{P}(m=x)\\
&\leq 1+\frac{t^2}{2}N^{n-1}\kappa_{2(n-1)} + \sum_{x:\,|x|\leq\log(N)^\beta}\sum_{k=2}^{\infty}\frac{1}{2^k}\frac{\left(t^2N^{n-1}x^{2(n-1)}\right)^k}{k!}\cdot\mathbb{P}(m=x)\\
&\leq 1+\frac{t^2}{2}N^{n-1}\kappa_{2(n-1)} + \sum_{k=2}^{\infty}\frac{1}{2^k}\frac{(t^2N^{n-1}\log(N)^{2\beta(n-1)})^k}{k!}\\
&\leq 1+\frac{t^2}{2}N^{n-1}\kappa_{2(n-1)} + \frac{t^4}{4}N^{2n-2}\log(N)^{4\beta(n-1)}(e-2)\\
&\leq \exp\left(\frac{t^2}{2}N^{n-1}\kappa_{2(n-1)} + t^4N^{2(n-1)}\log(N)^{4\beta(n-1)}\right)
\end{aligned}$$

Inserting $t = a_n/M$ into this result we obtain

$$\begin{aligned}
&\mathbb{P}(\exists i \leq N:\ \widehat{T}_i(\xi^1) \neq \xi_i^1)\\
&\leq \exp\left(\log(N) - tN^{n-1}\right)\exp\left(\frac{t^2}{2}N^{n-1}\kappa_{2(n-1)}M + t^4N^{2(n-1)}\log(N)^{4\beta(n-1)}M\right)\\
&= \exp\left(\left[1-a_nc_n\left(1-\frac{a_n\,\kappa_{2(n-1)}}{2}\right)\right]\log(N)\right)\cdot(1+o(1)).
\end{aligned}$$

for our choice of t and M.

The moments of the standard normal distribution is given by $\kappa_{2(n-1)} = (2n-3)!!$ for all $n \in \mathbb{N}$. Choose $a_n = (\kappa_{2(n-1)})^{-1}$. The term in brackets can be bounded by negative value if c_n satisfies

$$1-a_nc_n\left(1-\frac{a_n\,\kappa_{2(n-1)}}{2}\right) < 0$$

which is the case if and only if $c_n > 2(2n-3)!!$. This is exactly the memory capacity proposed by Hopfield and Krotov in Theorem 1.

We continue with the proof of our central result.

Proof of Theorem 3 We start by slightly reformulating the dynamics of the model. Indeed, an (almost) equivalent formulation is to say that a neuron i will remain unchanged after an application of the update rule if

$$\Delta_i E(\sigma) := \sum_{\mu=1}^{M} \left(F\left(\sigma_i \xi_i^\mu + \sum_{j \neq i} \xi_j^\mu \sigma_j \right) - F\left(-\sigma_i \xi_i^\mu + \sum_{j \neq i} \xi_j^\mu \sigma_j \right) \right) > 0 \qquad (4)$$

and the spin of neuron i will be changed if $\Delta_i E(\sigma)$ is less than zero. In the limit $N \to \infty$ the case $\Delta_i E(\sigma) = 0$ is negligible for the later purposes.

Starting in one of the patterns (without loss of generality the first one ξ^1) we want to show that it is an attractive fixed point of the update dynamics, i.e. we need to show that $T_i(\xi^\mu) = \xi_i^\mu$ and we want the model to correct up to ϱN random errors by updating each of the neurons once. Without loss of generality we focus on the pattern ξ^1, taking a corrupted version $\tilde{\xi}^1$ uniformly at random from $S(\xi^1, \rho N)$ for a fixed $\rho \in [0, \frac{1}{2})$, and the neuron i. Then we can interpret the summand for $\mu = 1$ in (4):

$$E_{\text{signal}} = F\left(\sum_{j=1}^{N} \xi_j^1 \widetilde{\xi}_j^1 \right) - F\left(-2\, \xi_i^1 \widetilde{\xi}_i^1 + \sum_{j=1}^{N} \xi_j^1 \widetilde{\xi}_j^1 \right)$$

as signal term and the rest of the sum in (4):

$$E_{\text{noise}} = \sum_{\mu=2}^{M} \left(F\left(\sum_{j=1}^{N} \xi_j^\mu \widetilde{\xi}_j \right) - F\left(-2 \widetilde{\xi}_i \xi_i^\mu + \sum_{j=1}^{N} \xi_j^\mu \widetilde{\xi}_j \right) \right)$$

as noise term. As we will show, in order to have neuron i not updated correctly, the noise term needs to have a bigger impact in (4) than the signal term. We want to show that the probability for this event vanishes for $N \to \infty$.

We need to distinguish two cases: first the neuron i can be correct, i.e. $\xi_i^1 = \widetilde{\xi}_i^1$, and we want the associative memory not to change this value (this means $\Delta_i E(\widetilde{\xi}^1) > 0$). In the other case the neuron is wrong, i.e. $\xi_i^1 = -\widetilde{\xi}_i^1$, and the network needs to correct this neuron (this means $\Delta_i E(\widetilde{\xi}^1) < 0$). In both cases the signal term supports the desired behavior. Indeed, we have:

$$F\left(\sum_{j=1}^{N} \xi_j^1 \widetilde{\xi}_j^1 \right) - F\left(\sum_{j=1}^{N} \xi_j^1 \widetilde{\xi}_j^1 - 2 \right) > 0 \text{ on the one hand, and}$$

$$F\left(\sum_{j=1}^{N} \xi_j^1 \widetilde{\xi}_j^1 \right) - F\left(\sum_{j=1}^{N} \xi_j^1 \widetilde{\xi}_j^1 + 2 \right) < 0 \text{ on the other}$$

depending on whether neuron i is correct or incorrect (since $\sum_{j=1}^{N} \xi_j^1 \widetilde{\xi}_j^1 \geq (1 - 2\varrho)N > 2$ and $F = \exp$). This means that in order for neuron i to update correctly, it must be that $\operatorname{sgn}\left(E_{\text{signal}} + E_{\text{noise}}\right) = \operatorname{sgn}\left(E_{\text{signal}}\right)$, which is fulfilled as soon as $|E_{\text{noise}}| < |E_{\text{signal}}|$. Thus, a necessary condition for the event $\{T_i(\widetilde{\xi}^1) \neq \xi_i^1\}$ is that $|E_{\text{noise}}| \geq |E_{\text{signal}}|$. Therefore:

$$\begin{aligned} \mathbb{P}\left(\exists \mu\ \exists i : T_i\big(\widetilde{\xi}^\mu\big) \neq \xi_i^\mu\right) &\leq N \cdot M \cdot \mathbb{P}\big(T_i(\widetilde{\xi}^1) \neq \xi_i^1\big) \\ &\leq N \cdot M \cdot \mathbb{P}\big(|E_{\text{noise}}| \geq |E_{\text{signal}}|\big). \end{aligned}$$

By using the straightforward fact that $|e^{a\pm 1} - e^{a\mp 1}| \leq [1 - e^{-2}]e^2 \cdot e^a$:

$$|E_{\text{noise}}| \leq \sum_{\mu=2}^{M} \left| \exp(\xi_i^\mu \xi_i^1 + \sum_{j\neq i} \xi_j^\mu \widetilde{\xi}_j^1) - \exp(-\xi_i^\mu \xi_i^1 + \sum_{j\neq i} \xi_j^\mu \widetilde{\xi}_j^1) \right|$$
$$\leq [1 - e^{-2}]e^2 \sum_{\mu=2}^{M} \exp\left(\langle \xi^\mu | \widetilde{\xi}^1 \rangle\right)$$

where $\langle x|y\rangle$ is the inner product on $\{\pm 1\}^N$. At the same time

$$|E_{\text{signal}}| > e^{N(1-2\varrho)}[1 - e^{-2}].$$

It follows that:

$$\mathbb{P}\left(\exists \mu\ \exists i : T_i\left(\widetilde{\xi}^\mu\right) \neq \xi_i^\mu\right) \leq N \cdot M \cdot \mathbb{P}\left(T_i(\widetilde{\xi}^1) \neq \xi_i^1\right)$$
$$\leq N \cdot M \cdot \mathbb{P}\left(\sum_{\mu=2}^{M} e^2 \exp(\langle \xi^\mu | \widetilde{\xi}^1 \rangle) > e^{N(1-2\varrho)}\right). \tag{5}$$

We will use two standard estimates from the theory of large deviations [5]: for a Binomially distributed random variable $\bar{S}_{m,p}$ with parameters m and p and for $\varepsilon > 0$, we have:

$$\mathbb{P}\left(\bar{S}_{m,p} \geq m(p+\varepsilon)\right) \leq \exp\left(-m\frac{\varepsilon^2}{2(p+\varepsilon)}\right) \tag{6}$$

and for a sum S_m of m i.i.d. random variables X_i with $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = \frac{1}{2}$ and $x \in (0, 1)$ we have

$$\mathbb{P}\left(S_m \geq mx\right) \leq \exp\left(-mI(x)\right) \tag{7}$$

as well as

$$\lim_{m\to\infty} \frac{1}{m} \log(\mathbb{P}\left(S_m > mx\right)) = -I(x) \tag{8}$$

where again $I(x) = \frac{1}{2}\left((1+x)\log(1+x) + (1-x)\log(1-x)\right)$. In fact, (8) is nothing but Cramérs theorem for fair, ± 1-Bernoullis.

Now let $\alpha < \frac{1}{2}I(1-2\varrho)$, $M = \exp(\alpha N) + 1$ and β_o be such that $I(\beta_o) = \alpha$. By continuity of I there exists an $\varepsilon > 0$, such that for all $x \in (1-2\varrho-\varepsilon, 1-2\varrho]$ we have that $\alpha < \frac{1}{2}I(x) \leq \frac{1}{2}I(1-2\varrho)$. Again by continuity of I we can choose $\beta < \beta_o$ such that:

$$\alpha - \frac{\varepsilon}{2} = I(\beta_o) - \frac{\varepsilon}{2} < I(\beta) < I(\beta_o) = \alpha.$$

Let us define

$$A = \{\mu \in \{2 \ldots M\} | \langle \xi^\mu | \widetilde{\xi}^1 \rangle \geq \beta N\},$$
$$p = \mathbb{P}\left(\langle \xi^2 | \widetilde{\xi}^1 \rangle \geq \beta N\right).$$

By (7), we have that: $p < \exp(-NI(\beta))$. On the other hand, by (8) we can conclude that for $\eta = \frac{1}{2}(\alpha - I(\beta)) > 0$ and N sufficiently large we have $p > \exp(-N(I(\beta) + \eta))$.

Now we compute the probability in (5) by picking those patterns ξ^μ with a significant overlap with ξ^1 (*i.e.* $\mu \in A$).

$$\mathbb{P}\left(\sum_{\mu=2}^{M} e^2 \exp(\langle \xi^\mu | \widetilde{\xi}^1 \rangle) > e^{N(1-2\varrho)}\right)$$
$$= \sum_{X \subset \{2 \dots M\}} \mathbb{P}\left(\left(\sum_{\mu=2}^{M} \exp(\langle \xi^\mu | \widetilde{\xi}^1 \rangle) > e^{N(1-2\varrho)-2}\right) \cap (A = X)\right)$$
$$\leq \sum_{k=0}^{M-1} \sum_{X \in P_k(\{2 \dots M\})} p^k (1-p)^{M-1-k} \cdot$$
$$\mathbb{P}\left(\sum_{\mu \in X} \exp(\langle \xi^\mu | \widetilde{\xi}^1 \rangle) + (M-1-k)e^{\beta N} > e^{N(1-2\varrho)-2} | A = X\right)$$

where P_k denotes the subsets of size k. Additionally we used that every overlap in A^c can be bounded by $\exp(\beta N)$. By the identical distribution of the patterns this is equal to

$$= \sum_{k=0}^{M-1} \binom{M-1}{k} p^k (1-p)^{M-1-k} \cdot$$
$$\mathbb{P}\left(\sum_{\mu=2}^{k+1} \exp(\langle \xi^\mu | \widetilde{\xi}^1 \rangle) > e^{N(1-2\varrho)-2} - (M-1-k)e^{\beta N} | A = \{2, \dots, k+1\}\right).$$

By using the maximal summand as an upper bound, a standard union bound and the identical distribution for all μ we arrive at the following term

$$\leq \sum_{k=0}^{M-1} \binom{M-1}{k} p^k (1-p)^{M-1-k} \cdot$$
$$\mathbb{P}\left(\max_{\mu \in \{2, \dots, k+1\}} \exp(\langle \xi^\mu | \widetilde{\xi}^1 \rangle) > \frac{1}{k}(e^{N(1-2\varrho)-2} - (M-1-k)e^{\beta N}) | A = \{2, \dots, k+1\}\right)$$
$$\leq \sum_{k=0}^{M-1} k \binom{M-1}{k} p^k (1-p)^{M-1-k} \cdot$$
$$\mathbb{P}\left(\exp(\langle \xi^2 | \widetilde{\xi}^1 \rangle) > \frac{1}{k}(e^{N(1-2\varrho)-2} - (M-1-k)e^{\beta N}) | A = \{2, \dots, k+1\}\right).$$

Denote by

$$r = r(k) = \mathbb{P}\left(\exp\left(\langle \xi^2 | \widetilde{\xi}^1 \rangle\right) \geq \frac{1}{k}(e^{N(1-2\varrho)-2} - (M-1-k)e^{\beta N})\right).$$

We then arrive at

$$\mathbb{P}\left(\left(\exp(\langle \xi^2 | \widetilde{\xi}^1 \rangle) > \frac{1}{k}(e^{N(1-2\varrho)-2} - (M-1-k)e^{\beta N})\right) | A = \{2, \dots, k+1\}\right)$$
$$= \frac{\mathbb{P}\left(\left(\exp(\langle \xi^2 | \widetilde{\xi}^1 \rangle) > \frac{1}{k}(e^{N(1-2\varrho)-2} - (M-1-k)e^{\beta N})\right); 2 \in A\right)}{\mathbb{P}(2 \in A)} \leq \frac{r(k)}{p}$$

because $\mathbb{P}(2 \in A) = p$.

Thus an upper bound is:

$$\mathbb{P}\Big(\sum_{\mu=2}^{M}\exp(-\xi_i^\mu\xi_i^1+\sum_{j\neq i}\xi_j^\mu\widetilde{\xi}_j^1)-\exp(\xi_i^\mu\xi_i^1+\sum_{j\neq i}\xi_j^\mu\widetilde{\xi}_j^1)>ce^{N(1-2\varrho)}\Big)$$
$$\leq\sum_{k=0}^{M-1}k\binom{M-1}{k}r(k)p^{k-1}(1-p)^{M-1-k}.$$

Now let us split this sum into two parts: the first one for $k\in\{0,\dots,\lfloor 2p(M-1)\rfloor\}$ and the second one the remaining part. We start with the second part. By using the identity $k\binom{M-1}{k}=(M-1)\binom{M-2}{k-1}$ and the trivial fact that $r(k)\leq 1$, we get:

$$\sum_{k=\lfloor 2p(M-1)\rfloor+1}^{M-1}k\binom{M-1}{k}r(k)p^{k-1}(1-p)^{M-1-k}$$
$$\leq(M-1)\sum_{k=\lfloor 2p(M-1)\rfloor+1}^{M-1}\binom{M-2}{k-1}p^{k-1}(1-p)^{M-2-(k-1)}$$
$$=(M-1)\mathbb{P}\left(S_{M-2}\geq\lfloor 2p(M-1)\rfloor\right)\leq(M-1)\mathbb{P}\left(S_{M-2}\geq\frac{3}{2}p(M-2)\right).$$

We used the bound $\lfloor 2p(M-1)\rfloor>\frac{3}{2}p(M-2)$, which is a consequence of the fact that $p(M-1)$ goes to infinity. This will be shown at the end of the proof. Then, by using (6), with $\varepsilon=\frac{p}{2}$ we obtain:

$$\sum_{k=\lfloor 2p(M-1)\rfloor+1}^{M-1}k\binom{M-1}{k}r(k)p^{k-1}(1-p)^{M-1-k}\leq(M-1)\exp\left(-\frac{p(M-2)}{12}\right).$$

We consider now the first part. Since:

$$\frac{1}{k}(e^{N(1-2\varrho)-2}-(M-1-k)e^{\beta N})=\frac{1}{k}(e^{N(1-2\varrho)-2}-e^{(\alpha+\beta)N})+e^{\beta N}$$

we clearly see that $r(k)$ is increasing in k if $\alpha+\beta<1-2\varrho$ is fulfilled. This condition will be proven later on. Thus:

$$\sum_{k=0}^{\lfloor 2p(M-1)\rfloor}\binom{M-1}{k}kr(k)p^{k-1}(1-p)^{M-1-k}$$
$$\leq\frac{1}{p}\max_{k\in\{0,\dots,\lfloor 2p(M-1)\rfloor\}}kr(k)$$
$$\leq 2(M-1)\cdot r(2p(M-1)).$$

Let us examine this last term: since $p<e^{-NI(\beta)}$, we observe

$$r(2p(M-1))=\mathbb{P}\Big(\exp(\langle\xi^2|\widetilde{\xi}^1\rangle)>\frac{1}{2p}\big(e^{N(1-2\varrho-\alpha)-2}-e^{\beta N}\big)+e^{\beta N}\Big)$$
$$\leq\mathbb{P}\Big(\exp(\langle\xi^2|\widetilde{\xi}^1\rangle)>\frac{1}{2}\big(e^{N(1-2\varrho-\alpha+I(\beta))-2}-e^{(\beta+I(\beta))N}\big)\Big).$$

First of all let us show that $1-2\varrho-\alpha+I(\beta)>\beta+I(\beta)$, so that the first term dominates the second term: indeed, this is equivalent to proving that $\alpha+\beta<1-2\varrho$. By concavity we have for $x\in(0,1)$:

$$I(x) \leq \log\left(\frac{(1+x)^2}{2} + \frac{(1-x)^2}{2}\right) = \log(1+x^2) \leq x^2 \leq x.$$

From this we obtain

$$\alpha + \beta < \alpha + \beta_o = \alpha + I(\alpha) \leq 2\alpha \leq I(1-2\varrho) \leq 1 - 2\varrho.$$

This proves that $1 - 2\varrho - \alpha + I(\beta) > \beta + I(\beta)$ and also concludes the statement that $r(k)$ is increasing (see above). Now take γ such that $1 - 2\varrho - \varepsilon < \gamma < 1 - 2\varrho - \frac{\varepsilon}{2}$ for an $\varepsilon > 0$. Then

$$1 - 2\varrho - \varepsilon < \gamma < 1 - 2\varrho - \alpha + I(\beta).$$

For N sufficiently large we get:

$$r(2p(M-1)) \leq \mathbb{P}\left(\exp(\langle \xi^2 | \widetilde{\xi}^1 \rangle) > e^{\gamma N}\right).$$

By applying (7) and for N sufficiently large one sees that:

$$r(2p(M-1)) \leq \mathbb{P}\left(\exp(\langle \xi^2 | \widetilde{\xi}^1 \rangle) > e^{\gamma N}\right) \leq \exp\left(-N I(\gamma)\right).$$

Now if we bring everything together, we finally arrive at:

$$\begin{aligned}
&\mathbb{P}\Big(\exists \mu\ \exists i : T_i\big(\widetilde{\xi}^\mu\big) \neq \xi_i^\mu\Big) \\
&\quad \leq N \cdot M \cdot \mathbb{P}\Big(\sum_{\mu=2}^{M} \exp\big(-\xi_i^\mu \xi_i^1 + \sum_{j \neq i} \xi_j^\mu \widetilde{\xi}_j^1\big) - \exp\big(\xi_i^\mu \xi_i^1 + \sum_{j \neq i} \xi_j^\mu \widetilde{\xi}_j^1\big) > c e^{N(1-2\varrho)}\Big) \\
&\quad \leq N \cdot M \Big(2(M-1)\exp\big(-N I(\gamma)\big) + (M-1)\exp\big(-\frac{p(M-2)}{12}\big)\Big) \\
&\quad \leq 2N \frac{M}{M-1} \exp\left(-N(I(\gamma) - 2\alpha)\right) + N \frac{M}{M-1} \exp\Big(2\alpha - \frac{M-2}{M-1} \cdot \frac{p(M-1)}{12}\Big).
\end{aligned}$$

But by the definition of γ, we have $I(\gamma) > 2\alpha$. So the first term clearly tends to 0. For the second term by using the lower bound on p, we have that:

$$p(M-1) \geq \exp(N(\alpha - I(\beta) - \eta)).$$

But we know that $I(\beta) < I(\beta_o) = \alpha$ and $\eta = \frac{1}{2}(\alpha - I(\beta))$, so $\alpha - I(\beta) - \eta > 0$. Therefore the second term converges also to 0. Also a straightforward consequence of this bound is that $p(M-1)$ goes to infinity and thus establishes the still open fact that $\lfloor 2p(M-1) \rfloor > \frac{3}{2}p(M-2)$, used above. So clearly if the condition $\alpha < \frac{I(1-2\varrho)}{2}$ is fulfilled, we obtain

$$\mathbb{P}\left(\exists \mu\ \exists i : T_i\left(\widetilde{\xi}^\mu\right) \neq \xi_i^\mu\right) \to 0$$

This finishes the proof. □

References

1. Amit, D.J., Gutfreund, H., Sompolinsky, H.: Spin-glass models of neural networks. Phys. Rev. A. **32**(2), 1007–1018 (1985a). doi:10.1103/PhysRevA.32.1007
2. Amit, D.J., Gutfreund, H., Sompolinsky, H.: Storing infinite numbers of patterns in a spin-glass model of neural networks. Phys. Rev. Lett. **55**, 1530–1533 (1985b). doi:10.1103/PhysRevLett.55.1530
3. Bovier, A.: Sharp upper bounds on perfect retrieval in the Hopfield model. J. Appl. Probab. **36**(3), 941–950 (1999)

4. Bovier, A., Niederhauser, B.: The spin-glass phase-transition in the Hopfield model with p-spin interactions. Adv. Theor. Math. Phys. **5**(6), 1001–1046 (2001). doi:10.4310/ATMP.2001.v5.n6.a2
5. Dembo, A., Zeitouni, O.: Large deviations techniques and applications. Stochastic Modelling and Applied Probability, vol. 38, p. 396. Springer, Berlin (2010) Corrected reprint of the second (1998) edition. ISBN 978-3-642-03310-0. doi:10.1007/978-3-642-03311-7
6. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. U.S.A. **79**(8), 2554–2558 (1982)
7. Krotov, D., Hopfield, J.J.: Dense associative memory for pattern recognition. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 29, pp. 1172–1180. Curran Associates, Inc., Red Hook (2016)
8. Loukianova, D.: Lower bounds on the restitution error in the Hopfield model. Probab. Theory Relat. Fields **107**(2), 161–176 (1997). doi:10.1007/s004400050081
9. Löwe, M.: The storage capacity of generalized Hopfield models with semantically correlated patterns. Markov Process. Relat. Fields **5**(1), 1–19 (1999)
10. Löwe, M.: On the storage capacity of Hopfield models with correlated patterns. Ann. Appl. Probab. **8**(4), 1216–1250 (1998). doi:10.1214/aoap/1028903378
11. Löwe, M., Vermet, F.: The storage capacity of the Hopfield model and moderate deviations. Stat. Probab. Lett. **75**(4), 237–248 (2005). doi:10.1016/j.spl.2005.06.001
12. Löwe, M., Vermet, F.: The capacity of q-state Potts neural networks with parallel retrieval dynamics. Stat. Probab. Lett. **77**(14), 1505–1514 (2007). doi:10.1016/j.spl.2007.03.030
13. McEliece, R.J., Posner, E.C., Rodemich, E.R., Venkatesh, S.S.: The capacity of the Hopfield associative memory. IEEE Trans. Inform. Theory **33**(4), 461–482 (1987). doi:10.1109/TIT.1987.1057328
14. Newman, C.M.: Memory capacity in neural network models: rigorous lower bounds. Neural Netw. **1**(3), 223–238 (1988)
15. Talagrand, M.: Rigorous results for the Hopfield model with many patterns. Probab. Theory Relat. Fields **110**(2), 177–276 (1998). doi:10.1007/s004400050148