Neural Networks 66 (2015) 22–35

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

# Hierarchical neural networks perform both serial and parallel processing

Elena Agliari [a], Adriano Barra [a], Andrea Galluzzi [b], Francesco Guerra [a,c], Daniele Tantari [b], Flavia Tavani [d,*]

[a] Dipartimento di Fisica, Sapienza Università di Roma, P.le A. Moro 2, 00185, Roma, Italy

[b] Dipartimento di Matematica, Sapienza Università di Roma, P.le A. Moro 2, 00185, Roma, Italy

[c] Istituto Nazionale di Fisica Nucleare, Sezione di Roma, Roma, Italy

[d] Dipartimento SBAI (Ingegneria), Sapienza Università di Roma, Via A. Scarpa 14, 00185, Roma, Italy

ARTICLE INFO

*Article history:*
Received 5 September 2014
Received in revised form 18 February 2015
Accepted 22 February 2015
Available online 2 March 2015

*Keywords:*
Multitasking associative networks
Serial processing
Parallel processing

ABSTRACT

In this work we study a Hebbian neural network, where neurons are arranged according to a hierarchical architecture such that their couplings scale with their reciprocal distance. As a full statistical mechanics solution is not yet available, after a streamlined introduction to the state of the art via that route, the problem is consistently approached through signal-to-noise technique and extensive numerical simulations. Focusing on the low-storage regime, where the amount of stored patterns grows at most logarithmical with the system size, we prove that these non-mean-field Hopfield-like networks display a richer phase diagram than their classical counterparts. In particular, these networks are able to perform serial processing (i.e. retrieve one pattern at a time through a complete rearrangement of the whole ensemble of neurons) as well as parallel processing (i.e. retrieve several patterns simultaneously, delegating the management of different patterns to diverse communities that build network). The tune between the two regimes is given by the rate of the coupling decay and by the level of noise affecting the system.

The price to pay for those remarkable capabilities lies in a network's capacity smaller than the mean field counterpart, thus yielding a new budget principle: the wider the multitasking capabilities, the lower the network load and vice versa. This may have important implications in our understanding of biological complexity.



## 1. Introduction

Statistical mechanics constitutes a powerful technique for the understanding of neural networks (Amit, 1992; Coolen, Kuhn, & Sollich, 2005; Sollich, Tantari, Annibale, & Barra, 2014), however overcoming the mean-field approximation is extremely hard (even beyond neural networks). Basically, the mean-field approximation lies in assuming that each spin/neuron $S_i$ in a network dialogs with *all* the other spin/neurons with the same strength.[1] For instance, if we consider a ferromagnetic model, once introduced $N$ spins $S_i = \pm 1$, $i \in (1, \dots, N)$, we have the two extreme scenarios of a nearest-neighbor model like the Ising lattice, whose Hamiltonian can be written as

$$H_{\text{Ising}} = -\sum_{\langle i,j \rangle} J S_i S_j, \tag{1}$$

where, crucially, the sum runs over all the couples $\langle i, j \rangle$ of *adjacent* sites, and the mean-field Curie–Weiss model, whose Hamiltonian can be written as

$$H_{\text{Curie–Weiss}} = -\sum_{i<j}^{N,N} J S_i S_j, \tag{2}$$

where the sum runs over *all* the $N(N-1)/2$ spin couples irrespective of any notion of distance; this is equivalent to think of spins interacting through nearest neighbor prescriptions but as they were embedded in an $N$-dimensional space. Clearly, solving the

* Corresponding author.
*E-mail address:* flavia.tavani@sbai.uniroma1.it (F. Tavani).

[1] Notice that this situation corresponds to a system embedded in a fully-connected (i.e. complete graph) topology. However, situations where we introduce some degree of dilution (e.g. Erdös–Rényi graph), yet preserving the homogeneity of the structure and an extensive coordination number, can be looked and treated as mean field models.

http://dx.doi.org/10.1016/j.neunet.2015.02.010
0893-6080/

statistical mechanics of the latter model is much simpler with respect to the former. The main route toward finite-dimensional descriptions has been paved by physicists in the study of condensed matter.[2] Indeed, incredible efforts have been spent from the 1970s in working out the renormalization-group (Wilson, 1971a), namely a technique which allows inferring the properties of three-dimensional ferromagnets starting from mean-field descriptions, but a straight solution of the Ising model in dimensions 3 is still out of the current mathematical reach.[3]

Actually, in the last decade some steps forward toward *more realistic* systems have been achieved merging statistical mechanics (Ellis, 1985; Gallavotti & Miracle-Sole', 1967; Mezard, Parisi, & Virasoro, 1987) and graph theory (Albert & Barabasi, 2002; Bollobas, 1998; Watts & Strogatz, 1998). In particular, mathematical methodologies were developed to deal with spin systems embedded in random graphs, where the ideal, full homogeneity among spins is lost (Agliari, Annibale, Barra, Coolen, & Tantari, 2013a, 2013b). Thus, networks of neurons arranged according to Erdös–Rényi (Barra & Agliari, 2008), small-world (Agliari & Barra, 2011), or scale-free (Perez-Castillo et al., 2004) topologies were addressed, yet finite-dimensional networks were still out of debate.

Focusing on neural networks, it should be noted that, beyond the difficulty of treating non-trivial topologies for neuron architecture, one has also to cope with the complexity of their coupling pattern, meant to encode the Hebbian learning rule. The emerging statistical mechanics is much trickier than that for ferromagnets; indeed neural networks can behave either as ferromagnets or as spin-glasses, according to the parameter settings: their phase space is split into several disconnected pure states, each coding for a particular stored pattern, so to interpret the thermalization of the system within a particular energy valley as the spontaneous retrieval of the stored pattern associated to that valley. However in the high-storage limit, where the amount of patterns scales linearly with the number of neurons, neural networks approach pure spin-glasses (losing retrieval capabilities at the blackout catastrophe Amit, 1992) and, as a simple Central Limit argument shows (Barra, Genovese, Guerra, & Tantari, 2012), when the amount of patterns diverge faster that the amount of neurons they become purely spin glasses. For the sake of exhaustiveness we also stress that, even in the retrieval region, neural networks are *exactly* linear combinations of two-party spin glasses (Barra, Contucci, Mingione, & Tantari, 2015; Barra, Genovese, & Guerra, 2010, 2012; Barra, Genovese, Guerra, Tantari et al., 2012; Barra, Genovese, Guerra, & Tantari, 2014): due to the combination of such difficulties, neural networks on a finite dimensional topology have not been extensively investigated so far.

However, very recently, a non-mean-field model, where a topological distance among spins can be defined and couplings can be accordingly rescaled, turned out to be, to some extent, treatable also for complex systems such as spin-glasses (Castellana & Parisi, 2011; Monthus & Garel, 2014). More precisely, spins are arranged according to a hierarchical architecture as shown in Fig. 1: each pair of nearest-neighbor spins form a "dimer" connected with the strongest coupling, then spins belonging to nearest "dimers" interact each other with a weaker coupling and so on recursively (Mukamel, 2008). In particular, the Sherrington–Kirkpatrick model for spin-glasses defined on the hierarchical topology has been investigated in Castellana, Decelle, Franz, Mezard, and Parisi (2010): despite a full analytic formulation of its solution still lacks, renormalization techniques, (Castellana & Parisi, 2011; Monthus & Garel, 2013), rigorous bounds on its free-energies (Castellana, Barra, & Guerra, 2014) and extensive numerics (Metz, Leuzzi, & Parisi, 2014; Metz, Leuzzi, Parisi, & Sacksteder, 2013) can be achieved nowadays and they give extremely sharp hints on the behavior of systems at large size defined on these peculiar topologies.

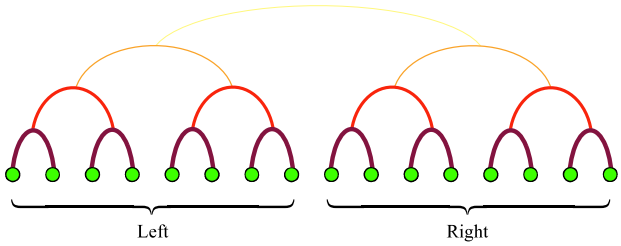


**Fig. 1.** Schematic representation of the hierarchical topology, that underlies the system under study: green spots represent nodes where spins/neurons live, while different colors and thickness for the links mimic different intensities in their mutual interactions: the brighter and thinner the link, the smaller the related coupling. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Remarkably, as we are going to show, when implementing the Hebb prescription for learning on these hierarchical networks, an impressive phase diagram, much richer than the mean-field counterpart, emerges. More precisely, neurons turn out to be able to orchestrate both serial processing (namely sharp and extensive retrieval of a pattern of information), as well as parallel processing (namely retrieval of different patterns simultaneously).

The remaining of the paper is structured as follows: in the next subsections we provide a streamlined description of mean-field serial and parallel processors, and we introduce the hierarchical scenario. Then, we split in three sections our findings according to the methods exploited for investigation: statistical mechanics, signal-to-noise technique and extensive numerical simulations. All these approaches consistently converge to the scenario outlined above. Seeking for clarity and completeness, each technique is first applied to a ferromagnetic hierarchical mode (which can be thought of as a trivial one-pattern neural network and acts as a test-case) and then for a low-storage hierarchical Hopfield model.

### 1.1. Mean-field processing: serial and parallel processors.

Probably the most famous model for neural networks is the Hopfield model presented in his seminal paper dated 1982 (Hopfield & Tank, 1987), counting nowadays more than twenty-thousand citations (Scholar). This is a mean-field model, where neurons are schematically represented as dichotomic Ising spins (state $+1$ represents firing while state $-1$ stands for quiescence) interacting via a (symmetric rearrangement of) the Hebbian rule for learning as masterfully shown by the extensive statistical–mechanical analysis that Amit, Gutfreund and Sompolinsky performed on the model (Amit, 1992; Amit, Gutfreund, & Sompolinsky, 1985).

More formally, once introduced $N$ neurons/spins $S_i$, $i \in (1, \dots, N)$, and $p$ quenched patterns $\xi_\mu$, with $\mu \in (1, \dots, p)$, whose entries are drawn once for all from the uniform distribution

$$P(\xi_i^\mu) = \frac{1}{2}\delta(\xi_i^\mu - 1) + \frac{1}{2}\delta(\xi_i^\mu + 1), \tag{3}$$

the Hopfield model is then captured by the following Hamiltonian $H_{\text{Hopfield}}(S|\xi)$:

$$H_{\text{Hopfield}}(S|\xi) = -\frac{1}{N}\sum_{i<j}^{N}\left(\sum_{\mu=1}^{p}\xi_i^\mu \xi_j^\mu\right) S_i S_j. \tag{4}$$

[2] In that context the long-range interactions are unacceptable because the involved couplings are of electromagnetic nature, hence displaying power-law decay with the distance.

[3] It is worth mentioning that the Wilson–Kadanoff renormalization equations (Wilson, 1971b, 1972, 1974) turn out to be exact in models with power law interactions as those built on the hierarchical lattice that we are going to consider.

Before proceeding with the description of the Hopfield model, it is very instructive to make a step beside and revisit the ferromagnetic system described by the Curie–Weiss Hamiltonian (Eq. (2)). The order parameter for the latter is given by the magnetization $m(S)$ defined as

$$m(S) = \frac{1}{N}\sum_{i=1}^{N} S_i, \tag{5}$$

which, indeed, can distinguish between a paramagnetic/disordered phase ($m = 0$) and a ferromagnetic phase characterized by spontaneous magnetization ($m \neq 0$). Moreover, we can write Eq. (2) also in terms of $m$ as

$$H_{\text{Curie–Weiss}}(S) = -\frac{1}{N}\sum_{i<j}^{N,N} S_i S_j \sim -\frac{N}{2}m^2, \tag{6}$$

where a sub-leading term $\sum_i (S_i)^2/(2N) = 1/2$ has been neglected and we set $J = 1$.

Restricting ourselves to the zero noise limit (for simplicity as entropy maximization can be discarded), following the minimum energy principle we see that the system tends to rearrange in such a way that $|m| \to 1$, corresponding to the configurations $\mathbf{S} = (+1, +1, \ldots, +1)$ or $\mathbf{S} = (-1, -1, \ldots, -1)$. If we read such a state as a neural configuration we would have a pathological state corresponding to *all* spins firing or quiescent. This point can be easily overcome (Mattis transformation) by replacing $S_i \to \xi_i^1 S_i$, where the set $\{\xi^1\}$ may be drawn e.g., according to (3). In this way the Hamiltonian (2) can be rewritten as

$$H_{\text{Mattis}}(S|\xi) = -\sum_{i<j}^{N,N} \xi_i^1 \xi_j^1 S_i S_j = -\frac{N}{2}m_1^2, \tag{7}$$

where $m_1$ is the Mattis magnetization defined as

$$m_1 = \frac{1}{N}\sum_{i=1}^{N} \xi_i^1 S_i. \tag{8}$$

Reasoning exactly as before, in the low noise limit, the system relaxes to the state with $|m_1| \to 1$, corresponding to a spin configuration $\mathbf{S}$ parallel (or anti-parallel) to the pattern $\xi_1$. The relaxation to such a minimum (which now is also the most likely and has only, on average, one half of the neurons firing) is seen as the *retrieval* of the (unique) stored pattern encoded by the string $\xi^1$.

Now, enhancing the network capability, in such a way that the stored patterns are $p > 1$, requires to abandon the ferromagnetic context as the system must be able to develop several free energy minima, each corresponding to the retrieval of a different pattern. This passage is formally straightforward: one simply introduces a sum over the patterns labeled as $\mu = 1, \ldots, p$ in the Mattis Hamiltonian, thus obtaining the Hopfield Hamiltonian (4).

When $p$ is large, that is comparable with the system size (thus in the so-called high-storage regime where $p$ scales as $N$, $p = \alpha N$ with $\alpha \in \mathbb{R}^+$, and as $N \to \infty$), for $\alpha > \alpha_c \sim 0.14$, retrieval properties are lost (and, for $p \to \infty$ quicker than $N$ the Hebbian coupling approaches a standard Gaussian $\mathcal{N}[0, 1]$), hence the model collapses to the Sherrington–Kirkpatrick model for spin-glasses (Amit, 1992; Barra, Genovese, Guerra, Tantari et al., 2012). In this regime neural capabilities are lost due the presence of too much disorder that splits the phase space into an amount of minima that scales exponentially with the system size (Mezard et al., 1987). In the present paper we will work away from this *black out* limit focusing on the low storage scenario, where $p$ is either finite or growing much slower than $N$ (e.g. logarithmical), in such a way that $\lim_{N\to\infty}(p/N) \to 0$.

As mentioned above, as long as the noise is low enough, the system can relax in a (free) energy minimum: for the Hopfield model described by (4) there exist overall $2p$ absolute minima corresponding to the configurations $S_i = \xi_i^\mu$ for all $i = 1, \ldots, N$; each minima encodes for the retrieval of a different pattern and the factor 2 accounts for symmetry $S_i \to -S_i$. The relaxation to the minimum corresponding to the, say, $k$-th pattern is evidenced by $m_k \neq 0$ and $m_i = 0$, $\forall i \neq k$ (the latter holding on the average as patterns $\xi$'s are orthogonal—in the infinite size limit). The particular minimum selected depends on the external field (if present) and on the initial state of the system.

We stress that, since each pattern is built of by $N$ bits of information $\xi_i^\mu = \pm 1$, its retrieval involves the coordination of the whole network and the system can only retrieve patterns singularly, that is, one pattern at a time. For this reason this kind of processing is referred to as *serial*.

This feature can be overcome and the neural network made able to perform *parallel* retrieval, thus giving rise to the so called *multitasking associative network* (Agliari, Barra, Galluzzi, Guerra, & Moauro, 2012), by allowing for blank entries in the Hebbian kernel, that is, pattern entries are extracted once for all from

$$P(\xi_i^\mu) = \left[\frac{1-a}{2}\delta(\xi_i^\mu - 1) + \frac{1-a}{2}\delta(\xi_i^\mu + 1)\right] + a\delta(\xi_i^\mu), \tag{9}$$

where $a \in [0, 1]$ tunes the amount of null-entries in the bit-strings.

Let us try to infer the effects of (9) on the retrieval process by focusing for simplicity on a simple case with $N = 8$ and two toy-patterns $\xi^1 = (+1, +1, +1, +1, 0, 0, 0, 0)$ and $\xi^2 = (0, 0, 0, 0, -1, -1, -1, -1)$, and with the external field (the stimuli) pointing to the first minimum. In suitable regions of phase space (where the network retrieves), the system will try to align with the first pattern, such that the first four neurons will be all firing. The remaining neurons do not receive any information from the pattern $\xi^1$, nevertheless, as the Hopfield Hamiltonian is a quadratic form in the Mattis magnetizations, (free)-energy minimization is better achieved if the remaining neurons align with the second pattern (instead of random reshuffling), such that the final state will be $\mathbf{S} = (+1, +1, +1, +1, -1, -1, -1, -1)$, and we say that the system has spontaneously perfectly retrieved the two patterns. An analogous behavior emerges for arbitrary $p$ patterns: the system tends to relax to a state where the Mattis magnetizations related to a subset of patterns are strictly non zero. The performance of this network crucially depends on how $a$ is tuned as analyzed in detail in Agliari et al. (2013a, 2013b), Agliari, Barra, De Antoni, and Galluzzi (2013) and Agliari et al. (2013) for the low-storage and the high-storage regimes, respectively.

### 1.2. The neural network on a hierarchical topology

We now start our investigation of a neural network embedded in the hierarchical topology depicted in Fig. 1. As mentioned, two main difficulties must be faced: the complexity of the emergent energy landscape (essentially due to frustration in the coupling pattern) and the non-mean-field nature of the model (essentially due to the inhomogeneity of the network architecture). It is therefore safer to proceed by steps discussing first the hierarchical ferromagnet (hence retaining only the second difficulty), known as Dyson hierarchical model (DHM). Then, via the Mattis transformation we reach a Mattis hierarchical model (MHN) and finally we extend to the Hopfield hierarchical model (HHM).

The Dyson hierarchical model (Dyson, 1969) is a system made of $N$ binary (Ising) spins $S_i = \pm 1$, $i = 1, \ldots, N$ in mutual interaction and built recursively in such a way that the system at the $(k+1)$-th iteration contains $N = 2^{k+1}$ spins and is obtained by taking two replicas of the system at the $k$-th iteration (each made of $2^k$ spins) and connecting all possible couples with overall $\binom{N}{2}$ couplings equal to $-J/2^{\sigma(k+1)}$, $J$ and $\sigma$ being real scalars tuning the

interaction strength: the former acts uniformly over the network, the latter triggers the decay with the "distance" among spins. The resulting Hamiltonian can be written recursively as

$$H_{k+1}^{\text{Dyson}}(S|J,\sigma) = H_k^{\text{Dyson}}(\mathbf{S}_1|J,\sigma) + H_k^{\text{Dyson}}(\mathbf{S}_2|J,\sigma) - \frac{J}{2^{2\sigma(k+1)}}\sum_{i<j}^{2^{k+1}} S_iS_j, \tag{10}$$

where $\mathbf{S}_1 = \{S_i\}_{i=1}^{2^k}$ and $\mathbf{S}_2 = \{S_j\}_{i=2^k+1}^{2^{k+1}}$, while $H_0^{\text{Dyson}} \equiv 0$.

Before proceeding it is worth stressing that the parameters $J$ and $\sigma$ are bounded as $J > 0$ and $\sigma \in (\frac{1}{2}, 1)$: the former trivially arises from the ferromagnetic nature of the model which makes neighboring spin to "imitate" each other, while the latter can be understood by noticing that for $\sigma > 1$ the interaction energy goes to zero in the $N \to \infty$ limit,[4] while for $\sigma < \frac{1}{2}$ the interaction energy is no longer linearly-additive implying instability.[5] Moreover, this model is intrinsically *non-mean-field* because a notion of metrics, or distance, has been implicitly introduced: two nodes are said to be at distance $d$ if they get first connected at the $d$th iteration. In general, calling $d_{ij}$ the *distance* between the spins $i$, $j$, (thus $d_{ij} = 1, \dots, k+1$), we can associate to each couple a distant-dependent coupling $J_{ij}$ and rewrite (10) in a more familiar form as

$$H_{k+1}^{\text{Dyson}}(S|J,\sigma) = -\sum_{i<j} J_{ij}S_iS_j, \tag{11}$$

where

$$J_{ij} = \sum_{l=d_{ij}}^{k+1} \frac{J}{2^{2\sigma l}} = J\frac{4^{\sigma-d_{ij}\sigma} - 4^{-k\sigma-\sigma}}{4^\sigma - 1}. \tag{12}$$

Once extracted quenched values for the pattern entries $(\xi_i^\mu)_{\mu=1}$ from the distribution

$$P(\xi_i^\mu) = \frac{1}{2}\delta(\xi_i^\mu - 1) + \frac{1}{2}\delta(\xi_i^\mu + 1), \tag{13}$$

the next step is to replace $S_i$ with $\xi^1 S_i$. This results in the following hierarchical Mattis model

$$H_{k+1}^{\text{Mattis}}(S|J,\sigma) = -\sum_{i<j} J_{ij}\xi_i^1\xi_j^1 S_iS_j. \tag{14}$$

Finally, summing over $p$ patterns, we obtain the Hopfield hierarchical model (HHM) that reads as (for $J = 1$)

$$H_{k+1}^{\text{Hopfield}}(S|\xi,\sigma) = H_k^{\text{Hopfield}}(S_1|\xi,\sigma) + H_k^{\text{Hopfield}}(S_2|\xi,\sigma) - \frac{1}{2}\frac{1}{2^{2\sigma(k+1)}}\sum_{\mu=1}^{p}\sum_{i,j=1}^{2^{k+1}} \xi_i^\mu\xi_j^\mu S_iS_j, \tag{15}$$

with $H_0^{\text{Hopfield}} \equiv 0$ and $\sigma$ still within the previous bounds, i.e. $\sigma \in (\frac{1}{2}, 1)$. As anticipated, here we restrict the analysis to low storage limit only: recalling $N = 2^{k+1}$, we can fix $p$ finite at first so to move straightforwardly from the DHM to the HHM (as the notion of distance is preserved) and, posing

$$J_{ij} = \frac{4^{\sigma-d_{ij}\sigma} - 4^{-k\sigma-\sigma}}{4^\sigma - 1}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu, \tag{16}$$

we can write equivalently the Hamiltonian (15) in the more compact form

$$H_{k+1}^{\text{Hopfield}}(S|\xi,\sigma) = -\sum_{i<j}^{2^{k+1}} J_{ij}S_iS_j. \tag{17}$$

Thus in the HHM the Hebbian prescription is coupled with (or "weighted by" Agliari, Barra, Del Ferraro, Guerra, & Tantari, in press, Pastur, Shcherbina, & Tirozzi, 1994) a function of the neuron's distance.

In the following, in order to analyze in depth the system performance and the properties of hierarchical retrieval, we tackle the problem from different perspectives, each developed in a dedicated section. In particular, the next section is devoted to the statistical–mechanical route, for which we report only results (as the methodologies underlying such achievements are still extremely technical and have been presented to the pertinent Community Agliari et al., 2015a, 2015b). As through this path a full analytical solution still lacks, further investigations must be addressed: indeed in Section 3 we largely exploit outcomes from signal-to-noise studies, while numerical simulations are presented in Section 4.

## 2. Insights from statistical mechanics

Here we summarize findings that can be achieved by suitably extending interpolation techniques (Guerra, 2003; Guerra & Toninelli, 2002) beyond the mean-field paradigm: it is important to stress once more that, as this strand gives only (not-mean-field) bounds on the free energy (and not the full solution), the self-consistencies that result are not the true self-consistencies of the model, thus motivating the next sections.

### 2.1. Pure/ferromagnetic and parallel/mixed free energies in the Dyson model

As the Hamiltonian $H_{k+1}(S|J,\sigma)$ is given (see Eq. (10)) and the noise level $\beta^{-1} = T$ (where $T$ stands for *noise* for historical reasons) introduced, it is possible to define the partition function $Z_{k+1}(\beta,J,\sigma)$ at finite volume $k+1$ as

$$Z_{k+1}(\beta,J,\sigma) = \sum_{\{S\}} \exp\left[-\beta H_{k+1}(S|J,\sigma)\right], \tag{18}$$

and the related free energy $f_{k+1}(\beta,J,\sigma)$, namely the intensive logarithm of the partition function, as

$$f_{k+1}(\beta,J,\sigma) = \frac{1}{2^{k+1}}\log\sum_{\{S\}}\exp\left[-\beta H_{k+1}(\vec{S}) + h\sum_{i=1}^{2^{k+1}} S_i\right], \tag{19}$$

where the sum runs over all possible $2^{2^{k+1}}$ spin configurations and $h$ tunes a possible homogeneous external field. Note that the usual free energy $\tilde{f}$ is related to $f$ by $\tilde{f}(\beta) = -\beta f(\beta)$, hence we will find the equilibrium states checking the maxima of $f(\beta)$ and not the minima.

We are interested in an explicit expression of the infinite volume limit of the intensive free energy, defined as

$$f(\beta,J,\sigma) = \lim_{k\to\infty} f_{k+1}(\beta,J,\sigma), \tag{20}$$

in terms of suitably introduced magnetizations $m$, that act as order parameters for the theory. To this task we introduce the global magnetization $m$, defined as the limit $m = \lim_{k\to\infty} m_{k+1}$ where

$$m_{k+1} = \frac{1}{2^{k+1}}\sum_{i=1}^{2^{k+1}} S_i, \tag{21}$$

[4] The sum $\sum_{i<j}^{2^{k+1}}$ brings a contribution scaling like $2^{2(k+1)} \sim N^2$, while the pre-factor scales as $2^{-2\sigma(k+1)} \sim N^{-2\sigma}$, thus, when $\sigma > 1$ the internal energy (the expectation of the Hamiltonian normalized over the system size) is overall vanishing in the infinite size limit $k \to \infty$.

[5] The sum $\sum_{i<j}^{2^{k+1}}$ brings a contribution scaling like $2^{2(k+1)} \sim N^2$, while the pre-factor scales as $2^{-2\sigma(k+1)} \sim N^{-2\sigma}$, thus, when $\sigma < \frac{1}{2}$ the intensive energy is overall divergent in the limit $k \to \infty$.

and, recursively and with a little abuse of notation, level by level (over $k$ levels) the $k$ magnetizations $\vec{m}_a, \ldots, \vec{m}_k$, as the same $k \to \infty$ limit of the following quantities (we write explicitly only the two upper magnetizations related to the two main clusters *left* and *right*—see Fig. 1):

$$m_k^1 = \frac{1}{2^k}\sum_{i=1}^{2^k} S_i, \qquad m_k^2 = \frac{1}{2^k}\sum_{i=2^k+1}^{2^{k+1}} S_i, \tag{22}$$

and so on. The *averages* are denoted by the brackets $\langle\cdot\rangle$ such that, e.g. for the observable $m_{k+1}(\beta, J, \sigma)$, we can write

$$\langle m_{k+1}(\beta, J, \sigma)\rangle = \frac{\sum_{\sigma} m_{k+1} e^{-\beta H_{k+1}(\vec{S}|J,\sigma)}}{Z_{k+1}(\beta, J, \sigma)}, \tag{23}$$

and clearly $\langle m(\beta, J, \sigma)\rangle = \lim_{k\to\infty}\langle m_{k+1}(\beta, J, \sigma)\rangle$.

Starting with the pure ferromagnetic case, which mirrors here the serial retrieval of a single pattern in the Hopfield counterpart, its free energy can be bounded as (see also Castellana et al., 2014)

$$f(h, \beta, J, \sigma) \geq \sup_m \Big\{\log 2 + \log\cosh\Big[h + \beta m J(C_{2\sigma-1} - C_{2\sigma})\Big] - \frac{\beta J}{2}(C_{2\sigma-1} - C_{2\sigma})m^2\Big\}, \tag{24}$$

where

$$C_{2\sigma} = \frac{1}{2^{2\sigma} - 1}, \tag{25}$$

$$C_{2\sigma-1} = \frac{1}{2^{2\sigma+1} - 1}. \tag{26}$$

Now, let us suppose that, instead of a global ordering, the system can be effectively split into two parts (the two largest communities called *left* and *right* in Fig. 1), with two different magnetizations $m_{left} = m_1$ and $m_{right} = m_2$; we also assume $m_{left} = -m_{right}$. Through the interpolative route we approach a bound for the free energy related to such a mixed state. We stress the fact that the upper link, connecting the two communities with opposite magnetization, remains and it gives a contribute $m$ in the system as (see also Agliari et al., 2015a)

$$\begin{aligned} f_{k+1} \geq\ & \frac{1}{2}\log\cosh\left\{h + \beta J\left[m(2^{(k+1)(1-2\sigma)}) + m_1\left(\sum_{l=1}^{k} 2^{l(1-2\sigma)} - \sum_{l=1}^{k+1} 2^{-2l\sigma}\right)\right]\right\} \\ & + \frac{1}{2}\log\cosh\left\{h + \beta J\left[m(2^{(k+1)(1-2\sigma)}) + m_2\left(\sum_{l=1}^{k} 2^{l(1-2\sigma)} - \sum_{l=1}^{k+1} 2^{-2l\sigma}\right)\right]\right\} \\ & - \frac{\beta J}{2}\left[\left(\sum_{l=1}^{k} 2^{l(1-2\sigma)} - \sum_{l=1}^{k+1} 2^{-2l\sigma}\right) \times \left(\frac{m_1^2 + m_2^2}{2}\right) - 2^{(k+1)(1-2\sigma)}m^2\right] + \log 2. \end{aligned} \tag{27}$$

Notice that, thanks to the symmetry $S_i \to -S_i$, the state considered mirrors the parallel retrieval of two patterns in the Hopfield counterpart. Identifying $m_1 = m_2 = m$ we recover the previous bound as expected, and, quite remarkably, in the infinite size limit the two free energies assume the same values, thus serial and parallel retrieval are both equally accomplished by the network.

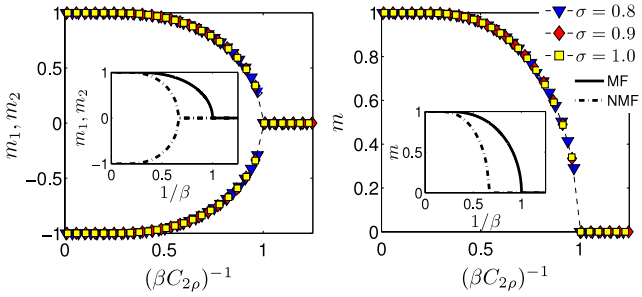


**Fig. 2.** Main plots: numerical solutions of the non-mean-field self-consistent equations for the parallel state (left panel) and for the pure state (right panel) of the Dyson model (see Eq. (28)) obtained for different values of $\sigma$ (as explained by the legend) and plotted versus a rescaled noise. Note that by rescaling the noise the dependence on $\sigma$ is lost and all curves are collapsed. Insets: comparison between the numerical solutions of the non-mean-field self-consistent equations (dashed line) and of the mean-field self-consistent equations (solid line) as a function of the noise and for fixed $\sigma = 1$ (see Eq. (28)). Notice that we have qualitatively the same behavior but with different critical noise level separating the retrieval region from the paramagnetic (where the noise is too high) one.

Optimizing the bound (27) we obtain

$$m_{1,2} = \tanh(h + \beta J m_{1,2}(C_{2\sigma-1} - C_{2\sigma})), \tag{28}$$

representing two disentangled self-consistent equations whose behavior is depicted in Fig. 2. Note that Eq. (28) also expresses the self-consistent equation for the pure state (using $m_{1,2} = m$), obtained by optimizing the bound (24).

### 2.2. Serial versus parallel retrieval in Hopfield hierarchical model

Guided by the ferromagnetic model just described, we now turn to the hierarchical Hopfield model (HHM) and start its analysis from a statistical–mechanical perspective, namely we infer the behavior of a system described by the following recursive Hamiltonian

$$H_{k+1}^{HHM}(S|\xi, \sigma) = H_k^{HHM}(S_1|\xi, \sigma) + H_k^{HHM}(S_2|\xi, \sigma) - \frac{1}{2}\frac{1}{2^{2\sigma(k+1)}}\sum_{\mu=1}^{p}\sum_{i,j}^{2^{k+1}} \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j. \tag{29}$$

To this task, we introduce suitably $p$ Mattis magnetizations (or Mattis overlaps), over the whole system, as

$$m^\mu = \frac{1}{2^{k+1}}\sum_{i=1}^{2^{k+1}} \xi_i^\mu S_i, \quad \mu \in [1, p]. \tag{30}$$

Even in this context, the definition above can account for the state of inner clusters by the sum over the (pertinent) spins. For instance, focusing on the two larger communities we have the $2p$ Mattis magnetizations

$$m_{left}^\mu = \frac{1}{2^k}\sum_{i=1}^{2^k} \xi_i^\mu S_i, \qquad m_{right}^\mu = \frac{1}{2^k}\sum_{i=2^k+1}^{2^{k+1}} \xi_i^\mu S_i, \tag{31}$$

with $\mu \in [1, p]$. Again, we will not enter in the mathematical details concerning non-mean-field bounds for the model free energy (as they can be found in Agliari et al., 2015a), while we streamline directly the physical results.

Still mirroring the previous section, we are interested in obtaining a bound limiting the free energy of the HHM, the latter being defined as the $k \to \infty$ limit of $f_{k+1}$, whose expression reads

$$f_{k+1}(\beta, \{h_\mu\}, \sigma) = \frac{1}{2^{k+1}}\log\sum_{\{S\}}\exp\left[-\beta H_{k+1}(\vec{S}) + \sum_{\mu=1}^{p} h^\mu \sum_{i=1}^{2^{k+1}} S_i\right], \tag{32}$$

where we accounted also for $p$ external stimuli $h^\mu$.

The non-mean field bound for serial processing free energy reads as

$$f(\beta,\{h^\mu\},p) \geq \sup_m \Bigg[ \log 2 + \Big\langle \log\cosh\Big(\sum_{\mu=1}^{p}\Big[h^\mu + \beta m^\mu \times (C_{2\sigma-1}-C_{2\sigma})\Big]\xi^\mu\Big)\Big\rangle_\xi - \frac{\beta}{2}\sum_{\mu=1}^{p}\langle (m^\mu)^2\rangle_\xi (C_{2\sigma-1}-C_{2\sigma})\Bigg], \tag{33}$$

with optimal order parameters fulfilling

$$\langle m^\mu\rangle_\xi = \left\langle \xi^\mu \tanh\left[\beta\sum_{\nu=1}^{p}[h^\nu + (C_{2\sigma-1}-C_{2\sigma})m^\nu]\,\xi^\nu\right]\right\rangle_\xi,$$

and whose critical noise is $\beta_c^{NMF} = C_{2\sigma-1} - C_{2\sigma}$, where the index *NMF* stresses that the estimate was obtained through a non mean field bound of the free energy.

Of course we can assume again that the two different families of Mattis magnetizations ($\{m^\mu_{1,2}\}_{\mu=1}^p$) (those playing for the two inner blocks of spins *left* and *right* lying under the $k+1$-th level) behave independently as the higher links connecting them go to zero quickly for $k \to \infty$ and we can start the interpolative machine: following this way we generalize the serial processing analysis to a two-pattern parallel retrieval analysis, which results in the following bound for the related free energy:

$$\begin{aligned} f(\beta,\{h_\mu\},p) \geq \sup_{\{m^\mu_{1,2}\}} \Bigg[ & \log 2 + \frac{1}{2}\Big\langle \log\cosh\Big\{\sum_{\mu=1}^{p}\Big[h^\mu + \beta m_1^\mu \times \Big(\sum_{l=1}^{k}2^{l(1-2\sigma)} - \sum_{l=1}^{k}2^{l(-2\sigma)}\Big) + \beta m^\mu 2^{(k+1)(1-2\sigma)}\Big]\xi^\mu\Big\}\Big\rangle_\xi \\ & + \frac{1}{2}\Big\langle \log\cosh\Big\{\sum_{\mu=1}^{p}\Big[h^\mu + \beta m_2^\mu\Big[\sum_{l=1}^{k}2^{l(1-2\sigma)} - \sum_{l=1}^{k}2^{l(-2\sigma)}\Big] + \beta m^\mu 2^{(k+1)(1-2\sigma)}\Big]\xi^\mu\Big\}\Big\rangle_\xi \\ & - \frac{\beta}{2}\Big[\sum_{l=1}^{k}2^{l(1-2\sigma)} - \sum_{l=1}^{k}2^{l(-2\sigma)}\Big] \cdot \sum_{\mu=1}^{p}\frac{\langle (m_1^\mu)^2\rangle_\xi + \langle (m_2^\mu)^2\rangle_\xi^2}{2} \\ & - \frac{\beta}{2}2^{(k+1)(1-2\sigma)}\sum_{\mu=1}^{p}\langle (m^\mu)^2\rangle_\xi\Bigg]. \end{aligned}$$

Here we do not investigate further the parallel retrieval of larger ensembles of patterns, as the way to proceed is identical to the outlined one, but we simply notice that, if we want the system to handle $M$ patterns, hence we assume it effectively splits $M$ times into sub-clusters until the $k+1-M$ level, then the procedure keeps on working as long as

$$\lim_{k\to\infty}\sum_{l=k+1-M}^{k+1}2^{l(1-2\sigma)}\sum_{\mu=1}^{p}m_l^\mu = 0. \tag{34}$$

Since the magnetizations are bounded, in the worst case we have

$$\begin{aligned} \sum_{l=k+1-M}^{k+1}2^{l(1-2\sigma)}\sum_{\mu=1}^{p}m_l^\mu &\leq p\sum_{l=k+1-M}^{k+1}2^{l(1-2\sigma)} \\ &\leq p\sum_{l=k+1-M}^{\infty}2^{l(1-2\sigma)} \propto 2^{(1-2\sigma)(k+1-M)}p. \end{aligned} \tag{35}$$

If we want the system to handle up to $p$ patterns, we need $p$ different blocks of spins and then $M = \log(p)$.

## 3. Insights from signal-to-noise techniques

Results from statistical mechanics gave stringent hints on the network's behavior, however they act as bounds only.

This requires further inspection via other techniques: the first route we exploit is signal-to-noise. Through the latter, beyond generally confirming the predictions obtained via the first path, we obtain sharper statements regarding the evolution of the Mattis order parameters. These two approaches are complementary: while statistical mechanics describes the system with $N \to \infty$ and $\beta < \infty$, with the signal-to-noise technique we inspect the regime $N < \infty$ and $\beta \to \infty$.

### 3.1. A glance at the fields in the Dyson network

Plan of this section is to look at the dynamically stable configurations of the neurons, that is to say, we investigate the configurations (global and local minima) that imply each neuron $S_i$ to be aligned with its corresponding field $h_i[S]$, i.e. $S_i\,h_i[S] > 0$, $\forall i$. This approach basically corresponds to a negligible-noise statistical–mechanical analysis but it is mathematically much more tractable.

We can rearrange the Dyson Hamiltonian in a useful form for such an investigation as follows:

$$\begin{aligned} & H^{\text{Dyson}}_{k+1}(\{S_1\dots S_{2^{k+1}}\}) \\ & \quad = -\frac{J}{2}\sum_{\mu=1}^{k+1}\sum_{i=1}^{2^{k+1}}S_i\left[\sum_{l=\mu}^{k+1}\left(\frac{1}{2^{2\sigma}}\right)^l\right]\sum_{\{j\}:d_{ij}=\mu}S_j, \end{aligned} \tag{36}$$

thus, highlighting the field $h_i$ insisting on the spin $S_i$ we can write

$$H^{\text{Dyson}}_{k+1}[\{S_1\dots S_{2^{k+1}}\}] = -\sum_{i=1}^{2^{k+1}}S_i h_i[\mathbf{S}], \tag{37}$$

$$h_i[\mathbf{S}] = J\sum_{\mu=1}^{k+1}\left[\sum_{l=\mu}^{k+1}\left(\frac{1}{2^{2\sigma}}\right)^l\right]\sum_{\{j\}:d_{ij}=\mu}S_j. \tag{38}$$

While Glauber dynamics will be discussed in Section 4 (dedicated to numerics), we just notice here that the microscopic law governing the evolution of the system can be defined as a stochastic alignment to local field $h_i[\mathbf{S}]$.

$$S_i(t+\delta t) = \text{sign}\left\{\tanh\left[\beta h_i\left[\mathbf{S}\left(t\right)\right]\right] + \eta_i(t)\right\},$$

where the stochasticity lies in the independent random numbers $\eta_i(t)$, uniformly distributed over the interval $[-1, 1]$ and tuned by $\beta$. The latter continues to rule the noise level even dynamically as it amplifies, or suppresses, the smoothness of the hyperbolic tangent; in particular, in the noiseless limit $\beta \to \infty$ we get

$$S_i(t+\delta t) = \text{sign}\left[h_i\left(\mathbf{S}(t)\right)\right]. \tag{39}$$

This is crucial for checking the stability of a state as, if $S_i h_i[\mathbf{S}] > 0\ \forall\, i \in [1, N]$, the configuration $\{\mathbf{S}\}$ is dynamically stable (at least for $\beta \to \infty$, as in the presence of noise there is a $\beta$-dependent probability to fluctuate away).

We keep the previous ensemble of non-independent order parameters $m_i^n$ defined in detail as

$$m_i^n[\mathbf{S}] = \frac{1}{2^n}\sum_{j=2^n\times i-(2^n-1)}^{2^n\times i}S_j \quad \text{with } i = 1, 2, \dots, 2^{k+1-n} \text{ and } n = 0, 1, 2, \dots, k+1, \tag{40}$$

namely

$$\begin{cases} m_i^0 = S_i & \text{with } i = 1, 2, \ldots, 2^{k+1}, \\ m_i^1 = \frac{1}{2}\sum_{j=2i-1}^{2i} S_j & \text{with } i = 1, 2, \ldots, 2^k \to m_1^1 = \frac{1}{2}\sum_{j=1}^{2} S_j, \\ m_i^2 = \frac{1}{2^2}\sum_{j=2^2 i-(2^2-1)}^{2^2 i} S_j & \text{with } i = 1, 2, \ldots, 2^{k-1} \to m_1^2 = \frac{1}{4}\sum_{j=1}^{4} S_j, \\ \ldots & \\ m_1^{k+1} = \frac{1}{2^{k+1}}\sum_{j=1}^{2^{k+1}} S_j. \end{cases}$$

From Eq. (38), we get the following fundamental expression for the fields

$$h_i[\mathbf{S}] = \left[ J \sum_{\mu=1}^{k+1} \left( \sum_{l=\mu}^{k+1} \frac{1}{2^{2\sigma}} \right)^{l} \right] 2^{\mu-1} m_{f(\mu,i)}^{\mu-1}, \tag{41}$$

where we used the relation $m_{f(\mu,i)}^{\mu-1} = \sum_{\{j\}:d_{ij}=\mu} S_j$. Thus the order parameters $m_{f(\mu,i)}^{\mu-1}$ represent the magnetizations assumed by spins that lie at distance $\mu$ from $S_i$. Note that the function $f(\mu, i)$ can be estimated through the floor function $\lfloor\cdot\rfloor$ (e.g., $\lfloor 3.14 \rfloor = 3$) as

$$f(\mu, i) = \left\lfloor \frac{i + (2^{\mu-1} - 1)}{2^{\mu-1}} \right\rfloor + (-1)^{\left( \lfloor \frac{i+(2^{\mu-1}-1)}{2^{\mu-1}} \rfloor + 1 \right)}.$$

Finally, we notice that the largest value allowed for a field – away from the boundary value $\sigma = 1/2$ – for large $k$ approaches a plateau (whose boundaries – in the $(k, \sigma)$ plane – are important for finite-size-scaling during numerical analysis), hence we can easily check the right field normalization

$$\begin{aligned} Q(\sigma, k+1) &= \sum_{\mu=1}^{k+1} J(\mu, k+1, \sigma) 2^{\mu-1} \\ &= J \frac{2^{-2(k+1)\sigma} \left( 2^{2(k+2)\sigma} - 2^{k+2\sigma+2} + 2^{k+2} + 4^{\sigma} - 2 \right)}{-3 \times 4^{\sigma} + 16^{\sigma} + 2}, \end{aligned} \tag{42}$$

as $Q(\sigma, k)$ represents the largest value allowed by a field.

Note that in the infinite size limit

$$\lim_{k\to\infty} Q(\sigma, k) = Q(\sigma) = J \frac{2^{2\sigma}}{-3 \times 4^{\sigma} + 4^{2\sigma} + 2}, \tag{43}$$

that is $Q$ is always bounded whenever $\sigma > \frac{1}{2}$.

### 3.2. Metastabilities in the Dyson network: noiseless case

We can now proceed to the stability analysis explaining in detail a few test cases that show how to proceed for any other case of further interest:

[a] the global ferromagnetic state, i.e. $S_i = +1, i \in (1, \ldots, 2^{k+1})$.
[b] the parallel/mixed state, i.e. the first half of spins up and the second half down, thus $S_i = +1, i \in (1, \ldots, 2^k)$ and $S_i = -1,\ i \in (2^k + 1, \ldots, 2^{k+1})$.
[c] the dimer, i.e. $S_1 = S_2 = +1$ while $S_i = -1$ for all $i \neq (1, 2)$.
[d] the square, i.e. $S_1 = S_2 = S_3 = S_4 = +1$ while $S_i = -1$ for all $i > 4$.

Let us go through each case analysis separately:

- [a] The global ferromagnetic state $S_i = +1\ \forall i \in [1, 2^{k+1}] \Rightarrow m_i^n[\mathbf{S}] = 1\ \forall i, n$ has fields

$$\begin{aligned} &\Rightarrow h_i[\mathbf{S}] \\ &= J \frac{4^{-(k+1)\sigma} \left[ 2^{2(k+2)\sigma} - 2^{k+2+2\sigma} + 2^{k+2} + 4^{\sigma} - 2 \right]}{-3 \times 4^{\sigma} + 16^{\sigma} + 2}, \end{aligned} \tag{44}$$

$$\Rightarrow h_i[\mathbf{S}] > 0 \quad \forall k,\ \sigma \in (1/2, 1). \tag{45}$$

Thus, the configuration $S_i = +1\ \forall i \in [1, 2^{k+1}]$ is stable in the noiseless limit $\forall \sigma \in [\frac{1}{2}, 1]$. In the limit $k \to \infty$ we have

$$h_i[\mathbf{S}] = J \frac{4^{\sigma}}{-3 \times 4^{\sigma} + 16^{\sigma} + 2}.$$

To address network's behavior in the presence of noise, fixing $J = 1$ without loss of generality, we can look at the solution of the following equation

$$\tanh(\beta h_i[\mathbf{S}]) \simeq 1 \Rightarrow \tanh\left( \beta \frac{4^{\sigma}}{-3 \times 4^{\sigma} + 16^{\sigma} + 2} \right) \simeq 1. \tag{46}$$

This allows to find the curve $\beta_c^{\text{no errors}}(\sigma)$ versus $\sigma$ (shown in Fig. 3). In fact, we know that, at the time $t + \delta t$, the system obeys the dynamics

$$S_i(t + \delta t) = \text{sign}(\tanh(\beta h_i(\mathbf{S})) + \eta_i),$$

where $\eta_i$ is a random variable, whose value is uniformly distributed in $[-1, 1]$. Imposing $\tanh(\beta h_i) \simeq 1$ we ask that $|h_i| \gg 1$, so the sign of the right hand side member of the equation is positive, thus the sign of $S_i$ at the time $t + \delta t$ is the same of the field $h_i$ at the time $t$. Then, fixed $\sigma$, for every $\beta > \beta_c^{\text{no errors}}(\sigma)$ the state $S_i = +1\ \forall i \in [1, 2^{k+1}]$ is stable without errors.

- [b] The parallel/mixed state $S_j = +1\ S_i = -1\ \forall j \in [1, 2^k]\ \forall i \in [2^k + 1, 2^{k+1}]$ has fields

$$\begin{aligned} &\Rightarrow h_j[\mathbf{S}] \\ &= J \frac{4^{-(k+1)\sigma} \left( 2^{2(k+2)\sigma} + 2^{k+1+2\sigma} - 2^{k+1+4\sigma} + 4^{\sigma} - 2 \right)}{-3 \times 4^{\sigma} + 16^{\sigma} + 2} \\ &= -h_i[\mathbf{S}] > 0 \quad \forall k + 1 \geq 2, \end{aligned} \tag{47}$$

$$\Rightarrow \lim_{k\to\infty} h_j[\mathbf{S}] = J \frac{1}{2^{1-2\sigma} + 4^{\sigma} - 3}, \tag{48}$$

thus the configuration $S_j = +1\ S_i = -1\ \forall j \in [1, 2^k]\ \forall i \in [2^k + 1, 2^{k+1}]$ is stable in the noiseless limit $\forall k + 1 > 2$, $\sigma \in (1/2, 1)$. Using the same arguments of the previous case, fixing $J = 1$ without loss of generality, to infer network's behavior in the presence of the noise we can look at the solution of the following equation

$$\tanh(\beta h_i[\mathbf{S}]) \simeq 1 \Rightarrow \tanh\left( \beta \frac{1}{2^{1-2\sigma} + 4^{\sigma} - 3} \right) \simeq 1. \tag{49}$$

This allows to find the curve $\beta_c^{\text{no-errors}}(\sigma)$ versus $\sigma$ (see Fig. 3). Then, fixed $\sigma$, for every $\beta > \beta_c^{\text{no-errors}}(\sigma)$ the state $S_j = 1\ S_i = -1\ \forall j \in [1, 2^k]\ \forall i \in [1 + 2^k, 2^{k+1}]$ is stable without errors. So we can see how, in the infinite size limit, the state with all spins aligned $S_j = +1\ \forall j \in [1, 2^{k+1}]$ and the state with half spins pointing upwards and half pointing downwards $S_j = +1\ \forall j \in [1, 2^k]\ S_i = -1\ \forall i \in [1 + 2^k, 2^{k+1}]$ are both robust. For an arbitrary finite value of $k$ it is possible to solve numerically Eq. (49) to get an estimate for $\beta_c^{\text{no-errors}}(\sigma)$ versus $\sigma$: in Fig. 3 $\beta_c^{\text{no-errors}}(\sigma)$ is plotted for the state $S_j = +1\ S_i = -1\ \forall j \in [1, 2^k]\ \forall i \in [1 + 2^k, 2^{k+1}]$ and the state $S_i = +1\ \forall i \in [1, 2^{k+1}]$.

- [c] The dimer $S_j = +1\ S_i = -1\ \forall j \in [1, 2]\ \forall i \in [3, 2^{k+1}]$ has fields

$$\begin{aligned} &h_1[\mathbf{S}] = h_2[\mathbf{S}] \\ &= \frac{2^{-2\sigma(k+1)} (2^{2\sigma(k+2)} + 2^{k+2+2\sigma} - 4^{1+(k+1)\sigma} - 2^{k+2} - 3 \times 4^{\sigma} + 6)}{(-3 \times 4^{\sigma} + 16^{\sigma} + 2)}, \end{aligned}$$

$$\lim_{k\to\infty} h_1[\mathbf{S}] = \lim_{k\to\infty} h_2[\mathbf{S}] = 2 \cdot \frac{4^{\sigma} - 4}{-3 \times 4^{\sigma} + 16^{\sigma} + 2} < 0$$

$$\forall \sigma \in (1/2, 1).$$

Therefore, the configuration $S_j = +1\ S_i = -1\ \forall j \in [1, 2]\ \forall i \in [3, 2^{k+1}]$, in the infinite size limit, is unstable $\forall\ \sigma \in (1/2, 1)$.

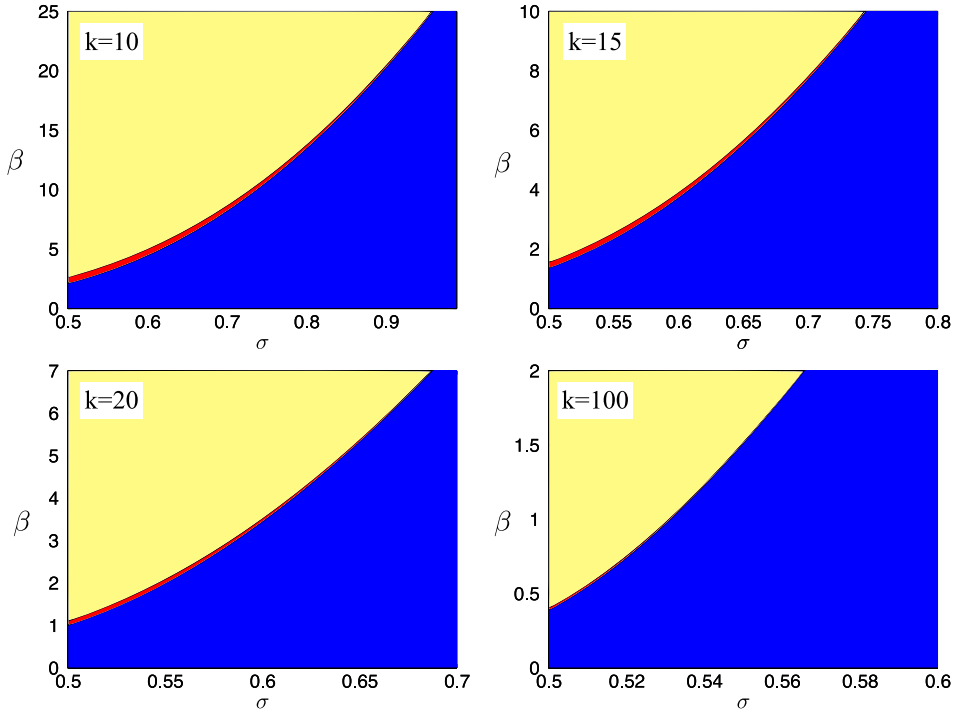


**Fig. 3.** Phase diagram for the perfect retrieval accomplished by a pure state ($S_i = +1\ \forall i = 1, \dots, 2^{k+1}$) and parallel state ($S_i = +1\ \forall i = 1, \dots, 2^k$ and $S_i = -1\ \forall i = 2^k + 1, \dots, 2^{k+1}$). The line separating different regions corresponds to numerical solution of $\beta_c^{\text{no errors}}[\sigma]$ versus $\sigma$, obtained from (46) and (49) for different values of $k$ (10, 15, 20, 100 respectively). In yellow, the area where both the pure and parallel states are perfectly retrieved, while in blue the area where none of them is retrieved. The red line represents the area where only the pure state is stable: this region vanishes as $k$ gets larger, hence confirming that the pure and the mixed state are both global minima. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- [d] The square $S_j = 1\ S_i = -1\ \forall j \in [1, 4]\ \forall i \in [5, 2^{k+1}]$ has fields

$$h_j[\mathbf{S}, k] = -\frac{2^{1-2(k+1)\sigma}\left(-2^{k+1+2\sigma} + 2^{2k\sigma+1} + 2^{k+1} + 2^{2\sigma+1} - 4\right)}{-3\times 4^{\sigma} + 16^{\sigma} + 2} - \frac{-3\times 4^{-(k+1)\sigma} + 2^{1-2\sigma} + 1}{1 - 4^{\sigma}}, \tag{50}$$

$$h_j[\mathbf{S}, k+1] = \frac{\left(2^{2(k+3)\sigma} - 2^{k+2+2\sigma} + 2^{k+2+4\sigma} - 2^{2(k+1)\sigma+3} + 7\times 2^{2\sigma+1} - 7\times 16^{\sigma}\right)}{(-3\times 4^{\sigma} + 16^{\sigma} + 2)/(2^{-2(k+2)\sigma})}$$

thus

$$\lim_{k\to\infty} h_j[\mathbf{S}] = \frac{4^{-\sigma}\,(16^{\sigma} - 8)}{-3\times 4^{\sigma} + 16^{\sigma} + 2} = \begin{cases} > 0, & \text{if } \sigma > \dfrac{3}{4} \\ < 0, & \text{if } \sigma < \dfrac{3}{4}. \end{cases}$$

Therefore, the configuration $S_j = +1\ S_i = -1\ \forall j \in [1, 4]\ \forall i \in [5, 2^{k+1}]$ in the limit ($k \to \infty$) for $T = 0$ is stable $\forall\ \sigma \in (\frac{3}{4}, 1)$.

It is worth noticing that beyond the extensive meta-stable states (e.g. the parallel/mixed one) already suggested by the statistical–mechanical route, stability analysis predicts that tightly connected modules (e.g. octagon, hexadecagon, etc.) with spins anti-aligned with respect to the bulk get dynamically stable in the infinite size limit (see Fig. 4): these *motifs* in turn are able to process small amount of information and an analysis of their capabilities can be found in Agliari et al. (2013a, 2013b), and their robustness is due to their intrinsic loopy structure.

### 3.3. Signal analysis for the Hopfield hierarchical model

Let us now consider the Hopfield hierarchical model (see Eq. (29)). As we are interested in obtaining an explicit prescription for the fields experienced by the neurons, we can rewrite its

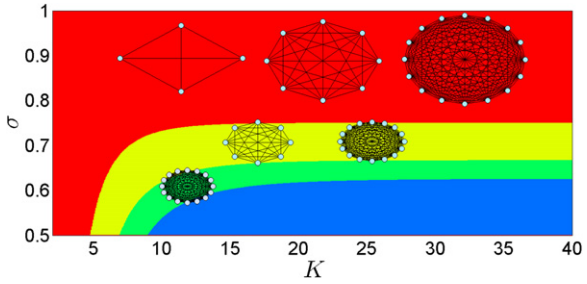


**Fig. 4.** Stability and instability zones for various configurations in the plane $(\sigma, k)$ when $\beta \to 0$, obtained by solving the inequality $S_i h_i(\sigma, k, [\mathbf{S}]) > 0$. In particular in the figure, the square represents the configuration $S_i = +1\ \forall i \in [1, 4]$ and $S_i = -1\ \forall i \in [5, 2^{k+1}]$, the octagon the configuration $S_i = +1\ \forall i \in [1, 8]$ and $S_i = -1\ \forall i \in [9, 2^{k+1}]$, and the hexadecagon the configurations $S_i = +1\ \forall i \in [1, 16]$ and $S_i = -1\ \forall i \in [17, 2^{k+1}]$. In red we can see the region where all of them are stable, in yellow the region where only the octagon and the hexadecagon are stable, in green the region where only the hexadecagon is stable, while in blue none of these reticular animals is stable. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Hamiltonian in terms of neural distance $d_{ij}$ as

$$H_{k+1}(S|\xi, \sigma) = \sum_{i<j} S_i S_j \Big[\sum_{l=d_{ij}}^{k+1} \left(\frac{-1}{2^{2\sigma l}}\right)\Big] \sum_{\mu=1}^{p} \xi_i^{\mu}\xi_j^{\mu} \tag{51}$$

or, splitting over all the possible distances $d$, and grouping all the neurons sharing the same distance from the $i$th neuron, as

$$H_{k+1}(S|\xi, \sigma) = -\sum_{d=1}^{k+1}\sum_{i=1}^{2^{k+1}} S_i \Big[\sum_{l=d}^{k+1} \left(\frac{1}{2^{2\sigma}}\right)^{l}\Big] \sum_{\{j\}:d_{ij}=d} S_j \sum_{\mu=1}^{p} \xi_i^{\mu}\xi_j^{\mu},$$

such that, paying attention to the fields we can write

$$H_{k+1}(S|\xi, \sigma) = -\sum_{i=1}^{2^{k+1}} S_i h_i[\mathbf{S}], \tag{52}$$

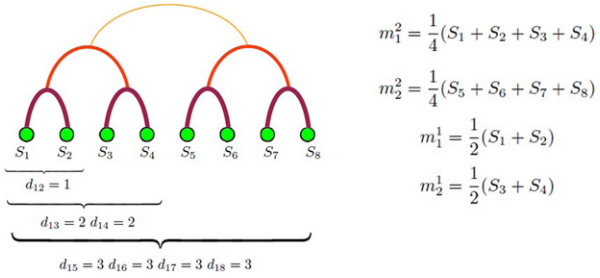


**Fig. 5.** The Hierarchical structure represented by embedding the system in a tree like topology: the distance $d_{ij}$ is the canonical distance on the tree and an example of some order parameters $m_i^n$ (dropping the label $\mu$ indicating the patterns' dependence) representing the magnetization of groups of spins up to a distance $n$.

$$h_i[\mathbf{S}]=\sum_{d=1}^{k+1}\Big[\sum_{l=d}^{k+1}\left(\frac{1}{2^{2\sigma}}\right)^l\Big]\sum_{\{j\}:d_{ij}=d}S_j\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu. \tag{53}$$

Mirroring the analysis carried on for the Dyson model (see Fig. 5), we introduce an ensemble of non-independent Mattis-like order parameters as

$$m_i^{\mu,n}[\mathbf{S}]=\frac{1}{2^n}\sum_{j=i\times 2^n-(2^n-1)}^{i\times 2^n}S_j\xi_j^\mu \quad \text{with } i=1,2,\dots,2^{k+1-n},$$
$$n=0,1,2,\dots,k+1, \tag{54}$$

where $n$ runs over the $k+1$ possible distances among neurons and $i$ runs over all the blocks of $2^n$ neurons that have $n$ as maximum distance, so that

$$\begin{cases} m_i^{\mu,0}=S_i\xi_i^\mu & \text{with } i=1,2,\dots,2^{k+1}\\ m_i^{\mu,1}=\frac{1}{2}\sum_{j=2i-1}^{2i}S_j\xi_j^\mu & \text{with } i=1,2,\dots,2^k \to m_1^{\mu,1}\\ \quad=\frac{1}{2}\sum_{j=1}^{2}S_j\xi_j^\mu & \\ m_i^{\mu,2}=\frac{1}{2^2}\sum_{j=2^2i-(2^2-1)}^{2^2i}S_j\xi_j^\mu & \text{with } i=1,2,\dots,2^{k-1} \to m_1^{\mu,2}\\ \quad=\frac{1}{4}\sum_{j=1}^{4}S_j\xi_j^\mu & \\ \dots & \\ m_1^{\mu,k+1}=\frac{1}{2^{k+1}}\sum_{j=1}^{2^{k+1}}S_j\xi_j^\mu. & \end{cases}$$

As we saw for the Dyson case, this allows writing the fields as

$$\begin{aligned} h_i[\mathbf{S}]&=\sum_{\nu=1}^{p}\xi_i^\nu\sum_{d=1}^{k+1}\Big[\sum_{l=d}^{k+1}\left(\frac{1}{2^{2\sigma}}\right)^l\Big]2^{d-1}m_{f(d,i)}^{\nu,d-1}\\ &=\sum_{\nu=1}^{p}\xi_i^\nu\sum_{d=1}^{k+1}J(d,k+1,\sigma)2^{d-1}m_{f(d,i)}^{\nu,d-1}, \end{aligned}$$

where

$$J(d,k+1,\sigma)2^{\mu-1}=\frac{4^{\sigma-d\sigma}-4^{-k\sigma-\sigma}}{4^\sigma-1}2^{d-1}. \tag{55}$$

The microscopic evolution of the system is defined as a stochastic alignment to local field $h_i[\mathbf{S}]$:

$$S_i(t+\delta t)=\text{sign}\{\tanh[\beta h_i[\mathbf{S}(\mathbf{t})]]+\eta_i(t)\}, \tag{56}$$

where the stochasticity lies in the independent random numbers $\eta_i(t)$ uniformly drawn over the interval $[-1,1]$. In the noiseless limit $\beta\to\infty$ we have

$$S_i(t+\delta t)=\text{sign}[h_i[\mathbf{S}(\mathbf{t})]] \tag{57}$$

and so if $S_ih_i[\mathbf{S}]>0\ \forall\, i\in[1,N]$, the configuration $[\mathbf{S}]$ is dynamically stable.

### 3.4. Signal to noise analysis for serial retrieval

Using Eqs. (52) and (54) and posing $S_i=\xi_i^\mu$ in order to check the robustness of the serial pure-state retrieval (of the test pattern $\mu$), we can write

$$\begin{aligned} \xi_i^\mu h_i[\mathbf{S}]&=\xi_i^\mu\sum_{\nu=1}^{p}\xi_i^\nu\sum_{d=1}^{k+1}J(d,k+1,\sigma)\sum_{j:d_{ij}=d}\xi_j^\nu\xi_j^\mu,\\ &=\sum_{d=1}^{k+1}J(d,k+1,\sigma)2^{d-1}+\xi_i^\mu\sum_{\nu\neq\mu}^{p}\xi_i^\nu\\ &\quad\times\sum_{d=1}^{k+1}J(d,k+1,\sigma)\sum_{j:d_{ij}=d}\xi_j^\nu\xi_j^\mu. \end{aligned} \tag{58}$$

We can decompose the previous equation into two contributions, a stochastic noisy term $R(\xi)$ and a deterministic signal $I$ as

$$\xi_i^\mu h_i[\mathbf{S}]=I+R(\xi). \tag{59}$$

The signal term $I$ is positive because

$$I=\sum_{d=1}^{k+1}J(d,k+1,\sigma)2^{d-1}\geq 0, \tag{60}$$

while the noise $R(\xi)$ has null average (the latter being denoted by standard brackets), namely

$$R(\xi)=\xi_i^\mu\sum_{\nu\neq\mu}^{p}\xi_i^\nu\sum_{d=1}^{k+1}J(d,k+1,\sigma)\sum_{j:d_{ij}=d}\xi_j^\nu\xi_j^\mu, \tag{61}$$

$$\langle R(\xi)\rangle_\xi=0. \tag{62}$$

Thus, in order to see the regions of the tunable parameters $\sigma$, $k+1$ where the signal prevails over the noise and the network accomplishes retrieval, we need to calculate the second moment of the noise over the distribution of quenched variables $\xi$ so to compare the signal amplitudes of $I$ and $|\sqrt{\langle R^2(\xi)\rangle_\xi}|$:

$$\begin{aligned} \langle R^2(\xi)\rangle_\xi&=\Big\langle\Big[\sum_{\nu\neq\mu}^{p}\xi_i^\nu\sum_{d=1}^{k+1}J(d,k+1,\sigma)\sum_{j:d_{ij}=d}\xi_j^\nu\xi_j^\mu\Big]\\ &\quad\times\Big[\sum_{\eta\neq\mu}^{p}\xi_i^\eta\sum_{d=1}^{k+1}J(d,k+1,\sigma)\sum_{j:d_{ij}=d}\xi_j^\eta\xi_j^\mu\Big]\Big\rangle_\xi. \end{aligned} \tag{63}$$

Neglecting off-diagonal terms (as they have null average), we get the following expressions for $\langle R^2(\xi)\rangle_\xi$:

$$\begin{aligned} \langle R^2(\xi)\rangle_\xi&=\Big\langle\sum_{\nu\neq\mu}^{p}(\xi_i^\nu)^2\Big(\sum_{d=1}^{k+1}J(d,k+1,\sigma)\sum_{j:d_{ij}=d}\xi_j^\nu\xi_j^\mu\Big)^2\Big\rangle_\xi\\ &=\Big\langle\sum_{\nu\neq\mu}^{p}\Big(\sum_{d=1}^{k+1}\Big(\frac{4^{\sigma-d\sigma}-4^{-(k+1)\sigma}}{4^\sigma-1}\Big)\sum_{j:d_{ij}=d}\xi_j^\nu\xi_j^\mu\Big)^2\Big\rangle_\xi, \end{aligned} \tag{64}$$

where we used $(\xi_i^\nu)^2=1\ \forall i,\nu$. Once again, as the $\xi$'s are symmetrically distributed, only even order terms give contributions, thus

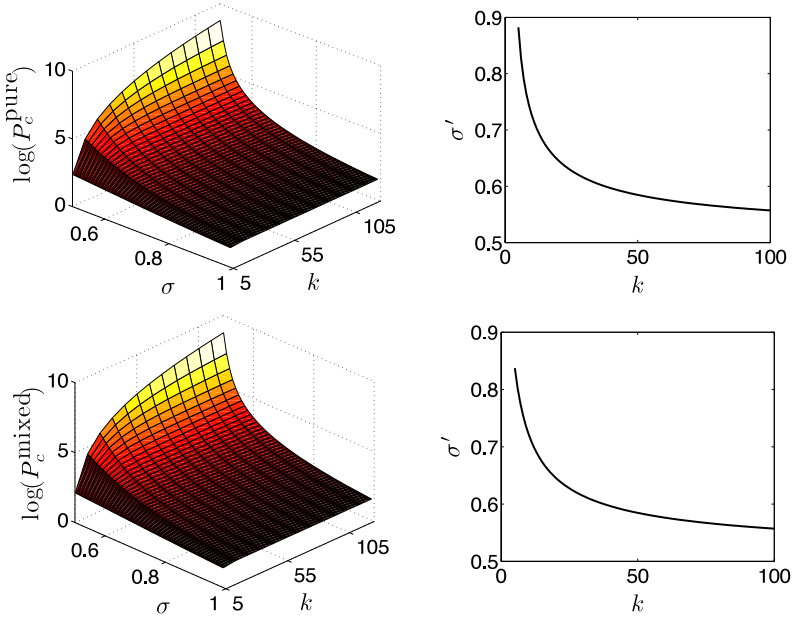


**Fig. 6.** Upper panel (serial retrieval): On the left we show the maximum value of storable patterns $P_c$ as a function of $k$ and of $\sigma$ (as results from Eq. (72), (91)) for the pure (upper panel)/parallel (lower panel) state in order to have signal's amplitude greater than the noise (i.e. retrieval). Note the logarithmic scale for $P_c$ highlighting its wide range of variability. On the right there is the maximum value of the neural interaction decay rate $\sigma'(k)$ versus $k$ allowed to the couplings if we want the retrieval to be possible with a logarithmic storage ($p = k$) of patterns, in the $\beta \to \infty$ limit.

we can safely neglect off-diagonal terms and write again

$$\begin{aligned}\langle R^2(\xi)\rangle_\xi &= (p-1)\sum_{d=1}^{k+1}\left\langle\left[\left(\frac{4^{\sigma-d\sigma}-4^{-k\sigma-\sigma}}{4^\sigma-1}\right)\sum_{j:d_{ij}=d}\xi_j^\nu\xi_j^\mu\right]^2\right\rangle_\xi,\\ &= (p-1)\sum_{d=1}^{k+1}\left(\frac{4^{\sigma-d\sigma}-4^{-k\sigma-\sigma}}{4^\sigma-1}\right)^2\\ &\quad\times\left\langle\sum_{j:d_{ij}=d,}\sum_{k:d_{ik}=d}\xi_j^\nu\xi_j^\mu\xi_k^\nu\xi_k^\mu\right\rangle_\xi. \end{aligned}\tag{65}$$

Therefore

$$\langle R^2(\xi)\rangle_\xi = (p-1)\sum_{d=1}^{k+1}J(d,\sigma,k+1)^2 2^{d-1}.\tag{66}$$

Exploiting the approximation $\langle|x|\rangle \sim |\sqrt{\langle x^2\rangle}|$, we can simplify the previous expression into

$$\langle|R(\xi)|\rangle \sim \sqrt{\langle R^2(\xi)\rangle_\xi} = \sqrt{(p-1)\sum_{d=1}^{k+1}J(d,\sigma,k+1)^2 2^{d-1}},\tag{67}$$

where we consider the positive branch of the serial retrieval only. We are now ready to check the stability of the pure retrieval: as long as

$$I > \sqrt{\langle R^2(\xi)\rangle_\xi} \Rightarrow \xi_i^\mu h_i[\mathbf{S}] = I + R(\xi) > 0,\tag{68}$$

the pure state is stable. Hence we need to calculate explicitly

$$\sqrt{\langle R^2(\xi)\rangle_\xi} = \sqrt{\frac{(p-1)16^{-k\sigma}}{(4^\sigma-2)(4^\sigma-1)^2(16^\sigma-2)}}\cdot\sqrt{\Psi_1+\Psi_2},$$

where

$$\begin{aligned}\Psi_1 &= (4^\sigma-2)4^{2(k+1)\sigma}-3\times 2^{k+2\sigma+1},\\ \Psi_2 &= 2^{k+6\sigma+1}-(16^\sigma-2)2^{2(k+1)\sigma+1}+2^{k+2}-64^\sigma\\ &\quad+2^{2\sigma+1}+2^{4\sigma+1}-4.\end{aligned}$$

The expression for the signal is much simpler, resulting in

$$I = \frac{4^{-(k+1)\sigma}\left(-2^{k+2\sigma+2}+4^{(k+2)\sigma}+2^{k+2}+4^\sigma-2\right)}{-3\times 4^\sigma+16^\sigma+2}.\tag{69}$$

Imposing $I = \sqrt{\langle R^2(\xi)\rangle_\xi}$ and solving for the variable $p$, we find the critical load allowed by the network, namely the function $P_c(\sigma, k)$, whose behavior is shown in Fig. 6:

$$I = \sqrt{\langle R^2(\xi)\rangle_\xi} \Rightarrow P_c(\sigma,k).\tag{70}$$

Now, imposing the relation

$$P_c(\sigma,k) = k$$

and solving numerically with respect to $\sigma$, we can plot the maximum value $\sigma_{\max}(k)$ that the variable $\sigma$ can reach such that the storage $P = k$ produces retrievable patterns, as shown in Fig. 6.

In the infinite size limit we get

$$\begin{aligned}I-\sqrt{\langle R^2(\xi)\rangle} &= \frac{2^{2\sigma}}{-3\times 4^\sigma+16^\sigma+2}\\ &\quad-\frac{\sqrt{(p-1)}2^{2\sigma}}{\sqrt{(4^\sigma-1)(16^\sigma-2)}},\end{aligned}\tag{71}$$

$$P_c(\sigma) = \frac{(4^\sigma-1)(16^\sigma-2)}{(-3\times 4^\sigma+16^\sigma+2)^2}+1.\tag{72}$$

### 3.5. Signal to noise analysis for parallel retrieval

Fixing $S_i = \xi_i^\mu\ \forall i \in [1, 2^k]$ and $S_i = \xi_i^\gamma\ \forall i \in [1+2^k, 2^{k+1}]$ for $\mu \neq \gamma$, namely selecting $\mu$ and $\gamma$ as test patterns to retrieve, we set the system in condition to handle contemporarily two patterns, the former managed by the first half of the neurons, the latter by the second half. The robustness of this state is addressed hereafter following the same prescription outlined so far. Namely, being

$$S_i h_i[\mathbf{S}] = S_i\sum_{\nu=1}^{p}\xi_i^\nu\sum_{d=1}^{k+1}J(d,k+1,\sigma)\sum_{j:d_{ij}=d}\xi_j^\nu S_j,\tag{73}$$

if $i \in [1, 2^k]$ we have

$$S_i h_i(S) = \xi_i^\mu \sum_{\nu=1}^{p} \xi_i^\nu \Bigg( \sum_{d=1}^{k} J(d, k+1, \sigma) \times \sum_{j:d_{ij}=d} \xi_j^\nu \xi_j^\mu + J(k+1, k+1, \sigma) \sum_{j:d_{ij}=k+1} \xi_j^\nu \xi_j^\gamma \Bigg),$$

while if $i \in [2^k+1, 2^{k+1}]$, the same equation still holds provided we replace $\mu$ with $\gamma$ and $\gamma$ with $\mu$, hence hereafter we shall consider only one of the two cases as they are symmetrical. Again, we can decompose the above expression in the sum of a constant, positive term – that plays as the signal – $I > 0$, and a stochastic term for the noise $R(\xi)$, namely we can write

$$S_i h_i[S] = I + R(\xi), \tag{74}$$

$$I = \sum_{d=1}^{k} \Big( J(d, k+1, \sigma) 2^{d-1} \Big),$$

$$\begin{aligned} R(\xi) &= J(k+1, k+1, \sigma) \sum_{j:d_{ij}=k+1} \xi_j^\mu \xi_j^\gamma + \xi_i^\mu \sum_{\nu \neq \mu}^{p} \xi_i^\nu \\ &\quad \times \Big( \sum_{d=1}^{k} J(d, k+1, \sigma) \sum_{j:d_{ij}=d} \xi_j^\nu \xi_j^\mu \\ &\quad + J(k+1, k+1, \sigma) \sum_{j:d_{ij}=k+1} \xi_j^\nu \xi_j^\gamma \Big). \end{aligned}$$

In order to get a manageable expression for the noise, it is convenient to reshuffle $R(\xi)$ distinguishing four terms such that

$$R(\xi) = a + b + c + d, \tag{75}$$

where

$$a = J(k+1, k+1, \sigma) \sum_{j:d_{ij}=k+1} \xi_j^\mu \xi_j^\gamma, \tag{76}$$

$$b = \xi_i^\mu \sum_{\nu \neq \mu}^{p} \xi_i^\nu \sum_{d=1}^{k} J(d, k+1, \sigma) \sum_{j:d_{ij}=d} \xi_j^\nu \xi_j^\mu, \tag{77}$$

$$c = \xi_i^\mu \sum_{\substack{\nu \neq \mu \\ \nu \neq \gamma}}^{p} \xi_i^\nu J(k+1, k+1, \sigma) \sum_{j:d_{ij}=k+1} \xi_j^\nu \xi_j^\gamma, \tag{78}$$

$$d = \xi_i^\mu \xi_i^\gamma J(k+1, k+1, \sigma) 2^k. \tag{79}$$

As $\mu \neq \gamma$, we have that $\langle R(\xi) \rangle_\xi = 0$, while $\langle R^2(\xi) \rangle_\xi$ turns out to be

$$\langle R^2(\xi) \rangle_\xi = \langle a^2 + b^2 + c^2 + d^2 + 2(ab + ac + ad + bc + bd + cd) \rangle_\xi. \tag{80}$$

Let us consider these terms separately: skipping lengthy, yet straightforward calculations, we obtain the following expressions

$$\begin{aligned} \langle a^2 \rangle_\xi &= \Big\langle J^2(k+1, k+1, \sigma) \sum_{j:d_{ij}=k+1} \sum_{n:d_{in}=k+1} \xi_j^\mu \xi_j^\gamma \xi_n^\mu \xi_n^\gamma \Big\rangle_\xi \\ &= J^2(k+1, k+1, \sigma) \times 2^k. \end{aligned} \tag{81}$$

$$\begin{aligned} \langle b^2 \rangle_\xi &= \Bigg\langle \Bigg( \xi_i^\mu \sum_{\nu \neq \mu}^{p} \xi_i^\nu \sum_{d=1}^{k} J(d, k+1, \sigma) \sum_{j:d_{ij}=d} \xi_j^\nu \xi_j^\mu \Bigg)^2 \Bigg\rangle_\xi \\ &= (p-1) \sum_{d=1}^{k} J^2(d, k+1, \sigma) 2^{d-1}. \end{aligned} \tag{82}$$

$$\begin{aligned} \langle c^2 \rangle_\xi &= \Bigg\langle \Bigg( \xi_i^\mu \sum_{\nu \neq \mu \& \nu \neq \gamma}^{p} \xi_i^\nu J(k+1, k+1, \sigma) \sum_{j:d_{ij}=k+1} \xi_j^\nu \xi_j^\gamma \Bigg)^2 \Bigg\rangle_\xi \\ &= (p-2) J^2(k+1, k+1, \sigma) 2^k. \end{aligned} \tag{83}$$

$$\begin{aligned} \langle d^2 \rangle_\xi &= \Big\langle \Big( \xi_i^\mu \xi_i^\gamma J(k+1, k+1, \sigma) 2^k \Big)^2 \Big\rangle_\xi \\ &= J^2(k+1, k+1, \sigma) 2^{2k}, \end{aligned} \tag{84}$$

and, since $a$ and $b$ and, analogously, $b$ and $c$, are defined over different blocks of spins, clearly

$$\langle 2ab \rangle_\xi = 0, \tag{85}$$

$$\langle 2bc \rangle_\xi = 0, \tag{86}$$

$$\langle 2bd \rangle_\xi = 0. \tag{87}$$

As a result, rearranging terms opportunely we finally obtain

$$\begin{aligned} &\langle R^2(\xi) \rangle_\xi \\ &= 4^{-2k\sigma} \Big( \frac{\left[ 4^k (4^\sigma - 1)^2 + 2^k (4^\sigma - 1)^2 + 2^k (p-2) (4^\sigma - 1)^2 \right]}{(4^\sigma - 1)^2} \\ &\quad + (2((-3 \times 2^{k+2\sigma+1} + 2^{k+6\sigma+1} + 2^{k+2} + 2^{2\sigma+1} + 2^{4\sigma+1} \\ &\quad - (4^\sigma - 2) 4^{2(k+1)\sigma} - (16^\sigma - 2) 2^{2(k+1)\sigma+1}) \\ &\quad - 64^\sigma)(p-1))((4^\sigma - 2)(16^\sigma - 2))^{-1} \Big), \end{aligned}$$

while the signal term reads as

$$I = \frac{2^{-2k\sigma-1} \left( -2^{k+2\sigma} - 2^{k+4\sigma} + 2^{2(k+1)\sigma+1} + 2^{k+1} + 2^{2\sigma+1} - 4 \right)}{-3 \times 4^\sigma + 16^\sigma + 2}. \tag{88}$$

Imposing $I = \sqrt{\langle R^2(\xi) \rangle_\xi}$, and solving with respect to the variable $p$ we can outline the function $P_c(\sigma, k+1)$ that returns the maximum allowed load the network may afford accomplishing parallel retrieval, that is shown in Fig. 6 and whose properties were checked by Monte Carlo simulation in Fig. 7:

$$I = \sqrt{\langle R^2(\xi) \rangle_\xi} \Rightarrow P_c(\sigma, k+1). \tag{89}$$

### 3.6. Insights from numerical simulations

Aim of this section is to present results from extensive numerical simulations to check the stability of parallel processing over the finite-size effects that is not captured by statistical mechanics or that can be hidden in the signal-to-noise analysis. Further this allows checking that the asymptotic behavior (in the volume) of the network is in agreement with previous findings.

All the simulations were carried out according to the following algorithm.

1. Building the matrix coupling, pattern storage. Once extracted randomly from a uniform prior over $\pm 1$ $p$ patterns of length $k+1$, and defined the distance between two spins $i$ and $j$ as $d_{ij}$ we build the matrix $\mathbf{J}$, for the HHM, as

$$J_{ij} = \frac{4^{\sigma - d_{ij}\sigma} - 4^{-(k+1)\sigma}}{4^\sigma - 1} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu,$$

for $i = 1, \ldots, 2^{k+1}$, $j = 1, \ldots, 2^{k+1}$, (90)

while for the DHM we use the form:

$$J_{ij} = \frac{4^{\sigma - d_{ij}\sigma} - 4^{-(k+1)\sigma}}{4^\sigma - 1},$$

for $i = 1, \ldots, 2^{k+1}$ and $j = 1, \ldots, 2^{k+1}$, (91)

where $k+1$ is the number of levels of the hierarchical construction of the network, and $\sigma \in (\frac{1}{2}, 1]$.

2. Initialize the network.

We used different initializations to test the stability of the resulting stationary configuration:

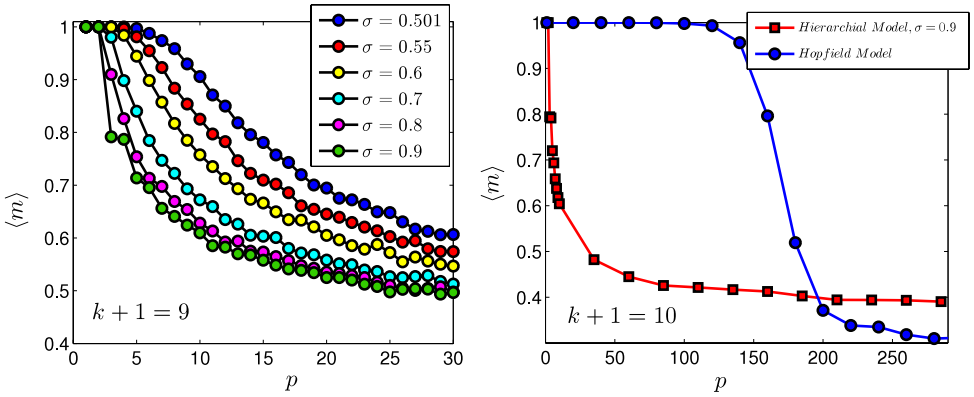


**Fig. 7.** Monte Carlo simulation of the Hierarchical Hopfield network at zero noise level. Left Panel: the averaged magnetization value of the retrieved pattern is plotted versus the amount $p$ of stored patterns for different values of $\sigma$. Accordingly to the analytic estimates, the decreasing of $\sigma$ improves the quality of the retrieval. Right Panel: comparing the standard Hopfield model with the Hierarchical one ($N = 2^{10}$), the averaged magnetization value is plotted versus the amount $p$ of stored patterns. As expected, the Hierarchical Hopfield network is not able to manage an extensive number of stored patterns as the standard Hopfield model does.

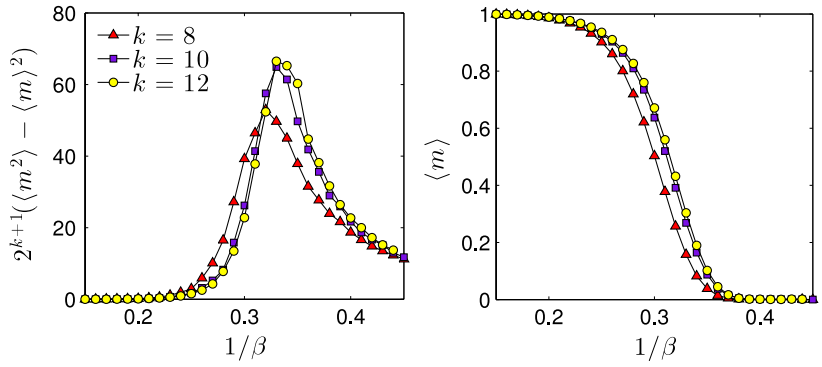


**Fig. 8.** Starting from the state $S_i = +1\ \forall i \in [1, 2^{k+1}]$ results of the simulations for DHM for $\sigma = 0.99$ and $N = 2^{k+1}$, $k+1 = 8, 10, 12$ are plotted. In the left panel, the rescaled magnetic susceptibility $2^{k+1}(\langle m^2 \rangle - \langle m \rangle^2)$ is plotted versus $\beta$ (one over the noise), showing the critical noise level of the paramagnetic–ferromagnetic transition to be the point where the fluctuation of the order parameter has a peak. In the right panel the magnetization $\langle m \rangle = \langle \frac{1}{N} \sum_{i=1}^{N} S_i \rangle$ is plotted versus $\beta$ (one over the noise).

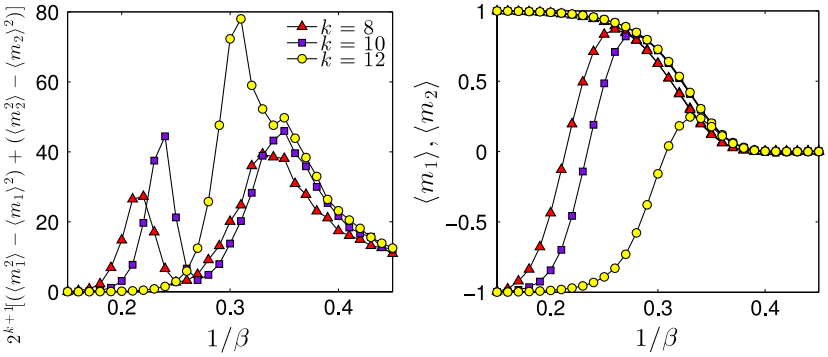


**Fig. 9.** Starting from the state $S_i = +1, S_j = -1\ \forall i \in [1, 2^k]$ and $\forall j \in [2^k + 1, 2^{k+1}]$ results of the simulations for DHM for $\sigma = 0.99$ and $N = 2^{k+1}$ are plotted. In the left panel, the rescaled magnetic susceptibility $2^{k+1}[(\langle m_1^2 \rangle - \langle m_1 \rangle^2) + (\langle m_2^2 \rangle - \langle m_2 \rangle^2)]$, i.e. the total fluctuation of the order parameters, is plotted versus $\beta$ (i.e. one over the noise) for $k+1 = 8, 10, 12$. In the right panel, the magnetizations $\langle m_1 \rangle = \langle \frac{1}{2^k} \sum_{i=1}^{2^k} S_i \rangle$ and $\langle m_2 \rangle = \langle \frac{1}{2^k} \sum_{i=1+2^k}^{2^{k+1}} S_i \rangle$ are plotted versus $\beta$ (i.e. one over the noise) for $k+1 = 8, 10, 12$. The figures show the existence of a dynamical phase transition, beyond the standard paramagnetic–ferromagnetic one, in which the mixed state is no more stable and the system switches on the pure state. This region of instability tends to disappear with the growth of the system size.

– Pure retrieval: We initialize the network in an assumed fixed point of the dynamics, namely $S_i = \xi_i^\mu$ with $i = 1, \ldots, 2^{k+1}$ and $\mu = 1$ for the HHM, while $S_i = +1$ with $i = 1, \ldots, 2^{k+1}$ in the DHM case, and we check the equilibrium as reported in Fig. 8.

– Parallel retrieval: Since we study the multitasking features shown by this hierarchical network, we can also assign different types of initial conditions with respect to the pure state, e.g.

(i) For the DHM, starting from the lowest energy level (after the standard one $S_i = 1\ \forall i$) we chose $S_i = +1$ for $i = 1, \ldots, 2^k$ and $S_i = -1$ for $i = 2^k + 1, \ldots, 2^{k+1}$ (vice versa is the same, and we check the equilibrium as reported in Fig. 9);

(ii) For the HHM, looking for multitasking features, we set in the case $p = 2$, we set $S_i = \xi_i^1$ for $i = 1, \ldots, 2^k$ and $S_i = \xi_i^2$ $i = 2^k + 1, \ldots, 2^{k+1}$ (Fig. 12); in the case $p = 4$ we set $S_i = \xi_i^\mu\ \forall i \in \left[1 + \frac{(\mu-1)N}{4}, \frac{\mu N}{4}\right]$ and $\mu \in [1, 4]$ (Fig. 11).

In this way, we have two or four communities (sharing the same size) building the network with a different order parameter.

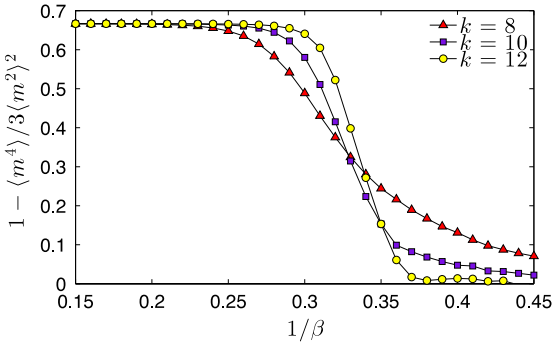


**Fig. 10.** Starting from the state $S_i = +1\ \forall i \in [1, 2^{k+1}]$ with $\sigma = 0.99$ for the DHM and $k + 1 = 8$, 10, 12. Binder cumulant $1 - \frac{\langle m^4 \rangle}{3\langle m^2 \rangle^2}$ versus noise $\frac{1}{\beta}$ for $k + 1 = 8$, 10, 12. Plotting the binder cumulant for different values of $k + 1$ permits to find the critical noise of this state (Binder, 1981).

3. Evolution: Glauber dynamics.
   The evolution of the spins follows a standard random asynchronous dynamics (Coolen et al., 2005) and the state of the network is updated according to the field acting on the spins at every step of iteration, that is,

   $$S_i(t + 1) = \text{sign}\{\tanh[\beta h_i(\mathbf{S}(t))] + \eta(t)\}, \quad \text{for } \beta = T^{-1}$$

   where $\eta(t)$ is the noise introduced as a random uniform contribution over the real interval $[-1, 1]$ in every step.
   For each noise the stationary mean values of the order parameters have been measured mediating over $O(10^3)$ different realizations. For the HHM the average of the order parameters is performed over the quenched variables. For DHM, to better highlight the stability of the parallel configuration, $S_i = +1$ for $i = 1, \ldots, 2^k$, $S_i = -1$ for $i = 2^k + 1, \ldots, 2^{k+1}$, during half of the relaxation period to equilibrium a small positive field is applied to the system.
4. Results.
   It is worth noting that – at difference with paradigmatic prototypes for phase transitions (i.e. the celebrated Curie–Weiss model), as we can see from Figs. 8, 9, 10 – in these models we studied here the critical noise level approaches its asymptotic value (obtained by analytical arguments in the infinite size limit) from above (i.e. from higher values of $\beta$s). This happens because the intensities of couplings are increasing functions (clearly upper limited) of the size of the system. As can be inferred from Fig. 9 (where we present results regarding simulations for the DHM at $\sigma = 0.99$, $k + 1 = 8, 10, 12$ [$S_i = +1, S_j = -1\ \forall i \in [1, 2^k]$ and $\forall j \in [2^k + 1, 2^{k+1}]$]), the mixed state is stable in the low noise region, as expected from theoretical arguments, and the noise region in which this configuration is not stable tends to disappear with the growth of the system size. Also in the HHM case (Fig. 11, 12) the stability of parallel configurations is verified (in the low noise region) for system's configurations shared by the two and four communities.

## 4. Conclusions and outlooks

Comprehension of biological complexity is one of the main aim of this century's research: the route to pave is long and scattered over countless branches. Restricting to neural networks, due to prohibitive constraints when dealing with their statistical–mechanical modeling beyond the mean field approximation (where a notion of distance – or metrics – in the space where neurons are embedded in, is lost), their theory has been largely developed without investigating the crucial degree of freedom of neural distance. However, research is nowadays capable of investigations toward more realistic and/or better performing models: indeed,

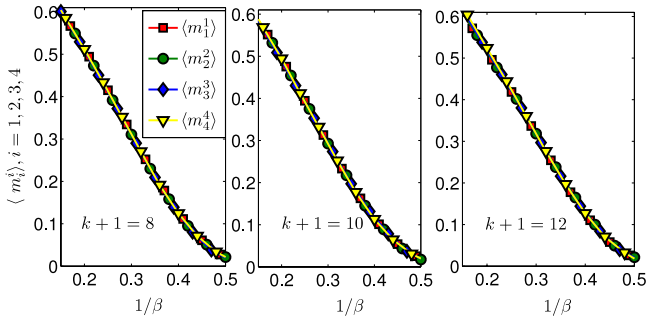


**Fig. 11.** Starting from the state $S_i = \xi_i^1$, $S_j = \xi_j^2, S_n = \xi_n^3, S_l = \xi_l^4\ \forall i \in [1, 2^{k-1}], \forall j \in [2^{k-1} + 1, 2^k]$, $\forall n \in [2^k + 1, \frac{3}{2}2^k]$, $\forall l \in [\frac{3}{2}2^k + 1, 2^{k+1}]$ results of the simulations for HHM for $\sigma = 0.99$ and $N = 2^{k+1}$ are plotted. The Mattis order parameters $\langle m_i^\mu \rangle$ for $i = \mu \in \{1, 4\}$ are different from zero (proving that the parallel retrieval state is stable) and are plotted versus noise. The others are below the noise threshold.

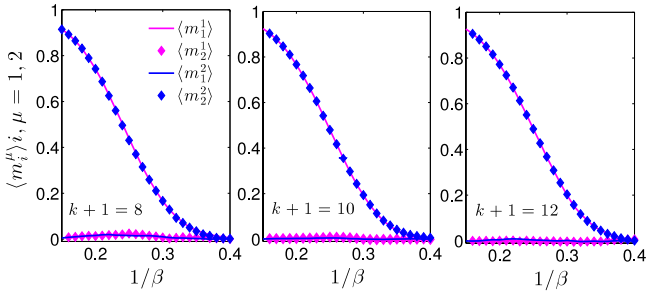


**Fig. 12.** Starting from the state $S_i = \xi_i^1$, $S_j = \xi_j^2\ \forall i \in [1, 2^k]$, $\forall j \in [2^k + 1, 2^{k+1}]$ results of the simulations for HHM for $\sigma = 0.99$ and $N = 2^{k+1}$ are plotted. The Mattis order parameters $\langle m_i^\mu \rangle = \langle \frac{1}{2^{k-2}} \sum_{j=1+(i-1)2^{k-2}}^{i2^{k-2}} S_j \xi_j^\mu \rangle$ for $i, \mu \in [1, 2]$ are plotted versus noise, from left we have $k + 1 = 8$, 10, 12. Again the Mattis magnetizations $m_1^1$ and $m_2^2$ remain different from zero, proving the stability of the parallel retrieval state in which different blocks of spins are aligned with different patterns.

while the mean-field scenario, mainly split among Hopfield network for retrieval and Boltzmann machines for learning, has been so far – at least partially – understood (nor heuristically too far beyond the replica symmetric approximation neither completely, at the rigorous level, within such a scheme), investigation of the non-mean-field counterpart however is only at the embryonal development.

In this work we tackled the problem of studying information processing (retrieval only) on hierarchical topologies, where neurons interact with a Hebbian strength (or simply ferromagnetically in their simplest implementation, namely the Dyson model) that decays with their reciprocal distance. While a full statistical–mechanical treatment is not yet achievable, stringent bounds for its free energy – intrinsically of non-mean-field nature – are however available and return a survey of network capabilities by far richer than the corresponding mean-field counterpart (that is the Hopfield model within the low storage regime). Indeed these network are able to retrieve one pattern at a time accomplishing an extensive reorganization of the whole neuronal state – mirroring serial processing in standard Hopfield networks – but they are also able to switch to multitasking behavior handling multiple patterns at once – without falling into spurious states – hence performing as parallel processors (note that here serial/parallel capabilities are not related to the activities of the neurons – that operate always contemporarily – rather to the patterns the network is able to retrieve at once).

Remarkably, as far as the low storage regime is concerned, the defragmentation of the whole network into effective cliques – crucial for switching to parallel processing – returns a phase space

that shares huge similarities with the multitasking associative networks (Agliari et al., 2012; Sollich et al., 2014).

However, as stringent theorems that definitively confirm this scenario are not yet fully available (we have bounds only), to give robustness to the statistical mechanics predictions, we performed the naive signal-to-noise analysis (Shiino & Fukai, 1992) checking whether those multitasking states – candidate by the first approach to mimic parallel retrieval – are indeed stable beyond the pure state related to serial processing and remarkably we found huge regions of the tunable parameters (strength of the interaction decay $\sigma$ and noise level $\beta$) where those states are extremely robust.

For this richness of behaviors there is however a price to pay: emergent multitasking features in these hierarchical, not-mean-field models require a substantial drop in network's capacity (intuitively because the effective amount of hard links where information may be stored is sensibly lower than in the standard Hopfield networks) thus implying a new balance associative networks beyond the mean-field scenario has to face.

While a satisfactory picture of the behavior of hierarchical neural networks is still to be achieved, we hope that this work may act as one of the first steps toward that direction.

## Acknowledgments

The authors acknowledge partial financial support from the GNFM (INdAM) – Gruppo Nazionale per la Fisica Matematica– [Progetto Giovani Agliari 2014], INFN – Istituto Nazionale di Fisica Nucleare – and Sapienza Universita' di Roma.

AB is grateful to LIFE group (Laboratories for Information, Food and Energy) for partial financial support through POR Calabria FESR 2007/2013 asse 1 programma INNOVA progetto MATCH.

## References

Agliari, E., Annibale, A., Barra, A., Coolen, A. C. C., & Tantari, D. (2013a). Immune networks: multi-tasking capabilities at medium load. *Journal of Physics A*, *46*(33), 335101.

Agliari, E., Annibale, A., Barra, A., Coolen, A. C. C., & Tantari, D. (2013b). Immune networks: multitasking capabilities near saturation. *Journal of Physics A*, *46*(41), 415003.

Agliari, E., & Barra, A. (2011). A Hebbian approach to complex-network generation. *Europhysics Letters*, *94*(1), 10002. http://dx.doi.org/10.1209/0295-5075/94/10002.

Agliari, E., et al. (2013). Parallel processing in immune networks. *Physical Review E*, *87*(4), 042701.

Agliari, E., Barra, A., De Antoni, A., & Galluzzi, A. (2013). Parallel retrieval of correlated patterns: from Hopfield networks to Boltzmann machines. *Neural Networks*, *38*, 52–63.

Agliari, E., Barra, A., Del Ferraro, G., Guerra, F., & Tantari, D. (2014). Anergy in self-directed B lymphocytes: a statistical mechanics perspective. *Journal of Theoretical Biology*, http://dx.doi.org/10.1016/j.jtbi.2014.05.006. in press.

Agliari, E., Barra, A., Galluzzi, A., Guerra, F., & Moauro, F. (2012). Multitasking associative networks. *Physical Review Letters*, *109*(26), 268101.

Agliari, E., Barra, A., Galluzzi, A., Guerra, F., Tantari, D., & Tavani, F. (2015a). Metastable states in Dyson hierarchical model drive parallel processing in Hopfield hierarchical network. *Journal of Physics A: Mathematical and Theoretical*, *48*, 015001.

Agliari, E., Barra, A., Galluzzi, A., Guerra, F., Tantari, D., & Tavani, F. (2015b). Retrieval capabilities of hierarchical networks: from Dyson to Hopfield. *Physical Review Letters*, *114*, 028103.

Albert, R., & Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*, 47–97.

Amit, D. J. (1992). *Modeling brain function: the world of attractor neural network*. Cambridge Univerisity Press.

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1985). Spin-glass models of neural networks. *Physical Review A*, *32*, 1007.

Barra, A., & Agliari, E. (2008). Criticality in diluted ferromagnet. *Journal of Statistical Mechanics*, *10*, 10003.

Barra, A., Contucci, P., Mingione, E., & Tantari, D. (2015). Multi-species mean-field spin-glasses. Rigorous results. *Annales Henri Poincaré*, *16*, 691–708.

Barra, A., Genovese, G., & Guerra, F. (2010). The replica symmetric approximation of the analogical neural network. *Journal of Statistical Physics*, *140*, 784–796.

Barra, A., Genovese, G., & Guerra, F. (2012). Equilibrium statistical mechanics of bipartite spin systems. *Journal of Physics A*, *44*, 245002.

Barra, A., Genovese, G., Guerra, F., & Tantari, D. (2012). How glassy are neural networks? *Journal of Statistical Mechanics*, *07*, 07009.

Barra, A., Genovese, G., Guerra, F., & Tantari, D. (2014). About a solvable mean field model of a Gaussian spin glass. *Journal of Physics A*, *47*(15), 155002.

Binder, K. (1981). Finite size scaling analysis of Ising model block distribution functions. *Zeitschrift fur Physik B Condensed Matter*, Springer.

Bollobas, B. (1998). *Modern graph theory*, *Vol. 184*. Springer Press.

Castellana, M., Barra, A., & Guerra, F. (2014). Free-energy bounds for hierarchical spin models. *Journal of Statistical Physics*, *155*, 211.

Castellana, M., Decelle, A., Franz, S., Mezard, M., & Parisi, G. (2010). The hierarchical random energy model. *Physical Review Letters*, *104*, 127206.

Castellana, M., & Parisi, G. (2011). A renormalization group computation of the critical exponents of hierarchical spin glasses. *Physical Review E*, *83*, 041134.

Coolen, A. C. C., Kuhn, R., & Sollich, P. (2005). *Theory of neural information processing systems*. Oxford University Press.

Dyson, F. J. (1969). Existence of a phase-transition in a one-dimensional Ising ferromagnet. *Communications in Mathematical Physics*, *12*, 91–107.

Ellis, R. S. (1985). *Entropy, large deviations and statistical mechanics*. Springer-Verlag.

Gallavotti, G., & Miracle-Sole', S. (1967). Statistical mechanics of lattice systems. *Communications in Mathematical Physics*, *5*(5), 317–323.

Guerra, F. (2003). Broken replica symmetry bounds in the mean field spin glass model. *Communications in Mathematical Physics*, *233*, 1–12.

Guerra, F., & Toninelli, F. L. (2002). The thermodynamic limit in mean field spin glass models. *Communications in Mathematical Physics*, *230*(1), 71–79.

Hopfield, J. J., & Tank, D. W. (1987). Computing with neural circuits: a model. *Science*, *233*(4764), 625.

Metz, F. L., Leuzzi, L., & Parisi, G. (2014). The renormalization flow of the hierarchical Anderson model at weak disorder. *Physical Review B*, *89*, 064201.

Metz, F. L., Leuzzi, L., Parisi, G., & Sacksteder, V. (2013). Transition between localized and extended states in the hierarchical Anderson model. *Physical Review B*, *88*, 045103.

Mezard, M., Parisi, G., & Virasoro, M. (1987). *Spin glass theory and beyond*. World Scientific Publishing.

Monthus, C., & Garel, T. (2013). Dynamical barriers in the Dyson hierarchical model via real space renormalization. *Journal of Statistical Mechanics*, P02023.

Monthus, C., & Garel, T. (2014). Scaling of the largest dynamical barrier in the one-dimensional long-range Ising spin-glass. *Physical Review B*, *89*, 014408.

Mukamel, D. (2008). Notes on the statistical mechanics of systems with long-range interactions, Les Houches Lecture Notes. arXiv:0905.1457.

Pastur, L., Shcherbina, M., & Tirozzi, B. (1994). The replica-symmetric solution without replica trick for the Hopfield model. *Journal of Statistical Physics*, *74*, 1161–1183.

Perez-Castillo, I., et al. (2004). Analytic solution of attractor neural networks on scale-free graphs. *Journal of Physics A*, *37*, 8789–8799.

Shiino, M., & Fukai, T. (1992). Self-consistent signal-to-noise analysis and its application to analogue neural networks with asymmetric connections. *Journal of Physics A: Mathematical and General*, *25*, L375.

Sollich, P., Tantari, D., Annibale, A., & Barra, A. (2014). Extensive parallel processing on scale free networks. *Physical Review Letters*, *113*, 238106.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of smallworld networks. *Nature*, *393*(6684), 440–442. http://dx.doi.org/10.1038/30918.

Wilson, K. G. (1971a). Renormalization group and critical phenomena. I. Renormalization group and the Kadanoff scaling picture. *Physical Review B*, *4*, 3174.

Wilson, K. G. (1971b). Renormalization group and critical phenomena. II. Phasespace cell analysis of critical behavior. *Physical Review B*, *4*, 3184.

Wilson, K. G. (1972). Feynman-graph expansion for critical exponents. *Physical Review Letters*, *28*, 548.

Wilson, K. G. (1974). Critical phenomena in 3.99 dimensions. *Physica*, *73*, 119.