Journal of Statistical Mechanics:
Theory and Experiment





# Parallel learning by multitasking neural networks

OPEN ACCESS

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Interdisciplinary statistical mechanics

# Parallel learning by multitasking neural networks

**Elena Agliari**[1], **Andrea Alessandrelli**[2,5], **Adriano Barra**[3,5,*] **and Federico Ricci-Tersenghi**[4,5,6]

[1] Dipartimento di Matematica, Sapienza Università di Roma, Piazzale Aldo Moro, 5, 00185 Roma, Italy
[2] Dipartimento di Informatica, Università di Pisa, Lungarno Antonio Pacinotti, 43, 56126 Pisa Italy
[3] Dipartimento di Matematica e Fisica, Università del Salento, Via per Arnesano, 73100 Lecce, Italy
[4] Dipartimento di Fisica, Sapienza Università di Roma, Piazzale Aldo Moro 2, I-00185 Roma, Italy
[5] Istituto Nazionale di Fisica Nucleare, Sezioni di Roma1 e Lecce, Italy
[6] CNR-Nanotec, Rome unit, 00185 Rome, Italy
E-mail: adriano.barra@gmail.com

Received 18 August 2023
Accepted for publication 22 October 2023
Published 27 November 2023

Online at stacks.iop.org/JSTAT/2023/113401
https://doi.org/10.1088/1742-5468/ad0a86

**Abstract.** *Parallel learning*, namely the simultaneous learning of multiple patterns, constitutes a modern challenge for neural networks. While this cannot be accomplished by standard Hebbian associative neural networks, in this paper we show how the multitasking Hebbian network (a variation on the theme of the Hopfield model, working on sparse datasets) is naturally able to perform this complex task. We focus on systems processing in parallel a finite (up to logarithmic growth in the size of the network) number of patterns, mirroring the low-storage setting of standard associative neural networks. When patterns to be reconstructed are mildly diluted, the network handles them hierarchically, distributing the amplitudes of their signals as power laws w.r.t. the pattern

*Author to whom any correspondence should be addressed.



1742-5468/23/113401+38$33.00

information content (hierarchical regime), while, for strong dilution, the signals pertaining to all the patterns are simultaneously raised with the same strength (parallel regime). Further, we prove that the training protocol (either supervised or unsupervised) neither alters the multitasking performances nor changes the thresholds for learning. We also highlight (analytically and by Monte Carlo simulations) that a standard cost function (i.e. the Hamiltonian) used in statistical mechanics exhibits the same minima as a standard loss function (i.e. the sum of squared errors) used in machine learning.

**Keywords:** machine learning, computational neuroscience, optimization over networks, systems neuroscience

J. Stat. Mech. (2023) 113401

## Contents

## 1. Introduction

Typically, neural networks have to deal with several inputs occurring at the same time: for instance, think about automatic driving, i.e. artificial neural networks, where they have to distinguish and react to different objects (e.g. pedestrians, traffic lights, riders, crosswalks) that may appear simultaneously. Likewise, when a biological neural network learns, it rarely has to deal with one single input at a time[7]: for instance, while trained in school to learn any single letter, we are also learning about the composition of our alphabets. From this perspective, when stating that neural networks operate *in parallel*, some caution on potential ambiguity should be paid. To fix these ideas, let us focus on the Hopfield model [27], the *harmonic oscillator* of associative neural networks accomplishing pattern recognition [12, 19]: its neurons indeed operate synergistically in parallel but with the purpose of retrieving one single pattern at a time, not several simultaneously [12, 14, 29]. Parallel processing, where multiple patterns are simultaneously retrieved, cannot be accessible to the standard Hopfield networks as long as each pattern is fully informative; namely, its vectorial binary representation is devoid of blank entries. On the other hand, when a fraction of entries can be blank [6], multiple-pattern retrieval is potentially achievable by the network. Intuitively, this can be explained by noticing that the overall number of neurons making up the networks—and thus available for information processing—equals the length of the binary vectors codifying the patterns to be retrieved. Hence, as long as these vectors contain information in all their entries, as one pattern is retrieved there will no longer be neurons available for retrieving other patterns at the same time. Conversely, the multitasking neural networks, introduced in [7], are able to overcome this limitation, and have been shown to succeed in retrieving multiple patterns simultaneously just by leveraging the presence of lacunæ in the patterns stored by the network. Their emerging pattern recognition properties have been extensively investigated at medium storage (i.e. on random graphs above the percolation threshold) [4] and at high storage (i.e. on random graphs below the percolation threshold) [5], as well as on scale-free [33] and hierarchical [8] topologies.

[7] It is enough to note that, should serial learning take place rather than parallel learning, Pavlov's classical conditioning would not be possible [6].

J. Stat. Mech. (2023) 113401

While the study of the parallel retrieval capabilities of these multitasking networks is now complete, comprehension of their parallel learning capabilities has just started, and it is the main focus of the present paper. In this regard, it is important to stress that the Hebbian prescription has been recently revised to turn it from a storing rule (built on a set of already definite patterns, as in the original Amit–Gutfreund–Sompolinksy (AGS) theory) into a genuine learning rule (where unknown patterns have to be inferred by experiencing solely a sample of their corrupted copies), see e.g. [2, 11, 22][8].

In this work, we merge these extensions of the bare AGS theory and use definite patterns, equipped with blank entries, to generate a sparse dataset of corrupted examples: that is the only information experienced by the network. Given this setting, we aim to highlight the role of lacunæ density and of the dataset size and quality on the network performancein particular, deepening the way the network simultaneously learns the patterns hidden behind the supplied examples. In this investigation, we focus on the low-storage scenario (where the number of definite patterns grows sub-linearly with the volume of the network) addressing both the *supervised* and the *unsupervised* setting.

The paper is structured as follows. In section 2, for the sake of completeness, we review the multitasking associative network; after briefly summarizing its parallel retrieval capabilities (section 2.1), we introduce a simple dataset that the network has to cope with in order to move from the simpler storing of patterns to their learning from examples (section 2.2). Next, in section 3 we provide an exhaustive statistical-mechanics picture of the network's emergent information-processing capabilities by taking advantage of Guerra's interpolation techniques [3, 16, 25]. In particular, focusing on the cost function (section 3.1), we face the *big-data* limit (section 3.1.1), we deepen the nature of the phase transition the network undergoes as ergodicity breaking spontaneously takes place (section 3.1.2), and we outline phase diagrams, namely plots in the space of the control parameters where different regions depict different global computational capabilities (section 3.2). Further, in section 3.3 we show how the network's cost function (typically used in statistical mechanics) can be strongly related to standard loss functions (typically used in machine learning) to appreciate how parallel learning effectively lowers several loss functions at once. Finally, in section 4 we summarize the results and discuss outlooks.

In the appendices we consider several subtleties: in appendix A we provide a more general setting for the sparse datasets considered in this research[9], while in appendix B we inspect the relative entropies of these datasets. Appendices C and D give details of the calculations, plots and proofs of the main theorems.

[8] While statistical learning theories appeared in the literature a long time ago, see e.g. [1, 23, 32] for the original works and [10, 18, 21, 30] for updated references, the statistical mechanics of Hebbian learning were not deepened in these studies and only generalization capabilities were addressed [24].

[9] In the main text we address the simplest kind of pattern dilution; namely, we just force to be blank the same fraction of their entries whose position is preserved in the generation of the datasets (hence, whenever the pattern has a zero, in all the examples it gives rise to, the zero will be kept), while in the appendix we relax this assumption (and blank entries can move along the examples still preserving their amount). As in the thermodynamic limit the theory is robust w.r.t. these structural details we present as a main theme the simplest setting and in appendix A the more cumbersome one.



J. Stat. Mech. (2023) 113401

## 2. Parallel learning in multitasking Hebbian neural networks

### 2.1. A preliminary glance at the emergent parallel retrieval capabilities

Hereafter, for the sake of completeness, we briefly review the retrieval properties of the multitasking Hebbian network in the low-storage regime, while we refer to [7, 9] for an extensive treatment.

**Definition 1.** Given $N$ Ising neurons $\sigma_i = \pm 1$ $(i = 1,\dots,N)$, and $K$ random patterns $\boldsymbol{\xi}^\mu$ $(\mu = 1,\dots,K)$, each of length $N$, whose entries are i.i.d. from

$$\mathbb{P}(\xi_i^\mu) = \frac{(1-d)}{2}\delta_{\xi_i^\mu,-1} + \frac{(1-d)}{2}\delta_{\xi_i^\mu,+1} + d\delta_{\xi_i^\mu,0}, \tag{2.1}$$

where $\delta_{i,j}$ is the Kronecker delta and $d \in [0,1]$, the Hamiltonian (or cost function) of the system reads as

$$\mathcal{H}_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) := -\frac{1}{2N}\sum_{\substack{i,j \\ i\neq j}}^{N,N}\left(\sum_{\mu=1}^{K}\xi_i^\mu\xi_j^\mu\right)\sigma_i\sigma_j. \tag{2.2}$$

The parameter $d$ tunes the 'dilution' in pattern entries: if $d = 0$ the standard Rademacher setting of AGS theory is recovered, while for $d = 1$ no information is retained in these patterns: otherwise stated, these vectors display, on average, a fraction $d$ of blank entries.

**Definition 2.** In order to assess the network retrieval performance, we introduce the $K$ Mattis magnetizations

$$m_\mu := \frac{1}{N}\sum_{i}^{N}\xi_i^\mu\sigma_i,\ \mu = 1,\dots,K, \tag{2.3}$$

which quantify the overlap between the generic neural configuration $\boldsymbol{\sigma}$ and the $\mu$th pattern.

Note that the cost function (2.2) can be recast as a quadratic form in $m_\mu$, namely

$$\mathcal{H}_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{N}{2}\sum_{\mu}m_\mu^2 + \frac{1}{2N}\sum_{\mu,i=1}^{K,N}(\xi_i^\mu)^2, \tag{2.4}$$

where the last term on the r.h.s. stems from diagonal terms $(i = j)$ in the sum on the r.h.s. of equation (2.2); its mean is $K(1-d)/2$ and in the low-load scenario (i.e. $K$ grows sub-linearly with $N$) can be neglected in the thermodynamic limit $N \to \infty$.

As we will explain, the dilution ruled by $d$ is pivotal for the network in order to perform parallel processing. It is instructive to first consider a toy model handling just $K = 2$ patterns. Let us assume, for simplicity, that the first pattern $\boldsymbol{\xi}^1$ contains



J. Stat. Mech. (2023) 113401

information (i.e. no blank entries) solely in the first half of its entries, and the second pattern $\boldsymbol{\xi}^2$ contains information solely in the second half of its entries; that is,

$$\boldsymbol{\xi}^1 = \left( \underbrace{\xi_1^1, \dots, \xi_{N/2}^1}_{\in\{-1,+1\}^{\frac{N}{2}}}, \underbrace{0, \dots, 0}_{\in\{0\}^{\frac{N}{2}}} \right), \quad \boldsymbol{\xi}^2 = \left( \underbrace{0, \dots, 0}_{\in\{0\}^{\frac{N}{2}}}, \underbrace{\xi_{N/2+1}^1, \dots, \xi_N^1}_{\in\{-1,+1\}^{\frac{N}{2}}} \right). \tag{2.5}$$

Unlike the standard Hopfield reference ($d=0$), where the retrieval of one pattern employs all the resources and there is no chance to retrieve any other pattern, not even partially (i.e. as $m_1 \to 1$ then $m_2 \approx 0$ because patterns are orthogonal for large $N$ values in the standard random setting), here nor $m_1$ neither $m_2$ can reach the value 1 and therefore the complete retrieval of one of the two still leaves resources for the retrieval of the other. In this particular case, the minimization of the cost function $\mathcal{H}_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = -\frac{N}{2}\left(m_1^2 + m_2^2\right)$ is optimal when *both* the magnetizations are equal to one-half, that is when they both saturate their upper bound. In general, for an arbitrary dilution level $d$, the minimization of the cost function requires the network to be in one of the following regimes:

- *Hierarchical scenario*: for values of dilution not too high (i.e. $d < d_c$, *vide infra*), one of the two patterns is fully retrieved (say $m_1 \approx 1-d$) and the other is retrieved to the largest extent given the available resources, these being constituted by, approximately, the $Nd$ neurons corresponding to the blank entries in $\boldsymbol{\xi}^1$ (thus, $m_2 \approx d(1-d)$), and so on if further patterns are considered.
- *Parallel scenario*: for large values of dilution (i.e. above a critical threshold $d_c$), the magnetizations related to all the patterns rise and the signals they convey share the same amplitude.

In general, in this type of neural network, the *pure state ansatz*[10] $\boldsymbol{m} = (1,0,0,\dots,0)$, that is, $\sigma_i = \xi_i^1$ for $i = 1,\dots,N$, barely works, and parallel retrieval is often favored. In fact, for $K \geqslant 2$, at relatively low values of pattern dilution $d_c$ and in the zero-noise limit $\beta \to \infty$, one can prove the validity of the so-called *hierarchical ansatz* [7] as we briefly discuss: one pattern, say $\boldsymbol{\xi}^1$, is perfectly retrieved and displays a Mattis magnetization $m^1 \approx (1-d)$; a fraction $d$ of neurons is not involved and is therefore available for further retrieval, with any remaining pattern, say $\boldsymbol{\xi}^2$, which yields $m_2 \sim (1-d)d$; proceeding iteratively, one finds $m_\ell = d^{\ell-1}(1-d)$ for $\ell = 1,\dots,\hat{K}$ and the overall number $\hat{K}$ of patterns simultaneously retrieved corresponds to the employment of all the resources. Specifically, $\hat{K}$ can be estimated by setting $\sum_{\ell=0}^{\hat{K}-1}(1-d)d^\ell = 1$, with the cutoff at finite $N$ as $(1-d)d^{\hat{K}-1} \geqslant N^{-1}$, due to discreteness: for any fixed and finite $d$, this implies $\hat{K} \lesssim \log N$, which can be thought of as a 'parallel low-storage' regime of neural networks. It is worth stressing that, in the above-mentioned regime of low dilution, the configuration leading to $m_\ell = d^{\ell-1}(1-d)$ for $\ell = 1,\dots,\hat{K}$ is the one that minimizes the

[10] In this state the neurons are aligned with one of the patterns and, without loss of generality, here we refer to $\mu = 1$.



J. Stat. Mech. (2023) 113401

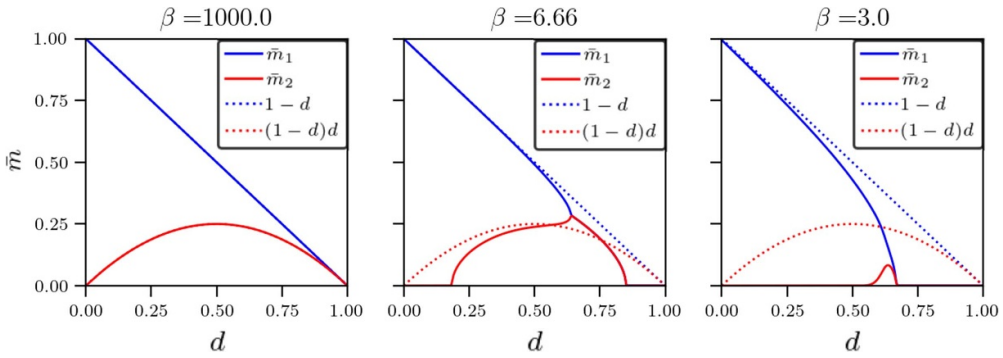


**Figure 1.** Numerical solutions of the two self-consistent equations (2.7) and (2.8) obtained for $K=2$, see [7], as a function of $d$ and for different choices of $\beta$: in the $d\to 0$ limit the Hopfield serial retrieval is recovered (one magnetization with intensity one and the other locked at zero), for $d\to 1$ the network ends up in the parallel regime (where all the magnetizations acquire the same value), while for intermediate values of dilution the hierarchical ordering prevails (both the magnetizations are raised, but their amplitude is different).

cost function. The hierarchical retrieval state $\boldsymbol{m}=(1-d)\left(1,d,d^2,d^3,\ldots\right)$ can also be specified in terms of neural configuration as [7]

$$\sigma_i^*=\xi_i^1+\sum_{\nu=2}^{\hat{K}}\xi_i^\nu\prod_{\rho=1}^{\nu-1}\delta_{\xi_i^\rho,0}\,. \tag{2.6}$$

This organization is stable until a critical dilution level $d_c$ is reached, where $m_1\sim\sum_{k>1}m_k$ [7]. Beyond that level the network undergoes a rearrangement and a new organization called a *parallel ansatz* supplants the previous one. Indeed, for high values of dilution (i.e. $d\to 1$) it is immediate to check that the ratio among the various intensities of all the magnetizations stabilizes to the value one, i.e. $(m_k/m_{k-1})\sim d^{k-1}(1-d)/d^{k-2}(1-d)\to 1$; hence, in this regime all the magnetizations are raised with the same strength and the network is operationally set in a fully parallel retrieval mode: the parallel retrieval state simply reads $\boldsymbol{m}=(\bar{m})\,(1,1,1,1,\ldots)$. This picture is confirmed by the plots shown in figure 1 and obtained by solving the self-consistency equations for the Mattis magnetizations related to the multitasking Hebbian network equipped with $K=2$ patterns that read as [7]

$$m_1=d(1-d)\tanh(\beta m_1)+\frac{(1-d)^2}{2}\left\{\tanh\left[\beta\left(m_1+m_2\right)\right]+\tanh\left[\beta\left(m_1-m_2\right)\right]\right\}, \tag{2.7}$$

$$m_2=d(1-d)\tanh(\beta m_1)+\frac{(1-d)^2}{2}\left\{\tanh\left[\beta\left(m_1+m_2\right)\right]-\tanh\left[\beta\left(m_1-m_2\right)\right]\right\} \tag{2.8}$$

where $\beta\in\mathbb{R}^+$ denotes the level of noise.

We remark that these hierarchical or parallel organizations of the retrieval, beyond emerging naturally within the equilibrium description provided by statistical mechanics,



J. Stat. Mech. (2023) 113401

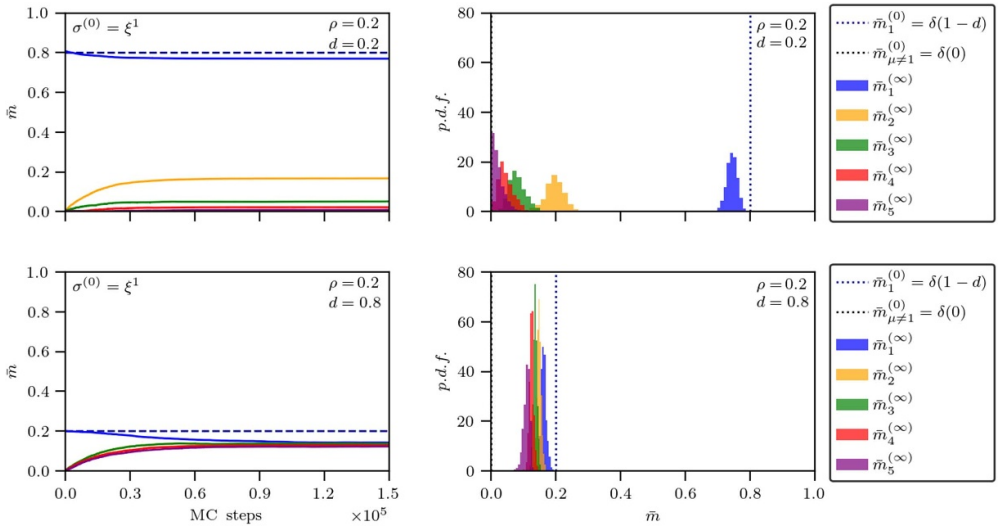


**Figure 2.** We report two examples of Monte Carlo dynamics until thermalization within the hierarchical (upper plots, dilution level $d = 0.2$) and parallel (lower plots, dilution level $d = 0.8$) scenarios, respectively. These plots confirm that the picture provided by statistical mechanics is actually dynamically reached by the network. We initialize the network sharply in a pattern as a Cauchy condition (represented as the dotted blue Dirac delta peaked at the pattern in the second columns) and, in the first column, we show the stationary values of the various Mattis magnetizations pertaining to different patterns, while in the second column we report their histograms achieved by sampling 1000 independent Monte Carlo simulations: starting from a sequential retrieval regime, the network ends up in a multiple retrieval mode, hierarchical vs parallel depending on the level of dilution in the patterns.

are actually the real stationary states of the dynamics of these networks at work with diluted patterns as shown in figure 2.

### 2.2. From parallel storing to parallel learning

In this section, we revise the multitasking Hebbian network [7, 9] in such a way that it can undergo a *learning* process instead of a simple *storing* of patterns. In fact, in the typical learning setting, the set of definite patterns, hereafter referred to as 'archetypes', to be reconstructed by the network is not available; rather, the network is exposed to examples, namely noisy versions of these archetypes.

As long as enough examples are provided to the network, this is expected to correctly form its own representation of the archetypes such that, in further expositions to a new example related to a certain archetype, it will be able to retrieve it and, since then, suitably generalize it. This generalized Hebbian kernel has recently been introduced to encode unsupervised [2] and supervised [11] learning processes and, in the present paper, these learning rules are modified in order to deal with diluted patterns.

First, let us define the dataset that these networks have to cope with: the archetypes are randomly drawn from the distribution (2.1). Each archetype $\boldsymbol{\xi}^\mu$ is then used to generate a set of $M_\mu$ perturbed versions, denoted as $\boldsymbol{\eta}^{\mu,a}$ with $a = 1, \ldots, M_\mu$ and



J. Stat. Mech. (2023) 113401

$\boldsymbol{\eta}^{\mu,a} \in \{-1,0,+1\}^N$. Thus, the overall set of examples to be supplied to the network is given by $\boldsymbol{\eta} = \{\boldsymbol{\eta}^{\mu,a}\}_{\mu=1,\dots,K}^{a=1,\dots,M_\mu}$. Of course, different ways to sample examples are conceivable: for instance, one can require that the position of blank entries appearing in $\boldsymbol{\xi}^\mu$ is preserved over all the examples $\{\boldsymbol{\eta}^{\mu,a}\}_{a=1,\dots,M_\mu}$, or one can require that only the number of blank entries $\sum_{i=1}^N \delta_{\xi_i^\mu,0}$ is preserved (either strictly or on average). Here, we face the first case because it requires a simpler notation, but we refer to appendix A for a more general treatment.

**Definition 3.** The entries of each example are drawn following

$$\mathbb{P}(\eta_i^{\mu,a}|\xi_i^\mu) = \frac{1+r_\mu}{2}\delta_{\eta_i^{\mu,a},\xi_i^\mu} + \frac{1-r_\mu}{2}\delta_{\eta_i^{\mu,a},-\xi_i^\mu}, \tag{2.9}$$

for $i = 1,\dots,N$ and $\mu = 1,\dots,K$. Notice that $r_\mu$ tunes the dataset quality: as $r_\mu \to 1$, examples belonging to the $\mu$th set collapse on the archetype $\boldsymbol{\xi}^\mu$, while as $r_\mu \to 0$, examples turn out to be uncorrelated with the related archetype $\boldsymbol{\xi}^\mu$[11].

As we will show in the next sections, the behavior of the system depends on the parameters $M_\mu$ and $r_\mu$ only through the combination $\frac{1-r_\mu^2}{M_\mu r_\mu^2}$; therefore, as long as the ratio $\frac{1-r_\mu^2}{M_\mu r_\mu^2}$ is $\mu$-independent, the theory is not affected by the specific choice of the archetype. Thus, for the sake of simplicity, hereafter we will consider $r$ and $M$ independent of $\mu$ and we will pose $\rho := \frac{1-r^2}{Mr^2}$. Remarkably, $\rho$ acts as an information-content control parameter [11]: to see this, let us focus on the $\mu$th pattern and $i$th digit, whose related block is $\boldsymbol{\eta}_i^\mu = (\eta_i^{\mu,1}, \eta_i^{\mu,2}, \dots, \eta_i^{\mu,M})$, the error probability for any single entry is $\mathbb{P}(\xi_i^\mu \neq 0)\mathbb{P}(\eta_i^{\mu,a} \neq \xi_i^\mu) = (1-d)(1-r_\mu)/2$ and, by applying the majority rule on the block, we get $\mathbb{P}(\xi_i^\mu \neq 0)\mathbb{P}(\text{sign}(\sum_a \eta_i^{\mu,a})\xi_i^\mu = -1) \underset{M\gg 1}{\approx} \frac{(1-d)}{2}\left[1 - \text{erf}\left(1/\sqrt{2\rho}\right)\right]$. Thus, by computing the conditional entropy $H_d(\xi_i^\mu|\boldsymbol{\eta}_i^\mu)$ that quantifies the amount of information needed to describe the original message $\xi_i^\mu$ given the related block $\boldsymbol{\eta}_i^\mu$, we get

$$\begin{aligned} H_d(\xi_i^\mu|\boldsymbol{\eta}_i^\mu) = &-\left[\frac{1+d}{2} + \frac{1-d}{2}\text{erf}\left(\frac{1}{\sqrt{2\rho}}\right)\right]\log\left[\frac{1+d}{2} + \frac{1-d}{2}\text{erf}\left(\frac{1}{\sqrt{2\rho}}\right)\right] \\ &-\left[\frac{1-d}{2} - \frac{1-d}{2}\text{erf}\left(\frac{1}{\sqrt{2\rho}}\right)\right]\log\left[\frac{1-d}{2} - \frac{1-d}{2}\text{erf}\left(\frac{1}{\sqrt{2\rho}}\right)\right] \end{aligned} \tag{2.10}$$

which is monotonically increasing with $\rho$. Therefore, with a slight abuse of language, in the following $\rho$ shall be referred to as *dataset entropy*.

The available information is allocated in the synaptic coupling among neurons (as in the standard Hebbian storing), as specified by the following supervised and unsupervised generalization of the multitasking Hebbian network:

[11] Strictly speaking, in this particular model, a minimal correlation still persists as $r_\mu \to 0$ in the sense that the zeros are always located at the same entries in the patterns. However, as proved in appendix A, by relaxing this assumption the same emerging picture is not altered, but it is mathematically more cumbersome.



J. Stat. Mech. (2023) 113401

**Definition 4.** Given $N$ binary neurons $\sigma_i = \pm 1$, with $i \in (1,\dots,N)$, the cost function (or *Hamiltonian*) of the multitasking Hebbian neural network in the supervised regime is

$$\mathcal{H}^{(sup)}_{N,K,d,M,r}(\boldsymbol{\sigma}|\boldsymbol{\eta}) = -\frac{1}{2N}\frac{1}{(1-d)(1+\rho)}\sum_{\mu=1}^{K}\sum_{i,j=1}^{N,N}\left(\frac{1}{Mr}\sum_{a=1}^{M}\eta_i^{\mu,a}\right)\left(\frac{1}{Mr}\sum_{b=1}^{M}\eta_j^{\mu,b}\right)\sigma_i\sigma_j. \tag{2.11}$$

**Definition 5.** Given $N$ binary neurons $\sigma_i = \pm 1$, with $i \in (1,\dots,N)$, the cost function (or *Hamiltonian*) of the multitasking Hebbian neural network in the unsupervised regime is

$$\mathcal{H}^{(\text{unsup})}_{N,K,d,M,r}(\boldsymbol{\sigma}|\boldsymbol{\eta}) = -\frac{1}{2N}\frac{1}{(1-d)(1+\rho)}\sum_{\mu=1}^{K}\sum_{i,j=1}^{N,N}\left(\frac{1}{Mr^2}\sum_{a=1}^{M}\eta_i^{\mu,a}\eta_j^{\mu,a}\right)\sigma_i\sigma_j. \tag{2.12}$$

**Remark 1.** The factor $(1-d)(1+\rho)$ appearing in (2.11) corresponds to $\mathbb{E}_{\xi},\mathbb{E}_{(\eta|\xi)}[\sum_a \eta_i^{\mu,a}/(Mr)]^2$ and it acts as a normalization factor. A similar factor is also inserted in (2.12).

**Remark 2.** By direct comparison between (2.11) and (2.12), the role of the 'teacher' in the supervised setting is evident: in the unsupervised scenario, the network has to handle all the available examples regardless of their archetype label, while in the supervised counterpart a teacher has previously grouped examples belonging to the same archetype together (whence the double sum on $a = (1,\dots,M)$ and on $b = (1,\dots,M)$ appearing in equation (2.11), that is missing in equation (2.12)).

We investigate the model within a canonical framework: we introduce the Boltzmann–Gibbs measure

$$\mathcal{P}^{(\text{sup,unsup})}_{N,K,\beta,d,M,r}(\boldsymbol{\sigma}|\boldsymbol{\eta}) := \frac{\exp\left[-\beta\mathcal{H}^{(\text{sup,unsup})}_{N,K,d,M,r}(\boldsymbol{\sigma}|\boldsymbol{\eta})\right]}{\mathcal{Z}^{(\text{sup,unsup})}_{N,K,\beta,d,M,r}(\boldsymbol{\eta})}, \tag{2.13}$$

where

$$\mathcal{Z}^{(\text{sup,unsup})}_{N,K,\beta,d,M,r}(\boldsymbol{\eta}) := \sum_{\boldsymbol{\sigma}}\exp\left[-\beta\mathcal{H}^{(\text{sup,unsup})}_{N,K,d,M,r}(\boldsymbol{\sigma}|\boldsymbol{\eta})\right] \tag{2.14}$$

is the normalization factor, also referred to as the partition function, and the parameter $\beta \in \mathbb{R}^+$ rules the broadness of the distribution in such a way that for $\beta \to 0$ (infinite noise limit) all the $2^N$ neural configurations are equally likely, while for $\beta \to \infty$ the distribution is delta-peaked at the configurations corresponding to the minima of the cost function.

The average performed over the Boltzmann–Gibbs measure is denoted as

$$\omega^{(\text{sup,unsup})}_{N,K,\beta,d,M,r}[\cdot] := \sum_{\boldsymbol{\sigma}}^{2^N} \cdot\, \mathcal{P}^{(\text{sup,unsup})}_{N,K,\beta,d,M,r}(\boldsymbol{\sigma}|\boldsymbol{\eta}). \tag{2.15}$$



J. Stat. Mech. (2023) 113401

Beyond this average, we shall also take the so-called *quenched* average; that is, the average over the realizations of archetypes and examples, namely over the distributions (2.1) and (2.9), and this is denoted as

$$\mathbb{E}\left[\cdot\right] = \mathbb{E}_{\xi}\mathbb{E}_{(\eta|\xi)}\left[\cdot\right]. \tag{2.16}$$

**Definition 6.** The quenched free energy of the network at finite network size $N$ reads as

$$-\beta\mathcal{F}^{(\mathrm{sup,unsup})}_{N,K,\beta,d,M,r} = \frac{1}{N}\mathbb{E}\log\mathcal{Z}^{(\mathrm{sup,unsup})}_{N,K,\beta,d,M,r}\left(\boldsymbol{\eta}\right). \tag{2.17}$$

In the thermodynamic limit we pose

$$\mathcal{F}^{(\mathrm{sup,unsup})}_{K,\beta,d,M,r} = \lim_{N\to\infty}\mathcal{F}^{(\mathrm{sup,unsup})}_{N,K,\beta,d,M,r}. \tag{2.18}$$

**Definition 7.** The network capabilities can be quantified by introducing the following order parameters, for $\mu = 1,\dots,K$,

$$\begin{aligned} m_{\mu} &:= \frac{1}{N}\sum_{i=1}^{N}\xi_i^{\mu}\sigma_i, \\ n_{\mu,a} &:= \frac{1}{(1+\rho)\,r}\frac{1}{N}\sum_{i=1}^{N}\eta_i^{\mu,a}\sigma_i, \\ n_{\mu} &:= \frac{1}{M}\sum_{a=1}^{M}n_{\mu,a} = \frac{1}{(1+\rho)\,r}\frac{1}{NM}\sum_{i,a=1}^{N,M}\eta_i^{\mu,a}\sigma_i, \end{aligned} \tag{2.19}$$

We stress that, beyond the fairly standard $K$ Mattis magnetizations $m_{\mu}$, which assess the alignment of the neural configuration $\boldsymbol{\sigma}$ with the archetype $\boldsymbol{\xi}^{\mu}$, we also need to introduce $K$ empirical Mattis magnetizations $n_{\mu}$, which compare the alignment of the neural configuration with the average of the examples labeled with $\mu$, as well as $K\times M$ single-example Mattis magnetizations $n_{\mu,a}$, which measure the proximity between the neural configuration and a specific example. An intuitive way to see the suitability of the $n_{\mu}$'s and $n_{\mu,a}$'s is by noticing that the cost functions $\mathcal{H}^{(\mathrm{sup})}$ and $\mathcal{H}^{(\mathrm{unsup})}$ can be written as a quadratic form in, respectively, $n_{\mu}$ and $n_{\mu,a}$; on the other hand, the $m_{\mu}$'s do not appear therein explicitly as the archetypes are unknowns.

Finally, notice that no spin-glass order parameters are needed here, since we are working in the low-storage regime [12, 19].

## 3. Parallel learning: the picture by statistical mechanics

### 3.1. Study of the cost function and its related free energy

To inspect the emergent capabilities of these networks, we need to estimate the order parameters introduced in equation (2.19) and analyze their behavior versus the control parameters $K,\beta,d,M,r$. For this task we need an explicit expression of the free energy in terms of these order parameters so as to extremize the former over the latter. In this



J. Stat. Mech. (2023) 113401

section we carry out this investigation in the thermodynamic limit and in the low-storage scenario by relying upon Guerra's interpolating techniques (see e.g. [3, 15, 16, 26]): the underlying idea is to introduce an interpolating free energy whose extrema are the original model (which is the target of our investigation but we are not able to address it directly) and a simple one (which is usually a one-body model that we can solve exactly). We then start by evaluating the solution of the latter and next we propagate the obtained solution back to the original model by the fundamental theorem of calculus, integrating on the interpolating variable. Usually, in this last passage, one assumes replica symmetry, namely that the order-parameter fluctuations are negligible in the thermodynamic limit as this makes the integral propagating the solution analytical. In the low-load scenario replica symmetry holds exactly, making the following calculation rigorous. In fact, as long as $K/N \to 0$ while $N \to \infty$, the order parameters self-average around their means [17, 34], which will be denoted by a bar; that is

$$\lim_{N\to\infty} \mathcal{P}_{N,K,\beta,d,M,r}(m_\mu) = \delta(m_\mu - \bar{m}_\mu), \quad \forall \mu \in (1,\dots,K), \tag{3.1}$$

$$\lim_{N\to\infty} \mathcal{P}_{N,K,\beta,d,M,r}(n_\mu) = \delta(n_\mu - \bar{n}_\mu), \quad \forall \mu \in (1,\dots,K), \tag{3.2}$$

where $\mathcal{P}_{N,K,\beta,d,M,r}$ denotes the Boltzmann–Gibbs probability distribution for the observables considered. We anticipate that the centers of these distributions are independent of the training (either supervised or unsupervised) underlying the Hebbian kernel.

Before proceeding, we slightly revise the partition functions (2.14) by inserting an extra term in their exponents because it allows us to apply the functional generator technique to evaluate the Mattis magnetizations. This implies the following modification, respectively, in the supervised and unsupervised settings, of the partition function:

**Definition 8.** Given the interpolating parameter $t \in [0,1]$, the auxiliary field $J$ and the constants $\{\psi_\mu\}_{\mu=1,\dots,K} \in \mathbb{R}$ to be set *a posteriori*, Guerra's interpolating partition function for the supervised and unsupervised multitasking Hebbian networks is given, respectively, by

$$\mathcal{Z}^{(\mathrm{sup})}_{N,K,\beta,d,M,r}(\boldsymbol{\eta}; J, t) = \sum_{\{\boldsymbol{\sigma}\}} \int \mathrm{d}\mu(z_\mu) \exp\left[ J \sum_{\mu,i} \xi_i^\mu \sigma_i + \frac{t\beta N(1+\rho)}{2(1-d)} \sum_\mu n_\mu^2(\boldsymbol{\sigma}) + (1-t)\frac{N}{2} \sum_\mu \psi_\mu n_\mu(\boldsymbol{\sigma}) \right]. \tag{3.3}$$

$$\mathcal{Z}^{(\mathrm{unsup})}_{N,K,\beta,d,M,r}(\boldsymbol{\eta}; J, t) = \sum_{\{\boldsymbol{\sigma}\}} \int \mathrm{d}\mu(z_\mu) \exp\left[ J \sum_{\mu,i} \xi_i^\mu \sigma_i + \frac{t\beta N(1+\rho)}{2(1-d)M} \sum_{\mu=1}^{K} \sum_{a=1}^{M} n_{\mu,a}^2(\boldsymbol{\sigma}) + (1-t) N \sum_{\mu,a} \psi_\mu n_{\mu,a}(\boldsymbol{\sigma}) \right]. \tag{3.4}$$

More precisely, we added the term $J \sum_\mu \sum_i \xi_i^\mu \sigma_i$ that allows us to 'generate' the expectation of the Mattis magnetization $m_\mu$ by evaluating the derivative w.r.t. $J$ of the quenched free energy at $J = 0$. This operation is not necessary for *Hebbian storing*, where



J. Stat. Mech. (2023) 113401

the Mattis magnetization is a natural order parameter (the Hopfield Hamiltonian can be written as a quadratic form in $m_\mu$, as standard in AGS theory [12]), while for *Hebbian learning* (whose cost function can be written as a quadratic form in $n_\mu$, not in $m_\mu$, as the network does not directly experience the archetypes) we need such a term as otherwise the expectation of the Mattis magnetization would not be accessible. This operation becomes redundant in the $M \to \infty$ limit, where $m_\mu$ and $n_\mu$ become proportional by a standard central limit theorem argument (see also section 3.1.1 and [11]). Clearly, $\mathcal{Z}^{(\text{sup,unsup})}_{N,K,\beta,d,M,r}(\boldsymbol{\eta}) = \lim_{J\to 0} \mathcal{Z}^{(\text{sup,unsup})}_{N,K,\beta,d,M,r}(\boldsymbol{\eta}; J)$ and these generalized interpolating partition functions, provided in equations (3.3) and (3.4), respectively, recover the original models when $t = 1$, while they return a simple one-body model at $t = 0$.

As for the $\psi_\mu$'s, their role is mimicking, as closely as possible, the true post-synaptic field perceived by the neurons.

The partition functions (3.3) and (3.4) can be used to define a generalized measure and a generalized Boltzmann–Gibbs average that we indicate by $\omega_t^{(sup,unsup)}[\cdot]$. Of course, when $t = 1$ the standard Boltzmann–Gibbs measure and related averages are recovered. Analogously, we can also introduce a generalized interpolating quenched free energy as:

**Definition 9.** The interpolating free energy for the multitasking Hebbian neural network is introduced as

$$-\beta \mathcal{F}^{(\text{sup,unsup})}_{N,K\beta,d,M,r}(J,t) := \frac{1}{N}\mathbb{E}\left[\ln \mathcal{Z}^{(\text{sup,unsup})}_{N,K,\beta,d,M,r}(\boldsymbol{\eta}; J, t)\right], \tag{3.5}$$

and, in the thermodynamic limit,

$$\mathcal{F}^{(\text{sup,unsup})}_{K,\beta,d,M,r}(J,t) := \lim_{N\to\infty} \mathcal{F}^{(\text{sup,unsup})}_{N,K,\beta,d,M,r}(J,t). \tag{3.6}$$

Obviously, by setting $t = 1$ in the interpolating free-energy, we recover the original ones, namely $\mathcal{F}^{(\text{sup,unsup})}_{K,\beta,d,M,r}(J) = \mathcal{F}^{(\text{sup,unsup})}_{K,\beta,d,M,r}(J, t = 1)$, which we finally evaluate at $J = 0$.

We are now ready to state the first theorem:

**Theorem 1.** *In the thermodynamic limit ($N \to \infty$) and in the low-storage regime ($K/N \to 0$), the quenched free energy of the multitasking Hebbian network—trained under supervised or unsupervised learning—reads as*

$$-\beta \mathcal{F}^{(\text{sup,unsup})}_{K,\beta,d,M,r}(J) = \mathbb{E}\left\{\ln\left[2\cosh\left(J\sum_{\mu=1}^{K}\xi^\mu + \frac{\beta}{1-d}\sum_{\mu=1}^{K}\bar{n}_\mu\hat{\eta}^\mu\right)\right]\right\} - \frac{\beta}{1-d}(1+\rho)\sum_{\mu=1}^{K}\bar{n}_\mu^2, \tag{3.7}$$

*where* $\mathbb{E} = \mathbb{E}_\xi\mathbb{E}_{(\eta|\xi)}$, $\hat{\eta}^\mu = \dfrac{1}{Mr}\sum_{a=1}^{M}\eta_i^{\mu,a}$, *and the values* $\bar{n}_\mu$ *must fulfill the following self-consistent equations:*

$$\bar{n}_\mu = \frac{1}{(1+\rho)}\mathbb{E}\left\{\tanh\left[\frac{\beta}{(1-d)}\sum_{\nu=1}^{K}\bar{n}_\nu\hat{\eta}^\nu\right]\hat{\eta}^\mu\right\}, \quad \forall\mu \in (1,\dots,K), \tag{3.8}$$

*as these values of the order parameters are extremal for the free energy* $\mathcal{F}^{(sup,unsup)}_{K,\beta,d,M,r}(J = 0)$.



J. Stat. Mech. (2023) 113401

**Corollary 1.** *By considering the auxiliary field $J$ coupled to $m_\mu$ and recalling that $\lim_{N\to\infty} m_\mu = \bar{m}_\mu$, we obtain a self-consistent equation also for the Mattis magnetization as $\bar{m}_\mu = -\beta\partial_J \mathcal{F}^{(\mathrm{sup,unsup})}_{K,\beta,d,M,r}(J)|_{J=0}$; thus, we have*

$$\bar{m}_\mu = \mathbb{E}\left\{\tanh\left[\frac{\beta}{(1-d)}\sum_{\nu=1}^{K}\bar{n}_\nu\hat{\eta}^\nu\right]\xi^\mu\right\}, \quad \forall\mu\in(1,\dots,K). \tag{3.9}$$

For the proof of proposition 1 and of corollary 1 we refer to appendix D.1.

We highlight that the expressions of the quenched free energy for a network trained with or without the supervision of a teacher do actually coincide. Intuitively, this happens because we are considering only a few archetypes (i.e. we work at low load); consequently, the minima of the cost function are well separated and there is only a negligible role of the teacher in shaping the landscape to avoid overlaps in their basins of attractions. Clearly, this is expected to no longer be true in the high-load setting and, indeed, it is proven not to hold for non-diluted patterns, where supervised and unsupervised protocols give rise to different outcomes [2, 11].

The self-consistent equation (3.9) has been solved numerically for several values of parameters and the results for $K=2$ and $K=3$ are shown respectively in figures 3 (where the values of the cost function are also reported) and 4. We also checked the validity of these results by comparing them with the outcomes of Monte Carlo simulations, finding an excellent asymptotic agreement; further, in the large $M$ limit, the magnetizations eventually converge to the values predicted by the theory developed in the storing framework; see equation (2.6). Therefore, in both the scenarios, the hierarchical or parallel organization of the magnetization's amplitudes are recovered: beyond the numerical evidence just mentioned, in appendix C an analytical proof is provided.

*3.1.1. Low-entropy datasets: the big-data limit.* As discussed in section 2.2, the parameter $\rho$ quantifies the amount of information needed to describe the original message $\boldsymbol{\xi}^\mu$ given the set of related examples $\{\boldsymbol{\eta}^{\mu,a}\}_{a=1,\dots,M}$. In this section we focus on the case $\rho\ll 1$ that corresponds to a highly informative dataset; we recall that in the limit $\rho\to 0$ we get a dataset where either the items ($r\to 1$) or their empirical average ($M\to\infty$, $r$ finite) coincide with the archetypes in such a way that the theory collapses to the standard low-load Hopfield reference.

As explained in appendix D.2, we start from the self-consistent equations (3.8) and (3.9) and we exploit the central limit theorem to write $\hat{\eta}^\mu \sim \xi^\mu\left(1+\lambda_\mu\sqrt{\rho}\right)$, where $\lambda_\mu\sim\mathcal{N}(0,1)$. In this way we reach the simpler expressions given by:

**Proposition 1.** *In the low-entropy dataset scenario, preserving the low storage and thermodynamic limit assumptions, the two sets of order parameters of the theory, $\bar{m}_\mu$ and $\bar{n}_\mu$, become related by the following equations:*

$$\bar{n}_\mu = \frac{\bar{m}_\mu}{(1+\rho)} + \beta'\frac{\rho\,\bar{n}_\mu}{(1+\rho)}\mathbb{E}_{\xi,Z}\left\{\left[1-\tanh^2\left(g\left(\beta,Z,\bar{\boldsymbol{n}}\right)\right)\right]\left(\xi^\mu\right)^2\right\}, \tag{3.10}$$

$$\bar{m}_\mu = \mathbb{E}_{\xi,Z}\left\{\tanh\left[g\left(\beta,\boldsymbol{\xi},Z,\bar{\boldsymbol{n}}\right)\right]\xi^\mu\right\}, \tag{3.11}$$



J. Stat. Mech. (2023) 113401

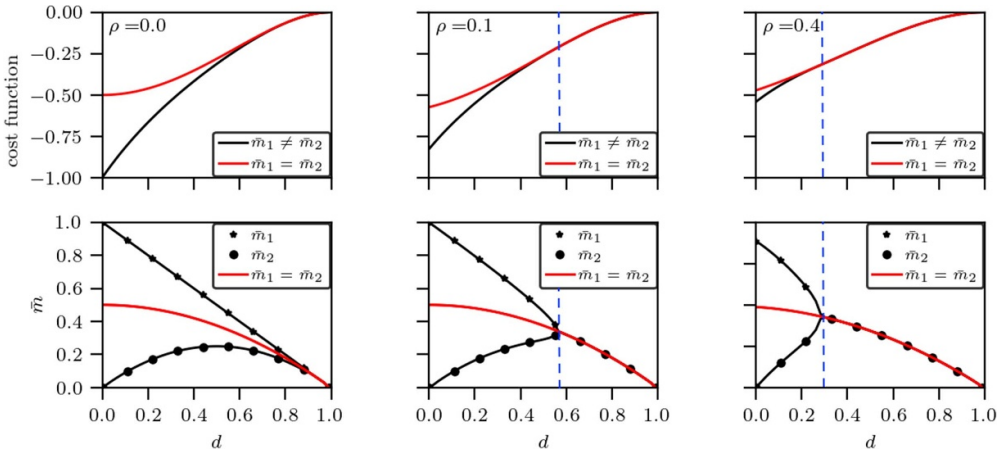


**Figure 3.** The dependence of the cost function (upper plots) and the magnetizations (lower plots) on the pattern dilution $d$ is shown, in the noiseless limit $\beta \to \infty$, for datasets generated by $K = 2$ archetypes and corresponding to different entropies $\rho$. Starting at $\rho = 0.0$ we see that the hierarchical regime (black lines) dominates at a relatively mild dilution value (i.e. the energy pertaining to this configuration is lower w.r.t. the parallel regime), while for $d \to 1$ the hierarchical ordering naturally collapses to the parallel regime (red lines), where all the magnetizations acquire the same values. Further note how, by increasing the entropy in the dataset (e.g. for $\rho = 0.1$ and $\rho = 0.4$), the domain of validity of the parallel regime is enlarged (much as increasing $\beta$ in the network, see figure 1). The vertical blue lines mark the transitions between these two regimes as captured by statistical mechanics: it corresponds to switching from the white to the green regions of the phase diagrams of figure 6.

*where*

$$g(\beta, \boldsymbol{\xi}, Z, \bar{\boldsymbol{n}}) = \beta' \sum_{\nu=1}^{K} \bar{n}_\nu \xi^\nu + \beta' Z \sqrt{\rho \sum_{\nu=1}^{K} \bar{n}_\nu^2 (\xi^\nu)^2} \tag{3.12}$$

*where* $Z \sim \mathcal{N}(0,1)$ *is a standard Gaussian variable and, to lighten the notation and assuming* $d \neq 1$ *with no loss of generality, we pose*

$$\beta' = \frac{\beta}{1-d}. \tag{3.13}$$

The regime $\rho \ll 1$, beyond being an interesting one (e.g. it can be seen as a *big data* $M \to \infty$ limit of the theory), offers a crucial advantage because of the above emerging proportionality relation between $\bar{n}$ and $\bar{m}$ (see equation (3.10)). In fact, the model is supplied only with examples—upon which the $n_\mu$'s are defined—while it is not aware of archetypes—upon which the $m_\mu$'s are defined—yet we can use this relation to recast the self-consistent equation for $\bar{n}$ into a self-consistent equation for $\bar{m}$ such that its numerical solution in the space of the control parameters allows us to get the phase diagram of such a neural network more straightforwardly.



J. Stat. Mech. (2023) 113401

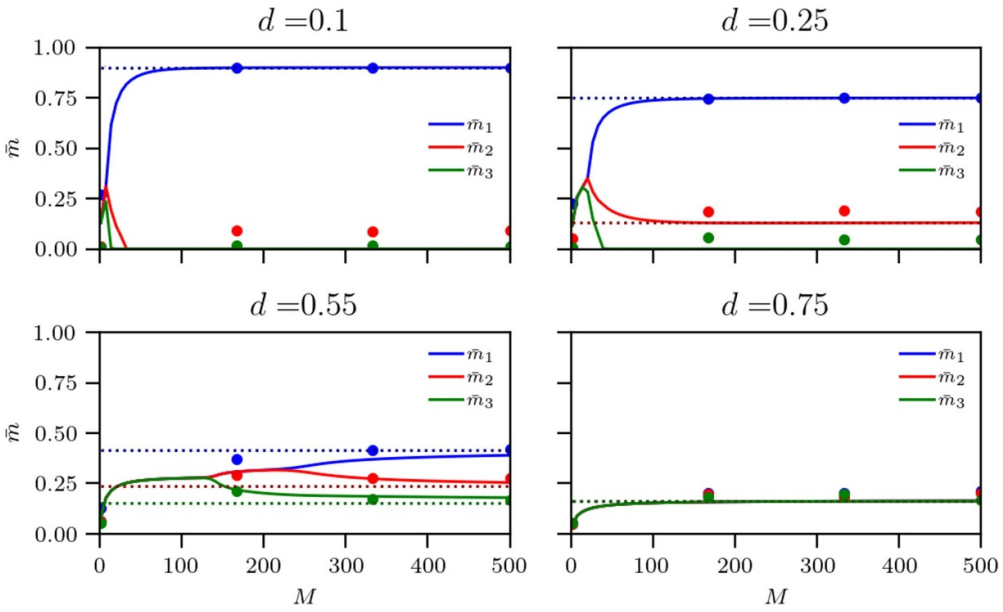


**Figure 4.** Behavior of the Mattis magnetizations as more and more examples are supplied to the network. Monte Carlo numerical checks at fixed temperature $\beta = 6.66$ (colored dots, $N = 6000$) for a diluted network with $r = 0.1$ and $K = 3$ are in plain agreement with the theory: solutions of the self-consistent equation for the Mattis magnetizations reported in corollary 1 are shown as solid lines. As the dilution increases, the network behavior departs from a Hopfield-like retrieval ($d = 0.1$) where just the blue magnetization is raised (serial pattern recognition) to the hierarchical regime ($d = 0.25$ and $d = 0.55$) where multiple patterns are simultaneously retrieved with different amplitudes, while for higher values of dilution the network naturally evolves toward the parallel regime ($d = 0.75$) where all the magnetizations are raised and with the same strength. Note also the asymptotic agreement with the dotted lines, whose values are those predicted by the multitasking Hebbian storage [7].

Retaining the condition $\rho \ll 1$, we now seek an estimate of the minimal number of examples (given the level of noise $r$, the number of archetypes to handle $K$, etc) that guarantee that the network can safely infer the archetype from the supplied dataset. To obtain these thresholds we have to deepen the ground-state structure of the network; that is, we now handle equations (3.10) and (3.11) to compute their zero fast-noise limit ($\beta \to \infty$). As detailed in appendix D.2 (see corollary 2), by taking the limit $\beta \to \infty$ in equations (3.10) and (3.11) we get

$$\bar{m}_\mu = \mathbb{E}_\xi \left\{ \mathrm{erf} \left[ \left( \sum_{\nu=1}^{K} \bar{m}_\nu \xi^\nu \right) \left( 2\rho \sum_{\nu=1}^{K} \bar{m}_\nu^2 (\xi^\nu)^2 \right)^{-1/2} \right] \xi^\mu \right\}. \tag{3.14}$$

Once we reach a relatively simple expression for $\bar{m}_\mu$, we can further manipulate it and try to get information about the existence of a lower-bound value for $M$, denoted by



J. Stat. Mech. (2023) 113401

$M_{\otimes}$, which ensures that the network has been supplied with sufficient information to learn and retrieve the archetypes.

Setting $\beta \to \infty$, we expect that the magnetizations fulfill the hierarchical organization, namely $(\bar{m}_1, \bar{m}_2, \ldots) = (1-d)(1, d, \ldots)$ and (3.14) becomes

$$\bar{m}_\mu \sim \frac{1-d}{2} \mathbb{E}_{\xi^{\nu \neq \mu}} \left\{ \operatorname{erf}\left[ \frac{d^\mu + \sum_{\nu \neq \mu}^{K} d^\nu \xi^\nu}{\sqrt{2\rho}\sqrt{d^{2\mu} + \sum_{\nu \neq \mu}^{K} d^{2\nu} (\xi^\nu)^2}} \right] + \operatorname{erf}\left[ \frac{d^\mu - \sum_{\nu \neq \mu}^{K} d^\nu \xi^\nu}{\sqrt{2\rho}\sqrt{d^{2\mu} + \sum_{\nu \neq \mu}^{K} d^{2\nu} (\xi^\nu)^2}} \right] \right\}, \tag{3.15}$$

where we highlight that the expectation is over all the archetypes but the $\mu$th one under inspection.

Next, we introduce a confidence interval, ruled by $\Theta$, and we require that

$$\bar{m}_\mu > (1-d)\, d^{\mu-1} \operatorname{erf}[\Theta]. \tag{3.16}$$

In order to quantify the critical number of examples $M_{\otimes}^{\mu}$ needed for a successful learning of the archetype $\mu$, we can exploit the relation

$$\mathbb{E}_{\xi^{\nu \neq \mu}} \{ \operatorname{erf}[f(\xi)] \} \geqslant \min_{\xi^{\nu \neq \mu}} \{ \operatorname{erf}[f(\xi)] \}, \tag{3.17}$$

where, in our case,

$$\begin{aligned} \min_{\xi^{\nu \neq \mu}} \{ \operatorname{erf}[f(\xi)] \} &= \min_{\xi^{\nu \neq \mu}} \left\{ \operatorname{erf}\left[ \frac{d^\mu + \sum_{\nu \neq \mu}^{K} d^\nu \xi^\nu}{\sqrt{2\rho}\sqrt{d^{2\mu} + \sum_{\nu \neq \mu}^{K} d^{2\nu} (\xi^\nu)^2}} \right] + \operatorname{erf}\left[ \frac{d^\mu - \sum_{\nu \neq \mu}^{K} d^\nu \xi^\nu}{\sqrt{2\rho}\sqrt{d^{2\mu} + \sum_{\nu \neq \mu}^{K} d^{2\nu} (\xi^\nu)^2}} \right] \right\} \\ &= 2 \operatorname{erf}\left[ \left( d^\mu - \sum_{\nu \neq \mu}^{K} d^\nu \right) \left( 2\rho \sum_{\nu=1}^{K} d^{2\nu} \right)^{-1/2} \right]. \end{aligned} \tag{3.18}$$

Thus, using the previous relation in (3.16), the following inequality must hold:

$$\operatorname{erf}\left[ \left( d^\mu - \sum_{\nu \neq \mu}^{K} d^\nu \right) \left( 2\rho \sum_{\nu=1}^{K} d^{2\nu} \right)^{-1/2} \right] = \operatorname{erf}\left[ \sqrt{\frac{1+d}{2\rho(1-d)}} \frac{2d^{\mu-1} - 1 - 2d^\mu + d^K}{\sqrt{1-d^{2K}}} \right] > d^{\mu-1} \operatorname{erf}[\Theta] \tag{3.19}$$

and we can write:

**Proposition 2.** *In the noiseless limit $\beta \to \infty$ and for relatively small dilutions $d < d_c(K)$, the critical threshold (in the number of required examples) for learning $M_{\otimes}$ depends on the dataset noise $r$, the dilution $d$, the number of archetypes to handle $K$, and the chosen confidence interval $\Theta$, and reads as*

$$M_{\otimes}^{\mu}(\Theta, r, d, K) > 2 \left( \operatorname{erf}^{-1}\left[ d^{\mu-1} \operatorname{erf}[\Theta] \right] \right)^2 \left( \frac{1-r^2}{r^2} \right) \frac{(1-d)\left(1-d^{2K}\right)}{(1+d)\left(2d^{\mu-1} - 1 - 2d^\mu + d^K\right)^2}. \tag{3.20}$$



J. Stat. Mech. (2023) 113401

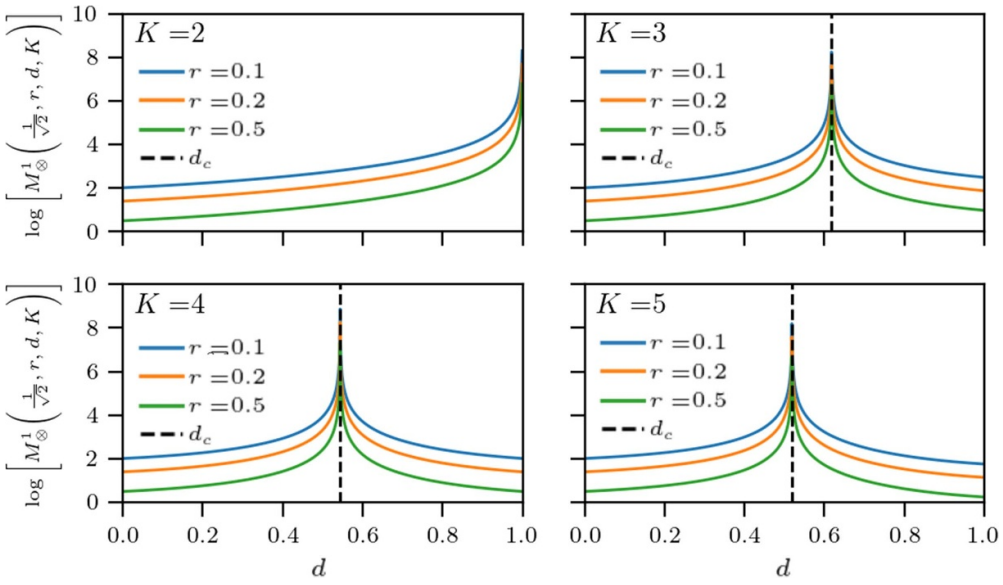


**Figure 5.** We plot the logarithm of the critical number of examples (required to raise the first magnetization) $M^1_\otimes$, evaluated for $\Theta = 1/\sqrt{2}$ and for different loads $K = 2,3,4,5$, as a function of the dilution of the networks, for different noise values of the dataset (as shown in the legend). Note the divergent behavior of $M^1_\otimes$ when approaching the critical dilution level $d_c(K)$, as predicted by the parallel Hebbian storage limit [7, 9]. This is the crossover between the two multitasking regimes, hierarchical vs parallel; hence, at $d_c$ there is no sharp behavior to infer and, consistently, the network cannot accomplish learning. This is confirmed by (3.20) where the critical amount of examples to correctly infer the archetype is reported: its denominator, for $\mu = 1$, reduces to $1 - 2d + d^K$ and it vanishes when $d \to d_c$.

*The choice* $\Theta = 1/\sqrt{2}$ *corresponds to the fairly standard condition* $\mathbb{E}_\xi \mathbb{E}_{(\eta|\xi)}[\xi^1_i h_i(\boldsymbol{\xi}^1)] > \sqrt{\mathrm{Var}[\xi^1_i h_i(\boldsymbol{\xi}^1)]}$ *when* $\mu = 1$, *where* $h_i(\boldsymbol{\xi}^1)$ *is the local field acting on a neuron* $i$ *when* $\boldsymbol{\sigma} = \boldsymbol{\xi}^1$.

To quantify these thresholds for learning, in figure 5 we report the required number of examples to learn the first archetype (out of $K = 2,3,4,50$, as shown in the various panels) as a function of the dilution of the network.

*3.1.2. Ergodicity breaking: the critical phase transition.* The main interest in the statistical mechanics approach to neural networks lies in inspecting their emerging capabilities that typically appear once ergodicity breaks down. As a consequence, finding the boundaries of validity of ergodicity is a classical starting point to deepen these aspects.

For this task, hereafter, we provide a systematic fluctuation analysis of the order parameters: the underlying idea is to check when, starting from the high-noise limit ($\beta \to 0$, where everything is uncorrelated and simple probabilistic arguments apply



J. Stat. Mech. (2023) 113401

straightforwardly), these fluctuations diverge, as that defines the onset of ergodicity breaking as stated in:

**Proposition 3.** *The ergodic region, in the space of the control parameters* $(\beta, d, \rho)$ *is confined to the half-plane defined by the critical line*

$$\beta_c = \frac{1}{1-d}, \tag{3.21}$$

*whatever the entropy of the dataset* $\rho$.

**Proof.** The proof is based on Guerra interpolation, but this time the target observable is rescaled fluctuations rather than the free energy. The rescaled fluctuations $\tilde{n}_\nu^2$ of the magnetizations are defined as

$$\tilde{n}_\nu = \sqrt{N}\,(n_\nu - \bar{n}_\nu)\,. \tag{3.22}$$

We recall that the interpolating framework we are using, for $t \in (0,1)$, is defined via

$$Z(t) = \sum_{\{\boldsymbol{\sigma}\}} \exp\left[\frac{\beta}{2} tN(1+\rho)\sum_{\mu=1}^{K} n_\mu^2 + N(1-t)\,\beta\,(1+\rho)\,N_\mu n_\mu\right], \tag{3.23}$$

and it is a trivial exercise to show that, for any smooth function $F(\sigma)$, the following relation holds:

$$\frac{\mathrm{d}\langle F\rangle}{\mathrm{d}t} = \frac{\beta}{2}(1+\rho)\left(\langle F\sum_\nu \tilde{n}_\nu^2\rangle - \langle F\rangle\langle\sum_\nu \tilde{n}_\nu^2\rangle\right), \tag{3.24}$$

such that, by choosing $F = \tilde{n}_\mu^2$, we can write

$$\begin{aligned}\frac{\mathrm{d}\langle \tilde{n}_\mu^2\rangle}{\mathrm{d}t} &= \frac{\beta}{2}(1+\rho)\left(\langle \tilde{n}_\mu^2\sum_\nu \tilde{n}_\nu^2\rangle - \langle \tilde{n}_\mu^2\rangle\langle\sum_\nu \tilde{n}_\nu^2\rangle\right)\\ &= \frac{\beta}{2}(1+\rho)\left(\langle \tilde{n}_\mu^4\rangle + \langle \bar{n}_\mu^2\sum_{\nu\neq\mu} \tilde{n}_\nu^2\rangle - \langle \tilde{n}_\mu^2\rangle^2 - \langle \tilde{n}_\mu^2\rangle\langle\sum_{\nu\neq\mu} \tilde{n}_\nu^2\rangle\right)\\ &= \beta(1+\rho)\,\langle \tilde{n}_\mu^2\rangle^2\end{aligned} \tag{3.25}$$

Thus, we have

$$\langle \tilde{n}_\mu^2\rangle_t = \frac{\langle \tilde{n}_\mu^2\rangle_{t=0}}{1 - t\beta(1+\rho)\,\langle \tilde{n}_\mu^2\rangle_{t=0}} \tag{3.26}$$



J. Stat. Mech. (2023) 113401

where the Cauchy condition $\langle \tilde{n}_\mu^2 \rangle_{t=0}$ reads

$$\langle \tilde{n}_\mu^2 \rangle_{t=0} = N \mathbb{E}_\xi \mathbb{E}_{(\eta|\xi)} \frac{\sum_{\{\boldsymbol{\sigma}\}} \left( \frac{1}{N^2 (1+\rho)^2} \sum_{i,j} \tilde{\eta}_i^\mu \tilde{\eta}_j^\mu \sigma_i \sigma_j + \tilde{n}_\mu^2 - 2 \frac{1}{N(1+\rho)} \sum_i \hat{\eta}_i^\mu \sigma_i \tilde{n}_\mu \right) \exp\left[ \beta \sum_\nu \tilde{n}_\nu \sum_i \tilde{\eta}_i^\nu \sigma_i \right]}{\sum_{\{\boldsymbol{\sigma}\}} \exp\left[ \beta \sum_\nu \tilde{n}_\nu \sum_i \tilde{\eta}_i^\nu \sigma_i \right]}$$

$$= \frac{1-d}{(1+\rho)} - N_\mu^2. \tag{3.27}$$

Evaluating $\langle \tilde{n}_\mu^2 \rangle_t$ for $t = 1$, we finally get

$$\langle \tilde{n}_\mu^2 \rangle_{t=1} = \frac{1 - d - (1+\rho) N_\mu^2}{\left[ 1 - \beta \left( 1 - d - (1+\rho) N_\mu^2 \right) \right]} \tag{3.28}$$

namely the rescaled fluctuations are described by a meromorphic function whose pole is

$$\beta = \frac{1}{(1 - d - (1+\rho) N_\mu^2)} \xrightarrow{N_\mu = 0} \beta_{\mathrm{C}} = \frac{1}{1-d}. \tag{3.29}$$

□

### 3.2. Stability analysis via standard Hessian: the phase diagram

The set of solutions to the self-consistent equations for the order parameters (3.10) describes as candidate solutions a number of different states whose stability must be investigated to understand which solution is preferred as the control parameters are made to vary: this procedure results in picturing the phase diagrams of the network. In order to evaluate the stability of these solutions, we need to check the sign of the second derivatives of the free energy. More precisely, we need to build up the Hessian, a matrix $\boldsymbol{A}$ whose elements are

$$\frac{\partial^2 \mathcal{F}(\bar{\boldsymbol{n}})}{\partial n^\mu \partial n^\nu} = A^{\mu\nu}. \tag{3.30}$$

Then, we evaluate and diagonalize $\boldsymbol{A}$ at a point $\tilde{\boldsymbol{n}}$, representing a particular solution of the self-consistency equation (3.10): the numerical results are reported in the phase diagrams provided in figure 6. Specifically, we find

$$A^{\mu\nu} = (1+\rho) \left[ [1 - \beta(1-d)] + \rho \beta \mathbb{E} \left\{ \mathcal{T}_{K\beta,\rho}^2 (\bar{\boldsymbol{n}}, z) (\xi^\mu)^2 \right\} \right] \delta^{\mu\nu} + Q^{\mu\nu} \tag{3.31}$$

where we set $\mathcal{T}_{K\beta,\rho}(\bar{\boldsymbol{n}}, z) = \tanh\left( \beta \sum_{\lambda=1}^K \bar{n}_\lambda \xi^\lambda + z\beta \sqrt{\rho \sum_{\lambda=1}^K (\bar{n}_\lambda \xi^\lambda)^2} \right)$ and



J. Stat. Mech. (2023) 113401

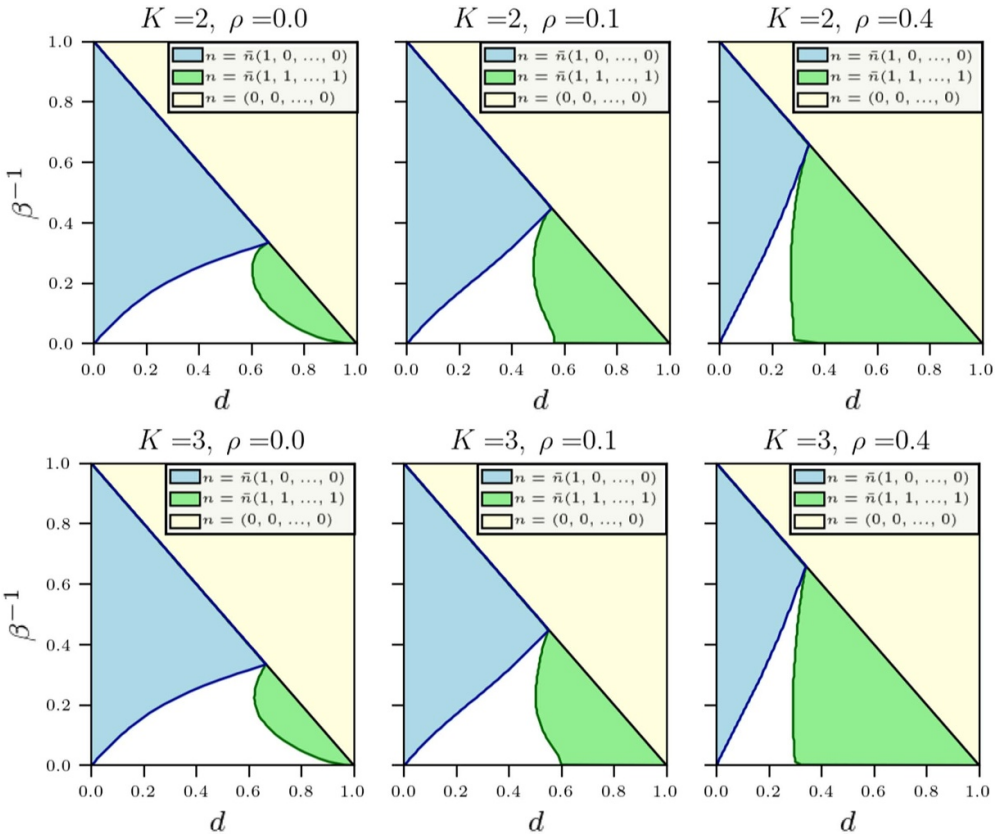


**Figure 6.** Phase diagram in the dilution-noise $(d, \beta^{-1})$ plane for different values of $K$ and $\rho$. We highlight that different regions, marked with different colors, represent different operational behaviors of the network: in yellow the ergodic solution, in light-blue the pure state solution (that is, solely one magnetization different from zero), in white the hierarchical regime (that is, several magnetizations differ from zero and they all assume different values) and in light-green the parallel regime (several magnetizations differ from zero but the amplitude is the same for all).

$$
\begin{aligned}
Q^{\mu\nu} = {} & \beta \mathbb{E}\left\{\left[\mathcal{T}^2_{K\beta,\rho}(\bar{\boldsymbol{n}}, z)\right]\xi^\mu\xi^\nu\right\}(1-\delta^{\mu\nu}) \\
& + 2\rho\beta^2 \mathbb{E}\left\{\left[\mathcal{T}_{K\beta,\rho}(\bar{\boldsymbol{n}}, z)\right]\left[1-\mathcal{T}^2_{K\beta,\rho}(\bar{\boldsymbol{n}}, z)\right]\left[\bar{n}_\nu\xi^\nu + \bar{n}_\mu\xi^\mu\right]\xi^\mu\xi^\nu\right\} \\
& + 2\rho^2\beta^3\bar{n}_\mu\bar{n}_\nu \mathbb{E}\left\{\left[1-3\mathcal{T}^2_{K\beta,\rho}(\bar{\boldsymbol{n}}, z)\right]\left[1-\mathcal{T}^2_{K\beta,\rho}(\bar{\boldsymbol{n}}, z)\right](\xi^\mu\xi^\nu)^2\right\}
\end{aligned}
\tag{3.32}
$$

In order to lighten the notation we will use $\mathcal{T}_{K,\beta,\rho}(\bar{\boldsymbol{n}}, z) = \mathcal{T}_K$ and we will omit the subscript whenever $\mathcal{T}_K$ occurs to be independent of $K$. We can now inspect the domain of stability of each possible solution of the self-consistency equation by plugging the structure of the candidate solution in (3.31).

*3.2.1. Ergodic state:* $\bar{\boldsymbol{n}} = \bar{n}_{d,\rho,\beta}(0, \ldots, 0)$. In this case the solution has the form $\bar{\boldsymbol{n}} = \bar{\boldsymbol{m}} = \boldsymbol{0}$; consequently, the Hessian matrix is diagonal and it reads as

$$
A^{\mu\nu} = \delta^{\mu\nu}(1+\rho)\left[1-\beta(1-d)\right]. \tag{3.33}
$$



J. Stat. Mech. (2023) 113401

As its eigenvalues are all equal to $(1+\rho)[1-\beta(1-d)]$, if we require the matrix to be positively definite, we must have

$$d > \frac{\beta-1}{\beta}. \tag{3.34}$$

Therefore, as $d > 1-\beta^{-1}$, the ergodic solution is stable; this scenario is reported in the phase diagrams provided in figure 6 as the yellow region. We stress that this result in the ergodic region is in plain agreement with the inspection of ergodicity breaking provided in proposition 3.

*3.2.2. Pure state:* $\bar{\boldsymbol{n}} = \bar{n}_{d,\rho,\beta}(1,0,\dots,0)$. In this case the structure of the solution has the form $\bar{m}_\mu = \bar{n}_\mu = 0$ for $\mu > 1$; thus, the only self-consistency equation different from zero is

$$\bar{n} = \frac{\mathbb{E}_{\xi,Z}\left[\mathcal{T}\,\xi^\mu\right]}{(1+\rho)} + \beta\frac{\rho\,\bar{n}}{(1+\rho)}\mathbb{E}_{\xi,Z}\left[\left(1-\mathcal{T}^2\right)\left(\xi^\mu\right)^2\right], \tag{3.35}$$

where $\mathcal{T} = \tanh\left[\beta\bar{n}\xi^\mu(1+z\sqrt{\rho})\right]$. It is easy to check that $\boldsymbol{A}$ becomes diagonal, with

$$A^{\mu\nu} = \begin{cases} (1+\rho) - \beta\,(1-d)\,(1+\rho)\,\mathbb{E}\left[1-\mathcal{T}^2\right] + 4\beta^2\rho\bar{n}\,(1-d)\,\mathbb{E}\left[\mathcal{T}\left(1-\mathcal{T}^2\right)\right] \\ \quad +2\beta^3\rho^2\bar{n}^2\,(1-d)\,\mathbb{E}\left[\left(1-3\mathcal{T}^2\right)\left(1-\mathcal{T}^2\right)\right] \quad \text{if } \mu=\nu \\ (1+\rho) - \beta\,(1-d)\,(1+\rho)\,\mathbb{E}\left[1-(1-d)\,\mathcal{T}^2\right] \qquad \text{if } \mu\neq\nu \end{cases}. \tag{3.36}$$

Notice that these eigenvalues do not depend on $K$ since $\mathcal{T}$ does not depend on $K$. Requiring positivity for all the eigenvalues, we get the region in the plane $(d,\beta^{-1})$, where the pure state is stable: this corresponds to the blue region in the phase diagrams reported in figure 6.

We stress that these pure state solutions, namely the standard Hopfield-type ones, in the ground state ($\beta^{-1}\to 0$) are never stable whenever $d\neq 0$ as the multi-tasking setting prevails. Solely at positive values of $\beta$, this single-pattern retrieval state is possible as the role of the noise is to destabilize the weakest magnetization of the hierarchical displacement, *vide infra*).

*3.2.3. Parallel state:* $\bar{\boldsymbol{n}} = \bar{n}_{d,\rho,\beta}(1,\dots,1)$. In this case, the structure of the solution has the form of a symmetric mixture state corresponding to the unique self-consistency equation for all $\mu = 1,\dots K$; namely,

$$\bar{n} = \frac{\mathbb{E}_{\xi,Z}\left\{\tanh\left[g\left(\beta,\boldsymbol{\xi},Z,\bar{n}\right)\right]\xi^\mu\right\}}{(1+\rho)} + \beta\frac{\rho\,\bar{n}}{(1+\rho)}\mathbb{E}_{\xi,Z}\left\{\left[1-\tanh^2\left(g\left(\beta,\boldsymbol{\xi}Z,\bar{n}\right)\right)\right]\left(\xi^\mu\right)^2\right\}, \tag{3.37}$$

where

$$g\left(\beta,\boldsymbol{\xi},Z,\bar{n}\right) = \beta\bar{n}\left[\sum_{\lambda=1}^{K}\xi^\lambda + \beta\,Z\sqrt{\rho\sum_{\lambda=1}^{K}\left(\xi^\lambda\right)^2}\right]. \tag{3.38}$$



J. Stat. Mech. (2023) 113401

Then, the matrix $\boldsymbol{A}$ displays the following structure:

$$\boldsymbol{A} = \begin{pmatrix} a & b & \cdots & b & b \\ b & a & \cdots & b & b \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ b & b & \cdots & a & b \\ b & b & \cdots & b & a \end{pmatrix} \tag{3.39}$$

with diagonal terms

$$\begin{aligned} a = (1+\rho) - \beta(1-d)(1+\rho) + \beta(1+\rho)\,\mathbb{E}\left[\mathcal{T}^2(\xi^\mu)^2\right] \\ + 4\beta^2\rho\bar{n}\mathbb{E}\left[\mathcal{T}\left(1-\mathcal{T}^2\right)\xi^\mu\right] + 2\beta^3\rho^2\bar{n}^2\mathbb{E}\left[\left(1-3\mathcal{T}^2\right)\left(1-\mathcal{T}^2\right)(\xi^\mu)^2\right], \end{aligned} \tag{3.40}$$

and off-diagonal terms

$$b = \beta\mathbb{E}\left[\mathcal{T}^2(\xi^\mu\xi^\nu)\right] + 2\rho^2\beta^3\bar{n}^2\mathbb{E}\left[\left(1-3\mathcal{T}^2\right)\left(1-\mathcal{T}^2\right)(\xi^\mu\xi^\nu)^2\right]. \tag{3.41}$$

This matrix always has only two kinds of eigenvalues, namely $a-b$ and $a+(K-1)b$; thus, for the stability of the parallel state, after computing (3.40) and (3.41), we have only to check for which point in the $(d,\beta^{-1})$ plane both $a-b$ and $a+(K-1)b$ are positive. In the phase diagrams of figure 6, the region where the parallel regime is stable is depicted in green.

*3.2.4. Hierarchical state:* $\bar{\boldsymbol{n}} = \bar{n}_{d,\rho,\beta}((1-d), d(1-d), d^2(1-d), \ldots)$. In this case the solution is of the form $\bar{\boldsymbol{n}} = \bar{n}_{d,\rho,\beta}((1-d), d(1-d), d^2(1-d), \ldots)$ and the region left untreated so far in the phase diagram, namely the white region in the plots of figure 6, is the room left to such hierarchical regime.

### 3.3. From the cost function to the loss function

We finally comment on the fact that, in the present approach, the quantifiers related to the assessment of pattern recognition of neural networks, i.e. the Mattis magnetization, are good quantifiers of the learning process too. The standard cost functions used in statistical mechanics of neural networks (e.g. the Hamiltonians) can be related one-to-one to standard loss functions used in machine learning (i.e. the squared-sum error functions); namely, after introducing the two loss functions $L_\mu^+ := (1/2N)||\xi^\mu - \sigma||^2 = 1 - m_\mu$ and $L_\mu^- = (1/2N)||\xi^\mu + \sigma||^2 = 1 + m_\mu$[12], it can immediately be shown that

$$\mathcal{H}_N(\boldsymbol{\sigma}|\boldsymbol{\xi}) = \frac{-1}{2N}\sum_{i,j}^{N,N}\sum_{\mu}^{K}\xi_i^\mu\xi_j^\mu\sigma_i\sigma_j \equiv -N\sum_{\mu}^{K}\left(1 - L_\mu^+ \cdot L_\mu^-\right).$$

Thus, minimizing the former implies minimizing the latter in such a way that, if we extremize w.r.t. the neurons we are performing machine retrieval (i.e. pattern recognition), while if we extremize w.r.t. the weights we are performing machine learning.

[12] Note that in the last passage we naturally highlighted the presence of the Mattis magnetization in these loss functions.



J. Stat. Mech. (2023) 113401

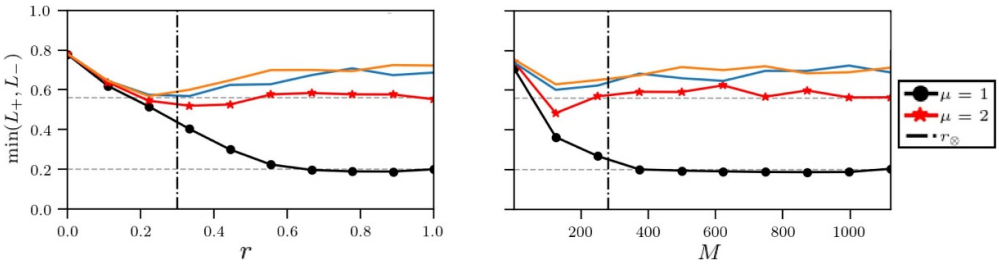


**Figure 7.** Left: parallel minimization of several (mean squared error) loss functions $L_\pm = ||\xi^\mu \pm \sigma||^2$ (each pertaining to a different archetype) as the noise in the dataset $r$ is varied. Here $M = 25$, $N = 10\,000$. The horizontal gray dashed lines are the saturation level of the loss functions, namely $1 - \frac{d}{2} - (1-d)d^{\mu-1}$. We get $r_\otimes$ (the vertical black line) by the inversion of (3.20). Right: parallel minimization of several (mean squared error) loss functions $L_\pm = ||\xi^\mu \pm \sigma||^2$ (each pertaining to a different archetype) as the dataset size $M$ is varied. As $M$ grows, the simultaneous minimization of more than one loss function takes place, unlike learning via standard Hebbian mechanisms where one loss function—dedicated to a single archetype—is minimized at a time. Orange and blue lines pertain to loss functions of other patterns that, at these levels of dilution and noise, cannot be minimized along with the previous ones.

Indeed, at least in this setting, learning and retrieval are two faces of the same coin (clearly the task here, from a machine learning perspective, is rather simple as the network is only asked to correctly classify the examples and possibly generalize).

In figure 7 we inspect what happens to these loss functions, pertaining to the various archetypes, as the cost function is minimized. We see that, in contrast to the standard Hopfield model (where only one loss function at a time diminishes its value), in this parallel learning setting, several loss functions (related to different archetypes) are simultaneously lowered, as expected with a parallel learning machine.

## 4. Conclusions

Since AGS seminal works in the mid '80s [13, 14], attractor neural networks have experienced unprecedented growth, and the bulk of techniques developed for spin glasses in the last four decades (e.g. replica trick, cavity method, message passage, interpolation) now act as a prosperous cornucopia for explaining the emergent information processing capabilities that these networks show as their control parameters are tuned. However, while the original AGS theory remains a solid pillar and a paradigmatic reference in the field, several extensions are required to keep it up to date and to face modern challenges.

The first generalization we need is to move from a setting where the machine stores already defined patterns (as in the standard Hopfield model) toward a more realistic learning procedure where these patterns are unknown and have to be inferred from

J. Stat. Mech. (2023) 113401

examples: the *Hebbian storage* rule of AGS theory quite naturally generalizes toward supervised and an unsupervised *Hebbian learning* prescriptions [2, 11]. This enlarges the space of the control parameters from $\alpha, \beta$ (or $K$, $N$, $\beta$) of the standard Hopfield model toward $\alpha, \beta, \rho$ (or $K$, $N$, $\beta$, $M$, $r$) as we now deal also with a dataset where we have $M$ examples of mean quality $r$ for each pattern (archetype) or, equivalently, we speak of a dataset produced at a given entropy $\rho$.

Once this is accomplished, the second strong limitation of the original AGS theory that must be relaxed is that patterns share the same length and, in particular, this equals the size of the network (namely in the standard Hopfield model there are $N$ neurons to handle patterns, whose length is exactly $N$ for all of them): a more general scenario encompasses patterns that contain different amounts of information; that is, patterns are diluted. The information-processing capabilities of Hebbian networks at work with diluted patterns have been extensively investigated in the last decade [4, 5, 7, 8, 20, 28, 31, 33, 35] and, in particular, they are shown to be able to retrieve several patterns in parallel (a key property of neural networks that is not captured by standard AGS theory).

In this paper, we combine these features and address the parallel learning of diluted patterns, focusing on the low-storage regime; that is, when the number of patterns scales at most logarithmically with the size of the network. Note that this further enlarges the space of the control parameters by introducing the dilution $d$: we have several control parameters because the network information processing capabilities are enriched w.r.t. the bare Hopfield reference[13].

Here we proved that, if we supply the network with a diluted dataset, containing on average a fraction $d$ of blank entries, the network spontaneously undergoes parallel learning and behaves as a multitasking associative memory able to learn, store and retrieve multiple patterns in parallel. In fact, the Hamiltonian of the model (acting as a cost function for the neuronal dynamics) can be recast as a mean squared error (acting as a loss function for synapsis dynamics), in such a way that—when experiencing a diluted dataset—the network can simultaneously lower the loss functions related to different patterns.

For mild values of dilution, the most favored displacement of the Mattis magnetizations is a *hierarchical ordering*; namely, the intensities of these signals scale as power laws w.r.t. their information content $m_K \sim d^K \cdot (1-d)$, while at high values of dilution a *parallel ordering*, where all these amplitudes collapse to the same value, prevails: the phase diagrams of these networks properly capture these different working regions.

Remarkably, in the low-storage regime (where glassy phenomena can be neglected), the presence (or the lack) of a teacher does not alter the above scenario and the threshold for a secure learning, namely the minimal required amount of examples (given the

[13] However, we clarify how it could be inappropriate to speak about structural differences among the standard Hopfield model and the present multitasking counterpart; ultimately, these huge behavioral differences are just consequences of the different nature of the datasets provided to the same Hebbian network during training.



J. Stat. Mech. (2023) 113401

constraints; that is, the noise in the dataset $r$, the amount of different archetypes $K$ to cope with, etc) $M_{\otimes}$ that guarantees that the network is able to infer the archetype and thus generalize, is the same for supervised and unsupervised protocols and we estimated its value.

## Acknowledgments

This research was supported by Ministero degli Affari Esteri e della Cooperazione Internazionale (MAECI) via the BULBUL grant (Italy-Israel), CUP Project No. F85F21006230001 and by PNRR MUR project no. PE0000013-FAIR, and received financial support from the Simons Foundation (Grant No. 454949, G Parisi) and ICSC—Italian Research Center on High Performance Computing, Big Data and Quantum Computing, funded by the European Union—NextGenerationEU.

Further, this work was partly supported by the Alan Turing Institute through the Theory and Methods Challenge Fortnights event *Physics-informed machine learning*, which took place on 16–27 January 2023 at the Alan Turing Institute headquarters.

E A acknowledges financial support from Sapienza University of Rome (RM120172B8066CB0).

E A, A A and A B acknowledge GNFM-INdAM (Gruppo Nazionale per la Fisica Mamematica, Istituto Nazionale d'Alta Matematica), A A further acknowledges UniSalento for financial support via PhD-AI and A B further acknowledges the PRIN-2022 Project *Statistical Mechanics of Learning Machines: from algorithmic and information-theoretical limits to new biologically inspired paradigms*.

## Appendix A. A more general sampling scenario

The way in which we add noise over the archetypes to generate the dataset in the main text (see equation (2.9)) is a rather peculiar one as, in each example, it preserves the number but also the positions of lacunæ already present in the related archetype. This implies that the noise cannot affect the amplitudes of the original signal, i.e. $\sum_i (\eta_i^{\mu,a})^2 = \sum_i (\xi_i^{\mu})^2$ holds for any $a$ and $\mu$, while we do expect that with more general kinds of noise this property is not preserved sharply.

Here we consider the case where the number of blank entries present in $\boldsymbol{\xi}^{\mu}$ is preserved on average in the related sample $\{\eta^{\mu,a}\}_{a=1,\dots,M}$ but lacunæ can move along the examples: this more realistic kind of noise gives rise to cumbersome calculations (still analytically treatable) but should not strongly affect the learning, storing and retrieving capabilities of these networks (as we now prove).

Specifically, here we define a new kind of example $\tilde{\eta}_i^{\mu,a}$ (that we can identify from the previous ones $\eta_i^{\mu,a}$ by labeling them with a tilde) in the following way:



J. Stat. Mech. (2023) 113401

**Definition 10.** Given $K$ random patterns $\boldsymbol{\xi}^\mu$ $(\mu = 1,\dots,K)$, each of length $N$, whose entries are i.i.d. from

$$\mathbb{P}(\xi_i^\mu) = \frac{(1-d)}{2}\delta_{\xi_i^\mu,-1} + \frac{(1-d)}{2}\delta_{\xi_i^\mu,+1} + d\delta_{\xi_i^\mu,0}, \tag{A.1}$$

we use these archetypes to generate $M \times K$ different examples $\{\tilde{\eta}_i^{\mu,a}\}^{a=1,\dots,M}$ whose entries are depicted following

$$\begin{aligned}\mathbb{P}(\tilde{\eta}_i^{\mu,a}|\xi_i^\mu = \pm 1) &= A_\pm(r,s)\,\delta_{\tilde{\eta}_i^{\mu,a},\xi_i^\mu} + B_\pm(r,s)\,\delta_{\tilde{\eta}_i^{\mu,a},-\xi_i^\mu} + C_\pm(r,s)\,\delta_{\tilde{\eta}_i^{\mu,a},0}\\ \mathbb{P}(\tilde{\eta}_i^{\mu,a}|\xi_i^\mu = 0) &= A_0(r,s)\,\delta_{\tilde{\eta}_i^{\mu,a},\xi_i^\mu} + B_0(r,s)\,\delta_{\tilde{\eta}_i^{\mu,a},+1} + C_0(r,s)\,\delta_{\tilde{\eta}_i^{\mu,a},-1}\end{aligned} \tag{A.2}$$

for $i = 1,\dots,N$ and $\mu = 1,\dots,K$, where we pose

$$\begin{aligned} A_\pm(r,s) &= \frac{1+r}{2}\left[1 - \frac{d}{1-d}(1-s)\right] + \frac{d(1-s)(1-r)}{4(1-d)}, & A_0(r,s) &= \frac{1+s}{2},\\ B_\pm(r,s) &= \frac{1-r}{2}\left[1 - \frac{d}{1-d}(1-s)\right] + \frac{d(1-s)(1+r)}{4(1-d)}, & B_0(r,s) &= \frac{1-s}{4},\\ C_\pm(r,s) &= \frac{d}{2(1-d)}(1-s), & C_0(r,s) &= \frac{1-s}{4}, \end{aligned} \tag{A.3}$$

with $r, s \in [0;1]$ (whose meaning we specify soon, *vide infra*).

Equation (A.2) codes for the new noise, and the values of the coefficients presented in (A.3) have been chosen in order that all the examples contain on average the same fraction $d$ of null entries as the original archetypes. To see this, it is enough to check that the following relation holds for each $a = 1,\dots,M$, $i = 1,\dots,N$ and $\mu = 1,\dots,K$

$$\mathbb{P}(\tilde{\eta}_i^{\mu,a} = 0) = \sum_{x\in\{-1,0,1\}} \mathbb{P}(\tilde{\eta}_i^{\mu,a} = 0|\xi_i^\mu = x)\,\mathbb{P}(\xi_i^\mu = x) = C_\pm(r,s)(1-d) + A_0(r,s)\,d = d. \tag{A.4}$$

After defining the dataset, the cost function follows straightforwardly in Hebbian settings as:

**Definition 11.** After introducing $N$ Ising neurons $\sigma_i = \pm 1$ $(i = 1,\dots,N)$ and the dataset considered in the definition above, the cost function of the multitasking Hebbian network equipped with not-preserving-dilution noise reads as

$$\mathcal{H}^{(sup)}_{N,K,M,r,s,d}(\boldsymbol{\sigma}|\tilde{\boldsymbol{\eta}}) = -\frac{1}{N}\frac{1}{(1-d)(1+\tilde{\rho})}\sum_{\mu=1}^{K}\sum_{i,j=1}^{N,N}\left(\frac{1}{\tilde{r}M}\sum_{a=1}^{M}\tilde{\eta}_i^{\mu,a}\right)\left(\frac{1}{\tilde{r}M}\sum_{b=1}^{M}\tilde{\eta}_j^{\mu,b}\right)\sigma_i\sigma_j, \tag{A.5}$$

J. Stat. Mech. (2023) 113401

where

$$\tilde{r} = \frac{r}{(1-d)}\left[1 - \frac{d}{2}(5-3s)\right] \tag{A.6}$$

and $\tilde{\rho}$ is the generalization of the dataset entropy, defined as

$$\tilde{\rho} = \frac{1-\tilde{r}^2}{M\tilde{r}^2}\,. \tag{A.7}$$

**Definition 12.** The suitably re-normalized example's magnetizations $n_\mu$ read as

$$n_\mu := \frac{1}{(1+\tilde{\rho})}\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\tilde{r}M}\sum_{a=1}^{M}\tilde{\eta}_i^{\mu,a}\right)\sigma_i\,. \tag{A.8}$$

En route toward the free energy, still preserving Guerra's interpolation as the underlying technique, we give the next definition:

**Definition 13.** After introducing the noise $\beta \in \mathbb{R}^+$, an interpolating parameter $t \in (0,1)$, the $K+1$ auxiliary fields $J$ and $\psi_\mu$ ($\mu \in (1,\dots,K)$), the interpolating partition function related to the model defined by the cost function (A.5) reads as

$$\begin{aligned}&\mathcal{Z}^{(\mathrm{sup})}_{\beta,N,K,M,r,s,d}(\boldsymbol{\xi},\tilde{\boldsymbol{\eta}};J,t)\\ &= \sum_{\{\boldsymbol{\sigma}\}}\exp\left[J\sum_{\mu,i=1}^{K,N}\xi_i^\mu\sigma_i + t\beta N\frac{(1+\tilde{\rho})}{2(1-d)}\sum_{\mu=1}^{K}n_\mu^2(\boldsymbol{\sigma}) + (1-t)N\sum_{\mu=1}^{K}\psi_\mu\, n_\mu(\boldsymbol{\sigma})\right].\end{aligned} \tag{A.9}$$

and the interpolating free energy $\mathcal{F}_{\beta,K,M,r,s,d} = \lim_{N\to\infty}\mathcal{F}_{\beta,N,K,M,r,s,d}$ induced by the partition function (A.9) reads as

$$-\beta\mathcal{F}_{\beta,N,K,M,r,s,d}(J,t) = \frac{1}{N}\mathbb{E}\left[\ln\mathcal{Z}^{(\mathrm{sup})}_{\beta,N,K,M,r,s,d}(\boldsymbol{\xi},\tilde{\boldsymbol{\eta}};J,t)\right] \tag{A.10}$$

where $\mathbb{E} = \mathbb{E}_\xi\mathbb{E}_{(\tilde{\eta}|\xi)}$.

**Remark 3.** Of course, as in the model studied in the main text still with Guerra's interpolation technique, we aim to find an explicit expression (in terms of the control and order parameters of the theory) of the interpolating free energy evaluated at $t=1$ and $J=0$.

We thus perform computations following the same steps as the previous investigation: the $t$ derivative of interpolating free energy is given by

$$-\beta\frac{\mathrm{d}\mathcal{F}_{\beta,K,M,r,s,d}(J,t)}{\mathrm{d}t} = \frac{\beta}{2(1-d)}(1+\tilde{\rho})\sum_{\mu=1}^{K}\langle n_\mu^2\rangle_t - \sum_{\mu=1}^{K}\psi_\mu\langle n_\mu\rangle_t. \tag{A.11}$$

fixing

$$\psi_\mu = \frac{\beta}{1-d}(1+\tilde{\rho})\,\bar{n}_\mu \tag{A.12}$$



J. Stat. Mech. (2023) 113401

and computing the one-body term

$$-\beta\mathcal{F}_{\beta,K,M,r,s,d}(J,t=0)=\mathbb{E}\ln\left[2\cosh\left(\sum_{\mu=1}^{K}\psi_\mu\frac{1}{(1+\tilde{\rho})}\frac{1}{\tilde{r}M}\sum_{a=1}^{M}\tilde{\eta}_i^{\mu,a}+J\sum_{\mu=1}^{K}\xi^\mu\right)\right]$$
$$=\mathbb{E}\ln\left\{2\cosh\left[\frac{\beta}{1-d}\sum_{\mu=1}^{K}\bar{n}_\mu\left(\frac{1}{\tilde{r}M}\sum_{a=1}^{M}\tilde{\eta}_i^{\mu,a}\right)+J\sum_{\mu=1}^{K}\xi^\mu\right]\right\}. \tag{A.13}$$

We get the final expression as $N\to\infty$, such that we can state the next theorem:

**Theorem 2.** *In the thermodynamic limit ($N\to\infty$) and in the low-load regime ($K/N\to 0$), the quenched free energy of the multitasking Hebbian network equipped with not-preserving-dilution noise, regardless of the presence of a teacher, reads as*

$$-\beta\mathcal{F}_{\beta,K,M,r,s,d}(J)=\mathbb{E}\left\{\ln\left[2\cosh\left(\beta'\sum_{\mu=1}^{K}\bar{n}_\mu\hat{\eta}^\mu+J\sum_{\mu=1}^{K}\xi^\mu\right)\right]\right\}-\frac{\beta'}{2}(1+\tilde{\rho})\sum_{\mu=1}^{K}\bar{n}_\mu^2. \tag{A.14}$$

*where $\beta'=\beta/(1-d)$, $\mathbb{E}=\mathbb{E}_\xi\mathbb{E}_{(\tilde{\eta}|\xi)}$ and $\hat{\eta}^\mu=\dfrac{1}{\tilde{r}M}\sum_{a=1}^{M}\tilde{\eta}_i^{\mu,a}$ and the values $\bar{n}_\mu$ must fulfill the following self-consistent equations:*

$$\bar{n}_\mu=\frac{1}{(1+\tilde{\rho})}\mathbb{E}\left\{\left[\tanh\left(\beta'\sum_{\nu=1}^{K}\bar{n}_\nu\hat{\eta}^\nu\right)\right]\hat{\eta}^\mu\right\}\qquad\text{for }\mu=1,\dots,K, \tag{A.15}$$

*that extremize the free energy $\mathcal{F}_{\beta,K,M,r,s,d}(J=0)$ w.r.t. them.*

Furthermore, the simplest path to obtain a self-consistent equation also for the Mattis magnetization $m_\mu$ is by considering the auxiliary field $J$ coupled to $m_\mu$; namely $\bar{m}_\mu=-\beta\nabla_J\mathcal{F}_{\beta,K,M,r,s,d}(J)|_{J=0}$, to get

$$\bar{m}_\mu=\mathbb{E}\left[\tanh\left(\beta'\sum_{\nu=1}^{K}\bar{n}_\nu\hat{\eta}^\nu\right)\xi^\mu\right]\qquad\text{for }\mu=1,\dots,K. \tag{A.16}$$

We do not plot these new self-consistency equations as, in the large $M$ limit, there are no differences w.r.t. those obtained in the main text (please refer to figure 8).

J. Stat. Mech. (2023) 113401

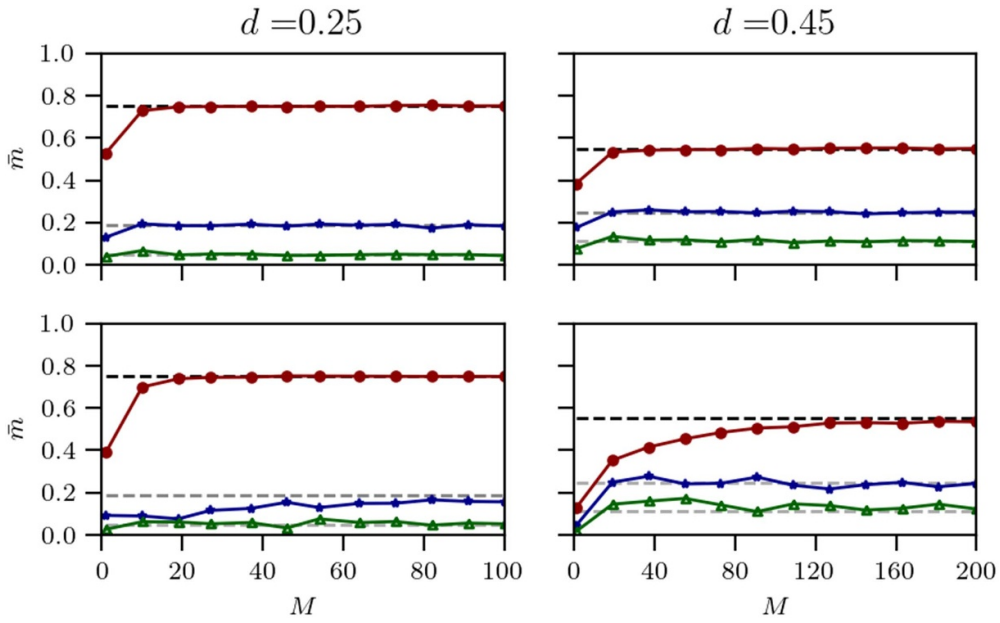


**Figure 8.** Comparison of the numerical solution of the selfconsistency equations related to the Mattis magnetization in the two models: the upper panel relates to the first model (reported in the main text), and the lower panel reports on the second model (deepened here). Aside from the different transient at small $M$, the two models behave essentially in the same way.

## Appendix B. On the dataset entropy $\rho$

In this appendix, focusing on a single generic bit, we deepen the relation between the conditional entropy $H(\xi_i^\mu|\boldsymbol{\eta}_i^\mu)$ of a given pixel $i$ regarding archetype $\mu$ and the information provided by the dataset regarding such a pixel, namely the block $(\eta_i^{\mu,1},\eta_i^{\mu,2},\dots,\eta_i^{\mu,M})$, to justify why we called $\rho$ the dataset entropy in the main text. As the calculations are slightly different between the two analyzed models (the one preserving the dilution position provided in the main text and the generalized one given in the previous appendix) we repeat them model by model for the sake of transparency.

### B.1. I: Multitasking Hebbian network equipped with not-affecting-dilution noise

Let us focus on the $\mu$th pattern and the $i$th digit, whose related block is

$$\eta_i^\mu = \left(\eta_i^{\mu,1},\eta_i^{\mu,2},\dots,\eta_i^{\mu,M}\right); \tag{B.1}$$

the error probability for any single entry is

$$\mathbb{P}(\xi_i^\mu \neq 0)\,\mathbb{P}(\eta_i^{\mu,a} \neq \xi_i^\mu) = (1-d)\,(1-r)\,/2 \tag{B.2}$$



J. Stat. Mech. (2023) 113401

and, by applying the majority rule on the block, it is reduced to

$$\mathbb{P}(\xi_i^\mu \neq 0)\,\mathbb{P}\left[\text{sign}\left(\sum_a \eta_i^{\mu,a}\xi_i^\mu\right) = -1\right] \underset{M\gg 1}{\approx} \frac{(1-d)}{2}\left[1-\text{erf}\left(\frac{1}{\sqrt{2\rho}}\right)\right]. \quad \text{(B.3)}$$

Thus,

$$H_{d,r,M}\left(\boldsymbol{\xi}^\mu|\boldsymbol{\eta}^\mu\right) = -\left[x\left(d,r,M\right)\log_2 x\left(d,r,M\right) + y\left(d,r,M\right)\log_2 y\left(d,r,M\right)\right] \quad \text{(B.4)}$$

where

$$x\left(d,r,M\right) = \frac{(1-d)}{2}\left[1-\text{erf}\left(\frac{1}{\sqrt{2\rho}}\right)\right], \quad y\left(d,r,M\right) = 1 - x\left(d,r,M\right)\ . \quad \text{(B.5)}$$

### B.2. II: Multitasking Hebbian network equipped with not-preserving-dilution noise

Let us focus on the $\mu$th pattern and the $i$th digit, whose related block is

$$\tilde{\eta}_i^\mu = \left(\tilde{\eta}_i^{\mu,1}, \tilde{\eta}_i^{\mu,2}, \ldots, \tilde{\eta}_i^{\mu,M}\right); \quad \text{(B.6)}$$

the error probability for any single entry is

$$\mathbb{P}(\xi_i^\mu \neq 0)\,\mathbb{P}(\tilde{\eta}_i^{\mu,a}\xi_i^\mu \neq +1|\xi_i^\mu \neq 0) + \mathbb{P}(\xi_i^\mu = 0)\,\mathbb{P}(\tilde{\eta}_i^{\mu,a} \neq 0|\xi_i^\mu = 0) = d(1-s)\,. \quad \text{(B.7)}$$

By applying the majority rule on the block, it is reduced to

$$\mathbb{P}(\xi_i^\mu \neq 0)\left[1-\mathbb{P}\left(\text{sign}\left(\hat{\eta}_i^\mu\xi_i^\mu\right) = +1\middle|\xi_i^\mu \neq 0\right)\right] + \mathbb{P}(\xi_i^\mu = 0)\,\mathbb{P}\left(\text{sign}\left[|\hat{\eta}_i^\mu|\right] = +1\middle|\xi_i^\mu = 0\right)$$

$$\underset{M\gg 1}{\approx} \frac{(1-d)}{2}\left\{1-\text{erf}\left[\left(2\tilde{\rho} - \frac{d(1-s)}{(1-d)M\tilde{r}^2}\right)^{-1/2}\right]\right\} + \frac{d}{2}\left\{1-\text{erf}\left[\left(\frac{1-s}{M\tilde{r}^2}\right)^{-1/2}\right]\right\}. \quad \text{(B.8)}$$

Thus,

$$H_{d,r,s,M}\left(\xi_i^\mu|\tilde{\boldsymbol{\eta}}_i^\mu\right) = -\left[x\left(d,r,s,M\right)\log_2 x\left(d,r,s,M\right) + y\left(d,r,s,M\right)\log_2 y\left(d,r,s,M\right)\right] \quad \text{(B.9)}$$

where

$$x\left(d,r,s,M\right) = \frac{(1-d)}{2}\left\{1-\text{erf}\left[\left(2\tilde{\rho} - \frac{d(1-s)}{(1-d)M\tilde{r}^2}\right)^{-1/2}\right]\right\} + \frac{d}{2}\left\{1-\text{erf}\left[\left(\frac{1-s}{M\tilde{r}^2}\right)^{-1/2}\right]\right\}$$

$$y\left(d,r,s,M\right) = 1 - x\left(d,\tilde{\rho}\right). \quad \text{(B.10)}$$

Whatever the model, the conditional entropies $H_{d,r,M}(\xi_i^\mu|\boldsymbol{\eta}_i^\mu)$ and $H_{d,r,s,M}(\xi_i^\mu|\tilde{\boldsymbol{\eta}}_i^\mu)$ are monotonic increasing functions of $\rho$ and $\tilde{\rho}$, respectively, hence the reason for calling $\rho$ and $\tilde{\rho}$ the entropy of the dataset.



J. Stat. Mech. (2023) 113401

## Appendix C. Explicit calculations and figures for the cases $K=2$ and $K=3$

In this appendix, we collect the explicit expression of the self-consistent equations in (3.10) and (3.11) (focusing only on the cases of $K=2$ and $K=3$) and some figures obtained from their numerical solution.

### C.1. $K=2$

Fixing $K=2$ and explicitly calculating the mean with respect to $\xi$, (3.10) and (3.11) read as

$$\bar{n}_1 = \frac{\bar{m}_1}{(1+\rho)} + \frac{\beta'(1-d)\rho\bar{n}_1}{(1+\rho)}\left[1 - d\mathcal{S}_2(\bar{n}_1,0) - \frac{(1-d)}{2}\mathcal{S}_2(\bar{n}_1,-\bar{n}_2) - \frac{(1-d)}{2}\mathcal{S}_2(\bar{n}_1,\bar{n}_2)\right]$$

$$\bar{n}_2 = \frac{\bar{m}_2}{(1+\rho)} + \frac{\beta'(1-d)\rho\bar{n}_2}{(1+\rho)}\left[1 - d\mathcal{S}_2(0,\bar{n}_2) - \frac{(1-d)}{2}\mathcal{S}_2(\bar{n}_1,-\bar{n}_2) - \frac{(1-d)}{2}\mathcal{S}_2(\bar{n}_1,\bar{n}_2)\right]$$

$$\bar{m}_1 = \frac{(1-d)^2}{2}\left[\mathcal{T}_2(\bar{n}_1,\bar{n}_2) + \mathcal{T}_2(\bar{n}_1,-\bar{n}_2)\right] + d(1-d)\,\mathcal{T}_2(\bar{n}_1,0)$$

$$\bar{m}_1 = \frac{(1-d)^2}{2}\left[\mathcal{T}_2(\bar{n}_1,\bar{n}_2) - \mathcal{T}_2(\bar{n}_1,-\bar{n}_2)\right] + d(1-d)\mathcal{T}_2(0,\bar{n}_2) \tag{C.1}$$

where we used

$$\begin{aligned}\mathcal{T}_2(x,y) &= \mathbb{E}_\lambda \tanh\left[\beta'\left(x+y+\lambda\sqrt{\rho(x^2+y^2)}\right)\right],\\ \mathcal{S}_2(x,y) &= \mathbb{E}_\lambda \tanh^2\left[\beta'\left(x+y+\lambda\sqrt{\rho(x^2+y^2)}\right)\right].\end{aligned} \tag{C.2}$$

Numerically solving this set of equations, we construct the plots presented in figure 9.

### C.2. $K=3$

Moving on to the case of $K=3$, by following the same steps as in the previous subsection, we get

$$\begin{aligned}\bar{n}_1 = \frac{\bar{m}_1}{(1+\rho)} + \frac{\beta'(1-d)\rho\bar{n}_1}{(1+\rho)}\Bigg\{1 - d\frac{(1-d)}{2}\left[\mathcal{S}_3(\bar{n}_1,\bar{n}_2,0) + \mathcal{S}_3(\bar{n}_1,0,\bar{n}_3) + \mathcal{S}_3(\bar{n}_1,-\bar{n}_2,0)\right.\\ \left. + \mathcal{S}_3(\bar{n}_1,0,-\bar{n}_3)\right] - d^2\mathcal{S}_3(\bar{n}_1,0,0) - \frac{(1-d)^2}{4}\left[\mathcal{S}_3(\bar{n}_1,\bar{n}_2,\bar{n}_3) + \mathcal{S}_3(\bar{n}_1,\bar{n}_2,-\bar{n}_3)\right.\\ \left. + \mathcal{S}_3(\bar{n}_1,-\bar{n}_2,\bar{n}_3) + \mathcal{S}_3(\bar{n}_1,-\bar{n}_2,-\bar{n}_3)\right]\Bigg\},\end{aligned}$$



J. Stat. Mech. (2023) 113401

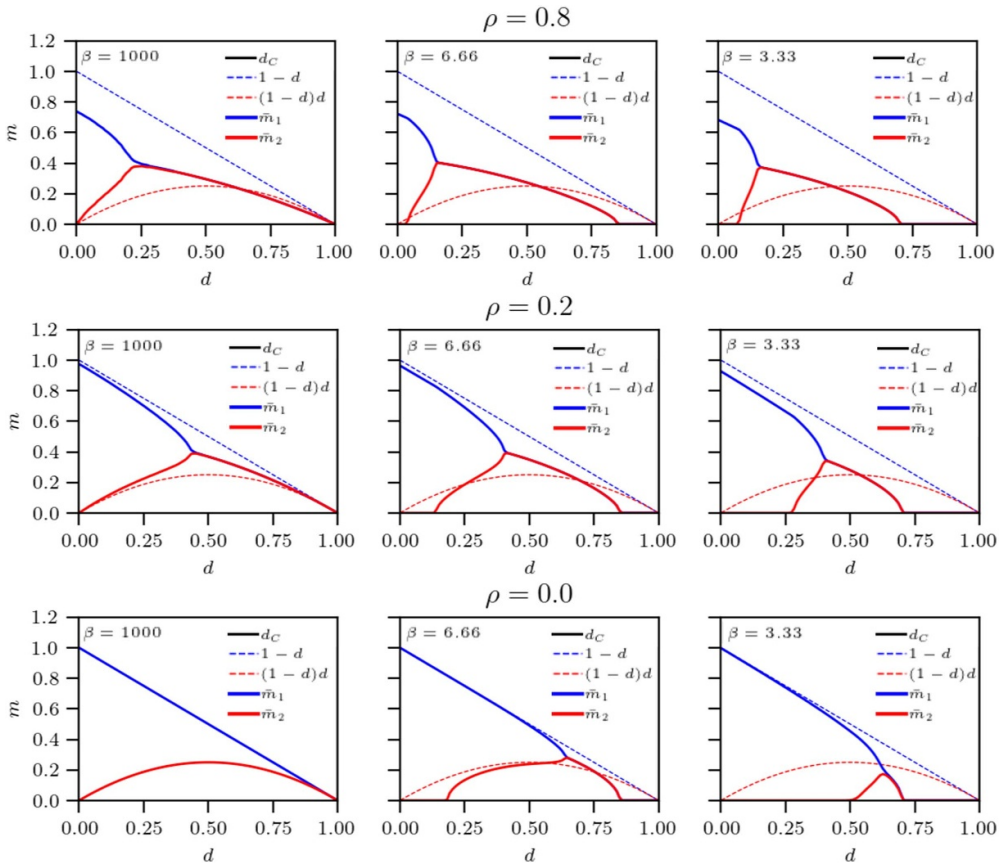


**Figure 9.** Numerical resolution of the system of equation (C.1) for $K=2$: we plot the behavior of the magnetization $\bar{m}$ versus the degree of dilution $d$ for fixed $r=0.2$ and different values of $\beta$ (from right to left $\beta=1000, 6.66, 3.33$) and $\rho$ (from top to bottom $\rho=0.8, 0.2, 0.0$). We stress that for $\rho=0.0$ we recover the standard diluted model presented in figure 1.

$$\begin{aligned}\bar{m}_1 = &\frac{(1-d)^3}{4}\left[\mathcal{T}_3(\bar{n}_1,\bar{n}_2,\bar{n}_3)+\mathcal{T}_3(\bar{n}_1,\bar{n}_2,-\bar{n}_3)+\mathcal{T}_3(\bar{n}_1,-\bar{n}_2,\bar{n}_3)+\mathcal{T}_3(\bar{n}_1,-\bar{n}_2,-\bar{n}_3)\right]\\ &+d\frac{(1-d)^2}{2}\left[\mathcal{T}_3(\bar{n}_1,\bar{n}_2,0)+\mathcal{T}_3(\bar{n}_1,0,\bar{n}_3)+\mathcal{T}_3(\bar{n}_1,-\bar{n}_2)+\mathcal{T}_3(\bar{n}_1,0,-\bar{n}_3)\right]\\ &+d^2(1-d)\mathcal{T}_3(\bar{n}_1,0,0)\,,\end{aligned} \tag{C.3}$$

where we used

$$\begin{aligned}\mathcal{T}_3(x,y,z) &= \mathbb{E}_\lambda \tanh\left[\beta'\left(x+y+z+\lambda\sqrt{\rho(x^2+y^2+z^2)}\right)\right],\\ \mathcal{S}_3(x,y,z) &= \mathbb{E}_\lambda \tanh^2\left[\beta'\left(x+y+z+\lambda\sqrt{\rho(x^2+y^2+z^2)}\right)\right].\end{aligned} \tag{C.4}$$

In order to lighten the presentation, we report only the expression of $\bar{m}_1$ and $\bar{n}_1$, and the related expressions of $\bar{m}_2(\bar{m}_3)$ and $\bar{n}_2(\bar{n}_3)$ can be obtained by making the simple

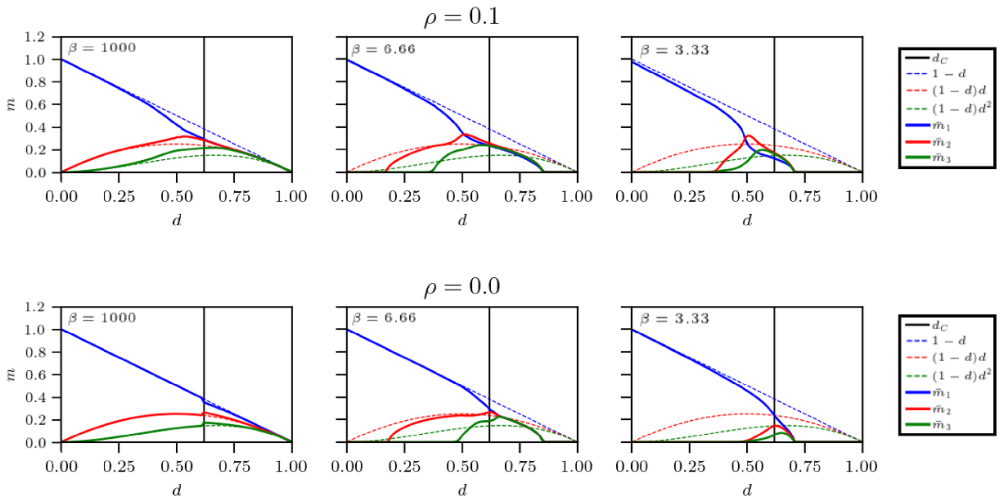


**Figure 10.** Numerical solution of the system of equation (C.3) for $K=3$: we plot the behavior of the magnetization $\bar{m}$ versus the degree of dilution $d$ for fixed $r=0.2$ and different values of $\beta$ (from left to right $\beta=1000, 6.66, 3.33$) and $\rho$ (from top to bottom $\rho=0.8, 0.2, 0.0$).

substitutions $\bar{m}_1 \longleftrightarrow \bar{m}_2(\bar{m}_3)$, and $\bar{n}_1 \longleftrightarrow \bar{n}_2(\bar{n}_3)$ in (C.3). The numerical solution of the previous set of equations is depicted in figure 10.

## Appendix D. Proofs

### D.1. Proof of theorem 1

In this subsection we show the proof of proposition 1. In order to prove the aforementioned proposition, we put in front of it the following:

**Lemma 1.** *The $t$ derivative of interpolating free energy is given by*

$$-\beta \frac{\mathrm{d}\mathcal{F}^{(\mathrm{sup,unsup})}_{N,K,\beta,d,M,r}}{\mathrm{d}t} = \frac{\beta}{2(1-d)}(1+\rho)\sum_{\mu=1}^{K}\mathbb{E}\omega_t\left[n_\mu^2\right] - \sum_{\mu=1}^{K}\psi_\mu \mathbb{E}\omega_t\left[n_\mu\right]. \tag{D.1}$$

Since the computation is lengthy but not cumbersome we omit it.

**Proposition 4.** *In the low-load regime, in the thermodynamic limit the distribution of the generic order parameter $X$ is centered at its expectation value $\bar{X}$ with vanishing fluctuations. Thus, since $\Delta X = X - \bar{X}$, in the thermodynamic limit, the following relation holds:*

$$\mathbb{E}\omega_t\left[(\Delta X)^2\right] \xrightarrow[N\to+\infty]{} 0. \tag{D.2}$$

**Remark 4.** We stress that afterwards we use the relations

$$\mathbb{E}\omega_t\left[(n_\mu - \bar{n}_\mu)^2\right] = \mathbb{E}\omega_t\left[n_\mu^2\right] - 2\bar{n}_\mu \mathbb{E}\omega_t\left[n_\mu\right] + \bar{n}_\mu^2. \tag{D.3}$$

which are computed with brute force with Newton's binomial.



J. Stat. Mech. (2023) 113401

Now, using these relations, if we fix the constants as

$$\psi_\mu = \frac{\beta}{1-d}(1+\rho)\,\bar{n}_\mu \tag{D.4}$$

in the thermodynamic limit, due to proposition 4, the expression of derivative w.r.t. $t$ becomes

$$-\beta\frac{\mathrm{d}\mathcal{F}^{(\mathrm{sup,unsup})}_{K,\beta,d,M,r}}{\mathrm{d}t} = -\frac{\beta}{2(1-d)}(1+\rho)\sum_{\mu=1}^{K}\bar{n}_\mu^2. \tag{D.5}$$

**Proof.** Let us start from finite-size $N$ expression. We apply the fundamental theorem of calculus:

$$\mathcal{F}^{(\mathrm{sup,unsup})}_{K,\beta,d,M,r} = \mathcal{F}^{(\mathrm{sup,unsup})}_{K,\beta,d,M,r}(t=1) = \mathcal{F}^{(\mathrm{sup,unsup})}_{K,\beta,d,M,r}(t=0) + \int_0^1 \partial_s \mathcal{F}^{(\mathrm{sup,unsup})}_{K,\beta,d,M,r}(s)\Big|_{s=t}\mathrm{d}t. \tag{D.6}$$

We have already computed the derivative w.r.t. $t$ in equation (D.5). It only remains to calculate the one-body term:

$$\begin{aligned}\mathcal{Z}^{(\mathrm{sup,unsup})}_{N,K,\beta,d,M,r}(t=0) &= \prod_{i=1}^{N}\sum_{\{\boldsymbol{\sigma}\}}\exp\left[\left(\sum_{\mu=1}^{K}\frac{\psi_\mu}{2(1+\rho)}\hat{\eta}^\mu + J\xi^\mu\right)\sigma_i\right]\\ &= 2^N\cosh^N\left(\sum_{\mu=1}^{K}\frac{\psi_\mu}{2(1+\rho)}\hat{\eta}^\mu + J\xi^\mu\right).\end{aligned} \tag{D.7}$$

Using the definition of quenched free energy (3.5) we have

$$\begin{aligned}-\beta\mathcal{F}^{(\mathrm{sup,unsup})}_{K,\beta,d,M,r}(J,t=0) &= \ln\left[2\cosh\left(\sum_{\mu=1}^{K}\frac{\psi_\mu}{2(1+\rho)}\hat{\eta}^\mu + J\xi^\mu\right)\right]\\ &= \mathbb{E}\left[\ln 2\cosh\left(\frac{\beta}{1-d}\sum_{\mu=1}^{K}\bar{n}_\mu\hat{\eta}^\mu + J\xi^\mu\right)\right]\end{aligned} \tag{D.8}$$

where $\mathbb{E} = \mathbb{E}_\xi\,\mathbb{E}_{(\eta|\xi)}$. Finally, adding in (D.6), (D.8) and (D.5), we reach the thesis. □

### D.2. Proof of proposition 1

In this subsection we show the proof of proposition 1.

**Proof.** For large datasets, using the central limit theorem we have

$$\hat{\eta}^\mu \sim \xi^\mu\left(1+\sqrt{\rho}\,Z_\mu\right). \tag{D.9}$$



J. Stat. Mech. (2023) 113401

where $Z_\mu$ is a standard Gaussian variable $Z_\mu \sim \mathcal{N}(0,1)$. Replacing equation (D.9) in the self-consistency equation for $\bar{n}$, namely equation (3.8), and applying Stein's lemma[14] in order to recover the expression for $\bar{m}_\mu$, we get the large dataset equation for $\bar{n}_\mu$, i.e. equation (3.10).

We will use the relation

$$\mathbb{E}_{\lambda_\mu}\left[F\left(a+\sum_{\mu=1}^{K} b_\mu \lambda_\mu\right)\right] = \mathbb{E}_Z\left[F\left(a+Z\sqrt{\sum_{\mu=1}^{K} b_\mu^2}\right)\right], \tag{D.11}$$

where $\lambda_\mu$ and $Z$ are i.i.d. Gaussian variables. In doing so, we obtain

$$g(\beta,\boldsymbol{\xi},Z,\bar{\boldsymbol{n}}) = \beta'\sum_{\nu=1}^{K}\bar{n}_\nu\xi^\nu + \beta'\sqrt{\rho}\sum_{\nu=1}^{K} Z_\nu \bar{n}_\nu^2 (\xi^\nu)^2 = \beta'\left(\sum_{\nu=1}^{K}\bar{n}_\nu\xi^\nu + Z\sum_{\nu=1}^{K}\sqrt{\rho\,\bar{n}_\nu^2(\xi^\nu)^2}\right), \tag{D.12}$$

and thus we reach the thesis. □

**Corollary 2.** *The self-consistency equations in the large dataset assumption and null-temperature limit are*

$$\bar{m}_\mu = \mathbb{E}_\xi\left\{\mathrm{erf}\left[\left(\sum_{\nu=1}^{K}\bar{m}_\nu\xi^\nu\right)\left(2\rho\sum_{\nu=1}^{K}\bar{m}_\nu^2(\xi^\nu)^2\right)^{-1/2}\right]\xi^\mu\right\}. \tag{D.13}$$

**Proof.** In order to lighten the notation, we rename

$$C = \tanh^2\left[g(\beta,\boldsymbol{\xi},Z,\bar{\boldsymbol{n}})\right]. \tag{D.14}$$

We start by assuming finite the limit

$$\lim_{\beta'\to\infty}\beta'(1-C) = D \in \mathbb{R} \tag{D.15}$$

and we stress that as $\beta'\to\infty$ we have $C\to 1$. As a consequence, the following reparametrization is found to be useful:

$$C = 1 - \frac{\delta C}{\beta'} \quad \text{as} \quad \beta'\to\infty. \tag{D.16}$$

[14] This lemma, also known as Wick's theorem, applies to standard Gaussian variables, say $J\sim\mathcal{N}(0,1)$, and states that, for a generic function $f(J)$ for which the two expectations $\mathbb{E}(Jf(J))$ and $\mathbb{E}(\partial_J f(J))$ both exist, then

$$\mathbb{E}(Jf(J)) = \mathbb{E}\left(\frac{\partial f(J)}{\partial J}\right). \tag{D.10}$$

.



J. Stat. Mech. (2023) 113401

Therefore, as $\beta' \to \infty$, it yields

$$\begin{aligned} \bar{n}_\mu &= \frac{\bar{m}_\mu}{1+\rho-\rho\,\delta C\,(1-d)} \\ \bar{m}_\mu &= \mathbb{E}_\xi \mathbb{E}_Z \left[ \mathrm{sign}\left( \sum_{\nu=1}^{K} \bar{n}_\nu \xi^\nu + Z \sum_{\nu=1}^{K} \sqrt{\rho\, \bar{n}_\nu^2 \left(\xi^\nu\right)^2} \right) \xi^\mu \right]; \end{aligned} \tag{D.17}$$

to reach this result, we have also used the relation

$$\mathbb{E}_z \mathrm{sign}\left[A + Bz\right] = \mathrm{erf}\left[\frac{A}{\sqrt{2}B}\right], \tag{D.18}$$

where $z$ is a Gaussian variable $\mathcal{N}(0,1)$ and the truncated expression $\bar{n}_\mu = \bar{m}_\mu/(1+\rho)$ for the first equation in (D.17). □

## References


[1] Ackley D H, Hinton G E and Sejnowski T J 1985 A learning algorithm for Boltzmann machines *Cogn. Sci.* **9** 147–69

[2] Agliari E, Alemanno F, Barra A and De Marzo G 2022 The emergence of a concept in shallow neural networks *Neural Netw.* **148** 232–53

[3] Agliari E, Alemanno F, Barra A and Fachechi A 2020 Generalized Guerra's interpolation schemes for dense associative neural networks *Neural Netw.* **128** 254–67

[4] Agliari E, Annibale A, Barra A, Coolen A and Tantari D 2013 Immune networks: multi-tasking capabilities at medium load *J. Phys. A: Math. Theor.* **46** 335101

[5] Agliari E, Annibale A, Barra A, Coolen A and Tantari D 2013 Immune networks: multitasking capabilities near saturation *J. Phys. A: Math. Theor.* **46** 415003

[6] Agliari E, Aquaro M, Barra A, Fachechi A and Marullo C 2023 From Pavlov conditioning to Hebb learning *Neural Comput.* **35** 930–57

[7] Agliari E, Barra A, Galluzzi A, Guerra F and Moauro F 2012 Multitasking associative networks *Phys. Rev. Lett.* **109** 268101

[8] Agliari E, Barra A, Galluzzi A, Guerra F, Tantari D and Tavani F 2015 Retrieval capabilities of hierarchical networks: from Dyson to Hopfield *Phys. Rev. Lett.* **114** 028103

[9] Agliari E, Barra A, Galluzzi A and Isopi M 2014 Multitasking attractor networks with neuronal threshold noise *Neural Netw.* **49** 19–29

[10] Agliari E, Barra A, Sollich P and Zdeborová L 2020 Machine learning and statistical physics: theory, inspiration, application *J. Phys. A: Math. Theor.* **53** 500401

[11] Alemanno F, Aquaro M, Kanter I, Barra A and Agliari E 2023 Supervised Hebbian learning *Europhys. Lett.* **141** 11001

[12] Amit D J and Amit D J 1989 *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge University Press)

[13] Amit D J, Gutfreund H and Sompolinsky H 1985 Spin-glass models of neural networks *Phys. Rev.* A **32** 1007

[14] Amit D J, Gutfreund H and Sompolinsky H 1985 Storing infinite numbers of patterns in a spin-glass model of neural networks *Phys. Rev. Lett.* **55** 1530

[15] Barbier J and Macris N 2019 The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference *Probab. Theory Relat. Fields* **174** 1133–85

[16] Barra A, Genovese G and Guerra F 2010 The replica symmetric approximation of the analogical neural network *J. Stat. Phys.* **140** 784–96

[17] Bovier A 2006 *Statistical Mechanics of Disordered Systems: A Mathematical Perspective* vol 18 (Cambridge University Press)

[18] Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N, Vogt-Maranto L and Zdeborová L 2019 Machine learning and the physical sciences *Rev. Mod. Phys.* **91** 045002




J. Stat. Mech. (2023) 113401

[19] Coolen A C, Kühn R and Sollich P 2005 *Theory of Neural Information Processing Systems* (Oxford University Press)
[20] Decelle A and Furtlehner C 2021 Restricted Boltzmann machine: recent advances and mean-field theory *Chin. Phys.* B **30** 040202
[21] Decelle A and Ricci-Tersenghi F 2014 Pseudolikelihood decimation algorithm improving the inference of the interaction network in a general class of Ising models *Phys. Rev. Lett.* **112** 070603
[22] Decelle A, Ricci-Tersenghi F and Zhang P 2016 Data quality for the inverse Lsing problem *J. Phys. A: Math. Theor.* **49** 384001
[23] Engel A 2001 *Statistical Mechanics of Learning* (Cambridge University Press)
[24] Fontanari J 1990 Generalization in a Hopfield network *J. Phys. France* **51** 2421–30
[25] Guerra F 2001 Sum rules for the free energy in the mean field spin glass model *Mathematical Physics in Mathematics and Physics: Quantum and Operator Algebraic Aspects* vol 30 (Fields Institute Communications)
[26] Guerra F 2003 Broken replica symmetry bounds in the mean field spin glass model *Commun. Math. Phys.* **233** 1–12
[27] Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities *Proc. Natl. Acad. Sci.* **79** 2554–8
[28] Huang H 2017 Statistical mechanics of unsupervised feature learning in a restricted Boltzmann machine with binary synapses *J. Stat. Mech.* 053302
[29] Kang L and Toyoizumi T 2023 A Hopfield-like model with complementary encodings of memories (arXiv:2302.04481)
[30] Ricci-Tersenghi F 2012 The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods *J. Stat. Mech.* P08015
[31] Roussel C, Cocco S and Monasson R 2021 Barriers and dynamical paths in alternating Gibbs sampling of restricted Boltzmann machines *Phys. Rev.* E **104** 034109
[32] Seung H S, Sompolinsky H and Tishby N 1992 Statistical mechanics of learning from examples *Phys. Rev.* A **45** 6056
[33] Sollich P, Tantari D, Annibale A and Barra A 2014 Extensive parallel processing on scale-free networks *Phys. Rev. Lett.* **113** 238106
[34] Talagrand M 1998 Rigorous results for the Hopfield model with many patterns *Probab. Theory Relat. Fields* **110** 177–275
[35] Tubiana J and Monasson R 2017 Emergence of compositional representations in restricted Boltzmann machines *Phys. Rev. Lett.* **118** 138301

J. Stat. Mech. (2023) 113401