J Stat Phys (2016) 164:105–129
DOI 10.1007/s10955-016-1530-z

A Comparative Study of Sparse Associative Memories

Vincent Gripon[1] · **Judith Heusel**[2] · **Matthias Löwe**[2] ·
Franck Vermet[3]

Received: 16 November 2015 / Accepted: 22 April 2016 / Published online: 14 May 2016

Abstract We study various models of associative memories with sparse information, i.e. a pattern to be stored is a random string of 0s and 1s with about $\log N$ 1s, only. We compare different synaptic weights, architectures and retrieval mechanisms to shed light on the influence of the various parameters on the storage capacity.

Keywords Neural networks · Associative memory · Sparse patterns · Storage capacity · Exponential inequalities

Mathematics Subject Classification Primary: 82C32 · 60K35, Secondary: 68T05 · 92B20

1 Introduction

Starting with the seminal paper [7], Gripon, Berrou and coauthors revived the interest in associative memory models, see e.g. [1,11–13]. Their approach is motivated by both biological considerations and ideas from information theory and leads to a neural network that is organized in clusters of interacting neurons. They state that their model (which we will refer to as the GB model) is more efficient (see [7]) and has by far a larger storage capacity than the benchmark model for associative memories, the Hopfield model introduced in [9]. Indeed, their considerations lead to a storage capacity of the order $N^2/(\log N)^2$ messages (or patterns or images; these words will be used synonymously) for their model with N neurons,

✉ Judith Heusel
jheus_01@uni-muenster.de

[1] Telecom Bretagne UMR CNRS Lab-STICC, Technopole Brest Iroise, 29238 Brest, France

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

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

while the standard Hopfield model with N neurons only has a capacity of $N/(2\log N)$ (see [4,19]).

However, the standing assumption of the GB model is that for N neurons there are c clusters of neurons with $1 \leq c \leq \log N$, and each message to be stored has only exactly one active neuron per cluster. This not only leads to a restriction on the number of storable messages, but also to them being very sparse (where sparsity is defined by a small number of active neurons). As a matter of fact, for sparse messages other models of associative memories have been proposed by Willshaw [24], Amari [10], Okada [20], or [2,8,16]. All these models have in common that their storage capacity is conjectured to be much larger than that of the Hopfield model. The Willshaw model has also been discussed in a number of papers by Palm, Sommer, and coauthors ([21–23] e.g.), with the difference that there the focus is rather on information capacity than on exact retrieval (and that many of the techniques are not rigorous). In [16] it has been rigorously proven for a sparse three-state network, the so called Blume–Emery–Griffiths model, that the capacity is indeed of the predicted order (even though there, strictly speaking the degree of sparsity is not allowed to depend on the number of neurons).

A natural question is thus to separate the various factors that can influence the storage capacity of a model: the sparseness of the messages, the storage mechanism, and the algorithm to retrieve the stored patterns. The objective of the present article is to analyze this question. To this end we will try to give bounds on the storage capacity of the Willshaw model, Amari's version of a sparse 0–1 Hopfield model, and the GB model. In particular, we will see that all these models achieve a storage capacity of the order of $N^2/(\log N)^2$ when the number of active neurons c satisfies $c = a \log N$ for some positive a. Also we will discuss the influence of model specificities to the absolute constants in the storage capacities.

More precisely, we organize our article in the following way. In the next section, we describe the three models we aim at studying and formally define what is meant by "storing a message". In Sect. 3 we give some insight why an order of $N^2/(\log N)^2$ for the number of stored messages is to be expected in a model with N neurons, of which about only $\log N$ are active. To this end we consider a certain event in the GB model that implies that a message cannot be retrieved correctly. In the fourth section we state our main results. These are proved in Sect. 5. Section 6 takes up ideas from Sect. 3 to show, that if the number of messages is too large, an erased message cannot be completed correctly in the GB model. Finally, Sect. 7 discusses some dynamical properties of the considered models and contains some simulations, in particular on the probability to correct an error in several steps. These probabilities are notoriously difficult to access analytically (see e.g. [5,17,18]). The simulations give an impression of the advantages and drawbacks of the several models.

2 The Models

We will now present the models that are in the center of our interest in the present paper. The reference model is always the Hopfield model on the complete graph (i.e. all neurons are interconnected), with M patterns $(\xi^\mu)_{\mu=1,\dots M} = (\xi_i^\mu)_{i=1,\dots N}^{\mu=1,\dots M} \in \{-1,+1\}^{N\times M}$. Here the so called synaptic efficacy J_{ij} is given by

$$J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu \qquad 1 \leq i \neq j \leq N$$

and an input $\sigma \in \{-1,+1\}^N$ is transformed by the dynamics

$$T_i(\sigma) = \operatorname{sgn}\left(\sum_{j \neq i} J_{ij}\sigma_j\right)$$

where sgn is the sign function (and the sign of 0 is chosen at random). This update can happen either synchronously or asynchronously in i. In [19] it was shown that for unbiased and i.i.d. random variables $((\xi_i^\mu)_{i=1,\dots N})_{\mu=1,\dots M}$ and $M = c\frac{N}{\log N}$ with $c < \frac{1}{2}$, an arbitrary message is stable under the dynamics with a probability converging to one. Of course, this model can be generalized to i.i.d. biased patterns with expectation a. In [15] the author suggests to replace the synaptic efficacy by $J_{ij} = \sum_\mu (\xi_i^\mu - a)(\xi_j^\mu - a)$ and shows that the storage capacity (in the sense that an arbitrary pattern is a fixed point of the above dynamics) decreases for a strong bias. More precisely, he gives a lower bound on the storage capacity of the Hopfield model with biased patterns of the form $Cp^2(1-p)^2N/\log N$, where C is an explicit constant that depends on the notion of storage capacity used and p is the probability that ξ_1^1 equals $+1$. Note that this behaviour is amazingly similar to the behaviour of Hopfield models with correlated patterns, cf. [14]. Another model for biased ± 1-patterns was proposed by Okada [20].

However, if we think of the bias as a certain sparsity of the patterns, it may be more natural to consider patterns $(\xi^\mu)_{\mu=1,\dots M}$ where the (ξ_i^μ) still are i.i.d. but take values 0 and 1 where $\mathbb{P}(\xi_i^\mu = 1) = p$ is small. We will henceforth consider such patterns and three such models.

2.1 Amari's Model

The model Amari proposed in [10] is closest in spirit to the Hopfield model. Here we take $J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$ and with this new setting, we consider input spin configurations $\sigma \in \{0,1\}^N$ and map their spins to either 0 or 1 with the help of a dynamics. Of course, one should only map an input spin σ_i to 1, if the so called local field $\sum_{j\neq i} J_{ij}\sigma_j$ is large enough, say larger than a given threshold. To compare Amari's results to the other models we choose

$$\mathbb{P}(\xi_i^\mu = 1) = p = \frac{\log N}{N}.$$

As a matter of fact, this is the case of extremely diluted patterns, since if p is even smaller, say $p = c/N$ for some c, with positive probability some of the patterns will entirely consist of 0's and will thus be indistinguishable.

We propose the following dynamics, where a spin σ_i will be 1, if the so called local field

$$S_i(\sigma) = \sum_{j\neq i} J_{ij}\sigma_j$$

is large enough, say larger than a given threshold.

$$T_i(\sigma) = \Theta(S_i(\sigma) - h)$$

where $\Theta(x) = \mathbf{1}_{\{x\geq 0\}}$ and we choose $h = \gamma \log N$ for some $\gamma > 0$. Note that this seems a reasonable choice if we want the (ξ^μ) to be fixed points of the dynamics. Consider for example the case $\xi_i^1 = 1$ we have that

$$\sum_{j\neq i} J_{ij}\xi_j^1 = \sum_{j\neq i}\xi_j^1 + \sum_{\mu\neq 1}\sum_{j\neq i}\xi_i^\mu \xi_j^\mu \xi_j^1$$

and the first term on the right hand side is of order $\log N$. Also note that Amari just considers the case of a fixed number $\log N$ of active neurons per message (which is similar), and states that the above model would perform much worse in the case we consider. We will see that this is not the case.

2.2 The Willshaw Model

The following model was proposed in a celebrated paper by Willshaw [24]. It corresponds to Amari's model with the restriction that the efficacy J_{ij} does not depend on the *number* of messages that use neurons i and j but just on whether there is any μ with $\xi_i^\mu \xi_j^\mu = 1$. In the case of the Hopfield model this procedure is known as "clipped" synapses.

Formally, we will now either assume that the (ξ_i^μ) are i.i.d $0 - 1$ random variables with success probability $p = \frac{\log N}{N}$ or we take the M messages to be realized uniformly at random from all sets of M messages with exactly $c = \log N$ active neurons. Both cases are similar, but the first one is mathematically more convenient, because in this case the images as well as all their spins are independent. Moreover, in the Willshaw model we choose

$$J_{ij} = \Theta\left(\sum_\mu \xi_i^\mu \xi_j^\mu - 1\right) = \begin{cases} 1 & \text{if } \exists \mu : \xi_i^\mu \xi_j^\mu = 1 \\ 0 & \text{otherwise,} \end{cases}$$

for all $i, j \in \{1, \ldots, N\}$. There are two different (yet similar) types of dynamics to be considered. The first one is the threshold dynamics also considered in Amari's model. So again for an input $\sigma \in \{0, 1\}^N$ we set

$$T_i\left(\sigma\right) = \Theta(\bar{S}_i(\sigma) - h\Big)$$

with $\bar{S}_i(\sigma) = \sum_j J_{ij}\sigma_j$ and $h = \gamma \log N$, for some $\gamma > 0$. This dynamics is applicable to both types of patterns (i.i.d. random variables (ξ_i^μ) or randomly chosen messages amongst all sets of M messages with exactly c active neurons). For the Willshaw model, we consider $\bar{S}_i$ instead of $S_i(\sigma) = \sum_{j \neq i} J_{ij}\sigma_j$, since simulations support that it improves performance to modify S_i in order to account for self influence of neurons. This modification is well known and will be referred as "memory effect".

In the latter case of exactly c active neurons per message and the messages being randomly chosen messages amongst all sets of M messages with exactly c active neurons there is another retrieval dynamics that requires the knowledge of all the $\bar{S}_i(\sigma)$ for $1 \leq i \leq N$. In this setting, for a given input $\sigma \in \{0, 1\}^N$ we compute all the $\bar{S}_i(\sigma)$ and order them: they will be denoted by $h_{(1)} \geq h_{(2)} \geq \cdots \geq h_{(c)} \geq \cdots \geq h_{(N)}$. Then we set all neurons i with $\bar{S}_i(\sigma) \geq h_{(c)}$ to 1 and the others to 0. Note that in case of a tie we may obtain more than c 1's after a step of the dynamics. This procedure was called "Winner takes all"-algorithm (WTA algorithm, for short) in [13] in a model that is closely related to the following cluster model.

Similarly, we may as well imagine that c is fixed but we do not know it. In this case we could just take the most active neurons, i.e. set all neurons with a value $\bar{S}_i(\sigma)$ lower than $h_{(1)}$ to 0. Interestingly, if we consider as input a partially erased version $\tilde{\xi}^\mu$ of a stored message ξ^μ, for the one step retrieval we consider theoretically in Sects. 4 and 5, this does not change anything as long as we consider the memory effect described above, since in this case $h_{(1)} = h_{(c)}$. This is because $h_{(1)}$ cannot be larger than the number of initial 1's in the dynamics input and this upper bound is reached for at least all the neurons that are active in the message ξ^μ we are looking for. Considering the performance of the model with several steps of the retrieval dynamics numerically, however, shows that the above threshold $h_{(c)}$ is superior to a threshold $h_{(1)}$. As a matter of fact, the dynamics using $h_{(1)}$ as threshold does not benefit from performing more than one iteration (see Theorem 7.4).

On the other hand using $h_{(c)}$ allows for improvement over the time. Also note that the WTA algorithm with $h_{(1)}$ as threshold can be applied in the case where the (ξ_i^μ) are i.i.d $0-1$ random variables with success probability $p = \frac{\log N}{N}$, as will be proven in Sect. 4.

2.3 The GB Model

Here we assume that $N = l \log l =: l \cdot c$ for some l. One tries to store M messages $\xi^1, \ldots, \xi^M$ in a network with a block structure. The messages are sparse in the sense that each message ξ^μ has c active neurons, only, one in each block of l neurons. To take into account the block structure, we will denote by (a, k) the k-th neuron of the a-th block.

For $a \neq a'$, an edge $e = ((a, k), (a', k'))$ is said to be active for the message ξ^μ if $\xi^\mu_{(a,k)}\xi^\mu_{(a',k')} = 1$. Let

$$\mathcal{E}\left((\xi^\mu)_{\mu=1,\ldots,M}\right) := \left\{e : e \text{ is an active edge of one of the } \xi^\mu\right\}.$$

We can also define the graph associated with or spanned by an arbitrary message ξ^0. This will be the (necessarily complete) graph with all vertices (a, k) such that $\xi^0_{(a,k)} = 1$ and edges $e = ((a, k), (a', k'))$ for all $a, a', k, k', a \neq a'$ such that $\xi^0_{(a,k)}\xi^0_{(a',k')} = 1$. Then a message ξ^0 is considered to be stored in the model if all edges of this complete graph spanned by ξ^0 are present in the set of edges $\mathcal{E}((\xi^\mu)_{\mu=1,\ldots,M})$.

Similar to the Willshaw model, we define the synaptic efficacy by

$$W_{(a,k),(a',k')} = \Theta\left(\sum_{\mu=1}^{M} \xi^\mu_{(a,k)}\xi^\mu_{(a',k')} - 1\right).$$

Thus for $a \neq a'$ $W_{(a,k),(a',k')} = 1$ if and only if (a, k) and (a', k') are activated simultaneously in one of the messages (both in the same message). On the other hand, for $a = a'$ we have $W_{(a,k),(a,k')} = 1$ if and only if $k = k'$ and there exists μ such that the k'th neuron in block a is 1. As a matter of fact, this description shows that the major difference to the Willshaw model is that in the GB model one has a restriction of the location of the 1's.

With this synaptic efficacy one can associate a dynamics T on $(\{0, 1\}^l)^c$: instead of the local field $S_i(\sigma)$ of the preceding models, we define

$$S_{(a,k)}(\sigma) = \sum_{b=1}^{c}\sum_{r=1}^{l} \Theta\left(W_{(a,k),(b,r)}\sigma_{(b,r)} - 1\right)$$

for $\sigma \in (\{0, 1\}^l)^c$, and the dynamics

$$T_{(a,k)}(\sigma) = \Theta(S_{(a,k)}(\sigma) - h).$$

Here again h is a threshold that needs to be adapted to the tasks we want the network to perform. E.g., choosing $h = c$ one readily verifies that all stored messages $\xi \in \mathcal{M} = \{\xi^1, \ldots, \xi^M\}$ are stable, i.e. we have $T(\xi) = \xi$. Obviously, this can only go to the expense of error tolerance of the network.

The dynamics described above is the equivalent of the threshold dynamics in the Willshaw model. As in the latter model, we can also define a WTA algorithm. This will respect the local nature of the GB model. To describe it, assume we want to update the values of the neurons in the a'th cluster $\sigma_{(a,k)}$, $k = 1, \ldots, l$. For each $k = 1, \ldots, l$ we then build

$$s_{(a,k)}(\sigma) = \sum_{b=1}^{c} \Theta\left(\sum_{r=1}^{l} W_{(a,k),(b,r)}\sigma_{(b,r)} - 1\right). \tag{1}$$

(This is called the SUM-OF-MAX rule in [25]; it accounts for the fact that in each message there only can be one active connection between two clusters). We then order the $s(a,k)$, $k = 1, \ldots, l$ and set the neuron(s) with the largest value to 1 and all others to 0.

3 Wrong Messages and a First Bound on the Storage Capacity

In this section we will approach the question: what could be the right order for the storage capacity of the above networks?

At first glance, storage capacity may refer to different properties of the network. E.g. from Sect. 4 we will ask ourselves: how many messages can we store such that they are fixed points of the network dynamics or how many messages can we register in our network such that even a certain number of errors can be corrected? On the other hand, in the previous section we already learned that in the GB model with a threshold dynamics, an arbitrary number of input messages is stable if we choose the threshold equal to c, the number of active neurons. It is intuitively clear that this can only have a negative effect on the error retrieval abilities of the network, if we store too many messages in the network.

An extreme case of such a lack of error tolerance is if we recognize an input as a stored message even if it is not. This property will be discussed in greater detail for the GB model and partially for the Willshaw model in this section. The insight we gain will provide us with an idea of how many messages we can store in the models.

We will prove the following theorem.

Theorem 3.1 *Consider the GB model with the threshold retrieval dynamics and threshold $h = c$. Take*

$$M = \alpha(\log c)l^2 = \alpha l^2 \log\log l.$$

If $\alpha > 2$, a random message (independent of the stored patterns) will be recognized as a stored message with probability converging to 1 as $l \to \infty$.

If $\alpha = 2$ and as $l \to \infty$, with strictly positive probability a random message will be recognized as a stored message.

On the other hand, if $\alpha < 2$ the probability that a random message will be recognized as stored goes to zero as $l \to \infty$.

We will use positive association of random variables (see e.g. [6]) to prove this theorem. Recall that a set of real valued random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is positively associated, if for any non-decreasing functions f and g from $\mathbb{R}^n$ to $\mathbb{R}$ for which the corresponding expectations exist we have

$$\mathrm{Cov}(f(\mathbf{X}), g(\mathbf{X})) \geq 0.$$

Also recall that independent random variables are positively associated and that non-decreasing functions of positively associated random variables remain positively associated.

For positively associated random variables we will repeatedly apply the following inequality.

Lemma 3.2 (see [3, Theorem 1]) *Let $X_1, X_2, \ldots, X_n$ be positively associated integer valued random variables. Then*

$$0 \le \mathbb{P}[X_i = 0, i = 1, \dots, n] - \prod_{i=1}^{n} \mathbb{P}[X_i = 0] \le \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j).$$

Proof of Theorem 3.1 Let ξ^0 be a random message. Without loss of generality we may (after relabelling) assume that $\xi^0_{(a,1)} = 1$, for all $a = 1, \dots, c$. Let $\mathcal{G}(\xi^0)$ be the event that ξ^0 is stored in the GB model. Its probability $\mathbb{P}(\mathcal{G}(\xi^0))$ is given by

$$\mathbb{P}\left(\mathcal{G}(\xi^0)\right) = \mathbb{P}\left(\forall a, b \in \{1, \dots, c\}, a \ne b, \exists \mu \in \{1, \dots, M\} : \xi^\mu_{(a,1)} \xi^\mu_{(b,1)} = 1\right).$$

Note that the latter can be rewritten as

$$\mathbb{P}\left(\forall a, b \in \{1, \dots, c\}, a \ne b : \max_\mu \xi^\mu_{(a,1)} \xi^\mu_{(b,1)} = 1\right).$$

Now the $(\xi^\mu_{(a,1)})$ are independent $0-1$-valued random variables, and taking their product and the maximum of these products are increasing functions of them. Thus $\{\max_\mu \xi^\mu_{(a,1)} \xi^\mu_{(b,1)}, a \ne b\}$ are positively associated (see e.g. [6]), which implies

$$\begin{aligned}\mathbb{P}\left(\forall a, b \in \{1, \dots, c\}, a \ne b : \max_\mu \xi^\mu_{(a,1)} \xi^\mu_{(b,1)} = 1\right) &\ge \mathbb{P}\left(\max_\mu \xi^\mu_{(a,1)} \xi^\mu_{(b,1)} = 1\right)^{c(c-1)/2} \\ &= \left(1 - (1 - 1/l^2)^M\right)^{c(c-1)/2}\end{aligned}$$

where on the right hand side of the above inequality a and b is an arbitrary pair of distinct variables.

Choosing $M = \alpha \log cl^2$ we see that the right hand side is approximately given by

$$\left(1 - \left(1 - 1/l^2\right)^M\right)^{c(c-1)/2} \approx \exp\left(-\frac{c^2}{2} e^{-\alpha \log c}\right)$$

which converges to 1 if $\alpha > 2$, and to $e^{-1/2}$ if $\alpha = 2$.

On the other hand, we can also use positive association for an upper bound. We put $X_e = \max\{\xi^\mu_{(a,1)} \xi^\mu_{(b,1)}, \mu = 1, \dots, M\}$ for $e = ((a, 1), (b, 1))$ and

$$Z = \sum_{e \in V} X_e \quad \text{with } V = \{((a, 1), (b, 1)), a, b \in \{1, \dots, c\}, a \ne b\}.$$

Trivially,

$$\mathbb{P}\left[\mathcal{G}(\xi^0)\right] = \mathbb{P}[Z = c(c-1)/2].$$

On the other hand, the random variables $Y_e = 1 - X_e$ are also positively associated integer valued, and we may use the above lemma to arrive at

$$\mathbb{P}[Z = L] \le \prod_e \mathbb{P}[Y_e = 0] + \sum_{e, e' \in V} \mathrm{Cov}(Y_e, Y_{e'})$$

i.e.

$$\mathbb{P}[Z = L] \le d^L + \sum_{e, e' \in V} \mathrm{Cov}(X_e, X_{e'}) \tag{2}$$

where we set $d := (1 - (1 - 1/l^2)^M)$ and we are left with computing the covariances. To this end notice that $\mathrm{Cov}(X_e, X_{e'}) = 0$, if e and e' are disjoint. So assume that $e = ((a, 1), (b, 1))$ and $e' = ((a, 1), (b', 1))$ and put $\mathcal{M}(a, 1) := \{\mu : \xi^\mu_{(a,1)} = 1\}$. Then

$$\begin{aligned}\mathbb{E}(X_eX_{e'}) &= \mathbb{P}\left(\exists \mu, \nu \in \mathcal{M}(a,1) : \xi^{\mu}_{(b,1)} = 1, \xi^{\nu}_{(b',1)} = 1\right)\\ &= \sum_{r=0}^{M} \mathbb{P}\left(\exists \mu, \nu \in \mathcal{M}(a,1) : \xi^{\mu}_{(b,1)}\xi^{\nu}_{(b',1)} = 1 \mid |\mathcal{M}(a,1)| = r)\mathbb{P}(|\mathcal{M}(a,1)| = r\right)\\ &= \sum_{r=0}^{M} \mathbb{P}\left(\exists \mu \in \mathcal{M}(a,1) : \xi^{\mu}_{(b,1)} = 1 \mid |\mathcal{M}(a,1)| = r)^2\mathbb{P}(|\mathcal{M}(a,1)| = r\right)\\ &= \sum_{r=0}^{M} \left(1-(1-1/l)^r\right)^2 \binom{M}{r}(1/l)^r(1-1/l)^{M-r},\end{aligned}$$

as on $\mathcal{M}(a,1)$ the events

$$\left\{\exists \mu \in \mathcal{M}(a,1) : \xi^{\mu}_{(b,1)} = 1\right\} \quad \text{and} \quad \left\{\exists \nu \in \mathcal{M}(a,1) : \xi^{\nu}_{(b',1)} = 1\right\}$$

are independent and have equal probabilities. The expression on the right hand side can be simplified to give

$$\begin{aligned}\mathbb{E}(X_eX_{e'}) &= 1 - 2\sum_{r=0}^{M}\binom{M}{r}(1/l)^r(1-1/l)^M + \sum_{r=0}^{M}\binom{M}{r}(1/l)^r(1-1/l)^{M+r}\\ &= 1 - 2(1-1/l)^M(1+1/l)^M + \left(1-1/l)^M(1+\frac{1}{l}(1-1/l)\right)^M\\ &= 1 - 2\left(1-1/l^2\right)^M + \left(1-2/l^2+1/l^3\right)^M.\end{aligned}$$

On the other hand,

$$(\mathbb{E}(X_e))^2 = (\mathbb{P}(X_e = 1))^2 = d^2 = \left(1 - \left(1 - \frac{1}{l^2}\right)^M\right)^2.$$

This yields

$$\begin{aligned}\mathrm{Cov}(X_e, X_{e'}) &= 1 - 2\left(1-1/l^2\right)^M + \left(1-2/l^2+1/l^3\right)^M - \left(1-\left(1-1/l^2\right)^M\right)^2\\ &= \left(1-2/l^2+1/l^3\right)^M - \left(1-2/l^2+1/l^4\right)^M\\ &= \exp\Big(M\log\big(1-2/l^2+1/l^3\big)\Big) - \exp\Big(M\log\big(1-2/l^2+1/l^4\big)\Big)\\ &= \exp\left(-2M/l^2\right)\left(M/l^3 + \mathcal{O}\left(M/l^4\right)\right),\end{aligned}$$

after expanding the logarithm and the exponential and taking into account that $M(\frac{1}{l})^3$ converges to 0 for our choice of the parameters. Thus for

$$M = \alpha(\log\log N)N^2/(\log N)^2$$

we obtain because of $c = \log l \approx \log N$.

$$\begin{aligned}\sum_{e,e'\in V}\mathrm{Cov}(X_e, X_{e'}) &\leq \alpha(\log\log N)c^4\exp(-2\alpha\log\log N)/N\\ &\approx \frac{1}{N}\alpha(\log\log N)(\log N)^4\exp(-2\alpha\log\log N)\end{aligned}$$

Inserting this into (2), we obtain

$$\begin{aligned}\mathbb{P}[\mathcal{G}(\xi^0)] &= \mathbb{P}[Z = c(c-1)/2] \\ &\leq d^L + \sum_{e,e' \in V} \mathrm{Cov}(X_e, X_{e'}) \\ &\leq d^L + \frac{1}{N}\alpha(\log\log N)(\log N)^4 \exp(-2\alpha \log\log N) \\ &\leq d^L + \frac{1}{N}\alpha(\log N)^{(4-2\alpha)} \log\log N\end{aligned}$$

The second summand on the right hand side clearly vanishes. But also d^L converges to 0 for $\alpha < 2$ (which can be seen as in the first part of the proof). Thus $\mathbb{P}[\mathcal{G}(\xi^0)]$ converges to 0, and we can remark that $\mathbb{P}[\mathcal{G}(\xi^0)]$ is exactly of order d^L for $\alpha \in]1, 2[$. □

Remark 3.3 The above computation also justifies a choice of c that is not of constant order. Indeed, for c being a constant independent of N the same approximation of $\mathbb{P}[\mathcal{G}(\xi^0)]$ by d^L is true. However d^L converges to a constant larger than 0, even if $M = l^2$.

A very similar theorem holds true, for the Willshaw model with an intensity of 1s given by $\mathbb{P}(\xi_i^\mu = 1) = \frac{\log N}{N}$.

Theorem 3.4 *Consider the Willshaw model with i.i.d. messages and coordinates such that* $\mathbb{P}(\xi_i^\mu = 1) = \frac{\log N}{N}$. *Consider the threshold retrieval dynamics with threshold* $h = c$. *Take* $M = \alpha \frac{N^2}{(\log N)^2} \log\log N$.

If $\alpha > 2$ *a random message with* c *active neurons (independent of the stored patterns) will be recognized as a stored message with probability converging to 1 as* $l \to \infty$.

If $\alpha = 2$ *and as* $l \to \infty$, *with strictly positive probability a random message will be recognized as a stored message.*

On the other hand, if $\alpha < 2$ *the probability that a random message will be recognized as stored goes to zero as* $l \to \infty$.

The proof is almost identical to the proof of the previous theorem. We therefore omit it.

4 Stability and Error Correction

In this section we will try to give lower and sometimes also upper bounds on the number of patterns we can store in the various models, such that the given messages are stable under the dynamics of the network and errors in the input can be corrected.

We saw that in the GB model and the Willshaw model, slightly more than $N^2/(\log N)^2$ already suffice to supersaturate the networks. We will therefore always assume that $M = \alpha N^2/(\log N)^2$.

We start with Amari's model.

Theorem 4.1 *Suppose that in Amari's model with threshold* $h = \gamma \log N$ ($\gamma < 1$ *to be chosen appropriately), we have that* $M = \alpha N^2/(\log N)^2$. *Then, if* $\alpha < e^{-2}$ *for any fixed* μ, *we have*

$$\mathbb{P}\left(\forall i : T_i(\xi^\mu) = \xi_i^\mu\right) \to 1$$

as $N \to \infty$.

Moreover, for any error rate $0 < \rho < 1$*, if* $\gamma < 1 - \rho$ *is chosen appropriately and* $\alpha < (1-\rho)e^{-(1+\frac{1}{1+\rho})}$*, for any fixed* μ*, and any* $\tilde{\xi}^\mu$ *obtained by deleting at random* $\rho \log N$ *of the 1's in* ξ^μ*, we have:*

$$\mathbb{P}\left(\forall i : T_i(\tilde{\xi}^\mu) = \xi_i^\mu\right) \to 1$$

as $N \to \infty$*.*

Finally, if $M > -\log(1-e^{-1})N^2/(\log N)^2$

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

as $N \to \infty$*.*

It is interesting to observe that the previous theorem also gives a result on the Willshaw model with a threshold dynamics.

Corollary 4.2 *In the Willshaw model with i.i.d. random variables* ξ_i^μ*, threshold* $h = \gamma \log(N)$*,* $\gamma < 1$ *and* $M = \alpha N^2/(\log N)^2$ *for* $\alpha < e^{-2}$ *we have for any fixed* μ

$$\mathbb{P}\left(\forall i : T_i(\xi^\mu) = \xi_i^\mu\right) \to 1$$

as $N \to \infty$*.*

Moreover, for any error rate $0 < \rho < 1$*, if* $\gamma < 1 - \rho$ *is chosen appropriately and* $\alpha < (1-\rho)e^{-(1+\frac{1}{1+\rho})}$*, for any fixed* μ*, and any* $\tilde{\xi}^\mu$ *obtained by deleting at random* $\rho \log N$ *of the 1's in* ξ^μ*, we have:*

$$\mathbb{P}\left(\forall i : T_i(\tilde{\xi}^\mu) = \xi_i^\mu\right) \to 1$$

as $N \to \infty$*.*

Finally, if $M > -\log(1-e^{-1})N^2/(\log N)^2$

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

as $N \to \infty$*.*

In computer simulations the threshold dynamics in the Willshaw model is outperformed by WTA. Our theoretical results are by now limited to the question of the stability of messages and one step of the retrieval dynamics.

Theorem 4.3 *Consider the Willshaw model with i.i.d. messages and coordinates such that* $\mathbb{P}[\xi_i^\mu = 1] = \frac{c}{N}$*, where* $c = \log(N)$*. Consider the WTA dynamics with threshold* $h_{(1)}$ *and let* $M = \alpha N^2/(\log N)^2$*. Then for* $\alpha < -\log(1-e^{-1})$ *we have for any fixed* μ

$$\mathbb{P}\left(\forall i : T_i(\xi^\mu) = \xi_i^\mu\right) \to 1$$

as $N \to \infty$*.*

This bound is sharp: For $\alpha > -\log(1-e^{-1})$ *we have for any fixed* μ

$$\mathbb{P}\left(\exists i : T_i(\xi^\mu) \neq \xi_i^\mu\right) \to 1$$

as $N \to \infty$*.*

Finally, if $\rho \log N$*,* $0 \leq \rho < 1$ *of the initial 1's of message* ξ^μ *are erased at random to obtain* $\tilde{\xi}^\mu$*, we can prove the following result:*
For $\alpha < -\log(1-e^{-1/(1-\rho)})$ *we have for any fixed* μ

$$\mathbb{P}\left(\forall i : T_i(\tilde{\xi}^\mu) = \xi_i^\mu\right) \to 1$$

as $N \to \infty$*.*

Again, this bound is sharp: For $\alpha > -\log(1 - e^{-1/(1-\rho)})$ we have for any fixed μ

$$\mathbb{P}\left(\exists i : T_i(\tilde{\xi}^\mu) \neq \xi_i^\mu\right) \to 1$$

as $N \to \infty$.

Remark 4.4 For mathematical convenience, we assumed in Th. 4.3 that the stored messages are independent, with i.i.d. coordinates (ξ_i^μ) such that

$$\mathbb{P}[\xi_i^\mu = 1] = \frac{c}{N}.$$

We can naturally expect the same results in the case where exactly c neurons are active in each stored message, but properties of independence are lacking to prove such results in this situation.

A very similar statement holds for the GB model with the WTA algorithm.

Theorem 4.5 *In the GB model with independent messages with WTA dynamics (which again is called T) let $M = \alpha l^2/c^2$. Then for $\alpha < -\log(1 - e^{-1})$ we have for any fixed μ*

$$\mathbb{P}\left(\forall (a, c) : T_{(a,c)}(\xi^\mu) = \xi_{(a,c)}^\mu\right) \to 1$$

as $N \to \infty$.

If $\rho \log N$ of the initial 1's of a message ξ^μ are erased at random to construct $\tilde{\xi}^\mu$, we obtain: For $\alpha < -\log(1 - e^{-1/(1-\rho)})$ we have for any fixed μ

$$\mathbb{P}\left(\forall (a, c) : T_{(a,c)}(\tilde{\xi}^\mu) = \xi_{(a,c)}^\mu\right) \to 1$$

as $N \to \infty$.

5 Proofs

This section contains the proofs of the results in the previous section. We start with Theorem 4.1.

Proof of Theorem 4.1 Recall the situation of the theorem. We choose $h = \gamma \log(N)$ with $\gamma \in (0, 1)$. Then, for each $\delta \in (0, 1)$,

$$\begin{aligned}
&\mathbb{P}\left(\exists 1 \le i \le N, T_i(\xi^1) \neq \xi_i^1\right)\\
&\le \mathbb{P}\Big(\Big|\log(N) - \sum_j \xi_j^1\Big| \ge (1-\delta)\log(N)\Big)\\
&\quad + \mathbb{P}\left(\left\{\exists 1 \le i \le N, T_i(\xi^1) \neq \xi_i^1\right\} \cap \left\{\Big|\log(N) - \sum_j \xi_j^1\Big| < (1-\delta)\log(N)\right\}\right)
\end{aligned}$$

and the first term disappears as $N \to \infty$ due to the law of large numbers, since the ξ_i^μ are Bernoulli random variables with success probability $p = \log N/N$.

Let $\delta > \gamma$. If

$$\Big|\log(N) - \sum_j \xi_j^1\Big| < (1-\delta)\log(N),$$

we have that $\sum_j \xi_j^1 > \delta \log(N)$, and for each i with $\xi_i^1 = 1$, we obtain

$$S_i(\xi^1) = \sum_{j \neq i} J_{ij} \xi_j^1 = \xi_i^1 \sum_{j \neq i} \xi_j^1 + \sum_{j \neq i} \xi_j^1 \sum_{\mu=2}^{M} \xi_i^\mu \xi_j^\mu \geq \sum_{j \neq i} \xi_j^1 \geq \delta \log(N) - 1 \geq \gamma \log(N),$$

for N large enough, i.e. $T_i(\xi^1) = 1$.

On the other hand, for each i with $\xi_i^1 = 0$, we get

$$\begin{aligned}
&\mathbb{P}\left(\{T_i(\xi^1) \neq \xi_i^1\} \cap \{\xi_i^1 = 0\} \cap \left\{\left|\log(N) - \sum_j \xi_j^1\right| < (1-\delta)\log(N)\right\}\right) \\
&\leq \sum_{k=\lfloor \delta \log(N) \rfloor}^{\lceil (2-\delta)\log(N) \rceil} \mathbb{P}\left(\left\{\sum_{j \neq i} \xi_j^1 \sum_{\mu=2}^{M} \xi_i^\mu \xi_j^\mu \geq \gamma \log(N)\right\} \cap \{\xi_i^1 = 0\} \cap \left\{\sum_j \xi_j^1 = k\right\}\right) \\
&= \sum_{k=\lfloor \delta \log(N) \rfloor}^{\lceil (2-\delta)\log(N) \rceil} \mathbb{P}\left(\left\{\sum_{j} \xi_j^1 \sum_{\mu=2}^{M} \xi_i^\mu \xi_j^\mu \geq \gamma \log(N)\right\} \cap \{\xi_i^1 = 0\} \Big| \sum_j \xi_j^1 = k\right) \cdot \mathbb{P}\left(\sum_j \xi_j^1 = k\right),
\end{aligned}$$

since for $j = i$, the term $\xi_j^1 \sum_{\mu=2}^{M} \xi_i^\mu \xi_j^\mu$ is equal to 0.

This yields

$$\begin{aligned}
&\mathbb{P}\left(\{T_i(\xi^1) \neq \xi_i^1\} \cap \{\xi_i^1 = 0\} \cap \left\{\left|\log(N) - \sum_j \xi_j^1\right| < (1-\delta)\log(N)\right\}\right) \\
&\leq \max_{\lfloor \delta \log(N) \rfloor \leq k \leq \lceil (2-\delta)\log(N) \rceil} \mathbb{P}\left(\sum_j \xi_j^1 \sum_{\mu=2}^{M} \xi_i^\mu \xi_j^\mu \geq \gamma \log(N) \Big| \sum_j \xi_j^1 = k\right) \cdot \\
&\sum_{k=\lfloor \delta \log(N) \rfloor}^{\lceil (2-\delta)\log(N) \rceil} \mathbb{P}\left(\sum_j \xi_j^1 = k\right) \\
&\leq \mathbb{P}\left(\sum_j \xi_j^1 \sum_{\mu=2}^{M} \xi_i^\mu \xi_j^\mu \geq \gamma \log(N) \Big| \sum_j \xi_j^1 = \lceil (2-\delta)\log(N) \rceil\right),
\end{aligned}$$

since the quantity $\sum_j \xi_j^1 \sum_{\mu=2}^{M} \xi_i^\mu \xi_j^\mu$ is increasing with $\sum_j \xi_j^1$, and the maximum is attained for $k = \lceil (2-\delta)\log(N) \rceil$.

Without loss of generality, $(2-\delta)\log(N) \in \mathbb{N}$ and $\xi_j^1 = 1$, $1 \leq j \leq (2-\delta)\log(N)$; $\xi_j^1 = 0$, $j > (2-\delta)\log(N)$. Then, for each $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sum_j \xi_j^1 \sum_{\mu=2}^{M} \xi_i^\mu \xi_j^\mu \geq \gamma \log(N) \Big| \sum_j \xi_j^1 = (2-\delta)\log(N)\right) \\
&= \mathbb{P}\left(\sum_{j=1}^{(2-\delta)\log(N)} \sum_{\mu=2}^{M} \xi_i^\mu \xi_j^\mu \geq \gamma \log(N)\right) \\
&\leq e^{-t\gamma \log(N)} \mathbb{E} \exp\left(t \sum_{j=1}^{(2-\delta)\log(N)} \sum_{\mu=2}^{M} \xi_i^\mu \xi_j^\mu\right)
\end{aligned}$$

$$
\begin{aligned}
&= e^{-t\gamma \log(N)} \left[\mathbb{E} \exp\left(t \sum_{j=1}^{(2-\delta)\log(N)} \xi_i^2 \xi_j^2 \right) \right]^{M-1} \\
&= e^{-t\gamma \log(N)} \left(1 - p + p\left(1 - p + pe^t\right)^{(2-\delta)\log(N)}\right)^{M-1} \\
&\le e^{-t\gamma \log(N)} \left(1 - p + pe^{p(e^t-1)(2-\delta)\log(N)}\right)^{M-1} \\
&\le \exp\left[-t\gamma \log(N) + (M-1)p\left(e^{p(e^t-1)(2-\delta)\log(N)} - 1\right)\right] \\
&= \exp\left[-t\gamma \log(N) + (M-1)p\left(p(e^t-1)(2-\delta)\log(N) + \mathcal{O}(\log(N)p^2)\right)\right] \\
&= \exp\left[-t\gamma \log(N) + Mp^2(e^t-1)(2-\delta)\log(N) + \mathcal{O}(M\log(N)p^3)\right],
\end{aligned}
$$

using $1 + u \le e^u$ for all $u \ge 0$, expanding the exponential and assuming t to be small.

Assuming $M = \alpha N^2/\log(N)^2$, we obtain that the last line is equal to

$$
\begin{aligned}
&\exp\left[-t\gamma \log(N) + Mp^2(e^t-1)(2-\delta)\log(N) + \mathcal{O}(M\log(N)p^3)\right] \\
&\quad = \exp\left[-t\gamma \log(N) + \alpha(e^t-1)(2-\delta)\log(N) + \mathcal{O}(\log(N)^2/N)\right] \\
&\quad = \exp\left[\log(N)(-t\gamma + \alpha(e^t-1)(2-\delta))\right](1 + o(1)).
\end{aligned}
$$

The function $-t\gamma + \alpha(e^t - 1)(2 - \delta)$ takes its minimum at $t = \log(\gamma/(\alpha(2-\delta)))$.

We aim at showing

$$
\mathbb{P}\left(\exists 1 \le i \le N, T_i(\xi^1) \ne \xi_i^1 \middle| \left|\log(N) - \sum_j \xi_j^1\right| < (1-\delta)\log(N) \right) \to 0.
$$

Following the lines above, this probability can be estimated by

$$
\begin{aligned}
&\mathbb{P}\big(\exists 1 \le i \le N, T_i(\xi^1) \ne \xi_i^1 \big| \left|\log(N) - \sum_j \xi_j^1\right| < (1-\delta)\log(N)\big) \\
&\quad \le \mathbb{P}\big(\exists 1 \le i \le N, \xi_i^1 = 0, T_i(\xi^1) \ne \xi_i^1 \big| \sum_j \xi_j^1 = (2-\delta)\log(N)\big) \\
&\quad \le N \cdot \exp\left[\log(N)(-t\gamma + \alpha(e^t-1)(2-\delta))\right] \\
&\quad \le N \cdot \exp\left[\log(N)(-\gamma\log(\gamma/((2-\delta)\alpha)) + \alpha(2-\delta)(\gamma/(\alpha(2-\delta)) - 1))\right] \\
&\quad = N \cdot \exp\left[\log(N)(-\gamma\log(\gamma/((2-\delta)\alpha)) + \gamma - \alpha(2-\delta))\right]
\end{aligned}
$$

and we need

$$
\gamma \log(\gamma/((2-\delta)\alpha)) - \gamma + \alpha(2-\delta) > 1,
$$

which is fulfilled if

$$
\alpha < \frac{\gamma}{2-\delta}\frac{1}{e^{1+1/\gamma}}.
$$

So for each $\alpha < e^{-2}$, we can find a threshold $h = \gamma \log(N)$ such that

$$
\mathbb{P}\left(\exists 1 \le i \le N, T_i(\xi^1) \ne \xi_i^1\right) \to 0.
$$

This proves the first part of the theorem.

For the second part notice that any fixed ξ^μ will have almost $\log N$ 1's such that we can delete $\rho \log N$ many of them and the statement of the theorem makes sense. The rest of the proof of part two consists of choosing h now as a value slightly smaller than $(1-\rho)\log N$ and repeating the above arguments. Indeed, call $\tilde{\xi}^1$ a configuration obtained from ξ^1 when deleting $\rho \log N$ 1's. Then, as above, the local field $S_i(\tilde{\xi}^1)$ splits into a signal term and a noise term:

$$S_i(\tilde{\xi}^1) = \sum_{j\neq i} \tilde{\xi}_j^1 J_{ij} = \xi_i^1 \sum_{j\neq i} \tilde{\xi}_j^1 + \sum_{j\neq i} \tilde{\xi}_j^1 \sum_{\mu\geq 2} \xi_i^\mu \xi_j^\mu.$$

In comparison to the first part of the proof the ingredient $\sum_{j\neq i} \tilde{\xi}_j^1$ of the signal term is decreased to a size of $(1-\rho)\log N$, while the noise term $\sum_{j\neq i} \tilde{\xi}_j^1 \sum_{\mu\geq 2} \xi_i^\mu \xi_j^\mu$ is treated in a similar fashion as in part one and is typically of order $\alpha(1-\rho)(\log N)$.

For the third statement of the theorem we will make use of an observation that is also useful in the proof of Theorem 4.3 and will actually be shown in this context: For a message (without loss of generality ξ^1) with active neurons $\xi_1^1 = \ldots = \xi_c^1 = 1$ and $\xi_i^1 = 0$ for all $i \geq c$ we show that for M large enough, i.e. $M = \alpha N^2/(\log N)^2$ and $\alpha > -\log(1-e^{-1})$ with probability converging to 1, there exists an $i \geq c+1$ such that for all $j \leq c$ there is a $\mu \geq 2$ such that $\xi_i^\mu \xi_j^\mu = 1$.

After borrowing this statement from the proof of Theorem 4.3 we can proceed as follows: Taking into account that with overwhelming probability c is larger than $(1-\delta)\log N$ for any $\delta > 0$ and N large enough, we see that in Amari's model for such an $i \geq c+1$

$$\begin{aligned} T_i(\xi^1) &= \Theta\left(\sum_{j\neq i} J_{ij}\xi_j^1 - \gamma \log N\right) \\ &= \Theta\left(\sum_{j\leq c} J_{ij} - \gamma \log N\right) \\ &\geq \Theta\left((1-\delta)\log N - \gamma \log N\right) = 1 \end{aligned}$$

if we choose $1-\delta > \gamma$. As we can choose $\delta > 0$ arbitrarily small, with any threshold $\gamma \log N$ with $\gamma < 1$ such a neuron will not be recovered correctly.

Proof of Corollary 4.2 The only thing one has to observe is that for each i with $\xi_i^1 = 1$ we again have $T_i(\xi^1) = 1$, because again $\sum_j \xi_j^1 \geq \gamma \log(N)$ for any $\gamma < 1$.

On the other hand for each i with $\xi_i^1 = 0$ we have that the probability that ξ_i^1 is turned into a 1 by the dynamics and thus not recovered correctly is given by $\mathbb{P}(\sum_j J_{ij}\xi_j^1 \geq \gamma \log(N))$. Now,

$$\sum_j J_{ij}\xi_j^1 < \sum_j \xi_j^1 \sum_{\mu\geq 2} \xi_i^\mu \xi_j^\mu$$

and the right hand side is the quantity considered in the previous proof. Thus the bound obtained in the previous proof is also a bound for the Willshaw model with threshold dynamics.

Remark 5.1 Of course, the previous proof underestimates the storage capacity of the Willshaw model with threshold dynamics. However, the difference between J_{ij} and $\sum_{\mu\geq 2} \xi_i^\mu \xi_j^\mu$ is not that huge. Indeed, for $M = \alpha N^2/(\log N)^2$ the latter is close to a Poisson random variable with parameter α and we will see in the next theorem, that even with a better performing dynamics we only reach a bound of $\alpha \leq 0.45$.

We continue with the Willshaw model with WTA dynamics.

Proof of Theorem 4.3 We start with proving the third statement of the theorem. This will automatically yield the first part by setting ρ to 0.

Using the same method as in the proof of Theorem 4.1, we can restrict the proof to the cases where c_1 neurons in the message ξ^1 are active, with $c_1 \in [(1-\varepsilon_1)c, (1+\varepsilon_1)c]$, for some small $\varepsilon_1 > 0$. Assume that f of the "1"-bits in ξ^1 are erased and $k = c_1 - f$ "1"s are known. Without loss of generality, we can assume that $\xi_i^1 = 1$ for $i \le c_1$ and $\xi_i^1 = 0$ for $i \ge c_1 + 1$.

Let $\tilde{\xi}^1 \in \{0,1\}^N$ be a version of ξ^1 corrupted as described above, such that $\tilde{\xi}_i^1 = 1$ for $i \le k$ and $\tilde{\xi}_i^1 = 0$ for $i \ge k+1$. We have trivially that,

$$h_i(\tilde{\xi}^1) = \sum_{j=1}^{k} J_{ij},$$

and thus $h_i(\tilde{\xi}^1) = k$ for all $i \le c_1$. Therefore $y = T(\tilde{\xi}^1)$ will satisfy $y_i = \xi_i^1$ for all $i \le c_1$.

Thus recalling the WTA we see that $y \ne \xi^1$, if there exist $i \ge c_1 + 1$, such that for all $j \le k$ there exists $\mu \ge 2$ such that $\xi_i^\mu \xi_j^\mu = 1$.

The probability of the latter event can be bounded as follows. Consider

$$\begin{aligned}
&\mathbb{P}[\exists i \ge c_1 + 1, \forall j \le k : \exists \mu \ge 2, \xi_i^\mu \xi_j^\mu = 1] \\
&\le N \sum_{l=0}^{M-1} \sum_{\substack{I \subset \{2,\dots,M\} \\ \mathrm{card}(I)=l}} \mathbb{P}[\forall j \le k : \exists \mu \ge 2, \xi_N^\mu \xi_j^\mu = 1 | \xi_N^\mu = 1 \Leftrightarrow \mu \in I] \mathbb{P}[\xi_N^\mu = 1 \Leftrightarrow \mu \in I] \\
&\le N \sum_{l=0}^{M-1} \sum_{I} \mathbb{P}[\forall j \le k : \exists \mu \in I, \xi_j^\mu = 1] \mathbb{P}[\xi_N^\mu = 1 \Leftrightarrow \mu \in I] \\
&= N \sum_{l=0}^{M-1} \binom{M-1}{l} \left(1 - \left(1 - \frac{c}{N}\right)^l\right)^k \left(\frac{c}{N}\right)^l \left(1 - \left(\frac{c}{N}\right)\right)^{M-l-1} \\
&= N \sum_{l=0}^{M-1} \binom{M-1}{l} \sum_{i=0}^{k} \binom{k}{i} (-1)^i \left(1 - \frac{c}{N}\right)^{il} \left(\frac{c}{N}\right)^l \left(1 - \left(\frac{c}{N}\right)\right)^{M-l-1} \\
&= N \sum_{i=0}^{k} \binom{k}{i} (-1)^i \left(1 - \frac{c}{N} + \frac{c}{N}\left(1 - \frac{c}{N}\right)^i\right)^{M-1}
\end{aligned}$$

by elementary transformations.

Now we expand the term in the brackets and use the bound $1 + x \le e^x$ for all x to obtain

$$\begin{aligned}
&\mathbb{P}[\exists i \ge c_1 + 1, \forall j \le k : \exists \mu \ge 2, \xi_i^\mu \xi_j^\mu = 1] \\
&\le N \sum_{i=0}^{k} \binom{k}{i} (-1)^i \left(1 - i\left(\frac{c}{N}\right)^2 + \frac{i(i-1)}{2}\left(\frac{c}{N}\right)^3 + \mathcal{O}\left(i^3 \left(\frac{c}{N}\right)^4\right)\right)^{M-1} \\
&\le N \sum_{i=0}^{k} \binom{k}{i} (-1)^i \exp\left(-iM\left(\frac{c}{N}\right)^2 + M\frac{i(i-1)}{2}\left(\frac{c}{N}\right)^3 + \mathcal{O}\left(Mi^3 \left(\frac{c}{N}\right)^4\right)\right)
\end{aligned}$$

$$\begin{aligned}
&= N\sum_{i=0}^{k}\binom{k}{i}(-1)^i e^{-iM(\frac{c}{N})^2}\left(1+M\frac{i(i-1)}{2}\left(\frac{c}{N}\right)^3+\mathcal{O}\left(Mi^3\left(\frac{c}{N}\right)^4\right)\right)\\
&\leq N\left(1-e^{-M(\frac{c}{N})^2}\right)^k+MN\left(\frac{c}{N}\right)^3\sum_{i=0}^{k}\binom{k}{i}(-1)^i e^{-iM(\frac{c}{N})^2}\frac{i(i-1)}{2}\\
&\qquad +N(1+e^{-M(\frac{c}{N})^2})^k\mathcal{O}\left(Mk^3\left(\frac{c}{N}\right)^4\right)\\
&= N(1-e^{-M(\frac{c}{N})^2})^k+MN(\frac{c}{N})^3e^{-2M(\frac{c}{N})^2}\frac{k(k-1)}{2}(1-e^{-M(\frac{c}{N})^2})^{k-2}\\
&\qquad +N(1+e^{-M(\frac{c}{N})^2})^k\mathcal{O}\left(Mk^3\left(\frac{c}{N}\right)^4\right).
\end{aligned}$$

If we choose

$$M=\alpha\left(\frac{N}{c}\right)^2 \quad \text{and } k=(1-\rho)\log(N) \quad \text{for some } \rho\in[0,1[$$

we arrive at

$$\begin{aligned}
&\mathbb{P}[\exists i\geq c_1+1,\forall j\leq k:\exists\mu\geq 2,\xi_i^{\mu}\xi_j^{\mu}=1]\\
&\leq N(1-e^{-\alpha})^{(1-\rho)\log(N)}+\alpha e^{-2\alpha}(\log N)^3(1-e^{-\alpha})^{(1-\rho)\log N-2}\\
&\qquad +(1+e^{-\alpha})^{(1-\rho)\log(N)}\mathcal{O}(\frac{(\log N)^5}{N}).
\end{aligned}$$

If $(1-\rho)\log(1-e^{-\alpha})<-1$, i.e. $\alpha<-\log(1-e^{-1/(1-\rho)})$, the first term converges to 0 and the two last terms also vanish for $N\to\infty$. This gives

$$\mathbb{P}[\exists i\geq c_1+1,\forall j\leq k:\exists\mu\geq 2,\xi_i^{\mu}\xi_j^{\mu}=1]\to 0$$

as desired.

It remains to prove the reverse bound on the storage capacity. The considerations are similar to what we did above. Now assume that $M\geq\alpha(\frac{N}{c})^2$ for some $\alpha>0$ and again that ξ^1 has entries $\xi_i^1=1$ for $i=1,\dots c_1$ and $\xi_i^1=0$ for $i>c_1$.

Again consider

$$\begin{aligned}
&\mathbb{P}[\exists i\geq c_1+1,\forall j\leq k:\exists\mu\geq 2,\xi_i^{\mu}\xi_j^{\mu}=1]\\
&=1-\mathbb{P}\Big[\bigcap_{i\geq c_1+1}\{\exists j\leq k:\forall\mu\geq 2,\xi_i^{\mu}\xi_j^{\mu}=0\}\Big]\\
&=1-\mathbb{P}_{\{\xi_j^{\mu},j\leq k,\mu\geq 2\}}\prod_{i=c_1+1}^{N}\mathbb{P}_{\{\xi_i^{\mu},\mu\geq 2\}}\Big[\exists j\leq k:\sum_{\mu\geq 2}\xi_i^{\mu}\xi_j^{\mu}=0\Big]
\end{aligned}$$

by independence after conditioning (and the $\mathbb{P}_{\{\xi_j^{\mu}\}}$ denote the probabilities with respect to the corresponding random variables). Now

$$\mathbb{P}_{\{\xi_i^{\mu},\mu\geq 2\}}\Big[\exists j\leq k:\sum_{\mu\geq 2}\xi_i^{\mu}\xi_j^{\mu}=0\Big]=1-\mathbb{P}_{\{\xi_i^{\mu},\mu\geq 2\}}\Big[\forall j\leq k:\sum_{\mu\geq 2}\xi_i^{\mu}\xi_j^{\mu}\geq 1\Big]$$

Let $X_j := \sum_{\mu \geq 2} \xi_i^\mu \xi_j^\mu$. We observe by similar arguments as in Sect. 3 that the (X_j) are positively associated with respect to $\mathbb{P}_{\{\xi_i^\mu, \mu \geq 2\}}$. Therefore, for $i \geq c_1 + 1$,

$$\mathbb{P}_{\{\xi_i^\mu, \mu \geq 2\}}[\forall j \leq k : X_j \geq 1] \geq \prod_{j=1}^{k} \left(\mathbb{P}_{\{\xi_i^\mu, \mu \geq 2\}}[X_j \geq 1] \right)$$

which gives

$$\begin{aligned} &\mathbb{P}[\exists i \geq c_1 + 1, \forall j \leq k : \exists \mu \geq 2, \xi_i^\mu \xi_j^\mu = 1] \\ &\geq 1 - \mathbb{P}_{\{\xi_j^\mu, j \leq k, \mu \geq 2\}} \prod_{i=c_1+1}^{N} \left(1 - \prod_{j=1}^{k} (\mathbb{P}_{\{\xi_i^\mu, \mu \geq 2\}}[X_j \geq 1]) \right) \end{aligned}$$

To compute the right hand side take e.g. $i = N$. Then for all $j \leq k$,

$$\begin{aligned} \mathbb{P}_{\{\xi_N^\mu, \mu \geq 2\}}[X_j \geq 1]) &= 1 - \mathbb{P}_{\{\xi_N^\mu, \mu \geq 2\}} \left(\sum_{\mu=1}^{M} \xi_N^\mu \xi_j^\mu = 0 \right) \\ &= 1 - \prod_{\mu : \xi_j^\mu = 1} \mathbb{P}_{\{\xi_N^\mu, \mu \geq 2\}} \left(\xi_N^\mu = 0 \right) \\ &= 1 - \left(1 - \frac{c}{N} \right)^{W_j}, \end{aligned}$$

where $W_j := \sum_{\mu=1}^{M} \xi_j^\mu$. With overwhelming probability

$$W_j \in \left[(1 - \varepsilon) \frac{Mc}{N}, (1 + \varepsilon) \frac{Mc}{N} \right]$$

for all N large enough, for all $j \leq k$. More precisely, for all $\varepsilon > 0$, $k = C \log(N)$, with $C > 0$,

$$\mathbb{P}\left[\forall j \leq k : W_j \in \left[(1 - \varepsilon) \frac{Mc}{N}, (1 + \varepsilon) \frac{Mc}{N} \right] \right] \geq 1 - 2C \log(N) e^{-\frac{Mc\varepsilon^2}{2N}}.$$

This justifies that we can restrict to these cases, and putting things together, we obtain for $M = \alpha(\frac{N}{c})^2$ that

$$\mathbb{P}[\exists i \geq c_1 + 1, \forall j \leq k : \exists \mu \geq 2, \xi_i^\mu \xi_j^\mu = 1] \geq 1 - \left(1 - \left(1 - e^{-\alpha} \right)^k \right)^{N - k_1}.$$

The right hand side converges to 1 if $\left(1 - \left(1 - e^{-\alpha} \right)^k \right)^N$ goes to 0, which is the case if and only if

$$\begin{aligned} N \log \left(1 - \left(1 - e^{-\alpha} \right)^k \right) &\approx -N \left(1 - e^{-\alpha} \right)^k \\ &= -N \exp \left(k \log(1 - e^{-\alpha}) \right) \\ &= -N^{1 + (1-\rho) \log(1 - e^{-\alpha})} \to -\infty. \end{aligned}$$

This is true if $1 + (1 - \rho) \log(1 - e^{-\alpha}) > 0$, which is true if and only if

$$\alpha > -\log(1 - e^{-1/(1-\rho)}).$$

This finishes the proof. □

Remark 5.2 Note that the previous proof reveals that not only we have upper and lower bounds on the storage capacity of the Willshaw model with WTA dynamics, but also that these bounds match. Such matching bounds can very rarely be proven. The only other model we are aware of where this is the case, is the Hopfield model (see [4,19]).

Proof of Theorem 4.5 The decisive observation here is that the GB model is "almost" a Willshaw model. As a matter of fact, as stated already in the description of the model in Sect. 2, the only difference is, that in the GB model there is a restriction on the location of the 1's. However, if we analyze the proof of Theorem 4.3, we find that the dynamics is in a sense ”non-spatial", i.e. a neuron is getting signals from all the other neurons in both of these models. Thus this detail does not influence the proof.

This observation, however, also raises the question, whether we can also prove the third statement in Theorem 4.3 for the GB model. It is, indeed, natural to conjecture that a similar statement holds true. However, in the proof of part 3 of Theorem 4.3 we make use of positive association. This property enters the proof in the Willshaw model, because with our setting we are having increasing functions of i.i.d. random variables (the spins ξ_i^μ), that are indeed positively associated. In the GB model, the extra condition that each pattern has exactly one 1 in each of the blocks implies that for each fixed μ the random variables $\xi_{(a,k)}^\mu$ are no longer independent. Hence we do not have positive association. □

6 The Wrong Message Revisited: A Limit of all Reconstruction Techniques

In this section we return to the question addressed in Sect. 3. There we showed that in the GB model with M too large a wrong message will be recognized with large probability as a correct one, which limits the confidence we can have into our associative memory.

A very similar consideration shows that we cannot reconstruct erased messages in the GB model, if M is too large. Indeed, in the GB model suppose we delete at random a proportion of $(1-\rho)c$ of active bits of a given message. If the remaining bits can be completed in more than one way to a message that is recognized by the system (N.B. not necessarily a message that is stored in the network), there is no way whatsoever, a reconstruction algorithm could find the correct message with probability one.

Using ideas from Sect. 3 one can prove a theorem on the probability to complete an erased message by a message on a given set of neurons. To formulate it, suppose that a message ξ^1 is stored in the network. Without loss of generality $\xi_{(a,1)}^1 = 1$ for all clusters $1 \le a \le c$ and all the other bits are 0. Assume we keep the $\xi_{(a,1)}^1 = 1$ for the clusters $1 \le a \le \rho c$, $0 < \rho < 1$ and set all other neurons to 0. Then for each cluster $\rho c + 1 \le a \le c$ we choose a neuron (a, i), $2 \le i \le l$ and set it to 1. Let G be the event that the message ζ having 1's in position $(a, 1)$, $1 \le a \le \rho c$ and (a, i) for $\rho c + 1 \le a \le c$ is recognized by the system as a stored message.

Theorem 6.1 *Suppose that in the GB model we store $M = \alpha l^2 \log c$ messages. Then $\mathbb{P}(G)$ tends to 0 if and only if $\alpha < 2$.*

Proof We only sketch the proof here as it is almost identical to the considerations in Sect. 3.

Other than there, we already know ρc bits of ζ are correct. Hence we only need to find messages that are active on the remaining $r(c,\rho) := \rho(1-\rho)c^2 + (1-\rho)c((1-\rho)c-1)/2 = \frac{c^2}{2}(1-\rho^2) - \frac{1}{2}c(1-\rho)$ edges.

Positive association bounds thus $\mathbb{P}(G)$ by $(1-(1-\frac{1}{l^2})^M)^{r(c,\rho)} =: d^{\,r(c,\rho)}$ from below. The same exponential inequality as in Sect. 3 also shows an upper bound for $\mathbb{P}(G)$ by $d^{\,r(c,\rho)}$

plus a vanishing term. Replacing $(1-\frac{1}{l^2})^M$ by $c^{-\alpha}$ we thus see that $d^{\,r(c,\rho)}$ is of order $\exp(-\frac{c^{2-\alpha}}{2}(1-\rho^2))$ and therefore goes to zero, if and only if, $c^{2-\alpha}(1-\rho^2)\to\infty$.

Remark 6.2 Similarly to Theorem 3.1, we get that $\mathbb{P}(G)$ is well approximated by $d^{\frac{c^2}{2}(1-\rho^2)}$, when the latter goes to 0, for $\alpha\in]1,2[$. This is not the case for $\alpha\in]0,1[$, since the additive error term in the upper bound vanishes, but slower than $d^{\frac{c^2}{2}(1-\rho^2)}$.

7 Dynamical Properties of the Models

An interesting question is the convergence of the proposed dynamics. Recall that we distinguish two types of dynamics: a) fixed threshold ones where h is fixed a priori and b) varying threshold ones where h is updated at each iteration of the dynamics (e.g. WTA). Note that in all cases we consider the memory effect described in Sect. 7.1.

Let us first consider the Willshaw model.

7.1 Willshaw Model

In this section we show the following results:

(1) Choosing a fixed h forces convergence of the dynamics,
(2) Choosing a varying h can lead to oscillations in the dynamics,
(3) Choosing the threshold $h_{(1)}$ as defined in Sect. 2, performance does not benefit from iterating more than once the dynamics.

Note that the major interest of varying thresholds is that they lead to better performance as illustrated in Sect. 7.3. There thus exists a tradeoff between performance and convergence guarantees for the Willshaw model.

Theorem 7.1 *Choosing a fixed threshold h forces the dynamics to converge.*

Proof Let us consider an input pattern $\tilde{\xi}^\mu$ where some 1s have been erased. Denote $c_\mu=\|\tilde{\xi}^\mu\|_0$ to be the number of 1's in $\tilde{\xi}^\mu$. Then it is immediate that if $h>c_\mu$ the dynamics converges in one iteration to a null vector.

On the other hand, let us introduce the sequence $\left(\tilde{\xi}^\mu(t)\right)_{t\geq 0}$:

$$\begin{aligned}\tilde{\xi}^\mu(0) &:= \tilde{\xi}^\mu\\ \tilde{\xi}^\mu(t+1) &:= T\left(\tilde{\xi}^\mu(t)\right) \qquad \text{and} \quad \text{for all } t\in\mathbb{N},\end{aligned}$$

and the sequence $(a^\mu(t))_{t\geq 0}$ such that $a^\mu(t)=\{i,\tilde{\xi}^\mu_i(t)=1\}$ for all $t\in\mathbb{N}_0$.

We now can show the following proposition:

Proposition 7.2 *If $h\leq c_\mu$, the sequence $(a^\mu(t))_{t\geq 0}$ is nondecreasing with respect to inclusion.*

Proof Let us proceed by induction.

First we have trivially that $a^\mu(0)\subseteq a^\mu(1)$. This is due to the fact that $\forall i,j\in a^\mu(0)$, $J_{ij}=1$.

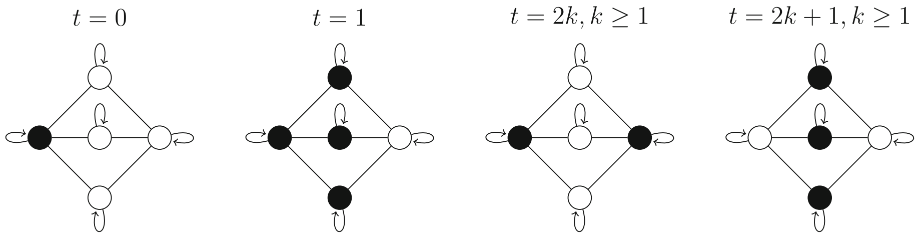

Fig. 1 Illustration of the oscillation of the dynamics when using WTA with the Willshaw model. Here the model contains $N=5$ neurons and the number of 1s in stored messages is $c=2$

Then let us suppose that for some t we have $a^\mu(t) \subseteq a^\mu(t+1)$. By definition, $\forall i \in a^\mu(t+1)$, we have $\#\{j \in a^\mu(t), J_{ij}=1\} \geq h$, where # denotes the cardinality operator.

Since $a^\mu(t) \subseteq a^\mu(t+1)$, it also holds that $\#\{j \in a^\mu(t+1), J_{ij}=1\} \geq h$ and we conclude that $a^\mu(t+1) \subseteq a^\mu(t+2)$.

A direct corollary is that $(a^\mu(t))_{t\geq 0}$ converges.

Theorem 7.3 *Choosing a varying h can lead to oscillations in the dynamics of the Willshaw model.*

Proof To illustrate this property, we propose an example where $N=5$ and $c=2$. We choose the threshold $h_{(1)}$ as defined in Sect. 2. Let us consider that:

$$\left(\xi^\mu\right)_{1\leq\mu\leq 6} = \left(\begin{pmatrix}1\\1\\0\\0\\0\end{pmatrix}\begin{pmatrix}1\\0\\1\\0\\0\end{pmatrix}\begin{pmatrix}1\\0\\0\\1\\0\end{pmatrix}\begin{pmatrix}0\\1\\0\\0\\1\end{pmatrix}\begin{pmatrix}0\\0\\1\\0\\1\end{pmatrix}\begin{pmatrix}0\\0\\0\\1\\1\end{pmatrix}\right).$$

Consider the input:

$$\tilde{\xi}^\mu(0) = \begin{pmatrix}1\\0\\0\\0\\0\end{pmatrix}.$$

One can easily check that:

$$\left(\tilde{\xi}^\mu(t)\right)_{0\leq t\leq 4} = \left(\begin{pmatrix}1\\0\\0\\0\\0\end{pmatrix}\begin{pmatrix}1\\1\\1\\1\\0\end{pmatrix}\begin{pmatrix}1\\0\\0\\0\\1\end{pmatrix}\begin{pmatrix}0\\1\\1\\1\\0\end{pmatrix}\begin{pmatrix}1\\0\\0\\0\\1\end{pmatrix}\right),$$

and thus $\tilde{\xi}^\mu(4) = \tilde{\xi}^\mu(2)$.

The same dynamics is illustrated in Fig. 1.

More generally, using the threshold $h_{(1)}$ as defined in Sect. 2, the performance of the model does not benefit from using more than one iteration, as expressed in the following theorem:

Theorem 7.4 *Consider a Willshaw network where the threshold is chosen as the maximum of the achieved scores ($h_{(1)}$). Choose as input a partially (but not completely) erased version $\tilde{\xi}^\mu$ of a stored message ξ^μ. Then the dynamics converges if and only if it converges in one step. In particular, it can only converge to ξ^μ if it does so in one iteration.*

Proof Let us use the same notations as in the proof of Theorem 7.1. We denote by $h_{(1)}(t)$ the value of the threshold at step t.

Let us discuss two cases:

(1) In the first case it holds for all i and j that $\tilde{\xi}_i^\mu(1) = 1$ and $\tilde{\xi}_j^\mu(1) = 1$ implies that $J_{ij} = 1$. In other words: All activated neurons after one iteration are connected one to another. In this case one can easily check that

$$h_{(1)}(1) = \text{card}(\{i, \tilde{\xi}_i^\mu(1) = 1\})$$

and thus we have for all $t \geq 1$ that $\tilde{\xi}^\mu(1) = \tilde{\xi}^\mu(t)$.

(2) There are i' and j' such that $\tilde{\xi}_{i'}^\mu(1) = 1$ and $\tilde{\xi}_{j'}^\mu(1) = 1$ but $J_{i'j'} = 0$, i.e. there are activated neurons that are not interconnected.
Note that by construction of J we then cannot have that $\tilde{\xi}^\mu(1) = \xi^\mu$. We fix such a pair i' and j'. By construction of J, we have for all i and j that $\xi_i^\mu = 1$ and $\xi_j^\mu = 1$ implies that $J_{ij} = 1$. As a direct consequence, we obtain that

$$h_{(1)}(0) = \text{card}(\{i, \tilde{\xi}_i^\mu(0) = 1\})$$

and therefore all neurons activated at step 0 are connected to all neurons activated at step 1 (note also that $\{i, \xi_i^\mu = 1\} \subsetneq \{i, \tilde{\xi}_i^\mu(1) = 1\}$). Thus we obtain

$$h_{(1)}(1) = \text{card}(\{i, \tilde{\xi}_i^\mu(1) = 1\}).$$

Consequently, $\tilde{\xi}_{i'}^\mu(2) = 0$ and $\tilde{\xi}_{j'}^\mu(2) = 0$ which leads to $\tilde{\xi}^\mu(1) \neq \tilde{\xi}^\mu(2)$. We conclude that the neurons activated in $\tilde{\xi}^\mu(2)$ are those connected to all neurons in $\tilde{\xi}^\mu(1)$. In particular we obtain $\{i, \tilde{\xi}_i^\mu(0) = 1\} \subseteq \{i, \tilde{\xi}_i^\mu(2) = 1\}$.
Similarly we have that $h_{(1)}(2) = \text{card}(\{i, \tilde{\xi}_i^\mu(2) = 1\})$.
We then observe that there cannot be a neuron active at step 3 that is not active at step 1, as the neurons activated at step 3 are connected to all neurons activated at step 2 and thus to all neurons activated at step 0. We conclude that for all $t \geq 1$ we have that

$$\tilde{\xi}^\mu(2i-1) = \tilde{\xi}^\mu(2i+1) \quad \text{and} \quad \tilde{\xi}^\mu(2i) = \tilde{\xi}^\mu(2i+2),$$

together with $\tilde{\xi}^\mu(1) \neq \tilde{\xi}^\mu(2)$.

□

7.2 GB Model

Interestingly, the specific GB structure can be exploited in order to provide good performance and to ensure at the same time convergence of the dynamics. This is thanks to the previously mentioned SUM-OF-MAX rule (see Equation (1)). Recall the SUM-OF-MAX dynamic rule:

$$T_{(a,k)}(\sigma) = \Theta(s_{(a,k)}(\sigma) - h(a)), \text{ where } h(a) = \max\{s(a,k), k = 1, \ldots, l\}.$$

This rule can be advantageously combined with a modification of the input when retrieving a partially erased image. This modification consists in activating all neurons in clusters where no neuron is active. Then we have trivially $h(a) = c$ for all a and this modification is such that the set of active neurons is non-increasing with iterations of the dynamics.

Here is a rapid sketch of the proof of this result: to be activated using the SUM-OF-MAX rule, a neuron has to be connected to at least one activated neuron in each cluster. In particular it has to be connected to an activated neuron in its own cluster. Due to the specific structure of the GB model, the only connection a neuron may have with a neuron in its own cluster is with itself. Therefore, to be activated, a neuron has to already be activated at the previous step of the dynamics.

We refer to this algorithm as "SOM" in Fig. 2.

7.3 Simulations

In order to compare the performance of the three above mentioned solutions, we run several simulations. We consider that the number of 1s in each message is c for the Willshaw model.

We propose to use three different families of algorithms: a) fixed threshold ones where h is determined a priori, b) varying threshold ones where h can be modified at each iteration and c) exhaustive search where solutions are looked for using a brute-force approach. This last option allows us to compare the different models intrinsically, thus removing any bias from chosen retrieval dynamics.

For case a) we define h as the number of 1s in the input pattern. This value appears to be optimal for most cases we simulated. For case b) we use the winner-takes-all algorithm previously described in which we select h so that the number of 1s in the obtained vector is minimum and at least c. For case c) we use an exhaustive search of potential candidates and select randomly one of them. Note that for Amari's model we select the clique (or one of the cliques) that achieve the maximum sum of inner edge weights. Finally, for each case we also plot the obtained curves when using SUM-OF-MAX with the GB model for easier comparison of performance.

We depict the evolution of the error rate for a given problem as a function of the number of stored patterns. This measure is not totally fair as:

- A stored pattern with c 1s using the Willshaw model or Amari's one made of N neurons has entropy $\log_2\left(\binom{N}{c}\right)$ whereas with the GB model its entropy is lesser: $C\log_2(l)$.
- The number of possible connections in a Willshaw model or Amari's one with N neurons is larger than that using a GB model with the same number of neurons. Moreover in the Amari model each connection can take up to M distinct values.

In order to account for these differences, we propose to depict also the evolution of the error rate as a function of the efficiency of the model, defined as the ratio between the entropy of the set of stored patterns and the number C of bits required for straightforward encoding of the used synaptic weights. The latter value C depends on the model parameters: for an Amari model made of N neurons and storing M patterns, it is equal to:

$$C_{Amari} = \binom{N}{2}\log_2(M+1)\ .$$

For the Willshaw model it becomes:

$$C_{Willshaw} = \binom{N}{2}\ .$$

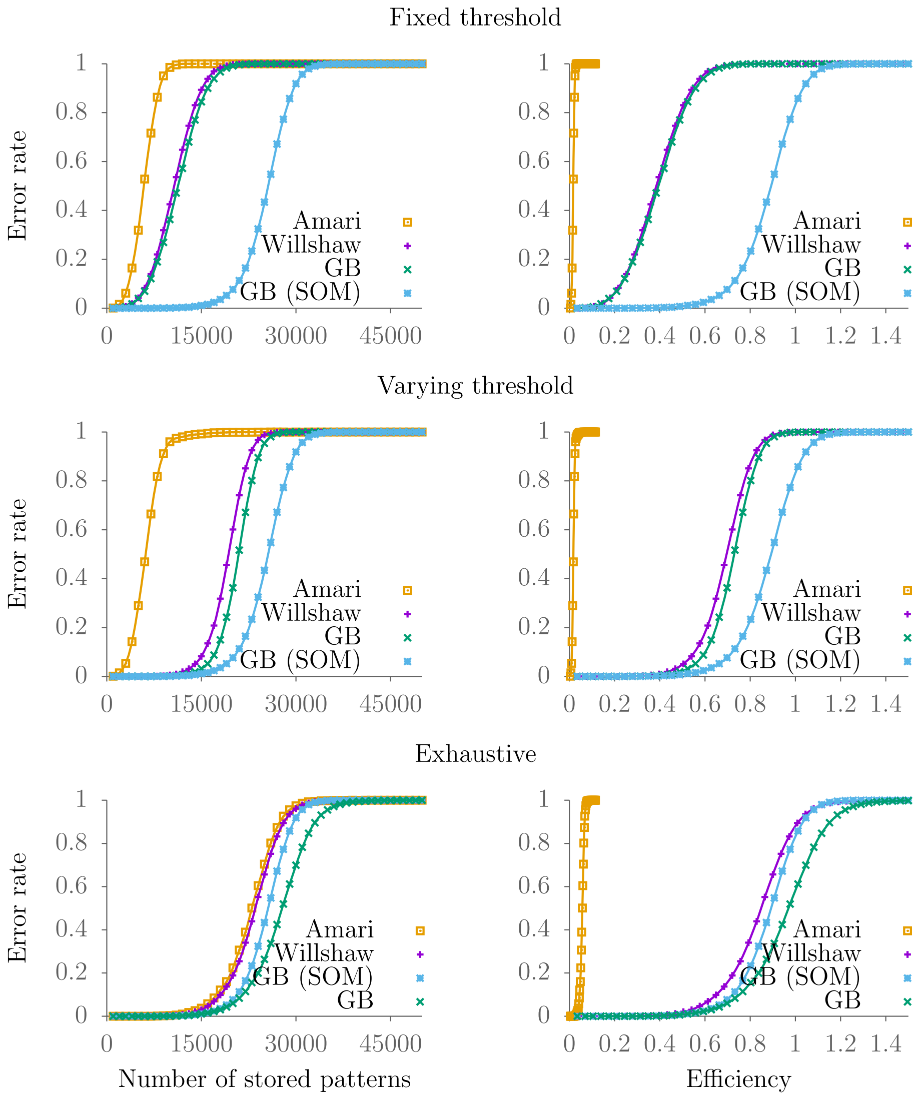

Fig. 2 Comparison of performance of Amari, Willshaw and GB models (with proposed dynamics and SUM-OF-MAX (SOM)). For all simulated point, there are $N = 2048$ neurons (grouped in $c = 8$ clusters of $l = 256$ neurons for the GB model), stored messages contain exactly $c = 8$ 1s each and the objective is to retrieve a previously stored pattern when 4 out of the initial 8 1s in stored messages are missing. Each point is the average of 100.000 tests. Figures in first column depict the evolution of the error rate as a function of the number of stored patterns. Figures in second column depicts the evolution of the error rate as a function of efficiency. First line correspond to fixed threshold dynamics, second line to varying threshold strategies and third line to exhaustive ones

For the GB model, it depends on the parameters c and l and becomes:

$$C_{GB} = \binom{c}{2} l^2 \, .$$

The results are depicted in Fig. 2. Some remarks about these results:

- No matter what algorithms are used, the performance of the methods clearly indicates that GB performs better than Willshaw that performs itself better than Amari's networks.
- The only difference between Amari's networks and Willshaw's is the fact the former use weighted connections instead of binary ones. Our simulations clearly indicates that weights offer no gain in performance.
- It appears clearly that fixed threshold algorithms perform worse than varying threshold ones.

Acknowledgments We are very grateful to two anonymous referees for a very careful reading of a first version of the manuscript and valuable remarks that helped to improve its readability significantly.

References

1. Aliabadi, B.K., Berrou, C., Gripon, V., Jiang, X.: Storing sparse messages in networks of neural cliques. IEEE Trans. Neural Netw. Learn. Syst. **25**, 980–989 (2014)
2. Bollé, D., Verbeiren, T.: Thermodynamics of fully connected Blume-Emery-Griffiths neural networks. J. Phys. A: Math. Gen. **36**(6), 295–305 (2003)
3. Boutsikas, M.V., Koutras, M.V.: A bound for the distribution of the sum of discrete associated or negatively associated random variables. Ann. Appl. Probab. **10**(4), 1137–1150 (2000)
4. Bovier, A.: Sharp upper bounds on perfect retrieval in the Hopfield model. J. Appl. Probab. **36**(3), 941–950 (1999)
5. Burshtein, D.: Nondirect convergence radius and number of iterations of the Hopfield associative memory. IEEE Trans. Inf. Theory **40**(3), 838–847 (1994)
6. Esary, J.D., Proschan, F., Walkup, D.W.: Association of random variables, with applications. Ann. Math. Stat. **38**, 1466–1474 (1967)
7. Gripon, V., Berrou, C.: Sparse neural networks with large learning diversity. IEEE Trans. Neural Netw. **22**(7), 1087–1096 (2011)
8. Heusel, J., Löwe, M., Vermet, F.: On the capacity of an associative memory model based on neural cliques. Stat. Probab. Lett. **106**, 256–261 (2015)
9. Hopfield, J.J.: Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA **79**(8), 2554–2558 (1982)
10. Amari, S.I.: Characteristics of sparsely encoded associative memory. Neural Netw. **2**(6), 451–457 (1989)
11. Jarollahi, H., Gripon, V., Onizawa, N., Gross, W.J.: Algorithm and architecture for a low-power content-addressable memory based on sparse-clustered networks. IEEE Trans. Very Large Scale Integr. Syst. **27**(2), 375–387 (2016)
12. Jarollahi, H., Onizawa, N., Gripon, V., Gross, W.J.: Algorithm and architecture of fully-parallel associative memories based on sparse clustered networks. J. Signal Process. Syst. **76**(3), 235–247 (2014)
13. Jiang, X., Gripon, V., Berrou, C., Rabbat, M.: Storing sequences in binary tournament-based neural networks. IEEE Trans. Neural Netw. Learn. Syst. **27**(5), 913–925 (2016)
14. Löwe, M.: On the storage capacity of Hopfield models with correlated patterns. Ann. Appl. Probab. **8**(4), 1216–1250 (1998)
15. Löwe, M.: On the storage capacity of the Hopfield model with biased patterns. IEEE Trans. Inf. Theory **45**(1), 314–318 (1999)
16. Löwe, M., Vermet, F.: The storage capacity of the Blume-Emery-Griffiths neural network. J. Phys. A **38**(16), 3483–3503 (2005)
17. Löwe, M., Vermet, F.: The capacity of q-state Potts neural networks with parallel retrieval dynamics. Stat. Probab. Lett. **77**(14), 1505–1514 (2007)
18. Löwe, M., Vermet, F.: Capacity of an associative memory model on random graph architectures. Bernoulli **21**(3), 1884–1910 (2015)
19. McEliece, R.J., Posner, E.C., Rodemich, E.R., Venkatesh, S.S.: The capacity of the Hopfield associative memory. IEEE Trans. Inf. Theory **33**(4), 461–482 (1987)
20. Okada, M.: Notions of associative memory and sparse coding. Four major hypotheses in neuroscience. Neural Netw. **9**(8), 1429–1458 (1996)
21. Palm, G.: On associative memory. Biol. Cybern. **36**(1), 19–31 (1980)

22. Palm, G.: Neural associative memories and sparse coding. Neural Netw., **37**(0), 165–171 (2013) (**Twenty-fifth Anniversay Commemorative Issue**)
23. Schwenker, F., Sommer, F., Palm, G.: Iterative retrieval of sparsely coded associative memory patterns. Neural Netw. **9**(3), 445–455 (1996)
24. Willshaw, D.J., Buneman, O.P., Longuet-Higgins, H.C.: Non-holographic associative memory. Nature **222**, 960–962 (1969)
25. Yao, Z., Gripon, V., Rabbat, M.: A GPU-based Associative Memory using Sparse Neural Networks. In: Proceedings of the PCNN-14 conference, pp. 688–692 (2014)