

Multitasking Associative Networks

Elena Agliari,^{1,2} Adriano Barra,³ Andrea Galluzzi,³ Francesco Guerra,^{3,4} and Francesco Moauro³

¹*Dipartimento di Fisica, Università degli Studi di Parma, viale Usberti 7/A, 43100 Parma, Italy*

²*Istituto Nazionale di Fisica Nucleare, Gruppo Collegato di Parma, 43100 Parma, Italy*

³*Dipartimento di Fisica, Sapienza Università di Roma, Piazzale Aldo Moro 5, 00182 Roma, Italy*

⁴*Istituto Nazionale di Fisica Nucleare, Gruppo di Roma, 00182 Roma, Italy*

(Received 25 February 2012; published 26 December 2012)

We introduce a bipartite, diluted and frustrated, network as a sparse restricted Boltzmann machine and we show its thermodynamical equivalence to an associative working memory able to retrieve several patterns in parallel without falling into spurious states typical of classical neural networks. We focus on systems processing in parallel a finite (up to logarithmic growth in the volume) amount of patterns, mirroring the low-level storage of standard Amit-Gutfreund-Sompolinsky theory. Results obtained through statistical mechanics, the signal-to-noise technique, and Monte Carlo simulations are overall in perfect agreement and carry interesting biological insights. Indeed, these associative networks pave new perspectives in the understanding of multitasking features expressed by complex systems, e.g., neural and immune networks.

DOI: [10.1103/PhysRevLett.109.268101](https://doi.org/10.1103/PhysRevLett.109.268101)

PACS numbers: 84.35.+i, 87.18.-h

Neural networks rapidly became the “harmonic oscillators” of parallel processing: Neurons, thought of as “binary nodes” (spins) of a network, behave collectively to retrieve information, the latter being spread over the synapses, thought of as the interconnections among nodes. However, common intuition of parallel processing is not only the underlying parallel work performed by neurons to retrieve, say, an image on a book, but rather, for instance, to retrieve the image and, while keeping the book securely in hand, noticing beyond its edges the room where we are reading, still maintaining available resources for further retrievals as a safety mechanism.

Standard Hopfield networks are not able to accomplish this kind of parallel processing [1]. Indeed, spurious states, conveying corrupted information, cannot be looked at as the contemporary retrieval of several patterns, but they are rather an unwanted outcome, yielding to a glassy blackout [6]. Such a limit of Hopfield networks can be understood by focusing on the deep connection (in both direct [7] and inverse [8] approach) with restricted Boltzmann machines (RBMs) [9]. In fact, given a machine with its set of visible (neurons) and hidden (training data) units, one gets, under marginalization over the latter, that the thermodynamic evolution of the visible layer is equivalent to that of a Hopfield network. It follows that an underlying fully connected bipartite RBM necessarily leads to bit strings of length equal to the system size and whose retrieval requires an orchestrated arrangement of the whole set of spins. This implies that no resources are left for further tasks, which is, from a biological point of view, too strong a simplification.

The goal of this Letter is to relax this constraint so as to extend standard neural networks toward multitasking capabilities, whose interest goes far beyond the artificial intelligence framework [10–12]. In particular, starting from a

RBM, we perform dilution on its links in such a way that nodes in the external layer are connected to only a fraction of nodes in the inner layer (Fig. 1, left). As we show, this leads to an associative network which, for nonextreme dilutions, is still embedded in a fully connected topology, but the bit strings encoding for information is sparse (i.e., their entries are $+1$, -1 , as well as 0) (Fig. 1, right); for relatively low and large degrees of dilution, this ultimately makes the network able to parallel retrieve without falling into spurious states.

More precisely, let us denote the P binary spins making up the external layer as $\tau_\mu = \pm 1$, $\mu \in [1, \dots, P]$ and the N binary spins making up the internal layer as $\sigma_i = \pm 1$, $i \in [1, \dots, N]$. RBMs admit the Hamiltonian description

$$H(\sigma, \tau; \xi) = \frac{1}{\sqrt{N}} \sum_{i,\mu} \xi_i^\mu \sigma_i \tau_\mu, \quad (1)$$

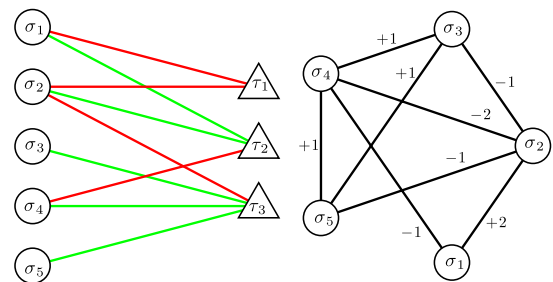


FIG. 1 (color online). Example of diluted RBM, whose layers are made up of $N = 5$ and $P = 3$ elements (left), and its corresponding weakened associative network (right). In the former, brighter (darker) links have positive (negative) coupling; in the latter, the patterns turn out to be $\xi^1 = [-1, -1, 0, 0, 0]$, $\xi^2 = [+1, +1, 0, -1, 0]$, $\xi^3 = [0, -1, +1, +1, +1]$ and the weight associated with each link (i, j) is $\sum_\mu \xi_i^\mu \xi_j^\mu$.

where we called ξ_i^μ the (quenched) interaction strength between the i th spin of the inner layer and the μ th spin of the external layer [possibly to be extracted from a proper probability distribution $P(\xi_i^\mu)$, meant as the outcome of a learning process]. Usually, one defines $\alpha = \lim_{N \rightarrow \infty} P/N$ as the storage value; in this work, we deal with the “low storage” regime, i.e., $P \sim \log N$, corresponding to $\alpha = 0$.

The thermodynamics of the system can be obtained by explicit calculation of the (quenched) free energy $f(\beta)$ via the partition function $Z_{N,P}(\beta)$ [13,14], which read off, respectively, as

$$f(\beta) = \lim_{N \rightarrow \infty} \frac{1}{\beta N} \mathbb{E} \log Z_{N,P}(\beta), \quad (2)$$

$$Z_{N,P}(\beta) = \sum_{\sigma, \tau} \exp[-\beta H(\sigma, \tau; \xi)], \quad (3)$$

\mathbb{E} being the average over the quenched variables ξ . A key point here is that the interaction is one body in each layer, such that marginalizing over one spin variable is straightforward and gives (expanding up to second order the hyperbolic cosine)

$$\begin{aligned} Z_{N,P}(\beta) &= \sum_{\sigma} \prod_{\mu=1}^P \left[\cosh \left(\frac{\beta}{\sqrt{N}} \sum_i \xi_i^\mu \sigma_i \right) \right] \\ &= \sum_{\sigma} e^{(\beta^2/2N) \sum_{i,j} \sum_{\mu=1}^P \xi_i^\mu \xi_j^\mu \sigma_i \sigma_j} \\ &= \sum_{\sigma} e^{-(N\beta^2/2) \sum_{\mu=1}^P m_\mu^2}, \end{aligned} \quad (4)$$

where we introduced the P Mattis magnetizations $m_\mu = N^{-1} \sum_i \xi_i^\mu \sigma_i$. When $P(\xi_i^\mu = +1) = P(\xi_i^\mu = -1) = 1/2$, the Hamiltonian implicitly defined in Eq. (5) recovers exactly the Hopfield model (at a rescaled noise level β^2) and the ansatz of the pure state, i.e., $\mathbf{m} = (1, 0, \dots, 0)$ (under permutational invariance) [6], correctly yields the proper minimization of the free energy in the low-noise limit. This means that, once equilibrium is reached, the system configuration σ is aligned (under gauge invariance) with pattern ξ^1 , relaxation that is understood as recovery of a pattern of information.

As anticipated, here we remove the hypothesis of full connection for the bipartite network, diluting randomly its links in such a way that the coupling distribution gets

$$P(\xi_i^\mu) = \frac{1-d}{2} \delta_{\xi_i^\mu, -1} + \frac{1-d}{2} \delta_{\xi_i^\mu, +1} + d \delta_{\xi_i^\mu, 0}, \quad (6)$$

where $d \in [0, 1]$ is a proper dilution parameter and $\delta_{i,j}$ is the Kronecker delta. It is easy to see that, with this distribution, after marginalizing over one layer as usual, we get an associative network, where the P patterns ξ^μ ($\mu = 1, \dots, P$) contain zeros, on average, for a fraction d of their length (a sparse coding can also be found in Willshaw’s model [15]). As a result, the pure state ansatz can no longer

work. In fact, now the retrieval of a pattern does not employ all spins and those corresponding to null entries can be used to recall other patterns.

In particular, as we will show (see also Fig. 2), at a relatively low degree of dilution ($d < d_{c_1}$), one pattern, say $\mu = 1$, is perfectly retrieved, while a fraction d of spins is still available and its overlap with any remaining pattern is, on average, $1-d$; hence, the second best-retrieved pattern, say $\mu = 2$, displays a (thermodynamical and quenched) average of the Mattis magnetization equal to $d(1-d)$. Proceeding analogously, one finds

$$m_k = d^{k-1}(1-d). \quad (7)$$

The overall number of retrieved patterns K therefore corresponds to $\sum_{k=0}^{K-1} (1-d)d^k = 1$, with the cutoff at finite N as $(1-d)d^{K-1} \geq N^{-1}$, due to discreteness. For any fixed and finite d , this implies $K \lesssim \log N$, which can be thought of as a “parallel low-storage” regime of neural networks.

On the other hand, at larger degrees of dilution ($d > d_{c_1}$) and $P > 2$, this state is no longer stable since no magnetization is large enough to yield a field $\xi_i^\mu m_\mu$ able to align all the related ($\xi_i^\mu \neq 0$) spins; as a result, the system falls into a spurious state where all patterns are partially retrieved, but none exactly. Finally, when dilution is extreme ($1-d \sim P^{-1}$), the retrieval of (nearly) all patterns can still be accomplished. Whenever the global minimum of the system corresponds to the perfect retrieval of at least one pattern, we refer to “multitasking capabilities” or, analogously, to “parallel retrieval.”

Before proceeding with the thermodynamic analysis, we stress that the dilution introduced here is deeply different from the one introduced earlier by Sompolinsky [16] or more recently by Coolen *et al.* [17], who worked out the Hopfield model embedded in random networks, ranging from Erdős-Rényi graphs to small worlds. In those

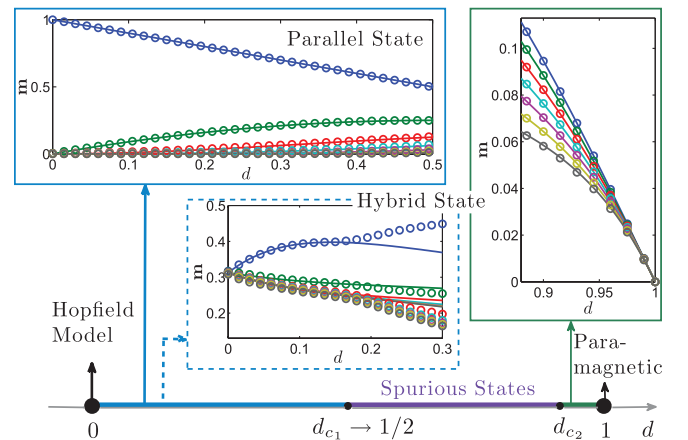


FIG. 2 (color). Schematic representation of the different regimes exhibited by the systems at $\beta \rightarrow \infty$; here, we fixed $P = 7$, for which $d_{c_1} \approx 0.51$, $d_{c_2} \approx 0.89$. Solid (dashed) line frames denote global (local) minima. The states depicted correspond to Eqs. (7) (parallel state) and (11) (hybrid state).

systems, obtained by directly diluting the Hopfield network, the exciting result was the robustness of the (single) retrieval under dilution.

Such different ways of performing dilution—either on links of the associative network (see Ref. [16]) or on pattern entries [see Eq. (6)]—yield dramatically different thermodynamic behaviors. To see this, let us consider the field insisting on each spin, namely, for the generic i th spin $\varphi_i = N^{-1} \sum_{i \neq j=1}^N \sum_{\mu=1}^P \xi_i^\mu \xi_j^\mu \sigma_j$, and analyze its distribution $P(\varphi|d)$ at zero-noise level. When dilution is realized on links in the direct σ - σ network (here d is the fraction of links cut), only an average fraction d of the N available spins participates to φ , in such a way that both the peak and the span of the distribution decrease with d (Fig. 3, left). Conversely, when dilution is realized on links in the underlying bilayer σ - τ network (here d is the fraction of null entries in a pattern), as $d > 0$, $P(\varphi|d)$ gets broader and peaked at smaller values of fields. Indeed, at β , N and P fixed, when dilution is introduced in bit strings, couplings are made *uniformly* weaker (this effect is analogous to a rise in the fast noise), so that the distribution of spin configurations, and consequently also $P(\varphi|d)$, gets broader. For small values of d this effect dominates, while at larger values the overall reduction of coupling range prevails and fields get not only smaller but also more peaked (Fig. 3, right). A topological dilution in the resulting σ - σ network can also be realized in this case, by taking d sufficiently close to 1 [10].

These different scenarios produce different physics, in particular, the latter field distribution can allow parallel retrieval of patterns. The robustness of these multiple basins of attractions can be checked by signal-to-noise

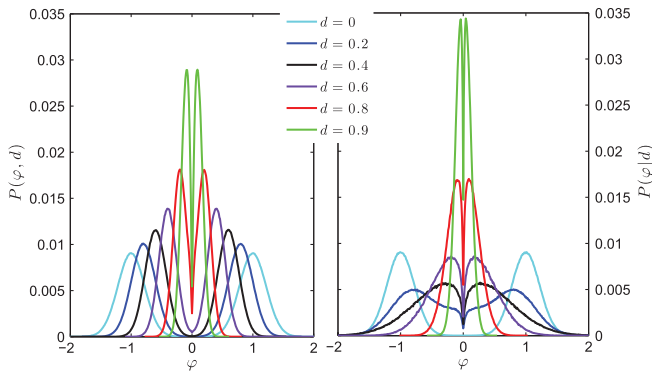


FIG. 3 (color). Distributions of fields φ for dilution *à la* Sompolinsky (left) and for our dilution performed on the bipartite network (right), shown for various degrees of dilution, as explained by the legend; for both systems we fixed $\beta \rightarrow \infty$, $N = 5000$, and $\alpha = 0.05$. In the former case, d represents the average fraction of cut links, in the latter case d represents the average fraction of null pattern entries. As d is tuned, on the left, $P(\varphi|d)$ behaves monotonically corresponding to a Hopfield model embedded on a random graph sparser and sparser, while, on the right, $P(\varphi|d)$ does not behave monotonically and the model is still defined on a fully connected topology.

analysis [10] and by solving the statistical mechanics of the model as sketched in the following. We underline that, as no slow noise due to an extensive amount of patterns is at work ($\alpha = 0$), replica tricks or techniques designed for disordered systems [13,14] are not necessary. We introduce a generic vector for Mattis magnetizations as $\mathbf{m} = (m_1, \dots, m_P)$, a density of the states $\mathcal{D}(\mathbf{m}) = 2^{-N} \sum_{\sigma} \delta[\mathbf{m} - \mathbf{m}(\sigma)]$, and we write the free-energy density as

$$f(\beta) = \frac{\ln 2}{\beta} + \frac{1}{\beta N} \log \int d\mathbf{m} \mathcal{D}(\mathbf{m}) \exp\left(\frac{1}{2} \beta N \mathbf{m}^2\right). \quad (8)$$

After introducing the P -component vector \mathbf{x} to allow integral representation of the P delta functions encoded in the density of the states, and after some algebra, this equation becomes

$$f(\mathbf{m}, \mathbf{x}) = -\frac{1}{\beta N} \int d\mathbf{m} d\mathbf{x} \exp[-N\beta \tilde{f}(\mathbf{m}, \mathbf{x})], \quad (9)$$

$$\tilde{f}(\mathbf{m}, \mathbf{x}) = -\frac{1}{2} \mathbf{m}^2 - i\mathbf{x} \cdot \mathbf{m} - \frac{1}{\beta} \langle \log 2 \cos[\beta \xi \cdot \mathbf{x}] \rangle_{\xi}, \quad (10)$$

whose minimization with respect to \mathbf{m} , \mathbf{x} gives the standard saddle point equation $\mathbf{m} = \langle \xi \tanh(\beta \xi \cdot \mathbf{m}) \rangle_{\xi}$, whose numerical solution for the case $P = 2$ is shown in Fig. 4. When $\beta \rightarrow \infty$, stable retrieved states of amplitude $m_1 = 1 - d$ and $m_2 = d(1 - d)$ are found, in agreement with Eq. (7). On the other hand, in the presence of (fast) noise, the dependence on d of the network performance gets more complex. In fact, for small d , only the first pattern can be retrieved (whenever the fast noise is greater than the signal on m_2) and the parallel ansatz $\mathbf{m} = (d, d(1 - d), d(1 - d)^2, \dots)$ recovers the standard pure one (which can be seen as a particular case of the former). This Hopfield-like behavior persists as long as $d(1 - d) < \beta^{-1}$, above which m_2 also starts to grow and approaches the related zero-noise curve. At intermediate degrees of dilution, the two magnetizations m_1 , m_2 collapse and their amplitude decreases monotonically towards zero. When $d \sim 1$, the signal on both magnetizations is smaller than fast noise so that retrieval is no longer possible and the system behaves paramagnetically. We now explain these features in more detail: We focus on the critical points corresponding to vanishing of magnetizations and to bifurcations, again for the simplest case $P = 2$. The self-consistency equations are

$$\begin{aligned} m_1 &= d(1 - d) \tanh(\beta m_1) \\ &\quad + \frac{(1 - d)^2}{2} [\tanh(\beta y) + \tanh(\beta x)], \\ m_2 &= d(1 - d) \tanh(\beta m_2) \\ &\quad + \frac{(1 - d)^2}{2} [\tanh(\beta y) - \tanh(\beta x)], \end{aligned}$$

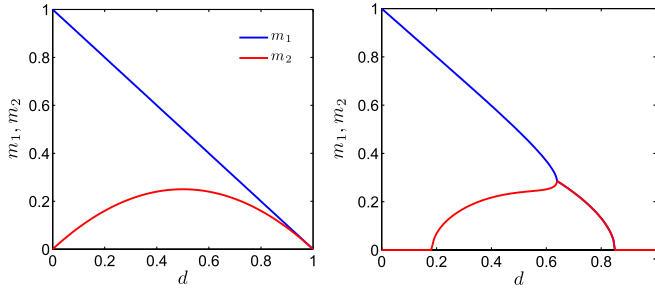


FIG. 4 (color online). Two patterns analysis. Analytical solution at $\beta = 10^4$ (left) and at $\beta = 6.66$ (right). All of these curves have been checked versus Monte Carlo simulations (not shown) with N up to 10^5 spins with overall perfect agreement.

where $y = m_1 + m_2$ and $x = m_1 - m_2$. The critical noise level at which the magnetizations disappear can be obtained by expanding the self-consistent equation for m_2 , namely, $m_2 \sim (1-d)\beta m_2 + \mathcal{O}(m_2^3)$. Therefore, from a standard fluctuation analysis, the critical noise level for the two patterns turns out as $\beta_c = (1-d)^{-1}$, which recovers $\beta_c = 1$ for the standard Hopfield model away from saturation [6]. Critical values of the noise level corresponding to bifurcations can be obtained by expanding for small x , and such calculations can be extended to the case $P > 2$ (see Fig. 5); an extensive treatment of the network performances can be found elsewhere [10].

In general, as mentioned above, the case $P > 2$ can be much more subtle as, even in the noiseless case, it exhibits several phases (see Fig. 2): the parallel ansatz [Eq. (7)] ceases to be stable when $m_1 \leq \sum_{k>1} m_k$, which corresponds to a critical dilution d_{c_1} approaching (exponentially from above) $1/2$ in the limit of large P . Within the same region, a “hybrid state” \tilde{s} , which is a hierarchical mixture of all patterns, is also found to be metastable. More precisely, being $\Xi = \sum_{\mu} \xi_i^{\mu}$,

$$\tilde{s}_i = (1 - \delta_{\Xi,0}) \text{sgn}(\Xi) + \delta_{\Xi,0} [\xi_i^1 + \delta_{\xi_i^1,0} \xi_i^2 + \delta_{\xi_i^1,0} \delta_{\xi_i^2,0} \xi_i^3 + \dots]. \quad (11)$$

This state gets the global minimum whenever $\sum_i (1 - \delta_{\Xi,0}) \text{sgn}(\Xi) \xi_i^{\mu} / N > \sum_{k=1}^{(P-1)/2} \varphi_k (P+1) / (P-k)$, where $\varphi_k = 2 \sum_l [(1-d)/2]^{2l} d^{P-2l} (P-k)! / [l!(l-1)!(P-k-2l+1)!]$ and P is odd. This condition corresponds to $d > d_{c_2}$, where d_{c_2} converges to 1 as P gets larger.

As a final remark, we underline that, although the steady state of the current model and an arbitrary spurious state both display nonzero overlap with several patterns, they are still deeply different. In particular, here the retrieval of multiple patterns corresponds to absolute energy minima (in the noiseless case this holds for any $d > 0$) and at least one pattern is exactly retrieved. However, the present model is not devoid of genuine spurious states, which are, in general, mixtures of all patterns. These states can be destabilized by decreasing β (analogously to the

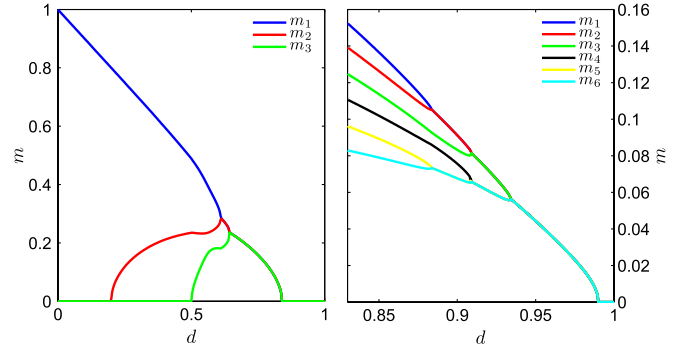


FIG. 5 (color online). Retrieval of three (left) and six (right) patterns at $\beta = 10^2$; for the latter, we zoomed on the region of high dilution where bifurcations occur. The number of such bifurcations is, in general, upper bounded by $P - 1$.

standard Hopfield model) or, interestingly, by either increasing or decreasing d .

In summary, the structural equivalence of associative networks and RBMs allows significant developments, both practical and theoretical. For instance, one can simulate the dynamics of these networks by dealing with an update of $N + P$ spins and a storage of only NP synapses, instead of updating N spins and storing $\sim N^2$ synapses. Moreover, the equivalence suggests that traditional associative networks, where the whole set of neurons needs to be properly arranged in order to achieve retrieval, are not optimal. We overcome this constraint by diluting the links of the RBM, which translates into partially blank patterns. Interestingly, the resulting associative network is not only still able to perform retrieval, but it can actually retrieve several contemporary patterns, without falling into spurious states. This is an important step toward real autonomous parallel processing and may find applications not only in artificial intelligence [12] but also in biological contexts [11]. For instance, when applied to the modeling of the adaptive immune system, this result allows us to see that the (lymphocyte) network is able to successfully respond to several pathogens at once [10].

The FIRB Grant No. RBFR08EKEV, Sapienza University, and INFN are acknowledged for financial support.

-
- [1] Notable examples of neural networks where (possibly partial) retrieval of more than one pattern of information is accomplished can be found in Refs. [2,3], where (a fixed amount of) patterns were contemporary retrieved owing to the correlation among them, in Ref. [4], where partial inhibition of a single pattern recovery is allowed by chemical modulation, and in Ref. [5], where the integrate-and-fire neurons are considered, but this approach is still far away from the associative network framework.
- [2] L. F. Cugliandolo and M. V. Tsodyks, *J. Phys. A* **27**, 741 (1994).

- [3] M. Griniasty, M.V. Tsodyks, and D.J. Amit, *Neural Comput.* **5**, 1 (1993).
- [4] A.C.C. Coolen and A.J. Noest, *J. Phys. A* **23**, 575 (1990).
- [5] D.J. Amit, A. Bernacchia, and V. Yakovlev, *Cereb. Cortex* **13**, 435 (2003).
- [6] D.J. Amit, *Modeling Brain Function* (Cambridge University Press, Cambridge, England, 1989).
- [7] A. Barra, A. Bernacchia, E. Santucci, and P. Contucci, *Neural Netw.* **34**, 1 (2012).
- [8] G. Tkacik, E. Schneidman, M.J. Berry II, and W. Bialek, [arXiv:0912.5409](https://arxiv.org/abs/0912.5409).
- [9] Y. Bengio, *Mach. Learn.* **2**, 1 (2009).
- [10] E. Agliari, A. Barra, A. Galluzzi, F. Guerra, and F. Moauro, [arXiv:1202.6326](https://arxiv.org/abs/1202.6326).
- [11] M.F. Bear, B.W. Connors, and M.A. Paradiso, *Neuroscience: Exploring the Brain* (Lippincott, Williams & Wilkins, Philadelphia, PA, 2007).
- [12] K. Hwang and A. Faye, *Computer Architecture and Parallel Processing* (McGraw-Hill, New York, 2008).
- [13] A. Barra, F. Guerra, and G. Genovese, *J. Stat. Phys.* **140**, 784 (2010).
- [14] M. Mézard, G. Parisi, and M.A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
- [15] D. Golomb, N. Rubin, and H. Sompolinsky, *Phys. Rev. A* **41**, 1843 (1990).
- [16] H. Sompolinsky, *Phys. Rev. A* **34**, 2571 (1986).
- [17] B. Wemmenhove and A. C. C. Coolen, *J. Phys. A* **36**, 9617 (2003).