Thermodynamics of Restricted Boltzmann Machines and Related Learning Dynamics

A. Decelle *G. Fissore* *C. Furtlehner*

Abstract

We analyze the learning process of the restricted Boltzmann machine (RBM), a certain type of generative models used in the context of unsupervised learning. In a first step, we investigate the thermodynamics properties by considering a realistic statistical ensemble of RBM, assuming the information content of the RBM to be mainly reflected by spectral properties of its weight matrix W. A phase diagram is obtained which seems at first sight similar to the one of the Sherrington-Kirkpatrick (SK) model with ferromagnetic couplings. The main difference resides in the structure of the ferromagnetic phase which may or may not be of compositional type, depending mainly on the distribution's kurtosis of the singular vectors components of W.

In a second step the learning dynamics of an RBM from arbitrary data is studied in thermodynamic limit. A "typical" learning trajectory is shown to solve an effective dynamical equation, based on the aforementioned ensemble average and involving explicitly order parameters obtained from the thermodynamic analysis. This accounts in particular for the dominant singular values evolution and how this is driven by the input data: in the linear regime at the beginning of the learning, they correspond to unstable deformation modes of W reflecting dominant covariance modes of the data. In the non-linear regime it is seen how the selected modes interact in later stages of the learning procedure, by eventually imposing a matching between order parameters with their empirical counterparts estimated from the data. Experiments on both artificial and real data illustrate these considerations, showing in particular how the RBM operates in the ferromagnetic compositional phase.

1 Introduction

The Restricted Boltzmann Machine (RBM) [1] is an important machine learning tool used in many applications, by virtue of its ability to model complex probability distributions. It corresponds to a certain type of neural networks called generative models in the sense that it defines a probability distribution able to approximate in principle any empirical distribution of data points living in some discrete or real space of dimension $N \gg 1$. One of its interest being that it can be seen as one of the simplest neural network generative model and that its probability distribution take a simple analytical form. In its discrete form and when data correspond to binary vectors, it is a bipartite heterogeneous Ising model composed of one layer of visible units (the observable variables) connected to one layer of hidden units (the latent or hidden vari-

ables building up the dependencies between the visible ones), with couplings and fields that are obtained from a learning procedure, given a set of examples. It can also be composed in order to form "deep" architecture by stacking many RBMs. In that case, it has been studied either as a multi-layer generative model, as the Deep Boltzmann Machine (DBM) [2], or, as a pre-training procedure for neural network by training each RBMs separatally [3]. The standard learning procedure called contrastive divergence [4] (CD) or the refined persistence CD [5] (PCD) are based on a quick Monte Carlo estimation of the response function of the RBM and are efficient and well documented [6]. Nevertheless, despite some interesting interpretation of CD in terms of non-equilibrium statistical physics [7], the learning of RBMs remains a set of obscure recipes from the statistical physics point of view: hyperparameters (like the size of the hidden layer) are supposed to be set empirically without any theoretical guidelines.

In similar models (like the Hopfield model), many works during the 1980s in statistical physics [8, 9, 10, 11] managed to define the learning capacity of such a model and in particular to compute how many independent patterns could be stored. It is worth noticing that, RBMs can be as well regarded as a statistical physics model (being defined as a Boltzmann distribution with pairwise interactions on a bipartite graph) and therefore can be studied in a similar way as the Hopfield model. The analogy is even stronger since connections between the Hopfield model and RBMs have been made explicitly when using Gaussian hidden variables [12], here the number of patterns of the Hopfield model corresponding to the number of hidden units. Motivated by the recent excitement for neural networks, recent works actually propose to exploit the statistical physics formulation of RBMs to understand what would be its learning capacity and how mean-field methods can be improved for such models. In [13, 14, 15], mean-field based learning methods using TAP equations are developed. TAP solutions are usually expected to define a decomposition of the measure in terms of pure thermodynamical states and are useful both as an algorithm to compute the marginals of the variables of the model but also to identify the pure states when they are yet unknown. For instance, in a sparse explicit Boltzmann machines, i.e. without latent variables, this implicit clustering can be done by means of belief propagation [1] fixed points with simple empirical learning rules [16]. In [17, 18], an analysis of the static properties of RBMs is done assuming a given weight matrix W, in order to understand collective phenomena in the latent representation, i.e. the way latent variables organize themselves in a compositional phase [19, 20] to represent actual data. These analysis, using the replica trick (or equivalent) make the common assumption that the components of the weight matrix W are i.i.d. This approximation is problematic since, as far as realistic RBM are concerned (RBM learned on data), the learning mechanism introduces correlations within the weights of W and yet it seems rather crude to continue to assume the independence to understand the statistical property of the machine. Concerning the learning procedure of neural networks, many recent statistical physics based analyses have been proposed, most of them within teacher-student setting [21]. This imposes a rather strong assumption on the data in the sense that it is assumed that these are generated from a model belonging to the parametric family of interest, hiding as a consequence the role played by the data themselves in the procedure. From the analysis of related linear models [22, 23], it is already a well established fact that a selection of the most important modes of the singular values decomposition (SVD) of the data is performed in the linear case. In fact in the simpler context of linear feed-forward models the learning dynamics can be fully characterized by means of the

[1] a somewhat different form of the TAP equations

SVD of the data matrix [24], showing in particular the emergence of each mode by order of importance regarding singular values.

First steps to follow this guideline have been done in [25], in the context of a general RBM and to address the shortcomings of previous analyses, in particular concerning the assumptions over the weights distribution. To this end it has been proposed to characterize both the learned RBM and the learning process itself by means of the SVD spectrum of the weight matrix in order to single out the information content of an RBM. This information content is assumed to be represented by singular values outside of a continuous bulk corresponding to noise. By doing this it is possible to go beyond the usual unrealistic assumption of i.i.d. weights made for analyzing RBMs. Proceeding along this direction, in the present work we first present a thermodynamic analysis of RBMs exploiting the proposed more realistic assumptions over the weight matrix. Then, on the same basis, the learning dynamics of RBMs are studied by direct analysis of the dynamics of the SVD modes, both in the linear and non-linear regimes.

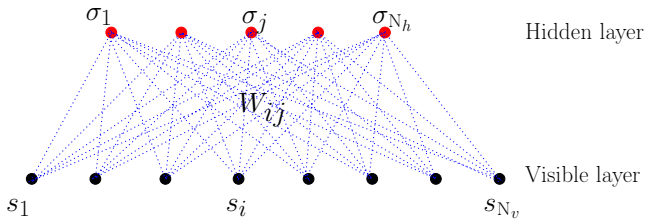

Fig. 1: RBM.

The paper is organized as follows: In Section 2 we introduce some basic facts about RBM and its associated learning procedures. The Section 3 is concerned with static thermodynamical properties of the RBM with realistic hypothesis on the RBMs weights: a statistical ensemble of weight matrices is discussed in Section 3.1; Mean-field equations in the replica-symmetric framework are given in Section 3.2 and the corresponding phase diagram is studied in Section 3.3 with a proper delimitation of the RS domain where the learning procedure is supposed to take place. The ferromagnetic phase is studied in great details in 3.4 by looking in particular at the conditions leading to a compositional phase. The Section 4 is devoted to the learning dynamics. A deterministic learning equation in the thermodynamic limit is proposed in Section 4.1, in which a set of dynamical parameters is shown to emerge naturally from the SVD decomposition of the weight matrix. This equation is subsequently analyzed for linear RBMs in Section 4.2 in order to identify the unstable deformation modes of W which result in the first patterns that emerge at the beginning of the learning process; the non-linear regime is described in Section 4.3 based on the thermodynamic analysis, by actually numerically solving the effective learning equations in simple cases. Our analysis is finally illustrated and validated in Section 5 by actual tests on the MNIST dataset.

2 The RBM and its associated learning procedure

An RBM is a Markov random field with pairwise interactions defined on a bipartite graph formed by two layers of non-interacting variables: the visible nodes and the hidden nodes representing respectively data configurations and latent representations

(see Figure 1). The former noted $\boldsymbol{s} = \{s_i, i = 1 \dots N_v\}$ correspond to explicit representations of the data while the latter noted $\boldsymbol{\sigma} = \{\sigma_j, j = 1 \dots N_h\}$ are there to build arbitrary dependencies among the visible units. They play the role of an interacting field among visible nodes. Usually the nodes are binary-valued (of Boolean type or Bernoulli distributed) but Gaussian distributions or more broadly arbitrary distributions on real-valued bounded support are also used [26], ultimately making RBMs adapted to more heterogeneous data sets. Here to simplify we assume that visible and hidden nodes will be taken as binary variables $s_i, \sigma_j \in \{-1, 1\}$ (using ± 1 values gives the advantage of working with symmetric equations hence avoiding to deal with the "hidden" biases on the variables that appear when considering binary $\{0, 1\}$ variables). Like in the Hopfield model [8], which can actually be cast into an RBM [12], an energy function is defined for a configuration of nodes

$$E(\boldsymbol{s}, \boldsymbol{\sigma}) = -\sum_{i,j} s_i W_{ij} \sigma_j + \sum_{i=1}^{N_v} \eta_i s_i + \sum_{j=1}^{N_h} \theta_j \sigma_j \tag{1}$$

and this is exploited to define a joint distribution between visible and hidden units, namely the Boltzmann distribution

$$p(\boldsymbol{s}, \boldsymbol{\sigma}) = \frac{e^{-E(\boldsymbol{s},\boldsymbol{\sigma})}}{Z} \tag{2}$$

where W is the weight matrix and $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ are biases, or external fields on the variables. $Z = \sum_{\boldsymbol{s},\boldsymbol{\sigma}} e^{-E(\boldsymbol{s},\boldsymbol{\sigma})}$ is the partition function of the system. The joint distribution between visible variables is then obtained by summing over hidden ones. In this context, learning the parameters of the RBM means that, given a dataset of M samples composed of N_v variables, we ought to infer values to W, $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ such that new generated data obtained by sampling this distribution should be similar to the input data. The general method to infer the parameters is to maximize the log likelihood of the model, where the pdf (2) has first been summed over the hidden variables

$$\mathcal{L} = \sum_j \langle \log(2\cosh(\sum_i W_{ij} s_i - \theta_j)) \rangle_{\text{Data}} - \sum_i \eta_i \langle s_i \rangle_{\text{Data}} - \log(Z). \tag{3}$$

Different learning methods have been set up and proven to work efficiently, in particular the contrastive divergence (CD) algorithm from Hinton [4] and more recently TAP based learning [13]. They all correspond to expressing the gradient ascent on the likelihood as

$$\Delta W_{ij} = \gamma \left(\langle s_i \sigma_j p(\sigma_j | \boldsymbol{s}) \rangle_{\text{Data}} - \langle s_i \sigma_j \rangle_{p_{\text{RBM}}} \right) \tag{4}$$

$$\Delta \eta_i = \gamma \left(\langle s_i \rangle_{p_{\text{RBM}}} - \langle s_i \rangle_{\text{Data}} \right) \tag{5}$$

$$\Delta \theta_j = \gamma \left(\langle \sigma_j \rangle_{p_{\text{RBM}}} - \langle \sigma_j p(\sigma_j | \boldsymbol{s}) \rangle_{\text{Data}} \right) \tag{6}$$

where γ is the learning rate. The main problem are the $\langle \cdots \rangle_{p_{\text{RBM}}}$ terms on the right hand side of (4-6). These are not tractable and the various methods basically differ in their way of estimating those terms (Monte-Carlo Markov chains, naive mean-field, TAP...). For an efficient learning the $\langle \cdots \rangle_{\text{Data}}$ terms also have to be approximated by making use of random mini-batches of data at each step.

3 Static thermodynamical properties of an RBM

3.1 Statistical ensemble of RBMs

When analyzing the thermodynamical properties of RBMs, a common assumption which is made consists in considering i.i.d. random variables for the weights W_{ij} like for example in [20, 17, 18]. This generally leads to a Marchenko-Pastur (MP) distribution [27] of the singular values of W, which is unrealistic. In order to fix some notation let us recall in passing the singular value decomposition (SVD) definition. As a generalization of eigenmodes decomposition to rectangular matrices, the SVD for a RBM is given by

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{7}$$

where $\mathbf{U}$ is an orthogonal $N_v \times N_h$ matrix whose columns are the left singular vectors $\mathbf{u}^\alpha$, $\mathbf{V}$ is an orthogonal $N_h \times N_h$ matrix whose columns are the right singular vectors $\mathbf{v}^\alpha$ and $\mathbf{\Sigma}$ is a diagonal matrix whose elements are the singular values w_α. The separation into left and right singular vectors is due to the rectangular nature of the decomposed matrix, and the similarity with eigenmodes decomposition is revealed by the following SVD equations

$$\mathbf{W}\mathbf{v}^\alpha = w_\alpha \mathbf{u}^\alpha$$
$$\mathbf{W}^T\mathbf{u}^\alpha = w_\alpha \mathbf{v}^\alpha$$

In [25] we argue that the MP distribution of SVD modes actually corresponds to the noise of the weight matrix, while the information content of the RBM is better expressed by the presence of SVD modes outside this bulk. This led us to write the weight matrix as

$$W_{ij} = \sum_{\alpha=1}^{K} w_\alpha u_i^\alpha v_j^\alpha + r_{ij} \tag{8}$$

where the $w_\alpha = O(1)$ are isolated singular values (describing a rank K matrix), the $\boldsymbol{u}^\alpha$ and $\boldsymbol{v}^\alpha$ are the eigenvectors of the SVD decomposition and the $r_{ij} = \mathcal{N}(0, \sigma^2/L)$ where $L = \sqrt{N_h N_v}$ are i.i.d. corresponding to noise. The $\{u^\alpha\}$ and $\{v^\alpha\}$ are two sets of respectively N_v and N_h-dimensional orthonormal vectors which means they are respectively $O(1/\sqrt{N_v})$ and $O(1/\sqrt{N_h})$ and $K \leq N_v, N_h$. We assume $N_h < N_v$ to be the rank of W and $w_\alpha > 0$ and $O(1)$ for all α. Note that all together, in the limit $N_v \to \infty$ and $N_h \to \infty$ with $\kappa \overset{\text{def}}{=} N_h/N_v$ fixed and $K/L \to 0$, WW^T has a spectrum density $\rho(\lambda)$ composed of a Marchenko-Pastur bulk of eigenvalues in addition to discrete modes:

$$\rho(\lambda) = \frac{L}{2\pi\sigma^2} \frac{\sqrt{(\lambda^+ - \lambda)(\lambda - \lambda^-)}}{\kappa\lambda} \mathbb{1}_{\{\lambda \in [\lambda^-, \lambda^+]\}} + \sum_{\alpha=1}^{K} \delta(\lambda - w_\alpha^2),$$

with

$$\lambda^\pm \overset{\text{def}}{=} \sigma^2 \left(\kappa^{\frac{1}{4}} \pm \kappa^{-\frac{1}{4}}\right)^2.$$

The meaning of the noise term r_{ij} is the presence of an extensive number of modes at the bottom of the spectrum, along which the variables won't be able to condense, but still contribute to the fluctuations. In the present form our model of RBM is similar to the Hopfield model and recent generalizations [28], the patterns being represented by the SVD modes outside the bulk. The main difference, in addition to the bipartite

structure of the graph, is the non-degeneracy of the singular values w_α. The choice made here is to consider K finite, so that $W_{ij} = O(1/N)$ which means that the thresholds θ_j, which have the meaning of feature detectors should be $O(1)$ because feature j is detected when an extensive number of spin S_i are aligned with W_{ij}. In addition, it then allows us to assume simple distributions for the components of $\boldsymbol{u}^\alpha$ and $\boldsymbol{v}^\alpha$ considered i.i.d. for instance. This altogether defines our statistical ensemble of RBM to which we restrict ourselves to study the learning procedure. Another approach would be to consider $K = N_h$ extensive, thereby assuming that all modes can potentially condense even though they are associated to dominated singular values. In that case, the separation between the condensed modes and the rest should be made when introducing order parameters and the noise would then correspond to uncondensed modes. If the number of condensed modes is assumed to be extensive, then we should instead consider an average over the orthogonal group which would lead to a slightly different mean-field theory [29, 30]. We want now to explore in depth the thermodynamic properties of model (8) by making various assumptions on the statistical properties of the u_i^α and v_j^α.

3.2 Replica symmetric Mean-field equation

Our analysis in the thermodynamic limit follows classical treatments using replicas, like [31, 9] for the Hopfield model or [17] for bipartite models. The starting point is to express the average over u, v and weights r_{ij} of the log partition function Z in (2) with the help of the replica trick:

$$\mathsf{E}_{u,v,r}[\log(Z)] = \lim_{p\to 0} \frac{d}{dp}\mathsf{E}_{u,v,r}[Z^p].$$

First the average over r_{ij} yields a term

$$\exp\Big[\frac{\sigma^2}{2L}\Big(\sum_a s_i^a\sigma_j^a\Big)^2\Big] = \exp\Big[\frac{\sigma^2}{2L}\Big(p + \sum_{a\neq b} s_i^a s_i^b \sigma_j^a\sigma_j^b\Big)\Big].$$

After this averaging, 4 sets of order parameters $\{(m_\alpha^a, \bar{m}_\alpha^a), a = 1,\dots p, \alpha = 1,\dots K\}$ and $\{(Q_{ab}, \bar{Q}_{ab}), a, b = 1,\dots p, a \neq b\}$ are introduced with help of two distinct Hubbard-Stratonovich transformations. The first one corresponds to

$$\exp\Big[\frac{\sigma^2}{2L}\Big(\sum_{i,j,a\neq b} s_i^a s_i^b \sigma_j^a\sigma_j^b\Big)\Big] = \int \prod_{a\neq b} \frac{dQ_{ab}d\bar{Q}_{ab}}{2\pi}$$
$$\times \exp\Big[-\frac{L\sigma^2}{2}\sum_{a\neq b}\Big(Q_{ab}\bar{Q}_{ab} - \frac{Q_{ab}}{N_v}\sum_i s_i^a s_i^b - \frac{\bar{Q}_{ab}}{N_h}\sum_j \sigma_j^a\sigma_j^b\Big)\Big].$$

The second one is aimed at extracting magnetizations contributions correlated with the modes:

$$\exp\Big(L\sum_\alpha w_\alpha s_\alpha^a\sigma_\alpha^a\Big) \propto \int\prod_\alpha \frac{dm_\alpha^a d\bar{m}_\alpha^a}{2\pi}$$
$$\times \exp\Big(-L\sum_\alpha w_\alpha\big(m_\alpha^a\bar{m}_\alpha^a - m_\alpha^a s_\alpha^a - \bar{m}_\alpha^a\sigma_\alpha^a\big)\Big),$$

with

$$s_\alpha^a \stackrel{\text{def}}{=} \frac{1}{\sqrt{L}} \sum_i s_i u_i^\alpha \qquad \text{and} \qquad \sigma_\alpha^a \stackrel{\text{def}}{=} \frac{1}{\sqrt{L}} \sum_j \sigma_j^a v_j^\alpha, \tag{9}$$

These variables represent the following quantities:

$$m_\alpha^a \sim E_{u,v,r}\big(\langle \sigma_\alpha^a \rangle\big) \qquad \bar{m}_\alpha^a \sim E_{u,v,r}\big(\langle s_\alpha^a \rangle\big)$$

$$Q_{ab} \sim E_{u,v,r}\big(\langle \sigma_i^a \sigma_i^b \rangle\big) \qquad \bar{Q}_{ab} \sim E_{u,v,r}\big(\langle s_j^a s_j^b \rangle\big),$$

namely the correlations of the hidden [resp. visible] states with the left [resp. right] singular vectors and the Edward-Anderson (EA) order parameters measuring the correlation between replicas of hidden or visible states. E_u and E_v denote an average w.r.t. to the rescaled components $u \simeq \sqrt{N_v} u_i^\alpha$ and $v \simeq \sqrt{N_h} v_j^\alpha$ of the SVD modes. The transformations involve pairs of complex integration variables because of the asymmetry introduced by the two-layers structure by contrast to fully connected models. They lead to the following representation:

$$\mathsf{E}_{u,v,r}[Z^p] = \int \prod_{a,\alpha} \frac{dm_\alpha^a d\bar{m}_\alpha^a}{2\pi} \prod_{a \neq b} \frac{dQ_{ab} d\bar{Q}_{ab}}{2\pi}$$

$$\times \exp\Big\{ -L\Big(\sum_{a,\alpha} w_\alpha m_\alpha \bar{m}_\alpha + \frac{\sigma^2}{2} \sum_{a \neq b} Q_{ab} \bar{Q}_{ab} - \frac{1}{\sqrt{\kappa}} A[m,Q] - \sqrt{\kappa} B[\bar{m},\bar{Q}] \Big) \Big\}$$

with $\kappa = N_h/N_v$ and

$$A[m,Q] \stackrel{\text{def}}{=} \log\Big[\sum_{S^a \in \{-1,1\}} \mathsf{E}_u \Big(e^{\frac{\sqrt{\kappa}\sigma^2}{2} \sum_{a \neq b} Q_{ab} S^a S^b + \kappa^{\frac{1}{4}} \sum_{a,\alpha} (w_\alpha m_\alpha^a - \eta_\alpha) u^\alpha S^a} \Big) \Big], \tag{10}$$

$$B[\bar{m},\bar{Q}] \stackrel{\text{def}}{=} \log\Big[\sum_{S^a \in \{-1,1\}} \mathsf{E}_v \Big(e^{\frac{\sqrt{\kappa}\sigma^2}{2} \sum_{a \neq b} \bar{Q}_{ab} \sigma^a \sigma^b + \kappa^{-\frac{1}{4}} \sum_{a,\alpha} (w_\alpha \bar{m}_\alpha^a - \theta_\alpha) v^\alpha \sigma^a} \Big) \Big], \tag{11}$$

$$\tag{12}$$

with

$$\theta_\alpha \stackrel{\text{def}}{=} \frac{1}{\sqrt{L}} \sum_j \theta_j v_j^\alpha = O(1).$$

Since $\{v^\alpha\}$ is an incomplete basis we have also to take care of a potential residual transverse part $\eta^\perp$ and $\theta^\perp$ such that the following decomposition hold:

$$\eta_i = \eta_i^\perp + \sqrt{L} \sum_\alpha \eta_\alpha u_i^\alpha, \tag{13}$$

$$\theta_j = \theta_j^\perp + \sqrt{L} \sum_\alpha \theta_\alpha v_j^\alpha. \tag{14}$$

To keep things tractable, both $\eta^\perp$ and $\theta^\perp$ will be considered to be negligible in the sequel. Taking into account these components would lead to add a random field to the effective RS field of the variables and eventually to richer set of saddle point solutions. Note that the order of magnitude of θ_α is at this stage an assumption. If the θ_j where

uncorrelated from the v_j^α they would instead scale as $1/\sqrt{L}$. Moreover, regarding ensemble average we will consider fixed θ_α in the sequel.

The thermodynamic properties are obtained by first letting $L \to \infty$ allowing for a saddle point approximation and then the limit $p \to 0$ is taken. We restrict here the discussion to replica symmetric (RS) saddle points [32]. The breakdown of RS can actually be determined by computing the so-called AT line [33] (See Appendix A). At this point we assume a non-broken replica symmetry. The set $\{(Q_{ab}, \bar{Q}_{ab}\}$ reduces then to a pair $(q, \bar{q})$ of spin glass parameters, i.e. $Q_{ab} = q$ and $\bar{Q}_{ab} = \bar{q}$ for all $a \neq b$ while quenched magnetization towards the SVD directions are now represented by $\{(m_\alpha, \bar{m}_\alpha), \alpha = 1, \dots K\}$.

Taking the limit $p \to 0$ yields the following limit for the free energy:

$$f[m, \bar{m}, q, \bar{q}] = \sum_\alpha w_\alpha m_\alpha \bar{m}_\alpha - \frac{\sigma^2}{2} q\bar{q} + \frac{\sigma^2}{2}(q + \bar{q}) \\ - \frac{1}{\sqrt{\kappa}} \mathsf{E}_{u,x}\Big[\log 2\cosh\big(h(x,u)\big)\Big] - \sqrt{\kappa}\mathsf{E}_{v,x}\Big[\log 2\cosh\big(\bar{h}(x,v)\big)\Big]. \tag{15}$$

Assuming a replica-symmetric phase, the saddle-point equations are given by

$$m_\alpha = \kappa^{\frac{1}{4}} \mathsf{E}_{v,x}\Big[v^\alpha \tanh\big(\bar{h}(x,v)\big)\Big], \qquad q = \mathsf{E}_{v,x}\Big[\tanh^2\big(\bar{h}(x,v)\big)\Big] \tag{16}$$

$$\bar{m}_\alpha = \kappa^{-\frac{1}{4}} \mathsf{E}_{u,x}\Big[u^\alpha \tanh\big(h(x,u)\big)\Big], \qquad \bar{q} = \mathsf{E}_{u,x}\Big[\tanh^2\big(h(x,u)\big)\Big] \tag{17}$$

where

$$h(x,u) \stackrel{\text{def}}{=} \kappa^{\frac{1}{4}}\Big(\sigma\sqrt{q}x + \sum_\gamma (w_\gamma m_\gamma - \eta_\gamma)u^\gamma\Big)$$

$$\bar{h}(x,v) \stackrel{\text{def}}{=} \kappa^{-\frac{1}{4}}\Big(\sigma\sqrt{\bar{q}}x + \sum_\gamma (w_\gamma \bar{m}_\gamma - \theta_\gamma)v^\gamma\Big).$$

and $\kappa = N_h/N_v$, and where $\mathsf{E}_{u,x}$ and $\mathsf{E}_{v,x}$ denote an average over the Gaussian variable $x = \mathcal{N}(0,1)$ and the rescaled components $u \sim \sqrt{N_v}u_i^\alpha$ and $v \sim \sqrt{N_h}v_j^\alpha$ of the SVD modes. The equations are course symmetric under the exchange $\kappa \to \kappa^{-1}$ simultaneously with $m \leftrightarrow \bar{m}$, $q \leftrightarrow \bar{q}$ and $\eta \leftrightarrow \theta$ given that u and v have same distribution. In addition, for independent distributed u_i^α and v_j^α, when the fields vanish ($\eta = \theta = 0$) solutions corresponding to non-degenerate magnetizations have symmetric counterparts: each pair of non-vanishing magnetizations can be negated independently as $(m_\alpha, \bar{m}_\alpha) \to (-m_\alpha, -\bar{m}_\alpha)$ to generate new solutions. So if one solution is obtained with n condensed modes there actually correspond 2^n distinct solutions.

3.3 Phase Diagram

These fixed point equations can be solved numerically to tell us how the variables condensate on the SVD modes within each equilibrium state of the distribution and whether a spin-glass or a ferromagnetic phase is present or not. The important point here is that with K finite and a non-degenerate spectrum the mode with highest singular value dominates the ferromagnetic phase.

In absence of bias ($\eta = \theta = 0$) the main properties of the phase diagram may then be summarized as follows. Once $1/\sigma$ is interpreted as temperature and w_α/σ as

ferromagnetic couplings, we get a phase diagram similar to that of the Sherrington-Kirkpatrick (SK) model with three distinct phases (see Figure 2)

- a paramagnetic phase ($q = \bar{q} = m_\alpha = \bar{m}_\alpha = 0$) (P),
- a ferromagnetic phase ($q, \bar{q}, m_\alpha, \bar{m}_\alpha \neq 0$) (F),
- a spin glass phase ($q, \bar{q} \neq 0$; $m_\alpha = \bar{m}_\alpha = 0$) (SG).

The lines separating the different phases correspond to second order phase transition and can be obtained by a stability analysis of the Hessian of the free energy. They are related to unstable modes of the linearized mean-field equations and correspond to an eigenvalue of the Hessian becoming negative.

The (SG-P) line is obtained similarly by looking at the Hessian in the $(q, \bar{q})$ sector:

$$H_{q\bar{q}} \underset{\substack{m=0\\ q=0}}{=} -\frac{1}{2}\begin{bmatrix} \sigma^2 & \frac{\sigma^4}{\sqrt{\kappa}} \\ \sqrt{\kappa}\sigma^4 & \sigma^2 \end{bmatrix}$$

from what results that the spin glass phase develop when $\sigma \geq 1$[2]. This transition line should be understood directly from the spectral properties of the weight matrix. This is obtained classically [32] with help of linearized TAP equations and the Marchenko-Pastur distribution. Indeed in our context the linearized TAP equations read

$$\begin{bmatrix} \mu \\ \nu \end{bmatrix} = \begin{bmatrix} -\sqrt{\kappa}\sigma^2 & W^T \\ W & -\frac{\sigma^2}{\sqrt{\kappa}} \end{bmatrix} \begin{bmatrix} \mu \\ \nu \end{bmatrix}$$

based on the variance σ^2/L of the weights in absence of dominant modes. The paramagnetic phase becomes unstable when the rhs matrix has its highest eigenvalue equal to one. If λ is a singular value of W the corresponding eigenvalues $\Lambda^\pm$ verify

$$\big(\frac{\Lambda^\pm}{\sqrt{\kappa}} \pm \sigma^2\big)\big(\sqrt{\kappa}\Lambda^\pm \pm \sigma^2\big) = \lambda^2.$$

From that it is clear that the largest eigenvalue Λ_{max} corresponds to the largest singular value λ_{max} and owing to the Marchenko-Pastur distribution verifies

$$\big(\frac{\Lambda_{max}}{\sqrt{\kappa}} + \sigma^2\big)\big(\sqrt{\kappa}\Lambda_{max} + \sigma^2\big) = \sigma^2(\sqrt{\kappa}+1)\big(\frac{1}{\sqrt{\kappa}}+1\big).$$

$\Lambda_{max} = 1$ is readily obtained for $\sigma^2 = 1$.

For the (F-SG) frontier we can look at the sector $(m_\alpha, \bar{m}_\alpha)$ corresponding to the emergence of a single mode α:

$$H_{\alpha\alpha} = \begin{bmatrix} w_\alpha & w_\alpha^2 \mathsf{E}_{v,x}\Big[(v^\alpha)^2 \operatorname{sech}^2\big(\bar{h}(x,v)\big)\Big] \\ w_\alpha^2 \mathsf{E}_{u,x}\Big[(u^\alpha)^2 \operatorname{sech}^2\big(h(x,u)\big)\Big] & w_\alpha \end{bmatrix}$$

$$\underset{m_\alpha=0}{=} \begin{bmatrix} w_\alpha & w_\alpha^2(1-q) \\ w_\alpha^2(1-\bar{q}) & w_\alpha \end{bmatrix}$$

[2] Note that in [17] a dependence $\sqrt{\kappa(1-\kappa)}$ ($\sqrt{\alpha(1-\alpha)}$ in their notations) is found. This dependence is hidden in our definition of σ^2 giving $L = \sqrt{N_v N_h}$ times the variance of r_{ij} in our case instead of $N_v + N_h$ times in their case.

written in the spin-glass phase. From this it is clear that the first mode to become unstable is the mode α with highest singular value w_α and this occurs when q and $\bar{q}$ solution of (16,17) verify

$$(1-q)(1-\bar{q})w_\alpha^2 = 1.$$

As for the SK model, as seen on Figure 2, this line appears to be well below to AT line to be computed in the next section. Therefore a replica symmetry breaking treatment would be necessary in principle to properly separate these two phases. Being mainly interested from their practical viewpoint, namely the ability of RBM to learn arbitrary data, we are mostly concerned with the ferromagnetic phase above the AT line, so that this point will be left aside.

For the (P-F) line considering the same sector in the Hessian but now from the paramagnetic phase, i.e. setting $q = 0$ above yields the emergence of the single mode α for $w_\alpha = 1$.

Note that all this is independent of the choice of statistical average over u and v. Instead, the way of averaging influences the nature of the ferromagnetic phase as we shall see later on.

The region where the RS solution is stable can also be computed by determining the so-called Anderson-Thouless (AT) line. Details of the computations can be found in Appendix. A. It is similar to the classical computation made for the SK model, though slightly more involved. In fact we were not able to fully characterize all the possible instabilities of the Hessian in replica space which would potentially lead to a breakdown of the replica symmetry. At least the one responsible for the ordinary SK model RS breakdown has a counter part in the bipartite case which reads:

$$\frac{1}{\sigma^2} > \sqrt{\mathsf{E}_{x,u}\Big(\mathrm{sech}^4\big(h(x,u)\big)\Big)\mathsf{E}_{x,v}\Big(\mathrm{sech}^4\big(\bar{h}(x,v)\big)\Big)},$$

as therefore a necessary condition for the stability of the RS solution. For $\kappa = 1$ terms below the radical become identical and the condition reduces to the same form as for the SK model, except for the u averages not present in the SK model. As seen on the Figure 2 the influence of κ and on the type of average made on u and v.

3.4 Nature of the Ferromagnetic phase

Some subtleties arise when considering various ways of averaging over singular vectors components. In [19, 20] is underlined the importance of the capability of networks to produce compositional states structured by combination of hidden variables. In our representation, we don't have a direct access to this property, but to the dual one in some sense, namely states corresponding to combination of modes. Their presence and their structure, are rather sensitive to the way the average over u and v is performed. In this respect the case where $\boldsymbol{u}^\alpha$ and $\boldsymbol{v}^\alpha$ are Gaussian i.i.d distributed is very special: all fixed points associated to dominated modes can be shown to be unstable and fixed points associated to combinations of modes are not allowed. To see this, first notice that the magnetization part of the saddle point equations (16,17) read in that case

$$m_\alpha = (w_\alpha \bar{m}_\alpha - \theta_\alpha)(1-q) \tag{18}$$

$$\bar{m}_\alpha = (w_\alpha m_\alpha - \eta_\alpha)(1-\bar{q}). \tag{19}$$

Since the role of the bias is mainly to introduce some asymmetry between otherwise degenerated fixed points obtained by sign reversal of at least one pair $(m_\alpha, \bar{m}_\alpha)$, let

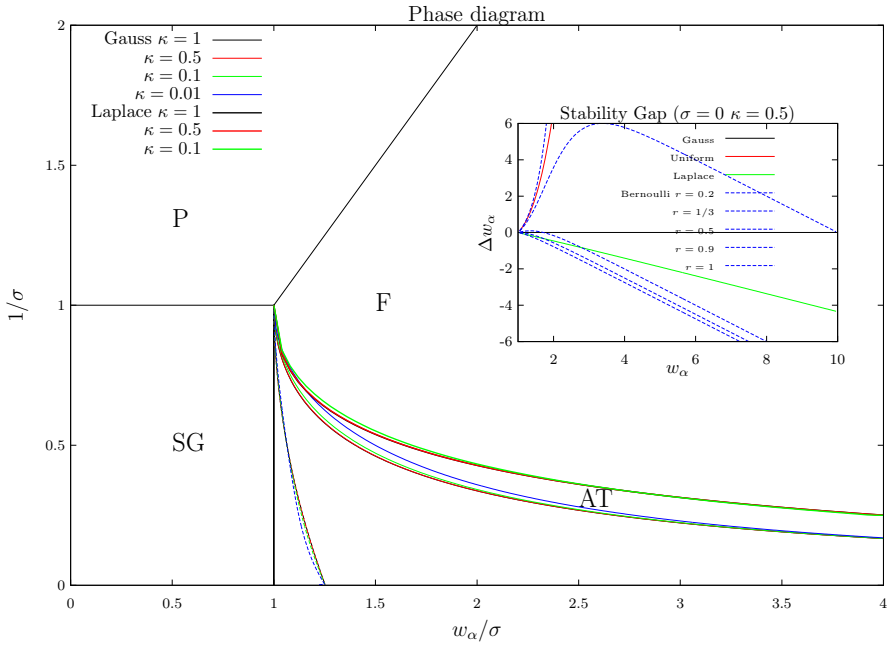

Fig. 2: Phase diagram in absence of bias and with a finite number of modes, with Gaussian and Laplace distributions for u and v. The dotted line separates the spin glass phase from the ferromagnetic phase under the replica symmetry hypothesis. The RS phase is unstable below the AT line. The influence of κ on the AT and (SG-F) lines is shown. In all cases the hypothetical SG-F line lies well inside the broken RS phase. Inset: high temperature stability gap at defined as Δw_α as a function of w_α for a fixed point associated to a mode β, corresponding to various distributions.

us analyze the situation without fields i.e. by setting $\eta = \theta = 0$. In such case we immediately see that as long as the singular values are non degenerate, only one single mode may condense at the same time. Indeed if the mode α condenses we have necessarily

$$w_\alpha^2(1-q)(1-\bar{q}) = 1,$$

which can be verified only by one mode at the time. Then looking at the stability of these we see as shown below that only the fixed point associated to the largest singular value is actually stable.

Instead, for other distributions like uniform Bernoulli or Laplace for instance, stable fixed points associated to many different single modes or combinations of modes can exist and contribute to the thermodynamics. In order to analyze this question in more general terms we first rewrite the mean-field equations in a convenient way which require some preliminary remarks. We restrict the discussion to i.i.d variables so that we may consider single variable distributions. Joint distributions will be distinguished

from single variable distribution by use of bold argument $\mathbf{u} = \{u^\alpha, \alpha = 1, \dots, K\}$, K being the (finite) number of modes susceptible of condensing.

Given the distribution p assumed to be even we introduce the following distribution:

$$p^\star(u) \stackrel{\text{def}}{=} -\int_{-\infty}^{u} xp(x)dx = \int_{|u|}^{\infty} xp(x)dx, \tag{20}$$

attached to mode α. This distribution has the following properties:

Lemma 3.1. *Given that p is centered of unit variance and kurtosis κ_u, $p^\star$ is a centered probability distribution with variance*

$$\int_{-\infty}^{\infty} u^2 p^\star(u)du = \frac{\kappa_u}{3}.$$

Proof. In order to show this consider the moments of $p^\star$. For any n odd they vanish while for n even they read:

$$\begin{aligned}
\int_{-\infty}^{+\infty} u^n p^\star(u)du &= 2\int_0^{\infty} u^n p^\star(u)du \\
&= 2\int_0^{\infty} duu^n \int_u^{\infty} xp(x)dx \\
&= 2\int_0^{\infty} xp(x)dx \int_0^{x} u^n du \\
&= \frac{1}{n+1}\int_{-\infty}^{\infty} x^{n+2} p(x)dx,
\end{aligned}$$

i.e. relate to moments of order $n+2$ of p. The property then follows from the fact that p has unit variance. ∎

In this respect, the Gaussian averaging is special because we have $\kappa_u = 3$ and $p^\star = p$ in that case. Then the mean-field equations (16,17) corresponding to the magnetization can be rewritten in a similar form as the Gaussian averaging case (18,19) by introducing the variable q_α and $\bar{q}_\alpha$:

$$m_\alpha = (w_\alpha \bar{m}_\alpha - \theta_\alpha)(1 - q_\alpha), \tag{21}$$

$$\bar{m}_\alpha = (w_\alpha m_\alpha - \eta_\alpha)(1 - \bar{q}_\alpha), \tag{22}$$

with

$$q_\alpha = \int dx \frac{e^{-x^2/2}}{\sqrt{2\pi}} d\mathbf{v} p_\alpha(\mathbf{v}) \tanh^2\Big(\kappa^{-\frac{1}{4}}\Big(\sigma\sqrt{\bar{q}}x + \sum_\gamma (w_\gamma \bar{m}_\gamma - \theta_\gamma)v^\gamma\Big)\Big), \tag{23}$$

$$\bar{q}_\alpha = \int dx \frac{e^{-x^2/2}}{\sqrt{2\pi}} d\mathbf{u} p_\alpha(\mathbf{u}) \tanh^2\Big(\kappa^{\frac{1}{4}}\Big(\sigma\sqrt{q}x + \sum_\gamma (w_\gamma m_\gamma - \eta_\gamma)u^\gamma\Big)\Big), \tag{24}$$

where

$$p_\alpha(\mathbf{u}) \stackrel{\text{def}}{=} p^\star(u^\alpha) \prod_{\beta \neq \alpha} p(u^\beta).$$

This rewriting will prove very useful also in the next section when analyzing the learning dynamics. Let us now assume, in absence of bias, a non-degenerate fixed point associated to some given mode β ($w_\beta < w_\alpha$) with finite $(m_\beta, \bar{m}_\beta)$ and $m_\alpha = \bar{m}_\alpha = 0, \forall \alpha \neq \beta$. The fixed point equation imposes the relation

$$w_\beta = \frac{1}{\sqrt{(1-q_\beta)(1-\bar{q}_\beta)}} \stackrel{\text{def}}{=} w(q_\beta, \bar{q}_\beta). \tag{25}$$

The stability of such a fixed point with respect to any other mode α is related to the positive definiteness of the following block part of the Hessian

$$H_{\alpha\alpha} = \begin{bmatrix} w_\alpha & w_\alpha^2 \mathsf{E}_{v,x}\Big[(v^\alpha)^2 \operatorname{sech}^2\big(\bar{h}(x,v)\big)\Big] \\ w_\alpha^2 \mathsf{E}_{u,x}\Big[(u^\alpha)^2 \operatorname{sech}^2\big(h(x,u)\big)\Big] & w_\alpha \end{bmatrix}$$

with

$$h(x,u) = \kappa^{\frac{1}{4}}\big(\sigma\sqrt{q}x + w_\beta \bar{m}_\beta u^\beta\big) \qquad \text{and} \qquad \bar{h}(x,v) = \kappa^{-\frac{1}{4}}\big(\sigma\sqrt{\bar{q}}x + w_\beta \bar{m}_\beta v^\beta\big),$$

in the present case. This in fact reduces to

$$H_{\alpha\alpha} = \begin{bmatrix} w_\alpha & w_\alpha^2(1-q) \\ w_\alpha^2(1-\bar{q}) & w_\alpha \end{bmatrix}.$$

Therefore for the Gaussian average case, since in that case $q_\beta = q$ and $\bar{q}_\beta = \bar{q}$ we necessarily have from (25)

$$1 - (1-q)(1-\bar{q})w_\alpha^2 = 1 - \frac{w_\alpha^2}{w_\beta^2} < 0 \qquad \text{for} \qquad w_\alpha > w_\beta,$$

i.e. the Hessian has negative eigenvalues. This means that if the mode β is dominated by some other mode α, the magnetization $(m_\alpha, \bar{m}_\alpha)$ will develop until $(1-q)(1-\bar{q})w_\alpha^2 = 1$, while m_β will vanish.

For the general case of i.i.d. variables, assuming u^α and v^α obey the same distribution p, let F and F_α the cumulative distributions associated respectively to p and p_α

$$F(u) \stackrel{\text{def}}{=} \int_{-\infty}^{u} p(x)dx$$

$$F_\alpha(u) \stackrel{\text{def}}{=} \int d\mathbf{u}\ \theta(u-u^\alpha)p_\alpha(\mathbf{u})dx = -\int_{\infty}^{u} du^\alpha \int_{-\infty}^{u^\alpha} xp(x)dx.$$

Given the values of $(q, \bar{q})$ obtained from the fixed point associated to mode β we have the following property:

Proposition 3.2. *If*

$$(i)\ F_\beta(u) < F(u), \qquad \forall u \in \mathbb{R}^+ \qquad \textit{then} \qquad q_\beta > q \qquad \textit{and} \qquad \bar{q}_\beta > \bar{q},$$

$$(ii)\ F_\beta(u) > F(u), \qquad \forall u \in \mathbb{R}^+ \qquad \textit{then} \qquad q_\beta < q \qquad \textit{and} \qquad \bar{q}_\beta < \bar{q},$$

which in turn implies

$$w(q,\bar{q}) < w_\beta \ (i) \qquad and \qquad w(q,\bar{q}) > w_\beta \ (ii)$$

with

$$w(q,\bar{q}) \stackrel{\text{def}}{=} \frac{1}{\sqrt{(1-q)(1-\bar{q})}}.$$

Proof. This is obtained by straightforward by parts integration over u and v respectively in equations (16,17) relative to magnetizations. ■

In other words if F_β dominates F on $\mathbb{R}^+$ then there is a positive stability gap defined as

$$\Delta w_\beta \stackrel{\text{def}}{=} w(q,\bar{q}) - w_\beta, \tag{26}$$

such that there is a non-empty range for higher values of $w_\alpha \in [w_\beta, w(q,\bar{q})[$ for which the fixed point associated to mode β corresponds to a local minimum of the free energy. Note that property (i) [resp. (ii)] goes in the same way (in the sense that it implies it) as p_β having a larger [resp. smaller] variance than p i.e. $\kappa_u > 3$ [resp. $\kappa_u < 3$]. Therefore distributions p with negative relative kurtosis ($\kappa_u - 3$) will tend to favor the presence of metastable states, while the situation will tend to be more complex for probabilities with positive relative kurtosis. Indeed in that case the fixed point associated to the highest mode α_{max} might not correspond to a stable state if lower modes are present in the range $[w(q,\bar{q}), w_{\alpha_{max}}[$ and fixed points associated to combinations of modes have to be considered. Note that in contrary to the Gaussian case, this can be achieve because the q_α are different for each mode and therefore more flexibility is offered by equations (21,22) than equations (18,19).

Let us give some examples. Denote by $\gamma_u \stackrel{\text{def}}{=} \kappa_u - 3$ the relative kurtosis. As already said the Gaussian distribution is a special case with $\gamma_u = 0$. In addition, for instance for p corresponding to Bernoulli, Uniform or Laplace we have the following properties illustrated on inset of Figure 2:

- Bernoulli ($\gamma_u = -2$):

$$p(u) = \frac{1}{2}\big(\delta(u+1) + \delta(u-1)\big), \qquad F(u) = \frac{1}{2}\big(\theta(u+1) + \theta(u-1)\big)$$
$$p_\alpha(u) = \frac{1}{2}\theta(1-u^2), \qquad F_\alpha(u) = \frac{1}{2}\theta(1-u^2)(u+1) + \theta(u-1)$$

  so that $F_\alpha(u) > F(u)$ for $u > 0$ yielding a positive stability gap.

- Uniform ($\gamma_u = -6/5$):

$$p(u) = \frac{1}{2\sqrt{3}}\theta(3-u^2), \qquad F(u) = \frac{1}{2\sqrt{3}}\theta(3-u^2)(u+\sqrt{3}) + \theta(u-\sqrt{3})$$
$$p_\alpha(u) = \frac{1}{4\sqrt{3}}\theta(3-u^2)(3-u^2), \qquad F_\alpha(u) = \frac{1}{4\sqrt{3}}\theta(3-u^2)(3u - \frac{u^3}{3} + 2\sqrt{3}) + \theta(u-\sqrt{3}).$$

  One can then verify that $F_\alpha(u) > F(u)$ for $u > 0$ yielding as well a positive stability gap.

- Laplace ($\gamma_u = 3$):

$$p(u) = \frac{1}{\sqrt{2}}e^{-\sqrt{2}|u|}, \qquad F(u) = \frac{1}{2} + \frac{u}{2|u|}\big(1 - e^{-\sqrt{2}|u|}\big)$$
$$p_\alpha(u) = \frac{1}{2}\big(|u| + \frac{1}{\sqrt{2}}\big)e^{-\sqrt{2}|u|}, \qquad F_\alpha(u) = F(u) - \frac{u}{2\sqrt{2}}e^{-\sqrt{2}|u|}.$$

Here we see instead that $F_\alpha(u) < F(u)$ for $u > 0$ yielding a negative stability gap.

These three examples fall either in condition (i) or (ii). In such cases the stability gap Δw_β is either always positive or always negative, independently of w_β. We can also provide examples for which the stability condition may vary with w_β. Consider for instance a sparse Bernoulli distribution, with $r \in [0,1]$ some sparsity parameter:

$$p(u) = \frac{r}{2}\big(\delta(u + \frac{1}{\sqrt{r}}) + \delta(u - \frac{1}{\sqrt{r}})\big) + (1-r)\delta(u).$$

The relative kurtosis reads in that case

$$\gamma_u(r) = \frac{1}{r} - 3.$$

Looking at $F(u)$ and $F_\alpha(u)$ it is seen that neither of conditions (i) or (ii) are fulfilled except for $r = 1$ which corresponds to the plain Bernoulli case. As we see on the inset of Figure 2 for $r < 1/3$, the stability gap is always negative, meaning that a single mode ferromagnetic phase is not stable, and is replaced by a compositional ferromagnetic phase at all temperature. Instead for $r > 1/3$ at sufficiently high temperature (low w_α)) the single mode fixed point dominate the ferromagnetic phase.

Laplace distribution: let us look at the properties of the phase diagram in this case where a negative stability gap is expected which may lead to a compositional phase. For this we need the expression for a sum of Laplace variables to compute the averages involved in (16,17). To this purpose we define the following distributions:

$$f(s) = \int \prod_\gamma du^\gamma \frac{\lambda_\gamma}{2} e^{-\lambda_\gamma |u^\gamma|}\ \delta(s - \sum_\gamma u^\gamma),$$

$$g_\alpha(s) = \int du^\alpha \frac{\lambda_\alpha}{4}(\lambda_\alpha |u^\alpha| + 1) e^{-\lambda_\alpha |u^\alpha|} \prod_{\gamma \neq \alpha} du^\gamma \frac{\lambda_\gamma}{2} e^{-\lambda_\gamma |u^\gamma|}\ \delta(s - \sum_\gamma u^\gamma).$$

Their Laplace transform upon decomposing into partial fractions read:

$$\tilde{f}(\omega) = \prod_\gamma \frac{\lambda_\gamma^2}{\lambda_\gamma^2 - \omega^2} = \sum_\gamma C_\gamma \frac{\lambda_\gamma^2}{\lambda_\gamma^2 - \omega^2}$$

and

$$\begin{aligned}\tilde{g}_\alpha(\omega) &= \frac{\lambda_\alpha^2}{\lambda_\alpha^2 - \omega^2} \prod_\gamma \frac{\lambda_\gamma^2}{\lambda_\gamma^2 - \omega^2} \\ &= C_\alpha \frac{\lambda_\alpha^4}{(\lambda_\alpha^2 - \omega^2)^2} + \sum_{\gamma \neq \alpha} C_\gamma \frac{\lambda_\gamma^2 \lambda_\alpha^2}{\lambda_\alpha^2 - \lambda_\gamma^2}\Big(\frac{1}{\lambda_\gamma^2 - \omega^2} - \frac{1}{\lambda_\alpha^2 - \omega^2}\Big).\end{aligned}$$

where

$$C_\gamma \stackrel{\text{def}}{=} \prod_{\delta \neq \gamma} \frac{\lambda_\delta^2}{\lambda_\delta^2 - \lambda_\gamma^2}.$$

From these decomposition we immediately identify

$$f(s) = \frac{1}{2}\sum_{\gamma} C_\gamma \lambda_\gamma e^{-\lambda_\gamma |s|},$$

$$g_\alpha(s) = \frac{\lambda_\alpha C_\alpha}{4}(\lambda_\alpha |s| + 1)e^{-\lambda_\alpha |s|} + \frac{1}{2}\sum_{\gamma\neq\alpha} C_\gamma \frac{\lambda_\gamma \lambda_\alpha}{\lambda_\alpha^2 - \lambda_\gamma^2}\big(\lambda_\alpha e^{-\lambda_\gamma |s|} - \lambda_\gamma e^{-\lambda_\alpha |s|}\big).$$

This then results in the following decomposition of the EA parameters:

$$q = \int dxds \frac{e^{-\sqrt{2}|s| - x^2/2}}{2\sqrt{\pi}} \sum_{\gamma} C_\gamma[\bar{m}] \tanh^2\big(\bar{h}_\gamma(x,s)\big) \tag{27}$$

$$q_\alpha = \int dxds \frac{e^{-\sqrt{2}|s| - x^2/2}}{2\sqrt{\pi}} \Big[\frac{1}{\sqrt{2}}(|s| + \frac{1}{\sqrt{2}}) C_\alpha[\bar{m}] \tanh^2\big(\bar{h}_\alpha(x,s)\big) \tag{28}$$

$$+ \sum_{\gamma\neq\alpha} C_\gamma[\bar{m}] \frac{(w_\gamma \bar{m}_\gamma - \theta_\gamma)^2 \tanh^2\big(\bar{h}_\gamma(x,s)\big) - (w_\alpha \bar{m}_\alpha - \theta_\alpha)^2 \tanh^2\big(\bar{h}_\alpha(x,s)\big)}{(w_\gamma \bar{m}_\gamma - \theta_\gamma)^2 - (w_\alpha \bar{m}_\alpha - \theta_\alpha)^2}\Big] \tag{29}$$

with

$$\bar{h}_\gamma(x,s) \stackrel{\text{def}}{=} \kappa^{-\frac{1}{4}}\big(\sigma\sqrt{\bar{q}}x + (w_\gamma \bar{m}_\gamma - \theta_\gamma)s\big)$$

and

$$C_\gamma[\bar{m}] \stackrel{\text{def}}{=} \prod_{\delta\neq\gamma} \frac{(w_\gamma \bar{m}_\gamma - \theta_\gamma)^2}{(w_\gamma \bar{m}_\gamma - \theta_\gamma)^2 - (w_\delta \bar{m}_\delta - \theta_\delta)^2}.$$

This allows for an efficient resolution of the mean-field equation (16,17,21,22). By doing so we are able to observe the dawning of a purely compositional phase in the ferromagnetic domain, when the modes at the top of the spectrum get close enough. In order to characterize it we consider the stability gap $\Delta^{(n)}(w_\alpha)$ giving the range $[w_a - \Delta^{(n)}(w_\alpha), w_a]$ below the highest mode w_a such that the ferromagnetic corresponds to the condensation of n distinct modes present on this interval, including the highest.

In addition this will prove useful when analyzing the learning dynamics described in the next section.

4 Dynamics of Learning an RBM

4.1 The learning dynamics in the thermodynamic limit

In [25] we propose a mean field analysis of the learning dynamics, in the form of a phenomenological equation obtained after averaging over some parameters of the RBM, i.e. by choosing a well defined statistical ensemble of RBMs and using self-averaging properties in the thermodynamic limit. In order to keep the paper self-consistent we recall how this equation is obtained with additional details and then explore its properties in the light of the preceding section. First we project the gradient ascent equation (4-6) onto the basis $\{u_\alpha(t) \in \mathbb{R}^{N_v}\}$ and $\{v_\alpha(t) \in \mathbb{R}^{N_h}\}$ defined by the SVD of W. Discarding stochastic fluctuations usually inherent to the learning procedure

and letting the learning rate $\gamma \to 0$, the continuous version of (4-6) can be recast as follows:

$$\frac{1}{L}\Big(\frac{dW}{dt}\Big)_{\alpha\beta} = \langle s_\alpha \sigma_\beta \rangle_{\text{Data}} - \langle s_\alpha \sigma_\beta \rangle_{\text{RBM}}, \tag{30}$$

$$\frac{1}{\sqrt{L}}\Big(\frac{d\eta}{dt}\Big)_{\alpha} = \langle s_\alpha \rangle_{\text{RBM}} - \langle s_\alpha \rangle_{\text{Data}}, \tag{31}$$

$$\frac{1}{\sqrt{L}}\Big(\frac{d\theta}{dt}\Big)_{\alpha} = \langle \sigma_\alpha \rangle_{\text{RBM}} - \langle \sigma_\alpha \rangle_{\text{Data}}, \tag{32}$$

with s_α and σ_α given already in (9). We also have

$$\Big(\frac{dW}{dt}\Big)_{\alpha\beta} = \delta_{\alpha,\beta}\frac{dw_\alpha}{dt} + (1-\delta_{\alpha,\beta})\Big(w_\beta(t)\Omega^v_{\beta\alpha}(t) + w_\alpha(t)\Omega^h_{\alpha\beta}\Big)$$

$$\frac{1}{\sqrt{L}}\Big(\frac{d\eta}{dt}\Big)_{\alpha} = \frac{d\eta_\alpha}{dt} - \sum_\beta \Omega^v_{\alpha\beta}\eta_\beta$$

$$\frac{1}{\sqrt{L}}\Big(\frac{d\theta}{dt}\Big)_{\alpha} = \frac{d\theta_\alpha}{dt} - \sum_\beta \Omega^h_{\alpha\beta}\theta_\beta$$

where

$$\Omega^v_{\alpha\beta}(t) = -\Omega^v_{\beta\alpha} \stackrel{\text{def}}{=} \frac{d\boldsymbol{u}^{\alpha,T}}{dt}\boldsymbol{u}^\beta$$
$$\Omega^h_{\alpha\beta}(t) = -\Omega^h_{\beta\alpha} \stackrel{\text{def}}{=} \frac{d\boldsymbol{v}^{\alpha,T}}{dt}\boldsymbol{v}^\beta$$

By eliminating $\big(\frac{dw}{dt}\big)_{\alpha\beta}$, $\big(\frac{d\eta}{dt}\big)_{\alpha}$ and $\big(\frac{d\theta}{dt}\big)_{\alpha}$ we get the following set of dynamical equations:

$$\frac{1}{L}\frac{dw_\alpha}{dt} = \langle s_\alpha \sigma_\alpha \rangle_{\text{Data}} - \langle s_\alpha \sigma_\alpha \rangle_{\text{RBM}} \tag{33}$$

$$\frac{d\eta_\alpha}{dt} = \langle s_\alpha \rangle_{\text{RBM}} - \langle s_\alpha \rangle_{\text{Data}} + \sum_\beta \Omega^v_{\alpha\beta}\eta_\beta \tag{34}$$

$$\frac{d\theta_\alpha}{dt} = \langle \sigma_\alpha \rangle_{\text{RBM}} - \langle \sigma_\alpha \rangle_{\text{Data}} + \sum_\beta \Omega^h_{\alpha\beta}\theta_\beta \tag{35}$$

along with the infinitesimal rotation generators of the left and right singular vectors

$$\Omega^v_{\alpha\beta}(t) = -\frac{1}{w_\alpha + w_\beta}\Big(\frac{dW}{dt}\Big)^{\text{A}}_{\alpha\beta} + \frac{1}{w_\alpha - w_\beta}\Big(\frac{dW}{dt}\Big)^{\text{S}}_{\alpha\beta} \tag{36}$$

$$\Omega^h_{\alpha\beta}(t) = \frac{1}{w_\alpha + w_\beta}\Big(\frac{dW}{dt}\Big)^{\text{A}}_{\alpha\beta} + \frac{1}{w_\alpha - w_\beta}\Big(\frac{dW}{dt}\Big)^{\text{S}}_{\alpha\beta} \tag{37}$$

where

$$\Big(\frac{dW}{dt}\Big)^{\text{A,S}}_{\alpha\beta} \stackrel{\text{def}}{=} \frac{1}{2}\Big(\langle s_\alpha \sigma_\beta \rangle_{\text{Data}} \pm \langle s_\beta \sigma_\alpha \rangle_{\text{Data}} \mp \langle s_\beta \sigma_\alpha \rangle_{\text{RBM}} - \langle s_\alpha \sigma_\beta \rangle_{\text{RBM}}\Big).$$

The dynamics of learning is now expressed in the reference frame defined by the singular vectors of W. The skew-symmetric rotation generators $\Omega^{v,h}_{\alpha\beta}(t)$ of the basis vectors induced by the dynamics tell us how the data rotate relatively to this frame. Given an initial condition these help us to keep track of data representation in this frame. Note that these equations become singular when some degeneracy occurs in W because then the SVD is not uniquely defined. Except from the numerical point of view, where some regularization might be needed, this does not constitute an issue. In fact only rotations among non-degenerate modes are meaningful, while the rest corresponds to gauge degrees of freedom.

At this point our set of dynamical equations (33-37) are written for a general form of an RBM. Our goal is to find from these equation the typical trajectory of an RBM within some statistical ensemble. For this we make the hypothesis that the learning dynamics is represented by a trajectory of $\{w_\alpha(t), \eta_\alpha(t), \theta_\alpha(t), \Omega^{v,h}_{\alpha\beta}(t)\}$, while the specific realization of the u_i^α and v_j^α and r_{ij} in (8) can be considered to be irrelevant up to the way they are distributed. We allow then ourselves to perform some averaging of the u_i^α and v_j^α and r_{ij} with respect to some simple distributions, as long as this average is correlated with the data. This means that the components s_α of any given sample configuration is assumed also to be kept fixed while averaging. What matters mainly is the strength given by $w_\alpha(t)$ and the rotation given by $\Omega^{v,h}_{\alpha\beta}(t)$ of the SVD modes. As a simplification and also by lack of understanding on what intrinsically drives their evolution, the distributions of u_i^α and v_j^α will be considered stationary in the sequel. Concerning r_{ij} we allow its variance σ^2/L to vary with time in order to describe in a minimal way how the MP bulk evolves during the learning. The dynamics of σ will be later specifically derived in Section 4.3. With the notations

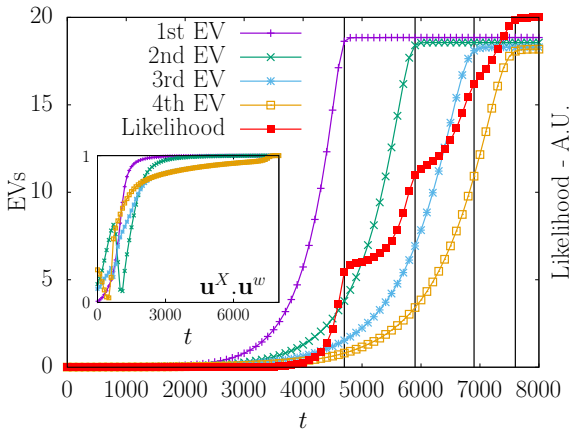

Fig. 3: Time evolution of the eigenvalues in the linear model and of the likelihood. We observe very clearly how the different modes emerge from the bulk and how the likelihood increases at each eigenvalue learned. In the inset, the scalar product of the vectors $\boldsymbol{u}$ obtained from the SVD of the data and of $\boldsymbol{w}$. The $\boldsymbol{u}$s of $\boldsymbol{w}$ are aligned with the SVD of the data at the end of the learning.

of Section 3.4 and in particular using the rescaling $v \sim \sqrt{N_h} v_i^\alpha$, the empirical terms

take the form:

$$\langle \sigma_\alpha \rangle_{\text{Data}} = \langle (s_\alpha w_\alpha - \theta_\alpha)\big(1 - q_\alpha[\mathbf{s}]\big)\rangle_{\text{Data}} \tag{38}$$

$$\langle s_\alpha \sigma_\beta \rangle_{\text{Data}} = \langle s_\alpha (s_\beta w_\beta - \theta_\beta)\big(1 - q_\beta[\mathbf{s}]\big)\rangle_{\text{Data}} \tag{39}$$

where

$$q_\alpha[\mathbf{s}] \stackrel{\text{def}}{=} \int dx \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} d\mathbf{v} p_\alpha(\mathbf{v}) \tanh^2\Big(\kappa^{-\frac{1}{4}}\big(\sigma x + \sum_\gamma (w_\gamma s_\gamma - \theta_\gamma) v^\gamma\big)\Big),$$

which actually depends on the activation function (an hyperbolic tangent in this case). The term σx correspond to $\sum_k r_{kj} s_k$ and is obtained by central limit theorem from the independence of the r_{kj}. $q_\alpha[\mathbf{s}]$ is the empirical counterpart to the EA parameters q and q_α already encountered in Section 3.4. It can be easily estimated for simple i.i.d. distributions like Gaussian or Laplace. The main point here is that the empirical terms (38,39) defines operators whose decomposition onto the SVD modes of W functionally depends solely on w_α, θ_α and on the projection of the data on the SVD modes of W. These terms are precisely driving the dynamics. The adaptation of the RBM to this driving force are given by RBM terms in (33,34,35). Those can be as well estimated in the thermodynamic limit (see Section 4.3), as a function of w_α, θ_α and η_α alone, by means of the order parameters $(m_\alpha, \bar{m}_\alpha)$ given in Section 3.2, once the mean-field equations (16,17) have been solved. This of course is based on the hypothesis that the RBM stays in the RS domain during learning. Experimental evidences are going to support this hypothesis later on.

4.2 Linear instabilities

When starting the learning, the weight matrix W is usually quite small and therefore it is tempting to analyze the linear behavior of the RBM in order to understand what happened at the beginning. In particular, we will see that the dynamics of a non-linear RBM at the beginning of the learning can be understood by looking at a linear stability analysis of the learning process. The purpose of this analysis is to identify which "deformation modes" of the weight matrix are the most unstable, and how they relate to the input data. Additionally, the good point with the linear case is that no averaging is needed, the dynamics being actually independent of the particular realization of the u_i^α and v_j^β. In addition no distinction has to be made between dominant modes and others one which should be then treated approximately as the noise component in (8). Instead we may treat them all on the same footing in the linear case.

So let us see how the linear regime is obtained from an RBM with binary units. It can be obtained by rescaling all the weights and fields by a common "inverse temperature" β factor and let this go to zero in equations (4). In principle the stability analysis would lead to assume both the weights and magnetizations to be small. In fact we can performed the analysis without approximation in a slightly more general case, by assuming only the magnetizations to be small. This is then equivalent to consider our RBM to be a linear one where magnetization undergo Gaussian fluctuations.

This limit is obtained by keeping up to quadratic terms for magnetizations in the

mean field free energy:

$$
\begin{aligned}
F_{MF}(\mu,\nu) \simeq & \frac{1}{2}\sum_{i=1}^{N}(1+\mu_i)\log(1+\mu_i)+(1-\mu_i)\log(1-\mu_i) \\
& + \frac{1}{2}\sum_{j=1}^{M}(1+\nu_j)\log(1+\nu_j)+(1-\nu_j)\log(1-\nu_j) \\
& - \sum_{i,j}\big(W_{ij}\mu_i\nu_j - \frac{1}{2}W_{ij}^2(\mu_i^2+\nu_j^2)\big) + \sum_{i=1}^{N}\eta_i\mu_i + \sum_{j=1}^{M}\theta_j\nu_j \\
& = \frac{1}{2\sigma_v^2}\sum_{i=1}^{N}\mu_i^2 + \frac{1}{2\sigma_h^2}\sum_{j=1}^{M}\nu_i^2 - \sum_{ij}W_{ij}\mu_i\nu_j + \sum_{i=1}^{N}\eta_i\mu_i + \sum_{j=1}^{M}\theta_j\nu_j.
\end{aligned}
$$

where the variance (σ_v^2, σ_h^2) of respectively visible and hidden variables read $(N_h < N_v)$:

$$\sigma_v^{-2} = 1 + \sum_j W_{ij}^2 \simeq 1 + \sum_\alpha w_\alpha^2 \tag{40}$$

$$\sigma_h^{-2} = 1 + \sum_i W_{ij}^2 = 1 + \sum_\alpha w_\alpha^2. \tag{41}$$

The quadratic term in W_{ij} which comes from the TAP contribution to the free energy is optional for our stability analysis. In absence of this term the modes evolve strictly independently, while taking into account this term leads to a correction to individual variances which couples all the modes together.

Magnetizations (μ, ν) of visible and hidden variables have now Gaussian fluctuations with covariance matrix

$$C(\mu_v, \mu_h) \stackrel{\text{def}}{=} \begin{bmatrix} \sigma_v^{-2} & -W \\ -W^T & \sigma_h^{-2} \end{bmatrix}^{-1}$$

In the linear regime the biases of the data and related fields $(\theta_\alpha, \eta_\alpha)$ can be simply discarded par proper centering of the variables. So we just consider equation (33) which now involves directly the covariance matrix of the data expressed in the frame defined by the SVD modes of W

$$\langle s_\alpha \sigma_\beta \rangle_{\text{Data}} = \sigma_h^2 w_\beta \langle s_\alpha s_\beta \rangle_{\text{Data}}.$$

From $C(\mu_v, \mu_h)$ we get the other terms yielding the following equations:

$$
\begin{aligned}
\frac{dw_\alpha}{dt} &= w_\alpha \sigma_h^2 \Big(\langle s_\alpha^2 \rangle_{\text{Data}} - \frac{\sigma_v^2}{1 - \sigma_v^2 \sigma_h^2 w_\alpha^2} \Big) \\
\Omega_{\alpha\beta}^{v,h} &= (1-\delta_{\alpha\beta})\sigma_h^2 \Big(\frac{w_\beta - w_\alpha}{w_\alpha + w_\beta} \mp \frac{w_\beta + w_\alpha}{w_\alpha - w_\beta} \Big) \langle s_\alpha s_\beta \rangle_{\text{Data}}
\end{aligned}
$$

Note that these equations are exact for a linear RBM, since they can be derived without any reference to the coordinates of u_α and v_α over which we average in the non-linear regime. These equations tell us that, during the learning the vectors $\boldsymbol{u}^\alpha$ (and also $\boldsymbol{v}^\alpha$)

will rotate until being aligned to the the principal components of the data, i.e. until $\langle s_\alpha s_\beta \rangle_{\text{Data}}$ becomes diagonal. Then calling $\hat{w}_\alpha^2$ the corresponding empirical variance given by the data, the system reach the following equilibrium values:

$$w_\alpha^2 = \begin{cases} \dfrac{\hat{w}_\alpha^2 - \sigma_v^2}{\sigma_v^2 \sigma_h^2 \hat{w}_\alpha^2} & \text{if} \quad \hat{w}_\alpha^2 > \sigma_v^2, \\ 0 & \text{if} \quad \hat{w}_\alpha^2 \leq \sigma_v^2. \end{cases}$$

assuming fixed (σ_v, σ_h) for the moment. From this we see that the RBM selects the strongest SVD modes in the data. The linear instabilities correspond to directions along whose the variance of the data is above the threshold σ_v^2. This determines the unstable deformations modes of the weight matrix which can develop during the learning and will eventually interact, following the usual mechanism of non-linear pattern formation encountered for instance in reaction-diffusion processes [34]. Other possible deformations are damped to zero. The linear RBM will therefore learn all (up to N_h) principal components that passed the threshold. Note that this selection mechanism is already known to occur for linear auto-encoders [23] or some other similar linear Boltzmann machines [22]. On Fig. 3 we can see the eigenvalues being learned one by one in a linear RBM.

If we take into account the dependence (40,41) of (σ_v, σ_h) then the system cannot reach a stable solution except if all modes are below threshold at the beginning. Otherwise modes which are excited first grow eventually like $\sqrt{t}$ for large time, and the excitation threshold tends to zero for all modes. In any cases this minimal non-linear analysis describe a unimodal distribution by definition of the multivariate Gaussian. So in order to perform the analysis for non-linear RBM, a well suited mean-field theory is required to understand the dynamics and the steady-state regime.

4.3 Non-linear regime

During the linear regime some specific modes are selected and at some point these modes start to interact in a non-trivial manner. The empirical terms in (4-6) involve higher order statistics of the data as explicitly seen in (39) and the Gaussian estimation with $\sigma_v^2 = \sigma_h^2 = 1$ of the RBM response terms $\langle s_\alpha \rangle_{\text{RBM}}$ and $\langle s_\alpha \sigma_\beta \rangle_{\text{RBM}}$ is no longer valid. Schematically the linear regime is valid as long as the state of the RBM is in the paramagnetic phase. But as soon as one mode passes the linear threshold, the system enters the ferromagnetic phase. Then the proper estimation of the response terms follows from the thermodynamic analysis performed in Section 3. It depends on the assumption which is made on the statistical properties of singular vectors components of the weight matrix. Suppose we assume these to be Gaussian i.i.d. for instance. From the analysis proposed in Section 3.4 of the ferromagnetic phase, this leads to the fact that the mode with highest singular value dominates completely this phase: we expect one single ferromagnetic state, characterized by magnetizations aligned with this mode only. Magnetizations correlated to other modes vanish. At least this is the correct picture without fields ($\eta = \theta = 0$). With non-vanishing fields, we don't expect this picture to be changed drastically. In fact, solving the mean field equations in presence of such fields show in some cases the appearance of meta-stable states, correlated with single dominated modes. Still the free energy difference with the ground state, i.e. correlated with the mode corresponding to the highest eigenvalue is of the order $O\big(L(w_\alpha - w_{max})\big)$, so that their contribution become rapidly negligible with large systems size.

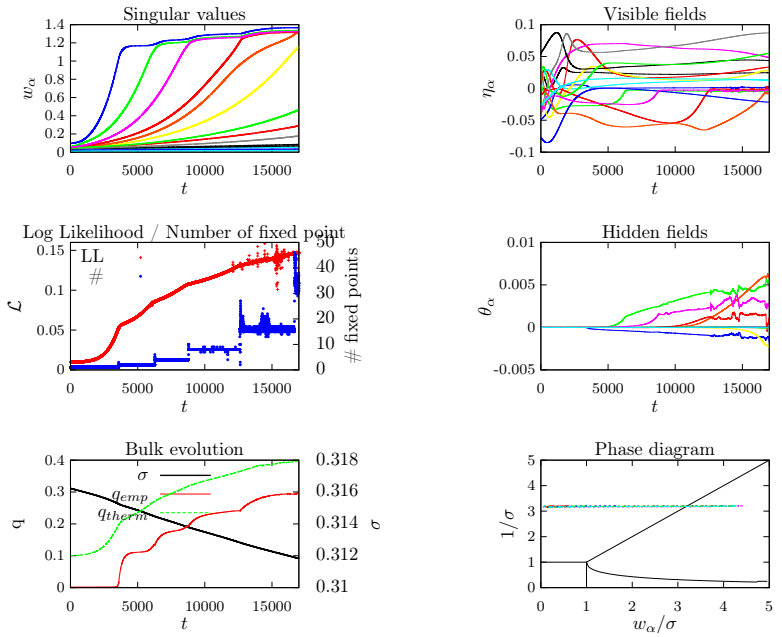

Fig. 4: Predicted mean Evolution of an RBM of size $(N_v, N_h) = (1000, 500)$ learned on a synthetic data set of 10^4 samples of size $N_v = 1000$ obtained from a multi modal distribution with 20 clusters randomly defined on a submanifold of dimension $d = 15$. The dynamics follows the projected magnetization in this reduced space with help of 15 modes. We observe a kind of pressure on top singular values from lower ones.

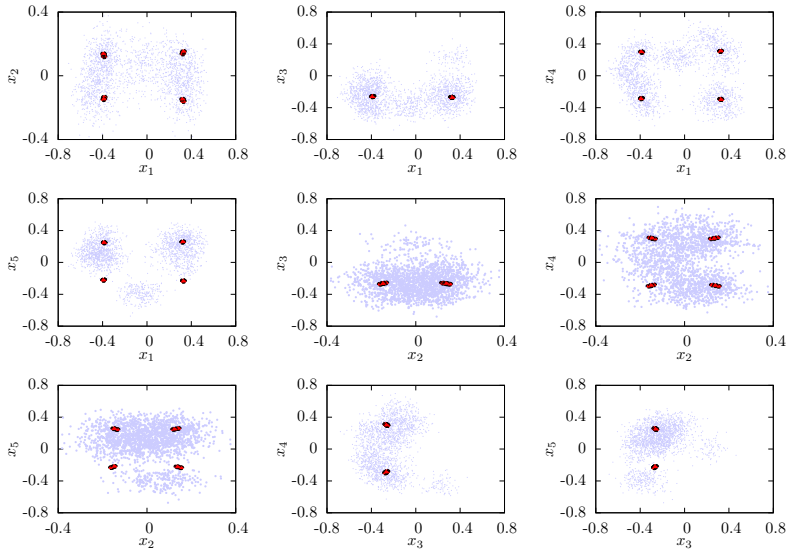

Fig. 5: Scatter plots of the mean-field magnetizations (in red) and the samples (in blue) in various plan projection defined by pairs of left eigenvectors of W. This case corresponds to an RBM of size $(N_v, N_h) = (100, 50)$ learned on a synthetic data set of 10^4 samples of size $N_v = 100$ obtained from a multi modal distribution with 11 clusters randomly defined on a submanifold of dimension $d = 5$. The scatter plot is obtained at a point where 5 modes have already condensed and 16 saddle point solutions have been found.

In order to have a realistic picture of the learning process, we consider instead a Laplace distribution for the SVD modes components which, as seen in Section 3.4 allows the ferromagnetic phase to be a compositional phase. The reason for this is that the Laplace distribution leads to less interference among modes than with a Gaussian distribution, so that they interact weakly in the mean-field equations. Solving the equations (21,22,27,29) in absence of fields yields instead the following picture: one fixed point solution will typically have non-vanishing magnetization $\{m_\alpha, \bar{m}_\alpha\}$ for all α such that $w_\alpha \in [w_{max} - \Delta w, w_{max}]$, where Δw is approximately the gap $\Delta w(q, \bar{q})$ define in (26). This solution is a degenerate ground state, all other solutions being obtained by symmetry thanks to simple independent reversing of the signs of the condensed magnetizations $(m_\alpha, \bar{m}_\alpha)$. Hence for K condensed modes we get a degeneracy of 2^K. When the fields are included, all these fixed points are moved in the direction of the fields, or depending on the strength of field some of them may disappear. At the end remains a potentially large amount of nearly degenerate states able at least in simple cases to cover in some way the empirical distribution.

Coming back to the learning dynamics, the first thing which is expected, already from the linear analysis, is that the noise term in (8) vanishes by condensing into a delta function of zero modes. Then the terms corresponding to the response of

the RBM in (4,6) are estimated in the thermodynamic limit by means of the order parameters defined previously:

$$\langle s_\alpha \rangle_{\text{RBM}} = \frac{1}{Z_{\text{Therm}}} \sum_\omega e^{-Lf(m^\omega, \bar{m}^\omega, q^\omega, \bar{q}^\omega)} \bar{m}_\alpha^\omega \stackrel{\text{def}}{=} \langle \bar{m}_\alpha \rangle_{\text{Therm}},$$

$$\langle s_\alpha \sigma_\beta \rangle_{\text{RBM}} = \frac{1}{Z_{\text{Therm}}} \sum_\omega e^{-Lf(m^\omega, \bar{m}^\omega, q^\omega, \bar{q}^\omega)} \bar{m}_\alpha^\omega m_\beta^\omega \stackrel{\text{def}}{=} \langle \bar{m}_\alpha m_\beta \rangle_{\text{Therm}}.$$

We have introduce the notation $\langle\rangle_{\text{Therm}}$ to denote the thermodynamical average and the partition function is expressed as

$$Z_{\text{Therm}} \stackrel{\text{def}}{=} \sum_\omega e^{-Lf(m^\omega, \bar{m}^\omega, q^\omega, \bar{q}^\omega)}$$

in this limit. The index ω runs over all stable fixed point solutions of (16,17) weighted accordingly to the free energy given by (15). These are the dominant contributions as long as free energy differences are $O(1)$, internal fluctuations given by each fixed point are comparatively of order $O(1/L)$. In addition, the dynamics of the bulk can be characterized by the evolution of σ^2 defined empirically from the weights as

$$\sigma^2 = \frac{1}{L} \sum_{ij} r_{ij}^2,$$

we have

$$\frac{d\sigma^2}{dt} = \frac{1}{L} \sum_{ij} r_{ij} \frac{dW_{ij}}{dt},$$

$$= \frac{1}{L} \sum_{ij} r_{ij} \left[\langle s_i \tanh\Big(\sum_k r_{kj} s_k + \kappa^{-\frac{1}{4}} \sum_\alpha (w_\alpha s_\alpha - \theta_\alpha) v_j^\alpha \sqrt{L} \Big) \rangle_{\text{Data}} - \langle s_i \sigma_j \rangle_{\text{RBM}} \right]$$

by independence of r_{i*} and r_{*j} with u_i^α and v_i^α respectively. Using self averaging properties of both empirical and response terms with respect to r_{ij}, u_i^α and v_j^α yields

$$\frac{1}{L^2} \sum_{ij} r_{ij} \langle s_i \sigma_j \rangle_{\text{Data}} = \frac{\sigma^2}{L} \big(1 - \langle q[\mathbf{s}] \rangle_{\text{Data}} \big)$$

$$\frac{1}{L^2} \sum_{ij} r_{ij} \langle s_i \sigma_j \rangle_{\text{RBM}} = \frac{\sigma^2}{L} \big(1 - \langle q \rangle_{\text{Therm}} \big),$$

with

$$q[\mathbf{s}] \stackrel{\text{def}}{=} \int dx \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} d\mathbf{v} p(\mathbf{v}) \tanh^2\Big(\kappa^{-\frac{1}{4}} \big(\sigma x + \sum_\gamma (w_\gamma s_\gamma - \theta_\gamma) v^\gamma \big) \Big).$$

In the end if we summarize, our equations take the suggestive form

$$\frac{1}{L}\frac{dw_\alpha}{dt} = \langle s_\alpha(w_\alpha s_\alpha - \theta_\alpha)(1-q_\alpha[\mathbf{s}])\rangle_{\text{Data}} - \langle \bar{m}_\alpha(w_\alpha \bar{m}_\alpha - \theta_\alpha)(1-q_\alpha)\rangle_{\text{Therm}}, \tag{42}$$

$$\frac{d\eta_\alpha}{dt} = \langle \bar{m}_\alpha\rangle_{\text{Therm}} - \langle s_\alpha\rangle_{\text{Data}} + \sum_\beta \Omega^v_{\alpha\beta}\eta_\beta, \tag{43}$$

$$\frac{d\theta_\alpha}{dt} = \langle (w_\alpha \bar{m}_\alpha - \theta_\alpha)(1-q_\alpha)\rangle_{\text{Therm}} - \langle (w_\alpha s_\alpha - \theta_\alpha)(1-q_\alpha[\mathbf{s}])\rangle_{\text{Data}} + \sum_\beta \Omega^h_{\alpha\beta}\theta_\beta, \tag{44}$$

$$\frac{d\sigma^2}{dt} = \sigma^2\Big(\langle q\rangle_{\text{Therm}} - \langle q[\mathbf{s}]\rangle_{\text{Data}}\Big), \tag{45}$$

with $\Omega^{v,h}$ taking as well the form of a difference between a data averaging $\langle\rangle_{\text{Data}}$ and a thermodynamical averaging $\langle\rangle_{\text{Therm}}$ involving only order parameters. Note here that w_α are faster variables than other ones, thye evolve at different time scales. This is the final and main result of this paper which possibly might help improving learning algorithms of RBM or maybe more complex models like deep Boltzmann machines (DBM). From this it is now clear what the learning of an RBM is aimed at. These equations have converged once the data set is clustered in such a way that each cluster is represented by a solution of the mean-field equations with magnetizations and EA parameters corresponding to their empirical counterparts. In particular, these clusters can somehow be seen as the attractors in the context of feed-forward networks, yielding in a similar way a partition of the data. This can be seen by starting from random configurations and let the system evolves (by using the TAP equations or a MCMC). At the end the system will end up in one of those clusters (characterized by a fixed point of the mean-field equations). Note that this is the reason why the RBM needs to reach a ferromagnetic phase with many states to be able to match the empirical term in (4) in order to converge. In addition the log likelihood (3) can also be estimated in the thermodynamic limit

$$\mathcal{L} = \Big\langle \sqrt{\kappa}\mathsf{E}_{x,v}\Big[\log\cosh\Big(\kappa^{-\frac{1}{4}}\Big(\sigma x + \sum_\alpha (w_\alpha s_\alpha - \theta_\alpha)v^\alpha\Big)\Big)\Big]\Big\rangle_{\text{Data}}$$

$$- \Big\langle \sum_\alpha \eta_\alpha s_\alpha\Big\rangle_{\text{Data}} - \frac{1}{L}\log\Big(Z_{\text{Therm}}\Big),$$

(after normalization by L). For instance, in the case of a multimodal data distribution with a finite number of clusters embedded in a high dimensional configuration space, the SVD modes of W which will develop are the one pointing in the almost surely orthogonal direction of the magnetizations defined by these clusters. In this simple case the RBM will evolve as in the linear case to a state such that the empirical term becomes diagonal, while the singular values adjust themselves until matching the proper magnetization in each fixed point.

We have integrated equations (42,43,44,45,36,37) in simple cases by using the Laplace averaging of the SVD modes components, based on the (27,29) expression of the EA parameters. Basically the hidden distribution to be modeled is defined by

$$P(s) = \sum_{c=1}^{C} p_c \prod_{i=1}^{N} \frac{e^{h_i^c s_i}}{2\cosh(h_i^c)}, \tag{46}$$

i.e. a multimodal distribution composed of C clusters, of independent variables, where the magnetization of each variable i in cluster c is given by $m_i^c = \tanh(h_i^c)$. Each cluster is weighted by some probability p_c. In addition we assume these magnetizations vectors m^c to be embedded in a low dimensional space of dimension $d << N$. d defines the rank of W. The initial condition for W is such that the left singular vectors $\{u_\alpha, \alpha = 1, \dots d\}$ span this low dimensional space. An example of the typical dynamics obtained in such cases is shown on Figure 4. By contrast with the linear problem where singular values evolve independently, we distinctively witness the interaction between singular values: a kind of pressure is exerted by lower modes on higher ones resulting in successive bumps in the dynamics of the top modes. The number of states is roughly multiplied by two each time a mode condense and get close enough from the top modes. Concerning the dynamic of the fields, we don't really observe convergence toward stable directions. Some (possibly numerical) instability is observed when many modes are condensed, between the fields and the number of fixed point solutions which become very noisy at some point. It is also interesting to see how the magnetization given by each state are distributed with respect to the dataset. On Figure 5 we see that the fixed points tend (as expected) to position themselves within dense regions of sample points. Our coarse description shows however some limitations for more complex situations, the number of adjustable parameters being too limited to be able to match arbitrary distribution of clusters. This behaviour is therefore to be taken in a mean sense. It is able to reproduce at least a realistic learning dynamics of the singular values of the weight matrix.

5 Numerical Experiments

Given the comprehensive theoretical analysis of the RBM model given in the previous sections, we are now able to provide a meaningful description of the learning dynamics for a RBM trained with k-steps contrastive divergence (CDk) [4]. The observations presented in this section will serve as a validation for the theoretical analysis. First, to provide a more direct comparison to section 4.3, we will look at the learning dynamics of an RBM trained on a set of simple synthetic data. Subsequently, we will test the model against real world data by training on the MNIST dataset.

5.1 Synthetic dataset

As a simple case, we trained the RBM over the same dataset defined in fig. 4, derived from the simple multimodal distribution in eq. 47 (see Appendix B for details). Thus we set $N_v = 1000$, $N_h = 500$ and we trained using 10^4 samples with an effective dimension $d = 15$ organized in 20 separate clusters. The weights are initialized from a Gaussian distribution with standard deviation $\sigma = 10^{-3}$, while the hidden bias is initialized to 0 and the visible bias is initialized with the empirical mean of the data

$$\eta_i = \frac{1}{2} \log \left(\frac{p_i}{1 - p_i} \right)$$

where p_i is the empirical probability of activation for the i_{th} hidden node.

Finally, the training set is divided into batches of size 20, 5 Gibbs sampling steps are used (CD5) and the learning rate γ is kept low in order to reduce noise, $\gamma = 5 \times 10^{-8}$. The results of the analysis are shown in fig. 6. We see that the dynamics of the singular values obtained by direct integration of the mean-field equations (Fig. 4) are very well

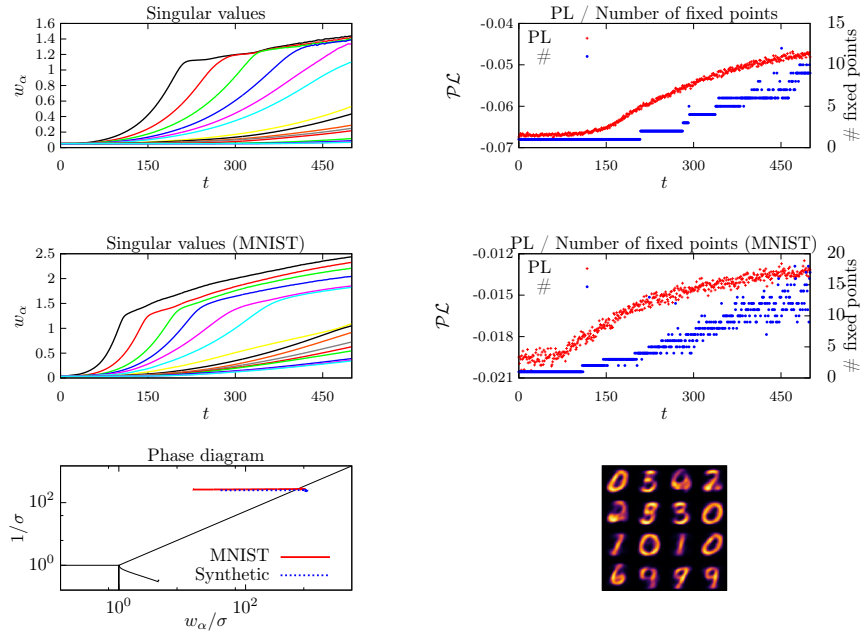

Fig. 6: Experimental evolution of an RBM during training for a synthetic dataset (top plots, to compare to Fig. 4) and for MNIST (central plots). The bottom left plot shows the learning trajectories in the phase diagram, while the bottom right image shows some examples of fixed point solutions for MNIST (we note the presence of some spurious fixed points).

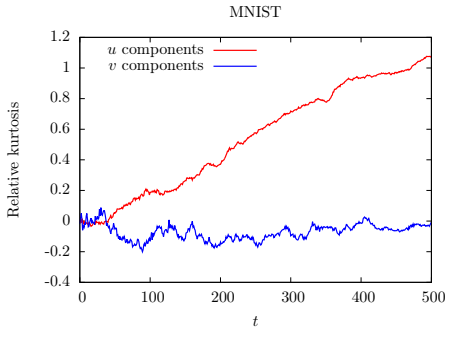

Fig. 7: Relative kurtosis for mode components trained on MNIST.

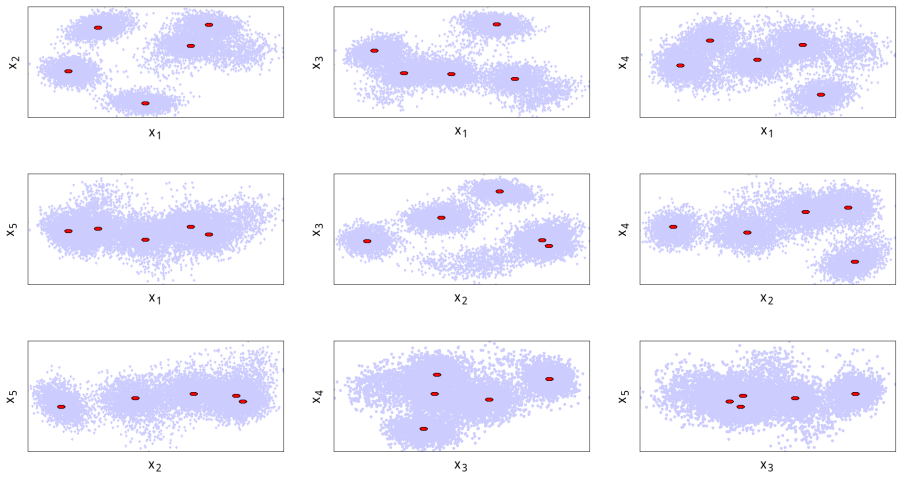

Fig. 8: Scatter plots of samples (blue) and fixed points (red) in various plan projections defined by pairs of left eigenvectors of W. The dataset is the same as in Fig. 5 and in this case 5 modes have condensed and 7 fixed points solutions have been found.

reproduced, the only difference being a slightly higher pressure on the strongest modes. The number of fixed point solutions also seems to follow the same trend but more noise is present, an indication of the fact that the RBM has a tendency to learn spurious fixed points during the training. The learning trajectory on the phase diagram is also of interest; we see that the RBM is initialized in the paramagnetic state as expected and the effect of the learning is to drive the model to the ferromagnetic phase. Once in the ferromagnetic phase, the trajectory slows down and the model is assessed near the critical line between paramagnetic and ferromagnetic states, where the estimate of the weights is most stable (according to [35]). Finally, in Fig. 8 we see how the RBM is able to generate a proper clustering of the data over the spectral modes. In particular, the TAP fixed points of the trained model are well distributed and able to cover the full data distribution, improving over the typical behaviour for Laplace distributed weights that emerged with our theoretical analysis (Fig. 5).

5.2 MNIST dataset

The MNIST dataset is composed by 70000 handwritten digits (60000 for training, 10000 for testing) of size 28×28 pixels. Being this dataset highly multimodal, we expect it to push the limits of our spectral analysis. For the training, the initialization of the model is the same one used for the synthetic data, 10000 training samples are used (taken at random from the dataset) and the values of the other hyperparameters are as follows: $N_v = 784$, $N_h = 100$, batch size $= 20$, $\gamma = 5 \times 10^{-7}$. With respect to the linear regime (described in section 4.2) we see in Fig. 9 how the RBM is able to learn the SVD of the dataset quite precisely at the beginning of the training, then the learning dynamics quickly enter the non-linear regime. Even in this highly

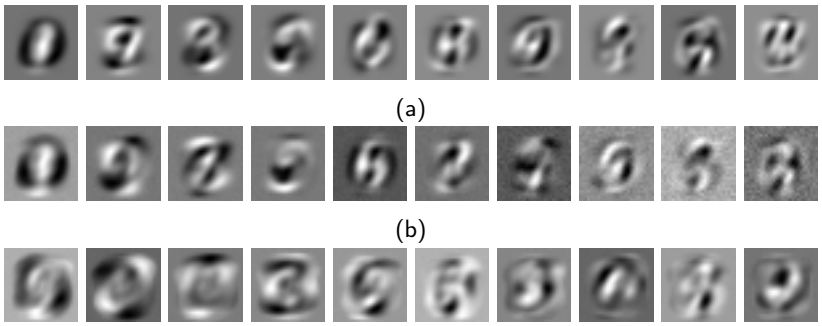

Fig. 9: **(a)** Principal components extracted from the training set (starting from the second, as the first one is encoded into the visible bias). **(b)** The first 10 modes of a RBM trained for 1 epoch (with $\gamma \simeq 0.1$). **(c)** Same as (b) but after a 10 epochs training.

multimodal scenario, our findings over simple synthetic data seem to be confirmed, as seen in Fig. 6. The high number of modes, however, determines an increase in the magnitude of the singular values of condensed modes and seems to destabilize a bit the learning, making the computation of fixed points less reliable. In fact, as a high number of modes are condensing, the model is not able to get rid of all the spurious fixed points. This problem can be mitigated by using an even smaller learning rate, at the cost of slowing down the training. Probably, using a variable learning rate could be a more practical solution (descreasing the learning rate from time to time to let the model eliminate unneeded fixed points). Concerning the (relative) kurtosis of the mode components distributions, we did not observe a very stable and systematic behavior. Either we see small fluctuations around zero, either some excursions occur and a finite value in the range $[0, 3]$ is building up either for the u or the v components, coherently to the compositional phase interpretation given previously. The latter is the case for MNIST, as shown in Fig. 7. Additionally the transverse part of the fields, meaning orthogonal to the condensed modes, is usually not completely negligible, in contrary to what we assume in (13,14). This clearly constitutes a limitation of our analysis. These transverse components offer more flexibility for generating and selecting fixed points and interfere in some non-trivial way with the kurtosis property, which possibly explains why we don't get a systematic behavior.

6 Discussion

Before drawing some perspectives, let us first summarize the main outcomes of the present work:

- **(i) thermodynamic properties of realistic RBM** by considering a non-i.i.d. ensemble of weight matrices based on empirical observations obtained by training RBM on real data.
- **(ii) RS equations and compositional phase**: we found in particular with equations (21,22,23,24) a way of writing the RS equations for the RBM, which

leads to a simple characterization of the ferromagnetic phase where the RBM is assumed to operate. Schematically, negative kurtosis of the singular vector components distribution favours the proliferation of meta stable states while positive kurtosis tend to favour a compositional phase. In particular we were able to address precisely a concrete case of compositional phase by considering a Laplace distribution of the singular vectors components.

- **(iii) a set of equations representing a typical learning dynamics** by a trajectory of $\{w_\alpha(t), \eta_\alpha(t), \theta_\alpha(t), \Omega^{v,h}_{\alpha\beta}(t), \sigma^2(t)\}$. The dominant singular values spectrum represented by $\{w_\alpha(t)\}$ expressing the information content of the RBM is playing the main role. The bulk of dominated modes corresponding to noise sees its dynamics summarized by the evolution of $\sigma^2(t)$. Rotations of dominant singular vectors during the learning process are given by $\Omega^{v,h}$ while projections of the bias along the main modes are given by η and θ. These equations have been obtained by averaging over the components of left and right SVD vectors of the weight matrix, keeping fixed a certain number of quantities considered to be the relevant ones, fully characterizing a typical RBM during the learning process. This averaging corresponds actually to a standard self-averaging assumption in a RS phase.
- **(iv) a clustering interpretation of the training process** is obtained through equations (42,43,44,45) where is explicitly seen the kind of matching the RBM is trying to perform between order parameters obtained from fixed point solutions and empirical counterparts of these order parameters in the non-linear regime. A natural clustering of the data can actually be defined by assigning to each sample data the fixed point obtained after initializing the fixed point equation with a visible configuration corresponding to that given sample.

The main picture which emerges from the present analysis is that of a set of clusters corresponding to the fixed points of the RBM, which try to uniformly cover the support of the dataset. A full understanding of the mechanism by which the RBM manages to cover properly the dataset is still lacking, even though the study case of the Laplace distribution for the singular vectors components gives some insight. By comparison real RBM have more flexibility than such a simple "mean Laplace RBM" considered in Section 3.4 to produce such a covering of the data manifold. We were not able yet to pinpoint precisely the main ingredient for that mechanism, even though we suspect transverse biases (orthogonal to the modes) of the hidden units to be the missing ingredient of our analysis.

From the theoretical point of view we would like to see how these results can be adapted to more complex models like DBM or generative models based on convolutional networks. In particular we would like to understand whether adding more layers can facilitate the covering of the dataset by fixed points. From the practical point of view these results might help to orientate the choice of hyper-parameters that are made when training an RBM and refine the criteria for assessing the quality of a learned RBM. For instance, the choice of the number of hidden variables is dictated by two considerations, the effective rank of W i.e. the number of relevant modes to be considered and the level of interaction between these modes. Less hidden variables goes into the direction of having more compact RBM and reducing the rank of W to its needed value, but leads to modes with stronger interactions, which means less flexibility for generating a good covering of fixed points.

A AT line

The stability of the RS solution to the mean field equation is studied along the lines of [33] by looking at the Hessian of the replicated version of the free energy and identifying eigenmodes from symmetry arguments. Before taking the limit $p \to 0$ the free energy reads

$$f[m,\bar{m},Q,\bar{Q}] = \sum_{a,\alpha} w_\alpha m_\alpha^a \bar{m}_\alpha^a + \frac{\sigma^2}{2}\sum_{a\neq b} Q_{ab}\bar{Q}_{ab} - \frac{1}{\sqrt{\kappa}} A_p[m,Q] - \sqrt{\kappa} B_p[\bar{m},\bar{Q}],$$

with A_p and B_p given in (10,11). Assuming the small perturbation

$$m_\alpha^a = m_\alpha + \epsilon_\alpha^a \qquad \bar{m}_\alpha^a = \bar{m}_\alpha + \bar{\epsilon}_\alpha^a$$
$$Q_{ab} = q + \eta_{ab} \qquad \bar{Q}_{ab} = \bar{q} + \bar{\eta}_{ab},$$

around the saddle point $(m_\alpha, \bar{m}_\alpha, q, \bar{q})$, the perturbed free energy reads

$$\begin{aligned}
\Delta f = &\sum_{a,\alpha} w_\alpha \bar{\epsilon}_\alpha^a \epsilon_\alpha^a + \frac{\sigma^2}{2}\sum_{a\neq b}\bar{\eta}_{ab}\eta_{ab} + \sum_{a,b,\alpha,\beta}\Big[\big(\delta_{ab}\bar{A}_{\alpha\beta} + \bar{\delta}_{ab}\bar{B}_{\alpha\beta}\big)\epsilon_\alpha^a\epsilon_\beta^b + CT\Big] \\
&+ \sum_{a\neq b,c,\alpha}\Big[\big((\delta_{ab}+\delta_{ac})\bar{C}_\alpha + (1-\delta_{ac}-\delta_{bc})\bar{D}_\alpha\big)\epsilon_\alpha^c\eta_{ab} + CT\Big] \\
&+ \sum_{a\neq b,c\neq d}\Big[\big(\delta_{(ab)(cd)}\bar{E}_0 + \mathbb{1}_{\{a\in(cd)\oplus b\in(cd)\}}\bar{E}_1 + \mathbb{1}_{\{(ab)\cap(cd)=\emptyset\}}\bar{E}_2\big)\eta_{ab}\eta_{cd} + CT\Big],
\end{aligned}$$

where CT means "conjugate term" in the sense $\epsilon \leftrightarrow \bar{\epsilon}$, $A_{\alpha\beta} \leftrightarrow \bar{A}_{\alpha\beta}\ldots$, where $\bar{\delta}_{ab} \stackrel{\text{def}}{=} 1-\delta_{ab}$ and the operators are given by

$$A_{\alpha\beta} \stackrel{\text{def}}{=} (\delta_{\alpha\beta} - m_\alpha m_\beta)w_\alpha w_\beta \qquad B_{\alpha\beta} \stackrel{\text{def}}{=} \Big(\mathsf{E}_{x,v}\big(v^\alpha v^\beta \tanh^2(\bar{h}(x,v))\big) - m_\alpha m_\beta\Big)w_\alpha w_\beta$$

$$C_\alpha \stackrel{\text{def}}{=} \frac{\kappa^{1/4}\sigma^2}{2} m_\alpha(1-q)w_\alpha \qquad D_\alpha \stackrel{\text{def}}{=} \frac{\kappa^{1/4}\sigma^2}{2}\Big(\mathsf{E}_{x,v}\big(v^\alpha \tanh^3(\bar{h}(x,v))\big) - m_\alpha q\Big)w_\alpha$$

$$E_0 \stackrel{\text{def}}{=} \frac{\sqrt{\kappa}\sigma^4}{4}(1-q^2) \qquad E_1 \stackrel{\text{def}}{=} \frac{\sqrt{\kappa}\sigma^4}{4}q(1-q) \qquad E_2 \stackrel{\text{def}}{=} \frac{\sqrt{\kappa}\sigma^4}{4}\Big(\mathsf{E}_{x,v}\big(\tanh^4(\bar{h}(x,v))\big) - q^2\Big)$$

with

$$h(x,u) \stackrel{\text{def}}{=} \kappa^{1/4}\Big(\sqrt{q}\sigma x + \sum_\alpha (m_\alpha w_\alpha - \eta_\alpha)u^\alpha\Big),$$

and conjugate quantities are obtained by replacing m_α by $\bar{m}_\alpha$, q by $\bar{q}$, u^α by v^α and η_α by θ_α and κ by $1/\kappa$. As for the SK model, the Hessian of dimension $2Kp\times 2Kp$ thereby defined can be diagonalized with help of three similar set of eigenmodes corresponding to different permutation symmetry in replica space.

The first family corresponds to $2K+2$ replica symmetric modes defined by $\eta_\alpha^a = \eta_\alpha$

and $\eta_{ab} = \eta$ solving the linear system

$$\big(\frac{w_\alpha}{2} - \lambda\big)\bar\epsilon_\alpha - \frac{1}{2}\bar A_{\alpha\alpha}\epsilon_\alpha + \sum_\beta \big(\bar A_{\alpha\beta} + (p-1)\bar B_{\alpha\beta}\big)\epsilon_\beta + \big((p-1)\bar C_\alpha + \frac{(p-1)(p-2)}{2}\bar D_\alpha\big)\eta = 0$$

$$\big(\frac{w_\alpha}{2} - \lambda\big)\epsilon_\alpha - \frac{1}{2}A_{\alpha\alpha}\bar\epsilon_\alpha + \sum_\beta \big(A_{\alpha\beta} + (p-1)B_{\alpha\beta}\big)\bar\epsilon_\beta + \big((p-1)C_\alpha + \frac{(p-1)(p-2)}{2}D_\alpha\big)\bar\eta = 0$$

$$\big(\frac{\sigma^2}{2} - \lambda\big)\bar\eta + \sum_\alpha \big(\bar C_\alpha + \frac{p-2}{2}\bar D_\alpha\big)\epsilon_\alpha + 2\big(\bar E_0 + 2(p-2)\bar E_1 + \frac{(p-2)(p-3)}{2}\bar E_2\big)\eta = 0$$

$$\big(\frac{\sigma^2}{2} - \lambda\big)\eta + \sum_\alpha \big(C_\alpha + \frac{p-2}{2}D_\alpha\big)\bar\epsilon_\alpha + 2\big(E_0 + 2(p-2)E_1 + \frac{(p-2)(p-3)}{2}E_2\big)\bar\eta = 0$$

with eigenvalue λ solving a polynomial equation of degree $2K+2$ corresponding to a vanishing determinant of the above system.

The second family corresponds to a broken replica symmetry where one replica a_0 is different from the others

$$(\epsilon^a_\alpha, \bar\epsilon^a_\alpha) = \begin{cases} (\epsilon_\alpha, \bar\epsilon_\alpha) & \text{for } a \neq a_0 \\ (1-p)(\epsilon_\alpha, \bar\epsilon_\alpha) & \text{for } a = a_0 \end{cases} \qquad (\eta_{ab}, \bar\eta_{ab}) = \begin{cases} (\eta, \bar\eta) & \text{for } a, b \neq a_0 \\ (1-\frac{p}{2})(\eta, \bar\eta) & \text{for } a = a_0 \text{ or } b = a_0 \end{cases}$$

This family correspond to a set of dimension $(2K+2)(p-1)$. Its parameterization is obtained by imposing orthogonality with the previous one. The corresponding system reads

$$\big(\frac{w_\alpha}{2} - \lambda\big)\bar\epsilon_\alpha - \frac{1}{2}\bar A_{\alpha\alpha}\epsilon_\alpha + \sum_\beta (\bar A_{\alpha\beta} - \bar B_{\alpha\beta})\epsilon_\beta + \frac{p-2}{2}\big(\bar C_\alpha - \bar D_\alpha\big)\eta = 0$$

$$\big(\frac{w_\alpha}{2} - \lambda\big)\epsilon_\alpha - \frac{1}{2}A_{\alpha\alpha}\bar\epsilon_\alpha + \sum_\beta (A_{\alpha\beta} - B_{\alpha\beta})\bar\epsilon_\beta + \frac{p-2}{2}\big(C_\alpha - D_\alpha\big)\bar\eta = 0$$

$$\big(\frac{\sigma^2}{2} - \lambda\big)\bar\eta + \sum_\alpha (\bar C_\alpha - \bar D_\alpha)\epsilon_\alpha + 2\big(\bar E_0 + (p-4)\bar E_1 - (p-3)\bar E_2\big)\eta = 0$$

$$\big(\frac{\sigma^2}{2} - \lambda\big)\eta + \sum_\alpha (C_\alpha - D_\alpha)\bar\epsilon_\alpha + 2\big(E_0 + (p-4)E_1 - (p-3)E_2\big)\bar\eta = 0$$

Finally the eigenmodes of the Hessian is made complete by considering a broken symmetry where two replicas a_0 and a_1 are different from the others, with the following parameterization dictated again by orthogonality constraints with the two previous ones:

$$(\epsilon^a_\alpha, \bar\epsilon^a_\alpha) = 0, \qquad (\eta_{ab}, \bar\eta_{ab}) = \begin{cases} (\eta, \bar\eta) & \text{for } a, b \neq a_0 \\ \frac{3-p}{2}(\eta, \bar\eta) & \text{for } a \in a_0, a_1 \text{ or } b \in a_0, a_1 \\ \frac{(p-2)(p-3)}{2}(\eta, \bar\eta) & \text{for } (a,b) = (a_0, a_1). \end{cases}$$

The dimension of this set is now $p(p-3)$, and represents eigenvectors iff the following system of equations is satisfied

$$\Big(\frac{\sigma^2}{2}-\lambda\Big)\bar{\eta}+2(\bar{E}_0-2\bar{E}_1+\bar{E}_2)\eta=0$$

$$\Big(\frac{\sigma^2}{2}-\lambda\Big)\eta+2(E_0-2E_1+E_2)\bar{\eta}=0$$

The corresponding eigenvalues reads

$$\lambda=\frac{\sigma^2}{2}\pm 2\sqrt{(\bar{E}_0-2\bar{E}_1+\bar{E}_2)(E_0-2E_1+E_2)},$$

of degeneracy $p(p-3)/2$. Finally the RS stability condition reads

$$\frac{1}{\sigma^2}>\sqrt{\mathsf{E}_{x,u}\Big(\mathrm{sech}^4\big(h(x,u)\big)\Big)\mathsf{E}_{x,v}\Big(\mathrm{sech}^4\big(\bar{h}(x,v)\big)\Big)},$$

which reduces to the same form of AT line as the SK model when $\kappa=1$, except for the u and v averages not present in the SK model. As seen on the left Figure 2 the influence of κ is very limited.

B Synthetic dataset

The multimodal distribution modeling the N-dimensional synthetic data is

$$P(s)=\sum_{c=1}^{C}p_c\prod_{i=1}^{N}\frac{e^{h_i^c s_i}}{2\cosh(h_i^c)},\tag{47}$$

where C is the number of clusters, p_c is a weight and $\boldsymbol{h}^c$ is a hidden field for cluster c. The values for p_c are taken at random and normalized, while to compute h_i^c we take into account the magnetizations $m_i^c=\tanh(h_i^c)$. Expanding over the spectral modes, we can set an effective dimension d by constraining the sum to the range $\alpha=1,\dots,d$

$$m_i^c=\sum_{\alpha=1}^{d}m_\alpha^c u_{i,\alpha}\tag{48}$$

Clusters' magnetizations m_α^c are drawn at random between $[-1,1]$ and normalized with the factor

$$Z=\sqrt{\frac{\sum_\alpha m_\alpha^2}{d\cdot r}},\quad r=\tanh(\eta)\tag{49}$$

where r is introduced to decrease the clusters' polarizations (in our simulations, we used $\eta=0.3$). The spectral basis $u_{i,\alpha}$ is obtained by drawing at random d N-dimensional vectors and applying the Gram-Schmidt process (which can be safely employed as N is supposedly big and thus the initial vectors are nearly orthogonal). The hidden fields are then obtained from the magnetizations

$$h_i^c=\tanh^{-1}(m_i^c)\tag{50}$$

and the samples are generated by choosing a cluster according to p_c and setting the visible variables to ±1 according to

$$p(s_i=1)=\frac{1}{1+e^{-2h_i^c}}\tag{51}$$

References

[1] P. Smolensky. *In Parallel Distributed Processing: Volume 1 by D. Rumelhart and J. McLelland*, chapter 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory. 194-281. MIT Press, 1986.

[2] R. Salakhutdinov and G. Hinton. Deep Boltzmann machines. In *Artificial Intelligence and Statistics*, pages 448–455, 2009.

[3] G.E. Hinton and R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[4] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14:1771–1800, 2002.

[5] T. Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1064–1071, New York, NY, USA, 2008. ACM.

[6] G.E. Hinton. *A Practical Guide to Training Restricted Boltzmann Machines*, pages 599–619. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[7] D.S.P. Salazar. Nonequilibrium thermodynamics of restricted Boltzmann machines. *Phys. Rev. E*, 96:022131, 2017.

[8] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8):2554–2558, 1982.

[9] D. J. Amit, H. Gutfreund, and H. Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of Physics*, 173(1):30–67, 1987.

[10] E. Gardner. Maximum storage capacity in neural networks. *EPL (Europhysics Letters)*, 4(4):481, 1987.

[11] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271, 1988.

[12] B. Barra, A. Bernacchia, E. Santucci, and P. Contucci. On the equivalence of Hopfield networks and Boltzmann machines. *Neural Networks*, 34:1–9, 2012.

[13] G. Marylou, E.W. Tramel, and F. Krzakala. Training restricted Boltzmann machines via the Thouless-Anderson-Palmer free energy. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 640–648, 2015.

[14] H. Huang and T. Toyoizumi. Advanced mean-field theory of the restricted Boltzmann machine. *Physical Review E*, 91(5):050101, 2015.

[15] C. Takahashi and M. Yasuda. Mean-field inference in gaussian restricted Boltzmann machine. *Journal of the Physical Society of Japan*, 85(3):034001, 2016.

[16] C. Furtlehner, J.-M. Lasgouttes, and A. Auger. Learning multiple belief propagation fixed points for real time inference. *Physica A: Statistical Mechanics and its Applications*, 389(1):149–163, 2010.

[17] A. Barra, G. Genovese, P. Sollich, and D. Tantari. Phase diagram of restricted Boltzmann machines and generalized Hopfield networks with arbitrary priors. arXiv:1702.05882, 2017.

[18] H. Huang. Statistical mechanics of unsupervised feature learning in a restricted Boltzmann machine with binary synapses. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(5):053302, 2017.

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

[20] R. Monasson and J. Tubiana. Emergence of compositional representations in restricted Boltzmann machines. *Phys. Rev. Let.*, 118:138301, 2017.

[21] L. Zdeborová and F. Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[22] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Comput.*, 11(2):443–482, 1999.

[23] H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59(4):291–294, 1988.

[24] A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv:1312.6120, 2014.

[25] A. Decelle, G. Fissore, and C. Furtlehner. Spectral dynamics of learning in restricted Boltzmann machines. *EPL*, 119(6):60001, 2017.

[26] E.W. Tramel, M. Gabrié, A. Manoel, F. Caltagirone, and F. Krzakala. A Deterministic and Generalized Framework for Unsupervised Learning with Restricted Boltzmann Machines. arXiv:1702.03260, 2017.

[27] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[28] M. Mézard. Mean-field message-passing equations in the Hopfield model and its generalizations. *Phys. Rev. E*, 95:022117, 2017.

[29] G. Parisi and M. Potters. Mean-field equations for spin models with orthogonal interaction matrices. *Journal of Physics A: Mathematical and General*, 28(18):5267, 1995.

[30] M. Opper and O. Winther. Adaptive and self-averaging Thouless-Anderson-Palmer mean field theory for probabilistic modeling. *Physical Review E*, 64:056131, 2001.

[31] D. J. Amit, H. Gutfreund, and H. Sompolinsky. Spin-glass models of neural networks. *Phys. Rev. A*, 32:1007–1018, 1985.

[32] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin Glass Theory and Beyond*. World Scientific, Singapore, 1987.

[33] J. R. L. Almeida and D. J. Thouless. Stability of the Sherrington-Kirkpatrick solution of a spin glass model. *J.Phy. A:Math. Gen*, 11(5):983–990, 1978.

[34] P. C. Hohenberg and M. C. Cross. *An introduction to pattern formation in nonequilibrium systems*, pages 55–92. Springer Berlin Heidelberg, Berlin, Heidelberg, 1987.

[35] I. Mastromatteo and M. Marsili. On the criticality of inferred models. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(10):P10012, 2011.